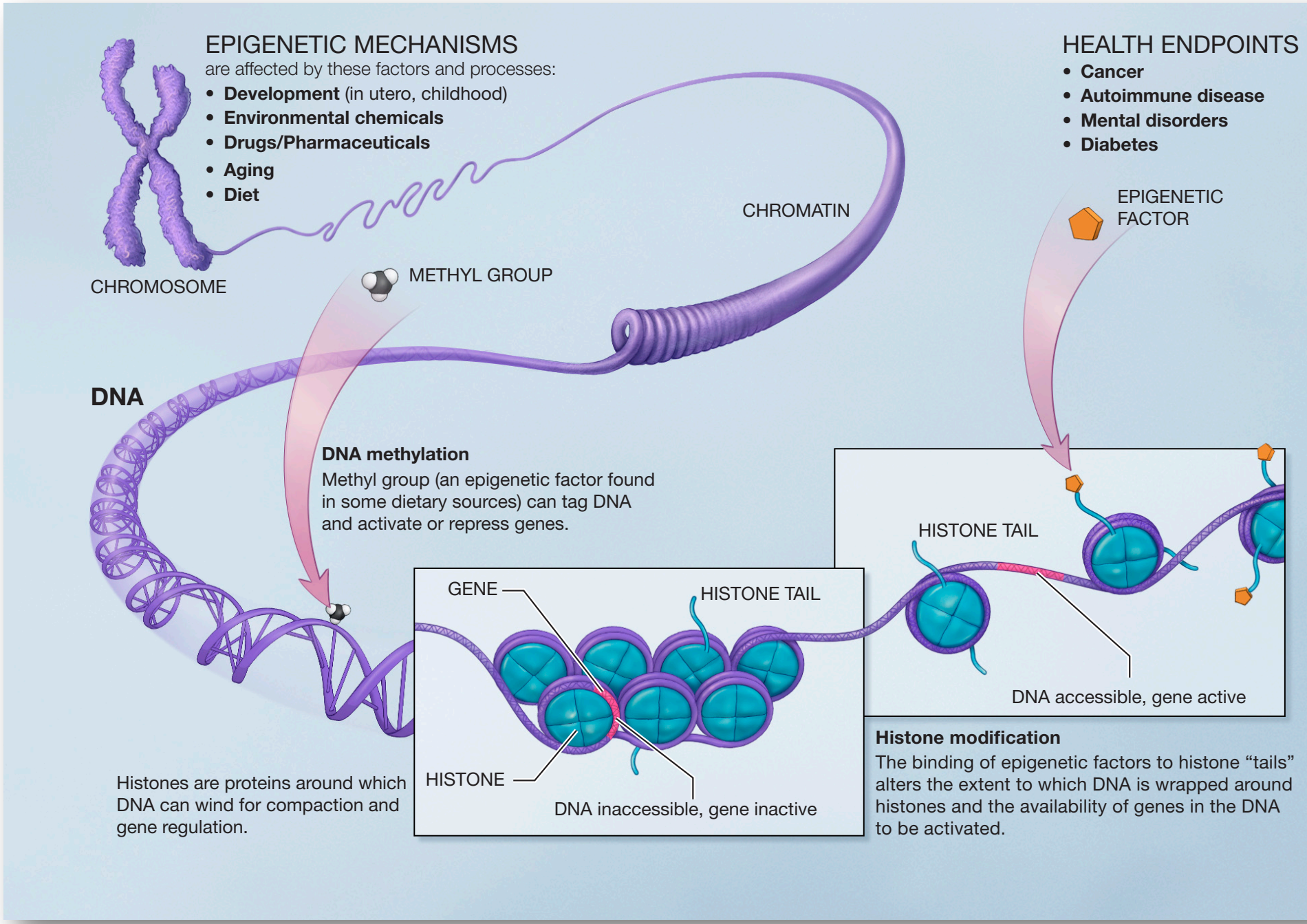
Schedule

- 09.30 - 09.45 **Introduction to methylation**
- 09.45 - 10.15 Computer exercises set-up + break
- 10.15 - 10.30 Methylation Exercises Overview I: Array workflow
- 10.30 - 12.00 Exercises Array
- 12.00 - 13.00 lunch
- 13.00 - 14.00 Methylation methods & technologies
- 14.00 - 14.30 Methylation Exercises Overview II: Methylation Sequencing
- 14.30 - 14.45 Break
- 14.45 - 16.30 Exercises Sequencing
- 16.30 - 17.00 Test yourself

Introduction to Methylation

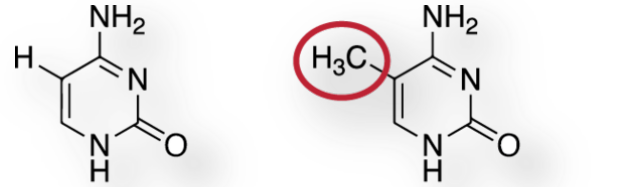
Epigenomics Workshop 2020

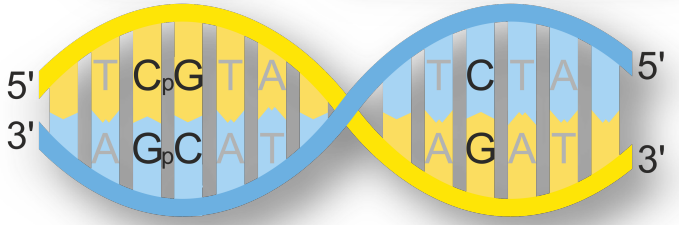

EPIGENETIC MECHANISMS
are affected by these factors and processes:
• Development (in utero, childhood)
• Environmental chemicals
• Drugs/Pharmaceuticals
• Aging
• Diet
HEALTH ENDPOINTS
• Cancer
• Autoimmune disease
• Mental disorders
• Diabetes
CHROMOSOME
CHROMATIN
METHYL GROUP
EPIGENETIC FACTOR
DNA
DNA methylation
Methyl group (an epigenetic factor found in some dietary sources) can tag DNA and activate or repress genes.
GENE
HISTONE TAIL
HISTONE
DNA inaccessible, gene inactive
HISTONE TAIL
DNA accessible, gene active
Histones are proteins around which DNA can wind for compaction and gene regulation.
Histone modification
The binding of epigenetic factors to histone "tails" alters the extent to which DNA is wrapped around histones and the availability of genes in the DNA to be activated.

A brief history of DNA methylation

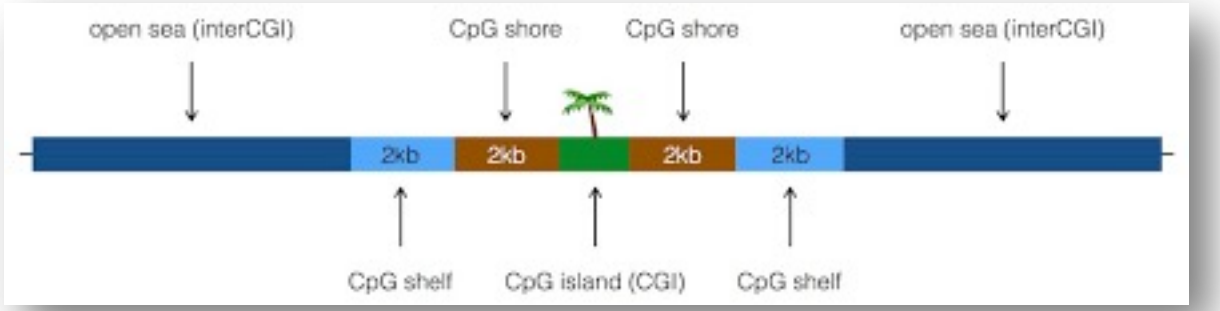
- 1944: Avery identifies DNA as gene material
- 1948: Hotchkiss discovered 5-methylcytosine

"minor constituent designated epicytosine [with] a migration rate somewhat greater than that of cytosine"

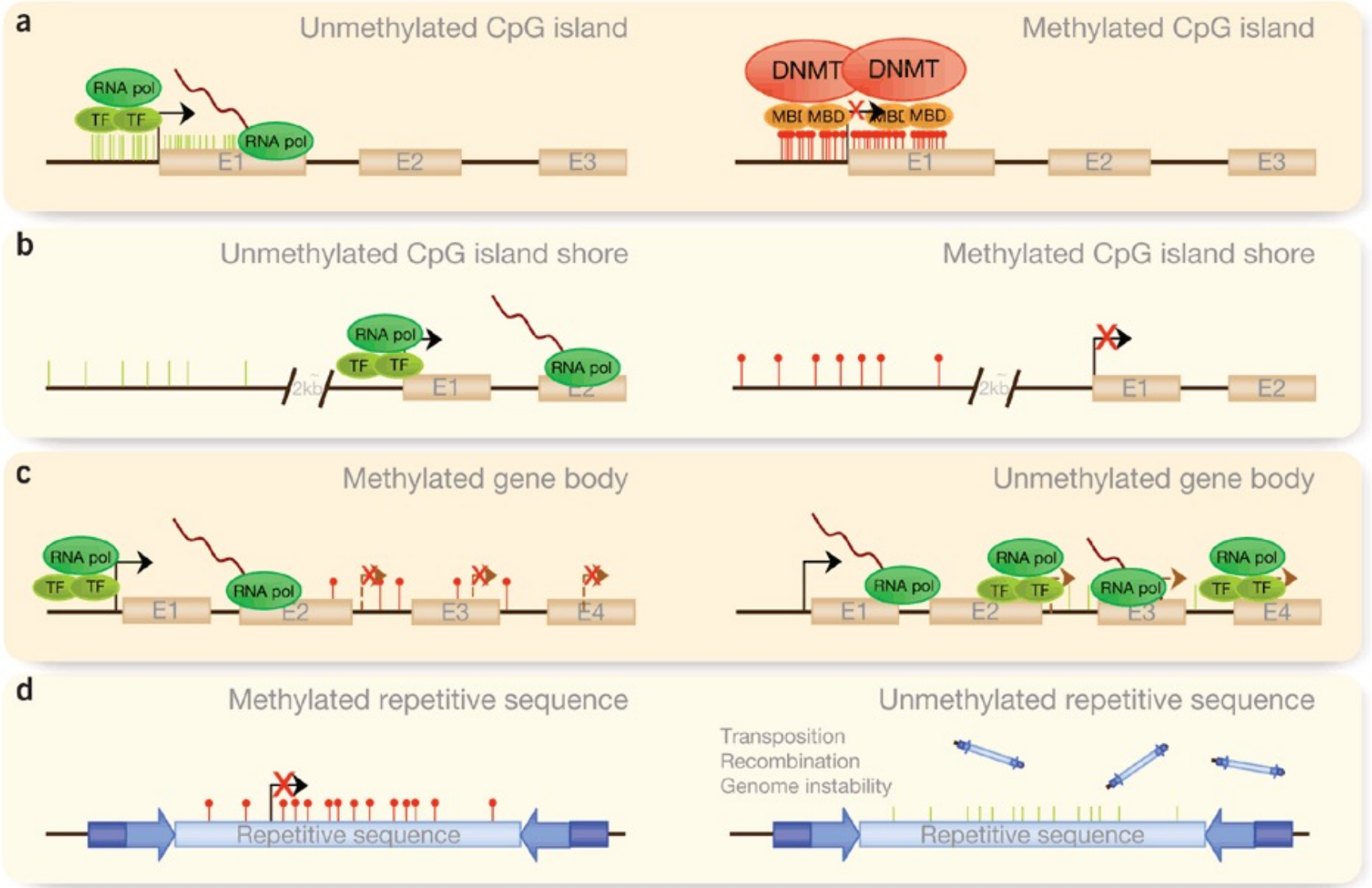




- 1953: DNA structure resolved
- 1962: Methylation in CpG dinucleotides
- Around 1980: DNA methylation regulates gene expression
  - McGhee & Ginder: Beta-globin expression
  - Jones & Taylor: Cytidine analogs
- Early 1980s: CpG Island discovery



Methylation in health and disease



Methods commonly used for detection of DNA methylation

A Site-Specific Methylation

Method	Refs	Outcome
* Methylation-specific PCR	Herman *et al.*,1996 Karouzakis *et al.*,2009	Detection of methylation of a spesific gene or a region
* Quantitative PCR (i.e.High Melting Resolution analysis)	Candiloro *et al.*, 2011 Newman *et al.*, 2012 Kristensen *et al.*, 2013	Methylation level of a specific gene/regions of the genome
* COBRA (combined bisulfite and restriction analysis)	Xiong and Laird, 1997 Lahtz *et al.*, 2013	
* Pyrosequencing	Candiloro *et al.*, 2011 Kristensen *et al.*, 2013	Quantification of methylation frequencies at individual consecutive CpG sites

B Global Methylation

Method	Refs	Outcome
* Methylation-specific PCR of repetetive sequences	Yang *et al.*, 2004	Methylation status of repetitive sequences of the genome
* HPLC	Kuo *et al.*, 1980 Ehrlich *et al.*, 1982	Total amount of methylated cytosines in the genome
* HPCE	Li et al., 2009	
* Mass spectrometry	Annan *et al.*, 1989 Coolen *et al.*, 2007	
* Anti-5meC immunological methods (Flow cytometry, microscopy etc.)	Habib *et al.*, 1999 Piyathilake *et al.*, 2004 Brown *et al.*, 2008 Karouzakis *et al.*, 2009 Schneider and Fagagna, 2012	
* Microarray	Weber *et al.*, 2005 Bar-Nur *et al.*, 2011 Bocker *et al.*, 2011 Walker *et al.*, 2011	Genome-mapping (methylation status of large DNA fragments)
* Next-generation sequencing (i.e. Illumina platform)	Bibikova *et al.*, 2009 Russnes *et al.*, 2011 Zong *et al.*, 2012 Glossop *et al.*, 2013 Renner *et al.*, 2013	(1) Methylation status of individiual CpG dinucleotides, (2) Methylation status of gene regions with sites in the promoter region, 5'UTR, first exon, gene body, 3'UTR, and (3) Methylation status of CpG islands, shore and shelf regions (distance from the CpG islands), and non-CpG islands of the genome

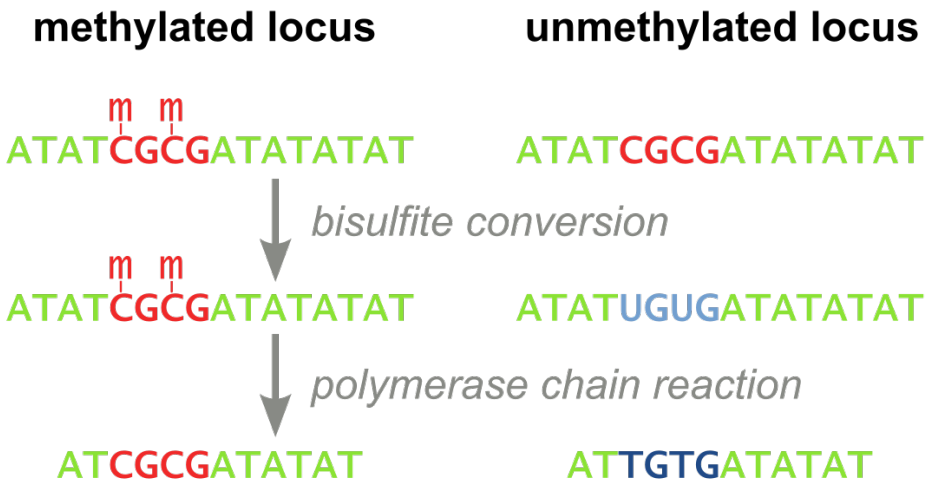
C Global Methylation Detection Using Proxy Markers

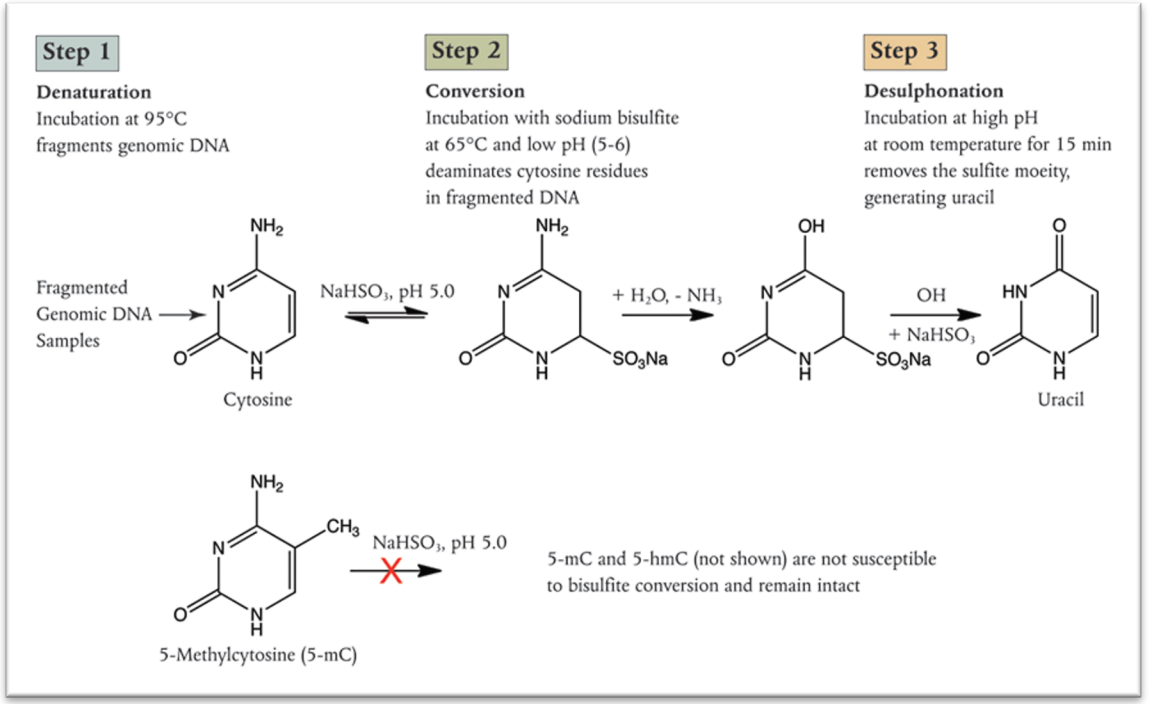
Marker	Example Methods	Interpretation	Refs
MBD1	* MBD domain of MBD1 attaches to a luciferase sensor (luminometer)	Global DNA methylation	Badran *et al.*, 2011
	* Dot blot analysis of MBD1 protein		Zhang *et al.*, 2012
	* Illumina sequencing of methylated DNA enriched by the MBD domain of MBD1		Morita *et al.*, 2012

Celik et al. (2014), Journal of Immunological Methods

Bisulfite Conversion

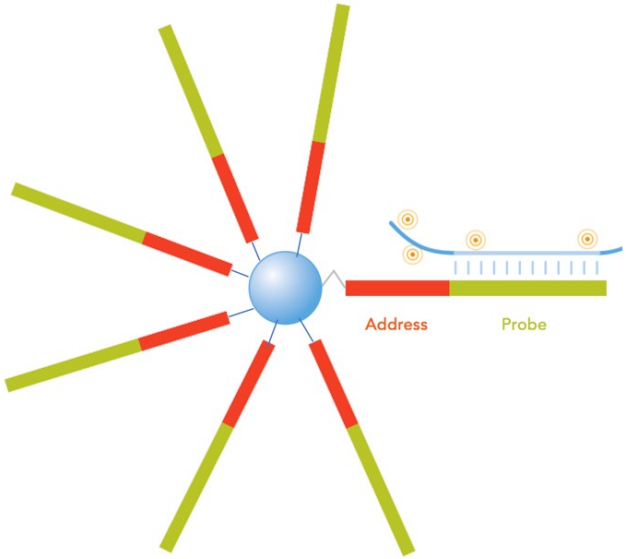
- Bisulfite treatment crucial for both arrays and NGS
- C -> U (-> T)
- mC -> mC (-> C)
- methylation-specific PCR, high resolution melting curve analysis, microarray-based approaches, and next-generation sequencing

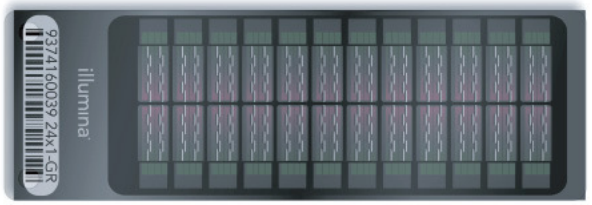




Methylation Arrays

- Infinium Methylation BeadChip arrays from Illumina: 27K, 450K and 850K (or EPIC)
- >480,000 CpG loci, covers 99% of RefSeq genes
- Distributed over various functional elements; covers 96% of CGI
- 50bp single-stranded DNA oligos (“probes”) attached to silica beads, 2 detection channels (red and green)
- Hybrid of 2 different probe designs!



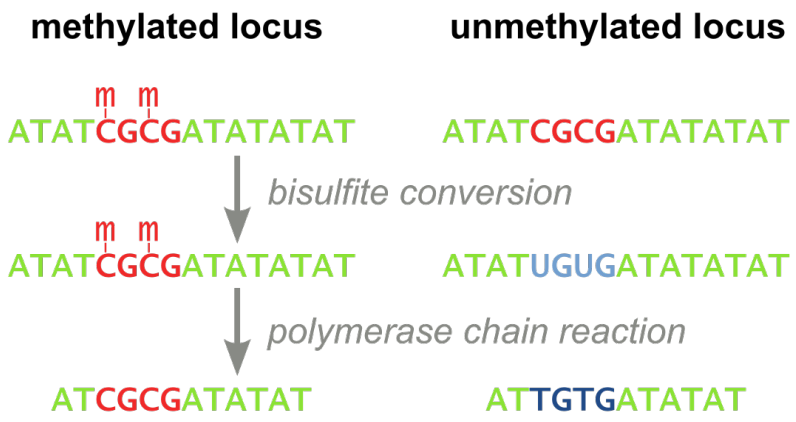


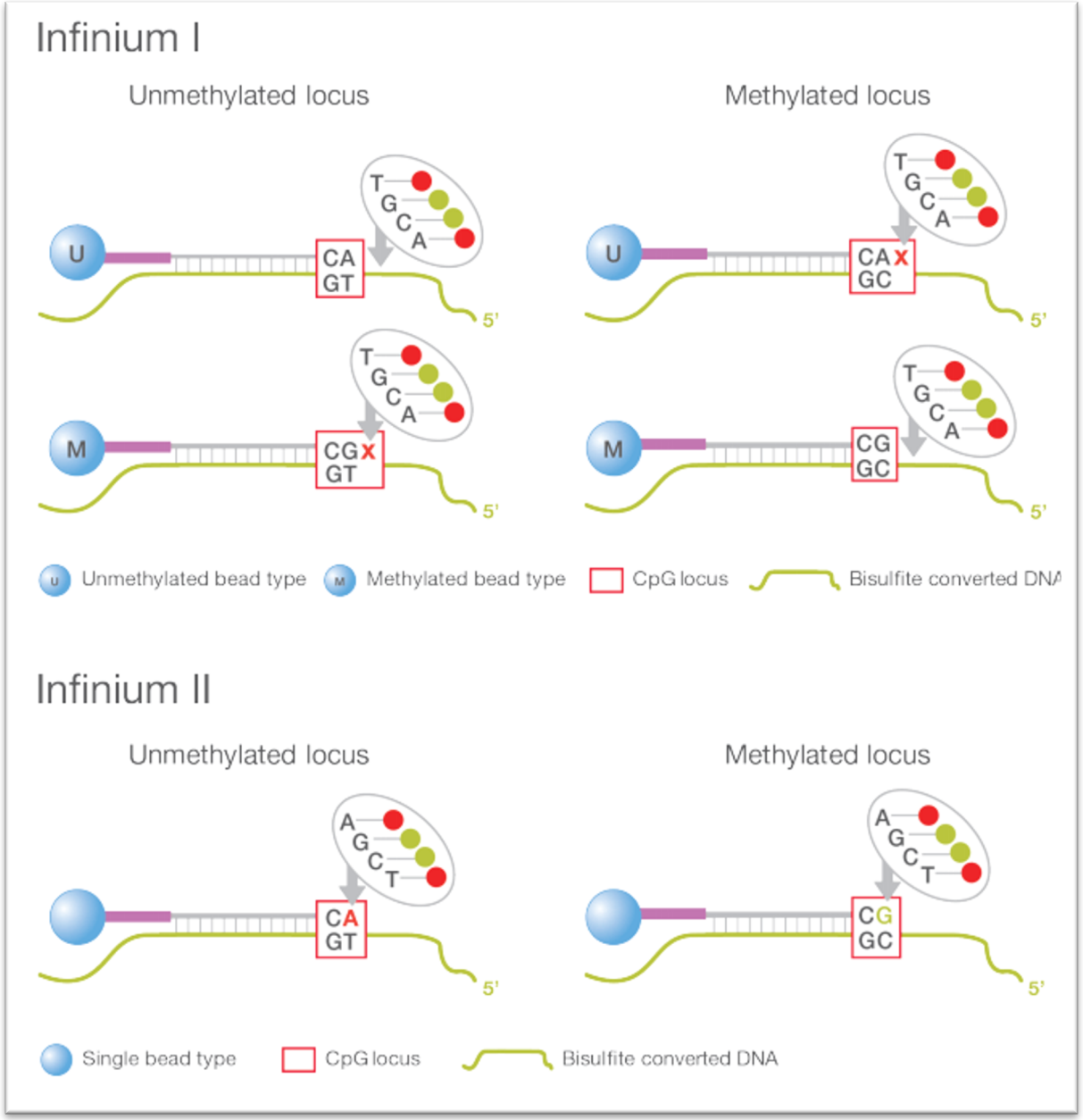
BeadChip Array

Infinium: Type I vs II

- Type I: single color detection, two beads
- Type II: two color detection, single bead

Type I	Type II
Only probes on 27K	New for 450K
2 beads/CpG	1 bead/CpG (fit more)
Better for CpG dense regions	Better for less CpG dense regions
More stable and reproducible	Lower dynamic range





From colors to methylation

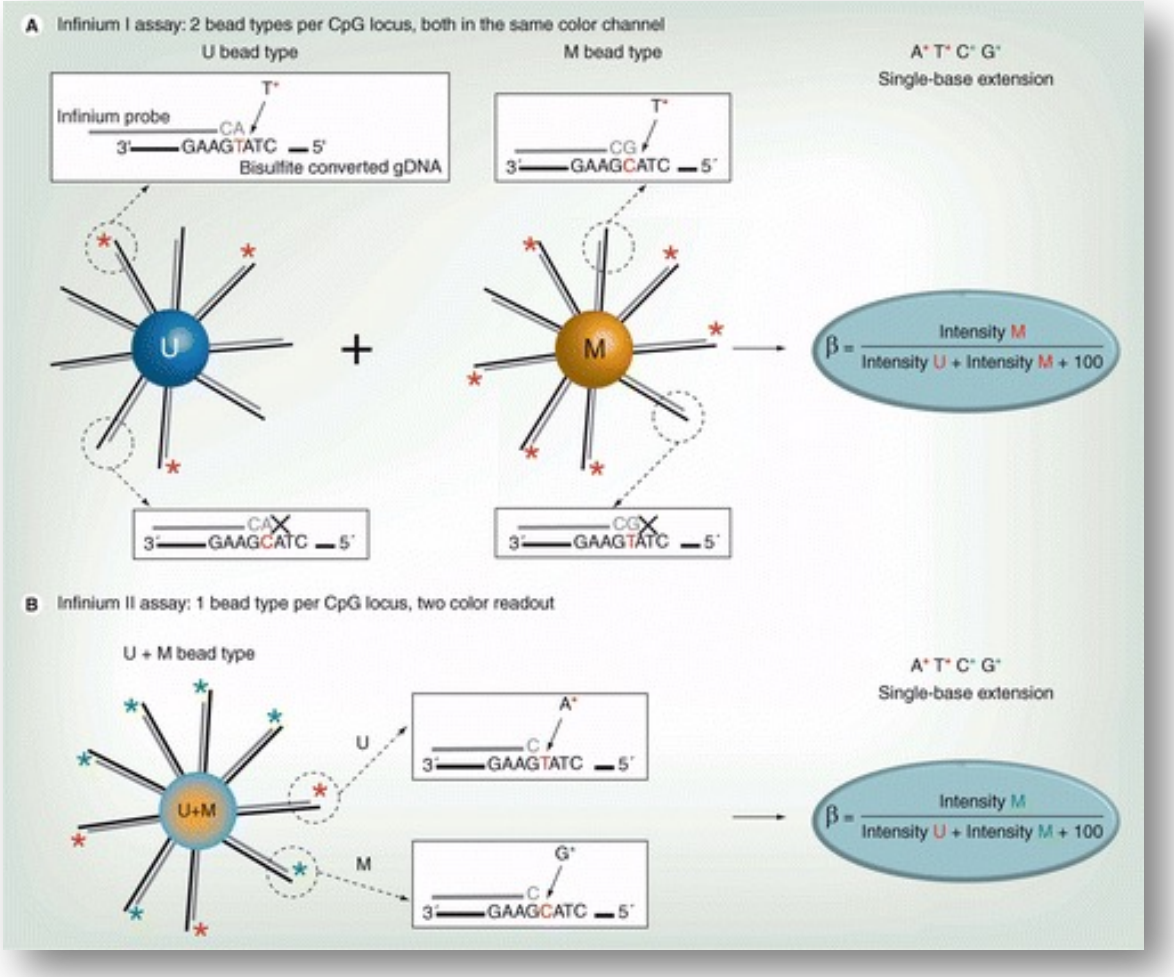
- Intensities are usted to estimate Beta values; for both probe designs

$$\beta = M/(M+U+100)$$

- Beta value between 0 and 1 (represents the fraction methylation)
- Easily interpretable, but related M value has better statistical properties

$$Mvalue = log2(M/U)$$

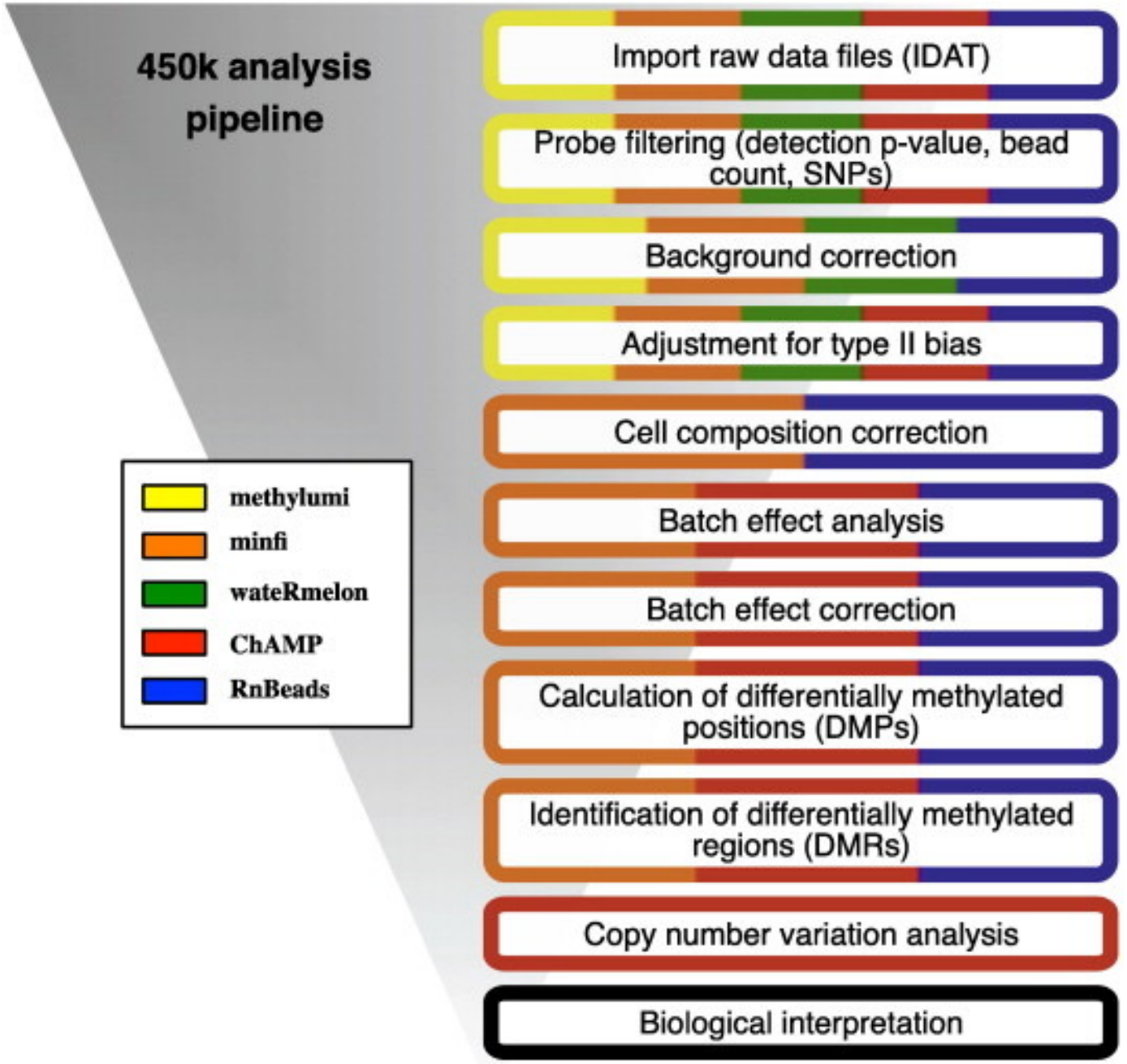
This step and the next will be part of the tutorial for downstream analysis, so time to get your setup ready!



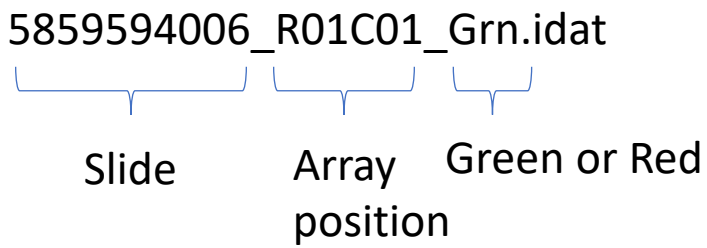
Methylation Array Workflow

Typical workflow

- Tools for analyzing Illumina arrays
- Provides tools for many of the steps presented here.
- minfi



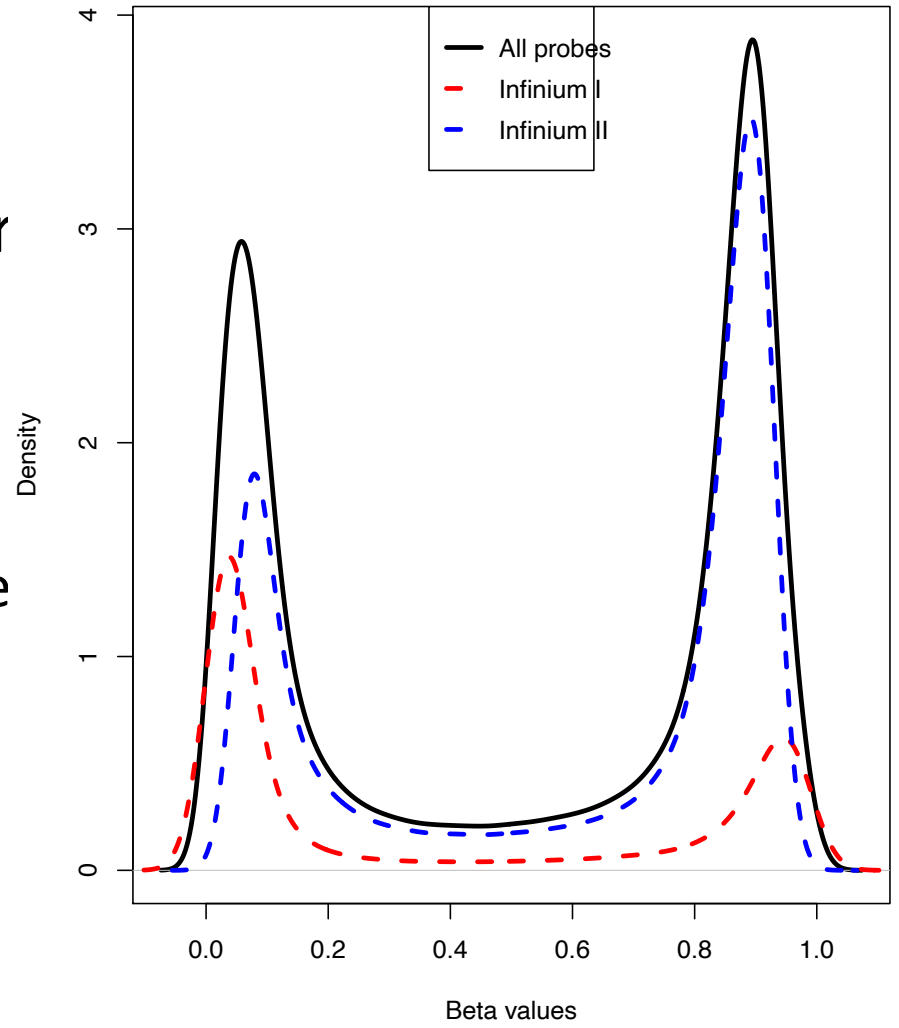
Import Data



- IDAT files; slide scanner output
- Needs a SampleSheet, usually accompanies array data (or can be made manually)
- Raw intensities -> *RGChannelSet*
- Needs to be converted to *MethylSet* for initial QC

QC + Filtering

- Aim: find outliers/batch effects and artifacts and try to remove or account for them
- Several metrics:
  - Plot distributions of the Beta values
  - Quality of probes: average detection p-value



QC + Filtering

- Aim: find outliers/batch effects and artifacts and try to remove or account for them
- Several metrics:
  - Plot distributions of the Beta values
  - Quality of probes: average detection p-value
  - Internal quality control probes
  - Remove probes with known SNPs
  - MDS/PCA plot

- STAINING CONTROLS
- BISULFITE CONVERSION CONTROLS
- EXTENSION CONTROLS
- SPECIFICITY CONTROLS
- HYBRIDIZATION CONTROLS
- TARGET REMOVAL CONTROLS
- NON-POLYMORPHIC CONTROLS
- NEGATIVE CONTROLS

Normalization

- Within and across array normalization

Considerations for normalization of DNA methylation data by Illumina 450K BeadChip assay in population studies

Paul Yousefi, Karen Huen, Raul Aguilar Schall, Anna Decker, Emon Elboudwarej, Hong Quach, ...show all

Between-array normalization for
a

nd Hermann Brenner [1, 2]

ogy and Aging Research, German Cancer Research Center (DKFZ), Heidelberg, Germany,

Functional normalization of 450k methylation array data improves replication in large cancer studies

Jean-Philippe Fortin[1], Aurélie Labbe[2,3,4], Mathieu Lemire[5], Brent W Zanke[6], Thomas J Hudson[5,7], Elana J Fertig[8], Celia MT Greenwood[2,9,10] and Kasper D Hansen[1,11*]

A systematic assessmen
approaches for the Infin
platform

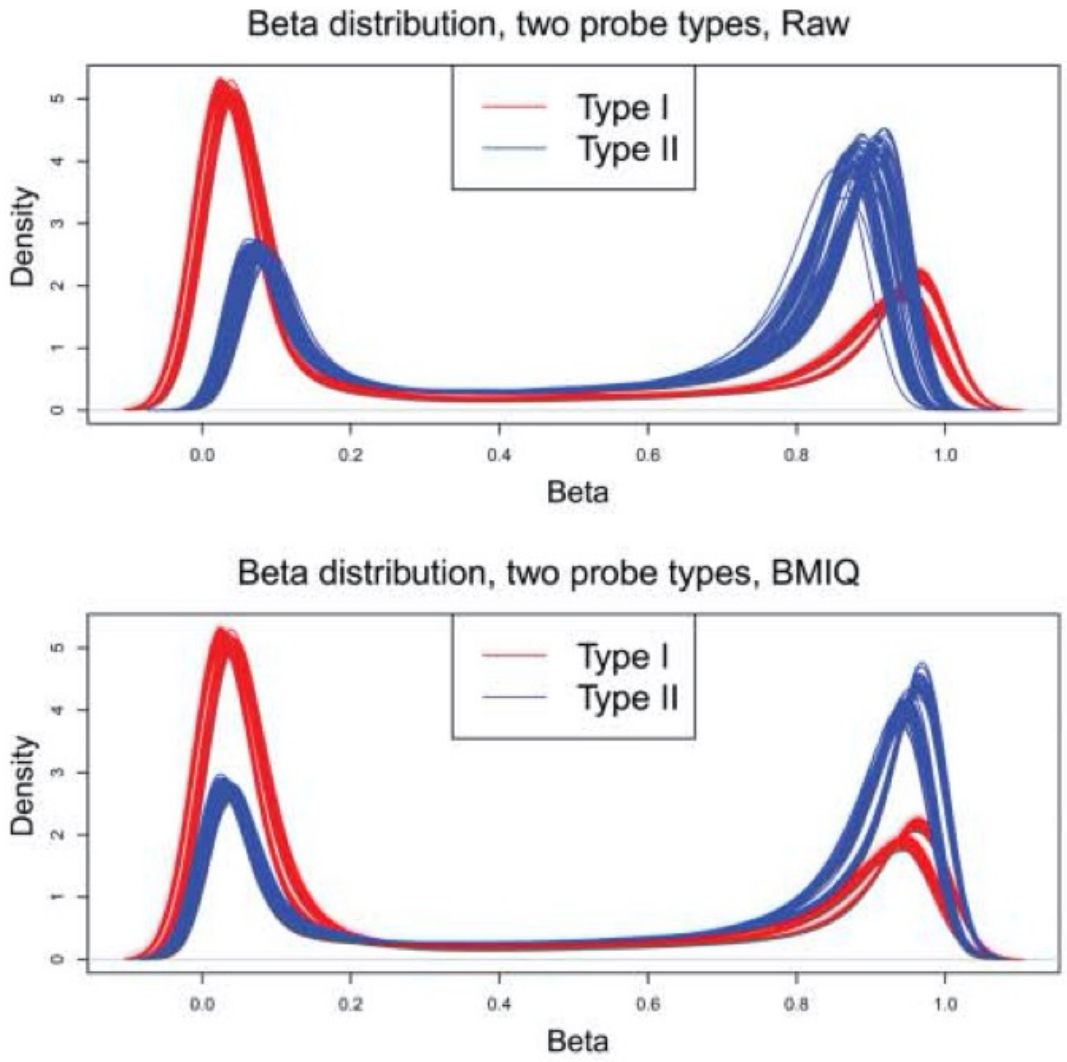
Michael C Wu, Bonnie R Joubert, Pei-fen Kuan, Siri E Håberg, Wenche Nystad, Shyamal D Peddada & Stephanie J London

on methods for
ethylation data using whole-
e sequencing data

Nadia Boutaoui, Glorisa Canino, Jianhua Luo, ...show all

Normalization

- Within and across array normalization
- Depends on biological signal.
- Within-array normalization not essential when doing CpG level analysis.

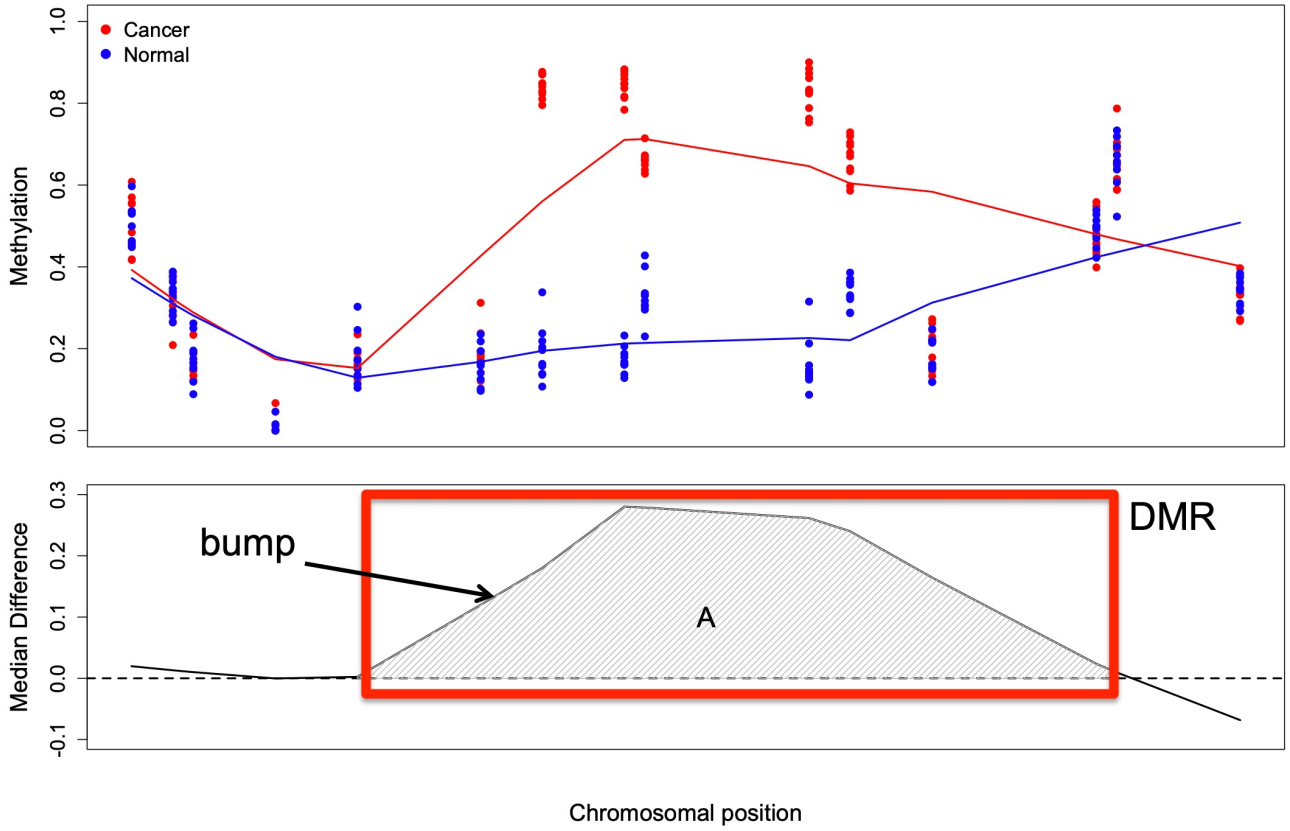


Differential Methylation

- Identification of systematic differences in methylation between groups of samples (i.e., case vs control, smokers vs non-smokers, ...)
- Countless ways to approach this, depending on:
  - Question(s) being asked
  - Available information on potential confounders
  - Nature/structure of the data (repeat measurements, …)
- Some possible approaches include:
  - T-tests and ANOVA models
  - Wilcoxon rank-sum and Kruskal Wallis tests
  - Linear, logistic and Cox regression
  - Mixed effects models
  - Surrogate Variable Analysis (SVA)
- Use M-values: $Mvalue{=}log2(M/U)$
  - More homoscedastic

Differential Methylation

- Single CpG can be useful (DMP), but often regions or block of CpGs (DMR)
- How to define region?
  - Sliding window
  - Heuristic cutoffs/Smoothing
  - Functional units



Gene Set Enrichment

- Long list of DMP or DMR… What does it mean?
- Gene expression -> gen set enrichment analysis (e.g. GO)
- Not so straightforward for methylation data!
  - CpG link to genes unclear
  - Directionality?
  - Extreme bias: number of CpG per gene differs

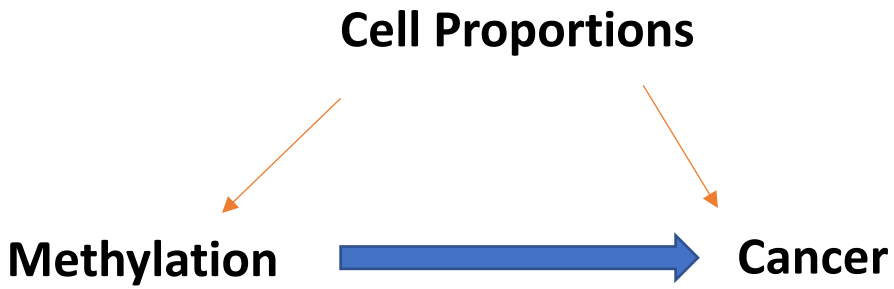
Gene-set analysis is severely biased when applied to genome-wide methylation data

Paul Geeleher[1,2], Lori Hartnett[3], Laurance J. Egan[3], Aaron Golden[4], Raja Affendi Raja Ali[3] and Cathal Seoighe[2,*]

- missMethyl, methylGSA, BioMethyl

Cell Type Deconvolution

- Estimates the relative proportion of pure cell types within a sample
- *Minfi*: RGChannelSet from a DNA methylation study of blood, and return the relative proportions of CD4+ and CD8+ T-cells, natural killer cells, monocytes, granulocytes, and b-cells in each sample.
- Most cohort studies currently analyse data from blood samples: can be used to correct for cell type heterogeneity



Datasets

- Small toy data
- IDAT files
- 10 samples in total: there are 4 different sorted T-cell types, collected from 3 different individuals :
  - Naïve
  - Treg
  - act_naive
  - act_Treg
- An additional sample is included from another study (GSE51180) to illustrate approaches for identifying and excluding poor quality samples.

Methylation Sequencing Workflow

Measuring DNA methylation by Bisulfite-sequencing

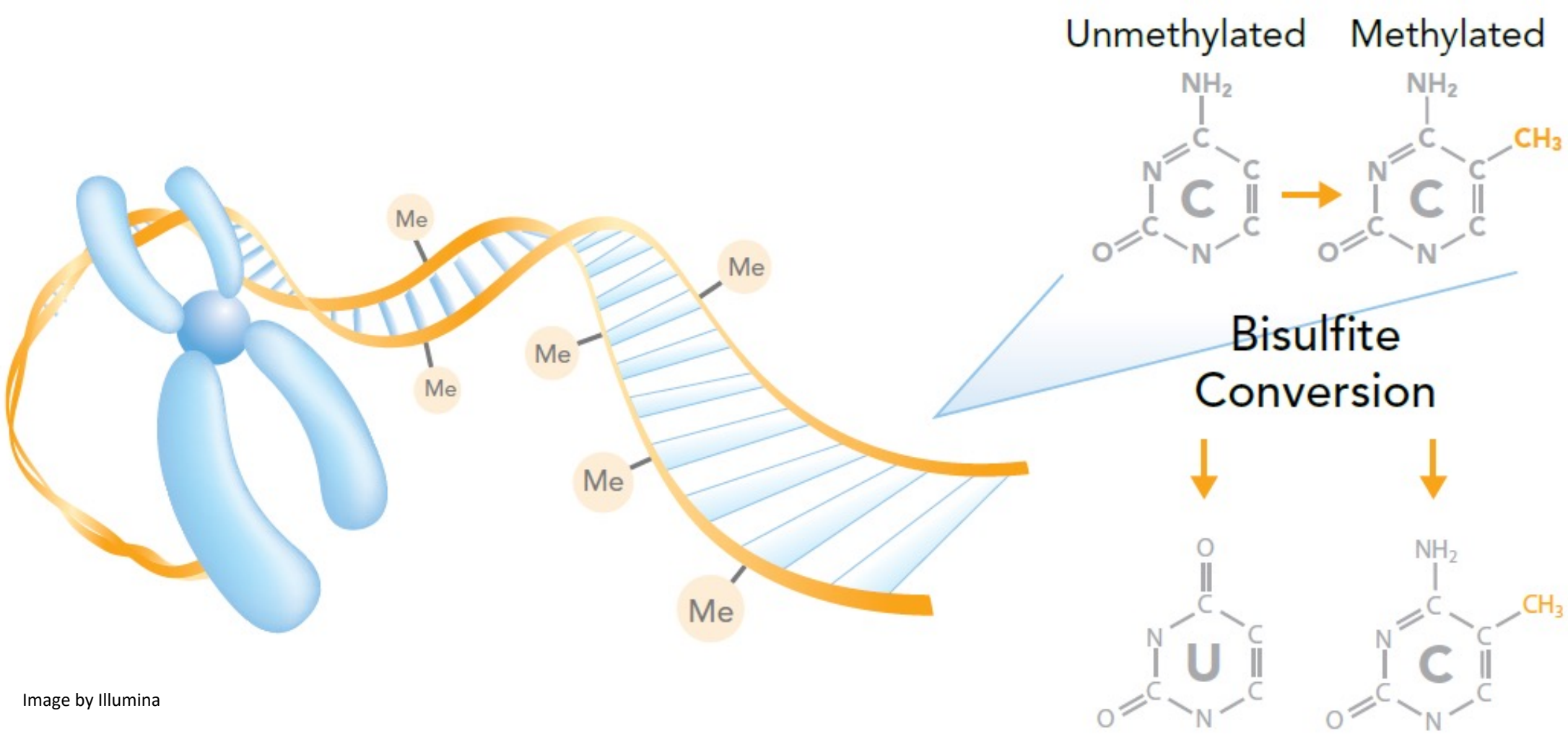
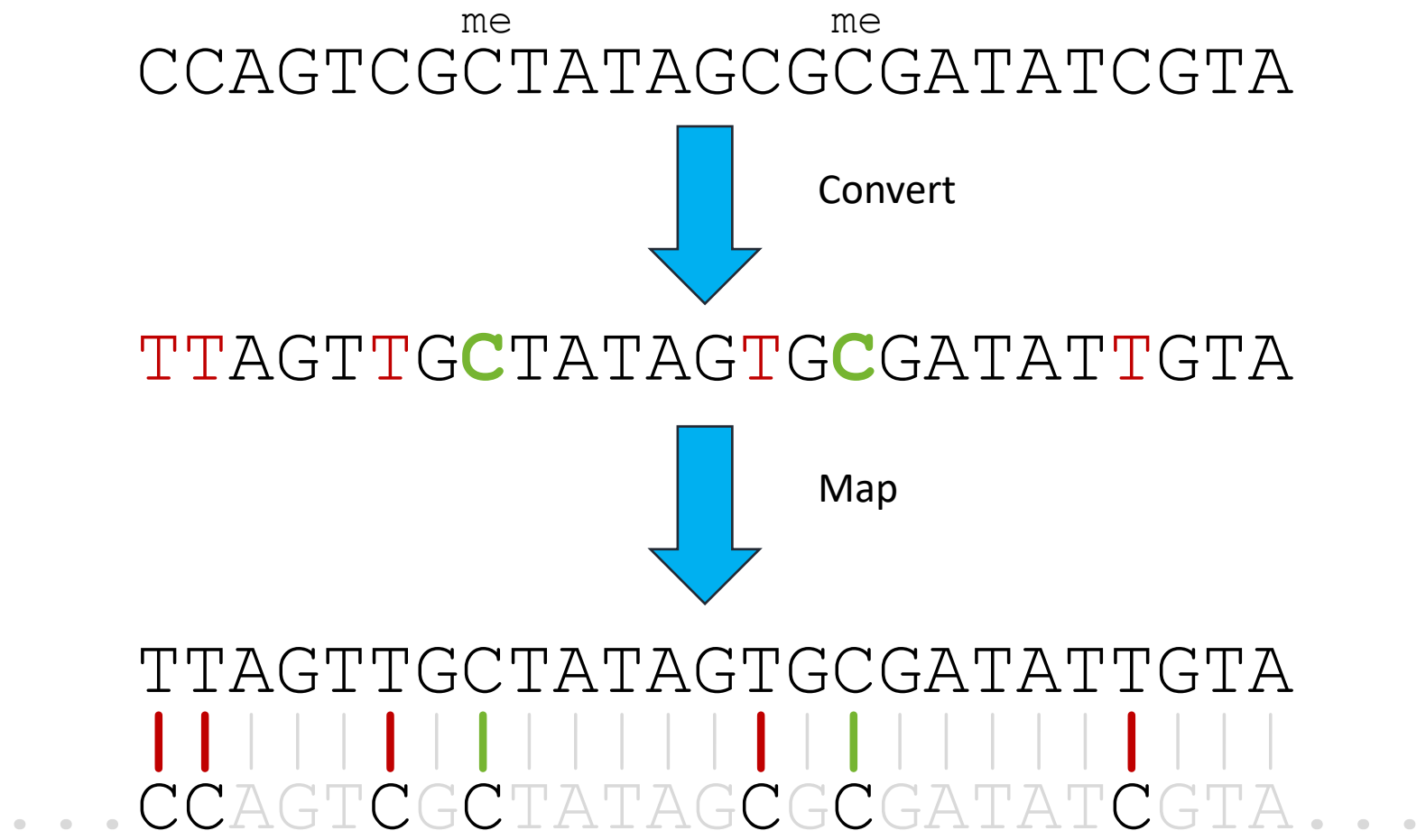
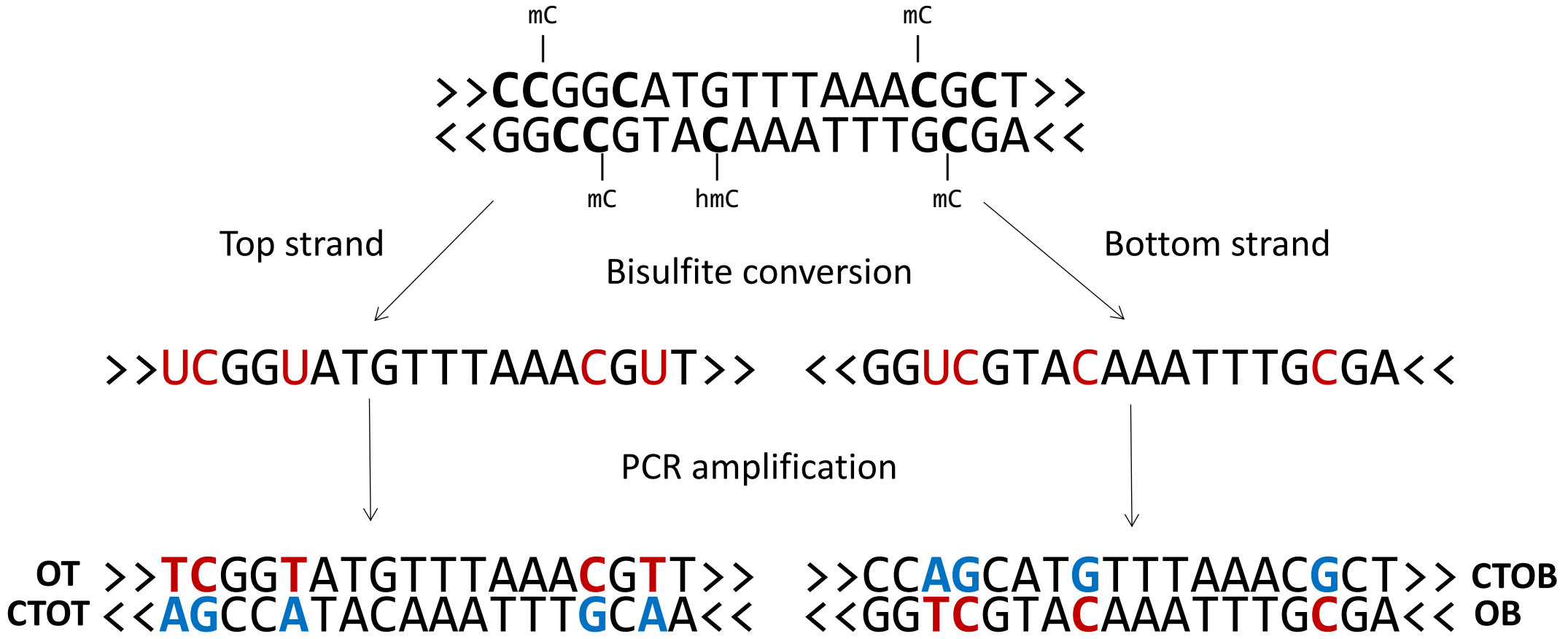


Image by Illumina

Easy readout… in theory



Bisulfite conversion of a genomic locus

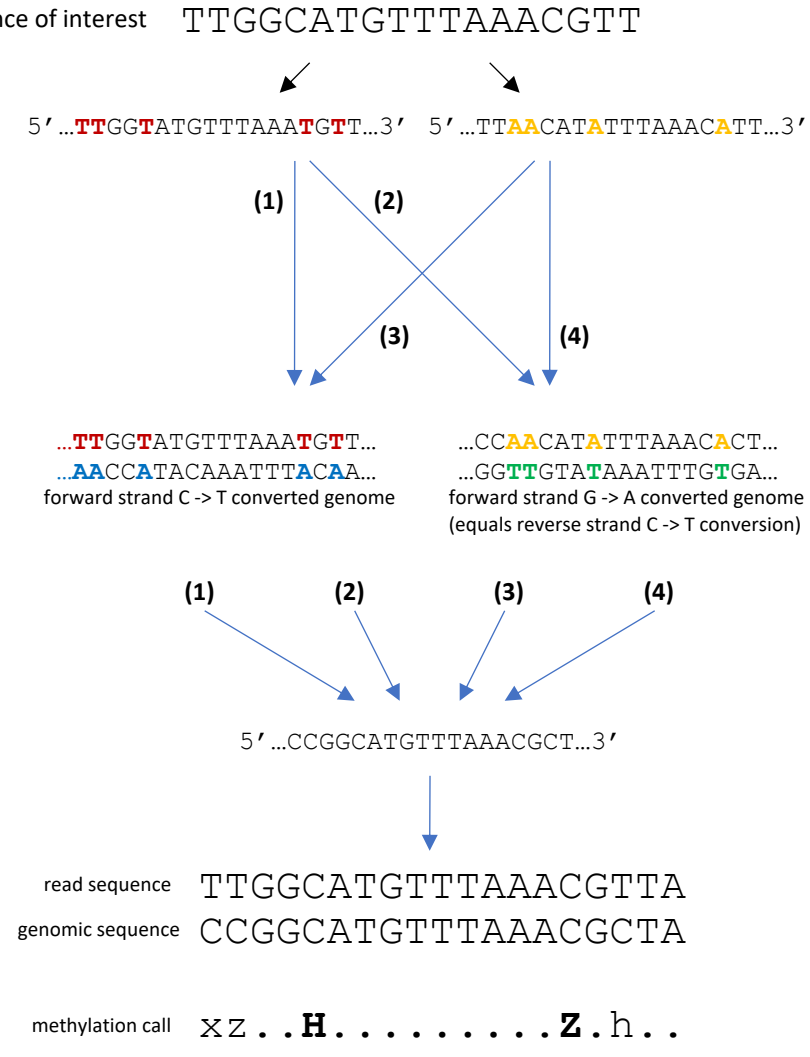


- 2 different PCR products and 4 possible different sequence strands from one genomic locus

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

3-letter alignment of Bisulfite-Seq reads

Bismark



bisulfite convert read (treat sequence as both forward and reverse strand)

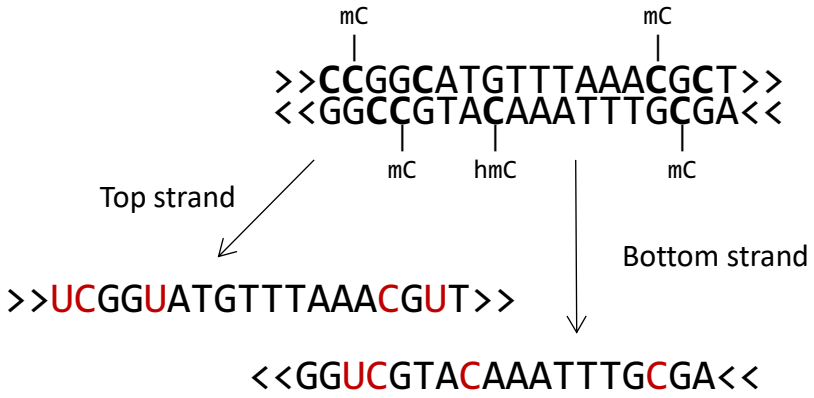
align to bisulfite converted genomes

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

Common sequencing protocols



1) Directional libraries

(vast majority of kits, also EpiGnome/Truseq)

OT >>TCGGTATGTTTAAACGTT>>
<<GGTCGTACAAATTTGCGA<< **OB**

2) PBAT libraries

CTOT <<AGCCATACAAATTTGCAA<<
>>CCAGCATGTTTAAACGCT>> **CTOB**

3) Non-directional libraries

(e.g. single-cell BS-Seq, Zymo Pico Methyl-Seq)

OT >>TCGGTATGTTTAAACGTT>>
CTOT <<AGCCATACAAATTTGCAA<<
>>CCAGCATGTTTAAACGCT>> **CTOB**
<<GGTCGTACAAATTTGCGA<< **OB**

Quality Control is important!

- Relies on accurate C > T detection
- Pre-alignment:
  - Base quality/composition
  - Duplication
  - Trim adapters

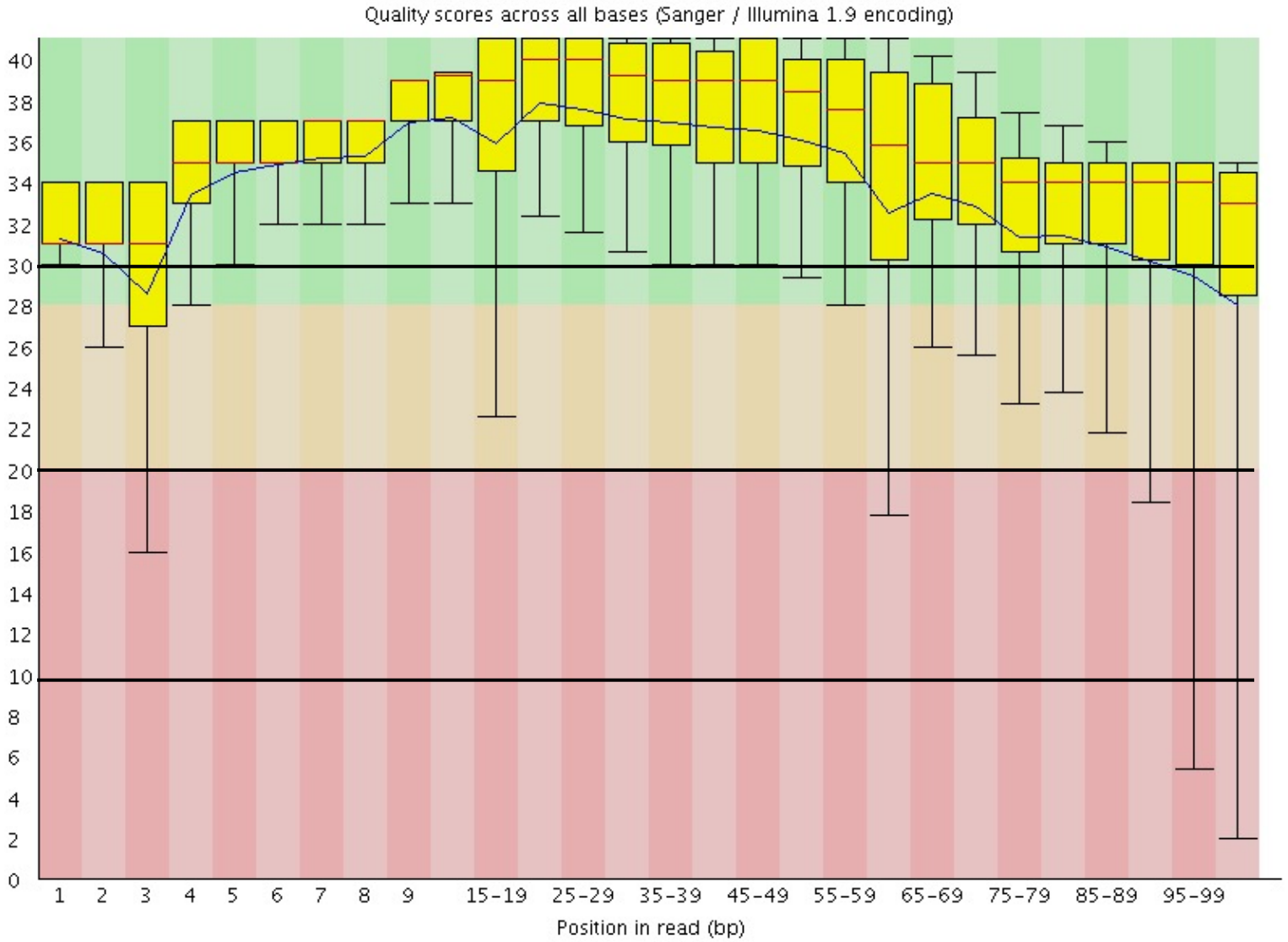
Average Base Quality

Error rate

0.1%

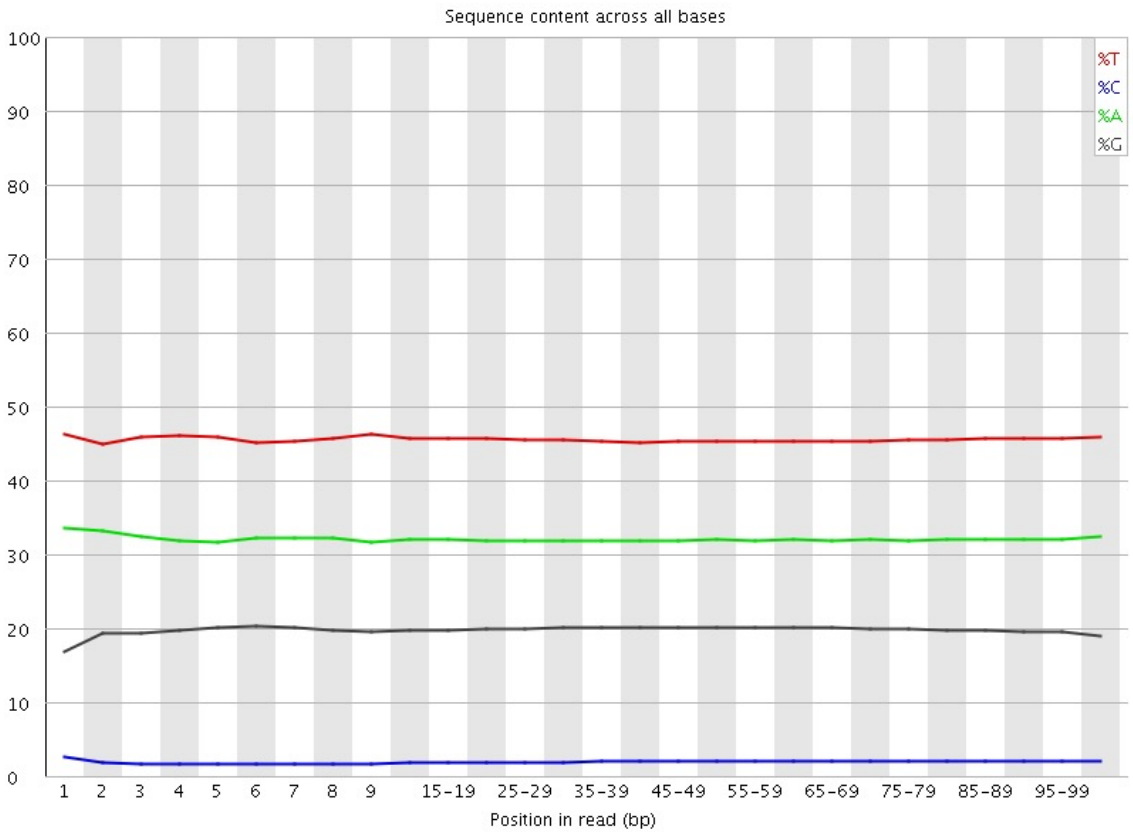
1%

10%

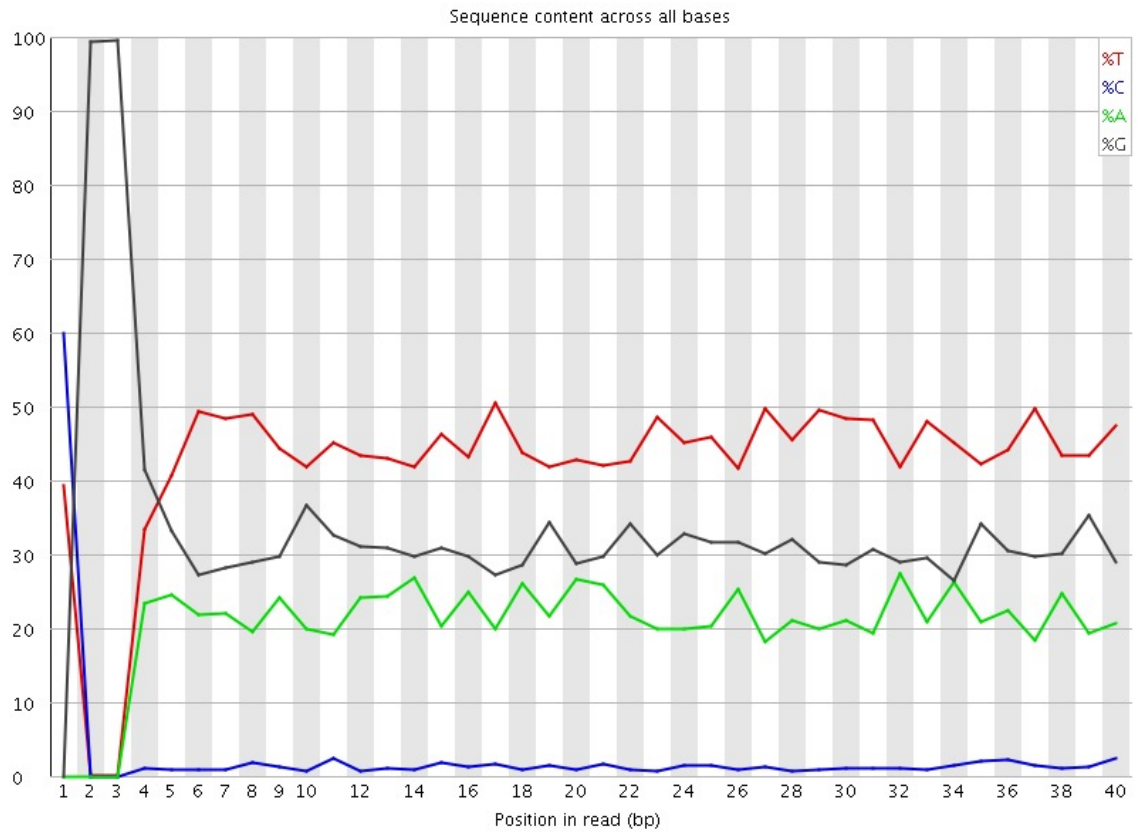


Base Composition

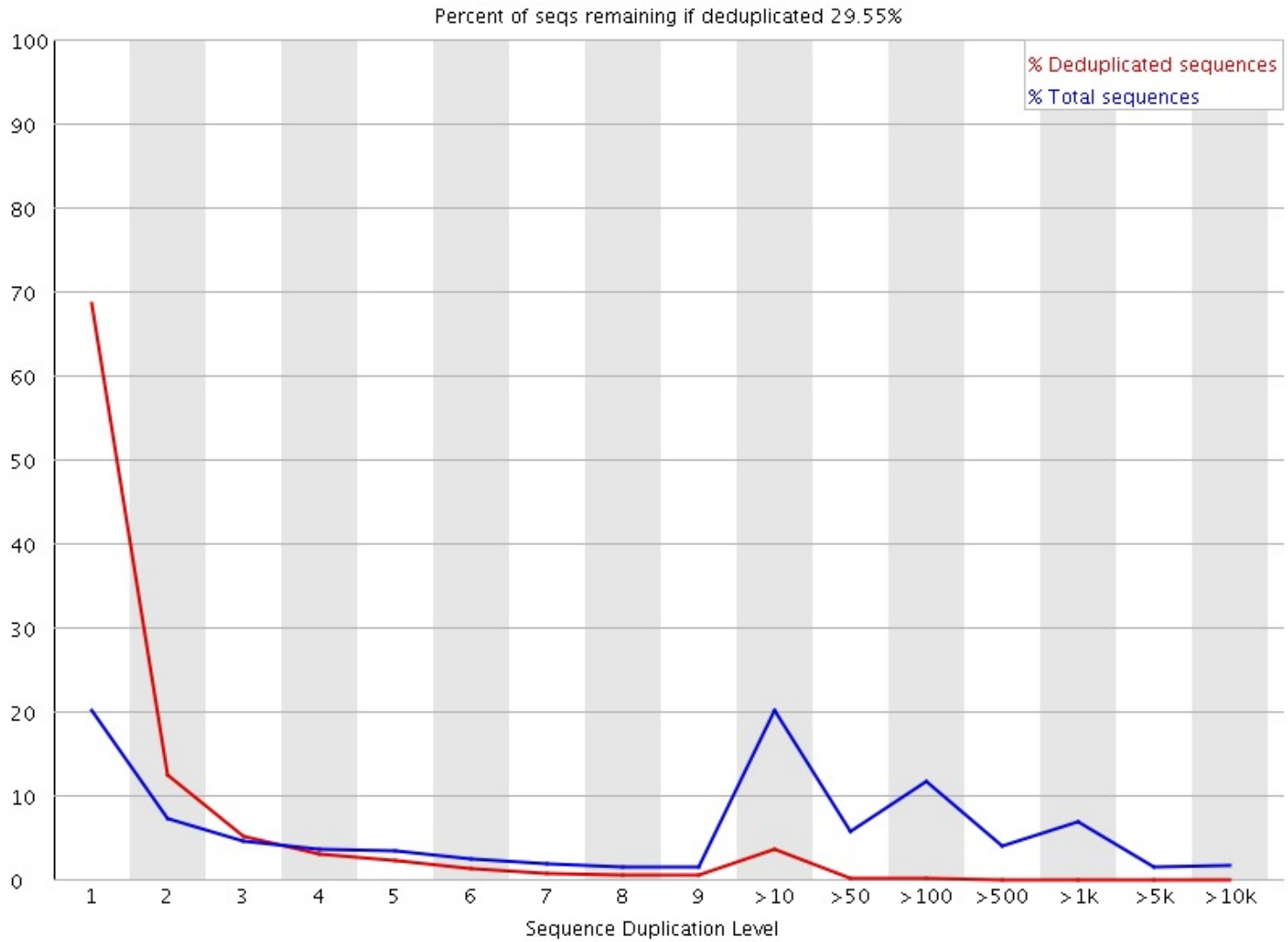
WGSBS



RRBS



Duplication rate

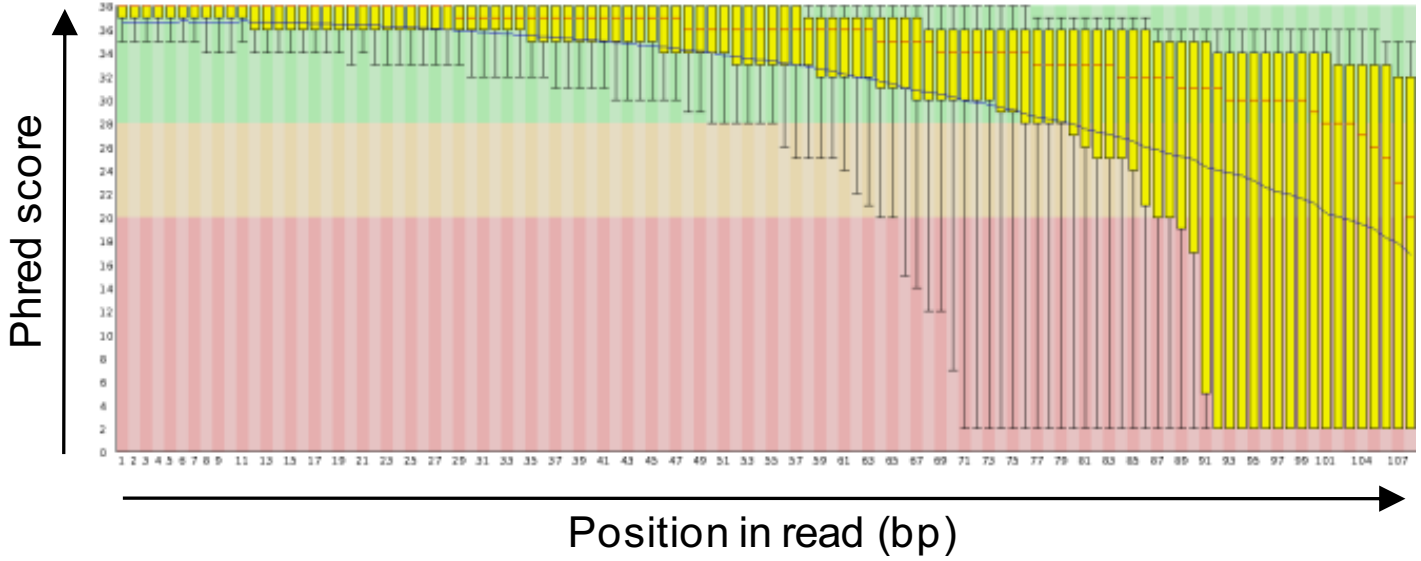


Overrepresented sequences

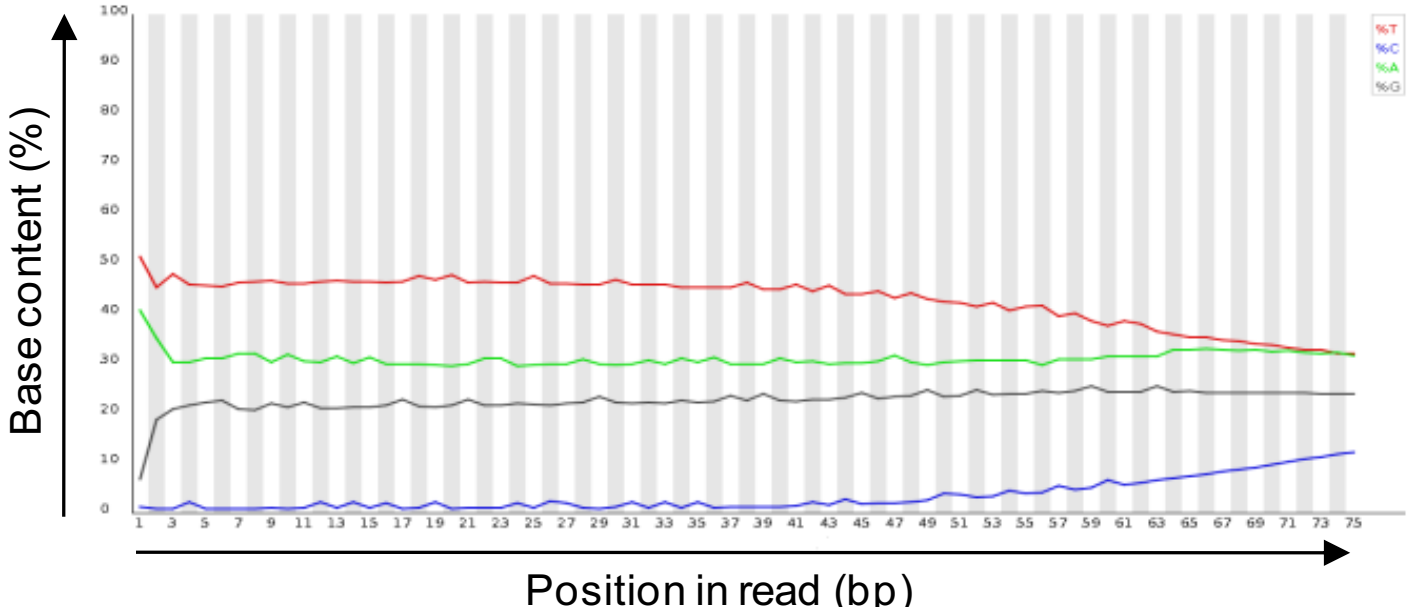
Overrepresented sequences

Sequence	Count	Percentage	Possible Source
GAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTCGTAT	6254891	23.52739098691508	Illumina Paired End PCR Primer 2 (100% over 40bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCT	1956005	7.357393503317777	Illumina Paired End PCR Primer 2 (100% over 40bp)
GAAGAGCGGTTCAGCAGGAATGCCGAGATCGGAAGAGCGG	774763	2.9142237687587667	Illumina Paired End PCR Primer 2 (96% over 31bp)
GAAGAGCGGTTCAGCAGGAATGCCGAGGATCGGAAGAGCG	140148	0.5271581538405985	Illumina Paired End Adapter 2 (100% over 27bp)
AAGAGCGGTTCAGCAGGAATGCCGAGATCGGAAGAGCGGT	105720	0.3976593317352233	Illumina Paired End PCR Primer 2 (96% over 30bp)
NAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTCGTAT	98639	0.37102458213233724	Illumina Paired End PCR Primer 2 (97% over 40bp)
AAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTCGTATG	82413	0.30999147281777295	Illumina Paired End PCR Primer 2 (100% over 40bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGATCGGAAG	53872	0.20263624214188372	Illumina Paired End PCR Primer 2 (97% over 36bp)
NNAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTCGTAT	36541	0.137446742725471	Illumina Paired End PCR Primer 2 (100% over 38bp)
ATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTC	35781	0.13458804908076072	Illumina Paired End PCR Primer 2 (100% over 40bp)
CGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTCGT	33905	0.1275315895051338	Illumina Paired End PCR Primer 2 (100% over 40bp)
NATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCT	30564	0.1149646217854272	Illumina Paired End PCR Primer 2 (97% over 40bp)
GAAGAGCGGTTCAGCAGGAATGCCGAGACGGATCTCGTAT	28274	0.10635092646123442	Illumina Paired End PCR Primer 2 (97% over 40bp)
CAAACAACTTCTAAAACAAACAAAAACACAAAACCACTAA	27952	0.10513974310123876	No Hit

Common problems in BS-Seq



- Declining base quality
- Common issue in NGS



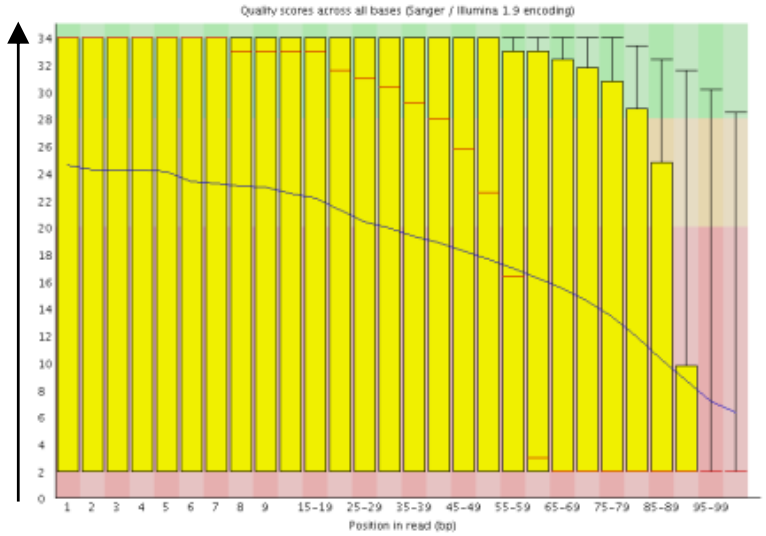
- Adapter contamination
- Present, but not as easily observed in ‘normal’ libraries

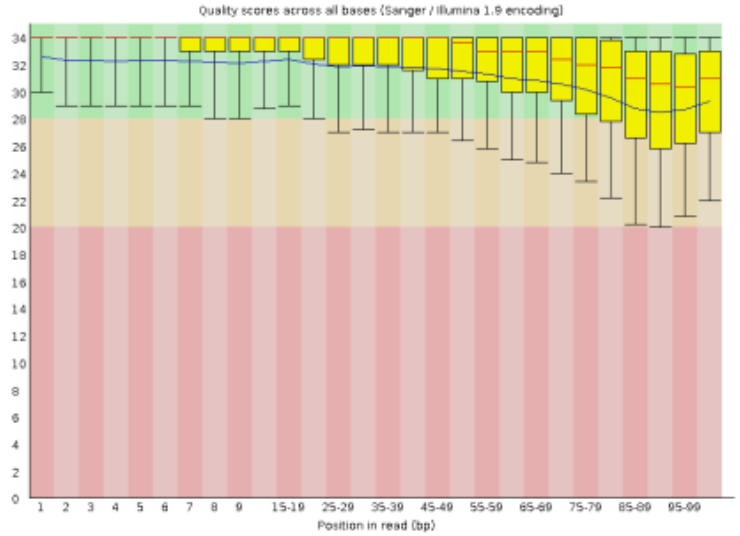
Removing poor quality basecalls

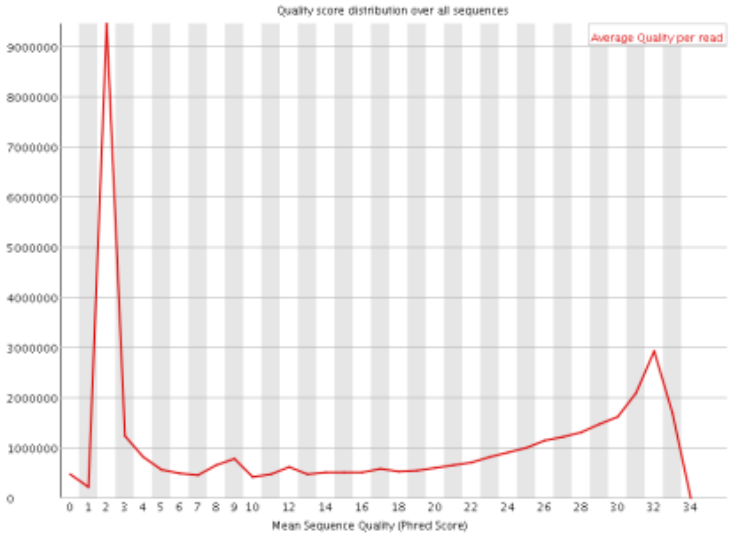
before trimming

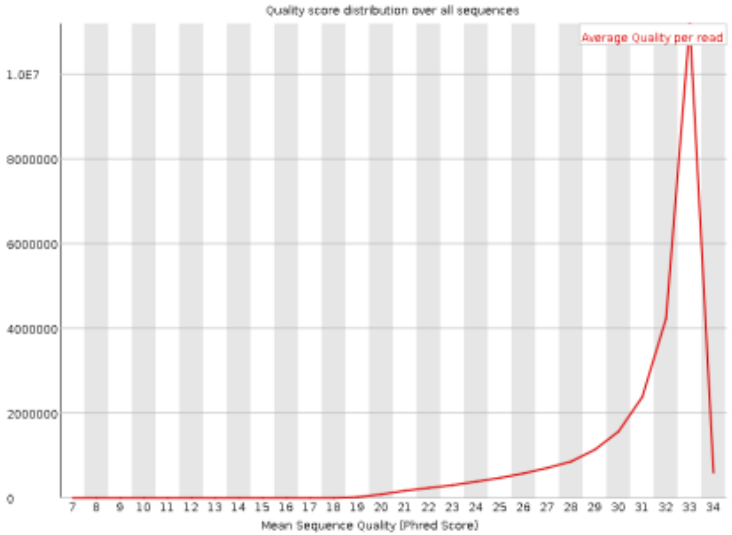
after trimming

Phred score



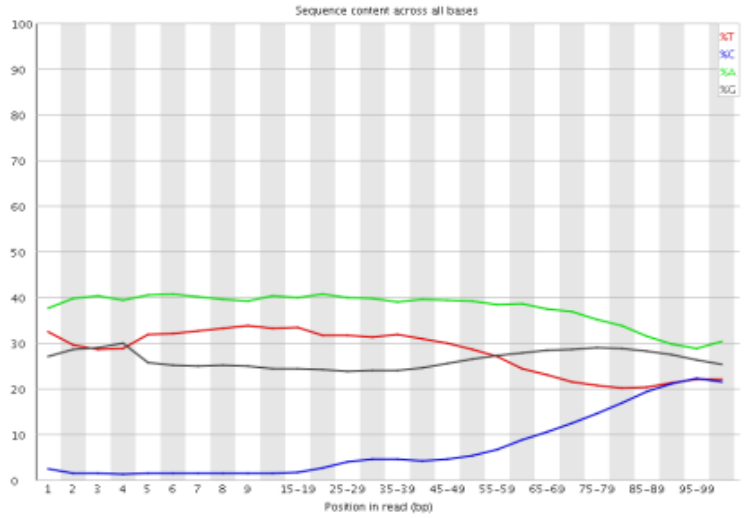




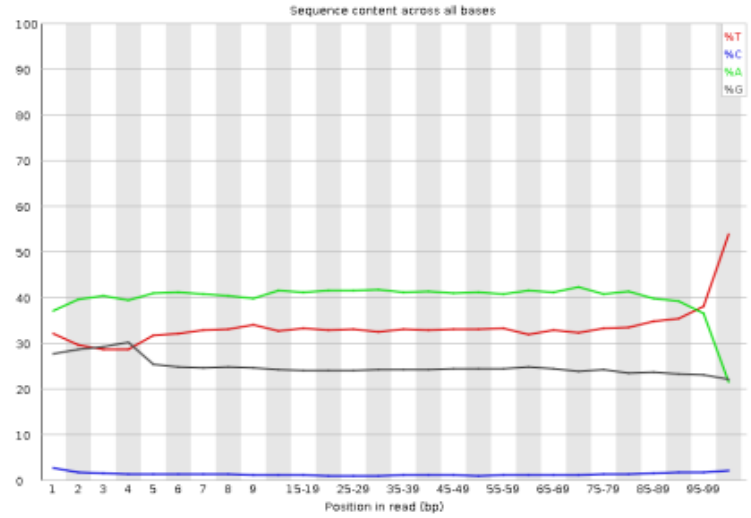


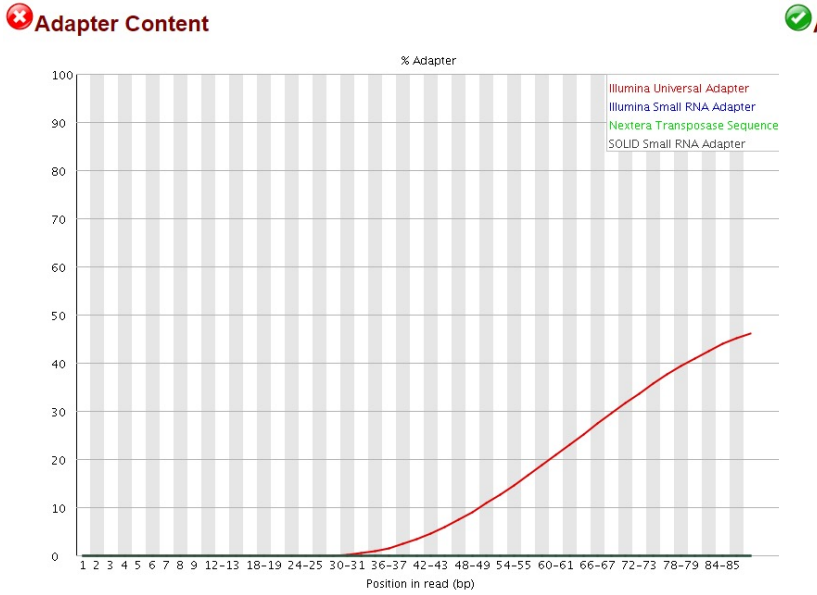
Removing adapter contamination

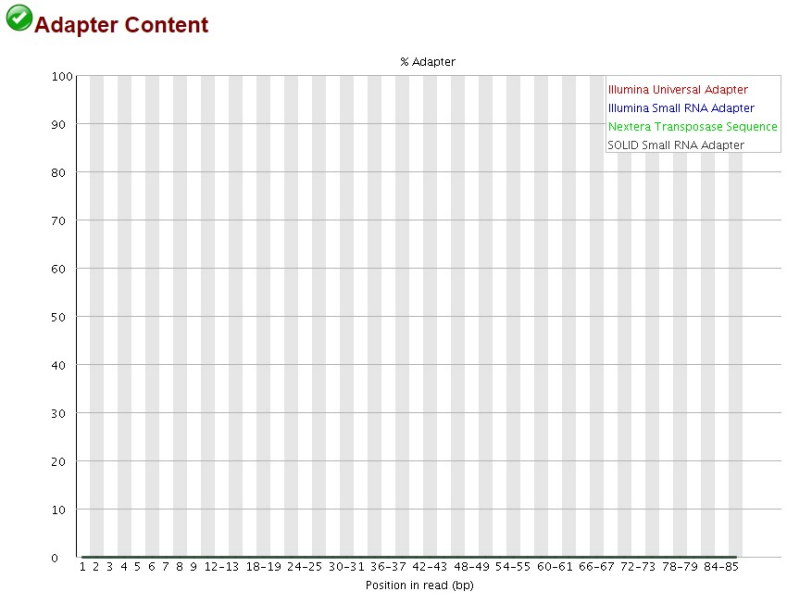
before trimming



after trimming







Summary Adapter/Quality Trimming

Important to trim because failure to do so might result in:

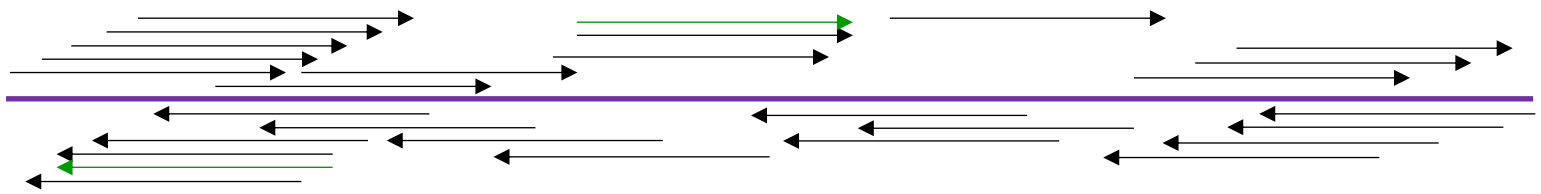
- Low mapping efficiency
- Mis-alignments
- Errors in methylation calls since adapters are methylated
- Basecall errors tend toward 50% (C:mC)

Quality Control is important!

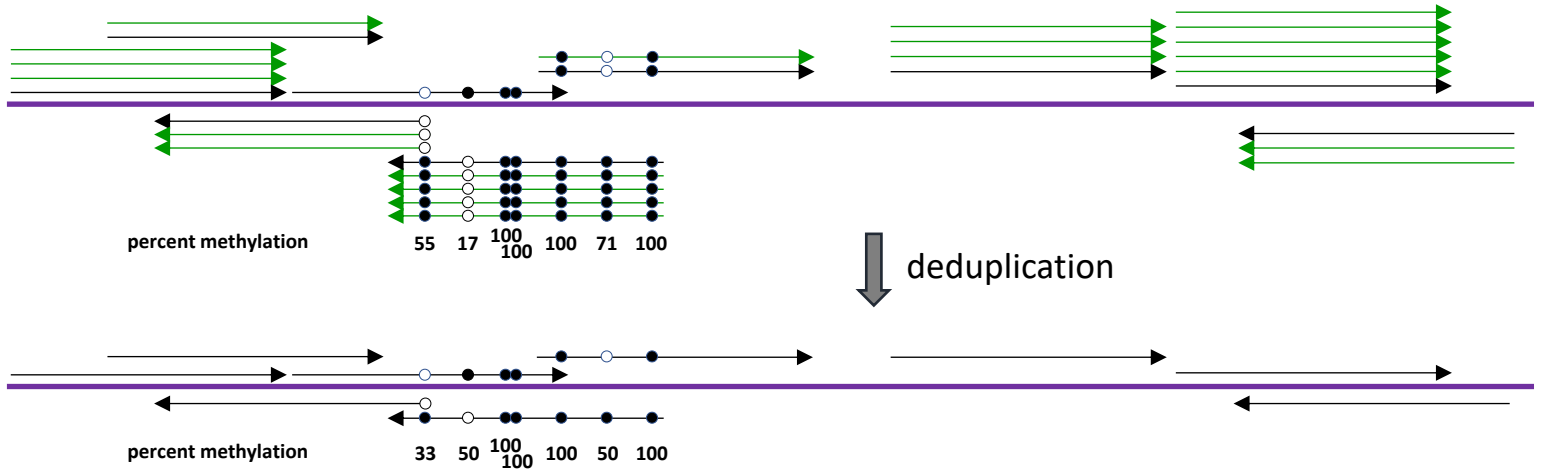
- Relies on accurate C > T detection
- Pre-alignment:
  - Base quality
  - Trim adapters
- Post-alignment:
  - Incomplete conversion? Check non-CpG methylation, should be near 100%
  - Degradation? Check alignment rates and read length
  - PCR bias? Perhaps deduplicate

Sequence duplication

Complex/diverse library:



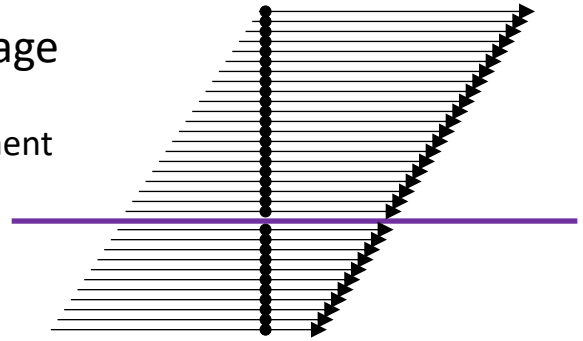
Duplicated library:



Deduplication - considerations

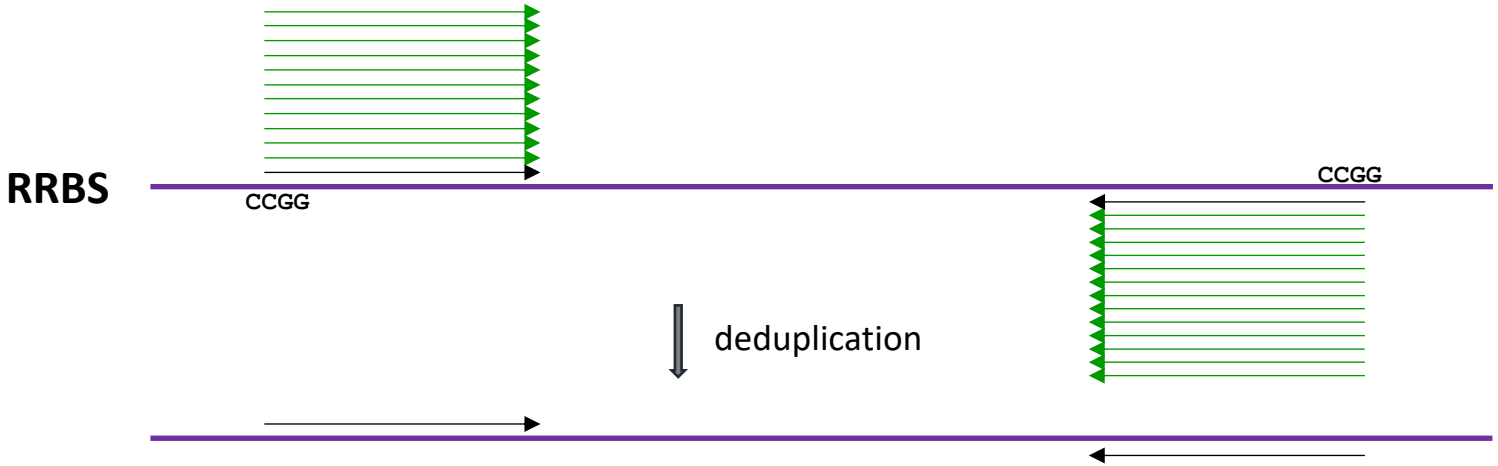
Advisable for large genomes and moderate coverage

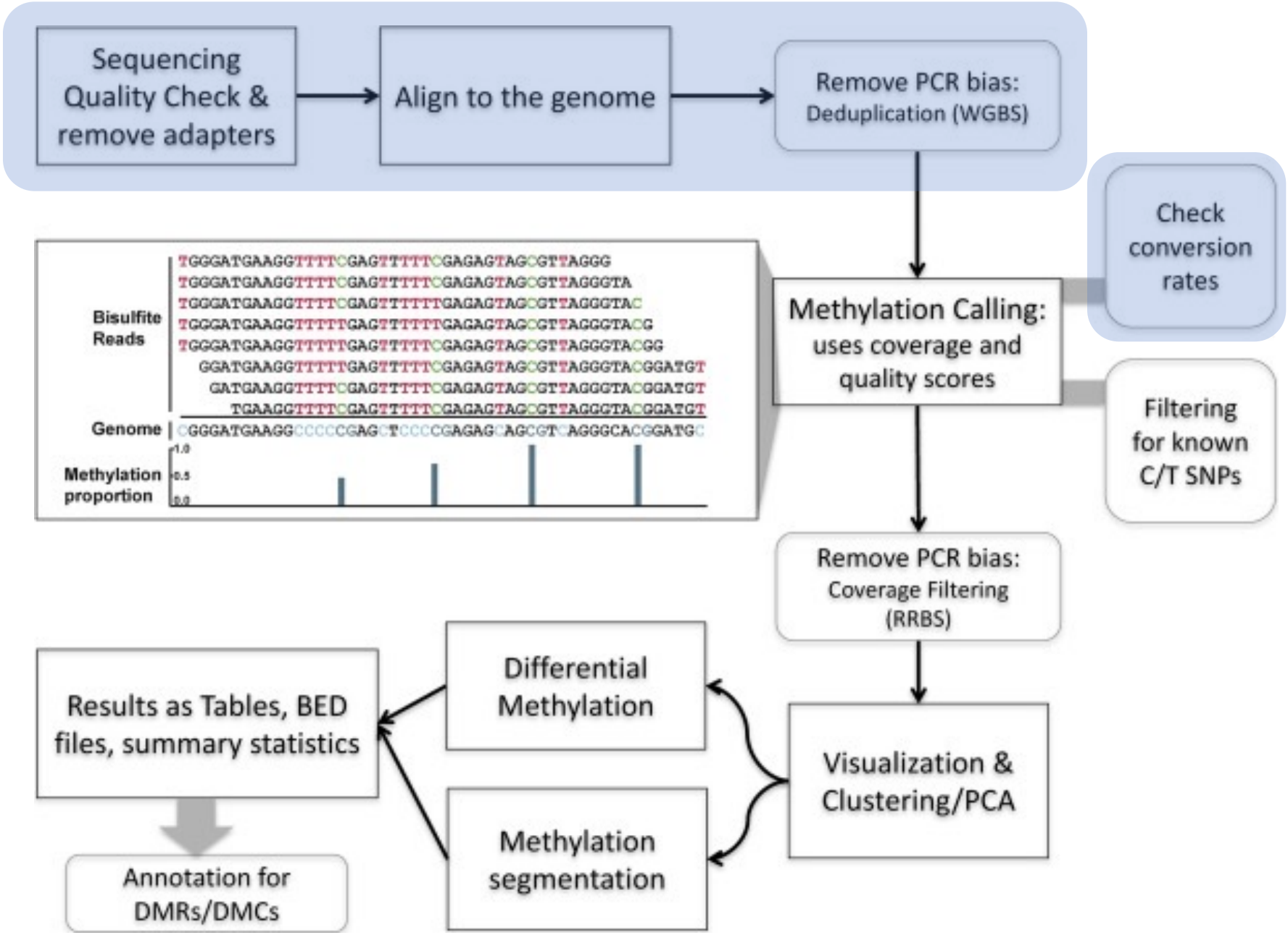
- unlikely to sequence several genuine copies of the same fragment amongst >5bn possible fragments with different start sites
- maximum coverage with duplication may still be (read length)-fold (even more with paired-end reads)



NOT advisable for RRBS or other target enrichment methods

where higher coverage is either desired or expected

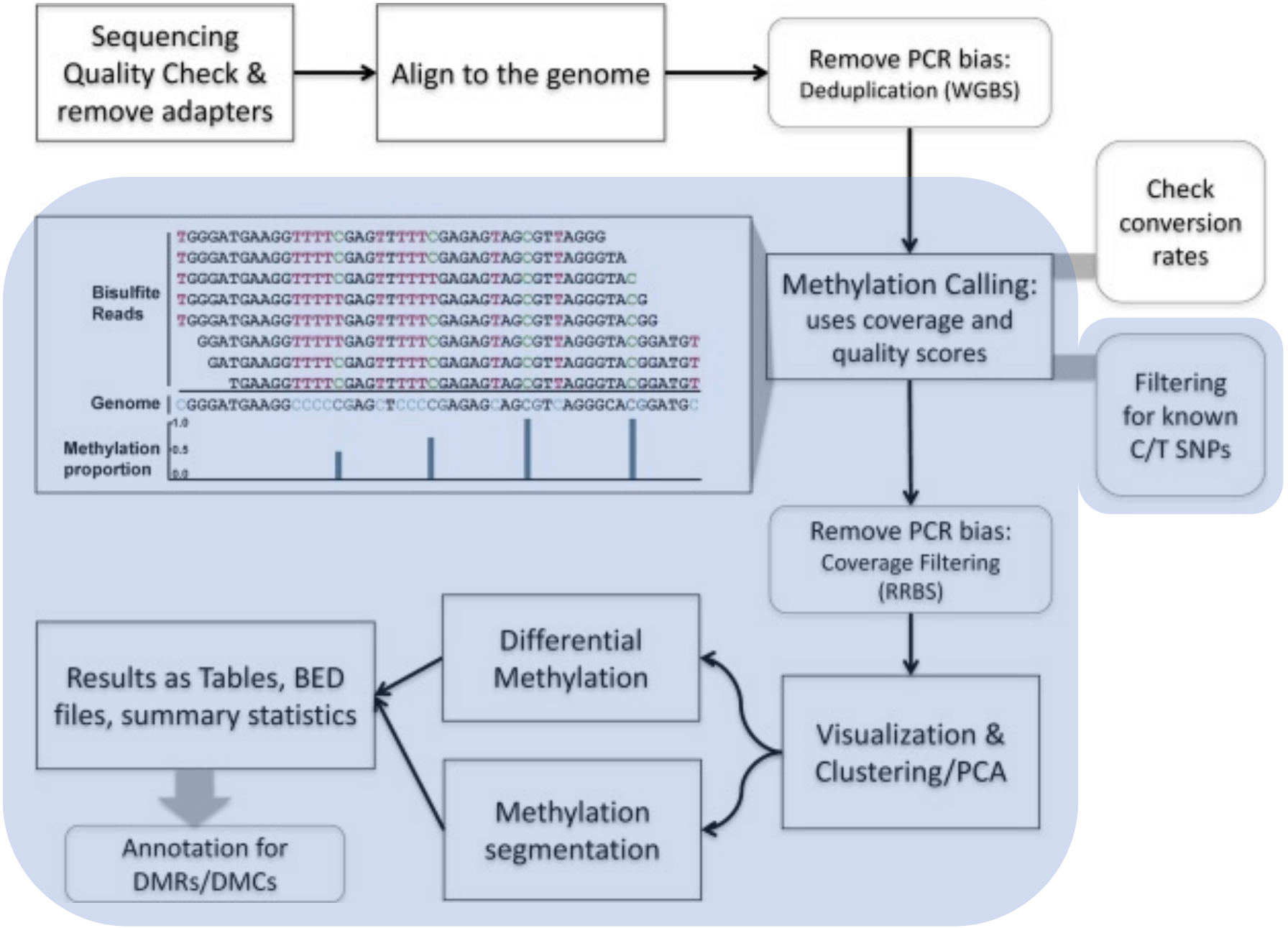




nf-core

A community effort to collect a curated set of analysis pipelines built using Nextflow

- Nf-core pipeline: methylseq (see Thursday)
- Preprocessing + QC
  - 2 aligners: Bismark or bwa-meth/MethylDackel
  - QC: qualimap, preseq and multiqc
- Output of this can be used as starting point downstream analysis



MethylKit

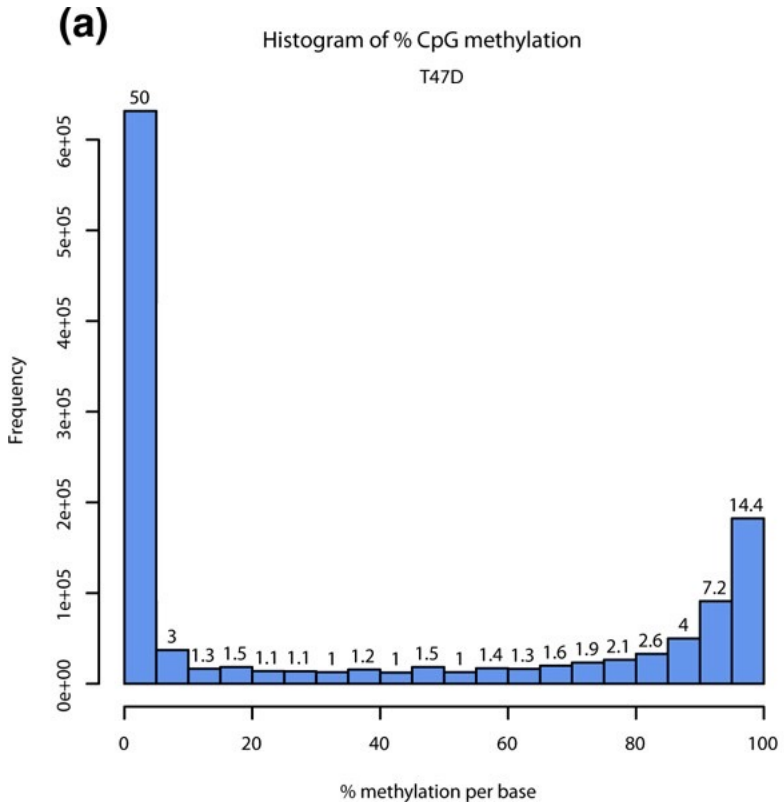
R package for downstream analysis of bisulfite data

Can read in Bismark coverage files as input…

Descriptive statistics

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

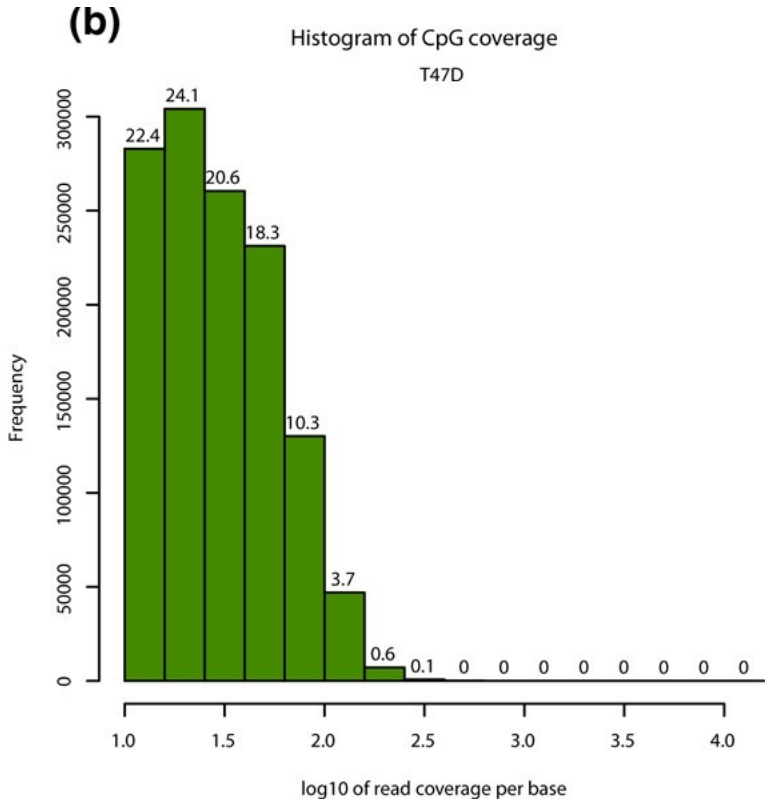
- After reading in the data…
- Distribution percentage methylation



- Equivalent to Beta value in array data
- Expect a peak at low and high ends of the distribution

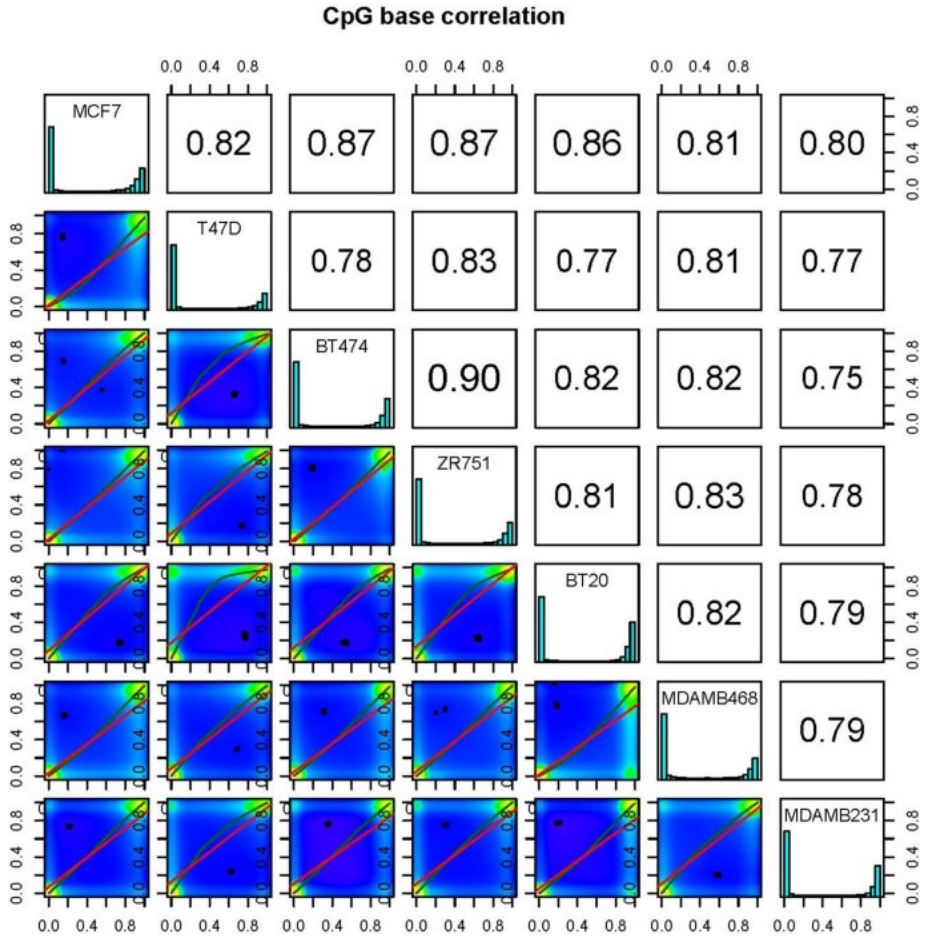
Descriptive statistics

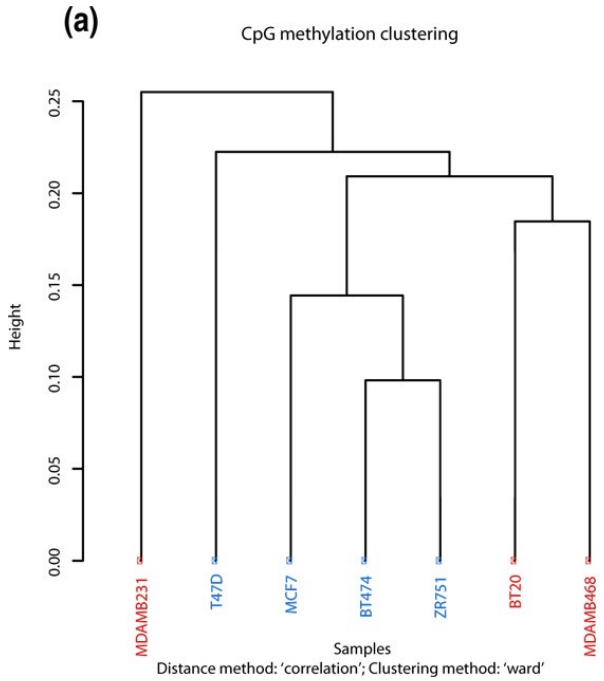
- After reading in the data…
- Coverage distribution

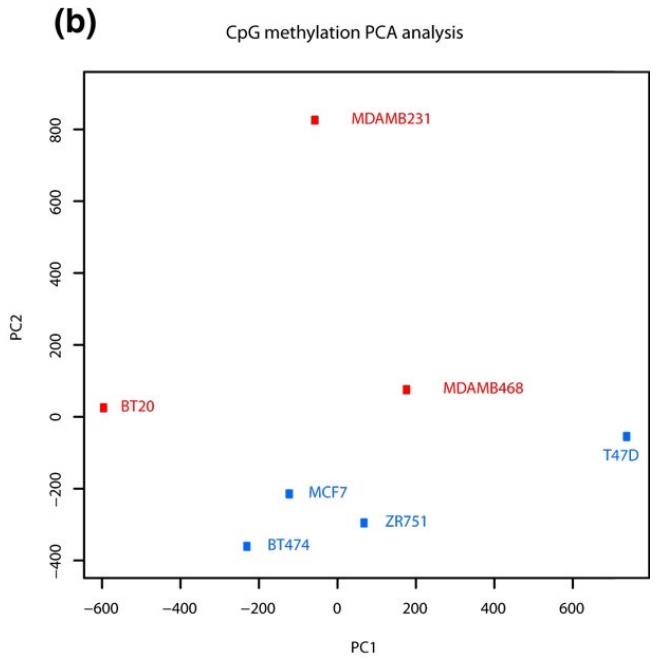


- Experiments that suffer from PCR duplication will have a secondary peak towards the right hand site of the histogram
- Can be used to determine filter cutoff; both very low and very high coverage

Sample Structure



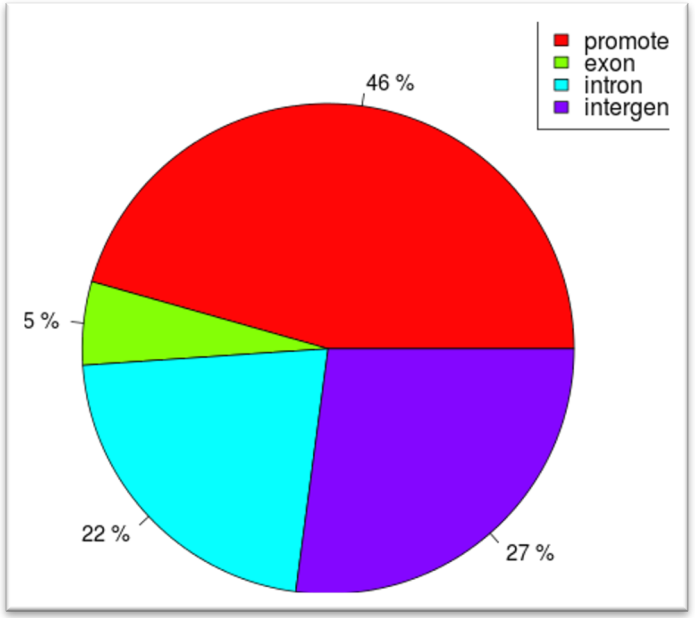




Differential analysis

- Again, many choices; usually calculated by comparing the proportion of methylated Cs in a test sample relative to a control
- No replicates: Fisher's Exact Test
- Replicates:
  - Linear regression (limma as for arrays)
  - Logistic regression (works with [0-1] data)
  - Beta-binomial (deals better with count data)
- Regression models can add covariates (batch, age, sex, …)
- Can also aggregate in regions (see tutorial)

Annotation



- How to interpret the DMR/DMPs?
- Integrate with genome annotation datasets
  - Where in relation to gene/regulatory region
  - Promoter or intron or exon?
- Genomation R package: toolkit for annotation and in bulk visualization of genomic intervals
- Tutorial: basic annotation transcripts and CpG islands
- Requires some knowledge of R (especially the GenomicRanges package)

Remarks…

- Normalization somewhat less important for bisulfite sequencing (Fisher’s Exact is sensitive to sequencing depth though)
- Gene enrichment is as difficult as for arrays; not many implemented methods (rGREAT, Goseq).

Dataset

- Small dataset of mouse mammary gland cells
- 4 samples: 2 luminal, 2 basal
- Bismark coverage files

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

(VIII) Notes about different library types and commercial kits

Here is a table summarising general recommendations for different library types and/or different commercially available kits. Some more specific notes can be found below.

Technique	5' Trimming	3' Trimming	Mapping	Deduplication	Extraction
BS-Seq	⬜	⬜	⬜	✅	`--ignore_r2 2`
RRBS	`--rrbs` (R2 only)	`--rrbs` (R1 only)	⬜	❌	(`--ignore_r2 2`)
RRBS (NuGEN Ovation)	special processing	special processing	⬜	❌	`--ignore_r2 2`
PBAT	6N / 9N	(6N / 9N)	`--pbat`	✅	⬜
single-cell (scBS-Seq)	6N	(6N)	`--non_directional`; single-end mode	✅	⬜
TruSeq (EpiGnome)	8 bp	(8 bp)	⬜	✅	⬜
Accel-NGS (Swift)	10 bp	(10 bp)	⬜	✅	⬜
Zymo Pico-Methyl	10 bp	(10 bp)	`--non_directional`	✅	⬜

- ⬜ - Default settings (nothing in particular is required, just use Trim Galore or Bismark default parameters)
- ✅ - Yes, please!
- ❌ - No, absolutely not!

5' **Trimming** can be accomplished with Trim Galore using:

`--clip_r1 <NUMBER>` (Read 1) or

`--clip_r2 <NUMBER>` (Read 2)

3' **Trimming** can be accomplished with Trim Galore using:

`--three_prime_clip_r1 <NUMBER>` (Read 1) or

`--three_prime_clip_r2 <NUMBER>` (Read 2).

SPECIFIC LIBRARY/KIT NOTES

RRBS