

The Rat Atlas of Tissue-specific and Enriched miRs (RATEmiRs) is a database web application to query microRNA-Seq (miRNA-Seq) data to measure expression in several tissues obtained from five female and five male Sprague Dawley rats 12-13 weeks in age. Specific details of the analysis are available in Smith et al., 2016. Briefly, independent rats were used to obtain replicate samples of each tissue. Sequencing of the miRNA extracted from the tissues was performed by Illumina HiSeq 2000 analysis generating 50 bp single end reads with 4-5 million reads per sample. Three separate bioinformatics pipelines processed the data as follows with Table 1 denoting the sample sizes for each pipeline:

Eli Lilly

FastQ files were processed to remove adaptor sequence and discard reads < 17bp in length with cutadapt v. 1.4.1 with options `-a TGG AATTCTCGGGTGCCAAGG -quality-base 33 -q 20 -match-read-wildcards -m 17`. All trimmed reads that contained an 'N' were discarded. Identical sequences from the same sample were combined into a single sequence in the form expected by miRDeep2. That is, with ids of the form "NNN_#_x#", where NNN is a unique alphabetic code identifying the sample, the first '#' is a unique integer, and 'x#' records the number of times this sequence occurred in the input for the sample. Quantifier.pl from the mirdeep2 package v. 2.0.0 was used to generate miR alignment files (.mrd files) against known miRs from miRbase 20 with options `-P -p <org>_hairpin.fa -m <org>_mature.fa -r trimmed_reads.fa -d`. The .mrd files were then parsed with a custom Perl script as follows. Each isomiR sequence in an alignment was associated with the corresponding mature miR identifier. If it aligned to a miRNA precursor, but not with an expected mature miR sequence, it was identified as <miR>-pre to indicate this. Usually these correspond with reverse strand miRs that have not yet been annotated as alternate mature forms. Sometimes these appear to be microRNA offset RNAs, and sometimes they appear to represent incompletely processed sequence. If a given sequence was identified as aligning with more than one precursor, it was associated with all potential names as a composite name. That is, a sequence that aligned to both let-7a-5p and let-7f-5p was assigned to the composite mature miR let-7a-5p;let-7f-5p to indicate for subsequent analyses that this identification was ambiguous. All sequences that had not yet been identified against known rat miRs were then looked for in the analysis with respect to known mouse miRs, and so on against human and finally C. elegans miRs. This process allows for identification of conserved rodent miRs that have not yet been annotated in rat, and against conserved mammalian miRs that have not yet been annotated in rat and mouse. In parallel to the above identification of known miRs, we used miRDeep2's novel miR identification process. Additional details are described in the supplementary methods of Smith et al., 2016.

Animal level data was obtained after preprocessing of the raw sequencing data. For each microRNA, there are measurements from 10 animals (5 males and 5 females) of 23 different tissues. These 23 tissues were from 14 organs. In order to find tissue-specific miRNAs, tissue level counts were obtained for each microRNA by summing over all the animals. The tissue level counts were aggregated to organ level by selecting the maximum of the tissue counts to represent the organ count. For example, brain (organ) has four different tissue types (Hippocampus, Brainstem, Cerebellum, and Cerebrum), the maximum of these four tissue level counts will be used to represent the expression levels of brain. After these data processing steps, a count data matrix is generated, where each row corresponds to a miR and each column corresponds to an organ type.

Non-negative Matrix Factorization (NMF) is a commonly used unsupervised learning algorithm to decompose a nonnegative matrix X into two non-negative matrices W and H . In our case, X is the observed count matrix of dimension $N \times 14$, where N is the number of miRs. Each column of W defines a miR group and each column of H represents the expression pattern of the miR group corresponding to the organ type. Based on this decomposition, we can find a representative set of miRs that corresponds to the specific high expression pattern for organ.

To identify tissue-specific miRs, it has to be organ specific (which means that it has to be highly expressed in one organ, and lowly in the rest of the organs). Thus we start from each organ-specific miR, and examine if it is also highly expressed in a particular type of tissue or a few tissues, which belong to the current organ. For example, a miRNA may be specific to brain (organ), but it may only be specific to the hippocampus (tissue). To determine the expression level of each miR, we fitted a two-component mixture of Poisson distribution to the animal level tissue counts data. The component with larger intensity of the Poisson distribution is naturally used to model high counts and the component with lower intensity is used to model background noise. The larger component of the Poisson mixture model naturally corresponds to high expression level, whereas the smaller component of the Poisson mixture model naturally corresponds to the low expression level. For an organ-specific microRNA to be tissue-specific, it has to be expressed at a high level in one tissue of this organ. After the mixture model is fitted and the intensities are estimated, we can determine which tissues the microRNA is expressed in according to their estimated intensities and then declare tissue-specific microRNAs afterwards. Additionally, the raw sequencing data was manually analyzed in each tissue for each miRNA and statistical analysis of the manually identified tissue-specific and enriched miRNAs was then conducted to verify or dispute the tissue-specificity or enrichment.

NIEHS

Raw fastq files with the miRNA-Seq reads were checked for quality using a TQE standard based on trimming the adapters, quality filtering at $Q \leq 20$ and the elimination of short reads < 14 bases long. In some cases when minor adapter sequences (indices) were observed after initial process, additional rounds of processing were applied until all the criteria were met. In the end, reads longer than 25 base pairs were also eliminated, resulting in quality filtered reads between 14-24 bp. Reads passing the TQE filtering were then aligned to the rat miRBase database v.19 using BWA version 0.6.1-r104q with “-n 2” and default parameters otherwise. Only perfect matches were further used in the quantification, which was performed by summarizing read count level measurements at each mature miRBase miRs. Data points with zeros were imputed with 1.0. Total counts for each miR were normalized by transcripts per million reads (TPM). Prior to statistical analysis, a transformation from floats to integers was performed by ceiling the data.

To detect tissue-enriched miRNAs, a one-vs-rest strategy was adopted when applying statistic models. “Differentially” expressed miRs (DEMiRs) in one tissue against the combination of all other tissues were identified using a quasi-Poisson approach (Quasi-Seq) with shrunken dispersion estimates. During the

modeling process, the offset was computed on the 3rd quartile. Significant DEMiRs as tissue-enriched were detected at a nominal p-value $< \alpha$ and with a positive difference.

To detect tissue-specific miRNAs, a percentile criteria was used to select any tissue-enriched miR which had a mean expression $> a\%$ of the max mean expression from any of the other tissue.

Maastricht

mirDeep2 was used to predict putative novel miRNAs. The raw fastq files from the rat samples were mapped to the Rat genome (version 5.0.73 from Ensembl). miRDeep2 output was then parsed to keep only the 496 predicted miRNA precursors with a score of 1 or above. Raw reads were trimmed off the 3' adapter sequence. The parameters were set to look for a perfect match with the first 8 bases of the 3' adapter to consider a hit. The trimmed sequences with a size below 16 or above 35 bp were discarded. During this trimming step, every identical sequence was pulled and each iteration was summed. The trimmed reads were then mapped to the list of rat precursor miRNA obtained either from miRBase, or generated de novo from the miRDeep2 prediction using a fast short read mapping software PatMaN, configured to allow no mismatches or gap. PatMan output was then parsed to 1) assign a unique name to each unique sequence consisting of the pre-miRNA name, the pre-miRNA arms (5p or 3p) and the mapping coordinates (for instance, miR-127_3p_57_78), 2) divide the total read count of each unique sequence by the number of assigned loci in the complete mapping and 3) create a mature output table by pulling the different isomiR count for each miRNA species (i.e. 5p and 3p). The parsed PatMaN output were normalized with the trimmed mean of M-values method (TMM) and filtered to remove all miRNAs where the normalized read counts were fewer than 10 reads in all the 215 samples. Total read counts were then computed for each individual miRNA.

A miRNA was called tissue-specific or tissue-enriched when the percentage of reads for a single tissue (or tissue group) was respectively higher than 90 or 50 % of the total read count.

In order to identify isomiRs, the raw count number of each given isomiR was first converted to the percentage of expression compared to the mature count observed for its corresponding miRNA. By comparing this percentage, we were able to report the miRNA for which the most expressed isomiRs differs between tissues.

Table 1. Sample sizes of each tissue for each pipeline

Tissues	Pipeline sample sizes		
	Lilly	NIEHS	Maastricht
Adrenal	10	10	10
Muscle biceps [‡]	10	10	10
Brainstem [†]	10	10	10
Cerebellum [‡]	10	10	10

Cerebrum [†]	10	10	10
Cortex [#]	10	10	10
Dorsal root ganglion (DGR/Uk)	10	4	10
Duodenum ⁺	10	9	10
Stomach glandular (Gln)*	10	10	10
Heart	10	10	10
Hippocampus [†]	10	10	10
Ileum ⁺	10	8	10
Jejunum ⁺	10	10	10
Kidney [#]	10	10	10
Liver	10	10	10
Medulla [#]	10	10	10
Stomach non-glandular (NGln)*	10	10	10
Ovary	5	5	5
Pancreas	10	10	10
Muscle soleus [‡]	10	10	10
Testicle	5	1	5
Uterus	5	5	5
Whole Blood	10	6	10

Denotation of tissues that comprise of an organ:

#:Kidney; *: Stomach; +: Intestine; †: Brain; ‡: Muscle