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** [4,5], **and Patricia Agudelo-Romero** [2,6,7]

1 DSI-NRF Centre of Excellence for Biomedical Tuberculosis Research; SAMRC Centre for Tuberculosis Research; Division of Molecular Biology and Human Genetics, Faculty of Medicine and Health Sciences, Stellenbosch University, Cape Town, South Africa. **2** Wal-yan Respiratory Research Centre, The Kids Research Institute Australia, 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.

Summary

Next-generation targeted genome sequencing allows the analysis of regions of interest within a genome. While it is possible to incorporate targeted sequencing into whole-genome sequencing (WGS) bioinformatics pipelines, there remains a gap in accurately converting WGS metrics into precise target sequencing metrics and filtering the raw BED files to the targeted regions. Here, we introduce the `target-methylseq-qc` pipeline (https://doi.org/10.5281/zenodo.13147688), designed to (i) collect metrics from alignment files generated in targeted methylation sequence analysis, using the `picard_profiler` mode, and (ii) filtering `bedGraph` files to 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 focused on the target regions, the inclusion of a FASTA reference file and BED intervals file is required. Upon completion of the analysis, a MultiQC report (Ewels et al., 2016) will be generated, encompassing the updated sequencing coverage data for the targeted regions along with reports from other tools as well. 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, the `bed_filter` mode of the pipeline is designed to filter the bedGraph files generated by nf-core/methylseq (Ewels et al., 2024), using a reference BED file from the sequencing panel. For our pipeline, we utilized the Twist Human Methylome panel (https://www.twistbioscience.com/resources/data-files/twist-human-methylome-panel-target-bed-file) for filtering the relevant reads using the bedtools filter command (Quinlan & Hall, 2010), as per best practices (*Twist Methylome, 2016a, 2016b*). Filtering raw BED files with the targeted regions is crucial because it ensures that the analysis focuses on specific

genomic targets accurately and efficiently. This step minimizes the inclusion of off-target sequences and reduces the potential for including sequencing artifacts, which can be introduced during capture-based targeted sequencing processes. Downstream analyses from the filtered BED files will enable the calculation of CpG ratios and the testing for differentially methylated cytosines (DMCs) or regions (DMRs).

Regardless of the usage mode of the pipeline, the final MultiQC report automatically collates the relevant reports from FastQC (Andrews, 2010), bedtools 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 execution platforms, such as local laptop or workstation machines, high-performance computing environments and cloud infrastructure. Although, `target-methylseq-qc` pipeline was originally created for calculating sequencing coverage in target sequencing as a follow-up step to the nf-core/methylseq pipeline (Ewels et al., 2024), with the initial objective of identifying children at risk of developing asthma before onset (Agudelo-Romero et al., 2024), the pipeline's versatility extends its application. It can be used with other sequencing panels from various next-generation methods.

Statement of need

The `target-methylseq-qc` pipeline is designed to streamline the quality control process for target methylation sequencing data. Researchers and bioinformaticians working with methylation sequencing data often face challenges in ensuring data quality and consistency across different samples and experiments. This pipeline addresses these challenges by providing a standardized and automated workflow for quality control, leveraging the capabilities of the nf-core framework. Key features of the `target-methylseq-qc` pipeline include (i) standardized input format: The pipeline expects a CSV samplesheet with specific fields tailored to different modes (`picard_profiler` and `bed_filter`), ensuring consistency and ease of use, (ii) flexible execution modes: Users can choose between different subworkflows (`picard_profiler` and `bed_filter`) based on their specific needs, enabling tailored quality control processes, (iii) comprehensive parameter control: Users can fine-tune the pipeline's behavior through a wide range of parameters, covering execution modes, input/output options, reference genome options, and infrastructural configuration. By automating and standardizing the quality control process, the `target-methylseq-qc` pipeline helps researchers save time, reduce errors, and ensure high-quality data for downstream analysis and clinically applicable insights.

Design principles and capabilities

The `target-methylseq-qc` pipeline builds upon the standardized pipeline template maintained by the nf-core community (Ewels et al., 2020) for Nextflow pipelines and makes use of the `nf-core/modules` project to install modules for FastQC, MultiQC (Ewels et al., 2016), bedtools, Picard tools as well as Samtools (Danecek et al., 2021) within the pipeline (Figure 1).

The use of the nf-core template facilitates keeping the design of the pipeline generic and portable across different execution platforms. Hence, 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.
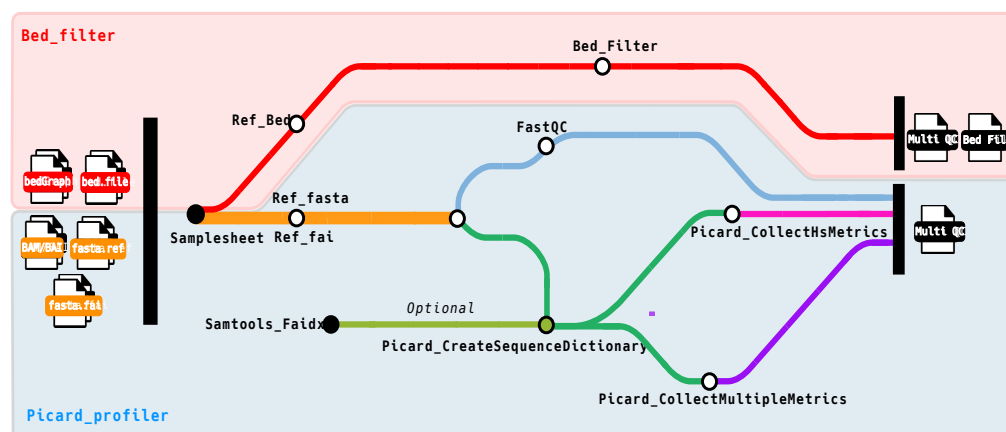
Figure 1: Subway map for various steps in the target-methylseq-qc pipeline.

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 of the users, the MultiQC configuration allows for the metrics collection for 10x, 20x, 30x and 50x coverage.

Input

Following the convention for standard input in Nextflow pipelines, `target-methylseq-qc` expects a CSV samplesheet as an input. Table 1 shows an example samplesheet for `target-methylseq-qc` in `picard-profiler` mode. The samplesheet contains three columns, capturing (i) the name of the sample, (ii) the path to BAM file, and (iii) the path to the BAM index (BAI) file.

Table 1: Samplesheet structure for `picard_profiler` mode .

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

The `bed_filter` mode requires a different set of columns in the input samplesheet CSV file, as shown in Table 2.

Table 2: Samplesheet structure for `bed_filter` mode .

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 stored in AWS S3, Azure Blob Storage or Google Cloud Storage buckets.

The behavior 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, and (iii) Reference Genome Options. In addition, generic parameters are inherited from the nf-core template, including (i) max job request options, (ii) generic options and (iii) institutional config options. A complete list of the parameters specific to the `target-methylseq-qc` pipeline is summarized in Table 3.

Table 3: Summary of pipeline-specific parameters for target-methylseq-qc pipeline .

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 two subworkflows generate different outputs which are presented together in the MultiQC file. For the `picard_profile` mode, a MultiQC file is produced, providing the relevant results related to the coverage metrics (Figure 2). For the `bed_filter` mode, a BED file is generated with the methylation positions filtered based on the BED intervals file from the targeted methylation profile (Figure 3).

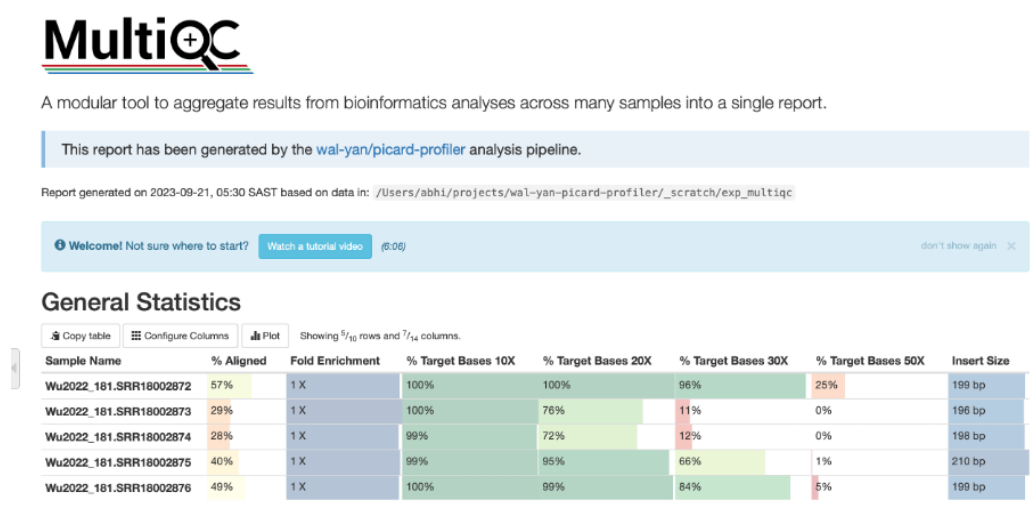**Figure 2:** MultiQC report generated for target-methylseq-qc, in `picard-profiler` highlighting the refine metrics from targeted sequencing at 10x, 20x, 30x and 50x coverage.

Figure 3: Filtered bedGraph file generated using the `bed_filter` mode of target-methylseq-qc.

Funding Statement

References

Agudelo-Romero, P., Iosifidis, T., Kicic-Starcevich, E., Hancock, D., Caparros-Martin, J., Martino, D., & Stick, S. (2024). Nasal and amnion methylomes: Biomarkers for smoke exposure and maternal asthma [Conference Proceedings]. *Journal of Allergy and Clinical Immunology*, *153*, AB243. https://doi.org/10.1016/j.jaci.2023.11.779

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

Ewels, P. A., 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

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

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

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

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