

Data and text mining

ChronQC: a quality control monitoring system for clinical next generation sequencing

Nilesh R. Tawari*, Justine Jia Wen Seow, Dharuman Perumal, Jack L. Ow, Shimin Ang, Arun George Devasia and Pauline C. Ng*

Computational and Systems Biology, Genome Institute of Singapore, Genome, #02-01, Singapore 138672, Singapore

*To whom correspondence should be addressed.

Associate Editor: Bonnie Berger

Received on August 4, 2017; revised on December 21, 2017; editorial decision on December 26, 2017; accepted on December 27, 2017

Abstract

Summary: ChronQC is a quality control (QC) tracking system for clinical implementation of next-generation sequencing (NGS). ChronQC generates time series plots for various QC metrics to allow comparison of current runs to historical runs. ChronQC has multiple features for tracking QC data including Westgard rules for clinical validity, laboratory-defined thresholds and historical observations within a specified time period. Users can record their notes and corrective actions directly onto the plots for long-term recordkeeping. ChronQC facilitates regular monitoring of clinical NGS to enable adherence to high quality clinical standards.

Availability and implementation: ChronQC is freely available on GitHub (<https://github.com/nilesh-tawari/ChronQC>), Docker (<https://hub.docker.com/r/nileshtawari/chronqc/>) and the Python Package Index. ChronQC is implemented in Python and runs on all common operating systems (Windows, Linux and Mac OS X).

Contact: tawari.nilesh@gmail.com or pauline.c.ng@gmail.com

Supplementary information: [Supplementary data](#) are available at *Bioinformatics* online.

1 Introduction

Clinical implementation of next-generation sequencing (NGS) requires high quality standards. This necessitates continuous monitoring of various quality control (QC) metrics and test performance (Doig *et al.*, 2017). QC monitoring serves two purposes: (1) it detects problems in the laboratory's processes or systems that could render a patient's result invalid (e.g. batch effects or machine instability) and (2) it identifies opportunities for system improvement.

QC is an important part of an NGS pipeline. A number of QC tools are designed for detecting problematic data at specific steps in the NGS pipeline. FastQC and Qualimap are designed for QC of FASTQ and BAM files, respectively, (<https://www.bioinformatics.babraham.ac.uk/projects/fastqc>; Okonechnikov *et al.*, 2016). AlmostSignificant (Ward *et al.*, 2016) aggregates QC metrics from Illumina sequencers, FastQC and FastQScreen. MultiQC is widely adopted by the community and consolidates the output of several QC tools in a single report (Ewels *et al.*, 2016). However, all the afore-mentioned tools lack historical tracking which is required for

long-term monitoring of laboratory test results. Therefore, there still exists a need for a QC monitoring tool.

A QC monitoring tool should have the ability to annotate, track and compare historical QC metrics. To the best of our knowledge, a standalone QC monitoring tool that can be integrated into the existing infrastructure of a laboratory is currently not available. In this light, we present ChronQC, an open-source, interactive, record-keeping QC monitoring system for clinical implementation of NGS.

2 Materials and methods

2.1 ChronQC workflow

The workflow of ChronQC is depicted in Figure 1. ChronQC works downstream of MultiQC. ChronQC stores a sample's MultiQC output along with the corresponding run ID and run date information in a ChronQC statistics database (`chronqc.stats.sqlite`) using Python code (Fig. 1, Step 1). Alternatively, ChronQC can work with a custom SQLite statistics database that contains information on NGS

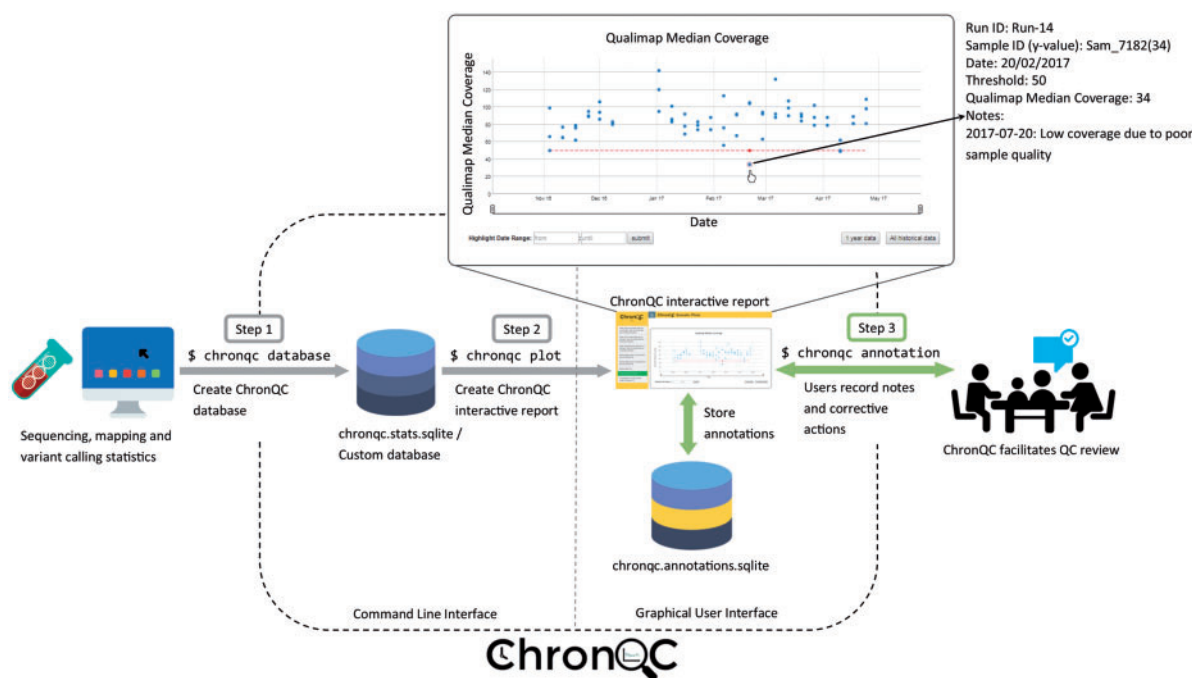


Fig. 1. ChronQC workflow. ChronQC has two components: a command line interface compatible with sequencing machines and a graphical user interface compatible with the clinical environment. HTML plots display metrics for each run or sample. Annotations are displayed on the right side of the plot and are stored in the `chronqc.annotations.sqlite` database for long-term record-keeping

sequencing runs, run dates and laboratory or bioinformatics QC metrics.

The next step is to generate a ChronQC interactive report (Fig. 1, Step 2). Using the statistics database and a configuration file, ChronQC generates time series plots for various metrics to create an interactive, self-contained HTML file. ChronQC plots are rendered using an open-source JavaScript charting library Dygraphs (<https://dygraphs.com>). ChronQC plots can be visualized in a browser to facilitate QC review.

2.2 ChronQC plots

ChronQC currently supports seven types of charts (Supplementary Tables S1 and S2). An example can be seen at the top of Figure 1 (more examples at <https://nilesh-tawari.github.io/chronqc>). The different chart types are associated with different QC tracking features based on Westgard rules for clinical validity (e.g. demarcating ± 2 SD) (Westgard et al., 1981), laboratory-defined thresholds and historical QC observations within a specified time period. ChronQC plots can assist in identifying trends, bias and excessive scatter in the clinical data so that corrective and preventive actions can be taken to ensure that patient results remain clinically valid.

2.3 Interactivity and tracking

ChronQC is designed to be interactive. ChronQC plots can be adjusted to a time period and are zoomable. Mousing over a point displays its associated data such as run ID, sample IDs and corresponding values (Fig. 1). Furthermore, users can record notes such as corrective actions on the plots by clicking on a point or selecting a date. User notes are stored for long-term recordkeeping in the SQLite ChronQC annotations database (Fig. 1, Step 3). The plots are interlinked so that when an individual point or date is annotated in one graph, the same annotation appears on other graphs. By using the ChronQC report with the ChronQC annotations database, users can see the notes that have been recorded previously.

3 Conclusion

ChronQC is built in a modular fashion and can leverage on existing tools. It can be used to track various laboratory and bioinformatics QC parameters over a period of time. Importantly, users can record their notes directly onto the plots for long-term record-keeping. Interactive QC data representation over a period of time can help with troubleshooting anomalies and spotting trends with the ultimate purpose of maintaining clinical standards.

Acknowledgements

The authors would like to thank Patrick Tan, Alexander Lezhava, Tony Lim, Christopher Wong and Sarah Ng for their scientific inputs.

Funding

This work was supported by the POLARIS program [Biomedical Research Council Strategic Positioning Fund, 2012/001].

Conflict of Interest: none declared.

References

- Doig, K.D. et al. (2017) PathOS: a decision support system for reporting high throughput sequencing of cancers in clinical diagnostic laboratories. *Genome Med.*, 9, 38.
- Ewels, P. et al. (2016) MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, 32, 3047–3048.
- Okonechnikov, K. et al. (2016) Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics*, 32, 292–294.
- Ward, J. et al. (2016) AlmostSignificant: simplifying quality control of high-throughput sequencing data. *Bioinformatics*, 32, 3850–3851.
- Westgard, J.O. et al. (1981) A multi-rule Shewhart chart for quality control in clinical chemistry. *Clin. Chem.*, 27, 493–501.