The Journal of Open Source Software

CheckQC: Quick quality control of Illumina sequencing runs

Matilda Åslin[1], **Monika Brandt**[1], and **Johan Dahlberg**[1]

1 Department of Medical Sciences, Molecular Medicine and Science for Life Laboratory Uppsala University, Uppsala, Sweden

DOI: 10.21105/joss.00556

Software

- Review
- Repository
- Archive

Submitted: 26 January 2018
Published: 05 February 2018

Introduction

The last quarter of 20th century has been marked by the drive to decode genes and later whole genomes, giving rise to the field of genomics (Lander et al. 2001). In order to study the genome it is essential to know the sequential order of nucleotides in the DNA sequence. The process of determining this order is referred to as DNA sequencing. These technologies are not limited to the study of DNA, but can also be used in a wide variety of other applications such as studying the transcriptome (RNA sequencing) and epigenetic modifications of the DNA itself. One of the most widely used technologies for DNA sequencing is the sequencing-by-synthesis solution provided by Illumina Inc (E. R. Mardis 2017). Sequencing is typically carried out at sequencing core facilities that offer sequencing as a service (Spjuth et al. 2016).

CheckQC was created to quickly assess the quality of a sequencing run specifically aimed at facilities using Illumina sequencing technology. CheckQC requires that the raw data is processed by the bcl2fastq software ("Bcl2fastq Conversion Software," n.d.) provided by Illumina. In just a few seconds, CheckQC gathers statistics from the sequencing run and returns warnings about any metrics not fulfilling previously specified quality control (QC) criteria. Furthermore, the exit code returned by the program indicates whether all the criteria were met. This feature makes it easy to plug CheckQC into a sequencing data processing workflow, which can be aborted if quality criterias are not met. The metrics to check and thresholds used for QC are specified in a configuration file, making CheckQC easily adapted to the specific needs of any user or core facility.

One widely used quality control software is FastqQC ("FastQC a Quality Control Tool for High Throughput Sequence Data," n.d.), which focus on assessing the quality of individual samples. While CheckQC also evaluates criteria on a sample level, it delivers two other perspectives on the data that are not addressed by the existing software. Firstly, CheckQC evaluates criteria for the run as a whole, which is generally important from a core facility perspective. An example of this is that it provides confirmation that a sequencing run as a whole has generated a sufficient amount of data according to the pre-defined QC thresholds. Secondly, it contains a feature of setting thresholds for different QC criteria based on instrument and/or run type and evaluates these criteria automatically, which to the best of our knowledge, is not supported by existing software.

Another software library which focuses on working with Illumina data from a sequencing run perspective is basecallQC (Carroll and Dore 2017). This is a library for the R (R Core Team 2017) programming language, which allows the user to perform utility actions related to a sequencing run, as well as create summary reports. It does not, however, provide built in functionality to assess whether quality control thresholds are fulfilled or not, nor does it provide a commandline interface to the user.

Features

- Support for the following Illumina instrument types: HiSeq2500, HiSeqX, MiSeq and NovaSeq.
- QC criteria are specified based on instrument type and/or run configuration.
- Thresholds are set on a error and/or warning level, where QC evaluations only reporting warnings will result in a successful run completion, while errors will not. This way CheckQC can easily be incorporated in an automated workflow, e.g. that implemented with Arteria (Dahlberg et al. 2017), where the exit code can be examined to decide whether to proceed in a workflow or not.
- CheckQC can be run as a command-line application or as a web service.
- CheckQC currently supports parsing of Illumina InterOp files and Stats.json (generated by Illumina software bcl2fastq2 since version 2.18).
- CheckQC is designed to be extendable, thus allowing new input file formats to be parsed, and new types of criteria to be added with ease.

References

"Bcl2fastq Conversion Software." n.d. https://support.illumina.com/sequencing/sequencing_software/bcl2fastq-conversion-software.html.

Carroll, Thomas, and Marian Dore. 2017. *BasecallQC: Working with Illumina Basecalling and Demultiplexing Input and Output Files*. https://doi.org/10.18129/B9.bioc.basecallQC.

Dahlberg, Johan, Johan Hermansson, Steinar Sturlaugsson, and Pontus Larsson. 2017. "Arteria: An Automation System for a Sequencing Core Facility." *bioRxiv*.

"FastQC a Quality Control Tool for High Throughput Sequence Data." n.d. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Lander, E S, L M Linton, B Birren, C Nusbaum, M C Zody, J Baldwin, K Devon, et al. 2001. "Initial Sequencing and Analysis of the Human Genome." *Nature* 409 (6822):860–921.

Mardis, Elaine R. 2017. "DNA Sequencing Technologies: 2006-2016." *Nat. Protoc.* 12 (2):213–18.

R Core Team. 2017. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Spjuth, Ola, Erik Bongcam-Rudloff, Johan Dahlberg, Martin Dahlö, Aleksi Kallio, Luca Pireddu, Francesco Vezzi, and Eija Korpelainen. 2016. "Recommendations on E-Infrastructures for Next-Generation Sequencing." *Gigascience* 5 (June):26.