A Tool for the Comparison of Transcript Differential Expression Analysis Pipelines

Stefano Beretta[1], Yuri Pirola[1], Valeria Ranzani[2], Grazisa Rossetti[2]
Raoul Bonnal[2], Raffaella Rizzi[1], Gianluca Della-Vedova[1],
Massimiliano Pagani[2], Paola Bonizzoni[1]

1 Dipartimento di Informatica, Sistemistica e Comunicazione, Università degli Studi di Milano - Bicocca, Milano, Italy
2 Istituto Nazionale di Genetica Molecolare (INGM), Milano, Italy

Corresponding Author:
Stefano Beretta[1]
Email address: stefano.beretta@disco.unimib.it

INTRODUCTION

Long non-coding RNAs (lncRNAs) have recently gained interest, especially for their involvement in controlling several cell processes, but a full understanding of their role is lacking. Differential Expression (DE) analysis is one of the most important tasks in the analysis of RNA-seq data, since it potentially points out genes involved in the regulation of the condition under study.

However, a classical analysis at gene level may disregard the role of Alternative Splicing (AS) in regulating cell conditions. This is the case, for example, when a given gene is expressed in all the different conditions, but the expressed isoform is significantly diverse in the different conditions (that is an isoform switch).

A transcript level analysis may better shed light on this case, especially in studies having as goal, for example, a better understanding of the behavior of lncRNAs in lymphocytes T cells, which are fundamental in studies of specific diseases, such as cancer.

After Cufflinks/Cuffdiff, several approaches for DE analysis at isoform/transcript level have been proposed. However, their results are often sensitive to the upstream analysis such as read mapping, transcript reconstruction and quantification, and it is often hard to choose "a priori" the most appropriate combination of tools.

This work presents a tool for assisting the user in this choice, and poses the bases for a study devoted to the characterization of lncRNAs and the identification of of isoform switch events. Our tool includes a framework for the description and the execution of a set of DE pipelines over the same input dataset, as well a set of tools for reconciling and comparing the results.

METHOD

We designed an automated and easily customizable tool which is able to execute a set of existing pipelines for DE analysis at transcript level starting from RNA-seq data. Our method is built upon Snakemake, a workflow management system, with the specific goal of reducing the complexity of creating workflows. This approach guarantees that the experimentation is fully replicable and easy to customize. Each considered pipeline is structured in three steps: (i) transcript assembly, (ii) quantification, and (iii) DE analysis. By default, our tool builds and compares 9 different pipelines, each taking as input the same set of RNA-seq reads, obtained by combining different state-of-the-art methods to perform the transcript assembly (TA step) with different state-of-the-art methods to perform quantification and differential expression analysis (Q+DE step). More precisely, the 9 pipelines are obtained by combining two tools (Cufflinks and StringTie) and a Reference Annotation (Ensembl annotated transcripts) for the TA step, with three tools (Cuffquant+Cuffdiff, StringTie-B+Ballgown and Kallisto+Sleuth) for the Q+DE step. Each pipeline produces for each transcript a p-value, giving an evaluation of the statistical significance of its expression variation among the different conditions (opposed to the null hypothesis of a random variation).

RESULTS

We have tested our tool on 15 datasets of RNA-seq reads consisting of 3 individuals sequenced under 5 different conditions, as a starting point in the characterization of specific lncRNAs. The datasets have been produced by an Illumina HiScanSQ sequencer: each dataset contains on average 23.5 million paired-end sequences spanning the entire genome.

We have computed the correlation between the two sets of p-values for each pair of pipelines, observing that that the correlation coefficients are larger for some pairs of pipelines using the same approach for the Q+DE step. More precisely, the couples using Cuffquant+Cuffdiff have correlations between 0.86 and 0.89, while those employing StringTie-B+Ballgown have correlations between 0.83 and 0.85.

The correlation coefficients of all other pairs of pipelines (included those using Kallisto+Sleuth) are smaller than 0.4 (hence much less significant). A likely explanation is that the choice of the Q+DE tools crucially influences the final results, and is more important than the choice of the tool for the TA step. Still, we plan to perform an in-depth analysis of this phenomenon.

Moreover, our experiments have confirmed that the datasets contain two specific differentially expressed isoforms of the gene PTPRC, which is known in literature to have a switch event between those isoforms. We have also confirmed other transcripts which are compatible with annotated lncRNAs. A further work is to develop a better method to compute the correlate the transcripts assembled by the different pipelines, exploiting their predicted intron-exon structure to compute the comparison, and introducing an ad-hoc and robust method to estimate the correlation coefficients. Finally, a future development is to amalgamate the outputs obtained by the different pipelines to produce more reliable predictions.