A Whirlwind tour of (Hydroxy)Methylation Analysis
Outline

● Intro to DNA methylation
  ○ Why and where
  ○ Assay

● Hydroxymethylation
  ○ What and why
  ○ Assays (affinity, TAB-seq, OXBS-seq, RRHP)

● An example OXBS-seq experiment
  ○ QC
  ○ Alignment
  ○ Quantitating
  ○ Modeling

● An example RRHP experiment
  ○ Relating to gene expression?
DNA methylation - why care?

- Methylated CpGs are highly mutable but are still maintained
- Chemically stable - covalent change, retained in cell division
- Global hypomethylation observed in cancer
- DNA methylation of repetitive elements prevents transposition
- Affects/correlates with transcription
  - critical for imprinting and X-inactivation
  - CpG island methylation correlates with gene activity
  - May act as “chromatin state memory” during cell replication
DNA methylation - genomic structure

- ~30 million sites within genome,
  - Underrepresented
  - Majority (60-80%) methylated
- **CG** (vast majority), **CHG** or **CHH** (where H correspond to A, T or C)
- Clustered upstream of TSS - CpG islands
  - CpG island methylation = lower transcription
    - Possibly by bulk hindrance of TFs?
  - ~30000 within genome
- Also found in gene bodies and enhancers
  - Methylation of gene bodies can be associated with upregulated transcription
  - Methylation can affect some TF bindings (e.g. CTCF)
  - Intragenic methylation is also associated with the use of alternative promoters or splice variants
DNA hydroxymethylation

5-hydroxymethylcytosine
- Formed by oxidation of 5mC by Tet enzymes
- Part of demethylation pathway
- Enriched in the brain, particularly neurons
- Enriched in stem cells
- Not only an intermediate in demethylation but thought to regulate transcription in its own right
Assays - Bisulfite conversion

GATCGACGATCGGAGC GTAGGTACGACCGTT

Bisulfite Conversion

GATCGAUGATCGGAGUGTAGGTATUGACCGTT

PCR Amplification

GATCGATGTGATCGGAGTG TAGGTATGACCGTT
Assays - Bisulfite conversion

GATCGACGATCGGAGCGTAGGTAACGACGTT

Methylation

GATCGAUUGATCGGAGUGTAGGTAUGACGTT

Bisulfite Conversion

GATCGATCGGAGTGTTAGGGTATTGACGTT

PCR Amplification
Assays

The Toolkit

1. Immunological
   - Antibodies (affinity based)
2. Enzymatic
   - MspI – cuts at CCGG
     - reduced representation (RRHP)
   - Tet oxidation (TAB-seq, affinity)
3. Chemical
   - Oxidation of 5hMC (OXBS-seq)
   - Glucosylation of 5hMC (RRHP, TAB-seq, affinity)
   - Bisulfite conversion of 5MC and 5hMC (OXBS-seq, TAB-seq, affinity)

Assay targets

1. Whole genome
   - Expensive
2. Reduced representation
   - Target regions enriched for CpGs
   - Cheaper but incomplete
     - Enzymatic:
       - MspI cuts at CCGG
       - Size select fragments
       More incomplete, messy, double stranded possible
     - Pulldown
       - Exome-style
       Complete, clean, single strand only

Resolution

1. Regional - affinity
2. Single base – bisulfite methods
Assays - Affinity based profiling

- “hMeDIP-seq”
- Either direct pulldown or via chemical conversion labelling intermediary
- CMS = cytosine-5-methylenesulfonate
  - Highly immunogenic
- biased towards highly hydroxymethylated regions

RRHP

- Reduced Representation Hydroxymethylation Profiling
- “MspI protection assay”
- Cheap
- Semi-quantitative

Assays - TAB-seq

- Tet-assisted bisulfite sequencing
- Can only know 5mC levels by subtraction
- Good if you only want 5hmC levels
- Expensive

Yu et al.. Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome. Cell. 2012 Jun 8;149(6):1368-80
Assays - OXBS-seq

- Oxidative bisulfite treatment
- Can only know 5hmC levels by subtraction
  - bisulfite seq and oxidative bisulfite seq done in parallel on same sample
- Expensive

An OXBS-seq experiment

- Whole genome OXBS-seq
- Occipital and frontal lobe material from controls and Alzheimer’s patients
- Run with CEGX Trumethyl kit

Quality metrics – base sequence content

Bisulfite

Oxy-bisulfite
Quality metrics - bisulfite conversion rate

Low CHH and CHB methylation

Low cytosine content on OT strand
CEGX controls – conversion

OxBS-seq
Preprocessing - trimming

**Usual reasons to trim**
- Remove adapters
- Remove low quality sequences
- Remove primer biased sequences

**Novel reasons to trim**
- 5` bisulfite conversion failure bias

**M-bias plots can help diagnose issues**
M-bias plots

Expect consistent methylation levels regardless of base position on read
Alignment

OXBS-seq
Choice of Aligner

**Bismark**
- Standard used in field
- Slower but accurate
- Well supported
- Best compromise of accuracy, coverage and speed

**BSmap**
- Higher mapping percentage
- Faster
- Lower coverage in bakeoffs
- Didn’t handle paired reads well in our hands
  - Ignored outer-inner read parameters, mapping percentage dropped to Bismark's when filtered
## Summarized alignment data

<table>
<thead>
<tr>
<th>chr</th>
<th>start</th>
<th>end</th>
<th>perc</th>
<th>C</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10470</td>
<td>10470</td>
<td>100</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>10472</td>
<td>10472</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>10485</td>
<td>10485</td>
<td>50</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>10490</td>
<td>10490</td>
<td>100</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>10494</td>
<td>10494</td>
<td>100</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>10498</td>
<td>10498</td>
<td>100</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>10526</td>
<td>10526</td>
<td>67</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>10543</td>
<td>10543</td>
<td>100</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>10564</td>
<td>10564</td>
<td>100</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>10572</td>
<td>10572</td>
<td>100</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>10578</td>
<td>10578</td>
<td>100</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>10580</td>
<td>10580</td>
<td>100</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>10590</td>
<td>10590</td>
<td>100</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>13417</td>
<td>13417</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>13644</td>
<td>13644</td>
<td>100</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>15749</td>
<td>15749</td>
<td>100</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>15769</td>
<td>15769</td>
<td>100</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>15789</td>
<td>15789</td>
<td>100</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>15912</td>
<td>15912</td>
<td>100</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>16059</td>
<td>16059</td>
<td>75</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>16071</td>
<td>16071</td>
<td>80</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>16083</td>
<td>16083</td>
<td>100</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>16140</td>
<td>16140</td>
<td>93</td>
<td>25</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>16142</td>
<td>16142</td>
<td>93</td>
<td>26</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>16243</td>
<td>16243</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>16244</td>
<td>16244</td>
<td>73</td>
<td>19</td>
<td>7</td>
</tr>
<tr>
<td>1</td>
<td>17379</td>
<td>17379</td>
<td>100</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>17407</td>
<td>17407</td>
<td>100</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>17452</td>
<td>17452</td>
<td>100</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>17453</td>
<td>17453</td>
<td>100</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>17478</td>
<td>17478</td>
<td>100</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>17479</td>
<td>17479</td>
<td>67</td>
<td>4</td>
<td>2</td>
</tr>
</tbody>
</table>

**Readout is:**
- Number of C’s (unconverted)
- Number of T’s (converted)

Methylation level is a ratio of unconverted C’s over total reads
- “Cytosine ratio”
- C/(C+T)
- “Beta value”

Beta is sometimes logit transformed for modeling
- Log2(C/T)
- “M value”
Measuring hydroxymethylation

To find the levels of hydroxymethylation:

\[ BS = 5mC + 5hMC \]

\[ OXBS = 5mC \]

(bisulfite cytosine ratio) - (oxybisulfite cytosine ratio)
Measuring hydroxymethylation – our first pass

Filter for >10 read coverage in both samples

Used ratio subtraction to measure 5hMC levels

Used Fisher’s exact test to test if ratios significantly different
Subtraction of ratios can give negative numbers ("overshoot")
• doesn’t account for varying read depth in assays!

In an OXBS-seq dataset - 17% of CpG sites captured by RRBS and oxRRBS experiments exhibited “overshoot”

MLML - maximum likelihood methylation levels
• estimate the % of mC, hmC and un-methylated for each site
• gives a confidence value that can help discard problematic sites
• Improved error rate at low coverage

Measuring hydroxymethylation – number of sites

Client expected ~4% of CpG sites to be “hydroxymethylated”

How do you define a “hydroxymethylated” site?

1. Any particular CpG cytosine in the data can be a mixture of
   • Unmethylated,
   • Methylated
   • Hydroxymethylated
2. Neither bisulfite or oxidative treatment are 100% effective

Cutoffs were informed by control probes included in assay

>0.7 cytosine ratio in the OX-Bsseq assay (low end of conversion rate of 5hMC sites)
>0.3 difference in cytosine ratios between bisulfite and OX-bisulfite assays
   (i.e. majority 5hMC)
Measuring hydroxymethylation – number of sites

- Number of hmC: 38639
- Number of mC: 2845193
- hmC = 1.339849% of the total methylated
Statistical models – incorporating replicates

- linear regression of ratio ~ x.
- 100 reads cover a nucleotide is not statistically equivalent to the case when 2 reads cover a nucleotide
- glm() function in R with the binomial distribution with weights representing the covering reads.

https://mikelove.wordpress.com/2013/09/02/binomial-glm-for-ratios-of-read-counts/

How to weight when we have two assays?

Conservative
- minimum number of reads between the assays

Liberal
- Mean number of reads between the assays
- 5400 sites that show an adjusted p.value of less than 0.05.
RRHP

- Reduced Representation Hydroxymethylation Profiling
- “MspI protection assay”
- cheap

RRHP experiment

RRHP after Tet2 knockdown in primary corneal epithelium line

- RRHP was single replicate
- Run in parallel with 3 replicates of RNA-seq
- Run and processed by Zymo
  - used their counts (and annotations when available)
Correlating RRHP results to gene expression levels

Effects of Tet2 kd on 5hMC levels in corneal epithelium
Differential Expression in Tet2 knockdowns
Genomic context of 5hMC changes

siRNA1 treatment

cgi | N | Y
---|---|---
exon | N | Y
intron | N | Y

Density vs. log2 fold change

- exon N
- exon Y
- intron N
- intron Y
- promoter N
- promoter Y
5hMC change versus gene expression changes

RNA-seq log2 fold change

5hmc log2 fold change
Regularized and bias adjusted 5hMC change versus expression change
Q16: Is RRHP a quantitative method? How does it differ from TAB-seq and oxBs-seq, which also looks at 5hmC?

RRHP is a semi-quantitative method as the level of 5hmC is deduced based on the number of sequencing reads a particular CCGG site receives (i.e. higher reads are correlated with higher 5hmC level). TAB-seq and oxBs-seq are quantitative methods in which you will determine the 5hmC%. However, these methods rely on bisulfite conversion and require higher sequencing depth. OxBs-seq requires double the amount of sequencing as it requires subtractive sequencing to determine the 5hmC%.
RRHP – future plans

- Convince the researcher to run replicate samples
- See if subsets of DE genes associated with different 5hMC changes are enriched in certain functional classes
- Abandon entirely and get them to run TAB-seq or RRTAB-seq
Acknowledgements

**OXBS-seq work**
Harvard NeuroDiscovery Center

**Shi Lab**
Yujiang Shi
Ira Fetahu
Feizhen Wu
Dingaliu Ma

**RRHP work**
Harvard Stem Cell Institute

**Frank Lab**
Natasha Frank
Lynn Guo
Jiang Yusheng

[Image of Lorena Pantano]

[Image of Rory Kirchner]
Working with the summarized data

Many different packages in R/bioC for working with bisulfite data

MethyKit is a suite of tools that does just about everything (though poorly)

Most of the other packages will accept the bsseq format (bsseq BioC package)

Bsseq::read.bismark - read in Bismark formatted data of different types (I like the CpG report as it has strand info)

Working with 30 million CpGs is difficult and wasteful as many loci have no coverage in any samples

- can pre-specify regions/loci or let bsseq find them (this takes longer)
- Some packages have disk based objects for access
- Parallelization options
SNPs

- Bismark aligns to a reference genome
- C/T SNPs (common mutations after methylated cytosine deamination events) can look like de-methylation events

Bis-SNP

- Combined DNA methylation and SNP calling for Bisulfite-seq data
  - Require double stranded data (not all methylation assays have this!)
<table>
<thead>
<tr>
<th></th>
<th>DMRs = Differentially Methylated Regions</th>
<th>DMLs = Differentially Methylation Loci</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Pros</strong></td>
<td>• Signals often cluster, particularly in CpG islands</td>
<td>• Won’t miss smaller signals • Easy to measure</td>
</tr>
<tr>
<td></td>
<td><strong>Pros</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Low multiple testing penalty • Easier to interpret • Less likely to be artefact</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Cons</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Can miss smaller signals • How to measure?</td>
<td>• High multiple testing correction penalty • Redundant information • Role of single DML often unclear • Can be a SNP artefact</td>
</tr>
</tbody>
</table>
DMR methods

Algorithmic

- Gathering clusters of DMLs - DSS
  - % significant DMLs in cluster
  - Significance cutoff
  - Distance to next CpG cutoff

Statistical

Predefined regions and sliding windows - methylKit
  - summarize C and T counts within region/window

Smoothing - bumphunter (array) and Bsmooth (NGS)
Some recent DMR results
Annotation - Typical approaches

The issues

- Goal is to map DML/DMR CpGs to biological functions
- Biological functions usually encoded at the gene level
- Have to relate CpG information to genes

Approaches

- Closest gene
- Changes to CpG islands in promoter
Annotation - problems

- Gene size - larger genes more likely to have nearby CpG
- Signal variability and integration of multiple signals
  - DMLs versus DMRs
  - What does diffuse methylation in a region mean?
  - Is one CpG in a region more important than another?
- Predicting effect on gene expression depends on context
  - CpG island methylation = gene transcription off
  - Gene body methylation = gene transcription on
  - Intergenic methylation = who knows?
- Formats
Technologies - microarray

1. Illumina 27K/450K/850K microarrays
   - Based on probe binding and extension
   - Two types of probes, with different distributions
     i. Type I - 2 probes, terminate on cytosine
        ● one probe for unmethylated, one for methylated
        ● positive signal if can extend probe to next base
        ● single channel
        ● signal based on probe binding
     ii. Type II - 1 probe, terminates prior to assayed cytosine
        ● two colors, one for unmethylated, one for methylated
        ● Signal based on probe binding and extension
   - Challenging normalization
   - Mainly targeted to CpG islands and surrounding regions
   - Lots of R and Bioconductor support
   - HUMAN ONLY
<table>
<thead>
<tr>
<th>Method</th>
<th>CpGi focus</th>
<th>Sensitivity</th>
<th>All CpGs</th>
<th>Price</th>
<th>Commercial mouse</th>
<th>Bioinfo support</th>
<th>Cell composition</th>
<th>Call SNPs</th>
</tr>
</thead>
<tbody>
<tr>
<td>WGBS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RRBS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Truseq</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>450/850K</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Bisulfite methods

Whole genome bisulfite sequencing

- Large files (120GB)
- Recommended 30 X coverage across all samples

Targeted methods

- Target CpG islands and other areas of interest

Enzymatic digestion based - Reduced Representation Bisulfite sequencing (RRBS) (3-5GB)

Capture based - Illumina TruSeq Methyl Capture EPIC (~15GB)
RRBS - Reduced Representation Bisulfite Sequencing

Single end (common) or paired end
low diversity
Illumina Truseq EPIC Methyl Capture

Paired end sequencing
Recommended 55 million reads = mean coverage >40X, max CpGs with >10X
CpG islands - definition and some facts

1. length greater than 200bp
2. G+C content greater than 50%
3. a ratio of observed to expected CpG greater than 0.6

Major regulatory units

- Methylation of CpG islands in promoters represses transcription
- 90% are constitutively unmethylated, enriched for H3K4 methylation
- methylated CpG Islands are often intergenic or intragenic
- 60-75% of genes have them in their promoters
CpG islands - genomic structure

- shores show variable tissue-specific DNA methylation patterns often altered in tumorigenesis
Alignment questions and issues

- Strand sequenced?
  - Capture methods may only capture one strand before bisulfite conversion

- Directional versus non-directional libraries?
  - Directional - only OT and OB post-bisulfite PCR strands sequenced
  - Non-directional - all 4 possible strands sequenced (OT, CTOT, OB, CTOB)
  - Typically only see directional, but Bismark has output to determine which it is

- Paired versus single end
  - Typically no reason to do paired end with RRBS or Truseq

- Align to full genome or to partial genome?
  - Whole genome = conservative approach as pulldowns and digests are not perfect
  - Reduced genome = better mapping percentages, but are they real? Found ~8% off target alignment with Truseq methyl capture method
  - What about duplicate reads?
    - Depends on application
      - Reduced representation - don’t dedupe, assayed regions too small
      - Whole genome - dedupe
Alignment

**Legend**
- CpG’s in red = original sequence
- CpG’s in blue = converted
- Positions corresponding to original C’s in CpG underlined

- OT = original top strand
- CTOT = complementary to OT
- OB = original bottom strand
- CTOB = complementary to OB

In IGV
- For OT, CTOT: C>T; C>C
- For CTOB, OB: G>A; G>G

**methylated locus**

ATACGCGTATT-3’
TATAAGCGCATATA-5’

**unmethylated locus**

ATATCGCGTATT-3’
TATAAGCGCATATA-5’

Bisulfite conversion

ATATCGUGTATT-3’
TATAAGGUATATA-5’

Polymerase chain reaction

ATATCGGTATT-3’ OT
TATAACGACATATA-5’ CTOT

ATATCGCATATT-3’ CTOB
TATAAGCGTATA-5’ OB
Cell composition

- Changes in DNA methylation may actually reflect changes in cellular composition, especially in blood.

- Need to adjust for cellular composition effects.
- Established methods to quantitate and adjust for cell composition effects exist for arrays.
- Adapted but not established methods exist for RRBS.
- For Truseq and WGBS, can only adjust (using SVA).
Genomic inflation

Population structure and hidden factors (eg SNPs and cell composition effects) can massively inflate pvalues.

Genomic methods not appropriate as assumption of few real effects is broken.
Statistical models for differential methylation

Goal

- find differences in C/T ratios at loci
- Incorporate read number variation (i.e. coverage variation) for both assays
- Incorporate replicates

General Methods

- Fisher’s exact test - does not account for replicates or variation
- Regression - logistic (MethylKit), beta after smoothing (BiSeq)
- Beta-binomial model (MOABS, DSS, MethylSig)
  - can account for both sampling and epigenetic variability
  - Bayesian, hierarchical (DSS)

Packages

BiSeq, MOABS, MethylSig, DSS, RADMeth, MLML