

Bisulfite sequencing

Mikhail Dozmorov

Spring 2018

Bisulfite sequencing in a nutshell

First treat the DNA with bisulfite. As a result

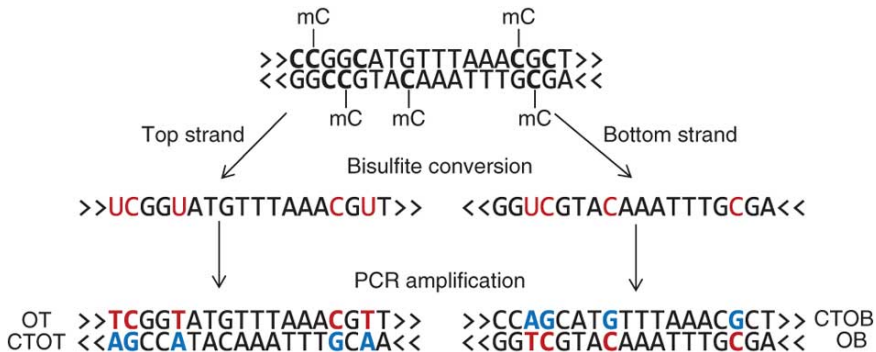
- Unmethylated C will be turned into T.
- Methylated C will be protected and still be C.
- No change for other bases.

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.

Krueger, Felix, Benjamin Kreck, Andre Franke, and Simon R. Andrews. "DNA Methylome Analysis Using Short Bisulfite Sequencing Data." *Nature Methods* 9, no. 2 (January 30, 2012): 145–51. <https://doi.org/10.1038/nmeth.1828>.

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

Guo, Junjie U., Yijing Su, Chun Zhong, Guo-li Ming, and Hongjun Song. "Hydroxylation of 5-Methylcytosine by TET1 Promotes Active DNA Demethylation in the Adult Brain." *Cell* 145, no. 3 (April 29, 2011): 423–34.
<https://doi.org/10.1016/j.cell.2011.03.022>.

Qu, Jiangnan, Meng Zhou, Qiang Song, Elizabeth E. Hong, and Andrew D. Smith. "MLML: Consistent Simultaneous Estimates of DNA Methylation and Hydroxymethylation." *Bioinformatics (Oxford, England)* 29, no. 20 (October 15, 2013): 2645–46.
<https://doi.org/10.1093/bioinformatics/btt459>.

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

Bisulfite sequencing mapping

Mapping of bisulfite-treated sequences to a reference genome constitutes a significant computational challenge due to the combination of:

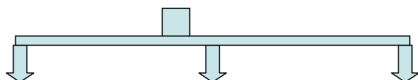
- The reduced complexity of the DNA code
- Up to four DNA strands to be analysed
- The fact that each read can theoretically exist in all possible methylation states.

Alignment of BS-seq

- 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.

Bisulfite Read

>>**ATTTCG**>>



Reference

>>AT**ACTTCG**ATGAT**CTCG**CAAG**ACTCCG**GC>>
>>AT**ACTTCG**ATGAT**CTCG**CAAG**ACTCCG**GC>>
>>AT**ACTTCG**ATGAT**CTCG**CAAG**ACTCCG**GC>>

Krueger, Felix, Benjamin Kreck, Andre Franke, and Simon R Andrews. "DNA Methylation Analysis Using Short Bisulfite Sequencing Data." *Nature Methods* 9, no. 2 (January 30, 2012): 145–51. <https://doi.org/10.1038/nmeth.1828>.

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 **C****CG****A**TGATGT**CG**CTGA**CG**CA**CGA**
DNA methylation level 100% 50% 50% 0%

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



Bisulphite-sequencing reads **A****CG****T**, **A****T****G****A**, **A****T****G****A**, **A****T****G****T**,
T**CG****A**, **T****CG****A**, **T****CG****T**, **T****T****G****T**

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

Three-base aligner

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

Three-base aligner

c Three-letter alignment

Reference sequence **T**TGATGATG**T**TG**T**TGATG**T**ATGA

Read alignment

TtGA TtGA

TtGA TtGA

TtGT TtGT

AtGT 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

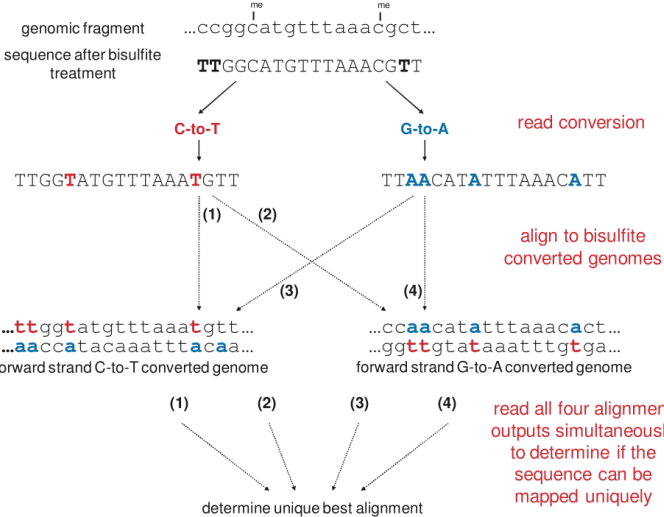
Bismark's approach to bisulfite mapping and methylation calling.

- 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/>

Bismark

A



Here, the best alignment has no mismatches and comes from thread (1)

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.

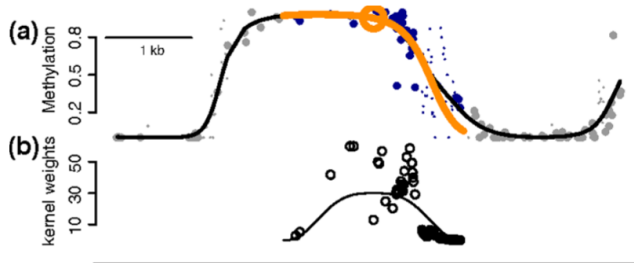
Single BS-seq dataset analysis

Detecting the methylation loci/regions:

- Estimate “methylation density” (percentage of cells have methylation) at each C position, which is simply $\frac{\#methyl}{\#total}$ 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



BSmooth smoothing

Notations at position j :

- N_j, M_j : total/methylated reads
- π_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 π_j

Bsmooth Bioconductor package: bsseq

- Mainly provide functions for smoothing and some visualization.
- Implemented in parallel computing environment to speed up the calculation.

```
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:

Sample/Methylation	Total	Methylated
Sample 1	40	2
Sample 2	25	19

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 σ_j^2 , then do Wald test

Hansen, Kasper D, Benjamin Langmead, and Rafael A Irizarry. "BSmooth: From Whole Genome Bisulfite Sequencing Reads to Differentially Methylated Regions." *Genome Biology* 13, no. 10 (2012): R83. <https://doi.org/10.1186/gb-2012-13-10-r83>.

Differential methylation in regions

- Multiple loci can be differentially methylated - need one p-value
- Fisher's method for combining p-values given K independent tests:

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

- $T \sim \chi_{2K}^2$
- Other methods: Stouffer-Liptak

Zaykin, D. V. "Optimally Weighted Z-Test Is a Powerful Method for Combining Probabilities in Meta-Analysis." *Journal of Evolutionary Biology* 24, no. 8 (August 2011): 1836–41. <https://doi.org/10.1111/j.1420-9101.2011.02297.x>.

DSS: Dispersion shrinkage for sequencing data

- 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

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

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

$$p_{ijk} \sim \text{Beta}(\mu_{ijk}, \phi_{ij})$$

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:

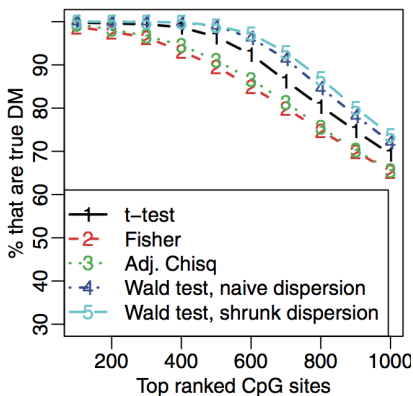
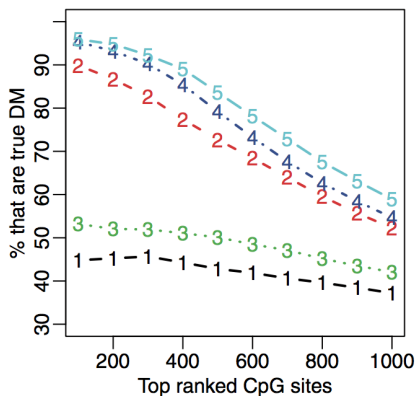
$$t_j = \frac{\hat{\mu}_{i1} - \hat{\mu}_{i2}}{\sqrt{\hat{var}_{i1} - \hat{var}_{i2}}}$$

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

Feng, Hao, Karen N. Conneely, and Hao Wu. "A Bayesian Hierarchical Model to Detect Differentially Methylated Loci from Single Nucleotide Resolution Sequencing Data." *Nucleic Acids Research* 42, no. 8 (April 2014): e69–e69. <https://doi.org/10.1093/nar/gku154>. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4005660/>

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.

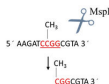
```
dmlTest <- DMLtest(BSobj, group1=c("C1", "C2", "C3"),  
                  group2=c("N1", "N2", "N3"),  
                  smoothing=TRUE, smoothing.span=500)  
dmrs <- callDMR(dmlTest)
```


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

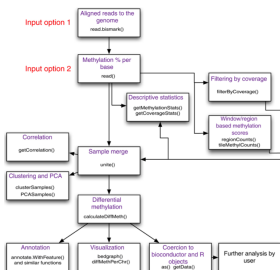
- Utilizes cutting pattern of MspI enzyme ($C^{\wedge}CGG$) to systematically digest CpG-poor DNA



- 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>

Akalin, Altuna, Matthias Kormaksson, Sheng Li, Francine E. Garrett-Bakelman, Maria E. Figueroa, Ari Melnick, and Christopher E. Mason. "MethylKit: A Comprehensive R Package for the Analysis of Genome-Wide DNA Methylation Profiles." *Genome Biology* 13, no. 10 (October 3, 2012): R87. <https://doi.org/10.1186/gb-2012-13-10-r87>.

Methods to detect differentially methylated loci or regions

Method	Citation	Designed for	Determines regions or uses predefined	Accounts for covariates	Statistical element used
Minfi	Aryee et al., 2014	450k	Determines	Yes	Bump hunting
IMA	Wang et al., 2012	450k	Predefined	No	Wilcoxon
COHCAP	Warden et al., 2013	450k or BS-seq	Predefined	Yes	FET, t-test, ANOVA
BSmooth	Hansen et al., 2012a	BS-seq	Determines	No	Bump hunting on sn
DSS	Feng et al., 2014	BS-seq	Determines	No	Wald
MOABS	Sun et al., 2014	BS-seq	Determines	No	"Credible methylation"
BiSeq	Hebestreit et al., 2013	BS-seq	Determines	Yes	Wald
DMAP	Stockwell et al., 2014	BS-seq	Predefined	Yes	ANOVA, χ^2 , FET
methylKit	Akalin et al., 2012	BS-seq	Predefined	Yes	Logistic regression
RADMeth	Dolzhenko and Smith, 2014	BS-seq	Determines	Yes	Likelihood-ratio
methylSig	Park et al., 2014	BS-seq	Predefined	No	Likelihood-ratio
Bumphunter	Jaffe et al., 2012	General	Determines	Yes	Permutation, smooth
ABCD-DNA	Robinson et al., 2012	MeDIP-seq	Predefined	Yes	Likelihood ratio
DiffBind	Ross-Innes et al., 2012	MeDIP-seq	Predefined	Yes	Likelihood ratio
M&M	Zhang et al., 2013	MeDIP-seq+MRE-seq	Determines	No	(Similar to) FET

BS-seq data SNP/methylation caller

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