Journal of Information Technology and Computing
https://jitc.sabapub.com
ISSN: 2709-5916

2022 Volume3, Issue 1 : 32 – 36
DOI : 10.48185/jitc.v3i1.509

Applying data mining steps to explore different small RNAs from buffalo milk transcriptome

Pooja Chhabra[1,*], Brij Mohan Goel[1]

[1]Department of Computer Science and Applications, Baba Mastnath University, Rohtak, India

Received: 23.05.2022 • Accepted: 25.06.2022 • Published: 30.06.2022 • Final Version: 30.06.2022

Abstract: This study is a first attempt to find different types of RNA in lactating buffalo's milk somatic cells. The molecular factors that regulate lactation need to be identified and understood in order to help milk production. By using data mining techniques, patterns and information hidden within a dataset can be identified. In order to detect the RNA, data of 12 samples of buffalo milk somatic cells were analyzed. For extraction of diverse RNAs COMPSRA (COMprehensive Platform for Small RNA Analysis) pipeline was used. We were able to identify several miRNAs, piRNAs, snRNAs, snoRNAs, circRNAs and tRNAs in buffalo milk somatic cells. circRNAs ranked highest among all the samples in our dataset, followed by piRNAs and then miRNAs. Understanding the RNA regulators of lactation will improve and facilitate management of buffalo milk production. Furthermore, our study contributes towards a complete annotation of the buffalo genome..

Keywords: RNAseq, Data mining, RNA, Next Generation Sequencing, RNA databases

1. Introduction

To understand the genome, it is important to understand its functional elements. It is the transcriptome, a set of various types of RNA like miRNA [1], piRNA [2], tRNA [3], snRNA [4], snoRNA [5], circRNA [6] and some non-coding sequences, that determine the function of the genome. Additionally, non coding RNAs that are not translated to proteins play critical roles. For example, snoRNAs regulate rRNA splicing, snRNAs participate in mRNA splicing, tRNA plays an important role in translation, miRNA is involved in translational repression, piRNAs control gene expression during transcription and post-transcriptional processes, while circRNA performs a significant role in gene regulation expression and biological development. To decipher the information about all these RNAs, data mining techniques play an important role. Thus, the present study aimed to discover different types of RNAs from milk somatic cells of lactating buffaloes using RNA sequencing data.

In order to extract different RNAs COMPSRA [7] pipeline was used in our study. The platform provides a comprehensive method for detecting and evaluating small RNAs from RNAseq data. It allows the analysis of small RNA sequences by integrating sequence processing tools and prebuilt RNA databases. It is a free software that allows noncommercial users to access various tailored RNA databases and RNA sequence processing tools.

* Corresponding Author: poojachhabra31@gmail.com

2. Materials and Methods

The steps performed for mining diverse RNAs (Fig. 1) are:-

Figure 1. Steps Performed for Data Mining

2.1. Data Selection

This work adopted a real data set that is available at the GenBank under BioProject PRJNA453843 with accession number GGRC00000000.1. A total of 12 fastq files containing the sequences of buffalo milk somatic cells were examined for the study.

2.2. Data cleaning

First, the quality of all the samples was checked using FastQC tool [8]. The results of FastQC gave the information like number of reads, GC content, quality of samples, presence or absence of adapters, overrepresented sequences, average length of the sequences, sequence duplication level etc. Based on this information the data was cleaned. First the adpater sequences were trimmed. The sequences of low quality (average phred score<=30) were removed. The minimum length of the reads were taken as 20bp for further analysis.

2.3. Data Reduction

By using the filters as discussed above the data was reduced by using the data mining tool Cutadapt [9]. The output of the Cutadapt again was used for quality check with FastQC tool to ensure accuracy of the samples for further processing.

2.4. Data transformation

It is a process of changing the structure, format or values of the data. For transformation, an alignment of qualified reads with the reference genome was performed using STAR (Spliced Transcripts Alignment to a Reference) aligner [10].

2.5. Data Mining and Pattern Evaluation

Different databases like miRBase [11] for the extraction of miRNAs; GENCODE [12] for mining snoRNA and snRNA;GtRNAdb [13] for mining tRNA;piRNAcluster [14], piRNABank [15] and piRBase [16] to extract piRNAs;circBase [17] for mining circRNAs were used. The mapped reads were annotated against these databases for pattern evaluation which resulted the number of miRNAs, piRNAs, tRNAs, snRNAs, snoRNAs and circRNAs as mined data.

2.6. Data Visualization

The distribution of mined RNAs is visualized using graph (Fig. 2).

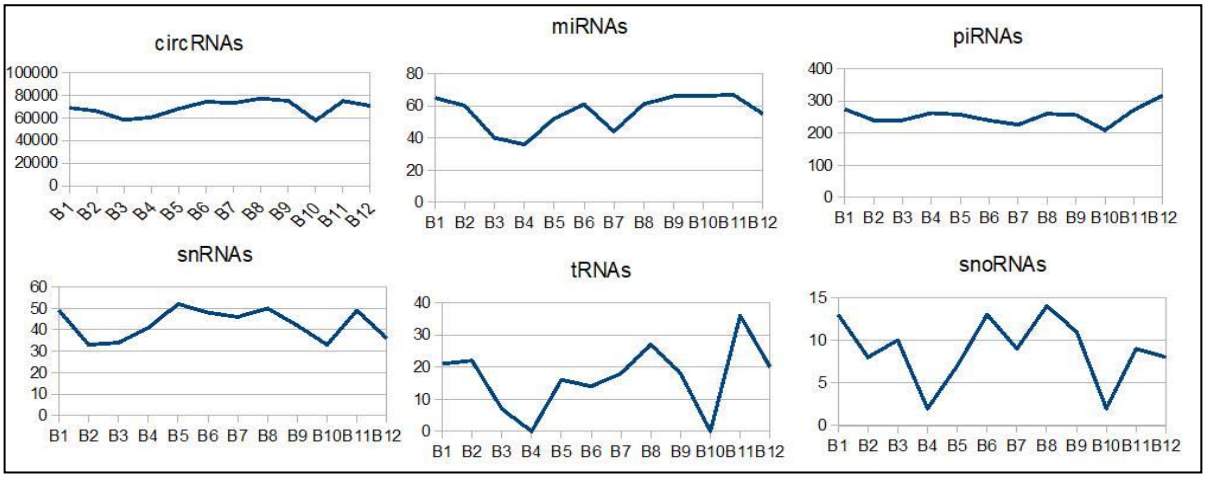

Figure 2. Distribution of mined RNAs across all the samples of buffalo milk somatic cells

3. Results and Discussions

The raw reads for all samples had a read length ranging from 1nt to 101nt. In this study major focus was on mining of miRNA, piRNA, tRNA, snRNA, snoRNA and circRNA. Since all of these RNAs are of length greater than 20nt, all reads having a length lesser than 20nt were removed. Each sample had an average of more than 47 million raw reads. More than 45 million reads were obtained for each sample after trimming and filtering. Approximately 32 million reads were trimmed for reasons such as poor quality, adapter content or short length.

Human reference genome was chosen for alignment of the reads as it is better annotated than buffalo genome. The mapping percentage of the samples with respect to the reference genome is given in Fig. 3. It was observed that all the samples aligned to the reference genome with a high mapping percent ranging from 85.87 to 99.70 (Fig. 3). The mapped reads were further evaluated against the different databases like miRBase, piRNABank, piRBase, piRNAcluster, GtRNAdb, GENCODE and circBase for mining different types of RNAs . The average number of circRNA, miRNA, piRNA, snoRNA, snRNA and tRNA discovered was 68642, 56, 254, 8, 42 and 16 respectively. Maximum number of circRNAs were identified in our dataset followed by piRNAs and then miRNAs.

Name of Sample	Mapping % with Reference Genome
B1	99.06%
B2	99.49%
B3	98.33%
B4	85.87%
B5	98.50%
B6	98.40%
B7	98.60%
B8	97.70%
B9	99.70%
B10	98.06%
B11	99.42%
B12	96.75%

Figure 3. Alignment of processed reads of milk somatic cells with Human reference genome.

It is necessary to identify and understand the molecular factors that regulate lactation, since lactation is a complex process regulated by multiple molecular RNA drivers. The understanding of RNA regulatory mechanisms in lactation will aid in improving and managing milk production in buffaloes. Several studies have also identified different types of RNA in plants as well as animal species. Among samples of human serum [7], circRNA was identified with the highest number, while 1182 miRNA [18] was detected in samples of human milk. There have been 249 miRNAs observed in bovine milk somatic cells [19] and approximately 400 miRNAs found in cattle milk infected with mastitis [20]. In this study, different RNAs like miRNA, piRNA, tRNA, circRNA, snRNA, snoRNA were discovered from RNAseq data of lactating buffalo's milk somatic cells using COMPSRA, a data mining pipeline. The highest count of RNA detected in the samples was circRNA followed by piRNA.

4. Conclusion

In addition to observing the distribution of RNAs in the human genome, our results reflect their availability or annotation. By discovering more RNA molecules of different types in other livestock species, the databases will be further enriched. Studying distinct RNA molecules like miRNA, piRNA, snRNA, snoRNA, circRNA and tRNA in buffalo milk somatic cells is a major step towards annotation of the buffalo genome.

References

[1] Bartel, D. P. (2004) . MicroRNAs: Genomics, biogenesis,mechanism, and function. Cell. 116(2): 281-297.
[2] Siomi, M. C. et al. (2011). PIWI-interacting small RNAs:The vanguard of genome defence. Nature Reviews Molecular Cell Biology. 12(4) : 246-258.
[3] O'Donoghue, P., Ling, J., & Soll, D. (2018). Transfer RNA function and evolution. *RNA biology. 15*(4-5): 423–426.

[4] Adachi, H. & Yu, Y. (2014). Insight into the mechanisms and functions of spliceosomal nRNA pseudouridylation. World J. Biol. Chem.5: 398–408.
[5] Maxwell, E. S. & Fournier, M. J. (1995). The small nucleolar RNAs. Annu Rev Biochem. 64: 897–934.
[6] Lu, M. (2020). Circular RNA: functions, applications and prospects. *ExRNA* **2:**1.
[7] Li, J. et al. (2020). COMPSRA: a COMprehensive Platform for Small RNA-Seq data Analysis. Sci Rep. 10: 4552.
[8] Andrews, S. (2010). FastQC: A Quality Control Tool for High Throughput Sequence Data [Online].web resource for piRNA producing loci". Nucleic acids research. 44(D1):223–230.
[9] Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal. 17(1): 10-12.
[10] Dobin, A. & Gingeras, T.R. (2015). Mapping RNA-seq Reads with STAR. Curr Protoc Bioinformatics. 51(1): 11-14.
[11] Kozomara, A., Birgaoanu, M. & Griffiths-Jones, S. (2019). miRBase: from microRNA sequences to function. Nucleic Acids Research.47:155-162.
[12] Frankish, A. et al. (2019). GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Research.47: 766–773.
[13] Chan, P.P & Lowe, T.M. (2009). GtRNAdb: a database of transfer RNA genes detected in genomic sequence. Nucleic acids research. 37: 93–97.
[14] Rosenkranz, D. (2016). piRNA cluster database: a web resource for piRNA producing loci. Nucleic acids research. 44: 223–230.
[15] Lakshmi, S.S. & Agrawal, S. (2018). piRNABank: a web resource on classified and clustered Piwi-interacting RNAs. Nucleic Acids Res. . 36:173-177.
[16] Wang, J. et al. (2019) . piRBase: a comprehensive database of piRNA sequences. Nucleic Acids Research. 47:175-180.
[17] Glazar, P., Papavasileiou, P. & Rajewsky, N. (2014). circBase: a database for circular RNAs. RNA. 20(11):1666–1670.
[18] Rubio, M. et al. (2018). Circulating miRNAs, isomiRs and small RNA clusters in human plasma and breast milk. PloS ONE. 13(3).
[19] Li, R. et al. (2016),.Comparative Analysis of the miRNome of Bovine Milk Fat,Whey and Cells. PLoS ONE. 11(4).
[20] Lai, Y. C. et al. (2020). Bovine milk transcriptome analysis reveals microRNAs and RNU2 involved in mastitis. FEBS J. 287(9): 1899-1918.