Avida Targeted Methylation Sequencing Analysis

Analysis of Avida Methyl sequencing data using open-source software packages

Technical Guide

Version A0, September 2024

For Research Use Only. Not for use in diagnostic procedures.

Notices

Manual Part Number

G9419-90001

Edition

Version A0, September 2024

Agilent Technologies, Inc.
5301 Stevens Creek Blvd
Santa Clara, CA 95051 USA

Technical Support

For US and Canada

Call (800) 227-9770 (option 3,4,4)

Or send an e-mail to:
ngs.support@agilent.com

For all other regions

Agilent's world-wide Sales and Support Center contact details for your location can be obtained at
www.agilent.com/en/contact-us/page.

Warranty

The material contained in this document is provided "as is," and is subject to being changed, without notice, in future editions. Further, to the maximum extent permitted by applicable law, Agilent disclaims all warranties, either express or implied, with regard to this manual and any information contained herein, including but not limited to the implied warranties of merchantability and fitness for a particular purpose. Agilent shall not be liable for errors or for incidental or consequential damages in connection with the furnishing, use, or performance of this document or of any information contained herein. Should Agilent and the user have a separate written agreement with warranty terms covering the material in this document that conflict with these terms, the warranty terms in the separate agreement shall control.

Safety Notices

CAUTION

A **CAUTION** notice denotes a hazard. It calls attention to an operating procedure, practice, or the like that, if not correctly performed or adhered to, could result in damage to the product or loss of important data. Do not proceed beyond a **CAUTION** notice until the indicated conditions are fully understood and met.

WARNING

A WARNING notice denotes a hazard. It calls attention to an operating procedure, practice, or the like that, if not correctly performed or adhered to, could result in personal injury or death. Do not proceed beyond a WARNING notice until the indicated conditions are fully understood and met.

In this Guide...

This guide provides an analysis pipeline for Illumina NGS sequencing files generated from libraries prepared with the Avida Methyl reagents.

Content

1
Introduction to Avida Methylation Analysis

This chapter provides an overview of Avida products and an introduction to the purpose of the Avida Targeted Methylation Sequencing Analysis Technical Guide.

Overview of the Avida Reagent Kits and Panels

Detection of tumor-specific variants and methylation patterns in cell-free DNA (cfDNA) samples can be challenging due to limiting sample quantities and low abundance of tumor DNA within a large amount of normal DNA. The Avida library preparation and target enrichment system was specifically designed to address these challenges to provide highly sensitive detection of somatic variants and the presence of methylated DNA using cfDNA (1–100 ng for detection of variants or 3–100 ng for detection of methylation changes) or 10–100 ng of sheared genomic DNA (gDNA).

The performance of the Avida platform is enabled through several technological innovations:

- A highly optimized library preparation formulation that efficiently converts cfDNA molecules and gDNA fragments into Illumina-sequenceable inserts.
- Dual unique molecule identifiers (UMIs) to tag top and bottom strands and allow for error correction.
- A PCR-free library preparation method that reduces bias and the risk of allele dropout.
- A target enrichment method that leverages cooperative probe binding to capture 70–80% of both strands of the DNA molecules.
- A soft conversion method that preserves and retrieves methylation information from virtually any region of interest, even at very low DNA input levels.

Together, these features allow for assessment of methylation changes in disease-specific genomic regions.

The Avida portfolio consists of three library preparation kit options that support DNA analysis (Avida DNA reagent kits), methylation analysis (Avida Methyl reagent kits), and DNA plus methylation analysis from a single sample (Avida Duo Methyl reagent kits). Along with the library preparation kits, three DNA catalog panels and one methylation-specific catalog panel (Avida Methyl 3400 DMR Cancer Panel) are available. The Avida Methyl 3400 DMR Cancer Panel is an 866-kb panel providing comprehensive coverage of approximately 3400 methylation regions established using public databases and internal sequencing studies. These regions point to biomarkers suggesting deviating methylation patterns in tumor-derived DNA as compared to normal cell-derived DNA.

About the Avida Targeted Methylation Sequencing Analysis Technical Guide

This technical guide provides step-by-step instructions on the bioinformatics analysis — from FASTQ input files to CpG base-level methylation status — using open-source software packages. Detailed protocols outlining the library prep, target capture, and soft conversion steps of the Avida Methyl and Avida Duo Methyl workflows can be found on the Agilent website using the links below. Refer to the Avida DNA Targeted Sequencing Analysis Technical Guide (publication number G9409-90001) for analysis instructions on the Avida DNA portion of the Avida Duo Methyl workflow.

https://www.agilent.com/cs/library/usermanuals/public/G9419-90000.pdf
https://www.agilent.com/cs/library/usermanuals/public/G9439-90000.pdf

The analysis pipeline for the Avida Methyl workflow starts with paired-end sequencing data files in FASTQ format generated by the Illumina sequencer. The pipeline consists of UMI analysis, sequence alignment, data QC metrics creation, and sample methylation result generation. This technical guide describes each of these analysis pipeline steps, with detailed step-by-step command line examples. The information is intended to guide you as you use your FASTQ input files to interrogate the targeted regions for the following features.

- Per-base methylation level of cytosines
- Methylation degree of each of the approximately 3400 methylation regions covered by the Avida Methyl 3400 DMR Cancer Panel

All steps described in this publication make use of publicly available open-source tools. Table 1 lists the open-source software packages used in the example scripts in Chapter 2, "Avida Methyl Analysis Pipeline Steps."

Open-Source Tools for Avida Methyl Analysis

Table 1 Open-source software packages*

Package name	Versions tested in this guide
BEDTools	2.25.0
Bismark	0.23.0
Bowtie2	2.4.5
Cutadapt	3.5
Java	1.8.0_322
Miniconda3	3.9.5
Picard	2.20.1
SAMtools	1.9
Seqtk	1.2-r94
Umi-Grinder	0.0.1

* The Software may utilize third party software made available under various open source and third party software licenses ("Third Party Components"). The terms associated with the Third Party Components are available in the comprehensive-license-disclosure.txt within the ExDViewer Program Files Directory. You agree to comply with all applicable terms. In addition to the warranty disclaimers contained in the terms associated with the Third Party Components, Agilent makes the following disclaimers regarding the Third Party Components on behalf of itself, and the copyright holders, contributors, and licensors of the Third Party Components: To the maximum extent permitted by applicable law, the Third Party Components are provided by the copyright holders, contributors, licensors, and Agilent "as is" and AGILENT MAKES NO WARRANTIES OR REPRESENTATIONS OF ANY KIND, WHETHER ORAL OR WRITTEN, WHETHER EXPRESS, IMPLIED, OR ARISING BY STATUTE, CUSTOM, COURSE OF DEALING, OR TRADE USAGE, INCLUDING WITHOUT LIMITATION THE IMPLIED WARRANTIES OF TITLE, MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE, AND NON-INFRINGEMENT. In no event will the copyright owner, contributors, licensors, or Agilent be liable for any direct, indirect, incidental, special, exemplary, or consequential damages (including, but not limited to, procurement of substitute goods or services; loss of use, data, or profits; or business interruption), however caused and on any theory of liability, whether in contract, strict liability, or tort (including negligence or otherwise) arising in any way out of the use of the Third Party Components.

2
Avida Methyl Analysis Pipeline Steps

This chapter describes the steps of the analysis pipeline for Avida targeted methylation sequencing data with shell script examples for each step.

The shell script examples are based on the following folder structures. If your computing node uses different folder structures, adjust the text in the examples accordingly.

- “/localdata" - the local storage mount where reference genome, software packages, input FASTQ files and output files are stored.
- “/localdata/tools" - the directory where all software packages such as Bismark, SAMtools, BEDTools etc. are installed.
- "$sample_dn" - the directory that contains the input FASTQ files and the output results are to be stored for the sample. For example, sample_dn="/localdata/analyses/run1/sample1".

Avida Targeted Methyl Sequencing Analysis Pipeline Summary

The sequence of steps for the Avida Methyl analysis pipeline is illustrated in Figure 1 and in Table 2.

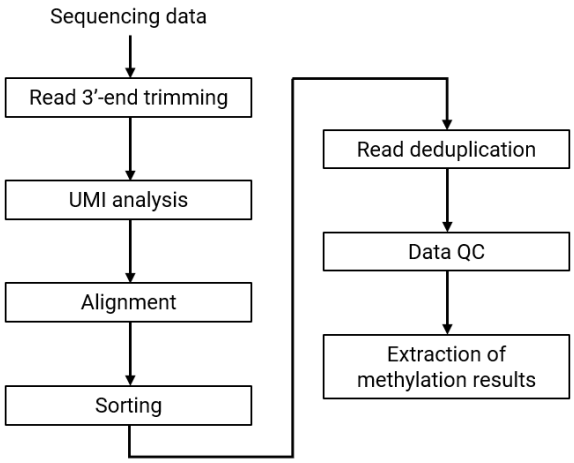

Figure 1 Diagram of the Avida Methyl analysis pipeline

Table 2 Steps of the Avida Methyl analysis pipeline

Step number and description (software package)
1 (Optional) Read trimming at 3' ends (seqtk)
2 UMI trimming and collection from the 5' end (custom*)
3 Adaptor trimming (Cutadapt)
4 Read alignment (Bismark)
5 Read sorting by coordinates and addition of read groups (Picard)
6 UMI-based deduplication (UmiBam) or positional deduplication (Picard)
7 QC of alignment and coverage (BEDTools, Picard, SAMtools)
8 Extraction of methylation status (Bismark methylation extractor)

* The script used is a custom script. See "Step 2. UMI trimming and collection from the 5' end (custom)" on page 10.

The steps shown in Table 2 are described in further detail in the following section, "Analysis Pipeline Steps."

Analysis Pipeline Steps

Step 1. (Optional) Read trimming at 3’ ends (seqtk)

If the read length is longer than 100 bp, trim the reads at the 3’ end. The aligner used in this pipeline uses global alignment, i.e., end-to-end alignment, through Bowtie2 in the back end. For DNA inserts that are relatively short compared to the read length, trimming may result in a higher alignment rate by removing alien sequences and low-quality bases that are not needed for methylation analysis. This step is optional.

```
seqtk trimfq -L 100 R1.fq > r1_trimmed.fq && seqtk trimfq -L 100 R2.fq > r2_trimmed.fq
```

Step 2. UMI trimming and collection from the 5’ end (custom)

The Avida Methyl and Avida Duo Methyl reagent kits add inline UMIs to both ends of the DNA inserts to enable UMI-based read deduplication and consensus matching. The inline UMIs are preserved at the 5’ end of both Read1 (R1) and Read2 (R2) of the sequencing reads. To collect the UMIs and process them for analysis, the recommended approach is to trim off 5 bases at the beginning of each read, take the first 3 of the 5 bases as the UMI, and discard the remaining 2 bases (sometimes referred to as “dark bases”).

```
python duplex_umi_fastq.py \
  --r1-in-fn r1_trimmed.fq --r2-in-fn r2_trimmed.fq \
  --r1-out-fn r1_dmi.fq --r2-out-fn r2_dmi.fq \
  -d $sample_dn/readstats
```

Note: The script “duplex_umi_fastq.py” is a custom script. The script trims 5 bases from the 5’ end of both R1 and R2, and then combines the first 3 bases of the 5 trimmed bases from each read to generate the UMI in the format of “ABC+DEF.” The script then attaches the combined UMI to the FASTQ header line of both R1 and R2.

For example, the FASTQ header starts as:

@VL00103:1:AACFWKHM5:1:1210:38852:46946 1:N:0:CACCTGTT+AGGATAGC

and ends as:

@VL00103:1:AACFWKHM5:1:1210:38852:46946:GGG+CTG 1:N:0:CACCTGTT+AGGATAGC

Note that, in this example, the combined UMI “GGG+CTG” was attached to the first section of the FASTQ header line.

Step 3. Adaptor trimming (Cutadapt)

Perform standard adaptor trimming using Cutadapt.

```
cutadapt -m 30 -q 15 -o r1_cutadapt.fq -p r2_cutadapt.fq \
  -a AGATCGGAAGAGCACACGTCTGAACTCCAGTCA \
  -A AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT \
  r1_dmi.fq r2_dmi.fq > $sample_dn/readstats/cutadapt.log
```

Step 4. Read alignment (Bismark)

Align bisulfite or enzymatically treated reads using Bismark, with Bowtie2 as the back end aligner.

Note: The directory that contains the Bowtie2 executable file (e.g., /localdata/tools/bowtie2-2.4.5/) must be included in the environment variable $PATH before running Bismark. The environment variable $PATH can be set with a shell script such as:

```
export PATH=$PATH:/localdata/tools/bowtie2-2.4.5
```

```
bismark --bowtie2 /localdata/references/hg19/Sequence/bismark --bowtie2 \
 /localdata/tools/samtools-1.9 -p 4 -N 1 -L 20 \
 -1 r1_cutadapt.fq -2 r2_cutadapt.fq --chunkmbs 2048 --nucleotide_coverage \
 --output_dir $sample_dn/bsout --temp_dir $sample_dn/bsout --basename bs_aln
```

Step 5. Read sorting by coordinates and addition of read groups (Picard)

Sort reads by chromosomal coordinates and add read group tags. Verify mate-pair information and repair as necessary.

```
java -Djava.io.tmpdir=$sample_dn -Xmx16G \
 -jar /localdata/tools/picard-tools-2.20.1/picard.jar AddOrReplaceReadGroups \
 VALIDATION_STRINGENCY=SILENT \
 INPUT=$sample_dn/bsout/bs_aln_pe.bam \
 OUTPUT=$sample_dn /picard_sorted.bam \
 RGID=demo_sample RGSM=demo_sample \
 SORT_ORDER=coordinate CREATE_INDEX=true \
 RGLB=pe RGPU=HiSeq RGCN=PDx RGPL=illumina
```

```
java -Djava.io.tmpdir=$sample_dn -Xmx16G \
 -jar /mnt/scratch/tools/picard-tools-2.20.1/picard.jar FixMateInformation \
 INPUT=$sample_dn/picard_sorted.bam \
 OUTPUT=$sample_dn/picard_fixed.bam \
 ADD_MATE_CIGAR=true CREATE_INDEX=true
```

Step 6. Deduplication, UMI-based (Umi-Grinder) or positional (Picard), and sorting/indexing (SAMtools)

Choose between UMI-based deduplication (step 6a) and positional deduplication (step 6b). If you use UmiBam for positional deduplication, then you will need to perform sorting and indexing (step 6c) using either SAMtools or Picard.

Step 6a. UMI-based deduplication (Umi-Grinder) In this method, read deduplication is performed with alignment positions with UMI information to achieve a more refined unique molecule identification. The reads need to be sorted by names first.

The read header line need to be modified in preparation to run UmiBam. The header starts as:

VL00103:1:AACFWKHM5:1:1210:38852:46946:GGG+CTG_1:N:0:CACCTGTT+AGGATAGC

and ends as:

VL00103:1:AACFWKHM5:1:1210:38852:46946:GGGCTG

Note that the trailing string beginning with "_1" was trimmed and the "+" in the combined UMI was dropped.

```
# May need to revise read names in bam file first
# to be compatible with the assumptions made by Umi-Grinder
samtools sort -n -O bam -m 16G -@ 4 -T $sample_dn \
 -o $sample_dn/name_sorted.bam \
 $sample_dn/name_revised.bam
UmiBam -p --bam --umi --mm 1 \
 --samtools_path /localdata/tools/samtools-1.9 name_sorted.bam
mv name_sorted.UMI_1mm_deduplicated.bam deduped.bam
```

Step 6b. Positional deduplication (Picard) In this method, read deduplication is performed based on the alignment chromosome and start and end positions, without the use of UMIs.

```
/usr/bin/java -Djava.io.tmpdir=$sample_dn -Xmx16G -jar \
 /localdata/tools/picard/picard.jar MarkDuplicates \
 INPUT=$sample_dn/picard_fixed.bam \
 OUTPUT=$sample_dn/deduped.bam \
 METRICS_FILE=$sample_dn/readstats/pos_dedup_metrics.txt \
 CREATE_INDEX=true REMOVE_DUPLICATES=true
```

Step 6c. Sorting and indexing (SAMtools) Reads are sorted by chromosomal position again and an index to the BAM file is generated. You can use either SAMtools or Picard to perform this step. The example script below uses SAMtools.

```
samtools sort -O bam -m 4G -@ 4 -T $sample_dn \
 -o $sample_dn/umi_deduped.bam $sample_dn/deduped.bam
 samtools index $sample_dn/deduped.bam
```

Step 7. QC of alignment and coverage (BEDTools, Picard, SAMtools)

Generate and collect alignment metrics for data QC. This step requires the target regions file available from Agilent SureDesign (Regions.bed).

```
samtools view -H $sample_dn/picard_fixed.bam | grep @SQ \
  | sed 's/@SQ  SN:\|LN://g' | cut -f1,2 > $sample_dn/genome_length.tsv

# base coverage
bedtools coverage -a omni1_designed_regions.bed \
  -b $sample_dn/deduped.bam -d -g $sample_dn/genome_length.tsv \
  > $sample_dn/ss_base.cover.tsv

#Picard Hsmetrics
java -Djava.io.tmpdir=$sample_dn -Xmx16G \
  -jar /localdata/tools/picard-tools-2.20.1/picard.jar BedToIntervalList \
  INPUT=omni1_designed_regions.bed \
  OUTPUT=$sample_dn/interval_list \
  SD=/localdata/references/hg19/Sequence/genome.fa.dict
java -Djava.io.tmpdir=$sample_dn -Xmx16G \
  -jar /localdata/tools/picard-tools-2.20.1/picard.jar CollectHsMetrics \
  INPUT=$sample_dn/deduped.bam \
  OUTPUT=$sample_dn/ss_hsmetrics.txt \
  R=/localdata/references/hg19/Sequence/genome.fa \
  BAIT_INTERVALS=$sample_dn/interval_list \
  TARGET_INTERVALS=$sample_dn/interval_list
# flagstat
samtools flagstat $sample_dn/picard_fixed.bam > $sample_dn/flagstats.txt
```

Step 8. Extraction of methylation status (Bismark methylation extractor)

Extract the methylation status of all C bases in each read using the script included in the Bismark package. Results are stored in text files (*.txt) based on C context and strand. The output from the methylation extractor can be further analyzed to summarize methylation levels at individual CpG sites or biomarker regions.

```
bismark_methylation_extractor --paired-end \
  --samtools_path /localdata/tools/samtools-1.9 --buffer_size 16G \
  --parallel 4 --bedGraph --counts --ignore 2 --ignore_r2 7 \
  --output $sample_dn/bsout $sample_dn/bs_dedup.bam
```

www.agilent.com

In This Book

This technical guide provides an analysis pipeline for Illumina NGS sequencing files generated from libraries prepared with the Avida Methyl or Avida Duo Methyl reagents.

Version A0, September 2024

G9419-90001

Agilent