TAPS+ Data Analysis Guide

For use on data generated with the Watchmaker DNA Library Prep Kit with TAPS+

Doug Wendel, Tyler Matheny, Thomas Harrison, Andrew Gaines, Eduard Casas, Kristina Giorda, and Brian Kudlow — Watchmaker Genomics, Boulder, CO USA

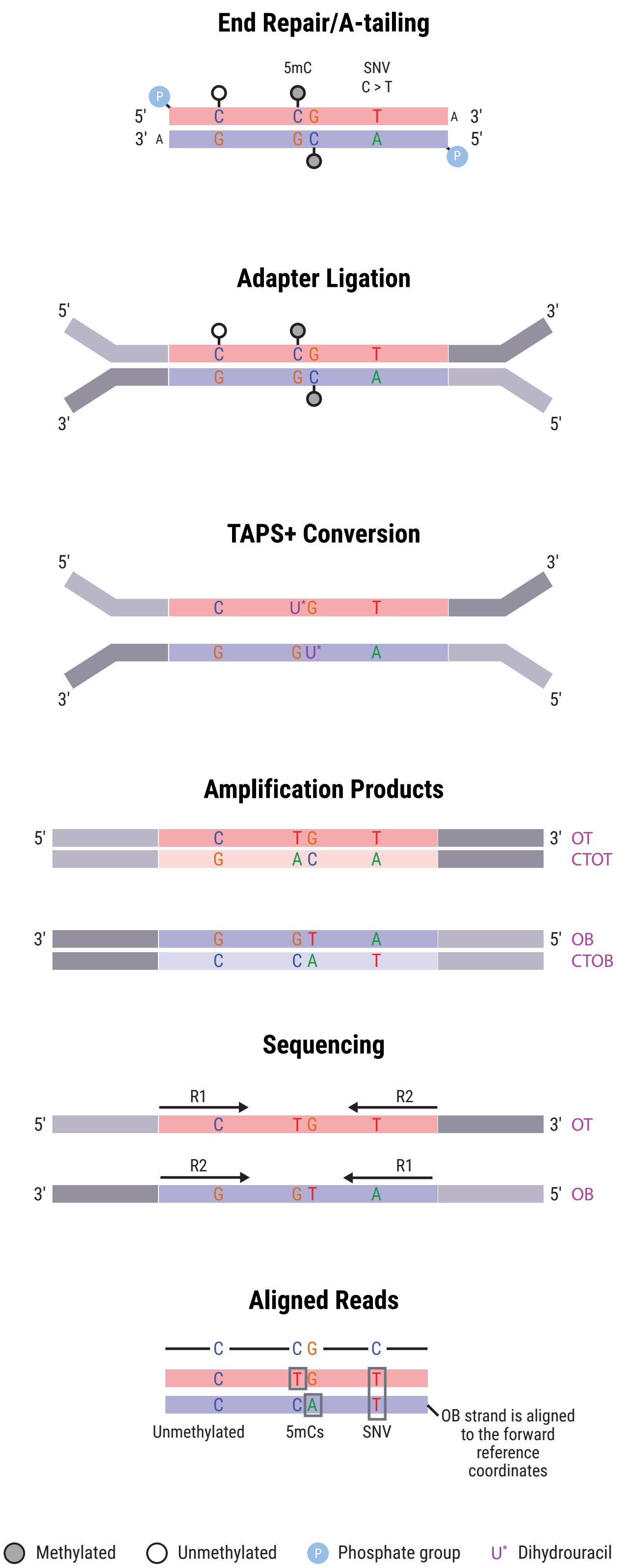
OVERVIEW

This guide provides an overview of methodologies for analyzing data generated with TAPS+ DNA methylation chemistry — a modified TAPS[1] (TET-assisted pyridine borane sequencing) approach. Unlike conventional bisulfite sequencing (BS-Seq) or enzymatic methylation sequencing (EM-Seq) methods, TAPS+ produces a positive readout by converting 5-methylcytosine (5mC) and 5-hydroxymethylcytosine (5hmC) to thymine, while unmethylated cytosines remain unchanged. This inversion requires analysis pipelines to interpret methylation signatures differently than with negative-readout methods (e.g., bisulfite treatment).

Figure 1 (on the following page) illustrates the core principles of TAPS+ library preparation and subsequent analysis. The schematic emphasizes critical considerations for methylation analysis, including end-repair-associated bias (mbias) and strandedness, both of which are essential for accurate methylation calling. It also shows how TAPS+ preserves sequence complexity for variant detection while simultaneously enabling base-resolved methylation analysis.

The guide is intended for bioinformaticians and researchers familiar with standard NGS analysis tools and workflows. It provides options ranging from full-featured pipelines to specialized methylation callers, enabling accurate interpretation of TAPS+ data across diverse applications.

TABLE OF CONTENTS



A. End Repair/A-tailing: During end repair, polymerase activity fills in 5' overhangs to create blunt ends, producing mbias at the 3' ends. After which, an A-base is added to the 3' ends to prepare for adapter ligation.

B. Adapter Ligation: Sequencing adapters are ligated to the A-tailed ends, adding priming sequences for later amplification.

C. TAPS+ Conversion: TAPS+ converts 5mC and 5hmC to dihydrouracil (DHU), leaving other cytosines unconverted.

D. Amplification: During PCR amplification, DHUs are read as thymine (T), while unmethylated cytosines remain as cytosines (C), enabling differentiation between methylated and unmethylated cytosines.

E. Sequencing: Bi-directional sequencing is performed, generating reads from both the original strands (original top [OT] and original bottom [OB]) and their complementary strands (complementary to the original top [CTOT] and complementary to the original bottom [CTOB]).

F. Alignment and Analysis: The reads are aligned to a reference genome. Strand information and C-to-T conversions in the aligned reads allow for the identification of methylated cytosines (5mC and 5hmC) based on TAPS conversion, distinguishing them from unmethylated cytosines and single nucleotide variants (SNVs).

Figure 1. Overview of the Watchmaker DNA Library Prep with TAPS+ process.

TAPS+ ANALYSIS OPTIONS OVERVIEW

Several software tools are available for analyzing TAPS+ data. Users may adopt existing pipelines or adapt custom workflows, provided that TAPS+ base conversion is accounted for.

- **nf-core/methylseq** – Best-practice, production-ready pipeline for most users. Supports multiple methylation chemistries, including TAPS+.
- **rastair** – Optimized for TAPS+. Efficient CpG-only methylation caller with variant calling support at CpG sites (C > T and G > A calls only).
- **asTair** – Flexible tool developed for TAPS+. Supports additional genomic contexts beyond CpGs; integrates easily into custom pipelines.
- **Illumina DRAGEN Methylation Pipeline** – High-performance, hardware-accelerated workflow available via Illumina's BaseSpace platform.

Each method has distinct advantages and limitations, as shown in Table 1. The following sections provide detailed instructions for each method.

Table 1. Overview of methods

Tool	Scope	Strengths	Limitations	Recommended use case
nf-core/methylseq	Full pipeline (TAPS+/BS-Seq/EM-Seq)	Community-maintained, reproducible, production-ready	Requires Nextflow; CpG-only context for TAPS+	Best general-purpose option
rastair	CpG context only (TAPS+)	Native TAPS support, efficient execution, optimized mbias handling, CpG variant calling	Limited context support	Targeted CpG methylation analysis
asTair	TAPS+/BS-Seq/multi-context*	Flexible, pipeline integration	Slower, unmaintained	Multi-context workflows, custom pipelines
DRAGEN	TAPS+/BS-Seq/EM-Seq	Hardware-accelerated, Illumina integration	Proprietary, BaseSpace-dependent	High-throughput DRAGEN environments

*Such as CHG and CHH contexts

nf-core/methylseq

Description

nf-core/methylseq is a community-maintained pipeline built on Nextflow. Originally developed for BS-Seq and EM-Seq, it now natively supports TAPS+. The pipeline integrates alignment, methylation calling, and QC steps, ensuring reproducible and scalable execution across local, HPC, and cloud environments (e.g., AWS, GCP, and Azure).

Advantages

- Unified support for TAPS+, BS-seq, and EM-seq
- Scalable, containerized execution with robust MultiQC reporting
- Community-maintained and aligned with best practices
- Reduces need for custom scripting

Limitations

- Less modularity for custom workflows

Procedure

1. Prepare and sequence libraries using the Watchmaker DNA Library Prep Kit with TAPS+.
2. Install Nextflow and (optionally) nf-core CLI tools.

```
# Install Nextflow
curl -s https://get.nextflow.io | bash

# Install nf-core CLI tools (recommended but not required)
pip install nf-core
```

3. Run nf-core/methylseq with TAPS+ parameters:

```
nextflow run nf-core/methylseq \
  --taps \
  --aligner bwamem \
  --genome GRCh38 \
  --input samplesheet.csv\
  --outdir ./results \
  -profile docker
```

For detailed control over execution, please refer to the TAPS section of the nf-core/methylseq documentation here: https://nf-co.re/methylseq.

rastair

Description

rastair is a stand-alone caller developed specifically for TAPS+ in the Schuster-Boeckler group at the Ludwig Institute for Cancer Research.[2] It is efficient, optimized for CpG methylation contexts, and includes logical handling of read masking and mbias. rastair also supports limited variant calling (C>T / G>A at CpG sites).

Note: A stand-alone data analysis script covering steps 3 through 7 is available in our GitHub repository.[3]

Advantages

- Native support for TAPS+ libraries
- Open source and broadly installable
- Efficient execution with streamlined resource usage
- Optimized handling of read masking and mbias
- Integrated CpG-specific variant calling

Limitations

- CpG-only context support
- Assumes a diploid genome for variant calling

Procedure

1. Prepare and sequence libraries using the Watchmaker DNA Library Prep Kit with TAPS+.
2. Install rastair via Docker.
   a. Documentation and installation instructions for rastair can be found here: https://bitbucket.org/bsblabludwig/rastair/src/v0.8.2/.

rastair has an official Dockerhub image:

```
docker pull sbludwig/rastair:version-0.8.2
```

Alternatively, you may choose to install rastair natively:

```
# Install cargo appropriate for your system:
# https://www.rust-lang.org/tools/install
cargo install --git https://bitbucket.org/bsblabludwig/rastair.git --tag v0.8.2
```

3. Trim adapters as necessary, e.g., (replace adapter sequences as appropriate):

```
cutadapt \
    --cores=$CPUS \
    -a AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC \
    -A AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTA \
    -o ${WORK_DIR}/${PREFIX}_R1_001_trimmed.fastq.gz \
    -p ${WORK_DIR}/${PREFIX}_R2_001_trimmed.fastq.gz \
    ${PREFIX}_R1_001.fastq.gz \
    ${PREFIX}_R2_001.fastq.gz \
    > ${WORK_DIR}/cutadapt.log
```

4. Align reads (e.g., with BWA-MEM).
   a. Replace the reference genome with one appropriate for your samples.

```
bwa mem \
    -t $CPUS \
    /path/to/reference_genome.fa \
    ${PREFIX}_R1_001_trimmed.fastq.gz \
    ${PREFIX}_R2_001_trimmed.fastq.gz \
    | samtools view --threads $CPUS -o ${PREFIX}_trimmed.bam -
```

5. Sort, remove duplicates, and filter the resulting BAM, for example:

```
samtools sort \
    -o ${PREFIX}_trimmed_sorted.bam \
    -O bam \
    -T ${PREFIX}_trimmed.bam.temp \
    ${PREFIX}_trimmed.bam

gatk \
    --java-options "-Xmx${MEMORY}g" \
    MarkDuplicates \
    --INPUT ${PREFIX}_trimmed_sorted.bam \
    --OUTPUT ${PREFIX}_trimmed_sorted_marked.bam \
    --METRICS_FILE ${PREFIX}_trimmed_sorted_marked.metrics \
    --TMP_DIR .

samtools view \
    -b \
    -f 3 \
    -F 1024 \
    -F 2048 \
    -q ${MAPQ} \
    -o ${OUTPUT_FILE}.bam \
    ${INPUT_FILE}
```

6. Perform mbias analysis to determine trimming parameters.
   a. This step can be run prior to methylation calling and is usually helpful in determining an appropriate mbias mask:

```
rastair mbias \
    --threads ${CPUS} \
    --fasta-file ${fasta} \
    ./${PREFIX}_trimmed_sorted_marked_deduped.bam >
${PREFIX}_trimmed_sorted_marked_deduped_rastair_mbias.txt
```

7. Call methylation using optimized parameters for your dataset.
   a. Adjust input values as appropriate. `-nOT/-nOB` defaults are suggested; however, we recommend substituting these values with library-specific cutoffs based on the results of the mbias analysis outlined above:

rastair

```
rastair call \
    --nOT 0,0,20,0 \ # r1_start, r1_end, r2_start, r2_end
    --nOB 0,0,20,0 \ # r1_start, r1_end, r2_start, r2_end
    --min-baseq 30 \
    --min-mapq 20 \
    --fasta-file /path/to/reference_genome.fa \
    ${PREFIX}_trimmed_sorted_marked_deduped.bam \
    > rastair_output.mods
```

asTair

Description

asTair was developed alongside TAPS+ chemistry and supports both TAPS+ and BS-seq libraries. It allows analysis of multiple genomic contexts beyond CpGs and integrates smoothly into custom bioinformatics pipelines. While functional, this software is no longer maintained, and we only recommend its use when methylation contexts beyond CpG are required.

Note: A stand-alone data analysis script covering steps 3 through 7 is available in our GitHub repository.[3]

Advantages

- Supports CpG, CHG, and CHH contexts
- Open source and broadly installable
- Good for pipeline integration
- Supports single-end workflows

Limitations

- Significantly slower than rastair for TAPS+ workflows
- Functional but not actively updated or maintained

Procedure

1. Prepare and sequence libraries using the Watchmaker DNA Library Prep Kit with TAPS+.
2. Install asTair, via pip.

```
pip install astair==3.3.2 # The version is important
```

3. Trim adapters as necessary, e.g., (replace adapter sequences as appropriate):

```
cutadapt \
    --cores=$CPUS \
    -a AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC \
    -A AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTA \
    -o ${WORK_DIR}/${PREFIX}_R1_001_trimmed.fastq.gz \
    -p ${WORK_DIR}/${PREFIX}_R2_001_trimmed.fastq.gz \
    ${PREFIX}_R1_001.fastq.gz \
    ${PREFIX}_R2_001.fastq.gz \
    > ${WORK_DIR}/cutadapt.log
```

4. Align reads (e.g., with BWA-MEM).
   a. Replace the reference genome with one appropriate for your samples.

```
bwa mem \
    -t $CPUS \
    /path/to/reference_genome.fa \
    ${PREFIX}_R1_001_trimmed.fastq.gz \
    ${PREFIX}_R2_001_trimmed.fastq.gz \
    | samtools view --threads $CPUS -o ${PREFIX}_trimmed.bam -
```

5. Sort, remove duplicates, and filter the resulting BAM, for example:

```
samtools sort \
    -o ${PREFIX}_trimmed_sorted.bam \
    -O bam \
    -T ${PREFIX}_trimmed.bam.temp \
    ${PREFIX}_trimmed.bam

gatk \
    --java-options "-Xmx${MEMORY}g" \
    MarkDuplicates \
    --INPUT ${PREFIX}_trimmed_sorted.bam \
    --OUTPUT ${PREFIX}_trimmed_sorted_marked.bam \
    --METRICS_FILE ${PREFIX}_trimmed_sorted_marked.metrics \
    --TMP_DIR .

samtools view \
    -b \
    -f 3 \
    -F 1024 \
    -F 2048 \
    -q ${MAPQ} \
    -o ${OUTPUT_FILE}.bam \
    ${INPUT_FILE}
```

6. Perform mbias analysis to determine trimming parameters.
   a. This step can be run prior to methylation calling and is usually helpful in determining an appropriate mbias mask:

```
astair mbias \
    --N_threads ${CPUS} \
    --no_information 0 \
    --input_file ${PREFIX}_trimmed_sorted_marked_deduped.bam \
    --reference ./${fasta} \
    --directory ./ \
    --read_length ${read_length} \
    --method ${WGBS}
```

7. Call methylation using optimized parameters for your dataset.
   a. Adjust input values as appropriate. `-start_clip/-end_clip` defaults are suggested; however, we recommend substituting these values with library-specific cutoffs based on the results of the mbias analysis outlined above:

```
# Create the output dir for AsTair
mkdir -p ${PREFIX}_astair_output

astair call \
    --N_threads $CPUS \
    --no_information 0 \
    --skip_clip_overlap true \
    --start_clip 7 \
    --end_clip 7 \
    --context CpG \ # context can be any of: [all|CpG|CHG|CHH]
    --max_depth 100000 \
    --minimum_base_quality 30 \
    --minimum_mapping_quality 20 \
    -i ./${PREFIX}_trimmed_sorted_marked_deduped.bam \
    -f /path/to/NA12878_lambda_2kb.fa \
    -d .${PREFIX}_astair_output
```

NOTES:

- It is critical that you include the `--no_information 0` flag for correct methylation conversion percentages when using asTair.
- If you want to inspect contexts other than CpGs, you can omit the “context” flag from the asTair call, or optionally choose another context from this list: *[all|CpG|CHG|CHH]*
- Finally, the `-max-depth` parameter for asTair may be important if you have deeply sequenced a smaller genome (default coverage depth cutoff is 250). The official asTair documentation has additional information on these and other parameters.

Illumina DRAGEN Pipeline

Description

Illumina's DRAGEN pipeline leverages FPGA hardware acceleration for rapid alignment and methylation calling. It supports TAPS mode and is accessible via Illumina BaseSpace. Documentation for the DRAGEN Methylation Pipeline can be found here.[4]

Advantages

- Hardware-accelerated, high-throughput analysis
- Integrated into Illumina BaseSpace platform for ease of use
- Capable of fast alignment and methylation calling in a single workflow
- Supports CpG, CHG, and CHH contexts

Limitations

- Proprietary and requires Illumina DRAGEN hardware/BaseSpace access
- Less transparent than open-source pipelines for parameter customization
- Methylation context flexibility is limited compared to nf-core/methylseq, rastair, or asTair

Procedure

1. Prepare and sequence libraries using the Watchmaker DNA Library Prep Kit with TAPS+.
2. Launch the Illumina DRAGEN Methylation Pipeline from Basespace.
3. Select either your Biosamples or Samples as appropriate for analysis, and add the appropriate biosamples/samples to the run.
4. Select "Map/Align + Methylation Reporting" for the pipeline configuration if you are starting from fastq files, otherwise "Reporting Only" if you've previously aligned your data and will use BAM files as input.
5. Select an appropriate reference genome (e.g., hg38), "BAM" for alignment output, and "Directional" (e.g.,) for Methylation protocol
6. Ensure you select "Enable TAPS Support" as this will reverse the logic used to determine which cytosines are methylated.
7. Fill out any other options that are applicable to your use case (e.g., targeted regions, etc.) and launch the job.

SPECIAL TOPICS AND ADVANCED ANALYSIS

Methylation Bias (mbias) Analysis

Preparation of DNA libraries often introduces bias at fragment ends due to end repair and A-tailing steps.[5,6] This positional bias, referred to as mbias, can affect methylation calling and must be corrected during analysis by masking bases at read ends.

- **nf-core/methylseq** includes automated selection of trimming ranges on a per-sample basis, simplifying workflow management.
- **rastair** provides a flexible masking approach. For high-quality or long-read libraries, no trimming of R1 is generally required. Masking 10 – 20 bases from the start of R2 is recommended. Libraries with shorter inserts may require some 3' masking on R1.
- **asTair** typically requires masking of 7 – 10 bases from each end as a starting point.

The nf-core/methylseq pipeline can dynamically select mbias filtering cutoffs for each sample, which is the simplest and most reproducible way to ensure appropriate trimming. Depending on your analysis goals, you may also choose to set cutoffs manually, informed by factors such as coverage depth and the specific mbias profile of your samples (Figure 2).

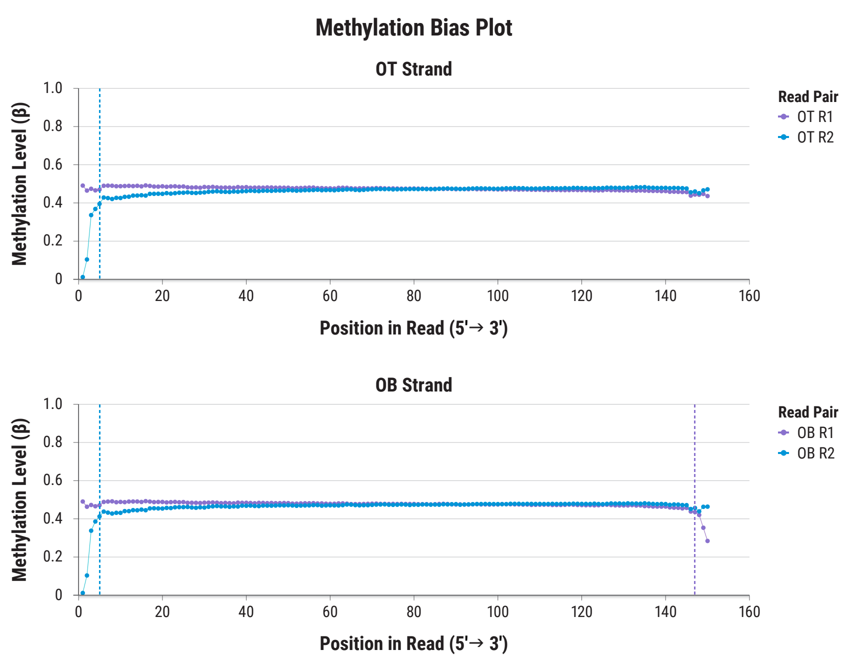


Figure 2. Example of mbias plot. Methylation levels measured across the OT and OB strands for both Reads 1 and 2. Vertical lines indicate suggested mbias masking thresholds for this library.

Coverage Considerations:

- Confident methylation calls generally require >5 reads covering both strands at a given site.
- Variant calling requires sufficient depth on both strands to avoid false positives.

Adjustments may be necessary depending on sample type and downstream analysis requirements.

Differentially Methylated Regions and methylKit

Epigenetic changes linked to disease, development, or environmental factors are often studied through the analysis of differentially methylated regions (DMRs). TAPS+ is well-suited to this application because it provides accurate base-resolution discrimination between methylated and unmethylated cytosines while preserving DNA integrity. Unlike BS-seq, TAPS+ does not degrade DNA, allowing retention of genomic complexity and enabling simultaneous analysis of single nucleotide variants (SNVs), insertions and deletions (indels), and copy number variations (CNVs) from the same library. This multimodal capability supports comprehensive characterization of both genomic and epigenomic alterations.

methylKit is a widely adopted bioinformatics tool for DMR analysis. It provides functionality for quality control, differential methylation testing, and annotation of methylated regions. Outputs from TAPS+ callers such as **rastair** can be converted into methylKit-compatible formats. The **nf-core/methylseq** pipeline generates this output automatically for TAPS+ data, reducing the need for custom scripts.

For workflows using rastair, a script for directly converting methylation calls to a methylKit-compatible input can be found in in our GitHub repository.[3]

Variant Calling

Variant calling with TAPS+ requires a modified workflow because methylated cytosines (5mC and 5hmC) are converted to thymine during library preparation. Standard SNV callers are not methylation-aware and may misinterpret these C>T transitions as somatic or germline variants. One effective strategy is to use rastair for CpG-aware modification calls alongside GATK HaplotypeCaller for traditional variant detection, followed by integrating the results from both sets of calls.

Workflow

1. **Library preparation and preprocessing**
   Prepare and sequence libraries using the Watchmaker DNA Library Prep Kit with TAPS+. Perform standard preprocessing: adapter trimming, alignment (e.g., BWA-MEM), sorting, duplicate marking, and BAM indexing.
2. **CpG calls with rastair**
   Run `rastair call` on preprocessed BAM files to generate modification calls at CpG positions.
3. **Convert rastair output to VCF**
   Transform the `rastair .mods` output into VCF format using a custom script or utility (an example of which can be found in our public GitHub repository[2]). This creates a reference call set of CpG sites with methylation represented as per-site variant-like records.

   Note: rastair currently calls only C>T / G>A SNVs and ignores other SNV types within CpG contexts.
4. **Variant calls with HaplotypeCaller**
   Run GATK HaplotypeCaller, ensuring that CpG positions are masked to avoid false C>T calls at canonical CpG sites. A copy of this CpG mask can be found in our public GitHub repository.[2]
5. **Recover non-reference CpGs**
   From the HaplotypeCaller output, identify variants that create new CpG sites (e.g., A>G or T>C substitutions adjacent to G/C bases). Recalculate methylation at these positions using the original BAM and context-aware logic from rastair or similar tools.

6. **Integrate results**
   Merge the VCFs from rastair and HaplotypeCaller into a unified call set, ensuring consistent formatting and annotation across both sources.

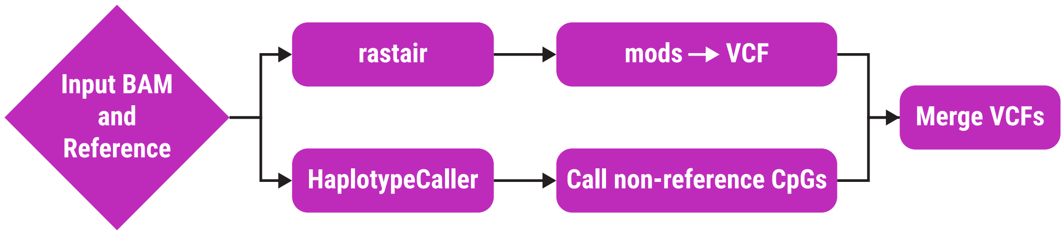


Figure 3. Variant calling workflow for TAPS+ data. Schematic overview of a combined strategy for variant and methylation analysis in TAPS+ libraries. CpG sites are first identified using rastair, with outputs converted to VCF. Conventional variants are then detected using GATK HaplotypeCaller with CpGs masked to prevent false C>T calls. Non-reference CpGs are recovered by re-interrogating HaplotypeCaller outputs, after which the CpG-specific and conventional variant sets are merged into a unified VCF for downstream analysis.

Reversed Read Orientation in TAPS+ Libraries

Certain TAPS+ libraries — particularly those prepared with xGen™ UDI-UMI Adapters — produce sequencing reads in the opposite orientation from what analysis tools such as rastair and asTair expect. This occurs because the adapters are designed as reverse complement sequences of standard designs, leading to strand orientation reversal during sequencing.

By default, rastair and asTair assume that Read 1 (R1) corresponds to the forward strand. In these libraries, however, R1 actually aligns to the reverse strand. A common indicator of this issue is unexpectedly low or absent methylation calls.

Correction options:

1. **Swap reads just prior to alignment** – Process fastqs normally, but swap R1 and R2 immediately before alignment; or
2. **Rename fastqs before processing** – Rename R1 and R2 up front, and adjust adapter trimming accordingly (e.g., in cutadapt, swap the sequences used for the "-a" and "-A" parameters).

REFERENCES

1. Liu, Y., Siejka-Zielińska, P., Velikova, G. et al. Bisulfite-free direct detection of 5-methylcytosine and 5-hydroxymethylcytosine at base resolution. *Nat Biotechnol* 37, 424 – 429 (2019). https://doi.org/10.1038/s41587-019-0041-2

2. Rastair homepage: https://www.rastair.com/installation.html

3. GitHub repository: https://github.com/watchmaker-genomics/taps-foundry

4. DRAGEN Methylation pipeline documentation: https://www.illumina.com/products/by-type/informatics-products/basespace-sequence-hub/apps/edico-genome-inc-dragen-methylation-pipeline.html

5. Library end-repair reaction introduces methylation biases in paired-end (PE) Bisulfite-Seq applications, QCFail.com. https://tinyurl.com/mw4x6jst

6. Ryan, D.P., Ehninger, D. Bison: bisulfite alignment on nodes of a cluster. *BMC Bioinformatics* 15, 337 (2014). https://doi.org/10.1186/1471-2105-15-337

WATCHMAKER GENOMICS

For Technical Support, please contact support@watchmakergenomics.com.

5744 Central Avenue, Suite 100
Boulder, CO 80301
www.watchmakergenomics.com