

## <SUPPLEMENTARY METHODS>

### Extraction of extended stem-loops in human genome

We extracted the extended stem-loops on human chromosomes 16, 17, 18 and 19 to screen miRNA precursors. We scanned the chromosomes with a window size of 100 nucleotides and with 90 nucleotides of overlap at both ends. RNA secondary structures of the window fragments were predicted with their reverse complementary sequences by the RNAfold program (available at <http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi>). We then screened the extended stem-loop structures using several criteria, including sequence length (64–90 nucleotides), stem length (above 22 nucleotides), bulge size (under 15 nucleotides), loop size (3–20 nucleotides) and free energy (under –25 kcal). We thereby we extracted 65539, 68458, 34853 and 62229 sequences of stem-loop structures on chromosomes 16, 17, 18 and 19, respectively.

### Method based on conservation

We first trained the scoring model with 68 positive and 1000 negative training data, which are already used for training ProMiR, based on the following characteristic features reported by a group (Lim *et al.*, 2003) ; (1) base pairing of the rest of the fold-back; (2) the conservation in the 5' half and 3' half of the miRNA; (3) sequence biases in the first five bases of the miRNA; (4) a symmetric tendency of internal bulges and loops in the miRNA region; (5) a GC ratio in the pre-miRNA. The scoring model has been implemented by the position weighted matrix (PWM) to apply the characteristic features. **The PWM was constructed by the following equation**

$$P_{ij} = \log \left( \frac{f(x_{ij}) + s(x)}{p(x_j)} \right),$$

where  $P_{ij}$  is a weighted score for  $j$ -th symbol (base) of the  $i$ -th position of sequence  $x$ .  $f(x_{ij})$  is a frequency for  $j$ -th symbol (base) of  $i$ -th position of sequence  $x$ .  $s(x)$  is a pseudo-count.  $p(x_j)$  is a background frequency of  $j$ -th symbol. Consequently, the PWM consists of an  $n \times m$  matrix when the length of sequence  $x$  is  $n$  and the number of symbols is  $m$ . The conservation extent is evaluated by calculating the probability of test sequences using the PWM.

### Construction of RNAi for Drosha

The siRNA duplex (siDrosha-4) was designed to anneal to Drosha mRNA, in which the target sequence was 5'-AAGGACCAAGUAUUCAGCAAG-3'. The siRNA duplexes

were transfected into HeLa cells using Oligofectamine reagent (Invitrogen, Carlsbad, CA, USA). Total RNA was prepared seven days after transfection.

### **RT-PCR**

After extraction of total RNA, synthesis of cDNA was carried out using Oligo-dT primer.

### **Real-time PCR**

Real-time PCR was performed in the iCycler IQ™ system for 50 cycles of 45 seconds at 95 °C followed by 30 seconds at each primer's T<sub>m</sub> and 30 sec at 72 °C. A PCR supermix from Bio-Rad containing *Taq* DNA polymerase, MgCl<sub>2</sub>, dNTP and SYBR Green I was used. The SYBR Green I is a fluorescent indicator of double-stranded DNA synthesis.

### **PCR primers**

>Drosha\_lower

CATACCAGGAAATGAGCTTG

>Drosha\_upper

ATGACATCAAGAAGGTGGTG

>let-7a-1\_lower

GCTGCACTACATCTCTTTAAGAC

>let-7a-1\_upper

CCTTCCTGTGGTGCTCAACT

>mir-345\_lower

GCAACCAAGTGGGTCAGAGA

>mir-345\_upper

GGTAGGTGTTGTCAGTAGTGCAG

>NC16-1C\_lower

AACAGTTCTGCCTAACTAGTCAA

>NC16-1C\_upper

TGCTCCACTGATGTAAAGGTATG

>NC16-2C\_lower

CTCCCTGGGGGTGGTAGCTCTC

>NC16-2C\_upper

TGCTGTCAGCCGTGAATGGGGAC

>NC16-3\_lower

AAACTTTCAAATACTAGCCCATT

>NC16-3\_upper

TTTCACAAAGGGAGAACGGAGG

>NC17-5\_lower

GACGTACATTCCAATCTGACCCC

>NC17-5\_upper

ACCAGCTTCACCTCGGCTCAT

>NC17-9\_lower

CAAGGAGAACAAGGCCCTCTAAG

>NC19-9\_upper

GATTCCCCTCCCCGTAC

>NC18-2\_lower

GACTGGCTGAGCGTCGAAT

>NC18-2\_upper

TGTAAACCAGTAGTTTTCAACCC

>NC18-3\_3'Lower 65.6

TGCAACTCAAACAGATTGTTA

>NC18-3\_5'Upper 65.6

TGTTTTAAACTGTAAGCAAGT

>NC19-5\_3'Lower 70.2

CATGCAGTAAATTCTGTTTATGG

>NC19-5\_5'Upper 68.4

AAATTGTGGTAATTGTGTAACCA

>NC19-6\_3Lower 70.8

CACACACACGCAAAGATAG

>NC19-6\_5Upper 72.7

GAGTGACGAGGGCAAAC