BioTrain.TV

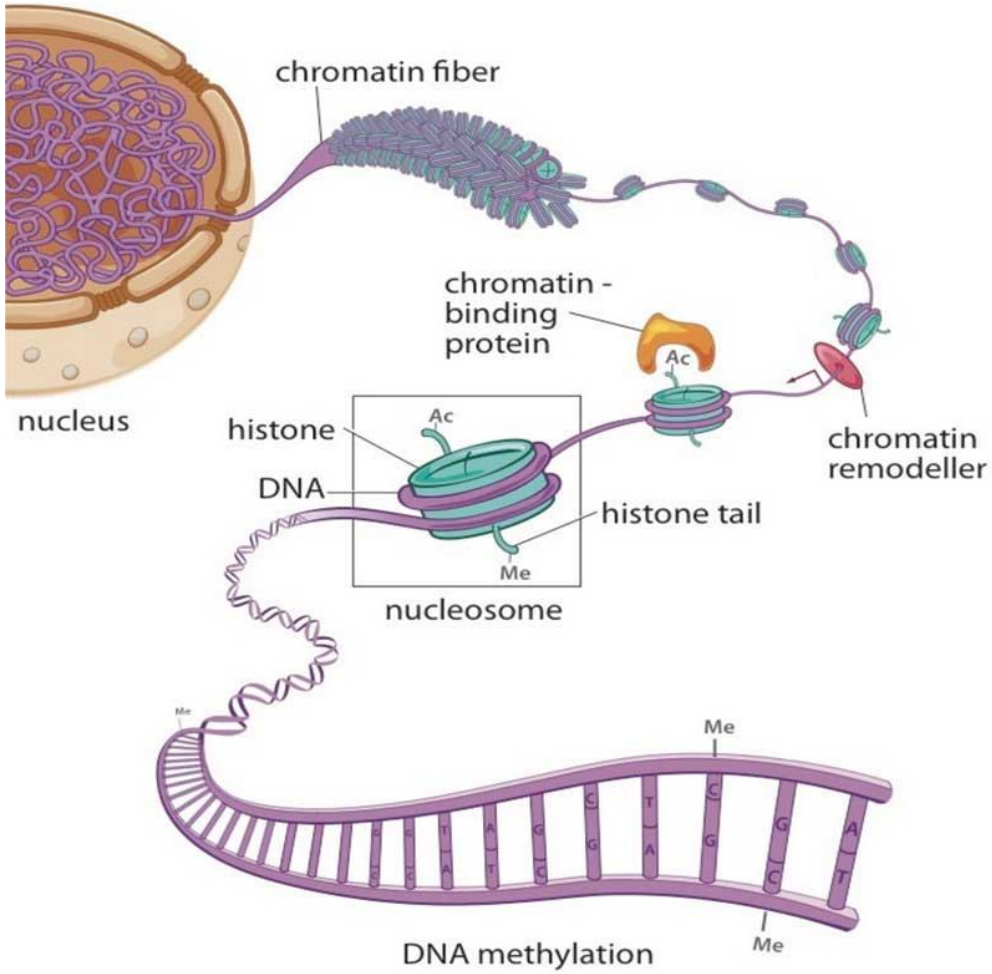
#Babraham Bioinformatics

Analysing DNA Methylation Data

Simon Andrews, Felix Krueger

v2026-06

Epigenetics



From The Cell Biology of Stem Cells (2010)

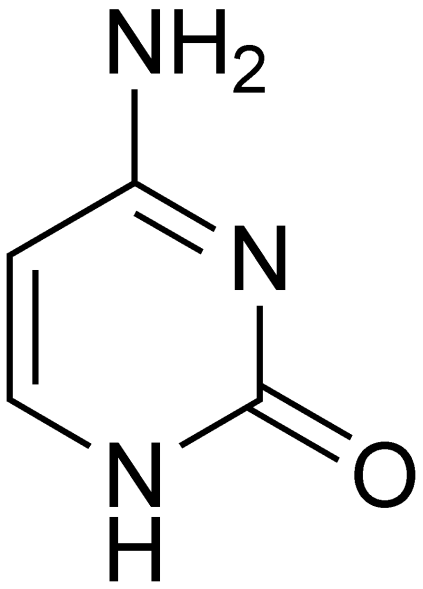
Studies changes in gene expression which are not encoded by the underlying DNA sequence

Chromatin

- histone modification
- non-coding RNAs
- higher order structure (accessibility/compaction)

DNA cytosine methylation

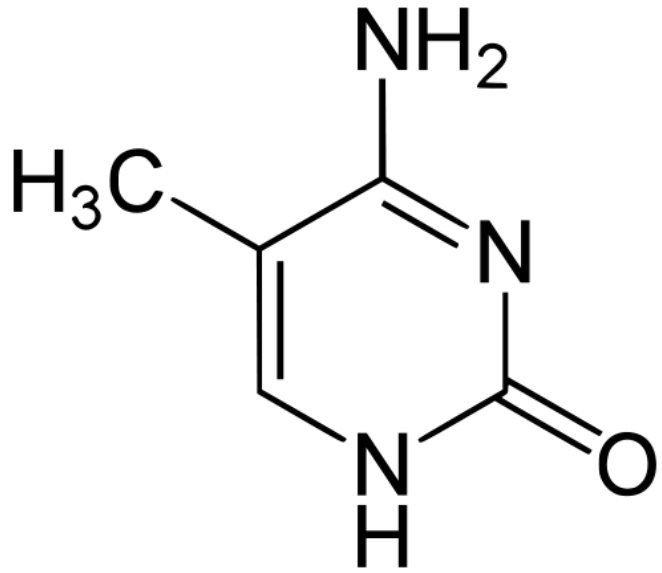
DNA Methylation



Cytosine

DNA methyl-transferases

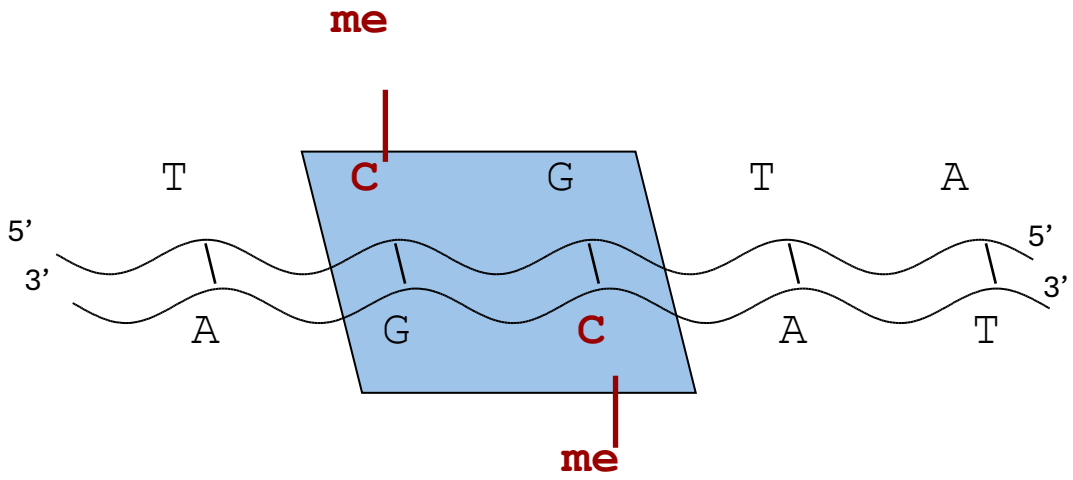
Active Demethylation
Passive demethylation



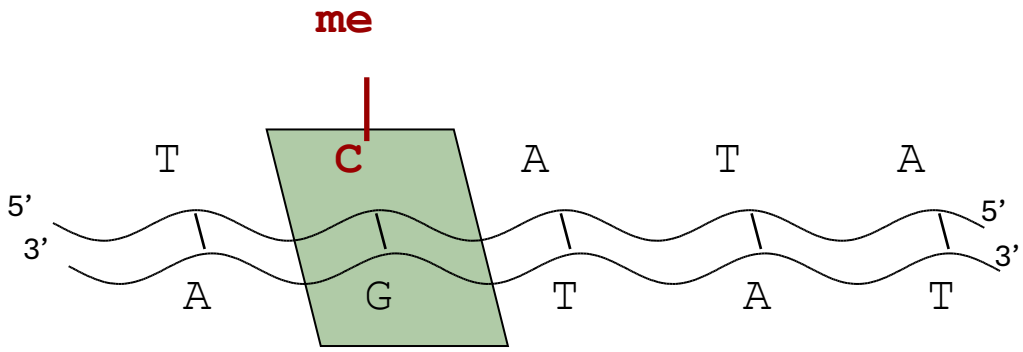
5-methyl Cytosine

Context of DNA methylation

CG context



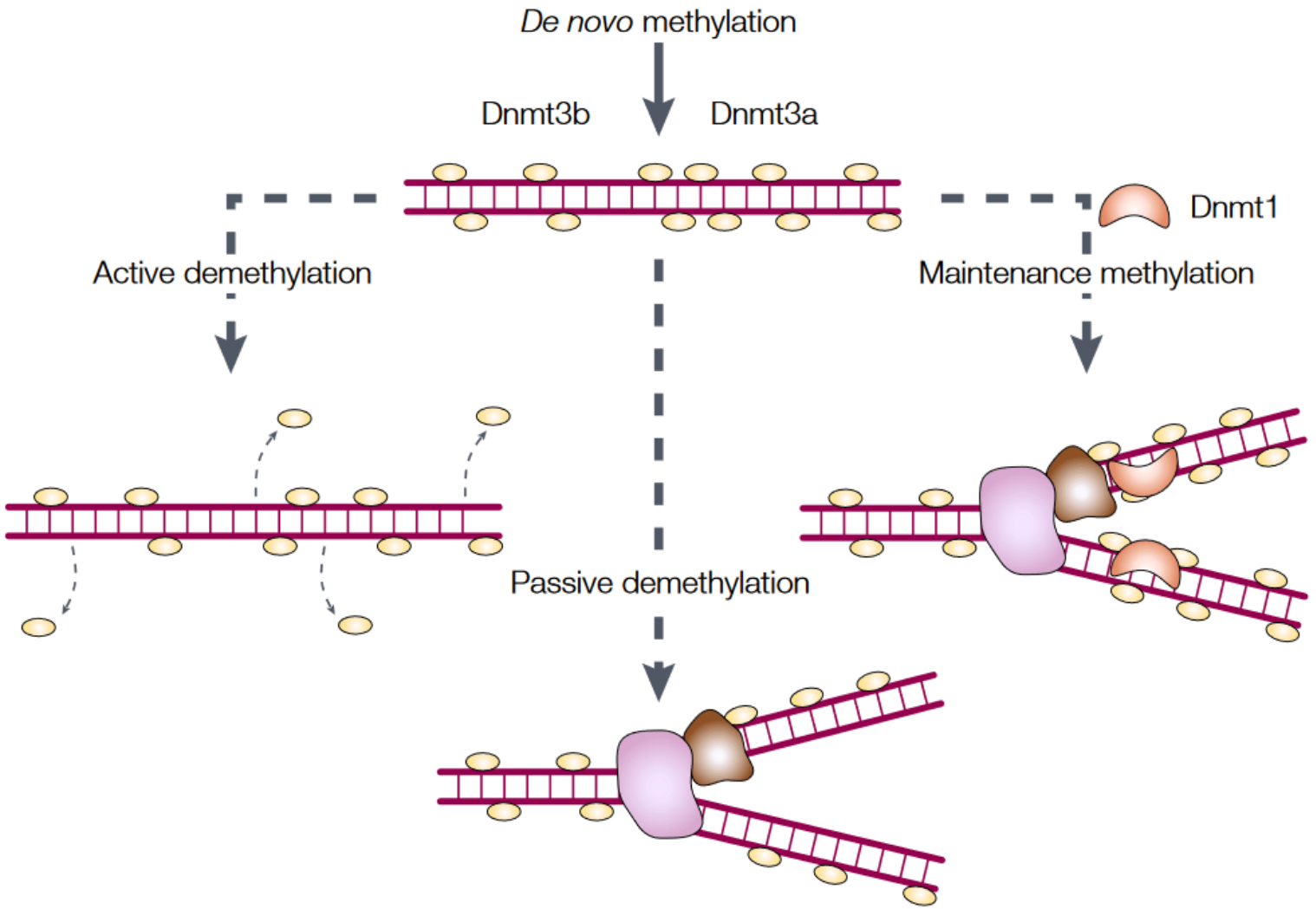
non-CG context



	Mammals	Plants
CG	present	present

Methylation Maintenance

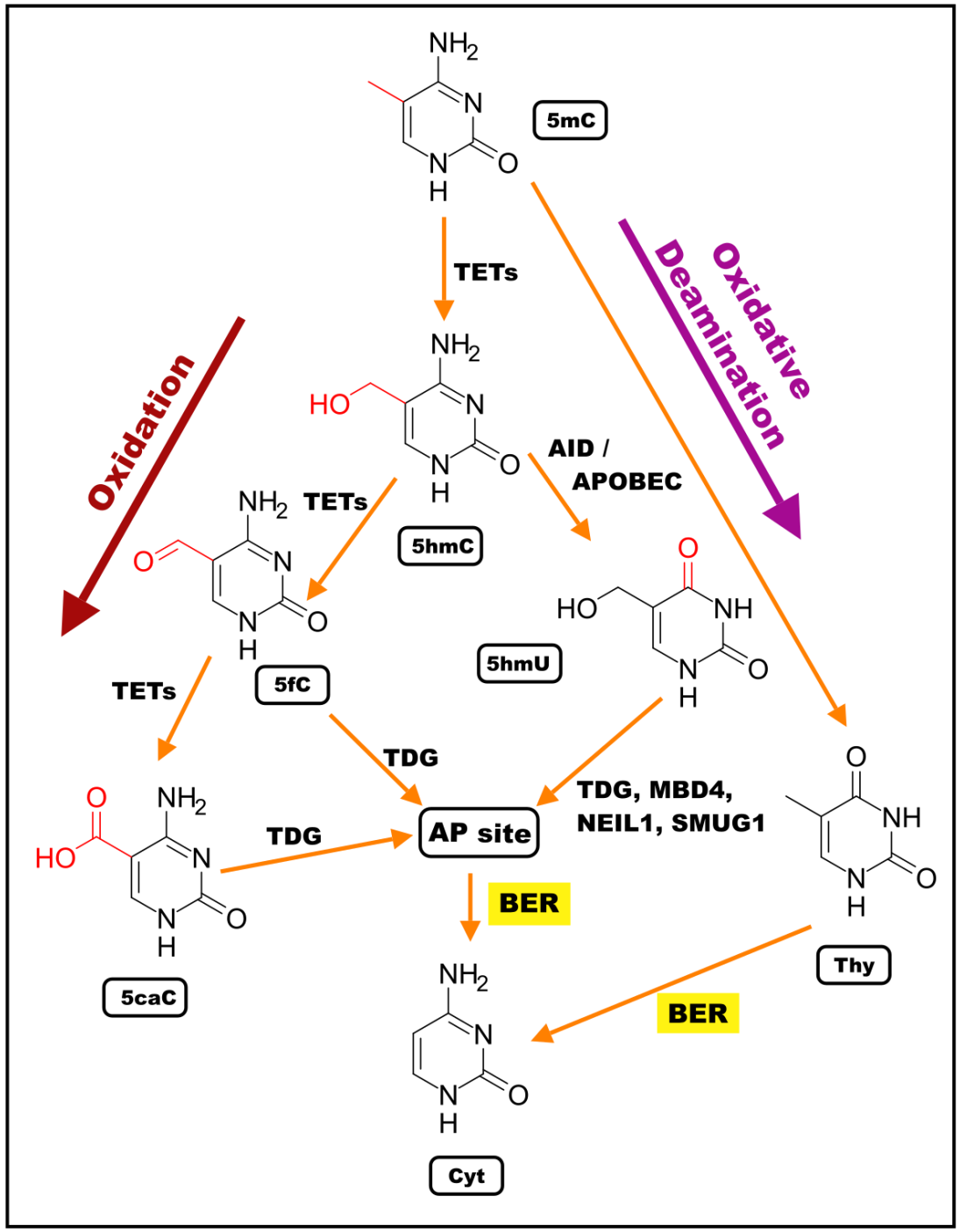
Passive Demethylation



from W. Reik & J. Walter, Nat. Rev. Genet. 2001

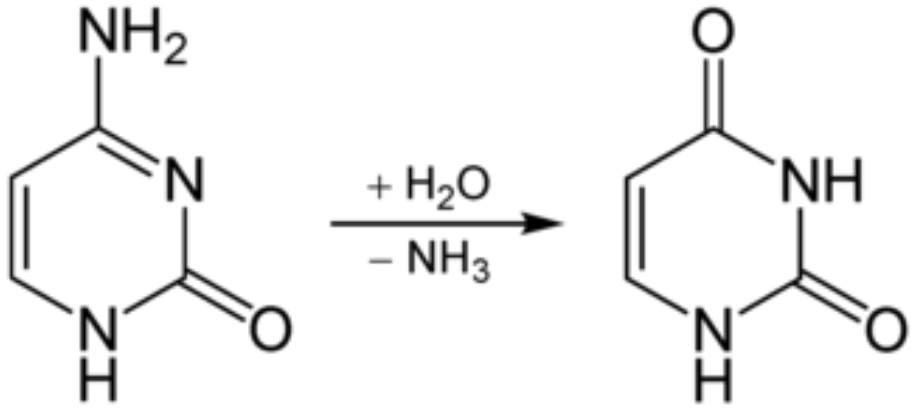
Active Demethylation

- C to mC is a direct reaction
- mC to C is indirect
  - mC to 5fC
  - mC to 5hmU
- Base Excision Repair then converts back to Cytosine



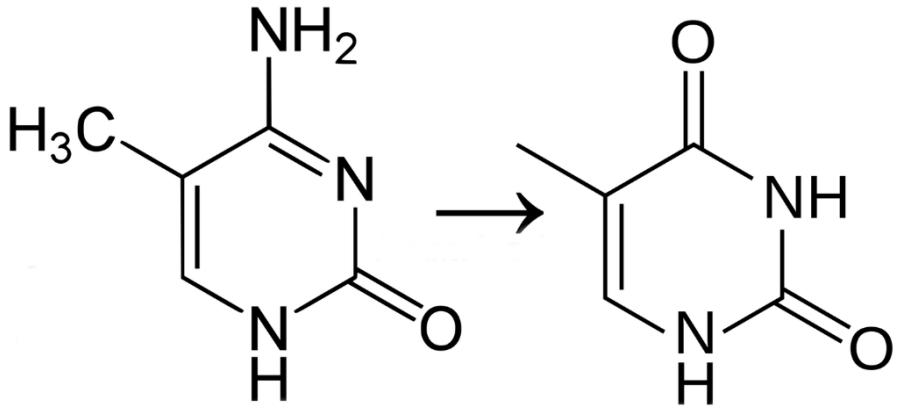
Cytosine Mutation

Spontaneous Deamidation



Cytosine **Uracil**

- Common occurrence
- Uracil is non-natural base
- Fixed by Uracil Degylcosylase



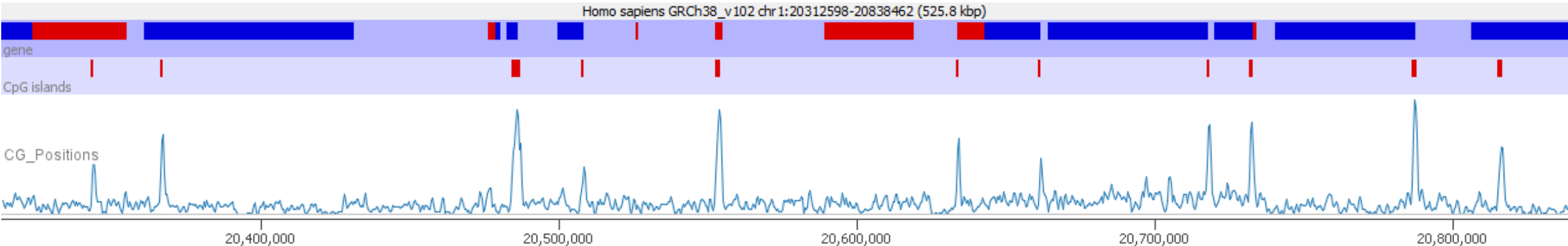
5m-Cytosine **Thymine**

- Less common occurrence
- Thymine is a natural base
- Often leads to C to T Mutation

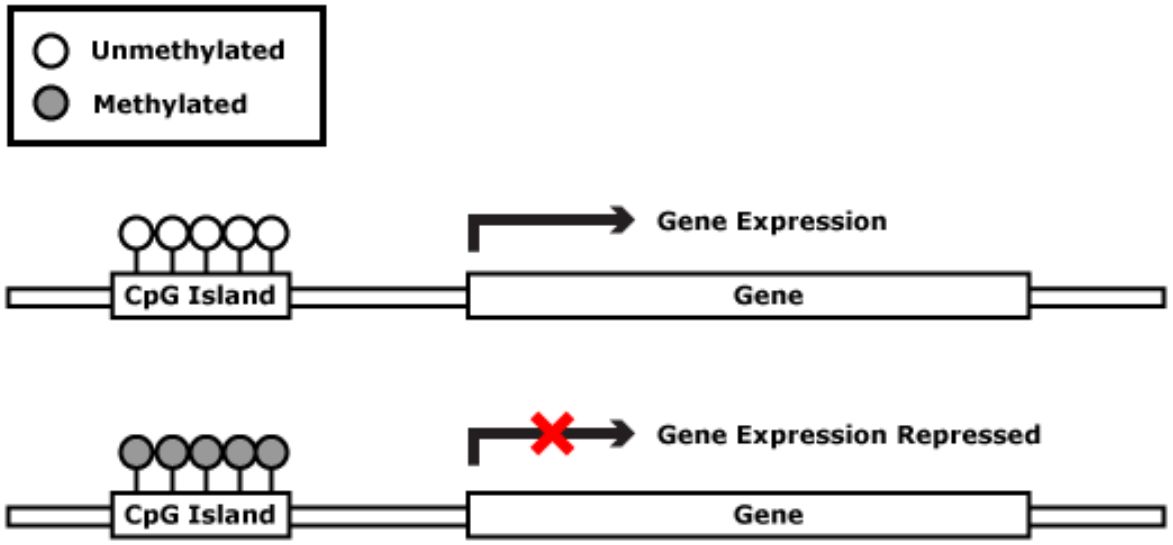
Bringing Things Together

- Methylation only persists on CpG dinucleotides
- Methylated Cytosines tend to mutate to Thymine (CpG > CpT)

- CpGs are rare in the genome – concentrate at **CpG islands**
- CpGs **inside** islands are mostly **unmethylated**
- CpGs **outside** islands are mostly **methylated**

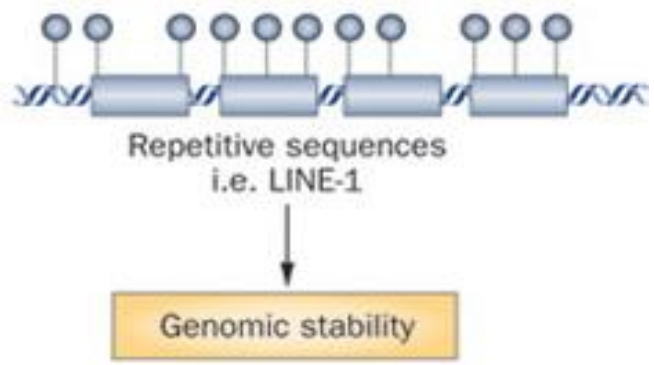


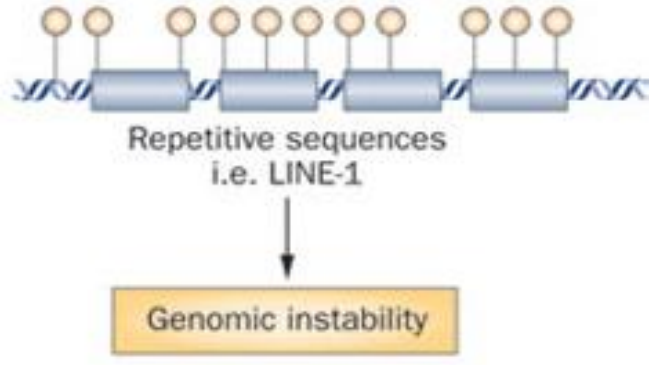
Regulation by DNA methylation



Silencing of gene expression
Tissue differentiation and embryonic development

Faults in correct DNA methylation may result in
- early development failure
- epigenetic syndromes
- cancer



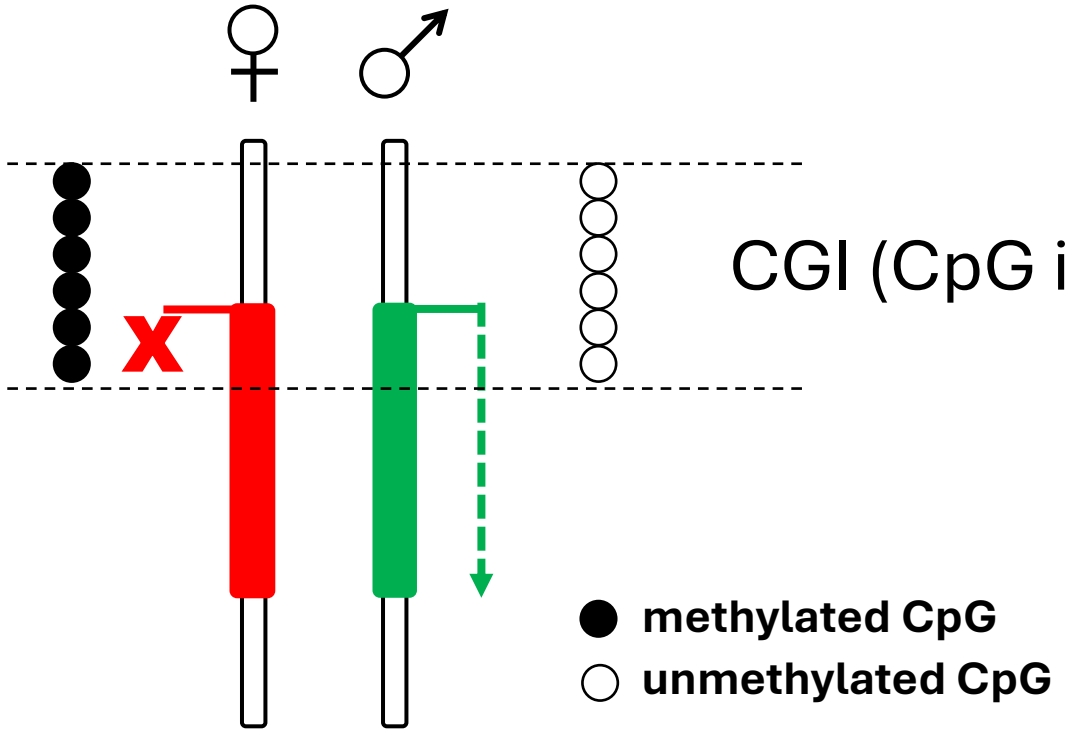


Repeat activity
Genomic stability

Genomic Imprinting

mono-allelic expression

Differential allelic
DNA methylation



Imprinted Genes: Mono-allelic expression with parent-of-origin specificity.
Have key roles in energy metabolism, placenta functions.

Measuring Cytosine Methylation by Sequencing

Short Read Sequencing
(Illumina, Aviti etc)

- Based on DNA Polymerase
- C and mC both pair to G
- Can’t directly sequence mC

- Need a chemical trick to make mC look different to C

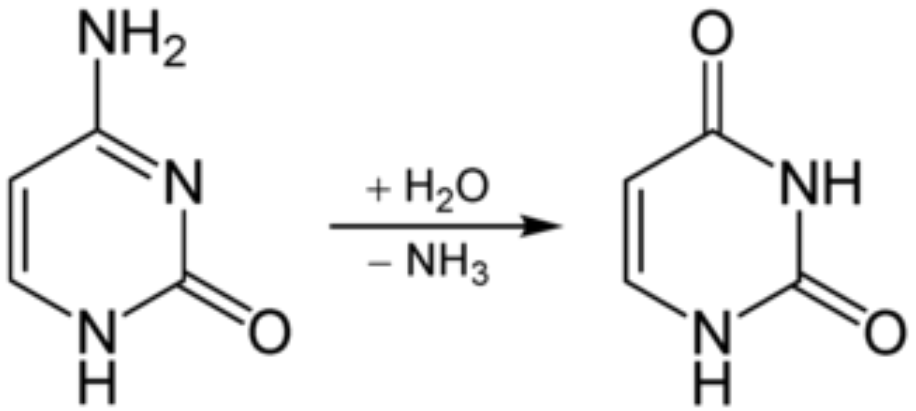
Long Read Sequencing
(Oxford Nanopore)

- Based on electrical signal as DNA passes through a pore
- C and mC give different signals
- Can sequence methylation directly

- Need a base caller trained for this

Bisulphite-Seq or EM-Seq

Unmethylated Cytosine

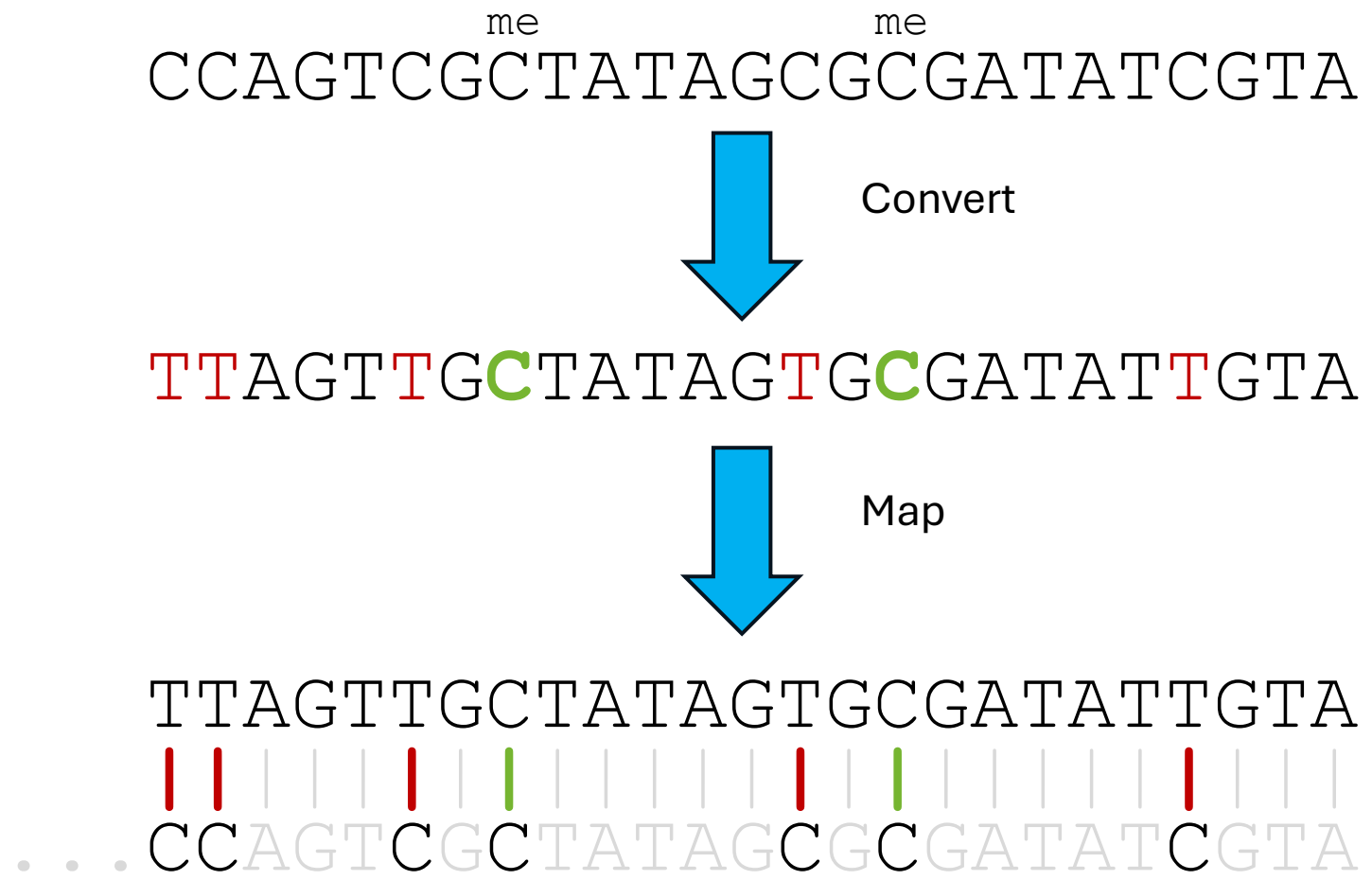


Uracil (Recognised as Thymine)

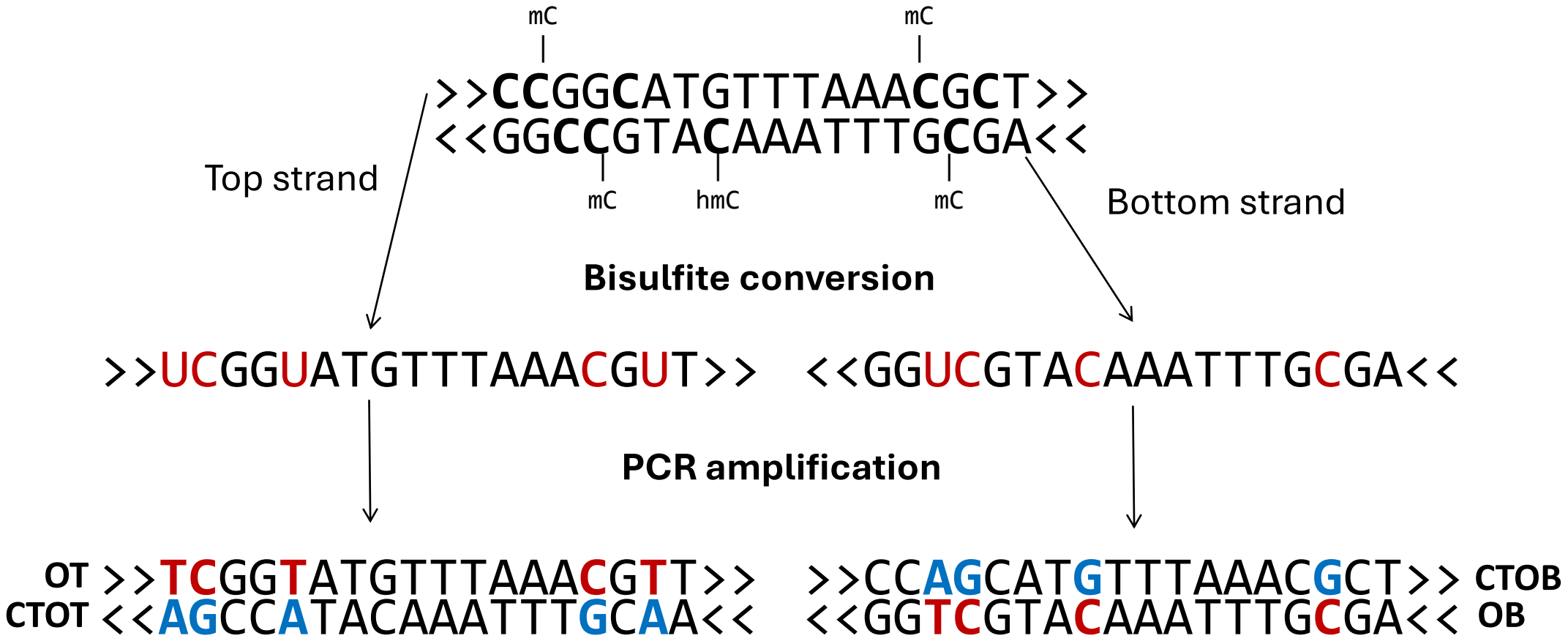
Methylated Cytosines are **NOT** affected

- **Chemical (Sodium Bisulphite):** Easy, cheap, harsh
- **Enzymatic (TET, T4BGT, APOBEC3A):** Efficient, more expensive!

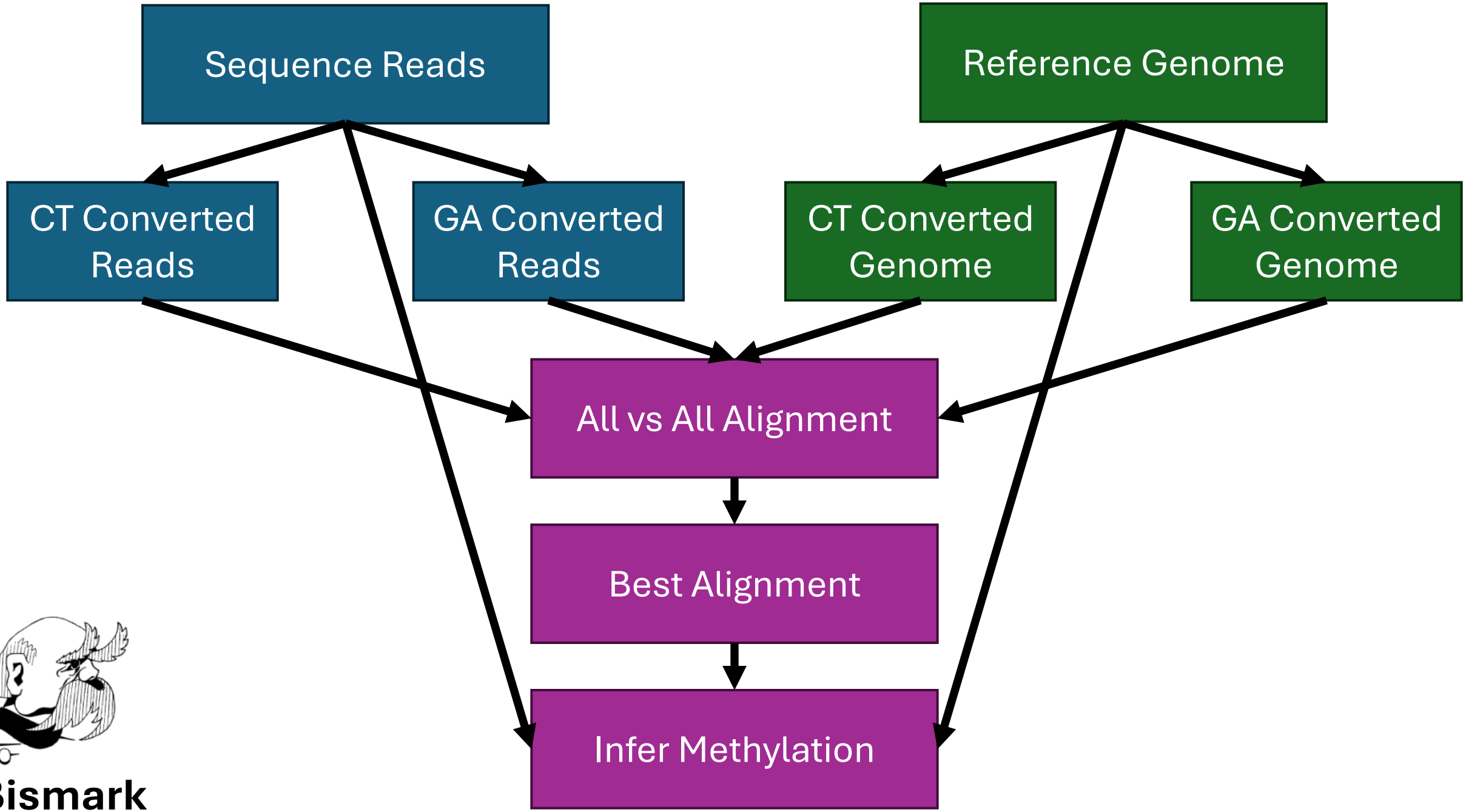
Bisulfite Sequencing



Sequencing Double Stranded DNA



FOUR possible sequences from the same original region


Sequence Reads
Reference Genome
CT Converted Reads
GA Converted Reads
CT Converted Genome
GA Converted Genome
All vs All Alignment
Best Alignment
Infer Methylation

Bismark

Raw Sequence Reads

Sequence QC Checks

Bismark Alignment

PCR Duplicate Removal

Extract Methylation Calls

Create Report

Bismark

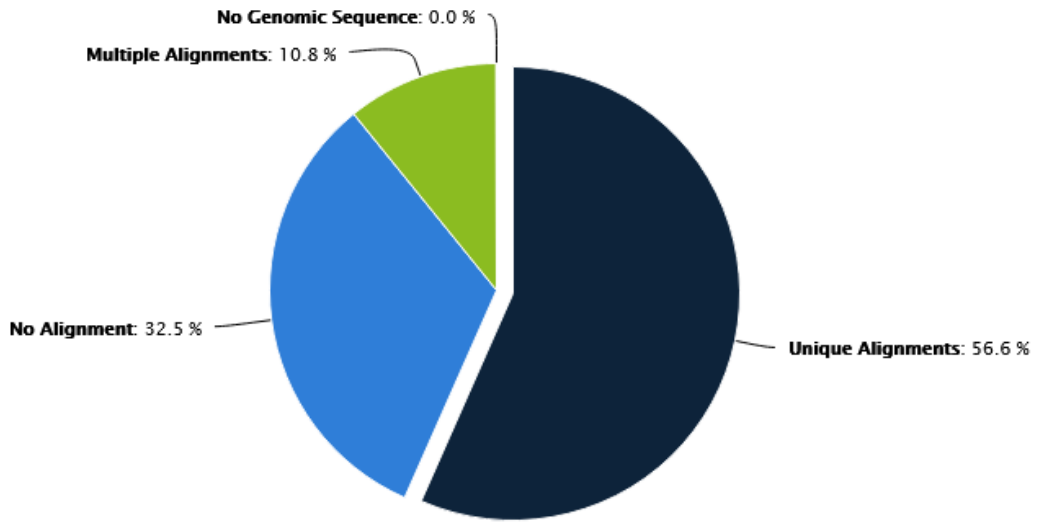
nextflow

Bismark Processing Report

lane1_R1.fq.gz

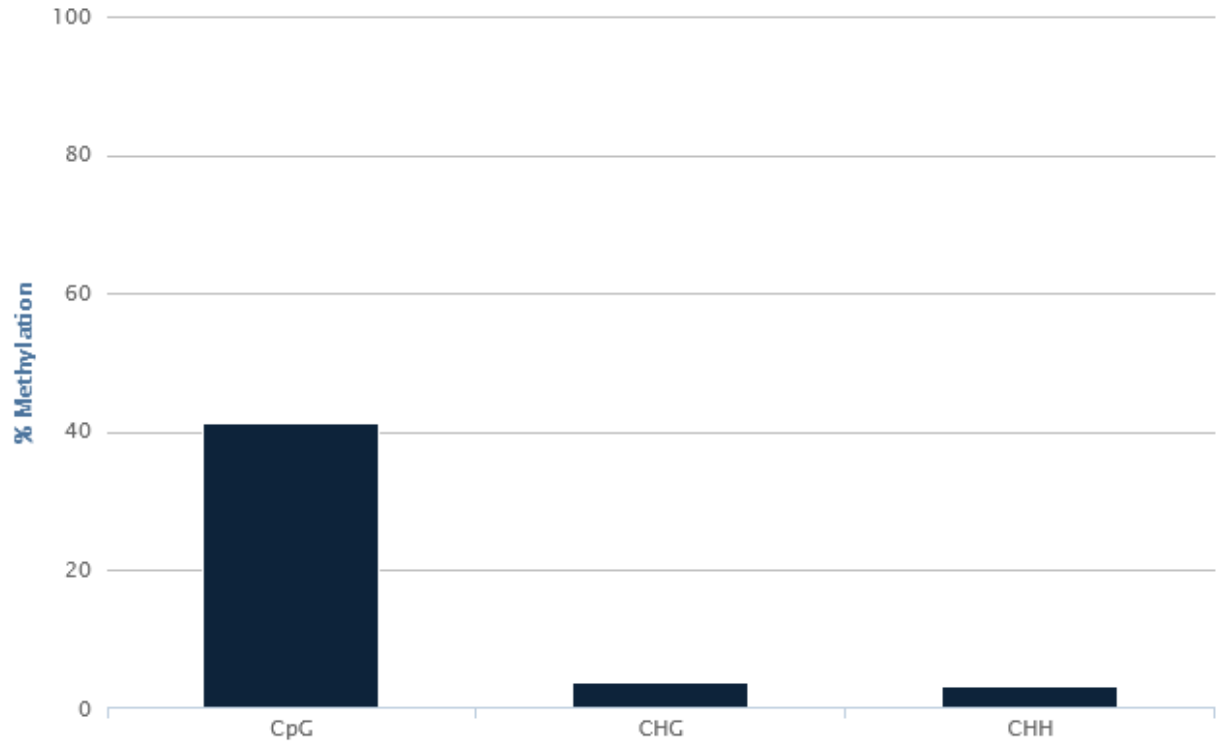
Alignment

Sequences analysed in total	447299373
Single-end alignments with a unique best hit	253328964
Sequences without alignments under any condition	145567291
Sequences that did not map uniquely	48403118
Genomic sequence context not extractable (edges of chromosomes)	17



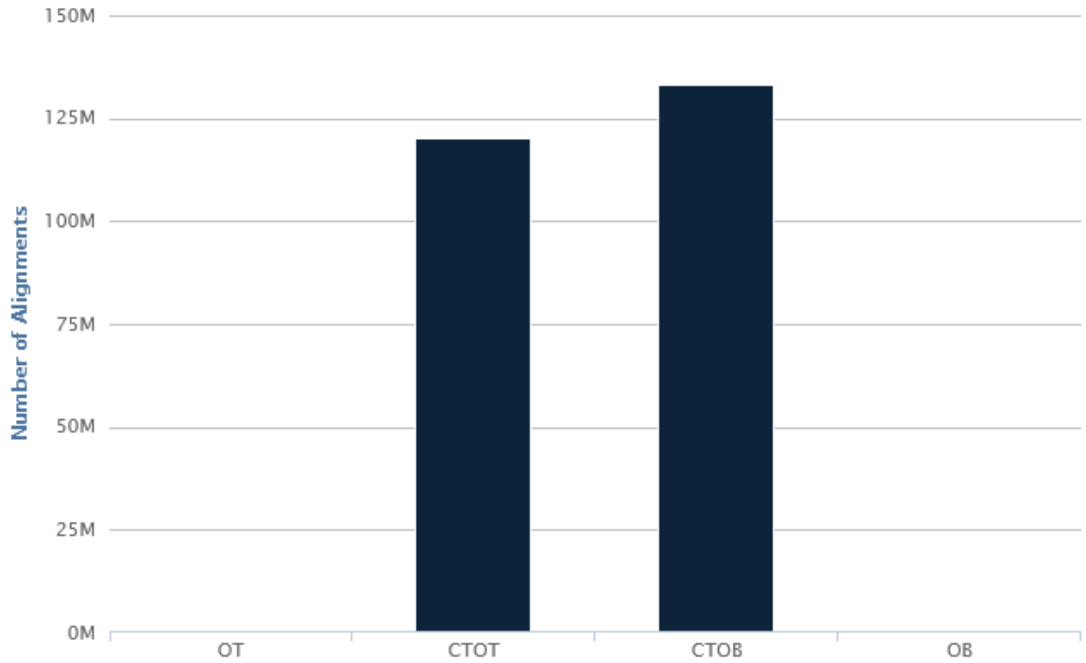
Cytosine Methylation

Total C's analysed	5232303260
Methylated C's in CpG context	114439471
Methylated C's in CHG context	47636473
Methylated C's in CHH context	118416606
Unmethylated C's in CpG context	164043167
Unmethylated C's in CHG context	1204069440
Unmethylated C's in CHH context	3583698103
Percentage methylation (CpG context)	41.1%
Percentage methylation (CHG context)	3.8%
Percentage methylation (CHH context)	3.2%



Alignment to Individual Bisulfite Strands

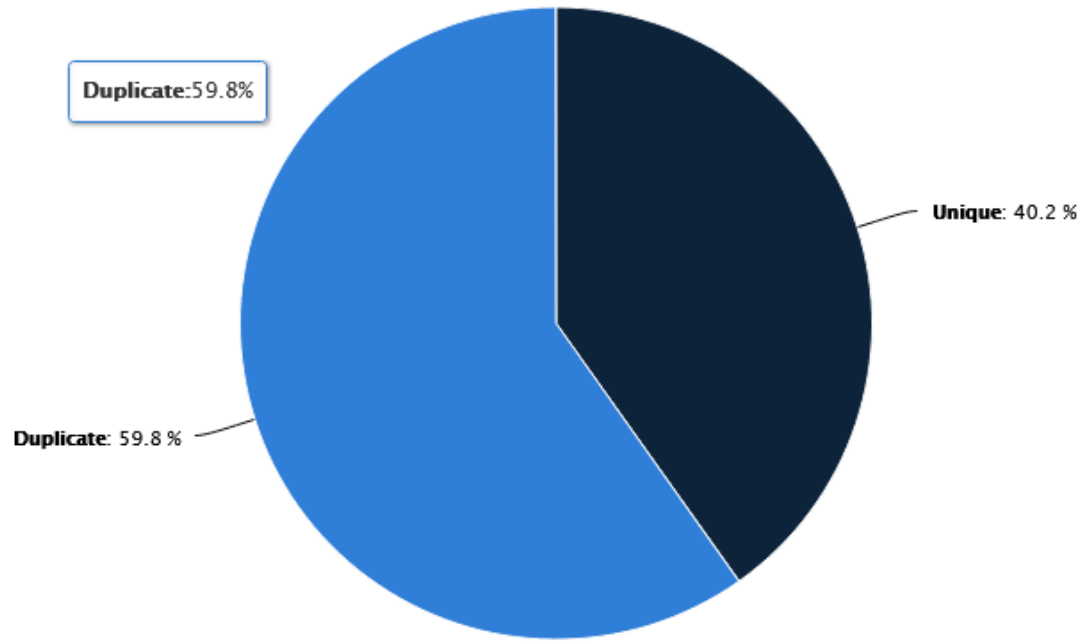
OT	0	original top strand
CTOT	120197022	complementary to original top strand
CTOB	133131942	complementary to original bottom strand
OB	0	original bottom strand



Deduplication

Alignments analysed	253328947
Unique alignments	101795844
Duplicates removed	151533103

Duplicated alignments were found at **54946655** different positions



Bismark Bisulfite Mapper

A tool to map bisulfite converted sequence reads and determine cytosine methylation states

View on GitHub

Bismark Bisulfite Mapper

User Guide - v0.23.0

30 September, 2020

This User Guide outlines the Bismark suite of tools and gives more details for each individual step. For troubleshooting some of the more commonly experienced problems in sequencing in general and bisulfite-sequencing in particular please browse through the sequencing section at QCFail.com.

Technique	5' Trimming	3' Trimming	Mapping	Deduplication	Extraction
BS-Seq	☐	☐	☐	✅	`--ignore_r 2 2`
RRBS	`--rrbs` (R2 only)	`--rrbs` (R1 only)	☐	❌	☐
RRBS (NuGEN Ovation)	special processing	special processing	☐	❌	`--ignore_r 2 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)	R1: 10, R2:15bp	(10 bp)	☐	✅	☐
Zymo Pico-Methyl	10 bp	(10 bp)	`--non_directional`	✅	☐

http://felixkrueger.github.io/Bismark/Docs/

Bismark Coverage Files

Chromosome	Start	End	%Meth	Unmeth Count	Meth Count
18	6638214	6638214	0	0	3
18	6638233	6638233	0	0	5
18	6638291	6638291	0	0	7
18	6638303	6638303	0	0	6
18	6638322	6638322	0	0	8
18	6638339	6638339	40	2	3
18	6638765	6638765	0	0	1

BioTrain.TV

#Babraham Bioinformatics

Visualising and Exploring Methylation Data

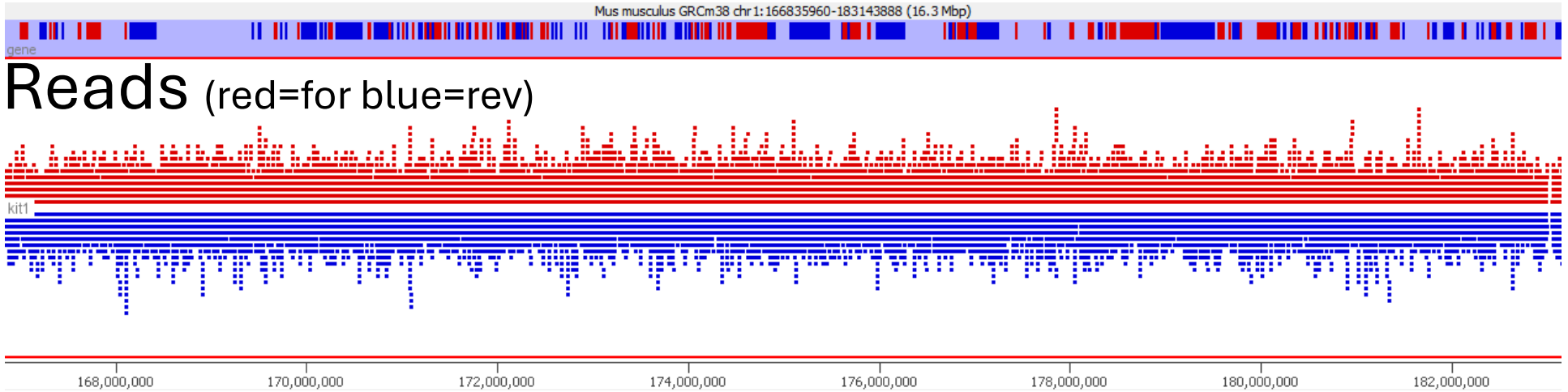
Simon Andrews

Which Data do you use?

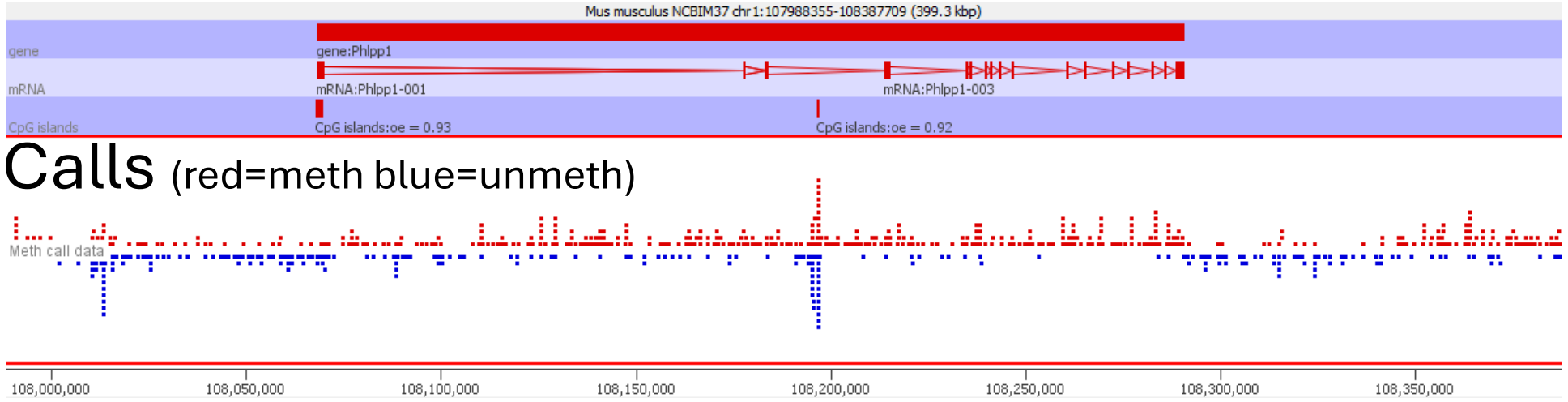
- Methylation contexts
  - CpG: Only generally relevant context for mammals
  - CHG: Only known to be relevant in plants
  - CHH: Generally unmethylated

- Methylation strands
  - CpG methylation is generally symmetric
  - Normally makes sense to merge OT / OB strands

Looking at your data



Reads (red=for blue=rev)



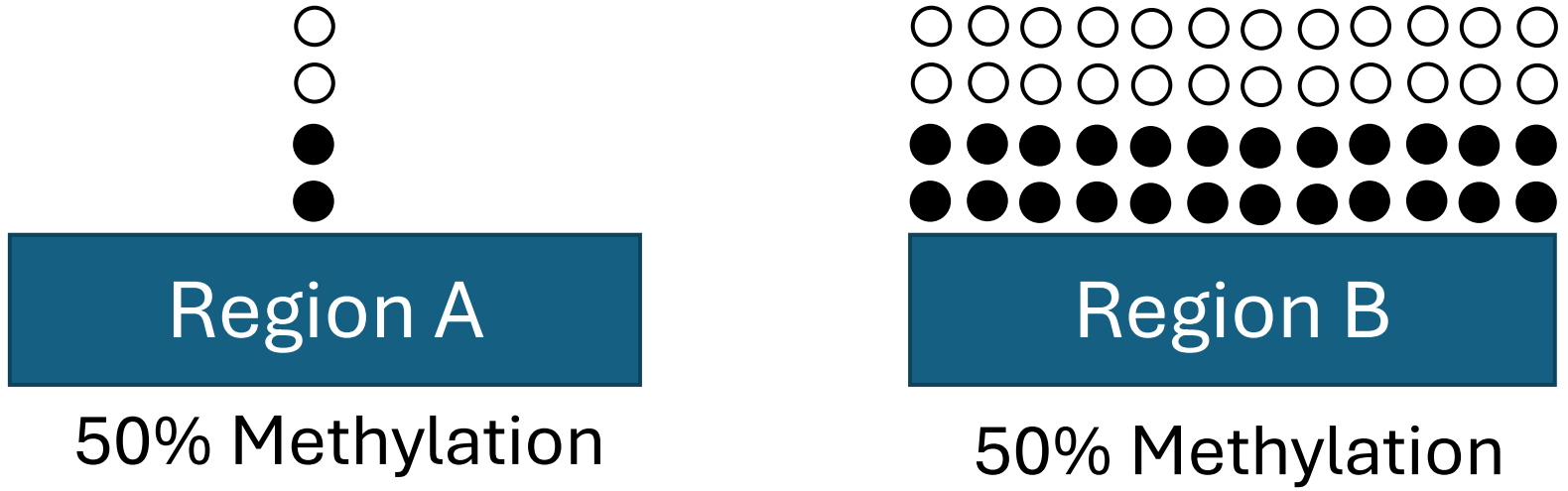
Calls (red=meth blue=unmeth)

Quantitating your Data (Percent Methylation)

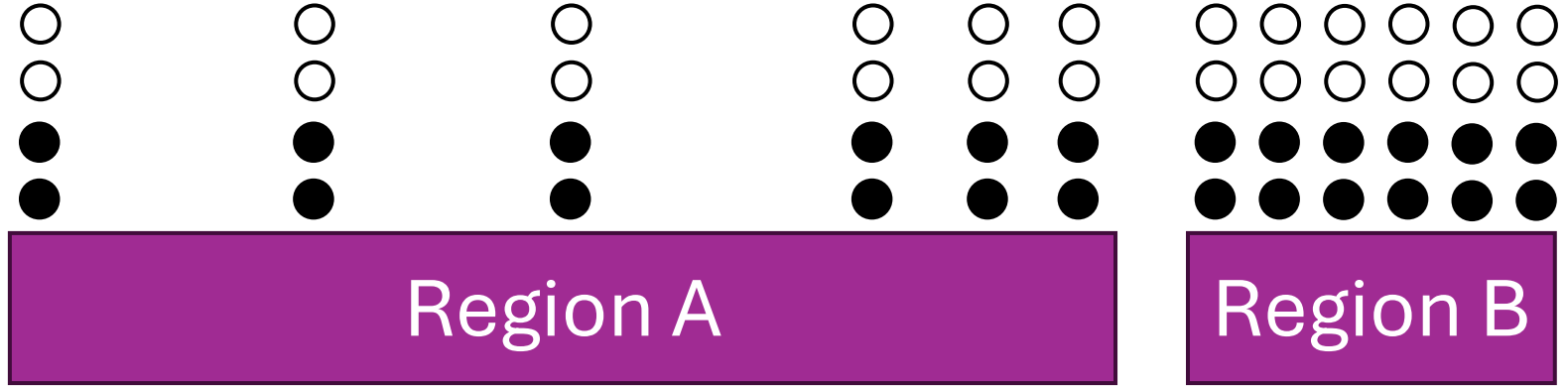
- Per Base
  - Is possible, but generally not enough data to be practical
- In Windows
  - Windows Tiled over the genome
  - Puts more data into each measurement
- Over Features
  - CpG Islands
  - Gene Bodies

Building Measurement Windows

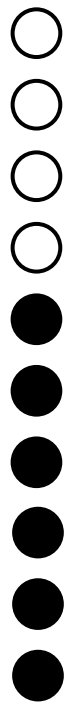
Fixed Size



Fixed Data



Calculating Percentage Methylation

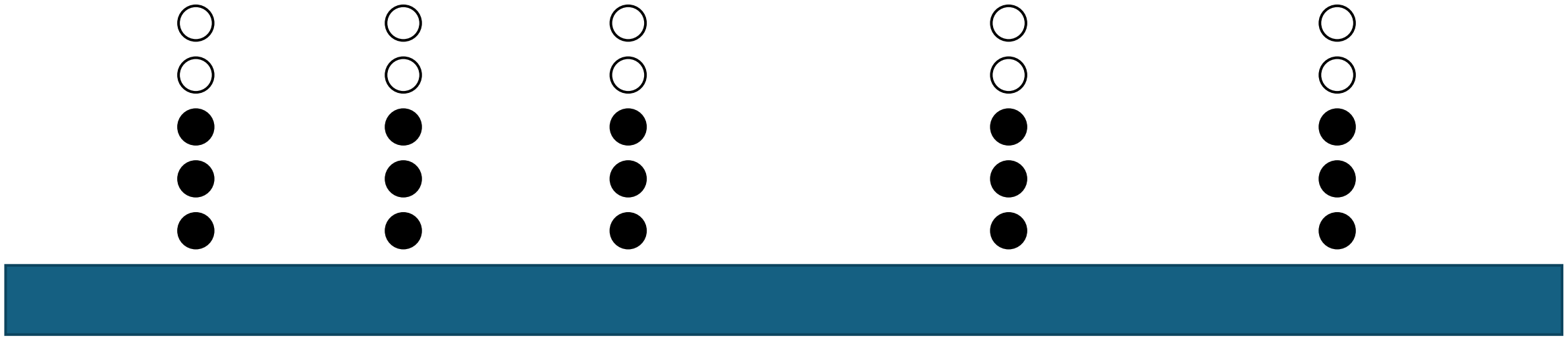


● = meth

○ = unmeth

(6/10) * 100 = 60% methylated

Calculating Percentage Methylation

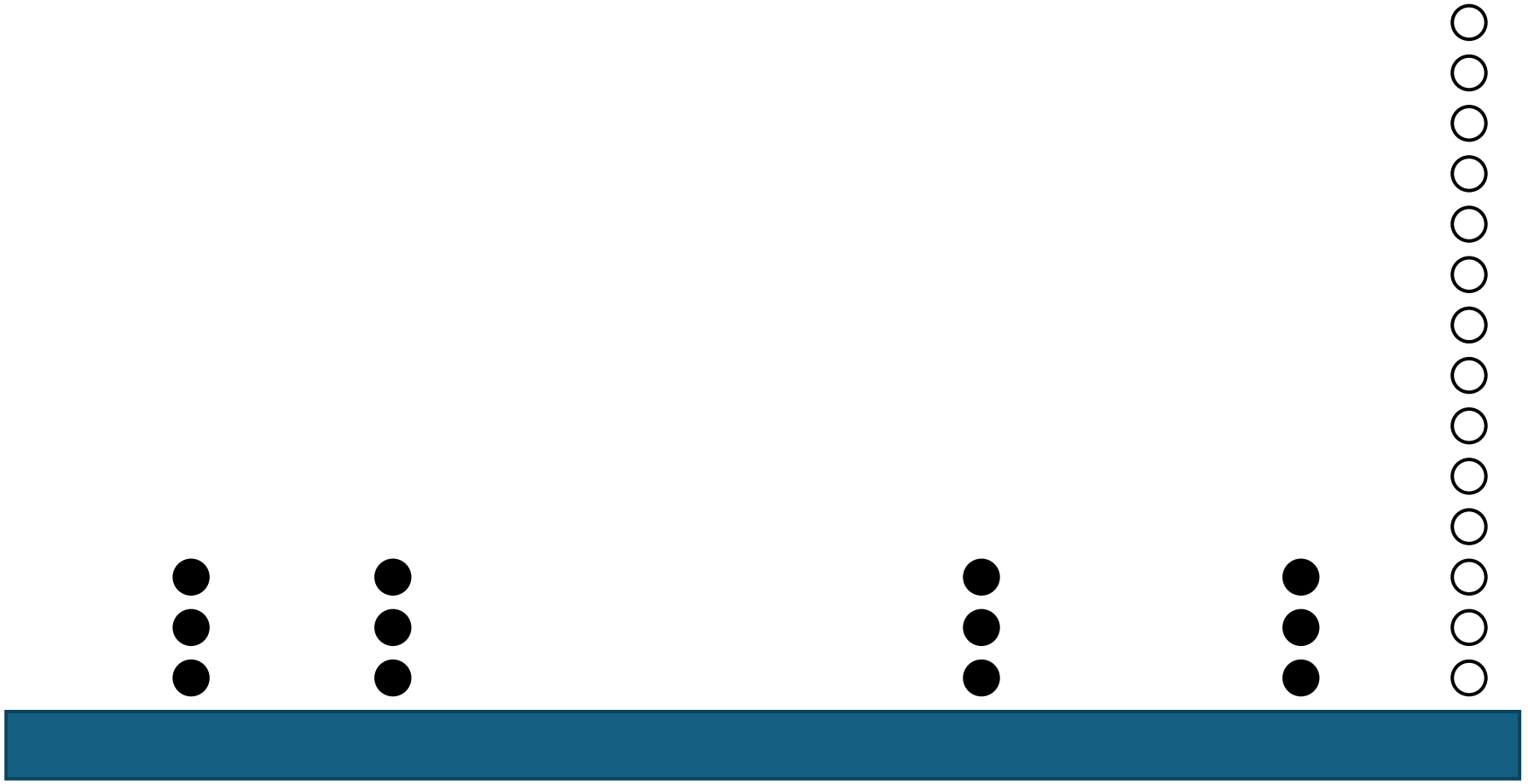


Total methylated calls = 15

Total unmethylated calls = 10

Methylation level = (15/(15+10))*100 = **60%**

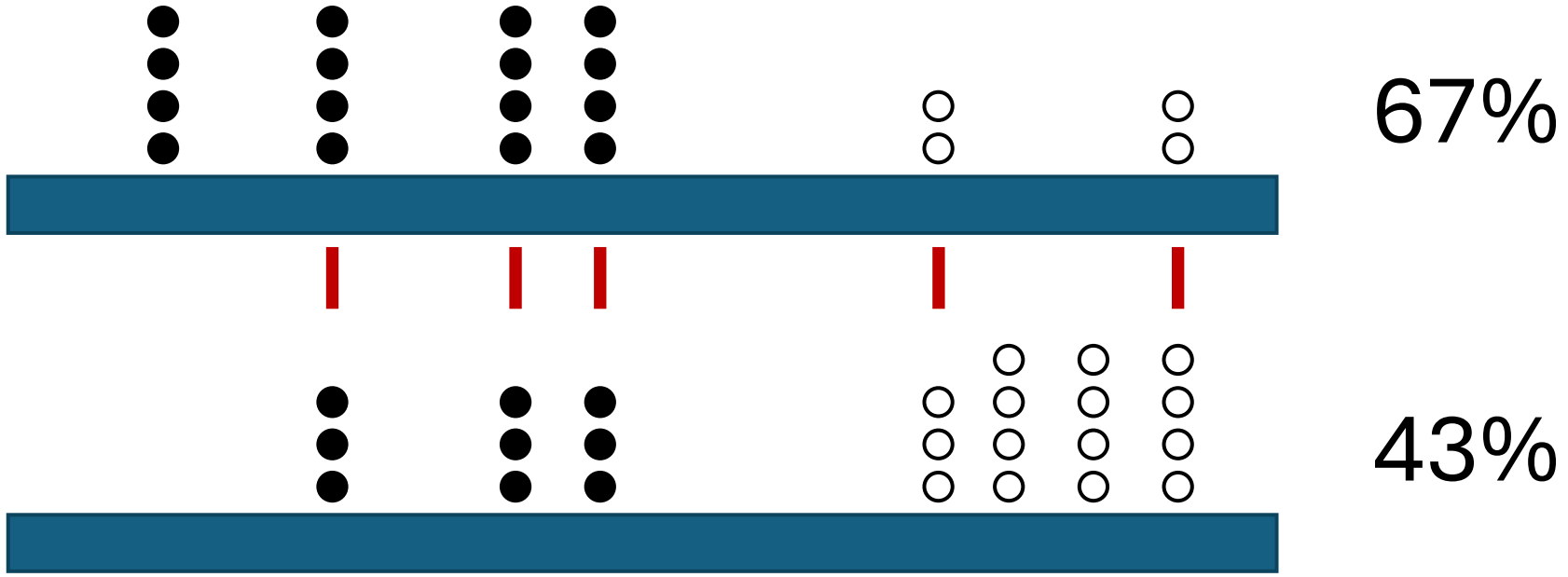
Calculating Percentage Methylation



Percentage methylation from all calls independently = **46%**

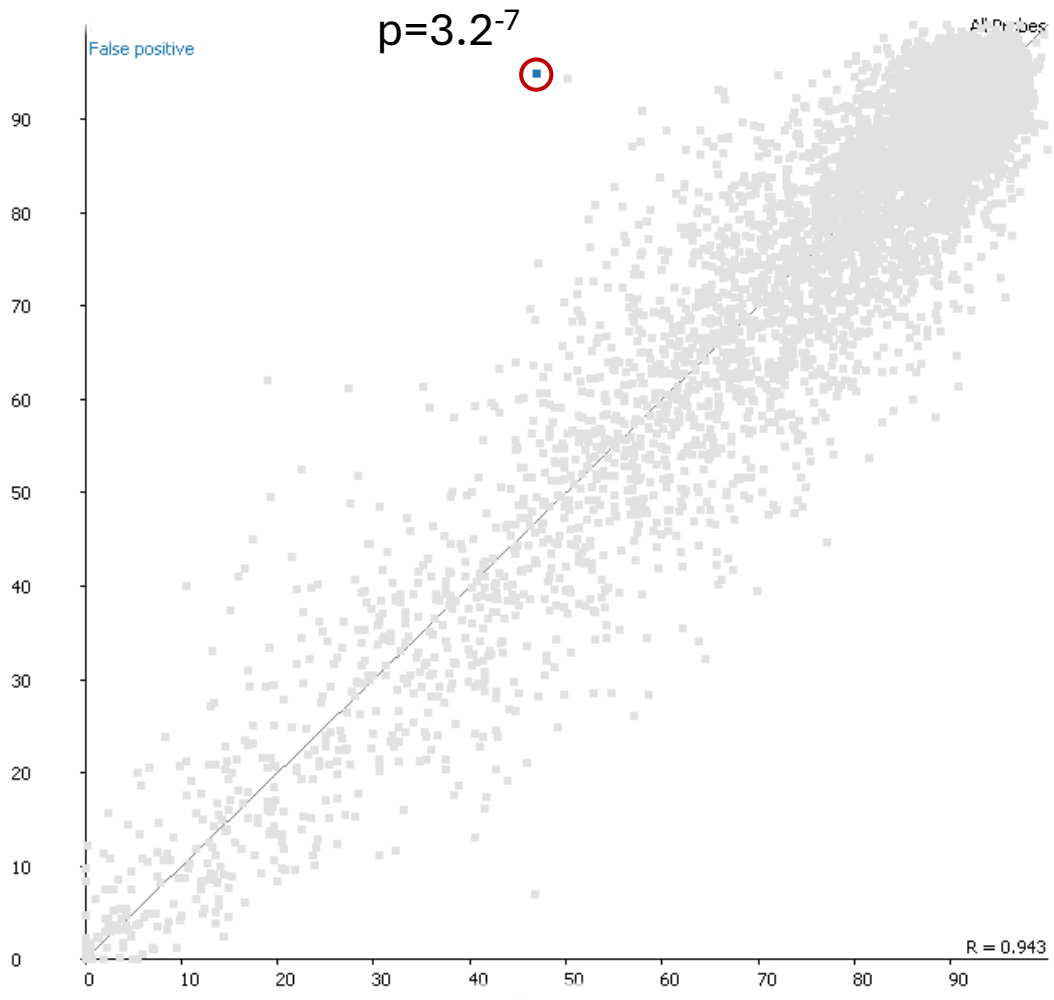
Percentage methylation from mean methylation per base = **80%**

Calculating Percentage Methylation

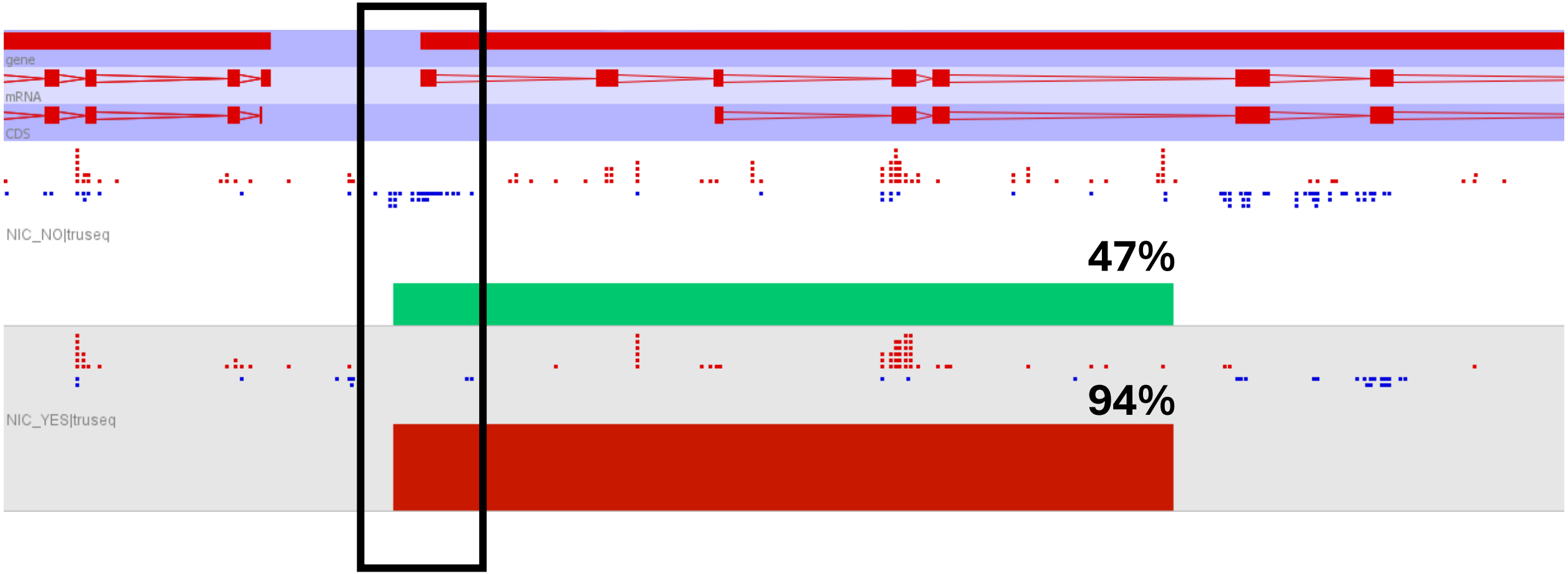


Common = 60% in both

This affects real data



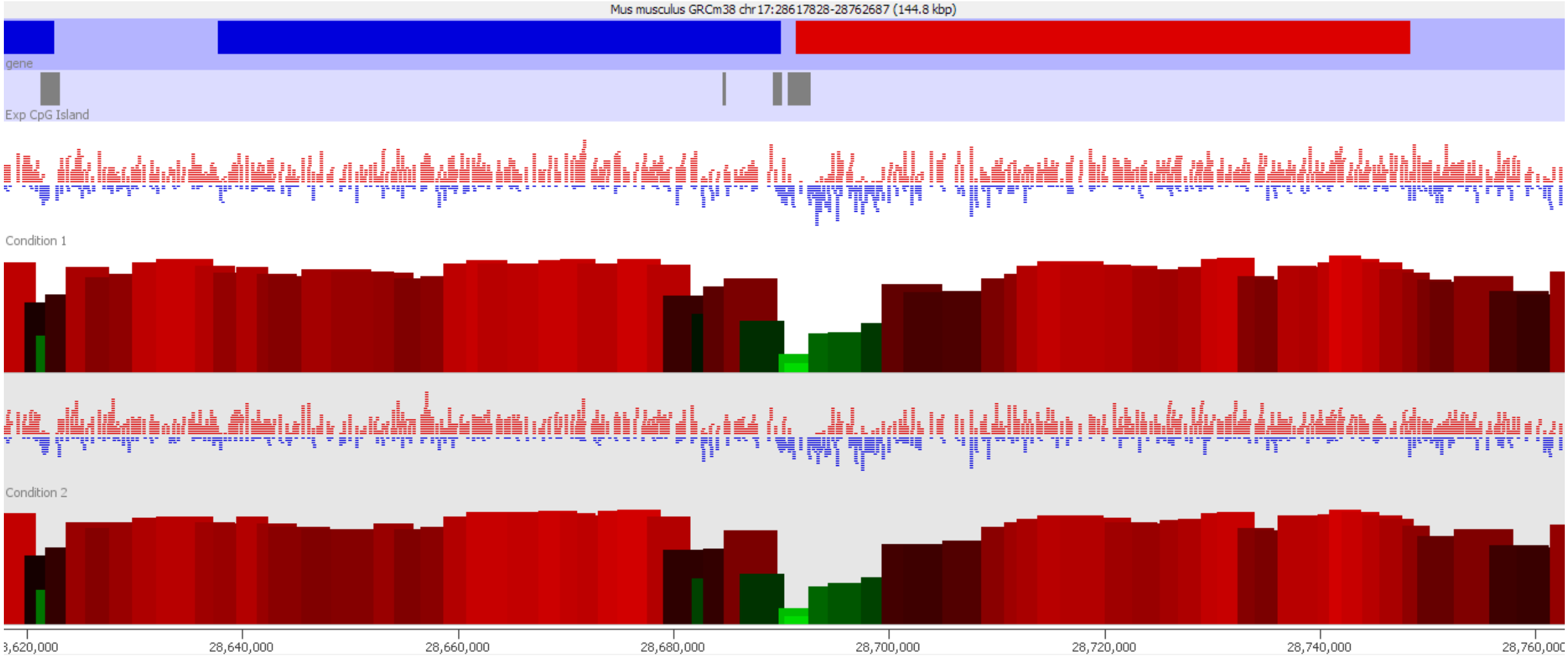
This affects real data



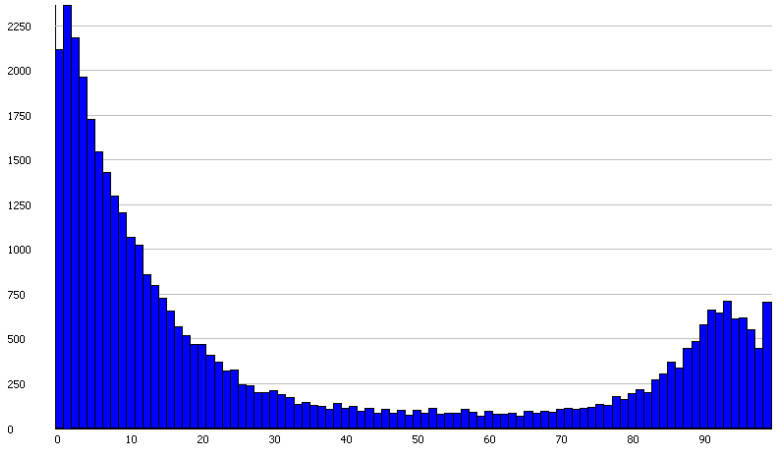
Answering Basic Questions

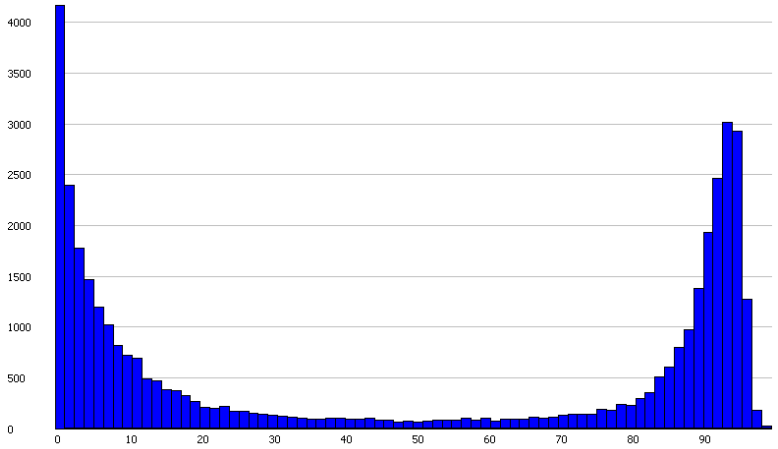
- Patterning
  - What sorts of changes in methylation do I observe along a chromosome
- Distributions
  - What are the overall levels and distributions of methylation values in my samples
- Comparisons
  - On a global scale what is the overall relationship between methylation levels in different samples

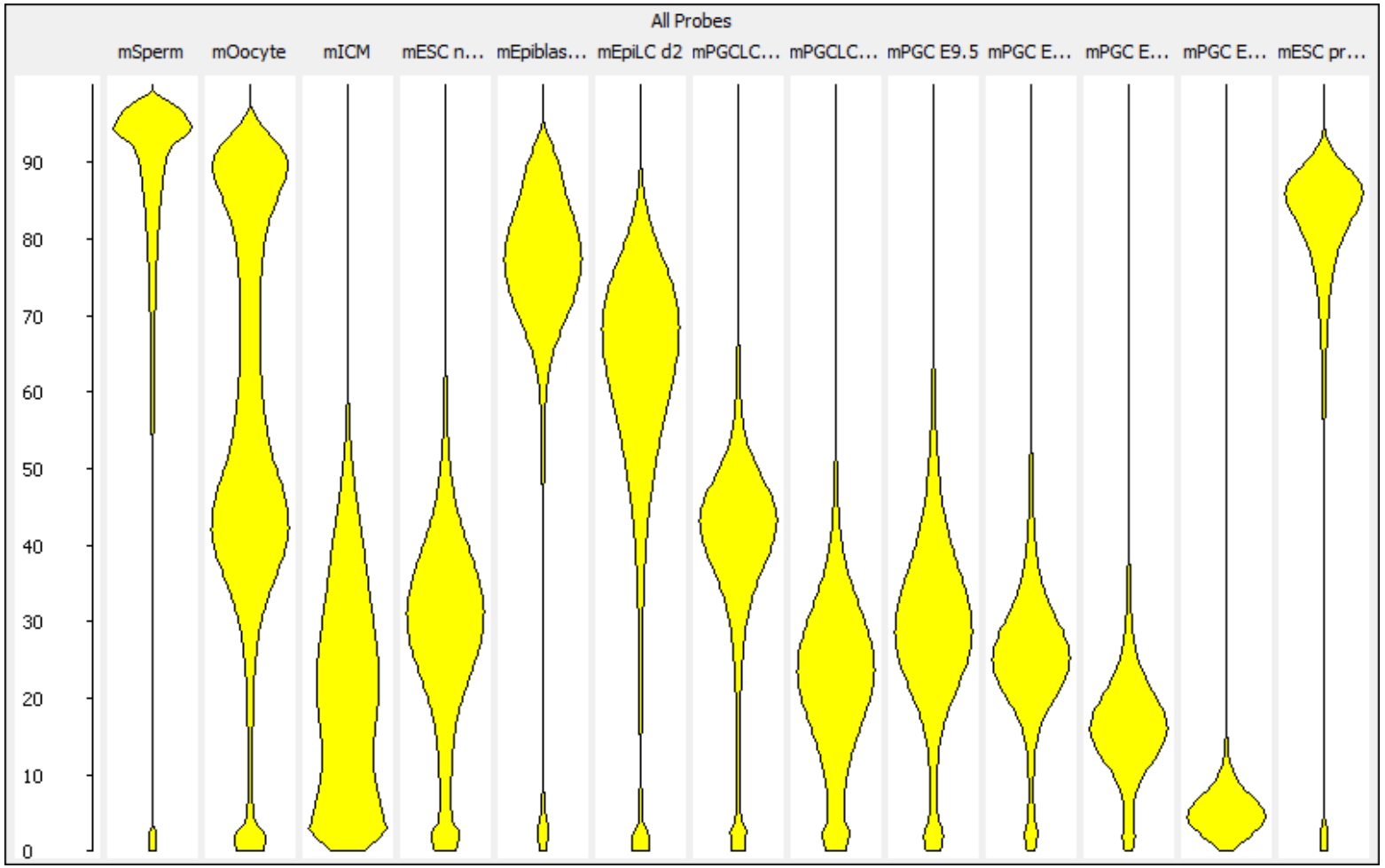
Patterning


Mus musculus GRCm38 chr17:28617828-28762687 (144.8 kbp)
gene
Exp CpG Island
Condition 1
Condition 2
28,620,000
28,640,000
28,660,000
28,680,000
28,700,000
28,720,000
28,740,000
28,760,000

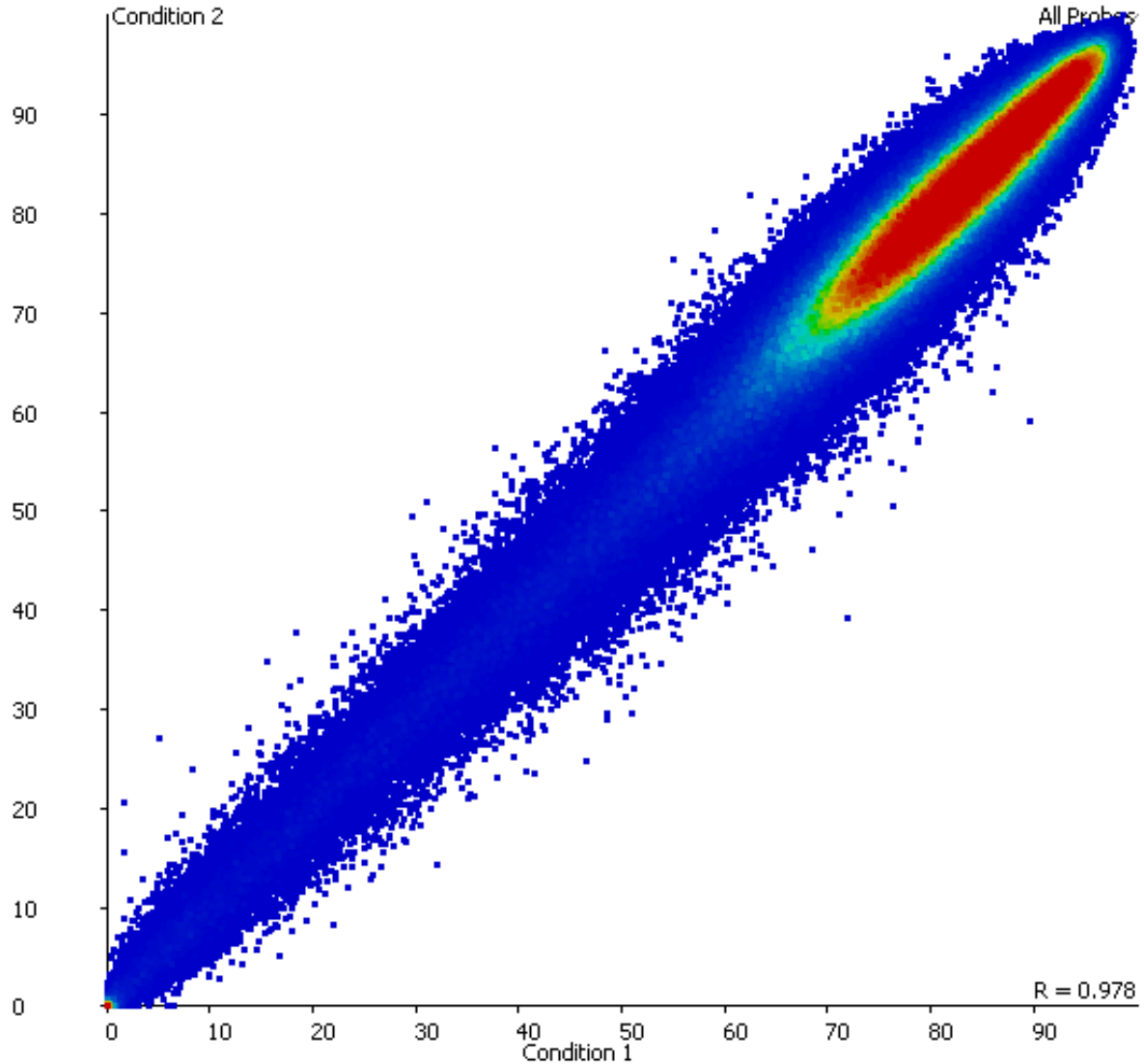
Distributions

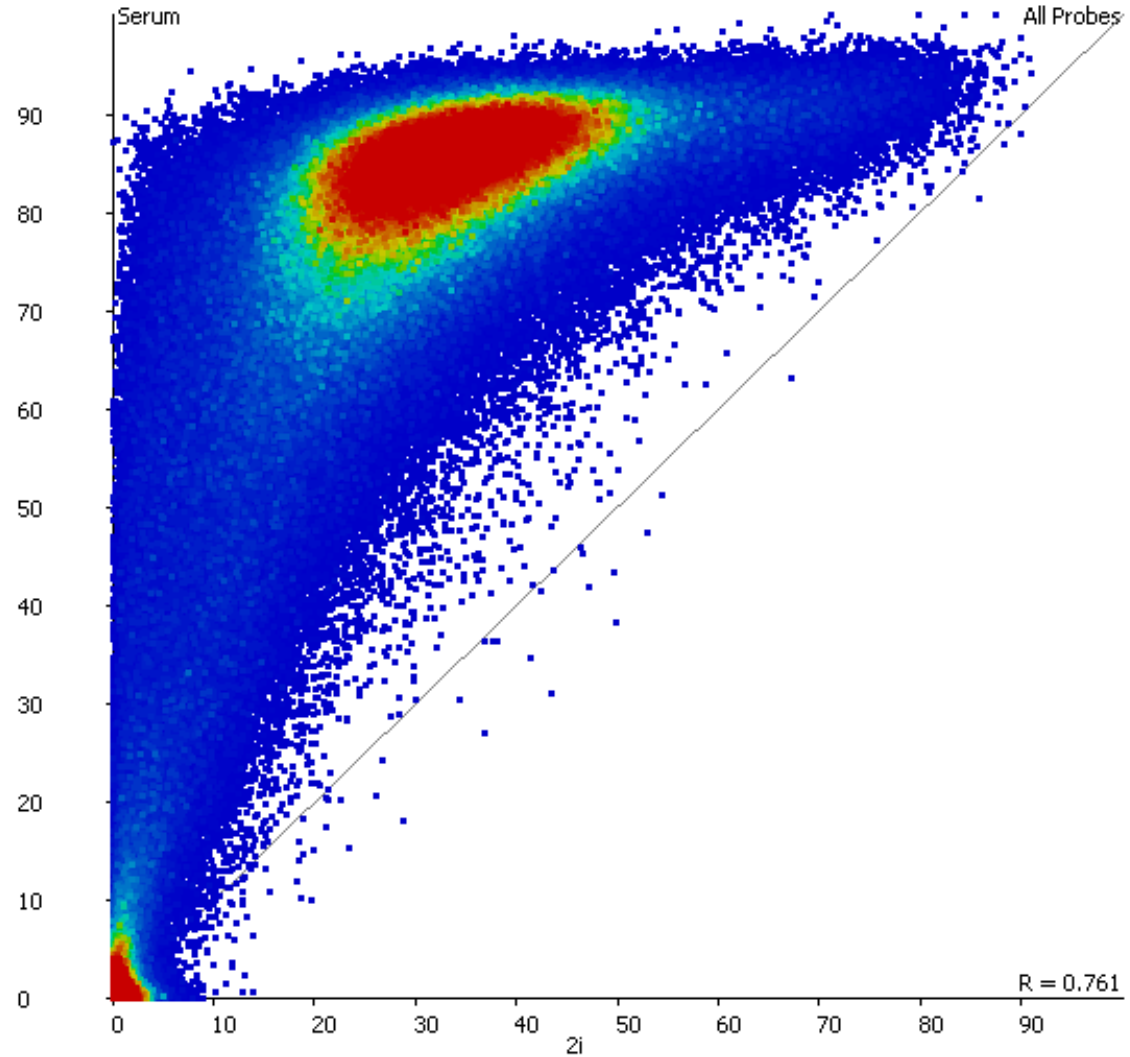




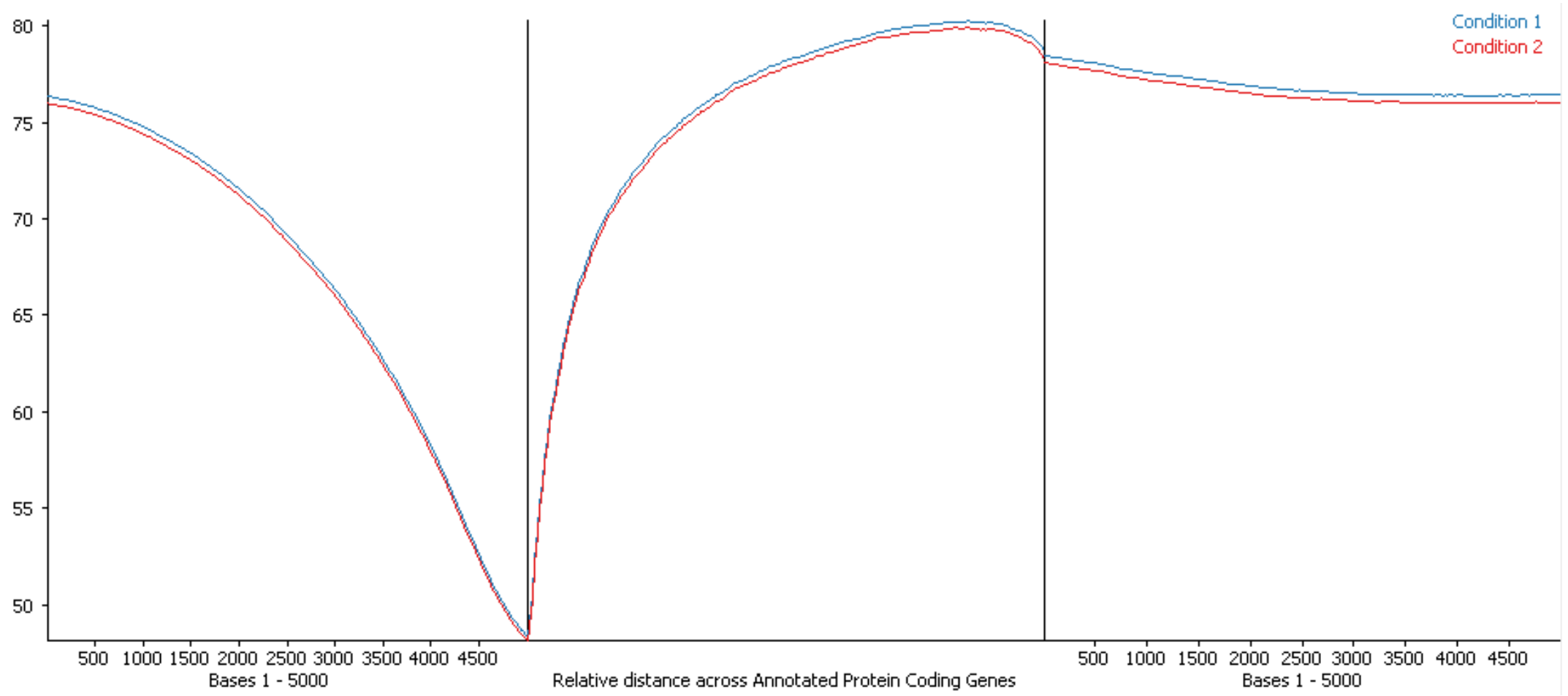


Comparisons





MetaGene Summaries



Upstream Region

Gene Body

Downstream Region

Exercise

Visualising and Exploring Methylation Data in SeqMonk

BioTrain.TV
#Babraham Bioinformatics

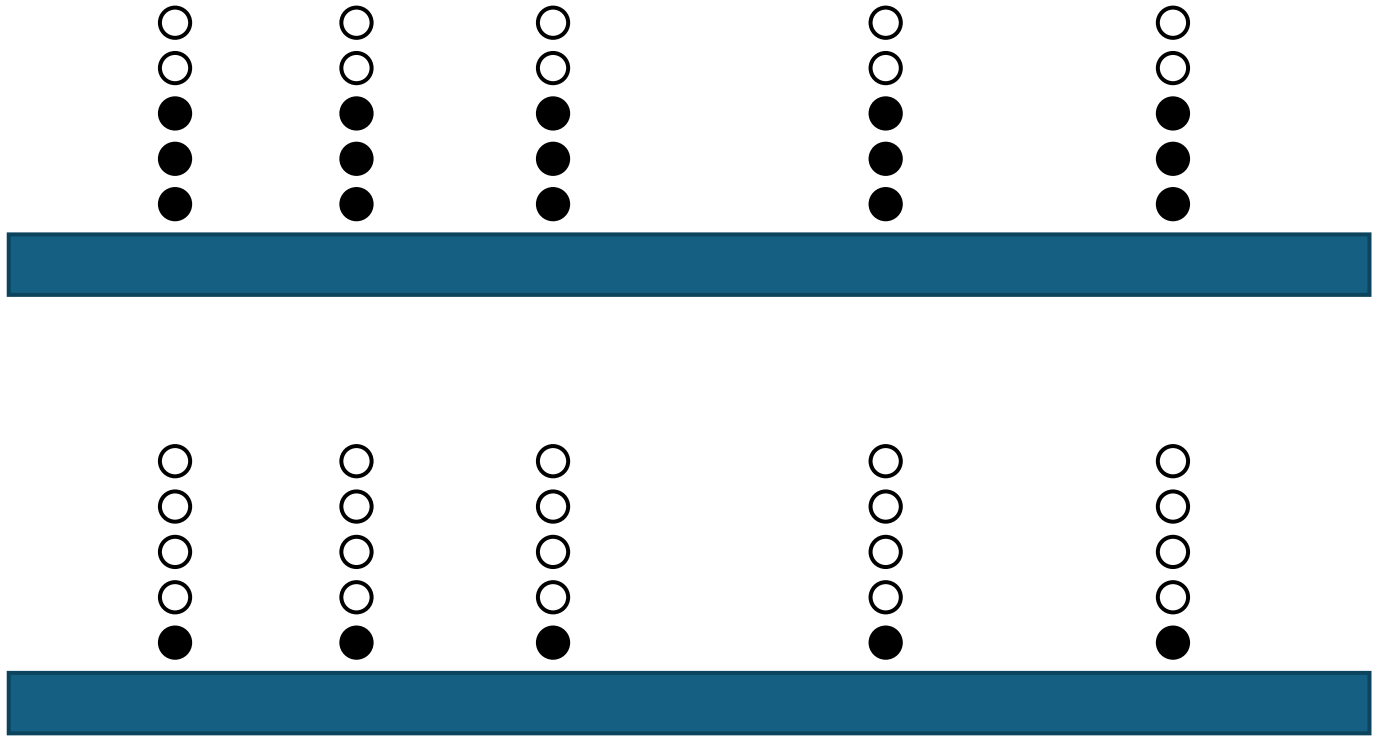
BioTrain.TV

#Babraham Bioinformatics

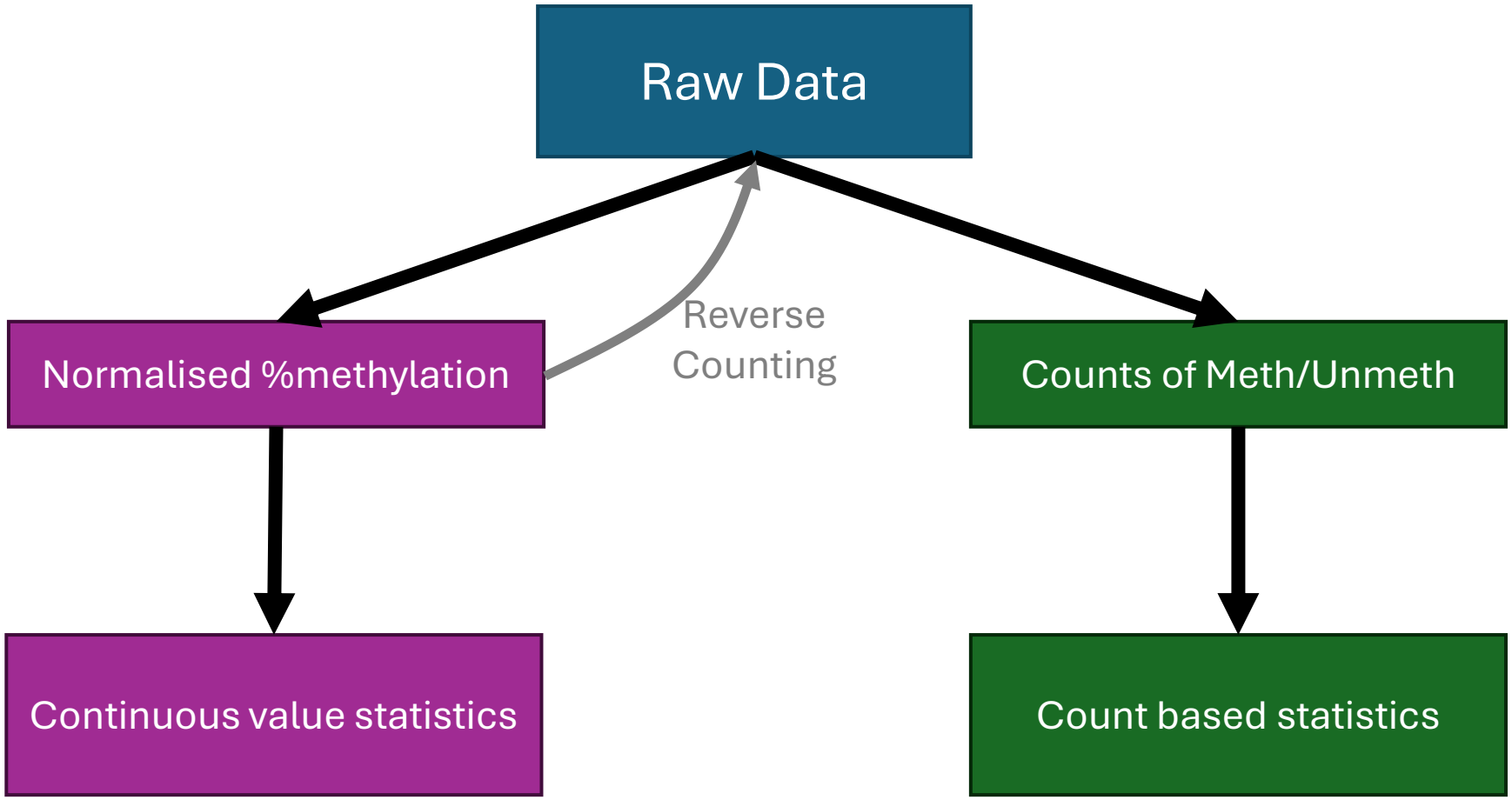
Differential Methylation Statistics

Simon Andrews

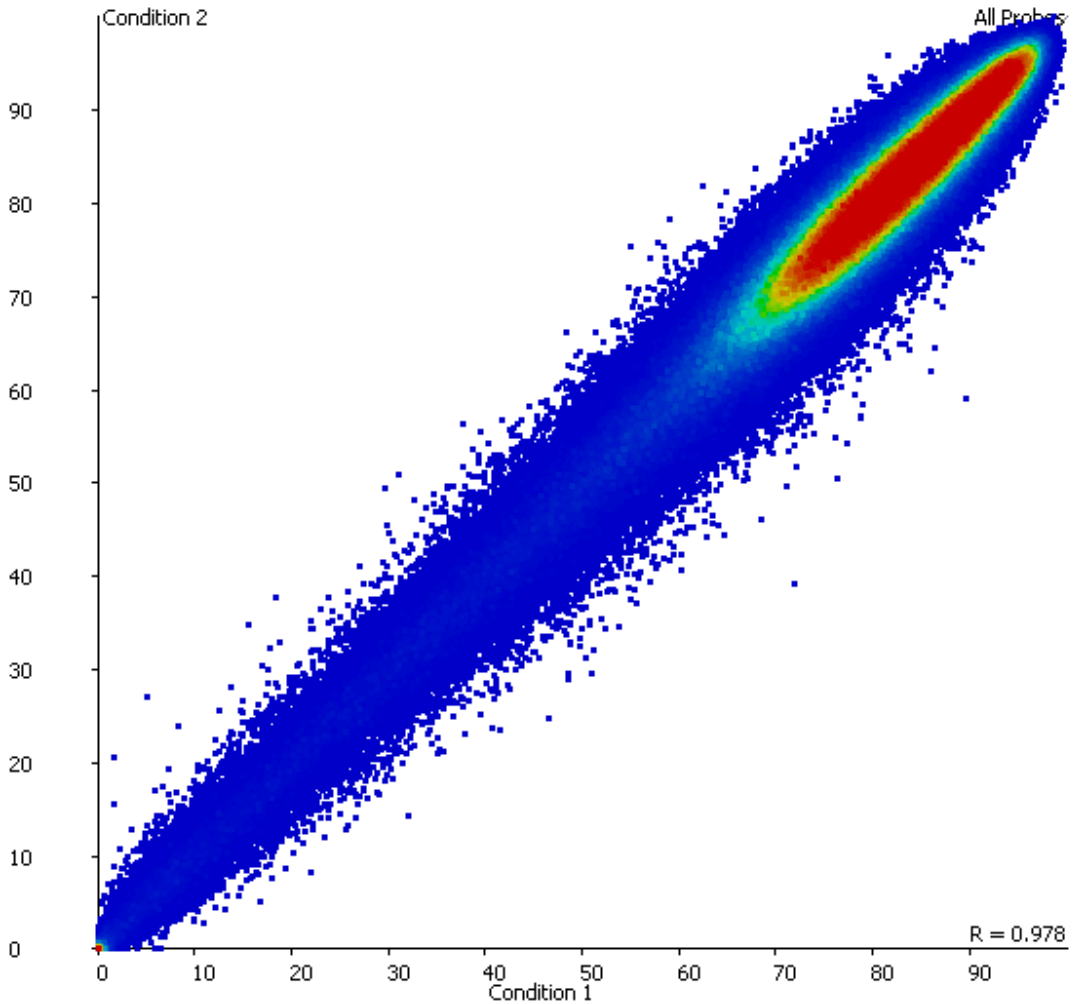
A simple question...



Two Strategies

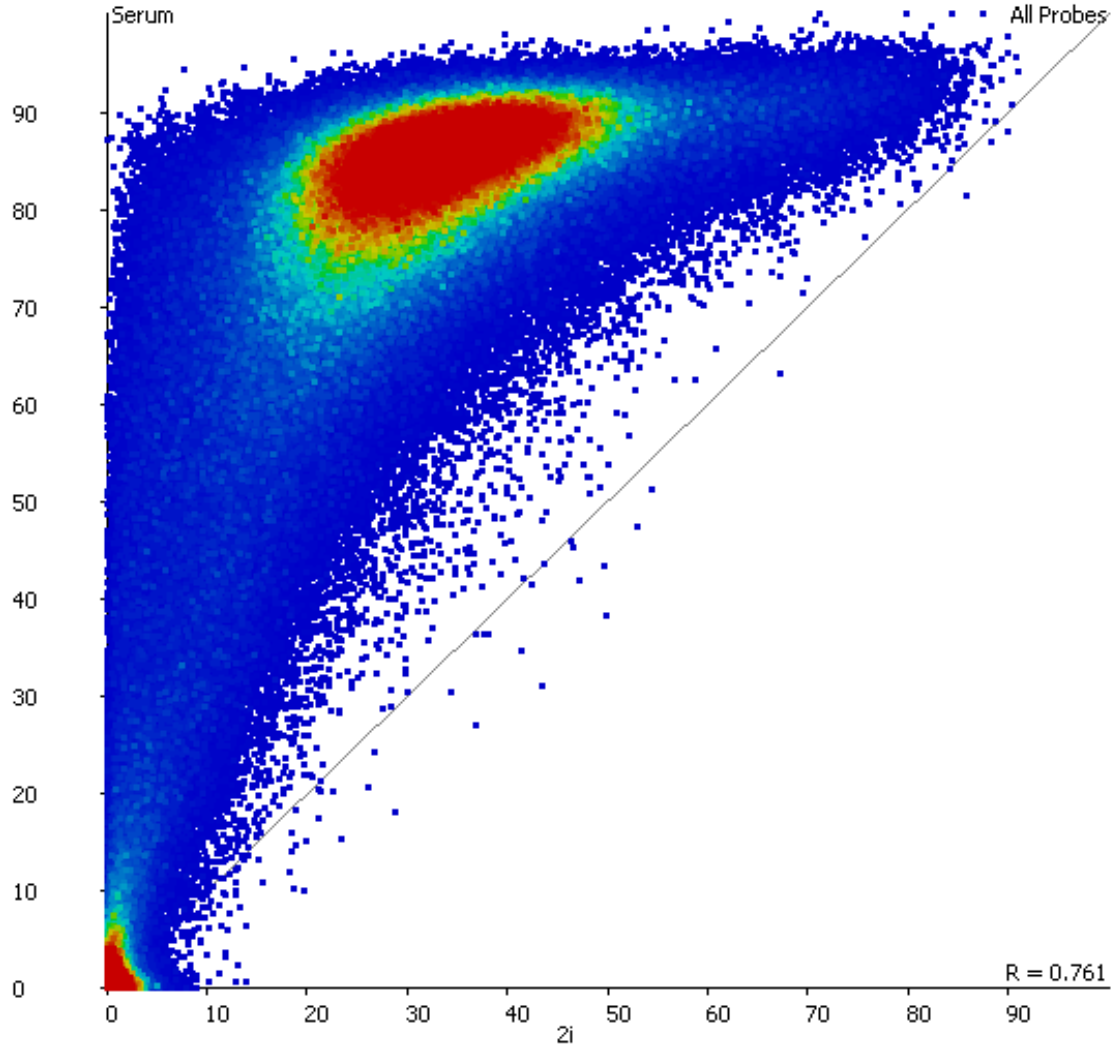


Sensible Questions



Which regions show a significant change in methylation level between the two conditions?

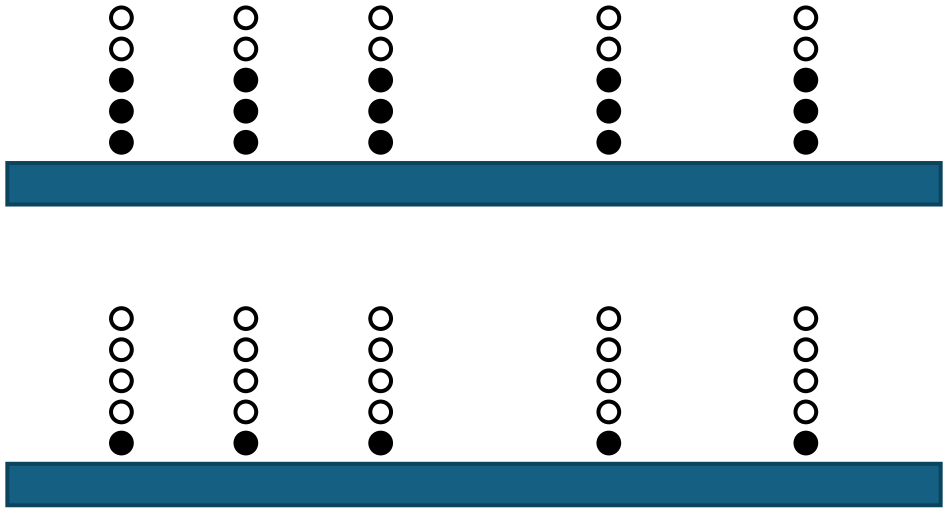
Sensible Questions



Which regions show a significant change in methylation level between the two conditions?

Which regions show a change in methylation which is larger or smaller than the global change beteen the samples overall?

The problem of power…



- Ideally we'd like to analyse every Cytosine (CpG) individually
- It's unlikely we'll have enough data to identify significantly changing bases per base
- Analysing in windows provides higher counts and more power

Power Analysis

(Assuming a human genome with $p<0.05$ and power of detection of 0.8)

Window Size (# CpG cytosines)

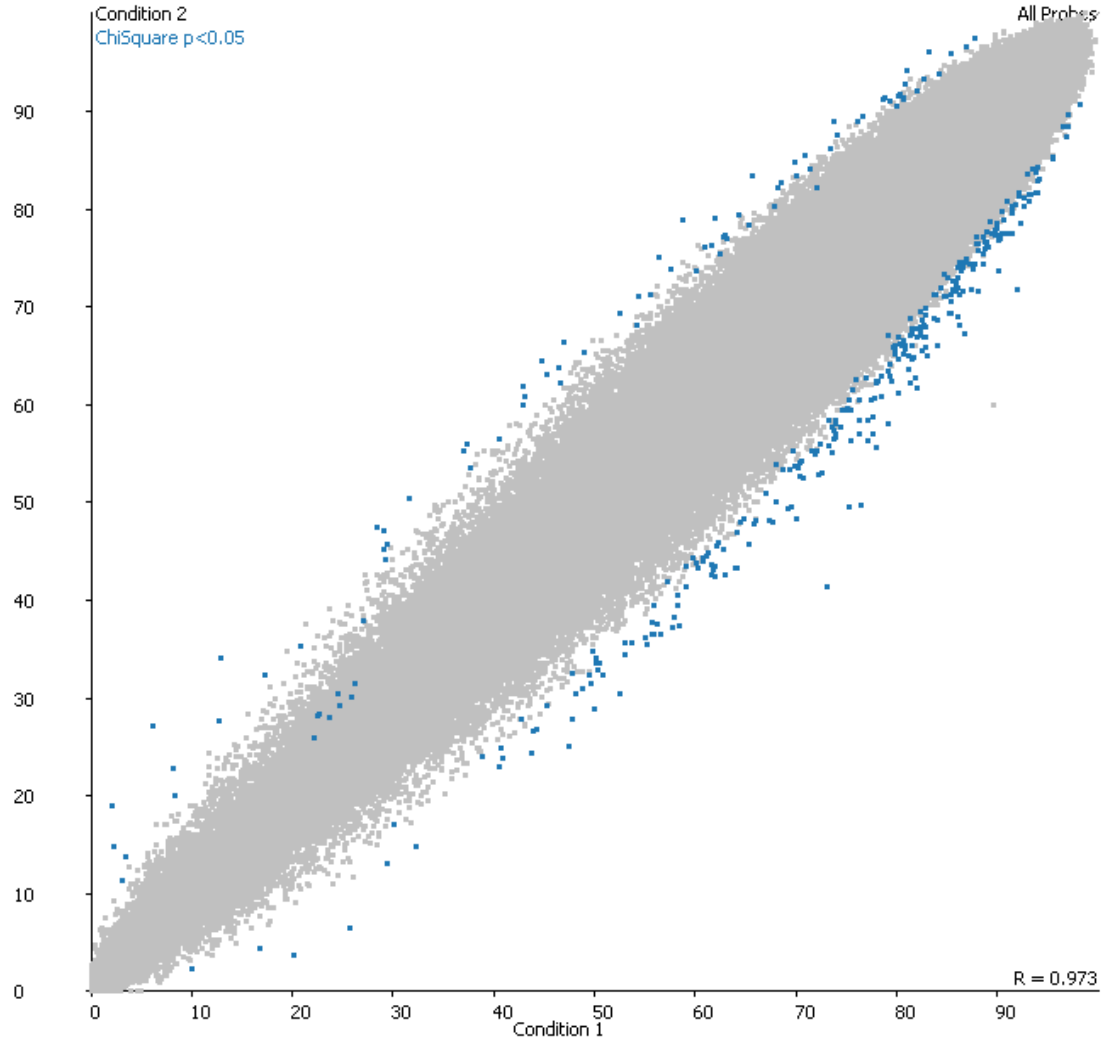
Absolute methylation change (from 80%)	1	10	25	50	100	200	500
1	158805	14212	5419	2609	1254	602	228
5	6794	608	232	112	54	26	10
10	1825	164	63	30	15	7	3
20	509	46	18	9	5	2	1
50	94	9	4	2	1	1	1

Required Fold Genome Coverage

Count Based Statistics

Contingency Statistics

- Count Based Stats
- Unreplicated design
  - Chi-Square
  - Fisher's Exact
- Replicated Design
  - Logistic Regression
  - Binomial Linear Model



Contingency Stats Example

Sample	**Unmeth Count**	**Meth Count**
Control	51	48
Treated	32	58

51% Methylated

36% Methylated

```
fisher.test(
  matrix(
    c(51,32,48,58),
    nrow = 2
  )
)
```

```
        Fisher's Exact Test for Count Data

data:  matrix(c(51, 32, 48, 58), nrow = 2)
p-value = 0.02897
```

*NB: This takes no account of running multiple tests

Samples with Replicates

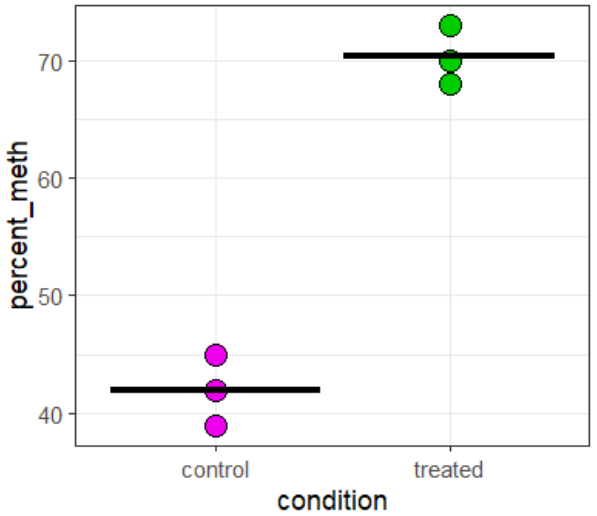
```
glm(
   cbind(M, U) ~ condition,
   data = dat,
   family = binomial()
) -> fit

summary(fit)
```

```
Call:
glm(formula = cbind(M, U) ~ condition,
family = binomial(), data = dat)

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)       -0.3228     0.1170  -2.759  0.00579 **
conditiontreated   1.1860     0.1722   6.887 5.71e-12 ***
```

	condition	M	U
1	control	42	58
2	control	39	61
3	control	45	55
4	treated	70	30
5	treated	68	32
6	treated	73	27



F1000Research
F1000Research 2017, 6:2055 Last updated: 20 APR 2018

Check for updates

METHOD ARTICLE

Differential methylation analysis of reduced representation bisulfite sequencing experiments using edgeR [version 1; referees: 2 approved, 1 approved with reservations]

Yunshun Chen[1,2], Bhupinder Pal[1,2], Jane E. Visvader[1,2], Gordon K. Smyth[2,3]

[1]Department of Medical Biology, The University of Melbourne, Melbourne, VIC, 3010, Australia
[2]The Walter and Eliza Hall Institute of Medical Research, Parkville, VIC, 3052, Australia
[3]School of Mathematics and Statistics, The University of Melbourne, Melbourne, VIC, 3010, Australia

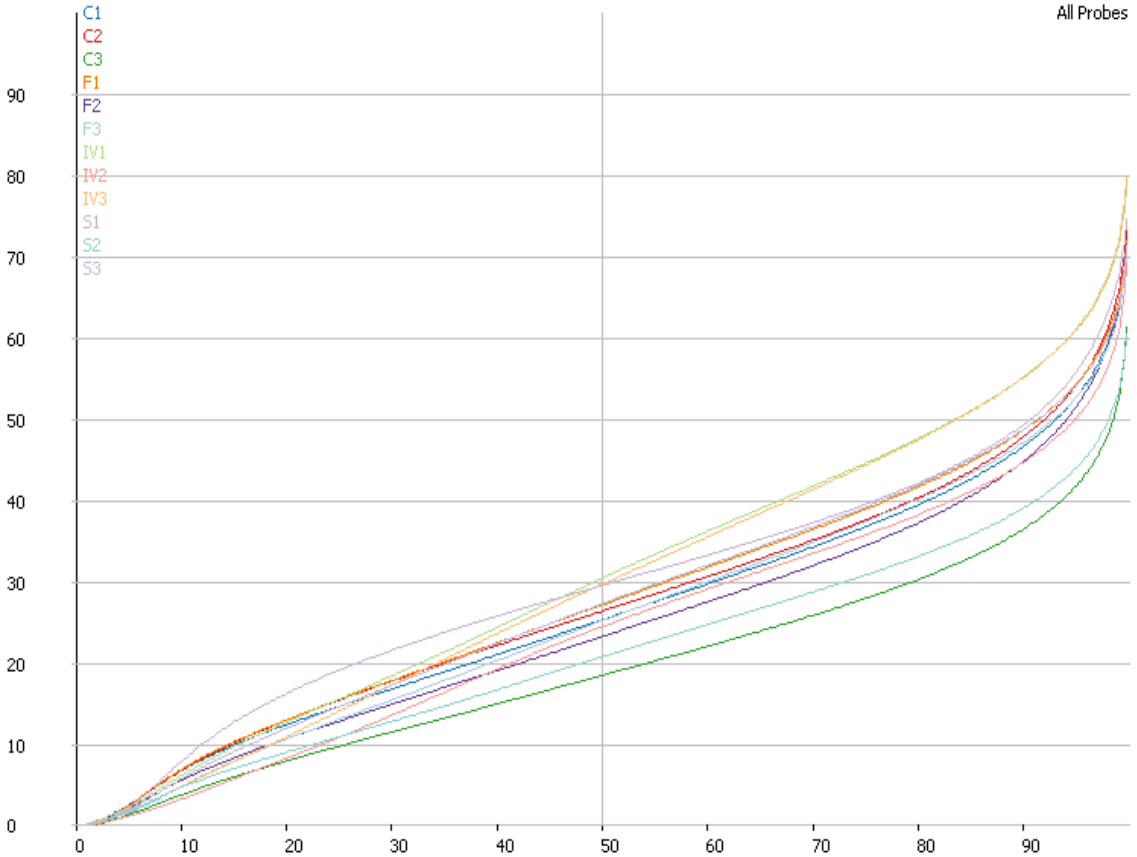
Continuous Statistics

Statistics on Continuous Values

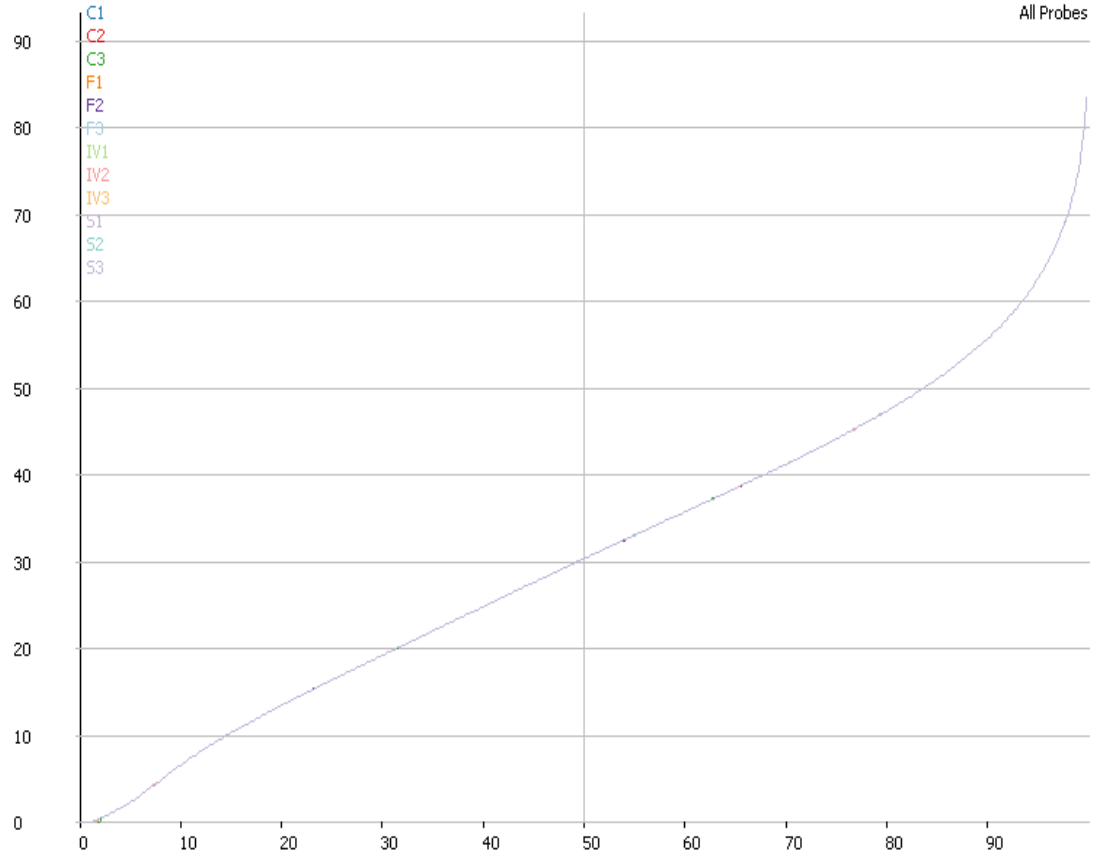
- Simple Models
- Must have replicates
  - T-Test
  - ANOVA
- Can’t use technical significance
- Less Powerful than count based
- Make assumptions about data behaviour

- Continuous value linear models
- More flexible – allow more complex designs
- LIMMA – allows better variance estimates
- Still less powerful than count stats

Global vs Local Effects



Original Methylation Levels



After Quantile Normalisation

Reverse counting

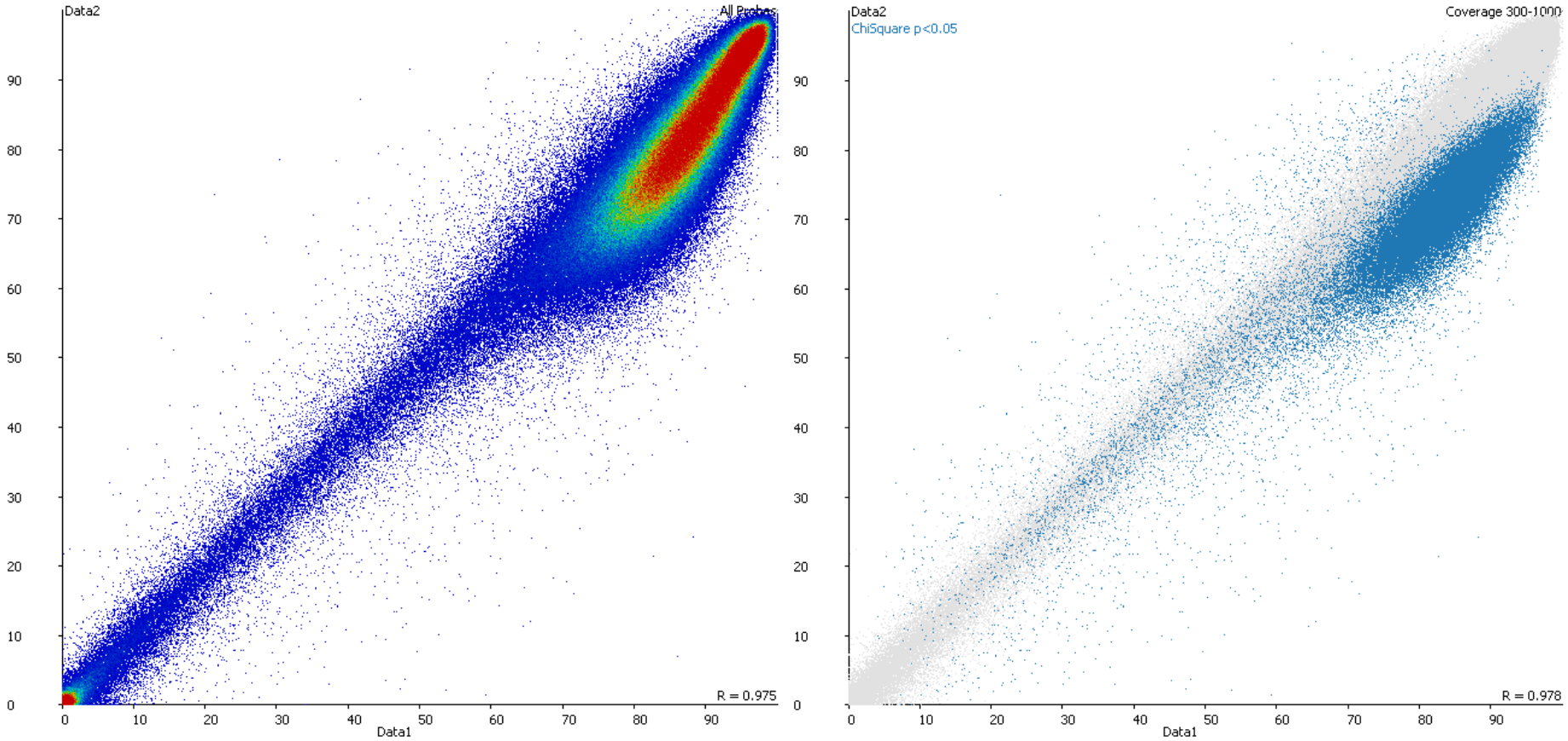
- Some packages offer a conversion from normalised methylation back to counts

True observations: Meth = **20** Umeth=**30** (40% meth)
Corrected % methylation = 50%
Reversed counts: Meth = **25** Unmeth=**25**

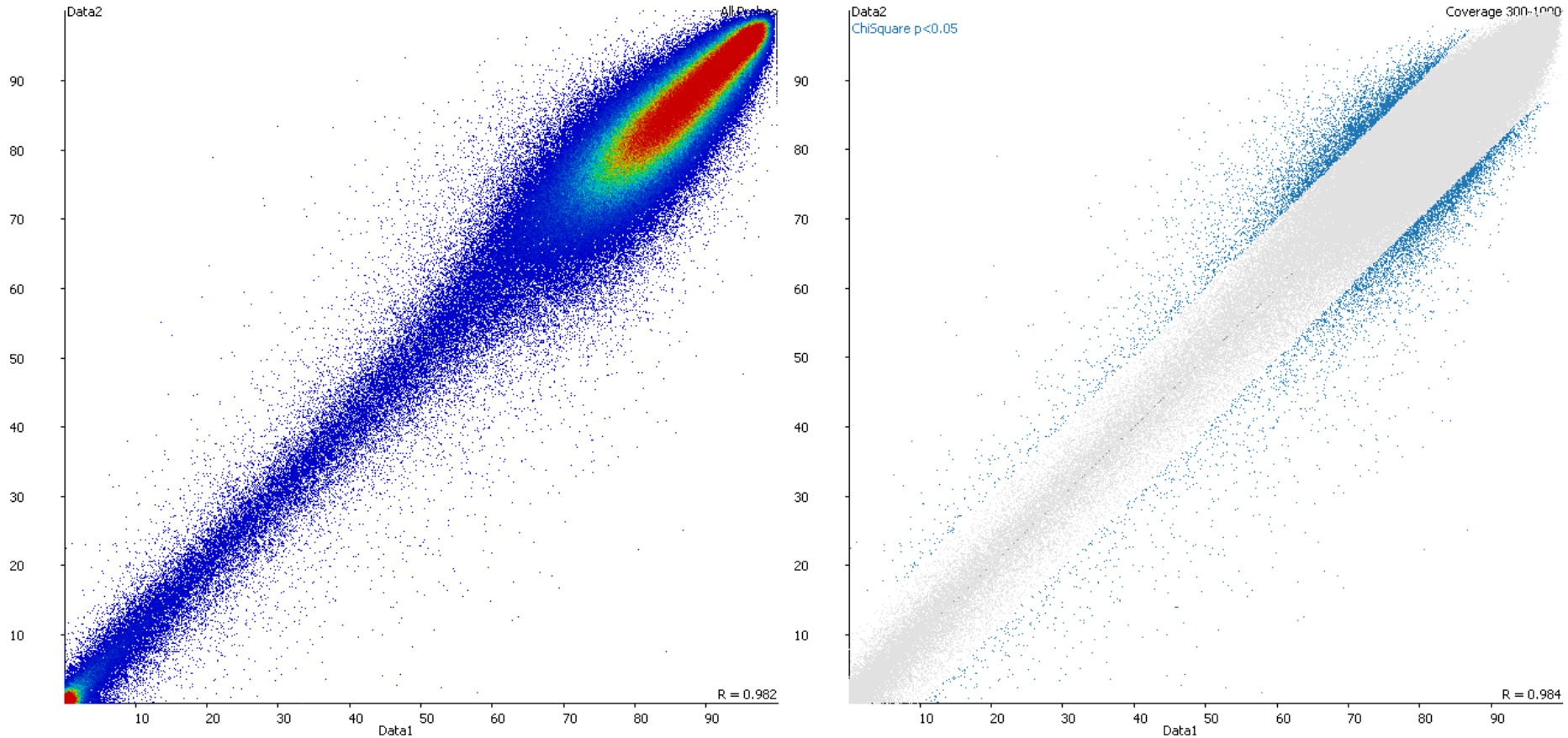
- Allows count based statistics – regains the lost power from normalisation

- Retains information about noise from the true observation level

Correction can have a big effect



Correction can have a big effect



Exercise

Differential Methylation Statistics in SeqMonk