Ingredients for in silico miRNA identification and annotation

Christine Gaspin, Olivier Rué, Matthias Zytnicki

SciMedCentral

JSM Biotechnology & Biomedical Engineering

Mini Review

Ingredients for *In silico* miRNA Identification and Annotation

Christine Gaspin[1,2]*, Olivier Rué[1,2], and Matthias Zytnicki[1]

[1]*MIAT, Université de Toulouse, France*
[2]*Plate-forme Genotoul Bioinfo, Université de Toulouse, France*

***Corresponding author**
Christine Gaspin INRA, MIAT & PF GenoToul Bioinfo, Université de Toulouse, Chemin de Borde Rouge, BP52 627, 31326 Castanet Tolosan cedex, France, Tel : 33561285282; Fax: 33561285335; Email: Christine.Gaspin@inra.fr

Submitted: 04 November 2016

Accepted: 08 December 2016

Published: 09 December 2016

ISSN: 2333-7117

Keywords
- miRNA
- Nomenclature
- Isoform
- Identification
- Alignment

Abstract

Annotation of microRNAs (miRNAs) is a prerequisite to the study of their functional analysis at the genome scale. The first list of criteria published for miRNA annotation was established in 2003 by the group of Thomas Tuschl and since then, a large number of bioinformatics resources relying on these criteria have been produced. Recently, the criteria and nomenclature considered for miRNA annotation were questioned by several groups with regard to available resources (databases, miRNA prediction software) but also considering all accumulated knowledge during the ten past years. In this paper, we revisit criteria for miRNA annotation and their importance to design relevant bioinformatics resources able to identify accurately members of known miRNA families as well as members of putative new families.

ABBREVIATIONS

miRNA: microRNAs; sRNA: small RNA

INTRODUCTION

Identification and annotation of miRNAs are the first steps of the study of miRNA functional analysis at the genome scale. Since the discovery of the lin-4 and let-7 miRNA genes in *Caenorhabditis elegans* [1,2], the number of discovered miRNAs exponentially increased in databases. The first uniform system dedicated to miRNA identification and annotation was proposed in 2003 [3], and the authors used a combination of expression and biogenesis criteria to distinguish between *bona fide* miRNAs and other classes of small RNAs. The first repository dedicated to miRNAs was miRBase [4], and it is still the most widely used by far. As the primary repository for miRNA sequence annotation, the initial goals of miRBase were i) to assign unique names to distinct miRNAs prior to publication of their discovery in order to maintain consistent gene nomenclature and ii) to provide a comprehensive and searchable database of all published mature miRNAs and related pre-miRNA hairpin sequences. As a result of bioinformatics screening and the increasing number of small RNA sequencing efforts, the database has grown exponentially from 506 entries covering 5 species (*C. elegans*, *Caenorhabditis briggsae*, *D. melanogaster*, human, mouse and *Arabidopsis thaliana*) in the first publication [4] to 28.645 entries in the most recent release 21 [5]. This accumulation of miRNA sequences in miRBase but also the continued evolution of this repository to meet the needs of the scientific community helped to agree on the characteristics that must be met for a sequence to be considered as a *bona fide* miRNA. These include miRNA biogenesis and expression characteristics, but also a strong conservation of the mature miRNA sequence in related species. The availability of such characteristics and the importance of the role of miRNA in essential processes have boosted the development of many resources, including specific databases but also *in silico* methods, and tools dedicated to the identification of *bona fide* miRNAs, whose accuracy has increased by integrating recently discovered characteristics. Indeed, more than fifty organism-specific or multi-organism miRNA databases and as many miRNA prediction software are now available to the scientific community [6].

Recently, the criteria considered for miRNA identification and annotation were questioned by several groups [7-10] with regard to available resources (databases, miRNA prediction software) but also considering all accumulated knowledge during the ten past years. They argue for a review of nomenclature guidelines and are developing alternative resources that meet their needs. As a result, scientists who are in charge of developing tools for miRNA prediction, annotation and functional analysis have to (re)-consider regularly the continuously evolving characteristics and resources that can be used to identify and annotate properly miRNA genes [11]. In this paper, we revisit criteria for miRNA annotation and we discuss their importance to design relevant bioinformatics resources able to identify and annotate accurately members of known miRNA families as well as members of putative new families.

miRNA characteristics

The essence of a miRNA identification and annotation tool is to learn the characteristics of the miRNAs, and to use this knowledge to provide high quality results. In this section, we will show what are the distinctive features that are specific to miRNAs, and how these features are used by bioinformatics tools.

There are many types of small RNAs present in every

eukaryotic cell, and several of them have common features with the miRNA class. The transcripts can be produced anywhere in the genome (intergenic, introns, UTRs, exons, transposable elements…) so identifying miRNAs is a very challenging task. In order to provide good predictions, current tools use a combination of different aspects related to miRNAs: knowledge from the miRNA biogenesis, knowledge from other species or validated miRNAs, and expression data (usually, sRNA-seq datasets).

Biogenesis related characteristics

Primary miRNAs (pri-miRNA) are mostly transcribed by RNA polymerase II. The pri-miRNA is folded into one or more long hairpins, with possibly several conserved patterns in the loop and the sequence that are at the extremities of the hairpin [7]. However, the pri-miRNA seems rapidly decayed, and RNA-Seq studies usually cannot detect evidences of pri-miRNAs [12]. This pri-miRNA is then cleaved by Drosha (in animals) or DCL (in plants), into one or more individual pre-miRNAs. Each pre-miRNA adopts a stem-loop secondary structure with a constrained size that differs between plants and animal, but is always present. The pre-miRNA is cleaved by a Dicer protein, leaving at determined positions three products including the two mature miRNAs and the loop. The two mature miRNAs form a duplex with 2 bp overhangs at their 3' ends, while the loop is precisely positioned between the mature miRNAs. During this process, sequence heterogeneity in size and content of the mature sequence may arise from imprecise cleavage and editing mechanisms. Resulting isoforms show length and sequence heterogeneity at their 5' and 3' extremities, but also edited positions of the mature miRNA even if edition internal to the mature are rare events that concern a few miRNA species [13]. Although isoforms usually differ by 1-3 bp, more substantial differences may be observed as well. Moreover, many miRNAs have been duplicated during genome evolution, and some of them have evolved. As a consequence, several (at least mature) miRNAs are found almost identical in different parts of a genome. In the first age of miRnome studies, it was thought that only one arm of the pre-miRNA was functional. It is now well accepted that both strands of the hairpin may be functional according to the conditions tested, tissue analyzed, etc.

All these characteristics, imposed by the biogenesis of the miRNAs, are crucial to validate a candidate. However, there is no strict rule, and exceptions are observed for most steps. For instance, other pathways also generate miRNAs: mirtrons, matured from the introns of genes, skip the first cleavage step. Moreover, the miRNA biogenesis may be very similar to other small RNA biogeneses. For instance, both miRNAs and silencing RNAs are processed by Dicer proteins, and loaded by Argonaute proteins to regulate the target region. As a consequence, these rules, derived from biogenesis, are sometimes insufficient to efficiently discriminate miRNAs.

Homology characteristics

The easiest method to validate a putative mature miRNA is to find a confirmed mature miRNA with high identity. The rationale is that similarity implies homology, and thus similar function. However, finding sequences similar to candidates of size ~20bp with a few mismatches does give false positives. To reduce the number of false positives, some empirical rules have been defined: for instance, the "seed region" of the miRNA (corresponding to nucleotides 2-8, but this region seems to vary from species to species) should match perfectly [14]. The seed is thought to be the region where the miRNA and the target start hybridizing, and contains, in principle, the core function of the miRNA.

This homology criteria also contributed to the definition of miRNA families, akin to gene families. A first definition of a miRNA family is that all members of a given family should share a common ancestor. Very little work has been done to reconstruct the synteny of miRNAs in the tree of life [15,16], so we have to resort to guessing that nearly identical sequences have a common ancestor. In practice, similarity of the seed region is often used, and a family may be defined by the set of miRNAs with a common seed region. However, given the size of the seed region (~7 bp), the similarity may be fortuitous. Alternatively, a miRNA family can also be defined as the set of the miRNAs with a common function. In practice, both definitions can be handled identically, if the hypothesis is that the function is encoded in the mature miRNA, or in its seed region. The only interesting difference is that this definition may also group together two miRNAs that are nearly identical due to random or convergent evolution. Studies on miRNA evolution also consider miRNA families according to global pre-miRNA sequence similarity [17,18] which is in accordance with miRBase family organization (Figure 1a,1b).

So far, there is no widely accepted definition of miRNA families, although these families are widely used. This leads to obvious problems in the attribution of families, including in miRBase. The authors of this database have developed a constant effort for assigning a consistent name to members of the same family [4]. However, with time, and discovery of previously unknown membership relations, the names of some miRNAs have changed, leading to substantial changes [19] in the database and ambiguities in family definition and naming.

Expression related characteristics

Small RNA-sequencing (sRNA-seq) is probably the most widely used method to find miRNAs. Accumulated knowledge suggests that a *bona fide* miRNA locus should contain two stacks of ~21-23 bp reads, each one corresponding to a distinct mature miRNA which is generated from the 3' and 5' arms of the pre-miRNA stem-loop structure, and are distant by a few dozen base pairs, sometimes a lot of more in plants. These distributions were confirmed by miRBase (v21, using the mature miRNAs and the precursors miRNAs) (Figure 2a,2b). When aligned and folded, the majority form of 5' and 3' arms should adopt a 2bp overhang. Moreover, it is expected that one of the stack is higher than the other, also called the miRNA*, and the 5' extremity of each stack should show an homogeneity of sequence start. However, we do not observe this pattern for a substantial fraction of miRNAs. Several alternative small RNAs may accumulate next to the canonical miRNA, with various sizes, (although close to the mature miRNA size). For these miRNAs, the signal is often harder to interpret. Sometimes, one of the two mature miRNAs is not expressed, and other times, both strands are found almost equally expressed. Moreover, some "young" miRNAs originate from silencing RNA (siRNA) loci [20], and thus the expression profile may resemble the siRNA profiles as well. Last, other small

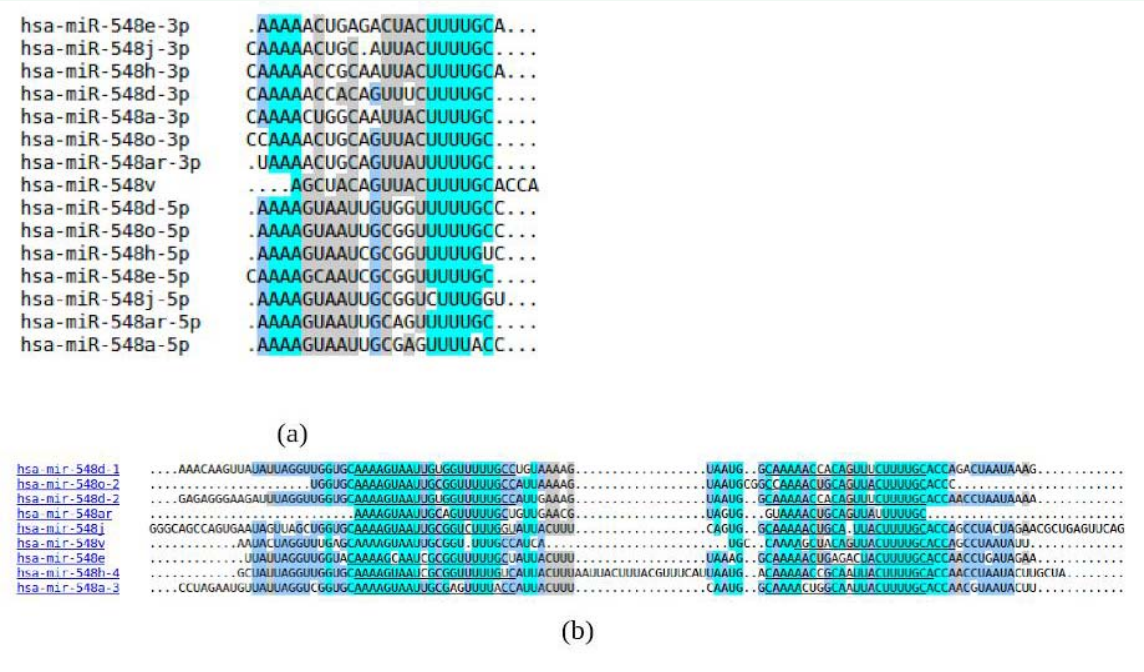

Figure 1 Alignments provided by miRBase for mir-548 family. Only sequences labeled as confident are given for miRNA (a) and pre-miRNA (b).

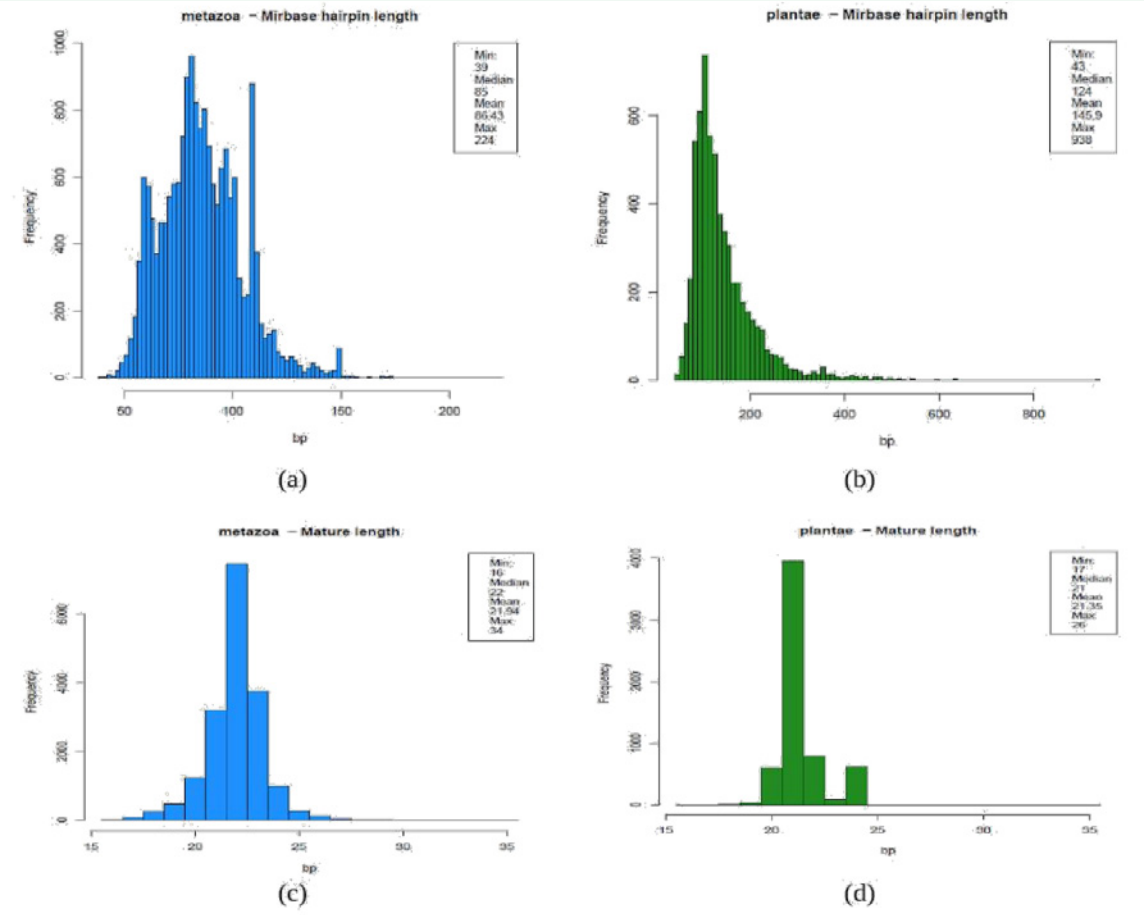

Figure 2 (a). Distribution of the pre-miRNA sizes in Metazoa, as given by miRBase. (b). Distribution of the pre-miRNA sizes in Plantae. (c). Distribution of the mature miRNAs in Metazoa. (d). Distribution of the mature miRNAs in Plantae.

RNAs, such tRFs (tRNAs derived fragments) exhibit sizes ranging from 15 to 32 bp [21] and may also exhibit a similar expression profile, leading to possible error during the annotation process. For all these reasons, even though sRNA-seq is the primary tool for miRNA discovery, the analysis should be performed with care. Even with this caveat, given the current knowledge on sRNAs, no annotation can be considered as 100% valid.

miRNA identification and annotation

Small RNA sequencing is now recognized as the most powerful technical approach for miRNA discovery. A generic pipeline for the identification of miRNA candidates involves a pre-processing step that provides a reduced dataset of good quality. This quality check is performed after the trimming of adapters and the removing of low complexity/quality reads. It evaluates the read length distribution which is expected to be centered at 22 (resp. 21) bases for a good animal (resp. plant) miRNA-seq library preparation. Sizes generally considered are in the range 18-25 nt in accordance with values obtained from miRBase (Figure 2c,2d). Finally, the dataset is reduced to unique sequences by eliminating redundancy while keeping reads counting by sequence.

In a next step, the reads are mapped against a reference genome or one or several databases, and miRNA candidate loci are selected and evaluated [22]. Here, each pipeline has its own strategy on several aspects: the miRNA variants (whether to include them or not), the mapping strategy (the alignment tool and its parameters), and the prediction of a pre-miRNA stem-loop at each mapped locus. We will discuss the two latter key aspects in the following sections.

Read mapping

Mapping reads onto the genome always trigger the question: "Should I accept reads with multiple (equally probable) mappings"? Accepting multiple mappings leads to difficulties in assigning reads to specific positions, identifying miRNA loci, and quantifying the expression of the miRNAs. On the contrary, uniquely mapping reads cannot provide information on duplicated miRNAs. Using miRBase, we found that 2082 mature miRNAs out of 2588 where unique in human, and 265 out of 427 in *A. thaliana*. Discarding multiple matches thus also precludes the analysis of 20% (in human) or 38% (in *A. thaliana*) of the miRNAs. Moreover, the repeatedness of these families may have been positively selected to perform crucial tasks in the regulation system, and skipping these genes may have damaging consequences. Accepting errors while mapping (mismatches and/or indels) obviously increases the sensitivity of the analysis. There are various reasons why accepting errors may be Preferred:

i) Sequencing error may prevent the detection of miRNAs. However, with current technology, the error rate is about 0.1% (about 1 mismatch every 40 reads), and sequencing errors are not expected to alter significantly the miRNA detection step;

ii) The genome/transcriptome of the organism under study differs from the sequenced genome and variants may exist at any position despite required conservation;

iii) Edition may alter mature miRNAs. Here, the user should specify that the alignment errors should be located at the (usually 3') extremities of the reads. Whereas this verification would be useful in practice, it is usually not implemented in alignment tools, and users have to resort to *ad hoc* analysis.

Accepting errors comes with a price: loss of specificity. We used Bowtie [11] with default parameters to quantify the repeatedness of the known miRNAs in *A. thaliana* and Human genomes. We mapped the *A. thaliana* mature miRNA against the genomic sequence, and we found that each miRNA has, in average, 2.3 possible locations with no mismatch. There are 5.6 possible locations per miRNA with 1 mismatch, 15.4 with 2 mismatches, and 45.7 with 3 mismatches. Likewise, in Human, the number of locations per known mature miRNA is 55, 348, 1221 and 3567 with 0, 1, 2 and 3 mismatches respectively. Thus, the number of spurious hits is expected to increase exponentially when the number of accepted mismatches increases.

Stem-loop prediction

Before sRNA-seq data were available, genome sequences were searched for loci that were conserved among several species and could fold into stem-loop structure that were scored using simple rules [23,24] or more selective machine learning techniques using known microRNAs as a training set [25]. These methods yielded many false positive loci that were not effectively transcribed. For instance, in the human genome, around 11 million loci were found to fold into a stem-loop structure [26].

Methods that aim at discovering new miRNAs from sRNA-seq data also have to consider all miRNA characteristics because many transcripts from sRNA-seq encode other types of non coding RNAs. However, compared to the computation of all stem-loop structures from a genomic sequence, the number of candidate loci to analyze is highly reduced by considering only mapped loci. Thus, at each locus mapped, the presence of a stem-loop structure in accordance with the pre-miRNA secondary structure will be a first requirement. Many tools exist in the literature for *ab initio* prediction of the required stem-loop structure of the pre-miRNA. Almost all use a secondary structure predictor like RNAfold [27] that is used on sliding windows, with size similar to pre-miRNA expected length. By considering that the most represented read of a locus encodes a mature miRNA candidate, some authors have proposed to anchor there the pre-miRNA and to search for a near-perfect alignment that corresponds to the duplex resulting from Dicer cleavage [28]. The latter approach provides a folding that fits better the expected duplex for long pre-miRNA sequences (such as those observed in plants), in much less time.

CONCLUSION

The sRNA-seq technology approach is now a popular method used to discover and annotate miRNAs at the genome scale. A large amount of bioinformatics resources were developed to analyze sRNA-seq data that cover key issues from annotation of new and known miRNA to the identification of their function. In this paper, we examined the accumulated knowledge and the impact of derived characteristics in the essential steps of mapping reads and evaluating the pre-miRNA secondary structure. Existing bioinformatics resources have contributed a lot to increase knowledge and to improve the content of reference repositories, and the continuous development of high quality

databases such as miRBase is crucial for the analysis and the annotation of miRNAs. However, since many aspects of miRNAs are still not fully understood (biogenesis, target identification, differences with other small RNAs, family reconstruction, etc.), the database is bound to be flawed with errors. The user should know the limitations of miRBase, and be careful when exploiting this resource. This includes the manual verification of the miRNA entries (proper expression profile and pre-miRNA folding), and the miRNA families used for the analysis (alignment of the mature miRNA members). It is all the more important as a mis-annotated miRNA could be used as an evidence to annotate another putative small RNA, and thus errors tend to expand. Besides this reference repository, there is still place for developing new mappers contributing to solve multi-mapping ambiguities, able to deal with any type of errors and facilitating the prioritization of candidates. On the biological side, the precise location of pri-miRNA transcripts and their content remains to explore for helping to solve mapping ambiguities.

ACKNOWLEDGEMENTS

OR was supported by France Génomique National infrastructure, funded as part of "Investissement d'avenir" program managed by Agence Nationale pour la Recherche (contrat ANR-10-INBS-09).

REFERENCES

1. Lee RC, Feinbaum RL, Ambros V. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell. 1993; 75: 843-854.
2. Reinhart BJ, Slack FJ, Basson M, Pasquinelli AE, Bettinger JC, Rougvie AE, et al. The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. Nature. 2000; 403: 901-906.
3. Ambros V, Bartel B, Bartel DP, Burge CB, Carrington JC, Chen X, et al. A uniform system for microRNA annotation. RNA. 2003; 9: 277-279.
4. Griffiths-Jones S. The microRNA Registry. Nucleic Acids Res. 2004; 32: 109-111.
5. Kozomara A, Griffiths-Jones S. miRBase: annotating high confidence microRNAs using deep sequencing data. Nucleic Acids Res. 2014; 42: 68-73.
6. Lukasik A, Wójcikowski M, Zielenkiewicz P. Tools4miRs - one place to gather all the tools for miRNA analysis. Bioinformatics. 2016; 32: 2722-2724.
7. Ha M, Kim VN. Regulation of microRNA biogenesis. Nat Rev Mol Cell Biol. 2014; 15: 509-524.
8. Desvignes T, Batzel P, Berezikov E, Eilbeck K, Eppig JT, McAndrews MS, et al. miRNA Nomenclature: A View Incorporating Genetic Origins, Biosynthetic Pathways, and Sequence Variants. Trends Genet. 2015; 31: 613-626.
9. Budak H, Bulut R, Kantar M, Alptekin B. MicroRNA nomenclature and the need for a revised naming prescription. Brief Funct Genomics. 2016; 15: 65-71.
10. Fromm B, Billipp T, Peck LE, Johansen M, Tarver JE, King BL, et al. A Uniform System for the Annotation of Vertebrate microRNA Genes and the Evolution of the Human microRNAome. Annu Rev Genet. 2015; 49: 213-242.
11. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009; 10: 25.
12. Chang TC, Pertea M, Lee S, Salzberg SL, Mendell JT. Genome-wide annotation of microRNA primary transcript structures reveals novel regulatory mechanisms. Genome Res. 2015; 25: 1401-1409.
13. Ameres SL, Zamore PD. Diversifying microRNA sequence and function. Nat Rev Mol Cell Biol. 2013; 14: 475-488.
14. Bartel DP. MicroRNAs: target recognition and regulatory functions. Cell. 2009; 136: 215-233.
15. Hertel J, Bartschat S, Wintsche A, Otto C; Students of the Bioinformatics Computer Lab, Stadler PF. Evolution of the let-7 microRNA family. RNA Biol. 2012; 9: 231-241.
16. Guerra-Assunção JA, Enright AJ. Large-scale analysis of microRNA evolution. BMC Genomics. 2012; 13: 218.
17. Meunier J, Lemoine F, Soumillon M, Liechti A, Weier M, Guschanski K, et al. Birth and expression evolution of mammalian microRNA genes. Genome Res. 2013; 23: 34-45.
18. Hertel J, Stadler PF. The Expansion of Animal MicroRNA Families Revisited. Life (Basel). 2015; 5: 905-920.
19. Van Peer G, Lefever S, Anckaert J, Beckers A, Rihani A, Van Goethem A, et al. miRBase Tracker: keeping track of microRNA annotation changes. Database. 2014; 2014.
20. Voinnet O. Origin, biogenesis, and activity of plant microRNAs. Cell. 2009; 136: 669-687.
21. Kumar P, Mudunuri SB, Anaya J, Dutta A. tRFdb: a database for transfer RNA fragments. Nucleic Acids Res. 2015; 43: 141-145.
22. Tam S, Tsao MS, McPherson JD. Optimization of miRNA-seq data preprocessing. Brief Bioinform. 2015; 16: 950-963.
23. Lim LP, Lau NC, Weinstein EG, Abdelhakim A, Yekta S, Rhoades MW, et al. The microRNAs of Caenorhabditis elegans. Genes Dev. 2003; 17: 991-1008.
24. Lai EC, Tomancak P, Williams RW, Rubin GM. Computational identification of Drosophila microRNA genes. Genome Biol. 2003; 4: 42.
25. Huang TH, Fan B, Rothschild MF, Hu ZL, Li K, Zhao SH. MiRFinder: an improved approach and software implementation for genome-wide fast microRNA precursor scans. BMC Bioinformatics. 2007; 8: 341.
26. Bentwich I, Avniel A, Karov Y, Aharonov R, Gilad S, Barad O, et al. Identification of hundreds of conserved and nonconserved human microRNAs. Nat Genet. 2005; 37: 766-770.
27. Hofacker IL, Fontana W, Stadler PF, Bonhoeffer S, Tacker M, Schuster P. Fast folding and comparison of RNA secondary structures. Monatsh Chem. 1994; 125: 167-188.
28. Higashi S, Fournier C, Gautier C, Gaspin C, Sagot MF. Mirinho: An efficient and general plant and animal pre-miRNA predictor for genomic and deep sequencing data. BMC Bioinformatics. 2015; 16: 179.

Cite this article

Gaspin C, Rué O, Zytnicki M (2016) Ingredients for In silico miRNA Identification and Annotation. JSM Biotechnol Bioeng 3(5): 1071.