

# Introduction to DNA Methylation Platforms and Data Analysis

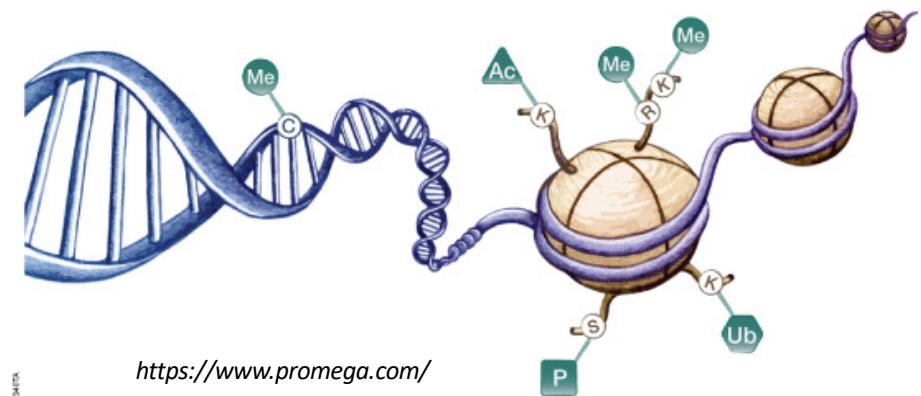
Dr. Katerina Kechris  
Professor  
Department of Biostatistics and Informatics  
University of Colorado Anschutz Medical Campus

# Outline

- Introduction
- Example
- Methylome Profiling
- Technologies (Array, Sequencing) and Data Processing
- Analysis and Resources

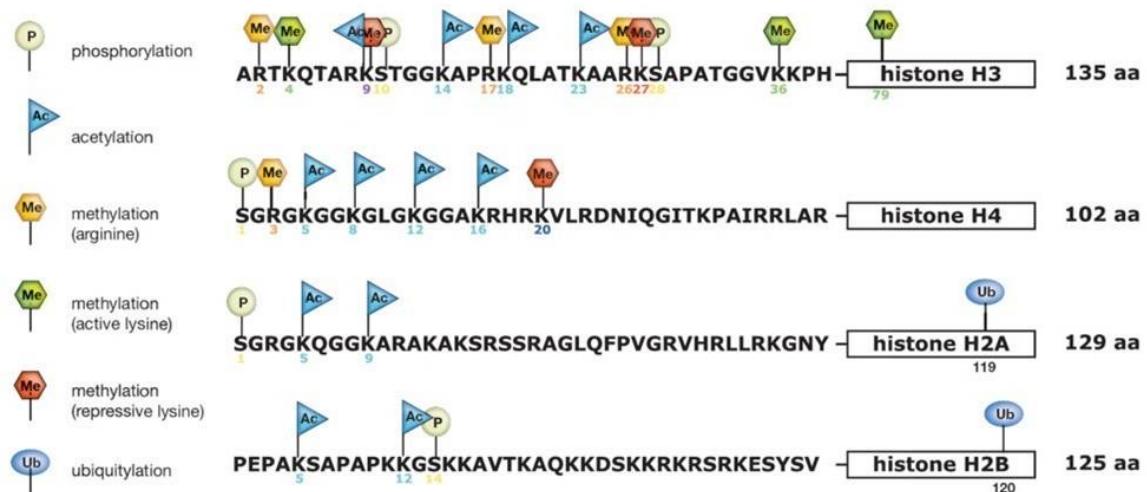
# Epigenetics

- “epi” = above/in addition to the genome
- Genomic modifications that do not involve a change in the nucleotide sequence
  1. Histone modification
  2. DNA methylation
- Influences regulation of gene expression
- Can occur in response to environmental exposures



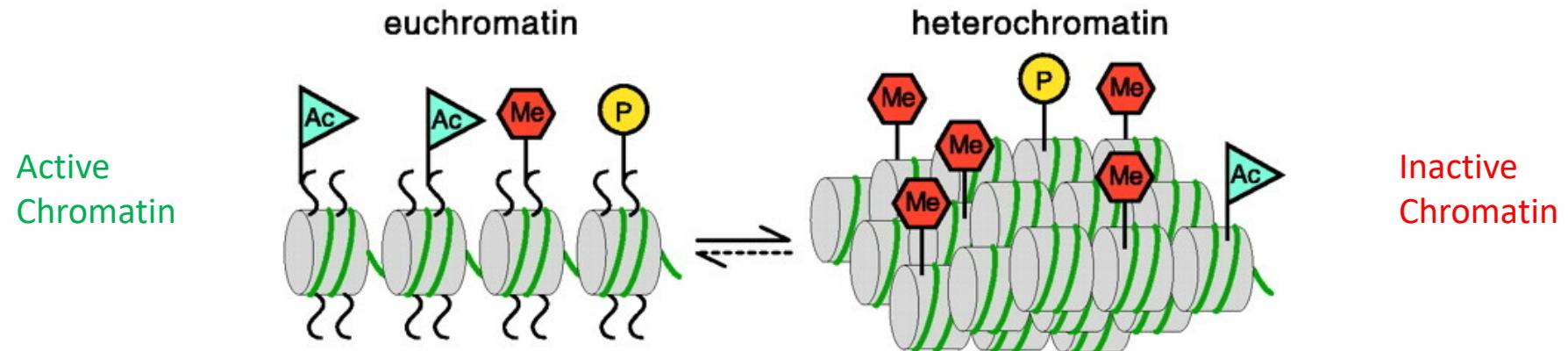
# Histone Modifications

- Modifications to histone proteins affect DNA accessibility
- Influence the activation or repression of gene expression
- Histone acetylation & methylation - most well characterized modifications



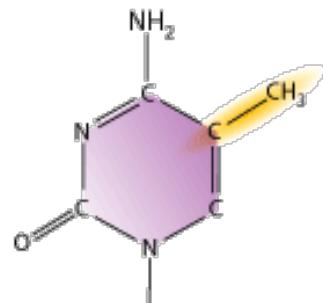
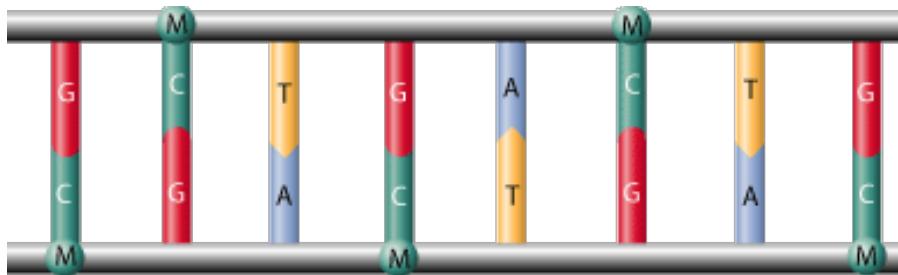
Allis et al (2006) *Epigenetics*

# DNA Accessibility



Jenuwein & Allis (2001) Science

# DNA Methylation



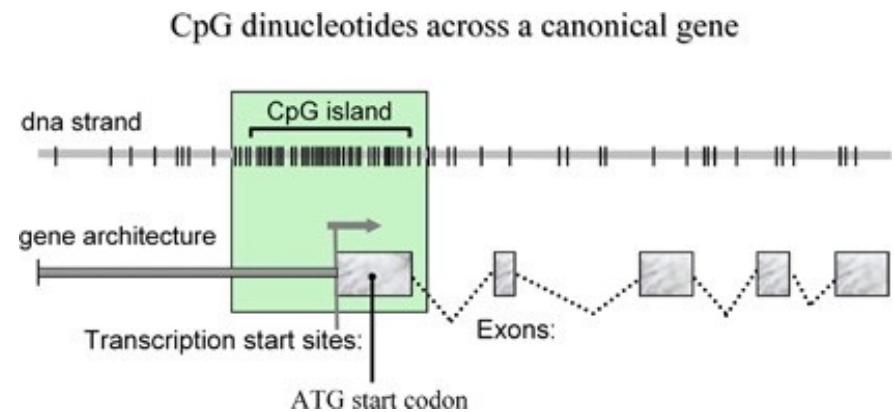
DNA methylation is the addition of a methyl group (M) to the DNA base cytosine (C).

<http://extremelongevity.net>

- Typically occurs at CpG dinucleotides (~28M in human genome)
- Rare non-CpG methylation of C followed by A,T or C (more common in plants)

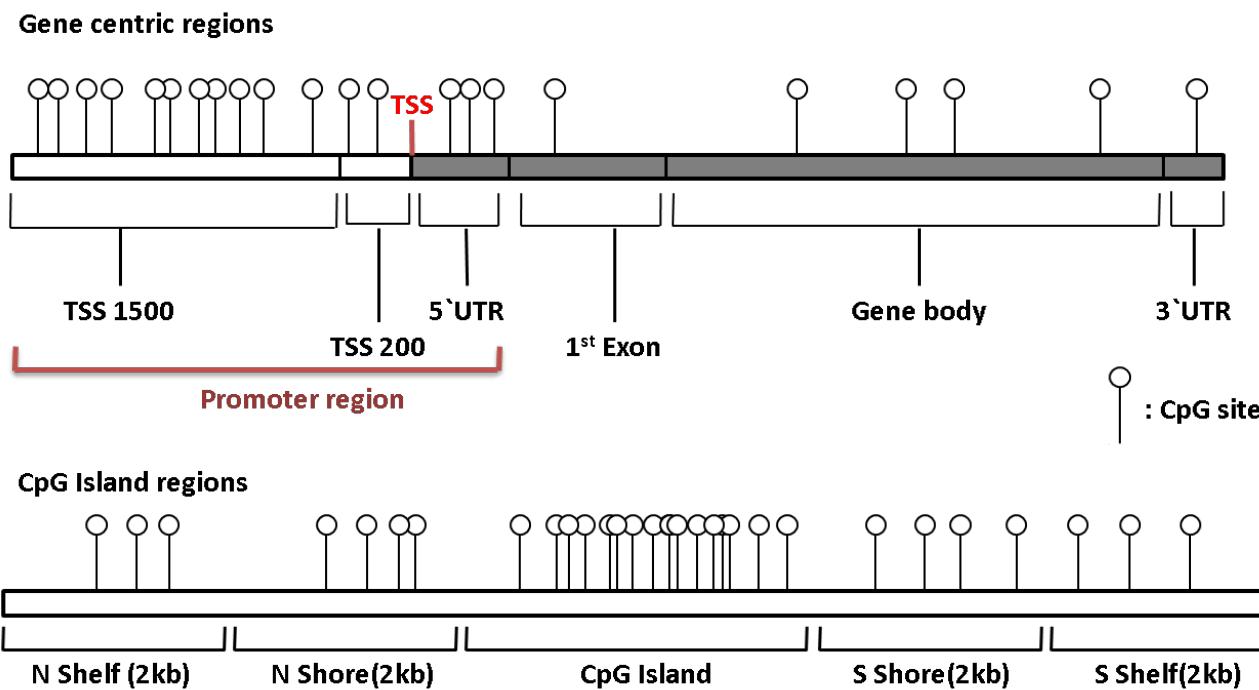
# CpG Islands

- CpG – C then G - linked by (p)phosphate bond
- 70%-80% of CpG cytosines can be methylated (mammals)
- CpG Island - clusters of unmethylated CpGs occurring near highly expressed genes
  - G+C content > 0.50, CpG dinucleotide ratio > 0.60, > 200bp window
- 40-50% genes have CpG islands in promoters (mammals)
- Genes usually expressed if CpG sites in CpG islands of gene promoter is unmethylated



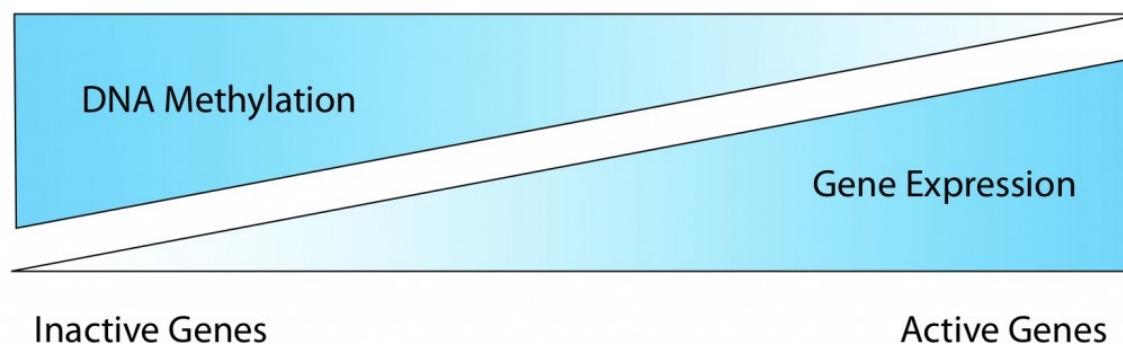
Sharma et al., (201) Neuropsychopharmacology

# CpG Island, Shore, Shelf



# DNA Methylation

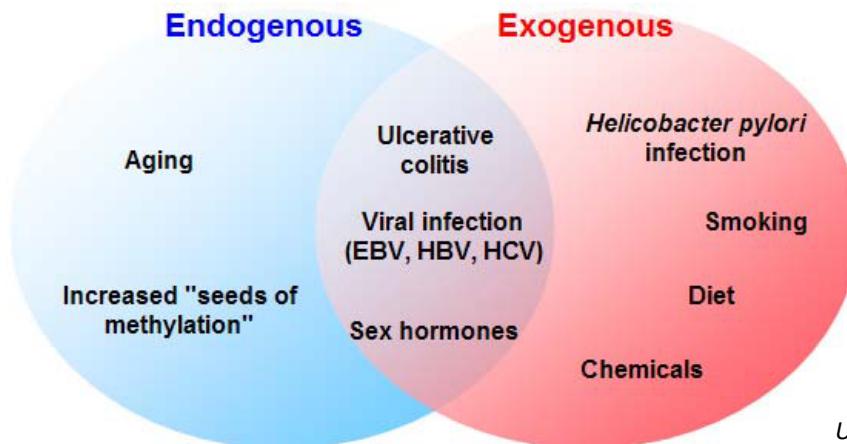
- Methyl group interferes with binding of transcriptional activators



*<http://www.precisionnutrition.com>*

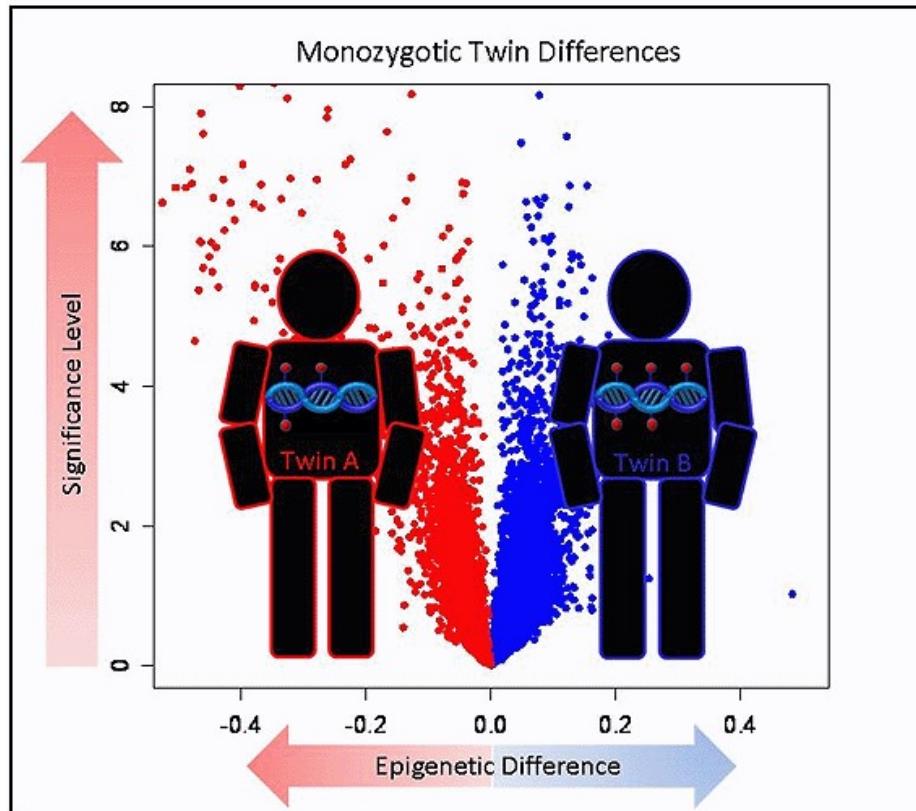
# DNA Methylation

- Stable: Heritable through cell division
- Plastic: May change due to environment
  - Diet, smoking, physical activity, intrauterine environment, medicines, pollutants, aging process, ...
- Provides record of environmental exposures



Ushijima & Okochi-Takada (2005) Cancer Sci.

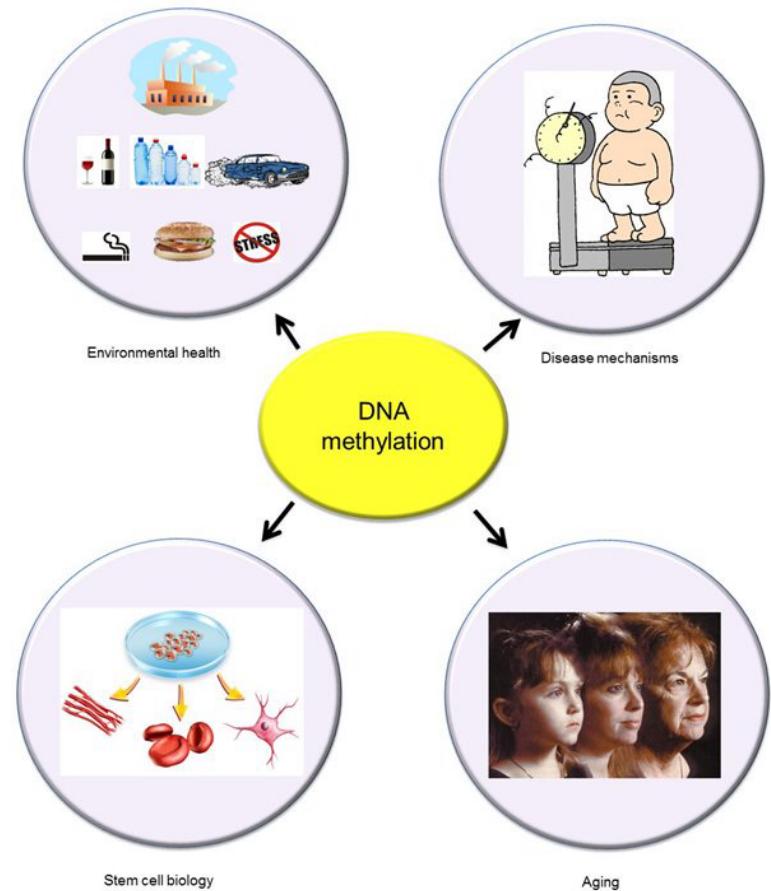
# Twin Studies



Adapted from Kaminsky (2009) *Nature Genetics*

# Studies of DNA Methylation

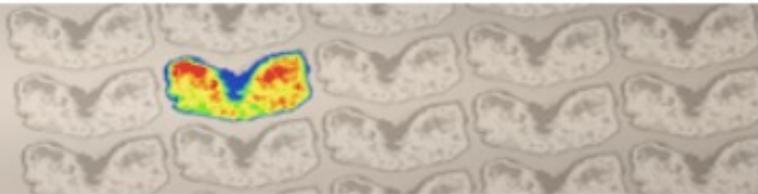
- Biological Processes
  - e.g., stem cell differentiation, embryonic development, inflammation, aging
- Disease Mechanisms
  - e.g., cancer, atherosclerosis, asthma, diabetes
- Environmental Exposures



<https://www.urmc.rochester.edu/>

# Outline

- Introduction
- Example
- Methylome Profiling
- Technologies (Array, Sequencing) and Data Processing
- Analysis and Resources



[Transl Psychiatry](#) 2017 Jan; 7(1): e994.

PMCID: PMC5545731

## **Alcohol-dose-dependent DNA methylation and expression in the nucleus accumbens identifies coordinated regulation of synaptic genes**

[R Cervera-Juanes](#),<sup>1</sup> [L J Wilhelm](#),<sup>1</sup> [B Park](#),<sup>2</sup> [K A Grant](#),<sup>1</sup> and [B Ferguson](#)<sup>1,\*</sup>

# Goals

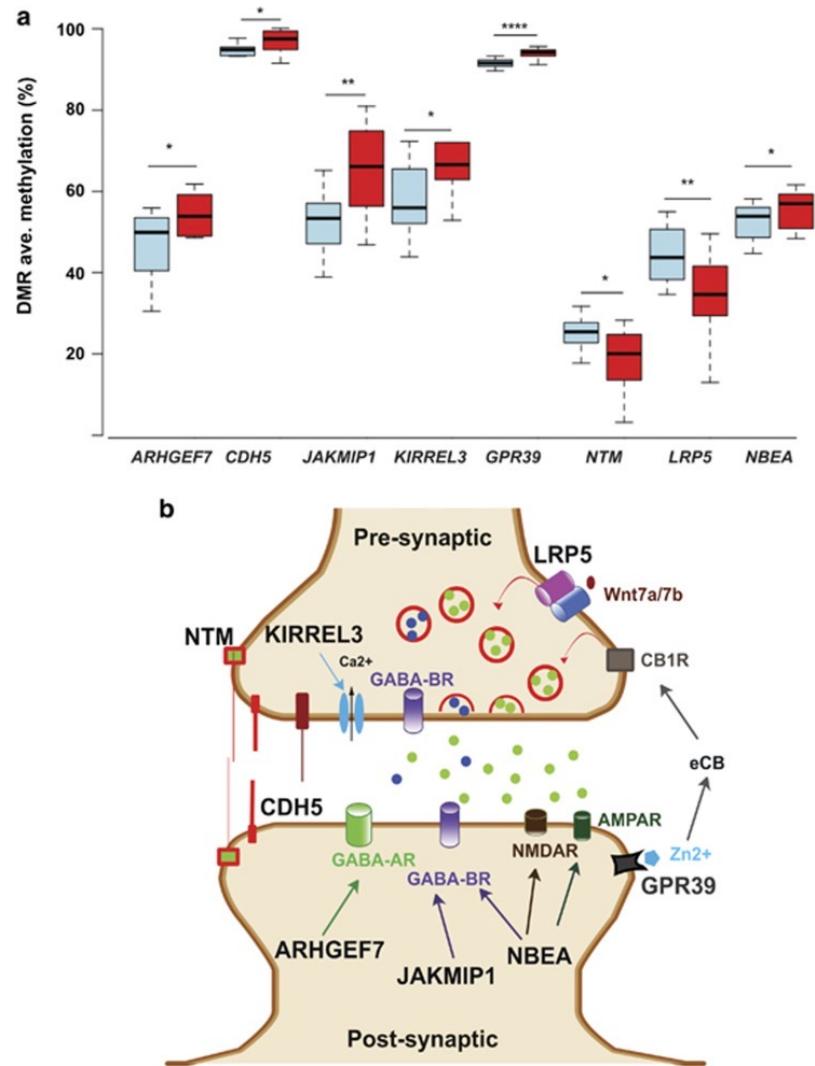
- Study the neuroadaptive changes associated with dependent or compulsive drinking
- Use non-human primate – overcomes logistical constraints of human studies
- Nucleus Accumbens Core (brain) - “relay station selecting and integrating the most relevant environmental stimuli among competing limbic and cortical afferents to drive behavioral output such as alcohol seeking”
- Identify biological mechanisms that cause changes due to long term alcohol use

# Study Design

- Male rhesus macaques ( $n=30$ , *Macaca mulatta*) - late adolescents, young adults and middle aged adults
- Voluntary and long-term ethanol self-administration (12 months)
  - Open access (22 h per day) to 4% alcohol and water *ad libitum*
  - Alcohol intake recorded
  - Subjects grouped into Low/Binge Drink ( $n=7$ ) or High/Very-heavy Drinkers ( $n=9$ ) for methylome profiling
- Nucleus Accumbens Core (NAcc) Methylome
  - Sequencing: Bisulfite Conversion + Methyl-Seq library preparation
- qPCR for target genes

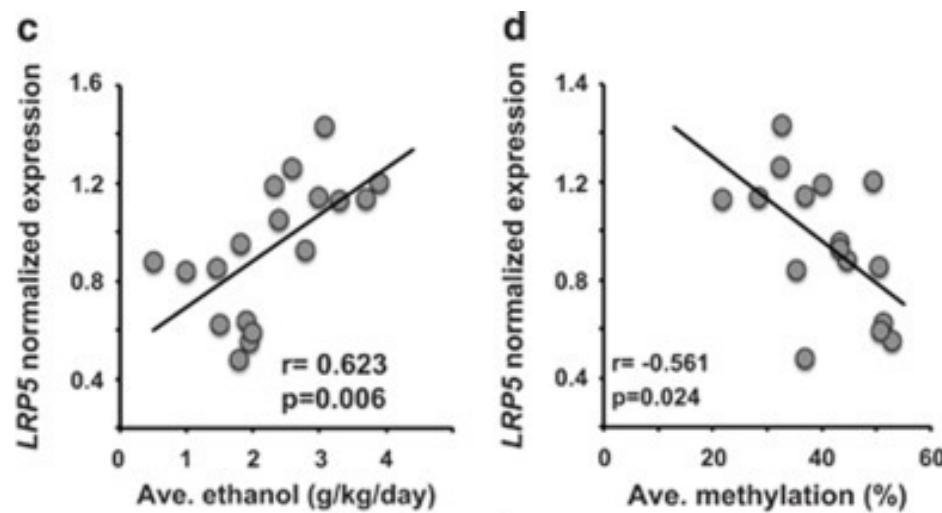
# Results

- 17 differential methylation regions (DMRs) - most correlated with average daily alcohol consumption
- 8 of the DMR mapped to genes implicated in modulating synaptic plasticity
- Validation confirmed the significant alcohol-dose-associated methylation of the DMRs



# Results

- Expression analysis of three of the DMR-associated genes, *LRP5*, *GPR39* and *JAKMIP1* - whole-gene or alternative transcript expression



# Outline

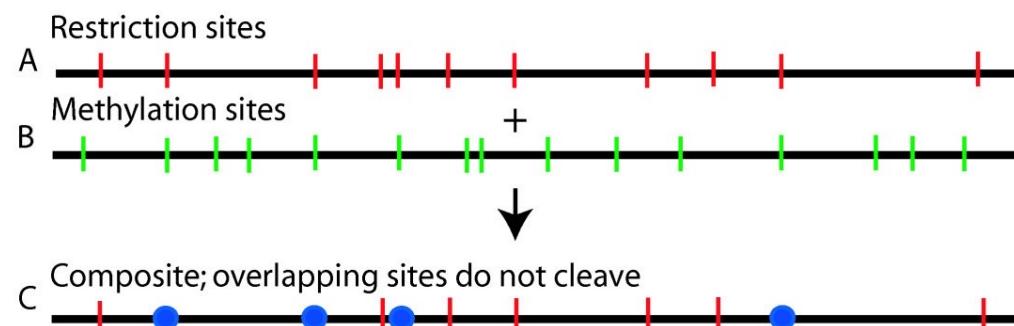
- Introduction
- Example
- Methylome Profiling
- Technologies (Array, Sequencing) and Data Processing
- Analysis and Resources

# Experimental Methods

- DNA methylation information erased by standard molecular biology techniques (e.g., cloning, PCR amplification)
- Pre-treatment needed
  - Restriction enzymes – methylation sensitive - unable to cut methylated DNA, leaves methylated DNA intact
  - Affinity Enrichment – specific antibodies, or methyl-binding proteins
  - Bisulfite Conversion – conversion of C to U

# Pretreatment: Restriction enzymes

- Several methylation sensitive restriction enzymes
- Cleave only unmethylated target sequences, leave methylated DNA intact
- Followed by array (CHARM) or sequencing
- Relatively low coverage
  - CpG containing recognition sites are limited



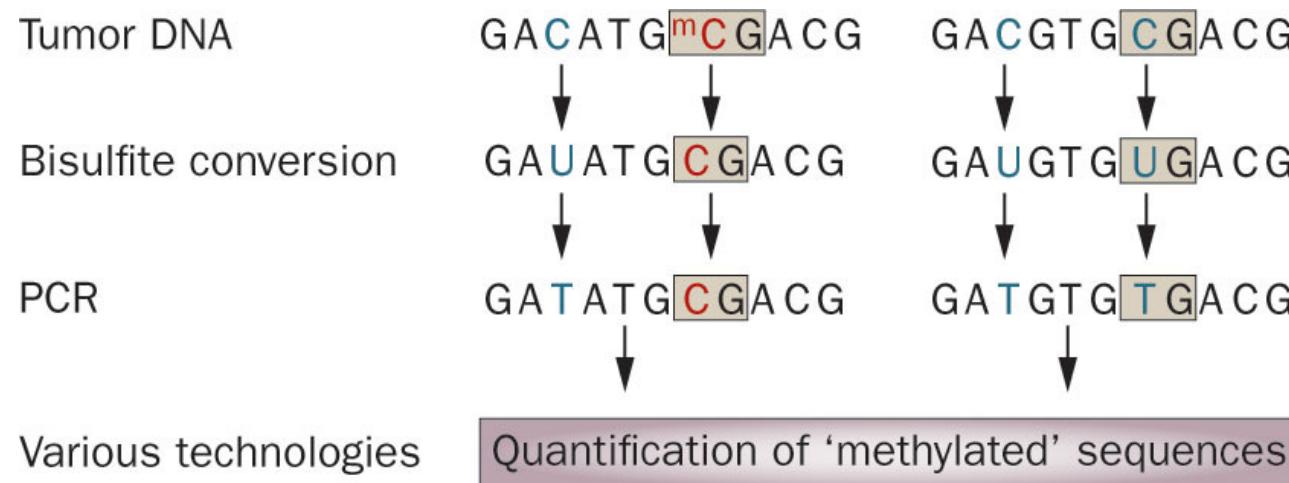
Ananiev, et al (2008) BMC Molecular Biol

# Pretreatment: Affinity Enrichment

- Specific antibodies or methyl-binding proteins to pull down methylated regions
- Examples:
  - Methyl-DNA Immunoprecipitation (meDIP)
  - Methyl-CpG Binding Domain protein (MBD)
- Followed by array or sequencing
- Low starting material
- Biased towards density of CpG
- Regional analysis (100-200bp, not single base resolution)

# Pretreatment: Bisulphite Conversion

Treatment of denatured genomic DNA with sodium bisulphite deaminates unmethylated cytosine:



**Now epigenetic difference looks like genetic difference**

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

ARRAY – most practical for large human studies

SEQUENCING

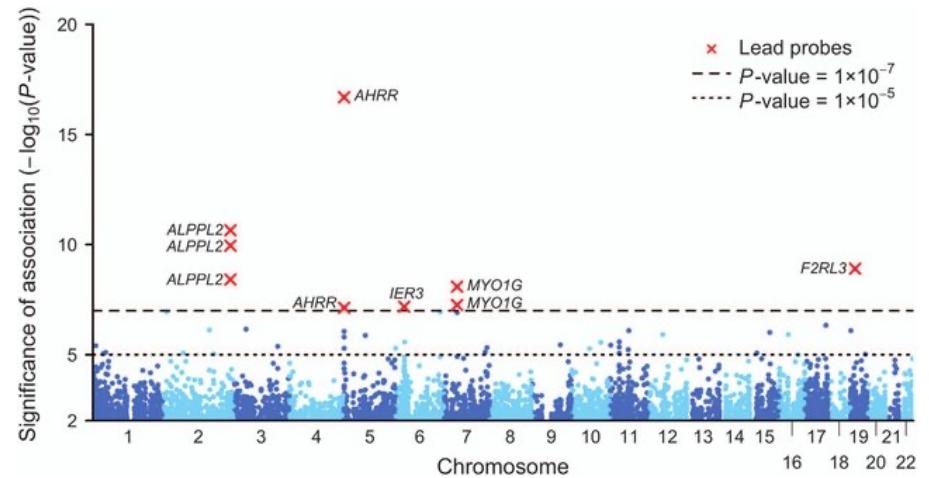
Laird (2010) Nature Review Genetics

# Epigenome-Wide Association Study (EWAS)

Goal:

- Identify methylated regions in a sample
- Identify percentage methylation
- Identify methylation differences between samples

Platforms: Array or Sequencing



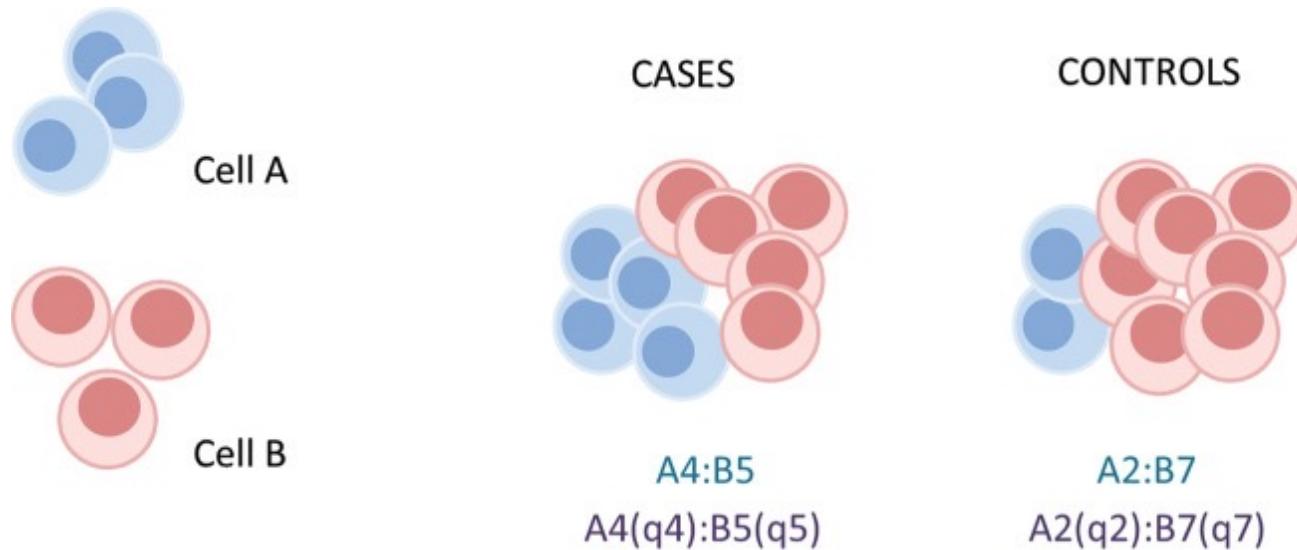
Linner et al., (2017) Molecular Psychiatry

# Sample Selection: Challenges of Cell Mixtures

- Often tissues have mixture of cell types (e.g., saliva, whole blood, placenta, tumors)
- Different cell types may have different methylation levels
- Sample groupings (disease) may have different cell types
  - e.g., autoimmune disease – likely to have different autoimmune cells than non-disease subjects

# Challenge: Cell Mixtures

Confounding – methylation differences due to disease state or cell mixtures?



# Methods

1. Measure complete blood counts (cbc) - if possible
  - Absolute and % of major cell types
  - Use as fixed effects in model
2. Reference data (Houseman et al., *BMC Bioinformatics* 2012)
  - For common tissue types (e.g., whole blood)
  - Methylation patterns for different cell types
3. Reference Free Approaches – Matrix Decompositions
  - Singular value decomposition (Houseman, *Bioinformatics* 2014)
  - Linear model w/ principal components (EWASher; ReFACTor)
  - Surrogate Variable Analysis (SVA), Remove Unwanted Variation (RUv) – more general purpose
  - New RUvM – specific for methylation array platforms (Maksimovic et al., (2015) *Nucleic Acids Research*)

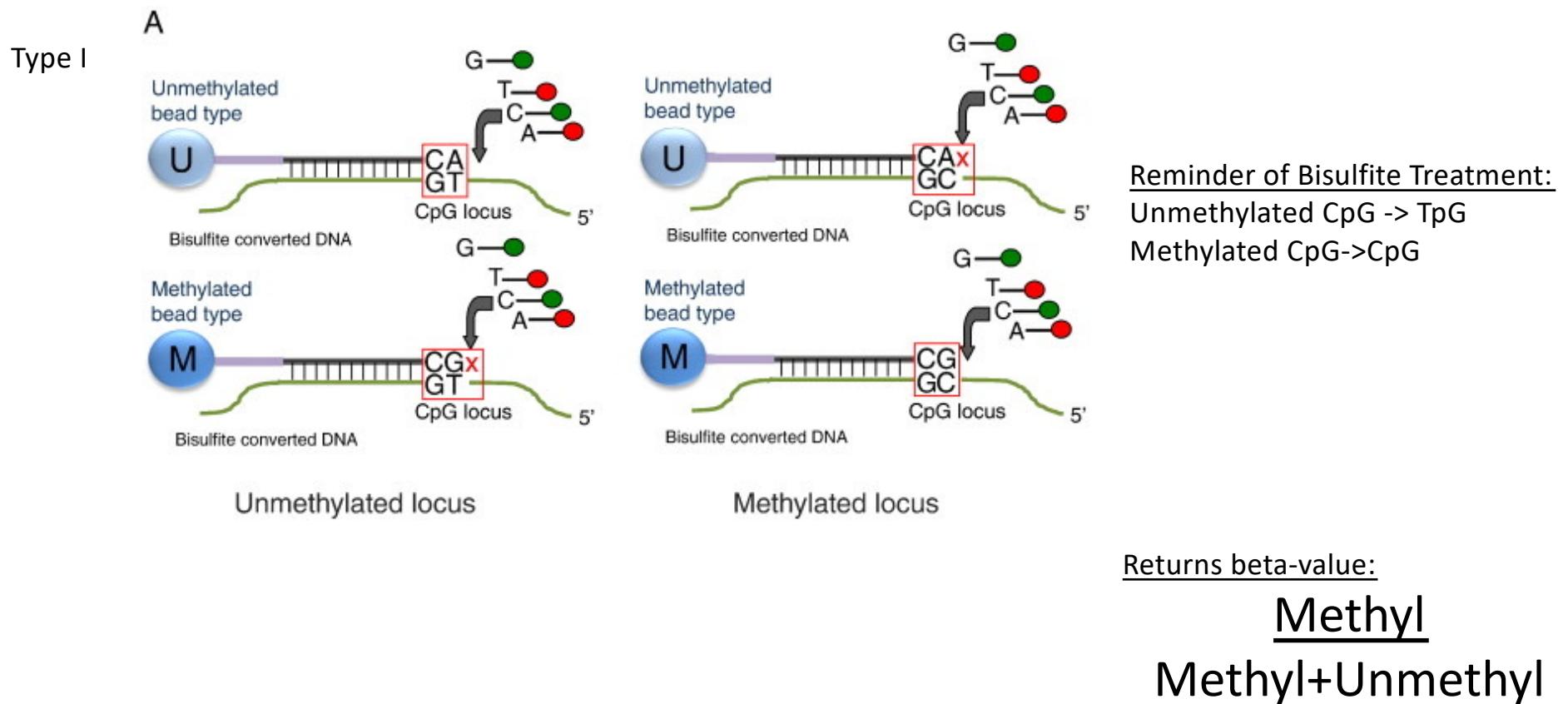
# Outline

- Introduction
- Example
- Methylome Profiling
- Technologies (Array, Sequencing) and Data Processing
- Analysis and Resources

# Arrays

# Array Based - Illumina Infinium Assay

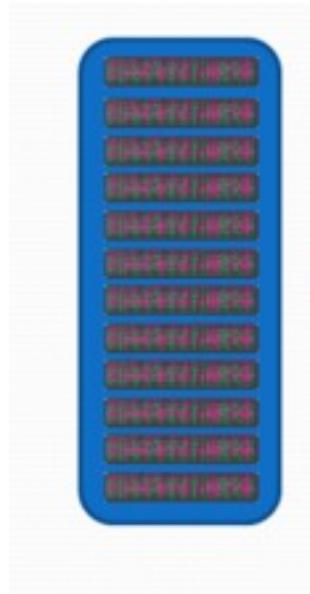
- Illumina 27K (2008)
  - 27,578 individual CpG sites (14,495 genes)
- Illumina 450K (2011)
  - 485,512 individual CpG sites (99% of RefSeq genes, ~17 CpG/gene)
  - 96% of CpG Islands, many shores, shelves, etc.
  - 1.5% of CpG in the genome
- Illumina EPIC (2016)
  - ~850,000
  - Most of 450K; more “open sea”, enhancer regions, transcription factor binding sites, open chromatin



# Data Processing – Image & Background

## Image Analysis

- Illumina GenomeStudio
- Returns for each site
  - average  $M$
  - average  $U$
  - average  $\beta = M/(M+U)$



## Background Correction

- Correct for each color – background noise
- Negative control probes, low intensity values, opposite color channel
- Deconvolution methods (separate signal from background)

# Data Processing - Filter (or flag) Probes

- High intensity probes
- Detection value  $P > 0.05$
- Probes with SNPs
- Cross-reactive probes

# Detection p-values

- Signal for probe is greater than average signal from the negative controls
- Estimate normal distribution (null) based on negative controls
- Detection p-value =  $1 - \Phi[(x - \mu_{\text{neg}})/\sigma_{\text{neg}}]$ 
  - $\Phi$  normal cumulative distribution
  - $x$  = sum of two beads (Type I) or sum of two-color intensities (Type II)
- Remove probes with detection p-value > 0.05

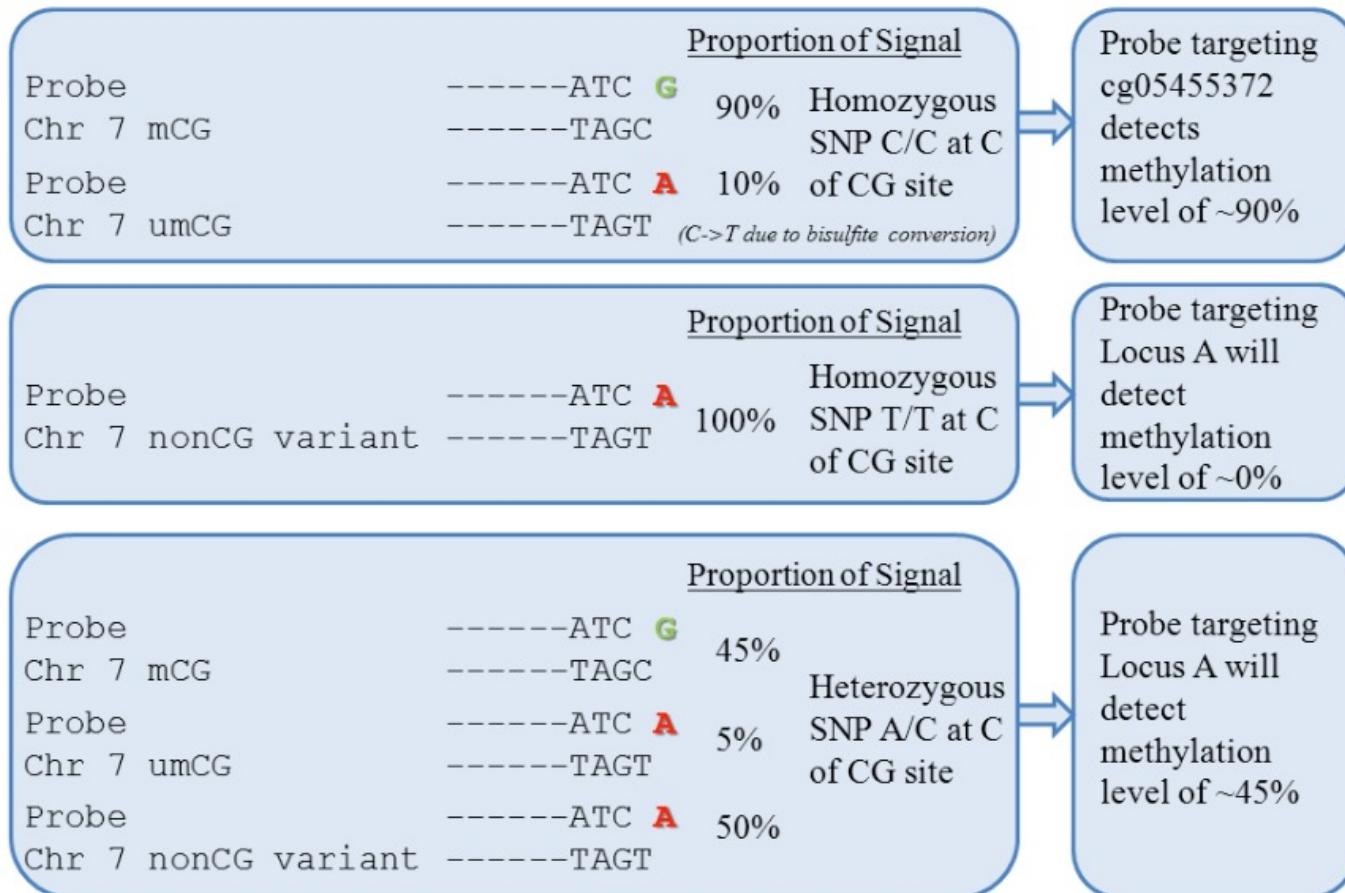
*Newer methods to estimate background/null: Heiss & Just (2019) Clinical Epigenetics,  
SeSAMe – Zhou et al. (2018) Nucleic Acids Research*

# Probes with SNPs

- Probes containing SNPs (single nucleotide polymorphism)
  - May detect C/T polymorphisms in the genome (genetic variation)
  - 9.4% of the Infinium I probes and 15.5% of the Infinium II probes
  - Majority of these SNPs are rare with low alternative allele frequencies

### Schematic representation of polymorphic CpG

Probe targets cg05455372 - a polymorphic CpG cytosine (rs2863984) (as shown in Figure 2)



Chen et al. (2013) Epigenetics

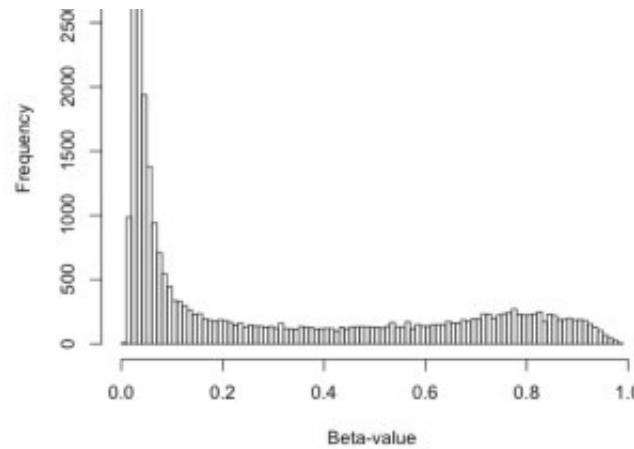
# Cross-Reactive Probes

- Cross-Reactivity
  - Bisulfite treatment
  - Initially ‘4-letter’ genome (A,C,T,G) to approx. ‘3-letter’ genome
  - Methylated C’s don’t change (~3.5% of C’s)
  - More likely to co-hybridize at other locations in the genome
  - 8.6-25% of probes may be non-specific (remove)

# Data Processing - Summarization

Beta-value

$$\text{Beta}_i = \frac{\max(y_{i,methy}, 0)}{\max(y_{i,unmethy}, 0) + \max(y_{i,methy}, 0) + \alpha}$$

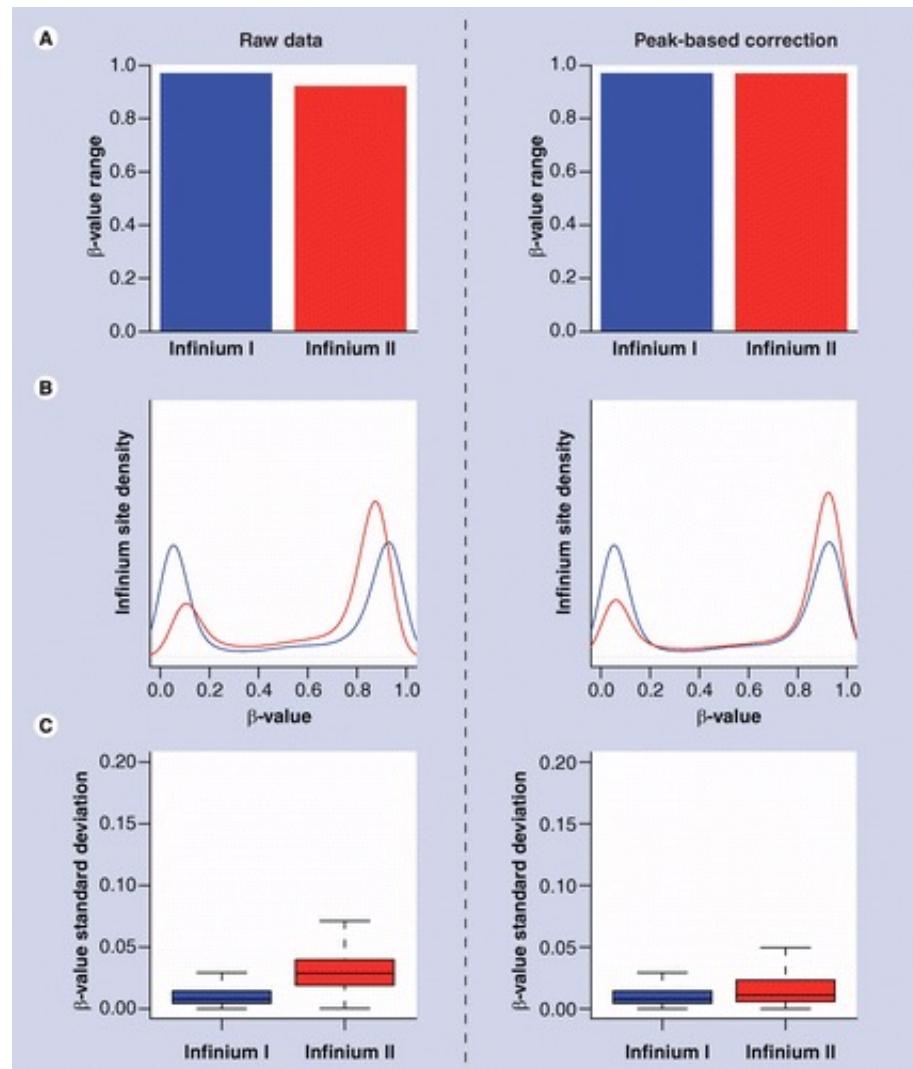


M-value

Analyze with M value, also report  $\beta$  value

# Data Processing - Normalization

- Within-array (Type I vs II probe differences)
- Across-array (methods take advantage of negative control probes)



Dedeurwaerder et al., (2011) Future Medicine

# Sequencing

# Pull-down Methods

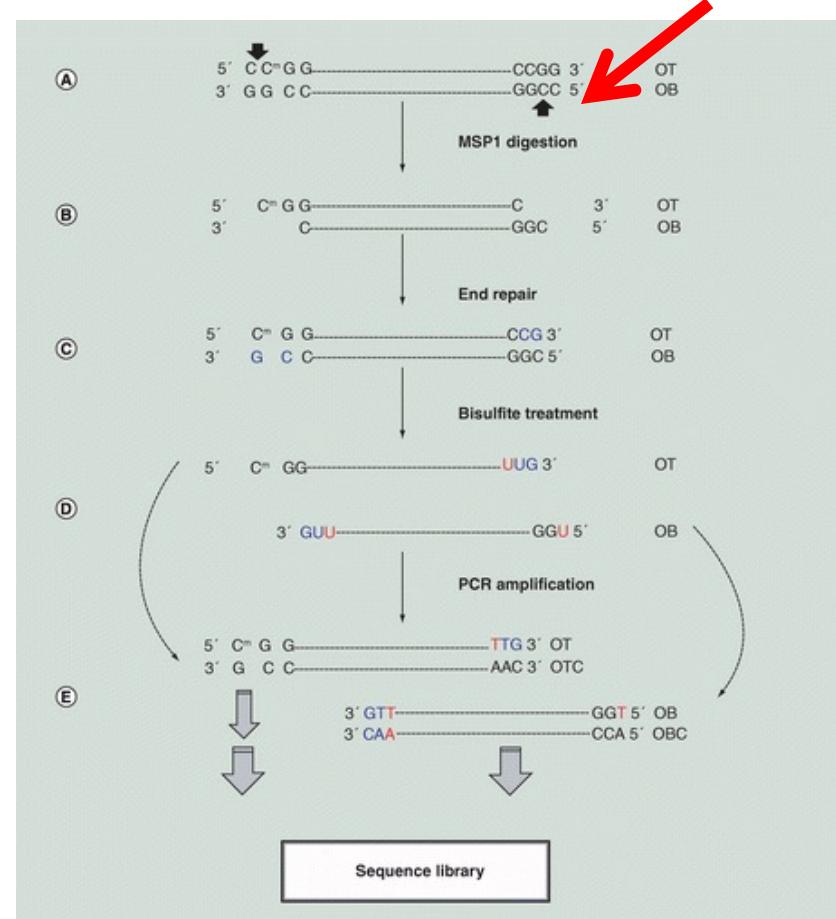
- In addition to pre-treatment (e.g., bisulfite conversion), for sequencing need to consider the coverage of the methylome (pull-down methods)
- Whole Genome – capturing all CpGs
- Reduced Representation
  - Targets CpGs in CpG rich regions by MSP1 digestion
- Methyl Capture
  - Capture specific genomic regions by kits
  - Agilent SureSelect Human Methyl-Seq
  - NimbleGen SeqCap Epi Enrichment System

# Whole Genome Bisulfite Sequencing (WGBS)

- Targets entire genome – sheared into fragments
- ~28 million CpGs in the human genome
- For one sample to get 30x coverage
  - Need 1B 100bp end reads
  - 4 lanes in Hi Seq, 1 NovaSeq
- In practice some sites have low coverage (1-10X) – estimates not reliable
  - “At the current recommended coverage of 30, the standard deviation (SD) observed for WGBS was 2-3X fold higher than that on the methylation array. This suggests that to achieve a level of precision that is broadly similar to that observed in methylation array, a coverage of at least **100x** will be necessary..” *Zhou et al., Scientific Reports 2019*
- Too costly for most studies (common with big consortiums: e.g., ENCODE, NIH Epigenomics Roadmap)

# Reduced Representation Bisulfite Sequencing (RRBS)

- Pull subset of DNA to be sequenced
- Restriction endonuclease (*Msp*1), cuts DNA into fragments
- Cuts between the 2 C's in “CCGG” sites
  - Enriched in CpG islands, promoters
- Not affected by methylation status
- 80% of CpG islands & 60% of promoter regions
- “CCGG” motif not uniform, some genes have more/less
- More cost-efficient than WGBS

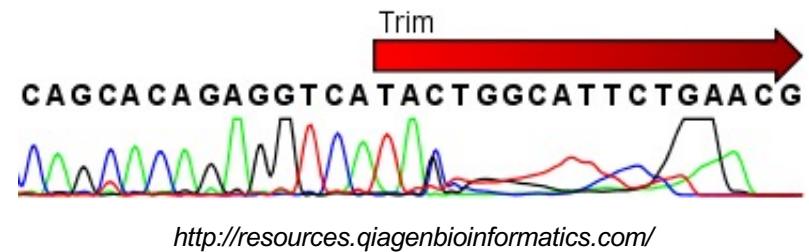


Sun et al., (2015) Epigenomics

# Methyl-Capture Methods

- Need less genomic material
- Agilent SureSelect:
  - 84Mb of the genome, 3.7M CpGs
  - Targets genomic regions where methylation known to impact gene regulation
  - CpG islands, shores, promoters, ....
- Nimblegen SeqCap:
  - 5.5M CpGs
  - Uses probes to hybridize both strands of fully, partially, non-methylated derivatives – pool them together

# Data Processing - Trimming



- Standard sequencing biases
  - Sequencing into the adaptor, 3' low sequencing quality
- Other trimming specific to methylation
  - Cause over/under estimation of methylation rates
  - Overhang end-repair
    - RRBS incorporates artificial unmethylated CG (lower methylation rates)
  - Bisulfite conversion failure known to be enriched at 5' end
    - So unmethylated C's not converted to T's (higher methylation rates)
  - Account for position for trimming

# Data Processing - Mapping

## Challenges of bisulfite conversion

- Reads are not complementary to reference genome
- Need to be able align T (read) to C (ref. genome), but not reverse
- Account for C->T conversions

Tumor DNA

Bisulfite conversion

PCR

Various technologies

GACATG[mC]GACG

GA[U]ATG[C]GACG

GATATG[C]GACG

Quantification of 'methylated' sequences

Wellner (2010) Nature Reviews Neurology

# Mapping Strategies

## 1. Wild Cards

- Substitute C with Y
  - Y= C or T
- Reads with both C's and T's can be aligned
- Better coverage, but inflates methylation levels
- e.g., GSNAP, BSMAP

### a Setup of the example

Genomic DNA sequence CCGATGATGT CGCTGACCGCACGA  
DNA methylation level 100% 50% 50% 0%

DNA fragmentation, selective conversion of unmethylated Cs into Ts, DNA sequencing

Bisulphite-sequencing reads ACGT, ATCA, ATGA, ATGT,  
TCGA, TCGA, TCGT, TTGT

### b Wild-card alignment

Reference sequence YYGATGATGTYGYTGAYGYAYGA  
Read alignment TCGA  
TCGA  
TCGT  
ATGT  
ACGT  
ATGT  
ATGA  
ATGA

DNA methylation level 100% 50% 100% 0%

Bock (2012) Nature Review Genetics

# Mapping Strategies

## 2. Three-letters

- Change all C's to T's in genome and reads (forward strand);  
Change all G's to A's (reverse strand)
- Reduced sequence complexity, use aligners w/ lower mappability
- Remove reads mapping to multiple locations
- e.g., BISMARK, BS Seeker 2, MethylCoder, BRAT-BW

### a Setup of the example

Genomic DNA sequence CCGATGATGT CGCTGACCGCACGA  
DNA methylation level 100% 50% 50% 0%

DNA fragmentation, selective conversion of unmethylated Cs into Ts, DNA sequencing

Bisulphite-sequencing reads ACGT, ATCA, ATGA, ATGT,  
TCGA, TCGA, TCGT, TTGT

### c Three-letter alignment

Reference sequence TTGATGATGT TGTTGATGTATGA  
Read alignment TtGA TtGA TtGA  
TtGA TtGA TtGA  
TtGT TtGT TtGT  
AtGT AtGT AtGT  
ATGT ATGT ATGT  
ATGA ATGA ATGA

Bock (2012) Nature Review Genetics

# Data Processing –QC/Filtering

- Standard read and alignment QC
  - Quality score, multiple mappings, etc.
- Some specifics to methylation:
  - Reads expected to be T rich, C poor
  - Examine % of non-CpG C conversion to T (unless relevant to particular tissues/species)
  - Remove CpGs with low coverage (< 10x)
  - Quality score cutoffs at CpG site

# Data Processing – Methylation Calling

- Counts

Reference AT**C**G**C**TTC**G**ACT

Reads

AT <b>C</b> G <b>T</b> TTTC <b>G</b> ATT
AT <b>C</b> G <b>T</b> TTTC <b>G</b> ATT
AT <b>T</b> G <b>T</b> TTTC <b>G</b> ATT
AT <b>T</b> G <b>T</b> TTTC <b>G</b> ATT
AT <b>T</b> G <b>T</b> TTTC <b>G</b> ATT
AT <b>T</b> G <b>T</b> TTTC <b>G</b> ATT
AT <b>T</b> G <b>T</b> TTTC <b>T</b> GATT
AT <b>T</b> G <b>T</b> TTTC <b>T</b> GATT

- Percentage

Methylation  
(red sites)       $2/8 = 25\%$

$$\frac{\text{readcounts}(\text{methylated})}{\text{readcounts}(\text{methylated}) + \text{readcounts}(\text{unmethylated})}$$

# Outline

- Introduction
- Example
- Methylome Profiling
- Technologies (Array, Sequencing) and Data Processing
- Analysis and Resources

# Differentially Methylated Position (DMP)

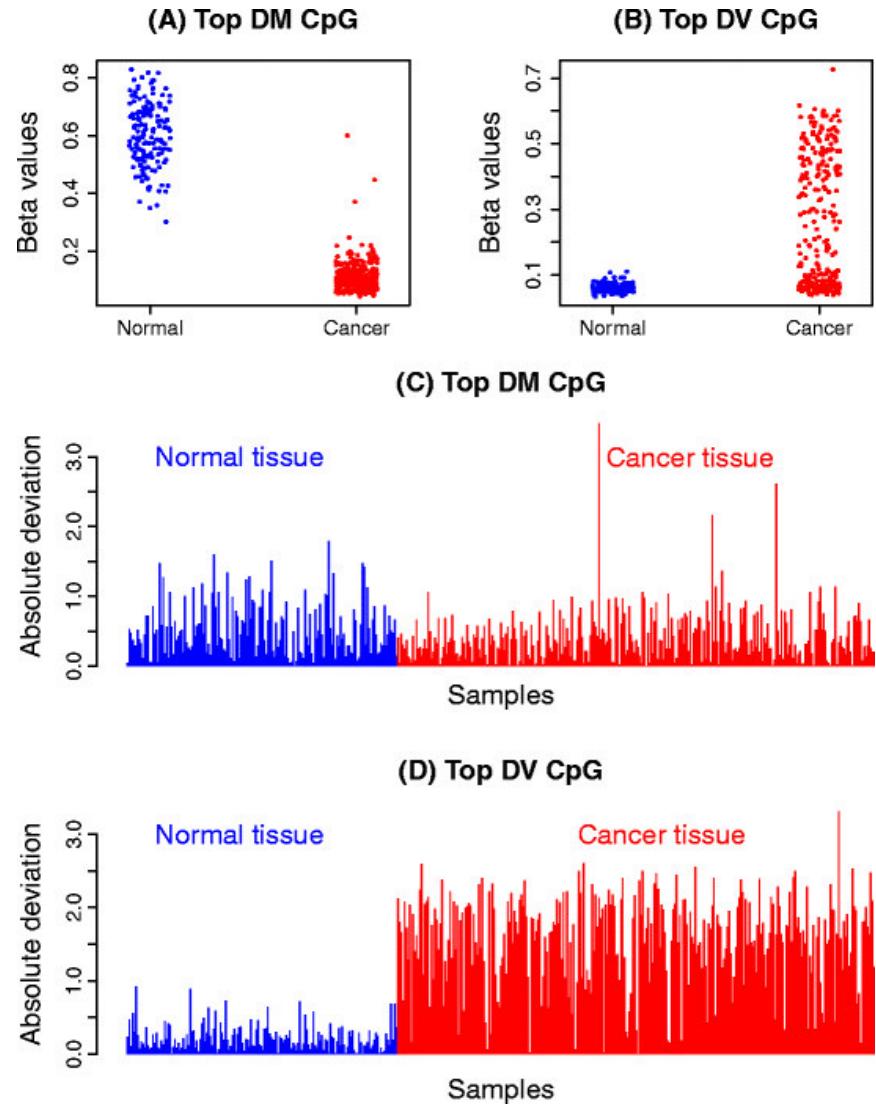
Single probe/site analysis

- M-value: Gaussian models (t-test); Non-parametric (Wilcoxon test, Kolmogorov-Smirnoff), Empirical Bayes (limma)
- $\beta$ -value: Beta regression (between 0-1)
  - Heteroscedastic (variance at extremes)
- Counts (sequencing only): careful due to unequal coverage for samples & over-dispersion; Beta-Binomial methods address these problems

**Don't forget multiple testing!**

# Differentially Variable Position (DVP)

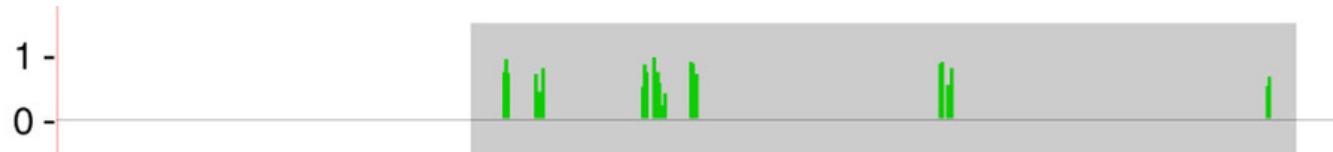
Levene's Test for Equal Variance  
*Phipson & Oshlack (2014) Genome Biology*



# Differentially Methylated Region (DMR)

## Multiple probe/site analysis

- Assume site close together behave similarly
- Improves statistical power; less sensitive to SNPs, outliers, bad probes
- Not ideal for 25% probes that are isolated



1. Group sites -> test for significance
2. Test for significance -> group sites

*Kuleshov et al., (2014) Nature Biotechnology*

# DMR: Group Sites First

- Group by predefined regions (e.g., CpG islands, shores, UTRs) or by group of sites that are highly correlated
- Smoothing may be used
  - Methylation of adjacent CpGs may be similar - borrow information from neighbors
  - Smooth raw methylation data
  - Avoid missing values, more robust to sequencing errors, helps with low coverage



Hansen et al., (2012) Genome Biology

# DMR: Test Sites First

- Group significant sites
  - DMRcate, ProbeLasso, Bumphunter, Comb-p, BiSeq, BSmooth
  - Modeling correlations among sites
  - Cutoffs:
    - distance between sites
    - p-value cutoff
    - # of sites in DMR

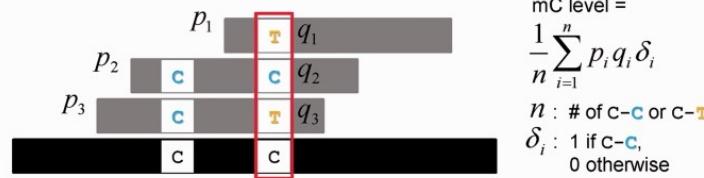
# Hidden Markov Models (HMM) for DMR Detection

e.g., Bisulfighter, HMM-Fisher, HMM-DM

Saito et al., (2014) Nucleic Acids Research

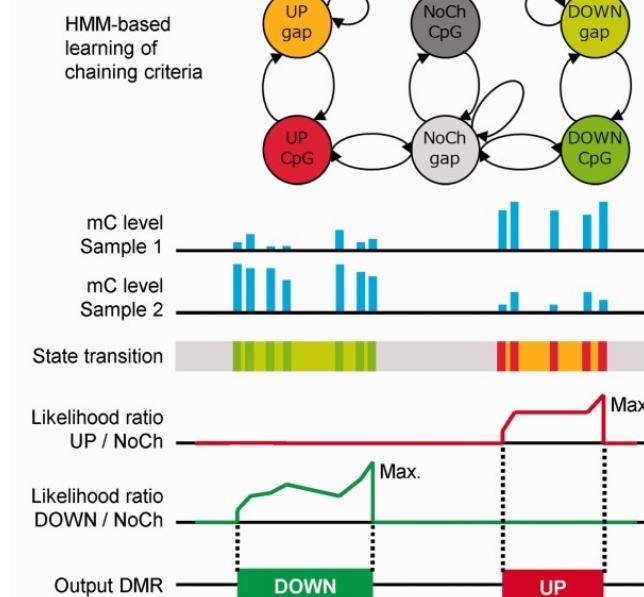
## (a) mC calling

c-C: methylated c  
c-T: unmethylated c  
 $p_i$ : Alignment probability  
 $q_i$ : Base quality



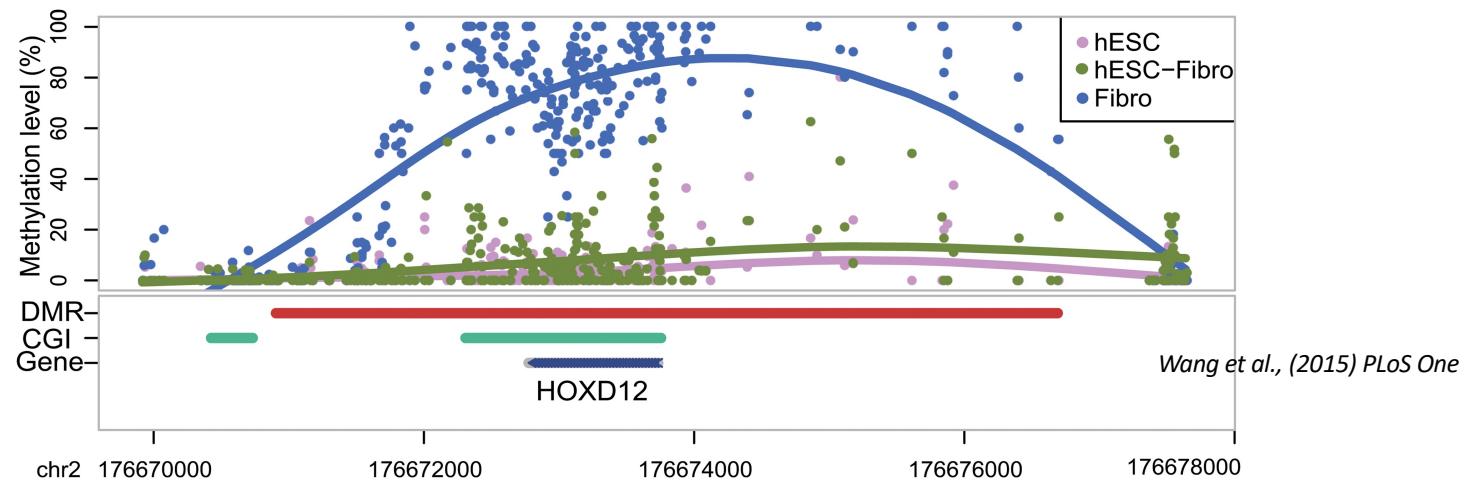
1. Align reads allowing c-C matches and c-T mismatches
2. Evaluate alignment probability based on read quality and multi-locus mapping
3. Discard small-probability alignments
4. Estimate the mC level for each C position

## (b) DMR detection



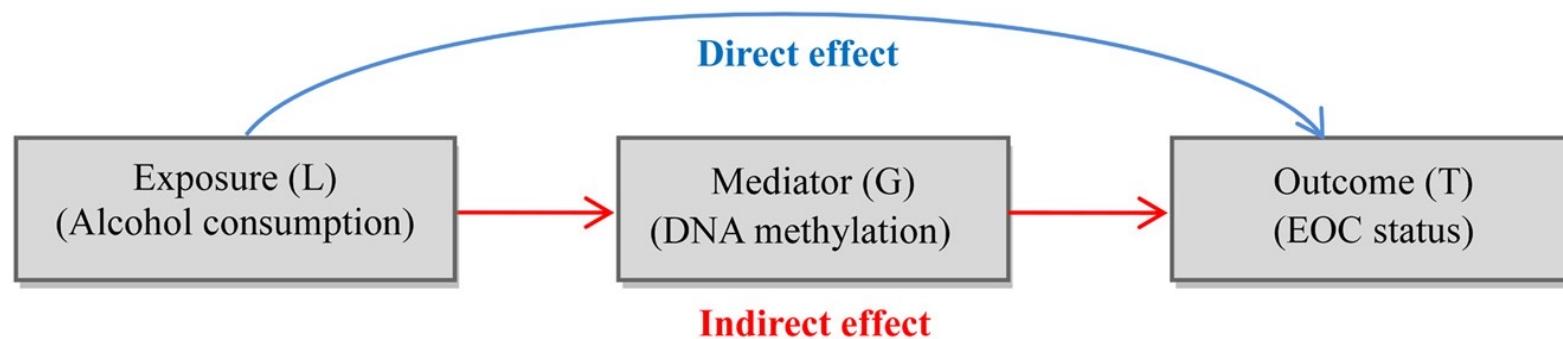
# Interpretation – Downstream Analyses

- Hyper/hypo methylation
- Location of DMP/DMRs in genome/annotation + Visualization
- Search for over-representation/enrichment of gene sets/pathways/regulatory regions near DMP/DMRs



# Interpretation – Downstream Analyses

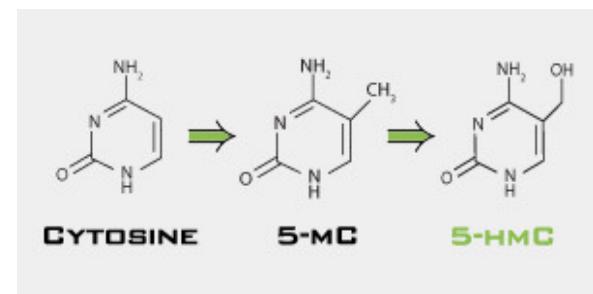
- Integration with other –omics (gene expression)
- Methylation QTL (SNPs ~ methylation)
- Mediation analysis (e.g., does methylation mediate the relationship between an exposure and an outcome)



Wu et al., (2018) Journal of Human Genetics

# Extensions

- Mapping for reads when no reference genome (e.g., RefFreeDMA)
- SNP calling (e.g., Bis-SNP, BS-SNPer)
- Reconstructing cell type methylation patterns (e.g., methylFlow)
- Single cell methylome (Farlik et al., Cell Rep 2015)
- DNA hydroxymethylation (5hmC vs 5mC)



<http://www.dnahydroxymethylation.com/>

# Resources - Epigenome

The image displays two web-based resources for epigenomic research:

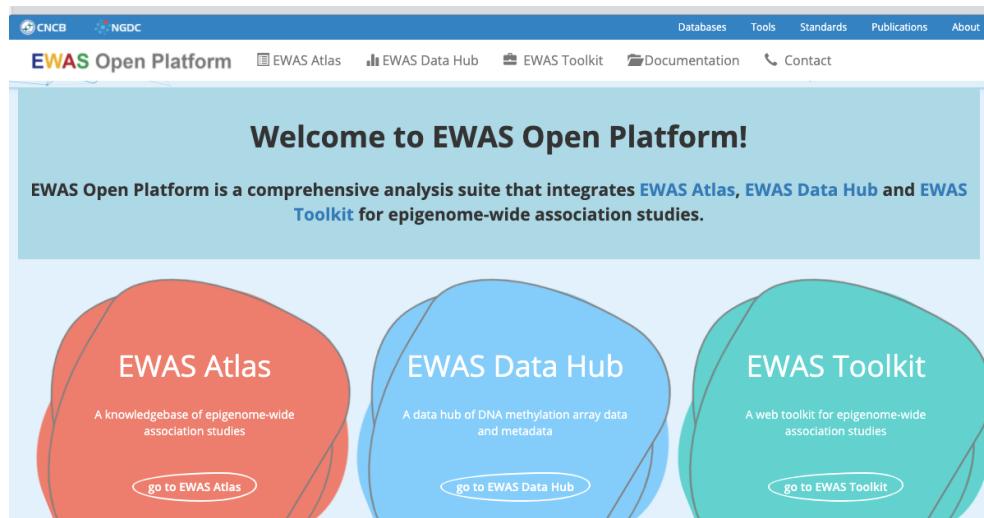
**NIH Roadmap Epigenomics Project (Top Left):**

- Header:** ROADMAP epigenomics PROJECT
- Navigation:** HOME, PARTICIPANTS, BROWSE DATA, PROTOCOLS, COMPLETE EPIGENOMES, TOOLS, PUBLICATIONS.
- Search:** Search: [ ] GO
- Content:** Overview, Project Data, Mapping Centers, Protocols & Standards, Publications.
- Diagram:** A detailed diagram of chromatin structure showing Genes, RNA, Chromatin, DNase I hypersensitive sites, DNA methylation, and Histone Modifications.
- Credit:** NIH Roadmap Epigenomics Mapping Consortium

**ENCODE Encyclopedia of DNA Elements (Bottom Right):**

- Header:** ENCODE Data Encyclopedia Materials & Methods Help
- Search:** Search... Sign in / Create account
- Section:** ENCODE: Encyclopedia of DNA Elements
- Diagram:** A complex diagram illustrating the ENCODE project's focus on DNA elements. It shows a chromosome with Hypersensitive Sites, RNA polymerase, and various modifications like CH<sub>3</sub> and CH<sub>3</sub>CO. Below, a grid maps these to 3D Chromatin Structure, Chromatin Accessibility, Chromatin Interactions, Methyome, Chromatin Modification, Transcriptome, and RNA Binding. Arrows point from these categories to specific genomic features: Enhancer-Like Elements, Promoter-Like Elements, Genes, and Transcripts.
- Links:** About ENCODE Project, Getting Started, Experiments, Search ENCODE portal, About ENCODE Encyclopedia, candidate Cis-Regulatory Elements, Search for candidate Cis-Regulatory Elements, Hosted by SCREEN, Human GRCh38 Q, Mouse mm10 Q, Visit hg19 site, Help

# Resources - Methylome



The screenshot shows the homepage of the EWAS Open Platform. At the top, there are logos for CNCB and NGDC, followed by a navigation bar with links to Databases, Tools, Standards, Publications, and About. Below the navigation bar, the main header reads "Welcome to EWAS Open Platform!". A sub-header below it states: "EWAS Open Platform is a comprehensive analysis suite that integrates **EWAS Atlas**, **EWAS Data Hub** and **EWAS Toolkit** for epigenome-wide association studies." The page features three large, rounded rectangular sections representing the integrated tools: "EWAS Atlas" (red), "EWAS Data Hub" (blue), and "EWAS Toolkit" (teal). Each section contains a brief description and a "go to" button.

**EWAS Open Platform**

Databases Tools Standards Publications About

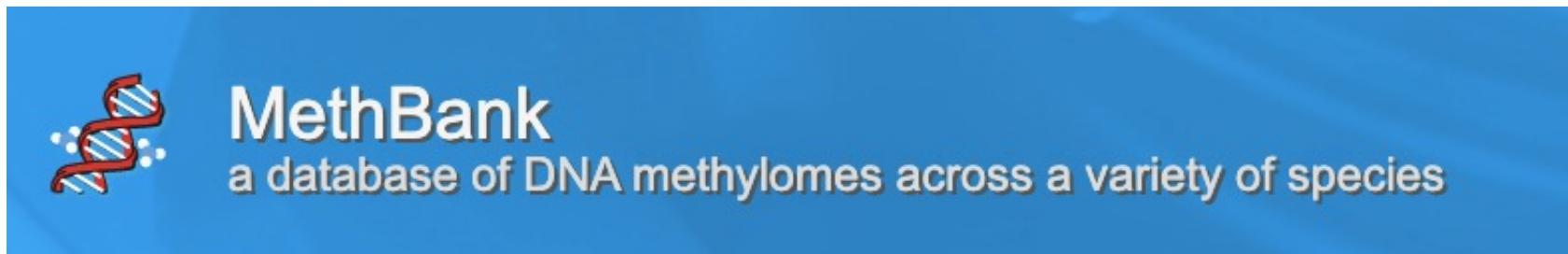
Welcome to EWAS Open Platform!

EWAS Open Platform is a comprehensive analysis suite that integrates **EWAS Atlas**, **EWAS Data Hub** and **EWAS Toolkit** for epigenome-wide association studies.

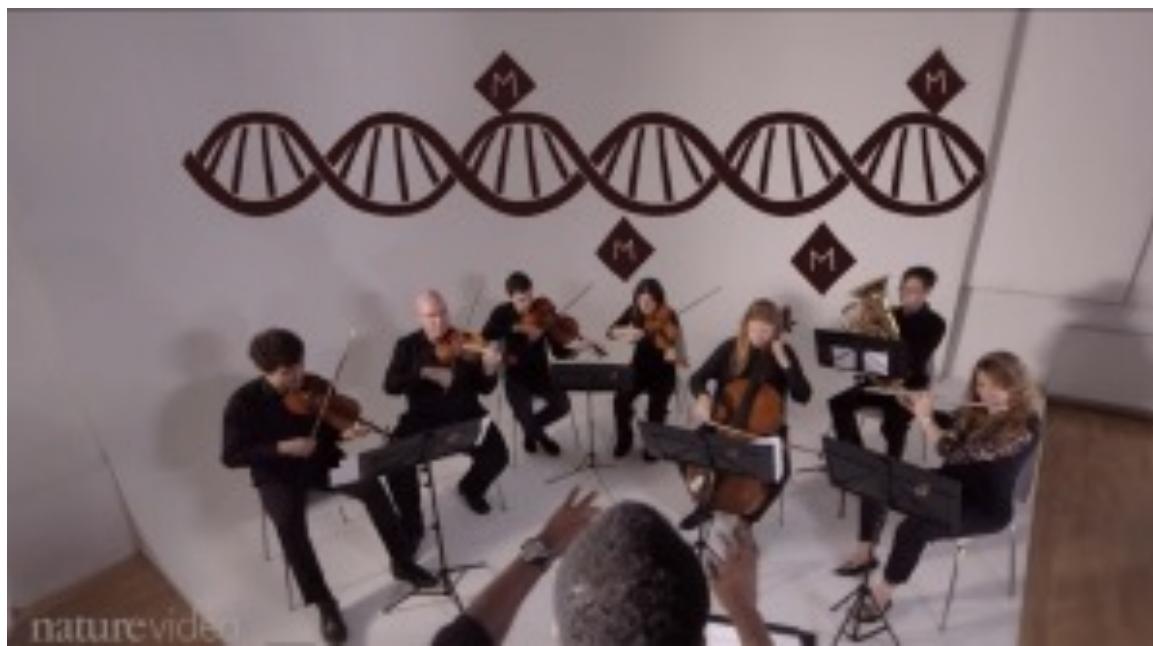
**EWAS Atlas**  
A knowledgebase of epigenome-wide association studies  
[go to EWAS Atlas](#)

**EWAS Data Hub**  
A data hub of DNA methylation array data and metadata  
[go to EWAS Data Hub](#)

**EWAS Toolkit**  
A web toolkit for epigenome-wide association studies  
[go to EWAS Toolkit](#)



# “Epigenome: The symphony in your cells”



*Smith (2015) Nature*