Identification and structural features analysis of long noncoding RNAs

Pronozin A.*[1], Afonnikov D.[1]

[1] *Kurchatov Genomic Center of the Institute of Cytology and Genetics of Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia*

Motivation

Long non-coding RNAs (lncRNAs) are defined as transcripts of more than 200 nucleotides length and without any protein coding potential. LncRNAs are involved in important plant development processes such as phosphate homeostasis, flowering, photomorphogenesis and stress response in this connection, their study is relevant. Information is obtained from transcriptomes, but bioinformatic annotation methods are not sufficiently presented, especially for plants. This raises the challenge of developing approaches to automatic annotation and prediction of lncRNA functions in plants.

Aim

- Development of automatic pipeline for identification and classification of lncRNA sequences of agricultural plants based on large-scale analysis of transcriptomes.
- Classification of lncRNA by localization in the genome, assessment of the expression of transcripts encoding lncRNA, assessment of the diversity of lncRNA characteristics in the maize genome.

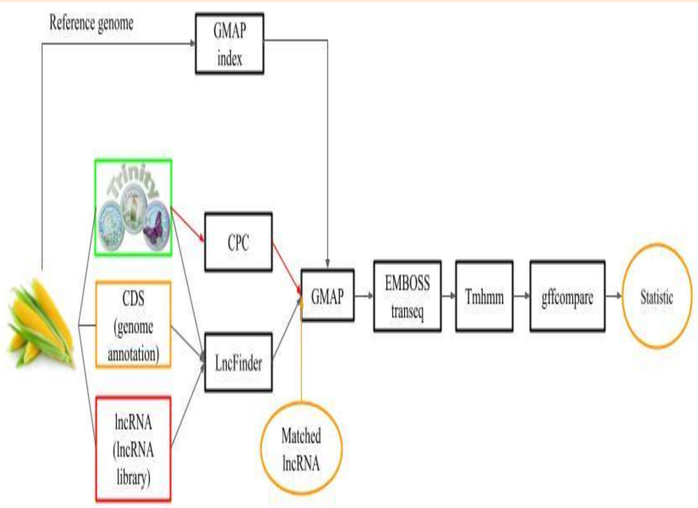
Materials and Methods

The pipeline includes the following steps:

1. lncRNAs prediction by the LncFinder [1] program and CPC [2].
2. Alignment of the predicted lncRNAs to the reference genome, by the GMAP program [3].
3. Transmembrane potential prediction, program TMHMM[4].
4. Classification of lncRNAs of their localization in the genome, gffcompare program [5].
5. Analysis of the structural features of lncRNAs.

The pipeline is implemented using the workflow management system **Snakemake** [6].

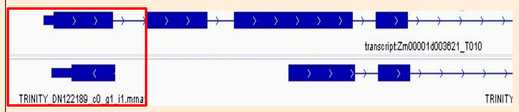
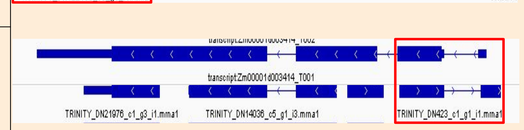
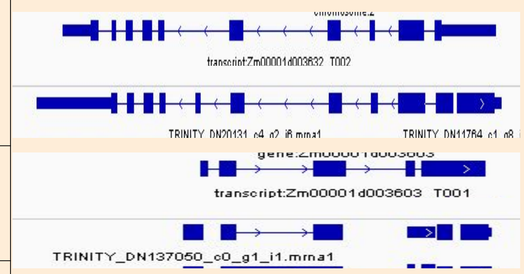
Figure 1: Pipeline



Results

- The model for lncFinder was trained on a sample consisting of known 40 thousand lncRNAs and 30 thousand CDS of corn, this sample was divided randomly into a training (80% of the total sample) and a test (20% of the total sample). As a result, the ***F1*** = 0.96 for both the training and test samples.
- The pipeline was applied to analyze the transcriptomic sequences of *Zea mays (*~800 transcriptome libraries, 3148430 sequences*)*.
- The 2741504 (87%) were identified as lncRNAs by lncFinder program and 1847554 (58%) were identified as lncRNAs by CPC. The 1608236 matched lncRNAs between two programs are analyzed.
- 1578817 (50%) lncRNAs were aligned to the reference genome.
- For 746868 lncRNAs the transmembrane potential was predicted and this transcripts were removed from analysis.

Table1: Distribution of lncRNAs with respect to the known maize genes encoding proteins

Exonic antisense	84375	
Intronic antisense	890	
Multi-exon with at least one junction match	58724	
Retained intron(s) compared to reference	18659	
Intergenic	189287	

Execution time

The execution time test was carried out on 30,000 investigated maize transcripts. The running time is 128 minutes, the predominant execution time is 87 minutes GMAP.

Statistic

The analysis of the structural and functional features of lncRNAs demonstrated:

- majority of lncRNAs have a single exon structure (~58%).
- approximately 70% of multi-exonic lncRNAs have an intron length of 1 to 500 nt.
- approximately 68% of lncRNAs have an exon size of 2 to 300 nt.

A more detailed analysis of the alignment features on the target gene structure was performed for antisense lncRNAs (exon and intron, 85265 in total). It turned out that the predominant number of lncRNAs is aligned within the first exon of the target gene.

Figure 2: The ratio of the number of exons per lncRNA.

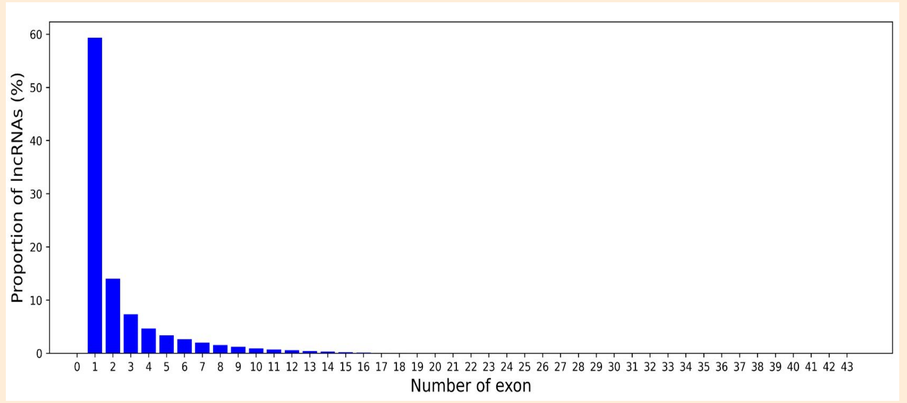
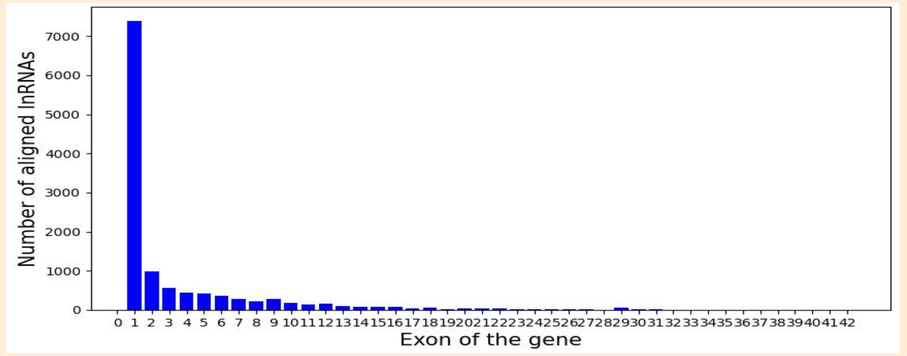


Figure 3: Distribution of lncRNA alignment to target gene structure.



Conclusion

In this paper, we analyzed ~800 transcriptome libraries of maize obtained using the TRINITY program. A pipeline based on the snakemake platform has been developed for the analysis. The pipeline allowed the identification of 351937 new lncRNAs. New lncRNAs are classified into classes depending on their localization in the genome. The features of the alignment of lncRNAs to protein-coding genes were revealed.

Work was funded by the Kurchatov Genome Center of the Federal Research Center IC&G SB RAS, agreement with the Ministry of Education and Science of the Russian Federation № 075-15-2019-1662.

References

1. Kong L. et al. CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine //Nucleic acids research. – 2007. – Т. 35. – №. suppl_2. – С. W345-W349.
2. Han S. et al. (2019) LncFinder: an integrated platform for long non-coding RNA identification utilizing sequence intrinsic composition, structural information and physicochemical property, *Briefings in bioinformatics*, **6**:2009-2027.
3. Wu T. D., Watanabe C. K. (2005) GMAP: a genomic mapping and alignment program for mRNA and EST sequences, *Bioinformatics*, **9**:1859-1875.
4. Krogh A. et al. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes //Journal of molecular biology. – 2001. – Т. 305. – №. 3. – С. 567-580.
5. Pertea G., Pertea M. (2020) GFF utilities: GffRead and GffCompare, *F1000Research*, **9**.
6. Köster J., Rahmann S. (2012) Snakemake—a scalable bioinformatics workflow engine, *Bioinformatics*, **19**:2520-2522.