University of Zurich[UZH]

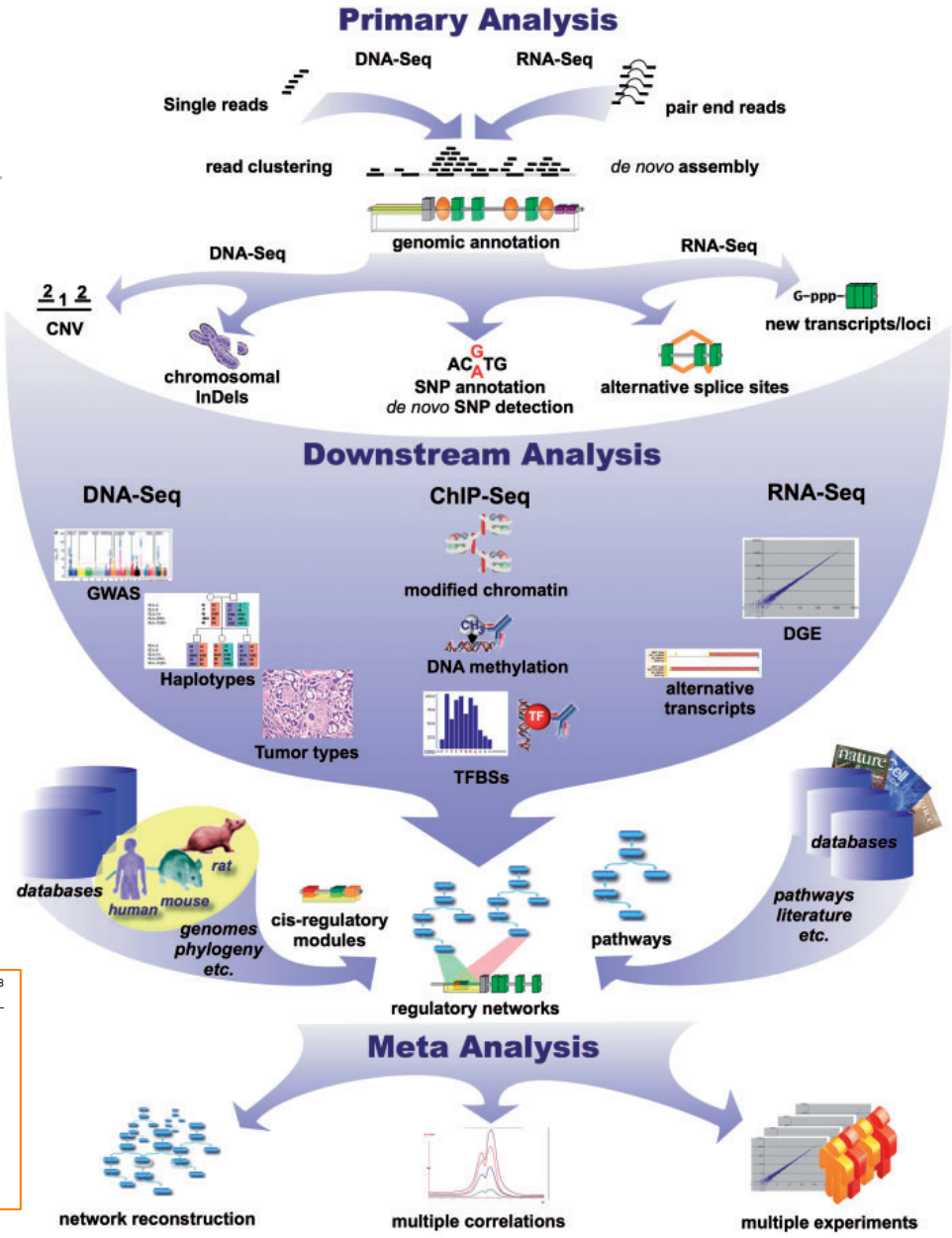
Institute of Molecular Life Sciences

# Epigenomics, DNA methylation methods

Mark D. Robinson, Institute of Molecular Life Sciences

- Three more lectures (+JCs):
  - DNA methylation
  - ChIP + genesets
  - single-cell analyses
- Epigenetics, CpGs, DNA methylation
- DNA methylation platforms
- DMR finding

# Where are we?




BRIEFINGS IN BIOINFORMATICS. VOL 11. NO 5. 499–511
Advance Access published on 25 May 2010
doi:10.1093/bib/bbq018

## Next generation sequencing in functional genomics

*Thomas Werner*

Submitted: 8th March 2010; Received (in revised form): 5th May 2010

# The expanding scope of DNA sequencing

Jay Shendure[1] & Erez Lieberman Aiden[2–4]

## Where are we?

| Method | Sequencing to determine: | Example reference | 'Subway' route as defined in Figure 3 |
|---|---|---|---|
| DNA-Seq | A genome sequence | 57 | Comparison, 'anatomic' (isolation by anatomic site), flow cytometery, DNA extraction, mechanical shearing, adaptor ligation, PCR and sequencing |
| Targeted DNA-Seq | A subset of a genome (for example, an exome) | 20 | Comparison, cell culture, DNA extraction, mechanical shearing, adaptor ligation, PCR, hybridization capture, PCR and sequencing |
| Methyl-Seq | Sites of DNA methylation, genome-wide | 34 | Perturbation, genetic manipulation, cell culture, DNA extraction, mechanical shearing, adaptor ligation, bisulfite conversion, PCR and sequencing |
| Targeted methyl-Seq | DNA methylation in a subset of the genome | 129 | Comparison, cell culture, DNA extraction, bisulfite conversion, molecular inversion probe capture, circularization, PCR and sequencing |
| DNase-Seq, Sono-Seq and FAIRE-Seq | Active regulatory chromatin (that is, nucleosome-depleted) | 113 | Perturbation, cell culture, nucleus extraction, DNase I digestion, DNA extraction, adaptor ligation, PCR and sequencing |
| MAINE-Seq | Histone-bound DNA (nucleosome positioning) | 130 | Comparison, cell culture, MNase I digestion, DNA extraction, adaptor ligation, PCR and sequencing |
| ChIP-Seq | Protein-DNA interactions (using chromatin immunoprecipitation) | 131 | Comparison, 'anatomic', cell culture, cross-linking, mechanical shearing, immunoprecipitation, DNA extraction, adaptor ligation, PCR and sequencing |
| RIP-Seq, CLIP-Seq, HITS-CLIP | Protein-RNA interactions | 46 | Variation, cross-linking, 'anatomic', RNase digestion, immunoprecipitation, RNA extraction, adaptor ligation, reverse transcription, PCR and sequencing |
| RNA-Seq | RNA (that is, the transcriptome) | 39 | Comparison, 'anatomic', RNA extraction, poly(A) selection, chemical fragmentation, reverse transcription, second-strand synthesis, adaptor ligation, PCR and sequencing |
| FRT-Seq | Amplification-free, strand-specific transcriptome sequencing | 119 | Comparison, 'anatomic', RNA extraction, poly(A) selection, chemical fragmentation, adaptor ligation, reverse transcription and sequencing |
| NET-Seq | Nascent transcription | 41 | Perturbation, genetic manipulation, cell culture, immunoprecipitation, RNA extraction, adaptor ligation, reverse transcription, circularization, PCR and sequencing |
| Hi-C | Three-dimensional genome structure | 71 | Comparison, cell culture, cross-linking, proximity ligation, mechanical shearing, affinity purification, adaptor ligation, PCR and sequencing |
| Chia-PET | Long-range interactions mediated by a protein | 73 | Perturbation, cell culture, cross-linking, mechanical shearing, immunoprecipitation, proximity ligation, affinity purification, adaptor ligation, PCR and sequencing |
| Ribo-Seq | Ribosome-protected mRNA fragments (that is, active translation) | 48 | Comparison, cell culture, RNase digestion, ribosome purification, RNA extraction, adaptor ligation, reverse transcription, rRNA depletion, circularization, PCR and sequencing |
| TRAP | Genetically targeted purification of polysomal mRNAs | 132 | Comparison, genetic manipulation, 'anatomic', cross-linking, affinity purification, RNA extraction, poly(A) selection, reverse transcription, second-strand synthesis, adaptor ligation, PCR and sequencing |
| PARS | Parallel analysis of RNA structure | 42 | Comparison, cell culture, RNA extraction, poly(A) selection, RNase digestion, chemical fragmentation, adaptor ligation, reverse transcription, PCR and sequencing |
| Synthetic saturation mutagenesis | Functional consequences of genetic variation | 93 | Variation, genetic manipulation, barcoding, RNA extraction, reverse transcription, PCR and sequencing |
| Immuno-Seq | The B-cell and T-cell repertoires | 86 | Perturbation, 'anatomic', DNA extraction, PCR and sequencing |
| Deep protein mutagenesis | Protein binding activity of synthetic peptide libraries or variants | 95 | Variation, genetic manipulation, phage display, *in vitro* competitive binding, DNA extraction, PCR and sequencing |
| PhIT-Seq | Relative fitness of cells containing disruptive insertions in diverse genes | 92 | Variation, genetic manipulation, cell culture, competitive growth, linear amplification, adaptor ligation, PCR and sequencing |

FAIRE-seq, formaldehyde-assisted isolation of regulatory elements–sequencing. MAINE-Seq, MNase-assisted isolation of nucleosomes-sequencing; RIP-Seq, RNA-binding protein immunoprecipitation-sequencing; CLIP-Seq, cross-linking immunoprecipitation-sequencing; HITS-CLIP, high-throughput sequencing of RNA isolated by cross-linking immunoprecipitation; FRT-Seq, on-flowcell reverse transcription–sequencing. NET-Seq, native elongating transcript sequencing. TRAP, translating ribosome affinity purification. PhIT-Seq, phenotypic interrogation via tag sequencing.

# Epigenetics definition

Epi - "on top of" or "in addition to"

“Epigenetics”:

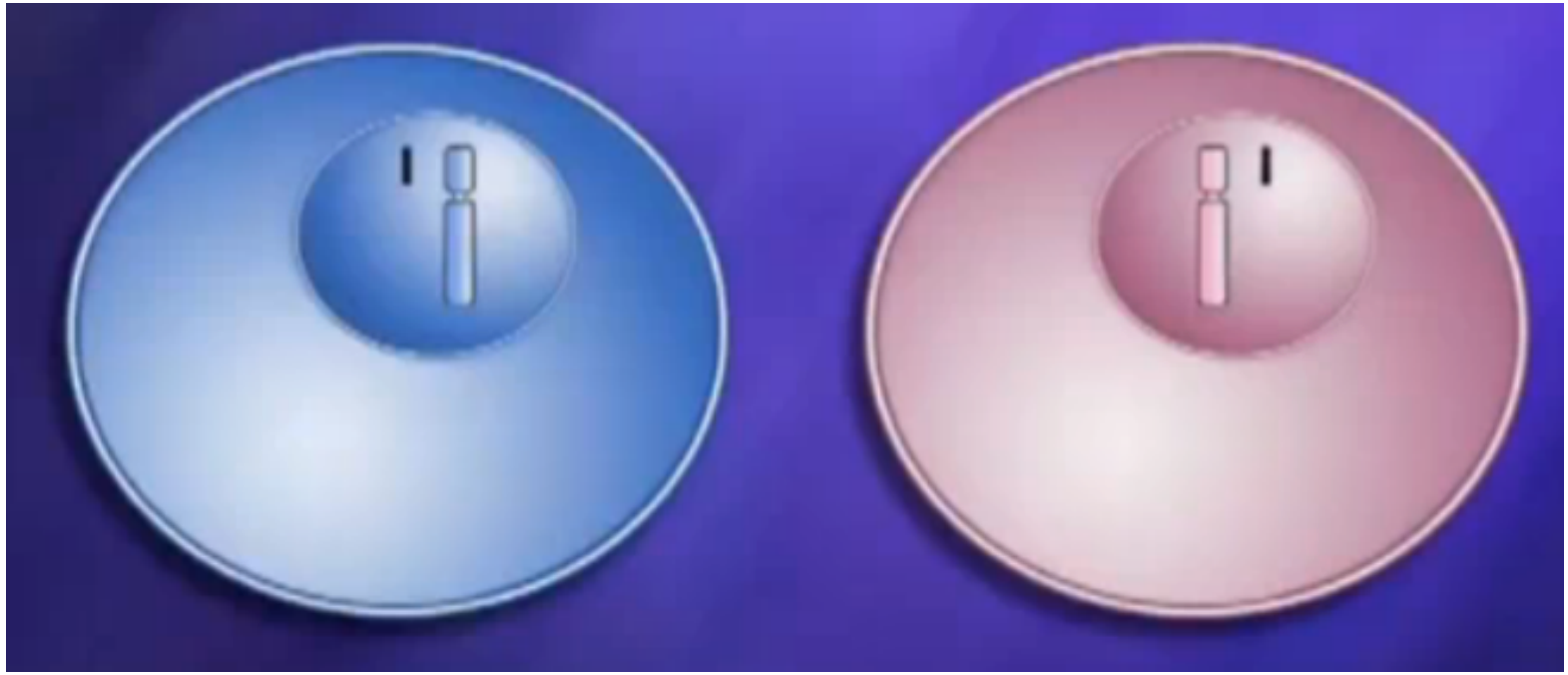
- **heritable alterations in gene expression caused by mechanisms other than changes in DNA sequence.**
- the study of the mechanisms of temporal and spatial control of gene activity during the development of complex organisms

# Example of epigentics: X-inactivation

Females have 2 **X**-chromosomes, but one of them is (mostly) silenced. In early embryogenesis, either the maternal or paternal allele is silenced at random, but any subsequent cell divisions will maintain the silenced X. For example, calico coat colour is determined by an X-inactivation outcome (gene is on the X-chromosome).

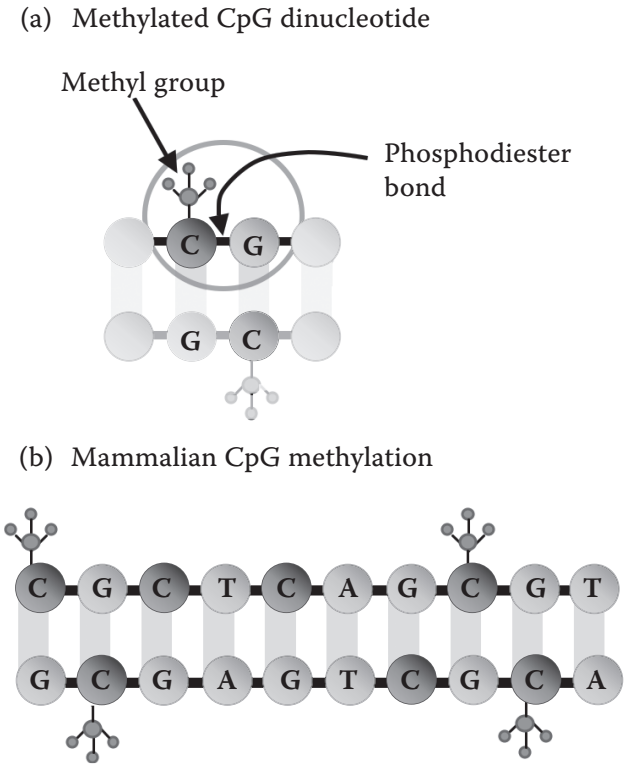
https://www.youtube.com/watch?v=n330FzHpI90

# X-inactivation



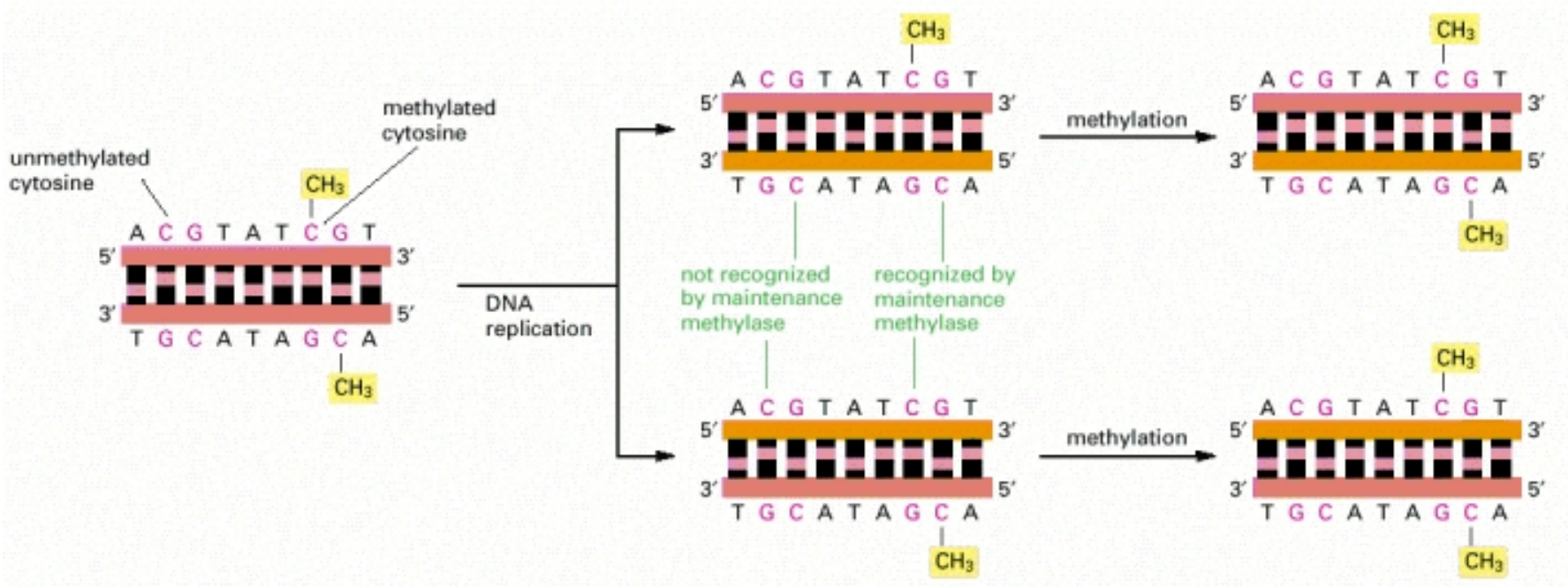
One of the X chromosomes is randomly silenced.
Subsequent cell divisions maintain the same silenced X.

# DNA methylation




Covalent addition of methyl group ($CH_3$) to cytosine (almost exclusively at CpG sites in mammals); **binary status** at individual sites

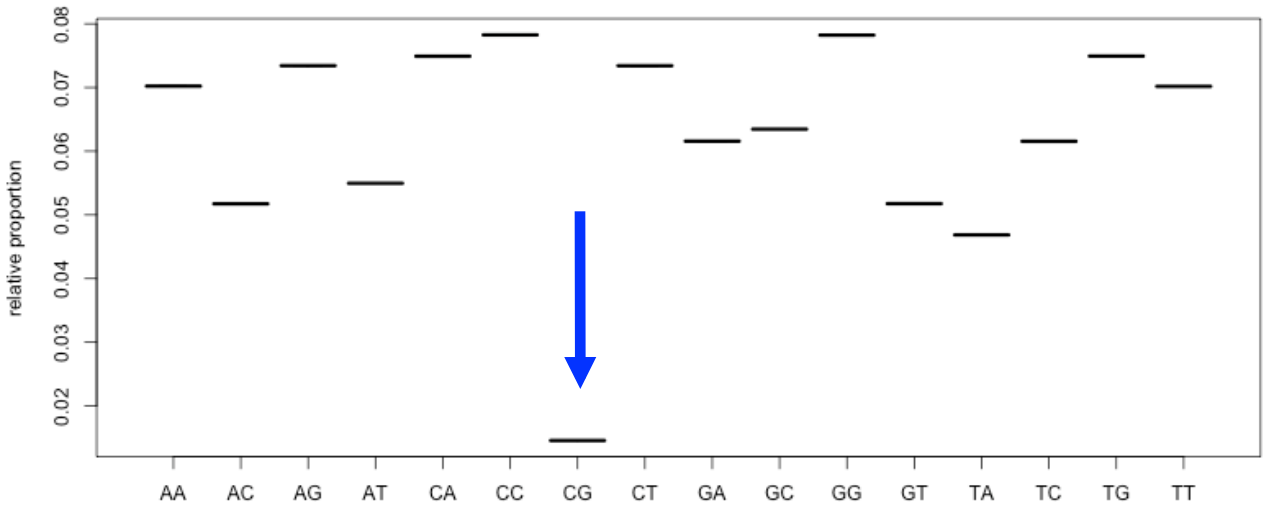
# DNA methylation stably inherited during cell division



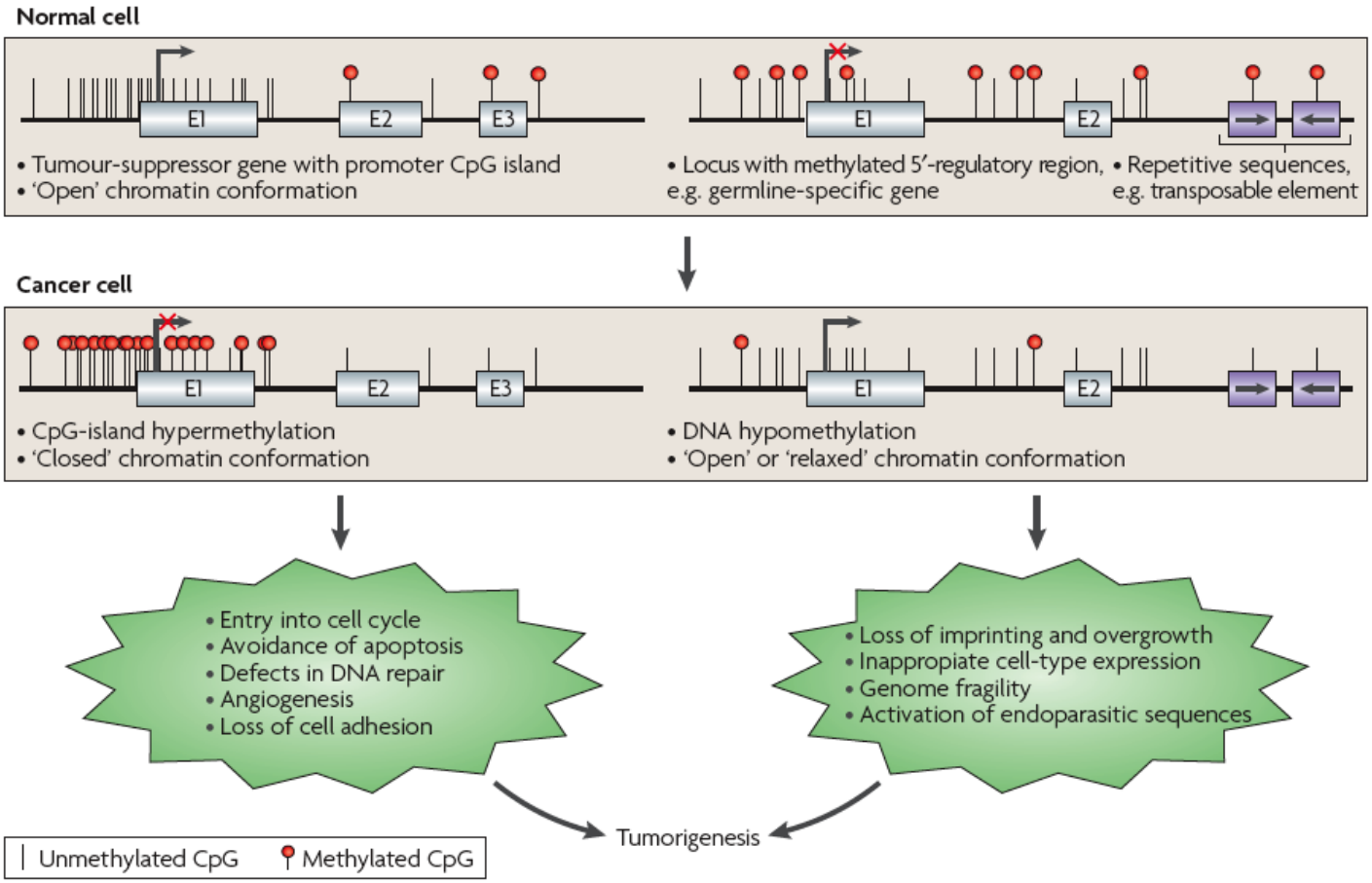

Various DNA methyltranferases (DNMTs), some for “maintenance” (e.g. DNMT1), some for “de novo” methylation.

# CpG islands

- CG dinucleotides are under-represented in the genome, but often occur in “clusters” called CpG islands (CGIs); various CGI definitions



# Dogma: CpG methylation and transcription (cancer)



# DNA methylation

Laird, Nature Reviews Genetics, March 2010

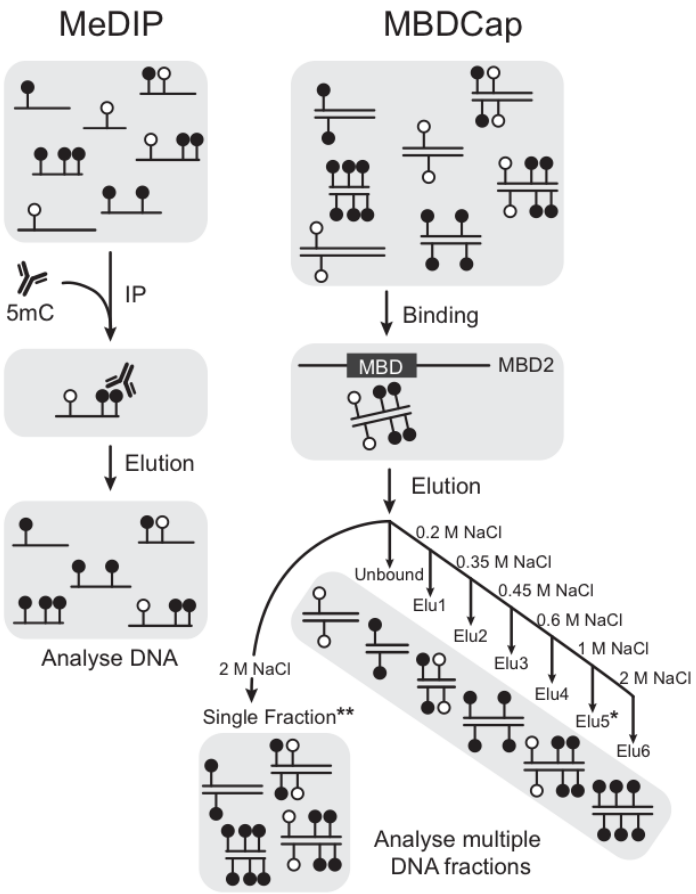
Table 1 | **Main principles of DNA methylation analysis**

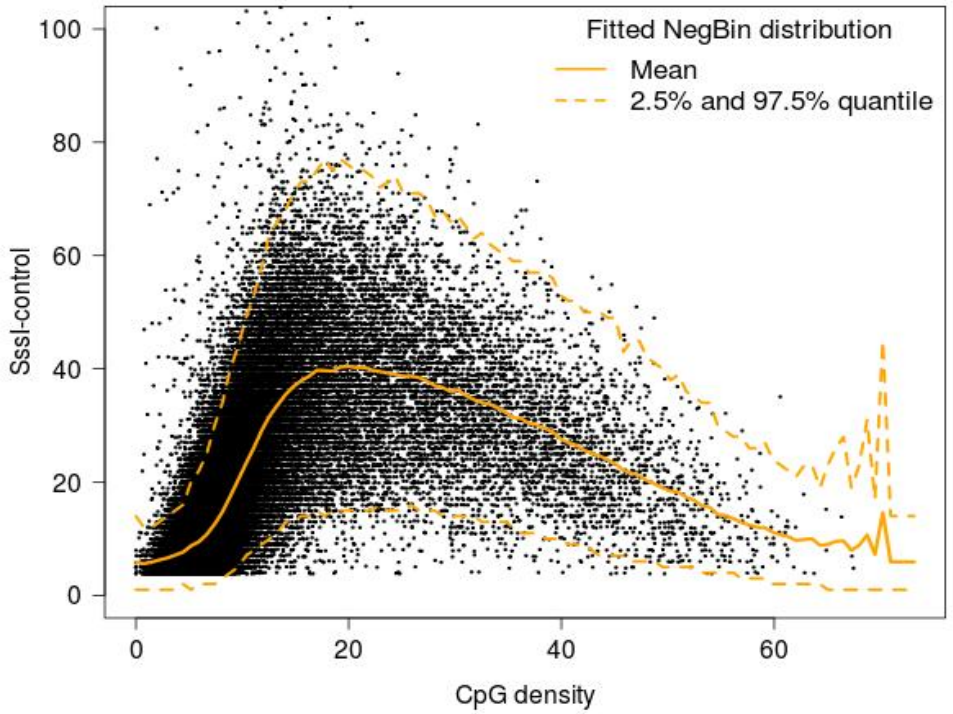
| Pretreatment | Analytical step | | | |
|---|---|---|---|---|
| | **Locus-specific analysis** | **Gel-based analysis** | **Array-based analysis** | **NGS-based analysis** |
| **Enzyme digestion** | • *Hpa*II-PCR | • Southern blot<br>• RLGS<br>• MS-AP-PCR<br>• AIMS | • DMH<br>• MCAM<br>• HELP<br>• MethylScope<br>• CHARM<br>• MMASS | • Methyl–seq<br>• MCA–seq<br>• HELP–seq<br>• MSCC |
| **Affinity enrichment** | • MeDIP-PCR | | • MeDIP<br>• mDIP<br>• mCIP<br>• MIRA | • MeDIP–seq<br>• MIRA–seq |
| **Sodium bisulphite** | • MethyLight<br>• EpiTYPER<br>• Pyrosequencing | • Sanger BS<br>• MSP<br>• MS-SNuPE<br>• COBRA | • BiMP<br>• GoldenGate<br>• Infinium | • RRBS<br>• BC–seq<br>• BSPP<br>• WGSBS |

**Direct sequencing**

**Oxford Nanopore Pacific Biosciences etc.**

# Methods for affinity enrichment (MeDIP-seq, MBD-seq) DNA methylation data




Efficiency of capture in a fully methylated sample, is strongly associated with CpG density.

# BATMAN - Bayesian tool for methylation analysis

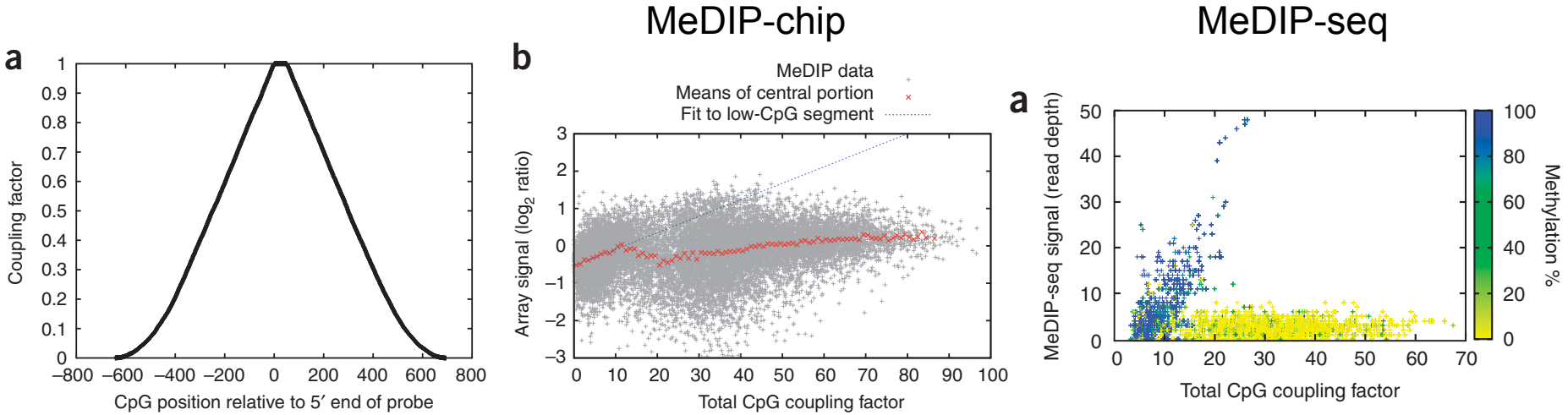
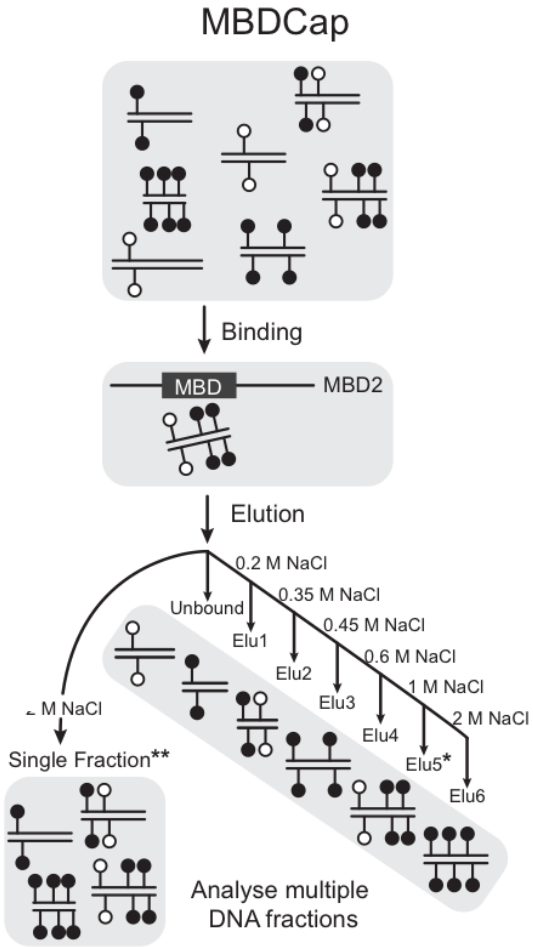
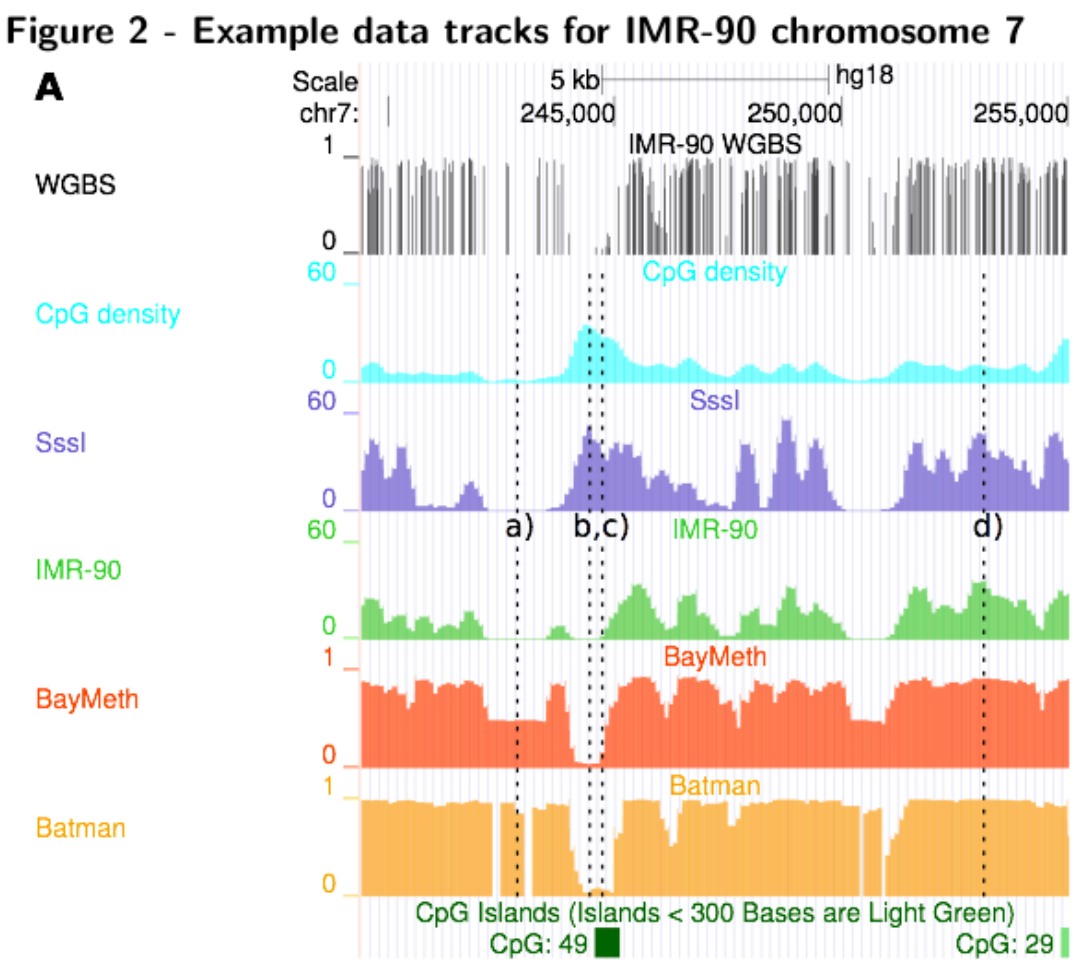



**Figure 1** Calibration of the Batman model against MeDIP-chip data. (**a**) Estimated CpG coupling factors for a MeDIP-chip experiment as a function of the distance between a CpG dinucleotide and a microarray probe. (**b**) Plot of array signal against total CpG coupling factor, showing a linear regression fit to the low-CpG portion, as used in the Batman calibration step. This plot shows all data from one array on chromosome 6.

# Motivation for BayMeth: Estimating DNA methylation levels using fully methylated control




Riebler et al. 2014, Genome Biology

# Model formulation: BayMeth

We consider genomic regions $i = 1, \ldots, n$ and define

- $y_{i,S}$: Number of reads for the sample of interest.
- $y_{i,C}$: Number of reads for the *Sssl-control.*

Then,

$$y_{i,S}|\mu_i, \lambda_i \sim \text{Poisson}(f \times \mu_i \times \lambda_i); \qquad y_{i,C}|\lambda_i \sim \text{Poisson}(\lambda_i)$$

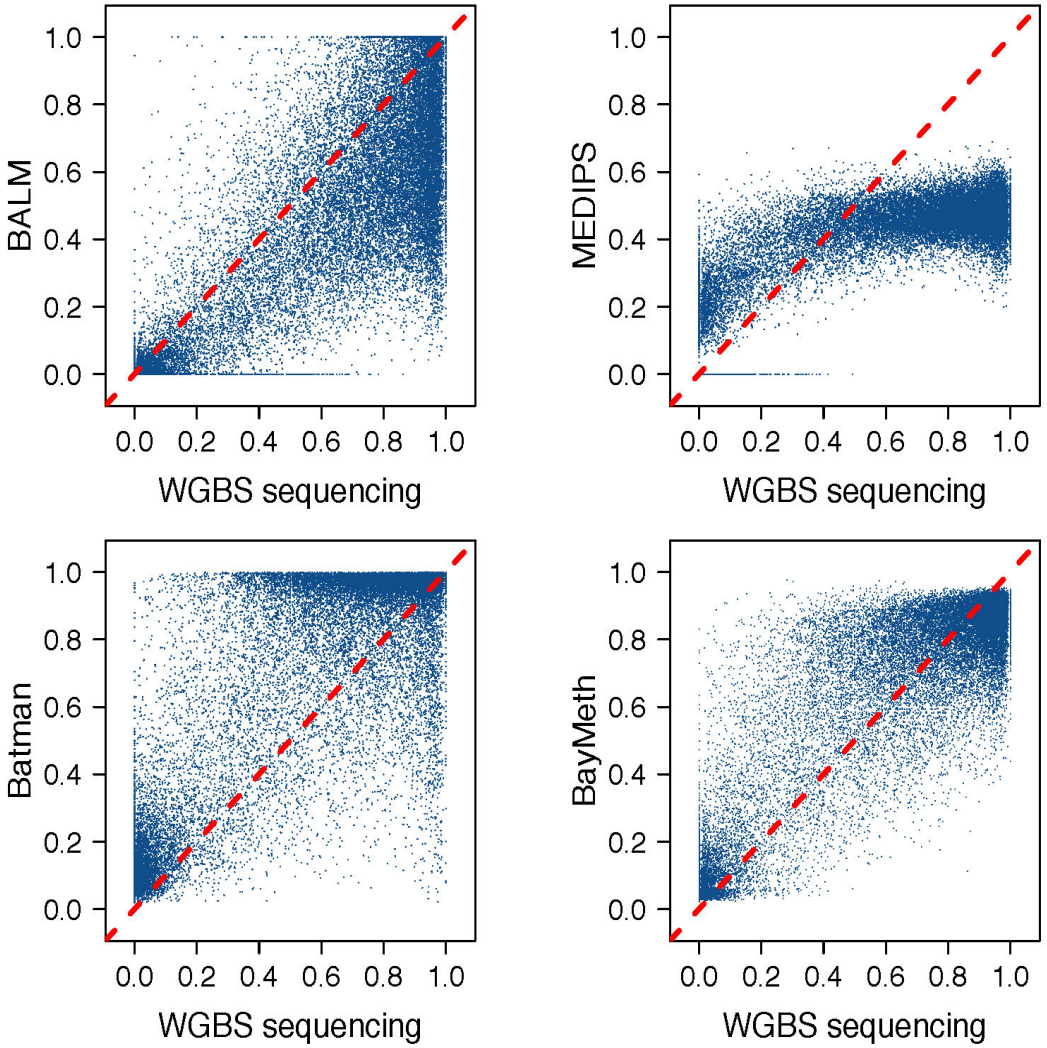
f: offset

$\lambda_i$: region-specific read density, and

$\mu_i$: the regional methylation level (Main parameter of interest)

Riebler et al. 2014, Genome Biology

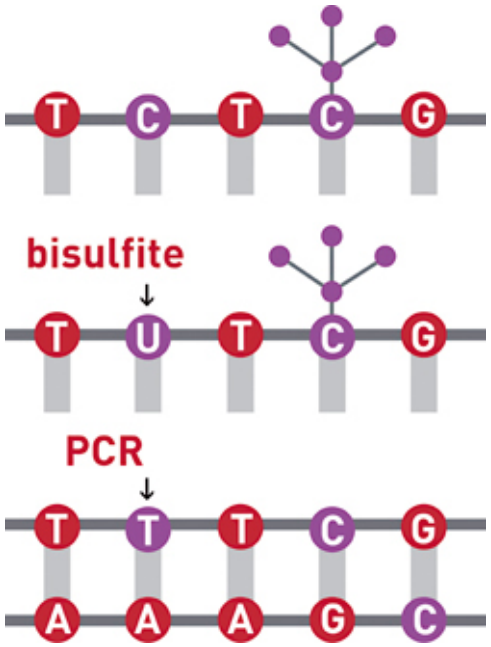
# Improvements can be made: compare to "truth"




← BayMeth

Riebler et al. 2014, Genome Biology

# Bisulphite sequencing



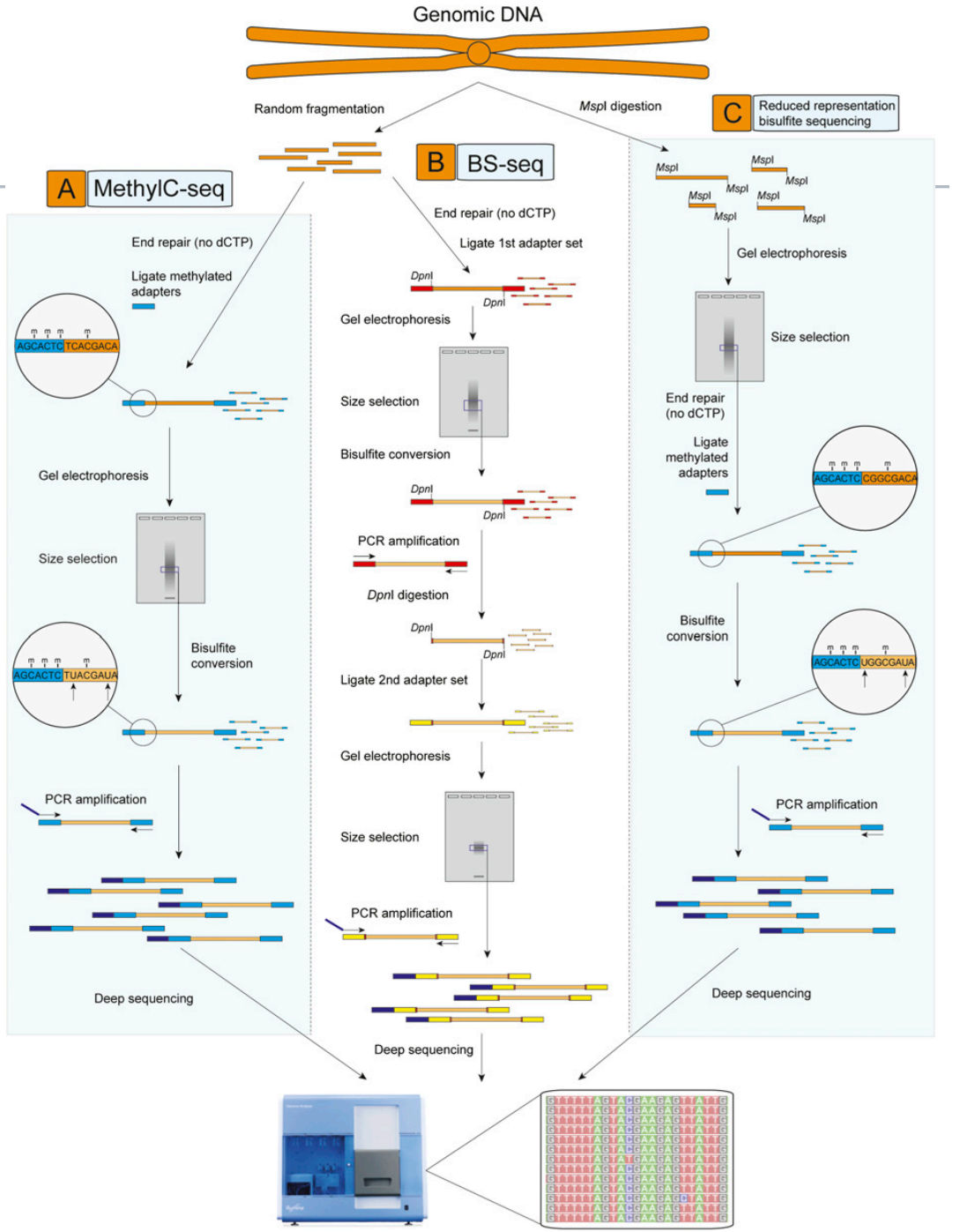

Sodium bisulphite converts methylated **C**ytosine into **U**racil, which can be read as **T**hymine after PCR

In combination with sequencing (Sanger or NGS), can achieve methylation mapping at single base resolution

Can be nicely combined with genotyping arrays (e.g. Illumina HumanMethylation 450k) or with high-throughput sequencing.

http://www.diagenode.com/en/applications/bisulfite-conversion.php

# DNAme methods that use bisulphite conversion with sequencing



BIOINFORMATICS ORIGINAL PAPER Vol. 28 no. 13 2012, pages 1698–1704
doi:10.1093/bioinformatics/bts254

Sequence analysis Advance Access publication May 10, 2012

**Fast and sensitive mapping of bisulfite-treated sequencing data**

Christian Otto[1,2], Peter F. Stadler[1,2,3,4,5,6] and Steve Hoffmann[1,2,*]

[1]Interdisciplinary Center for Bioinformatics and Bioinformatics Group, Department of Computer Science, University Leipzig, 04107 Leipzig, Germany, [2]Transcriptome Bioinformatics Group, LIFE — Leipzig Research Center for Civilization Diseases, University Leipzig, 04107 Leipzig, Germany, [3]RNomics Group, Fraunhofer Institute for Cell Therapy and Immunology, 04103 Leipzig, Germany, [4]Santa Fe Institute, Santa Fe, NM 87501 USA, [5]Department of Theoretical Chemistry, University of Vienna, A-1090 Vienna, Austria and [6]Max-Planck-Institute for Mathematics in Sciences, 04103 Leipzig, Germany

Associate Editor: Michael Brudno

# Bisulphite sequencing analyses: mapping

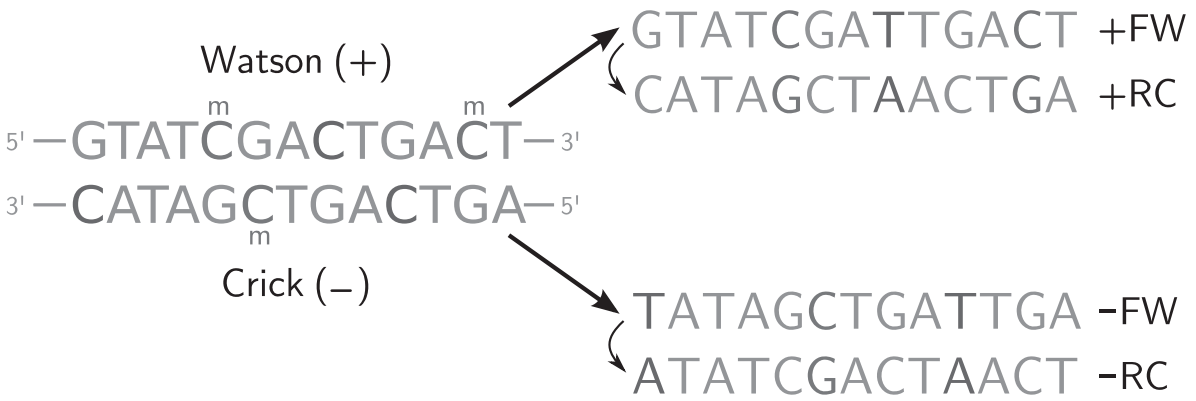



**Fig. 1.** Possible read types (+FW, +RC, −FW and −RC) in bisulfite sequencing protocols. Methylated and unmethylated cytosines in the genomic sequence (left) are coloured in red and blue, respectively, and positions in the read sequences (right) derived from genomic cytosines are coloured correspondingly. Note that the intermediate conversion of unmethylated cytosines into uracils after bisulfite treatment is omitted

University of Zurich[UZH]

Institute of Molecular Life Sciences

# Genotypes with BS-seq data

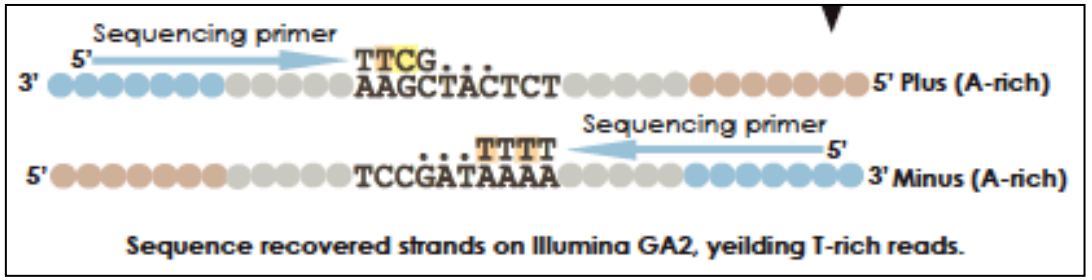
| | Positive Strand | Negative Strand |
|---|---|---|
| Reference genome: | TCCGATGAGA | TCTCATCGGA |
| Add optional methylation: | TC**CG**ATGAGA | TCTCAT**CG**GA |
| Actual read: | TTCGATGAGA | TTTTATCGGA |

Rule:

T G A C **CG**

↓ ↓ ↓ ↓ ↓

T G A T CG




Sequence recovered strands on Illumina GA2, yeilding T-rich reads.

University of Zurich[UZH]
Institute of Molecular Life Sciences

# Genotypes with BS-seq data

| | Positive Strand | Negative Strand |
|---|---|---|
| Reference: | TCCGATGAGA | TCTCATCGGA |
| What if the genome was: | GCCGATGAGA | TCTCATCGGC |
| | CCCGATGAGA | TCTCATCGGG |
| | ACCGATGAGA | TCTCATCGGT |
| Actual read: | TTCGATGAGA | TTTTATCGGA |
| | GCCGATGAGA | TTTTATCGGT |
| | TCCGATGAGA | TTTTATCGGG |
| | ACCGATGAGA | TTTTATCGGT |

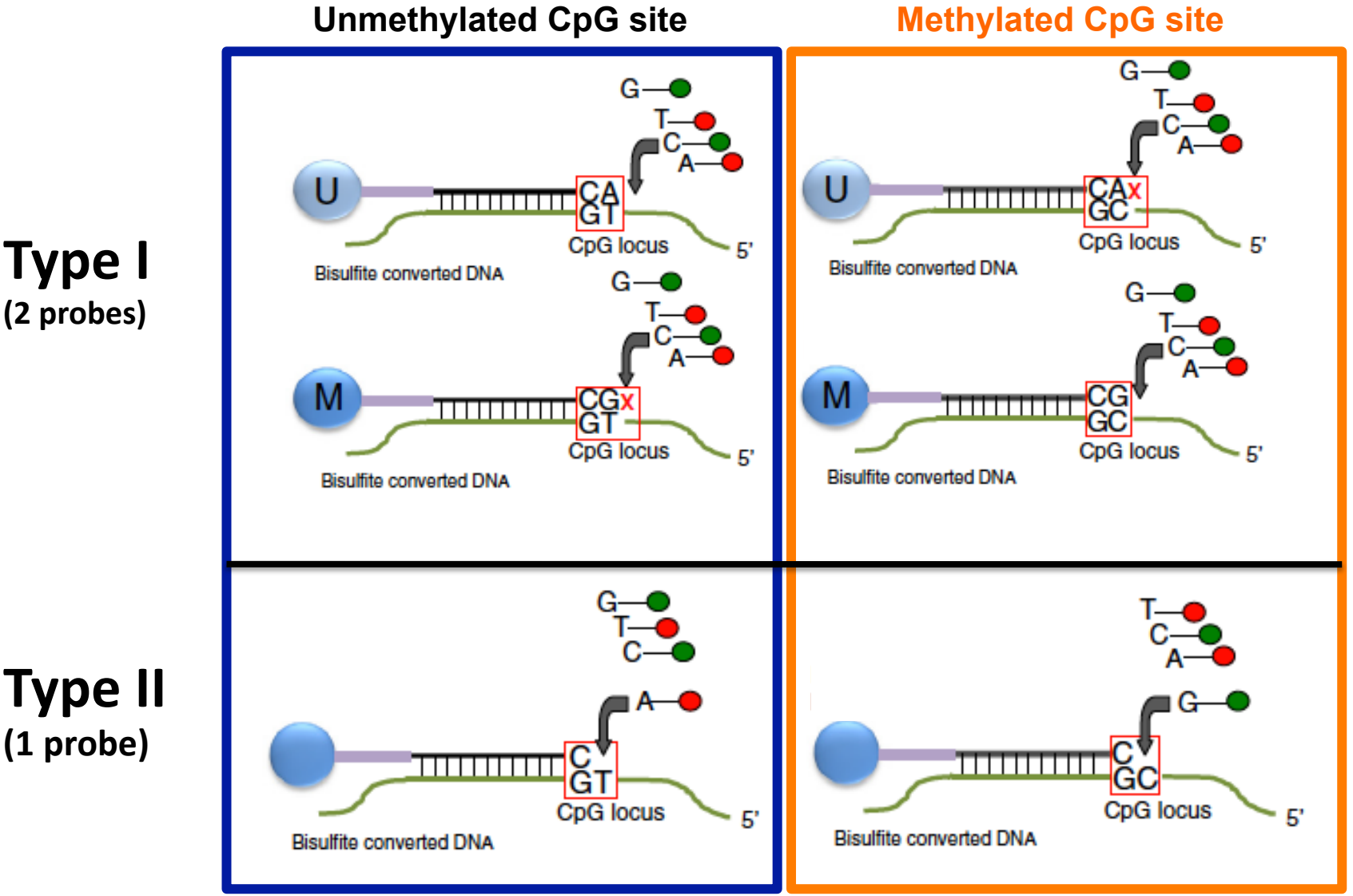
T G A C CG
↓ ↓ ↓ ↓ ↓
T G A T CG

You can reconcile the ambiguity with the read from the opposite strand.

University of Zurich[UZH]

# Bisulphite conversion **but not sequencing**, rather genotyping (450k array)

Institute of Molecular Life Sciences

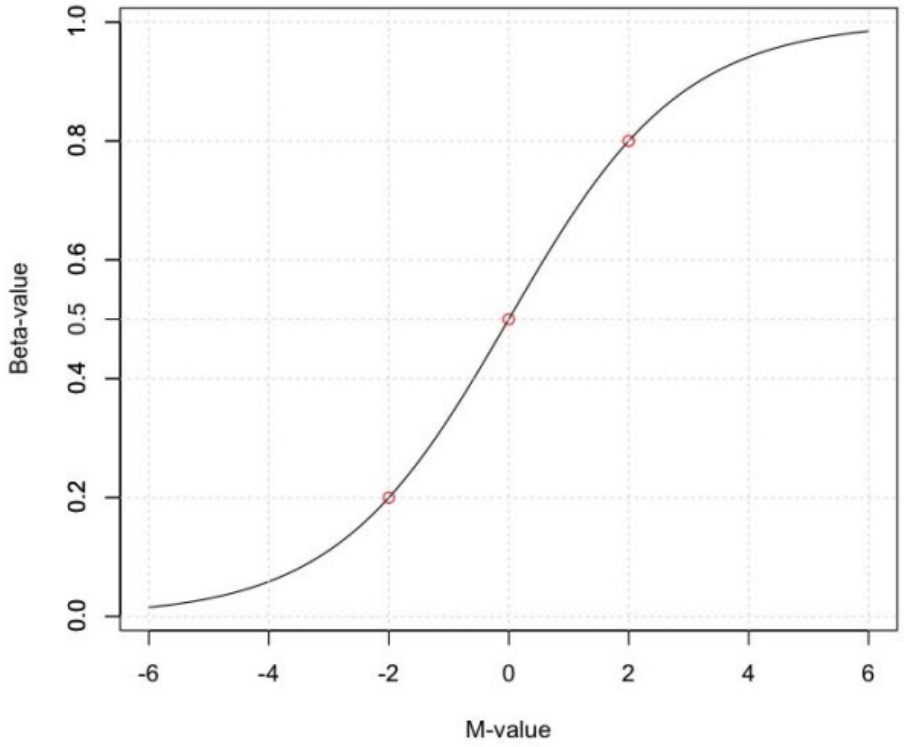



from Bibikova et al. Genomics 2011

# Intuitive measure of methylation: “beta” values

beta = M / (M+U+e)

If e=0, logistic relationship between fold change and beta




Du *et al. BMC Bioinformatics 2010*

# Analysis of 450k arrays

Overall, very good correspondence between 450k platform and others (e.g. BS-seq)

**Normalization** issues for different probe types

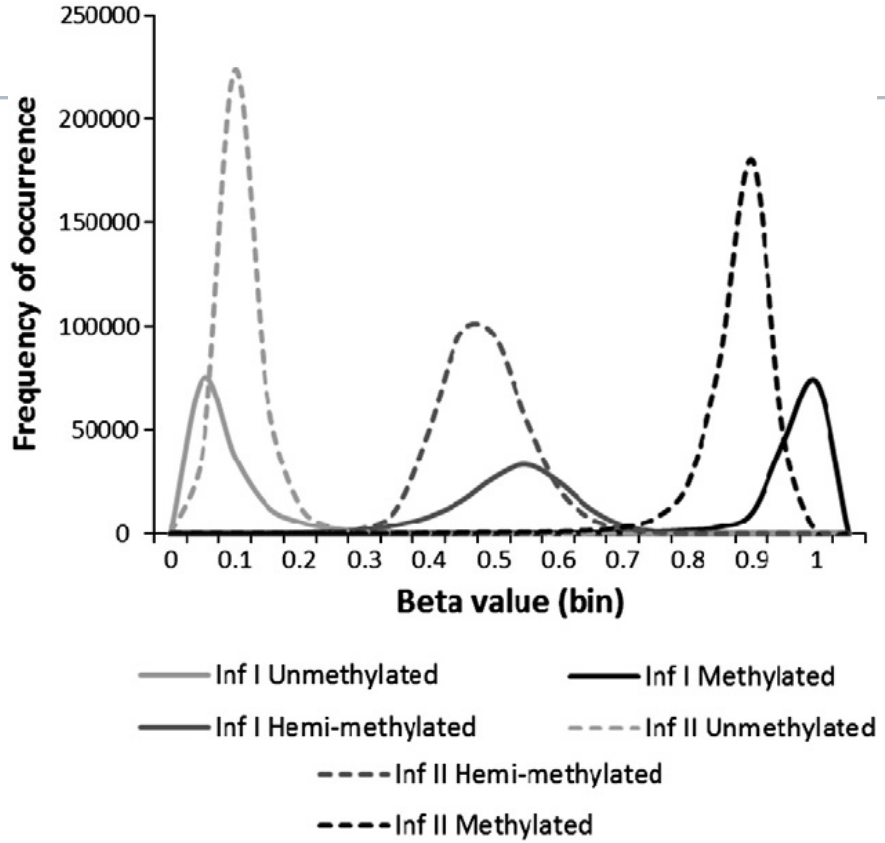
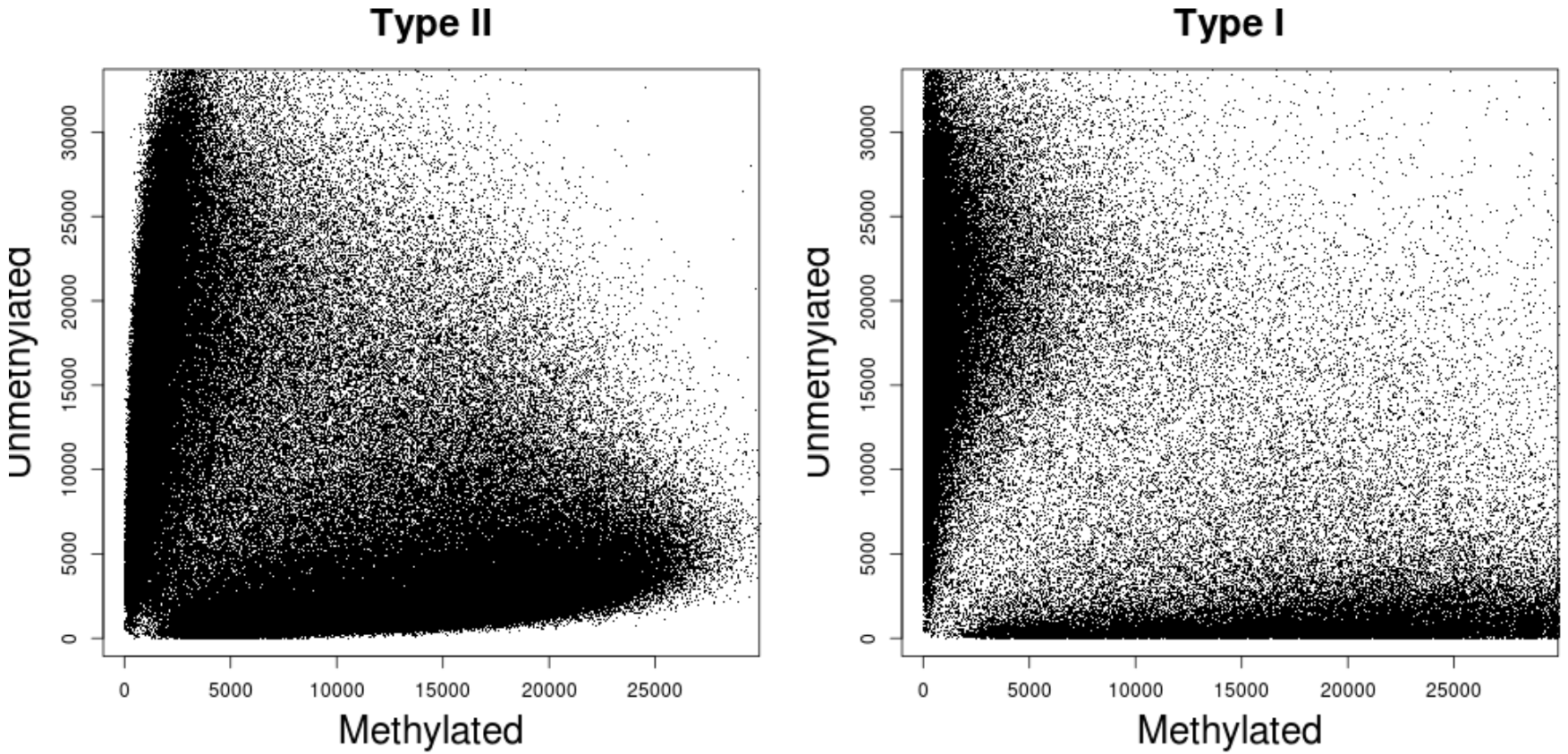



**Fig. 3.** Distribution of Methylation values for Infinium I and Infinium II loci. Unmethylated (U), Hemi-methylated (H), and Methylated (M) reference standards were created from Coriell genomic DNA sample as discussed in Methods. Note slightly different performance of Infinium I and Infinium II assays in regard to beta value distribution.

# 450k array data

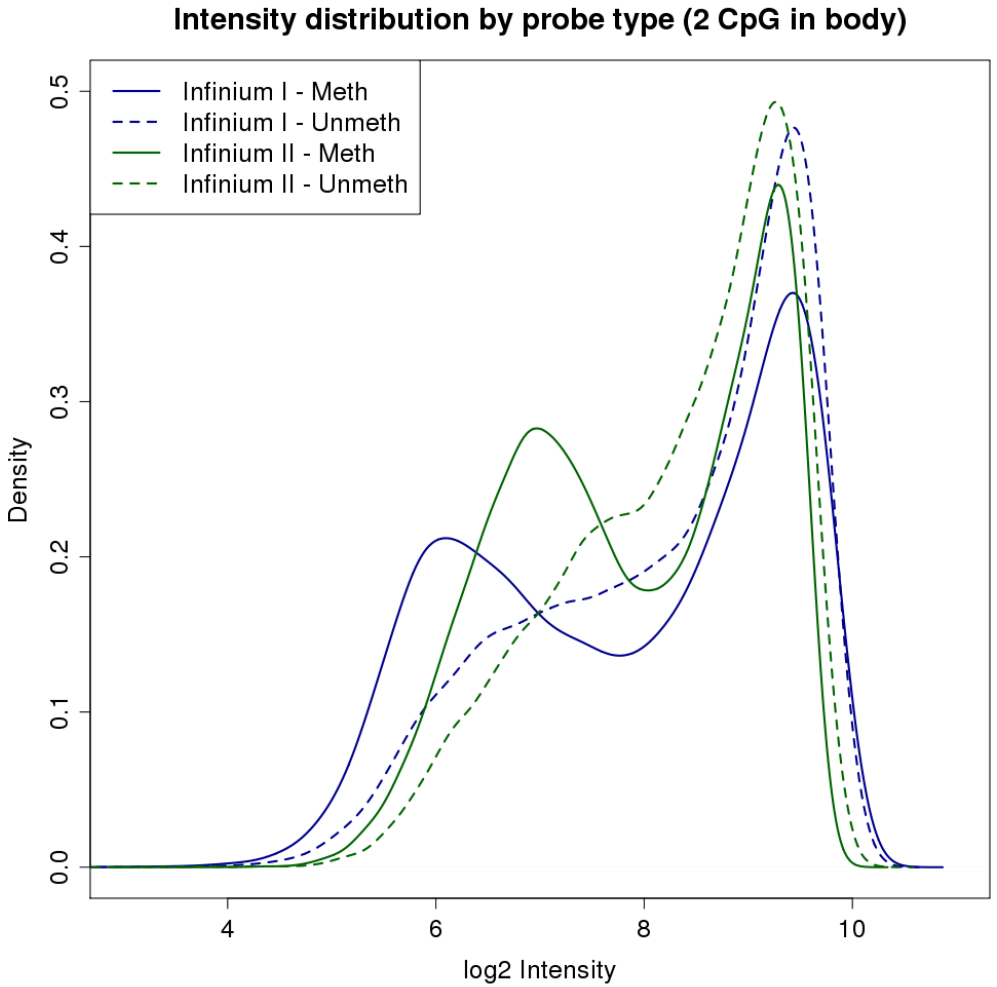
Very different behaviour of Type I and Type II probes



# Intensity distribution of probes with 2 CpGs

Not only are type I and type II probes distributed very differently, the presence of CpG sites (which can be unmethylated or methylated) can affect the observed signal.

Also, present of SNPs in probe may differentially affect human samples




Maksimovic et al. 2012, Genome Biology

# Methods for differential methylation: "bump hunting"

Methods for differential methylated sites use: i) log-ratios of methylated to unmethylated signal (450k array); ii) difference in binomials (BS-seq)

Methods are in active development for going from differentially methylated sites to differentially methylated **regions** (e.g. bump hunting)

Jaffe et al. (2012) Int. Journal of Epidemiology

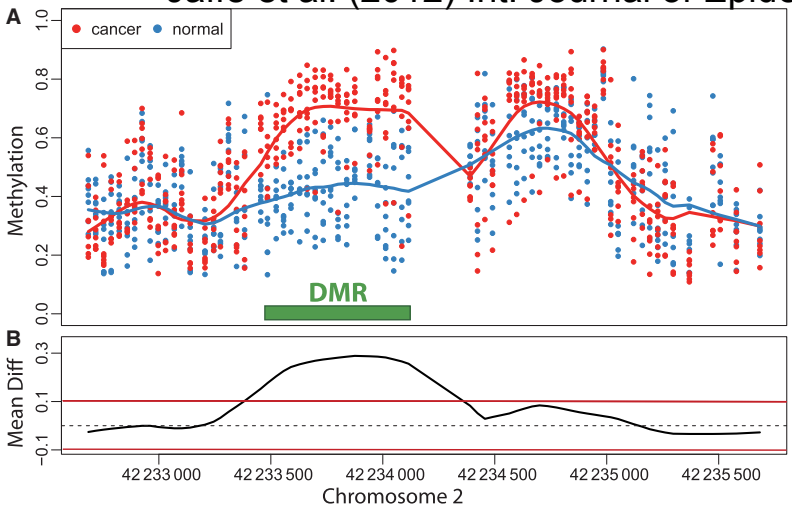


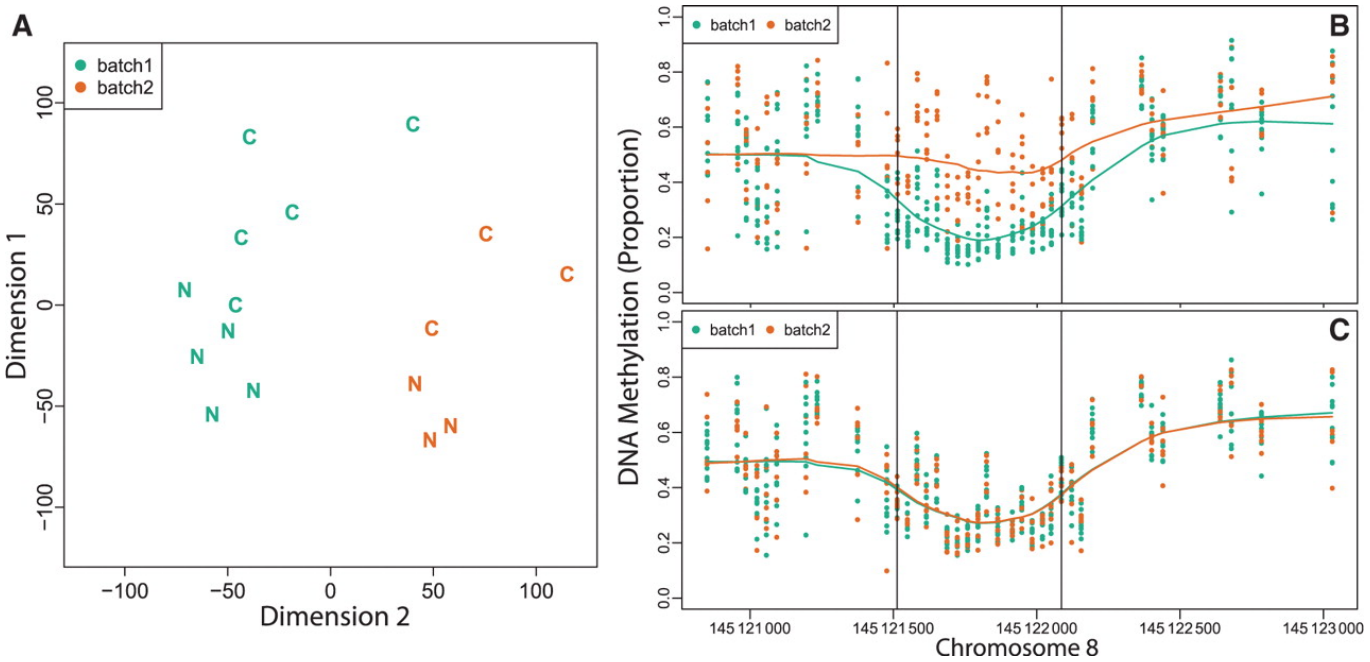

**Figure 1** Example of a differentially methylation region (DMR). (A) The points show methylation measurements from the colon cancer dataset plotted against genomic location from illustrative region on chromosome 2. Eight normal and eight cancer samples are shown in this plot and represented by eight blue points and eight red points at each genomic location for which measurements were available. The curves represent the smooth estimate of the population-level methylation profiles for cancer (red) and normal (blue) samples. The green bar represents a region known to be a cancer DMR.[20] (B) The black curve is an estimate of the population-level difference between normal and cancer. We expect the curve to vary due to measurement error and biological variation but to rarely exceed a certain threshold, for example those represented by the red horizontal lines. Candidate DMRs are defined as the regions for which this black curve is outside these boundaries. Note that the DMR manifests as a *bump* in the black curve

$$Y_{ij} = \mu(t_j) + \beta(t_j)X_i + \sum_{k=1}^{p} \Upsilon_k(t_j)Z_{i,k} + \sum_{l=1}^{q} a_{l,j}W_{i,l} + \varepsilon_{i,j}$$

# Methods for differential methylation

Batch effects are ever-present

**charm::dmrFind()**




Jaffe et al. (2012) Int. Journal of Epidemiology

# Probe-level methylation → region methylation

$$Y_{ij} = \mu(t_j) + \beta(t_j)X_i + \sum_{k=1}^{p} \Upsilon_k(t_j)Z_{i,k} + \sum_{l=1}^{q} a_{l,j}W_{i,l} + \varepsilon_{i,j}$$

i – individual
j – loci

Includes surrogate variable analysis

Outcome of interest (e.g. cancer versus normal)

Measured confounders

Unmeasured confounders

Jaffe et al. (2012) Int. Journal of Epidemiology

# Mechanics of DMR finding: charm package

Steps:

1. Get normalized data
2. For each probe (CpG site), calculate (differential) statistics at each probe
3. Apply a smoothing technique to these statistics
4. Set threshold and call regions as those that persist beyond threshold