The Journal of Open Source Software

target-methylseq-qc: a lightweight pipeline for collecting 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Published: unpublished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 sequencing metrics. Here, we introduce the target-methylseq-qc pipeline (Sharma et al., 2024) , designed to (i) collects metrics from alignment files generated in targeted-methylation sequence analysis using the `picard_profiler` mode and (ii) filtering `bedGraph` for features overlapping with the reference BED file using the `bed_filter` mode, both of these modes are subworkflows written using the Nextflow (Di Tommaso et al., 2017) workflow language.

The target-methylseq-qc pipeline, when used in the `picard_profiler` mode 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. Upon completion of the analysis, a MultiQC report (Philip Ewels et al., 2016) will be generated, encompassing the updated sequencing coverage data for the targeted regions with some extras. The `picard_profiler` mode of the pipeline 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.

On the other hand, `bed_filter` mode of the pipeline is designed to accommodate the use-case of filtering bedGraph files as per the reference bed panel, such as Twist Human Methylome panel and best practices *Twist Methylome* (2016b) using bedtools (Quinlan & Hall, 2010) filter command. FIXME (**?**) can you help explain better the downstream usage of these files?

Regardless of the usage mode of the pipeline, the final MultiQC report automatically collates the relevant reports from FastQC (Andrews, 2010), Bedtool and Picard tools in an HTML document, which could be shared with collaborators or added as supplementary material in publications.

target-methylseq-qc is a portable pipeline compatible with multiple platforms, such as local laptop or workstation machines, high-performance computing environments and cloud infrastructure. Although target-methylseq-qc was originally created for calculating coverage in target sequencing as a follow-up step to the `nf-core/methylseq` pipeline (Phil Ewels et al., 2024), 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.

Design principles and capabilities

The target-methylseq-qc 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 (Philip Ewels et al., 2016) , Bedtools, 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 target-methylseq-qc 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.

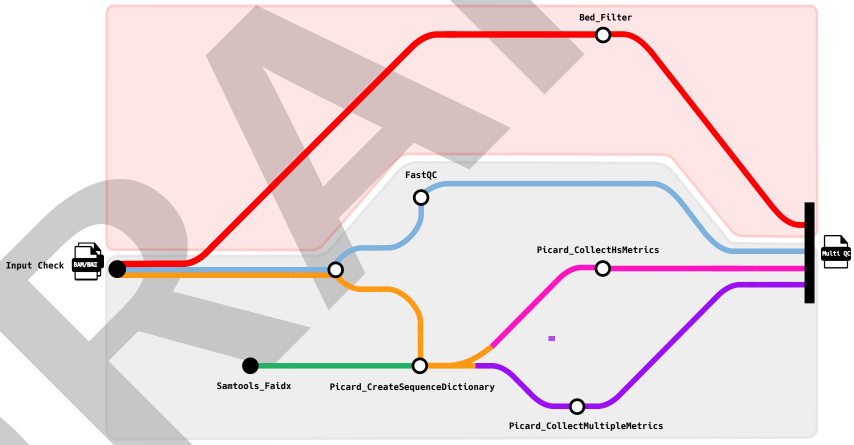


Figure 1: Subway map for various steps in the target-methylseq-qc pipeline.

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.

Tutorials and documentation

The steps needed to configure the pipeline inputs and configuration for the relevant infrastructure are available in the documentation within the Github repository as well as a dedicated documentation website (*Target-Methylseq-Qc Website*, 2024).

Pre-requisites

To ensure proper operation of the target-methylseq-qc pipeline, three dependencies must be available in the execution environment: Java (LTS $>$ 11), Nextflow ($>$ 24.04), and a package manager such as conda (Gruning et al., 2018) or a container system such as docker or singularity (Veiga Leprevost et al., 2017).

Getting started with the pipeline setup is straightforward given that (i) Java (LTS $>$ 11) (ii) Nextflow ($>$ 24.04) and (iii) a package manager (e.g. conda) or a container system (e.g. docker or singularity) 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.

Pipeline installation

target-methylseq-qc pipeline can be downloaded from the GitHub code repository using the git command line tool or directly through using the Nextflow command line tool using the following commands

```
# Git based download
$ git clone github https://github.com/wal-yan/target-methylseq-qc

# Nextflow based download
$ nextflow pull https://github.com/wal-yan/target-methylseq-qc
```

Test profiles

Two built-in test profiles are available in target-methylseq-qc pipeline for each mode of execution. These profiles can be used to run tests on the relevant infrastructure using the bundled test datasets (Agudelo-Romero, 2024), helping users to identify and resolve any infrastructural issue before the analysis stage.

```
# picard_profiler mode
$ nextflow run wal-yan/target-methylseq-qc \
  -profile docker,test_picard_profiler


# bed_filter mode
$ nextflow run wal-yan/target-methylseq-qc \
  -profile docker,test_bed_filter
```

Input

Following the convention for standard input in the Nextflow pipelines, target-methylseq-qc expects a CSV samplesheet as an input with the following fields.

Table 1: An example of a samplesheet for target-methylseq-qc in picard-profiler mode containing three columns, capturing the (i) name of the sample (ii) path to BAM file and (iii) path to the BAM index (BAI) file.

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

sample	bam	bai

Whereas the `bed_filter` mode requires a different set of columns in the input samplesheet CSV file, as shown in Table

sample	bedGraph
sample-01	/path/to/sample-01.bedGraph
sample-02	/path/to/sample-02.bedGraph

Execution

The pipeline initialization step, as per the best practices of the nf-core template, checks the validity of the file paths specified to be either a POSIX compliant file system or a cloud object storage path for files storaged in AWS S3, Azure Blob Storage or Google Cloud Storage buckets.

The behaviour of the pipeline can be controlled through the pipeline parameters which are divided into different groups such as (i) Execution Mode, (ii) Input/Output Options (iii) Reference Genome Options in addition to the generic parameters inherited from the nf-core template such as (i) Max job request options (ii) Generic options and (iii) Institutional config options. A complete list of the parameters specific to target-methylseq-qc pipeline are summarised in Table .

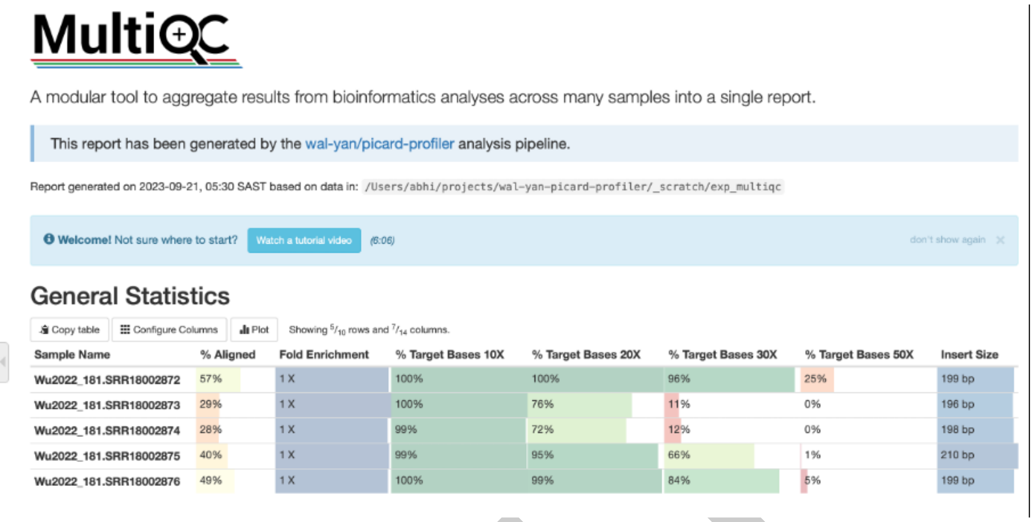
Parameter Name	Description
picard_profiler	Enable this boolean option to use the picard_profiler subworkflow
bed_filter	Enable this boolean option to use the bed_filter subworkflow
input	Path to comma-separated file containing information about the samples in the experiment.
outdir	The output directory where the results will be saved.
ref_fasta	Path to FASTA genome file.
ref_fai	Path to the FASTA index file.
ref_bed	Path to the BED file for the reference panel.

Output

Upon completion, the pipeline generates a MultiQC file with the relevant results of the analysis Figure 2.

Upon completion, the two subworkflows generate different outputs which are presented together in the MultiQC file. For picard_profile mode, a MultiQC file is produced, providing the relevant results related to the coverage metrics (Figure 2A). For the bed_filter mode mode, a BED file is generated with the methylation positions filtered based on the BED intervals file from the targeted methylation profile (Figure 2B).

Figure 2: Examples of the target-methylseq-qc pipeline modes. (A) MultiQC report generated for picard-profiler mode, highlighting refined metrics from targeted sequencing at 10X, 20X, 30X and 50X coverage. (B) Filtered BED file produced after run Bed_profiler mode.



A modular tool to aggregate results from bioinformatics analyses across many samples into a single report.

This report has been generated by the wal-yan/picard-profiler analysis pipeline.

Report generated on 2023-09-21, 05:30 SAST based on data in: `/Users/abhi/projects/wal-yan-picard-profiler/_scratch/exp_multiqc`

Welcome! Not sure where to start?

General Statistics

Showing 5/10 rows and 7/14 columns.

Sample Name	% Aligned	Fold Enrichment	% Target Bases 10X	% Target Bases 20X	% Target Bases 30X	% Target Bases 50X	Insert Size
Wu2022_181.SRR18002872	57%	1 X	100%	100%	96%	25%	199 bp
Wu2022_181.SRR18002873	29%	1 X	100%	76%	11%	0%	196 bp
Wu2022_181.SRR18002874	28%	1 X	99%	72%	12%	0%	198 bp
Wu2022_181.SRR18002875	40%	1 X	99%	95%	66%	1%	210 bp
Wu2022_181.SRR18002876	49%	1 X	100%	99%	84%	5%	199 bp

Figure 2: MultiQC report generated for target-methylseq-qc, in `picard-profiler` highlighting the refine metrics from targeted sequencing at 10X, 20X, 30X and 50X coverage.

Funding Statement

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

References

Agudelo-Romero, P. (2024). *Wal-yan/target-methylseq-qc test-data* [Data set]. Zenodo. https://doi.org/10.5281/zenodo.13601364

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, Phil, Hüther, P., Miller, E., Sateesh_Peri, Spix, N., bot, nf-core, Peltzer, A., F., S., Alneberg, J., Garcia, M. U., Krueger, F., Tommaso, P. D., Hörtenhuber, M., Ajith, V., Davenport, C., Patel, H., Salam, W., Cochetel, N., Alessia, … Céline, N. (2024). *Nf-core/methylseq: Huggy mollusc* (Version 2.6.0). Zenodo. https://doi.org/10.5281/zenodo.10463781

Ewels, Philip, 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

Gruning, B., Dale, R., Sjodin, A., Chapman, B. A., Rowe, J., Tomkins-Tinch, C. H., Valieris, R., Koster, J., & Bioconda, T. (2018). Bioconda: Sustainable and comprehensive software distribution for the life sciences [Journal Article]. *Nat Methods*, *15*(7), 475–476. https://doi.org/10.1038/s41592-018-0046-7

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/

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841–842. https://doi.org/10.1093/bioinformatics/btq033

Sharma, A., Conradie, T., Martino, D., Stick, S., & Agudelo-Romero, P. (2024). *Wal-yan/target-methylseq-qc* (Version v2.0.0). Zenodo. https://doi.org/10.5281/zenodo.13147688

Target-methylseq-qc website. (2024). https://wal-yan.github.io/target-methylseq-qc/usage.html

Twist methylome. (2016a). https://www.twistbioscience.com/products/ngs/fixed-panels/human-methylome-panel

Twist methylome. (2016b). https://www.twistbioscience.com/resources/technical-note/analyzing-twist-targeted-methylation-sequencing-data-using-twist-human

Veiga Leprevost, F. da, Gruning, B. A., Alves Aflitos, S., Rost, H. L., Uszkoreit, J., Barsnes, H., Vaudel, M., Moreno, P., Gatto, L., Weber, J., Bai, M., Jimenez, R. C., Sachsenberg, T., Pfeuffer, J., Vera Alvarez, R., Griss, J., Nesvizhskii, A. I., & Perez-Riverol, Y. (2017). BioContainers: An open-source and community-driven framework for software standardization [Journal Article]. *Bioinformatics*, *33*(16), 2580–2582. https://doi.org/10.1093/bioinformatics/btx192