Sequencing-based methylation technologies and analysis

Mikhail Dozmorov

2021-04-12

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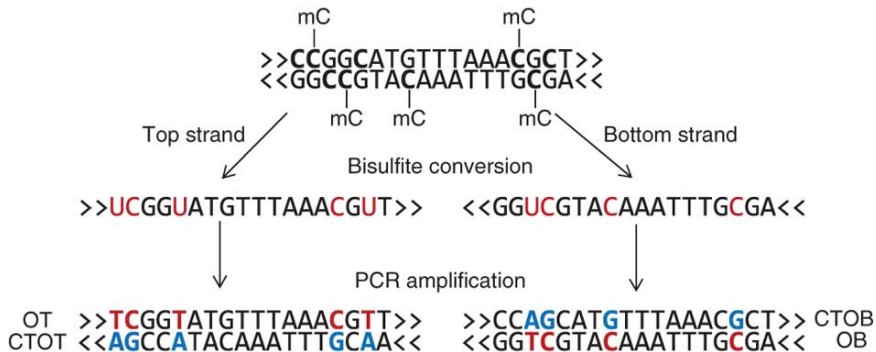
OT, original top strand; **CTOT**, strand complementary to the original top strand; **OB**, original bottom strand; and **CTOB**, strand complementary to the original bottom strand.

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 demethylation 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 the 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, Jianghan, 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 analyzed
- 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.
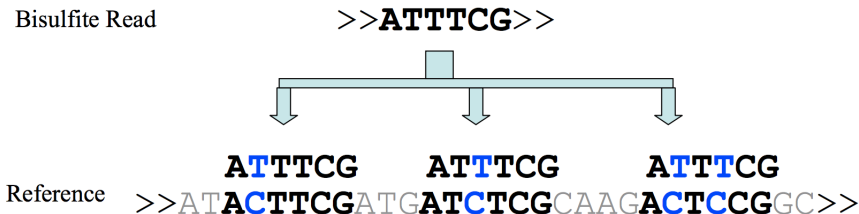
Bisulfite Read >>**ATTTCG**>>

Reference

ATTTCG **ATTTCG** **ATTTCG**
>>ATACTTCGATGATCTCGCAAGACTCCGGC>>

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.

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**ATGATGT**CG**CTGA**CG**CA**CG**A

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

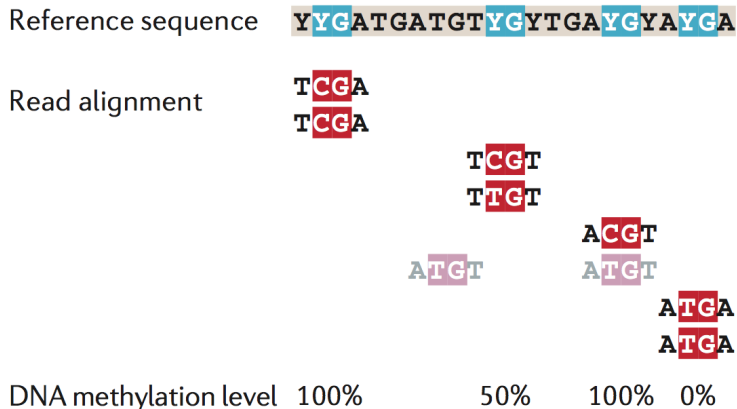DNA fragmentation, selective conversion of unmethylated Cs into Ts, DNA sequencing

Bisulphite-sequencing reads: A**CG**T, AT**G**A, AT**C**A, AT**G**T,
T**CG**A, T**CG**A, T**CG**T, TT**G**T

Wild-card aligners

- Replace Cs in the genomic DNA sequence with 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.
- Solware: BSMAP, GSNAP, Last/bisulfighter, Pash, RMAP, RRBSMAP and segemehl

Wild-card alignment

b Wild-card alignment

Reference sequence YYGATGATGTYGYTGAYGYAYGA

Read alignment

TCGA
TCGA

TCGT
TTGT

ACGT
ATGT

ATGT

ATGA
ATGA

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

Three-base aligner

c Three-letter alignment

Reference sequence TTGATGATGTTGTTGATGTATGA

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, thereby decreasing their sequence complexity and becoming ambiguous.
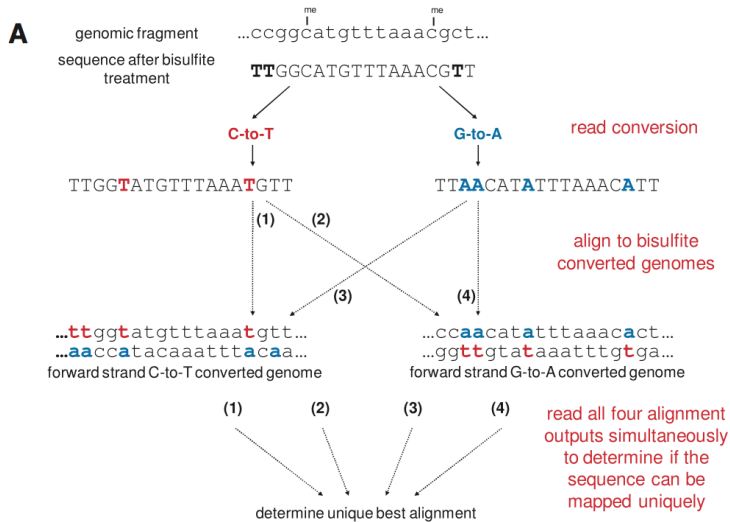
- Wild-card aligners typically have higher genomic coverage but at the cost of introducing some bias towards increased DNA methyla1on levels because the extra Cs in a methylated sequencing read can raise the sequence complexity.
- These problems are more prevalent in repetitive regions of the genome and are reduced with longer reads.

Bismark

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/

A genomic fragment ...ccggcatgtttaaacgct...

sequence after bisulfite treatment **TT**GGCATGTTTAAACG**T**T

C-to-T   G-to-A   read conversion

TTGG**T**ATGTTTAAA**T**GTT   TT**AA**CAT**A**TTTAAAC**A**TT

align to bisulfite converted genomes

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

...**tt**gg**t**atgtttaaa**t**gtt...   ...cc**aa**cat**a**tttaaac**a**ct...
...**aa**cc**a**tacaaattt**a**c**a**a...   ...gg**tt**gta**t**aaatttg**t**ga...
forward strand C-to-T converted genome   forward strand G-to-A converted genome

(1)   (2)   (3)   (4)   read all four alignment
outputs simultaneously
to determine if the
sequence can be
mapped uniquely

determine unique best alignment

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

BS-seq data analysis

Compared with gene expression technologies, still in a relatively early stage.

Questions include:

- Single dataset analysis:
  - Segment genome according to methylation status.
- Comparison of multiple datasets:
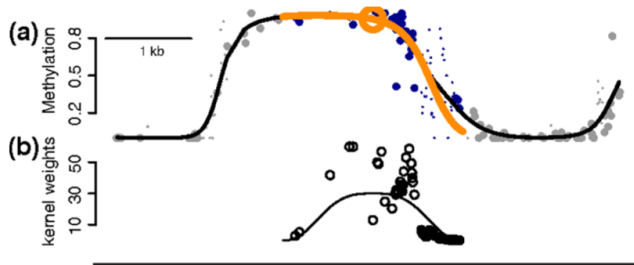 - Differential methylation (DM) analysis.

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 a 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

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:
      $\pi_{ij} = \alpha(l_j) + \beta(l_j)X_i + \epsilon_{i,j}, \ \epsilon_{i,j} \sim N(0, \sigma_j^2)$
    - 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:

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

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

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.

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 smoothed
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, smoot
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

Robinson, Mark D., Abdullah Kahraman, Charity W. Law, Helen Lindsay, Malgorzata Nowicka, Lukas M. Weber, and Xiaobei

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.

Section 1

Variants of methylation technologies

(m)RRBS: (multiplexed) Reduced Representation Bisulfite Sequencing

- Utilizes cutting pattern of MspI enzyme (C^CGG) to systematically digest CpG-poor DNA
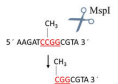
- Covers the majority of CpG islands and promoters, and a reasonable number of exons, shores and enhancers
- Advantages:
  - Only need 50-200ng DNA
  - Can be from any species
  - Cost and time

Methylation and Hydroxymethylation

- The two major epigenetic modifications of cytosines, 5-methylcytosine (5-mC) and 5-hydroxymethylcytosine (5-hmC), coexist with each other in a range of mammalian cell populations.
- 5-hmC was discovered in 2009, abundant in embryonic stem cells and neurons.
- Sequence- and strand-specific mark
- Near but not on transcription factor binding sites
- 5-hmC is reciprocal to 5-mC - high level of one suggests a low level of the other
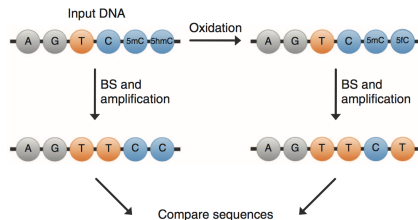
Hydroxymethylation

- 5-hydroxymethylcytosine (5-hmC) is an oxidation product of the extensively studied 5-methylcytosine (5-mC) modification
- It has been observed at substantial levels in both somatic and embryonic stem cells (Kriaucionis and Heintz, 2009; Tahiliani et al., 2009).
- 5-hmC is involved in epigenetic regulation

Technology

- **Bisulfite sequencing (BS-seq)** - treatment with sodium bisulfite converts unmethylated cytosines to uracils but does not distinguish between 5-mC and 5-hmC (Huang et al., 2010), and consequently, the yield of methylation from BS-seq is the sum of 5-mC and 5-hmC levels
- Two recently developed techniques, **oxidative bisulfite sequencing (oxBS-seq)** (Booth et al., 2012) and **Tet-Assisted Bisulfite sequencing (TAB-seq)** (Yu et al., 2012), provide high-throughput single-base resolution measurements of 5-mC and 5-hmC, respectively.
- Combining two technologies can be used to estimate joint levels of 5-mc and 5-hmc

oxBS-seq

- Oxidative bisulfite sequencing
- Chemical oxidation of 5hmC to 5-formylcytosine (5fC) enables bisulfite conversion of 5fC to uracil


Base	Sequence	BS Sequence	oxBS Sequence
C	C	T	T
5mC	C	C	C
5hmC	C	C	T

Booth, Michael J., Miguel R. Branco, Gabriella Ficz, David Oxley, Felix Krueger, Wolf Reik, and Shankar Balasubramanian. "Quantitative Sequencing of 5-Methylcytosine and 5-Hydroxymethylcytosine at Single-Base Resolution." Science (New York, N.Y.) 336, no. 6083 (May 18, 2012): 934–37. https://doi.org/10.1126/science.1220671.

TAB-seq

- Tet-assisted bisulfite sequencing
- TET proteins oxidize 5mc fo 5hmc to 5caC - reads as uracil after bisulfite treatment
- 5hmC glycosylation protects 5hmC from TET oxidation - reads as C after bisulfite treatment
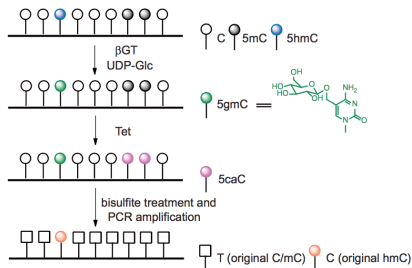
Yu, Miao, Gary C. Hon, Keith E. Szulwach, Chun-Xiao Song, Liang Zhang, Audrey Kim, Xuekun Li, et al. "Base-Resolution Analysis of 5-Hydroxymethylcytosine in the Mammalian Genome." Cell 149, no. 6 (June 2012): 1368–80. https://doi.org/10.1016/j.cell.2012.04.027.

Analysis

- When combining BS-seq with TAB-seq or oxBS-seq, the 5-mC level at a given CpG site can be estimated by subtracting the 5-hmC level (TAB/oxBS-seq) from the combined 5-mC:5-hmC level (BS-seq).
- Maximum likelihood methylation levels (MLML) for simultaneous estimation of 5-mC and 5-hmC, combining data from any two of BS-seq, TAB-seq or oxBS-seq, or all three when available.

http://smithlabresearch.org/software/mlml/

Qu, Jianghan, 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.