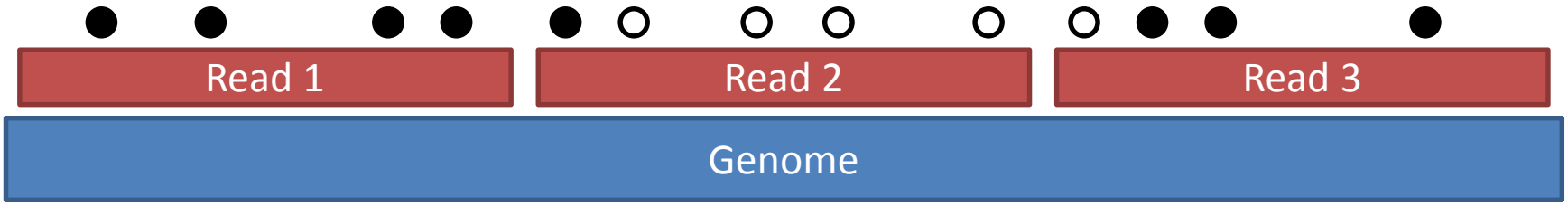
Visualising and Exploring BS-Seq Data

Simon Andrews
simon.andrews@babraham.ac.uk
@simon_Andrews

V2017-01

Babraham Bioinformatics

Starting Data




```
L001_bismark_bt2_pe.deduplicated.bam

CHG_OB_L001_bismark_bt2_pe.deduplicated.txt.gz
CHG_OT_L001_bismark_bt2_pe.deduplicated.txt.gz
CHH_OB_L001_bismark_bt2_pe.deduplicated.txt.gz
CHH_OT_L001_bismark_bt2_pe.deduplicated.txt.gz
CpG_OB_L001_bismark_bt2_pe.deduplicated.txt.gz
CpG_OT_L001_bismark_bt2_pe.deduplicated.txt.gz

L001_bismark_bt2_pe.deduplicated.cov.gz
```

Babraham Bioinformatics

Which data to use?

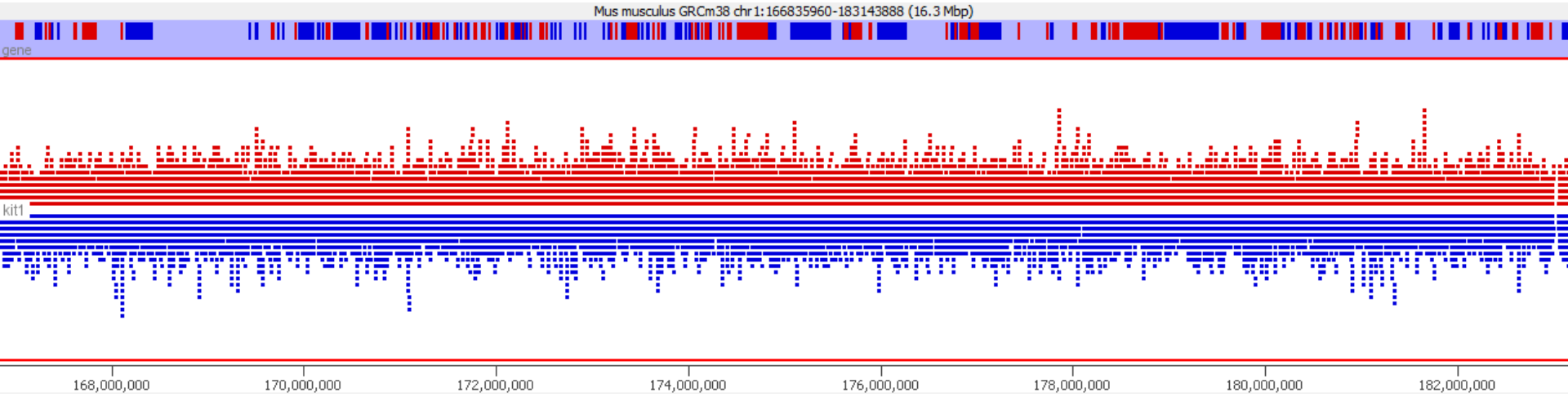
- Methylation contexts
  - CpG: Only generally relevant context for mammals
  - CHG: Only known to be relevant in plants
  - CHH: Generally unmethylated

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

Babraham Bioinformatics

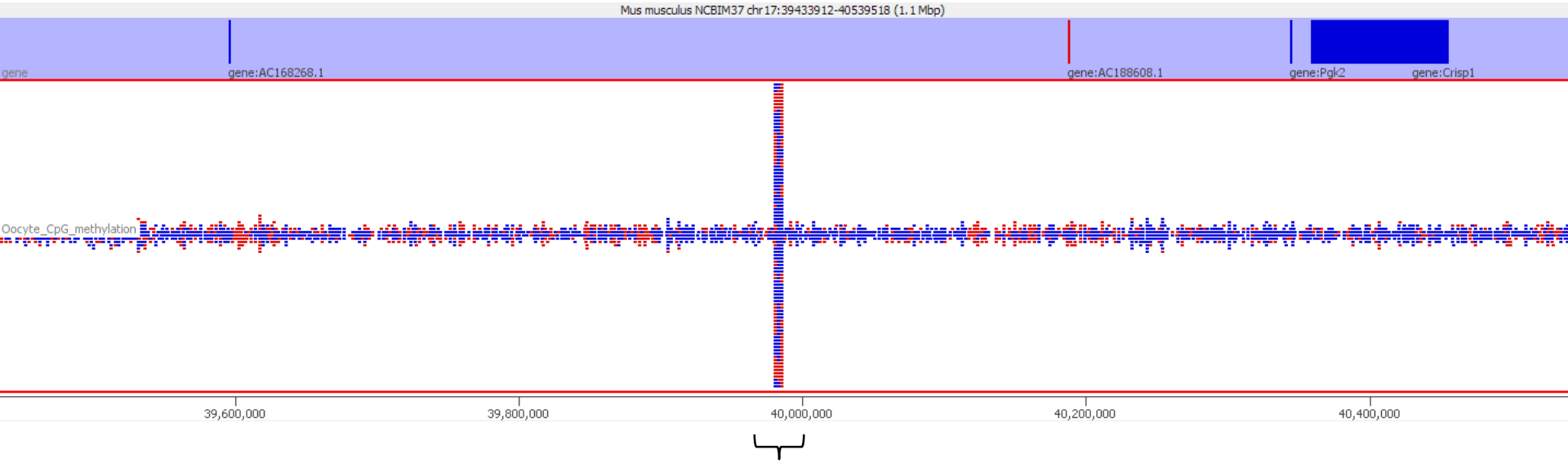
Coverage Expectations

- Whole genome covered equally
- Both strands present equally



Babraham Bioinformatics

Coverage Outliers



Around 600x average genome density

Babraham Bioinformatics

Coverage Outliers

Babraham Bioinformatics

Coverage Outliers

- Normally the result of mis-mapping repetitive sequences not in the genome assembly
- Centromeric / temomeric sequences are common
- Can be a significant proportion of all data
- Can throw off calculations of overall methylation
- Should be removed

Babraham Bioinformatics

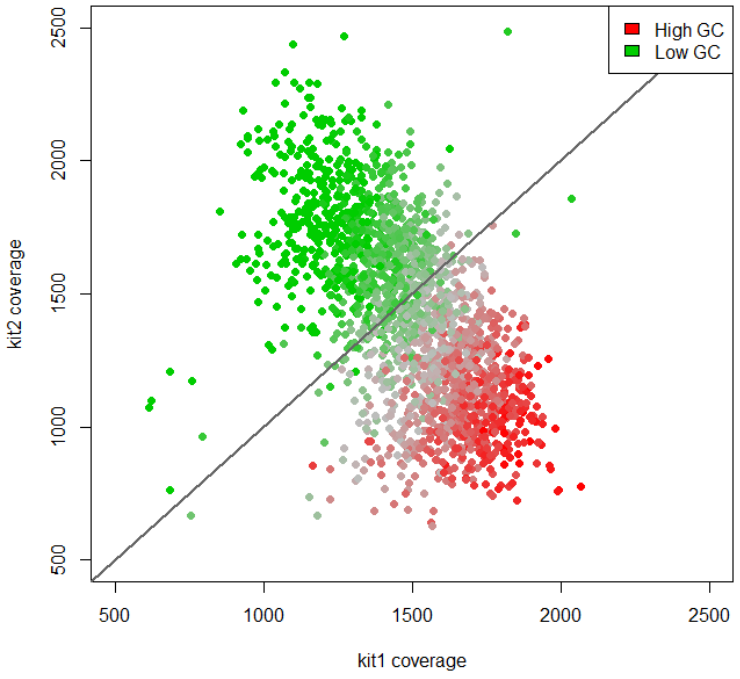
Coverage Biases

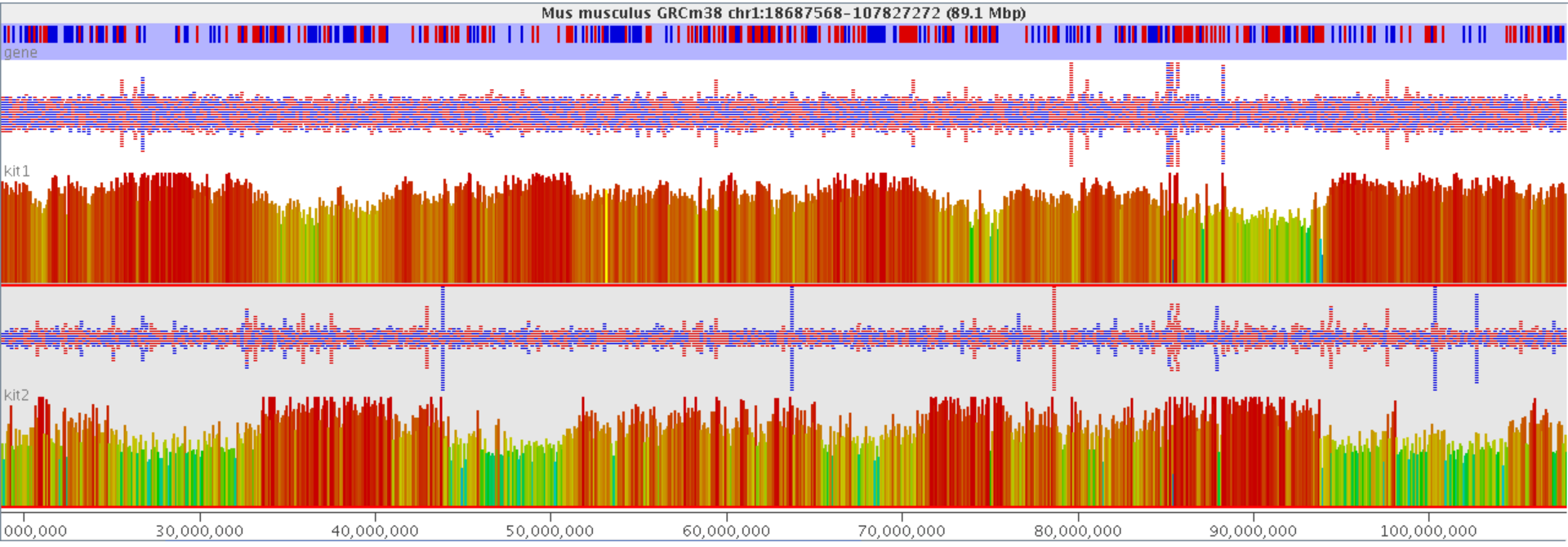
- BS-Seq read level coverage should be flat
- Different kits have different biases
- Most bias is based on composition (normally GC)
- Link between methylation and GC content can cause problems.

Babraham Bioinformatics

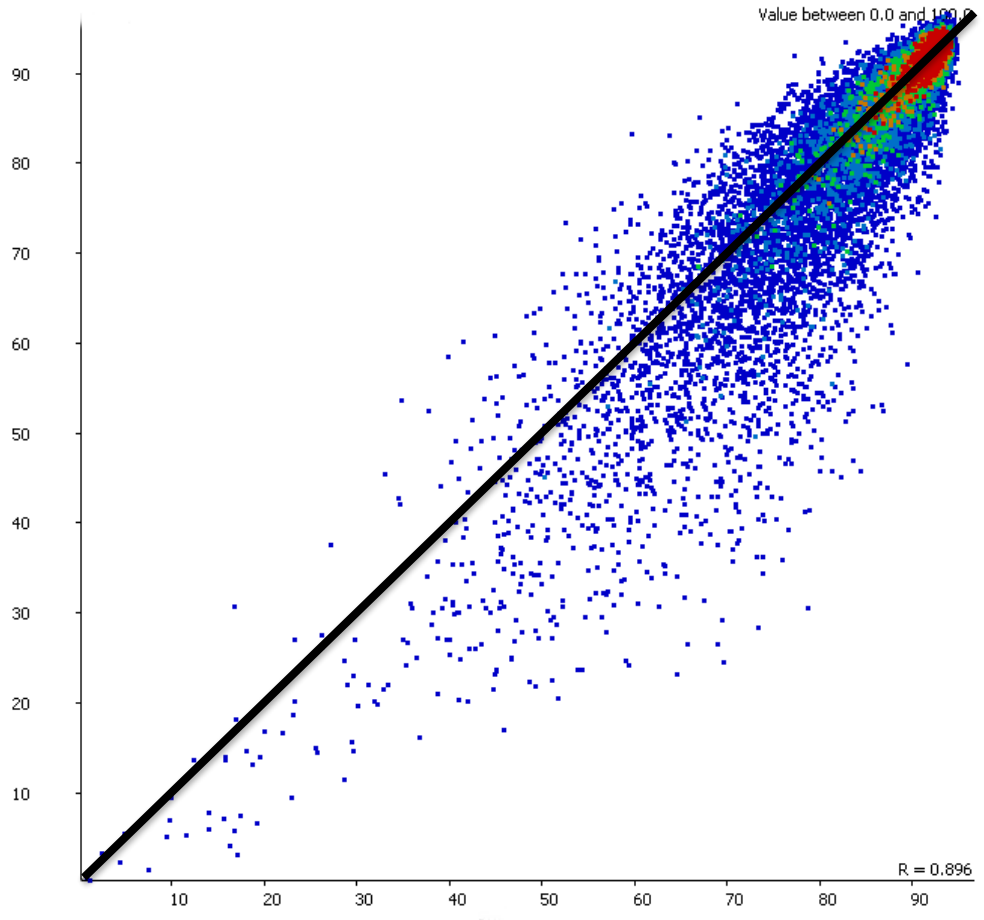
Coverage Bias

GC Content is most likely but others could exist



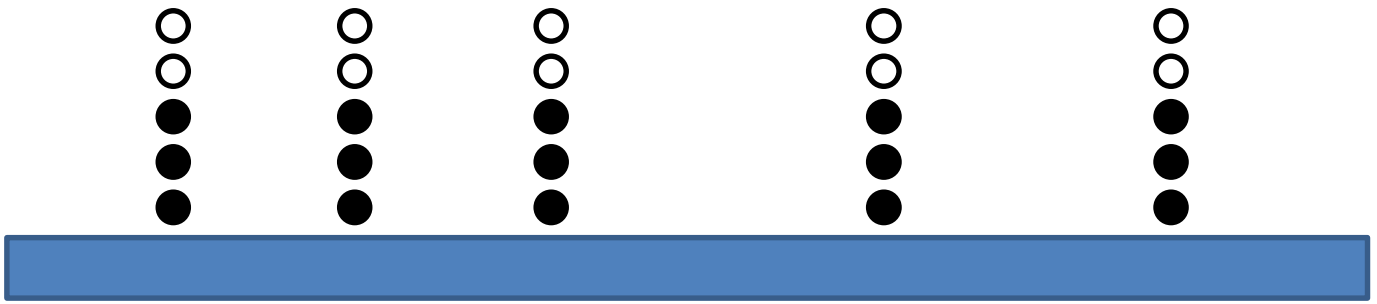


Coverage bias can lead to Methylation bias



Babraham Bioinformatics

Quantitating Methylation



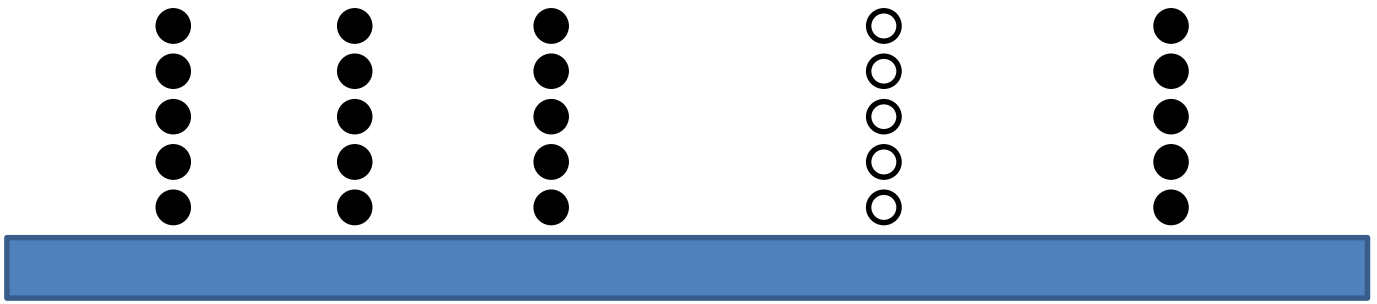
Total methylated calls = 15

Total unmethylated calls = 10

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

Babraham Bioinformatics

Quantitating Methylation

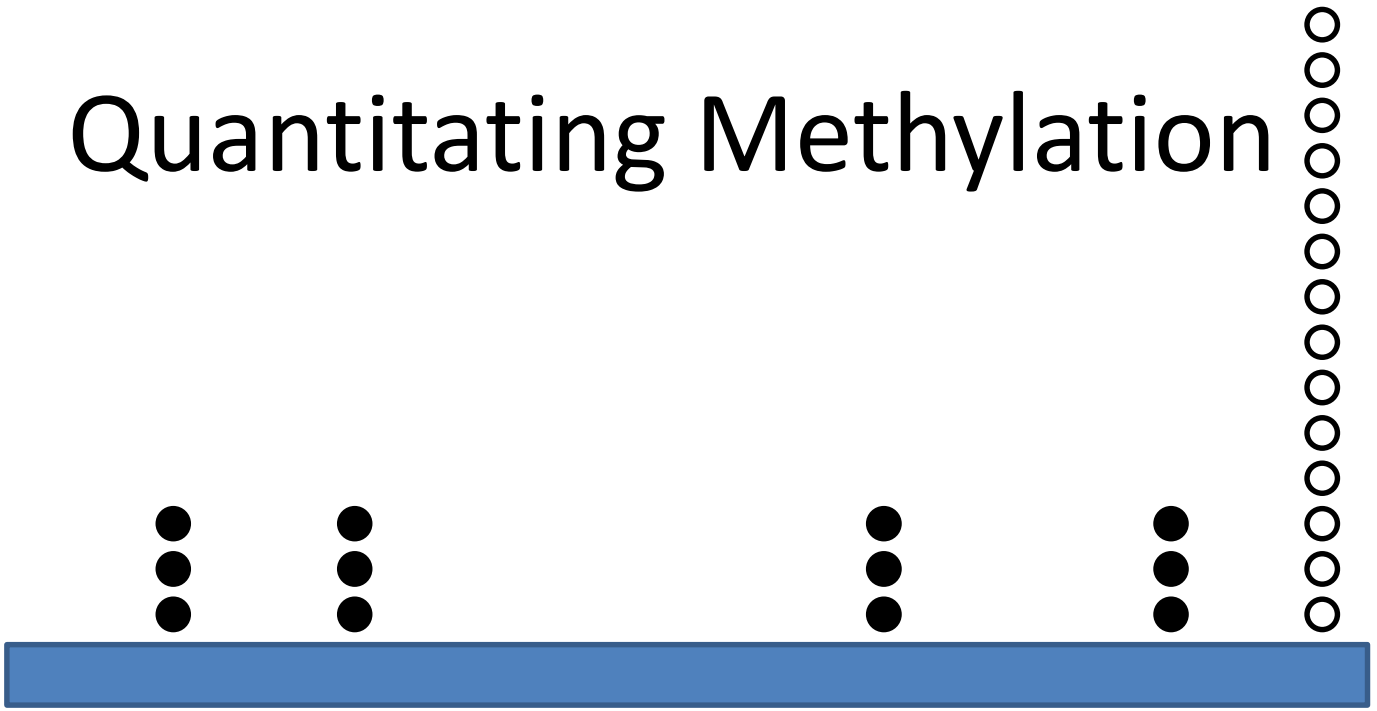


Total methylated calls = 20
Total unmethylated calls = 5

Methylation level = (20/(20+5))*100 = 80%

Babraham Bioinformatics

Quantitating Methylation



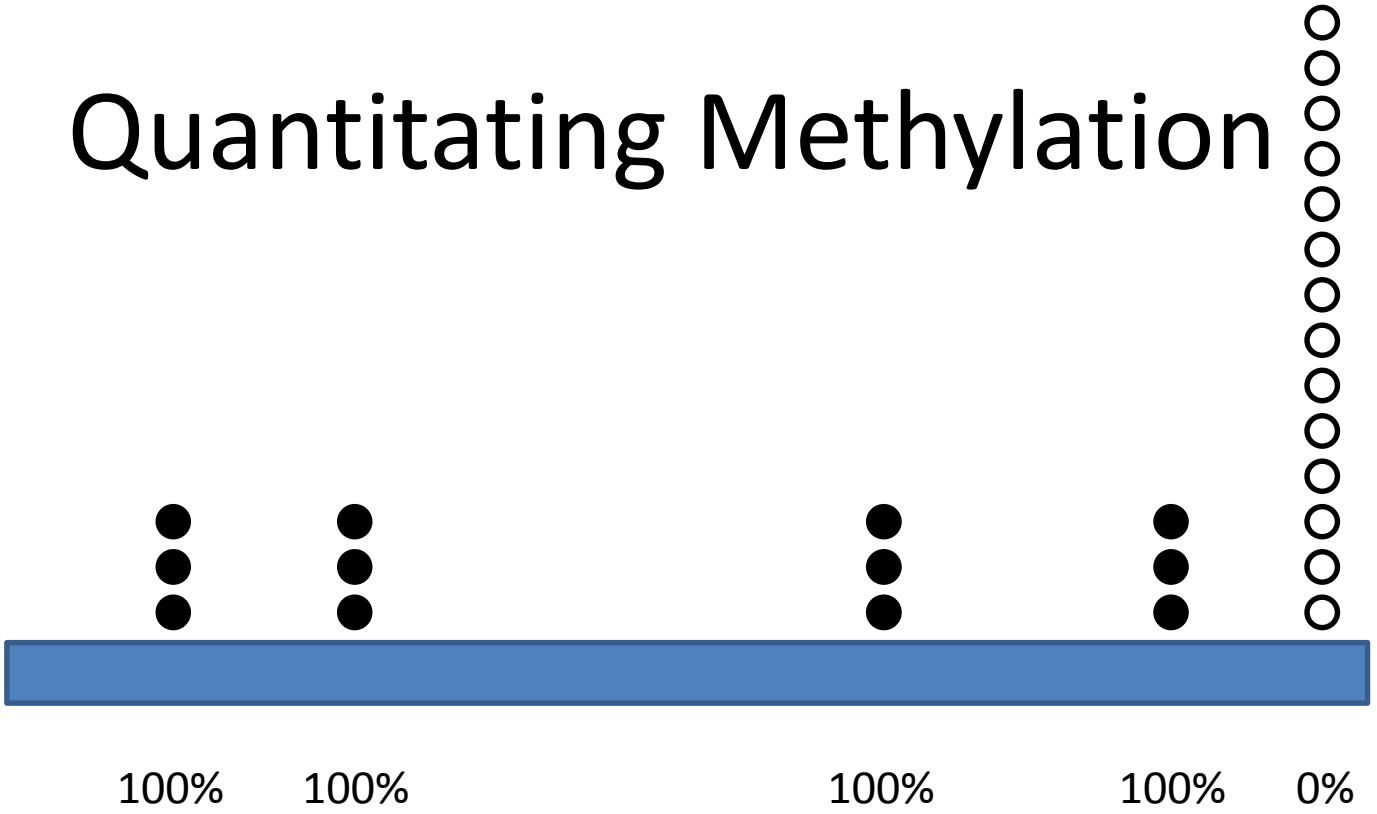
Total methylated calls = 12

Total unmethylated calls = 14

Methylation level = (12/(12+14))*100 = 46%?

Babraham Bioinformatics

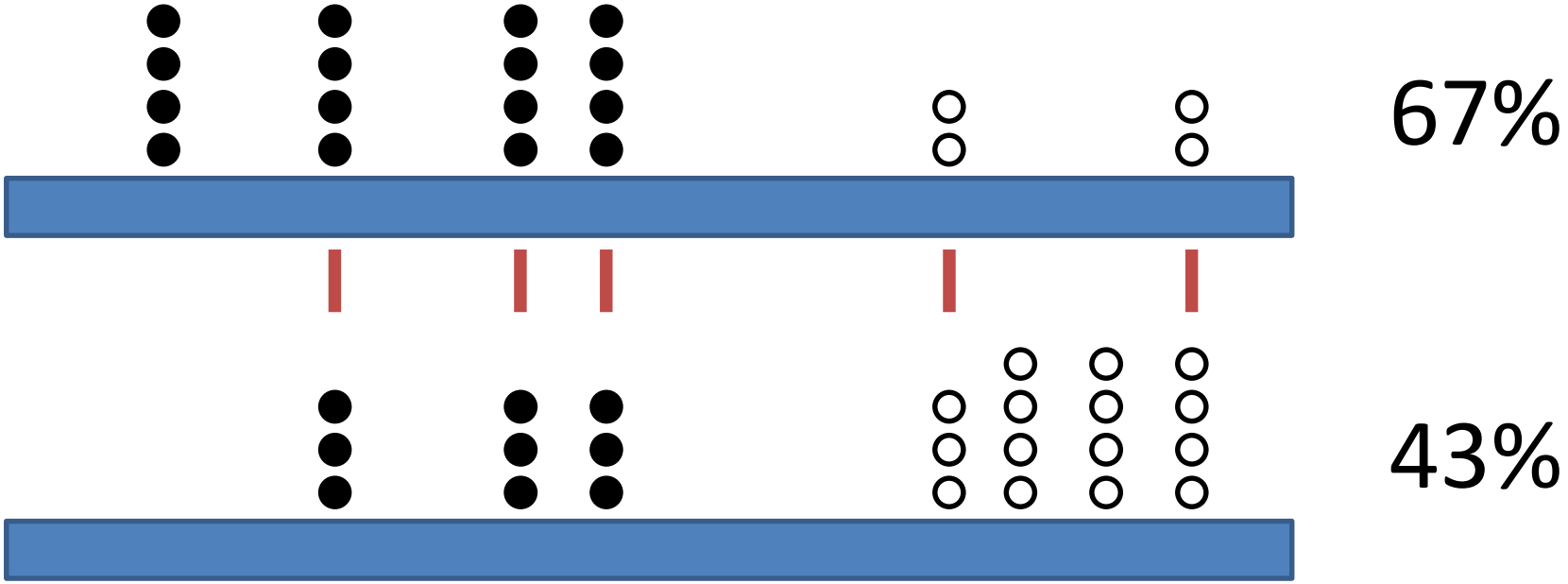
Quantitating Methylation



Methylation level = (100+100+100+100+0)/5 = 80%

Babraham Bioinformatics

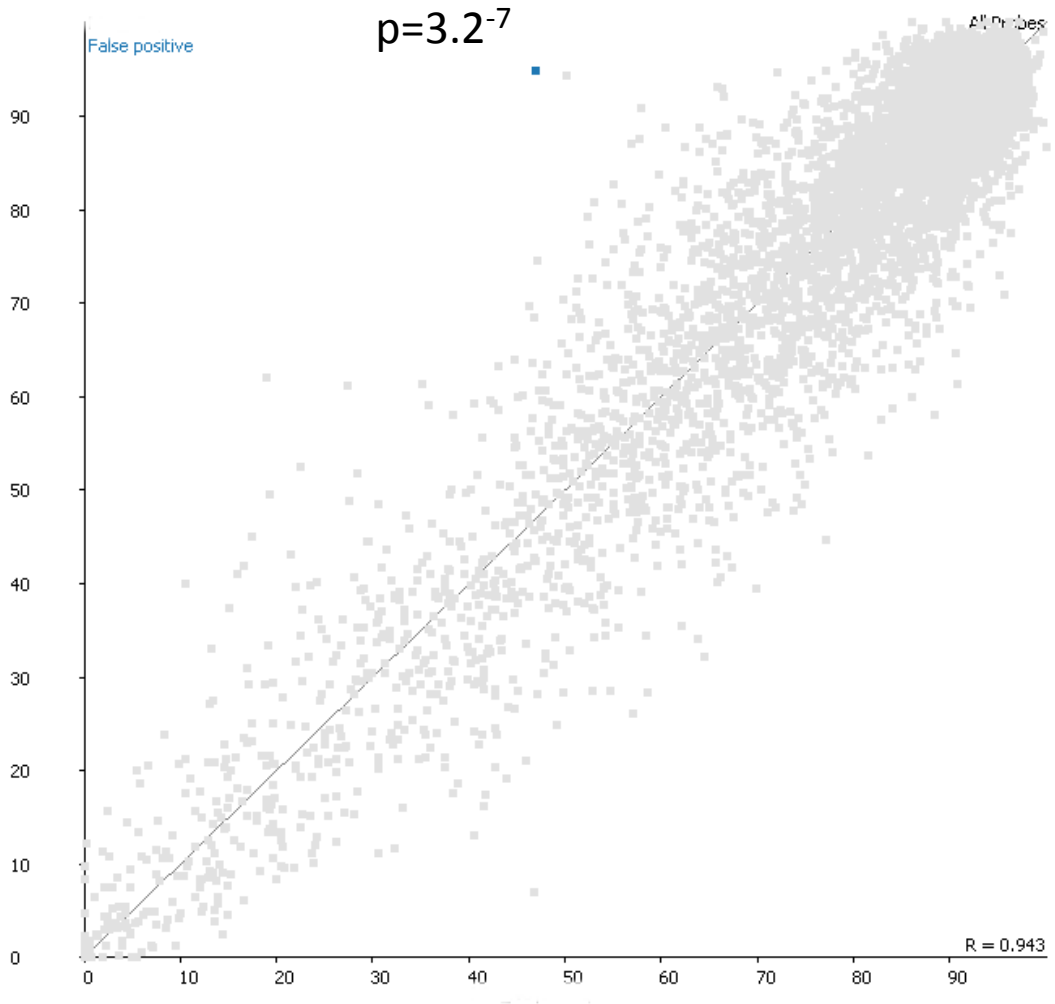
Quantitating Methylation



Common = 60% in both

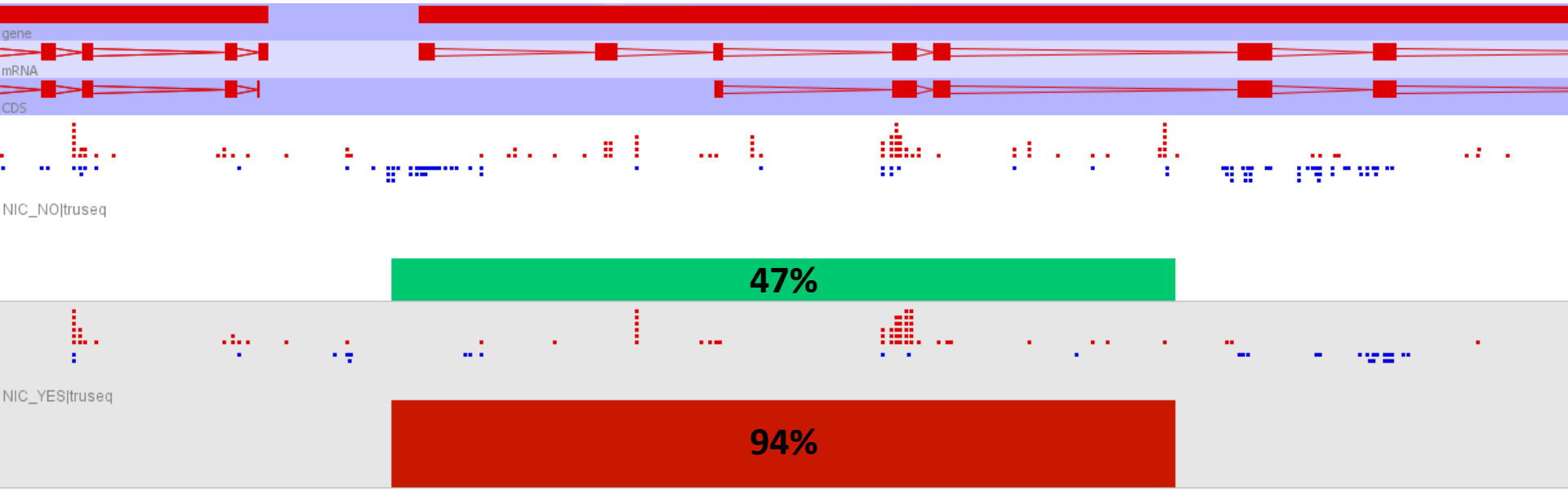
Babraham Bioinformatics

Example



Babraham Bioinformatics

Example



Babraham Bioinformatics

More Complex Methods

- Factors to consider during quantitation
  - Surrounding methylation levels
    - Assume that methylation doesn't change over very short distances
  - Coverage of individual bases
    - Down-weight very low or high values
  - Density of CpGs
    - Apply the level to all bases within a region

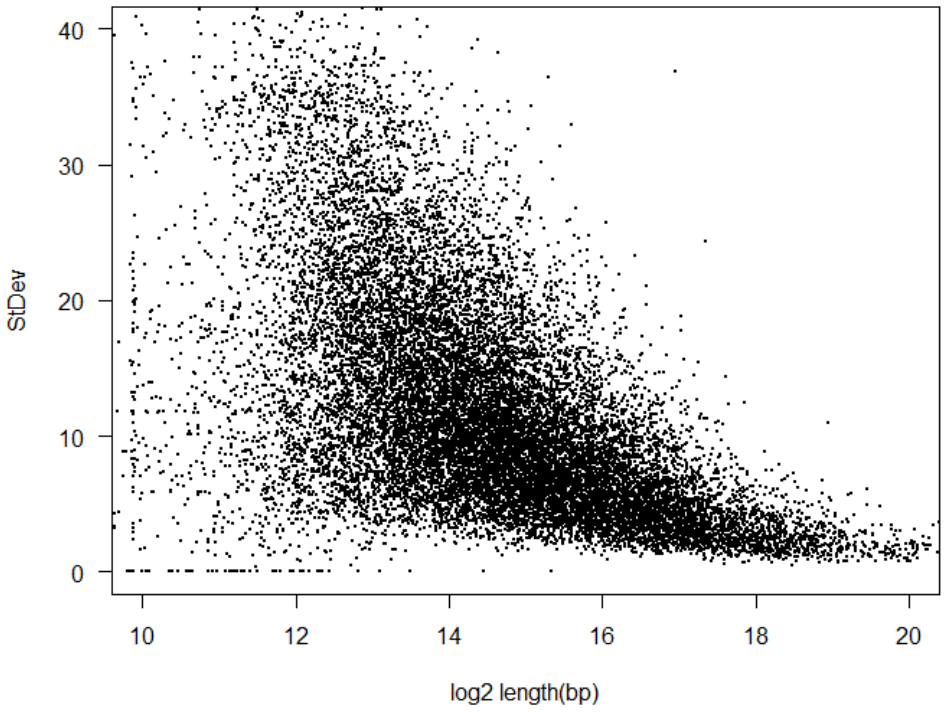
Babraham Bioinformatics

Where to make measures

- Per base
  - Very large number of measures
  - Poor accuracy for individual bases
- Unbiased windows
  - Tiled over whole genome
  - Need to decide how they will be defined
- Targeted regions
  - Which regions
  - What context

Babraham Bioinformatics

Making Comparable Measures



- Observation level correlates with stability.
- Want to try to have similar amounts of data in each measurement window.
- Equalises noise for visualisation and power for analysis.

Babraham Bioinformatics

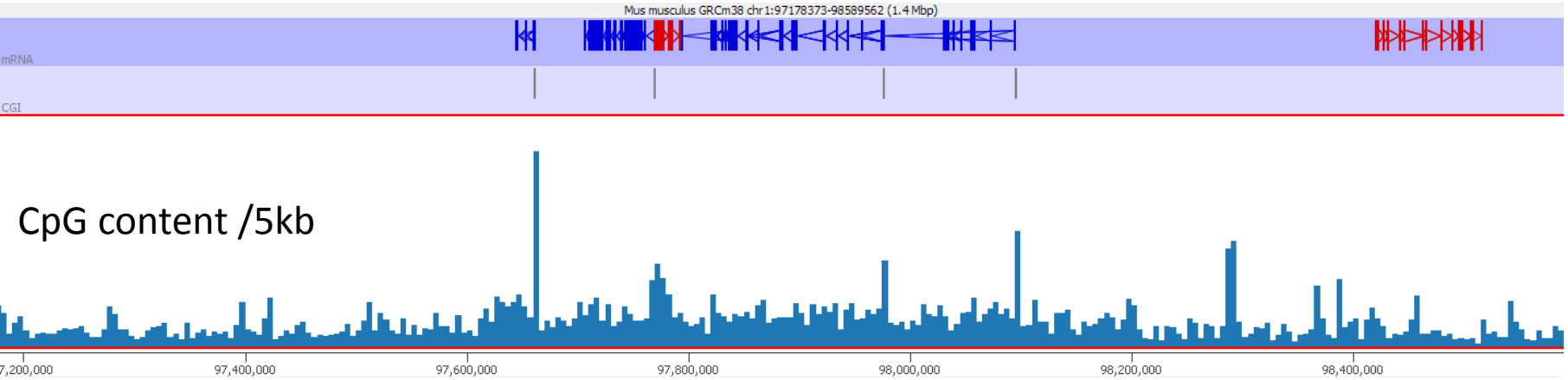
Unbiased analysis

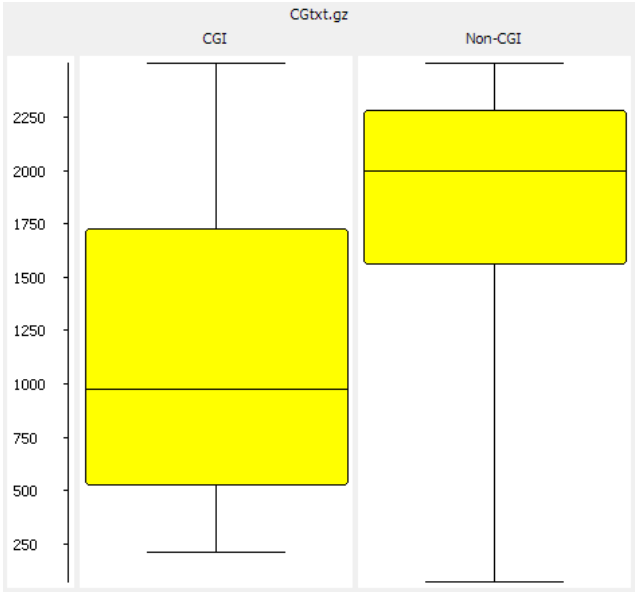
- What basis do you use for creating the windows?
  - Fixed size?
    - Uneven CpG distribution can be problematic
  - Fixed CpG count?
    - Different resolution

- Need to tailor the window sizes to the data density
- Can relate windows to features later

Babraham Bioinformatics

Fixed size vs Fixed CpG





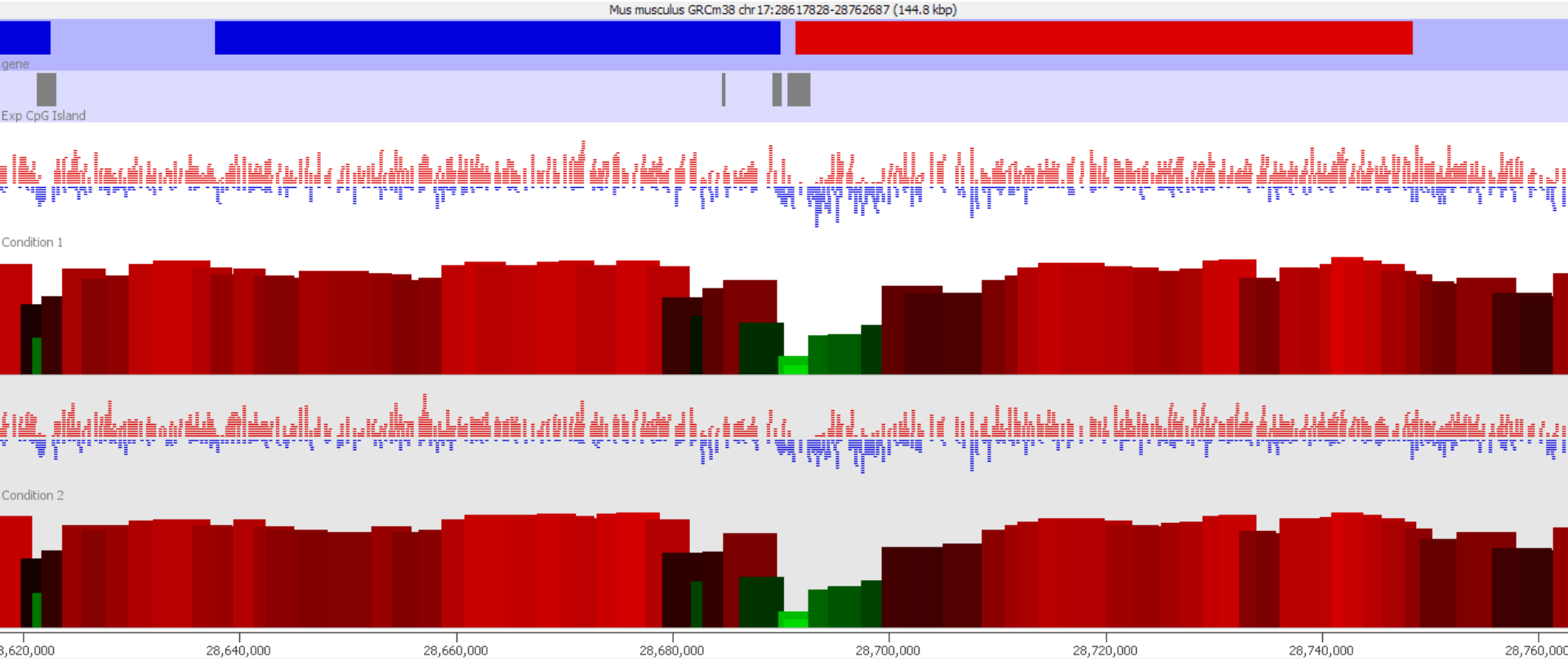
50 CpG window lengths

Targeted Quantitation

- Measure over features
  - CpG islands
    - Be careful where you get your locations
    - Try to fix sizes
  - Promoters
    - Should probably split into CpG island and non-CpG island
    - Try to fix sizes
  - Gene bodies
    - Filter by biotype to remove small RNA genes?

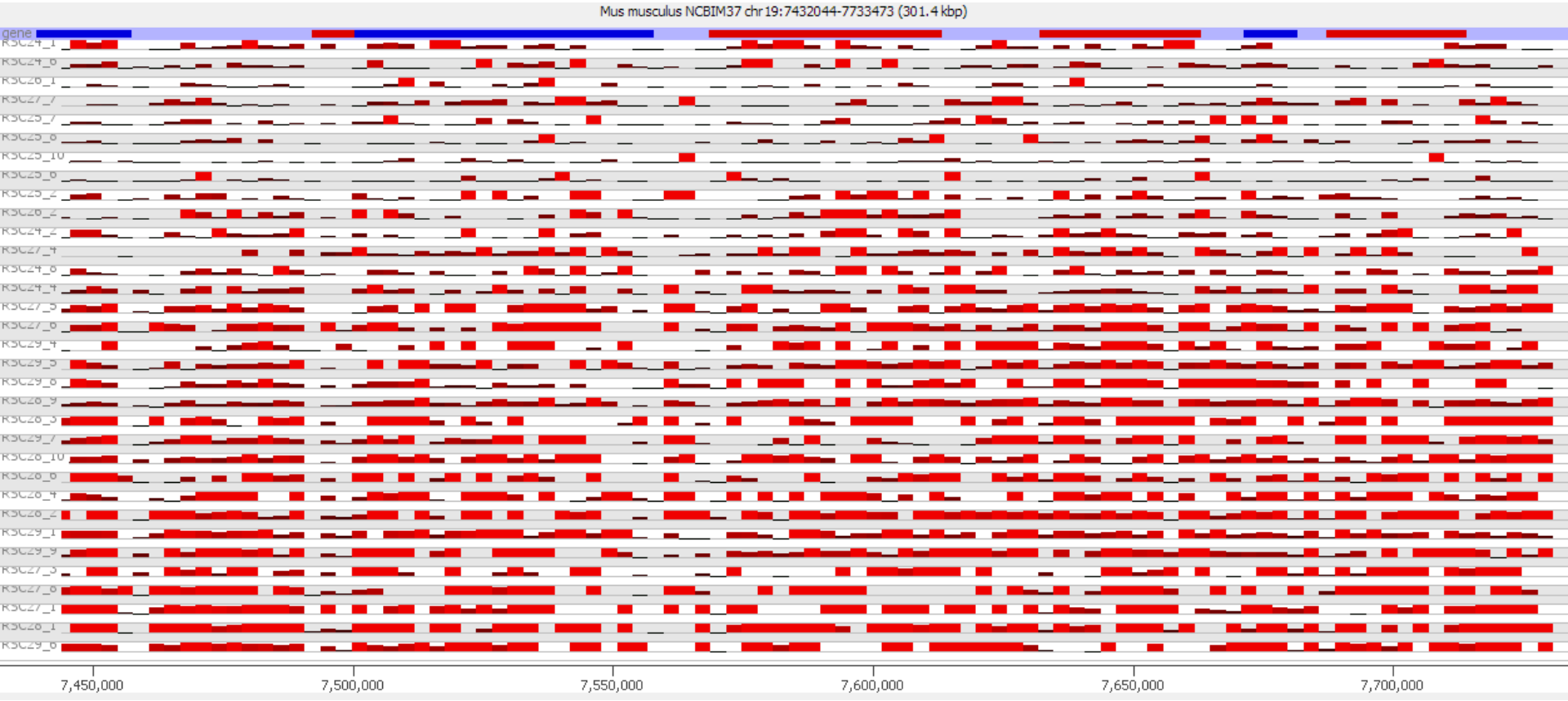
Babraham Bioinformatics

Viewing quantitated methylation



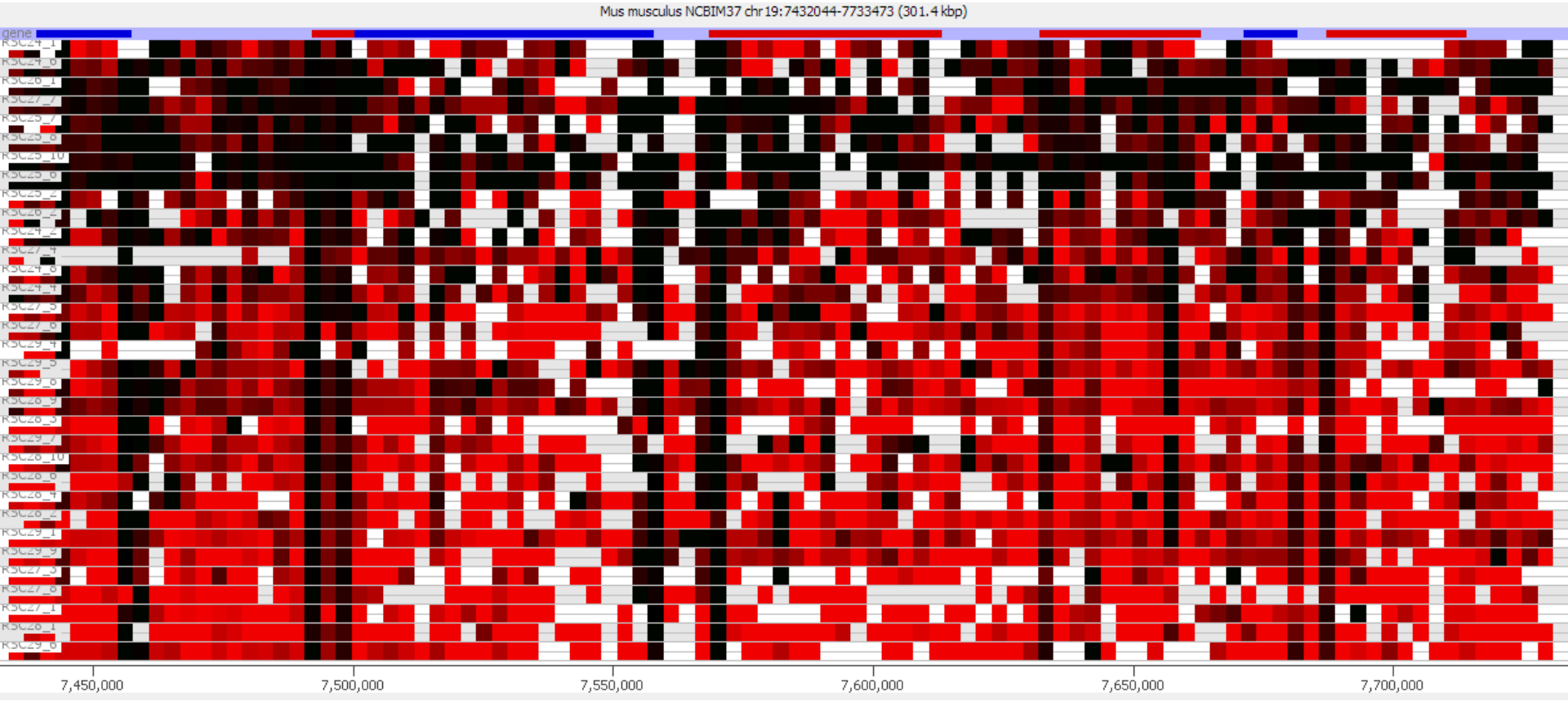
Babraham Bioinformatics

Large sample numbers



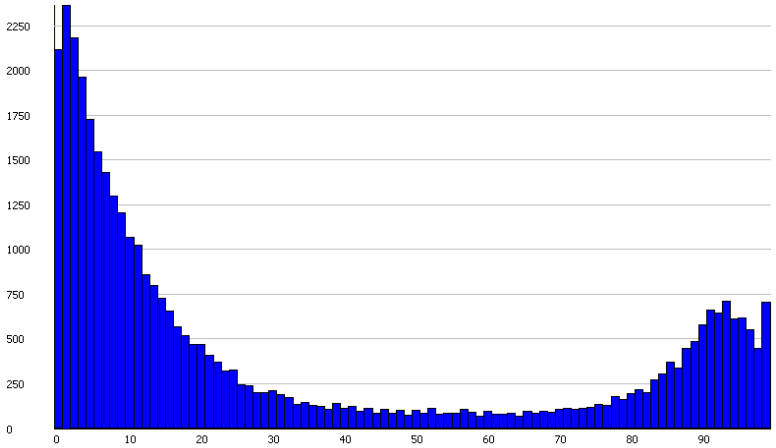
Babraham Bioinformatics

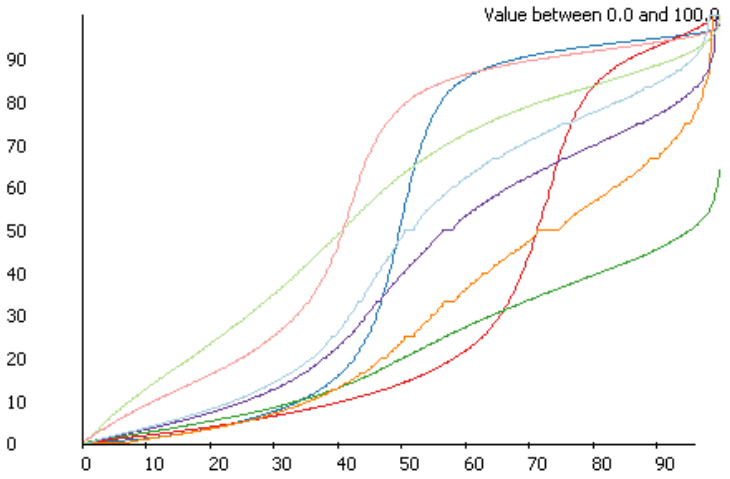
Large sample numbers

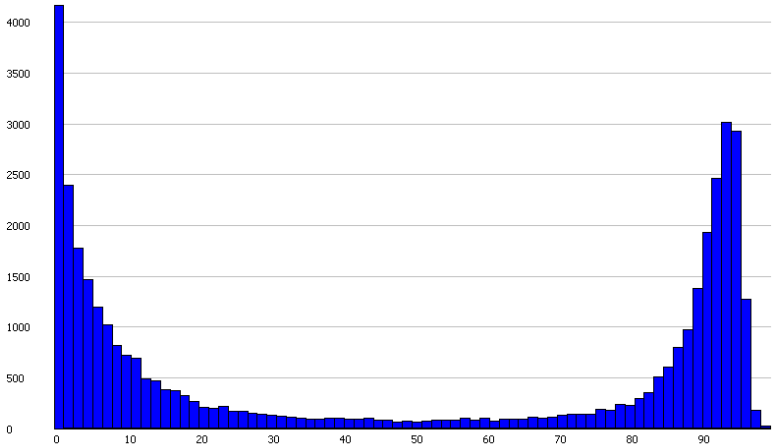


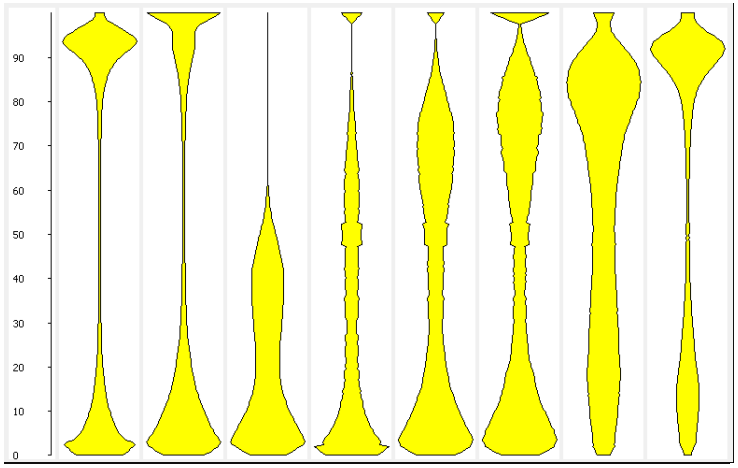
Babraham Bioinformatics

Viewing and Comparing Distributions



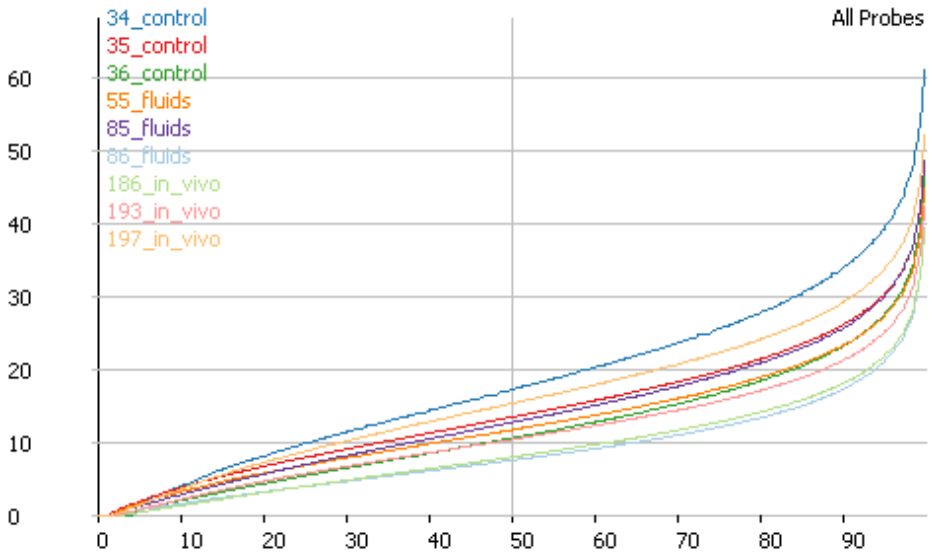


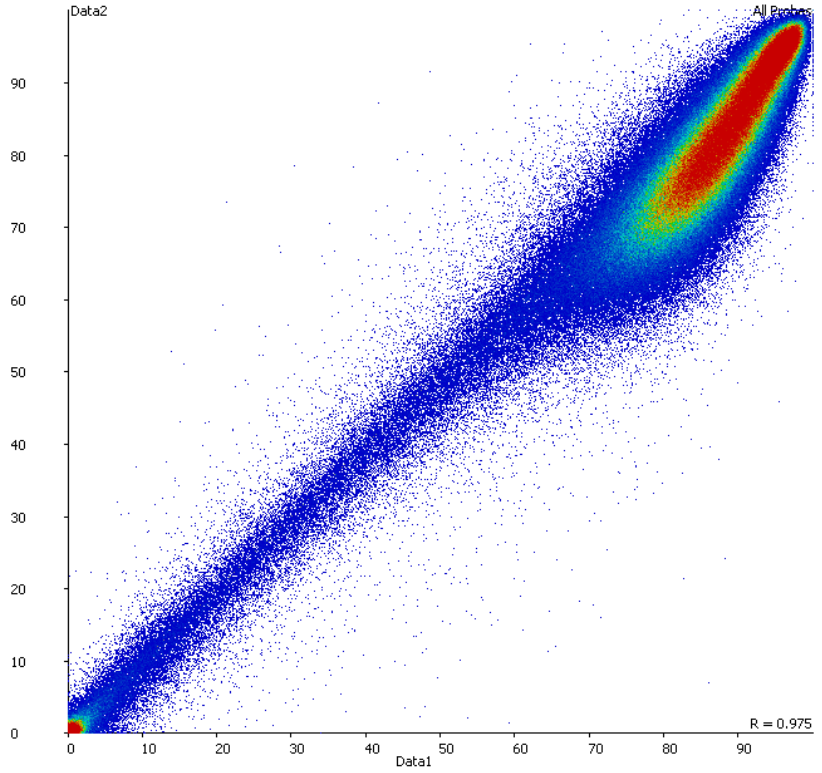




Babraham Bioinformatics

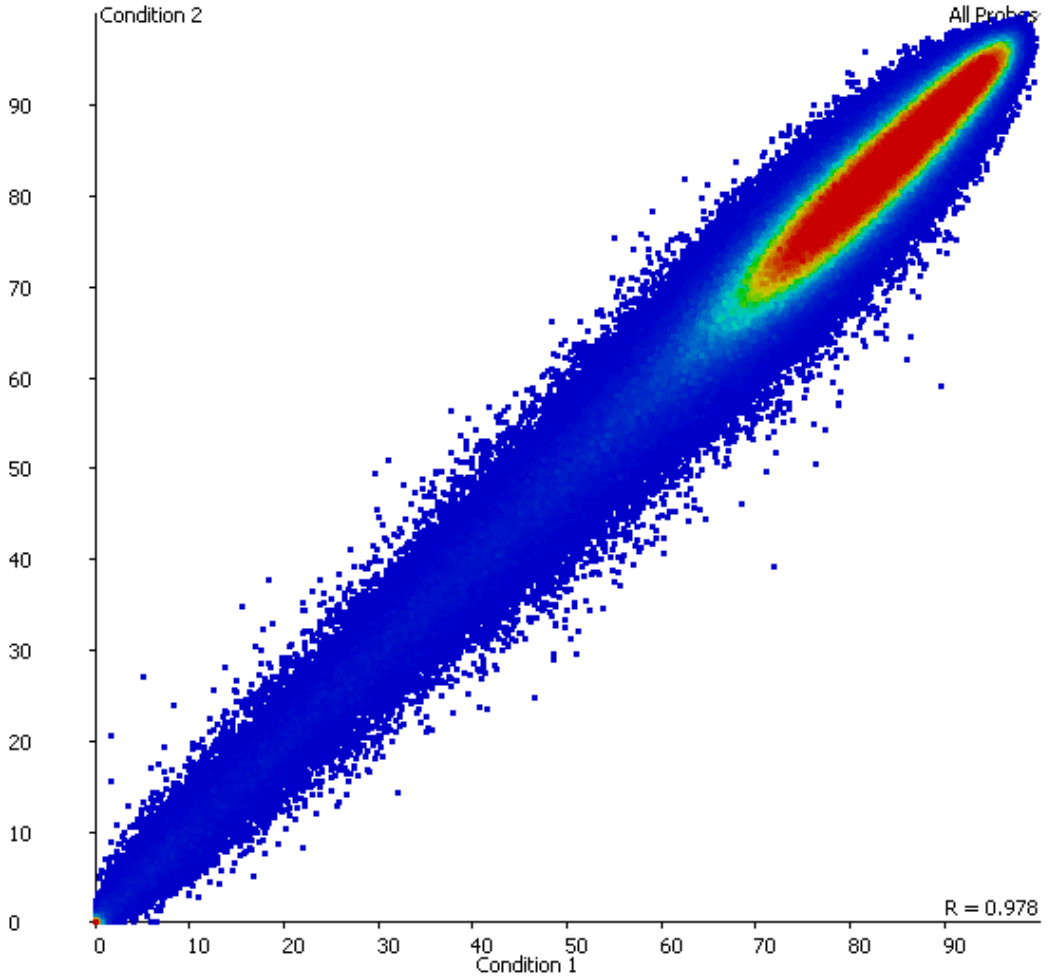
Normalising Methylation Values





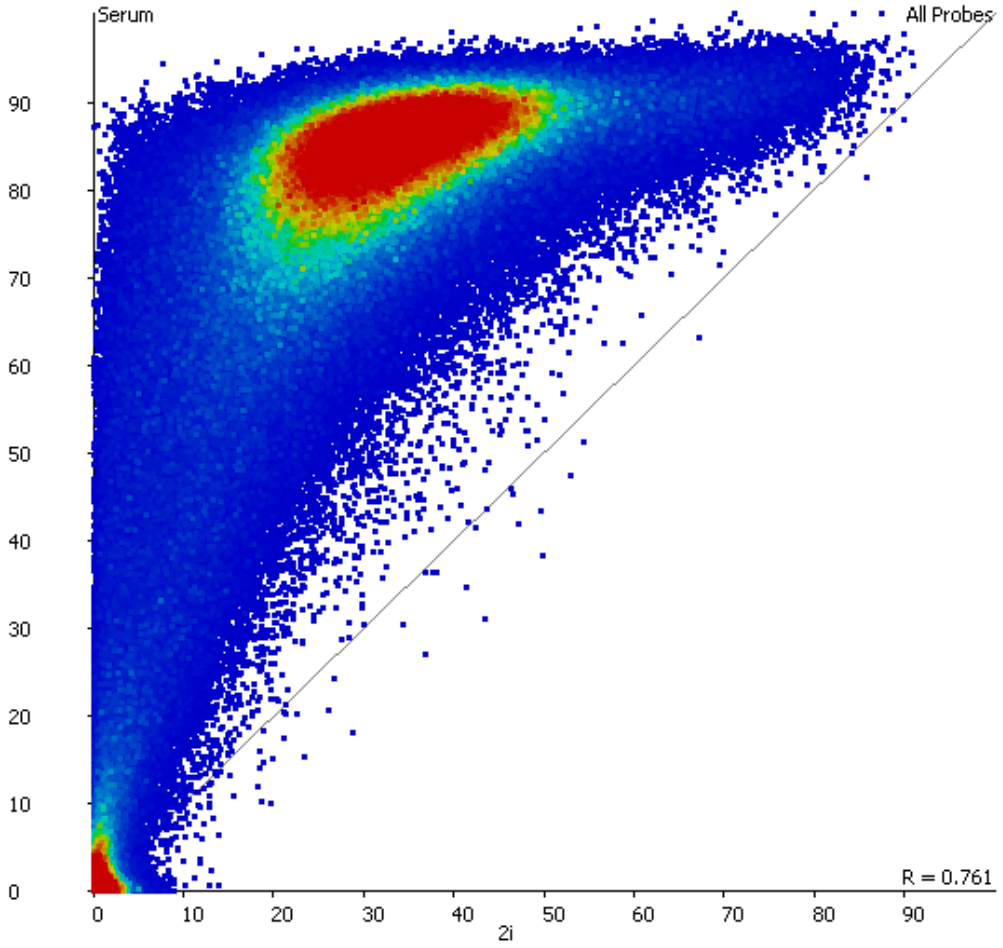
Babraham Bioinformatics

Viewing comparisons



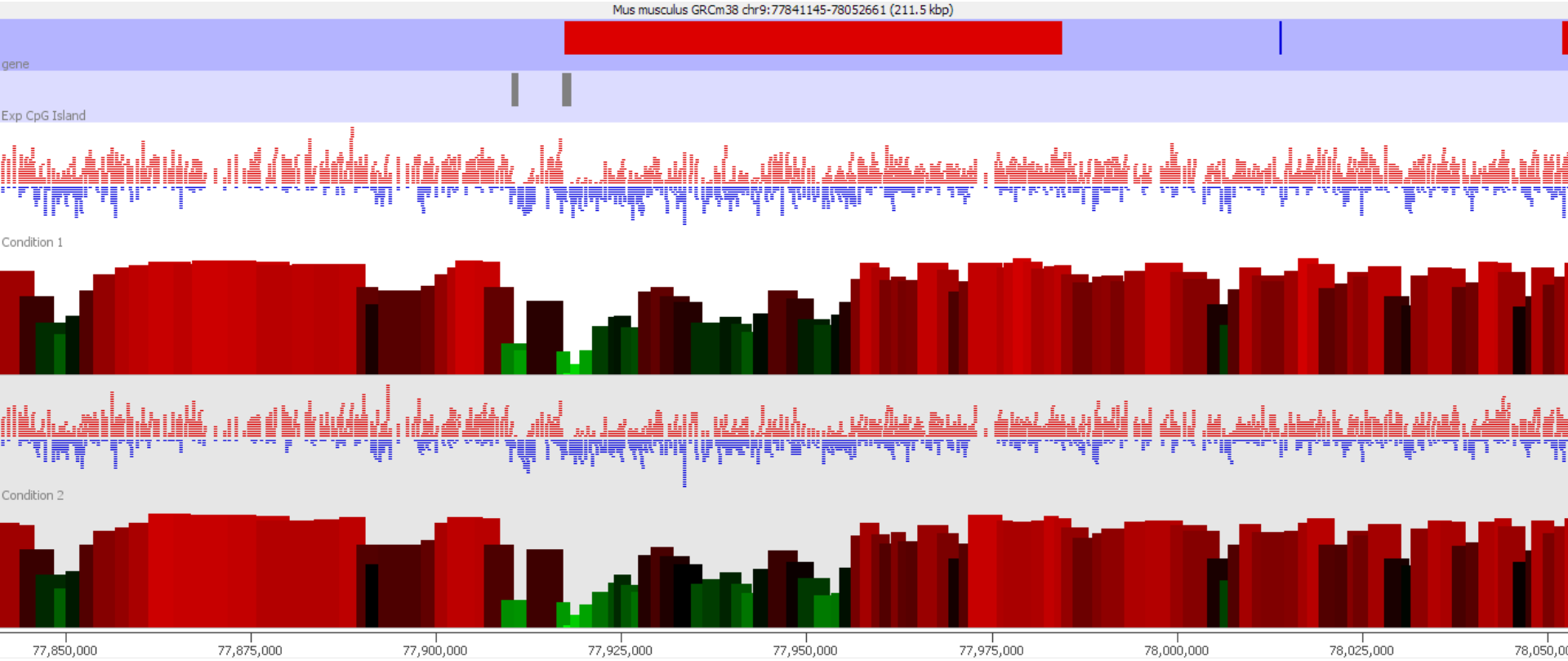
Babraham Bioinformatics

Viewing comparisons



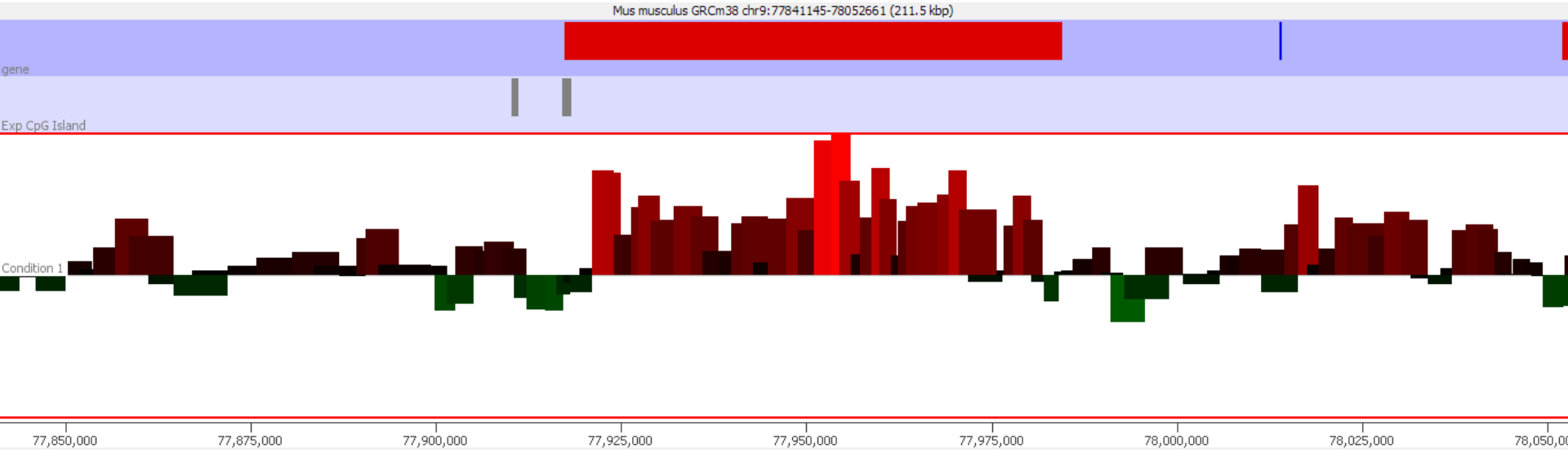
Babraham Bioinformatics

Viewing differences



Babraham Bioinformatics

Viewing differences



Babraham Bioinformatics

Trends

- Effects at individual loci can be subtle
- Want to find more generalised effect
- Collate information across whole genome
- Look for general trend

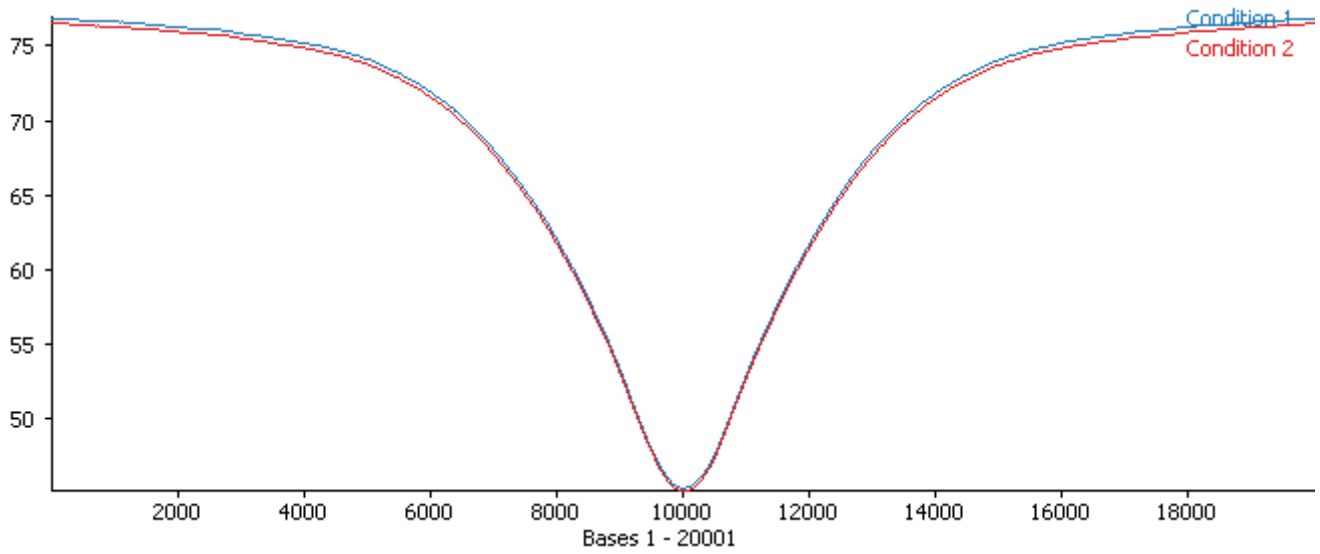
Babraham Bioinformatics

Considerations for trend plots

- What features to use
  - Fixed vs relative scale
- How much context
  - Variable scales
- How to calculate base measures
  - What window size
  - Aligned vs unaligned windows
- Missing values
- Scale

Babraham Bioinformatics

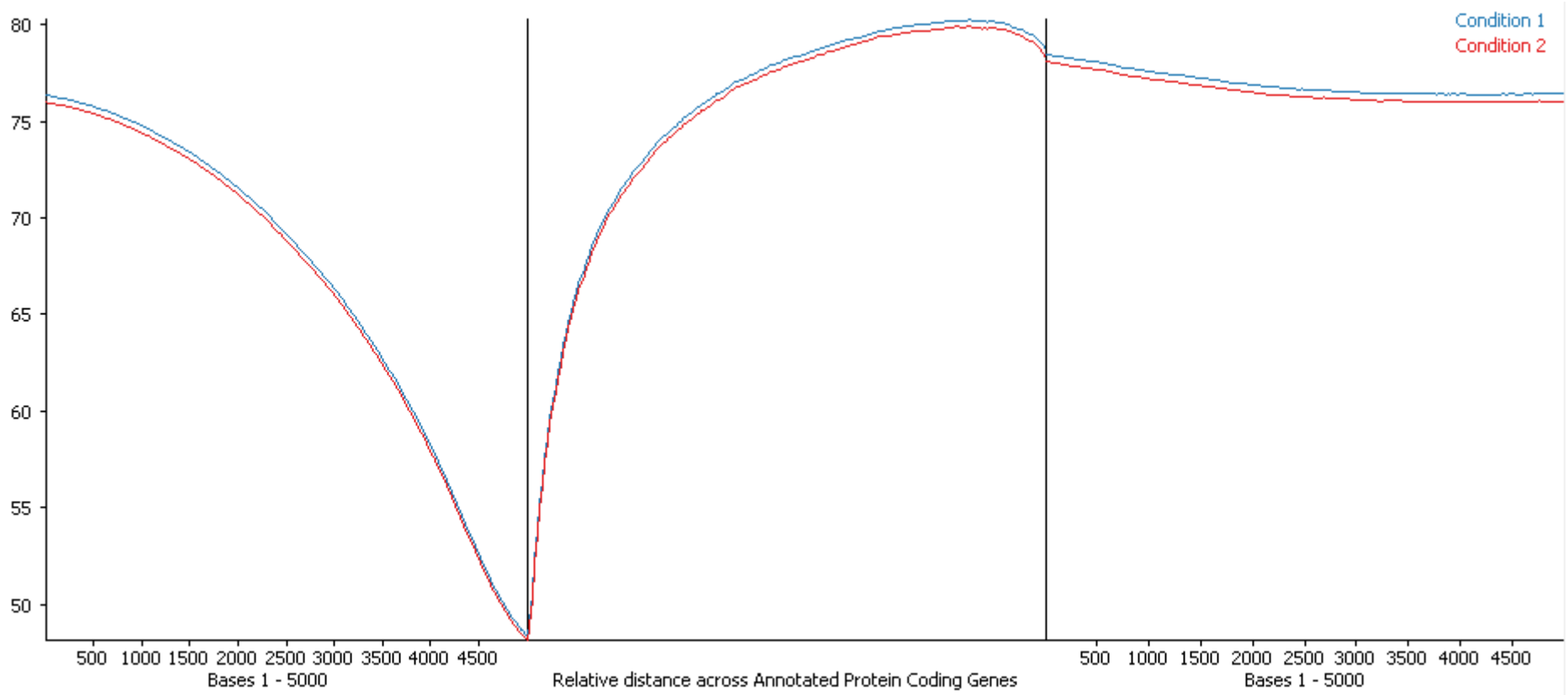
Simple Example



Methylation profile centred on CpG islands +/- 10kb

Babraham Bioinformatics

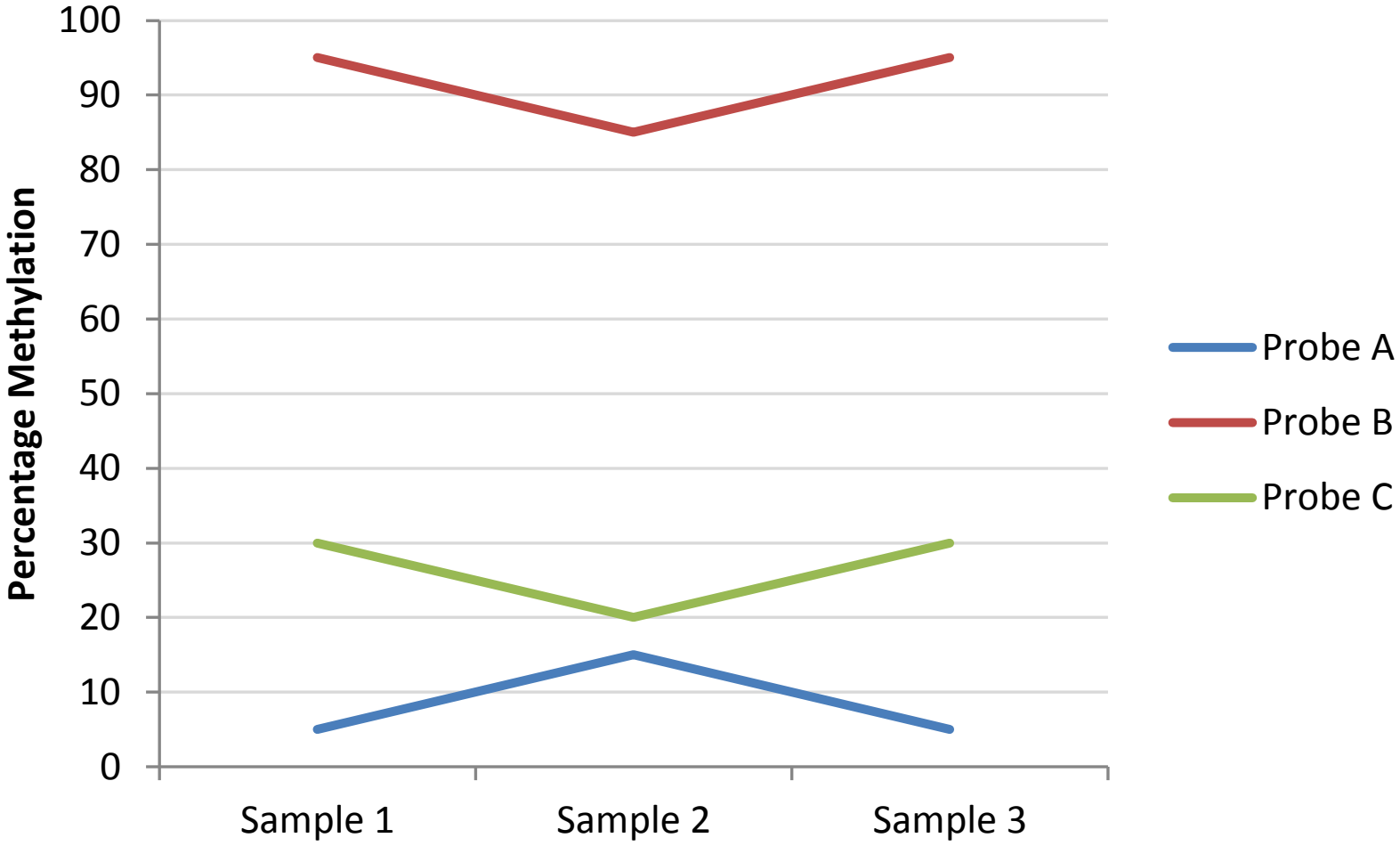
More complex example



Methylation profile over genes +/- 5kb

Babraham Bioinformatics

Clustering



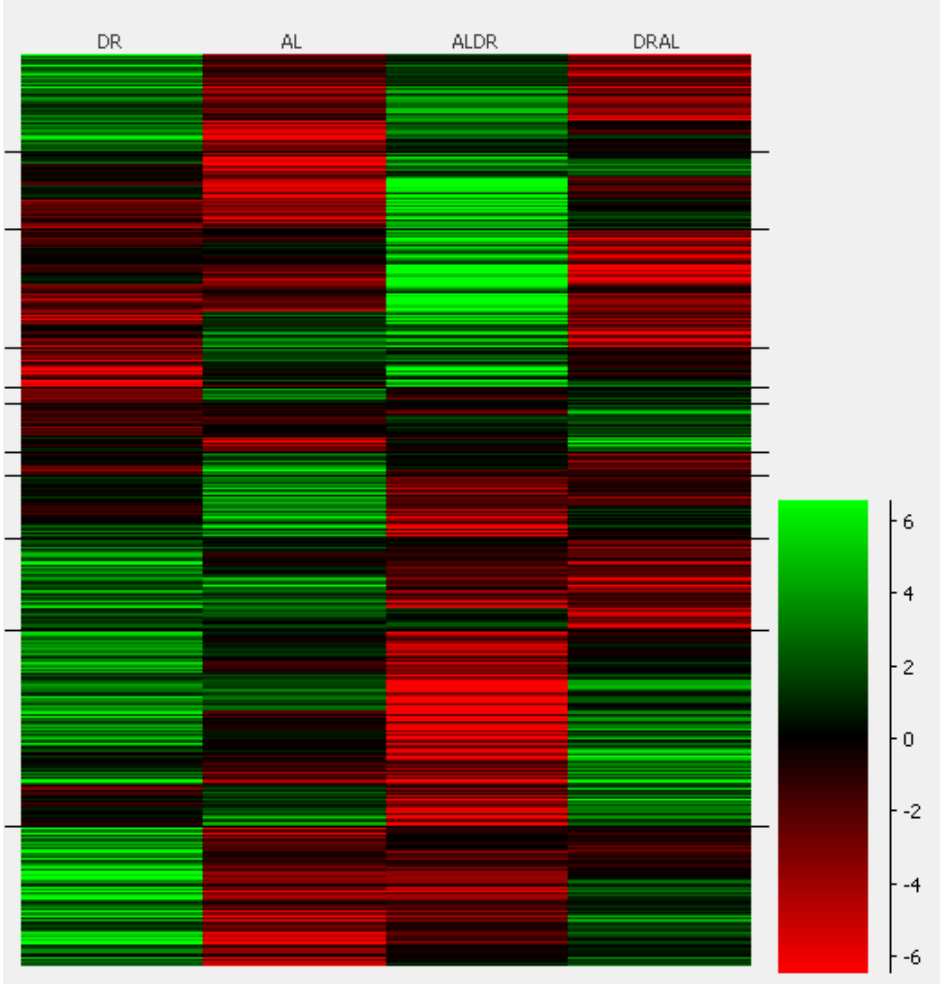
Babraham Bioinformatics

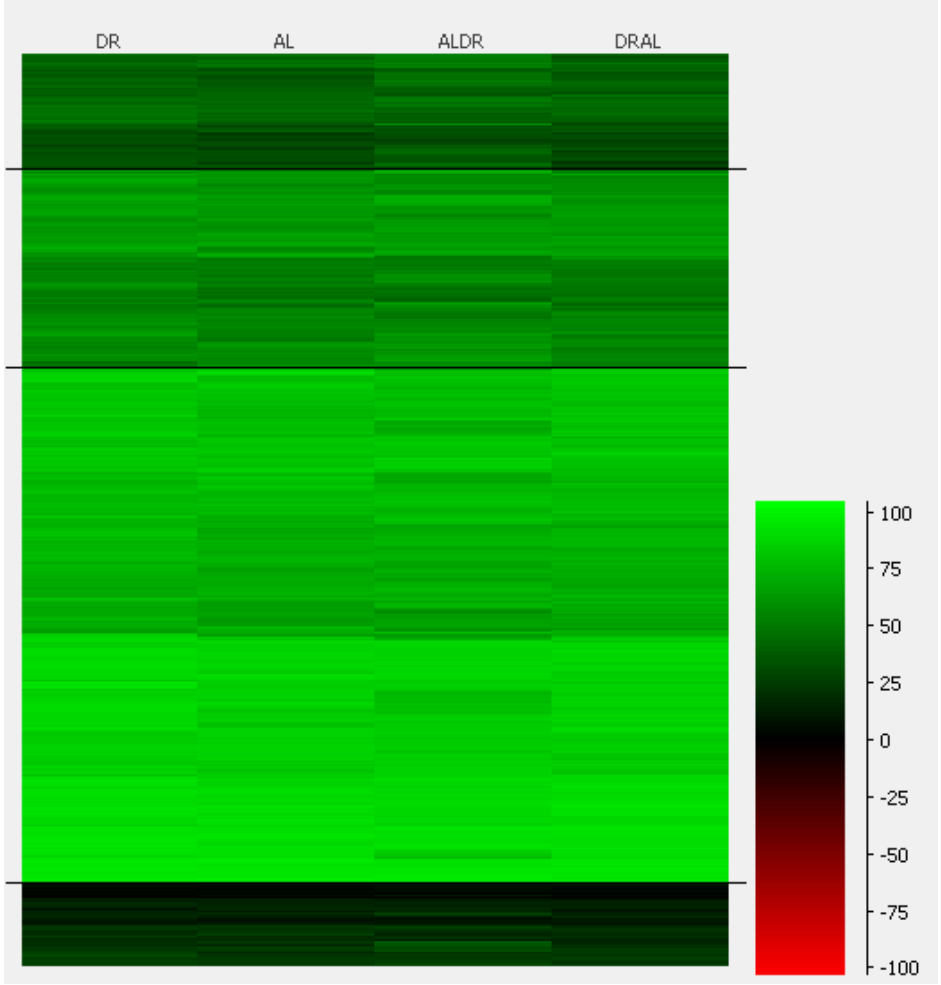
Clustering

- Correlation Clustering
  - Focusses on the differences between conditions
  - Absolute values not important
  - Look for similar trends
  - Show median normalised values
- Euclidean Clustering
  - Focusses on absolute differences between conditions
  - Look for similar levels
  - Show raw values

Babraham Bioinformatics

Clustering





Babraham Bioinformatics