How Sequencing Experiments Fail

v1.0

Simon Andrews
simon.andrews@babraham.ac.uk

Babraham Institute

Babraham Bioinformatics

Classes of Failure

Something went wrong with a machine

Samples aren't what they're supposed to be

Problems during sequencing library preparation

Unexpected material in your libraries

Samples didn't behave the way you expected

Drawing the wrong conclusion from the data

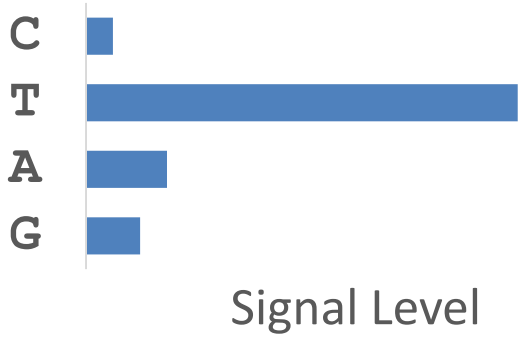
Babraham Bioinformatics

Technical

Babraham Bioinformatics

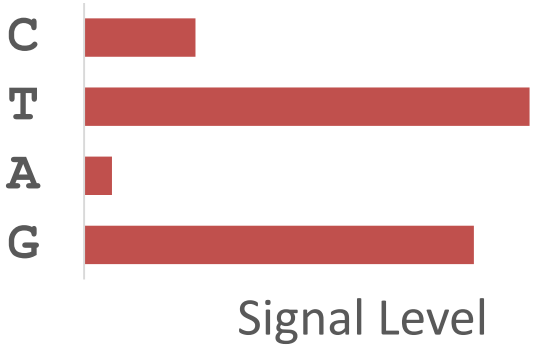
Technical

Technical Failures



Call = T

Confidence = High



Call = T

Confidence = Low

Call = T

Confidence = Low

Babraham Bioinformatics

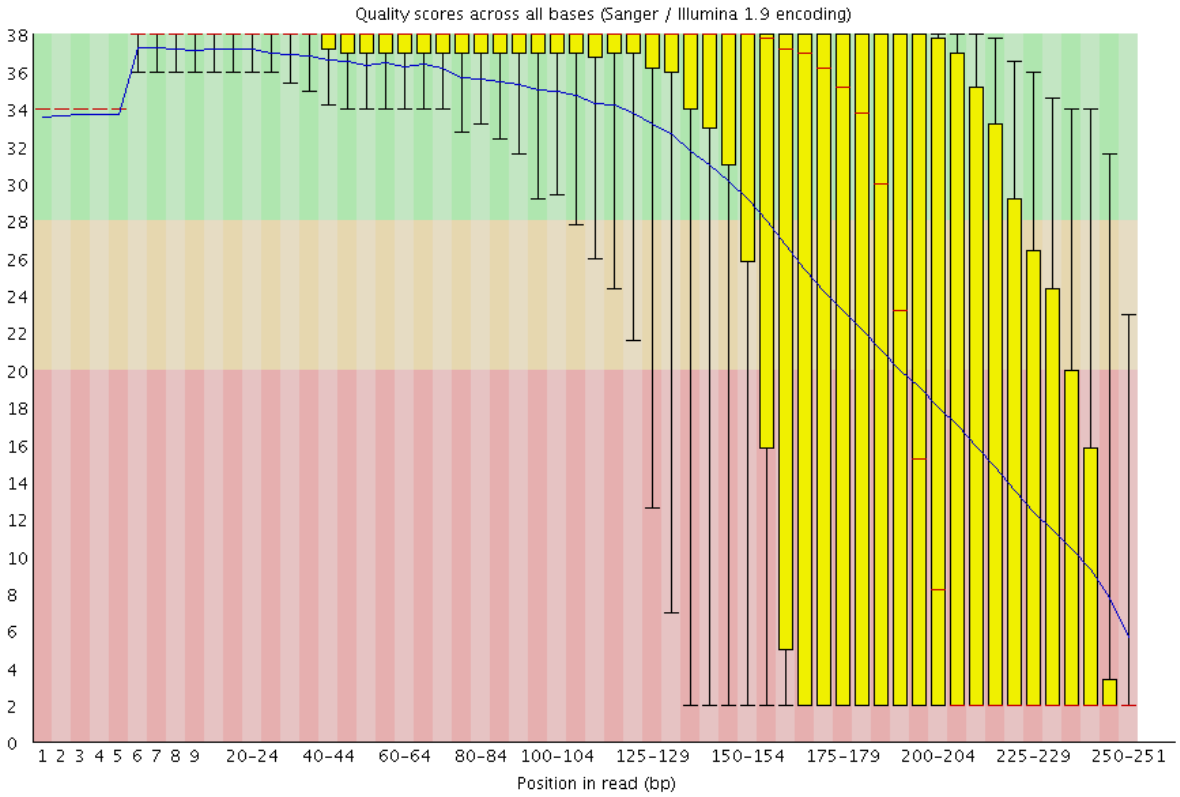
Technical

Phred Scores

$$\text{Phred} = -10 \log_{10} p$$

p = Probability call is incorrect

10% error	Phred10
1% error	Phred20
0.1% error	Phred30



Babraham Bioinformatics

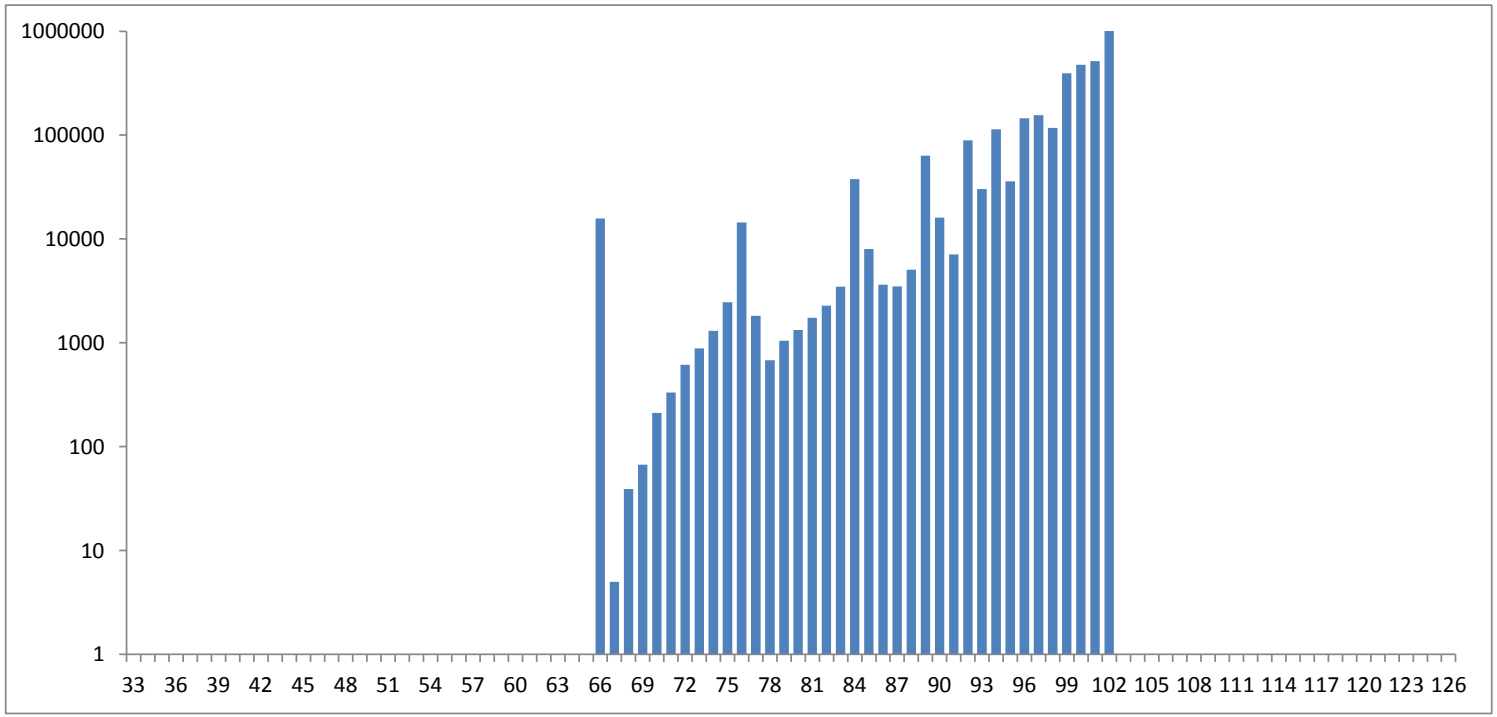
Technical

Incorrect Encoding

Phred64

```
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefgh
```

Phred33



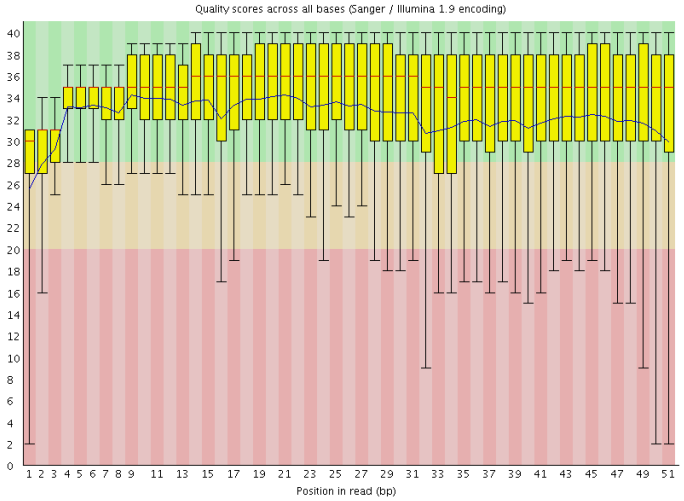
Phred64 (Illumina)

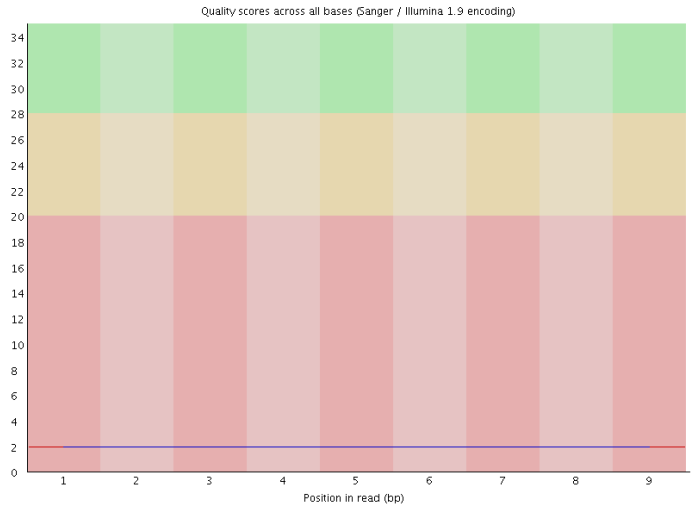
Phred33 (Sanger)

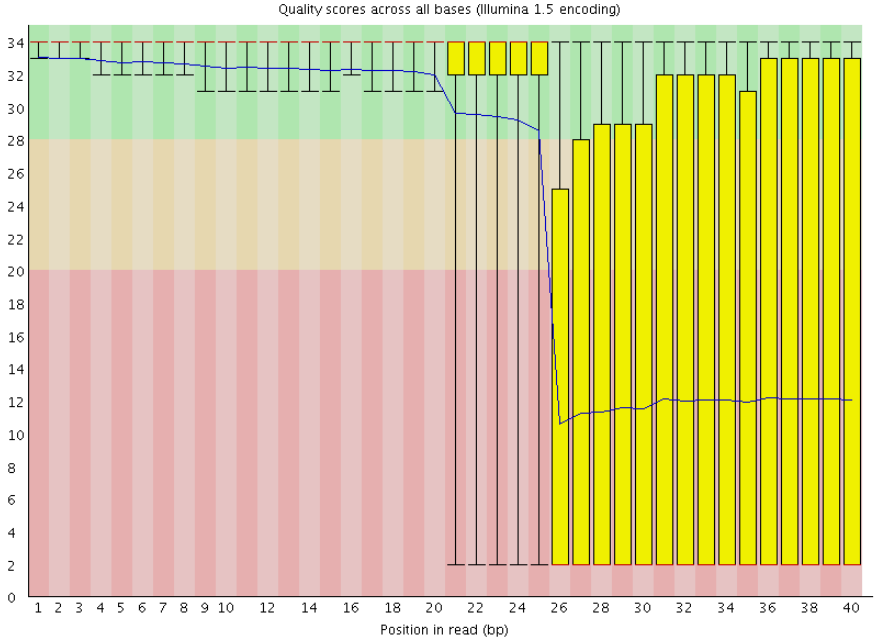
Babraham Bioinformatics

Technical

Phred Scores



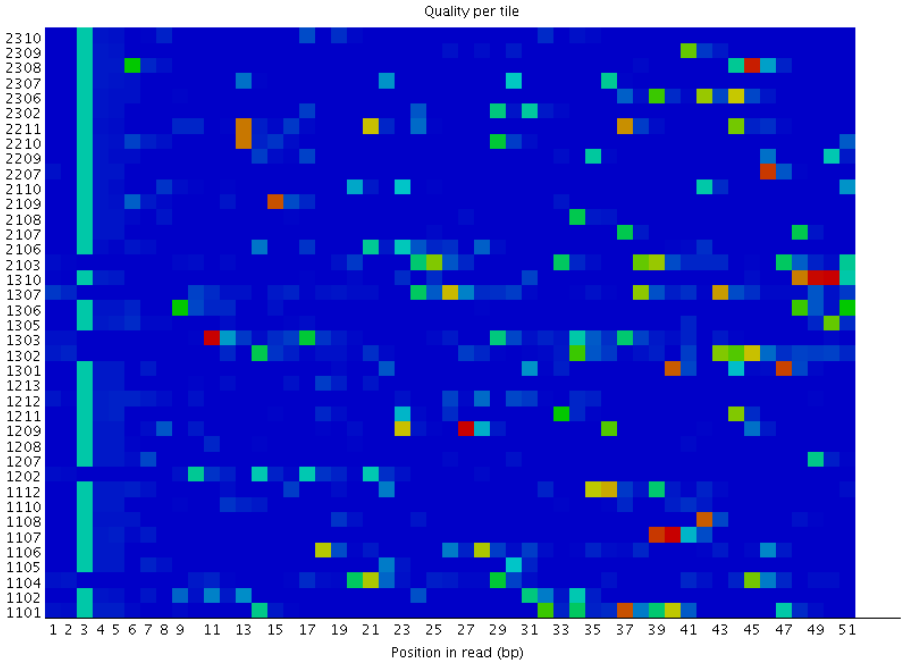


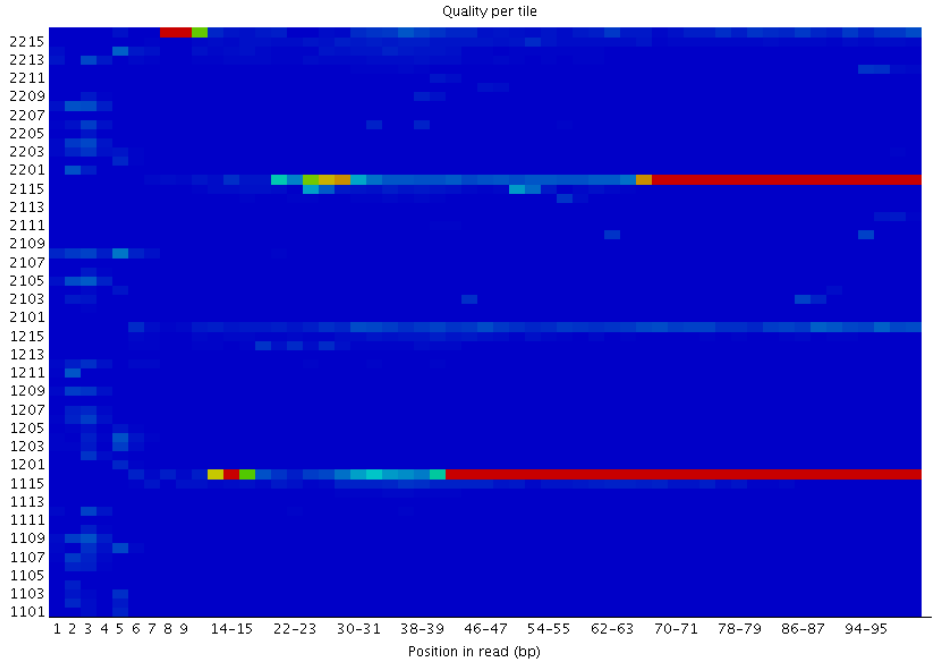


Babraham Bioinformatics

Technical

Positional Phred Scores

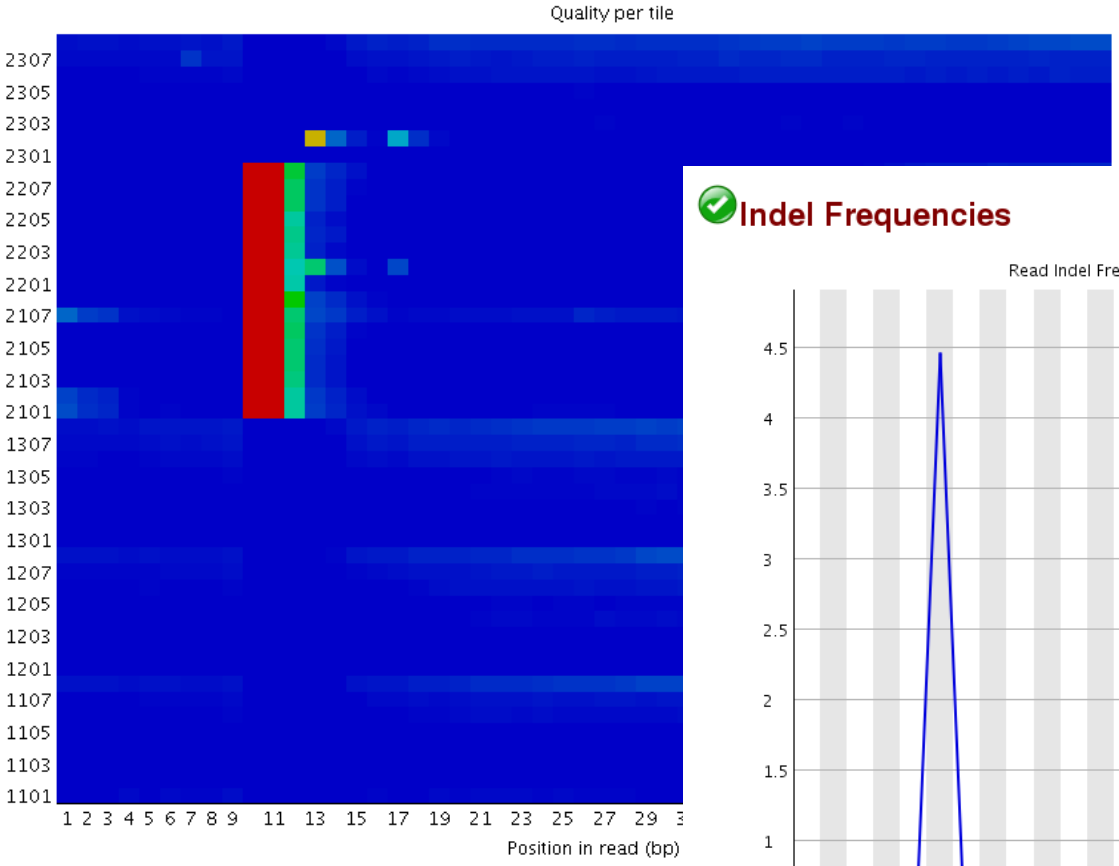


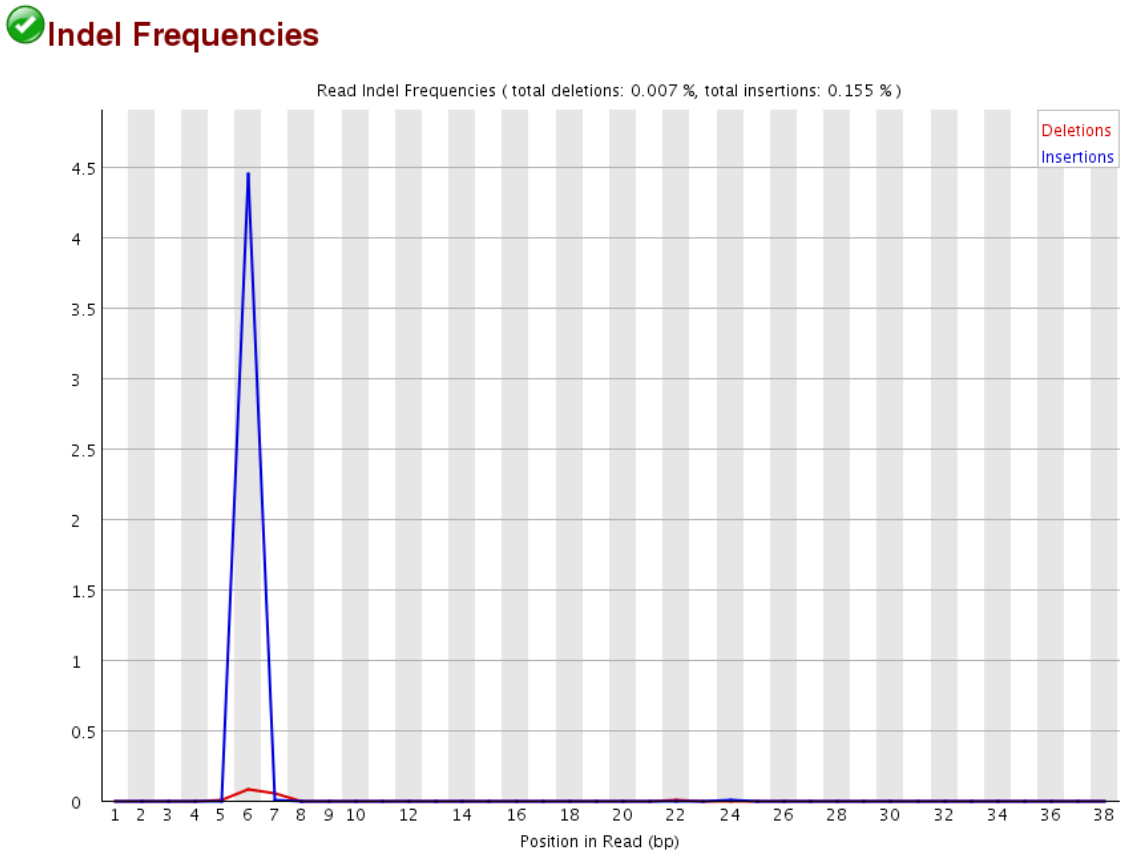


Babraham Bioinformatics

Technical

Positional Phred Scores



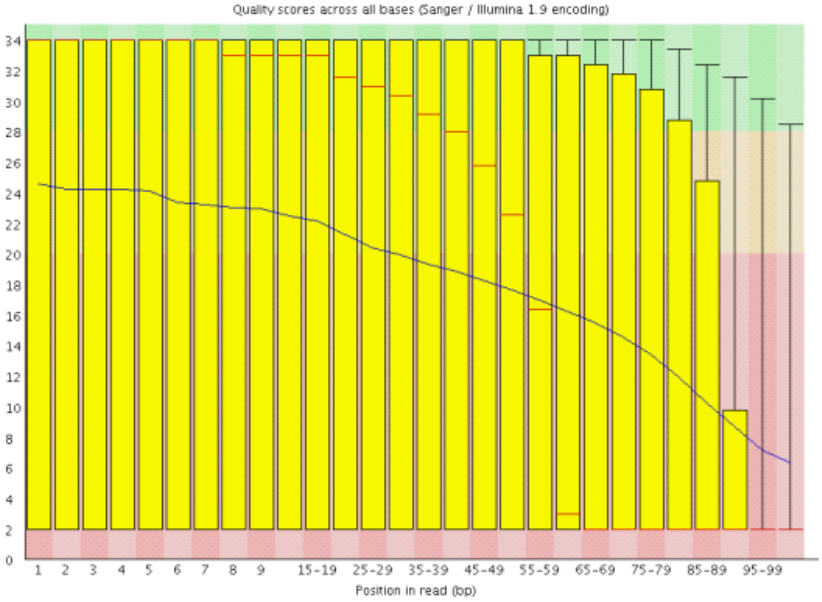


Technical

Biased Phred Scores

HindIII sites

GGATCCGTATGCGATGCTAGCGT
GGATCATATATATGCTAGCGTAT
GGATCTATATTGCGCGATACTGG
GGATCCCGTAGCTGCGATGCTGA
GGATCAAGGATAGCGCGTCTAGA
GGATCTATATAGTTGCCGTATCG
GGATCGGAGCGGGATCGGATGCG

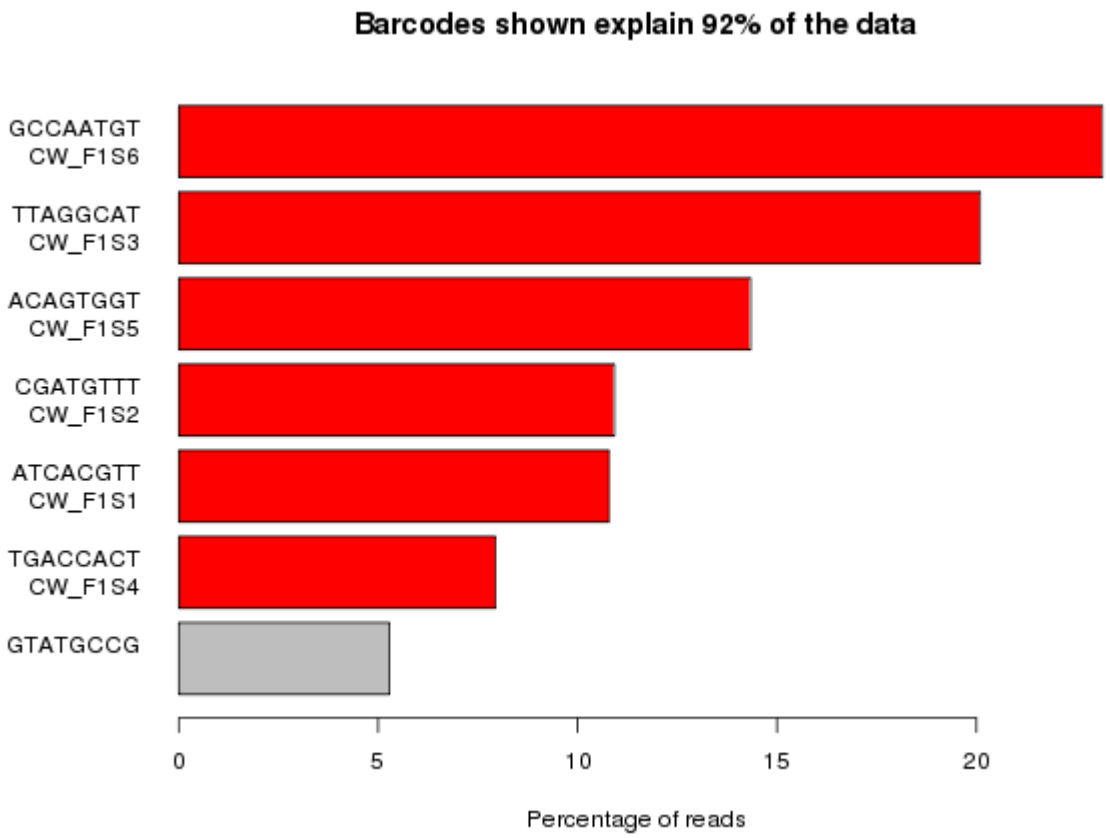


Babraham Bioinformatics

Tracking

Tracking

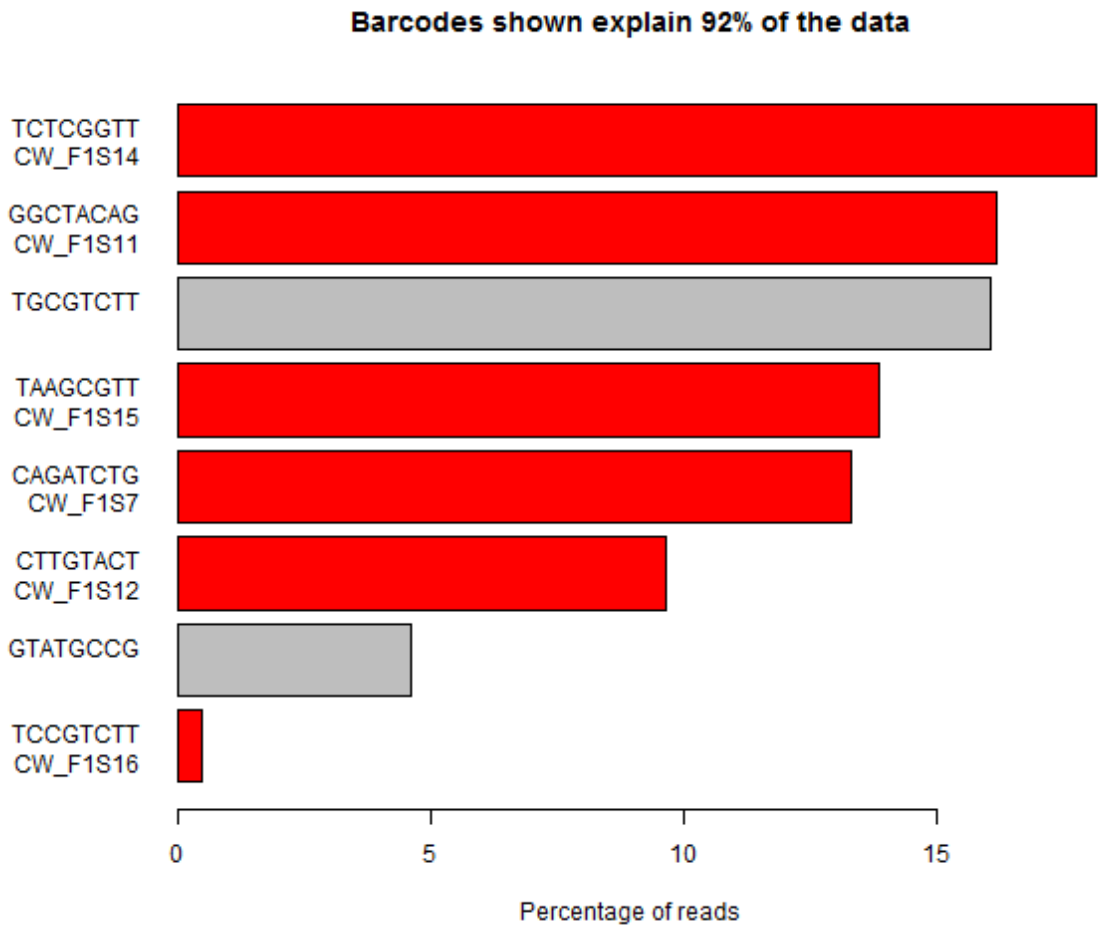
Tracking - Barcodes



Babraham Bioinformatics

Tracking

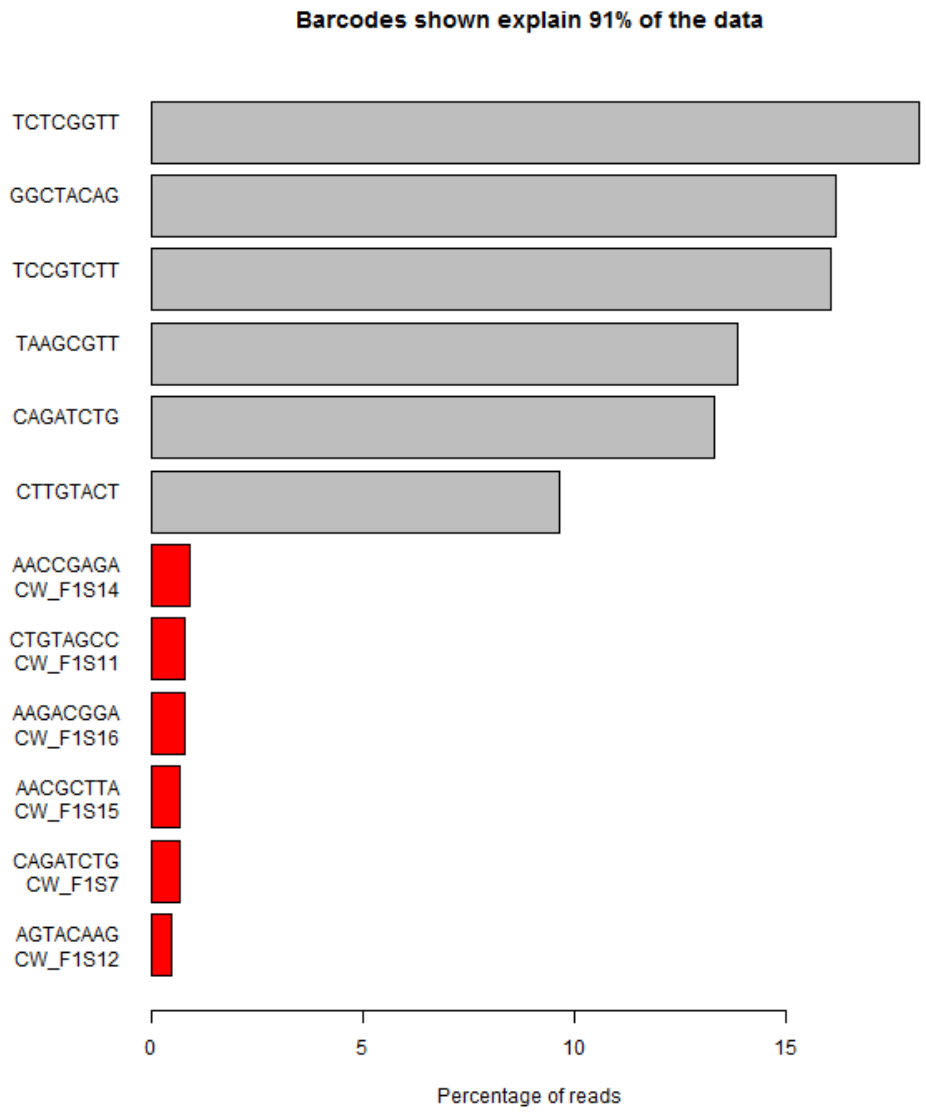
Tracking - Barcodes



Babraham Bioinformatics

Tracking

Tracking - Barcodes



Babraham Bioinformatics

Tracking

Tracking Exercise

You have some barcode statistics from a set of runs from the same group.

Red = Expected barcode
Grey = Unexpected barcode

Can you see if you can spot a problem within this data set,
and say how many of the lanes it might affect?

Babraham Bioinformatics

Tracking

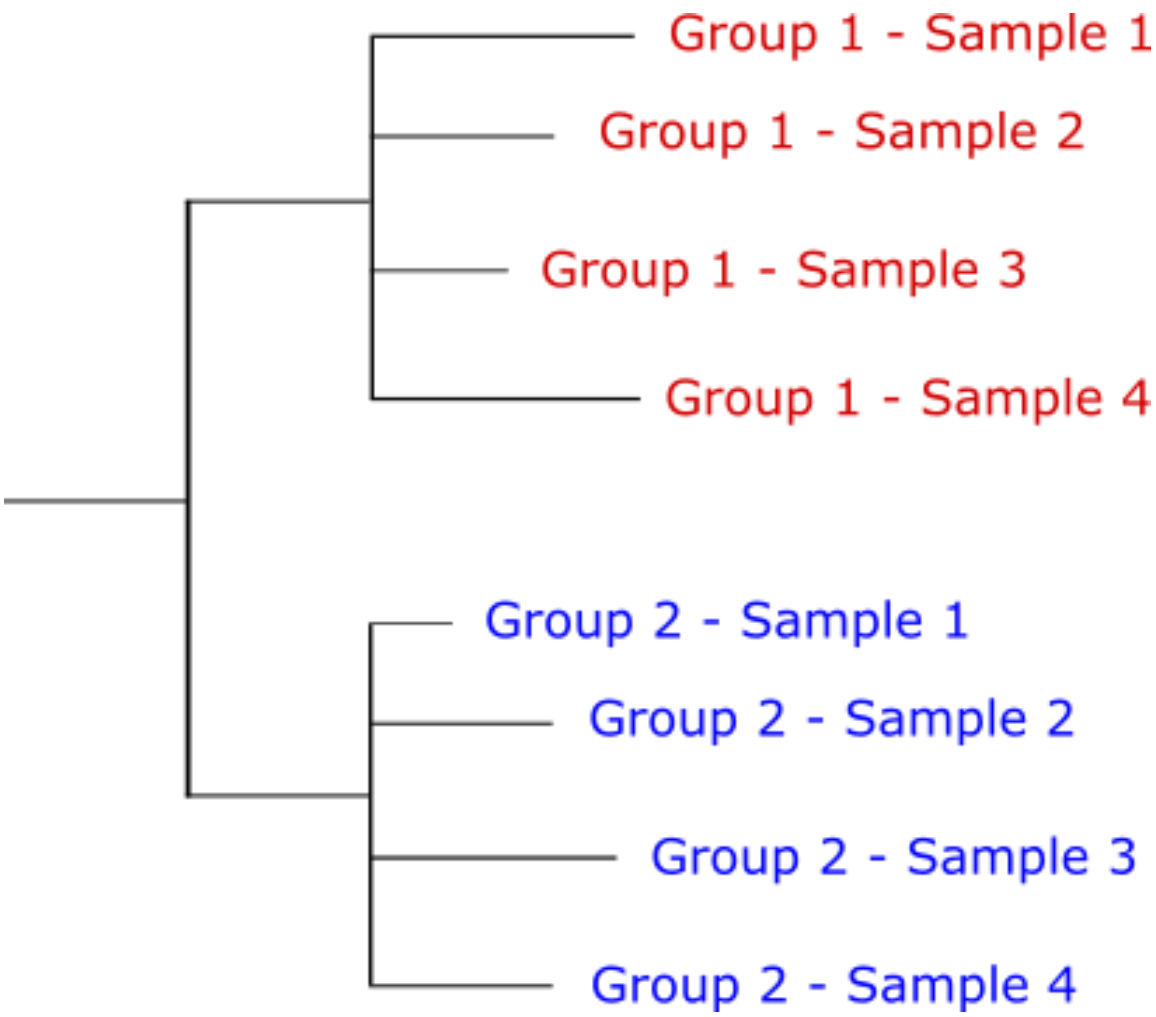
Tracking – Swapped Samples

- Swapped between users
  - Different sample type
  - Different species
  - See later Contamination / Biology sections
- Swapped within experiment
  - Look for consistent biological signal
  - Use other knowledge of the samples to validate

Babraham Bioinformatics

Tracking

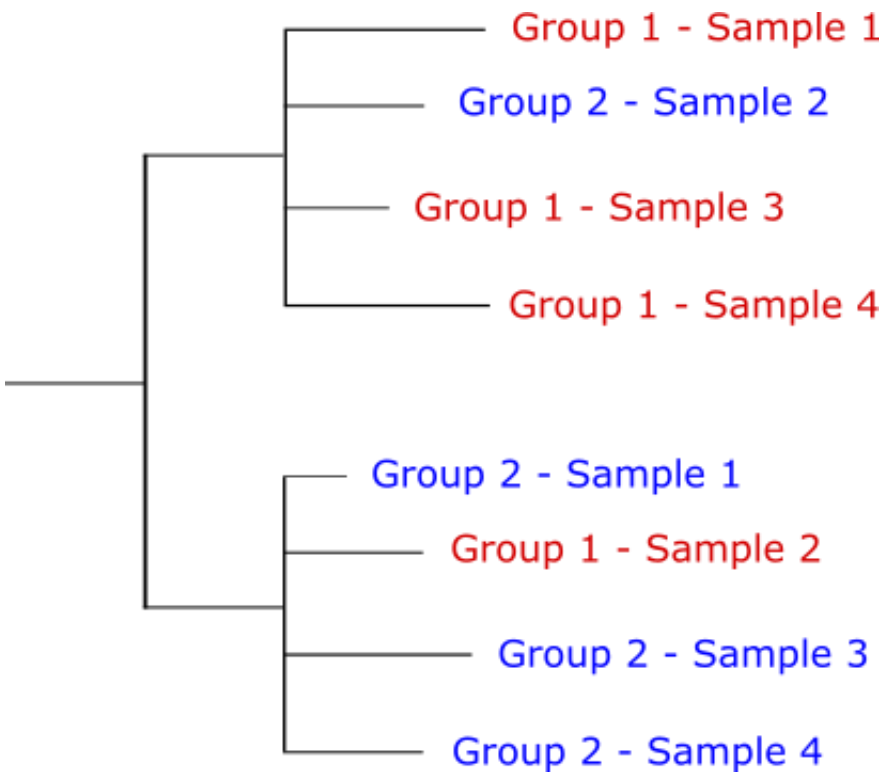
Tracking – Sample groups

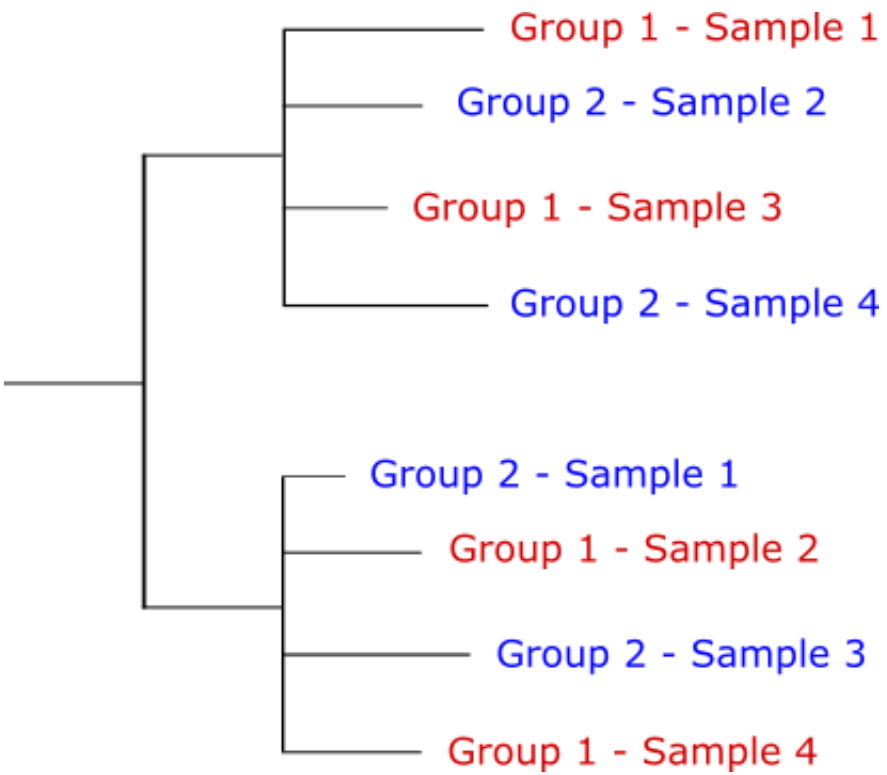


Babraham Bioinformatics

Tracking

Tracking – Sample groups

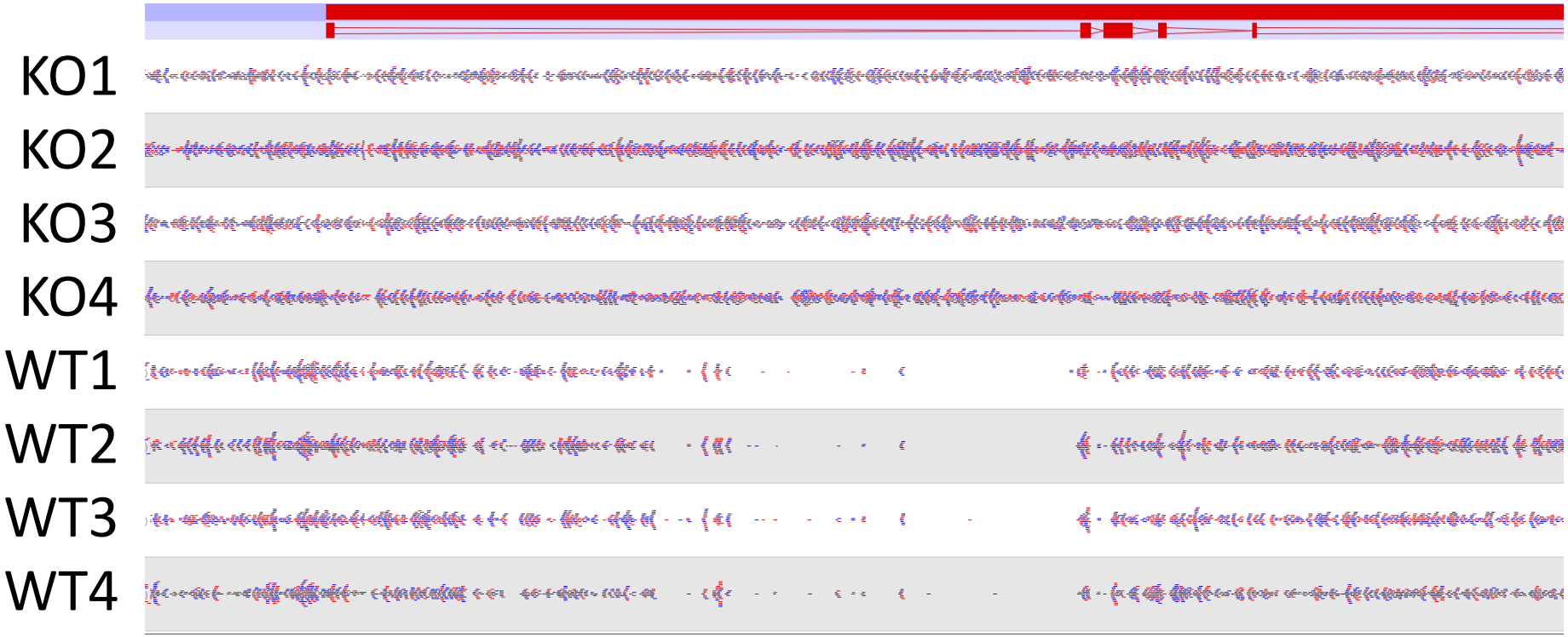




Babraham Bioinformatics

Tracking

Tracking – Sample swaps



Babraham Bioinformatics

Babraham Bioinformatics

Library

Library Problems

- Material Lost
  - Overamplification
    - Duplication
- Biases in selection
  - Priming bias
  - GC bias
  - Methylation bias
  - Size selection bias
- Technical contamination
  - Read through adapter
  - Adapter dimers

Babraham Bioinformatics

Library

Duplication

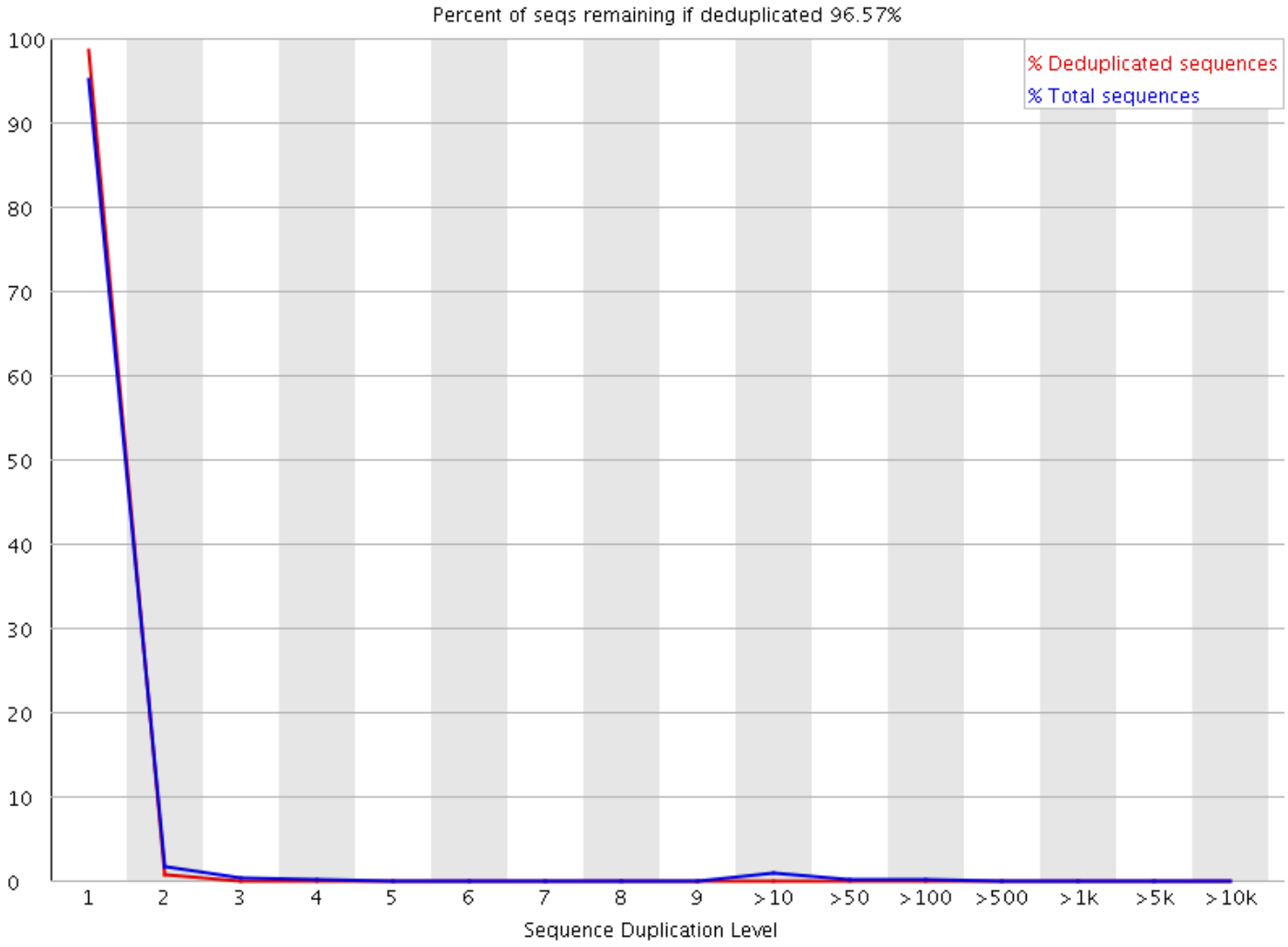
- Over-sequencing of library complexity
- Too little material or too much PCR
- Can be difficult to assess

- Why does duplication matter?
  - Potentially biased
  - Over-estimates measurement accuracy

Babraham Bioinformatics

Library

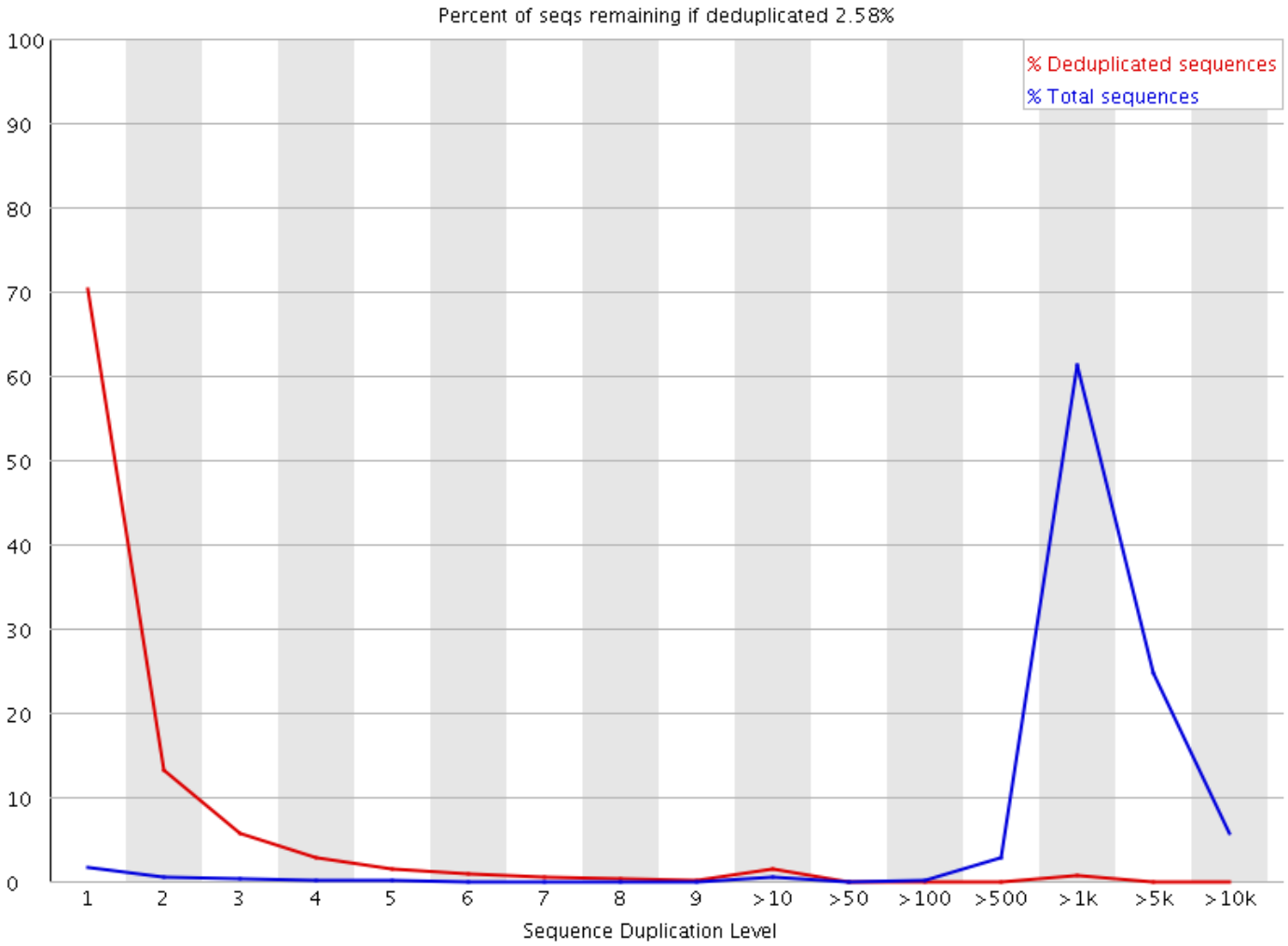
Duplication



Babraham Bioinformatics

Library

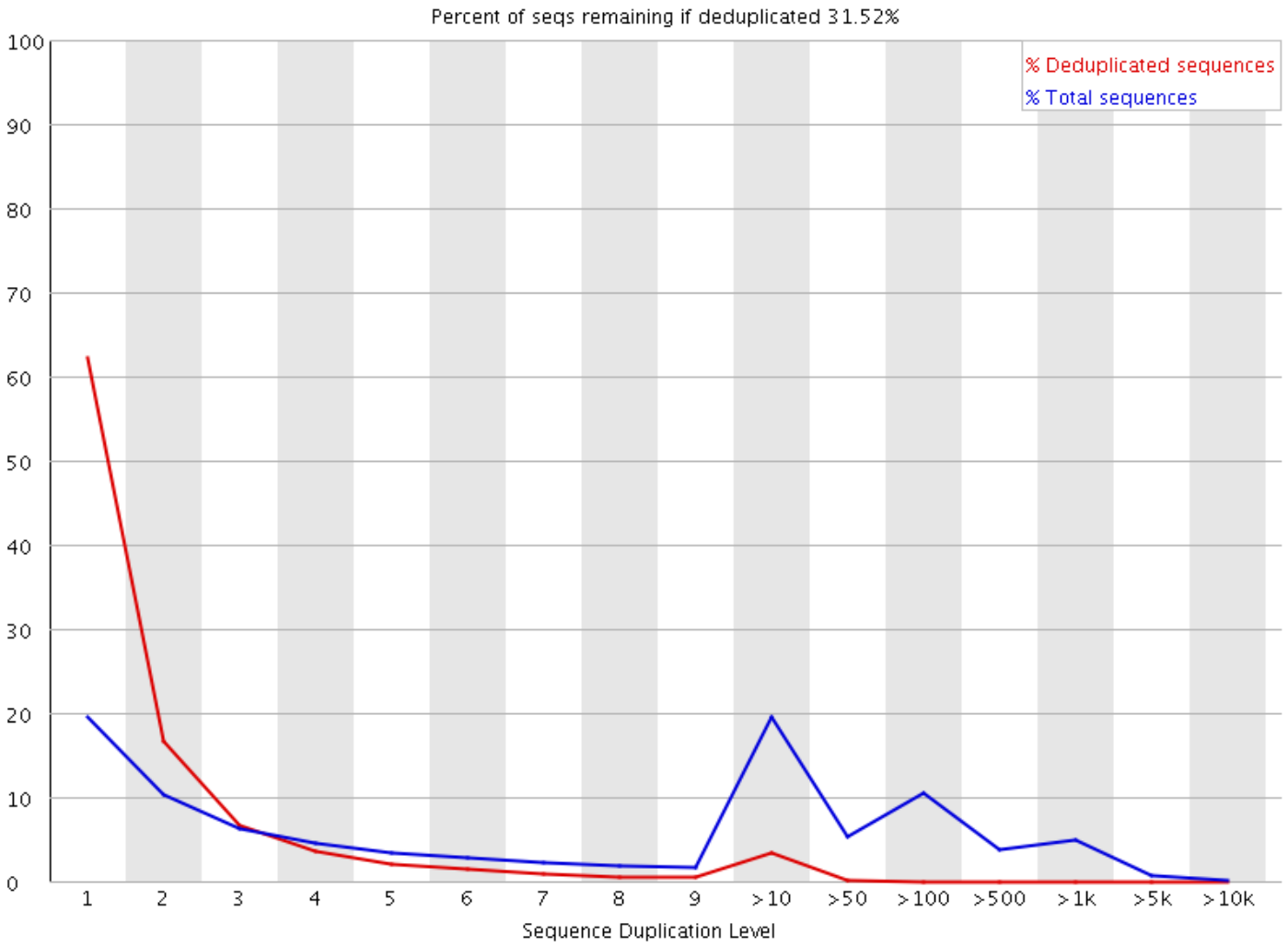
Duplication



Babraham Bioinformatics

Library

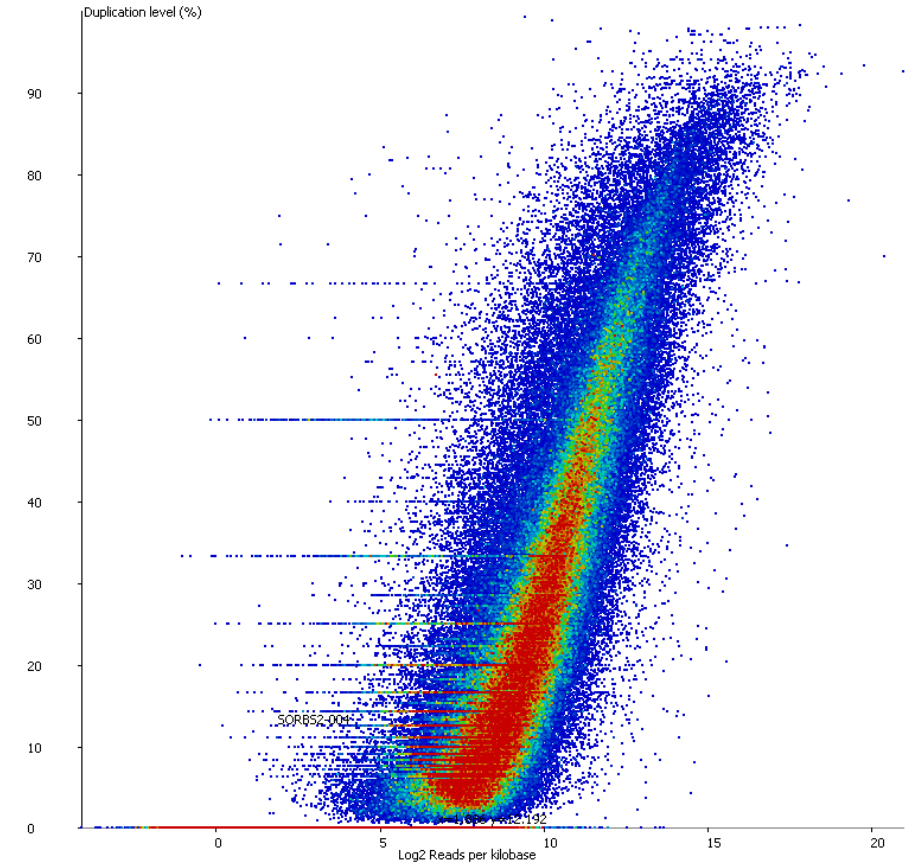
Duplication

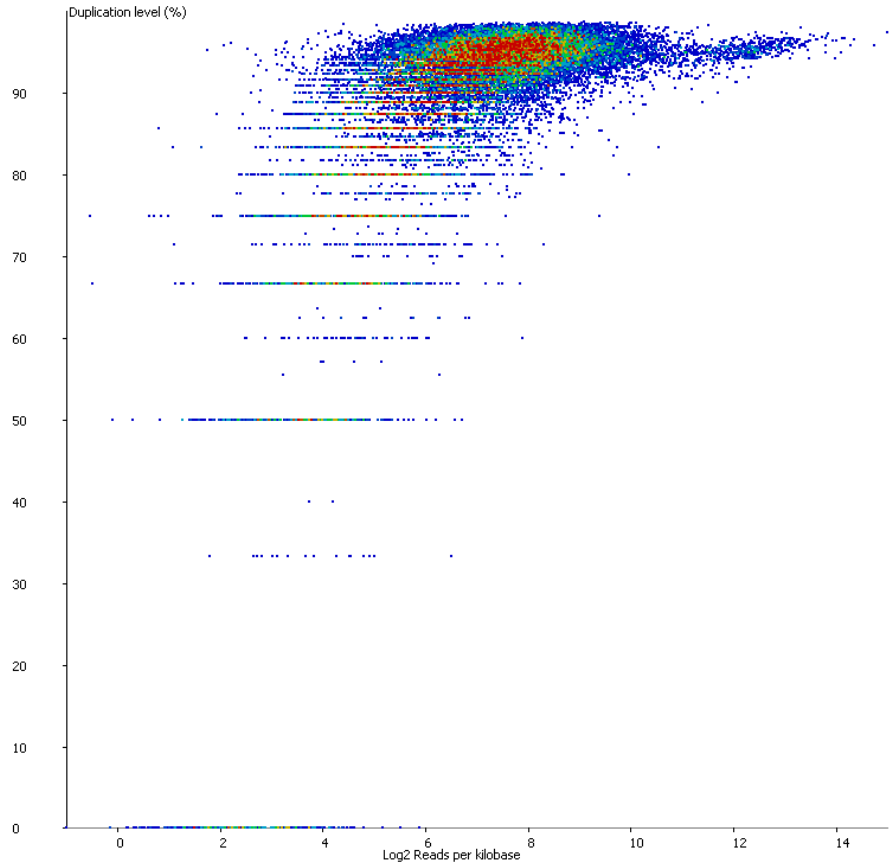


Babraham Bioinformatics

Library

Duplication

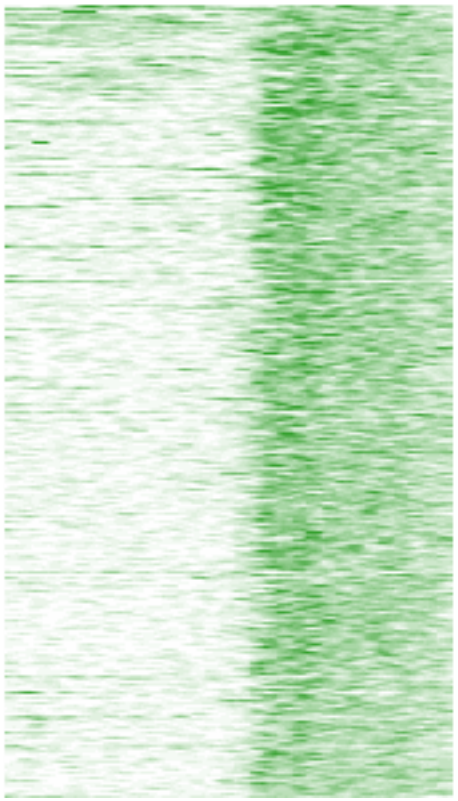


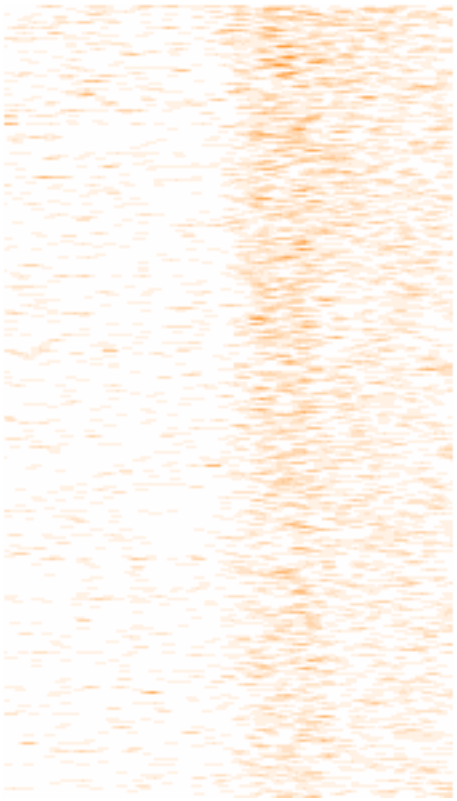


Babraham Bioinformatics

Library

Duplication - Repeats





Repeat

Repeat

Real

Mapped

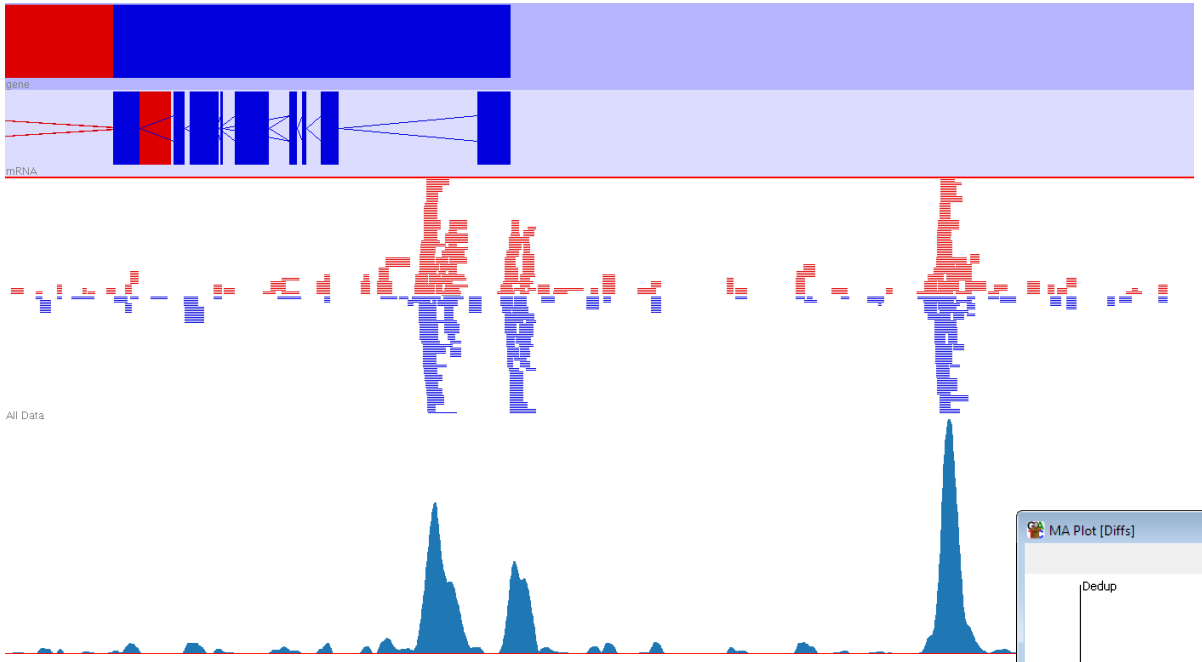
Deduplicated

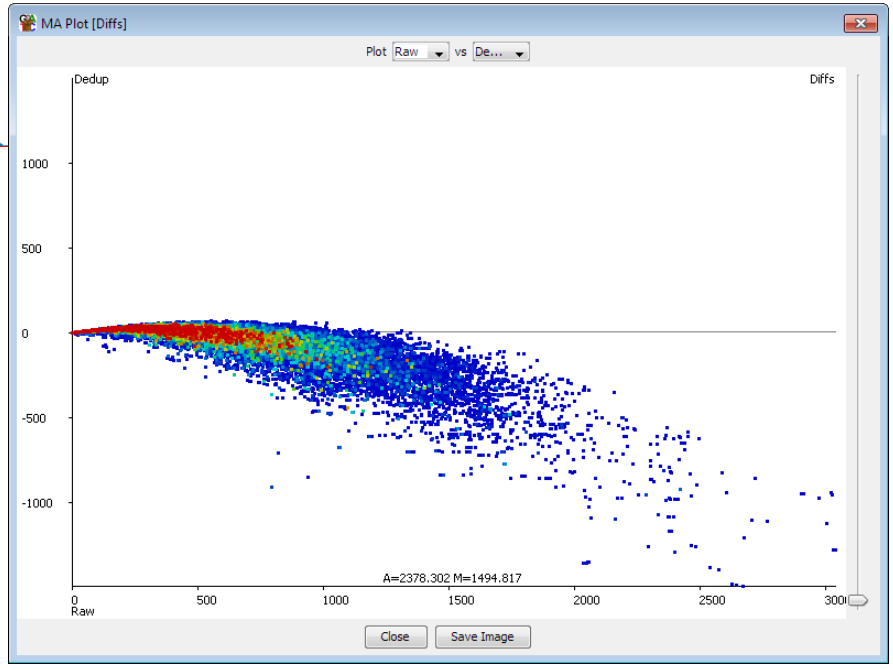
Peak callers (MACS for example) deduplicate internally, so you don't have to consciously do this.

Only avoided by using uniquely mapped reads.

Babraham Bioinformatics

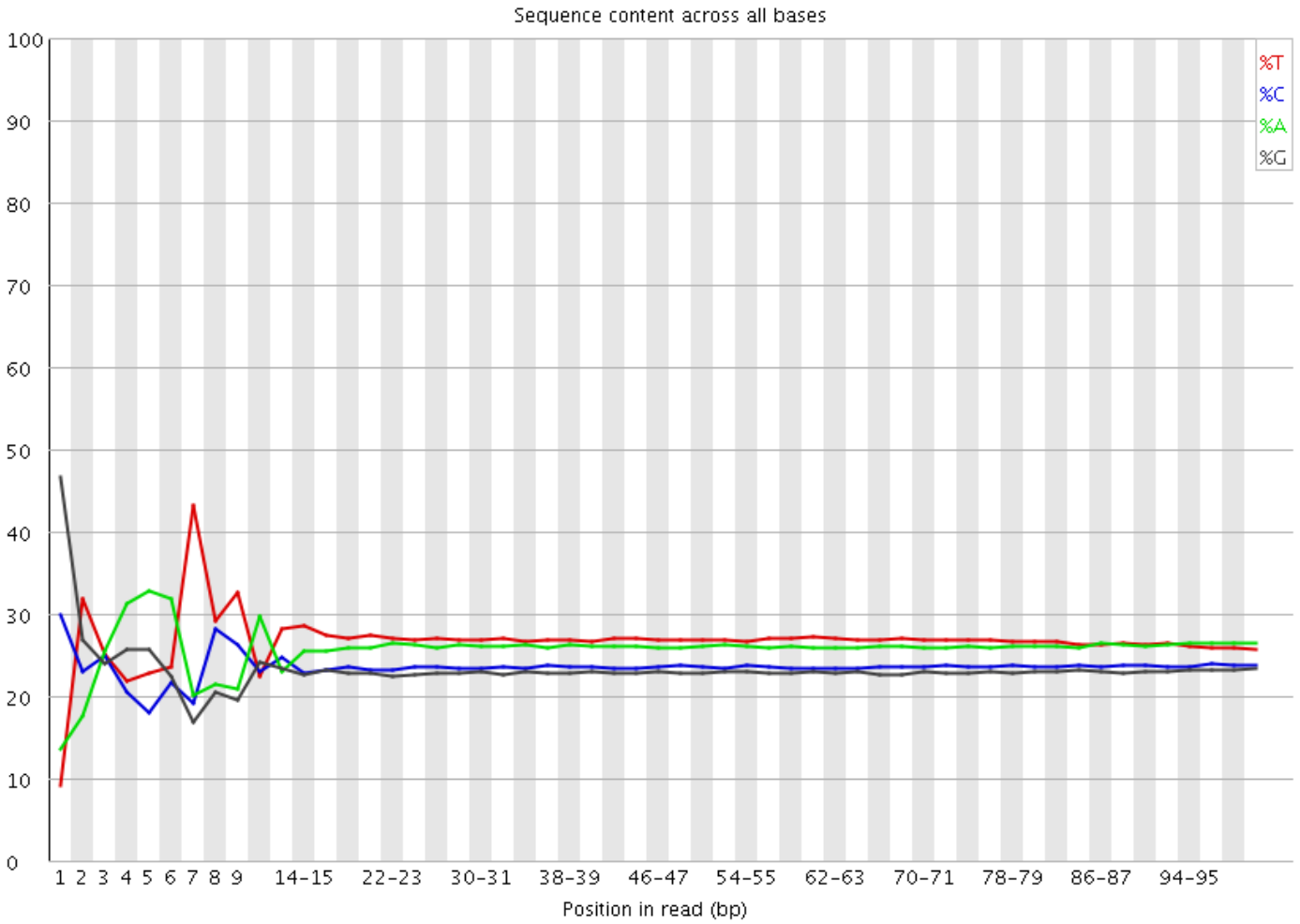
Deduplication





Library

Priming Bias



Babraham Bioinformatics

Contamination

Contamination

Contamination

- Different kinds
  - Technical contamination
    - Adapter dimers
  - Contamination with a species you might expect
    - *E.coli* in a mouse sample
  - Contamination with something unexpected
  - Contamination with the wrong material
    - DNA in an RNA-Prep
  - Mixed samples

Babraham Bioinformatics

Contamination

Mapping Efficiency

- Know what to expect
  - Data type (genomic / transcriptomic)
  - How good / complete is the genome

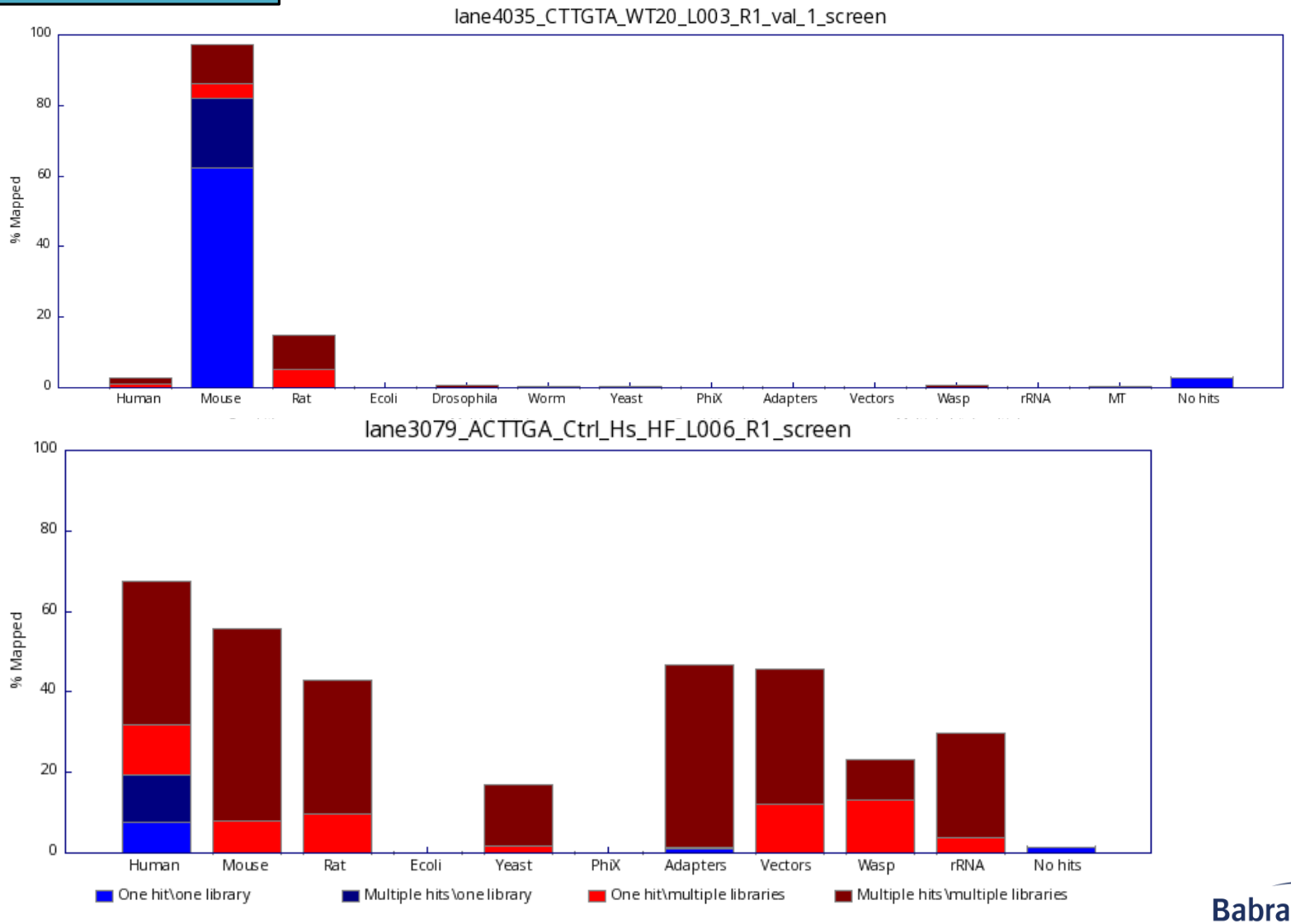
- Distinguish unique / multi-mapped reads
  - Understand the mapping process

```
Reads:
                Input:    5725730
               Mapped:    4703342 (82.1% of input)
            of these:     471516 (10.0%) have multiple  alignments
                                 (471516 have >1)
82.1% overall read alignment rate.
```

Babraham Bioinformatics

Contamination

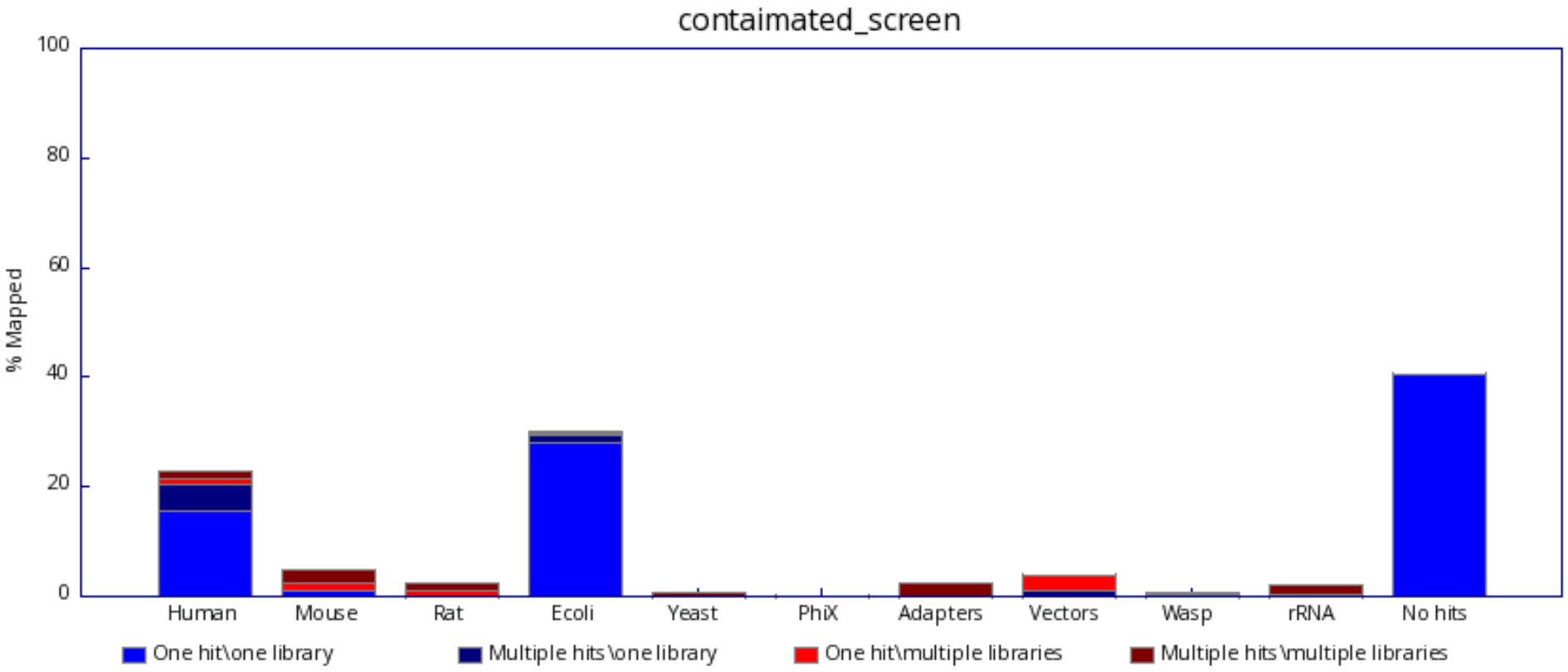
Species Screen



Babraham Bioinformatics

Contamination

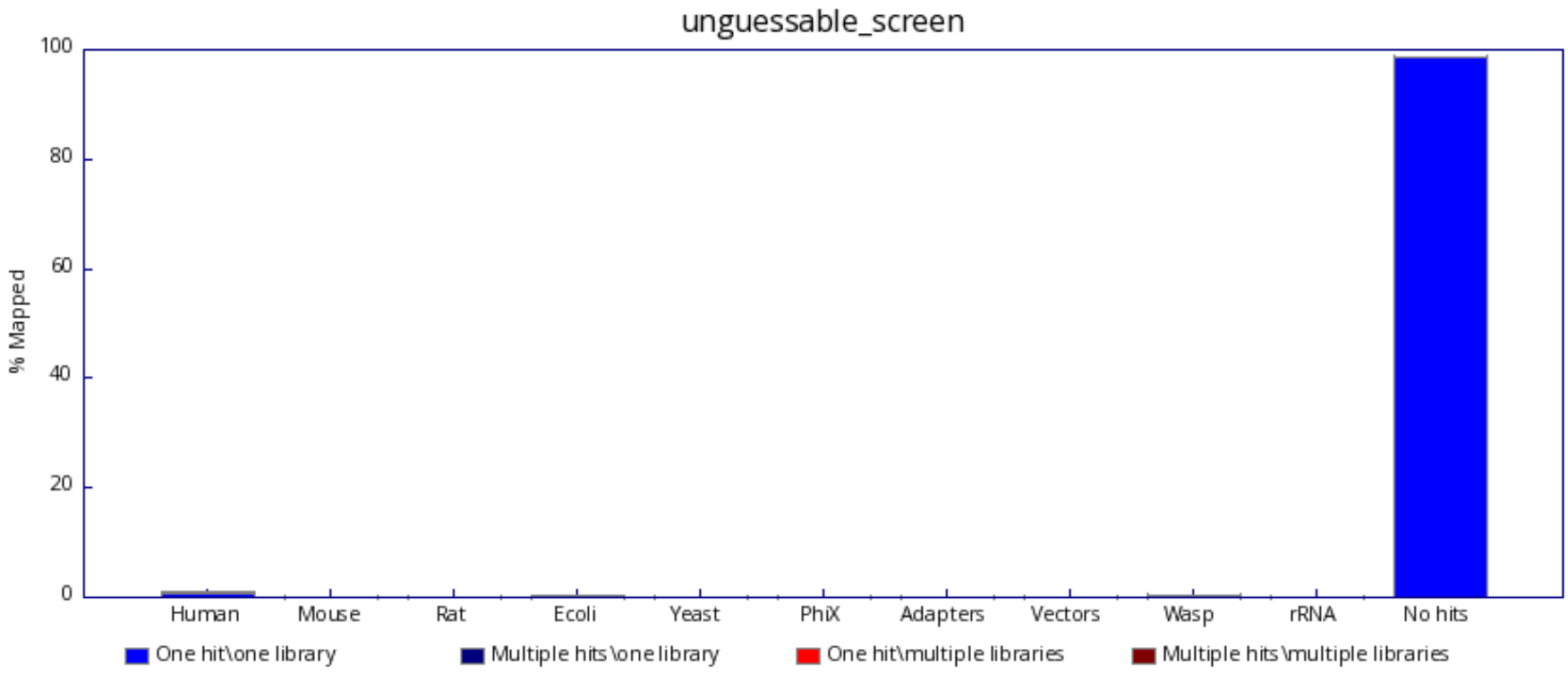
Species Screen



Babraham Bioinformatics

Contamination

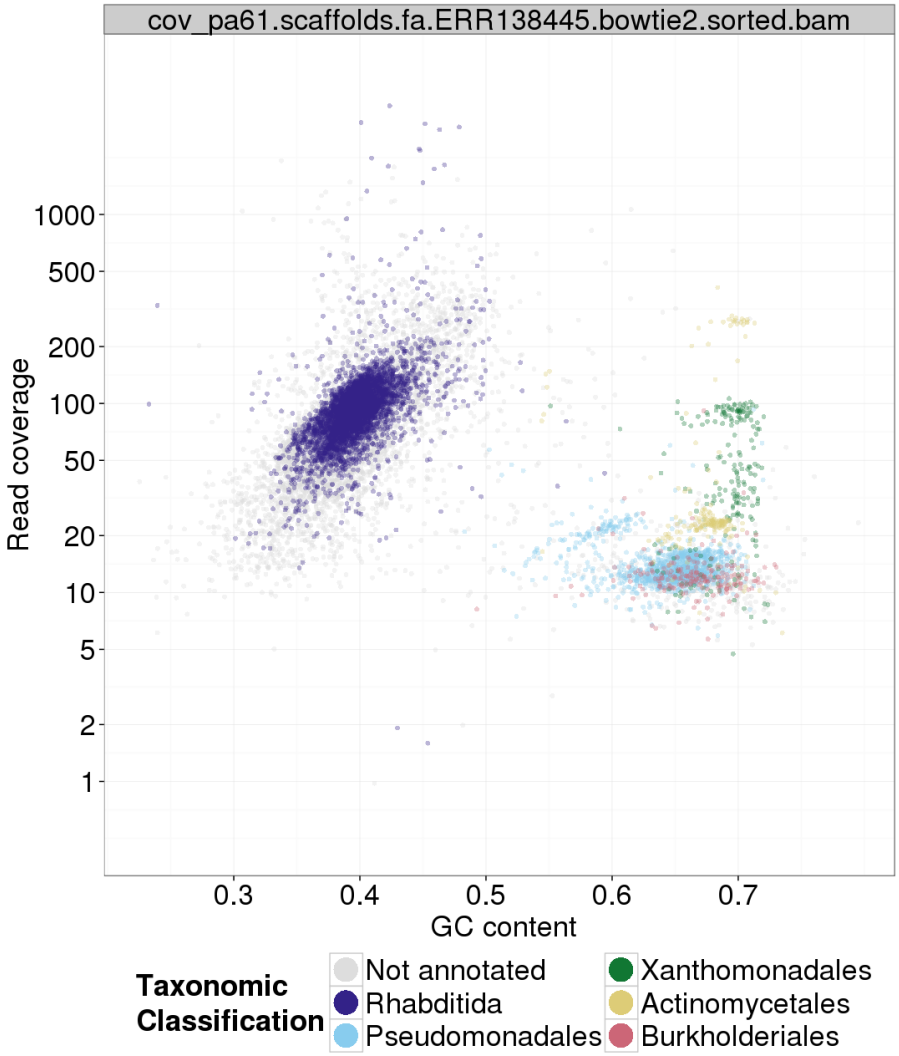
Species Screen



Babraham Bioinformatics

Contamination

TAGC Plots



Assemble

↓

Filter contigs

↓

Plot %GC vs Coverage

↓

Sample and blast

Babraham Bioinformatics

Contamination

Contamination

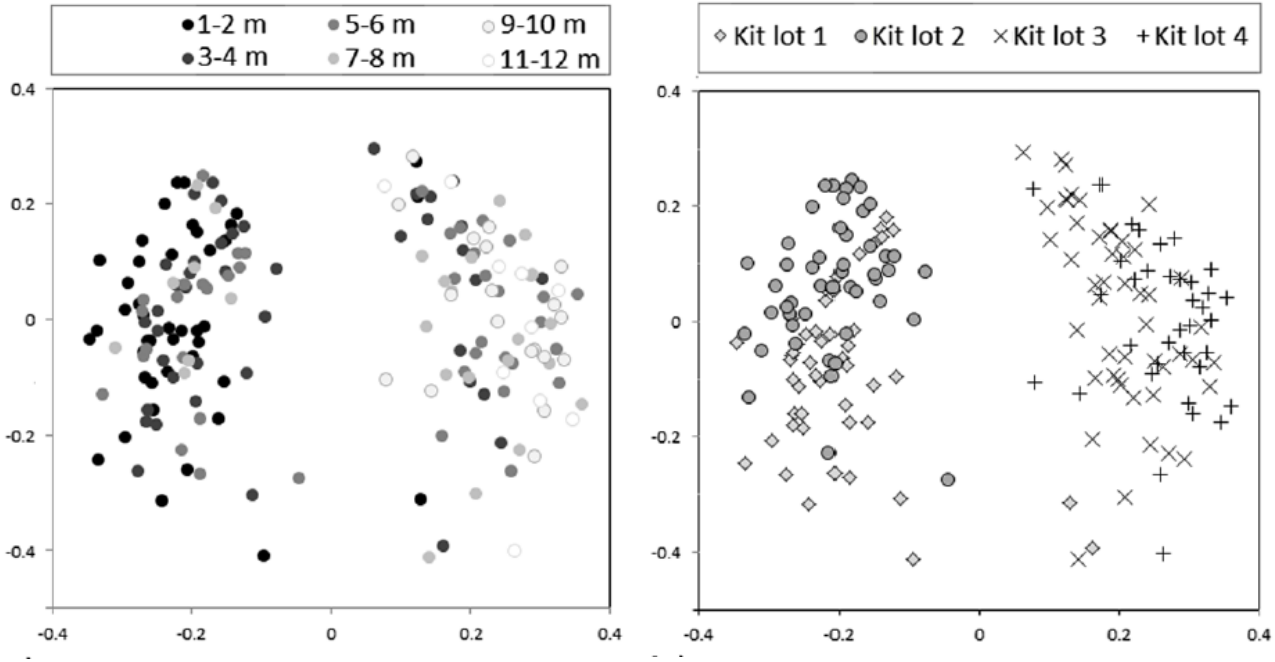
Salter *et al. BMC Biology* 2014, **12**:87
http://www.biomedcentral.com/1741-7007/12/87

BMC Biology

RESEARCH ARTICLE Open Access

Reagent and laboratory contamination can critically impact sequence-based microbiome analyses

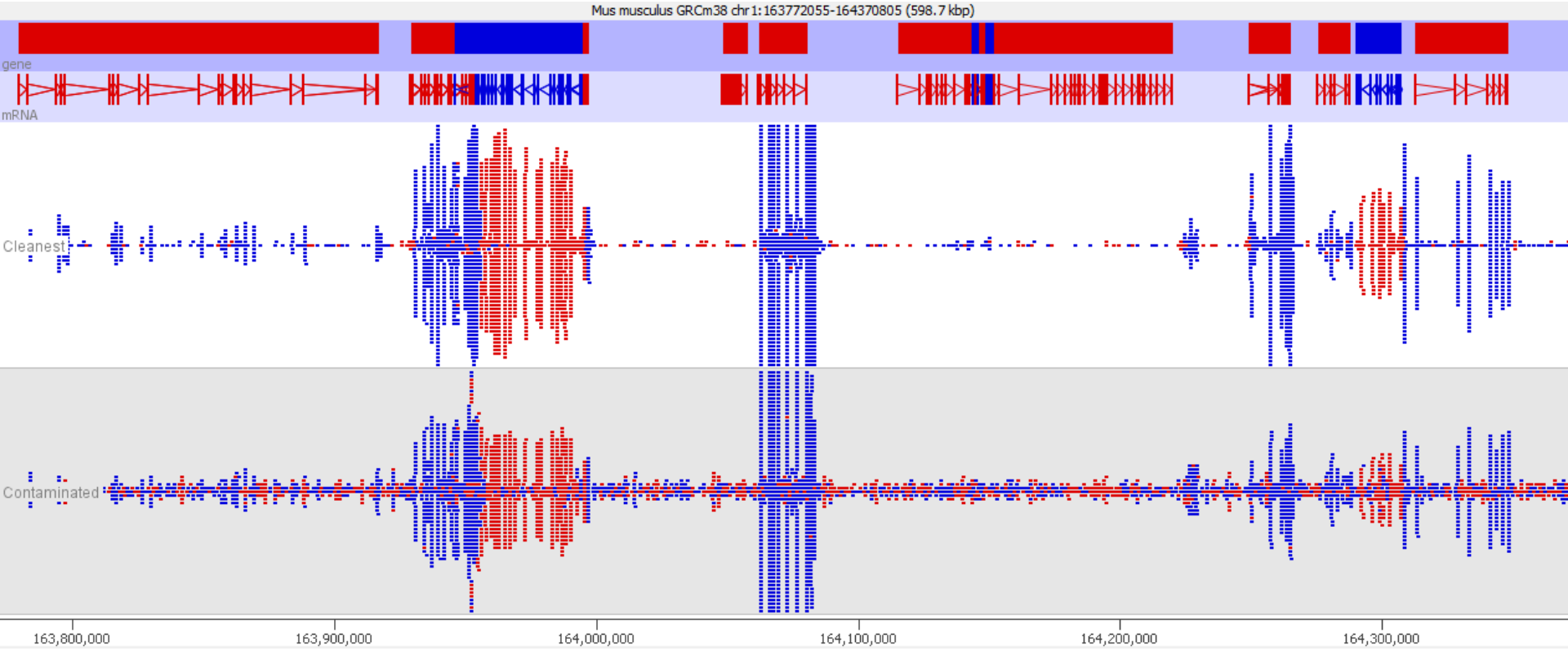
Susannah J Salter[1*], Michael J Cox[2], Elena M Turek[2], Szymon T Calus[3], William O Cookson[2], Miriam F Moffatt[2], Paul Turner[4,5], Julian Parkhill[1], Nicholas J Loman[3] and Alan W Walker[1,6*]



Babraham Bioinformatics

Contamination

Internal Contamination



Babraham Bioinformatics

Biological

Babraham Bioinformatics

Biological

Samples Don't Behave

- All samples come with a set of expectations
  - Biological effect
  - Sample source
  - Rough biological behaviour
- If these aren't met
  - Samples may not be what you expect
  - Statistical analyses may be invalid
  - Larger biological picture may be missed

Babraham Bioinformatics

Biological

Exercise

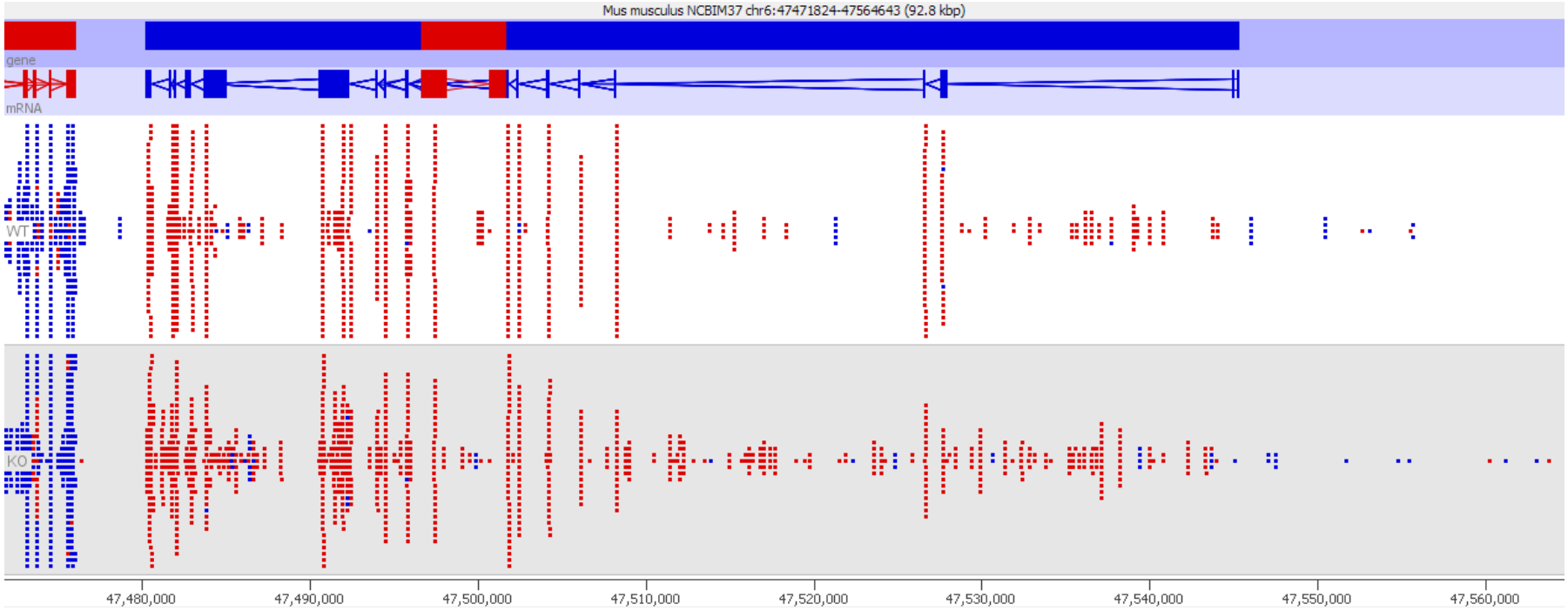
You have been given a set of QC and visualisation results for a knockout in male black6 mice (same genotype as the reference) of a single gene.

Have a look through the plots and see if there is anything which would cause you concern regarding the behaviour of the samples.

Babraham Bioinformatics

Biological

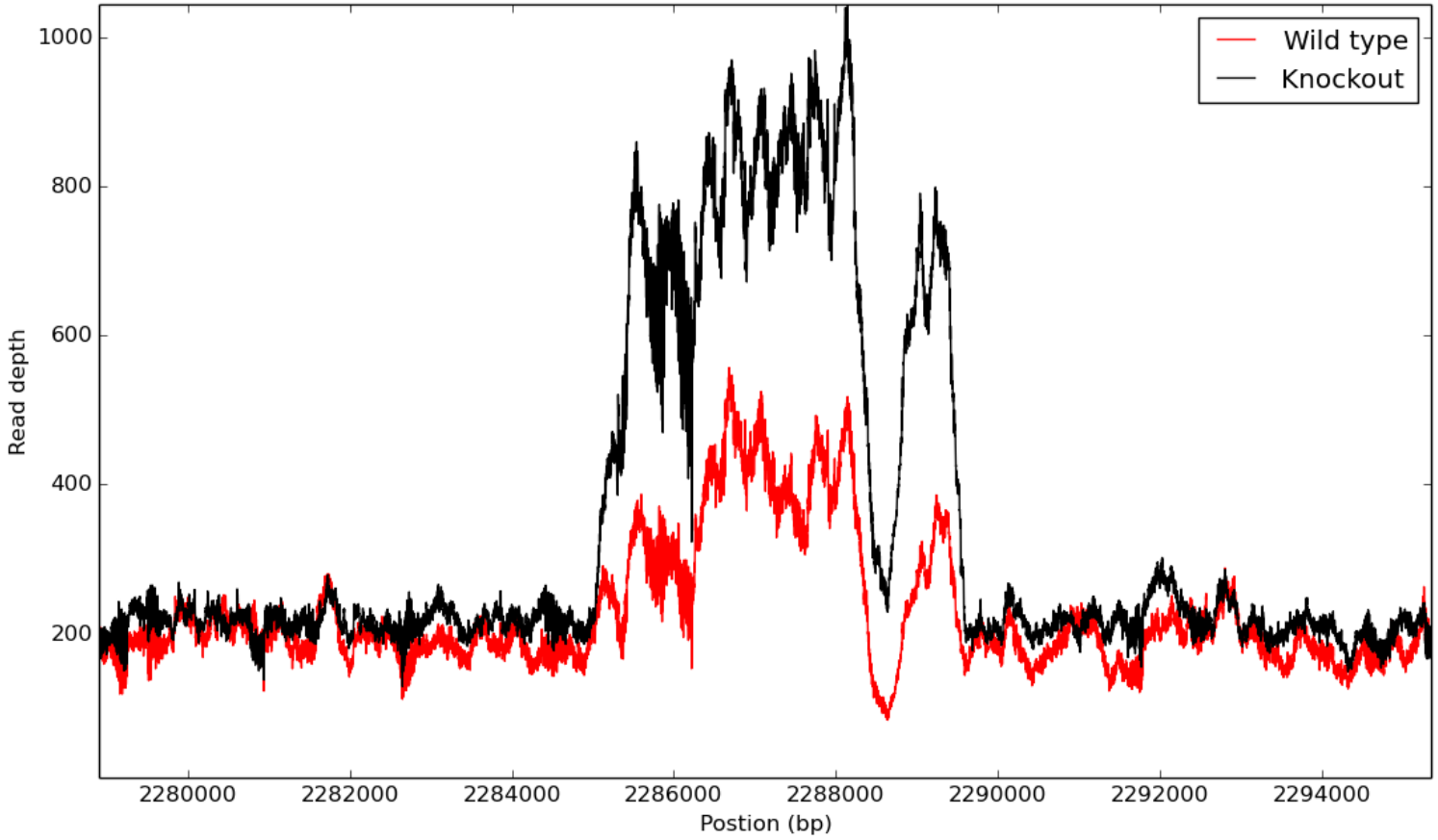
Expected effects missing



Babraham Bioinformatics

Biological

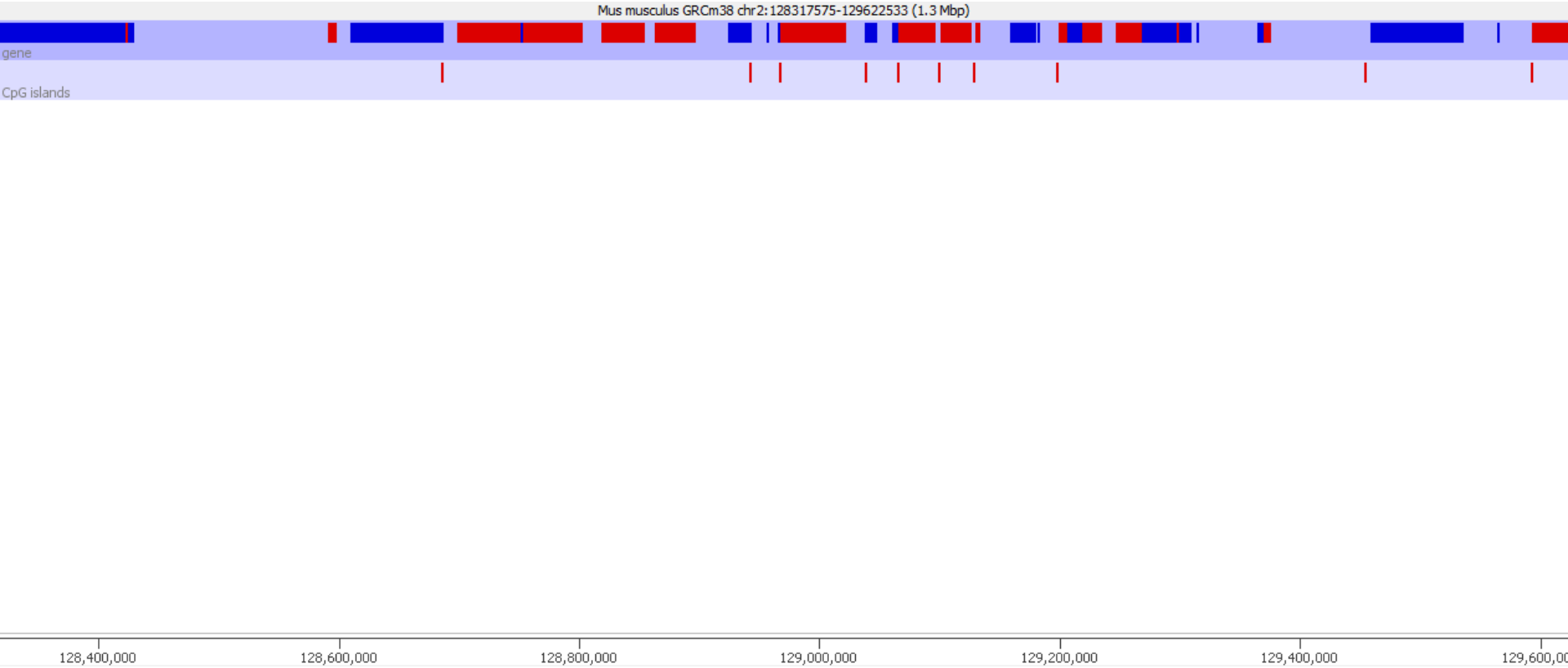
Confounded effects



Babraham Bioinformatics

Biological

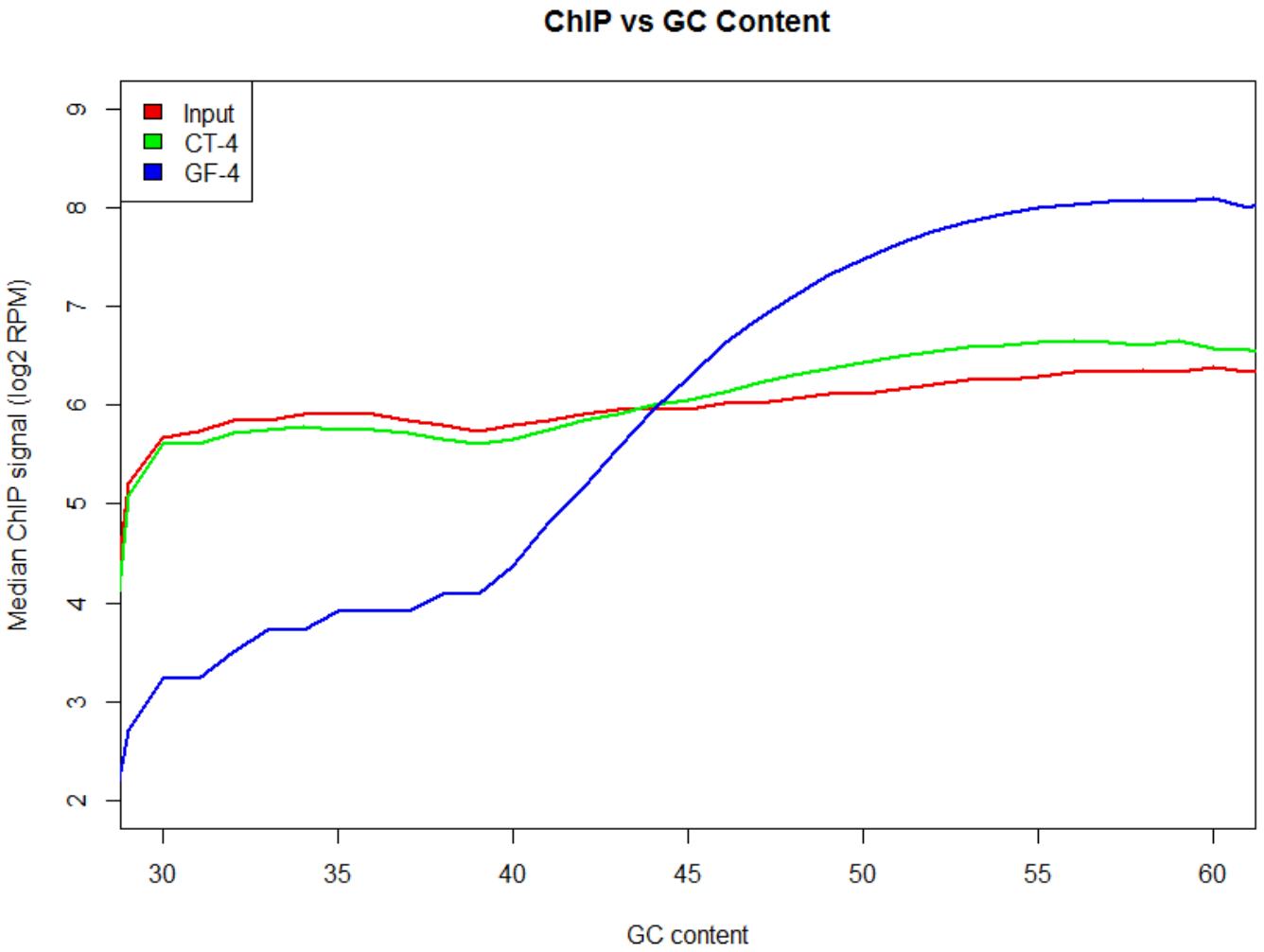
ChIP doesn't behave



Babraham Bioinformatics

Biological

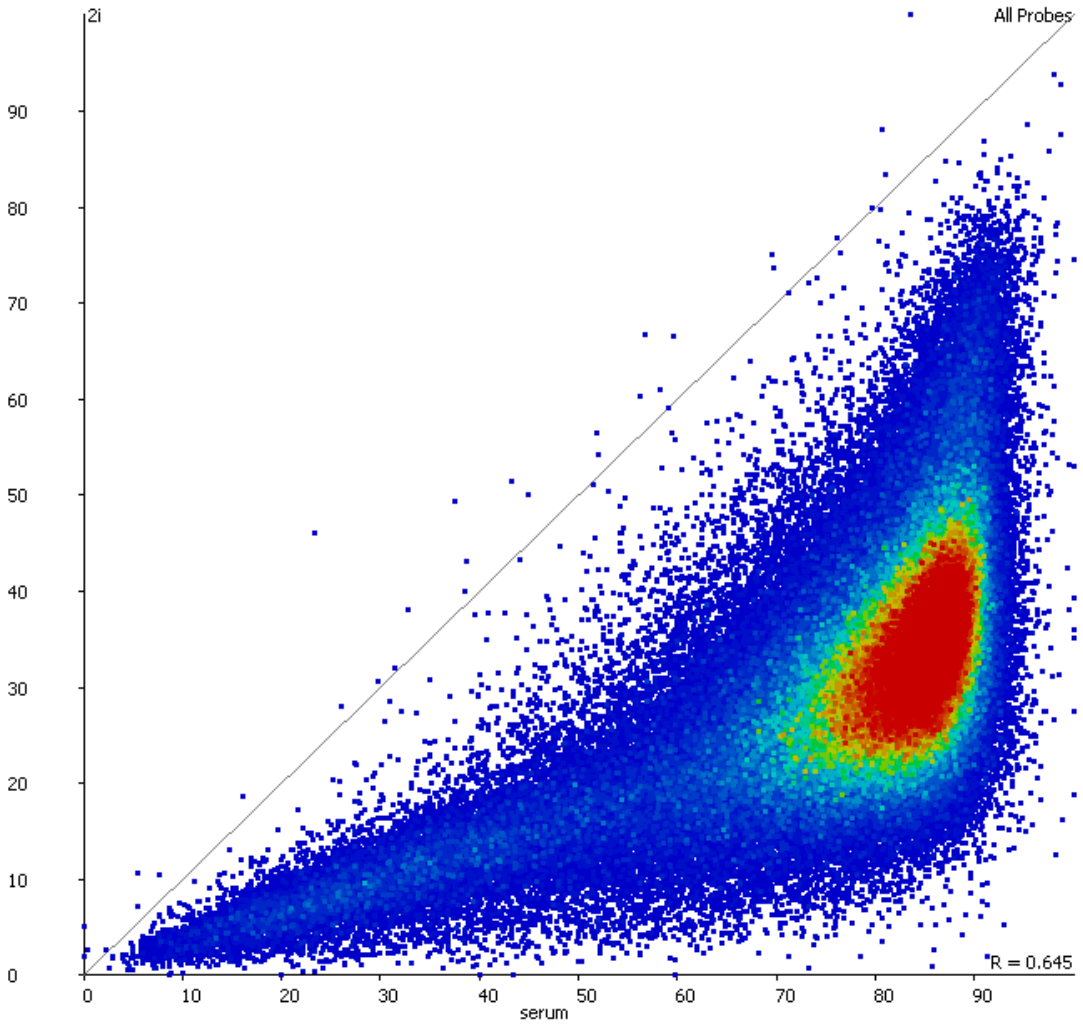
ChIP doesn't behave



Babraham Bioinformatics

Biological

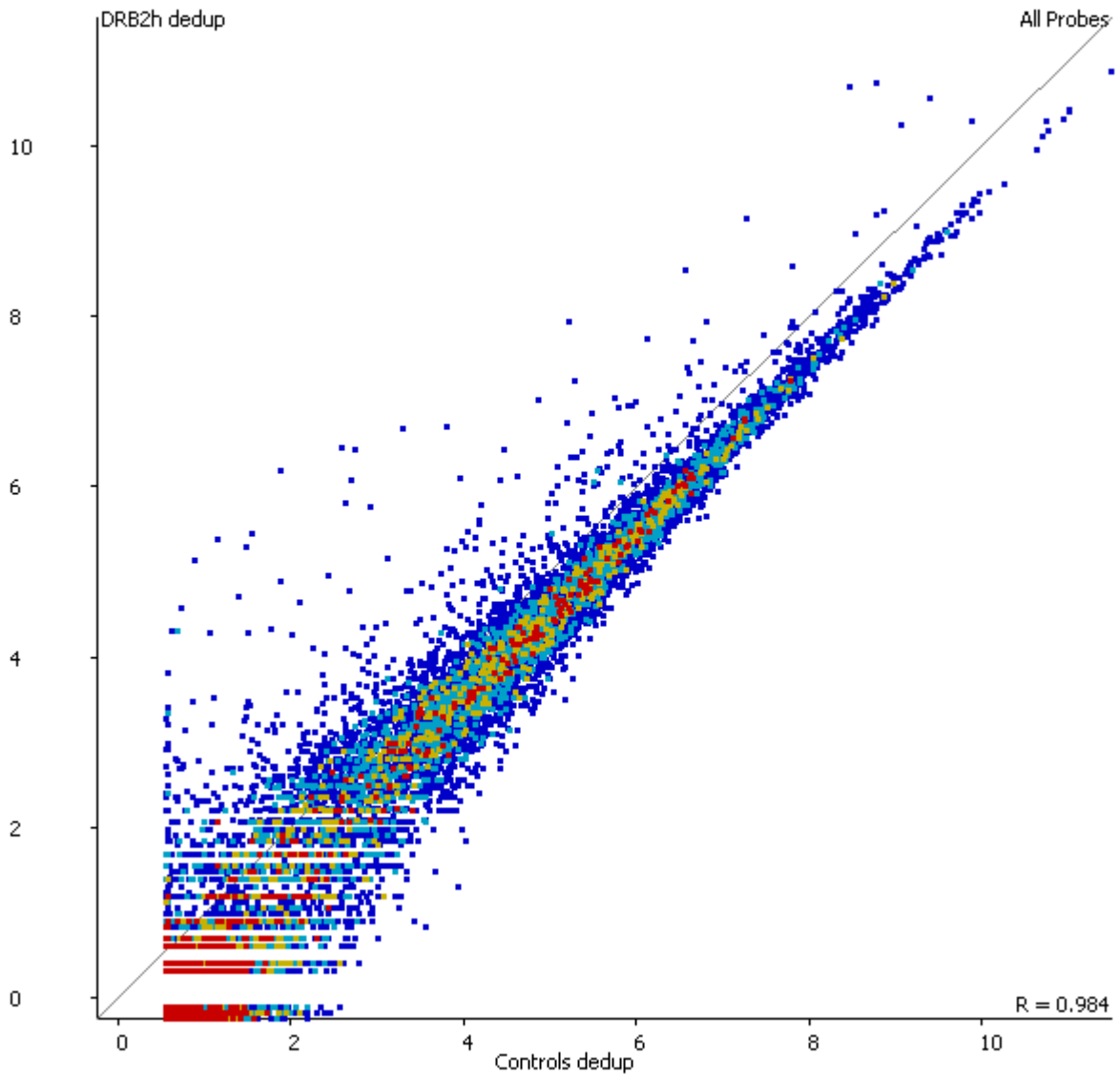
Methylation doesn't behave



Babraham Bioinformatics

Biological

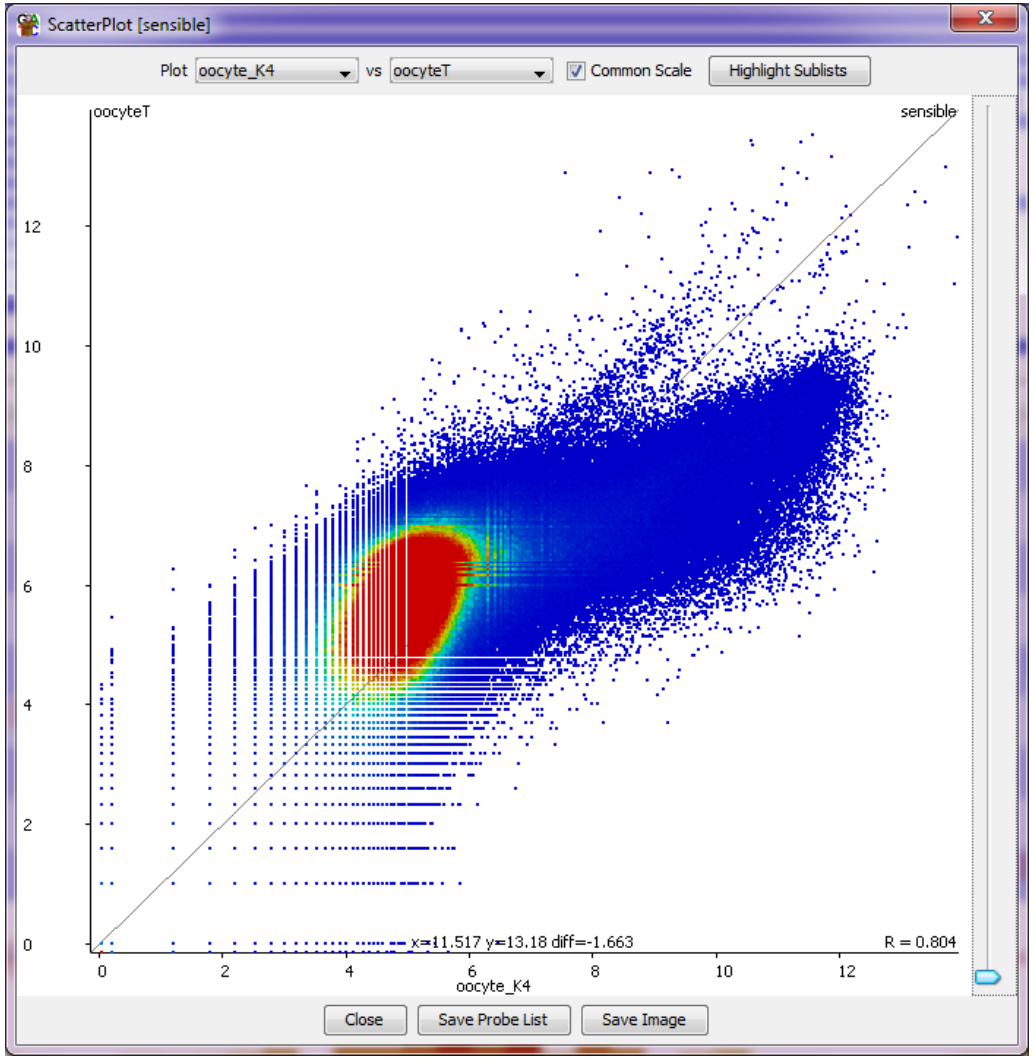
RNA-Seq doesn't behave



Babraham Bioinformatics

Biological

Multiple subgroups



Babraham Bioinformatics

Interpretation

Babraham Bioinformatics