cfMeDIP-seq: Protocol, Computational Pipeline, and Classifier Methodologies

Executive Summary

This document provides comprehensive technical details on the cell-free methylated DNA immunoprecipitation and high-throughput sequencing (cfMeDIP-seq) methodology, including:

1. The initial protocol as described in Shen et al., Nature Protocols 2019
2. Sources of technical variability in the protocol
3. The MEDIPIPE computational analysis pipeline
4. Machine learning classifiers developed for cancer detection
5. Protocol and classifier improvements over time

---

1. Initial cfMeDIP-seq Protocol (Nature Protocols 2019)

1.1 Overview

The cfMeDIP-seq protocol was published by Shen, Burgener, Bratman, and De Carvalho in Nature Protocols (2019) as an adaptation of conventional MeDIP-seq for low-input cell-free DNA samples. The protocol enables methylome profiling from 1-10 ng of input DNA through the addition of "filler DNA" before immunoprecipitation.

Timeline: The protocol can be completed in approximately 3-4 days by a standard molecular biology laboratory.

Key Innovation: Addition of filler DNA (unmethylated lambda phage DNA or methylated genomic DNA) acts as a carrier to improve the efficiency and specificity of antibody-mediated capture of methylated cfDNA fragments.

1.2 Protocol Steps

The cfMeDIP-seq protocol consists of four major stages:

STAGE 1: Library Preparation from cfDNA (Day 1)

Step 1: cfDNA End-Repair and A-Tailing

- Input: 1-10 ng of plasma cfDNA
- Process: Repair DNA ends to create blunt ends with 5' phosphates
- Add adenosine (A) overhangs to 3' ends to enable adapter ligation
- Time: ~2-3 hours

- **Variability Sources**:
  - DNA input quantity affects downstream efficiency
  - Temperature control during enzymatic reactions
  - Timing of enzymatic reactions

Step 2: Adapter Ligation

- Ligate methylated sequencing adapters to DNA fragments
- Adapters contain unique molecular identifiers (UMIs) in some implementations
- Time: ~2-3 hours
- **Variability Sources**:
  - Adapter concentration and quality
  - Ligation efficiency (temperature, enzyme activity)
  - Adapter dimer formation

Step 3: Purification

- Remove excess adapters and reaction components
- Use SPRI beads (Solid Phase Reversible Immobilization)
- Time: ~1 hour
- **Variability Sources**:
  - Bead-to-DNA ratio affects recovery
  - Washing steps may cause sample loss
  - Ethanol quality and evaporation timing

STAGE 2: Methylated DNA Immunoprecipitation (Day 2)

Step 4: Addition of Filler DNA and Spike-In Controls

- Add filler DNA to reach total DNA amount of ~1 μg
- Filler DNA options:
  - Lambda phage DNA (unmethylated)
  - Human genomic DNA (methylated)
- Add Arabidopsis thaliana spike-in controls (methylated and unmethylated)
- Or add synthetic spike-in controls (Wilson et al., 2022)

- Time: ~30 minutes
- **Variability Sources** (HIGH IMPACT):
  - **Type of filler DNA** (methylated vs unmethylated) significantly affects results
  - Filler DNA quality and methylation status
  - Spike-in control quantity and integrity
  - This is a major source of batch effects between labs

Step 5: DNA Denaturation

- Heat DNA to 95°C for 10 minutes to denature double-stranded DNA
- Immediately place on ice
- Time: ~15 minutes
- **Variability Sources**:
  - Temperature accuracy
  - Denaturation time consistency
  - Cooling rate

Step 6: Immunoprecipitation

- Incubate denatured DNA with anti-5-methylcytosine (5mC) antibody
- Typically use 5 μg of antibody per reaction
- Incubate overnight at 4°C with rotation
- Antibody binds to methylated DNA fragments
- Time: ~16-18 hours (overnight)
- **Variability Sources** (HIGH IMPACT):
  - **Antibody lot-to-lot variation** (different antibody lots show different binding efficiencies)
  - Antibody concentration
  - Incubation time and temperature
  - Rotation speed/mixing efficiency
  - DNA concentration affects antibody binding

Step 7: Capture with Protein A/G Beads

- Add Protein A/G magnetic beads to capture antibody-DNA complexes

- Incubate for 2 hours at 4°C with rotation
- Time: ~2 hours
- **Variability Sources**:
  - Bead quality and binding capacity
  - Incubation time
  - Non-specific binding to beads

Step 8: Washing

- Wash beads 3-5 times with immunoprecipitation buffer
- Remove non-specifically bound DNA
- Critical step for specificity
- Time: ~30 minutes
- **Variability Sources** (MODERATE IMPACT):
  - Number and vigor of washes
  - Buffer composition and pH
  - Temperature during washing
  - Magnetic separation efficiency

Step 9: Elution

- Elute DNA from beads using proteinase K digestion
- Incubate at 50°C for 2-3 hours
- Time: ~3 hours
- **Variability Sources**:
  - Proteinase K activity
  - Elution efficiency
  - Temperature and timing

Step 10: Purification of Immunoprecipitated DNA

- Purify eluted DNA using SPRI beads
- Time: ~1 hour
- **Variability Sources**:

- Recovery efficiency
- Contamination with residual proteins

STAGE 3: Library Amplification (Day 3)

Step 11: qPCR Quality Control

- Perform qPCR on a small aliquot to assess:
  - Enrichment of methylated regions (positive control: HIST1H2BA)
  - Depletion of unmethylated regions (negative control: GAPDH)
  - Fold-enrichment ratio should be >25
- Calculate specificity using spike-in controls
- Specificity should be >95% (i.e., <5% of unmethylated DNA captured)
- Time: ~2 hours
- **Variability Sources**:
  - qPCR technical variability
  - Primer efficiency
  - Quality thresholds affect decision to proceed

Step 12: PCR Amplification

- Amplify library using indexed primers
- Typically 12-15 cycles of PCR
- Use high-fidelity polymerase
- Time: ~2 hours
- **Variability Sources** (MODERATE IMPACT):
  - **PCR cycle number** affects library complexity and duplication rate
  - PCR bias (GC-content, fragment length)
  - Primer concentration
  - Polymerase fidelity and processivity

Step 13: Final Purification and Size Selection

- Purify PCR products with SPRI beads
- Perform size selection (typically 200-600 bp)

- Time: ~1 hour
- **Variability Sources**:
  - Size selection stringency
  - Recovery efficiency

Step 14: Library Quantification and Quality Assessment

- Quantify library concentration using Qubit or qPCR
- Assess library size distribution using Bioanalyzer or TapeStation
- Check for adapter dimers
- Time: ~1 hour
- **Variability Sources**:
  - Quantification method accuracy
  - Fragment size distribution affects sequencing

STAGE 4: High-Throughput Sequencing (Day 4 and beyond)

Step 15: Sequencing

- Pool libraries with unique barcodes
- Perform paired-end sequencing (typically 2 × 75 bp or 2 × 150 bp)
- Platform: Illumina (HiSeq, NovaSeq, or NextSeq)
- Target depth: 5-20 million paired-end reads per sample
- **Variability Sources** (MODERATE IMPACT):
  - **Sequencing depth** significantly affects detection sensitivity
  - Sequencing quality (Q-scores)
  - Base-calling errors
  - Cluster density on flow cell

1.3 Quality Control Metrics

The protocol includes several QC checkpoints:

1. **Fold-Enrichment Ratio**: Ratio of methylated (HIST1H2BA) to unmethylated (GAPDH) regions
   - Threshold: >25-fold enrichment
   - Measures immunoprecipitation specificity

2. **Specificity**: Percentage of reads mapping to methylated spike-in controls vs unmethylated
    - Threshold: >95% specificity
    - Calculated using Arabidopsis thaliana spike-ins
3. **CpG Enrichment**: Relative frequency of CpGs in enriched regions vs genome
    - Indicates successful capture of methylated regions
4. **Saturation Analysis**: Assesses whether sequencing depth is adequate
    - Uses MEDIPS package
    - Determines if additional sequencing would provide more information
5. **Fragment Size Distribution**: Should match cfDNA size profile
    - Peak at ~166 bp (mononucleosomal)
    - Range: 100-700 bp

1.4 Key Sources of Variability

Based on the protocol and subsequent validation studies, the major sources of variability are:

Consistent (Systematic) Variability Sources:

1. **Type of Filler DNA** (HIGH IMPACT)
    - Methylated vs unmethylated filler creates systematic differences
    - Wilson et al. (2022) showed this accounts for ~5% of variance
    - Can be corrected with spike-in controls
2. **Antibody Lot** (HIGH IMPACT)
    - Different antibody lots have different binding characteristics
    - Affects enrichment efficiency consistently across a batch
    - Difficult to correct without spike-in standards
3. **Sequencing Depth** (MODERATE-HIGH IMPACT)
    - Lower depth reduces sensitivity for rare methylation events
    - Affects detection of low tumor fraction samples
    - Predictable effect that can be modeled
4. **PCR Cycle Number** (MODERATE IMPACT)
    - More cycles increase duplication rate
    - Introduces GC-content and length biases

- Systematic across samples in same batch

5. **Fragment Length, GC Content, and CpG Density** (MODERATE IMPACT)
    - Biophysical properties affect antibody binding
    - Longer fragments, higher GC, and higher CpG density are preferentially captured
    - Can be modeled and corrected with appropriate normalization

Sporadic (Random) Variability Sources:

1. **DNA Input Quality and Quantity** (HIGH IMPACT)
    - Degraded DNA reduces library quality
    - Very low input (<1 ng) increases stochasticity
    - Patient-to-patient variation in cfDNA characteristics
2. **Temperature Fluctuations** (LOW-MODERATE IMPACT)
    - During overnight incubation
    - During enzymatic reactions
    - Causes run-to-run variation
3. **Bead Handling** (LOW-MODERATE IMPACT)
    - Incomplete resuspension
    - Magnetic separation timing
    - Sample loss during transfers
4. **Operator Technique** (LOW-MODERATE IMPACT)
    - Pipetting accuracy and precision
    - Timing of steps
    - Contamination
5. **Sequencing Quality** (LOW-MODERATE IMPACT)
    - Flow cell quality variation
    - Cluster density effects
    - Lane effects on Illumina platforms

1.5 Improvements to Mitigate Variability

Synthetic Spike-In Controls (Wilson et al., 2022)

A major improvement to the protocol was the development of synthetic spike-in controls with defined characteristics:

- 54 DNA fragments with combinations of:
  - Methylation status (methylated and unmethylated)
  - Fragment length (80, 160, 320 bp)
  - GC content (35%, 50%, 65%)
  - CpG fraction (1/80 bp, 1/40 bp, 1/20 bp)
- Addition of 0.01 ng spike-in controls enables:
  - Absolute quantification of methylated cfDNA
  - Batch effect correction
  - Correction for fragment length, GC, and CpG biases
  - Quality control per sample

Benefits Demonstrated:

- Mitigates batch effects between laboratories
- Corrects for filler DNA type (methylated vs unmethylated)
- Provides absolute quantification standard
- Enables robust inter-batch comparisons

Unique Molecular Identifiers (UMIs) (Burgener et al., 2021)

Addition of UMIs to sequencing adapters enables:

- Removal of PCR duplicates while preserving biological duplicates
- Correction for sequence errors
- Improved quantification accuracy
- Essential for MRD applications with very low tumor fractions

---

2. MEDIPIPE Computational Analysis Pipeline

2.1 Overview

MEDIPIPE is an automated, end-to-end pipeline for cfMeDIP-seq data quality control, methylation quantification, and sample aggregation, published by Zeng et al. in Bioinformatics (2023). The pipeline was developed using Snakemake for reproducibility and scalability.

Key Features:

- Automated dependency installation via Conda
- Support for single-end or paired-end sequencing
- Handles spike-in controls and UMIs
- Parallel processing for large-scale datasets
- Comprehensive quality control reporting

2.2 Pipeline Architecture

MEDIPIPE consists of four major modules:

```
Input: FASTQ files + Configuration file
↓
Module 1: Read Trimming and QC
↓
Module 2: Processed Read Alignment and QC
↓
Module 3: DNA Methylation Quantification
↓
Module 4: Data Aggregation and Filtering
↓
Output: QC reports, BAM files, Methylation matrices
```

2.3 Module 1: Read Trimming and Quality Control

Tools Used:

- **UMI-tools** (v1.0.1): Extract UMI barcodes if present
  - Uses regex method for variable barcode length
  - Filters reads without expected barcode
- **Trim Galore** (v0.6.7): Adapter trimming and quality filtering
  - Removes adapter sequences
  - Trims low-quality bases
  - Works with both single-end and paired-end data
- **FastQC** (v0.11.9): Quality assessment
  - Analyzes both raw and processed reads
  - Generates quality metrics and visualizations

Outputs:

- Trimmed FASTQ files
- FastQC reports (HTML and data files)
- UMI extraction statistics (if UMIs present)

Key Metrics:

- Read counts before and after trimming
- Base quality distributions
- Adapter content
- Sequence duplication levels

2.4 Module 2: Processed Read Alignment and QC

Tools Used:

- **BWA-MEM** (v0.7.17): Read alignment
  - Maps processed reads to reference genome
  - For spike-in experiments, spike-in sequences appended to reference
  - Paired-end mode preserves fragment information
- **SAMtools** (v1.17): Alignment processing
  - Filters unmapped reads
  - Removes secondary alignments
  - Filters improperly paired reads (for paired-end)
  - Marks or removes duplicates (or uses UMI-tools dedup if UMIs present)
- **Picard** (v2.26.6): Fragment size metrics (paired-end only)
  - CollectInsertSizeMetrics for fragment length distribution
  - Computes short-to-long fragment ratios in 1 Mb and 5 Mb windows
- **MEDIPS**: cfMeDIP-seq specific QC
  - Saturation analysis
  - Coverage metrics
  - Enrichment scores
  - Spike-in control quantification

Outputs:

- Aligned BAM files (sorted and indexed)
- Alignment statistics
- Duplication metrics
- Fragment size metrics (paired-end)
- cfMeDIP-seq QC metrics
- Extracted spike-in reads (if applicable)

Key Metrics:

- Mapping rate
- Duplication rate
- Fragment size distribution (paired-end)
- Saturation curves
- CpG enrichment scores
- Spike-in control recovery

2.5 Module 3: DNA Methylation Quantification

MEDIPIPE implements three methods for methylation estimation, each using different models to correct for CpG density bias:

Method 1: MEDIPS (v1.46.0) - Linear Regression Model

Approach:

- Divides genome into fixed-size windows (typically 300-500 bp)
- Counts reads in each window
- Applies linear regression to model relationship between read counts and CpG density
- Calculates Relative Methylation Score (RMS) for each window

Normalization:

- Can normalize using spike-in controls
- Corrects for sequencing depth
- Adjusts for CpG density

Output:

- Methylation values per genomic window

- CpG enrichment scores
- Saturation metrics

Best For:

- Initial methylation profiling
- When spike-in controls are available
- Comparative studies

Method 2: QSEA (v1.20.0) - Sigmoidal Model with Empirical Knowledge

Approach:

- Uses sigmoidal model to capture non-linear relationship between enrichment and CpG density
- Incorporates empirical calibration curve
- Models enrichment as function of:
  - CpG density
  - Fragment length
  - GC content

Advantages:

- Better handling of CpG-poor regions
- More accurate absolute methylation estimates
- Reduced bias in low-density regions

Output:

- Calibrated methylation scores
- Can estimate absolute methylation levels

Best For:

- Studies requiring absolute methylation quantification
- Analysis of CpG-poor regions
- Cross-study comparisons

Method 3: MEDStrand (v0.0.0.9000) - Stranded Sigmoid Model

Approach:

- Extension of sigmoidal model that accounts for strand-specific effects

- Models forward and reverse strand separately
- Improves accuracy for asymmetric methylation patterns

Advantages:

- Most sophisticated modeling approach
- Accounts for strand-specific biases
- Highest accuracy for absolute methylation

Output:

- Strand-aware methylation estimates
- Improved absolute quantification

Best For:

- High-precision applications
- Studies investigating strand-specific methylation
- Clinical diagnostic applications

2.6 Module 4: Data Aggregation and Filtering

Aggregation Functions:

1. **QC Report Aggregation**
   - Uses **MultiQC** (v1.0.dev0) to aggregate:
     - FastQC reports
     - SAMtools statistics
     - Picard metrics
     - MEDIPS QC metrics
   - Generates interactive HTML summary report
   - Enables cross-sample QC visualization
2. **Methylation Matrix Aggregation**
   - Combines methylation quantifications across all samples
   - Creates bin-by-sample matrices for each quantification method
   - Outputs as TAB-delimited TXT files
3. **Filtering**
   - Removes problematic regions:

    - Sex chromosomes (X, Y)
    - Mitochondrial chromosome
    - ENCODE blacklist regions
    - Regions with low mappability (Umap k100 < 0.5)
  - Creates both filtered and unfiltered output files

4. **Indexing**
  - Uses **Tabix** (v1.17) to index aggregated files
  - Enables rapid retrieval of specific genomic regions
  - Facilitates downstream analysis

Outputs:

- Comprehensive MultiQC report (HTML)
- Summarized QC metrics spreadsheet
- Aggregated methylation matrices (original and filtered)
- Tabix-indexed files for rapid access

2.7 Output File Structure

```
output/
├── fastqc/                 # FastQC reports per sample
├── trimmed/                # Trimmed FASTQ files
├── aligned/                # BAM files per sample
├── qc/                     # Sample-level QC metrics
│   ├── samtools_stats/
│   ├── picard_metrics/
│   └── medips_qc/
├── fragmentomics/          # Fragment size metrics (paired-end)
├── methylation/            # Methylation quantifications per sample
│   ├── MEDIPS/
│   ├── QSEA/
│   └── MEDStrand/
└── aggregated/             # Project-level aggregated outputs
    ├── multiqc_report.html
    ├── qc_summary.csv
    ├── methylation_MEDIPS.txt.gz
    ├── methylation_MEDIPS_filtered.txt.gz
    ├── methylation_QSEA.txt.gz
```

```
├── methylation_QSEA_filtered.txt.gz
├── methylation_MEDStrand.txt.gz
└── methylation_MEDStrand_filtered.txt.gz
```

2.8 Key Quality Control Metrics Tracked

MEDIPIPE tracks comprehensive QC metrics throughout the pipeline:

Sequencing Quality:

- Total read counts
- Read quality scores (Q20, Q30)
- Adapter content
- Duplication rates
- UMI diversity (if applicable)

Alignment Quality:

- Mapping rate
- Properly paired rate (paired-end)
- Secondary alignment rate
- Supplementary alignment rate

cfMeDIP-seq Specific:

- CpG enrichment score
- Saturation estimates
- Spike-in recovery rates
- Specificity calculations
- Fold-enrichment ratios

Fragment Characteristics:

- Fragment size distribution
- Short-to-long fragment ratios
- Fragmentation patterns

2.9 Configuration and Flexibility

MEDIPIPE uses a single YAML configuration file that specifies:

Experimental Settings:

- Single-end or paired-end sequencing
- Presence of UMIs
- Presence of spike-in controls
- Window size for quantification

File Paths:

- Input FASTQ files
- Reference genome
- Spike-in sequences (if applicable)
- Output directory

Parameters:

- Trimming parameters
- Alignment parameters
- Quantification method selection
- Filtering thresholds

Computational Resources:

- Number of threads per step
- Memory allocation
- Cluster submission parameters

2.10 Advantages of MEDIPIPE

1. **Reproducibility**: Containerized execution with automated dependency management
2. **Scalability**: Parallel processing of independent samples
3. **Flexibility**: Single configuration file handles various experimental designs
4. **Comprehensive**: End-to-end solution from raw reads to quantified methylation
5. **Standardization**: Consistent processing across batches and studies
6. **Quality Control**: Extensive QC metrics at each step
7. **Multiple Algorithms**: Provides three quantification methods for comparison

---

3. Machine Learning Classifiers for cfMeDIP-seq Data

3.1 Overview

Multiple machine learning approaches have been developed for cancer classification using cfMeDIP-seq data. This section details the classifiers from the key publications.

3.2 Shen et al. 2018 (Nature) - Foundational Classifier

Study: "Sensitive tumour detection and classification using plasma cell-free DNA methylomes"

3.2.1 Data Processing

Feature Definition:

- Genome divided into 300 bp non-overlapping windows
- Windows mapped to:
  - CpG islands
  - CpG shores (regions flanking CpG islands)
  - CpG shelves (regions flanking shores)
  - FANTOM5 enhancers (regulatory elements)
- This captures ~2.8 million windows genome-wide

Feature Selection:

- Identified Differentially Methylated Regions (DMRs) between cancer and controls
- Selected top 1,000 most variable windows
- Variability measured by median absolute deviation (MAD)
- Features: Read counts in selected 300 bp windows

3.2.2 GLMnet Classifier

Algorithm: Elastic Net Regularized Logistic Regression (GLMnet)

Model Details:

- GLMnet combines L1 (Lasso) and L2 (Ridge) regularization
- Regularization path: $\alpha = 0.5$ (equal weight to L1 and L2)
- λ (regularization strength) selected by cross-validation
- Performs automatic feature selection via L1 penalty
- Typically retains 50-200 features from the 1,000 input features

Training:

- 10-fold cross-validation for hyperparameter tuning
- Nested cross-validation for performance estimation
- Class imbalance handled by:
  - Weighted loss function
  - Or downsampling of majority class

Output:

- Probability score (0-1) for cancer presence
- Feature coefficients (which DMRs are most informative)

Performance (Pancreatic Cancer Example):

- Discovery cohort AUROC: 0.96
- Validation cohort AUROC: 0.94
- Sensitivity at 98% specificity: 69%

3.2.3 Random Forest Classifier

Algorithm: Ensemble of Decision Trees

Model Details:

- Number of trees: 500-1000
- mtry (features per split): $\sqrt{}$(number of features) $\approx$ 31 for 1,000 features
- Minimum node size: 5
- Out-of-bag (OOB) error used for internal validation

Training:

- Bootstrap sampling for each tree
- Features randomly selected at each split
- Majority voting for final prediction
- Variable importance calculated by permutation

Output:

- Probability score (proportion of trees voting for cancer)
- Variable importance rankings

- Proximity matrix (sample similarity)

Performance:

- Similar to GLMnet in most cancer types
- More robust to outliers
- Better captures non-linear relationships

3.2.4 Multi-Cancer Classification

Approach:

- Hierarchical classification strategy:
  1. **Binary Classifier**: Cancer vs. Non-Cancer
  2. **Multi-Class Classifier**: Classify cancer type

Cancer vs. Non-Cancer:

- Used same GLMnet approach
- Top 1,000 DMRs distinguishing cancer from healthy

Cancer Type Classification:

- Separate GLMnet model
- Top 300 DMRs for each pairwise cancer type comparison
- One-vs-rest strategy for multi-class
- Combined probabilities using softmax

Performance:

- Cancer detection AUROC: 0.90-0.99 across types
- Tissue-of-origin prediction accuracy: 80-95%

3.3 Halla-aho & Lähdesmäki 2022 (BMC Bioinformatics) - Comprehensive Method Comparison

Study: "Probabilistic modeling methods for cell-free DNA methylation based cancer classification"

This study systematically compared various statistical methods for cfMeDIP-seq classification.

3.3.1 Feature Selection Methods Evaluated

1. t-Test

- Classical two-sample t-test

- Identifies regions with significant mean differences
- Fast and simple
- Assumes normality

2. Moderated t-Test (limma)

- Borrows information across features
- More stable with small sample sizes
- Better false discovery rate control
- Implementation: Bioconductor limma package

3. Fisher's Exact Test

- Non-parametric approach
- Converts continuous methylation to binary (hyper/hypomethylated)
- Tests association using contingency tables
- More robust to outliers
- **Finding: Most robust feature selection method**

4. Differentially Methylated Regions (DMRs)

- Top N regions by test statistic
- Typical N: 300-1,000
- Selected separately for each cancer type

3.3.2 Feature Generation Methods Evaluated

1. Raw Features

- Direct use of read counts in genomic windows
- No dimensionality reduction
- Baseline approach

2. Principal Component Analysis (PCA)

- Unsupervised dimensionality reduction
- Projects features onto principal components
- Captures maximum variance
- Limitation: May not capture cancer-relevant variation

3. Iterative Supervised PCA (ISPCA)

- Supervised variant of PCA
- Iteratively selects features most associated with outcome
- Performs PCA on selected features
- Combines feature selection with dimensionality reduction
- **Finding: Most robust feature generation method**

ISPCA Algorithm:

1. Initialize: Select features by univariate test
2. Iteration:
   a. Perform PCA on selected features
   b. Extract top K principal components
   c. Fit classifier using PCs
   d. Rank features by association with outcome
   e. Update selected features
3. Converge when feature set stabilizes

3.3.3 Classification Methods Evaluated

1. GLMnet (Elastic Net)

- Same as Shen et al. 2018
- Best with raw features
- Automatic feature selection

2. Logistic Regression with L1 Penalty (Lasso)

- Simpler than GLMnet ($\alpha = 1$)
- Performs feature selection
- More interpretable coefficients

3. Logistic Regression with L2 Penalty (Ridge)

- No feature selection ($\alpha = 0$)
- All features retained with shrinkage
- Better with highly correlated features

4. Logistic Regression with Bayesian Priors

a) **Horseshoe Prior:**

- Heavily shrinks uninformative features to zero
- Minimal shrinkage for informative features
- Automatic feature selection
- Implemented in rstanarm package

b) **Regularized Horseshoe Prior:**

- Extension with additional regularization
- Better finite-sample properties
- More stable than standard horseshoe

5. Simple Logistic Regression

- Features: Number of hyper- and hypomethylated regions
- Only 2 features!
- Regions defined by Fisher's exact test
- Surprisingly effective
- **Finding: Simple and robust**

3.3.4 Key Findings from Systematic Comparison

Best Overall Methods:

1. **Fisher's Exact Test + ISPCA + Logistic Regression**: Most robust across datasets
2. **Simple Logistic Regression** (2 features): Nearly as good, much simpler
3. **GLMnet**: Good performance but more variable

Effect of Sequencing Depth:

- All methods degrade with lower depth
- Fisher's exact test most robust to low depth
- ISPCA maintains performance better than PCA
- Depth < 5M reads significantly impacts sensitivity

Effect of Tumor Fraction:

- Critical factor for all methods
- <1% tumor fraction: All methods struggle
- 1-5% tumor fraction: Best methods achieve 70-90% sensitivity

- 5% tumor fraction: >95% sensitivity achievable

Recommendations:

- Use Fisher's exact test for feature selection
- Consider ISPCA for dimensionality reduction
- Simple logistic regression sufficient for many applications
- GLMnet when interpretability less critical

3.4 Nuzzo et al. 2019/2020 - Renal Cell Carcinoma Classifier

Studies:

- ASCO 2019 Abstract: "Cell-free methylated DNA immunoprecipitation and high throughput sequencing technology in patients with clear cell renal cell carcinoma"
- Nature Medicine 2020: "Detection of renal cell carcinoma using plasma and urine cell-free DNA methylomes"

3.4.1 Model Architecture

Feature Definition:

- 300 bp genomic windows
- Mapped to CpG islands, shores, shelves, FANTOM5 enhancers
- Same as Shen et al. 2018

Feature Selection:

- Top 300 DMRs distinguishing RCC from controls
- Selected by statistical significance (adjusted p-value)

Classifier:

- GLMnet (elastic net regularized logistic regression)
- Same architecture as Shen et al. 2018

3.4.2 Performance

Plasma cfDNA:

- Discovery cohort: AUROC = 0.990 (95% CI: 0.984-0.997)
- Validation cohort: Independent set
- Stage I sensitivity: ~75-80%

- Overall sensitivity: >90%

Urine cfDNA (Novel Finding):

- First demonstration of urine cfMeDIP-seq for cancer detection
- AUROC = 0.791 (95% CI: 0.759-0.823)
- Lower than plasma but still clinically useful

RCC vs. UBC (Bladder Cancer) Discrimination:

- AUROC = 0.954 (95% CI: 0.940-0.969)
- Demonstrates tissue-of-origin capability

Key Innovation:

- Demonstrated cfMeDIP-seq works in urine
- Important for urological cancers
- Non-invasive alternative to cystoscopy

3.5 De Pascali et al. 2024 - BRCA1/2 Breast Cancer Classifier

Study: "Differential methylation of circulating free DNA assessed through cfMeDiP as a new tool for breast cancer diagnosis and detection of BRCA1/2 mutation" (Journal of Translational Medicine)

3.5.1 Study Design

Cohort:

- 23 breast cancer patients with BRCA1/2 germline mutations
- 19 healthy controls (no BRCA1/2 mutation)
- 2 healthy BRCA1/2 carriers
- Blood samples at diagnosis

Feature Definition:

- 300 bp differentially methylated regions (DMRs)
- Identified between BRCA carriers and controls

Feature Selection:

- 7,095 hypermethylated regions identified
- 212 hypomethylated regions identified

- Biological focus: DNA repair and cell cycle genes

3.5.2 Classifiers Implemented

Classifier 1: GLMnet

- Elastic net logistic regression
- L1 and L2 penalties
- Cross-validation for parameter selection

Classifier 2: Random Forest

- Ensemble of 500 decision trees
- Variable importance by permutation
- Out-of-bag error estimation

Comparison with Public Datasets:

- DMRs compared to TCGA tumor methylation data
- Validation of tumor-specific signatures

3.5.3 Performance

Cancer Detection:

- High accuracy discriminating BRCA-positive from healthy
- Both classifiers showed comparable performance
- Specific details not provided in abstract

Biological Insights:

- BRCA1/2 breast cancers show hypomethylation of DNA repair genes
- Cell cycle regulatory genes differentially methylated
- Potential for subtype-specific biomarkers

Clinical Implications:

- Could identify BRCA carriers at high risk
- Potential for early detection in high-risk population
- Response to PARP inhibitors may correlate with methylation profile

---

4. Protocol and Classifier Improvements Over Time

4.1 Timeline of Major Improvements

2018 - Shen et al. (Nature)

- Original cfMeDIP-seq protocol
- Arabidopsis thaliana spike-ins for QC
- GLMnet and Random Forest classifiers
- Demonstrated feasibility in multiple cancer types

2019 - Shen et al. (Nature Protocols)

- Detailed protocol publication
- Standardized 3-4 day workflow
- Quality control metrics defined
- Made method accessible to community

2021 - Burgener et al.

- Addition of Unique Molecular Identifiers (UMIs)
- Improved duplicate removal
- Sequence error correction
- Critical for MRD applications

2022 - Wilson et al. (Cell Reports Methods)

- Synthetic spike-in controls
- 54 fragments with defined characteristics
- Absolute quantification capability
- Batch effect correction
- Major improvement for clinical applications

2022 - Halla-aho & Lähdesmäki (BMC Bioinformatics)

- Systematic comparison of classifiers
- Identified optimal methods
- Simple logistic regression found highly effective
- Guidelines for method selection

2023 - Zeng et al. (Bioinformatics)

- MEDIPIPE automated pipeline
- Standardized processing
- Multiple quantification algorithms
- Large-scale analysis capability

2024 - De Pascali et al. (J Translational Med)

- Application to BRCA1/2 breast cancer
- Demonstrated subtype specificity
- Integration with genomic mutation data

4.2 Key Protocol Improvements

Improvement 1: Synthetic Spike-In Controls

Problem Addressed:

- Batch effects between laboratories
- Variation in filler DNA type
- Inability to absolutely quantify methylation
- Fragment length, GC, and CpG biases

Solution (Wilson et al., 2022):

- 54 synthetic DNA fragments
- Systematic variation in:
  - Methylation status
  - Fragment length (80, 160, 320 bp)
  - GC content (35%, 50%, 65%)
  - CpG fraction (1/80, 1/40, 1/20 bp)
- 0.01 ng added per reaction

Benefits:

- Absolute quantification via linear model
- Batch effect correction
- Per-sample quality control

- Correction for biophysical biases
- Successfully mitigated inter-lab variation

Improvement 2: Unique Molecular Identifiers (UMIs)

Problem Addressed:

- PCR duplicates vs biological duplicates
- Sequence errors introduced during amplification
- Quantification accuracy at low tumor fractions

Solution (Burgener et al., 2021):

- UMI barcodes in adapter sequences
- Molecular-level tracking of DNA molecules
- Computational deduplication using UMI-tools

Benefits:

- Accurate duplicate removal
- Sequence error correction (consensus calling)
- Improved quantification, especially for MRD
- Essential for detecting low-level disease

Improvement 3: Automated Pipeline (MEDIPIPE)

Problem Addressed:

- Lack of standardized analysis workflow
- Different labs using different processing steps
- Difficult to compare across studies
- Manual analysis time-consuming and error-prone

Solution (Zeng et al., 2023):

- End-to-end automated pipeline
- Snakemake workflow management
- Containerized execution
- Comprehensive QC reporting

Benefits:

- Standardized processing
- Reproducibility
- Scalability to large datasets
- Reduced analysis time
- Facilitates cross-study comparisons

4.3 Classifier Evolution

From Single-Method to Systematic Comparison

Early Approach (2018):

- GLMnet proposed as standard
- Random Forest as alternative
- Limited comparison of methods

Systematic Evaluation (2022):

- Halla-aho & Lähdesmäki compared:
  - Multiple feature selection methods
  - Multiple feature generation methods
  - Multiple classifiers
  - Bayesian approaches

Key Finding:

- Simpler can be better
- 2-feature logistic regression nearly as good as complex methods
- Fisher's exact test most robust feature selection

From Binary to Multi-Class Classification

Progression:

1. Single cancer type detection (2018)
2. Multiple independent binary classifiers
3. Hierarchical classification:
   - First: Cancer vs. No Cancer
   - Then: Cancer type classification

4. Multi-class probabilistic output

From Research to Clinical Application

Evolution:

- Academic proof-of-concept (2018)
- Method standardization (2019)
- Clinical validation studies (2020-2024)
- Commercial development (Adela, Inc.)

Clinical Requirements Driving Improvements:

- Specificity >99% to minimize false positives
- Sensitivity across early stages
- Tissue-of-origin prediction
- Batch effect robustness
- Quality control per sample
- Absolute quantification

4.4 Remaining Challenges and Future Directions

Current Limitations:

1. **Sensitivity at Very Low Tumor Fractions (<0.5%)**
   - All methods struggle below this threshold
   - Important for earliest stage cancers
   - MRD monitoring challenges
2. **Standardization Across Laboratories**
   - Despite improvements, inter-lab variation remains
   - Antibody lot variation difficult to control
   - Need for universal reference standards
3. **Cost**
   - Sequencing depth requirement (5-20M reads per sample)
   - Antibody and reagent costs
   - Not yet cost-competitive with other screening methods

4. **Computational Complexity**
    - Large-scale data processing requirements
    - Need for bioinformatics expertise
    - Analysis time for clinical turnaround

Future Directions:

1. **Enhanced Spike-In Standards**
    - More comprehensive coverage of biophysical space
    - Universal reference materials
    - Certified reference standards for clinical use
2. **Machine Learning Advances**
    - Deep learning approaches
    - Integration with other liquid biopsy modalities
    - Pan-cancer models
3. **Single-Molecule Approaches**
    - Long-read sequencing integration
    - Direct methylation detection
    - Reduced amplification bias
4. **Clinical Integration**
    - FDA approval pathways
    - Health economics studies
    - Integration into screening guidelines
    - Real-world evidence generation
5. **Multi-Omics Integration**
    - Combine methylation with:
        - cfDNA fragmentation patterns
        - Mutation detection
        - Protein biomarkers
    - Improve sensitivity and specificity

---

5. Comparison with Other Methylation Detection Methods

5.1 cfMeDIP-seq vs Bisulfite Sequencing

cfMeDIP-seq Advantages:

- No DNA degradation (preserves cfDNA)
- Lower input requirements (1-10 ng vs 10-100 ng)
- Genome-wide coverage
- Lower cost per sample
- Faster library preparation

Bisulfite Sequencing Advantages:

- Single-CpG resolution
- Absolute methylation quantification (without modeling)
- Gold standard for validation
- No antibody-dependent variability

cfMeDIP-seq vs WGBS/RRBS:

- WGBS (Whole Genome Bisulfite Sequencing):
  - Comprehensive but very expensive
  - High input requirements
  - cfMeDIP-seq 5-10× lower cost
- RRBS (Reduced Representation Bisulfite Sequencing):
  - Focuses on CpG-rich regions
  - Misses enhancers and CpG-poor regions
  - cfMeDIP-seq provides broader coverage

5.2 cfMeDIP-seq vs Targeted Methylation Methods

Targeted Methylation (e.g., GRAIL Galleri):

- Focuses on ~100,000 selected regions
- Deep coverage of informative regions
- Pre-defined feature set
- High specificity

cfMeDIP-seq (Adela):

- Genome-wide discovery approach
- Single assay for all applications
- Flexible feature selection
- Platform adaptability

---

6. Summary and Recommendations

6.1 Protocol Best Practices

For Optimal Results:

1. **Use Synthetic Spike-In Controls** (Wilson et al., 2022)
   - Essential for batch effect correction
   - Enables absolute quantification
   - Provides per-sample QC
2. **Implement UMIs for MRD Applications**
   - Critical for low tumor fraction detection
   - Improves quantification accuracy
   - Enables error correction
3. **Maintain Consistent Antibody Lots**
   - Major source of batch effects
   - Validate new lots with standards
   - Consider bulk purchasing
4. **Use Methylated Filler DNA**
   - More consistent than unmethylated
   - Better performance in most studies
   - Can be corrected with spike-ins if unmethylated used
5. **Follow Quality Control Thresholds Strictly**
   - Fold-enrichment >25
   - Specificity >95%
   - CpG enrichment >1.5

6.2 Computational Analysis Recommendations

For Reproducibility and Quality:

1. **Use MEDIPIPE or Standardized Pipeline**
   - Ensures consistent processing
   - Comprehensive QC
   - Community-validated approach
2. **Select Quantification Method Based on Application**
   - MEDIPS: General comparative studies
   - QSEA: When absolute quantification needed
   - MEDStrand: High-precision applications
3. **Implement Multiple Classifier Approaches**
   - Test GLMnet and simpler methods
   - Use cross-validation appropriately
   - Consider ensemble approaches
4. **Validate on Independent Cohorts**
   - Avoid overfitting
   - Test across sequencing depths
   - Assess performance at different tumor fractions

6.3 Future Research Priorities

1. **Development of Universal Standards**
   - Reference materials for inter-lab QC
   - Certified spike-in controls
   - Proficiency testing programs
2. **Integration with Clinical Workflows**
   - Turn-around time optimization
   - Automated result reporting
   - Clinical decision support tools
3. **Machine Learning Innovation**
   - Deep learning for feature learning

- Multi-task learning for multiple cancers
- Interpretable AI for clinical acceptance

4. **Large-Scale Validation Studies**
   - Prospective clinical trials
   - Real-world evidence generation
   - Health economics analysis

---

7. References

1. Shen SY, Singhania R, Fehringer G, et al. Sensitive tumour detection and classification using plasma cell-free DNA methylomes. Nature. 2018;563(7732):579-583. doi:10.1038/s41586-018-0703-0
2. Shen SY, Burgener JM, Bratman SV, De Carvalho DD. Preparation of cfMeDIP-seq libraries for methylome profiling of plasma cell-free DNA. Nat Protoc. 2019;14(10):2749-2780. doi:10.1038/s41596-019-0202-2
3. Wilson SL, Shen SY, Harmon L, et al. Sensitive and reproducible cell-free methylome quantification with synthetic spike-in controls. Cell Rep Methods. 2022;2(9):100294. doi:10.1016/j.crmeth.2022.100294
4. Burgener JM, Zou J, Zhao Z, et al. Tumor-naïve multimodal profiling of circulating tumor DNA in head and neck squamous cell carcinoma. Clin Cancer Res. 2021;27(15):4230-4244.
5. Zeng Y, Ye W, Stutheit-Zhao EY, et al. MEDIPIPE: an automated and comprehensive pipeline for cfMeDIP-seq data quality control and analysis. Bioinformatics. 2023;39(7):btad423. doi:10.1093/bioinformatics/btad423
6. Halla-aho V, Lähdesmäki H. Probabilistic modeling methods for cell-free DNA methylation based cancer classification. BMC Bioinformatics. 2022;23:138. doi:10.1186/s12859-022-04651-9
7. Nuzzo P, Spisak S, Chakravarthy A, et al. Cell-free methylated DNA (cfMeDNA) immunoprecipitation and high throughput sequencing technology (cfMeDIP-seq) in patients with clear cell renal cell carcinoma (ccRCC). J Clin Oncol. 2019;37(15_suppl):3052.
8. Nuzzo PV, Berchuck JE, Korthauer K, et al. Detection of renal cell carcinoma using plasma and urine cell-free DNA methylomes. Nat Med. 2020;26(7):1041-1043. doi:10.1038/s41591-020-0933-1
9. De Pascali SA, et al. Differential methylation of circulating free DNA assessed through cfMeDiP as a new tool for breast cancer diagnosis and detection of BRCA1/2 mutation. J Transl Med. 2024;22:934. doi:10.1186/s12967-024-05734-2

10. Nassiri F, Chakravarthy A, Feng S, et al. Detection and discrimination of intracranial tumors using plasma cell-free DNA methylomes. Nat Med. 2020;26(7):1044-1047. doi:10.1038/s41591-020-0932-2

Document prepared: November 2025

Note: This document is intended for research and educational purposes. The cfMeDIP-seq technology and associated tests are not FDA approved or cleared for clinical use. This document reflects published methodologies and may not represent current commercial implementations.