

Approaches for sRNA analysis of human rna-seq data: comparison, benchmarking

Bezuglov V.^{1,2*}, Skakov I.², Shtratnikova V.¹, Sergeyev O.^{1**}, Stupnikov A.^{3,4}

¹ Belozersky Institute of Physico-Chemical Biology, Lomonosov Moscow State University, Moscow, Russia

² Faculty of Bioengineering and Bioinformatics, Lomonosov Moscow State University, Moscow, Russia

³ Moscow Institute of Physics and Technology, Moscow, Russia

⁴ National Medical Research Center for Endocrinology, Moscow, Russia

*vitaly1530@gmail.com; ** olegsergeyev1@yandex.ru

Key words: transcript annotation; Differential Gene Expression; small RNA; microRNA; piRNA; small RNA fragments; transcriptomics; benchmarking

Motivation and Aim: Transcriptomics analysis of various small RNA (sRNA) biotypes is a new and rapidly developing field. However, the bioinformatic analysis of NGS data for sRNA is prone to many challenges and not yet well established. Here, we attempt to identify the optimal pipeline configurations for each step of the sRNA analysis of human data, including read trimming, filtering, mapping, transcript abundance quantification, and differential expression (DE) analysis. Also we calculated the quality of obtained DE analysis signatures to estimate robustness of the obtained gene signature with robust and efficient rank statistics based approach.

Methods and Algorithms: Small RNA raw sequencing data from 7 published human studies [1–7]. All read adapters were removed, following lab kit protocols. To adjust the pipeline we used several trimming options for the upper bound. The trimmed reads were processed with various alignment (bowtie, hisat2, STAR, rsem) and pseudoalignment (salmon, kallisto) methods. Then three filtering strategies (by mean, median count and default DE analysis methods filtering) with several thresholds were applied to reduce noise. DE analysis was conducted using DESeq2, edgeR, limma, EBSeq and NOISeq packages. The number of DE transcripts cannot be used as a robust metric for DE analysis quality, as this approach does not account for false-positive results. To evaluate the expression signature quality, we applied the previously published Hobotnica approach [8]. Also we conducted an analysis with permuted group labels to detect false positive signatures and calculate FPR and Jaccard index.

Results: The most popular and usable methods of sRNA analysis were carried out. Based on our analysis, we suggest a pipeline that produces robust DE analysis results for sRNA transcripts, at least for categorical factors and two-group comparisons of biosamples (Table 1). We used existing tools to construct an optimal pipeline for quality sequencing data analysis, regardless of the differences in input data. To account for data variation in the original datasets, flexible trimming thresholds were applied. We suggest lower bound and upper bound of read length for trimming, thresholds for filtering expression data for the analysis of two groups. According to the results of DE analyses we suggest the DGE method for data with strong and well-detected signals.

Conclusion: The effect of various factors that impact the expression analysis of human sRNA at different stages of data processing were investigated. The optimal pipeline (trimming, aligning, assigning, filtering, DE analysis) parameters were suggested on Table 1, and an optimized pipeline for setting and running sRNA expression analysis

was proposed. Assessing the resulting expression signatures with rank-statistics-based inference suggests a way to estimate the quality of resulting signatures and performance of bioinformatic analysis for particular biological data. Results of this work were published [9] and new approaches (filtering thresholds and DE methods) were added to improve our suggestions.

Table 1. Recommended pipeline for sRNA analysis

Stage	Pipeline Command	Justification
Trimming	lower bound: 15 upper bound: Read length – 40% of adapter length	Retain sufficient and the same number of reads after trimming for downstream analyses for all datasets
Aligning	bowtie aligner with 1 mismatch allowed	The high alignment rate and H-score for all datasets
Assigning	ITAS	Optimized annotation for small RNA
Filtering	default DE method filter	Sufficient number of transcripts for the downstream analysis and higher H-score
DE analysis	DESeq2	Sufficient number of significant transcripts and high H-score

Funding: The study was supported by the Russian Science Foundation grant number 18-15-00202, <https://rscf.ru/project/18-15-00202>.

References

- Donkin I., Versteyhe S., Ingerslev L.R. et al. Obesity and bariatric surgery drive epigenetic variation of spermatozoa in humans. *Cell Metab.* 2016;23(2):369-378
- Ingerslev L.R., Donkin I., Fabre O. et al Endurance training remodels sperm-borne small RNA expression and methylation at neurological gene hotspots. *Clin Epigenet.* 2018;10:12. doi 10.1186/s13148-018-0446-7
- Hua M., Liu W., Chen Y. et al. Identification of small non-coding RNAs as sperm quality biomarkers for in vitro fertilization. *Cell Discovery.* 2019;5:20. doi 10.1038/s41421-019-0087-9
- Wong R.K., MacMahon M. et al. A comparison of RNA extraction and sequencing protocols for detection of small RNAs in plasma. *BMC Genomics.* 2019;20(1):446
- Huang G., Cao M. et al. Small RNA-sequencing identified the potential roles of neuron differentiation and MAPK signaling pathway in dilated cardiomyopathy. *Biomed Pharmacother.* 2019;114:108826
- Kanth P., Hazel M.W. et al. Small RNA sequencing of sessile serrated polyps identifies microRNA profile associated with colon cancer. *Genes Chromosom Cancer.* 2019;58:23-33
- Morgan C.P., Shetty A.C. et al. Repeated sampling facilitates within- and between-subject modeling of the human sperm transcriptome to identify dynamic and stress-responsive sncRNAs. *Sci Rep.* 2020;10:17498
- Stupnikov A., Sizykh A., Budkina A. et al. Hobotnica: exploring molecular signature quality. *F1000Res.* 2021;10:1260. doi 10.12688/f1000research.74846.2
- Bezuglov V., Stupnikov A., Skakov I. et al. Approaches for sRNA analysis of human RNA-seq data: comparison, benchmarking. *Int J Mol Sci.* 2023;24:4195. doi 10.3390/ijms24044195