CS-E5875 High-Throughput Bioinformatics
DNA methylation analysis

Harri Lähdesmäki

Department of Computer Science
Aalto University

November 13, 2020

Contents

- DNA methylation
- Bisulfite sequencing (BS-seq) protocol
- Alignment and quantification of BS-seq data
- Statistical analysis of BS-seq data

DNA methylation

- Epigenetic changes are reversible modifications on DNA, or “on top of DNA”, which do not change the DNA sequence itself
- DNA methylation is an epigenetic modification where methyl group is added to the 5 position of a cytosine in DNA
- Methyl group is added enzymatically by DNA methyl transferases (DNMT)
- By far the most extensively studied epigenetic modification on DNA

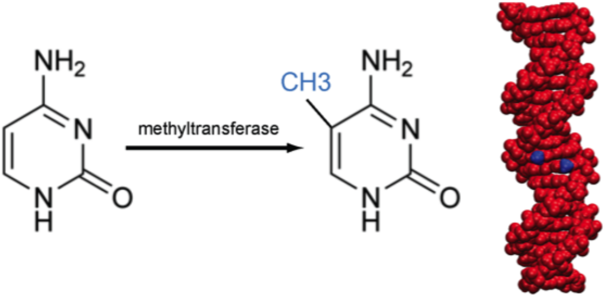


Figure from http://www.ks.uiuc.edu/Research/methylation/

DNA methylation

- In mammaling genomes, DNA methylation primarily occurs in the context of CpG dinucleotides
- Non-CpG methylation found e.g. in stem cells and brain
- CpGs occur with a smaller frequency than expected
  - Human genome GC content is 42%
  - CpGs are expected to occur 4.41% of the time
  - The frequency of CpG dinucleotides is 1%
  - Methylated CpGs are prone to spontaneous deamination to thymines

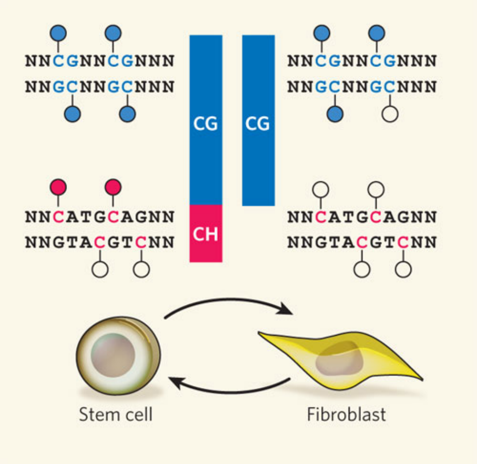


Figure from (Schubeler, 2009)

DNA methylation

- Two general classes of enzymatic methylation activities
  - De novo methylation
  - Maintenance methylation

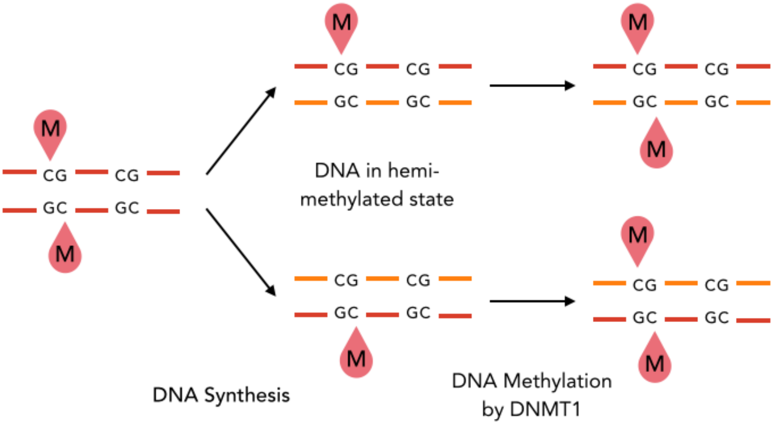


Figure from http://2014.igem.org/Team:Heidelberg/Project/PCR_2.0

DNA methylation in gene regulation and various traits

- CpG islands (C+G dense $\gtrsim$500 long regions) are present in the 5' regulatory regions of many genes
- Hypermethylation (=overmethylation) of CpG islands near gene promoters contributes to transcriptional silencing by
  - Affecting binding of transcription factors (DNA binding protein that regulate gene transcription)
  - Binding proteins with methyl-CpG-binding domains (MBDs), and recruiting e.g. histone deacetylases and other chromatin remodellers

DNA methylation in gene regulation and various traits

- CpG islands (C+G dense $\gtrsim$500 long regions) are present in the 5' regulatory regions of many genes
- Hypermethylation (=overmethylation) of CpG islands near gene promoters contributes to transcriptional silencing by
  - Affecting binding of transcription factors (DNA binding protein that regulate gene transcription)
  - Binding proteins with methyl-CpG-binding domains (MBDs), and recruiting e.g. histone deacetylases and other chromatin remodellers
- DNA methylation differences are associated with many diseases
- DNA methylation is also known to associate with e.g. age of an individual and smoking

DNA methylation

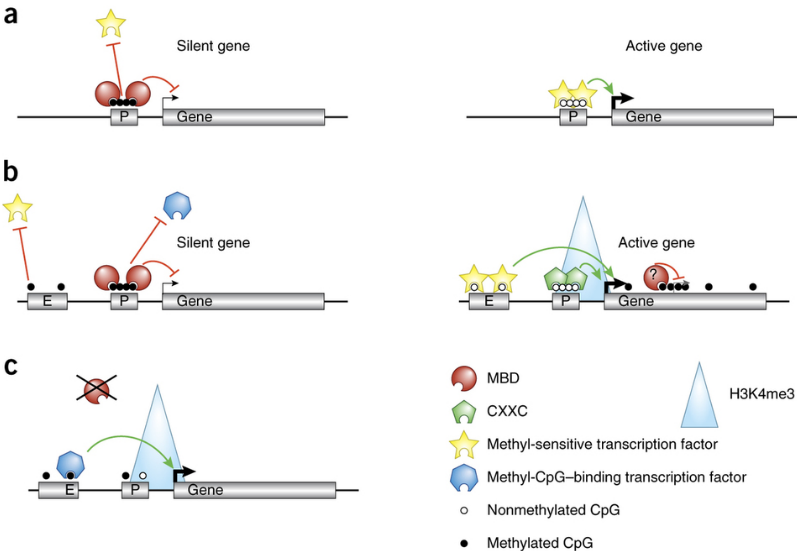
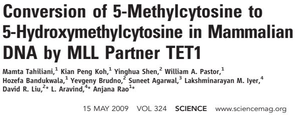


Figure from (Spruijt & Vermeulen, 2014)

DNA demethylation

- Until recently, it was believed that methylated DNA can be unmethylated only by dilution during cell differentiation/DNA replication
- Recently, TET family proteins were shown to be dioxygenases that converted 5mC to 5hmC, 5fC and 5caC, which can be further converted back to unmethylated C
- TETs thus contribute to active demethylation, but 5hmC, 5fC and 5caC can also have multiple functions

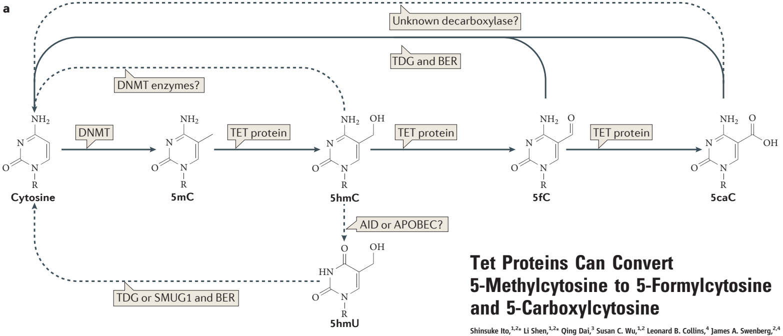
DNA demethylation



Conversion of 5-Methylcytosine to 5-Hydroxymethylcytosine in Mammalian DNA by MLL Partner TET1

Mamta Tahiliani,[1] Kian Peng Koh,[1] Yinghua Shen,[2] William A. Pastor,[1] Hozefa Bandukwala,[1] Yevgeny Brudno,[2] Suneet Agarwal,[3] Lakshminarayan M. Iyer,[4] David R. Liu,[2*] L. Aravind,[4*] Anjana Rao[1*]

15 MAY 2009 VOL 324 **SCIENCE** www.sciencemag.org



Tet Proteins Can Convert 5-Methylcytosine to 5-Formylcytosine and 5-Carboxylcytosine

Shinsuke Ito,[1,2*] Li Shen,[1,2*] Qing Dai,[3] Susan C. Wu,[1,2] Leonard B. Collins,[4] James A. Swenberg,[2,4] Chuan He,[3] Yi Zhang[1,2†]

2 SEPTEMBER 2011 VOL 333 **SCIENCE** www.sciencemag.org

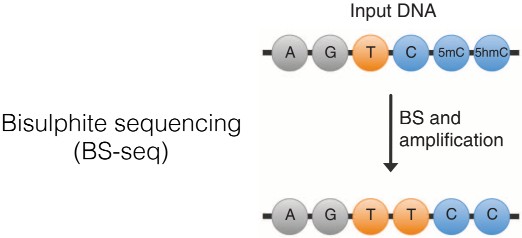
BER := base excision repair
TDG := thymine DNA glycosylase
AID := activation-induced deaminase
APOBEC := apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like

Contents

- DNA methylation
- Bisulfite sequencing (BS-seq) protocol
- Alignment and quantification of BS-seq data
- Statistical analysis of BS-seq data

Bisulfite sequencing (BS-seq) protocol

- Bisulfite treatment of genomic DNA converts unmethylated cytosines to urasils which are read as thymine during sequencing
- Methylated (and hydroxymethylated) cytosines are resistant to the conversion and are read as cytosine



	Cytosine modification		
	C	5mC	5hmC
BS-seq	T	C	C

Figure from (Booth et al, 2012)

Bisulfite sequencing (BS-seq) protocol

- Bisulfite treatment of genomic DNA converts unmethylated cytosines to urasils which are read as thymine during sequencing
- Methylated (and hydroxymethylated) cytosines are resistant to the conversion and are read as cytosine

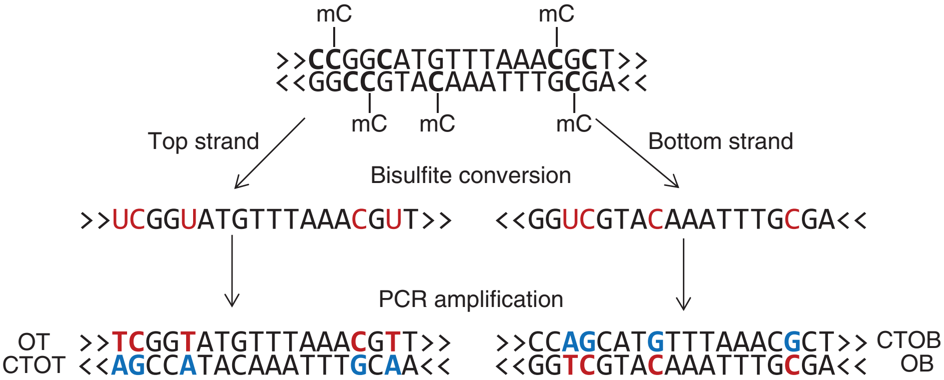


Figure from (Krueger et al, 2012)

Reduced representation BS-seq (RRBS-seq)

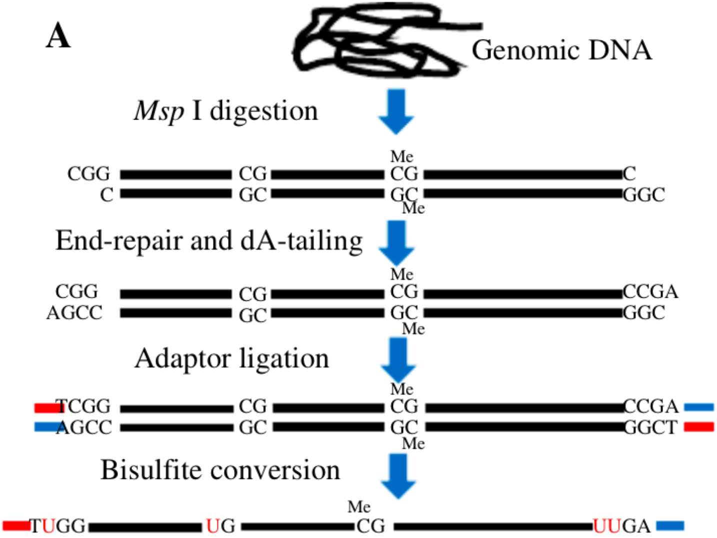
- BS-seq provides an accurate map of methylation state at single nucleotide resolution
- Whole genome analysis is expensive because only about 1% of the human genome contains CpGs

→ Experimental techniques to enrich for the areas of the genome that have a high CpG content

- Reduced representation BS-seq (RRBS-seq) uses restriction enzymes prior to bisulfite sequencing
  - MspI digests genomic DNA in a methylation-insensitive manner
  - MspI targets 5'CCGG3' sequences and cleaves the phosphodiester bonds upstream of CpG dinucleotide.

  → Each fragment will have a CpG at each end
- RRBS-seq will cover majority of promoters and GC rich regions

Reduced representation BS-seq (RRBS-seq)



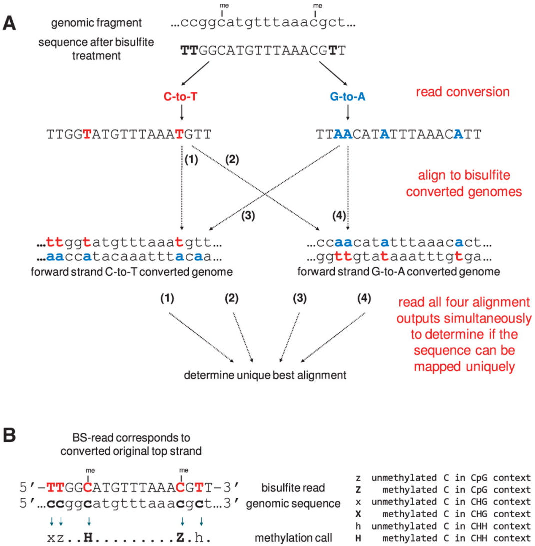
Contents

- DNA methylation
- Bisulfite sequencing (BS-seq) protocol
- Alignment and quantification of BS-seq data
- Statistical analysis of BS-seq data

Aligning BS-seq reads

- Bisulfite treatment introduces mutations into genomic DNA in a methylation dependent manner
  - Alignment of BS-seq reads is more challenging
  - Standard alignment methods cannot be used directly
- Bismark tool uses the following approach to map BS-seq reads
  - Reads from a BS-seq experiment are converted into a C-to-T version and a G-to-A version
  - The same conversion for the genome
  - Bowtie alignment in the genome that has reduced complexity
  - A unique best alignment is determined from four parallel alignment processes (see next page)

Bismark tool


A
genomic fragment
me
me
...ccggcatgtttaaacgct...
sequence after bisulfite treatment
TTGGCATGTTTAAACGTT
C-to-T
G-to-A
read conversion
TTGGTATGTTTAAATGTT
TTAACATATTTAAACATT
(1)
(2)
align to bisulfite converted genomes
(3)
(4)
...ttggtatgtttaaatgtt...
...aaccatacaaatttacaa...
forward strand C-to-T converted genome
...ccaacatatttaaacact...
...ggttgtataaatttgtga...
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
B
BS-read corresponds to converted original top strand
me
me
5'-TTGGCATGTTTAAACGTT-3'
bisulfite read
5'...ccggcatgtttaaacgct...3'
genomic sequence
xz..H.........Z.h.
methylation call

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

Figure from (Krueger & Andrews, 2011)

Quantifying BS-seq data

- Bismark outputs, among others, one line per read containing useful information
  - Mapping position, alignment strand, the bisulfite read sequence, its equivalent genomic sequence and a methylation call string
- Bismark automatically extracts the methylation information at individual cytosine positions
  - For different sequence contexts (CpG, CHG, CHH; where H can be either A, T or C)
  - Strand-specific or strands merged
- That is, for each cytosine Bismark outputs
  - n_i the number of reads covering the cytosine in sample i
  - m_i the number of methylated readouts (i.e., “C”) for the cytosine in sample i
- One way to quantify methylation proportion is

$$\hat{p}_i = \frac{m_i}{n_i} = \frac{\text{the number of C reads overlapping the cytosine}}{\text{the number of C or T reads overlapping the cytosine}}$$

Contents

- DNA methylation
- Bisulfite sequencing (BS-seq) protocol
- Alignment and quantification of BS-seq data
- Statistical analysis of BS-seq data

Beta-binomial model

- At the end, one is typically interested in testing a hypothesis, e.g. is there a statistically significant difference in methylation levels between group A and group B
- Some early methods applied e.g. the t-test on the estimated methylation fractions $\hat{p}_i$ (or their logit transformations)
- We will look at RadMeth tool (Dolzhenko and Smith, 2014)
- RadMeth uses the beta-binomial regression model, where beta-binomial is a compound distribution obtained from the binomial by assuming that its probability of success parameter follows a beta distribution

Beta-binomial model

- $i = 1, \dots, s$, where s is the number of samples
- For each cytosine in the genome we have the following model
  - n_i: the number of reads covering the cytosine in sample i
  - m_i: the number of reads that contain "C" readout (i.e. methylated) at the cytosine in sample i ($0 \leq m_i \leq n_i$)
  - If we knew the underlying methylation level p_i, then: $M_i \sim \mathrm{Binom}(p_i, n_i)$

Beta-binomial model

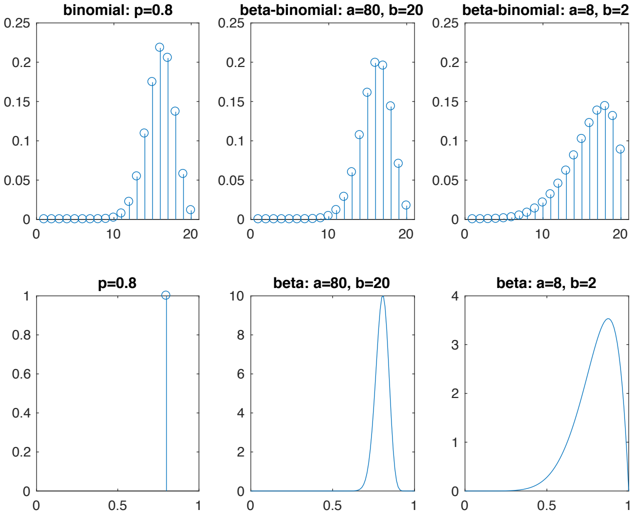
- $i = 1, \ldots, s$, where s is the number of samples
- For each cytosine in the genome we have the following model
  - n_i: the number of reads covering the cytosine in sample i
  - m_i: the number of reads that contain "C" readout (i.e. methylated) at the cytosine in sample i ($0 \leq m_i \leq n_i$)
  - If we knew the underlying methylation level p_i, then: $M_i \sim \mathrm{Binom}(p_i, n_i)$
  - p_i: the unknown methylation level of the cytosine in sample i
  - Instead of assuming a fixed (unknown) methylation level, assume p_i has a compounding distribution $p_i \sim \mathrm{Beta}(\alpha, \beta),\ \alpha \geq 0, \beta \geq 0$
  - The probability of observing methylation level $M_i = m_i$ for a coverage n_i follows so called beta-binomial model

$$\begin{aligned} P(M_i = m_i | n_i, \alpha, \beta) &= \int_0^1 \mathrm{Binom}(m_i | p_i, n_i) \mathrm{Beta}(p_i | \alpha, \beta) \mathrm{d} p_i \\ &= \binom{n_i}{m_i} \frac{\mathrm{B}(m_i + \alpha, n_i - m_i + \beta)}{\mathrm{B}(\alpha, \beta)}, \end{aligned}$$

where B is the beta function

Beta-binomial model

- An illustration of binomial / beta / beta-binomial densities



Binomial and beta-binomial densities

Beta-binomial model

- Mean and variance of the beta-binomial model are

$$\mu = \frac{n_i \alpha}{\alpha + \beta} \quad \text{and} \quad \sigma^2 = \frac{n_i \alpha \beta (\alpha + \beta + n_i)}{(\alpha + \beta)^2 (\alpha + \beta + 1)}$$

- Reparameterization
  - $\pi = \frac{\alpha}{\alpha+\beta}$ is the the average methylation level of a set of replicate samples
  - $\gamma = \frac{1}{\alpha+\beta+1}$ is the common dispersion parameter

  allows us to write the same model as

$$M_i \sim \mathrm{BetaBinomial}(n_i, \pi, \gamma)$$

  where the mean and the variance are now defined as
  - $\mathrm{E}(M_i) = n_i \pi$
  - $\mathrm{Var}(M_i) = n_i \pi (1 - \pi)(1 + (n_i - 1)\gamma)$
- Recall that the variance of the binomial distribution is $n_i \pi (1 - \pi)$ which is smaller than $\mathrm{Var}(M_i)$ for $n_i \geq 2$

Generalized beta-binomial model

- In most of the real world applications, methylation levels can be confounded by one or more factors (e.g. age and smoking)
- The generalized linear model (GLM) generalizes the ordinary linear regression to allow for response variables that have likelihood models other than a normal distribution

Generalized beta-binomial model

- For each sample i (and for each cytosine), the mean methylation level π_i depends on covariates $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{it})^T$

  $$g(\pi_i) = \sum_{j=1}^{t} x_{ij}\eta_j = \mathbf{x}_i^T \boldsymbol{\eta}$$

  where η is a $t \times 1$ parameter vector and

  $$\begin{aligned} g(\pi) &= \mathrm{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right) \\ \pi_i &= \mathrm{logit}^{-1}(\mathbf{x}_i^T \boldsymbol{\eta}) = \mathrm{logistic}(\mathbf{x}_i^T \boldsymbol{\eta}) = \frac{\exp(\mathbf{x}_i^T \boldsymbol{\eta})}{\exp(\mathbf{x}_i^T \boldsymbol{\eta}) + 1} \end{aligned}$$

- $\mathrm{logit}(\cdot) :]0,1[\to \mathbb{R}$, thus $\mathrm{logit}(\cdot)^{-1} : \mathbb{R} \to]0,1[$

Model fitting and inference

- The beta-binomial regression is fit separately for each CpG site
- The parameters $\boldsymbol{\eta}$ and γ are estimated using maximum likelihood
  - Iteratively reweighted least squares algorithm using a Newton-Raphson method
- Test the differential methylation w.r.t. a test factor η_j:
  - Learn the full model and the reduced model without the test factor
  - Compare the models using log-likelihood ratio test

$$D = -2 \ln \left(\frac{\text{likelihood of the reduced model}}{\text{likelihood of the full model}} \right)$$

- p-value from chi-square test with $d_{full} - d_{reduced}$ degrees of freedom, where d_{full} denotes the number of free parameters in the full model

RadMeth application

- Neuron and non-neuron RRBS-seq samples from mouse frontal cortex: $x_{i1} \in \{0, 1\}$
- 6 samples: $s = 6$
- Two additional factors: age ($x_{i2} \in \mathbb{R}_+$), sex ($x_{i3} \in \{0, 1\}$)
- 72 000 differentially methylated (DM) regions between neuron and non-neuron samples that contain at least 10 CpGs
- DM regions with minimum methylation difference above 0.55
  - 1708 lowly methylated (active) regions in neurons
  - These regions are associated with (located close to) 1089 genes
  - GO enrichment analysis by DAVID found a strong association of these genes with various aspects of neuronal development and function

RadMeth application

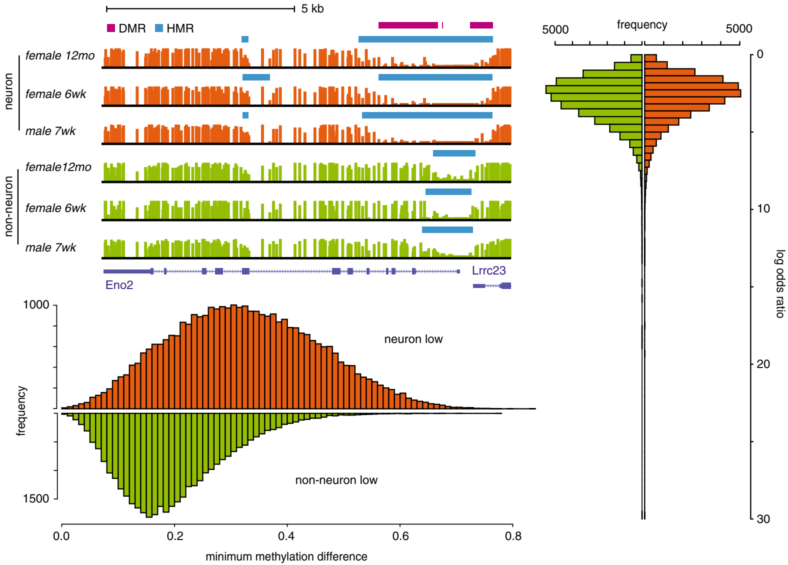


Figure from (Dolzhenko and Andrew, 2014)

References

- Michael J. Booth et al., Quantitative Sequencing of 5-Methylcytosine and 5-Hydroxymethylcytosine at Single-Base Resolution, Science, 336(6063):934-937, 2012
- Jeremy J Day & J David Sweatt, DNA methylation and memory formation, Nature Neuroscience 13:1319-1323, 2010
- Egor Dolzhenko and Andrew D Smith, Using beta-binomial regression for high-precision differential methylation analysis in multifactor whole-genome bisulfite sequencing experiments, BMC Bioinformatics, 15:215, 2014
- Eckhardt F et al., DNA methylation profiling of human chromosomes 6, 20 and 22, Nature Genetics,38(12):1378-85, 2006.
- Felix Krueger and Simon R. Andrews, Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications, Bioinformatics, 27(11):1571-1572, 2011.
- Felix Krueger et al., DNA methylome analysis using short bisulfite sequencing data, Nature Methods 9, 145-151, 2012
- Jialong Lianga et al., Single-Cell Sequencing Technologies: Current and Future, Journal of Genetics and Genomics, 41(10):513-528, 2014
- Alexander Meissner, et al., Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis, Nucleic Acids Res., 33(18):5868-77, 2005.
- Christoph Plass, et al., Mutations in regulators of the epigenome and their connections to global chromatin patterns in cancer, Nature Reviews Genetics 14, 765-780, 2013
- Dirk Schubeler, Epigenomics: Methylation matters, Nature 462:296-297, 2009
- Cornelia G Spruijt & Michiel Vermeulen, DNA methylation: old dog, new tricks?, Nature Structural & Molecular Biology 21, 949-954, 2014