UNIVERSITY OF CALIFORNIA
RIVERSIDE

Non-Coding RNAs Play Significant Roles in Host-Virus Interactions

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Genetics, Genomics and Bioinformatics

by

Lichao Li

December 2019

Dissertation Committee:
Dr. Weifeng Gu, Chairperson
Dr. Xuemei Chen
Dr. Shou-wei Ding

The Dissertation of Lichao Li is approved:

Committee Chairperson

University of California, Riverside

ACKNOWLEDGEMENTS

I would like to express my sincere appreciation to each one who helped me during my PhD studies at the University of California, Riverside (UCR). This work would not have been possible without their generous help.

Firstly, I would like to express my deep gratitude to my PhD advisor, Dr. Weifeng Gu. I am honored to be a member of his lab and his first student to be graduating. He taught me not only all the detailed skills by himself, but also the attitude and motivation to be a scientist.

I would like to thank Dr. Daniel Chaves for the excellent previous work on PIR-1 and Dr. Weifeng Gu who trust me to handle the PIR-1 project. Thank Dr. Darryl Conte who is currently editing the article.

I thank all members in our lab. Ruidong Li guided me a lot when I was a beginner. An-Phong Nguyen, Kathy Yu, Randy Quinn, Xuan Bui and Yesica Ayon were my best helpers for my experiments. I thank other members include Hui Dai, James Randolph, Myoung Ho, Jenna Xu who discussed with me in the lab or during the meetings.

I thank the members of my dissertation committee, Dr. Xuemei Chen and Dr. Shou-wei Ding, for their effort and time in keeping me on track in the past three years. I also thank the following UCR faculty for their guidance in the first two years of my PhD training: Dr. Sika Zheng, Dr. Wenxiu Ma, Dr. Tomas Girke and Dr. Ted Karginov. Their advices gave insights to my research.

I also thank the following friends, Dr. Gina Maduro from Dr. Morris Maduro lab, Dr. Yi Zhai from Dr. Wenbo Ma lab, Dr. Qingxia Han from Dr. Shou-wei Ding lab, Dr. Yike Ding from Dr. Alexander Raikhel lab, for their training on special experiment that I was not familiar with.

Lastly, to my dearest family, my parents and my grandma who always support me to pursuit my Ph.D. degree and keep on going in science field.

ABSTRACT OF THE DISSERTATION

Non-Coding RNAs Play Significant Roles in Host-Virus Interactions

by

Lichao Li

Doctor of Philosophy, Graduate Program in Genetics, Genomics and Bioinformatics
University of California, Riverside, December 2019
Dr. Weifeng Gu, Chairperson

Previous Dicer immunoprecipitation (IP) discovered an RNA polyphosphatase PIR-1 interacting with Dicer, may participate in RNAi but the mechanism is unrevealed. Here we demonstrate that *C. elegans* PIR-1 is involved in the RNAi-mediated silencing of Orsay virus via promoting the biogenesis of 23-mer RNAs and the loading of 23-mer RNAs to RDE-1. We also showed that PIR-1 acts as a *de facto* RNA phosphatase *in vivo* to regulate triphosphorylated RNAs (ppp-RNAs). Thus, PIR-1 is a conserved master regulator of ppp-RNAs and plays important roles in silencing viral ppp-RNAs and modifying cellular ppp-RNAs.

Next we apply PIR-1 in our small RNA cloning strategy. The high-throughput sequencing has become a standard tool for analyzing RNA and DNA. We have developed a new strategy to clone modified/unmodified small RNA in an all-liquid-based reaction carried out in a single PCR tube with as little as 16 ng total RNA. The 7-hour cloning process only needs ~1-hour labor. Moreover, this method can also clone mRNA, simplifying the need to prepare two cloning systems for small RNA and mRNA.

At last, we study the function of non-coding RNA in influenza A virus. It utilizes a special process, cap-snatching, to obtain a host capped small RNA for priming viral mRNA synthesis, generating hybrid capped mRNA for translation. Previous studies have been focusing on cap-snatching at the

5' end of viral mRNA. Here we report two non-canonical cap-snatching regions: one 300-nt upstream of the 3' end of each mRNA generating capped mRNA/ncRNA, and the other in the 5' region of vRNA and mapped primarily at the 2-nt, likely generating ncRNA. We also demonstrate that the influenza virus snatches virus-derived capped RNA in addition to host capped RNA. These findings expand our understanding of the cap-snatching mechanism and suggest that the influenza A virus may utilize this process to diversify its mRNA/ncRNA.

TABLE OF CONTENTS

INTRODUCTION 1
INTRODUCTION 1
REFERENCES 16
CHAPTER 1 22
A Highly Conserved RNA Polyphosphatase PIR-1 Is Required for Development and Antiviral Response in *C. elegans* 22
ABSTRACT 22
INTRODUCTION 23
MATERIALS AND METHODS 24
RESULTS 28
DISCUSSION 45
REFERENCES 50
CHAPTER 2 52
A Convenient Strategy to Clone Modified/unmodified small RNA and mRNA for High Throughput Sequencing 52
ABSTRACT 52
INTRODUCTION 53
MATERIALS AND METHODS 54
RESULTS 55
DISCUSSION 72
REFERENCES 76
CHAPTER 3 78
Influenza A Virus Utilizes Non-Canonical Cap-Snatching to Diversify mRNA/ncRNA Pools . 78
ABSTRACT 78
INTRODUCTION 79
MATERIALS AND METHODS 81
RESULTS 83
DISCUSSION 104
REFERENCES 110
APPENDICES 114
Appendix A 114

Appendix B 115
Appendix C 118
Appendix D 121
Appendix E...122
Appendix F...123
Appendix G 124
Appendix H 126

LIST OF FIGURES

INTRODUCTION

Figure I. 1. The exogenous RNAi pathway 2

Figure I. 2. General RNA cloning protocol 9

Figure I. 3. The 5' and 3' sequences of IAV mRNA, cRNA and vRNA...........................15

CHAPTER 1

Figure 1. 1. PIR-1 is an RNA polyphosphatase...30

Figure 1. 2. PIR-1 is essential for somatic and germline development. 32

Figure 1. 3. PIR-1 is required for suppressing the Orsay virus. 38

Figure 1. 4. PIR-1 promotes the loading of viral primary siRNAs to RDE-1 43

CHAPTER 2

Figure 2. 1. The linker and primer design. 57

Figure 2. 2. Strategy to make a small RNA library. 59

Figure 2. 3. Analysis of sensitivity and reliability of the cloning strategy. 63

Figure 2. 4. Cloning of ppp-RNA and csRNA. 66

Figure 2. 5. Cloning mRNA. 67

Figure 2. 6. Quantification using a native PAGE gel. 71

CHAPTER 3

Figure 3. 1. The IAV RNA levels at varied postinfection timepoints. 84

Figure 3. 2. The histogram of the IAV reads..87

Figure 3. 3. The histogram of vRNA 5' regions. 91

Figure 3. 4. The U2 cap realignment at mRNA +1 and vRNA +2. 98

Figure 3. 5. The cleavage/priming motif of non-canonical cap-snatching. 100

Figure 3. 6. A coding frame is shared by multiple capped RNAs. 103

Figure 3. 7. A unified model for canonical and non-canonical cap-snatching. 105

APPENDICES

Figure S1. 1. PIR-1 is required for suppressing the Orsay virus. Related to figure 1.4...114

Figure S3. 1. Characterization of IAV caps utilized in non-canonical cap-snatching.....121

Figure S3. 2. Verifying non-canonical cap-snatching using RT-PCR............................122

Figure S3. 3. Characterization of non-canonical cap-snatching. 123

Figure S3. 4. The cleavage/priming motif of IAV U2 caps..125

LIST OF TABLES

CHAPTER 1

Table 1. 1. List of PIR-1 interactors obtained in an independent MudPIT experiment....45

CHAPTER 3

Table 3. 1. mRNA cap-snatching rates and ratios from different sites..............................95

APPENDICES

Supplementary File 1. Linkers and Primers...115
Supplementary File 2. A simplified working protocol for RNA cloning 118
Supplementary File 3. IAV protein reads from high-throughput sequencing 126

INTRODUCTION

INTRODUCTION

1. Antiviral RNA interference in *C. elegans*

RNA interference (RNAi) is a gene silencing process that was first revealed in *C. elegans* by Andrew Fire and Craig C. Mello in 1998 (Fire et al., 1998). RNAi is responsible for transposon silencing, gene regulation and antiviral response in most eukaryotes. RNAi process which involved in silencing of foreign genes is called exo-RNAi (exogenous RNAi) pathway. It plays important roles in counteracting RNA viruses in organisms including human (Ashe et al., 2013; Guo & Lu, 2013; Maillard et al., 2013; X.-H. Wang et al., 2006). And RNAi that relates to transposon silencing and self gene regulation is called endo-RNAi (endogenous RNAi) (Merkling & van Rij, 2013; Paddison & Hannon, 2002; Toh et al., 2016). During RNAi process based on small interfering RNAs (siRNAs), double-stranded RNAs (dsRNAs) are cleaved by Dicer into 20-30 nt siRNAs, which interact with various argonaute proteins to form RNA induced silencing complex(RISC), mediate both transcriptional(TGS) and post-transcriptional (PTGS) gene silencing.

1.1 Exo-RNAi Pathway

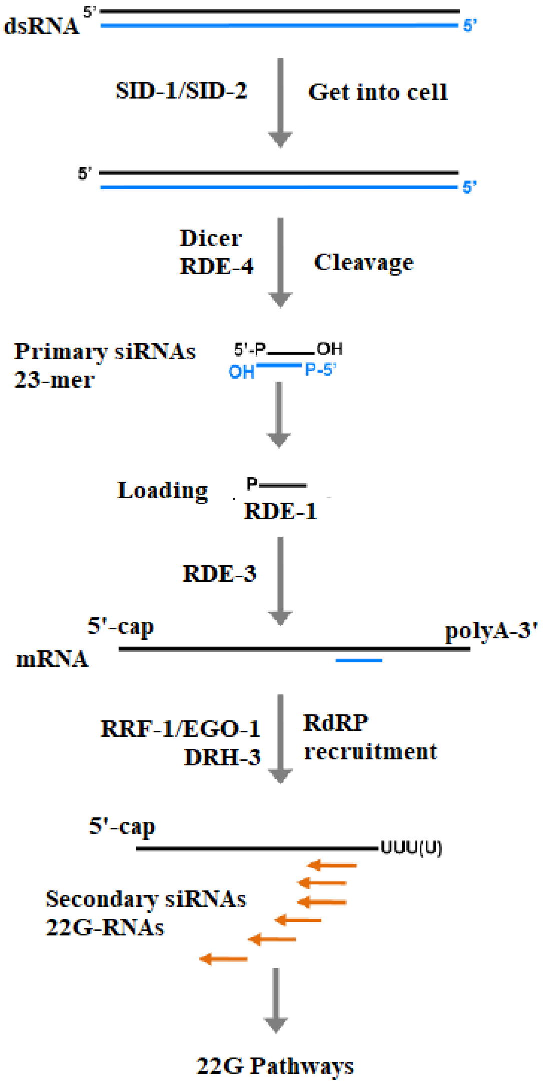


Figure I. 1. The exogenous RNAi pathway

Exo-RNAi is triggered by dsRNAs from the environment which get into cell via SID-2, or injected and passed by SID-1. The primary siRNAs are direct products of Dicer cleavage while the secondary siRNAs, the result of RdRP synthesis triggered by RDE-1-bound primary siRNAs in somatic cells (by RRF-1), and in the germline (by both RRF-1 and EGO-1). Downstream are 22G-RNAs pathways.

Exo-RNAi can be triggered in *C. elegans* by exposing cells to double-stranded RNA (dsRNA) directly by injection or through feeding of dsRNA-expressing bacteria (Figure I.1). The process is strictly dependent on the Argonaute RDE-1, which is loaded with siRNAs coming from the cleavage of dsRNA by Dicer complex. With the dsRNA-binding protein RDE-4 (RNAi-deficient 4; (Hiroaki Tabara, Yigit, Siomi, & Mello, 2002), DCR-1 cuts the dsRNA to generate 23-nt duplex RNAs with 2-nt 3'-hydroxyl overhangs and 5'-monophosphate termini, called 23-mer RNAs(Grishok, 2000; Ketting, 2001; Knight & Bass, 2001). The 23-mer duplexes are loaded into the Argonaute protein RDE-1, which also stably interacts with DCR-1 (H. Tabara et al., 1999). Then RDE-1 cleaves one of the strands (the passenger strand) using its RNase H activity, to leave a single strand that will serve as the guide strand to recognize and bind RNA targets (Steiner, Okihara, Hoogstrate, Sijen, & Ketting, 2009). Target RNA is cleaved near the RDE-1 binding site by the endonuclease RDE-8, followed by the addition of a 3' poly-uridine tract synthesized by the polynucleotidyl transferase RDE-3 (Chen et al., 2005; Tsai et al., 2015). This event lead to the recruitment of the RNA-dependent RNA polymerases (RdRP) RRF-1 (RdRP family 1), and EGO-1 (Enhancer of Glp-One 1) . RdRPs use the RNA template to catalyze the primer-independent synthesis of antisense 22-23 nt siRNAs bearing a 5'-triphosphorylated guanosine, termed 22G-RNAs (Aoki, Moriguchi, Yoshioka, Okawa, & Tabara, 2007; Pak & Fire, 2007; Sijen et al., 2001).

1.2 22G-RNA Pathways

All RNAi pathways, whether they are exogenously or endogenously triggered, both need RdRP-synthesized antisense 22G-RNAs. The regulatory outcome of 22G-RNAs depends mostly on which Argonautes they interact with. There are two main 22G-RNA pathways, the WAGO pathway and the CSR-1 pathway. The simultaneous deletion of the 12 WAGOs leads to a loss of the majority of 22G-RNAs (Claycomb et al., 2009; Gu et al., 2009). Generally, ~50% (~9,000) of all protein-coding genes in *C. elegans* are targeted by 22G-RNAs from both pathways and the majority of endogenous 22G-RNAs are produced in the germline (Gu et al., 2009). The WAGO pathway uses transcriptional and post-transcriptional mechanisms to silence genes and transposons (Gu et al., 2009). The CSR-1 pathway targets essentially all germline-expressed protein coding genes (~4,000), not to silence but promote their expression and protect them from silencing. (Cecere, Hoersch, O'Keeffe, Sachidanandam, & Grishok, 2014; Claycomb et al., 2009; Gu et al., 2009; Seth et al., 2013; Wedeles, Wu, & Claycomb, 2013).

WAGO pathway 22G-RNAs are triggered by primary small RNAs. Since PRG-1 and ALG-3 have been shown to concentrate in nuclear pore-associated P granules, these are thought to be the sites where primary Argonaute binds to their targets in the germline (Batista et al., 2008; Conine et al., 2010). WAGO 22G-RNAs are then amplified in Mutator foci, which harbor EGO-1 RdRP complexes, as well as a variety of additional proteins required for the enrichment of 22G-RNAs (Phillips et al., 2012, Yang et al., 2012; Zhang et al., 2012). In contrast, the production of CSR-1 22G-RNAs does not require Mutator foci (Zhang et al., 2011). Instead, it depends mostly on EGO-1-mediated synthesis in P granules, where both CSR-1 and EGO-1 colocalize (Claycomb et al., 2009; Gu et al., 2009).

The mechanism of 22G-RNA-mediated post-transcriptional silencing is related to RDE-10 and RDE-11. A complex composed of nematode-specific proteins RDE-10 and RDE-11 with potential RNA binding and nuclease properties, respectively, was shown to be required for RDE-1 and ERGO-1-driven 22G-RNA amplification in both germline Mutator foci and the cytoplasm of somatic cells. Notably, RDE-10 was shown to bind exo-RNAi targeted mRNAs, promoting their deadenylation (Yang et al., 2012). The endonuclease RDE-8 was shown to exhibit a similar pattern of localization as RDE-10/RDE-11 and to generally promote the accumulation of 22G-RNAs in the germline and soma (Tsai et al., 2015). In association with the nucleotidyltransferase RDE-3, target RNAs are cleaved by RDE-8 and untemplated uridines are added to the 3' end of the 5'-cleavage fragment. This event may lead to the subsequent degradation of the mRNA fragment, as 3'-uridylation promotes RNA decay (Lee et al., 2014). The degradation activity of these complexes may contribute to the post-transcriptional silencing of exo- and endo-RNAi targets.

At the transcriptional level, silencing occurs via two WAGO Argonautes between the nucleus and the cytoplasm: NRDE-3(WAGO-12) and HRDE-1(WAGO-9). NRDE-3 (Nuclear RNAi-defective) mediates nuclear RNAi in somatic cells, while HRDE-1 (Heritable RNAi-defective) acts in germline nuclei (Guang et al., 2008; Buckley et al., 2012). The association of these Argonautes with 22G-RNAs, causes their translocation into the nucleus. With other factors, namely NRDE-1, NRDE-2 and NRDE-4, they are able to simultaneously interact with chromatin and target transcripts to inhibit Pol II elongation, and subsequently to establish a silent chromatin state through repressive histone H3K9 trimethylation marks (Burkhart et al., 2011; Burton et al., 2011; Ashe et al., 2012; Buckley et al., 2012; Gu, S.G., et al., 2012; Luteijn et al., 2012; Shirayama et al., 2012). Mutants of *hrde-1*, show a mortal germline phenotype (Mrt) at 25˚C in which the ability of the germline to produce oocytes and sperm decreases gradually in a few generations, finally cause animals that are completely sterile (Buckley et al., 2012). And HRDE-1 was demonstrated to direct transgenerational silencing of germline-expressed transcripts targeted by exo-RNAi by maintaining H3K9 trimethylation on the corresponding genes (Ashe et al., 2012; Buckley et al., 2012; Gu, S.G. et al., 2012). Cloning of HRDE-1-associated endogenous 22G-RNAs revealed a large overlap with the 22G-RNAs associated with the WAGO-1 Argonaute (Shirayama et al., 2012). Since WAGO-1 shows a preference for 22G-RNAs targeting transposons and repetitive elements (Gu et al., 2009), the Mrt phenotype of the *hrde-1* mutant is thought to arise from the inability to maintain repressive epigenetic marks on these elements.

The Argonaute CSR-1 (Chromosome Segregation and RNAi-defective) is highly expressed in the germline, oocytes and embryos (Claycomb et al., 2009). As its name indicates, *csr-1* mutants display lethal chromosome segregation defects during mitotic divisions in embryos, exhibit partial resistance to exo-RNAi in the germline, as well as changes in the morphology and proliferation of germline nuclei (Yigit et al, 2006; Claycomb et al., 2009). CSR-1 concentrates throughout the germline in P granules, localizes to meiotic chromosomes of oocytes and later to mitotic chromosomes of developing embryos (Claycomb et al., 2009). CSR-1 22G-RNAs depend strictly on the RdRP EGO-1, DRH-3 and EKL-1, mutants of which exhibit several overlapping phenotypes with *csr-1* (Duchaine et al., 2006; Nakamura et al., 2007; Gu et al., 2009; She et al., 2009). As mentioned earlier, CSR-1 22G-RNAs target germline protein-coding mRNAs without leading to their downregulation (Claycomb et al., 2009).

The CSR-1 pathway has recently been shown to promote Pol II transcription of the germline expressed genes (Cecere et al., 2014). This is consistent with prior observations that CSR-1 regulates the distribution of repressive H3K9 dimethyl chromatin marks (Maine et al., 2005; She et al., 2009). In addition, CSR-1 was shown to counteract the silencing effects of the PRG-1/21U-RNA pathway in the germline (Lee et al., 2012; Seth et al., 2013; Wedeles et al., 2013). These studies demonstrated that both transgenes and endogenous genes targeted by CSR-1 are completely immune to recognition and silencing by PRG-1. This immediately suggested the hypothesis that the CSR-1 and the PRG-1 pathways act simultaneously and in opposition of each other, as part of a system that discriminates “self” from “non-self” gene expression. This system would ensure proper germline function while protecting its genetic material from the deleterious effects of invading or existing foreign sequences.

Lastly, very little is known regarding the events that trigger CSR-1 22G-RNA synthesis. The first clue that some CSR-1 22G-RNAs may also be triggered by primary sRNAs came from a recent study of CSR-1 in males (Conine et al., 2013). CSR-1 was shown to interact with 22G-RNAs derived from ALG-3/4 target mRNAs, and to promote the transcription of those mRNAs in developing spermatocytes. Accordingly, *csr-1* mutant males were shown to have the same temperature sterility phenotype of *alg-3; alg-4* mutant males. When *csr-1* homozygous males were mated to heterozygous hermaphrodites for successive generations, their fertility decreased gradually until the sixth generation, which was completely sterile. Since 22G-RNA-loaded CSR-1 was found in mature sperm, it was proposed that CSR-1 passes along 22G-RNAs that ensure proper expression of ALG-3/4-class genes in the next generation (Conine et al., 2013).

The aforementioned examples illustrate the connections between sRNA pathways in *C. elegans*. The discoveries made thus far in *C. elegans*, especially those that implicated sRNAs as inheritable agents that propagate epigenetic programs from one generation to the next, have propelled an enthusiastic search for similar mechanisms in vertebrates.

2. RNA cloning

The high-throughput sequencing has become a revolutionary technique for analyzing DNA/RNA. As the cost has significantly decreased over the past decade, it is becoming a standard tool for gene analyses in biomedical fields. Therefore, any technical advance in this technique will have a broad impact. Since a single sequencing run is usually sufficient for analyzing multiple samples, individual samples are usually cloned with specific barcodes, pooled, and sequenced as a single library for cost-sharing, and then debarcoded to obtain sample-specific sequences (Gu, Claycomb, Batista, Mello, & Conte, 2011; Stiller, Knapp, Stenzel, Hofreiter, & Meyer, 2009). Although this strategy significantly decreases the sequencing cost, it can not solve the problem that the overall library construction cost using commercial kits could easily surpass the sequencing cost. Usually DNA, mRNA and small RNA sequencing libraries are constructed with distinct linkers and/or chemistry, limiting customers to purchase different kits (Reuter, Spacek, & Snyder, 2015). Lack of compatibility among these expensive kits and transparency for the protocol details often lead to confusion and waste. Most commercial kits only provide just enough primers for a few rounds of library constructions, allowing no mistakes or errors during cloning. Since the linkers for DNA, mRNA, and RNA libraries are different, the corresponding PCR primers for obtaining the final amplicons are also different. In most sequencing platforms, a DNA or cDNA library is made by ligating fragmented DNA/RNA with platform-specific linkers, which are then used to design primers for amplifying and reading sequences/barcodes (Goodwin, McPherson, & Richard McCombie, 2016).

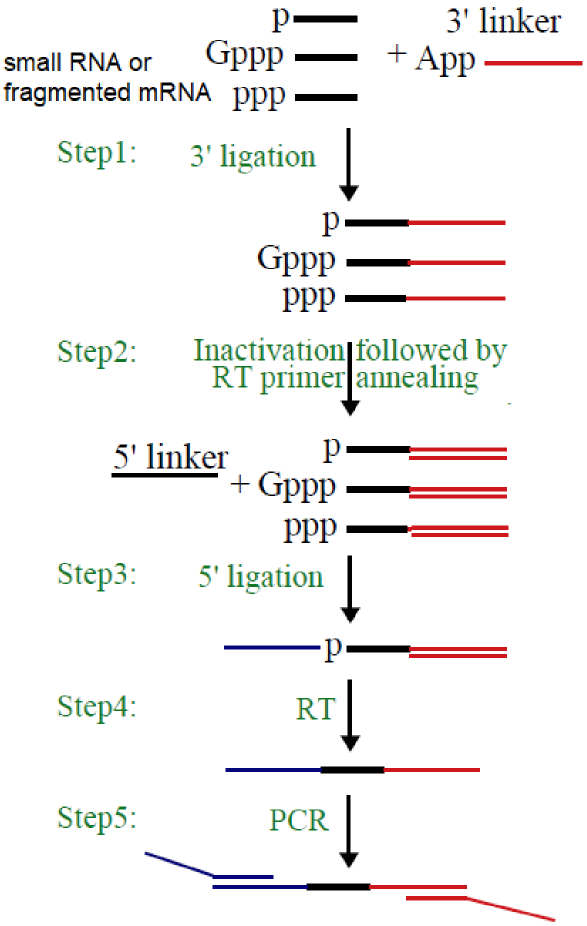


Figure I. 2. General RNA cloning protocol

RNAs were first ligated with a 3’ DNA linker which is 5’ adenylylated. Then ligate to a 5’ RNA linker or anneal with the reverse transcription primer in advance to avoid linker-linker ligation. The DNA-RNA hybrid is then reverse transcribed to cDNA and amplified by PCR using primers that contain barcodes. 5’ modified RNAs have to be enzyme treated to be ligatable.

In the ligation based methods, the 5' ligation needs a 5' monophosphate (p) and the 3' ligation needs a 3'OH on target RNA (Figure I.2). The RNA was first subjected to the ligation of the 18-nt 3' linker. This RNA oligo is “activated” through adenylation of the 5' end (App), which results in extremely efficient ligation and precludes the use of ATP due to the skip of rate-determining step. At the 3' end it possesses a dideoxy cytidine modification (ddC) in order to prevent ligation between linkers. For 5' ligation an RNA linker was used. cDNA synthesis followed, using a DNA oligo complementary to the 3' linker, the oligo can be added in our protocol to avoid free 3’ linker ligate with 5’ linker. Then we can use PCR primers with variable barcode to multiplex several samples in one deep-sequencing run.

Most small RNA species contain modifications at the ends (Batista et al., 2008; Conine et al., 2010; Gu et al., 2012; Kirino & Mourelatos, 2007; Pak & Fire, 2007; Ruby et al., 2006; Taft, Kaplan, Simons, & Mattick, 2009). For example, *C. elegans* 22G-RNA (22G) bears a 5' triphosphorylation, which is incompatible with 5' ligation (Gu et al., 2009; Pak & Fire, 2007); capped small RNA or promoter-associated small RNA has a 5' cap, which is also incompatible with 5' ligation (Gu et al., 2012; Taft et al., 2009). Strategies including enzymatic treatments have been developed to remove these modifications or overcome the inhibitory effects.

For cloning of 5'-monophosphorylated sRNAs, sRNAs are not subjected to any enzymatic treatment prior to 5' adaptor ligation. The direct cloning method is suited for analysis of Dicer products such as miRNAs, primary sRNAs (including 26G-RNAs and viral-23-mers), and 21U-RNAs.

For 5'-triphosphorylated sRNAs, RNA was treated with Tobacco Acid Pyrophosphatase (TAP), which removes 5' cap moieties and terminal phosphates, leaving only one 5'-terminal phosphate group. Since it does not alter sRNAs that were originally 5'-monophosphorylated, this approach clones the widest range of sRNA types. This method constitutes an improvement from an earlier method which completely dephosphorylated all RNAs using CIP, followed by the addition of one 5' phosphate by T4 Polynucleotide Kinase (PNK). This had the disadvantage of allowing the cloning of unphosphorylated RNA degradation fragments, lowering the representation of sequences of interest in the sRNA libraries. The use of TAP minimizes this problem.

Also cloning strategies have been developed to adopt a bypass mechanism to avoid a direct 5' RNA ligation. For example, after 3' ligation, RNA is converted to cDNA and then cDNA is ligated with a 3' linker, which corresponds to a 5' linker ligated to RNA (Pak and Fire 2007). A second bypass mechanism uses cDNA circularization and then the 3' linker is cut into two parts: one flanking the 5' end and the other flanking the 3' end of insert cDNA (Kwon, 2011). The third bypass mechanism utilizes the template switch activity of reverse transcriptase to add a 3' linker to cDNA, which corresponds to a 5' linker ligated to RNA (Ko & Lee, 2006). These bypass mechanisms can overcome the ligation difficulty associated with the RNA 5' modifications since the corresponding 3' end of cDNA does not have any modification.

3. Influenza A virus

Influenza A virus (IAV) causes influenza in birds and mammals. They are negative-sense, single-stranded, segmented RNA viruses. most subtypes are labeled according to an H (Hemagglutinin) and an N number (Neuraminidase). There are 18 different known H antigens and 11 different known N antigens (Tong et al., 2013). Each subtype mutated to multiple strains with different pathogenic profiles.

3.1 Structure and genetics

The virions of all IAVs are similar in composition. They are all made up of a viral envelope containing two main types of proteins, wrapped around a central core. The two proteins around the viral particles are hemagglutinin (HA) and neuraminidase (NA). HA is a protein that helps binding of virus and entry of the viral genome into the target cell. NA is related to release from the attachment sites present in mucus and infected cells (Cohen et al., 2013; Suzuki, 2005). These proteins are the targets for both host antibodies and antiviral drugs.

The entire Influenza A virus genome is 13,588 bps long and contains eight negative sense RNA segments that code for at least 10 but up to 14 proteins. The relevance or presence of alternate gene products are various (Eisfeld, Neumann, & Kawaoka, 2015). Segment 1 encodes RNA polymerase subunit (PB2). Segment 2 encodes RNA polymerase subunit (PB1) and the PB1-F2 protein, which induces cell death, by shift of reading frames from the same RNA segment. Segment 3 encodes RNA polymerase subunit (PA) and the PA-X protein, which relate to host transcription shutoff (Khaperskyy, Schmaling, Larkins-Ford, McCormick, & Gaglia, 2016). Segment 4 encodes for HA. About 500 molecules of HA are needed to form a virion. HA determines the extent and severity of the infection in the host. Segment 5 encodes NP, which is a nucleoprotein. Segment 6 encodes NA. There are about 100 molecules of NA on a virus. Segment 7 encodes two matrix proteins (M1 and M2) by using different reading frames. Segment 8 encodes two distinct non-structural proteins (NS1 and NEP) by using different reading frames from the same RNA segment.

The RNA segments of the viral genome have complementary base sequences at the terminal ends (Suzuki, 2005). IAV utilizes a viral RNA-dependent RNA polymerase (RdRP) complex to generate positive-sense mRNA and complementary RNA (cRNA) from template vRNA, and negative-sense vRNA from template cRNA (Kobayashi et al. 1996; Shi et al. 1995; Shih and Krug 1996; Bouvier and Palese 2008; Guilligay et al. 2008; Sugiyama et al. 2009; Reich et al. 2014). vRNA and cRNA are exactly reverse complementary and both bear 5' triphosphate (ppp) but no poly(A) tail (Bouvier and Palese 2008; Desselberger et al. 1980). In contrast, IAV mRNA is a hybrid RNA composed of a host capped small RNA, IAV-encoded sequence, and poly (A) tail obtained via a stuttering mechanism.

IAV RdRP is composed of three subunits including polymerase basic protein 1 (PB1), polymerase basic protein 2 (PB2), and polymerase acidic protein (PA) (Guilligay et al., 2008; Kobayashi, Toyoda, & Ishihama, 1996; Reich et al., 2014; Shih & Krug, 1996; Sugiyama et al., 2009). Transcription of the viral RNA (vRNA) can only initiate after the PB2 protein binds to host capped RNAs, recruit the PA subunit to cleave several nucleotides after the cap. This host-derived cap and following nucleotides are the primers for viral transcription initiation. Transcription proceeds along the vRNA until it reaches poly(U), initiating a 'stuttering' where the nascent viral mRNA is poly-adenylated, making a mature transcript for nuclear export and translation by host machinery (Velthuis, te Velthuis, & Fodor, 2016). Most IAV mRNA molecules are one nucleotide (nt) shorter at the 5' end than corresponding cRNA molecules since the RdRP usually initiates IAV mRNA synthesis using the penultimate nt of template vRNA (Bouloy, Plotch, & Krug, 1978; Krug, Broni, & Bouloy, 1979; Poon, Pritlove, Fodor, & Brownlee, 1999; Pritlove, Poon, Fodor, Sharps, & Brownlee, 1998).

Once the viral mRNA is made and translated, the viral proteins are assembled into virions and leave the nucleus towards the cell membrane. The host cell membrane has groups of viral transmembrane proteins (HA, NA, and M2) and an layer of the M1 protein which assists the assembled virions to budding through the membrane, releasing fresh mature viruses into the extracellular area (Smith, 2004).

3.2 Cap-snatching mechanism

To obtain a host cap, IAV utilizes cap-snatching, in which 1) PB2 binds host capped RNA; 2) PA cleaves at positions 10-15 nt from the 5' end; and 3) the RdRP utilizes the last nt, usually G of a host cap to base-pair with the penultimate nt, always C, of a template vRNA to prime mRNA synthesis(Bouloy et al., 1978; Bouvier & Palese, 2008; Dias et al., 2009; Fodor et al., 2002;

Guilligay et al., 2008; Kobayashi et al., 1996; Krug et al., 1979; Plotch, Bouloy, & Krug, 1979; Reich et al., 2014; Shi, Summers, Peng, & Galarz, 1995; Sugiyama et al., 2009; Yuan et al., 2009). Although this G appears as encoded by a template C, it actually belongs to host capped RNA (Beaton & Krug, 1981; Bouvier & Palese, 2008; Hagen, Tiley, Chung, & Krystal, 1995; Rao, Yuan, & Krug, 2003). ~20% of IAV mRNA contains additional nts, which appear as a small repeat of IAV mRNA 5' UTR sequences, between a host cap and virus-coded sequence, as estimated using U1/U2-cap-containing IAV mRNA (Decroly, Ferron, Lescar, & Canard, 2011; Gu et al., 2015; Koppstein, Ashour, & Bartel, 2015). At least three models may account for these additional nts. In the currently preferred model 'prime-cis-realignment' or 'prime-realignment', 1) a cap is cleaved by IAV RdRP, annealed to a template vRNA using the base-pairing between the cap last nt, usually G, and the template penultimate nt, C, and extended up to 9-nt (predominantly 4 or less); and 2) a fraction of nascent mRNA detaches from the template, reanneals with the template using the base-pairing between the mRNA last nt, usually A, and the template last nt, U, and then is extended again (Gu et al., 2015; Koppstein et al., 2015; Te Velthuis & Oymans, 2018). Consistent with this model, the cap cleaved by PA usually prefers G as the last nt and the first extension predominantly ends with A, perfectly matching the 3'UC---5' template sequence (Gu et al., 2015; Koppstein et al., 2015). A second model 'prime-random-realignment' assumes that the first step utilizes the same annealing/extension mechanism above but in the second step, the released nascent mRNA anneals with both cis and trans templates. Although there has been no evidence supporting this model, it generates similar results as the first model because the 5'UTRs of IAV mRNAs are almost identical (Figure I.3).

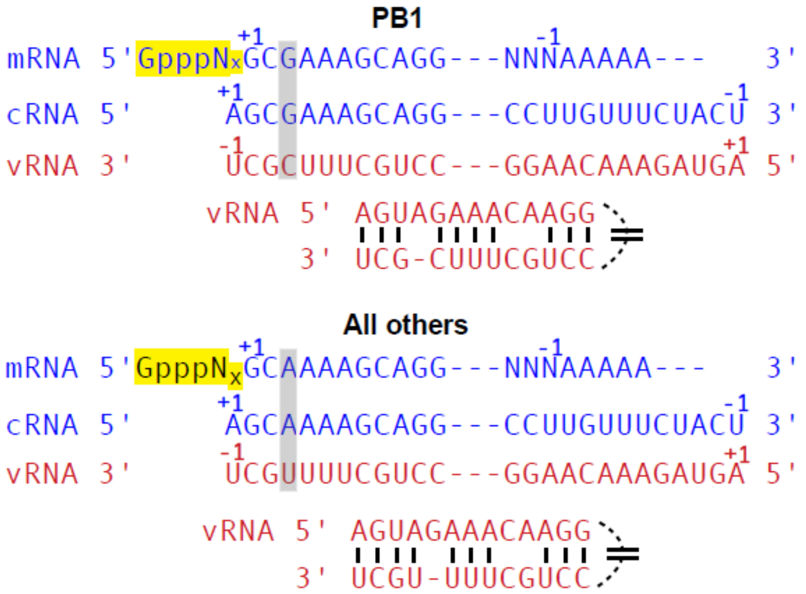

PB1
+1
-1
mRNA 5'GpppNxGCGAAAGCAGG---NNNAAAAA--- 3'
+1
-1
cRNA 5' AGCGAAAGCAGG---CCUUGUUUCUACU 3'
-1
+1
vRNA 3' UCGCUUUCGUCC---GGAACAAAGAUGA 5'
vRNA 5' AGUAGAAACAAGG
3' UCG-CUUUCGUCC
All others
+1
-1
mRNA 5'GpppNxGCAAAAGCAGG---NNNAAAAA--- 3'
+1
-1
cRNA 5' AGCAAAAGCAGG---CCUUGUUUCUACU 3'
-1
+1
vRNA 3' UCGUUUUCGUCC---GGAACAAAGAUGA 5'
vRNA 5' AGUAGAAACAAGG
3' UCGU-UUUCGUCC

Figure I. 3. The 5' and 3' sequences of IAV mRNA, cRNA and vRNA.

Unlike other RNAs, PB1 contains a single nt variation at the 5' end of mRNA and cRNA and 3' end of vRNA, as highlighted in gray. The 5' and 3' ends of each vRNA is able to form a stem structure, a.k.a., 'panhandle'. The IAV coded part of each RNA is read from 5' to 3' as +1, +2, etc and from 3' to 5' as -1, -2 and etc.

REFERENCES

Aoki, K., Moriguchi, H., Yoshioka, T., Okawa, K., & Tabara, H. (2007). In vitro analyses of the production and activity of secondary small interfering RNAs in *C. elegans*. *The EMBO Journal*, *26*(24), 5007–5019.

Ashe, A., Bélicard, T., Le Pen, J., Sarkies, P., Frézal, L., Lehrbach, N. J., … Miska, E. A. (2013). A deletion polymorphism in the Caenorhabditis elegans RIG-I homolog disables viral RNA dicing and antiviral immunity. *eLife*, *2*, e00994.

Batista, P. J., Graham Ruby, J., Claycomb, J. M., Chiang, R., Fahlgren, N., Kasschau, K. D., … Mello, C. C. (2008). PRG-1 and 21U-RNAs Interact to Form the piRNA Complex Required for Fertility in *C. elegans*. *Molecular Cell*, Vol. 31, pp. 67–78. https://doi.org/10.1016/j.molcel.2008.06.002

Beaton, A. R., & Krug, R. M. (1981). Selected host cell capped RNA fragments prime influenza viral RNA transcription in vivo. *Nucleic Acids Research*, *9*(17), 4423–4436.

Bouloy, M., Plotch, S. J., & Krug, R. M. (1978). Globin mRNAs are primers for the transcription of influenza viral RNA in vitro. *Proceedings of the National Academy of Sciences of the United States of America*, *75*(10), 4886–4890.

Bouvier, N. M., & Palese, P. (2008). The biology of influenza viruses. *Vaccine*, *26 Suppl 4*, D49–D53.

Cecere, G., Hoersch, S., O'Keeffe, S., Sachidanandam, R., & Grishok, A. (2014). Global effects of the CSR-1 RNA interference pathway on the transcriptional landscape. *Nature Structural & Molecular Biology*, *21*(4), 358–365.

Chen, C.-C. G., Simard, M. J., Tabara, H., Brownell, D. R., McCollough, J. A., & Mello, C. C. (2005). A Member of the Polymerase *β* Nucleotidyltransferase Superfamily Is Required for RNA Interference in *C. elegans*. *Current Biology*, Vol. 15, pp. 378–383. https://doi.org/10.1016/j.cub.2005.01.009

Claycomb, J. M., Batista, P. J., Pang, K. M., Gu, W., Vasale, J. J., van Wolfswinkel, J. C., … Mello, C. C. (2009). The Argonaute CSR-1 and its 22G-RNA cofactors are required for holocentric chromosome segregation. *Cell*, *139*(1), 123–134.

Cohen, M., Zhang, X.-Q., Senaati, H. P., Chen, H.-W., Varki, N. M., Schooley, R. T., & Gagneux, P. (2013). Influenza A penetrates host mucus by cleaving sialic acids with neuraminidase. *Virology Journal*, *10*, 321.

Conine, C. C., Batista, P. J., Gu, W., Claycomb, J. M., Chaves, D. A., Shirayama, M., & Mello, C. C. (2010). Argonautes ALG-3 and ALG-4 are required for spermatogenesis-specific 26G-RNAs and thermotolerant sperm in Caenorhabditis elegans. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(8), 3588–3593.

Decroly, E., Ferron, F., Lescar, J., & Canard, B. (2011). Conventional and unconventional mechanisms for capping viral mRNA. *Nature Reviews. Microbiology*, *10*(1), 51–65.

Dias, A., Bouvier, D., Crépin, T., McCarthy, A. A., Hart, D. J., Baudin, F., … Ruigrok, R. W. H. (2009). The cap-snatching endonuclease of influenza virus polymerase resides in the PA subunit. *Nature*, *458*(7240), 914–918.

Eisfeld, A. J., Neumann, G., & Kawaoka, Y. (2015). At the centre: influenza A virus ribonucleoproteins. *Nature Reviews. Microbiology*, *13*(1), 28–41.

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*, Vol. 391, pp. 806–811. https://doi.org/10.1038/35888

Fodor, E., Crow, M., Mingay, L. J., Deng, T., Sharps, J., Fechter, P., & Brownlee, G. G. (2002). A single amino acid mutation in the PA subunit of the influenza virus RNA polymerase inhibits endonucleolytic cleavage of capped RNAs. *Journal of Virology*, *76*(18), 8989–9001.

Goodwin, S., McPherson, J. D., & Richard McCombie, W. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, Vol. 17, pp. 333–351. https://doi.org/10.1038/nrg.2016.49

Grishok, A. (2000). Genetic Requirements for Inheritance of RNAi in *C. elegans*. *Science*, Vol. 287, pp. 2494–2497. https://doi.org/10.1126/science.287.5462.2494

Guilligay, D., Tarendeau, F., Resa-Infante, P., Coloma, R., Crepin, T., Sehr, P., … Cusack, S. (2008). The structural basis for cap binding by influenza virus polymerase subunit PB2. *Nature Structural & Molecular Biology*, *15*(5), 500–506.

Guo, X., & Lu, R. (2013). Characterization of Virus-Encoded RNA Interference Suppressors in Caenorhabditis elegans. *Journal of Virology*, Vol. 87, pp. 5414–5423. https://doi.org/10.1128/jvi.00148-13

Gu, W., Claycomb, J. M., Batista, P. J., Mello, C. C., & Conte, D. (2011). Cloning Argonaute-associated small RNAs from *Caenorhabditis elegans*. *Methods in Molecular Biology* , *725*, 251–280.

Gu, W., Gallagher, G. R., Dai, W., Liu, P., Li, R., Trombly, M. I., … Finberg, R. W. (2015). Influenza A virus preferentially snatches noncoding RNA caps. *RNA* , *21*(12), 2067–2075.

Gu, W., Lee, H.-C., Chaves, D., Youngman, E. M., Pazour, G. J., Conte, D., & Mello, C. C. (2012). CapSeq and CIP-TAP Identify Pol II Start Sites and Reveal Capped Small RNAs as piRNA Precursors. *Cell*, Vol. 151, pp. 1488–1500. https://doi.org/10.1016/j.cell.2012.11.023

Gu, W., Shirayama, M., Conte, D., Vasale, J., Batista, P. J., Claycomb, J. M., … Mello, C. C. (2009). Distinct Argonaute-Mediated 22G-RNA Pathways Direct Genome Surveillance in the Germline. *Molecular Cell*, Vol. 36, pp. 231–244. https://doi.org/10.1016/j.molcel.2009.09.020

Hagen, M., Tiley, L., Chung, T. D., & Krystal, M. (1995). The role of template-primer interactions in cleavage and initiation by the influenza virus polymerase. *The Journal of General Virology*, *76 (Pt 3)*, 603–611.

Ketting, R. F. (2001). Dicer functions in RNA interference and in synthesis of small RNA involved in developmental timing in . *Genes & Development*, Vol. 15, pp. 2654–2659. https://doi.org/10.1101/gad.927801

Khaperskyy, D. A., Schmaling, S., Larkins-Ford, J., McCormick, C., & Gaglia, M. M. (2016). Selective Degradation of Host RNA Polymerase II Transcripts by Influenza A Virus PA-X Host Shutoff Protein. *PLOS Pathogens*, Vol. 12, p. e1005427. https://doi.org/10.1371/journal.ppat.1005427

Kirino, Y., & Mourelatos, Z. (2007). Mouse Piwi-interacting RNAs are 2’-O-methylated at their 3' termini. *Nature Structural & Molecular Biology*, *14*(4), 347–348.

Knight, S. W., & Bass, B. L. (2001). A role for the RNase III enzyme DCR-1 in RNA interference and germ line development in Caenorhabditis elegans. *Science*, *293*(5538), 2269–2271.

Kobayashi, M., Toyoda, T., & Ishihama, A. (1996). Influenza virus PB1 protein is the minimal and essential subunit of RNA polymerase. *Archives of Virology*, *141*(3-4), 525–539.

Ko, J.-H., & Lee, Y. (2006). RNA-conjugated template-switching RT-PCR method for generating an Escherichia coli cDNA library for small RNAs. *Journal of Microbiological Methods*, *64*(3), 297–304.

Koppstein, D., Ashour, J., & Bartel, D. P. (2015). Sequencing the cap-snatching repertoire of H1N1 influenza provides insight into the mechanism of viral transcription initiation. *Nucleic Acids Research*, *43*(10), 5052–5064.

Krug, R. M., Broni, B. A., & Bouloy, M. (1979). Are the 5’ ends of influenza viral mRNAs synthesized in vivo donated by host mRNAs? *Cell*, *18*(2), 329–334.

Kwon, Y.-S. (2011). Small RNA library preparation for next-generation sequencing by single ligation, extension and circularization technology. *Biotechnology Letters*, *33*(8), 1633–1641.

Maillard, P. V., Ciaudo, C., Marchais, A., Li, Y., Jay, F., Ding, S. W., & Voinnet, O. (2013). Antiviral RNA Interference in Mammalian Cells. *Science*, Vol. 342, pp. 235–238. https://doi.org/10.1126/science.1241930

Merkling, S. H., & van Rij, R. P. (2013). Beyond RNAi: antiviral defense strategies in Drosophila and mosquito. *Journal of Insect Physiology*, *59*(2), 159–170.

Paddison, P. J., & Hannon, G. J. (2002). RNA interference: the new somatic cell genetics? *Cancer Cell*, *2*(1), 17–23.

Pak, J., & Fire, A. (2007). Distinct populations of primary and secondary effectors during RNAi in . *Science*, *315*(5809), 241–244.

Plotch, S. J., Bouloy, M., & Krug, R. M. (1979). Transfer of 5'-terminal cap of globin mRNA to influenza viral complementary RNA during transcription in vitro. *Proceedings of the National Academy of Sciences of the United States of America*, *76*(4), 1618–1622.

Poon, L. L., Pritlove, D. C., Fodor, E., & Brownlee, G. G. (1999). Direct evidence that the poly(A) tail of influenza A virus mRNA is synthesized by reiterative copying of a U track in the virion RNA template. *Journal of Virology*, *73*(4), 3473–3476.

Pritlove, D. C., Poon, L. L., Fodor, E., Sharps, J., & Brownlee, G. G. (1998). Polyadenylation of influenza virus mRNA transcribed in vitro from model virion RNA templates: requirement for 5' conserved sequences. *Journal of Virology*, *72*(2), 1280–1286.

Rao, P., Yuan, W., & Krug, R. M. (2003). Crucial role of CA cleavage sites in the cap-snatching mechanism for initiating viral mRNA synthesis. *The EMBO Journal*, *22*(5), 1188–1198.

Reich, S., Guilligay, D., Pflug, A., Malet, H., Berger, I., Crépin, T., … Cusack, S. (2014). Structural insight into cap-snatching and RNA synthesis by influenza polymerase. *Nature*, *516*(7531), 361–366.

Reuter, J. A., Spacek, D. V., & Snyder, M. P. (2015). High-throughput sequencing technologies. *Molecular Cell*, *58*(4), 586–597.

Ruby, J. G., Graham Ruby, J., Jan, C., Player, C., Axtell, M. J., Lee, W., … Bartel, D. P. (2006). Large-Scale Sequencing Reveals 21U-RNAs and Additional MicroRNAs and Endogenous siRNAs in . *Cell*, Vol. 127, pp. 1193–1207. https://doi.org/10.1016/j.cell.2006.10.040

Seth, M., Shirayama, M., Gu, W., Ishidate, T., Conte, D., & Mello, C. C. (2013). The *C. elegans* CSR-1 Argonaute Pathway Counteracts Epigenetic Silencing to Promote Germline Gene Expression. *Developmental Cell*, Vol. 27, pp. 656–663. https://doi.org/10.1016/j.devcel.2013.11.014

Shih, S. R., & Krug, R. M. (1996). Surprising function of the three influenza viral polymerase proteins: selective protection of viral mRNAs against the cap-snatching reaction catalyzed by the same polymerase proteins. *Virology*, *226*(2), 430–435.

Shi, L., Summers, D. F., Peng, Q., & Galarz, J. M. (1995). Influenza A virus RNA polymerase subunit PB2 is the endonuclease which cleaves host cell mRNA and functions only as the trimeric enzyme. *Virology*, *208*(1), 38–47.

Sijen, T., Fleenor, J., Simmer, F., Thijssen, K. L., Parrish, S., Timmons, L., … Fire, A. (2001). On the role of RNA amplification in dsRNA-triggered gene silencing. *Cell*, *107*(4), 465–476.

Smith, A. E. (2004). How Viruses Enter Animal Cells. *Science*, Vol. 304, pp. 237–242. https://doi.org/10.1126/science.1094823

Steiner, F. A., Okihara, K. L., Hoogstrate, S. W., Sijen, T., & Ketting, R. F. (2009). RDE-1 slicer activity is required only for passenger-strand cleavage during RNAi in *Caenorhabditis elegans*. *Nature Structural & Molecular Biology*, *16*(2), 207–211.

Stiller, M., Knapp, M., Stenzel, U., Hofreiter, M., & Meyer, M. (2009). Direct multiplex sequencing (DMPS)--a novel method for targeted high-throughput sequencing of ancient and highly degraded DNA. *Genome Research*, *19*(10), 1843–1848.

Sugiyama, K., Obayashi, E., Kawaguchi, A., Suzuki, Y., Tame, J. R. H., Nagata, K., & Park, S.-Y. (2009). Structural insight into the essential PB1-PB2 subunit contact of the influenza virus RNA polymerase. *The EMBO Journal*, *28*(12), 1803–1811.

Suzuki, Y. (2005). Sialobiology of influenza: molecular mechanism of host range variation of influenza viruses. *Biological & Pharmaceutical Bulletin*, *28*(3), 399–408.

Tabara, H., Sarkissian, M., Kelly, W. G., Fleenor, J., Grishok, A., Timmons, L., … Mello, C. C. (1999). The rde-1 gene, RNA interference, and transposon silencing in *C. elegans*. *Cell*, *99*(2), 123–132.

Tabara, H., Yigit, E., Siomi, H., & Mello, C. C. (2002). The dsRNA Binding Protein RDE-4 Interacts with RDE-1, DCR-1, and a DExH-Box Helicase to Direct RNAi in *C. elegans*. *Cell*, Vol. 109, pp. 861–871. https://doi.org/10.1016/s0092-8674(02)00793-6

Taft, R. J., Kaplan, C. D., Simons, C., & Mattick, J. S. (2009). Evolution, biogenesis and function of promoter-associated RNAs. *Cell Cycle* , *8*(15), 2332–2338.

Te Velthuis, A. J. W., & Oymans, J. (2018). Initiation, Elongation, and Realignment during Influenza Virus mRNA Synthesis. *Journal of Virology*, *92*(3). https://doi.org/10.1128/JVI.01775-17

Toh, C.-X. D., Chan, J.-W., Chong, Z.-S., Wang, H. F., Guo, H. C., Satapathy, S., … Loh, Y.-H. (2016). RNAi Reveals Phase-Specific Global Regulators of Human Somatic Cell Reprogramming. *Cell Reports*, *15*(12), 2597–2607.

Tong, S., Zhu, X., Li, Y., Shi, M., Zhang, J., Bourgeois, M., … Donis, R. O. (2013). New World Bats Harbor Diverse Influenza A Viruses. *PLoS Pathogens*, Vol. 9, p. e1003657. https://doi.org/10.1371/journal.ppat.1003657

Tsai, H.-Y., Chen, C.-C. G., Conte, D., Jr, Moresco, J. J., Chaves, D. A., Mitani, S., … Mello, C. C. (2015). A ribonuclease coordinates siRNA amplification and mRNA cleavage during RNAi. *Cell*, *160*(3), 407–419.

Velthuis, A. J. W. te, te Velthuis, A. J. W., & Fodor, E. (2016). Influenza virus RNA polymerase: insights into the mechanisms of viral RNA synthesis. *Nature Reviews Microbiology*, Vol. 14, pp. 479–493. https://doi.org/10.1038/nrmicro.2016.87

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

Wedeles, C. J., Wu, M. Z., & Claycomb, J. M. (2013). Protection of Germline Gene Expression by the *C. elegans* Argonaute CSR-1. *Developmental Cell*, Vol. 27, pp. 664–671. https://doi.org/10.1016/j.devcel.2013.11.016

Yuan, P., Bartlam, M., Lou, Z., Chen, S., Zhou, J., He, X., … Liu, Y. (2009). Crystal structure of an avian influenza polymerase PA(N) reveals an endonuclease active site. *Nature*, *458*(7240), 909–913.

CHAPTER 1

A Highly Conserved RNA Polyphosphatase PIR-1 Is Required for Development and Antiviral Response in *C. elegans*

ABSTRACT

Previous Dicer immunoprecipitation (IP) discovered an RNA polyphosphatase PIR-1 interacting with Dicer, may participate in RNAi but the mechanism is unrevealed. Here we demonstrate that *C. elegans* PIR-1 is involved in the RNAi-mediated silencing of Orsay virus via promoting the biogenesis of 23-mer RNAs and the loading of 23-mer RNAs to RDE-1. We also showed that a catalytically-dead PIR-1 strain has the same growth defects as the null mutants, suggesting that PIR-1 acts as a *de facto* RNA phosphatase *in vivo* to regulate 5' triphosphorylated RNAs (ppp-RNAs). This is consistent with our *in vitro* observation that the recombinant wild type (WT) PIR-1 is an RNA polyphosphatase while the catalytically-dead PIR-1 cannot dephosphorylate ppp-RNAs but binds them tightly. PIR-1 immunoprecipitation analysis suggests that PIR-1 interacts with Dicer and other proteins of the ERI complex throughout development. In conclusion, PIR-1 is a conserved master regulator of ppp-RNAs and plays important roles in silencing viral ppp-RNAs and modifying cellular ppp-RNAs. Insights obtained from this study may be transformed into new tools for gene regulation and antivirus.

INTRODUCTION

RNA interference (RNAi) plays important role in regulating genes in all domains of life. In *C. elegans*, during RNAi process based on small interfering RNAs (siRNAs), double stranded RNAs (dsRNAs) are cleaved by Dicer into 20-30 nt siRNAs, which interact with various argonaute proteins to form RNA induced silencing complex(RISC), mediate both transcriptional(TGS) and posttranscriptional (PTGS) gene silencing. Previous proteomic survey of Dicer not only confirmed previously known Dicer interactors, but also revealed that the worm homolog of a human protein named PIR-1 (Phosphatase that Interacts with RNA and Ribonucleoprotein Particle 1) abundantly copurified with Dicer (Duchaine et al. 2006).

Previous study suggested that PIR-1 may belongs to a group of Dual Specificity Protein Phosphatases (Deshpande et al. 1999). However, several studies have demonstrated that *C. elegans* PIR-1 homologs, Baculovirus RNA 5'-triphosphatase (BVP) and human PIR-1 (hPIR-1), have a novel RNA polyphosphatase activity, which removes *β* and *γ* phosphates from triphosphorylated RNAs (ppp-RNAs) (Changela et al. 2005; Sankhala, Lokareddy, and Cingolani 2014). *C. elegans* PIR-1 also shares significant sequence similarity with the triphosphatase domain of mRNA capping enzymes, which only removes *γ*-phosphate from 5' triphosphorylated pre-mRNAs. Therefore, PIR-1 may belong to either the novel polyphosphatase family or the triphosphatase family.

To investigate the antiviral role of PIR-1, we introduce Orsay virus into *C. elegans*. Orsay virus is an RNA virus closely related to Nodaviruses (Félix et al. 2011). Its genome contains two positive single-stranded RNAs (ssRNA), RNA1 and 2, both of which are likely 5' capped and 3' non-polyadenylated (Franz et al. 2014). RNA1 encodes an RdRP for transcription and replication, both generating triphosphorylated double-stranded RNA intermediates. RNA2 encodes a capsid protein alpha and a novel protein delta (Jiang et al. 2014). Orsay virus only infects a few intestinal cells and cannot be vertically transmitted through germlines. Somatic RNAi provides robust anti-Orsay viral responses in which both primary (23-mer RNAs) and secondary siRNAs (22G-RNAs) are generated (Ashe et al. 2013; Guo and Lu 2013). Observing that *pir-1* null mutant promotes Orsay virus replication in comparison to wild type, we presumed that PIR-1 may function within the antiviral RNAi response.

MATERIALS AND METHODS

Purification of PIR-1 protein

The coding sequence of *C. elegans* PIR-1 gene, T23G7.5a.1, was integrated into pET-28a between the NdeI and BamHI restriction sites. The resulting plasmid was transformed into BL21 (DE3)-RIL strain for protein expression. To express the recombinant PIR-1, a single colony was grown in 5 ml Terrific Broth (TB) medium containing 50 μg/ml Kanamycin and 20 μg/ml chloramphenicol at 37 °C for 8 hours; the whole culture was inoculated into 1 L TB media containing 50 μg/ml Kanamycin and 20 μg/ml chloramphenicol to grow ~ 24 hours at room temperature, reaching OD600 0.5; and the protein expression was induced using 0.5 mM IPTG at 16 °C for 15 hours. The cells were pelleted and re-suspended in 25 ml lysis/wash buffer containing 50 mM Tris-HCl (pH 7.5), 0.5 M NaCl, 5 mM 2-mercaptoethanol, 5% glycerol, 10 mM imidazole, 0.01% NP40, and 1 mM PMSF. The cells were sonicated on ice 20 using 30 cycles of 20-second sonication followed by 40-second pause. The resulting solution was centrifuged at 20,000 g at 4 ºC for 10 minutes and the supernatant was mixed with 2.5 ml HisPur Ni-NTA (Thermo Fisher Scientific) beads which were prewashed 3 times with the lysis/wash buffer without PMSF. After 1-hour nutation at 4 ºC, the beads were washed with 200 ml lysis/wash buffer in a 20 ml column and the protein was eluted using 10 of 0.5 ml lysis/wash buffer containing 0.4 M imidazole. The fraction 5-7, which contained the majority of the protein, were pooled and loaded onto a HiPrep Sephacryl S-100 HR column and was fractionated with the imidazole-free lysis/wash buffer using the NGC Quest 100 Plus Chromatography System (Bio-Rad). The FPLC fractions contained a recombinant PIR-1 of high homogeneity. The pooled fractions were dialyzed for storage at -80 °C using a buffer containing 20 mM Tris (pH 7.5), 100 mM NaCl, 1 mM DTT, 0.01% NP-40, 0.1 mM EDTA and 50% glycerol.

Ivermectin-Based Counter-Selection

Counter-selection using ivermectin was employed to obtain large numbers of *pir-1* arrested homozygote animals to supply enough RNA and protein for the experiments performed. Ivermectin is a class of macrocyclic lactones called avermectins. The *pir-1* deletion alleles were crossed into an ivermectin-resistant triple mutant genetic background comprising the genes *avr-14, avr-15* and *glc*-1, which encode glutamate-gated chloride channel subunits (Dent et al. 2000). This mutant combination is referred to as *'avr3x*'. Exist of the wild-type sequence of any one of the three genes is sufficient to recover sensitivity to ivermectin. The genetic balancer *mnC1** used to propagate *pir-1* mutants contains a wild-type copy of *avr-15*. This makes heterozygote carriers and balancer homozygote worms to get sensitive to ivermectin. Conversely, balancer-free pir-1 homozygotes are able to grow in the presence of the drug. Balanced pir-1 animals were expanded in regular NGM in order to obtain sufficient synchronous L1 larvae to plate on NGM supplemented with 15 µg/L of ivermectin. After plating, only *pir-1* homozygotes grew while all others never passed the L1 stage. Worms were precipitate in M9 buffer for up to 5 times. While this led to loss of some pir-1 arrested worms, it eliminated most live L1s.

RNA Extraction and qRT-PCR

Worms were collected from plates and washed 3x with 10 ml of M9 buffer in 15 ml conical tubes with 1min centrifugations at 4000x g. They were then incubated with 10 ml of M9 buffer for 15-30 minutes in a rocking platform to allow removal of virus around worms. At least 5 volumes of TRI Reagent (MRC, Molecular Research Center) were added to pellets, which were either flash frozen in a dry ice/ethanol bath or processed immediately. Lysis was performed by crushing with a metal dounce for 30 strokes. Samples were aliquoted into 1.5 ml tubes and processed according to the manufacturer. BCP separation reagent was used. RNA pellets were phenol chloroform purified and dissolved in Tris-HCl pH 7.5.

Before cDNA synthesis, 100 µg of RNA were pre-treated with 6 U of Turbo DNase (Ambion) in a 100 µl volume at 37℃ for 15 mins. For each 20 µl cDNA reaction, 2 µg of RNA were used, using Superscript III Reverse Transcriptase (Invitrogen) with random hexamers according to the manufacturer's instructions. qRT-PCR reactions were carried out using the ABI Prism 7500 Fast Sequence Detection System with Fast SYBR Green PCR Master Mix (Applied Biosystems). Each 15 µl reaction contained 7.5 µl of SYBR Green reagent, 400 nM of each primer and 2 µl of cDNA. The standard fast thermocycling program was used and for each primer pair a standard curve with 1:4 cDNA dilutions was generated to calculate mRNA amounts in samples. Per sample, 2-3 technical replicates were run.

Western Blotting

For most experiments, proteins were resolved by SDS-PAGE. Optimal resolution of PIR-1 isoforms was achieved using 10% gels. Proteins were transferred to Amersham Hybond P 0.45 PVDF membranes (GE Healthcare) using a Bio-Rad Trans-Blot SD semi-dry apparatus with protein transfer buffer (20% methanol, 0.1% SDS, 48 mM Tris, 39 mM glycine). Membranes were blocked with TBST with 3% (w/v) of BSA for 30 minutes to 1 hour at room temperature. Antibodies were diluted in TBST/BSA. Primary antibodies were incubated at 4˚C overnight room temperature for 1 hour and secondary antibodies at room temperature for 1-3 hours. Membranes were washed 5x 10 minutes in TBST after each incubation. Chemiluminescence from the HRP-conjugated secondary antibodies was developed with Chemiluminescence and colorimetric detection kits (Bio-rad). Anti-Flag primary antibody (Sigma-Aldrich) were diluted 1:5000, anti-GFP (Invitrogen) were diluted 1:2000, mouse HRP-conjugated secondary antibodies (Santa Cruz Biotechnologies) were incubated at 1:10000 dilutions.

Immunoprecipitation

Protein extraction was generally performed in 2 worm-pellet volumes of lysis/IP buffer containing 25 mM HEPES-KOH pH 7.5, 2 mM EDTA, 0.5% NP-40, and 150 mM KCl, supplemented with protease and phosphatase inhibitors (Roche protease and phosphatase inhibitor cocktail tablets). All steps were performed on ice or 4˚C. Animals were crushed with 75-100 strokes in a metal dounce. Lysates were cleared by centrifugation at 20,000x g for 15 minutes and filtered by 0.22 um PES low binding filters. Protein concentration was measured using Bradford Assay. We generally obtained concentrations of 20 mg/mL. Prewashed antibody beads slurry is added to the cleared lysate and incubated for 1.5 hour. Beads were washed at least 5x 10 minutes with lysis buffer. After complete removal of the buffer, the immunoprecipitates were eluted by incubating the beads with 50 μl of 2X protein sample buffer (1X is 50 mM Tris-HCl pH 6.8, 10% glycerol, 2% SDS, 0.01% (w/v) bromophenol blue, 100 mM DTT) for 10 minutes at 75℃. Inputs were typically diluted to 5 μg/μl in 2X protein sample buffer and 50-200 μg were loaded on gels. Antibodies for IPs were used as follows: 40 μl of anti-Flag M2 Magnetic Beads (Sigma-Aldrich) for Flag IPs.

RESULTS

C. elegans PIR-1 (cPIR-1) possesses an RNA polyphosphatase activity *in vitro*.

hPIR1 has an RNA polyphosphatase activity, which removes the γ and β phosphates of ppp-RNAs. Both hPIR1 and *c*PIR-1 are also highly homologous to the triphosphatase domain of RNA capping enzymes, which removes the γ phosphate of ppp-RNAs. Baculovirus phosphatase (BVP), a PIR-1 homolog, is an RNA triphosphatase, which rescued the lethality caused by a yeast RNA capping enzyme mutant lack of a triphosphatase activity. To dissect the activity of *c*PIR-1 *in vitro*, we purified N-terminal His-tagged recombinant proteins using the WT and catalytically-dead (C150S) *pir-1*, and examined the activity on *in vitro* transcribed ppp-RNAs (Figure 1.1). The C150S is designed based on the sequence homology of PIR-1 family members containing the catalytic motif H$\underline{\text{C}}$X$_5$RXG (Figure 1.1A), and the hPIR1 structure (Jeong et al. 2014; Sankhala et al. 2014). The C150 is critical for catalysis, basically forming a thiol-phosphate ester. Since the substrate ppp-RNAs and the product p- or pp-RNAs co-migrate on PAGE gels, we added Terminator exonuclease (Terminator), which destroys only p-RNAs. If cPIR-1 is a polyphosphatase, the product p-RNAs will be degraded by Terminator, and if PIR-1 is a triphosphatase, the product pp-RNAs will be resistant to Terminator. We found that WT PIR-1 is an RNA polyphosphatase converting ppp-RNAs to p-RNAs, which were then degraded by Terminator (Figure 1.1B). In contrast, C150S is not able to generate p-RNAs, indicating that C150 is required for catalysis.

C150S binds ppp-RNAs tightly. WT PIR-1 modifies ppp-RNAs (Figure 1.1B) but has little affinity for the product p-RNA (Figure 1.1C). In contrast, the C150S cannot dephosphorylate ppp-RNAs (Figure 1.1B), but binds it tightly (Figure 1.1C). The interaction is dependent on the 5'ppp since the same RNA with 5'p cannot bind C150S (Figure 1.1C). This suggests that the C150S/RNA interaction is caused by the expected activity of a catalytically-dead protein, i.e., binding without catalysis, but not by a gain of function of the C150S, since it is difficult to gain a 5'ppp-binding activity. This interaction was not caused by the contaminating proteins or denatured proteins, since the C150S/RNA complex migrated as a clear band in the native PAGE gel and was detected by RNA staining and a western blot for His tags from His-tagged C150S (Figure 1.1D). In summary, these experiments indicate that PIR-1 is an authentic RNA polyphosphatase.

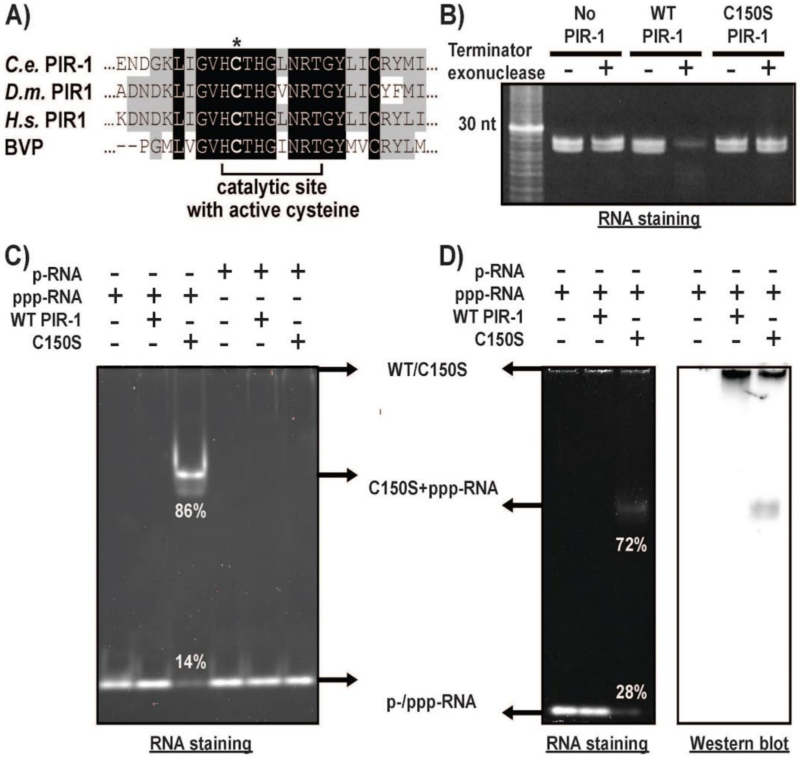
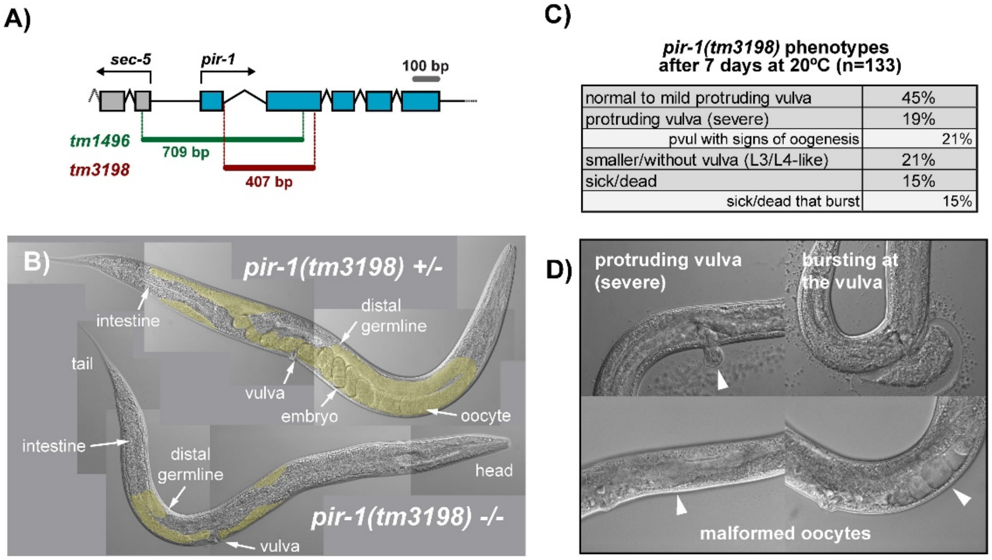


Figure 1. 1. PIR-1 is an RNA polyphosphatase.

A) The active site alignment of PIR-1 orthologs including Baculovirus phosphatase (BVP) and the proteins in *C. elegans* (*C.e.*), *Drosophila* (*D.m.*), and human (*H.s.*). Asterisk indicates the catalytic cysteine. The gray part indicates an amino acid identity among two to three sequences and the black part indicates an identity for all the sequences. **B)** WT PIR-1 is able to convert ppp-RNAs to p-RNAs *in vitro*. The 26-27 nt ppp-RNAs were incubated with the WT PIR-1 or C150S PIR-1, followed by a Terminator treatment, which specifically destroys p-RNAs. **C)** C150S PIR-1 may bind ppp-RNAs tightly. The WT or C150 PIR-1 was incubated with p- or ppp-RNAs and resolved by a 15% native PAGE gel with SYBR Gold RNA staining. **D)** SYBR Gold RNA staining and Western blot verify that C150S rather than some contaminating proteins co-migrates with ppp-RNAs. The ppp-RNAs were incubated with the WT or C150 PIR-1 and resolved on a 15% native PAGE gel. Then one half of the gel was used for SYBR Gold RNA staining and the other half for Western blot to detect the His-tagged C150S and WT PIR-1 using a His-tag Ab. The arrows indicate the migration positions of the substrate or product in the reaction.

PIR-1 Is essential for growth and development

We obtained a new *pir-1* deletion allele to confirm the growth defects. As described previously, the *pir-1* deletion (*tm1496*) worms mostly arrest at L4 larval stage with underdeveloped germlines. This allele, however, also deletes the whole promoter of a three-gene operon and the 5' end of the first gene, *sec-5*, including 35-nt in the coding region (Figure 1.2A). To verify the phenotype, we obtained a new allele, *tm3198*, which only deletes 407-nt including the first intron and second exon of *pir-1* while retaining the rest of the gene coding an out-of-frame product (Figure 1.2A). Like *tm1496*, the *tm3198* homozygotes are also sterile. However, most worms grow further than *tm1496* worms and arrest at young adult stage with underdeveloped germlines (Figure 1.2B). Since RNAi of *sec-5* caused embryonic lethality (data not shown), we speculate that the earlier developmental arrest of *tm1496* worms is likely due to the combined loss-of-function of *pir-1*, *sec-5*, and other genes of the operon.



***pir-1(tm3198)* phenotypes after 7 days at 20ºC (n=133)**

normal to mild protruding vulva	45%
protruding vulva (severe)	19%
pvul with signs of oogenesis	21%
smaller/without vulva (L3/L4-like)	21%
sick/dead	15%
sick/dead that burst	15%

Figure 1. 2. PIR-1 is essential for somatic and germline development.

A) Diagram of the WT pir-1 genomic locus and the deletion allele *tm1496* and *tm3198*. B) Image of a live heterozygous pir-1 mutant balanced with mnC1 (top) and a terminally arrested homozygous *pir-1* mutant (bottom) grown at 20˚C for 96 hours. Germlines are highlighted in yellow and are partly concealed by intestinal tissue. C) Quantification of visible phenotypes exhibited by 133 *tm3198* homozygotes grown for seven days. D) Images of live *pir-1* mutant animals exhibiting major phenotypes scored in C.

The limited development of the *tm3198* homozygotes is likely due to a maternal load of *pir-1* mRNA or protein. Since *pir-1* is essential, the mutant worms are maintained as heterozygotes using the *mnC1* balancer. The *tm3198* homozygotes arrest at stages between L3-like larvae and young adults (Figure 1.2C) with the majority (~64%) displaying adult features, including the presence of a vulva, 16 pairs of seam cells, and continuous alae (Figure 1.2B and data not shown). The vulva is frequently protruding and occasionally leads to the burst of the worms, indicating that PIR-1 is required for somatic development (Figure 1.2D). Although a tiny fraction of worms exhibit vestiges of oogenesis, all animals fail to produce any progeny (Figure 1.2D). The development-arrested worms are active and feed normally with a WT life-span, ~16-18 days at 20˚C. Attempts to silence *pir-1* via RNAi failed to produce any visible phenotype, precluding us from studying PIR-1 throughout development. The resistance may reflect that PIR-1, a Dicer interactor, is essential for RNAi. The germline is underdeveloped in the *tm3198*. In all, PIR-1 promotes somatic growth, germline proliferation, and sperm differentiation.

The RNA polyphosphatase activity of PIR-1 plays important roles *in vivo*.

To examine the *in vivo* effects of the catalytically-dead C150S, we used CRISPR to convert the WT *pir-1* to the *C150S* mutant. From the two lines obtained, we observed the same germline defects as in the *tm3198*. However, the growth rate is similar to that of the WT, reaching adulthood within ~ 3 days at 20 ºC, while the *tm3198* usually requires 7-days. Since we are able to fully rescue the *tm3198* with a WT *pir-1* transgene, the slow growth is likely caused by lack of PIR-1 protein. Although the C150S cannot rescue the germline defect, it can rescue this slow growth, suggesting that the phosphatase activity is not the only function that PIR-1 plays in the cells. In addition, we failed to rescue the *tm3198* using a *pir-1* (*C150S*)*::gfp* transgene, further demonstrating that the catalytic activity of PIR-1 play important roles.

PIR-1 is required for silencing the Orsay virus via RNAi

Since the PIR-1 interactor DCR-1 is essential for silencing the Orsay virus via RNAi in *C. elegans* (Ashe et al., 2013; Franz et al., 2014), we examined whether PIR-1 is also involved. Although restricted by RNAi, the virus is able to replicate in a few intestinal cells in WT worms; however, the viral RNAs can increase 100 folds in RNAi-deficient worms (Ashe et al., 2013). An RT-qPCR analysis indicated that the levels of Orsay RNA1 were ~100 folds higher in the *tm3198* than in the WT at 20 ºC (one-tailed TTest, P <0.001, Figure 1.3A). The increase was comparable to that in the known RNAi mutants such as *rde-1* (*ne300*), the Argonaute binding 23-mer primary siRNAs. PIR-1 and RDE-1 likely work in the same pathway since a *pir-1*; *rde-1* mutant exhibited a similar increase as each single mutant (Figure 1.3A). The *pir-1*(WT)*::gfp* transgene restored the viral suppression in the mutant, suggesting that lack of PIR-1 caused the RNAi deficiency, disilencing the virus. The viral load is much higher in the catalytically-dead C150S than in the WT, indicating that the phosphatase activity is required for the antiviral role (Figure 1.3A).

We next examined how PIR-1 functions in RNAi. Since PIR-1 interacts with Dicer, we speculated that the mutants may have compromised Dicer activities, making less 23-mer primary siRNAs. We directly cloned 23-mer RNAs using 5' ligation without the TAP treatment, since 23-mer RNAs bear 5'p. Both strands of dsRNA1 exhibited a 23-mer peak, indicating our method did clone primary siRNAs (Appendix A-Figure S1.1A). The antisense (anti) RNA1 exhibited a much sharper peak (38% are 23-mer) than the sense (15% are 23-mer); as for the read number, there were more sense siRNAs than the anti siRNAs across the RNA1, a result different from the theoretical 1:1 ratio (Appendix A-Figure S1.1A). These discrepancies were likely caused by the contamination of degraded RdRP mRNA (sense RNA1). The degradation of the anti RNA1 is minimal since it is expressed at a much lower level than the sense RNA1. Consistent with this model, the sense minus antisense represents the degradation products of sense RNA1, exhibiting a much evener size distribution (Appendix A-Figure S1.1A, green). A similar analysis on the Orsay RNA2 reached the same conclusion. Based on this analysis, we decided to use 23-mer on the anti RNA1 to represent primary siRNAs. We used the miRNA reads as the normalization standard, since the *pir-1* mutants do not display obvious miRNA defects (Figure 1.3C). We found that the N2 and *avr3x* exhibited very few primary siRNAs across the RNA1 (Figure 1.3B, yellow and green). While the *avr3x* peaks at 23-nt, the N2 peaks at 22-nt. It is normal to see a tiny 22-nt peak in N2 worms since: 1) the virus is silenced, thus generating very few dsRNA for making 23-mer RNAs; and 2) although most of 22G-RNAs, which are expressed at much higher levels than 23-mer, cannot be efficiently cloned by 5' ligation due to the 5'ppp, a small fraction may be dephosphorylated to 5'p-RNAs by various factors, becoming clonable and contaminating the primary siRNA profile. Regardless, the number of primary siRNAs in the N2 and *avr3x* are very small and comparable, consistent with the observation that the Orsay virus is silenced (Figure 1.3A). In contrast, the RNAi mutants, *pir-1, rde-1*, and *drh-3*, all exhibited much more primary siRNAs (for the *pir-1*, ~6 folds over the N2 and *avr3x*; for the *rde-1*, ~20 folds; for the *drh-3*, ~8 folds) and apparent peaks at 23-nt (Figure 1.3B). As expected, these 23-mer bear all the features of Dicer products: 1) 5' p and 3' OH; 2) 2-nt 3' overhangs when paired sense and antisense siRNAs are aligned; 3) roughly equal sense and antisense siRNAs (Appendix A-Figure S1.1C); 4) 23-mer are the major peak, ~ 40% of all primary siRNAs (Figure 1.3B); and 5) these 23-mer RNAs do not exhibit an obvious 5' nucleotide preference (Figure 1.3C).

The efficiency of 23-mer RNAs biogenesis is likely compromised in the *pir-1* mutant. The increase of 23-mer RNAs (~6 folds) in the *tm3198* did not match that of viral RNAs (~100 folds) (Figure 1.3A vs. B). It is likely that lack of PIR-1 affects the processing of dsRNAs by Dicer, as indicated by a lower 23-mer/viral RNA ratio (~16 folds less). To explain the accumulation of 23-mer RNAs, we propose that the compromised Dicer activity caused by lack of PIR-1 may be compensated by more dsRNAs, generating more 23-mer RNAs.

We then examined whether these increased 23-mer RNAs are able to trigger more 22G-RNAs. As secondary siRNAs made by RdRPs, 22G-RNAs bear 5'ppp, which is incompatible for 5' ligation. We used a commercial RNA polyphosphatase (Tobacco Acid Pyrophosphatase, TAP) to convert ppp-22G-RNAs to p-22G-RNAs. However, primary siRNAs are also cloned since they are 5' p-RNAs. A fraction of them appear as 22G-RNAs (~5% in the *pir-1* and *rde-1* mutants) since 23-mer only represent the peak. Therefore, 22G-RNAs cloned using the TAP treatment are composed of primary and secondary 22G-RNAs. To obtain the number of secondary 22G-RNAs rather than a mixture, first we considered only 22G-RNAs with a TSS motif YG (G, the 5' G of 22G-RNAs;

Y, an upstream pyrimidine encoded but not transcribed), since secondary 22G-RNAs *de novo* transcribed by RdRPs contain TSS (Gu et al. 2012), while primary 22G-RNAs generated by Dicer do not. Then we subtracted p-22G-RNAs cloned without the TAP treatment from all 22G-RNAs cloned with the treatment to obtain the ‘adjusted YG-22G-RNAs', i.e., ppp-22G-RNAs. This way we are able to minimize the background noise, especially p-22G-RNAs in the RNAi-deficient strains (Figure 1.3B). In addition, the degradation of viral RNAs also contributes to the total number of 22G-RNAs. The usage of YG motif can remove half of the degradation-derived 22G-RNAs (Y is ~ 50%). To further minimize this background, we used 22G-RNAs mapped to the anti RNA1 to represent the total 22G-RNAs, since: 1) They are actually generated using the sense RNA1 template by RdRPs and much more abundant than those on the sense RNA1 (20 folds more in the WT); 2) the degradation of the anti RNA1, which is expressed at much lower levels than the sense RNA1, generates less sRNAs. The sRNA profile on anti RNA1 exhibits a much sharper and higher 22mer peak than that on sense RNA1.

Surprisingly, we found the increased 23-mer RNAs did not trigger more 22G-RNAs in the *pir-1* mutant. The N2 and *avr3x* sRNAs cloned using the TAP treatment exhibited sharper and higher 22mer peaks, while the *tm3198* also exhibited a 22mer peak of dramatically reduced abundance and the *rde-1* and *drh3* each exhibited a 23-mer peak (Figure 1.3D). The sRNA size profile indicates that the cloning method worked as intended to clone both primary and secondary siRNAs. As discussed above, we used the adjusted YG-22G-RNAs to represent secondary 22G-RNAs. We found that the *pir-1* mutants exhibited much less YG-22G-RNAs than the N2 worms (4.5 folds less, TTest, one-tailed P <0.01) and the *avr3x*, and that the *rde-1* and *drh-3* worms were almost depleted of YG-22G-RNAs (~50 folds less than N2, TTest, one-tailed P<0.0004 for *rde-1*) (Figure 1.3E). We used the 22G/23-mer RNAs ratio as the triggering potential to measure how efficiently 23-mer RNAs are able to promote the biogenesis of 22G-RNAs. One 23-mer RNAs in N2 and avr3x is able to trigger 40 and 11 22G-RNAs respectively, while one 23-mer RNAs in the *pir-1* mutant is only able to generate 1 22G and 23-mer RNAs in the *rde-1* and *drh-3* mutants barely trigger any 22G-RNAs (Figure 1.3F). DRH-3 is required for promoting RdRPs to make 22G-RNAs, but likely not for generating 23-mer RNAs. In the *drh-3* mutant, more 23-mer RNAs are generated since the desilenced virus provides more dsRNAs. However, these 23-mer RNAs are not able to trigger 22G-RNAs because of the compromised RdRP activity due to lack of DRH-3. In the *rde-1* mutant, 22G-RNAs decreased because there was no RDE-1/23-mer RNAs complex, which is required for triggering 22G-RNAs. Since it is assumed that RDE-1 is required for the stability of 23-mer RNAs, we expected to observe less 23-mer RNAs in the *rde-1* mutant. However, we observed more 23-mer RNAs, which bear all the features of primary siRNAs as those in the *drh-3* and *pir-1* mutants (Figure 1.3B-C). We speculate that these 23-mer RNAs were generated from the upregulated dsRNAs due to RNAi deficiency. At least two models can explain the stability: 1) 23-mer RNAs bind other Argonautes which cannot trigger 22G-RNAs; 2) 23-mer RNAs exist as free duplex siRNAs without Argonautes.

However, it is hard to understand why more 23-mer RNAs and less 22G-RNAs coexist in the *pir-1* mutant, since RDE-1 exists. This phenotype is not related to the germline defects, since the Orsay virus replicates in somatic cells. The negative correlation of 23-mer RNAs vs. 22G-RNAs is less dramatic in the *pir-1* mutant than other mutants (Figure 1.3F). We reasoned that the *pir-1* nulls derived from the balanced heterozygous worms contain a maternal load of WT PIR-1, while the *rde-1* and *drh-3* mutants do not.

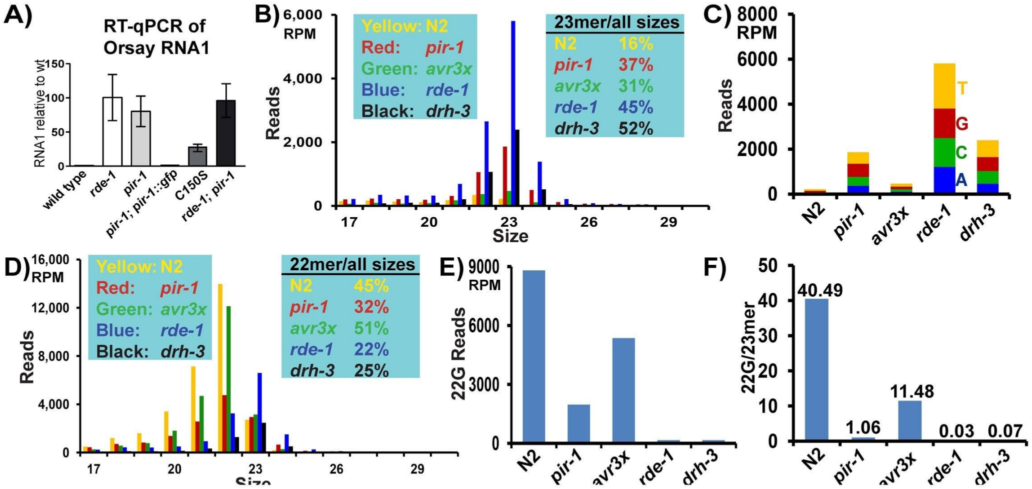


Figure 1. 3. PIR-1 is required for suppressing the Orsay virus.

A) RT-qPCR analysis of the Orsay virus RNA1 levels, as first normalized to the gpd-2 mRNA level and then compared to the RNA1 level in the WT. B) The size distribution of sRNAs cloned without the TAP treatment (only cloning p-RNAs) and mapped to antisense RNA1 in the mutants and controls. The right table lists the rate of 23mers in each sample. The normalization standard for B-F is the total miRNAs, C) The first nt distribution of 23mers in B. D) The size distribution of sRNAs cloned with the TAP treatment (cloning p- & ppp-RNAs) and mapped to anti RNA1 in the mutants and controls. The right table lists the rate of 22mers in each sample. E) Comparison of YG-22Gs, which are derived from the 22mers in D. F) Comparison the ratios of YG-22Gs over 23mers, as shown in C and E.

PIR-1 promotes the loading of viral primary siRNAs to RDE-1

The low 22G-RNAs/23-mer ratio contradicts what we have learned from the non-viral RNAi, which showed that 23-mer RNAs are required triggering 22G-RNAs. We speculated that 23-mer RNAs in the *pir-1* mutant were not loaded to RDE-1. We constructed strains containing a non-integrated *gfp::rde-1* transgene of high transmission under the control of the native regulatory sequence of the *rde-1* gene in an *rde-1* mutant ('*wt*') and in a *pir-1; rde-1* mutant ('*pir*-1'). We used the tagged-WT-*rde-1* to pull down the RDE-1/23-mer complex. '*pir-1*' were counter-selected for seven days prior to infection to ensure the depletion of maternal PIR-1 and infected for three days, when 'wt' control grow on ivermectin until the population consisted of a range of L4 larvae to young adults. We then IP-pulled down RDE-1 and cloned the bound RNAs using the TAP treatment.

The Orsay reads were significantly enriched in the IPs. The virus contributed 9% and 13% of the total reads in the '*wt*' and '*pir-1'* inputs respectively, while contributing 41% and ~28% in the corresponding IPs. This result indicates that the IP worked. As discussed above, we analyzed the sRNA profile on the anti RNA1 primarily due to less degradation (Figure 1.4A). The '*wt*' input exhibited a 22mer peak (49% of the total reads), while the IP exhibited a 23-mer peak (42%) (Figure 1.4A). To better measure how much 23-mer RNAs are enriched in the IP, we normalized the viral reads based on the total number of miRNAs, most of which do not bind RDE-1. The number of 23-mer RNAs increased ~ 39 folds in the IP (Figure 1.4A), as compared to the input.

23-mer RNAs were loaded less efficiently to RDE-1 in the '*pir-1*' mutant. We did observe a couple of differences in the IPs: 1) the miRNA-normalized 23-mer RNAs were ~10-fold less in the '*pir-1*' IP than the '*wt*' IP (Figure 1.4A); 2) compared to the input, the RDE-1 IP in the '*pir-1*' enriched ~ 22-fold 23-mer RNAs, lower than the ~39 folds in the '*wt*'. These results suggest that 23-mer RNAs were loaded less efficiently to RDE-1 in the '*pir-1*'. However, since miRNAs usually do not bind RDE-1, miRNAs in the IP may only reflect the residual miRNAs binding the beads non-specifically during IPs, which varies due to unpredictable factors. For example, a 2.5% ('*wt*') vs. 5% ('*pir-1*') residual miRNAs would lead to 40-fold and 20-fold 23-mer RNAs enrichments in the IP's, close to the number we obtained. However, this enrichment difference does not reflect whether RDE-1 is loaded efficiently or not. To better examine this question, we normalized the samples using mir-243, which, unlike other miRNAs, specifically binds RDE-1 rather than ALG-1/ALG-2. Since our data indicates that PIR-1 does not affect the biogenesis of miRNAs including mir-243, we may assume mir-243 is loaded to RDE-1 normally in the mutant. Using this normalization, we found that RDE-1 in the '*wt*' IP bound 6-fold 23-mer RNAs as many as in the '*pir-1*' IP (261818 vs. 42957 reads).

PIR-1 affects the preference of 23-mer RNAs species loaded to RDE-1. The RNAi mutants, *pir-1*, *rde-1* and *drh-3*, all exhibited increased 23-mer RNAs without an obvious first preference in the total sRNAs (Figure 1.4C). However, more 23As (average 42%) and less 23Ts (average 17%) were loaded to RDE-1 in the '*wt*'. In contrast, less 23As (average 28%) and more 23Ts (average 44%) were loaded to RDE-1 in the '*pir-1*' (Figure 1.4B). This result further demonstrates that the loading of 23-mer RNAs is compromised in the *pir-1* mutant.

To explain the stability of non-RDE-1-bound 23-mer RNAs, we speculated that 23-mer RNAs may be loaded to other Argonautes. ALG-1 is the top candidate since: 1) it is the closest paralog to RDE-1; 2) ALG-1 binds miRNAs and RDE-1 binds a special miRNA, miR-243; 3) 23-mer RNAs and miRNAs are both made by Dicer and peak at 23-nt with 5' p and 3' OH. Other Argonautes are excluded since: 1) ALG-3/ALG-4/ALG-5 are not expressed in somatic cells; 2) PRG-1 is only expressed in germline cells and bind 21U-RNAs, and PRG-2 is closely related to PRG-1; 3) ERGO-1 only binds 26G-RNAs; 4) CSR-1 and WAGO1-12 bind 22G-RNAs; and 5) C04F12.1 likely binds 22G-RNAs since it is closely related to CSR-1. Instead of testing the competition of RDE-1 and ALG-1 for binding 23-mer RNAs in the *pir-1* mutant, we answered a more general question, i.e., whether ALG-1 is able to bind viral 23-mer RNAs in the presence or absence of RDE-1 in a WT *pir-1* background. Like the *pir-1* mutant, the *rde-1* mutant accumulated even more 23-mer RNAs (Figure 1.4C). RDE-1 functions downstream of Dicer, likely not affecting the processing of dsRNAs. If 23-mer RNAs can be loaded to ALG-1, the *rde-1* mutant would be the best to study this loading since both the accumulated 23-mer RNAs and lack of

RDE-1 favor the loading of 23-mer RNAs to ALG-1. We pulled down ALG-1 from the Orsay-infected WT and *rde-1* mutant using IP, and analyzed the bound sRNAs treated with TAP.

A fraction of the 23-mer RNAs in the '*pir-1*' mutant was loaded to ALG-1. ALG-1 IP worked perfectly since ~99% sRNAs in the WT and 97% in the *rde-1* were miRNAs. In contrast, the inputs contained only ~18% (WT) and 25% (*rde-1*), while the rest was other endogenous and viral sRNAs. We then analyzed the sRNAs from the virus. In the WT input, the peak was 22G-RNAs; in the IP, 23-mer RNAs either became the peak (sense) or relatively increased (anti) (Figure 1.4C). To estimate if primary siRNAs are enriched in the IP, we need to obtain the number of authentic primary 23-mer RNAs in both the input and the IP. Although the 23-mer RNAs in the IP sample may represent secondary siRNAs (maybe over-represented since the peaks are not sharp, indicating contaminations), the 23-mer RNAs in the input may not, since our method cloned both primary and secondary 23-mer RNAs, and the abundant secondary 23-mer RNAs may overwhelm the primary 23-mer RNAs (Figure 1.4C**,** the WT: input). Due to this technical issue, we did not assess whether ALG-1 enriches 23-mer RNAs in the WT. In the *rde-1* mutant, which lacks secondary siRNAs (Figure 1.3F), we observed clear 23-mer RNAs peaks in the input (Figure 1.4C), ~10,000 reads on the anti RNA1, which is equal to 1% of the miRNA reads (a perfect normalization for the ALG-1 load since ALG-1 binds miRNAs); in the IP, we obtained ~1,200 23-mer RNAs on the anti RNA1, corresponding to 0.12% of the total miRNA reads; we also observed that the IP sample exhibited much sharper 23-mer peaks, indicating that the IP preferentially pulled down 23-mer RNAs over other sizes. Based on these, we concluded that: 1) ALG-1 does bind viral 23-mer RNAs but prefers miRNAs over 23-mer RNAs. We obtained the same conclusion using sense RNA1 (Figure 1.4C) or using RNA2 (data not shown).

Interestingly, the ALG-1-23-mer in the *rde-1* mutant exhibits the same first nucleotide preference as the RDE-1-23-mer in the *pir-1* mutant (Figure 1.4B/D). We aligned the first nt A and T rates of 23-mer RNAs mapped to the four viral strands (sense/anti RNA1/2) in the ALG-1 IP using the *rde-1* worms and RDE-1 IP using the *pir-1* worms. The rates are highly correlated (two-tailed P < 0.0004, spearman r = 0.98, close to 1, the perfect positive correlation). We speculate that the loading of primary siRNAs is compromised in both *rde-1* and *pir-1* with an accumulation of duplex siRNAs, and the siRNAs bound with ALG-1 and RDE-1 in these mutants may be loaded using an alternative mechanism.

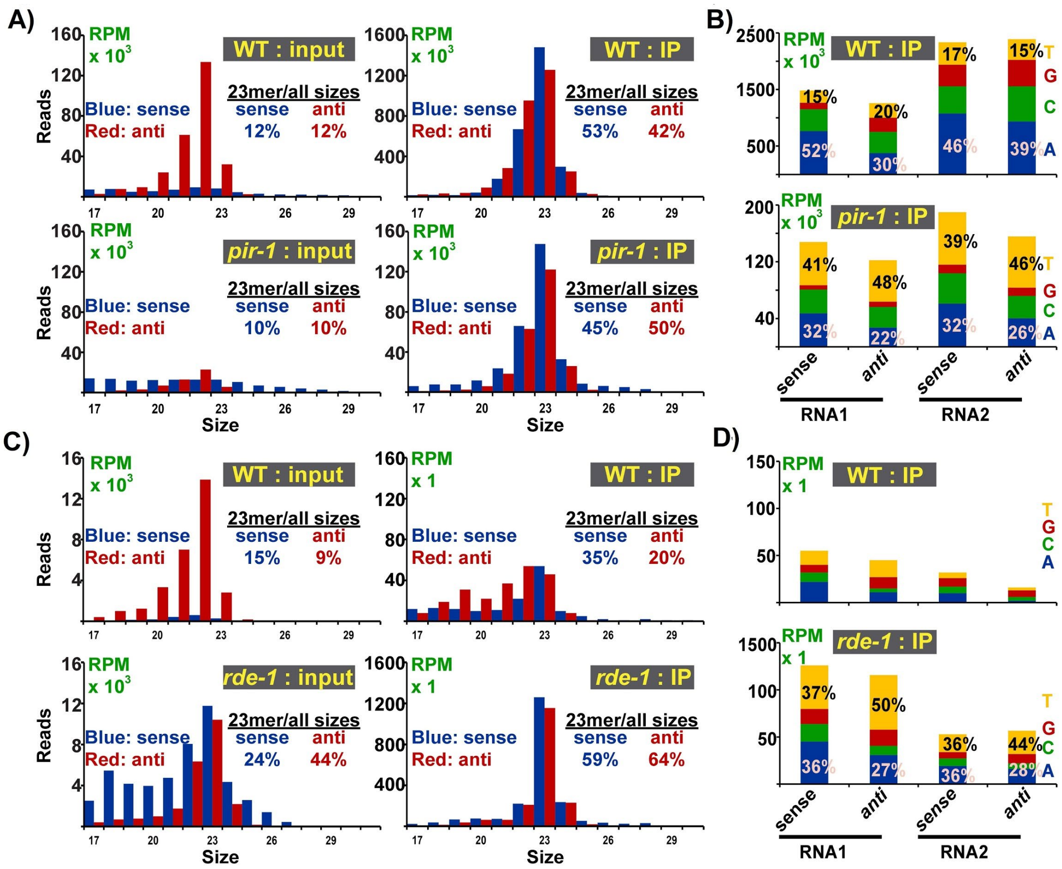


Figure 1. 4. PIR-1 promotes the loading of viral primary siRNAs to RDE-1

A) The size distribution of sRNAs cloned with the TAP treatment and mapped to both sense and anti RNA1 in the RDE-1 IPs and the inputs using the WT and pir-1 mutant worms. The right table lists the rate of 23mers mapped to sense and anti RNA1. B) The first nt distribution of 23mers mapped to RNA1 (A) in the RDE-1 IPs. C) The size distribution of sRNAs cloned with the TAP treatment and mapped to both sense and anti RNA1 in the ALG-1 IPs and the inputs using the WT and rde-1 mutant worms. The right table lists the rate of 23mers mapped to sense and anti RNA1. D) The first nt distribution of 23mers mapped to RNA1 (C) in the ALG-1 IPs. See also Appendix A-Figure S1.1.

PIR-1 interacts with Dicer and other proteins of the ERI complex throughout development

To understand how PIR-1 regulates sRNAs, we employed multidimensional protein identification technology to identify proteins that co-immunoprecipitate (co-IP) with PIR-1. We performed immunoprecipitations (IPs) using gravid adult worms grown with 3xFlag::PIR-1. We incorporated the protein extracts from the WT animals in the analysis. The ERI components are among the proteins of top spectral counts, namely DCR-1, DRH-3, RRF-3, RDE-4, and ERI-3. They rank among the highest abundance proteins recovered, as judged by the obtained spectral counts, the measurement that most strongly correlates with relative protein abundance. Meanwhile, we observed several DNA replication related proteins such as RNH-1.0, DNA replication licensing factor MCM-2, MCM4-7, and RNA pol II subunit proteins are co-precipitate with PIR-1. And most small nuclear ribonucleoprotein (SNR) are detected too. These SNR proteins express in gonad specifically where PIR-1 is enriched too. They coordinate not only aspects of pre-mRNA processing, mRNA metabolism and transport, but also DNA replication, DNA damage repair and telomere maintenance. Since there is no available antibody for RNH-1.0 and SNR proteins, we are creating HA-tagged strains to verify the interactions and functions.

Table 1. 1. List of PIR-1 interactors obtained in an independent MudPIT experiment

Protein	Length in aa/MW in kDa	Spectral Count (abundance)	Peptide Count	Protein Coverage
PIR-1	233 / 27	2062	21	73%
DCR-1	1910 / 218.3	1031	87	52%
DRH-3	1119/129	32	12	17%
ERI-3	578 / 66.3	108	14	23%
RDE-4	385/43.4	233	29	71%
RRF-3	1765/201.3	85	21	16%
SNR-3	126/13.6	49	5	52%
RNH-1.0	155/18.4	126	9	57%

DISCUSSION

PIR-1 belongs to a family of novel RNA phosphatases with diverse functions. The homolog BVP and the triphosphatase domain of RNA capping enzymes only remove the *γ* phosphate of ppp-RNAs, which are then converted to capped RNAs, thus stabilizing them. However, hPIR1 and *C. elegans* PIR-1 are able to remove both *β* and *γ* phosphates of ppp-RNAs, potentially making them susceptible to 5'-to-3' exonucleases. Therefore, ppp-RNAs including nascent nuclear RNAs and viral RNAs can be regulated by different phosphatases, resulting totally opposite fates. A recent report using mammalian cell lines has identified a couple of lincRNAs as the hPIR-1 targets. However, this discovery cannot account for the hPIR1-related phenotypes observed previously including: 1) hPIR1 interacts with the components of spliceosome; 2) the over-expression of hPIR1 inhibits cell growth or the downregulation of hPIR1 promotes cell growth. Lack of convenient and specific phenotype makes it difficult to understand how hPIR1 regulates important biological processes. Here we demonstrate that PIR-1 is an essential component of the RNAi-mediated pathways including ERI and antiviral responses. We also dissect the molecular mechanism of the antiviral pathway and demonstrate that PIR-1 is required for the efficient production of primary siRNAs and the loading of these siRNAs to Argonautes.

PIR-1 modifies viral ppp-RNAs and is required for generating primary siRNAs and for loading them to the Argonautes. The antiviral pathways generate ppp-dsRNAs using RdRPs. And a PIR-1/Dicer complex is required for processing these dsRNAs into primary siRNA 23-mer RNAs and 26G-RNAs respectively. Also primary siRNAs are used to trigger the biogenesis of 22G-RNAs. In the antiviral pathway, the ratio of 23-mer/viral RNA is ~ 10-fold lower in the *pir-1* mutant than in the WT, suggesting that lack of PIR-1 may compromise the ppp-dsRNA-processing ability of Dicer. Based on our *in vitro* data, PIR-1 may dephosphorylate ppp-dsRNAs to p-dsRNAs *in vivo*, thus making them better substrates for Dicer. Consistent with this model, the catalytically-dead PIR-1 loses the virus-silencing capability.

In addition, PIR-1 is also required for the efficient loading of primary siRNAs to RDE-1 in the antiviral pathway. Since PIR-1 interacts with Dicer, it may be required for the integrity of the Dicer complex which also contains DRH-1 and RDE-4, thus affecting the loading of the Dicer product, i.e., 23-mer RNAs.

In the antiviral pathway, DRH-1, another Dicer-interacting protein, is required for the production of 23-mer RNAs. In the *pir-1* mutant, the biogenesis rate of 23-mer RNAs (23-mer/viral RNA) is lower. This contrast does not necessarily mean that PIR-1 is only partially required for the biogenesis of 23-mer RNAs or PIR-1 works downstream of DRH-1, since the *pir-1* mutants contain maternally-loaded WT PIR-1, while the *drh-1* mutants do not. The same reasoning applies to the *dcr-1* mutant, which also contains maternally-loaded Dicer. We propose PIR-1/Dicer/DRH-1/RDE-4 works at the very beginning step of RNAi, basically cutting dsRNAs, based on: 1) PIR-

1 interacts with DCR-1; 2) PIR-1 is required for the loading of 23-mer RNAs, the immediately following step.

PIR-1 may serve as a sensor for recruiting viral ppp-RNAs to the Dicer complex. Previous studies have implicated RIG-1 as a sensor for RNA viruses, since RIG-1 has a triphosphate-binding domain which recognizes viral ppp-RNAs. However, RIG-1 is only required for restricting specific RNA viruses, but not for all or most RNA viruses. Since all the RNA viruses generate ppp-RNAs during their life cycles, apparently the ppp group is not sufficient for recruiting RIG-1. In the RIG-1-mediated silencing of viruses, the underlying mechanism is not based on RNAi but on interferon-mediated pathways. Interestingly in *C. elegans*, DRH-1, the RIG-1 homolog, is required for silencing the Orsay virus in an RNAi-dependent manner. However, DRH-1 appears required for the translocation of Dicer along dsRNA templates, since 23-mer RNAs are still made by Dicer but limited to the very end of dsRNAs. Based on our *in vitro* evidence that PIR-1 is able to recognize ppp-RNAs and remove two phosphate groups, it is tempting to speculate that PIR-1 may recruit ppp-RNAs to Dicer because PIR-1 also interacts with Dicer. If PIR-1 is able to dock Dicer to ppp-dsRNAs, we would expect that the catalytically-dead C150S mutant is able to silence the virus since it binds ppp-RNAs tightly. Apparently our result is the opposite. However, it is still possible that PIR-1 recruits Dicer to ppp-dsRNAs. First, PIR-1 can recognize, bind and modify the ppp group, a handle on viral RNAs. Second, the loss of the antiviral capacity in the *C150S* mutant may reflect that the mutant tightly binds the ppp-RNAs, thus keeping Dicer from processing dsRNAs and sequestering it from other substrates. This means that the *C150S* could have silenced the virus if Dicer had been released. In summary, PIR-1 may help engage Dicer on ppp-dsRNAs.

PIR-1 plays other functions likely in RNAi-independent manners. The *pir-1* mutant is sterile at all temperatures. However, the abolishment of the ERI pathway only causes temperature-sensitive sterile phenotype due to sperm defects. Other essential components for RNAi, such as MUT-7 and RDE-3, are not required for fertility at room temperature either. These lines of evidence indicate that RNAi is not essential for fertility in *C. elegans* and the essential roles PIR-1 plays are likely RNAi-independent. It is also possible that PIR-1 may regulate nuclear nascent ppp-RNAs including pre-mRNAs, since PIR-1 is primarily localized in the nucleus and may associate with chromatin. Therefore, PIR-1 may function as a master phosphatase to modify ppp-RNAs in several complexes or pathways.

The C150S PIR-1 can be used to purify ppp-RNAs and WT PIR-1 can be used to modify ppp-RNAs. The PIR-1 family uses a conserved Cysteine to form a thiol-phosphate ester with ppp-RNAs for catalysis. As expected, the C150S abolishes the catalytic activity of PIR-1. However, it still binds ppp-RNAs and the binding is triphosphate-specific. This suggests that either the 150S forms an O-linked ester bond with ppp-RNAs without further catalysis, thus causing the binding, or other determinants bind the ppp group of ppp-RNAs. Regardless the mechanism, the C150S mutant can be used to enrich nascent or viral ppp-RNAs. Since the level of nascent RNAs may represent the transcription rate, it is possible to develop a convenient method to analyze the transcription rate using this mutant protein. In addition, the recombinant PIR-1 can be used to modify ppp-RNAs for RNA cloning. Actually we are using this enzyme to clone ppp-22G-RNAs. One of the advantages of using PIR-1 is that PIR-1 works in the ligation buffer at the ligation temperature, thus significantly simplifying the cloning procedure.

In summary, we demonstrate PIR-1, as a novel class RNA polyphosphatase, is required for the RNAi-mediated small RNA and antiviral pathways. In addition, PIR-1 may also play non-RNAi-mediated roles required for larval development. We also dissect the molecular mechanism of PIR-1-mediated antiviral roles. This is the first comprehensive study demonstrating that PIR-1, as a master RNA phosphatase, clearly plays important roles in regulating ppp-RNAs. Based on its functions, PIR-1 will become an important target in viral studies and cancer studies.

REFERENCES

Ashe, Alyson, Tony Bélicard, Jérémie Le Pen, Peter Sarkies, Lise Frézal, Nicolas J. Lehrbach, Marie-Anne Félix, and Eric A. Miska. 2013. "A Deletion Polymorphism in the Caenorhabditis Elegans RIG-I Homolog Disables Viral RNA Dicing and Antiviral Immunity." *eLife*. https://doi.org/10.7554/elife.00994.

Changela, Anita, Alexandra Martins, Stewart Shuman, and Alfonso Mondragón. 2005. "Crystal Structure of Baculovirus RNA Triphosphatase Complexed with Phosphate." *The Journal of Biological Chemistry* 280 (18): 17848–56.

Dent, J. A., M. M. Smith, D. K. Vassilatis, and L. Avery. 2000. "The Genetics of Ivermectin Resistance in Caenorhabditis Elegans." *Proceedings of the National Academy of Sciences of the United States of America* 97 (6): 2674–79.

Deshpande, Tarangini, Toshimitsu Takagi, Luning Hao, Stephen Buratowski, and Harry Charbonneau. 1999. "Human PIR1 of the Protein-Tyrosine Phosphatase Superfamily Has RNA 5′-Triphosphatase and Diphosphatase Activities." *Journal of Biological Chemistry*. https://doi.org/10.1074/jbc.274.23.16590.

Duchaine, Thomas F., James A. Wohlschlegel, Scott Kennedy, Yanxia Bei, Darryl Conte Jr, Kaming Pang, Daniel R. Brownell, et al. 2006. "Functional Proteomics Reveals the Biochemical Niche of *C. Elegans* DCR-1 in Multiple Small-RNA-Mediated Pathways." *Cell* 124 (2): 343–54.

Félix, Marie-Anne, Alyson Ashe, Joséphine Piffaretti, Guang Wu, Isabelle Nuez, Tony Bélicard, Yanfang Jiang, et al. 2011. "Natural and Experimental Infection of Caenorhabditis Nematodes by Novel Viruses Related to Nodaviruses." *PLoS Biology* 9 (1): e1000586.

Franz, Carl J., Hilary Renshaw, Lise Frezal, Yanfang Jiang, Marie-Anne Félix, and David Wang. 2014. "Orsay, Santeuil and Le Blanc Viruses Primarily Infect Intestinal Cells in Caenorhabditis Nematodes." *Virology*. https://doi.org/10.1016/j.virol.2013.09.024.

Guo, X., and R. Lu. 2013. "Characterization of Virus-Encoded RNA Interference Suppressors in Caenorhabditis Elegans." *Journal of Virology*. https://doi.org/10.1128/jvi.00148-13.

Gu, Weifeng, Heng-Chi Lee, Daniel Chaves, Elaine M. Youngman, Gregory J. Pazour, Darryl Conte, and Craig C. Mello. 2012. "CapSeq and CIP-TAP Identify Pol II Start Sites and Reveal Capped Small RNAs as *C. Elegans* piRNA Precursors." *Cell*. https://doi.org/10.1016/j.cell.2012.11.023.

Jiang, Hongbing, Carl J. Franz, Guang Wu, Hilary Renshaw, Guoyan Zhao, Andrew E. Firth, and David Wang. 2014. "Orsay Virus Utilizes Ribosomal Frameshifting to Express a Novel Protein That Is Incorporated into Virions." *Virology* 450-451 (February): 213–21.

McDonald, W. Hayes, David L. Tabb, Rovshan G. Sadygov, Michael J. MacCoss, John Venable, Johannes Graumann, Jeff R. Johnson, Daniel Cociorva, and John R. Yates 3rd. 2004. "MS1, MS2, and SQT-Three Unified, Compact, and Easily Parsed File Formats for the Storage of Shotgun Proteomic Spectra and Identifications." *Rapid Communications in Mass Spectrometry: RCM* 18 (18): 2162–68.

Park, Sung Kyu, John D. Venable, Tao Xu, and John R. Yates. 2008. “A Quantitative Analysis Software Tool for Mass Spectrometry–based Proteomics.” *Nature Methods*. https://doi.org/10.1038/nmeth.1195.

Peng, Junmin, Joshua E. Elias, Carson C. Thoreen, Larry J. Licklider, and Steven P. Gygi. 2003. “Evaluation of Multidimensional Chromatography Coupled with Tandem Mass Spectrometry (LC/LC-MS/MS) for Large-Scale Protein Analysis: The Yeast Proteome.” *Journal of Proteome Research* 2 (1): 43–50.

Sankhala, Rajeshwer Singh, Ravi Kumar Lokareddy, and Gino Cingolani. 2014. “Structure of Human PIR1, an Atypical Dual-Specificity Phosphatase.” *Biochemistry*. https://doi.org/10.1021/bi401240x.

Xu, T., S. K. Park, J. D. Venable, J. A. Wohlschlegel, J. K. Diedrich, D. Cociorva, B. Lu, et al. 2015. “ProLuCID: An Improved SEQUEST-like Algorithm with Enhanced Sensitivity and Specificity.” *Journal of Proteomics* 129 (November): 16–24.

CHAPTER 2

A Convenient Strategy to Clone Modified/unmodified small RNA and mRNA for High Throughput Sequencing

ABSTRACT

The high-throughput sequencing has become a standard tool for analyzing RNA and DNA. This method usually needs a cDNA/DNA library ligated with specific 5' and 3' linkers. Unlike mRNA, small RNA often contains modifications including 5' cap or triphosphate and 2'-O'Methyl, requiring additional processing steps before linker additions during cloning processes; due to low expression levels, it is difficult to clone small RNA with a small amount of total RNA. Here we have developed a new strategy to clone modified/unmodified small RNA in an all-liquid-based reaction carried out in a single PCR tube with as little as 20 ng total RNA. The 7-hour cloning process only needs ~1-hour labor. Moreover, this method can also clone mRNA, simplifying the need to prepare two cloning systems for small RNA and mRNA. Since the linkers are derived from the linkers used for DNA library construction, the barcoded PCR primers, which are based on the linker sequences, can be used to obtain cDNA/DNA amplicons for small RNA/mRNA/DNA high-throughput libraries. Not only is our method more convenient for cloning modified RNA than available methods, but it is also more sensitive, versatile and cost-effective. Moreover, the all-liquid-based reaction can be performed in an automated manner.

INTRODUCTION

To clone RNA, RNA is converted to cDNA with 5' and 3' linkers added for cDNA amplification and sequencing. mRNA can be fragmented and then used to make cDNA with 5' and 3' linkers added using various methods including ligation and reverse transcription; small RNA is usually first ligated with 5' and 3' linkers and then converted to cDNA. In the ligation based methods, the 5' ligation needs a 5' monophosphate (p) and the 3' ligation needs a 3'OH on target RNA.

In the introduction, we reviewed several ligation strategies that bypass 5’ ligation difficulty since the corresponding 3' end of cDNA does not have any modification. However, the 5' ligation of RNA usually serves as a selection for 5'p-RNA including miRNA, Dicer-dependent siRNA and piRNA. Without it the final library may contain a significant fraction of cDNA derived from degraded RNA, which usually bears 5'-OH. Not only does this contamination reduce the cloning yield of authentic RNA, but it also generates artifacts, leading to wrong conclusions. Moreover, bypass mechanisms equalize RNA 5' ends, meaning that specific modification information is lost. Therefore, these methods can not be used to only clone or enrich a specific group of 5' modified RNA under normal conditions. We have developed a unified strategy for constructing high-throughput small RNA/mRNA libraries. Since the linkers are derived from a strategy for making high-throughput DNA libraries, the barcoded PCR primers here can be used to amplify high-throughput DNA libraries too. Our strategy is capable of cloning modified and unmodified small RNAs using the enzymatic treatments including dephosphorylation of the RNA by *C. elegans* PIR-1 and decapping RNA by human Decap2 (hDcp2). Unlike the previously reported methods, these treatments are co-performed in the linker ligation reactions, avoiding the steps for enzyme removal/inactivation and/or buffer exchange. All the steps are performed inside a PCR tube in an all-liquid-based manner, significantly reducing labor time. This method is able to construct a small

RNA library using as little as ~20 ng of total RNA, a level much lower than the amount required by available commercial kits. Since it is all-liquid-based, it can be adapted to automation. The cost is minimal since the method only needs a few common enzymes, which can be purchased or easily purified using a single His-tag purification.

In all, our strategy is more sensitive, convenient, versatile and cost-effective than most of available methods.

MATERIALS AND METHODS

Adenylylation of the 3' linker

T4 DNA ligase was used to adenylylate the 3' linker oligo 5'p-AGATCGGAAGAGCACACGTCTGAACTCCAGTCA/ddC/, which together with 5'ACGGCATACGAGGGAAG/ddC/ was annealed to 5'CTCTTCCGATCTGCTTCCCTCGT A/ddC/ at 95 °C for 2 minutes followed by a slow cooling (0.1 °C/second) to room temperature, forming a nicked double-stranded DNA at 10 µM concentration in 10 mM Tris pH 7.5 and 30 mM KCl. Then 2.5 µM of the annealed DNA was incubated with 2.5 µM of T4 DNA ligase in 1X DNA ligation buffer at 37 °C for one hour, followed by phenol extraction, DNA precipitation and purification using a 15% PAGE/7M urea gel.

Partial digestion of RNA using nuclease P1

8-2000 ng total RNA was partially digested with 0.01 unit of nuclease P1 in a buffer containing 50 mM sodium citrate (pH 7.0) and 10 mM MgCl2 at 60 °C for 10 minutes, generating RNA fragments with a median size of 150-nt.

Bioinformatic analysis

High-throughput sequences were analyzed using our previous custom PERL (5.10.1) scripts and Bowtie 0.12.7 (Langmead et al. 2009; Gu et al. 2012). For *C. elegans* analyses, reads were mapped to the genome (WormBase release WS215) and the Generic Genome Browser was used to visualize the alignments (Stein et al. 2002). The software package is stored at https://github.com/guweifengucr/Wglab_small_RNA_analysis; the high throughput data GSE129664 is accessible using https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE129664 with token axebgasejniltox; and the genome browser track is accessible using http://wglabpred.dyn.ucr.edu/cgi-bin/gbrowse/wg120.

RESULTS

Designing linkers and primers for cloning small RNA and mRNA based on the DNA cloning system

Different linkers are used for cloning small RNA, mRNA and DNA due to distinct reaction mechanisms. We aimed to use the same linker system to clone all of them so that we can amplify the final DNA/cDNA amplicon using the same PCR primers including 64 barcoded 3' primers and one 5' primer. Our design is based on the Illumina platform simply due to its popularity. For DNA cloning, the last 13 nucleotides (nt) of the 5' linker and the first 13-nt of the 3' linker ligated to each target DNA strand form double-stranded DNA, a commonly used strategy required for DNA-based ligation (Figure I.2) (Illumina 2012). Since we desired a unified linker system, our linkers for RNA cloning was designed based on a DNA cloning system. However, the 13-nt sequences are not used for designing PCR primer simply to avoid primer specificity issues. The 5' linker 5'OH-ACACUCUUUCCCUACACGACGCUCUUCCGAUCUOH is an RNA oligo with 5'OH blocking the RNA ligase-mediated activation (adenylylation with ATP) (Figure 2.1). Therefore, this linker can only serve as a ligation acceptor using the 3'OH but not a ligation donor with the 5' OH. In contrast, the 3' linker

5'App- AGATCGGAAGAGCACACGTCTGAACTCCAGTCAddC

is an activated DNA oligo (adenylylation) which allows for ligation without ATP, i.e., serving as a ligation donor with the 3' ddC (dideoxyC) blocking its usage as a ligation acceptor (Figure 2.1). This design determines the ligation direction, i.e., the 3' linker to the target and then the 3' linker-target to the 5' linker. These linkers are further extended to contain the full-size linkers in PCR reactions using

5'AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGA and
5'CAAGCAGAAG ACGGCATACGAGAT-NNNNNNNN-GTGACTGGAGTTCAGACGTGT,

in which N represents an 8-nt barcode and the italic parts represent the sequences derived from the ligation linkers (Figure 2.1B).

A complete list of the ligation linkers, reverse transcription primer and PCR primers including 64 barcoded 3' primers are provided in Appendix B-Supplementary File 1.

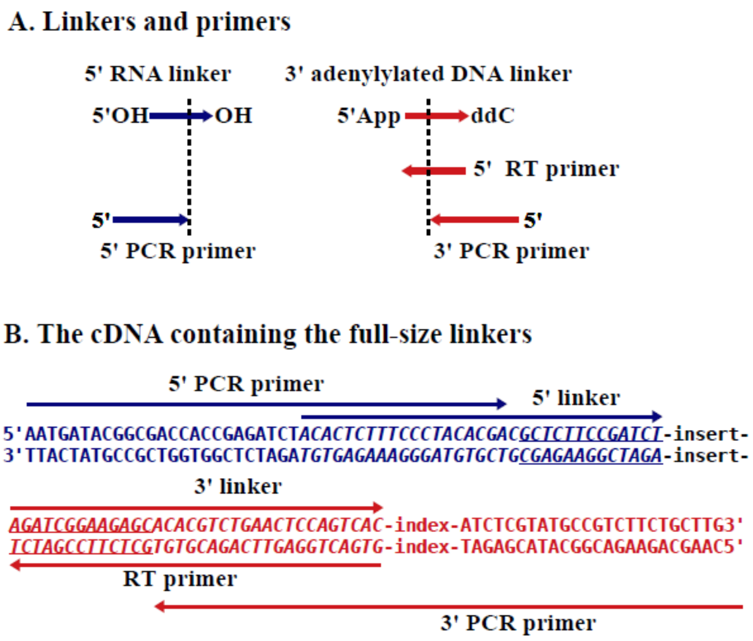

A. Linkers and primers
5' RNA linker
3' adenylylated DNA linker
5'OH
OH
5'App
ddC
5' RT primer
5'
5'
5' PCR primer
3' PCR primer
B. The cDNA containing the full-size linkers
5' PCR primer
5' linker
5'AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT-insert-
3'TTACTATGCCGCTGGTGGCTCTAGATGTGAGAAAGGGATGTGCTGCGAGAAGGCTAGA-insert-
3' linker
AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC-index-ATCTCGTATGCCGTCTTCTGCTTG3'
TCTAGCCTTCTCGTGTGCAGACTTGAGGTCAGTG-index-TAGAGCATACGGCAGAAGACGAAC5'
RT primer
3' PCR primer

Figure 2. 1. The linker and primer design.

A) The linkers and primers: the 5' linker is an RNA oligo with 5'OH and 3'OH; the 3' linker is an adenylylated DNA oligo with 5'App and 3' ddC (dideoxyC); the reverse transcription primer is reverse complementary to the 3' linker; and the 5' and 3' PCR primers partially overlap with the 5' and 3' linkers respectively. B) the sequence feature of the final cDNA: the amplicon is drawn as a double-stranded DNA with 5 arrows indicating the primer/linker sequences; the underlined parts are the 13-nt inverted repeats flanking the inserts; and the index represents an 8-nt barcode.

Constructing a small RNA high-throughput sequencing library

We first developed a protocol to clone small RNAs. As shown in Figure 2.2 and (Appendix C-Supplementary File 2), small RNA is first ligated with 0.5 µM activated 3' linker using 0.5 µM truncated T4 RNA ligase 2 in 10 µl buffer containing 50 mM Tris (pH 7.5), 10 mM DTT and 10 mM MgCl2. Since the 3' linker is 3' ddC-modified primarily for blocking 3' linker self-ligation, it can only be ligated to the 3'OH of target RNA. This strategy also avoids the activation of the 5' phosphate of target RNA since 1) no ATP is used and 2) the truncated T4 RNA ligase 2 does not possess the adenylylation activity (Ho and Shuman 2002; Nandakumar et al. 2004; Aravin and

Tuschl 2005). To clone the 3'-end 2'-O-methylated RNA including piRNA and some miRNA, 25% PEG-8000 is added (Munafó and Robb 2010). T4 RNA ligase 1 can substitute for the truncated T4 RNA ligase 2 (Gu et al. 2011). However, it can activate (adenylylation) target RNA even without the addition of ATP likely because: 1) at least a fraction of T4 RNA ligase 1 is bound with ATP or adenylylation; and 2) T4 RNA ligase 1 transfers AMP from the activated 3' linker to adenylate target RNA (Gu et al. 2011). This activation may cause target RNA circularization or target RNA-RNA ligation, decreasing the yield of target RNA-linker ligation. However, by controlling the amount of T4 RNA ligase 1 and ligation time, a satisfactory result could be easily achieved (Gu et al. 2009). T4 RNA ligase 1 does have an advantage since it can ligate 2'-O-methylated RNA more efficiently than the truncated T4 RNA ligase 2 (Gu et al. 2011, 2012).

After the 2-hour 3' ligation, the reaction is first heat-inactivated for: 1) denaturing the truncated T4 RNA ligase 2; and 2) annealing with ~0.5 μM reverse transcription (RT) oligo to the 3' linker either ligated or unligated by adding 0.5 μl of a 10 μM RT oligo. And then 0.4 μM 5' linker, 0.5 mM ATP, 0.25 μM RNA ligase 1 and water are added to the reaction reaching a final volume of 20 μl for the 5' ligation. T4 RNA ligase 1 activates the 3'-ligated target RNA if it contains or has been enzymatically treated to expose 5' monophosphate (5'p). In contrast, the 5' linker lacking 5'p cannot be activated. Therefore, the 5' linker can only serve as a ligation acceptor ligated to the 3'-ligated and 5'-activated target RNA. The annealing of the RT oligo with the 3' linker may reduce the ligation of the free 3' linker, if any left after the 3' ligation, with the 5' linker, since T4 RNA ligase 1 prefers single-stranded substrates. This strategy may significantly reduce the formation of 5' linker-3' linker ligation especially when the RNA substrate is much less than the 3' linker, generating excessive 3' linkers, as discussed below in the single worm RNA cloning.

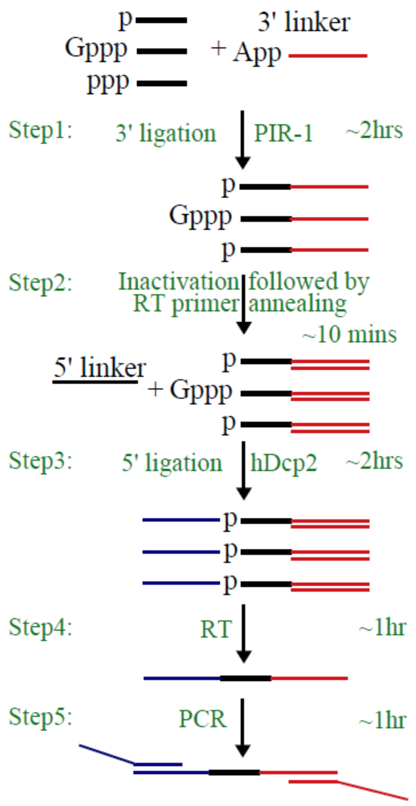


Figure 2. 2. Strategy to make a small RNA library.

RNA is ligated to an adenylylated 3' linker using the truncated RNA ligase 2 while 5' ppp-RNA is dephosphorylated with PIR-1 at step 1; the reaction is inactivated at 65 °C for 10 minutes and a reverse transcription (RT) primer is annealed to the 3' linker at 65 °C for 5 minutes at step 2; a 5' linker is ligated with the target RNA using T4 RNA ligase 1 while hDcp2 is added to decap capped RNA at step 3; a reverse transcription is performed to obtain the first strand cDNA at step 4; and cDNA is amplified and extended to obtain full-size 5' and 3' linkers at step 5.

A 30-minute reverse transcription step follows the 5' ligation with addition of ~0.2 µM Superscript II or III (Invitrogen), ~4 mM additional DTT, ~0.4 mM dNTP and ~2 µl 10X RT dilution buffer (0.25 M Tris pH 8.8 and 0.75 M KCl) reaching a final volume of ~24 µl. Then the enzymes are inactivated at 85 °C for 5 minutes. The obtained cDNA is amplified and extended to obtain the full-size linkers by PCR. 8-nt barcodes are inserted in the 3' PCR primers for labeling individual samples. A typical 50 µl PCR reaction is composed of 1X PFU buffer, 15 mM tetramethylammonium chloride for reducing primer dimmer, 0.1 mM dNTPs, 0.1 µM 5' and 3' primers, 5 µl RT reaction, and 1X PFU polymerase. The PCR is first amplified for 5 cycles (94 °C 20 s; 53 °C 20 s; 68 °C 30 s) and then amplified for 11 cycles (94 °C 20 s; 68 °C 40 s). Additional 0.6 µM 5' and 3' primers are added and the PCR is amplified for 2 more cycles (94 °C 20 s; 68 °C 40 s).

In all, the whole cloning process is performed in a single PCR tube with liquid components added sequentially and can be easily finished within ~7 hours including ~ 1-hour labor time. A formulated working protocol is presented in Appendix C-Supplementary File 2.

Achieving high reliability and sensitivity

Our ultimate goal is to use the above strategy to clone small RNA including miRNA, piRNA and siRNA, and fragmented mRNA. Since fragmented mRNA is basically small RNA, we only optimized the conditions to achieve high reliability and sensitivity using in vivo small RNA, and then applied these conditions to clone fragmented mRNA. We first examined our method by cloning small RNA using total RNA isolated from mouse testes, mouse ovaries, and *C. elegans* adult worms, and using purified *C. elegans* small RNA of size less than 200-nt, which contain miRNA, 22G-RNA (siRNA), 21U-RNA (piRNA), tRNA and 5.8S/5S rRNA (Ruby et al. 2006; Gu et al. 2009). As expected (Figure 2.3A), all the samples generated a ~150-base pair (bp) cDNA band which contains both the linker and the RNA insert. The size of the band in the testis sample is ~5 bps bigger than the bands in other samples since the major RNA species in the testes is the 25-30 nt piRNA while other samples contain the 20-23 nt small RNA (Ruby et al. 2006; Kirino and Mourelatos 2007; Gu et al. 2009; Tushir et al. 2009). We sequenced the library constructed from 0.5 µg *C. elegans* total RNA, and found that 94% of the cDNA was derived from authentic small RNA including miRNA, siRNA and piRNA. The size distribution and first nucleotide preference of each small RNA species are the same as reported (Gu et al. 2009). For example, miRNA peaks at 22-nt and prefers 5' U; 22G-RNA peaks at 22-nt and prefers 5' G; and 21U-RNA peaks at 21-nt and prefers 5' U (Fig. 3B).

The above experiment indicates that our method is very sensitive since we were able to make a small RNA library using 0.4 μg mouse ovary total RNA, which contains the least small RNA in all the samples (Figure 2.3A). To further determine the sensitivity threshold, we used a 2-fold titration of *C. elegans* total RNA ranging from 62.5 to 2000 ng as the substrate and observed a clear cDNA band with small RNA inserts from all the samples (Figure 2.3C). We also found that the method worked with 31 and 16 ng total RNA in a separate titration experiment (data not shown). All these experiments were performed using our standard condition in a single-tube-liquid-based manner.

To examine the reliability of our method especially when using a tiny amount of total RNA, we decided to establish a complete protocol starting from isolating total RNA from single worms, cloning it and using bioinformatics to compare the results. A single worm contains 10- 20 ng total RNA from ~ 2,000 cells, as estimated from the total RNA yield extracted from ~10,000 adult worms. This amount is equal to the total RNA derived from ~500-1,000 mammalian cells. Based on the intensity of the ~22 nt small RNA band stained with Ethidium Bromide or SyBr Gold, we estimated that there is ~1 ng small RNA per 10 μg total RNA in adult worms. We isolated total RNA from single worms using proteinase K digestion and used it to clone the ~22 nt small RNA, which was ~ 1.5 pg or 0.21 fmole, as estimated using the above parameters. Given it is unlikely to obtain a 100% yield for single-worm RNA isolations, the actual amount of RNA used for cloning should be less than 0.21 fmole. For such a tiny amount of starting material, we decided to use half enzymes and linkers in the 3'/5' ligation and RT steps as compared to our standard procedure, minimizing the formation of linker-linker (no RNA insert) or other byproducts. The result clearly indicates that our method is able to clone small RNA at fmole level from single worms and deliver high reliability at this level, as the two individual worm samples exhibited consistent small RNA profiles (Figure 2.3D).

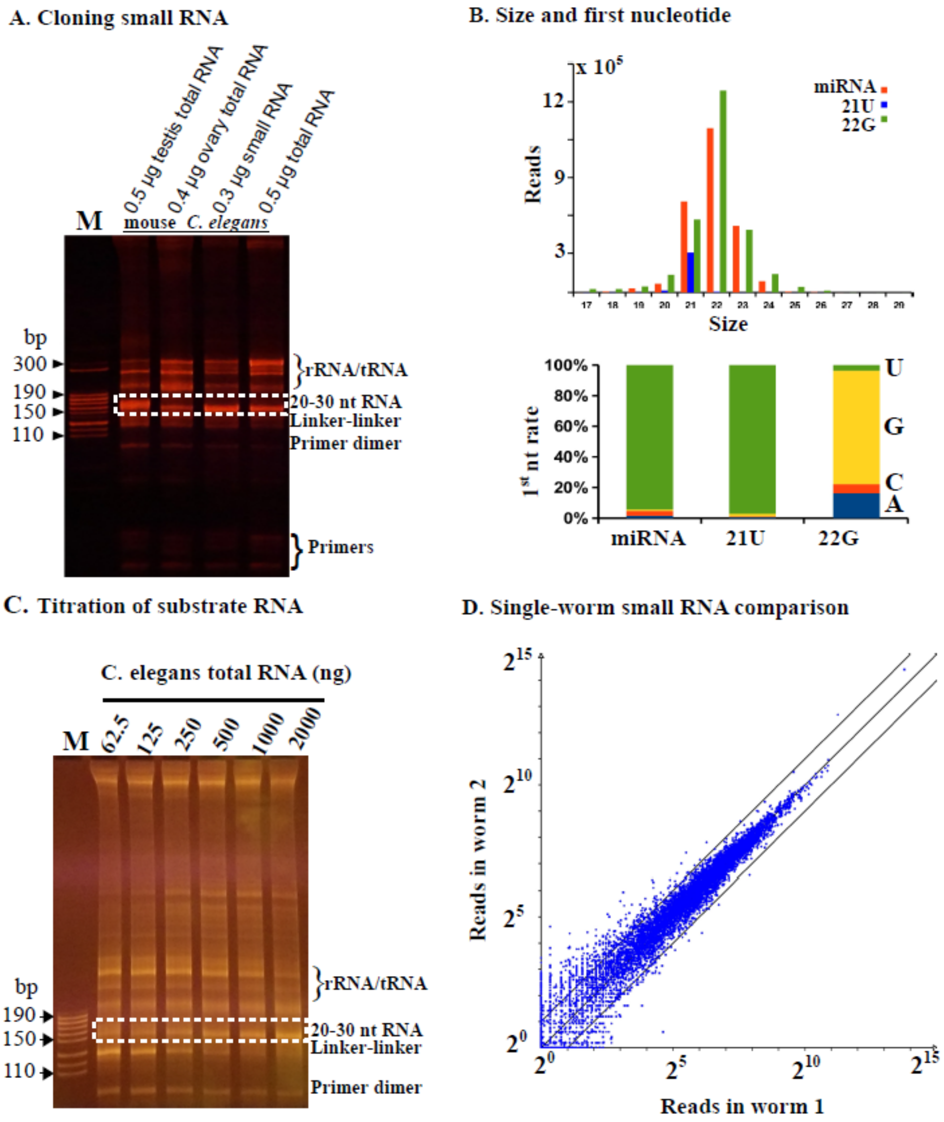


Figure 2. 3. Analysis of sensitivity and reliability of the cloning strategy.

A) 0.5 µg mouse testis total RNA, 0.4 µg mouse ovary total RNA, 0.3 µg *C. elegans* small RNA of size less than 200-nt, and 0.5 µg *C. elegans* total RNA were used to clone small RNA, and 5 µl of each PCR product was resolved on an 8% native PAGE gel. The inserted RNA in the PCR product is labeled on the right as well as the linker-linker ligation product, primer dimers and free primers, as compared to the DNA size marker (M); the dotted box represents the 20-30 nt RNA inserts. B) The size and first nucleotide analyses of miRNA, 21U and 22G cloned using 0.5 µg *C. elegans* total RNA treated with PIR-1. C) a 2-fold serial titration of total RNA (62.5 to 2,000 ng) was used to examine the cloning sensitivity threshold with the dotboxed area representing the desired amplicon and 'M' representing the DNA size marker. D) The comparison of 22G derived from two single worms with each blue dot representing one gene with 'X' mapped 22G reads in worm 1 and 'Y' reads in worm 2.

Cloning 5' triphosphorylated RNA (ppp-RNA)

Small ppp-RNA can not be directly ligated at the 5' end. To clone it, ppp-RNA is usually treated with commercial RNA polyphosphatase, generating 5' monophosphorylated RNA (p-RNA), which is compatible with 5' ligation (Gu et al. 2009). However, the available enzymes need special buffers and temperature incompatible with RNA ligation. Therefore, ppp-RNA is usually pre-treated enzymatically for dephosphorylation, extracted with organic reagents for enzyme removal, and precipitated for buffer exchange. This process is tedious, time consuming and counterproductive for cloning efficiency (Gu et al. 2009). If we have adopted this strategy, the labor time for the cloning procedure would have doubled and the cloning efficiency would have decreased due to sample loss. We aimed to utilize an RNA polyphosphatase which works efficiently in the ligation condition. Previous studies have shown that human PIR1 dephosphorylates ppp-RNA, generating p-RNA. However, it works at 37 °C, a temperature incompatible with the ligation condition. Since PIR-1 is highly conserved and PIR-1 works at 20 °C, it is a perfect candidate for the desired activity. We obtained a recombinant PIR-1 of homogeneity from E. coli. To examine its dephosphorylation activity, we co-applied this enzyme in the 3' ligation step with the truncated T4 RNA ligase 2, generating 3' ligated and 5' p-RNA using the same reaction. Then the RNA was further 5' ligated. The high-throughput sequencing analysis confirmed that this strategy worked efficiently for cloning ppp-RNA. As shown in Figure 2.4A, p-RNA including miRNA and 21U-RNA was cloned efficiently with and without PIR-1. In contrast, 22G-RNA (ppp-RNA) was only efficiently cloned with PIR-1 (Figure 2.4A). The 22G proportion in the PIR-1 treated sample was ~ 50%, a ratio very close to the one reported previously using the samples pretreated with Tobacco Acid Pyrophosphatase (Gu et al. 2009).

Cloning capped small RNA

Capped small RNA (csRNA) or promoter-associated small RNA is generated during transcription initiation and bears a 5' G cap with size less than 200-nt (Gu et al. 2012; Affymetrix ENCODE Transcriptome Project and Cold Spring Harbor Laboratory ENCODE Transcriptome Project 2009; Taft et al. 2009). The cap structure is incompatible with 5' ligation by RNA ligases. Tobacco acid pyrophosphatase (TAP) had been routinely used to decap csRNA, generating 5' p-RNA, before the manufacturer discontinued this enzyme (Gu et al. 2012). Several alternative decapping enzymes have been developed including Edc1-fused Dcp1-Dcp2 (Paquette et al. 2018). Our goal is to find a decapping enzyme compatible with the ligation condition. We obtained a hDcp2 construct from Dr. Megerditch Kiledjian (Wang et al. 2002), purified a recombinant hDcp2 expressed from E. coli and examined its decapping activity in the ligation buffer. hDcp2 worked well in the 5' ligation step with T4 RNA ligase 1 (Figure 2.4B), as the capped RNA substrate can not be ligated (RNA-RNA in Figure 2.4B) or circularized with the ligase but became ligatable with the hDcp2 treatment. We then used hDcp2 to construct a csRNA library. Since csRNA is not as abundant as other small RNA species, we first dephosphorylated the abundant small RNA species using Calf Intestinal Phosphatase (CIP), making them 5' unligatable with T4 RNA ligase 1, and then applied hDcp2 in the 5' ligation reaction to decap csRNA, making it 5' ligatable. As shown in Figure 2.4C, both sense and antisense (anti) csRNAs around the promoter area of pab-1 were cloned and they were separated by around 150-nt, a typical distance as reported (Gu et al. 2012; Affymetrix ENCODE Transcriptome Project and Cold Spring Harbor Laboratory ENCODE Transcriptome Project 2009; Taft et al. 2009).

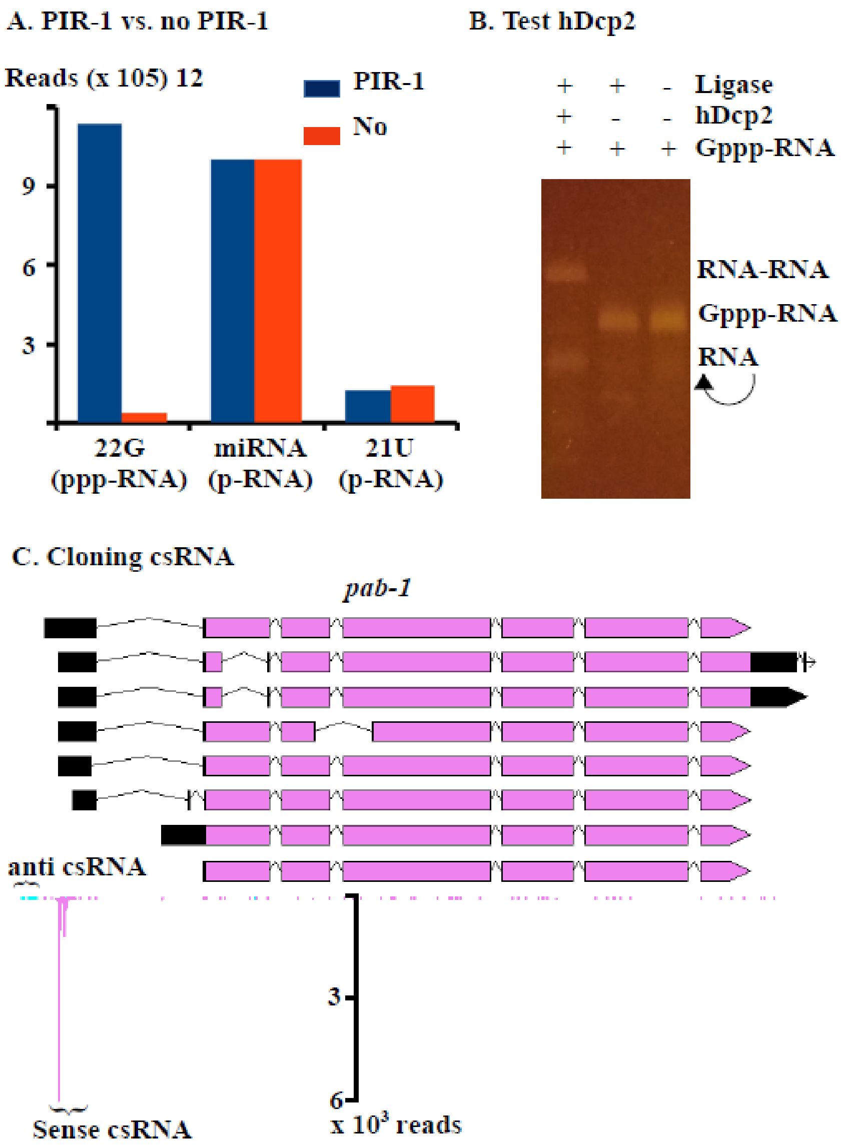


Figure 2. 4. Cloning of ppp-RNA and csRNA.

A) the comparison of small RNAs cloned with PIR-1 or no PIR-1, as normalized to the total miRNA, which is equally cloned in both methods; B) a capped RNA substrate is self-ligated (RNA-RNA) or circularized (a single RNA alone) by T4 RNA ligase 1 with hDcp2; C) cloned csRNAs within the pap-1 promoter region with 'pink' representing sense reads and 'blue' representing anti-sense reads.

In our condition, hDcp2 does not work well when added in the 3' ligation step (data not shown). Since the buffer conditions for the 3' and 5' ligations are almost the same, we suspect that the hDcp2 we purified may be contaminated with a low level of RNA phosphatases, which dephosphorylated the p-RNA generated by hDcp2 in the 2-hour 3' ligation step, making it incompatible for the following 5' ligation. In contrast, when hDcp2 is added in the 5' ligation, the ligase may add a linker to the 5'p of the decapped RNA before dephosphorylation by any RNA phosphatase.

Cloning mRNA

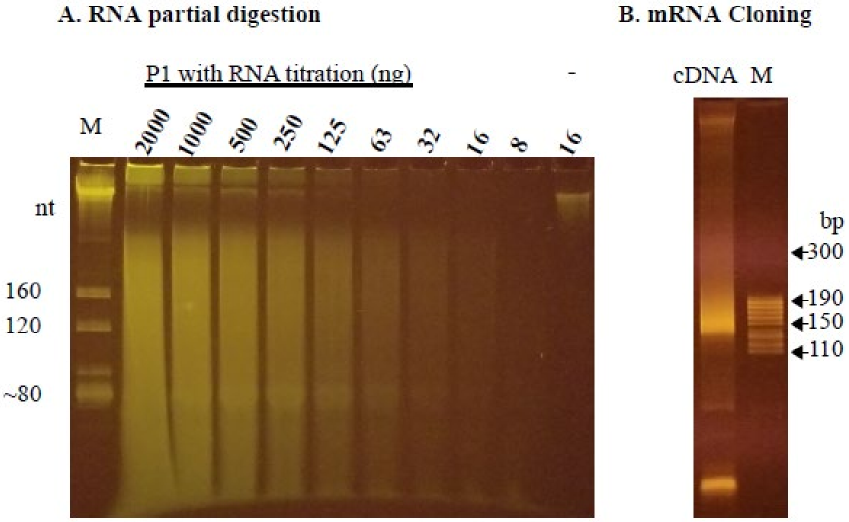


Figure 2. 5. Cloning mRNA.

A) Partial digestion of RNA using nuclease P1. 'M', a total RNA size marker with the estimated sizes using 5.8S rRNA (~160-nt), 5S rRNA (~120-nt) and tRNA (~80-nt); a 2-fold serial titration of substrate total RNA from 2,000 to 8 ng, was partially digested with nuclease P1; '-' is the negative control using ~16 ng purified mRNAs incubated without P1. B) fragmented mRNA was cloned as a cDNA library and resolved on an 8% native PAGE gel, as compared to the size marker M.

In addition to cloning small RNA, we aimed to use the same method to clone mRNA. In general, for mRNA cloning, mRNA is fragmented to small RNA, ligated with linkers, and converted to cDNA. However, a partial digestion of mRNA either chemically or enzymatically usually generates small RNA with 5'OH and 2'p or 3'p or cyclic phosphate at the 3' end. For ligation-based cloning methods, these ends have to be converted to 5'p and 3'OH. To avoid these conversion steps, we utilized nuclease P1, an enzyme cutting single-stranded RNA or DNA, generating small RNA or DNA with 5'p and 3'OH (Fujimoto et al. 1974). A potential challenge for enzyme-mediated partial digestion is how to control over/under-digestion when RNA/enzyme ratios vary. It is very inconvenient to figure out a specific condition for each sample especially when sample concentrations are not quantifiable. By diluting the enzyme and examining different buffers, we eventually obtained a condition under which the substrate RNA amount from 8 ng to 2000 ng did not affect the size of fragmented RNA with a given amount of enzyme (Figure 2.5). We speculate that this condition was achieved because of an extremely quick interaction (digestion) between P1 and substrate RNA, meaning P1 almost behaves like a free-moving enzyme which, regardless of substrate RNA concentrations, crosses any given unit of distance at the same frequency. To minimize the effect of secondary RNA structures on digestion, the reaction is performed at 60 °C, a temperature at which nuclease P1 works very efficiently. A neutral pH buffer is adopted to minimize RNA degradation by hydrolysis at 60 °C. In all, we have established an RNA fragmentation method which is substrate-concentration-independent, insensitive to RNA secondary structures, and ligation-compatible. Moreover, this method is very quick, taking 10 minutes, very convenient, requiring one enzyme and a simple buffer, and inexpensive, costing a few cents per reaction.

We then constructed a library with poly (A)-containing mRNA fragmented using nuclease P1. The obtained mRNA profile is similar to that of a previous one made using alkali hydrolysis. For example, the reads spread randomly along the length of mRNA and the individual gene expression levels correlate well between the two samples. Moreover, our method was capable of cloning at least 1 ng mRNA (data not shown), a sensitivity level sufficient for most routine work.

Minimizing bulged PCR products to reduce amplification bias caused by PCR overcycle

For any given amount of template DNA, it is difficult to predict how many PCR cycles are needed to generate enough products with a minimal amount of undesirable byproducts such as primer dimers. In laboratory routines, PCR reactions are usually empirically overcycled to maximize yields. Under such conditions, primers are used up first and then product DNA is denatured and renatured without amplification, resulting in futile cycles. This may not cause any observable effect on PCR specificity or even yields, since PCR conditions are usually so stringent that only one product is generated per reaction, and thus denaturing/renaturing processes in overcycled PCR reactions do not change the product. However, overcycle may cause serious issues for constructing a library containing more than one product. For example, since different insert DNA/cDNA molecules are flanked by same 5' and 3' linkers in high throughput sequencing libraries, linker regions anneal to each other intramolecularly and/ or intermolecularly once PCR reactions are overcycled. However, insert DNA/cDNA strands may not find their complementary strands, resulting in bulged products, i.e., molecules with perfectly base-paired linkers flanking two single-stranded or partially annealed insert strands. Bulged products appear as smears running much slower on native PAGE gels than perfectly matched products since they contain inserts with different DNA compositions and secondary structures (Figure 2.6A). Depending on overcycle number, bulged products could extend to a broad size range, leading to more purification work and reduced yield.

Bulged products are biased for less abundant DNA. In a PCR reaction generating only 2 products, AA and BB of the same size, three products AA, AB, and BB are generated by overcycling. If AA is 99% of AA+BB and BB is 1% before overcycling, AA, AB and BB are 98.01%, 1.98% and 0.01% respectively after overcycling. In the bulged products, both the A and B compositions are 50%, while in the perfectly matched products, AA is ~99.99% (98% / (98%+0.01%)) and BB is ~0.01% (0.01% / (98%+0.01%)). Therefore, if perfectly matched products are selected, abundant DNA species are over-represented, and if bulged products are selected, less abundant DNA species are over-represented.

To solve the overcycle issue, theoretically both perfectly matched and bulged products can be purified together. However, PCR products often contain primer dimers and undesirable products, which form bulged products with desirable PCR products once PCR reactions are overcycled, basically making this purifying-all strategy ineffective. To solve this issue, we used to examine the PCR product at cycle number 12, 16, 20, and 24 from the same PCR reaction on a PAGE gel, obtain the maximal PCR number without overcycle, and use this number to mass-produce the product (Gu et al. 2011). Although this solution worked well, it was tedious and the PCR reactions were usually overcycled at cycle 24. So we had to run another PCR reaction and gel purification to obtain the non-overcycled product. We have developed a new strategy to solve the overcycle problem with much less workload. We first amplify cDNA using 0.1 μM primers for 16 cycles and then add 0.6 μM primers for 2 more cycles. If a PCR reaction is overcycled at cycle 16, the two additional cycles convert it back to a nonovercycled reaction; if a reaction is not overcycled, the two more cycles can not make it overcycled since the added 0.6 μM primers are excessive.

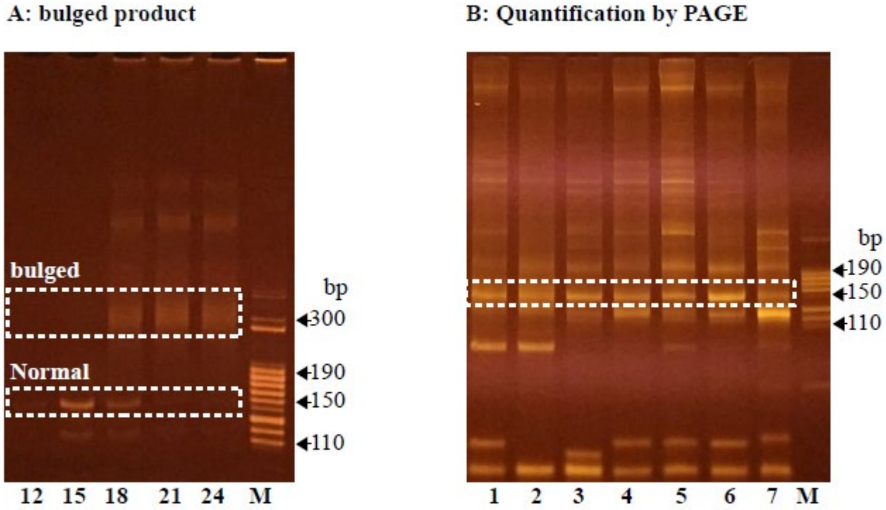


Figure 2. 6. Quantification using a native PAGE gel.

A) PCR reactions produced perfectly base paired normal (dotted box) products at low cycle numbers (12 and 15) and then bulged PCR products (dotted box) when the reactions were overcycled (18, 21 and 24); B) Quantification using a PAGE gel: 5 µl PCR product was resolved using an 8% native PAGE gel. Dot-boxed are the small RNA amplicons for sample 1-7; M, the size marker.

A convenient quantification and pooled purification strategy

A high-throughput sequencing library usually contains multiple barcoded samples for cost sharing. To obtain a specific composition, individual samples are purified, quantified and then mixed as a pooled sample. Since samples may have primer dimers and/or other byproducts including cDNA derived from rRNA and tRNA, a gel purification is required for obtaining target cDNA of specific sizes. This process is tedious and time-consuming (Gu et al. 2011). To reduce workload, we developed a two-step method: 1) visually comparing the relative concentration of target DNA of specific sizes, say 140-170 bps containing miRNA, siRNA and piRNA; 2) pooling samples according to a desired ratio and gel-purifying the pooled sample (Figure 2.6B). This way we were able to easily make a library consisting of 60 samples within 8 hours. Although the visual quantification may not be perfect, we found that the variation was usually within 3 folds.

DISCUSSION

Our major goal is to design a simple, convenient and cost-effective method for cloning small RNA. For this purpose, we have developed an all-liquid-based multi-step reaction in a single PCR tube. The whole procedure takes ~ 7 hours with only ~ 1-hour labor time and the rest primarily used for increasing ligation efficiency of modified RNA. It can be shortened to ~4 hours for cloning unmodified miRNA (0.5 and 1-hour for the 3' and 5' ligation respectively, 30 minutes for RT and 1 hour for PCR). Like the commercial kits, our method starts with total RNA, a strategy providing convenience but generating byproducts derived from tRNA and rRNA. In addition, the whole procedure involves no gel purification or DNA precipitation (the pooled DNA amplicons are still gel-purified, as discussed below). Our method is much more sensitive than the commercial kits, working with as little as 16 ng total RNA. We speculate that we could start with even less total RNA if using one gel purification to enrich target cDNA amplicons while removing primer dimmers and linker-linker ligation byproducts, followed by a second round of PCR amplification for mass-producing target cDNA amplicons.

We prefer ligation reactions for adding 5' linkers to target RNA over other techniques, e.g., cDNA circularization or template switching via reverse transcriptase. Since most degraded RNA bears a 5'OH, it can not be cloned using ligation, which requires a 5'p. Therefore, not only does the 5' ligation add a linker for cDNA application and sequencing, but also serves as a selection mechanism for enriching authentic small RNAs. Due to this selection, our method works well with samples heavily contaminated with degraded RNA including some immunoprecipitated

RNA samples. At the same time, our method can be easily adapted to clone RNA bearing a 5'OH simply by adding T4 Polynucleotide Kinase (PNK) in the 5' ligation step (data not shown).

Our method only needs a few common enzymes such as T4 RNA ligases, reverses transcriptase, DNA polymerases and PIR-1 and hDcp2. The last two are required for processing 5' modified RNA but not for p-RNA including miRNA, Dicer-dependent siRNA and piRNA. We easily obtained RNase-free enzymes using a single His-tag-mediated purification. The first three enzymes above are also commercially available. However, for some reasons, when tested by other labs, the protocol did not work well with some commercial enzymes, but worked well with our purified enzymes. We believe that the failure may be caused by the commercial T4 RNA ligase since commercial SuperScript III and Taq worked in our method. We speculated that the commercial T4 RNA ligases used may be contaminated with a trace amount of RNases or phosphatases. We estimate that the cost per library construction is negligible if using home-made enzymes and activated 3' linkers, and less than $10 if using commercial enzymes and 3' activated linkers.

Our method clones all types of small RNA including 5' and/or 3' modified RNA. Commercial kits are usually optimized for cloning small RNA with 5'p and 3'OH. Additional enzymatic steps and conditions are needed for processing csRNA, 5'ppp-RNA, 5'OH-RNA and 3'-modified RNA (Gu et al. 2012, 2009). These steps could easily double labor time and lead to reduced cloning efficiency. In our method, all steps are performed in one single liquid based reaction, significantly reducing workload. piRNA usually contains 2'-O-methyl at the 3' end, which decreases the 3' ligation efficiency. This inhibitory effect is overcome by the addition of 25% PEG-8000 (Munafó and Robb 2010). PIR-1 is used to modify ppp-RNA in the 3' ligation, generating p-RNA which is compatible with 5' ligation. This is a convenient strategy for cloning small RNA since ~50% of it is ppp-RNA (22G-RNA) (Gu et al. 2009). csRNA is decapped into 5' p-RNA by hDcp2 and ligated in the 5' ligation step. However, csRNA is usually expressed at an extremely low level, decapping only makes them 5' ligatable but not enriched (Gu et al. 2012). To enrich csRNA, 5' p-RNA, usually the major small RNA species, is dephosphorylated, making it 5' unligatable, and then csRNA is decapped for 5' ligation (Gu et al. 2012). 5' OH-RNA is cloned by addition of PNK in the 5' ligation step, as discussed below for mRNA cloning. Theoretically, if all these enzymes are used, the method allows for cloning of all cellular small RNAs.

We aimed to develop a versatile method for cloning both mRNA and small RNA and to use the same sets of PCR primers to amplify libraries derived from DNA and RNA. Since the 5' and 3' linkers are derived from the DNA cloning system, the PCR primers definitely work with DNA libraries. Actually we used these primers to amplify DNA libraries constructed using a commercial kit (data not shown) without purchasing the expensive barcoded primers. For mRNA cloning, we first used alkali-hydrolysis to fragment mRNA, generating small RNA with 5'OH and 3' cyclic phosphate. Then we used PNK to fix the ends in the 3' ligation step by: 1) removing the 3' cyclic phosphate at room temperature for ~5 hours; and 2) phosphorylating the 5' end with the addition of ATP at 37 °C for 1 hour (data not shown). Although it works well, the 3' ligation step takes ~ 6 hours, making it impossible to finish the cloning process within one day. This prompts us to use nuclease P1, which generates ligation-compatible ends. The cost of nuclease P1 is close to alkali-hydrolysis, basically nothing. Nuclease P1 works well at 60 °C, significantly reducing the effect of RNA secondary structures on the digestion pattern and efficiency. Moreover, our partial digestion condition is insensitive to the RNA amount so that there is no need to optimize the enzyme/RNA ratio for getting fragmented RNA of desired size.

All in all, our method is convenient, sensitive, and versatile. Any lab with basic molecular techniques can establish an integrated system to clone small RNA and mRNA.

REFERENCES

Affymetrix ENCODE Transcriptome Project, Cold Spring Harbor Laboratory ENCODE Transcriptome Project. 2009. Post-transcriptional processing generates a diversity of 5'- modified long and short RNAs. *Nature* **457**: 1028–1032.

Aravin A, Tuschl T. 2005. Identification and characterization of small RNAs involved in RNA silencing. *FEBS Lett* **579**: 5830–5840.

Batista PJ, Ruby JG, Claycomb JM, Chiang R, Fahlgren N, Kasschau KD, Chaves DA, Gu W, Vasale JJ, Duan S, et al. 2008. PRG-1 and 21U-RNAs interact to form the piRNA complex required for fertility in . *Mol Cell* **31**: 67–78.

Chen X. 2005. MicroRNA biogenesis and function in plants. *FEBS Lett* **579**: 5923–5931.

Conine CC, Batista PJ, Gu W, Claycomb JM, Chaves DA, Shirayama M, Mello CC. 2010. Argonautes ALG-3 and ALG-4 are required for spermatogenesis-specific 26G-RNAs and thermotolerant sperm in Caenorhabditis elegans. *Proc Natl Acad Sci USA* **107**: 3588–3593.

Fujimoto M, Fujiyama K, Kuninaka A, Yoshino H. 1974. Mode of action of nuclease P-1 on nucleic acids and its specificity for synthetic phosphodiesters. *Agricultural and Biological Chemistry* **38**: 2141–2147.

Goodwin S, McPherson JD, McCombie WR. 2016. Coming of age: ten years of next- generation sequencing technologies. *Nat Rev Genet* **17**: 333–351.

Gu W, Claycomb JM, Batista PJ, Mello CC, Conte D. 2011. Cloning Argonaute-associated small RNAs from Caenorhabditis elegans. *Methods Mol Biol* **725**: 251–280.

Gu W, Lee H-C, Chaves D, Youngman EM, Pazour GJ, Conte D, Mello CC. 2012. CapSeq and CIP-TAP identify Pol II start sites and reveal capped small RNAs as *C. elegans* piRNA precursors. *Cell* **151**: 1488–1500.

Gu W, Shirayama M, Conte D, Vasale J, Batista PJ, Claycomb JM, Moresco JJ, Youngman EM, Keys J, Stoltz MJ, et al. 2009. Distinct argonaute-mediated 22G-RNA pathways direct genome surveillance in the *C. elegans* germline. *Mol Cell* **36**: 231–244.

Ho CK, Shuman S. 2002. Bacteriophage T4 RNA ligase 2 (gp24.1) exemplifies a family of RNA ligases found in all phylogenetic domains. *Proc Natl Acad Sci USA* **99**: 12709–12714.

Illumina. 2012. Illumina Sequence Information for Customers.

Kircher M, Sawyer S, Meyer M. 2012. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Res* **40**: e3.

Kirino Y, Mourelatos Z. 2007. Mouse Piwi-interacting RNAs are 2'-O-methylated at their 3' termini. *Nat Struct Mol Biol* **14**: 347–348.

Ko J-H, Lee Y. 2006. RNA-conjugated template-switching RT-PCR method for generating an Escherichia coli cDNA library for small RNAs. *J Microbiol Methods* **64**: 297–304.

Kwon Y-S. 2011. Small RNA library preparation for next-generation sequencing by single ligation, extension and circularization technology. *Biotechnol Lett* **33**: 1633–1641.

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**: R25.

Munafó DB, Robb GB. 2010. Optimization of enzymatic reaction conditions for generating representative pools of cDNA from small RNA. *RNA* **16**: 2537–2552.

Nandakumar J, Ho CK, Lima CD, Shuman S. 2004. RNA substrate specificity and structure-guided mutational analysis of bacteriophage T4 RNA ligase 2. *J Biol Chem* **279**: 31337– 31347.

Pak J, Fire A. 2007. Distinct populations of primary and secondary effectors during RNAi in *C. elegans*. *Science* **315**: 241–244.

Paquette DR, Mugridge JS, Weinberg DE, Gross JD. 2018. Application of a Schizosaccharomyces pombe Edc1-fused Dcp1-Dcp2 decapping enzyme for transcription start site mapping. *RNA* **24**: 251–257.

Reuter JA, Spacek D, Snyder MP. 2015. High-Throughput Sequencing Technologies. *Mol Cell* **58**: 586–597.

Ruby JG, Jan C, Player C, Axtell MJ, Lee W, Nusbaum C, Ge H, Bartel DP. 2006. Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in *C. elegans*. *Cell* **127**: 1193–1207.

Stein LD, Mungall C, Shu S, Caudy M, Mangone M, Day A, Nickerson E, Stajich JE, Harris TW, Arva A, et al. 2002. The generic genome browser: a building block for a model organism system database. *Genome Res* **12**: 1599–1610.

Stiller M, Knapp M, Stenzel U, Hofreiter M, Meyer M. 2009. Direct multiplex sequencing (DMPS)--a novel method for targeted high-throughput sequencing of ancient and highly degraded DNA. *Genome Res* **19**: 1843–1848.

Taft RJ, Kaplan CD, Simons C, Mattick JS. 2009. Evolution, biogenesis and function of promoter-associated RNAs. *Cell Cycle* **8**: 2332–2338.

Tushir JS, Zamore PD, Zhang Z. 2009. SnapShot: Fly piRNAs, PIWI proteins, and the ping- pong cycle. *Cell* **139**: 634, 634.e1.

Wang Z, Jiao X, Carr-Schmid A, Kiledjian M. 2002. The hDcp2 protein is a mammalian mRNA decapping enzyme. *Proc Natl Acad Sci USA* **99**: 12663–12668.

CHAPTER 3

Influenza A Virus Utilizes Non-Canonical Cap-Snatching to Diversify mRNA/ncRNA Pools

ABSTRACT

Influenza A virus utilizes a special process, cap-snatching, to obtain a host capped small RNA for priming viral mRNA synthesis, generating hybrid capped mRNA for translation. Previous studies have been focusing on cap-snatching at the 5' end of viral mRNA. Here we report two non-canonical cap-snatching regions: one 300-nt upstream of the 3' end of each mRNA generating capped mRNA/ncRNA, and the other in the 5' region of vRNA and mapped primarily at the second nt, likely generating ncRNA. In both regions, cap-snatching mainly utilizes a base-pairing between the last nt G of a host capped RNA and a nt C in a template RNA to prime viral RNA synthesis, as the canonical mRNA cap-snatching. However, the nt upstream of this C is usually A/U rather than just U; prime-realignment occurs much less frequently. We also demonstrate that the influenza virus snatches virus-derived capped RNA in addition to host capped RNA. The non-canonical cap-snatching likely generates novel mRNA with the start AUG encoded in the viral RNA or host capped small RNA. These findings expand our understanding of the cap-snatching mechanism and suggest that the influenza A virus may utilize this process to diversify its mRNA/ncRNA.

INTRODUCTION

Influenza A virus (IAV) often causes epidemic and pandemic respiratory infection in humans and animals. Its genome contains eight negative-sense viral RNAs (vRNA). IAV utilizes a viral RNA-dependent RNA polymerase (RdRP) complex to generate positive-sense mRNA and complementary RNA (cRNA) from template vRNA, and negative-sense vRNA from template cRNA (Kobayashi et al. 1996; Shi et al. 1995; Shih and Krug 1996; Bouvier and Palese 2008; Guilligay et al. 2008; Sugiyama et al. 2009; Reich et al. 2014).

In the era of Sanger sequencing, it had been long held that IAV snatches 10-18 nt caps from host pre-mRNA (Krug et al. 1979; Bouloy et al. 1978; Plotch et al. 1979; Beaton and Krug 1981; Caton and Robertson 1980; Dhar et al. 1980; Plotch et al. 1981; Shaw and Lamb 1984). However, this conclusion may be incomplete at best since: 1) very limited sequencing data and gene annotations were available at that time; and 2) mRNA may be preferentially selected as cap donors out of multiple genomic matches resulted from the small query size of snatched host caps. Recently three groups including us used high-throughput sequencing to obtain a more comprehensive spectrum of host cap donors (Gu et al. 2015; Koppstein et al. 2015; Sikora et al. 2014, 2017). Both our group and Koppstein et al. independently demonstrated that noncoding RNA (ncRNA) is the top cap donor. For example, we showed that U1 or U2 snRNAs alone provided ~7% of viral caps and that ncRNAs including U1 and U2 provided ~55% (Gu et al. 2015). We also used *in situ* hybridization to verify the high-throughput sequencing result, considering the possibility that the U1/U2 caps were added to IAV mRNA via cloning artefacts. Unlike the other two groups which only sequenced the 5' ends of IAV mRNA using IAV-specific primers, we sequenced the 5' ends of all host and IAV capped RNA including capped ncRNA, allowing us to perform more comprehensive analyses and obtaining unique matches. Promoter associated small RNA (PASR) or capped small RNA (csRNA) constitutes a new ncRNA species, which is generated during Pol II-mediated transcription initiation (Seila et al. 2008; Affymetrix/Cold Spring Harbor Laboratory ENCODE Transcriptome Project 2009; Nechaev et al. 2010; Gu et al. 2012). PASR usually exhibits a bimodal distribution where sense PASR co-mapped to the same regions as the 5' end of pre-mRNA/mRNA and antisense PASR mapped ~150nts upstream in the antisense direction (Seila et al. 2008; Affymetrix/Cold Spring Harbor Laboratory ENCODE Transcriptome Project 2009; Nechaev et al. 2010; Gu et al. 2012). Although we cannot single out sense PASR from the host mRNA/pre-mRNA/sense PASR mixture to determine its contribution to IAV mRNA caps, we were to able to demonstrate that antisense PASR contributed ~7% of all IAV caps (Gu et al. 2015) and the snatching rate (IAV cap / (IAV cap + host cap) was much higher than that of mRNA/pre-mRNA/sense PASR. This raises a possibility that IAV caps which appear to be snatched from pre-mRNA may be derived from an alternative source, i.e., sense PASR. Another question is whether IAV allows for the transcription initiation and elongation of Pol II, generating pre-mRNA and PASR, or just for the initiation, generating PASR. Regardless, PASR is an important IAV cap source.

Our previous study focused on canonical cap-snatching occurring primarily at the first nt of IAV mRNA, defined as 'mRNA +1' for convenience (Gu et al. 2015). As shown in Figure 3.5, IAV mRNA, cRNA and vRNA are read from 5' to 3' as +1, +2, etc, and from 3' to 5' as -1, -2, etc. Here we identified non-canonical cap-snatching in two regions: one as a cluster of loci ~300-nt upstream of each mRNA 3' end (mRNA 3' cluster) and the other covering the +1 to +10 of each vRNA (vRNA 5' region). mRNA 3' clusters generate novel mRNA and capped ncRNA sharing same strands with annotated mRNA while vRNA 5' regions likely only produce ncRNA. The novel mRNA is usually in-frame with annotated mRNA but encoding shorter proteins with 0-3 amino acids derived from a host cap. We demonstrate that non-canonical cap-snatching also prefers G/C base-pairing for priming RNA synthesis as canonical cap-snatching at IAV mRNA +1. For all cap-snatching events, the initial priming prefers a G/C base-pairing between the last (-1) nt, usually G, of a cap and the template -2 C (mRNA+1 and vRNA 5' regions) or internal nt (mRNA 3' clusters), usually C. We also demonstrate that the nt downstream of the template C, usually A or U (collectively as W), plays important roles in selecting internal template C's in mRNA 3' clusters. Although cis-realignment, which shares same templates with initial priming/extension, utilizes the same A/U base-paring mechanism in canonical and non-canonical regions, it occurs less frequently in non-canonical regions. Our evidence indicates that trans-alignment does occur between two IAV templates, suggesting one IAV mRNA could contain three sequences, one derived from a host cap and the other two derived from two IAV mRNAs. In conclusion, this study provides further insight into the cap-snatching mechanism and suggests that IAV may use cap-snatching to diversity its mRNA and ncRNA.

MATERIALS AND METHODS

IAV infection

The cell culture and virus infection conditions were described previously (Gu et al. 2015). Briefly, A549 cells were incubated with influenza A/Brisbane/59/2007 (H1N1) at a multiplicity of infection 1 at 37°C for 1 hour, washed and cultured for 6, 12, 24 and 48 hours.

RNA extraction

RNA was extracted from infected cells using TRI reagents (Sigma) according to the manufacturer's protocol. The resulting aqueous solution was phenol/chloroform extracted and co-precipitated with 20 μg glycogen.

Obtaining the 5' end sequences of host and IAV RNA

To simultaneously analyze the 5' end of host and IAV RNA, high-throughput sequencing libraries were constructed using CapSeq and then sequenced (Gu et al. 2015). In brief, 2 µg of total RNA was processed using Terminator exonuclease (Epicentre) to remove rRNA and decapped using Tobacco Acid Pyrophosphatase (Epicentre). The linkers required for high- throughput sequencing were added to the 5' end of target RNA using ligation and to the 3' end using random priming in reverse transcription. We obtained cDNA containing 50-200-nt RNA inserts and sequenced the first 50 or 100 nts using HiSeq 2000. The data was stored at

http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?token=apwnwimmvnupfkr&acc=GSE67493 (Gu et al. 2015).

Bioinformatic analyses

The RefSeq IAV sequences lack parts of the 5' and 3' UTRs, which are critical for our analyses (Sikora et al. 2014; Pruitt et al. 2014). We used a custom PERL script to assemble the 5' UTRs of IAV mRNA, cRNA, and vRNA based on our CapSeq data and then obtained the corresponding 3' UTRs using reverse complement, as shown in Figure 3.5. The bioinformatic analyses were performed using custom PERL scripts and Bowtie 0.12.7, as described previously (Gu et al. 2015; Langmead et al. 2009). Since the previous analyses focused on the 5' end of mRNA, we modified the scripts to fit the analyses of vRNA 5' regions and mRNA 3' clusters. In brief, we mapped the first 40-nt sequence of each read to human genome and annotations, the Ensembl GRCh37 release 71 (Zerbino et al. 2018). We obtained the unmatched reads and then split them into two parts, position 1-20 and 21-40. The latter was mapped to IAV RNA, and the resulting full-size match was extended towards the 5' end of the reference sequence using the position 20-1 of the read with a score +1 and - 3 for each match and mismatch respectively. The longest extension was selected using the maximum score, and the sequence 5' of the matched part was extracted as a non-IAV sequence. We also removed the most frequent extension stops caused by indels due to sequencing errors or mutations. Gbrowse was used to generate histograms of IVA reads (Stein et al. 2002). We used a custom PERL script to predict the coding frame encoded by non-canonical capped RNA with the criteria: 1) it contains at least 50 amino acids; 2) there is a Kozak motif (A/G)NNAUGG in which AUG is the start codon encoded by IAV or host caps; 3) a poly (A) tail can be added using the 'stuttering' mechanism (Luo et al. 1991; Pritlove et al. 1998; Poon et al. 1999; Zheng et al. 1999).

RESULTS

CapSeq highly enriches reads mapped to the 5' ends of mRNA and vRNA

We previously used CapSeq to analyze canonical cap-snatching in the IAV-infected A549 cells. Unlike other analyses, which used PCR to specifically enrich the 5' ends of IAV mRNA, our method obtained a global profile of the 5' ends of host and IAV capped or triphosphorylated RNA (ppp-RNA) (Gu et al. 2012, 2015; Hainer et al. 2015). CapSeq utilizes sequential enzymatic treatments to enrich capped RNA in a one-pot reaction, including excess Terminator exonuclease (Terminator) to almost completely remove monophosphorylated RNA (p-RNA), predominantly rRNA, CIP to remove any residual p-RNA after the Terminator treatment, and Tobacco acid pyrophosphatase (TAP) to generate p-RNA from capped RNA and ppp-RNA at the last step. The CIP treatment is not a required step, only serving as a 'fail-safe' measure. We had used CapSeq to obtain data of high quality mapping the 5' end of capped RNA in non-infected cells (Gu et al. 2012, 2015; Hainer et al. 2015). However, these analyses did not allow us to assess if CIP played any accessory role in addition to the preceding Terminator treatment step for removing p-RNA. We could have used ppp-RNA, if existing, as an internal control to assess the efficacy of the CIP treatment since only CIP but not Terminator modifies ppp-RNA, making it unclonable. Moreover, we have never tried to optimize the CIP condition since our CapSeq protocol worked pretty well, removing 99.5% of rRNA and generating 50,000 to 100,000-fold enrichment for reads derived from the 5' end over any internal position of capped RNA (Gu et al. 2012, 2015). In this study, we found that CIP did not work in our protocol since we effectively cloned ppp-RNA (vRNA) from the IAV-infected cells (Figure 3.1 and Figure 3.2). This failure was likely resulted from insufficient enzymes since the added CIP, which was calculated empirically based on the concentration of intact substrate RNA, was highly likely overwhelmed by excessive monophosphorylated nucleotides (~3,000 folds as much as intact RNA), the product of the preceding Terminator- mediated rRNA depletion step in the 'one-pot' reaction. Another contributing factor is the double-stranded (ds) structures formed using vRNA 5' and 3' ends (Figure 3.5).

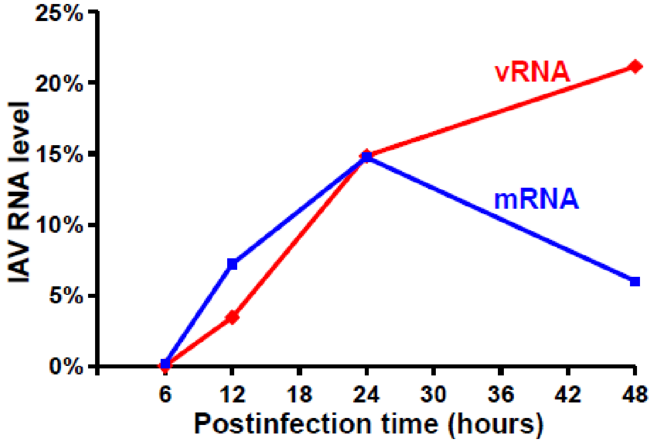


Figure 3. 1. The IAV RNA levels at varied postinfection timepoints.

The IAV mRNA (blue) and vRNA (red) levels, defined as 'IAV reads divided by total non-rRNA host/IAV reads', were measured as at 6, 12, 24 and 48-hr postinfection timepoints using the high-throughput sequencing data.

Although the CIP step did not work, we cloned little *in vivo* p-RNA, which was likely derived from specific RNA degradation pathways and usually mapped to the internal sites of IAV RNA. For example, ~97.9% of IAV mRNA reads were mapped to +1 and only ~2.1% were mapped to internal sites; ~92.0% of vRNA reads were mapped to +1 and 8.0% were mapped to internal sites (Figure 3.2). This constitutes ~77,000 and ~20,000-fold (97.9/2.1 X 1700 and 92/8 X 1,700) enrichment for reads starting at +1 over reads starting at any internal position, calculated based on the average size, ~1,700-nt, of IAV RNA. We believe that only a fraction of these internal reads represent p-RNA, since 99.5±0.3% of rRNA, the major p-RNA species in total RNA, was depleted by Terminator in the five samples analyzed. Consistent with this hypothesis, the majority of these internal sites on mRNA and a significant fraction of them on vRNA represent capped RNA synthesized via non-canonical cap-snatching, as discussed below. Since here we focus on non-canonical cap-snatching, the reads at vRNA +1, which contain a triphosphate group rather than a host cap, serve as 'unexpected' byproducts and the reads derived from p-RNA surviving the enzymatic depletion step may serve as an extremely low but sensible background noises. As discussed below, these byproducts and noises do not affect our capped RNA analysis. Moreover, they serve as 'perfect' negative internal controls for cloning artefacts since they are predominantly uncapped.

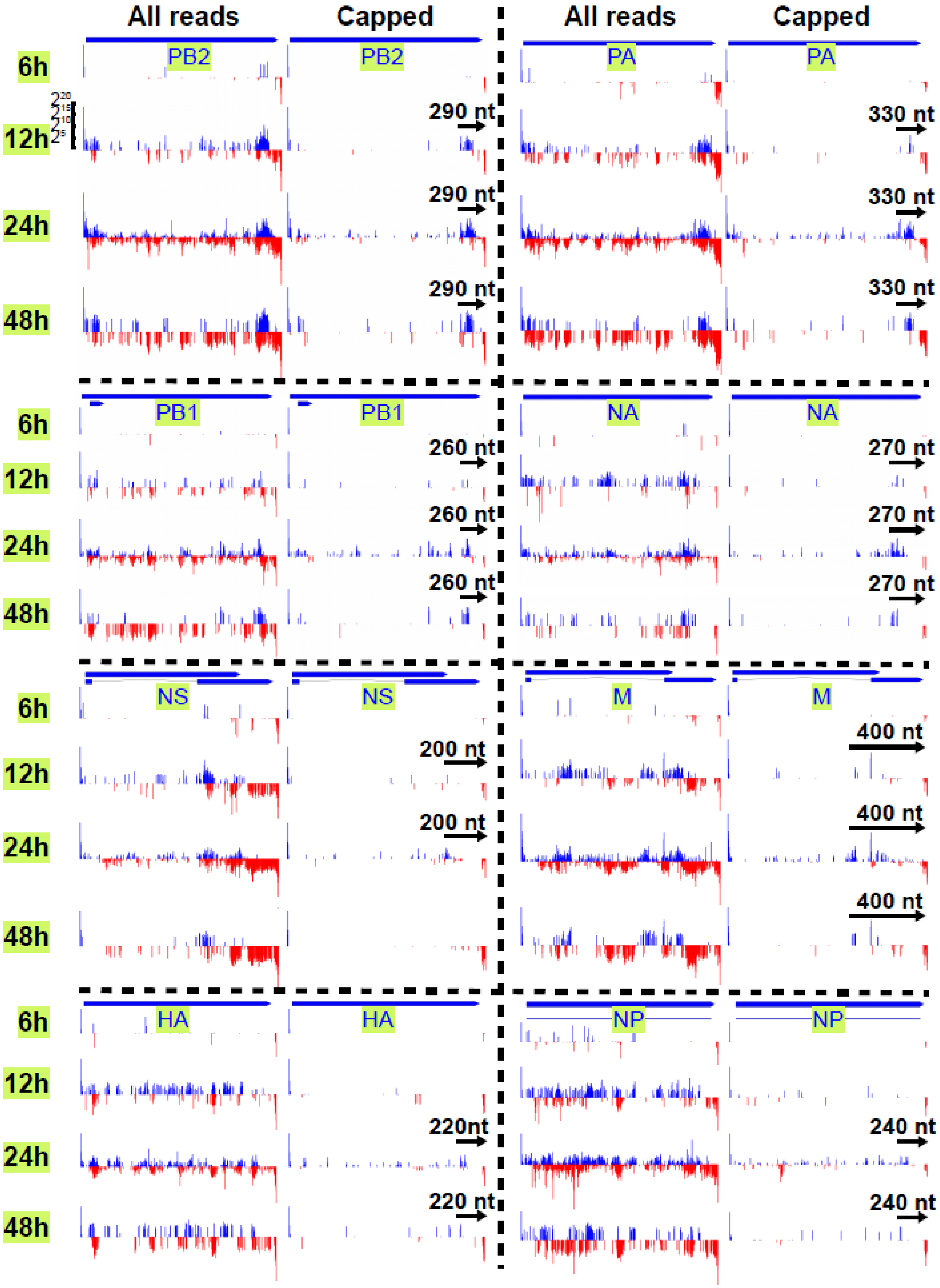

All reads
Capped
All reads
Capped
PB2
PB2
PA
PA
6h
12h
24h
48h
2^{20}
2^{15}
2^{10}
2^{5}
290 nt
290 nt
290 nt
330 nt
330 nt
330 nt
PB1
PB1
NA
NA
260 nt
260 nt
260 nt
270 nt
270 nt
270 nt
NS
NS
M
M
200 nt
200 nt
400 nt
400 nt
400 nt
HA
HA
NP
NP
220nt
220 nt
240 nt
240 nt

Figure 3. 2. The histogram of the IAV reads.

IAV RNA reads were represented by their first mapped nts with 'blue' indicating mRNA/cRNA and 'red' indicating vRNA, and each IAV position was visualized using the combined read number (Y-axis) derived from the co-mapped reads and normalized to the total non-rRNA host/IAV reads. For each IAV RNA strand, the blue annotations at the top represent the coding frames, which are a little bit smaller than the corresponding IAV RNA strands; the left part of each panel represents the histogram of all the IAV reads and the right panel represents that of only capped RNA reads at 6-48 hr postinfection timepoints. The black arrows indicate the distance between the left edges of mRNA 3' clusters and cRNA 3' ends. Each IVA RNA strand is represented with the same width (X-axis), generating different unit sizes. However, each panel uses the same log scale of the Y axis, as labeled in the top left panel.

Examination of the IAV infection time course

We previously analyzed canonical cap-snatching at the 12 and 24-hr postinfection timepoints (Gu et al. 2015). Here we added 6 and 48-hr timepoints to obtain a broader time course. Under our condition, IAV mRNA and vRNA levels were very low at the 6-hr timepoint, accounting for ~0.2 % and 0.04% of total host/IAV RNA, and increased almost linearly at the 12 and 24-hr timepoints, indicating active transcription and replication, and vRNA levels continuously increased while mRNA levels dropped dramatically at the 48-hr timepoint, indicating the late stage of IAV infection (Figure 3.1). We repeated the 6-hr timepoint and only reached a similar conclusion with the mRNA level at 0.5% and vRNA level at 0.08%. Apparently, these low expression levels prevents us from obtaining sufficient IAV reads for bioinformatic analyses at affordable cost using high-throughput sequencing. Therefore, our data analyses focused on the 12 to 24-hr timepoints. Since we obtained similar conclusions from all the three timepoints (Figure 3.2 and Appendix D-Figure S3.1) and the 12 and 24-hr timepoints exhibited higher levels of IAV mRNA (7.2% and 14.7% of host/IAV total RNA respectively), we used these two timepoints to represent the data in most of the analyses. The uninfected controls were also sequenced but excluded from the analyses simply due to lack of any viral reads (data not shown).

Identification of non-canonical cap-snatching

As a 5' ligation-dependent method, CapSeq allows us to obtain a directional library with cDNA sequence explicitly representing RNA, avoiding confusion when mapping reads derived from dsRNA. Since the RefSeq database lacked IAV 5' and 3' UTRs, we used the CapSeq reads and a PERL script to assemble the 5' UTRs, generating AGC(A/G)AAAGCAGG (G for PB1 and A for the rest) for cRNA, GC(A/G)AAAGCAGG for mRNA, most of which lacks 5' A as compared to cRNA, and AGUAGAAACAAGG for vRNA (Figure 3.5). Based on the reciprocal template/product relationship, the 3' UTRs of cRNA and vRNA contain CCUUGUUUCUACU and CCUGCUUU(U/C)GCU. As reported, the 5' and 3' UTRs of each vRNA and cRNA are basically inverted repeats, which are able to form imperfectly base-paired dsRNA within the same molecule (Figure 3.5) (Desselberger et al. 1980).

We aimed to investigate whether cap-snatching occurs at loci other than IAV mRNA +1. We modified our previous method to map the CapSeq reads to the full length of IAV cRNA and vRNA. And extracted non-IAV 5' portions as potential host-derived caps and 3' portions completely matching IAV RNA (Gu et al. 2015). Any IAV site could contain both non-capped RNA reads and capped RNA reads, defined as 'containing a host cap of at least 5-nt long' in this paper. We use the first matched nt of a read to represent the mapped site of the whole read and the part 5' of the matched nt to serve as a potential host cap. For example, if a 40-nt read is mapped to vRNA +2 to +41, the mapped position is defined as +2 and the read does not contain a host cap; if the last 30 nts of this read is mapped to vRNA +2 to +21, the mapped position is defined as +2 and the read contains a 10-nt host cap.

We observed a cluster of non-canonical capped RNA reads mapped ~300-nt upstream of each mRNA 3' end in addition to canonical capped RNA reads at mRNA +1 (Figure 3.2). These reads form obvious clusters on PA, PB1, PB2, and NA mRNAs; capped RNA reads are also mapped to similar regions in other mRNAs, forming either apparent clusters of low abundance (M and NS) or no apparent cluster (HA and NP) (Figure 3.2). Although we observed obvious 3' clusters in the 12-48-hr timepoints, the 6-hr sample exhibited tiny clusters only on PA, PB1 and PB2 likely due to the extremely low expression level of IAV RNA (0.24% of total host/IAV RNA) (Figure 3.2). Overall, the capped RNA reads in mRNA 3' clusters are ~0.3% of those mapped to IVA mRNA +1. However, this ratio varies dramatically among individual mRNAs with the highest for NA (8.52%) and lowest for M (0.04%) at the 24-hr timepoint (Appendix D-Figure S3.1A). We observed an even higher ratio (9.4%) for NA at the 48-hr timepoint. In addition to mRNA +1 and 3' clusters, other mRNA regions also contain many capped RNA reads of less abundance as well as some non-capped RNA reads (Figure 3.2). However, the non-5' regions of vRNA contain much less capped RNA reads but more non-capped RNA reads (Figure 3.2). This stark contrast suggests that the non-canonical capped RNA reads mapped to IAV mRNA were not generated by cloning artefacts, as discussed below.

The vRNA 5' region, defined as +1 to +10, contains a high rate of capped RNA reads (capped reads divided by all reads). Among the 9 million non-rRNA reads in the 24-hr sample, ~14.7% and ~70.4% were mapped to IAV and host mRNA respectively, and ~14.9% were mapped to IAV vRNA. The 12-hr sample contains relatively more IAV mRNA (7.2%) than vRNA (3.5%), both of which are much lower than those in the 24-hr sample. Among the reads mapped to the first 10-nt of vRNA, 97.7%, 1.3% and 1.0% were assigned to the +1, +2, and rest 8-nt (Figure 3.3). Overall, ~0.7% of the total reads in this region contain a host cap. However, the distribution of capped RNA reads does not follow that of the total reads as 46.6%, 46.3% and 7.1% were mapped to the +1, +2 and rest 8-nt respectively. As a consequence, the rate of capped RNA reads is much higher at +2 (26.4%) and the rest 8-nt (5.2%) than +1 (0.3%) (Figure 3.3). This capped RNA rate at vRNA +2 is closer to that (99%) at mRNA +1 or cRNA +2. Technically they are all the same positions encoding the same nucleotide, G, but on different RNA strands (Figure 3.5). The start nts of cRNA/mRNA and vRNA/ capped vRNA constitute a symmetry with ppp-AG for cRNA/vRNA and a host cap followed by G for mRNA/capped vRNA (Figure 3.5). We found that the 5' region of each vRNA strand all bears a high capped RNA rate in all the four timepoints including the 6-hr one (Figure 3.3). Reads mapped downstream of this region barely contain host caps, as discussed below (Table 3.1).

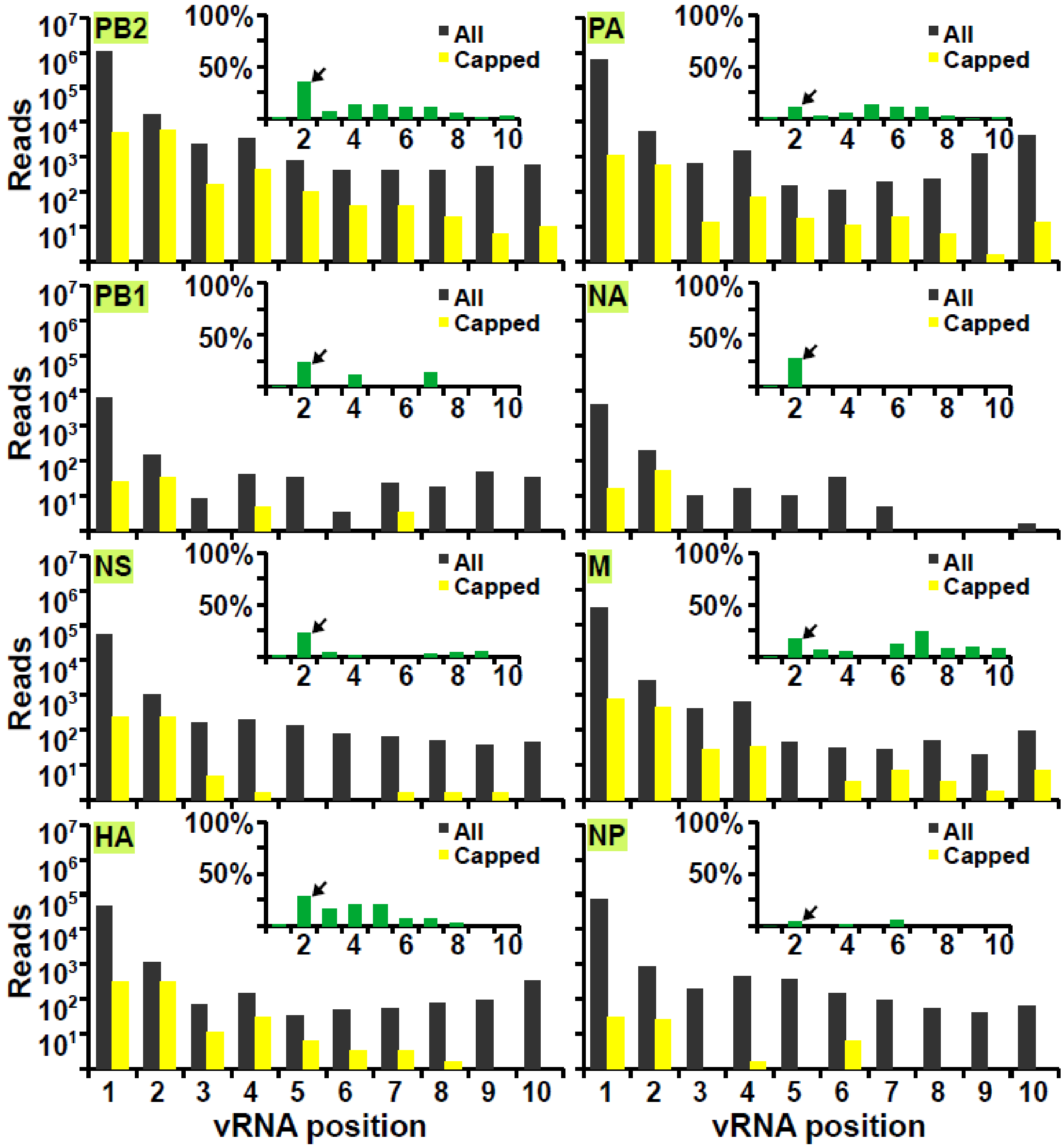


Figure 3. 3. The histogram of vRNA 5' regions.

The reads are normalized, mapped, combined and visualized using their start sites, as described in Figure 3.2, with 'yellow' indicating capped RNA reads and 'black indicating all reads at the position +1 (the first) to +10 of vRNA. The inlet figures represent the ratio of capped RNA reads to all reads at each position with arrows indicating vRNA +2.

U1 and U2 snRNAs are the top cap donors for non-canonical cap-snatching

We previously reported that U1 and U2 were the top host cap donors, contributing 3.3% and 3.5% of all caps at IAV mRNA +1 (Gu et al. 2015). Here we demonstrate that they are also the top donors for caps mapped to mRNA 3' clusters and vRNA 5' regions since U1 and U2 contributed 3.3% and 5.1% caps in mRNA 3' clusters, and 1.6% and 5.2% caps in vRNA 5' regions (Appendix D-Figure S3.1B). These rates are at least 10 folds as much as those of the top host mRNA donors. These canonical and non-canonical rates appear to be within the same range with variations likely caused by the low read number of capped RNA in non-canonical regions.

We observed a positive correlation between the levels of host cap donors and IAV caps at the 24-hr timepoint (p < 1 x 10-4 in Appendix D-Figure S3.1B). This weak correlation (r =0.26 instead of 1) may be caused by: 1) the low IAV capped RNA read number; and 2) the host cap number includes pre-RNA, mature RNA and PASR, not all of which are cap-snatching substrates. ncRNA appears to have a higher cap-snatching rate (IAV cap / [IAV cap+host cap]) than host mRNA/pre-mRNA/PASR (Appendix D-Figure S3.1B). For example, U1 and U2 have higher cap-snatching rates than most mRNAs (Appendix D-Figure S3.1B). We reached similar conclusions using the 12-hr timepoint (Appendix D-Figure S3.1C). All these observations are consistent with our previous conclusion that cap-snatching at mRNA +1 prefers ncRNA, especially U1 and U2 snRNAs, suggesting that IAV RdRP utilizes the same substrate pool for canonical and non-canonical cap-snatching (Gu et al. 2015).

Verifying non-canonical cap-snatching using U1 and U2 snRNAs

The non-IAV 5' portion of a read can be generated by cloning artefacts of low frequency rather than by cap-snatching. For example, host caps can be ligated to IVA RNA by RNA ligases, substituting for 5' ligation linkers. Such events are very rare since 1) the 5' linker amount used is

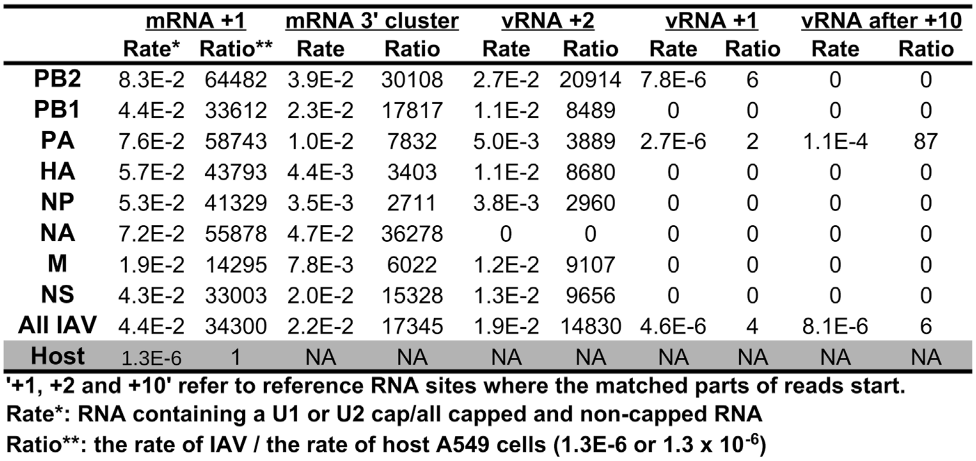
50 pmole while the total host capped RNA is only ~0.06 pmole, likely only a small fraction of which is degraded to generate 11-nt caps (the average size on IAV); and 2) degraded RNA usually contains 3' cyclic phosphate or 3' monophosphate, incompatible with RNA ligation. Non-IAV 5' portions can be generated by reverse transcriptase jumping along templates, generating non-continuous IAV sequences which were dissected as non-IAV and IAV parts by our algorithm, as in the defective interfering (DI) particles (Saira et al. 2013). This model is disfavored since 1) jumping events cannot de novo generate ~11-nt U1/U2 sequences out of IAV sequences; and 2) jumping events cannot generate cap-snatching signatures including the cap size, cleavage motif, priming motif and realignment feature, as discussed below. In addition, non-IAV 5' portions can be generated by reverse transcriptase via template-switching, in which a reverse transcriptase utilizes two templates, leading to a ligation-like behavior (Cocquet et al. 2006). Again, this mechanism occurs at a very low frequency and cannot account for cap-snatching signatures either.

To explicitly support our non-canonical cap-snatching model, we provided three negative internal controls. The host mRNA cloned in the same reaction serves as a 'perfect' negative control since it was treated exactly the same way as IAV capped RNA. Here we found that the rate of U1/U2 cap-containing host mRNA reads, defined as "U1/U2-containing reads divided by all capped and uncapped RNA reads", is extremely low (1.3 X 10-6) in the 24-hr sample. In contrast, IAV mRNA +1 and non-canonical sites (mRNA 3' clusters and vRNA 5' regions) all contain similar rates of U1/U2 caps, ~15,000-30,000 folds higher than that of host mRNA (Table 3.1). This general conclusion also applies to almost every individual RNA strand, including 8 mRNA and 7 vRNA strands (Table 3.1). The only exception is the NA vRNA 5' region, which lacks an enough read coverage for obtaining a U1/U2 cap rate. A second negative control is the non-5' region of vRNA, in which the U1/U2 cap rate is very close to that of host mRNA, i.e., the 'background rate' (Table 3.1). A third negative control is vRNA +1, which contains ~92% of all vRNA reads. Usually the

+1 or 5' end is the hot spot for template-switching since it is the last template nt, as we utilized this mechanism for cloning small RNA (Gu et al. 2009). However, we only observed a rate a little bit higher than the background (Table 3.1). In all, these negative controls serve as robust evidence supporting that non-canonical cap-snatching is not caused by cloning artefacts. Consistent with this, the ratio of U1/U2 caps to all caps in canonical and non-canonical cap-snatching regions are almost the same (Appendix D-Figure S3.1B), as discussed above. Moreover, both share the same cap-snatching features, as discussed below. Therefore, the canonical cap-snatching results serve as positive controls for the non-canonical ones.

We also used a non-ligation-based method, basically only RT/PCR, to confirm our result. We used random hexamers and oligo (dT)12-18 to generate IAV cDNA respectively and then amplified the cDNA using a shared reverse primer with different 10-nt forward primers 5'-attached with an 11-nt U2 5' sequence. Three positive forward primers were picked from the sites containing a U2 cap in the 3' region of PB2 mRNA and three negative control primers were randomly picked from the upstream sites containing no U2 cap (Appendix E-Figure S3.2). We had to use a second shared reverse primer in nested PCR reactions to achieve product specificity. As expected, we can easily detect the targets, as confirmed by Sanger sequencing, in the positive PCR reactions even at the low PCR cycle number while failing to detect any product in the negative controls even at the high PCR cycle number. Interestingly, we also obtained a truncated product for each positive reaction, all of which contains the same 33-nt deletion (Appendix E-Figure S3.2). This deletion does not affect our conclusion since it is ~60-70 nt downstream of the start sites of the capped RNA reads. Moreover, we observed the same positive results at the same PCR cycle number using the cDNA templates made with either random hexamers or oligo (dT)12-18, suggesting at least a significant fraction of the capped RNA in IAV mRNA 3' regions contains a poly(A) tail.

Table 3. 1. mRNA cap-snatching rates and ratios from different sites.



	mRNA +1		mRNA 3' cluster		vRNA +2		vRNA +1		vRNA after +10	
	Rate*	Ratio**	Rate	Ratio	Rate	Ratio	Rate	Ratio	Rate	Ratio
PB2	8.3E-2	64482	3.9E-2	30108	2.7E-2	20914	7.8E-6	6	0	0
PB1	4.4E-2	33612	2.3E-2	17817	1.1E-2	8489	0	0	0	0
PA	7.6E-2	58743	1.0E-2	7832	5.0E-3	3889	2.7E-6	2	1.1E-4	87
HA	5.7E-2	43793	4.4E-3	3403	1.1E-2	8680	0	0	0	0
NP	5.3E-2	41329	3.5E-3	2711	3.8E-3	2960	0	0	0	0
NA	7.2E-2	55878	4.7E-2	36278	0	0	0	0	0	0
M	1.9E-2	14295	7.8E-3	6022	1.2E-2	9107	0	0	0	0
NS	4.3E-2	33003	2.0E-2	15328	1.3E-2	9656	0	0	0	0
All IAV	4.4E-2	34300	2.2E-2	17345	1.9E-2	14830	4.6E-6	4	8.1E-6	6
Host	1.3E-6	1	NA	NA	NA	NA	NA	NA	NA	NA

'+1, +2 and +10' refer to reference RNA sites where the matched parts of reads start.
Rate*: RNA containing a U1 or U2 cap/all capped and non-capped RNA
Ratio: the rate of IAV / the rate of host A549 cells (1.3E-6 or 1.3 x 10^{-6})**

Realignment occurs much less frequently in mRNA 3' clusters and vRNA 5' regions

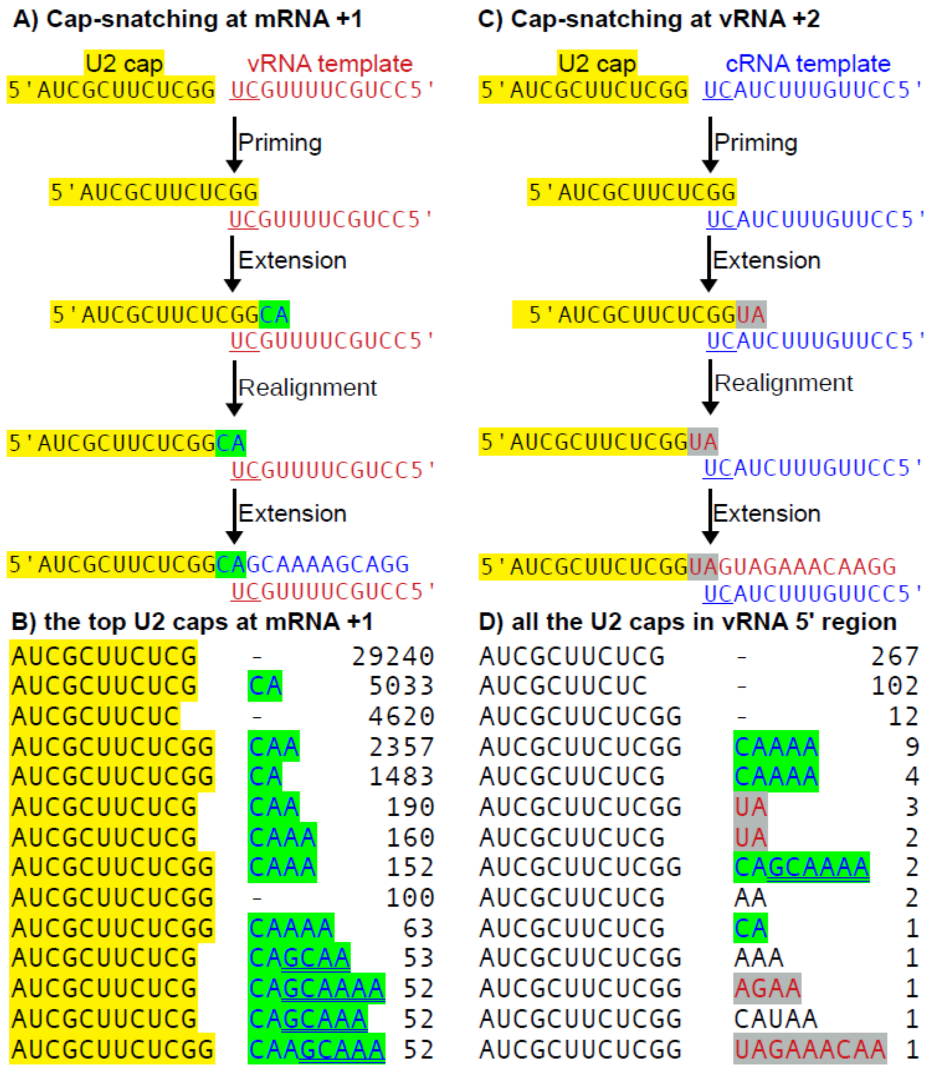
We analyzed the size distribution of IAV caps including the extra nts added via the realignment mechanism. Both mRNA 3' clusters and vRNA 5' regions exhibit an almost symmetric bell-like size distribution of IAV caps of 7-16 nts long peaking at 11-nt (Appendix F-Figure S3.3A). In contrast, mRNA +1 sites exhibit a 12-nt peak with a skewed size distribution, as the slope left of the peak is much steeper than that right of the peak. This asymmetric distribution was apparently caused by the extra nts added via the realignment mechanism since we obtained a symmetric distribution peaking at 11-nt after removing the most abundant species of these nts (Appendix F-Figure S3.3A). There were no dramatic size changes after removal of the realigned extra nts in mRNA 3' clusters and vRNA 5' regions (Appendix F-Figure S3.3A), suggesting that realignment occurs much less frequently in non-canonical cap-snatching.

To explicitly compare realignment rates, we analyzed U1 and U2 caps on IAV RNA. We only considered mRNA +1 and vRNA 5' regions, since all IAV mRNAs but PB1 and all cRNAs share the first 11 and 13 nts respectively (Figure 3.5), allowing us to easily figure out prime-realignment patterns. Moreover, these loci represent at least 90% cap-snatching events on the corresponding

RNAs. We found that 16% and 23% of U1 and U2 caps at IAV mRNA +1 contain extra nts, and at least 15% and 22% clearly contain recognizable realignment or re-realignment patterns, suggesting that almost all the extra nts were added via realignment. In contrast, ~7% each of U1 and U2 caps in IAV vRNA 5' regions contain extra nts, and only ~1% and ~5.6% possibly contain recognizable realignment patterns.

Identification of trans-realignment

In the canonical prime-cis-realignment model, prime and realignment steps are coupled on same RNA templates (Koppstein et al. 2015; Decroly et al. 2011; Te Velthuis and Oymans 2018; Geerts-Dimitriadou et al. 2011b). To examine if a realignment step can utilize a different RNA template, a process defined here as 'trans-realignment', we first analyzed the realignment patterns at mRNA +1 and extracted the most abundant 'extra nts' generated via realignment. As shown in Figure 3.4A, the first priming step utilizes the base-pairing of the host cap -1 G and template vRNA -2 C; the cap is usually extended for 2-4 nts, ending at 'A' (first extension); the extended sequence is realigned using the base-pairing of the cap -1 A and template vRNA -1 U (realignment or second priming); the RNA is extended again (second extension); and in rare cases, a third round of priming-extension may occur. Since all IAV mRNAs share almost identical 11-nt 5' UTRs and extra nts, usually 2-4 nts generated by realignment, are copied from 5' UTRs, technically prime-cis-realignment using one template twice and prime-trans-realignment using two templates each once generate the same results (Figure 3.4). Since the 5' UTRs of mRNA and vRNA bear different sequences (Figure 3.5), any trans-realignment between them can be identified using unique sequence information. Although ~5.6% of the U2 caps in vRNA 5' regions exhibit recognizable realignment patterns, only ~1.7% were generated by prime-cis-realignment via the G/C and A/U base-pairings (Figure 3.4C and D). The top realignment patterns bear the signature sequences (CA, CAAAA and CAGCAAAA) derived from caps at IAV mRNA +1 which are primed, extended and prematurely terminated (Figure 3.4B and D). We estimate that trans-realignment contributed ~4% of the U2 caps in the vRNA 5' regions.



B) the top U2 caps at mRNA +1

AUCGCUUCUCG	-	29240
AUCGCUUCUCG	CA	5033
AUCGCUUCUC	-	4620
AUCGCUUCUCGG	CAA	2357
AUCGCUUCUCGG	CA	1483
AUCGCUUCUCG	CAA	190
AUCGCUUCUCG	CAAA	160
AUCGCUUCUCGG	CAAA	152
AUCGCUUCUCGG	-	100
AUCGCUUCUCGG	CAAAA	63
AUCGCUUCUCG	CAGCAA	53
AUCGCUUCUCG	CAGCAAAA	52
AUCGCUUCUCG	CAGCAAA	52
AUCGCUUCUCGG	CAAGCAAA	52

D) all the U2 caps in vRNA 5' region

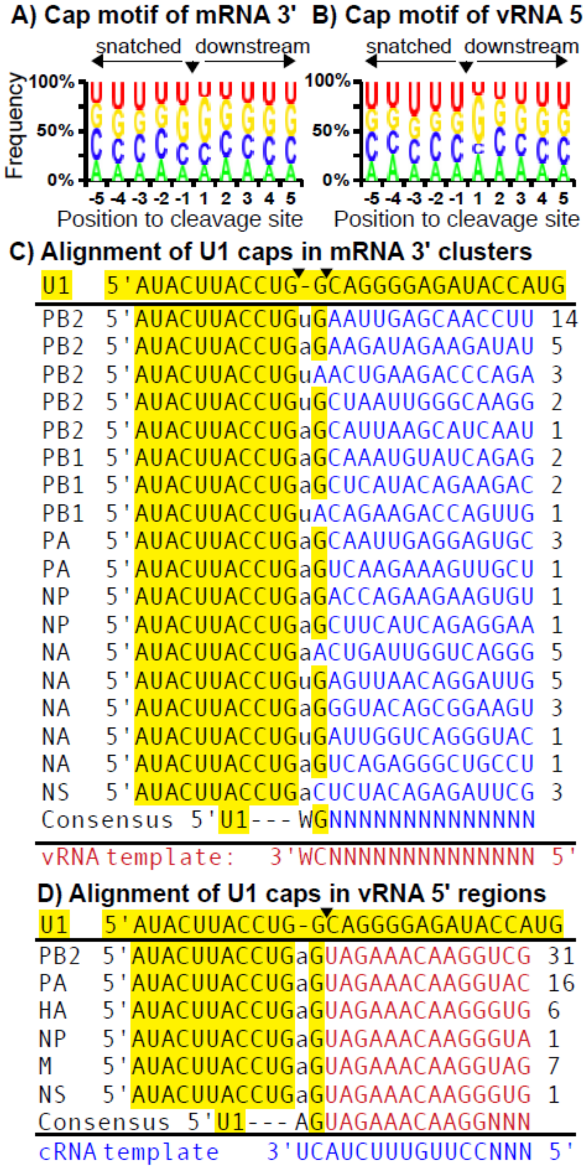
AUCGCUUCUCG	-	267
AUCGCUUCUC	-	102
AUCGCUUCUCGG	-	12
AUCGCUUCUCGG	CAAAA	9
AUCGCUUCUCG	CAAAA	4
AUCGCUUCUCGG	UA	3
AUCGCUUCUCG	UA	2
AUCGCUUCUCGG	CAGCAAAA	2
AUCGCUUCUCG	AA	2
AUCGCUUCUCG	CA	1
AUCGCUUCUCGG	AAA	1
AUCGCUUCUCGG	AGAA	1
AUCGCUUCUCGG	CAUAA	1
AUCGCUUCUCGG	UAGAAACAA	1

Figure 3. 4. The U2 cap realignment at mRNA +1 and vRNA +2.

A) & C) a U2 cap highlighted in yellow is annealed with a template vRNA (A) or cRNA (C) via the base-pairing between the cap -1 G and template -2 C in the initial priming step, extended (green in A or gray in C) but only prematurely ended with A, realigned with the template via the base-pairing between the extended cap -1 A and template -1 U, and then extended to a full size RNA (blue fonts for mRNA and red fonts for vRNA); B) the top 14 U2 caps containing at least AUCGCUUCUC at mRNA +1 with extra nts added via realignment (green), re-realignment (green double-underlined), no realignment (-); D) all the U2 caps containing at least AUCGCUUCUC in vRNA 5' regions with extra nts added via trans-realignment (green), cis-realignment (gray), unknown mechanisms (black fonts), and no realignment (-).

IAV utilizes a more general WG motif for priming capped RNA synthesis

Previous studies have found that cap-snatching at IAV mRNA +1 usually utilizes the base pairing between the -1 G of a host cap and the -2 C of a template vRNA to prime mRNA synthesis (Gu et al. 2015; Koppstein et al. 2015; Geerts-Dimitriadou et al. 2011b). Therefore, although the +1 G of IAV mRNA appears to be encoded, it is actually derived from a host cap via priming. To examine whether a specific base paring plays a similar role in synthesizing capped RNA in vRNA 5' regions and mRNA 3' clusters, we first divided hybrid RNA reads into two parts, IAV RNA sequences and host caps. 56% and 66% of IAV-encoded parts in mRNA 3' clusters and vRNA 5' regions start with G (Appendix F-Figure S3.3B & C). We speculated that IAV may prefer cleavage sites 3' of G on host RNA, generating this +1 G preference on capped IAV RNA via priming. To examine this hypothesis, we mapped the snatched host caps to human genome and analyzed the nt preference surrounding the last (-1) nt of these caps. There is an obvious preference for G immediately after the -1 nt of host caps used in cap-snatching in mRNA 3' clusters and vRNA 5' regions (Figure 3.5A and B). A reasonable explanation for this preference is that IAV RdRP prefers to cleave host caps 3' of G and utilizes this G to base- pair with a template C, priming RNA synthesis. Since we assigned this priming G to the IAV RNA parts, the host caps losing this G appeared to be cleaved 5' of G when mapped to human genome (Figure 3.5A and B).


A) Cap motif of mRNA 3'
B) Cap motif of vRNA 5'
snatched
downstream
Frequency
100%
50%
0%
-5 -4 -3 -2 -1 1 2 3 4 5
Position to cleavage site
C) Alignment of U1 caps in mRNA 3' clusters
D) Alignment of U1 caps in vRNA 5' regions

```
C) Alignment of U1 caps in mRNA 3' clusters
U1     5'AUACUUACCUG-GCAGGGGAGAUACCAUG
PB2    5'AUACUUACCUGuGAAUUGAGCAACCUU 14
PB2    5'AUACUUACCUGaGAAGAUAGAAGAUAU 5
PB2    5'AUACUUACCUGuAACUGAAGACCCAGA 3
PB2    5'AUACUUACCUGuGCUAAUUGGGCAAGG 2
PB2    5'AUACUUACCUGaGCAUUAAGCAUCAAU 1
PB1    5'AUACUUACCUGaGCAAAUGUAUCAGAG 2
PB1    5'AUACUUACCUGaGCUCAUACAGAAGAC 2
PB1    5'AUACUUACCUGuACAGAAGACCAGUUG 1
PA     5'AUACUUACCUGaGCAAUUGAGGAGUGC 3
PA     5'AUACUUACCUGaGUCAAGAAAGUUGCU 1
NP     5'AUACUUACCUGaGACCAGAAGAAGUGU 1
NP     5'AUACUUACCUGaGCUUCAUCAGAGGAA 1
NA     5'AUACUUACCUGaACUGAUUGGUCAGGG 5
NA     5'AUACUUACCUGuGAGUUAACAGGAUUG 5
NA     5'AUACUUACCUGaGGGUACAGCGGAAGU 3
NA     5'AUACUUACCUGuGAUUGGUCAGGGUAC 1
NA     5'AUACUUACCUGaGUCAGAGGGCUGCCU 1
NS     5'AUACUUACCUGaCUCUACAGAGAUUCG 3
Consensus 5'U1---WGNNNNNNNNNNNNNN
vRNA template:  3'WCNNNNNNNNNNNNNN 5'

D) Alignment of U1 caps in vRNA 5' regions
U1     5'AUACUUACCUG-GCAGGGGAGAUACCAUG
PB2    5'AUACUUACCUGaGUAGAAACAAGGUCG 31
PA     5'AUACUUACCUGaGUAGAAACAAGGUAC 16
HA     5'AUACUUACCUGaGUAGAAACAAGGGUG 6
NP     5'AUACUUACCUGaGUAGAAACAAGGGUA 1
M      5'AUACUUACCUGaGUAGAAACAAGGUAG 7
NS     5'AUACUUACCUGaGUAGAAACAAGGGUG 1
Consensus 5'U1---AGUAGAAACAAGGNNN
cRNA template   3'UCAUCUUUGUUCCNNN 5'
```

Figure 3. 5. The cleavage/priming motif of non-canonical cap-snatching.

A and B) IAV capped RNA reads derived from mRNA 3' clusters and vRNA 5' regions were mapped to human genome and the nt frequency of each position (x axis) surrounding the last nt (-1, the reference point) of the host caps was displayed (y axis) with '▼' indicating the 'apparent' cut site; C and D) all the capped RNA reads containing the U1 sequence AUACUUACCUG were mapped/aligned to each IAV mRNA 3' cluster and vRNA 5' region with 'yellow' indicating the U1 derived sequence in U1 snRNA (top) and each hybrid capped RNA read, 'blue' and 'red' indicating IAV mRNA/cRNA and vRNA sequences respectively, 'lowercase' indicating virtual nts not transcribed but converted from template nts, 'W' representing A/U, the last column representing the read number, and the arrows indicating the cap cleavage sites.

We then examined which C's on the template RNA were selected for the cap-mediated priming. For synthesizing capped RNA in mRNA 3' clusters, multiple C's on the template vRNA were used (Figure 3.5C and Appendix G-Figure S3.4A). For synthesizing capped RNA in vRNA 5' regions, cRNA templates have two C's, -2 and -4, and almost all priming events occur at -2. For example, 100% of U1 caps and 99% of U2 caps utilized the template -2 C for priming and only 0.7% of U2 caps utilized the template -4 C (Figure 3.5D and Appendix G-Figure S3.4B). This template -2 C preference also occurred in canonical cap-snatching at mRNA +1 since mRNA synthesis was predominantly primed using the cap -1 G and almost exclusively the template -2 C instead of the -4 and -9 C's (Gu et al. 2015; Koppstein et al. 2015; Geerts-Dimitriadou et al. 2011b, 2011a).

The priming events during IAV capped RNA synthesis prefer U or A (collectively as W) followed by G or WG. In IAV mRNA 5' regions, at least 95% of cap-snatching utilized the -2 C of template vRNA for initial priming events (Gu et al. 2015), generating the start nt G preceded by a virtual nt A not transcribed but converted from the -1 U of template vRNA. In other words, the last two nts of IAV vRNA, CU, serve as the template for making AG in at least 95% of IAV mRNA, in which G is the first nt and A is the virtual nt upstream. In vRNA 5' regions, ~66% of capped RNAs start with G, and ~92% of the virtual nts upstream of the start nts, which are converted from template cRNA, are A's (Appendix F-Figure S3.3C). Based on these two observations, we concluded that the initial priming events prefer an AG motif, in which G is the first nt and A is the virtual nt. Because the 5' ends of vRNA and mRNA are similar, this AG motif is deduced from a very limited sequence diversity and thus may not represent an authentic motif. The mRNA 3' clusters contain sufficient sequence diversity, allowing us to obtain a more general rule. We found that the priming events in mRNA 3' cluster also prefer a (A/U)G or WG (W representing A/U) motif since 56% of capped RNA starts with G, and 56% and 25% of the virtual nts upstream of the start nts are A and U respectively (Appendix F-Figure S3.3B). For example, the U1/U2 caps mapped to IAV mRNA 3' clusters clearly utilize this WG motif (Figure 3.5C and Appendix G-Figure S3.4A). Since the WG motif is based on multiple loci and includes the AG motif, it represents a more general rule for cap-mediated initial priming events.

Cap-snatching likely generates mRNA encoding truncated or new IAV proteins

We speculate that at least a fraction of the non-canonical capped IAV RNA in mRNA 3' clusters bears all the required structure elements of translation-capable mRNA. First, the cap is stolen from host Pol II transcripts, most of which are used for translation; second, the capped RNA in mRNA 3' clusters likely contains a poly(A) tail (Appendix E-Figure S3.2); third, the cap or internal AUG of annotated IAV mRNA may provide a start AUG; and fourth, since the consensus Kozak sequence (A/G)NNAUGG is very short, 12.5% of any given AUG-containing sequence bears a Kozak motif (Kozak 1987; Cavener 1987; Hamilton et al. 1987; Kozak 1986). Considering all these structure requirements, we developed a custom PERL script to predict potential mRNA in mRNA 3' clusters. We found several capped RNAs potentially encoding 17 proteins, all of which belong to the C-terminal parts of annotated IAV proteins and are composed of at least 50 amino acids (Appendix H-Supplementary File 3). The 5' UTR size is usually less than 50-nt and each protein may be coded by several capped RNAs with 5' UTRs of various sizes (Appendix H-Supplementary File 3, column 3). Among the 17 proteins, the host caps provide the start AUG for 10 proteins (Appendix H-Supplementary File 3, 'hybrid' in column 2), generating 1-3 amino acids, while IAV RNA provides the start AUG for the rest of the proteins. Since mRNA 3' clusters are very close to the 3' end of mRNA, only 3 proteins contain more than 100 amino acids.

Several larger proteins can be generated by other non-canonical capped RNAs. We observed capped RNAs upstream of mRNA 3' clusters but well downstream of annotated IAV mRNAs (Figure 3.2). These RNAs are likely authentic capped RNAs generated by cap-snatching since: 1) the cap size is ~11-nt; 2) most of them use a G/C base paring-mediated priming; 3) most host caps start with A, a signature start nt of host caps (Gu et al. 2015); and 4) the same IAV locus can have multiple caps derived from different host RNAs (Figure 3.6). Using the same method above, we identified several bigger proteins (data not shown). For example, one internal AUG of PA can be used by several capped RNAs to encode a protein composed of 343 amino acids (Figure 3.6). Capped RNA mapped to vRNA 5' regions likely belongs to ncRNA. Although its size could be as long as that of vRNA, we can only identify a few short coding frames, all of which have long 5' and 3' UTRs (data not shown). Moreover, we don't know whether the capped RNA contain a poly(A) tail, and if so, how the tail is added. Therefore, we propose that capped RNA in vRNA 5' regions may serve as ncRNA instead of mRNA.

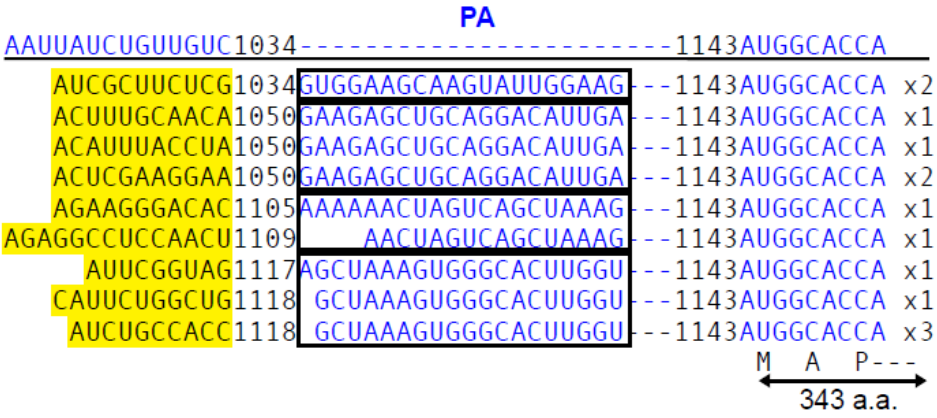

PA
AAUUAUCUGUUGUC1034----------------------1143AUGGCACCA
AUCGCUUCUCG1034GUGGAAGCAAGUAUUGGAAG--1143AUGGCACCA x2
ACUUUGCAACA1050GAAGAGCUGCAGGACAUUGA--1143AUGGCACCA x1
ACAUUUACCUA1050GAAGAGCUGCAGGACAUUGA--1143AUGGCACCA x1
ACUCGAAGGAA1050GAAGAGCUGCAGGACAUUGA--1143AUGGCACCA x2
AGAAGGGACAC1105AAAAAACUAGUCAGCUAAAG--1143AUGGCACCA x1
AGAGGCCUCCAACU1109AACUAGUCAGCUAAAG--1143AUGGCACCA x1
AUUCGGUAG1117AGCUAAAGUGGGCACUUGGU--1143AUGGCACCA x1
CAUUCUGGCUG1118GCUAAAGUGGGCACUUGGU--1143AUGGCACCA x1
AUCUGCCACC1118GCUAAAGUGGGCACUUGGU--1143AUGGCACCA x3
M A P---
343 a.a.

Figure 3. 6. A coding frame is shared by multiple capped RNAs.

Multiple capped RNA reads mapped to the internal positions of IAV PA mRNA were aligned with 'yellow', 'boxed blue' and 'non-boxed blue' indicating the host caps, 5'UTRs and shared coding frame respectively. The start positions of the 5' UTRs and coding frames are labeled in the sequences, the read number is labeled on the right, and the encoded amino acids are labeled at the bottom.

DISCUSSION

We provide multifaceted evidence to demonstrate that IAV utilizes cap-snatching to generate capped RNA in non-canonical regions. We found that both canonical and non-canonical cap-snatching share similar mechanisms. However, non-canonical cap-snatching bears unique features including infrequent realignment events and a more general priming motif, WG. (Figure 3.7). We also found that cap-snatching promotes the diversity of IAV mRNA and ncRNA.

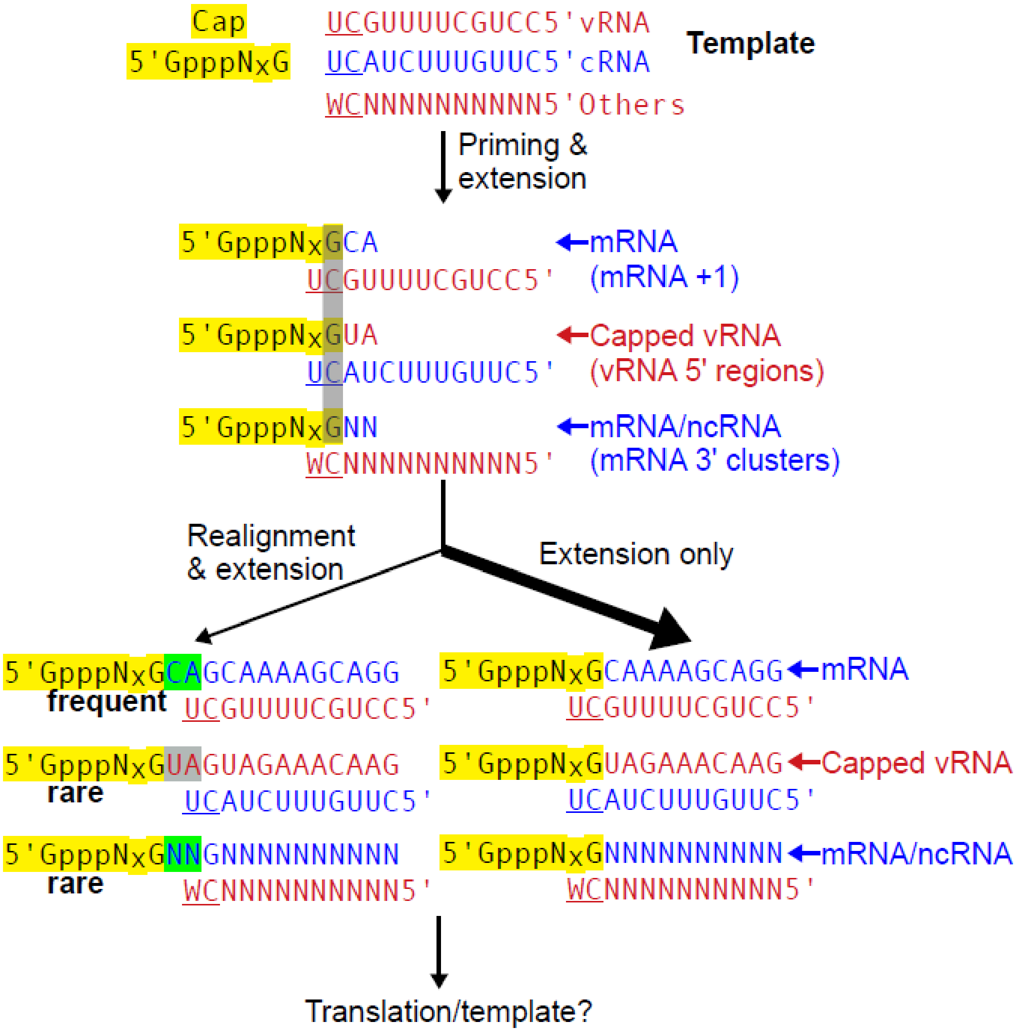


Figure 3. 7. A unified model for canonical and non-canonical cap-snatching.

Host caps are annealed to IAV RNA templates using the base-pairing between the cap -1 G and template -2 C in the initial priming step; the majority of host caps are extended to make full-size IAV RNA; a small fraction of host caps are extended for a few nts usually ending with 'A', realigned via the base-pairing between the -1 A of the extended sequences and the template -1 U, and extended again to generate full-size mRNA and ncRNA. W represents 'A' or 'U'.

Non-canonical cap-snatching is authentic and not caused by cloning artefacts. We demonstrated that non-canonical and canonical cap-snatching share several 'signature' features including: 1) the median size of snatched host caps is ~ 11-nt (Appendix F-Figure S3.3A); 2) U1 and U2 are the top cap donors totally contributing 7-8% of host caps snatched (Appendix D-Figure S3.1B); 3) a G/C base pair and WG motif are preferred in the initial priming step (Figure 3.4-3.5 & S3.4-5); 4) the cap-mediated priming prefers the -2 C of template RNA in both mRNA +1 and vRNA 5' regions (Figure 3.3-3.5 & S3.4-5). In addition, we provided three negative internal controls, including vRNA +1, non-5' regions of vRNA and host mRNA, to demonstrate that cloning artefacts barely generate any reads containing host caps in these regions (Table 3.1). Were caps added via cloning artefacts, we would expect: 1) no G/C base pair preference and no WG motif; 2) a random size distribution of caps snatched; 3) cap-snatching-mediated cap additions to host mRNA and IAV vRNA since they represent more than 80% of the cloned reads; and 4) no preference for template -2C. In conclusion, our data clearly supports that these non-canonical cap-snatching events are authentic and not caused by cloning artefacts.

We identified a more general cap-snatching/priming motif, WG. IAV RdRP utilizes cap-independent and dependent manners to synthesize RNA (Desselberger et al. 1980; Te Velthuis and Oymans 2018). Unlike DNA polymerases, most RNA polymerases do not require primers for initiating RNA synthesis. Although IAV RdRP appears to utilize a host cap to prime capped RNA synthesis, the single nt base-pairing is technically equal to de novo synthesizing RNA using a modified G in a primer-independent manner. In the cap-independent mode, IAV RdRP synthesizes vRNA using the -1 U of template cRNA and vice versa; in the cap-dependent mode, IAV RdRP primarily utilizes the -2 C of template vRNA to synthesize mRNA. Here we demonstrate that IAV RdRP also utilizes the -2 C of template cRNA to synthesize capped RNA in vRNA 5' regions, basically establishing a symmetry in synthesizing capped/uncapped RNA using vRNA and cRNA templates. However, this symmetry is imperfect since vRNA templates are predominantly used for synthesizing capped mRNA while cRNA templates are predominantly for synthesizing non-capped vRNA. In both cases, the level of the minor RNA species is only ~1-2% of that of the major species. In summary, the template -2 C is preferentially used for synthesizing capped RNA in both cases and the template -1 U is preferentially used for synthesizing non-capped RNA. Moreover, we demonstrate that IAV RdRP also preferentially utilizes the template -2 C, defined as a relative position, to synthesize capped RNA in mRNA 3' clusters.

We demonstrate that the cleavage site preference on host caps likely play a critical role in preferentially selecting template C sites, since: 1) previous studies have demonstrated that cap-snatching preferentially cuts host caps 3' of G; 2) our data shows the cleavage preference, 50%, for 3' of G on host caps (Figure 3.5A & B), is very close to ~60%, the start nt preference for G in mRNA 3' clusters and vRNA 5' regions (Appendix F-Figure S3.3B & C).

The nts 3' of template C's affect the selection of C sites for G/C base-pair-mediated priming events. Based on canonical and non-canonical cap-snatchings, we obtained a general motif CW on template RNA, corresponding to a WG motif on capped RNA, in which 'W' is encoded but not expressed. This WG motif is not caused by realignment since realignment is infrequent (less than 5%) in mRNA 3' clusters while ~50% of priming events utilize this motif. We speculate that this motif may help develop a novel therapeutic strategy.

Realignment occurs much less frequently in non-canonical cap-snatching. Realignment or re-realignment events constitute ~20% cap-snatching with recognizable sequence patterns at IAV mRNA +1. In contrast, realignment events with recognizable sequence patterns only constitute ~1.5% of cap-snatching in non-canonical regions. This suggests that IAV RdRP may use different modes to synthesize capped RNA at different loci. Or this discrepancy is caused by template sequence differences, i.e., 5'CUUUUGCU for mRNA +1, 5'UUUCUACU for vRNA 5' regions, and various sequences for mRNA 3' clusters. A sequence swap assay may help address this hypothesis.

Trans-realignment may utilize IAV-derived caps. It has been hypothesized that IAV cap-snatching does not target IAV mRNA as cap donors (Shih and Krug 1996). However, we clearly show that some caps utilized in vRNA 5' regions are likely derived from IAV mRNA based on the specific realignment patterns. Here we propose two models. In model 1, cap-snatching targets both host and IAV capped RNAs as cap donors. As a simple and straightforward model, it has serious caveats since 1) IAV mRNA is exported to cytoplasm for translation, generating a physical barrier for cap-snatching; and 2) evidence showed that IVA RdRP does not target its own mRNA (Shih and Krug 1996). In model 2, the realignment process may fail to prime with the same template RNA, jumping onto a second template, a process called 'trans-realignment' here. We prefer this model because it is well known that RNA/DNA polymerases often stall on promoters, generating PASR (Seila et al. 2008; Affymetrix/Cold Spring Harbor Laboratory ENCODE Transcriptome Project 2009; Nechaev et al. 2010; Gu et al. 2012). IAV RdRP may have the same feature, generating partially extended caps and then falling off templates. These caps primarily anneal back to same template RNA using the -1 U simply because 1) the initial priming/extension usually ends with A due to four template U's within the -2 to -7 positions on template vRNA; 2) the selection of same template RNA is due to physical proximity. These caps may be used to prime with another template at a lower frequency, as we observed. We detected this phenomenon simply because the two templates, cRNA and vRNA, bear different 5' sequences. If trans-realignment were to occur between two cRNA or two vRNA templates, the result would appear as cis-realignment on the same template.

Non-canonical cap-snatching diversifies IAV mRNA and ncRNA. We demonstrate that in mRNA 3' clusters, cap-snatching generates capped RNA bearing all the features of functional mRNA including a cap, poly(A) tail, start AUG and Kozak sequence. Actually, we can identify more translation-capable capped RNAs if we include other mRNA regions or relax the Kozak motif requirement. In many cases, host caps provide a start AUG and a coding frame for 1-3 amino acids (Appendix H-Supplementary File 3). Although non-canonical cap-snatching only generates ~1% of capped RNAs mapped to IAV mRNA, the expression levels of some non-canonical mRNA may reach to the median level of host mRNA because IAV mRNA reads constitutes 18% of all reads mapped to host and IAV capped RNA. Interestingly, we found that on NA mRNA, the non-canonical capped RNA level reaches ~9% of the canonical mRNA level (Figure 3.2 and S3.2A). In addition, cap-snatching may promote the diversity of IAV mRNA and ncRNA via 1) introducing a new AUG and Kozak signal to the internal sequences of annotated mRNA; and 2) obtaining new coding frames, especially on vRNA, due to the high mutation rate of IAV RdRP.

In summary, we provide a comprehensive profile of IAV cap-snatching and a more general mode for the priming-realignment mechanism. We also propose that cap-snatching promotes the diversity of IAV mRNA and ncRNA. Insights from this study may help better understand the cap-snatching mechanism and design research and therapeutic tools.

REFERENCES

Affymetrix/Cold Spring Harbor Laboratory ENCODE Transcriptome Project. 2009. Post-transcriptional processing generates a diversity of 5'-modified long and short RNAs. Nature 457: 1028–32.

Beaton AR, Krug RM. 1981. Selected host cell capped RNA fragments prime influenza viral RNA transcription in vivo. *Nucleic Acids Res* **9**: 4423–36.

Bouloy M, Plotch SJ, Krug RM. 1978. Globin mRNAs are primers for the transcription of influenza viral RNA in vitro. *Proc Natl Acad Sci U S A* **75**: 4886–90.

Bouvier NM, Palese P. 2008. The biology of influenza viruses. *Vaccine* **26 Suppl 4**: D49-53.

Caton AJ, Robertson JS. 1980. Structure of the host-derived sequences present at the 5' ends of influenza virus mRNA. *Nucleic Acids Res* **8**: 2591–603.

Cavener DR. 1987. Comparison of the consensus sequence flanking translational start sites in Drosophila and vertebrates. *Nucleic Acids Res* **15**: 1353–1361.

Cocquet J, Chong A, Zhang G, Veitia RA. 2006. Reverse transcriptase template switching and false alternative transcripts. *Genomics* **88**: 127–131.

Datta K, Wolkerstorfer A, Szolar OHJ, Cusack S, Klumpp K. 2013. Characterization of PA-N terminal domain of Influenza A polymerase reveals sequence specific RNA cleavage. *Nucleic Acids Res* **41**: 8289–8299.

Decroly E, Ferron F, Lescar J, Canard B. 2011. Conventional and unconventional mechanisms for capping viral mRNA. *Nat Rev Microbiol* **10**: 51–65.

Desselberger U, Racaniello VR, Zazra JJ, Palese P. 1980. The 3' and 5'-terminal sequences of influenza A, B and C virus RNA segments are highly conserved and show partial inverted complementarity. *Gene* **8**: 315–328.

Dhar R, Chanock RM, Lai CJ. 1980. Nonviral oligonucleotides at the 5' terminus of cytoplasmic influenza viral mRNA deduced from cloned complete genomic sequences. *Cell* **21**: 495–500.

Dias A, Bouvier D, Crepin T, McCarthy AA, Hart DJ, Baudin F, Cusack S, Ruigrok RW. 2009. The cap-snatching endonuclease of influenza virus polymerase resides in the PA subunit. *Nature* **458**: 914–8.

Fodor E, Crow M, Mingay LJ, Deng T, Sharps J, Fechter P, Brownlee GG. 2002. A single amino acid mutation in the PA subunit of the influenza virus RNA polymerase inhibits endonucleolytic cleavage of capped RNAs. *J Virol* **76**: 8989–9001.

Geerts-Dimitriadou C, Goldbach R, Kormelink R. 2011a. Preferential use of RNA leader sequences during influenza A transcription initiation in vivo. *Virology* **409**: 27–32.

Geerts-Dimitriadou C, Zwart MP, Goldbach R, Kormelink R. 2011b. Base-pairing promotes leader selection to prime in vitro influenza genome transcription. *Virology* **409**: 17–26.

Gu W, Gallagher GR, Dai W, Liu P, Li R, Trombly MI, Gammon DB, Mello CC, Wang JP, Finberg RW. 2015. Influenza A virus preferentially snatches noncoding RNA caps. *RNA* **21**: 2067–2075.

Gu W, Lee HC, Chaves D, Youngman EM, Pazour GJ, Conte D, Mello CC. 2012. CapSeq and CIP-TAP identify Pol II start sites and reveal capped small RNAs as *C. elegans* piRNA precursors. *Cell* **151**: 1488–500.

Gu W, Shirayama M, Conte D, Vasale J, Batista PJ, Claycomb JM, Moresco JJ, Youngman EM, Keys J, Stoltz MJ, et al. 2009. Distinct argonaute-mediated 22G-RNA pathways direct genome surveillance in the *C. elegans* germline. *Mol Cell* **36**: 231–244.

Guilligay D, Tarendeau F, Resa-Infante P, Coloma R, Crepin T, Sehr P, Lewis J, Ruigrok RW, Ortin J, Hart DJ, et al. 2008. The structural basis for cap binding by influenza virus polymerase subunit PB2. *Nat Struct Mol Biol* **15**: 500–6.

Hagen M, Tiley L, Chung TD, Krystal M. 1995. The role of template-primer interactions in cleavage and initiation by the influenza virus polymerase. *J Gen Virol* **76 (Pt 3)**: 603–611.

Hainer SJ, Gu W, Carone BR, Landry BD, Rando OJ, Mello CC, Fazzio TG. 2015. Suppression of pervasive noncoding transcription in embryonic stem cells by esBAF. *Genes Dev* **29**: 362–378.

Hamilton R, Watanabe CK, de Boer HA. 1987. Compilation and comparison of the sequence context around the AUG startcodons in Saccharomyces cerevisiae mRNAs. *Nucleic Acids Res* **15**: 3581–3593.

Kobayashi M, Toyoda T, Ishihama A. 1996. Influenza virus PB1 protein is the minimal and essential subunit of RNA polymerase. *Arch Virol* **141**: 525–539.

Koppstein D, Ashour J, Bartel DP. 2015. Sequencing the cap-snatching repertoire of H1N1 influenza provides insight into the mechanism of viral transcription initiation. *Nucleic Acids Res*.

Kozak M. 1987. An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs. *Nucleic Acids Res* **15**: 8125–8148.

Kozak M. 1986. Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. *Cell* **44**: 283–292.

Krug RM, Broni BA, Bouloy M. 1979. Are the 5' ends of influenza viral mRNAs synthesized in vivo donated by host mRNAs? *Cell* **18**: 329–34.

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**: R25.

Luo GX, Luytjes W, Enami M, Palese P. 1991. The polyadenylation signal of influenza virus RNA involves a stretch of uridines followed by the RNA duplex of the panhandle structure. *J Virol* **65**: 2861–2867.

Nechaev S, Fargo DC, dos Santos G, Liu L, Gao Y, Adelman K. 2010. Global analysis of short RNAs reveals widespread promoter-proximal stalling and arrest of Pol II in Drosophila. *Science* **327**: 335–8.

Pflug A, Lukarska M, Resa-Infante P, Reich S, Cusack S. 2017. Structural insights into RNA synthesis by the influenza virus transcription-replication machine. *Virus Res* **234**: 103–117.

Plotch SJ, Bouloy M, Krug RM. 1979. Transfer of 5’-terminal cap of globin mRNA to influenza viral complementary RNA during transcription in vitro. *Proc Natl Acad Sci USA* **76**: 1618– 1622.

Plotch SJ, Bouloy M, Ulmanen I, Krug RM. 1981. A unique cap(m7GpppXm)-dependent influenza virion endonuclease cleaves capped RNAs to generate the primers that initiate viral RNA transcription. *Cell* **23**: 847–58.

Poon LL, Pritlove DC, Fodor E, Brownlee GG. 1999. Direct evidence that the poly(A) tail of influenza A virus mRNA is synthesized by reiterative copying of a U track in the virion RNA template. *J Virol* **73**: 3473–3476.

Pritlove DC, Poon LLM, Fodor E, Sharps J, Brownlee GG. 1998. Polyadenylation of Influenza Virus mRNA Transcribed In Vitro from Model Virion RNA Templates: Requirement for 5′ Conserved Sequences. *J Virol* **72**: 1280–1286.

Pruitt KD, Brown GR, Hiatt SM, Thibaud-Nissen F, Astashyn A, Ermolaeva O, Farrell CM, Hart J, Landrum MJ, McGarvey KM, et al. 2014. RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res* **42**: D756-63.

Rao P, Yuan W, Krug RM. 2003. Crucial role of CA cleavage sites in the cap-snatching mechanism for initiating viral mRNA synthesis. *EMBO J* **22**: 1188–98.

Reich S, Guilligay D, Pflug A, Malet H, Berger I, Crépin T, Hart D, Lunardi T, Nanao M, Ruigrok RWH, et al. 2014. Structural insight into cap-snatching and RNA synthesis by influenza polymerase. *Nature* **516**: 361–366.

Saira K, Lin X, DePasse JV, Halpin R, Twaddle A, Stockwell T, Angus B, Cozzi-Lepri A, Delfino M, Dugan V, et al. 2013. Sequence analysis of in vivo defective interfering-like RNA of influenza A H1N1 pandemic virus. *J Virol* **87**: 8064–8074.

Seila AC, Calabrese JM, Levine SS, Yeo GW, Rahl PB, Flynn RA, Young RA, Sharp PA. 2008. Divergent transcription from active promoters. *Science* **322**: 1849–51.

Shaw MW, Lamb RA. 1984. A specific sub-set of host-cell mRNAs prime influenza virus mRNA synthesis. *Virus Res* **1**: 455–67.

Shi L, Summers DF, Peng Q, Galarz JM. 1995. Influenza A virus RNA polymerase subunit PB2 is the endonuclease which cleaves host cell mRNA and functions only as the trimeric enzyme. *Virology* **208**: 38–47.

Shih SR, Krug RM. 1996. Surprising function of the three influenza viral polymerase proteins: selective protection of viral mRNAs against the cap-snatching reaction catalyzed by the same polymerase proteins. *Virology* **226**: 430–435.

Sikora D, Rocheleau L, Brown EG, Pelchat M. 2014. Deep sequencing reveals the eight facets of the influenza A/HongKong/1/1968 (H3N2) virus cap-snatching process. *Sci Rep* **4**: 6181.

Sikora D, Rocheleau L, Brown EG, Pelchat M. 2017. Influenza A virus cap-snatches host RNAs based on their abundance early after infection. *Virology* **509**: 167–177.

Stein LD, Mungall C, Shu S, Caudy M, Mangone M, Day A, Nickerson E, Stajich JE, Harris TW, Arva A, et al. 2002. The generic genome browser: a building block for a model organism system database. *Genome Res* **12**: 1599–610.

Sugiyama K, Obayashi E, Kawaguchi A, Suzuki Y, Tame JR, Nagata K, Park SY. 2009. Structural insight into the essential PB1-PB2 subunit contact of the influenza virus RNA polymerase. *EMBO J* **28**: 1803–11.

Te Velthuis AJW, Oymans J. 2018. Initiation, Elongation, and Realignment during Influenza Virus mRNA Synthesis. *J Virol* **92**.

Yuan P, Bartlam M, Lou Z, Chen S, Zhou J, He X, Lv Z, Ge R, Li X, Deng T, et al. 2009. Crystal structure of an avian influenza polymerase PA(N) reveals an endonuclease active site. *Nature* **458**: 909–913.

Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, Billis K, Cummins C, Gall A, Girón CG, et al. 2018. Ensembl 2018. *Nucleic Acids Res* **46**: D754–D761.

Zheng H, Lee HA, Palese P, García-Sastre A. 1999. Influenza A virus RNA polymerase has the ability to stutter at the polyadenylation site of a viral RNA template during RNA replication. *J Virol* **73**: 5240–5243.

APPENDICES

Appendix A

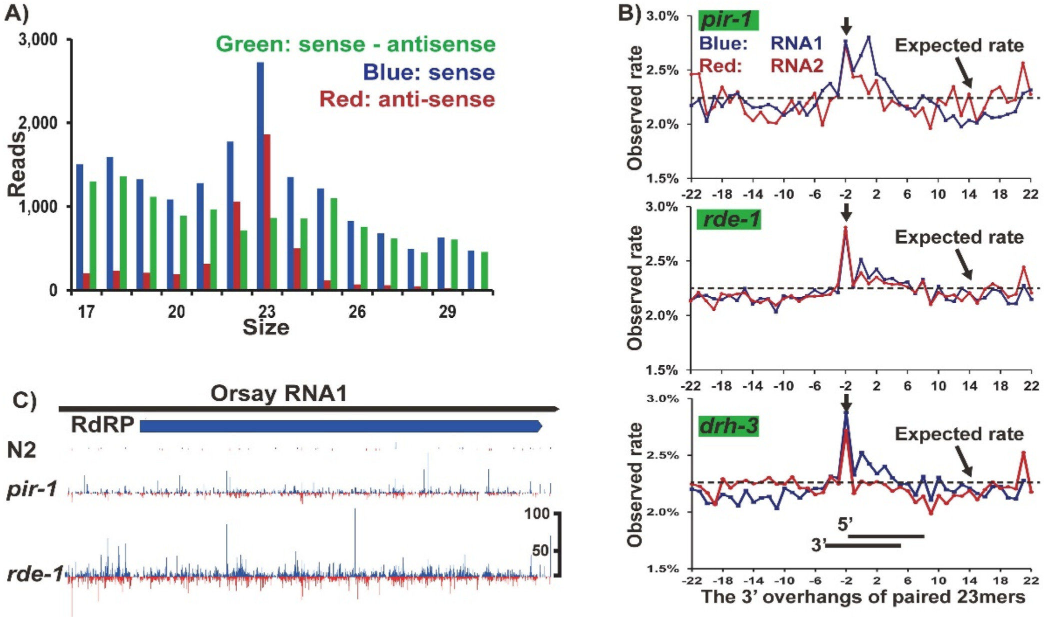


Figure S1. 1. PIR-1 is required for suppressing the Orsay virus. Related to Figure 1.4.

A) The size distribution of sRNAs cloned without the TAP treatment (only cloning p-RNAs) and mapped to sense and anti RNA1 in the *pir-1* mutant. The difference of sense and anti reads at each size is indicated as 'green: sense - anti'. Total miRNA reads as the normalization standard. B) The overhang sizes (3' end position of one 23mer – the 5' end position of the other 23mer in the paired 23mers. A paired 23mer is defined as one 23mer on the sense strand overlapping with another 23er on the anti strand. The overhang size is calculated for each pair and the total pairs of each overhang size are obtained. The rate on Y represents the ratio of the 23mer pairs with a specific overhang size over all the pairs with overhang sizes ranging from -22 to 22. The dotted line is the expected rate for each overhang size. RNA1 and RNA2 are calculated independently. C) The 23mer distribution along RNA1. Top is the RNA1 genome and the RdRP gene encoded; bottom, 23mer distribution with 'blue' representing sense reads and 'red' representing anti reads. Each bar represents the start position of a 23mer with height representing the read number using the same scale for all the samples, as indicated.

Appendix B

Supplementary File 1. Linkers and Primers

Table S1: Linkers and Primers

3' linker:	AppAGATCGGAAGAGCACACGTCTGAACTCCAGTCA/dideoxyC/
5' linker:	ACACUCUUUCCCUACACGACGCUCUUCCGAUCU
RT primer:	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
5' PCR primer:	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGA
3' PCR primer:	CAAGCAGAAGACGGCATACGAGAT ATCACGCA GTGACTGGAGTTCAGACGTGT
	CAAGCAGAAGACGGCATACGAGAT CGATGTCT GTGACTGGAGTTCAGACGTGT
	CAAGCAGAAGACGGCATACGAGAT TTAGGCGT GTGACTGGAGTTCAGACGTGT
	CAAGCAGAAGACGGCATACGAGAT TGACCACC GTGACTGGAGTTCAGACGTGT
	CAAGCAGAAGACGGCATACGAGAT ACAGTGCG GTGACTGGAGTTCAGACGTGT
	CAAGCAGAAGACGGCATACGAGAT GCCAATTT GTGACTGGAGTTCAGACGTGT
	CAAGCAGAAGACGGCATACGAGAT CAGATCGG GTGACTGGAGTTCAGACGTGT
	CAAGCAGAAGACGGCATACGAGAT ACTTGATG GTGACTGGAGTTCAGACGTGT
	CAAGCAGAAGACGGCATACGAGAT GATCAGTT GTGACTGGAGTTCAGACGTGT
	CAAGCAGAAGACGGCATACGAGAT TAGCTTTT GTGACTGGAGTTCAGACGTGT
	CAAGCAGAAGACGGCATACGAGAT GGCTACGG GTGACTGGAGTTCAGACGTGT
	CAAGCAGAAGACGGCATACGAGAT CTTGTAAT GTGACTGGAGTTCAGACGTGT
	CAAGCAGAAGACGGCATACGAGAT AGTCAAGT GTGACTGGAGTTCAGACGTGT
	CAAGCAGAAGACGGCATACGAGAT AGTTCCAC GTGACTGGAGTTCAGACGTGT
	CAAGCAGAAGACGGCATACGAGAT CCGTCCCA GTGACTGGAGTTCAGACGTGT
	CAAGCAGAAGACGGCATACGAGAT GTAGAGCA GTGACTGGAGTTCAGACGTGT
	CAAGCAGAAGACGGCATACGAGAT GTCCGCTC GTGACTGGAGTTCAGACGTGT
	CAAGCAGAAGACGGCATACGAGAT GTGAAACT GTGACTGGAGTTCAGACGTGT
	CAAGCAGAAGACGGCATACGAGAT GTGGCCGG GTGACTGGAGTTCAGACGTGT
	CAAGCAGAAGACGGCATACGAGAT GTTTCGCC GTGACTGGAGTTCAGACGTGT
	CAAGCAGAAGACGGCATACGAGAT CGTACGTA GTGACTGGAGTTCAGACGTGT
	CAAGCAGAAGACGGCATACGAGAT GAGTGGCG GTGACTGGAGTTCAGACGTGT
	CAAGCAGAAGACGGCATACGAGAT GGTAGCTA GTGACTGGAGTTCAGACGTGT
	CAAGCAGAAGACGGCATACGAGAT ATGAGCGA GTGACTGGAGTTCAGACGTGT
	CAAGCAGAAGACGGCATACGAGAT CAAAAGGC GTGACTGGAGTTCAGACGTGT
	CAAGCAGAAGACGGCATACGAGAT CAACTACC GTGACTGGAGTTCAGACGTGT
	CAAGCAGAAGACGGCATACGAGAT CACCGGAT GTGACTGGAGTTCAGACGTGT
	CAAGCAGAAGACGGCATACGAGAT CACTCATA GTGACTGGAGTTCAGACGTGT
	CAAGCAGAAGACGGCATACGAGAT CATTTTCG GTGACTGGAGTTCAGACGTGT
	CAAGCAGAAGACGGCATACGAGAT CCAACAGC GTGACTGGAGTTCAGACGTGT
	CAAGCAGAAGACGGCATACGAGAT CGGAATTC GTGACTGGAGTTCAGACGTGT
	CAAGCAGAAGACGGCATACGAGAT CTATACTC GTGACTGGAGTTCAGACGTGT
	CAAGCAGAAGACGGCATACGAGAT CTCAGAGG GTGACTGGAGTTCAGACGTGT
	CAAGCAGAAGACGGCATACGAGAT GACGACGA GTGACTGGAGTTCAGACGTGT
	CAAGCAGAAGACGGCATACGAGAT TAATCGTA GTGACTGGAGTTCAGACGTGT
	CAAGCAGAAGACGGCATACGAGAT TACAGCTA GTGACTGGAGTTCAGACGTGT
	CAAGCAGAAGACGGCATACGAGAT TATAATAG GTGACTGGAGTTCAGACGTGT
	CAAGCAGAAGACGGCATACGAGAT TCATTCTG GTGACTGGAGTTCAGACGTGT
	CAAGCAGAAGACGGCATACGAGAT TCCCGAAC GTGACTGGAGTTCAGACGTGT
	CAAGCAGAAGACGGCATACGAGAT TCGAAGCC GTGACTGGAGTTCAGACGTGT
	CAAGCAGAAGACGGCATACGAGAT TCGGCATG GTGACTGGAGTTCAGACGTGT
	CAAGCAGAAGACGGCATACGAGAT CTTCGGCC GTGACTGGAGTTCAGACGTGT
	CAAGCAGAAGACGGCATACGAGAT TTGTTACA GTGACTGGAGTTCAGACGTGT
	CAAGCAGAAGACGGCATACGAGAT AATCCAAT GTGACTGGAGTTCAGACGTGT

	CAAGCAGAAGACGGCATACGAGAT TCCTTTAA GTGACTGGAGTTCAGACGTGT
	CAAGCAGAAGACGGCATACGAGAT GTACAATC GTGACTGGAGTTCAGACGTGT
	CAAGCAGAAGACGGCATACGAGAT CCTGACTC GTGACTGGAGTTCAGACGTGT
	CAAGCAGAAGACGGCATACGAGAT TAGCACAC GTGACTGGAGTTCAGACGTGT
	CAAGCAGAAGACGGCATACGAGAT TCGGCGGG GTGACTGGAGTTCAGACGTGT
	CAAGCAGAAGACGGCATACGAGAT TCAATTTC GTGACTGGAGTTCAGACGTGT
	CAAGCAGAAGACGGCATACGAGAT GTAGGGAA GTGACTGGAGTTCAGACGTGT
	CAAGCAGAAGACGGCATACGAGAT AAGGGTGG GTGACTGGAGTTCAGACGTGT
	CAAGCAGAAGACGGCATACGAGAT ATAGTCTT GTGACTGGAGTTCAGACGTGT
	CAAGCAGAAGACGGCATACGAGAT TGGAGGGG GTGACTGGAGTTCAGACGTGT
	CAAGCAGAAGACGGCATACGAGAT TCAGGAAG GTGACTGGAGTTCAGACGTGT
	CAAGCAGAAGACGGCATACGAGAT ACCACAAC GTGACTGGAGTTCAGACGTGT
	CAAGCAGAAGACGGCATACGAGAT TTCAAGCC GTGACTGGAGTTCAGACGTGT
	CAAGCAGAAGACGGCATACGAGAT TACTCAGA GTGACTGGAGTTCAGACGTGT
	CAAGCAGAAGACGGCATACGAGAT TTCCTACT GTGACTGGAGTTCAGACGTGT
	CAAGCAGAAGACGGCATACGAGAT TTATATGG GTGACTGGAGTTCAGACGTGT
	CAAGCAGAAGACGGCATACGAGAT CCGTGACA GTGACTGGAGTTCAGACGTGT
	CAAGCAGAAGACGGCATACGAGAT TAGCGCCA GTGACTGGAGTTCAGACGTGT
	CAAGCAGAAGACGGCATACGAGAT CCTGGAGC GTGACTGGAGTTCAGACGTGT
	CAAGCAGAAGACGGCATACGAGAT CCCGCAAT GTGACTGGAGTTCAGACGTGT

RNA in red; all oligos are written from 5' to 3'; blue are barcodes

Appendix C

Supplementary File 2. A simplified working protocol for RNA cloning

Table S2: A simplified working protocol for RNA cloning

Note:

A. Each reaction is carried out in a single PCR tube with multiple steps by adding components sequentially to carry out different functionalities.

B. The final concentration of each component at an individual step is based on the accumulative (total) volume (number in red on line 1 of each table), including the volume of the previous step and the volume of the added components (starting from line 2 of each table).

Step 1: 3' ligation (0.1-1 μg total RNA or 0.1 μg small RNA <200nt)

	stock conc.	final conc.	10 μl/Rx	4 Rxs
H2O/small RNA			3.25 μl	13 μl
Ligation buffer without ATP	10 X	1 X	1 μl	4 μl
PEG-8000	50%	25%	5 μl	20 μl
3' linker	20 μM	0.5 μM	0.25 μl	1 μl
Truncated T4 RNA ligase 2	20 μM	0.5 μM	0.25 μl	1 μl
PIR-1 (optional)	10 μM	0.25 μM	0.25 μl	1 μl

Mix thoroughly by pipetting 20 times and incubate at room temperature for 2 hrs

The volume of each reaction is 10 μl.

Step 2: annealing

Inactivate at 65 °C for 10 mins, add 0.5 μl of 10 μM RT primer and anneal at 65 °C for 5 mins

Cool to room temperature by 0.1 °C/s

The volume of each reaction is 10.5 μl due to the addition of the RT primer.

Step 3: 5' ligation

	stock conc.	final conc.	20 μl/Rx	4 Rxs
H2O			8.1 μl	32.4 μl
ATP	20 mM	0.5 mM	0.5 μl	2 μl
5' linker	20 μM	0.4 μM	0.4 μl	1.6 μl
T4 RNA ligase 1	20 μM	0.25 μM	0.25 μl	1 μl
hDCP2 (optional)	20 μM	0.25 μM	0.25 μl	1 μl

Add 9.5 μl of the mixture to each reaction, and incubate at room temp. for 2 hrs

The volume of each reaction is 20 μl including 10.5 μl from the previous step.

Step 4: Reverse transcription (RT)

	stock conc.	final conc.	24.25 μl/Rx	4 Rxs
dNTP	10 mM	0.41 mM	0.99 μl	3.98 μl
RT dilution buffer	12 X	1 X	2.02 μl	8.08 μl
DTT	100 mM	4 mM	0.97 μl	3.88 μl
SSII	20 μM	0.21 μM	0.25 μl	1.02 μl

Add 4.25 μl of the mixture to each reaction, and incubate at 42 °C 30 mins & 85 °C 5mins

The volume of each reaction is 24.25 μl including 20 μl from the previous step.

Step 5: PCR

	stock	final	50 μl/Rx	4 Rxs
H2O			37.25 μl	149 μl
PFU bμffer	10 X	1 X	5 μl	20 μl
TMAC	1000 mM	15 mM	0.75 μl	3 μl
dNTP	10 mM	0.1 mM	0.5 μl	2 μl

5' PCR primer	10 µM	0.1 µM	**0.5** µl	**2** µl
Indexed 3' PCR primer	10 µM	0.1 µM	**0.5** µl	**2** µl
The above cDNA	10 X	1 X	**5** µl	**20** µl
PFU	100 X	1 X	**0.5** µl	**2** µl

1 cycle	**94 °C 1min**
5 cycles	**94 °C 20s; 53 °C 20s; 68 °C 40s**
11 cycles	**94 °C 20s; 68 °C 40s**
	4 °C forever

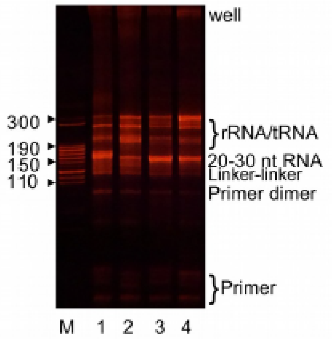
The volume of each reaction is 50 µl.

Add 3 µl each of 5' and indexed 3' PCR primer of 10 µM

2 cycles	**94 °C 20s; 68 °C 40s**
	4 °C forever

The volume of each reaction is 56 µl including the additional primers.

Indexed PCR products were compared visually using a 8% native PAGE gel, pooled according to a custom-specified ratio, phenol-extracted, precipitated and gel-purified as one sample.



M: 10 bp marker (110-200, 260 and 300)
1: 0.5 µg mouse testis total RNA
2: 0.4 µg mouse ovary total RNA
3: 0.5 µg *C. elegans* small RNA (<200 nt)
4: 1 µg *C. elegans* total RNA

Buffers

10 X ligation buffer without ATP:
0.5 M Tris pH7.5, 0.1 M $MgCl_2$, 0.1 M DTT

12 X RT dilution buffer:
250 mM Tris pH 8.8, 0.75 M KCl

Enzymes:
T4 RNA ligase 1 ((homemade)
Truncated T4 RNA ligase 2 (homemade)
Superscript II (SSII) from Invitrogen or homemade
PFU (commercial or homemade)
PIR-1 (homemade)
hDCP2 (homemade)
TMAC (Tetramethylammonium chloride, Sigma)

The theoretical size of a cDNA containing a 22-nt RNA insert is 146 bps. However, the apparent size on a PAGE gel may appear as 140-160 bps due to the sequence heterogeneity.

Appendix D

A) mRNA 3'/mRNA +1

RNA	rate
PB2	0.47%
PB1	1.98%
PA	0.33%
HA	0.89%
NP	1.67%
NA	8.52%
M	0.04%
NS	0.33%

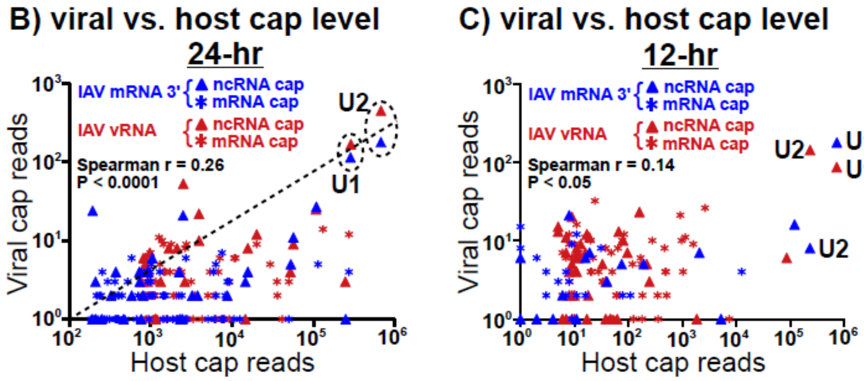


Figure S3. 1. Characterization of IAV caps utilized in non-canonical cap-snatching.

A) the ratio of capped RNA reads in mRNA 3' clusters to those at mRNA +1 for each IAV RNA strand; B) and C) the correlation between the levels of host and IAV caps at the 24 and 12-hr timepoints with 'blue' and 'red' representing caps utilized in mRNA 3' clusters and vRNA 5' regions respectively, and '▲' and '*' representing caps snatched from host ncRNA and mRNA respectively; the r and one-tail P values were calculated using non-parametric Spearman correlation.

Appendix E

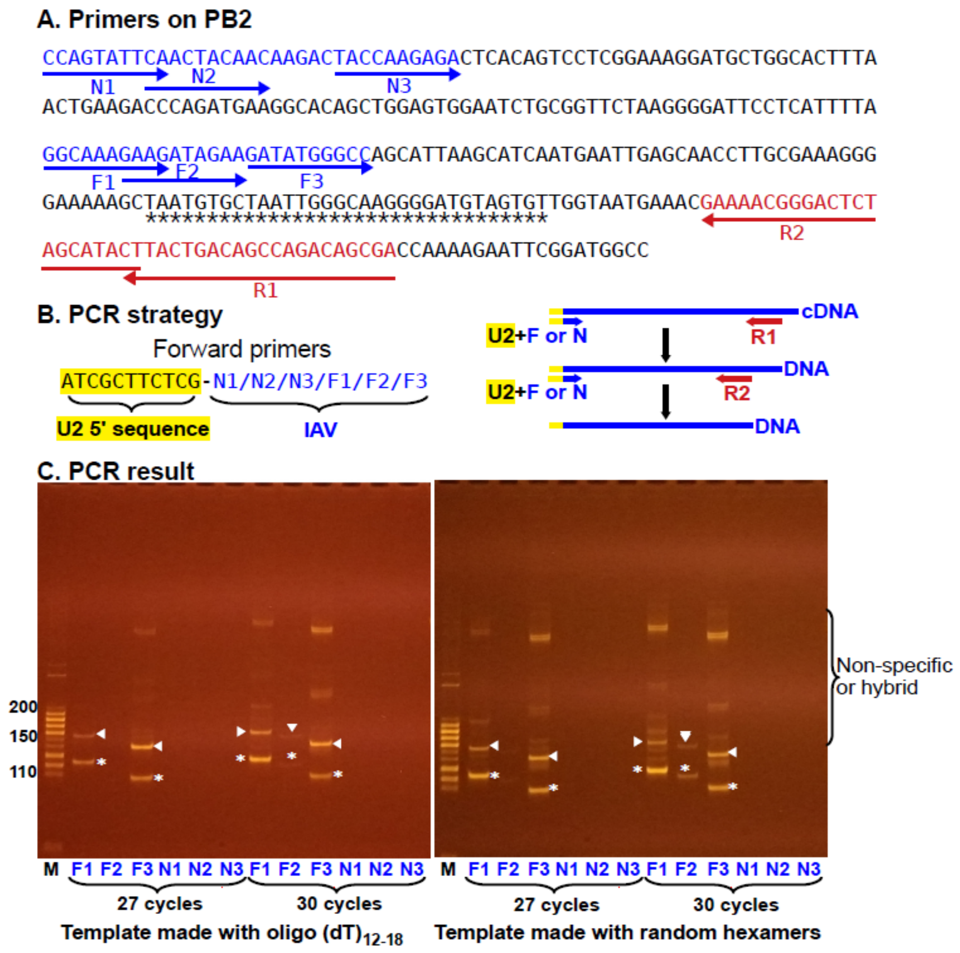

A. Primers on PB2
CCAGTATTCAACTACAACAAGACTACCAAGAGACTCACAGTCCTCGGAAAGGATGCTGGCACTTTA
N1
N2
N3
ACTGAAGACCCAGATGAAGGCACAGCTGGAGTGGAATCTGCGGTTCTAAGGGGATTCCTCATTTTA
GGCAAAGAAGATAGAAGATATGGGCCAGCATTAAGCATCAATGAATTGAGCAACCTTGCGAAAGGG
F1
F2
F3
GAAAAAGCTAATGTGCTAATTGGGCAAGGGGATGTAGTGTTGGTAATGAAACGAAAACGGGACTCT
****
R2
AGCATACTTACTGACAGCCAGACAGCGACCAAAAGAATTCGGATGGCC
R1
B. PCR strategy
Forward primers
ATCGCTTCTCG-N1/N2/N3/F1/F2/F3
U2 5' sequence
IAV
cDNA
U2+F or N
R1
DNA
U2+F or N
R2
DNA
C. PCR result
200
150
110
Non-specific
or hybrid
M F1 F2 F3 N1 N2 N3 F1 F2 F3 N1 N2 N3
27 cycles
30 cycles
Template made with oligo (dT)12-18
M F1 F2 F3 N1 N2 N3 F1 F2 F3 N1 N2 N3
27 cycles
30 cycles
Template made with random hexamers

Figure S3. 2. Verifying non-canonical cap-snatching using RT-PCR.

A) the PB2 sequence was used to design reverse primers R1 and R2 and forward primers F1-3 and N1-3 for the positive targets and negative controls respectively; B) Each forward primer is composed of a 11-nt U2 5' end sequence and and a 10-nt IAV sequence, generating similar Tm's for PCR; a forward primer and the reverse primer R1 were used in the first round of PCR and the resulting cDNA was used to carry out the second round of PCR (nested PCR) with the same forward primer and the reverse primer R2 for a total PCR cycle number of 27 and 30 (first + second); and C) the PCR results amplified using 1st strand cDNA template made with oligo (dT)12-18 or random hexamers were resolved on an 8% native PAGE gel with '▼' and '*' representing the normal and truncated products respectively; the bigger bands may represent nonspecific products or hybrid products, which contains one strand of the normal products and one strand of truncated products, when PCR reactions were overcycled.

Appendix F

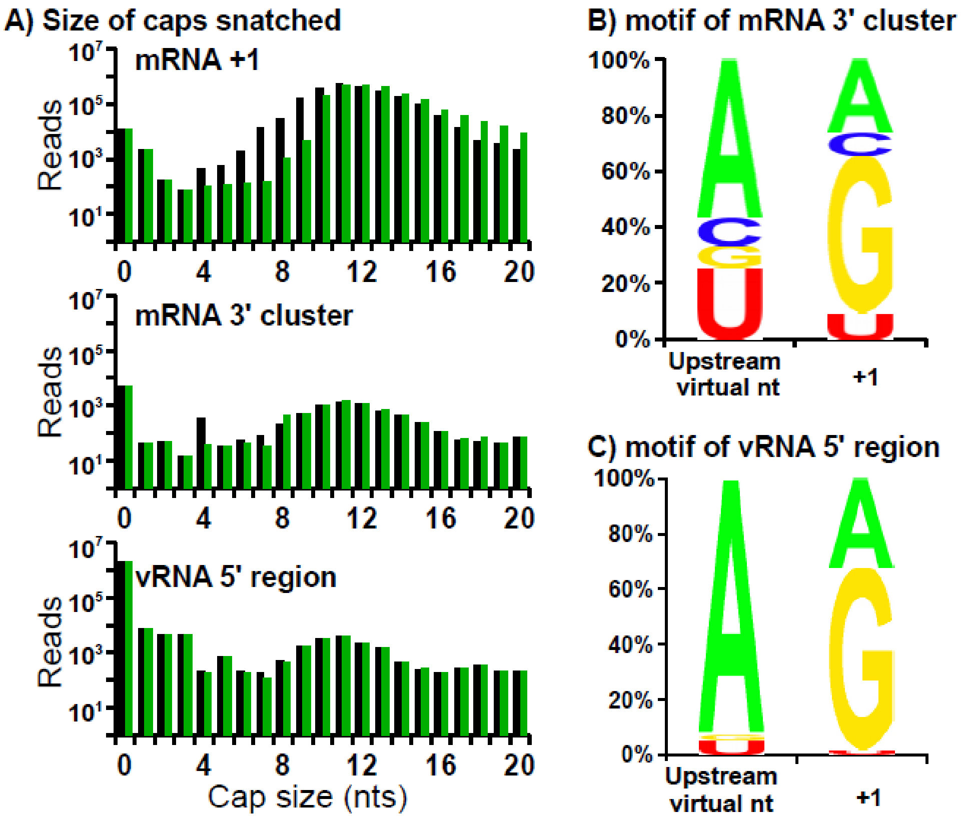
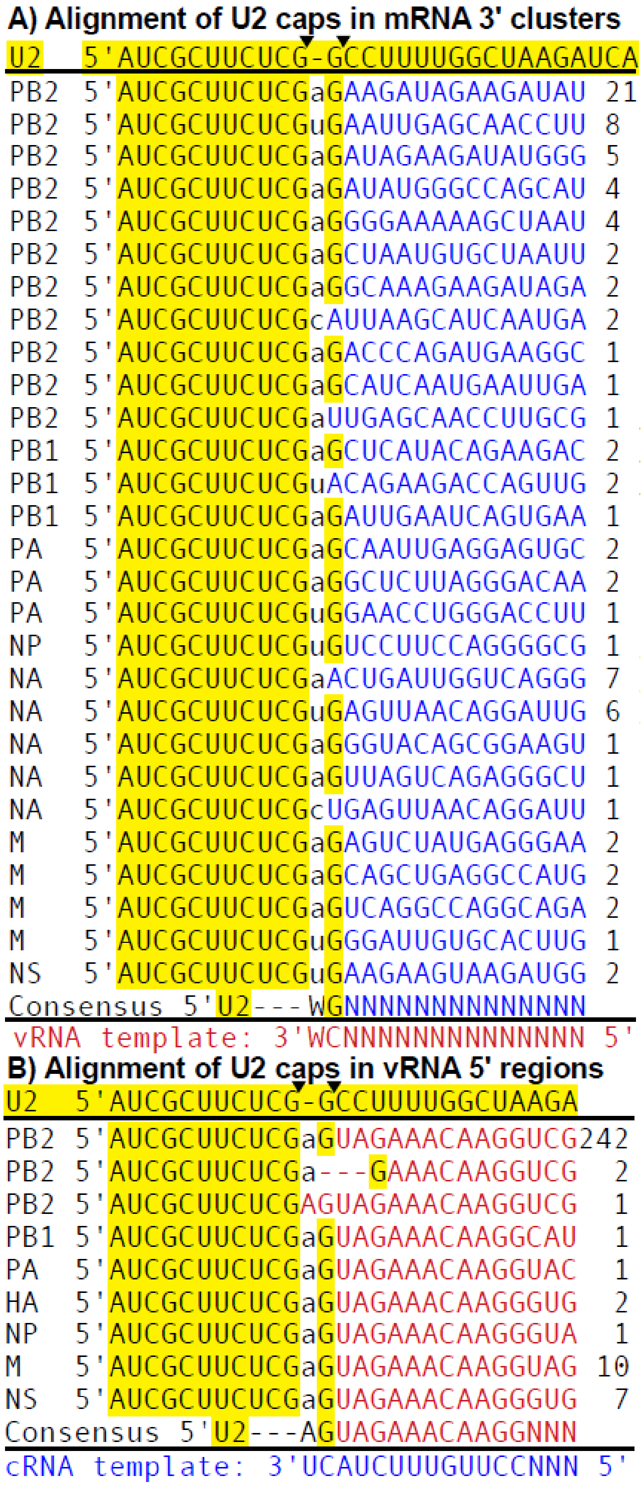


Figure S3. 3. Characterization of non-canonical cap-snatching.

A) the size distributions of IAV caps derived from IAV mRNA +1 (top), mRNA 3' clusters (middle) and vRNA 5' regions (bottom) including (green) and excluding (black) nts added via realignment. B and C) represent the nt frequencies of the first mapped nts of capped RNA reads and of the 'virtual nts' upstream of the mapped positions in mRNA 3' clusters and vRNA 5' regions respectively. When capped RNA +1 is mapped to vRNA +1, the virtual nt can not be defined since there is no upstream nt available.

Appendix G



A) Alignment of U2 caps in mRNA 3' clusters

```
U2    5'AUCGCUUCUCG-GCCUUUUGGCUAAGAUCA
PB2  5'AUCGCUUCUCGaGAAGAUAGAAGAUAU  21
PB2  5'AUCGCUUCUCGuGAAUUGAGCAACCUU  8
PB2  5'AUCGCUUCUCGaGAUAGAAGAUAUGGG  5
PB2  5'AUCGCUUCUCGaGAUAUGGGCCAGCAU  4
PB2  5'AUCGCUUCUCGaGGGAAAAAGCUAAU  4
PB2  5'AUCGCUUCUCGaGCUAAUGUGCUAAUU  2
PB2  5'AUCGCUUCUCGaGGCAAAGAAGAUAGA  2
PB2  5'AUCGCUUCUCGcAUUAAGCAUCAAUGA  2
PB2  5'AUCGCUUCUCGaGACCCAGAUGAAGGC  1
PB2  5'AUCGCUUCUCGaGCAUCAAUGAAUUGA  1
PB2  5'AUCGCUUCUCGaUUGAGCAACCUUGCG  1
PB1  5'AUCGCUUCUCGaGCUCAUACAGAAGAC  2
PB1  5'AUCGCUUCUCGuACAGAAGACCAGUUG  2
PB1  5'AUCGCUUCUCGaGAUUGAAUCAGUGAA  1
PA   5'AUCGCUUCUCGaGCAAUUGAGGAGUGC  2
PA   5'AUCGCUUCUCGaGGCUCUUAGGGACAA  2
PA   5'AUCGCUUCUCGuGGAACCUGGGACCUU  1
NP   5'AUCGCUUCUCGuGUCCUUCCAGGGGCG  1
NA   5'AUCGCUUCUCGaACUGAUUGGUCAGGG  7
NA   5'AUCGCUUCUCGuGAGUUAACAGGAUUG  6
NA   5'AUCGCUUCUCGaGGGUACAGCGGAAGU  1
NA   5'AUCGCUUCUCGaGUUAGUCAGAGGGCU  1
NA   5'AUCGCUUCUCGcUGAGUUAACAGGAUU  1
M    5'AUCGCUUCUCGaGAGUCUAUGAGGGAA  2
M    5'AUCGCUUCUCGaGCAGCUGAGGCCAUG  2
M    5'AUCGCUUCUCGaGUCAGGCCAGGCAGA  2
M    5'AUCGCUUCUCGuGGGAUUGUGCACUUG  1
NS   5'AUCGCUUCUCGuGAAGAAGUAAGAUGG  2
Consensus 5'U2---WGNNNNNNNNNNNNNN
 vRNA template: 3'WCNNNNNNNNNNNNNN 5'
```

B) Alignment of U2 caps in vRNA 5' regions

```
U2   5'AUCGCUUCUCG-GCCUUUUGGCUAAGA
PB2  5'AUCGCUUCUCGaGUAGAAACAAGGUCG242
PB2  5'AUCGCUUCUCGa---GAAACAAGGUCG   2
PB2  5'AUCGCUUCUCGAGUAGAAACAAGGUCG   1
PB1  5'AUCGCUUCUCGaGUAGAAACAAGGCAU   1
PA   5'AUCGCUUCUCGaGUAGAAACAAGGUAC   1
HA   5'AUCGCUUCUCGaGUAGAAACAAGGGUG   2
NP   5'AUCGCUUCUCGaGUAGAAACAAGGGUA   1
M    5'AUCGCUUCUCGaGUAGAAACAAGGUAG  10
NS   5'AUCGCUUCUCGaGUAGAAACAAGGGUG   7
Consensus 5'U2---AGUAGAAACAAGGNNN
cRNA template: 3'UCAUCUUUGUUCCNNN 5'
```

Figure S3. 4. The cleavage/priming motif of IAV U2 caps.

A and B) The capped reads with the U2 sequence AUCGCUUCUCG were mapped to each IAV mRNA 3' cluster with 'yellow' highlighting the U2-derived sequence in U2 snRNA (top) and each hybrid IAV sequence, 'blue' indicating IAV mRNA/cRNA, 'red' representing vRNA, 'lowercase' indicating a virtual nt not transcribed but converted from the template nt, the last column representing the read number, arrows indicating the cap cleavage sites, and 'W' in the consensus sequence representing A or U.

Appendix H

Supplementary File 3. IAV protein reads from high-throughput sequencing

>peptide_sequence and protein type (hybrid contains amino acids encoded by the caps)
IAV_strand, read_number, the_size_range_of_5'_UTR, the_first_IAV_nt_position
A bracket indicates amino acids encoded by snatched host caps
Kozak(A/G)NNAUGG is used to identify the start AUG

In the cRNA 3' cluster

>(MA)EDRRYGPALSINELSNLAKGEKANVLIGQGDVVLVMKRKRDSSILTDSQTATKRIRMAIN	hybrid_PB2		
PB2	1	5-5	2123
>(M)DRRYGPALSINELSNLAKGEKANVLIGQGDVVLVMKRKRDSSILTDSQTATKRIRMAIN	hybrid_PB2		
PB2	2	9-9	2126
>(M)EDPDEGTAGVESAVLRGFLILGKEDRRYGPALSINELSNLAKGEKANVLIGQGDVVLVMKRKRDSSILTDSQTATKRIRMAIN	hybrid_PB2		
PB2	1	11-11	2054
>(MED)RRYGPALSINELSNLAKGEKANVLIGQGDVVLVMKRKRDSSILTDSQTATKRIRMAIN	hybrid_PB2		
PB2	5	8-12	2123
>(M)EGTAGVESAVLRGFLILGKEDRRYGPALSINELSNLAKGEKANVLIGQGDVVLVMKRKRDSSILTDSQTATKRIRMAIN	hybrid_PB2		
PB2	1	9-9	2066
>MEYDAVATTHSWVPKRNRSILNTSQRGILEDEQMYQRCCNLFEKFFPSSSYRRPVGISSMVEAMVSRARIDARIDFESGRIKK EEFAEIMKICSTIEDLRRQK	PB1		
PB1	55	40-111	1986
>(MVG)SYRRPVGISSMVEAMVSRARIDARIDFESGRIKKEEFAEIMKICSTIEDLRRQK	hybrid_PB1		
PB1	1	3-3	2131
>MENLNKKVDDGFIDIWTYNAELLVLLENERTLDFHDSNVKNLYEKVKSQLKNNAKEIGNGCFEFYHKCNDECMESVKNGTYD YPKYSEESKLNREKIDGVKLESMGVYQILAIYSTVASSLVLLVSLGAISFWMCSNGSLQCRICI	HA		
HA	1	12-12	1289
>MDAIVSSTLELRSRYWAIRTRSGGNTNQQRASAGQISTQPTFSVQRNLPFDKATIMAAFSGNTEGRTSDMRAEIIKMMESAR PEEVSFQGRGVFELSDERATNPIVPSFDMSNEGSYFFGDNAEEYDN	NP		
NP	13	10-78	1155
>MAAFSGNTEGRTSDMRAEIIKMMESARPEEVSFQGRGVFELSDERATNPIVPSFDMSNEGSYFFGDNAEEYDN	NP		
NP	21	5-136	1320
>MESARPEEVSFQGRGVFELSDERATNPIVPSFDMSNEGSYFFGDNAEEYDN	NP		
NP	17	4-49	1386
>(M)DCIRPCFWVELVRGLPRENTTIWTSGSSISFCGVNSDTANWSWPDGAELPFTIDK	hybrid_NA		
NA	1	11-11	1263
>(MGG)GSFVQHPELTGLDCIRPCFWVELVRGLPRENTTIWTSGSSISFCGVNSDTANWSWPDGAELPFTIDK	hybrid_NA		
NA	1	3-3	1227
>MVLASTTAKAMEQMAGSSEQAAEAMEVASQARQMVQAMRAIGTHPSSSTGLKNDLLENLQAYQKRMGVQMQRFK	M1		
M	5	11-14	559
>MEVASQARQMVQAMRAIGTHPSSSTGLKNDLLENLQAYQKRMGVQMQRFK	M1		
M	11	10-20	631
>(MAD)DSSDPLVVAASIIGIVHLILWIIDRLFSKSIYRIFKHGLKRGPSTEGVPESMREEYREEQQNAVDADDDHFVSIELE	hybrid_M		
M	1	4-4	771
>(MD)DSSDPLVVAASIIGIVHLILWIIDRLFSKSIYRIFKHGLKRGPSTEGVPESMREEYREEQQNAVDADDDHFVSIELE	hybrid_M		
M	1	11-11	771

In the vRNA 5' region

>METPKTVLNILNMPIIIPGLNDAVPSINNGRIFSIFFLVESLKYFKSMLASISAGI	no_match		
PB1	34	1046-1049	1050
>MVKEISVSWELVIIFLTTFANFAFFSFPPTGNPDCSSFSHMLLASVSTKYTNPLICIPGVAIALRFSFPLSASLVIVFRVNALIR	no_match		
PB1	34	1409-1412	1413
>MDRFLLGTQECVVATASYSIFLAGPCAGIITALFTDSISLWLTNGFNGLHKRPW	no_match		
PB1	34	293-296	297
>MGLDTWPIALLRSISPISRTQYFSHLCGSSLGSVRENSMLTKFTTSVSFLK	no_match		
PA	748	518-527	528
>MVKHDPFIQTHSDSCVLNICFFQLFMVPVIIPLYFSTATAPLSGPEIPIVNQPMPSWHALADHATDSNFELYGDGASPRGQLIKAL	no_match		
NA	37	719-720	721