miRPipe: A robust framework for identification of novel miRNAs from RNA-Seq data

VIVEK RUHELA

SBILab, IIIT-Delhi

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY
DELHI

Problem Statements

1. How to do the performance assessment of newly developed bioinformatics pipeline/tool/method?

2. How to identify the novel miRNAs and functionally similar miRNAs from RNA-Seq data with least false positive and false negative error?

- The recently published RNA-Seq analysis pipelines for the identification of novel miRNAs are as follows:
  - Mirnovo [3]
  - miRPro [4]
  - miRge2.0 [5]
  - sRNAtoolbox [6]
  - Mir&More2 [7]
  - Mirdeep2 (standalone tool) [8]
  - mirdeep* (standalone tool) [9]

Drawback:

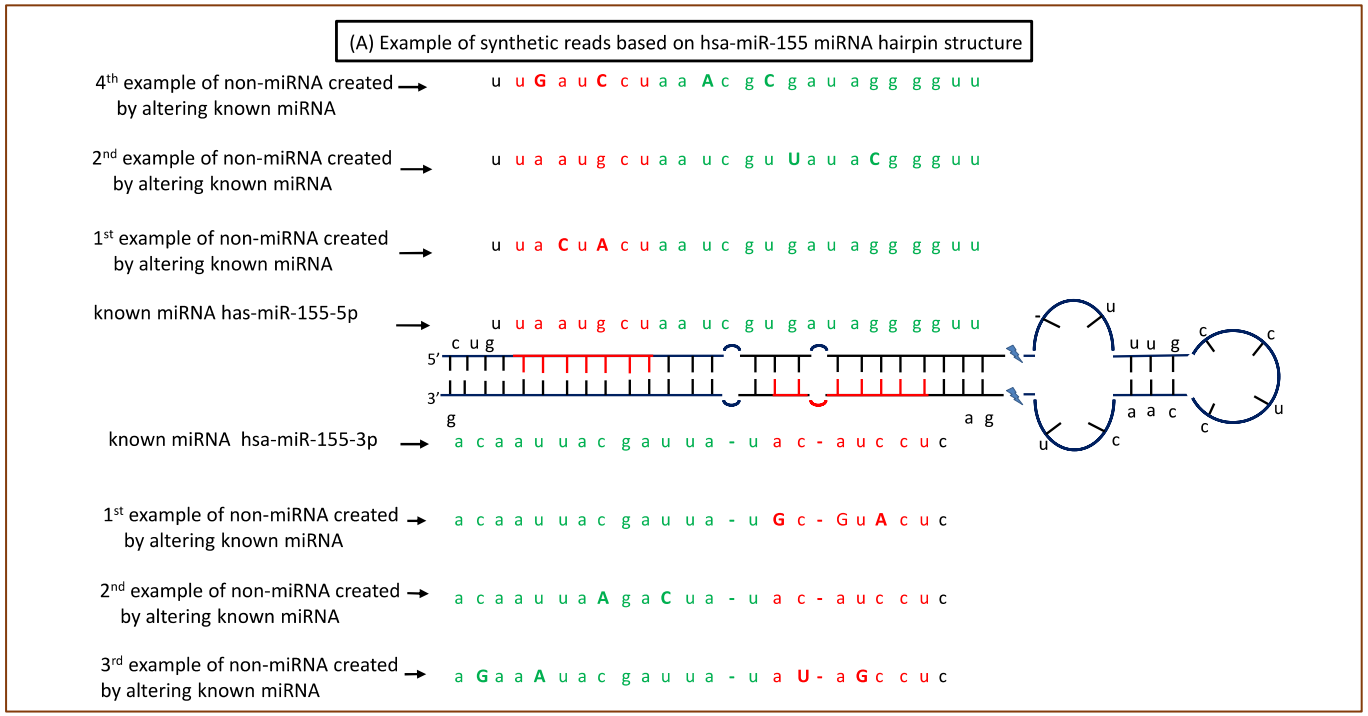
1. In general, the pipelines are benchmarked with the real RNA-Seq expression data where ground truth information is not available.
2. Further, recently published RNA-Seq pipelines do not identify the functionally similar miRNAs, reverse complement sequence of miRNAs as known miRNAs.

Workflow of miRSim: A synthetic RNA-Seq simulator

Step-1: Reference Files	❖ Provide an input file with the names and sequences of known miRNA collected from miRBase and, names and sequences of known piRNA collected from piRNADB.
Step-2: Provide specifications of synthetic data file	❖ Provide inputs for generating a synthetic data FASTQ file as follows: ✓ Total number of reads (N) ✓ % of known miRNAs (as percentage of total reads N), % of novel miRNAs, % of non-miRNA, % of piRNAs, and % of non-piRNA. ✓ Quality score encoding (33/64) and Adapter sequence, etc.
Step-3: No. of Sequences per Chromosome	❖ Compute the number of reads per chromosome (number of reads/chromosome are proportional to the number of miRNA present in every chromosome). ❖ Compute the number of miRNAs per chromosome such that each RNA sequence depth is greater than or equal to the minimum depth specified in Step-2 above.
Step-4: Expression Split	❖ Split the total read count per chromosome into the number of RNAs obtained in previous step such that the final expression counts of RNAs follow Poisson/gamma distribution.
Step-5: Sequence Generation	❖ Generate reads by randomly selecting the RNAs and by assigning expression counts.
Step-6: Generate Output files	❖ Prepare sequences of length 75 by adding adaptor, primer, and quality string such that the mean phred quality score is always greater than 20. ❖ Write fastq/fasta file with multiple threads where each thread is assigned to a small chunk ❖ Merged all the chunk output into a single fastq/fasta file.

Figure-4: Workflow of miRSim: synthetic RNA-Seq data simulator

Example of miRSim simulated synthetic RNA-Seq data



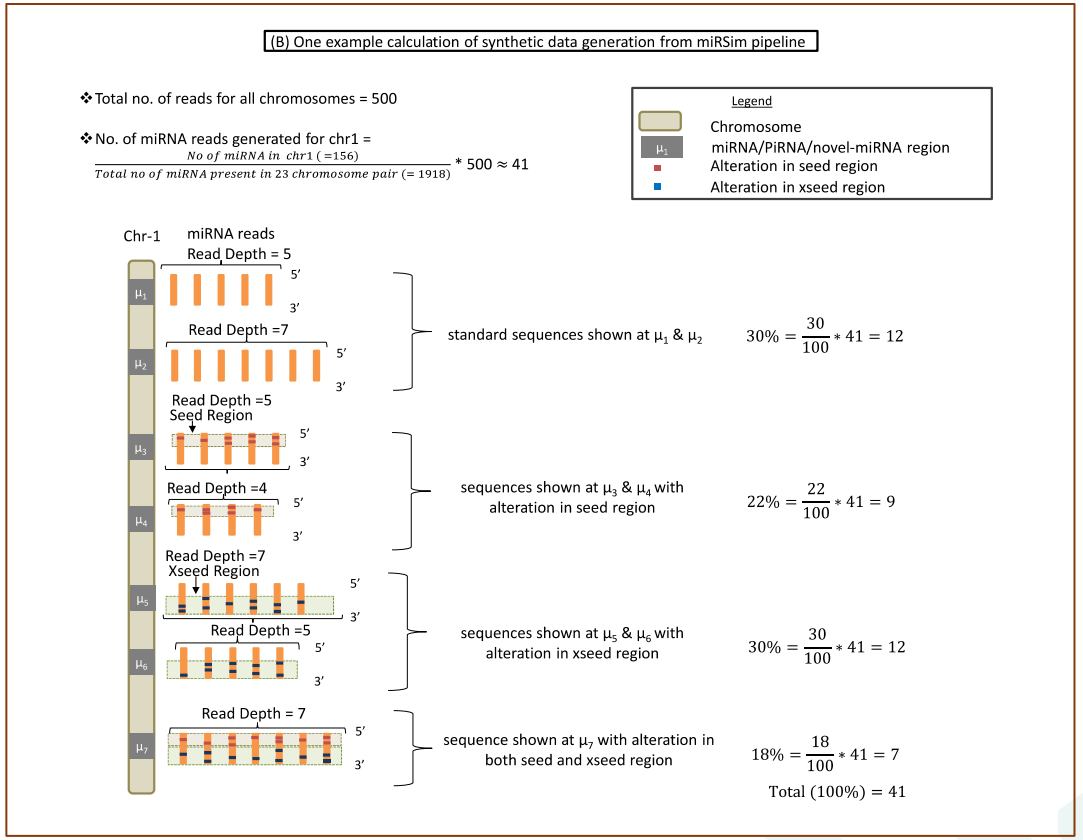


Figure-5: (A) Example of synthetic reads based on hsa-miR-155 miRNA hairpin structure. Red color shows seed region, green color shows xseed region and capital letters denote altered nucleotide, (B) One example data from miRSim pipeline. Here, the miRNA/piRNA region is represented by μ_1, μ_2... Here μ_1 represents original miRNA and μ_2 - μ_7 are derived from μ_1 by alterations in the seed and xseed sequence of μ_1 and may or may not constitute a valid miRNA. The number of miRNAs present in the chromosome-1 and total number of miRNAs in all chromosomes are taken from miRBase (version22)

Workflow of miRPipe pipeline for sncRNA identification from RNA-Seq data

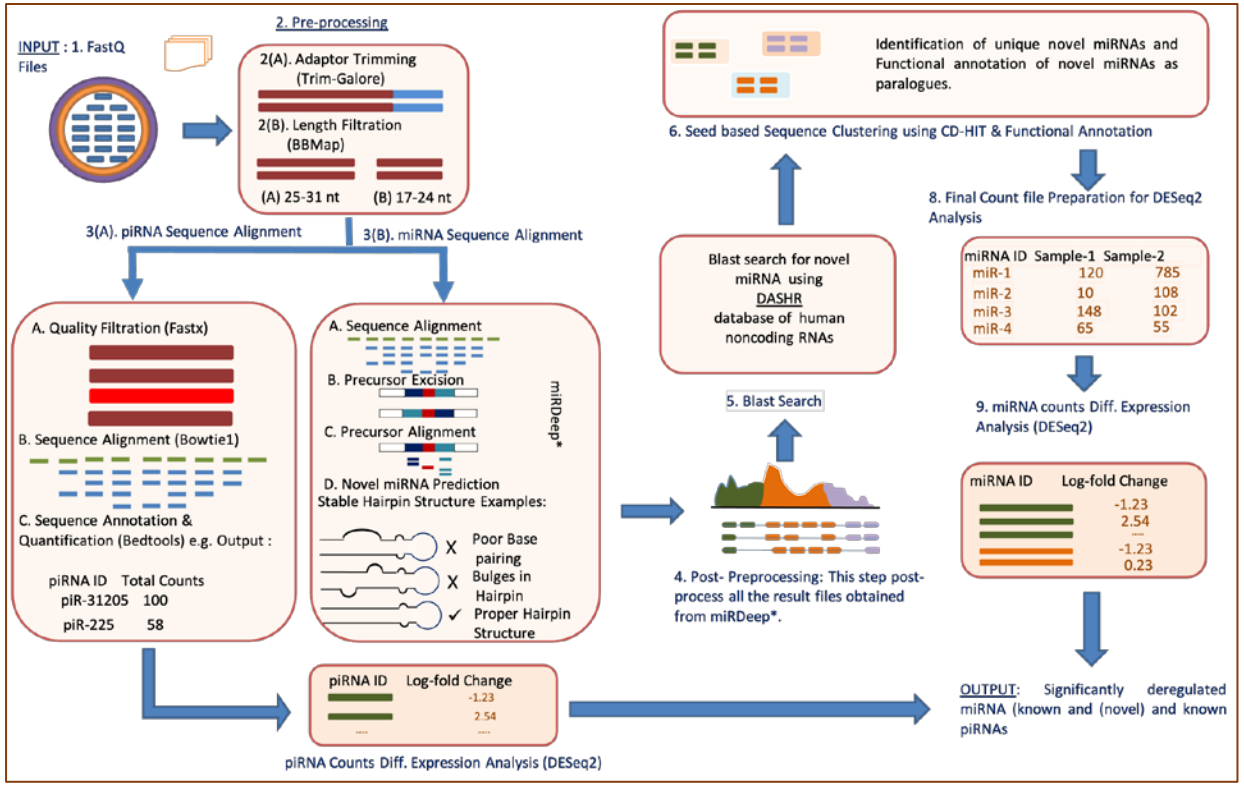


Figure-6: Workflow of miRPipe for the identification of sncRNAs from RNA-Seq data.

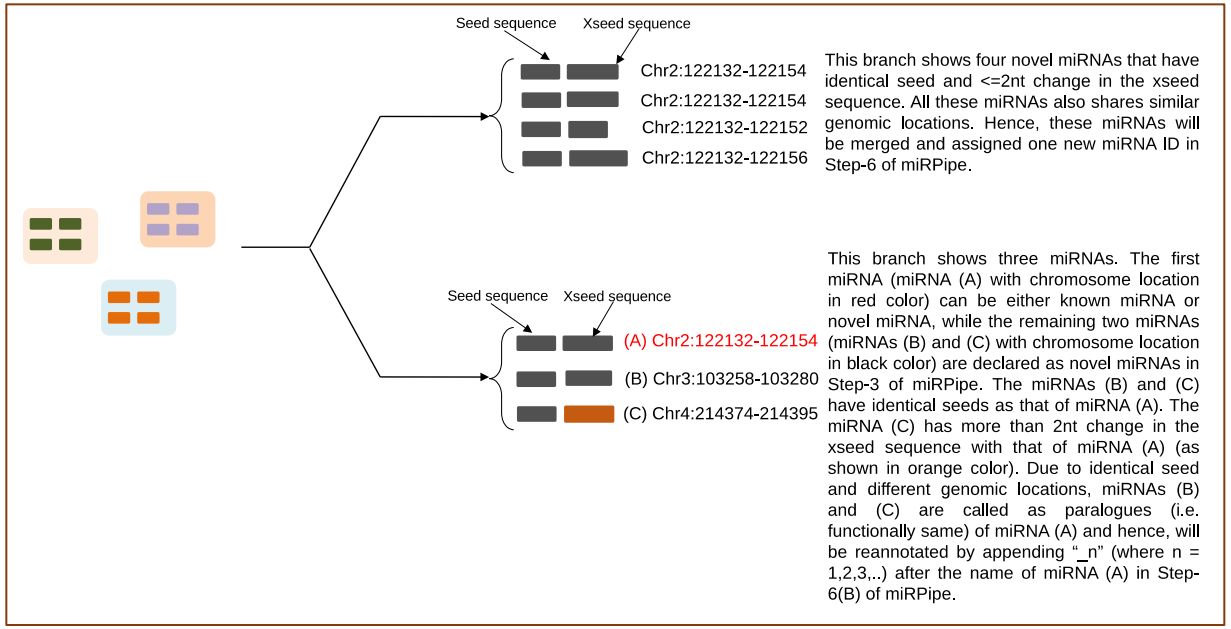


Figure-7: Functional annotation of novel miRNAs using seed-based clustering (Step-6 of miRPipe workflow).

Benchmarking of miRPipe pipeline on synthetic RNA-Seq data

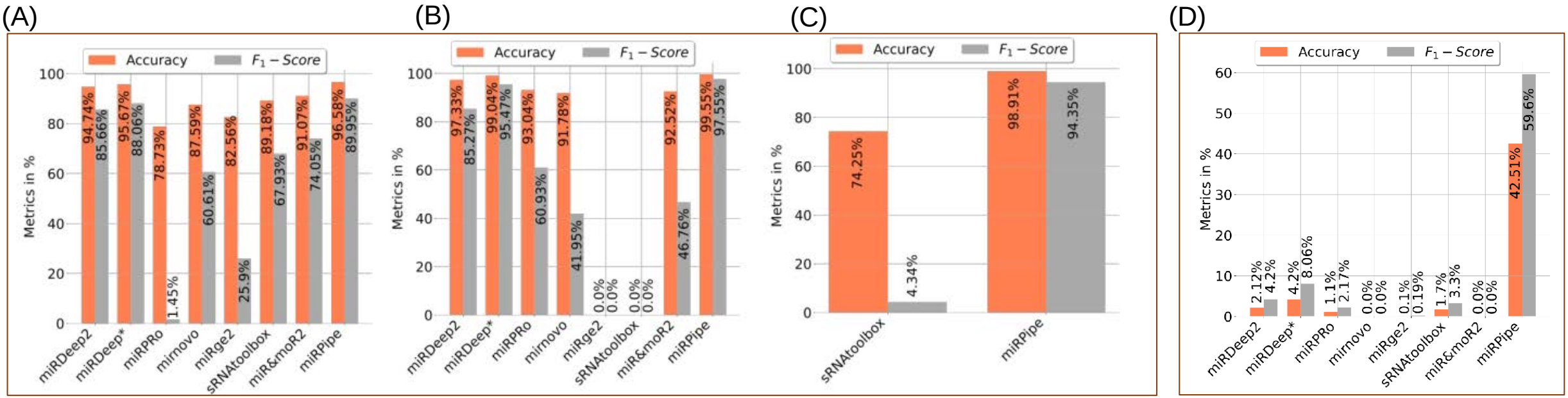


Figure-8: Average performance of miRPipe and existing pipelines for all read-depth categories (50K, 0.1 million, and 1 million) for identification of (A) known miRNAs (B) novel miRNAs, (C) known piRNAs and (D) reverse complement miRNA sequence as known miRNA

RT-qPCR and literature validation of miRPipe pipeline on real RNA-Seq data (Lung, Breast and CLL cancer data)

Table-2: (A) RT-qPCR validation of miRPipe pipeline with CLL RNA-Seq expression dataset, (B) Literature validation of miRPipe on three publicly available real RNA-Seq expression datasets (Lung cancer (GSE37764), Breast cancer (GSE171282), and CLL(GSE123436)). The cells with '-' indicates that pipeline does not identify that particular type of RNA.

Pipeline	Number of dysregulated miRNAs identified by pipeline	Number of miRNAs validated with RT-qPCR results	% False Positives	% False Negative
miRDeep2	29	9	68.96551724	70.96774194
miRDeep*	22	10	54.54545455	67.74193548
miRPro	34	12	64.70588235	61.29032258
mirnovo	32	6	81.25	80.64516129
miRge2	25	4	84.00	87.09677419
sRNAToolbox	5	1	80.00	96.77419355
miR&moR2	5	1	80.00	96.77419355
miRPipe	**31**	**17**	**45.16129032**	**45.16129032**

Pipeline	No. of dysregulated piRNAs identified by pipeline in lung cancer dataset (GSE37764)	No. of piRNAs reported in lung cancer related studies	No. of dysregulated miRNAs identified by pipeline in breast cancer dataset (GSE171282)	No. of miRNAs reported in breast cancer related studies	No. of dysregulated miRNAs identified by pipeline in CLL (GSE123436)	No. of miRNAs reported in CLL cancer related studies
miRDeep2	-	-	22	9 (40.90%)	29	17 (58.62%)
miRDeep*	-	-	8	7 (87.5%)	22	18 (81.81%)
miRPro	-	-	31	10 (32.5%)	36	21 (58.33%)
mirnovo	-	-	29	8 (27.58%)	25	14 (56%)
miRge2	-	-	34	19 (55.88%)	25	8 (32%)
sRNAToolbox	18	0 (0%)	14	12 (85.71%)	5	2 (40%)
miR&moR2	-	-	42	23 (54.76%)	5	2 (40%)
miRPipe	**20**	**6 (33.33%)**	**21**	**19 (90.47%)**	**31**	**27 (87.09%)**

Why miRPipe?

- Identification of functionally similar miRNAs and miRNA paralogues.
- Identification of reverse complement miRNA sequences as known miRNAs.
- Automated differential Expression analysis of sncRNAs (miRNAs and piRNAs).
- Simultaneous/standalone pipeline execution for dysregulated miRNA and piRNA identification.
- Nomenclature of novel miRNAs based on miRBase Nomenclature System.
- Also applicable to non-human genome RNA-Seq data processing after replacing the reference genome sequence.
- Available in dockerized version to ensure reproducibility.
- Developed in an interactive jupyter notebook to enhance interpretability and scalability.

Why miRSim?

- Generate the synthetic RNA-Seq data in a readable format with ground truth.
- Precise performance assessment of newly developed tool/method/pipeline.
- Also applicable to non-human genome with proper parameter tuning.
- Source codes are available at Zenodo and GitHub repositories.

Publications

1.Vivek Ruhela, Anubha Gupta, Sriram Krishnamachari, Gaurav Ahuja, Gurvinder Kaur, and Ritu Gupta. "miRPipe: A Unified Computational Framework for a Robust, Reliable, and Reproducible Identification of Novel miRNAs from the RNA Sequencing Data." Frontiers in Bioinformatics: 71.

2.Vivek Ruhela, Ritu Gupta, Sriram Krishnamachari,, Gaurav Ahuja, and Anubha Gupta (2021). “miRSim: Seed-based Synthetic Small Non-coding RNA Sequence Simulator.” Zenodo. https://doi.org/10.5281/zenodo.6546356.