Evaluation of de novo RNA-Seq assemblers in differential expression experiments

Lucas Miguel Carvalho, Zanoni Dia, Felipe Rodrigues da Silva

Unicamp, Embrapa Informática Agropecuária

Abstract

Evaluation of de novo RNA-Seq assemblers in differential expression experiments. RNA-Seq is a technology developed from Next-Generation Sequencing data (NGS) for transcriptome studies. It usually generates millions of short fragments of mRNA for contrasting treatments. Its reliability is undisputed for model organisms for which there are well assembled reference genome sequence and annotation available. However, generating an eukaryotic reference genome still is a difficult and expansive task. Assembling RNA-seq reads to describe an organism transcriptome without aligning the reads to its reference genome is called de novo transcriptomics. The objective of this study is to evaluate methodologies applied on de novo transcriptomics studies, proposing criteria for ranking the data, in order to maximize the chance of correctly identifying a differentially expressed transcript. The classification can help eliminate false positives transcripts usually found on the expensive and laborious downstream analysis, such as Real Time PCR. In this work, several parameters were tested with 3 assemblers: Trinity, Oases and IDBA-Tran. Actual RNA-Seq data from Arabidopsis thaliana and Canis vulgaris experiments were used for as input for the 3 assemblers. The list of differentially expressed genes was ranked using 15 different criteria and compared to the genes identified by the Tuxedo suite (BowTie, TopHat, Cufflinks, CummeRbund). Contrary to our expectation, the results show that the amount of true differentially expressed identified transcripts do not change significantly with reduction of input data. The assembler that consistently delivered best results was Trinity. The best ranking criteria was the transcript number of reads combined with p-value.