Journal of Entomology and Zoology Studies

Available online at www.entomoljournal.com

E-ISSN: 2320-7078
P-ISSN: 2349-6800
JEZS 2019; 7(3): 174-177

Received: 01-03-2019
Accepted: 05-04-2019

Aruna Pandey
Department of Computational Biology and Bioinformatics, SHUATS, Allahabad, Uttar Pradesh, India

Shikha Saxena
Division of Veterinary Biotechnology, ICAR-Indian Veterinary Research Institute (IVRI), Izatnagar, Bareilly, Uttar Pradesh, India

Raja Ishaq Nabi Khan
Division of Veterinary Biotechnology, ICAR-Indian Veterinary Research Institute (IVRI), Izatnagar, Bareilly, Uttar Pradesh, India

Pramod W Ramteke
Department of Biological Sciences, SHUATS, Allahabad, Uttar Pradesh, India

Correspondence
Pramod W Ramteke
Department of Biological Sciences, SHUATS, Allahabad, Uttar Pradesh, India

Computational pipeline for the annotation of long non-coding RNAs in a newly assembled genome

Aruna Pandey, Shikha Saxena, Raja Ishaq Nabi Khan and Pramod W Ramteke

Abstract
Long non-coding RNAs (lncRNAs) has been studied in mammalian genome. However, it remains a challenge to uncover lncRNAs function in a newly assembled species. In the present study, we have predicted a computational pipeline for analysing lncRNAs extracted from public repository. We have applied our pipeline to a goat lncRNAs transcripts dataset, and discussed the possible functions of the lncRNAs transcripts. The lncRNAs transcripts were extracted from Ensembl database and filtered with length parameter (length >=200 nt) and coding potentiality, using a CPC2 tool. The filtered lncRNAs transcripts were further evaluated against MEME and TomTom tool for motif and transcription factors (TF's) discovery. In addition, the lncRNAs transcripts were classified and annotated further using Infernal and blast tool, respectively. Using our stringent pipeline a total of 1304 lncRNAs transcripts were predicted. 3 motifs and 39 transcription factors were found to be regulating these lncRNAs transcripts. Moreover the lncRNAs studied here, belongs to different ncRNA families and also act as precursors for various microRNAs (miRNAs) such as chi-mir-23b, chi-mir-29a etc. In conclusion, our method can annotate lncRNAs transcripts in goat genome and the predicted pipeline can also be applied to other newly assembled genomes for understanding the function of lncRNAs.

Keywords: lncRNAs, miRNAs, goat, RNA-Seq, motifs, TF's, genome, infernal

1. Introduction
The structure of eukaryotic genome is very complex, a large portion of genome is considered to be non-coding. Long non-coding RNAs are RNA transcripts, longer than 200 nt, and lack protein-coding capacity [1]. The involvement of lncRNAs in various biological processes [2, 3] and animal diseases [4] has led to intense interest in these transcripts. Moreover, the huge expansion in the non-coding genome and transcriptome in complex organisms, makes non-coding RNA a natural candidate for the informational molecule underlying the increase in organismal complexity [5]. In the recent years, advances in transcriptome and deep sequencing technology led to the discovery of large amounts of non-protein coding transcription, particularly of long noncoding RNAs (lncRNAs) in mammalian genome. For example using next generation sequencing lncRNAs were identified in human [6], mouse [7], goat [8], cattle, pig [9] and sheep [10] species.

Although thousands of lncRNAs have been identified in mammalian species but only a few of them were functionally characterized. Having a difficulty in experimentally characterize the biological function of the lncRNAs and the enormous flow of genomics data becoming available relevant to lncRNAs' biological functions, it is interesting to predict lncRNAs' functions computationally.

In the present study, we have suggested a computational pipeline to elucidate the complex role of lncRNAs transcripts, yielded from newly assembled goat genome. The current research will provide a deeper functional annotation of the goat genome and can also be applied to other newly assembled genomes for understanding the function of lncRNAs.

2. Material and Methods
2.1 Long non-coding RNAs data retrieval
lncRNAs sequences and their genomic coordinates were downloaded from the Ensembl database (release 95) [11]. To obtain long non-coding RNAs, only sequences classified as, lincRNA, were kept. Further, the transcripts were filtered according to transcript length ≥ 200. Finally, filtered transcripts were evaluated using coding potential calculator 2 (CPC2) [12].

2.2 Transcription factors identification and binding affinity with lncRNAs

To identify enriched transcription factor binding sites in the promoters of lncRNAs, the promoter sequences were extracted from Ensembl biomart database [13] and further examined for motif identification using the MEME suite (V5.0.5) [14]. The resulting motifs were searched for transcription factor identification against the JASPAR database [15] using TomTom [16]. To analyse the lncRNAs binding affinity with transcription factors, the 3'UTR target site of TF's were extracted from Ensembl biomart database and analyzed in miRanda [17] tool to evaluate the strength of interaction using the parameters ΔG and total score value.

2.3 lncRNAs classification and functional prediction

Filtered lncRNAs were classified into different non-coding RNAs (ncRNAs) families using INFERNAL (V1.0), which categorizes ncRNAs and their conserved primary sequence and RNA secondary structure through the use of multiple sequence alignments (MSAs), consensus secondary structure annotation and covariance models (CMs) [18]. For Potential involvement of lncRNAs as a precursors of miRNAs, the lncRNAs were aligned with the precursors of known miRNAs in the mirbase database [19] using BLAST with default parameters [20]. In house Perl script was written to extract only goat precursors sequences. The lncRNAs homologous to miRNAs with >=90% coverage were eventually defined as miRNA precursors. The present study was conducted at CBG lab, IVRI, Izatnagar, Bareilly, in the session 2018-2019.

3. Results

3.1 Long non-coding RNAs data analysis

The schematic representation of lncRNAs annotation is shown in Figure 1. A total of 4675 lncRNAs transcripts were extracted from Ensembl database. Among them, 4558 transcripts were filtered with length >=200. These filtered lncRNAs transcript were subjected to CPC2 tool to reconfirm their coding potentiality. Of these 4558 filtered transcripts, 1304 transcripts were identified as non-coding with coding potentiality score >=0.1.

3.2 TF's identification and interaction with long non-coding RNAs and miRNAs

In this study we have identified TF's among the 1304 lncRNAs that contribute to the transcriptional control of the lncRNAs genes. A total of 39 TF's were predicted from 3 overrepresented motifs. Of these 39 TFs, 16 unique TFs (KLF4, TCF7L2, ZNF263, E2F6, SP1, STAT5A, E2F4, COG1, IRF3, KLF9, STAT3, ID1, GATA4, RELB, RREB1 and SOX10) were found to have a high binding affinity (ΔG value >= −291 kCal/Mol and total score >= 300) with two lncRNAs- ENSCHIG00000000537 and ENSCHIG00000000546. The top ten TF's interaction with lncRNAs are shown in Table 1

It has also been known that miRNAs and TF's are often highly interacted in a dependent or independent manner [21]. A total of 15 TF's were found interacting with 23 miRNAs with ΔG value >= −22 kCal/Mol and total score >= 152. Interestingly, these miRNAs regulating TFs were found to be common among lncRNAs regulating TF's.

3.3 Classification of lncRNAs into non-coding RNA families and analysis

To better annotate lncRNAs from an evolutionary point of view, we classified the predicted lncRNAs into different ncRNA families using INFERNAL tool. Based on a consensus secondary structure annotation using a covariance model, we identified 13 unique non-coding sequences belonging to 3 conserved lncRNA families: - rRNA. tRNA and miR (Table 2). Among the conserved lncRNA families, one familiy (tRNA) accounted for more than 10 members. In total, 58 lncRNAs were detected as precursors for the 55 unique miRNAs. In addition three miRNA precursor- chi-mir-1271, chi-mir-23b and chi-mir-29a with 100 % coverage could be aligned with more than one lncRNAs:- ENSCHIG00000007630, ENSCHIG00000003605, ENSCHIG00000004052, ENSCHIG00000003530, ENSCHIG00000004553, ENSCHIG00000005287, respectively.

Table 1: lncRNAs binding with transcription factors

lncRNA-ID	TF's	total score	ΔG value
ENSCHIG00000000537	TCF7L2	331	-318.71
ENSCHIG00000000537	SP1	301	-344.67
ENSCHIG00000000537	ZNF263	300	-348.97
ENSCHIG00000000537	STAT5A	335	-366.67
ENSCHIG00000000537	KLF9	310	-379.39
ENSCHIG00000000546	STAT3	340	-564.8
ENSCHIG00000000546	COG1	362	-577.77
ENSCHIG00000000546	E2F4	303	-580.45
ENSCHIG00000000546	IRF3	347	-586.66
ENSCHIG00000000546	SOX10	301	-609.07

Table 2: Classification of lncRNAs in to different non-coding RNAs family

lncRNAs-ID	Family_Name	Family_Accession
ENSCHIG00000000806	tRNA	RF00005
ENSCHIG00000001136	tRNA	RF00005
ENSCHIG00000001303	tRNA	RF00005
ENSCHIG00000001454	tRNA	RF00005
ENSCHIG00000001542	tRNA	RF00005
ENSCHIG00000002007	tRNA	RF00005
ENSCHIG00000002048	tRNA	RF00005
ENSCHIG00000003530	tRNA	RF00005
ENSCHIG00000006842	tRNA	RF00005
ENSCHIG00000007078	tRNA	RF00005
ENSCHIG00000007766	tRNA	RF00005
ENSCHIG00000003816	LSU_rRNA_eukarya	RF02543
ENSCHIG00000006842	miR-563	RF00005

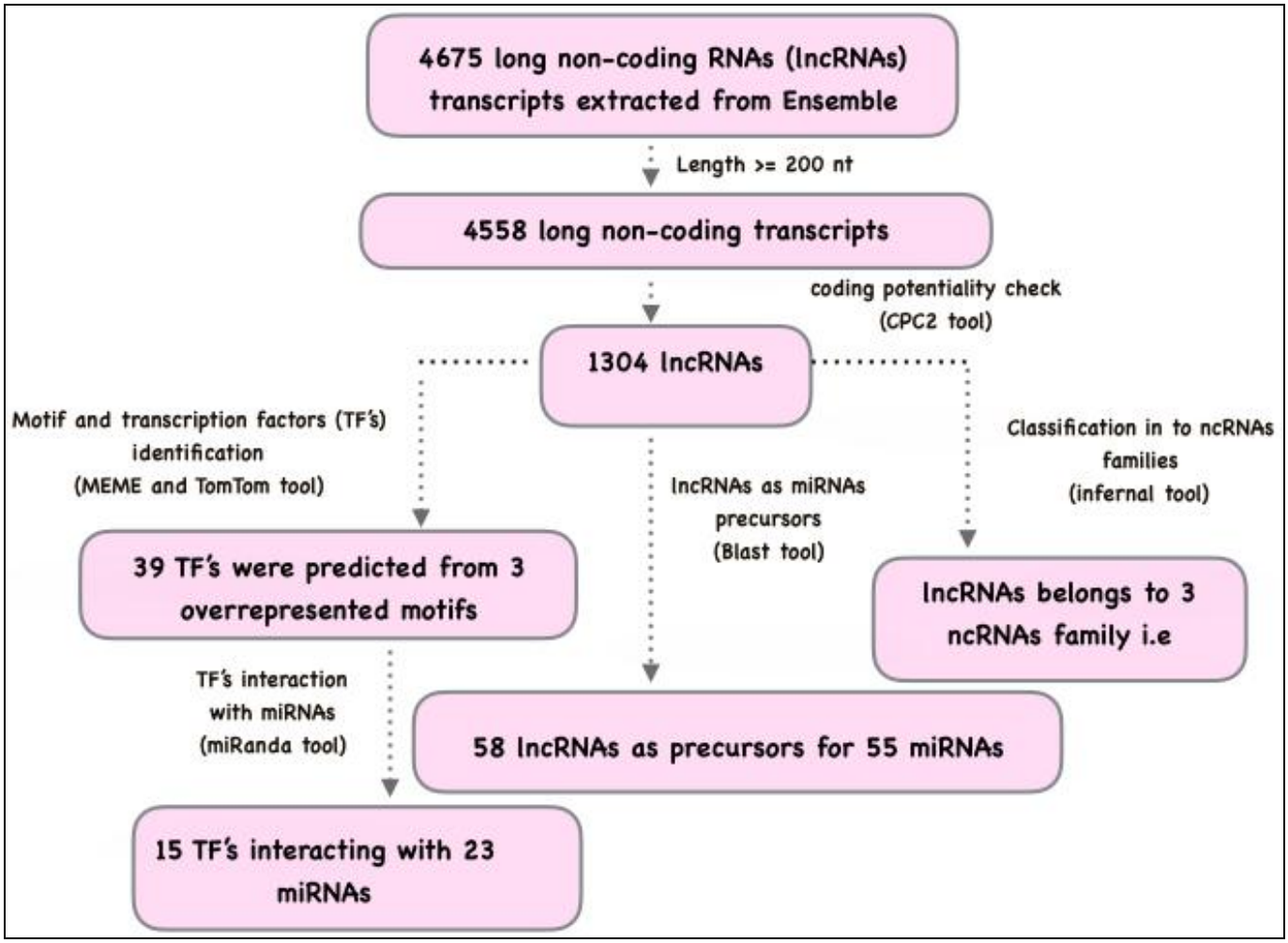

Fig 1: Annotation pipeline of long non-coding RNAs

4. Discussion

In the past years bioinformatics analysis has became the mainstream method for the analysis of lncRNAs. Herein we describe, a comprehensive lncRNAs analysis pipeline, which takes advantages of the annotation of lncRNAs retrieved from public repositories. However, there are many other methods besides the ones used in this study to improve the annotation accuracy of lncRNAs. For instance, the RNA-Seq method such as PLAR (http://webhome.weizmann.ac.il/home/igoru/PLAR/) and UClncR [22] could also be used to obtain a more stringent result in the analysis process.

Long non-coding RNAs are known to execute their functions through interacting with other types of molecules such as TF's [23, 24]. In the present study, 39 TF's were identified and 16 of them were found highly interacting with two lncRNAs- ENSCHIG00000000537 and ENSCHIG00000000546. These highly interacting transcription factor such as IRF3, known to have a role in immune response [25]. IRF3 also found to be regulated by long non-coding RNAs [26]. STAT3, another highly interacting immune regulating transcription factor [27] was found to be governed by lncRNA-UICC [28]. Similarly, SOX10 was regulated by lncRNAs in various studies [29, 30]. On the basis of above findings, the lncRNAs- ENSCHIG00000000537 and ENSCHIG00000000546 might work as immune regulators via TF's IRF3, STAT3 and SOX10. Further, long non-coding RNAs were classified into different ncRNAs family such as tRNA. Consistent with our result, it is possible that goat lncRNAs might serve as precursors for different types of functional tRNAs.

Studies also suggest that long non-coding transcripts in eukaryotes function as precursors for miRNAs [31, 32]. In our study, lncRNAs- ENSCHIG00000004052 and ENSCHIG00000004553 act as precursor for chi-mir-23b and chi-mir-29a, respectively. miR-23b is a central regulator of inflammation during autoimmunity [33], while miR-29a regulate apoptosis [34], thus ENSCHIG00000004052 and ENSCHIG00000004553 may involved in immune response and apoptosis.

5. Conclusion

The above observations might open the possibilities that the functions of lncRNAs can be inferred through their interactions with other molecules –TF's and miRNAs. In addition, the predicted pipeline just relieves the daunting work of initial analysis of lncRNAs, however researchers need to spend more time on further analysis and interpretation of these findings.

6. Acknowledgement

The authors are thank full to ICAR Indian Veterinary Research Institute (IVRI) Izatnagar, Bareilly, for providing necessary laboratory facilities for carry out the present work successfully.

7. References

1. Kim T, Croce CM. Long noncoding RNAs: undeciphered cellular codes encrypting keys of colorectal cancer

pathogenesis. Cancer letters. 2018; 417:89-95.
2. Mercer TR, Dinger ME, Mattick JS. Long non-coding RNAs: insights into functions. Nature reviews genetics. 2009; 10(3):155.
3. Ma H, Hao Y, Dong X, Gong Q, Chen J, Zhang J *et al*. Molecular mechanisms and function prediction of long noncoding RNA. The Scientific World Journal. 2012.
4. Zhao P, Liu S, Zhong Z, Jiang T, Weng R, Xie M *et al*. Analysis of expression profiles of long noncoding RNAs and mRNAs in brains of mice infected by rabies virus by RNA sequencing. Scientific reports. 2018; 8(1):11858.
5. Cao H, Wahlestedt C, Kapranov P. Strategies to annotate and characterize long noncoding RNAs: advantages and pitfalls. Trends in Genetics, 2018.
6. Hao S, Yao L, Huang J, He H, Yang F, Di Y *et al*. Genome-wide analysis identified a number of dysregulated long noncoding RNA (lncRNA) in human pancreatic ductal adenocarcinoma. Technology in cancer research & treatment. 2018; 17:1533034617748429.
7. Bhattarai S, Pontarelli F, Prendergast E, Dharap A. Discovery of novel stroke-responsive lncRNAs in the mouse cortex using genome-wide RNA-seq. Neurobiology of disease. 2017; 108:204-12.
8. Song S, Min Y, Yefang L, Marhaba R, Qianjun Z, Yabin P *et al*. Genome-wide discovery of lincRNAs with spatiotemporal expression patterns in the skin of goat during the cashmere growth cycle. BMC genomics. 2018; 19(1): 495
9. Kern C, Ying W, James C, Ian K, Mary D, Hans C *et al*. Genome-wide identification of tissue-specific long non-coding RNA in three farm animal species. BMC genomics. 2018; 9(1):684.
10. Li Q, Ruizao L, Huijing ZRD, Zengkui L, Enmin L *et al*. Identification and Characterization of Long Noncoding RNAs in Ovine Skeletal Muscle. Animals. 2018; 8(7):127.
11. Hubbard T, Daniel B, Ewan B, Graham C, Yuan C, Clark L *et al*. The Ensembl genome database project. Nucleic acids research. 2002; 30(1):38-41.
12. Kang YJ, Yang DC, Kong L, Hou M, Meng YQ, Wei L *et al*. CPC2: A fast and accurate coding potential calculator based on sequence intrinsic features. Nucleic acids research. 2017; 45(W1):W12-16.
13. Kinsella RJ, Kähäri A, Haider S, Zamora J, Proctor G, Spudich G *et al*. Ensembl BioMarts: a hub for data retrieval across taxonomic space. Database, 2011.
14. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L *et al*. MEME SUITE: tools for motif discovery and searching. Nucleic acids research. 2009; 37(2):W202-8.
15. Sandelin A, Alkema W, Engström P, Wasserman WW, Lenhard B. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. Nucleic acids research. 2004; 32(1):D91-4.
16. Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble WS. Quantifying similarity between motifs. Genome biology. 2007; 8(2):R24.
17. Enright AJ, John B, Gaul U, Tuschl T, Sander C, Marks DS. MicroRNA targets in Drosophila. Genome biology. 2003; 5(1):R1.
18. Nawrocki EP, Kolbe DL, Eddy SR. Infernal 1.0: inference of RNA alignments. Bioinformatics. 2009; 25(10):1335-7.
19. Griffiths-Jones S, Grocock RJ, Van Dongen S, Bateman A, Enright AJ. miRBase: microRNA sequences, targets and gene nomenclature. Nucleic acids research. 2006; 34(1):D140-4.
20. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Journal of molecular biology. 1990; 215(3):403-10.
21. Martinez NJ, Walhout AJ. The interplay between transcription factors and microRNAs in genome-scale regulatory networks. Bioessays. 2009; 31(4):435-45.
22. Sun Z, Nair A, Chen X, Prodduturi N, Wang J, Kocher JP. UClncR: Ultrafast and comprehensive long non-coding RNA detection from RNA-seq. Scientific reports. 2017; 7(1):14196.
23. Herriges MJ, Swarr DT, Morley MP, Rathi KS, Peng T, Stewart KM *et al*. Long noncoding RNAs are spatially correlated with transcription factors and regulate lung development. Genes & development. 2014; 28(12):1363-79.
24. Zhou L, Sun K, Zhao Y, Zhang S, Wang X, Li Y *et al*. Linc-YY1 promotes myogenic differentiation and muscle regeneration through an interaction with the transcription factor YY1. Nature communications. 2015; 6:10026.
25. Yanai H, Chiba S, Hangai S, Kometani K, Inoue A, Kimura Y *et al*. Revisiting the role of IRF3 in inflammation and immunity by conditional and specifically targeted gene ablation in mice. Proceedings of the National Academy of Sciences. 2018; 115(20):5253-8.
26. Valadkhan S, Plasek LM. Long Non-Coding RNA-Mediated Regulation of the Interferon Response: A New Perspective on a Familiar Theme. Pathogens & immunity. 2018; 3(1):126.
27. Hillmer EJ, Zhang H, Li HS, Watowich SS. STAT3 signaling in immunity. Cytokine & growth factor reviews. 2016; 31:1-5.
28. Su K, Zhao Q, Bian A, Wang C, Cai Y, Zhang Y. A novel positive feedback regulation between long noncoding RNA UICC and IL-6/STAT3 signaling promotes cervical cancer progression. American journal of cancer research. 2018; 8(7):1176.
29. Raveendra BL, Swarnkar S, Avchalumov Y, Liu XA, Grinman E, Badal K *et al*. Long noncoding RNA GM12371 acts as a transcriptional regulator of synapse function. Proceedings of the National Academy of Sciences. 2018; 115(43):E10197-205.
30. Coe EA, Tan JY, Shapiro M, Louphrasitthiphol P, Bassett AR, Marques AC *et al*. The MITF-SOX10 regulated long non-coding RNA DIRC3 is a melanoma tumour suppressor. bioRxiv, 2019, 591065.
31. Kapranov P, Cheng J, Dike S, Nix DA, Duttagupta R, Willingham AT *et al*. RNA maps reveal new RNA classes and a possible function for pervasive transcription. Science. 2007; 316(5830):1484-8.
32. Cai W, Li C, Liu S, Zhou C, Yin H, Song J *et al*. Genome Wide Identification of Novel Long Non-coding RNAs and Their Potential Associations With Milk Proteins in Chinese Holstein Cows. Frontiers in genetics, 2018, 9.
33. Hu R, O'connell RM. MiR-23b is a safeguard against autoimmunity. Nature medicine. 2012; 18(7):1009.
34. Mott JL, Kobayashi S, Bronk SF, Gores GJ. mir-29 regulates Mcl-1 protein expression and apoptosis. Oncogene. 2007; 26(42):6133.