RESEARCH

Evaluation of transcriptome assembly platforms for RNA-seq method using non-model organisms

Naoaki Ono[1*], Yuki Okuda[1], Masanori Arita[2], Daisaku Ohota[3], Tetsuo Sato[1], Tadao Sugiura[1], Md. Altaf-UI Amin[1] and Shigehiko Kanaya[1]

[*]Correspondence: nono@is.naist.jp
[1]Graduate School of Information Science, Nara Institute of Science and Technology, 8916-5 Takayama, Ikoma, NARA, 630-0192, Japan
Full list of author information is available at the end of the article

Abstract

Massive sequencing of mRNA using next generation sequencers have become popular methods to analyze a non-model organism whose genome is not open. In order to identify genes expressed in the sample cell, it is necessary to assemble the short reads obtained by next generation sequencers and various methods and softwares have been proposed recently. In this study, we compared the sizes and numbers of assembled contig sequences using three different platforms and also evaluated data size to obtain gene sequnce efficently.

Keywords: Next generation sequencers; RNA-Seq; *de novo* assembly

Background

Recently, rapid development of massively parallel short read sequencers allows us to analyze whole transcriptome. However, algorithms to assemble those short read fragments are still under development. In this study, we compared three assemble softwares, Trinity[1], OASES[2], and SOAP-denovo-trans[3] developed for application of RNA-seq methods to understand advantage and disadvantage of each algorithms.

Another important parameter of RNA-seq analysis is the size of read samples. Though it is clear that more reads are required to assemble genes more correctly, we have to need to estimate appropriate size of read samples to save experimental costs. We evaluated the effect of data size on the assembled contig sizes and numbers by using randomly sampled read data.

Methods

The total mRNA of *Euglena gracilis* was sampled after 2 day cultivation and sequenced by illumina Hiseq 2000 as 100 bp paired-end library. The whole reads over 80M pairs were accumulated as a total read pool. To evaluate data size, paired reads were randomly sampled from the total read pool by 1M, 10M, 20M, 30M 40M, 50M, 60M, 70M, and 80M read pairs.

Results

The numbers and length of assembled contigs using each softwares were shown in Table 1. The number of contigs assembled by Trinity were the largest, while SOAP-denovo assembled the longest contig. Figure 1 showed how the assemble results depended on the data size. The largest contig length, N50 length and mean length saturate together around 40M read pairs, which is about 60 fold of the

estimated genome size of *E. gracilis*. On the other hand, the number of contigs are still increasing even at 80M read pairs. This result implies that too large sample size does not contribute to assemble of full length genes but increase diverse short fragments. Further analysis of variation of assembled sequences will provide useful information to optimize the method of RNA-seq.

Table 1 Comparison of assemble softwares

software	Trinity	Oases	SOAP
# of contigs	210,234	162,632	183,799
Longest contig len.	12,367	4,752	13,879
N50	1,226	550	887
mean len.	809	402	363

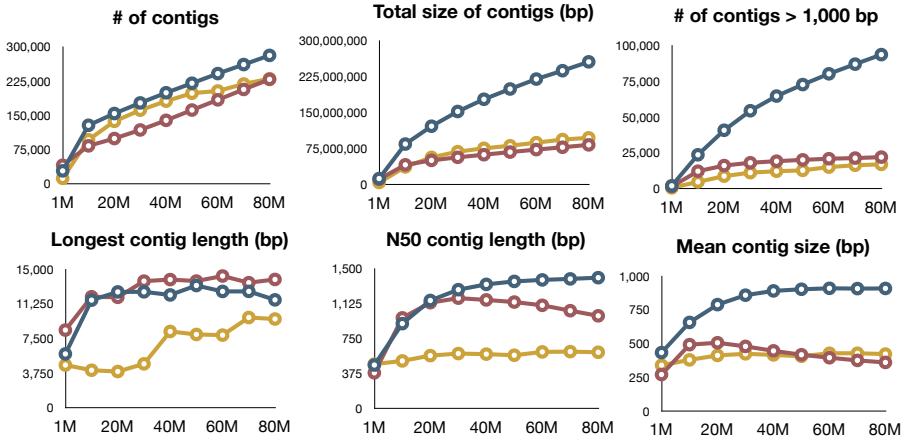

Figure 1 Evaluation of data size. Number of contigs and length as functions of number of sample reads. Blue, yellow, and red lines represent Trinity, OASES, and SOAP-denovo-trans, respectively. The results of Trinity were the longest and largest in almost all cases. The longest contig length and N50 were saturated around 40M pair reads, while the number of contigs still increases.

Author details
[1]Graduate School of Information Science, Nara Institute of Science and Technology, 8916-5 Takayama, Ikoma, NARA, 630-0192, Japan. [2]Department of Biophysics and Biochemistry Graduate School of Science, University of Tokyo, 2-11-16 Yayoi, Bunkyo-ku, Tokyo, 113-0033 Japan. [3]Graduate School of Life and Environmnet Science, Osaka Prefecture University, 1-1 Gakuen-cho, Nakaku, Sakai, Osaka, 599-8531 Japan.

References

1. Grabherr, M.G., *et al.*: Full-length transcriptome assembly from rna-seq data without a reference genome. Nat. Biotechnol. **29**, 644–652 (2011)
2. Schulz, M.H., *et al.*: Oases: robust *de novo* rna-seq assembly across the dynamic range of expression levels. Bioinformatics **28**, 1086–1092 (2012)
3. Francis, W.R., *et al.*: A comparison across non-model animals suggests an optimal sequencing depth for de novo transcriptome assembly. BMC Genomics **14**, 167 (2013)