A non-biased framework for the annotation and classification of the non-miRNA small RNA transcriptome

Pantano, L [1], Estivill X [12,*] and Marti E [12*]

1 Genetic Causes of Disease Group, Genes and Disease Program, Centre for Genomic Regulation (CRG), UPF, Barcelona.

2 Centro de Investigacin Biomedica en Red en Epidemiologa y Salud Publica (CIBERESP)

Implementation

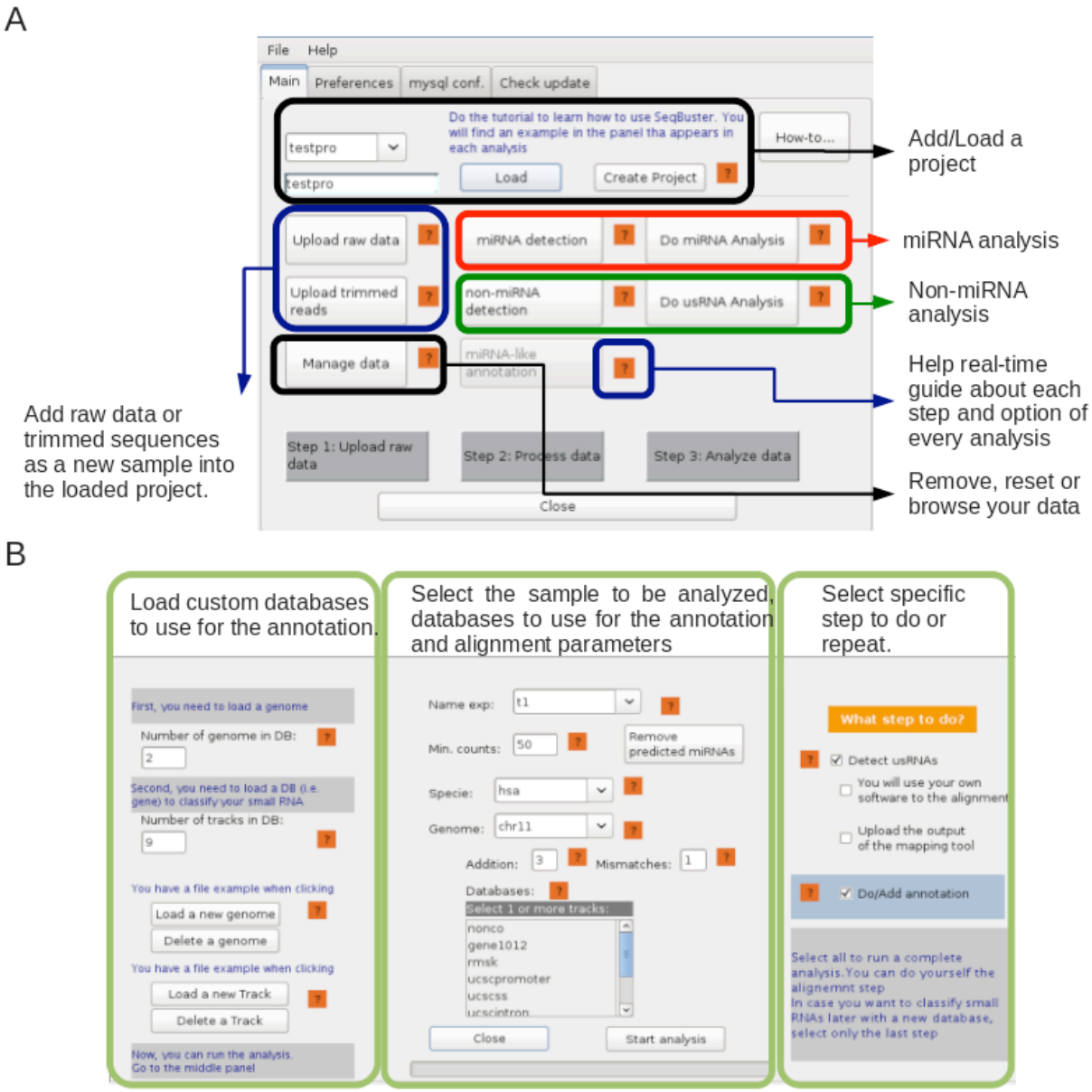
All the steps in the pipeline are automatized and implemented in SeqBuster 2.0 developed in Java platform. The new SeqBuster 2.0 has a main panel showing 5 different sections aimed at: 1) create new projects or load an existent one; 2) upload raw data or trimmed sequences; 3) run miRNA analyses; 4) run non-miRNA analyses and 5) to manage data to reset, remove or browse your processed data. Moreover, a real-time guide has been integrated offering help for each option and step. Help options display a short guide on how to perform a simple analysis of your data (Supplementary figure 1A). Specifically the SeqCluster extension panel that appears after clicking on the 'non-miRNA detection' button, is divided in 3 sections (Supplementary figure 1B): 1) load new genomes and annotation files; 2) set parameters for the analysis, such as the sample to be selected or the genome to use; and 3) select which steps to do for sRNA detection (or repeat in case of re-analysis).

Fasta files containing the custom genome may be loaded to the tool. Moreover, any track (genes, repeats, mRNA,etc.) corresponding with a loaded genome can be integrated into the extension, permitting a complete custom analysis of the non-miRNAs small RNA transcriptome. In the case of UCSC gene tracks the extension extracts different types of information, including the transcription start site, splicing sites, exons, introns and promoters.

External sources are needed to have a complete functionality: blast repository is required for the mapping step, MySQL for the data storage and R statistical packages (standard, RmySQL and RXML) for posterior analysis.

The time consumed is directly related to the second process in the framework where sequences are mapped onto a genome. The size of the genome and

the number of cores used here are the determinants for the total time that one sample requires to be completely analyzed. Each sample was processed in 4 hours, using a workstation (hp xw9300) with 8Gb of RAM memory and 4 cores, and the mapping step consumed 3 of the 4 hours.



Supplementary figure 1:. SeqBuster and SeqCluster integration

SeqBuster 2.0 and SeqCluster extension interface is developed using Java platform (1.6 version). A) SeqBuster panel shows 5 different sections for: 1) creating new projects (black upper box); 2) uploading raw data or trimmed sequences (blue box), 3) miRNA analysis (red box), 4) non-miRNA analysis (green box); and 5) managing data to reset, remove or browse

your processed data (black lower box). Moreover, a real-time guide has been integrated offering help for each option and step of each analysis. B) SeqCluster extension panel is divided in 3 sections for: 1) loading new genomes and annotation files (left box); 2) setting parameters for the analysis, such as sample to be selected, genome to use ... (middle box); and 3) selection of steps to do for the non-miRNA detection (or to repeat in case of re-analysis) showed in the right box.

miRNA detection

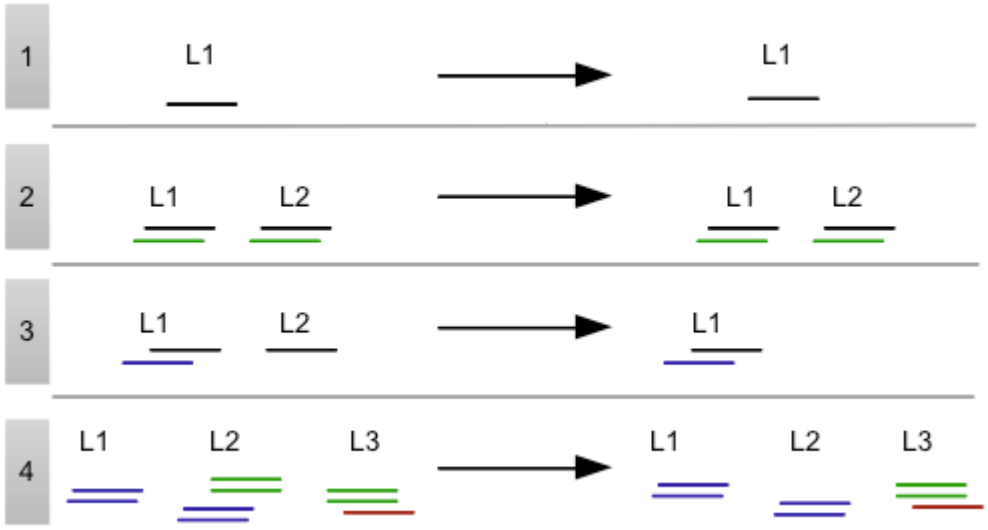
Each fragment of 8 nt in length (seed) that appears on read sequences will be saved in a hash table structure to speed up the searching process. Then, the miRNA precursor sequences are read using windows of 8 nucleotides in order to find the previous stored seeds. When a seed is spotted at a precursor sequence, the read sequence that contains such seed and the current precursor region are fully compared to decide whether this position can be recorded as a hit of the given read sequence. A hit will be considered if the read sequence is found on the precursor allowing 0/1 mismatch and up to 3 nt trimmed at the 3' end of the sequence without matching the precursor sequence. The output stores the best hits for each read sequence with additional information about variation of the sequence when compared to the annotated miRNA. The algorithm takes 1 minute to map 200,000 sequences onto miRBase database, retrieving 99% of concordance with the blast tool (Supplementary table 1). A command-line version of this module has been released for users only interested in this step. This module needs a fasta file with the sequences to be mapped, the precursor sequences in fasta format and the position of the miRNAs on the precursors (files named hairpin.fa and miRNA.str in the miRBase repository, respectively). After that and as an optional step, the user may load data from any custom miRNA prediction pipeline to remove these sequences from the analysis. The unique information needed is the sequences predicted as putative miRNAs.

Program	Num Sequences	Time	Annotated	Pre-indexation
Blast	250000	65 m	17550	YES
ZOOM	250000	6 s	16351	NO
SeqBuster v2	250000	15 s	21350	NO

Supplementary table 1: Comparison of mapping tools. Blast was used with the following parameters: word size = 7 and threshold identity= 85%. Zoom was run allowing 3 mismatches. The last column refers to the needed of database index creation previously to the mapping step.

Decision algorithm

When two pre-usRNAs map on two different locations and these pre-usRNAs only share one of the two loci (inconsistency), they go through a recursive algorithm to converge to a consistent usRNA (all sequences in a usRNA cluster share all the genome location where they map). The algorithm uses the percentage of overlap between pre-usRNAs as the main parameter to solve the inconsistency and select the more realistic scenario. This module will retrieve the minimal number of usRNAs made of the maximum number of pre-usRNAs with a consistent genome location, using as the main parameter the percentage of overlap between pre-usRNAs. A score will be assigned indicating the reliability of the usRNA according to the number of cycles needed to solve inconsistency, giving 1 to the best score and increasing the value proportionally to the complexity of solving the cross-mapping event. For instance, in the latter case, where two pre-usRNAs have two genome locations and only one shared by both, the algorithm will assign the common location as the unique genome hit removing the remaining locations (see Supplementary figure 2). When pre-usRNAs are not solved by the previous modules, they are ignored for downstream analyses due to an inconsistent common origin. This algorithm solves the problem generated by cross-mapping events, but the expression related to each locus is not approached. When using differential expression analysis, the expression of each cluster will be compared between samples ignoring the number of loci that each cluster has associated.



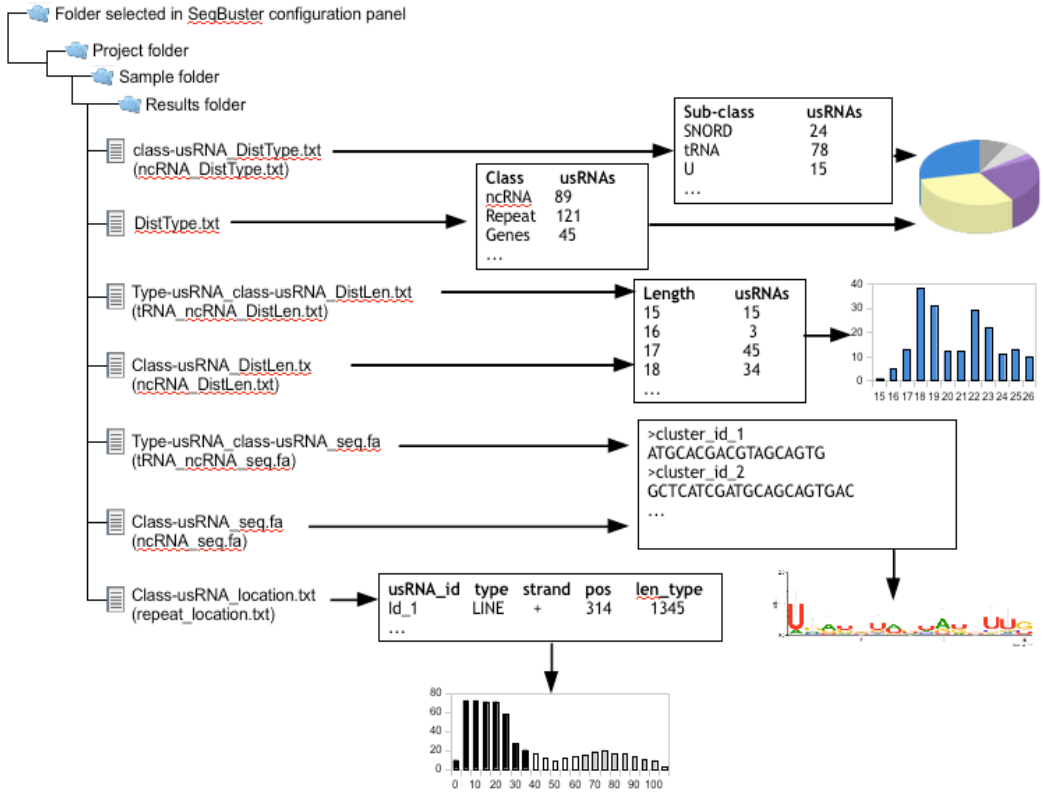
Supplementary figure 2: usRNA definition. Pre-usRNAs are represented by horizontal lines of different colors indicating different pre-usRNAs. Loci are indicated by the label "L" +"Number". 1) One pre-usRNA mapping into one location will be defined as one consistent unambiguous usRNA. 2) Two pre-usRNAs mapping in two locations (ambiguous usRNA), and sharing these two locations will be merged and defined as one consistent, ambiguous usRNA. 3) Two pre-usRNAs mapping in two locations, and sharing one location will be solved recursively until achieve the more realistic situation. In this case, the two pre-usRNAs will be merged into one consistent unambiguous usRNA. 4) A more complex case of the previous situation showing inconsistency. In a consistent scenario the three pre-usRNAs should map on the same locations, however, the short sequences produce wrong mapping that may be corrected using nearest sequences information (considering pre-usRNAs mapping on the same region). Here, green group merges with the red group due to a higher overlap with red sequence than with purples sequence. As a consequence, one of the locations (L2) will be removed from the green pre-usRNA information since the more realistic scenario is that red and green pre-usRNAs go together. At the end, this group of pre-usRNAs will be defined as two consistent usRNAs, one mapping on L3 and the other on L1 and L2.

usRNA Classification

The challenge in this step of the analysis is to avoid the exclusion of us-RNAs showing multiple locations onto the genome. This is, for instance, an important problem in miRNAs detection since many currently known miRNAs (45 human miRNAs) share genome location with TEs. This produces multiple locations in the annotation step and consequently they are not detected using a normal framework. For this reason, SeqCluster extension has integrated a final step to classify those usRNAs having multiple genome locations. In this step the consistency context is studied, meaning that if all the locations share the same context, that usRNA is directly classified. For instance, if all the locations of an usRNA map onto a ncRNA, that usRNA is labeled as ncRNA-usRNA. On the contrary, if the consistency is not observed and the usRNAs map onto several types of databases, usRNAs remain unclassified but not removed from the database, thus permitting the visualization and analysis of these specific cases. Furthermore, each class of genome context may be divided in several groups to extend the classification to more specific subtypes and etc provide a complete view of the genome context for each usRNAs. For instance, transposon elements may be divided into: LINE (long interspersed repetitive DNA), SINE (short intersperded repetitive DNA) or LTR (long terminal repeat).

Output scheme

The framework generates a MySQL table showing, for each usRNA, a unique identifier, number of reads, number of locations, coordinates and the identification according to the annotation step (Supplementary figure 3).



Supplementary figure 3: Output scheme: All files generated by SeqBuster will be stored in a folder that the user configured the first time the program is launched. For each project one folder will be created and for each sample added to the project another folder inside the project folder will appear. All the results will be stored in the sample folder, in the result section. For the non-miRNA sRNA analysis the following files will appear containing: proportion of usRNAs classified in each class/ sub-class with the suffix 'DistType.txt', length distribution of the usRNAs for each class/sub-class with the suffix 'DistLen.txt', usRNAs sequences for each class/sub-class with the suffix 'seq.fa', and finally the location of each usRNAs in their corresponding source molecule when they are classified to some group with the suffix 'location.txt'.

SeqCluster performs a general analysis, offering standard outputs using the previous table with the following graphic and plain-text files: distribution of the different types of usRNAs, length distribution for each type of usRNAs,

position in the putative precursors of each usRNAs type, and fasta files divided by types of usRNAs. Additionally, a BED file compatible with genome browsers will be generated with all the information to allow a user-friendly visualization and better comprehension. Moreover, the expression profile for each usRNA cluster can be visualized to further understand its biogenesis. If the sRNAs that form part of a cluster cover a big region and the expression profile of the different sequences fits a flat distribution (meaning that all the sequences that cover this region are equally represented) these sRNAs may constitute RNA degradation products. Otherwise, if the expression profile covers a small region, and its distribution shows a peak (some members of the cluster are highly expressed, similarly to the expression profile of sequences in a miRNA gene), these sRNAs could be functional sRNAs. However, since mRNA structure appears to influence its stability (Sendler et al., 2011), a possibility exists that a peak distribution reflects degradation products at regions corresponding to less stable, non-stem structures of mRNAs. In order to further study this, SeqCluster generates secondary structures (ss) visualization using RNAfold (Hofacker *et al*, 2003) considering approximately 40 nt upstream and 40 nt downstream of the context in which the usRNA maps. In the case of usRNA mapping on genes, SeqCluster permits a visualization of all usRNAs mapped on the same gene (in HTML format) to help in the decision on whether small RNAs may be RNA degradation products. In this analysis, the ss of a usRNA is compared with the ss of 100 random sequences within the mRNA. The ss of the 100 random sequences are compared between them. SeqCluster generates a graphic for each usRNA showing the distribution of the ss distance scores between the usRNA and the 100 random sequences (group 1), and the distance scores between the 100 random sequences (group 2). Similar group 1 and group 2 distance score distributions (determined using the t-test) indicate that the ss of the usRNAs does not differ from any random sequence of the same gene, which may suggest that the ss is not the unique factor influencing on the generation of that usRNA.

Another factor that could influence the detection of an sRNA is the sequence

depth or sequencing coverage. If the sequencing depth is low, mRNA degradation fragments could show peaks covering a small region sRNAs derived from low expressed mRNA. Thus, SeqCluster includes a p-value (detection score) for each usRNA to provide information about the probability to detect a usRNA. The p-value is calculated using binomial statistics, taking into consideration the frequency of the usRNA and sequencing coverage of the experiment.

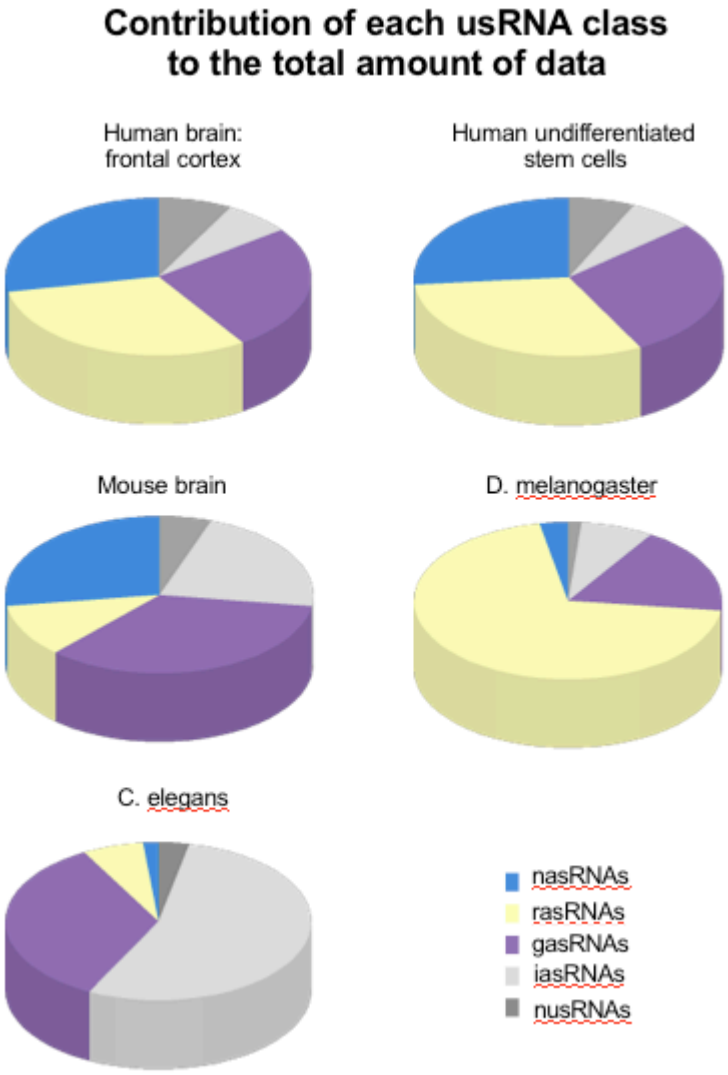
SeqCluster permits differential expression analysis between two samples or two groups of samples in different biological contexts to highlight those sRNAs with possible relevant functions. Published methods previously used for miRNA differential expression were integrated for usRNA differential expression: Binomial statistics, Bayesian method and z-score statistics (Reinartz et al., 2002; Berninget et al., 2008; Creighton et al., 2009). We adapted these analyses for multiple samples in case/control experiments, considering only usRNAs that did no vary within a group. All samples of each group are compared between them in pairs using the selected method, and a ratio and p-value is calculated for each usRNA. If the comparisons of a given usRNA do not have a significant differential expression across all or a subset of the group (parameter selected by the user), that usRNA is kept for the comparison with the other group. usRNAs have to pass those filters in both groups to be included in the analysis. In addition, SAMseq strategy (Fahlgreen et al, 2009) was also implemented, were a R package for array analysis was adapted to the sequencing data. Moreover, two methods for the comparison of the means were implemented: t-test and one-way-ANOVA (analysis of variance). These methods provide also the parameter 'effect of sample sizes' to determine the power of the statistical test (where 0 is small effect and 1 is high effect). Finally, datasets involving time series can be also analyzed using specific statistical methods to determine which sRNAs correlate with a selective time point according to Somel *et al.*, 2010.

In all the cases usRNA frequencies are considered as RPM (reads per million) values, where the frequency of each usRNA is divided by the total amount of reads mapped on the genome multiplied by 1 million.

SeqCluster application to real datasets

We have applied SeqCluster extension to human brain samples sequenced by illumina 1G in our previous work (Martí *et al.*,2010). These corresponded to the frontal cortex (FC) and the striatum (ST). Furthermore, we also used published data from different species sequenced through the same technology to avoid methodology biases. The data selected were human stem cells (SC) (Morin *et al.*,2008), mouse cell lines (MB) (Tam *et al.*,2008), D.melanogaster embryos (DM) (Chung *et al*., 2008) and C. elegans cells (CE) (Batista *et al*., 2008).

The usRNAs were classified according to genome content: nasRNAs when mapping to non-coding, rasRNAs when mapping on repeat elements, gasRNAs when mapping on genes, iasRNAs when mapping on inter-genic regions and nusRNAs if they were not classified (Supplementary figure 4). In mammals, nasRNAs were highly represented with a 30% of the total data, followed by rasRNAs in human samples but not in mouse, where the proportion of this type is lower. In mouse, the second most abundant class was represented by gasRNAs (40%). In D.melanogaster and C.elegans, nasRNAs were poorly represented (6%), with the more important class being rasRNAs for D.melanogaster (75%) and iasRNAs for C.elegans (50%). Deeper analyses were run by SeqCluster, retreiving the more abundant classes in nasRNAs. In this case, tRNAs (78 uRNAs) and SNORD (45 uRNAs) mapping on the sense transcript were the more abundant, suggesting that those molecules could generate small RNAs, as previously described described (Kiss, 2002; Bachellerie *et al.*, 2002; Niwa and Slack, 2007; Luo and Li, 2007; Weber, 2006; Smalheiser and Torvik, 2005; Scott *et al.*, 2009; Kawaji *et al*., 2008; Taft *et al*., 2009a; Ender *et al.*, 2008; Ono *et al.*, 2011). The LINE class was the more represented type (53 uRNAs) among the rasRNAs, with 36 complementary to this TE sequences and 17 lying on sense transcripts. Other minor classes of usRNAs have been found with particular features. Among them, we highlight usRNAs that lay onto exon/exon gene junctions as reported previously (Affymetrix ENCODE Transcriptome Project, 2009). Other types of usRNAs were annotated at the splicing site region of the genes, also described in previous work (Taft *et al.*, 2010).



Supplementary figure 4: General results of SeqCluster extension

The proportion of each usRNA classes in each sample is represented here. The classes used were gasRNA (genomic region), nasRNA (non-coding RNA), rasRNA (repeat elemetns), iasRNAs (inter-genic regions) and nusRNAs (unclassified). The samples showed are 1) frontal cortex from human, 2) undifferentiated stem cells from human, 3) cell lines NIH3T3 from mouse, 4) adult female head from D. melanogaster and 5) larves from C.elegans. (See methods fro further information). Striatum from human brain showed a more equal distribution than the frontal cortex sample.

References

Affymetrix ENCODE Transcriptome Project and Cold Spring Harbor Laboratory ENCODE Transcriptome Project (2009). Post-transcriptional processing generates a diversity of 5'-modified long and short rnas. Nature, 457(7232), 1028–1032.

Bachellerie, J. P., Cavaille´, J., and Hünttenhofer, A. (2002). The expanding snorna world. Biochimie, 84(8), 775–790.

Batista, P. J., Ruby, J. G., Claycomb, J. M., Chiang, R., Fahlgren, N., Kasschau, K. D., Chaves, D. A., Gu, W., Vasale, J. J., Duan, S., Conte, D., Luo, S., Schroth, G. P., Carrington, J. C., Bartel, D. P., and Mello, C. C. (2008). Prg-1 and 21u-rnas interact to form the pirna complex required for fertility in c. elegans. Mol Cell, 31(1), 67–78.

Berninger P, Gaidatzis D, van Nimwegen E, Zavolan M. Biozentrum, Computational analysis of small RNA cloning data. Methods. 2008 Jan;44(1):13-21.

Carninci, P., Sandelin, A., Lenhard, B., Katayama, S., Shimokawa, K., Ponjavic, J., Semple, C. A., Taylor, M. S., Engstro m, P. G., Frith, M. C., Forrest, A. R., Alkema, W. B., Tan, S. L., Plessy, C., Kodzius, R., Ravasi, T., Kasukawa, T., Fukuda, S., Kanamori-Katayama, M., Kitazume, Y., Kawaji, H., Kai, C., Nakamura, M., Konno, H., Nakano, K., Mottagui-Tabar, S., Arner, P., Chesi, A., Gustincich, S., Persichetti, F., Suzuki, H., Grimmond, S. M., Wells, C. A., Orlando, V., Wahlestedt, C., Liu, E. T., Harbers, M., Kawai, J., Bajic, V. B., Hume, D. A., and Hayashizaki, Y. (2006). Genome-wide analysis of mammalian promoter architecture and evolution. Nat Genet, 38(6), 626–635.

Chung, W. J., Okamura, K., Martin, R., and Lai, E. C. (2008). Endogenous rna interference provides a somatic defense against drosophila transposons. Curr Biol, 18(11), 795–802.

Creighton CJ, Reid JG and Gunaratne PH. (2009) Expression profiling of microRNAs by deep sequencing. Brief Bioinform.;10(5):490-7..

Hofacker IL. (2003) Vienna RNA secondary structure server. Nucleic Acids Res.;31(13):3429-31.

Kawaji, H., Nakamura, M., Takahashi, Y., Sandelin, A., Katayama, S., Fukuda, S., Daub, C. O., Kai, C., Kawai, J., Yasuda, J., Carninci, P., and Hayashizaki, Y. (2008). Hidden layers of human small rnas. BMC Genomics, 9, 157–157.

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., and Haussler, D. (2002). The human genome browser at ucsc. Genome Res, 12(6),996–1006.

Kiss, T. (2002). Small nucleolar rnas: an abundant group of noncoding rnas with diverse cellular functions. Cell, 109(2), 145–148.

Luo, Y. and Li, S. (2007). Genome-wide analyses of retrogenes derived from the human box h/aca snornas. Nucleic Acids Res, 35(2), 559–571.

Marti, E., Pantano, L., Bañez-Coronel, M., Llorens, F., Miñones-Moyano, E., Porta, S., Sumoy, L., Ferrer, I., and Estivill, X. (2010). A myriad of mirna variants in control and huntington's disease brain regions detected by massively parallel sequencing. Nucleic Acids Res, 38(20), 7219–7235.

Niwa, R. and Slack, F. J. (2007). The evolution of animal microrna function. Curr Opin Genet Dev, 17(2), 145–150.

Morin, R. D., O'Connor, M. D., Griffith, M., Kuchenbauer, F., Delaney, A., Prabhu, A. L., Zhao, Y., McDonald, H., Zeng, T., Hirst, M., Eaves, C. J., and Marra, M. A. (2008). Application of massively parallel sequencing to microrna profiling and discovery in human embryonic stem

cells. Genome Res, 18(4), 610–621.

Reinartz J, Bruyns E, Lin JZ, Burcham T, Brenner S, Bowen B, Kramer M, Woychik R. Massively parallel signature sequencing (MPSS) as a tool for in-depth quantitative gene expression profiling in all organisms. Brief. Funct. Genomic Proteomic. 2002;1:95–104.

Sendler E, Johnson GD and Krawetz SA, (2011) Local and global factors affecting RNA sequencing analysis. Anal Biochem. (in press).

Somel, M., Guo, S., Fu, N., Yan, Z., Hu, H. Y., Xu, Y., Yuan, Y., Ning, Z., Hu, Y., Menzel, C., Hu, H., Lachmann, M., Zeng, R., Chen, W., and Khaitovich, P. (2010). Microrna, mrna, and protein expression link development and aging in human and macaque brain. Genome Res, 20(9), 1207–1218.

Scott, M. S., Avolio, F., Ono, M., Lamond, A. I., and Barton, G. J. (2009). Human mirna precursors with box h/aca snorna features. PLoS Comput Biol, 5(9).
Smalheiser, N. R. and Torvik, V. I. (2005). Mammalian microrna derived from genomic repeats. Trends Genet, 21(6), 322–326.

Taft, R. J., Glazov, E. A., Cloonan, N., Simons, C., Stephen, S., Faulkner, G. J., Lassmann, T., Forrest, A. R., Grimmond, S. M., Schroder, K., Irvine, K., Arakawa, T., Nakamura, M., Kubosaki, A., Hayashida, K., Kawazu, C., Murata, M., Nishiyori, H., Fukuda, S., Kawai, J., Daub, C. O., Hume, D. A., Suzuki, H., Orlando, V., Carninci, P., Hayashizaki, Y., and Mattick, J. S. (2009b). Tiny rnas associated with transcription start sites in animals. Nat Genet, 41(5), 572–578.

Taft, R. J., Simons, C., Nahkuri, S., Oey, H., Korbie, D. J., Mercer, T. R., Holst, J., Ritchie, W., Wong, J. J., Rasko, J. E., Rokhsar, D. S., Degnan, B. M., and Mattick, J. S. (2010). Nuclear-localized tiny rnas are associated with transcription initiation and splice sites in metazoans. Nat Struct Mol Biol, 17(8), 1030–1034.

Tam, O. H., Aravin, A. A., Stein, P., Girard, A., Murchison, E. P., Cheloufi, S., Hodges, E., Anger, M., Sachidanandam, R., Schultz, R. M., and Hannon, G. J. (2008). Pseudogene-derived small interfering rnas regulate gene expression in mouse oocytes. Nature, 453(7194), 534–538.

Weber, M. J. (2006). Mammalian small nucleolar rnas are mobile genetic elements. PLoS Genet, 2(12).