TRANSLATIONAL RESEARCH
IN VETERINARY SCIENCE
Vol 2, No 1, 2019

DOI: http://dx.doi.org/10.12775/TRVS.2019.001

ISSN 2657-4489 (online)

Quality control assessment of the RNA-Seq data generated from liver and pituitary transcriptome of Hereford bulls using StrandNGS software

Chandra Shekhar Pareek[1,2*], Mateusz Sachajko[1,2], Adrian Szczepański[¥], Magdalena Buszewska-Forajta[¥], Katarzyna Żarczyńska[3], Przemysław Sobiech[3], Edyta Juszczuk-Kubiak[¥], Qaisar Shahzad[4], Yang Qing Lu[4], Magdalena Ogłuszka[5], Ewa Polawska[5], Mariusz Pierzchała[5]

[1]Centre for Modern Interdisciplinary Technologies, Nicolaus Copernicus University, Toruń, Poland
[2]Centre of Veterinary Sciences, Inter-University Centre of Veterinary Medicine, Nicolaus Copernicus University, Toruń, Poland
[3]Department and Clinic of Internal Diseases, Faculty of Veterinary Medicine, University of Warmia and Mazury in Olsztyn Poland
[4]State Key Laboratory for Conservation and Utilisation of Subtropical Agro-bioresources, Guangxi University, Nanning, Guangxi, 530004, China
[5]Institute of Genetics and Animal Breeding, Polish Academy of Sciences, Jastrzębiec, Poland
[¥]Voluntary Authors

***Corresponding author:**
Prof. zw. dr hab. Chandra S. Pareek, Head, Division of Functional Genomics in Biological and Biomedical Research, Centre for Modern Interdisciplinary Technologies, Nicolaus Copernicus University, ul. Wileńska 4, 87-100 Toruń, Poland
E-mail: pareekcs@umk.pl

Abstract. Background: Quality control (QC) assessment is the most critical step in the high-throughput RNA-seq data analysis to characterize the in-depth understanding of genome and transcriptome assembling to a given reference genome. It provides not only a quick insight into the RNA-seq data quality to allow early identification of good or bad RNA-seq data samples, but also to verify the alignment QC checks for further essential high-throughput bioinformatics analysis such as, identification of novel genetic variants, differentially expressed genes (DEGs), gene network and metabolic pathways.

Method: After isolation of total RNA from liver (n=15) and pituitary gland (n=15) tissues of young Hereford bulls, the pooled total RNA (n=30) were fragmented using GeneRead rRNA depletion kit (Qiagen, Hilden, Germany) and cDNA library preparation were preformed using ScriptSeq™ v2 RNA-Seq library preparation kit (Epicentre, illumina, USA), followed by high-throughput sequencing of combined liver and pituitary transcriptome using MiSeq reagent kit v2 (illumina, USA) to obtain high quality of paired-end RNA-seq reads of 251 base-pairs (bps). In this paper, the QC assessment of obtained RNA-seq raw data as well as post-alignment QC of processed RNA-seq data of combined liver and pituitary transcriptome (n=30) of Hereford bulls were performed using the strand NGS software v1.3 (Agilent; http://www.strand-ngs.com/) data analysis package. The reads were aligned with Bowtie using default settings against both Bull and Cow genome assembly.

Results: Using two runs of MiSeq platform, a total of over 60 million paired-end RNA-seq reads were successfully obtained and submitted to NCBI SRA resources (https://www.ncbi.nlm.nih.gov/sra?linkname=bioproject_sra_all&from_uid=312148). Library complexity plot results revealed 72.02% of duplicate reads with a low library complexity value of 0.28. The pre-alignment QC analysis of raw RNA-seq data revealed the sequence read lengths ranged from 35-251 bp size with more than 50% of all reads with length over 200bp and 10% of reads below 100bp.

Conclusion: By testing the RNA-seq methodology on Illumina platform, two MiSeq sequencing runs yielded significantly high quality of 30 million sequencing reads per single MiSeq run. Our initial pre-alignment and post-alignment analysis of RNA-seq data analysis revealed that mapping of the Hereford liver and pituitary gland transcriptome to reference *Bos taurus* genome was successfully performed, however, more than 50% of all reads with length over 200bp were recovered. Therefore, obtained results concludes that liver and pituitary transcriptome sequencing with rRNA depletion method is less effective than mRNA RNA-seq method.

Keywords: RNA-seq; NGS; quality control; *Bos Taurus*; cattle; liver; pituitary gland; Hereford; transcriptome; strandNGS.

Introduction

Advancement in next-generation genome sequencing (NGS) technologies have greatly improved our ability to detect wide range of novel genomic and transcriptomic discoveries in the field of biomedical and veterinary science researches [1]. However, advancement in NGS technologies has also created significant challenges in bioinformatics and experimental design [2], particularly the major challenges is systematic evaluation of quality control of the RNA-seq data [3]. Monitoring and surveying QC metrics of NGS transcriptomic data provides unique and independent evaluations of RNA-seq data quality from differing perspectives [4]. Moreover, properly conducting of QC protocols and correctly interpreting the QC results are crucial to ensure a successful and meaningful RNA-seq study in veterinary and biomedical sciences [5]. Recent studies have proposed several web-based bioinformatics workflow for QC analysis [6], namely: Quality Control for RNA-Seq (QuaCRS) [7], FastqPuri [8], RNA-QC-chain [9], and clinQC [10], NGSCheckMate [11], QC for illumina RNA-seq data [12], FastQ Screen [13], NGS-QC Toolkit [14] and SEQC/MAQC consortium [15-16]. In this paper, we have evaluated a comprehensive QC analysis of liver and pituitary RNA-seq data of Hereford bulls using the web-based bioinformatics workflow of strandNGS software.

Materials and Methods

Experimental animals: After slaughtering, the liver (n=15) and pituitary gland (15) tissues samples from Hereford bulls were collected from Institute of Genetics and Animal Breeding, Polish Academy of Science (PAS), Jastrzębiec, Poland and the collected tissues were immediately kept in liquid nitrogen, and finally stored in deep freezer at –80°C. All procedures involving handling of animals were in accordance with the guiding principles of care and use of research animals. All experimental animals were reared in a closed herd with uniform feeding and environmental condi-

tions, and were approved by the local ethics commission (permission No. 3/2005) of Institute of Genetics and Animal Breeding, PAS, Jastrzębiec, Poland.

Experimental design and methodological procedures of NGS and QC bioinformatics analysis: Isolation of Total RNA were performed from 50–60 mg of frozen liver and pituitary tissues (n= 30) of Hereford bulls using TRIzol reagent (Thermo Fisher Scientific Inc., Waltham, MA, USA) and according to the guanidinium thiocyanate method [17]. Isolated total RNA were purified to remove the genomic DNA contamination using the RNase-free DNase clean-up kit (Thermo Fisher Scientific Inc., Waltham, MA, USA). The integrity of RNA were assessed using Bioanalyzer 2100 with RNA 6000 Nano Lab-chips according to manufacturer's instructions (Agilent Technologies, Palo Alto, CA, USA). The RNA integrity number (RIN) values of all investigated Hereford bulls were ranged from 7.5 to 8.5 After the quality check, total RNA were pooled in a single tube and fragmentation of pooled total RNA were performed using GeneRead rRNA Depletion Kits (Qaigen, Germany) and M220 Focused-ultrasonicator™ (Covaris, Inc. USA). The library preparation of liver and pituitary samples were performed using ScriptSeq™ v2 RNA-Seq library preparation kit (Epicentre, Illumina, USA). At the end of NGS laboratory procedure, two runs of MiSeq illumina sequencing of liver and pituitary transcriptome of Hereford bulls were performed using MiSeq® reagent sequencing kit v2 (Illumina, USA) and 251 bp paired-end sequence reads were generated. Finally, the obtained paired-end sequencing data (Liv-Pit transcriptome) generated from MiSeq Illumina instrument were loaded to the StrandNGS workbench to perform the quality control (QC) assessment of the RNA-Seq data of liver and pituitary transcriptome of Hereford bulls using StrandNGS software. The overall workflow of QC procedure is illustrated in Figure 1.

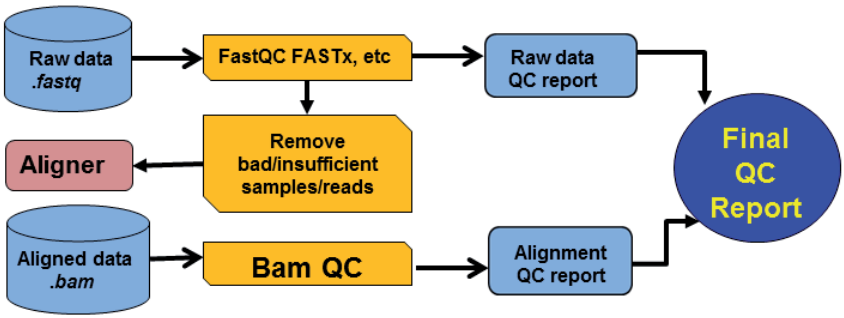


Figure 1. Overall workflow of quality control (QC) analysis of the generated RNA-seq data of bovine liver and pituitary gland transcriptome of Hereford bulls

Results

Raw RNA-seq data and NCBI submission: After two MiSeq run, raw RNA-seq data of 7.11GB with an average of 72% duplicate reads were obtained from liver and pituitary transcriptome of Hereford bulls (Table 1). The raw RNA-seq FASTQ sequencing data were deposited in the NCBI database under submission BioProject: PRJNA312148 (http://www.ncbi.nlm.nih.gov/sra?linkname=bioproject_sra_all&from_uid=312148).

Table 1. The QC matrices of the generated RNA-seq data of bovine liver and pituitary gland transcriptome of Hereford bulls

RNA-seq matrices parameters	First Miseq run RNA-seq data	Second Miseq run RNA-seq data
RNA-seq data size	3.62 GB	3.49 GB
Paired end reads length	251bp	251bp
Total clusters PF	15,477,651	14,736,047
Total reads raw data	30,955,302	29,472,094
Duplicate reads (%)	72.02	71.69
NCBI SRA submission	BioProject: PRJNA312148	BioProject: PRJNA312148

Assessment of Library QC (Figure 2): The Library QC section includes metrics that pertain to the sequence library. This makes it possible to determine if all the targeted regions are represented in the sequence library and that there is no redundancy in all aligned reads. The library QC metrics results presented as library complexity plot illustrated the distribution of uniquely mapped reads to the total set of mapped paired-end reads (Figure 2). The plot displayed an X = Y line (slope m = 1) to verify the complexity of the library. Results revealed a lot of duplicate reads in the RNA-seq library, resulting in low library complexity (0.28) in the NGS experiment (Figure 2).

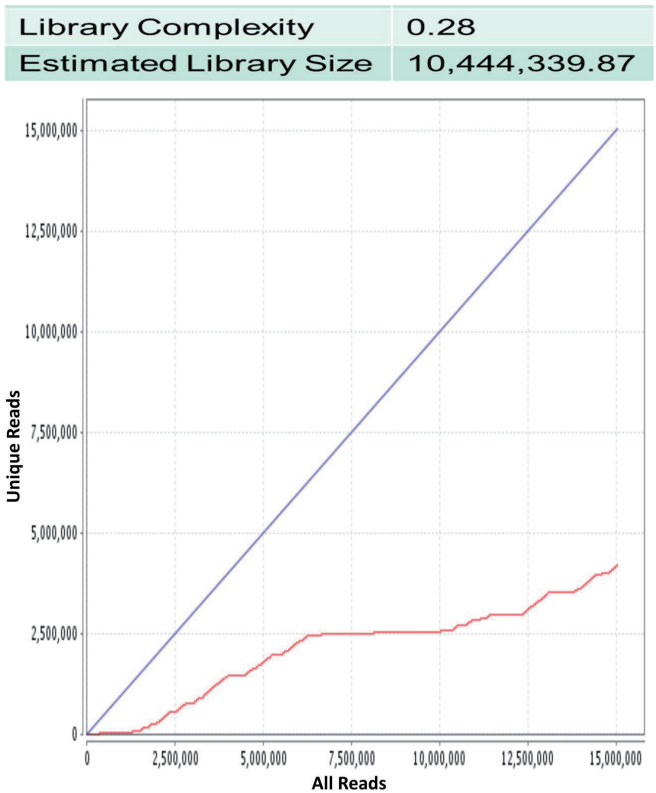
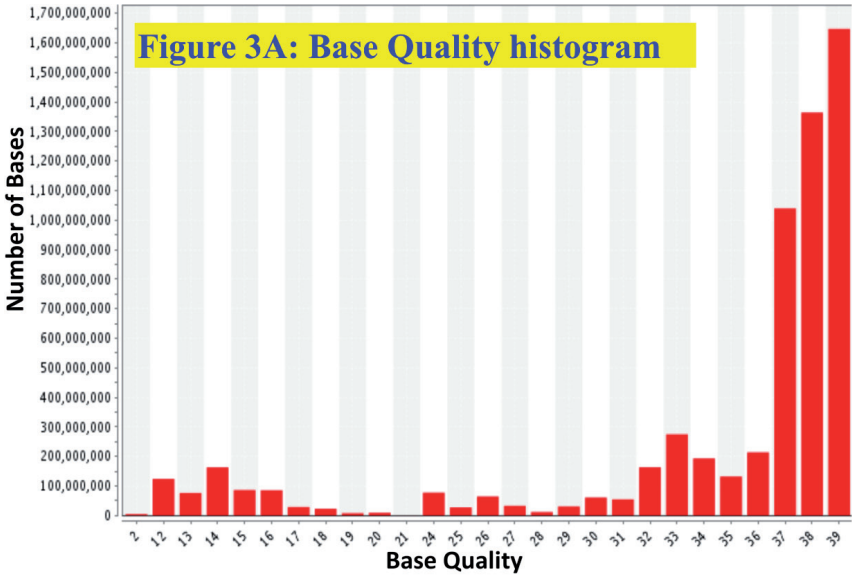


Figure 2. Library complexity plot of the generated RNA-seq data of bovine liver and pituitary gland transcriptome of Hereford bulls

Assessment of QC of RNA-seq reads before alignment (pre-alignment) using Strand NGS software (**Figure 3**): The QC assessment of over 60 million raw RNA-seq reads before alignment were measured based on the histogram plot of number of bases against the base-quality as shown in Figure 3A and histogram plot of number of read count against the average base quality in read as shown in Figure 3B.



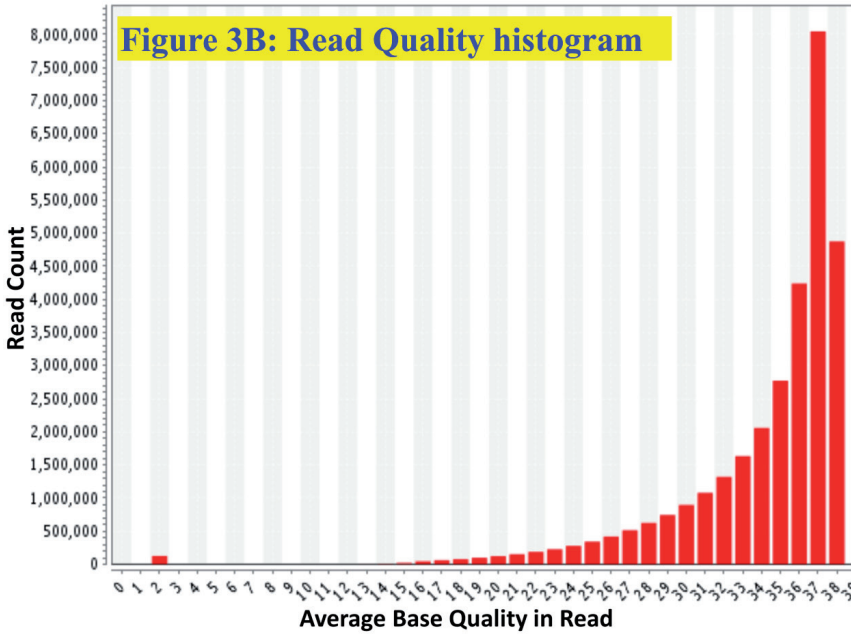


Figure 3. Showing the histograms of numbers of bases against base quality (Figure 3A) and read count against average base quality in read (Figure 3B) histogram before alignment of the generated RNA-seq data of bovine liver and pituitary gland transcriptome of Hereford bulls

Assessment of QC of RNA-seq reads after alignment (post-alignment) using Strand NGS software (Figure 4): The post-alignment QC assessment were performed based on the number of QC RNA-seq paired-end reads to Bos Taurus reference genome. After alignment of QC paired-end reads to Bos Taurus reference genome, the obtained post-alignment QC results of liver and pituitary RNA-seq data were presented in five histograms figures and a pie chart figure (Figure 4A-F): i) the number of paired-end reads against alignment score (Figure 4A); ii) the number of paired-end reads against mapping quality (Figure 4B); the number of paired-end reads against report match count (Figure 4C); iv) the match status displayed as a pie chart (Figure 4D) representing the proportion of reads with different read statuses for paired alignment data; v) the number of paired-end reads against read length (Figure 4E); and the number of paired-end reads against read quality (Figure 4F).

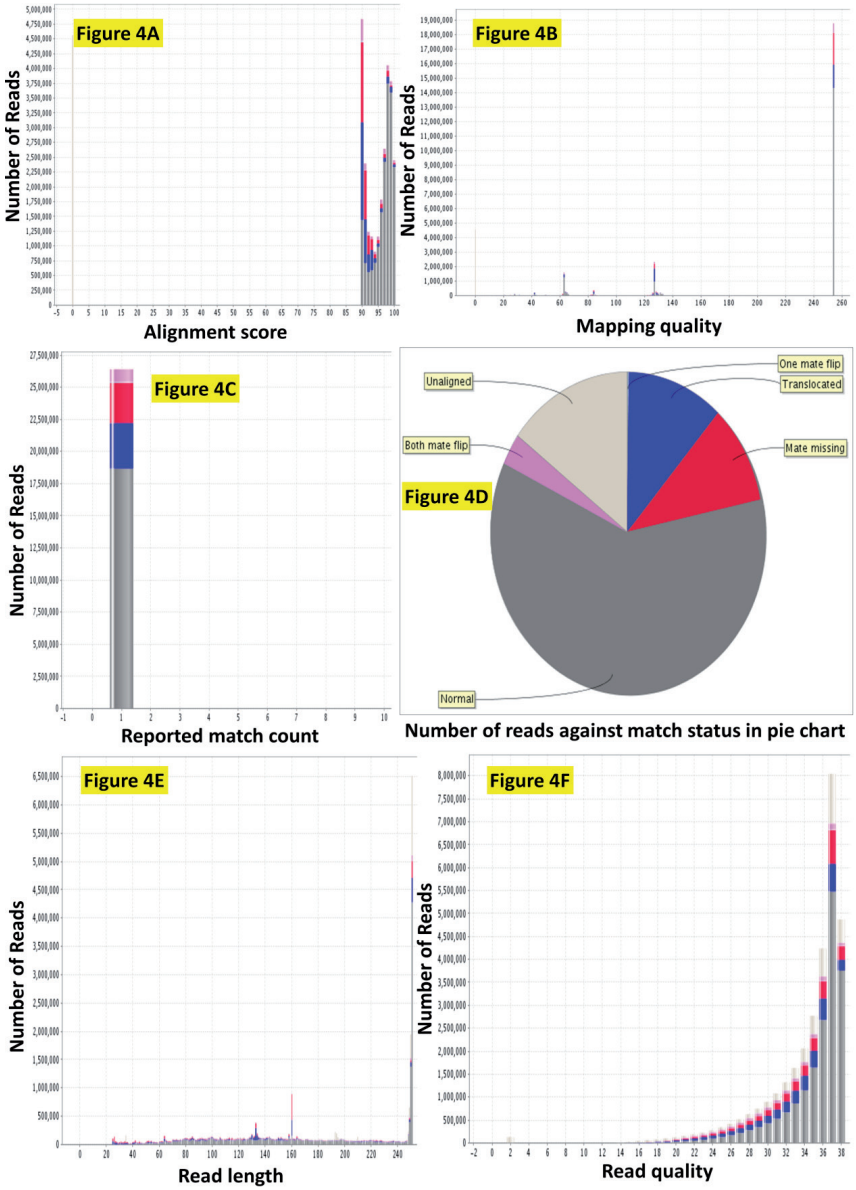


Figure 4. Post-alignment QC results based on the number of QC RNA-seq paired-end reads to *Bos Taurus* reference genome. Figure representing number of reads against alignment score (Figure 4A); against mapping quality (Figure 4B); against match count (Figure 4C); against match status in pie chart (Figure 4D); against read length (Figure 4E); against read quality (Figure 4F), respectively

Figure Legends color indicates:

Light blue: One mate flip: One member of the read pair has an orientation flip, indicating an inversion.

Blue: Translocated: The read's mate is present on a different chromosome.

Red: Mate missing: The read's mate is not present in the sample.

Grey: Normal: The read is within its mate, and orientation of the reads agrees with the model.

Pink: Both mate flip: Both members undergo an orientation change.

Light grey: Unaligned: not assigned.

Discussions

The QC of RNA-seq data before and after alignment is an essential prerequisite in any NGS experiments because of experimental biases in nucleotide composition, library preparation issues, PCR biases – all of which influence the RNA-seq analysis [2-6].Therefore, before any RNA-seq analysis or sequence alignment is done, a read-level analysis of RNA-seq data must be performed in all NGS experiments as the first essential step to start the bioinformatics analysis. In literatures, several bioinformatics tools that are publically available to conduct the QC analysis on raw FastQ files *viz*., Musket [18], HiTEC [19] and SHREC [20] and FastQC package developed by the Babraham Institute bioinformatics group (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). In our study, we utilized the commercially available StrandNGS v1.3 tools (Agilent; http://www.strand-ngs.com/) to perform the QC analysis of combined liver and pituitary transcriptome (n=30) of Hereford bulls. In pre-alignment QC analysis, StrandNGS tool was successfully performed to check the library complexity, base quality and nucleotide distribution, the average base quality score per read, identification of the most duplicated reads (PCR artefact, same fragment read twice or excessive presence of a fragment), unique reads and all reads mapped to the targeted region (Figure 2–3). The library complexity plot (Table 1) helps to identify the percentage of total aligned reads that are unique and also to identify the reads represent the entire range of targeted regions. Alternatively, this plot also informs if

there is a redundancy of reads in the sequenced data. Furthermore, in the context of library complexity plot, duplicate reads are those that map to the same location with same orientation. However, in the post-alignment QC analysis (Figure 4), StrandNGS tool was successfully performed the alignment or mapping of the processed pre-alignment RNA-seq data on the *Bos Taurus* reference genome to identify the number of aligned reads, uniquely aligned reads, which was very critical and crucial for the genetic variants or single nucleotide polymorphisms (SNPs) discoveries [21, 22] and identification of DEGs, gene network and metabolic pathways in bovine liver and pituitary tissues [23, 24].

Conclusion

By testing the RNA-seq methodology on Illumina platform, two MiSeq sequencing runs yielded significantly high quality of 60 million sequencing reads. Bioinformatics QC analysis on raw RNA-seq data of Hereford bulls revealed that the pre-alignment and post-alignment analysis of raw RNA-seq data were successfully performed to align and assemble the Hereford liver and pituitary gland transcriptome to the *Bos Taurus* reference genome, However, only more 50% RNA-seq data generated from total RNA library rRNA depletion kit were mapped. Based on the obtained results one can conclude that liver and pituitary transcriptome sequencing with rRNA depletion method is less effective than mRNA RNA-seq method.

Acknowledgement

This research paper was financially supported by National Science Centre, Krakow, Poland, Project No. 2012/05/B/NZ2/01629; entitled "Analiza transkryptomów genomu Bos taurus przy zastosowaniu technologii sekwencjonowania kolejnej generacji".

References

1. Pareek CS, Smoczynski R, Tretyn A. Sequencing technologies and genome sequencing. *J Appl Genet*. 2011; 52:413–435.
2. Williams AG, Thomas S, Wyman SK, Holloway AK. RNA-seq Data: Challenges in and Recommendations for Experimental Design and Analysis. Curr Protoc Hum Genet. 2014; 83:1–20.
3. Chao HP, Chen Y, Takata Y, Tomida MW, Lin K, Kirk JS, Simper MS, Mikulec CD, Rundhaug JE, Fischer SM, Chen T, Tang DG, Lu Y, Shen J. Systematic evaluation of RNA-Seq preparation protocol performance. BMC Genomics. 2019; 20:571.
4. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, Szcześniak MW, Gaffney DJ, Elo LL, Zhang X, Mortazavi A. A survey of best practices for RNA-seq data analysis. Genome Biol. 2016; 17:13.
5. Xu J, Gong B, Wu L, Thakkar S, Hong H, Tong W. Comprehensive Assessments of RNA-seq by the SEQC Consortium: FDA-Led Efforts Advance Precision Medicine. Pharmaceutics. 2016; 8 pii: E8.
6. Li W, Richter RA, Jung Y, Zhu Q, Li RW. Web-based bioinformatics workflows for end-to-end RNA-seq data computation and analysis in agricultural animal species. BMC Genomics. 2016; 17:761.
7. Kroll KW, Mokaram NE, Pelletier AR, Frankhouser DE, Westphal MS, Stump PA, Stump CL, Bundschuh R, Blachly JS, Yan P. Quality Control for RNA-Seq (QuaCRS): An Integrated Quality Control Pipeline. Cancer Inform. 2014; 13:7–14. doi: 10.4137/CIN.S14022. eCollection 2014.
8. Pérez-Rubio P, Lottaz C, Engelmann JC. FastqPuri: high-performance preprocessing of RNA-seq data. BMC Bioinformatics. 2019; 20:226. doi:10.1186/s12859-019-2799-0.
9. Zhou Q, Su X, Jing G, Chen S, Ning K. RNA-QC-chain: comprehensive and fast quality control for RNA-Seq data. BMC Genomics. 2018; 19:144.
10. Pandey RV, Pabinger S, Kriegner A, Weinhäusel A. ClinQC: a tool for quality control and cleaning of Sanger and NGS data in clinical research. BMC Bioinformatics. 2016; 17:56.

11. Lee S, Lee S, Ouellette S, Park WY, Lee EA, Park PJ. NGSCheckMate: software for validating sample identity in next-generation sequencing studies within and across data types. Nucleic Acids Res. 2017; 45:e103.
12. Sheng Q, Vickers K, Zhao S, et al. Multi-perspective quality control of Illumina RNA sequencing data analysis. Brief Funct Genomics. 2017; 16:194–204.
13. Wingett SW, Andrews S. FastQ Screen: A tool for multi-genome mapping and quality control. F1000Res. 2018; 7:1338
14. Patel RK, Jain M. NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. PLoSOne 2012; 7:e30619.
15. SEQC/MAQC-III Consortium. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. Nat Biotechnol. 2014; 32:903–914.
16. Xu J, Gong B, Wu L, Thakkar S, Hong H, Tong W. Comprehensive Assessments of RNA-seq by the SEQC Consortium: FDA-Led Efforts Advance Precision Medicine. Pharmaceutics. 2016; 8:1.
17. Chomczynski P, Sacchi N. Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. Anal Biochem. 1987; 162:156–159.
18. Liu Y, Schroder J, Schmidt B. Musket: a multistage k-mer spectrum-based error corrector for Illumina sequence data. Bioinformatics 2013; 29:308–315.
19. Ilie L, Fazayeli F, Ilie S. HiTEC: accurate error correction in high-throughput sequencing data. Bioinformatics 2011; 27: 295–302.
20. Schroder J, Schroder H, Puglisi SJ, etal. SHREC: a short-read error correction method. Bioinformatics 2009; 25:2157–2163.
21. Pareek CS, Smoczyński R, Kadarmideen HN, Dziuba P, Błaszczyk P, Sikora M, Walendzik P, Grzybowski T, Pierzchała M, Horbańczuk J, Szostak A, Ogluszka M, Zwierzchowski L, Czarnik U, Fraser L, Sobiech P, Wąsowicz K, Gelfand B, Feng Y, Kumar D. Single Nucleotide Polymorphism Discovery in Bovine Pituitary Gland Using RNA-Seq Technology. PLoS One. 2016; 11: e0161370.
22. Pareek CS, Błaszczyk P, Dziuba P, Czarnik U, Fraser L, Sobiech P, Pierzchała M, Feng Y, Kadarmideen HN, Kumar D. Single nucleotide polymorphism discovery in bovine liver using RNA-seq technology. PLoS One. 2017; 12: e0172687.

23. Wysocka D, Sobiech P, Herudzińska M, Sachajko M, Pareek CS. Investigation of candidate genes for metabolic disorders expressed in liver and pituitary gland by comparing the RNA-seq data of Polish-HF and Polish-Red cattle. Trans Res Vet Sci. 2018; 1: 69–83.
24. Pareek CS, Sachajko M, Jaskowski JM, Herudzinska M, Skowronski M, Domagalski K, Szczepanek J, Czarnik U, Sobiech P, Wysocka D, Pierzchala M, Polawska E, Stepanow K, Ogłuszka M, Juszczuk-Kubiak E, Feng Y, Kumar D. Comparative Analysis of the Liver Transcriptome among Cattle Breeds Using RNA-seq. Vet Sci. 2019; 6:36.