Determination of primary microRNA processing in clinical samples by targeted pri-miR-sequencing

THOMAS CONRAD,[1] EVGENIA NTINI,[2] BENJAMIN LANG,[3] LUCA COZZUTO,[3] JESPER B. ANDERSEN,[4] JENS U. MARQUARDT,[5] JULIA PONOMARENKO,[3,6] GIAN GAETANO TARTAGLIA,[3,6] and ULF A. VANG ØROM[7]

[1]Berlin Institute for Medical Systems Biology, 10115 Berlin, Germany

[2]Max Planck Institute for Molecular Genetic, 14195 Berlin, Germany

[3]Centre for Genomic Regulation (CRG), The Barcelona Institute for Science and Technology, 08003 Barcelona, Spain

[4]Biotech Research and Innovation Centre (BRIC), Department of Health and Medical Sciences, University of Copenhagen, 2200 Copenhagen, Denmark

[5]Universitätsklinikum Schleswig-Holstein, 23538 Lübeck, Germany

[6]University Pompeu Fabra (UPF), 08002 Barcelona, Spain

[7]Aarhus University, Department of Molecular Biology and Genetics, 8000 Aarhus, Denmark

ABSTRACT

MicroRNA expression is important for gene regulation and deregulated microRNA expression is often observed in diseases such as cancer. The processing of primary microRNA transcripts is an important regulatory step in microRNA biogenesis. Due to low expression level and association with chromatin, primary microRNAs are challenging to study in clinical samples where input material is limited. Here, we present a high-sensitivity targeted method to determine processing efficiency of several hundred primary microRNAs from total RNA that requires relatively few RNA sequencing reads. We validate the method using RNA from HeLa cells and show the applicability to clinical samples by analyzing RNA from normal liver and hepatocellular carcinoma. We identify 24 primary microRNAs with significant changes in processing efficiency from normal liver to hepatocellular carcinoma, among those the highly expressed miRNA-122 and miRNA-21, demonstrating that differential processing of primary microRNAs is occurring and could be involved in disease. With our method presented here we provide means to study pri-miRNA processing in disease from clinical samples.

Keywords: miRNA biogenesis; primary miRNAs; RNA sequencing; liver; HCC; clinical samples

INTRODUCTION

MicroRNAs (miRNA) are small RNAs that regulate gene expression at the posttranscriptional level (Kim et al. 2009). miRNAs are transcribed as primary miRNA (pri-miRNA) that can be several kilobases long. These transcripts are processed in the nucleus to 60–90 nt long precursor miRNA hairpins (pre-miRNA) by the Microprocessor complex. The pre-miRNAs are subsequently exported to the cytoplasm by export factors where they are processed into 19–23 nt long mature miRNAs by Dicer and incorporated into the RISC (RNA-induced silencing complex) to exert their regulatory function (Kim et al. 2009).

We have previously shown that the endogenous Microprocessor activity toward individual pri-miRNAs can be determined using RNA sequencing (Conrad et al. 2014). We identified a signature in the pri-miRNA where the Microprocessor cleaves the transcript, which can be quantified to determine processing efficiency (0—not processed to 1—fully processed) of individual pri-miRNAs (Fig. 1A). In our previous work we found that the pri-miRNA transcript remains rather stable after cleavage by the Microprocessor and leaves a specific signature (Fig. 1B) that can be used to quantify the fraction of pri-miRNA cleaved on a steady-state level (Conrad et al. 2014) and dynamically using pulse-chase RNA sequencing (Louloupi et al. 2017). With these findings we showed on a transcriptome-wide scale that processing efficiency is highly variable among different canonical pri-miRNAs and a major determinant of the expression levels of individual mature miRNAs. In addition, we have shown that
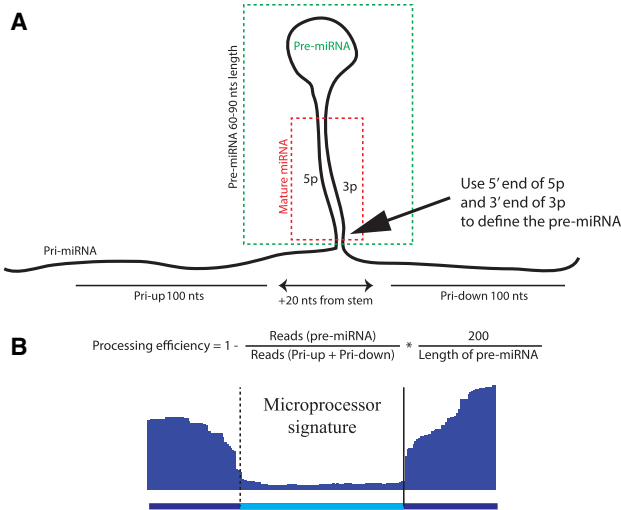
FIGURE 1. The Microprocessor signature. (*A*) The general structure of a pri-miRNA with the hairpin that will become the pre-miRNA upon Microprocessor cleavage and the mature miRNA sequence are indicated. The equation used to calculate processing efficiency is shown. Here the ratio of pri-miRNAs covering the pre-miRNA region (representing unprocessed pri-miRNA) to flanking regions (representing both processed and unprocessed pri-miRNA) is calculated to show the fraction of the pri-miRNA that has been cleaved. The processing efficiency is then calculated as 1 − this ratio. Processing efficiency will have a value between 1 (fully processed) and 0 (not processed). (*B*) Example of the signature where the Microprocessor has cleaved in an RNA sequencing read density plot, showing the profile that can be quantified to determine processing efficiency.

the kinetics of pri-miRNA processing vary between individual transcripts, including within polycistronic transcripts (Louloupi et al. 2017). We have shown that the processing efficiency of individual pri-miRNAs is similar between three different cell lines (HeLa, HEK293 and A549), suggesting that diversity in processing is largely dictated by the diverse substrate sequences. On the other side, several studies have shown that cofactors and signaling pathways can regulate processing of individual pri-miRNAs or groups of pri-miRNAs (Davis et al. 2008; Suzuki et al. 2009; Auyeung et al. 2013; Mori et al. 2014; Witteveldt et al. 2018).

In vitro assays looking at individual pri-miRNA transcripts have shown pri-miRNA processing to be an actively regulated process that is responsive to TGFß signaling (Davis et al. 2008), DNA damage (Suzuki et al. 2009) and cell density (Mori et al. 2014), and a recent study applied our pri-miRNA processing approach to demonstrate that activation of the Type I interferon response in cells affects general pri-miRNA processing (Witteveldt et al. 2018). Several Microprocessor cofactors have been identified that influence pri-miRNA processing (Conrad et al. 2014), modification of pri-miRNA with m6A has been suggested to affect processing of pri-miRNA transcripts (Alarcón et al. 2015; Han et al. 2020) as has A-to-I editing (Li et al. 2018), and

composition of the sequence flanking the pre-miRNA hairpin has been shown to affect efficiency of Microprocessor cleavage (Auyeung et al. 2013; Conrad et al. 2014).

The major limitation of our established approach to profile pri-miRNA processing transcriptome-wide has been a requirement for purification of chromatin-associated RNA and large sequencing depth (for the initial methods development we used 200 million reads per sample), making analysis of clinical samples unfeasible. In the work reported here we aimed to improve the approach to enable the analysis of pri-miRNA processing in clinical samples directly from whole cell total RNA isolated from tissue. We used a targeted sequencing approach allowing for quantification of 277 pri-miRNAs with as little as 500 thousand Illumina HiSeq reads, and demonstrate the applicability to clinical samples using RNA from normal liver (NL) and hepatocellular carcinoma (HCC) samples. We also identified differentially processed pri-miRNAs between NL and HCC with a potential implication in disease.

RESULTS

We developed a targeted approach to determine processing efficiency of a library of selected pri-miRNAs from small amounts of starting material of total RNA and low sequencing depth. We designed enrichment probes (xGEN lockdown probes, Integrated DNA Technologies) to enrich for pri-miRNA sequences from RNA sequencing libraries (targeting the cDNA of the libraries) covering the sequence of the pri-miRNA flanking the region of the pre-miRNA both upstream and downstream (Fig. 2A). We targeted 32 pri-miRNAs that we had already determined the processing efficiency for using chromatin-associated RNA and high sequencing depth (Conrad et al. 2014). For each pri-miRNA we designed two probes upstream and two probes downstream from the Microprocessor cleavage site. Each probe was designed to be complementary to 120 nt in the region immediately upstream or downstream from the Microprocessor cleavage sites, respectively. We performed two independent enrichment experiments on a TruSeq stranded RNA sequencing library generated from chromatin-associated RNA from HeLa cells (from Conrad et al. 2014) to test the error range of the enrichment approach (Fig. 2B). We see a highly reproducible processing efficiency (low error range) between replicates with a correlation, $R = 0.995$, of processing efficiencies for the 32 pri-miRNAs, demonstrating that the approach is robust and generates reproducible data.

We next asked how well the enrichment approach recapitulates the processing efficiencies determined by high-depth sequencing without enrichment (data from Conrad et al. 2014). From the two independent probe enrichment replicates we see a high correlation ($R = 0.92$) in processing efficiency when comparing to previous data of pri-miRNA processing from purified chromatin-associated RNA (Fig.
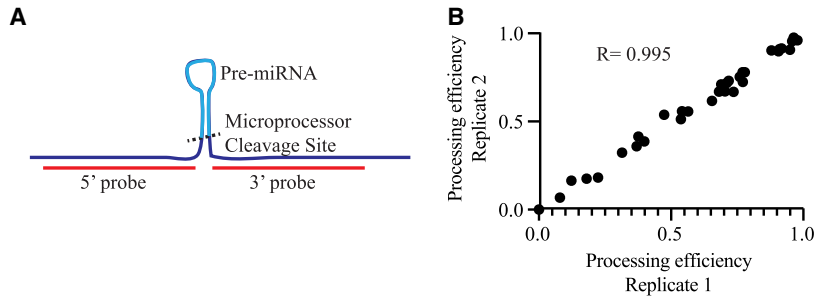
A

Pre-miRNA

Microprocessor
Cleavage Site

5' probe   3' probe

B

R= 0.995

Processing efficiency
Replicate 2

Processing efficiency
Replicate 1

FIGURE 2. Enrichment of pri-miRNAs for determination of processing efficiency. (*A*) Schematic of a pri-miRNA and the localization of our enrichment probes. (*B*) Reproducibility of processing efficiency for 32 pri-miRNAs in HeLa cells between two independent enrichment experiments from an RNA sequencing library generated from chromatin-associated HeLa RNA.

3A; Conrad et al. 2014). Processing efficiencies and coverage for the 32 pri-miRNAs are shown in Supplemental Table 1.

To assess the degree of enrichment achieved with the targeted approach we calculated RPKM (Reads Per Kilobase per Million reads) for each of the 32 pri-miRNAs. As shown in Figure 3B the enrichment is linear across the expression range of assayed pri-miRNAs ($R = 0.998$) with a slope of 6670-fold enrichment when using targeted sequencing of pri-miRNA transcripts. This means that, in principle, 30,000 reads should be sufficient to determine processing efficiency for selected pri-miRNAs when enriched from the chromatin fraction. Assuming that chromatin-associated RNA constitutes 5% of the total cellular (nonribosomal) RNA 600,000 reads is necessary to achieve robust determination of pri-miRNA processing from total RNA after targeted enrichment.

To address differential pri-miRNA processing in clinical samples we used RNA from 40 HCC tumors and nine NL samples. We selected 277 miRNAs with a described involvement in cancer (Farazi et al. 2011; Volinia et al. 2012; Wong et al. 2018) and designed 120-nt xGEN enrichment probes as illustrated in Figure 2A. For each RNA sample we prepared a TruSeq stranded RNA sequencing library for Illumina HiSeq 2500 sequencing and enriched with the probe library as described in Materials and Methods. We aimed for 1 million reads per library when multiplexing and the final output per library was between 500 thousand and 3 million reads. With this sequencing depth we detect the processing efficiency for 209 pri-miRNAs in at least two samples, and show differential processing between HCC and NL for 24 of the pri-miRNAs included in the analysis ($P < 0.05$,

Wilcoxon rank test) (Fig. 4A; Supplemental Table 2). Following targeted enrichment of pri-miRNAs, 33% of uniquely mapped reads align to pri-miRNA sequences, demonstrating an efficient targeted capture despite low expression levels of the pri-miRNA transcripts. We see the most variance in capture from HCC samples ($33 \pm 22\%$ of uniquely mapping reads) compared to NL ($33 \pm 1\%$ of uniquely mapped reads), suggesting more variability in sample quality of HCC samples. Of particular interest, we see the most statistically significant changes in the processing of the liver-specific miR122 and the oncomiR miR21 (Fig. 4B), where processing becomes more efficient in HCC. In NL, around 80% of the pri-miRNA is cleaved by the Microprocessor whereas in HCC we can observe almost complete processing (close to 100% of the pri-miRNA is processed).

DISCUSSION

Here, we have developed a targeted version of our approach to measure endogenous pri-miRNA processing for hundreds of different pri-miRNA at a time. Our targeted approach offers high sensitivity to the selected pri-miRNAs and allows assessing numerous pri-miRNAs at a time in a large collection of tissue and clinical samples. The method is particularly feasible to complement total RNA sequencing studies as existing RNA sequencing libraries can directly be used for targeted sequencing of pri-miRNA transcripts at very low cost once the enrichment probe library has been designed.
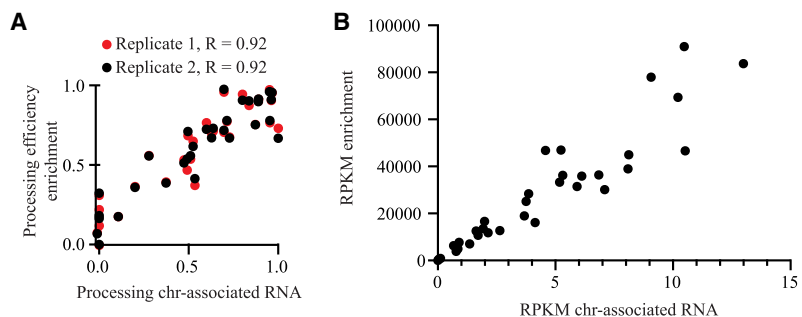
A

● Replicate 1, R = 0.92
● Replicate 2, R = 0.92

Processing efficiency
enrichment

Processing chr-associated RNA

B

RPKM enrichment

RPKM chr-associated RNA

FIGURE 3. Reproducibility and sensitivity of pri-miRNA processing efficiency determination. (*A*) Correlation of processing efficiency determined from high-depth sequencing of chromatin-associated RNA (Conrad et al. 2014) and low-depth sequencing of targeted sequencing of total RNA for two technical replicates. (*B*) Sensitivity of each approach determined by RPKM at pri-miRNAs. Slope shows a 6670-fold increase in sensitivity and a corresponding decreased need for sequencing depth. RPKM for targeted sequencing is calculated as the average of two independent enrichment replicates.
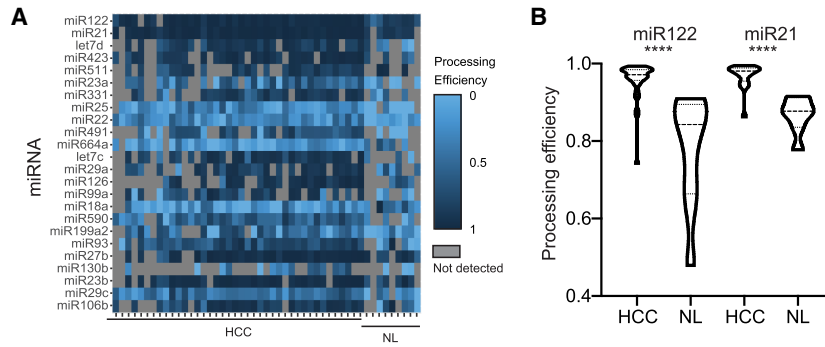
FIGURE 4. Differential pri-miRNA processing between NL and HCC. (*A*) Heat-map of pri-miRNA levels of the 24 differentially processed pri-miRNAs between HCC and NL. miRNAs are shown with the most statistically significant changes between NL and HC at the *top* (miR122) and increasing *P*-values toward the *bottom* (miR106b). (*B*) Increased processing efficiency of pri-miRNAs in HCC compared to NL for miR-122 and miR-21. Shown is the processing efficiency as determined by RNA sequencing in all 40 HCC or all nine NL samples. (****) *P* < 0.0001, Wilcoxon rank test.

While the method provides a reliable measure for processing efficiency, it should be noted that some factors of variability exist, such as stability of pri-miRNAs, degradation of pri-miRNA ends after cleavage by the Microprocessor and sample preparation and storage. The processing efficiency should therefore be taken as a relative measure that can identify differential processing efficiencies between samples.

We see processing efficiencies to be reproducible across samples, with most variability in the HCC samples. This is likely a consequence of partial degradation of the processed pri-miRNA during surgery, purification and storage of patient samples and RNA. This illustrates the need to include a reasonable panel size of samples to overcome this variability. Based on mapping of pri-miRNAs in both NL and HCC samples, the 40 samples used in this study provide a sufficient panel size. This panel size could vary dependent on the quality of the samples used.

We show that the targeted approach works well for clinical samples by demonstrating reproducible differential processing of a number of pri-miRNAs between NL and HCC. Here, we show that differential pri-miRNA processing occurs between NL and HCC suggesting that the primary processing step in miRNA biogenesis can impact gene expression in disease. Our targeted approach offers the possibility for further studies of pri-miRNA processing from clinical samples that have so far not been feasible, allowing identification of the molecular consequences of deregulated pri-miRNA processing in disease.

MATERIALS AND METHODS

Cell culture

HeLa cells were cultured at 37°C with 5% CO_2 in Dulbecco's modified Eagle's medium containing 10% FCS and antibiotics.

Chromatin fractionation and RNA isolation

RNA was purified using TRIzol according to the manufacturer's recommendations. For isolation of chromatin-associated RNA we used one confluent 10 cm dish of HeLa cells. Cells were trypsinized, washed with cold PBS and resuspended in cold lysis buffer (10 mM Tris pH 7.4; 150 mM NaCl; 0.15% NP-40; RNase inhibitor). The suspension was layered over 1 mL sucrose buffer (10 mM Tris pH 7.4; 150 mM NaCl; 24% sucrose) after incubation for 5′ on ice and centrifuged for 5 min at 3500*g*. Supernatant was discarded and the remaining nuclear pellet was briefly rinsed with cold PBS and resuspended in 250 µL Glycerol buffer (20 mM Tris pH 7.4; 75 mM NaCl; 0.5 mM EDTA; 50% Glycerol). We added 250 µL nuclear lysis buffer (10 mM Tris pH 7.4, 7.5 mM $MgCl_2$, 0.2 mM EDTA, 0.3 M NaCl, 1 M Urea, 1% NP-40) followed by vortexing for 4 sec. Samples were centrifuged for 2 min at full speed after incubation on ice for 2′ and the supernatant containing the nucleoplasmic fraction discarded. The remaining chromatin pellet was briefly rinsed with cold PBS and resuspended in TRIzol for RNA isolation.

Clinical samples and RNA isolation

The clinical samples were obtained directly from the OR (Operating Room) on ice. All samples were snap frozen in liquid nitrogen within 30′. The samples were then stored at −80°C and total RNA isolated with Qiagen RNA purification columns.

Design of capture probes

Capture probes were designed as xGEN Lockdown probes from IDT with the minimal recommended length of 120 nt. For each pri-miRNA targeted by the library, we made four probes, two upstream and two downstream from the Microprocessor processing site. The two probes on either side of the Microprocessing site were designed to be partially overlapping. For some pri-miRNAs, for example, miR222-3p and miR34-5p, we could only design one probe with sufficient specificity. Probe sequences are given in Supplemental Table 3.

Enrichment of RNA sequencing library and NGS protocol

Targeted RNA sequencing libraries were enriched according to the xGEN guidelines with minor modifications. In brief, 12 barcoded RNA sequencing libraries were pooled and 500 ng of combined libraries used for each target enrichment. Blocking oligos were added to the pooled libraries, samples were dried in a SpeedVac and resuspended in Hybridization Buffer including the custom enrichment probes and hybridized at 65° for 4 h after a 30 sec incubation at 95°. Probes were received at 3 pmole/

probe and resuspended in 30 µL TE to a storage stock at 0.1 pmole/probe. A total of 2 µL of the storage stock was mixed with 198 µL TE to have the dilution stock at 1 fmole/probe. We obtained the working stock by mixing 4 µL of the dilution stock with 6 µL TE to reach a concentration of 400 amole/probe. For each capture experiment we added 1 µL of the working stock and 3 µL TE.

Streptavidin beads were washed twice in Bead Wash Buffer before the capture reaction of custom enrichment probes and the targeted sequencing library. Samples were incubated 45 min at 65° followed by a normal wash and two times high-stringency heated washes followed by an additional wash at room temperature and elution of the captured library with nuclease-free water. The captured libraries were amplified with a polymerase chain reaction (PCR) using KAPA HiFi Hotstart polymerase for 14 cycles. The postcapture PCR fragment was purified with Agencourt AMPure XP beads before sequencing of the targeted libraries. Library preparation was performed using the TruSeq Stranded Total RNA Kit (Illumina). Sequencing was performed on an Illumina HiSeq 2500 instrument as 2 × 50 bp stranded sequencing.

Data analysis and calculation of processing efficiency

Raw reads were inspected for quality using FastQC (v0.11.5). Reads were then trimmed with skewer (version 0.2.2) for removing the adapters and low quality reads. Processed reads were aligned to the reference genome (human, GENCODE release 27, GRCh38.p10) using STAR aligner (version 2.5.3a). bedtools "genomecov" was run on the aligned reads with the "-split" option for calculating the coverage. The coverage was multiplied for a scaling factor (using the parameter "-scale") that is obtained by dividing 1 billion/(number of mapped reads × read size). The BED graph was converted to bigWig using bedGraphToBigWig. Custom scripts were used to calculate read count averages and the processing efficiency was calculated by counting the reads covering the pre-miRNA region divided by the reads covering the two 100 nt flanking regions of the pri-miRNA (starting 20 nt from the stem to avoid noise in read coverage), and normalized to the length of each pri-miRNA in nts (Fig. 1A). Pre-miRNAs were annotated from the sequence of the mature miRNAs so that each pre-miRNA starts with the first base of the 5P miRNA and ends with the last base of the 3P miRNA.

DATA DEPOSITION

All sequencing data have been deposited to GEO under accession numbers GSE148756 and GSE149631.

SUPPLEMENTAL MATERIAL

Supplemental material is available for this article.

ACKNOWLEDGMENTS

Author contributions: Conceived experiments: T.C., U.A.V.Ø.; performed experiments: T.C., U.A.V.Ø.; analyzed data: E.N., B.L., L.C.; contributed clinical samples: J.B.A., J.U.M.; supervised computational work: E.N., J.P., G.G.T.; supervised experimental work: U.A.V.Ø.; interpreted data: T.C., E.N., B.L., L.C., J.P., G.G.T., U.A.V.Ø.; secured funding: U.A.V.Ø.; wrote the manuscript: T.C., B.L., U.A.V.Ø.; commented on the manuscript and approved the final version: all authors.

REFERENCES

Alarcón CR, Lee H, Goodarzi H, Halberg N, Tavazoie SF. 2015. N^6-methyladenosine marks primary microRNAs for processing. *Nature* **519**: 482–485. doi:10.1038/nature14281

Auyeung VC, Ulitsky I, McGeary SE, Bartel DP. 2013. Beyond secondary structure: primary-sequence determinants license Pri-miRNA hairpins for processing. *Cell* **152**: 844–858. doi:10.1016/j.cell.2013.01.031

Conrad T, Marsico A, Gehre M, Ørom UA. 2014. Microprocessor activity controls differential miRNA biogenesis in vivo. *Cell Rep* **9**: 542–554. doi:10.1016/j.celrep.2014.09.007

Davis BN, Hilyard AC, Lagna G, Hata A. 2008. SMAD proteins control DROSHA-mediated microRNA maturation. *Nature* **454**: 56–61. doi:10.1038/nature07086

Farazi TA, Horlings HM, Ten Hoeve JJ, Mihailovic A, Halfwerk H, Morozov P, Brown M, Hafner M, Reyal F, Van Kouwenhove M, et al. 2011. MicroRNA sequence and expression analysis in breast tumors by deep sequencing. *Cancer Res* **71**: 4443–4453. doi:10.1158/0008-5472.CAN-11-0608

Han B, Chu C, Su X, Zhang N, Zhou L, Zhang M, Yang S, Shi L, Zhao B, Niu Y, et al. 2020. N^6-methyladenosine-dependent primary microRNA-126 processing activated PI3K-AKT-mTOR pathway drove the development of pulmonary fibrosis induced by nano-scale carbon black particles in rats. *Nanotoxicology* **14**: 1–20. doi:10.1080/17435390.2019.1661041

Kim VN, Han J, Siomi MC. 2009. Biogenesis of small RNAs in animals. *Nat Rev Mol Cell Biol* **10**: 126–139. doi:10.1038/nrm2632

Li L, Song Y, Shi X, Liu J, Xiong S, Chen W, Fu Q, Huang Z, Gu N, Zhang R. 2018. The landscape of miRNA editing in animals and its impact on miRNA biogenesis and targeting. *Genome Res* **28**: 132–143. doi:10.1101/gr.224386.117

Louloupi A, Ntini E, Liz J, Ørom UA. 2017. Microprocessor dynamics shows co- and post-transcriptional processing of pri-miRNAs. *RNA* **23**: 892–898. doi:10.1261/rna.060715.117

Mori M, Triboulet R, Mohseni M, Schlegelmilch K, Shrestha K, Camargo FD, Gregory RI. 2014. Hippo signaling regulates microprocessor and links cell-density-dependent miRNA biogenesis to cancer. *Cell* **156**: 893–906. doi:10.1016/j.cell.2013.12.043

Suzuki HI, Yamagata K, Sugimoto K, Iwamoto T, Kato S, Miyazono K. 2009. Modulation of microRNA processing by p53. *Nature* **460**: 529–533. doi:10.1038/nature08199

Volinia S, Galasso M, Sana ME, Wise TF, Palatini J, Huebner K, Croce CM. 2012. Breast cancer signatures for invasiveness and prognosis defined by deep sequencing of microRNA. *Proc Natl Acad Sci* **109**: 3024–3029. doi:10.1073/pnas.1200010109

Witteveldt J, Ivens A, Macias S. 2018. Inhibition of microprocessor function during the activation of the type I interferon response. *Cell Rep* **23**: 3275–3285. doi:10.1016/j.celrep.2018.05.049

Wong NW, Chen Y, Chen S, Wang X. 2018. *OncomiR*: an online resource for exploring pan-cancer microRNA dysregulation. *Bioinformatics* **34**: 713–715. doi:10.1093/bioinformatics/btx627