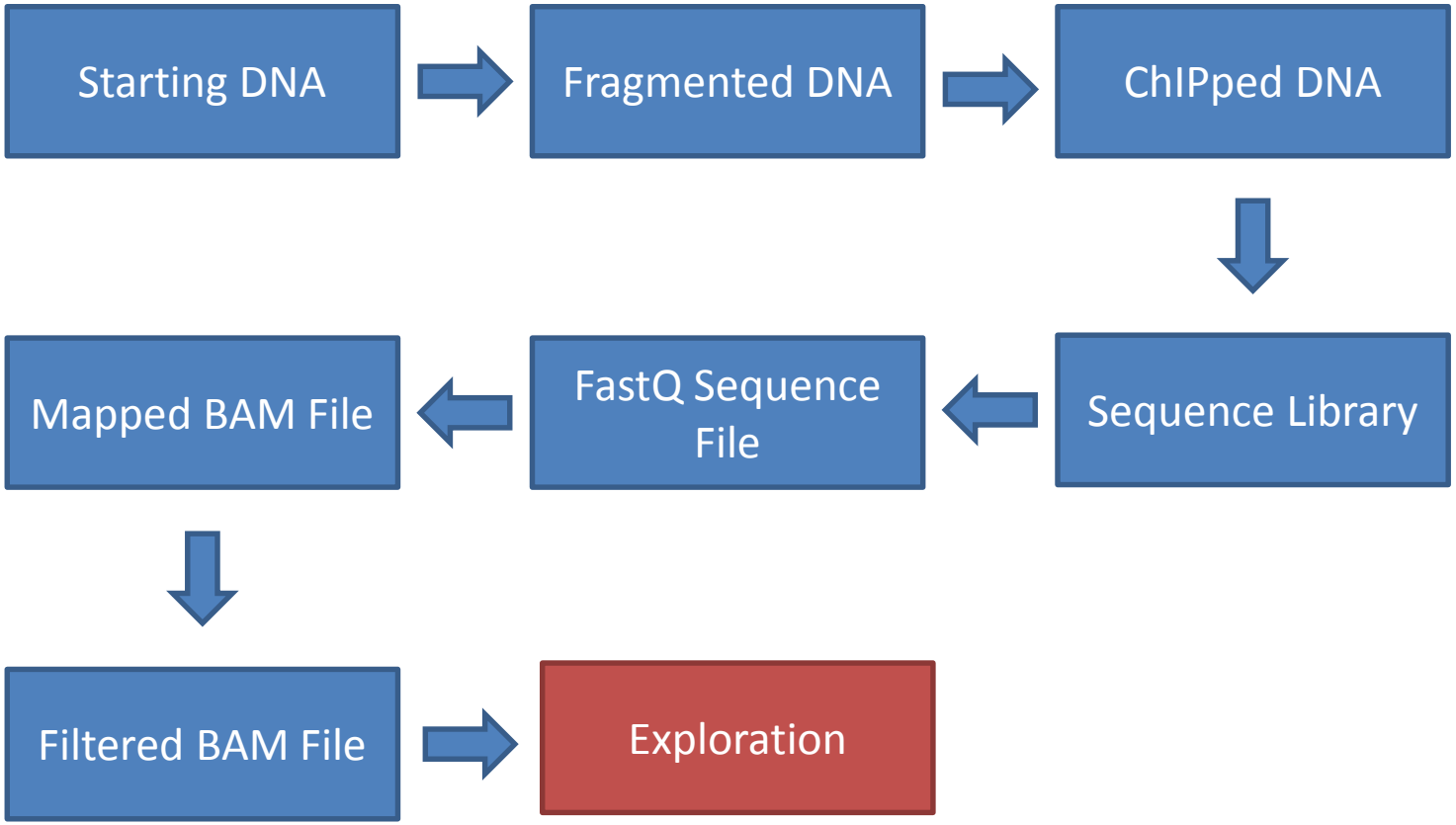
Exploring and Understanding ChIP-Seq data

Simon Andrews
simon.andrews@babraham.ac.uk
@simon_andrews
v2018-03-02

Babraham Bioinformatics

Data Creation and Processing



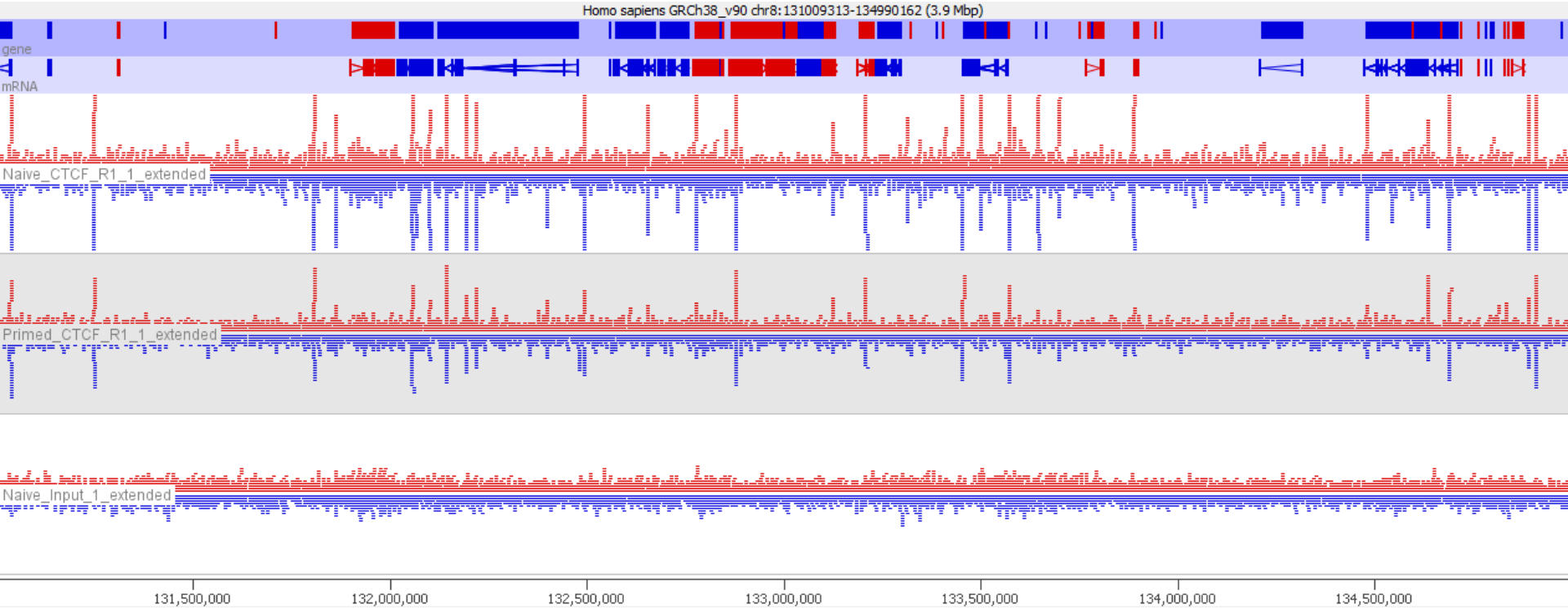
Babraham Bioinformatics

Some Basic Questions

- Is there any enrichment?
- What is the size / patterning of enrichment?
- How well are my controls behaving?
- What is the best way to quantitate this data?
- Are there any technical artefacts?

Babraham Bioinformatics

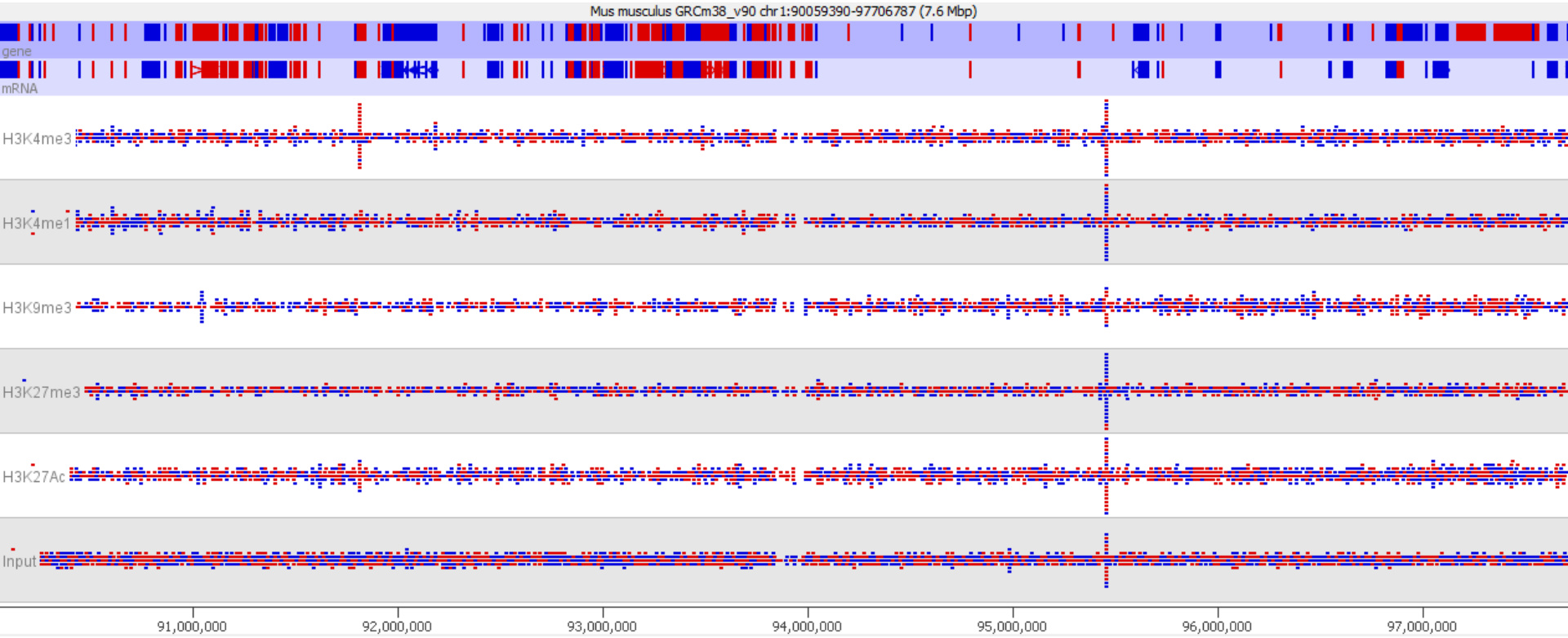
Start with a visual inspection



- Is there any enrichment?
- What is the size / patterning of enrichment?
- How well are my controls behaving?

Babraham Bioinformatics

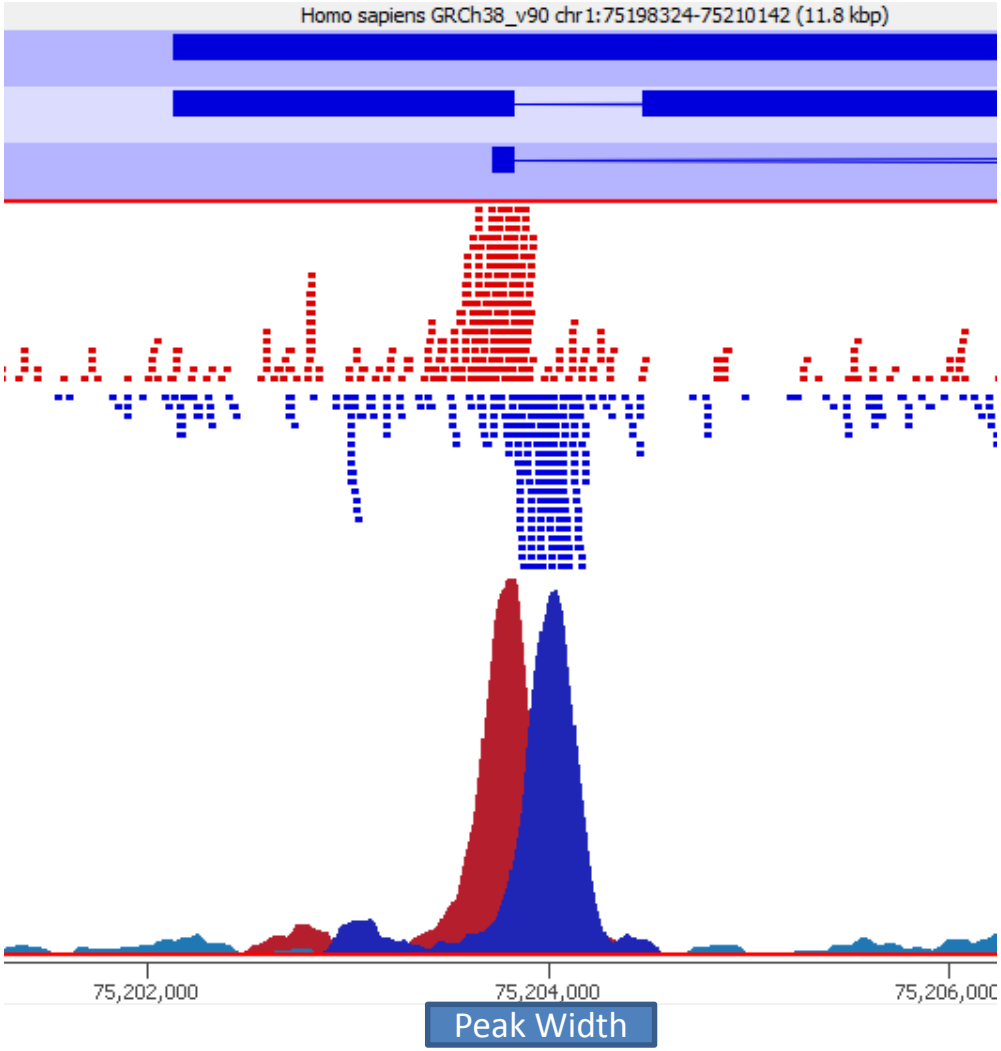
Start with a visual inspection



- Is there any enrichment?
- What is the size / patterning of enrichment?
- How well are my controls behaving?

Babraham Bioinformatics

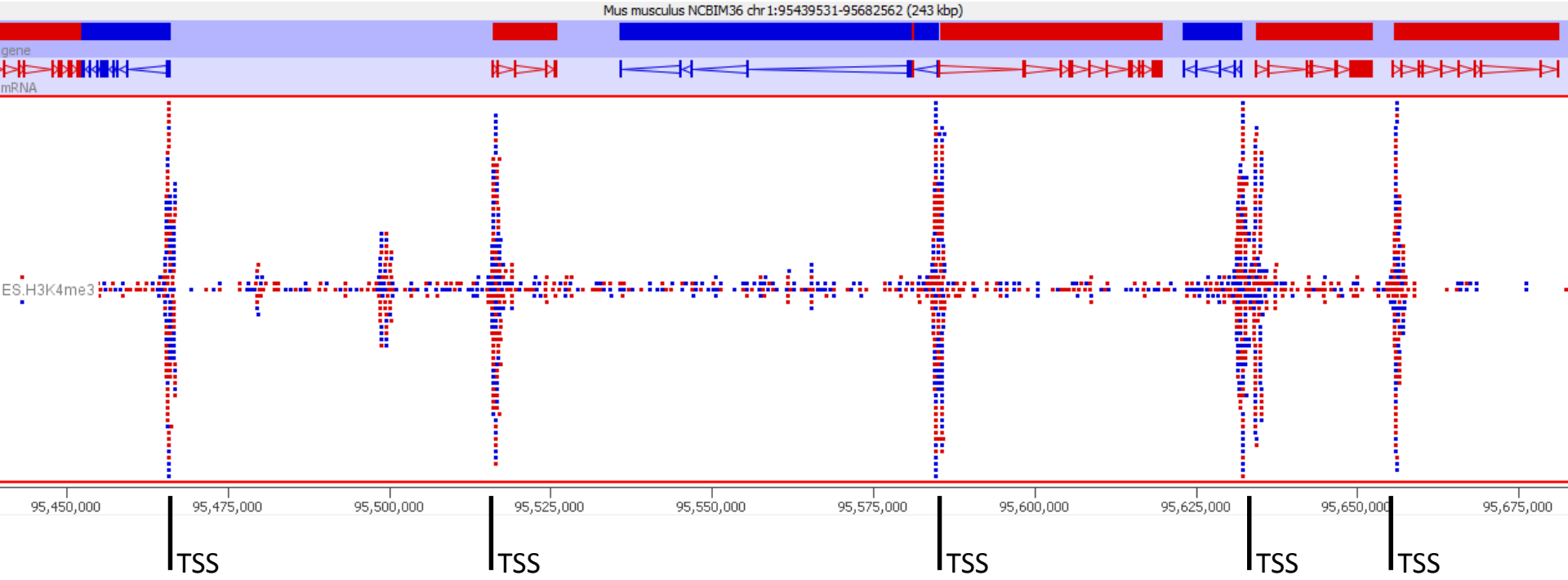
Extending reads if necessary



Babraham Bioinformatics

For point enrichment, insert size is roughly peak width/2

Look for peaks
Associate with features



- Are my peaks narrow or broad
- Do peak positions obviously correspond to existing features?

Babraham Bioinformatics

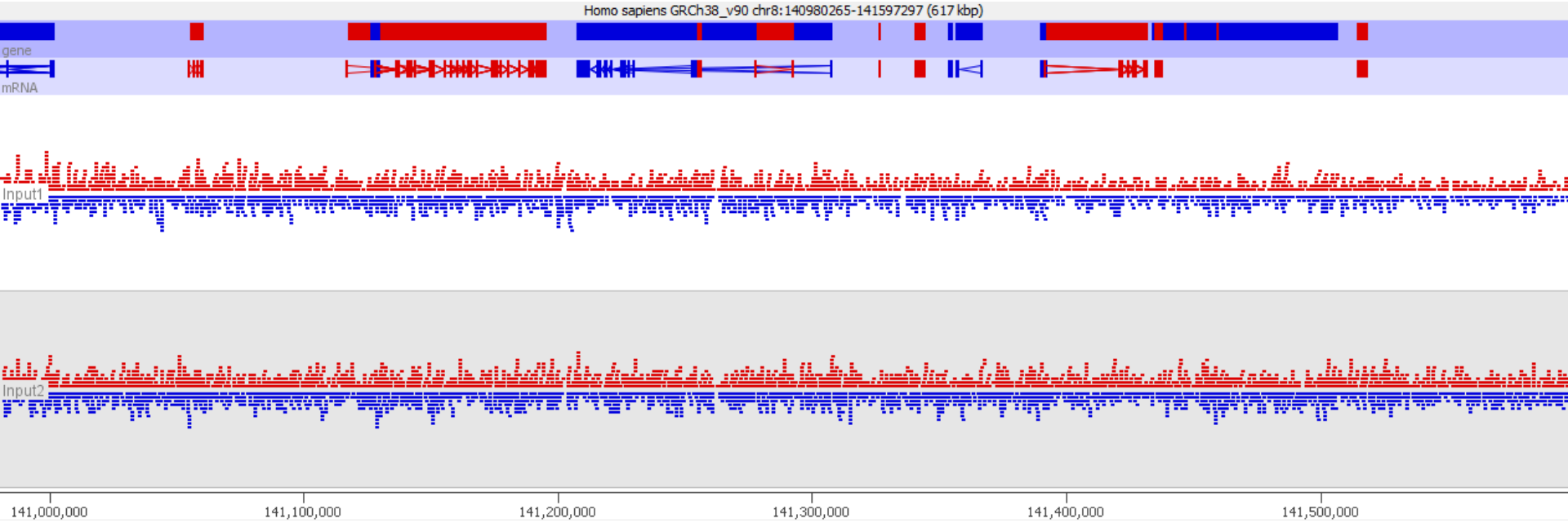
Examine Controls

- IgG or other Mock IP
  - Good result is no material at all
  - Not worth sequencing. Reads are only informative if the ChIP hasn't worked.

- Input material (sonicated / Mnase etc)
  - Genomic library - everywhere equally
  - Technical issues can cause variation

Babraham Bioinformatics

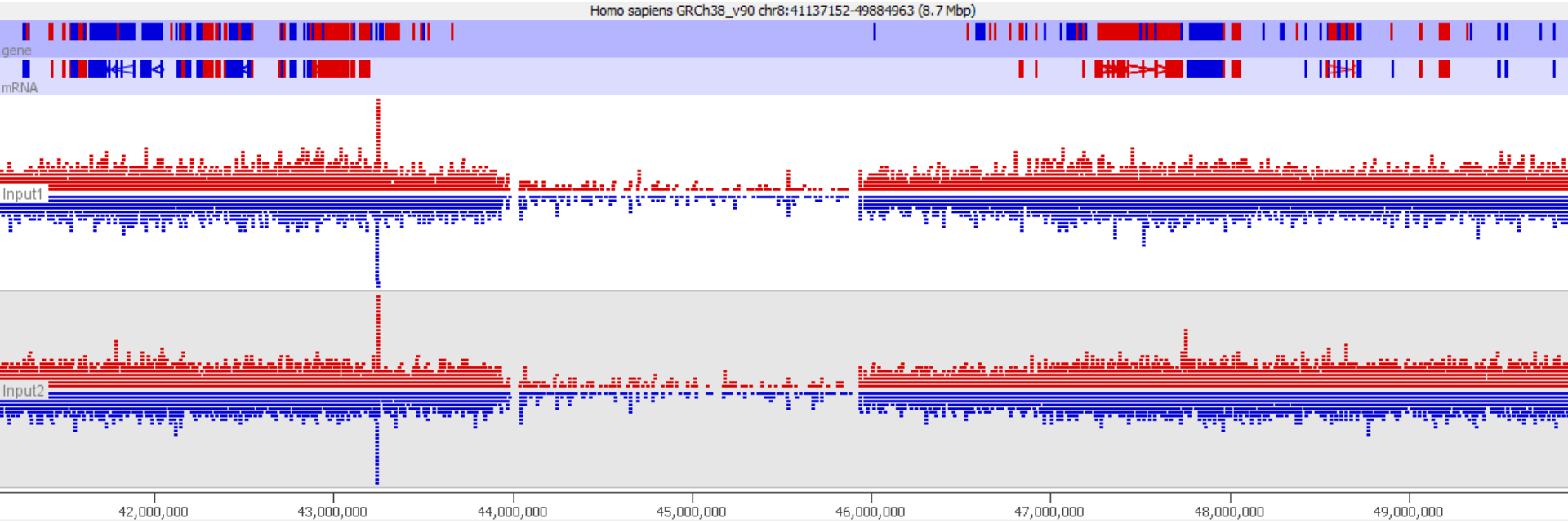
Examine Controls



- Does the coverage look even
- If there are multiple inputs to do they look similar

Babraham Bioinformatics

Examine Controls



Babraham Bioinformatics

Why do controls misbehave?

- Low coverage
  - Repetitive unmappable regions
  - Holes in the assembly
- High coverage
  - Mismapped reads from outside the assembly
- Biases
  - DNA stability
  - GC content
  - Segmental Duplication

Babraham Bioinformatics

Types of input problem

- Categorical
  - Mismapped reads
  - Indicates that the region can't be trusted
  - Blacklist and remove - don't try to correct

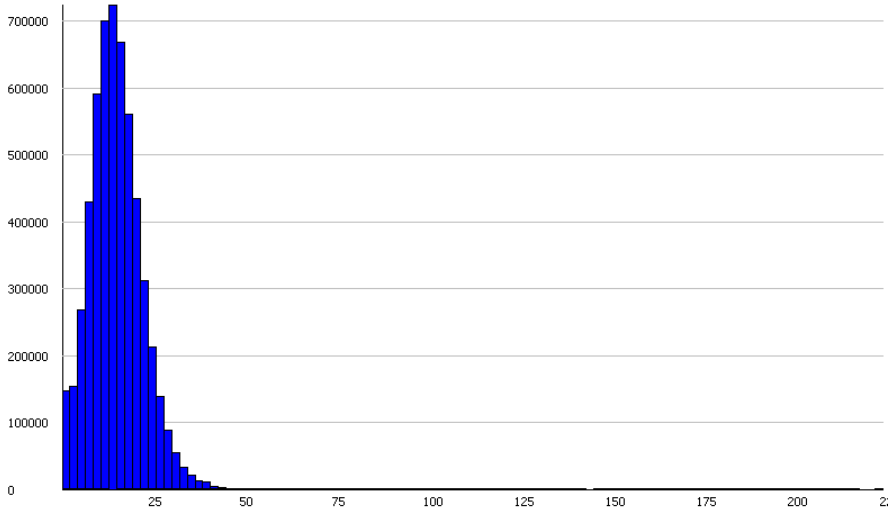
- Quantitative
  - Other biases (most often GC)
  - Some potential to correct, but difficult
  - Hopefully consistent, so will cancel out

Babraham Bioinformatics

Making Blacklists

- Unusual Coverage
  - Outlier detection (boxplots etc.)
  - Often only filter over-representation (maybe also zero counts)

- Pre-built lists
  - Large projects often build these (ENCODE)
  - Not for all species



Babraham Bioinformatics

Pre-Existing Blacklists

- ENCODE
- modENCODE
- UCSC

frontiers in GENETICS

TECHNOLOGY REPORT ARTICLE
published: 10 April 2014
doi: 10.3389/fgene.2014.00075

Impact of artifact removal on ChIP quality metrics in ChIP-seq and ChIP-exo data

Thomas S. Carroll[1*†], Ziwei Liang[2†], Rafik Salama[1†], Rory Stark[1] and Ines de Santiago[1*]

[1] Cambridge Institute CRUK, University of Cambridge, Cambridge, UK
[2] Lymphocyte Development, MRC Clinical Sciences Centre, Imperial College, London, UK

- Check assembly versions

Babraham Bioinformatics

https://sites.google.com/site/anshulkundaje/projects/blacklists

Comparison of samples

Babraham Bioinformatics

Initial Quantitation

- Always start with a simple unbiased quantitation (not focussed on features/peaks)
- Tiled measures over the whole genome
  - Use approximate insert size as window size
  - Something around 500bp is normally sensible
- Linear read count quantitation corrected for total library size

Babraham Bioinformatics

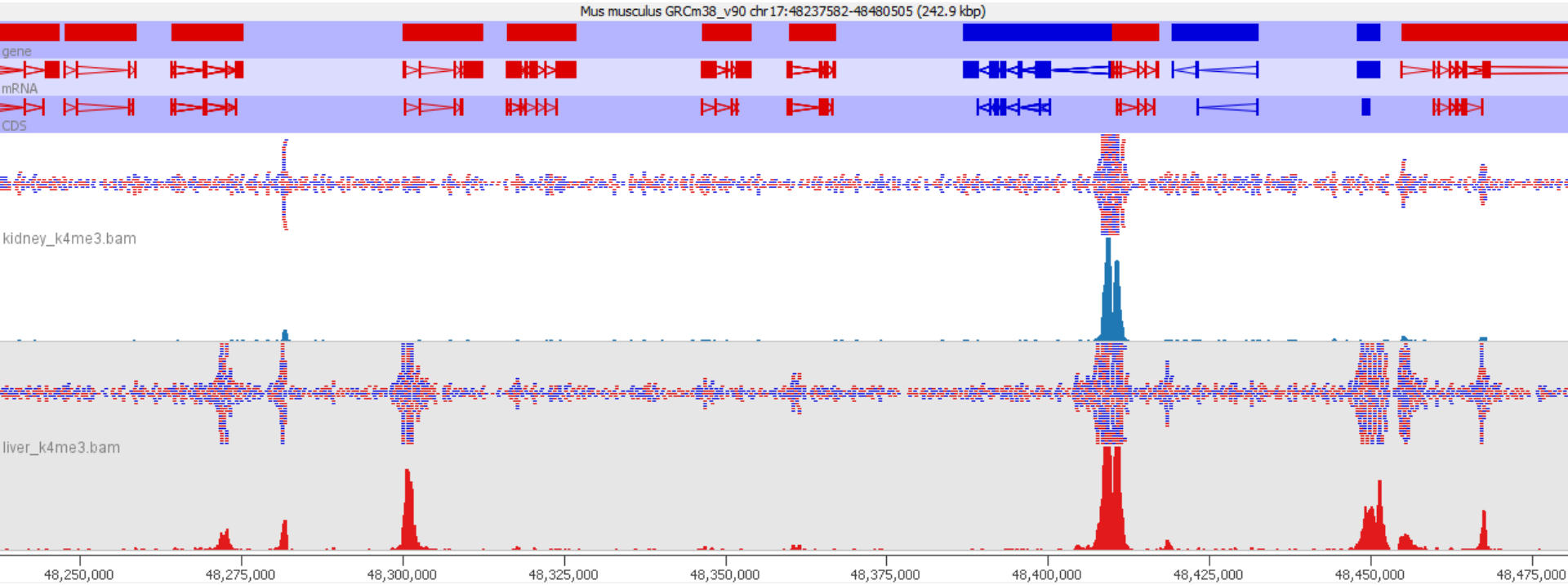
Quantitation Questions

- Initially similar to what we did with the raw data, but with some values behind it
  - Can we observe enrichment in our ChIP above our controls
  - Do we see similar enrichment between replicates, or between conditions

Babraham Bioinformatics

Compare samples

Visual comparison against raw data

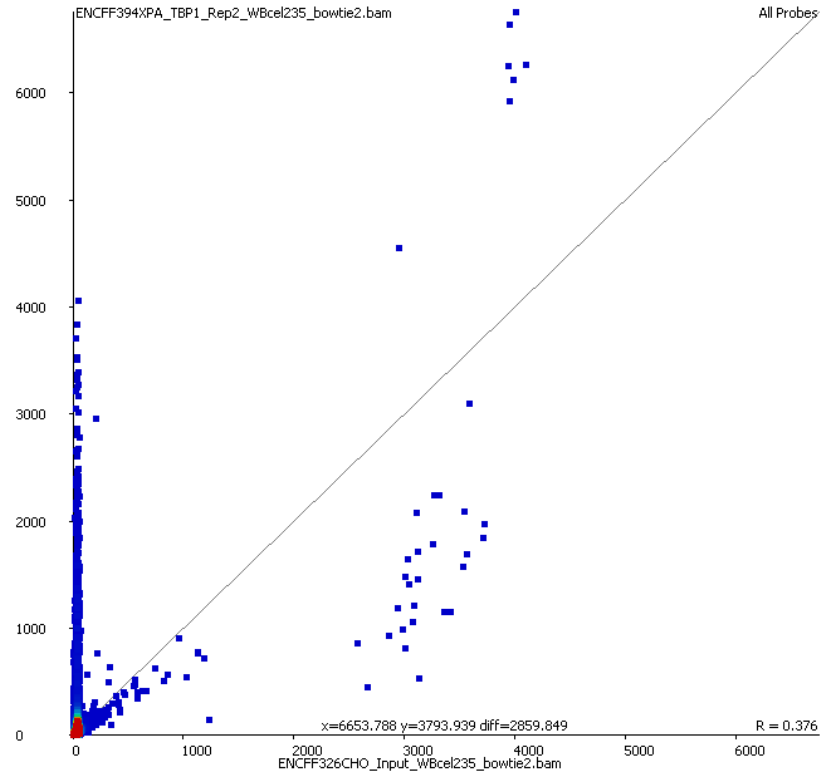


Babraham Bioinformatics

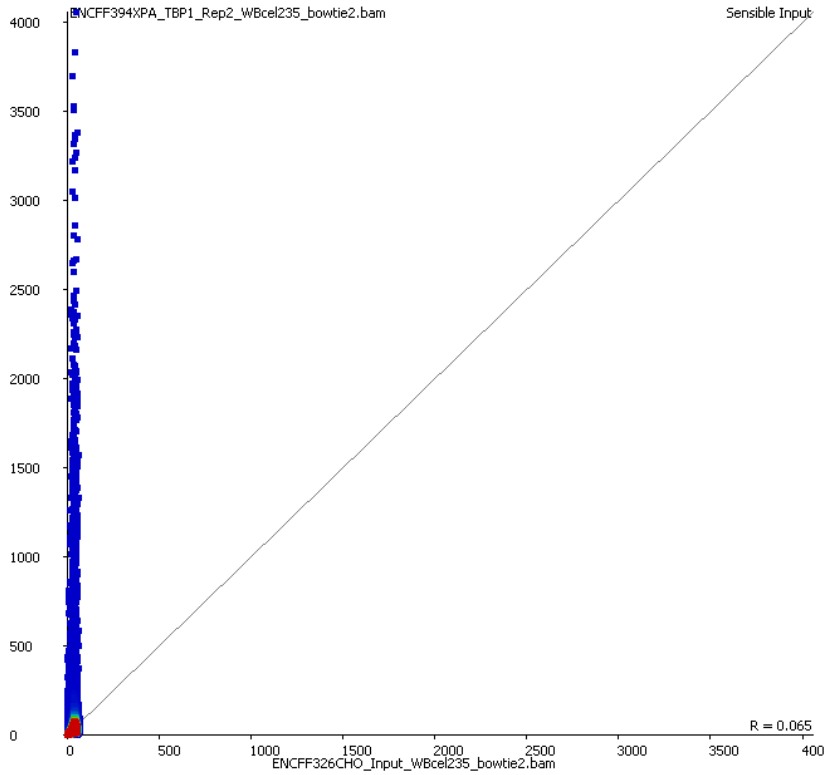
- Similar apparent overall enrichment
- Any obvious differences

Compare samples

Scatterplot input vs ChIP



Raw



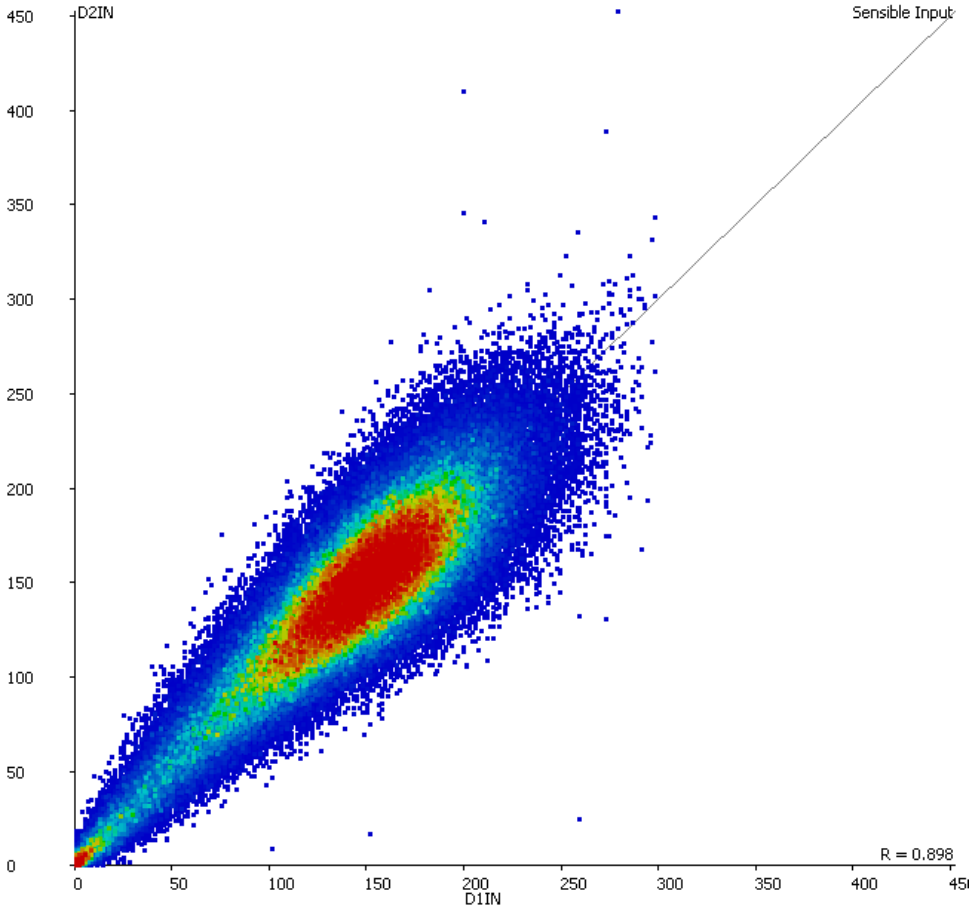
Filtered

- Any relationship between input and ChIP
- What proportion enriched
- How enriched over input

Babraham Bioinformatics

Compare samples

Scatterplot input vs input

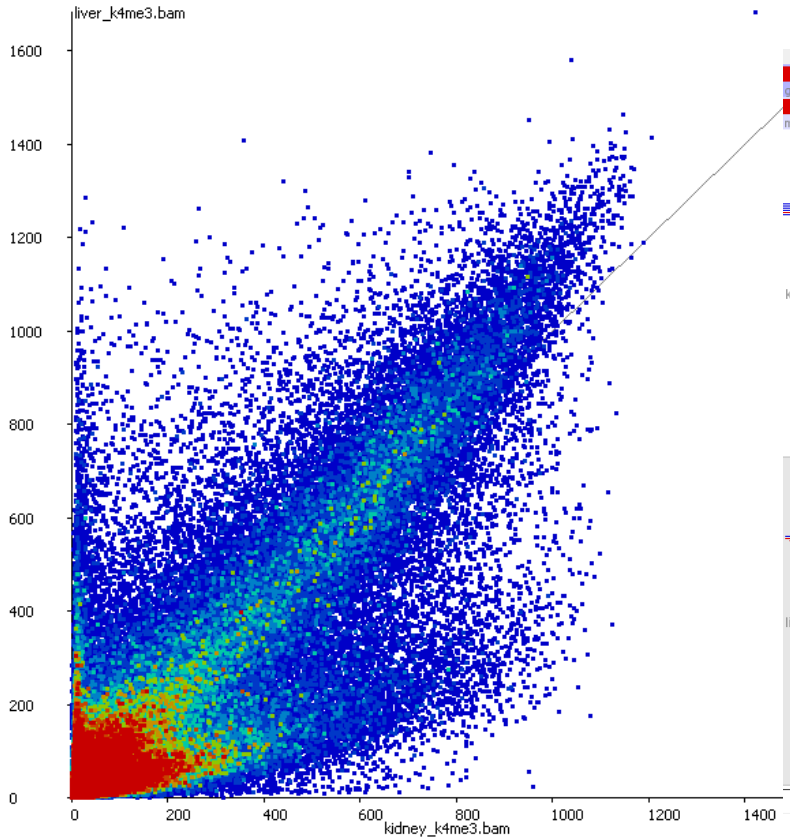


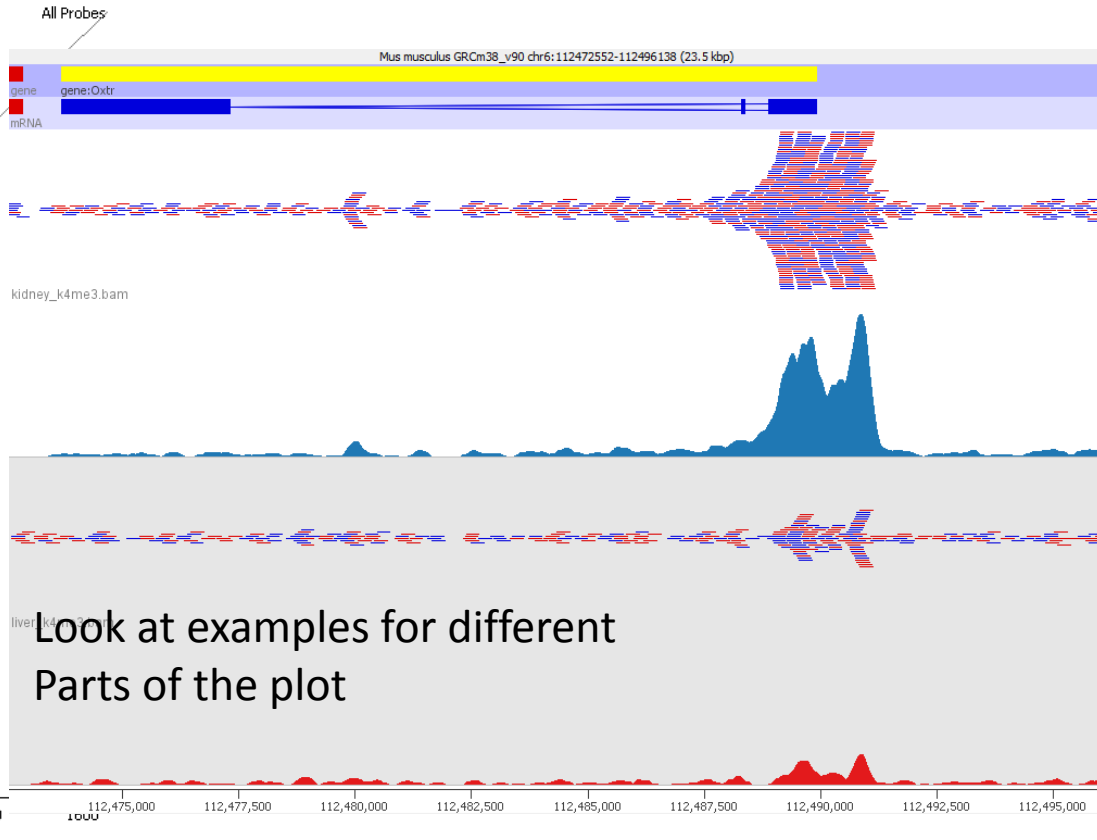
- Any suggestion of differential biases in inputs
- Can we merge them to use as a common input

Babraham Bioinformatics

Compare samples

Scatterplot ChIP vs ChIP



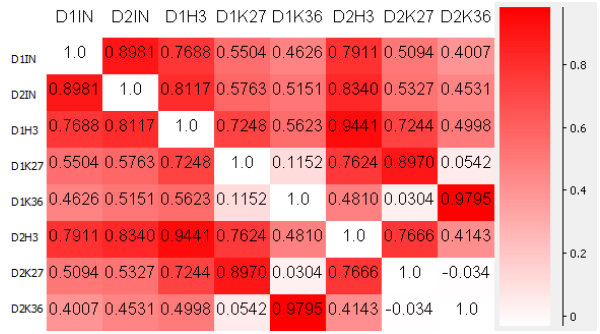


Look at examples for different Parts of the plot

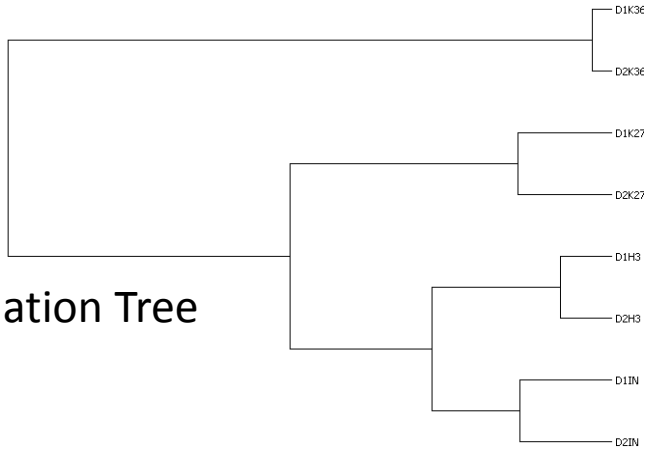
- Look for outgroups (differentially enriched)
- Compare level of enrichment (compare to diagonal)

Compare samples

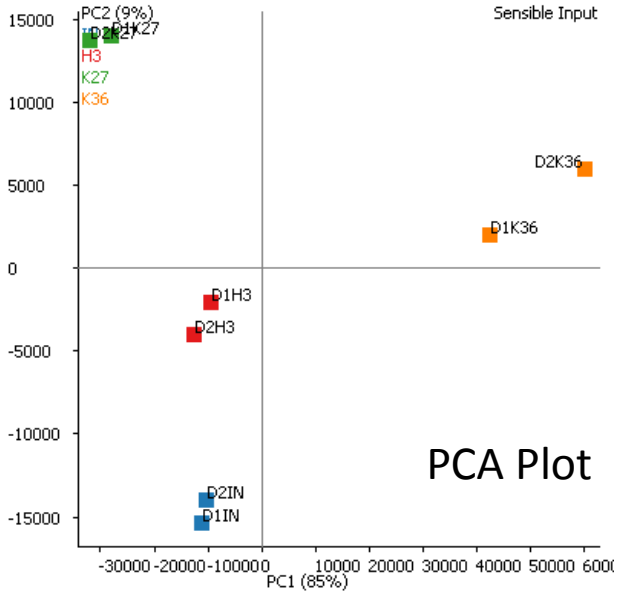
Higher level clustering



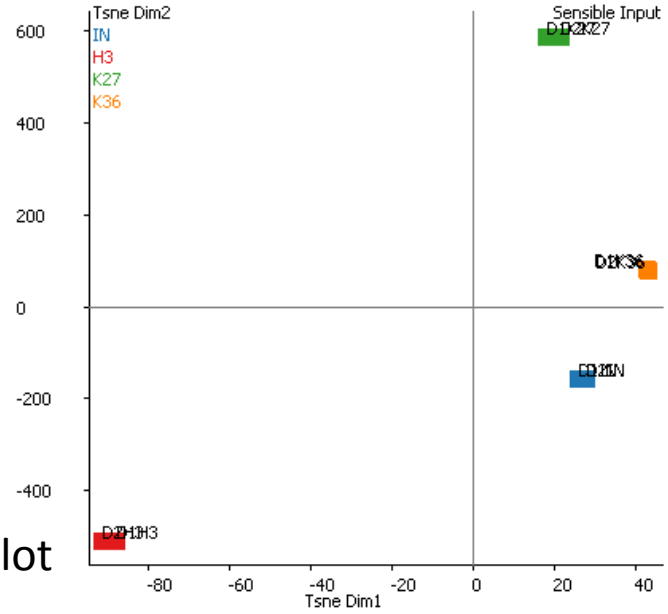
Correlation Matrix



Correlation Tree



PCA Plot

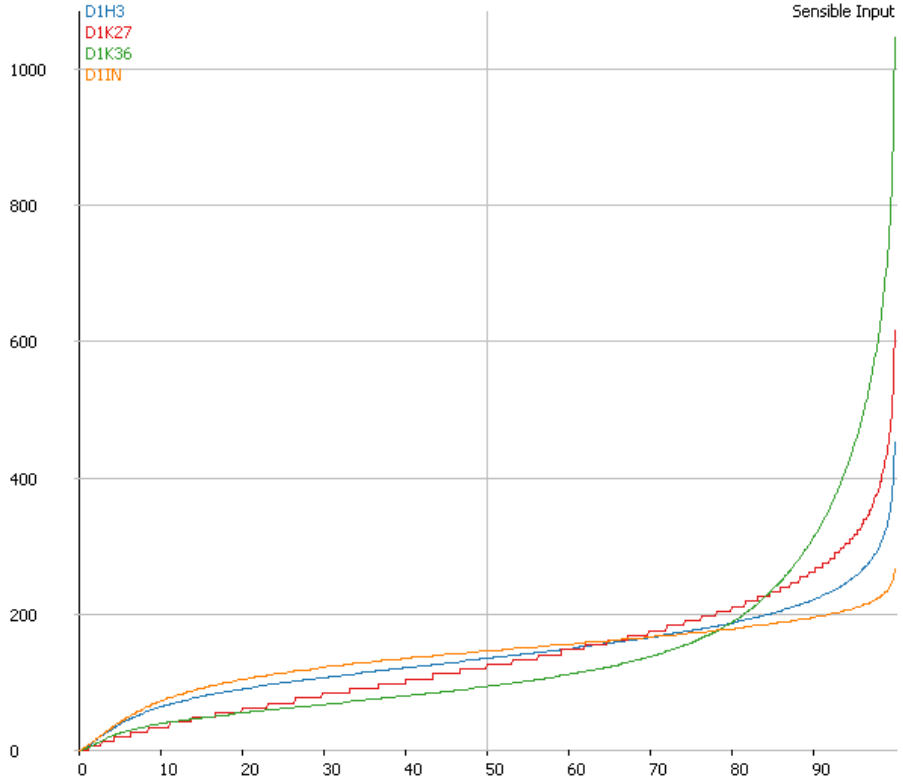


tSNE Plot

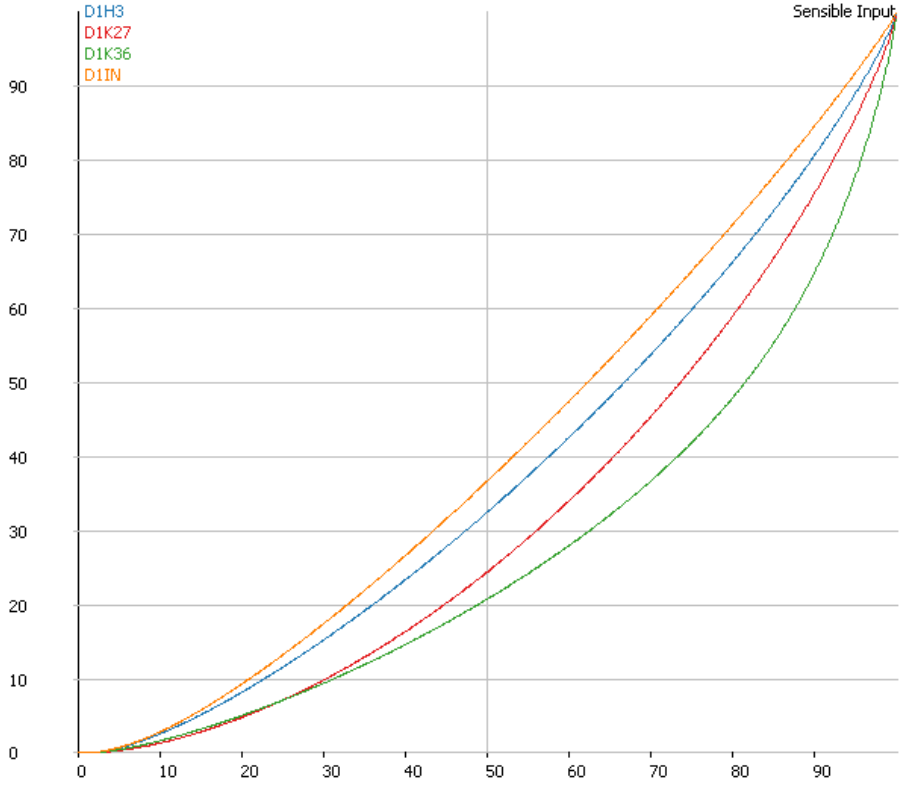
Babraham Bioinformatics

Compare samples

Summarise distributions



Cumulative Distribution Plot



QQ Plot

- Flatness of input
- Separation of ChIPs
- Consistency of ChIPs

Babraham Bioinformatics

Associate enrichment with features

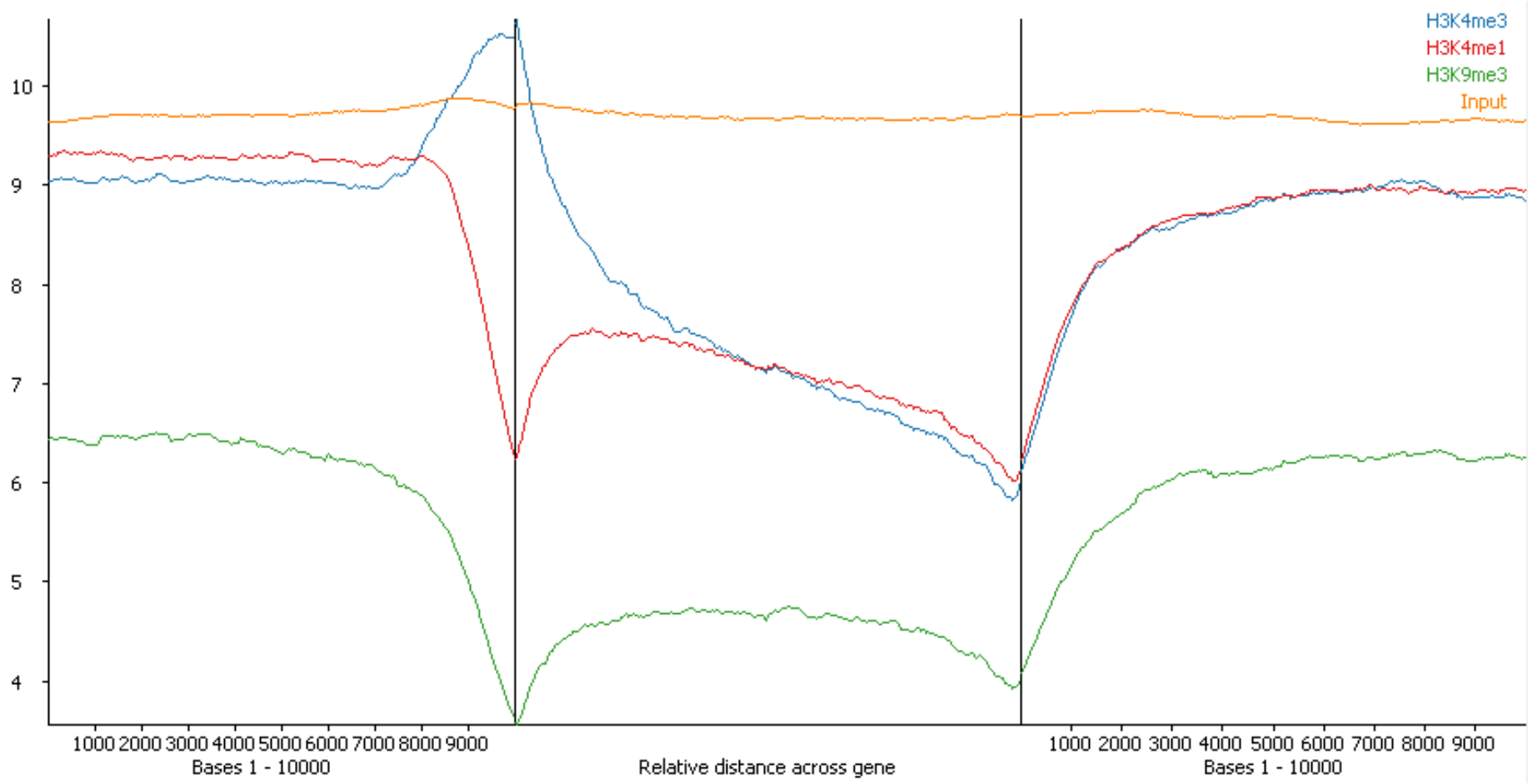
Babraham Bioinformatics

Trend Plots

- Graphical way to look at overall enrichment relative to positions in features
  - Gene bodies
  - Promoters
  - CpG islands
- May influence how we later quantitate and analyse the data
  - Analyse per feature
  - Look for exceptions to the general rule

Babraham Bioinformatics

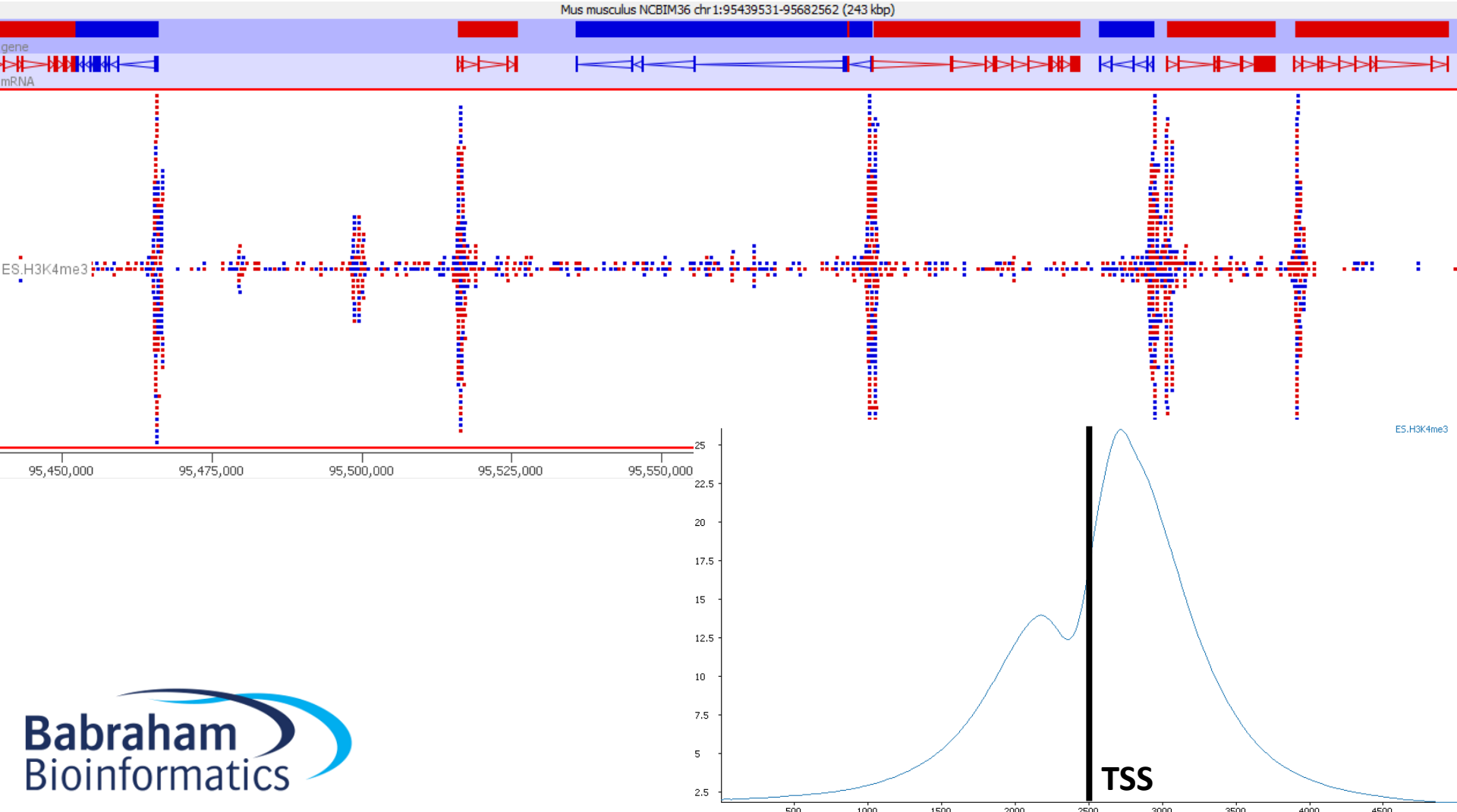
Trend Plot Example



Babraham Bioinformatics

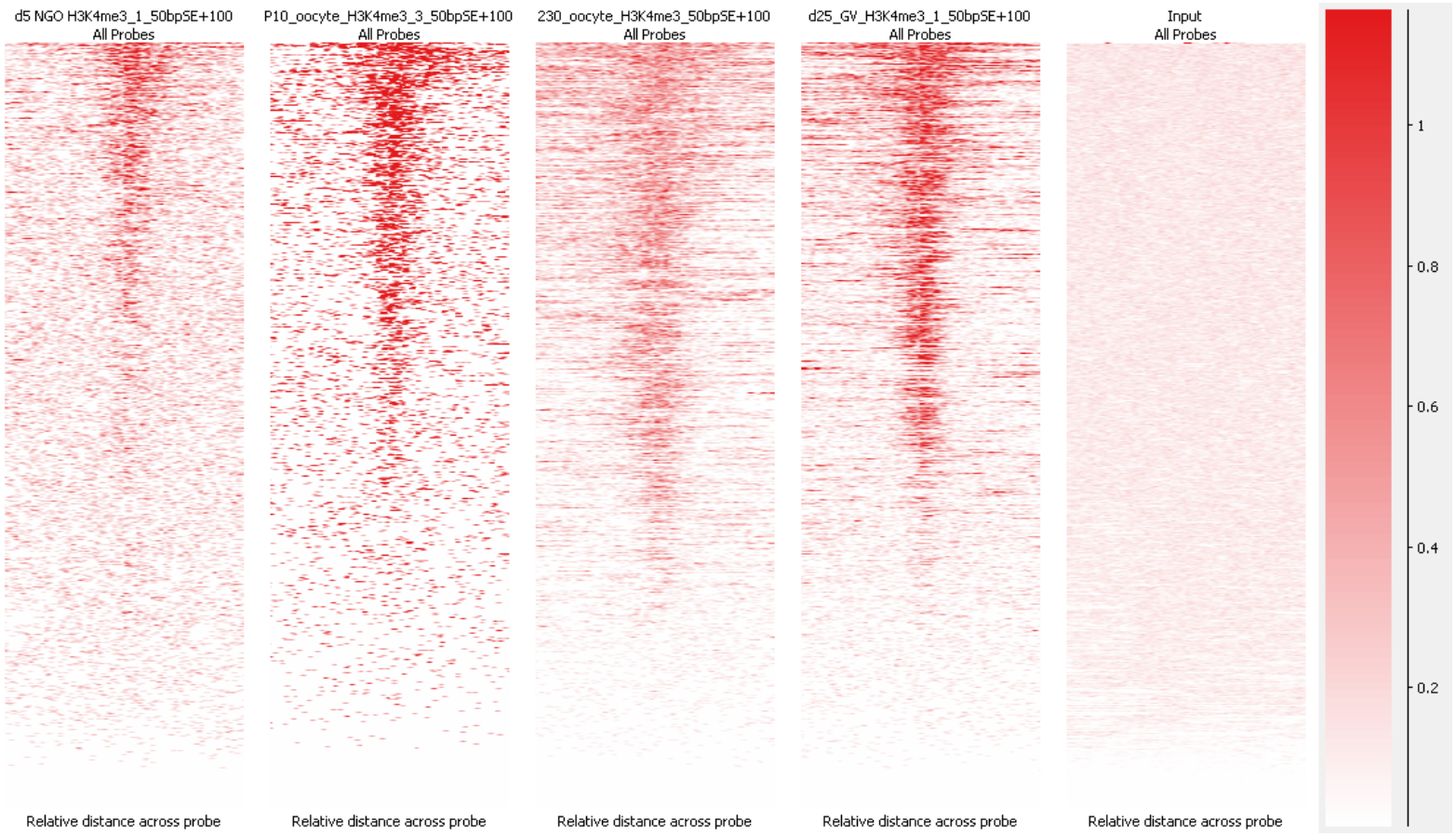
- Overall average
- Says nothing about proportion of features affected

Trend plots should match the data



Babraham Bioinformatics

Aligned Probes Plots give more detail



Babraham Bioinformatics

- Information per feature instance
- Comparison of equivalent features in different marks/samples

After exploration you should..

- Know whether your ChIP is really enriched
- Know the nature / shape of the enrichment
- Know whether your controls behave well
- Know whether you're likely to have differential enrichment
- Know if you will need additional normalisation
- Know the best strategy to measure your data

Babraham Bioinformatics