Author(s): Chen, Liang; Heikkinen, Liisa; Knott, Emily; Liang, Yanchun; Wong, Garry

Title: Evolutionary conservation and function of the human embryonic stem cell specific miR-302/367 cluster

Year: 2015

Version:

Please cite the original version:

Chen, L., Heikkinen, L., Knott, E., Liang, Y., & Wong, G. (2015). Evolutionary conservation and function of the human embryonic stem cell specific miR-302/367 cluster. Comparative Biochemistry and Physiology D: Genomics and Proteomics, 16, 83-98. https://doi.org/10.1016/j.cbd.2015.08.002

Comparative Biochemistry and Physiology, Part D 16 (2015) 83–98

Contents lists available at ScienceDirect

Comparative Biochemistry and Physiology, Part D

journal homepage: www.elsevier.com/locate/cbpd

Evolutionary conservation and function of the human embryonic stem cell specific miR-302/367 cluster

Liang Chen [a,b], Liisa Heikkinen [c], K. Emily Knott [c], Yanchun Liang [a,*], Garry Wong [b,d,**]

[a] Key Laboratory of Symbolic Computation and Knowledge Engineering of the Ministry of Education, College of Computer Science and Technology, Jilin University, Changchun 130012, China
[b] A.I. Virtanen Institute, Faculty of Health Sciences, University of Eastern Finland, PL 1627, Kuopio 70211, Finland
[c] University of Jyväskylä, Department of Biological & Environmental Science, P.O. Box 35, FI-40014, University of Jyväskylä, Finland
[d] Faculty of Health Sciences, University of Macau, Taipa, Macau S.A.R., China

ARTICLE INFO

Article history:
Received 15 May 2015
Received in revised form 14 August 2015
Accepted 19 August 2015
Available online 29 August 2015

Keywords:
Stem cells
Cancer
miRNA
Functional genomics
Target analysis

ABSTRACT

miRNA clusters define a group of related miRNAs closely localized in the genome with an evolution that remains poorly understood. The miR-302/367 cluster represents a single polycistronic transcript that produces five precursor miRNAs. The cluster is highly expressed and essential for maintenance of human embryonic stem cells. We found the cluster to be highly conserved and present in most mammals. In primates, seed sequence and miRNA structure are conserved, but inter-precursor sequences are evolving. Insertions of new miRNAs, deletions of individual miRNAs, and a cluster duplication observed in different species suggest an actively evolving cluster. Core transcriptional machinery consisting of NANOG and OCT-4 transcription factors that define stem cells are present upstream of the miR-302/367 cluster. Interestingly, we found the miR-302/367 cluster flanking region to be enriched as a target site of other miRNAs suggesting a mechanism for feedback control. Analysis of miR-302 and miR-367 targets demonstrated concordance of gene set enrichment groups at high gene ontology levels. This cluster also expresses isomiRs providing another means of establishing sequence diversity. Finally, using three different kidney tumor datasets, we observed consistent expression of miR-302 family members in normal tissue while adjacent tumor tissue showed a significant lack of expression. Clustering expression levels of miR-302 validated target genes showed a significant correlation between miR-302/367 cluster miRNAs and a subset of validated gene targets in healthy and adjacent tumor tissues. Taken together, our data show a highly conserved and still evolving miRNA cluster that may have additional unrecognized functions.

1. Introduction

microRNAs (miRNAs) represent a class of small noncoding RNAs originally discovered in *Caenorhabditis elegans* but now known to be present in most metazoan species. miRNAs function as essential gene regulators in the development and maturity of plants and animals including humans. miRNAs are initially transcribed as primary transcripts (pri-miRNAs) by RNA polymerase II or III, and are subsequently cleaved to 65–70 nucleotide precursors (pre-miRNAs) by RNA endonuclease Drosha (Bartel, 2004; Ha and Kim, 2014). Pre-miRNAs occur as thermodynamically stable hairpin structures that are exported to the cytoplasm via the protein exportin and processed to active 21–22 nucleotide mature miRNAs by RNA endonuclease Dicer within the RNA silencing complex (RISC). Mature miRNAs are further processed in RISC by selecting one of two strands (−5p or −3p) and guided to their cognate mRNAs where they facilitate destabilization and eventual degradation of the target, prevent translation via recruitment of inhibitory factors, or in some cases can enhance RNA stability (Krol et al., 2010; Li et al., 2013). Humans express an estimated 1881 pre-miRNAs which produce 2588 mature miRNAs (Kozomara and Griffiths-Jones, 2014).

While miRNA genes are found throughout the genome, their localization is not random. Initial cloning and mapping of miRNAs to the chromosomal locations found that up to 30% of miRNAs were located within close proximity to other miRNAs and these genomic locations were termed miRNA clusters (Lagos-Quintana et al., 2001; Lau et al., 2001; Lee and Ambros, 2001). The distance between miRNAs within a cluster can vary. While some clusters have been defined as multiple miRNAs within a 1 Mbp region, correlation of expression between adjacent miRNA pairs has been shown to decrease beyond 10 kbp in *Drosophila melanogaster* (Marco et al., 2013). The number of miRNAs in a cluster can vary with some containing over 70 miRNAs (Kozomara and Griffiths-Jones, 2014). The evolutionary mechanism by which miRNAs are located in clusters appears to be via gene duplication, emergence of new hairpins, or via mechanisms involving genomic

* Corresponding author.
** Correspondence to: G. Wong, Faculty of Health Sciences, Building E12 Room 3005, Avenida da Universidade, Taipa, Macau, China. Tel.: +853 88224979 (office), +853 62822701 (mobile); fax: +853 88222314.
E-mail address: GarryGWong@umac.mo (G. Wong).

http://dx.doi.org/10.1016/j.cbd.2015.08.002

rearrangements (Marco et al., 2013; Mohammed et al., 2014). As a result, clusters can contain highly homologous miRNAs of the same miRNA family, as well as those from other miRNA families. However, miRNA family members do not necessarily remain clustered, and can be distributed randomly throughout the genome; the *let-7* miRNA family is a prime example.

Primary miRNAs within clusters can be transcribed as single transcripts that contain multiple pre-miRNAs (Saini et al, 2007, 2008). These polycistronic units are rare in mammals, but common in more densely populated genomes such as bacteria, the classic example being the lac operon (Jacob and Monad, 1961). The principle advantage of a polycistronic unit is the coordinated expression of multiple genes, implying that the gene products carry out different but coordinated functions, such as separate enzymes within the same biochemical pathway. Similarly, it has been proposed that miRNAs within miRNA clusters, whether transcribed as single or polycistronic units, share

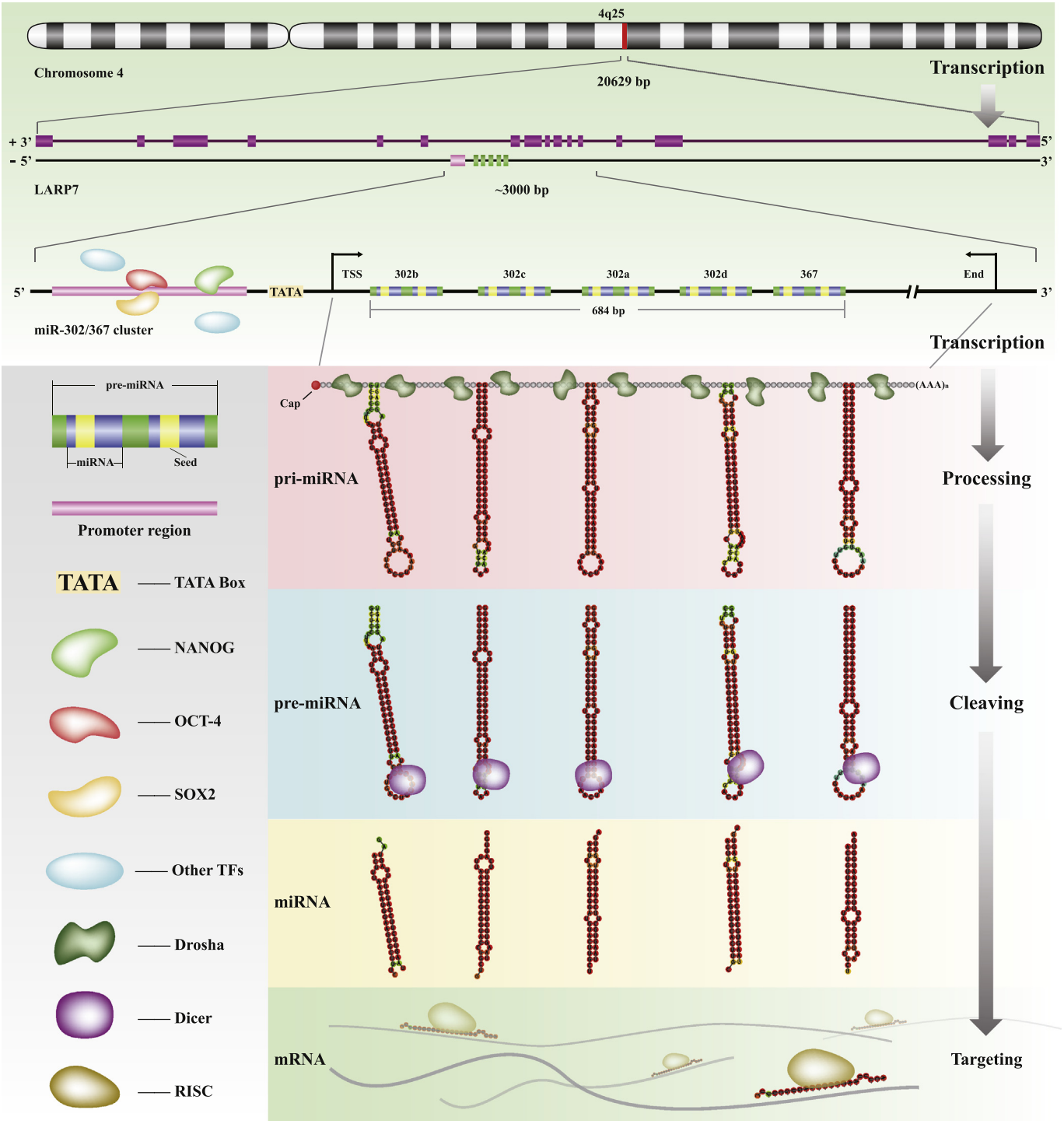

Fig. 1. miR-302/367 cluster biogenesis. The human miR-302/367 cluster contains five precursor miRNAs within 684 bp on chromosome band 4q25. This cluster is located within an intron of the LARP7 gene and transcribed on the opposite strand. The core promoter of the miRNA cluster has been defined previously (Barroso-delJesus et al., 2008). Pre-miRNA structures are lowest free energy forms predicted (Lorenz et al., 2011). Drosha, Dicer and RISC complex size are not to scale.

common molecular functions via regulation of genes involved in similar biological processes (Kim and Nam, 2006). As can be expected from genes derived from duplication events, miRNAs within miRNA clusters often, but not always, share similar structures. In fact, mature miRNAs within a cluster may share similar seed sequences and thus belong to the same miRNA family. However, different structural regions (seed, loop, mature miRNA, 5′ and 3′ arms) within a miRNA gene appear to be under different evolutionary pressure (Saunders et al., 2007; Warthmann et al., 2008). Currently, it is estimated that humans express 153 miRNA clusters dispersed throughout the genome to produce 468 pre-miRNAs (miRBase v21, inter-miRNA distance < 10 kb).

The miR-302/367 cluster consists of five pre-miRNAs: mir-302b, mir-302c, mir-302a, mir-302d (from the mir-302 family), and mir-367. The cluster is encoded as a single polycistronic transcriptional unit and follows a canonical biogenesis pathway (Fig. 1) (Suh et al., 2004). The core miR-302/367 cluster region is 684 bp long located in an intron of the LARP7 gene. The primary miRNA (about 2500 nucleotides) is transcribed from a region located on the opposite strand of the LARP7 gene on chromosome 4 in humans (Barroso-delJesus et al., 2008; Marson et al., 2008). The core promoter for this transcriptional unit has been defined to include embryonic stem cell factors NANOG, OCT3/4, SOX2, and REX1 (Barroso-delJesus et al., 2008). The structure of the gene coding for the human miR-302/367 cluster has been characterized (Barroso-delJesus et al., 2008). Early studies on human embryonic stem cells (hESC) indicated that this cluster is highly expressed and, therefore, constitutes a signature miRNA of stem cells (Laurent, 2008). It is also highly expressed in induced pluripotent stem cells (iPSC) and its expression decreases during differentiation (Laurent, 2008). Overexpression of the miR-302/367 cluster promotes cellular reprogramming and maintains the stemness of hESCs (Liao et al., 2011; Miyoshi et al., 2011). Genetically engineered knockouts have also shown that it has an important role in maintaining pluripotency (Zhang et al., 2013). Gene expression analysis has shown that miRNAs in this cluster control cell cycle regulation, epigenetic factors, TGF-β signaling, and BMP inhibitors (Lakshmipathy et al., 2010; Lipchina et al., 2011; Kim et al., 2014). The miR-302/367 cluster may also act in cancer processes. The miR-302/367 cluster can act as a tumor suppressor in cervical carcinoma cells (Cai et al., 2013) and unrestricted somatic stem cells (Jamshidi-Adegani et al., 2014). miR-302/367 clusters are overexpressed in malignant germ cell tumors (GCTs) independent of histologic subtype, tumor site (ovary, testis, or extragonadal), or patient age (Palmer et al., 2010) and could be potential serum biomarkers of malignant germ cell tumors (Murray et al., 2011; Rijlaarsdam et al., 2015). The miR-302/367 cluster may also act in coordination with other miRNA clusters. Along with miR-20 and miR-92 families, the miR-302 genes can control mitochondrial apoptosis machinery via regulation of proapoptotic protein BIM levels (Pernaute et al., 2014).

While the role of miR-302/367 in hESC and iPSC function is well known, its origin and distribution in organisms other than human remains poorly characterized. Moreover, the conservation of specific structural elements remains to be defined. By investigating and identifying conserved regions of the miRNA cluster, insight can be gained into the important functional units and in which organisms they may take place. Furthermore, analysis of miRNA cluster targets and enrichment of gene ontology groups can provide insight into molecular functions of the miRNA cluster. Hierarchical clustering of validated miRNA cluster targets in healthy and tumor tissues can also provide insight into specific genes and their associated pathways regulated by this miRNA cluster. In this study, we analyze the presence of the miR-302/367 cluster across species using available sequence data. Based on its presence and structure we infer its origin. Using functional genomics tools we describe its potential physiological functions. Finally, we provide evidence that the cluster's validated molecular targets might prove useful as kidney cancer biomarkers.

2. Materials and methods

2.1. Data compilation

All information of pre-miRNAs and mature miRNAs in the miR-302/367 cluster were extracted from miRBase (Release 21, http://www.mirbase.org/), including accessions/symbols, coordinates, sequences and family information. Full length miR-302/367 cluster sequences were obtained from Ensembl database with the coordinates as follows: from start of mir-302b to the end of mir-367, GRCh38:4:112647874:112648557:-1, length 684 bp (Cunningham et al., 2015). The miRNA clusters used as control were also selected from miRBase by setting the inter-miRNA distance < 500 bp and requiring at least 4 miRNAs in the same strand. The miR-371/372/373 cluster described in this study is homologous with mouse miR-290/295 cluster which produces the most abundant miRNAs in mESCs and is regulated by master transcription factors OCT-4, SOX2, and NANOG (Whyte et al., 2013).

We applied the miR-302/367 cluster sequence as a query sequence in searching the Ensembl site (which focuses on vertebrates and also contains invertebrate model organisms, such as *C. elegans* and *D. melanogaster*) using BLAT with default parameters (Kent, 2002). We also used BLAT to search for the 684 bp long sequence in the UCSC Genome Browser for those species that were not available in Ensembl. We also searched for the cluster sequence in Ensembl specific databases EnsemblMetazoa and EnsemblPlants. Based on the BLAT results, the full length cluster sequences of the relevant species were batched and obtained from Ensembl by REST API using a Perl script. For those species only in the UCSC Genome Browser (Rosenbloom et al., 2015), we obtained the sequence directly using tools (view- > DNA) on the website. We then recorded the coordinates of the cluster sequence used in each species. At the same time, we determined whether the cluster was located in the intron of LARP7 by inspecting the gene annotation track in Ensembl Genome Browser and UCSC Genome Browser. In some species, the host gene name was not annotated as LARP7, but could be annotated as homologous to LARP7 by Ensembl and GenBank. We obtained the 5′ and 3′ flanking sequences based on the coordinates of the cluster for those species.

SNPs were collected from Biomart (http://asia.ensembl.org/biomart/martview/). Queries were performed using the R package biomaRt (Durinck et al., 2009).

Next generation sequencing data for stem cells, including miRNA-seq and Chip-seq, were downloaded from ENCODE at UCSC and ENCODE (ENCODE Project Consortium, 2012). We filtered out the H1hesc datasets and downloaded the sequence read alignments in the BAM format. miRNA-seq bam files links are: http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeCshlShortRnaSeq/wgEncodeCshlShortRnaSeqH1hescCellShorttotalTapAlnRep2.bam, http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeCshlShortRnaSeq/wgEncodeCshlShortRnaSeqH1hescCytosolShorttotalTapAlnRep2.bam, and http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeCshlShortRnaSeq/wgEncodeCshlShortRnaSeqH1hescNucleusShorttotalTapAlnRep2.bam.

Chip-seq bam file accession numbers in ENCODE were: ENCFF000ORJ and ENCFF000OPV. Samtools 1.0 was used to generate the bam file, including printing all alignments into SAM format, splitting out certain region from the genome, and preparing the data for the depth plot. An in-house Perl script was used for filtering out the isomiRs based on the output of Samtools (Li et al., 2009).

Thirty-four cancer types including a total of 10,551 datasets were collected from the public database TCGA (http://cancergenome.nih.gov/). The list of 34 cancers were: Acute myeloid leukemia (LAML), adrenocortical carcinoma (ACC), bladder urothelial carcinoma (BLCA), brain lower grade glioma (LGG), breast invasive carcinoma (BRCA), cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC), cholangiocarcinoma (CHOL), colon adenocarcinoma (COAD),

esophageal carcinoma (ESCA), FFPE pilot phase II (FPPP), glioblastoma multiforme (GBM), head and neck squamous cell carcinoma (HNSC), kidney chromophobe (KICH), kidney renal clear cell carcinoma (KIRC), kidney renal papillary cell carcinoma (KIRP), liver hepatocellular carcinoma (LIHC), lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), lymphoid neoplasm diffuse large B-cell lymphoma (DLBC), mesothelioma (MESO), ovarian serous cystadenocarcinoma (OV), pancreatic adenocarcinoma (PAAD), pheochromocytoma and paraganglioma (PCPG), prostate adenocarcinoma (PRAD), rectum adenocarcinoma (READ), sarcoma (SARC), Skin cutaneous melanoma (SKCM), stomach adenocarcinoma (STAD), testicular germ cell tumors (TGCT), thymoma (THYM), thyroid carcinoma (THCA), uterine carcinosarcoma (UCS), uterine corpus endometrial carcinoma (UCEC), uveal melanoma (UVM). We only focused on the tumor and normal samples in the datasets which code for 01 (primary solid tumor) and 11 (solid tissue normal) in the TCGA database, except for LAML, for which we kept samples which code for 03 (primary blood derived cancer). miRNA-seq data were used for detecting the expression of the miR-302/367 cluster, and the reads per million (RPM) miRNA mapped were directly extracted from tab-delimited (.txt) data files in level 3 (TCGA provides three levels of data: level 3 is processed data). For some cancers, such as KIRC, KICH, and KIRP, we also downloaded their matched RNA-seq datasets (the same sample was used for both miRNA-seq and RNA-seq) and normalized counts were used to denote gene expression.

Validated miRNA targets were queried from miRTarBase 4.5 (Hsu et al., 2014), Tarbase 7.0 (Vlachos et al., 2015), miRecords 4 (Xiao et al., 2009) and predictions were downloaded from TargetScan 6.1 (Lewis et al., 2005) and microrna.org (Betel et al., 2008). Outdated miRBase miRNA symbols were mapped to the newest v21 miRBase symbols.

The coordinator bed file of all human genes was downloaded from the UCSC Genome Browser. One GFF file with genome coordinates of all human miRNA were downloaded from miRBase v21. The accession or identifier conversions between different databases were achieved by db2db (Mudunuri et al., 2009). We used UCSC Genome Browser tools LiftOver to convert other human assembly coordinates to Hg19.

2.2. Evolutionary analysis

We performed multiple sequence alignment using ClustalX 2 (Larkin et al., 2007) and edited the results with BioEdit. MEGA 6.0 was used for constructing Neighbor-Joining trees based on Kimura's 2-parameter distance (Kimura, 1980; Saitou and Nei, 1987; Tamura et al., 2013) and illustrated using iTOL (Letunic and Bork, 2007). The secondary structure and distance between structures were predicted and calculated using programs (RNAfold and RNAdistance) in the ViennaRNA package version 2.1.7 (Lorenz et al., 2011). A sequence divergence index, using Kimura's 2-parameter distance was calculated with the function dist.dna() in the R package ape version 3.2 with parameter model equal to K80 (Paradis et al., 2004).

We first used ClustalX 2 to create multiple alignments of all cluster sequences. Because the cluster miRNAs had uniform distribution on sequences spanning different lengths and because there were four miRNAs from the same family, we set gap extension to a small value of 0.5 and kept others as default parameters. We then split the multi-alignment results based on the human pre-miRNA coordinates. Every cluster sequence from 58 species was cleaved into nine segments, including five putative pre-miRNA and four inter-precursor sequences. Since, the whole process was based on the multi-alignment result file output from ClustalX, the gaps were reserved.

We built two evolutionary trees using MEGA 6.0: one analysis was based on the multi-alignment of the complete cluster sequences and the other was based on the joined sequences of the five pre-miRNAs (inter-precursor sequence sequences were removed). The neighbor joining algorithm was used for tree reconstruction, using Kimura's 2-parameter model to calculate the divergence. The number of bootstrap replications was set to 1000. The resulting output Newick format tree files and annotation files were uploaded into iTOL for visualization.

Based on the same multi-alignment of the complete cluster sequence produced in ClustalX, we compared sequence differences among all species against human. We used a sliding window of 77 nucleotides along the alignment of human and other species. This parameter was selected since it is the average size of a pre-miRNA (Li et al., 2010). By moving the sliding window base by base, we calculated the rate of nucleotide substitution within the window. A sequence divergence index based on Kimura's 2-parameter distance was calculated and recorded for plotting. Likewise, comparison among all species and human were made based on the split alignment result, but this time we removed all the gaps. From these data, we predicted secondary structure using RNAfold with default parameters, and finally, we used RNAdistance to make the 2D distance comparison with humans for each part of the alignment.

To identify orthologs of miR-302-3p and miR-367-3p, we identified their 7-mer seed sequences and searched all miRBase v21 mature sequences with the same seed (Mohammed et al., 2014). For consideration of miRNAs with potential seed shifting, we only searched starting at 0, 1, and 2 nt offsets for perfect sequence matches.

2.3. NGS data analysis

We applied Chipster 3.0 (Kallio et al., 2011) for generating miRNA-seq data from ENCODE. Bam files and miRNA coordinates files were uploaded to Chipster and we chose HTSeq to perform miRNA expression analysis. Normalized counts were used as Reads per Million (RPM).

$$RPM = \frac{miRNA_{read} \times 10^6}{total\ mapped_{read}}$$

miRNA and gene expression associations with cancer were analyzed directly using the Level 3 dataset in TCGA. For KIRC, KIRP and KICH datasets, we not only chose matched miRNA-seq and RNA-seq, but also tumor and normal matched datasets. Samtools-1.0 was applied to generate the depth data for Chip-seq data. For identifying and counting isomiRs, we wrote in-house Perl scripts. We filtered isomiR sequences with the set length threshold and set the region to be located within 5 bp of the 3′ and 5′ arm of the canonical mature miRNA coordinate, mismatches were not allowed.

2.4. Target analysis

We used miRTarBase, Tarbase, and miRecords as well-known and regularly updated validated miRNA target databases for identifying high quality validated target genes. We downloaded the full list of miRNA-target gene pairs from miRTarBase and miRecords and filtered the targets of miR-302a/b/c/d and 367. Tarbase could only be manually searched and we downloaded the target gene for every miRNA member in the cluster. Old versions of miRNA symbols were converted to the new version miRBase v21 and target gene names were converted to the official gene symbol. All validated target genes from public databases were retained.

We used miRanda and TargetScan for miRNA target gene prediction (Enright et al., 2003; Lewis et al., 2005). The miRanda algorithm focusses on three properties: complementarity, thermodynamic stability, and conservation (Enright et al., 2003). The source code and prediction of targets were obtained from http://www.microrna.org/. Human miRNA targets from microRNA.org (August 2010 Release) with a good mirSVR score and conserved miRNA were chosen. The target list was filtered by setting another criterion score: align_score > 120, for all the targets that had already obtained a favorable mirSVR score (Betel et al., 2010).

TargetScan algorithm seeks out conserved 8mer and 7mer sites that match the seed region of each miRNA. The predicted conserved targets list was downloaded from TargetScan (Release 6.1, http://www.targetscan.org/cgi-bin/targetscan/data_download.cgi?db=vert_61). All predictions from TargetScan were saved. Only targets of miR-302a/b/c/d-3p and miR-367-3p were available in both databases and we applied them to strengthen the functional analysis of the dominant miR-302 family and miR-367-3p in the cluster. Targets of miR-302a/b/c/d-5p and miR-367-5p were extracted from validated targets database, most of them records in Tarbase. A Perl script was used for extracting, ranking, and combining the target genes. DAVID (http://david.abcc.ncifcrf.gov/) was used for GO and KEGG enrichment analysis (Huang da et al., 2009).

For analysis of the interaction of miRNAs with the miR-302/367 cluster, we searched for all possible miRNA target sites both upstream and downstream 3 kb of the miR-320/367 cluster. The source code (Release 6.1) of TargetScan, miRNA sequences and families were obtained from http://www.targetscan.org/cgi-bin/targetscan/data_download.cgi?db=vert_61. Based on the miRNA sequences and families available in TargetScan, 12 species were chosen: chicken (*Gallus gallus*), chimpanzee (*Pan troglodytes*), cow (*Bos taurus*), dog (*Canis familiaris*), horse (*Equus caballus*), human (*Homo sapiens*), macaque (*Macaca mulatta*), mouse (*Mus musculus*), opossum (*Monodelphis domestica*), platypus (*Ornithorrhynchus anatinus*), rat (*Rattus norvegicus*) and *X. tropicalis* (*Xenopus tropicalis*). We did not find the cluster in teleost fish. Multiple sequence alignment was applied to 5′ and 3′ 3 kb flanking regions and the cluster region from these 12 species separately. TargetScan could detect conserved 8mer and 7mer target sites in the alignment, and could score them by Pct value and context score (Grimson et al., 2007; Friedman et al., 2009). Circos 0.64 was used for illustrating the interaction between miRNAs and their target sites (Krzywinski et al., 2009).

2.5. Differential expression analysis

We simply employed a t-test to identify genes differentially expressed (Callow et al., 2000). Because the data are tumor and tumor adjacent paired, here we apply two-sample t-test to rank the differentially expressed genes.

Normalized counts vector of gene i in tumor adjacent tissue is $N_i = [N_1, N_2, \ldots, N_n]$, while $T_i = [T_1, T_2, \ldots, T_n]$ is for tumor tissue, and n is the size of sample. $\overline{N}$ and $\overline{T}$ are the sample mean from N_i and T_i.

The t statistic to test whether gene expressions (normalized counts) are different can be calculated as follows:

$$t = \frac{(\overline{T} - \overline{N})}{S_{TN} \cdot \sqrt{\frac{1}{n}}}$$

where:

$$S_{TN} = \sqrt{S_T^2 + S_N^2}$$

S_{TN} is the grand standard deviation.

We perform function t.test() in the basic stats R package to test genes in two types of sample. Each gene could get a p-value as return and rank them on the basis of the p-value.

3. Results

3.1. Evolutionary analysis

Using annotated sequences in miRBase, sequences downloadable from Ensembl, and sequences searched from the UCSC Genome browser followed by structure prediction, we observed the presence of the miR-302/367 cluster in 58 species with all members present in nearly all mammalian species, amphibians, and birds (Table 1). A notable discovery was the platypus, a semi-aquatic egg laying mammal. We were not able to detect the cluster in fish, insects, nematodes, bacteria, or plants. Many of the miRNAs in the cluster had been previously confidently characterized and curated in miRBase (14 species) (Table 1, Supplementary Table 1). For species not curated in miRBase, the cluster could be predicted based on available sequences in Ensembl, based on BLASTN and RNAFold (32 species). For the remaining 12 species available from the UCSC Genome Browser, the pre-miRNA sequences were obtained based on multiple sequence alignment. We applied miPred to predict whether the homologous pre-miRNAs which are not annotated by miRBase and Ensembl are real pre-miRNAs (Jiang et al., 2007). Only a small number of species appear to have lost miR-302 family members within the cluster. Rat has lost part of mir-302c, anole lizard is missing mir-302a/c while *X. tropicalis* is missing mir-302c and part of mir-302b. Dolphin is missing mir-302a/d in the cluster, but it appears with another copy of mir-302a in a sequenced scaffold. Minke whale has all members of the cluster with extra copies of mir-302a and mir-367 in another scaffold (Supplementary Table 2). We were not able to detect any instances of clusters where the miR-302 family was present, but mir-367 was missing. We also did not find instances of mir-367 independent from the cluster. Several species had incomplete sequences, with ambiguous “NN” sequences in the cluster region and, therefore, our analyses of these species (orangutan, tree shrew and sloth) were not determinable (Supplementary Table 1). In chicken and turkey, another miRNA, mir-1811, appears to have emerged in the miR-302/367 cluster (Glazov et al., 2008). In gorilla, we find a duplication of the same cluster in a nearby region (gorGor3.1: 4:122081981:122082677:1; gorGor3.1: 4:122095855:122096551:-1; Coordinate format is Assembly: chromsome:steart:end:strand), due to an apparent LARP7 duplication in gorilla (Supplementary Table 2). We were able to detect the LARP7 gene or a homologous LARP7 gene in 56 species (Table 1). In species where the LARP7 gene was absent, we were not able to detect the cluster.

A high level conservation in a miRNA gene among distantly related organisms reflects various evolutionary pressures acting on these regions, and suggests that these gene regions are essential for processing and function (Warthmann et al., 2008). We compared the divergence between the human miR-302/367 cluster and the other 57 species using Kimura's 2-parameter distance. We observed more significant divergence among species in inter-precursor sequence regions of the cluster (i.e. regions between precursor miRNAs), not only between closely related species, such as gorilla and chimpanzee, but between human and mouse (Fig. 2A). Comparisons to all analyzed species are shown in Supplementary Fig. 1. The seed region appears to be the most conserved, and other precursor miRNA regions are also well conserved.

We also extended this analysis to secondary structure. After predicting the 2D structure for the precursor miRNAs and inter-precursor sequence regions separately, we compared human to the other 57 species separately and created a heat map (Fig. 2B). The figure clearly shows the large differences in inter-precursor sequence compared to precursor miRNA regions. We also noticed that some miRNA precursors are more conserved structurally such as mir-302a, whereas both of the inter-precursor sequence regions flanking mir-302a appear to be the most structurally diverged (Fig. 2B). We made pairwise comparisons between species for each pre-miRNA's 2D structure, represented in five heat maps for mir-302a/b/c/d and mir-367, respectively (Supplementary Fig. 2). The results show different conserved patterns in the five pre-miRNAs as follows: 1) Secondary structure of mir-302a is most conserved in mammals except for megabat and platypus; 2) Secondary structure of mir-302b is similar in 8 primates and all reptiles; 3) Secondary structure of mir-302c is well conserved within taxonomic classes, especially in mammals; 4) Secondary structure of mir-302d is conserved in primates, aves and reptiles; 5) Secondary structure of mir-367 appears to be more diverse in comparison to the other pre-miRNAs, although it achieves the best sequence similarity performance between species.

Table 1
The information of miR-302/367 cluster in 58 species.

	Species Latin name	Common name	Class	Remark	miRBase	LARP7	miR-302b	miR-302c	miR-302a	miR-302d	miR-367
1	*Gallus gallus*	Chicken	Av	IN	+	+	+++	+++	+++	+++	+++
2	*Pan troglodytes*	Chimpanzee	M	CO	+	+	+++	+++	+++	+++	+++
3	*Bos taurus*	Cow	M	CO	+	+	+++	+++	+++	+++	+++
4	*Canis familiaris*	Dog	M	CO	+	+	+++	+++	+++	+++	+++
5	*Equus caballus*	Horse	M	CO	+	+	+++	+++	+++	+++	+++
6	*Homo sapiens*	Human	M	CO	+	+	+++	+++	+++	+++	+++
7	*Macaca mulatta*	Macaque	M	CO	+	+	+++	+++	+++	+++	+++
8	*Mus musculus*	Mouse	M	CO	+	+	+++	+++	+++	+++	+++
9	*Monodelphis domestica*	Opossum	M	CO	+	+	+++	+++	+++	+++	+++
10	*Oryctolagus cuniculus*	Rabbit	M	CO	+	+	+++	+++	+++	+++	+++
11	*Anolis carolinensis*	Anole lizard	R	DE	+	+	+++	–	–	++	+++
12	*Ornithorhynchus anatinus*	Platypus	M	CO	+	+	+++	++	+++	++	++
13	*Xenopus tropicalis*	Xenopus	Am	DE	+	+	–	–	+++	+	+++
14	*Taeniopygia guttata*	Zebra finch	Av	CO	+	+	++	++	++	++	+++
15	*Vicugna pacos*	Alpaca	M	CO	–	+	++	++	++	++	++
16	*Dasypus novemcinctus*	Armadillo	M	CO	–	+	++	++	++	++	++
17	*Otolemur garnettii*	Bushbaby	M	CO	–	+	++	++	++	++	++
18	*Felis catus*	Cat	M	CO	–	+	++	++	++	++	++
19	*Pelodiscus sinensis*	Chinese softshell turtle	R	CO	–	+	++	++	++	++	++
20	*Anas platyrhynchos*	Duck	Av	OU	–	+	++	++	++	++	++
21	*Loxodonta africana*	Elephant	M	CO	–	+	++	++	++	++	++
22	*Mustela putorius furo*	Ferret	M	CO	–	+	++	++	++	++	++
23	*Ficedula albicollis*	Flycatcher	Av	CO	–	+	++	++	++	++	++
24	*Nomascus leucogenys*	Gibbon	M	CO	–	+	++	++	++	++	++
25	*Gorilla gorilla*	Gorilla	M	DU	–	+	++	++	++	++	++
26	*Cavia porcellus*	Guinea pig	M	CO	–	+	++	++	++	++	++
27	*Erinaceus europaeus*	Hedgehog	M	CO	–	+	++	++	++	++	++
28	*Procavia capensis*	Hyrax	M	CO	–	+	++	++	++	++	++
29	*Dipodomys ordii*	Kangaroo rat	M	CO	–	+	++	++	++	++	++
30	*Callithrix jacchus*	Marmoset	M	CO	–	+	++	++	++	++	++
31	*Myotis lucifugus*	Microbat	M	CO	–	+	++	++	++	++	++
32	*Microcebus murinus*	Mouse lemur	M	CO	–	+	++	++	++	++	++
33	*Papio anubis*	Olive baboon	M	CO	–	+	++	++	++	++	++
34	*Ailuropoda melanoleuca*	Panda	M	CO	–	+	++	++	++	++	++
35	*Ochotona princeps*	Pika	M	CO	–	+	++	++	++	++	++
36	*Ovis aries*	Sheep	M	CO	–	+	++	++	++	++	++
37	*Ictidomys tridecemlineatus*	Squirrel	M	CO	–	+	++	++	++	++	++
38	*Tarsius syrichta*	Tarsier	M	CO	–	+	++	++	++	++	++
39	*Meleagris gallopavo*	Turkey	Av	IN	–	+	++	++	++	++	++
40	*Chlorocebus sabaeus*	Vervet-AGM	M	CO	–	+	++	++	++	++	++
41	*Echinops telfairi*	Lesser hedgehog tenrec	M	CO	–	+	++	++	++	+	++
42	*Pteropus vampyrus*	Megabat	M	CO	–	+	++	++	+	+	++
43	*Sus scrofa*	Pig	M	DU	–	+	++	++	++	+	++
44	*Rattus norvegicus*	Rat	M	DE	–	+	++	–	++	+	++
45	*Sorex araneus*	Shrew	M	CO	–	+	++	+	++	++	++
46	*Sarcophilus harrisii*	Tasmanian devil	M	DE	–	+	++	++	++	–	++
47	*Macropus eugenii*	Wallaby	M	DE	–	+	++	++	++	–	++
48	*Alligator mississippiensis*	American alligator	R	CO	–	+	+	+	+	+	+
49	*Melopsittacus undulatus*	Budgerigar	Av	CO	–	?	+	+	+	+	+
50	*Cricetulus griseus*	Chinese hamster	M	CO	–	+	+	+	+	+	+
51	*Trichechus manatus*	Manatee	M	CO	–	+	+	+	+	+	+
52	*Geospiza fortis*	Medium ground finch	Av	CO	–	+	+	+	+	+	+
53	*Balaenoptera acutorostrata*	Minke whale	M	OU	–	+	+	+	+	+	+
54	*Heterocephalus glaber*	Naked mole-rat	M	CO	–	+	+	+	+	+	+
55	*Chrysemys picta*	Painted turtle	R	CO	–	+	+	+	+	+	+
56	*Saimiri boliviensis*	Squirrel monkey	M	CO	–	+	+	+	+	+	+
57	*Ceratotherium simum*	White rhinoceros	M	CO	–	?	+	+	+	+	+
58	*Tursiops truncatus*	Dolphin	M	OU/DE	–	+	+	+	–	–	+

Species Latin name: The Latin name of the species we used in this study.
Common name: The common name of the species we used based on Ensembl and UCSC Genome Browser.
Classes: M = Mammalia; R = Retilia; Av = Aves; Am = Amphibia.
Remark: CO = common cluster; IN = miRNA insertion; DE = miRNA deletion; DU = cluster duplication; OU = cluster members could be found outside of the cluster.
miRBase: If the miRNA exists in miRBase, this column is marked by +, otherwise it is marked –. The latest miRBase version 21 is used.
LARP7: If the cluster is located in the intron of gene LARP7, this column will be marked by +. If the gene LARP7 gene is missing or unannotated in this species, this column will be marked by ?.
miR-302b, miR-302c, miR-302a, miR-302d and miR-367: If this miRNA exists in miRBase, this column is marked by +++. If this miRNA is annotated in Ensembl as predicted, this column is marked by ++. If it is predicted to be a true microRNA precursor by miPred, this column is marked by +. If this miRNA is deleted or not present, – is used.

To obtain a detailed view of precursor miRNA divergence, we looked for a human miRNA precursor with SNP data available. mir-367 has 4 SNPs, three are located in the seed region of the 5p mature miRNA and one in the loop region. We aligned the mir-367 precursor with other available sequences (Fig. 2C). Consistent with our previous findings, the SNPs in the seed region were highly conserved, while the SNP in the loop region was highly diverged. In the 3p seed region, we found only *X. tropicalis*, which has a one base difference. Other aligned pre-miRNAs are shown in Supplementary Fig. 3. Based on the alignments, we observed three different patterns relating to loss of miRNAs: 1) 3′ arm of mir-302 is conserved but 5′ arm is lost or not existent, for example, seen in *X. tropicalis* (302b), tasmanian devil (302d), and

wallaby (302d); 2) only loop and seed region in 3′ arm is lost, as seen in rat (302c); 3) totally lost or not existent, for example, seen in *X. tropicalis* (302c), anole lizard (302a, 302c), and dolphin (302a, 302d).

As our analysis indicted different divergence rates depending upon the region studied, we constructed evolutionary trees based upon the entire miR-302/367 cluster, and miR-302/367 precursor sequence only without inter-precursor sequences (Fig. 2D). While species clustered together well according to class for the miRNA precursor regions only, mammals were separated when clustering the entire sequence. The marsupials (opossum, wallaby, and tasmanian devil) and platypus are clustered in a clade with aves and reptilia.

We also studied the cluster's length difference among various species. In Supplementary Fig. 4, we show that the scale of the cluster tends to be steady in mammalian clade, especially among primates.

Based on our datasets, *X. tropicalis* was the outgroup in our phylogenetic analysis. We observed that the cluster was evolutionarily stable, especially in primates, with some deletion events of certain miRNAs in specific taxonomic groups. To trace a possible earlier origin for the cluster, we followed methods described by Mohammed et al. (2014). We broadened the concept of miRNA family and searched for mature miRNAs which have exactly the same seed sequence and are also homologous in other regions among all the species in miRBase v21. We found a large number of homologous mature miRNA of miR-302 and a small number of miR-367 sequences. Interestingly, these include highly homologous miR-367 mature miRNA in *Ciona intestinalis* (cin-miR-367) and *Saccoglossus kowalevskii* (sko-miR-4818a/b/c/d/e/f/g) (Supplementary Fig. 5).

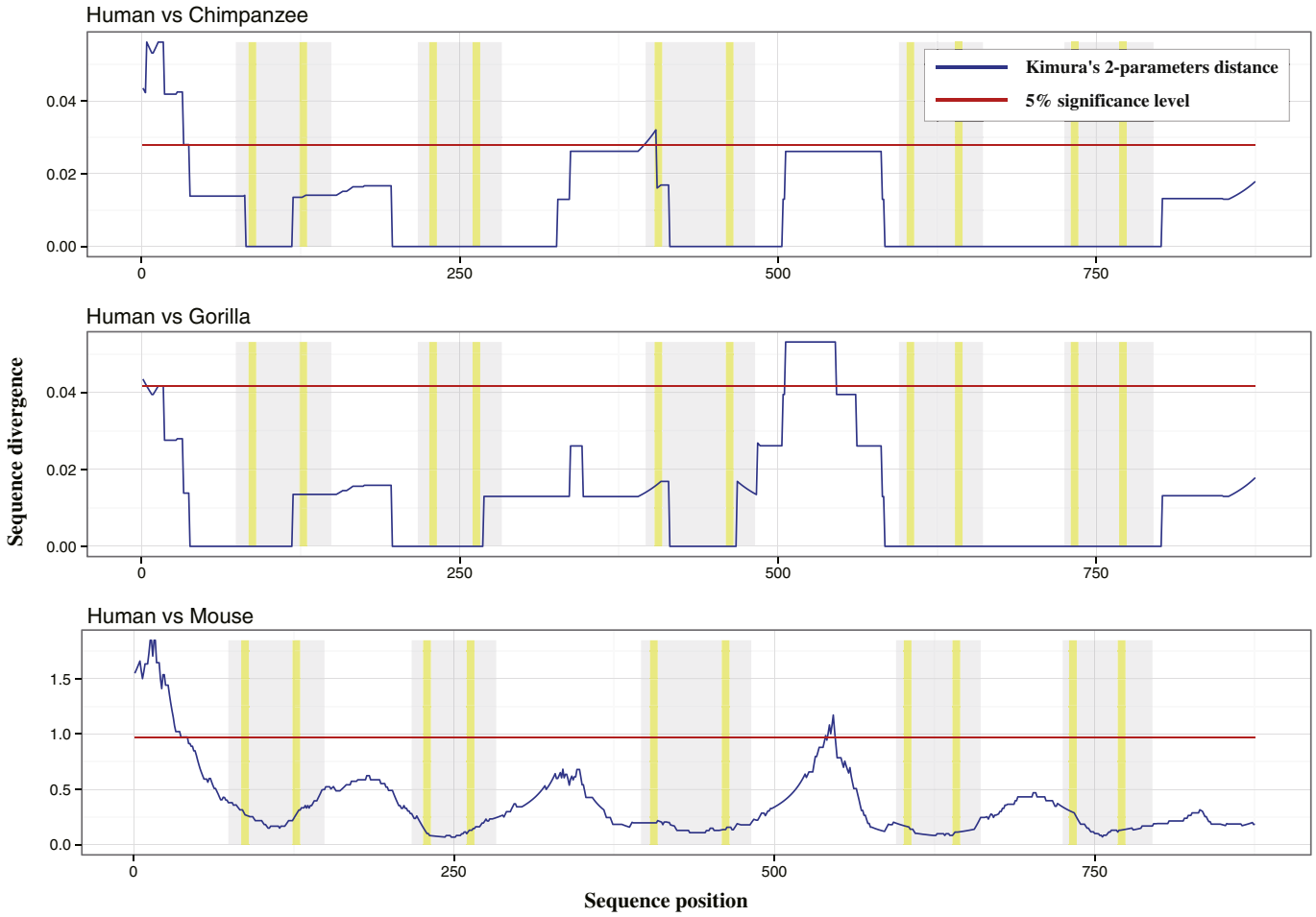

Fig. 2. A. Sequence divergence plots. Human miR-302/367 cluster sequence was compared to homologous sequences from gorilla, chimpanzee, and mouse, and Kimura's 2-parameter distance was plotted as described in Methods. A 77 nt long sliding window analysis reveals that pre-miRNAs in the cluster are more conserved than inter-precursor sequence regions. Similar plots were produced comparing other species (Supplementary Fig. 1). Gray bands represent pre-miRNA and yellow bands represent the seed region. X axis is the sequence position in the alignment, y axis is the Kimura's 2-parameter distance. Kimura's 2-parameter distance was calculated by function dist.dna() in the R package ape version 3.2 with parameter model equal to K80. The plot was created by ggplot2 (Wickham and Wickham, 2009). B. Secondary structure distance heat map. From left to right, each column represents the pre-miRNA and the inter-precursor sequence between pre-miRNAs. Shown are 57 rows representing the species studied ordered by taxonomic class. Classes are distinguished by color and a white frame highlights the primates. We first predicted the 2D structure for each region sequence (pre-miRNA region and between region) across all 58 species (human included). Then we compared human to the other 57 species and calculated the 2D structure distance as described in Methods. 2D structures and distances between structures were predicted by ViennaRNA-2.1.7. The distance matrix was normalized and the Heat map was created in Matlab. C. mir-367 multi-alignment across 58 species. Gray bands represent mature miRNA and yellow represents the seed region. Black frames indicate the location of SNPs in human. Different taxonomic classes are distinguished by color, white frame highlights primates. Consensus in represented at the bottom by "*" symbol. Alignment of other pre-miRNAs are shown in Supplementary Fig. 3. From left to right the SNPs are rs373558217, rs369906379, rs150161032 (MAF = 0.000459137), rs376629554, and rs369489772, respectively. WebLogo 3.4 (http://weblogo.berkeley.edu/) was used to create the sequence logos (Crooks et al., 2004). D. Sequence and evolutionary trees. The beginning of mir-302b to the end of mir-367 without any extension was chosen to define the miR-302/367 cluster. This region is 684 bp long. From the cluster region, we joined the head to tail of each of the five aligned pre-miRNA sequences, called the Joint pre-miRNA sequence, as shown in the figure. Evolutionary trees were built using the whole cluster (top), and the Joint pre-miRNA sequence (bottom). The lateral sector heat map, from the inside to outside, respectively, denotes the reliability that the sequence in the cluster is a true pre-miRNA. Based on Table 1, we defined that if the pre-miRNA is recorded in miRBase, the reliability is 1, and for pre-miRNA predicted by Ensembl and miPred, reliabilities are 0.9 and 0.8, respectively. If the pre-miRNA is lost in the cluster the reliability is 0. The trees were built using MEGA 6 and illustrated by iTOL as described in the Methods.

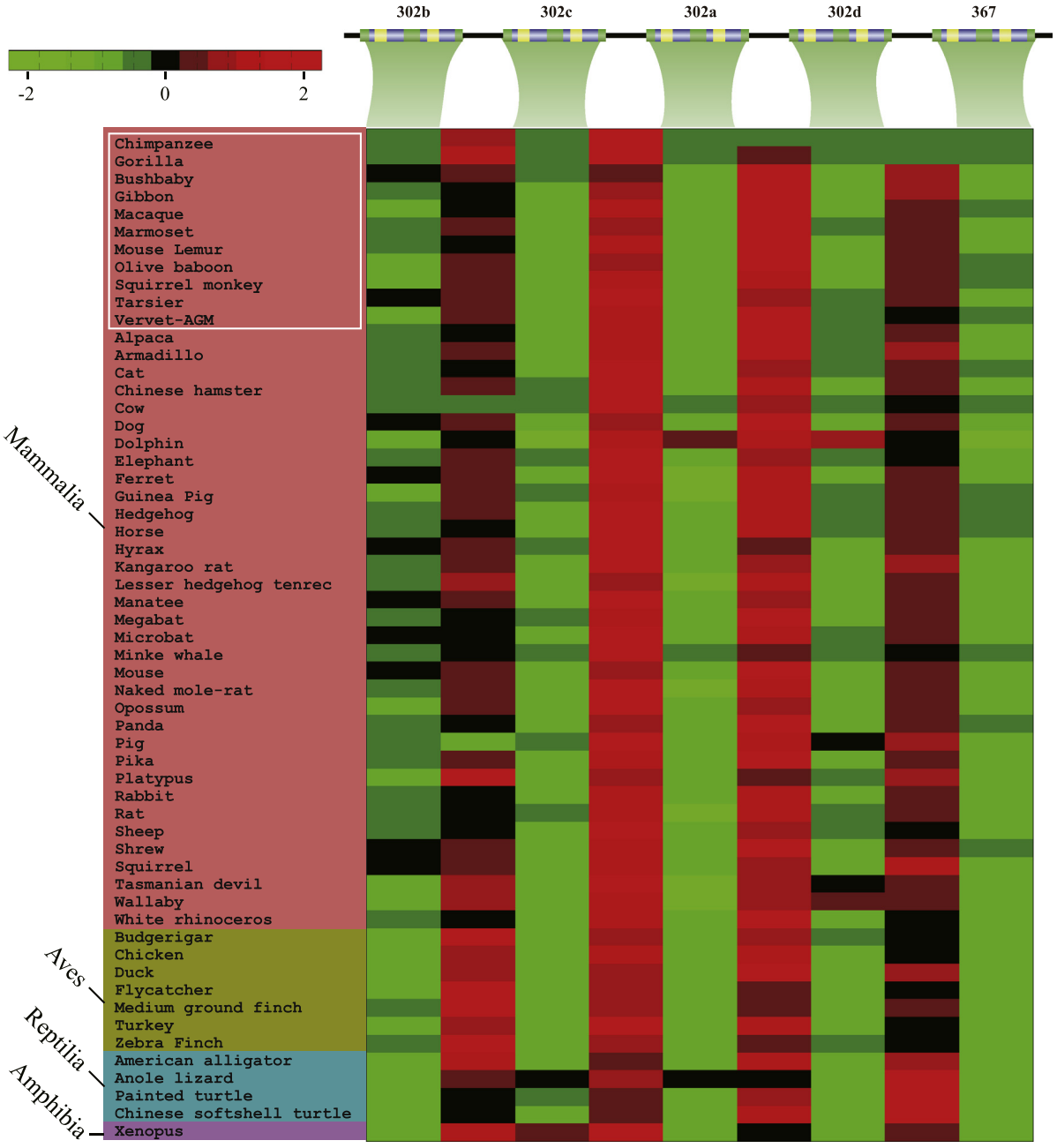

Fig. 2 (*continued*).

3.2. Targets analysis of miR-302/367 cluster

We downloaded target genes of the miR-302/367 cluster (both predicted targets and those validated by biological experiments) from public databases. miRTarBase 4.5, Tarbase 7.0 and miRecords 4 were used for validated targets, and they yielded a total of 412 targets for miR-302a/b/c/d-3p and 94 targets for miR-367-3p. The 5′ arm products miR-302a-5p, miR-302c-5p, and miR-302d-5p had 73, 40 and 50 validated targets, respectively. The TargetScan algorithm, which seeks out conserved 7mer and 8mer sites that match the seed region of each miRNA, predicted 893 and 844 conserved target genes for miR-367-3p and miR-302a/b/c/d-3p, respectively. Target genes predicted by miRanda are based on three properties: sequence complementarity, free energies of RNA-RNA duplexes, and conservation of target sites. Under the criterion integrating mirSVR score and conserved character, 2843 and 3283 target genes for miR-367-3p and miR-302a/b/c/d-3p, were observed, respectively. It should be noted that miR-302a/b/c/d-3p have different target gene sets in miRanda and the intersection set size is 3283, but miR-302a/b/c/d-3p targets the same 844 by the prediction method of TargetScan.

In order to obtain more credible predicted target genes and fewer false positives, we further strengthened our selection by reducing the candidate target gene set size based on: 1) top 200 target genes with higher reliability from TargetScan and miRanda which rank target genes according to a specific strategy; 2) target genes appearing in both databases. Genes that satisfied either 1) or 2) could be reserved, and finally we obtained 885 target genes for miR-367-3p and 1181 target genes for miR-302a/b/c/d-3p via adding experimental validated genes. After a union of the validated targets of miR-302-5p, there were 157 targets for miR-302a/c/d-5p, although they did not share the

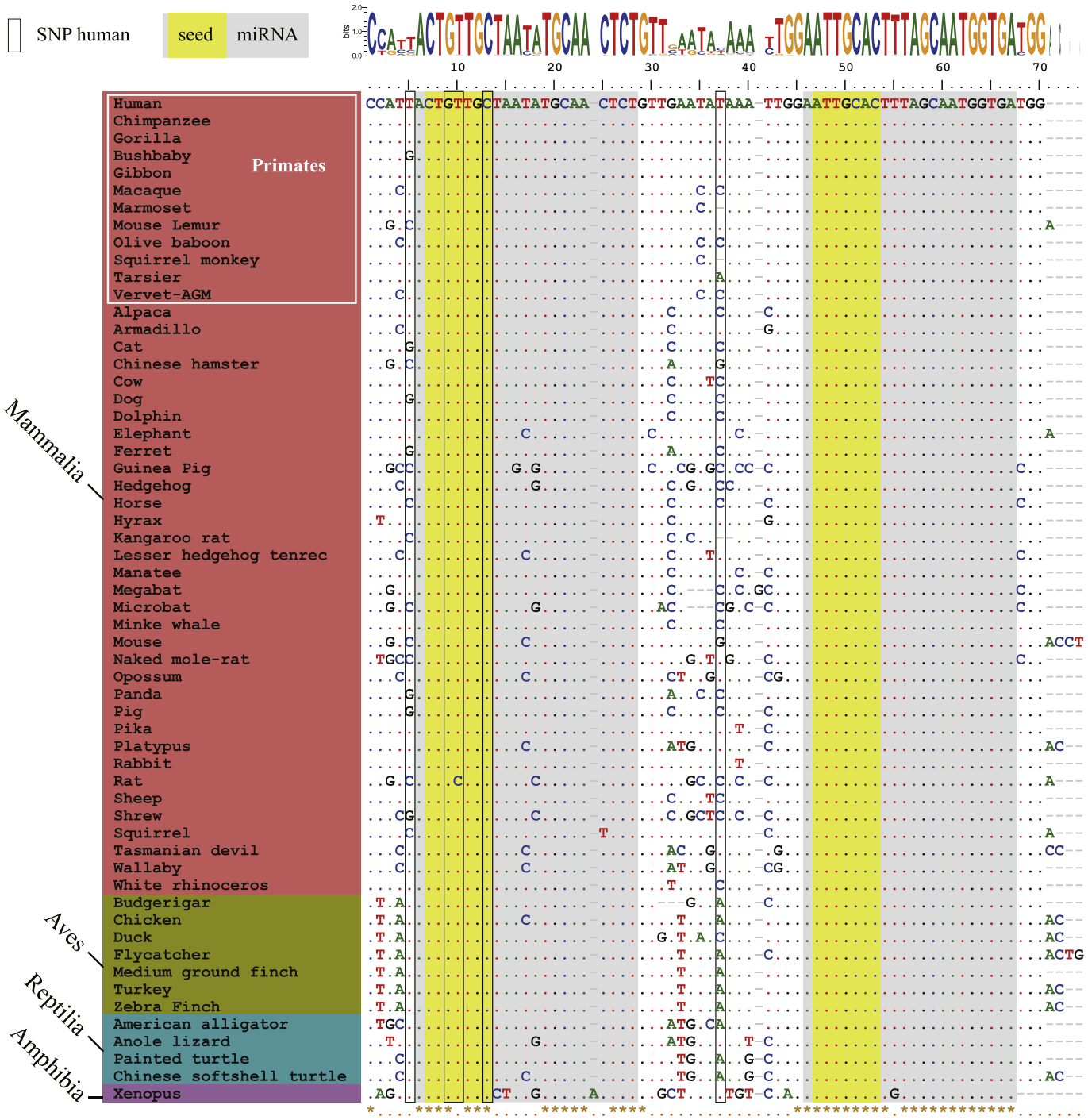

Fig. 2 (*continued*).

same seed. Cluster function analysis by GO and pathway enrichment was applied on the intersection sets, different sets and union sets of miR-367-3p, miR-302a/b/c/d-3p and miR-302a/c/d-5p target genes (Fig. 3).

Both miR-302a/b/c/d-3p and miR-367-3p unique target sets were enriched for regulation of macromolecule metabolic process (GO:0060255). Using the GO molecular function ontology, miR-302-3p targets were enriched for DNA binding (GO:0003677), miR-367-3p targets were enriched for protein binding (GO:0005515) while miR-302-5p targets were enriched for RNA binding (GO:0003723). The union of targets indicated protein binding (GO:0005515), binding (GO:0005488), and transcription regulator activity (GO:0030528). KEGG enrichment revealed that cluster targets play a role in cell cycle and cancer. Overall, the enriched target categories show enriched transcription regulation activity suggesting a coordinated function of their targets. The full list of enriched GO and KEGG categories are shown in Supplementary Table 3.

3.3. Regulation of miR-302/367 cluster

Because the miR-302/367 cluster is transcribed as a single transcriptional unit, we wanted to gain insight into its regulation. Previous studies have delineated the core promoter region and important transcription factors in the 5′ flanking region (Barroso-delJesus et al., 2008; Marson et al., 2008). We were also able to observe the binding

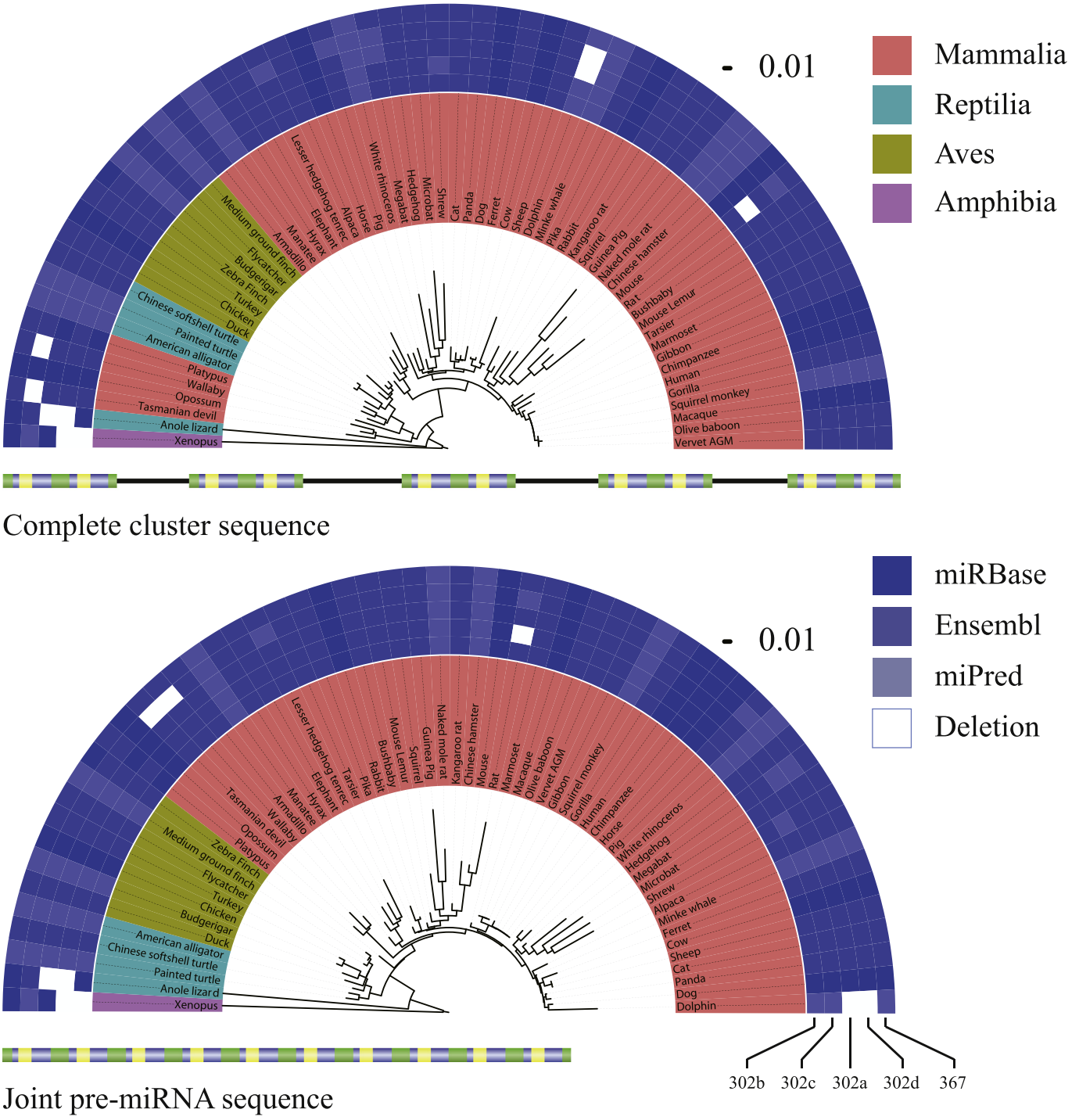

Fig. 2 (*continued*).

of NANOG and OCT-4 to the 5′ flanking region from public Chip-seq data confirming earlier studies that used classical molecular biology methods (Supplementary Fig. 6). We also hypothesized that other forms of transcriptional control could be used. We analyzed whether the miR-302/367 cluster itself could be under control of miRNAs or function as a miRNA sponge and regulate other RNA transcripts by competing for shared microRNAs. We observed hundreds of miRNAs that could target the cluster and region both 5′ and 3′ around the cluster. For example, in a putative pri-miRNA region, defined as from TSS start to the end in Fig. 4, we could find that the 3′ downstream and cluster region were enriched in target sites for miRNAs. No miRNAs were detected 5′ upstream of the pri-miRNA. To find many target sites near the TSS start is not surprising since the promotor regions are highly conserved across the 58 species. In addition, 14 miRNAs belonging to two families (family 130 ac/301ab/301b/301b-3p/454/721/4295/3666, and 17/17-5p/20ab/20b-5p/93/106ab/427/518a-3p/519d, which are defined by TargetScan base on seed), were found that targeted the end of the cluster (Fig. 4).

We then analyzed the expression levels of the individual miRNAs within the cluster. All miRNAs in both the miR-302 family and miR-367 had abundant counts from small RNAs isolated from cells. However, some members of the miR-302 family, miR-302b for example, have specific isomiRs expressed in nucleus or cytosol (Fig. 5). IsomiRs were abundantly expressed throughout the family as well, with the 3p

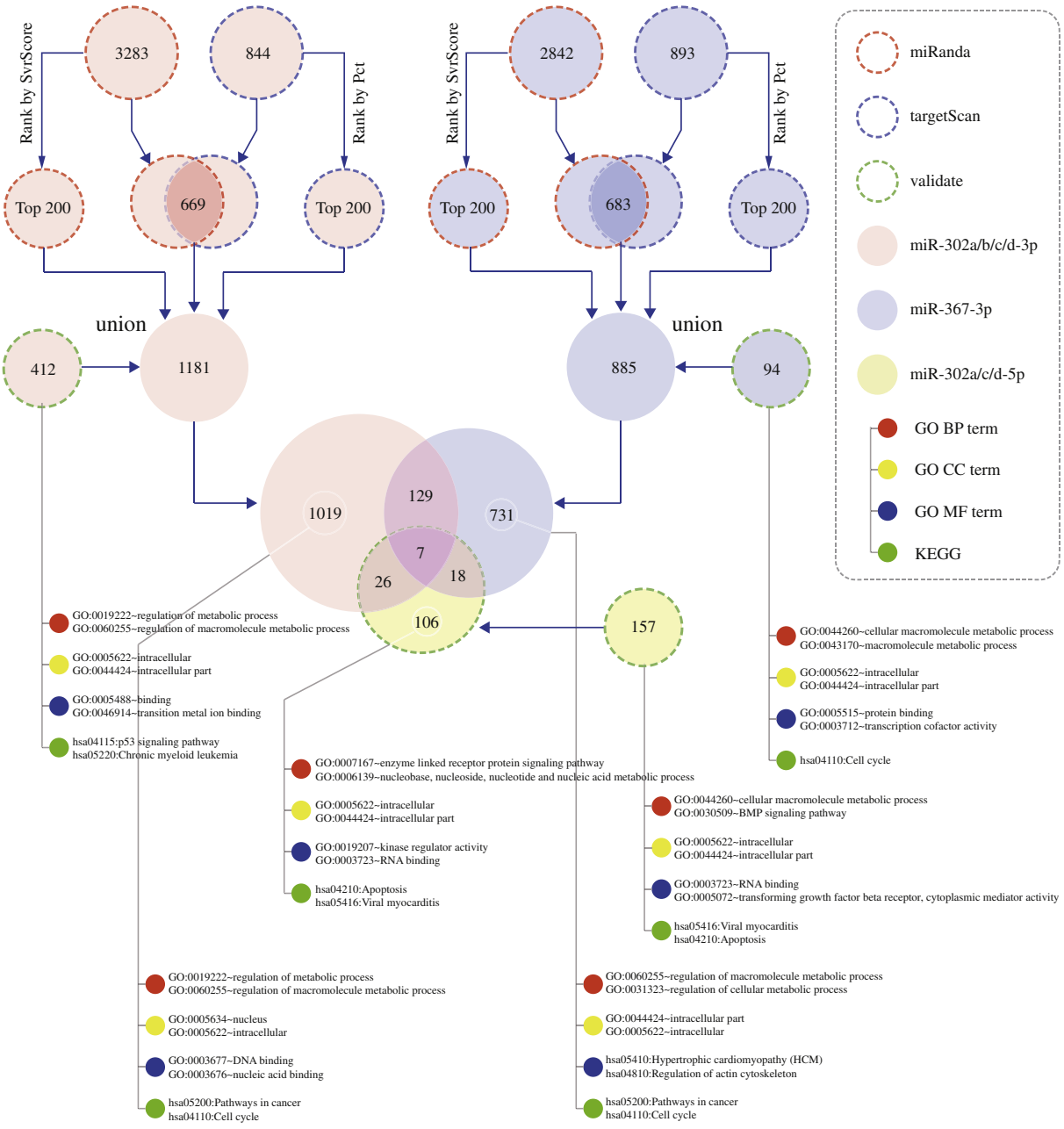

Fig. 3. Workflow of miR-302/367 function analysis. Red dashed circles refer to the targets predicted by miRanda, blue dashed circles represent the targets according to the prediction of TargetScan, and green dashed circle is the target set validated by biological experiment. The circles filled with pink represents the target set of miR-302-3p, orange for miR-302-5p, while the blue circles are the target sets of miR-367. David was used for enrichment analysis. Only the top two significant GO terms and pathways are listed in the figure (detail shown in Supplementary Table 3). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

being the dominant form, with the exception of miR-302a, which abundantly utilizes both 5p and 3p mature miRNA arms (Fig. 5). The most dominant isomiRs and a comprehensive list of isomiRs can be found in Supplementary Fig. 7.

3.4. miR-302/367 cluster targets as cancer biomarkers

Previous studies have suggested that the miR-302/367 cluster has a role in cancer. By analyzing available miRNA-seq data from cancer tissues we were able to confirm the presence or absence of mature miRNAs from this cluster. We found the ESC specific miR-302/367 cluster is preferentially expressed in tumor samples of LGG, TGCT and UVM (Supplementary Fig. 8). The cluster also tends to be expressed (but, expression is very low) in the tissues adjacent to cancer (here, called normal sample) in KIRC, KIRP and KICH. A good example was from a data set that contained Kidney renal clear cell carcinoma tumor tissue and adjacent normal tissue from 71 patients. These tissues showed clear expression of miR-302 family members in normal tissue, but few instances of expression in tumor tissues from these patients (Fig. 6A). We were not able to detect the presence of miR-367 in either normal or tumor tissues. Taking the top 50 significantly regulated validated targets provided us with 13 down-regulated and 37 up-regulated genes in normal tissues. For miR-302-5p, the top ten significantly expressed validated target genes are shown in the heat map in Fig. 6. Clustering of these genes demonstrated that 68 of the 71 tissues could be classified as tumor or normal. A t-test indicated that these top 50 validated targets were significantly regulated in the normal versus tumor samples. The same analysis for KIRP and KICH is shown in Supplementary Fig. 9.

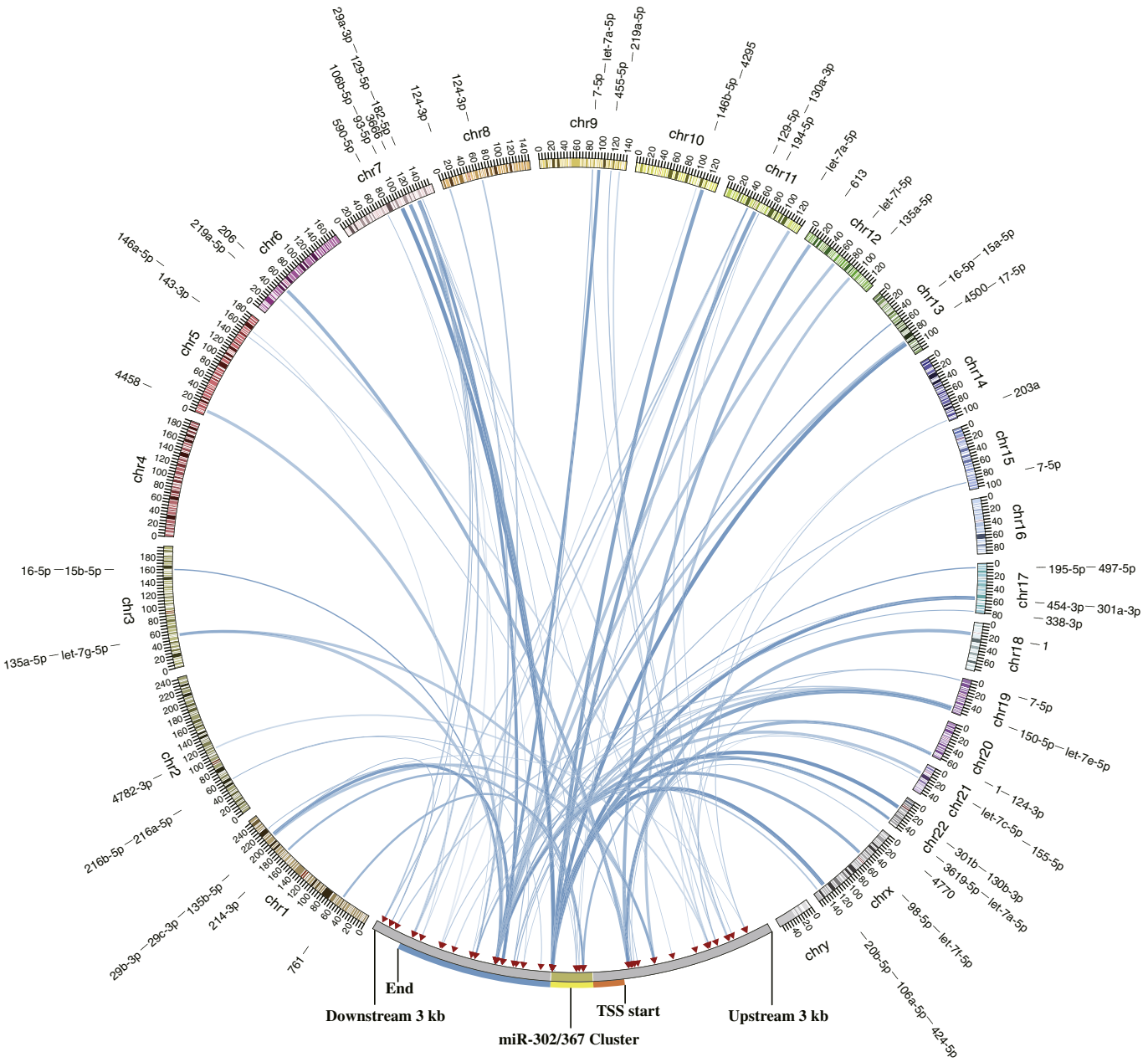

Fig. 4. miRNA target sites in flanking regions and cluster site of miR-302/367. miRNA target sites in the cluster region and its upstream 3 kb, and downstream 3 kb are shown and filled with gray. The target sites are predicted by TargetScan. The upstream 0.5 kb, cluster transcript and downstream 2.5 kb regions are highlighted with orange, yellow, and blue, respectively. Interaction of miRNA and cluster transcript is indicated with a blue line and ends with a red arrow. The thickness and transparency of the line follows the Pct value and context score, respectively. The direction of arrow is from miRNA to target sites. The chosen coordinate of TSS start and end is based on included references (Barroso-delJesus et al., 2008; Marson et al., 2008). The primary coordinates used for the miRNA-302/367 cluster was chr4:113569030-113569713 as indicated in yellow. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

4. Discussion

Previous studies have suggested that miRNA clusters evolve by tandem duplication, construction by putting together miRNAs, construction by left together miRNAs, cluster fissioning followed by mRNA acquisition, or by formation of new hairpins (Marco et al., 2013; Mohammed et al., 2014). We observed an emerging mir-302b in the cluster in *X. tropicalis*. We found few instances of mir-367 present in other species and located in other loci, so our data support a tandem duplication model. This is further supported by analysis of the function of the miR-302 family and miR-367, based on validated and predicted target analysis. Functions of both the miR-302 family and miR-367 appear to be similar. We did uncover support for *de novo* hairpin formation within the cluster as well, as a complete duplication of the cluster in a primate species.

To address the question of whether the cluster co-evolves within the mRNA where it is located, even though they are expressed on different strands, we analyzed the coincident presence of LARP7 and miR-302/367. We found 71 species (including fish) with the LARP7 gene present and among these, 56 species also had the miR-302/367 cluster located within the intron. In only a few species, such as budgerigar (*Melopsittacus undulatus*) and white rhinoceros (*Ceratotherium simum*), we found the cluster but not the LARP7 gene. The remaining species that had LARP7 but not the miR-302/367 cluster belonged to the classes: *Actinopterygii*, *Hyperoartia* and *Sarcopterygii*. Thirteen of these species are verified

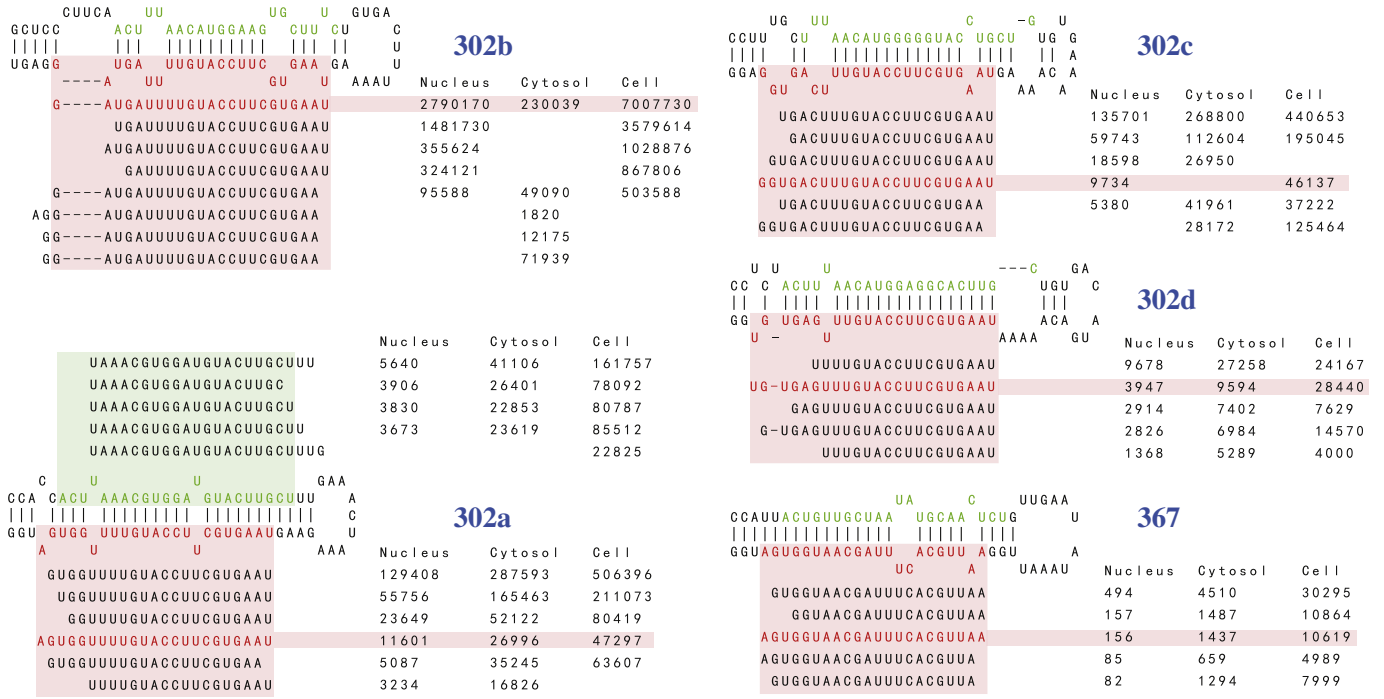

Fig. 5. isomiRs of the miR-302/367 cluster and their relative expression levels in hESC. The canonical miRNA sequence and reads are marked with red. Beside each figure, a table shows the read count number of those miRNA in hESCs from different samples (nucleus, cytosol and cell). Here, we only list the dominant expression isomiRs, and only mir-302a could detect high expression 5′ arm products. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

by BLAT (no similar sequence found). They are tetraodon (*Tetraodon nigroviridis*), amazon molly (*Poecilia formosa*), cave fish (*Astyanax mexicanus*), stickleback (*Gasterosteus aculeatus*), spotted gar (*Lepisosteus oculatus*), platyfish (*Xiphophorus maculatus*), medaka (*Oryzias latipes*), atlantic cod (*Gadus morhua*), Nile tilapia (*Oreochromis niloticus*), zebrafish (*Danio rerio*), fugu (*Takifugu rubripes*), lamprey (*Petromyzon marinus*), coelacanth (*Latimeria chalumnae*). Such a discordance would not support the hypothesis that the LARP7 gene and miRNA cluster are co-evolving.

Next, we looked at conservation of different structural regions of the miRNA cluster. miR-367 is highly conserved across species. miR-302a/b/c/d remains conserved in the seed region, but appears to be more highly variable in inter-precursor sequence regions. These more diverged inter-precursor sequence regions could be used to generate new miRNAs or to harbor sequence variation. In addition, SNPs between miR-302d and miR-367 were also found to be variable in other species, and we found another SNP rs13136737 with a high minor allele frequency (MAF = 0.391644). The context is palindromic (GCAA TTGCGTTAACG, SNP is underlined). The extent to which this site has functional significance is currently under study by other groups and may provide a further tool to view cluster divergence.

We observed another potential source of novel miRNAs that could allow novel functions via expression of different arms and isomiRs. The 3′ arm products are the dominant expression form in this cluster. While 5′ products were expressed, they had lower levels of expression and also had very diverse isomiRs. This type of miRNA isomiRs have been found in a wide range of human tissues, including embryonic stem cells (Neilsen et al., 2012; Tan et al., 2014), and may have functional consequences. Since we performed functional analysis for every miRNA form of the miR-302/367 cluster, we were able to find similar functions of the miR-302 family and miR-367 despite the differences in their seed sequences. This supports the notion that the miR-302/367 cluster evolved to perform specific related functions. Thus, the individual miRNAs from the cluster can perform coordinated gene regulatory events during specific biological processes.

Other studies have shown a role for the miR-302/367 cluster in cancer (Murray et al., 2011; Khalili et al., 2012; Cai et al., 2013; Jamshidi-Adegani et al., 2014). Our search through the TCGA cancer data sets revealed only low expression in most cancer tissues, as well as in normal adjacent tissues. A notable exception was in the KIRC set. The results suggest that miR-302/367 might have a role in only a limited number of cancers. High levels of expression in adjacent normal versus tumor tissues in patients were observed. When gene expression levels of validated targets were evaluated, only a limited subset was found to have the profile (down regulated in the presence of a miR-302 family member). This suggests that miR-302 might target only a subset of known validated targets, which depends upon the tissue studied and the physiologic state. Moreover, a highly significant concordance between the lack of miR-302 family members and up-regulation of validated target genes suggests the identity of specific genes regulated by miR-302 in KIRC. These genes or miR-302 itself might therefore be useful as biomarkers for kidney tumors.

5. Conclusions

The analysis presented here shows a highly conserved but still evolving miRNA cluster. Tandem duplication and *de novo* construction of new miRNAs within the cluster allows continued evolution of miR-302/367. As expected of a polycistronic unit with a common function, tight regulation of its expression appears to be maintained via transcriptional control. The correlated expression of miR-302/367 and its targets together with the inverse correlation in healthy and adjacent tumor tissue suggests a role for this miRNA in kidney renal cell carcinomas. As more sequence data becomes available, the precise targets of miR-302/367 in specific contexts and therefore its precise function(s) will be better understood.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.cbd.2015.08.002.

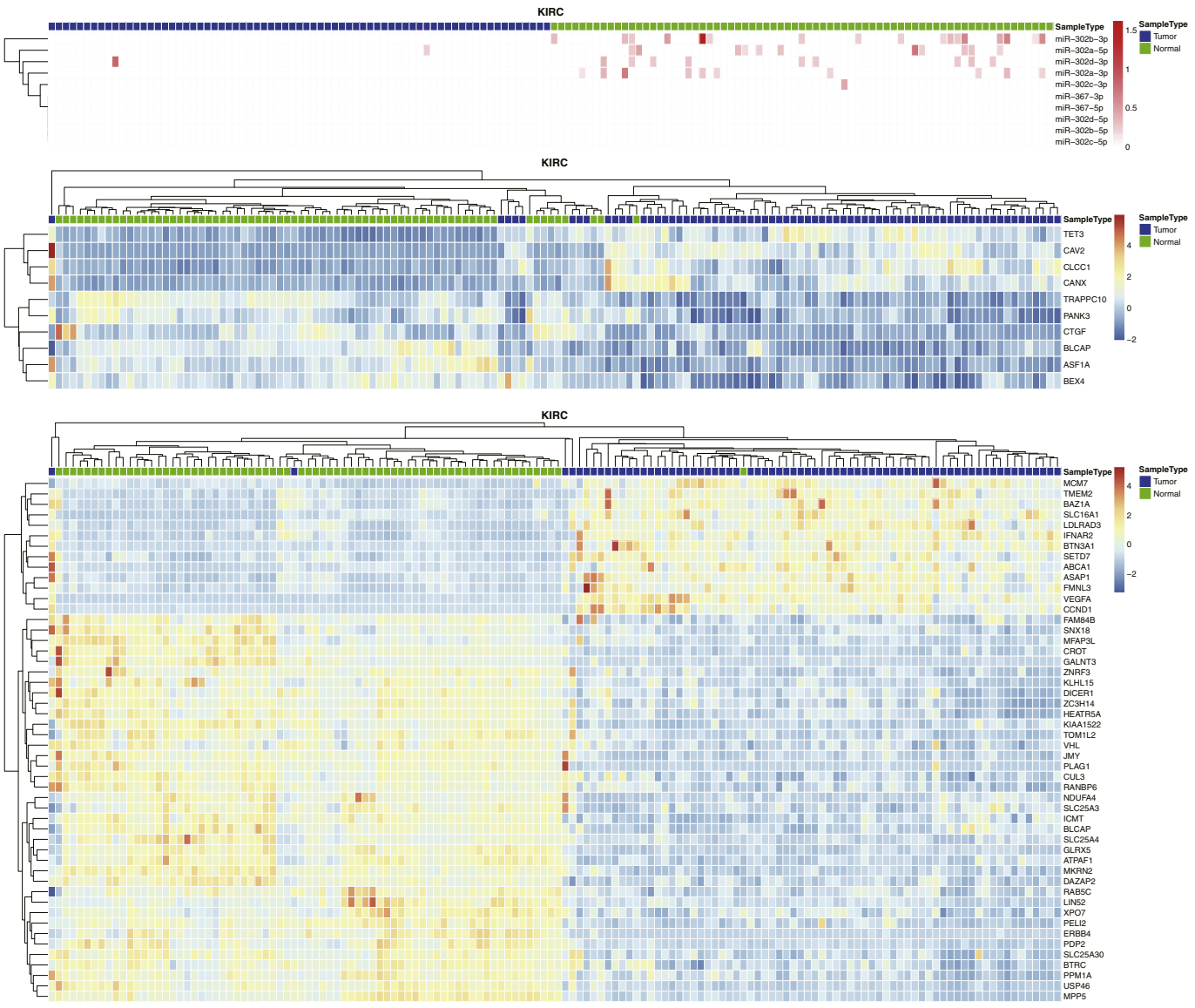

Fig. 6. Expression data of the miR-302/367 cluster and its' targets in the KIRC dataset. We obtained matched miRNA-seq and RNA-seq data for this cancer from TCGA with data from tumor tissue and adjacent normal tissue from the same 71 patients. The first heat map shows the expression of the miRNA in cluster, red indicates expression and white no expression. Rows are clustered while columns are not, and columns are separated by normal (green) and tumor (blue) samples, with the samples in the same order in each area. The second heat map shows the expression of the top 10 differentially expressed validated target genes of miR-302a-5p. The third heat map shows the top 50 differentially expressed validated target genes of miR-302a/b/c/d-3p. The target genes are selected from validated target genes and the two sets of expression values are significantly different in tumor and normal tissues (ranked by t-test p-value). The heat maps were generated using the R package pheatmap v 1.0.2. Supplementary Fig. 8 displays the same figures for analysis of KICH and KIRP. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Abbreviations

ENCODE encyclopedia of DNA elements
GO gene ontology
hESC human embryonic stem cells
iPSC induced pluripotent stem cells
KIRC kidney renal clear cell carcinoma
KICH kidney chromophobe
KIRP kidney renal papillary cell carcinoma
LARP7 la ribonucleoprotein domain family, member 7
MAF minor allele frequency
miRNA microRNA
NANOG nanog homeobox
OCT-4 (POU5F1) POU class 5 homeobox 1
Pre-miRNA Precursor microRNA
Pri-miRNA primary microRNA
REX1 (ZFP42) ZFP42 zinc finger protein
RISC RNA silencing complex
RPM reads per million
SNP single nucleotide polymorphism
SOX2 SRY (sex determining region Y)-box 2
TCGA the cancer genome atlas

Role of funding sources

The funders had no role in study design, data analysis, or preparation of the manuscript.

Acknowledgments

The authors wish to thank the 'Bridging the Gap' and 'Swap and Transfer' Erasmus Mundus project (Reference 545648) supported by the European Union for making this study possible. This work was

partially supported by the Academy of Finland (Project 62340), the National Natural Science Foundation of China (no. 61272207, 61472158), the Science-Technology Development Project from Jilin Province (20120730), and University of Macau Faculty of Health Sciences (MYRG2015-00231-FHS).

References

Barroso-delJesus, A., Romero-Lopez, C., Lucena-Aguilar, G., Melen, G.J., Sanchez, L., Ligero, G., Berzal-Herranz, A., Menendez, P., 2008. Embryonic stem cell-specific miR302-367 cluster: human gene structure and functional characterization of its core promoter. Mol. Cell. Biol. 28, 6609–6619.

Bartel, D.P., 2004. MicroRNAs: genomics, biogenesis, mechanism, and function. Cell 116, 281–297.

Betel, D., Wilson, M., Gabow, A., Marks, D.S., Sander, C., 2008. The microRNA.org resource: targets and expression. Nucleic Acids Res. 36, D149–D153.

Betel, D., Koppal, A., Agius, P., Sander, C., Leslie, C., 2010. Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. Genome Biol. 11, R90.

Cai, N., Wang, Y.D., Zheng, P.S., 2013. The microRNA-302-367 cluster suppresses the proliferation of cervical carcinoma cells through the novel target AKT1. RNA (New York, N.Y.) 19, 85–95.

Callow, M.J., Dudoit, S., Gong, E.L., Speed, T.P., Rubin, E.M., 2000. Microarray expression profiling identifies genes with altered expression in HDL-deficient mice. Genome Res. 10, 2022–2029.

Crooks, G.E., Hon, G., Chandonia, J.M., Brenner, S.E., 2004. WebLogo: a sequence logo generator. Genome Res. 14, 1188–1190.

Cunningham, F., Amode, M.R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S., Gil, L., Girón, C.G., Gordon, L., Hourlier, T., Hunt, S.E., Janacek, S.H., Johnson, N., Juettemann, T., Kähäri, A.K., Keenan, S., Martin, F.J., Maurel, T., McLaren, W., Murphy, D.N., Nag, R., Overduin, B., Parker, A., Patricio, M., Perry, E., Pignatelli, M., Riat, H.S., Sheppard, D., Taylor, K., Thormann, A., Vullo, A., Wilder, S.P., Zadissa, A., Aken, B.L., Birney, E., Harrow, J., Kinsella, R., Muffato, M., Ruffier, M., Searle, S.M.J., Spudich, G., Trevanion, S.J., Yates, A., Zerbino, D.R., Flicek, P., 2015. Ensembl 2015. Nucleic Acids Res. 43, D662–D669.

Durinck, S., Spellman, P.T., Birney, E., Huber, W., 2009. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nat. Protoc. 4, 1184–1191.

ENCODE Project Consortium, 2012. An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74.

Enright, A.J., John, B., Gaul, U., Tuschl, T., Sander, C., Marks, D.S., 2003. MicroRNA targets in *Drosophila*. Genome Biol. 5, R1.

Friedman, R.C., Farh, K.K., Burge, C.B., Bartel, D.P., 2009. Most mammalian mRNAs are conserved targets of microRNAs. Genome Res. 19, 92–105.

Glazov, E.A., Cottee, P.A., Barris, W.C., Moore, R.J., Dalrymple, B.P., Tizard, M.L., 2008. A microRNA catalog of the developing chicken embryo identified by a deep sequencing approach. Genome Res. 18, 957–964.

Grimson, A., Farh, K.K., Johnston, W.K., Garrett-Engele, P., Lim, L.P., Bartel, D.P., 2007. MicroRNA targeting specificity in mammals: determinants beyond seed pairing. Mol. Cell 27, 91–105.

Ha, M., Kim, V.N., 2014. Regulation of microRNA biogenesis. Nat. Rev. Mol. Cell Biol. 15, 509–524.

Hsu, S.D., Tseng, Y.T., Shrestha, S., Lin, Y.L., Khaleel, A., Chou, C.H., Chu, C.F., Huang, H.Y., Lin, C.M., Ho, S.Y., Jian, T.Y., Lin, F.M., Chang, T.H., Weng, S.L., Liao, K.W., Liao, I.E., Liu, C.C., Huang, H.D., 2014. miRTarBase update 2014: an information resource for experimentally validated miRNA-target interactions. Nucleic Acids Res. 42, D78–D85.

Huang da, W., Sherman, B.T., Lempicki, R.A., 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protoc. 4, 44–57.

Jacob, F., Monad, J., 1961. Genetic regulatory mechanisms in the synthesis of proteins. J. Mol. Biol. 3, 318–356.

Jamshidi-Adegani, F., Langroudi, L., Shafiee, A., Mohammadi-Sangcheshmeh, A., Ardeshirylajimi, A., Barzegar, M., Azadmanesh, K., Naderi, M., Arefian, E., Soleimani, M., 2014. Mir-302 cluster exhibits tumor suppressor properties on human unrestricted somatic stem cells. Tumour Biol. 35, 6657–6664.

Jiang, P., Wu, H., Wang, W., Ma, W., Sun, X., Lu, Z., 2007. MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features. Nucleic Acids Res. 35, W339–W344.

Kallio, M.A., Tuimala, J.T., Hupponen, T., Klemela, P., Gentile, M., Scheinin, I., Koski, M., Kaki, J., Korpelainen, E.I., 2011. Chipster: user-friendly analysis software for microarray and other high-throughput data. BMC Genomics 12, 507.

Kent, W.J., 2002. BLAT—the BLAST-like alignment tool. Genome Res. 12, 656–664.

Khalili, M., Sadeghizadeh, M., Ghorbanian, K., Malekzadeh, R., Vasei, M., Mowla, S.J., 2012. Down-regulation of miR-302b, an ESC-specific microRNA, in gastric adenocarcinoma. Cell J. 13, 251–258.

Kim, V.N., Nam, J.W., 2006. Genomics of microRNA. Trends Genet. 22, 165–173.

Kim, J.Y., Shin, K.K., Lee, A.L., Kim, Y.S., Park, H.J., Park, Y.K., Bae, Y.C., Jung, J.S., 2014. MicroRNA-302 induces proliferation and inhibits oxidant-induced cell death in human adipose tissue-derived mesenchymal stem cells. Cell Death Dis. 5, e1385.

Kimura, M., 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. 16, 111–120.

Kozomara, A., Griffiths-Jones, S., 2014. miRBase: annotating high confidence microRNAs using deep sequencing data. Nucleic Acids Res. 42, D68–D73.

Krol, J., Loedige, I., Filipowicz, W., 2010. The widespread regulation of microRNA biogenesis, function and decay. Nat. Rev. Genet. 11, 597–610.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J., Marra, M.A., 2009. Circos: an information aesthetic for comparative genomics. Genome Res. 19, 1639–1645.

Lagos-Quintana, M., Rauhut, R., Lendeckel, W., Tuschl, T., 2001. Identification of novel genes coding for small expressed RNAs. Science 294, 853–858.

Lakshmipathy, U., Davila, J., Hart, R.P., 2010. miRNA in pluripotent stem cells. Regen. Med. 5, 545–555.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., Thompson, J.D., Gibson, T.J., Higgins, D.G., 2007. Clustal W and Clustal X version 2.0. Bioinformatics 23, 2947–2948.

Lau, N.C., Lim, L.P., Weinstein, E.G., Bartel, D.P., 2001. An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans*. Science 294, 858–862.

Laurent, L.C., 2008. MicroRNAs in embryonic stem cells and early embryonic development. J. Cell. Mol. Med. 12, 2181–2188.

Lee, R.C., Ambros, V., 2001. An extensive class of small RNAs in *Caenorhabditis elegans*. Science 294, 862–864.

Letunic, I., Bork, P., 2007. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics 23, 127–128.

Lewis, B.P., Burge, C.B., Bartel, D.P., 2005. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell 120, 15–20.

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

Li, J., Liu, Y., Dong, D., Zhang, Z., 2010. Evolution of an X-linked primate-specific micro RNA cluster. Mol. Biol. Evol. 27, 671–683.

Li, Y., Masaki, T., Yamane, D., McGivern, D.R., Lemon, S.M., 2013. Competing and noncompeting activities of miR-122 and the 5′ exonuclease Xrn1 in regulation of hepatitis C virus replication. Proc. Natl. Acad. Sci. U. S. A. 110, 1881–1886.

Liao, B., Bao, X., Liu, L., Feng, S., Zovoilis, A., Liu, W., Xue, Y., Cai, J., Guo, X., Qin, B., Zhang, R., Wu, J., Lai, L., Teng, M., Niu, L., Zhang, B., Esteban, M.A., Pei, D., 2011. MicroRNA cluster 302–367 enhances somatic cell reprogramming by accelerating a mesenchymal-to-epithelial transition. J. Biol. Chem. 286, 17359–17364.

Lipchina, I., Elkabetz, Y., Hafner, M., Sheridan, R., Mihailovic, A., Tuschl, T., Sander, C., Studer, L., Betel, D., 2011. Genome-wide identification of microRNA targets in human ES cells reveals a role for miR-302 in modulating BMP response. Genes Dev. 25, 2173–2186.

Lorenz, R., Bernhart, S.H., Honer Zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P.F., Hofacker, I.L., 2011. ViennaRNA Package 2.0. Algorithms for molecular biology. AMB 6, 26.

Marco, A., Ninova, M., Ronshaugen, M., Griffiths-Jones, S., 2013. Clusters of microRNAs emerge by new hairpins in existing transcripts. Nucleic Acids Res. 41, 7745–7752.

Marson, A., Levine, S.S., Cole, M.F., Frampton, G.M., Brambrink, T., Johnstone, S., Guenther, M.G., Johnston, W.K., Wernig, M., Newman, J., Calabrese, J.M., Dennis, L.M., Volkert, T.L., Gupta, S., Love, J., Hannett, N., Sharp, P.A., Bartel, D.P., Jaenisch, R., Young, R.A., 2008. Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells. Cell 134, 521–533.

Miyoshi, N., Ishii, H., Nagano, H., Haraguchi, N., Dewi, D.L., Kano, Y., Nishikawa, S., Tanemura, M., Mimori, K., Tanaka, F., Saito, T., Nishimura, J., Takemasa, I., Mizushima, T., Ikeda, M., Yamamoto, H., Sekimoto, M., Doki, Y., Mori, M., 2011. Reprogramming of mouse and human cells to pluripotency using mature microRNAs. Cell Stem Cell 8, 633–638.

Mohammed, J., Siepel, A., Lai, E.C., 2014. Diverse modes of evolutionary emergence and flux of conserved microRNA clusters. RNA (New York, N.Y.) 20, 1850–1863.

Mudunuri, U., Che, A., Yi, M., Stephens, R.M., 2009. bioDBnet: the biological database network. Bioinformatics 25, 555–556.

Murray, M.J., Halsall, D.J., Hook, C.E., Williams, D.M., Nicholson, J.C., Coleman, N., 2011. Identification of microRNAs From the miR-371 ~ 373 and miR-302 clusters as potential serum biomarkers of malignant germ cell tumors. Am. J. Clin. Pathol. 135, 119–125.

Neilsen, C.T., Goodall, G.J., Bracken, C.P., 2012. IsomiRs—the overlooked repertoire in the dynamic microRNAome. Trends Genet. 28, 544–549.

Palmer, R.D., Murray, M.J., Saini, H.K., van Dongen, S., Abreu-Goodger, C., Muralidhar, B., Pett, M.R., Thornton, C.M., Nicholson, J.C., Enright, A.J., Coleman, N., 2010. Malignant germ cell tumors display common microRNA profiles resulting in global changes in expression of messenger RNA targets. Cancer Res. 70, 2911–2923.

Paradis, E., Claude, J., Strimmer, K., 2004. APE: analyses of phylogenetics and evolution in R language. Bioinformatics 20, 289–290.

Pernaute, B., Spruce, T., Smith, K.M., Sanchez-Nieto, J.M., Manzanares, M., Cobb, B., Rodriguez, T.A., 2014. MicroRNAs control the apoptotic threshold in primed pluripotent stem cells through regulation of BIM. Genes Dev. 28, 1873–1878.

Rijlaarsdam, M.A., van Agthoven, T., Gillis, A.J., Patel, S., Hayashibara, K., Lee, K.Y., Looijenga, L.H., 2015. Identification of known and novel germ cell cancer-specific (embryonic) miRs in serum by high-throughput profiling. Andrology 3, 85–91.

Rosenbloom, K.R., Armstrong, J., Barber, G.P., Casper, J., Clawson, H., Diekhans, M., Dreszer, T.R., Fujita, P.A., Guruvadoo, L., Haeussler, M., Harte, R.A., Heitner, S., Hickey, G., Hinrichs, A.S., Hubley, R., Karolchik, D., Learned, K., Lee, B.T., Li, C.H., Miga, K.H., Nguyen, N., Paten, B., Raney, B.J., Smit, A.F., Speir, M.L., Zweig, A.S., Haussler, D., Kuhn, R.M., Kent, W.J., 2015. The UCSC genome browser database: 2015 update. Nucleic Acids Res. 43, D670–D681.

Saini, H.K., Griffiths-Jones, S., Enright, A.J., 2007. Genomic analysis of human microRNA transcripts. Proc. Natl. Acad. Sci. U. S. A. 104, 17719–17724.

Saini, H.K., Enright, A.J., Griffiths-Jones, S., 2008. Annotation of mammalian primary microRNAs. BMC Genomics 9, 564.
Saitou, N., Nei, M., 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4, 406–425.
Saunders, M.A., Liang, H., Li, W.H., 2007. Human polymorphism at microRNAs and microRNA target sites. Proc. Natl. Acad. Sci. U. S. A. 104, 3300–3305.
Suh, M.R., Lee, Y., Kim, J.Y., Kim, S.K., Moon, S.H., Lee, J.Y., Cha, K.Y., Chung, H.M., Yoon, H.S., Moon, S.Y., Kim, V.N., Kim, K.S., 2004. Human embryonic stem cells express a unique set of microRNAs. Dev. Biol. 270, 488–498.
Tamura, K., Stecher, G., Peterson, D., Filipski, A., Kumar, S., 2013. MEGA6: molecular evolutionary genetics analysis version 6.0. Mol. Biol. Evol. 30, 2725–2729.
Tan, G.C., Chan, E., Molnar, A., Sarkar, R., Alexieva, D., Isa, I.M., Robinson, S., Zhang, S., Ellis, P., Langford, C.F., Guillot, P.V., Chandrashekran, A., Fisk, N.M., Castellano, L., Meister, G., Winston, R.M., Cui, W., Baulcombe, D., Dibb, N.J., 2014. 5′ isomiR variation is of functional and evolutionary importance. Nucleic Acids Res. 42, 9424–9435.
Vlachos, I.S., Paraskevopoulou, M.D., Karagkouni, D., Georgakilas, G., Vergoulis, T., Kanellos, I., Anastasopoulos, I.L., Maniou, S., Karathanou, K., Kalfakakou, D., Fevgas, A., Dalamagas, T., Hatzigeorgiou, A.G., 2015. DIANA-TarBase v7.0: indexing more than half a million experimentally supported miRNA:mRNA interactions. Nucleic Acids Res. 43, D153–D159.
Warthmann, N., Das, S., Lanz, C., Weigel, D., 2008. Comparative analysis of the MIR319a microRNA locus in *Arabidopsis* and related Brassicaceae. Mol. Biol. Evol. 25, 892–902.
Whyte, W.A., Orlando, D.A., Hnisz, D., Abraham, B.J., Lin, C.Y., Kagey, M.H., Rahl, P.B., Lee, T.I., Young, R.A., 2013. Master transcription factors and mediator establish super-enhancers at key cell identity genes. Cell 153, 307–319.
Wickham, H., Wickham, H., 2009. ggplot2. Elegant graphics for data analysis. Wiley Interdisciplinary Rev.: Comp. Wiley Interdiscip. Rev. Comput. Stat. 3, 180–185.
Xiao, F., Zuo, Z., Cai, G., Kang, S., Gao, X., Li, T., 2009. miRecords: an integrated resource for microRNA-target interactions. Nucleic Acids Res. 37, D105–D110.
Zhang, Z., Xiang, D., Heriyanto, F., Gao, Y., Qian, Z., Wu, W.S., 2013. Dissecting the roles of miR-302/367 cluster in cellular reprogramming using TALE-based repressor and TALEN. Stem Cell Rep. 1, 218–225.