Nucleic Acids Research, 2025, **53**, gkaf070
https://doi.org/10.1093/nar/gkaf070
Advance access publication date: 18 February 2025
Genomics

The human genome encodes a multitude of novel miRNAs

Fan Gao[1,2,†], Fang Wang[3,4,†], Yue Chen[3,†], Bolin Deng[3], Fujian Yang[3], Huifen Cao[3], Junjie Chen[1], Huiling Chen[2], Fei Qi[1,*], Philipp Kapranov[1,*]

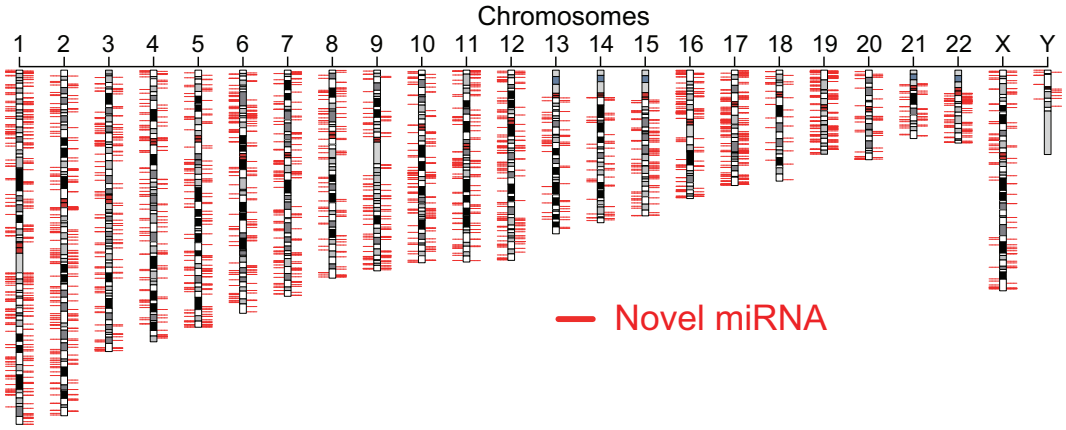
[1]State Key Laboratory of Cellular Stress Biology, School of Life Sciences, Xiamen University, Xiamen 361102, China
[2]Xiamen Institute for Food and Drug Quality Control, 33 Haishan Road, Xiamen 361012, China
[3]Institute of Genomics, School of Medicine, Huaqiao University, 668 Jimei Road, Xiamen 361021, China
[4]Institute of Rare Diseases, West China Hospital of Sichuan University, Chengdu, Sichuan 610041, China

[*]To whom correspondence should be addressed. Email: philippk08@hotmail.com
Correspondence may also be addressed to Fei Qi. Email: qifei@xmu.edu.cn
[†]These authors contributed equally to this work.

Abstract

Human cells generate a vast complexity of noncoding RNAs, the "RNA dark matter," which includes a vast small RNA (sRNA) transcriptome. The biogenesis, biological relevance, and mechanisms of action of most of these transcripts remain unknown, and they are widely assumed to represent degradation products. Here, we aimed to functionally characterize human sRNA transcriptome by attempting to answer the following question—can a significant number of novel sRNAs correspond to novel members of known classes, specifically, microRNAs (miRNAs)? By developing and validating a miRNA discovery pipeline, we show that at least 2726 novel canonical miRNAs, majority of which represent novel miRNA families, exist in just one human cell line compared to just 1914 known miRNA loci. Moreover, potentially tens of thousands of miRNAs remain to be discovered. Strikingly, many novel miRNAs map to exons of protein-coding genes emphasizing a complex and interleaved architecture of the genome. The existence of so many novel members of a functional class of sRNAs suggest that the human sRNA transcriptome harbors a multitude of novel regulatory molecules. Overall, these results suggest that we are at the very beginning of understanding the true functional complexity of the sRNA component of the "RNA dark matter."

Graphical abstract



Introduction

In the recent two decades, the existence of pervasive transcription of the human genome which results in generation of a very complex pattern of long noncoding (lnc) RNAs, defined by transcripts >200 nt, has been well-established [1–3]. On the other hand, multiple transcriptome mapping efforts have also shown that mammalian cells also possess a highly complex small RNA (sRNA, <200 nt) transcriptome [4–13]. The first indication of the complexity of the mammalian sRNA transcriptome was provided by a 2007 study that detected ~450K sRNAs in two human cell lines using high-density tiling arrays [4]. The high complexity of the human sRNA transcriptome was then confirmed in multiple subsequent transcriptome studies that relied on next-generation sequencing (NGS) [5–11]. For example, the study by the ENCODE consortium detected >150K novel sRNAs in just one human cell line [7]. While multiple mammalian sRNAs belong to the well-characterized functional classes of sRNAs, such

Received: April 16, 2024. **Revised:** January 22, 2025. **Editorial Decision:** January 25, 2025. **Accepted:** January 28, 2025

as microRNAs (miRNAs), small nucleolar RNAs (snoRNAs), small nuclear RNAs (snRNAs), and others, the vast majority of sRNAs identified in these transcriptome surveys represented novel sRNAs that have not been previously annotated [14–16]. For example, the ENCODE consortium found ∼7K annotated sRNAs compared to >150K novel sRNAs [7].

Thus, the situation with the mammalian sRNA transcriptome resembles that of the lncRNA transcriptome where most of the detected transcription has been found outside of the annotated exons and a number of different transcript classes could be identified [17, 18]. However, even though the functions and mechanisms of action of lncRNAs remain a subject of debate, the biogenesis of many of them is well-understood [1], and a number of different models that explain functionality and mechanism of action have been proposed [19, 20]. On the other hand, much less is known about biogeneses, functions, or mechanisms of action of most of the components of novel sRNA transcriptome leading to the prevailing assumption that most of them represent stable products of degradation of long transcripts or transcription by-products [21]. Nonetheless, we have recently shown using high-throughput phenotypic assays in cultured human cells that a surprisingly large fraction of novel sRNAs can have biological significance [22]. These results imply that novel sRNA transcriptome might represent a major source of hitherto unexplored regulatory RNA molecules and warrant further exploration.

Considering the many unknowns of the novel sRNA transcriptome, it is very hard to annotate and classify these transcripts into various sRNA classes based on the shared biogenesis, function or mechanistic aspects. In fact, novel sRNAs have been mostly classified based on the genomic features with which they associate [23, 24], e.g. several classes of novel sRNAs, such as promoter-associated small RNAs (PASRs) [4], transcription start site (TSS)-associated RNAs (TSSa-RNAs) [5], and transcription initiation RNAs (tiRNAs) [9], have been found associated with promoters and transcriptional start sites of genes. Novel sRNAs have also been found associated with gene termini, as exemplified by termini-associated short RNAs (TASRs) [4, 10] and splice sites as represented by splice-site RNAs (spliRNAs) [8]. However, such annotations provide limited insight into functionality of sRNAs since it is quite possible that sRNAs associated with the same genomic features might have very different properties.

Therefore, in an effort to functionally annotate the novel sRNA transcriptome, we explored a possibility that many novel sRNAs might represent novel members of a known functionally relevant class of sRNAs—miRNAs. To achieve this, we developed a pipeline to discover novel canonical miRNAs which have to satisfy two conditions: (i) being generated *in vivo* using the canonical Drosha/DGCR8 miRNA biogenesis pathway and (ii) possessing correct RNA secondary structure of Drosha-dependent miRNAs. We further validated the pipeline by detecting *in vivo* signatures of Dicer cleavage in the novel miRNAs. Using this approach, we discovered thousands of novel miRNAs in just one human cell line, including many that map to exons of known genes, and we estimate that tens of thousands of others may exist. Overall, our results are consistent with a highly complex and interleaved genomic organization where in multiple functional elements and transcripts share the same genomic sequence [25] and further highlight the complexity of the transcriptional output and architecture of the functional elements in the human genome.

Materials and methods

Biological material

Human chronic myelogenous leukemia and embryonic kidney cell lines K562 and 293FT were obtained from the Cell Bank of Chinese Academy of Sciences and National Infrastructure of Cell Line Resource, respectively. Cells were cultured in Roswell Park Memorial Institute (RPMI) 1640 medium (Thermo Fisher Scientific, USA) and Dulbecco's Modified Eagle Medium (DMEM, Sigma), respectively, supplemented with 10% (v/v) fetal bovine serum (Thermo Fisher Scientific, USA) and 1% pen/strep (v/v) (Thermo Fisher Scientific, USA) at 37°C in 5% CO_2.

Generation of a stable inducible Drosha knockdown system and small RNA-seq

The lentiviral plasmid pHS-ASR-ZQ021 that contained Drosha small hairpin RNA (shRNA) sequence under the control of a Dox-inducible H1 promoter and TetR protein stabilized by fusion with enhanced green fluorescent protein (EGFP) was generated by SyngenTech (Beijing, China) and confirmed by Sanger sequencing. The lentivirus particles were generated by transfecting the 293FT packaging cell line with the pHS-ASR-ZQ021 and used to transfect the K562 cell line at the multiplicity of infection (MOI) of 15. Mixed population of transfected K562 cells containing the Drosha shRNA sequence was selected by the flow cytometry (BD CytoFLEX) using EGFP fluorescence and expanded. From this population, individual cells were then selected by the flow cytometry (BD CytoFLEX) and cultured 5–7 days to generate monoclonal populations. Out of those, six monoclonal populations were tested for the Drosha knockdown efficiency and the best one was selected to establish a stable inducible Drosha knockdown system. These experiments were outsourced to SyngenTech (Beijing, China).

For validation of the system and discovery of novel miRNAs, the K562 cells harboring Drosha shRNA were plated into 3 ml of culture medium supplemented with 1 μg/ml Dox (doxycycline, Macklin Inc., 24390-14-5) at a density of 5×10^5 cells/ml in six-well plates and incubated for 3, 13, and 31 days, with Dox continuously present in the medium. The cells were also grown in parallel without Dox as a control. For each well, the medium with or without Dox was changed every 2 days. Two independent biological replicates were performed for the 3-day treatment, and three for the 13- and 31-day treatments. After the incubation, aliquots of two million cells from each biological replicate were harvested and used for total RNA isolation with TRNzol Universal reagent (TIANGEN, Beijing, China), following the manufacturer's protocol. The knockdown of Drosha mRNA was assessed using reverse-transcription quantitative polymerase chain reaction (RT-qPCR) using primers TCTACAGTGGTTGGAACGAG and ACTCACACTCGGATTCACTG against Drosha mRNA and GAPDH mRNA as a control. Each RT-qPCR reaction was performed in three separate wells. The RT-qPCR was performed on the Mx3005P cycler (Agilent Technologies, Inc.). The C_t values were analyzed using MxPro software (Agilent Technologies, Inc.) with Comparative Quantitation (Calibrator) settings.

Construction of sRNA-seq libraries was conducted with NEB Next® Multiplex Small RNA Library Prep Set for Illumina® (NEB). The procedure selects for sRNAs with 5′ phosphate and 3′ hydroxyl termini. After the library construc-

tion, the polymerase chain reaction (PCR) products corresponding to sRNAs in the range of 20–40 nt were purified by denaturing polyacrylamide gel electrophoresis (PAGE) and used for sRNA-seq on the Illumina platform (Hiseq 2500) and single-end 50-base strategy. The library construction and sequencing were performed by Novogene Inc. (Beijing).

Small RNA-seq data analysis

The raw reads that contained any of the following features were removed: (i) >10% of bases were *N*, (ii) >50% of bases had low (≤5) quality scores, (iii) sequence adapters, or (iv) long homopolymeric tracks. The quality-filtered reads were then aligned to the reference human genome (GRCh37/hg19) using Bowtie2 (v2.2.9) with the default settings. The alignments with unique 5′ and 3′ coordinates were generated by Samtools (v1.9) in a strand-specific fashion and used for the subsequent analyses after removing the alignments that overlapped the RNA family of repeats as annotated by the RepeatMasker track (https://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/rmsk.txt.gz) of the UCSC Genome Browser.

The RPKM values for each annotated miRNA from miRBase v22 (https://www.mirbase.org/) or a unique sRNA alignment *i* were calculated for each sample as shown below, where *RC* denotes the read counts for the sRNA *i* in the sample *j*, *TRL* is the total read length in the sample *j*, and *TRC* is the total read count in sample *j*. Only reads 5′ and 3′ ends of which matched the corresponding termini of the annotated miRBase miRNAs or unique sRNA alignment were used for the reads per kilobase per million mapped reads (RPKM) calculations.

$$\mathrm{RPKM}_{i,j} = \frac{\mathrm{RC}_{i,j} \times 10^9}{\mathrm{TRL}_j \times \mathrm{TRC}_j}$$

To determine the effect of Drosha knockdown on annotated miRBase miRNAs, we used adjusted fold change (*AFC*). *AFC* avoids a problem of division by zero for miRNAs with no read counts in the control samples. *AFC* of 0.5 represents no change, while down- and upregulated transcripts would have *AFC* in the ranges of [0, 0.5) and (0.5, 1], respectively. The formula to calculate *AFC* for each sRNA *i* in each sample *j* (each biological replicate for each time point of 3, 13, or 31 days) is shown below.

$$AFC_{i,j} = \frac{\mathrm{RPKM}(+\mathrm{Dox})_{i,j}}{\mathrm{RPKM}(-\mathrm{Dox})_{i,j} + \mathrm{RPKM}(+\mathrm{Dox})_{i,j}}$$

To discover novel Drosha-dependent sRNAs, we used two metrics. First, the expression fold change (*FC*) of each unique sRNA alignment *i* in each sample *j* (each biological replicate for each time point of 13 or 31 days) was calculated as shown below.

$$FC_{i,j} = \frac{\mathrm{RPKM}(+\mathrm{Dox})_{i,j}}{\mathrm{RPKM}(-\mathrm{Dox})_{i,j}}$$

We then calculated the average fold change ($ave.FC_{i,t}$) for each alignment *i* across in each time point *t* (13 or 31 days).

Second, we calculated statistical significance of depletion for each sRNA alignment in response to Dox using two-sided Student's *t*-test with six-pair samples (three biological replicates for each time point of 13 or 31 days). Only sRNAs with detectable expression (RPKM >0) in at least two of the six −Dox samples were used in this analysis. The *P*-values were then adjusted for multiple comparisons with the Benjamini–Hochberg method in the R environment to get the false discovery rate (FDR). Drosha-dependent sRNAs were then defined as those that satisfied two conditions: (i) *ave.FC* had to be <1 in both the 13- and 31-day time points and (ii) FDR had to be <0.2.

Known genes and lncRNAs were downloaded from GENCODE release 42 (https://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_42/). The coordinates of 407 vlincRNAs identified in K562 cell line were taken from St Laurent *et al.* [26]. The ENCODE small RNA-seq data were downloaded from the UCSC Genome Browser (http://www.genome.ucsc.edu/cgi-bin/hgFileUi?db=hg19&g=wgEncodeCshlShortRnaSeq). For the evolutionary conservation analysis, the genomic coordinates of the homologous sequences in the genomes of mouse (mm10), marmoset (calJac3), and bushbaby (otoGar1) were determined by the LiftOver tool from the UCSC Genome Browser. The coordinates of the sRNAs bound to the Argonaute (AGO) proteins and Dicer binding sites were downloaded from Gene Expression Omnibus (GEO) using the following GEO accession numbers: GSE55331 [27], GSM721075 [13], and GSE55324 [27]. Since AGO proteins bind mature miRNAs, the overlap was performed using a 2-base shift at either the 5′ or 3′ ends. However, since Dicer binds to a longer pre-miRNA, the overlap with the Dicer dataset required at least a 1-base overlap. The AGO-qCLASH data [28, 29] were downloaded from GEO using the following accession numbers: GSM5015717, GSM5015718, GSM5015719, GSM5015720, and GSM5015721. A sequence of each hybrid read was split into the first 25 nt and the remaining sequence that were then separately aligned to the genome and only unique alignments were kept. The alignments corresponding to the first 25 nt of the hybrid reads were overlapped with the coordinates of the known and novel miRNAs. The alignments of the remaining sequences were overlapped with the coordinates of the annotated genes. The BEDTools (v2.25.0) suite was used for all operations related to genomic overlap. All overlaps were performed in a strand-specific fashion with the exception of overlaps with the different classes of repeats.

To assign novel miRNAs to the existing miRNA families, the covariance models of 4170 RNA families were downloaded from the Rfam (release 14.10) database [30]. The nucleotide sequences of all the novel miRNAs together with their 100 nt upstream and 100 nt downstream genomic sequences were compared with these models using the Infernal software (version 1.1.4) [31] with the default or the "-rfam" parameter that sets stringent filters. For each miRNA, only hits on the same genomic strand and with the minimal *E*-value were kept.

The odds ratio (*OR*) of enrichment of the overlap of sRNAs with a genomic element *i* relative to the random chance was calculated as follows:

$$OR_i = \frac{LS_i/TS_i}{TL_i/TLG},$$

where LS_i is the length of the sRNAs overlapping the genomic element *i*, TS_i is the total length of the sRNAs mapping to genome, TL_i is the total length of the genomic element *i*, and *TLG* is the total length of the genome. Note that the total length of the RNA family of repeats were removed from L_i and *TLG*.

Results

Establishing a stable inducible Drosha knockdown system

To identify novel sRNAs that could potentially correspond to novel miRNAs, we have chosen an *in vivo* approach by profiling sRNA transcriptome in cultured cells depleted of the RNase III endonuclease Drosha that together with the DGCR8 protein forms the Microprocessor complex responsible for the first critical step in biogenesis of the majority of known miRNA [32–35]. During this step, pre-miRNA hairpins are cleaved from the long primary (pri-miRNA) transcripts and later processed to mature miRNAs by Dicer [36], and therefore depletion of the Drosha enzyme should also deplete the mature miRNAs. To achieve Drosha depletion, we transfected human leukemia cell line K562 with a lentiviral construct encoding shRNA targeting the Drosha mRNA under the control of Dox-inducible H1 promoter and TetR protein (Fig. 1A; "Materials and methods" section). After selecting stable transfected clones, we have tested depletion of Drosha mRNA using RT-qPCR in cells grown for 3, 13, and 31 days in the presence of Dox and in the control cells grown in parallel without Dox (Fig. 1B). We found that the 3-day incubation time was not sufficient to see significant depletion of Drosha mRNA (Fig. 1B and Supplementary Table S1). Instead, a longer incubation time of 13 days was required to observe a significant (on average 94%) depletion of the transcript in response to Dox (Fig. 1B and Supplementary Table S1). Considering that miRNAs are relatively stable [37] and therefore long time of Drosha depletion might be required to observe a change in the steady-state levels of these molecules, we also tested a longer incubation time point of 31 days and found that Drosha mRNA was also significantly (on average 91%) depleted in these samples as well (Fig. 1B and Supplementary Table S1).

To assess the sensitivity and specificity of the effect of the Drosha knockdown on its true targets in our system, we have first selected two subsets of known miRNAs annotated in the public miRNA database, miRBase [38]. One group represented 1289 mature miRNAs corresponding to 763 pre-miRNAs (Supplementary Table S2) that were proven to be generated by the Drosha/DGCR8 complex based on *in vitro* cleavage assays and will be referred to as "Drosha-dependent" miRNAs [39]. The second group consisted of 486 atypical mature miRNAs that have been previously shown not to require Drosha activity for their biogenesis and were mostly (466/486) represented by "mirtrons," generated by splicing of short introns [40–42], as well as some others such as sno- and tRNA-derived miRNAs and 5′ capped miRNAs (Supplementary Table S2). Atypical miRNAs share many features with the typical miRNAs such as downstream processing of pre-miRNAs by Dicer, export to cytosol via Exportin-5, loading into the RNA-induced silencing complex (RISC) and regulation of their targets [43]. Therefore, since the major difference between the typical and atypical miRNAs is the "Drosha-dependent" processing step, the latter group should represent a strict control for the specificity of the discovery of the "Drosha-dependent" sRNAs and will be referred to as "Drosha-independent" miRNAs. In addition, miRBase contains 972 mature miRNAs (Supplementary Table S2) that were shown not to be efficiently processed *in vitro* from the corresponding pri-miRNAs by the purified Drosha/DGCR8 complex [39] but not classified as atypical. Still, since it is conceivable that such "Drosha-inefficient" miRNAs could theoretically be generated by the Drosha/DGCR8 complex *in vivo*, we did not include them in the analysis of the specificity of the Drosha-depletion system (also see below). Finally, the Drosha-dependent status of the remaining 133 miRBase miRNAs has not be characterized and therefore these miRNAs were also not included in the analysis.

The effect of Drosha depletion in cells grown with and without Dox for 3, 13, and 31 days was evaluated using sRNA-seq analysis performed on sRNAs in the range of 20–40 nt. Consistent with little Drosha depletion observed at 3-day, no decrease in either group of miRNAs could be observed at that time point (Fig. 1C and Supplementary Table S3). However, the "Drosha-dependent" miRNAs exhibited a stark decrease in abundance in response to Dox at both 13- and 31-day time points while the "Drosha-independent" miRNAs have shown the opposite trend (Fig. 1C and Supplementary Table S3). These patterns were highly reproducible in each of the three biological replicas done for each of these time points (Fig. 1C and Supplementary Table S3). The results above strongly supported high level of specificity of our Drosha knockdown system towards true Drosha/DGCR8 substrates. However, downregulation of sRNAs in response to Drosha depletion could also be caused by secondary effects and therefore such sRNAs would not represent real substrates of the Microprocessor complex. Therefore, we have developed additional filtering steps to select true miRNAs based on specific features of their RNA folding patterns.

Developing a deep learning model to select true novel Microprocessor substrates

Proper processing by the Microprocessor complex requires presence of specific sequence and structural features in the pre-miRNA hairpins as well in their immediate flanking DNA sequence [44–48]. Therefore, we have tested whether the true substrates of the Drosha/DGCR8 complex could be distinguished based on the presence of such features by the multi-branch convolutional neural network (MuStARD) method that can learn specific sequence-structure patterns from a specific class of user-defined sRNAs [49]. We used MuStARD to build a model that can recognize proper sequence-structure patterns around the "Drosha-dependent" miRNAs using sequences of 695 out of the 763 "Drosha-dependent" pre-miRNAs from miRBase and 5000 randomly-selected sequences from the human genome as respectively the positive and negative components of the input training set ("Materials and methods" section and Supplementary Table S4). The MuStARD algorithm works with a 100 base-long sequences; therefore, the sequences of the pre-miRNAs (on the order of 70 nt) were extended equally in each direction using flanking genome sequence. Based on the current knowledge of the recognition of its targets by the Drosha/DGCR8 complex [36], such 100 base-long sequences should contain most of the sequence–structure information required for proper recognition by the Microprocessor complex.

To validate the model, however, we used mature miRNAs since they would more properly represent the situation with the novel sRNAs found in this study for which pre-miRNA sequences are not known. However, since the relative position of a novel sRNA in the RNA structure is not known—the sRNA could be located close to either 5′ or 3′ end of

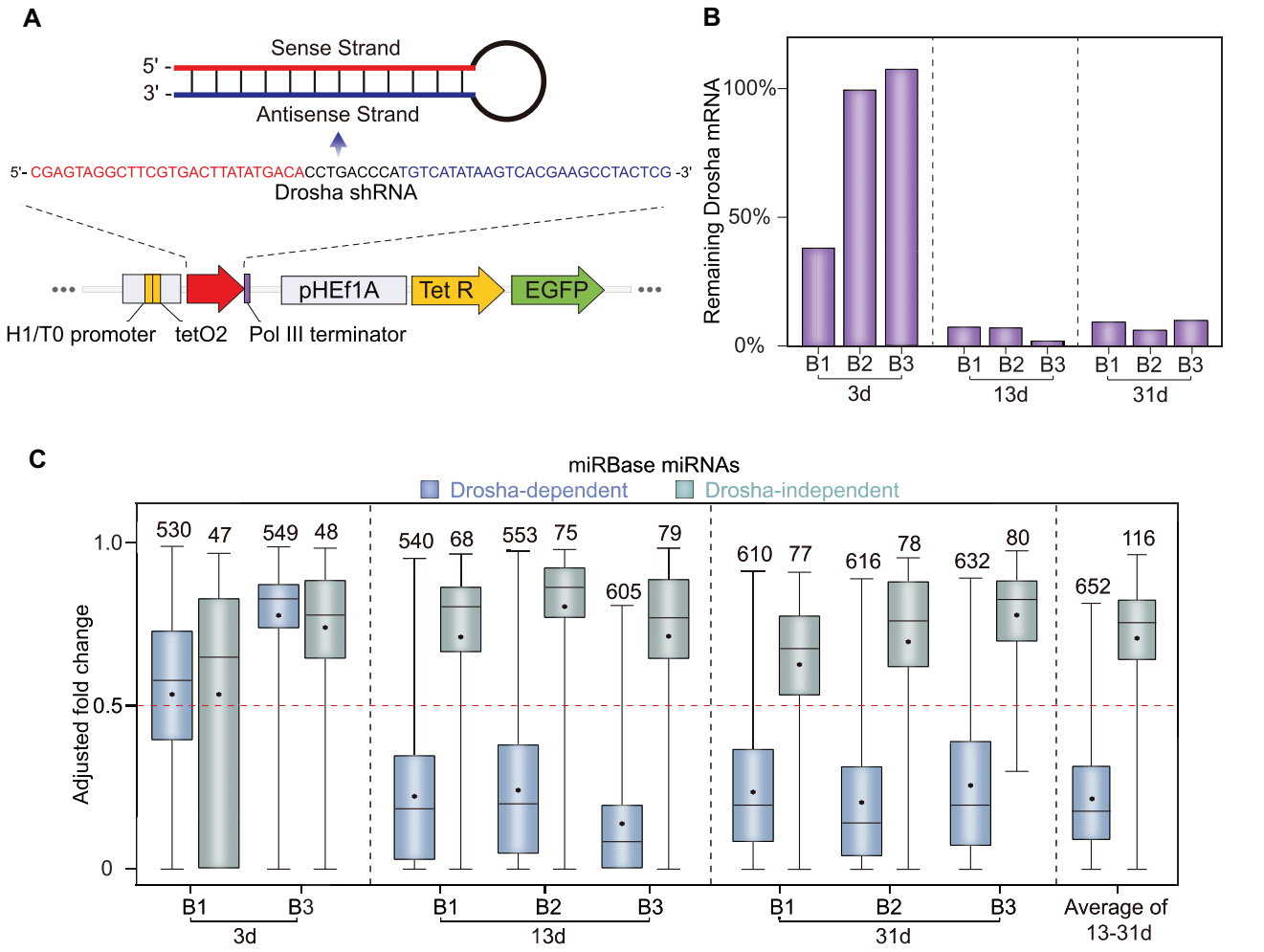


Figure 1. Establishing the inducible Drosha knockdown system. (**A**) Schematics of the stable Drosha shRNA expressing lentiviral cassette under the control of a Tet-On promoter. (**B**) Depletion of Drosha mRNA in the human K562 cells containing stably integrated Drosha shRNA cassette in response to growth in presence of Dox for variable amounts of time relative to the cells grown without Dox in parallel for each biological replicate (B1–B3). (**C**) Depletion of Drosha-dependent and Drosha-independent miRBase miRNAs in the Drosha depleted cells. The AFC (*Y*-axis) of [0, 0.5) and (0.5, 1] represents, respectively, decrease or increase in the expression levels while 0.5 (the dashed red line) means no change (see "Materials and methods" section for more details). Box plots indicate median (middle line), 25th, 75th percentile (box), and 1.5× interquartile range (whiskers) as well as the average (asterisk) for each biological replicate (B1–B3).

the hairpin recognized by the Drosha/DGCR8 complex—we have extended sequence of each miRNA to 100 bases using the flanking genomic sequence in three different ways: (i) extending by 10 nt at the 5′ end and extending to 100 nt by adding the remaining sequence at the 3′ end, (ii) extending by 10 nt at the 3′ end and extending to 100 nt by adding the remaining sequence at the 5′ end, and (iii) extending equally at both ends (Fig. 2A). For each sequence, MuStARD returned a score in the range of [0, 1] with the score of 1 signifying the highest probability of a sequence having the sequence–structure property of a positive training set, in this case "Drosha-dependent" miRNA. For each of the 695 "Drosha-dependent" pre-miRNAs, we used their corresponding 1173 mature miRNAs to calculate MuStARD score for each of the three extended sequences and selected the highest (Fig. 2A). In parallel, we also generated the MuStARD scores calculated using the same strategy for the 116 "Drosha-dependent" mature miRNAs not used in the model construction (the validation set). As can be seen from Fig. 2B, the model performed very well on both the training and the validation "Drosha-dependent" miRNA sequences—94.1% (1104/1173) and 89.7% (104/116) of miRNAs in these two respective categories had maximum MuStARD scores in the range of (0.5, 1] (Supplementary Table S4). For comparison, only 0.44% (22/5000) of the negative sequences had MuStARD scores in this range (Supplementary Table S4). Since ~90% of the "Drosha-dependent" miRNAs from the validation set had the maximum MuStARD score >0.5, we have selected this threshold for the downstream analysis.

We then further tested the specificity of the model on the "Drosha-independent" miRNAs. Importantly, none of the miRNAs from this control group were used in the model training either as part of the positive or negative sets. As shown in the Fig. 2B, the model could clearly separate the "Drosha-dependent" miRNAs from this control group: only 33.5% (163/486) of the mature "Drosha-independent" miRNAs had maximum MuStARD score >0.5 (Supplementary Table S4). Altogether, these results argue for a high level of specificity of the MuStARD model in identification of sequences that could be efficiently processed by the Microprocessor complex.

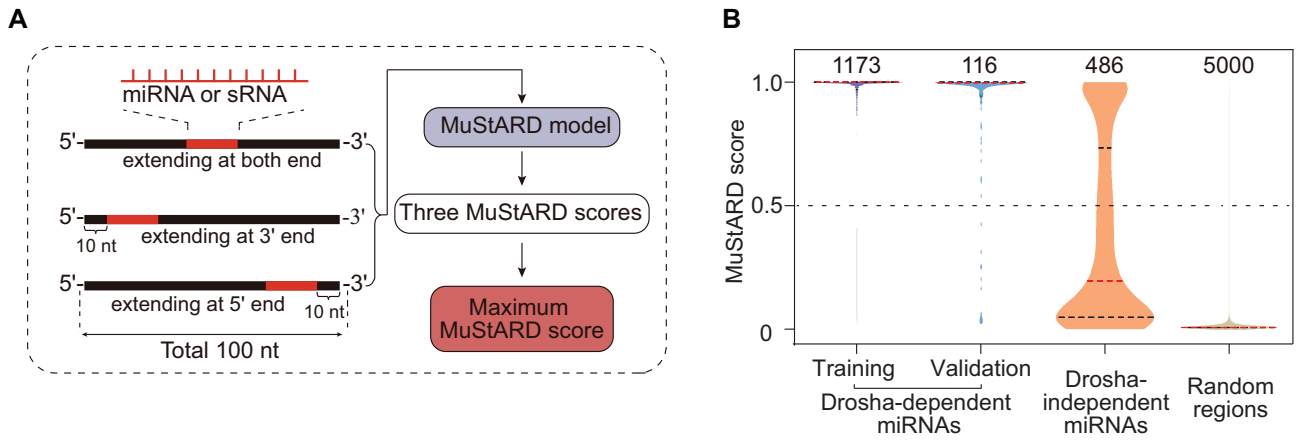


Figure 2. Establishing and validating the MuStARD model for predicting proper RNA secondary structure of Drosha-dependent miRNAs. (**A**) Schematic diagram of obtaining the MuStARD score for a miRBase miRNA or an sRNA. (**B**) Violin plots of the distribution of MuStARD scores obtained using a MuStARD model on the training and validation sets of miRBase Drosha-dependent miRNAs, miRBase Drosha-independent miRNAs, and random genomic regions. The dashed lines within the plots indicate (from bottom to top) 25th percentile, median, and 75th percentile. The numbers above indicate total numbers of sequences in each category.

Development of the genome-wide miRNA discovery pipeline

Given the high specificity of the genetic and AI-based miRNA discovery strategies, represented by respectively the Drosha knockdown system and MuStARD model, we have developed a discovery pipeline of novel miRNA that is based on application of both of these strategies as described in Fig. 3. The first part of the pipeline identifies sRNAs that are potentially processed by the Drosha enzyme. The second part further filters the candidates to identify sRNAs that have sequence and structure features of true and efficient substrates of the Drosha/DGCR8 complex. Finally, the performance of the model is validated based on the fraction of the resulting novel miRNAs that have evidence of *in vivo* Dicer cleavage as determined by the presence of the expected sRNAs derived from the opposite strand of the RNA duplex (Fig. 4A–C).

To identify "Drosha-dependent" novel sRNAs, we used the sRNA-seq data from the six pairs of the Dox treated 13- and 31-day samples and the corresponding −Dox controls represented by a total of 747 170 505 NGS reads, of which 650 556 475 could be mapped to the genome (Fig. 3 and Supplementary Fig. S1). For each of the 12 samples, we then further filtered the alignments to select a total of 642 057 268 uniquely mapping reads, coordinates of which were further collapsed to generate 20 620 641 alignments (∼1 M to ∼4.7 M per sample) with unique 5′ and 3′ coordinates that were used for the downstream analysis (Fig. 3 and Supplementary Fig. S1). As the next step, we excluded the alignments that overlapped the "RNA" class of repeats as annotated by the RepeatMasker database [50] to remove potential degradation products of highly abundant cellular RNAs, such as ribosomal RNAs (rRNAs), snRNAs, or transfer RNAs (tRNAs), that compose this class of genomic repeats. Then, for each of the 20 014 675 remaining alignments, we calculated fold change (FC) between each Dox-treated sample and the corresponding control (Fig. 3 and Supplementary Fig. S1; "Materials and methods" section). Therefore, for each sRNA alignment, we calculated FC in each of the six pairs of Dox-treated and control samples and then used these values to calculate the average FC (Supplementary Fig. S1 and "Materials and methods" section). Then, for each sRNA alignment, we calculated statistical significance of the FC using paired Student's *t*-test. We then used the raw *P*-values as input into the Benjamini–Hochberg procedure to select 435 158 alignments that decreased in response to the Drosha depletion with the average FC <1 and FDR <0.2 (Fig. 3 and "Materials and methods" section). Since different sRNA alignments could be derived from the same pre-miRNAs, we merged the coordinates of the 435 158 sRNA alignments to obtain 323 257 clusters (Fig. 3 and Supplementary Fig. S1). Of those, 322 101 clusters did not overlap miRBase pre-miRNAs and thus potentially represented novel miRNAs, while 1156 clusters corresponded to annotated miRBase miRNAs. Of the latter, most (1095/1156) clusters corresponded to mature miRBase miRNAs.

As the final step of the part 1, we have selected a single representative alignment with the maximum read depth for each cluster and only kept alignments in the length range of 20–25 nt consistent with the length distribution of the known miRNAs (Fig. 3 and Supplementary Fig. S1). As the result, we obtained 41 806 and 860 sRNAs that represented respectively potential novel miRNAs and annotated mature miRBase miRNAs (Fig. 3; Supplementary Tables S5 and 6). As expected, majority, 646/860 or 75.1%, of the miRBase miRNAs detected by our pipeline were represented by the "Drosha-dependent" miRNAs (Supplementary Table 6). Interestingly, another 169/860 (19.7%) were represented by the "Drosha-inefficient" miRNAs (Fig. 3 and Supplementary Table S6), suggesting that these transcripts could indeed be processed in a Drosha-dependent fashion *in vivo*, even though exhibiting poor processing *in vitro* [39]. The remaining 26 and 19 miRNAs were represented by respectively the "Drosha-independent" miRNAs and the miRNAs Drosha dependency of which has not been characterized *in vitro* (Fig. 3 and Supplementary Table S6).

To estimate the performance of the first part of the pipeline, we calculated two metrics. First, we estimated precision of the detection of boundaries of the annotated miRBase miRNAs by the alignments generated at the final part of the pipeline. As summarized in Fig. 3 and shown in the Fig. 5A–C for several examples, the boundaries of 45.4% of the 646 mature "Drosha-dependent" miRBase miRNAs were detected

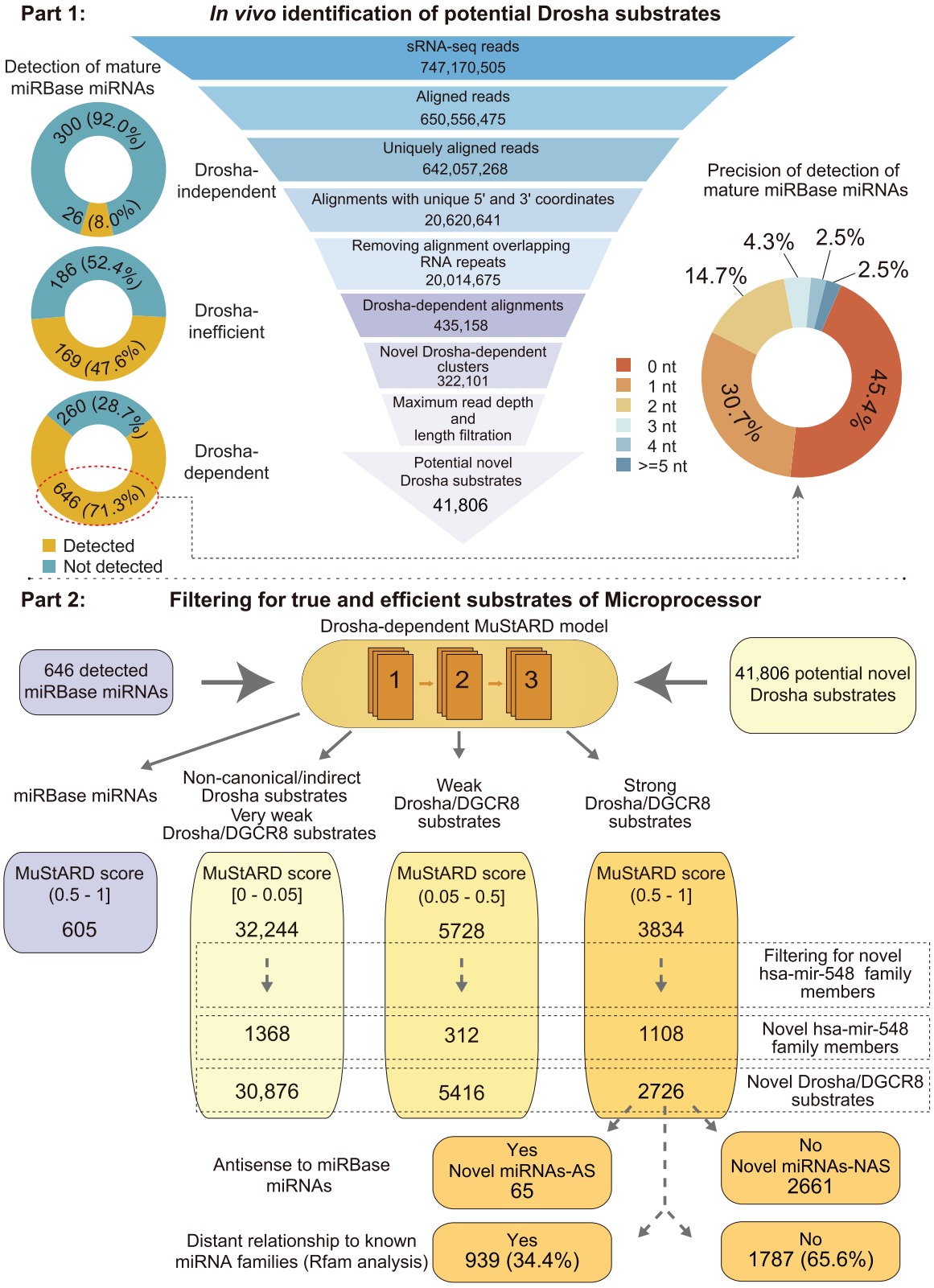


Figure 3. Schematic representation of the miRNA discovery pipeline. The two steps of the pipeline and the key numbers and metrics are shown.

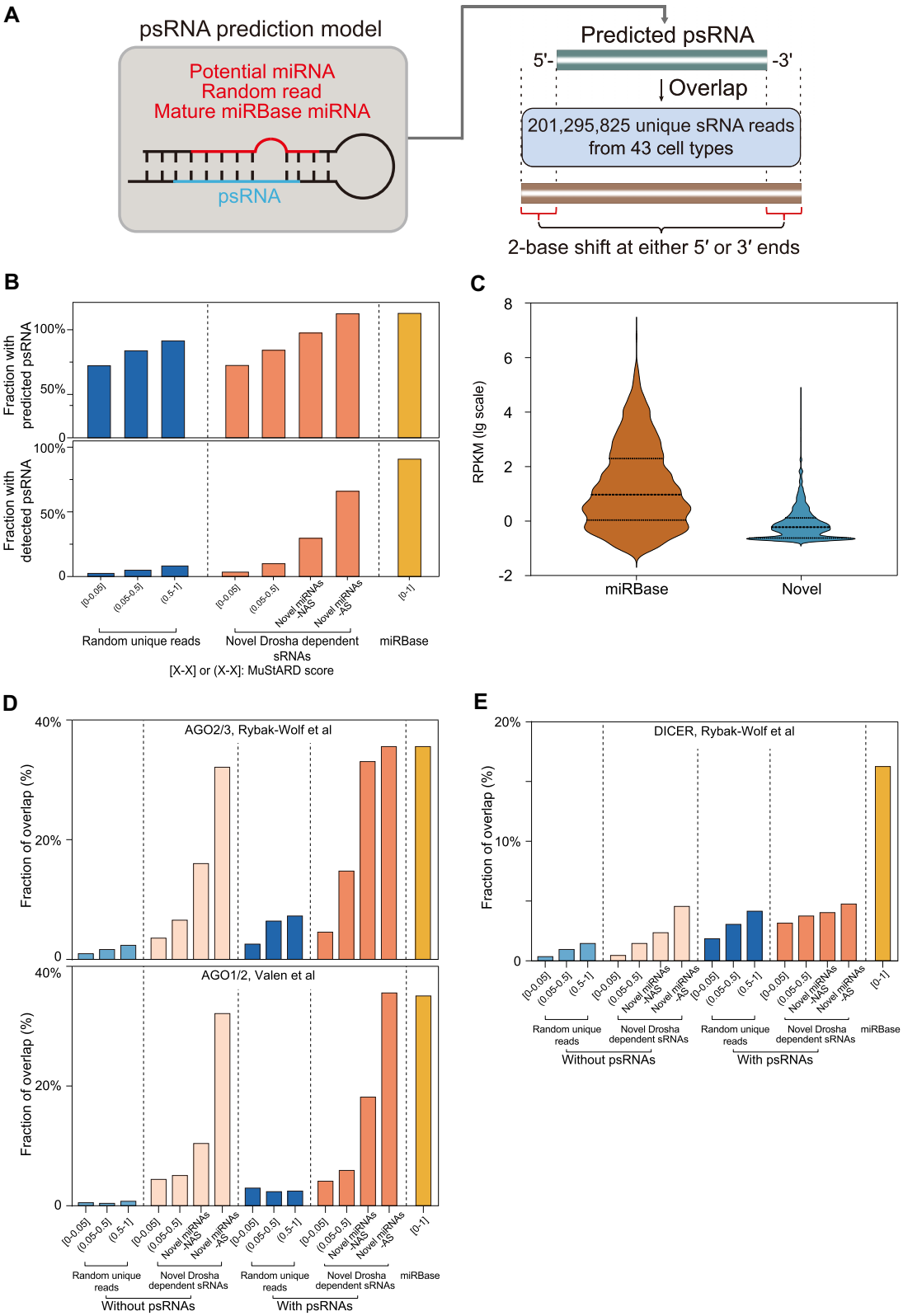


Figure 4. Obtaining evidence of the *in vivo* Dicer cleavage of various groups of miRNAs and sRNAs. (**A**) Schematics of the prediction and detection of psRNAs. (**B**) Fraction of predicted (top) and detected (bottom) psRNAs for each indicated group of miRNAs or sRNAs. (**C**) Violin plots showing the distributions of expression levels of miRBase miRNAs and novel miRNAs in K562 cells. The dashed lines within the plots indicate (from bottom to top) 25th percentile, median, and 75th percentile. (**D** and **E**) Fractions of sRNAs from each indicated group of miRNAs or sRNAs that overlap with either the AGO- (D) or Dicer-associated (E) sRNAs from the studies of Rybak-Wolf *et al.* [27] and Valen *et al.* [13].

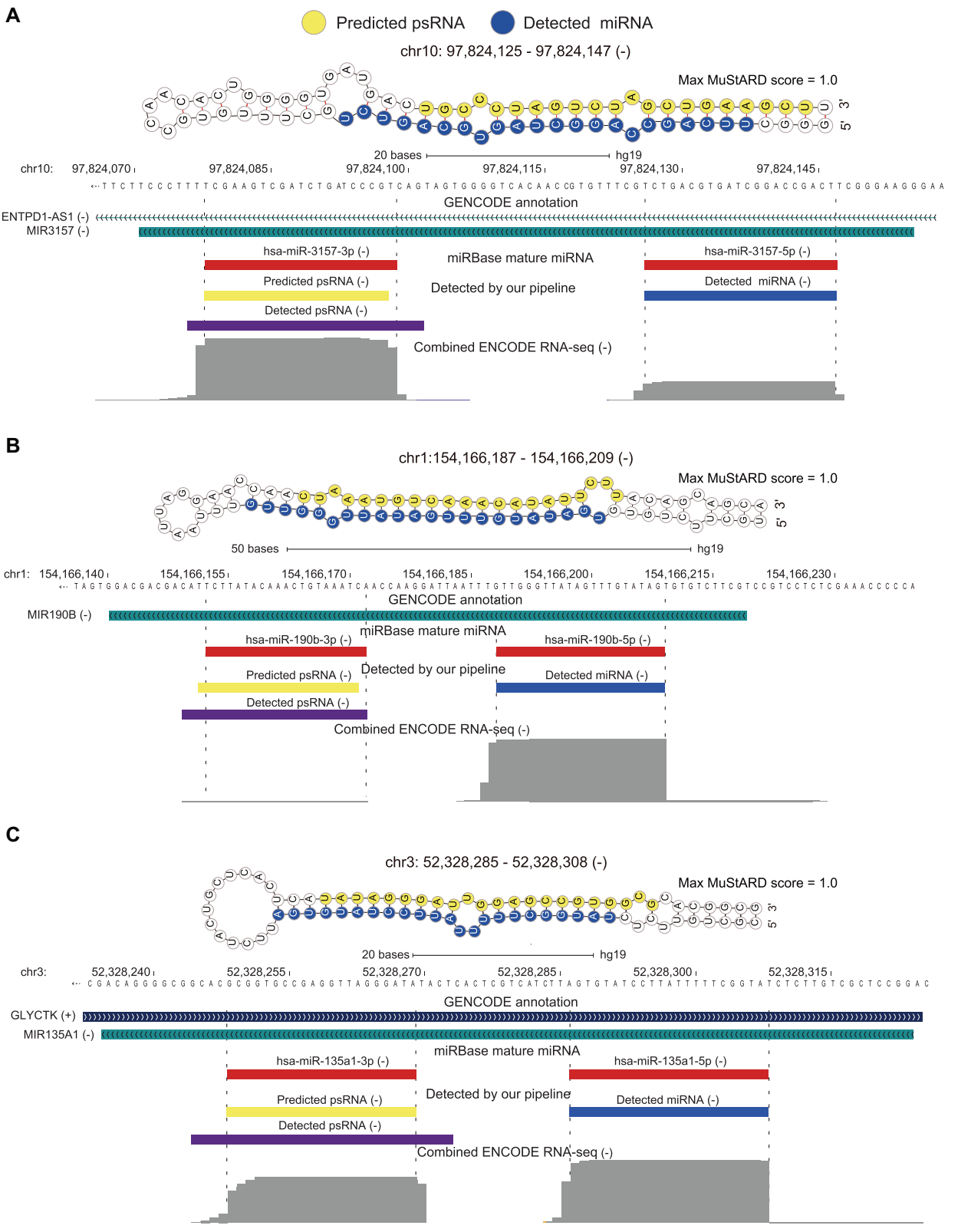


Figure 5. Example of detection of Drosha-dependent miRBase miRNAs using the pipeline developed in this study. Genomic contexts, miRBase annotations, and predicted targeting and psRNAs are shown for hsa-miR-3157 (**A**), hsa-miR-190b (**B**), and hsa-miR-135a1 (**C**) miRNAs. The secondary RNA structures are based on the predictions generated by the MuStARD program.

precisely (0 base shift) by our pipeline, while those of additional 30.7% and 14.7% were detected with respectively only 1- or 2-base shifts from both 5′ and 3′ ends combined (Supplementary Table S7). Thus, the boundaries of ∼91% of the annotated "Drosha-dependent" miRNAs were detected with relatively high precision (within two bases).

Second, we estimated the sensitivity and specificity of the first part of the pipeline by determining the fraction of the "Drosha-dependent" or "Drosha-independent" miRBase miRNAs that were classified as "Drosha-dependent" sRNAs. To estimate this, we first determined how many "Drosha-dependent" or "Drosha-independent" miRNAs were expressed in our samples using the total 20 620 641 sRNA alignments that were obtained from the Drosha-depleted or control samples prior to selecting Drosha-dependent sRNAs (Fig. 3). Overall, we could detect respectively 906 and 326 "Drosha-dependent" and "Drosha-independent" miRNAs (Supplementary Table S6). Of the 906 "Drosha-dependent" miRNAs, the pipeline correctly classified the aforementioned 646 or 71.3% as "Drosha-dependent" miRNAs. On the other hand, only 26 or 8% of the 326 atypical miRNAs were incorrectly classified as "Drosha-dependent" in our pipeline (Fig. 3 and Supplementary Table S6). In addition, we could detect expression of 355 "Drosha-inefficient" miRNAs, of which 169 (47.6%) could be classified as "Drosha-dependent" by our pipeline (Fig. 3 and Supplementary Table S6). Thus, the first part of the pipeline could correctly classify the majority (71.3%) of the truly positive "Drosha-dependent" miRNAs, while incorrectly classifying a small fraction (8%) of the truly negative "Drosha-independent" miRNAs (Supplementary Table S6). And, as expected, it had somewhat higher fraction of positive calls in an ambiguous category of "Drosha-inefficient" miRNAs that might contain true Drosha substrates.

As shown in the Supplementary Fig. S2, the 260 expressed "Drosha-dependent" miRBase miRNAs that were not detected by our pipeline also tended to be depleted in response to Drosha knockdown; however, the fold depletion was not as high as for the 646 "Drosha-dependent" miRNAs that were correctly classified by our pipeline (Supplementary Table S8). Such miRNAs could represent very stable sRNA species steady state levels of which are less responsive to Drosha knockdown. In summary, the part 1 of the pipeline could identify majority of the annotated miRBase miRNAs expressed in K562 and generated by the Drosha/DGCR8 complex with high precision, sensitivity and specificity.

We then applied the MuStARD model to the 41 806 novel miRNA candidates (Supplementary Table S5) and classified them into three groups. As expected, the vast majority (93.7% or 605/646) of the Drosha-dependent miRBase miRNAs detected in the part 1 of the pipeline had the maximum MuStARD scores >0.5 (Fig. 3). The corresponding fraction for the novel miRNA candidates was significantly smaller—only 3834 out of 41 806 (9.2%) sRNAs had the maximum MuStARD scores of >0.5 and will be referred to as "strong Drosha/DGCR8 substrates" (Fig. 3 and Supplementary Table S5). Still, as mentioned above, only 0.44% of the random sequences had such score (Supplementary Table S4), thus novel miRNA candidates generated by the part 1 of the pipeline were enriched 20.8-fold (9.2% versus 0.44%) in the sequences with the high (>0.5) maximum MuStARD scores compared to the random genomic regions. On the other hand, majority of the novel miRNA candidates, 32 244 out of 41 806 (∼77%), had the scores in the range of [0, 0.05] (Fig. 3 and Supplementary Table S5). In this regard, most of the random genomic regions (96.6% and Supplementary Table S4) also had scores in this range. Therefore, the miRNAs candidates with such low scores likely did not represent true miRNAs. Drosha has been reported to have noncanonical substrates and thus have other functions in addition to miRNA processing [51, 52]. It is therefore likely that sRNAs in this group represent such noncanonical substrates or, alternatively, they represent sRNAs levels of which are indirectly affected by Drosha depletion. Therefore, we will refer to this group as "noncanonical/indirect Drosha substrates." Conceivably, this category can also include very weak substrates of the Drosha/DGCR8 complex. Additional 5728 out of 41 806 (13.7%) novel miRNA had the scores in the range of (0.05, 0.5] (Fig. 3 and Supplementary Table S5). This group could contain weak substrates of Drosha/DGCR8 complex and will be referred to as "weak Drosha/DGCR8 substrates."

Based on sequence analysis, we realized that many novel miRNA candidates represented novel members of a large miRNA gene family hsa-mir-548. Therefore, we developed an additional filtering step in the pipeline to remove such sequences (Fig. 3). First, using the VSEARCH program [53], we found that all known members of the hsa-mir-548 family could be clustered together using 70% sequence identity level. Therefore, we used this threshold to cluster the miRNA candidates together with the known members of the hsa-mir-548 family and found that 1108/3834 (28.9%) of the candidates in the "strong Drosha/DGCR8 substrates" category corresponded to novel members of the hsa-mir-548 family. The novel hsa-mir-548 miRNAs were then removed from the subsequent analysis, resulting in a category of 2726 sRNAs representing "novel strong Drosha/DGCR8 substrates" that will be further referred to as "novel miRNAs" (Fig. 3 and Supplementary Table S9). Furthermore, using the same approach, we also removed novel members of the mir-548 gene family from the categories of 5728 "weak Drosha/DGCR8 substrates" and 32 244 "noncanonical/indirect Drosha substrates" to generate respectively 5416 and 30 876 "novel weak Drosha/DGCR8 substrates" and "novel noncanonical/indirect Drosha substrates" categories (Fig. 3).

Properties of novel miRNAs

Majority (∼2/3) of the 2726 "novel miRNAs" represented novel miRNA families while 939 of them (∼1/3) could be assigned to 289 distinct known miRNA families in the Rfam database (Supplementary Table S10 and "Materials and methods" section). Most of these miRNAs represented distant members of the families since the same search done using stringent parameters resulted in only 32 novel miRNAs assigned to 16 distinct miRNA families (Supplementary Table S10).

Interestingly, 65 out of 2726 "novel miRNAs" mapped antisense to miRBase miRNAs (Supplementary Table S9). These sRNA are likely processed from natural antisense transcripts which overlap some miRBase miRNAs. Such transcripts have been reported previously [54]. The same study also reported that the antisense regions could form hairpin structures based on the patterns of RNA editing mediated by adenosine deaminases acting on RNA (ADARs) [54]. Currently, only four sense-antisense miRNA pairs are annotated in the miRBase. Of the additional 65 antisense miRNAs found in this work,

one overlapped an antisense miRNA Hsa-Mir-337-as annotated in the MirGeneDB 2.1 database [55]. We will refer to the 65 novel antisense miRNAs as "novel miRNAs-AS." The remaining 2661 sRNAs will be referred to as "novel miRNAs-NAS," which also contained one sense-antisense miRNA pair. Among "novel miRNAs-NAS," two have been previously reported in the MirGeneDB 2.1 database under the names of Hsa-Novel-2 and Hsa-Novel-3 (Supplementary Table S9).

Evolutionary conservation of RNA structure is considered as one of the attributes of functional RNAs. Therefore, we tested presence of conserved pre-miRNA structures in three other species with different evolutionary distances from humans—a New World monkey (marmoset), a prosimian (bushbaby), and a rodent (mouse)—which split from the human lineage respectively ∼43, ∼64, and ∼76 million years ago [56, 57]. For each of the 2726 "novel miRNAs," we used the extended 100 nt sequences with the maximum MuStARD score (Fig. 2A) to search and extract the homologous sequences in the genomes of the three species ("Materials and methods" section). Then, we estimated how many of the homologous sequences might fold into Drosha-dependent pre-miRNA structures by calculating the MuStARD score for each sequence. A homologous sequence was considered to have conserved RNA fold if its MuStARD score was >0.5. Based on this analysis, we found that 19.8%, 4.1%, and 3.7% of the original 2726 human "novel miRNAs" are conserved in, respectively, the marmoset, bushbaby, and mouse genomes (Table 1 and Supplementary Table S11).

We then performed the same analysis for the 1286 miRBase Drosha-dependent miRNAs which contain well-studied miRNAs with widely accepted functions and biological significance [38]. Not surprisingly, the miRBase miRNAs exhibited a much higher conservation level—for the three species, the fractions of conserved RNA fold among the miRBase miRNAs ranged from ∼3- to ∼10-fold higher than those for the novel miRNAs (Table 1 and Supplementary Table S11). Nevertheless, the novel miRNAs showed a significantly higher conservation level than the 5000 random genomic regions as evidenced by ∼21- to 61-fold higher fractions of conserved RNA fold among the former (Table 1 and Supplementary Table S11).

A total of 648 (23.8%) novel human miRNAs exhibited conserved RNA folds in at least one of the three species with each species containing a unique set of conserved miRNAs (Supplementary Fig. S3 and Supplementary Table S11). A similar pattern was also observed for the miRBase miRNAs, while the corresponding number was 861 (67.0%, Supplementary Fig. S3 and Supplementary Table S11). Therefore, it is quite likely that more novel miRNAs identified in this work are conserved in genomes of species not tested in this work. Overall, these results suggest that the novel miRNAs are less conserved than the known miRBase miRNAs yet have a much higher conservation level than the random genomic regions (see "Discussion" section). This notion is further emphasized by the fact that only 21 (0.8%) novel human miRNAs had conserved RNA fold in all the three species compared to 293 (22.8%) for miRBase and 0 for the random regions (Table 1, Supplementary Fig. S3 and Supplementary Table S11).

Validation of the miRNA discovery pipeline

A true pre-miRNA generated by the Drosha/DGCR8 cleavage is subjected to further processing by Dicer riboendonuclease. This step results in an RNA duplex, containing the "guide" and "passenger" strands, of which typically only the former is loaded into the RISC complex [36]. Therefore, the evidence of Dicer processing as manifested by the presence of the passenger miRNA (psRNA) is a hallmark property of a true miRNA [58]. Thus, we used this property to validate the *bona fide* nature of the novel miRNAs discovered in this work using the following three steps outlined in the Fig. 4A: (i) prediction of genomic coordinates of potential psRNA for each miRNA candidate, (ii) detection of the potential psRNAs in sRNA-seq data, and (iii) comparing the results to the negative control sRNAs that are not expected to represent miRNAs.

To accomplish the first step, we have developed an automatic algorithm ("Materials and methods" section) to estimate genomic coordinates of a potential psRNA for each miRNA candidate based on the predicted RNA structure of the corresponding pre-miRNA, location of the pipeline-detected miRNA candidate in that structure, and the pattern of Dicer cleavage that creates 2 nt overhangs between the passenger and driver miRNAs [36]. Using this algorithm, we could predict genomic coordinates for 1247/1289 (97%) of the "Drosha-dependent" mature miRBase miRNAs (Fig. 4B and Supplementary Table S12). In this analysis, each miRBase miRNA was treated as a "guide" miRNA. Of the 1247 miRNAs, miRBase contained the coordinates of the corresponding psRNAs for 1012 miRNAs (Supplementary Table S13). The coordinates of the 840/1012 (83%) of the predicted psRNA were detected with 4-base shifts from both 5′ and 3′ ends combined compared to the coordinates provided by the miRBase, arguing for the high precision of our psRNA predicting algorithm (Supplementary Table S13).

Of the 2726 "novel miRNAs," only 118 mapped on the same strand and within 60 bp of each other and thus potentially representing the mature products of the same 59 pre-miRNAs. Thus, the vast majority (2667/2726 or 97.8%) of "novel miRNAs" represented different pre-miRNAs. Therefore, we used each of the 2726 "novel miRNAs" to predict the coordinates of the corresponding potential psRNAs using the extended sequence that produced that maximum MuStARD score. As the result, we successfully predicted psRNAs for 63/65 (96.9%) of "novel miRNAs-AS" and 2177/2661 (81.8%) of the "novel miRNAs-NAS" categories of miRNA candidates (Supplementary Table S12). In addition, using the same procedure, we predicted potential psRNAs for 3707/5416 (68.4%) and 17 501 out of 30 876 (56.7%) for the miRNA candidates from the respective categories of "novel weak Drosha/DGCR8 substrates" and "novel non-canonical/indirect Drosha substrates" (Fig. 4B and Supplementary Table S12). Interestingly, the fraction of predicted psRNAs was the highest for the miRBase miRNAs, and it correlated with the MuStARD score of the novel miRNAs and random reads (Fig. 4B and Supplementary Table S12). The lack of predicted psRNA is likely explained by the inability of an RNA to fold in a structure that even remotely resembles a pre-miRNA. As expected, the fraction of such sequences would be lowest in the miRBase miRNAs and it would increase among the novel sRNAs with the decrease in the MuStARD score.

Passenger miRNAs can be unstable, have lower abundance levels than the driver miRNAs and therefore are not always detectable even for known miRNAs [59]. In order to increase the sensitivity of the psRNA detection in the second step of the validation, we have included multiple additional

Table 1. Conservation of novel miRNAs

Human sequences	Species	Homologous sequences			Conserved Drosha-dependent pre-miRNA structures		
		Number	Fraction	Shared by three species	Number	Fraction	Shared by three species
Novel miRNAs (2726)	Marmoset	1654	60.7%	388 (14.2%)	541	19.8%	21 (0.8%)
	Bushbaby	767	28.1%		113	4.1%	
	Mouse	710	26.0%		100	3.7%	
Drosha-dependent miRBase miRNAs (1286)	Marmoset	995	77.4%	437 (34.0%)	776	60.3%	293 (22.8%)
	Bushbaby	578	44.9%		441	34.3%	
	Mouse	609	47.4%		465	36.2%	
Random genomic regions (5000)	Marmoset	2890	57.8%	363 (7.3%)	25	0.5%	0 (0%)
	Bushbaby	1292	25.8%		10	0.2%	
	Mouse	737	14.7%		3	0.1%	

sRNA-seq datasets from the ENCODE consortium [7] represented by 181 libraries from 43 cell types with a total of 171 994 643 unique reads. After removing reads overlapping the RNA repeats, we combined the filtered reads with the unique reads from sRNA-seq data obtained in this study to generate a total of 201 295 825 unique sRNA reads. A psRNA was considered detected if its genomic coordinates matched those of a unique sRNA read within a 2-base shift at either the 5′ or 3′ end (total shift ≤ 4 bases) as shown in Fig. 4A.

As the third step of the validation, we generated a true negative control to estimate the performance of the psRNA prediction as a proper metric in selecting true Drosha-dependent miRNAs. We randomly selected 50 000 novel sRNAs from the 20 014 675 novel sRNAs with unique 5′ and 3′ ends that were used as the input into the pipeline (Fig. 3). After removing sRNAs that corresponded to the novel members of the mir-548 family and miRBase miRNAs, this number was reduced to 49 229 of which our algorithm could predict coordinates of potential psRNA for 29 271 (59.5%) sRNAs (Supplementary Table S12). We then applied the MuStARD model to these 29 271 sRNAs and found that respectively 21 579, 5574, and 2118 sRNAs had scores in the ranges of [0, 0.05], (0.05, 0.5], and (0.5, 1] (Supplementary Table S12). The 21 579 random sRNAs with the very low MuStARD scores of [0, 0.05] would thus represent the true negative controls that are expected to have the lowest fraction of real Drosha-dependent miRNAs.

As expected, the "Drosha-dependent" miRBase miRNAs had the highest associations with psRNAs. Of the 646 such miRNAs detected by our pipeline (Fig. 3), we could predict potential psRNAs for 96.6% (624/646, Fig. 4B and Supplementary Table S12) and detect the presence of 91.5% (571/624) of these psRNAs in the RNA-seq data as shown in Fig. 4B and exemplified in the Fig. 5A–C for several miRNAs. On the other hand, also as expected, the negative control sRNAs had the smallest fraction of potential psRNAs that were detected in the sRNA-seq data—699 out of 21 579 or 3.2% (Fig. 4B and Supplementary Table S12). Strikingly, as shown in Fig. 4B and Supplementary Table S12, we could detect potential psRNAs for 30.4% (662/2177) of the "novel miRNAs-NAS," which was 9.5-fold higher than the aforementioned fraction for the random sRNAs (3.2%). Several such examples are shown in Fig. 6A–D. In addition, the detected psRNA fraction for the "novel miRNAs-AS" category was even higher—42/63 (66.7%, Fig. 4B and Supplementary Table S12), arguing that these miRNAs are also generated by Dicer. Interestingly, the corresponding fraction for the "novel noncanonical/indirect Drosha substrates" was 739 out of 17 501 (4.2%), somewhat higher than the random sRNAs (Fig. 4B and Supplementary Table S12), suggesting that some members of this category were true miRNAs. Furthermore, as expected, the detected psRNA fraction for the "novel weak Drosha/DGCR8 substrates" was even higher, 396/3707 (10.7%, Fig. 4B and Supplementary Table S12), indicating that some of the miRNA candidates in this category were also true miRNAs.

Overall, taken together, the Drosha-dependency, the high MuStARD score and the strong enrichment in association with psRNAs, strongly argue that the 2726 sRNAs in both the "novel miRNAs-AS" and "novel miRNAs-NAS" categories represent *bona fide* miRNAs. The failure to detect psRNAs for 1536/2240 (68.6%) of these sRNAs is likely due to the sensitivity of detection since, as shown in Fig. 4C and Supplementary Table S14, the "novel miRNAs" tend to have much lower expression levels than the annotated "Drosha-dependent" miRNAs.

However, our results suggest that additional miRNAs do exist. Based on the fractions of the sRNAs associated with psRNAs, it appears that besides the 2726 novel miRNAs, additional miRNAs exist in the "novel weak Drosha/DGCR8 substrates" and even "novel non-canonical/indirect Drosha substrates" categories. Furthermore, strikingly, we observed increase in the fractions of the psRNA-associated sRNAs among the randomly selected sRNAs with the increase in the MuStARD scores. Specifically, random sRNAs with the scores in the ranges of (0.05, 0.5] and (0.5, 1] contained respectively 5.7% and 9.0% psRNAs, which were higher than the 3.2% which was found for the background random sRNAs (Fig. 4B and Supplementary Table S12). Interestingly, the group of random sRNAs with the MuStARD scores in the range of (0.5, 1] and with psRNA was represented by 190 entities, which corresponded to 0.38% of the original 50 000 randomly chosen sRNAs (Supplementary Table S12). Considering that the random sRNAs were chosen from 20 014 675 sRNAs, this would mean that the total number of Drosha-dependent miRNAs in just one human cell line could be as high as ∼80K (see "Discussion" section).

Canonical miRNAs function by interacting with the AGO proteins to form the RISC RNA–protein complex [60]. Therefore, we further explored whether the novel miRNAs identified in this work also interact with AGO proteins. We took advantage of two publicly available datasets that identified sRNAs associated with either AGO2/3 from the study by Rybak-Wolf *et al.* [27] or AGO1/2 proteins from the study by

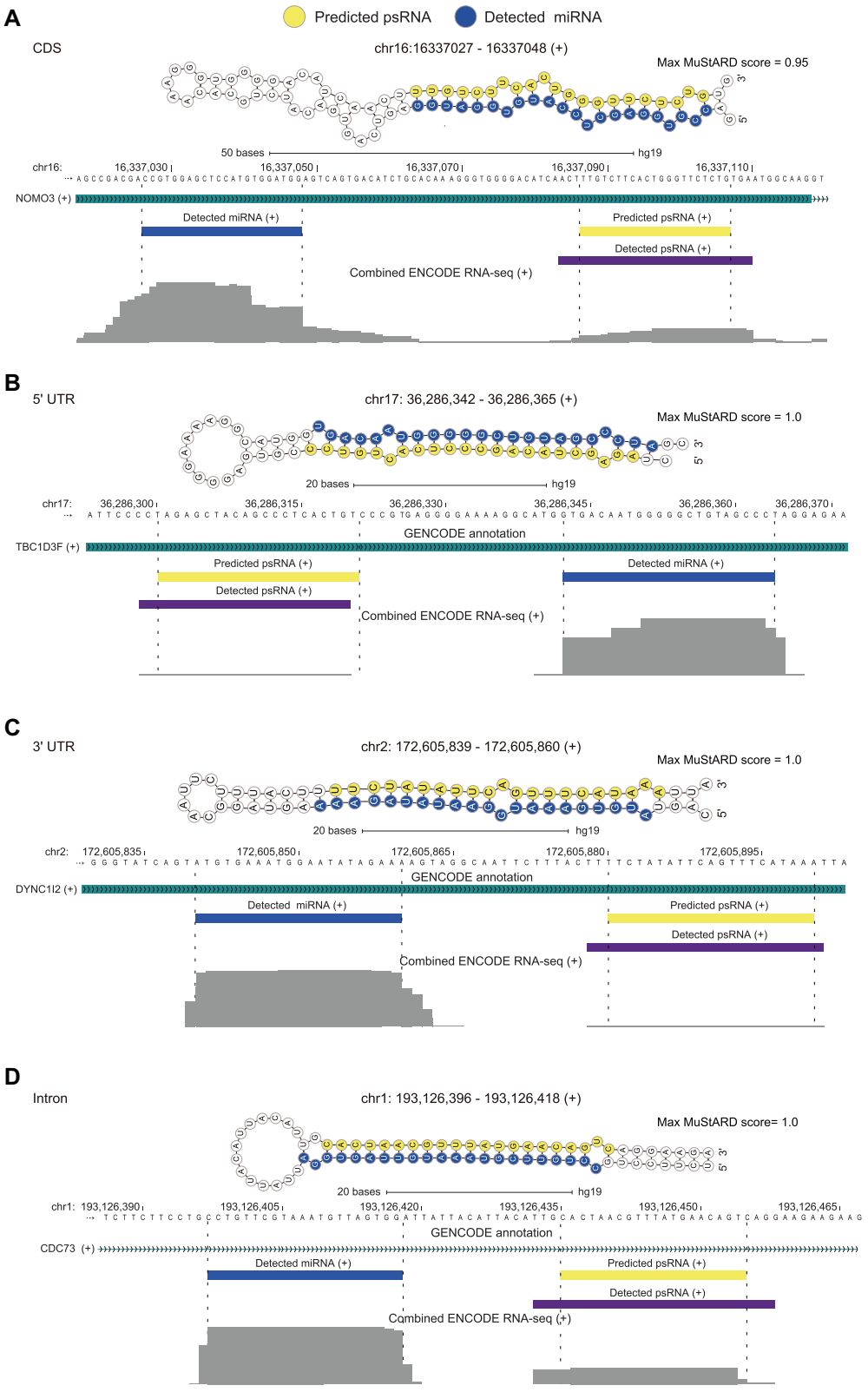


Figure 6. Example of detection of novel miRNAs using the pipeline developed in this study. Genomic contexts and predicted targeting and psRNAs are shown for novel miRNAs detected in CDS of gene NOMO3 (**A**), 5′ UTR of gene TBC1D3F (**B**), 3′ UTR of gene DYNC1I2 (**C**), and intron of gene CDC73 (**D**). The secondary RNA structures are based on the predictions generated by the MuStARD program.

Valen *et al.* [13]. We then calculated the fractions of the overlap between the various aforementioned categories of sRNAs (those identified using our pipeline, random sRNAs, and the control miRBase miRNAs) and the AGO-interacting sRNAs ("Materials and methods" section). As shown in the Fig. 4D, we have observed an increase in the fraction of overlap with the increasing MuStARD scores (Supplementary Table S15). While certain amount of nonspecific signal could be expected in such experiments, it would not be expected to correlate with the propensity of the RNA structures to fold into RNA structures recognized by the Drosha/DGCR8 complex which is measured by the MuStARD score. Instead, this trend argued for a high specificity of the observed interactions between the sRNAs and the AGO proteins.

The additional proof for specificity of the interactions came from the observation that the fractions of the overlaps were always higher for the novel Drosha dependent sRNAs compared to the random sRNAs reads with the same MuStARD score (Fig. 4D and Supplementary Table S15). Finally, the sRNAs with psRNAs had higher fractions of the overlap than those without psRNAs (Fig. 4D and Supplementary Table S15). In fact, based on the Rybak-Wolf *et al.* dataset, the fraction of the "novel miRNAs-NAS" sRNAs with psRNAs that associated with AGO2/3 proteins was almost as high as that for the miRBase miRNAs—33.2% versus 35.7% (Fig. 4D and Supplementary Table S15). All of the above general trends were found in both datasets (Fig. 4D and Supplementary Table S15). Overall, we found that 613/2661 and 34/65 sRNAs from the respectively "novel miRNAs-NAS" and "novel miRNAs-AS" datasets were found to interact with the AGO proteins in at least one dataset (Supplementary Table S15). The failure to detect all sRNAs in these categories could be explained by the cell-type specificity of miRNA expression since the two datasets were obtained in cell lines other than K562. This is a likely reason why only 1682/2880 miRBase miRNAs were found to be associated with AGO proteins in these datasets (Supplementary Table S15).

We then explored whether novel miRNAs could bind and potentially regulate other transcripts similar to the canonical miRNAs. Transcriptome-wide miRNA-binding sites could be identified by a combination of AGO immunoprecipitation of RISC complexes followed by ligation of the miRNAs that are base-paired to their targets inside the complexes using crosslinking, ligation, and sequencing of hybrids on argonaute (AGO-CLASH) approach [61, 62]. To identify binding sites of our novel miRNAs, we took advantage of a publicly available dataset from a human HCT116 cell line based on a modified AGO-qCLASH method [28, 29]. Indeed, we could detect 123/2661 and 2/65 sRNAs from the respectively "novel miRNAs-NAS" and "novel miRNAs-AS" interacting with 3′ untranslated regions (UTRs) of respectively 390 and 24 genes. The failure to detect all novel miRNAs likely stems from the cell-type specific expression of these molecules since only 570/2880 miRBase miRNAs could be detected in this dataset. However, these results suggest that in principle, novel miRNAs appear to participate in the same interactions as the canonical miRNAs.

The study Rybak-Wolf *et al.* also provided genome-wide information of Dicer binding using a PAR-CLIP assay [27], which allowed for an independent confirmation of Dicer-mediated processing of novel miRNAs found in this work. The fractions of overlap were appreciably lower for each category, even for the known miRBase miRNAs — >35% for AGO versus 16.3% for Dicer (Fig. 4E and Supplementary Table S15)—which is likely due to the transitory interactions between Dicer and its substrates. Still, we observed very similar general trends as in the AGO datasets (Fig. 4E and Supplementary Table S15), arguing that the parameters used in this work such as the Drosha-dependency, MuStARD score, and the association with psRNA tend to select for true miRNAs that are processed by Dicer and interact with AGO proteins.

Support by multiple sequencing reads and evidence of consistent processing at the 5′ ends constitute two important criteria that are currently used for defining known high confidence miRNAs [38, 58]. To define a high confidence miRNA, miRBase requires ≥ 20 reads to overlap each the "guide" and "passenger" miRNA with ≥50% of these reads having the same 5′ ends [38]. Using these criteria, 201/2661 (7.6%) "novel miRNAs-NAS" and 19/65 (29.2%) "novel miRNAs-AS" could be classified as high confidence (Supplementary Table S16). In this analysis, the same 5′ end was defined within a 2-base shift to allow for imprecision of mapping and sequencing. Overall, 220/2726 (8.1%) of the novel miRNAs reported here could be classified as high confidence using these criteria (Supplementary Table S16). However, for comparison, only ∼500 (26%) out of 1914 human miRBase miRNA loci could be classified as high confidence [38]. The miRBase developers note that not being classified as "high confidence" does not mean that a miRNA is not real since the main reason for this is the lack of expression data [38]. In fact, 1225/1914 (64%) human miRBase microRNA loci do not have sufficient expression data to reach the ≥20 reads requirement [38]. Likewise, only 17.6% (481/2726) novel miRNAs reached the ≥20 reads requirement, however based on a much smaller sRNA-seq dataset than miRBase: ∼200 million used in this study versus 5.5 billion reads used by miRBase [38]. Therefore, it is highly likely that the number of high confidence novel miRNAs will increase with including more sRNA-seq data in the analysis. Furthermore, it is important to note that many of our novel miRNAs are derived from exons of mRNAs and therefore, they would also overlap by abundant small degradation products of these exons. This in turn would reduce the fraction of reads having the same 5′ ends compared with the miRBase miRNAs most of which do not map to exons. Indeed, we found this to be the case since the fraction of novel miRNAs with ≥50% reads having the same 5′ end increased from 45.7% (220/481) for all novel miRNAs to 56.2% (95/169) for those mapping outside of exons (Supplementary Table S16).

Novel miRNAs are enriched in exons of mRNAs

Analysis of the genome-wide distribution of the 2726 "novel miRNAs" revealed a strong enrichment in exons of mRNAs with the corresponding odds ratio of enrichment being 14.6 (Fig. 7A and B; Supplementary Table S17). For comparison, the corresponding odds ratio for the "Drosha-dependent" miRBase miRNAs expressed in the same cell line was 2.7 (Fig. 7A and B; Supplementary Table S17). Interestingly, as exemplified in Fig. 6A–D, members of the "novel miRNAs" category were enriched in both coding and noncoding portions of mRNAs (Fig. 7B and Supplementary Table S17). Strikingly, the strongest odds ratios for the of enrichment "novel miRNAs" were found in the coding regions (CDSs) followed by the 5′ UTRs and then 3′ UTRs (Fig. 7B and Supplementary Table S17). In contrast, the "Drosha-dependent" miRBase miRNAs

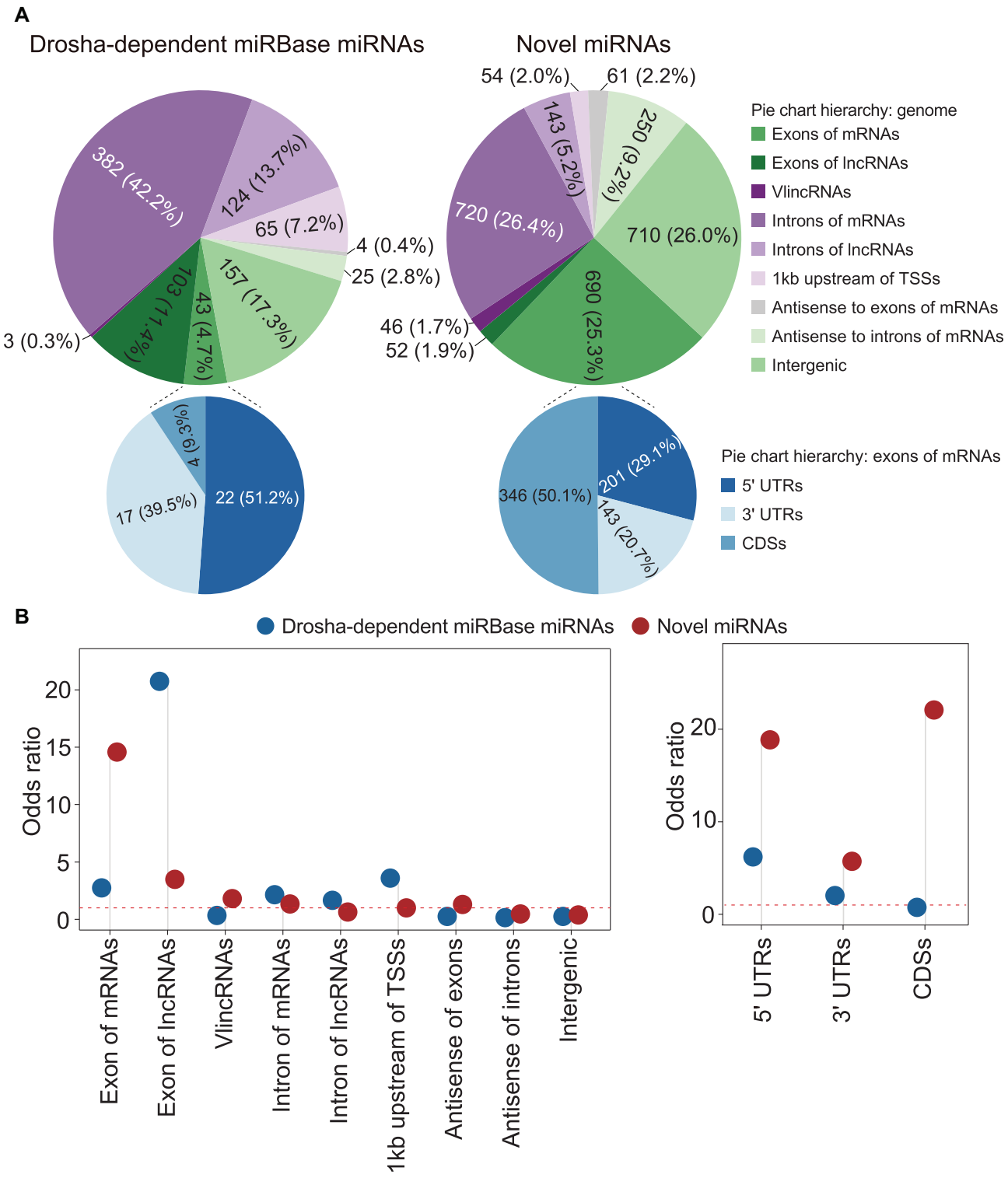


Figure 7. Distribution of Drosha-dependent miRBase miRNAs expressed in K562 and novel miRNAs among the different genomic features. (**A**) Total numbers and fractions of the miRBase miRNAs expressed in K562 (left) and novel miRNAs (right) mapping to the various genomic features are shown. Each miRNA was counted just once according to the hierarchy shown on the right. (**B**) Odds ratios of enrichment of the miRBase and novel miRNAs in the indicated genomic features.

were not enriched in the CDSs (Fig. 7B and Supplementary Table S17). In fact, 50.1% (346) of the "novel miRNAs" located in exons were found in the CDSs of 325 genes, while the remaining 29.1% (201) and 20.7% (143) "novel miRNAs" were found in respectively 5′ and 3′ UTRs of 202 and 139 genes (Fig. 7A and Supplementary Table S17). Overall, we found 690 "novel miRNAs" mapping to exons of 651 genes.

The "novel miRNAs" were also enriched in the exons of lncRNAs; however, much less so than the "Drosha-dependent" miRBase miRNAs with the corresponding odds ratios of 3.5 and 20.7 (Fig. 7B and Supplementary Table S17). Of all other genomic features tested, the "novel miRNAs" have shown moderate enrichment in the vlinc (very long intergenic noncoding) subclass of lncRNAs [26, 63] while the "Drosha-dependent" miRBase miRNAs were depleted in this class of transcripts as evidenced by the corresponding the odds ratios of 1.8 and 0.4 (Fig. 7B and Supplementary Table S17).

Discussion

In this work, we established a pipeline for discovery of novel miRNAs. Using this pipeline, we identified 2726 novel miRNAs representing 2667 novel miRNA loci (excluding the miRNAs that are likely derived the same pre-miRNA) in a single human cell line. It is important to emphasize that we focused on discovering only the canonical, Drosha-dependent miRNAs. For comparison, miRBase contains 1914 miRNA loci (represented by 2880 mature miRNAs), including both the Drosha-dependent and Drosha-independent, atypical miRNAs. Furthermore, we estimate that tens of thousands of other miRNA loci exist, suggesting that the true complexity of the sRNA transcriptome is currently significantly under-estimated even for the known classes of sRNAs.

The criteria for defining true miRNAs have evolved over time to reflect the progress of the data generation technologies in order to separate this class of sRNAs from other transcripts which might share some similar features [38, 58, 64, 65]. The novel miRNAs found in this work satisfy five such conditions as follows. First, the size range: the lengths of the novel miRNAs fall within the appropriate range for the mature miRNAs. Second, the RNA fold: the sequences of novel miRNAs are part of the appropriate RNA structures that have the same folding properties as those for the known canonical miRNAs as determined by the deep learning MuStARD algorithm.

Third, biogenesis: canonical miRNAs are produced by the sequential action of Drosha and Dicer riboendonucleases [36]. The sRNA candidates were initially selected by the virtue of being dependent on Drosha. Then, the validity of the novel miRNAs identified using the pipeline was confirmed by identification of the *in vivo* signature indicative of the final step of the miRNA biogenesis common to all known miRNAs—Dicer cleavage of the stem-loop pre-miRNA precursors [36]. During this step, the Dicer riboendonuclease-mediated cleavage in the stem region of a pre-miRNA hairpin structure generates a pair of sRNAs—the guide miRNA and the psRNA [36]—the latter of which is typically much less stable than the former but can be still detectable and functional *in vivo* [66, 67]. Indeed, we have found that the fraction of the psRNA-associated novel miRNAs obtained by the pipeline was significantly higher than found in randomly sampled sRNAs that were used as input in the pipeline: 31.4% versus 3.2%. Since the search for the associated psRNAs was conducted using exactly the same conditions and datasets for both the randomly sampled sRNAs and pipeline-derived novel miRNA candidates, the higher fraction of detected potential psRNAs for the novel miRNA candidates is highly suggestive of the Dicer-mediated cleavage. And, since the psRNAs were found in the cellular sRNA population, the psRNA association is indicative of Dicer cleavage which takes place *in vivo*—a strong argument that the 2726 novel miRNAs identified in this work do represent *bona fide* miRNAs. In addition, we could directly detect interaction with Dicer for a subset of the novel miRNAs.

Fourth, interaction with the AGO proteins: all true miRNAs function via interactions with these proteins. In this work, we have provided strong evidence that our novel miRNAs indeed interact with AGO proteins. Furthermore, for a subset of the novel miRNAs, we have provided evidence for direct binding to target mRNAs inside the RISC complex. Fifth, a subset of novel miRNAs satisfied criteria associated with high confidence miRNAs. Finally, an appreciable fraction (∼1/3) of the novel miRNAs appear to represent distal members of known miRNA families. Taken together, these properties argue against the possibility that the novel miRNAs reported here represent some novel class of sRNAs that are not miRNAs. However, given the complexity of the human sRNA transcriptome, this formal possibility exists and additional studies are required to unambiguously confirm the true miRNA natures of these entities.

However, despite sharing multiple features with the known miRNAs, the novel miRNAs identified in this work have three characteristic features that differentiate them from the annotated miRBase miRNAs. First, novel miRNAs have relatively low expression levels. However, this does not necessarily mean that these transcripts represent non-functional noise. As we have recently shown using high-throughput phenotypic assays based on over-expression of sRNA in cultured human cells, functional sRNAs had a statistically significant tendency to have lower expression than sRNAs with no function [22]. A possible explanation for biological relevance of these low-expressed sRNAs could be a restricted pattern of expression and functionality limited to a specific subpopulation of cells [22]. In fact, cell-to-cell heterogeneity in expression of miRNAs have been detected in previous studies utilizing high-throughput single-cell sequencing methods [68–71]. Second, novel miRNAs are less conserved than the annotated miRNAs. This however could indicate involvement of these miRNAs in a human-specific function, as recently shown for a human-specific miRNAs miR-1229–3p [72]. Still, 100 novel human miRNAs appear to be conserved in as distantly related species as mouse and hundreds more are conserved in the genomes of primates.

Third, novel miRNAs are significantly enriched in exons of mRNAs, including the protein-coding regions. For example, previously, among all 1289 Drosha-dependent miRBase miRNAs, only 52 were found in exons, compared to 690 found in this work in just one human cell line. These observations are consistent with the complex, interleaved organization of the information encoded in the human genome in which a single base pair could be part of multiple functional elements [25], for example, protein-coding and regulatory RNA species as shown in this work. Furthermore, as we have shown recently, multiple other novel classes of functional sRNAs likely exist [22], which significantly expands the complexity of the information encoded by the genome.

Our work suggested that many other miRNAs exist that could be detected using the pipeline developed in this study which consists of two main parts. The first part identifies sRNAs that are depleted *in vivo* in response to depletion of the key miRNA biogenesis enzyme, Drosha riboendonuclease. The second part further refines such sRNAs to identify those that have structure-sequence features of true Drosha-dependent miRNAs using a deep learning-based approach. We found that both parts of the pipeline are necessary. First, not all sRNAs that are depleted in response to Drosha represent true Drosha-dependent miRNAs as shown by two lines of evidence. First, only a small fraction (13%, 41 806 out of 322 101) of the novel (not overlapping miRBase) Drosha-depleted sRNAs had the miRNA-like size distribution of 20–25 nt. This contrasted with most (78.5%, 860/1095) of the miRBase miRNAs found to be depleted in response to Drosha falling within this size range. Second, only 9.2% of the Drosha-depleted sRNAs had high MuStARD scores of (0.5, 1] compared to the corresponding fraction of 93.7% for the Drosha-dependent miRBase miRNAs. Therefore, only a small fraction of sRNAs, levels of which decreased in response to Drosha depletion, represented true Drosha-dependent miRNAs. These results are not surprising and are likely due to the following two major reasons. One, indirect effects—not all sRNAs that respond to the change in Drosha level are generated by this enzyme. Rather, they could represent downstream effects of the changes in the levels of the true Drosha targets. Also, Drosha has been shown to generate products other than miRNAs [51, 52]. Two, the accuracy of the deep learning-based approach alone was also not as high as that of the whole pipeline. The fraction of the randomly selected sRNAs that had high (0.5, 1] MuStARD scores and for which the predicted psRNAs were detected was 9.0% (190/2118). This value was substantially lower than the corresponding fraction of 31.4% for the pipeline-derived novel miRNAs. Altogether, these results suggest that a combination of the two approaches is required to properly refine novel miRNA candidates.

Still, the deep learning-based approach alone showed promise as evidenced by the analysis of the randomly selected sRNAs—with the increase in the MuStARD score, we observed a significant increase in the fraction of the detected psRNAs. Specifically, 3.2% (699 out of 21 579) of the random sRNAs with the scores in the range of [0, 0.05] had detectable psRNA. However, this fraction increased to 5.7% (318/5574) and 9.0% (190/2118) for the random sRNAs whose MuStARD scores were in the ranges (0.05, 0.5] and (0.5, 1], respectively (Supplementary Table S12). Since all categories of sRNAs were subjected to the psRNA analysis using the same analytical approach and datasets, it is hard to imagine that the increase in the psRNA association is spurious especially considering that it correlates well with the strength of the deep learning-based predictions. Instead, the increase in the psRNA-detected fraction most likely reflects the increase in the true novel miRNAs as evidenced by the corresponding increase in the MuStARD scores. Strikingly, extrapolating from the fraction of the randomly-selected sRNAs with the high MuStARD scores and associated psRNA, we can estimate that as much as ∼80K novel miRNAs still await discovery. The inability to detect them in this study most likely stems from their low expression in K562—in fact, majority (62.5%, 12 508 130 out of 20 014 675) of all sRNAs with unique 5′ and 3′ coordinates that were used for selection of random sRNAs were represented by single NGS reads. Still, since non-coding RNAs tend to be highly cell type specific [3], it is quite possible that these novel miRNAs are highly expressed in other cell types.

One important caveat of this work is that we have not directly shown that the novel miRNAs regulate other transcripts and have biological relevance. However, identification of these transcripts represents an important first step in answering the questions about the function and mechanism of action. Overall, our results show that we are clearly at the very beginning of the full understanding the depth of complexity of the sRNA transcriptome and that additional in-depth studies are required and fully warranted to fully understand its complexity and functionality.

Acknowledgements

Author contributions: P.K. conceived the project and supervised the analytical and wet lab parts of the project with contribution from F.Q. F.G. performed all wet lab experiments and part of the bioinformatics analyses. F.W. performed part of the bioinformatic analyses. B.D., F.Q., F.Y., J.C., and H. Cao assisted with the bioinformatics analyses. P.K. wrote the manuscript with contributions from F.G. Y.C. assisted with the bioinformatics analyses and manuscript preparation. H. Chen assisted with the manuscript preparation.

Supplementary data

Supplementary data is available at NAR online.

Conflict of interest

None declared.

Funding

P.K. is supported by the Research Fund for International Senior Scientists from the National Natural Science Foundation of China (grant number 32150710525) and National Natural Science Foundation of China (grant number 32170619). Y.C. is supported by the National Natural Science Foundation of China (grant number 32201055) and the Scientific Research Funds of Huaqiao University (grant number 21BS127). F.Q. is supported by the National Natural Science Foundation of China (grant number 32000462). H.C. is supported by the National Science Foundation of China (grant number 32000476), Youth Innovation Grant of Xiamen, Fujian Province, China (grant number 3502Z20206015), the Fundamental Research Funds for the Central Universities of Huaqiao University (grant number ZQN-922), and the Scientific Research Funds of Huaqiao University (grant number 600005-Z17Y0043). Funding to pay the Open Access publication charges for this article was provided by the Xiamen University.

Data availability

The NGS data and the coordinates of 41806 novel Drosha substrates generated in this study have been deposited in the GEO database under accession code GSE259288. The processed data generated in this study are provided in the Supplementary Data files and referred to in the main text, figure legends, and Materials and methods.

Code availability

Custom scripts constituting the pipeline used to generate genomic coordinates of Drosha substrates are available from GitHub (https://github.com/Gaofan315/small-RNA/releases/tag/smallRNA) and Zenodo (DOI: 10.5281/zenodo.10503904). The additional adjustments and downstream analyses involve standard packages in the R environment and BEDTools suite as described in "Materials and methods" section.

References

1. St Laurent G, Wahlestedt C, Kapranov P. The landscape of long noncoding RNA classification. *Trends Genet* 2015;**31**:239–51. https://doi.org/10.1016/j.tig.2015.03.007
2. Yang L, Froberg JE, Lee JT. Long noncoding RNAs: fresh perspectives into the RNA world. *Trends Biochem Sci* 2014;**39**:35–43. https://doi.org/10.1016/j.tibs.2013.10.002
3. Mattick JS, Amaral PP, Carninci P *et al.* Long non-coding RNAs: definitions, functions, challenges and recommendations. *Nat Rev Mol Cell Biol* 2023;**24**:430–47. https://doi.org/10.1038/s41580-022-00566-8
4. Kapranov P, Cheng J, Dike S *et al.* RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* 2007;**316**:1484–88. https://doi.org/10.1126/science.1138341
5. Seila AC, Calabrese JM, Levine SS *et al.* Divergent transcription from active promoters. *Science* 2008;**322**:1849–51. https://doi.org/10.1126/science.1162253
6. Affymetrix E, Transcriptome P, Cold S *et al.* Post-transcriptional processing generates a diversity of 5′-modified long and short RNAs. *Nature* 2009;**457**:1028–32. https://doi.org/10.1038/nature07759
7. Djebali S, Davis CA, Merkel A *et al.* Landscape of transcription in human cells. *Nature* 2012;**489**:101–8. https://doi.org/10.1038/nature11233
8. Taft RJ, Simons C, Nahkuri S *et al.* Nuclear-localized tiny RNAs are associated with transcription initiation and splice sites in metazoans. *Nat Struct Mol Biol* 2010;**17**:1030–4. https://doi.org/10.1038/nsmb.1841
9. Taft RJ, Glazov EA, Cloonan N *et al.* Tiny RNAs associated with transcription start sites in animals. *Nat Genet* 2009;**41**:572–8. https://doi.org/10.1038/ng.312
10. Kapranov P, Ozsolak F, Kim SW *et al.* New class of gene-termini-associated human RNAs suggests a novel RNA copying mechanism. *Nature* 2010;**466**:642–6. https://doi.org/10.1038/nature09190
11. Kawaji H, Nakamura M, Takahashi Y *et al.* Hidden layers of human small RNAs. *BMC Genomics* 2008;**9**:157. https://doi.org/10.1186/1471-2164-9-157
12. de Hoon M, Bonetti A, Plessy C *et al.* Deep sequencing of short capped RNAs reveals novel families of noncoding RNAs. *Genome Res* 2022;**32**:1727–35. https://doi.org/10.1101/gr.276647.122
13. Valen E, Preker P, Andersen PR *et al.* Biogenic mechanisms and utilization of small RNAs derived from human protein-coding genes. *Nat Struct Mol Biol* 2011;**18**:1075–82. https://doi.org/10.1038/nsmb.2091
14. Mattick JS, Makunin IV. Small regulatory RNAs in mammals. *Hum Mol Genet* 2005;**14**:R121–32. https://doi.org/10.1093/hmg/ddi101
15. Aalto AP, Pasquinelli AE. Small non-coding RNAs mount a silent revolution in gene expression. *Curr Opin Cell Biol* 2012;**24**:R121–32333–40. https://doi.org/10.1016/j.ceb.2012.03.006
16. Mattick JS. RNA out of the mist. *Trends Genet* 2023;**39**:187–207. https://doi.org/10.1016/j.tig.2022.11.001
17. Cao H, Wahlestedt C, Kapranov P. Strategies to annotate and characterize long noncoding RNAs: advantages and pitfalls. *Trends Genet* 2018;**34**:704–21. https://doi.org/10.1016/j.tig.2018.06.002
18. Mattick JS, Rinn JL. Discovery and annotation of long noncoding RNAs. *Nat Struct Mol Biol* 2015;**22**:5–7. https://doi.org/10.1038/nsmb.2942
19. Mercer TR, Mattick JS. Structure and function of long noncoding RNAs in epigenetic regulation. *Nat Struct Mol Biol* 2013;**20**:300–7. https://doi.org/10.1038/nsmb.2480
20. Kung JT, Colognori D, Lee JT. Long noncoding RNAs: past, present, and future. *Genetics* 2013;**193**:651–69. https://doi.org/10.1534/genetics.112.146704
21. Tuck AC, Tollervey D. RNA in pieces. *Trends Genet* 2011;**27**:422–32. https://doi.org/10.1016/j.tig.2011.06.001
22. Gao F, Wang F, Cao H *et al.* Evidence for existence of multiple functional human small RNAs derived from transcripts of protein-coding genes. *Int J Mol Sci* 2023;**24**:4163. https://doi.org/10.3390/ijms24044163
23. Taft RJ, Kaplan CD, Simons C *et al.* Evolution, biogenesis and function of promoter-associated RNAs. *Cell Cycle* 2009;**8**:2332–8. https://doi.org/10.4161/cc.8.15.9154
24. Yu D, Ma X, Zuo Z *et al.* Classification of transcription boundary-associated RNAs (TBARs) in animals and plants. *Front Genet* 2018;**9**:168. https://doi.org/10.3389/fgene.2018.00168
25. Kapranov P, Willingham AT, Gingeras TR. Genome-wide transcription and the implications for genomic organization. *Nat Rev Genet* 2007;**8**:413–23. https://doi.org/10.1038/nrg2083
26. St Laurent G, Shtokalo D, Dong B *et al.* VlincRNAs controlled by retroviral elements are a hallmark of pluripotency and cancer. *Genome Biol* 2013;**14**:R73. https://doi.org/10.1186/gb-2013-14-7-r73
27. Rybak-Wolf A, Jens M, Murakawa Y *et al.* A variety of dicer substrates in human and *C. elegans*. *Cell* 2014;**159**:1153–67. https://doi.org/10.1016/j.cell.2014.10.040
28. Stribling D, Lei Y, Guardia CM *et al.* A noncanonical microRNA derived from the snaR-A noncoding RNA targets a metastasis inhibitor. *RNA* 2021;**27**:694–709. https://doi.org/10.1261/rna.078694.121
29. Fields CJ, Li L, Hiers NM *et al.* Sequencing of argonaute-bound microRNA/mRNA hybrids reveals regulation of the unfolded protein response by microRNA-320a. *PLoS Genet* 2021;**17**:e1009934. https://doi.org/10.1371/journal.pgen.1009934
30. Kalvari I, Nawrocki EP, Ontiveros-Palacios N *et al.* Rfam 14: expanded coverage of metagenomic, viral and microRNA families. *Nucleic Acids Res* 2021;**49**:D192–200. https://doi.org/10.1093/nar/gkaa1047
31. Nawrocki EP, Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* 2013;**29**:2933–5. https://doi.org/10.1093/bioinformatics/btt509
32. Gregory RI, Yan KP, Amuthan G *et al.* The Microprocessor complex mediates the genesis of microRNAs. *Nature* 2004;**432**:235–40. https://doi.org/10.1038/nature03120
33. Han J, Lee Y, Yeom KH *et al.* The Drosha-DGCR8 complex in primary microRNA processing. *Genes Dev* 2004;**18**:3016–27. https://doi.org/10.1101/gad.1262504
34. Landthaler M, Yalcin A, Tuschl T. The human DiGeorge syndrome critical region gene 8 and Its D. melanogaster homolog are required for miRNA biogenesis. *Curr Biol* 2004;**14**:2162–7. https://doi.org/10.1016/j.cub.2004.11.001
35. Denli AM, Tops BB, Plasterk RH *et al.* Processing of primary microRNAs by the Microprocessor complex. *Nature* 2004;**432**:231–5. https://doi.org/10.1038/nature03049
36. Treiber T, Treiber N, Meister G. Regulation of microRNA biogenesis and its crosstalk with other cellular pathways. *Nat Rev Mol Cell Biol* 2019;**20**:5–20. https://doi.org/10.1038/s41580-018-0059-1
37. Bail S, Swerdel M, Liu H *et al.* Differential regulation of microRNA stability. *RNA* 2010;**16**:1032–9. https://doi.org/10.1261/rna.1851510

38. Kozomara A, Birgaoanu M, Griffiths-Jones S. miRBase: from microRNA sequences to function. *Nucleic Acids Res* 2019;**47**:D155–62. https://doi.org/10.1093/nar/gky1141
39. Kim K, Baek SC, Lee YY *et al.* A quantitative map of human primary microRNA processing sites. *Mol Cell* 2021;**81**:3422–39. https://doi.org/10.1016/j.molcel.2021.07.002
40. Ladewig E, Okamura K, Flynt AS *et al.* Discovery of hundreds of mirtrons in mouse and human small RNA data. *Genome Res* 2012;**22**:1634–45. https://doi.org/10.1101/gr.133553.111
41. Okamura K, Hagen JW, Duan H *et al.* The mirtron pathway generates microRNA-class regulatory RNAs in Drosophila. *Cell* 2007;**130**:89–100. https://doi.org/10.1016/j.cell.2007.06.028
42. Ruby JG, Jan CH, Bartel DP. Intronic microRNA precursors that bypass Drosha processing. *Nature* 2007;**448**:83–6. https://doi.org/10.1038/nature05983
43. Yang JS, Lai EC. Alternative miRNA biogenesis pathways and the interpretation of core miRNA pathway mutants. *Mol Cell* 2011;**43**:892–903. https://doi.org/10.1016/j.molcel.2011.07.024
44. Zeng Y, Yi R, Cullen BR. Recognition and cleavage of primary microRNA precursors by the nuclear processing enzyme Drosha. *EMBO J* 2005;**24**:138–48. https://doi.org/10.1038/sj.emboj.7600491
45. Auyeung VC, Ulitsky I, McGeary SE *et al.* Beyond secondary structure: primary-sequence determinants license pri-miRNA hairpins for processing. *Cell* 2013;**152**:844–58. https://doi.org/10.1016/j.cell.2013.01.031
46. Zeng Y, Cullen BR. Efficient processing of primary microRNA hairpins by Drosha requires flanking nonstructured RNA sequences. *J Biol Chem* 2005;**280**:27595–603. https://doi.org/10.1074/jbc.M504714200
47. Zhang X, Zeng Y. The terminal loop region controls microRNA processing by Drosha and Dicer. *Nucleic Acids Res* 2010;**38**:7689–97. https://doi.org/10.1093/nar/gkq645
48. Ma H, Wu Y, Choi JG *et al.* Lower and upper stem-single-stranded RNA junctions together determine the Drosha cleavage site. *Proc Natl Acad Sci USA* 2013;**110**:20687–92. https://doi.org/10.1073/pnas.1311639110
49. Georgakilas GK, Grioni A, Liakos KG *et al.* Multi-branch Convolutional Neural Network for Identification of Small Non-coding RNA genomic loci. *Sci Rep* 2020;**10**:9486. https://doi.org/10.1038/s41598-020-66454-3
50. Jurka J. Repbase update: a database and an electronic journal of repetitive elements. *Trends Genet* 2000;**16**:418–20. https://doi.org/10.1016/S0168-9525(00)02093-X
51. Lee D, Shin C. Emerging roles of DROSHA beyond primary microRNA processing. *RNA Biol* 2018;**15**:186–93. https://doi.org/10.1080/15476286.2017.1405210
52. Pong SK, Gullerova M. Noncanonical functions of microRNA pathway enzymes—Drosha, DGCR8, Dicer and Ago proteins. *FEBS Lett* 2018;**592**:2973–86. https://doi.org/10.1002/1873-3468.13196
53. Rognes T, Flouri T, Nichols B *et al.* VSEARCH: a versatile open source tool for metagenomics. *PeerJ* 2016;**4**:e2584. https://doi.org/10.7717/peerj.2584
54. Song Y, Li L, Yang W *et al.* Sense-antisense miRNA pairs constitute an elaborate reciprocal regulatory circuit. *Genome Res* 2020;**30**:661–72. https://doi.org/10.1101/gr.257121.119
55. Fromm B, Hoye E, Domanska D *et al.* MirGeneDB 2.1: toward a complete sampling of all major animal phyla. *Nucleic Acids Res* 2022;**50**:D204–10. https://doi.org/10.1093/nar/gkab1101
56. Foley NM, Mason VC, Harris AJ *et al.* A genomic timescale for placental mammal evolution. *Science* 2023;**380**:eabl8189. https://doi.org/10.1126/science.abl8189
57. Chatterjee HJ, Ho SY, Barnes I *et al.* Estimating the phylogeny and divergence times of primates using a supermatrix approach. *BMC Evol Biol* 2009;**9**:259. https://doi.org/10.1186/1471-2148-9-259
58. Desvignes T, Batzel P, Berezikov E *et al.* miRNA nomenclature: a view incorporating genetic origins, biosynthetic pathways, and sequence variants. *Trends Genet* 2015;**31**:613–26. https://doi.org/10.1016/j.tig.2015.09.002
59. Meijer HA, Smith EM, Bushell M. Regulation of miRNA strand selection: follow the leader? *Biochem Soc Trans* 2014;**42**:1135–40. https://doi.org/10.1042/BST20140142
60. Nakanishi K. Anatomy of four human Argonaute proteins. *Nucleic Acids Res* 2022;**50**:6618–38. https://doi.org/10.1093/nar/gkac519
61. Moore MJ, Scheel TK, Luna JM *et al.* miRNA-target chimeras reveal miRNA 3′-end pairing as a major determinant of Argonaute target specificity. *Nat Commun* 2015;**6**:8864. https://doi.org/10.1038/ncomms9864
62. Helwak A, Kudla G, Dudnakova T *et al.* Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. *Cell* 2013;**153**:654–65. https://doi.org/10.1016/j.cell.2013.03.043
63. Kapranov P, St Laurent G, Raz T *et al.* The majority of total nuclear-encoded non-ribosomal RNA in a human cell is 'dark matter' un-annotated RNA. *BMC Biol* 2010;**8**:149. https://doi.org/10.1186/1741-7007-8-149
64. Ambros V, Bartel B, Bartel DP *et al.* A uniform system for microRNA annotation. *RNA* 2003;**9**:277–9. https://doi.org/10.1261/rna.2183803
65. Fromm B, Zhong X, Tarbier M *et al.* The limits of human microRNA annotation have been met. *RNA* 2022;**28**:781–5. https://doi.org/10.1261/rna.079098.122
66. Ghildiyal M, Xu J, Seitz H *et al.* Sorting of Drosophila small silencing RNAs partitions microRNA* strands into the RNA interference pathway. *RNA* 2010;**16**:43–56. https://doi.org/10.1261/rna.1972910
67. Okamura K, Liu N, Lai EC. Distinct mechanisms for microRNA strand selection by Drosophila argonautes. *Mol Cell* 2009;**36**:431–44. https://doi.org/10.1016/j.molcel.2009.09.027
68. Faridani OR, Abdullayev I, Hagemann-Jensen M *et al.* Single-cell sequencing of the small-RNA transcriptome. *Nat Biotechnol* 2016;**34**:1264–6. https://doi.org/10.1038/nbt.3701
69. Xiao Z, Cheng G, Jiao Y *et al.* Holo-Seq: single-cell sequencing of holo-transcriptome. *Genome Biol* 2018;**19**:163. https://doi.org/10.1186/s13059-018-1553-7
70. Wang N, Zheng J, Chen Z *et al.* Single-cell microRNA-mRNA co-sequencing reveals non-genetic heterogeneity and mechanisms of microRNA regulation. *Nat Commun* 2019;**10**:95. https://doi.org/10.1038/s41467-018-07981-6
71. Isakova A, Neff N, Quake SR. Single-cell quantification of a broad RNA spectrum reveals unique noncoding patterns associated with cell types and states. *Proc Natl Acad Sci USA* 2021;**118**:e2113568118. https://doi.org/10.1073/pnas.2113568118
72. Soutschek M, Bianco AL, Galkin S *et al.* A human-specific microRNA controls the timing of excitatory synaptogenesis. bioRxiv, https://doi.org/10.1101/2023.10.04.560889, 5 October 2023, preprint: not peer reviewed.

Received: April 16, 2024. **Revised:** January 22, 2025. **Editorial Decision:** January 25, 2025. **Accepted:** January 28, 2025