Targets Identification and Characterization of Tramp Comples in Mouse

Fengchao Wang
Marquette University

TARGETS IDENTIFICATION AND CHARACTERIZATION OF TRAMP COMPLEX IN MOUSE

By
Fengchao Wang, B.S

A Thesis submitted to the Faculty of the Graduate School,
Marquette University,
in Partial Fulfillment of the Requirements for
the Degree of Master of Science

Milwaukee, Wisconsin
May, 2015

ABSTRACT
TARGETS IDENTIFICATION AND CHARACTERIZATION OF TRAMP COMPLEX IN MOUSE

Fengchao Wang, B.S

Marquette University, 2015

RNA surveillance and degradation play an important role in the development and growth of organisms by eliminating RNA that contains errors, or that is no longer needed by the cell. In some processes, RNAs designated to be degraded are first labeled and then specifically recognized by the exosome, which performs the final degradation. One of the key labeling factors in yeast is the TRAMP complex, a three-subunit complex composed of Air2, Trf4 and Mtr4. Air2 facilitates TRAMP binding of RNA, Trf4 appends a 3′ end polyA tail and Mtr4 regulates the rate of adenylation and modifies RNA structures for ease of degradation through its RNA helicase activity. Though TRAMP has been studied extensively in yeast and its biochemistry and RNA recognition functions well delineated, the recent identification of TRAMP in mammals has made it possible for work to characterize mammalian TRAMP function in tissue culture cells.

The mammalian transcriptome is much more complex and diverse compared to yeast in that a large portion is consisted of non-coding RNAs such as lncRNA, snoRNA, miRNA and so on. To understand the role of TRAMP complex in gene expression regulation, we knockdown the SKIV2L2 (mouse Mtr4) subunit and performed a polyA sequencing. With bioinformatics tools such as Bowtie, F-Seqq, MEDIPS, miRCompare, we constructed a data pipeline and identified several categories of targets including snoRNA, rRNA, miRNA and long-noncoding RNA in mouse cells.

These data suggests that the targets of TRAMP are widely spread along the genome, and these targets involve a myriad of regulatory pathways. Understanding the relationship between the targets will help reveal the function and effect of this complex. Also, a more accurate and comprehensive target identification method remains to be developed.

ACKNOWLEDGMENTS

Fengchao Wang, B.S

I would like to thank Dr. James Anderson, for his mentorship and insight in experimental technology. The opportunity to work with NGS and bioinformatics here completely changed my career and life path.

Thanks all my current and past committee members: Dr. Allison Abbott, who gave me advice on research as well as time management. Dr. Serdar Bozdag, who gave me valuable instructions on programming design. Dr. Pinfen Yang, Dr. Munroe and Dr. Karrer, for their help with my academic progression.

Thanks to all lab members, Yan, Alexis and Dr. Dorweiler, who guided me into the bioinformatics world and gave me practical help in the beginning phase of my research.

Lastly, I wish to thanks all department members, my friends, especially by best buddy Yi. He answered me tons of biology questions when we just met and helped me a lot when I left Milwaukee.

TABLE OF CONTENTS

I. Introduction

A. TRAMP complex in yeast

In the yeast *Saccharomyces ceresvisiae*, the **Tr**f4/**A**ir2p/**M**tr4p **p**olyadenylation complex (TRAMP) recognizes and targets a diverse set of RNAs for exosome-mediated RNA degradation or processing by appending a 3′ oligoadenylate tract [1-3]. TRAMP complex consists of the polyA polymerase, Trf4p, the zinc-knuckle domain protein, Air2, and the RNA-dependent ATPase, SKIV2L2p. Studies on Air2 show only two of its five zinc-knuckle domains are required for full TRAMP activity *in vitro* and *in vivo* [4], and this is likely because of a failure of Air2p to bind or position the RNA substrate so Trf4p can polymerize addition of the 3′ adenylate tract [4]. Trf4p/Pap2p is a nucleotidyltransferase with preference for polymerizing adenylates [5, 6], but it is unable to polyadenylate RNA substrates in the absence of Air2p [7, 8] which suggests Air2p might bind RNA substrates for Trf4p, which Trf4p needs given it has no recognizable RNA-binding domain like the canonical polyA polymerase, Pap1p [8]. After adenylation of RNA substrates by Trf4p, SKIV2L2p removes RNA secondary structure through ATP-hydrolysis and facilitates exosome degradation or processing by a yet unknown mechanism. Recent research suggests the exosome RNA binding protein complex Nrd1p/Nab3p/Sen1p (NNS) triggers transcription termination of RNA PolII and deliver the free 3’ end to TRAMP [9, 10].

Mammalian cells are more complex and sophisticated than yeast in RNA types and regulation pathways [11]. Some non-coding RNA exist in mammalian cells but are not present in yeast, such as miRNA and some long non-coding RNA. The diverse cellular behavior like differentiation and apoptosis demand delicate control of RNA synthesis and decay. This complexity potentially requires additional functionality of the exosome as well as its targeting complexes. The function advancement in mammalian cells comes with changes in structure and composition compared to their yeast counterpart. The move of investigation from yeast to mammalian cells may reveal more targets of TRAMP and provide a broader and thorough insight into the function and mechanism of TRAMP-induced RNA degradation pathway. Elucidating the function of RNA degradation pathways may bring deeper understanding of critical pathologies. Investigating the targeting preference and pattern in mammalian cells may exert a great impact on disease diagnostics and treatment. It’s already known that the targets of yeast TRAMP include hypomodified pre-tRNA, aberrant 5SrRNA and 7S pre 5.8S rRNAs, SRP RNAs and cryptic unstable transcripts (CUTs) (Allmang, Mitchell et al. 2000, Kadaba, Krueger et al. 2004, Wyers, Rougemaille et al. 2005). The increasing number and diversity of TRAMP RNA targets in yeast coupled with the pervasive transcription in mammals, underscores the importance of uncovering mouse SKIV2L2 (SKIV2L2) targets in mammals.

B. TRAMP in mammalian cells

A TRAMP-like RNA degradation and processing complex has been identified and characterized in human. The composition of human TRAMP complexes (Lubas, Christensen et al. 2011) is similar to TRAMP complexes in yeast [12], and contains human Trf4p/5p homolog hTRF4-2, ZCCHC7 (Air1p/2p homolog) and SKIV2L2 (Mtr4 homolog) [12]. Human SKIV2L2 is not limited to TRAMP but also have been found in novel Nuclear Exosome Targeting (NEXT) complexes [12]. NEXT works in parallel with TRAMP, but it appears a division of labor is accomplished by restricting NEXT to the nucleoplasm and TRAMP mostly in the nucleolus [12]. Recent research reported that NEXT complex is loaded to newly synthesized RNAs including snoRNAs via RBM7 and a cap-binding protein complex (CBP) [13].

Non-coding RNAs, including small and long non-coding RNAs, account for a large portion of transcriptome (Mattick and Makunin 2006). It has been reported that 98% of the genome output is ncRNAs, and almost all regions of a genome are expressed as non-coding RNAs resulting from what has been termed pervasive transcription (Huttenhofer, Schattner et al. 2005, Hangauer, Vaughn et al. 2013). Non-coding RNAs are being recognized important as more ncRNAs are directly implicated playing critical roles in various cellular pathways and processes. For example, miRNAs are involved in a series of developmental processes and function in regulating gene expression (PILLAI 2005). Long non-coding RNAs are linked to X chromosome inactivation, imprinting, and control of pluripotency (Wan and

Bartolomei 2008, Lee 2011). In addition, ncRNAs can serve in signal transduction, and structural scaffolding (Fatica and Bozzoni 2014). In spite of the undeniable importance of ncRNAs, the biogenesis, maturation, processing and degradation of most non-coding RNAs remains uncharacterized.

On the other hand, dysfunction of exosome degradation pathway and abnormal targeting may cause serious diseases and disorder including cancer, neurological disorders and liver diseases [14-16]. For example, C-Myc is a cell proliferation and differentiation regulating protein that is also considered as an oncogene. The regulation of C-Myc is primarily through posttranscriptional RNA decay [17, 18]. Von Hippel-Lindau (VHL) tumor, a clear-cell renal carcinoma is caused by the dysregulation of VHL tumor suppressor which regulates the decay of a growth-factor encoding transcript.

C. rRNA processing and maturation

Ribosomal RNA is a critical component of ribosome, and is required for protein synthesis. Eukaryotic ribosome composes of three rRNAs: 18S, 5.8S and 28S. These rRNAs are produced form the 45kb rRNA complete repeat sequence that is distributed across the genome. Biogenesis of rRNA is a lengthy and diverged process that requires the participation of a series of endoribonucleases and exoribonucleases as well as some long and small non-coding RNAs [19]. In general, 10 cleavage events occurs during the maturation process (Fig.I-1). The 5′ external transcribed spacer

(ETS) have two cleavage events. One at 600bp from the 5′ end, named A′, the primary cleavage. Another at 1600bp downstream from the 5′ end, named A_0. This cleavage is critical for the maturation of 18S rRNA. The processing event produces the 5′-A′ and A′-A_0 fragments which are both degraded rapidly by exosome [20, 21]. The maturation of 5.8S rRNA happens after the cleavage of A_0. Two forms of cleavage occurs at the 5′ of 5.8S rRNA: one cleaves the complete internal transcribed spacer 1 (ITS1) (short form) and another leave a trunk of nucleotides at the 5′ end (long form). After this, the cleavage in ITS2 produces the pre-28S rRNA and 12S (7S in yeast) rRNA, the precursor of 5.8S, with an extended 3' tail. This tail is trimmed by exosome in yeast to form the mature 5.8S rRNA. [22]

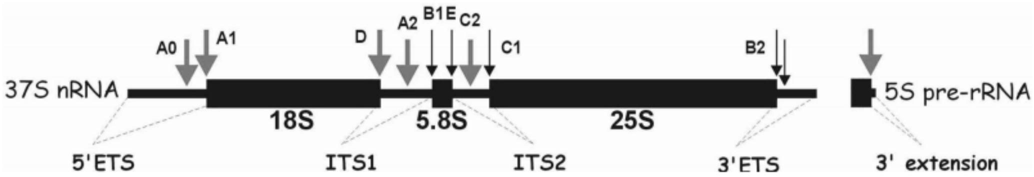

Figure I-1 Structure of rRNA gene and clevage events of eukaryotic rRNA maturation pathway. Arrows indicate cleavage sites. Picture from [19]

Our lab has reported that the 5′ ETS of 45s rRNA has been targeted by the mouse SKIV2L2 subunit [23]. Based on the role that TRAMP promotes the maturation of 5.8S rRNA in yeast [24], it is reasonable to speculate that TRAMP plays a more pervasive role in addition to the processing of 5′ ETS fragments. The maturation of

5.8S rRNA involves the cleavage at the 5′ end which splits it from the 18S rRNA, and the cleavage at the 3′ end that forms the 7S rRNA (precursor 5.8S). TRAMP is required for the removal of the 3′ end of the 7S rRNA that forms the matura 5.8S rRNA in yeast.[20, 24, 25]. The maturation process is elusive and requires exploration.

D. miRNA processing and maturation

miRNA is a small non-coding RNA that is involved in expression regulation and gene silencing. miRNA biogenesis have two intermediates: pri-miRNA and pre-miRNA. The miRNA primary transcript forms a stem-loop structure. In the first step, the Drosha clevaes the 5′and 3′end of precursor hairpin miRNA (pri-miRNA), and produces the pre-miRNA (Fig. I-2). The pre-miRNA are processed by Dicer and then forms RISC to degrade and silence mRNAs [26-28]. It is believed that the 5’ and 3’ fragments after Drosha cleavage have been degraded. How and where are they degraded remain elusive.

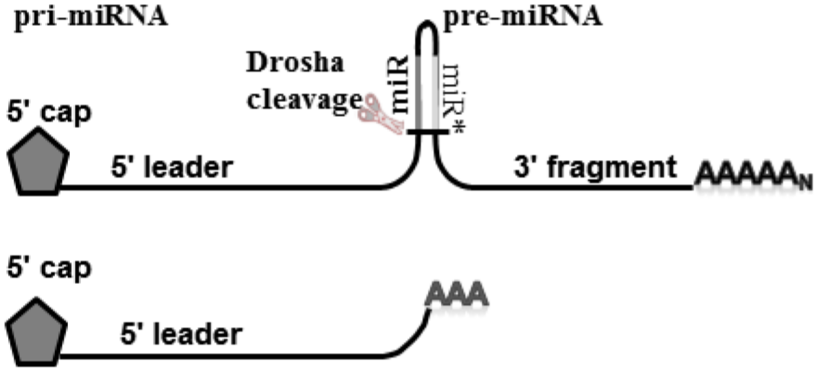

Figure I-2. Maturation process of miRNA. Picture from [26]

Transcription in the mouse is comparable to other mammals and is known to be quite pervasive, highlighting the need to eliminate processing byproducts, aberrant RNAs and RNAs with no functional significance. The human NEXT complex facilitates degradation of a class of antisense promoter-associated transcripts termed PROMPTS that may inhibit normal transcription of the cognate gene if allowed to accumulate [12]. Proteomic and protein interactome studies have shown that hSKIV2L2 copurifies with ARS2, a protein that is linked to miRNA biogenesis (Lubas, Christensen et al. 2011). This observation suggests a potential role of TRAMP in miRNA biogenesis.

E. snoRNA processing and maturation

Small nucleolar RNA (snoRNA) is a category of small RNA that participates rRNA, snRNA and tRNA modification and processing [29-32]. snoRNA can be categorized into two groups: C/D Box and H/ACA Box snoRNA according to their unique sequence motif. The C/D box snoRNA is related to methylation while H/ACA box to pseudouridylation [33, 34]. In general, snoRNA forms snoRNP with associating proteins, and its special sequence guided itself to the target rRNA for modification. Each snoRNA/snoRNP has its own modification location and type [35]. For example, snoRNA U3 functions in the preribosomal RNA processing, specifically the 5' ETS [36]. In mammalian cells, snoRNA reside in introns of coding and non-coding genes [37], and is released via splicing events [38].

It is intriguing to unveil the role TRAMP played in mammalian cells. This thesis addressed the following questions and tasks:

1. Evaluate the degradation effect of SKIV2L2 on coding genes.

2. Identify targets of SKIV2L2

3. Understand the behavior and mechanism of SKIV2L2-mediated RNA surveillance in mammalian cells.

II. Results

A. PolyA sequencing information and internal (A) filtration

The increasing understanding that ncRNA have functions in regulating gene expression on a micro and macro genomic scale underscores the need to fully understand the steady-state expression of ncRNAs. The establishment of TRAMP in targeting a wide variety of RNAs in yeast and mammals, prompted us to interrogate the mouse transcriptome for changes in polyadenylated RNA levels after depleting the TRAMP subunit, SKIV2L2. We used poly-A Seq, a derivative of RNA-Seq, to address the question of how depletion of SKIV2L2 from mouse cell lines affects RNA degradation and processing by identifying polyA+ RNAs that accumulate upon SKIV2L2 depletion. The library was constructed as described [39] and paired-end sequencing using an Illumina Hi-Seq 2000 platform was performed. RNA fragmented to ~200bp were reverse transcribed using an special oligo-d(T) primer generating cDNAs bearing bar-coded sequences for identification of the 3′ and 5′ ends. cDNA sequencing reads from N2A cells (75%) depleted of SKIV2L2 were digitalized and stored in text files as raw reads, which contains the read quality score and the sequence for each read. Bowtie2 was used to map the raw reads to genomic position in the UCSC mm10 mouse genome assembly [40]. A total of two sequencing runs were done independently, and throughout this thesis they are called original and replicate or sample 1 or 2. All samples were mapped with an 80% or greater mapping rate. The mapping information of the four samples is shown in Table II-1.

	Control1	Control2	SKIV2L21	SKIV2L22
Total Reads	18,248,934	32,697,560	15,990,484	30,270,156
Mapped Reads	15,318,665	27,516,039	11,567,908	25,526,134
Mapping Rate	0.84	0.84	0.72	0.84
5′ Reads	7,882,519	13,916,229	5,827,590	12,972,682
3′ Reads	7,436,146	13,599,810	5,740,318	12,553,452

Table II-1 Summary of read number and mapping rate. Four samples were sequenced and the sample name consists the sample type and batch. The original sequencing batch have roughly half sequencing depth compared to the second batch. There out of four samples have the mapping rate at 84% except for SKIV2L21. Number of 5′ reads are slightly greater than that of 3′ reads in all samples.

We used Oligo-d(T) to select and reverse transcribe RNAs that have a polyadenylated tail. While cleaning the mapped data to remove duplicate reads, a random sample of reads mapping to introns were visualized at the nucleotide level. It was discovered that distributed across the genome there are reads in the library with 3′ ends that represent genomically encoded oligo-A tracts that do not reflect bonafide posttranscriptional polyadenylation. Instead of eliminating these transcripts experimentally, we designed a read quality assurance program to address this problem in a more efficient and computationally economical way. Every 3′ end read was searched for the presence of either 5 contiguous adenosines, or 15 adenosine interspersed upstream or downstream within 20 nucleotides of the first nucleotide of the 3′ read. If the sequence reveals either of these criteria, the read was removed from further analyses. On average, 18.2% of the 3′ reads were identified as internally primed, which if left in during analysis would have reported a significant number of false positives.

B. Mouse SKIV2L2 knockdown does not result in the accumulation of protein coding mRNAs

The role SKIV2L2 plays in mouse RNA surveillance is predictable but was to an extent unconfirmed. A series of papers have been published on yeast TRAMP and SKIV2L2 (yeast SKIV2L2) substrates and biochemical function [3, 7, 41], and a few papers on how human SKIV2L2 impacts RNA turnover [12]. We fully expected to find that SKIV2L2 is required for RNA turnover in mouse, but the limited targets so far reported is much smaller than would be predicted. Interestingly, research in fungi and mammals has been suggestive that there are both common RNA substrates between these diverse eukaryotes, as well as substrates unique to each. As the first step toward data analysis, we collected reads at the 3′ end of each RefSeq coding genes and compared the number of reads each pair of control experimental samples to understand the effect SKIV2L2 exerts on protein coding genes.

The ~85% depletion of *Skiv2l2* in N2A cells was confirmed by q-RT PCR and western blotting [26]. Considering the limited role of TRAMP in mediating mRNA processing and degradation [42], a prediction was tested that the number of sequencing reads from the annotated 3′ end of protein coding genes would not change between control knockdown and SKIV2L2 knockdown. The initial analysis of SKIV2L2 knockdown was focused on identification and characterization of all mRNAs that accumulate upon depletion of SKIV2L2 relative to a control knockdown. We designed a data pipeline that transform the read count data to address this question. An overview of the analysis workflow is shown in Fig. II-1:

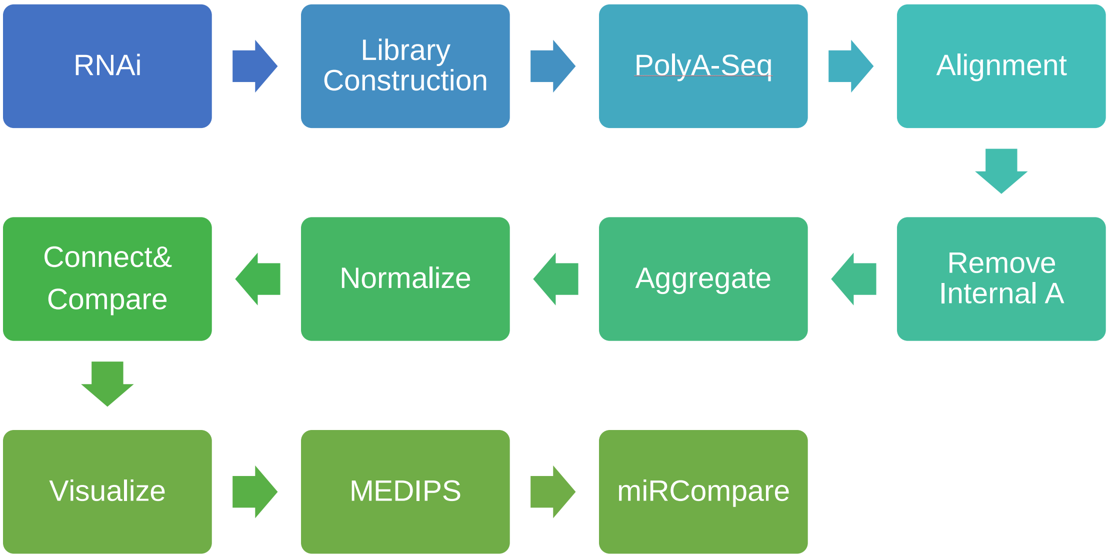

Figure II-1 Overview of the workflow of differential polyadenylation analysis. Analysis steps are represented by rectangles with the step name inside. Arrows indicates the analysis sequence. Samples were first knockdown by siRNA, and the sequencing library were constructed. PolyA-Seq, polyadenylation sequencing were conducted to sequence transcripts with a polyA tail. The result were mapped to UCSC mm10 genome assembly. Due to the aberrant annealing of oligo d(T) primers, some sequencing artifact were removed from the results. The sequencing reads were mapped to coding Refseq genes based on the genomic location. Identical procedures were used to process all the samples before normalizing with FPKM method and comparison between same Refseq entries. Data was visualized in histogram and subject to further analysis by MEDIPS and miRCompare (discussed below).

Control and SKIV2L2 knockdown from replicate sequencing were used to compute the number of reads at each genomic location. The polyadenylation level at each nucleotide across the genome was calculated by counting the number of reads that ends at that location, then all data normalized to allow direct comparisons of read abundance between samples. To normalize the number of reads at each nucleotide of all samples, fragments per kilobase per million mapped reads (FPKM) was calculated as described [43].

To compare the read abundance of each sample to the protein-coding RefSeq gene annotations, first, reads at the 3′ annotated end of each RefSeq protein coding gene were collected. Taking into account the mapping accuracy of Refseq annotation and the resolution of polyA-seq, it was decided to collect all reads 50bp up/downstream of the 3′ end of the transcript. This is a lenient range since the target transcript is supposed to reside right at the 3′ end with only a few nucleotides offset. This setting tolerates the inaccurate mapping and ensured all the potential reads for a given RefSeq would be counted. A script (shown in supplementary data) was written to transform the primitive RefSeq data to the customized RefSeq data containing the 101 base pair search window. The Galaxy bioinformatics platform was deployed for joining the coding Refseq genes into sample sequencing reads [44]. The ratio of SKIV2L2 replicate knockdown read count to that of control was calculated using the formula below to indicate the difference between the replicate control and SKIV2L2 knockdown samples (Control knockdown replicate and SKIL2L2 knockdown replicate):

$f_{(i)}=\log_2(\mathrm{SKIV2L2}_{(i)}/\mathrm{control}_{(i)})$ **i=RefSeq genes**

Based on the transformed differential polyA reads value, a histogram graph was plotted to demonstrate the trend of change (Figure II-2)

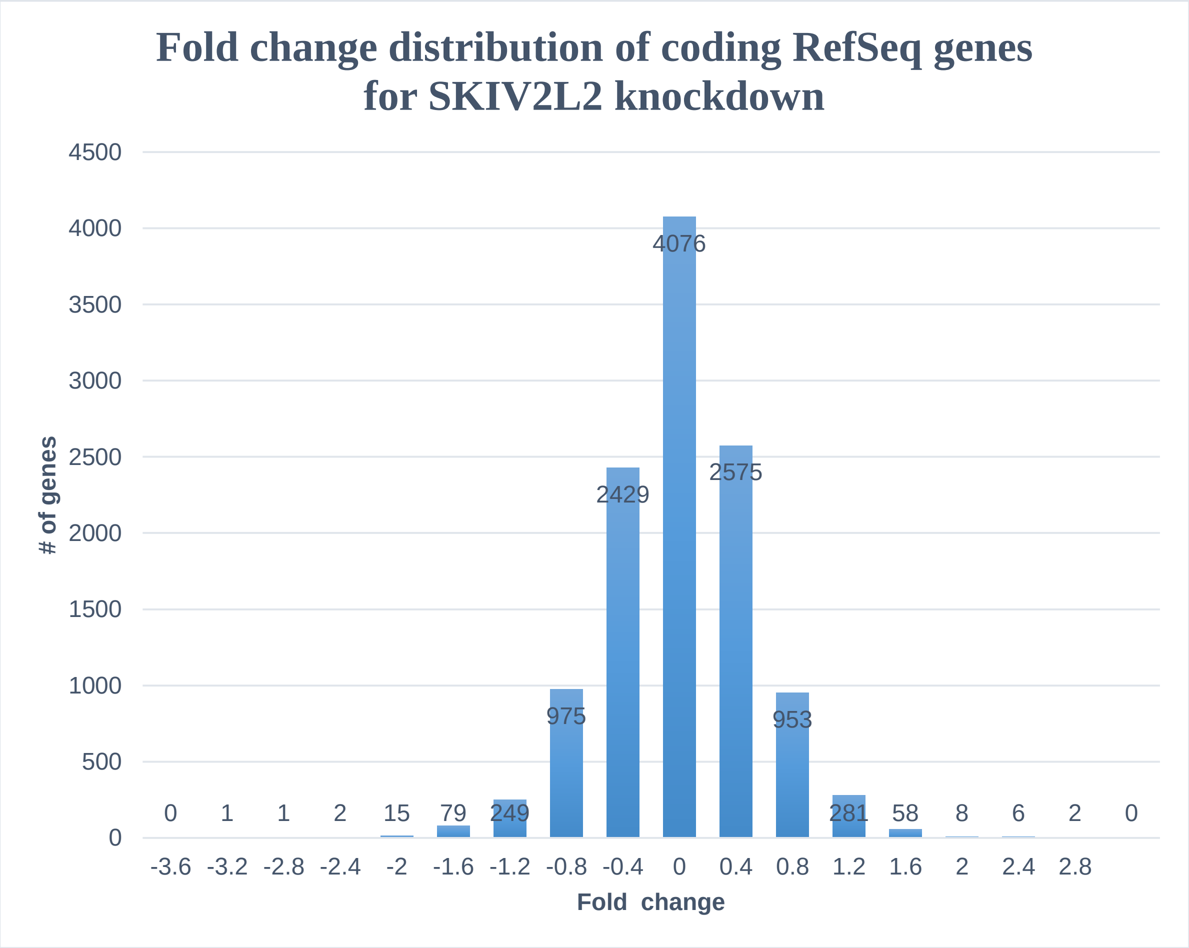

Figure II-2 Fold change distribution of coding RefSeq genes for SKIV2L2 knockdown. X-axis represent fold change in base 2 logarithm. The bin size is 0.4, the number below each bar shows the median of the bin. 0 means SKIV2L2 and control have equal level of polyadenylation. Positive value indicates SKIV2L2 is higher than control. Negative value indicates control over SKIV2L2. Value 2 means 4 times fold change. Y-axis denotes the number of RefSeq genes.

The fold change distribution of the RefSeq genes roughly follows normal distribution, only slightly biased toward SKIV2L2 knockdown, suggesting that knockdown of SKIV2L2 does not dramatically affect mRNA levels that are subject to polyadenylation by the canonical polyadenylation machinery. We cannot rule out that this slightly biased shift toward an increase/decrease in gene expression found in N2A cells depleted of SKIV2L2 may have indicate changes in gene expression that are dependent upon either degradation of an RNA that influences mRNA expression, or a failure to participate in the turnover of a small subset of mRNAs.

C. SKIV2L2 subunit targets a wider range of ncRNAs than in yeast

In the previous section, we performed a genome wide analysis to evaluate the effect of SKIV2L2 knockdown on ncRNA abundance. This analysis requires annotation and only showed a highly extracted summary of the sequencing data. To probe deeper and gain a more detailed picture of regions across the genome that exhibit differential RNA expression it was decided to perform an annotation independent analysis to highlight all regions of the genome. After evaluating various free toolkits such as Cufflinks, Crossbow, EdgeR, DESeq and MEDIPS, it was decided to proceed with MEDIPS for its high-performance and ease of operation.

We further performed genome-wide differential polyadenylation profiling with MEDIPS to identify genomic regions that show a statistically significant increase in the accumulation of reads across the genome. MEDIPS is a statistical analysis tool designed for Chip-Seq analysis and can be used for identifying differential gene expression [45]. For each sample, the genome was divided into a series of 100 bp regions, and the number of reads were counted within each region of the genome. A statistical calculation was performed for each defined region to identify those regions that show differences in RNA abundance between the control and SKIV2L2 knockdown. All four samples (Control Knockdown original and replicate, SKIV2L2 knockdown original and replicate) were used to increase statistical confidence. Parameters were set to default unless otherwise noted. With the 100bp window and a 90% probablility (P<= 0.1), we identified eleven genomic regions that exhibited statistically significant increases in the abundance of reads in SKIV2L2 knockdown when compared to the Control knockdown. The result of MEDIPS analysis is listed in table II-2:

Regions that have more reads in SKIV2L2 knockdown

Gene symbol	Location	Adjusted p-value	Type
Gas5	chr1:161033166-161040537	0.0592	lncRNA
Vgf	chr5:137032601-137032700	0.0031	mRNA
Rpsa	chr9:120129101-120129200	0.0123	mRNA
Mir-138-1	chr9:122682701-122682800	0.0512	miRNA
Unannotated	chr10:7977001-7977100	0.0023	NA
Tex14 (U3)	chr11:87443601-87443700	0.0382	snoRNA
Mir-17	chr14:115043414-115043783	0.0059	miRNA
Unannotated	chr15:85702925-85703233	0.0640	NA
let7c-2	chr15:85706501-85706600	0.0015	miRNA
let7-b	chr15:85707201-85707300	0.0088	miRNA
Rn45s	chr17:39843001-39843100	0.0008	rRNA

Table II-2 Result of MEDIPS analysis on Skiv2l2KD’s differential polyadenylated regions of MEDIPS analysis. The MEDIPS reported 11 genes (regions) as significant under 0.1 false discovery rate.

One prominent target in the list is Growth Arrest-Specific 5 (GAS5) gene at chr1:161033166-161040537 with p-value 0.0592 encompassing two positions from this gene (Fig. II-3). One of the peaks is located at the 3′ end of the Gas5 full transcript (Fig. II-3). The 3′ end of sequencing reads aligns neatly to the last nucleotide of the Gas5 transcript, suggesting that these reads may represent a processed fragment at the 3′ end of the transcript destined for degradation, or more likely this accumulation of reads in both samples represent the full length transcript of Gas5 that has been polyadenylated by the canonical polyA polymerase Pap1p. The abundance of reads in the two samples from this region is 855 Control knockdown, 584 SKIV2L2 knockdown, higher than overall genomic level. The Control replicate sample have 855 reads, higher than the SKIV2L22 which have 584. This difference is unconventional since most transcripts affected by SKIV2L2 knockdown accumulates in SKIV2L2 knockdown sample, not in Control. In spite of this, the analysis provides us insight into the expression level of this Gas5 gene, and that SKIV2L2 controls the level of Gas5 transcript. Gas5 is a long-noncoding RNA that accumulates in cell growth arrest and plays a role in apoptosis and prostate cancer [46, 47]. It is also known as a host gene for several snoRNAs that reside in 9 out of the 12 introns (U47, U74, U75, U76, U77, U78, U79, U80 and U81) [48]. The second differential peak of reads is located at 580bp downstream of the annotated Gas5 transcription start site (chr1:161,035,750), which is the 3′ end of the first intron that hosts the putative snoRNA U74 (Fig. II-4). In SKIV2L22 sample, the majority of reads aligns to the intron end, and the rest aligns upstream by 1 to 30 nucleotides. The length of the read pairs range from 100 to 300 bp, covering from the 3′ end of the intron to two third of the intron (Fig. II-5). The putative snoRNA is within the range of most of the reads in SKIV2L22 sample. This is a strong evidence suggesting that these reads comes from cleaved introns, rather than the snoRNA processing by-product. The SKIV2L2 knockdown 2 has 45 reads that is much greater than 1 read in Control knockdown2 sample. The difference in read abundance indicates that SKIV2L2 efficiently degrades this Gas5-derived transcript. The distribution and abundance information implies the regulation function of SKIV2L2 on snoRNA maturation and processing.

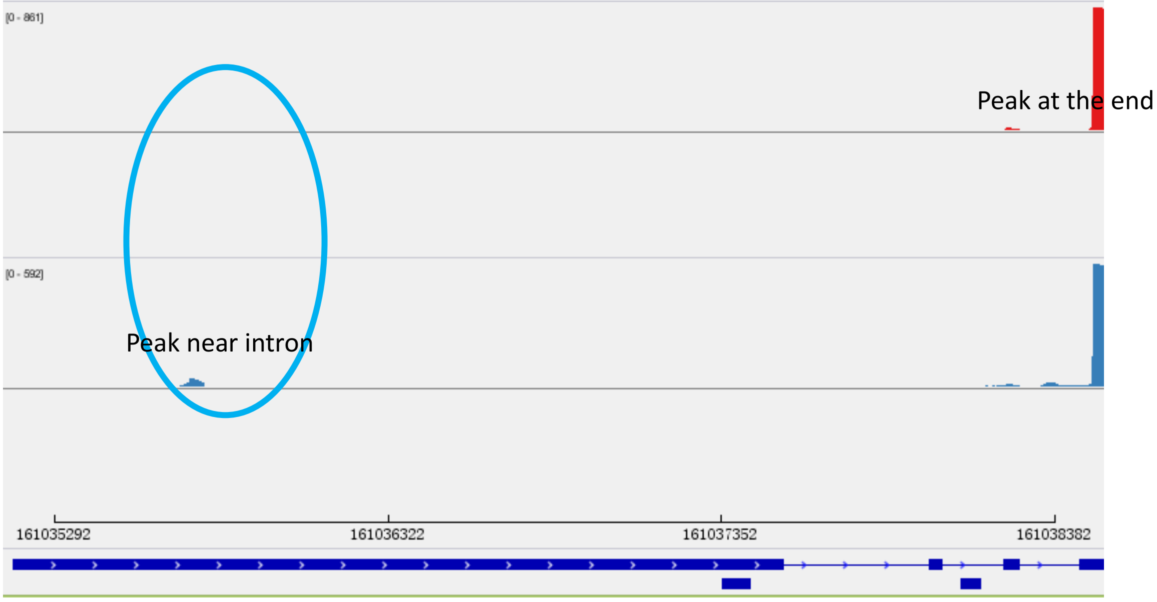

Figure II-3 The location and abundance of the two peaks found in Gas5. Top track: Control knockdown replicate. Bottom track: SKIV2L2 knockdown. The peak at the end of Gas5 transcript (Referred as “first peak”) have high expression abundance. The peak at the end of the first intron is relatively lower and less consistent in its location. This peak also showed a greater abundance in SKIV2L2 knockdown compared to control knockdown.

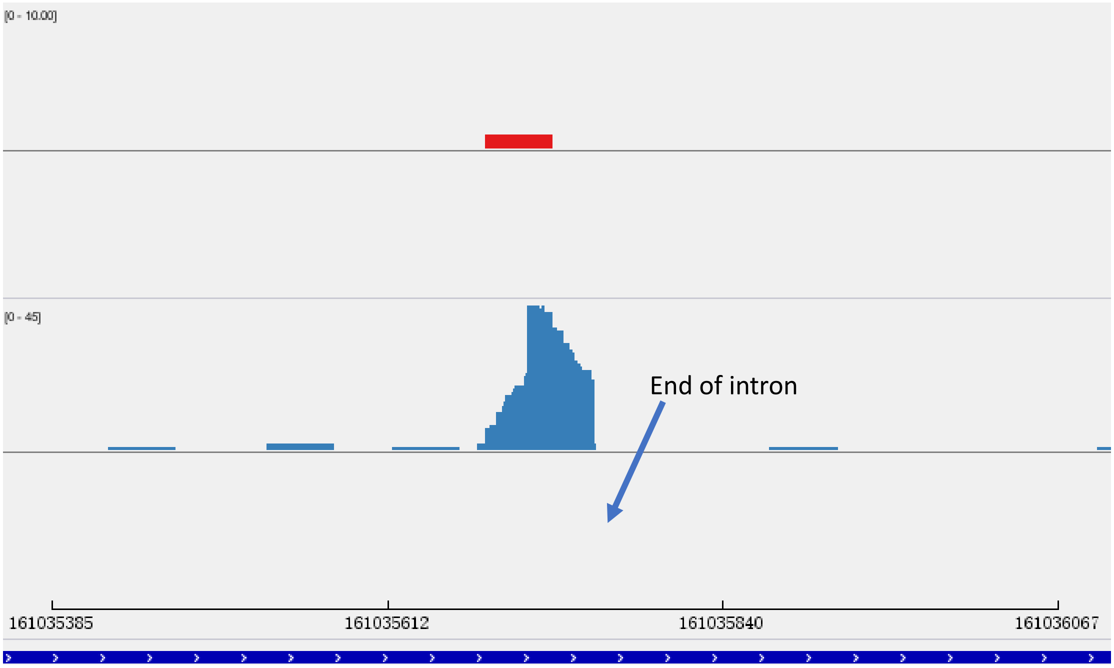

Fig. II-4 Location and abundance of the two peaks found in Gas5. Top track: Control knockdown replicate. Bottom track: SKIV2L2 knockdown. The peak is located at the end of 1st intron (as shown by the arrow). Large amounts of reads accumulates in SKIV2L2 knockdown compared to control.

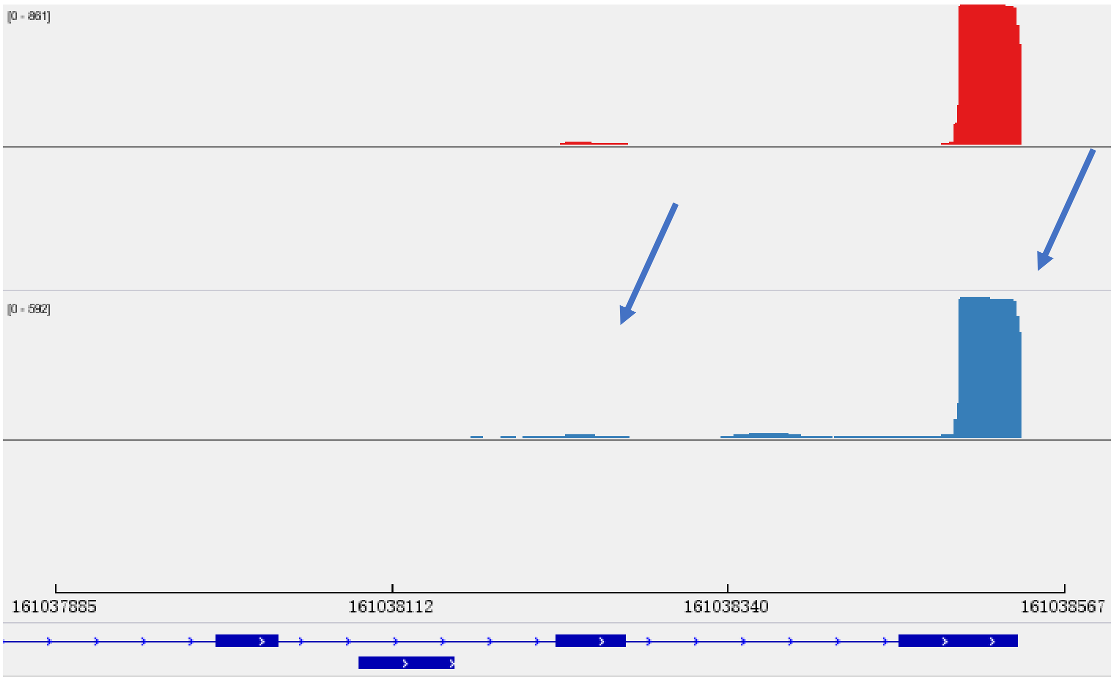

Fig. II-5 Location and abundance of the two peaks found in Gas5. Top track: Control knockdown replicate. Bottom track: SKIV2L2 knockdown. Second Peak have slightly greater reads in Control knockdown. All 3′ reads consistently aligned to the end of the Gas5 transcript.

Similar pattern is also found in RPSA gene. We identified differential polyadenylation peaks at the end of the second intron in RPSA gene (Fig. II-6). RPSA is a ribosomal protein coding gene whose product is involved in the assembly of the 40S ribosomal subunit, and in the maturation of the 20S rRNA precursor [49]. RPSA also works as a cell surface receptor for laminin, functions in cell adhesion and signal transduction [49]. The peak of different abundance locates at the end of the second intron (chr9:120,129,151) with the ratio of 1: 20 (Control knockdown: SKIV2L2 knockdown). All of the reads in this peak aligned to the 3′ end of intron 2, and the read pairs ranged from 100 to 250 bp, coverd upstream up to 300 bp from this point (Fig. II-7). This intron contained snoRNA6 in human and was predicted to contain its homolog in mouse [50, 51]. As in Gas5, the location and abundance of reads suggests that SKIV2L2 is involved in the metabolism of snoRNA, or at least the degradation of processing by-product. The next peak resides within the third intron (chr9:120,129,553), coincidentally have the same coordinate with the putative snoRNA 73 (Fig. II-8). The read abundance in SKIV2L2 knockdown is slightly greater than Control. It is difficult to determine whether this difference is caused by SKIV2L2's regulation or just expression fluctuation by itself. However when taking other snoRNA spots into consideration, this event is possibly caused by SKIV2L2 knockdown. The third peak is located within the fifth intron (chr9:120,130,566), at the 3′ end of snora62 (chr9:120,130,572) (Fig. VI-7). Though statistically this peak is not considered different, based on the distribution of reads it is safe to consider it as 'real' and can still provide some insight into the processing events in this intron. The

SKIV2L2 knockdown sample have 6 reads while the Control knockdown have 0, suggests that the snoRNA62 is very likely to be a target of SKIV2L2.

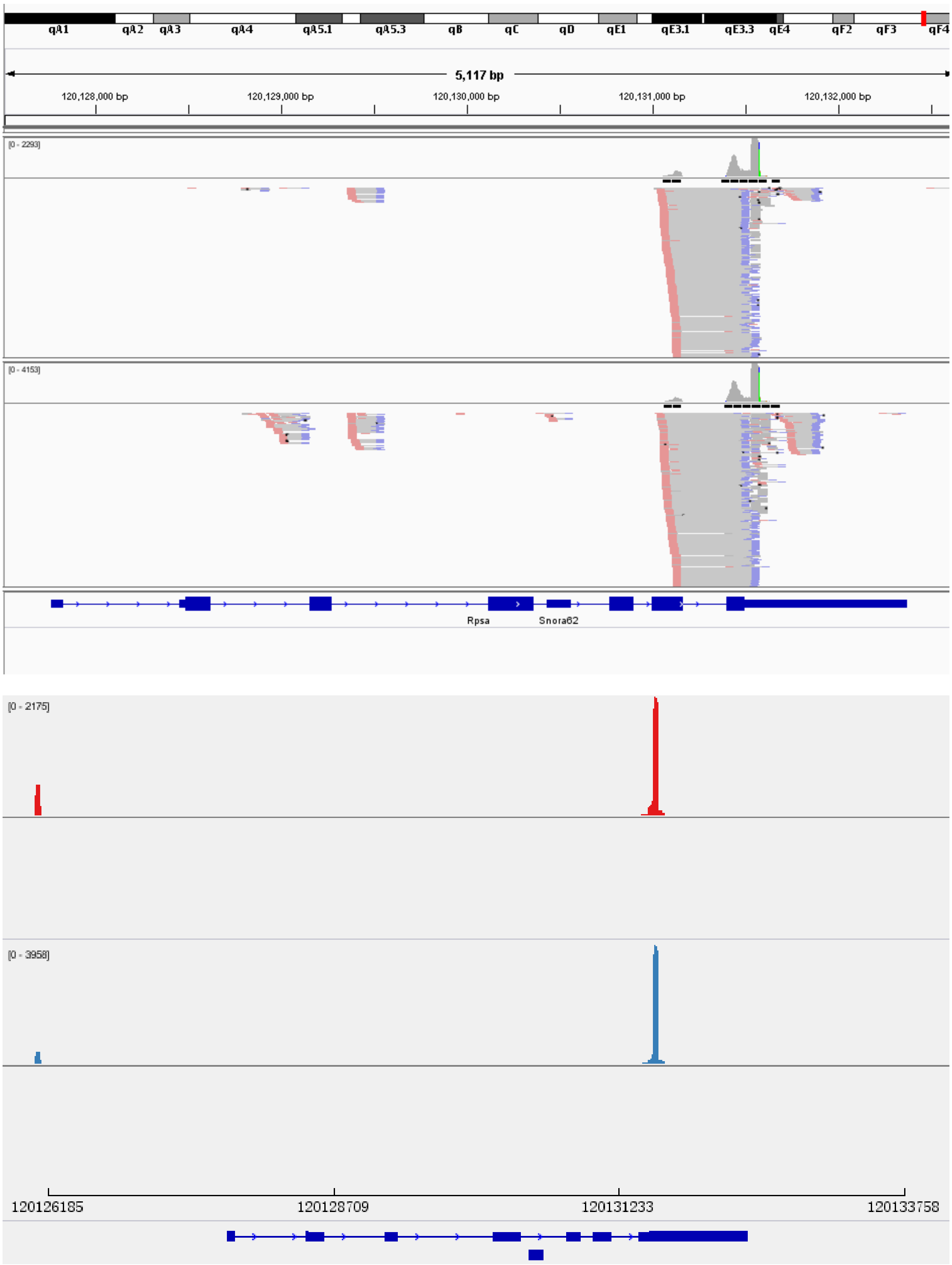

Figure II-6 Overall view polyadenylation status of RPSA. Top figure: Sequencing read view; Bottom figure: read density view. In both view, the top track shows control knockdown, the bottom track shows SKIV2L2 knockdown. This convention applies to all figures in this article. Three major peaks were found in *RpsA* gene.

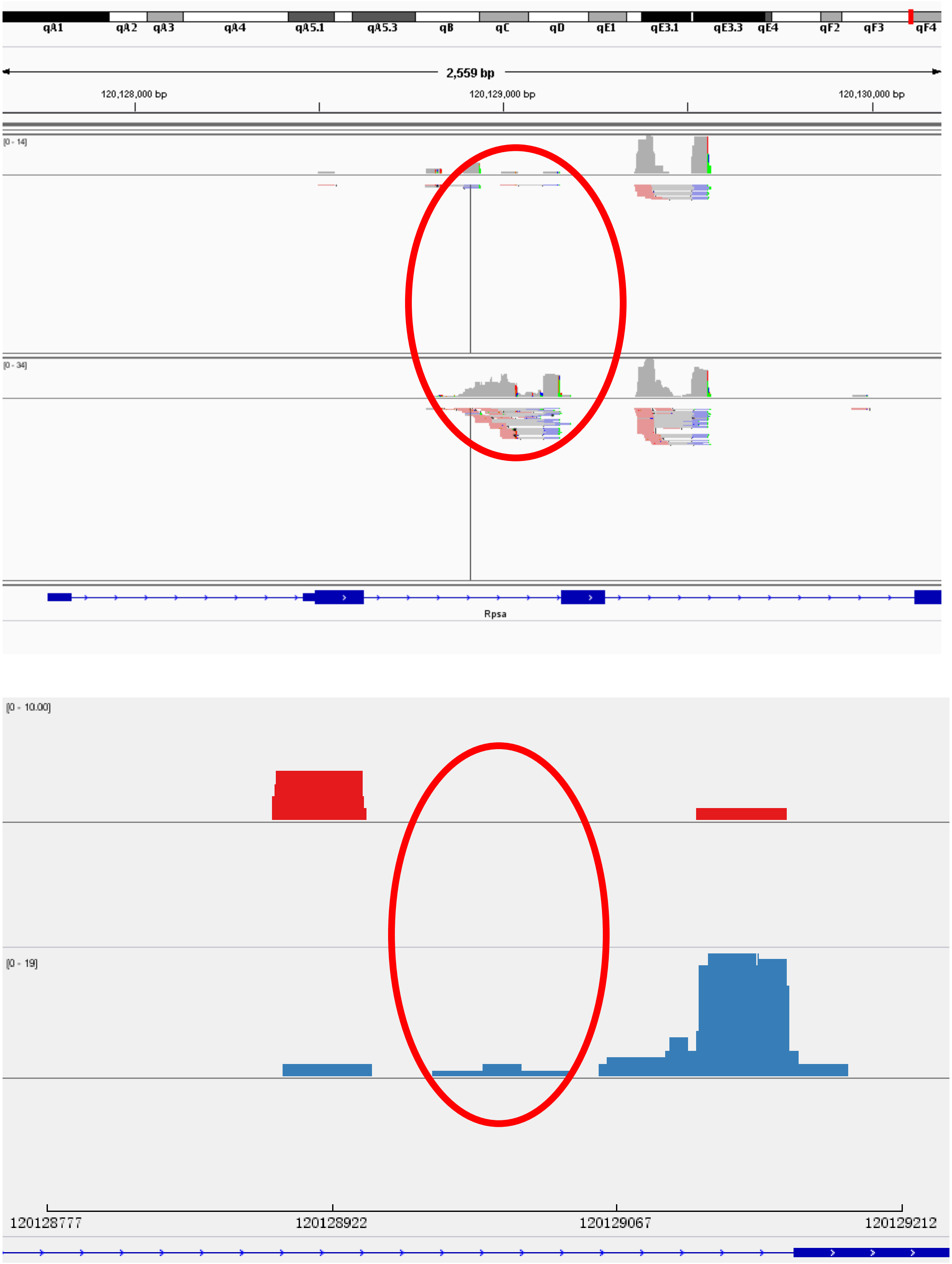

Figure II-7 Differential reads found at the second intron of RPSA, this intron contains putative snoRNA6. Reads accumulates in SKIV2L2 knockdown significantly. The red circle indicates the differential snoRNA6 hosting intron.

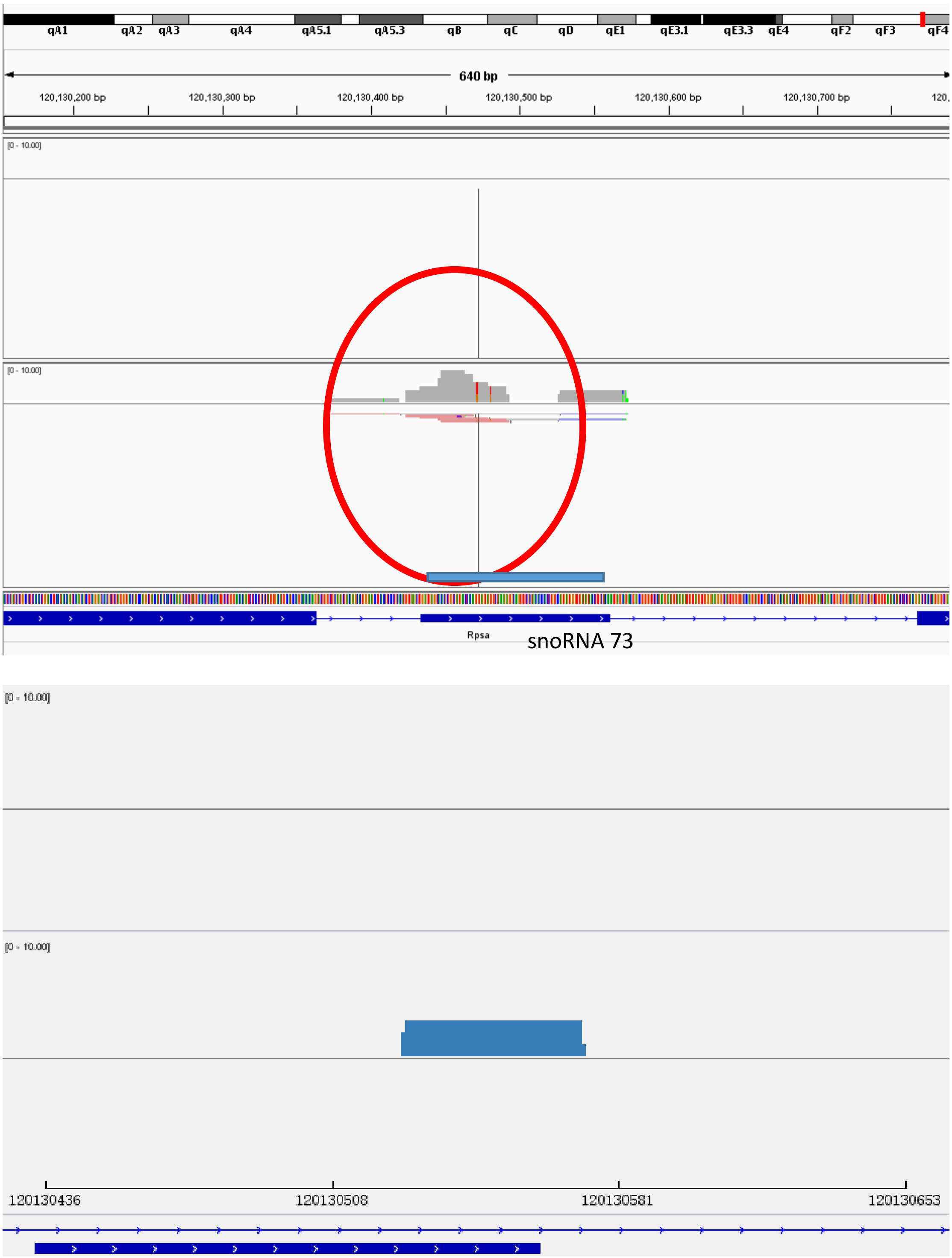

Figure II-8 Reads near snoRNA73. The polyadenylation level is relatively low in both samples, but traces of reads can still be detected at the snoRNA73 and the SKIV2L2 knockdown had more reads than control knockdown.

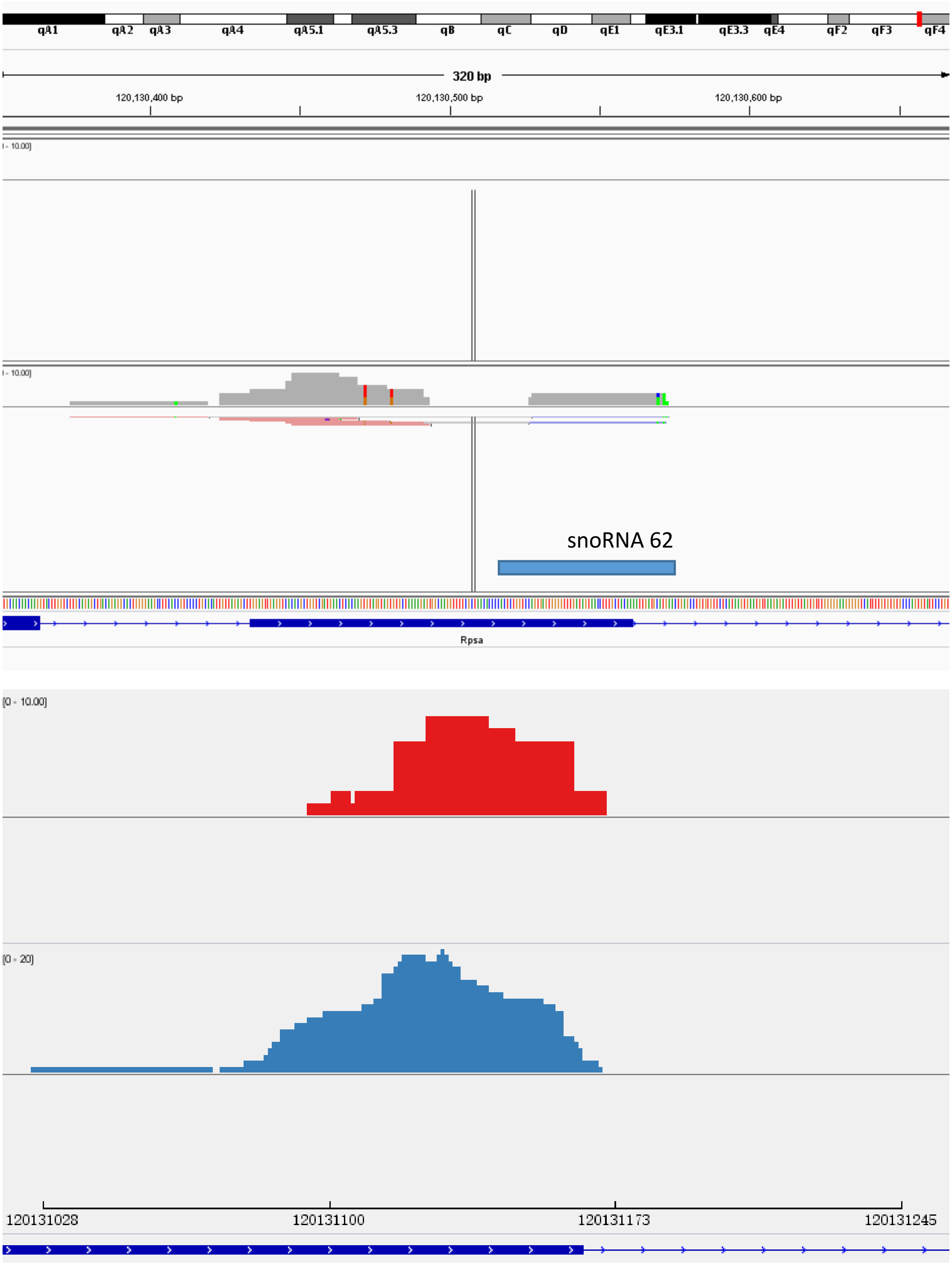

Figure II-9 Reads near snoRNA 62. Similar to snoRNA73, read abundance is low in SKIV2L2 knockdown and no reads had been found in control.

Another identified snoRNA-containing target was Tex14 (Testis Expressed 14) (Fig II-10). Tex14 is required for spermatogenesis, and more importantly is the host gene of snoRNA U3. The identified region (chr11:87443601-87443700) had 0 read pairs in Control knockdown and 3 read pairs in SKIV2L2 knockdown. The identity of this transcript remains unclear due to limited annotation. We previously reported that snoRNA U3 is processed by SKIV2L2 in mammalian cells [26]. This analysis confirmed our discovery and brought in more information on SKIV2L2 and snoRNA. All three copies of U3 snoRNA that reside in Tex14 (chr11:87462286-87462500, chr11:87471368-87471582, chr11:96032678-96032892) have been found in our sample (Fig. II-11, 12, 13). The first two share similar patterns of read distribution and abundance. All the 3′ end of 3′ reads align exactly to the 3′ end of U3, and the 5′ read pair have a varying location from 40 to 200 upstream of the 3′ end. The longest read pair fits to the full length of U3 snoRNA (Fig II-11). In both loci, the SKIV2L2 knockdown have slightly greater abundance than Control knockdown. The third locus is different in that the read is not paired, but only have the 5′ piece. Even though, the SKIV2L2 knockdown still have more reads in the third locus. We believe that these U3 spots as valid SKIV2L2 target because of the read distribution pattern and experiment verification. More importantly, the U3 snoRNA read distribution pattern serves as a model for identifying other snoRNAs that are processed by SKIV2L2.

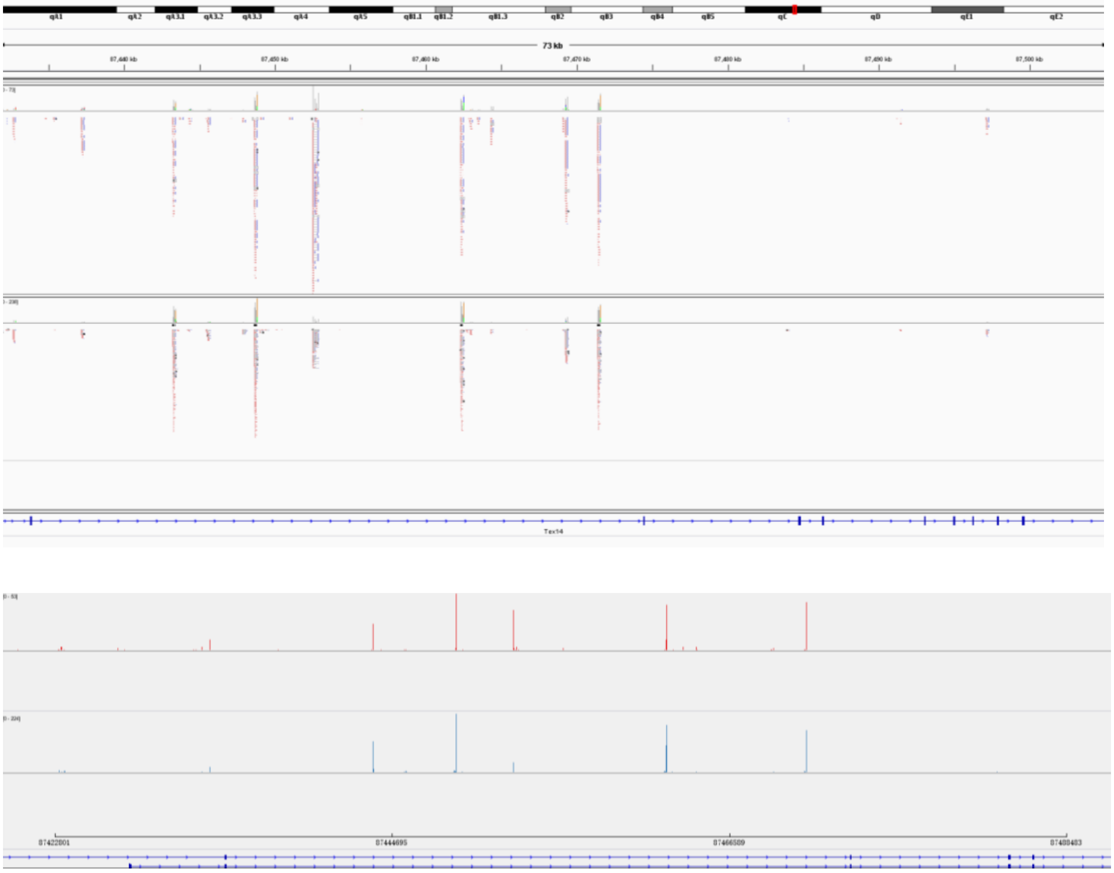

Figure II-10 Global overview of Tex14 gene. This gene is 14kb long that the browser can barely display its full length. 8 peaks were found in Tex14 and most of them have equal abundance. Tex14 also hosts 3 copies of the snoRNA U3.

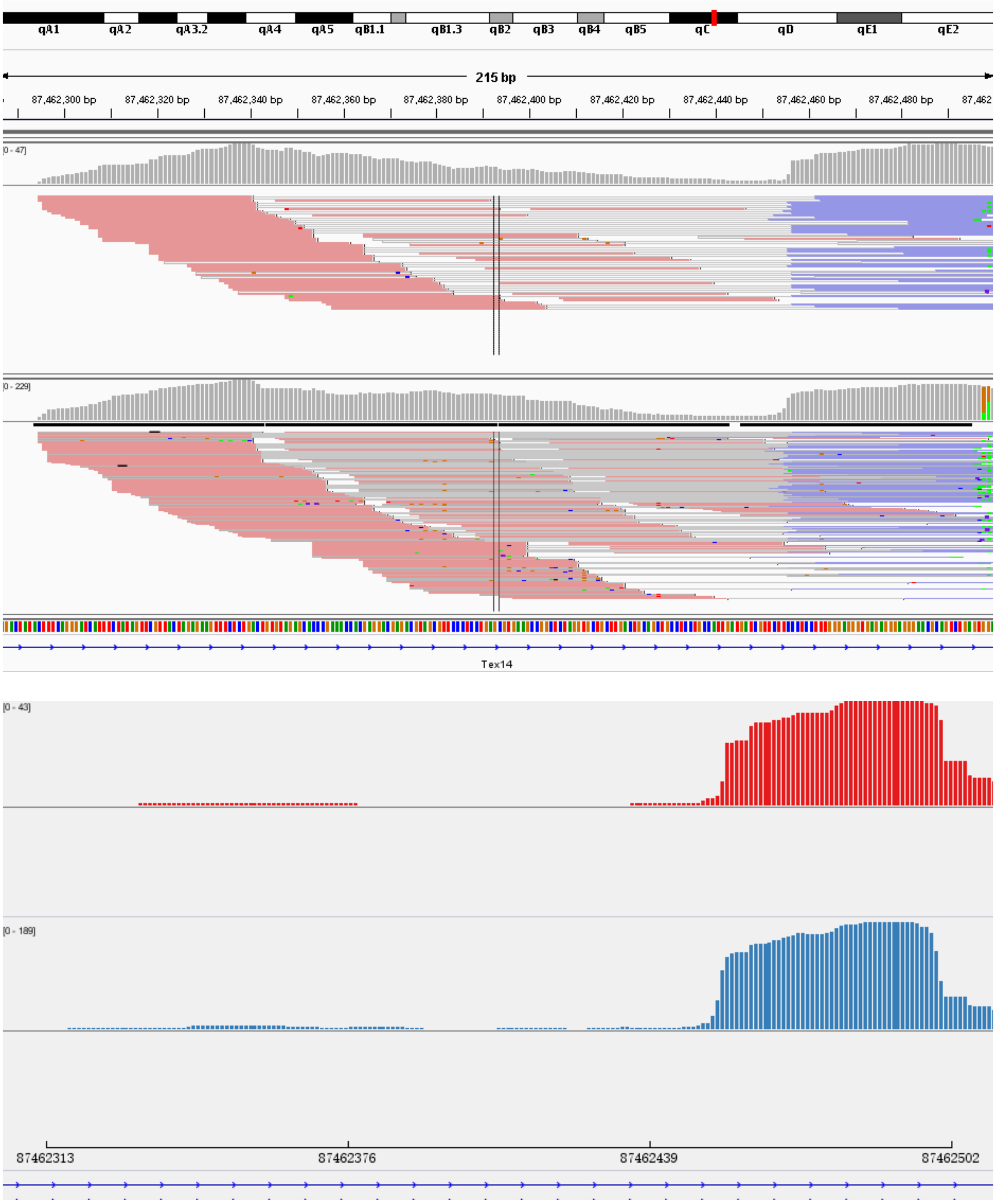

Figure II-11 The first U3 copy have rich expression and different abundance. The number of reads in control knockdown is less than SKIV2L2 knockdown. The bottom density graph showed a similar density due to different scale (see upper left corner of each lane).

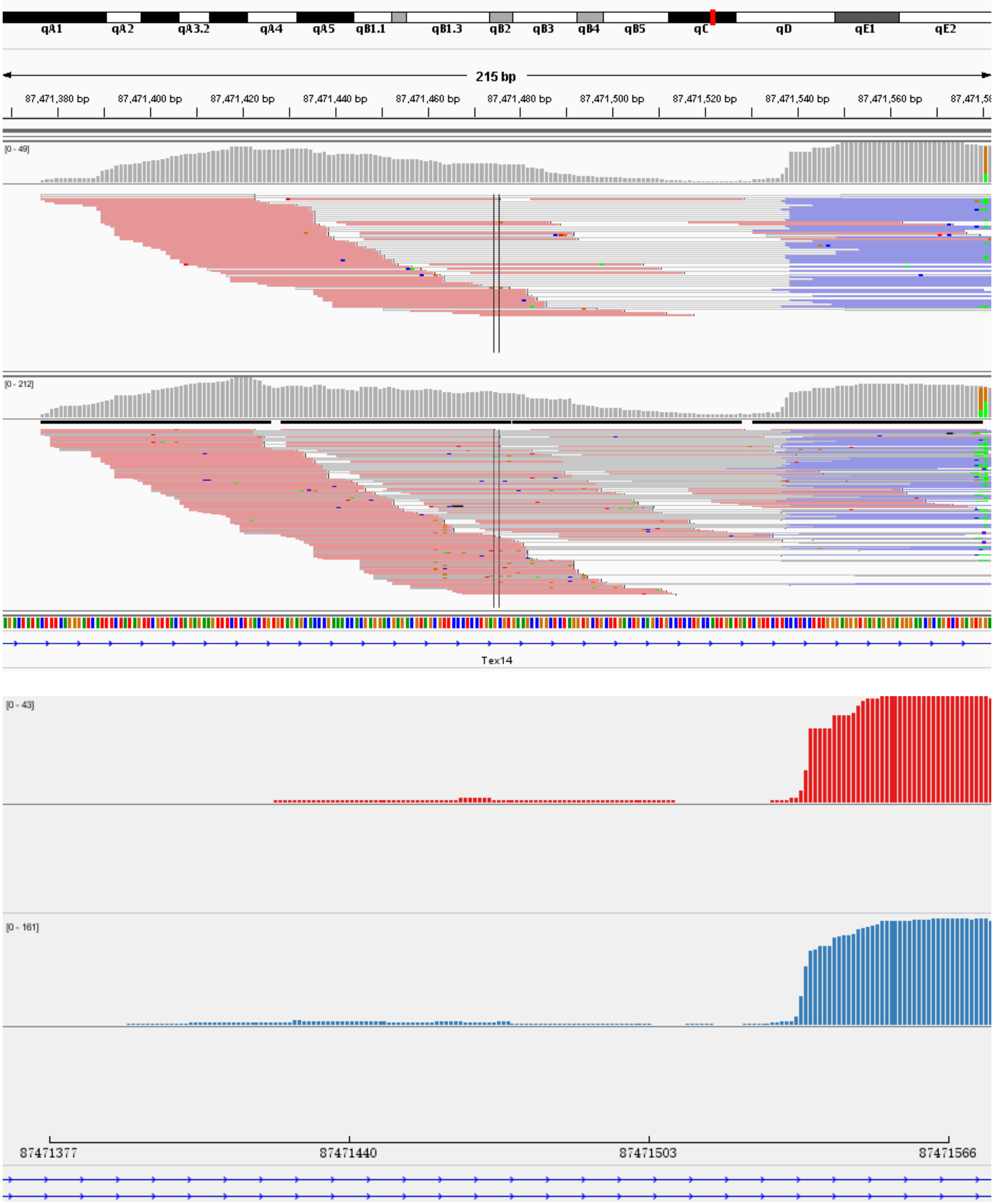

Figure II-12 The first U3 copy have rich expression and different abundance. The number of reads in control knockdown is less than SKIV2L2 knockdown.

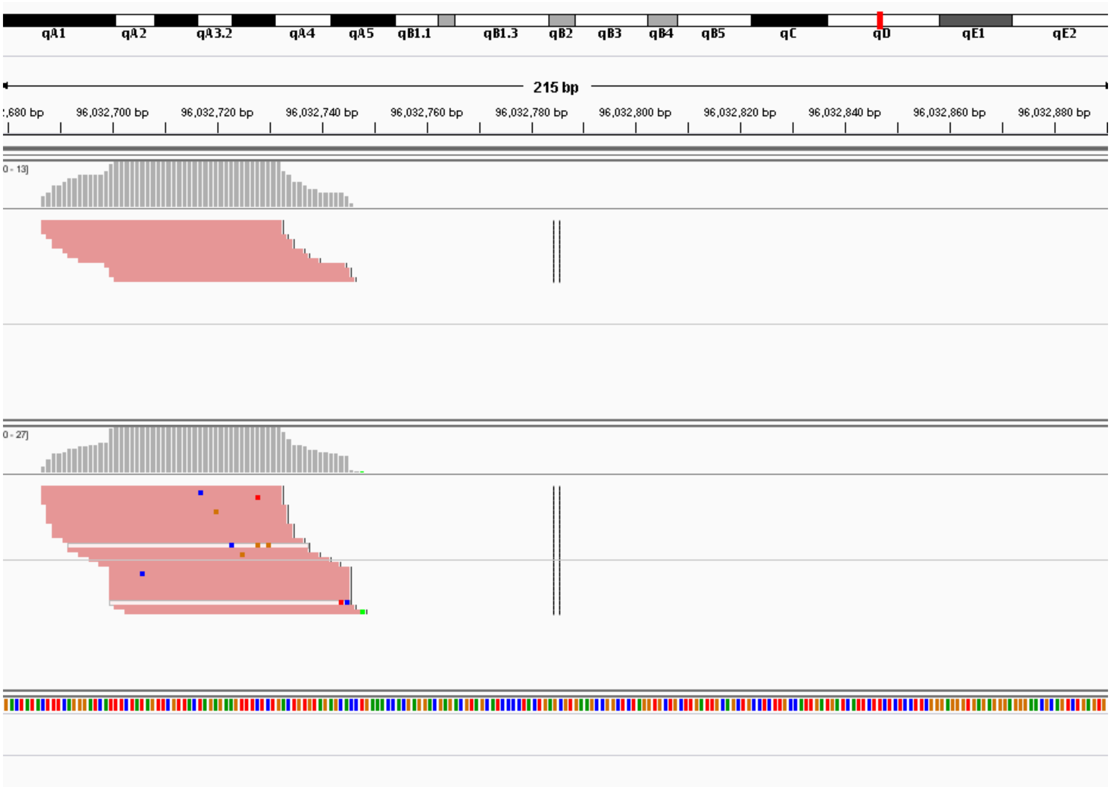

Figure II-13 The third copy of U3 are poorly expressed compared to other U3 copies in Tex14. This U3 copy showed a different polyadenylation pattern compared to others.

Several miRNAs were found as target in MEDIPS analysis. miR-138-1, mir-17, let7c-2 and let7b were identified as differentially polyadenylated with high confident level (Fig VI-11, 12, 13, 14). These miRNAs roughly shared a similar reads distribution and abundance pattern. The 3′end of 3′ read pair aligned consistently to the nucleotide 20bp downstream of the pre-miRNA’s 5′ end, which is the Drosha cleavage site. The SKIV2L2 knockdown sample had greater number of reads than Control knockdown. These two patterns were found in all four miRNAs, which inspiered us to conduct further investigation on whether all miRNA’s 5′ leader sequence is targeted by SKIV2L2.

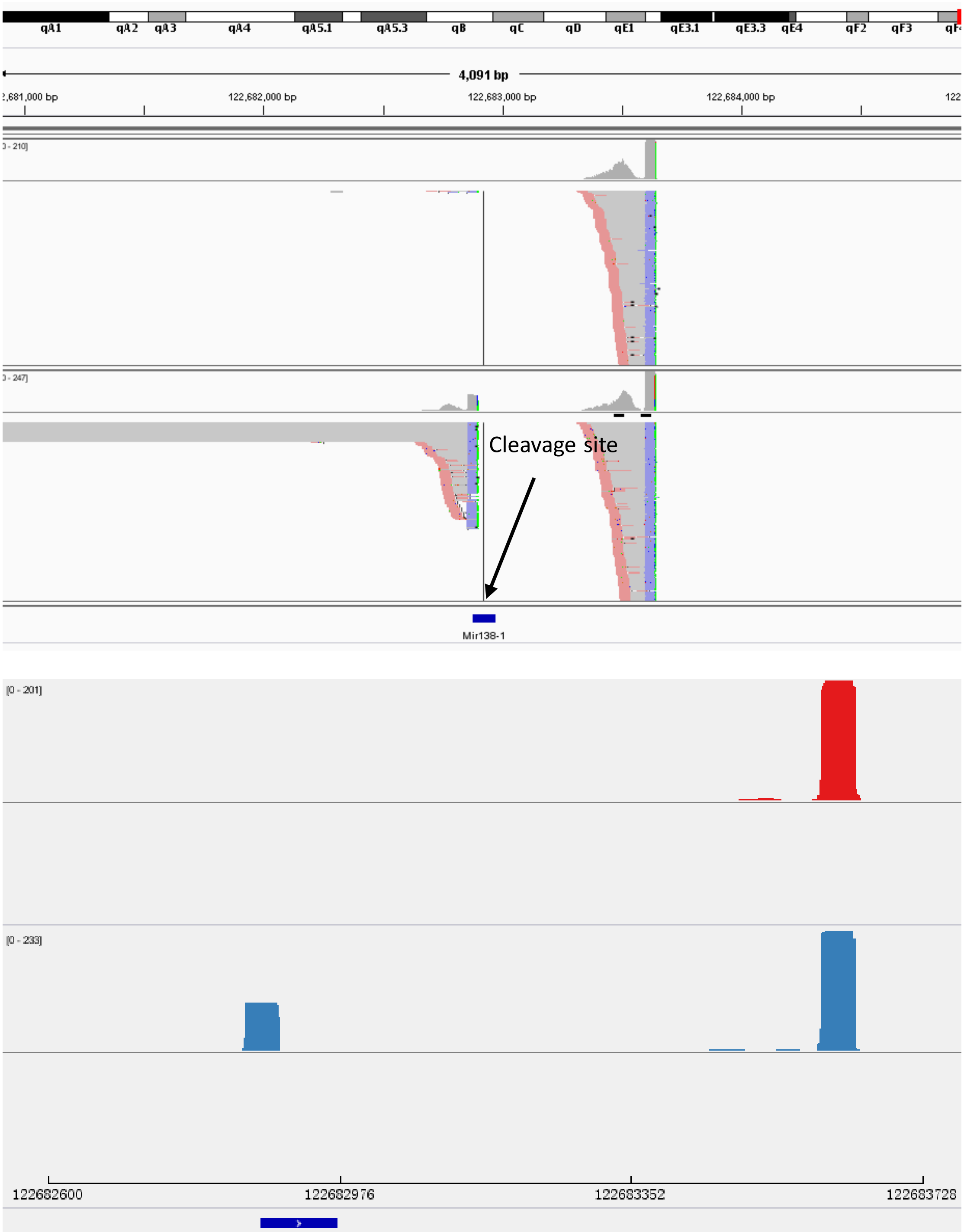

Figure II-13 Read distribution and abundance near mir-138-1. From left to right, the first group of reads represent the 5′ leader sequence of miR-138. The second group represent the end of the host transcript. Large amount of reads accumulate in SKIV2L2 knockdown at the Drosha cleavage site.

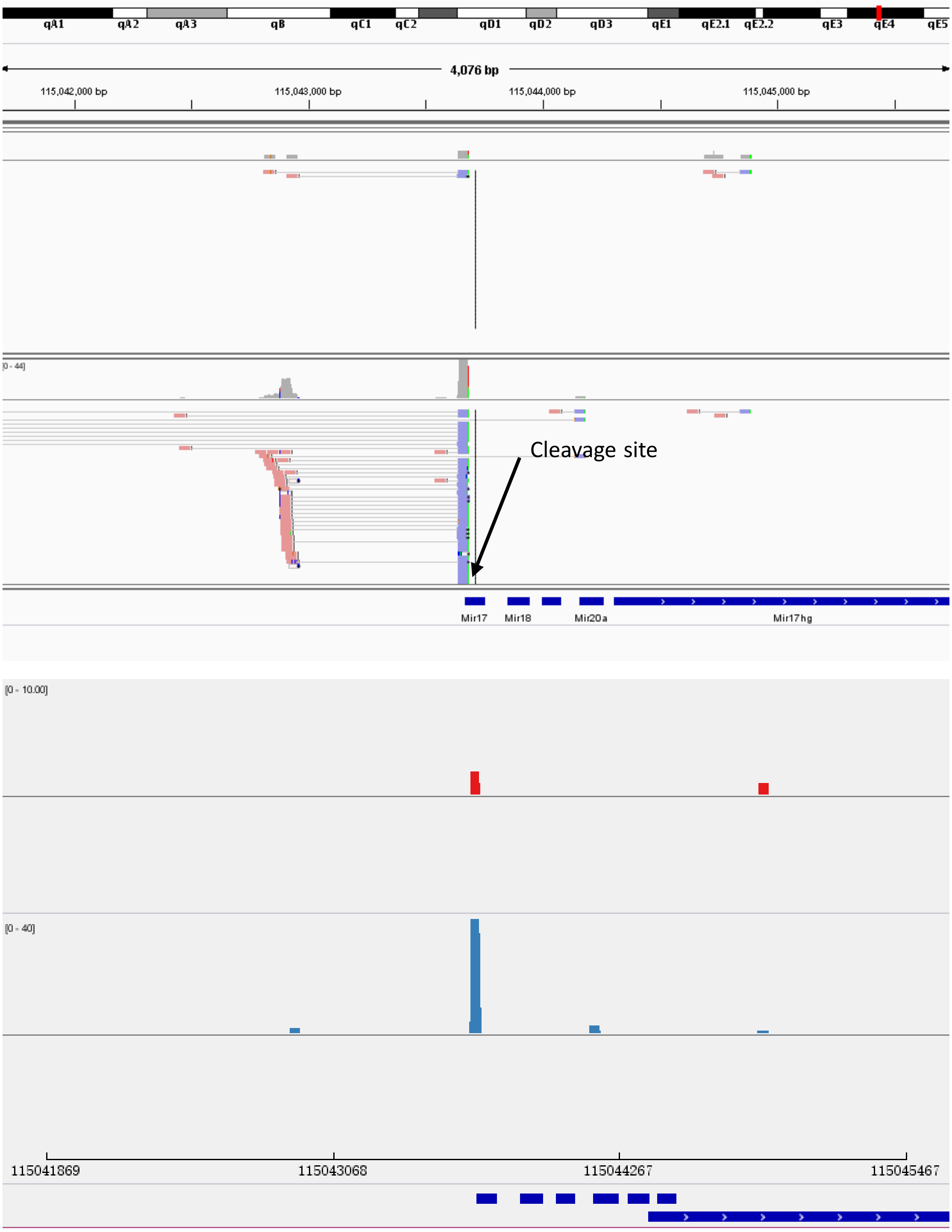

Figure II-14 Read distribution and abundance near mir-17, which have a typical read distribution of a SKIV2L2 mediated miRNA. The Drosha accumulation is shown at the center of the view as pointed.

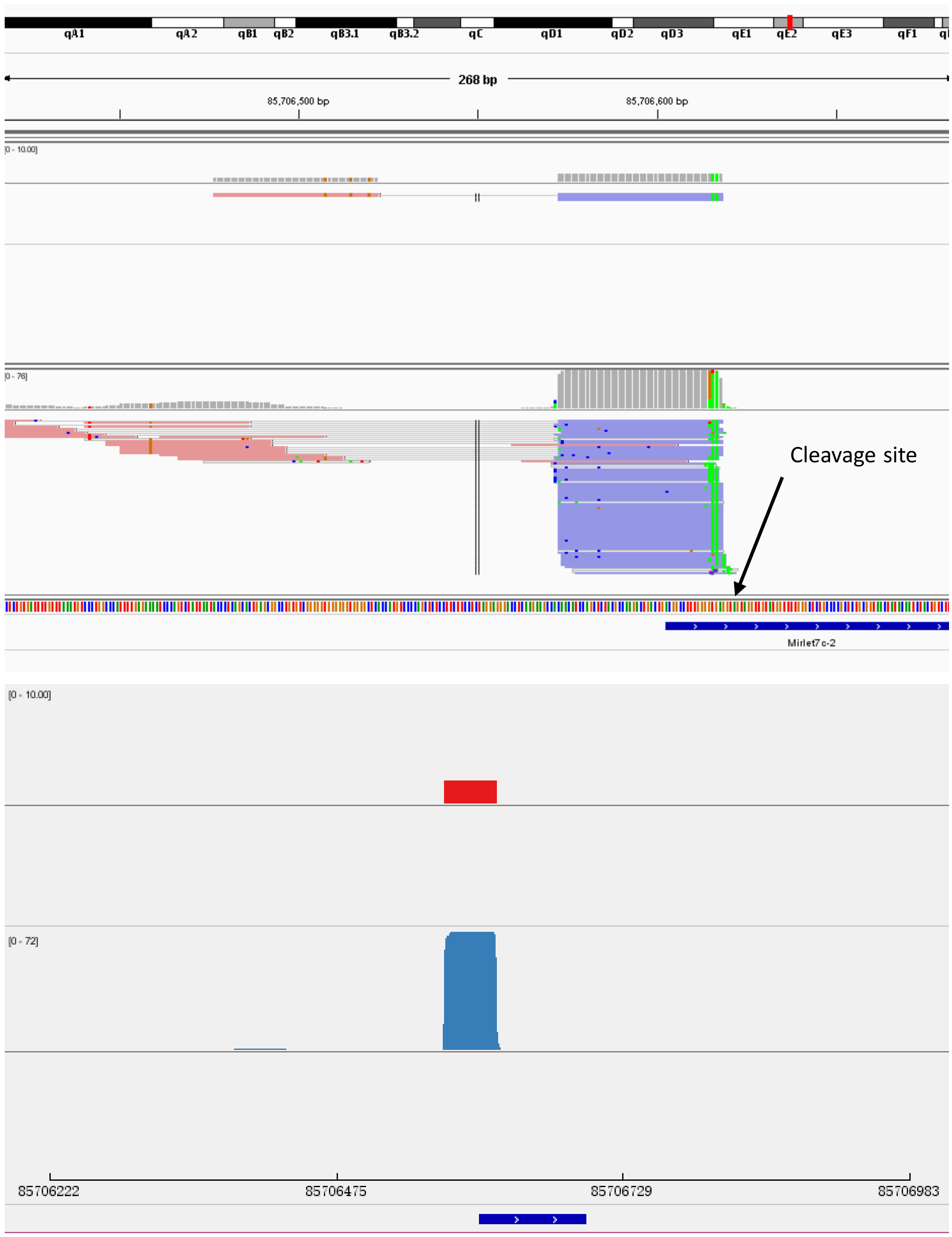

Figure II-15 Read distribution and abundance near let7c-2. This miRNA demonstrated a similar pattern to miR-138. 5' leader sequence was found for this miRNA.

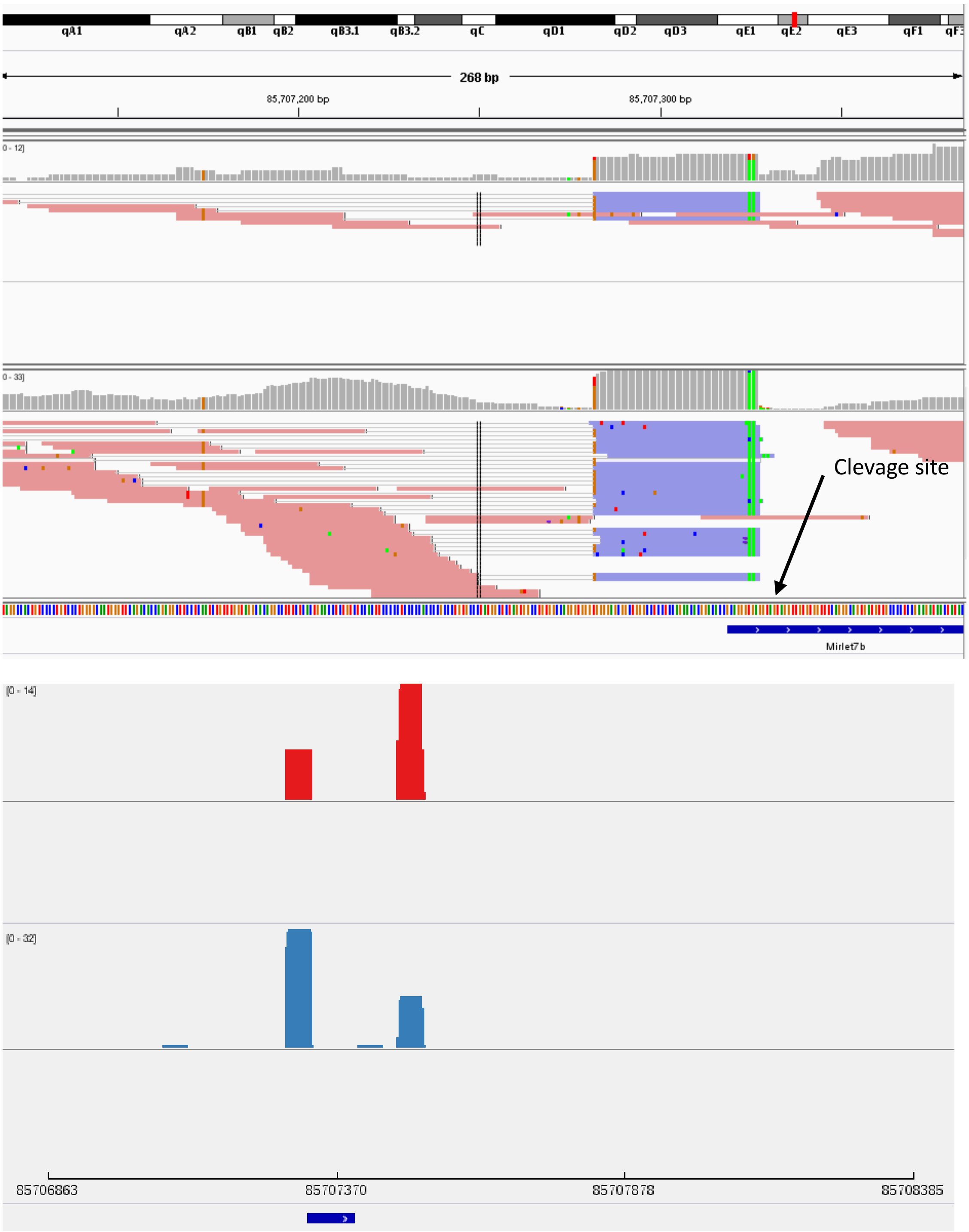

Figure II-16 Read distribution and abundance near let7b. This miRNA demonstrated a similar pattern to miR-138. 5’ leader sequence was found for this miRNA.

A special target is RN45S gene, a copy of the ribosomal DNA. There are 30 to 40 copies of rDNA spreading along the genome, and the RN45S gene is a representation of this rDNA gene in mm10 assembly. Also because of duplication, all RN45S rRNA was mapped to this single loci, producing a very high expression level and density (Fig. II-17). It's hard to interpret the read distribution and abundance by just reading the highly intensive alignment graph. To address this, we introduced a density estimation tool, F-seq, to extract and summarize this data for easier interpretation and analysis (Figure II-17) .

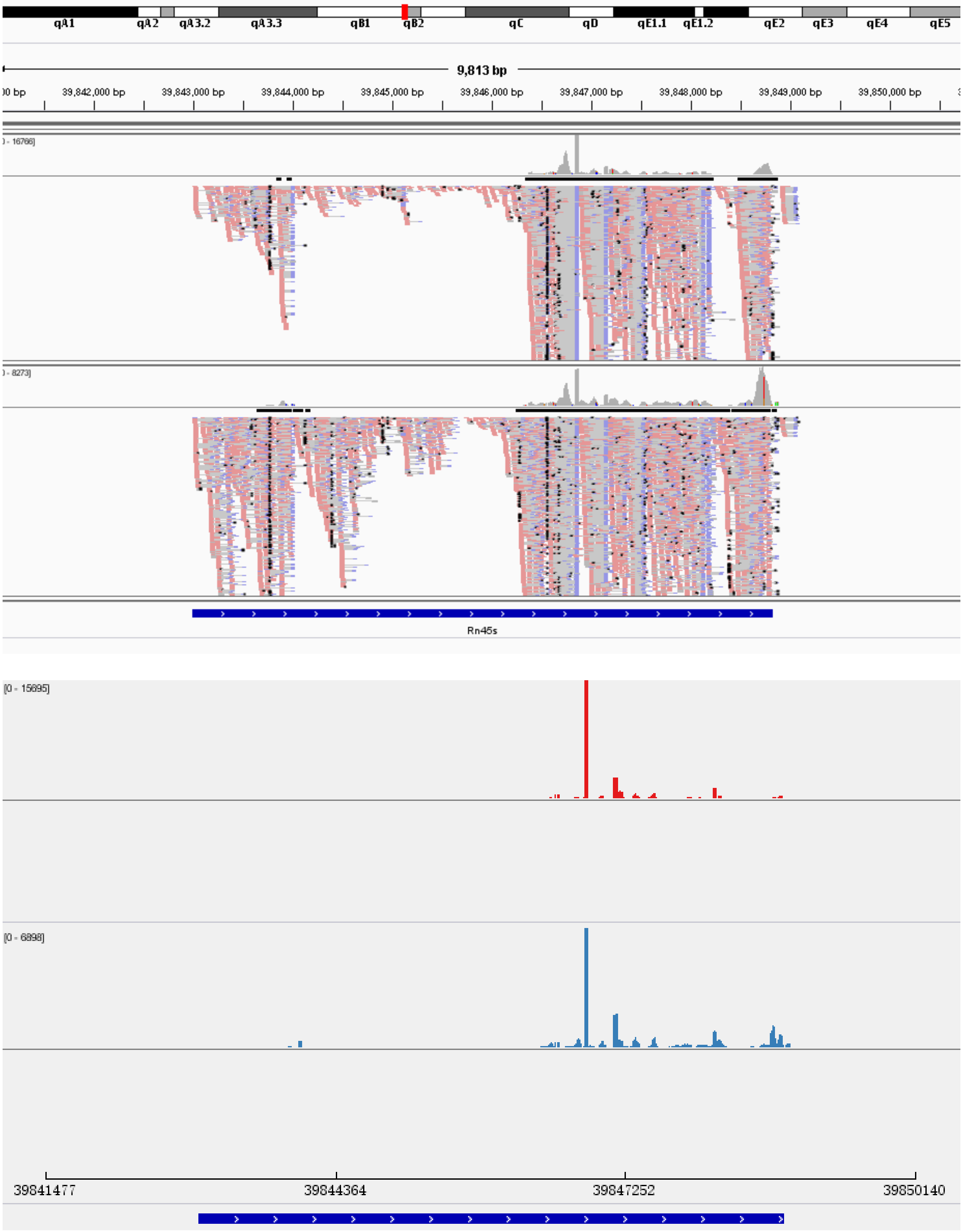

Figure II-17 Read distribution and abundance near RN45S. This graph shows all the rRNA reads that mapped to RN45s. Due to the high copy number the read looks very dense and is unreadable. The bottom density graph showed a summarized view.

Lastly, we found differential peaks in VGF gene (chr5:137032601-137032700). Nerve growth factor inducible (VGF) is a pro-protein from which a few neuropeptides are derived from that regulate energy homeostasis and nutrition [52]. PolyA RNAs were identified near the annotated transcript 3′-end as well as in the last of total 8 exons. The amount of pA-seq reads at or near the 3′end of the transcript did not vary significantly between control and SKIV2L2 knockdown, suggesting that the primary transcript of VGF is not affected by SKIV2L2 knockdown. The differential polyadenylated cluster lay in the middle of the transcript, at an exon. As with other two loci, it is unclear that what kind of transcript it represents, and what role SKIV2L2 plays in regulating this transcript.

MEDIPS profiling analysis revealed the complexity and variety of SKIV2L2's targets. mRNA, miRNA, snoRNA, rRNA and their processing derivatives were found to be processed or degraded by SKIV2L2. The diversified read-pair distribution in differential polyadenylated loci implied the different mechanism with which SKIV2L2 participates these processing events. This analysis provides us a high-level glimpse of the data and pointed out the most critical spots that needs attention.

D.Ribosomal RNA is subject to processing by mTRAMP

In MEDIPS analysis, the 45S Ribosomal RNA transcript region was identified as a target of SKIV2L2. A fine mapping is needed to articulate the processing events of the 45S rRNA and the role SKIV2L2 played in these processes. The 14kb rRNA primary transcript sequence serves as a genome assembly, and then the polyA-Seq data was mapped against it. Due to the high copy number of rRNA, the number of reads in this region was too high and covers every region of the rRNA primary transcript. A peak-calling step is required to summarize the reads abundance information to the intensity of polyadenylation signals.

We used F-Seq as an abundance density estimator to summarize the high-intensity data and generated a polyadenylation signal intensity graph (Fig. II-18).

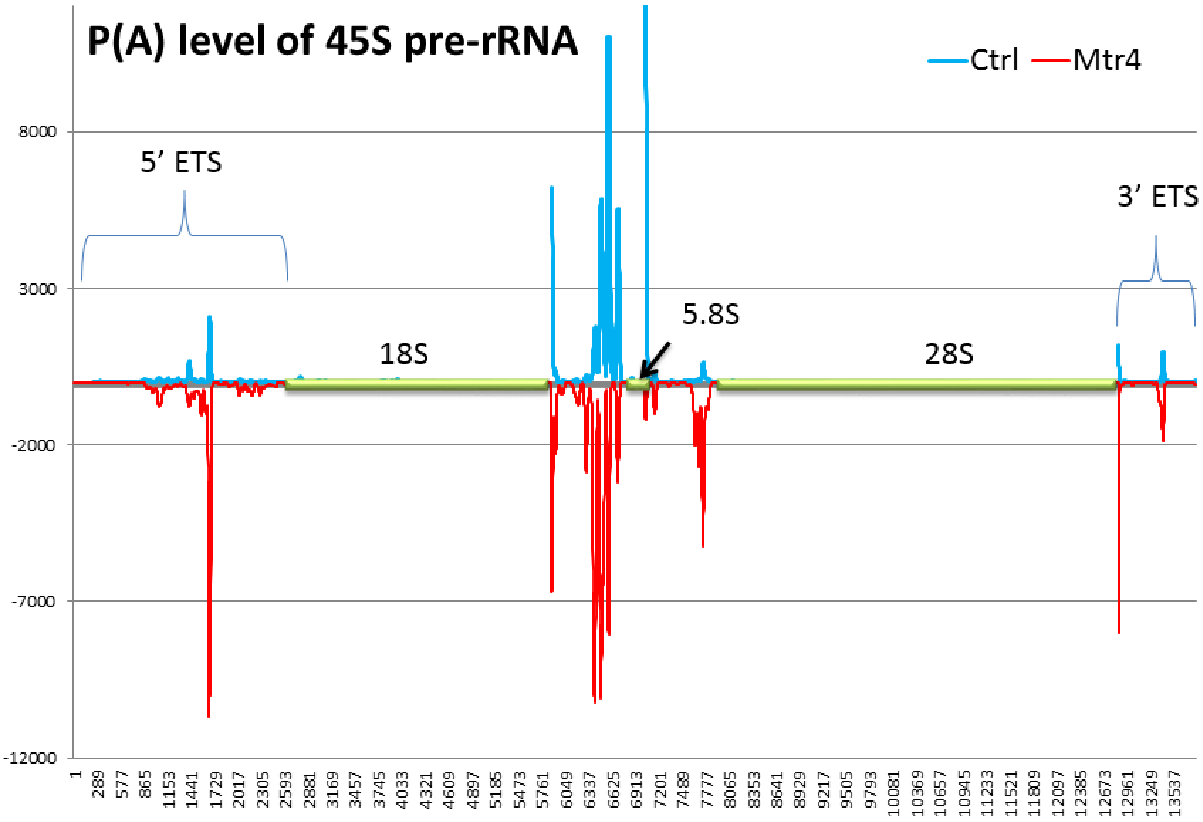

Figure II-18 Polyadenylation peaks in primary rRNA region. Reads inside the rRNA regions were removed. Blue indicates Control. Red, Skiv2l2 knockdown. Height of peaks reflect the polyadenylation intensity. The peak in 5' external transcribed spacer had a huge peak at 1600bp. This suggests that SKIV2L2 is responsible for the processin g of the A'-A_0 fragment. The one at the internal transcribed spacer is the intermediate of 5.8S rRNA called 12S rRNA. The processing of 12S was blocked by the SKIV2L2 knockdown and caused the accumulation of 12S.

Seven strong polyadenylation peaks were found in primary rRNA transcript. The one in the 5′ ETS is located at about 1600bp from the 5′ end, and the abundance in SKIV2L2 knockdown is almost five times greater than control. The location of this peak matches the A0 cleavage site that is critical to 18S rRNA maturation. This cleavage follows the primary cleavage at A’ site and produces an A’-A0 fragment [20]. The polyA-Seq data strongly suggests that this fragment was a substrate of SKIV2L2.

The second peak appears right after the 18S rRNA. The SKIV2L2 knockdown and control have similar amount of polyadenylation signal. It has been reported that mammalian ribosomal RNAs are pervasively polyadenylated especially at the 3′ end of the 18S rRNA[53]. This indicates that aberrant 18S rRNA go through a degradation pathway independent of mTRAMP.

Several broad and intense peaks clustered in the internal transcribed spacer 1 (ITS1), and the intensity equals between SKIV2L2 knockdown and control. Its already known that at least four cleavage events happens in ITS1 region, including the cleavage at the 3′ end of 18S and 5′ end of 5.8S rRNA [25, 54]. It seems that these cleavage fragments are polyadenylated and then degraded by a non-mTRAMP pathway. Theoretically, only two cleavage peaks should be identified within ITS1 region, while the data demonstrates a continuous polyadenylation ‘band’. One possibility is that nascent ITS1 fragments are first degraded by 3′-5′ exonuclease until a dimensional structure prevents the proceeding of the exonuclease. The degradation pause causes the polyadenylation of these fragments and finally degraded by an alternative pathway.

Two peaks were found in ITS2 region. The one immediately adjacent to the 3′ end of 5.8S rRNA has a huge peak in control compared to SKIV2L2 knockdown, while the one close to the 5′ end of 28S rRNA have a greater peak in SKIV2L2 knockdown. The location of the second peak is close to the cleavage site called 4b, which is required for 5.8S rRNA maturation. The 4b cleavage produces a 12S rRNA that is the precursor 5.8s rRNA except a tail at the 3′ end [20]. It is likely that the removal of the 3′ tail requires the mTRAMP. The SKIV2L2 knockdown caused accumulation of polyadenylated 12S rRNA, which could not be processed to mature 5.8S rRNA. The 5.8S rRNA is subsequently decreased in SKIV2L2 knockdown.

Two peaks were found in the 3′ ETS. The first one appears immediately after the 28S rRNA end, and the other in the middle of the 3′ ETS. Both have a stronger signal in SKIV2L2 knockdown. It was reported that 28S rRNA went through some polyadenylation at the 3′ end, as well as in the transcript body. The stronger signal in SKIV2L2 knockdown implies that the degradation of the 28S rRNA is subject to TRAMP pathway. Not much is known about the 3′ETS processing, and the signal in the middle of this region strongly suggesting an unknown cleavage event happens in 3′ETS.

An internal-A priming peak at the 5′ ETS drew us attention. This peak stem from the negative strand and have almost equal abundance in Control2 and SKIV2L22.

This implied transcription activities presents in this area on the negative strand, and the equal abundance excludes SKIV2L2's role in regulating this transcript.

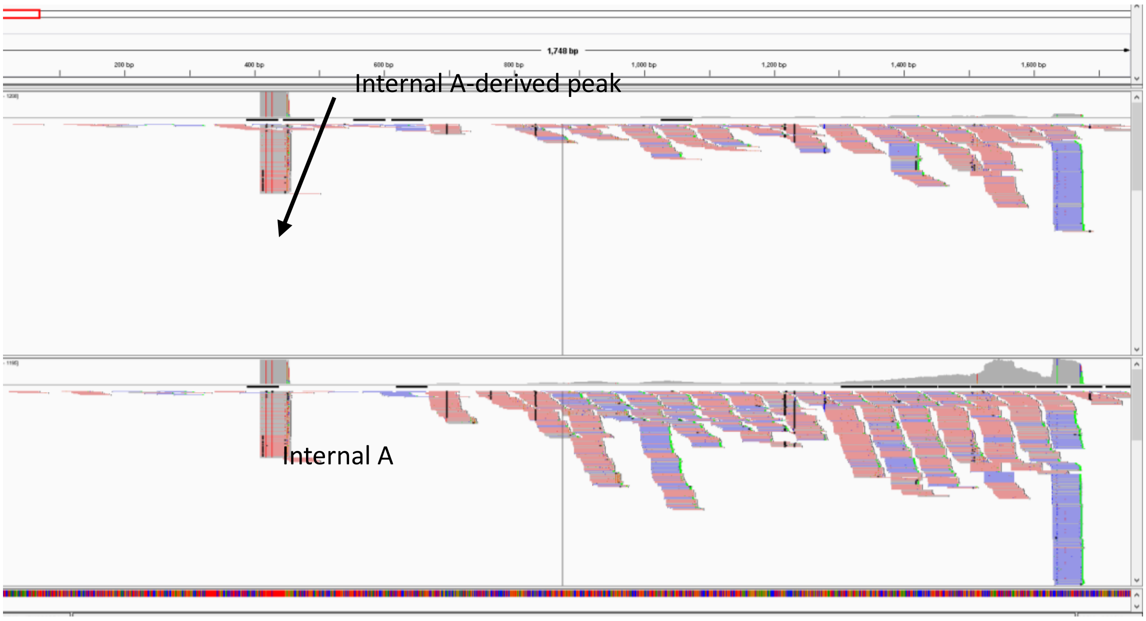

Figure II-19 Peaks in 5′ ETS due to internal-A priming. The red stretch in sequence indicates continuous, repetitive adenosine at the 5′ ETS, causing oligodT priming and result in the reads at the location.

E. 5′ leader sequence of miRNA is a common target of mTRAMP

Several miRNA-derived transcripts were identified in MEDIPS analysis as well as in our previous publication[26]. The frequent appearance of miRNA and its derivatives in SKIV2L2 target list prompted us to investigate whether they constitute a target category. The high affinity between SKIV2L2 and ARS2, a RNA cap-binding complex that is involved in miRNA biogenesis, further directed us to explore the regulation function of SKIV2L2 on miRNAs. To understand whether there miRNAs are just rare cases or miRNA 5′ leader sequence are pervasively subject to TRAMP processing, we designed miRCompare, a read abundance calculation and feature identification tool to compare the read abundance of miRNA related regions.

3′ sequencing reads within 20bps flanking the Drosha cleavage site were counted, and the distance between cleavage site and 3′ read end was calculated. The output was shown in table A-1.

Of all 859 miRNAs that were analyzed, 59 were found to have at least one read in either of the two samples and 54 were found to have a greater number of reads in SKIV2L2 knockdown than Control. The distance between cleavage site and 3′ of reads are very close, usually within 2 nucleotides, confirming the mapping accuracy of these reads. The result was achieved by directly comparing the raw reads since the mapping rate of the control and SKIV2L2 is almost identical (84%). Direct comparison could prevent bias from FPKM normalization. A large portion of miRNA in the result list have only 1 or 2 reads due to inadequate sequencing depth. Though statistically insignificant, they are ‘real’ cleavage products when taking the distance in to consideration. All the reads consistently aligns to the Drosha cleavage site (shown in the distance column), which indirectly implies that these reads are the 5′ leader sequencing of miRNA. It very unlikely that they were just random transcription ‘noises’. This massive miRNA profiling provides evidence that SKIV2L2 is involved in degrading the 5′ leader sequence.

We also observed higher read abundance at the end of mir-322 in SKIV2L2 knockdown. Due to the limitation of polyA-Seq, it’s unclear where it originates. It could possibly came from aberrant full length pre-miRNA, or from 3′ piece after cleavage. Either of these suggests that SKIV2L2 knockdown potentially degrades the miRNA processing by-products.

III. Discussion and Conclusion

Poly-A Seq enabled us to gather information of the transcriptome and gain insight into the activity of SKIV2L2 on the genomic scale. In our experimental design, we selectively picked the isolated PolyA-tailed transcripts and only sequenced the last 200bp with paired-end sequencing. This design is efficient and computationally friendly because only a 3′ trunk of the full transcript was sequenced. The data produced will be much less than the full length RNA-Seq, leaving much less pressure on downstream analysis procedures especially for some computing intensive steps such as intersecting the reads with Refseq. However, this design makes it harder to interpret the reads and identify to the transcript it is derived from, as well as to collect information regarding the length of the transcript. PolyA-Seq is ideal for a general profiling of transcriptome and in future studies, RNA-Seq analysis on full length transcripts will reveal more information about the transcripts that is usually degraded.

Our analysis showed a high rate of internal-A priming in our sample (18.2% average). This high rate greatly affected the accuracy of the downstream analysis procedures. For example, the miRCompare generated a more accurate average distance between the read end and Drosha cleavage site after removing the internally primed reads. Removing these false signal greatly improved the quality and efficiency of the differential analysis.

On the other hand, the internal-A priming provides extra information about the transcription activities in some cases. A polyA signal from internal oligo-d(T) priming is an indirect evidence of active transcription at this genome location. By design, a transcript that does not have a poly-A tail will not be reported. However, if the A-rich motif present in the transcript, it will be primed and generate a polyA signal, though the reported location is not the 3′end of that transcript. This trait complements the limitation that only the last 200bp of a transcript were sequenced, and act as a hint for inferring the length of the transcript. For example, a 600bp long transcript can only be sequence for its last 200bp. If a poly a signal exists in 200bp location, there will be a polyadenylation signal which indicates this transcript exists in the 200bp location.

A. Mouse SKIV2L2 targets a wide spectrum of non-coding RNAs

RNA processing and degradation have been well studied in yeast compared to in mammals. In order to elucidate the targets of mouse TRAMP complex, specifically the targets of SKIV2L2, we knock down the SKIV2L2 levels with siRNA in mouse N2A cells, and transformed the transcriptome with PolyA-Seq [26]. In this work, a more extensive bioinformatics analysis on the PolyA-Seq data was described using differential calling algorithms MEDIPS and F-Seq and miRNA feature recognition algorithm that was developed herein named miRCompare. The MEDIPS statistical analysis discovered 17 additional differential polyadenylated locations under 0.1 FDR (False Discovery Rate), and 14 of these are consider novel SKIV2L2 RNA targets. In addition to TRAMP targets previously identified in yeast, this analysis identified miRNA 5′ leader sequence, snoRNA derivatives and some long non-coding RNAs as SKIV2L2 targets. Our analysis also confirmed that mouse SKIV2L2 is involved in the 5.8S rRNA maturation process. These evidence strongly suggests that SKIV2L2 possesses a wider spectrum of RNA targets than yeast and the TRAMP complex.

SKIV2L2 associates with multiple complexes in mammalian cells and plays a central role in regulating the metabolism of RNAs. The yeast SKIV2L2 functions independently or in a complex with RBM7 and ZCCHC8. SKIV2L2 associates with RBM7 and ZCCHC8 to form the NEXT complex, and with ARS2 and Cap Binding Complex (CBC) to form an RNA binding complex that regulates the PROMPTs degradation [55]. In mouse ES cells, SKIV2L2 copurifies with NANOG, a pluripotency regulating and programming protein [56] suggesting that SKIV2L2 is a co-transcriptional modification protein that couples transcription, quality control and degradation of a wide range of RNAs. This observation also provided a clue on the cellular compartmentalization of SKIV2L2. The diverse association of SKIV2L2 with nuclear and cytoplasm complexes indicates a more diverse behavior. The diversity of SKIV2L2 targets and the presence in various cellular locations strongly suggests

SKIV2L2 as a central component of the polyadenylation-dependent RNA quality assurance system.

B. SKIV2L2 regulates apoptosis and cell differentiation by adjusting the abundance of Gas5 via miR-21

We identified Gas5 as one of the targets in MEDIPS analysis. Generally two peaks were found within Gas5, one for a snoRNA in intron and another for the Gas5 transcript. It has been reported that pro-apoptosis activity of Gas5 reside within the mature form of Gas5, which suggests that the two peaks have distinct function and impacts [47]. The peak at the end of the transcript have slightly higher reads in control compared to SKIV2L2, contrary to our expectation that knockdown will result in the accumulation of reads as happened in most cases. This reversed abundance inspired us to thinks about indirect regulation between SKIV2L2 and Gas5. A potential link between SKIV2L2 and Gas5 is miR-21. It is a Gas5 repressor that decreases the abundance of Gas5 [57]. Our miRNA data indicates the miR-21 as a target of SKIV2L2, since accumulation was found in knockdown sample. SKIV2L2 knockdown caused the accumulation of miR-21, which repressed the abundance of Gas5.

Gas5 is a long non-coding RNA ~700bp long that serves as the host gene for 11 snoRNAs [48]. It was reported that Gas5 promotes cell apoptosis by acting as a ligand of glucocorticoid receptor [58]. Cell apoptosis and differentiation are highly similar cellular processes [59], and glucocorticoid attenuates differentiation [60]. In our preliminary experiment, the SKIV2L2 knockdown cells were more prone to differentiation rather than apoptosis when induced. It is likely that decreased level of Gas5 promoted glucocorticoid for receptors and stimulated differentiation.

The function of Gas5 is believed to be regulated by NMD pathways [46]. Our data revealed one more quality assurance pathway that regulate mTRAMP and SKIV2L2. This novel discovery will greatly deepen the significance of SKIV2L2 in regulating cellular processes and expand the role of SKIV2L2 from a surveillance component to a RNA metabolism regulator.

C. SKIV2L2 facilitates snoRNA maturation and quality control.

We found several instances that SKIV2L2 affects the abundance of snoRNA as well as pre-snoRNA. We identified pre-snoRNA 74 in Gas5, snoRNA 6, 62, 73 in RPSA as well as U3. The targeted snoRNAs possesses diversed properties. Some of them belongs to C/D box snoRNA and some H/ACA box snoRNA, and U3 have its own promotor while snoRNA 74 relies on the host gene. The heterogeneous of this sample implied the regulation function of SKIV2L2 is possibly universal in mammalian cells as it is in yeast. It has been reported that snoRNA maturation and degradation in yeast relies on TRAMP [1, 61], and TRAMP helps the snoRNA transcription termination, intermediate transcript polyadenylation, as well as aberrant snoRNA degradation [62]. There are two types of peak in our datasets, one have peak at the end of intron, the other at the end of the actual snoRNA. Gas5 is type 1, which could result from a transcription termination that is induced by mTRAMP, or polyadenylation for endonucleases processing. Further research is required to determine the mechanism. snoRNA 62 belongs to type 2, which is more likely a snoRNA turnover due to erroneous transcription.

Our data suggests that like in yeast, a similar process occurs in mouse cells. It is likely that in addition to the typical aberrant transcript turnover function, SKIV2L2 also participates the maturation of snoRNA by facilitating the transcription termination and removal of 3′ extended nucleotides.

D. 5.8S rRNA maturation requires SKIV2L2.

We reported that the 5′ ETS of pre-rRNA are processed by SKIV2L2, which mediates its degradation at A0 cleavage site [23]. To fully characterize the function of SKIV2L2 in rRNA maturation, we further character the polyadenylation level of the 45Kb rRNA complete repeat sequence. After data filtration and processing, a major peak was found in ITS2 in non-overlapping positions in control and SKIV2L2 knockdown. The sequencing reads accumulating in the control are located near the 3′ end of the 5.8S rRNA, and the peak in SKIV2L2 knockdown is located near the predicted C2 cleavage site. Eukaryotic rRNA maturation is a complicated series of processes that requires multiple endo/exonucleases in various cellular compartments [24]. The 7S rRNA is generated after cleavage at C2 site in the middle of ITS2, and B1 sites in ITS1 [30]. The 7S rRNA is then processed to 5.8S+30nt form, and further to 5.8S+5~8nt form (6S rRNA) [63]. The sequencing reads in the control sample is near the 3′ end of 6S rRNA while the signal in SKIV2L2 knockdown is close to the beginning of 28S. This observation suggests that SKIV2L2 is required for the processing of 7S rRNA to form the 5.8S+30 and that SKIV2L2 knockdown blocked the further processing of 7S rRNA. This result is consistent with the yeast model, that the 3′ trimming of pre-5.8S rRNA requires the participation of TRAMP. The knockdown of SKIV2L2 did not completely block the formation of 6S rRNA, but evidently reduced the amount. There might be an alternative processing pathway, or the polyadenylation of the 7S rRNA greatly accelerated the processing speed. This result also indicates that RNA Pol I products could also be targeted by mTRAMP. It was reported that two TRAMP-like complexes exists in mammalian cells [12] with different cellular compartmentalization. The maturation process of 5.8 rRNA is probably a joint processing event of the two in both the nucleus and cytoplasm. We also observed transcription activity at the 5′ ETS region on the negative strand. It is believed that this event is involved in controlling the promotor of rDNA [64], and in turn the formation of rRNAs.

E. SKIV2L2 participates in the maturation of miRNAs.

The 5′ leader sequence of few miRNAs was reported as a novel target of mTRAMP in our last publication. To confirm this result and evaluate the universality, a transcriptome data profiler was developed to estimate the read abundance of all miRNAs. Our data shows the majority of miRNA 5′ leader sequence are targeted by mTRAMP by demonstrating the abundance difference between the control and knockdown and the location match of sequencing reads at the Drosha cleavage site. The limitation in sequencing technology leads to inadequate sensitivity to lowly expressed RNAs. A portion of miRNA have a relatively low read abundance and not suitable for statistical tests. The location match indirectly proved the robustness and supports our proposition. The involvement of SKIV2L2 in miRNA biogenesis is also supported by the fact that SKIV2L2 co-immunoprecipitates with ARS2 [12]. ARS2 is a multi-functional RNA silencing regulator that stimulates miRNA processing. ARS2 plays an indispensable in RNAi gene silencing and co-purified with Drosha but not dicer [65]. These facts links SKIV2L2 to the RNAi gene silencing pathway and the biogenesis of miRNAs.

Out of 713 mouse miRNAs listed in miR-database, we found 59 miRNA in our sample. Since we use oligo-d (T) to prime transcripts with poly-A tail, it is possible that there are some miRNAs whose 5’ leader sequence is not polyadenylated. Our current experiment design cannot detect such transcripts. To address this issue, we did a new set of sequencing with RNA-Seq instead of polyA-Seq. The random priming will capture any transcripts and may provide a more comprehensive results in our new analysis.

Our analysis identified snoRNA hosting introns, snoRNA, rRNA 5’ ETS, rRNA ITS2, as well as miRNA 5’ leader sequence as targets of mouse TRAMP. The unprecedented depth and breadth of TRAMP’s involvement in ncRNA processing drives us to mine deeper into the dataset and extract the association status with external datasets.

IV. Materials and Methods

Cell Culture Techniques

N2a cell line was cultured in DMEM (GIBCO, CA) with 10%FBS (GIBCO, CA), and transferred with TEDTA in all the analysis in this article. Cells were plated at 500,000 count/cm3 in P60 plates and continued growing for 48 hrs. in 37 °C CO_2 incubator. Lipofectamine (Life Technologies, CA) were incubated with SiRNA (Ambion, Life Technologies, CA) and OPTI.MEM (GIBCO, Life Technologies, CA), respectively, and then added to the plates. The cells were harvested 24 hrs later with Trizol Reagent (Invitrogen, Carlsbad, CA). Total RNA and total protein was isolated following the standard Trizol protocol in the manual.

RNA Techniques

Reverse Transcription

Reverse transcription was performed with Oligo-dT and gene specific primers (GSP, shown in sup. table), respectively. Total RNA was reverse transcribed with M-MLV reverse transcriptase (Promega) as described in the manual.

Quantitative PCR

To measure the quantity of the miRNA 5′ leader sequence, equal amounts of cDNA were synthesized using the M-MLV reverse transcriptase (described above). The product was mixed with SYBR® Green Supermixes (BIO-RAD, Hercules, CA) together with 10 pM primer sets. Cyclophillin B was set as the reference gene. qPCR reactions was performed at 55 °C for 3min, 95 °C for 10min, 40 cycles of 95 °C for 30s and 55 °C for 2min.

Protein Techniques

Western Blot

Equal amounts of protein was loaded onto 10% SDS-PAGE gel and ran for 2hr under 80V. Gel was transferred to a nitrocellulose membrane in 25V for 2.5 hr at 4 °C transfer buffer, using NuPage Novex Gel System (Invitrogen). Blot was rinsed in PBS (pH 7.5) containing 0.1% NP-40, and blocked in 5% milk for 1hr at room temperature (RT). SKIV2L2 and bActin antibodies were added in 1:5000 ratio, and the blot was incubated overnight at 4°C. After rinsing, the blot was developed with ECL reagents and exposed to films for 1s, 10s, and 1min in dark.

Reagents formula for western blot

10X Tris-Glycine Transfer Buffer	288 g Glycine, 60.4 g Tris base 1.8 L ddH2O Add 200ml methanol when using
ECL reagents I	1.875 M Tris-Cl, pH 8.8 265µl

	Luminol (44mg/ml DMSO) 50µl ρ-coumaric acid (15mg/ml DMSO) 22µl d_2H_2O 4.66ml
ECL reagents II	1.875 M Tris-Cl, pH 8.8 265µl H_2O_2 30% solution 3µl d_2H_2O 4.73ml

Table A-1

Computational Techniques

Computing Platform

All the analysis was conducted on Ubuntu Linux 13.04 or Microsoft Windows 8.1. All the packages and sources ran well on Python 2.7.1, JRE 1.7u45, R package 3.0.1.

Datasets, Sequence alignment and preparation

Raw reads of all samples were stored in FASTQ format. Two replicates were used in mapping, and each replicate had three conditions. Of each condition, the 5′ and 3′ reads were stored in a text file, respectively. The code of samples in this article are shown below:

	Control	SKIV2L2 Knockdown	Rrp6 Knockdown
Original	Control1	SKIV2L21	Rrp61
Replicate	Control2	SKIV2L22	Rrp62

Mapping was performed with Bowtie 2.1.0 [40], a short read alignment tool. GRCm38/mm10 mouse genome assembly and rRNA complete repeat unit (GenBank: BK000964.3) was used as the reference genome. Bowtie2 index was generated by the index generation tool in the bowtie2 package with the command below:

```
$BT2_HOME/bowtie2-build $BT2_HOME/reference/assembly genome.fa genome_version
```

Reads mapping was performed with default parameters for paired-end mapping except for the multi-thread function, as shown below:

```
$BT2_HOME/bowtie2 –p 8 –x example_assembly -1 $BT2_HOME/example/reads/reads_5p.fq -2
$BT2_HOME/example/reads/reads_3p.fq -S example_condition.sam
```

The output of mapping is a text-based SAM file. To achieve a better performance in downstream analysis and visualization, SAMtools (0.1.19) [66] were used to convert SAM to BAM, the compressed binary version of SAM. The BAM file was also sorted during the conversion.

```
samtools view -bS file.sam | samtools sort - file_sorted
```

BAM files were converted to a BED file under certain circumstances. To produce a high quality BED file, reads that were not mapped need to be filtered from the BAM file.

```
samtools view -h -F 4 -b blah.bam > blah_only_mapped.bam
```

Conversion from BAM to BED is conducted with BEDTools 2.19.0[67]:

```
bamToBed -i reads.bam > reads.bed
```

Internal-A identification

Sequence reads originate from internally primed oligod (T) was filtered from alignment files to a separate file. This process was achieved through IAFilter, a genetic sequence pattern identification and extraction tool designed and implemented by me. Alignment files were loaded by IAFilter and every read was subject to a property check: whether this is a 3′ read and whether there is a continuous adenosine pattern appears after the 3′ end of the read being screened. If both criteria were met, the read will be send to the result file used for down streaming analysis.

A Python (2.7.1) script was written to identify potential internal-A reads from BED files. The script takes as input a genome sequence file (FASTA format) and a polyA read file (BED format). The output consists of two BED files. One contains potential internal-A reads while the other normal reads. A FASTA format genome sequence file is a text file recording the full sequence of the genome. Each character denotes a base pair in the genome. Conventionally, a FASTA file is further divided into sections representing different chromosomes. The beginning location of each chromosome had been pre-calculated and hard-coded into the program. The program first reads the location of a read from the BED file, fetches a 20bp sequence upstream of this location, and examines if there are five continuous ‘A’s at the start of this fetched sequence, or if there are more than 15 ‘A’s in this 20bp sequence. If this sequence meets either of these criteria, this read will be considered potential internal-A then written into the internal-A result file. Following analysis are based on the internal-A-filtered BED read file。

RefSeq based differential polyadenylation analysis

Refseq [68] annotation (PubDate 8-10-2011, mm9) was downloaded from UCSC table browser in BED format containing chromosome, transcription start/end, translation start/end, gene symbol, gene id, and strand. Based on this data, a 100bp interval centered at the transcription end site was created for each entry in the Refseq data as the reads collection window. The sequence reads was pre-aligned by BWA [69] and converted to BED format with internal-A removed. A python script was written to count the number of reads at each genomic position, and the result of this count script was imported to Galaxy [44, 70, 71] Bioinformatics analysis platform. By executing the 'Join' function in 'Operate on Genomic Intervals' category, and set the minimal overlap to 1bp, the Galaxy computation platform returned a table with gene name and reads associated with that gene. This result was further imported into MS Excel and subtotaled by the gene name. The transcription variants were removed by 'Remove Duplicate' function to prevent duplicate counting. Both control and SKIV2L2 samples are subject to this pipeline and a ratio for each gene between control, and SKIV2L2 was calculated with the formula below:

$$f_{(i)}=\log_2(\mathrm{SKIV2L2}_{(i)}/\mathrm{control}_{(i)}) \quad \mathbf{i=RefSeq\ genes}$$

A histogram was plotted with the graph tool of MS Excel.

F-Seq analysis

Mus musculus ribosomal DNA complete repeating unit sequence (GenBank: BK000964.3) was downloaded from NCBI in FASTA format. The ribosomal RNA starts at 1 and ends at 13403. To get a solid result, 1-14000 region was extracted and saved as a new FASTA file named rrt.fasta (ribosomal RNA truncated). Bowtie2-index was called to generate index files for mapping with default parameters. Control1, control2, SKIV2L21, SKIV2L22, rrp61, rrp62 was mapped to this custom rDNA sequence and the output SAM files were sorted and converted to a BED file following the protocol stated above. For each sample, the BED entries were separated by strand. F-Seq [72], a feature density estimator, was employed to generate the density signal. The fragment size was set to 0, feature window to 60, and output format to wig. Each strand of a sample produced an output wig file, and all the samples were processed by this protocol. The wig density files were imported to MS Excel and aligned according to their coordinate, and a line chard was plotted based on density signal.

MEDIPS analysis

R package 3.0.1 was downloaded from the official website. BioConductor [73] and MEDIPS packages [45] were installed following the software manual. UCSC mm10 reference genome was loaded with the command. For SKIV2L2 knockdown and rrp6 knockdown analysis, both sequencing samples were used to improve statistical confidence. The parameters for this analysis were slightly adjusted to the read distribution of polyA-seq. The window size of this analysis was the default value, 100bp. All other parameters were set as recommended by the manual. Three R sets were constructed and each of the sets contains original and replicate sequencing data for one condition. For the statistical parameters, Bonferroni method was used for adjusting the p-value, edgeR package was employed for statistical calculation, poisson was the probability model, both MeDIP and CNV was turned off for MEDIPS. Regions with a p-value less than 0.1 were reported.

miRCompare analysis

The 5′ leader sequence of all mmu miRNA was collected and profiled. miRCompare, an read abundance calculation tool was designed and implemented by me. A list of miRNA 5′ duplex coordination containing the drosha cleavage site in mmu was curated and loaded into miRCompare. The tool collect reads adjacent to the drosha cleavage site, calculate abundance, distance to the site, as well as other properties provisioned by the designer.

miRCompare is a general purpose RNA-Seq abundance estimation tool. It scans a list of intervals along the genome and report the read count and other information for each interval. miRCompare takes a BAM sequence alignment file as input data source, and a tab-delimited text file as intervals of interest source. The analysis result will be reported in a text file. This Java-based software incorporates Picard (http://picard.sourceforge.net), a SAM/BAM manipulation tool, as BAM data access implementation. Picard allowed miRCompare to query intervals, iterate sequence alignment, and get attributes of the reads. A GFF3 parser was used to iterate the file and encapsulate coordinates into BED record objects. miRCompare fetches all the reads that fall within (or partially within) the specified interval, and filter out some reads according to instruction. Currently, two filtration options are available. One option controls whether to report reads of both strands or just report reads mapped to the same strand as the inquired RNA. The other option is whether to report the 5′ read or just 3′ read. Both of these functions are achieved by calling the methods of SAMRecord class imported from Picard. An interval adaptor was designed in the query function. It changes the range of the query interval by the number of bps supplied by the user. This function conveniences the investigation of upstream and downstream regions of an interval, without modifying the original interval file. The software stores the returned reads as the value of a hashtable and the name of the queried record as the key. This hashtable is kept for further calculation.

For the specific use of miRNA 5′ leader sequence analysis, several modifications were made and functions were added to adapt to the features of miRNA. Sequence alignments will be reported if they are 3′ end reads and on the same strand as the inquiry. Read count will be calculated by calling the size of the storage array. A distance to the start of the inquiry interval will also be calculated. The calculation formula differs between positive strand and negative strand.

Control2 and SKIV2L22 alignment file (BAM) were used as the alignment source. The miRNA coordinate list of mouse was obtained from miRBase. (ftp://mirbase.org/pub/mirbase/CURRENT/genomes/mmu.gff3) The annotation file contains ‘primary transcript,’ 5′ and 3′ miRNA. To focus the search on the 5′ ETS region, ‘primary transcript,’ 3′ entries was filtered out with GNU GREP. The analysis result was exported as single txt files which were further processed with MS Excel for numerical calculation and visualization.

VI. BIBLIOGRAPHY

1. LaCava, J., et al., *RNA Degradation by the Exosome Is Promoted by a Nuclear Polyadenylation Complex.* Cell, 2005. **121**(5): p. 713-724.
2. Vaňáčová, Š., et al., *A New Yeast Poly(A) Polymerase Complex Involved in RNA Quality Control.* PLoS Biol, 2005. **3**(6): p. e189.
3. Callahan, K.P. and J.S. Butler, *TRAMP Complex Enhances RNA Degradation by the Nuclear Exosome Component Rrp6.* Journal of Biological Chemistry, 2010. **285**(6): p. 3540-3547.
4. Fasken, M.B., et al., *Air1 Zinc Knuckles 4 and 5 and a Conserved IWRXY Motif Are Critical for the Function and Integrity of the Trf4/5-Air1/2-Mtr4 Polyadenylation (TRAMP) RNA Quality Control Complex.* Journal of Biological Chemistry, 2011. **286**(43): p. 37429-37445.
5. Hamill, S., S.L. Wolin, and K.M. Reinisch, *Structure and function of the polymerase core of TRAMP, a RNA surveillance complex.* Proc Natl Acad Sci U S A, 2010. **107**(34): p. 15045-50.
6. Haracska, L., et al., *Trf4 and Trf5 Proteins of Saccharomyces cerevisiae Exhibit Poly(A) RNA Polymerase Activity but No DNA Polymerase Activity.* Molecular and Cellular Biology, 2005. **25**(22): p. 10183-10189.
7. Jia, H., et al., *RNA unwinding by the Trf4/Air2/Mtr4 polyadenylation (TRAMP) complex.* Proc Natl Acad Sci U S A, 2012. **109**(19): p. 7292-7.
8. Holub, P., et al., *Air2p is critical for the assembly and RNA-binding of the TRAMP complex and the KOW domain of Mtr4p is crucial for exosome activation.* Nucleic Acids Research, 2012. **40**(12): p. 5679-5693.
9. Arigo, J.T., et al., *Termination of cryptic unstable transcripts is directed by yeast RNA-binding proteins Nrd1 and Nab3.* Mol Cell, 2006. **23**(6): p. 841-51.
10. Tudek, A., et al., *Molecular Basis for Coordinating Transcription Termination with Noncoding RNA Degradation.* Molecular Cell, 2014. **55**(3): p. 467-481.
11. Houseley, J. and D. Tollervey, *The Many Pathways of RNA Degradation.* Cell, 2009. **136**(4): p. 763-776.
12. Lubas, M., et al., *Interaction profiling identifies the human nuclear exosome targeting complex.* Mol Cell, 2011. **43**(4): p. 624-37.
13. Lubas, M., et al., *The Human Nuclear Exosome Targeting Complex Is Loaded onto Newly Synthesized RNA to Direct Early Ribonucleolysis.* Cell Reports. **10**(2): p. 178-192.
14. Raijmakers, R., et al., *PM–Scl-75 is the main autoantigen in patients with the polymyositis/scleroderma overlap syndrome.* Arthritis & Rheumatism, 2004. **50**(2): p. 565-569.
15. Wan, J., et al., *Mutations in the RNA exosome component gene EXOSC3 cause pontocerebellar hypoplasia and spinal motor neuron degeneration.* Nature genetics, 2012. **44**(6): p. 704-708.

16. Kerr, T.A. and N.O. Davidson, *Therapeutic RNA Manipulation in Liver Disease.* Hepatology (Baltimore, Md.), 2010. **51**(3): p. 1055-1061.
17. Henriksson, M. and B. Lüscher, *Proteins of the Myc Network: Essential Regulators of Cell Growth and Differentiation*, in *Advances in Cancer Research*, F.V.W. George and K. George, Editors. 1996, Academic Press. p. 109-182.
18. Wisdom, R. and W. Lee, *Translation of c-myc mRNA is required for its post-transcriptional regulation during myogenesis.* Journal of Biological Chemistry, 1990. **265**(31): p. 19015-21.
19. Nazar, R., *Ribosomal RNA Processing and Ribosome Biogenesis in Eukaryotes.* IUBMB Life, 2004. **56**(8): p. 457-465.
20. Kent, T., Y.R. Lapik, and D.G. Pestov, *The 5' external transcribed spacer in mouse ribosomal RNA contains two cleavage sites.* RNA, 2009. **15**(1): p. 14-20.
21. Craig, N., S. Kass, and B. Sollner-Webb, *Nucleotide sequence determining the first cleavage site in the processing of mouse precursor rRNA.* Proc Natl Acad Sci U S A, 1987. **84**(3): p. 629-33.
22. Strezoska, Ž., D.G. Pestov, and L.F. Lau, *Bop1 Is a Mouse WD40 Repeat Nucleolar Protein Involved in 28S and 5.8S rRNA Processing and 60S Ribosome Biogenesis.* Molecular and Cellular Biology, 2000. **20**(15): p. 5516-5528.
23. Dorweiler, J.E., et al., *Certain Adenylated Non-Coding RNAs, Including 5' Leader Sequences of Primary MicroRNA Transcripts, Accumulate in Mouse Cells following Depletion of the RNA Helicase MTR4.* PLoS ONE, 2014. **9**(6): p. e99430.
24. Thomson, E. and D. Tollervey, *The final step in 5.8S rRNA processing is cytoplasmic in Saccharomyces cerevisiae.* Mol Cell Biol, 2010. **30**(4): p. 976-84.
25. Lamanna, A.C. and K. Karbstein, *An RNA Conformational Switch Regulates Pre-18S rRNA Cleavage.* Journal of molecular biology, 2011. **405**(1): p. 3-17.
26. Dorweiler, J.E., et al., *Certain adenylated non-coding RNAs, including 5' leader sequences of primary microRNA transcripts, accumulate in mouse cells following depletion of the RNA helicase MTR4.* PLoS One, 2014. **9**(6): p. e99430.
27. Brameier, M., et al., *Human box C/D snoRNAs with miRNA like functions: expanding the range of regulatory RNAs.* Nucleic Acids Res, 2011. **39**: p. 675 - 86.
28. Scott, M. and M. Ono, *From snoRNA to miRNA: dual function regulatory non-coding RNAs.* Biochimie, 2011. **93**: p. 1987 - 92.
29. Clouet d'Orval, B., et al., *Box C/D RNA guides for the ribose methylation of archaeal tRNAs. The tRNATrp intron guides the formation of two ribose-methylated nucleosides in the mature tRNATrp.* Nucleic Acids Res, 2001. **29**: p. 4518 - 29.

30. Allmang, C., et al., *Functions of the exosome in rRNA, snoRNA and snRNA synthesis.* Embo J, 1999. **18**(19): p. 5399-410.

31. Brown, J., M. Echeverria, and L. Qu, *Plant snoRNAs: functional evolution and new modes of gene expression.* Trends Plant Sci, 2003. **8**: p. 42 - 9.

32. Decatur, W. and M. Fournier, *rRNA modifications and ribosome function.* Trends Biochem Sci, 2002. **27**: p. 344 - 51.

33. Ganot, P., M. Caizergues-Ferrer, and T. Kiss, *The family of box ACA small nucleolar RNAs is defined by an evolutionarily conserved secondary structure and ubiquitous sequence elements essential for RNA accumulation.* Genes & Development, 1997. **11**(7): p. 941-956.

34. Samarsky, D.A., et al., *The snoRNA box C/D motif directs nucleolar targeting and also couples snoRNA synthesis and localization*. Vol. 17. 1998. 3747-3757.

35. Decatur, W.A., et al., *Identifying effects of snoRNA-guided modifications on the synthesis and function of the yeast ribosome.* Methods Enzymol, 2007. **425**: p. 283-316.

36. Kass, S., et al., *The U3 small nucleolar ribonucleoprotein functions in the first step of preribosomal RNA processing.* Cell, 1990. **60**(6): p. 897-908.

37. Weinstein, L.B. and J.A. Steitz, *Guided tours: from precursor snoRNA to functional snoRNP.* Curr Opin Cell Biol, 1999. **11**(3): p. 378-84.

38. Terns, M.P. and R.M. Terns, *Small nucleolar RNAs: versatile trans-acting molecules of ancient evolutionary origin.* Gene Expr, 2002. **10**(1-2): p. 17-39.

39. Ni, T., et al., *A paired-end sequencing strategy to map the complex landscape of transcription initiation.* Nat Meth, 2010. **7**(7): p. 521-527.

40. Langmead, B. and S.L. Salzberg, *Fast gapped-read alignment with Bowtie 2.* Nat Meth, 2012. **9**(4): p. 357-359.

41. Jia, H., et al., *The RNA Helicase Mtr4p Modulates Polyadenylation in the TRAMP Complex.* Cell. **145**(6): p. 890-901.

42. Schmidt, K. and J.S. Butler, *Nuclear RNA surveillance: role of TRAMP in controlling exosome specificity.* Wiley Interdisciplinary Reviews: RNA, 2013. **4**(2): p. 217-231.

43. Mortazavi, A., et al., *Mapping and quantifying mammalian transcriptomes by RNA-Seq.* Nat Methods, 2008. **5**(7): p. 621-8.

44. Blankenberg, D., et al., *Galaxy: A Web-Based Genome Analysis Tool for Experimentalists*, in *Current Protocols in Molecular Biology*. 2001, John Wiley & Sons, Inc.

45. Lienhard, M., et al., *MEDIPS: genome-wide differential coverage analysis of sequencing data derived from DNA enrichment experiments.* Bioinformatics, 2013.

46. Tani, H., M. Torimura, and N. Akimitsu, *The RNA Degradation Pathway Regulates the Function of GAS5 a Non-Coding RNA in Mammalian Cells.* PLoS ONE, 2013. **8**(1): p. e55684.

47. Pickard, M.R., M. Mourtada-Maarabouni, and G.T. Williams, *Long non-coding RNA GAS5 regulates apoptosis in prostate cancer cell lines.*

Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease, 2013. **1832**(10): p. 1613-1623.

48. Smith, C.M. and J.A. Steitz, *Classification of gas5 as a multi-small-nucleolar-RNA (snoRNA) host gene and a member of the 5'-terminal oligopyrimidine gene family reveals common features of snoRNA host genes.* Mol Cell Biol, 1998. **18**(12): p. 6897-909.

49. Satoh, K., et al., *Cloning of 67-kDa laminin receptor cDNA and gene expression in normal and malignant cell lines of the human lung.* Cancer Letters, 1992. **62**(3): p. 199-203.

50. Hubbard, T., et al., *The Ensembl genome database project.* Nucleic Acids Res, 2002. **30**(1): p. 38-41.

51. Kiss, A.M., et al., *Human Box H/ACA Pseudouridylation Guide RNA Machinery.* Molecular and Cellular Biology, 2004. **24**(13): p. 5797-5807.

52. Hahm, S., et al., *Targeted Deletion of the Vgf Gene Indicates that the Encoded Secretory Peptide Precursor Plays a Novel Role in the Regulation of Energy Balance.* Neuron. **23**(3): p. 537-548.

53. Slomovic, S., et al., *Polyadenylation of ribosomal RNA in human cells.* Nucleic Acids Research, 2006. **34**(10): p. 2966-2975.

54. Gerbi SA, B.A., *Pre-Ribosomal RNA Processing in Multicellular Organisms.* Curie Bioscience Database [Internet].

55. Andersen, P.R., et al., *The human cap-binding complex is functionally connected to the nuclear RNA exosome.* Nat Struct Mol Biol, 2013. **20**(12): p. 1367-76.

56. Costa, Y., et al., *NANOG-dependent function of TET1 and TET2 in establishment of pluripotency.* Nature, 2013. **495**(7441): p. 370-374.

57. Zhang, Z., et al., *Negative regulation of lncRNA GAS5 by miR-21.* Cell Death Differ, 2013. **20**(11): p. 1558-68.

58. Kino, T., et al., *Noncoding RNA gas5 is a growth arrest- and starvation-associated repressor of the glucocorticoid receptor.* Sci Signal, 2010. **3**(107): p. ra8.

59. Lanneau, D., et al., *Apoptosis versus cell differentiation: role of heat shock proteins HSP90, HSP70 and HSP27.* Prion, 2007. **1**(1): p. 53-60.

60. Rauch, A., et al., *Glucocorticoids suppress bone formation by attenuating osteoblast differentiation via the monomeric glucocorticoid receptor.* Cell Metab, 2010. **11**(6): p. 517-31.

61. van Hoof, A., P. Lennertz, and R. Parker, *Yeast Exosome Mutants Accumulate 3'-Extended Polyadenylated Forms of U4 Small Nuclear RNA and Small Nucleolar RNAs.* Molecular and Cellular Biology, 2000. **20**(2): p. 441-452.

62. Grzechnik, P. and J. Kufel, *Polyadenylation linked to transcription termination directs the processing of snoRNA precursors in yeast.* Mol Cell, 2008. **32**(2): p. 247-58.

63. Briggs, M.W., K.T. Burkard, and J.S. Butler, *Rrp6p, the yeast homologue of the human PM-Scl 100-kDa autoantigen, is essential for efficient 5.8 S rRNA 3' end formation.* J Biol Chem, 1998. **273**(21): p. 13255-63.

64. Mayer, C., et al., *Intergenic transcripts regulate the epigenetic state of rRNA genes.* Mol Cell, 2006. **22**(3): p. 351-61.

65. Gruber, J.J., et al., *Ars2 Links the Nuclear Cap-Binding Complex to RNA Interference and Cell Proliferation.* Cell, 2009. **138**(2): p. 328-339.

66. Li, H., et al., *The Sequence Alignment/Map format and SAMtools.* Bioinformatics, 2009. **25**(16): p. 2078-9.

67. Quinlan, A.R. and I.M. Hall, *BEDTools: a flexible suite of utilities for comparing genomic features.* Bioinformatics, 2010. **26**(6): p. 841-842.

68. Pruitt, K.D., et al., *RefSeq: an update on mammalian reference sequences.* Nucleic Acids Res, 2014. **42**(Database issue): p. D756-63.

69. Li, H. and R. Durbin, *Fast and accurate short read alignment with Burrows-Wheeler transform.* Bioinformatics, 2009. **25**(14): p. 1754-60.

70. Giardine, B., et al., *Galaxy: A platform for interactive large-scale genome analysis.* Genome Research, 2005. **15**(10): p. 1451-1455.

71. Goecks, J., et al., *Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences.* Genome Biology, 2010. **11**(8): p. R86.

72. Boyle, A.P., et al., *F-Seq: a feature density estimator for high-throughput sequence tags.* Bioinformatics, 2008. **24**(21): p. 2537-2538.

73. Gentleman, R.C., et al., *Bioconductor: open software development for computational biology and bioinformatics.* Genome Biol, 2004. **5**(10): p. R80.

VII. Appendix

miRNA	Control 2 Ct.	Control2 avg. dist.	SKIV2L22 Ct.	Mtr2 avg. dist.
ID=MIMAT0000548;Alias=MIMAT0000548;Name=mmu-miR-322-5p;Derives_from=MI0000590	4	1.75	256	1.421875
ID=MIMAT0000150_1;Alias=MIMAT0000150;Name=mmu-miR-138-5p;Derives_from=MI0000722	2	2	95	1.336842105
ID=MIMAT0000523;Alias=MIMAT0000523;Name=mmu-let-7c-5p;Derives_from=MI0000560	2	2	72	2.013888889
ID=MIMAT0000649;Alias=MIMAT0000649;Name=mmu-miR-17-5p;Derives_from=MI0000687	2	1.5	40	1.475
ID=MIMAT0000522;Alias=MIMAT0000522;Name=mmu-let-7b-5p;Derives_from=MI0000558	6	2	32	2.15625
ID=MIMAT0002104;Alias=MIMAT0002104;Name=mmu-miR-463-5p;Derives_from=MI0002398	0	0	11	1.909090909
ID=MIMAT0000534;Alias=MIMAT0000534;Name=mmu-miR-26b-5p;Derives_from=MI0000575	0	0	8	2.375
ID=MIMAT0000383;Alias=MIMAT0000383;Name=mmu-let-7d-5p;Derives	1	1	7	1

_from=MI0000405				
ID=MIMAT0000530;Alias=MIMAT0000530;Name=mmu-miR-21a-5p;Derives_from=MI0000569	0	0	6	0.5
ID=MIMAT0000386;Alias=MIMAT0000386;Name=mmu-miR-106b-5p;Derives_from=MI0000407	1	2	5	2.4
ID=MIMAT0004527;Alias=MIMAT0004527;Name=mmu-miR-124-5p;Derives_from=MI0000716	0	0	5	1.8
ID=MIMAT0004631;Alias=MIMAT0004631;Name=mmu-miR-29a-5p;Derives_from=MI0000576	0	0	4	2
ID=MIMAT0000218;Alias=MIMAT0000218;Name=mmu-miR-24-1-5p;Derives_from=MI0000231	0	0	4	2
ID=MIMAT0000677;Alias=MIMAT0000677;Name=mmu-miR-7a-5p;Derives_from=MI0000728	0	0	4	1.75
ID=MIMAT0004841;Alias=MIMAT0004841;Name=mmu-miR-871-5p;Derives_from=MI0005471	0	0	4	2
ID=MIMAT0004523;Alias=MIMAT0004523;Name=mmu-miR-29b-1-5p;Derives_from=MI0000143	0	0	4	0.5
ID=MIMAT0004848;Alias=MIMAT0004848;Name=mmu-miR-883a-5p;Derives_from=MI0005476	0	0	3	-2
ID=MIMAT0000529;Alias=MIMAT0000529;Name=mmu-miR-20a-5p;Derives_from=MI0000568	0	0	3	1
ID=MIMAT0004838;Alias=MIMAT0004838;Name=mmu-miR-742-5p;Derives_from=MI0005206	0	0	3	0.333333333
ID=MIMAT0000526;Alias=MIMAT0000526;Name=mmu-miR-15a-5p;Derives_from=MI0000564	0	0	3	2
ID=MIMAT0003727;Alias=MIMAT0003727;Name=mmu-miR-374b-5p;Derives_from=MI0004125	0	0	2	2
ID=MIMAT0004873_1;Alias=MIMAT0004873;Name=mmu-miR-465c-5p;Derives_from=MI0005501	0	0	2	7
ID=MIMAT0000667;Alias=MIMAT0000667;Name=mmu-miR-33-5p;Derives_from=MI0000707	0	0	2	2
ID=MIMAT0000130;Alias=MIMAT0000130;Name=mmu-miR-30b-5p;Derives_from=MI0000145	0	0	2	1
ID=MIMAT0000128;Alias=MIMAT0000128;Name=mmu-miR-30a-5p;Derives_from=MI0000144	0	0	2	0
ID=MIMAT0000747;Alias=MIMAT0000747;Name=mmu-miR-382-5p;Derives_from=MI0000799	0	0	2	-0.5
ID=MIMAT0000221;Alias=MIMAT0000221;Name=mmu-miR-191-5p;Derives_from=MI0000233	1	2	1	2
ID=MIMAT0000654;Alias=MIMAT0000654;Name=mmu-miR-32-5p;Derives_from=MI0000691	1	1	1	-1
ID=MIMAT0003731;Alias=MIMAT0003731;Name=mmu-miR-671-5p;Derives_from=MI0004133	1	2	1	2

ID=MIMAT0000527;Alias=MIMAT0000527;Name=mmu-miR-16-5p;Derives_from=MI0000565	0	0	1	2
ID=MIMAT0029822;Alias=MIMAT0029822;Name=mmu-miR-7658-5p;Derives_from=MI0024998	0	0	1	0
ID=MIMAT0029906;Alias=MIMAT0029906;Name=mmu-miR-7688-5p;Derives_from=MI0025041	0	0	1	-10
ID=MIMAT0004629;Alias=MIMAT0004629;Name=mmu-miR-22-5p;Derives_from=MI0000570	0	0	1	2
ID=MIMAT0000525;Alias=MIMAT0000525;Name=mmu-let-7f-5p;Derives_from=MI0000562	0	0	1	2
ID=MIMAT0016980;Alias=MIMAT0016980;Name=mmu-miR-23b-5p;Derives_from=MI0000141	0	0	1	2
ID=MIMAT0004884;Alias=MIMAT0004884;Name=mmu-miR-466h-5p;Derives_from=MI0005511	0	0	1	2
ID=MIMAT0004526;Alias=MIMAT0004526;Name=mmu-miR-101a-5p;Derives_from=MI0000148	0	0	1	3
ID=MIMAT0017172;Alias=MIMAT0017172;Name=mmu-miR-410-5p;Derives_from=MI0001161	0	0	1	0
ID=MIMAT0000525_1;Alias=MIMAT0000525;Name=mmu-let-7f-5p;Derives_from=MI0000563	0	0	1	1
ID=MIMAT0000132;Alias=MIMAT0000132;Name=mmu-miR-99b-5p;Derives_from=MI0000147	0	0	1	0
ID=MIMAT0004522;Alias=MIMAT0004522;Name=mmu-miR-27b-5p;Derives_from=MI0000142	0	0	1	1
ID=MIMAT0000215;Alias=MIMAT0000215;Name=mmu-miR-186-5p;Derives_from=MI0000228	0	0	1	1
ID=MIMAT0001419;Alias=MIMAT0001419;Name=mmu-miR-433-5p;Derives_from=MI0001525	0	0	1	2
ID=MIMAT0004850;Alias=MIMAT0004850;Name=mmu-miR-883b-5p;Derives_from=MI0005477	0	0	1	2
ID=MIMAT0014834;Alias=MIMAT0014834;Name=mmu-miR-3064-5p;Derives_from=MI0014026	0	0	1	-6
ID=MIMAT0017327;Alias=MIMAT0017327;Name=mmu-miR-669f-5p;Derives_from=MI0006287	0	0	1	2
ID=MIMAT0003740;Alias=MIMAT0003740;Name=mmu-miR-674-5p;Derives_from=MI0004611	0	0	1	2
ID=MIMAT0004664;Alias=MIMAT0004664;Name=mmu-miR-214-5p;Derives_from=MI0000698	0	0	1	3
ID=MIMAT0001418;Alias=MIMAT0001418;Name=mmu-miR-431-5p;Derives_from=MI0001524	0	0	1	2
ID=MIMAT0000663;Alias=MIMAT0000663;Name=mmu-miR-218-5p;Derives_from=MI0000701	0	0	1	2
ID=MIMAT0000210;Alias=MIMAT0000210;Name=mmu-miR-181a-5p;Derives_from=MI0000697	0	0	1	2

ID=MIMAT0014822;Alias=MIMAT0014822;Name=mmu-miR-3057-5p;Derives_from=MI0014020	0	0	1	2
ID=MIMAT0003128;Alias=MIMAT0003128;Name=mmu-miR-485-5p;Derives_from=MI0003492	0	0	1	-1
ID=MIMAT0009441;Alias=MIMAT0009441;Name=mmu-miR-1968-5p;Derives_from=MI0009965	0	0	1	0
ID=MIMAT0017063;Alias=MIMAT0017063;Name=mmu-miR-29b-2-5p;Derives_from=MI0000712	2	4.5	0	0
ID=MIMAT0027343;Alias=MIMAT0027343;Name=mmu-miR-6516-5p;Derives_from=MI0022266	1	-7	0	0
ID=MIMAT0005859;Alias=MIMAT0005859;Name=mmu-miR-1198-5p;Derives_from=MI0006306	1	14	0	0
ID=MIMAT0027770;Alias=MIMAT0027770;Name=mmu-miR-6935-5p;Derives_from=MI0022782	1	2	0	0
ID=MIMAT0017053;Alias=MIMAT0017053;Name=mmu-miR-212-5p;Derives_from=MI0000696	1	2	0	0

Table A-1 miRNA 5′ leader sequence abundance and average distance of Control2 and SKIV2L22knockdown. Column left to right, miRNA name, read count in Control2, average distance of Control2, read count in SKIV2L22knockdown, average distance of SKIV2L22knockdown.

IAFilter.java

```
package edu.marquette.bio.andersonlab.iafilter;

import htsjdk.samtools.SAMFileReader;
import htsjdk.samtools.SAMFileWriter;
import htsjdk.samtools.SAMFileWriterFactory;
import htsjdk.samtools.SAMRecord;

import java.io.File;



public class InternalAFilter {

  private File inputSAMFile;
  private File outputSAMFile;
  private File outputSAMFile2;
  private File inputFastaFile;
  private File inputFastaIndexFile;
  FastaDAO fastadao;
  int contThre;
  int ratioThre;
  int ratioRange;

  public InternalAFilter(File insam, File outsam, File outsam2, File infasta,
          File infastaindex, int conthre, int ratthre, int ratrange) {
      inputSAMFile = insam;
      outputSAMFile = outsam;
      outputSAMFile2 = outsam2;
      inputFastaFile = infasta;

      inputFastaIndexFile = infastaindex;
      contThre=conthre;
      ratioThre=ratthre;
      ratioRange=ratrange;

      fastadao = new FastaDAO(infasta.toPath(), infastaindex);
  }

  public static void main(String[] args) {

      File insamfile = new File(args[0]);
      File outsamfile = new File(args[1]);
      File outsamfile2 = new File(args[2]);
      File outsamindex = new File(args[3]);
      File mm10 = new File(args[4]);
      File mm10index = new File(args[5]);
      InternalAFilter iaf = new InternalAFilter(insamfile, outsamfile, outsamfile2,mm10,mm10index, 5, 15,20);
      long starttime = System.currentTimeMillis();
      iaf.scan();
      long time = System.currentTimeMillis()-starttime;
      System.out.println("Time used" + time/1000);


  }

  public void scan() {
      SAMFileReader inputSAM = new SAMFileReader(inputSAMFile);
      SAMFileWriter outputSAM = new SAMFileWriterFactory().makeBAMWriter(
              inputSAM.getFileHeader(), true, outputSAMFile);
      SAMFileWriter outputSAM2 = new SAMFileWriterFactory().makeBAMWriter(
              inputSAM.getFileHeader(), true, outputSAMFile2);
          String str ="";

          long IACount = 0;
          long TotalCount = 0;


          for (SAMRecord samRecord : inputSAM) {
              str ="";

              if ((!samRecord.getReadUnmappedFlag())
                      && samRecord.getSecondOfPairFlag()) {
                  //System.out.println(samRecord.getReferenceName()+" "+samRecord.getAlignmentEnd());
                  if(samRecord.getReadNegativeStrandFlag())
                  str = fastadao.getSequence(samRecord.getReferenceName(),samRecord.getAlignmentEnd(),true);
                  else if(!samRecord.getReadNegativeStrandFlag()){
                      str                                                          =
  fastadao.getSequence(samRecord.getReferenceName(),samRecord.getAlignmentStart(),false);
                  }
                  //System.out.println(str);
                  //str = fastadao.getSequence("chr1", 10024867, true);
                  //System.out.println(str);

              } else {
                  continue;
              }

              boolean flag = IAcheck(str,contThre, ratioThre,ratioRange,samRecord.getReadNegativeStrandFlag());
              if (flag){
                  IACount++;
                  outputSAM.addAlignment(samRecord);
                  flag = false;
                  //System.out.println(flag);
              } else {
                  outputSAM2.addAlignment(samRecord);
              }

              if(TotalCount%200000==0){
                  System.out.print(".");
                  if (TotalCount%4000000==0) System.out.println();
              }
              TotalCount++;
          }

          System.out.println("IA:" + IACount);
          System.out.println("Total 3′ Reads:" + TotalCount);

          inputSAM.close();
          outputSAM.close();
          outputSAM2.close();
          fastadao.close();
  }

  /**
    * @param seq
    * @param contthre         Continuous A threshold
    * @param ratiothre        Ratio threshold
    * @param ratiorange
    * @param strand
    * @return return true if is an internalA; false if not
    */
  public static boolean IAcheck(String seq, int contthre, int ratiothre,
                  int ratiorange, boolean strand) {
          String feature = "";
          if (strand) {
                  feature = "A";
          } else if (!strand) {
                  feature = "T";
          }

          StringBuilder contbldr = new StringBuilder();
          contbldr.append("(?i)^");
          contbldr.append(feature);
          contbldr.append("{");
          contbldr.append(contthre);
          contbldr.append(",}\\w*");

          StringBuilder ratiobldr = new StringBuilder();
          ratiobldr.append(feature);
          ratiobldr.append(feature.toLowerCase());

          if (seq.matches(contbldr.toString())) {
                  return true;
          } else {
                  seq = seq.substring(0, ratiorange);
                  int counter = 0;
                  for (int i = 0; i < seq.length(); i++) {
                          if (ratiobldr.toString()
                                          .contains(String.valueOf(seq.charAt(i)))) {
```

```
                counter++;
            }
        }
        if (counter >= ratiothre) {
            return true;
        } else
            return false;
    }
  }

}
```

MiRCompare.java

```
package edu.marquette.biology.andersonlab.miRComp;
import htsjdk.samtools.SAMFileReader;
import htsjdk.samtools.SAMRecord;

import java.io.BufferedReader;
import java.io.File;
import java.io.FileReader;
import java.io.FileWriter;
import java.util.ArrayList;
import java.util.Hashtable;
import java.util.Iterator;

import edu.marquette.biology.andersonlab.domain.BedRecord;
import edu.marquette.biology.andersonlab.logic.QuerySAM;
import edu.marquette.biology.andersonlab.tabparser.Gff3impl;
import edu.marquette.biology.andersonlab.tabparser.tabParser;

/**     main program for intersecting the BAM data.
 * @author Fengchao
 *
 */
public class MiRComp {

  public static void main(String[] args) {
    // fixed for testing purpose
    //File resultfile = new File("C:/Users/Fengchao/Desktop/IASearch/control2_20141130_02.txt");
    File resultfile = new File("C:/Users/Fengchao/Desktop/IASearch/SKIV2L22_sorted_F_2_20141201.txt");
    File bedfile = new File("C:/Users/Fengchao/Desktop/IASearch/mmu_mirbase_5p.gff3");
    //File samfile = new File("C:/Users/Fengchao/Desktop/IASearch/control2_sorted_F_2.bam");
    File samfile = new File("C:/Users/Fengchao/Desktop/IASearch/SKIV2L22_sorted_F_2_20141201.bam");
    try {
            miRCompare(samfile, bedfile, resultfile);
        } catch (Exception e) {
            // TODO Auto-generated catch block
            e.printStackTrace();
        } finally {
        }
    }

    /**
     * @param samfl SAM/BAM input
     * @param bedfl BED input
     * @param resultfl output result file
     * @throws Exception
     */
    public static void miRCompare(File samfl, File bedfl, File resultfl)
            throws Exception {

        SAMFileReader samfilereader = new SAMFileReader(samfl);
        // need to integrate this part into parser
        FileReader in = new FileReader(bedfl);
        BufferedReader bedbf = new BufferedReader(in);
        tabParser gffinput = new Gff3impl(bedbf);

        FileWriter resultoutput = new FileWriter(resultfl);

        Hashtable<String, ArrayList<Double>> resulttbl = new Hashtable<String, ArrayList<Double>>();

        Hashtable<String, ArrayList<String>> readstbl = new Hashtable<String, ArrayList<String>>();

        BedRecord bedrecord = null;
        BedRecord record = new BedRecord();

        // ArrayList<SAMRecord> arrayl = new ArrayList<SAMRecord>();

        while ((bedrecord = gffinput.getNextRecord()) != null) {

//          System.out.println(bedrecord.getStart());
//          System.out.println(bedrecord.getChrom());
//          System.out.println(bedrecord.getName());

            ArrayList<SAMRecord> arrayl = QuerySAM.querySam(bedrecord, samfilereader,
                    10, 0, true, true);
            double distance = 0;
            if (arrayl.size()>0) {
```

```
        for (SAMRecord samrecord : arrayl) {
            distance += calcDistance(samrecord, bedrecord);
            // construct read-miRNA table
            if (readstbl.containsKey(samrecord.getReadName())) {
                readstbl.get(samrecord.getReadName()).add(
                        bedrecord.getName());
            } else {
                ArrayList<String> namearrltemp = new ArrayList<String>();
                namearrltemp.add(bedrecord.getName());
                readstbl.put(samrecord.getReadName(), namearrltemp);
            }
        }
    }
    //System.out.println(distance);

    if (arrayl.size() != 0) {
        distance = distance / arrayl.size();
    }
    //prepare result table
    ArrayList<Double> data = new ArrayList<Double>();
    data.add((double) arrayl.size());
    data.add(distance);
    resulttbl.put(bedrecord.getName(), data);
}
// calculate adjusted miRNA count
Iterator<ArrayList<String>> readsit = readstbl.values()
        .iterator();
while (readsit.hasNext()) {
    ArrayList<String> namearry = readsit.next();
    Iterator<String> mirit = namearry.iterator();
    mirit.next();
    while (mirit.hasNext()) {
        String miname = mirit.next();
        double counttemp = resulttbl.get(miname).get(0);
        counttemp = counttemp - 1;
        resulttbl.get(miname).set(0, counttemp);
    }
}
//System.out.print(bedrecord.getName());
// System.out.print(" ");
// System.out.println(arrayl.size());
// System.out.print(" ");
// System.out.println(distance);
// System.out.println("--------------");
```

```
    Iterator<String> resultit = resulttbl.keySet().iterator();
    while (resultit.hasNext()) {
        String namekey = resultit.next();
        StringBuilder sb = new StringBuilder();
        sb.append(namekey);
        sb.append("\t");
        sb.append(resulttbl.get(namekey).get(0));
        sb.append("\t");
        sb.append(resulttbl.get(namekey).get(1));
        sb.append("\r\n");
        resultoutput.write(sb.toString());

    }

    // Test code
    // record.setChrom("chrX");
    // record.setStart(53053755);
    // record.setEnd(53054349);
    // record.setStrand("-");
    // ArrayList<SAMRecord> arrayl= QuerySAM.querySam(record, saminput, 0,
    // 0,true,true);
    // for(SAMRecord samrecord:arrayl){
    // System.out.println(samrecord.getReadName());
    // }
    // System.out.print(arrayl.size());
    // System.out.print(arrayl.get(1).getFirstOfPairFlag());
    // System.out.print(arrayl.get(1).getReadName());
    resultoutput.close();
    bedbf.close();
    in.close();
}

/**
 * calculate the average distance of each read to the drosha clevage site
 *
 * @param samr
 * @param bedr
 * @return
 */
public static int calcDistance(SAMRecord samr, BedRecord bedr) {
    int dist = 0;
    if (bedr.getStrand().equals("+")) {
        dist = samr.getAlignmentEnd() - bedr.getStart();
```

```
            }
            if (bedr.getStrand().equals("-")) {
                    dist = bedr.getEnd() - samr.getAlignmentStart();
            }
            return dist;
    }

}
```

QuerySam.java

```
package edu.marquette.biology.andersonlab.logic;
import htsjdk.samtools.SAMFileReader;
import htsjdk.samtools.SAMRecord;
import htsjdk.samtools.util.CloseableIterator;

import java.util.ArrayList;
import java.util.Iterator;

import edu.marquette.biology.andersonlab.domain.BedRecord;

/**
  * @author Fengchao
  * An utility class to query the SAM Record.
  */
public class QuerySAM {

  /**
    * @param record    a BedRecord to provide the coordinates for query
    * @param samin    input SAM file for query
    * @param offs    offset bp for the start region. - means upstream and + means downstream
    * @param offe    offset bp for the end region
    * @param tpr_flag    three prime reads selection
    * @param strandFilter_flag      strand specific selection
    * @return
    */
  public static ArrayList<SAMRecord> querySam(BedRecord bedrcd, SAMFileReader samrdr, int offs, int offe, boolean tpr_flag, boolean strand_flag){
            CloseableIterator<SAMRecord> iter = null;
            ArrayList<SAMRecord> samrlist = new ArrayList<SAMRecord>();
            SAMRecord samrtemp = null;
            int newstart=0;
            int newend=0;
            //                  record.changeStartEnd(offs, offe); disabled due to calculation of distance
```

```
        //change offset for downstream and upstream query
//      try {
//          bedrcd.printall();
//      } catch (IllegalArgumentException | IllegalAccessException e) {
//          // TODO Auto-generated catch block
//          e.printStackTrace();
//      }
        if(bedrcd.getStrand().equals("+")){
            newstart=bedrcd.getStart()-offs;
            newend=bedrcd.getStart()+offs;
        }else if (bedrcd.getStrand().equals("-")){
            newstart=bedrcd.getEnd()-offs;
            newend=bedrcd.getEnd()+offs;
        }//this section is really tricky


        iter = samrdr.query(bedrcd.getChrom(), newstart, newend, false); //If true, each SAMRecord returned is will have its alignment completely contained in the interval of interest. If false, the alignment of the returned SAMRecords need only overlap the interval of interest.
        if (iter.hasNext()) {
            while (iter.hasNext()) {
                samrlist.add(iter.next());
            }
            iter.close();

            if (tpr_flag == true) {
                Iterator<SAMRecord> it = samrlist.iterator();
                while (it.hasNext()) {
                    if (it.next().getFirstOfPairFlag() == true) {
                        it.remove();
                    }
                }
            }

            if(strand_flag==true){
                Iterator<SAMRecord> it = samrlist.iterator();
                if(bedrcd.getStrand()=="+"){
                    while(it.hasNext()){
                        if(it.next().getReadNegativeStrandFlag()==false){
                            it.remove();
                        }
                    }
                } else if(bedrcd.getStrand()=="-"){
                    while(it.hasNext()){
```

```
                    if(it.next().getReadNegativeStrandFlag()==true){
                        it.remove();
                    }
                }
            }
        }

    }
    iter.close();
    return samrlist;
}
```