Benchmarks for transcript assembly and differential gene expression (new isoforms) analysis in the context of transposable elements

Tina Alaeitabar, Nicolas Francillonne, Mikaël Loaec, Hadi Quesneville, Joelle J. Amselem

Benchmarks for transcript assembly and differential gene expression (new isoforms) analysis in the context of transposable elements

T. ALAEITABAR [1], N. FRANCILLONNE[1], M. LOAEC [1], H. QUESNEVILLE[1], J. AMSELEM[1]

1 INRA, UR1164 URGI - Unité de Recherche Génomique Info, INRA de Versailles, Route de Saint-Cyr, Versailles, 78026, France

Introduction

The project aim is to analyze RNA-Seq in the context of gene expression associated with structural changes (transcript isoforms) in different conditions, especially to study the hypothesis that transposable elements (TEs) inserted in the vicinity of genes may affect structural changes of transcripts. In the last few years, a number of transcriptome assemblers have been developed, but the real challenge is to choose one of the existing assemblers that perform well enough for data with different transcriptome complexities.
We show here preliminary results of our benchmarking to compare three different RNA assemblers, Cufflinks, Trinity and Grit. We also discuss some case studies of relation between TEs and modification of transcripts structure, especially at their start (TSS).

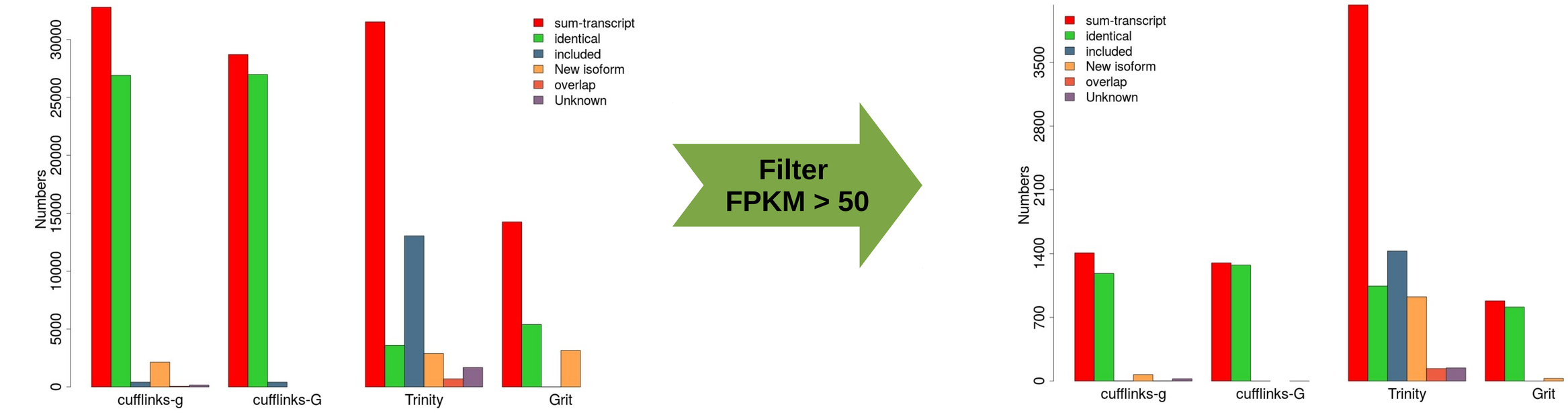
Figure 1: Pipelines used for benchmarking

The above flowchart (Fig.1) was run for two strategies of transcriptome assembly: The reference-based assembly using Cufflinks and Grit, *de novo* assembly using Trinity. The step of RNA-seq reads mapping to a reference genome, is done only for Cufflinks and Grit.

Strategy for RNASeq data analyzes

Here, we present the preliminary results (Fig. 2) of benchmarking using the well-annotated *Drosophila melanogaster* gene dataset against RNA-Seq transcript assemblies using Cufflinks (Trapnell, C. et al, 2010), Trinity (Grabherr, M. G. et al, 2011) and Grit (Boley, N et al, 2014). Then we performed Cuffcompare to compare the isoforms founded by each tool against the reference gene dataset. The result of Cuffcompare is the categorization of isoforms according to different predefined class codes (identical, included etc…).
Moreover, the resulting transcriptome assembly was cleaned by removing all transcripts nearly not expressed (FPKM < 50) to compare them again. Figure 2 reports the distributions of different class codes. Cufflinks was launch using –G/g by using reference annotation. Option –g allows to find potentially novel isoform fragment (at least one splice junction shared with a reference transcript). The results show that (i) Cufflinks recovers more isoforms identical to the reference (before and after FPKM cut-off) (ii) Trinity *de novo* assembly finds more incomplete isoforms (class “Included”) and (iii) Grit globally recovers less isoforms. For Grit, we could expect better results when polyA and TSS information (CAGE data) are used, as described in the publication. Therefore, due to better performance of Cufflinks, before and after FPKM cut-off, we use this tool in our analysis.

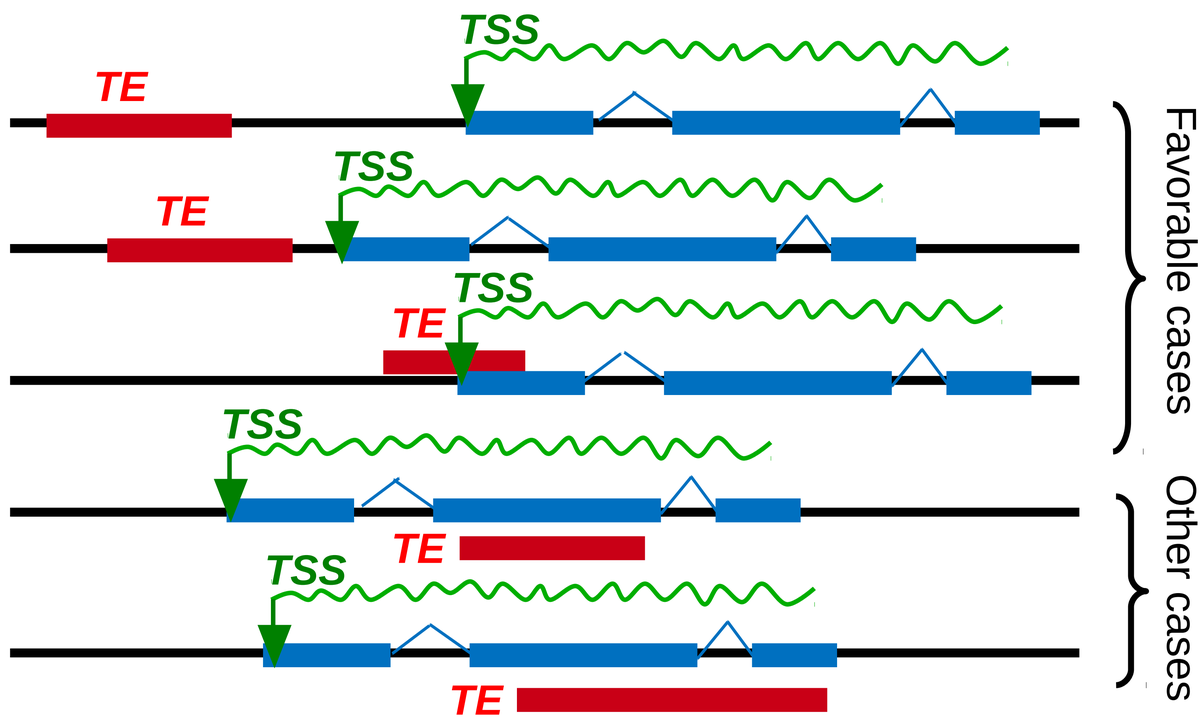
Figure 2: Comparison of isoform classes using CuffCompare. (A) Before FPKM filtering (B) After FPKM filtering



Impact of TE on transcription start site (TSS)

The existence of multiple TSS for a gene is a key event to create diversity and flexibility in the regulation of gene expression under differential conditions (biotic or abiotic). The genomes of most eukaryotes are composed of transposable elements (TEs). TEs are also reported to be carrier of significant signals for the initiation of RNA synthesis and processing. Thus, the presence/absence of TE near the 5’ region of a gene may have a role in creation of new TSS leading to the expression of a new isoform.
In this context, we studied the structural differences between isoforms at their 5'-flanking region i.e. near or not to TEs. Figure 3 shows different locations of TEs according to RNA-isoforms. The interesting case is when TE sets near TSS or overlaps with TSS in one the condition tested. We used bedtools-closest to calculate the distance between transcript and nearest TEs. In addition, we developed scripts to extract favorable cases from the results of bedtools-closest to compare transcript isoforms expressed under different conditions.

Figure 3: The different locations of TEs according to RNA-isoforms



Perspective:

We are currently developing an executable pipeline (SSH command line mode) to auto-mate all above processes. This pipeline will involve the following steps: (i) Alignments of RNA-seq reads to a reference genome using TopHat (ii) Assembly of RNA-seq reads into transcripts and report of the FPKM using Cufflinks (iii) Comparison of assembled transcripts of two distinct conditions (or to a reference annotation) using Cuffcompare (iv) Finding distances between TSS and nearest TEs, using Bedtools-closest. Then, we plan to apply this pipeline for Oak data from France-Genomique OakAdapt project. Additionally, we will develop a suite of Galaxy wrappers for this complete pipeline that will be accessible on a Galaxy instance.