Bisulfite sequencing

Mikhail Dozmorov

Spring 2018
Bisulfite sequencing in a nutshell

First treat the DNA with bisulfite. As a result

- Unmethylated C will be turned into T.
- Methylated C will be protected and still be C.
- No change for other bases.

Amplify, then sequence the treated DNA segments.

- The mismatches between C-T measures the methylation strength.

Raw data: sequence reads, but not exactly from the reference genome.
**Bisulfite sequencing in a nutshell**

OT, original top strand; CTOT, strand complementary to the original top strand; OB, original bottom strand; and CTOB, strand complementary to the original bottom strand.

Bisulfite limitations

- Bisulfite sequencing experiments do not distinguish an additional type of cytosine methylation, the 5-hydroxy-methylcytosine (hmC), which is a critical intermediary in active de-methylation pathways.
- Specific experimental methods for the identification of this mark at the base-resolution were developed.
- MLML, http://smithlabresearch.org/software/mlml/, is a popular computational method for a first analysis of these data.


Workflow for analyzing BS-data

Processing of bisulfite-sequencing data

- Quality control and pre-processing
- Bisulfite sequence alignment
- Quantification of absolute DNA methylation

Data visualization and statistical analysis

- Visual inspection in a genome browser of selected regions
- Visualization of global distribution of methylation values
- Clustering of samples based on similarity

Downstream analysis

- Identification of Differentially Methylated Regions (DMRs)
- Global analysis of DMRs
Mapping of bisulfite-treated sequences to a reference genome constitutes a significant computational challenge due to the combination of:

- The reduced complexity of the DNA code
- Up to four DNA strands to be analysed
- The fact that each read can theoretically exist in all possible methylation states.
The reads from BS-seq cannot be directly aligned to the reference genome.
There are four different strands after bisulfite treatment and PCR.
T could be aligned to T or C.
The search space for alignment is bigger.

3 main strategies for processing WGBS reads

- Wild-card alignment
- Three-letter alignment
- Reference-free processing
**Example of bisulfite alignment**

**a  Setup of the example**

Genomic DNA sequence: \( \text{CGATGATGTCGCTGA} \)

DNA methylation level: 100% 50% 50% 0%

DNA fragmentation, selective conversion of unmethylated Cs into Ts, DNA sequencing

Bisulphite-sequencing reads: \( \text{ACGTA, ATGAA, ATGATG, TCGA, TCGA, TCGA, TTTG} \)
**Wild-card aligners**

- Replace Cs in the genomic DNA sequence by the wild-card letter Y, which matches both Cs and Ts in the read sequence.
- Or modify the alignment scoring matrix in such a way that mismatches between Cs in the genomic DNA sequence and Ts in the read sequence are not penalized.
- Software: BSMAP, GSNAP, Last/bisulfighter, Pash, RMAP, RRBSMAP and segemehl
b Wild-card alignment

Reference sequence: YYGATGATGTYGYTGAYGYAYGA

Read alignment:
- TCGA
- TCGA
- TCGT
- TTGT
- ATGT

DNA methylation level:
- 100%
- 50%
- 100%
- 0%
Three-base aligner

- Simplify bisulfite alignment by converting all Cs into Ts in the reads and for both strands of the genomic DNA sequence
- Solware: Bismark, BRAT, BS-Seeker and MethylCoder
c Three-letter alignment

Reference sequence: TTGATGATGTTGTTGATGTATGA

Read alignment:
- TTGA
- TTGA
- TTGA
- TTGA
- ATGT
- ATGT
- ATGT
- ATGA
- ATGA

DNA methylation level:
- N/A
- 50%
- N/A
- 0%
Strengths and weaknesses

- Three-letter aligners have lower coverage in highly methylated regions because they purge the remaining Cs from the bisulfite-sequencing reads and thereby decrease their sequence complexity and they become ambiguous.
- Wild-card aligners typically have higher genomic coverage but at the cost of introducing some bias towards increased DNA methylation levels because the extra Cs in a methylated sequencing read can raise the sequence complexity.
- These problems are more prevalent in repetitive regions of the genome and are reduced with longer reads.
Bismark’s approach to bisulfite mapping and methylation calling.

- Reads from a BS-Seq experiment are converted into a C-to-T and a G-to-A version and are then aligned to equivalently converted versions of the reference genome.
- A unique best alignment is then determined from the four parallel alignment processes.

Bismark A tool to map bisulfite converted sequence reads and determine cytosine methylation states
https://www.bioinformatics.babraham.ac.uk/projects/bismark/
Here, the best alignment has no mismatches and comes from thread (1)
## Bismark

The methylation state of positions involving cytosines is determined by comparing the read sequence with the corresponding genomic sequence. Depending on the strand a read mapped against this can involve looking for C-to-T (as shown here) or G-to-A substitutions.

- **BS-read corresponds to converted original top strand**

```
5'–TTGGC\_ATGTTAAA\_CGTT–3'
5'...\_ccggc\_atgtttaa\_a\_c\_g\_t...3'
```

- **Bismark: A tool to map bisulfite converted sequence reads and determine cytosine methylation states**

https://www.bioinformatics.babraham.ac.uk/projects/bismark/

- **Methylation call**

```
xz.\_H....\_z.h.
```

- **Unmethylated** C in CpG context
- **Methylated** C in CpG context
- **Unmethylated** C in CHG context
- **Methylated** C in CHG context
- **Unmethylated** C in CHH context
- **Methylated** C in CHH context
BS-seq data analysis

Compared with ChIP-seq and RNA-seq, still in relatively early stage.

Questions include:

- **Single dataset analysis:**
  - Segment genome according to methylation status.

- **Comparison of multiple datasets:**
  - Differential methylation (DM) analysis.
Single BS-seq dataset analysis

Detecting the methylation loci/regions:

- Estimate “methylation density” (percentage of cells have methylation) at each C position, which is simply \#methyl/\#total at each CpG site, but:
  - Background error rates need to be considered.
  - Spatial correlation among nearby CpG sites can be utilized to improve estimation.

- Methylated regions (or states) can be determined by smoothing based method (e.g., moving average) using the estimated percentage as input.
Smoothing method

- Can directly smooth the percentages, but that doesn’t consider the uncertainty in percentage estimates.
- A better approach: BSmooth model (Hansen et.al. 2012 Genome Biology).
  - Assumes the true methylation level is a smooth curve of genomic coordinates.
  - The observed counts follow a binomial distribution.
  - Estimate smoothing function with local smoothing estimator

![Graph showing methylation levels and kernel weights](image-url)
BSmooth smoothing

Notations at position $j$:

- $N_j$, $M_j$: total/methylated reads
- $\pi_j$: underlying true methylation level
- $l_j$: location

Model:

- $M_j \sim \text{Bin}(N_j, \pi_j)$
- $\log(\pi_j/(1 - \pi_j)) = \beta_0 + \beta_1 l_j + \beta_2 l_j^2$

Fitting: weighted glm in each 2kb window, where the weights depend on the variances of estimated $\pi_j$
Mainly provide functions for smoothing and some visualization.

Implemented in parallel computing environment to speed up the calculation.

```r
M <- matrix(0:8, 3, 3) # Methylation evidence
Cov <- matrix(1:9, 3, 3) # Coverage
BS1 <- BSseq(chr = c("chr1", "chr2", "chr1"),
              pos = c(1,2,3), M = M, Cov = Cov,
              sampleNames = c("A","B", "C"))
BS1 <- BSmooth(BS1)
```
Differential methylation analysis

Comparison of methylation profiles under different biological conditions is of great interests.

- Results from such analysis are: differentially methylated loci (DML) or regions (DMR).

Strategy to detect DML:

- Hypothesis testing at each CpG site.

Strategy to detect DMR:

- Need to combine data from nearby CpG sites because of the spatial correlation.
DML detection based on 2x2 table

At each CpG site, summarize the counts from two samples into a 2x2 table:

<table>
<thead>
<tr>
<th>Sample/Methylation</th>
<th>Total</th>
<th>Methylated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample 1</td>
<td>40</td>
<td>2</td>
</tr>
<tr>
<td>Sample 2</td>
<td>25</td>
<td>19</td>
</tr>
</tbody>
</table>

Chi-square or Fisher’s exact test can be applied. bsseq has function fisherTests for this: fisherTests(BSobj, group1, group2)
Wald-test based

- Uses data with replicates
- The key is to estimate within-group variances
- BSmooth approach (for two-group comparison):
  - Denote the group assignment for \( i^{th} \)
  - Number of replicates in two groups are \( n_1 \) and \( n_2 \)
  - Frame the estimated values into a two-group testing framework:
    \[
    \pi_{ij} = \alpha(I_j) + \beta(I_j)X_i + \epsilon_{i,j}, \quad \epsilon_{i,j} \sim N(0, \sigma_j^2)
    \]
  - Use SAM-like method to estimate \( \sigma_j^2 \), then do Wald test

Multiple loci can be differentially methylated - need one p-value

Fisher’s method for combining p-values given $K$ independent tests:

$$T = -2 \sum_{k=1}^{K} \ln(p_k)$$

$T \sim \chi^2_{2K}$

Other methods: Stouffer-Liptak

Similar to RNA-seq DE analysis, the BS-seq data can be modeled as beta-binomial distribution.

For $i^{th}$ CpG site, $j^{th}$ group and $k^{th}$ replicate, $X_{ijk}$ is the number of reads that show methylation, $N_{ijk}$ is the total number of reads that cover this position and $p_{ijk}$ is the underlying “true” methylation proportion.

$$X_{ijk} | p_{ijk}, N_{ijk} \sim Binomial(N_{ijk}, p_{ijk})$$

Since the true methylation proportions among replicates can be anywhere between 0 and 1, we assume that they follow a beta distribution.

$$p_{ijk} \sim Beta(\mu_{ijk}, \phi_{ij})$$
DSS: Shrinkage-based method

- Beta distribution is parameterized by mean and dispersion, and impose a log-normal prior on dispersion $\phi_{ij} \sim \text{lognormal}(m_{0j}, r_{0j}^2)$, $m_{0j}$ mean and $r_{0j}^2$ can be estimated from the data.
- Wald test procedure can be derived. For two-group comparison:

$$t_i = \frac{\hat{\mu}_{i1} - \hat{\mu}_{i2}}{\sqrt{\hat{\text{var}}_{i1} - \hat{\text{var}}_{i2}}}$$

- where $\hat{\mu}_{ij}$ are mean methylation levels and $\hat{\text{var}}_{ij}, (j = 1, 2)$ is the estimated variance for group 1 or 2.

Simulation results

- The Wald test with shrunk dispersion performs favorably compared with other methods (2 replicates, 5 replicates)
Things to consider in DMR calling

Coverage depth:

- Should one filter out sites with shallower coverage?

Biological replicates:

- CpG-specific biological variances.
- Small sample estimate of the variance.

Spatial correlation of methylation levels among nearby CpG sites.

- Is smoothing appropriate?
- What if data has low spatial correlation, like in 5hmC.
Differential Methylation analysis using bsseq

- First create BSseq objects
- Use BSmooth function to smooth.
- fisherTests performs Fisher’s exact test, if there’s no replicate.
- BSmooth.tstat performs t-test with replicates.
- dmrFinder calls DMRs based on BSmooth.tstat results.

BSobj = BSmooth(BSobj)
dmlTest=fisherTests(BSobj, group1=c("C1", "C2", "C3"),
    group2=c("N1", "N2", "N3"))
dmr <- dmrFinder(dmlTest)
Differential Methylation analysis using DSS

- Input data has the same format as bsseq.
- DMLtest performs Wald test at each CpG.
- callDML/callDMR calls DML or DMR.
- More options in DML/DMR calling.

```r
dmlTest <- DMLtest(BSobj, group1=c("C1", "C2", "C3"),
                   group2=c("N1", "N2", "N3"),
                   smoothing=TRUE, smoothing.span=500)

dmrs <- callDMR(dmlTest)
```
Conclusion on BS-seq analyses

- Careful in alignments.
- Data modeling is different from ChIP/RNA-seq: Poisson/NB vs. Binomial models.
- DMR calling needs to consider spatial correlation, coverage and biological variances.
- Single read analysis could be very useful.
- A lot of room for method development.
(m)RRBS: (multiplexed) Reduced Representation Bisulfite Sequencing

- Utilizes cutting pattern of MspI enzyme (C^CGG) to systematically digest CpG-poor DNA

- Covers the majority of CpG islands and promoters, and a reasonable number of exons, shores and enhancers

- Advantages:
  - Only need 50-200ng DNA
  - Can be from any species
  - Cost and time
methylKit R package

- Technology: (RB)BS-seq and derivatives, including 5hmC
- Input: Bismark-aligned SAM files, or text-summarized % methylation
- Functionality: QC, clustering, differential methylation of sites/regions, visualization

https://github.com/al2na/methylKit

# Methods to detect differentially methylated loci or regions

<table>
<thead>
<tr>
<th>Method</th>
<th>Citation</th>
<th>Designed for</th>
<th>Determines regions or uses predefined</th>
<th>Accounts for covariates</th>
<th>Statistical element used</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minfi</td>
<td>Aryee et al., 2014</td>
<td>450k</td>
<td>Determines</td>
<td>Yes</td>
<td>Bump hunting</td>
</tr>
<tr>
<td>IMA</td>
<td>Wang et al., 2012</td>
<td>450k</td>
<td>Predefined</td>
<td>No</td>
<td>Wilcoxon</td>
</tr>
<tr>
<td>COHCAP</td>
<td>Warden et al., 2013</td>
<td>450k or BS-seq</td>
<td>Predefined</td>
<td>Yes</td>
<td>FET, t-test, ANOVA</td>
</tr>
<tr>
<td>BSmooth</td>
<td>Hansen et al., 2012a</td>
<td>BS-seq</td>
<td>Determines</td>
<td>No</td>
<td>Bump hunting on small</td>
</tr>
<tr>
<td>DSS</td>
<td>Feng et al., 2014</td>
<td>BS-seq</td>
<td>Determines</td>
<td>No</td>
<td>Wald</td>
</tr>
<tr>
<td>MOABS</td>
<td>Sun et al., 2014</td>
<td>BS-seq</td>
<td>Determines</td>
<td>No</td>
<td>“Credible methylation”</td>
</tr>
<tr>
<td>BiSeq</td>
<td>Hebestreit et al., 2013</td>
<td>BS-seq</td>
<td>Determines</td>
<td>Yes</td>
<td>Wald</td>
</tr>
<tr>
<td>DMAP</td>
<td>Stockwell et al., 2014</td>
<td>BS-seq</td>
<td>Predefined</td>
<td>Yes</td>
<td>ANOVA, $\chi^2$, FET</td>
</tr>
<tr>
<td>methylKit</td>
<td>Akalin et al., 2012</td>
<td>BS-seq</td>
<td>Predefined</td>
<td>Yes</td>
<td>Logistic regression</td>
</tr>
<tr>
<td>RADMeth</td>
<td>Dolzhenko and Smith, 2014</td>
<td>BS-seq</td>
<td>Determines</td>
<td>Yes</td>
<td>Likelihood-ratio</td>
</tr>
<tr>
<td>methylSig</td>
<td>Park et al., 2014</td>
<td>BS-seq</td>
<td>Predefined</td>
<td>No</td>
<td>Likelihood-ratio</td>
</tr>
<tr>
<td>Bumphunter</td>
<td>Jaffe et al., 2012</td>
<td>General</td>
<td>Determines</td>
<td>Yes</td>
<td>Permutation, smoothing</td>
</tr>
<tr>
<td>ABCD-DNA</td>
<td>Robinson et al., 2012</td>
<td>MeDIP-seq</td>
<td>Predefined</td>
<td>Yes</td>
<td>Likelihood ratio</td>
</tr>
<tr>
<td>DiffBind</td>
<td>Ross-Innes et al., 2012</td>
<td>MeDIP-seq</td>
<td>Predefined</td>
<td>Yes</td>
<td>Likelihood ratio</td>
</tr>
<tr>
<td>M&amp;M</td>
<td>Zhang et al., 2013</td>
<td>MeDIP-seq+MRE-seq</td>
<td>Determines</td>
<td>No</td>
<td>(Similar to) FET</td>
</tr>
</tbody>
</table>
BS-seq data SNP/methylation caller

- Bis-SNP
- MethylExtract
- BS-SNPer
- Etc.