Quantitative Biology 2017, 5(4): 352–367
DOI 10.1007/s40484-017-0123-4

RESEARCH ARTICLE

Cap-seq reveals complicated miRNA transcriptional mechanisms in *C. elegans* and mouse

Jiao Chen[1], Dongxiao Zhu[2] and Yanni Sun[1,*]

[1] Department of Computer Science and Engineering, Michigan State University, East Lansing, MI 48824, USA
[2] Department of Computer Science, Wayne State University, Detroit, MI 48202, USA
* Correspondence: yannisun@msu.edu

Received February 26, 2017; Revised May 16, 2017; Accepted June 21, 2017

***Background*: MicroRNAs (miRNAs) regulate target gene expression at post-transcriptional level. Intense research has been conducted for miRNA identification and the target finding. However, much less is known about the transcriptional regulation of miRNA genes themselves. Recently, a special group of pre-miRNAs that are produced directly by transcription without Drosha processing were validated in mouse, indicating the complexity of miRNA biogenesis.**

***Methods*: In this work, we detect clusters of aligned Cap-seq reads to find the transcription start sites (TSSs) for intergenic miRNAs and study their transcriptional regulation in *Caenorhabditis elegans* and mouse.**

***Results*: In both species, we have identified a class of special pre-miRNAs whose 5′ ends are capped, and are most probably generated directly by transcription. Furthermore, we distinguished another class of special pre-miRNAs that are 5′-capped but are also part of longer primary miRNAs, suggesting they may have more than one transcription mechanism. We detected multiple cap reads peaks within miRNA clusters in *C. elegans*. We surmised that the miRNAs in a cluster may either be transcribed independently or be re-capped during the microprocessor cleavage process. We also observed that H3K4me3 and Pol II are enriched at those identified miRNA TSSs.**

***Conclusions*: The Cap-seq datasets enabled us to annotate the primary TSSs for miRNA genes with high resolution. Special class of 5′-capped pre-miRNAs have been identified in both *C. elegans* and mouse. The capping patter of miRNAs in a cluster indicate that clustered miRNA transcripts probably undergo a re-capping procedure during the microprocessor cleavage process.**

Keywords: miRNA; Cap-seq; transcriptional regulation

INTRODUCTION

MicroRNAs (miRNAs) are a large family of ~21 nucleotide-long RNAs that have been uncovered as key regulators of gene expression at post-transcriptional level in metazoans, plants, and viruses [1–3]. In metazoans, mature miRNAs and argonaute (AGO) proteins form into the miRNA-induced silencing complex (miRISC), within which miRNAs base-pair to the 3′-UTR of target mRNAs and inhibit protein synthesis by either repressing translation or promoting mRNA degradation. It was inferred that more than one-third of all protein-coding genes are regulated by miRNAs [4]. miRNAs have also been discovered to play a crucial role in precision medicine. Precision medicine attempts to characterize the genetic background of patients and classify them into subpopulations that differ in their susceptibility to a particular disease [5–7]. The capability of modulating a vast number of protein-coding genes makes miRNA powerful regulators of the different cellular processes involved in the pathogenesis of various types of diseases, including cardiovascular diseases and cancer. For example, liver miRNA miR-122, as the most abundant and most specific liver miRNA, is most likely to represent a novel biomarker for cardiovascular and metabolic diseases as it plays a central role in lipid and glucose homeostasis and is detectable in serum and plasma [8]. Differential expression of miRNAs has also been observed in tumor

tissues. Their alteration expression in prostate cancer has been well documented [9]. Because of their important regulatory functions, many studies have focused on miRNA annotation and their targets finding [10–12]. However, how miRNAs themselves are expressed and regulated is not fully understood.

In the canonical miRNA biogenesis pathway, miRNAs are processed from longer transcripts, which are referred to as primary miRNAs (pri-miRNAs) [2]. Pri-miRNAs are either transcribed by polymerase II (Pol II) from independent genes or derived from the introns of protein-coding genes [13,14]. Two members of the RNase III family of enzymes, Drosha and Dicer, further process pri-miRNAs to mature miRNAs [15–17]. First, Drosha cleaves the hairpin structure of a pri-miRNA to a ~70-nucleotide precursor miRNA (pre-miRNA) in the nucleus. Pre-miRNAs are then exported to the cytoplasm by XPO5, where Dicer cleaves off the loop region of the hairpin and further processes it to ~21 bp mature miRNA(s). Recent studies have uncovered several non-canonical ways of generating miRNAs, demonstrating the complexity of miRNA biogenesis. One class of unconventional miRNAs is called mirtrons, which are encoded in introns, bypass Drosha processor but rely on splicing machinery for pre-miRNA generation [18,19]. miRNAs in mammals have been shown to frequently utilize alternative promoters in different cell types, and pri-miRNAs may encode subsets of clustered miRNAs [20]. Pri-miRNA transcripts can be cleaved by cytoplasmic Drosha in human cells [21]. Another study on mice has uncovered a second class of non-canonical miRNAs, of which the pre-miRNAs are 5′-capped and generated directly by transcription [22].

Although the genomic coordinates of mature and precursor miRNAs have been annotated in databases such as miRBase [23], very little is known about the coordinates of pri-miRNAs. RNA-seq technology [24] has been proved as an efficient way to annotate protein-coding genes. Mature mRNAs contain a 5′ 7-methylguanosine (m^7G) cap and a long 3′ polyadenylated (poly(A)) tail and are relatively stable, so they can be well extracted from cells and sequenced. The sequenced RNA fragments are then mapped to the reference genome for gene annotation. However, since the original 5′ ends of primary miRNA transcripts are rapidly cleaved off by Drosha during miRNA maturation, regular RNA-seq technology cannot be used to find the primary TSSs of miRNA genes. Pri-miRNAs are usually transcribed by Pol II and also contain a 5′ m^7G cap and a 3′ poly(A) tail [13], indicating that the biological features related to Pol II transcription can be used to identify the transcription initiation sites for miRNA genes.

To identify the primary TSSs of miRNAs, some computational methods have been implemented based on features related to Pol II transcribed genes, such as transcription factor binding sites (TFBSs), Pol II binding, and chromatin states including histone modifications and nucleosome positioning [25–28]. Typically, Pol II and H3K4me3 are highly enriched at active promoters, while nucleosomes are depleted at the TSSs. Wang *et al*. [29] designed a statistical model to mimic Pol II binding patterns at the promoters of highly expressed protein-coding genes and used it to search for similar Pol II binding patterns upstream of all intergenic miRNAs in human breast cancer cells to identify primary promoters. They verified their findings by checking the conservation, CpG content, and activating histone marks in the identified promoter regions. Ozsolak *et al*. combined nucleosome mapping with ChIP-chip screens for H3K4me3, H3K9/14ac, Pol II and Pol III signatures to identify the proximal promoter regions of pri-miRNAs in human genome [26]. They tested their algorithm on human annotated protein-coding genes and predicted the transcription initiation regions to a resolution of 150 bp. With the same method, the transcription initiation regions of 175 transcriptionally active miRNAs were determined. Saini *et al*. [30] predicted the 5′ ends of intergenic pri-miRNAs in human, mouse and rat genomes by combining the features of TSSs predictions, CpG islands and 5′ cap analysis of gene expression (CAGE) tags. miRStart [28] built an SVM model using the features of CAGE tags, TSS Seq libraries and H3K4me3 chromatin signature from ChIP-seq to identify the TSSs of human miRNAs. The model was trained on 7,268 protein-coding genes with unique TSS and identified 847 putative TSSs for the 940 human pre-miRNAs obtained from miRBase.

While the methods discussed above [26,29,30] have predicted TSSs for miRNA genes in mammal genomes, their prediction results have low resolution (hundreds of bps) because the typical distribution patterns of Pol II and chromatin features surrounding promoters may not hold for any particular gene. Even for some actively transcribed genes (29/85), the distance between the TSS and the closest Pol II peak can be over 1,000 bp (Supplementary Figure S1B). A more accurate method is to take advantage of the cap structure at 5′ ends of Pol II transcribed RNAs. Mapping the capped sequences to reference genomes will enable direct discovery of the TSSs. CAGE and Cap-seq have been used to directly sequence RNAs with 5′ m^7G caps, which are used to identify the candidate TSSs. CAGE [31] uses a so-called "cap trapper" method to capture full-length mRNAs and sequence the 5′ ends with Sanger sequencing technology. However, CAGE is not widely used to map TSSs for each gene because of the cost and sequencing depth. With the development of high throughput sequencing technology, enriching capped RNA transcripts followed by next-generation sequencing (NGS) technology (Cap-seq or

deepCAGE [32]) has been used to sequence the capped RNAs in the whole genome.

The mouse is a popular mammalian model system for genetic research, for which some Cap-seq datasets have been generated [22,33]. Recently the Cap-seq study in mice [22] has uncovered a non-canonical way of generating pre-miRNAs, in which the pre-miRNAs are generated directly by transcription and their 5′ ends are m^7G capped. These 5′ m^7G capped pre-miRNAs prefer to be exported from the nucleus to the cytoplasm by exportin 1 (XPO1) and after Dicer processing, only 3p-miRNA is efficiently loaded onto the AGO complex. This special class of 5′-capped pre-miRNAs have also been discovered in the human genome (miR-320a) [22], but whether they also exist in other non-mammalian species is still unknown.

Caenorhabditis elegans is also a well established model organism for genomic studies. The worm is a simple multicellular organism but with a variety of tissue types and a short life cycle [34]. Therefore, many functional genomic sequencing datasets, including Cap-seq [33,35,36], have been generated on this species. The transcription regulation in this animal is quite different from that in mammals. For example, the primary transcripts of about 70% of its protein-coding genes undergo trans-splicing [37]; and its pri-miRNA transcripts are exported by XPO1 and possibly processed in nuclear pore [38].

In this study, we utilized available Cap-seq datasets to study the transcription regulation of miRNAs in *C. elegans* and mouse. The main results are summarized below.

- We identified a group of candidate 5′ m^7G capped pre-miRNAs in *C. elegans*.
- We classified another class of miRNAs with non-canonical transcription mechanisms, for which the pre-miRNAs may be generated by both the canonical miRNA pathway with Drosha and the non-canonical pathway without Drosha.
- Based on the capping signals for miRNA genes in clusters, we proposed a hypothesis that these pri-miRNA transcripts might undergo cytoplasmic re-capping during the pre-miRNAs generation process.
- We developed a method to separate these identified primary miRNA promoters as broad or divergent and characterized them by analyzing the H3K4me3 and Pol II binding surrounding them.

RESULTS

Identification of primary miRNA TSSs in *C. elegans* and mouse

Overview of primary miRNA TSSs annotation

We used the single-linkage clustering method to detect transcription initiation clusters (TICs) [35] from Cap-seq data in *C. elegans* and mice. For each TIC, we also used a Poisson distribution to model the local background noise and test whether it is significantly enriched with Cap-seq reads by calculating a *p*-value (see Materials and Methods). The intronic miRNAs are usually processed as part of their host-gene mRNA [3] and thus their transcriptions coordinate with the protein-coding genes. Therefore, in this study, we have focused on identifying the primary TSSs for intergenic miRNAs. In both species, we first identified the intergenic miRNAs that are not covered by protein-coding genes, non-coding RNA genes, small nucleolar RNA (snoRNA) genes, or small nuclear RNA (snRNA) genes. Then the flanking region between 5′ end of pre-miRNA and the closest upstream gene was searched for TICs. The identified TICs were annotated as candidate primary TICs (Figure 1). The region within pre-miRNA was also searched for TICs, and the identified TICs were annotated as pre-cap TICs (Figure 1). Moreover, we also searched for miRNA clusters from miRBase with a distance threshold of 1,000 bp. For miRNA clusters, primary TSS(s) are annotated as the TIC(s) upstream of the 5′ end of the first pre-miRNA in the cluster and pre-cap TIC(s) are annotated as the TIC(s) within the pre-miRNAs in the cluster.

In *C. elegans*, we identified 134 intergenic miRNAs and 16 miRNA clusters. With the retrieved Cap-seq data [35], 70 intergenic miRNAs were identified with at least one candidate primary TIC in the flanking regions, and 9 miRNAs were identified with at least one pre-cap TIC inside of the precursors. For those miRNAs with candidate primary TICs or pre-cap TICs, 8 were identified with both of them. The modes (highest coverage of reads' 5′ ends) of the primary TICs are used to represent the candidate TSSs of miRNAs. In the 16 miRNA clusters, 6

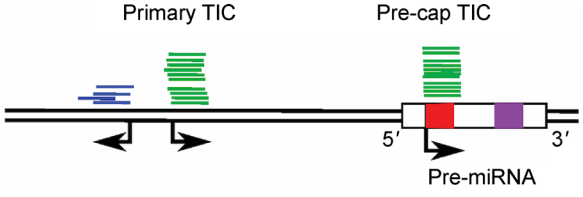

Figure 1. Primary TIC and pre-cap TIC. Primary TIC is located upstream of the pre-miRNA, while pre-cap TIC is located inside of the pre-miRNA. The annotation of a pre-miRNA is usually a stem-loop that includes the pre-miRNA and the lower stems. However, the real pre-miRNA only includes the red, purple sequences (mature miRNAs) and the loop between them. Therefore, the pre-cap TIC starts from the 5′ end of a 5p-miRNA. Each blue or green bar corresponds to a mapped read, where green indicates the plus strand and blue the minus strand. The Cap-seq datasets were sequenced in a strand-specific way. Only the reads on the same strand of the miRNA are considered.

were found to contain both primary TSSs and pre-cap TICs, and 2 clusters were found to contain only primary TSSs (Supplementary Table S1).

In mice, we used the Cap-seq data from the study by Xie *et al*. [22]. They performed Cap-seq to find the unconventional pre-miRNAs whose 5′ ends are generated directly by transcription initiation with Pol II and thus 5′ m^7G capped. Based on their results, there are the highest number of 5′-capped pre-miRNAs on chromosome 7, so we focused on identifying the primary TSSs for miRNAs within mouse chromosome 7. We identified 80 intergenic miRNAs on this chromosome, of which 10 miRNAs were identified with at least one pre-cap TIC and 37 miRNAs were identified with at least one candidate primary TICs. Of those miRNAs with candidate primary TICs or pre-cap TICs, 2 were found to contain both of them (Supplementary Table S2).

Comparison with previous work

In the previous study [33], Gu *et al*. developed the CapSeq protocol to enrich and sequence longer (70–90 nt) 5′-capped RNA transcripts, and CIP-TAP cloning to isolate and sequence 5′-capped small (18–40 nt) RNAs. They applied these two 5′ anchored RNA deep-sequencing approaches onto *C. elegans* and mouse genomes and annotated the primary TSSs for miRNAs in both of them. As a result, they identified at least one TSS for 55 individual pre-miRNAs and 9 miRNA clusters in *C. elegans*, and 134 individual pre-miRNAs in mice, with 7 miRNAs annotated on chromosome 7. Comparing with their results, we identified the primary TSSs for 24 overlapping miRNAs, 6 overlapping miRNA clusters and 41 additional miRNAs, 2 additional miRNA clusters in *C. elegans* and 37 additional miRNAs on mouse chromosome 7 (Supplementary Table S3).

In another study [36], Kruesi *et al*. devised a global run-on cap sequencing (GRO-cap) method to capture and sequence only those 5′ m^7G capped RNAs in *C. elegans* embryos, starved L1 larvae, and L3 larvae. With the GRO-cap sequencing data, they annotated the primary TSSs for 52 individual pre-miRNAs and 5 miRNA clusters. We identified the primary TSSs for 21 overlapping miRNAs and 5 overlapping miRNA clusters. Furthermore, we also annotated the TSSs for 46 more individual miRNAs and 3 more miRNA clusters (Supplementary Table S4).

To better evaluate the primary TSSs identified, we also calculated the coordinate differences between our results and the work from Gu *et al*. and Kruesi *et al*. For miRNAs with multiple candidate TSSs, we only kept the minimum coordinate difference. The results were shown in Supplementary Table S3 and Table S4. For most of the overlapping miRNAs or miRNA clusters, the coordinate differences are less than 100 bp, validating our results.

5′ m^7G capped pre-miRNAs are identified in *C. elegans*

As shown in Figure 1, reads in pre-cap TICs are usually aligned very well with the 5′ ends of pre-miRNAs. There is a possibility that these reads were actually sequenced from uncapped pre-miRNAs rather than capped RNAs. In this section, we first investigate whether usually uncapped pre-miRNAs are highly enriched by Cap-seq protocol.

Possible 5′ recessed RNAs enriched by Cap-seq

To enrich for capped RNA in Cap-seq experiments for *C. elegans*, exonuclease terminator and calf intestinal alkaline (CIP) were used to remove the uncapped RNAs [35]. However, some uncapped RNAs, such as pre-miRNAs and tRNAs, may not be accessed efficiently by terminator/CIP because of their 5′ recessed ends in their secondary structures. As a result, some of the Cap-seq reads in pre-cap TICs may actually come from pre-miRNAs. To investigate the contamination of pre-miRNA reads in Cap-seq data, we downloaded small RNA-seq data (GSM916519) for *C. elegans* and mapped the reads to the reference genome as the control. Those miRNAs that are detected as expressed by small RNA-seq data might be observed in Cap-seq data as well. We then quantified the number of Cap-seq reads aligned to those expressed pre-miRNAs (mapped with sufficient small RNA-seq reads) (Supplementary Figure S2A). The results showed that the numbers of Cap-seq reads mapped to pre-miRNAs are not proportional to the numbers of small RNA-seq reads mapped. For many highly expressed miRNAs (22/33), there are few or no Cap-seq reads aligned to their precursors, indicating that many Cap-seq reads may not be pre-miRNAs. In addition, pre-miRNAs, serving as an intermediate during miRNA maturation, are quickly processed by Dicer and thus tend not to be enriched by Cap-seq. Previous work has shown that the data of carefully designed pre-miRNA sequencing only contains less than 1% reads that can be mapped to pre-miRNAs [39].

According to Chen *et al*. [35], there was no step to remove tRNAs in the Cap-seq protocol. We then did the similar comparison for tRNAs because tRNAs may escape treatment of the terminator and CIP for the same reason as pre-miRNAs. The results showed that, although almost all the annotated tRNA genes in the worm were expressed (604/605), most of these tRNAs (463/605) do not have any Cap-seq reads mapped (Supplementary Figure S2B). A two tailed paired sample *t*-test was used to

estimate the mean difference between the normalized Cap-seq reads and small RNA-seq reads mapped to tRNAs. We got a *p*-value as 1.12e-185, showing that the tRNA reads captured by two protocols are significantly different (alpha level as 0.01). Pearson's correlation and Spearman's rank correlation between two mapping results were also calculated. The correlation coefficient results (-0.042 and -0.125, respectively) suggest that they are poorly correlated. Most of those tRNAs mapped with Cap-seq reads have less than 10 reads (124/142, Supplementary Figure S3), implying that these reads might be caused by random contamination. We then used the Poisson distribution to model the background noise (see Materials and Methods). 31 tRNAs were reported as significantly enriched with Cap-seq reads at the cutoff of 10^{-5}. Those tRNAs that are highly enriched for Cap-seq reads are usually overlapped with the repeat regions. Considering that most of the other tRNAs have few or no Cap-seq reads mapped, we infer that these regions might be transcribed by Pol II and produce capped RNAs under certain conditions.

We also suspected that some of the Cap-seq reads within pre-miRNAs might be mature miRNAs. However, mature miRNAs are short and usually do not possess complex secondary structures. They could be efficiently removed by terminator/CIP treatment [40]. The deficit of Cap-seq reads on most mature miRNAs adds support to this. Hence the pre-cap TICs are not likely formed by mature miRNAs either. The Cap-seq study on mice have also shown that uncapped RNAs constitute less than 10% of the total sequenced reads [22]. Considering the above analysis together, we posited that although Cap-seq inevitably contains some uncapped RNAs, many of the reads mapped to pre-miRNAs are likely sequenced from capped RNAs.

Defining 5′ m^7G capped pre-miRNAs with pre-cap TICs

RNAs synthesized by Pol II are 5′ m^7G capped cotranscriptionally. A previous study [22] has documented a new class of unusual miRNAs in new-born mice, for which the 5′ ends of the pre-miRNAs are m^7G capped and coincide with their TSSs. These miRNAs were suggested to be generated without Drosha processing, with their 5′ ends determined directly by transcription initiation and the 3′ ends generated by transcription termination.

To examine whether there are the same class of miRNAs in *C. elegans*, we analysed those pre-miRNAs with pre-cap TICs mapped inside. The 5′ ends of pre-cap TICs are usually consistent with the 5′ ends of pre-miRNAs or 5p-miRNAs. Therefore, those pre-miRNAs that were detected with pre-cap TICs should acquire m^7G cap at their 5′ ends and were annotated as 5′-capped miRNAs. We also applied our method to mouse Cap-seq data obtained from the study by Xie *et al*. [22] and compared our results with theirs. Our results have uncovered all the 9 intergenic 5′-capped pre-miRNAs on mouse chromosome 7 as shown in the paper, validating our method. Besides, we also annotated the primary TSSs for 37 miRNAs with the same dataset. Strikingly, new candidate TSSs were also found upstream of two 5′ capped pre-miRNAs (mmu-mir-344c and mir-344i). In total, we identified 9 5′-capped miRNAs in C. elegans (Table S1) and 10 on mouse chromosome 7 (Supplementary Table S2).

We also used statistical analysis to evaluate the enrichment of capped RNA reads on the 5′ ends of pre-miRNAs. With the calculated *p*-values, 7/9 5′-capped pre-miRNAs in *C. elegans* and 9/10 in mice are significantly enriched with Cap-seq reads (p-value $< 10^{-5}$; Supplementary Table S1).

To look for sequence motifs surrounding these identified putative miRNA TSSs, we plotted the nucleotide composition around them by Weblogo [41]. A strong YR motif was observed at both the putative primary miRNA TSSs or pre-cap TSSs of independent miRNAs, in which Y represents pyrimidine, R represents purine and R locates at the TSSs (+1 position, Figure 2).

M^7G capped pre-miRNAs often have upstream primary TICs

In both *C. elegans* and mice, we noticed that some of these 5′-capped pre-miRNAs also have primary TICs in the region from the pre-miRNA 5′ end to the closest upstream gene. In *C. elegans*, 8 out of 9 5′-capped pre-miRNAs have been detected with primary TICs; while in mice, 2 out of 10 5′-capped pre-miRNAs have upstream primary TICs. These upstream TICs are usually not very far from the pre-miRNAs and the Pol II binding at these TICs elongate to the downstream miRNAs, indicating that they are connected with the miRNAs. To our knowledge, this phenomenon has not been described in other studies. Two examples in *C. elegans* are shown in Figure 3. Since pre-miRNAs can be generated both in the canonical way as from a long pri-miRNA by Drosha and in the non-canonical way by transcription initiation and termination, we ask which TIC corresponds to the real primary TSS of the miRNA or can both of them produce pre-miRNAs?

We proposed two possible explanations for this phenomenon. The first explanation is that there could be multiple isoforms for these miRNA genes. For example, the second discovered miRNA in *C. elegans* (let-7) has been detected with at least three primary transcripts [43]. It was also reported that genes often use alternative promoters in a developmental stage or cell type specific

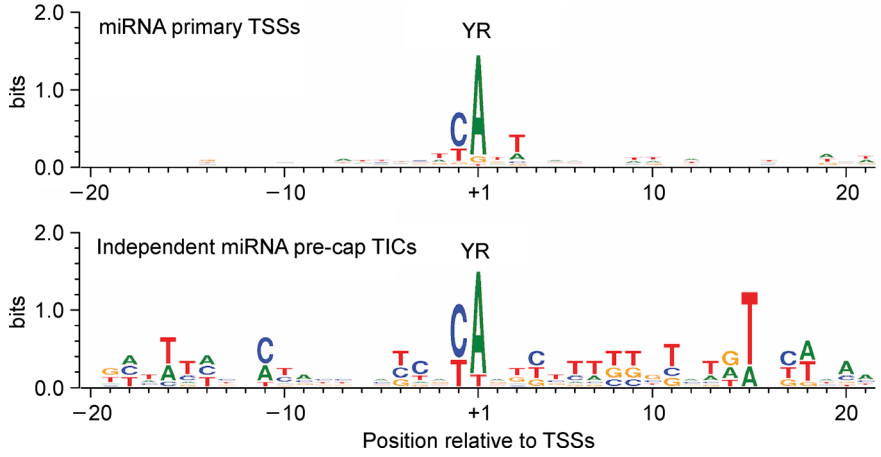

Figure 2. Sequence motif analysis of nucleotides around putative miRNA TSSs (+1). The nucleotide height (in bits) stands for the log2 ratio of the observed nucleotides frequency relative to the background genomic nucleotide composition. The YR motif is observed at the putative primary miRNA TSSs and independent miRNA pre-cap TSSs.

way that can spread up to thousands of kilobases. As an example, about one half of the protein-coding genes in human and mouse genomes have multiple alternative promoters [44]. Therefore, these miRNA genes may also contain mulitple alternative promoters that can generate several isoforms (Figure 4A). The other primary TICs may produce longer miRNA transcripts that are subject to the canonical miRNA processing procedures involving Drosha. Since the 5′-capped pre-miRNAs are most likely generated directly by transcription, two paths may be able to lead to the maturation of the same miRNA: one with Drosha and the other without. Because we used the dataset from mixed-stage embryos, these miRNA genes may employ alternative promoters and produce diverse isoforms at different stages. We also observed that for some miRNAs (Supplementary Figures S5–S7: cel-mir-235, cel-mir-244, cel-mir-238, and cel-mir-228), the upstream TICs are very close to the pre-miRNA, likely that they are generated from the same promoter as the pre-cap TIC.

The second explanation is that the TICs upstream of pre-miRNA may be transcribed enhancers or promoters,

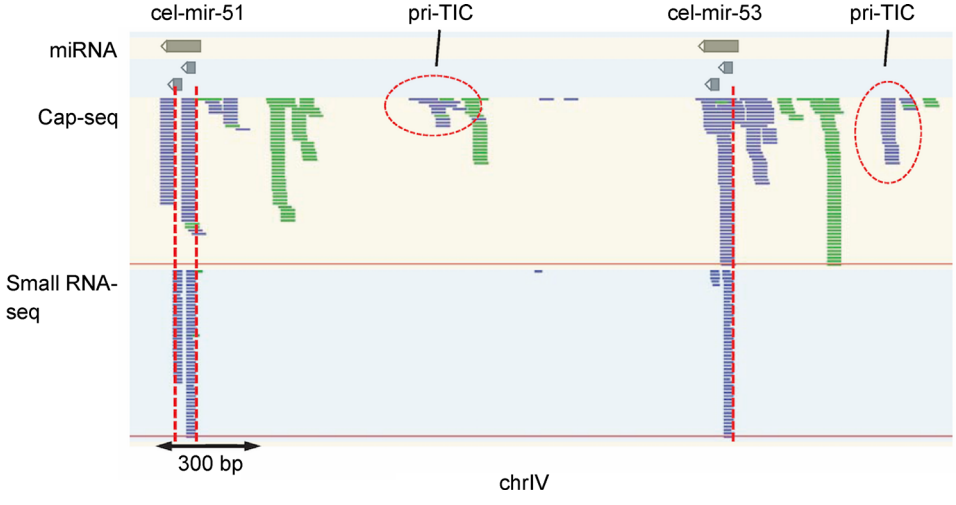

Figure 3. Primary TICs are detected upstream of 5′-capped miRNAs cel-mir-51 and cel-mir-53. Multiple Cap-seq peaks are observed in pre-miRNA regions. The mapped reads were visualized by GenomeView [42]. Each blue or green bar corresponds to a mapped read, where green indicates the plus strand and blue the minus strand. The reads in upper panel are from Cap-seq dataset, with uniform length of 36 nt. The reads in lower panel are from small RNA-seq dataset, with length in range from 14 nt to 26 nt.

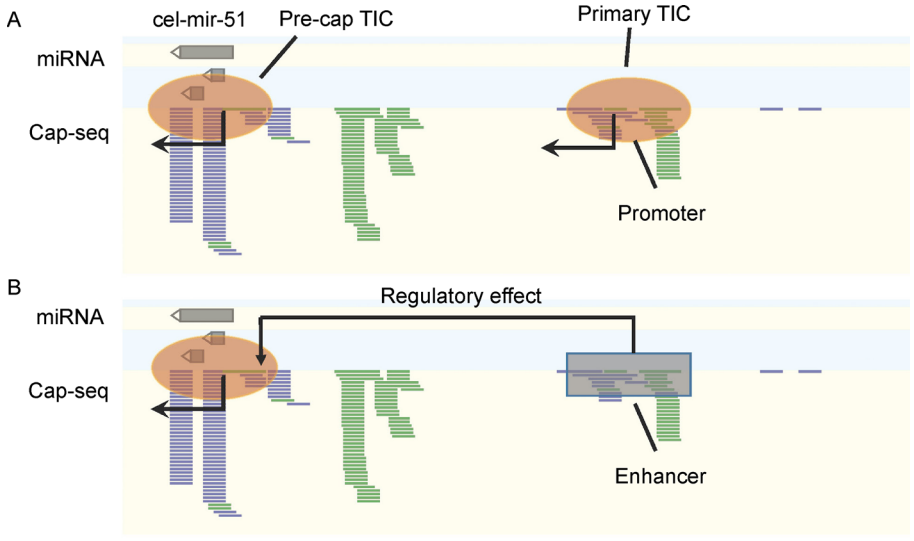

Figure 4. Upstream primary TICs also exist for m^{7}G capped pre-miRNAs. (A) Two alternative promoters for the same miRNA. Both of them are able to generate the transcripts that produce the same mature miRNA. (B) The miRNA transcript is generated by the pre-cap TIC. Upstream TIC(s) correspond to transcribed enhancer(s).

which generate transcripts that will not produce miRNAs. Recently several studies have shown that promoter and enhancer regions can be transcribed in human, mouse and *C. elegans* genomes [35,45,46]. The transcription in promoters and enhancers are usually associated with downstream genes. It is suggested in *C. elegans* that the elongation from an upstream enhancer toward a downstream gene may have the potential to deliver Pol II to a proximal promoter, or alternatively function directly as a distal promoter [35]. Thus, the upstream TICs may be transcribed in the enhancer regions and have a regulatory effect on the downstream miRNA genes (Figure 4B). To verify this hypothesis, we then downloaded the enhancer annotations for *C. elegans* from database Super Enhancer Archive (SEA) [47] and calculated whether these enhancers overlap or are close to those TICs upstream of pre-capped miRNAs. The SEA database has 904 computationally or experimentally identified enhancers for *C. elegans*. The results showed that these annotated enhancers are usually far away (over 50,000 bp) to the upstream TICs. Therefore, the transcribed enhancer model is not well supported by the available enhancer annotations.

5p-miRNAs are produced from the identified m^{7}G capped pre-miRNAs

The m^{7}G capped pre-miRNAs have been reported to produce single 3p mature miRNAs in mice [22]. The main explanation is that the capped 5p-miRNA is not efficiently loaded onto Ago complex. However, we noticed that some identified 5′-capped pre-miRNAs in *C. elegans* also generate 5p mature miRNAs based on the annotations in miRBase (Supplementary Table S1). While there is a possibility that some of the Cap-seq reads mapped to pre-miRNAs may be uncapped due to contamination or technical artifacts, most of these identified m^{7}G capped pre-miRNAs are significantly enriched with capped RNA reads according to our previous analysis. We surmise that there might be alternative pathways for producing 5p-miRNAs from those 5′ m^{7}G capped pre-miRNAs. Similar observations were also made for identified 5′-capped pre-miRNAs in mice: four of them (mmu-mir-484, mmu-mir-1903, mmu-mir-344f and mmu-mir-344i) actually prefer to generate 5p mature miRNAs [22].

The 5p mature miRNAs could be generated by alternative primary TSSs. As many of these identified 5′-capped pre-miRNAs also have candidate primary TSSs, canonical primary miRNA transcripts may be generated from them and produce 5p-miRNAs. Studies have shown that mature miRNA selection from 5′ and 3′ strands of the same precursor is highly regulated and varies under different cell types, developmental stages and disease states [48,49]. Capped 5p mature miRNAs may be produced under specific conditions to promote its target gene's expression.

Multiple transcription initiation sites for miRNA clusters in *C. elegans*

miRNAs in a cluster are close to each other, usually coexpressed and transcribed as a single pri-miRNA

[3,50]. Previously each miRNA cluster in *C. elegans* has been annotated with one primary TSS using Cap-seq [33] or GRO-cap [36] datasets. Strikingly, the datasets we used here have shown that the TICs for miRNA clusters in *C. elegans* can have a broad distribution with multiple strong peaks across the whole cluster. We identified primary TICs for 8 clusters, of which 7 have TICs inside of the clusters. One example of cluster cel-mir-35 – 41 is shown in Figure 5. The similar phenomenon is also observed in individual pre-miRNAs, as shown in Figure 3, the Cap-seq signal within cel-mir-51 and cel-mir-53 also displays multiple strong peaks. We noticed that these capped reads peaks inside of clusters were located on both arms of pre-miRNAs, with 5p-peaks have the same 5′ ends of 5p-miRNAs and 3p-peaks start from the 3′ ends of 3p-miRNAs (Figure 5). As analyzed above, not many pre-miRNAs or mature miRNAs were kept in the Cap-seq experiments and many of the reads mapped are likely to be capped RNAs. This is further supported by the coordinate differences between the capped peaks and the miRNA/miRNA* on 3p arms. We proposed three hypotheses to explain this phenomenon.

5′ re-cap after post-transcriptional processing

It has been reported that mature long transcripts of both protein-coding mRNAs and long ncRNAs in human cells can be processed post-transcriptionally to yield small RNAs, which are then modified by the addition of a 5′-cap structure [40]. Later on, a cytoplasmic capping enzyme, which is able to add 5′-cap to the ends of cleaved RNAs was identified in murine erythroid and nonerythroid cells [51]. This cytoplasmic capping enzyme, together with a kinase, can transfer covalently bound GMP onto a 5′-monophosphate RNA to create a 5′-GpppX RNA, but it can not function on RNAs with 5′-hydroxyl ends [52]. This phenomenon is known as cytoplasmic capping, which has been found in both murine and human cells [51,53,54].

The well coordinated capped reads at 5p and 3p arms of pre-miRNA hairpins indicate that the exposed 5′ ends of the miRNA transcripts after Drosha cleavage may be re-capped: that is why the capped reads were observed at the 5′ ends of 5p miRNAs and the 3′ ends of 3p miRNAs. Certainly, in the canonical miRNA biogenesis, pre-miRNAs are produced from pri-miRNAs in the nucleus by Drosha. Therefore, the 5′ monophosphate ends at the cleavage sites may be re-capped by nuclear capping enzyme in the nucleus. However, pri-miRNAs in *C. elegans* are exported to the cytoplasm by XPO1 and be processed to pre-miRNAs either in the nuclear pore or in the cytoplasm [38]. Considering the discovery of cytoplasmic capping enzyme, the pre-miRNA re-capping is more likely to happen in the cytoplasm.

The processing of pri-miRNAs in the cytoplasm requires cytoplasmic miRNA processors. It has been shown that cytoplasmic RNA viruses which encode miRNAs were able to produce functional miRNAs in the cytoplasm of BHK-21 cells [55]. The processing of these virus-generated cytoplasmic pri-miRNAs also relies on Drosha but takes place in the cytoplasm [56]. Based on this discovery, although without direct experimental

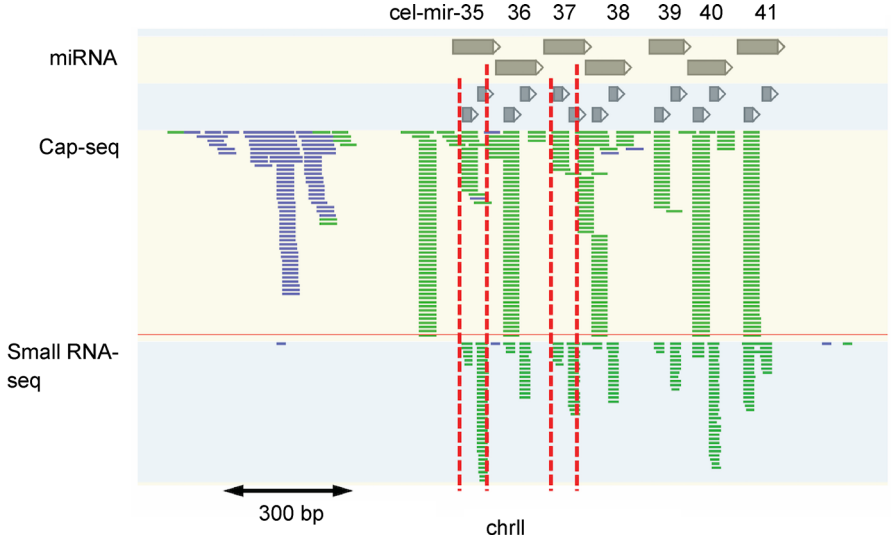

Figure 5. Capped TICs distribution in miRNA cluster cel-mir-35 – 41. Multiple strong capped peaks have been observed in pre-miRNAs in the cluster. As illustrated by the red dashed line, the Cap-seq peaks on the 5p arm have the same start position as the 5p-miRNAs, while the peaks on the 3p arm start from the end of the 3p-miRNAs. The length of Cap-seq reads is 36 nt.

verification, there is a possibility that the similar cytoplasmic miRNA processors involving Drosha also exist in *C. elegans* cells. All these findings support a new model of miRNA biogenesis in *C. elegans* in which pri-miRNAs are exported to the cytoplasm by XPO1, where they are cleaved by Drosha and further processed. Transcripts of miRNA clusters may undergo cytoplasmic re-capping during the cleavage process (Figure 6).

The cytoplasmic recapping of cluster miRNAs could produce capped 5p-miRNAs which may not be efficiently loaded on Ago. Indeed, from the annotations in miRBase, most of the clusters with capped TICs inside favorably generate 3p mature miRNAs. However, the clusters mir-41-44 and mir-86-8211 do prefer to generate 5p mature miRNAs. Here similar to the 5′-capped independent pre-miRNAs, the recapping may be a controlled process which only occur at certain cell types, developmental stages, or disease states. For example, with the capped short RNAs data from *C. elegans* young adult stage [33], we did not observe capped reads peaks inside of the miRNA clusters. The recapping could serve to regulate the strand selection on these miRNAs, suppressing the 5p-miRNAs and promoting the expression of their targets.

Multiple TSSs can be generated from the same promoter

Previous studies have shown that most core promoters do not have a single TSS, but multiple start sites that are closely located [57,58]. The broad TSS promoters in *C. elegans* are often enriched for CpG island, while sharp TSS promoters often have TATA-box. For some promoters, although TSSs are distributed over a large region, most transcription initiates at one specific nucleotide position [57]. Therefore, multiple capped peaks in individual pre-miRNAs or miRNA clusters may be generated from the broad promoter as multiple TSSs. To find evidence supporting this, we scanned the proximal promoters of these miRNA genes for motifs of TATA-box, Inr, DPE and BRE with position weight matrices derived from database JASPAR [59]. GC content and CpG number are also searched in the promoter regions (Table 1). We observe that the promoters of these miRNA clusters and individual 5′-capped miRNAs are usually GC rich: the GC content on each chromosome in *C. elegans* is almost the same, with a value of 36% ± 1%. But the GC content is usually above 40% for promoters of individual pre-miRNAs, and above 50% for miRNA clusters (Table 1). In addition, there tends to be a positive correlation between the GC content/CpG number and the number of capped peaks: higher GC content/more CpG result in more strong capped peaks.

These capped reads peaks are correlated to the mature miRNA sequences (Figures 3 and 5). There is a low probability that multiple peaks within miRNA cluster are simply generated randomly. It has been suggested that transcription start usage of each nucleotide can be predicted from local DNA sequence [58]. Therefore, it is likely that the positions of these TSSs are mainly determined by nucleotide sequence.

Pre-miRNAs in a cluster can be transcribed independently

Pre-miRNAs from the same genomic cluster can be transcribed and regulated independently [60]. For example, although the primary transcripts of mouse mir-433 and mir-127 were detected as overlapping in a 5′–3′ unidirectional way, experiments have verified that they were transcribed independently from each other [61].

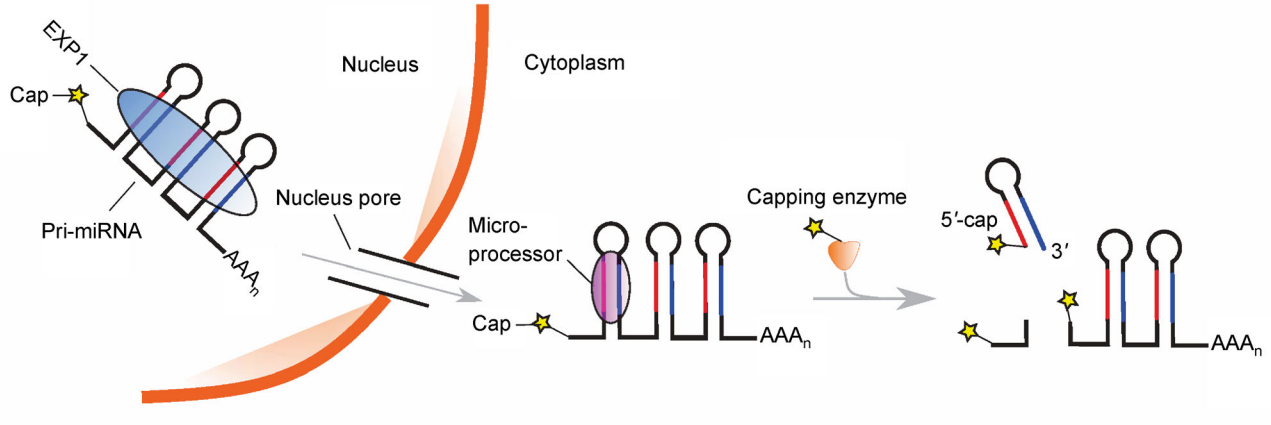

Figure 6. Model for miRNA cytoplasmic re-capping. In this non-canonical miRNA pathway, pri-miRNAs are exported to cytoplasm by XPO1 and processed there. During the pre-miRNA generating process, m^7G-caps are added to the 5′ ends of newly generated pre-miRNA and pri-miRNA left over by cytoplasmic capping enzyme.

Table 1. GC content, OE ratio and CpG number for miRNA promoters with multiple TSSs

Promoters	Location	GC content	OE ratio	CPG number
(A) Cluster miRNAs	II:11537326-11537576	0.55	1.18	18
	II:11889425-11889675	0.56	1.09	13
	III:11937020-11937270	0.52	0.93	9
	III:2172175-2172425	0.55	1.00	15
	X:2368553-2368803	0.51	0.76	9
	X:13145204-13145454	0.60	1.34	24
(B) Individual miRNAs	cel-mir-244 I:4684364-4684614	0.40	0.82	8
	cel-mir-235 I:6162337-6162587	0.41	0.76	8
	cel-mir-238 III:8867375-8867625	0.40	0.66	6
	cel-mir-228 IV:5561825-5562075	0.43	1.18	12
	cel-mir-51 IV:11026062-11026312	0.51	1.09	17
	cel-mir-53 IV:11027641-11027891	0.45	1.38	16
	cel-mir-49 X:9989082-9989332	0.51	0.75	12

According to this model, the pre-miRNAs in the same cluster may transcribe independently, producing multiple 5′ m^7G capped RNA peaks located inside of the cluster. The capped RNA peaks at the 5p arms indicate that the 5′ end of the pre-miRNA may be determined by transcription initiation [22], while the capped peaks at the 3p arms may correspond to the TSSs for the subsequent pre-miRNAs. We noticed that many pre-miRNAs in the cluster have both the capped RNA peaks on its own 5′ end and the 3p arm of upstream pre-miRNA. We have speculated that pre-miRNAs might be able to be generated in the canonical way with Drosha or in the non-canonical way by transcription initiation and termination, so here again, the same pre-miRNA may be generated by different mechanisms: one with Drosha and the other without.

Chromatin and Pol II profiles of primary miRNA promoters

Identification of divergent and multiple TSSs promoters

To characterize these identified pri-miRNA TSSs, we analysed the distribution of chromatin and Pol II features surrounding them. We observed that promoters in *C. elegans* often generate divergent or multiple transcripts with the Cap-seq datasets used. To identify these promoters, we defined the divergent/bidirectional and broad promoters (promoters with multiple TSSs) from those transcription initiation clusters (TICs). If the distance between two adjacent plus strand and minus strand TICs is less than or equal to 300 bp, they are combined as from the same divergent promoter. TICs on the same strand are clustered together if the distance between two adjacent TICs is within 500 bp. These clustered TICs on the same strand will define the broad promoters (see details in Materials and Methods). In the whole genome of *C. elegans*, we detected 11,272 promoters, of which 6,149 are divergent promoters and 2,359 are broad promoters (Figure 7A). The most upstream 5′ ends of both plus and minus strand TICs were used to represent the TSSs of the promoter.

With all the identified promoters, we focused on those with either multiple transcripts (at least 3 TSSs) or bidirectionality. That is, the broad promoters are non-bidirectional and bidirectional promoters do not contain multiple TSSs. We aligned the TSSs of these promoters, and plotted H3K4me3 and Pol II signal profiles surrounding them (Figure 8A and 8B)). The result shows that H3K4me3 signal is much stronger in the downstream of broad promoters than bidirectional promoters, indicating that H3K4me3 is strongly correlated with transcription initiations. Interestingly, we observe that Pol II signal in the downstream of broad promoters is also slightly stronger than in the upstream, which would not be observed if we use all promoters with multiple TSSs (including bidirectional promoters (Supplementary Figure S2D). The higher level of downstream Pol II signal may be caused by the Pol II pausing at the transcription initiation sites. It has been shown that Pol II promoter-proximal pausing is rare in *C. elegans* [36], which may explain why the Pol II signal difference between upstream and downstream is not big.

Chromatin and Pol II profiles surrounding miRNA promoters

We then assigned these identified promoters to the intergenic miRNAs in *C. elegans*. For each of them, the flanking region between its 5′ end and the most close

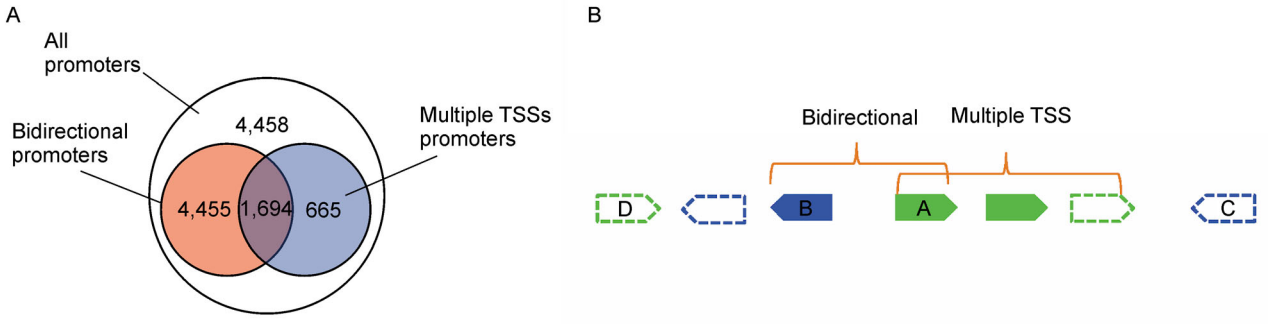

Figure 7. Promoters in *C. elegans*. (A) Promoters distribution in *C. elegans*. (B) Detecting bidirectional and multiple transcription promoters in *C. elegans*.

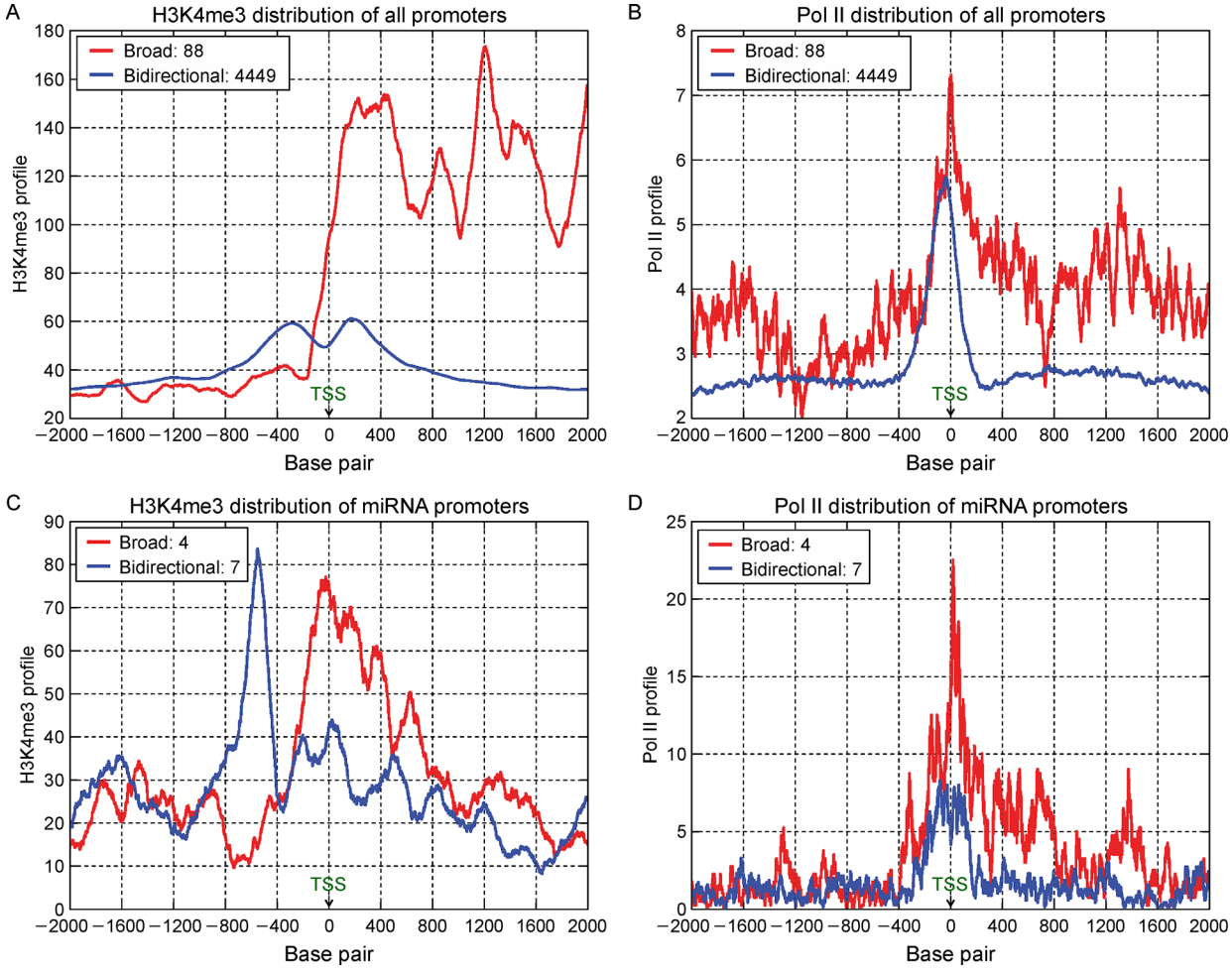

Figure 8. H3K3me3 and Pol II distributions around active promoters. (A and B), H3K4me3 and Pol II signal profiles surrounding bidirectional and broad promoters. (C and D), H3K4me3 and Pol II signal profiles surrounding miRNA bidirectional and broad promoters.

upstream gene's 3′ end is searched for promoters, and the most close promoter is assigned to the miRNA as its primary promoter. Again, we aligned these miRNAs with their primary TSSs acquired from the assigned promoters and plotted the H3K4me3 and Pol II signal surrouding the TSSs. The results are shown in Figure 8C and 8D.

H3K4me3 signal surrounding miRNA broad promoters peaks at the identified TSSs while the signal surrounding bidirectional promoters peaks at the upstream 600 bp position, indicating that there may be much stronger minus transcripts for these bidirectional promoters. For Pol II signal, the intensities are similar in the upstream of both miRNA broad promoters and bidirectional promoters. However, again the downstream of broad promoters has a stronger Pol II signal, suggesting Pol II is also paused in the proximal promoters of miRNA genes.

DISCUSSION

In this study, we used Cap-seq datasets to annotate the primary TSSs for intergenic miRNA genes to 1 base resolution in *C. elegans* and mouse. In total, we annotated the primary TSSs for 70 miRNAs and 8 miRNA clusters in *C. elegans*, and 37 miRNAs on mouse chromosome 7. Comparing with previous work using capped RNA-seq methods [33,36], we have annotated the TSSs for many more miRNAs in both species. We noticed that the Cap-seq datasets we used are either generated from mixed-stage embryos (*C. elegans*) or mixed tissues (new born mouse), while the works we compared with only utilized few staged datasets in *C. elegans* [33,36] or single tissue dataset in mouse [33]. Therefore, perhaps less miRNA genes were expressed and detected in their studies. We also searched for the Cap-seq TICs in the whole flanking regions upstream of pre-miRNAs. Without the details of how they detected the primary TSSs using the sequencing data, our methods may be able to capture more capped RNA signal for miRNAs.

Similar to the previous study [22], which detected a special class of pre-miRNAs that are 5′ m^7G capped in mice, here we also identified this type of pre-miRNAs in *C. elegans*. 5′-capped pre-miRNAs in mice prefer to be exported to the cytoplasm by XPO1 instead of XPO5 [22]. Comparing to mammals, XPO5 or its homologue is not encoded in *C. elegans* but XPO1 is [62]. Therefore, many more 5′-capped pre-miRNAs were expected to be identified in *C. elegans*. But using the Cap-seq datasets of mixed-stage embros of *C. elegans*, we have only detected 9 candidate 5′-capped pre-miRNAs. The XPO1 and cap-binding complex (CBC) have been reported to act jointly to export pri-miRNAs in *C. elegans* and *Drosophila*, but how the subsequent pre-miRNAs are processed from these pri-miRNAs in the cytoplasm remains unclear [38]. Accordingly, both 5′-capped pre-miRNAs and normal pri-miRNAs can be exported to the cytoplasm and their subsequent processing may be different from that in the canonical miRNA biogenesis pathway. Since XPO1 does not specifically export 5′-capped pre-miRNAs in *C. elegans*, its existence does not necessarily result in more 5′-capped pre-miRNAs. Thus, it is not odd that only a small number of these unusual pre-miRNAs are observed in *C. elegans*.

It has been suggested that XPO1-dependent m^7G capped pre-miRNAs may represent a group of ancient miRNAs that appeared before the emergence of XPO5 [22]. Therefore, these m^7G capped pre-miRNAs may be well conserved in different lineages. We checked the conservation of those identified 5′-capped pre-miRNAs in *C. elegans* and mouse from miRviewer [63]. Surprisingly, almost all of these identified m^7G capped pre-miRNAs are lineage specific: they are detected either only in caenorhabditis or muroidea.

In both mouse and *C. elegans*, we have identified a class of pre-miRNAs that are 5′ m^7G mapped but also posses upstream candidate primary TSSs. It has been suggested that most genes in mammals are not transcribed from a single TSS, but multiple TSSs that are closely located over 50 to 100 nucleotides [57,58,64,65]. Indeed, we observed that some upstream TICs are very close to the pre-miRNAs, suggesting that the proximal core promoter of these miRNA genes may generate multiple transcription initiations that are closely located to each other. However, the other further upstream TICs (over 500 nt) are unlikely to be generated from the same promoter as the pre-cap TICs. It has been reported that many genes have alternative promoters and can generate multiple isoforms in different cell types, tissues or developmental stages [66,67]. Since we used the Cap-seq datasets of mixed staged embryos of *C. elegans*, these miRNA genes may utilize alternative promoters at different stages to generate diverse transcripts which are subject to distinct processing procedures. That is, these 5′-capped pre-miRNAs may be only generated in specific cell types, developmental stages or disease states. Considering that these XPO1-dependent 5′ m^7G capped pre-miRNAs may belong to a group of ancient miRNAs, it is reasonable that at least some of them should be conserved along different lineages during evolution. However, we have showed that almost all of these identified 5′ m^7G capped pre-miRNAs are lineage specific. There is a possibility that these miRNA genes are newly generated and acquire their ability of producing 5′-capped pre-miRNAs later in the evolution. In extreme cases, most of these 5′-capped pre-miRNAs may have upstream distal promoters that can produce the normal primary transcripts at the other time. The situation where m^7G capped pre-miRNAs are expressed may due to the lack of Drosha or other related microprocessor. The ability of generating the same miRNAs with or without Drosha may help the organism to better adapt to the complex and changeable environment.

Many of the identified 5′ m^7G capped pre-miRNAs in

C. elegans preferentially generate 5p-miRNAs, which is different from that in mice as previously reported. These pre-miRNAs usually also have upstream candidate primary TSSs. We surmised that these 5′-capped pre-miRNAs may produce the 5p-miRNAs using the alternative upstream primary TSSs. However, this still needs further efforts to clarify.

Multiple capped peaks have been observed within pre-miRNAs in a cluster, and the well coordinated relationship between the capped peaks and the mature miRNAs suggest that the 5′ m^7G cap may be added during the pre-miRNA generating process. We proposed a new model of miRNA biogenesis in which the pri-miRNAs are cleaved and capped in the cytoplasm simultaneously (Figure 6). We noticed that the cytoplasmic capping is prominent in miRNA clusters, likely because the exposed 5′ end of miRNA cluster transcripts during cleavage need to be protected. Therefore, the m^7G cap is added and protects the cleaved transcripts from being degraded during the cleavage process. Then, the subsequent pre-miRNAs can be generated successfully.

The pre-miRNAs in clusters might be transcribed independently. To gain more evidence for this model, we also looked at the the sequence motif surrounding the modes of pre-cap TICs within clustered pre-miRNAs. The YR motif, which was shown at the putative primary miRNA TSSs and pre-cap TSSs of independent pre-miRNAs, is not observed (Supplementary Figure S8). Instead, a weak consensus sequence of "TNGG" is detected, in which "N" locates at the +1 position representing the modes of the TICs. Therefore, at least for some miRNA clusters, the independent transcription model is not supported by the YR motif.

MATERIALS AND METHODS

Datasets and processing

The small RNA Cap-seq data for new born mouse was retrieved from the study by Xie *et al*. [22], and was downloaded from Sequence Read Archive (SRA) with run number SRR1022391. The small RNA Cap-seq data for mixed staged embryos of *C. elegans* is from the study by Chen *et al*. [35], and was downloaded from NCBI Gene Expression Omnibus (GEO) with accession number GSE42819. The small RNA-seq data of *C. elegans* L4 stage was also downloaded from GEO database with accession number GSM916519. The ChIP-seq data of H3K4me3 and Pol II for *C. elegans* were also obtained from NCBI GEO database, with accession number GSE28770 and GSE15535, respectively. Small RNA-seq data for detecting mature miRNAs in *C. elegans* was downloaded from GEO with accession number GSM916519.

Raw sequencing data sets are in SRA format and were dumped to FASTQ format by SRA Toolkit [68]. The FASTQ files were then mapped to *C. elegans* (WBcel235) and Mouse (GRCm38) reference genome using bowtie [69] allowing 2 mismatches and 3 mismatches for reads length of 36 nt and 50 nt, respectively. Only uniquely mapped reads were reported in the output SAM files and were visualized by GenomeView [42].

Clustering of 5′ end reads

Small RNA Cap-seq data for *C. elegans* and mouse are strand specific. Mapped reads on forward and reverse strand were analyzed independently. We identified the transcription initiation clusters (TICs) with similar methods from the study by Chen *et al* [35]. First, mapped reads with same strand and 5′ end positions were combined and denoted as cap-stacks. Second, all cap-stacks containing five or more tags were clustered using a single-linkage approach: two or more stacks were clustered together if the distance between two adjacent stacks is less or equal to 50 bp. The position covered by the most 5′ ends within the TIC was defined as the mode, which represents the TSS for the TIC. In the case of two or more positions with the same number of tags, the one furthest upstream was selected as the mode. Here in total we obtained 32,530 TICs in *C. elegans* and 4,903 TICs on mouse chromosome 7.

Statistical analysis

Because the Cap-seq is not perfect and may involve contamination or artifacts, we used a Poisson distribution to model the background noise following the previous work [70]. Those TICs that are significantly enriched with sequencing reads are reported (Poisson distribution p-value based on λ). Since the reads of Cap-seq are not randomly distributed along the genome, we estimate a dynamic parameter λlocal, defined for each TIC as:

$$\lambda_{\text{local}} = \min[\lambda_{5k},\ \lambda_{10k}],$$

where λ_{5k} and λ_{10k} are estimated from the 5 kb or 10 kb window centered at the mode of the TIC. Our model is built on the assumption that a TIC is reliable if the number of reads enriched inside of the TIC is significantly higher (default with p-value $< 10^{-5}$) than the number of reads located in the TIC when they are randomly distributed in the local region.

Results of Cap-seq and small RNA-seq reads mapped to tRNAs were subject to a two-tailed paired sample t-test to estimate their differences, with p-value < 0.01 considered statistically significant. Both Cap-seq and small RNA-seq mapping results on tRNAs were normalized by their coverage depths on the genome. Since they only cover a

small part of the whole genome, the coverage depth was calculated as the total number of reads mapped multiply the reads length and divided by the length of covered genome region. The similar Poisson model is also used to estimate the enrichment of Cap-seq reads on tRNAs.

miRNA cluster and intergenic pre-miRNAs identification

miRNA clusters were retrieved from miRBase with a distance threshold of 1,000 bp. The distance between pre-miRNAs in mice are usually longer than 1,000 bp. Thus, we did not find any miRNA cluster in mice with our threshold.

To identify intergenic miRNAs from *C. elegans* and mouse genome, we downloaded the miRNA annotations from miRBase Release 21 and gene annotaion gff3 files. The gene annotation files for *C. elegans* and mouse were downloaded from Ensembl website (http://www.ensembl.org/index.html). Pre-miRNAs from both miRBase and Ensembl gene annotation files were isolated and those miRNAs that are not covered by protein-coding genes, non-coding RNA genes, snoRNA genes and snRNA genes were identified as intergenic miRNAs. In the case that same pre-miRNA is annotated both in miRBase and gene annotation file, we only kept the one in miRBase. In total, we identified 134 intergenic miRNAs in *C. elegans* and 80 intergenic miRNAs on mouse chromosome 7. The flanking region upstream of the 5′ end of intergenic pre-miRNA was also identified, and TICs detected in this region were annotated as the candidate primary TSSs for the miRNA. The TICs identified within the pre-miRNA region were annotated as the pre-cap TICs.

Finding bidirectional and multiple TSSs promoters

With the capped RNA reads, we identified transcription initiation sites and annotated their promoters as bidirectional or broad promoters (promoters with multiple TSSs) in *C. elegans*. We used the identified TICs to annotate these promoters. First for each TIC (*A* in Figure 7B) on the plus strand, the closest downstream (*C* in Figure 7B) and upstream (*B* in Figure 7B) minus strand TICs were searched. If the distance between the upstream minus TIC (*B*) and the plus strand TIC (*A*) is less than threshold (here we use 300 bp), the two TICs were treated as from the same promoter and the promoter was annotated as bidirectional. The downstream minus strand capped peak (*C*) acts as the boundary for detecting multiple transcripts. All the plus strand TICs upstream of it were annotated as from the same promoter. To identify the multiple transcripts on the minus strand, the most close upstream plus strand TIC (*D*) of *B* was found and all the minus strand TICs between *D* and *B* were annotated as from this promoter. Those left minus strand TICs were annotated with the similar method.

With this approach, we detected 11,272 promoters in *C. elegans*, of which, 6,149 are bidirectional promoters and 2,359 are broad promoters. The most upstream 5′ ends of both plus and minus strand TICs were used to represent for the TSSs of these promoters. Only one TSS on each strand was selected as representing TSS for one promoter.

SUPPLEMENTARY MATERIALS

The supplementary materials can be found online with this article at DOI 10.1007/s40484-017-0123-4.

ACKNOWLEDGEMENTS

We thank David Arnosti for critical discussions and helpful comments on the manuscript. We are also grateful to Weifeng Gu, Ron Chen, and Mingyi Xie for providing the explicit explanations to questions related to 5′-capped pre-miRNAs and 5′ recessed RNAs. This work was supported by NSF CAREER Grant DBI-0953738.

COMPLIANCE WITH ETHICS GUIDELINES

The authors Jiao Chen, Dongxiao Zhu, and Yanni Sun declare they have no conflict of interests. All the data sets the authors used are from public repositories.

REFERENCES

1. Kim, V. N. and Nam, J.-W. (2006) Genomics of microRNA. Trends Genet., 22, 165–173
2. Krol, J., Loedige, I. and Filipowicz, W. (2010) The widespread regulation of microRNA biogenesis, function and decay. Nat. Rev. Genet., 11, 597–610
3. Berezikov, E. (2011) Evolution of microRNA diversity and regulation in animals. Nat. Rev. Genet., 12, 846–860
4. Mallanna, S. K. and Rizzino, A. (2010) Emerging roles of microRNAs in the control of embryonic stem cells and the generation of induced pluripotent stem cells. Dev. Biol., 344, 16–25
5. Collins, F. S. and Varmus, H. (2015) A new initiative on precision medicine. N. Engl. J. Med., 372, 793–795
6. Larry Jameson, J. and Longo, D. L. (2015) Precision medicine—personalized, problematic, and promising. Obstet. Gynecol. Surv., 70, 612–614
7. Lüscher, T. F. (2016) Frontiers in precision medicine: genes and their modulation by miRNAs. Eur. Heart J., 37, 3247–3250
8. Willeit, P., Skroblin, P., Kiechl, S., Fernández-Hernando, C. and Mayr, M. (2016) Liver microRNAs: potential mediators and biomarkers for metabolic and cardiovascular disease? Eur. Heart J., 37, 3260–3266
9. Matin, F., Jeet, V., Clements, J. A., Yousef, G. M. and Batra, J. (2016) MicroRNA theranostics in prostate cancer precision medicine. Clin. Chem., 62, 1318–1333
10. Coronnello, C. and Benos, P. V. (2013) ComiR: combinatorial

microRNA target prediction tool. Nucleic Acids Res., 41, W159–W164
11. Yuan, C. and Sun, Y. (2013) RNA-CODE: a noncoding RNA classification tool for short reads in NGS data lacking reference genomes. PLoS One, 8, e77596
12. Lei, J. and Sun, Y. (2014) miR-PREFeR: an accurate, fast and easy-to-use plant miRNA prediction tool using small RNA-Seq data. Bioinformatics, 30, 2837–2839
13. Lee, Y., Kim, M., Han, J., Yeom, K.-H., Lee, S., Baek, S. H. and Kim, V. N. (2004) MicroRNA genes are transcribed by RNA polymerase II. EMBO J., 23, 4051–4060
14. Borchert, G. M., Lanier, W. and Davidson, B. L. (2006) RNA polymerase III transcribes human microRNAs. Nat. Struct. Mol. Biol., 13, 1097–1101
15. Lee, Y., Ahn, C., Han, J., Choi, H., Kim, J., Yim, J., Lee, J., Provost, P., Rådmark, O., Kim, S., *et al.* (2003) The nuclear RNase III Drosha initiates microRNA processing. Nature, 425, 415–419
16. Chendrimada, T. P., Gregory, R. I., Kumaraswamy, E., Norman, J., Cooch, N., Nishikura, K. and Shiekhattar, R. (2005) TRBP recruits the Dicer complex to Ago2 for microRNA processing and gene silencing. Nature, 436, 740–744
17. Kuehbacher, A., Urbich, C., Zeiher, A. M. and Dimmeler, S. (2007) Role of Dicer and Drosha for endothelial microRNA expression and angiogenesis. Circ. Res., 101, 59–68
18. Berezikov, E., Chung, W.-J., Willis, J., Cuppen, E. and Lai, E. C. (2007) Mammalian mirtron genes. Mol. Cell, 28, 328–336
19. Ruby, J. G., Jan, C. H. and Bartel, D. P. (2007) Intronic microRNA precursors that bypass Drosha processing. Nature, 448, 83–86
20. Chang, T.-C., Pertea, M., Lee, S., Salzberg, S. L. and Mendell, J. T. (2015) Genome-wide annotation of microRNA primary transcript structures reveals novel regulatory mechanisms. Genome Res., 25, 1401–1409
21. Dai, L., Chen, K., Youngren, B., Kulina, J., Yang, A., Guo, Z., Li, J., Yu, P. and Gu, S. (2016) Cytoplasmic Drosha activity generated by alternative splicing. Nucleic Acids Res., 44, 10454–10466
22. Xie, M., Li, M., Vilborg, A., Lee, N., Shu, M.-D., Yartseva, V., Šestan, N. and Steitz, J. A. (2013) Mammalian 5′-capped microRNA precursors that generate a single microRNA. Cell, 155, 1568–1580
23. Griffiths-Jones, S., Grocock, R. J., van Dongen, S., Bateman, A. and Enright, A. J. (2006) miRBase: microRNA sequences, targets and gene nomenclature. Nucleic Acids Res., 34, D140–D144
24. Wang, Z., Gerstein, M. and Snyder, M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet., 10, 57–63
25. Saini, H. K., Griffiths-Jones, S. and Enright, A. J. (2007) Genomic analysis of human microRNA transcripts. Proc. Natl. Acad. Sci. USA, 104, 17719–17724
26. Ozsolak, F., Poling, L. L., Wang, Z., Liu, H., Liu, X. S., Roeder, R. G., Zhang, X., Song, J. S. and Fisher, D. E. (2008) Chromatin structure analyses identify miRNA promoters. Genes Dev., 22, 3172–3183
27. Corcoran, D. L., Pandit, K. V., Gordon, B., Bhattacharjee, A., Kaminski, N. and Benos, P. V. (2009) Features of mammalian microRNA promoters emerge from polymerase II chromatin immunoprecipitation data. PLoS One, 4, e5279
28. Chien, C.-H., Sun, Y.-M., Chang, W.-C., Chiang-Hsieh, P.-Y., Lee, T.-Y., Tsai, W.-C., Horng, J.-T., Tsou, A.-P. and Huang, H.-D. (2011) Identifying transcriptional start sites of human microRNAs based on high-throughput sequencing data. Nucleic Acids Res., 39, 9345–9356
29. Wang, G., Wang, Y., Shen, C., Huang, Y. W., Huang, K., Huang, T. H., Nephew, K. P., Li, L. and Liu, Y. (2010) RNA polymerase II binding patterns reveal genomic regions involved in microRNA gene regulation. PLoS One, 5, e13798
30. Saini, H. K., Enright, A. J. and Griffiths-Jones, S. (2008) Annotation of mammalian primary microRNAs. BMC Genomics, 9, 564
31. Kodzius, R., Kojima, M., Nishiyori, H., Nakamura, M., Fukuda, S., Tagami, M., Sasaki, D., Imamura, K., Kai, C., Harbers, M., *et al.* (2006) CAGE: cap analysis of gene expression. Nat. Methods, 3, 211–222
32. de Hoon, M. and Hayashizaki, Y. (2008) Deep cap analysis gene expression (CAGE): genome-wide identification of promoters, quantification of their expression, and network inference. Biotechniques, 44, 627–632
33. Gu, W., Lee, H.-C., Chaves, D., Youngman, E. M., Pazour, G. J., Conte, D. Jr and Mello, C. C. (2012) CapSeq and CIP-TAP identify Pol II start sites and reveal capped small RNAs as *C. elegans* piRNA precursors. Cell, 151, 1488–1500
34. Corsi, A. K. (2006) A biochemist's guide to *Caenorhabditis elegans*. Anal. Biochem., 359, 1–17
35. Chen, R. A.-J., Down, T. A., Stempor, P., Chen, Q. B., Egelhofer, T. A., Hillier, L. W., Jeffers, T. E. and Ahringer, J. (2013) The landscape of RNA polymerase II transcription initiation in *C. elegans* reveals promoter and enhancer architectures. Genome Res., 23, 1339–1347
36. Kruesi, W. S., Core, L. J., Waters, C. T., Lis, J. T. and Meyer, B. J. (2013) Condensin controls recruitment of RNA polymerase II to achieve nematode X-chromosome dosage compensation. eLife, 2, e00808
37. Spieth, J., Lawson, D., Davis, P., Williams, G. and Howe, K. (2014) Overview of gene structure in *C. elegans*. In WormBook, 1–18. http://www.wormbook.org
38. Büssing, I., Yang, J. S. Jr, Lai, E. C. and Grosshans, H. (2010) The nuclear export receptor XPO-1 supports primary miRNA processing in *C. elegans* and *Drosophila*. EMBO J., 29, 1830–1839
39. Li, N., You, X., Chen, T., Mackowiak, S. D., Friedländer, M. R., Weigt, M., Du, H., Gogol-Döring, A., Chang, Z., Dieterich, C., *et al.* (2013) Global profiling of miRNAs and the hairpin precursors: insights into miRNA processing and novel miRNA discovery. Nucleic Acids Res., 41, 3619–3634
40. Fejes-Toth, K., Sotirova, V., Sachidanandam, R., Assaf, G., Hannon, G. J., Kapranov, P., Foissac, S., Willingham, A. T., Duttagupta, R., Dumais, E., *et al.* (2009) Post-transcriptional processing generates a diversity of 5′-modified long and short RNAs. Nature, 457, 1028–1032
41. Crooks, G. E., Hon, G., Chandonia, J.-M. and Brenner, S. E.

(2004) WebLogo: a sequence logo generator. Genome Res., 14, 1188–1190
42. Abeel, T., Van Parys, T., Saeys, Y., Galagan, J. and Van de Peer, Y. (2012) GenomeView: a next-generation genome browser. Nucleic Acids Res., 40, e12
43. Bracht, J., Hunter, S., Eachus, R., Weeks, P. and Pasquinelli, A. E. (2004) Trans-splicing and polyadenylation of let-7 microRNA primary transcripts. RNA, 10, 1586–1594
44. Davuluri, R. V., Suzuki, Y., Sugano, S., Plass, C. and Huang, T. H.-M. (2008) The functional consequences of alternative promoter use in mammalian genomes. Trends Genet., 24, 167–177
45. Djebali, S., Davis, C. A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., *et al.* (2012) Landscape of transcription in human cells. Nature, 489, 101–108
46. Sigova, A. A., Mullen, A. C., Molinie, B., Gupta, S., Orlando, D. A., Guenther, M. G., Almada, A. E., Lin, C., Sharp, P. A., Giallourakis, C. C., *et al.* (2013) Divergent transcription of long noncoding RNA/mRNA gene pairs in embryonic stem cells. Proc. Natl. Acad. Sci. USA, 110, 2876–2881
47. Wei, Y., Zhang, S., Shang, S., Zhang, B., Li, S., Wang, X., Wang, F., Su, J., Wu, Q., Liu, H., *et al.* (2016) SEA: a super-enhancer archive. Nucleic Acids Res., 44, D172–D179
48. Biasiolo, M., Sales, G., Lionetti, M., Agnelli, L., Todoerti, K., Bisognin, A., Coppe, A., Romualdi, C., Neri, A. and Bortoluzzi, S. (2011) Impact of host genes and strand selection on miRNA and miRNA* expression. PLoS One, 6, e23854
49. Meijer, H. A., Smith, E. M. and Bushell, M. (2014) Regulation of miRNA strand selection: follow the leader? Biochem. Soc. Trans., 42, 1135–1140
50. Lau, N. C., Lim, L. P., Weinstein, E. G. and Bartel, D. P. (2001) An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans*. Science, 294, 858–862
51. Otsuka, Y., Kedersha, N. L. and Schoenberg, D. R. (2009) Identification of a cytoplasmic complex that adds a cap onto 5′-monophosphate RNA. Mol. Cell. Biol., 29, 2155–2167
52. Schoenberg, D. R. and Maquat, L. E. (2009) Re-capping the message. Trends Biochem. Sci., 34, 435–442
53. Mukherjee, C., Patil, D. P., Kennedy, B. A., Bakthavachalu, B., Bundschuh, R. and Schoenberg, D. R. (2012) Identification of cytoplasmic capping targets reveals a role for cap homeostasis in translation and mRNA stability. Cell Rep., 2, 674–684
54. Kiss, D. L., Oman, K., Bundschuh, R. and Schoenberg, D. R. (2015) Uncapped 5′ ends of mRNAs targeted by cytoplasmic capping map to the vicinity of downstream CAGE tags. FEBS Lett., 589, 279–284
55. Rouha, H., Thurner, C. and Mandl, C. W. (2010) Functional microRNA generated from a cytoplasmic RNA virus. Nucleic Acids Res., 38, 8328–8337
56. Shapiro, J. S., Langlois, R. A., Pham, A. M. and Tenoever, B. R. (2012) Evidence for a cytoplasmic microprocessor of pri-miRNAs. RNA, 18, 1338–1346
57. Sandelin, A., Carninci, P., Lenhard, B., Ponjavic, J., Hayashizaki, Y. and Hume, D. A. (2007) Mammalian RNA polymerase II core promoters: insights from genome-wide studies. Nat. Rev. Genet., 8, 424–436
58. Frith, M. C., Valen, E., Krogh, A., Hayashizaki, Y., Carninci, P. and Sandelin, A. (2008) A code for transcription initiation in mammalian genomes. Genome Res., 18, 1–12
59. Sandelin, A., Alkema, W., Engström, P., Wasserman, W. W. and Lenhard, B. (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. Nucleic Acids Res., 32, D91–D94
60. Winter, J., Jung, S., Keller, S., Gregory, R. I. and Diederichs, S. (2009) Many roads to maturity: microRNA biogenesis pathways and their regulation. Nat. Cell Biol., 11, 228–234
61. Song, G. and Wang, L. (2008) MiR-433 and miR-127 arise from independent overlapping primary transcripts encoded by the miR-433-127 locus. PLoS One, 3, e3574
62. Murphy, D., Dancis, B. and Brown, J. R. (2008) The evolution of core proteins involved in microRNA biogenesis. BMC Evol. Biol., 8, 92
63. Kiezun, A., Artzi, S., Modai, S., Volk, N., Isakov, O. and Shomron, N. (2012) miRviewer: a multispecies microRNA homologous viewer. BMC Res. Notes, 5, 92
64. Carninci, P., Sandelin, A., Lenhard, B., Katayama, S., Shimokawa, K., Ponjavic, J., Semple, C. A., Taylor, M. S., Engström, P. G., Frith, M. C., *et al.* (2006) Genome-wide analysis of mammalian promoter architecture and evolution. Nat. Genet., 38, 626–635
65. Danino, Y. M., Even, D., Ideses, D. and Juven-Gershon, T. (2015) The core promoter: at the heart of gene expression. Biochim. Biophys. Acta, 1849, 1116–1131
66. Landry, J.-R., Mager, D. L. and Wilhelm, B. T. (2003) Complex controls: the role of alternative promoters in mammalian genomes. Trends Genet., 19, 640–648
67. Baek, D., Davis, C., Ewing, B., Gordon, D. and Green, P. (2007) Characterization and predictive discovery of evolutionarily conserved mammalian alternative promoters. Genome Res., 17, 145–155
68. Staff, S. (2011) Using the sra toolkit to convert. sra files into other formats. National Center for Biotechnology Information (US)
69. Langmead, B., Trapnell, C., Pop, M. and Salzberg, S. L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol., 10, R25
70. Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W., *et al.* (2008) Model-based analysis of ChIP-Seq (MACS). Genome Biol., 9, R137

