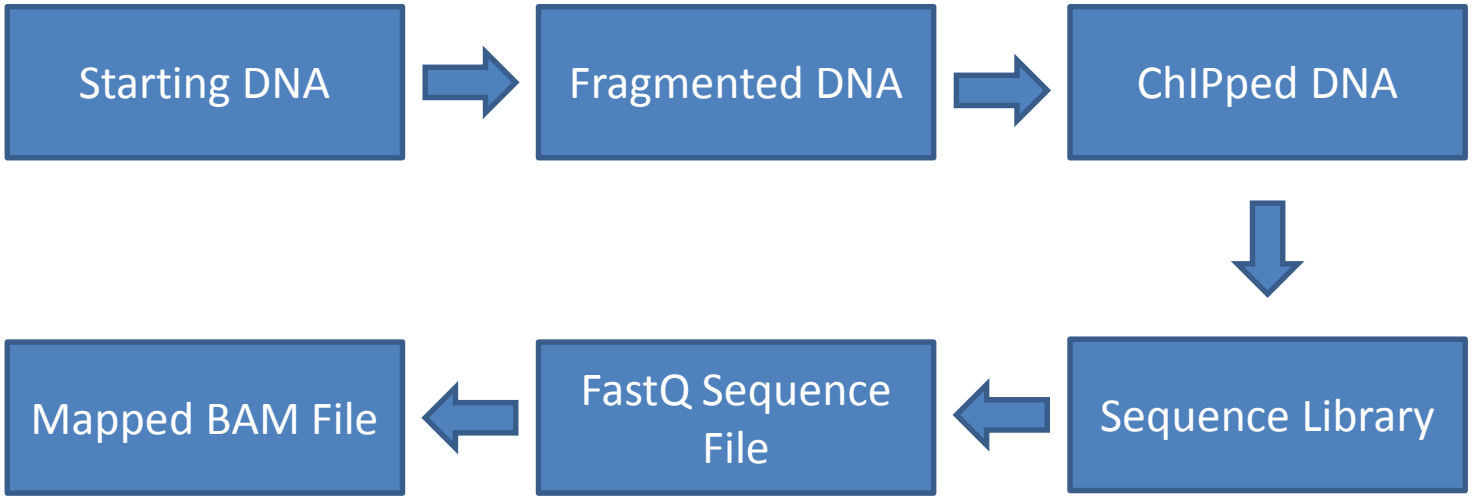
ChIP-Seq Data Processing and QC

Simon Andrews
simon.andrews@babraham.ac.uk
@simon_andrews
v2018-01

Babraham Bioinformatics

Data Creation and Processing



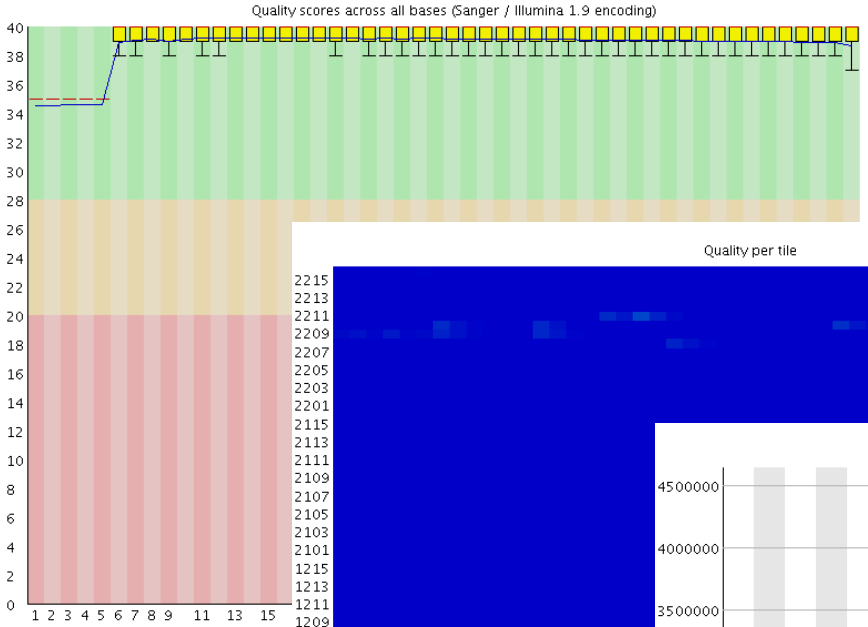
Babraham Bioinformatics

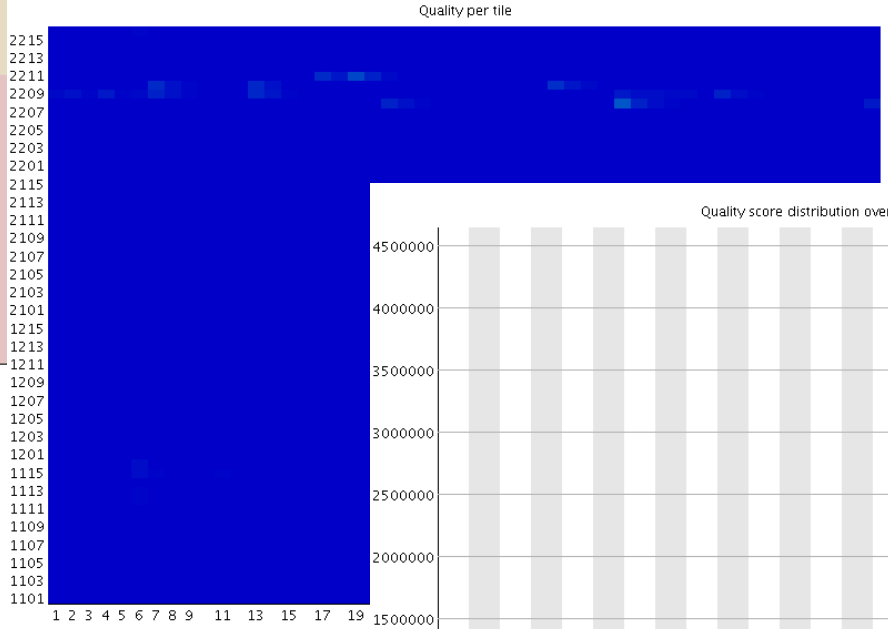
A typical ChIP Library

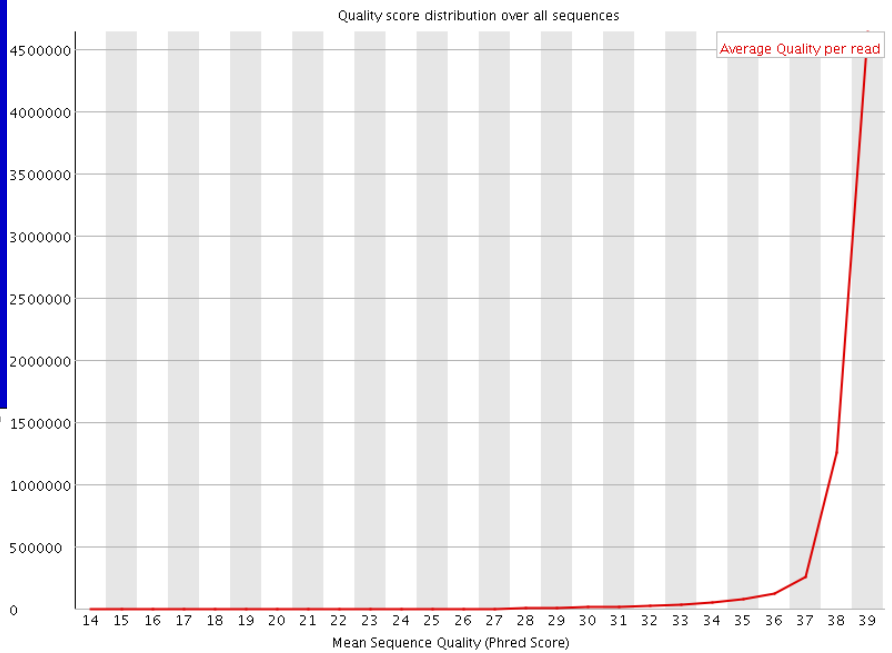
- Potential technical problems
  - Adapter contamination
  - PCR Duplication
- Potential biological problems
  - Lack of enrichment
  - Other selection biases

Babraham Bioinformatics

QC of raw sequence
Base Call Quality

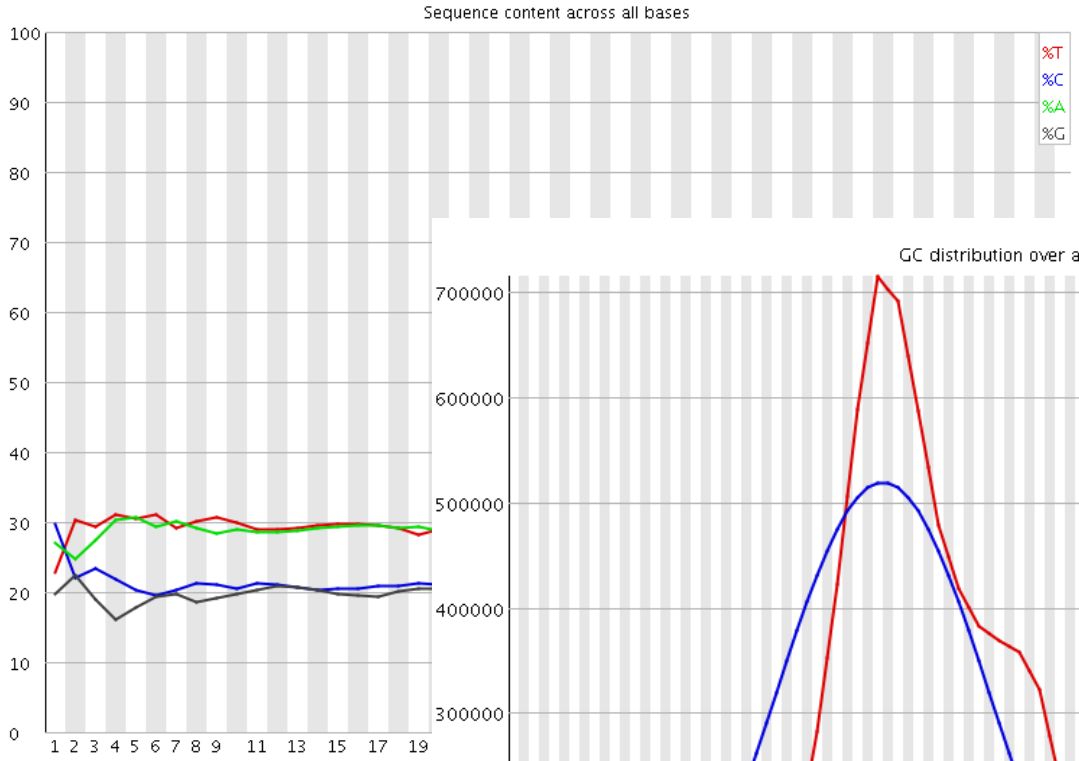


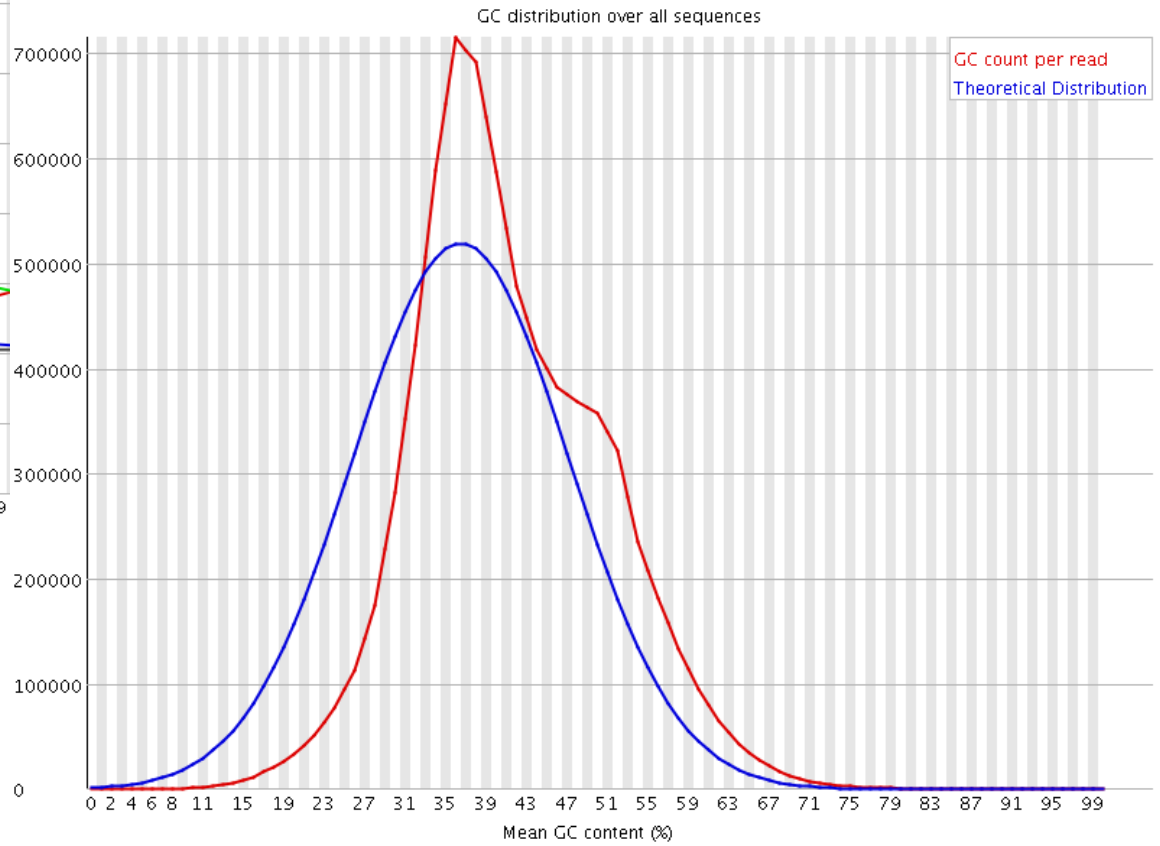




Babraham Bioinformatics

QC of raw sequence
Sequence Composition





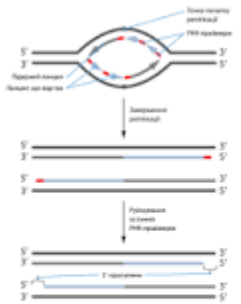
Babraham Bioinformatics

QC of raw sequence
Sequence Composition

Overrepresented sequences

Sequence	Count	Percentage	Possible Source
GGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGG	11933	0.17178556359211344	No Hit
CCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCC	8917	0.1283677089207136	No Hit
CCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACC	7298	0.10506084329969362	No Hit
CTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCT	7282	0.104830509853161	No Hit
GTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGT	7040	0.10134671647435503	No Hit

A **telomere** is a region of repetitive nucleotide **sequences** at each end of a chromosome, which protects the end of the chromosome from deterioration or from fusion with neighboring chromosomes. ... This **sequence** of TTAGGG is **repeated** approximately 2,500 times in humans.

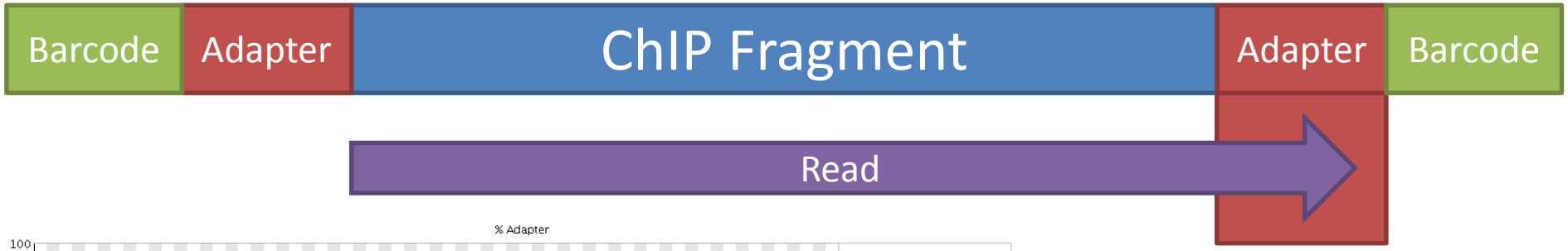


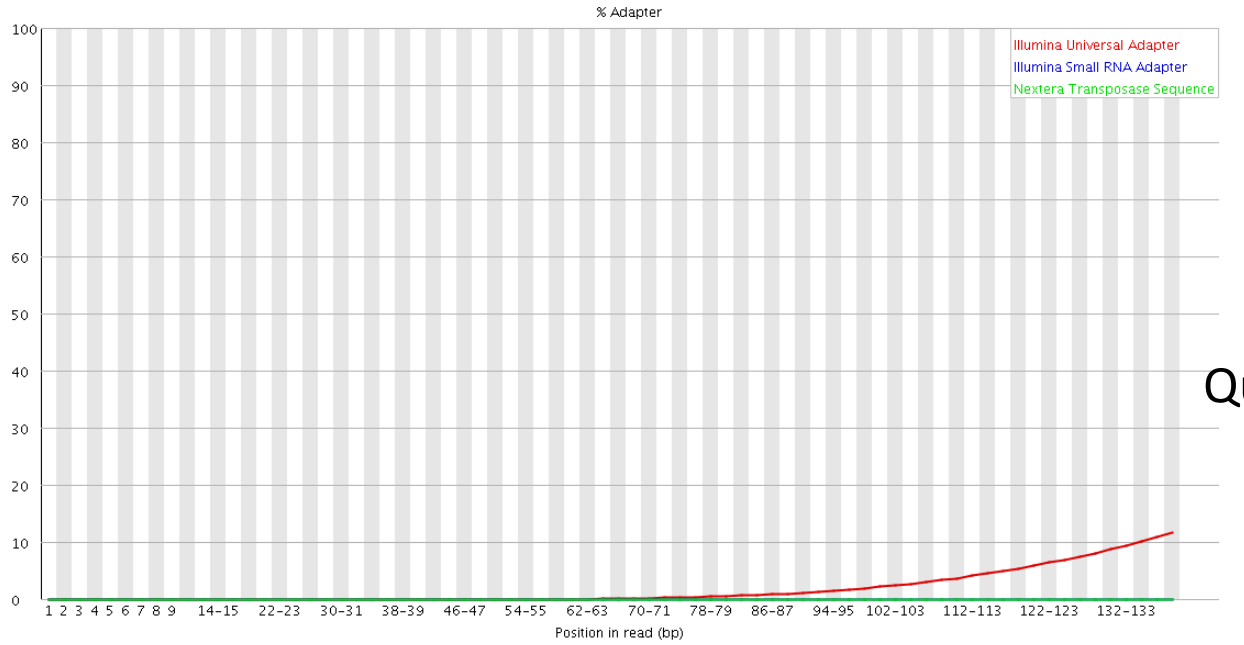
Telomere - Wikipedia
https://en.wikipedia.org/wiki/Telomere

About this result Feedback

Babraham Bioinformatics

QC of raw sequence
Adapter Contamination





Trim Galore!

Quality and Adapter Trimming

Babraham Bioinformatics

Mapping ChIP Data

- All regions should be linear genomic stretches

- Standard genomic aligners are fine
  - Bowtie2 http://bowtie-bio.sourceforge.net/bowtie2/
  - BWA http://bio-bwa.sourceforge.net/

Babraham Bioinformatics

Example Bowtie2 Mapping

- Create Genome Index (once - slow!)

```
bowtie2-build yeast_genome.fa yeast_index
```

- Map a single FastQ file

```
bowtie \
-x yeast_index \
-U data.fastq.gz \
| samtools view \
-bS \
-o data.bam
```

Babraham Bioinformatics

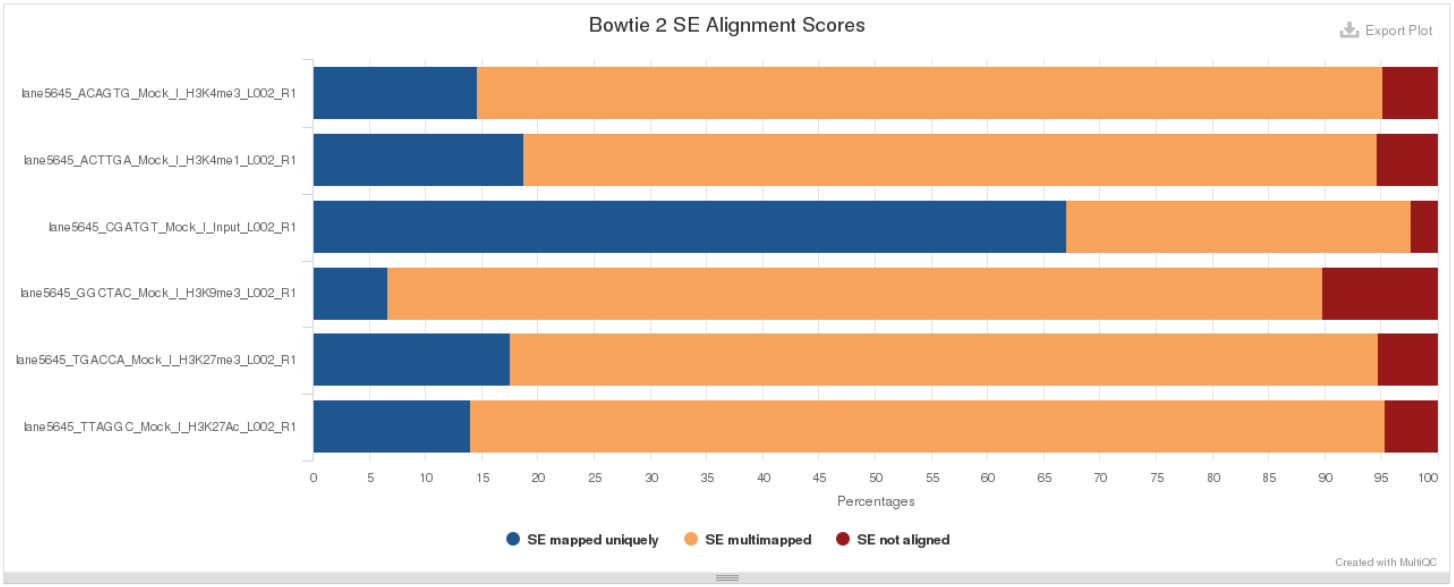
Post Alignment QC
Mapping Statistics

```
sample1:    2264052  (14.66%) aligned exactly 1 time
sample2:    2698005  (18.79%) aligned exactly 1 time
sample3:    13434392 (67.08%) aligned exactly 1 time
sample4:    1108477  (6.70%)  aligned exactly 1 time
sample5:    2143911  (17.58%) aligned exactly 1 time
sample6:    2980154  (13.98%) aligned exactly 1 time
```

Bowtie 2

Bowtie 2 is an ultrafast and memory-efficient tool for aligning sequencing reads to long reference sequences.

Number of Reads | Percentages



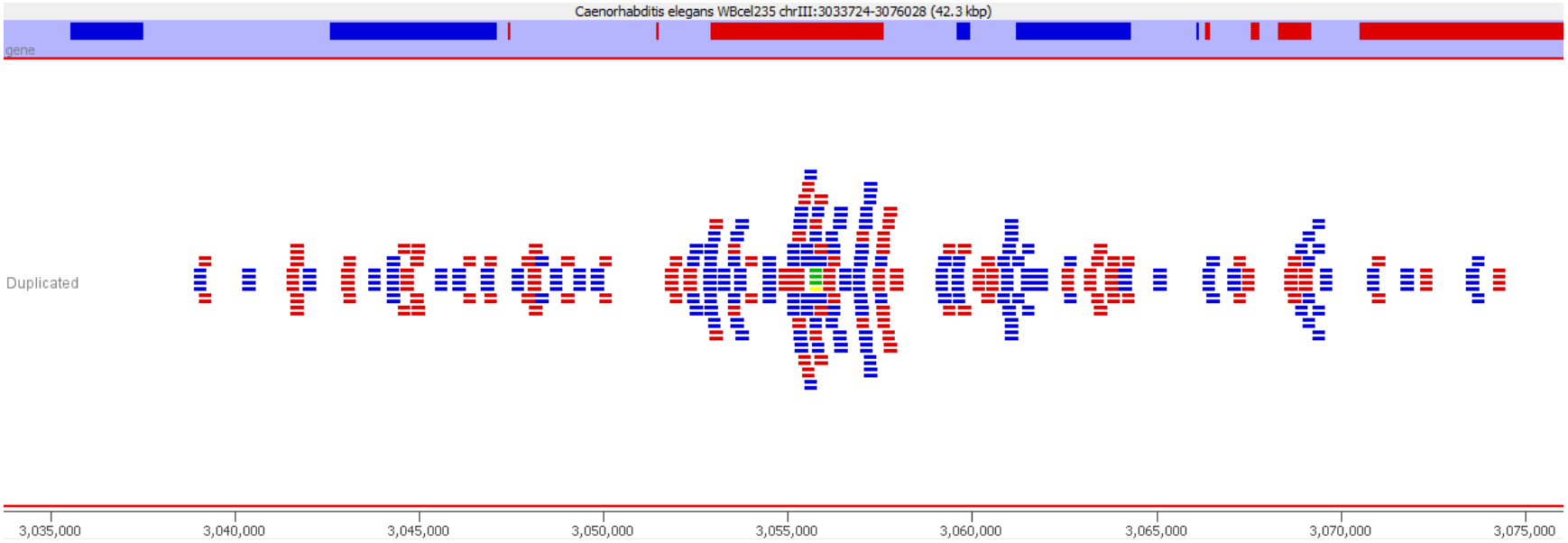
Post Alignment Processing
MAPQ Filtering

- ChIP-Seq relates sequences to positions in a reference genome
- You need to be confident that the reported position is correct
- Filtering on MAPQ value (likelihood of reported position being incorrect) is an easy way to do this
- MAPQ filtering should be performed in most cases

```
samtools view -q 20 -b -o filtered.bam data.bam
```

Babraham Bioinformatics

Post Alignment Processing
Deduplication




```
java -jar picard.jar SortSam \
    INPUT=filtered.bam \
    OUTPUT=sorted.bam \
    SORT_ORDER=coordinate
```

```
java -jar picard.jar MarkDuplicates \
    INPUT=sorted.bam \
    OUTPUT=dedup.bam \
    METRICS_FILE=metrics.txt
```

DO NOT DEDUPLICATE AS A MATTER OF COURSE! THINK FIRST!

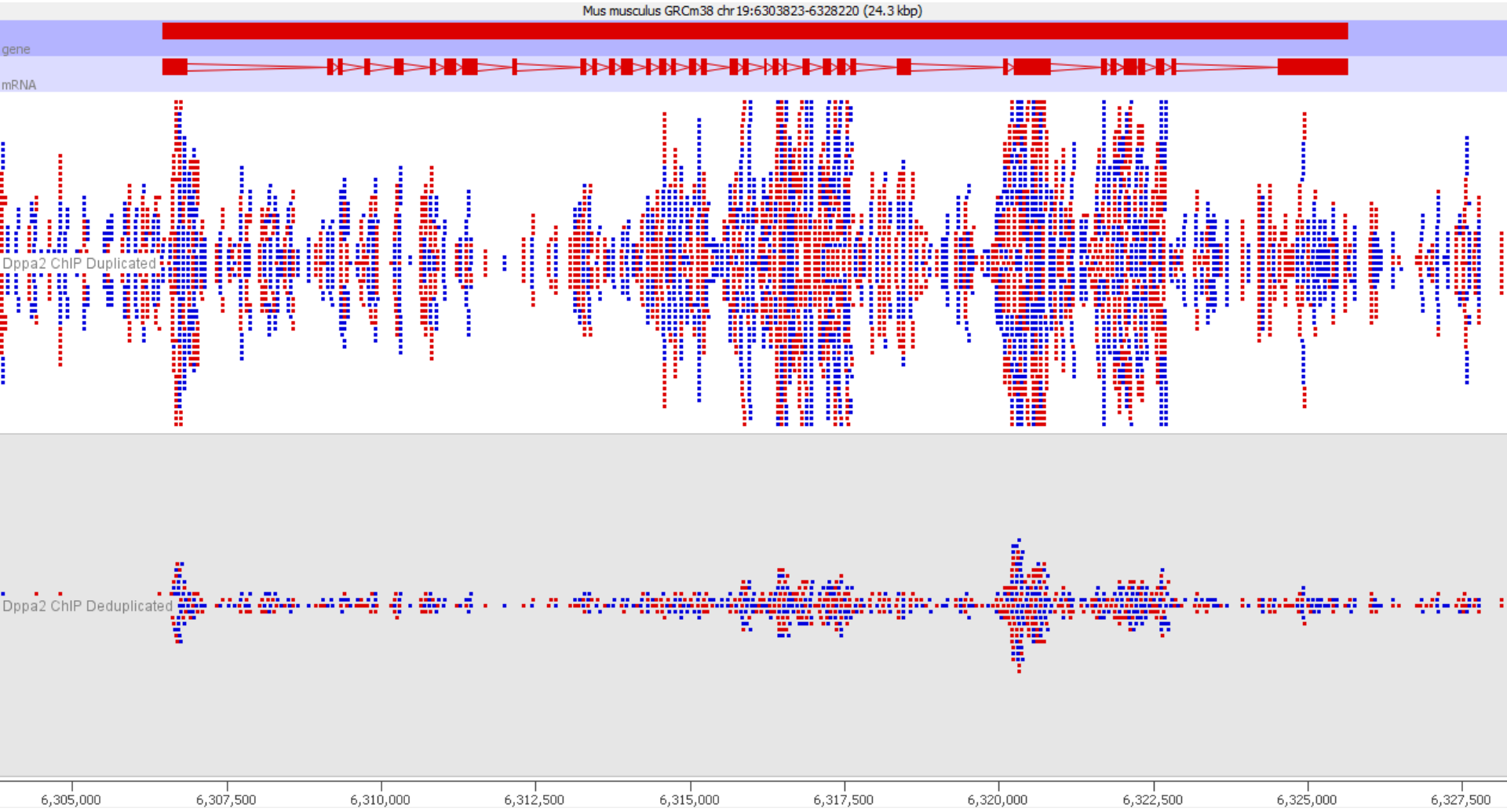
To Deduplicate or Not?

- Deduplication can make enrichment visually clearer and help to spot truly enriched regions

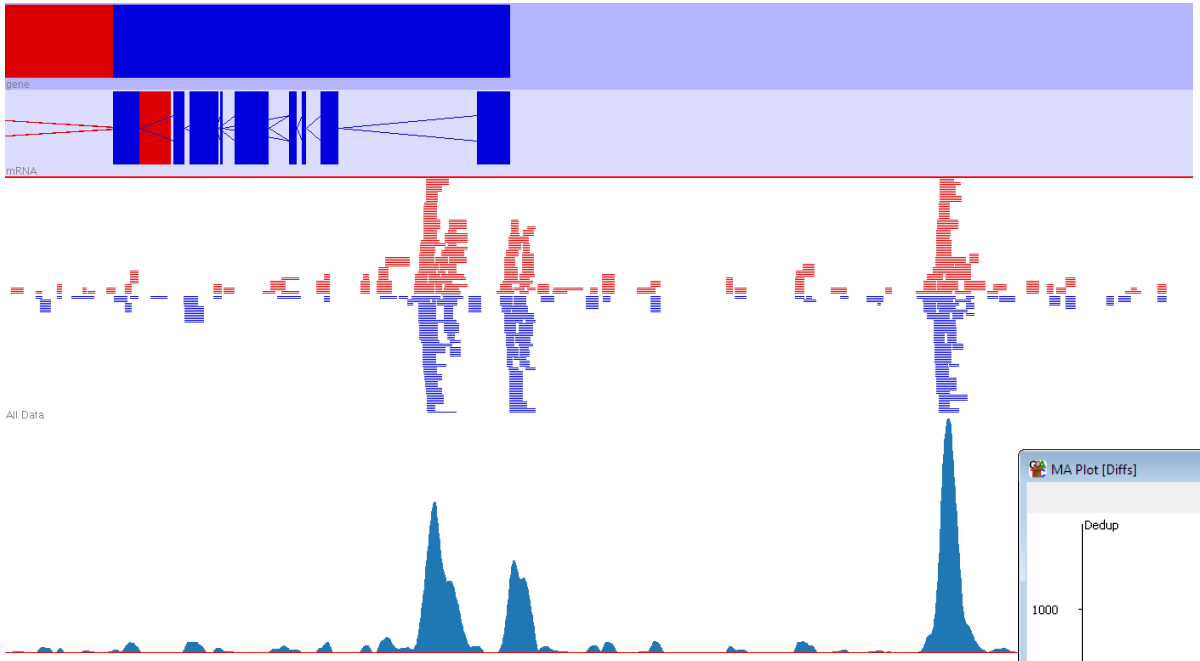
- Why not just deduplicate everything?
  - Quantitation compression
  - Repeat enrichment

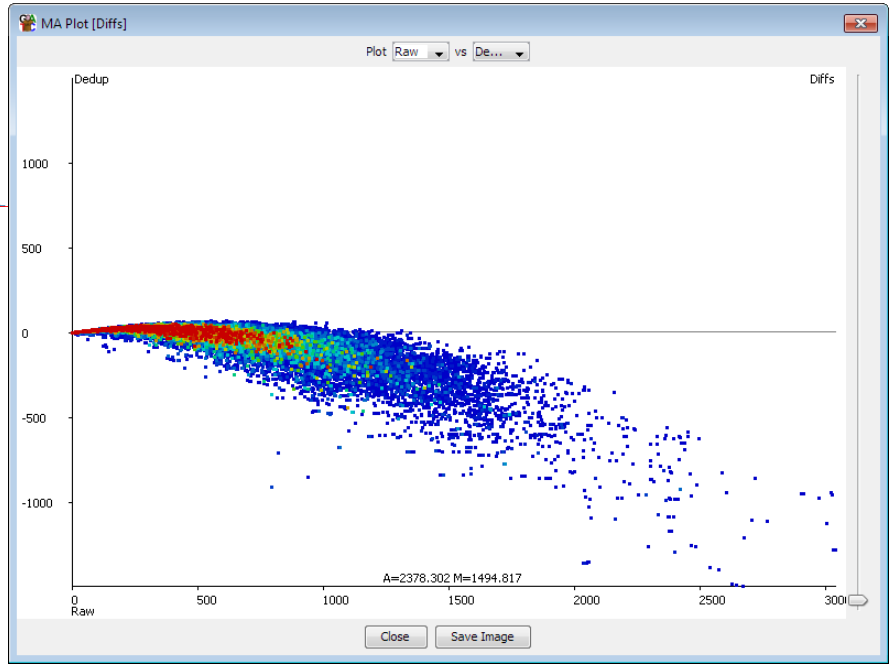
Babraham Bioinformatics

Deduplication



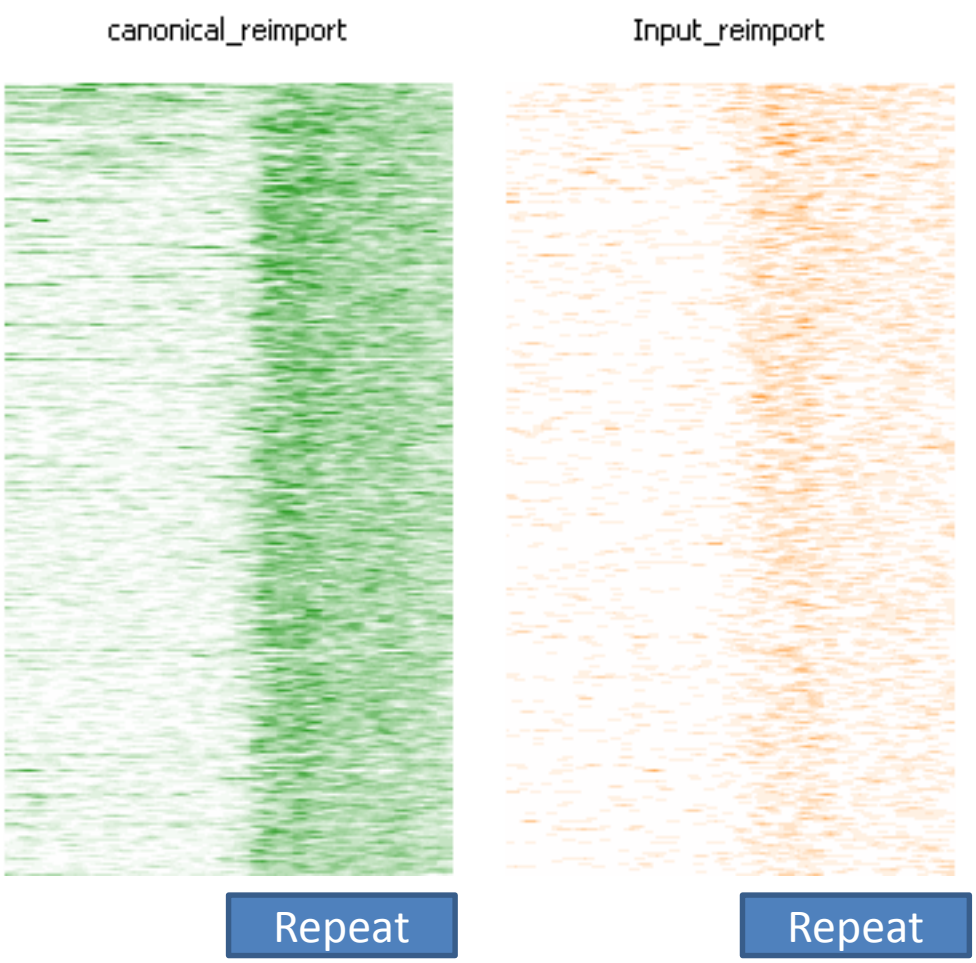
Quantitation Compression from Deduplication

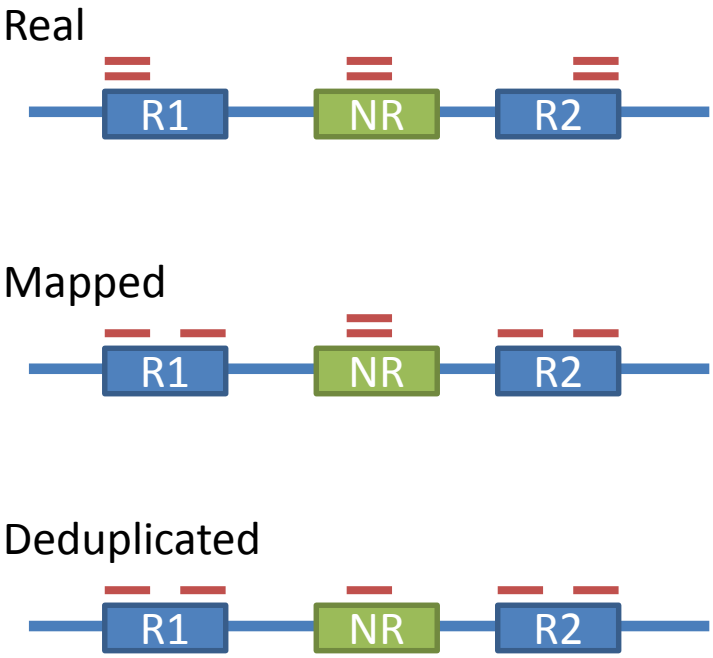




Babraham Bioinformatics

Repeat Enrichment from Deduplication

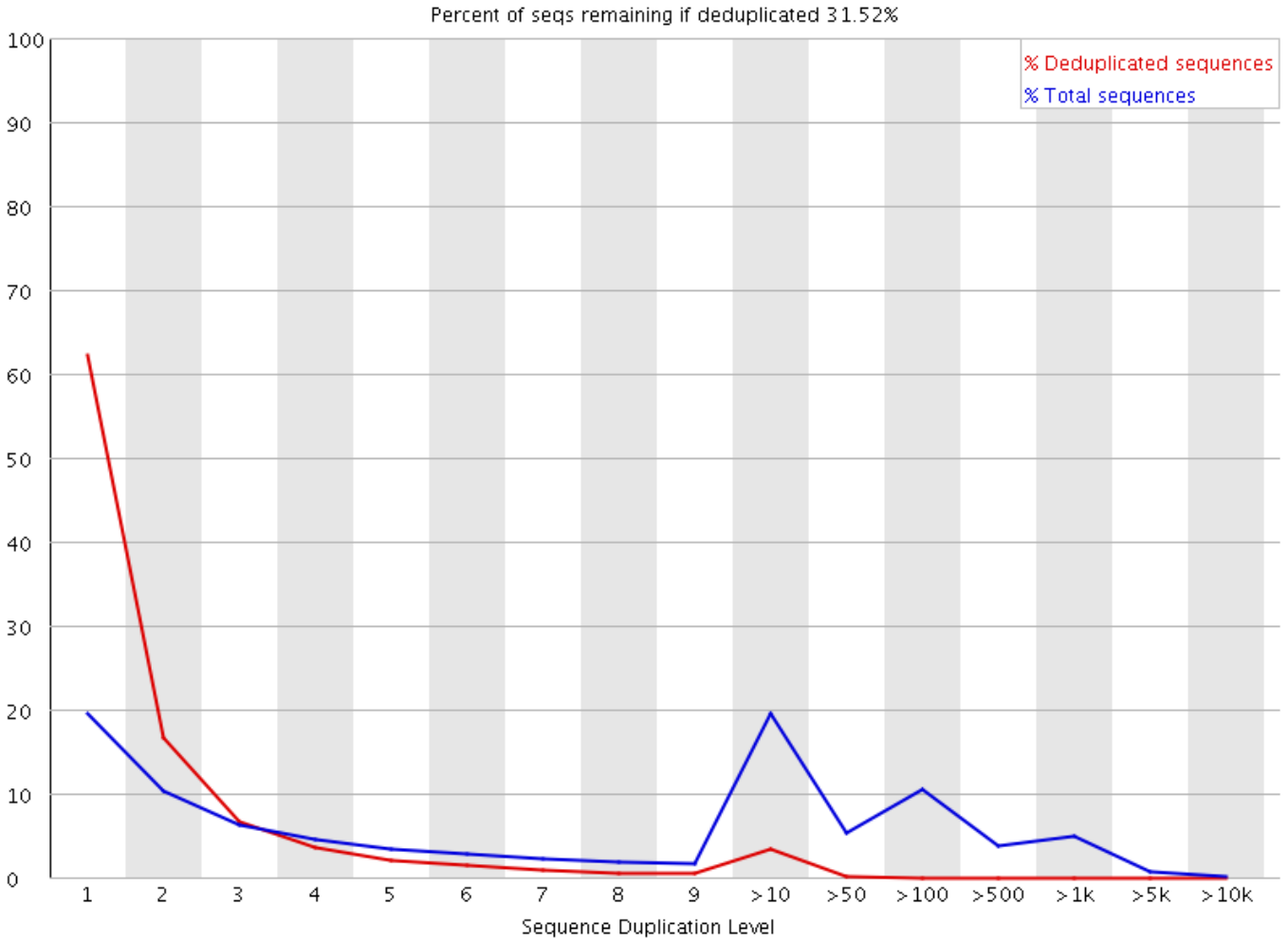




Peak callers (MACS for example) deduplicate internally, so you don't have to consciously do this.

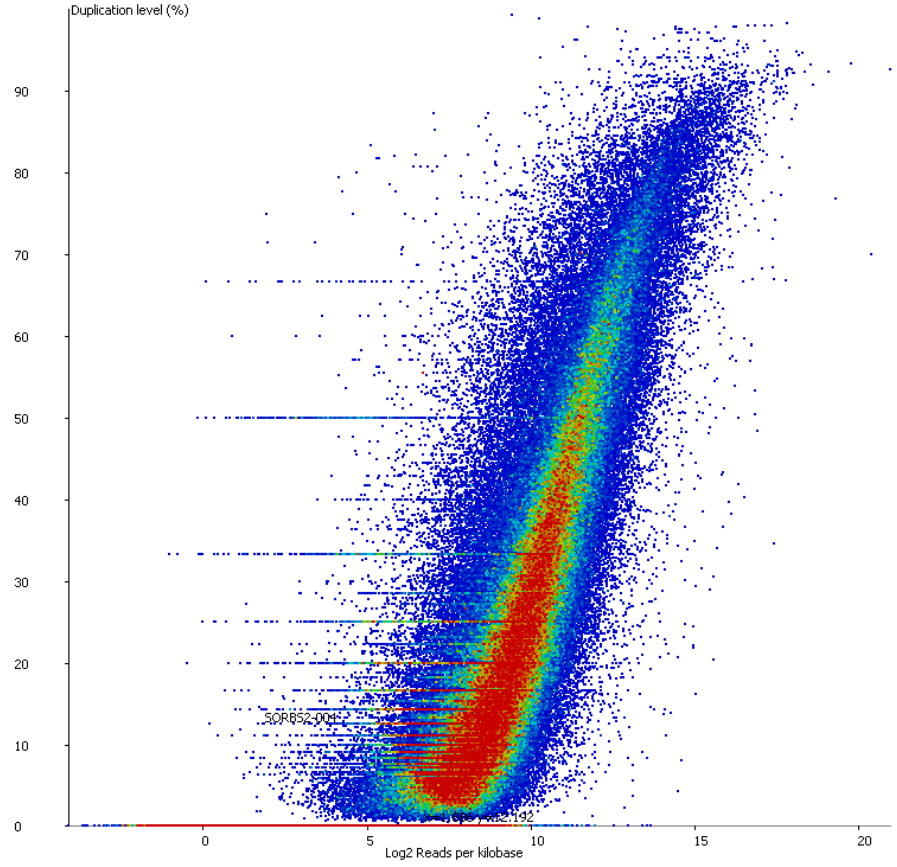
Only avoided by using uniquely mapped reads.

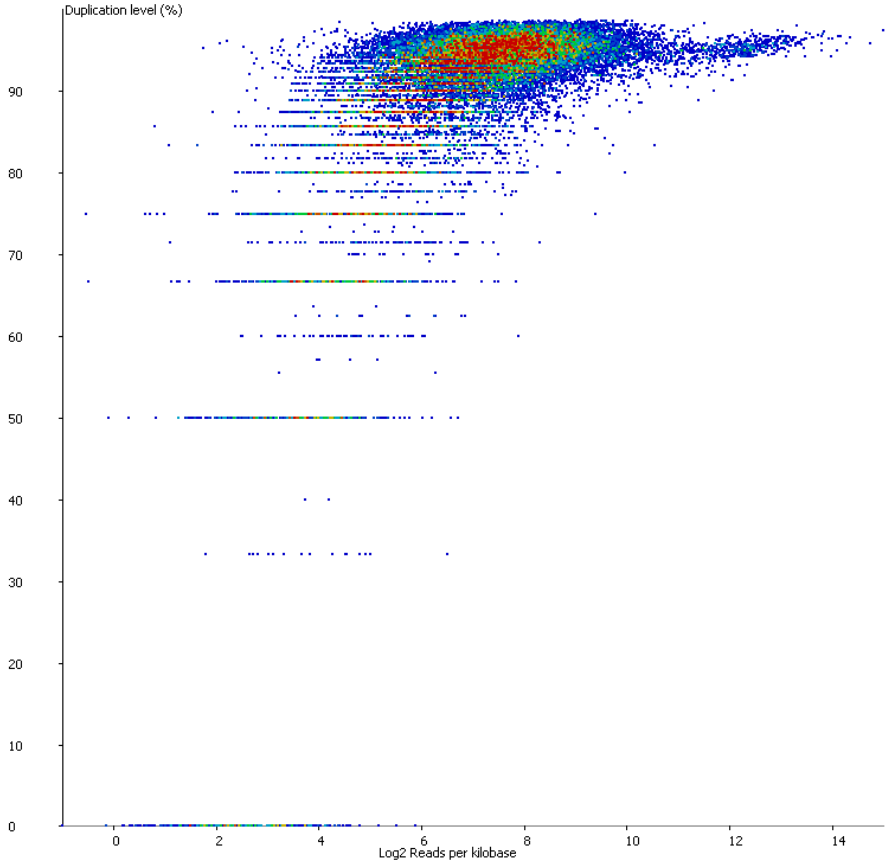
Assessing Duplication



Babraham Bioinformatics

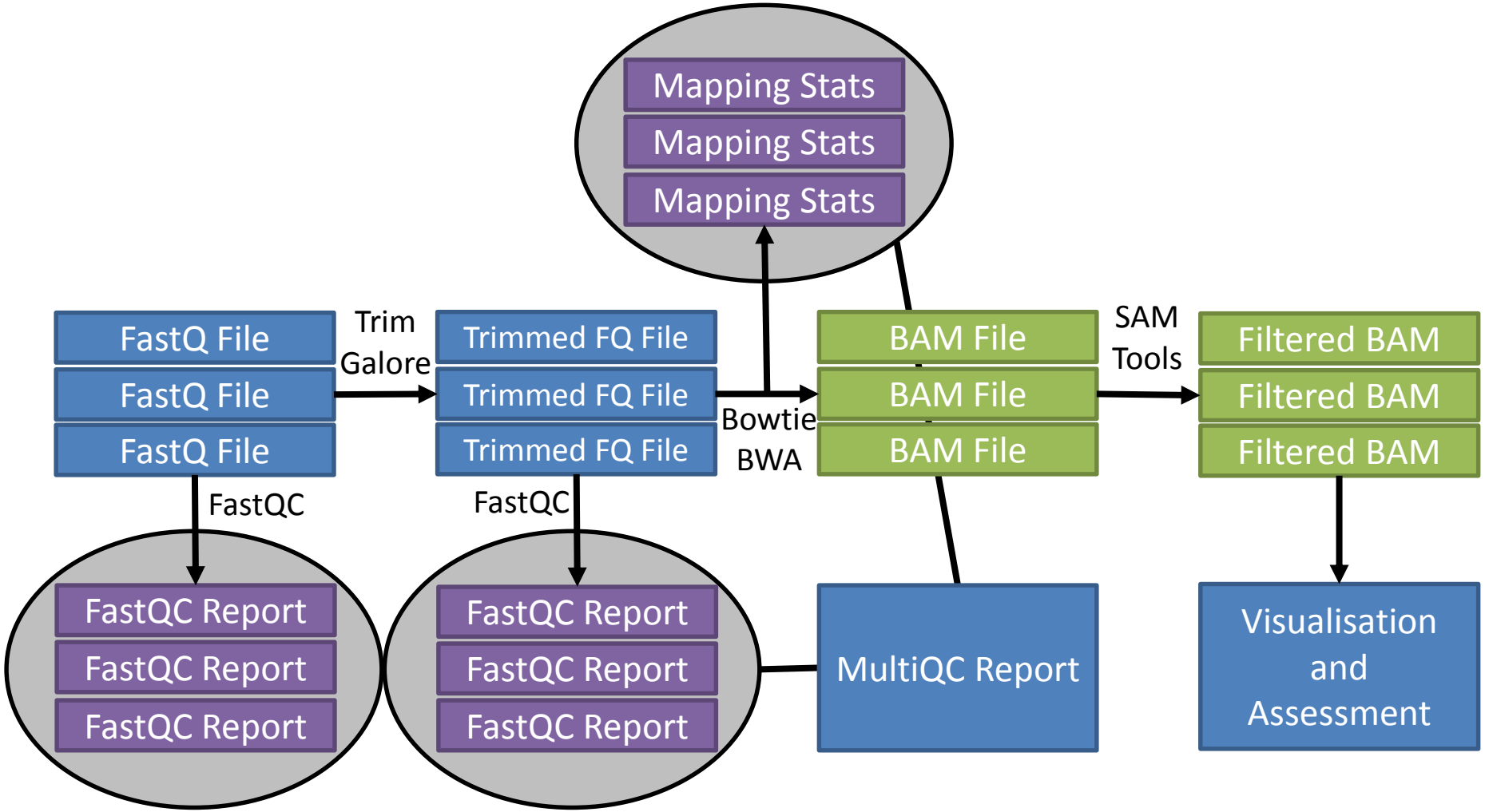
Assessing Duplication





Babraham Bioinformatics

Standard Processing Workflow



Babraham Bioinformatics