miRSim: Seed-based Synthetic Small Non-coding RNA Sequence Simulator

Vivek Ruhela[1], Ritu Gupta[3]*, Sriram K.[1], Gaurav Ahuja[1], and Anubha Gupta[2]*

[1]Dept. of Computational Biology & Centre for Computational Biology, Indraprastha Institute of Information Technology-Delhi (IIIT-D), India

[2]SBILab, Deptt. of ECE & Centre of Excellence in Healthcare, Indraprastha Institute of Information Technology-Delhi (IIIT-D), India

[3]Laboratory Oncology Unit, IRCH, All India Institute of Medical Sciences (AIIMS), New Delhi, India.

*To whom correspondence should be addressed

Abstract

Summary: Micro RNAs (miRNA) regulate a number of cellular functions and are centrally classified in the networks of oncogenes and tumor suppressor genes. To extract the microRNAs from sequencing data, it's important to use a robust, stable and flexible pipeline that provides results with high accuracy and precision. However, due to the unavailability of a synthetic sequence simulator with known ground truth, it is difficult to assess the performance of existing bioinformatics pipelines in identifying dysregulated miRNAs from RNA-seq data. Here, we present miRSim tool that generates synthetic sequence reads of miRNAs, piRNAs, and altered sequences of miRNAs in the form of a FASTQ file. miRSim provides the functionality for the performance assessment of bioinformatics tools on the accurate identification of miRNAs, functionally similar miRNAs (paralogues), miRNAs belonging to the same phylogenetic trees, and piRNAs. miRSim also generates synthetic sequences of miRNAs altered in the seed and xseed (region other than the seed region of an miRNA sequence) regions of the known miRNA and piRNA sequences to allow researchers test the efficacy of the pipelines in rejecting altered reads.

Availability and Implementation: The source codes of synthetic miRNA sequence simulator tool, miRSim, are available at http://doi.org/10.5281/zenodo.4560585. The tool was developed on hardware configuration of Single Intel(R) Core(TM) i5-8400 CPU 2Cores, 4Threads, @Base frequency of 2.80GHz, 8GB DDR4 RAM.

Contact: vivekr@iiitd.ac.in

1 Introduction

The availability of a huge amount of high throughput next-generation sequencing data requires reliable and accurate bioinformatics tools for drawing correct inferences for disease diagnostics and treatment. The performance assessment of existing or the development of any new bioinformatics tool such as a sequence aligner or a sequence quantification tool is robustly possible only when the ground truth is available. This paper provides one such tool, namely miRSim, for the generation of synthetic FASTQ file containing synthetically created reads of miRNAs, piRNAs, and altered sequences of miRNAs and piRNAs. Presently,some of the synthetic sequence simulators available for the generation of synthetic sequencing data such as ART (1), pIRS (2), Flux (3), polyester (4), and RNA-Seq simulator (5), that can also be used to generate reads containing miRNA sequences. Tools ART, pIRS, and Flux are generic tools that can generate either single-end or paired-end DNA/RNA-seq data, while and polyester, RNA-Seq Simulator are specifically designed for synthetic RNA-seq data generation that is dependent on platform specific properties such as fragment size, PCR amplification parameters, error distribution

*Corresponding authors

model, etc. However, these tools do not provide the ground truth data for the validation of a bioinformatics pipeline. miRSim addresses this issue and provides synthetic reads with the known ground truth. Besides the known/novel miRNAs and piRNAs, it also generates synthetic sequences in the 'Other' category by altering sequences of miRNAs and piRNAs. This category is explained in the next Section.

2 Methods

The design and working of the miRSim tool are illustrated in Fig.1A. Mechanistically, the standard miRNA sequences can be stored in FASTA or GFF file formats (gff3) as the reference input to the miRSim tool. The miRNA sequences, piRNA sequences, and novel miRNAs sequences were collected from the miRBase (6), piRNAdb database (version 1.7.6) (https://www.pirnadb.org/), and the recent literature (7; 8), respectively. For the robustness of any bioinformatics pipeline meant for finding up-regulated or down-regulated miRNAs, it is important for a pipeline to robustly detect miRNAs, novel miRNAs, as well as highly similar paralogues miRNAs. Hence, miRSim provides the option to add a selected percentage of altered sequences of miRNAs and piRNAs as the 'Other' category. In this category, one can add altered sequences of known miRNAs. The seed and xseed regions (region other than the seed region) of an miRNA govern the functionality of miRNA in biological processes (9). Hence, this 'Other' category is created by altering the a) seed regions (red region of the miRNA sequence shown in Fig.1B), b) xseed regions (blue region in Fig.1B) of more than 2nt change, and in both regions (both red and blue region in Fig.1B) of the standard miRNA sequences. The resulting sequence will not be a miRNA or piRNA. This 'Other' category can act as the true negatives to assess a pipeline on its efficacy in identifying true positives (miRNA) from true negatives. The fraction of sequences for each of these error types is provided in the form of an error profile as input to the miRSim tool by the user. One example is shown in Fig.1C.

miRSim also provides the ground truth in a readable comma-separated file format that contains information about known miRNAs, piRNAs, and novel miRNAs along with their sequences, chromosome location, the expression counts, and the CIGAR string. For the 'Other' category sequences, it specifies the known miRNAs/piRNAs (with chromosome location) from which these altered sequence reads are generated, the sequences of altered miRNAs/piRNAs, the expression counts, and the CIGAR string (when compared to the known miRNA/piRNA from which these have been generated). These noisy reads should be discarded by any robust pipeline. miRSim delivers output in the form of a compressed FASTQ or FASTA file format.

3 Discussion

The synthetic data generated by miRSim tool can be used to assess the pipeline performance for accurate determination of small non-coding RNA identification. miRSim tool has been utilized recently to rigorously test and benchmark eight most promising state-of-the-art pipelines of known and novel miRNA detection from the RNA-seq data. This includes the newly proposed miRPipe (10), miRDeep2 (11), miRDeep* (12), miRPro (13), mirnovo (14), miRge2.0 (15), sRNAtoolbox (16), and miR&moR2 (17) in (10). The source codes of the miRSim tool (18) are available at zenodo (https://zenodo.org/) and GitHub (www.github.com) so that users can use it and add more functionalities, if required.

4 Conclusion

We have developed a synthetic sequence simulator, miRSim, that generates synthetic FASTQ files of sncRNA. This tool can be helpful in benchmarking pipelines on their ability to determine the biologically valid miRNAs and piRNAs.

Funding

This research was supported by grant from Department of Science and Technology, Govt. of India [Grant: DST/ICPS/CPS-Individual/2018/279(G)] and the Department of Biotechnology, Govt. of India [Grant: BT/MED/30/SP11006/2015]. Authors would also like to thank the Centre of Excellence in Healthcare, IIIT-Delhi, India.
Conflict of interest statement. None declared.

References

[1]W. Huang, L. Li, J. R. Myers, and G. T. Marth, "ART: a next-generation sequencing read simulator," *Bioinformatics*, vol. 28, no. 4, pp. 593–594, 2012.

[2]X. Hu, J. Yuan, Y. Shi, J. Lu, B. Liu, Z. Li, Y. Chen, D. Mu, H. Zhang, N. Li, *et al.*, "pIRS: Profile-based Illumina pair-end reads simulator," *Bioinformatics*, vol. 28, no. 11, pp. 1533–1535, 2012.

[3]T. Griebel, B. Zacher, P. Ribeca, E. Raineri, V. Lacroix, R. Guigó, and M. Sammeth, "Modelling and simulating generic RNA-Seq experiments with the flux simulator," *Nucleic acids research*, vol. 40, no. 20, pp. 10073–10083, 2012.

[4]A. C. Frazee, A. E. Jaffe, B. Langmead, and J. T. Leek, "Polyester: simulating RNA-seq datasets with differential transcript expression," *Bioinformatics*, vol. 31, no. 17, pp. 2778–2784, 2015.

[5]G. R. Grant, M. H. Farkas, A. D. Pizarro, N. F. Lahens, J. Schug, B. P. Brunk, C. J. Stoeckert, J. B. Hogenesch, and E. A. Pierce, "Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM)," *Bioinformatics*, vol. 27, no. 18, pp. 2518–2528, 2011.

[6]A. Kozomara, M. Birgaoanu, and S. Griffiths-Jones, "miRBase: from microRNA sequences to function," *Nucleic acids research*, vol. 47, no. D1, pp. D155–D162, 2019.

[7]G. Kaur, V. Ruhela, L. Rani, A. Gupta, K. Sriram, A. Gogia, A. Sharma, L. Kumar, and R. Gupta, "RNA-Seq profiling of deregulated miRs in CLL and their impact on clinical outcome," *Blood cancer journal*, vol. 10, no. 1, pp. 1–9, 2020.

[8]A. Wallaert, W. Van Loocke, L. Hernandez, T. Taghon, F. Speleman, and P. Van Vlierberghe, "Comprehensive miRNA expression profiling in human T-cell acute lymphoblastic leukemia by small RNA-sequencing," *Scientific reports*, vol. 7, no. 1, pp. 1–8, 2017.

[9]T. Kehl, C. Backes, F. Kern, T. Fehlmann, N. Ludwig, E. Meese, H.-P. Lenhof, and A. Keller, "About miRNAs, miRNA seeds, target genes and target pathways," *Oncotarget*, vol. 8, no. 63, p. 107167, 2017.

[10]V. Ruhella, A. Gupta, K. Sriram, G. Ahuja, and R. Gupta, "miRPipe: A Unified Computational Framework for Robust, Reliable, and Reproducible Identification of Novel miRNAs from Sequencing Data," *Co-Submitted to Nucleic Acids Research*, pp. 1–12, 2021.

[11]M. R. Friedländer, S. D. Mackowiak, N. Li, W. Chen, and N. Rajewsky, "miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades," *Nucleic acids research*, vol. 40, no. 1, pp. 37–52, 2012.

[12]J. An, J. Lai, M. L. Lehman, and C. C. Nelson, "miRDeep*: an integrated application tool for miRNA identification from RNA sequencing data," *Nucleic acids research*, vol. 41, no. 2, pp. 727–737, 2013.

[13]J. Shi, M. Dong, L. Li, L. Liu, A. Luz-Madrigal, P. A. Tsonis, K. Del Rio-Tsonis, and C. Liang, "mirPRo–a novel standalone program for differential expression and variation analysis of miRNAs," *Scientific reports*, vol. 5, p. 14617, 2015.

[14]D. M. Vitsios, E. Kentepozidou, L. Quintais, E. Benito-Gutiérrez, S. van Dongen, M. P. Davis, and A. J. Enright, "Mirnovo: genome-free prediction of microRNAs from small RNA sequencing data and single-cells using decision forests," *Nucleic Acids Research*, vol. 45, no. 21, pp. e177–e177, 2017.

[15]Y. Lu, A. S. Baras, and M. K. Halushka, "miRge 2.0 for comprehensive analysis of microRNA sequencing data," *BMC bioinformatics*, vol. 19, no. 1, p. 275, 2018.

[16]E. Aparicio-Puerta, R. Lebrón, A. Rueda, C. Gómez-Martín, S. Giannoukakos, D. Jaspez, J. M. Medina, A. Zubkovic, I. Jurak, B. Fromm, *et al.*, "sRNAbench and sRNAtoolbox 2019: intuitive fast small RNA profiling and differential expression," *Nucleic acids research*, vol. 47, no. W1, pp. W530–W535, 2019.

[17]E. Gaffo, M. Bortolomeazzi, A. Bisognin, P. Di Battista, F. Lovisa, L. Mussolin, and S. Bortoluzzi, "MiR&moRe2: A Bioinformatics Tool to Characterize microRNAs and microRNA-Offset RNAs from Small RNA-Seq Data," *International journal of molecular sciences*, vol. 21, no. 5, p. 1754, 2020.

[18]V. Ruhela, R. Gupta, K. Sriram, G. Ahuja, and A. Gupta, "vivekruhela/miRSim v1.0.0 (Version v1.0.0)," *Zenodo. https://doi.org/10.5281/zenodo.4560585*, Feb 2021.

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 the inputs for generating a synthetic fastq file containing reads of known miRNA, non-miRNA, novel miRNA, known piRNA and non-piRNA: ✓ Total number of reads (N); % of known miRNAs (as percentage of total reads N), % of novel miRNAs, % of altered-miRNA, % of piRNAs, and % of altered-piRNA. ✓ Quality score encoding (33/64) and Adapter sequence. ✓ Minimum depth and expression profile distribution (Poisson/gamma)
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 ❖ Merge all the chunk output into a single fastq/fasta file.

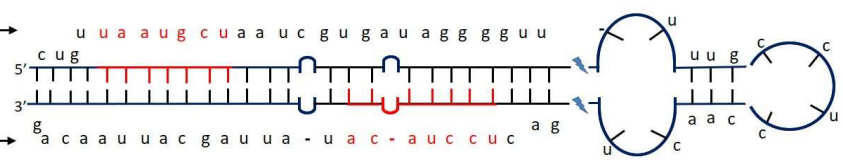
(A) miRSim Workflow

R4:4th example of non-miRNA created by altering known μ RNA → u u **G** a u **C** c u a a **A** c g **C** g a u a g g g g u u

R3: 2nd example of non-miRNA created by altering known μ RNA → u u a a u g c u a a u c g u **U** a u a **C** g g g u u

R2: 1st example of non-miRNA created by altering known μ RNA → u u a **C** u **A** c u a a u c g u g a u a g g g g u u

R1: known μ RNA has-miR-155-5p → u u a a u g c u a a u c g u g a u a g g g g u u



r1: known μ RNA hsa-miR-155-3p → g a c a a u u a c g a u u a - u a c - a u c c u c a g

r2: 1st example of non-miRNA created by altering known μ RNA → a c a a u u a c g a u u a - u **G** c - G u **A** c u c

r3: 2nd example of non-miRNA created by altering known μ RNA → a c a a u u a **A** g a **C** u a - u a c - a u c c u c

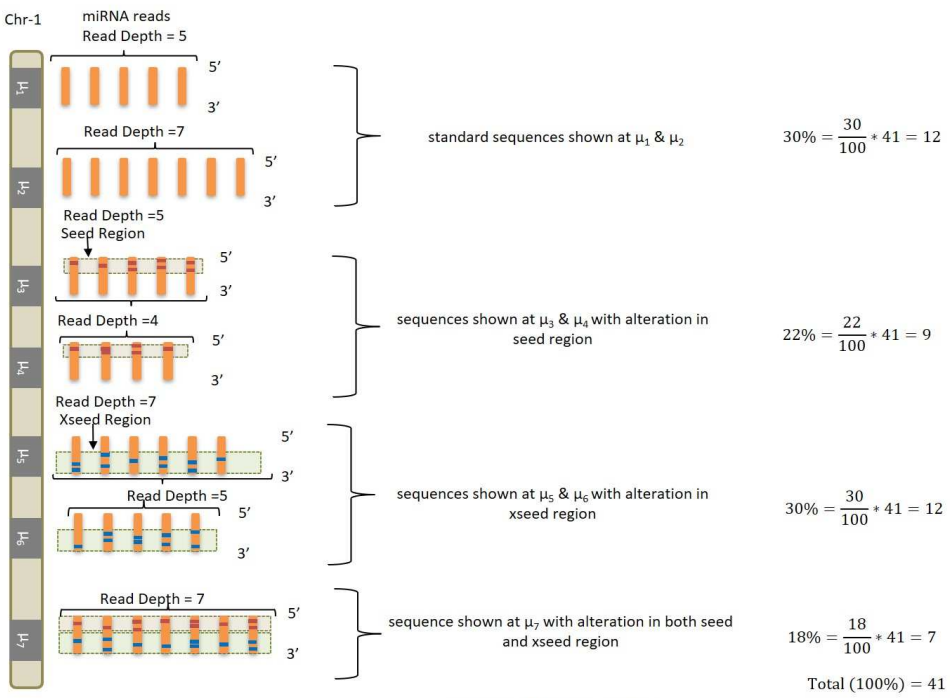
r4: 3rd example of non-miRNA created by altering known μ RNA → a **G** a a **A** u a c g a u u a - u a **U** - a **G** c c u c

(B) Example of synthetic reads based on hsa-miR-155 miRNA hairpin structure

❖ Total no. of reads for all chromosomes = 500

❖ No. of miRNA reads generated for chr1 = $\frac{No\ of\ miRNA\ in\ chr1\ (=156)}{Total\ no\ of\ miRNA\ present\ in\ 23\ chromosome\ pair\ (=1918)} * 500 \approx 41$

Legend
- Chromosome
- μ_1 miRNA/PiRNA/novel-miRNA region
- Alteration in seed region
- Alteration in xseed region



(C) Outline of miRSim pipeline

Fig. 1. (A) Workflow of the newly designed synthetic sncRNA read simulator miRSim Tool, (B) Example of synthetic reads based on hsa-miR-155 miRNA hairpin structure, (C) Outline of miRSim pipeline. Here, the miRNA/piRNA region is represented by μ_1, μ_2, ... The number of miRNAs present in the chromosome-1 and total number of miRNAs in all 23 chrmosomes are taken from miRBase (version22) (6)