Supplemental Material

Dynamics of microRNA expression during mouse prenatal development

Sorena Rahmanian, Rabi Murad, Alessandra Breschi, Weihua Zeng, Mark Mackiewicz, Brian Williams, Carrie Davis, Brian Roberts, Sarah Meadows, Dianna Moore, Diane Trout, Chris Zaleski, Alex Dobin, Lei-Hoon Sei, Jorg Drenkow, Alex Scavelli, Thomas Gingeras, Barbara Wold, Richard M. Myers, Roderic Guigó, Ali Mortazavi

Supplemental Fig. S1: Overview of mouse ENCODE miRNA data sets …… 3

Supplemental Fig. S2: Sequencing depth analysis of microRNA-seq data ……. 4

Supplemental Fig. S3: Coverage of miRNAs from different databases during mouse embryonic development ……. 5

Supplemental Fig. S4: Domination of expression by highly expressed miRNAs ……... 6

Supplemental Fig. S5: Characterization of the top 10 highly expressed miRNAs …….. 7

Supplemental Fig. S6: Alternative analysis of tissue specificity profiles …… 8

Supplemental Fig. S7: Analysis of miRNAs that change their specificity …….. 9

Supplemental Fig. S8: Analysis of miRNAs that are highly specific …….. 10

Supplemental Fig. S9: Analysis of miRNAs that are mostly ubiquitous …… 11

Supplemental Fig. S10: MicroRNA maSigPro clusters …….. 12

Supplemental Fig. S11: Expression profiles of NanoString miRNA clusters ……. 13

Supplemental Fig. S12: MicroRNA cluster tissue-specificity …… 14

Supplemental Fig. S13: Overview of human ENCODE miRNA data sets …… 15

Supplemental Fig. S14: Human fetal development miRNA transcriptome …… 16

Supplemental Fig. S15: Comparative dynamics of miRNAs during human and mouse development ……. 17

Supplemental Fig. S16: Comparison of miRNAs and their primary transcripts using GENCODE annotations augmented with *ab initio* models …… 18

Supplemental Fig. S17: Distribution of StringTie transcript models's TSS that have overlapping H3K4me3 peaks …….. 19

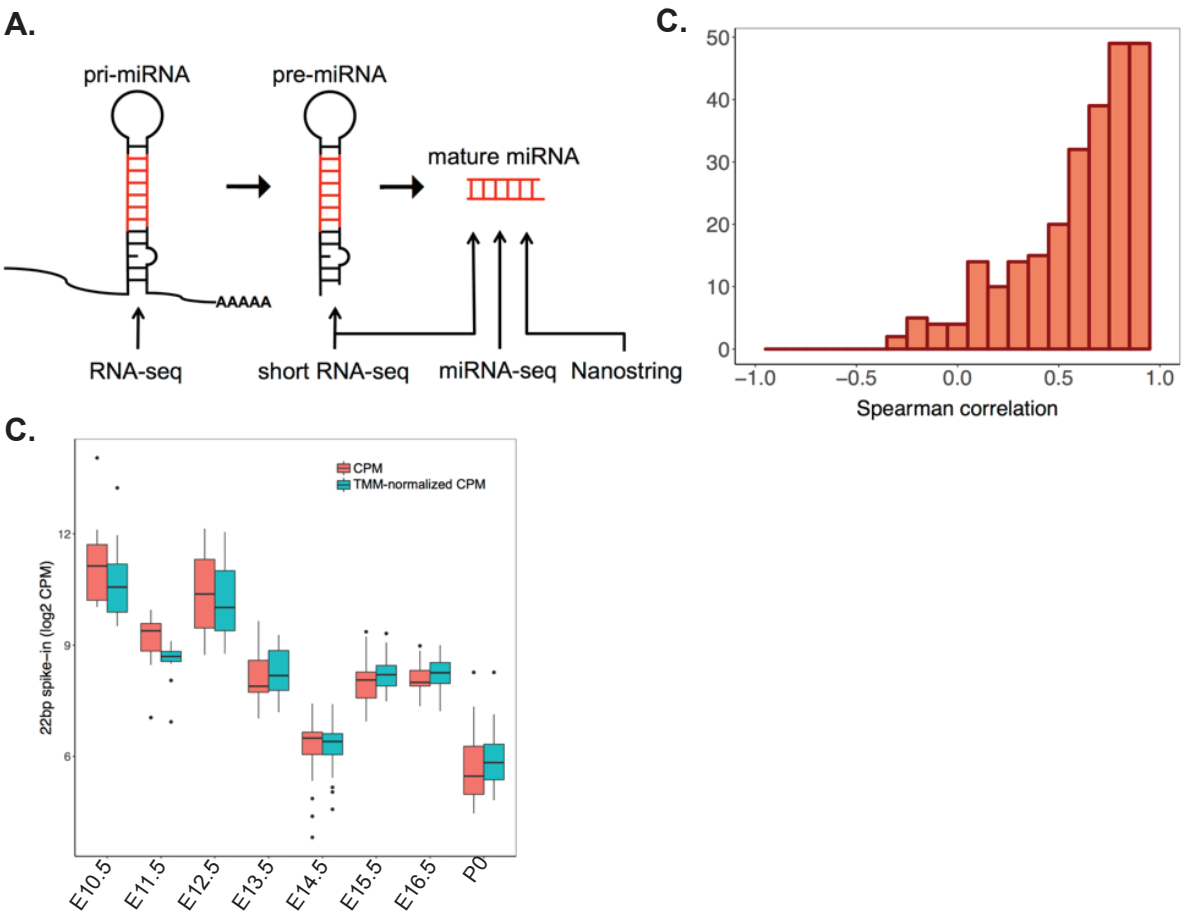
Supplemental Fig. S18: mRNA maSigPro cluster expression profiles ……. 20

Supplemental Fig. S19: Expression anti-correlation analysis ……... 21

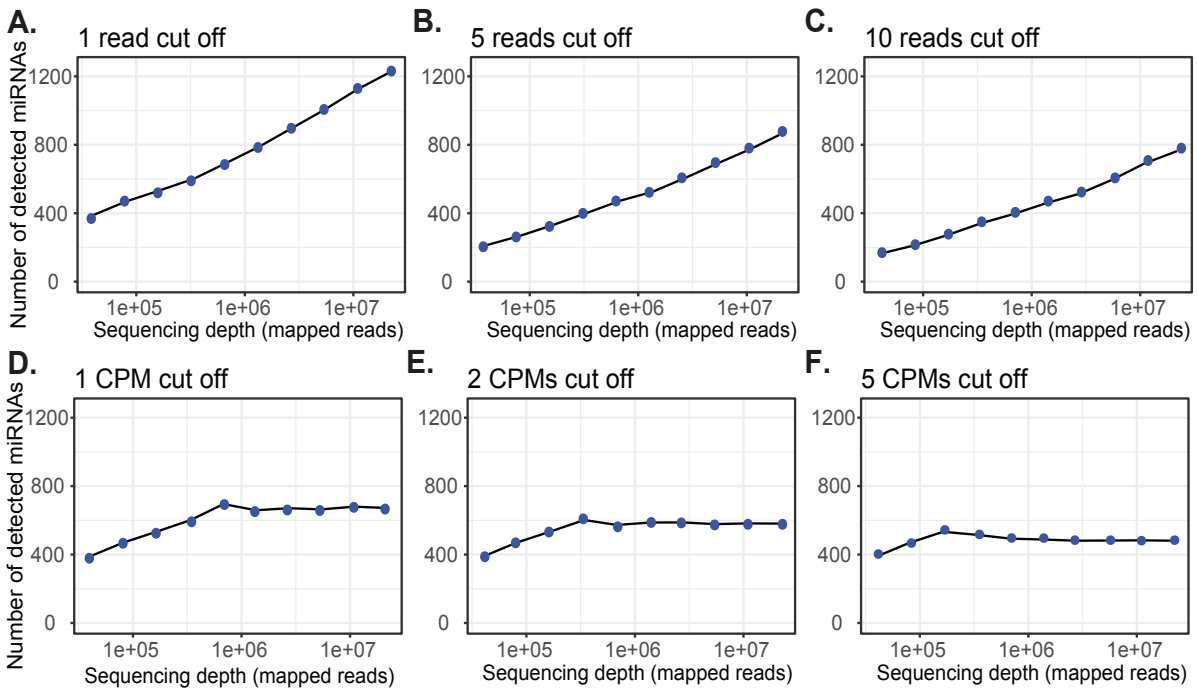
Supplemental Fig. S20: Expression profiles of all the significant interactions ……..…… 22

Supplemental Fig. S21: Conservation of microRNAs target sites ……..... 23

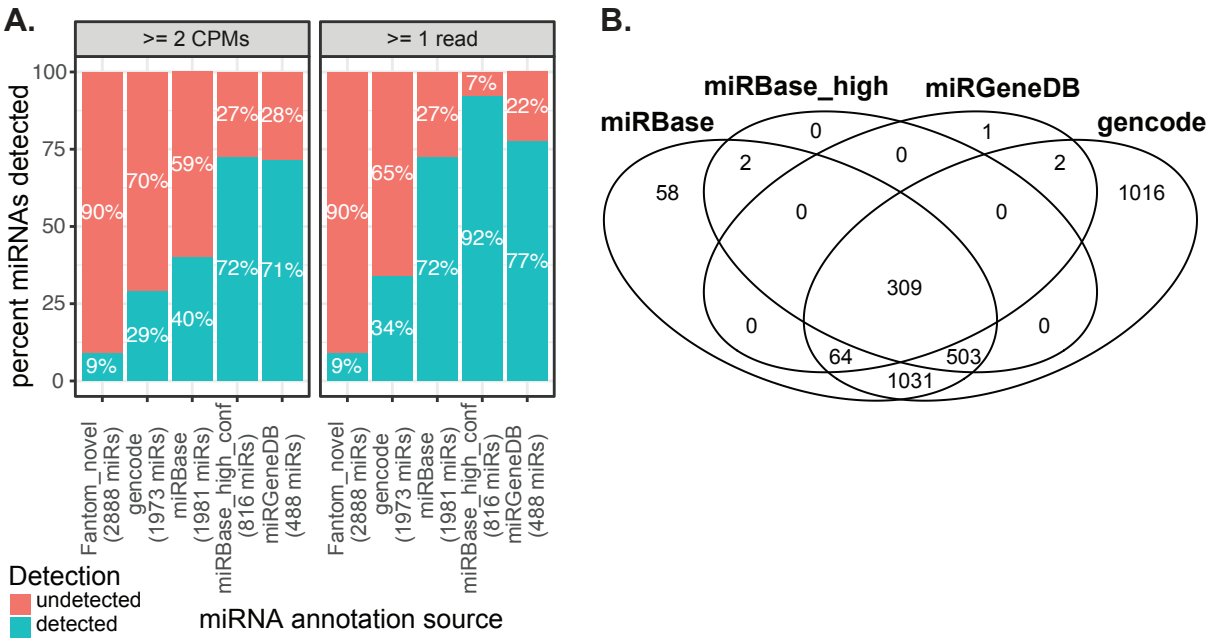
Supplemental Methods ……..… 24



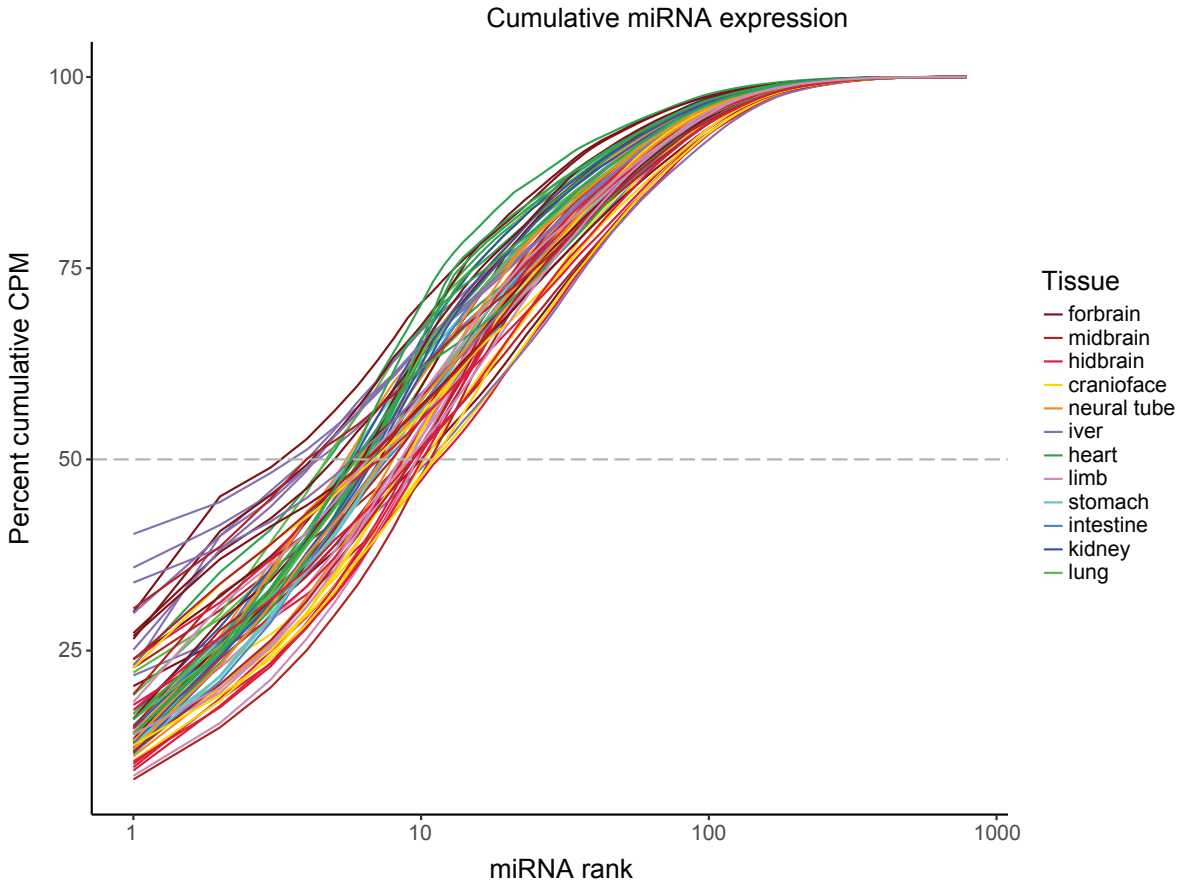
Supplemental Fig. S1: Overview of mouse ENCODE miRNA data sets. (A) Primary miRNAs were profiled using mRNA-seq (> 200 nt) in human and mouse, pre-miRNAs and mature miRNAs were profiled in human using short RNA-seq (< 200 nt), and mature miRNAs in mouse were profiled using microRNA-seq (< 30 nt) and NanoString. **(B)** Distribution of Spearman correlations (median = 0.68) between microRNA-seq and NanoString for miRNAs included in the NanoString coreset. **(C)** Expression levels of the 22 bp spike-in in mouse microRNA-seq samples across different stages of embryonic development for non-normalized counts-per-million (CPM, red boxplots) and TMM-normalized CPMs (cyan boxplots).



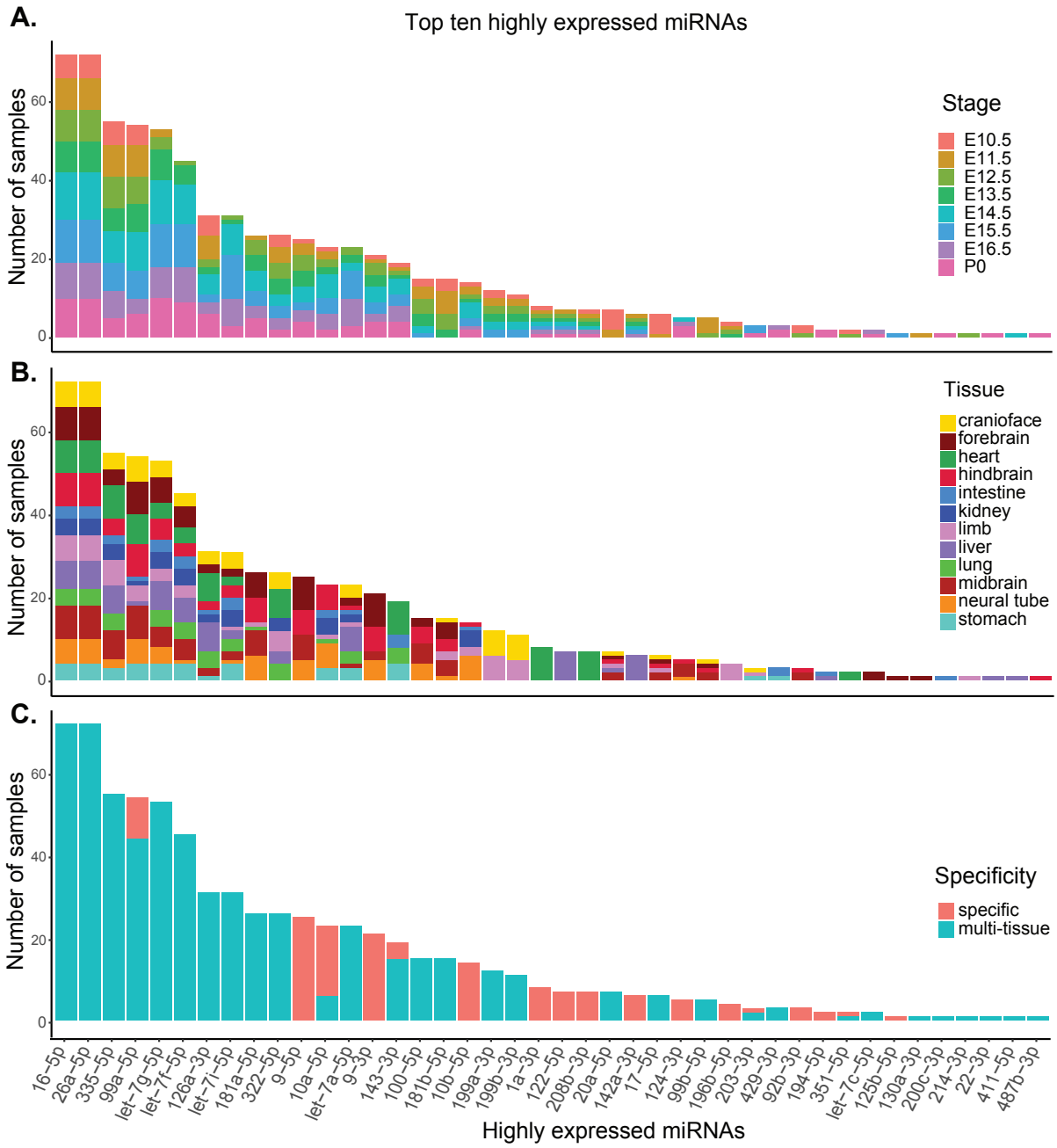
Supplemental Fig. S2: Sequencing depth analysis of microRNA-seq data. The deeply sequenced (~ 22 M mapped reads) heart E11.5 replicate 1 sample was subsampled to assess miRNA detection at different sequencing depths (11M, 5.4M, 2.7M, 1.4M, 680K, 340K, 160K, 80K and 40K) using different cutoffs: **(A)** 1 read, **(B)** 5 reads and **(C)** 10 reads, **(D)** 1 CPM , **(E)** 1M for 2 CPMs and **(F)** 5 CPMs.



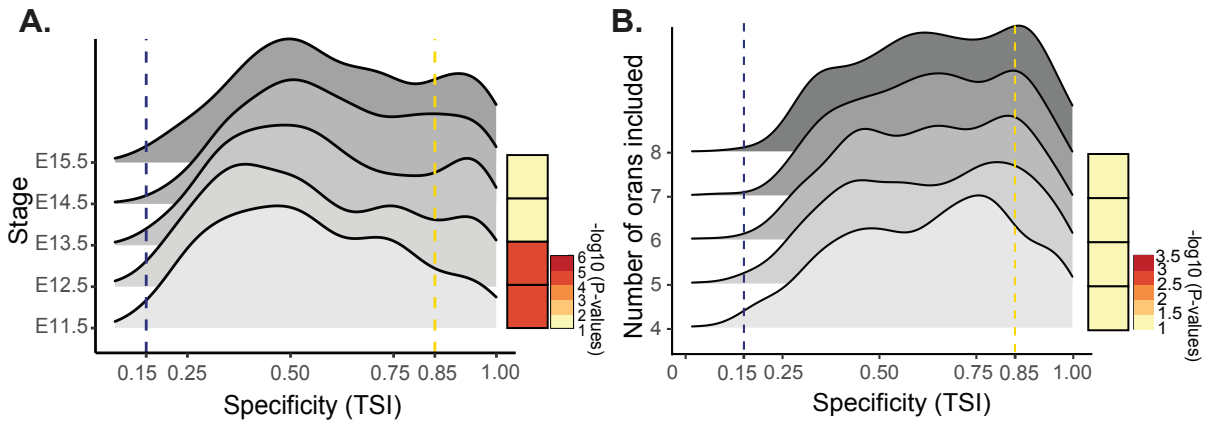
Supplemental Fig. S3: Coverage of miRNAs from different databases during mouse embryonic development. (A) We mapped our microRNA-seq data to mature miRNAs from different sources such as miRbase, miRbase high confidence set, miRGeneDB and the Fantom 5 novel miRNA set. Our data includes more than 70% of miRNAs annotated in the high confidence set of miRBase and miRGeneDB at a minimum of 2 CPM for at least one of the organs at one of the developmental time points. These numbers were even higher when using a softer cut off of 1 read **(B)** Overlap of annotated miRNAs in different databases.



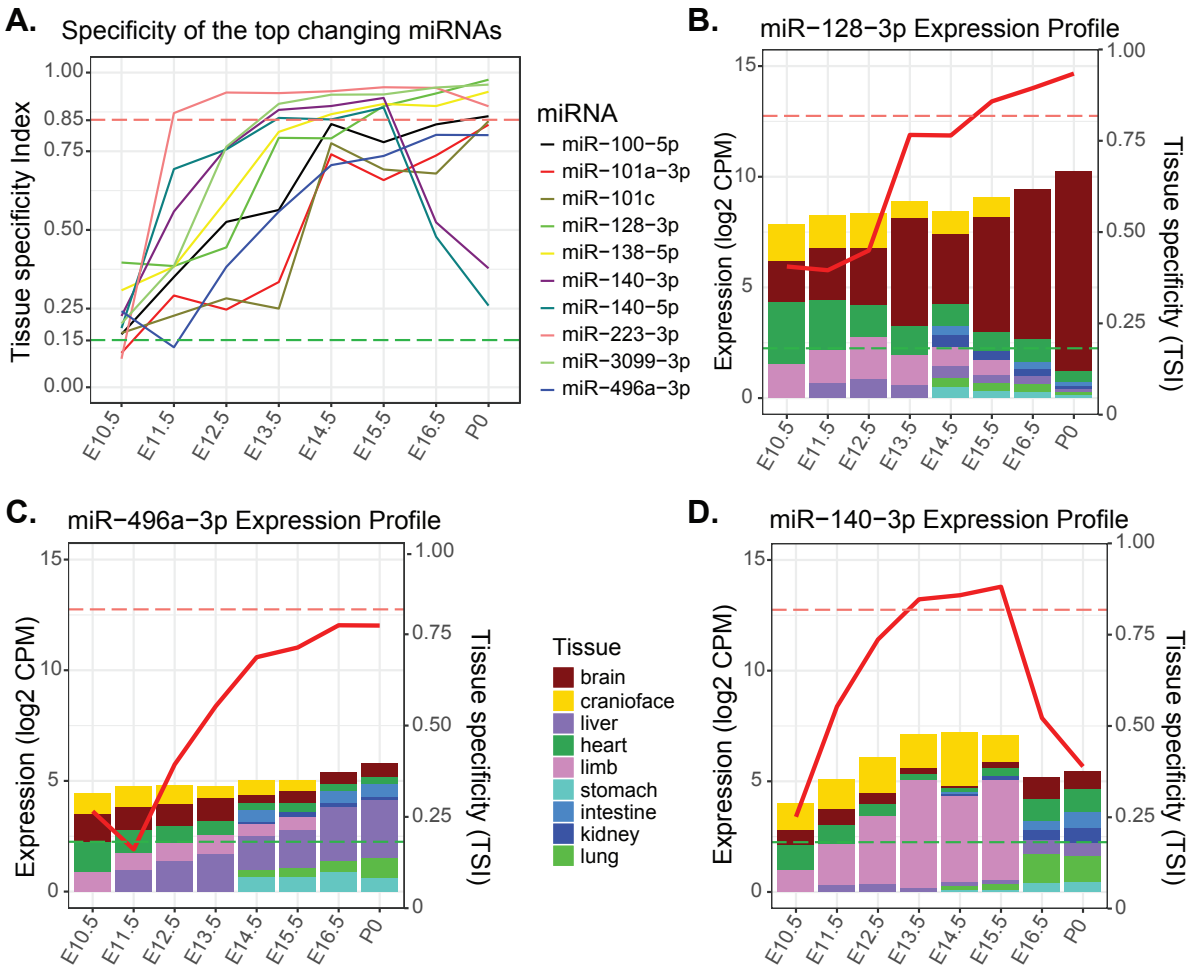
Supplemental Fig. S4: A few highly expressed miRNAs dominate miRNA-seq datasets. Cumulative distribution of sequencing reads, using miRNAs ranked by expression levels. The top 10 miRNAs account for more than half of the miRNA sequencing reads.



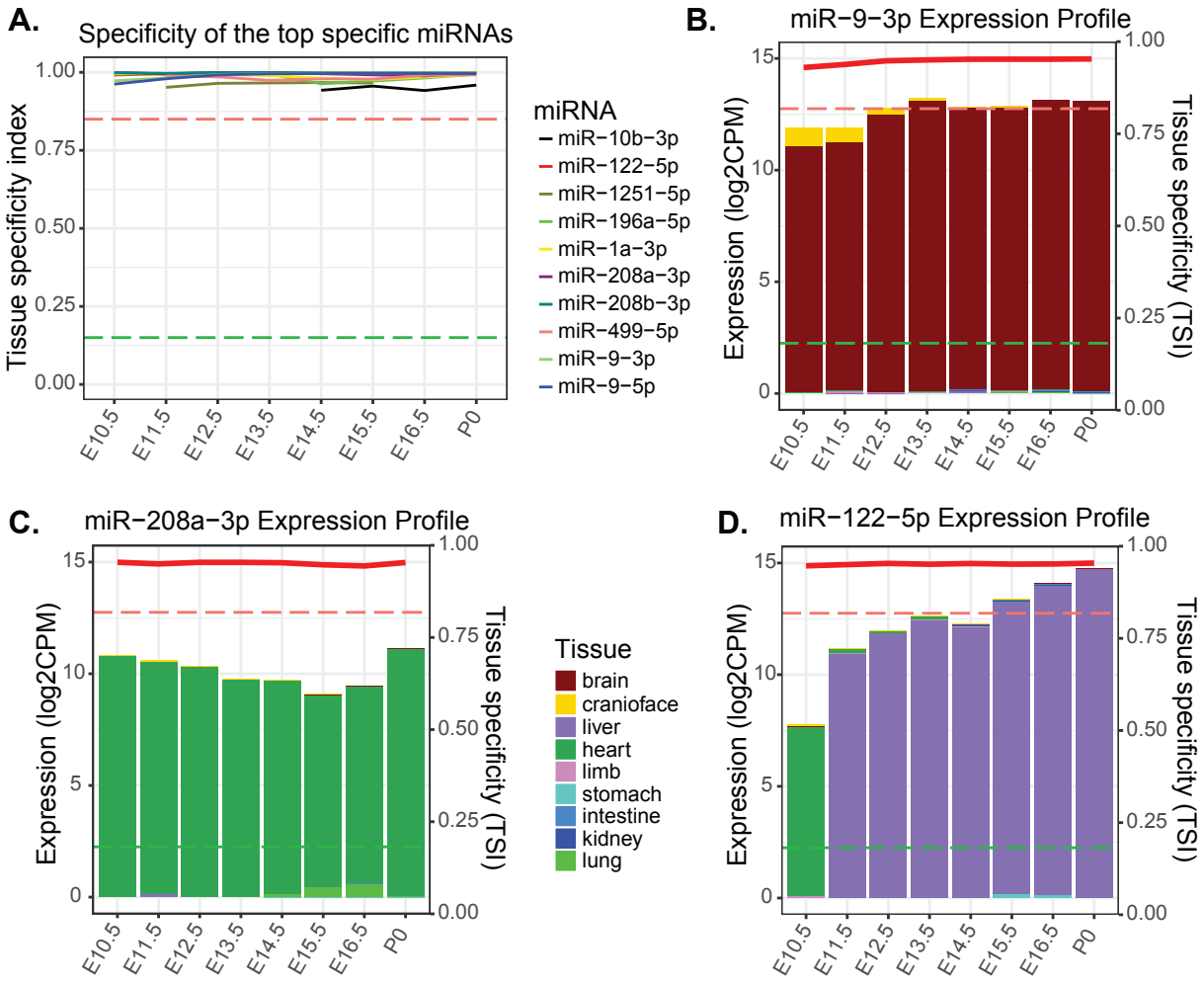
Supplemental Fig. S5: Characterization of the specificity of highly expressed miRNAs. Histograms of the 43 miRNAs that form the top ten highly expressed miRNAs in all the samples, ranked by the number of samples they are highly expressed in; colored by their **(A)** stage, **(B)** organ, and **(C)** specificity.



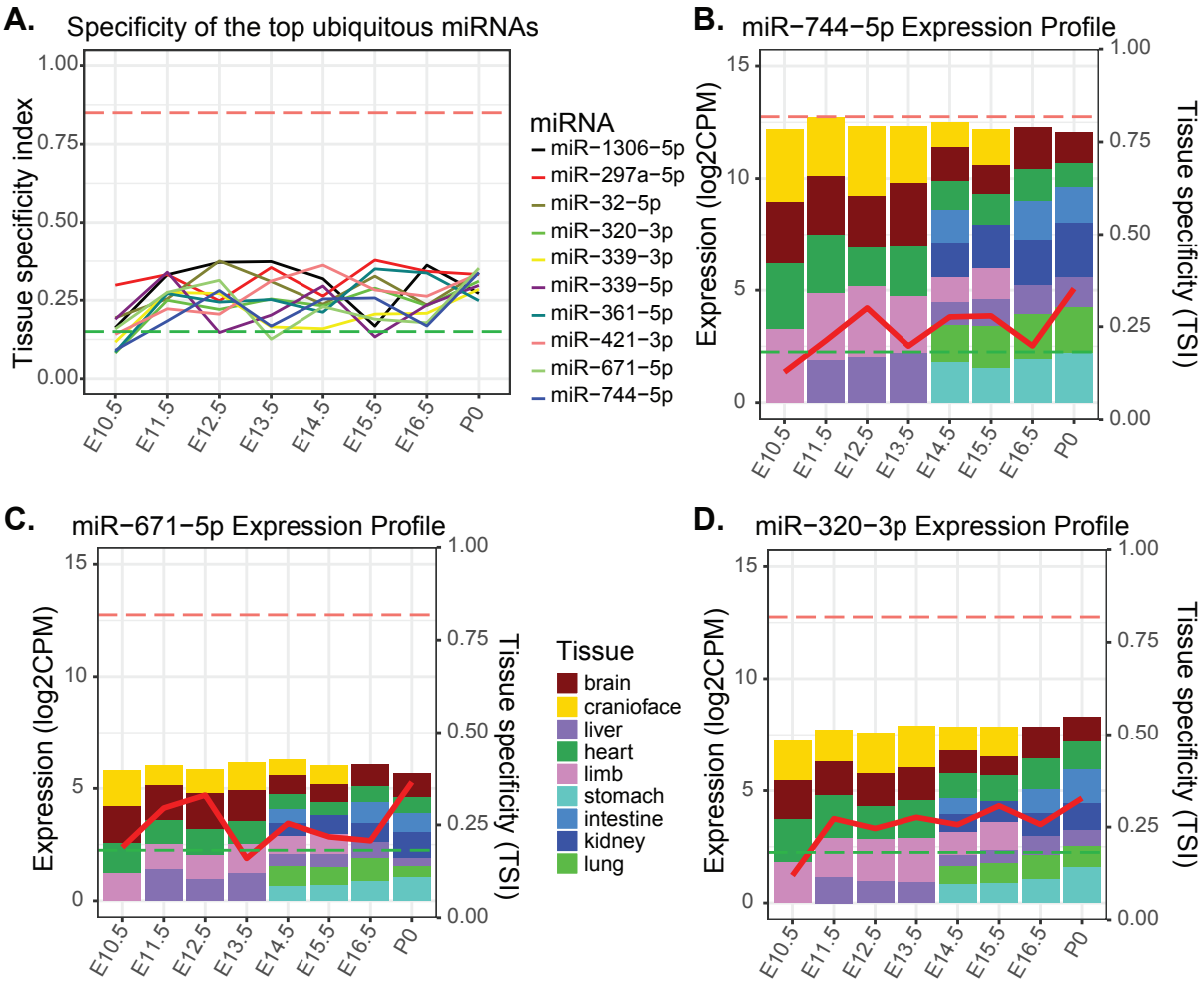
Supplemental Fig. S6: Alternative analysis of tissue specificity profiles. (A) In order to ensure that the observed differences in the distribution of miRNA tissue specificity is not an artifact of different number of organs used for each stage, we looked at the different distributions of the five intermediate stages, constraining the calculation of TSI to the same organs (forebrain, heart, cranioface, liver and limb) and observed that there is still statistically significant differences between TSI distribution of earlier stages and later stages **(B)** Then we looked at the TSI profile of P0 samples and re-calculated the TSI using different number of tissues, which did not result in any significant differences in the re-calculated distributions.



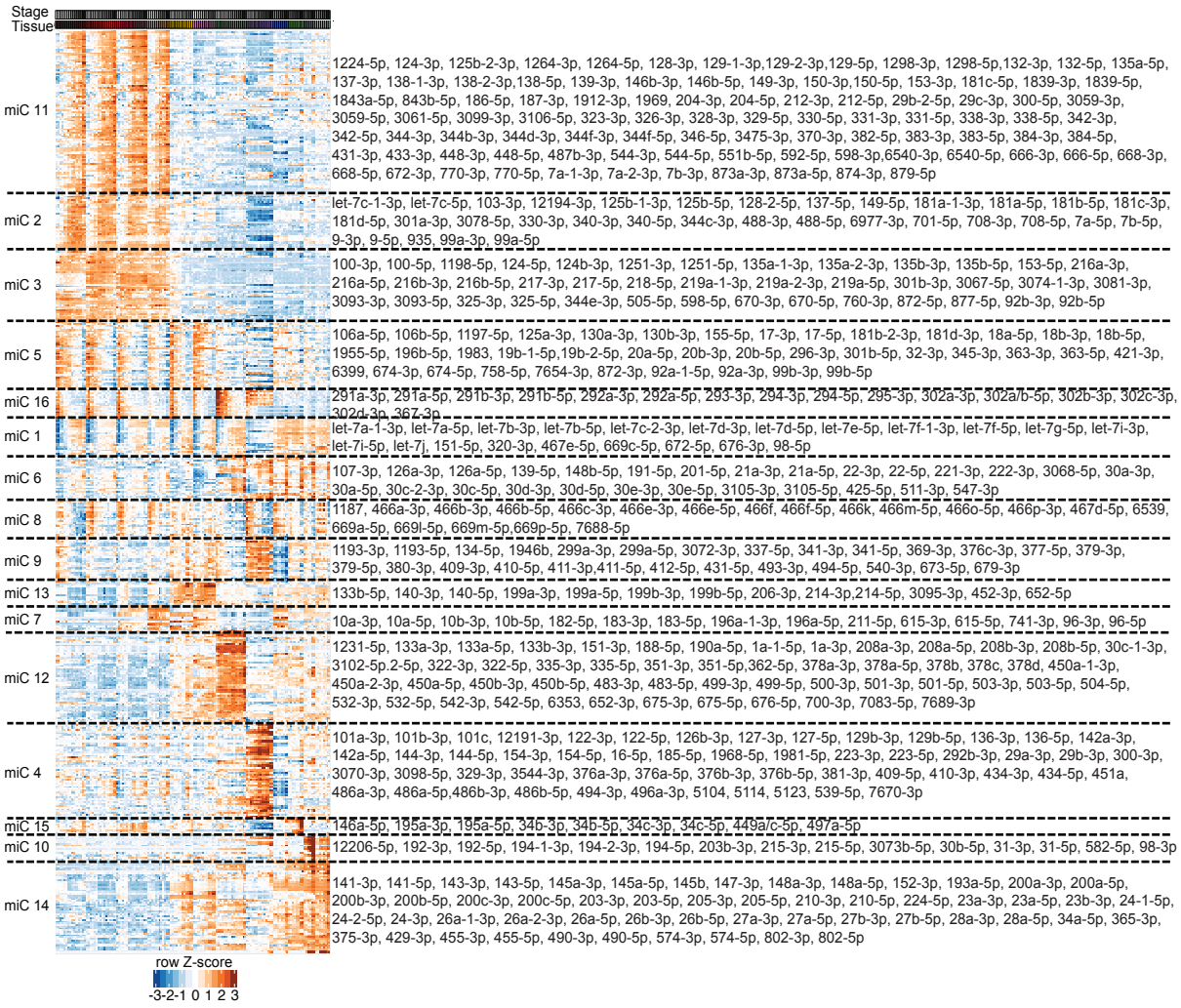
Supplemental Fig. S7: Analysis of miRNAs that change their specificity. (A) The tissue specificity profile of ten miRNA with the most change in their specificity through the embryonic development. **(B-D)** Expression profile and bar graphs for three of these miRNAs: **(B)** miR-128-3p, **(C)** miR-496a-3p, **(D)** miR-140-3p. The red curve traces the tissue specificity (TSI) on the second axis and each bar is colored proportional to the expression of the miRNA in different tissues. The fall in tissue specificity of miR-140-3p at the last two time point may be due to the lack of limb samples for those two time-point.



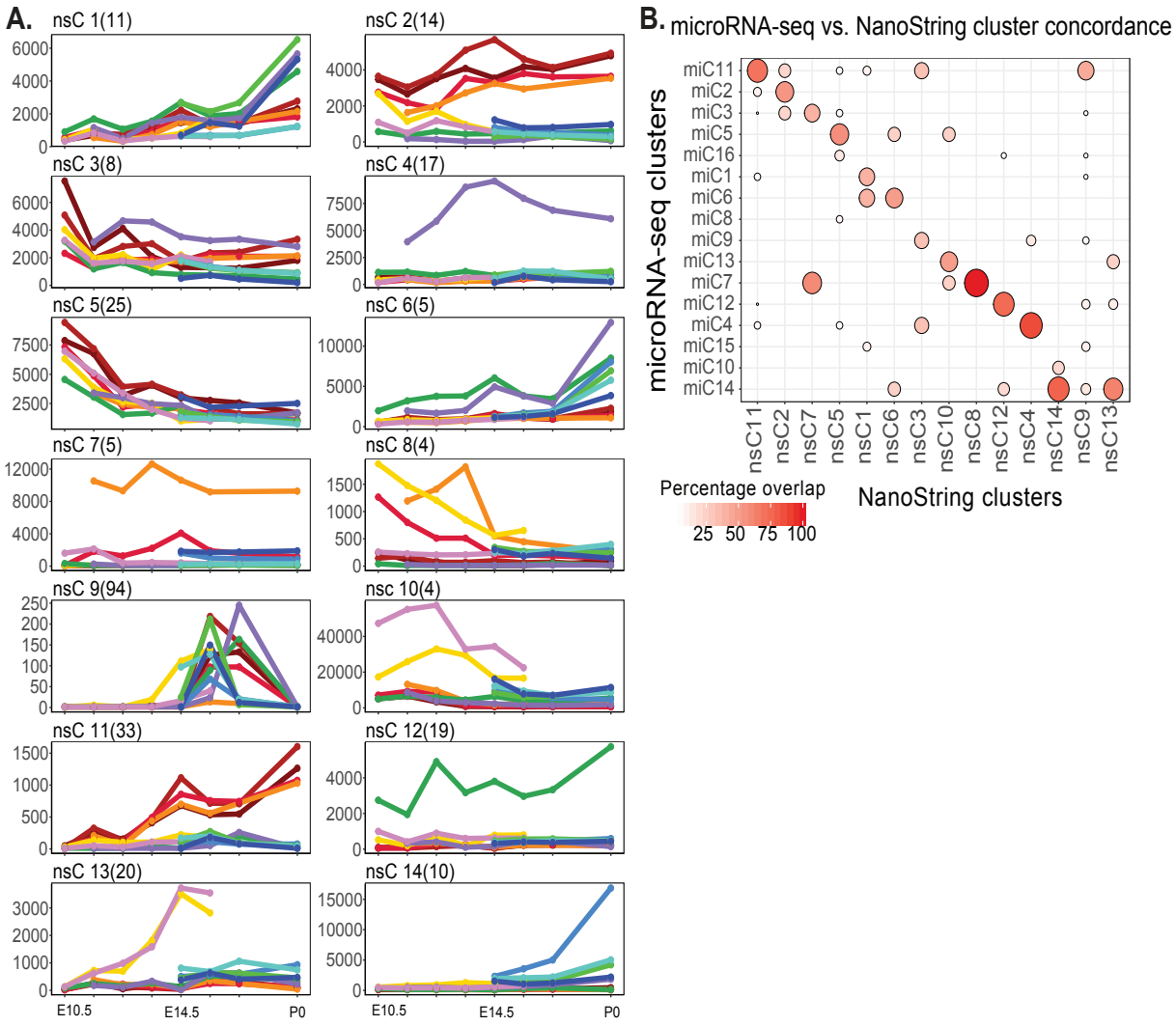
Supplemental Fig. S8: Analysis of miRNAs that are highly specific. (A) The tissue specificity profile of ten miRNA with the highest specificity through the embryonic development. **(B-D)** The bar graphs show the expression profile of three of these miRNAs: **(B)** miR-9-3p, **(C)** miR-208a-3p, **(D)** miR-122-5p. The red curve traces the tissue specificity (TSI) on the second axis and each bar is colored proportional to the expression of the miRNA in different tissues. miR-122-5p is highly specific at all stages but is shown as specific to heart at the E10.5 and liver specific at every other stage, possibly due to lack of a liver sample for stage E10.5.



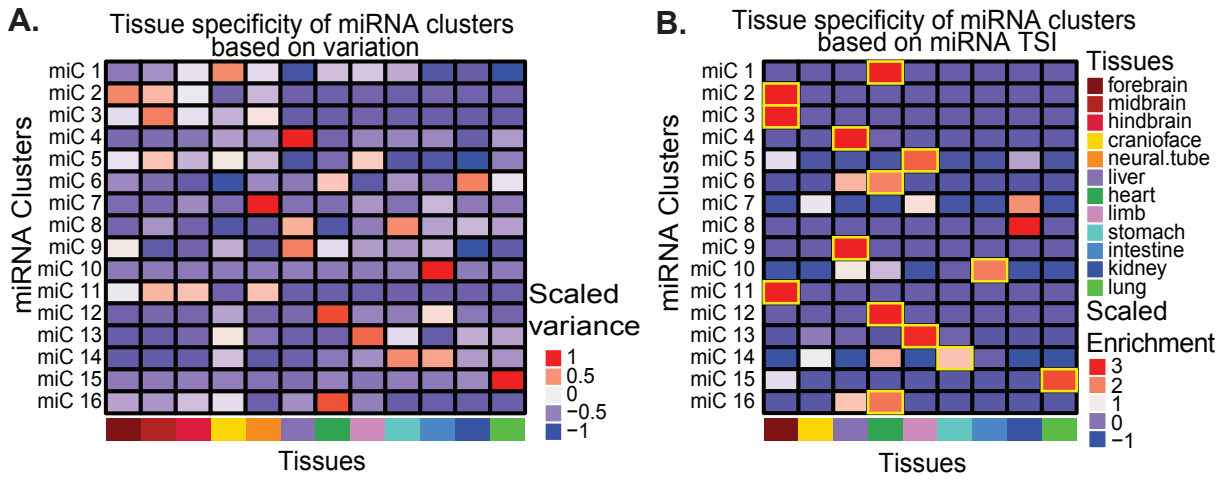
Supplemental Fig. S9: Analysis of miRNAs that are mostly ubiquitous. (A) The tissue specificity profile of ten miRNA with the lowest specificity through the embryonic development. **(B-D)** The bar graphs show the expression profile of three of these miRNAs: **(B)** miR-744-5p, **(C)** miR-671-5p, **(D)** miR-320-3p. The red curve traces the tissue specificity (TSI) on the second axis and each bar is colored proportional to the expression of the miRNA in different tissues.



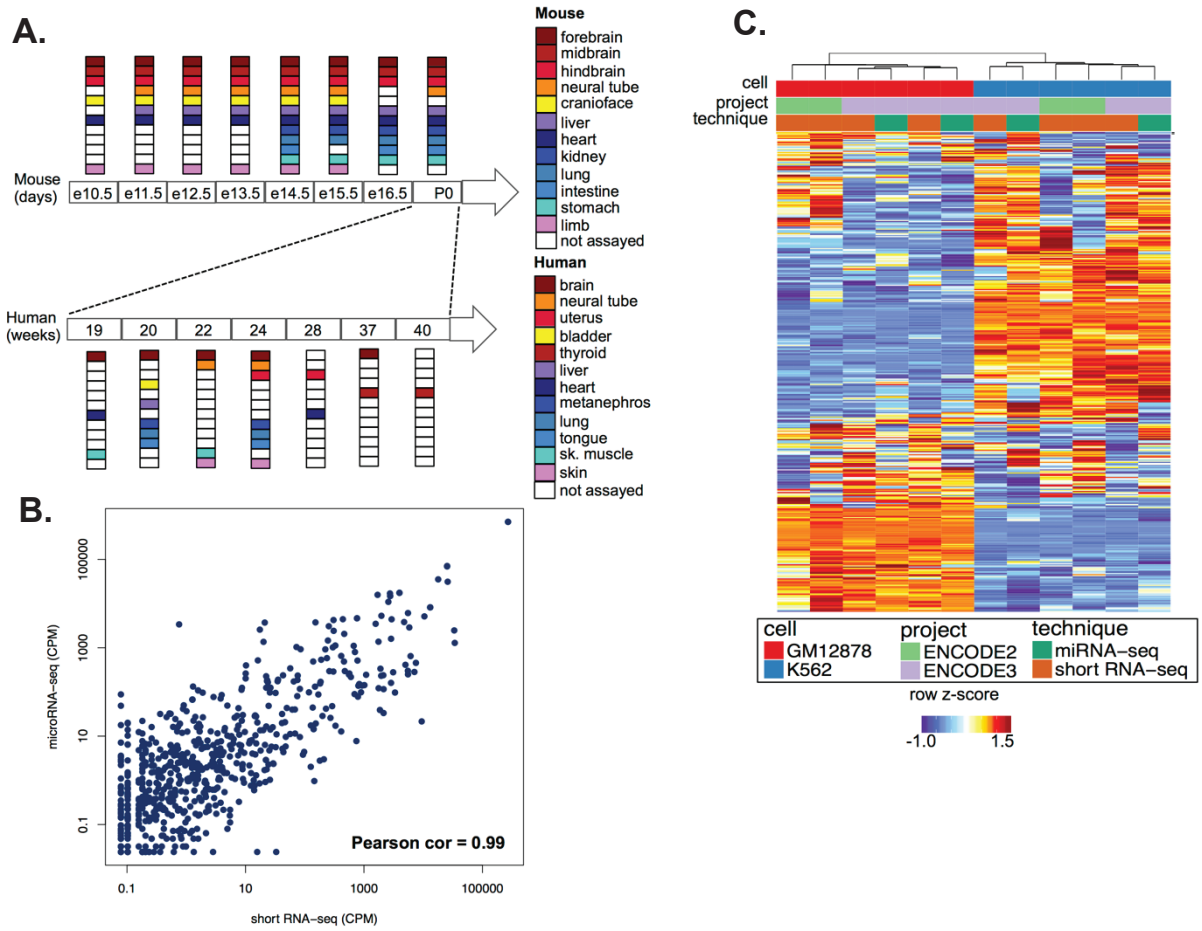
Supplemental Fig. S10: MicroRNA maSigPro clusters. Heatmap of the differentially expressed miRNAs clustered with maSigPro with the names of the miRNAs belonging to the different clusters listed.



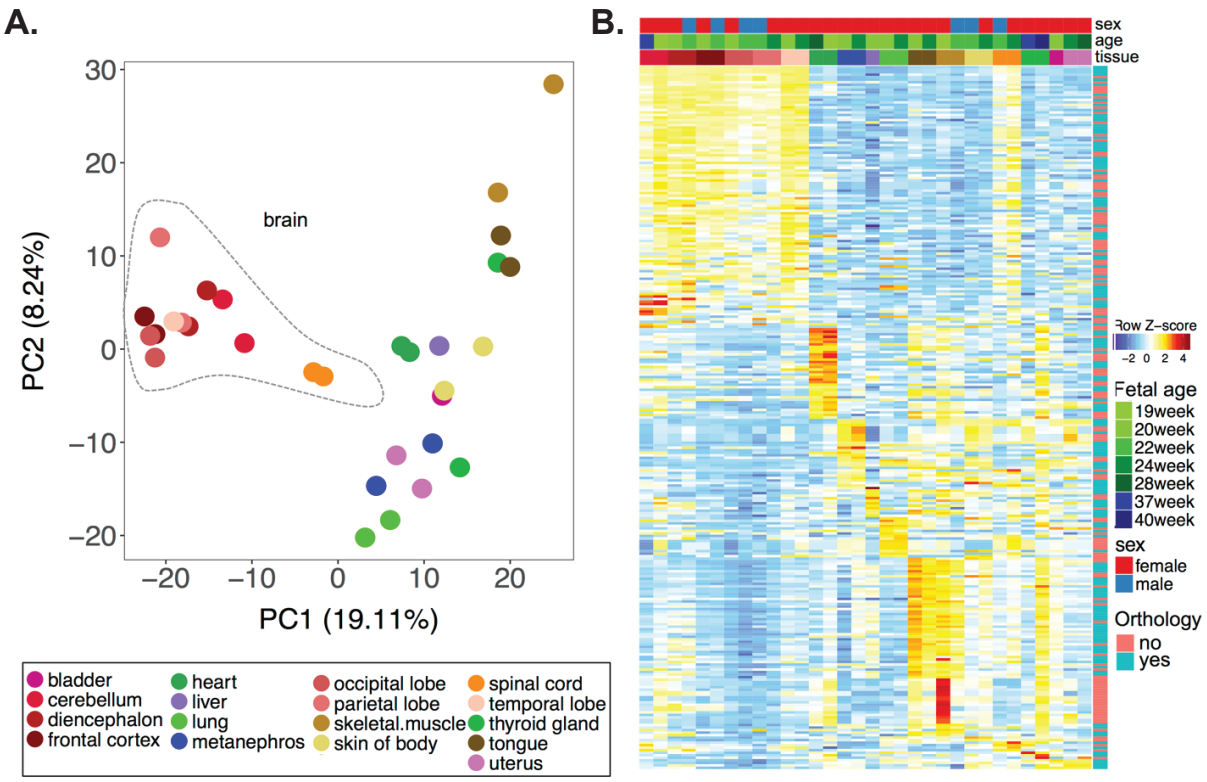
Supplemental Fig. S11: Expression profiles of NanoString miRNA clusters. (A) Median expression profiles for each of 14 miRNA clusters, measured using NanoString in matching samples, that were identified as differentially expressed by the linear regression-based algorithm maSigPro**. (B)** Comparison of NanoString and microRNA-seq shows overall good concordance between the clusters derived using the two orthogonal technologies.



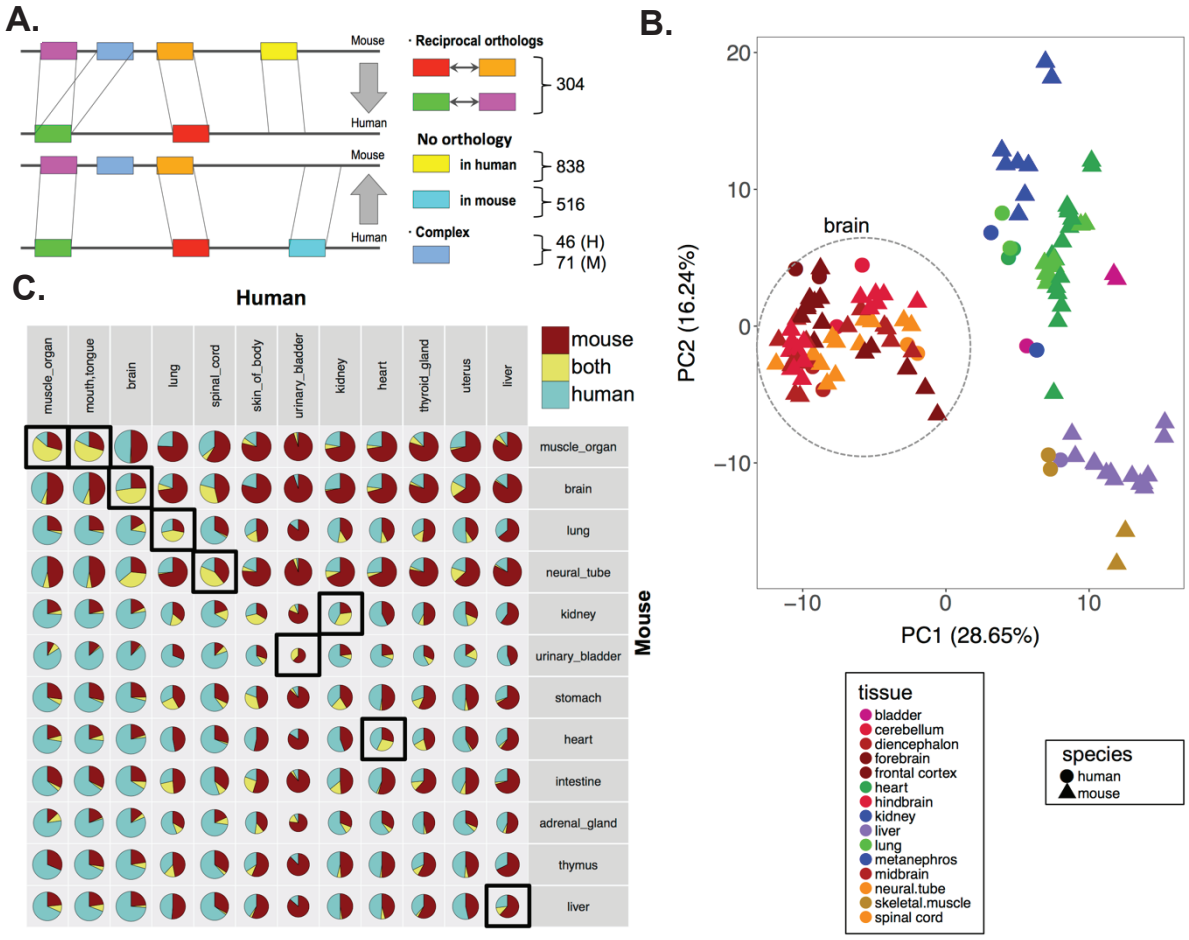
Supplemental Fig. S12: MicroRNA cluster tissue-specificity. (A) Heatmap of the variance of miRNA cluster mean expression during the time course in each organ. These values were scaled for each cluster separately to identify the tissue specificity of that cluster. **(B)** We further determined the tissue-specificity of the miRNA clusters by enrichment analysis of the tissue-specific miRNAs in each cluster. Tissue-specificity of the majority of the miRNA clusters correspond with what has been determined using the variance of the average miRNA expression of each cluster within each tissue, with gold boxed clusters indicating where the tissue specificities are concordant with cluster tissue specificity in panel A.



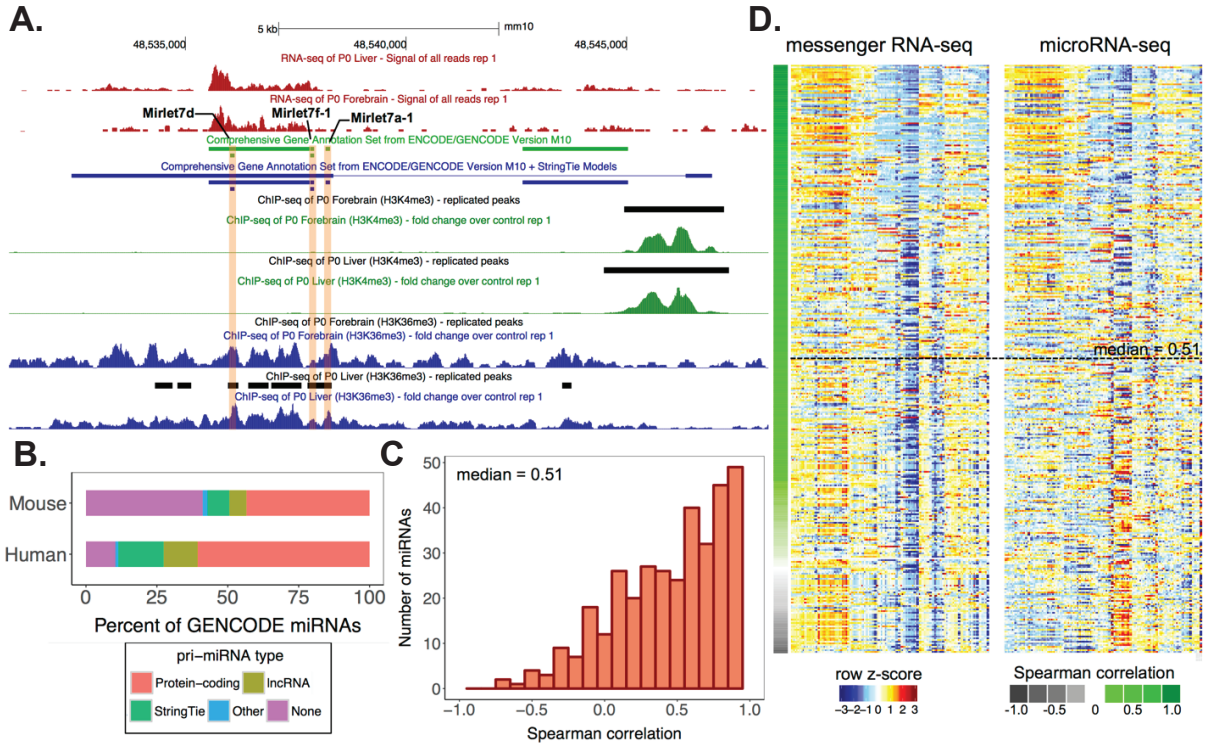
Supplemental Fig. S13: Overview of human ENCODE miRNA data sets. (A) Primary tissues representative of major organ systems were profiled in human along various stages of fetal development (weeks 19-40). **(B)** Comparison of normalized miRNA counts for GM12878, profiled using microRNA-seq and short RNA-seq, demonstrates high correlation between the two assays. **(C)** Heatmap of miRNA normalized counts in GM12878 and K562 cell lines shows that the samples cluster by cell type irrespective of profiling technique used or the date of sample preparation.



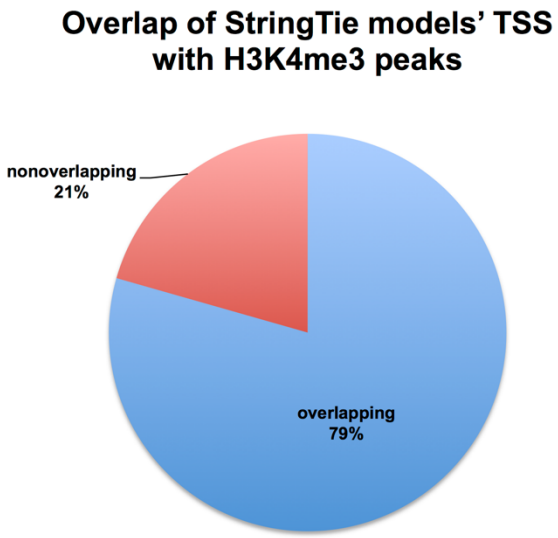
Supplemental Fig. S14: Human fetal development miRNA transcriptome. (A) Principal component analysis (PCA) of human tissues reveals distinct miRNA expression patterns in brain samples compared to other tissues. Colors denote different tissues. **(B)** Normalized expression levels of human tissue-specific miRNAs. Differential expression analysis reveals the largest set of differentially expressed miRNAs in brain and muscle samples.



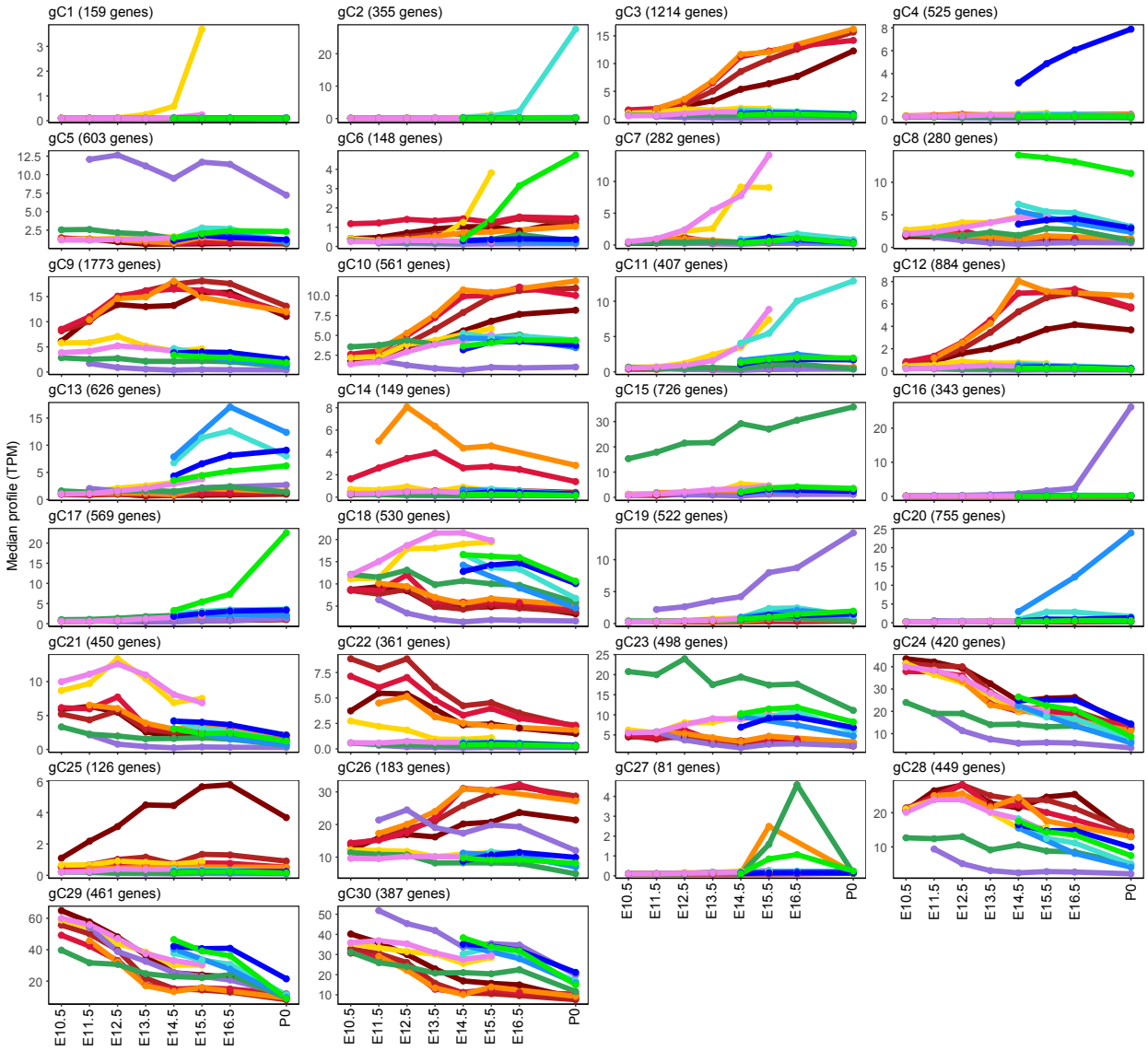
Supplemental Fig. S15: Comparative dynamics of miRNAs during human and mouse development. (A) 304 one-to-one orthologous human and mouse miRNAs were identified using reciprocal search. **(B)** Combined principal component analysis (PCA) of human and mouse samples. Triangles and circles denote mouse and human tissues respectively. Tissues are denoted by different colors. **(C)** Intersection of human and mouse tissue-specific miRNAs. For each pair of tissues in human and mouse we report the fraction of tissue-specific miRNAs in mouse only (red), human only (blue), or in both (yellow) within the 304 orthologous miRNAs. The sizes of the pie chart are in proportion to the numbers of tissue-specific miRNAs in the corresponding tissues.



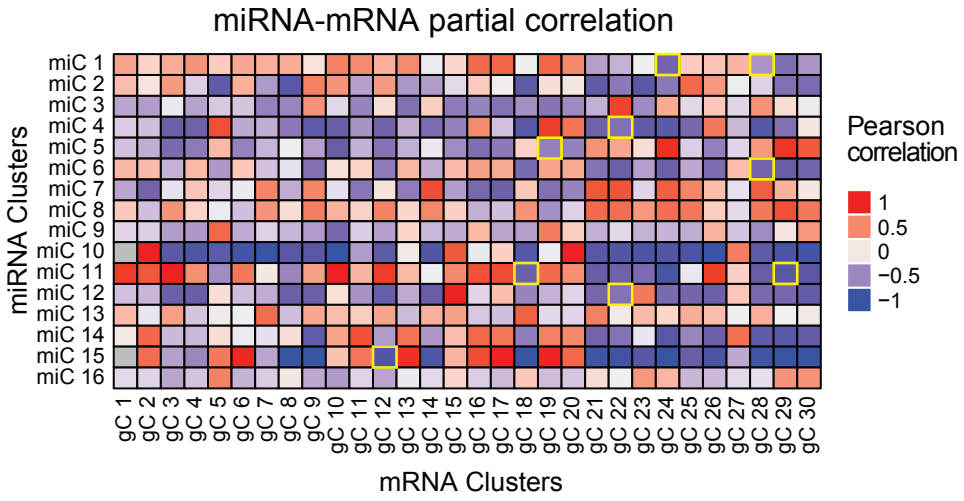
Supplemental Fig. S16: Comparison of miRNAs and their primary transcripts using GENCODE annotations augmented with *ab initio* models. (A) A genome browser snapshot of a representative *ab initio* gene model generated using mouse mRNA-seq data for 2 miRNAs (Mir-let7f-1 and Mir-let7a-1) that lack annotated GENCODE v. M10 pri-miRNAs. Additional evidence for the gene models is provided using H3K4me3 and K3K36me3 data in matching mouse samples. **(B)** Distribution of types of pri-miRNAs biotypes (protein-coding, lncRNA, *ab initio* gene models, and others) in mouse and human. Improvements in pri-miRNA annotations using the *ab initio* gene models are denoted by color green in both human and mouse. **(C)** Distribution of Spearman correlation (median = 0.51) among miRNAs and their corresponding pri-miRNAs. **(D)** Comparison of expression levels of pri-miRNAs and their corresponding mature miRNAs. The rows are sorted with decreasing Spearman correlation top to bottom.



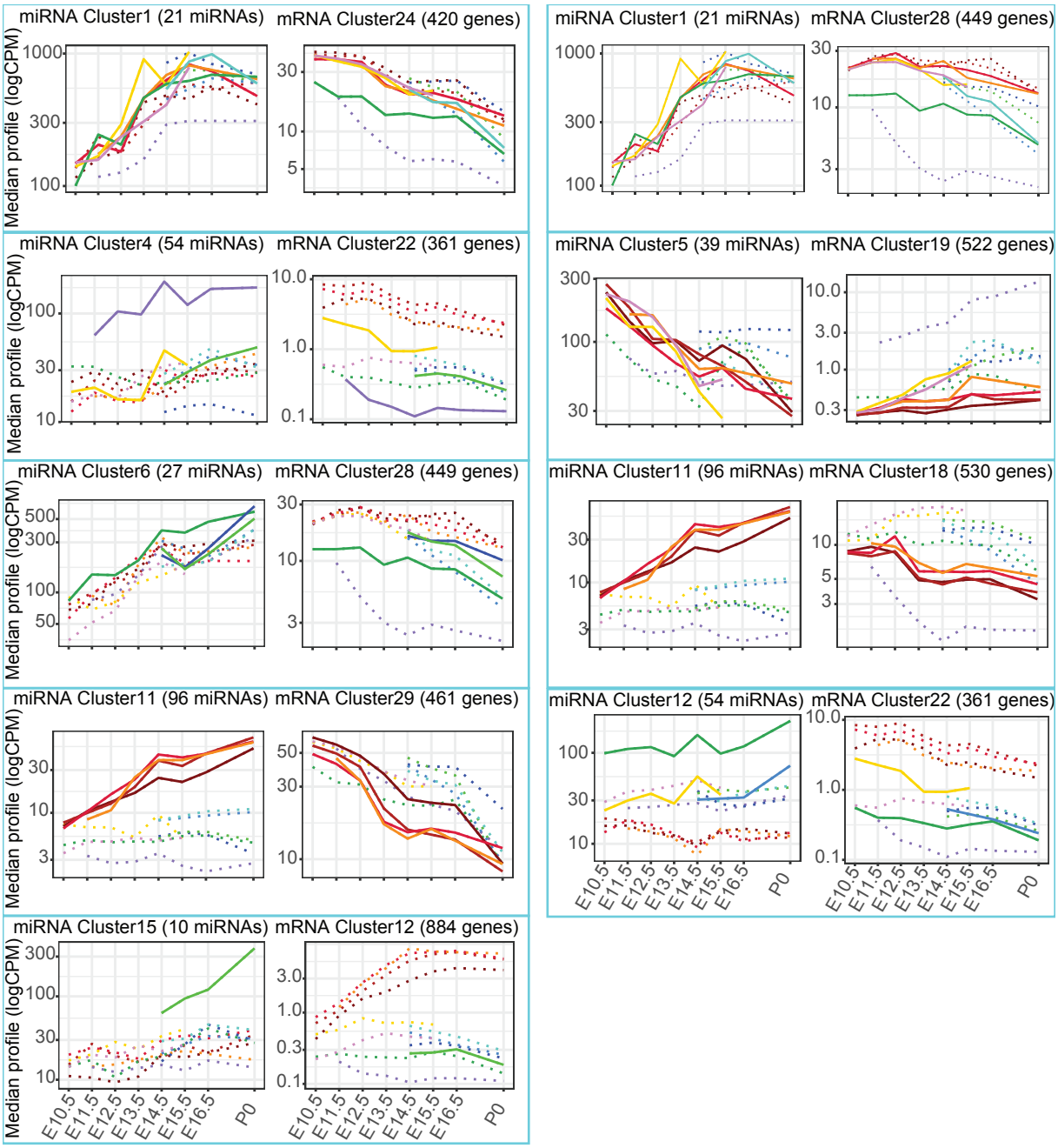
Supplemental Fig. S17: Distribution of StringTie transcript models's TSS that have overlapping H3K4me3 peaks Proportion of overlap between the TSS of StringTie *ab initio* transcript models and the H3K4me3 peaks in matching samples. The TSS of the transcript models were defined as ±250 bp regions from the 5'-end of the transcripts.



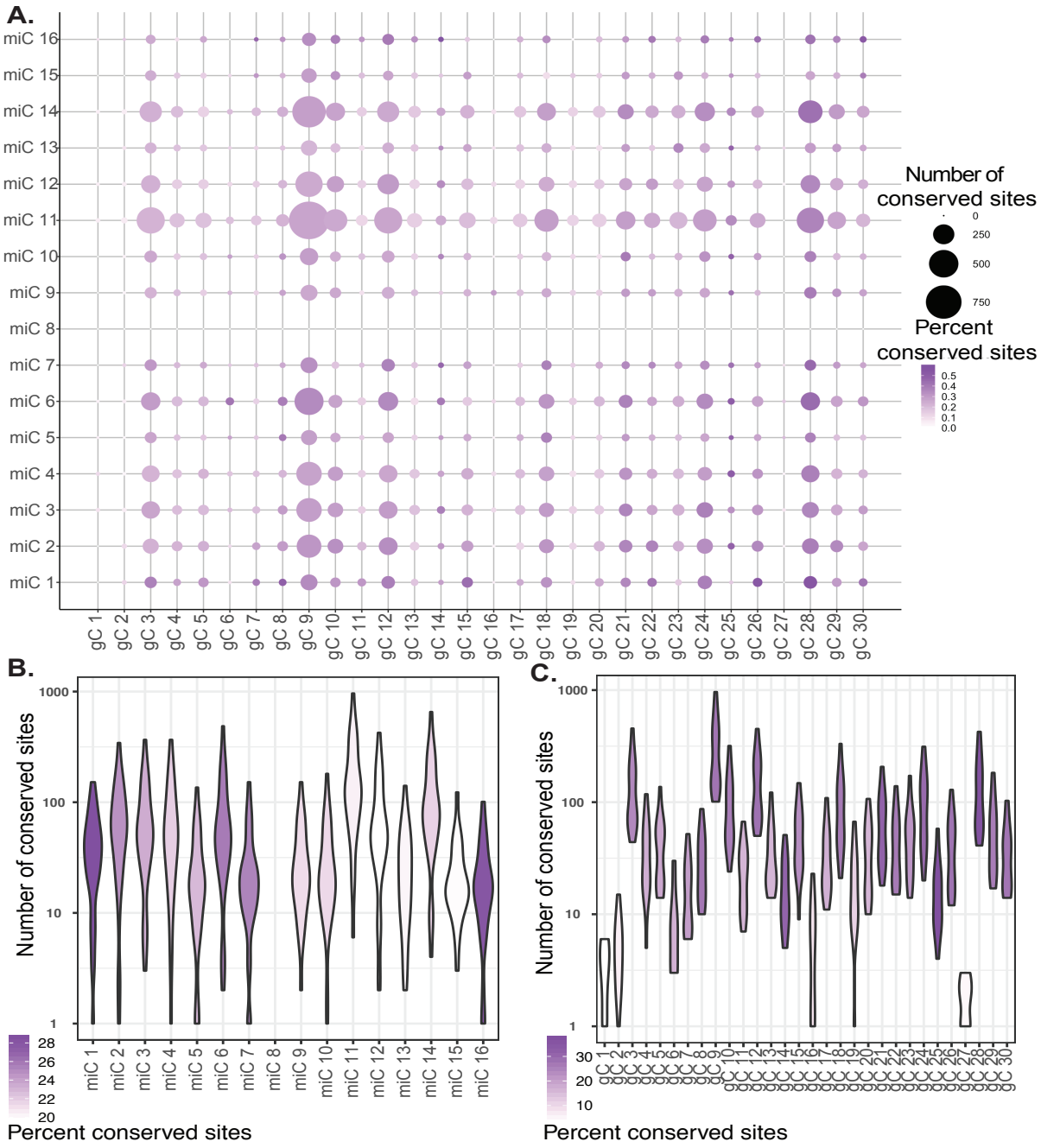
Supplemental Fig. S18: mRNA maSigPro cluster expression profiles. The median expression profiles of 30 mRNA clusters obtained by maSigPro.



Supplemental Fig. S19: Expression anti-correlation analysis: Tissue-wise partial Pearson correlation between miRNAs clusters and each mRNA clusters identifies significant anti-correlations. The highlighted boxes indicate the significant interactions identified by target enrichment analysis and have a negative partial correlation.



Supplemental Fig. S20: Expression profiles of all the significant interactions. The miRNA and mRNA expressions for the miRNA and mRNA clusters in the identified significant interactions.



Supplemental Fig. S21: Conservation of microRNAs target sites. We studied the number of conserved target sites of conserved miRNAs in each miRNA cluster in the 3'UTRs of each gene cluster. A criterion of PhastCons score of minimum 0.9 for at least 4 nucleotides of the target site was used to call the site conserved. Panel **(A)** shows the number of conserved miRNAs as the size of the points and the percentage of target sites conserved as the color (darker shade corresponds to the higher conserved fraction) **(B)** The analysis of conserved targets in gene clusters shows gene cluster 25 and 29 with the highest fraction of conserved targets while clusters 9 and 12 contains the highest number of conserved sites. (**C**) On the other hand, miRNA clusters 1 and 16 has the highest fraction of conserved targets, whereas clusters 6,11 and 14 have the highest number of conserved targets.

Supplemental Methods

Generation of mouse microRNA-seq libraries

The construction of microRNA-seq libraries was based on the previously published protocol (Roberts et al. 2015; Alon et al. 2011) without the highly abundant miRNA blocking step. Briefly, 500 ng of total RNA with RIN (RNA integrity number) score of higher than 9.0 was used as input material, together with spike-in controls. 3' adapter was ligated to the sample with T4 RNA ligase 2, truncated (NEB Cat. #M0242L), then reverse transcription primer was annealed to the 3' adapter in order to reduce the 5' and 3' adapter dimer. Subsequently, 5' adapter was ligated to the product with T4 RNA ligase 1 (NEB Cat. #M0437M). Here, we used a pool of four multiplex 5' adapters. At the end of the 5' adapter, there is a six-nucleotide spacer, which was present as the first six nucleotides in read 1 of the sequence data in order to provide base diversity during the crucial first cycles. Ligation product was reverse transcribed with Superscript II (Invitrogen Cat. #18064014) and the cDNA was further amplified using Phusion high-fidelity PCR master mix (NEB Cat. #M0530S). Primers used at the PCR stage introduce a barcode, used later for sample demultiplexing. PCR products were purified with Ampure XP beads (Beckman Coulter Cat. #A63881). To get rid of adapter-dimer and the other non-miRNA product, size selection of the microRNA-seq libraries was performed using 10% TBE-urea polyacrylamide gel (Bio-Rad Cat. #4566033) in hot (70C) TBE running buffer for 45 mins. The 140-nt denatured microRNA-seq library band was excised, eluted from the gel slice, precipitated by isopropanol and resuspended with 10ul EB buffer (QIAGEN Cat. #19086).

Generation of mouse NanoString samples

100 ng total RNA was used as starting material. Together with "spike-in" positive and negative controls, each target miRNA was ligated to a specific miRNAtag molecule and the chimeric miRNA:miRNAtag molecule was hybridized with fluorescent-labeled probes overnight. The miRNA:miRNAtag chimeric molecule is long enough to ensure the efficiency and specificity of probe hybridization. After the samples were processed in NanoString nCounter PrepStation to remove unhybridized probes, they were immobilized and aligned in scanning cartridges and scanned in NanoString nCounter digital analyzer with the maximal resolution setting to achieve the counts of each individual target miRNA molecule recognized by probe.

NanoString analysis of the mouse samples

The mouse NanoString samples were prepared with NanoString miRNA kit versions 1.2, 1.3, or 1.5 (based on miRBase v.15) following the manufacturer's protocol (see Supplemental Methods for details).

NanoString raw data for all samples were normalized with NanoStringNorm v.1.2.1 (Waggott et al. 2012) using the function "NanoStringNorm" with the following parameters: CodeCount = 'geo.mean', Background = 'max', SampleContent = 'top.geo.mean', round.values = TRUE, take.log = FALSE.

MicroRNA profiling of human fetal samples

The human fetal samples (weeks 19-40) were received from BioChain and the small RNA was isolated using mirVana miRNA isolation kit. The RNA was treated with DNase I and Ribominus as described in detail in the following protocol: https://www.encodeproject.org/documents/d897e6b3-9aa4-4a42-b43c-f1709b9a4798/@@download/attachment/SID38464.pdf

The libraries were prepared following Illumina TruSeq library protocol and sequenced as 100 bp single-end reads on Illumina HiSeq 2000.

Human short RNA-seq read adapter trimming and mapping

Reads were initially trimmed for TGGAATTCTC adapters and Ns with cutadapt (Martin 2014) using the parameters: "-m 16 --trim-n". In the case of polyA adapters, three additional parameters were used: "-a A{10} -e 0.1 -n 10" (to iteratively remove longer polyA tails).

Trimmed reads were mapped to the human genome (assembly hg38) with STAR v. 2.5.1b with parameters: "--outFilterMultimapNmax 10 --outFilterMultimapScoreRange 0 --outFilterScoreMinOverLread 0 --outFilterMatchNminOverLread 0 --outFilterMatchNmin 16 --outFilterMismatchNmax 1 --alignSJDBoverhangMin 1000 --alignIntronMax 1".

Finally, miRNA hairpins from GENCODE v. 25 were quantified by summing the reads that have 100% overlap with the hairpins.

Mouse microRNA-seq read adapter trimming and mapping

5’ and 3’ adapters were sequentially excised from raw reads with Cutadapt v. 1.7.1 (Martin 2014) and Python v. 2.7.10. The 3’ and 5’ adapter sequences are as follows:

3’_adapter_seq = "ACGGGCTAATATTTATCGGTGGAGCATCACGATCTCGTAT"
5’_adapter_seq1 = "^CAGTCG"
5’_adapter_seq2 = "^TGACTC"
5’_adapter_seq3 = "^GCTAGA"
5’_adapter_seq4 = "^ATCGAT"

The Cutadapt parameters used are as follows: “cutadapt -a 3’_adapter_seq -e 0.25 --match-read-wildcards --untrimmed-output=$NO_3AD_FILE input.fastq | cutadapt -e 0.34 --match-read-wildcards --no-indels -m 15 -O 6 -n 1 -g 5’_adapter_seq1 –g 5’_adapter_seq2 -g 5’_adapter_seq3 -g 5’_adapter_seq4 –untrimmed output=$NO_5AD_FILE --too-short-output = $TOO_SHORT_FILE - > trimmed_reads.fastq”.

Trimmed reads were mapped to mouse miRBase v. 22 (Kozomara & Griffiths-Jones 2011) mature miRNA sequences with STAR v. 2.4.2a (Dobin et al. 2013) with parameters: “--runThreadN 16 --alignEndsType EndToEnd –outFilterMismatchNmax 1 --outFilterMultimapScoreRange 0 --quantMode GeneCounts --outReadsUnmapped Fastx --outSAMtype BAM SortedByCoordinate --outFilterMultimapNmax 10 --outSAMunmapped Within --outFilterScoreMinOverLread 0 --outFilterMatchNminOverLread 0 --outFilterMatchNmin 16 --alignSJDBoverhangMin 1000 --alignIntronMax 1”.

Generation of ab initio transcripts models from mRNA-seq reads

mRNA-seq reads were mapped to the mouse genome (assembly mm10) using STAR v. 2.4.2a with the following parameters: “--genomeDir star.gencodeM10.index --readFilesIn $fastqs –sjdbGTFfile gencode.vM10.annotation.gtf --readFilesCommand zcat --runThreadN 8 --outFilterMultimapNmax 20 --alignSJoverhangMin 8 --alignSJDBoverhangMin 1 --outFilterMismatchNmax 999 --outFilterMismatchNoverReadLmax 0.04 --alignIntronMin 20 --alignIntronMax 1000000 --alignMatesGapMax 1000000 --outSAMunmapped Within --outFilterType BySJout --outSAMattributes NH HI AS NM MD XS --outSAMstrandField intronMotif --outSAMtype BAM SortedByCoordinate --sjdbScore 1”. The alignments to the

genome were assembled into *ab initio* transcripts using StringTie v. 1.2.4 (Pertea et al. 2015) using the following parameters: "-G gencode.vM10.annotation.gtf -c 3 -p 8". The transcript models for each sample were merged using StringTie with the following options: "stringtie --merge -G gencode.vM10.annotation.gtf -o merged.gtf -m 200 -F 1.0 -p 8". Single exon transcripts with no strand information were excluded. The expression levels of the GENCODE M10 and the new StringTie model transcripts were obtained using RSEM v. 1.2.25 (Li & Dewey 2011) with the following parameters: "rsem-calculate-expression --star --star-path ~/STAR-STAR_2.4.2a/bin/Linux_x86_64/ -p 10 --gzipped-read-file fastqs RSEM_Index_GENCODE_M10_Plus_StringTieModels".

Identification of orthologous miRNAs in mouse and human

Orthologous miRNAs between human and mouse were identified through genomic alignment. Human miRNAs were lifted over to the mouse genome and vice versa, with a minimum overlap requirement of 50%. If a human miRNA maps within 10 kb of a mouse miRNA, and that mouse miRNA also maps within 10 kb of the initial human miRNA, those miRNAs are defined to be in reciprocal orthologous relationship.

Differential analysis of human short RNA-seq data from human

MicroRNA hairpins specific of a given organ in human were identified with the glmQLFit() and glmQLFTest() function() from the R package edgeR. We used a deviation coding system for contrasts which compares one organ against all the others, with the function constr.sum(). P-values were adjusted for multiple testing with the Benjamini-Hochberg correction. Only hairpins with FDR<0.01 and a fold-change > 2 were considered tissue-specific.

Identification of tissue specificity of the miRNA clusters:

The average expression of miRNAs in each cluster was calculated for each sampling point (for each tissue at each of the time points). Then the standard deviation of each cluster was calculated for each tissue across different time points. The standard deviations obtained for different tissues were then scaled for each miRNA cluster and the tissues with positive values were considered as the tissue specificity of the miRNA cluster. The code to create this tissue specificity matrix and the corresponding plot is available at:

< https://github.com/sorenar/mouse_embryonic_miRNAs/blob/master/PartialCorrelation.R >

Building the partial correlation matrix:

The average expression of mRNAs in each cluster was calculated for each tissue at all stages. For each pairs of miRNA-mRNA clusters only the sample points corresponding to the tissues identified as specific to the miRNA cluster were used to find the Pearson correlation. The code to generate this partial correlation matrix is provided at: < https://github.com/sorenar/mouse_embryonic_miRNAs/blob/master/PartialCorrelation.R >

Conservation analysis of microRNA targets

In order to investigate the targeting of gene clusters by the miRNA clusters, we looked at the conservation of miRNA target sites within 3'UTR of each gene cluster. We first identified the 8mer seeds of all the miRbase mature miRNAs and looked for their corresponding target sites in the 3'UTRs (with perfect complementarity). We used the PhastCons scores of the Euarchontoglires subset of mm10 multi-alignments (downloaded from UCSC Genome Browser) to determine the conservation of these target seeds. We extracted the 3' UTR regions with the PhastCons score of more than 0.9 and then counted the target sites that have more than 4 nucleotides within that region as conserved and the remaining target sites as non-conserved. Finally, we plotted the total number of conserved target sites and the fraction of target sites conserved for each microRNA-gene cluster pair.

References

Alon, S. et al., 2011. Barcoding bias in high-throughput multiplex sequencing of miRNA. *Genome research*, 21(9), pp.1506–1511.

Dobin, A. et al., 2013. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1), pp.15–21.

Kozomara, A. & Griffiths-Jones, S., 2011. MiRBase: Integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Research*, 39(SUPPL. 1), pp.152–157.

Li, B. & Dewey, C.N., 2011. RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, 12.

Martin, M., 2014. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*.

Pertea, M. et al., 2015. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology*, 33(3), pp.290–295.

Roberts, B.S. et al., 2015. Blocking of targeted microRNAs from next-generation sequencing libraries. *Nucleic Acids Research*, 43(21), pp.1–8.

Waggott, D. et al., 2012. NanoStringNorm: An extensible R package for the pre-processing of nanostring mRNA and miRNA data. *Bioinformatics*.