Introduction to DNA Methylation Data Analysis

Jenny Wu
Director of Bioinformatics
Genomics Research and Technology Hub
Chao Family Comprehensive Cancer Center
UC Irvine
Outline

• Why DNA methylation
• Technologies for genome wide DNA methylation profiling
• Experimental design and data analysis pipeline
  • Infinium array based
  • Sequencing based: BS-Seq (RRBS, WGBS), EM-Seq, Pacbio, Nanopore
• DNA Methylation data analysis
  ✓ Pre-processing: platform dependent
  ✓ *Downstream Analysis: DMS, DMR*
  ✓ *Software choices*
Why DNA Methylation

• One of the most well studied epigenetic mechanisms

• Plays important roles in genomic imprinting, transposon inactivation, stem cell pluripotency and differentiation
Commonly Used Profiling Methods

- Illumina Infinium Array
- Short and Long read sequencing

- MBD Enrichment
- Antibody Enrichment
- Bisulphite Conversion
- TET Oxidation

- Genomic DNA
- Sequencing
Experimental Design

• Choose the platforms/protocols based on the biological question you wish to address
  • Array based or Sequencing based, cost, throughput
    • Targeted (RRBS, twist) or whole methylome
• For BS-Seq and EM-Seq 30x coverage (PE100, 500M+ reads)
• 3+ replicates per group
• Do not confound sample groups by sex, age, or batch
Infinium Methylation BeadChip

• Two types of beads to detect methylation

• Target site matches with the probe, enabling single-base extension and detection. Single-base mismatch inhibits extension

• M value: Ratio of the intensity from M/U bead probes
Illumina Methylation Arrays

- **Infinium Human MethylationEPIC Array**
  - V2 has ~935k methylation sites with coverage of RefSeq genes, CpG islands, enhancer regions, open chromatin sites etc.
  - Compatible with FFPE samples
  - 8 Samples per BeadChip

- **Infinium Mouse Methylation Array**
  - >285k methylation sites per sample at single-nucleotide resolution
  - 24 samples per BeadChip
Infinium Methylation Assay Workflow

Bisulfite Conversion  →  Whole Genome amplification and enzyme fragmentation  →  Array Processing Scanning

Template: (Double Stranded)
A: 5’-GACCGTCCAGCTGCAGTCGTGCT-3’
B: 3’-CTGGCAACGCAGTCGTGCTGAGGA-5’

Bisulfite Converted: (Single Stranded)
A: 5’-GATCATTCTTTAGCTACTGTTT-3’
B: 3’-TTCGTAACCTTTTAGCTACTGTTT-5’
Methylation Array Data Analysis Work Flow

**Data Generation**
- Lab Preparation
- Array Processing
- Idat file

**QC and Pre Processing**
- Quality Control and Normalization
- Bisulfite conversion
- Beta value density
- SWAN
- GenomeStudio

**Methylation Calling**
- M values and Beta values for each site
- GenomeStudio, minfi, Champ, missMethyl

**Differential Methylation**
- Identify differentially methylated sites and regions
- Limma, bumphunter, DMRFinder

*Common Software Tools*
Illumina GenomeStudio

- Basic QC, visualization and analysis and generate reports.
- Does not container control probes
- No advanced analysis

https://support.illumina.com/array/array_software/genomestudio/downloads.html
GenomeStudio Methylation Module v2011

• Basic QC, methylation calling and two group comparison

• View CpG island methylation status across the genome

• Generate plots for the single-site resolution data
R Based Tools for Methylation

• End to End solution for methylation Analysis
  • Preprocessing, QC assessment, normalization, methylation calling, plotting functionality (MDS)
  • DMC and DMR analysis
• **Minfi, Champ, SeSaMe**
Methylation Array Pricing

• Reagent Expenses:
  • Human: Infinium Methylation EPIC v2.0 Kit – 8 samples per chip
  • Mouse: Infinium Mouse Methylation BeadChip kit - 24 samples per chip

• Labor Expense: $425 per chip

<table>
<thead>
<tr>
<th># samples</th>
<th>Human Kit ($)</th>
<th>Mouse Kit ($)</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>2,392</td>
<td>5,366</td>
</tr>
<tr>
<td>16</td>
<td>4,576</td>
<td>9,734</td>
</tr>
<tr>
<td>32</td>
<td>8,320</td>
<td>19,468</td>
</tr>
<tr>
<td>96</td>
<td>24,960</td>
<td></td>
</tr>
</tbody>
</table>
WGBS Workflow

1. Sample QC
   - Sample Preparation
2. Bisulfite Treatment
3. Library QC
   - Library Construction
4. Sequencing
5. Data QC
   - Bioinformatics Analysis
Drawbacks of BS-Seq

• Harsh chemical reaction causes severe DNA degradation, therefore short DNA fragment

• Reduced sequence complexity caused by C-to-T conversion of unmodified cytosines, which accounts for ~95% of all cytosines in the human genome. Poor sequencing quality, low mapping rate and uneven genome coverage.
EM-Seq Mechanism and Workflow
Enzymatic Methyl Sequencing (EM-Seq)

- Superior sensitivity of detection of 5-mC and 5-hmC
- More uniform GC coverage
- Detection of more CpGs with fewer sequence reads
- Uniform dinucleotide distribution
- High-efficiency library preparation, with larger library insert sizes
- Good for challenging samples and lower DNA inputs, opens new avenues for research and clinical applications
Methyl-Seq Data Analysis Work Flow

**Data Generation**
- Lab Preparation
- Sequencing
- FASTQ files

**QC and Alignment**
- Quality Control and Alignment to ref genome
  - FastQC, Qualimap
  - bowtie2, bwa-meth

**Methylation Calling**
- Deduplication M/Beta values for each site
  - Picard-tools, bismark, BSMAP, MethylDakel

**Differential Methylation**
- Identify differentially methylated sites and regions
  - bisseq, camel, DSS, methylSeekR

*Common Software Tools*
Bismark Alignment and Methylation Calling

A genomic fragment sequence after bisulfite treatment

...ccggctgtttaatgtct...
TTGGCATGTTAACGT

C-to-T

TTGGTATGGTTAATGTT

G-to-A

TTAACATTTAACAATTT

read conversion

align to bisulfite converted genomes

forward strand C-to-T converted genome

forward strand G-to-A converted genome

(1) (2) (3) (4)

determine unique best alignment

read all four alignment outputs simultaneously to determine if the sequence can be mapped uniquely

B BS-read corresponds to converted original top strand

5’-TTGGCATGTTAACGT-3’  5’...ccggctgtttaatgtct...3’
bisulfite read  genomic sequence

z unmethylated C in CpG context
Z methylated C in CpG context
x unmethylated C in CHG context
X methylated C in CHG context
h unmethylated C in CHH context
H methylated C in CHH context

xz...Hz...........Zhz.
methylation call
M values and Beta values

Beta Value

\[ \text{Beta Value} = \frac{M}{M + U} \times 100 \]

M Value

\[ \text{M Value} = \log_2\left(\frac{M+a}{U+a}\right) \]

In the above example, the methylation level of the locus is 40%
Methylation Table

<table>
<thead>
<tr>
<th>CpG 1</th>
<th>Sample 1</th>
<th>Sample 2</th>
<th>...</th>
<th>Sample N</th>
</tr>
</thead>
<tbody>
<tr>
<td>K_{11}</td>
<td></td>
<td></td>
<td></td>
<td>K_{1N}</td>
</tr>
<tr>
<td>K_{21}</td>
<td></td>
<td></td>
<td></td>
<td>K_{2N}</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td>...</td>
</tr>
<tr>
<td>K_{p1}</td>
<td></td>
<td></td>
<td></td>
<td>K_{pN}</td>
</tr>
</tbody>
</table>

- $K_{ij}$ is between 0 and 1 for beta values, or normally distributed for M values.
- $p \gg N$ small number of replicates for bulk samples.
- Downstream dimension reduction is similar to other analysis such as in RNA-seq
PacBio Long Read Sequencing for Methylation

• Genome-wide detection and phasing of genetic and epigenetic variants from a single library prep
• No bisulfite treatment needed!
• Need high coverage (250X for 5mC) and high amount starting material!
PacBio Long Read Sequencing for Methylation

SMRTbell® library ➔ PacBio® long-read systems ➔ 5-base HiFi sequencing with A, C, G, T, +5mC

5mC encoded with standard BAM tags:
MM:ZC+m,4,12,16,4,16,19,44,10
ML:8:O,249;4,247;777,210,228,245,244

The PacBio long-read systems directly output long, highly accurate HiFi reads with annotation of 5mC methylation at all CpG sites. No special library preparation like bisulfite treatment is required.
### Pacbio SMRT Link v11.0

**Run QC**

<table>
<thead>
<tr>
<th>Instrument Name</th>
<th>Instrument Status</th>
<th>SMRT Cell Status</th>
<th>Run Completion</th>
<th>Sequencing ZMWs</th>
</tr>
</thead>
<tbody>
<tr>
<td>64002</td>
<td>Ready</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>64009</td>
<td>Ready</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>64012</td>
<td>Ready</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>64263e</td>
<td>Running</td>
<td></td>
<td>in 4 days and 14 hours</td>
<td></td>
</tr>
<tr>
<td>64303e</td>
<td>Running</td>
<td></td>
<td>in 12.2 hours</td>
<td></td>
</tr>
</tbody>
</table>
Pacbio Methylation Tools Under Development

Kinetic tags

CallCpG.py <bamIn> <bamOut>
(available to early access collaborators)

Pr(methylated) at CpG
Useful Links

• Minfi manual
  

• Workshop data and notebooks:
  
  /dfs8/commondata/workshop/methylation/

• Bismark manual
  

• Illumina GenomeStudio
  
  https://support.illumina.com/array/array_software/genomestudio/downloads.html
Figure 2. Principle of detecting modified DNA bases during SMRT sequencing. The presence of the modified base in the DNA template (top), shown here for 6-mA, results in a delayed incorporation of the corresponding T nucleotide, i.e., longer interpulse duration (IPD), compared to a control DNA template lacking the modification (bottom).
General Workflow