The Journal of Open Source Software

Picard-profiler: a lightweight pipeline for collecting Picard metrics from targeted sequence mapping files.

Abhinav Sharma [1], **Talya Conradie** [2,3], **David Martino** [2], **Stephen Stick** [2,4,5], **and Patricia Agudelo-Romero** [2,6,7]

1 Division of Molecular Biology and Human Genetics, Faculty of Medicine and Health Sciences, Stellenbosch University, Cape Town. **2** Wal-yan Respiratory Research Centre, Telethon Kids Institute, WA, Australia **3** Medical, Molecular and Forensic Sciences, Murdoch University, WA, Australia **4** Department of Respiratory and Sleep Medicine, Perth Children's Hospital for Children, WA, Australia. **5** Centre for Cell Therapy and Regenerative Medicine, School of Medicine and Pharmacology, WA, Australia. **6** Australian Research Council Centre of Excellence in Plant Energy Biology, School of Molecular Sciences, The University of Western Australia, WA, Australia **7** European Virus Bioinformatics Center, TH, Germany.

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor: Open Journals

Reviewers:

- @openjournals

Submitted: 01 January 1970
Published: unpublished

Summary

Next-generation targeted genome sequencing offers the opportunity to analyse regions of interest within a genome. While it is possible to incorporate targeted sequencing into whole-genome sequencing (WGS) pipelines, there remains a gap in accurately converting WGS metrics into precise target metrics. Here, we introduce the Picard-profiler pipeline (https://doi.org/10.5281/zenodo.8251379), designed to collects metrics from alignment files in targeted sequencing written in Nextflow (Di Tommaso et al., 2017). Picard-profiler accepts inputs in various alignment formats, including SAM, BAM and CRAM files (*HTS Format Specifications*, 2023). Additionally, to refine the metrics to the target regions the inclusion of a FASTA reference file and BED intervals file is required. Subsequently, a MultiQC report (P. Ewels et al., 2016) will be generated, encompassing the updated sequencing coverage data for the targeted regions with some extras.

Picard-profiler was built using Nextflow workflow manager and integrates Picard metrics from GATK picard tools (McKenna et al., 2010; *Picard Toolkit*, 2019), using two specific metrics: (i) collecthsmetrics (*CollectHsMetrics (Picard)*, 2019), which relies upon the hybrid-selection technique to capture exon sequences for targeted sequencing experiments; and (ii) collectmultiplemetrics (*CollectMultipleMetrics (Picard)*, 2021), which captures closely related metrics such as alignment summary, insert size, and quality score. The final MultiQC report automatically collates the report from FastQC (Andrews, 2010) and Picard tools in an HTML document, which could be shared with collaborators or added as supplementary material in publications.

Picard-profiler is a portable pipeline compatible with multiple platforms, such as local desktop or workstation machines, high-performance computing environments and cloud infrastructure. Although Picard-profiler was originally created for calculating coverage in target sequencing as a follow-up step to the nf-core/methylseq pipeline, within the Airway Epithelium Respiratory Illnesses and Allergy (AERIAL) paediatric cohort study(Kicic-Starcevich et al., 2023); its versatility allows for extending its application to other sequencing panels from various next-generation methods.

The Journal of Open Source Software

Design principles and capabilities

Picard-profiler pipeline builds upon the standardised pipeline template maintained by the nf-core community (P. A. Ewels et al., 2020) for Nextflow pipelines as well as makes use of the nf-core/modules project to install modules for FastQC, MultiQC (P. Ewels et al., 2016) , Picard as well as Samtools (Danecek et al., 2021) within the pipeline Figure 1.

The use of the nf-core template facilitates in keeping the design of the pipeline generic and portable across different execution platforms, therefore the Picard-profiler pipeline can be used on local machines, HPC orchestrators (e.g. SLURM, PBS), and cloud computing systems such as AWS Batch, Azure Batch, Google Batch, in addition to the more generic Kubernetes distribution.

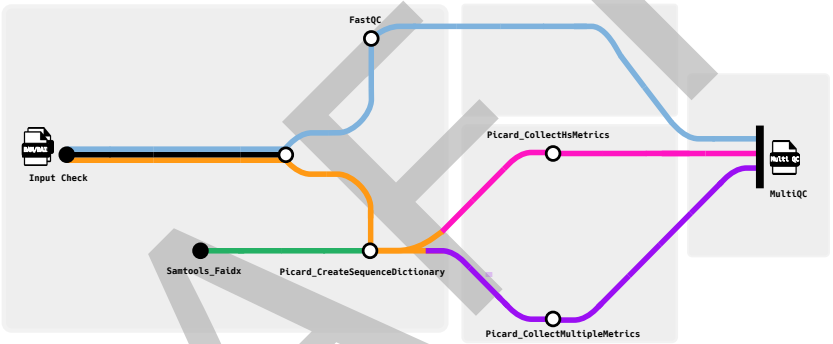

Figure 1: Subway map for various steps in the picard-profiler pipeline.

In addition to the base workflow as mentioned in Figure 1, the pipeline also includes optional picard/createsequencedictionary (*CreateSequenceDictionary (Picard)*, 2022) and Samtools modules to aid users in automatically generating the required genome dictionary (DICT) file, in case they have only the reference FASTA and BED files but intend to use the pipeline. Furthermore, depending on the quality check requirements by the users, we have enabled the metrics collection for 10x, 20x, 30x and 50x coverage.

Input and output

As standard input in the Nextflow pipelines, picard-profiler expects a CSV samplesheet as an input with the following fields.

Table 1: An example of a samplesheet for picard-profiler with three required columns, capturing the (i) name of the sample (ii) path to BAM index file and (iii) path to the BAM file.

sample	bai	bam
sample-01	/path/to/sample-01.bai	/path/to/sample-01.bam
sample-02	/path/to/sample-02.bai	/path/to/sample-02.bam

The very first step in the pipeline, as per the best practices of the nf-core template, is to check

the validity of the file paths specified to be either a POSIX compliant file system or a cloud object storage path. Upon completion, the pipeline generates a MultiQC file with the relevant results of the analysis Figure 2.

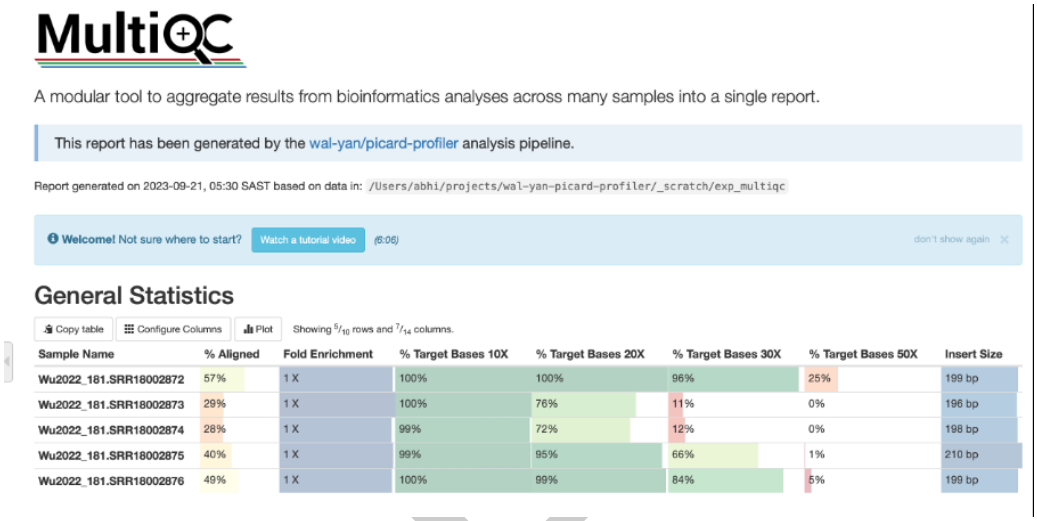

Figure 2: MultiQC report generated for Picard-profiler highlighting the refine metrics from targeted sequencing at 10X, 20X, 30X and 50X coverage.

Tutorials and documentation

The steps needed to configure the pipeline inputs and configuration for your infrastructure are available in the documentation within the Github repository itself. Getting started with the pipeline setup is straightforward given that (i) `Java` (LTS $>$ 11) (ii) `Nextflow` ($>$ 23.04) and (iii) a package manager such as `conda` or a container system such as `docker` are available in the execution environment. The in-built test profile from the pipeline can then be used to execute the profile on the relevant infrastructure with some test dataset.

```
$ nextflow run wal-yan/picard-profiler -profile test,docker -outdir test_profile_results
```

Funding Statement

This work was supported by the National Health and Medical Research Council of Australia (NHMRC115648).

References

Andrews, S. (2010). *Babraham bioinformatics - FastQC a quality control tool for high throughput sequence data*. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

CollectHsMetrics (picard). GATK. (2019, November 25). https://gatk.broadinstitute.org/hc/en-us/articles/360036856051-CollectHsMetrics-Picard-

CollectMultipleMetrics (picard). GATK. (2021, February 22). https://gatk.broadinstitute.org/hc/en-us/articles/360057440491-CollectMultipleMetrics-Picard-

CreateSequenceDictionary (picard). GATK. (2022, November 12). https://gatk.broadinstitute.org/hc/en-us/articles/360036729911-CreateSequenceDictionary-Picard-

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*(2), giab008. https://doi.org/10.1093/gigascience/giab008

Di Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., & Notredame, C. (2017). Nextflow enables reproducible computational workflows. *Nature Biotechnology*, *35*(4), 316–319. https://doi.org/10.1038/nbt.3820

Ewels, P. A., Peltzer, A., Fillinger, S., Patel, H., Alneberg, J., Wilm, A., Garcia, M. U., Di Tommaso, P., & Nahnsen, S. (2020). The nf-core framework for community-curated bioinformatics pipelines. *Nature Biotechnology*, *38*(3), 276–278. https://doi.org/10.1038/s41587-020-0439-x

Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, *32*(19), 3047–3048. https://doi.org/10.1093/bioinformatics/btw354

HTS format specifications. (2023). https://samtools.github.io/hts-specs/

Kicic-Starcevich, E., Hancock, D. G., Iosifidis, T., Agudelo-Romero, P., Caparros-Martin, J. A., Silva, D., Turkovic, L., Souef, P. N. L., Bosco, A., Martino, D. J., Kicic, A., Prescott, S. L., & Stick, S. M. (2023). *Airway epithelium respiratory illnesses and allergy (AERIAL) birth cohort: Study protocol*. medRxiv. https://doi.org/10.1101/2023.04.29.23289314

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., & DePristo, M. A. (2010). The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–1303. https://doi.org/10.1101/gr.107524.110

Picard toolkit. (2019). https://broadinstitute.github.io/picard/