

INTRODUCTION TO EPIGENOME-WIDE ASSOCIATION STUDIES (EWAS)

2. PRE-PROCESSING OF DNA METHYLATION DATA (THEORY)

EPIGENOME-WIDE ASSOCIATION STUDY (EWAS)

Workflow

1. Scientific question
2. Study population
3. Biological sample
4. DNA methylation data acquisition
5. Quality control of DNA methylation data
6. Epigenome-wide association study (EWAS)
7. Meta-EWAS or replication / validation
8. Biological interpretation

EPIGENOME-WIDE ASSOCIATION STUDY (EWAS)

Workflow

1. Scientific question

- Association study: Which are the CpGs (pathways) associated with tobacco smoking?
- Prediction study: Which is the smoking status of an individual (based on their methylation)?

2. Study population

3. Biological sample

4. DNA methylation data acquisition

5. Quality control of DNA methylation data

6. Epigenome-wide association study (EWAS)

7. Meta-EWAS or replication / validation

8. Biological interpretation



EPIGENOME-WIDE ASSOCIATION STUDY (EWAS)

Workflow

1. Scientific question
2. **Study population**
 - Observational study
 - Case-control study
 - Cohort study
 - Experimental study
3. Biological sample
4. DNA methylation data acquisition
5. Quality control of DNA methylation data
6. Epigenome-wide association study (EWAS)
7. Meta-EWAS or replication / validation
8. Biological interpretation





	Key advantage	Key disadvantage
Case versus control (singletons) 	Many cohorts exist	Cannot easily control for environmental and genetic confounders
Families 	Can study potential inheritance	Few large cohorts of this type exist
Disease-discordant monozygotic twins 	Can control for genetics	Few large cohorts of this type exist
Prospectively sampled, longitudinal 	Can establish causality	Slow and difficult to establish

Figure 1 | The different types of sample cohorts that could be used in an epigenome-wide association study. Refer to the main text for a full discussion.

EPIGENOME-WIDE ASSOCIATION STUDY (EWAS)

Workflow

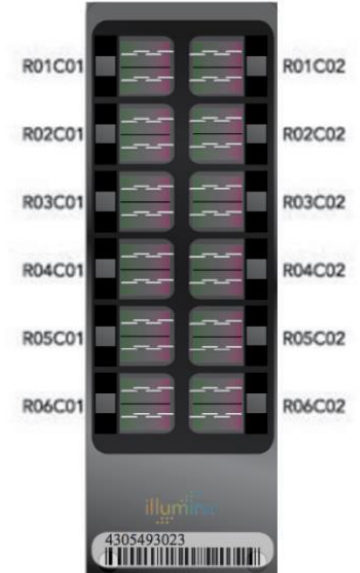
1. Scientific question
2. Study population
- 3. Biological sample**
4. DNA methylation data acquisition
5. Quality control of DNA methylation data
6. Epigenome-wide association study (EWAS)
7. Meta-EWAS or replication / validation
8. Biological interpretation



EPIGENOME-WIDE ASSOCIATION STUDY (EWAS)

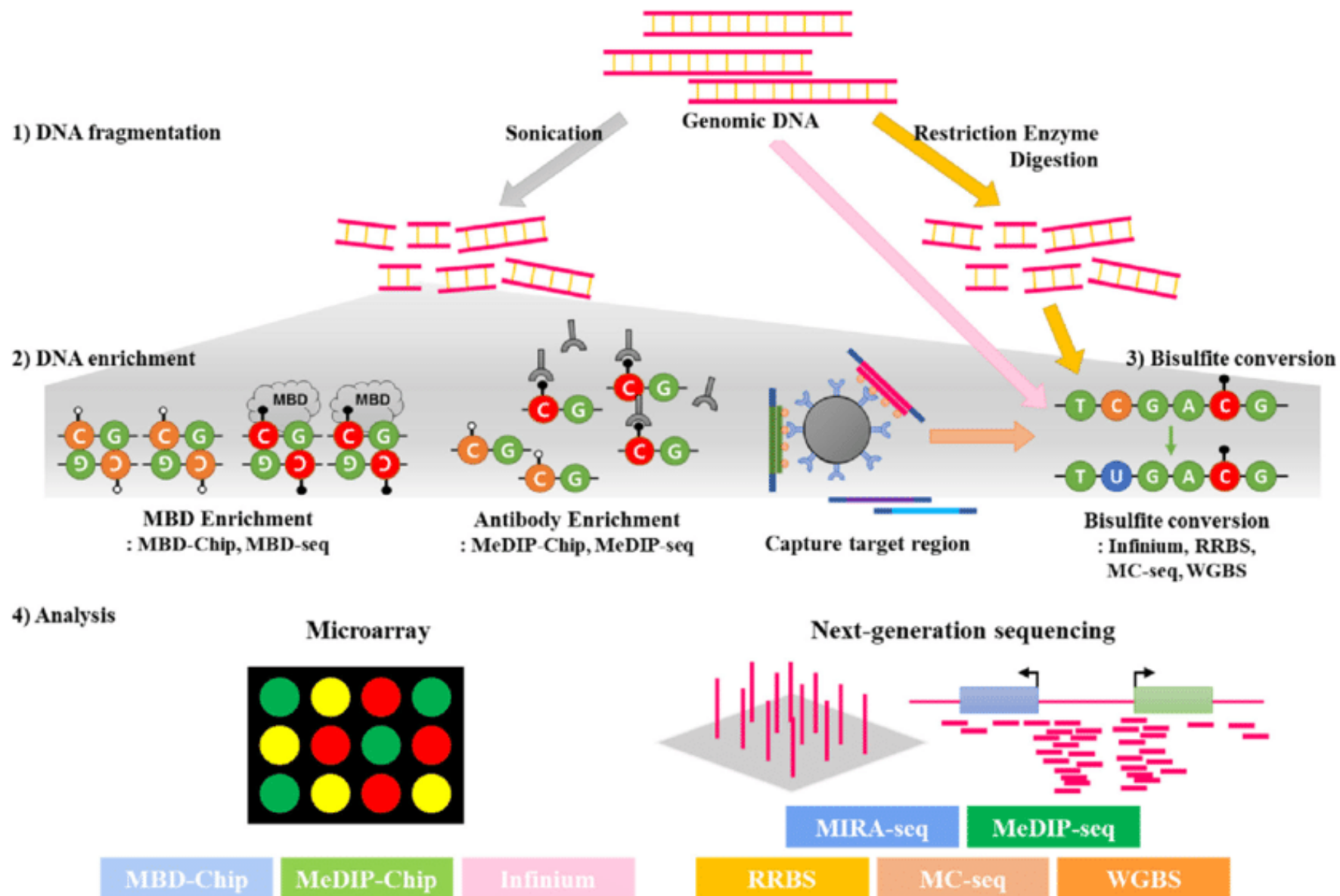
Workflow

1. Scientific question
2. Study population
3. Biological sample
- 4. DNA methylation data acquisition**
5. Quality control of DNA methylation data
6. Epigenome-wide association study (EWAS)
7. Meta-EWAS or replication / validation
8. Biological interpretation



5. DNA METHYLATION QUANTIFICATION

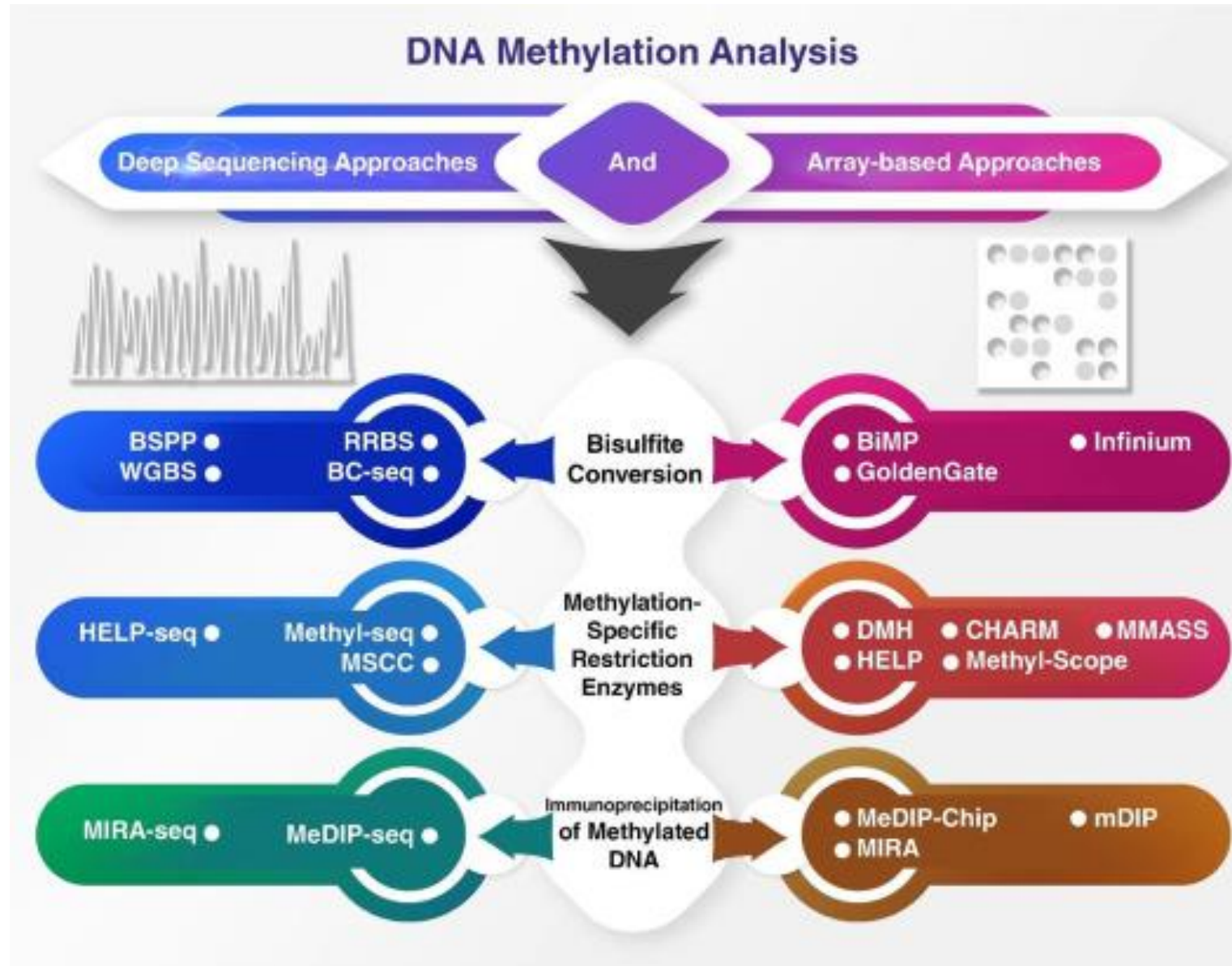
Methods



MBD: methyl-CpG-binding domain
MeDIP: methylated DNA immunoprecipitation
Infinium: Illumina Infinium 450/EPIC BeadChips
RRBS: reduced-representation bisulfite-sequencing
MC-seq: methyl-capture sequencing
WGBS: whole-genome bisulfite sequencing

5. DNA METHYLATION QUANTIFICATION

Methods



5. DNA METHYLATION QUANTIFICATION

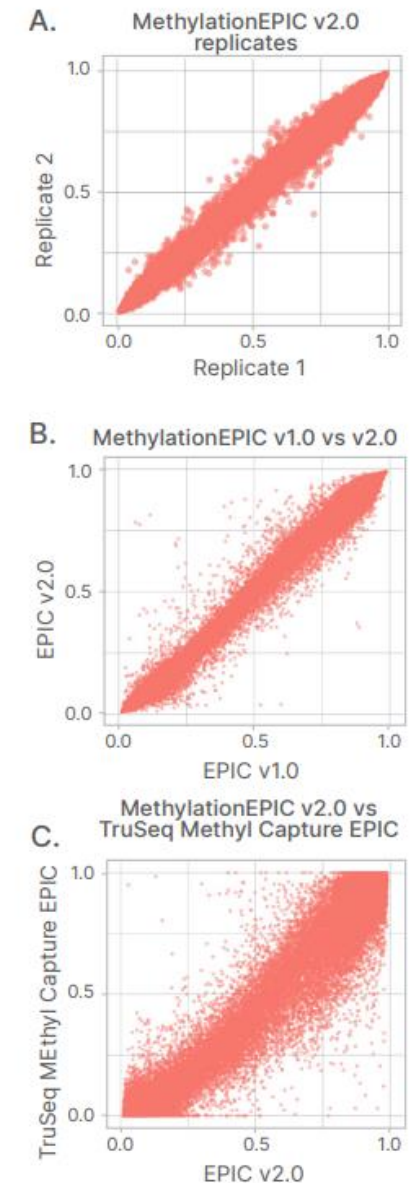
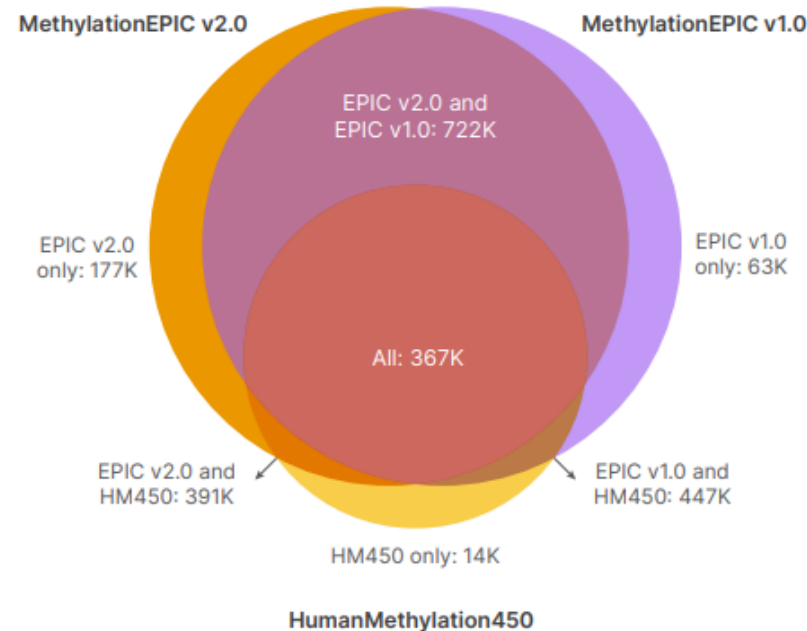
Illumina Infinium BeadChips

Advantages:

- Comprehensive genome-wide coverage
- Assay reproducibility
- User-friendly, streamlined workflow
- Available in many cohort studies

Arrays:

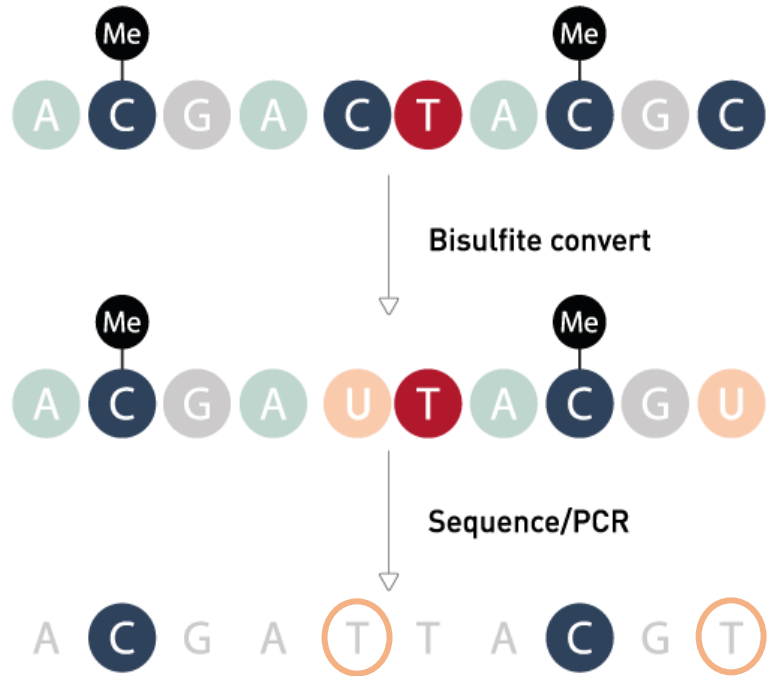
- GoldenGate: 1500 CpGs (cancer)
- 27K: 27K CpGs (promoters of 14K genes)
- 450K: 480K CpGs (27K + 99% RefSeq genes + others)
- EPIC v1: 850K CpGs (90% 450K + Regulatory elements)
- EPIC v2: 850K CpGs (90% 450K + Regulatory elements + cancer)



5. DNA METHYLATION QUANTIFICATION

Illumina Infinium technology

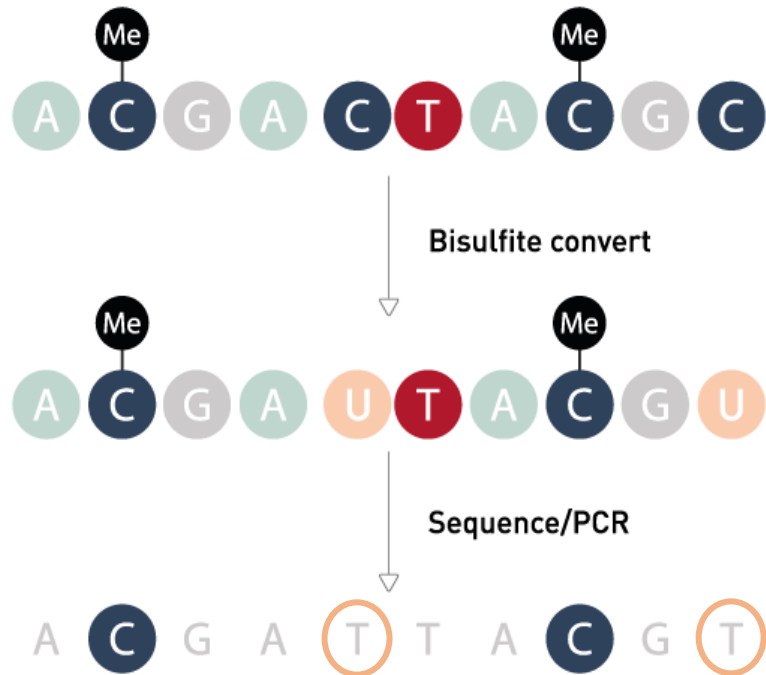
1) Bisulfite conversion



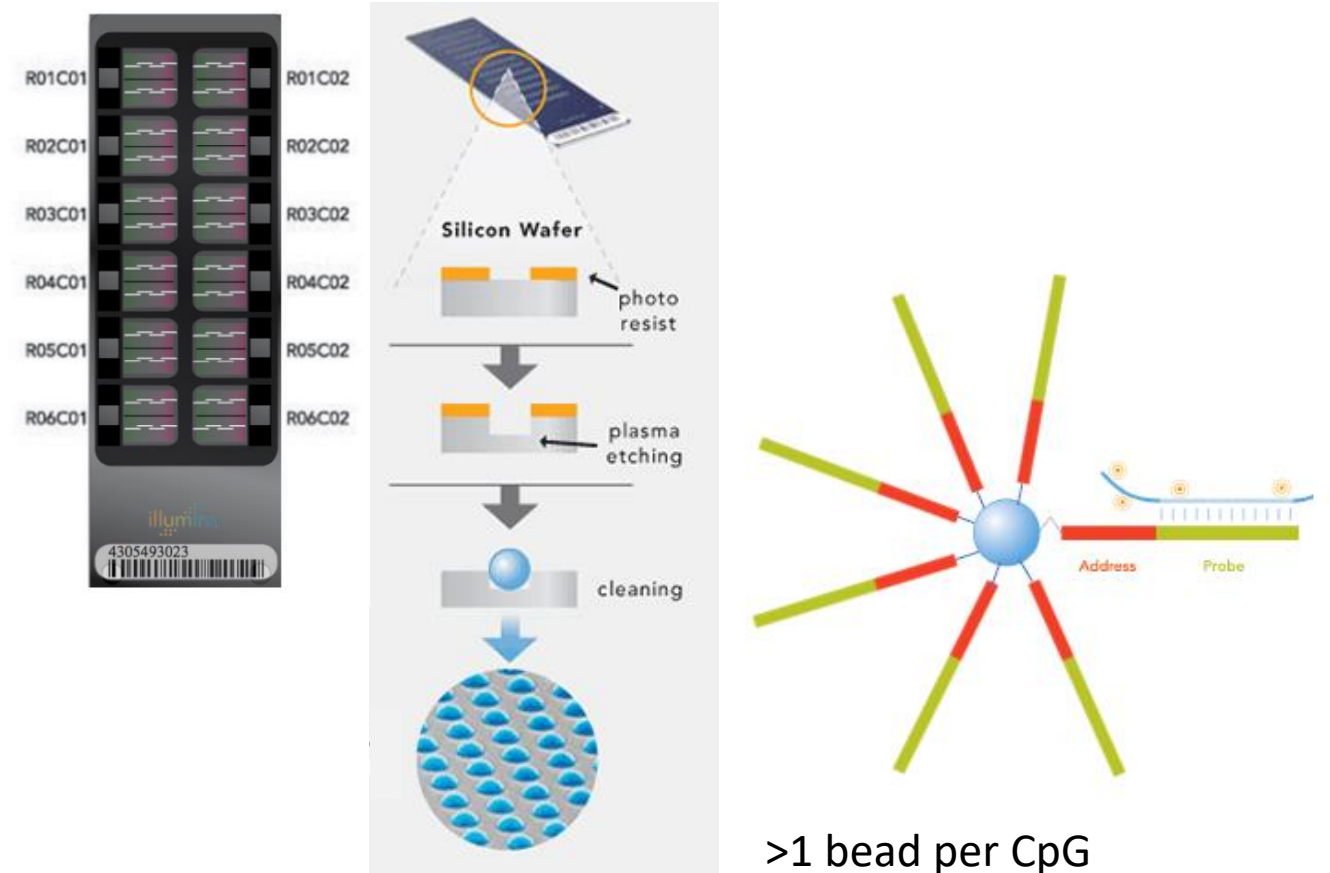
5. DNA METHYLATION QUANTIFICATION

Illumina Infinium technology

1) Bisulfite conversion



2) BeadChip hybridization

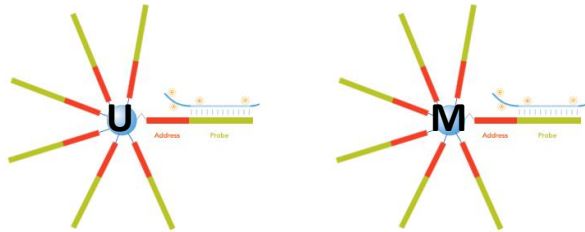


5. DNA METHYLATION QUANTIFICATION

u Unmethylated bead type m Methylated bead type CpG locus Bisulfite converted DNA

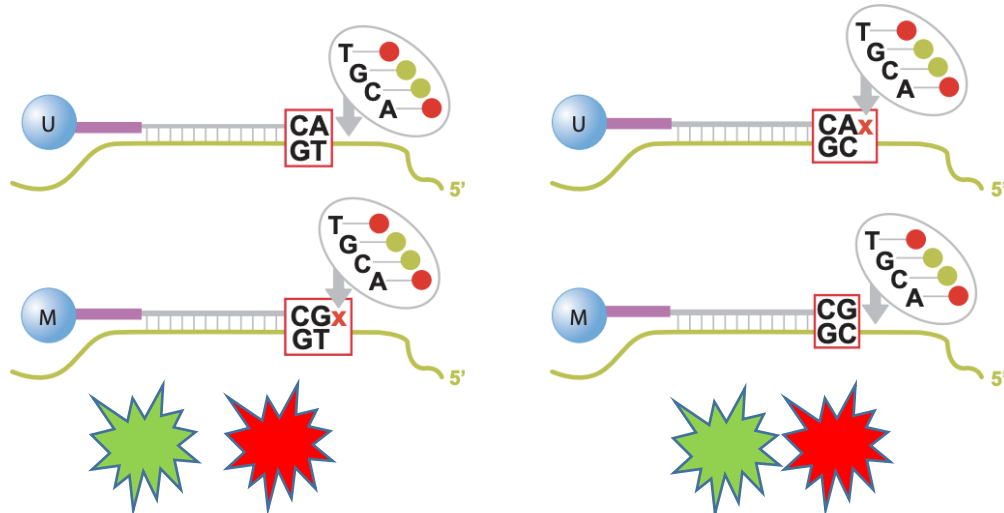
Illumina Infinium technology

Infinium I probes

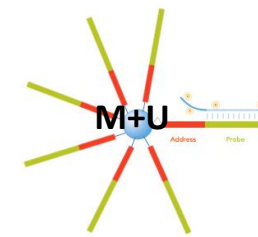


Unmethylated locus

Methylated locus

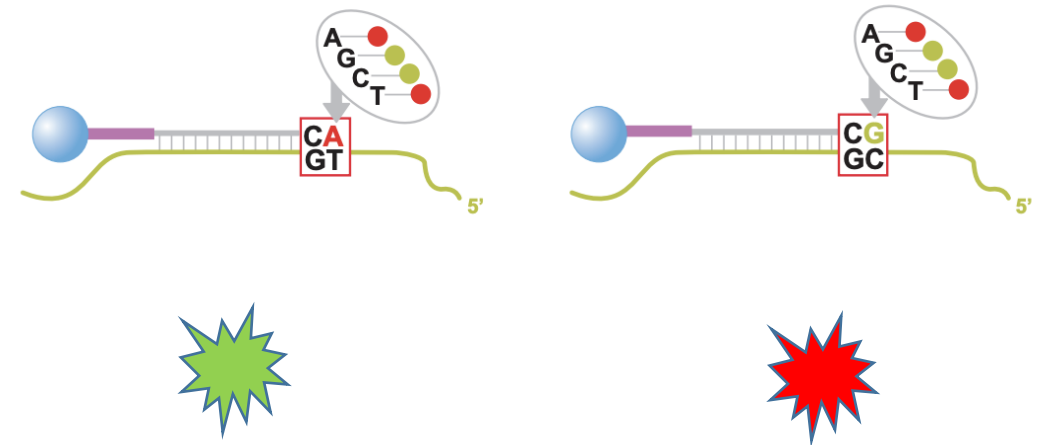


Infinium II probes



Unmethylated locus

Methylated locus



5. DNA METHYLATION QUANTIFICATION

Illumina Infinium probes

HumanMethylation450 array content.

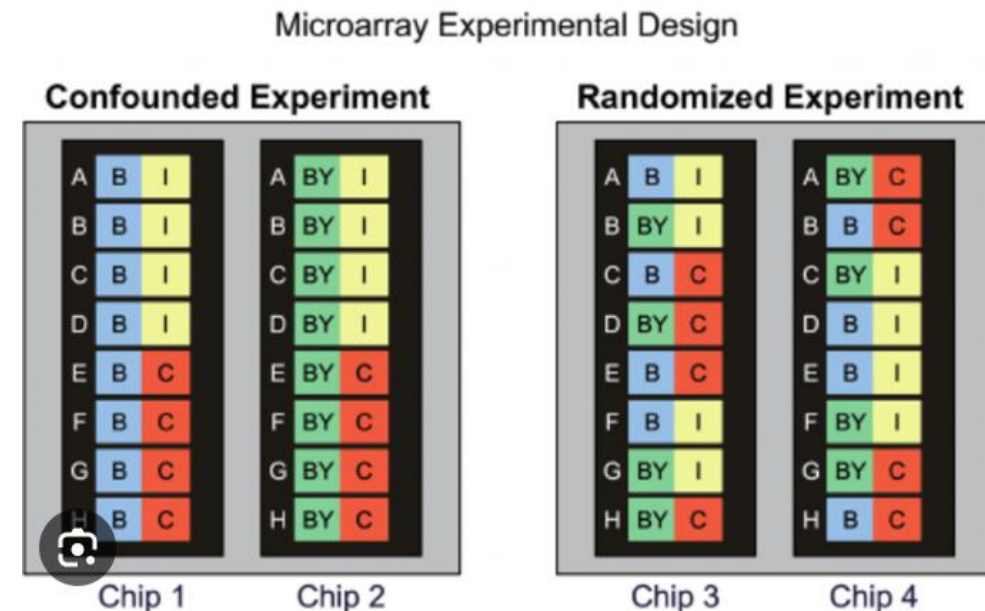
Feature type	Included on array
Total number of sites	485,577
RefSeq genes	21,231 (99%)
CpG islands	26,658 (96%)
CpG island shores (0–2 kb from CGI)	26,249 (92%)
CpG island shelves (2–4 kb from CGI)	24,018 (86%)
HMM islands ^a	62,600
FANTOM 4 promoters (High CpG content) ^a	9426
FANTOM 4 promoters (Low CpG content) ^a	2328
Differentially methylated regions (DMRs) ^a	16,232
Informatically-predicted enhancers ^a	80,538
DNase hypersensitive sites	59,916
Ensemble regulatory features ^a	47,257
Loci in MHC region	12,334
HumanMethylation27 loci	25,978
Non-CpG loci	3091

Category	Description	No. of Types	No. of Probes
Bisulfite Conversion	Methylation at a site known to be methylated	3	10
Normalisation	Randomly permuted bisulphite-converted sequences containing no CpGs; Determines system background	4	186
Staining	Efficiency and sensitivity of staining step	2	2
Extension	Extension efficiency of A, T, C, and G nucleotides from a hairpin probe	4	4
Hybridisation	Hybridisation efficiency using synthetic targets instead of amplified DNA	3	3
Target Removal	Efficiency of stripping step after extension reaction	1	2
Specificity	Methylation at non-polymorphic T sites	3	9
Non-polymorphic	Methylation at a base in a non-polymorphic region of the genome	4	4

5. DNA METHYLATION QUANTIFICATION

Experimental design

- Randomize samples across technical batch variables
 - Collection
 - Hospital
 - Time of the day
 - DNA extraction batch
 - Bisulfite conversion batch
 - Array
- > complete randomization
- Include replicates
- Include positive controls



EPIGENOME-WIDE ASSOCIATION STUDY (EWAS)

Workflow

1. Scientific question
2. Study population
3. Biological sample
4. Experimental design
5. DNA methylation data acquisition
- 6. Quality control of DNA methylation data**
7. Epigenome-wide association study (EWAS)
8. Meta-EWAS or replication / validation
9. Biological interpretation



6. QUALITY CONTROL OF DNA METHYLATION DATA

Quality control of DNA methylation data

Reduce variability introduced during the experimental process, while keeping true biological variation.

1. Import raw IDAT files
2. Sample quality control
3. Probe quality control
4. Normalization
5. PCA and technical batch effect correction
6. Control of outlier values

6. QUALITY CONTROL OF DNA METHYLATION DATA

Quality control of DNA methylation data

Reduce variability introduced during the experimental process, while keeping true biological variation.

1. Import raw IDAT files
2. Sample quality control
3. Probe quality control
4. Normalization
5. PCA and technical batch effect correction
6. Control of outlier values

6. QUALITY CONTROL OF DNA METHYLATION DATA

1. Import raw idat files

Raw data

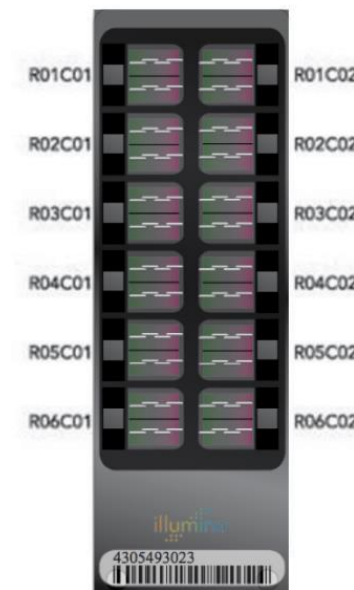
- IDAT files (2 files / sample)
- 4305493023_R01C01_Grn.idat
- 4305493023_R01C01_Red.idat



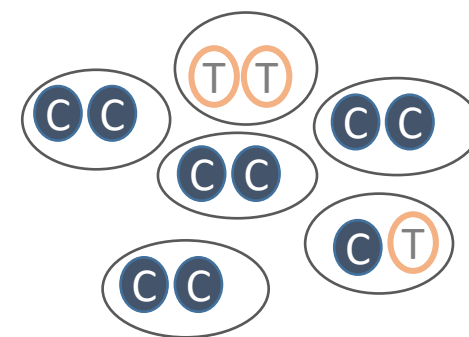
Beta values

$$\beta = \frac{M}{M + U + \alpha}, \quad 0 \leq \beta \leq 1$$

- M = methlyated signal
- U = unmethylated signal
- α = offset (usually 100) to stabilise beta-values



1 CpG in a tissue

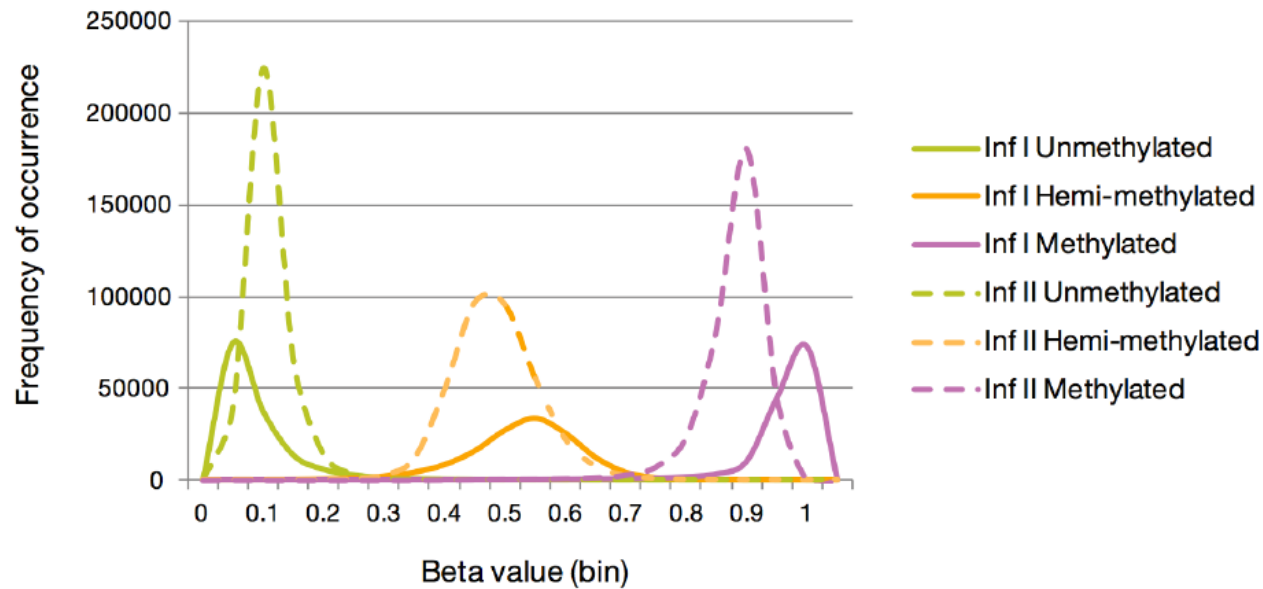


6. QUALITY CONTROL OF DNA METHYLATION DATA

Beta values

- from 0 to 1

Methylation reference samples

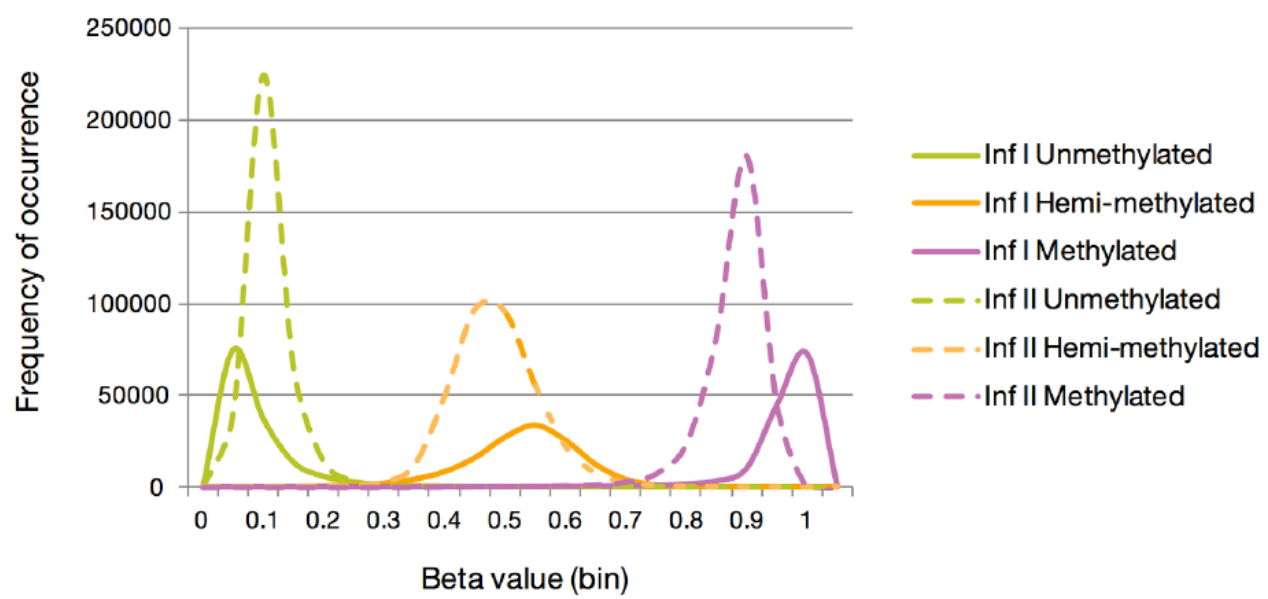


6. QUALITY CONTROL OF DNA METHYLATION DATA

Beta values

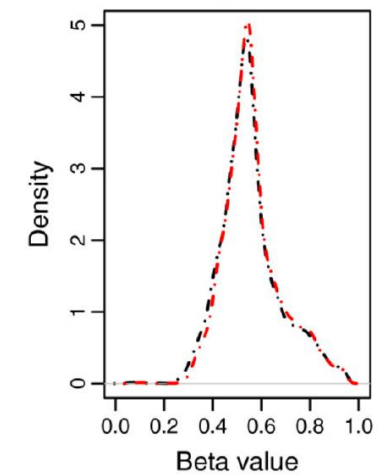
- from 0 to 1

Methylation reference samples

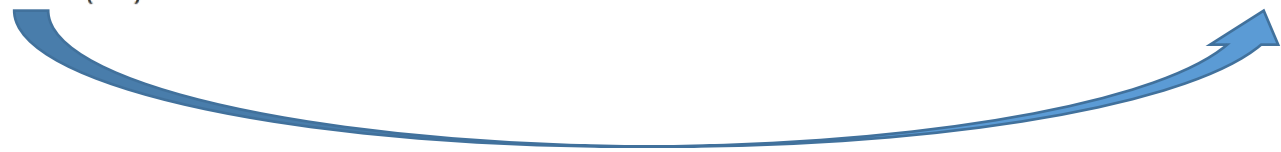
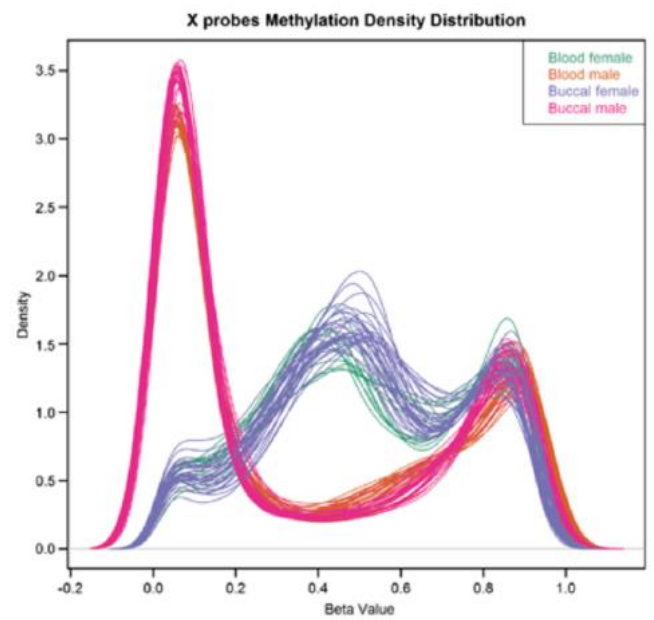


Hemi-methylated regions

Imprinted CpGs
N=250 approx



ChrX in females



6. QUALITY CONTROL OF DNA METHYLATION DATA

Beta values

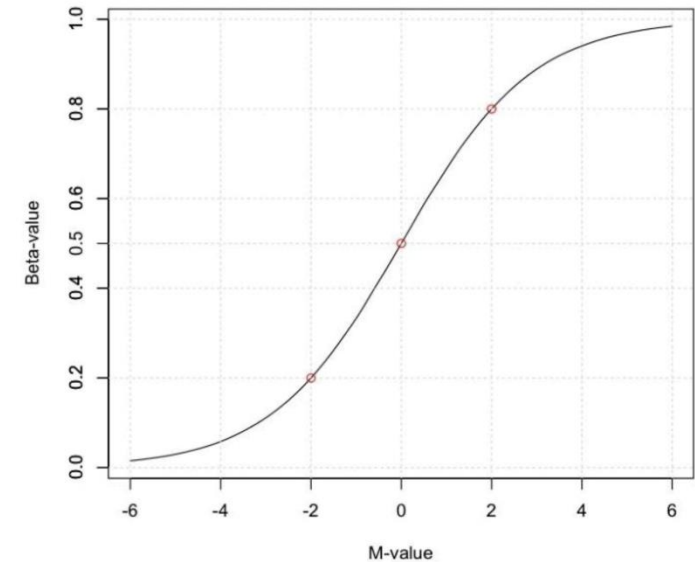
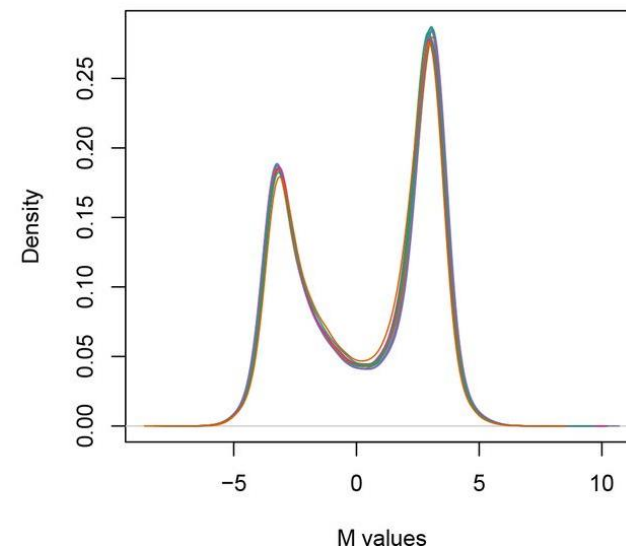
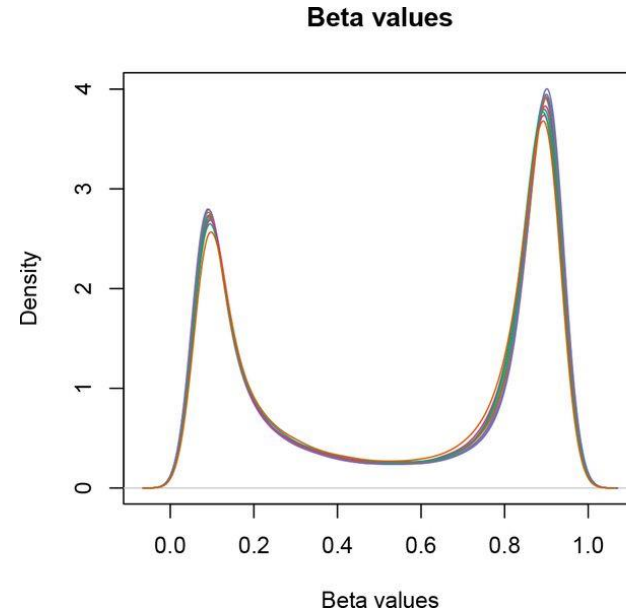
- from 0 to 1
- more intuitive interpretation

M values

- -inf to + inf
- more statistically valid
- less intuitive interpretation

$$M = \log_2 \left(\frac{\beta}{1 - \beta} \right)$$

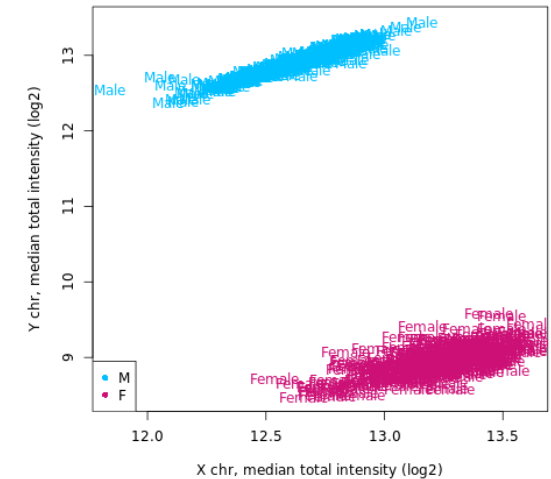
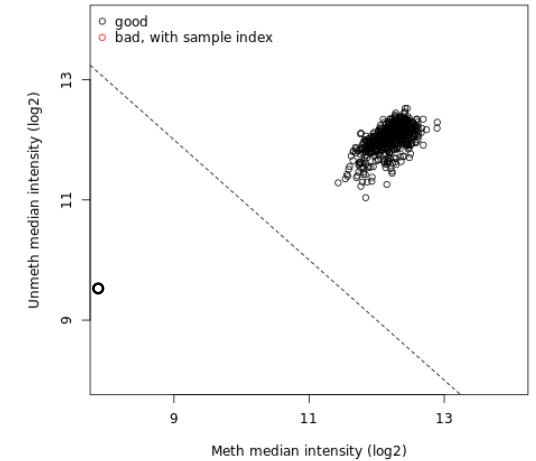
Relationship
between Beta-
value and M-value
is a logit
transformation



6. QUALITY CONTROL OF DNA METHYLATION DATA

2. Sample quality control

- 2.1. Overall low quality (methylated vs unmethylated signals)
- 2.2. Sample call rate (filter samples with % detected probes <95-98%)
- 2.3. Number of detected beads (filter samples with too few detected beads <3)
- 2.4. Sex consistency (sex chromosomal probes)
- 2.5. Technical duplicates (SNP probes)
- 2.6. DNA contamination (SNP probes)
- 2.7. Genetic consistency (SNP probes vs GWAS)



6. QUALITY CONTROL OF DNA METHYLATION DATA

3. CpG probe quality control

- 3.1. CpG probe call rate (filter probes with % detected probes <95-98%)
- 3.2. Number of detected beads (filter probes with too few detected beads <3)
- 3.3. Problematic probes (later in the QC pipeline)
 - Array control probes
 - SNP probes
 - Non-CpG methylation probes
 - CpG probes in sex chromosomes
 - CpG probes with cross-hybridization problems
 - CpG probes with SNPs in the CpG site
 - CpG probes with SNPs in other positions

Illumina manifest: <https://support.illumina.com/downloads/infinium-methylationepic-v1-0-product-files.html>

Zhou's list: <https://github.com/zhou-lab/InfiniumAnnotation>

Annotation

Recommended list of probes to eliminate

6. QUALITY CONTROL OF DNA METHYLATION DATA

4. Normalization

4.1. Background noise correction

4.2. Color bias correction

4.3. Probe bias correction

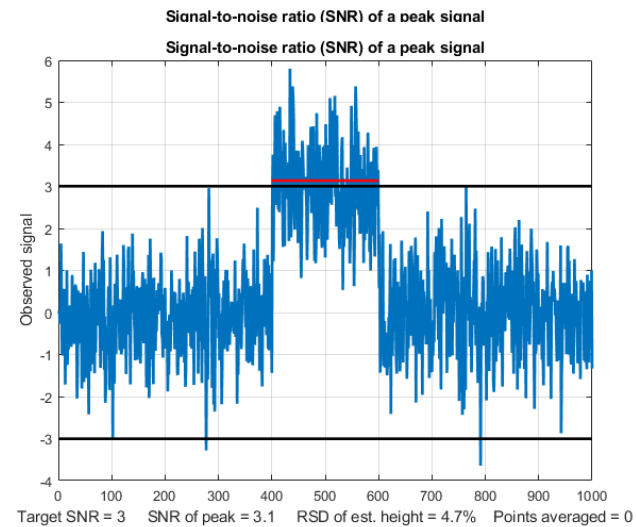
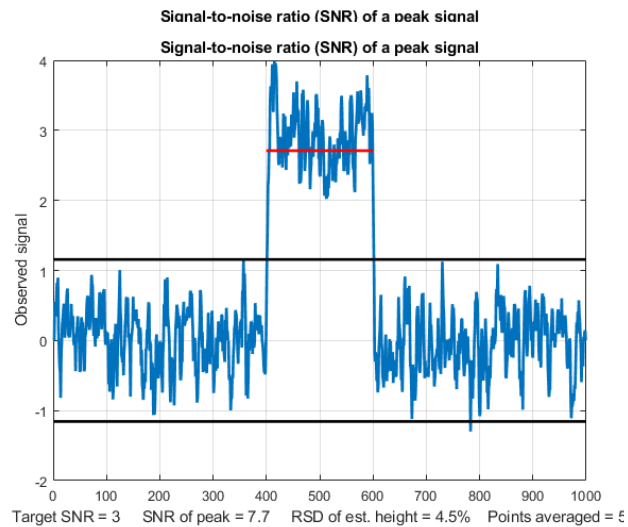
4.4. Across array normalization

No consensus on best method

6. QUALITY CONTROL OF DNA METHYLATION DATA

4.1. Background noise correction

- To remove noise from the data
- Often use negative control probes to remove this noise, but also other methods
- Tools: GenomeStudio (Illumina) or some R packages



6. QUALITY CONTROL OF DNA METHYLATION DATA

4.2. Color bias

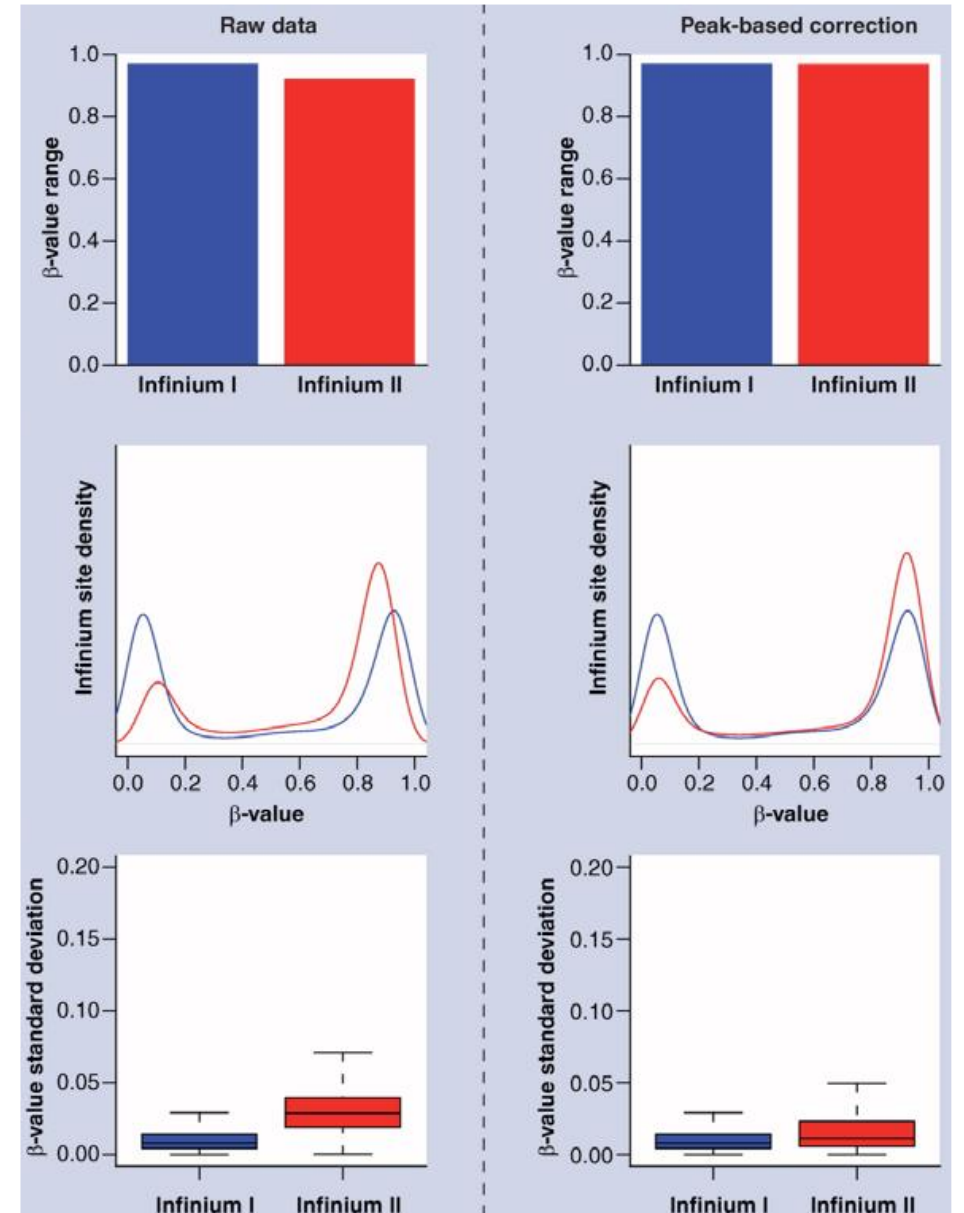
- The two color channels are known to perform differently (red>green)
- Method: signal / average signal of the internal normalisation control for that color
- Tools: GenomeStudio, and several R packages

- After background noise and color bias are removed, beta values are calculated

6. QUALITY CONTROL OF DNA METHYLATION DATA

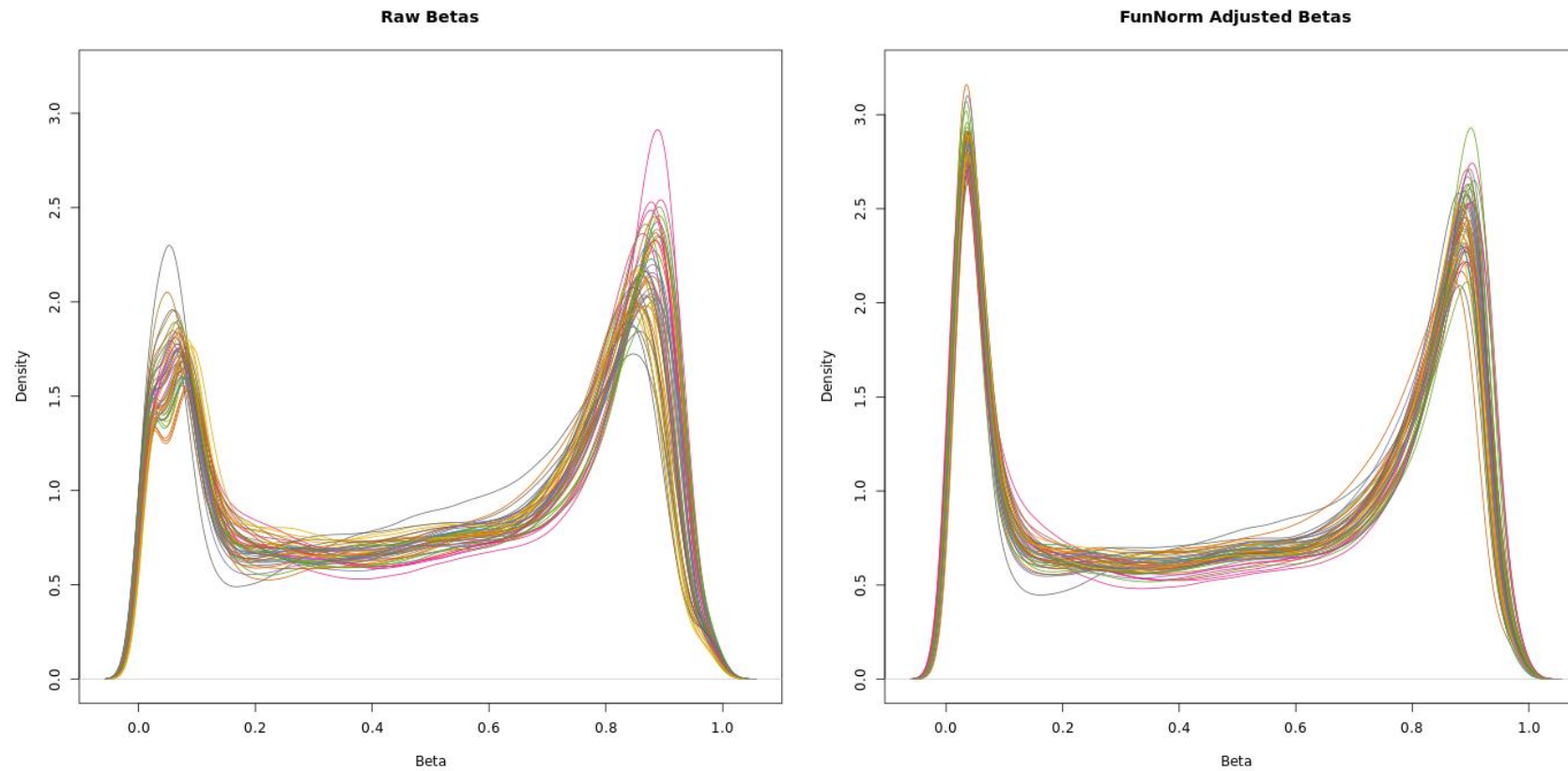
4.3. Probe bias correction

- Type I and II probes behave different
- It can be a problem in some type of analysis where we rank or combine probes (clustering, regional analysis...), but not for single CpG analyses
- Methods: peak based correction, BMIQ, SWAN



6. QUALITY CONTROL OF DNA METHYLATION DATA

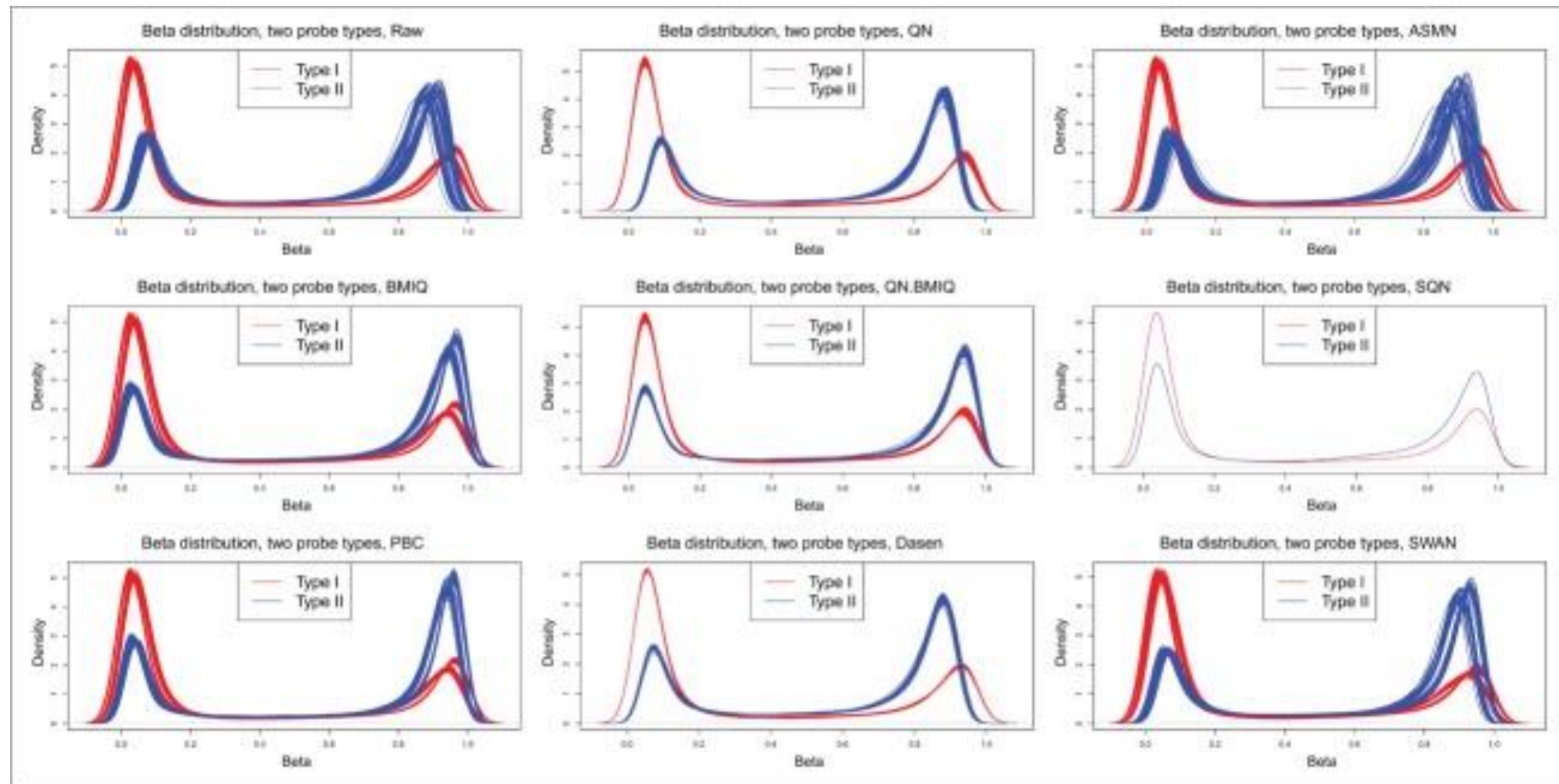
4.4. Across array normalization



6. QUALITY CONTROL OF DNA METHYLATION DATA

4.4. Across array normalization

There are many different methods:



6. QUALITY CONTROL OF DNA METHYLATION DATA

4.4. Across array normalization

- **Quantile normalization**

Normalises data to average/median of all observations

From gene expression arrays

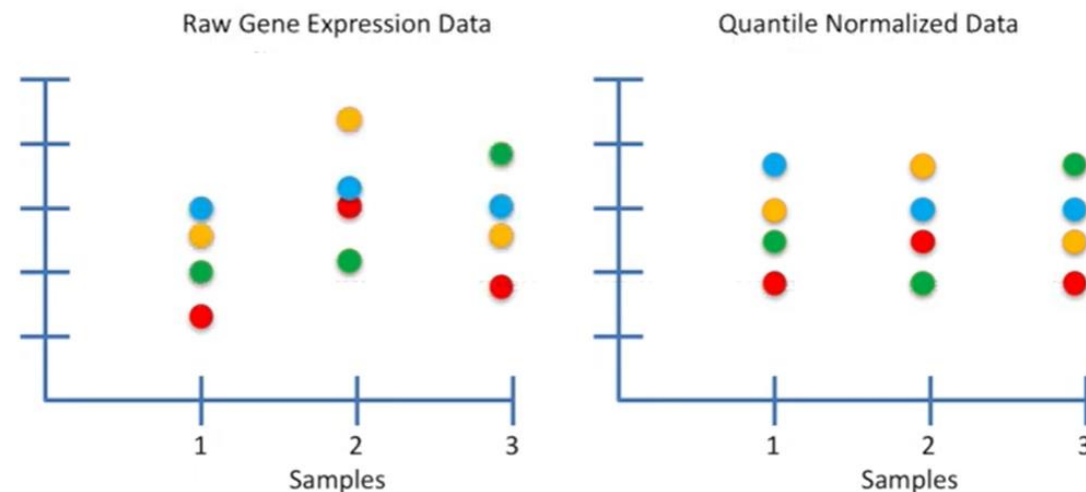
Not best option for DNA methylation

- **Functional normalization**

Quantile normalisation of control probes only

- **Functional normalization + residuals EWAS PCs (meffil)**

- Estimate quantiles
- Residualize EWAS PCs on the quantiles (fixed or random effects)



<https://www.youtube.com/watch?v=ecjN6Xpv6SE>

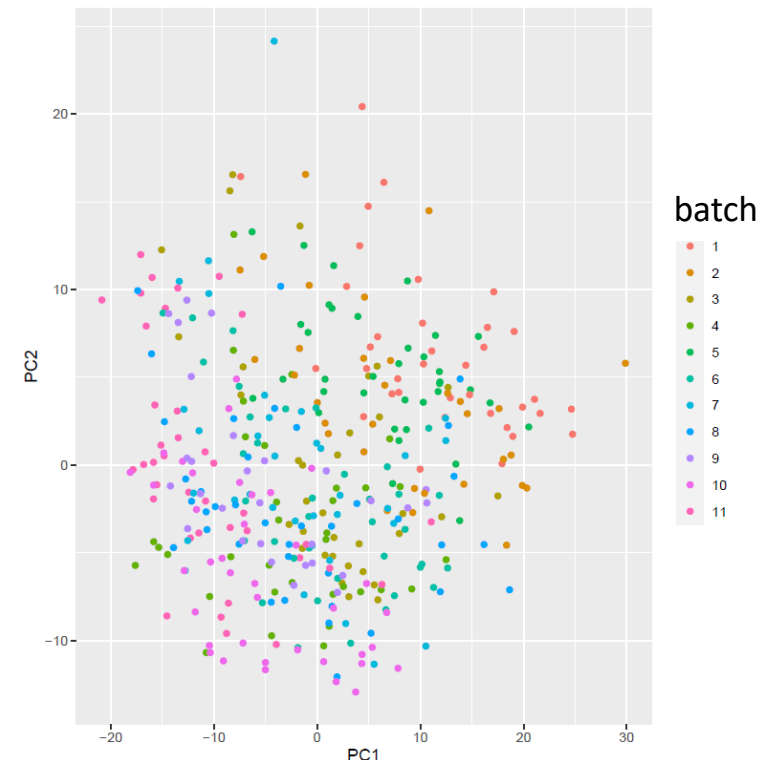
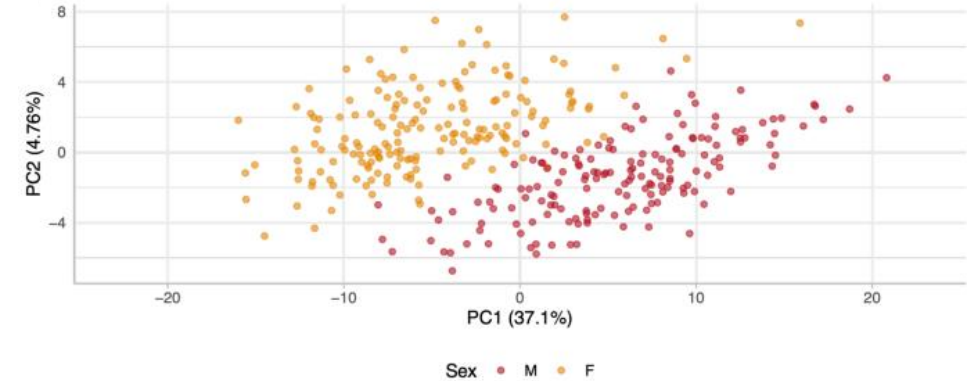
6. QUALITY CONTROL OF DNA METHYLATION DATA

5. PCA and technical batch effect correction

Principal component analyses (PCA)

Main variables explaining variance:

- Biological variables
 - Tissue
 - Sex
 - Age
 - Ancestry
 - Disease (ie. cancer)
- Technical batch variables
 - DNA extraction batch
 - Bisulfite conversion batch
 - Array
 - Position in array

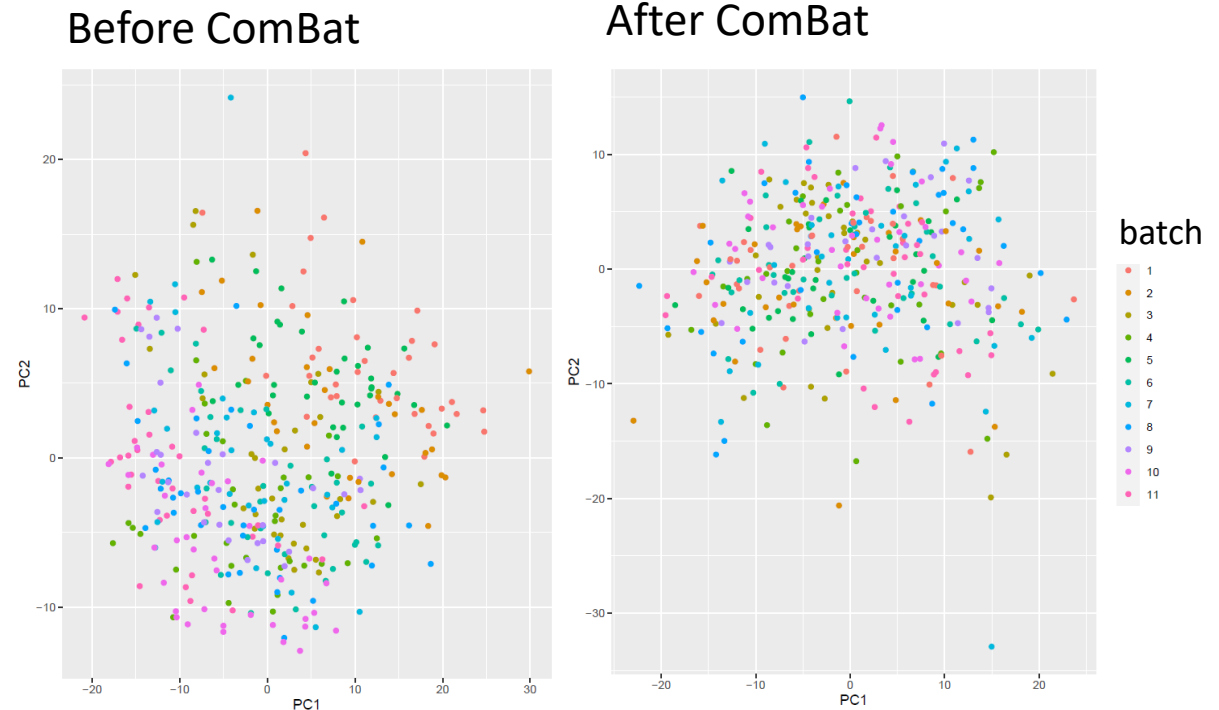


6. QUALITY CONTROL OF DNA METHYLATION DATA

5. Technical batch effect correction

Technical batch correction methods:

- Known technical batch variable:
 - Add variable in the regression models
 - Omics R package: residuals of known variable
 - ComBat R package: Bayesian approach
- Unknown technical batch variables:
 - Surrogate variable analysis (SVA)
 - Residuals
 - Add SVs in the regression model (**this one!**)



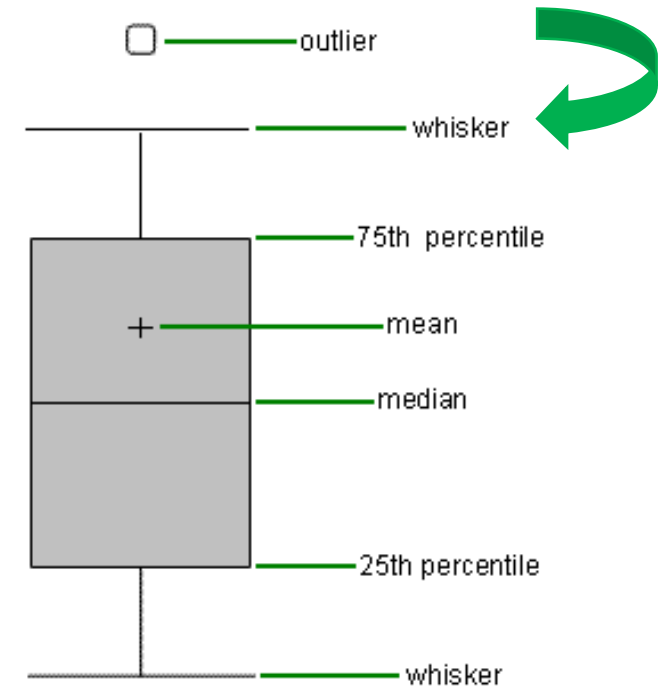
6. QUALITY CONTROL OF DNA METHYLATION DATA

6. Control of outlier values

- Extrem values in the data
- Problematic in DNA methylation

Outlier correction methods:

- Trimming
 - To delete values that we do not believe
 - Define outlier (ie. $4 \times \text{IQR}$)
- Winsorising
 - To retain the high-value responses but not take them too literally
 - Take top values and bring them to lower values (ie. p99)



meffil R package

meffil R package

<https://pubmed.ncbi.nlm.nih.gov/29931280/>

<https://github.com/perishky/meffil>

1. Import raw IDAT files: **yes**
2. Sample quality control
 - 2.1. Overall low quality: **yes**
 - 2.2. Sample call rate: **yes**
 - 2.3. Number of detected beads: **yes**
 - 2.4. Sex consistency: **yes** (Aryee et al., 2014)
 - 2.5. Technical duplicates: **NA**
 - 2.6. DNA contamination: **NA**
 - 2.7. Genetic consistency: **yes**
3. Probe quality control:
 - 3.1. CpG probe call rate: **yes**
 - 3.2. Number of detected beads: **yes**
 - 3.3. Problematic probes: **control, sex chr**
4. Normalization:
 - 4.1. Background noise correction: **'noob' method** (Triche et al., 2013)
 - 4.2. Color bias correction: **'noob' method** (Triche et al., 2013)
 - 4.3. Probe bias correction: **NA**
 - 4.4. Across array normalization: **functional normalization** (Fortin et al., 2014) + extension to fixed and random effects
5. PCA ant technical batch effect correction: **PCA, PC associations, SVA** during analysis
6. Control of outlier values: **outside the package**

methyAnalysis	Pan Du, Lei Huang, Gang Feng	DNA methylation data analysis and visualization
MethylAid	M. van Iterson	Visual and interactive quality control of large Illumina DNA Methylation array data sets
methyKit	Altuna Akalin	DNA methylation analysis from high-throughput bisulfite sequencing results
MethylMix	Olivier Gevaert	MethylMix: Identifying methylation driven cancer genes
methyMnM	Yan Zhou	detect different methylation level (DMR)
methyPipe	Kamal Kishore	Base resolution DNA methylation data analysis
MethylSeekR	Lukas Burger	Segmentation of Bis-seq data
methylumi	Sean Davis	Handle Illumina methylation data
minfi	Kasper Daniel Hansen	Analyze Illumina Infinium DNA methylation arrays
missMethyl	Belinda Phipson, Jovana Maksimovic	Analysing Illumina HumanMethylation BeadChip Data
MoonlightR	Antonio Colaprico, Catharina Olsen	Identify oncogenes and tumor suppressor genes from omics data
MPFE	Conrad Burden	Estimation of the amplicon methylation pattern distribution from bisulphite sequencing data
normalize450K	Jonathan Alexander Heiss	Preprocessing of Illumina Infinium 450K data

INTRODUCTION TO EPIGENOME-WIDE ASSOCIATION STUDIES (EWAS)

2. PRE-PROCESSING OF DNA METHYLATION DATA (PRACTICAL SESSION)

QUALITY CONTROL OF DNA METHYLATION DATA

Data: Subset from GEO GSE42861 (N=294)

- Array: 450K
- Tissue: blood
- Ancestry: White European
- Sex: males and females
- Smoking: never, former, current
- Age: yes
- Array batch: yes

Input: IDAT files

Output: ExpressionSet with matrix of beta values + covariates dataframe (exposure, covariates, cells)

Tool: meffil R package

Questions:

1. Is there any sample that is excluded due to Methylated vs Unmethylated (low quality)?
2. Is there any sample that is excluded due to inconsistent sex?
3. Which are the main biological and technical variables associated with PC1?