

Apr 09, 2025

Version 2

TwistMethNext V.2

DOI

<https://dx.doi.org/10.17504/protocols.io.n92ld59rov5b/v2>

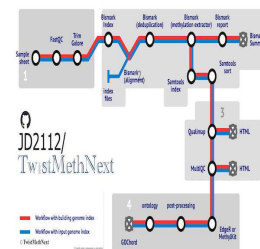
Jyotirmoy Das¹

¹Linköping University



Jyotirmoy Das

Linköping University



Create & collaborate more with a free account

Edit and publish protocols, collaborate in communities, share insights through comments, and track progress with run records.

Create free account

OPEN ACCESS



DOI: <https://dx.doi.org/10.17504/protocols.io.n92ld59rov5b/v2>

External link: <https://github.com/JD2112/TwistMethNext>

Protocol Citation: Jyotirmoy Das 2025. TwistMethNext. protocols.io

<https://dx.doi.org/10.17504/protocols.io.n92ld59rov5b/v2> Version created by [Jyotirmoy Das](#)

License: This is an open access protocol distributed under the terms of the **Creative Commons Attribution License**, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited



Protocol status: Working

We use this protocol and it's working

Created: April 08, 2025

Last Modified: April 09, 2025

Protocol Integer ID: 126395

Keywords: Twist NGS, DNA methylation, NextFlow, Singularity, Docker, Conda, analysis of twist ngs dna methylation data, twist ngs dna methylation data, twistmethnext dna methylation, dna methylation analysis, genome index, methylkit for differential methylation analysis, fly reference genome indexing, identification of methylation signature, built genome index, methylation signature, bioinformatics pipeline, differential methylation analysis, based bioinformatics pipeline, sequencing data necessitate, genome, crucial epigenetic modification, methylation calling, methylkit, twistmethnext, making twistmethnext, sequencing, gene expression, gene

Disclaimer

TwistMethNext makes use of publicly available open-source software and tools for genomic data processing and analysis. These tools are included and referenced under the terms of their respective licenses.

While we aim to ensure that all components work correctly and reliably, we provide this pipeline as-is, without warranty of any kind. The user is responsible for verifying the accuracy and suitability of the results for their specific use case.

Included Tools & Licenses

This pipeline integrates several third-party tools (e.g., FastQC, Trim Galore, Bismark, methylKit, etc.), each of which is governed by its own license agreement. Please refer to the official documentation of each tool for licensing and usage terms.

No Liability

The authors and maintainers of this pipeline assume no responsibility or liability for any errors or inaccuracies that may arise from the use of the software. By using this pipeline, you acknowledge that:

You are using publicly available tools at your own discretion.

You are responsible for complying with any applicable licensing agreements.

You understand that performance may vary depending on computing environments and input data.

User Responsibility

We strongly recommend users to:

- Review and cite the individual tools used in your analysis.
- Validate outputs through independent quality control measures.
- Contact respective tool developers for support regarding their software-specific issues.

Abstract

DNA methylation is a crucial epigenetic modification that regulates gene expression and is associated with various diseases, including cancer. The advent of next-generation sequencing (NGS) technologies has revolutionized the identification of methylation signatures on a large scale. However, the vast volume of sequencing data necessitates well-structured analytical workflows to ensure reproducibility and scalability. We introduce TwistMethNext, an end-to-end Nextflow-based bioinformatics pipeline that streamlines the analysis of Twist NGS DNA methylation data. TwistMethNext provides a comprehensive solution for quality control, read alignment, data normalization, methylation calling, differential methylation analysis, and visualization. The pipeline supports both on-the-fly reference genome indexing and the use of pre-built genome indexes, ensuring flexibility and efficiency. It incorporates robust tools such as edgeR and methylKit for differential methylation analysis and performs functional enrichment analysis using the clusterProfiler package. TwistMethNext is designed to be highly customizable, allowing users to configure input files, pipeline starting points, and skip various workflows as needed. This modular approach ensures reproducibility, scalability, and containerization, making TwistMethNext a valuable tool for researchers in the field of DNA methylation analysis.

Image Attribution

Metromap of the TwistMethNext pipeline.



Guidelines

```
#!/bin/bash
#SBATCH --job-name=TwistMethNext
#SBATCH --time=3-00:00:00 # 3 days (D-HH:MM:SS)
#SBATCH --cpus-per-task=16
#SBATCH --output=twistmethnext%j.out
#SBATCH --error=twistmethnext%j.err

set -e

# Load modules
module load singularity-4.1.1
module load nextflow-24.04.4

# Set Nextflow and Singularity directories
export NXF_HOME=/mnt/data/
export NXF_WORK=/mnt/data/nextflow_work
export SINGULARITY_CACHEDIR=/mnt/data/singularity_cache

# Ensure directories exist
mkdir -p $NXF_WORK $SINGULARITY_CACHEDIR

# Run Nextflow
nextflow run TwistNext \
  -profile singularity \
  --sample_sheet samplesheet.csv \
  --genome_fasta /mnt/reference/Reference_Genomes/ucsc_hg38/hg38.fa \
  --run_both_methods \
  --refseq_file hg38_RefSeq.bed.gz \
  --gtf_file Homo_sapiens.GRCh38.104.gtf \
  --outdir Results_TwistMethNext
```



Materials

A. Software Dependencies:

- Nextflow >22.04
- Singularity ≥ 3.8
- Docker ≥ 24
- Conda
- Java (≥ 8)

B. Hardware Requirements:

High-performance computing environment with sufficient memory and processing power to handle large NGS datasets.

Per process requirements -

- **cpus:** 12 cores
- **RAM:** 6 GB - 200 GB
- **Computational time:** ~2days for 24 paired-end samples
- **Storage:** ~3 TB for 24 paired-end samples

Safety warnings

- ❗ 1. Bismark Alignment is a computationally heavy step and requires long time to complete the process.
- 2. Running NextFlow pipeline directly from terminal in a single node could cause all available memory and cores and system can get slow or unresponsive.

Before start

To run the command, users need to make sure -

- NextFlow (≥ 22.04) is available;
- JAVA (≥ 8) is loaded properly;
- Singularity/Aptainer and/or Docker is ready to use;



NextFlow command

1 One-step solution

Command

run nextflow (Ubuntu 24.04.2 LTS)

```
nextflow run JD2112/TwistMethNext \  
  -profile singularity \  
  --sample_sheet Sample_sheet_twist.csv \  
  --genome_fasta /data/reference_genome/hg38/hg38.fa \  
  --run_both_methods \  
  --gtf_file /data/Homo_sapiens.GRCh38.104.gtf \  
  --refseq_file /data/hg38_RefSeq.bed.gz \  
  --outdir Results/TwistMethNext_both
```

Step-by-step procedure

2 **STEP-1. Quality Control and Trimming:** Assess the quality of raw sequencing data

2.1 **FastQC:** Raw Data Quality Control

FASTQC (Andrews ([2010](#))) is a widely used tool for assessing the quality of raw and processed sequencing data. It provides a comprehensive quality check, including metrics like per-base quality scores, GC content, and adapter contamination.

```
fastqc $args --threads $task.cpus $reads
```

2.2 **Trim galore:** Adapter Sequence Trimming

Trim Galore (Felix Krueger et al. ([2023](#))) is a versatile tool for trimming sequencing reads and removing adapter sequences. It's particularly useful for preparing raw sequencing data for downstream applications like alignment or differential expression/methylation

analysis. Trim Galore combines the functionalities of Cutadapt and FastQC for quality control and trimming.

```
trim_galore --paired --cores $task.cpus $args $reads
```

3 **STEP-2. Bismark Processes:** genome preparation, alignment, methylation calls

3.1 **Generate Reference Genome Index**

Bismark (F. Krueger and Davies ([2019](#))) needs to prepare the bisulfite index for the genome. In the current pipeline, user can provide the genome.fast. and the pipeline can index it. Optionally, user can provide the index files directly, and the pipeline will use it without indexing the genome again.

```
bismark_genome_preparation --bowtie2 --parallel 4 <genome.fasta>
```

3.2 **Alignment:** Align Reads

This step aligns bisulfite-treated sequencing reads to a reference genome.

```
bismark --genome <path_to_reference_genome> -1 <reads_R1.fq> -2  
<reads_R2.fq> -o <output_directory>
```

3.3 **Deduplicate Alignments**

This step removes duplicate reads to avoid overestimating methylation levels.

```
deduplicate_bismark ${paired_end} $args --bam $bam
```

3.4 **Sort and Index Alignments**

Use Samtools to sort and index the aligned reads for downstream analysis.

```
samtools sort $args -@ $task.cpus -o ${prefix}.sorted.bam $bam
```

3.5 Extract Methylation Calls

Extract methylation data from deduplicated BAM files.

```
bismark_methylation_extractor  
--bedGraph --gzip  
-o <output_directory>  
<deduplicated.bam>
```

3.6 Bismark Summary Reports

Generate a summary report of alignment and methylation statistics.

```
bismark2report
```

4 STEP-3: Post-processing Quality Checks

4.1 QualiMap: Alignment Quality Control

The main module for assessing alignment quality is qualimap bamqc (Okonechnikov, Conesa, and García-Alcalde ([2016](#))).

```
qualimap bamqc  
-bam <input.bam>  
-outdir <output_directory>  
-outformat <html>  
--java-mem-size
```

4.2 MultiQC: Quality Control Reporting

MultiQC (Ewels et al. ([2016](#))) is used for the QC reporting combining all results from the FastQC, Trim galore, Bismark Alignment, Bismark Deduplication, Bismark summary report, and Qualimap results.

```
multiqc -f $args temp_multiqc_input > multiqc.log 2>&1
```

5 **STEP-4: Downstream Processing**

Differential Methylation Analysis

To calculate the differential methylation from the input samples, two different methods can be used -

EdgeR or MethylKit

5.1 **EdgeR**

EdgeR (Chen et al. ([2025](#))) is a R-based (Ihaka and Gentleman ([1996](#))) Bioconductor (Huber et al. ([2015](#))) package primarily used for RNA-seq differential expression analysis but can also handle differential methylation analysis when paired with bisulfite sequencing data. This requires pre-processed methylation data, such as counts of methylated (M) and unmethylated (U) reads at each cytosine position or region of interest.

```
Rscript $baseDir/bin/edgeR_analysis.R
--coverage_files '${coverage_files}'
--design "${design_file}"
--compare "${compare_str}"
--output .
--threshold ${coverage_threshold}
```

5.2 **MethylKit**

MethylKit (Morgan and Huber ([2020](#))) is an R package designed for analyzing bisulfite sequencing data, particularly for differential methylation analysis. It supports genome-wide methylation data and is ideal for CpG, CHH, and CHG methylation studies.

```
Rscript $baseDir/bin/run_methylkit.R
--coverage_files '${coverage_files}'
--design "${design_file}"
--compare "${compare_str}"
--output .
--threshold ${coverage_threshold}
```

5.3 **Post-Processing**

Perform additional analyses, such as visualization using ggplot2 (Wickham ([2016](#))) and statistical assessments, to interpret methylation data. The process generates A. Volcano Plot, B. MA Plot and C. Summary Statistics from the differential methylation analysis results.

```
Rscript ${workflow.projectDir}/bin/post_processing.R
--results ${results_file}
--compare "${compare_str}"
--output .
--method ${method}
--logfc_cutoff ${logfc_cutoff}
--pvalue_cutoff ${pvalue_cutoff}
--hyper_color "${hyper_color}"
--hypo_color "${hypo_color}"
--nonsig_color "${nonsig_color}"
```

5.4 Gene Ontology Analysis

The pipeline has also a module to perform the Gene Ontology analysis from the top n corresponding genes from the differential methylation results (EdgeR/MethylKit) using the *clusterProfiler* (Xu et al. ([2024](#))) and *GOChord* (Walter, Sánchez-Cabo, and Ricote ([2015](#))) package. The results generates a full table with all Biological Processes and a Chord diagram with top 10 functions identified in the analysis.

```
Rscript ${workflow.projectDir}/bin/go_analysis.R
--results ${results}
--output .
--method ${method}
--logfc_cutoff ${logfc_cutoff}
--pvalue_cutoff ${pvalue_cutoff}
--top_n ${top_n}
```

Protocol references

- Andrews, S. 2010. "FastQC: A Quality Control Tool for High Throughput Sequence Data." <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>.
- Chen, Y., L. Chen, A. T. L. Lun, P. Baldoni, and G. K. Smyth. 2025. "edgeR V4: Powerful Differential Analysis of Sequencing Data with Expanded Functionality and Improved Support for Small Counts and Larger Datasets." *Nucleic Acids Research* 53 (2): gkaf018. <https://doi.org/10.1093/nar/gkaf018>.
- Di Tommaso, P., M. Chatzou, E. W. Floden, et al. 2017. "Nextflow Enables Reproducible Computational Workflows." *Nat Biotechnol* 35 (4): 316–19. <https://doi.org/10.1038/nbt.3820>.
- Ewels, P., M. Magnusson, S. Lundin, and M. Käller. 2016. "MultiQC: Summarize Analysis Results for Multiple Tools and Samples in a Single Report." *Bioinformatics* 32 (19): 3047–48. <https://doi.org/10.1093/bioinformatics/btw354>.
- Grüning, B., R. Dale, A. Sjödin, et al. 2018. "Bioconda: Sustainable and Comprehensive Software Distribution for the Life Sciences." *Nat Methods* 15 (7): 475–76. <https://doi.org/10.1038/s41592-018-0046-7>.
- Huber, W., V. Carey, R. Gentleman, et al. 2015. "Orchestrating High-Throughput Genomic Analysis with Bioconductor." *Nat Methods* 12: 115–21. <https://doi.org/10.1038/nmeth.3252>.
- Ihaka, R., and R. Gentleman. 1996. "R: A Language for Data Analysis and Graphics." *Journal of Computational and Graphical Statistics* 5 (3): 299–314.
- Krueger, F., and R. Davies. 2019. "Bismark: A Versatile Tool for Aligning Sequencing Data to Bisulfite-Converted Genomes." *Bioinformatics* 35 (4): 623–25. <https://doi.org/10.1093/bioinformatics/bty724>.
- Krueger, Felix, Frankie James, Phil Ewels, Ebrahim Afyounian, Michael Weinstein, Benjamin Schuster-Boeckler, Gert Hulselmans, and sclamons. 2023. "TrimGalore: V0.6.10 - Add Default Decompression Path." <https://doi.org/10.5281/zenodo.7598955>.
- Kurtzer, G. M., V. Sochat, and M. W. Bauer. 2017. "Singularity: Scientific Containers for Mobility of Compute." *PLoS One* 12 (5): e0177459. <https://doi.org/10.1371/journal.pone.0177459>.
- Merkel, D. 2014. "Docker: Lightweight Linux Containers for Consistent Development and Deployment." *Linux Journal* 2014 (239): 2. <https://doi.org/10.5555/2600239.2600241>.
- Morgan, M., and W. Huber. 2020. "methyKit: A Package for the Analysis of DNA Methylation Data in R." *Bioinformatics* 36 (1): 214–16. <https://doi.org/10.1093/bioinformatics/btz262>.
- Okonechnikov, Konstantin, Ana Conesa, and Fernando García-Alcalde. 2016. "Qualimap 2: Advanced Multi-Sample Quality Control for High-Throughput Sequencing Data." *Bioinformatics* 32 (2): 292–94. <https://doi.org/10.1093/bioinformatics/btv566>.



Walter, Wencke, Fátima Sánchez-Cabo, and Mercedes Ricote. 2015. "GOplot: An r Package for Visually Combining Expression Data with Functional Analysis." *Bioinformatics*. <https://doi.org/10.1093/bioinformatics/btv300>.

Wickham, H. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. <https://ggplot2.tidyverse.org>.

Xu, S., E. Hu, Y. Cai, et al. 2024. "Using clusterProfiler to Characterize Multiomics Data." *Nature Protocols* 19 (11): 3292–3320. <https://doi.org/10.1038/s41596-024-01020-z>.

Acknowledgements

I would like to acknowledge the Core Facility, Faculty of Medicine and Health Sciences, Linköping University, Linköping, Sweden and Clinical Genomics Linköping, Science for Life Laboratory, Sweden for their support. We are grateful to PDC (KTH, Sweden) support for computational support to test and validate the pipeline on the Dardel HPC.