DOI 10.18699/SBB-2019-05

Transcriptome assembling of non-model organisms on example holothuria Eupentacta fraudatrix

Boyko A.*, Dolmatov I.

National Scientific Center of Marine Biology of FEB RAS, Vladivostok, Russia

** e-mail: alteroldis@gmail.com*

Key words: transcriptome assembling, non-model orginasms, cardiac mechanics

Motivation and Aim: Over the past 5 years, more and more work on high-throughput sequencing of transcriptomes of non-model organisms has appeared. In such studies, there are many problems, for example a huge number of sequences in the assembly and complexity of predicted protein annotations. In connection with the study of the mechanisms of regeneration and development of echinoderms in our laboratory, we also encountered this problem.

Methods and Algorithms: Over the 400 millions clean reads were assembled using SPAdes 3.13[1] with 2 iterations for read correction step and with k-mer length of 25, 33 and 49. Of all the obtained contigs the Coding Sequences (CDSs) were extracted using TransDecoder 5.5.0[2]. The code of TransDecoder was modified in such a way that stop-codon or the beginning of the sequence, but not "ATG" (Met), was taken for the beginning of CDS. All CDSs were verified using BLAST-search by the SwissProt and Echinobase. Then all the obtained sequences were clustered into CD-HIT 4.7 [5] with three iterations. After each iteration, sequences in the clusters were assembled with an identity threshold of 80 %, using the own Python script, defined by us as HomoloCAP3. This script is a software add-on to CAP3[6] that makes it possible to use the data of pre-clustering of sequences and automatically select CAP3 parameters such as overlap and gap lengths and clipping range.

Results: The first stage of assembling in SPAdes resulted in a total of 703,169 contigs. This was unsatisfactory, as the level of fragmentation, the percentage of redundancy of almost identical contigs was high. Apparently, this situation arose due to the variability came from 5` and 3` untranslated regions. For this reason it was decided to use only CDSs for further assembling. As a result of the clustering and assembling with HomoloCAP3, filtering of the contaminant sequences and subsequent clustering with the aim to identify isoforms, we obtained a total of 85,805 contigs and 72,204 genes.

Conclusion: This approach to finalizing the assembly has been used for the first time and can significantly reduce the number of sequences with simultaneous increase in the number of full-length transcripts.

References

1. Bankevich A. et al. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol*. 2012;19:455-477.
2. Haas B.J. et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols*. 2013;8(8):1494-512.
3. Fu L. et al. CD-HIT : accelerated for clustering the next-generation sequencing data. *Bioinformatics*. 2012;28:3150-3152.
4. Huang X., Madan A. CAP3: A DNA Sequence Assembly Program. *Gen. Res*. 1999;9:868-877.