Fall 12-5-2018

GMAim: an analytical pipeline for microRNA splicing profiling using generative model

Kan Liu
University of Nebraska-Lincoln, liukan.big@gmail.com

GMAim: an analytical pipeline for microRNA splicing profiling using generative model

Kan Liu
Advisor: Dr. Juan Cui
Dec-05-2018

Department of Computer Science and Engineering
University of Nebraska-Lincoln

Abstract

MicroRNAs (miRNAs) are a class of short (~22 nt) single strand RNA molecules predominantly found in eukaryotes. Being involved in many major biological processes, miRNAs can regulate gene expression by targeting mRNAs to facilitate their degradation or translational inhibition. The imprecise splicing of miRNA splicing which introduces severe variability in terms of sequences of miRNA products and their corresponding downstream gene expression regulation. For example, to study biogenesis of miRNAs, usually, biologists can deplete a gene in the miRNA biogenesis pathway and study the change of miRNA sequences, which can cause impression of miRNAs. Although high-throughput sequencing technologies such as small RNA-seq provide unprecedented quantitative readouts for miRNA expression analysis, existing standalone tools developed for miRNA-seq analysis usually do not provide comprehensive and detailed information on miRNA splicing variations. To advance our understanding of miRNA variability by identifying significant miRNA imprecise splicing with statistical power, and to present a complete scenario of miRNA splicing regulation, we proposed a pipeline called GMAim as a Genome-wide MiRNA Imprecise splicing detection including read cataloging, miRNA family expression quantification, imprecise splicing identification and categorization. This pipeline was implemented with R and an R package was developed. Based on the cross validation, this tool is proved to be more powerful and accurate than other modern miRNA annotation tools by avoiding over-counting of sequence reads and quantification as well as fast implementation time.

Introduction

MicroRNAs (miRNAs) are basically a type of short (~22 nt) single strand RNA molecules predominantly shown in eukaryotes. With the rapid breakthrough and findings in the past 20 years, many significant revelations have been reported in discovering miRNA biogenesis, target prediction and ab initio annotation[1]. For example, with the increasing development of the small RNA-sequencing (sRNA-seq) technology, thousands of miRNAs can be discovered with the high sequencing confidence, which provides an ebb of new data to analyze as well as a huge challenge for miRNA prediction and annotation[2]. miRNAs are known to incorporate regulations on multi-functional post-transcriptional gene expression[3]. Hence, the precise identification and annotation of plant miRNAs that connect diversified biological processes for growth and development are still quite mysterious for us and also deserves to much more analytic attention.

Without precise splicing, the miRNA products and their corresponding downstream gene expression regulation might be severely affected[4]. Therefore, miRNA imprecise splicing is an overlooked yet important area in miRNA prediction and annotation. Different conditions using treatment such as knock-out argo factors, miRNA splicing precision can be severely affected by such treatments and this imprecise splicing regulation is quite condition-specific throughout species. According to Liu et al.[5], imprecise miRNAs were defined as those that did not fall within ±2 bases of the annotated mature miRNA(s) or miRNA*(s) positions. Assessment of miRNA processing precision depends critically upon sequencing depth. Here, we focused on the most highly expressed miRNA loci for this analysis. Precision was assessed by computing the ratio of imprecise small RNAs to the total miRNA.

The advent of more comprehensive and species-specific miRNA splicing isoform prediction is on the horizon. Many miRNA analysis tools use miRNA sequencing data to identify known and novel miRNAs and detect their differential expression profiles, e.g., miRDeep2, omiRas, miRanalyzer and miRExpress. Among them, miRDeep2 appears to be the most popular program and was widely used for quantifying known miRNAs and predicting novel miRNAs. However, miRDeep2 has the disadvantage of over-counting sequence reads and the inability to deal with mapped reads containing indels, affecting the detection of miRNA splicing variants. Moreover, miRDeep2 does not allow close examination of miRNA variations like miRNA imprecise splicing, which appear to be indispensable to fully understand the biogenesis and biological functions of miRNAs. To overcome such disadvantages, and to develop a new comprehensive pipeline for the miRNA analysis for the current challenges, we developed an R/Bioconductor package, GMAim, for Genomewide analyzing MiRNA Imprecise splicing profiles and differential expressed miRNA imprecise splicing products. By comparing to other tools, such as miRDeep2 and sRNAbench, we demonstrated that, by applying it to published Arabidopsis thaliana sRNA-seq data, GMAim can avoid over-counting of sequence reads and has quantification as well as fast implementation time.

Material and methods

Reads alignment of small RNA-seq

The single end RNA-seq was aligned against reference hairpin sequences using bowtie[6] (v1.0.0) (other aligner with the same functionality should also work) for a minimum alignment length of 15-nt without mismatch allowed. Only unique alignment of each read was kept for further counting. The hairpin reference sequences that we used were downloaded from miRBase[7].

Identification of imprecise spliced miRNA variants

For each miRNA region, we used the counts of sRNA-Seq reads mapped to the hairpin region between samples to estimate ratio of precise/imprecise splicing products. The imprecise splicing levels of all mature miRNAs are collected using: GAlignments, IRanges, Rsamtools[8] and GenomicFeatures Bioconductor packages.

Statistical Model for the imprecise splicing test

We took a generative model approach for testing the miRNA imprecise splicing. In detail, let Y_{gbki} be the count of the reads from replicate i under condition k that support conclusion $b \in \{precision, imprecision\}$ for gene g. We model this count with the following negative binomial (NB) regression model with log link function

$$Y_{gbki} NB\left(\mu_{gbki}, \phi_g\right)$$

and

$$log\left(\mu_{gbki}\right) = \alpha_{gki} + \beta_{gk}\delta_{gk}$$

where α_{gki} could be viewed as the intercept term for gene g from sample i in group k for both of the precise and imprecise splicing counts, β_{gk} is the log-odds of imprecise splicing variants for gene g in all samples in group k, δ_{gk} is an indicator function that take value 1 if $b = imprecision$.

The testing the existence of differential imprecise splicing between two groups $k_1$and k_2 could be formulated as testing $H_0: \beta_{gk_1} = \beta_{gk_2}$. We adjust the testing multiplicity across genes for the comparison of each two groups by Benjamini-Hochberg (BH) procedure.

We implemented our test in R utilizing some of the functionalities of Bioconductor package edgeR[9] using a slightly different but equivalent parameterization. In our R implementation, the intercept of the first sample and the log-odds of imprecise splicing of the first group were set as the baseline, and the other parameters were simply contrasting between the others and the baselines. We remark that we did not use the column-wise library size as an offset when fitting the generalized linear model, because we already included the sample specific intercepts. We also provide an option of using common over-dispersion parameter ($\phi_g = \phi$).

For the case study, we used Arabidopsis thaliana (plant) to test the performance of GMAim. The genome that we used for test is Arabidopsis thaliana (TAIR10). In Arabidopsis thaliana genome, there are 326 precursors and 428 mature miRNAs in miRbase. The test RNA-seq data were obtained from our previous result[10]. In this study, two conditions were collected and sequenced for the miRNA. On condition is the wild type and the other is mutant plant which mutant the Smx gene, which also regulates miRNA splicing in the biogenesis pathway in plant. We name those two conditions as control and Smx, separately. The samples were sequenced using Illumina platform with the single-end read at the length of 150 bp. After adapter removal and other quality control, we got an average mature miRNA with the length about 22 bp.

Results

The required inputs for GMAim are the small RNA-seq alignment files (sorted or unsorted Bam files, ideally) generated from aligning sRNA-seq against hairpin genomic sequences of certain species and universal annotation file (miRNA.dat) downloaded from miRBase[7] which present a sequencing platform and RNA-seq aligner independent methodology for miRNA imprecise splicing analysis.

The pipeline to use GMAim has four major steps (Figure 1). First, a user needs to get pre-procced miRNA-seq datat by conducting quality control, adaptor removal and alignment of the filtered reads back to the reference hairpin and genomic sequences. Typical length distribution of processed and mapped Arabidopsis miRNA-seq reads is shown in Figure 2. This result proves that the success of RNA-seq library construction and high quality for further miRNA profile analysis. The length distribution of the reads is centered on 22 nt, this illustrates microRNA sequencing is reliable. Then, the alignment files in BAM can be used for GMAim directly. Besides bam files, the ready-to-use miRNA stem-loop annotation file from miRbase and the reference hairpin sequence are required to import to GMAim. Then, GMAim can identify imprecise splicing of miRNAs reads and conduct differential analysis using negative binomial model. We will discuss them in detail later. The last step is results visualization and output.

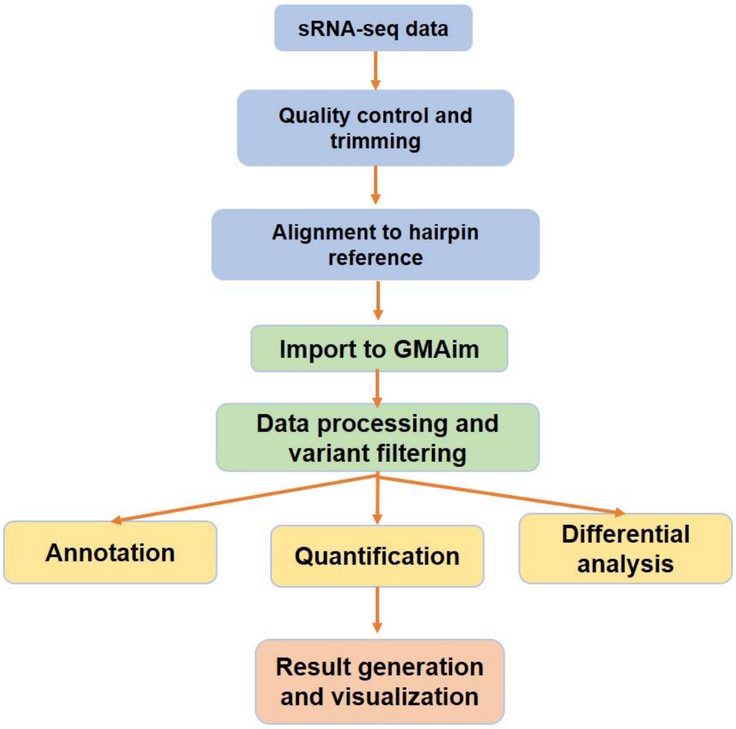

sRNA-seq data
Quality control and trimming
Alignment to hairpin reference
Import to GMAim
Data processing and variant filtering
Annotation
Quantification
Differential analysis
Result generation and visualization

Figure 1. Workflow of GMAim package

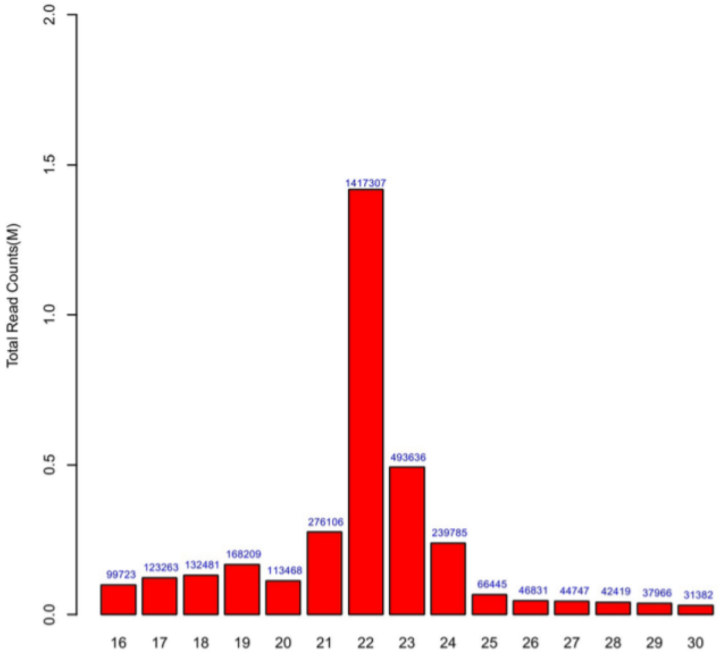
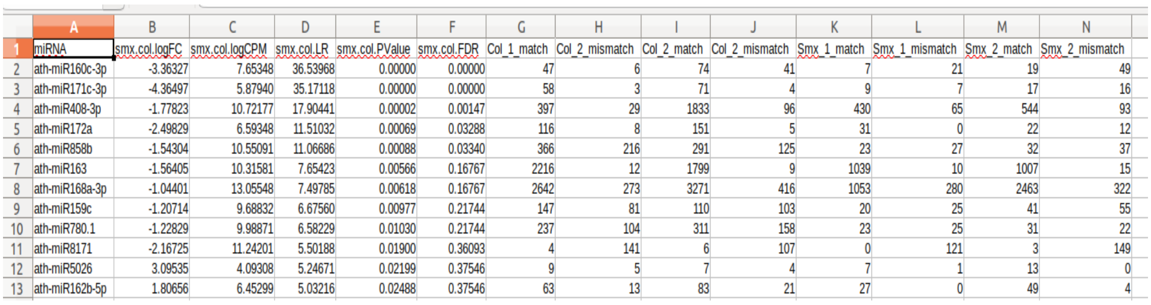


Figure 2. Typical length distribution of processed and mapped Arabidopsis miRNA-seq reads (peak~ 22nt)

Using negative binomial distribution model in GMAim, differential miRNA imprecise splicing results will be stored as local csv files with statistical significance included as well as miRNA imprecise splicing profile plots (Figure 3). It contains mature miRNA IDs, Fold change, p-value and FDR adjusted p-value in ascending order. The detailed precise and imprecise spliced read mapped to each mature miRNA count are listed at the end. Using such information, users can easily check the most significant or deviated miRNA splicing between different conditions. Using such information, GMAim presents a detailed prioritized significant splicing variant list identified between control and Smx samples. Also, from the results, we found an overall of the miRNA expression decrease compared from control to Smx samples, which further confirmed the affect from mutating Smx gene in miRNA splicing regulation.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	miRNA	smx.col.logFC	smx.col.logCPM	smx.col.LR	smx.col.PValue	smx.col.FDR	Col_1_match	Col_2_mismatch	Col_2_match	Col_2_mismatch	Smx_1_match	Smx_1_mismatch	Smx_2_match	Smx_2_mismatch
2	ath-miR160c-3p	-3.36327	7.65348	36.53968	0.00000	0.00000	47	6	74	41	7	21	19	49
3	ath-miR171c-3p	-4.36497	5.87940	35.17118	0.00000	0.00000	58	3	71	4	9	7	17	16
4	ath-miR408-3p	-1.77823	10.72177	17.90441	0.00002	0.00147	397	29	1833	96	430	65	544	93
5	ath-miR172a	-2.49829	6.59348	11.51032	0.00069	0.03288	116	8	151	5	31	0	22	12
6	ath-miR858b	-1.54304	10.55091	11.06686	0.00088	0.03340	366	216	291	125	23	27	32	37
7	ath-miR163	-1.56405	10.31581	7.65423	0.00566	0.16767	2216	12	1799	9	1039	10	1007	15
8	ath-miR168a-3p	-1.04401	13.05548	7.49785	0.00618	0.16767	2642	273	3271	416	1053	280	2463	322
9	ath-miR159c	-1.20714	9.68832	6.67560	0.00977	0.21744	147	81	110	103	20	25	41	55
10	ath-miR780.1	-1.22829	9.98871	6.58229	0.01030	0.21744	237	104	311	158	23	25	31	22
11	ath-miR8171	-2.16725	11.24201	5.50188	0.01900	0.36093	4	141	6	107	0	121	3	149
12	ath-miR5026	3.09535	4.09308	5.24671	0.02199	0.37546	9	5	7	4	7	1	13	0
13	ath-miR162b-5p	1.80656	6.45299	5.03216	0.02488	0.37546	63	13	83	21	27	0	49	4

Figure 3. Significantly differentially spliced miRNAs

GMAim can also provide the details flanking size at ±5 nt of both the 5’ and 3’ end of the mature miRNA splicing boundaries. Users can easily identify the highest frequent imprecise splicing boundary of both ends of mature miRNA. “0” means the perfect match of the 5’ or 3’ end of mature miRNAs. “-” denotes the upstream shift distance compared with the annotation and “+” denotes the downstream shift distance between identified miRNA and its corresponding reference. The miRNA imprecise splicing analysis can be plotted such profile using ggplot2 package to present the whole scenario of miRNA splicing under different samples (Figure 4).

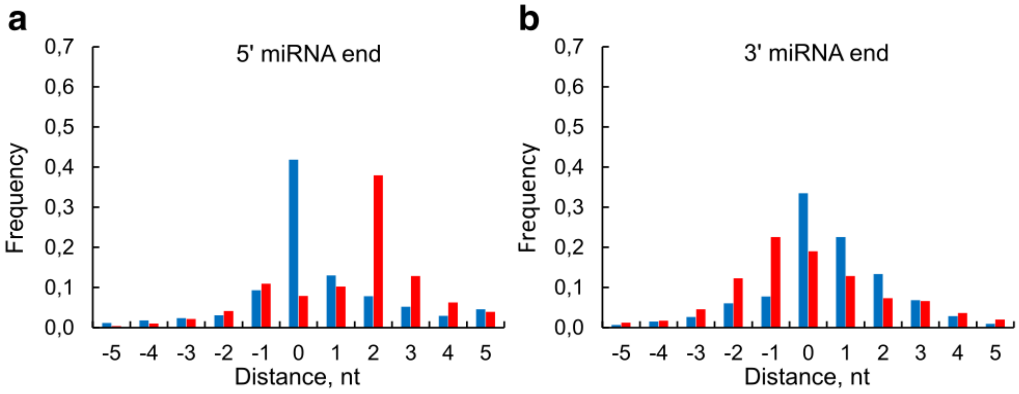
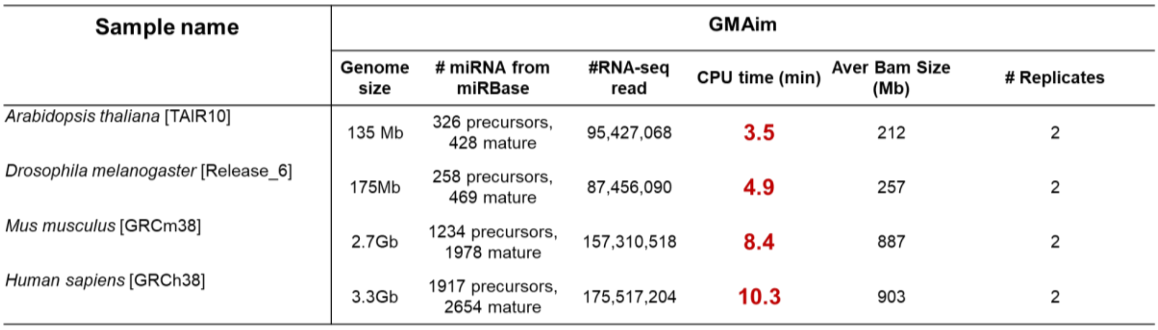


Figure 4. The overall miRNA imprecise splicing profile of Arabidopsis.

We further test our pipeline’s runtime and other performance using some well-annotated model species including human, mouse and fly. We chose them as model species because the miRNA family list in those species are thoroughly studied and curated to present a better benchmark.

By comparing different model species (Table 1), from the small genome such as plant and more complex genome such as fly, the runtime basically spans from 3 to 5 minutes of the whole pipeline. When it goes to much higher eukaryotic species such as human and mouse, due the large genome size and a higher number of miRNA list, we found the analytical runtime can be up to over 10 minutes. Those tests were conducted on Xeon E7 v4 server under Ubuntu 16.04 64-bit operation system.

Table 1. GMAim runtime for imprecise splicing analysis using various test datasets.



Sample name	GMAim					
	Genome size	# miRNA from miRBase	#RNA-seq read	CPU time (min)	Aver Bam Size (Mb)	# Replicates
Arabidopsis thaliana [TAIR10]	135 Mb	326 precursors, 428 mature	95,427,068	3.5	212	2
Drosophila melanogaster [Release_6]	175Mb	258 precursors, 469 mature	87,456,090	4.9	257	2
Mus musculus [GRCm38]	2.7Gb	1234 precursors, 1978 mature	157,310,518	8.4	887	2
Human sapiens [GRCh38]	3.3Gb	1917 precursors, 2654 mature	175,517,204	10.3	903	2

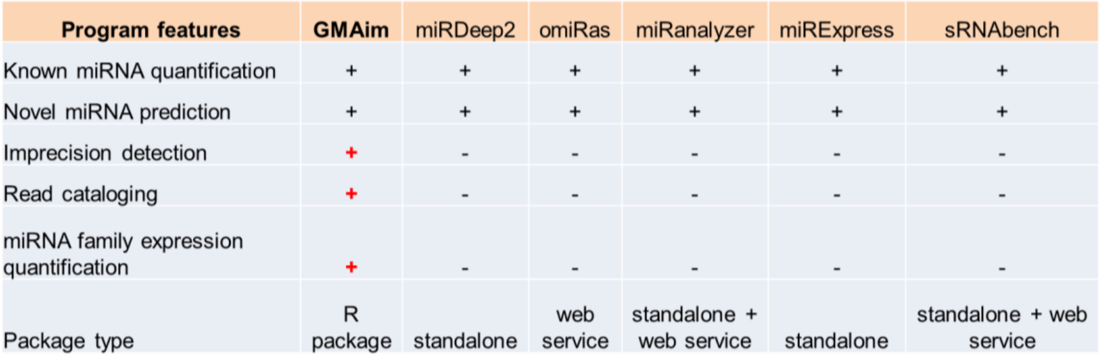
Discussion

Based on the rapid breakthrough and findings in the past 20 years, discovering miRNA isoforms, target prediction and ab initio annotation are becoming hot topics in better understanding the biogenesis and downstream regulatory role of miRNA. At present, prediction from sRNA-seq is still the most popular strategy for novel miRNA discovery and annotation. Therefore, it is necessary to develop new statistical methodology for miRNA imprecise splicing analysis to improve the knowledge both regulation of miRNA splicing and novel miRNA annotation.

Here we first present GMAim, a comprehensive miRNA imprecise splicing detection and differential analysis Bioconductor/R package from sRNA-seq data that is species independent and biological replicates supported. The package and methods are general to use for analysis sRNA-seq alignment files that are generated from independent of sequencing platforms and preprocessing alignment methods. The model incorporated in our package is a modified negative binomial distribution that allows for larger biological replicates variance without any read normalization needed. Our package is powerful and versatile in miRNA imprecise splicing analysis with a simple and robust way of estimating the variance from the data as well as presenting useful diagnostics, such as imprecise splicing plots and detailed miRNA splicing variant analytical tool.

By comparing to other most widely used standalone or web-based tools for miRNA analysis such as miRDeep2[11], omiRas[12], miRanalyzer[13], sRNAbench[14], etc, we conducted in-depth comparative analysis between GMAim and other tools using the datasets of Arabidopsis thaliana. From the comparison results (Table 2), we found that GMAim has a comprehensive advantage over other tools in miRNA imprecise splicing annotation and differential expression of those miRNA variant products.

Table 2. Comparisons in major features and functions between GMAim and other popular tools



Program features	**GMAim**	miRDeep2	omiRas	miRanalyzer	miRExpress	sRNAbench
Known miRNA quantification	+	+	+	+	+	+
Novel miRNA prediction	+	+	+	+	+	+
Imprecision detection	+	-	-	-	-	-
Read cataloging	+	-	-	-	-	-
miRNA family expression quantification	+	-	-	-	-	-
Package type	R package	standalone	web service	standalone + web service	standalone	standalone + web service

Further work involves statistics model development including all pair-wise comparison involved more than two conditions. Also, improvement on the R code to speed up the implementation time will be under further development. Since small RNA-seq will also induce a lot of noise for the miRNA enrichment, we will as well incorporate the noise filtration step by removing the sRNA by length filter. Current GMAim code is available on the website of our lab (http://sysbio.unl.edu/GMAim). Users can download it and modify the input file and parameters to use it on your PC or workstation.

Acknowledgments

This project was supported by funding under the NSF ABI (Award #: DBI-1564621) to Chi Zhang (C.Z.). This work was supported by the Nebraska Soybean Board (Award #1728) and the National Science Foundation (Award #: OIA-1557417) to C.Z. This work was completed utilizing the Holland Computing Center of the University of Nebraska.

References

1. Bartel, D.P., *MicroRNAs: genomics, biogenesis, mechanism, and function.* Cell, 2004. **116**(2): p. 281-97.
2. Cuperus, J.T., N. Fahlgren, and J.C. Carrington, *Evolution and functional diversification of MIRNA genes.* Plant Cell, 2011. **23**(2): p. 431-42.
3. Ameres, S.L. and P.D. Zamore, *Diversifying microRNA sequence and function.* Nat Rev Mol Cell Biol, 2013. **14**(8): p. 475-88.
4. Pritchard, C.C., H.H. Cheng, and M. Tewari, *MicroRNA profiling: approaches and considerations.* Nat Rev Genet, 2012. **13**(5): p. 358-69.
5. Liu, C., M.J. Axtell, and N.V. Fedoroff, *The helicase and RNaseIIIa domains of Arabidopsis Dicer-Like1 modulate catalytic parameters during microRNA biogenesis.* Plant Physiol, 2012. **159**(2): p. 748-58.
6. Langmead, B., et al., *Bowtie: An ultrafast memory-efficient short read aligner.* Genome Biol, 2009. **10**(3): p. R25.
7. Kozomara, A. and S. Griffiths-Jones, *miRBase: annotating high confidence microRNAs using deep sequencing data.* Nucleic acids research, 2013. **42**(D1): p. D68-D73.
8. Morgan, M., et al., *Rsamtools: Binary alignment (BAM), FASTA, variant call (BCF), and tabix file import.* R package version, 2016. **1**(0).
9. Robinson, M.D., D.J. McCarthy, and G.K. Smyth, *edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.* Bioinformatics, 2010. **26**(1): p. 139-40.
10. Li, S., et al., *SMA1, a homolog of the splicing factor Prp28, has a multifaceted role in miRNA biogenesis in Arabidopsis.* Nucleic acids research, 2018. **46**(17): p. 9148-9159.
11. Friedlander, M.R., et al., *miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades.* Nucleic Acids Res, 2012. **40**(1): p. 37-52.
12. Muller, S., et al., *omiRas: a Web server for differential expression analysis of miRNAs derived from small RNA-Seq data.* Bioinformatics, 2013. **29**(20): p. 2651-2.
13. Hackenberg, M., et al., *miRanalyzer: a microRNA detection and analysis tool for next-generation sequencing experiments.* Nucleic Acids Res, 2009. **37**(Web Server issue): p. W68-76.
14. Barturen, G., et al., *sRNAbench: profiling of small RNAs and its sequence variants in single or multi-species high-throughput experiments.* Methods in Next Generation Sequencing, 2014. **1**(1).