transcAnalysis: A Snakemake Pipeline for Differential Expression and Post-transcriptional Modification Analysis

Pedro H. A. Barros[1,2](✉), Waldeyr M. C. Silva[1], and Marcelo M. Brigido[1]

[1] Department of Cellular Biology, Institute of Biology, University of Brasilia, Brasĺlia 70910-900, Brazil
[2] Graduate Program in Molecular Biology, University of Brasilia, Brasĺlia, Brazil
pedroa_barros1@hotmail.com

Abstract. The transcAnalysis pipeline is a comprehensive tool that allows the analysis of transcriptome data. The pipeline allows for analysis of differential expression, alternative splicing, lncRNA and RNA editing analysis, with a specific focus on A-to-I editing mediated by the ADAR protein. This type of RNA editing is widespread and can significantly affect gene regulation and function. The results from these analyses are integrated, and the events are associated with each gene. The pipeline also integrates results that can help correlate gene expression and post-transcriptional events. This allows for a comprehensive understanding of the functional impact and provides insight into the biological processes and pathways associated with these events. One of the significant advantages of the transcAnalysis pipeline is its ability to perform all these analyses with a single command using the Snakemake package. This feature simplifies the analysis process and makes it accessible to researchers with limited bioinformatics expertise. Its user-friendly ability to perform multiple analyses with a single command make it an ideal choice for researchers looking to analyze transcriptome data.

Keywords: Pipeline · Differential Gene Expression · Alternative Splicing · RNA Editing

1 Introduction

The advent of next-generation sequencing (NGS) technologies has brought about a wealth of data that can be used to gain insights into biological systems. RNA sequencing (RNA-seq) has become the cornerstone for studying gene expression, with a primary focus on differential expression analysis [1]. However, this approach often overlooks other important information related to post transcriptional modifications and the expression of long non-coding RNAs (lncRNAs), which can also be included in the analysis.

M. S. Reis and R. C. de Melo-Minardi (Eds.): BSB 2023, LNBI 13954, pp. 106–111, 2023.
https://doi.org/10.1007/978-3-031-42715-2_10

One of the critical post-transcriptional modifications is alternative splicing (AS), which involves the removal of introns and non-canonical joining of exons from pre-mRNA, leading to the production of different proteins and transcriptional control. There are five primary forms of AS: exon skipping (ES), alternative 5' splice site (A5SS), alternative 3' splice site (A3SS), mutually exclusive exons (MXE), and intron retention (IR) [2]. These processes play a critical role in regulating gene expression and contribute to the diversity of the proteome.

RNA editing (RED) is mainly carried out by the protein ADAR in mammals, leading to the editing of adenine to inosine (A-to-I), which modifies the secondary structure of transcripts and alters the binding of RNA-binding proteins and miRNAs. RED can also lead to the generation of transcript isoforms through AS. These events are related to specific conditions and can aid in understanding biological phenomena [3]. By considering AS and RED in RNA-seq analyses, researchers can gain a more comprehensive understanding of the transcriptional and post-transcriptional mechanisms underlying gene expression.

2 Material and Methods

2.1 Workflow

The transcAnalysis pipeline performs mRNA, lncRNA, AS, and RE expression analysis from a BAM (Binary Alignment Map) file created after the alignment of RNA-seq reads from fastq files. Additionally, the pipeline requires a metadata file that is filled out by the user with their desired preferences to be executed, including the path to each sample. The pipeline facilitates the analysis by allowing the execution of each step with only one command line, which is possible due to the use of the snakemake, a workflow management system that provides integration with Conda and Docker, two popular tools in the bioinformatics community, to enhance reproducibility and portability of analysis pipelines [4]. The pipeline outputs the data integration related to the analyzed event (Fig. 1) and is available at: github.com/PHAB1/transcAnalysis.

2.2 Transciptome Analysis Pipeline

From the output obtained by the pipeline, the integration of the data related to each gene simultaneously to differential expression, RED (A-to-I), and AS is performed, considering each AS category separately (Table 1).

Differential Expression. Differential expression analysis was performed using the R package DESeq2[1] (v1.38.2), developed by Bioconductor project. The program specializes in normalization, visualization, and differential gene expression analysis, and uses empirical Bayes techniques to estimate the log of fold change, the dispersion, and related estimates such as the p-value and the False Discovery Rate (FDR) [5]. The program uses the gene expression array and a table containing the experimental design and performs differential expression analysis from two or more different conditions.

[1] https://bioconductor.org/packages/release/bioc/html/DESeq2.html.

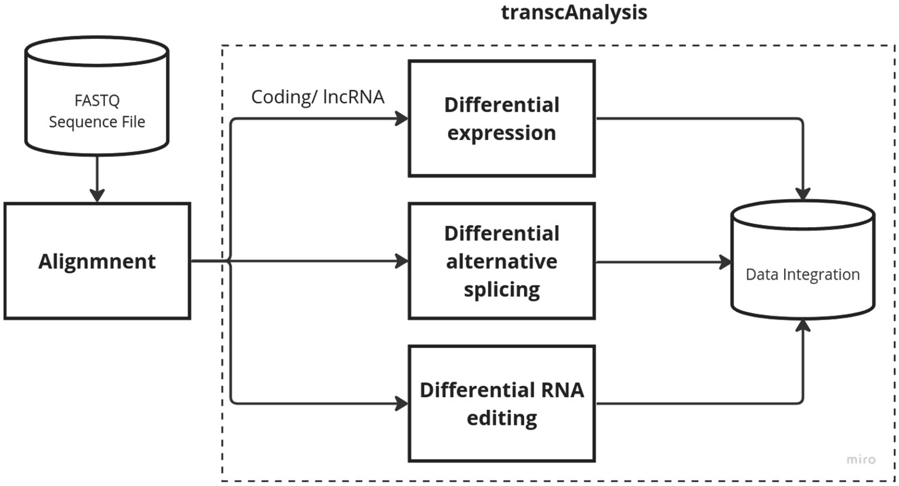

Fig. 1. Workflow and steps in the transcAnalysis pipeline

Table 1. Integration of transcriptome data. Only Fold change (FC) of which had $FDR < 0.05$ are shown. Similarly, the event counts of Exon Skkiping (ES), Intron Retention (IR) and RNA editing shown are significant with $FDR < 0.05$.

Gene	FC	SE	RI	...	RED
CTSS	0	4	0		42
APOBEC3C	0	0	0		13
METTL7A	0	0	0		16
⋮	⋮	⋮	⋮		⋮
IFITM2	2.35	0	0		3
RBM39	0	8	1		2

Alternative Splicing. For detection of differential AS, the rMATS[2] turbo program (v4.1.2) was used. The program is based on using mapped reads in regions indicating different isoforms to detect and estimate the proportion of different types of splicing between different conditions. Here, the five possible splicing types are identified, these being ES, IR, MXE, A3SS, and A5SS [6]. As an example, in ES, the reads mapped in the regions used as AS markers are the junction reads (S), positioned between the upstream exon and the downstream exon with single exon skipping, forming the isoform, while the reads related to the canonical form of the gene, with the inclusion of the exon, are the inclusion reads (I). Inclusion levels (ψ) are defined as the percentage of transcript inclusion [6].

$$\psi = (I/LI)/(I/LI + S/LS),$$

[2] https://github.com/Xinglab/rmats-turbo.

where:

$\psi =$ Inclusion Level (IncLevel),
$I =$ number of reads mapped to the inclusion isoform,
$S =$ number of reads mapped to the ES isoform,
$LI =$ effective length of the inclusion isoform exon,
$LS =$ effective length of the ES isoform.

For filtering of the splicing events, after identification with rMATS, $FDR < 0.05$ and $|\psi_{it} - \psi_{ic}| > 0.1$ were used, where ψ_{ic} and ψ_{it} indicates the relative mean ψ in each event between the (t) treatment and (c) control.

RNA Editing. For the detection of RED, the program SPRINT[3] (v0.1.8) was used. The SPRINT program preprocesses the reads by removing annotated single nucleotide genetic variants (SNPs), aiming to remove false positives from genetic variants, and subsequently identifies ER candidates [7]. Only candidates containing nucleotides with high quality $q > 20$ and mapped reads in regions with indistinguishable repeats are retained, removing those with mapping or sequencing errors.

LncRNA. Annotation of lncRNAs is performed and differential expression analysis is done with the DESeq2 program. For pathway enrichment analysis, the R package LncRNAs2Pathways[4] (v.1.1), developed to associate pathways related to lncRNAs, is used. The package uses a network library, which relates lncRNAs to the expression of mRNAs [8]. The interaction network was created from the analysis of 28 different studies and takes into account protein-protein interactions annotated in different databases such as REACTOME [9] and HPRD [10], and is able to relate differentially expressed lncRNAs to “KEGG” [11] or “Reactome” [9] pathways.

2.3 Experimental Transcriptome Data

The transcriptome samples in fastq format were obtained from the Sequence Read Archive (SRA) database. The experimental files used were all paired-end files. We used monocyte samples from patients with severe-stage COVID-19 (PRJNA699856). 6 treated and 6 controls were used. The STAR[5] program (v2.7) was used to align each sample to the reference genome GRCh38 (hg20) [13].

3 Results and Discussion

The pipeline performs statistical analysis between two distinct groups, including differential analysis of gene expression, lncRNA, AS, and RED. The pipeline is

[3] https://github.com/jumphone/SPRINT.
[4] https://cran.r-project.org/web/packages/LncPath/.
[5] https://github.com/alexdobin/STAR.

designed to integrate these events, allowing a complementary analysis to conventional expression analysis, as shown in Fig. 2, where the intersection of enriched terms between differentially expressed genes, differential alternative splicing, and differentially edited RNAs in monocytes from patients with severe COVID-19 versus the healthy patient is shown.

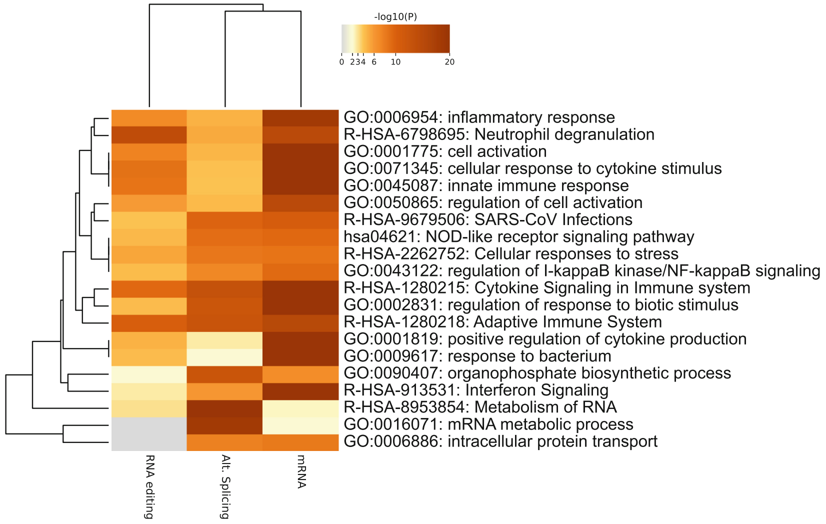

Fig. 2. Metascape functional analysis Heatmap of the transcriptome. Intersection of the most relevant terms in mRNA-related enriched genes, alternative splicing or RNA editing events.

3.1 Conclusion

There is a large amount of information in transcriptome data that is not normally used, programs that identify and analyze post-transcriptional modifications have no trivial use and require computer skills. The transcAnalysis pipeline was created with the intention of allowing the acquisition of data related to both gene expression and post-transcriptional modifications for the utilization of the data and integration, allowing association between the events. In addition, the Snakemake pipeline manager was used to create a user-friendly approach.

References

1. Hardwick, S., Deveson, I., Mercer, T.: Reference standards for next-generation sequencing. Nat. Rev. Genet. **18**(8), 473–484 (2017)
2. Marasco, L.E., Kornblihtt, A.R.: The physiology of alternative splicing. Nat. Rev. Mol. Cell Biol. **24**(4), 242–254 (2023)

3. Song, B., Shiromoto, Y., Minakuchi, M., Nishikura, K.: The role of RNA editing enzyme ADAR1 in human disease. WIRES **13**(1), e1665 (2023)
4. Mölder, F., et al.: Sustainable data analysis with Snakemake. F1000Res **18**, 10–33 (2021)
5. Love, M.I., Huber, W., Anders, S.: Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. Genome Biol. **15**(12), 550 (2014)
6. Shen, S., et al.: rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. Proc. Natl. Acad. Sci. U.S.A. **111**(51), E5593–E5601 (2014)
7. Zhang, F., Lu, Y., Yan, S., Xing, Q., Tian, W.: SPRINT: an SNP-free toolkit for identifying RNA editing sites. Bioinformatics **33**(22), 3538–3548 (2017)
8. Han, J., et al.: LncRNAs2Pathways: identifying the pathways influenced by a set of lncRNAs of interest based on a global network propagation method. Sci. Rep. **7**, 46566 (2017)
9. Gillespie, M., et al.: The Reactome pathway knowledgebase 2022. Nucleic Acids Res. **50**(D1), D687–D692 (2022)
10. Mishra, G.R., et al.: Human protein reference database-2006 update. Nucleic Acids Res. (34), D411–D414 (2006)
11. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., Morishima, K.: KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. **45**(D1), D353–D361 (2017)
12. Nishimura, D.: BioCarta. Biotech Softw. Internet Rep. **2**(3), 117–120 (2001)
13. Dobin, A., et al.: STAR: ultrafast universal RNA-Seq aligner. Bioinformatics **29**(1), 15–21 (2013)