DE NOVO SEQUENCING, ASSEMBLY AND CHARACTERIZATION OF GENOME IN TETRAPLOID PLANT CAPSELLA BURSA-PASTORIS

Kasianov A.S.[1*], Logacheva M.D.[2], Makeev V.J.[1], Penin A.A.[2]

1. Vavilov Institute of General Genetics, RAS Moscow, Russia

2. Lomonosov Moscow State University, Moscow, Russia

* Corresponding author

Key words: *plants, polyploidy, genome, sequence assembly, Capsella bursa-pastoris*

Motivation and Aim

Genome sequencing data are a basic part of modern genetics, genomics and evolutionary biology. Evolution of sequencing technologies gives us capabilities for characterization of genomes in many non-model species. However, de novo assembly of genomes of flowering plants is still a very difficult task, because many of them are polyploids. As a result their similar multiple paralogs presents in their genomes and it complicates the assembly.

Subject of our study was *Capsella bursa-pastoris.* It is a tetraploid plant with uncertain origin. *C. bursa-pastoris* may be a recent allotetraploid or more ancient autotetraploid. *It* is a perfect model for the studies of gene and genome evolution after duplication events due to the fact that *C. bursa-pastoris* has a close relationship with a model plant *Arabidopsis thaliana (Capsella* belongs to the same family*).*

Methods and Algorithms

We have sequenced DNA from *Capsella bursa-pastoris* using Illumina sequencing platform. Nearly 300 million paired-end reads were generated, . Coverage of genome was 150x. Insert lengths of paired-end libraries were 150bp and 450bp. We tried to assembly obtained data set by different programs (Velvet, CLC Genomics workbench, SOAPdenovo), but none of them demonstrated the capacity to assemble paralogous genome regions separately because of their high similarity. To get over this problems, the new algorithm for partition of reads

into subsets conforming to each of the paralogous genome regions, with using information of genome assembly of *Capsella rubella*(closely-related to *Capsella bursa-pastoris* plant) was developed. Generated subsets were used to obtain paralogous genome regions. Constructed sequences were annotated with using additonal sequenced data of *Capsella bursa-pastoris* transcriptome.

Results and conclusion

Genome of *C. bursa-pastoris* was sequenced and assembled using newly developed algorithm for separation of reads into subsets corresponding to the paralogous genome regions. Patterns of molecular evolution in paralogous genes were inferred.