**The snoRNAome of zebrafish (*Danio rerio*)**

Renáta Hamar1 and Máté Varga1

*1Department of Genetics, ELTE Eötvös Loránd University, Budapest, Hungary*

ABSTRACT Small nucleolar RNAs (snoRNAs) are one of the most abundant and evolutionary ancient group of functional non-coding RNAs. They were originally described as guides of post-transcriptional rRNA modifications, but snoRNAs fulfill an impressive variety of cellular functions. However, the characterization of the complete repertoire of snoRNAs in a given cellular context and the functional annotation of the zebrafish transcriptome are far from complete. Here, we use size-fractionated RNA sequencing data from adult zebrafish tissues to define the snoRNAome for this species. Our approach allowed us to identify several hitherto unannotated snoRNAs in the zebrafish genome. We created an analysis pipeline where the raw reads were aligned to the latest zebrafish genome assembly (GRCz11) using *Bowtie2* aligner and we identified putative snoRNA sequences using the *blockbuster* algorithm. Using several snoRNA predictor methods we were able to confirm the presence of multiple previously predicted snoRNAs in the data and also identified 68 new snoRNA-like sequences missing from the Ensembl database (v103). Based on our preliminary survey our snoRNAome dataset represents the most reliable set of snoRNAs to date in this species. We also present our results that show the dynamic expression of some snoRNAs during the early stages of zebrafish development and tissue-specific expression patterns for others in adults. Finally, we describe a novel database, snoDanio, which we have developed to facilitate the comprehensive annotation and characterization of small RNAs from transcriptomic data. The current release of snoDanio contains deep sequencing data from 30 RNA libraries from diverse tissues and developmental stage of zebrafish.

1. INTRODUCTION

Small nucleolar RNAs (snoRNAs) form one of the most abundant and ancient group of functional non-coding RNAs. They play critical roles in multiple regulatory processes, including guiding chemical modifications in several RNA classes, affecting the nucleolytic processing of ribosomal RNA (rRNA), the chromatin architecture and alternative splicing to modulate gene expression.

Based on common sequence motifs and structural features snoRNAs are classified in two major families, C/D box and H/ACA box snoRNAs (fig1).

There is emerging evidence showing that snoRNAs can be involved in various critical biological processes. Zebrafish is a fully developed model system being used in a variety of basic research and biomedical studies. Hence, it is an ideal model organism to study the functions and mechanisms of snoRNAs. Interestingly, the ratio of known zebrafish snoRNA genes to the total number of genes is relatively low compared to other species (fig2). This suggests that the annotated snoRNA pool of the species is not yet complete, and that part of the zebrafish snoRNAome is still to be discovered.

2. RESULTS

2.1 Discovery of new unannotated zebrafish snoRNAs in adult tissues and during development

2.1.1 Annotation of de novo small RNA transcripts using read profiles

One of the most common applications of the small RNA-seq method aims the discovery of novel small ncRNAs. Our protocol uses small RNA-seq data that are size-selected to capture only RNAs of the required length. Mapping the small RNA-seq data reveals interesting patterns and each blockgroup should represent a single ncRNA read profile. In fact, by analysing the read profiles extracted this way, new snoRNAs can be predicted (fig3). Finally, we assigned the reads to genes and made a brief overview of the detected biotypes of RNAs. Known snoRNAs were strongly represented in the mean distribution of detected RNA species by tissue type (fig4).

2.1.2 Most of the novel snoRNAs supported by all algorithms are located in coding genes and some of them show tissue or sex or developmental differential expression

The identification of snoRNAs is often difficult due to the lack of overall sequence conservation, small size (60-300 bp) and short sequence motifs. n principle, snoRNA prediction tools can be very different in their methods. Some rely more heavily on purely structural features (cmsearch), while others focus more on sequence similarities (snoreport) and some mix the two approaches (snogps, snoscan). The redetection of snoRNAs known by the largest databases has yielded surprising results. Different programs could detect the described snoRNAs with different efficiencies (sfig1). At present, we believe that the different methods complement each other well and therefore we take the most rigorous approach and consider only the sequences supported by all three tools as snoRNAs. Nevertheless, this topic certainly warrants further investigation in this area. Using several snoRNA predictor methods we were able to confirm the presence of multiple previously predicted snoRNAs in the data and also identified 68 well supported new snoRNA-like sequences missing from the current Ensembl database (v103). One of them was predicted as belonging to both classes, 27 were C/D box (sfig2A) and 39 were H/ACA box snoRNAs (sfig2B).

Our results show a dynamic expression of some snoRNAs during the zebrafish development (fig5A) and tissue-specific expression patterns for others in adults (fig5B). The expression of one of the newly identified snoRNA genes differs strongly between early developmental stages and adult animals. These suggest that some snoRNAs only function in the early stages of development and others in the adult stage. Another seven newly annotated snoRNAs show strong expression differences in different tissues. Their expression in endodermal tissues, such as the liver and the gut is more similar to each other, and differs markedly from neuroectoderm. Moreover we found some snoRNAs that mainly differs between male and female samples (fig5C). These results suggest that some snoRNAs only function in the early stages of development and others in the adult stage and may have sex-specific expression.

2.2 The vast majority of newly identified snoRNAs are produced from intron-embedded genes

We also compared all putative, novel snoRNAs with ncRNA sets from RNAcentral to identify already annotated zebrafish snoRNAs. We noted that twelve of these were already included in some (non-Ensemble) databases. SnoRNAs are often located in the intronic region of so called host genes. Of the predicted novel snoRNAs 47 are located within host genes and 8 appear in intergenic (fig6).

2.2.1 snoRNAs can be categorized by their abundance into two groups

Next we categorize snoRNAs by their abundance profile or patterns during development. For this reason,weused a metric called the equation of variation. A low computed result of variation corresponds to a more uniform expression across stages, whereas a high correction of variation corresponds to a more enriched expression in either one or a few stages. The derivative of this function gave me a threshold, sowecan separate uniformly expressed snoRNAs from the stage enriched (sfig). These two abundance profiles had a distinct characteristics. The stage specific snoRNAs are mostly encoded in the intergenic region, whereas a uniformly expressed snoRNAs are mostly encoded in protein-coding host gene (fig).

2.2.2 Some snoRNA candidates and their corresponding host genes are co-expressed without significant correlation of their respective levels

The expression of the intron encoded snoRNA is depend on the transcription and splicing of its host gene. So it would be fair to think that the expression of snoRNA is always correlated positively with thte hosts. However its not always the case. Whenwecalculated the pearson correlation between the snoRNA and its host, it was surprising that some snoRNAs are non-correlated or negatively-correlated with their host (fig). Based on the literature,wefound one snoRNA that can control its host through the host protein level. This snoRNA can influence the splicing and promote the NMD of the host if its level is enough high in the nucleus. So we assume, that the regulation of the splicing is the basis of this phenomenon. This is supported by the fact that non-correlating or negatively correlating snoRNAs are almost exclusively found in protein-coding RNAs (fig). Another important explanatory factor could be the biological function of the host. We can see that, there is a clear enrichment for positively-correlating snoRNAs in processes such as RNA metabolism, ribosome biogenesis or ribosomal protein functions (fig). So there are very strong differences between the two groups from this aspect.

2.2.3 NEED TO DO: The snoRNA abundance classes 🡪 distinct RNA levels, target preference, and conservation patterns

2.3 snoDanio: The zebrafish snoRNAome database

In this study, we have updated the zebrafish snoRNAome (snoDanio, link!), an openly licensed resource that facilitates integrative and interactive display and analysis of the expression, evolution, and functions of various snoRNAs by mining thousands of high-throughput sequencing data from tissue and developmental stages of zebrafish (fig).

2.2.4 IDEA: In order to classify the newly identified snoRNA candidates as genuine snoRNAs, it must have its own promoter.

2.3 IDEA: Intergenic snoRNAs as a hallmark of yet unannotated transcriptional units

2.4 IDEA: check some random new snoRNA incorporation into snoRNP complexes (IP + RT-PCR) OR do some ribo-probe experiment

3. DISCUSSION

The combination of multiple annotation algorithms (e.g. snoReport, snoGPS, snoscan and cmsearch) on size-fractionated datasets seems an optimal method for the high confidence identification of snoRNA genes. In summary, using our pipeline, we were able to predict 55 new snoRNA loci in zebrafish genome. Some of these are situated in intergenic regions, but the majority of them are encoded with the introns of protein coding and lncRNA host genes. To date, multiple mechanisms have been identified that explain the effect of snoRNAs on gene expression. Our update of the zebrafish snoRNAome has also indentified multiple snoRNAs with sex, stage and tissue specific expression, suggesting a hitherto underappreciated dimension of gene regulation.

4. MATERIALS AND METHODS

4.1 RNA extraction

Tissue were isolated from adult zebrafish euthanized using ice-cold water. Two male and two female individual fish were sacreficed for isolation of total RNA from brain, gut, heart and liver to produce 4 biological replicants. For two brain sample, was needed to collect RNA an additional individual. Extreme care was taken to avoid contamination to obtain pure homogenous tissue samples. The tissues were repeatedly washed in PBS to remove contaminating debris. TRI reagent.

4.2 RNA library preparation

All samples were subjected to ribodepletion and size selection before sequencing.

4.3 snoRNA annotation pipeline

We describe here a novel, easily accessible and cloud-based pipeline, which greatly simplifies the identification of new snoRNA candidate sequences. Our pipeline is available on the Galaxy web platform via this link: (link). Briefly, HISAT2 is the most popular splice-aware alignment algorithm, so we use it to perform alignments directly to the genome. Bowtie2, on the other hand, is designed to map reads in their continuum to the indexed reference. So the former is used to align to the whole genome, the latter to the transcriptome for best results. we considered searching for and filtering out rRNA reads, but snoRNAs are processed from rRNA, so we preferred to keep this fraction as well. Bowtie2 mapped the reads to the genome in -sensitive-local mode (https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-015-2221-x). There are some people who select out reads that match rRNA (https://academic.oup.com/gbe/article/8/3/840/2574143) but since rRNA can be excised from snoRNA (link), we did not take this step. The mapping stat of bowtie2 was quite low, question if this is just a by-product or if small RNAs created by alternative splice-ing exist in this fraction for a long time. we merged these files with the blockCluster algorithm, which also characterized the amount of reads. wes et the filtering threshold to a minimum depth of 100read per cluster and set the distance between clusters to 50bp, as suggested in a previous study. (Reducing struct article) In the consolidated file, we defined blocks from these reads and clusters from the blocks using blockbuster. we extracted the sections that overlapped with the annotations downloaded from ensembleV103.wedid this for both time-series and tissue (both tRNA-seq) data. We merged the overlapping clusters by their genomic coordinates. For the 60-300 long ones, we downloaded the genomic sequence and ran it through trnascan and the 3 known snoRNA scan applications to detect the corresponding sequences based on their structure. This method is good at detecting false positives, but it is worth analysing the transcriptome-matched portion of the sequencing to see if previously unknown snoRNA variants of genes described as pre-miR or lnc-RNA exist. So, for these reads, we also need to look at the read profile of each gene to determine whether processed snoRNAs are present whose annotation is obscured by the parental non-coding gene. Based on the results of the rnacentral sequence alignment run on 169 sequences (supp-table), we removed those that contained sequences that were at least 95% identical to zebrafish non-coding sequences and could be identified as snoRNA.

4.4 Different types of RNA-seq data processing pipelines

Our database includes RNA sequencing data of different types and from different laboratories to study the expression of each snoRNA. For consistency, for each data set, we performed the alignment starting from the raw data and mapped it to the most recent zv11 genome. We expand the transcriptome table with the new snoRNAs and used this current version for the analysis. We created our own galaxy pipeline for each type of RNA-seq experiment (supp-table).