Genomics Data 7 (2016) 46–53

Contents lists available at ScienceDirect

Genomics Data

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

# *In silico* genome wide mining of conserved and novel miRNAs in the brain and pineal gland of *Danio rerio* using small RNA sequencing data

Suyash Agarwal [a], Naresh Sahebrao Nagpure [a], Prachi Srivastava [b,*], Basdeo Kushwaha [a], Ravindra Kumar [a], Manmohan Pandey [a], Shreya Srivastava [a]

[a] Division of Molecular Biology and Biotechnology, ICAR-National Bureau of Fish Genetic Resources, Lucknow 226 002, Uttar Pradesh, India
[b] AMITY Institute of Biotechnology, AMITY University Uttar Pradesh, Lucknow Campus, Lucknow 226 028, India

ARTICLE INFO

*Article history:*
Received 23 September 2015
Received in revised form 2 November 2015
Accepted 18 November 2015
Available online 26 November 2015

*Keywords:*
MiRNA
NcRNAs
Novel miRs
Phylogenetic analysis
Zebrafish

ABSTRACT

MicroRNAs (miRNAs) are small, non-coding RNA molecules that bind to the mRNA of the target genes and regulate the expression of the gene at the post-transcriptional level. Zebrafish is an economically important freshwater fish species globally considered as a good predictive model for studying human diseases and development. The present study focused on uncovering known as well as novel miRNAs, target prediction of the novel miRNAs and the differential expression of the known miRNA using the small RNA sequencing data of the brain and pineal gland (dark and light treatments) obtained from NCBI SRA. A total of 165, 151 and 145 known zebrafish miRNAs were found in the brain, pineal gland (dark treatment) and pineal gland (light treatment), respectively. Chromosomes 4 and 5 of zebrafish reference assembly GRCz10 were found to contain maximum number of miR genes. The miR-181a and miR-182 were found to be highly expressed in terms of number of reads in the brain and pineal gland, respectively. Other ncRNAs, such as tRNA, rRNA and snoRNA, were curated against Rfam. Using GRCz10 as reference, the subsequent bioinformatic analyses identified 25, 19 and 9 novel miRNAs from the brain, pineal gland (dark treatment) and pineal gland (light treatment), respectively. Targets of the novel miRNAs were identified, based on sequence complementarity between miRNAs and mRNA, by searching for antisense hits in the 3′-UTR of reference RNA sequences of the zebrafish. The discovery of novel miRNAs and their targets in the zebrafish genome can be a valuable scientific resource for further functional studies not only in zebrafish but also in other economically important fishes.



## 1. Introduction

The discovery of the first miRNAs in *Caenorhabditis elegans* in 1993 paved the way for unearthing of thousands of the mature miRNAs in a wide variety of organisms, including animals, plants and even viruses [1–3]. In recent years, intensive studies have been carried out on identification of novel miRNAs and their targets, with the addition of newly discovered miRNAs to miRBase and subsequently resulting in its new version release. The current miRBase release 21 has miR information on 9 fish species. MiRNAs are a distinct class of endogenous RNA molecules, which do not code for any protein and are about 22 nucleotides in length [4]. MicroRNAs (miRNAs) are small, non-coding RNA molecules that function as master regulators of the genome. They bind to the mRNA of the target genes; thus, regulating the gene expression at the post-transcriptional level. Many miR-related discoveries have come from zebrafish investigations [5–10]. After the release of Zv9 (zebrafish genome draft), the zebrafish genome project joined the Genome Reference Consortium (GRC) for further improvement and ongoing maintenance. The GRC has now released a new reference assembly, GRCz10. Correlating the miR functions from a model organism to that of human health largely depends on recognizing true orthologs of human miRs. Thus, for the benefit of the entire miR community, a better understanding of the miRNome is essential [11].

Each miRNA appears to regulate the expression of tens to hundreds of genes to efficiently coordinate multiple cellular pathways. Precursor-miRNAs are usually 60–80 nucleotides in length with a hairpin secondary structure while mature miRNAs are mostly 18–26 nucleotides [12]. Many miRNAs are conserved across vertebrates [13]. Mutation in miRNA genes or the improper miRNA and target gene interaction may become a cause of various genetic diseases. With the advent of high-throughput sequencing technologies, non-conserved or weakly expressed miRNAs, along with species-specific miRNAs can be identified from a wide range of organisms [14–18]. Recent studies have focused on bioinformatic analysis of the NGS data obtained from small RNA sequencing, where algorithms predict miRNA precursor molecules based on the presence of hairpins and other associated parameters and how they are processed into mature miRNAs. This sort of analysis can

* Corresponding author.
*E-mail address:* psrivastava@amity.edu (P. Srivastava).

http://dx.doi.org/10.1016/j.gdata.2015.11.013

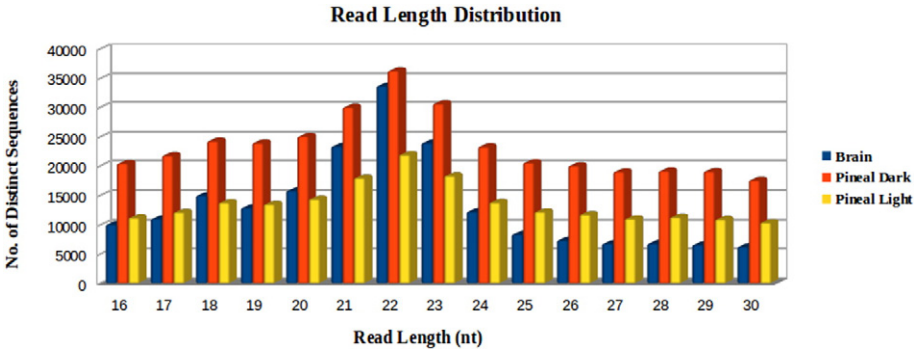


**Fig. 1.** Read length distribution. The figure represents the read length distribution of the 3 data sets after length filtering.

lead to the discovery of both novel and evolutionary conserved miRNAs. Kloosterman et al. [19] reported 66 new miRNAs and 11 star sequences corresponding to 116 potential miRNA hairpins in the zebrafish genome by deep sequencing of two small RNA cDNA libraries. Bizuayehu et al. [20] worked on the Atlantic Halibut and the results indicate a wide conservation of miRNA precursors and involvement of miRNA in multiple regulatory pathways. Despite enormous research on zebrafish, the annotation of miR-producing genes remains limited.

Zebrafish is a fish species of freshwater ecosystem and considered popular organism for studying the gene functions of the vertebrate, especially human development and genetic diseases. It is a favored model organism due to their specific features such as virtually transparent embryos, small size, ability to keep them together in large numbers, ease of breeding and easy to maintain/manipulate/observe in the lab experiments. The critical role of miRNAs in gene expression is highly evident from the recent studies in zebrafish. The miRNAs play key roles in zebrafish organ formation and their expression at different time points.

In the present work, we used Illumina HiSeq2000 small RNA sequencing data from the brain and pineal gland (dark and light treatments) of zebrafish from NCBI SRA database. The total number of reads in the data obtained from NCBI SRA was ~6 million, ~10.4 million and ~14.8 million for the pineal light, pineal dark and brain, respectively. An integrative bioinformatic strategy was applied to detect and analyze the whole miRNA transcriptome of zebrafish. The present study led to the discovery of novel miRNAs in the brain and pineal gland of zebrafish, which will contribute for a better understanding of the role miRNAs play in regulating diverse biological processes.

## 2. Materials and methods

### 2.1. Raw data retrieval and pre-processing

The small RNA sequencing data from three mature miRNA libraries (pineal light, pineal dark and brain) of zebrafish were downloaded from NCBI SRA (*SRX363296*, *SRX363297* and *SRX363298*) and were subsequently used for analysis in the present study. The downloaded data was in SRA format, which was subsequently converted to fastq format using sratoolkit (version 2.3.4-2) [30], fastq-dump option. The obtained data was generated on HiSeq2000 using standard Illumina sequencing workflow with the multiplexing option. A custom Perl script was written to remove low quality bases, adaptor sequences, count the number of occurrences of each read, and eliminate reads outside the targeted size range (≥16 and ≤30).

### 2.2. Identification of conserved miRNAs and other ncRNAs in zebrafish and other fishes

The filtered reads were further aligned onto the latest released version of zebrafish genome GRCz10, using bowtie [31] with two mismatches and zero gaps. Only the aligned reads were used for the downstream analysis. MiRBase [32–33] release 21 was used for annotation of known miRNA. A custom based Perl script was written in order to extract only the fish miRNA from the miRBase, along with the preparation of unique fish miRNA database, and its annotation files. Aligned reads were annotated against the unique fish miRNA database, RefSeq database, as well as noncoding RNA sequences of zebrafish from Rfam (version 11) [34]. A custom Perl script was written in order to extract the best read hit, which depicts the fish miRNA, and to segregate the miRNA hits into known zebrafish miRNA and other fish miRNA.

### 2.3. Differential expression of known miRNAs

The individual read counts of the 3 data sets were fed into a custom based Perl script to prepare the final read count table, which was taken as input for DeSeq [35]. The results were further segregated into up-regulated, down-regulated and neutral miRs based on the log2 fold change value. For the log2 fold change greater than 1, less than −1 and between 1 and −1, the miRs were designated as up, down and neutral miRs. A heatmap of few highly regulated miRNAs was drawn using R script.

### 2.4. Identification of novel miRNA candidates

The reads with no hits in the unique fish miRNA database and Rfam were further used for the prediction of putative new miRs using Mireap (version 0.2). All novel pre-miRNAs were identified based on the presence of a classic hairpin structure [36]. These filtered small RNA reads were aligned with the zebrafish genome using bowtie with strict parameters (number of mismatch; −v = 0). Mireap was used for the detection of novel miRs based on alignment, secondary structure, free energy and location on the precursor arm. The parameters used for Mireap prediction included: i) minimal miRNA length = 18; ii) maximal miRNA length = 26; iii) minimal miRNA reference length = 20; iv) maximal miRNA

**Table 1**
Statistics of the 3 data sets before and after length filtering.

| Parameter | Brain | Pineal gland (light treatment) | Pineal gland (dark treatment) |
|---|---|---|---|
| Total number of sequences | 14,781,569 | 5,933,638 | 10,385,264 |
| Total number of distinct sequences | 259,897 | 265,007 | 463,789 |
| Total number of sequences after length filtering | 14,347,869 | 5,724,361 | 10,008,434 |
| Total number of distinct sequences after length filtering | 200,520 | 204,696 | 351,638 |