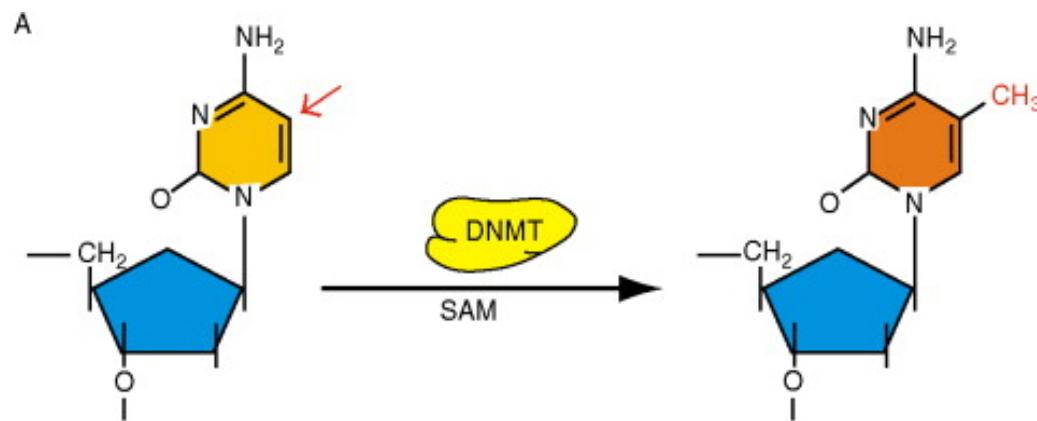


Bisulfite sequencing

DNA methylation

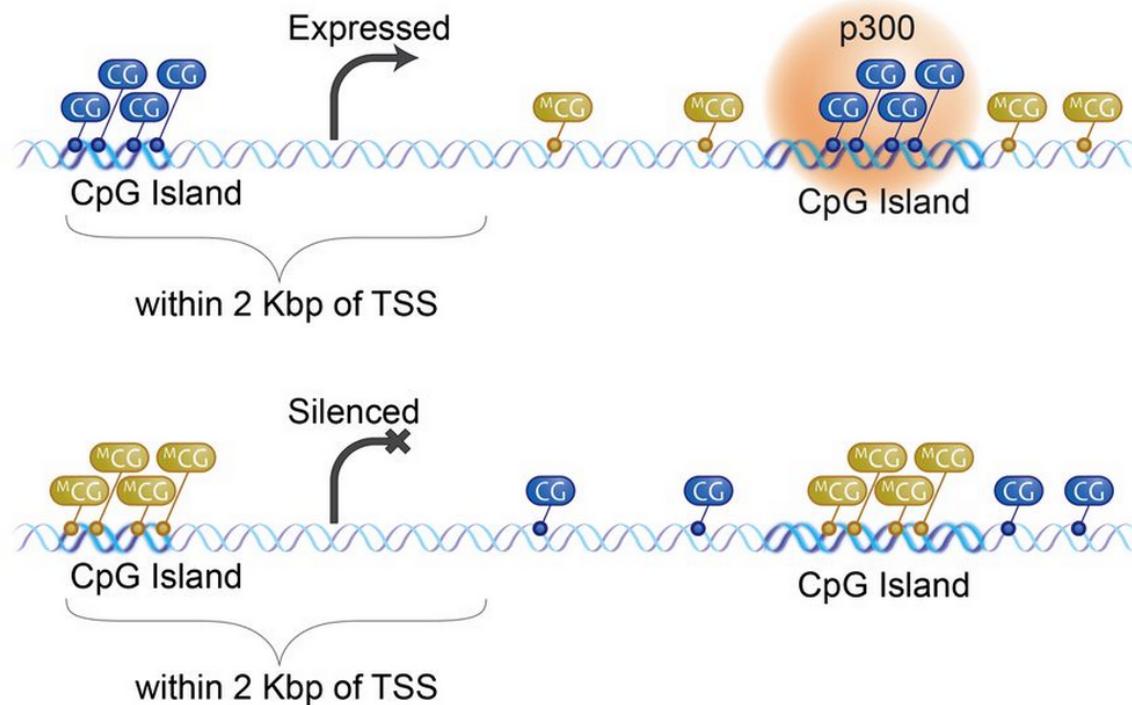
An epigenetic modification of the DNA sequence: adding a methyl group to the 5 position of cytosine (5mC)



Primarily happens at **CpG sites** (C followed by a G),
although non-CG methylation exists

DNA methylation

In human genome, >90% of CpG sites are fully methylated, except at CpG islands where methylation levels are typically low



Varley K E et al. Genome Res. 2013;23:555-567

Methylation of CpG islands in/near promoter region of gene can silence gene expression

Function of DNA methylation

- Important in gene regulation
 - Methylation of promoter regions can suppress gene expression
- Plays crucial role in development
 - Heritable during cell division
 - Helps cells establish identity during cell/tissue differentiation
- Can be influenced by environment
 - Good candidate to mediate GxE interactions

Sequencing approaches for DNA methylation

- Can be divided into two categories
 - Capture-based or enrichment-based sequencing
 - Use methyl-binding proteins or antibodies to capture methylated DNA fragments, then sequence fragments
 - **Resolution is low:** can typically quantify the amount of DNA methylation in 100-200 bp regions
 - Bisulfite-conversion-based sequencing
 - Bisulfite treatment converts unmethylated C's to T's
 - Sequencing converted data gives **single-bp resolution**
 - Can measure methylation status of each CpG site
 - Until recently, not possible to distinguish 5mC from 5hmC
- Focus of this lecture: **bisulfite sequencing**

Capture-based sequencing approaches

- All involve capture of methylated DNA followed by sequencing
- MeDIP-seq (Methylated DNA ImmunoPrecipitation)¹
 - Like ChIP-seq, but uses antibody against methylated DNA
 - Assesses relative rather than absolute methylation levels
 - Problem: don't observe unmethylated DNA fragments, only methylated ones
 - Another problem: immunoprecipitation may be affected by CpG density
 - MEDIPS² is a popular tool for analysis
- Capture via methyl-binding domain proteins: MBD-seq³/MIRA-seq⁴, methylCap-seq⁵
- Capture via methyl-sensitive restriction enzymes (MRE-seq)⁶

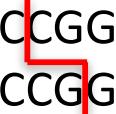
¹Weber et al. (2005) *Nat Genet*; ²Chavez et al. (2010) *Gen Res*; ³Serre et al. (2010) *NAR*

⁴Rauch et al. (2010) *Methods*; ⁵Brinkman et al. (2010) *Methods*; ⁶Maunakea et al. (2010) ₆*Nature*

Bisulfite sequencing (BS-seq)

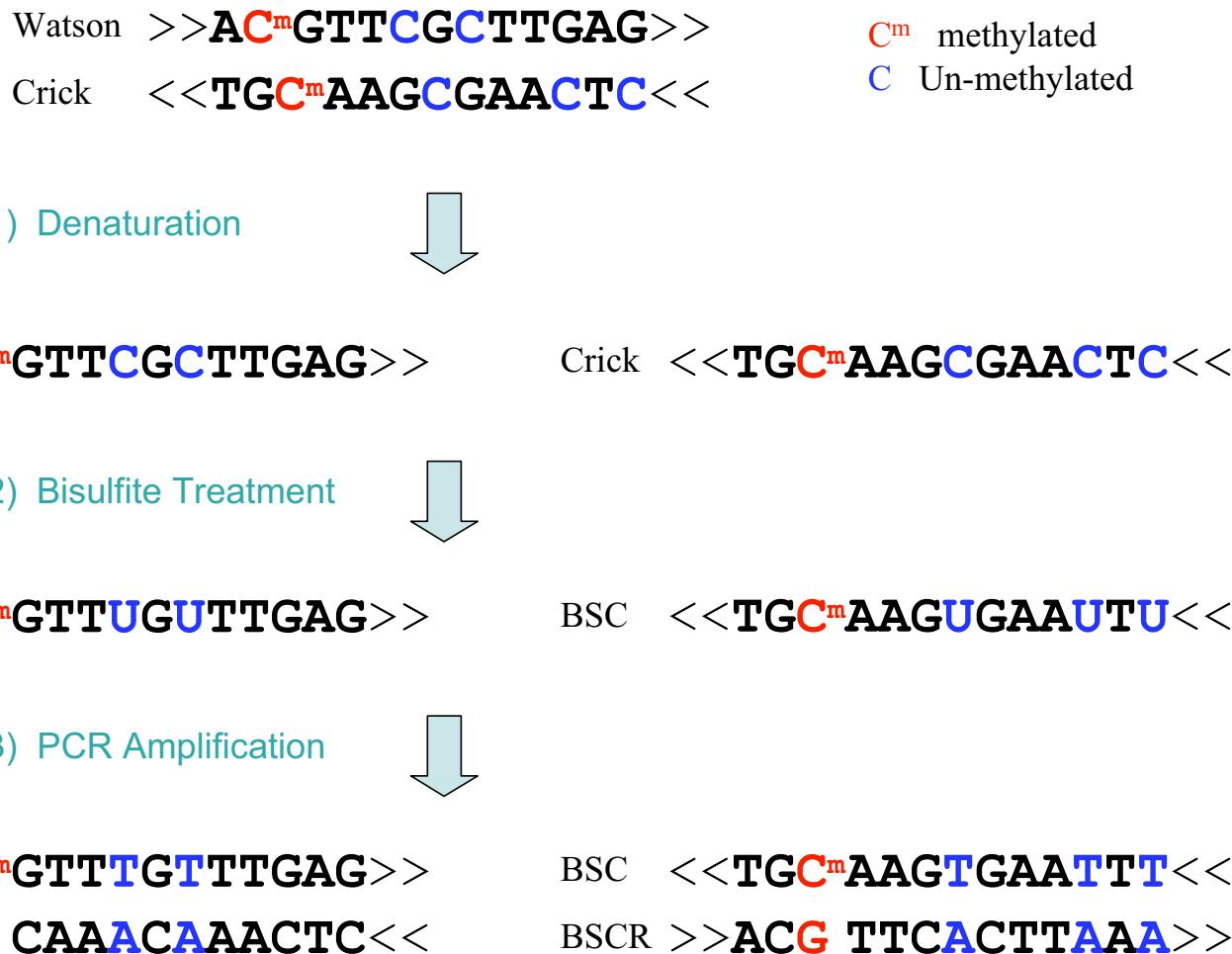
- Technology in a nutshell:
 - Treat fragmented DNA with bisulfite
 - Unmethylated C will be converted to U, amplified as T
 - Methylated C will be protected and remain C
 - No change for other bases
 - Amplify the treated DNA
 - Sequence the DNA segments
 - Align sequence reads to genome

Reduced representation bisulfite sequencing (RRBS)^{1,2}

- Goal: affordable alternative to genome-wide sequencing
 - By narrowing focus to CpG-rich areas, reduce # of reads necessary to obtain deep coverage of promoter regions
 - Interrogates ~1% of the genome but 5-10% of CpG sites
- Approach: enrich for CpG-rich segments of genome
 - Mspl restriction enzyme cuts at CpG sites, leaving fragments with CpGs at either end:

 - Size selection for fragments of 40-220bp maximizes coverage of promoter regions and CpG islands
 - Bisulfite treat, amplify, end-sequence, and align fragments to genome

¹Meissner (2005) *NAR*; ²Gu et al. (2011) *Nat Protoc*

Illustration of bisulfite conversion

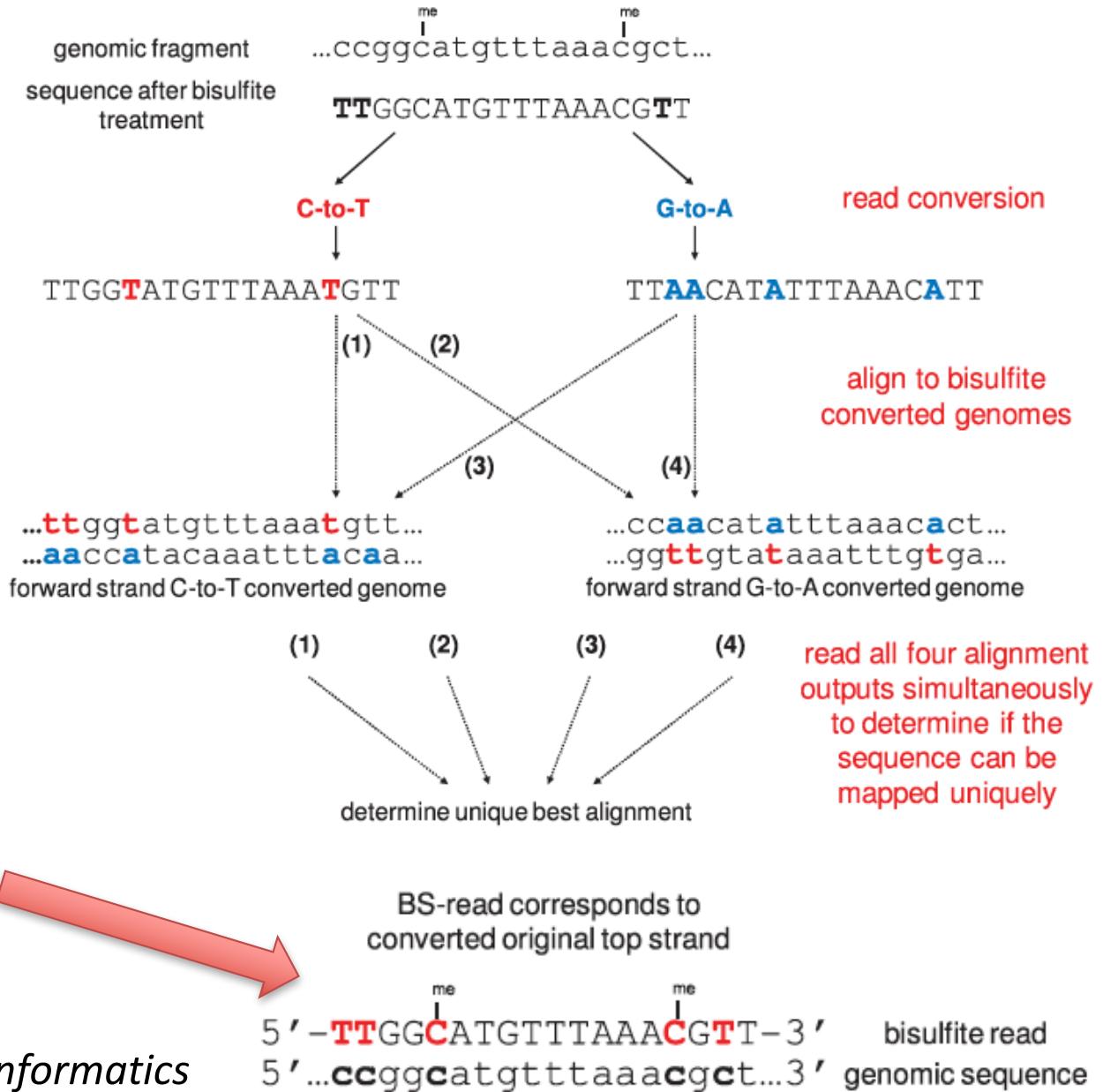


Alignment of BS-seq

- Problem: reads cannot be directly aligned to the reference genome.
 - Four different strands after bisulfite treatment and PCR
 - C-T mismatches will mean unmethylated reads can't be aligned to the correct position
 - Unmethylated CpGs will align with TpGs or likely not at all
 - Will lead to a strong bias in favor of methylated reads
- One possible solution *in silico* bisulfite conversion
 - Switch all C's to T's in both reads and reference sample
 - Use this for alignment, then change back to original

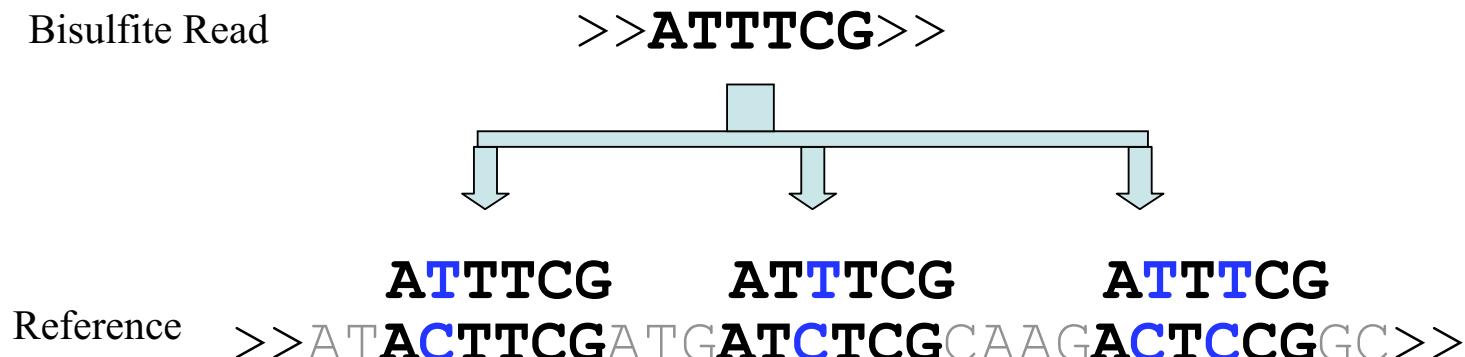
Strategy used by BISMARCK¹

- *In silico* bisulfite conversion of fragments **and** reference genome
 - Convert all C's to T's
 - Make complementary strand by converting all G's to A's
 - Align both strands to the four possible reference genomes
 - Choose best alignment
- Once aligned, convert back to original bases
- Compare to ref. genome to assess methylation



Alignment issues

- Possible problems with *in silico* approach
 - By converting all C's to T's, reduce sequence complexity to 3 bases
 - Larger search space for possible alignments
 - Could lead to mismatches or non-unique mapping



Strategy used by BSMAP¹

- Consider methylation status during alignment
 - create multiple versions of reference seed with C's converted to T's
 - compare each read to all possible seeds
 - do the same for complementary strand
- This approach reduces search space compared to *in silico* conversion of all C's to T's
 - T's in reads can match to C's or T's in reference
 - C's in reads can only match to C's in reference
- Computationally more intensive

Reference

>>**ACGTCGCT**TGATAGCT>>

Coordinate: 4875362

Seed Table

original seed

bisulfite seeds

	key	value
ACGTTCGCT	→	4875362, ...
ACGTTCGTT	→	4875362, ...
ACGTTTGCT	→	4875362, ...
ACGTTTGT	→	4875362, ...
ATGTCGCT	→	4875362, ...
ATGTCGTT	→	4875362, ...
ATGTTGCT	→	4875362, ...
ATGTTTGT	→	4875362, ...

Read >>**ATGTCGCT**TGAGAGCT>>

¹Xi and Li (2009) BMC Bioinformatics

Which alignment software is best?

- Advantages of BSMP:
 - reduces search space by eliminating mapping of C's to T's
 - greater proportion of uniquely mapping reads¹
- Advantages of BISMARK:
 - much faster than BSMP and other programs¹
 - uniqueness of mapping independent of methylation status¹
 - more user-friendly in terms of extracting data, interfacing with other software¹
- In general, BISMARK seems to be the popular choice

¹Chatterjee et al. (2012) *NAR*

Other aligners

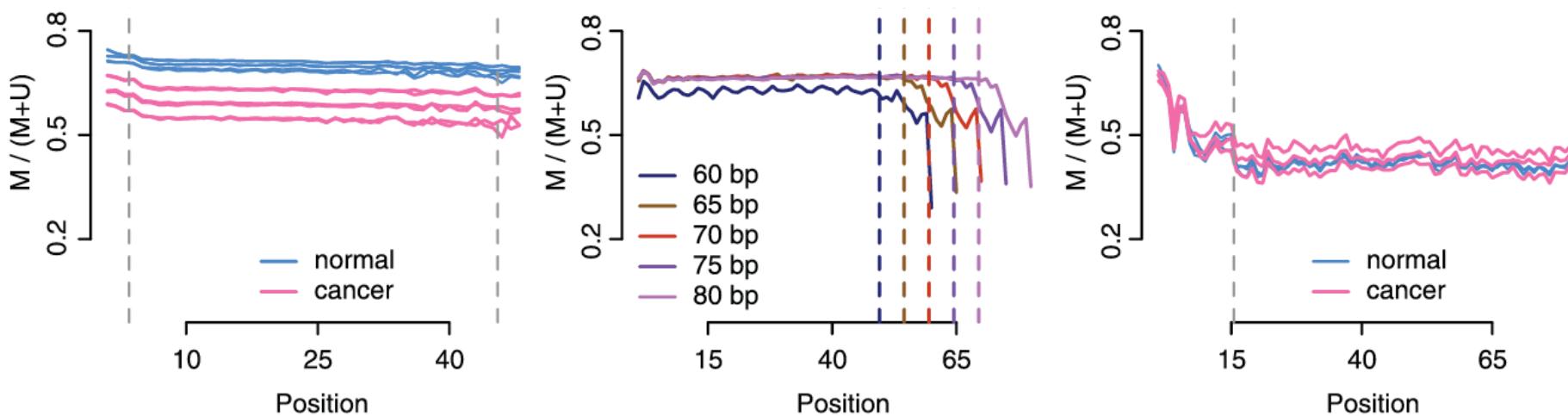
- Alignment of RRBS data
 - Chatterjee et al. notes it is much faster if we use information on Mspl cutpoints to “reduce” reference genome *in silico*¹
 - RRBSMAP: a version of BSMAP that does exactly that²
 - Has option to work with different restriction enzymes
- Many other aligners for bisulfite sequencing data
 - One useful review of these is Hackenberg et al.³

¹Chatterjee et al. (2012) NAR; ²Xi et al. (2012) *Bioinformatics*;

³Hackenberg et al. (2012): Chapter 2 in “DNA Methylation – From Genomics to Technology” Tatarinova (Ed.) <http://www.intechopen.com/books>

Another way to improve alignment

- Quality control of sequenced reads prior to alignment
- Issue: nucleotides towards the ends of reads can have greater rates of sequencing error
- Can assess this with M-bias plots post-alignment¹



- Solution: “trim” reads to remove less reliable sequence before aligning² (can also be done after alignment¹)

¹Hansen *et al.* 2012 *Genome Biology*; ²Chatterjee *et al.* (2012) *NAR*

What do the resulting data look like?

- Post-alignment, BS-seq data have a very simple form
- At each position, we have the total number of reads, and the methylated number of reads:

Position of CpG site	Total # reads	# methylated reads
chr1 3010874	22	18
chr1 3010894	31	27
chr1 3010922	12	10
chr1 3010957	7	6
chr1 3010971	6	6
chr1 3011025	7	5

Study design for BS-seq studies

- High costs → few samples typically analyzed
- Two common study designs
 - Analysis of a single sample:
 - Goal: observe methylation patterns across genome
 - Commonly done to **characterize methylome** for a particular cell type or species
 - Comparison of several samples:
 - Typical goal: compare methylation levels between groups
 - **Differential methylation analysis**
 - Compared with ChIP-seq and RNA-seq, methods are still in early stage, and are often *ad hoc*

Study design for BS-seq studies

- Because so few samples are involved in most studies, it is crucial to avoid all forms of heterogeneity
 - In large studies we can adjust for differences via covariates
 - With small N models often cannot accommodate covariates
- Heterogeneity = differences between samples other than variable of interest
 - Inadvertent differences in tissue sampled
 - Differences in cell type mixing proportions
 - Genetic differences between individuals
 - Age differences between samples
 - Different # of passages for cell lines

Avoiding heterogeneity

- Can avoid heterogeneity with careful study design
 - Stringent control of tissue dissection for tissue sampling
 - Analysis of homogeneous cell types whenever possible
 - Use of within-individual comparisons to avoid genetic and demographic differences
 - Example: paired tumor and normal samples from same patients
 - If not possible, match carefully for ethnicity, age, gender
 - Careful control of cell line experiments

Quality control of aligned BS-seq data

- Goal: remove sites likely to be low-quality or non-informative
 - Best filtering strategy will depend on study design and goals
- Filtering based on non-unique alignment
 - Will mostly happen naturally during alignment process
 - Post-alignment, CpG sites with unusually high read count are suspect
- Removal of sites with low coverage (often <5 or 10 total reads)
 - Appropriate cutoff will vary depending on analysis method used
 - For methods that model read count, can set cutoff lower
- Filtering based on lack of variability
 - If the goal is **differential methylation analysis**, remove sites with 0% of reads methylated in all samples, or 100% methylated in all samples
 - In contrast, if goal is to **characterize methylation patterns** in a particular genome, keep these sites!

Differential methylation analysis

- Typical goal: compare methylation levels between two groups
 - Example: tumor vs. normal tissue samples
 - Important: do groups contain biological replicates?
 - Some studies may compare 1 tumor to 1 normal sample
 - Other studies will include 2 or more replicates of each
- Popular *ad hoc* approaches for this comparison are Fisher's exact test and two-group t-test
- We will show why these can be problematic

Fisher's exact test with 2 samples

- If we have only one sample per group (no biological replicates), Fisher's exact test is a natural choice
- Example: single CpG site sequenced for 2 samples
 - For tumor sample, 32/44 methylated reads
 - For normal sample, 8/12 methylated reads
- Can then perform Fisher's exact test on the following table:

	Methylated	Unmeth.	Total reads
Tumor	32	12	44
Normal	8	4	12
Total	40	16	56

- OR = 1.33
- p = .73

Fisher's exact test in methylKit

- For comparisons between two samples, Fisher's exact test is a reasonable choice
 - Easy to carry out in R using `fisher.test()` function
 - Alternatively, methylKit¹ is a suite of R functions that facilitates analysis of genome-wide methylation data
 - Differential methylation analysis via either
 - Fisher's exact test (for comparisons between two samples)
 - Logistic regression based on methylation proportions
 - Analogous to two-group t-test, but with covariates
 - Can perform analysis in user-defined tiling windows
 - However, based on simple collapsing of information across sites rather than smoothing

¹Akalin *et al.* 2012 *Genome Biology*

Fisher's exact test with >2 samples

- For Fisher's exact test with biological replicates, need to collapse read information within groups
- Example: single CpG site sequenced for 4 samples
 - For 2 tumor samples, 32/44 and 4/10 methylated reads
 - For 2 normal samples, 8/12 and 12/34 methylated reads
- Could then perform Fisher's exact test on the following table:

		Methylated	Unmeth.	Total reads
	Tumor	36 = 32+4	18	54 = 44+10
	Normal	20 = 8+12	26	46 = 12+34
	Total	56	44	100

- OR = 2.6
- p = .0264

Problem with Fisher's exact test

- To perform Fisher's exact test for >2 samples, we have to collapse read information across samples within each group
- By doing this, we are ignoring information on biological variation between samples
 - **Biological variation:** natural variation in underlying fraction of DNA methylated between samples in the same condition
 - **Technical variation:** variation in estimation of methylation levels due to random sampling of DNA during sequencing¹
- By collapsing, we are assuming that:
 - samples within a group inherently have the same underlying fraction of DNA methylated
 - any variation between samples is due to technical variation

¹Hansen *et al.* 2012 *Genome Biology*

Naïve t-test

- Example: single CpG site sequenced for 4 samples
 - For 2 tumor samples, 32/44 and 4/10 methylated reads
 - For 2 normal samples, 8/12 and 12/34 methylated reads
- For t-test, compute a proportion for each sample
 - .727 and .400 for tumor samples
 - .667 and .353 for normal samples
- Difference in mean proportions = $.563 - .510 = .053$
- T-statistic = 0.2375
- $p = .834$

Problem with t-test

- To perform t-test, computed a proportion for each sample
 - Test inherently gives equal weight to each sample
 - Does not account for technical variation in proportion estimates
 - Recall: Technical variation = variation in estimation of methylation levels due to random sampling of DNA
 - Can expect this variation to be lower for samples with more reads
- One possible solution would be to incorporate weights based on read count
- However, another issue with this approach is the small number of samples
 - With N=4, the t-test has very little power due to low df

Fisher's exact vs. t-test

- The two tests yielded very different results
 - Fisher's exact p = .0264
 - T-test p = .834
- Main difference: unit of observation (reads vs. samples)
- Fisher's test was based on 100 “independent” reads
 - Reads are actually not independent if there is biological variation
 - Correlated within each sample, since samples have different methylation fractions
- T-test was based on 4 samples
 - Treated samples as equally informative, when really they are not
 - For 2 tumor samples, **32/44** and 4/10 methylated reads
 - For 2 normal samples, 8/12 and **12/34** methylated reads

Need better approaches

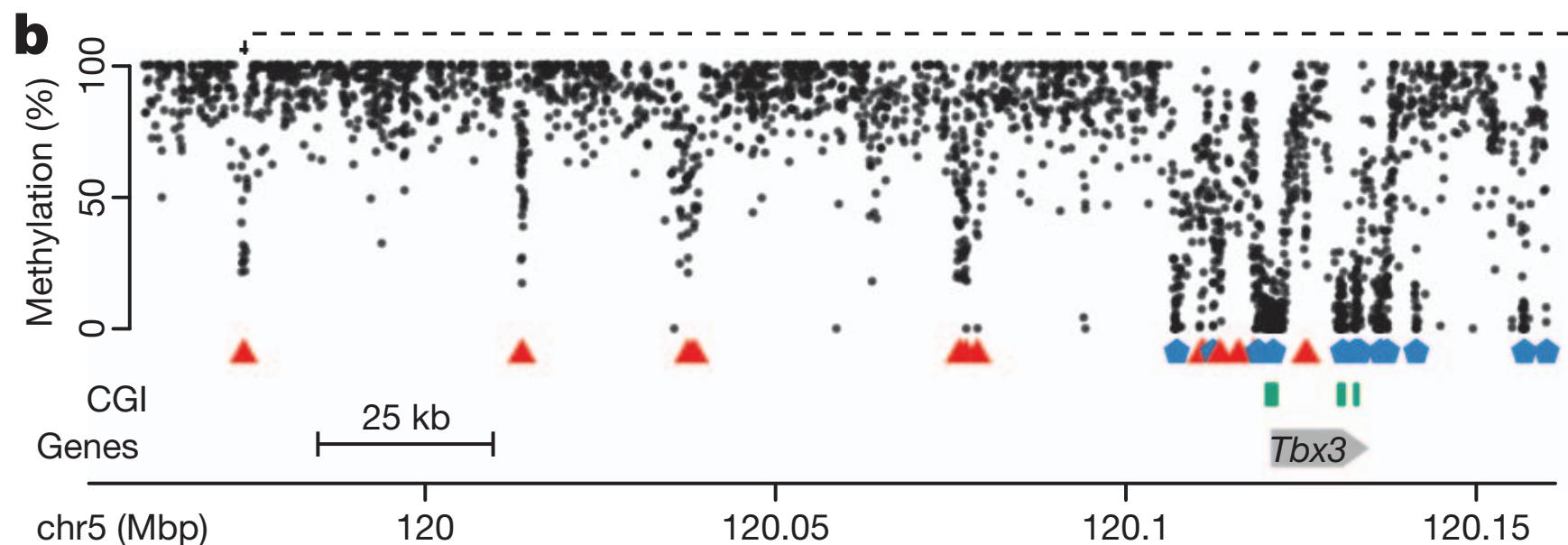
- Problem: want to test many sites with few samples
 - Limited information available at each site due to low # of samples
- Solution: approaches that borrow information across sites
 - Smoothing approaches that share information across nearby sites
 - Useful in single sample analyses that aim to **characterize the genome**
 - Useful for detecting **differential methylated regions (DMRs)** of the genome
 - Bayesian hierarchical model that borrows information across the genome
 - Useful for detecting **differentially methylated loci (DMLs)**

Smoothing approaches

- First consider **analysis of a single sample**
- Goal here is to identify methylated regions or loci:
 - Can estimate proportion of reads that are methylated at each C position, but:
 - Variability in estimation needs 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 methods using the estimated methylation proportion as input

HMM: Hidden Markov model

- Model switches between states along a chromosome
- Could model 3 methylation states: FMR, LMR, UMR
 - Stadler et al.¹ used estimated proportions to identify regions in mouse methylome corresponding to 3 states



¹Stadler et al. (2012) *Nature*

Smoothing sequencing data

- Problem with directly smoothing the proportions:
 - Doesn't consider the uncertainty in proportion estimates
 - Estimates more variable for CpG sites with low read counts
 - May want to put less weight on these estimates
- A better approach: BSmooth model¹
 - A local-likelihood smoothing approach
 - Key assumptions:
 - True methylation level π_j is a smooth curve of genomic coordinates.
 - The observed counts M_j follow a binomial(N_j, π_j) distribution.
 - Binomial assumption accounts for differences in variation for samples with different total read counts N_j

¹Hansen *et al.* 2012 *Genome Biology*

BSmooth smoothing

- Notation for CpG site j :
 - N_j, M_j : # total and # methylated reads
 - π_j : underlying true methylation level
 - l_j : location
- Model: $M_j \sim \text{Bin}(N_j, \pi_j)$
$$\log(\pi_j / (1 - \pi_j)) = \beta_0 + \beta_1 l_j + \beta_2 l_j^2$$
where β_0, β_1 , and β_2 vary smoothly along the genome.
- Fit this as a weighted generalized linear model (glm)
- Obtain a smoothed methylation estimate for each position along the genome using sliding window approach

Sliding window approach

- Choose window size (either distance or # CpG sites)
- For every genomic location l_j , use data in window surrounding l_j
- Fit weighted glm for all data in window, where weight for data point k depends inversely on:
 - the variance of estimated π_k , estimated as $\pi_k(1-\pi_k)/N_k$
 - distance of CpG site from window center $|l_k - l_j|$

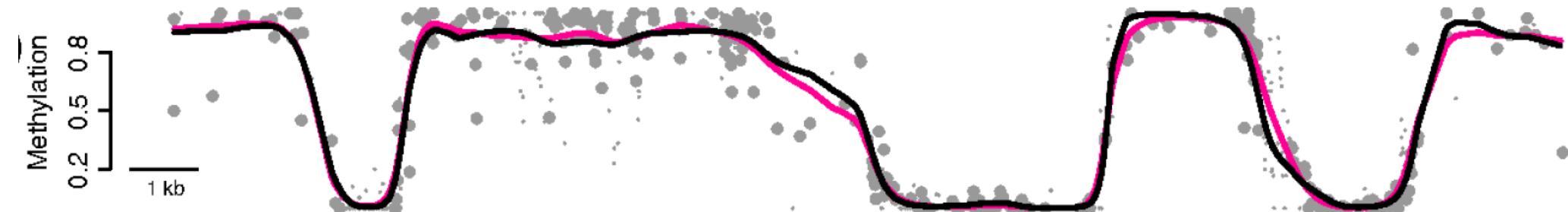
$$M_j \sim \text{Bin}(N_j, \pi_j)$$

$$\log(\pi_j / (1 - \pi_j)) = \beta_0 + \beta_1 l_j + \beta_2 l_j^2$$

- Estimation of β_0 , β_1 , and β_2 in window surrounding l_j provides estimate of π_j

Benefits of smoothing dense data

- By borrowing information across sites, can achieve high precision even with low coverage
 - Pink line is from smoothing full 30x data
 - Black line is from smoothing 5x version of data
 - Correlation = .90 across entire dataset
 - Median absolute difference of .056



Smoothed differential methylation analysis

- Goal: identify regions **differentially methylated** (DMRs) between groups
- BSmooth computes a t-test-like statistic
 - Signal-to-noise ratio based on smoothed data for multiple samples
 - Essentially the average difference between smoothed profiles from 2 groups, divided by estimated standard error
 - When biological replicates are included, this statistic correctly accounts for biological variation
- Identify DMRs as regions where this statistic exceeds some cutoff

Bsmooth functions implemented in Bioconductor package bsseq¹

- Functions for
 - Smoothing
 - Smoothed t-tests
 - DMR identification
 - Visualization of results
 - Fisher’s exact test (not smoothed)
- Can be implemented in parallel computing environment to speed up calculation

Use 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.

```
library(bsseq)
library(bsseqData)

## take chr21 on BS.cancer.ex to speed up calculation
data(BS.cancer.ex)
ix = which(seqnames(BS.cancer.ex)=="chr21")
BS.chr21 = BS.cancer.ex[ix,]

## use BSmooth to smooth and call DMR
BS.chr21 = BSmooth(BS.chr21) ## this takes 1-2 minutes

## perform t-test
BS.chr21.tstat = BSmooth.tstat(BS.chr21,
  c("C1","C2","C3"),c("N1","N2","N3"))

## call DMR
dmr.BSmooth <- dmrFinder(BS.chr21.tstat, cutoff = c(-4.6, 4.6))
```

Another approach: Bayesian hierarchical model¹

- Hierarchical model to separately model biological and technical variation
 - **Biological variation:** natural variation in underlying fraction of DNA methylated between samples in the same condition
 - **Technical variation:** variation in estimation of methylation levels due to random sampling of DNA during sequencing¹
 - Many methods only capture one or the other
 - Fisher's exact test: technical variation only
 - Naïve t-test: biological variation only
- Shrinkage approach allows us to borrow information about variation across genome
 - Especially useful when information per CpG site is limited by low number of samples

¹Feng *et al.* 2014 *Nucleic Acids Research*

Beta-binomial hierarchical model

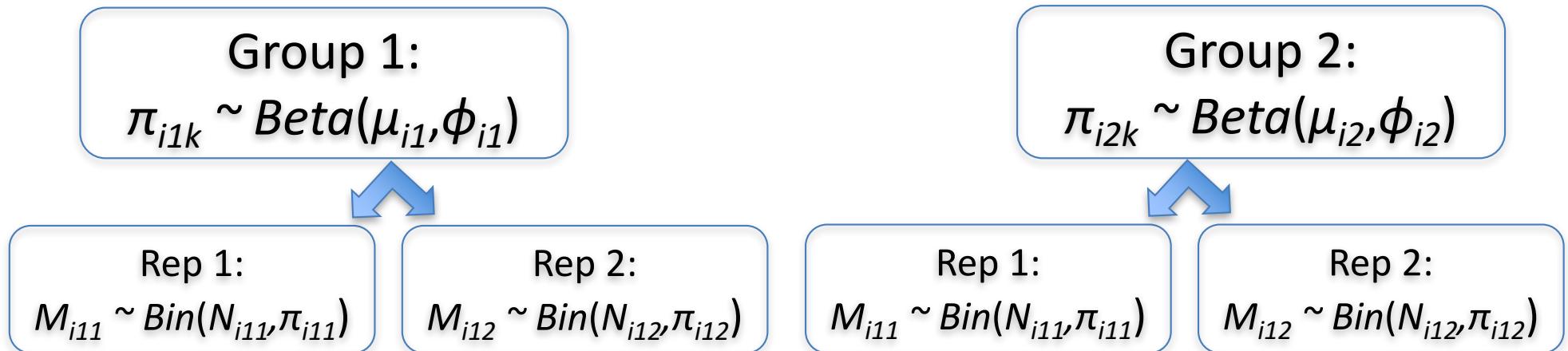
- “The most natural statistical model for replicated BS-seq DNA methylation measurements”¹
- Sampling of reads for each CpG site will follow a binomial distribution
 - Out of N reads covering a particular site, how many are methylated?
 - This number will follow a binomial(N, π) distribution
 - However, π may vary across replicates
- To model the biological variation of π across replicates, the beta distribution is a natural choice
- Beta-binomial distribution used to model methylated reads in DSS², BiSeq³, MOABS⁴, RADMeth⁵, MethylSig⁶

¹Robinson et al. 2014; ²Feng et al. 2014; ³Hebestreit et al. 2013; ⁴Sun et al. 2014;

⁵Dolzhenko & Smith 2014; ⁶Park et al. 2014

Beta-binomial hierarchical model

- Example: CpG site i , two groups $j=1$ (cancer) and 2 (normal), two replicates per group ($k = 1, 2$)



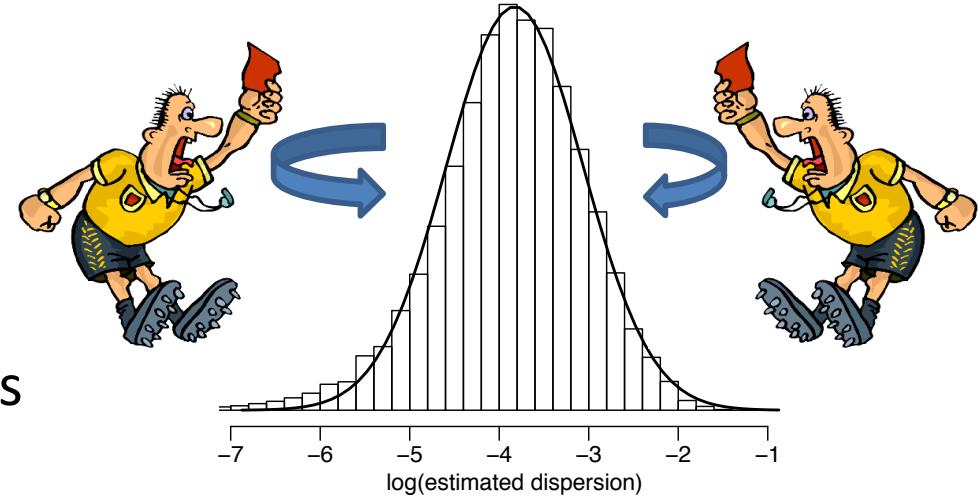
- Biological variation** modeled by dispersion parameter ϕ_{ij}
 - Replicates in each group may vary in true methylation proportion π_{ijk}
- Technical variation:** given N_{ijk} and π_{ijk} , number of methylated reads M_{ijk} varies due to random sampling of DNA
- Goal: test whether μ_{i1} and μ_{i2} are significantly different**

Motivation for shrinkage approach

- Hierarchical model:
$$M_{ijk} \sim Binomial(N_{ijk}, \pi_{ijk})$$
$$\pi_{ijk} \sim Beta(\mu_{ij}, \phi_{ij})$$
- **Goal: after correctly modeling different sources of variation, test whether μ_{i1} and μ_{i2} are significantly different at CpG i**
- Possible limitation of model: with small number of samples, estimation of parameters may be poor
 - In particular, difficult to accurately estimate dispersion ϕ_{ij} with only 2-3 replicates per group
 - Estimates may vary wildly due to small numbers
- Solution: borrow information from CpG sites across the genome to obtain reasonable estimates of ϕ_{ij}

Estimating dispersion parameter

- To obtain stable estimates of dispersion with few samples, we:
 - impose a log-normal prior on ϕ : $\phi_{ij} \sim \text{log normal}(m_j, r_j^2)$
 - use information from all CpGs in the genome to estimate the parameters m_j and r_j^2
- Choice of log-normal prior was motivated by distribution of dispersion in bisulfite sequencing data
 - RRBS data from mouse embryogenesis study
(Smith *et al.* 2012 *Nature*)
 - Estimation robust to departure from log-normality
 - Prior provides a good “referee”
 - Encourages dispersion estimates to stay within bounds



Wald test for DML, based on hierarchical model¹

- DML: Differentially Methylated Loci
 - Test for differential methylation at each CpG site
- At site i , test: $H_0 : \mu_{i1} = \mu_{i2}$
- Basic algorithm:
 - Use naïve estimates of ϕ across genome to estimate prior
 - For each site i , estimate μ_{i1} and μ_{i2} as proportion of methylated reads for each group
 - Bayesian estimation of ϕ_{ij} based on data and prior
 - Plug in estimates of μ_{ij} and ϕ_{ij} to create Wald statistic of form $t_i = \frac{\hat{\mu}_{i1} - \hat{\mu}_{i2}}{\sqrt{Var(\hat{\mu}_{i1} - \hat{\mu}_{i2})}}$

Using DSS to call DML and DMRs

- DSS can identify differentially methylated *loci* (DML) and *regions* (DMRs)
 - DML identified via Wald test, based on p-value threshold
 - DMRs called from DML based on user-specified criteria (region length, p-value and effect size thresholds)
- New features in DSS
 - Accommodates single-replicate studies by smoothing data from nearby CpG sites to form “pseudo-replicates”¹
 - Inclusion of design matrix to allow covariates and a more general experimental design²

¹Wu et al. *Nucleic Acids Research* 2015.

²Park et al. *Bioinformatics* 2016.

BS-seq experiment under general design

- General experimental design:
 - Multiple groups.
 - Multiple factors, crossed/nested.
 - Continuous covariates.
- Limited data analysis methods with not so good properties:
 - BiSeq and RADMeth, both based on generalized linear model (GLM).
 - Computationally demanding.
 - Numerically unstable.

DSS-general

- Suppose the input data include N CpG sites and D samples.
- Notations:
 - Y_{id}, m_{id} : methylated and total counts for i^{th} CpG and d^{th} data set.
 - π_{id}, ϕ_i : mean and dispersion.
 - \mathbf{X} : full ranked design matrix of dimension D by p .
- Counts are modeled by a beta-binomial regression:
$$Y_{id} \sim \text{beta-bin}(m_{id}, \pi_{id}, \phi_i)$$
$$g(\pi_{id}) = \mathbf{x}_d \boldsymbol{\beta}_i$$
- DML detection is achieved by a general hypothesis testing:
$$H_0 : C^T \boldsymbol{\beta}_i = 0, \text{ where } C \text{ is a p-vector.}$$

Our approach - approximation

- Beta-binomial regression.
- Transformation:
 - $g(Y/m)$ as response or data
 - What is $g(\cdot)$?
- Applying generalized (weighted) least square to estimate parameters, but with caution!

Choice of the link function

- arcsine link: $g(x) = \arcsin(2x - 1)$
- “Variance stabilization transformation” for binomial proportion:
 - Variance of the transformed data does not depend on mean (but on dispersion), so least square approach is possible.
 - logit or probit transformed data needs iterative procedure since variance depends on mean.
 - More linear than logit or probit, especially at the boundaries.

Parameter estimation

- Model: $Y_{id} \sim \text{beta-bin}(m_{id}, \pi_{id}, \phi_i)$

$$g(\pi_{id}) = \mathbf{x}_d \boldsymbol{\beta}_i$$

- Transformation:

$$Z_{id} = \arcsin(2Y_{id}/m_{id} - 1).$$

$$E[Z_{id}] \approx \arcsin(2E[Y_{id}]/m_{id} - 1) = \arcsin(2\pi_{id} - 1) = \mathbf{x}_d \boldsymbol{\beta}_i$$

$$\text{var}(Z_{id}) \approx \frac{1 + (m_{id} - 1)\phi_i}{m_{id}}.$$

$$V_i = \text{diag} \left(\frac{1 + (m_{id} - 1)\phi_i}{m_{id}} \right)$$

- Least square estimator:

$$\hat{\boldsymbol{\beta}}_i = (\mathbf{X}^T V_i^{-1} \mathbf{X})^{-1} \mathbf{X}^T V_i^{-1} Z.$$

Two-step estimation

- Dispersion estimation
 - Estimate $\hat{\beta}_i^{(0)}$ by setting dispersion to 0.
 - Estimate variance based on Pearson's chi-square statistics:
$$\chi_i^2 = \sum_d m_{id} (Z_{id} - \mathbf{x}_d \hat{\beta}_i^0)^2, \quad \hat{\sigma}_i^2 = \chi_i^2 / (D - p),$$
 - Dispersion can be derived as:
$$\hat{\phi}_i = \frac{D(\hat{\sigma}_i^2 - 1)}{\sum_d (m_{id} - 1)}.$$
 - Restriction: $1 < \hat{\sigma}_i^2 < \frac{\sum_d (m_{id} - 1)}{D} + 1$.
- Parameter estimation using GLS based on $\hat{\phi}_i$

Hypothesis testing

- For testing
 - Variance/covariance matrix estimates:
$$\hat{\Sigma}_i \equiv \widehat{\text{var}(\hat{\beta}_i)} = (\mathbf{X}^T \hat{V}_i^{-1} \mathbf{X})^{-1}.$$
 - Wald test statistics for $H_0 : C^T \beta_i = 0,$

$$t_i = \frac{C^T \hat{\beta}_i}{\sqrt{C^T \hat{\Sigma}_i C}}$$

Use DSS

- Input data object has the same format as `bsseq`.
- `DMLtest` performs Wald test at each CpG.
- `callDML`/`callDMR` calls DML or DMR.

```
## two group comparison
dmlTest <- DMLtest(BSobj, group1=c("C1", "C2", "C3"),
                     group2=c("N1", "N2", "N3"),
                     smoothing=TRUE, smoothing.span=500)
dmrs <- callDMR(dmlTest)
## A 2x2 design
DMLfit = DMLfit.multiFactor(RRBS, design, ~case+cell)
DMLtest = DMLtest.multiFactor(DMLfit, term="case")
```

Conclusions

- Analysis of genome-wide bisulfite sequencing data presents some unique challenges
 - Alignment of reads can be complicated
 - Many tests to be performed, but number of samples sequenced is limited by costs in most experiments
 - Beta-binomial model is widely used.

References

For software/analysis

- Akalin *et al.* 2012 *Genome Biology* 13:R87. **MethylKit paper**.
- Chatterjee *et al.* (2012) *Nucleic Acids Research*. 40(10):e79. **Compares aligners**.
- Chavez *et al.* (2010) *Genome Research* 20:1441-50. **MEDIPS software**.
- Dolzhenko and Smith (2014) *BMC Bioinformatics* 15:215. **RADMeth**.
- Feng, Conneely, and Wu (2014) *Nucleic Acids Research* 42(8):e69, **DSS for two-group**.
- Hansen *et al.* (2012) *Genome Biology* 13:R83. **Bsmooth paper**.
- Hebestreit, Dugas, and Klein (2013) *Bioinformatics* 29:1647-53. **BiSeq**.
- Krueger and Andrews (2011) *Bioinformatics* 27(11):1571-2. **BISMARK aligner**.
- Park *et al.* (2014) *Bioinformatics* 30:2414-22. **MethylSig**.
- Robinson *et al.* (2014) *Frontiers in Genetics* 5:324. **Review of methods for DML and DMR**
- Stadler *et al.* (2012) *Nature* 480:490-6. **Mouse methylome paper that used HMM**.
- Sun *et al.* (2014) *Genome Biology* 15:R38. **MOABS**.
- Wu *et al.* (2015) *Nucleic Acids Research*. 43(21):e141. **DSS-single for single replicates**.
- Park and Wu (2016) *Bioinformatics* 32 (10), 1446-1453. **DSS-general for general design**.
- Xi and Li (2009) *BMC Bioinformatics* 10:232. **BSMAP aligner**.
- Xi *et al.* (2012) *Bioinformatics* 28(3):430-2. **RRBSMAP aligner**.

References

For different sequencing technologies

- Bock et al. (2010) *Nat Biotech* 28(10):1106-16. **Compares RRBS, MeDIP-seq, others**
- Brinkman et al. (2010) *Methods* 52:232-236. **MethylCap-seq.**
- Gu et al. (2011) *Nat Protoc* 6(4):468-81. **Genome-wide RRBS protocol.**
- Maunakea et al. (2010) *Nature* 466:253-7. **MRE-seq.**
- Meissner (2005) *Nucleic Acids Research*. 33:5868-77. **Original RRBS paper.**
- Rauch et al. (2010) *Methods* 52:213-7. **MIRA-seq.**
- Serre et al. (2010) *Nucleic Acids Research*. 38:391-9. **MBD-seq.**
- Weber et al. (2005) *Nat Genet* 37:853-62. **Original MeDIP paper.**