MEDIPIPE: an automated and comprehensive pipeline for cfMeDIP-seq data quality control and analysis

Yong Zeng[1, #], Ye Wenbin[1], Eric Y. Stutheit-Zhao[1], Ming Han[1], Scott V. Bratman[1, 2], Trevor J. Pugh[1, 2, 3, #], Housheng Hansen He[1, 2, #]

[1] Princess Margaret Cancer Center, University Health Network, Toronto, Ontario, Canada

[2] Department of Medical Biophysics, University of Toronto, Toronto, Ontario, Canada

[3] Ontario Institute for Cancer Research, Toronto, Ontario, Canada

[#] To whom correspondence should be addressed.

Abstract

Summary: cell-free methylated DNA immunoprecipitation and high-throughput sequencing (cfMeDIP-seq) has emerged as a promising non-invasive technology to detect cancers and monitor treatments. Several bioinformatics tools are available for cfMeDIP-seq data analysis. However, an easy to implement and flexible pipeline, particularly, for large-scale cfMeDIP-seq profiling, is still lacking. Here we present the MEDIPIPE, which provides a one-stop solution for cfMeDIP-seq data quality control, methylation quantification and sample aggregation. The major advantages of MEDIPIPE are: 1) it is easy to implement and reproduce with automatically deployed execution environments; 2) it can handle different experimental settings with a single input configuration file; 3) it is computationally efficient for large-scale cfMeDIP-seq profiling data analysis and aggregation.

Availability and implementation: This pipeline is an open-source software under the MIT license and it is freely available at https://github.com/yzeng-lol/MEDIPIPE.

Contact: yzeng@uhnresearch.ca or trevor.pugh@utoronto.ca or hansenhe@uhnresearch.ca

Supplementary information: Supplementary data are appended.

1 Introduction

cell-free DNA (cfDNA) in blood has been a promising analyte for cancer prognosis and treatment monitoring (Corcoran and Chabner, 2018). Next-generation sequencing (NGS)-based technologies have also been tailored to identify cfDNA genomic and epigenomic signatures associated with cancer phenotypes. Among them, the epigenomic profiling through cell-free methylated DNA immunoprecipitation and high-throughput sequencing (cfMeDIP-seq) (Shen *et al.*, 2018) has proven capable of ultrasensitive tumor detection and classification, particularly in the scenario of early-stage cancer or minimal tumor residual after treatments (Shen *et al.*, 2018; Burgener *et al.*, 2021; Nassiri *et al.*, 2020). Refined cfMeDIP-seq protocols incorporating spike-in controls and/or unique molecular identifiers (UMIs) also offer improved capabilities for batch effect correction and error-suppression (Shen *et al.*, 2019; Burgener *et al.*, 2021). We have also been witnessing an increased number of large-scale cfMeDIP-seq profiling studies conducted in different cancer types (Shen *et al.*, 2018; Liu *et al.*, 2021; Burgener *et al.*, 2021; Nuzzo *et al.*, 2020; Nassiri *et al.*, 2020; Chen *et al.*, 2022). However, an easy-to-implement and flexible pipeline for large-scale cfMeDIP-seq profiling data processing and analysis is still lacking in the field.

Here we present an open-source pipeline, MEDIPIPE, to provide an automated and comprehensive solution for cfMeDIP-seq data quality control (QC), methylation quantification and sample aggregation. This pipeline was developed using Snakemake (Mölder *et al.*, 2021), ensuring all dependencies are automatically installed and execution is seamless and reproducible. This pipeline is able to handle various experimental settings, such as sequence layout and whether spike-in controls and/or UMI were added, via specifying a single input configuration file. Moreover, it can efficiently deal with large-scale cfMeDIP-seq profiling data on high performance computing clusters, because all independent steps for individual samples are run in parallel.

2 Pipeline description

MEDIPIPE consists of four modules, which starts with parsing the customized input configuration file, and subsequently conducting corresponding workflows (**Fig. 1 and Supplementary Fig. S1**). In general, the raw cfMeDIP-seq sequencing reads will be preprocessed, aligned, quantified and quality assessed per sample in the first three modules. Then, users can activate the final module to generate aggregated QC reports and quantification matrices.

2.1 Input of pipeline

A single input configuration file in YAML format is required for successfully running this pipeline, which specifies the paths to the working environment, genome references table, samples' sequencing data, and aggregation tables, as well as parameters for different experimental settings. Instructions and a detailed template are included in the repository. Notably, the sequencing reads can be either single-end or paired-end gzip-compressed FASTQ files, and multiple FASTQ files (e.g. multiple sequencing runs) for the same biological sample can be merged by the pipeline as well. We also provide shell scripts for automatically downloading ENCODE pre-build genome references (e.g. BWA index and annotated regions) (Luo *et al.*, 2020), or building customized genome references (e.g. when spike-in sequences were added to the primary genome).

2.2 Reads trimming and QC

Trim Galore (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) is employed to trim adapter sequences and low-quality bases of raw reads. If UMI barcodes are added into the library to suppress deduplication error, UMI-tools (Smith *et al.*, 2017) will be executed prior to Trim Galore to extract barcodes and remove those reads which failed to match the barcode pattern. Lastly, FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) is applied to check the qualities of both raw and preprocessed reads. All of these steps can be executed in either single-end or paired-end mode (**Fig. 1**).

2.3 Reads alignment and QC

Next, the preprocessed reads will be mapped back to corresponding genome(s) via BWA-MEM (Li, 2013). For the scenario of reads coming along with spike-in controls, we recommend appending spike-in sequences to the primary genome for alignment simultaneously. Aligned spike-in reads can be extracted for separate quantity assessment and quantification, which is specified in the configuration file. Then, SAMtools (Li *et al.*, 2009) is applied to filter out unmapped reads and secondary alignments, as well as improperly paired mates for paired-end reads. Duplicated reads will be removed either by SAMtools markdup (Li *et al.*, 2009) or UMI-tools dedup (Smith *et al.*, 2017), depending on whether UMI barcodes were added or not (**Fig. 1**). Specific to paired-end reads, the fragment size will be estimated by CollectInsertSizeMetrics from Picard tool kit (http://broadinstitute.github.io/picard/), and the fragmentation profiles, which is defined as the ratio of short to long fragments in consecutive windows (Cristiano *et al.*, 2019), will be computed at 1 megabase pair (Mb) and 5Mb resolution as well. Lastly, QC metrics derived from the aligned reads, including BAM files statistics via SAMtools (Li *et al.*, 2009) and cfMeDIP-seq saturation, coverage, and enrichment scores via MEDIPS (Lienhard *et al.*, 2014), are collected.

2.4 Methylation quantification

The cfMeDIP-seq offers higher coverage of methylated CpG dinucleotides throughout the genome with much lower cost compared to bisulfite-based methods, however, the absolute methylation levels can only be estimated using a computational model for this enrichment based profiling method. MEDIPIPE applies three successively developed methods for relative and absolute methylation estimation (**Supplementary Table S1**): MEDIPS (Lienhard *et al.*, 2014), QSEA (blind) (Lienhard *et al.*, 2017) and MeDStrand (Xu *et al.*, 2018). These methods were designed to eliminate CpG density bias with a linear regression model, sigmoidal model with empirical knowledge, and stranded sigmoid model, respectively. This allows the users to choose the quantifications as they need and enables a comprehensive comparison among different estimations. If there are spike-in controls, MEDIPS (Lienhard *et al.*, 2014) is used to separately quantify the methylation levels for them as well.

2.5 Aggregation and filtering

Since many cfMeDIP-seq profiling projects generate data for a large number of samples in multiple groups or batches, the final module of MEDIPIPE can aggregate QC metrics and methylation matrices across all samples for each such project. Namely, first, all QC reports generated by FASTQC, SAMtools, Picard and MEDIPS are aggregated by either MultiQC (Ewels *et al.*, 2016) or an embedded script. On top of that, a summarized QC report in HTML format with selected QC metrics is generated, allowing users to interactively examine sample QC metrics across different groups or batches (**Supplementary Fig. S2**). Meanwhile, different methylation quantifications are aggregated into corresponding TAB-delimited bin-sample TXT files, as well as uniformly filtered out sex chromosomes, mitochondrial chromosome, and ENCODE blacklist regions (Amemiya *et al.*, 2019). Both original and filtered aggregated quantification files are indexed by Tabix (Li, 2011), enabling rapid retrieval of data for genomic regions of interest for downstream analysis.

2.6 Outputs of pipeline

Since many All outputs of MEDIPIPE are organized in corresponding folders. Specifically, the main output files per sample can be grouped into four categories: QC reports, aligned BAM files, fragmentomic features (for paired-end reads only) and methylation quantifications (**Supplementary Table. S1**). Moreover, the aggregated outputs include multiQC reports, aggregated QC reports and the aggregated methylation quantification before and after filtering (**Supplementary Table. S2**)

3 Implementation

MEDIPIPE was developed with Snakemake following a clean, modular and robust design in accordance with best practice coding standards; other specialized tools can be easily added in the future. Detailed instructions on how to install and run this pipeline are presented in (https://github.com/yzeng-lol/MEDIPIPE). This pipeline is highly flexible thanks to the input configuration file, which comes along with options and parameters for different experimental settings and analyses. The pipeline can be run locally or submitted to the high performance computing (HPC) for efficient scheduling within a multiprocessor environment. We had also successfully run it with a published cfMeDIP-seq dataset (Nassiri *et al.*, 2020), which consists of 163 samples from 6 brain cancer subtypes, to get aggregated quantifications and QC reports (**Supplementary Fig. S2).** Lastly, although this pipeline was originally developed for cfMeDIP-seq data, it can also be used for the analysis of methylated DNA immunoprecipitation followed by sequencing (MeDIP-seq) data.

Acknowledgement

We like to thank Sasha Main, Emma Bell, Nicholas Cheng, Althaf Singhawansa and Justin Burgener for acting as the pipeline's beta-tester and their feedback. This is a pilot project for the initiative of cell-free Multiomics Data Coordination Centre (cfMOS-DCC), we would also like to thank the cfMOS-DCC members for sharing their cfMeDIP-seq data for pipeline test and fine-tuning.

Funding

This work was supported by The Cancer Genetics and Epigenetics (TCGE) program at Princess Margaret Cancer Centre. EYZ was funded by a Cancer Research Institute Irvington Postdoctoral Fellowship.

Conflicts of interest

Scott V. Bratman: Stock ownership in Adela; leadership position in Adela; patents licensed to Roche, Adela; and royalties from Roche. Others: NA.

References

Amemiya,H.M. *et al.* (2019) The ENCODE Blacklist: Identification of Problematic Regions of the Genome. *Sci. Rep.*, **9**, 9354.

Burgener,J.M. *et al.* (2021) Tumor-Naïve Multimodal Profiling of Circulating Tumor DNA in Head and Neck Squamous Cell Carcinoma. *Clin. Cancer Res.*, **27**, 4230–4244.

Chen,S. *et al.* (2022) The cell-free DNA methylome captures distinctions between localized and metastatic prostate tumors. *Nat. Commun.*, **13**, 6467.

Corcoran,R.B. and Chabner,B.A. (2018) Application of Cell-free DNA Analysis to Cancer Treatment. *N. Engl. J. Med.*, **379**, 1754–1765.

Cristiano,S. *et al.* (2019) Genome-wide cell-free DNA fragmentation in patients with cancer. *Nature*, **570**, 385–389.

Ewels,P. *et al.* (2016) MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, **32**, 3047–3048.

Lienhard,M. *et al.* (2014) MEDIPS: genome-wide differential coverage analysis of sequencing data derived from DNA enrichment experiments. *Bioinformatics*, **30**, 284–286.

Lienhard,M. *et al.* (2017) QSEA-modelling of genome-wide DNA methylation from sequencing enrichment experiments. *Nucleic Acids Res.*, **45**, e44.

Li,H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.

Li,H. (2011) Tabix: fast retrieval of sequence features from generic TAB-delimited files. *Bioinformatics*, **27**, 718–719.

Li,H. *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Liu,J. *et al.* (2021) Genome-wide cell-free DNA methylation analyses improve accuracy of non-invasive diagnostic imaging for early-stage breast cancer. *Mol. Cancer*, **20**, 36.

Luo,Y. *et al.* (2020) New developments on the Encyclopedia of DNA Elements (ENCODE) data portal.

Nucleic Acids Res., **48**, D882–D889.

Mölder,F. *et al.* (2021) Sustainable data analysis with Snakemake. *F1000Res.*, **10**, 33.

Nassiri,F. *et al.* (2020) Detection and discrimination of intracranial tumors using plasma cell-free DNA methylomes. *Nat. Med.*, **26**, 1044–1047.

Nuzzo,P.V. *et al.* (2020) Detection of renal cell carcinoma using plasma and urine cell-free DNA methylomes. *Nat. Med.*, **26**, 1041–1043.

Shen,S.Y. *et al.* (2019) Preparation of cfMeDIP-seq libraries for methylome profiling of plasma cell-free DNA. *Nat. Protoc.*, **14**, 2749–2780.

Shen,S.Y. *et al.* (2018) Sensitive tumour detection and classification using plasma cell-free DNA methylomes. *Nature*, **563**, 579–583.

Smith,T. *et al.* (2017) UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Res.*, **27**, 491–499.

Xu,J. *et al.* (2018) MeDEStrand: an improved method to infer genome-wide absolute methylation levels from DNA enrichment data. *BMC Bioinformatics*, **19**, 540.

Figures

Figure 1

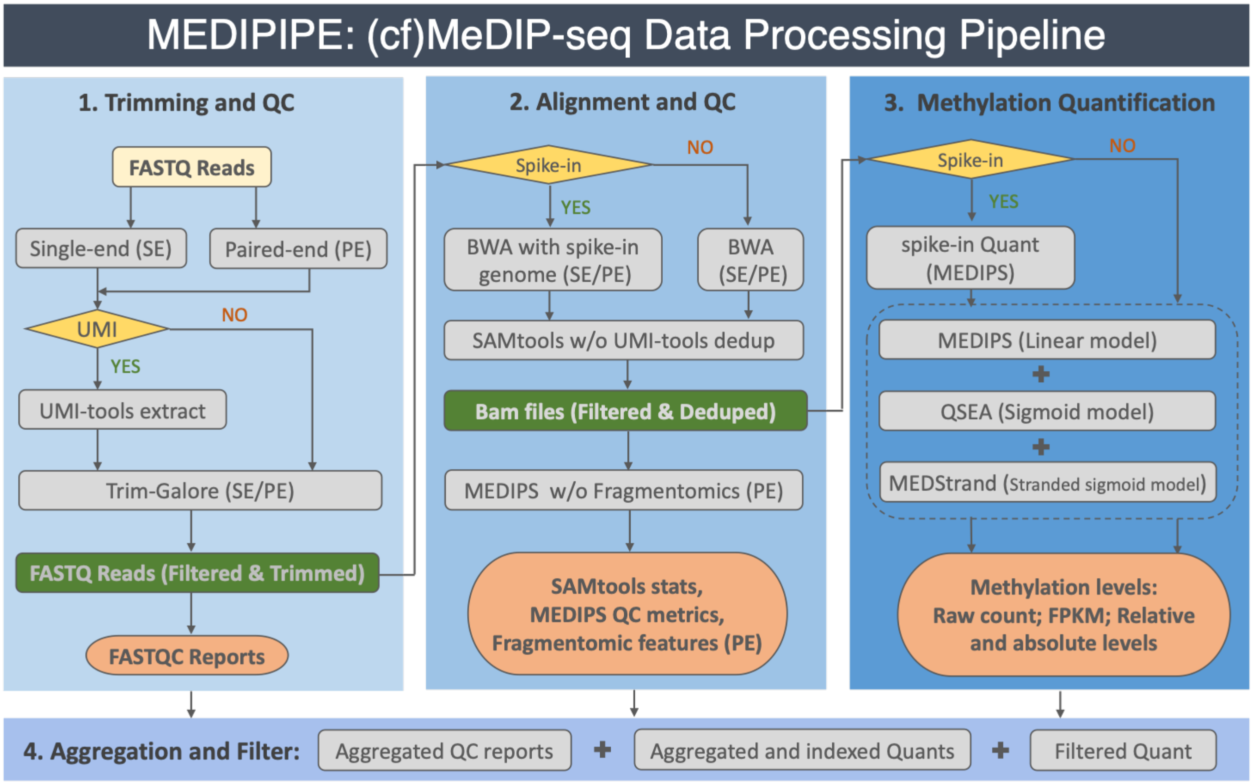

MEDIPIPE: (cf)MeDIP-seq Data Processing Pipeline
1. Trimming and QC
FASTQ Reads
Single-end (SE)
Paired-end (PE)
UMI
NO
YES
UMI-tools extract
Trim-Galore (SE/PE)
FASTQ Reads (Filtered & Trimmed)
FASTQC Reports
2. Alignment and QC
Spike-in
NO
YES
BWA with spike-in genome (SE/PE)
BWA (SE/PE)
SAMtools w/o UMI-tools dedup
Bam files (Filtered & Deduped)
MEDIPS w/o Fragmentomics (PE)
SAMtools stats, MEDIPS QC metrics, Fragmentomic features (PE)
3. Methylation Quantification
Spike-in
NO
YES
spike-in Quant (MEDIPS)
MEDIPS (Linear model)
QSEA (Sigmoid model)
MEDStrand (Stranded sigmoid model)
Methylation levels: Raw count; FPKM; Relative and absolute levels
4. Aggregation and Filter:
Aggregated QC reports
Aggregated and indexed Quants
Filtered Quant

Fig.1. Flowchart of the MEDIPIPE pipeline. UMI, unique molecular identifiers; QC, quality control.

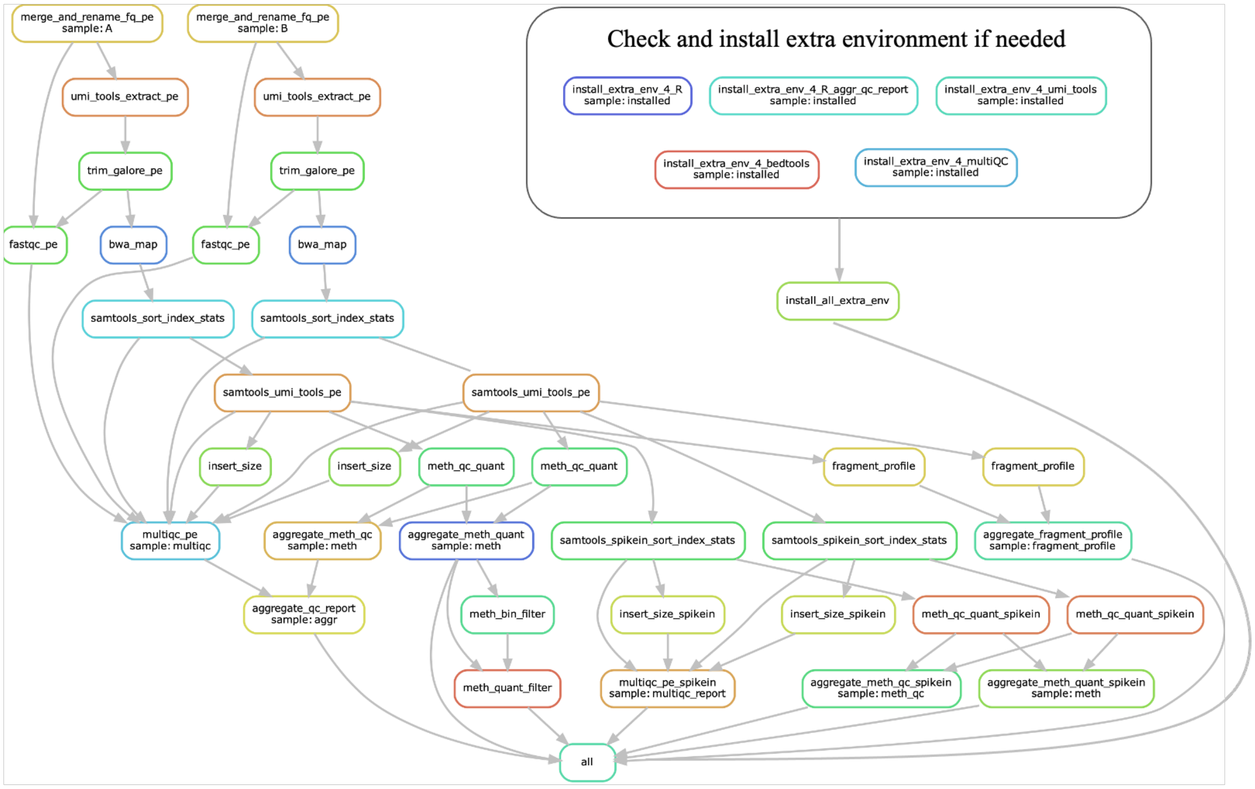
Supplementary Data

MEDIPIPE: an automated and comprehensive pipeline for cfMeDIP-seq data quality control and analysis

Yong Zeng[1, #], Ye Wenbin[1], Eric Y. Stutheit-Zhao[1], Ming Han[1], Scott V. Bratman[1, 2], Trevor J. Pugh[1, 2, 3, #], Housheng Hansen He[1, 2, #]

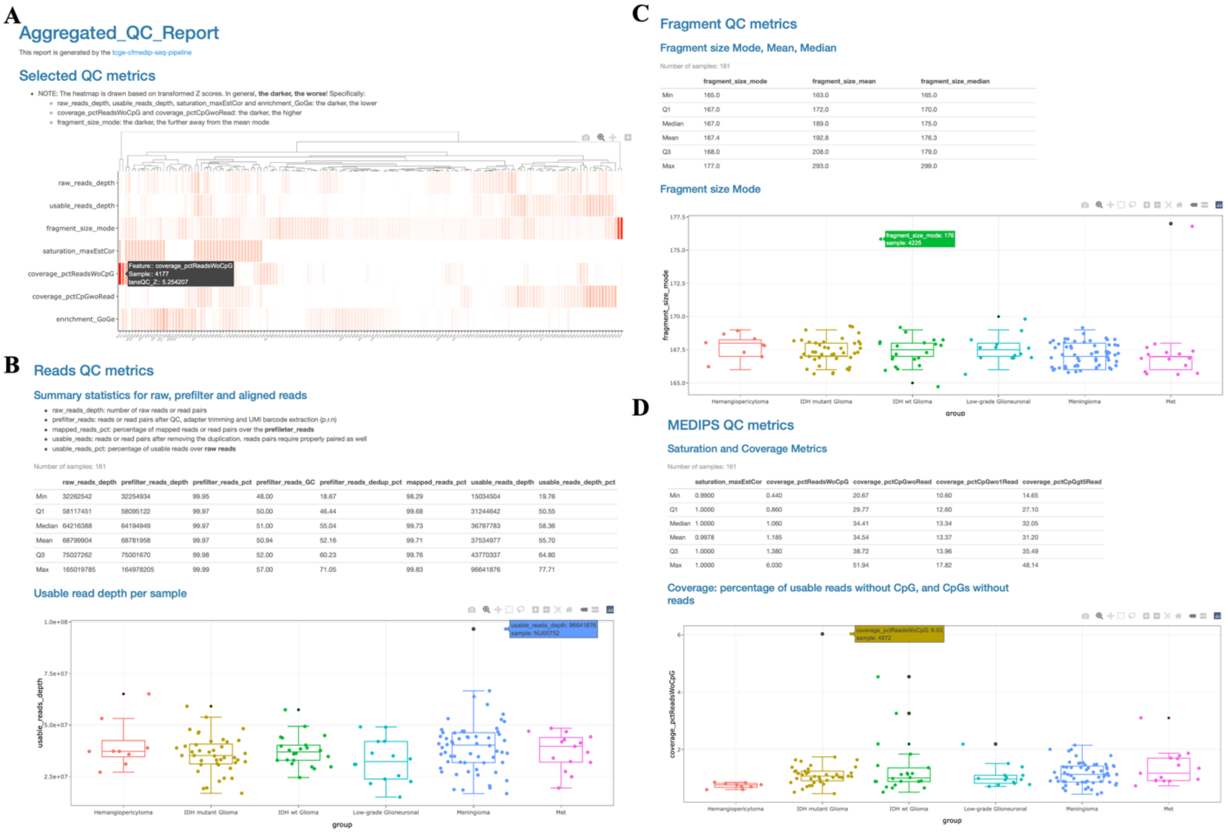
Supplementary Figures

Supplementary Figure 1



Supplementary Fig. S1. Example of directed acyclic graph (DAG) for MEDIPIPE dealing with two paired-end cfMeDIP-seq samples with UMI barcodes and spike-in controls.

Supplementary Figure 2



Aggregated_QC_Report

This report is generated by the tcge-cfmedip-seq-pipeline

Selected QC metrics

- NOTE: The heatmap is drawn based on transformed Z scores. In general, **the darker, the worse!** Specifically:
  - raw_reads_depth, usable_reads_depth, saturation_maxEstCor and enrichment_GoGe: the darker, the lower
  - coverage_pctReadsWoCpG and coverage_pctCpGwoRead: the darker, the higher
  - fragment_size_mode: the darker, the further away from the mean mode

Reads QC metrics

Summary statistics for raw, prefilter and aligned reads

- raw_reads_depth: number of raw reads or read pairs
- prefilter_reads: reads or read pairs after QC, adapter trimming and UMI barcode extraction (p.r.n)
- mapped_reads_pct: percentage of mapped reads or read pairs over the **prefileter_reads**
- usable_reads: reads or read pairs after removing the duplication. reads pairs require properly paired as well
- usable_reads_pct: percentage of usable reads over **raw reads**

Number of samples: 161

	raw_reads_depth	prefilter_reads_depth	prefilter_reads_pct	prefilter_reads_GC	prefilter_reads_dedup_pct	mapped_reads_pct	usable_reads_depth	usable_reads_depth_pct
Min	32262542	32254934	99.95	48.00	18.67	98.29	15034504	19.76
Q1	58117451	58095122	99.97	50.00	46.44	99.68	31244642	50.55
Median	64216388	64194949	99.97	51.00	55.04	99.73	36797783	58.36
Mean	68799904	68781958	99.97	50.94	52.16	99.71	37534977	55.70
Q3	75027262	75001670	99.98	52.00	60.23	99.76	43770337	64.80
Max	165019785	164978205	99.99	57.00	71.05	99.83	96641876	77.71

Usable read depth per sample

Fragment QC metrics

Fragment size Mode, Mean, Median

Number of samples: 161

	fragment_size_mode	fragment_size_mean	fragment_size_median
Min	165.0	163.0	165.0
Q1	167.0	172.0	170.0
Median	167.0	189.0	175.0
Mean	167.4	192.8	176.3
Q3	168.0	208.0	179.0
Max	177.0	293.0	299.0

Fragment size Mode

MEDIPS QC metrics

Saturation and Coverage Metrics

Number of samples: 161

	saturation_maxEstCor	coverage_pctReadsWoCpG	coverage_pctCpGwoRead	coverage_pctCpGwo1Read	coverage_pctCpGgt5Read
Min	0.9900	0.440	20.67	10.60	14.65
Q1	1.0000	0.860	29.77	12.60	27.10
Median	1.0000	1.060	34.41	13.34	32.05
Mean	0.9978	1.185	34.54	13.37	31.20
Q3	1.0000	1.380	38.72	13.96	35.49
Max	1.0000	6.030	51.94	17.82	48.14

Coverage: percentage of usable reads without CpG, and CpGs without reads

Supplementary Fig. S2. Example of an aggregated QC report, which is interactive in HTML format. **(A)** Summary heatmap based on transformed Z scores of selected QC metrics. **(B, C, D)** Summary statistics for Reads, Fragment and MEDIPS QC metrics, as well as boxplots with jittered samples per groups for corresponding QC metrics.

Supplementary Tables

Supplementary Table. S1: Output files per individual samples.

	Files	**Note**
Quality Control Reports	./fastqc_se/Sample_ID_fastqc.html. ./fastqc_pe/Sample_ID_R1_fastqc.html ./fastqc_pe/Sample_ID_R2_fastqc.html	FASTQC for raw reads or read pairs
	./barcode_fq_se/Sample_ID_extract.log ./barcode_fq_pe/Sample_ID_extract.log	UMI barcode extraction log (p.r.n)
	./trimmed_fq/Sample_ID.fastq.gz_trimming_report.txt ./trimmed_fq/Sample_ID_R1.fastq.gz_trimming_report.txt ./trimmed_fq/Sample_ID_R2.fastq.gz_trimming_report.txt	Automatic adapter trimming report
	./fastqc_se/Sample_ID_val_1_fastqc.html ./fastqc_pe/Sample_ID_R1_val_1_fastqc.html ./fastqc_pe/Sample_ID_R2_val_1_fastqc.html	FASTQC for preprocessed reads or read pairs
	./raw_bam/Sample_ID_sorted.bam.stats.txt	Alignment statistics for unfilter BAM file
	./dedup_bam_(umi)_se/Sample_ID_dedup.bam.stats.txt ./dedup_bam_(umi)_pe/Sample_ID_dedup.bam.stats.txt	Alignment statistics for deduped BAM file
	./dedup_bam_spikein/Sample_ID_spikein.bam.stats.txt	Alignment statistics for deduped spike-ins BAM file (p.r.n)
	./meth_qc_quant/Sample_ID_meth_qc.txt	MEDIPS QC : Saturation, Coverage and Enrichment Scores
Aligned Sequence	./raw_bam/Sample_ID_sorted.bam	unfilter Bam file
	./dedup_bam_(umi)_se/Sample_ID_dedup.bam ./dedup_bam_(umi)_pe/Sample_ID_dedup.bam	Deduped BAM file
	./dedup_bam_spikein/Sample_ID_spikein.bam	Deduped Spike-ins BAM file (p.r.n)
Frangmentomic Features (Paired-end reads only)	./fragment_size/Sample_ID_insert_size_metrics.txt	Picard fragment size estimation
	./fragment_size_spikein/Sample_ID_insert_size_metrics.txt	Picard fragment size estimation for spike-ins (p.r.n)
	./fragment_profile/Sample_ID_10_100kb_fragment_profile_GC_corrected_Ratio.txt ./fragment_profile/Sample_ID_50_100kb_fragment_profile_GC_corrected_Ratio.txt	Fragment profile (Short / Long reads ratio)

Methylation Quantification	./meth_qc_quant/Sample_ID_Granges_CpGs.bed	Bins coordinates and #CpG
	./meth_qc_quant/Sample_ID_count.txt	MEDIPS
	./meth_qc_quant/Sample_ID_rpkm.txt	
	./meth_qc_quant/Sample_ID_rms_medips.txt	
	./meth_qc_quant/Sample_ID_rms_medestrand.txt	MEDStrand
	./meth_qc_quant/Sample_ID_nrpm_qsea.txt	QSEA
	./meth_qc_quant/Sample_ID_beta_qsea.txt	
	./meth_qc_quant/Sample_ID_logitbeta_qsea.txt	
	./meth_qc_quant/Sample_ID_CNV_qsea.txt	
	./meth_qc_quant/Sample_ID_meth_quant.RData	Combined data frame per sample

Supplementary Table. S2: Output files for aggregated samples

	Files	Note
Aggregated Quality Control Reports	./aggregated/QC_se/multiqc_data ./aggregated/QC_pe/multiqc_data	MultiQC data
	./aggregated/QC_se/multiqc_report.html ./aggregated/QC_pe/multiqc_report.html	MultiQC HTML report
	./aggreated/meth_qc.txt	Combined MEDIPS QC metrics
	./aggreated/aggr_qc_report.csv	Aggregated selected QC metrics
	./aggreated/aggr_qc_report.html	Aggregated QC HTML report
Aggregated methylation quantification (Indexed with Tabix)	./aggregated/meth_bin.bed	Bins coordinates and #CpG
	./aggregated/meth_count.txt.gz	MEDIPS
	./aggregated/meth_rpkm.txt.gz	
	./aggregated/meth_rms_medips.txt.gz	
	./aggregated/meth_rms_medestrand.txt.gz	MEDStrand
	./aggregated/meth_nrpm_qsea.txt.gz	QSEA
	./aggregated/meth_beta_qsea.txt.gz	
	./aggregated/meth_logitbeta_qsea.txt.gz	
	./aggregated/meth_CNV_qsea.txt.gz	
	./aggregated/meth_meth_quant.RData	Combined data frame per sample
Aggregated methylation quantification and filtered out sex chromosomes, chromosome mitochodria and ENCDODE blacklist regions (Indexed with Tabix)	./autos_bfilt/meth_autos_bfilt_bin.bed	Bins coordinates and #CpG
	./autos_bfilt/meth_count_autos_bfilt.txt.gz	MEDIPS
	./autos_bfilt/meth_rpkm_autos_bfilt.txt.gz	
	./autos_bfilt/meth_rms_medips_autos_bfilt.txt.gz	
	./autos_bfilt/meth_rms_medestrand_autos_bfilt.txt.gz	MEDStrand
	./autos_bfilt/meth_nrpm_qsea_autos_bfilt.txt.gz	QSEA
	./autos_bfilt/meth_beta_qsea_autos_bfilt.txt.gz	
	./autos_bfilt/meth_logitbeta_qsea_autos_bfilt.txt.gz	
	./autos_bfilt/meth_CNV_qsea_autos_bfilt.txt.gz	