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 G. Devasia, Pauline C. Ng\*

Computational and Systems Biology, Genome Institute of Singapore, 60 Biopolis Street, Genome, #02-01, Singapore 138672

\*To whom correspondence should be addressed.

**Supplementary Table 1:** ChronQC chart types. Examples of these chart types can be seen at <https://nilesh-tawari.github.io/chronqc>

|  |  |  |
| --- | --- | --- |
| Chart type | Description  | Use case |
| Time series plot with mean and standard deviation | A time series plot of numerical data with historical runs. Rolling mean and ±2 standard deviations are shown.  | Can be used to track metrics such as total number of reads. The window to compute rolling mean and ±2 standard deviations can be set to either a specified duration (e.g. runs in the past year) or number of historical runs (e.g. past 10 runs).  |
| Time series plot with an absolute threshold | A time series plot of numerical data with user-defined lower and upper thresholds. | Can be used to track metrics such as depth of coverage, Ti/Tv ratio, and GC content per sample. Lower and upper thresholds can be based on empirical values.  |
| Time series plot with percentage of samples above a threshold | A time series plot representing percentage of numerical data above a user-defined threshold. | Can be used to track metrics such as percentage of samples in a run that exceed a certain threshold.  |
| Time series plot with percentage of samples with a category label | A time series plot of categorical data representing % of samples in a run with y-value equal to a category label. | Can be used to track percentage of samples in a run with a certain label. E.g. % of samples labeled “PASS”.  |
| Time series box-and-whisker plot of numerical data | A monthly time series box-and-whisker plot of numerical data. | Can be used to track number of single nucleotide variants (SNVs) and indels observed for each month.  |
| Time series with stacked bar plot | A stacked bar plot of categorical data summarized for each month. | Can be used to track number of observed mutations in clinically actionable genes per month. |
| Time series with bar and line plot | A bar and line representation of categorical data. | Can be used to track number of observed mutations in clinically actionable genes per month.  |

**Supplementary Table 2:** NGS tools with their corresponding QC metrics and their chart types implemented in ChronQC’s default configuration.

|  |  |  |
| --- | --- | --- |
| Tool name | QC metrics  | Chart type implemented in default JSON (config file) |
| FastQC | FastQC\_percent\_gc | time\_series\_with\_mean\_and\_stdev |
| FastQC\_total\_sequences | time\_series\_with\_mean\_and\_stdev |
| FastQC\_percent\_duplicates | time\_series\_with\_mean\_and\_stdev |
| FastQC\_percent\_fails  | time\_series\_with\_mean\_and\_stdev |
| FastQC\_avg\_sequence\_length  | time\_series\_with\_mean\_and\_stdev |
| QualiMap | QualiMap\_30\_x\_pc | time\_series\_with\_mean\_and\_stdev |
| QualiMap\_percentage\_aligned | time\_series\_with\_mean\_and\_stdev |
| QualiMap\_avg\_gc  | time\_series\_with\_mean\_and\_stdev (if FastQC\_percent\_gc is present this plot is omitted to avoid duplication) |
| QualiMap\_mapped\_reads  | time\_series\_with\_mean\_and\_stdev |
| QualiMap\_median\_coverage | time\_series\_with\_percentage\_of\_samples\_above\_threshold (Default threshold 30) and time\_series\_with\_absolute\_threshold (Default threshold 30) |
| QualiMap\_total\_reads  | time\_series\_with\_mean\_and\_stdev (if FastQC\_total\_sequences is present this plot is omitted to avoid duplication) |
| Bamtools | Bamtools\_mapped\_reads\_pct  | time\_series\_with\_mean\_and\_stdev (if QualiMap\_percentage\_aligned is present this plot is omitted to avoid duplication) |
| Samtools | SamtoolsFlagstat\_mapped\_passed  | time\_series\_with\_mean\_and\_stdev (if QualiMap\_mapped\_reads is present this plot is omitted to avoid duplication) |
| Bcftools | BcftoolsStats\_number\_of\_MNPs | time\_series\_with\_box\_whisker\_plot |
| BcftoolsStats\_number\_of\_SNPs | time\_series\_with\_box\_whisker\_plot |
| Bcftools\_Stats\_number\_of\_indels | time\_series\_with\_box\_whisker\_plot |
| BcftoolsStats\_number\_of\_records | time\_series\_with\_box\_whisker\_plot |
| BcftoolsStats\_tstv | time\_series\_with\_mean\_and\_stdev |
| Peddy | Peddy\_error | time\_series\_with\_percentage\_category (Default category: True) |
| Any other tool | Columns with numeric data | time\_series\_with\_mean\_and\_stdev |

References

Barnett, D. (2011). BamTools: a C++ API and toolkit for analyzing and managing BAM files. *Bioinformatics*, **27**, 1691-1692.

Brown, J. et al. (2017) FQC Dashboard: integrates FastQC results into a web-based, interactive, and extensible FASTQ quality control tool. *Bioinformatics*, **33**, 3137–3139.

Danecek, P. et al. (2017). BCFtools/csq: haplotype-aware variant consequences. *Bioinformatics*, **33**, 2037-2039.

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, **27**, 2987-2993.

Li, H. et al. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078-2079.

Narasimhan, V. et al. (2016). BCFtools/RoH: a hidden Markov model approach for detecting autozygosity from next-generation sequencing data. *Bioinformatics*, **32**, 1749-1751.

Pedersen, B. et al. (2017). Who’s Who? Detecting and Resolving Sample Anomalies in Human DNA Sequencing Studies with Peddy. *Am J Hum Genet*, **100**, 406-413.