# Computational Tools for Identification of microRNAs in Deep Sequencing Data Sets

Manuel A. S. Santos and Ana Raquel Soares
*University of Aveiro*
*Portugal*

## 1. Introduction

MicroRNAs (miRNAs) are a class of small RNAs of approximately 22 nucleotides in length that regulate eukaryotic gene expression at the post-transcriptional level (Ambros 2004; Bartel 2004; Filipowicz *et al.* 2008). They are transcribed as long precursor RNA molecules (pri-miRNAs) and are successively processed by two key RNAses, namely Drosha and Dicer, into their mature forms of ~22 nucleotides (Kim 2005; Kim *et al.* 2009). These small RNAs regulate gene expression by binding to target sites in the 3′ untranslated region of mRNAs (3′UTR). Recognition of the 3′UTR by miRNAs is mediated through complementary hybridization at least between nucleotides 2-8, numbered from the 5′ end (seed sequences) of the small RNAs, and complementary sequences present in the 3′UTRs of mRNAs (Ambros 2004; Bartel 2004; Zamore and Haley 2005). Perfect or nearly perfect complementarities between miRNAs and their 3′UTRs induce mRNA cleavage by the RNA-induced silencing complex (RISC), whereas imperfect base pair matching may induce translational silencing through various molecular mechanisms, namely inhibition of translation initiation and activation of mRNA storage in P-bodies and/or stress granules (Pillai *et al.* 2007).

This class of small RNAs is well conserved between eukaryotic organisms, suggesting that they appeared early in eukaryotic evolution and play fundamental roles in gene expression control. Each miRNA may repress hundreds of mRNAs and regulate a wide variety of biological processes, namely developmental timing (Feinbaum and Ambros 1999; Lau *et al.* 2001), cell differentiation (Tay *et al.* 2008), immune response (Ceppi *et al.* 2009) and infection (Chang *et al.* 2008). For this reason, their identification is essential to understand eukaryotic biology. Their small size, low abundance and high instability complicated early identification, but these obstacles have been overcome by next generation sequencing approaches, namely the Genome Sequencer™ FLX from Roche, the Solexa/Illumina Genome Analyzer and the Applied Biosystems SOLiD™ Sequencer which are currently being routinely used for rapid miRNA identification and quantification in many eukaryotes (Burnside *et al.* 2008; Morin *et al.* 2008; Schulte *et al.* 2010).

As in other vertebrates, miRNAs control gene expression in zebrafish, since defective miRNA processing arrest development (Wienholds *et al.* 2003). Also, a specific subset of miRNAs is required for brain morphogenesis in zebrafish embryos, but not for cell fate determination or axis formation (Giraldez *et al.* 2005). In other words, miRNAs play an important role in zebrafish organogenesis and their expression at specific time points is relevant to organ formation and differentiation. Since identification of the complete set of miRNAs is fundamental to fully understand biological processes, we have used high throughput 454 DNA pyrosequencing technologies to fully characterize the zebrafish miRNA population (Soares *et al.* 2009). For this, a series of cDNA libraries were prepared from miRNAs isolated at different embryonic time points and from fully developed organs sequenced using the Genome Sequencer™ FLX. This platform yields reads of up to 200 bases each and can generate up to 1 million high quality reads per run, which provides sufficient sequencing coverage for miRNA identification and quantification in most organisms. However, deep sequencing of small RNAs may pose some problems that need to be taken into consideration to avoid sequencing biases. For example, library preparation and computational methodologies for miRNA identification from large pool of reads need to be optimized. There are many variables to consider, namely biases in handling large sets of data, sequencing errors and RNA editing or splicing. If used properly, deep sequencing technologies have enormous analytical power and have been proven to be very robust in retrieving novel small RNA molecules. One of the major challenges when analyzing deep sequencing data is to differentiate miRNAs from other small RNAs and RNA degradation products.

Different research groups are developing dedicated computational methods for the identification of miRNAs from large sets of sequencing data generated by next generation sequencing experiments. miRDeep (http://www.mdc-berlin.de/en/research/research_teams/systems_biology_of_gene_regulatory_elements/projects/miRDeep/index.html) (Friedlander *et al.* 2008) and miRanalizer (http://web.bioinformatics.cicbiogune.es/microRNA/miRanalyser.php) (Hackenberg *et al.* 2009) can both detect known miRNAs annotated in miRBase and predict new miRNAs (although using different prediction algorithms) from small RNA datasets generated by deep sequencing. Although these online algorithms are extremely useful for miRNA identification, custom-made pipeline analysis of deep sequencing data may be performed in parallel to uncover the maximum number of small non-coding RNA molecules present in the RNA datasets.

In this chapter, we discuss the tools and computational pipelines used for miRNA identification, discovery and expression from sequencing data, based on our own experience of deep sequencing of zebrafish miRNAs, using the Genome Sequencer™ FLX from Roche. We show how a combination of a public available, user-friendly algorithm, such as miRDeep, with custom-built analysis pipelines can be used to identify non-coding RNAs and uncover novel miRNAs. We also demonstrate that population statistics can be applied to statistical analysis of miRNA populations identified during sequencing and we demonstrate that robust computational analysis of the data is crucial for extracting the maximum information from sequencing datasets.

## 2. miRNA identification by next-generation sequencing

### 2.1 Extraction of next-generation sequencing data

Next generation sequencing methods have been successfully applied in the last years to miRNA identification in a variety of organisms. However, the enormous amount of data generated represents bioinformatics challenges that researchers have to overcome in order to extract relevant data from the datasets.

We have used the Genome Sequencer™ FLX system (454 sequencing) to identify zebrafish miRNAs from different developmental stages and from different tissues. For this, cDNA libraries are prepared following commonly used protocols (Droege M and Hill B. 2008; Soares *et al.* 2009). These libraries contain specific adaptors for the small RNA molecules containing specific priming sites for sequencing. After sequencing, raw data filtration and extraction is performed using specialist software incorporated into the Genome Sequencer™ FLX system (Droege M and Hill B. 2008). Raw images are processed to remove background noise and the data is normalized. Quality of raw sequencing reads is based on complete read through of the adaptors incorporated into the cDNA libraries. The 200 base pair of 454 sequencing reads provide enough sequencing data for complete read through of the adaptors and miRNAs. During quality control, the adaptors are trimmed and the resulting sequences are used for further analysis. Sequences ≥ 15 nucleotides are kept for miRNA identification, and constitute the small RNA sequencing data.

Other sequencing platforms, such as Illumina/Solexa and SOLiD™, also have specialist software for raw data filtration. DSAP, for example, is an automated multiple-task web service designed to analyze small RNA datasets generated by the Solexa platform (Huang *et al.* 2010). This software filters raw data by removing sequencing adaptors and poly-A/T/C/G/N nucleotides. In addition, it performs non-coding RNA matching by sequence homology mapping against the non-coding RNA database Rfam (rfam.sanger.ac.uk/) and detects known miRNAs in miRBase (Griffiths-Jones *et al.* 2008), based on sequence homology.

The SOLiD™ platform has its own SOLiD™ System Small RNA Analysis Pipeline Tool (RNA2MAP), which is available online (http://solidsoftwaretools.com/gf/project/rna2map). This software is similar to DSAP, as it filters raw data and identifies known miRNAs in the sequencing dataset by matching reads against miRBase sequences and against a reference genome. Although these specialist software packages are oriented for miRNA identification in sequencing datasets they are not able to identify novel miRNAs. For this, datasets generated from any of the sequencing platforms available have to be analyzed using tools that include algorithms to identify novel miRNAs.

### 2.2 miRNA identification from next generation sequencing databases

miRNA identification (of both known and novel molecules) from datasets generated by deep-sequencing has been facilitated by the development of public user friendly algorithms, such as miRDeep (Friedlander *et al.* 2008), miRanalyzer (Hackenberg *et al.* 2009) and miRTools (Zhu *et al.* 2010).

We used miRDeep to identify miRNAs in our sequencing datasets (Figure 1). miRDeep was the first public tool available for the analysis of deep-sequencing miRNA data. This software was developed to extract putative precursor structures and predict secondary structures using RNAfold (Hofacker 2003) after genome alignment of the sequences retrieved by next-generation sequencing. This algorithm relies on the miRNA biogenesis model. Pre-miRNAs are processed by DICER, which originates three different fragments, namely the mature miRNA, the star and the hairpin loop sequences (Kim *et al.* 2009). miRDeep scores the compatibility of the position and frequency of the sequenced RNA with the secondary structures of the miRNA precursors and identifies new, conserved and non-conserved miRNAs with high confidence. It distinguishes between novel and known miRNAs, by evaluating the presence or absence of alignments of a given sequence with the stem loop sequences deposited in miRBase. The sequence with the highest expression is always considered as the mature miRNA sequence by the miRDeep algorithm. All hairpins that are not processed by DICER will not match a typical secondary miRNA structure and are filtered out.

After aligning the sequences against the desired genome using megaBlast, the blast output is parsed for miRDeep uploading. As sequencing errors, RNA editing and RNA splicing may alter the original miRNA sequence, one can re-align reads that do not match the genome using SHRiMP (http://compbio.cs.toronto.edu/shrimp/). The retrieved alignments are also parsed for miRDeep for miRNA prediction. miRDeep itself allows up to 2 mismatches in the 3′ end of each sequence, which already accounts with some degree of sequencing errors that might have occurred.

Reads matching more than 10 different genome loci are generally discarded, as they likely constitute false positives. The remaining alignments are used as guidelines for excision of the potential precursors from the genome. After secondary structure prediction of putative precursors, signatures are created by retaining reads that align perfectly with those putative precursors to generate the signature format. miRNAs are predicted by discarding non-plausible DICER products and scoring plausible ones. The latter are blasted against mature miRNAs deposited in miRBase, to extract known and conserved miRNAs. The remaining reads are considered novel miRNAs.

In order to evaluate the sensitivity of the prediction and data quality, miRDeep calculates the false positive rate, which should be below 10%. For this, the signature and the structure-pairings in the input dataset are randomly permutated, to test the hypothesis that the structure (hairpin) of true miRNAs is recognized by DICER and causes the signature.

miRanalizer (Hackenberg *et al.* 2009) is a recently developed web server tool that detects both known miRNAs annotated in miRBase and other non-coding RNAs by mapping sequences to non-coding RNA libraries, such as Rfam. This feature is important, as more classes of small non coding RNAs are being unravelled and their identification can provide clues about their functions. At the same time, by removing reads that match other non coding RNA classes, it reduces the false positive rate in the prediction of novel miRNAs, as these small non coding RNAs can be confused with miRNAs. For novel miRNA prediction, miRanalizer implements a machine learning approach based on the random forest method, with the number of trees set to 100 (Breiman 2001). miRanalyzer can be applied to miRNA discovery in different models, namely human, mouse, rat, fruit-fly, round-worm, zebrafish and dog, and uses datasets from different models to build the final prediction model. In comparison to miRDeep, this is disadvantageous as the latter can predict novel miRNAs from any model. All pre-miRNAs candidates that match known miRNAs are extracted from the experimental dataset and labelled as positive instances. Next, an equal amount of pre-miRNA candidates from the same dataset are selected by random selection with the known miRNAs removed and labelled as negative. Pre-processing of reads corresponding to putative new miRNAs includes clustering of all reads that overlap with the genome, testing whether the start of the current read overlaps less than 3 nucleotides with the end position of previous reads. This avoids DICER products grouping together and be considered non-miRNAs products, which would increase false negatives. Besides, clusters of more than 25 base pairs in length are discarded and the secondary structure of the miRNA is predicted via RNAfold (Hofacker 2003). Structures where the cluster sequence is not fully included and where part of the stem cannot be identified as a DICER product are discarded.

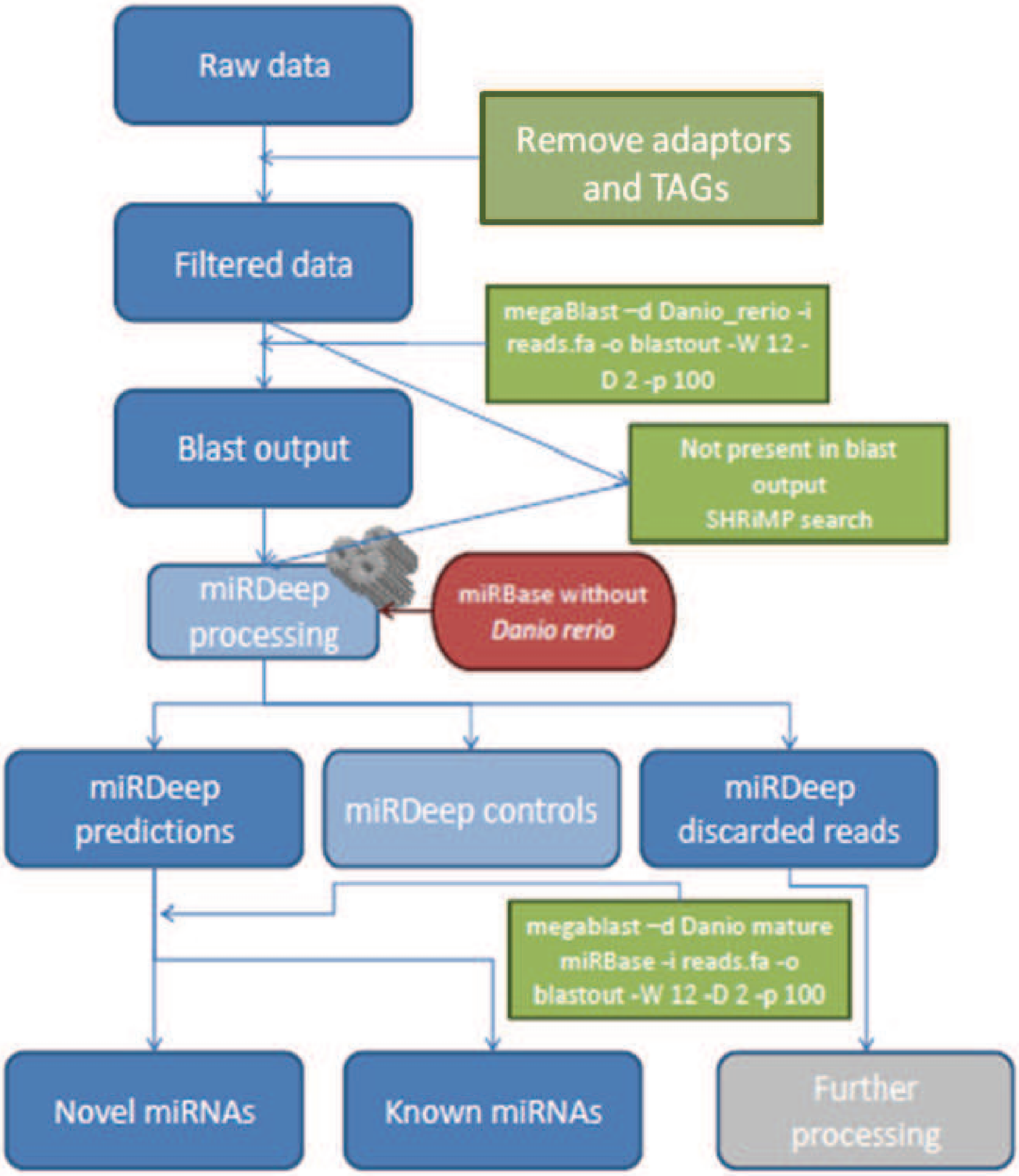

Raw data
Remove adaptors and TAGs
Filtered data
megaBlast –d Danio_rerio -i reads.fa -o blastout -W 12 -D 2 -p 100
Blast output
Not present in blast output
SHRiMP search
miRDeep processing
miRBase without Danio rerio
miRDeep predictions
miRDeep controls
miRDeep discarded reads
megablast –d Danio mature miRBase -i reads.fa -o blastout -W 12 -D 2 -p 100
Novel miRNAs
Known miRNAs
Further processing

Fig. 1. Data pipeline analysis using miRDeep.

miRTools is a comprehensive web server that can be used for characterization of the small RNA transcriptome (Zhu *et al.* 2010). It offers some advantages relative to miRDeep and miRanalyzer, since it integrates multiple computational approaches including tools for raw data filtration, identification of novel miRNAs and miRNA expression profile generation. In order to detect novel miRNAs, miRTools analyze all sequences that are not annotated to known miRNAs, other small non-coding RNAs and genomic repeats or mRNA that match the reference genome. These sequences are extracted and their RNA secondary structures are predicted using RNAfold (Hofacker 2003) and novel miRNAs are identified using miRDeep.

### 2.3 Analysis of discarded reads by miRNA identification algorithms can identify new miRNAs

Since miRDeep and miRanalyzer are highly stringent algorithms, some miRNAs may escape detection. The false negative discovery rate can, however be calculated by simply performing a megaBlast search of the sequencing data against the miRNAs deposited in miRBase. Perfect alignments are considered true positives. The list of known miRNAs identified by this method is compared to the list of known miRNAs identified by miRDeep or miRanalyzer. False negatives are those miRNAs present in the blast analysis, but which were missed by the miRNA prediction algorithms. This is, in our opinion, an essential control, as it gives information about the percentage of miRNAs that may have escaped miRDeep or miRanalyzer analysis. We have detected ~19% of false negatives, which prompted us to develop a parallel pipeline to analyze reads that may have been incorrectly discarded by the original algorithm (Figure 2). This analysis can and should be performed independently of the algorithm used to retrieve miRNAs from deep sequencing data.

To overcome the lack of sensitivity of miRDeep, our parallel bioinformatics pipeline includes a megaBlast alignment between the dataset of discarded reads by miRDeep and mature sequences deposited in miRBase. Besides, novel transcripts encoding miRNAs predicted by computational tools can be retrieved from the latest Ensembl version using BioMart and also from literature predictions. These sequences are then used to perform a megaBlast search against the sequencing data. The transcripts with perfect matches and alignment length > 18 nucleotides are kept for further processing. These transcripts are then compared with the mature miRNAs deposited in miRBase and those that produce imperfect alignments or do not produce alignments are considered novel miRNAs. Imperfect alignments may identify conserved miRNAs if there is a perfect alignment in the seed region.

Complementary alignments of our dataset reads against the zebrafish genome with SHRiMP alignments and complementary miRDeep analysis with an analysis of the reads discarded by this algorithm, allowed us to identify 90% of the 192 zebrafish miRNAs previously identified, plus 107 miRNA star sequences and 25 novel miRNAs.

### 2.4 Generation of miRNA profiles from deep sequencing data

Deep sequencing of miRNAs can also be used to generate miRNA expression profiles as the absolute number of sequencing reads of each miRNA is directly proportional to their relative abundance. miRNA profiles can be generated based on the number of reads of each particular miRNA. However, a normalization step is essential to compare miRNA expression levels between different samples. The variation in the total number of reads between samples leads to erroneous interpretation of miRNA expression patterns by direct comparison of read numbers (Chen *et al.* 2005). Normalization assumes that the small RNA population is constant and is represented by an arbitrary value (e.g. 1000), and can be calculated as indicated below:

$$\text{miRNA relative expression} = \frac{1000 \text{ x } (\text{NRmiRNA}_X{}^Y)}{\text{TNRmiRNAs}^Y}$$

where $\text{NRmiRNA}_X{}^Y$ is the number of reads of $\text{miRNA}_X$ (X = any miRNA) in sample Y, and $\text{TNRmiRNAs}^Y$ is the total number of miRNAs in sample Y. 1000 is an arbitrary number of reads that allows for data normalization across different samples. This calculates the relative expression of a specific miRNA in a given sample, relative to all miRNAs expressed.

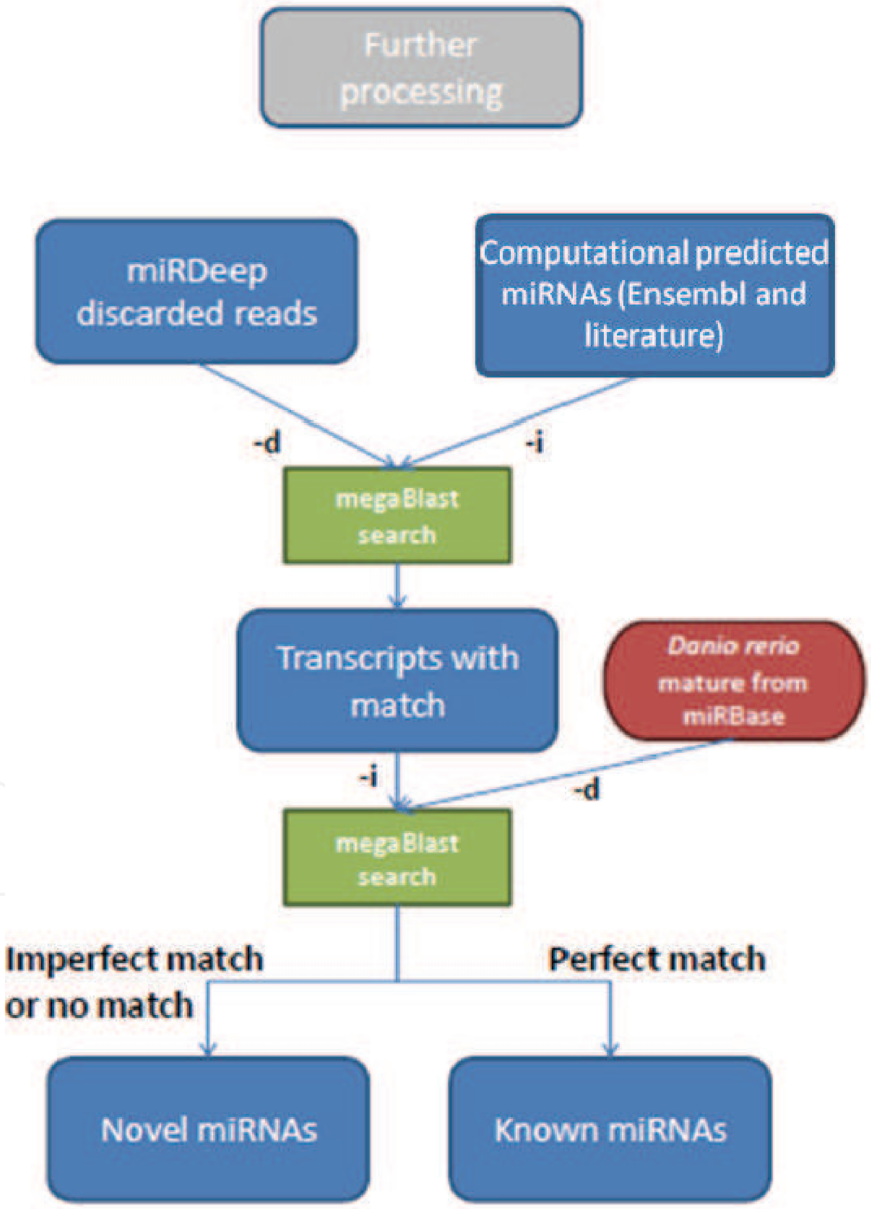


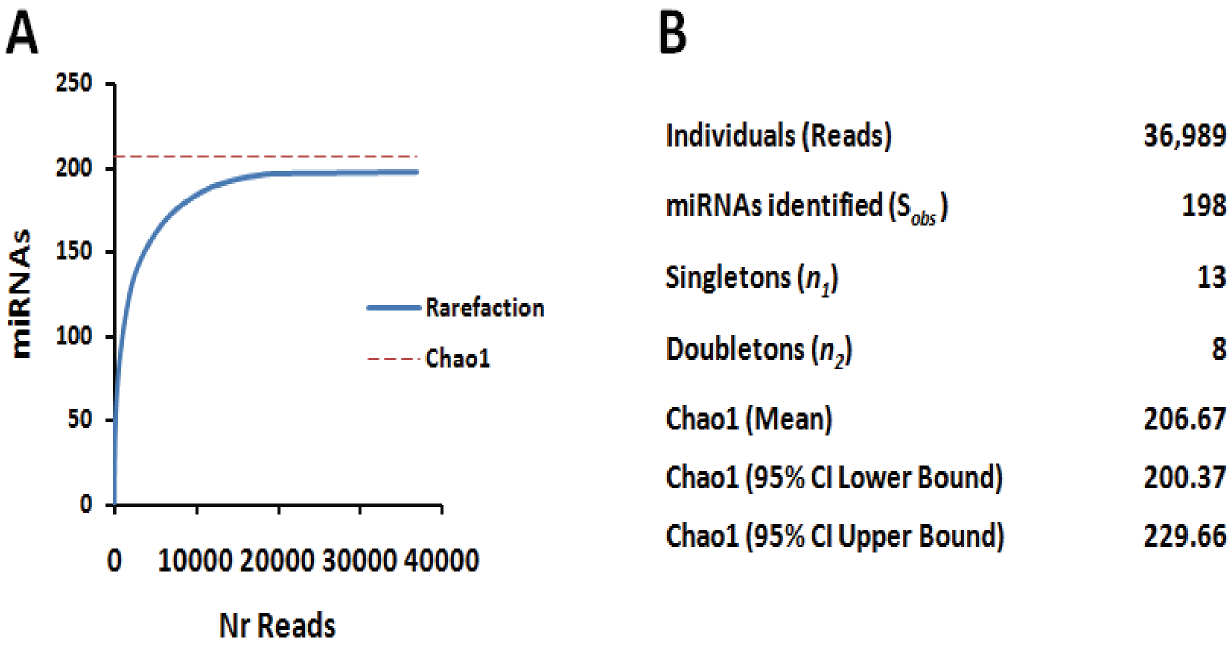
Fig. 2. Bioinformatics pipeline of reads discarded by miRDeep (-i and –d stand for query file and database respectively).

Using this formula it is possible to generate miRNA profiles for each sample sequenced. These profiles provide valuable information about relative miRNA expression, which is essential to understand miRNA function in different tissues. In order to compare miRNA profiles of two deep sequencing samples (e.g. condition vs control), a two-side t-test can be applied to determine miRNA levels. Sequence count values should be log-transformed to stabilize variance (Creighton *et al.* 2009). miRTools already include a computational approach to identify significantly differentially expressed miRNAs (Zhu *et al.* 2010). It compares differentially expressed miRNAs in multiple samples after normalization of the read count of each miRNA with the total number of miRNA read counts which are matched to the reference genome. The algorithm calculates statistical significance (P-value) based on a Bayesian method (Audic and Claverie 1997), which accounts for sampling variability of tags with low counts. Significantly differentially expressed miRNAs are those that show P-values <0.01 and at least 2-fold change in normalized sequence counts.

### 2.5 Statistical analysis of miRNA population

The platforms available for miRNA sequencing offer different sequencing coverage, ranging from thousands to millions of reads. In principle, higher sequencing coverage will enable discovery of more miRNA molecules in a sequencing run. However, technical problems during sample preparation can interfere with good quality sequencing of small RNAs. One of the most common problems is the generation of primer dimmers during PCR amplification of cDNA libraries. This may indicate an excess of primers during amplification, when compared to the miRNA levels in a given cDNA library or low annealing temperature. This problem is often only detected after sequencing. When this happens, a large number of reads do not pass quality control filters and the number of reads corresponding to small RNAs is considerably lower than the initial sequencing coverage. Besides this, quality control filters do not consider reads with sequencing errors in the adaptors or without recognizable adaptors. For these reasons, a tool that verifies if the sequencing coverage is sufficient to retrieve most miRNAs in a given sample is important.

A useful approach to assess the representativeness of miRNA reads in a sequencing experiment is to apply population statistics to the overall miRNA population. We have developed a statistical tool to calculate how many miRNAs are expected in a given sequencing experiment and how many reads are needed to identify them. Rarefaction curves of the total number of reads obtained versus the total number of miRNA species identified are plotted and the total richness of the miRNA population is determined. Chao1, a non-parametric richness estimator (Chao 1987), can be used to determine the total richness of the miRNA population, as a function of the observed richness ($S_{obs}$), and the number of total sequences obtained by sequencing. The value obtained represents the number of different miRNAs that can be identified in a specific sequencing experiment. The rarefaction curve estimates the number of reads needed to identify the different miRNAs that may be present in a sequencing run. For example, 206 miRNAs are expected to be present in a sequencing experiment that retrieves approximately 40000 reads (Figure 3). The steep curve levels off towards an asymptote, indicating the point (~20000 reads) where additional sampling will not yield extra miRNAs. As that critical point is below the total number of reads obtained, we can conclude that the sequencing coverage is sufficient to identify all miRNAs predicted in the particular sample. Rarefaction curves and the Chao1 statistical estimator are computed using EstimateS8.0 (Colwell and Coddington 1994).



| | |
|---|---|
| Individuals (Reads) | 36,989 |
| miRNAs identified ($S_{obs}$) | 198 |
| Singletons ($n_1$) | 13 |
| Doubletons ($n_2$) | 8 |
| Chao1 (Mean) | 206.67 |
| Chao1 (95% CI Lower Bound) | 200.37 |
| Chao1 (95% CI Upper Bound) | 229.66 |

Fig. 3. Statistical analysis of miRNA population. **A)** A rarefaction curve of the total number of reads generated by deep sequencing versus the total number of miRNA species identified is shown. The steep curve levels off towards an asymptote, which indicates the point where additional sampling will not yield new miRNAs. **B)** Homogeneity of the miRNA population was assessed using population statistics and by determining the Chao1 diversity estimator. The Chao1 reached a mean stable value of 207, with lower and upper limits of 200.37 to 229.66, respectively, for a level of confidence of 95%.

## 3. Conclusion

Small non-coding RNAs are a class of molecules that regulate several biological processes. Identification of such molecules is crucial to understand the molecular mechanisms that they regulate. There are already several deep sequencing approaches to identify these molecules. However, correct interpretation of sequencing data depends largely on the bioinformatics and statistical tools available. There are online algorithms that facilitate identification of miRNAs and other small non-coding RNAs from large datasets. However, there are no tools to predict novel small non-coding RNAs beyond miRNAs. As those additional RNA classes, namely piRNAs, snRNAs and snoRNAs are processed differently, the development of algorithms based solely on their biogenesis is challenging. Moreover, the available algorithms have some limitations and additional data analysis should be performed with the discarded reads that can potentially hold non-conventional miRNA molecules. Analysis of deep sequencing data is a powerful methodology to identify novel miRNAs in any organism and determine their expression profiles. The challenge is to deal with increasing dataset size and to integrate the information generated by small RNA sequencing experiments. This will be essential to understand how different RNA classes are related. Computational tools to integrate small non-coding RNA data with gene expression data and target predictions are pivotal to understand the biological processes regulated by miRNAs and other small non-coding RNA classes.

## 4. References

Ambros, V. (2004). The functions of animal microRNAs. *Nature* 431(7006), 350-355.

Audic, S., and Claverie, J. M. (1997). The significance of digital gene expression profiles. *Genome Res.* 7(10), 986-995.

Bartel, D. P. (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* 116(2), 281-297.

Breiman, L. Random forests. Machine Learning (2001). 45:28.

Burnside, J., Ouyang, M., Anderson, A., Bernberg, E., Lu, C., Meyers, B. C., Green, P. J., Markis, M., Isaacs, G., Huang, E., and Morgan, R. W. (2008). Deep sequencing of chicken microRNAs. *Bmc Genomics* 9.

Ceppi, M., Pereira, P. M., Dunand-Sauthier, I., Barras, E., Reith, W., Santos, M. A., and Pierre, P. (2009). MicroRNA-155 modulates the interleukin-1 signaling pathway in activated human monocyte-derived dendritic cells. *Proc. Natl. Acad. Sci. U. S. A* 106(8), 2735-2740.

Chang, J. H., Cruo, J. T., Jiang, D., Guo, H. T., Taylor, J. M., and Block, T. M. (2008). Liver-specific MicroRNA miR-122 enhances the replication of hepatitis C virus in nonhepatic cells. *Journal of Virology* 82(16), 8215-8223.

Chao, A. (1987). Estimating the Population-Size for Capture Recapture Data with Unequal Catchability. *Biometrics* 43(4), 783-791.

Chen, P. Y., Manninga, H., Slanchev, K., Chien, M. C., Russo, J. J., Ju, J. Y., Sheridan, R., John, B., Marks, D. S., Gaidatzis, D., Sander, C., Zavolan, M., and Tuschl, T. (2005). The developmental miRNA profiles of zebrafish as determined by small RNA cloning. *Genes & Development* 19(11), 1288-1293.

Colwell, R. K., and Coddington, J. A. (1994). Estimating Terrestrial Biodiversity Through Extrapolation. *Philosophical Transactions of the Royal Society of London Series B-Biological Sciences* 345(1311), 101-118.

Creighton, C. J., Reid, J. G., and Gunaratne, P. H. (2009). Expression profiling of microRNAs by deep sequencing. *Brief. Bioinform.* 10(5), 490-497.

Droege M, and Hill B. The Genome Sequencer FLX trade mark System-Longer reads, more applications, straight forward bioinformatics and more complete data sets. J.Biotechnol. 136 (1-2): 3-10. 2008.

Feinbaum, R., and Ambros, V. (1999). The timing of lin-4 RNA accumulation controls the timing of postembryonic developmental events in Caenorhabditis elegans. *Dev. Biol.* 210(1), 87-95.

Filipowicz, W., Bhattacharyya, S. N., and Sonenberg, N. (2008). Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? *Nat. Rev. Genet.* 9(2), 102-114.

Friedlander, M. R., Chen, W., Adamidi, C., Maaskola, J., Einspanier, R., Knespel, S., and Rajewsky, N. (2008). Discovering microRNAs from deep sequencing data using miRDeep. *Nature Biotechnology* 26(4), 407-415.

Giraldez, A. J., Cinalli, R. M., Glasner, M. E., Enright, A. J., Thomson, J. M., Baskerville, S., Hammond, S. M., Bartel, D. P., and Schier, A. F. (2005). MicroRNAs regulate brain morphogenesis in zebrafish. *Science* 308(5723), 833-838.

Griffiths-Jones, S., Saini, H. K., van, D. S., and Enright, A. J. (2008). miRBase: tools for microRNA genomics. *Nucleic Acids Res.* 36(Database issue), D154-D158.

Hackenberg, M., Sturm, M., Langenberger, D., Falcon-Perez, J. M., and Aransay, A. M. (2009). miRanalyzer: a microRNA detection and analysis tool for next-generation sequencing experiments. *Nucleic Acids Res.* 37(Web Server issue), W68-W76.

Hofacker, I. L. (2003). Vienna RNA secondary structure server. *Nucleic Acids Res.* 31(13), 3429-3431.

Huang, P. J., Liu, Y. C., Lee, C. C., Lin, W. C., Gan, R. R., Lyu, P. C., and Tang, P. (2010). DSAP: deep-sequencing small RNA analysis pipeline. *Nucleic Acids Res.* 38(Web Server issue), W385-W391.

Kim, V. N. (2005). MicroRNA biogenesis: coordinated cropping and dicing. *Nat. Rev. Mol. Cell Biol.* 6(5), 376-385.

Kim, V. N., Han, J., and Siomi, M. C. (2009). Biogenesis of small RNAs in animals. *Nat. Rev. Mol. Cell Biol.* 10(2), 126-139.

Lau, N. C., Lim, L. P., Weinstein, E. G., and Bartel, D. P. (2001). An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans. *Science* 294(5543), 858-862.

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 Research* 18(4), 610-621.

Pillai, R. S., Bhattacharyya, S. N., and Filipowicz, W. (2007). Repression of protein synthesis by miRNAs: how many mechanisms? *Trends Cell Biol.* 17(3), 118-126.

Schulte, J. H., Marschall, T., Martin, M., Rosenstiel, P., Mestdagh, P., Schlierf, S., Thor, T., Vandesompele, J., Eggert, A., Schreiber, S., Rahmann, S., and Schramm, A. (2010). Deep sequencing reveals differential expression of microRNAs in favorable versus unfavorable neuroblastoma. *Nucleic Acids Res.* 38(17), 5919-5928.

Soares, A. R., Pereira, P. M., Santos, B., Egas, C., Gomes, A. C., Arrais, J., Oliveira, J. L., Moura, G. R., and Santos, M. A. S. (2009). Parallel DNA pyrosequencing unveils new zebrafish microRNAs. *Bmc Genomics* 10.

Tay, Y. M. S., Tam, W. L., Ang, Y. S., Gaughwin, P. M., Yang, H., Wang, W. J., Liu, R. B., George, J., Ng, H. H., Perera, R. J., Lufkin, T., Rigoutsos, I., Thomson, A. M., and Lim, B. (2008). MicroRNA-134 modulates the differentiation of mouse embryonic stem cells, where it causes post-transcriptional attenuation of Nanog and LRH1. *Stem Cells* 26(1), 17-29.

Wienholds, E., Koudijs, M. J., van Eeden, F. J., Cuppen, E., and Plasterk, R. H. (2003). The microRNA-producing enzyme Dicer1 is essential for zebrafish development. *Nat. Genet.* 35(3), 217-218.

Zamore, P. D., and Haley, B. (2005). Ribo-gnome: the big world of small RNAs. *Science* 309(5740), 1519-1524.

Zhu, E., Zhao, F., Xu, G., Hou, H., Zhou, L., Li, X., Sun, Z., and Wu, J. (2010). mirTools: microRNA profiling and discovery based on high-throughput sequencing. *Nucleic Acids Res.* 38(Web Server issue), W392-W397.