doi 10.18699/PlantGen-2025-274

A pan-transcriptome scale study of long non-coding RNA diversity in maize

A.Y. Pronozin*, N.A. Shmakov, D.A. Afonnikov
Institute of Cytology and Genetics, Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
** pronozinartem95@gmail.com*

Motivation and aim: Long non-coding RNAs (lncRNAs) are a class of linear or circular RNA molecules over 200 nucleotides long that do not encode proteins. Experimental studies have shown a wide range of functions of lncRNAs in plants. They are involved in the regulation of resistance to cold, salt and heat stress, influence hypoxia tolerance and participate in fruit, root and leaf development. This diversity of functions makes lncRNAs an important target for research in the context of crop resistance and productivity.

Despite progress in the study of the structural and functional characteristics of lncRNAs, their evolutionary origin remains poorly understood. This is due to the many factors that need to be considered when identifying lncRNAs. The size of lncRNAs is similar to the size of some protein-coding genes, which leads to overprediction errors. Also, lncRNA sequences rapidly accumulate mutations, so that even with sequences from closely related species of organisms they have weak homology. These features of lncRNAs demonstrate the necessity of analyzing the transcriptomes of multiple members of the same species, and the concept of pan-genome and pan-transcriptome was developed for this purpose. However, to date, the works devoted to the study of pan-genomes and pan-transcriptomes are mainly focused on the identification and study of new protein-coding genes, and the detection of lncRNAs is a minor study.

The use of pan-transcriptomes will improve the study efficiency, accuracy, and completeness of predicted lncRNAs compared to using the transcriptome of a single species representative. This work focuses on the pan-transcriptome analysis of maize, comparing its coding and non-coding parts. In order to identify patterns of evolution of protein-coding genes and lncRNAs.

Materials and methods: A total of 503 libraries of maize inbred lines were used in this work. A bioinformatics pipeline was developed for pan-transcriptome assembly and analysis, which includes three processing steps: transcriptome assembly, pan-transcriptome assembly, and pan-transcriptome analysis. Pan-transcriptome analysis, includes: gene family clustering and expression analysis, pan-transcriptome structure analysis, SNP search and analysis, and synonymous and non-synonymous substitution analysis.

Results: As a result, two maize pan-transcriptomes were obtained. The pan-transcriptome containing the coding part counts 245436 protein-coding genes. The pan-transcriptome containing the non-coding part counts 595198 lncRNAs. Structure analysis revealed that both pan-transcriptomes are closed. However, a more deep investigation, showed that the pan-transcriptome of lncRNAs is mainly (52%) composed of transcripts represented in less than 10% of the investigated transcriptomes. Whereas the pan-transcriptome of protein coding genes mainly consists of genes represented in all transcriptomes or between 70–90% of the total. Polymorphism analysis confirmed the literature data on lncRNA variability. Analysis of nucleotide diversity (π) showed that lncRNAs have more mutations compared to mRNAs, indicating their rapid evolution. Also, mRNAs exhibit a more stable evolutionary pattern with neutral changes, where mutations and genetic drift are less pronounced. Whereas lncRNAs, show greater variability in values, and are shifted to the region of balanced selection. The analysis of synonymous and nonsynonymous substitutions in mRNAs and the frequency of nucleotide substitutions for lncRNAs also showed greater variability in lncRNAs, which once again confirmed our conclusions.

Conclusions: In this study, two maize pan-transcriptomes containing coding (protein-coding genes) and non-coding (lncRNA) parts were assembled. Although both pan-transcriptomes are closed, however, the lncRNA pan-transcriptome is closer to the open type because it stops expanding upon addition of a new transcriptome only at 500 samples, whereas the protein-coding genes pan-transcriptome stops expanding at 100 samples. Analysis of polymorphisms and synonymous/nonsynonymous substitutions showed that lncRNAs have more mutations compared to mRNAs, indicating their rapid evolution as well as greater variability between different members of the same species.

Funding: The work was carried out with funding from the budget project FWNR-2022-0020. The calculations were performed using the resources of the “Bioinformatics” Center.