Strategies for data analysis of bacterial RNA-Seq

Márlon G. F. Custódio[1]*, Martin S. Grecco[2]*, Vitor L. Coelho[1]*, Marisa F. Nicolás[1]

[1]National Laboratory for Scientific Computing
Petrópolis – RJ – Brazil

[2]Laboratorio de Virología Molecular, Centro de Investigaciones Nucleares,
Facultad de Ciencias, Universidad de la República
Montevideo – Uruguay
{marlonc,vitorlc,marisa}@lncc.br, msonora@cin.edu.uy
*These authors contributed equally to this work.

Abstract. *The analysis of transcriptomes by RNA-seq technique is currently the best way to assemble the transcriptome profile of an organism. In this paper, we present a protocol for analysis of RNA-seq, considering possible problems in the samples, such as contamination by ribosomal RNA, what happens with some frequency in preparations of libraries by various RNA-seq platforms.*

1. Introduction

Analysis of DNA using Next Generation Sequencing (NGS) techniques has increased substantially recent years. Nowadays, sequencing plataforms generate a large amount of data, allowing a deep analysis of several organisms. The study of transcriptome give us general view of transcripts sets in organism or cell. This enables us to evaluate gene expressions in specific conditions. RNA-seq is a technique for transcriptome analysis that allows us a large scale gene expresion evaluation. This paper presents a methodology for transcriptome study using RNA-seq data of bacteria subspecies of *Pseudomonas aeruginosa*, *Acinetobacter baumannii* and *Klebsiela pneumoniae* obtained by FLX Roche 454 sequencing plataform. It was verified the transcriptome profile under some conditions that directly influence the set of expressed genes.

The methodology is divided in two stages: pre-processing and gene expression analysis. Pre-processing stage manipulates the data that are received directly from the sequencer. In general, this step requires a greater amount of time and computational processing. Gene expression stage consists of the analysis of gene expression levels.

2. Related work

This section cites some related studies that inspired the methodology described in this paper.

The bacterial genome researches require specific protocols and softwares for their execution. The software Rockhopper [McClure et al. 2013] runs several stages of bacterial RNA-seq data analysis, including alignment, quantification, diferential expression test, and characterization of operon structures. The software EDGE-Pro [Anders et al. 2013] is used to quantify expression levels in bacteria and some prokaryotes and provides needed estimatives for diferential expression calculation. This calculation is performed by DESeq (http://bioconductor.org/packages/release/bioc/html/DESeq.html), a Bioconductor package. DESeq does not need the pre-processing stage, because it uses

coordinates file of coding genes (*ptt*) and coordinates file of rRNAs and tRNAs (*rnt*), including reference genome. Therefore, this paper presents a strategie allowing bacterial RNA-seq data analysis when the input data has a large amount of contaminants and low coverage.

3. Material and methods

In this section, we describe the stages of pre-processing of data sequencing and gene expression analysis.

3.1. Pre-processing

Pre-processing is composed by five steps: extracting of libraries, trimming, sequence alignment, rRNA filtering and duplicate deletion. Thereby, this stage is critical to have better balance of cost and benefit of reads to the next step. The following sections describe its steps and used softwares.

3.1.1. Extracting libraries

The read libraries were extracted from sequencer data volumes using 454 Sequencing System Off-Instrument Software Applications suite (http://454.com/products/analysis-software/index.asp). Among other algorithms, this software provides the tool sffile to decompress read libraries and ssinfo to generate the FASTA and FASTQ files from uncompressed libraries.

3.1.2. Trimming

The software Lucy [Chou and Holmes 2001] was applied in FASTA and FASTQ libraries to trimming and to filter low quality reads.

3.1.3. Sequence alignment

In this step, the software Bowtie 2 [Langmead and Salzberg 2012] aligns libraries and reference genome. Firstly, Bowtie2-build algorithm creates mapping index between reads libraries and reference genome. Next, Bowtie2-align algorithm aligns filtered sequences by Lucy using the created index. The output is a SAM format file. Then, the tool SAM (Sequence Alignment/Map) [Li et al. 2009] compresses the output SAM files to BAM format. The aim is to create files with ordered and indexed sequences.

3.1.4. Filtering for ribossomal RNA

The filtering of libraries is performed against rRNA contamination. Initially, we obtain a BED format file with rRNA coordinates in reference genome. The localization and indexes of rRNAs is identified in the libraries (BAM format, indexed and ordered) with these coordinates. The rRNAs are removed from libraries using the indexes. The BED file is generated by native commands of Linux and AWK language for text manipulation.

3.1.5. Removing duplicates

The CD-HIT-454 [Fu et al. 2012] is used for removing duplicate sequences. Duplicates are either exactly identical or meet these criteria includes: they start at the same position; their lengths can be different, but shorter one must be fully aligned with the longer one (the seed); they can only have 4% mismatches (insertion, deletion, and substitution)(can be adjusted); and only 1 base is allowed per insertion or deletion (can be adjusted). This process reduces the nonspecifically generated replicates during one of the stages of the sequencing workflow, which eliminates errors in the analysis of differential gene expression.

3.2. Diferential expression analysis

The stage of differential expression analysis consists of reads quantification and evaluation of gene expression levels. The quantification is performed by the tool HTSeq-count [Anders et al. 2014], a algorithm of HTSeq python package. The SAM files of libraries and genome annotation are the input. HTSeq-count generates a output file with the quantification of reads per feature. Then, the package edgeR calculates differential expression following the steps described in [Anders et al. 2013]:

- filtering of features with low expression or not aligned (step 2, section 14 A);
- estimation of normalization factors (steps 4 e 5, section 14 A);
- estimation of tagwise dispersion (step 7, section 14 A);
- differential expression test (step 9, section 14 A);
- statistics of differential expression test (step 6, section 14 B);

4. Results and discussion

This protocol was executed using a PC with CPU Intel® Core™ i5-3550 3.30GHz, with 6 GB RAM memory, operating system Linux Ubuntu 12.04. The runtime for pre-processing was $\simeq 3$ hours.

The sample consist of 3 runs, containing 2 libraries each one, resulting a total of $4,530,492$ reads. Approximately 95% of reads $(4,321,072)$ show low quality according to software Lucy. In addiction, $3,972,311$ reads were removed in rRNA filtering, demonstrating a great contamination by rRNA. After removal of duplicates with CD-HIT-454 remained $311,711$ reads. This problem becomes unusable the libraries as input in Rockhopper, because the software needs a larger number of reads. For our approach, although decreasing the coverage, this datasets could be analyzed. In addiction, EDGE-Pro could not be used because avaliable file formats were incompatible with the needed files to run it.

5. Conclusion

In this paper, we presented a methodology to study bacterial RNA-Seq data obtained from FLX Roche 454 sequencing plataform. However, this approach can be extended to another sequencers data. Moreover, except the rRNA removal, the steps can be used to study eukaryoto transcriptome. In pre-processing, a high rRNA contamination was verified (approximately 93%), culminating in a drastic decrease of reads. However, in the

diferential expression stage, we could verify aceptable results, even with low coverage in sequencing. This shows efficacy in analysis of transcriptome using RNA-seq data.

The authors acknowledge the assistance derived from FAPERJ, process No. E-26/110.873/2013. Dr. William Loss, Dr. Rafael Guedes and MSc. Guadalupe del Rosario Saji for advice during the stage of pre-processing of data.

References

Anders, S., McCarthy, D. J., Chen, Y., Okoniewski, M., Smyth, G. K., Huber, W., and Robinson, M. D. (2013). Count-based differential expression analysis of rna sequencing data using r and bioconductor. *Nat. Protocols*, 8(9):1765–1786.

Anders, S., Pyl, P. T., and Huber, W. (2014). Htseq; a python framework to work with high-throughput sequencing data. *bioRxiv*.

Chou, H.-H. and Holmes, M. H. (2001). Dna sequence quality trimming and vector removal. *Bioinformatics*, 17(12):1093–1104.

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). Cd-hit: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, 28(23):3150–3152.

Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with bowtie 2. *Nat Meth*, 9(4):357–359.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Subgroup, . G. P. D. P. (2009). The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–2079.

McClure, R., Balasubramanian, D., Sun, Y., Bobrovskyy, M., Sumby, P., Genco, C. A., Vanderpool, C. K., and Tjaden, B. (2013). Computational analysis of bacterial rna-seq data. *Nucleic Acids Research*, 41(14):e140.