DOI 10.18699/PlantGen2019-006

Analysis of out of the reference transcripts from RNA-seq libraries in crops

Afonnikov D.A.[1, 2]*, Genaev M.A.[1], Shmakov N.A.[1], Mustafin Z.S.[1], Mukhin A.M.[1, 2], Konstantinov D.K.[1, 2], Doroshkov A.V.[1, 2], Lashin S.A.[1, 2]
[1] *Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia*
[2] *Novosibirsk State University, Novosibirsk, Russia*
* *e-mail: ada@bionet.nsc.ru*

The analysis of crop gene expression based on RNA-seq experiments is one of the most effective ways to search for genes of biological significance. The results are important for geneticists and breeders in the breeding new lines and cultivars of improved stress response, the search for markers of new useful genes. However, most of the results of gene expression analysis published in articles and databases are based only on reference genomic sequences. For agricultural plants, there are more and more data on transcriptomes of varieties and lines, the genotype of which differs from the genotype of the reference organism. Most of these transcriptomes contain sequences that are not detected in the reference genome and can only be obtained by the *de novo* assembly method. In this work, a large-scale analysis of transcriptomes of 5 crops (maize, rice, tomato, potato and barley) taken from the available SRA archives of NCBI and EBI (over 1200 libraries in total) was carried out. We aimed at identification of "Out Of the Reference Transcripts" (OORT) in RNA-seq libraries and their annotation. For each of the libraries de novo transcript sequences were reconstructed and aligned to reference genome. Sequences of two types were identified: (1) transcripts aligned to the unannotated reference genome loci; (2) transcripts unaligned to reference genome. It is shown that the proportion of transcripts that are aligned to the reference unannotated loci varies from 20 to 25 %. Proportion of transcripts unaligned to the reference genome is up to 5 %. For sequences of "new" transcripts not aligned to the genome, the identification of ORFs and amino acid sequences was carried out and their annotation was performed. Some of such transcript were identified as non-coding RNAs, viral and pathogen sequences. We also identified candidate for plant resistance genes among OORT: 181 for unaligned transcripts and more than 1500 for unannotated. Transcripts that are homologous to genes of plant stress response to drought, oxidative stress, high temperatures and genes of plant resistance to pathogens were also identified.
Acknowledgements: The work was supported by RSF grant 18-14-00293.