Evaluation of quality-related parameters in raw NGS data and implementing tools to obtain them

H. Mohammadi[a,*], M. Sehhati[b] , A. Vaez[c]

a Student Research Committee, Department of Bioinformatics, School of Advanced Medical Technologies in Medicine, Isfahan University of Medical Sciences, Isfahan, 81746-73461, Iran

b Department of Bioinformatics, School of Advanced Technologies in Medicine, Isfahan University of Medical Sciences, Isfahan, 81746-73461, Iran

c Department of Bioinformatics, School of Advanced Technologies in Medicine, Isfahan University of Medical Sciences, Isfahan, 81746-73461, Iran

*hannane.mohammadi@gmail.com

Abstract: Given the low accuracy of Next Generation Sequencing (NGS) compared to Sanger sequencing, it is likely to misinterpret the data, if no primary quality control is performed. Quality control (QC) of raw data is considered as an important initiative step for overcomming instrumental artifacts. However, the main problem is that there is niether specific guideline nor gold standard parameters. This study aims to re-introduce parameters related to QC and suggests combination of existing tools for efficient quality checking. The suggested parameters of pre-processing to focus on, namely Quality Score, Read Complexity, Duplicate Reads were extracted for the data. The very first parameter to investigate, quality score as measure of uncertainty of basecall, depended on instrumental variables. Due to variation of length and arrangement of bases in each read, it is necessary to observe base composition visually for further decisions like adapter trimming or defining a quality score cut-off. Another important parameter to consider was read complexity which could cause mistaken alignment.Third one to investigate was duplicate read. Removing duplicate reads, believed to be a result of experimental errors, may cause loss of uniqe biological information. Also efficiency of the sequencer, bases of high quality, Primer/Adapter contamination and N base count were helpful for decision making. Tools used to obtain effective factors and implement an effective pipeline were a suggested combination of PPR Plot program[1], FaQCs[2], AfterQC[3] and NGS QC Toolkit[4]. A lot of information is generated by using QC tools that can help deciding on properties of secondary step of NGS analysis, utilizing our implemented combination of aforementioned tools, data-specific features like Quality Score, Read Complexity and Duplicate Reads could be quantified to simplify quality control for an expert.

Keywords: NGS; Quality Control; DNA-Seq; Pre-processing

References

[1] L. J. Manley, D. Ma, and S. S. Levine, “Monitoring error rates in Illumina sequencing,” *J. Biomol. Tech.*, vol. 27, no. 4, pp. 125–128, 2016.

[2] C.-C. Lo and P. S. G. Chain, “Rapid evaluation and quality control of next generation sequencing data with FaQCs,” *BMC Bioinformatics*, vol. 15, no. 1, p. 366, 2014.

[3] S. Chen, T. Huang, Y. Zhou, Y. Han, M. Xu, and J. Gu, “AfterQC: automatic filtering, trimming, error removing and quality control for fastq data,” *BMC Bioinformatics*, vol. 18, no. S3, p. 80, 2017.

[4] R. K. Patel and M. Jain, “NGS QC toolkit: A toolkit for quality control of next generation sequencing data,” *PLoS One*, vol. 7, no. 2, 2012.