Finders keepers, nobody weepers! Unraveling novel genes in transcriptomes.

Marina Pupke Marone[1], Felipe Rodrigues da Silva[2],

1 Institute of Biology, University of Campinas
2 Embrapa Informática Agropecuária

Abstract

RNA-seq has become a standard procedure for measuring gene expression levels as it is a very sensitive and accurate tool. Studies involving RNA-seq generate a big amount of data and most of them are focused on finding the differentially expressed known genes, ignoring novel ones that might be present in the dataset. We are trying to devise an optimal pipeline to find novel genes in transcriptomes from organisms which have their genomes sequenced and several datasets available on the SRA database. In order to do this, we are going to use data from A. thaliana because of its importance as a model organism, making it easier to work with, added to the fact that there is a lot of data available. Among the 664 RNA-seq datasets found online for A. thaliana, we chose two very distinct to be analyzed. Both sets use the wild-type Col-0 ecotype, but the RNA of one of them was collected from the inflorescences, while the other was collected from the whole plant. Those sets were chosen to test whether RNA extracted from less complex tissues increase the chance of finding new genes. Transcriptomes sets were assembled using StringTie and Trinity in order to evaluate which one is the most appropriate for this task, considering the number of ESTs found and their performances. A transcript is deemed novel when it is described on our assembly but missing on the most recent genome annotation for that species. The presence of "old school" ESTs increases the confidence on the existence of the transcript.

Funding: CNPq