International Journal of Research in Science & Engineering
ISSN: 2394-8299
Vol: 03, No. 02, Feb-Mar 2023
http://journal.hmjournals.com/index.php/IJRISE
DOI: https://doi.org/10.55529/ijrise.32.26.32

Comparable Analysis of COMPSRA and Excerpt Pipelines for Mining Distinct Molecules of RNA

Pooja Chhabra[1*], Brij Mohan Goel[2], Reena Arora[3]

[1,2]Department of Computer Science and Applications, Baba Mastnath University, Rohtak, Haryana, India.*
[3]Department of Biotechnology, ICAR-National Bureau of Animal Genetic Resources, Karnal, Haryana, India.

Corresponding Email: [1]poojachhabra31@gmail.com*

Received: 17 October 2022 **Accepted:** 06 January 2023 **Published:** 10 February 2023

Abstract: ***The COMPSRA and exceRpt pipelines that are used explicitly for quantifying RNAseq data were compared and evaluated in this study. In both pipelines, various tools are used to extract various kinds of RNAs from a given sample. Small RNA sequencing data from milk somatic cell samples from 12 buffaloes were compared using the COMPSRA and exceRpt analyses. The two selected pipelines were also evaluated from a variety of angles, including the length of time required for comparing the sequences, the types of supported databases for annotation, and the number of distinct RNAs produced as results. The output varies even though the pipelines are used for similar purposes, because different quantification techniques are used for transcriptomic data. When the two pipelines were compared, it became clear that both had drawbacks. For example, exceRpt's analysis time was very high, while COMPSRA's count of generated specific RNA was lower. In contrast to exceRpt, which detected abundance of tRNA, rRNA, miRNA, snRNA, snoRNA, and lncRNA, COMPSRA found circRNA and piRNA to have a higher level of diversity and abundance.***

Keywords: COMPSRA, Pipelines, RNA Databases, Excerpt, RNA.

1. INTRODUCTION

Understanding the regulation of genes requires knowledge of the functional components of a genome. The function of a gene is determined by different types of RNA. A critical component of Next Generation Sequencing (NGS) based small RNA analysis is the identification and quantification of the small RNAome components. The parts of RNA can now be identified more precisely and effectively due to a variety of new data mining techniques. An integrated pipeline is required for evaluating small RNAomes, which is a requirement for downstream

International Journal of Research in Science & Engineering
ISSN: 2394-8299
Vol: 03, No. 02, Feb-Mar 2023
http://journal.hmjournals.com/index.php/IJRISE
DOI: https://doi.org/10.55529/ijrise.32.26.32

analyses. RNAs come in a variety of forms, including fragmented mRNA, snoRNA, snRNA, miRNA, piRNA, and other noncoding RNAs. A reliable and integrated pipeline is necessary for identifying and profiling the RNA. A great deal of research has been done on the full range of small RNAs found in biological systems. A versatile and effective tool for processing and analysing miRNA-seq data is Cap-miRSeq [1]. RNAseq data analysis is available on the web servers Oasis 2 [2] and miRMaster [3]. Potential exogenous miRNAs can be found using miRMaster. SRNAnalyzer [4] can also be used to analyse miRNA. Small RNAs can be searched for using web-based tools from sRNAtoolbox [5]. MicroRNAs (miRNAs) from NGS data can be annotated and functionally analysed using the CPSS [6] web server. Small RNA non-coding RNA profiling is carried out using a pipeline known as ncPRO-seq [7]. These pipelines were created to find particular RNA molecules. RNA profiling is also carried out using the DARIO [8] web server, but the size of submitted data is constrained. A primary goal of this paper was to compare two pipelines, COMPSRA [9] and exceRpt [10], in order to find one that could produce various types of RNAs with minimal computational resources. The exceRpt, a web-based pipeline that works with the Extracellular RNA Communication Consortium to analyse RNA profiles, is superior to COMPSRA, an interactive command-based pipeline. A comparison of this kind would make it possible to choose the best pipeline for RNA molecule extraction from a dataset without sacrificing the quantity or quality of the data.

2. METHODS AND MATERIALS

Hardware Requirements

Both pipelines were implemented on hardware with a minimum of 16 GB of memory and a 4 core CPU. The alignment may not run smoothly on machines with fewer than four CPU cores. Moreover, each reference index after unzipping took up about 35GB of disc space.

Description of pipelines

The exceRpt [10] is a web-based pipeline, while COMPSRA [9] is a command-based pipeline. JRE (Java Runtime Environment) is necessary for COMPSRA's command-based execution. Unlike COMPSRA, which only supports .fq files, exceRpt can work with.fq, .sra and .fa files. The count profile of each RNA is output by both COMPSRA and exceRpt as a text file. Small RNA references databases used by COMPSRA for annotation of mapped reads include: GENCODE version 27 [12] for mining snoRNA and snRNA; GtRNAdb [11] for extracting tRNA; piRNA cluster [13], piRBase [15], piRNABank [14] for piRNA discovery; miRBase [16] for miRNA extraction; and circBase [17] for extracting circular RNA. However, exceRpt mines piRNAs from piRNABank; rRNA from 45S, 5S; circular RNAs from circBase; tRNAs from GtRNAdb; miRNAs from miRBase and GENCODE version 24 annotations in addition to other sources.

Preprocessing Module or Quality Control (QC), Alignment, Annotation, Microbe (optional), and Function (optional) are the five modules that make up COMPSRA [9], each with their own functionality and customization options. In contrast, exceRpt [10] consists of preprocessing filtering (QC), endogenous alignment, filtering, and exogenous alignment.

International Journal of Research in Science & Engineering
ISSN: 2394-8299
Vol: 03, No. 02, Feb-Mar 2023
http://journal.hmjournals.com/index.php/IJRISE
DOI: https://doi.org/10.55529/ijrise.32.26.32

Preprocessing Module (QC)

12 buffalo milk somatic cells' RNA sequencing (RNAseq) reads were used as input in the form of .fq files. Prior to discarding reads with low quality (average phred 30), the adapter portion of the reads were first trimmed in COMPSRA [9]. The reads that were used as input for additional processing had to be at least 20 bp long. After the adapters were removed, the reads' lengths varied from 20 nt to 101 nt from their original length of 101 nt. To enable bias-free comparison of the pipelines, similar trimmed files produced by putting them through the aforementioned filters were fed into exceRpt and aligned to the 45S, 5S, and mt rRNAs. In order to find RNA molecules longer than 200 nucleotides, such as lncRNA, the overlapping reads were combined and mapped to the appropriate databases.

Alignment with Genome

While the COMPSRA [9] pipeline adopts STAR [19] v2.5.3a for alignment, the exceRpt [10] adopts STAR v2.4.2a and Bowtie [20] v2.2.6. In the COMPSRA, host genome (hg38) was first aligned with the filtered reads produced by the preprocessing module, and after that, the annotation module quantified and annotated the mapped reads. But in exceRpt, exogenous alignment came after endogenous alignment. RNAseq reads were mapped to the transcriptome and host genome in endogenous alignment. After that, nonaligned reads were annotated in relation to the various databases. The exogenous alignment was furthered by aligning unmapped reads to exogenous miRNAs and rRNAs, and after that unmapped reads were mapped to all genomes in Ensembl and NCBI.

Annotation Module

The mapped reads were annotated by COMPSRA [9] to various small RNA reference databases. In exceRpt [10], the annotation and alignment modules were completed simultaneously.

3. RESULTS

Twelve buffalo milk somatic cells' small RNAseq .fq files were processed through COMPSRA [9] to assess how well it performed with regard to various types of RNAs, and they were then distinguished using the exceRpt [10] pipeline. The same server with 64 GB RAM was used to run both pipelines. When compared to COMPSRA, exceRpt was found to have a higher time complexity. Both pipelines produced output that was different from one another.

The various RNAs were found using both pipelines. Table 1 shows the distribution of RNAs produced by the two pipelines. Fig. 1 compares the various RNA molecules that were extracted from both pipelines.

Various RNA molecule types, including miRNA, piRNA, tRNA, snRNA, snoRNA, rRNA, circRNA, and lncRNA, were mined from both pipelines. miRNAs can range in length from 21 to 23 nucleotides, they play a role in translational repression and gene regulation. Numerous miRNAs play a crucial role in the development of cancer and other diseases. While only 112 miRNAs were found by COMPSRA, exceRpt found 236. Seven miRNA were shared by the two, while 222 and 70 miRNA were exclusive to COMPSRA and exceRpt, respectively. The piRNA plays a significant role in regulating the gene expression both during and after

International Journal of Research in Science & Engineering
ISSN: 2394-8299
Vol: 03, No. 02, Feb-Mar 2023
http://journal.hmjournals.com/index.php/IJRISE
DOI: https://doi.org/10.55529/ijrise.32.26.32

transcriptional processes. piRNA has a length that varies from 26 to 32 nucleotides. 475 piRNAs could be found using COMPSRA, which is more than could be found using exceRpt (305). There were 380 piRNA that were specific to COMPSRA, 211 that were specific to exceRpt, and 94 that were shared by the two. tRNA can range in length from 76 to 90 nt, they are important in the translation process. COMPSRA found 13 tRNA, which is 10 fewer than exceRpt. 11 tRNAs were shared by both pipelines, but only 2 by COMPSRA and 12 by exceRpt.

The snRNAs, which have a length of 100 to 300 nucleotides, are crucial for mRNA splicing. Both pipelines produced a nearly identical number of snRNAs. 35 snRNA were specific to COMPSRA, 38 to exceRpt, and 34 to both, while 35 were common to both. The snoRNAs, which range in length from 60 to 400 nucleotides, control rRNA splicing. COMPSRA produced only 19 snoRNA, whereas exceRpt found 52. 8 snoRNAs were shared by both pipelines, whereas COMPSRA had 11 and exceRpt had 43 unique snoRNAs. Circular RNAs (circRNAs) have a length of 100 to 1000 nucleotides and significantly influence the regulation of gene expression and biological development. Compared to exceRpt, COMPSRA produced more circular RNAs. Longer RNAs include lncRNA and rRNA. rRNAs are 120–4500 nucleotides long and regulate rRNA splicing, whereas lncRNAs are longer than 200 nucleotides and play significant roles in epigenetic regulation. exceRpt detected 12 rRNAs and 1988 lncRNAs in all 12 buffalo samples after identifying transcripts with a length greater than 200 and mapping them to appropriate databases. However, COMPSRA did not support any lncRNA or rRNA databases.

A read that has already been aligned to one type will not be annotated to another type, because exceRpt annotated the RNA types in priority order (circRNA< snRNA < snoRNA <piRNA <tRNA <miRNA). Contrarily, COMPSRA annotated a mapped read to every type of RNA without regard to priority. Because COMPSRA reduced the time and effort required by online tools for offsite data transfer, it was found to be superior to exceRpt. But exceRpt would be necessary for the extraction of lncRNA and rRNA information.

4. DISCUSSION

In this study, the same RNAseq dataset comprised of samples of milk somatic cells from buffalo was used to compare the efficacy of the COMPSRA [9] and exceRpt [10] pipelines in quantifying various RNA molecules. The pipeline's computational techniques played a significant role in how much information was available. Significant usability differences were seen, especially in the length of time it took to analyse the samples and the quantity of RNAs produced. A similar investigation using a data set of 12 healthy human serum control samples from RNAseq revealed COMPSRA to be a superior pipeline [9]. However, in this study the number of RNAs generated as well as the time complexity of the pipelines were compared, which also assisted in determining which pipeline was most pertinent for particular requirements.

The pipelines COMPSRA [9] and exceRpt [10] are both appropriate for RNA profiling. To map the reads to the reference, the COMPSRA adopts STAR, and the exceRpt employs STAR and Bowtie 2. In comparison to Bowtie 2, STAR Aligner has a higher time complexity. Endogenous alignment with the host genome and different databases was finished first in

International Journal of Research in Science & Engineering
ISSN: 2394-8299
Vol: 03, No. 02, Feb-Mar 2023
http://journal.hmjournals.com/index.php/IJRISE
DOI: https://doi.org/10.55529/ijrise.32.26.32

exceRpt. Unmapped reads were then once more aligned with exogenous miRNAs and rRNAs. Finally, all of the genomes in Ensembl and NCBI were mapped by the non-aligned reads. Exogenous alignment refers to the final two alignment steps. For exogenous alignment, STAR aligner is utilized. As a result, more snRNA, snoRNA, and miRNA molecules were detected. Additionally, it took about the same amount of time to upload the .fq files to the web-based pipeline (exceRpt) as it did to conduct the analysis. However, COMPSRA was able to save this time since it uses a command line pipeline and uploads data constantly. As a result, the time required by exceRpt was three times greater than that of COMPSRA. However, because COMPSRA does not support their respective databases, it was unable to find lncRNA and rRNA. Additionally, COMPSRA uses three databases to detect piRNAs, whereas exceRpt only uses piRNABank. These three databases are piRNABank, piRBase, and piRNA cluster. As a result, when compared to exceRpt, the COMPSRA found more piRNAs.

5. CONCLUSION

According to our research, there is a trade-off between the output generated and the time complexity of COMPSRA [9] and exceRpt [10]. If we are not interested in lncRNAs and rRNAs, it is best to use COMPSRA because it will produce tRNAs, miRNAs, piRNAs, snRNAs, snoRNAs, and circRNAs with less time complexity. However, exceRpt is the best method for extracting information on lncRNA and rRNA despite its high computational complexity. Therefore, it can be concluded that exceRpt is superior to COMPSRA in terms of output generated if more information needs to be mined from the samples.

6. REFERENCES

1. Sun, Z. et al. (2014).CAP-miRSeq: a comprehensive analysis pipeline for microRNA sequencing data, BMC Genomics 15, 423, https://doi.org/10.1186/1471-2164-15-423
2. Rahman, R. U. et al. (2018). Oasis 2: improved online analysis of small RNA-seq data , BMC Bioinforma. 19, 54, https://doi.org/10.1186/s12859-018-2047-z
3. Fehlmann, T. et al. (2017) . Web-based NGS data analysis using miRMaster: a large-scale meta-analysis of human miRNAs , Nucleic Acids Res. 45, pp. 8731–8744
4. Wu, X. et al. (2017). sRNAnalyzer-a flexible and customizable small RNA sequencing data analysis pipeline , Nucleic Acids Res. 45, pp.12140–12151, https://doi.org/10.1093/nar/gkx999
5. Rueda , A. et al.(2015). sRNAtoolbox: an integrated collection of small RNA research tools , Nucleic Acids Res. 43, pp. 467–473, https://doi.org/10.1093/nar/gkv555
6. Zhang , Y. et al. (2012) .CPSS : a computational platform for the analysis of small RNA deep sequencing data , Bioinformatics ,28(14) , doi: 10.1093/bioinformatics/bts282
7. Chen , C.J. et al. (2012) .ncPRO-seq: a tool for annotation and profiling of ncRNAs in sRNA-seq data , Bioinformatics, Oxford University Press (OUP) 28 (23), pp.3147-3149
8. Fasold , et al. (2011) .DARIO: A ncRNA detection and analysis tool for next-generation sequencing experiments , Nucleic acids research, 39 , W112-7. 10.1093/nar/gkr357
9. Li , J. et al. (2020).COMPSRA: a COMprehensive Platform for Small RNA-Seq data Analysis , Sci Rep 10, 4552 , https://doi.org/10.1038/s41598-020-61495-0

International Journal of Research in Science & Engineering
ISSN: 2394-8299
Vol: 03, No. 02, Feb-Mar 2023
http://journal.hmjournals.com/index.php/IJRISE
DOI: https://doi.org/10.55529/ijrise.32.26.32

10. ROZOWSKY , Joel , et al. (2019). exceRpt: a comprehensive analytic platform for extracellular RNA profiling , Cell systems , 8.4: 352-357. e3
11. Chan , P. P. & Lowe, T. M. (2016)..GtRNAdb 2.0: an expanded database of transfer RNA genes identified in complete and draft genomes , Nucleic Acids Res. 44 , pp. 184–189, https://doi.org/10.1093/nar/gkv1309
12. Harrow J, et al. (2012). GENCODE: the reference human genome annotation for The ENCODE Project, Genome Res. 22(9):1760-74 , doi: 10.1101/gr.135350.111. PMID: 22955987; PMCID: PMC3431492
13. Rosenkranz, D. (2016).piRNA cluster database: a web resource for piRNA producing loci, Nucleic Acids Res. 44, pp. 223–230, https://doi.org/10.1093/nar/gkv1265
14. Sai Lakshmi, S. & Agrawal, S. (2008) .piRNABank: a web resource on classified and clustered Piwi-interacting RNAs , Nucleic Acids Res. 36,pp. 73–177, https://doi.org/10.1093/nar/gkm696
15. Zhang , P. et al.(2014).piRBase: a web resource assisting piRNA functional study, Database 2014, bau110, https://doi.org/10.1093/database/bau110
16. Kozomara , A. et al.(2011) .miRBase: integrating microRNA annotation and deep-sequencing data , Nucleic Acids Res. **39**, pp.152–157, https://doi.org/10.1093/nar/gkq1027
17. Glazar , P. , Papavasileiou, P. & Rajewsky, N. (2014).circBase: a database for circular RNAs, RNA 20, pp. 1666–1670, https://doi.org/10.1261/rna.043687.113
18. Szymanski , M. et al (2002) . 5S Ribosomal RNA Database, Nucleic Acids Res., 176-80 , doi: 10.1093/nar/30.1.176. PMID: 11752286; PMCID: PMC99124
19. Dobin , A. et al. (2013).STAR: ultrafast universal RNA-seq aligner , Bioinforma. 29, pp. 15–21, https://doi.org/10.1093/bioinformatics/bts635
20. Langmead B. (2010) . Aligning short sequencing reads with Bowtie, Curr Protoc Bioinformatics, doi:10.1002/0471250953.bi1107s32

Figure Legend:

Table 1: Distribution of extracted RNAs in 12 samples of buffalo milk somatic cells using COMPSRA and exceRpt pipelines

Type of RNA	COMPSRA	ExceRpt
miRNA	77	229
piRNA	474	305
lncRNA	NA	1988
rRNA	NA	12
tRNA	13	23
snRNA	69	72
snoRNA	19	51
circRNA	99895	114

International Journal of Research in Science & Engineering
ISSN: 2394-8299
Vol: 03, No. 02, Feb-Mar 2023
http://journal.hmjournals.com/index.php/IJRISE
DOI: https://doi.org/10.55529/ijrise.32.26.32

Fig 1. Types and counts of RNAs mined from COMPSRA and exceRpt pipelines.