DOI 10.18699/SBB-2022-024

ICAnnoLncRNA: automatic pipeline of identification, classification and annotation of long noncoding RNAs

Pronozin A.*, Afonnikov D.

Kurchatov Genomic Center of the Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia
** pronozinartem95@gmail.com*

Long non-coding RNAs (lncRNAs) typically defined as transcripts of more than 200 nucleotides in length and without any protein coding potential. These RNAs are involved in important plant development processes such as phosphate homeostasis, flowering, photomorphogenesis and stress response. Information about lncRNA sequences and their expression typically obtained from transcriptome sequencing. To process these data relevant tools for automated lncRNA are required.

Here, we propose a ICAnnoLncRNA, pipeline for automatic identification, classification and annotation of plant lncRNAs The pipeline includes: (1) identification of lncRNAs by lncFinder and CPC2; (2) transmembrane potential prediction by TMHMM; (3) Mapping lncRNA to reference genome by GMAP; lncRNAs classification by its localization in the genome using gffcompare; (5) analysis of the lncRNAs structural features; (6) search for conservative lncRNA with homology to sequences from other organisms. The pipeline implemented using the Snakemake [1] workflow management system language.

We analyzed 45 million Z. mays transcripts. Identify ~31 million lncRNAs by LncFinder and CPC. For ~10 million (26 %) lncRNAs the significant transmembrane potential was predicted. 21 million (47 %) lncRNAs were aligned to the reference genome. We identified antisense – 3.565.080, intron – 2.785.105, intergenic – 2.856.326 lncRNAs. The 3 million lncRNAs in maize have homologs and 297.299 of them have homologs among known lncRNAs from other plants. Tissue-specificity analysis showed that the vast majority of lncRNAs are expressed in ovary tissues.

The proposed automatic pipeline made it possible to identify 9.206.511 new lncRNAs in maize genome, classified into classes depending on their localization in the genome, annotate them and evaluate their structural features.

Acknowledgements: The work was funded by the Kurchatov Genomic Center of the Institute of Cytology and Genetics, SB RAS, agreement with the Ministry of Education and Science of the Russian Federation No. 075-15-2019-1662.

References

1. Köster J., Rahmann S. Snakemake – a scalable bioinformatics workflow engine. *Bioinformatics*. 2012;19:2520-2522.