Gene 563 (2015) 165–171

Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Research paper

The impact of quality filter for RNA-Seq

Pablo H.C.G. de Sá [a], Adonney A.O. Veras [a], Adriana R. Carneiro [a], Kenny C. Pinheiro [a], Anne C. Pinto [b], Siomar C. Soares [b], Maria P.C. Schneider [a], Vasco Azevedo [b], Artur Silva [a], Rommel T.J. Ramos [a,*]

[a] Institute of Biological Sciences, Federal University Pará, Belém, Pará, Brazil
[b] Institute of Biological Sciences, Federal University Minas Gerais, Belo Horizonte, Minas Gerais, Brazil

ARTICLE INFO

Article history:
Received 22 November 2014
Received in revised form 6 March 2015
Accepted 13 March 2015
Available online 18 March 2015

Keywords:
RNA-Seq
Quality filter
Reference approach
Differential gene expression
Prokaryotes

ABSTRACT

Background: With the emergence of large-scale sequencing platforms since 2005, there has been a great revolution regarding methods for decoding DNA sequences, which have also affected quantitative and qualitative gene expression analyses through the RNA-Sequencing technique. However, issues related to the amount of data required for the analyses have been considered because they affect the reliability of the experiments. Thus, RNA depletion during sample preparation may influence the results. Moreover, because data produced by these platforms show variations in quality, quality filters are often used to remove sequences likely to contain errors to increase the accuracy of the results. However, when reads of quality filters are removed, the expression profile in RNA-Seq experiments may be influenced.
Result: The present study aimed to analyze the impact of different quality filter values for *Corynebacterium pseudotuberculosis* (sequenced by SOLiD platform), *Microcystis aeruginosa* and *Kineococcus radiotolerans* (sequenced by Illumina platform) RNA-Seq data. Although up to 47.9% of the reads produced by the SOLiD technology were removed after the QV20 quality filter is applied, and 15.85% were removed from *K. radiotolerans* data set using the QV30 filter, Illumina data showed the largest number of unique differentially expressed genes after applying the most stringent filter (QV30), with 69 genes. In contrast, for SOLiD, the acid stress condition with the QV20 filter yielded only 41 unique differentially expressed genes. Even for the highest quality *M. aeruginosa* data, the quality filter affected the expression profile. The most stringent quality filter generated a greater number of unique differentially expressed genes: 9 for high molecular weight dissolved organic matter condition and 12 for low P conditions.
Conclusion: Even high-accuracy sequencing technologies are subject to the influence of quality filters when evaluating RNA-Seq data using the reference approach.

1. Introduction

Since 2005, genomic data collection has greatly increased due to the emergence of second-generation sequencing platforms that are able to produce large volumes of data at low costs, which boosted the number of complete genome sequencing projects. Among these platforms, SOLiD and Illumina have been widely used for gene expression studies. The advantages of those two platforms over the previous generation of sequencing machines include the high data quality, especially for Illumina, and the throughput, with approximately 600 Mb in the HiSeq 2000 v3 platform and 155 Mb with the SOLiD technology (Pinto et al., 2014; Wang et al., 2014; Isabella and Clark, 2011; Scholz et al., 2012; Leggett et al., 2013). This output level differentiates these methods even from third generation sequencers, which use the Single Molecule Real Time (SMRT) DNA decoding method; examples include the PacBio, Helicos and Oxford Nanopore sequencers as well as the more recent base detection through pH variation (Ion Torrent PGM and Ion Proton). However with the increasing throughput of these technologies, previous-generation platforms will most likely be substituted (Schuster, 2008; Lam et al., 2012).

High-throughput platforms also improved transcriptomics analyses through the RNA-Seq technique (Martin and Wang, 2011). RNA-Seq is an approach that enables an assessment of the transcriptional profile of complete organisms without requiring previous knowledge of the sequences, as the Real Time Polymerase Chain Reaction (RT-PCR) and microarray technique do. It also allows the identification of new transcripts and genome annotation corrections (Mutz et al., 2013). Currently, there are studies that aim to determine the amount of RNA-Seq data required to evaluate expression in bacteria (Haas et al., 2012; Liu et al.,

Abbreviations: cDNA, DNA complementary to RNA.

* Corresponding author.
E-mail addresses: pablogomesdesa@gmail.com (P.H.C.G. de Sá), allanverasce@gmail.com (A.A.O. Veras), carneiroar@gmail.com (A.R. Carneiro), kennybiotec@gmail.com (K.C. Pinheiro), acybelle@gmail.com (A.C. Pinto), siomars@gmail.com (S.C. Soares), mariapaulacruzschneider@gmail.com (M.P.C. Schneider), vascoariston@gmail.com (V. Azevedo), arturluizdasilva@gmail.com (A. Silva), rommelthiago@gmail.com (R.T.J. Ramos).

http://dx.doi.org/10.1016/j.gene.2015.03.033
0378-1119/

Table 1
Summary of data obtained for the conditions and quality filter evaluated for *Corynebacterium pseudotuberculosis*.
Evaluation of the QV10, QV15, QV20 and QV0 quality filters regarding the number of reads, the percentage decrease in the number of reads, the number of bases and the sequencing coverage reached for *C. pseudotuberculosis*.

Condition	Filter	Reads	Decrease	Bases	Coverage
Control	QV0	25,235,478	–	1,287,009,378	643.50
	QV10	24,142,299	4.4%	1,231,257,249	615.63
	QV15	20,497,865	18.8%	1,045,391,115	522.70
	QV20	13,148,231	47.9%	670,559,781	335.28
2 M	QV0	18,783,810	–	957,974,310	478.99
	QV10	17,371,379	7.6%	885,940,329	442.97
	QV15	15,963,699	15.1%	814,148,649	407.07
	QV20	12,657,661	32.7%	645,540,711	322.77
50 °C	QV0	21,622,844	–	1,102,765,044	551.38
	QV10	20,874,392	3.5%	1,064,593,992	532.30
	QV15	18,770,148	13.2%	957,277,548	478.64
	QV20	13,177,122	39.1%	672,033,222	336.02
pH	QV0	17,393,077	–	887,046,927	443.52
	QV10	16,758,069	3.7%	854,661,519	427.33
	QV15	15,821,587	9.1%	806,900,937	403.45
	QV20	13,567,760	22%	691,955,760	345.98

Table 2
Summary of data obtained for the conditions and quality filter evaluated for *Kineococcus radiotolerans*.
Evaluation of the effects of the application of the QV10, QV15, QV20, QV25, QV30 and QV0 quality filters regarding the number of reads, the reduction percentage, the number of bases and the sequencing coverage reached for *K. radiotolerans*.

Condition	Filter	Reads	Decrease	Bases	Coverage
Control	QV0	28,508,412	–	2,907,858,024	642.07
	QV10	28,165,539	1.20	2,872,884,978	634.35
	QV15	27,769,423	2.59	2,832,481,146	625.43
	QV20	27,264,904	4.36	2,781,020,208	614.07
	QV25	26,327,539	7.65	2,685,408,978	592.96
	QV30	23,992,240	15.84	2,447,208,480	540.36
Radioactive	QV0	36,145,816	–	3,686,873,232	814.09
	QV10	35,718,661	1.18	3,643,303,422	804.47
	QV15	35,215,269	2.57	3,591,957,438	793.13
	QV20	34,579,294	4.33	3,527,087,988	778.81
	QV25	33,410,613	7.57	3,407,882,526	752.48
	QV30	30,463,631	15.72	3,107,290,362	686.11

2014). However, RNA depletion method has a great influence on the throughput of the experiments and impacts transcript representation (Castro et al., 2013).

Data produced by high-throughput platforms are often submitted to sequencing error correction and quality filters to increase the accuracy and to prevent errors in sequence assembly and in analyses, such as SNP calling and variant calling (Li and Homer, 2010; Loman et al., 2012). The effects of poor data quality on sequence alignment have already been discussed and showed the importance of removing low quality sequences before data processing (Li and Homer, 2010).

The effects of quality filters on coding sequence representation when conducting genome assembly were analyzed, instead of simply evaluating the results just with statistical metrics such as N50 and the longest and shortest sequences. This approach demonstrated that an increase in the quality of sequences increases the accuracy of the contigs produced during assembly (Carneiro et al., 2012).

The *de novo* representation of transcripts from RNA-Seq also requires processing steps such as quality filtering and trimming to improve the throughput of the assemblies (Mbandi et al., 2014), even when these processes decrease the number of analyzed reads. However, this approach influences the gene expression evaluation in cDNA sequencing (RNA-Seq), modifying the transcriptional profile while increasing the accuracy of mapping (Li and Homer, 2010).

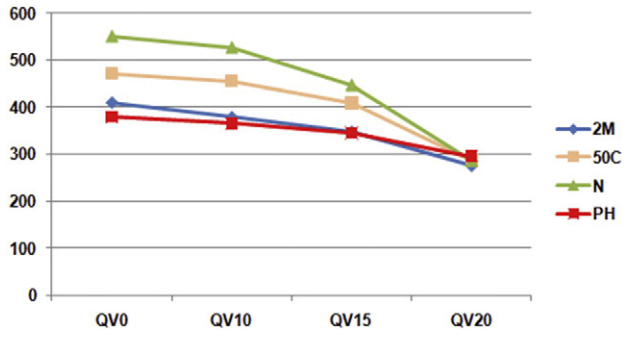

Fig. 1. Depth coverage after applying PHRED quality filter for *Corynebacterium pseudotuberculosis*. Sequencing coverage of *C. pseudotuberculosis* 1002 regarding the quality filter applied (QV0, QV10, QV15 and QV20) for the control (N) and osmotic (2 M), heat (50 °C) and acid (pH) stress conditions.

Thus, the present study evaluates the effects of different quality filters (PHRED) on the results of gene expression from reference-based RNA-Seq using transcriptome data from three prokaryotes, separately: *Corynebacterium pseudotuberculosis*, *Microcystis aeruginosa* and *Kineococcus radiotolerans*, obtained on different high-throughput sequencing platforms.

2. Methods

2.1. RNA-Seq and reference data

For *C. pseudotuberculosis* the genome of the strain 1002 (*C. pseudotuberculosis* 1002) was used as reference (GenBank: NC_017300). This bacterium was grown under osmotic (2 M), acid (pH) and heat (50 °C) stress and under control (Normal) conditions, which simulate the conditions faced by the bacteria during the infectious process. Subsequently, cDNA of each condition was sequenced by the SOLiD 3 Plus platform using the RNA-Seq technique. The sequencing data are available in the European Bioinformatics Institute (EBI) repository under accession number E-MTAB-2017 (Pinto et al., 2012).

RNA-Seq data regarding the response of *K. radiotolerans* to ionizing radiation (SRA: SRX357580) and data from the same strain without inducing radiation (control condition, SRA: SRX357579) were used, both available in the Sequence Read Archive database (SRA) and sequenced by the Illumina HiSeq 2000 platform. For this data the genome of *K. radiotolerans* strain SRS30216 (ATCC BAA-149) was used as reference (GenBank: NC_009664).

M. aeruginosa experiment was previously performed and data is available at SRA, as described below. Briefly, *M. aeruginosa* LE-3 was grown under low P condition (LowP), high molecular weight dissolved organic matter condition (HMWDOM) and control condition (control). Three replicates were made for each condition, which were sequenced by Illumina HiSeq 2000 using a RNA-Seq protocol described in SRA database.

For the LowP condition, the accession numbers of the data in SRA are LowP_R1 (SRA: SRX271690), LowP_R2 (SRA: SRX271784) and LowP_R3 (SRA: SRX271864). For HMWDOM condition, the SRA accession numbers are HMWDOM_R1 (SRA: SRX271903), HMWDOM_R2 (SRA: SRX271917) and HMWDOM_R3 (SRA: SRX271918). And for control condition, the SRA accession numbers are Control_R1 (SRA: SRX272017), Control_R2 (SRA: SRX272019) and Control_R3 (SRA: SRX272020). The reference genome of *M. aeruginosa* strain NIES-843

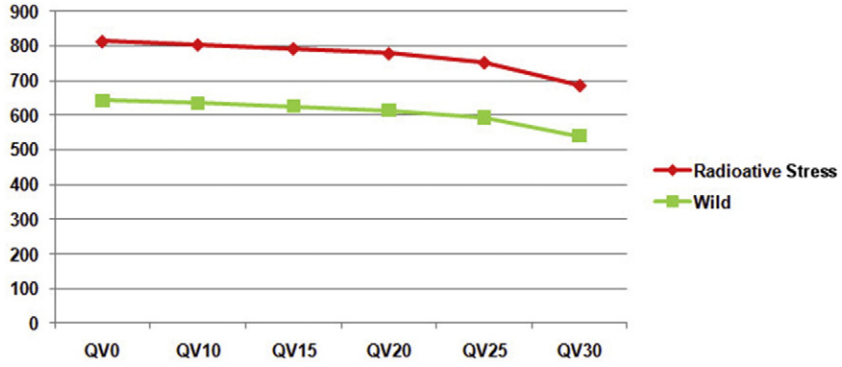

Fig. 2. Depth coverage after applying PHRED quality filter for *Kineococcus radiotolerans*. Sequencing coverage evaluation of the *K. radiotolerans* genome under radioactive stress and control conditions (Wild).

used here is available in NCBI under the accession number (GenBank: NC_010296).

The reference genome of each organism was used in FASTA format and the annotation file in the GTF format, so each organism had two reference files, one was the sequence FASTA and the other consisted of the genes annotated in the GTF format.

The details about the experiments performed to produce the RNA-Seq data above can be accessed through the EBI and SRA numbers cited previously.

2.2. Data processing

The reads obtained for the three organisms were submitted to the quality filtering step through the Quality Assessment (Ramos et al., 2011) program, using different PHRED filter values. Thus, four data sets were created for *C. pseudotuberculosis* according to the PHRED quality values (QV) using QV10, QV15 and QV20 as filters, in addition to QV0, which represents original data with no filters applied. For *K. radiotolerans*, the same filters already mentioned were used in addition to QV25 and QV30, and for *M. aeruginosa* the QV0, QV40, QV50 and QV60 data sets were generated due to the high quality of data produced by the sequencing platform Illumina HiSeq 2000, which allowed the application of more stringent quality filters.

2.3. Read mapping and gene expression

The TopHat program was used to map the reads of each data set. All reads were mapped only against their respective reference genome using TopHat default options and the parameter "–no-novel-juncs" in order to map reads only to coding regions defined by the gene annotation file, GTF file, of each organism. As a result, a .bam file was created containing only the mapped reads as output. Each data set obtained with the different quality filters was submitted to the Cuffdiff program for gene expression (FPKM) and differential expression analysis, which allowed the evaluation of the gene expression profiling and identification of differentially expressed genes of each different data set for each organism (Trapnell et al., 2012).

3. Results and discussion

3.1. Sequencing coverage

After applying the quality filter to the *C. pseudotuberculosis* 1002 data corresponding to the osmotic (2 M), heat shock (50 °C), acid (pH) stress and control conditions, there was a significant decrease in the number of reads when filter QV20 was applied, especially for the control and heat shock stress samples, in which 47.9% and 39.1% of the reads were removed, respectively (Table 1).

When evaluating the number of bases after applying the filters for all conditions, an increasing trend toward coverage approximation was observed as the quality filter increased, and the closest point was reached with the QV20 quality filter (Fig. 1). This fact corroborates data from other studies related to sequencing with SOLiD, in which PHRED 20 quality filters were applied (Carneiro et al., 2012).

Because data from *K. radiotolerans* were of high quality in the control and radioactive stress conditions (data quality panel), the QV0, QV10, QV15, QV20, QV25 and QV30 filters were used (Table 2). In both conditions, the percentage decrease due to the quality filter was similar. The greatest loss (15.84% of the reads) occurred when the QV30 quality filter was used on the data from the control condition (Fig. 2), although this reduction was not significant when considering the volume of data produced by the sequencing platform.

The difference between the coverage levels observed for each quality filter in different *K. radiotolerans* samples reduced as the filter quality increased and reached a minimum value of 145 X with QV30, a trend also observed for data from *C. pseudotuberculosis*.

The data of *M. aeruginosa* presented higher quality than the one from *K. radiotolerans*, and the filters QV0, QV40, QV50 and QV60 were applied for all replicates of each condition (Table 3). The proportion of removed reads was approximately the same in all replicates, and the control condition showed the greatest reduction after applying the QV60 filter (Fig. 3). For both organisms, the amount of reads decreased due to the application of high quality filters.

The RNA-Seq data used has the coverage required to RNA-Seq analysis as suggested (Haas et al., 2012), for all organisms and in all filters. To increase the confidence about our hypothesis we used an Illumina data (high quality reads) using replicated samples.

3.2. Transcript mapping

When evaluating the mapping of *C. pseudotuberculosis*, *K. radiotolerans* and *M. aeruginosa* reads compared with reference values (QV0), the greatest difference in the number of mapped reads occurred for *C. pseudotuberculosis* in the control condition with the QV0 and QV20 filters, when 25,075,687 and 13,130,035 reads were aligned, respectively. This result corresponded to a 47.63% decrease (11,945,652 reads) in the number of mapped sequences with the QV20 filter (Table 4), especially as a result of the decrease in the number of sequences in the quality filter step.

For *K. radiotolerans*, the greatest difference regarding the number of reads occurred in the radiation condition between filters QV0 and QV30, where a difference of 5,676,879 fewer reads (15.70%) was observed for

Table 3
Summary of data obtained for the conditions and quality filter evaluated for *Microcystis aeruginosa*.
Evaluation of the effects of the application of the QV40, QV50, QV60 and QV0 quality filters regarding the number of reads, the reduction percentage, the number of bases and the sequencing coverage reached for *M. aeruginosa* for each replicate.

Condition	Filter	R1_reads	R1_decrease	R1_bases	R1_coverage	R2_reads	R2_decrease	R2_bases	R2_coverage	R3_reads	R3_decrease	R3_bases	R3_coverage
Control	QV0	13,055,201	–	1,318,575,301	225,68	12,967,695	–	1,309,737,195	224,16	13,277,124	–	1,340,989,524	229,51
	QV40	13,029,658	0%	1,315,995,458	225,23	12,941,434	0%	1,307,084,834	223,71	13,252,062	0%	1,338,458,262	229,08
	QV50	12,337,033	6%	1,246,040,333	213,26	12,227,719	6%	1,234,999,619	211,37	12,593,670	5%	1,271,960,670	217,70
	QV60	10,150,010	22%	1,025,151,010	175,46	10,159,790	22%	1,026,138,790	175,62	10,562,991	20%	1,066,862,091	182,59
HWMDOM	QV0	13,993,702	–	1,413,363,902	241,90	12,393,729	–	1,251,766,629	214,24	15,095,239	–	1,524,619,139	260,94
	QV40	13,968,553	0%	1,410,823,853	241,46	12,357,403	0%	1,248,097,703	213,61	15,051,893	0%	1,520,241,193	260,19
	QV50	13,295,206	5%	1,342,815,806	229,82	11,948,377	4%	1,206,786,077	206,54	14,562,040	4%	1,470,766,040	251,72
	QV60	11,253,887	20%	1,136,642,587	194,54	9,920,046	20%	1,001,924,646	171,48	12,142,656	20%	1,226,408,256	209,90
LowP	QV0	17,363,994	–	1,753,763,394	300,16	20,184,291	–	2,038,613,391	348,91	17,938,546	–	1,811,793,146	310,09
	QV40	17,332,140	0%	1,750,546,140	299,61	20,150,314	0%	2,035,181,714	348,32	17,906,249	0%	1,808,531,149	309,53
	QV50	16,702,374	4%	1,686,939,774	288,72	19,470,442	4%	1,966,514,642	336,57	17,288,957	4%	1,746,184,657	298,86
	QV60	14,005,426	19%	1,414,548,026	242,10	16,456,513	18%	1,662,107,813	284,47	14,535,837	19%	1,468,119,537	251,27

filter QV30 (Table 5). For the control condition data, the difference between the same quality filters was 4,511,942 reads, which represented a 15.83% decrease.

After mapping the reads of *M. aeruginosa*, individually for each condition, the decrease of mapped reads in different quality filters was small, in general, for all replicates, even when stringent filters were used. The condition LowP_R1 showed the greatest difference for filters QV0 (5,150,359 mapped reads) and QV60 (4,890,800 mapped reads) where 259,559 reads, approximately (5.04%), were not mapped in QV60 (Table 6). This result demonstrates the high quality of Illumina data compared to data produced by the SOLiD platform.

3.3. Differential expression

After applying different quality filters (QV0, QV10, QV15 and QV20) for *C. pseudotuberculosis*, the data were submitted to RNA-Seq analyses to evaluate the differential expression for each stress condition compared to the control condition. Thus, for every stress condition, the QV20 filter showed the greatest number of differentially expressed genes exclusives in that condition (Fig. 4), even though it eliminated the greatest number of reads (Fig. 1).

When evaluating differentially expressed genes in the samples from the 2 M, pH and 50 °C conditions, which were filtered with the same quality parameters, the number of differentially expressed genes decreased as the quality filter increased, except in the data from the pH condition.

The differential expression analysis for the *K. radiotolerans* data was performed only for filters QV0, QV15, QV20 and QV30 due to the low variation in the intermediate filters QV10 and QV25 compared to the other filters. Filter QV30 resulted in the lowest number of differentially expressed genes for *K. radiotolerans* (Table 7), and the same result was observed for most of the stress conditions for *C. pseudotuberculosis* when the filter QV20 was applied.

Nonetheless, for Illumina data, filter QV30 resulted in the greatest number of unique differentially expressed genes in *K. radiotolerans*, 69 (Fig. 5). Thus, for SOLiD data, the most stringent filter (QV20) also produced the greatest numbers of unique genes for all conditions evaluated in *C. pseudotuberculosis* (Fig. 4).

Even for the high quality data of *M. aeruginosa*, the high quality values used in the filter reduced the amount of differentially expressed genes for LowP and HWMDOM conditions (Table 7). The filter QV60 presented the greatest numbers of unique differentially expressed genes: 12 and 9 genes for HWMDOM and LowP respectively (Fig. 6).

Although data used in the present study were obtained for different organisms (*C. pseudotuberculosis*, *K. radiotolerans* and *M. aeruginosa*) that have specific characteristics, such as guanine–cytosine (GC) content and the presence of mobile elements (Phillips and Wiegel, 2002; Orellaa et al., 2006), in addition to having been sequenced with different sequencing technologies, when the quality filter QV20 was applied for *C. pseudotuberculosis*, the number of differentially expressed genes decreased in 2 of the 3 conditions compared with QV0. For *K. radiotolerans* and *M. aeruginosa* this number decreased as the quality filter increased, except for *M. aeruginosa* condition HWMDOM filter QV50 which presented a slight increase in the number of genes (Table 7). Regardless of these discrepancies, considering that filtered data are more reliable, the mapping of these data showed the highest accuracy, which tends to produce better results (Li and Homer, 2010).

As demonstrated in the present study, the quality filter for the transcriptome analyses triggers variations in gene expression profiles, changing the characterization of the expression profile and demonstrating the impact of quality filters on RNA-Seq based-reference analyses.

Even though the decrease in the number of reads mapped occurred for *C. pseudotuberculosis*, *K. Radiotolerans* and *M. aeruginosa* after the

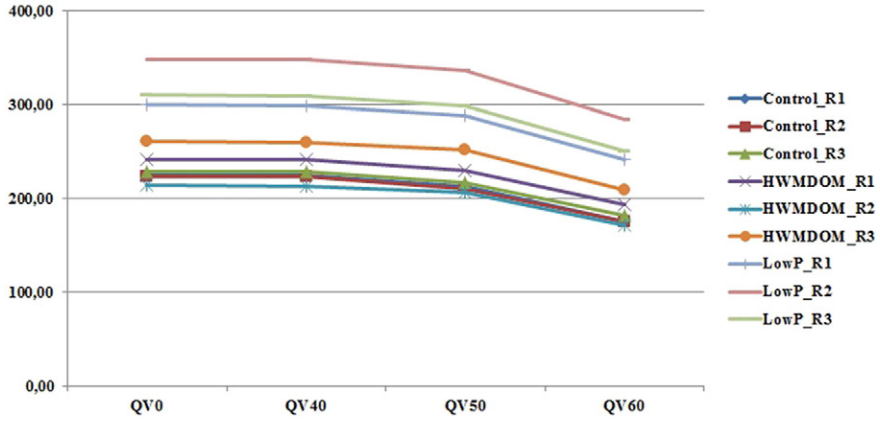

Fig. 3. Depth coverage after applying PHRED quality filter for *Microcystis aeruginosa*. Sequencing coverage evaluation for each replicate of the *M. aeruginosa* genome under LowP, HWMDOM and control conditions.

Table 4
Amount of reads mapped against the reference by condition and quality filter for *Corynebacterium pseudotuberculosis*.
Total mapped reads compared with reference values for conditions evaluated in *C. pseudotuberculosis* 1002 after quality filters QV0, QV10, QV15 and QV20 were applied for each condition.

Mapped reads	Control	2 M	50 °C	pH
QV0	25,075,687	18,652,648	21,484,367	17,219,895
QV10	24,025,238	17,275,206	20,770,240	16,615,718
QV15	20,439,750	15,914,660	18,714,371	15,730,533
QV20	13,130,035	12,645,792	13,159,534	13,527,289

Table 5
Amount of reads mapped against the reference by condition and quality filter for *Kineococcus radiotolerans*.
Total mapped reads compared with reference values for conditions evaluated in *K. radiotolerans* under quality filters QV0, QV10, QV15, QV20, QV25 and QV30.

Mapped reads	Control	Radioactive induced
QV0	28,500,467	36,135,864
QV10	28,158,912	35,710,371
QV15	27,763,266	35,207,564
QV20	27,259,203	34,572,184
QV25	26,322,511	33,404,295
QV30	23,988,525	30,458,985

QV20, QV30 and QV60 quality filters, respectively, all organisms presented the greatest number of unique differentially expressed genes with these filters, which shows that even a decrease considered insignificant regarding amount of reads, like that occurred in *K. Radiotolerans* and *M. aeruginosa*, has a great influence on the differential expression analyses.

Thus, there is a need to apply more stringent quality filters to increase the reliability of gene expression analyses and to prevent errors due to low quality sequences, especially considering the data quality and the specific errors of each platform (Scholz et al., 2012) as well as the number of sequences required for the analyses (Haas et al., 2012). In addition, when reference-based RNA-Seq is used, the mapping criteria of reads may influence the results, and the raw data evaluation step becomes essential to define the most adequate strategy.

With the increase in the quality of RNA-Seq data and the use of replicates, it is possible that in the future, the next-generation sequencing (NGS) approach will have the same reliability as RT-PCR experiments, which are still widely used to validate results because they represent a gold standard for gene expression evaluation.

4. Conclusions

Minimal changes in the quantity of RNA-Seq data may affect differential expression results. Thus, in addition to considering the throughput required for the expression experiment (Haas et al., 2012), data processing should be performed to make analyses more reliable.

Despite the high quality of data obtained through the HiSeq platform, which resulted in a less significant coverage decrease after the quality filter was applied compared to the SOLiD data, there was a large change in the expression profile, especially among data without quality filters and data with the QV30 and QV60 quality filter, which demonstrates the influence of the quality filter step on the results, even for high-performance and accurate platforms.

Finally, the quality filter parameters can change based on the quality of the reads produced by the sequencing platforms, then defining the best values is a complex task, but its importance was demonstrated, by providing evidence that quality filters affect the gene expression profile even for the high accuracy sequencing platforms.

Competing interests

The authors declare that they have no competing interests.

Table 6
Amount of reads mapped against the reference by condition and quality filter for *Microcystis aeruginosa*.
Total mapped reads compared with reference values for conditions evaluated in *M. aeruginosa* under quality filters QV0, QV40, QV50 and QV60.

Mapped reads	Control_R1	Control_R2	Control_R3	HWMDOM_R1	HWMDOM_R2	HWMDOM_R3	LowP_R1	LowP_R2	LowP_R3
QV0	3,781,504	3,601,794	3,712,980	3,626,370	3,285,552	2,807,519	5,150,359	6,265,458	5,248,897
QV40	3,781,420	3,601,714	3,712,887	3,626,298	3,285,445	2,807,447	5,150,221	6,265,273	5,248,747
QV50	3,773,491	3,593,990	3,705,081	3,619,136	3,276,920	2,801,762	5,134,557	6,247,354	5,234,048
QV60	3,640,436	3,469,049	3,574,584	3,500,063	3,136,524	2,696,559	4,890,800	5,961,314	4,999,736

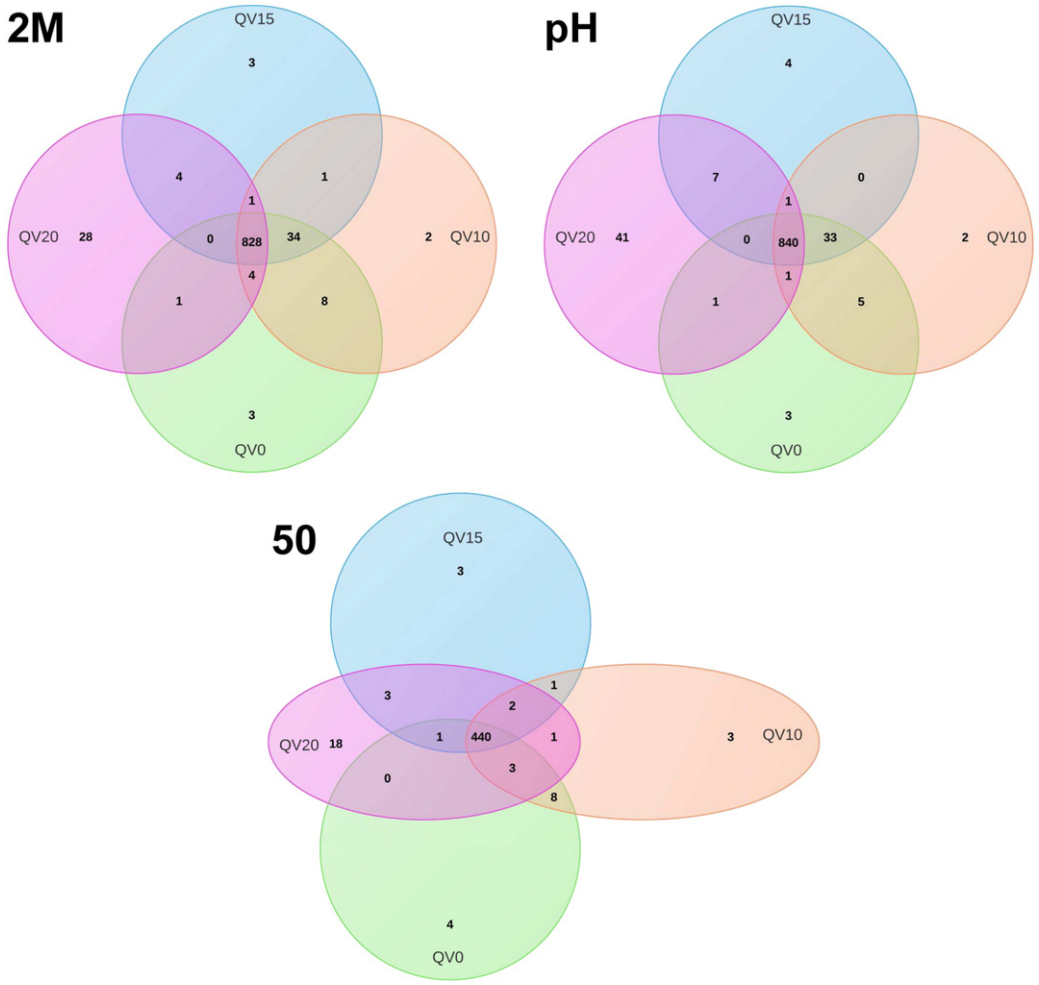

Fig. 4. Venn diagram of evaluation of the differential expression genes exclusive of *Corynebacterium pseudotuberculosis* for the quality filters. Venn diagram for the differential expression analysis of samples from stress conditions (50 °C, 2 M and pH) showing the number of unique differentially expressed genes and genes shared among the quality filters used.

Table 7
Summary of the amount of differential genes expressed by organism, condition and quality filter (PHRED).
Quantification of differentially expressed (DE) genes by stress condition and quality filter for *C. pseudotuberculosis* 1002, *K. radiotolerans* and *M. aeruginosa*.

Organism	Condition	Quality filter	DE genes
C. pseudotuberculosis	Osmotic (2 M)	0	878
	Osmotic (2 M)	10	878
	Osmotic (2 M)	15	871
	Osmotic (2 M)	20	866
	Thermal (50 °C)	0	489
	Thermal (50 °C)	10	491
	Thermal (50 °C)	15	483
	Thermal (50 °C)	20	468
	Acid (pH)	0	883
	Acid (pH)	10	882
	Acid (pH)	15	885
	Acid (pH)	20	891
K. radiotolerans	Radioactive	0	2386
	Radioactive	10	2381
	Radioactive	15	2380
	Radioactive	20	2360
	Radioactive	25	2330
	Radioactive	30	2194
M. aeruginosa	HWMDOM	0	956
	HWMDOM	40	955
	HWMDOM	50	958
	HWMDOM	60	944
	LowP	0	1363
	LowP	40	1361
	LowP	50	1354
	LowP	60	1346

Authors' contributions

PHGS, AAOV and KCP performed quality filters to generate the data sets, mapping of the data sets against the references and differential gene expression analyses. RTJR created the graphs from the differential gene expression results. PHGS, AAOV, VA, AS and RTJR wrote and revised the manuscript. ARC, ACP, SCS and MPCS revised the manuscript. VA, AS and RTJR supervised all aspects of the project. RTJR conceived of the study and participated in its design and coordination.

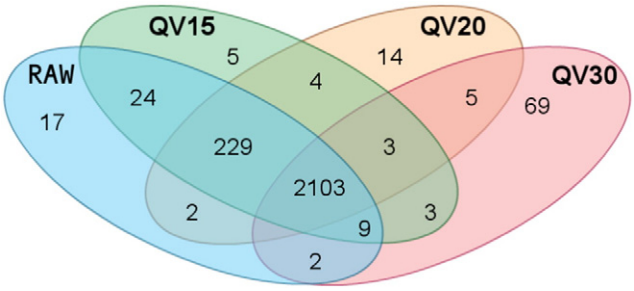

Fig. 5. Venn diagram of evaluation of the differential expression genes exclusive of *Kineococcus radiotolerans* for the quality filters. Venn diagram showing the differential expression analysis of the samples from the radioactive stress condition submitted to different quality filters. The data indicate the number of unique differentially expressed genes and genes shared by the quality filters.

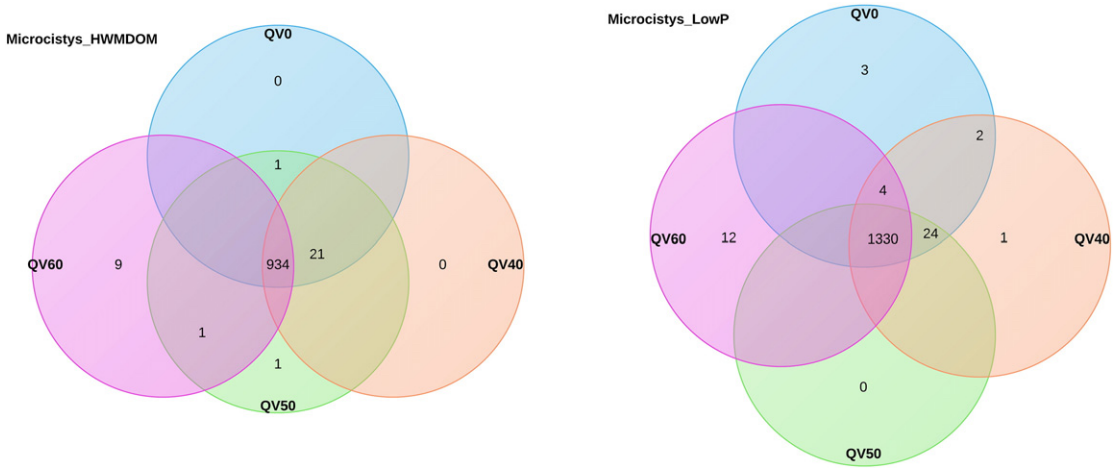

Fig. 6. Venn diagram of evaluation of the differential expression genes exclusive of *Microcystis aeruginosa* for the quality filters. Venn diagram showing the differential expression analysis of the samples from the LowP and HWMDOM condition submitted to different quality filters. The data indicate the number of unique differentially expressed genes and genes shared by the quality filters.

Acknowledgments

This work was part of the Rede Paraense de Genômica e Proteômica supported by Fundação de Amparo a Pesquisa do Estado do Pará. PHGS, AAOV, ARC, ACP, SCS, MPCS, VA, and AS were supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), and RTJR was supported by CNPq grant #482799/2013-7. KCP was supported by Fundação Amazônia Paraense de Amparo à pesquisa (FAPESPA).

References

Carneiro, A.R., Ramos, R.T.J., Barbosa, H.P.M., Schneider, M.P.C., Barh, D., Azevedo, V., Silva, A., 2012. Quality of prokaryote genome assembly: indispensable issues of factors affecting prokaryote genome assembly quality. Gene 505, 365–367.

Castro, T.L.P., Seyffert, N., Ramos, R.T.J., Barbosa, S., Carvalho, R.D.O., Pinto, A.C., Carneiro, A.R., Silva, W.M., Pacheco, L.G.C., Downson, C., Schneider, M.P.C., Miyoshi, A., Azevedo, V., Silva, A., 2013. Ion Torrent-based transcriptional assessment of a *Corynebacterium pseudotuberculosis* equi strain reveals denaturing high-performance liquid chromatography a promising rRNA depletion method. Microb. Biotechnol. 6, 168–177.

Haas, B.J., Chin, M., Nusbaum, C., Birren, B.W., Livny, J., 2012. How deep is deep enough for RNA-Seq profiling of bacterial transcriptomes? BMC Genomics 13, 734.

Isabella, V.M., Clark, V.L., 2011. Deep sequencing-based analysis of the anaerobic stimulon in *Neisseria gonorrhoeae*. BMC Genomics 12, 51.

Lam, H.Y.K., Clark, M.J., Chen, R., Chen, R., Natsoulis, G., O'Huallachain, M., Dewey, F.E., Habegger, L., Ashley, E.A., Gerstein, M.B., Butte, A.J., Ji, H.P., Snyder, M., 2012. Performance comparison of whole-genome sequencing platforms. Nat. Biotechnol. 30, 78–82.

Leggett, R.M., Ramirez-Gonzalez, R.H., Clavijo, B.J., Waite, D., Davey, R.P., 2013. Sequencing quality assessment tools to enable data-driven informatics for high throughput genomics. Front. Genet. 4, 288 (December).

Li, H., Homer, N., 2010. A survey of sequence alignment algorithms for next-generation sequencing. Brief. Bioinform. 11, 473–483.

Liu, Y., Zhou, J., White, K.P., 2014. RNA-seq differential expression studies: more sequence or more replication? Bioinformatics 30, 301–304.

Loman, N.J., Misra, R.V., Dallman, T.J., Constantinidou, C., Gharbia, S.E., Wain, J., Pallen, M.J., 2012. Performance comparison of benchtop high-throughput sequencing platforms. Nat. Biotechnol. 30, 434–439.

Martin, J.A., Wang, Z., 2011. Next-generation transcriptome assembly. Nat. Rev. Genet. 12, 671–682.

Mbandi, S.K., Hesse, U., Rees, D.J.G., Christoffels, A., 2014. A glance at quality score: implication for de novo transcriptome reconstruction of Illumina reads. Front. Genet. 5, 17 (February).

Mutz, K.-O., Heilkenbrinker, A., Lönne, M., Walter, J.-G., Stahl, F., 2013. Transcriptome analysis using next-generation sequencing. Curr. Opin. Biotechnol. 24, 22–30.

Orellaa, F.A.D., Gustavo, L., Achecoa, C.P., Liveirab, S.C.O., Iyoshia, A.M., Zevedoa, V.A., 2006. Review article *Corynebacterium pseudotuberculosis*: microbiology, biochemical properties, pathogenesis and molecular studies of virulence. 37, 201–218.

Phillips, R., Wiegel, J., 2002. *Kineococcus radiotolerans* sp. nov., a radiation-resistant, gram-positive bacterium. System 52, 933–938.

Pinto, A.C., Ramos, R.T.J., Silva, W.M., Rocha, F.S., Barbosa, S., Miyoshi, A., Schneider, M.P.C., Silva, A., Azevedo, V., 2012. The core stimulon of *Corynebacterium pseudotuberculosis* strain 1002 identified using ab initio methodologies. Integr. Biol. (Camb.) 4, 789–794.

Pinto, A.C., de Sá, P.H.C.G., Ramos, R.T.J., Barbosa, S., Barbosa, H.P.M., Ribeiro, A.C., Silva, W.M., Rocha, F.S., Santana, M.P., de Paula Castro, T.L., Miyoshi, A., Schneider, M.P.C., Silva, A., Azevedo, V., 2014. Differential transcriptional profile of *Corynebacterium pseudotuberculosis* in response to abiotic stresses. BMC Genomics 15, 14.

Ramos, R.T., Carneiro, A.R., Baumbach, J., Azevedo, V., Schneider, M.P., Silva, A., 2011. Analysis of quality raw data of second generation sequencers with Quality Assessment Software. BMC Res. Notes 4, 130.

Scholz, M.B., Lo, C.-C., Chain, P.S.G., 2012. Next generation sequencing and bioinformatic bottlenecks: the current state of metagenomic data analysis. Curr. Opin. Biotechnol. 23, 9–15.

Schuster, Stephan C., 2008. Next-generation sequencing transforms today's biology. 5, 16–18.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., Pachter, L., 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat. Protoc. 7, 562–578.

Wang, Y., Lupiani, B., Reddy, S.M., Lamont, S.J., Zhou, H., 2014. RNA-seq analysis revealed novel genes and signaling pathway associated with disease resistance to avian influenza virus infection in chickens. Poult. Sci. 93, 485–493.