Reducing the number of artifactual repeats in *de novo* assembly of RNA-Seq data by optimizing the assembly pipeline

ABSTRACT

One of the problems of de novo assembly is the occurrence of artifactual direct or inverted repeats that are mainly formed by misassembly of short sequencing reads and cannot be differentiated from real sequence repeats. In this study, we compared the frequency of artifactual repeats generated by four de novo assembly pipelines: (1) Velvet-Oases-The Gene Index Clustering Tool (TGICL), (2) Velvet-TGICL, (3) Trinity-TGICL and (4) SOAPdenovo-Trans-TGICL by analysing the RNA-Seq data of Gracilaria changii. The overall completeness of these four de novo assemblies were in the range of 85.2–90.0% for complete Benchmarking Universal Single-Copy Orthologs (BUSCOs), with the Velvet-TGICL assembly having the highest percentage of single copy and complete BUSCOs (78.9%). When Velvet-Oases-TGICL was used, a total of 2510 (8.44%) direct and 1967 (6.61%) inverted artifactual repeats were found among the assembled sequences. Polymerase chain reaction (PCR) analysis of 15 unigenes containing direct or inverted repeats confirmed that the repeats were due to assembly artifacts. When Oases was omitted from the assembly pipeline (i.e. Vetvet-TGICL), the number of unigenes containing artifactual direct and inverted repeats reduced significantly to 238 (1.63%) and 8 (0.06%), respectively. Among the four de novo assemblies, the Velvet-Oases-TGICL and Velvet-TGICL assemblies had the highest and the lower percentage of unigene containing artifactual repeats, respectively. The occurrence of artifactual repeats in the transcriptome data may complicate downstream analyses such as identification of splice variants and gene fusion, but the differential gene expression was less affected by the presence of artifactual repeats in this study. The information provided in this paper (based on a non-model seaweed G. changii) could be useful for designing and optimizing assembly pipeline for future analysis of RNA-Seq data from organisms without a reference genome.

Keyword: *De novo* assembly; Direct repeat; Inverted repeat; RNA-Seq; Transcriptome