NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

SciLifeLab

Epigenomics Data Analysis: DNA Methylation

Louella Vasquez

25-09-22

Schedule for Monday

Monday (22 September 2025)

DNA methylation with Illumina array and Bisulfite-seq

09:00 - 09:30 Welcome

09:30 - 10:15 Introduction to DNA methylation + Overview Array Exercises

10:15 - 10:30 Uppmax set up + break

10:30 - 12:00 Exercises Array Workflow

12:00 - 13:00 lunch *(offline)*

13:00 - 14:00 Methylation methods & technologies (Jessica Nordlund)

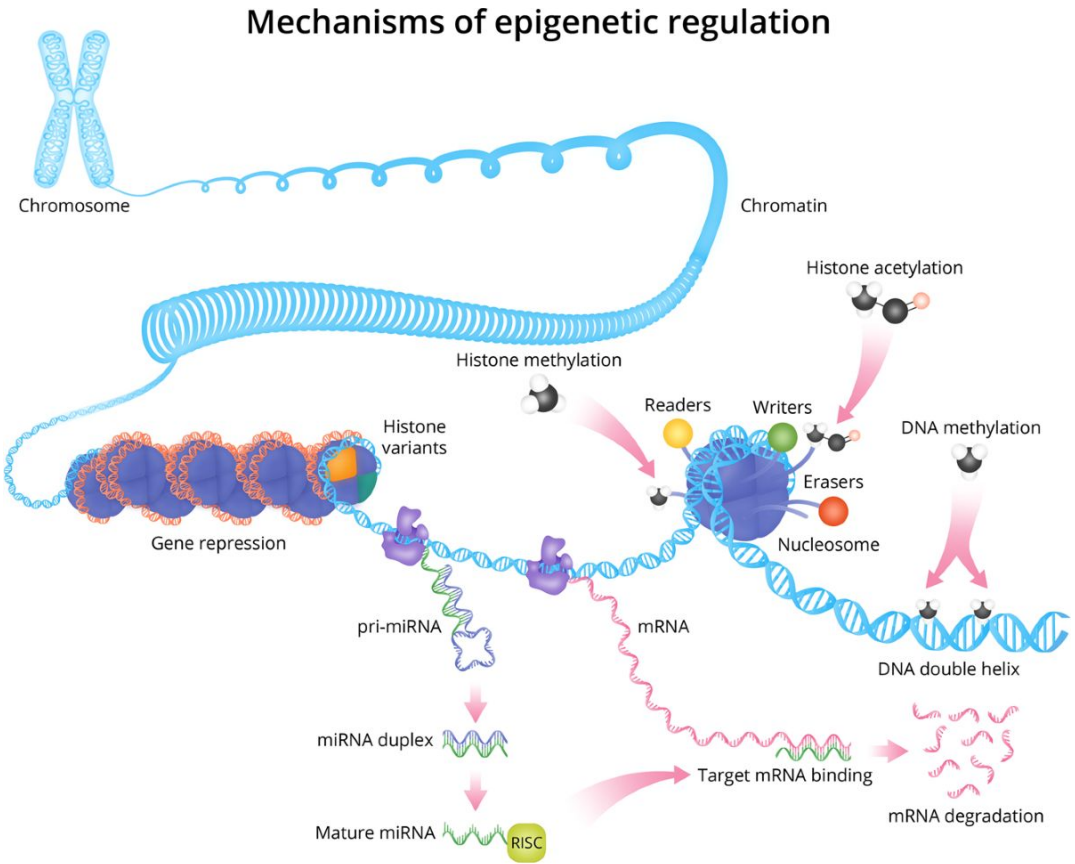
14:00 - 14:15 Break

14:15 - 14:30 Methylation Exercises Overview II: Methylation Sequencing

14:30 - 16:00 Exercises Methylation Sequencing

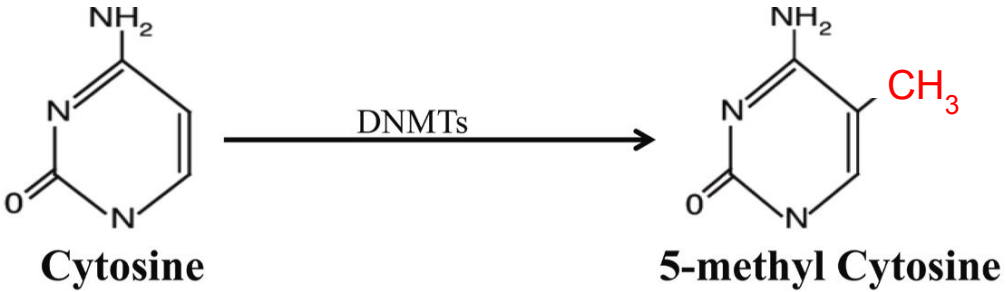
16:00 - 17:00 Daily challenge *(offline)*

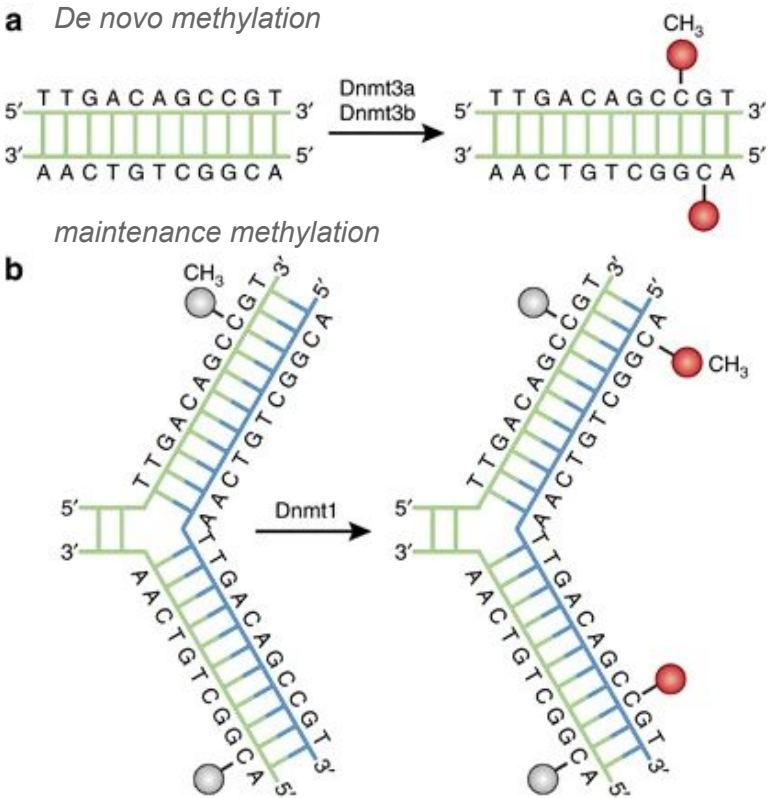
Epigenome regulates gene expression via chromosomal alteration that does not involve changes in the DNA sequence



Ref: *Mol Psychiatry* 25, 1718–1748 (2020)

DNA methylation is a stable, heritable chemical modification



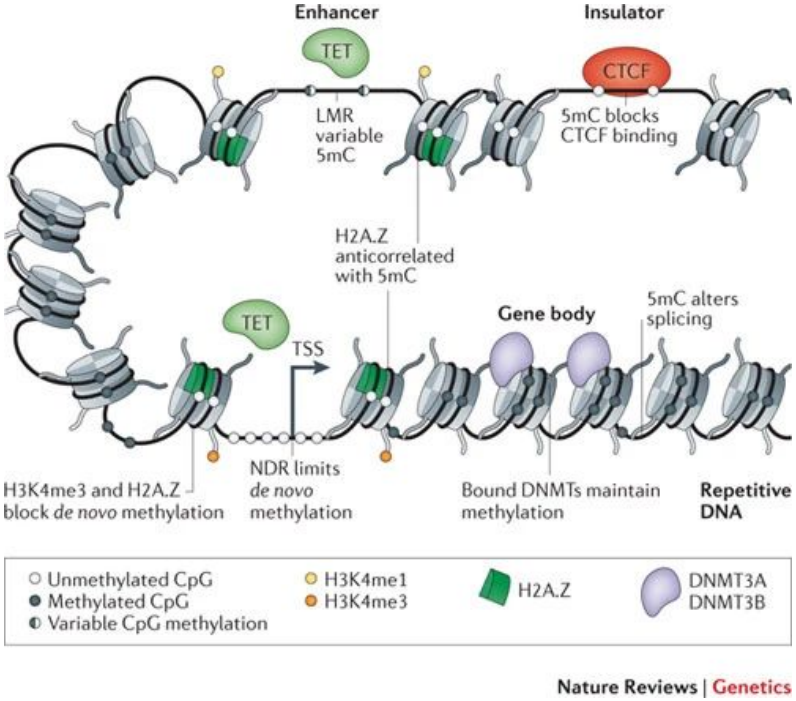


5mC in the genome

- predominantly in CpG (**C**ytosine-**p**hosphate-**G**uanine) dinucleotides in metazoan genomes
  - ~28M CpGs in humans
  - 60–80% methylated in somatic cells

- CpGs in CG-dense regions are CpG islands (CGIs)
  - 200-2000 bp with >50% GC-content
  - CGIs tend to be in promoters, unmethylated for transcribed genes

- non-CpG methylation has been mainly observed in hESCs and neuronal cells in humans
  - CHH, CHG where H = A, C or T



Key functions of DNA methylation

Tissue specific gene regulation

Suppression of transposable elements

Essential for normal development

Genomic imprinting

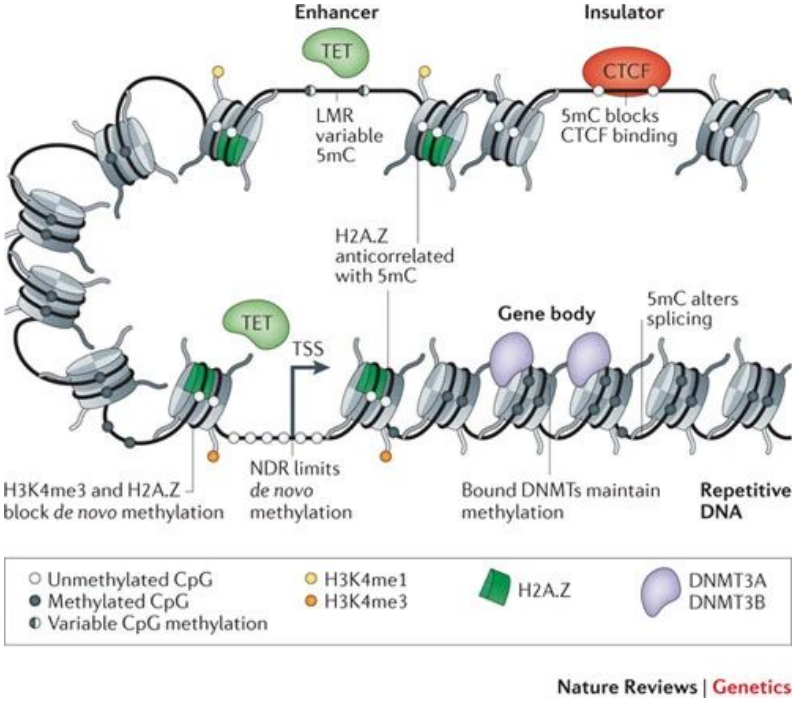
X-chromosome inactivation

Ageing

- Global hypomethylation is proportional to age

Cancer

- Global hypomethylation and locus-specific hypermethylation of CpG islands



Consortia with large-scale profiling of DNA methylation

Project	Website	Goals/Aims
ENCODE/Roadmap	http://www.roadmapepigenomics.org/	Reference epigenomes across a variety of human cell types
ICGC	https://icgc.org/	Comprehensive catalogs of genomic abnormalities in tumors in 50 different cancer types (some DNA methylation)
TCGA	https://tcga-data.nci.nih.gov/tcga/	Twenty-five tumor types; gene expression profiling, copy number variation profiling, SNP genotyping, DNA methylation profiling, microRNA profiling
BLUEPRINT	http://www.blueprint-epigenome.eu/	Distinct types of haematopoietic cells from healthy individuals and malignant leukaemic counterparts; at least 100 reference epigenomes

DNAM measurement

Differentiate mC from C

C > T Bisulfite conversion

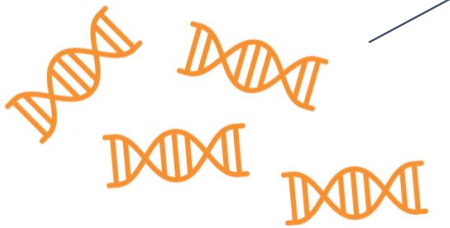
A: 5'-GACCGTTCCAGGTCCAGCAGTGCGCT-3'

B: 3'-CTGGCAAGGTCCAGGTCGTCACGCGA-5'

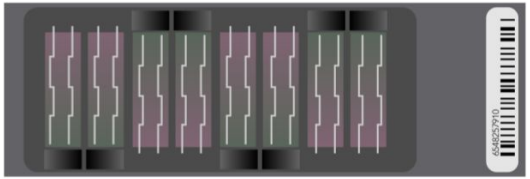
A: 5'-GATCGTTTTAGGTTTAGTAGTGCGTT-3'

B: 3'-TTGGCAAGGTTTAGGTTGTTATGCGA-5'

DNA amplification

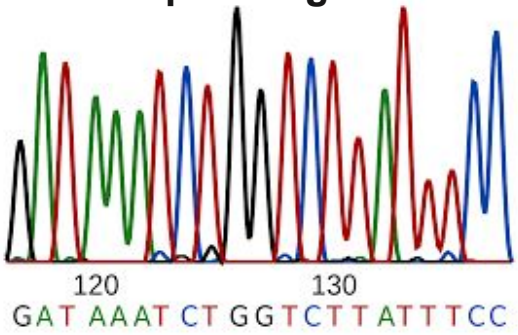


Methylation Array



DNA Sequencing

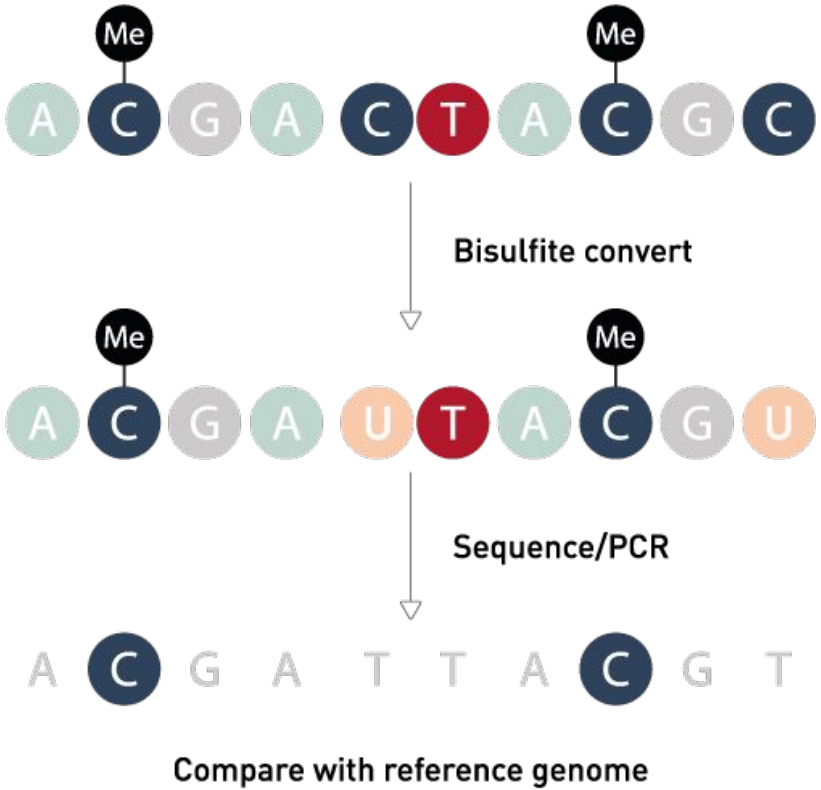
Select regions



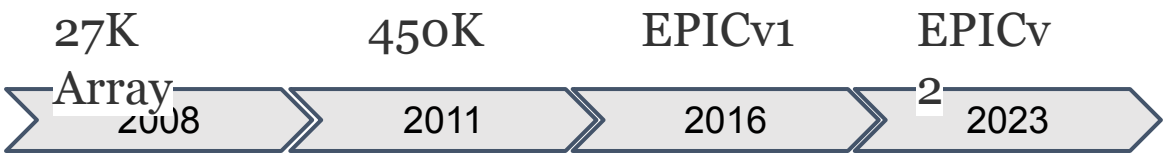
Bisulfite conversion

- Used for both array and sequencing

- C → U → (PCR) →T
- mC → C → (PCR) →C



DNA Methylation Array： Human Methylation BeadChip



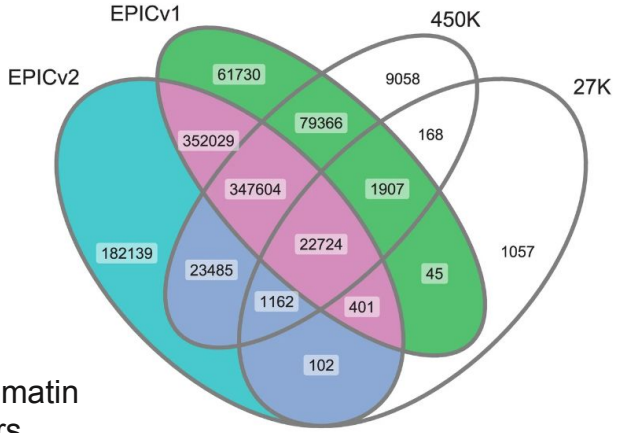
25K probes
CGI in promoters and cancer genes

485K probes
+
CGI shores & shelves
99% RefSeq genes
FANTOM4 promoters
MHC region enhancers

867K probes
+
FANTOM5 enhancers
miRNA promoters

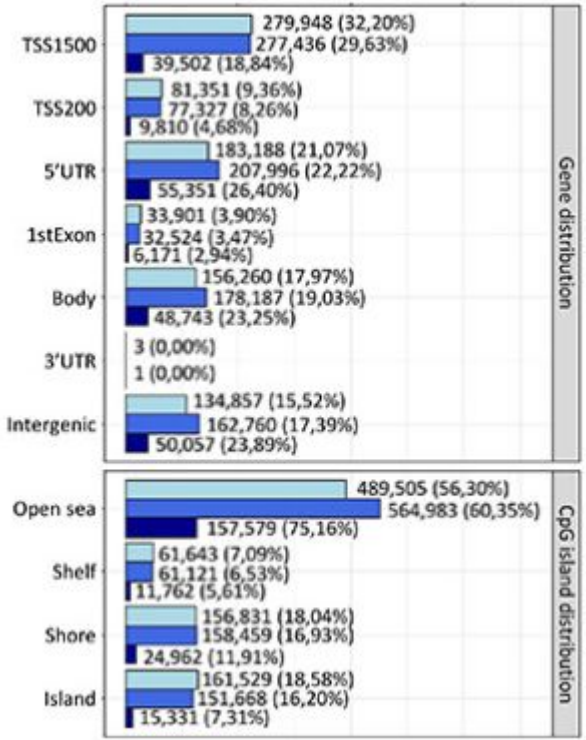
935K probes
+
ENCODE open chromatin
FANTOM5 enhancers
Superenhancers
Common cancer driver mutations

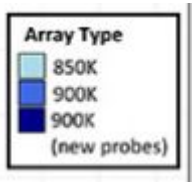
77.63% homologous to v1
200K new probes
>99% correlation in methylation values with v1

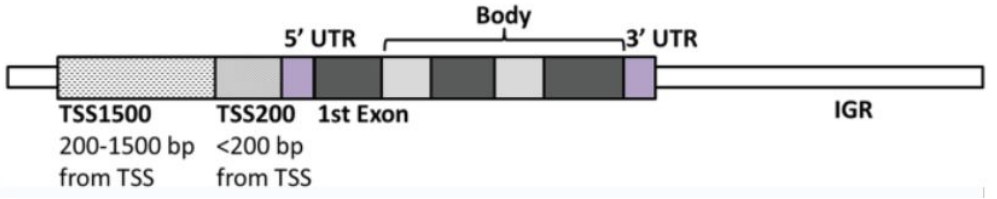


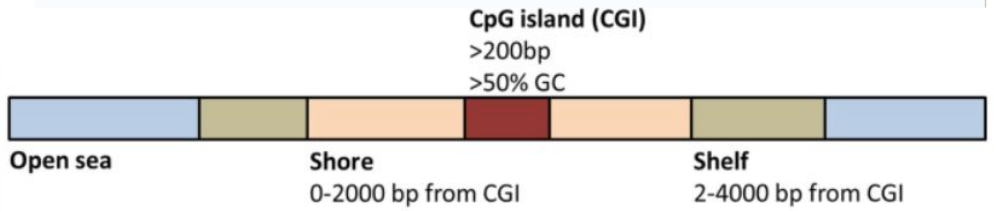
Ref: https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-024-1002

DNA Methylation Array： Human Methylation BeadChip









Ref: https://doi.org/10.1080/15592294.2023.2185742

DNA Methylation Array： Human Methylation BeadChip

- ✓ genome-wide
- ✓ single nucleotide resolution
- ✓ low cost, ease of use
- ✓ widely used in population scale studies
- ✓ high reproducibility and reliability
- ✓ compatible with FFPE tissues
- ✓ a few non CpG probes
- ✓ well established bioinformatics solutions

	HM450	EPICv1	EPICv2 (unique prefixes)	EPICv2 (all probes)
Total Probes	**486,427**	**866,553**	**931,293**	**937,690**
cg probes	482,421 (99.18%)	862,927 (99.6%)	926,858 (99.5%)	933,252 (99.5%)
ch probes	3,091 (0.641%)	2,932 (0.34%)	2,914 (0.31%)	2,914 (0.31%)
rs probes	65 (0.013%)	59 (0.0068%)	62 (0.0067%)	65 (0.0069%)
ct probes	850 (0.175%)	635 (0.073%)	635 (0.068%)	635 (0.068%)
nv probes	0 (0%)	0 (0%)	824 (0.88%)	824 (.088%)
Infinium-I	135,501 (27.9%)	142,158 (16.4%)	127,028 (13.6%)	128,295 (13.7%)
Infinium-II	350,926 (72.1%)	724,395 (83.6%)	804,362 (86.4%)	809,395 (86.3%)

"cg": CpG cytosine methylation probes; "ch": non-CG cytosine methylation probes; "rs": common SNP probes; "nv": probes for somatic mutations found in cancer; and "ct": quality control probes.

Infinium Bead Technology: Type I and Type II probes

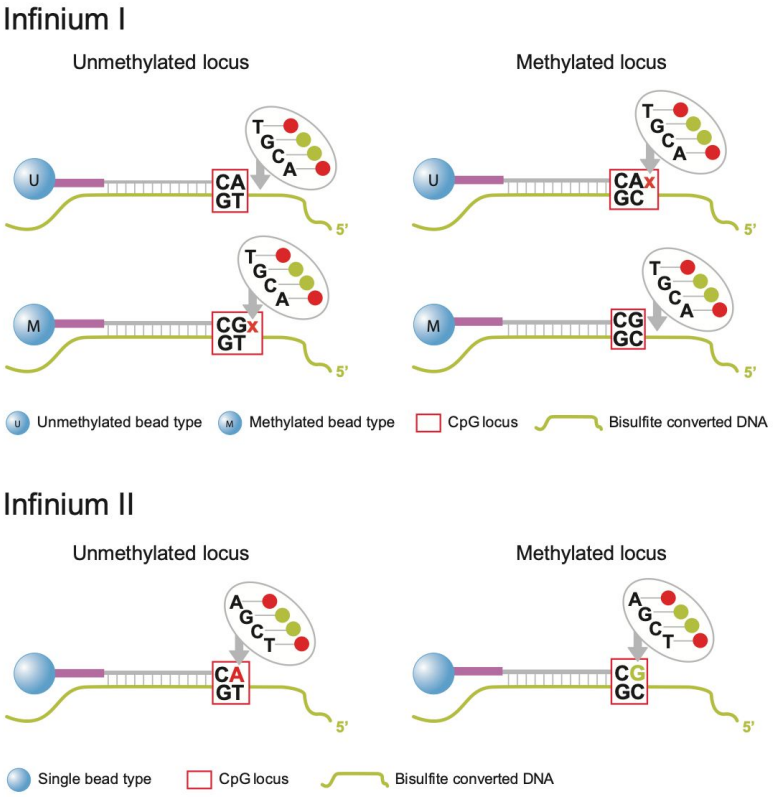
- Silica bead affixed with oligonucleotide containing 23 base address and 50 base probe sequence

  `CTACAAATACGACACCCGCAACCCATATTTCATATATTATCTCATTTAAC`

- Bisulfite converted - DNA is hybridised to the probe
- Single base extension of the probe with fluorescence ddNTP

- Fluorescence signal detects the C as thymine (T) if originally unmethylated or C if methylated.

- Infinium-I chemistry has one color channel & two bead types
  - Methylated (M) and Unmethylated (U) beads
- Infinium-II uses one bead type and two color channels
  - green for M, red for U



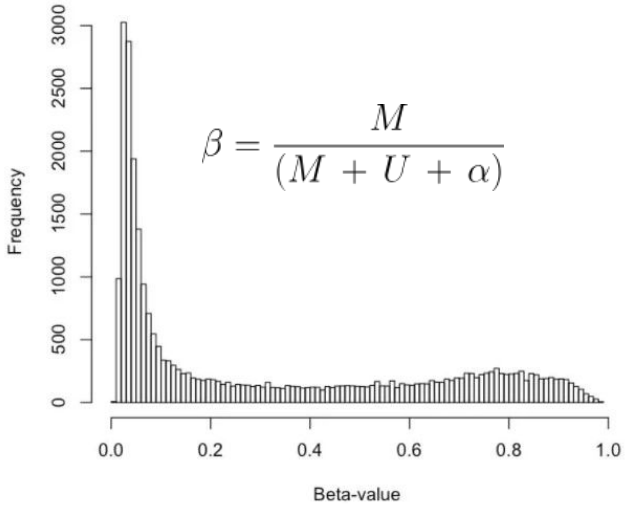
Measurement of Methylation Values

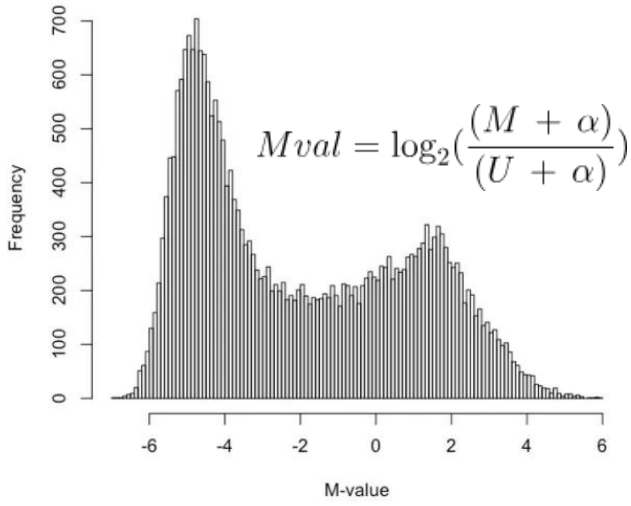
Beta value

- Bounded value between 0 to 1
- Fraction of cells with a methylated C
- Easier to interpret

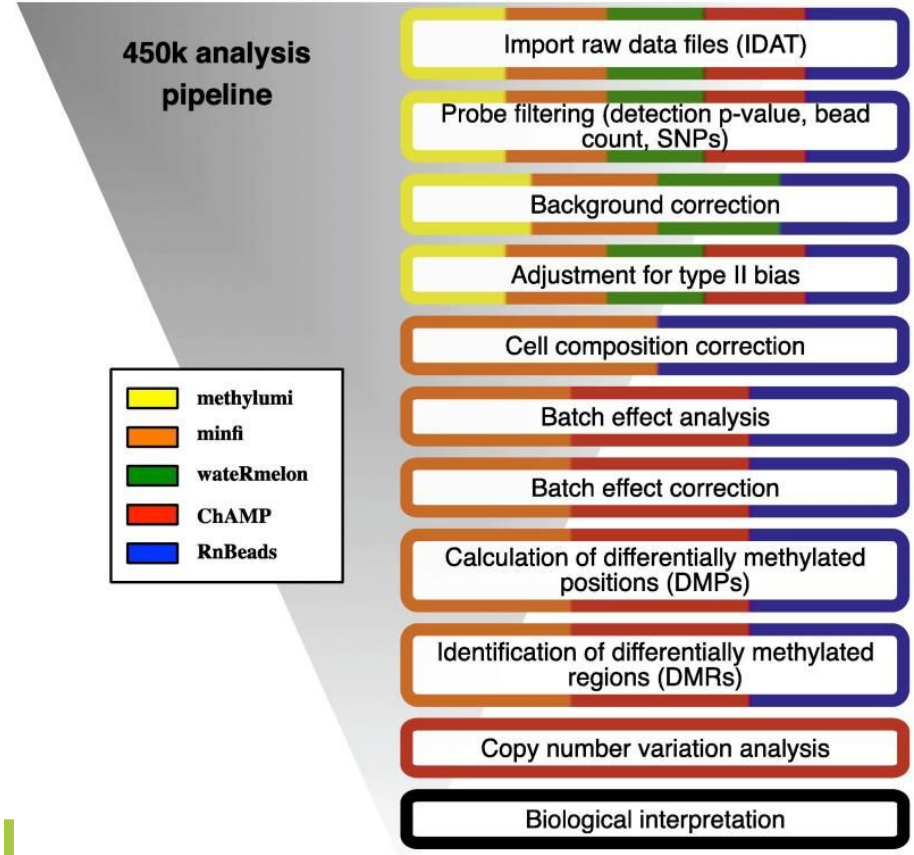
M value

- Log ratio between M and U intensities
- +M value, M > U
- - M value, M < U
- Used in statistical model testing





Analysis Pipeline



Freely available packages for Infinium 450k data analysis. *as of 2018*

Package	Use
ChAMP	Comprehensive suite of functions; automated pipeline
COHCAP	CpG island analysis and gene expression data integration
Comb-p	DMR calling
DMRcate	DMR calling
Epigenetic clock	Predictor of sample age
EWasher	Reference-free cell composition correction
FastDMA	Quantile normalisation and DMP/DMR calling
IMA	Preprocessing including normalisation methods; Pipeline option
Lumi	Background correction, general normalisation
Marmal-aid	450k database for data integration
MethylAid	Interface for interactive sample QC
Methylumi	Comprehensive suite of functions
Minfi	Comprehensive suite of functions
NIMBL	Matlab code for QC and DMP calling
RefFreeEWAS	Reference-free cell composition correction
RnBeads	Comprehensive suite of functions
shinyMethyl	Interface for interactive sample QC
wateRmelon	Preprocessing including performance metrics and numerous normalisation methods
SeSAMe	

Import raw methylation data

- Raw IDAT files are in folder named after chip ID
- Red/Green signal intensity files per sample

- IDAT file name format

  `5975827018_R06C02_Grn.idat`

  `5975827018_R06C02_Red.idat`

  `<chip barcode>_<chip position>_<channel>`

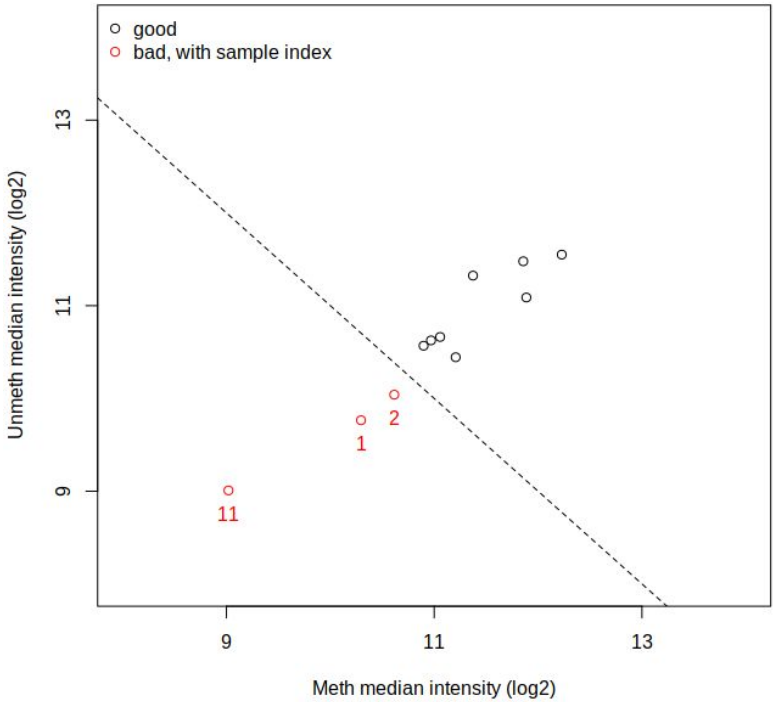
- Sample annotation CSV file

```
dataDirectory <- "/sw/courses/epigenomics/DNAmethylation/array_data/"
# read in the sample sheet for the experiment
targets <- read.metharray.sheet(dataDirectory, pattern="SampleSheet.csv")
# read in the raw data from the IDAT files
rgSet <- read.metharray.exp(targets=targets)
# Go from intensity data to methylation levels
MSet <- preprocessRaw(rgSet)
```

Quality Control to flag poor quality and outlier samples

- Median intensity of M vs U
- Beta value distribution
- Probes detection P value
- Internal control probes
  - Bisulfite conversion
  - Hybridization
  - Extension
  - Negative controls
- Gender check

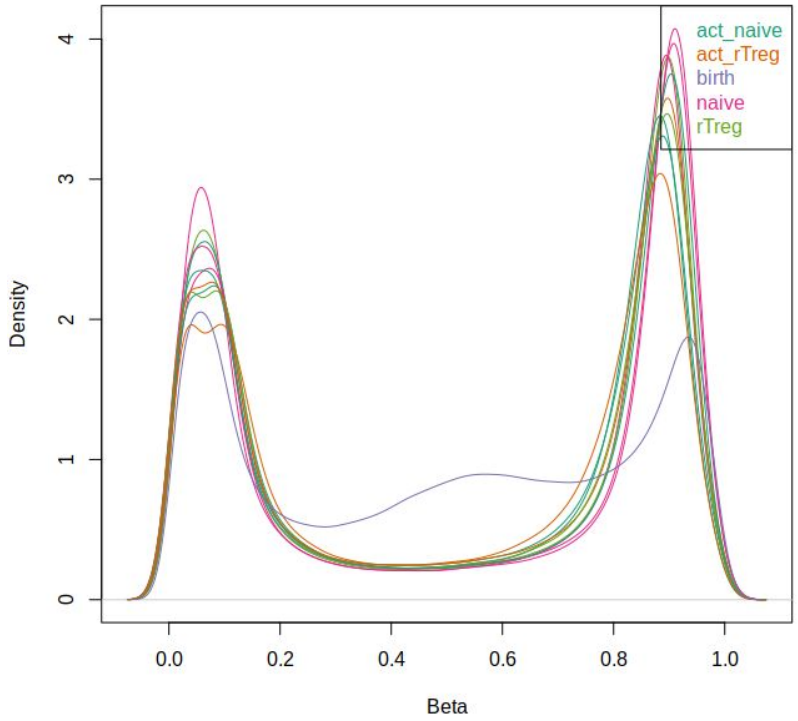
```
qc <- getQC(MSet)
plotQC(qc)
```



Quality Control to flag poor quality and outlier samples

- Median intensity of M vs U
- Beta value distribution
- Probes detection P value
- Internal control probes
  - Bisulfite conversion
  - Hybridization
  - Extension
  - Negative controls
- Gender check

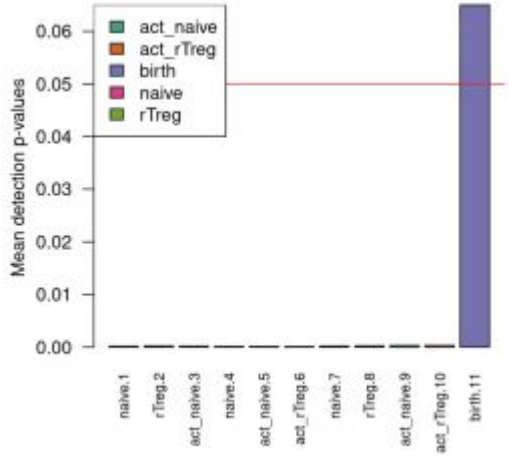
```
phenoData <- pData(MSet)
densityPlot(MSet, sampGroups = phenoData$Sample_Group)
```



Quality Control to flag poor quality and outlier samples

- Median intensity of M vs U
- Beta value distribution
- Probes detection P value
- Internal control probes
  - Bisulfite conversion
  - Hybridization
  - Extension
  - Negative controls
- Gender check

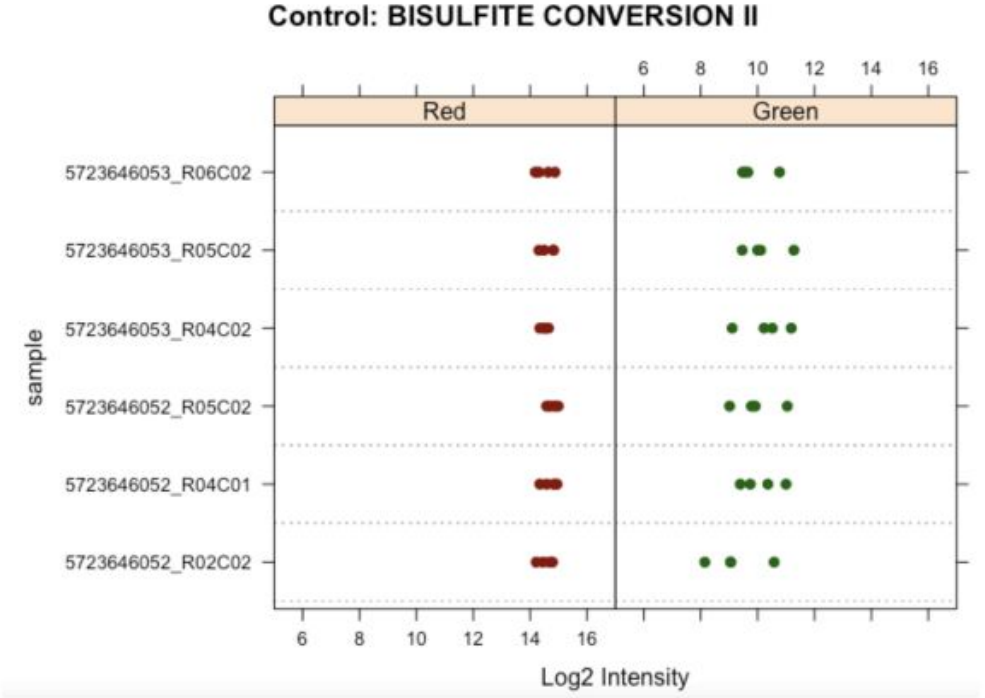
```
# Calculate the detection p-values
detP <- detectionP(rgSet)
# examine mean detection p-values across all samples to identify any failed
barplot(colMeans(detP), las=2, cex.names=0.8, ylab="Mean detection p-values
abline(h=0.05,col="red")
```



Quality Control to flag poor quality and outlier samples

- Median intensity of M vs U
- Beta value distribution
- Probes detection P value
- Internal control probes
  - Bisulfite conversion
  - Hybridization
  - Extension
  - Negative controls
- Gender check

```
1 controlStripPlot(RGSet, controls="BISULFITE CONVERSION II")
```



Quality Control to flag poor quality and outlier samples

- Median intensity of M vs U
- Beta value distribution
- Probes detection P value
- Internal control probes
  - Bisulfite conversion
  - Hybridization
  - Extension
  - Negative controls
- Gender check

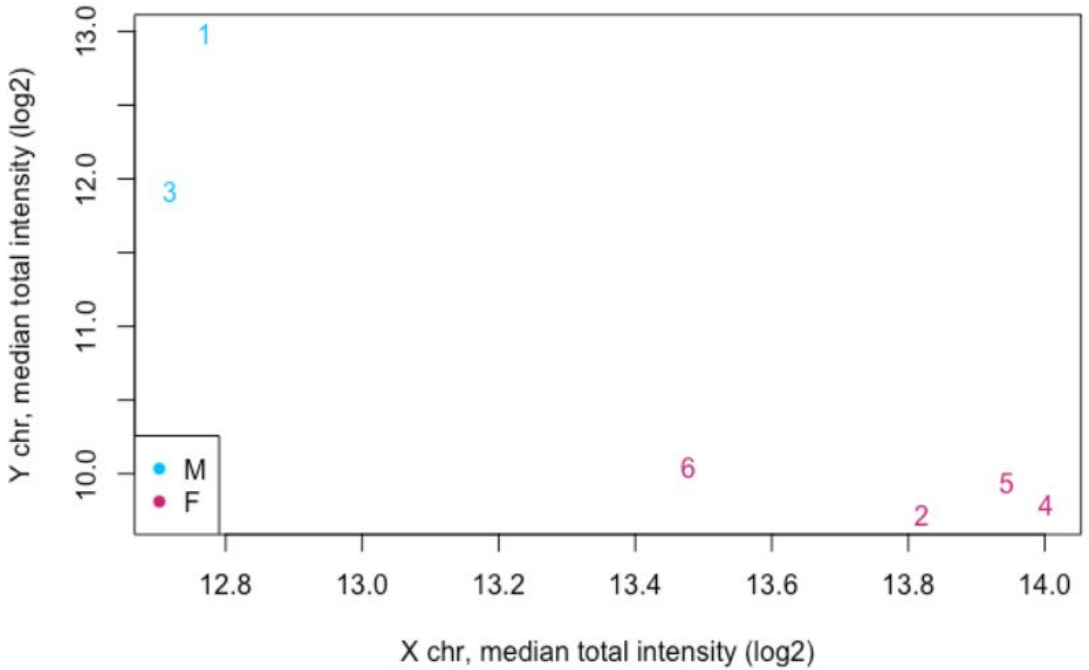
QC Probes

Control Probe	Purpose
Staining	measure efficiency and sensitivity of staining step (independent of hybridisation/extension step)
Extension	test efficiency of extension of A, T, C and G nucleotides from a hairpin probe (sample-independent). The perfect match hairpin controls should result in high signal, and the mismatch probes in low signal
Hybridization	test the overall performance of Infinium assay using synthetic targets (not DNA) at 3 concentrations
Target removal	test efficiency of stripping step after extension
Bisulphite conversion	test efficiency of bisulphite conversion by query of C/T polymorphism
Specificity controls	check for non-specific detection of methylation signal over unmethylated background. Specificity controls are designed against non-polymorphic T sites (G/T mismatch)
Non-polymorphic	query a non polymorphic base A, T, C and G to test overall performance of the assay from amplification to detection
Negative	randomly permutated bisulphite-converted sequences containing no CpGs. They should not hybridise to DNA. The mean of these probes determines the system background

Quality Control to flag mislabelled samples

- Median intensity of M vs U
- Beta value distribution
- Probes detection P value
- Internal control probes
  - Bisulfite conversion
  - Hybridization
  - Extension
  - Negative controls
- Gender check

```
predictedSex <- getSex(GRset, cutoff = -2)$predictedSex
plotSex(getSex(GRset, cutoff = -2))
```



Probe Filtering

remove probes with high detection P value

- median `P > 0.01` across samples, `P > 0.01` in nth% of samples

remove probes overlapping SNPs

- `minfi::dropLociWithSnps`
- `MAF > 0.05`

drop probes in X, Y chromosome

remove cross reactive probes

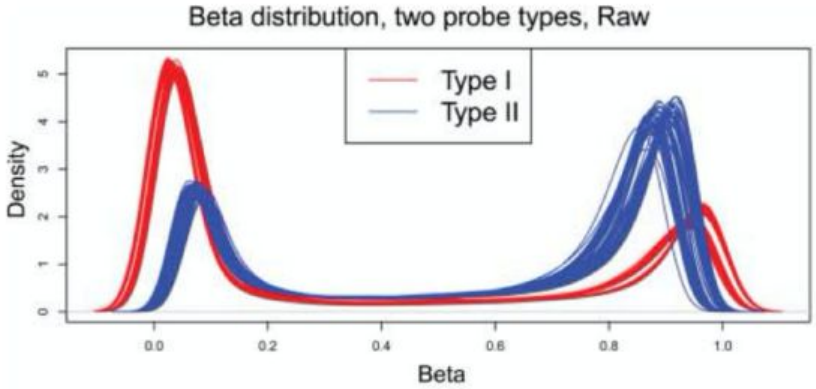
Preprocessing, correction and normalisation

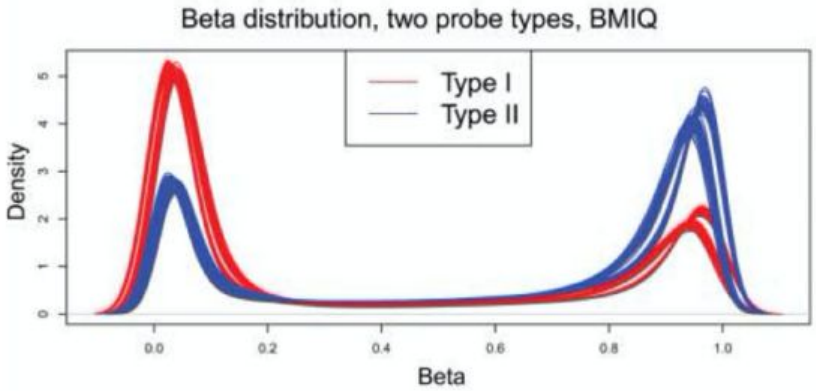
Within array:

- Dye bias correction
- Background correction
- Type I/II probe bias correction

Across array:

- starting material
- labelling efficiency





Normalisation in minfi

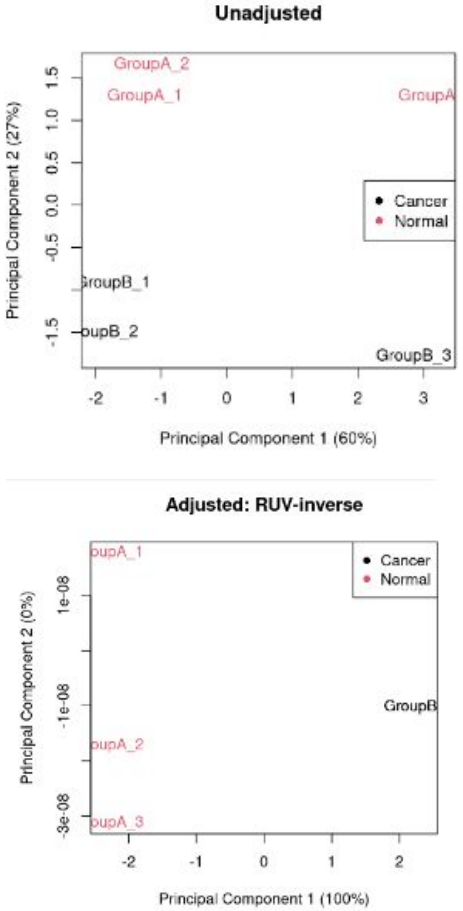
	Dye Bias	Background Correction	Type I/II Probe bias	Within array normalisation	Across array normalisation
preprocessRaw	-	-	-	-	-
preprocessIllumina	✔	✔	-	✔	✔
preprocessSWAN	-	-	✔	✔	-
preprocessQuantile	-	-	✔	✔	✔
preprocessNoob	✔	✔	-	-	-
preprocessFunnorm	✔	✔	✔	✔	✔

Differentially Methylated Probes (DMPs)

- Perform a statistical test to find any significant association between the methylation state of a CpG and the phenotype of interest
  - T-test, ANOVA
  - Wilcoxon rank-sum, Kruskall Wallis test, Fisher's exact test
  - Linear model, logistic regression, mixed effects model, Beta Binomial

- `minfi` uses `limma` functionality
  - Use M-values
- Phenotype could be categorical or continuous
  - cancer vs normal, between tissue types, smokers vs not
  - age, blood pressure, BMI

- Other sources of variation can be accounted for
  - E.g., plate-to-plate, lot-to-lot variance
  - covariate in the model
  - ComBat
  - SVA (Surrogate Variable Analysis)
  - RUVSeq



Differentially Methylated Regions (DMRs)

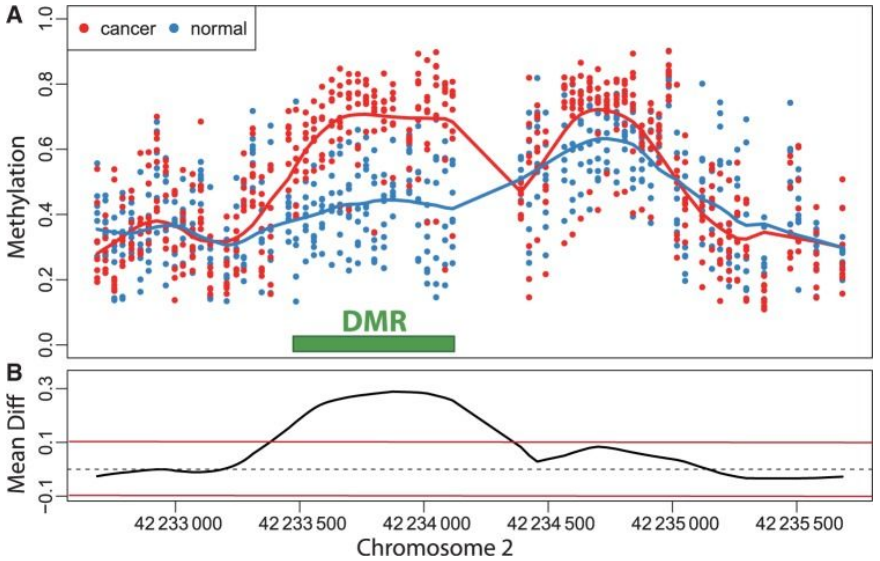
DMRs can be defined:

heuristically / De novo

Bumphunter, DMRcate, ProbeLasso, comb-p

functional units

CpG islands, gene bodies, promoters



WGBS

Sequencing-based methylation profiling

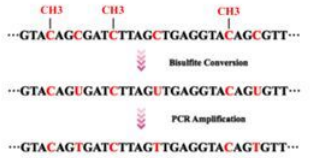
	Enzyme digestion	Affinity enrichment	Sodium bisulfite
Principles	Some restriction enzymes, such as *HpaII* and *SmaI*, are inhibited by $5^{me}C$ in the CpG.	Affinity enrichment uses antibodies specific for $5^{me}C$ or methyl-binding proteins with affinity for profiling of DNA methylation.	Sodium bisulfite chemically turns unmethylated cytosine into uracil, hence enabling methylation detection.
Method example	Methyl-seq *MCA-seq *HELP-seq *MSCC	*MeDIP-seq *MIRA-seq	*RRBS *WGBS *BSPP

*MCA: methylated CpG island amplification; *HELP: HpaII tiny fragment enrichment by ligation-mediated PCR; *MSCC: methylation-sensitive cut counting; *MeDIP-seq: methylated DNA immunoprecipitation; *MIRA: methylated CpG island recovery assay; *RRBS: reduced representation bisulfite sequencing; *WGBS: whole genome bisulfite sequencing; *BSPP: bisulfite padlock probes.

WGBS workflow

Tissue, Blood, etc

DNA Extraction



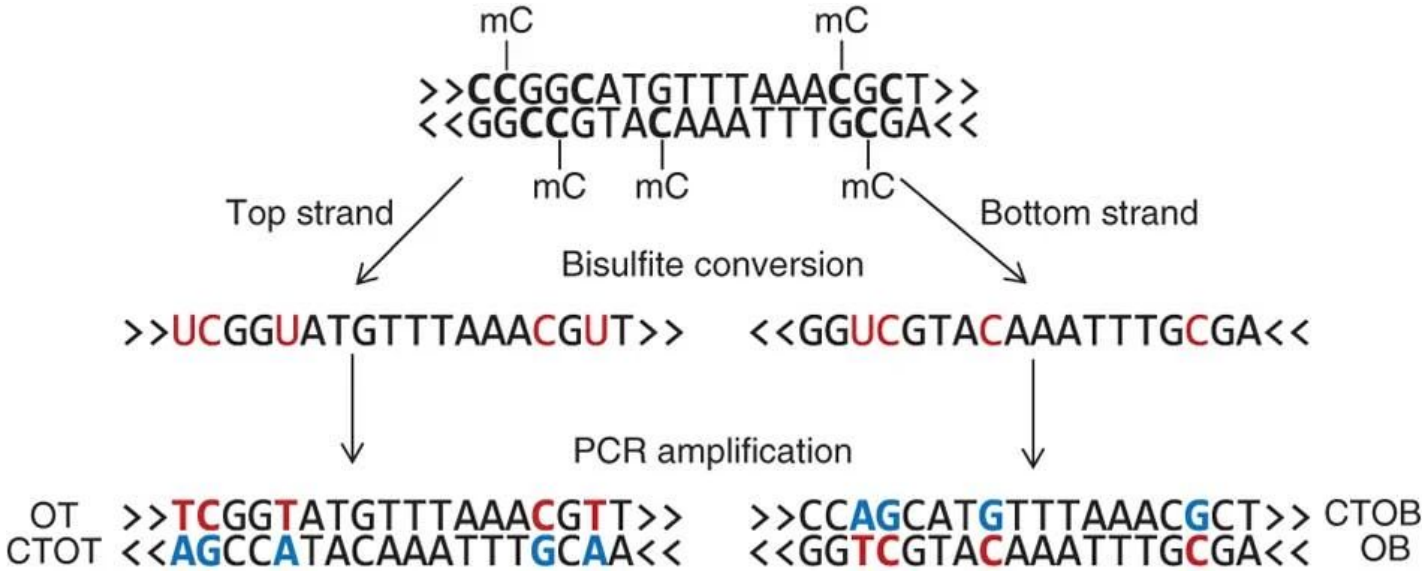
Bisulfite Treatment

Library Construction

Sequencing

Data Analysis

Effect of bisulfite treatment of DNA

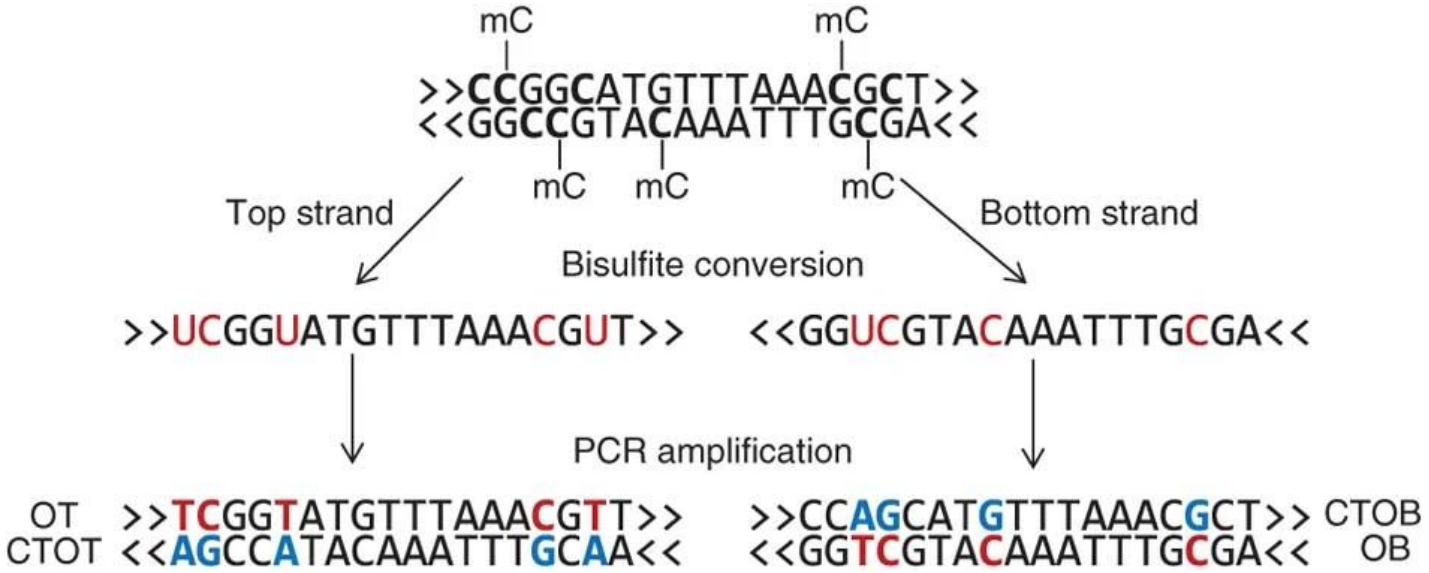


2 different PCR products and 4 different sequence strands from one genomic locus

Each of these 4 sequence strands can theoretically exist in any possible conversion state

Ref: https://www.nature.com/articles/nmeth.1828

Effect of bisulfite treatment of DNA



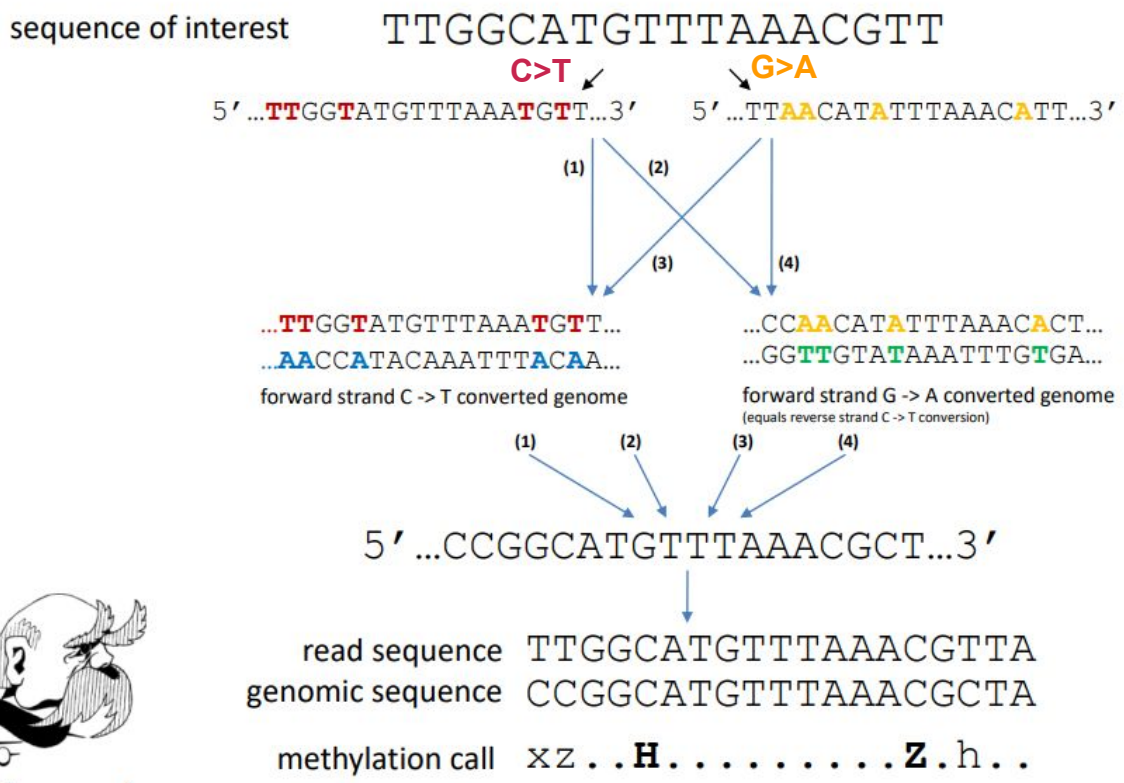
Non-directional library: all four strands

Directional library: OT and OB

PBAT library: CTOT and CTOB

Ref: https://www.nature.com/articles/nmeth.1828

3 letter alignment of BS-converted reads



Fully bisulfite convert read
(as both forward and reverse strand)

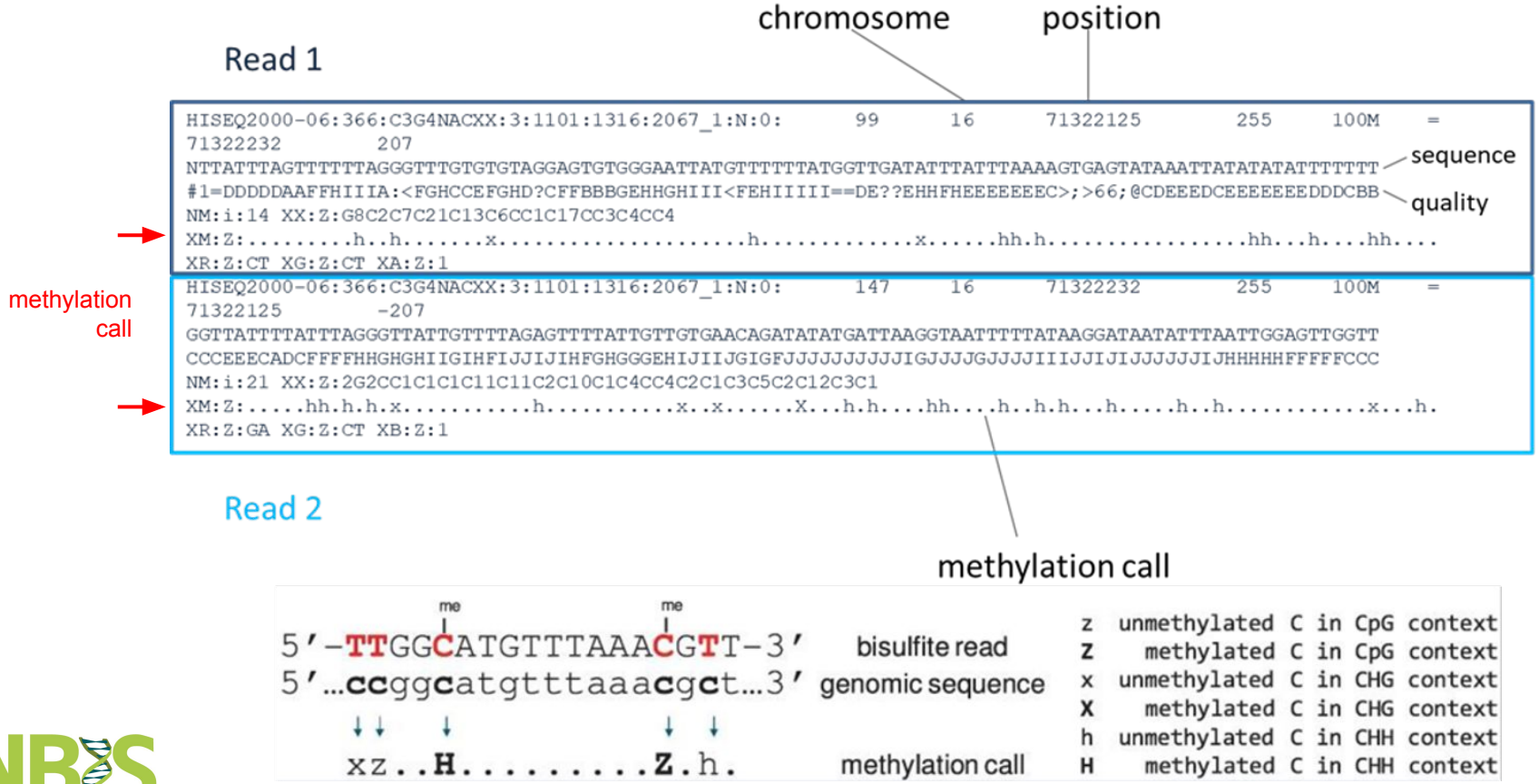
Align to bisulfite converted genomes

Read all 4 alignment outputs and extract the unmodified genomic sequence if the sequence could be mapped uniquely

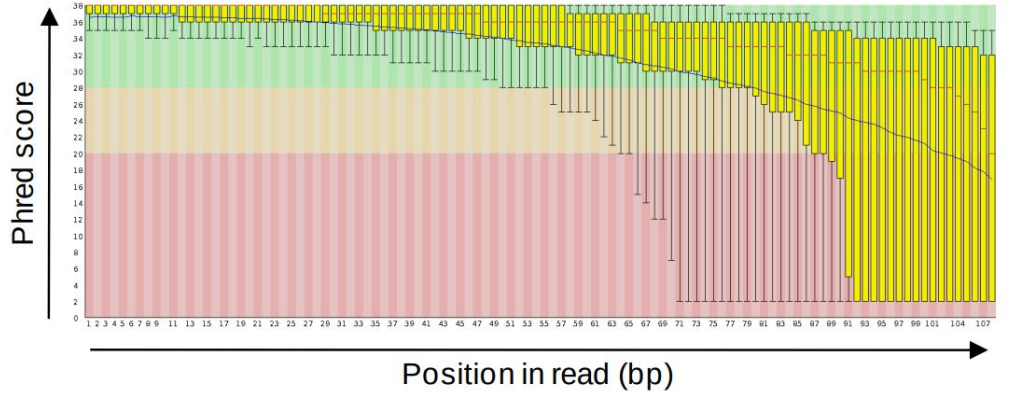
Methylation Call

h unmethylated C in CHH context
H methylated C in CHH context
x unmethylated C in CHG context
X methylated C in CHG context
z unmethylated C in CpG context
Z methylated C in CpG context

Methylation calls

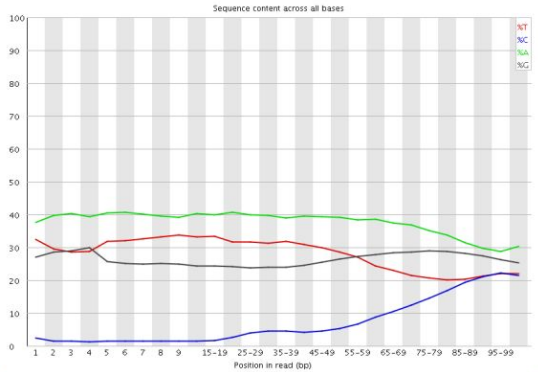


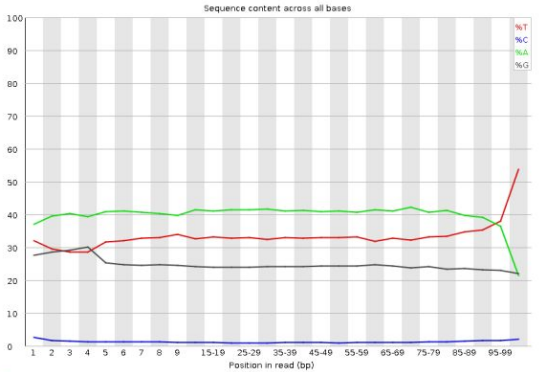
Quality control: base call quality



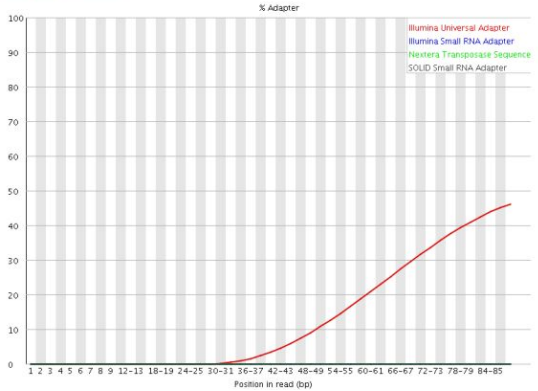
Quality control: adapter contamination

before trimming **after trimming**

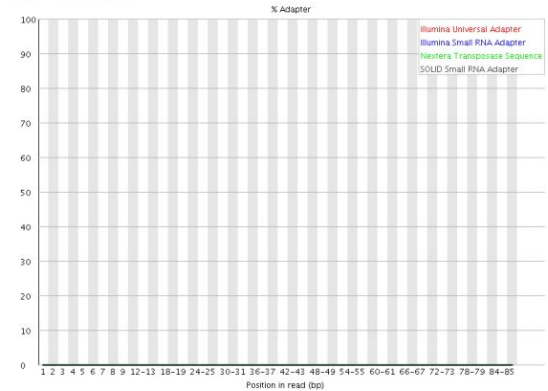




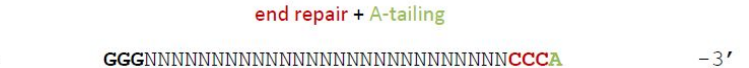
Adapter Content



Adapter Content



mapped Quality control - end repair bias



mapped Quality control

Bisulfite conversion efficiency

- non CpG sites in mammalian genome should have > 99.5% conversion in a good experiment
- Spike-in controls e.g., phage Lambda

DNA degradation during bisulfite conversion

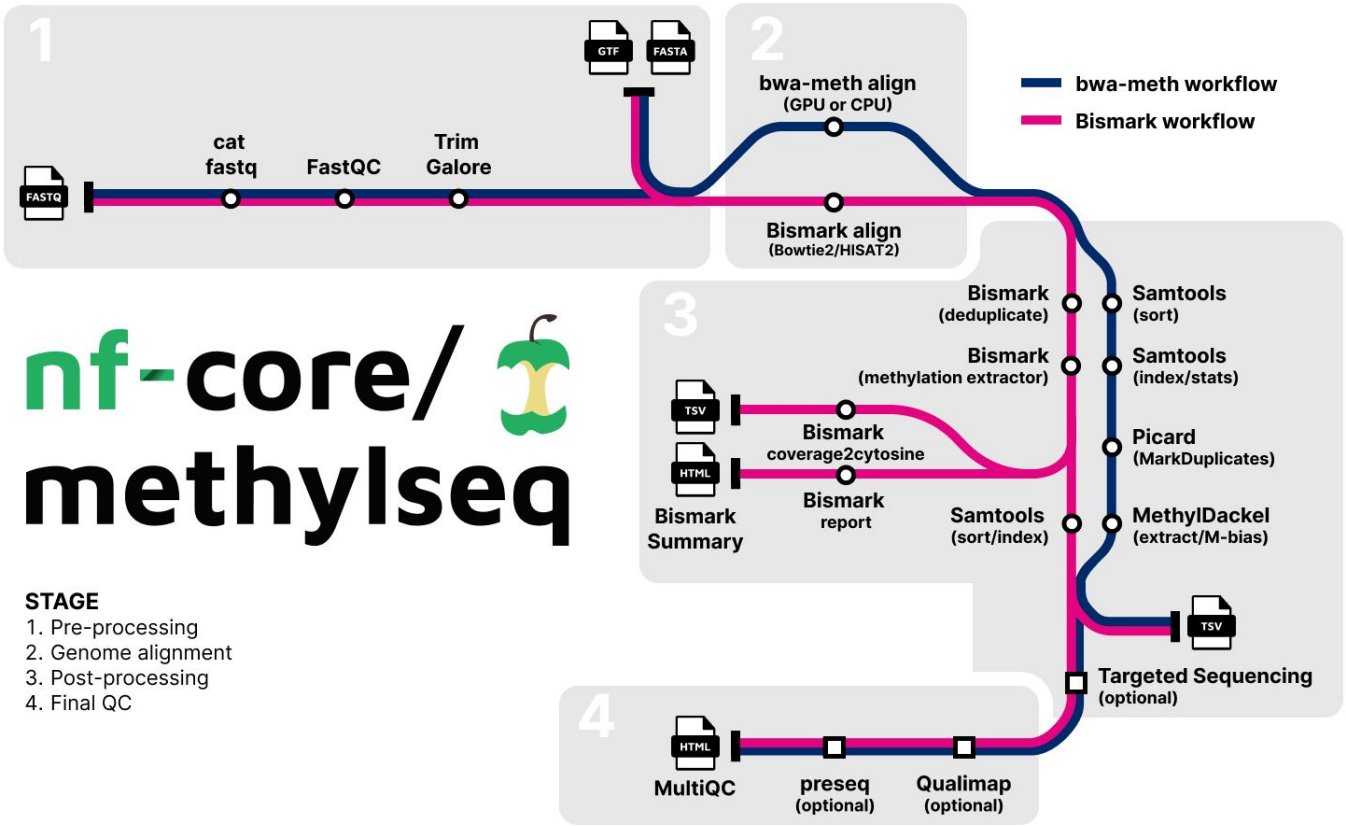
- unique alignment rates
- read length after trimming

Remove C>T SNPs

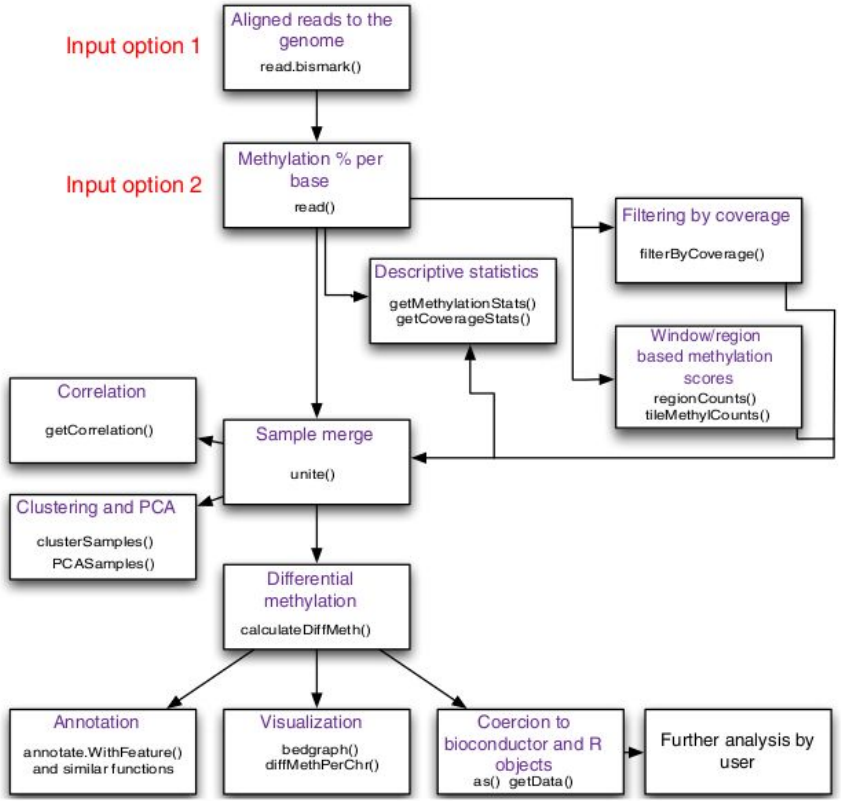
Deduplication

- recommended for WGBS, not for RRBS

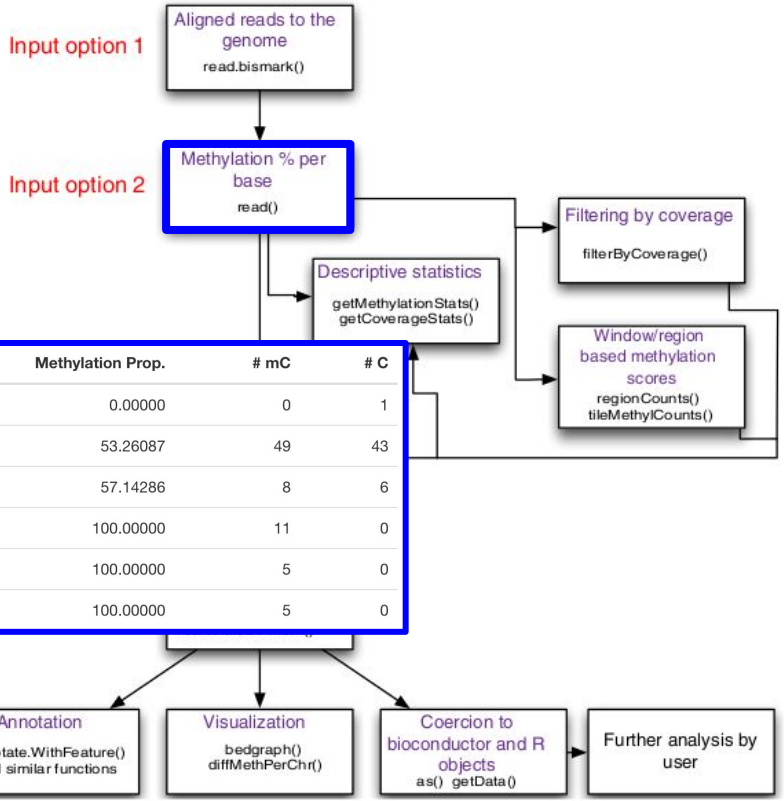
nf-core/methylseq is a bioinformatics analysis pipeline for BS-seq data



Analysis workflow for WGBS data

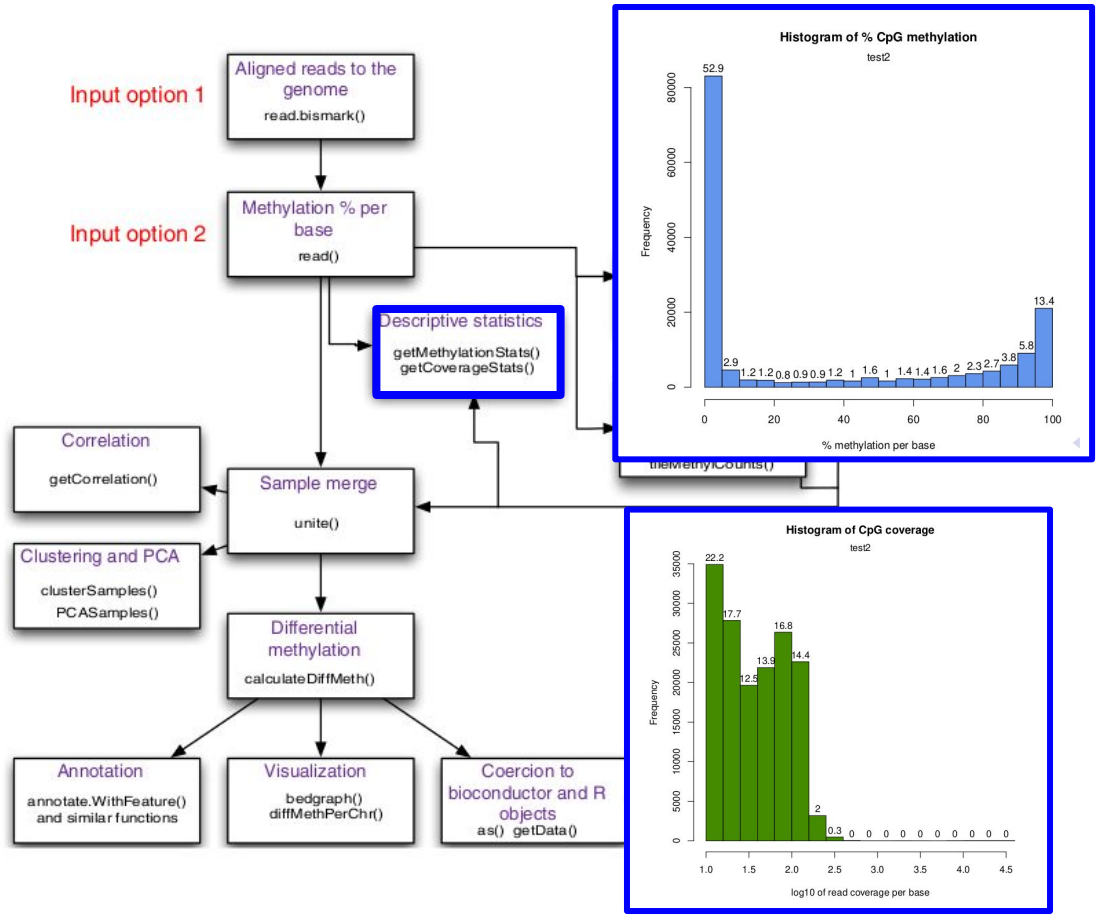


Analysis workflow for WGBS data



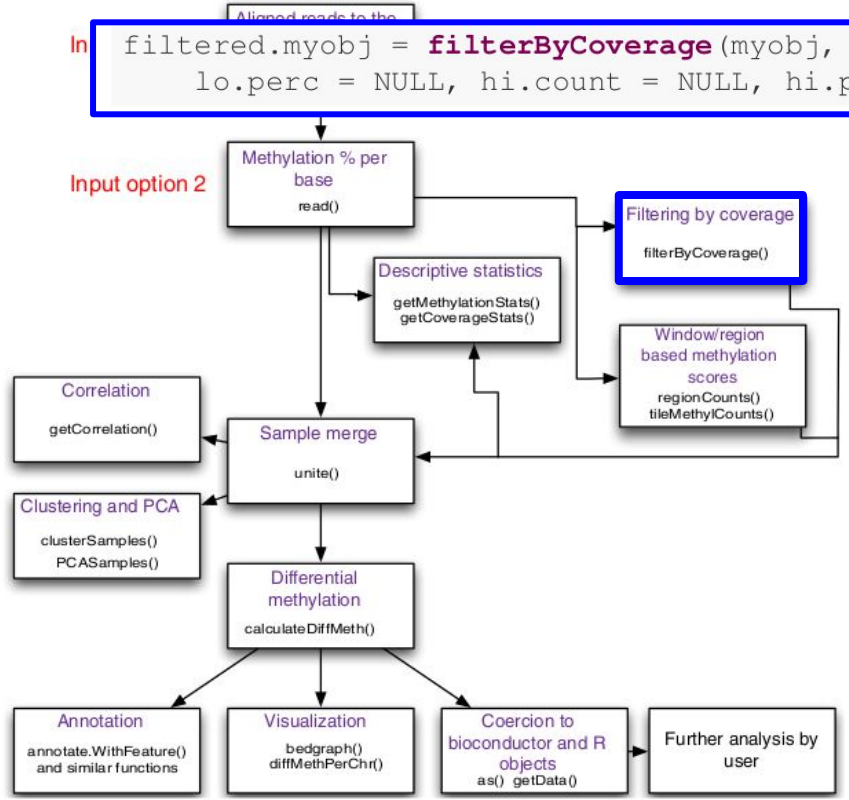
Chr	Start	End	Methylation Prop.	# mC	# C
chr8	3052997	3052997	0.00000	0	1
chr8	3052998	3052998	53.26087	49	43
chr8	3068732	3068732	57.14286	8	6
chr8	3068733	3068733	100.00000	11	0
chr8	3089948	3089948	100.00000	5	0
chr8	3089984	3089984	100.00000	5	0

Analysis workflow for WGBS data

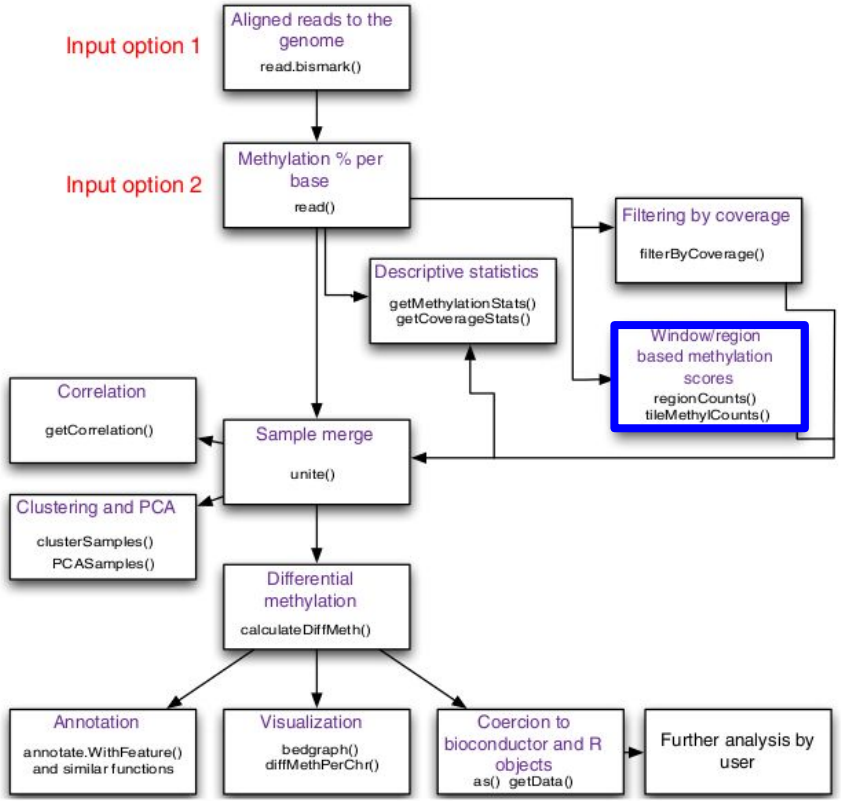


Analysis workflow for WGBS data

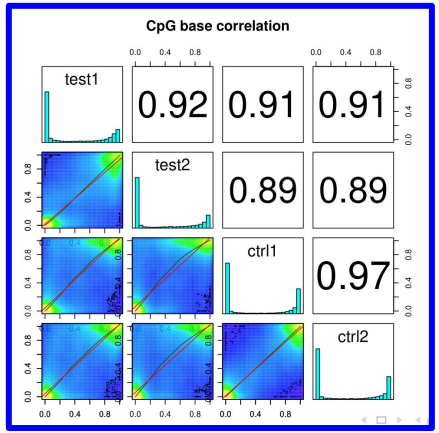
```
filtered.myobj = filterByCoverage(myobj, lo.count = 10,
    lo.perc = NULL, hi.count = NULL, hi.perc = 99.9)
```

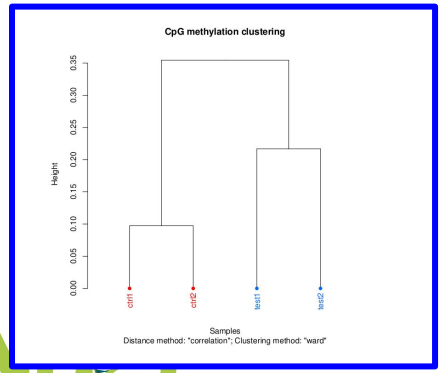


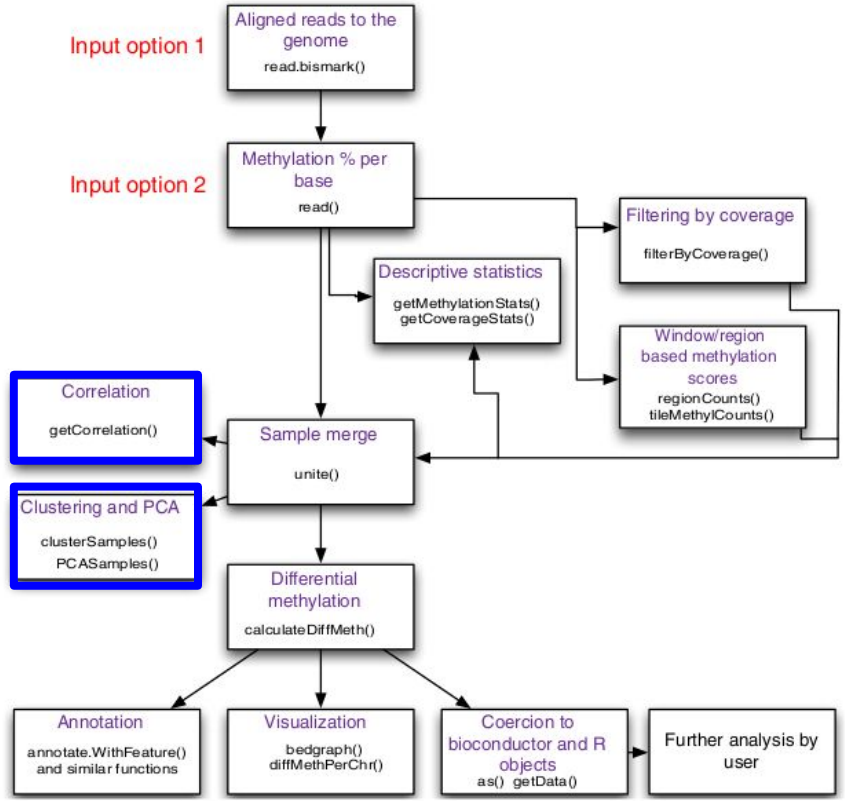
Analysis workflow for WGBS data

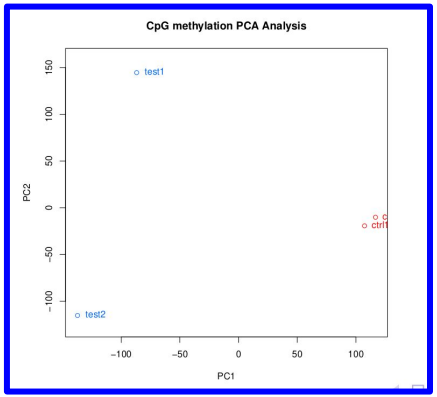


Analysis workflow for WGBS data









Differential Methylation

Remove CpGs with little variation

Remove CpGs that overlap C>T SNPs

No replicates: Fisher's exact test

With replicates:

- Logistic regression
- Beta Binomial

Overdispersion correction

Covariates can be included in the model