Caring is Sharing – Exploiting the Value in Data for Health and Innovation
M. Hägglund et al. (Eds.)

doi:10.3233/SHTI230343

OmicSDK-Transcriptomics: A Web Platform for Transcriptomics Data Analysis

Aurora SUCRE[a, b, 1], Maria MARTINEZ [b] and Alba GARIN-MUGA [a, b]
[a] *Biodonostia Health Research Institute, Basque Research and Technology Alliance (BRTA), 20014 Donostia-San Sebastián, Spain*
[b] *Fundación Vicomtech, Basque Research and Technology Alliance (BRTA), 20009 Donostia-San Sebastián, Spain*
ORCiD ID: Aurora SUCRE https://orcid.org/0000-0002-4078-9275
ORCiD ID: Alba GARIN-MUGA https://orcid.org/0000-0002-7160-1191

Abstract. Omics sciences, especially transcriptomics, have grown exponentially since the first human genome was sequenced in 2003. Different tools have been developed in the past years for the analysis of this kind of data, but many of them require specific programming knowledge to be used. In this paper, we present omicSDK-transcriptomics, the transcriptomics module of OmicSDK, a comprehensive tool for omics data analysis that combines pre-processing, annotation and visualization tools to be used with omics data. OmicSDK comprises a command-line tool and a user-friendly web solution, so researchers having different backgrounds can take advantage of all its functionalities.

Keywords. Transcriptomics, bioinformatics, data analysis, visual analytics

1. Introduction

Since the sequencing of the first human genome in 2003, omics sciences have grown exponentially. One of the most developed sciences in the field is transcriptomics, which focuses on the study of the transcriptome, the sum of all the RNA transcripts of an organism [1]. Transcriptomics studies usually focus on the levels of expression of these transcripts (i.e., how many copies are found), and this information can be used to define disease biomarkers and predict risks or potential treatment responses [2].

Although different tools for transcriptomics analysis exist, these require bioinformatics expertise, and normally, multiple tools are needed to perform an analysis. Therefore, there is a lack of easy-to-use solutions that allow users with little experience and limited programming knowledge to perform an entire analysis using a single tool.

To fill this gap and provide an easy-to-use solution for transcriptomics analysis, our goal was to assess a typical transcriptomics pipeline and explore tools widely used in this area, and finally develop the transcriptomic analysis modules of OmicSDK. OmicsSDK is a comprehensive software library for the analysis and visualization of omics data. The new modules allow the processing of raw RNA-seq data, differential expression analysis (DEA), functional analysis, and visualization of results using charts.

[1] Corresponding Author: Aurora Sucre, Vicomtech Foundation; E-mail: amsucre@vicomtech.org.

2. Methods

OmicSDK-transcriptomics was developed to ease the analysis and visualization of transcriptomics data and integrate into a single platform all the tools needed for a typical analysis. Initially, a review of literature was followed, then the different requirements were defined and finally the solution was developed, following an iterative process.

2.1. Literature review and definition of requirements

An extensive literature review was performed to define a common processing pipeline for gene expression data, detect which visual analytics approaches could be useful in that context, and define all functional requirements. During this process, no tools were found to perform all the expected steps, but many open-source libraries focusing on specific tasks were detected[2].

Besides the functional requirements defined during the review, it is a key factor that the solution is intuitive and simple enough, so researchers coming from diverse backgrounds and having different programming skills can exploit it.

2.2. OmicSDK-transcriptomics design and development

OmicSDK is a software library composed by three main modules for the ***1) analysis, 2) screening*** and ***3) visualization*** of omics data. For each omics field, different functionalities are included in the three modules for the comprehensive analysis of various datasets. This paper focuses on the sub-modules for transcriptomic analysis, which exploit and combine existing open-source tools to provide a complete pipeline for mRNA and miRNA data.

2.2.1. Analyse-OMICS

This module puts together all functions needed for primary analysis, to process raw sequencing files (FASTQ) and finally obtain total counts or normalized expression values. Different pipelines were implemented for mRNA and miRNA data, using the following tools: ***STAR*** [3] for RNA-seq data alignment, ***CUFFLINKS*** [4] to process RNA-Seq data to obtain the relative abundance of the transcripts, and ***miRge 3.0*** [5], which was designed specifically for miRNA-Seq data analysis and performs all relevant pre-processing steps using other third-party tools such as Cutadapt (adapter trimming), Samtools (managing sequencing files), and Bowtie (short reads alignment).

2.2.2. Screen-OMICS

The second module includes functionalities for secondary and tertiary analysis. Some tools were developed to obtain further insight into the data, and others to compare multiple datasets. The module is based on three R packages: ***TCGAbiolinks*** (differential expression analysis), ***ClusterProfiler*** (functional annotation) & ***GOSim*** (GO terms clustering and classification).

[2] Due to format constraints, we are unable to go into as much depth in this section as expected. For more details on the review, please contact the authors.

2.2.3. Visual-OMICS

The final module encompasses functions for data visualization using plots. Most functions were based on R packages such as ***ggplot2*** (bar plots, heat maps, and dot plots), ***circlize*** (chord diagrams) and ***networkD3*** (sankey diagrams).

2.3. Web development

To make the tool available to researchers with different backgrounds, an intuitive web solution was designed so programming knowledge is not required to use OmicSDK.

The solution is expected to grow iteratively; thus, a modular architecture was chosen to ease the inclusion of new functions without affecting the behaviour of other modules.

The web platform was developed using Angular and different Angular libraries such as Bootstrap to create visual interfaces and D3 to create interactive charts. The Angular front-end is connected to the OmicSDK functionalities through a Flask-based backend.

3. Results

OmicSDK is a software library composed by different modules to analyse several types of omics data, covering all stages of analysis. The functionalities are sorted into three main modules according to the analytical stage where they are used. Regarding transcriptomics analysis, OmicSDK can be used for raw data pre-processing, DEA, functional analysis, and/or data visualization, as shown in **Figure 1**.

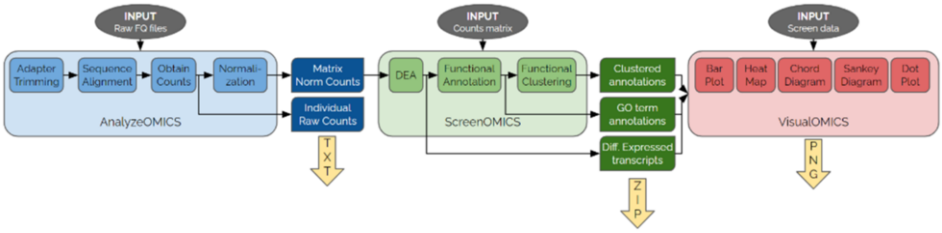

Figure 1. OmicSDK-transcriptomics pipeline: Modules, functionalities, and possible paths to follow.

3.1. Analyse-OMICS

This module allows users to process raw sequencing data to obtain total counts or normalized expression values. It is available for both mRNA and miRNA sequencing data. The following steps are followed to analyse each sample independently: ***1) Adapter trimming*****:** Input FASTQ reads may have adapter sequences at the ends. This initial step ensures only the actual sequence is kept by removing all adapters. ***2) Sequence alignment:*** Given a specific reference genome, all reads are aligned against it to determine how many correspond to each transcript. ***3) Obtain counts:*** During this step, the read counts (number of reads per transcript) are obtained. ***4) Normalization:*** Raw read count values for different samples are not comparable, so a final normalization step is needed. For miRNA data, the RPM (reads per million maps read) value is obtained, but for mRNA data, the RPKM (reads per kilobase per million mapped reads) value is obtained instead, because, in this case, the read length must be considered.

In case the user wants to further analyse and compare a dataset of multiple samples, the module also allows to merge all normalized values in a single matrix or obtain two independent matrices corresponding to different conditions to be compared.

3.2. Screen-OMICS

This module was intended for the comparison of expression data of two well-distinguished groups, as is commonly done in transcriptomics. To do so, three main functionalities were developed: ***1) DEA:*** During differential expression analysis, the normalized counts of two separate groups are statistically analysed to detect significant changes in expression levels between the two groups. For each transcript, a fold change value is obtained, representing how overexpressed or under expressed the transcript is in one group with respect to the other. Also, the statistics, associated p-values and FDR values are obtained, and results can be filtered according to this. ***2) Functional annotation:*** This step is performed to gain further insight into the differentially expressed transcripts and their biological function. The user must provide the list of transcripts, the algorithm for p-value adjustment and which ontologies to consider for annotation (molecular functions, cellular components and/or biological processes). After annotation, a list of GO terms is given. ***3) Functional clustering:*** The last step is to analyse the list of functional annotations to group them according to their similarities. Clustering is performed using the method chosen by the user (options are given).

3.3. Visual-OMICS

The final module includes the tools developed for result visualization, providing users with different graphical elements for a better understanding of the output. The results generated in previous modules are shown in tables and output files, but this module allows plotting these results, specifically using: ***bar plot*** (mean expression values for each transcript, for each group of samples), ***heat map*** (mean expression values, using a colour map), ***chord diagram*** (circular plot showing the relation between genes and their GO functions), ***Sankey diagram*** (2-column plot showing relations between GO terms and transcripts) and ***dot plot*** (mean expression for different group of samples. Distinct groups of genes are plot with different colours). Plot examples are shown in **Figure 2**.

3.4. Web tool and typical pipeline

All functions included in the software library can be used from the command line, so more experienced users can exploit all the benefits. Nevertheless, an angular-flask web solution was also developed, so OmicSDK is available to researchers with different backgrounds, including those with limited programming skills.

Each analytical module was embedded in a different section of the web tool, so they could be used independently. Nevertheless, a communication service was defined, so all elements are connected, and a complete analysis can be performed easily. Users may begin and end their analysis at any desired stage (see **Figure 1**), and all resulting files and plots are downloadable.

A module for interactive visualization was developed exclusively for the web-based tool, where all the described plots can be generated, but interactive utilities that allow users to customize their plots are available.

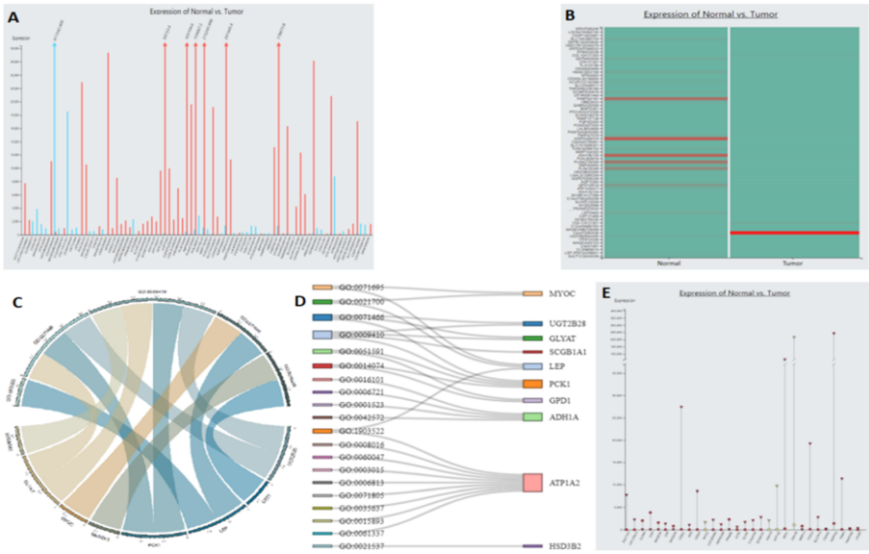

Figure 2. Visual-OMICS plots: A) Bar plot, B) Heat map, C) Chord diagram, D) Sankey plot, E) Dot plot

4. Conclusions

Even if multiple tools are available for RNA-Seq analysis, most of them only implement certain steps of the process or are overly complex to use. In other cases where more comprehensive tools were found, important functionalities were still missing.

OmicSDK-transcripts was designed to cover the main steps of a transcriptomics analysis. Its web implementation, having a user-centred design, should ease the usage for people with different backgrounds. It is expected to be an intuitive and easy-to-use solution, but usability and validation tests are still under design, so its real potential cannot be assured until that stage is reached. However, preliminary tests with subjects with no expertise (students, volunteers with no experience…) returned positive feedback.

The modular nature of the solution allowed its quick integration within the OmicSDK framework, taking advantage of its background capabilities. It will also ease the implementation of other functionalities in the future, when new necessities are detected. Finally, it makes possible to combine the available modules as needed.

References

[1] Ganguly P. Transcription [Internet]. Talking Glossary of Genomic and Genetic Terms. Available from: https://www.genome.gov/genetics-glossary/Transcription

[2] Borràs DM, Janssen B. The use of transcriptomics in clinical applications. Integration of Omics Approaches and Systems Biology for Clinical Applications. 2018;49–66.

[3] Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. Star: Ultrafast universal RNA-seq aligner. Bioinformatics. 2012;29(1):15–21.

[4] Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, et al. Differential gene, and transcript expression analysis of RNA-seq experiments with TopHat and cufflinks. Nature Protocols. 2012;7(3):562–78.

[5] Patil AH, Halushka MK. Mirge3.0: A comprehensive microRNA and TRF Sequencing Analysis Pipeline. NAR Genomics and Bioinformatics. 2021;3(3).