Optimizing a Small RNAseq Analysis Pipeline for NASA GeneLab Using Open-Source Tools and Libraries

Authors: Richard Barker, Amanda M. Saravia-Butler, Alexis Torres, Mike Lee, Lauren M. Sanders, Samrawit Gebre, Sylvain V. Costes

NASA Ames Research Center Moffett, Field, CA, USA

Blue Marble Space Institute of Science, Seattle, WA, USA Space Biosciences Division

To collaboration on this pipeline please contact richard.barker@NASA.gov

Introduction

Small RNA sequencing (small RNAseq) is a powerful tool for studying the regulation of gene expression in various organisms. Small RNAseq has been leveraged in space biology research to study how expression of small RNAs such as microRNAs (miRNAs), small interfering RNAs (siRNAs), and piwi-interacting RNAs (piRNAs), change upon exposure to the space environment. NASA GeneLab currently hosts both total RNAseq and small RNAseq raw data derived from space-relevant experiments within the Open Science Data Repository (OSDR). In addition to hosting raw data, which is only interpretable by bioinformaticians, GeneLab plans to process all small RNAseq datasets and make those processed data available to the scientific community to maximize the accessibility of these data to a broader scientific community. In this study, we present the development of the GeneLab standardized pipeline for processing small RNAseq datasets.

The diagram below summarizing small RNA's production and maturation of miRNAs involve transcription of pri-miRNA in the nucleus, processing of pri-miRNA into pre-miRNA by Drosha, export of pre-miRNA to the cytoplasm, processing of pre-miRNA into mature miRNA by Dicer, and loading of mature miRNA onto the RISC complex for target mRNA regulation. The main difference between plants and animals is that Dicer functions in the nucleus in plants (Campos ÁL, *et al.*, 2016).

Using publicly available synthetic small RNAseq datasets, we interrogated various open-source software to evaluate their accuracy and reproducibility in each step of the pipeline. For adapter trimming and quality filtering, we evaluated TrimGalore, FASTX, and Cutadapt. We evaluated alignment to both small RNA references (miRBase, MirGeneDB) and reference genomes (Ensembl) and compared the BWA, Bowtie, and Bowtie2 alignment tools using different parameter combinations to determine the optimal alignment method. To quantify the aligned data, we compared SAMtools, HTSeq, FeatureCount, and RSEM. Finally, we evaluated various software, including DESeq2, EdgeR, Limma trend, and Limma voom, for data normalization and subsequent differential expression analysis.

Small RNA processing

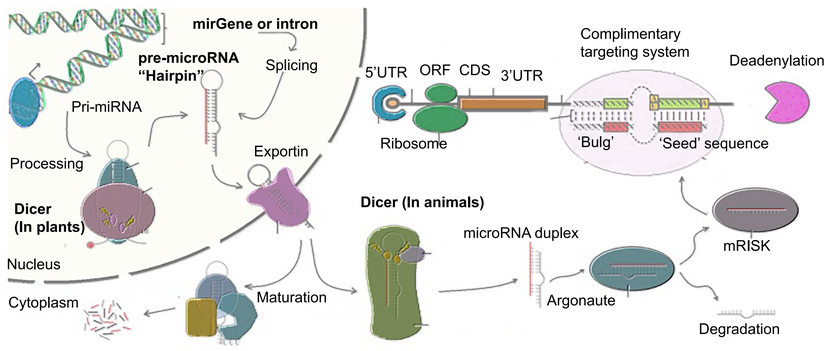

Figure 1: Regulator ncRNAs biogenesis landscape and functions in plants. The inside of the partial circle represents the nucleus, and the outside represents cytoplasm. The locations of pri-miRNAs, precursor hairpin non-coding RNAs and their processing by Dicer, Argonaut, and mRISK that lead to mature miRNA production and a functional complimentary targeting system. Reproduced and customized with permission from review by Campos ÁL, et al., (2016).

Methods summary

Trimming and filtering

- Tools compared: Cutadapt**, TrimGalore**, Seqkit
- Purpose: Remove adapter sequences, trim low-quality bases, and filter out low-quality reads
- Output: Trimmed/Filtered FASTQ files

Alignment

- Tools compared: BWA, Bowtie (BT), **Bowtie2 (BT2)**, BT_BestStrata, BT2_VerySenitive, BT2_VerySensitiveLocal
- Purpose: Compare alignment accuracy using a range of miRNA specific templates from miRBase and MirGeneDB
- Output: Alignment files (SAM, BAM)

Quantification

- Tools compared: **samtools idxstats**, RSEM, FeatureCount, HTSeq
- Purpose: Evaluate different quantification methods of alignment data from different combinations of trimming/filtering and alignment tools.
- Output: miRNA transcript raw counts

Differential Expression Analysis

- Tools compared: **DESeq2**, EdgeR, Limma Trend, Limma Voom.
- Purpose: Identify differentially expressed (DE) small RNAs between different conditions or treatments
- Output: DE metrics, log-fold changes (FC), p-values, and adjusted p-values

Annotations

- Databases compared: **mirbase.db,** mirGeneDB, TarBase, and BioMART
- Purpose: Assign descriptions and details related to specific genomic features (e.g., miRNA, hairpin, targets etc…)
- Output: Annotations for microRNAs, mature/hairpins, and their potential targets

Visualization

- Visual assessment tools: MDS, Bar plot, PCA, Heatmap, ideogram, and an example of Knowledge network mining (Knetminer)
- Purpose: Visualize and compare results from different tool combinations.
- Output: Data visualizations

Figure 2: General analysis workflow used for processing small RNAseq data using different combinations of tools for pipeline optimization. The best performing combination highlight in **bold**.

Draft small RNAseq pipeline

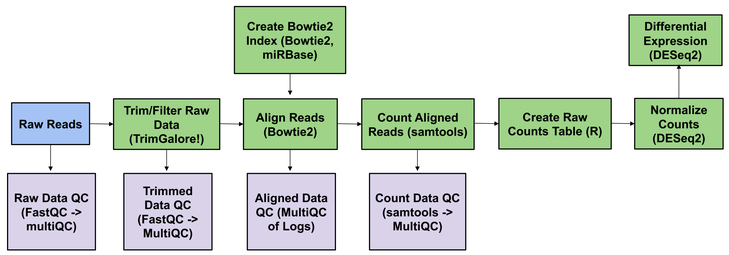

Figure 3: Semantic representation of the proposed GeneLab pipeline. Tools used for each step are indicated. Processed data outputs from each step will be published on the Open Science Data Repository (OSDR), and expression and statistical tables produced by these analysis tools will be visualized using the OSDR visualization platform.

Results

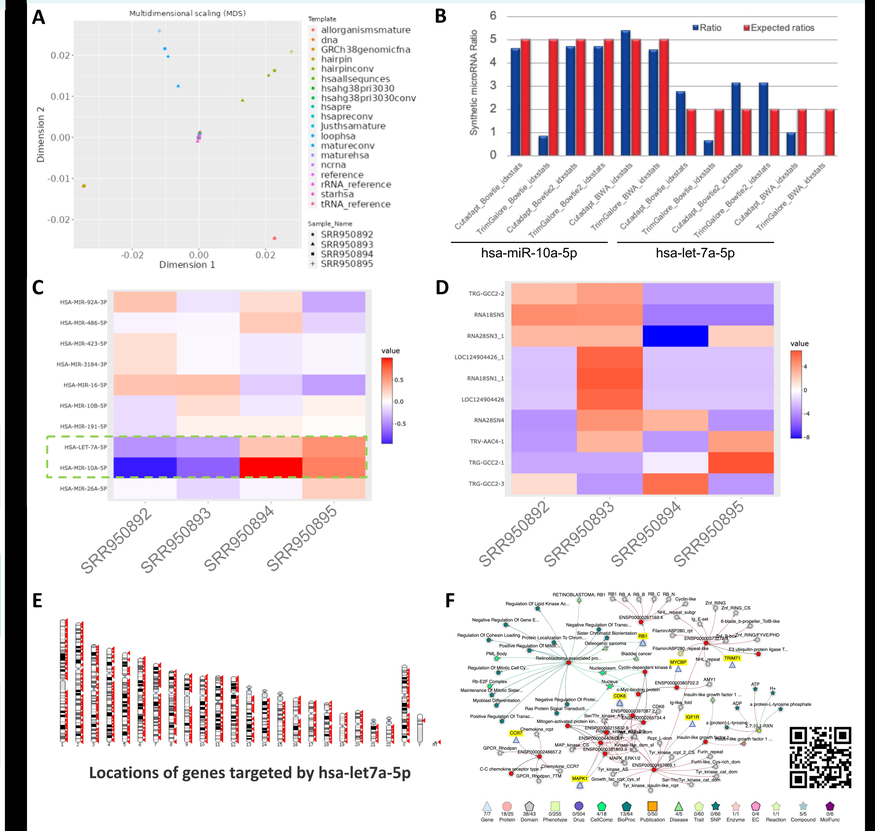

Figure 4: (A) Multidimensional scaling comparison of different references using the Cudatapt -> Bowtie2 -> idxstats -> log(CPM) pipeline. (B) Bar plot comparing the transcript ratios of synthetic spiked microRNAs using different pipelines. (C-D) Top 10 most DE miroRNAs based on Euclidean hierarchical clustering from the C) Cudatapt -> Bowtie2 -> idxstats pipeline using the miRBase mature reference and the D) Cudatapt -> Bowtie2 -> RSEM pipeline using the Ensembl genome reference. (E) Ideogram show potential has-let7a-5p targets from TarBase. (F) Knetminer knowledge network from experimentally supported targets (QRcode for the interactive version is provided).

Conclusion

Various tools and references were evaluated to optimize a small RNAseq analysis pipeline for small RNAseq data hosted on OSDR. The pipeline includes trimming/filtering, alignment, quantification, differential expression analysis, and quality assessment after each step. The accuracy of the results were compared to select the best approach, revealing increased accuracy and easier interpretation when microRNA-specific libraries were used as templates instead of whole genomes. The pipeline provides an efficient solution for small RNA analysis using *.fastq files from either small RNAseq or total RNAseq as potential inputs.

References

Campos ÁL, et al., (2016). **Tools for Sequence-Based miRNA Target Prediction: What to Choose**? Int J Mol Sci. 9;17(12):1987.

Ziemann M, Kaspi A, El-Osta A. (2016) **Evaluation of microRNA alignment techniques.** RNA. 22(8): 1120–1138.

Griffiths-Jones, Grocock, van Dongen, Bateman, and Enright (2006). **miRBase: microRNA sequences, targets and gene nomenclature.** Nucleic Acids Res. 1;34 (Database issue).

Sethupathy, Corda, and Hatzigeorgiou (2006). **TarBase: A comprehensive database of experimentally supported animal microRNA targets**. RNA. 12(2):192-7.

Love MI, Huber W, Anders S (2014). **Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2**. *Genome Biology*, **15**, 550.

Hassani-Pak et al., (2021**). KnetMiner: a comprehensive approach for supporting evidence-based gene discovery and complex trait analysis across species.** Plant Biotechnol Journel. 1670-1678. doi: 10.1111/pbi.13583.