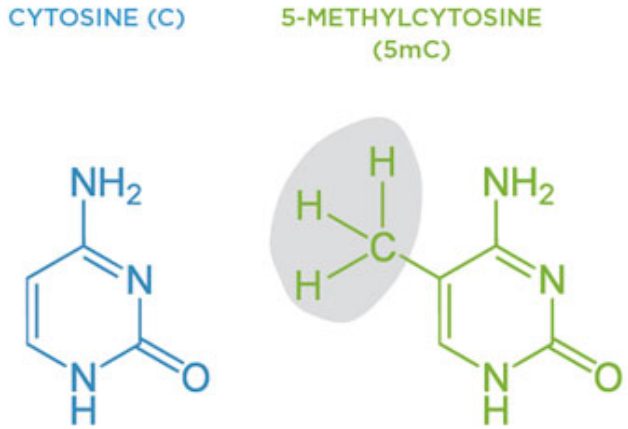
A Whirlwind tour of (Hydroxy)Methylation Analysis

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 (eg. CTCF)
  - intragenic methylation is also associated with the use of alternative promoters or splice variants

Epigenetic Regions[a]	Genome
Total Size	3000 Mb
CpG Sites	28,217,009
CpG Islands	27,718
WGBS DMRs	1,853,606
GenCode Promoters	149,463
Open Chromatin (synthesis)	303,304
TFBS	229,067
MEDs	264,834
FANTOM5 Enhancers	28,563
CTCF ChIP Peaks	433,466

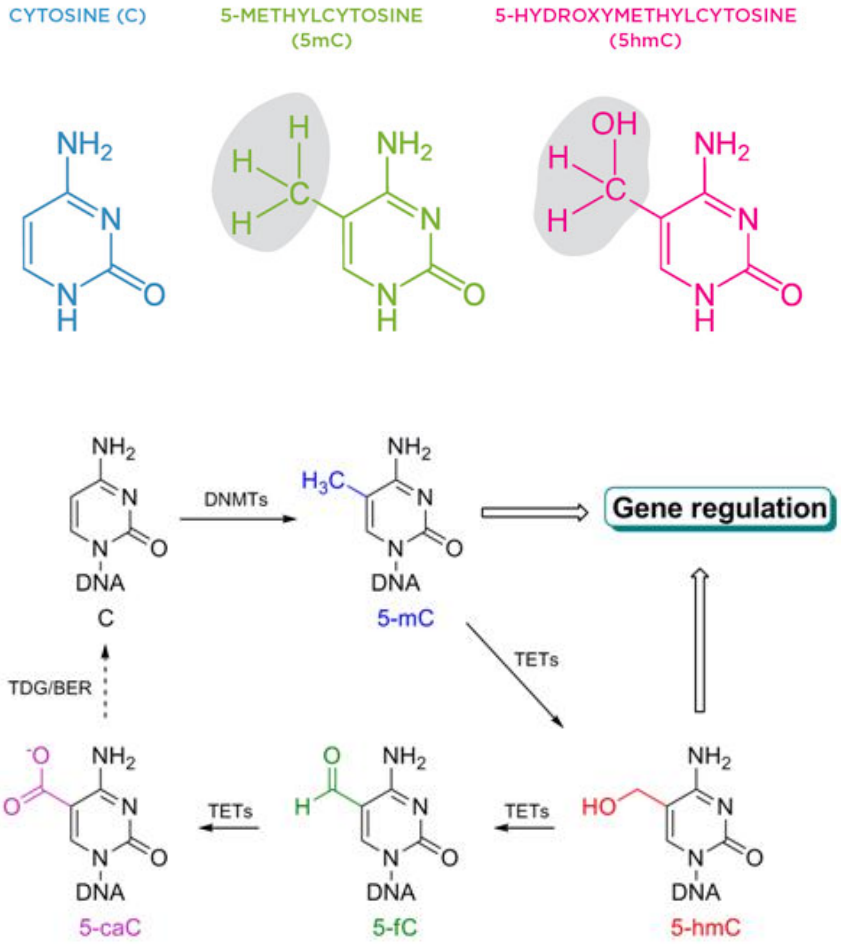
a. Cytosine phosphate guanine (CpG), whole-genome bisulfite sequencing (WGBS), differentially methylated regions (DMRs), transcription factor binding sites (TFBS), multifunctional epigenetic domains (MEDs), the *CTCF* gene encodes a transcription factor

DNA hydroxymethylation

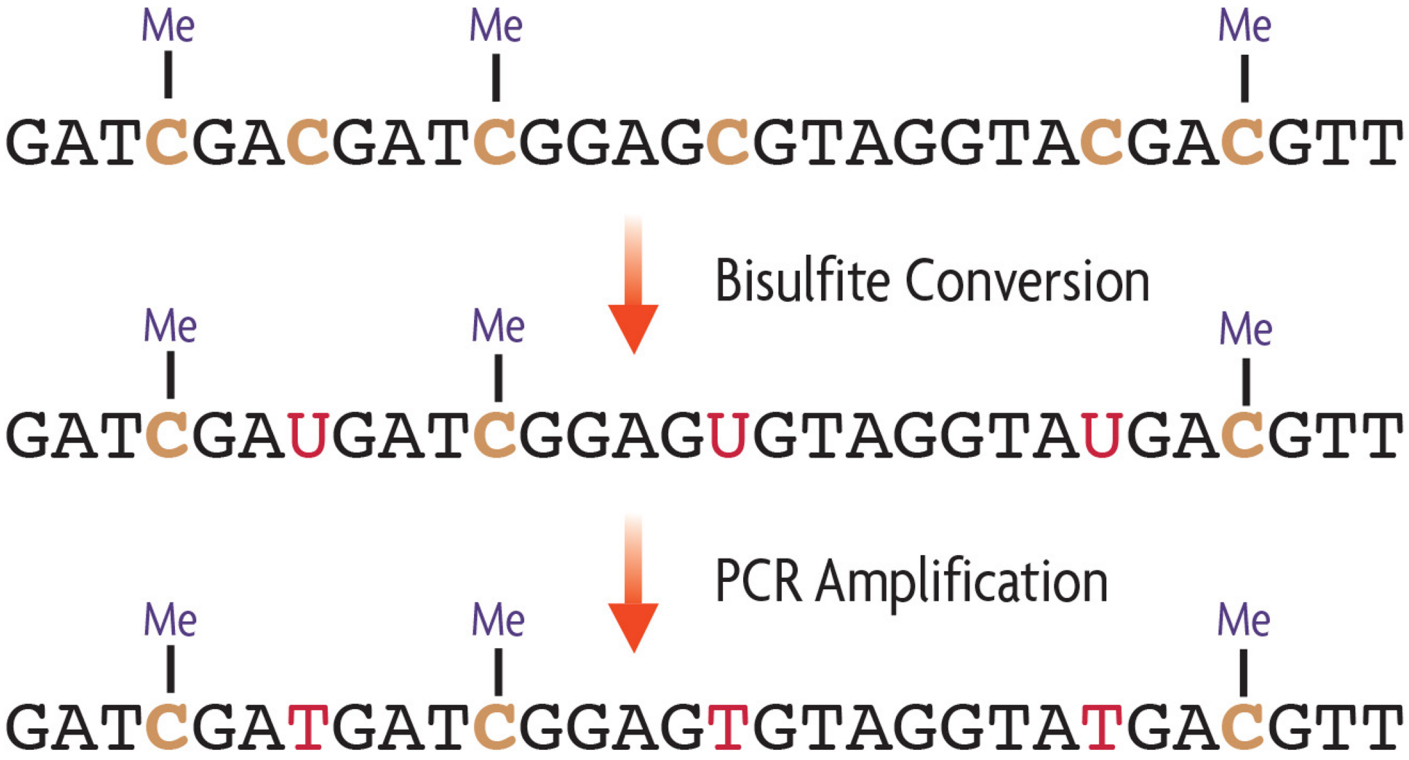
5-hydroxymethylcytosine

- Formed by oxidation of 5mC by Tet enzymes
- Part of demethylation pathway

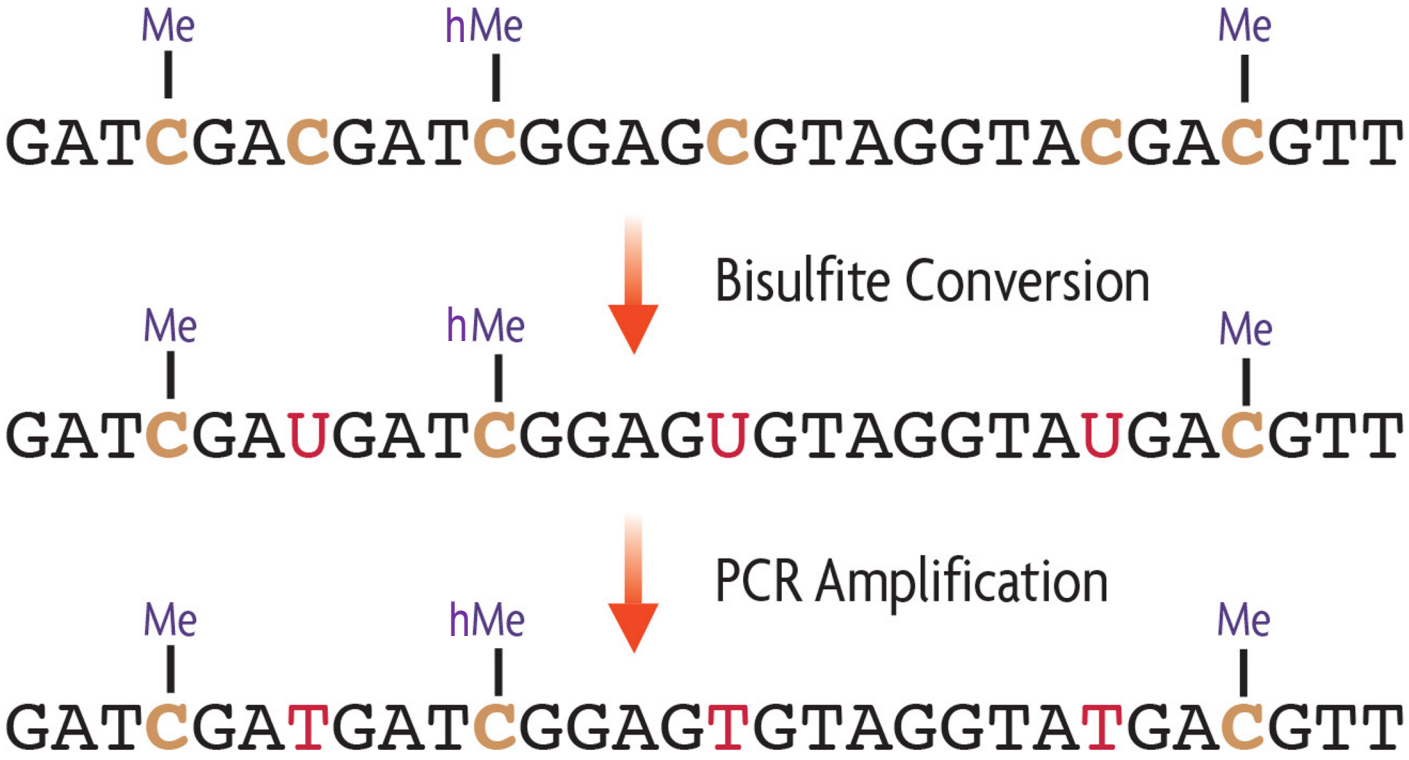
- Enriched in the brain, particularly neurons
- Enriched in stems cells
- Not only an intermediate in demethylation
  - thought to regulate transcription in its own right



Assays - Bisulfite conversion



Assays - Bisulfite conversion



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)

Resolution

1. **Regional - affinity**
2. **Single base – bisulfite methods**

Assay targets

1. **Whole genome**
    - Expensive
2. **Reduced representation**

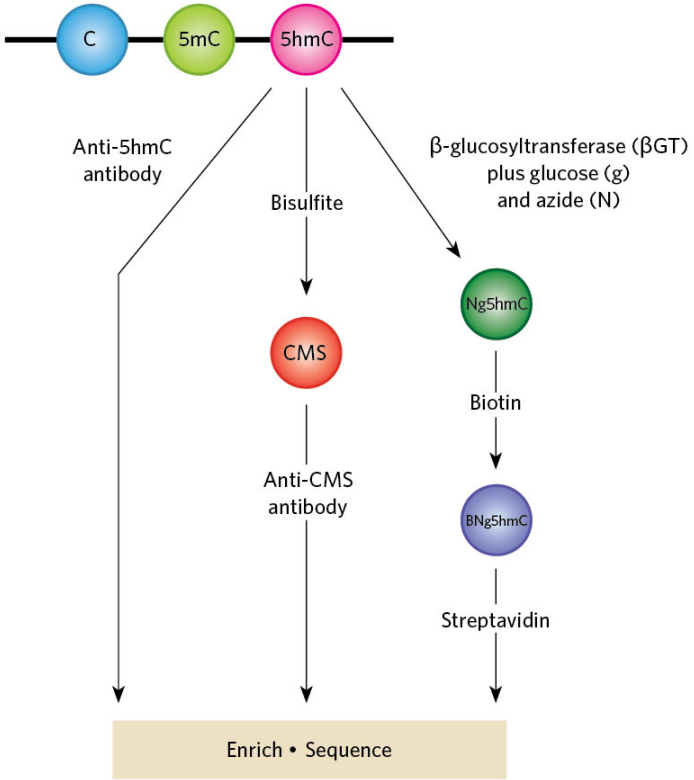
- Target regions enriched for CpGs
- Cheaper but incomplete
    - Enzymatic:
        - MspI cuts at CCGG
        - Size select fragments

      More incomplete, messy, double stranded possible
    - Pulldown
        - Exome-style

      Complete, clean, single strand only

Assays - Affinity based profiling

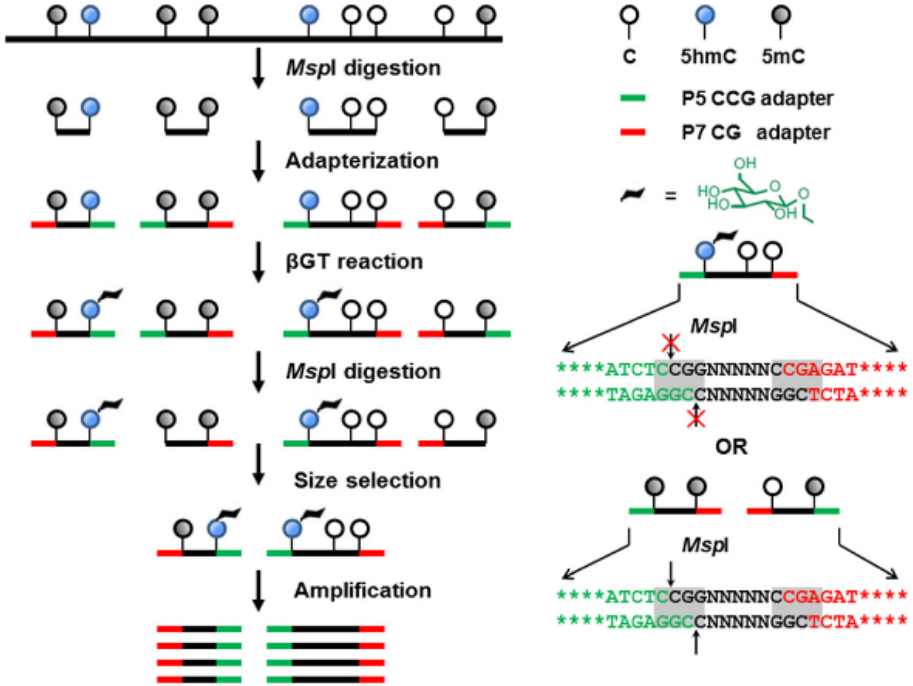
- “hMeDIP-seq”
- Either direct pulldown or via chemical conversion labelling intermediary
- CMS = cytosine-5-methylenesulfonate
    - Highly immunogenic
- biased towards highly hydroxymethylated regions



Ficz et al. Dynamic regulation of 5-hydroxymethylcytosine in mouse ES cells and during differentiation. Nature. 2011 May 19;473(7347):398-402.

RRHP

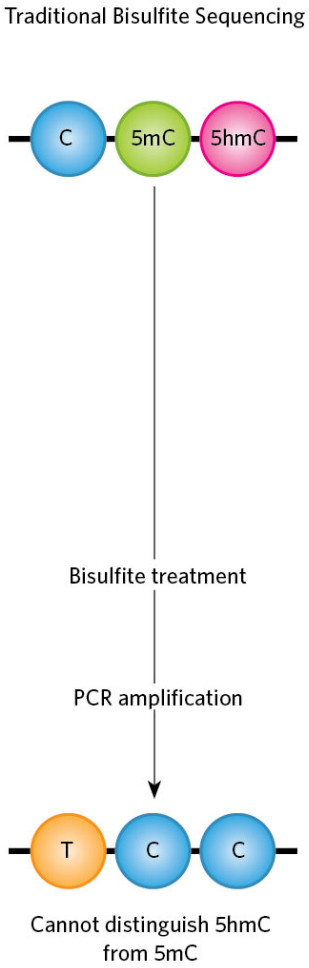
- Reduced Representation Hydroxymethylation Profiling
- “MspI protection assay”
- Cheap
- Semi-quantitative

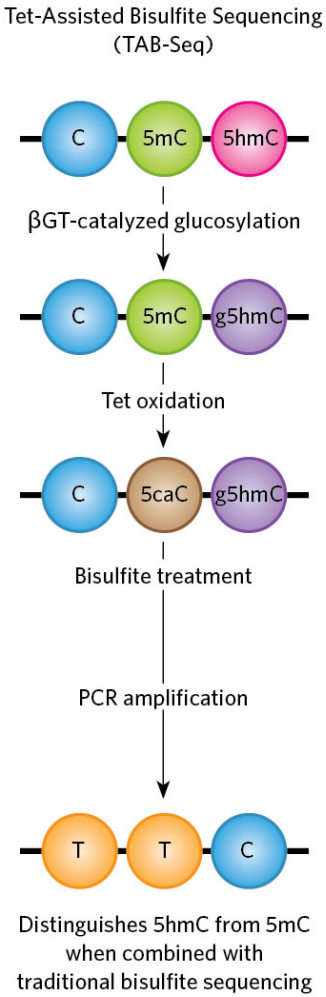


Petterson et al. RRHP: a tag-based approach for 5-hydroxymethylcytosine mapping at single-site resolution. Genome Biol. 2014 Sep 24;15(9):456.

Assays - TAB-seq

- Tet-assisted bisulfite sequencing
- Can only know 5mC levels by subtraction
- good if you only want 5hMC levels
- expensive

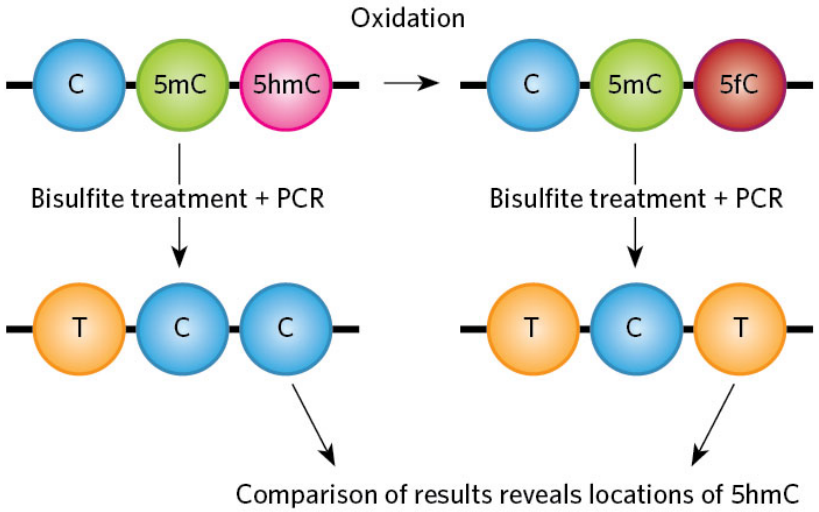




Yu et al.. Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome. Cell. 2012 Jun 8;149(6):1368-80

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



Booth et al. Oxidative bisulfite sequencing of 5-methylcytosine and 5-hydroxymethylcytosine. Nat Protoc. 2013 Oct;8(10):1841-51

An OXBS-seq experiment

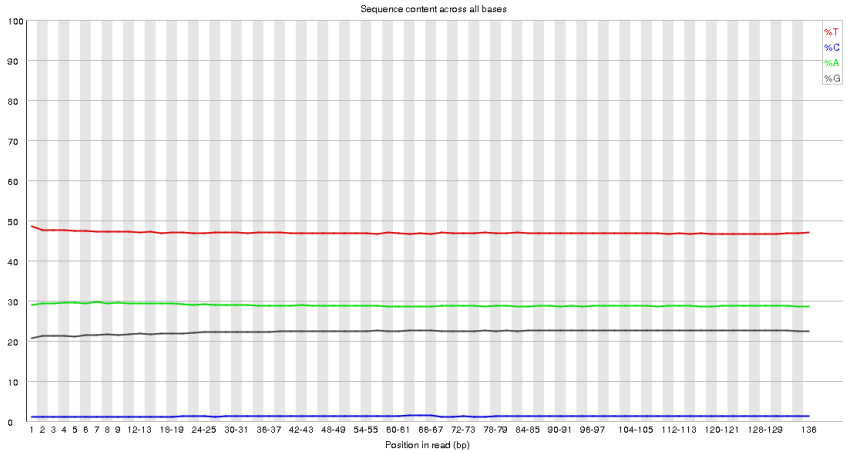
- Whole genome **OXBS-seq**
- Occiptal and frontal lobe material from controls and Alzheimer’s patients

- Run with CEGX Trumethyl kit

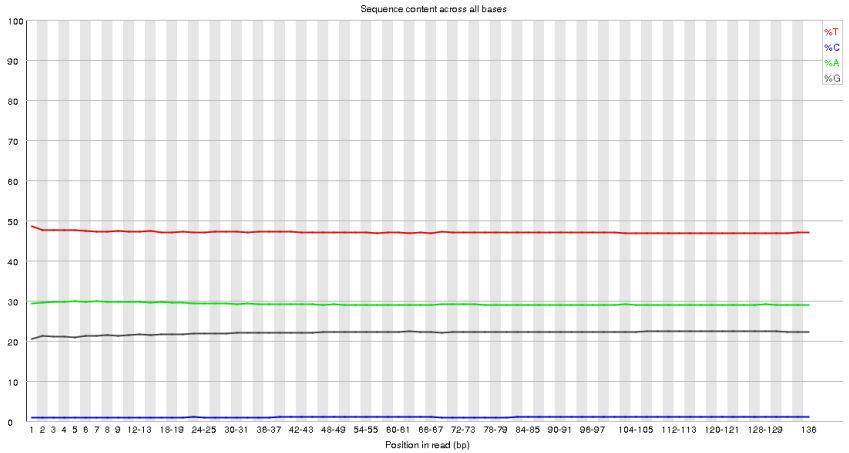
Fetahu IS, Ma D, Rabidou K, Argueta C, Smith M, Liu H, Wu F, Shi YG. Epigenetic signatures of methylated DNA cytosine in Alzheimer's disease. Sci Adv. 2019 Aug 28;5(8):eaaw2880.

Quality metrics – base sequence content

Bisulfite



Oxy-bisulfite

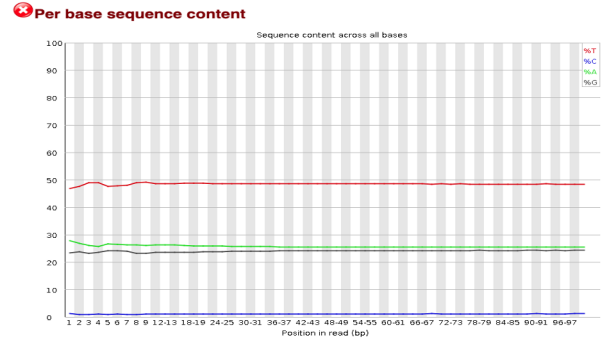


Quality metrics - bisulfite conversion rate

Low CHH and CHB methylation

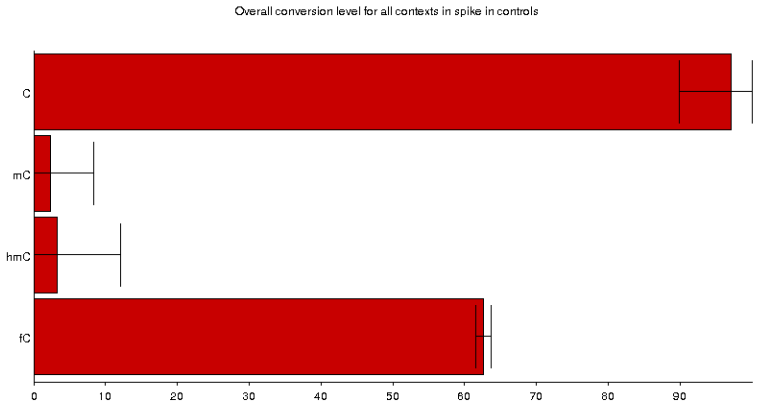
sampleID	numprocessed reads	% trimmed reads	% quality truncated reads	% reads too short after trimming	posttrim basic statistics	overrepresented seq	mapping efficiency	% methylated CpG	% methylated CHG	% methylated CHH
1	1838294	66	61.4	44.3	pass	pass	72.7	44.6	0.2	0.1
2	1449515	42.9	43.8	5.7	pass	warn	73.4	44.5	0.2	0.1
3	449644	46.3	44.5	6.1	pass	warn	71.4	41.9	0.2	0.1
4	3089154	44.7	44.1	6.4	pass	warn	72.3	44.2	0.2	0.1
5	477358	49.6	45.2	7.8	pass	pass	71.8	38.5	0.2	0.1
6	39507856	44	43.4	5.5	pass	warn	71.7	46.3	0.2	0.1
7	34598391	44.2	44.1	5.5	pass	warn	72.4	47	0.2	0.1
8	1681112	43.6	43.6	5.6	pass	warn	73.5	43.9	0.2	0.1
9	1824813	44.8	43.6	5.7	pass	warn	73.2	43.3	0.2	0.1
10	24765404	43.8	43.9	5.5	pass	pass	73.7	44.1	0.2	0.1
11	1106857	46.9	44.9	6.8	pass	pass	72.3	42.6	0.2	0.1
12	9232929	44.8	44	5.5	pass	warn	72.9	42	0.2	0.1
13	6815070	49.1	46.2	15.7	pass	pass	72.4	46.9	0.2	0.1
14	9629944	47.9	53.5	16.5	pass	pass	72	47.4	0.2	0.1
15	7708406	42.9	40.8	4.4	pass	pass	74.3	41.5	0.2	0.1
16	11283330	42.8	49.5	7.1	pass	pass	73.9	42.1	0.2	0.1
17	1773224	44.5	41.4	4.6	pass	pass	73.4	40.2	0.2	0.1
18	2567044	44	50.4	7.4	pass	warn	73.1	40.7	0.2	0.1
19	8550728	45.1	41.7	5.1	pass	pass	73.4	40.1	0.2	0.1
20	12353762	44.6	50.6	7.9	pass	warn	73.1	40.7	0.2	0.1
21	7370975	45.4	40.5	4.7	pass	pass	72.5	42.5	0.2	0.1
22	10618419	44.8	49.4	7.4	pass	pass	72.1	43.1	0.2	0.1
23	16882723	44.3	41.2	4.7	pass	warn	71.7	45.2	0.2	0.1
24	24396582	44.1	50.2	7.6	pass	warn	71.3	45.8	0.2	0.1
25	22860469	47.1	41.2	4.7	pass	warn	71.7	40.8	0.2	0.1
26	32966651	46.6	50.3	7.7	pass	warn	71.5	41.5	0.2	0.1
27	8820038	45.3	41	4.5	pass	pass	74	41.7	0.2	0.1
28	12754805	44.9	49.9	7.4	pass	pass	73.7	42.3	0.2	0.1

low cytosine content on OT strand

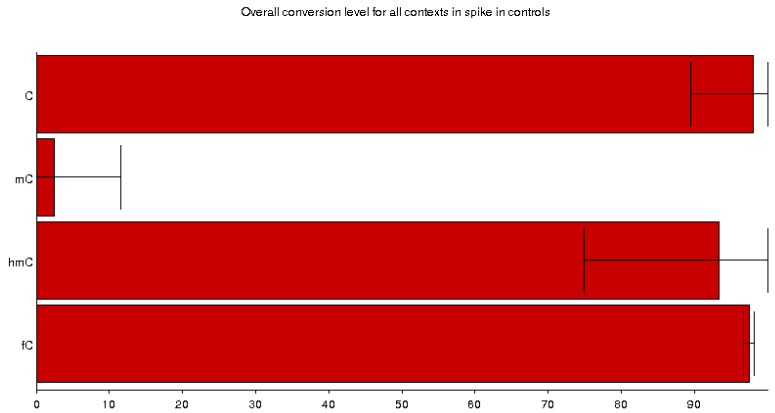


CEGX controls – conversion

Bisulfite



Oxy-bisulfite



Preprocessing - trimming

Usual reasons to trim

- Remove adapters
- Remove low quality sequences
- Remove primer biased sequences

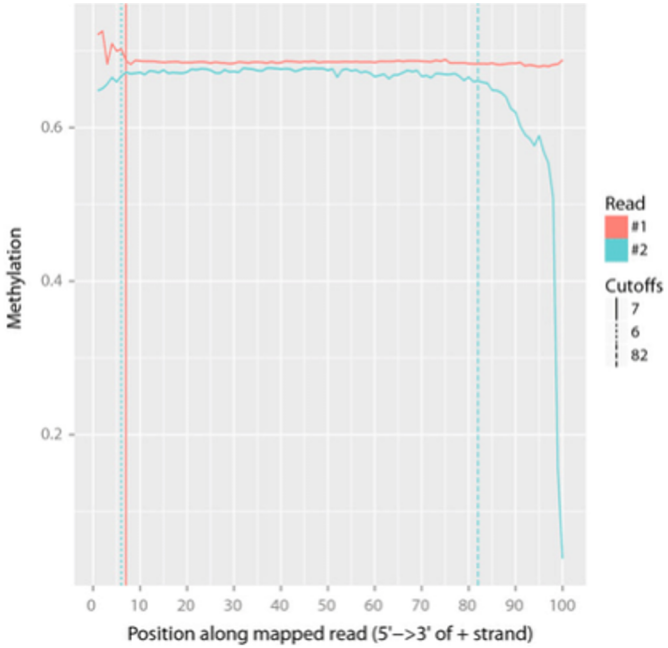
Novel reasons to trim

- 5` bisulfite conversion failure bias

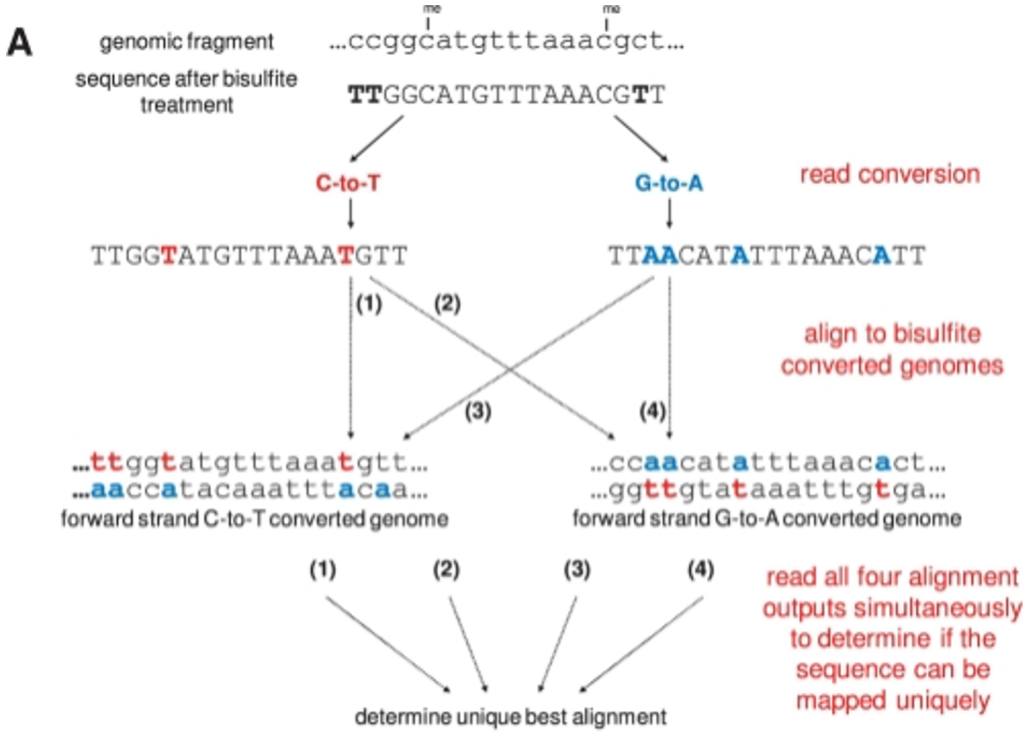
M-bias plots can help diagnose issues

M-bias plots

Expect consistent methylation levels regardless of base position on read



Alignment


A
genomic fragment
...ccggcatgtttaaacgct...
me
me
sequence after bisulfite treatment
TTGGCATGTTTAAACGTT
C-to-T
G-to-A
read conversion
TTGGTATGTTTAAATGTT
TTAACATATTTAAACATT
(1)
(2)
(3)
(4)
align to bisulfite converted genomes
...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

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

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

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"

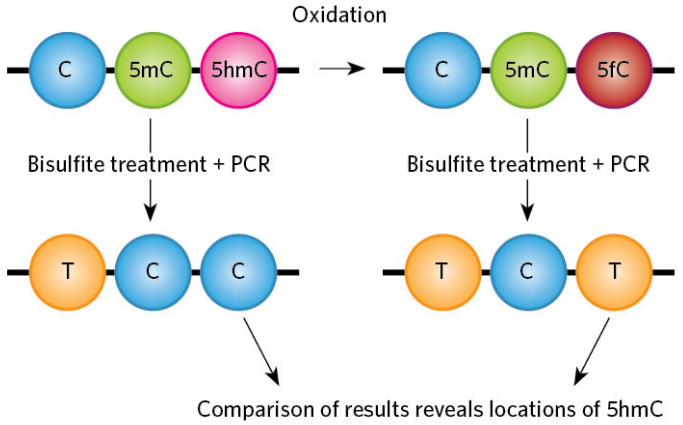
Measuring hydroxymethylation

To find the levels of hydroxymethylation:

BS = 5mC + 5hMC

OXBS = 5mC

(bisulfite cytosine ratio) - (oxybisulfite cytosine ratio)



Measuring hydroxymethylation – our first pass

```
library(readr)
library(ggvis)
library(dplyr)

setwd("~/Desktop/")
alldat <- read_delim("s409_hmc_fixed.tsv", delim="\t")

#dat <- filter(alldat, chr!="chrM")
dat <- alldat

dat <- filter(dat, (CT_counts)>=10, (ox_CT_count)>=10)   #coverage filter
dat <- rowwise(dat) %>%  mutate(., fisher=fisher.test(x=matrix(c(C_counts, (CT_counts-C_counts), ox_C_count, (ox_CT_count-ox_C_count)),ncol=2), alternative="greater")$p.value)
dat <-   dplyr::select(dat, chr, pos, strand, C_counts, CT_counts, ox_C_count, ox_CT_count, fisher)
dat <- mutate(dat, ratiobs=(C_counts)/(CT_counts), ratiooxbs=(ox_C_count)/(ox_CT_count)  )
dat <- mutate(dat, diff=ratiobs-ratiooxbs)
dat$padjust <- p.adjust(dat$fisher, method="fdr")

#randomly sample table for plotting
ggplot(sample_n(dat,1e6), aes(x=-log10(padjust), y=CT_counts))+geom_point(alpha=0.05)
```

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

Measuring hydroxymethylation – better method

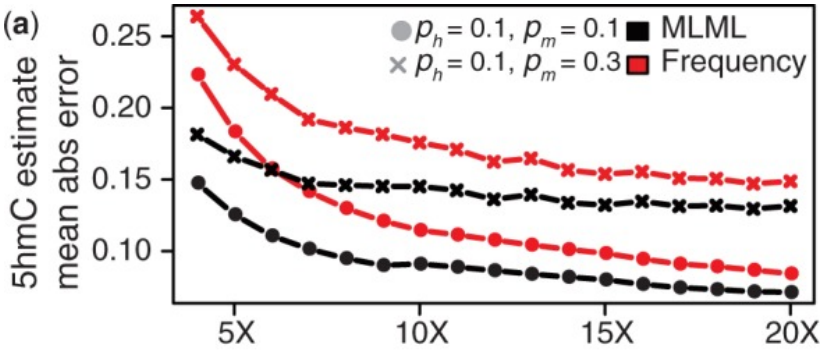
Subtraction of ratios can give negative numbers (”overshoot”)

- 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



Qu et al. MLML: consistent simultaneous estimates of DNA methylation and hydroxymethylation. Bioinformatics. 2013 Oct 15;29(20):2645-6

Measuring hydroxymethylation – number of sites

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

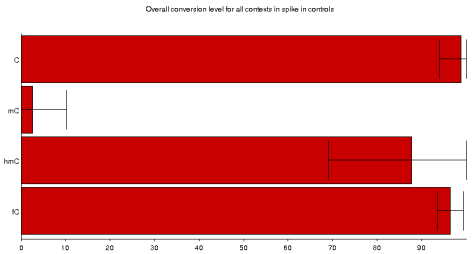
How do you define a “hydroxymethylated” site?

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

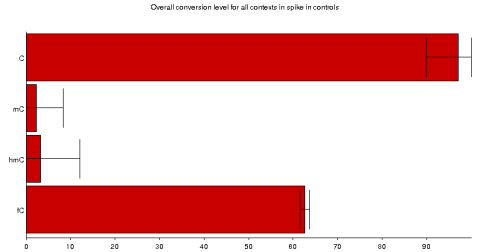
Cutoffs were informed by control probes included in assay

>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)

OXBS-seq

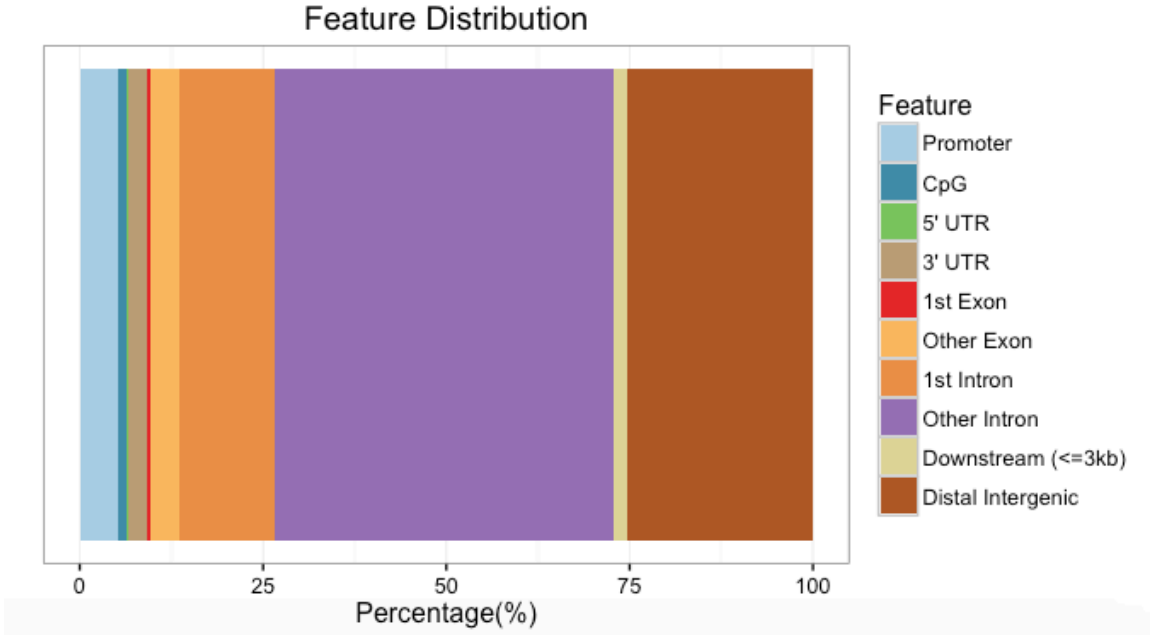


BS-seq



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.

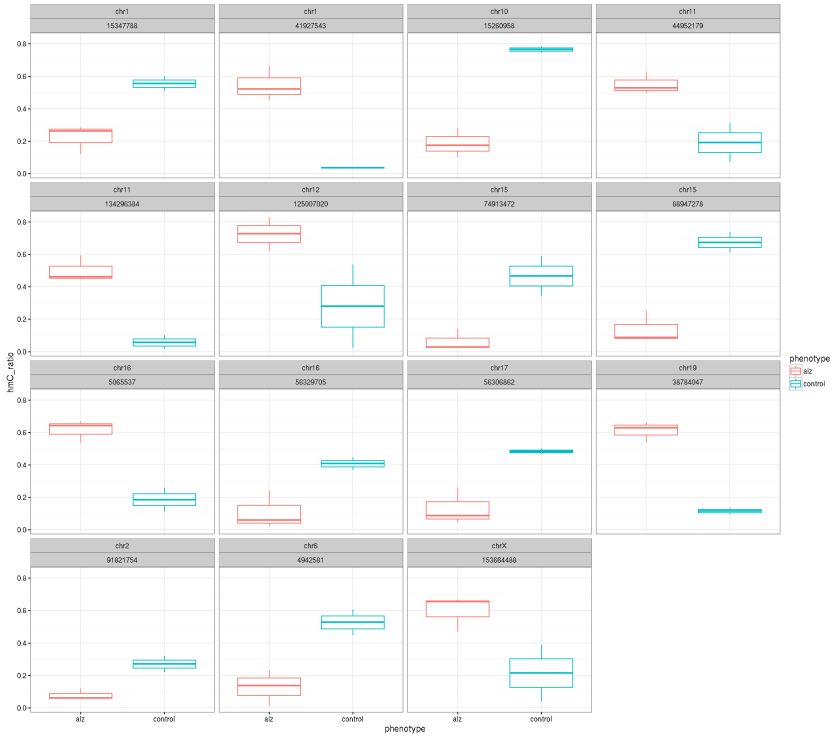
https://mikelove.wordpress.com/2013/09/02/binomial-glm-for-ratios-of-read-counts/

How to weight when we have two assays?

Conservative
- minimum number of reads between the assays

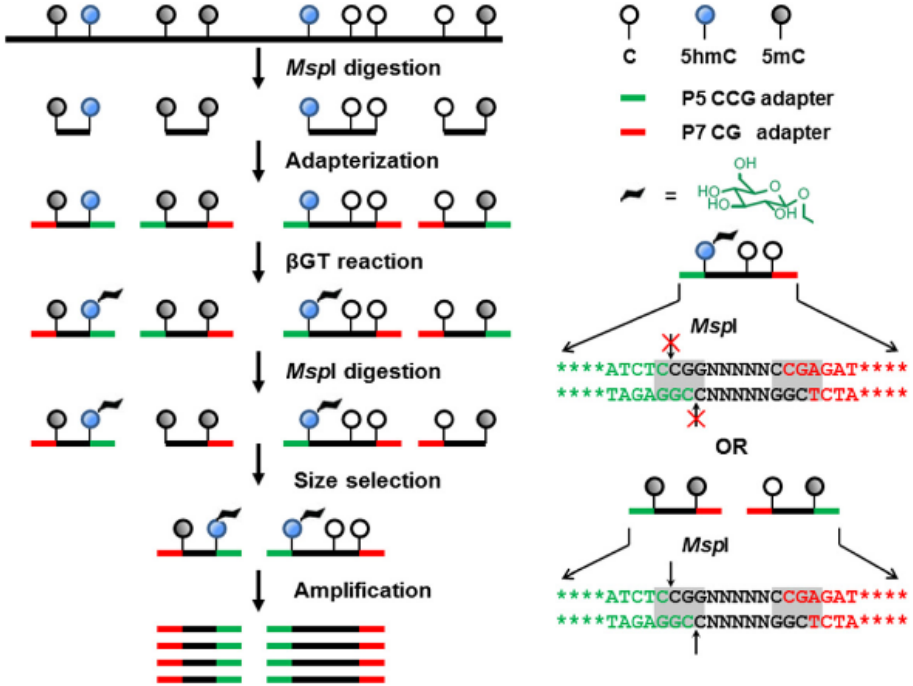
Liberal
- Mean number of reads between the assays

- 5400 sites that show an adjusted p.value of less than 0.05.



RRHP

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



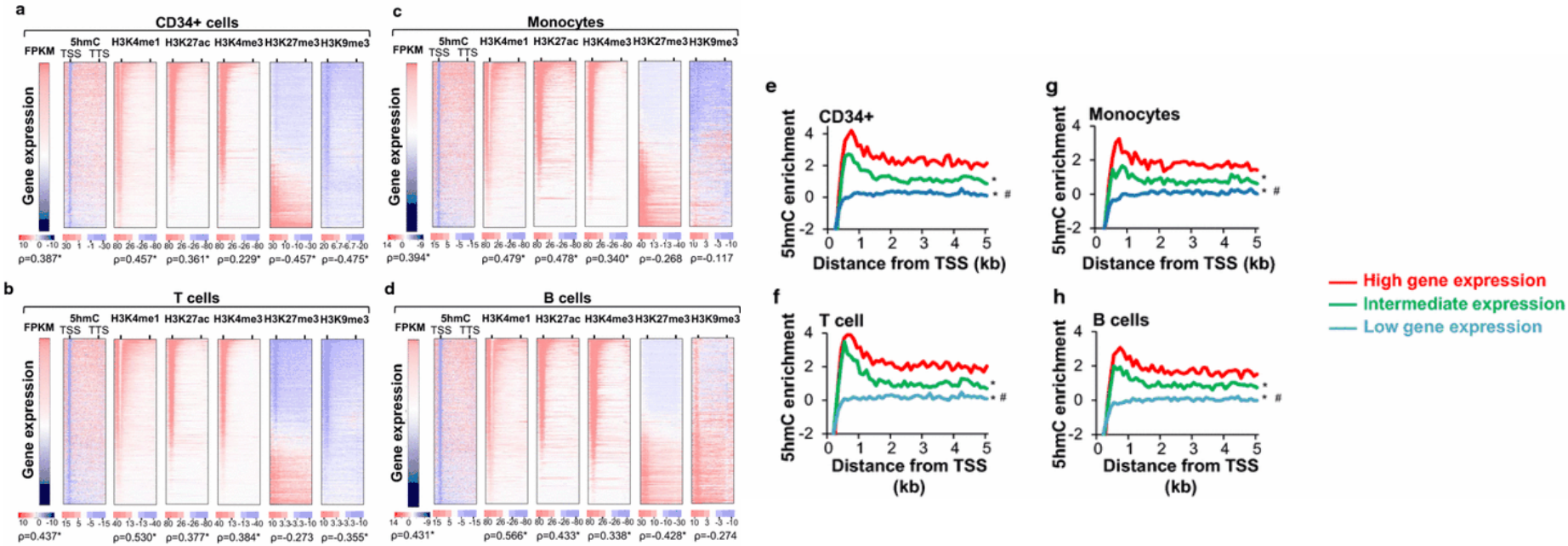
Petterson et al. RRHP: a tag-based approach for 5-hydroxymethylcytosine mapping at single-site resolution. Genome Biol. 2014 Sep 24;15(9):456.

RRHP experiment

RRHP after Tet2 knockdown in primary corneal epithelium line

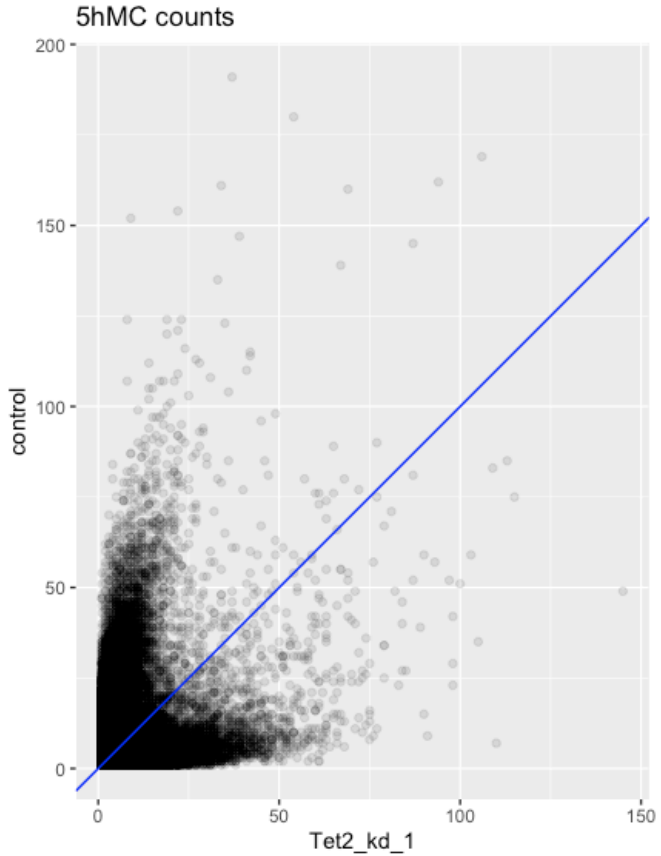
- RRHP was single replicate
- Run in parallel with 3 replicates of RNA-seq
- Run and processed by Zymo
    - used their counts (and annotations when available)

Correlating RRHP results to gene expression levels

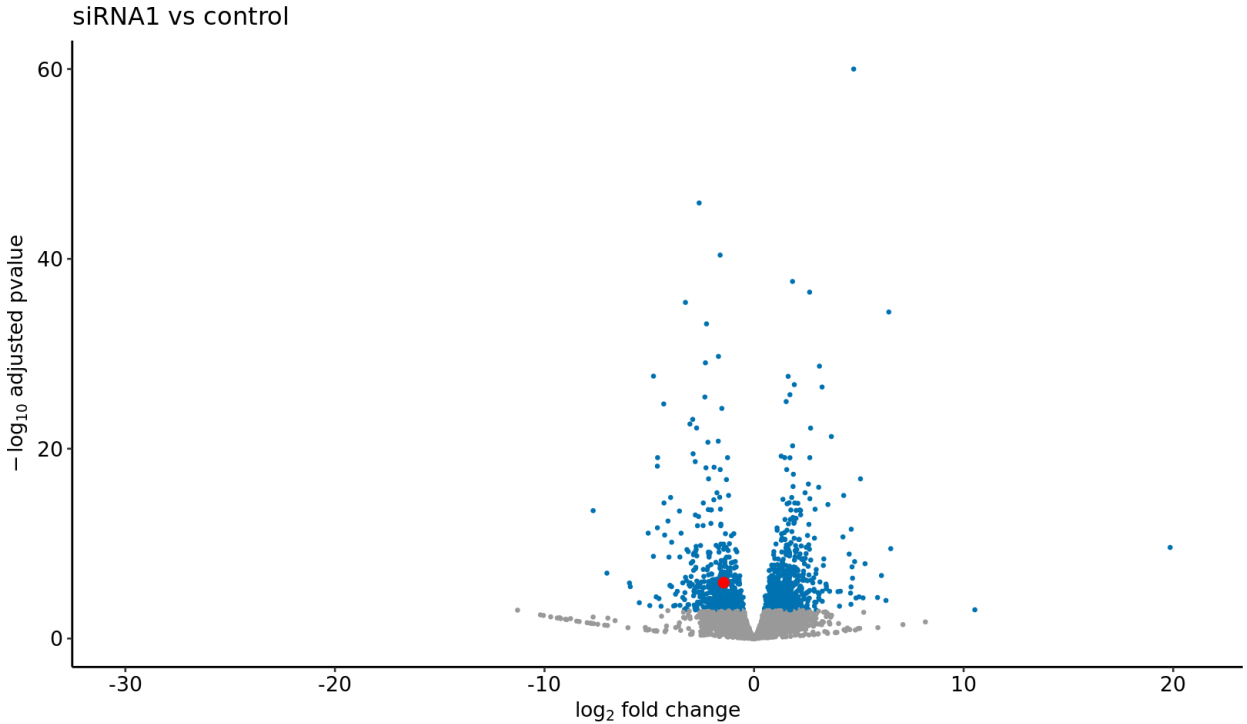


Tekpli et al. Changes of 5-hydroxymethylcytosine distribution during myeloid and lymphoid differentiation of CD34+ cells. Epigenetics Chromatin. 2016 May 31;9:21.

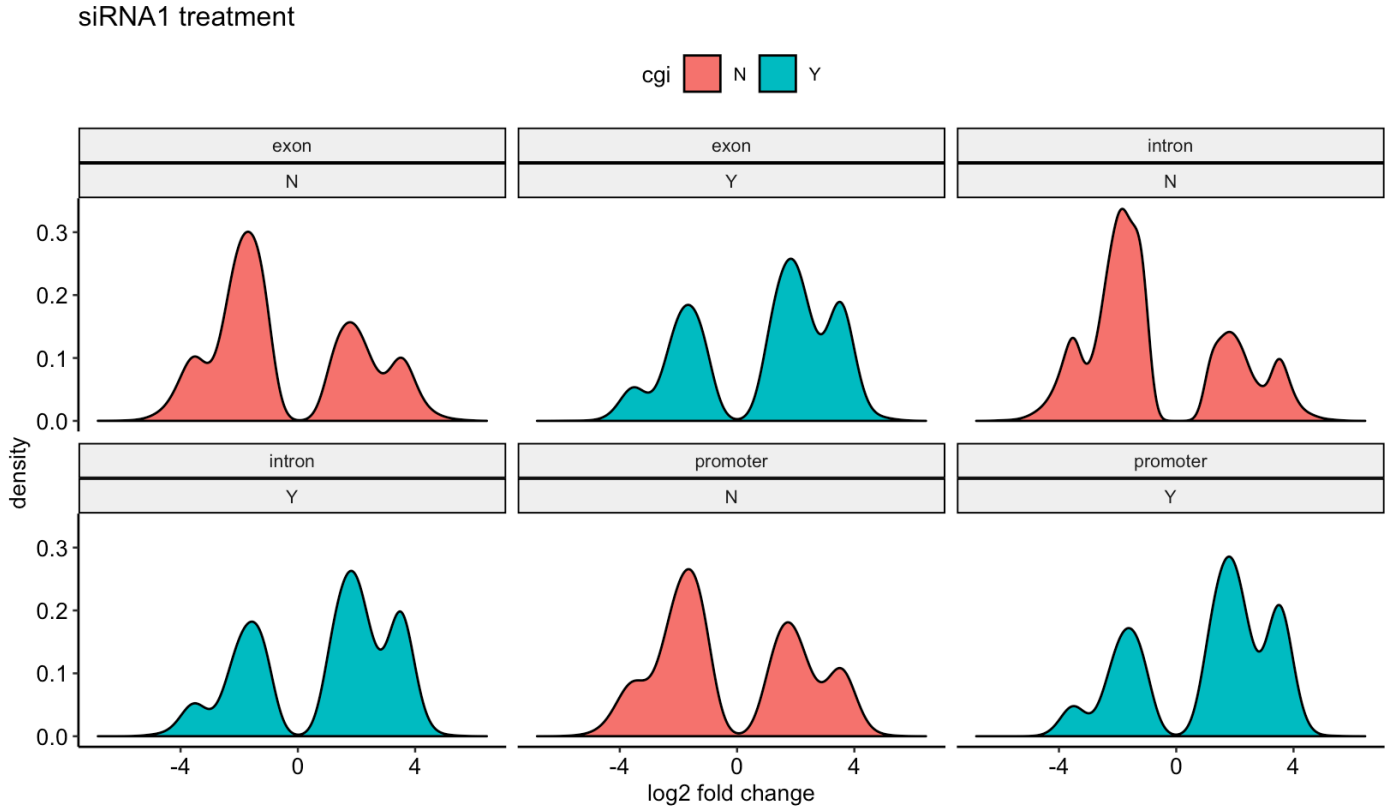
Effects of Tet2 kd on 5hMC levels in corneal epithelium



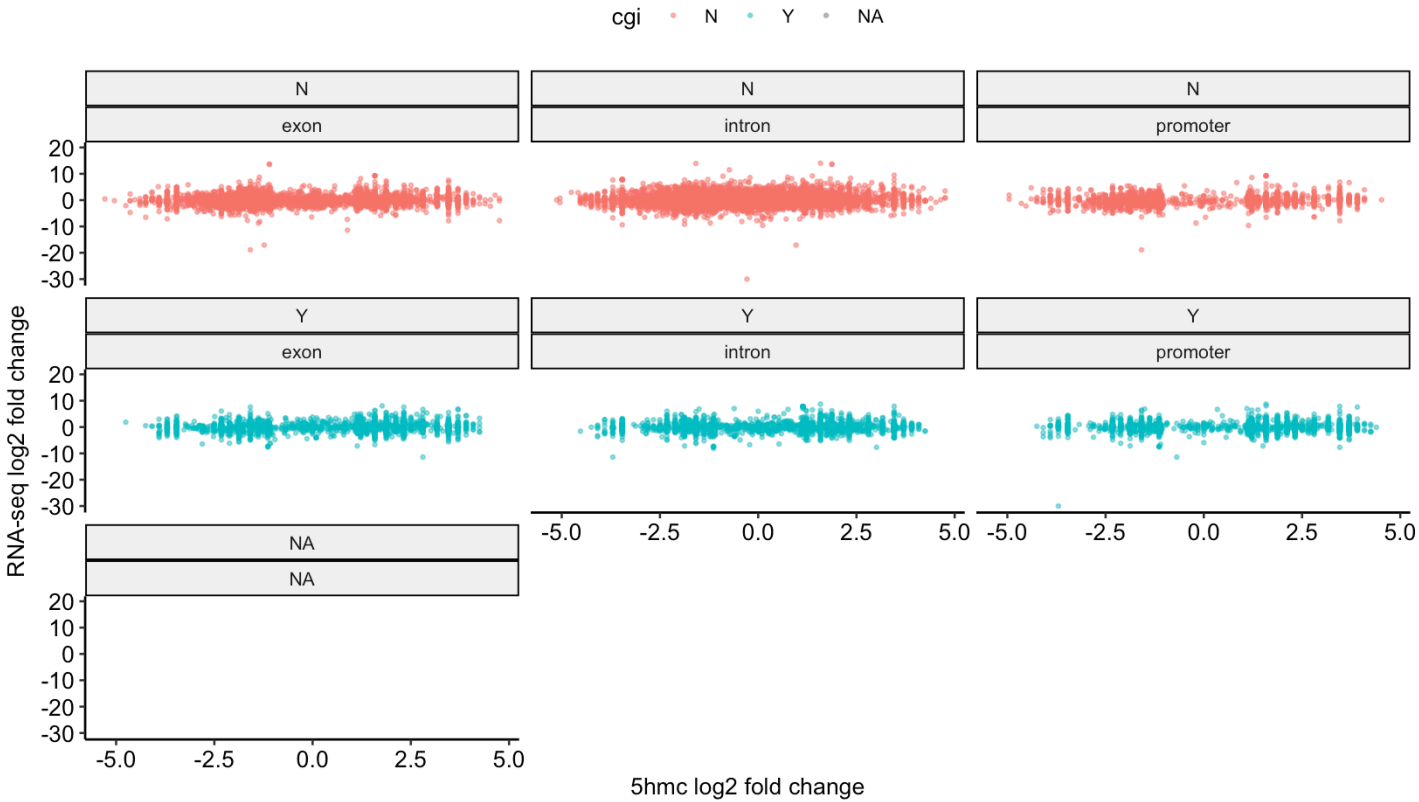
Differential Expression in Tet2 knockdowns



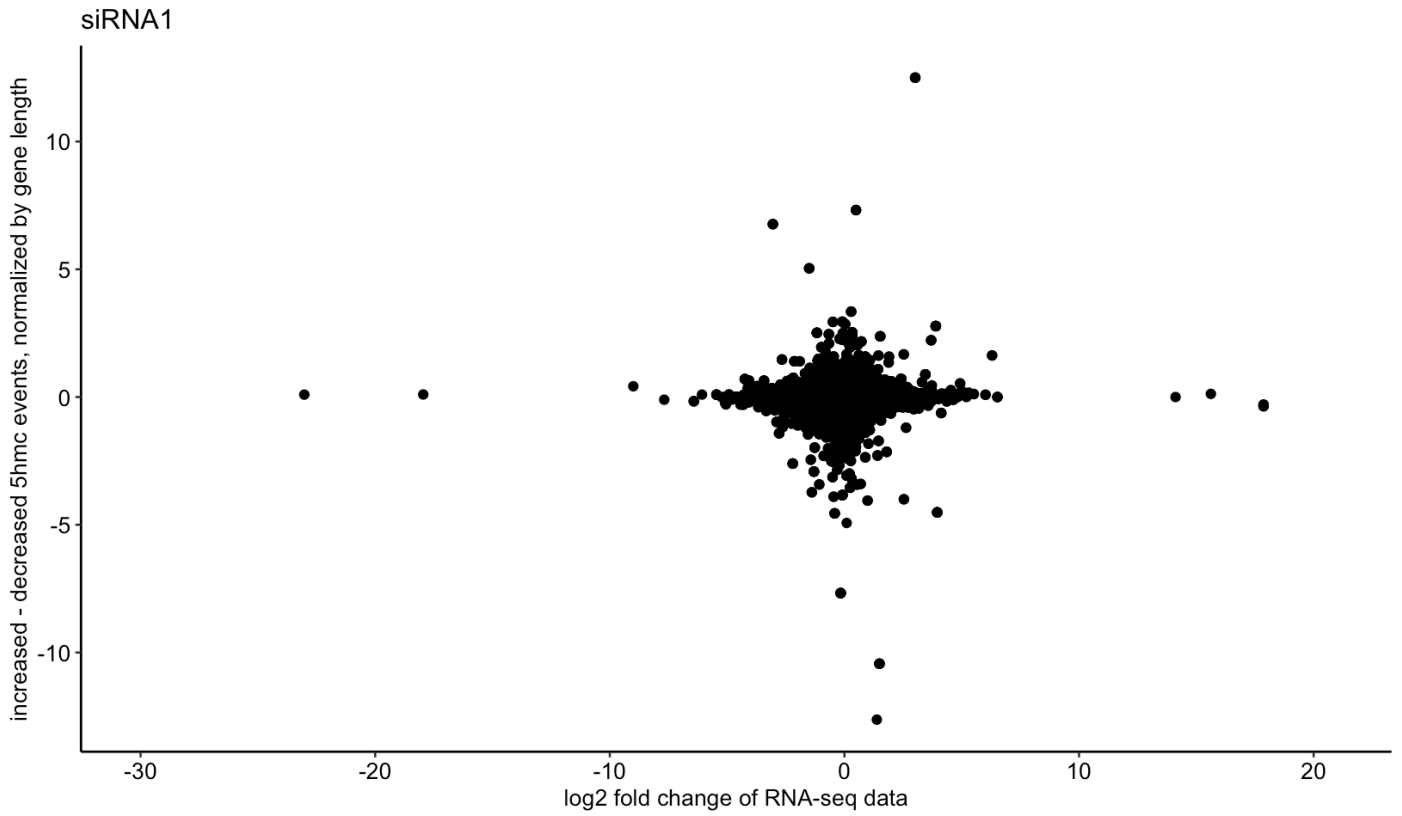
Genomic context of 5hMC changes



5hMC change versus gene expression changes



Regularized and bias adjusted 5hMC change versus expression change



…from the Zymo RRHP FAQ

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

Lorena Pantano

RRHP work

Harvard Stem Cell Institute

Frank Lab
Natasha Frank
Lynn Guo
Jiang Yusheng

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

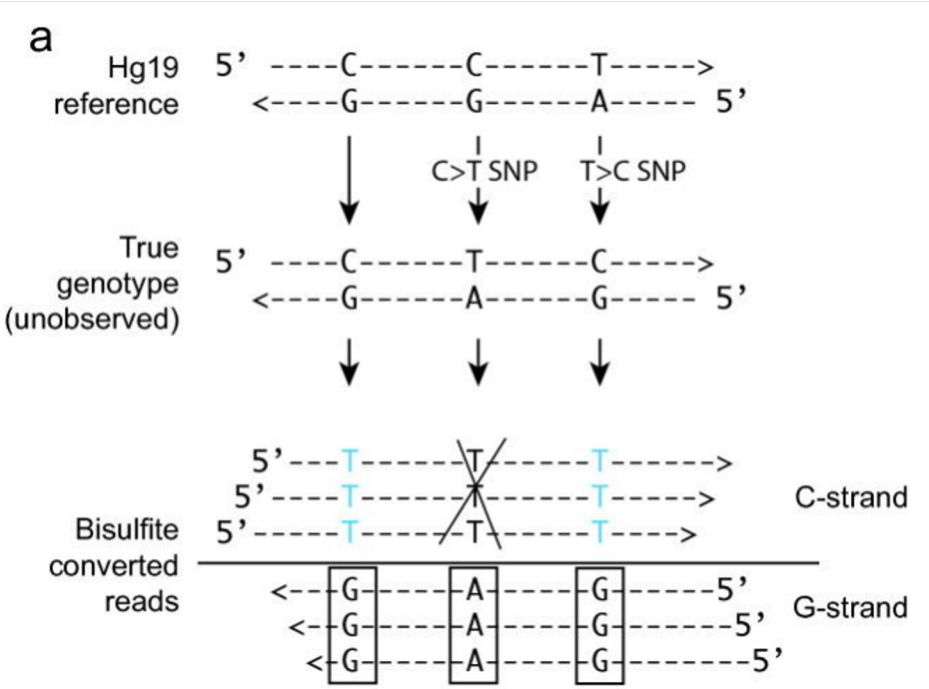
SNPs

- 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!)

BIS-SNP



DMRs versus DMLs

DMLs = Differentially Methylation Loci

Pros

- Won’t miss smaller signals
- Easy to measure

Cons

- High multiple testing correction penalty
  - There can be millions of hits
- Redundant information
- Role of single DML often unclear
- Can be a SNP artefact

DMRs = Differentially Methylated Regions

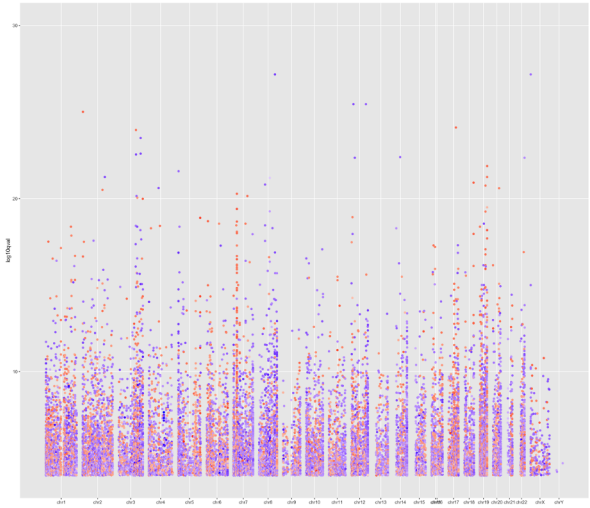
- Signals often cluster, particularly in CpG islands

Pros

- Low multiple testing penalty
- Easier to interpret
- Less likely to be artefact

Cons

- Can miss smaller signals
- How to measure?



DMR methods

Algorithmic

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

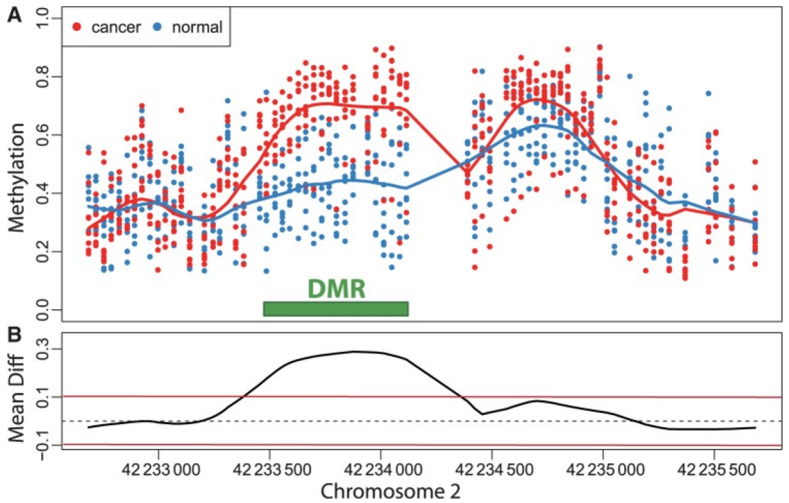
Statistical

Predefined regions and sliding windows - *methylKit*

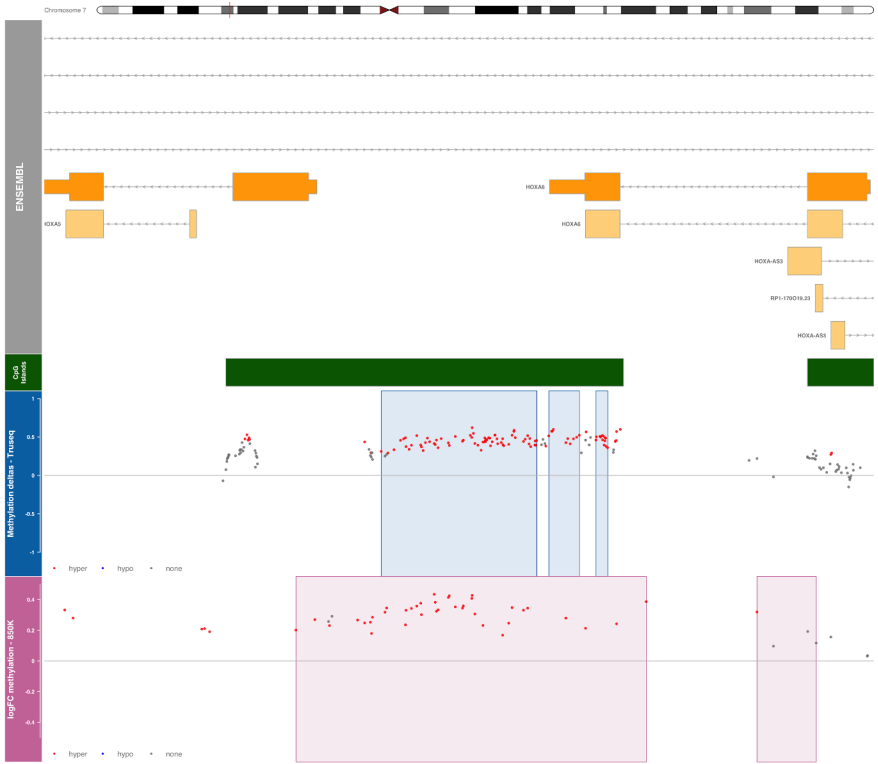
- summarize C and T counts within region/window

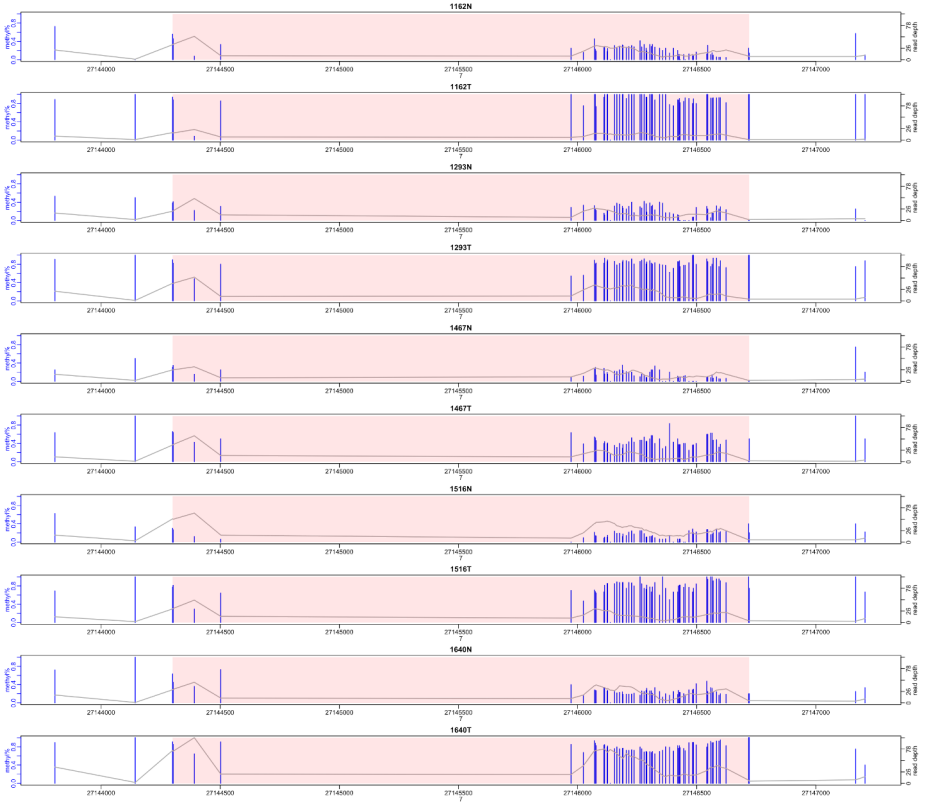
Smoothing - bumphunter (array) and Bsmooth (NGS)

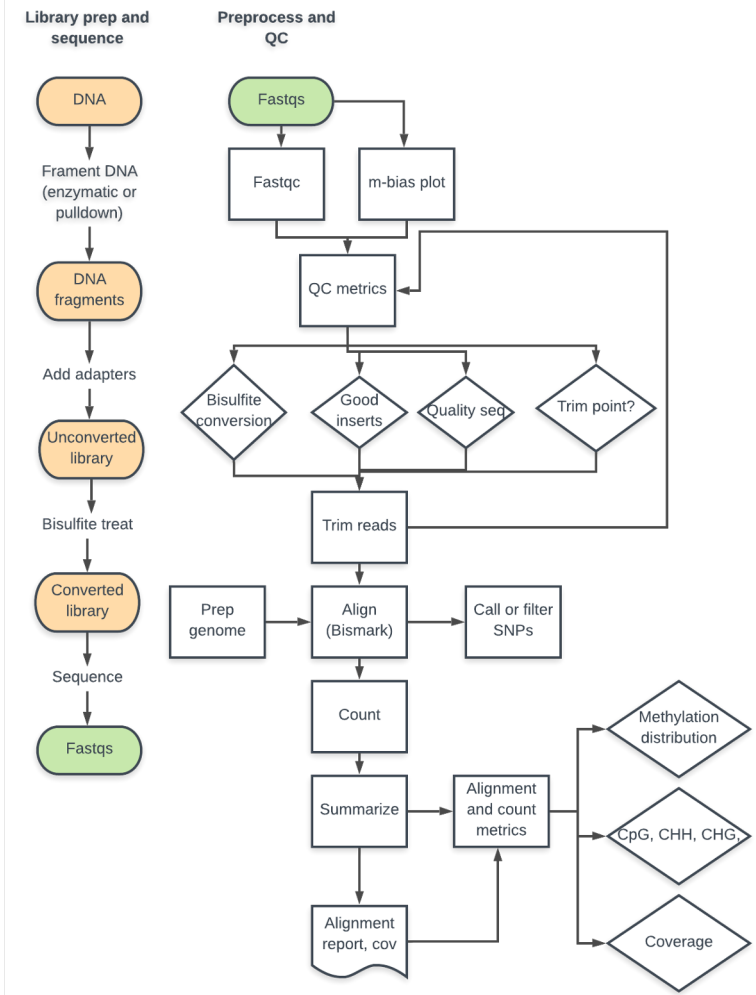
bumphunter



Some recent DMR results






Library prep and sequence
Preprocess and QC
DNA
Frament DNA (enzymatic or pulldown)
DNA fragments
Add adapters
Unconverted library
Bisulfite treat
Converted library
Sequence
Fastqs
Fastqs
Fastqc
m-bias plot
QC metrics
Bisulfite conversion
Good inserts
Quality seq
Trim point?
Trim reads
Prep genome
Align (Bismark)
Call or filter SNPs
Count
Summarize
Alignment and count metrics
Alignment report, cov
Methylation distribution
CpG, CHH, CHG,
Coverage

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

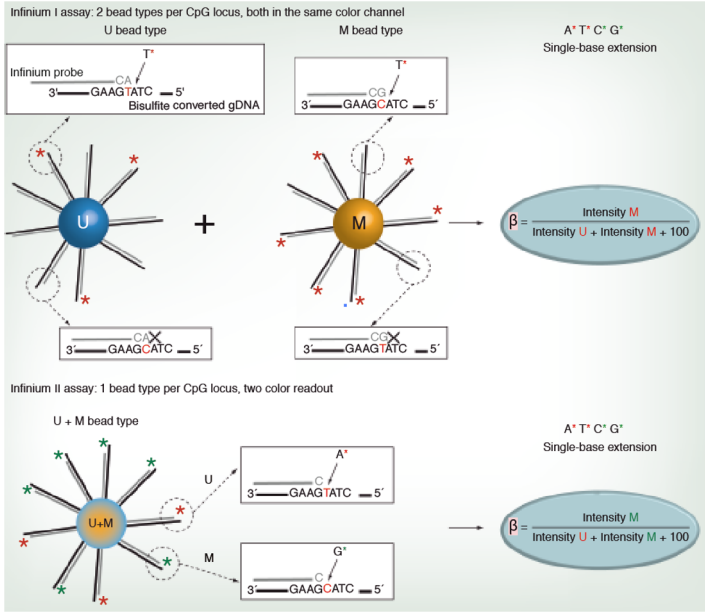
- Closest gene
- Changes to CpG islands in promoter

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 cystosine
            - 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**



CpGI focus
Sensitivity
All CpGs
Price
Commercial
mouse
Bioinfo support
Cell composition
Call SNPs
WGBS
RRBS
Truseq
450/850K

Bisulfite methods

Whole genome bisulfite sequencing

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

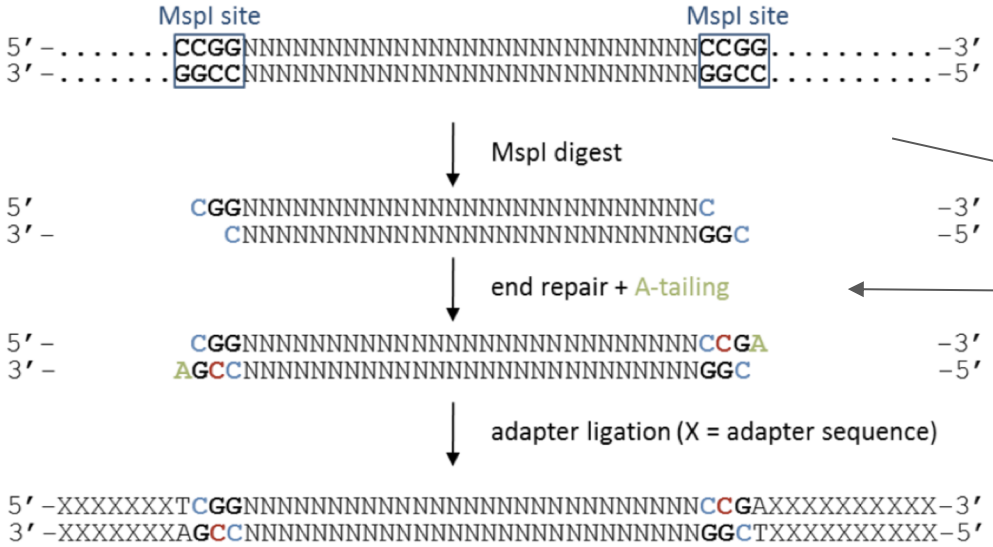
Targeted methods

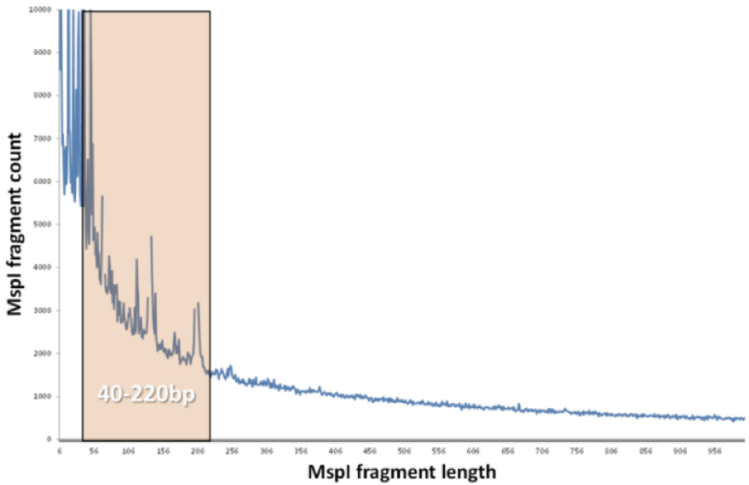
- Target CpG islands and other areas of interest

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

Capture based - Illumina TruSeq Methyl Capture EPIC (~15GB)

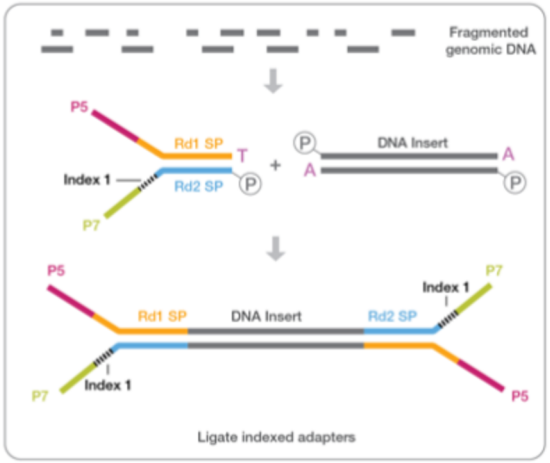
RRBS - Reduced Representation Bisulfite Sequencing

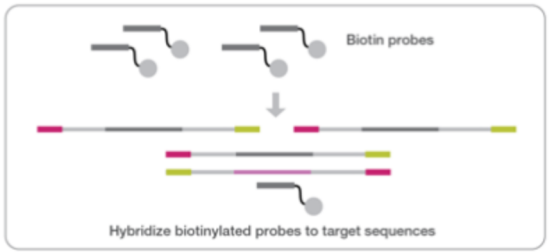


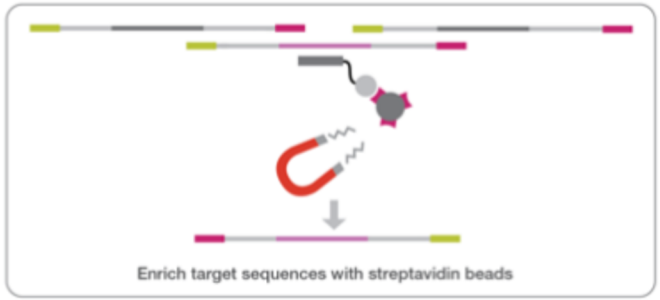


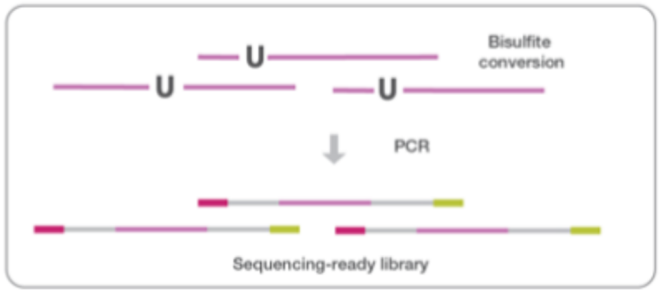
Single end (common) or paired end
low diversity

Illumina Truseq EPIC Methyl Capture









Paired end sequencing
Recommended 55 million reads = mean coverage >40X, max CpGs with >10X

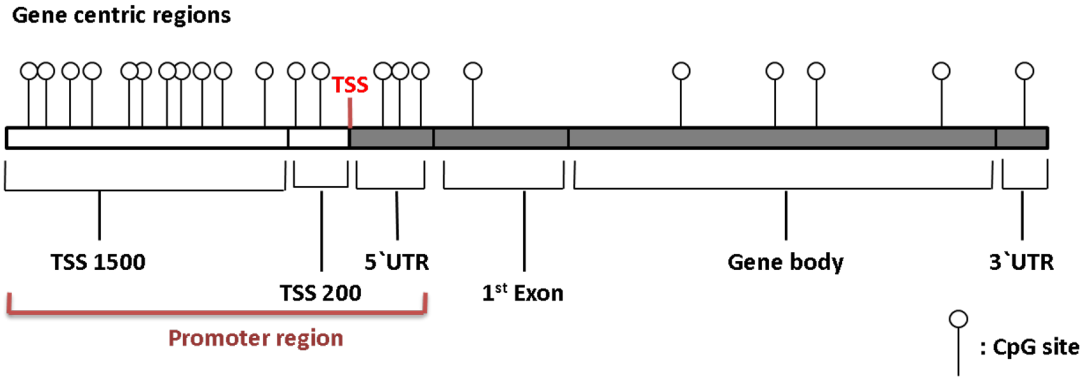
CpG islands - definition and some facts

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



Open sea

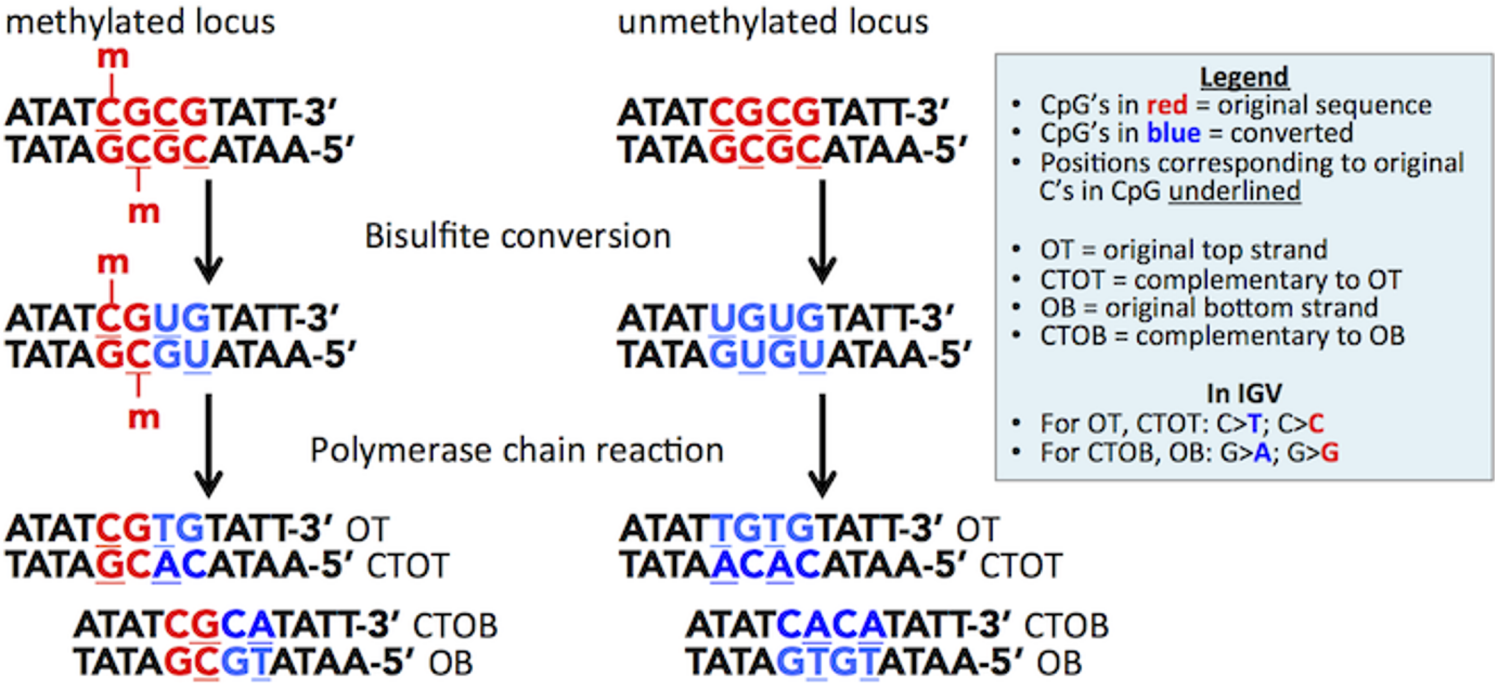
Open sea

- shores show variable tissue-specific DNA methylation patterns often altered in tumorigenesis

Alignment questions and issues

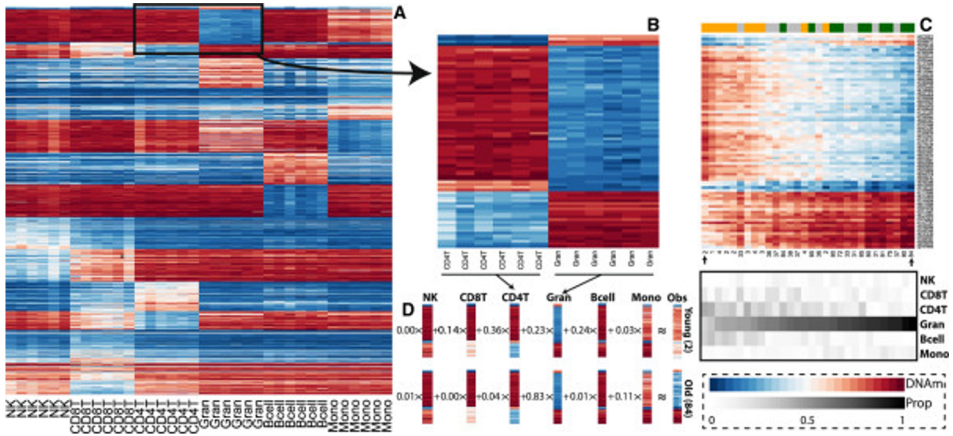
- Strand sequenced?
  - Capture methods may only capture one strand before bisulfite conversion
- Directional versus non-directional libraries?
  - Directional - only OT and OB post-bisulfite PCR strands sequenced
  - Non-directional - all 4 possible strands sequenced (OT, CTOT, OB, CTOB)
  - Typically 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


methylated locus
m
ATATCGCGTATT-3′
TATAGCGCATAA-5′
m
m
ATATCGUGTATT-3′
TATAGCGUATAA-5′
m
Bisulfite conversion
Polymerase chain reaction
ATATCGTGTATT-3′ OT
TATAGCACATAA-5′ CTOT
ATATCGCATATT-3′ CTOB
TATAGCGTATAA-5′ OB
unmethylated locus
ATATCGCGTATT-3′
TATAGCGCATAA-5′
ATATUGUGTATT-3′
TATAGUGUATAA-5′
ATATTGTGTATT-3′ OT
TATAACACATAA-5′ CTOT
ATATCACATATT-3′ CTOB
TATAGTGTATAA-5′ OB
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

Cell composition

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

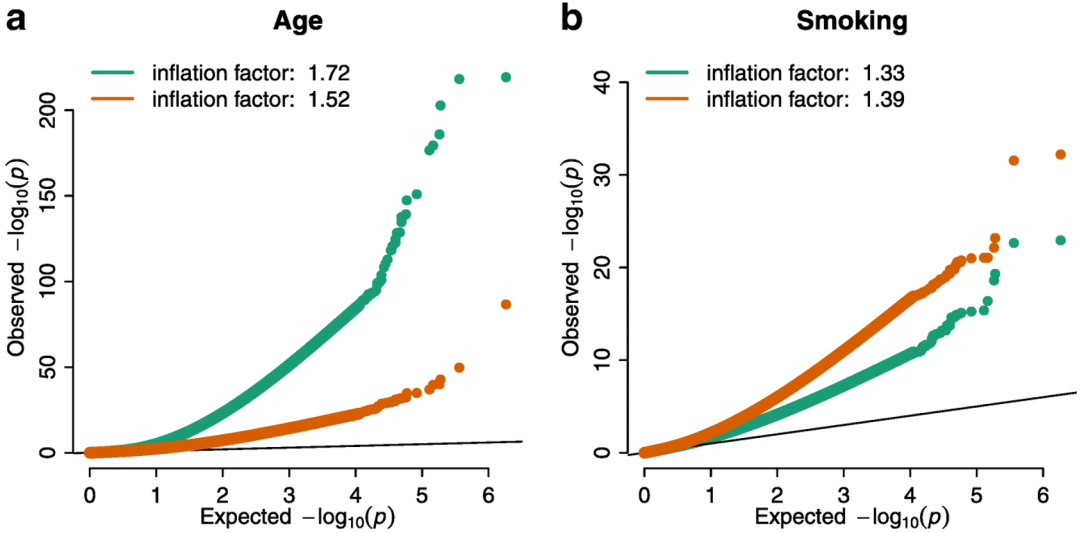


- Need to adjust for cellular composition effects
- Established methods to quantitate and adjust for cell composition effects exist for arrays
- Adapted but not established methods exist for RRBS
- For Truseq and WGBS, can only adjust (using SVA)

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



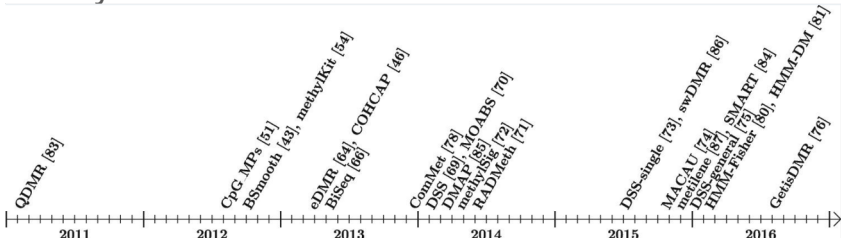
Statistical models for differential methylation

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**