Dicer-Processed Small RNAs: Rules and Exceptions

David Langenberger[a,b], M. Volkan Çakir[b], Steve Hoffmann[a,b], Peter F. Stadler[a,b,c,d,e,f,g,h,i]

[a]*LIFE, Leipzig Research Center for Civilization Diseases, University Leipzig, Philipp-Rosenthal-Strasse 27, D-04107 Leipzig, Germany*
[b]*Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelstraße 16-18, D-04107, Leipzig, Germany*
[c]*Bioinformatics Group, Department of Computer Science, Härtelstraße 16-18, D-04107, Leipzig, Germany*
studla@bioinf.uni-leipzig.de
[d]*Max Planck Institute for Mathematics in the Sciences, Inselstraße 22, D-04103 Leipzig, Germany*
[e]*RNomics Group, Fraunhofer Institut für Zelltherapie und Immunologie, Deutscher Platz 5e, D-04103 Leipzig, Germany*
[f]*Department of Theoretical Chemistry, University of Vienna, Währingerstraße 17, A-1090 Wien, Austria*
[g]*Center for non-coding RNA in Technology and Health, University of Copenhagen, Grønnegårdsvej 3, DK-1870 Frederiksberg, Denmark*
[h]*Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, NM 87501, USA*
[i]*Corresponding author*

Abstract

Canonical microRNAs are excised from their hairpin-shaped precursors by *Dicer*. In order to find possible exceptions to this rule and to identify additional substrates for *Dicer* processing we re-evaluate the small RNA sequencing data of the *Dicer* knockdown experiment in MCF-7 cells orignally published by Friedländer *et al.* [Nucleic Acids Res. **40**: 37-52 (2012)]. While the well-known non-*Dicer* mir-451 is not sufficiently expressed in these experiments, there are several additional *Dicer*-independent microRNAs, among them the important tumor supressor mir-663a. We recover previously described examples of non-miRNA *Dicer* substrates such as tRNA-Gln and several snoRNAs. Interestingly, sdRNAs derived from box C/D snoRNAs are *Dicer*-independent, while those derived from box H/ACA snoRNAs are often *Dicer* dependent. Several pol-III transcripts, in particular the vault RNAs and the great ape specific snaRs are processed by *Dicer*, while the small RNAs originating from Y RNAs seem to be *Dicer* independent.

Keywords: sdRNAs, snoRNAs, Y RNAs, vault RNAs, snaRs

1. Introduction

Canonical microRNAs are processed from a primary pol II transcript by means of the *Drosha*-dependent microprocessor complex (Gregory *et al.*, 2004), resulting in a characteristic hairpin of length 60–120 nucleotides. This pre-microRNA is then transported by Exportin-5 to the cytoplasm (Lund *et al.*, 2004), where the hairpin is cut by *Dicer* into a double stranded RNA about 22nt in length with a 2nt 3'-overhang (Murchison and Hannon, 2004). Several alternative pathways that bypass Drosha have been reported. The most prominent example are mirtrons (Okamura *et al.*, 2007; Ruby *et al.*, 2007), whose precursor hairpins are produced by splicing. A related, mirtron-like source of small RNAs requires both splicing and exosome-mediated trimming to extract the pre-microRNA hairpin (Flynt *et al.*, 2010; Chong *et al.*, 2010). More recently, it was shown that a few microRNAs, in particular mir-451, are matured without the help of *Dicer* (Cheloufi *et al.*, 2010; Cifuentes *et al.*, 2010). For a recent review of the many alternative pathways for the biogenesis of microRNAs and other, microRNA-like small RNA species see e.g. (Yang and Lai, 2011).

Apparently, *Dicer* is not only involved in microRNA biogenesis, but appears to be involved also in the processing of other small RNA species. Short, microRNA-like RNAs are processed from a diverse set of usually well-structured non-coding RNAs that includes tRNAs (Lee *et al.*, 2009; Cole *et al.*, 2009; Haussecker *et al.*, 2010; Findeiß *et al.*, 2011; Sobala and Hutvagner, 2011), snoRNAs (Kawaji *et al.*, 2008; Taft *et al.*, 2009; Langenberger *et al.*, 2010; Brameier *et al.*, 2011), vault RNAs (Stadler *et al.*, 2009; Persson *et al.*, 2009), Y RNAs (Langenberger *et al.*, 2010; Meiri *et al.*, 2010; Verhagen and Pruijn, 2011), and snRNAs (Langenberger *et al.*, 2010). Not much is known about the maturation of most of these small RNAs. The importance of *Dicer* for these small RNAs has been demonstrated in only a few cases: the small RNAs derived from human tRNA(Gln) are dependent on *Dicer* both *in vivo* and *in vitro* (Cole *et al.*, 2009), see also (Babiarz *et al.*, 2008). Some snoRNA derived sdRNAs show altered expression in mouse Dicer1 and Dgcr8 mutants (Taft *et al.*, 2009), and processing of ACA45 derived sdRNAs requires *Dicer* activity but not *Drosha*/*DGCR8* (Ender *et al.*, 2008). Endogenous siRNAs resulting from *Dicer* cleavage of long hairpins, typically deriving from SINEs with tandem inverted repeat structure have been reported in (Babiarz *et al.*, 2008).

Here we reevaluate a previously published set of RNA sequencing data (GSE31069) that compare the expression of small, microRNA-sized RNAs before and after *Dicer* knock-down in a MCF-7 cell line (Friedländer *et al.*, 2012). Our analysis focusses on the identification in particular of microRNAs that fail to respond to the depletion of *Dicer*, and conversely on those loci that are strongly *Dicer*-dependent but are not classified as microRNAs.

2. Methods

2.1. Data and Mapping

We downloaded a previously published sequencing data set series (GSE31069, (Friedländer *et al.*, 2012)) from the Gene Expression Omnibus (GEO) database (Edgar *et al.*, 2002). The data consists of four different samples, two containing short reads from the total cell content and two containing reads from the cytoplasmic fraction only. Both pairs contrast small RNA expression before and after *Dicer* knock-down in a MCF-7 cell line. The analysis reported here uses only the cytoplasmic sample pair (GSM769509 and GSM769511). Since short RNA processing takes place in this compartment we expect to reduce the noise from the nucleus.

All the adapter-free reads were mapped against the human genome (NCBI36.50 Release of July 2008) using `segemehl` (Hoffmann *et al.*, 2009): we activated the poly-A clipping, required small RNAs to map with an accuracy of at least 90% and selected the "best scoring hit strategy". With these settings we mapped 8,743,377 of 15,493,265 reads (56%) of the control sample and 5,471,242 of 9,237,490 reads (59%) of the *Dicer* knock-down sample. The resulting `sam` files were converted to `bam` format, using samtools (Li *et al.*, 2009) and subsequently translated to `bigWig` files using a custom perl script. The read density at each position in the `bigWig` files was normalized by the number of multiple hits of each read and the absolute number of mapped reads of each experiment (RPM) in order to make the two experiments comparable. We provide custom tracks for the UCSC Genome Browser (Kent *et al.*, 2002) to make the mapping results publicly available.

2.2. Expressed Sites and Annotation

In order to identify previously un-annotated loci with small RNA expression we created sorted bed files and then used `blockbuster` (Langenberger *et al.*, 2009) with default parameters to identify regions showing accumulations of at least 50 reads in at least one of control or *Dicer* knock-down data. We used `mergeBed` from `BEDtools` (Quinlan and Hall, 2010) to obtain the final

list of expressed regions of interest (1,946 for control and 1,798 for the knock-down set), which we call "sites" from now on.

We downloaded the latest annotations from different sources (1523 microRNA loci from miRBase v18 (Griffiths-Jones, 2004); 631 tRNA loci from gtRNAdb (Chan and Lowe, 2009); 402 snoRNA loci as well as 4528 other RNAs from UCSC annotation (Karolchik *et al.*, 2004)). This combined annotation track comprising 7,084 annotated ncRNA loci was compared with with our list of sites using `intersectBed` (Quinlan and Hall, 2010).

Furthermore, all reads were overlapped with the UCSC repeat masker track (Jurka *et al.*, 2000) and as soon as one read was mapped to a repeat associated region, all multiple hits of it were flagged with the type of repeat. If more than 50% of the expression of one site is caused by reads which are flagged as repeat associated, the whole site was flagged accordingly. In order to remove low-complexity sequences, which have a high probability of being random matches in short read data, we discarded all sites with a Shannon entropy of less than 1.6 bit.

2.3. Expression Levels

The expression level of each site, expressed in reads per kilobase of locus per million mapped reads (RPKM) was computed using the UCSC tool `bigWigAverageOverBed` (Kent *et al.*, 2002). From these values we derived, for each site, the $\log_2$-fold change λ between the Dicer knock-down sample and the RPKM of the control sample. All sites with $\lambda < 0$ are interpreted as *Dicer* processed. All sites, together with their annotations, their expression values, their λ and a link to the UCSC Genome Browser can be found at the supplement page http://www.bioinf.uni-leipzig.de/supplements/12-005.

2.4. Processing Pattern

Cleavage of a nearly double-stranded RNA by *Dicer* leads to a characteristic 2 nt overhang at the 3'end, see e.g. (Ji, 2008). In order to assess how important the thermodynamic stability of the precursor structure is for processing, we computed for a pair of putative single-stranded cleavage products, the following stability measure: `RNAcofold` (Bernhart *et al.*, 2006) is used to compute the energy of the duplex with the constraint that the joint structure exhibits the 2nt overhang at the 3'ends. Then the inner part of both sequences is shuffled 100 times so that the dinucleotide composition is preserved, while the terminal base pairs and overhanging nucleotides were left untouched. The resulting z-score of the co-folding energies is recorded. For each site we considered the two consecutive tags with the largest expression as candidates for *Dicer* processing.

Table 1: Fraction of *Dicer* processed sites among the annotated ncRNAs.

	***Dicer* processed**			
type	**yes**	**no**	**all**	**processed**
miRNA	255	10	265	96.2 %
tRNA	32	376	408	7.9 %
H/ACA snoRNA	8	4	12	66.0 %
C/D snoRNA	0	53	53	0.0 %
misc RNA	3	2	5	60.0 %
snRNA	2	92	94	2.1 %
scRNA	10	90	100	10.0 %
rRNA	28	254	282	9.9 %

In order to assess the overall similarity of a site with canonical microRNAs we use `RNAmicro` (Hertel and Stadler, 2006). This tool evaluates structural features as well as the pattern sequence conservation. We retrieved alignments of all sites with 20nt flanking sequence on both sides from the 8way-multiZ alignmemnt (human, chimp, orangutan, rhesus macaque, marmoset, mouse, opossum, platypus) (Blankenberg *et al.*, 2011). We extracted sequences from 8way-multiZ file, re-alignes them using clustalw (Larkin *et al.*, 2007) and used it to run RNAmicro. Then, the RNAmicro decision value (decV) was used to rate the sites, if they microRNA-like structures and conservations.

Dicer is well known to generate products in the narrow length range 21–28 nt, see e.g. (Starega-Roslan *et al.*, 2011). We therefore recorded the distribution of read lengths for each locus. In addition, we determined the lengths of blocks of reads `blockbuster` (Langenberger *et al.*, 2009) with default parameters. Read blocks summarize groups of reads that overlap nearly perfectly, hence its lengths is typically larger than that of individual reads.

3. Results

3.1. Identification of Dicer*-dependent small RNAs*

The *Dicer* knock-down (GSM769509) and control (GSM769511) datasets (Friedländer *et al.*, 2012) together identify 2,115 expressed sites. Of these, 1,048 overlap with the 7,084 annotated ncRNAs and 1,067 remain unannotated. After filtering out the low-complexity sites, we retain 1,002 annotated and 539 unknown sites for further analysis.

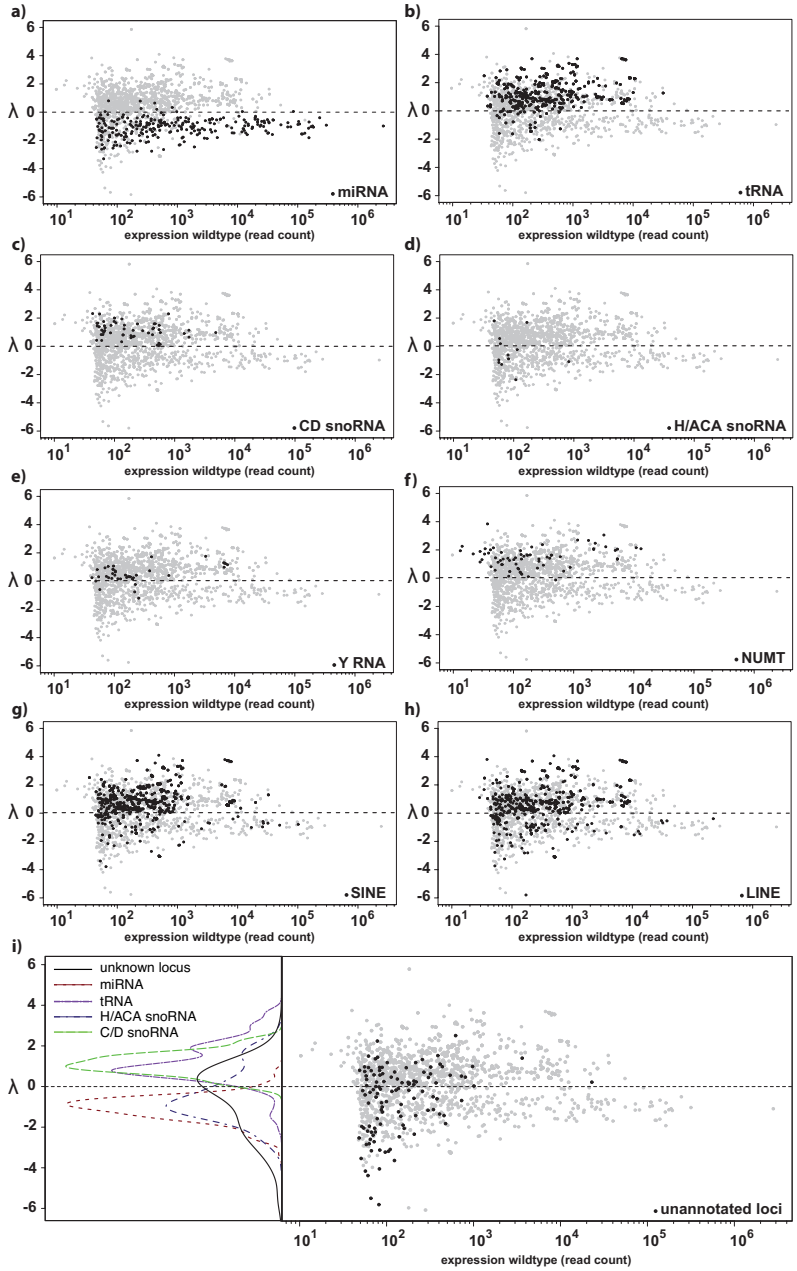

Figure 1: Summary of expression changes of small, microRNA-sized RNAs in response to a knock-down of *Dicer*. The entire dataset is shown in grey, specific groups are highlighted as black dots. (a) Almost all annotated microRNAs are down-regulated, i.e., exhibit $\log_2$-fold changes $\lambda < 0$. (b) Only a few tRNAs are downregulated. (c) None of the sdRNAs derived from box C/D snoRNAs is depeleted in response to the *Dicer* knock-down, while (d) the majority of the (small number of) sdRNAs derived from box H/ACA snoRNAs is *Dicer* dependent. (e) The small RNAs originating from Y RNAs and almost all Y RNA derived loci are not downregulated in response to *Dicer* knock down. (f) Mitochondrial transcripts and/or NUMTs are also a prolific source of small RNAs. These are independent of *Dicer* processing. Among repetitive elements, a substantial fraction of (g) expressed SINEs and (h) expressed LINEs shows *Dicer* dependent processing.

Fig. 1 summarizes the response of the small RNA sites to *Dicer* knockdown. The $\log_2$-fold change λ exhibits the expected bi-modal distribution separating in particular microRNAs from other small RNA products. Consistent with the original analysis of these datasets (Friedländer *et al.*, 2012), microRNAs are strongly reduced upon reduction of *Dicer* activity. A closer inspection, however, shows a more differentiated picture.

On the one hand, a small subgroup of microRNAs does not respond to the knockdown of *Dicer*. On the other hand, a sizable number of unannotated sites (some of which might constitute previously undescribed microRNAs) are associated with well-known structured RNAs exhibiting large negative values of λ, see Table 1.

A substantial fraction of sites expressing small RNAs are annotated repetitive elements, Table 2. Disregarding a moderate number of simple repeats and low complexity regions, which cannot be unambiguously distinguished

Table 2: Fraction of *Dicer* processed sites among the NUMTs and repeat associated regions.

	Dicer processed			
type	**yes**	**no**	**all**	**processed**
NUMT / chrM	1	66	67	1.5 %
SINE	126	427	553	22.8 %
LINE	141	368	509	27.7 %
LTR	81	306	387	20.9 %
DNA	27	106	133	20.3 %
Simple repeat	18	66	84	21.4 %
Low complexity	15	15	30	50.0 %
Other	1	13	14	7.1 %
Satellite	1	2	3	33.3 %
RNA	1	2	3	33.3 %
tRNA	0	2	2	0.0 %

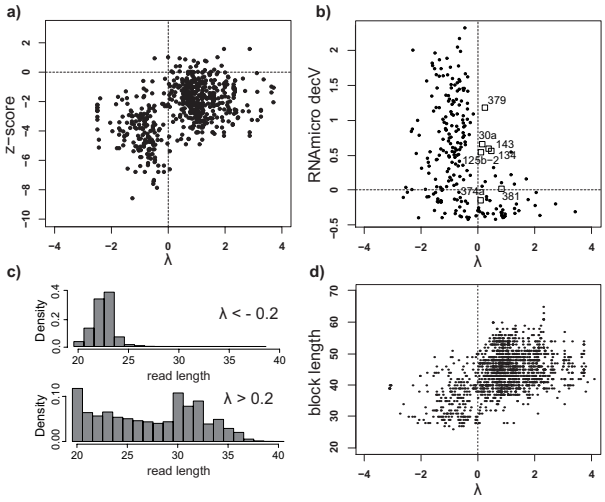

Figure 2: Correlation of *Dicer*-response λ with sequence-derived descriptors on the entire data set: (a) Free energy z-scores of constrained duplex structures with the 2-nt overhangs conforming to the canonical *Dicer* processing pattern. (b) Decision value of `RNAmicro`, a SVM-based machine learning tool trained to recognized canonical microRNAs. Its decision value combines the stability of the hairpin structure with patterns of sequence conservation but is agnostic about the location of the small RNA products. Only sites that form a hairpin structure can be scored by this method. (c) Reads with arising from *Dicer* processed regions, i.e. those with small values of λ, have a length within the size range between 20 and 25nt typical for microRNAs. The length distribution of reads without evidence for *Dicer*-processing is much broader and most reads fall outside the expected size range. (d) The reads orginating from *Dicer*-responsive RNA form shorter, more coherent read blocks.

from artefacts without further experimental evidence, we observe that about one fifth of repeat-associated small RNAs react to *Dicer*. This is not unexpected, as repetitive elements are one of the documented sources of novel microRNAs. Smalheiser and Torvik (2005), for instance showed that a subset of conventional mammalian microRNAs is derived from LINE-2 transposable elements. A family of miRNAs deriving from miniature inverted-repeat transposable elements (MITES) has been characterized by (Piriyapongsa and Jordan, 2007). A recent comprehensive analysis of microRNAs originating from transposable elements can be found in (Borchert *et al.*, 2011), see also (Yuan *et al.*, 2011).

3.2. Characterization of Dicer*-processed sites*

Dicer-processed small RNAs typically derive from helical regions that are significantly more stable than the precursor secondary structures of *Dicer*-independent small RNAs. Fig. 2a shows that in particular putative precursor structures that give rise to the typical processing patterns with 2nt overhangs are substantially stabilized *Dicer*-responsive small RNAs.

Canonical microRNAs also exhibit a characteristic pattern of sequence conservation that can help to distinguish them from other, similar, sources of small RNAs and from hairpin-like structures that are not processed into small RNAs, see e.g. (Lai *et al.*, 2003). `RNAmicro` (Hertel and Stadler, 2006) implements such a classifier based on a Support Vector Machine taking only a small number of structural and conservation based descriptors as input. Only sites that form hairpin structures can be scored by `RNAmicro`'s SVM. We have previously used `RNAmicro` to distinguish microRNA-like from snoRNA-like small RNA sites (Langenberger *et al.*, 2011). Fig. 2b shows that the `RNAmicro` decision value is also correlated with λ. With few exceptions, large decision values are limited to *Dicer* responsive sites.

Fig 2c summarizes the distribution of read lengths. As expected, nearly all reads arising from sites with $\lambda < -0.2$ have a lengths between 20 and 25nt, consistent with *Dicer* processing (Starega-Roslan *et al.*, 2011). In contrast, short reads from sites with $\lambda > 0.2$, i.e., those that are clearly not resulting from *Dicer* cleavage, are typically longer and show a flat distribution. We also observe a difference in the length of read blocks as determined by `blockbuster` (Langenberger *et al.*, 2009). Sites with $\lambda < 0$ have on average much shorter block sizes, often consisting only of a single block of microRNAs, Fig. 2d. Since the start and end position of mature microRNAs can vary by a couple of nucleotides (Ebhardt *et al.*, 2010) such that overlapping microRNAs read blocks have a length of around 30 nt.

3.3. *Structured regions processed by* Dicer *indicate potential microRNA candidates*

The data set used here has been generated specifically for the purpose of detecting novel microRNAs (Friedländer *et al.*, 2012). Among the un-annotated, non-repetitive sites with $\lambda < 0$ six additional structured regions (Figure 3) were found. One of them (Figure 3a) is located in an intergenic region Figure 3a, is located in intergenic region far away from any annotation. Three of four intronic sequences (Figure 3d-f) fold into hairpins and the short reads map to the stem positions expected for mirtrons (Okamura *et al.*, 2007; Ruby *et al.*, 2007). The 3' end of a candidate located in a SYT12-intron is determined by the splice acceptor (Figure 3b). Given its stable hairpin structure these findings suggest that it belongs to the recently described class of "semi-mirtrons" that require both splicing and exosome-mediated trimming for maturation (Flynt *et al.*, 2010; Chong *et al.*, 2010). The structure and the positions of mapped reads of the remaining candidate (Figure 3c) do not conform to a typical microRNA. Nevertheless, the ten-fold reduction of the read coverage in the *Dicer* knockdown experiment indicates *Dicer*-processing. Since these reads perfectly but not uniquely map to the intronic region, this candidate is of particular interest for further analysis.

3.4. Dicer-*processed non-microRNAs*

Surprisingly, there is a large number of well-known structured non-coding RNAs from which *Dicer*-sensitive small RNAs are produced.

A prominent example are the vault RNAs. The largest response is observed for vtRNA2-1 with $\lambda = -2.12$. This locus was originally classified as hsa-mir-886 but later-on recognized as a polymerase-III transcript (Canella *et al.*, 2010) and vault RNA paralog (Nandy *et al.*, 2009; Stadler *et al.*, 2009). The other three vault RNA loci also give rise to short RNAs (Persson *et al.*, 2009) and respond negatively to the *Dicer* depletion: λ(vtRNA1-1) = -0.14, λ(vtRNA1-2) = -0.76. The vtRNA1-3 locus is not sufficiently expressed.

The snaR ncRNAs (Parrott and Mathews, 2007) are pol-III transcripts that emerged in the ancestor of the African Great Apes from an Alu-derived precursor (Raha *et al.*, 2010; Parrott *et al.*, 2011). Fig. 4 shows that microRNA-like small RNAs are processed from the lower end of the stem-loop structure, which resembles a canonical pre-microRNA hairpin except for its length of more than 100nt. The snaR-derived small RNAs show the typical 2 nt 3' overhangs. Their expression depends very strongly on the *Dicer* concentration.

The situation is more complex for tRNAs and snoRNAs. While many of them give rise to small RNA products, the majority is not influenced by the *Dicer* knockdown. A small subset of tRNAs, on the other hand is clearly subject to *Dicer* processing. These include in particular tRNA-Gln-CTG with $\lambda = -2.05$ as noted already previously by Cole *et al.* (2009). Other tRNAs with a clear *Dicer* signature are tRNA-Asn-GTT ($\lambda = -1.47$), tRNA-Asn-ATT ($\lambda = -0.83$), tRNA-Ala-CGC ($\lambda = -1.28$), tRNA-Ile-TAT ($\lambda = -1.19$), tRNA-Glu-TTC ($\lambda = -0.79$). None of the four `mirbase` "microRNAs" that are derived from tRNAs (mir-1274/tRNA-Lys, mir-1280/tRNA-Leu, mir-720/tRNA-Thr, mir-1308/tRNA-Gly) are expressed at sufficiently high levels to estimate λ.

Small nucleolar RNAs can share several characteristics with microRNAs, including similar components in their processing, see (Scott and Ono, 2011) for a recent review. The structural similarities between H/ACA snoRNAs and microRNAs are most obvious and have been noticed in several computational studies. Scott *et al.* (2009), for instance, report twenty miRNA precursors that show significant similarity to H/ACA snoRNAs; of these miR-151, miR-605, mir-664 = SNORA36B, miR-215, and miR-140 even bind to dyskerin, a component of the H/ACA snoRNP. On the other hand, *Dicer* processing has been demonstrated previously for SNORA45 (Ender *et al.*, 2008). Consistently, we find λ(SNORA45) = -1.55. Of the 12 H/ACA snoRNAs with sufficient expression 8 have $\lambda < 0$ (Table 3), indicating that short reads from H/ACA snoRNAs are typically a product of *Dicer* processing. Interestingly, two H/ACA snoRNAs were classified as novel microRNAs by `mirdeep2` (Friedländer *et al.*, 2012): SNORA36A ($\lambda = -1.33$) and SNORA33 $\lambda = 0.15$. We emphasize, however, that only a small minority of H/ACA snoRNAs leads to abundant processing products. In addition these small RNAs are independent of Drosha (Ender *et al.*, 2008; Taft *et al.*, 2009; Brameier *et al.*, 2011), and in some cases Drosha even inhibits sdRNA formation (Taft *et al.*, 2009), emphasizing that the snoRNAs and (canonical) microRNAs are in general clearly distinguished entities.

A quite different picture emerges for box C/D snoRNAs. Although small RNAs are abundantly produced from box C/D snoRNAs in our data set, Tab. 1, there is no indication that any of them is a *Dicer* substrate. The box C/D snoRNAs that are discussed as possibly microRNA-like in (Langenberger *et al.*, 2011) show only marginal expression levels and no indication for *Dicer* processing. On the other hand, of the five microRNAs that resemble box C/D snoRNAs (having C and D boxes

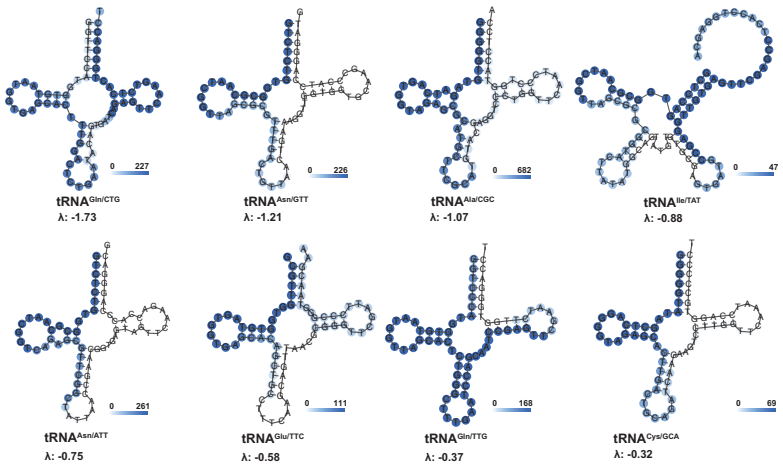

Figure 5: Several tRNAs are processed into small RNAs by *Dicer*. The processing patterns shown some similarities, in particular a tendency to have large short read coverage on the 3' side of the tRNA clover leaf. With the exception of tRNA-Ile-TAT the small RNAs are derived from within the mature tRNA.

Table 3: Box H/ACA snoRNAs processed by *Dicer*. SNORA36B (Ender *et al.*, 2008) (also annotated as mir-664) does not reach a sufficent expression level in MCF-7 cells.

snoRNA	λ	snoRNA	λ
SNORA45	-1.55	SNORA36B	*
SNORA51	-2.42	SNORA46	-1.00
SNORA36A	-1.33	SNORA56	-0.93
SNORA17	-1.19	SNORA7B	-0.66
SCARNA3	-1.13	SNORA7A	-0.28

in close proximity in the precursor and binding to fibrillarin) (Ono *et al.*, 2011), four are *Dicer* substrates (miR-27b $\lambda = -0.90$, miR-16-1 $\lambda = -0.40$, mir-28 $\lambda = -0.95$, and let-7g $\lambda = -1.16$) and the fifth (mir-31) is not sufficiently expressed in MCF-7 cells. It appears, thus, that *Dicer*-processing clearly distinguished between *bona fide* microRNAs and small RNAs derived from box C/D snoRNAs.

Y RNAs are small pol-III transcripts that originate from RNA component of the Ro RNP particle and have a role in DNA replication (Christov *et al.*, 2006). The four paralogous human Y RNAs form a cluster on Chr.7(148M) (Mosig *et al.*, 2007; Perreault *et al.*, 2007). The canonical loci show no evidence of *Dicer* processing hY3 $\lambda = 0.12$, hY4 $\lambda = 1.25$, hY1 $\lambda = 1.70$, hY5 $\lambda = 1.74$. We note that fragments from hY5 have also been annotated as mir-1975.

In addition to the canonical Y RNA cluster, however, there are more than a thousand Y RNA pseudogenes scattered across the genome (Perreault *et al.*, 2007). The deep sequencing data shows that several of these loci form a source of short reads. A few of the Y4-derived loci sites have negative values of λ. We note, however, these have relatively low expression levels and might be confounded by mapping artefacts. In total, 11 sites that are derived from Y RNA sequences are classified as microRNAs by `RNAmicro`, six of which have moderate negative values of λ.

3.5. MicroRNA not processed by Dicer

The best-studied microRNA that is not processed by *Dicer* is mir-451. Unfortunately this site is not significantly expressed in MCF-7 cells, so that we cannot use it as a control. There are ten additional microRNAs with $\lambda > 0$. Six of them (mir-30a, mir-143, mir-374a, mir-379, mir-381, and mir-134) derive from precursor hairpins that are recognized by `RNAmicro`. Two of these, mir-30a and mir-374a, exhibit exceptionally high levels of expression and feature short RNAs derived from both sides of the precursor stem, Figure 7. We suspect that they are exceptionally good substrates for *Dicer* so that their maturation is least affected by *Dicer* concentrations. The evolutionarily ancient mir-125b-2 also exhibits both a canonical read pattern and a canonical pattern of sequence conservation. Nevertheless, it shows no reaction to *Dicer* knockdown, $\lambda = 0.03$.

For mir-143, mir-381, mir-134, mir-4417, and mir-4516 no mir* reads were detectable. Mir-4417 is present in monkeys only (Supplemental Material), and no homologs are detectable for hsa-mir-4516, precluding the analysis of patterns of sequence conservation for these two microRNAs.

The entire precursor hairpin of mir-3676 is covered by small RNA sequences. A closer inspection shows, however, that mir-3676 coincides with tRNAThr-AGT and is thus clearly an erroneous annotation. The mis-annotated

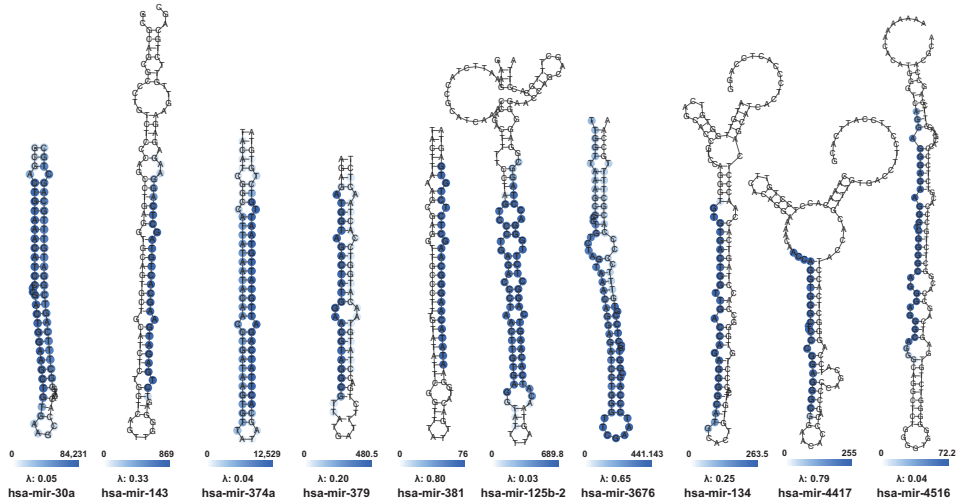

Figure 7: MicroRNAs that are not processed by Dicer.

"mir-3195", furthermore, corresponds to a GC-rich low-complexity region located with the first exon of the TAF4 gene.

The sequence of mir-663a is very GU-rich and does not meet our exclusion criterion for low-complexity sequences. We retained it in our data set because it is well document as an important tumor suppressor (Pan *et al.*, 2010; Yi *et al.*, 2012). In contrast to canonical microRNAs, its primary sequence is quite poorly conserved although it can be found throughout the major eutherian groups. Its read pattern also strongly deviates from the expectation for microRNAs. Similar to mir-451, the precursor is covered with a background of short reads, as also seen in the cummulative read patterns provided by the MicroRNA Registry, albeit there is dominating, most frequently produced "mature microRNAs".

4. Discussion

A rapidly expanding zoo of diverse small RNA species has emerged following the discovery of RNA interference (Fire *et al.*, 1998) and microRNAs (Lee *et al.*, 1993) almost two decades ago. With the rapid increase of high throughput sequencing data the boundaries between the different subdivisions of small RNAs have become more and more blurry.

Here we have focussed on the generation of small RNAs from their double-stranded precursors. Making use of a publicly available dataset (Friedländer *et al.*, 2012) we find, consistent with the well-established knowledge, that the overwhelming majority of miRBase microRNAs is processed by *Dicer*. There are, however, several notable exceptions. Cole *et al.* (2009) argue that *Dicer* knockdown with siRNAs for a short period of time sometimes does not result in a significant change in the miRNA steady state level due to slow microRNA turnover. At least some of the *Dicer*-unresponsive miRNAs, however, exhibit unusual structural features and/or read patterns that deviate substantially from canonical microRNAs. While $\lambda > 0$ in itself is course not sufficient proof for *Dicer*-independence, it is at least a strong indication and helps to identify candidates for further analysis.

Dicer-processing is not limited to microRNAs. Several polymerase-III transcripts are prolific *Dicer* substrates, including human vault RNAs, the great ape specific snaRs, and a small set of about a dozen tRNAs. While the vault RNAs products function like microRNAs, small RNAs derived from tRNA-Gln-CTG do not function in this way: they do not associate with argonaut presumably due to the fact that these small RNAs are just too small (Cole *et al.*, 2009). Despite their similarity with vault RNAs, including a secondary structure with a long terminal stem, there is no evidence that the abundant small RNAs deriving from Y RNAs are produced by *Dicer* cleavage. Both main classes of small nucleolar RNAs are sources of abundant small RNAs. While all of the highly expressed box C/D snoRNAs are processed independently of *Dicer*, the situation is different for H/ACA snoRNAs. Most box H/ACA snoRNAs are a source of small RNAs, but in most cases the expression levels are small, at least in the invesitaged MCF-7 libraries. Among the highly expressed ones, however, the majority clearly is a *Dicer* substrate.

In summary, there does not seem to be a clear separation between between processing pathways resulting in

small RNAs. Instead, the picture of an intricate network of interleaved alternatives emerges, in which the individual processing steps can be freely combined. As a consequence, it appears that a particular sequence of processing steps is neither a sufficient nor a necessary condition for a particular role. Small RNA sequencing data such as the ones analyzed here reveal only the end points of a likely more complex processing cascade. Longer potential intermediates, such as pre-microRNA hairpins or the parts of tRNAs resulting from stress-related cleavage in the anticodon loop (Jöchl *et al.*, 2008; Thompson *et al.*, 2008) are invisible here. It will be an interesting topic for future research to investigate if and how the generation of small RNAs is linked to other RNA processing mechanisms.

Acknowledgments

This work is supported in part by the *Deutsche Forschungsgemeinschaft* (SPP-1258 "Sensory and Regulatory RNAs in Prokaryotes": STA 850/7-2 to PFS). LIFE – Leipzig Research Center for Civilization Diseases, Universität Leipzig is funded by means of the European Social Fund and the Free State of Saxony.

Conflict of interest

The authors declare no conflict of interest.

References

Babiarz JE, Ruby JG, Wang Y, Bartel DP, Blelloch R, 2008. Mouse ES cells express endogenous shRNAs, siRNAs, and other Microprocessor-independent, Dicer-dependent small RNAs. Genes Dev 22:2773–2785.

Bernhart SH, Tafer H, Mückstein U, Flamm C, Stadler PF, Hofacker IL, 2006. Partition function and base pairing probabilities of RNA heterodimers. Algorithms Mol Biol 1:3.

Blankenberg D, Taylor J, Nekrutenko A, Team TG, 2011. Making whole genome multiple alignments usable for biologists. Bioinformatics 27:2426–2428.

Borchert GM, Holton NW, Williams JD, Hernan WL, Bishop IP, Dembosky JA, Elste JE, Gregoire NS, Kim JA, Koehler WW, Lengerich JC, Medema AA, Nguyen MA, Ower GD, Rarick MA, Strong BN, Tardi NJ, Tasker NM, Wozniak DJ, Gatto C, Larson ED, 2011. Comprehensive analysis of microRNA genomic loci identifies pervasive repetitive-element origins. Mob Genet Elements 1:8–17.

Brameier M, Herwig A, Reinhardt R, Walter L, Gruber J, 2011. Human box C/D snoRNAs with miRNA like functions: expanding the range of regulatory RNAs. Nucleic Acids Res 39:675–686.

Canella D, Praz V, Reina JH, Cousin P, Hernandez N, 2010. Defining the RNA polymerase III transcriptome: Genome-wide localization of the RNA polymerase III transcription machinery in human cells. Genome Res 20:710–721.

Chan P, Lowe T, 2009. GtRNAdb: a database of transfer RNA genes detected in genomic sequence. Nucleic Acids Res 37:D93–D97.

Cheloufi S, Dos Santos CO, Chong MM, Hannon GJ, 2010. A dicer-independent miRNA biogenesis pathway that requires Ago catalysis. Nature 465:584–589.

Chong MMW, Zhang G, Cheloufi S, Neubert TA, Hannon GJ, Littman DR, 2010. Canonical and alternate functions of the microRNA biogenesis machinery. Genes Dev 24:1951–1960.

Christov CP, Gardiner TJ, Szüts D, Krude T, 2006. Functional requirement of noncoding Y RNAs for human chromosomal DNA replication. Mol Cell Biol 26:6993–7004.

Cifuentes D, Xue H, Taylor DW, Patnode H, Mishima Y, Cheloufi S, Ma E, Mane S, Hannon GJ, Lawson ND, Wolfe SA, Giraldez AJ, 2010. A novel miRNA processing pathway independent of Dicer requires Argonaute2 catalytic activity. Science 328:1694–1698.

Cole C, Sobala A, Lu C, Thatcher SR, Bowman A, Brown JW, Green PJ, Barton GJ, Hutvagner G, 2009. Filtering of deep sequencing data reveals the existence of abundant Dicer-dependent small RNAs derived from tRNAs. RNA 15:2147–2160.

Ebhardt HA, Fedynak A, Fahlman RP, 2010. Naturally occurring variations in sequence length creates microRNA isoforms that differ in argonaute effector complex specificity. Silence 1:12.

Edgar R, Domrachev M, Lash A, 2002. Gene expression omnibus: Ncbi gene expression and hybridization array data repository. Nucleic Acids Res 30:207–210.

Ender C, Krek A, Friedländer MR, Beitzinger M, Weinmann L, Chen W, Pfeffer S, Rajewsky N, Meister G, 2008. A human snoRNA with microRNA-like functions. Mol Cell 32:519–528.

Findeiß S, Langenberger D, Stadler PF, Hoffmann S, 2011. Traces of post-transcriptional RNA modifications in deep sequencing data. Biol Chem 392:305–313.

Fire A, Xu S, Montgomery MK, Kostas SA, Driver SE, Mello CC, 1998. Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. Nature 391:806–811.

Flynt AS, Greimann JC, Chung WJ, Lima CD, Lai EC, 2010. MicroRNA biogenesis via splicing and exosome-mediated trimming in *Drosophila*. Mol Cell 38:900–907.

Friedländer MR, Mackowiak SD, Li N, Chen W, Rajewsky N, 2012. `miRDeep2` accurately identifies known and hundreds of novel microRNA genes in seven animal clades. Nucleic Acids Res 40:37–52.

Gregory RI, Yan KP, Amuthan G, Chendrimada T, Doratotaj B, Cooch N, Shiekhattar R, 2004. The Microprocessor complex mediates the genesis of microRNAs. Nature 432:235–240.

Griffiths-Jones S, 2004. The microRNA registry. Nucleic Acids Res 32:D109–D111.

Haussecker D, Huang Y, Lau A, Parameswaran P, Fire AZ, Kay MA, 2010. Human tRNA-derived small RNAs in the global regulation of RNA silencing. RNA 16:673–695.

Hertel J, Stadler PF, 2006. Hairpins in a haystack: Recognizing microRNA precursors in comparative genomics data. Bioinformatics 22:e197–e202.

Hoffmann S, Otto C, Kurtz S, Sharma CM, Khaitovich P, Vogel J, Stadler PF, Hackermüller J, 2009. Fast mapping of short sequences with mismatches, insertions and deletions using index structures. PLoS Comput Biol 5:e1000502.

Ji X, 2008. The mechanism of rnase III action: how dicer dices. Curr Top Microbiol Immunol 320:99–116.

Jöchl C, Rederstorff M, Hertel J, Stadler PF, Hofacker IL, Schrettl M, Haas H, Hüttenhofer A, 2008. Small ncRNA transcriptome analysis from *Aspergillus fumigatus* suggests a novel mechanism for regulation of protein-synthesis. Nucleic Acids Res 36:2677–2689.

Jurka J, *et al.*, 2000. Repbase update: a database and an electronic journal of repetitive elements. Trends Genet 16:418–420.

Karolchik D, Hinrichs A, Furey T, Roskin K, Sugnet C, Haussler D, Kent W, 2004. The UCSC table browser data retrieval tool. Nucleic Acids Res 32:D493–D496.

Kawaji H, Nakamura M, Takahashi1 Y, Sandelin A, Katayama S, email Fukuda S, Daub C, Kai C, Jun Kawai J, Yasuda J, Carninci P, Hayashizaki Y, 2008. Hidden layers of human small RNAs. BMC Genomics 9:157.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D, *et al.*, 2002. The human genome browser at ucsc. Genome Res 12:996–1006.

Lai EC, Tomancak P, Williams RW, Rubin GM, 2003. Computational identification of drosophila microRNA genes. Genome Biol 4:R42.

Langenberger D, Bartschat S, Hertel J, Hoffmann S, Tafer H, Stadler PF, 2011. Microrna or not microrna? In: de Souza ON, Telles GP, Palakal MJ, editors, Advances in Bioinformatics and Computational Biology, 6th Brazilian Symposium on Bioinformatics, BSB 2011, vol. 6832 of *Lecture Notes in Computer Science*, (pp. 1–9). Berlin, Heidelberg: Springer.

Langenberger D, Bermudez-Santana C, Hertel J, Hoffmann S, Khaitovich P, Stadler PF, 2009. Evidence for human microrna-offset rnas in small rna sequencing data. Bioinformatics 25:2298–2301.

Langenberger D, Bermudez-Santana C, Stadler PF, Hoffmann S, 2010. Identification and classification of small RNAs in transcriptome sequence data. Pac Symp Biocomput 15:80–87.

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson J, Gibson TJ, Higgins DG, 2007. Clustal w and clustal x version 2.0. Bioinformatics 23:2947–2948.

Lee RC, Feinbaum RL, Ambros V, 1993. The *C. elegans* heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell 75:843–854.

Lee YS, Shibata Y, Malhotra A, Dutta A, 2009. A novel class of small RNAs: tRNA-derived RNA fragments (tRFs). Genes Dev 23:2639–2649.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, *et al.*, 2009. The sequence alignment/map format and samtools. Bioinformatics 25:2078–2079.

Lund E, Güttinger S, Calado A, Dahlberg J, Kutay U, 2004. Nuclear export of microRNA precursors. Science 303:95–98.

Meiri E, Levy A, Benjamin H, Ben-David M, Cohen L, Dov A, Dromi N, Elyakim E, Yerushalmi N, Zion O, Lithwick-Yanai G, Sitbon E, 2010. Discovery of microRNAs and other small RNAs in solid tumors. Nucleic Acids Res 38:6234–6246.

Mosig A, Guofeng M, Stadler BMR, Stadler PF, 2007. Evolution of the vertebrate Y RNA cluster. Th Biosci 126:9–14.

Murchison EP, Hannon GJ, 2004. miRNAs on the move: miRNA biogenesis and the RNAi machinery. Curr Opin Cell Biol 16:223–229.

Nandy C, Mrázek J, Stoiber H, Grässer FA, Hüttenhofer A, Polacek N, 2009. Epstein-Barr virus-induced expression of a novel human vault RNA. J Mol Biol 388:776–784.

Okamura K, Hagen JW, Duan H, Tyler DM, Lai EC, 2007. The mirtron pathway generates microRNA-class regulatory RNAs in *Drosophila*. Cell 130:89–100.

Ono M, Scott MS, Yamada K, Avolio F, Barton GJ, Lamond AI, 2011. Identification of human miRNA precursors that resemble box C/D snoRNAs. Nucleic Acids Res 39:3879–3891.

Pan J, Hu H, Zhou Z, Sun L, Peng L, Yu L, Sun L, Liu J, Yang Z, Ran Y, 2010. Tumor-suppressive mir-663 gene induces mitotic catastrophe growth arrest in human gastric cancer cells. Oncol Rep 24:105–112.

Parrott AM, Mathews MB, 2007. Novel rapidly evolving hominid RNAs bind nuclear factor 90 and display tissue-restricted distribution. Nucleic Acids Res 35:6249–6258.

Parrott AM, Tsai M, Batchu P, Ryan K, Ozer HL, Tian B, Mathews MB, 2011. The evolution and expression of the snaR family of small non-coding RNAs. Nucleic Acids Res 39:1485–1500.

Perreault J, Perreault JP, Boire G, 2007. Ro-associated Y RNAs in metazoans: evolution and diversification. Mol Biol Evol 24:1678–1689.

Persson H, Kvist A, Vallon-Christersson J, Medstrand P, Borg A, Rovira C, 2009. The non-coding RNA of the multidrug resistance-linked vault particle encodes multiple regulatory small RNAs. Nat Cell Biol 11:1268–1271.

Piriyapongsa J, Jordan IK, 2007. A family of human microRNA genes from miniature inverted-repeat transposable elements. PLoS One 2:e203.

Quinlan A, Hall I, 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26:841–842.

Raha D, Wang Z, Moqtaderi Z, Wu L, Zhong G, Gerstein M, Struhl K, Snyder M, 2010. Close association of RNA polymerase II and many transcription factors with Pol III genes. Proc Natl Acad Sci USA 107:3639–3644.

Ruby GJ, Jan CH, Bartell DP, 2007. Intronic microRNA precursors that bypass Drosha processing. Nature 48:83–86.

Scott MS, Avolio F, Ono M, Lamond AI, Barton GJ, 2009. Human miRNA precursors with box H/ACA snoRNA features. PLoS Comput Biol 5:e1000507.

Scott MS, Ono M, 2011. From snoRNA to miRNA: Dual function regulatory non-coding RNAs. Biochimie 93:1987–1992.

Smalheiser NR, Torvik VI, 2005. Mammalian microRNAs derived from genomic repeats. Trends Genet 21:322–326.

Sobala A, Hutvagner G, 2011. Transfer RNA-derived fragments: origins, processing, and functions. Wiley Interdiscip Rev RNA 2:853–862.

Stadler PF, Chen JJL, Hackermüller J, Hoffmann S, Horn F, Khaitovich P, Kretzschmar AK, Mosig A, Prohaska SJ, Qi X, Schutt K, Ullmann K, 2009. Evolution of vault RNAs. Mol Biol Evol 26:1975–1991.

Starega-Roslan J, Krol J, Koscianska E, Kozlowski P, Szlachcic WJ, Sobczak K, Krzyzosiak WJ, 2011. Structural basis of microRNA length variety. Nucleic Acids Res 39:257–268.

Taft RJ, Glazov EA, Lassmann T, Hayashizaki Y, Carninci P, Mattick JS, 2009. Small RNAs derived from snoRNAs. RNA 15:1233–1240.

Thompson DM, Lu C, Green PJ, Parker R, 2008. tRNA cleavage is a conserved response to oxidative stress in eukaryotes. RNA 14:2095–2103.

Verhagen AP, Pruijn GJ, 2011. Are the Ro RNP-associated Y RNAs concealing microRNAs? Y RNA-derived miRNAs may be involved in autoimmunity. Bioessays 33:674–682.

Yang JS, Lai EC, 2011. Alternative miRNA biogenesis pathways and the interpretation of core miRNA pathway mutants. Mol Cell 43:892–903.

Yi C, Wang Q, Wang L, Huang Y, Li L, Liu L, Zhou X, Xie G, Kang T, Wang H, Zeng M, Ma J, Zeng Y, Yun JP, 2012. MiR-663, a microRNA targeting p21(WAF1/CIP1), promotes the proliferation and tumorigenesis of nasopharyngeal carcinoma. Oncogene Doi: 10.1038/onc.2011.629.

Yuan Z, Sun XS, Liu H, Xie J, 2011. MicroRNA genes derived from repetitive elements and expanded by segmental duplication events in mammalian genomes. PLoS ONE 6:e17666.

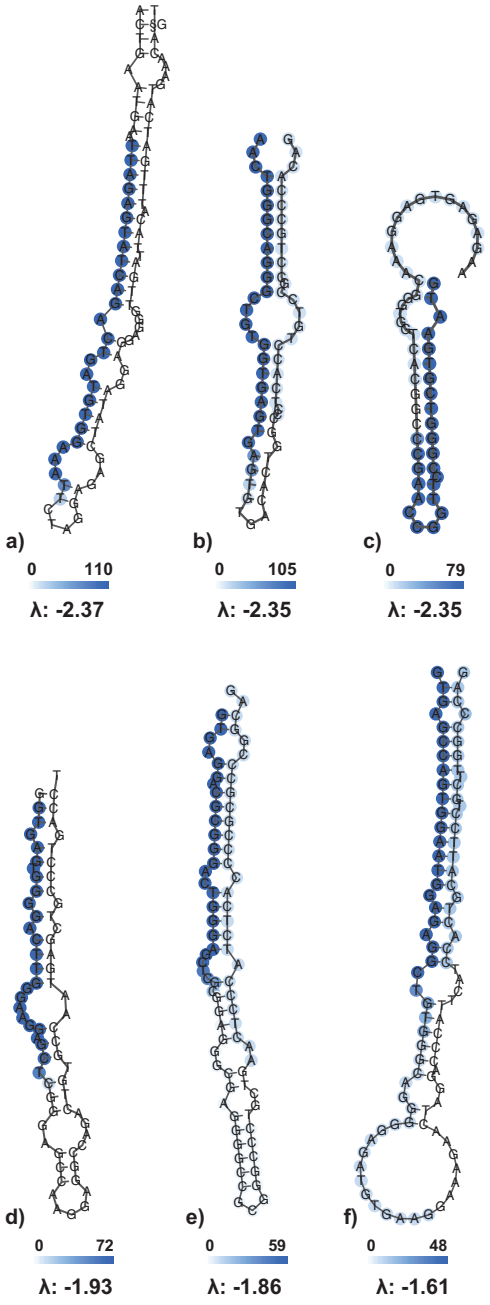

Figure 3: Six un-annotated non-repetitive loci that are processed by *Dicer*. (a) intergenic chr2:81,100,049-81,100,134(+); (b) a "semi-mirtron" in an intron of SYT12 chr11:66,569,729-66,569,790(+); (c) a source in an intron of SLC4A2 chr7:150,394,782-150,394,835(+); three mirtrons: (c) SLC4A2 chr7:150,394,782-150,394,835(+); (d) FLNA chrX:153,235,873-153,235,943(-); (e) MAP3K4 chr1:27,559,917-27,559,998(-); (f) TRIM28 chr19:63,753,464-63,753,555(+). The color scale represents the coverage on a logarithmic scale.

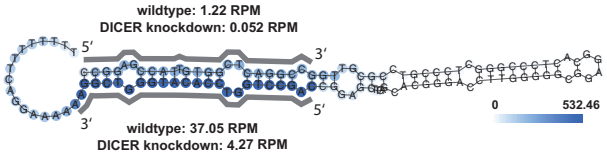

Figure 4: snaRs are processed by Dicer. Highlighted are the tags showing the highest expression.

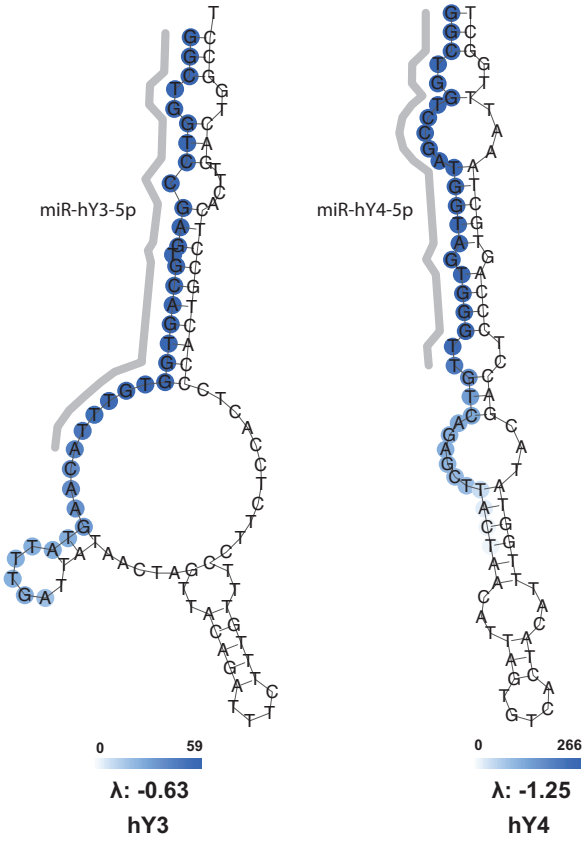

Figure 6: Small RNAs derived from Y RNAs. miR-hY3-5p and miR-hY4-5p (Verhagen and Pruijn, 2011) are highlighted.