The impact of quality filter for RNA-Seq data over differential expression profile

Pablo Gomes de Sá[1], Siomar de Castro Soares[2], Adonney Allan de Oliveira Veras[1], Anne Cybelle Pinto[2], Luis Guimarães[2], Vasco Azevedo[2], Artur Silva[1#], and Rommel Thiago Jucá Ramos[1#]

[1] Federal University of Pará, Biological Science Institute, Belém/PA, Brazil
{pablogomesdes,allanverasce,arturluizdasilva,rommelthiago}@gmail.com
[2] Federal University of Minas Gerais, Biological Science Institute, Belo Horizonte/MG, Brazil
{siomars,acybelle,luisguimaraes.bio,vascoariston}@gmail.com

# The authors contributed equally to this work.

The advent of new genome sequencing platforms in 2005, generally named next-generation sequencers (NGS), has boosted the number of complete genomes deposited in public databases, mainly due to the ability of those platforms to generate a huge volume of genomic data in a faster and cheaper fashion than the previous methodologies. Concomitantly, gene expression analyses have also been favored due to the development of RNA-Seq technique [1], which evaluates the whole transcriptional profile of an organism and is also used to identify new transcripts and to correct genome annotations. More interesting, the genome sequencing is not required to be finished in order to perform RNA-Seq analyses, like in Real Time technologies [2].

Data generated by high-throughput platforms are normally submitted to correction of sequence errors and read removal using quality filter in order to increase the accuracy and prevent errors during sequence assembling [3,4]. The effects of bad quality reads have already been addressed, showing the importance of removing bad quality sequences before processing the data [4].

In genome assembling, the different types of quality filters affect the proper representation of coding sequences due to the change in sequencing coverage [5]. Considering that the sequencing depth is used to measure the gene expression on cDNA sequencing (RNA-seq), the removal of reads through the use of quality filters may affect the gene transcriptional profile. In this work, we evaluate the effects of applying different quality filters (Phred) in gene expression analyses of *Corynebacterium pseudotuberculosis* 1002.

The rRNA was extracted from *Corynebacterium pseudotubeculosis* 1002 under four conditions: control, heat shock (50°C), osmotic stress (2M NaCl) and acidic stress (pH). The cDNA was then synthesized, and sequenced in SOLID platform. The generated data was submitted to three different quality filters (QV10, QV15 and QV20) and also analyzed in a standard condition, without quality filter (QV0). After

this step, the data was submitted to a pipeline with cufflinks and tophat [6] and the differential expression of the different datasets was evaluated.

As a result, we have seen a significant decrease in data amount above QV20, mainly in control and heat shock condition, where 47.9% and 39.1% of the reads were removed, respectively. On the other hand, QV20 was the most congruent filter in terms of read coverage in all evaluated conditions (Figure 1).

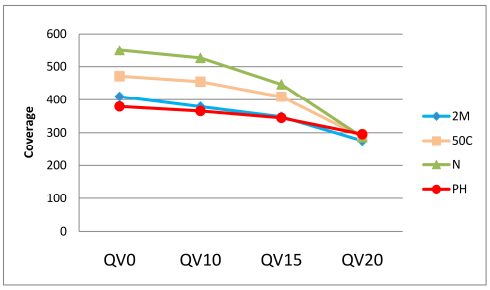

Fig. 1. Fig. 1. Evaluation of sequencing coverage for the genome of *C. pseudotuberculosis* 1002, using cDNA sequencing data from four conditions and four quality filters. N, Control condition; 2M, osmotic stress; 50, heat shock; pH, acidic stress; QV0, no filter; QV10, Phred 10; QV15, Phred 15; and QV20, Phred20.

The datasets were also submitted to RNA-seq analyses in order to evaluate the differential expression in each stress condition compared to the control condition. Although QV20 had the lowest number of sequencing reads (Figure 1), it presented the highest amount of exclusive differentially expressed genes (Figure 2).

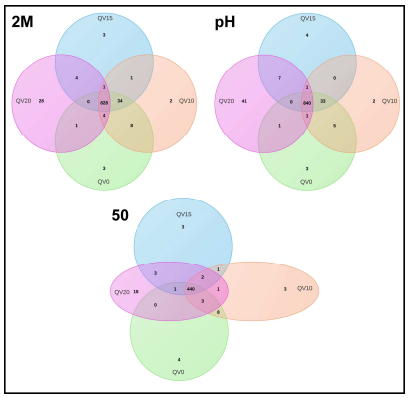

Fig. 2. Venn Diagram showing the exclusive and shared genes of the four quality filters under the stress conditions. 2M, osmotic stress; 50, heat shock; pH, acidic stress; QV0, no filter; QV10, Phred 10; QV15, Phred 15; and QV20, Phred20.

The number of genes differentially expressed between the heat shock, osmotic and acidic stress, submitted to the same quality filter, has not varied significantly. The major difference was observed in the acidic stress condition, which present 285, 283, 285 and 294 differentially expressed genes in QV0, QV10, QV15 and QV20, respectively. However, although the number of genes was very similar in all filters, the resulting composition of genes in the dataset was different in each filter.

Regarding differentially expressed genes, we observed that important genes related to survival of the bacteria in stress conditions have varied for the different filters, which points for the influence of the filter in identification of differentially expressed genes. The gene sigE, for instance, was considered as differentially in osmotic stress under QV0, QV10 and QV20, but did not appear under the quality filter QV15. On the other hand, sigC was only considered as differentially expressed under QV20. Interestingly, sigE was the first extracitoplasmatic function sigma (sigma ECF) identified in Escherichia coli, which was also the second sigma factor recognized to play a role in heat shock stress in this organism [7]. Besides, under osmotic stress the same profile was also found for the gene groEL, which was detected as differentially expressed under QV15 and QV20, but was disregarded under QV0 and QV10. This gene codes for a chaperon, whose function is to protect and assist other proteins in three-dimensional folding under normal and stress conditions. Those results point for the real need of using assaying different quality filter, which have been shown to directly influence data interpretation of differentially expressed genes.

From the list with exclusively expressed genes in different quality filters, we have identified 16, 12 and 4 genes under acidic, osmotic and heat shock conditions, respectively, which are harbored inside pathogencicty islands (PAIs) of *C. pseudotuberculosis* 1002. In acidic stress, we have found the following CDSs inside PAIs: *ciuE*, Cp1002_0058, Cp1002_0064, Cp1002_0554, Cp1002_0980, Cp1002_1925, Cp1002_1932, *fadF*, *cmtR*, Cp1002_1556, Cp1002_1867, Cp1002_1877, *mmpL11*, *potA*, *senX3* and *ureE*. In osmotic stress, we have found *ciuC*, Cp1002_0058, Cp1002_0554, Cp1002_0573, *crtB*, *pmmB*, *cas5*, Cp1002_0059, Cp1002_1694, Cp1002_1911, *pdxS* and *ureG* inside PAIs. And, in heat shock condition, the following genes from PAIs were differentially expressed: *ansP*, Cp1002_1631, Cp1002_1867 and *opuBA*.

The *Corynebacterium* iron uptake proteins (*ciuE* and ciuC) and the product of *fadF* all code for iron ABC-type transporters, which are higly common inside PAIs due to their role in surviving inside the host under different kinds of stress. Interestingly, potA codes for an ATP-binding protein, which plays a role in importing Spermidine/putrescine from the extracellular medium. Spermidine and putrescine are polyamines, which are conversely required for optimal growth of eukaryotic and prokaryotic cells and the levels of ornithine decarboxylase and polyamines are normally higher in rapidly growing cells [8,9,10]. Studies have already shown involvement of polyamines at the end of G1 phase, at the initiation of DNA synthesis and also affecting the rate of movement of the DNA replication fork [11]. Also, it was proposed an ability of spermine and spermidine to precipitate DNA in order to protect DNA from denaturation by heat [8], [9], [11]. Besides DNA synthesis and protection, intracellular levels of polyamines and of ornithine decarboxylase are normally increased during

stimulation of RNA synthesis [9,12]; polyamines can bind to specific sites on the tRNA molecule; and polyamines also affect the overall rate of protein synthesis, the fidelity of the translation by binding to ribosomes, and other steps [12,14].
In view of the pivotal role of polyamines in surviving inside the host and the presence of two potA genes in the genome, one of which inside a pathogenicity island, we postulated that the potA gene from the pathogenicity island was recently acquired and maintained due to its possible role in coping with different stresses, mainly during the intra-macrophagic life, explaining the differential expression during acidic stress.
The rRNA depletion and the sequencing coverage required in order to represent the transcriptome are highly important for a good efficiency of the RNA-seq approach [15,16]. However, the impact of quality filter stringency applied in pre-processing data step was never evaluated for RNA-Seq data. This work has showed the influence of applying different quality filters to RNA-Seq data, mainly in evaluating the differential expression. Finally, we also observed that, even with highly stringent filters, like Phred 20, the balance between sequencing coverage and high quality reads may increase the reliability of the analyses.

1. Martin J a & Wang Z. Next-generation transcriptome assembly. Nature reviews. Genetics 12, 671–682 (2011).
2. Mutz, K.-O., Heilkenbrinker, A., Lönne, M., Walter, J.-G., Stahl, F.. Transcriptome analysis using next-generation sequencing. Current opinion in biotechnology 24, 22–30 (2013).
3. Loman NJ, Misra R V, Dallman TJ, Constantinidou C, Gharbia SE, Wain J & Pallen MJ. Performance comparison of benchtop high-throughput sequencing platforms.Nature Biotechnology 30, 434-439 (2012).
4. Li H & Homer N. A survey of sequence alignment algorithms for next-generation sequencing. Briefings in bioinformatics 11,473–483 (2010).
5. Carneiro AR, Ramos RTJ, Barbosa HPM, Schneider MPC, Barh D, Azevedo V & Silva A. Quality of prokaryote genome assembly: indispensable issues of factors affecting prokaryote genome assembly quality. Gene 505, 365–367 (2012).
6. Trapnell C et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nature protocols 7, 562–578 (2012).
7. Wang, Q. P., and J. M. Kaguni. A novel sigma factor is involved in expression of the rpoH gene of Escherichia coli. J. Bacteriol. 171:4248–4253 (1989).
8. Tabor CW & Tabor H. Polyamines. Annual Review of Biochemistry 53: 749-790 (1984).
9. Tabor CW & Tabor H. 1,4-Diaminobutane (Putrescine), Spermidine, and Spermine. Annual Review of Biochemistry 45: 285-306 (1976).
10. Karvonen E, Kauppinen LT Partanen & Pösö H Irreversible inhibition of putrescine-stimulated S-adenosyl-L-methionine decarboxylase by berenil and pentamidine. Biochemical Journal 231,165-9 (1985).
11. Tabor H & Tabor CW. Biosynthesis and metabolism of 1,4-diaminobutane, spermidine, spermine, and related amines. Adv Enzymol Relat Areas Mol Biol 36, 203-268 (1972).
12. Caldarera CM, Casti A, Guarnieri C & Moruzzi G. Regulation of ribonucleic acid synthesis by polyamines. Reversal by spermine of inhibition by methylglyoxal bis(guanylhydrazone) of ribonucleic acid synthesis and histone acetylation in rabbit heart. Biochem J 152, 91-8 (1975).

13. Igarashi K, Hashimoto S, Miyake A, Kashiwagi K & Hirose S. Increase of Fidelity of Polypeptide Synthesis by Spermidine in Eukaryotic Cell-Free Systems. 128(2-3): 597–604 (1982).

14. Kurland CG. Translational accuracy in vitro. Cell 28(2):201-202 (1982).
15. Haas BJ, Chin M, Nusbaum C, Birren BW & Livny J. How deep is deep enough for RNA-Seq profiling of bacterial transcriptomes? BMC genomics 13, 734 (2012).
16. Wang Y, Ghaffari N, Johnson CD, Braga-Neto UM, Wang H, Chen R & Zhou H. Evaluation of the coverage and depth of transcriptome by RNA-Seq in chickens. BMC bioinformatics 12 Suppl 10, S5 (2011).