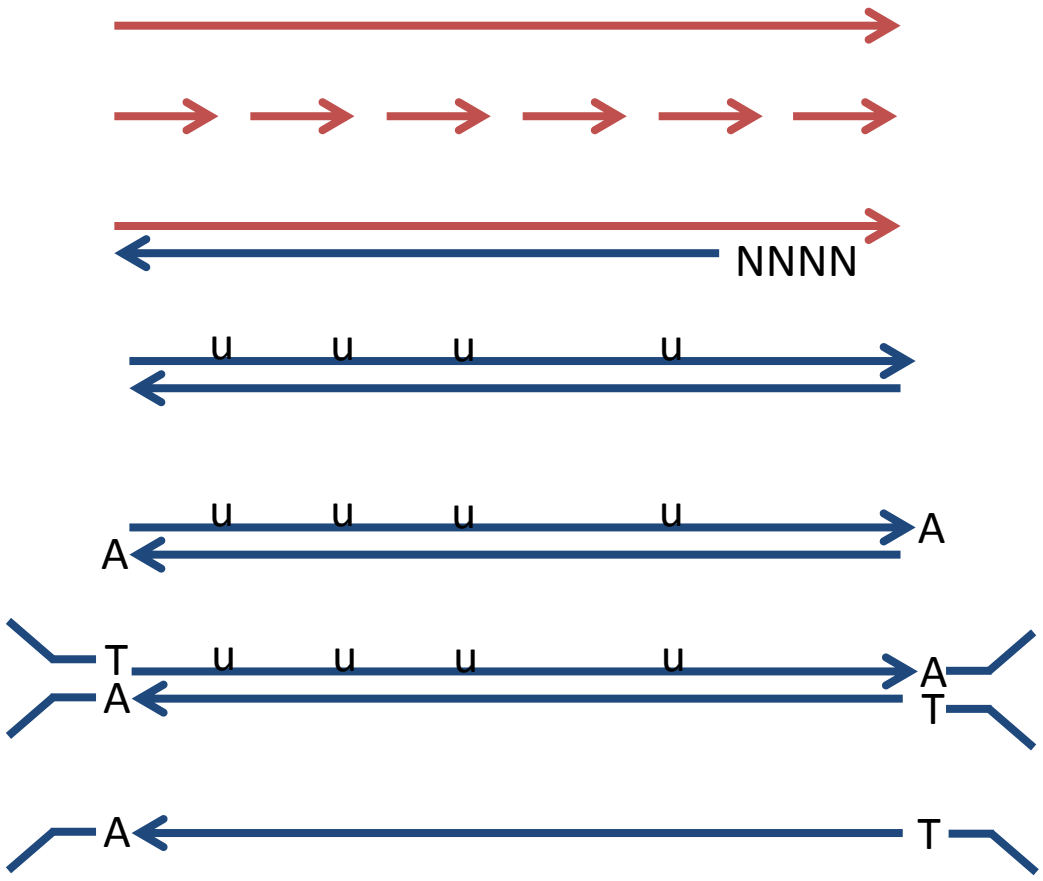
RNA-Seq Analysis

Simon Andrews
simon.andrews@babraham.ac.uk
@simon_andrews
v2017-03

Babraham Bioinformatics

RNA-Seq Libraries



rRNA depleted mRNA

Fragment

Random prime + RT

2nd strand synthesis (+ U)

A-tailing

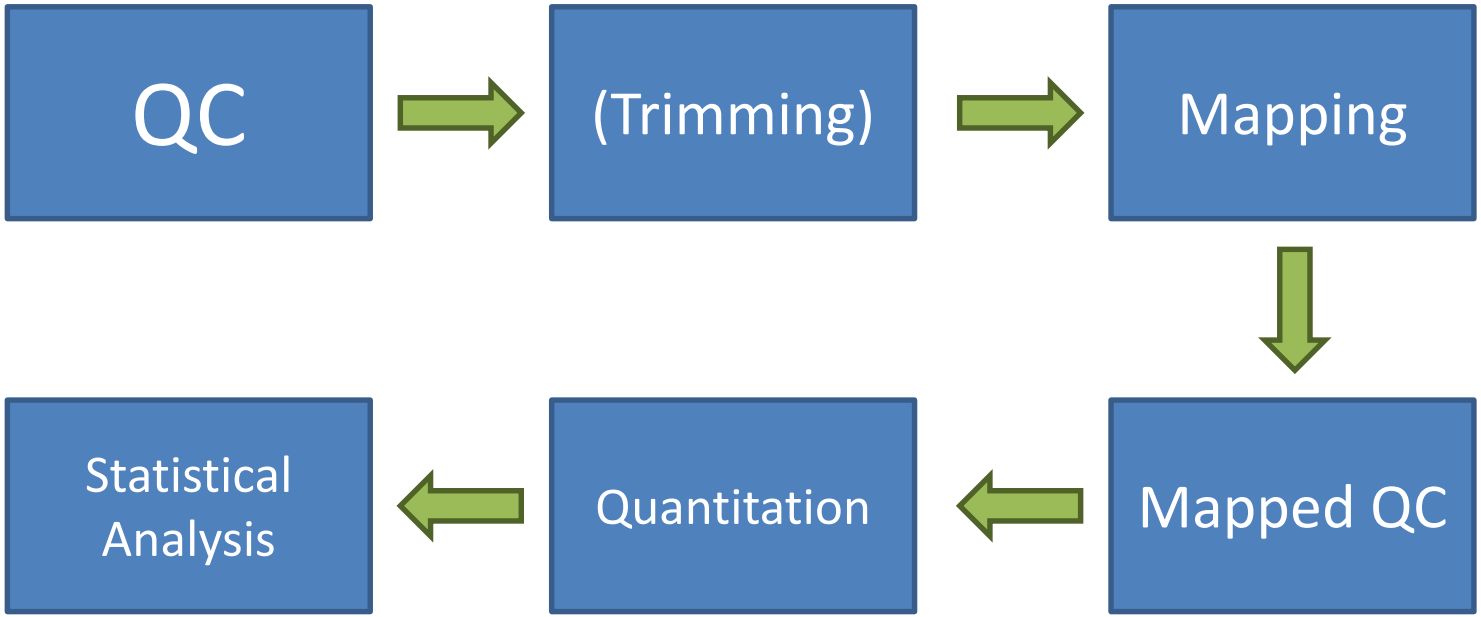
Adapter Ligation

(U strand degradation)

Sequencing

Babraham Bioinformatics

Reference based RNA-Seq Analysis



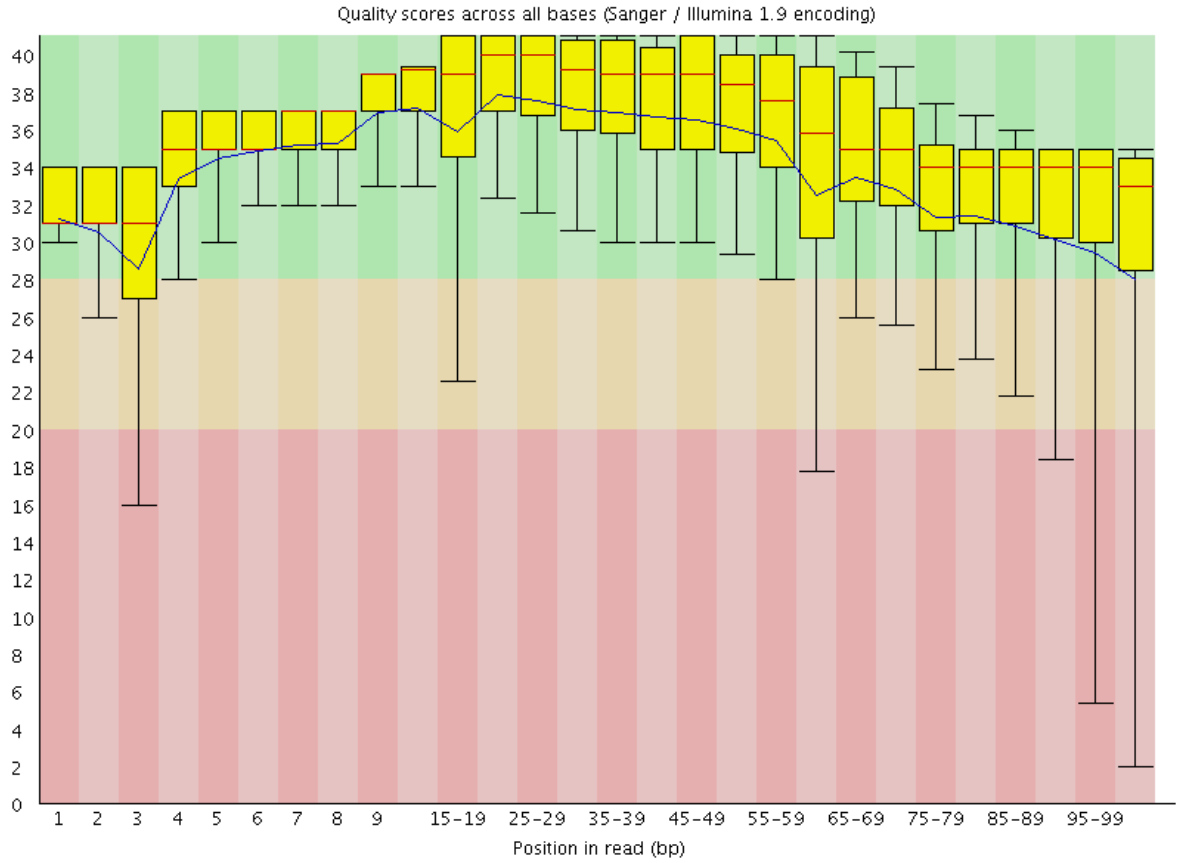
Babraham Bioinformatics

Quality Control

Babraham Bioinformatics

QC: Raw Data

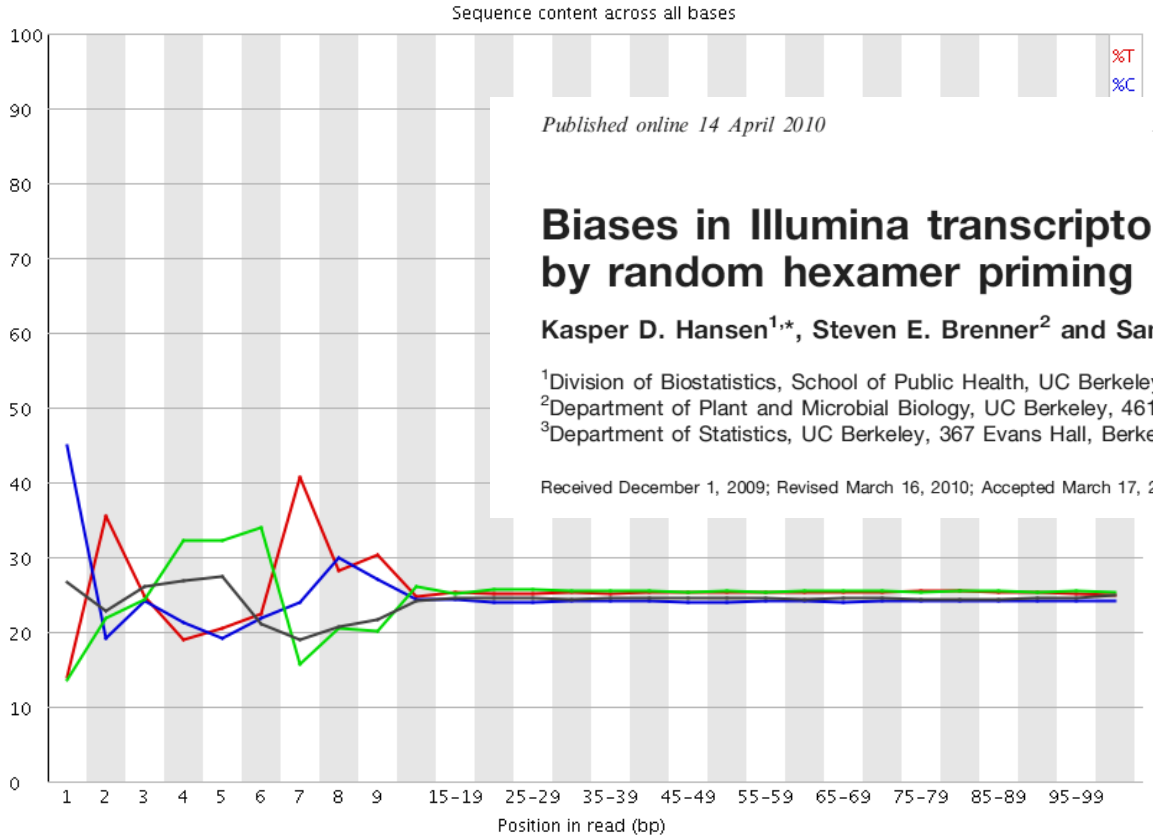
- Sequence call quality



Babraham Bioinformatics

QC: Raw Data

- Sequence bias



Published online 14 April 2010

Nucleic Acids Research, 2010, Vol. 38, No. 12 ***e131***
doi:10.1093/nar/gkq224

Biases in Illumina transcriptome sequencing caused by random hexamer priming

Kasper D. Hansen[1,*], Steven E. Brenner[2] and Sandrine Dudoit[1,3]

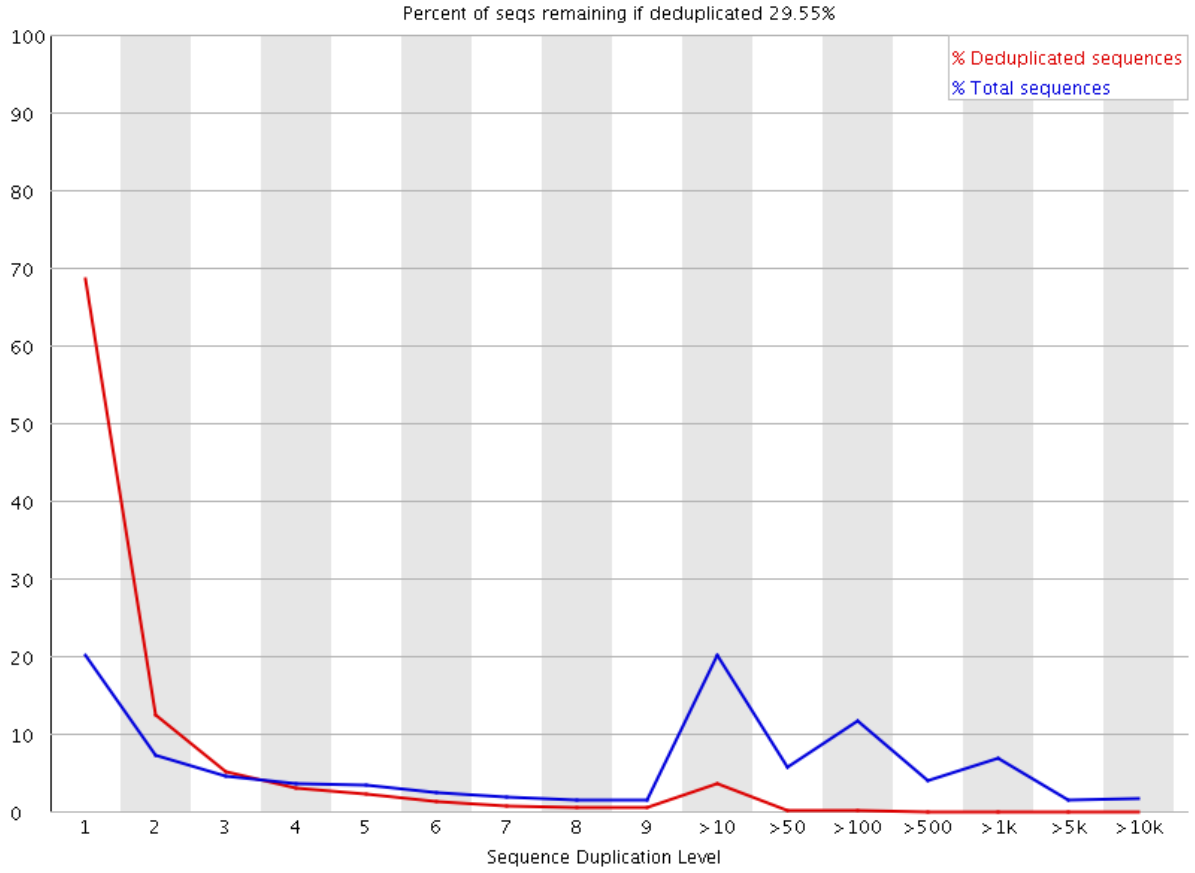
[1]Division of Biostatistics, School of Public Health, UC Berkeley, 101 Haviland Hall, Berkeley, CA 94720-7358, [2]Department of Plant and Microbial Biology, UC Berkeley, 461 Koshland Hall, Berkeley, CA 94720-3102 and [3]Department of Statistics, UC Berkeley, 367 Evans Hall, Berkeley, CA 94720-3860, USA

Received December 1, 2009; Revised March 16, 2010; Accepted March 17, 2010

Babraham Bioinformatics

QC: Raw Data

- Duplication level

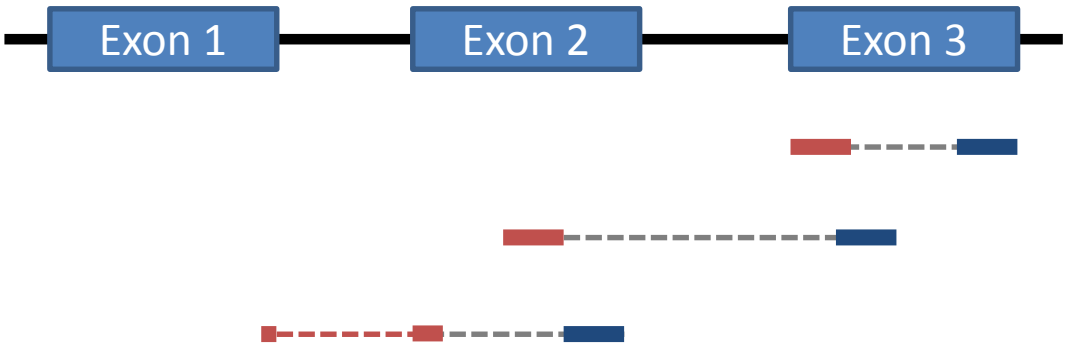


Babraham Bioinformatics

Mapping to a reference

Babraham Bioinformatics

Mapping



Genome

Simple mapping within exons

Mapping between exons

Spliced mapping

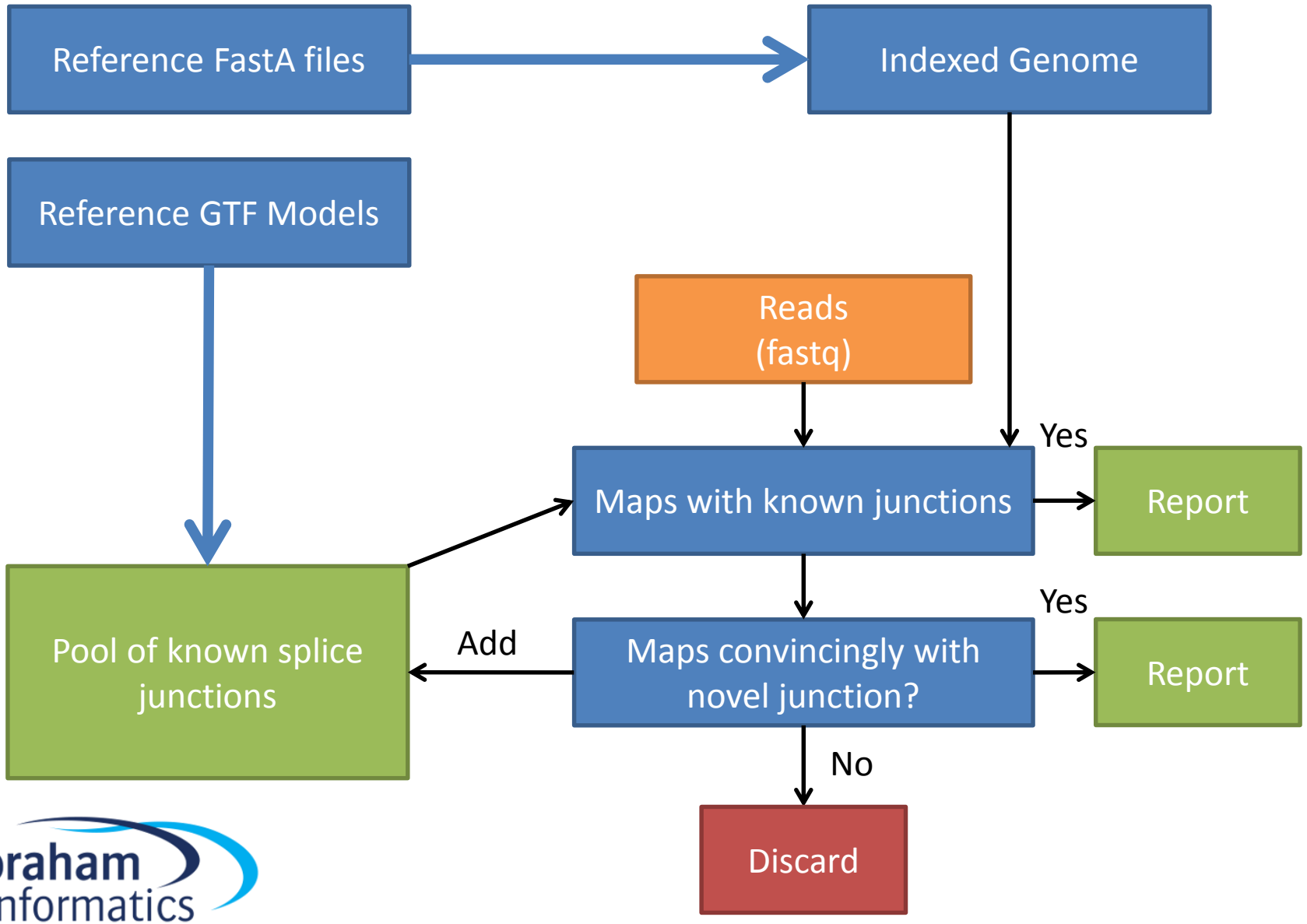
Babraham Bioinformatics

RNA-Seq Mapping Software

- HiSat2 (https://ccb.jhu.edu/software/hisat2/)
- Star (http://code.google.com/p/rna-star/)
- Tophat (http://tophat.cbcb.umd.edu/)

Babraham Bioinformatics

HiSat2 pipeline



Babraham Bioinformatics

Quality Control (mapped data)

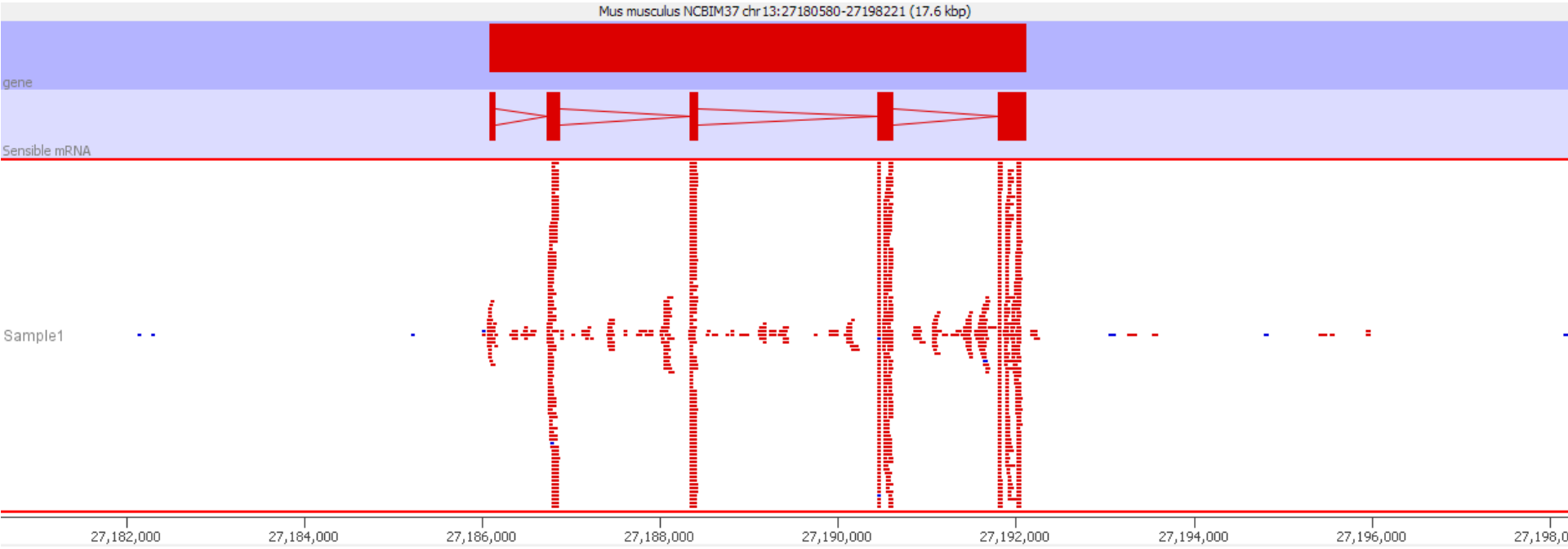
Babraham Bioinformatics

Mapping Statistics

```
2221186 reads; of these:
2221186 (100.00%) were paired; of these:
710157 (31.97%) aligned concordantly 0 times
794711 (35.78%) aligned concordantly exactly 1 time
716318 (32.25%) aligned concordantly >1 times
----
710157 pairs aligned concordantly 0 times; of these:
3924 (0.55%) aligned discordantly 1 time
----
706233 pairs aligned 0 times concordantly or discordantly; of these:
1412466 mates make up the pairs; of these:
1280507 (90.66%) aligned 0 times
52740 (3.73%) aligned exactly 1 time
79219 (5.61%) aligned >1 times
71.18% overall alignment rate
```

Babraham Bioinformatics

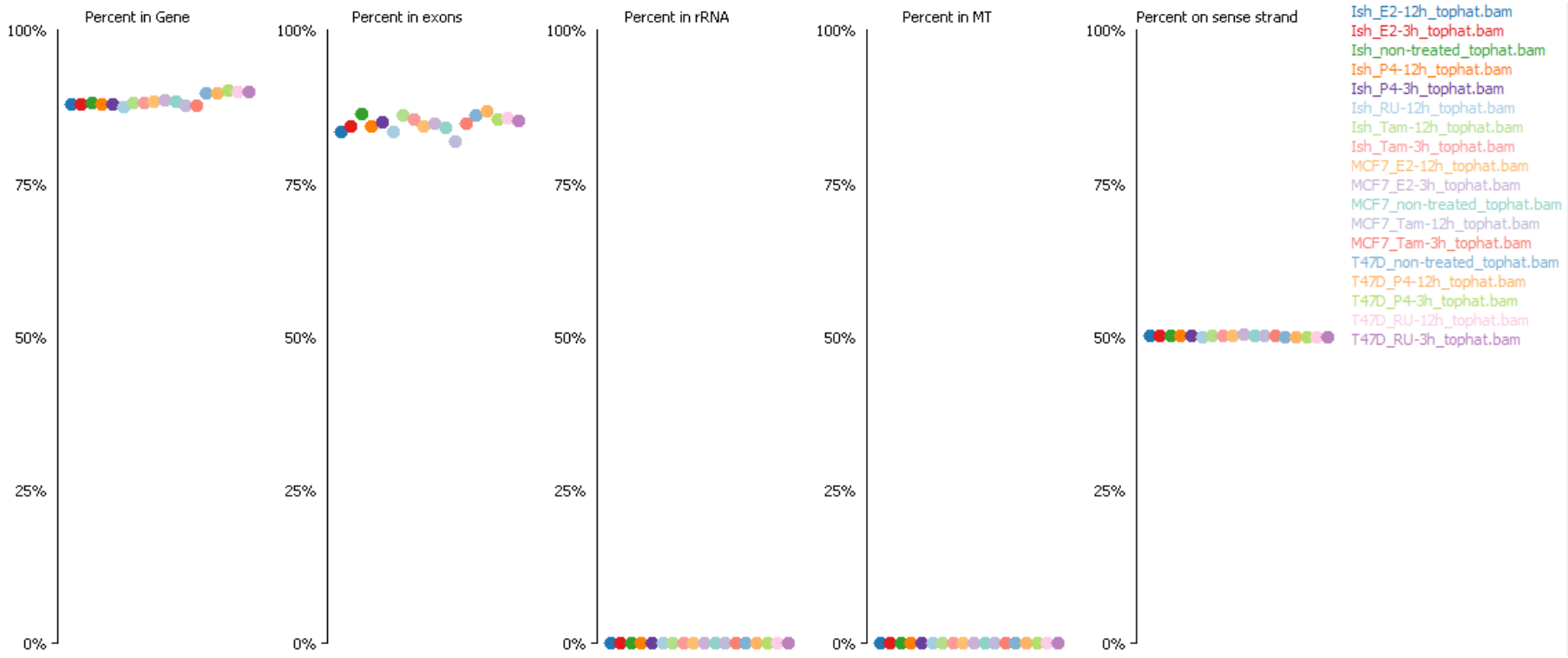
Viewing Mapped Data



- Reads over exons
- Reads over introns
- Reads in intergenic
- Strand specificity

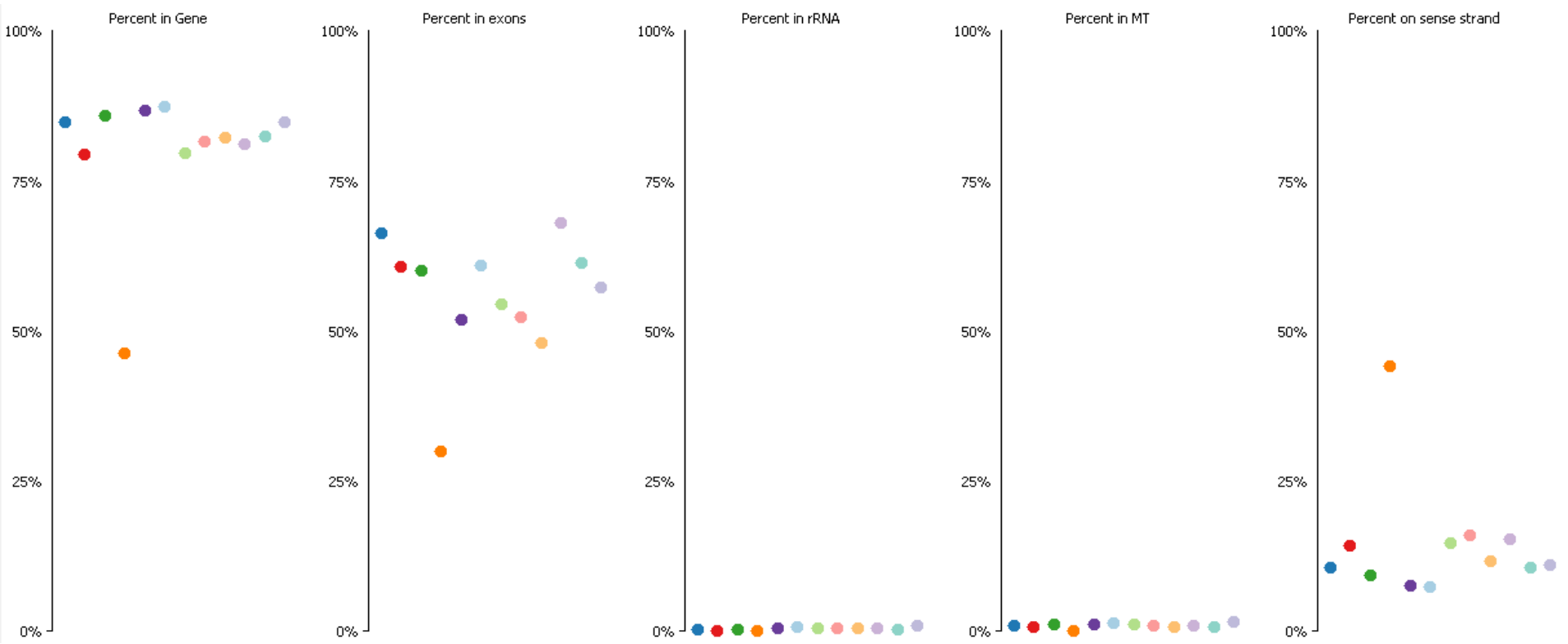
Babraham Bioinformatics

SeqMonk RNA-Seq QC (good)



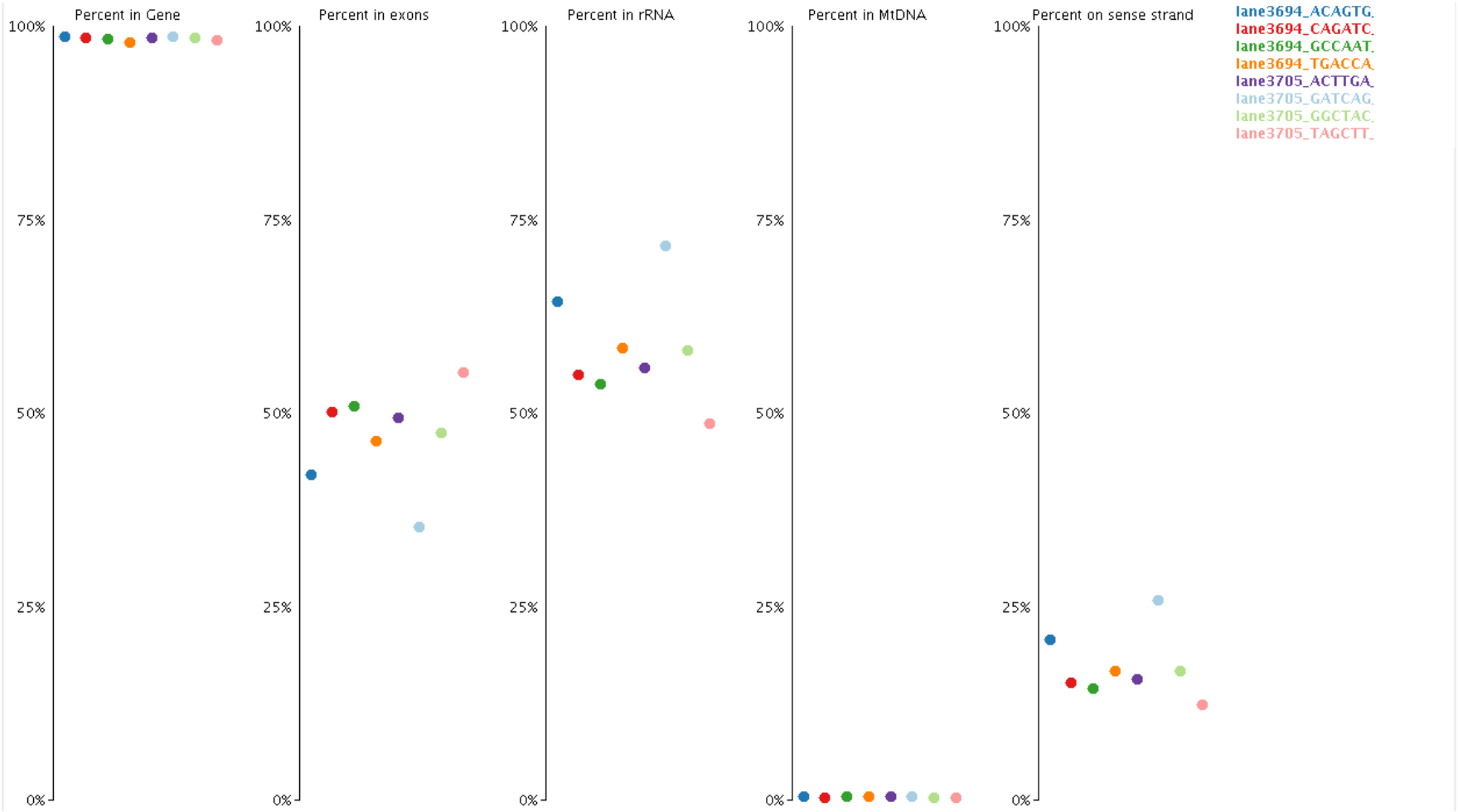
Babraham Bioinformatics

SeqMonk RNA-Seq QC (bad)



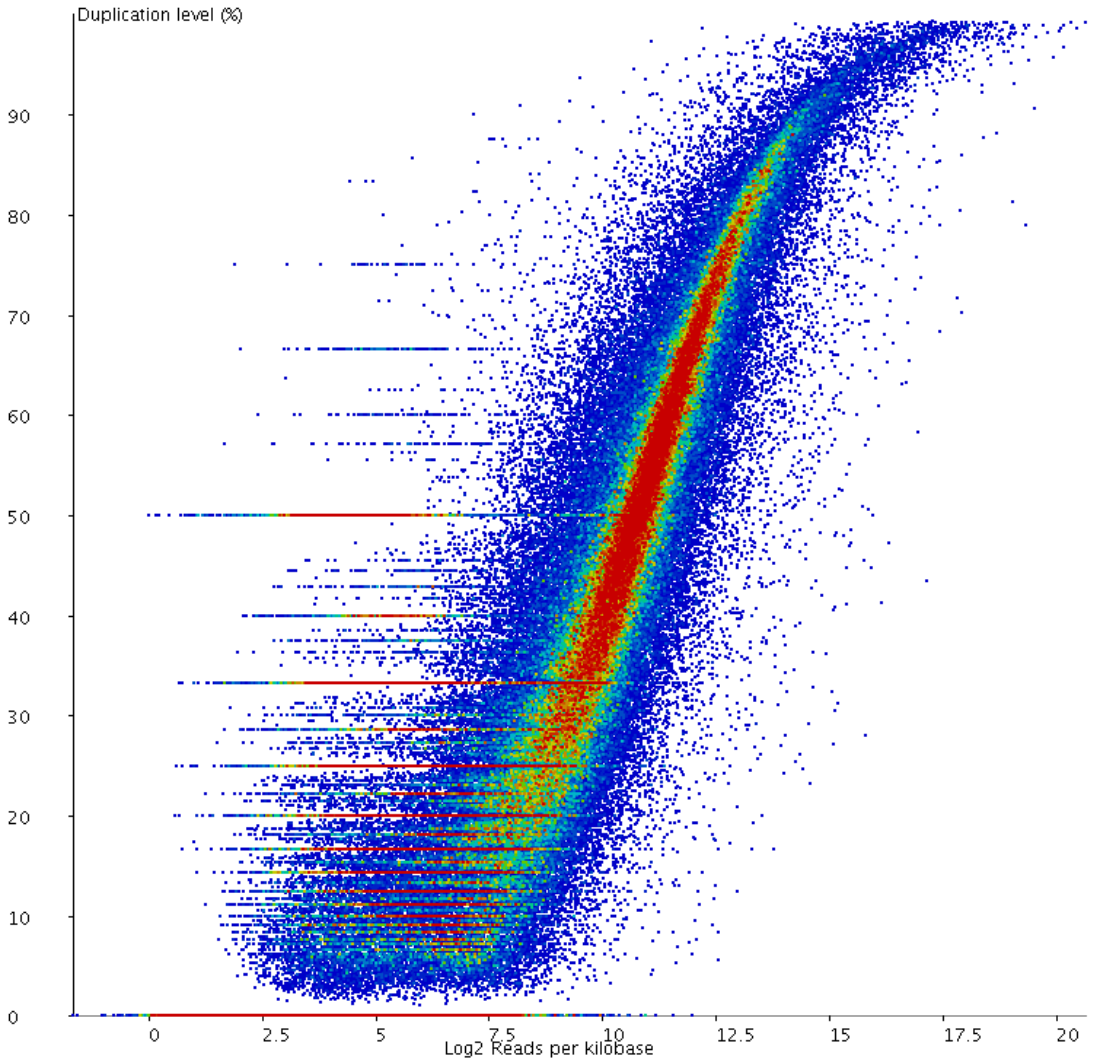
Babraham Bioinformatics

SeqMonk RNA-Seq QC (bad)



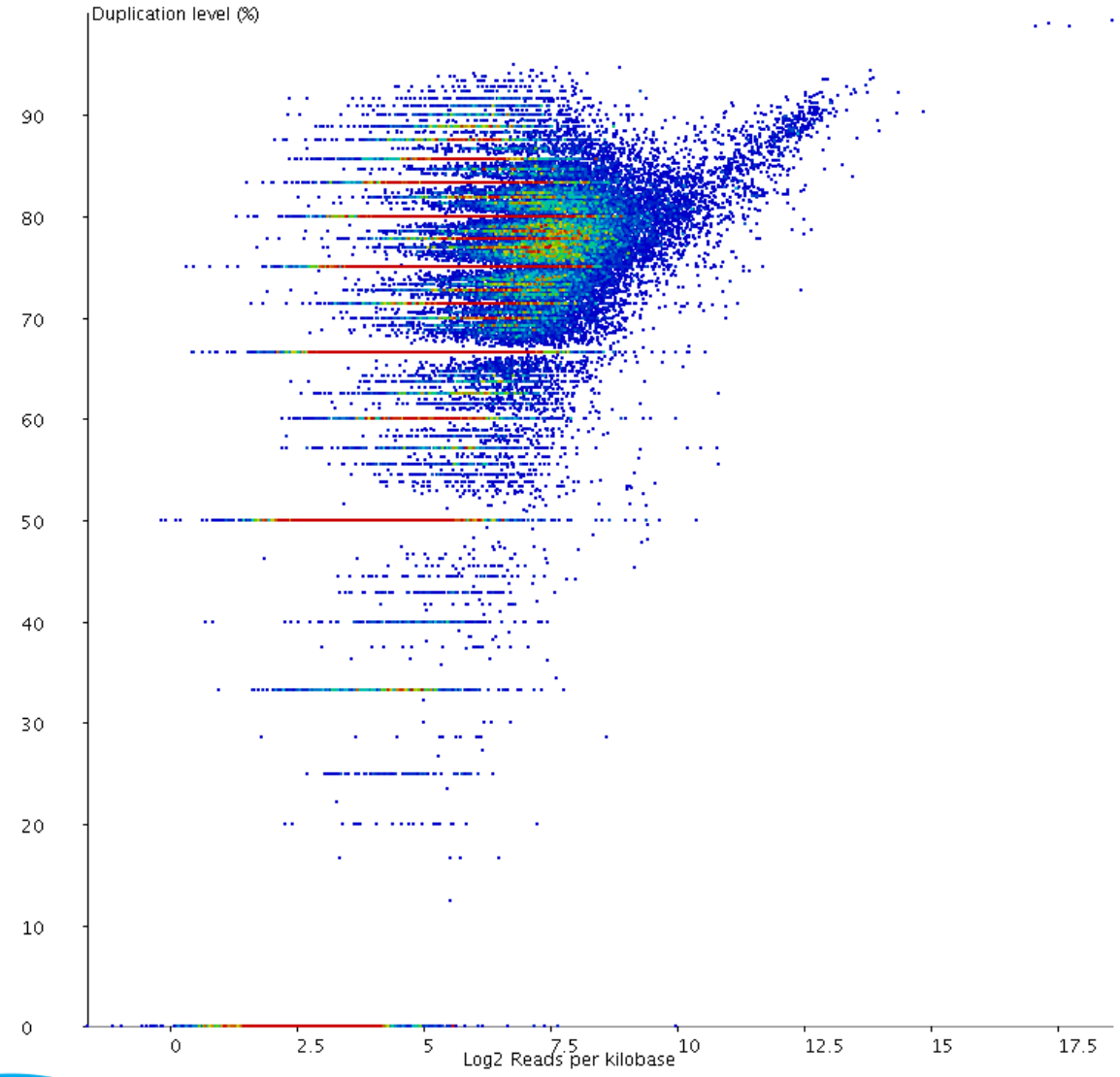
Babraham Bioinformatics

Duplication (good)



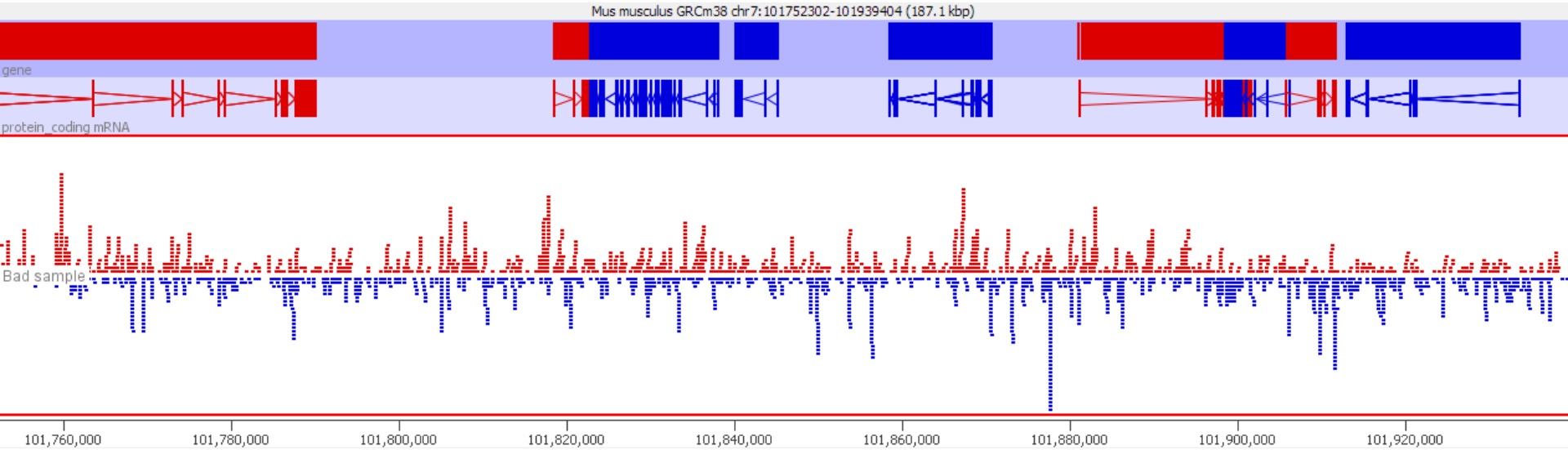
Babraham Bioinformatics

Duplication (bad)



Babraham Bioinformatics

Look at poor QC samples

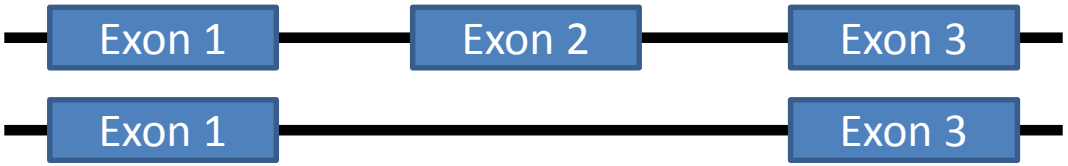


Babraham Bioinformatics

Quantitation

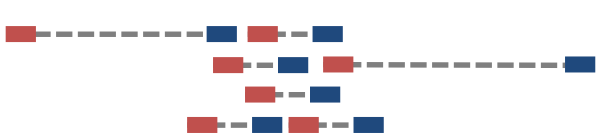
Babraham Bioinformatics

Quantitation



Splice form 1

Splice form 2



Definitely splice form 1

Definitely splice form 2

Ambiguous

Babraham Bioinformatics

Simple Quantitation

- Count read overlaps with exons of each gene
  - Consider library directionality
  - Simple
  - Gene level quantitation
  - Many programs
    - Seqmonk (graphical)
    - Feature Counts (subread)
    - BEDTools
    - HTSeq

Babraham Bioinformatics

Analysing Splicing

- Estimate transcript level expression
  - Cufflinks, RSEM, bitSeq etc.
- Compare splicing decision ratios
  - rMATS, logistic regression
- Compare exon expression to gene expression
  - DEXSeq
- Analyse junction usage directly
  - Counts + DESeq/EdgeR

Babraham Bioinformatics

RPKM / FPKM / TPM

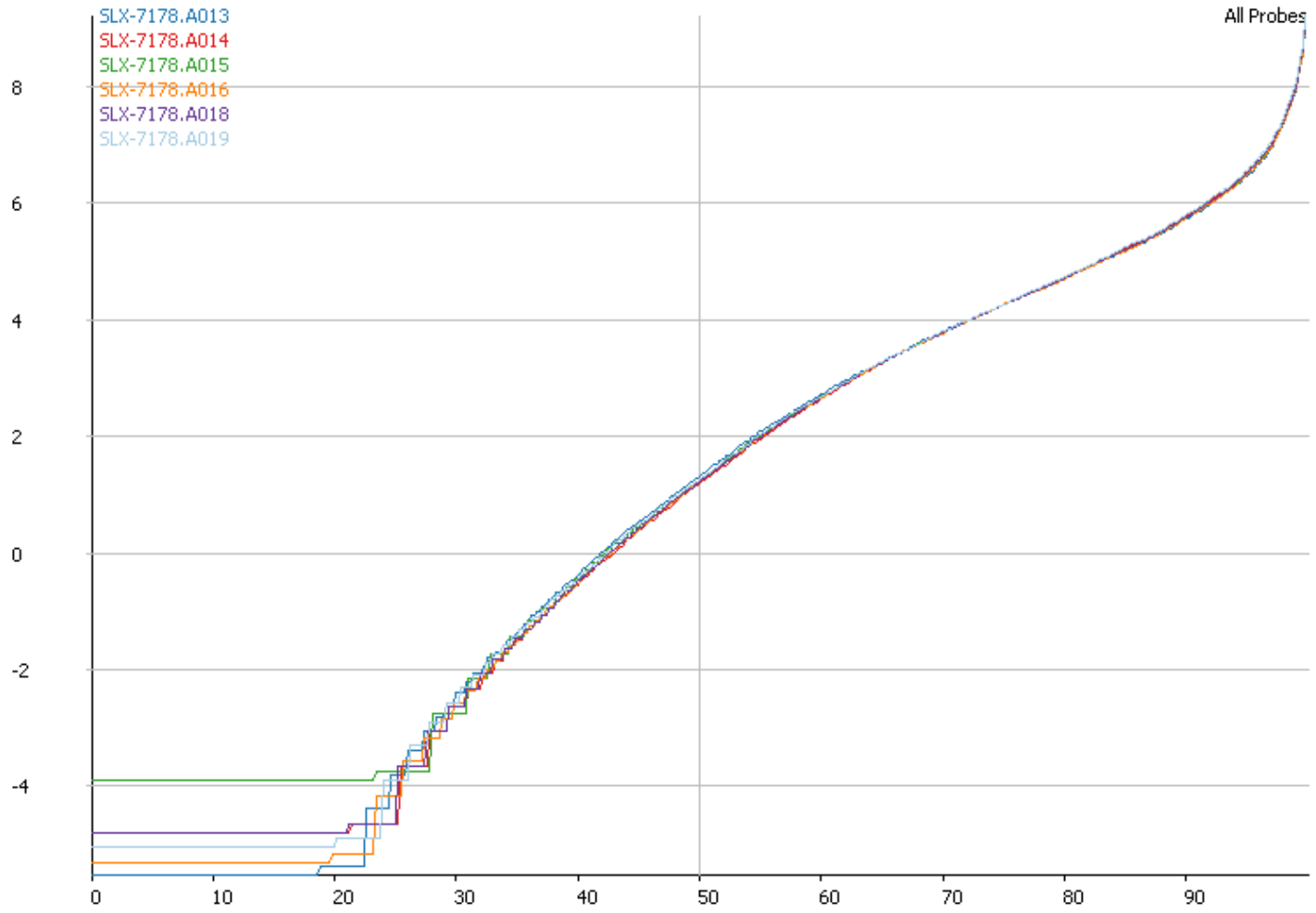
- RPKM (Reads per kilobase of transcript per million reads of library)
  - Corrects for total library coverage
  - Corrects for gene length
  - Comparable between different genes within the same dataset

- FPKM (Fragments per kilobase of transcript per million reads of library)
  - Only relevant for paired end libraries
  - Pairs are not independent observation
  - Effectively halves raw counts

- TPM (transcripts per million)
  - Normalises to transcript copies instead of reads
  - Corrects for cases where the average transcript length differs between samples

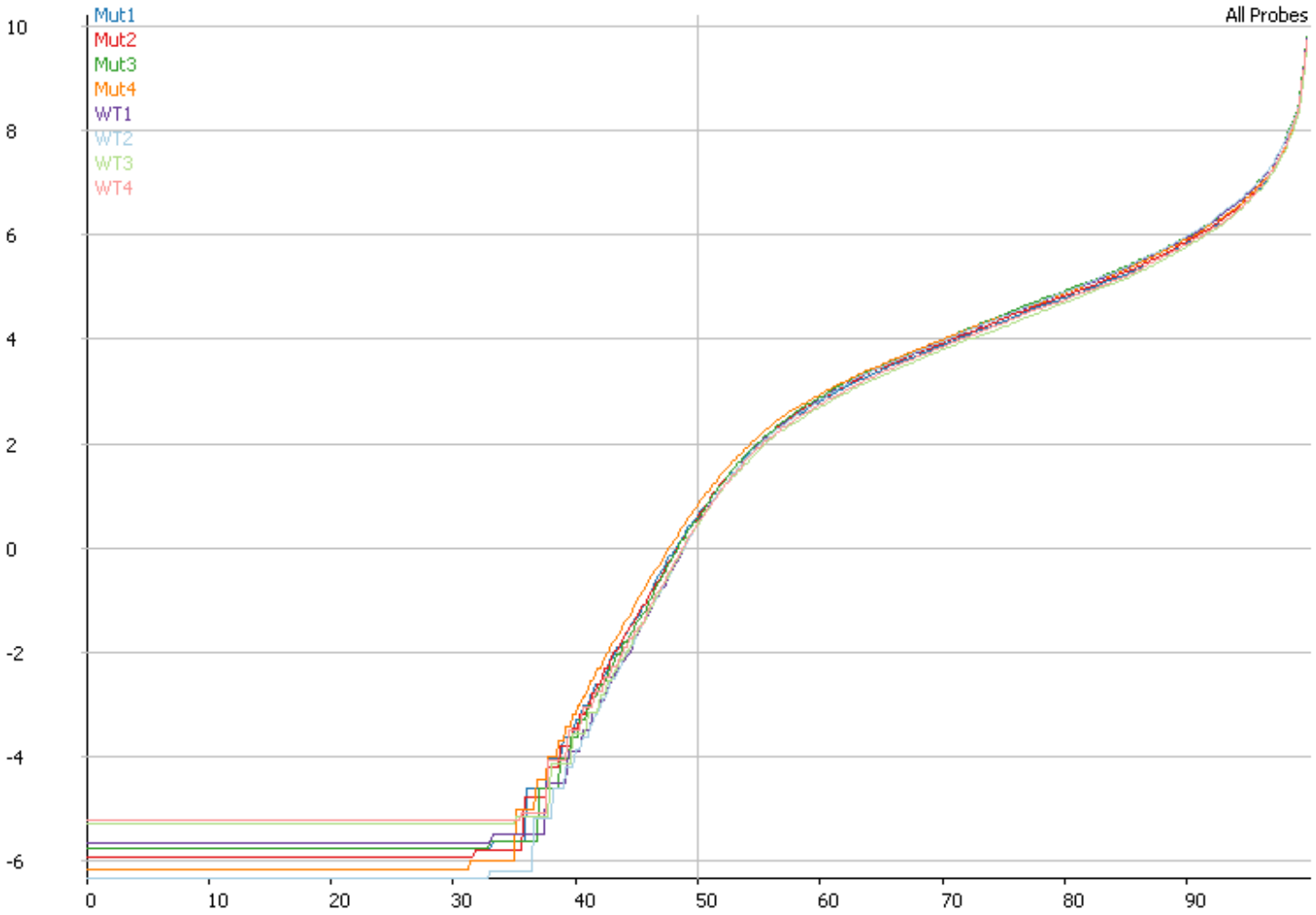
Babraham Bioinformatics

Normalisation – Coverage Outliers



Babraham Bioinformatics

Normalisation – DNA Contamination



Babraham Bioinformatics

Filtering Genes

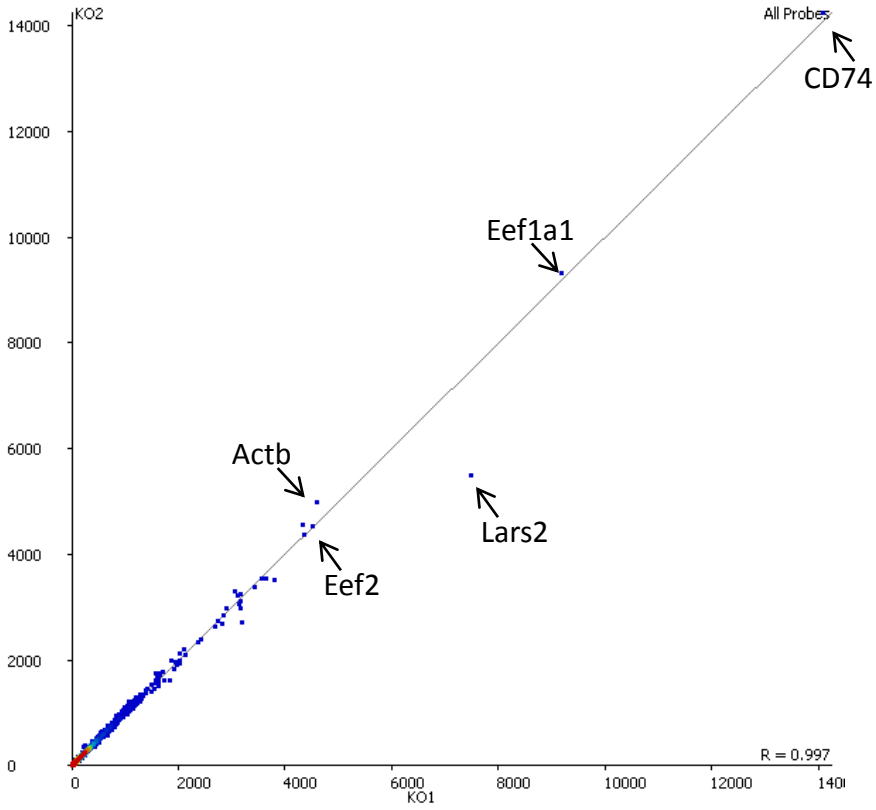
- Remove things which are uninteresting or shouldn't be measured

- Reduces noise – easier to achieve significance
  - Non-coding (miRNA, snoRNA etc) in RNA-Seq
  - Known mis-spliced forms (exon skipping etc)
  - Mitochondrial genes
  - X/Y chr genes in mixed sex populations
  - Unknown/Unannotated genes

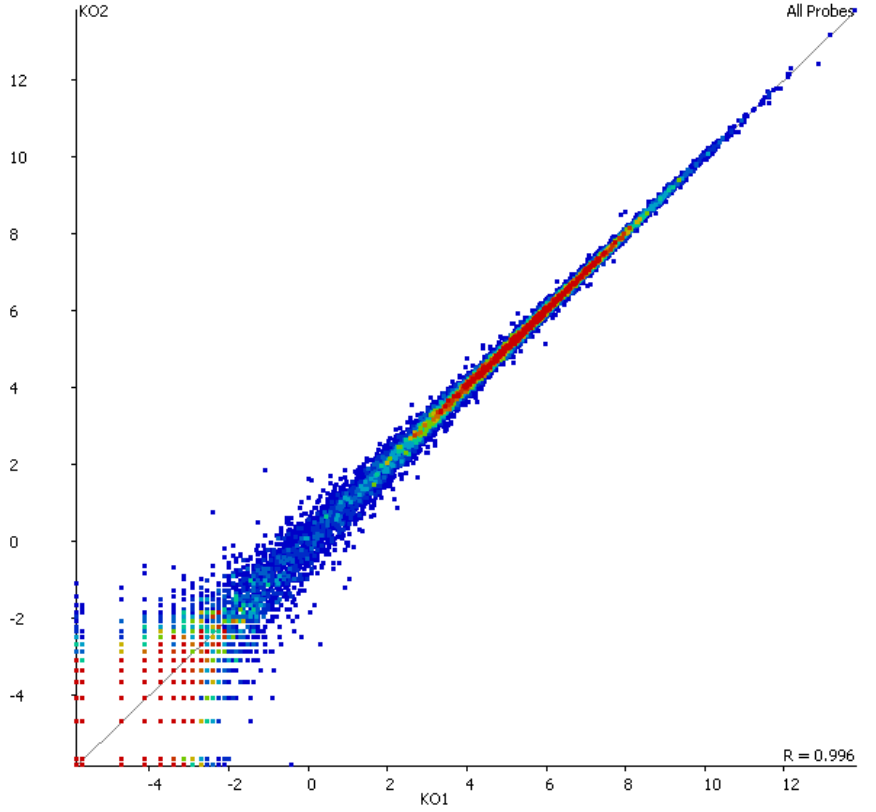
Babraham Bioinformatics

Visualising Expression

Linear



Log2

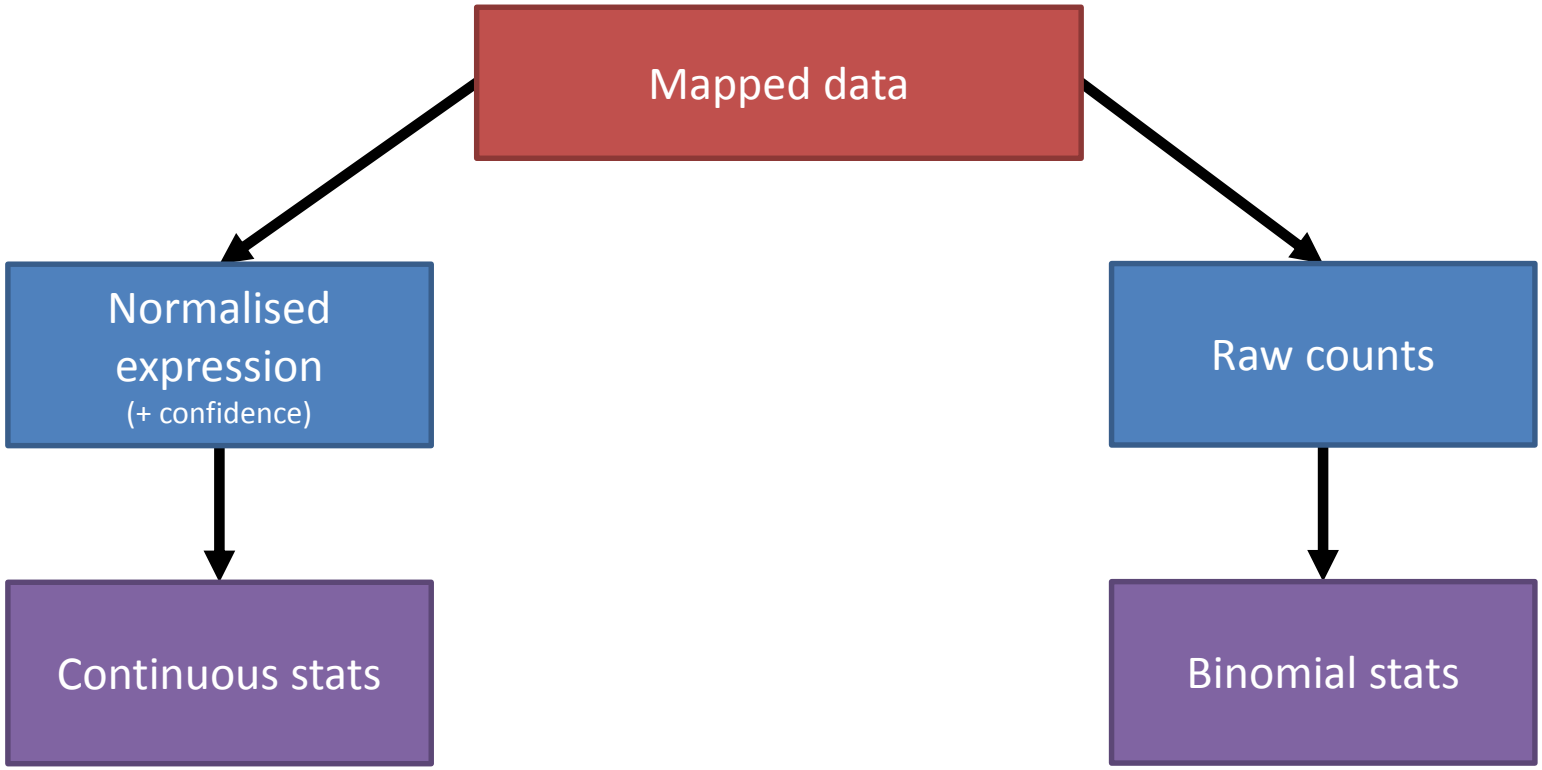


Babraham Bioinformatics

Differential Expression

Babraham Bioinformatics

Differential Expression



Babraham Bioinformatics

DE-Seq binomial Stats

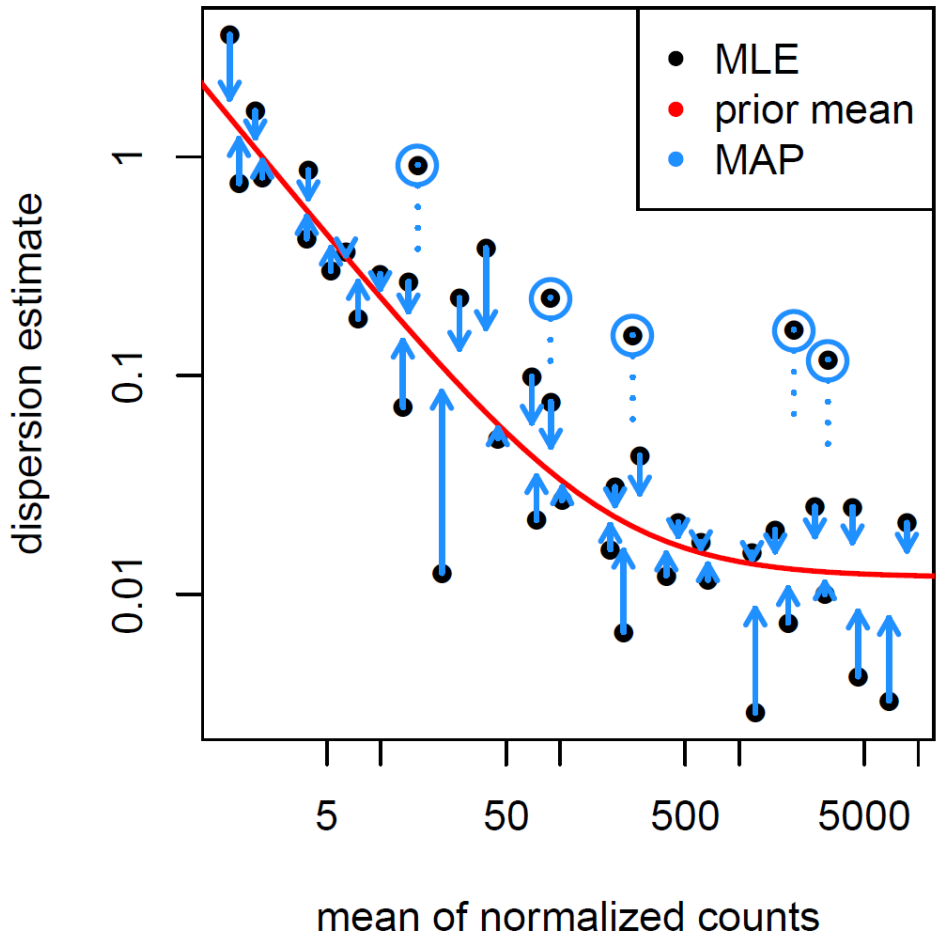
- Are the counts we see for gene X in condition 1 consistent with those for gene X in condition 2?

- Size factors
  - Estimator of library sampling depth
  - More stable measure than total coverage
  - Based on median ratio between conditions

- Variance – required for NB distribution
  - Insufficient observations to allow direct measure
  - Custom variance distribution fitted to real data
  - Smooth distribution assumed to allow fitting

Babraham Bioinformatics

Dispersion shrinkage



- Plot observed per gene dispersion
- Calculate average dispersion for genes with similar observation
- Individual dispersions regressed towards the mean. Weighted by
  - Distance from mean
  - Number of observations
- Points more than 2SD above the mean are not regressed

Babraham Bioinformatics

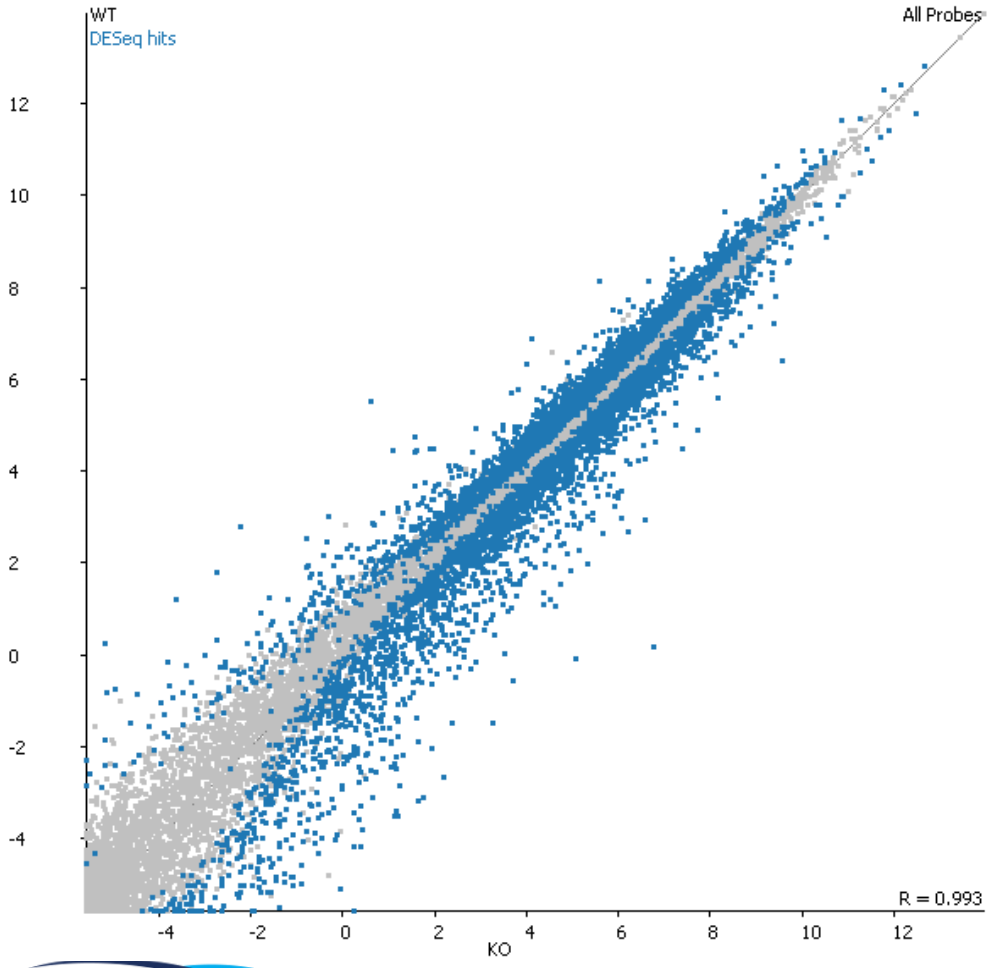
Other filters

- Cook's Distance – Identifies high variance
  - Effect on mean of removal of one replicate
  - For n<3 test not performed
  - For n = 3-6 failures are removed
  - For n>6 outliers removed to make trimmed mean
  - Disable with cooksCutoff=FALSE

- Hit count optimisation
  - Low intensity reads are removed
  - Limits multiple-testing to give max significant hits
  - Disable with independentFiltering=FALSE

Babraham Bioinformatics

Visualising Differential Expression Results



5x5 Replicates

5,000 out of 22,000 genes (23%) identified as DE using DESeq ($p<0.05$)

Need further filtering!

Babraham Bioinformatics

Magnitude of effect filtering Intensity difference test

- Fold change biases to low expression

- Need a test which
  - Has a statistical basis
  - Doesn’t bias by expression level
  - Returns sensible numbers of hits

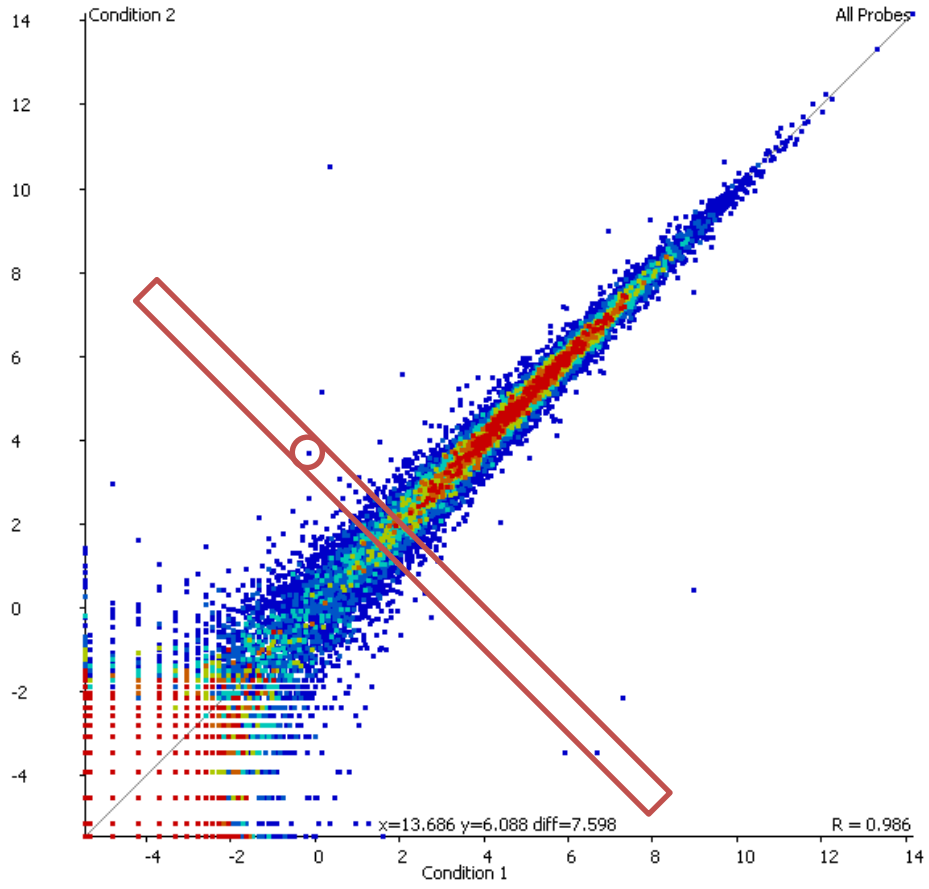
Babraham Bioinformatics

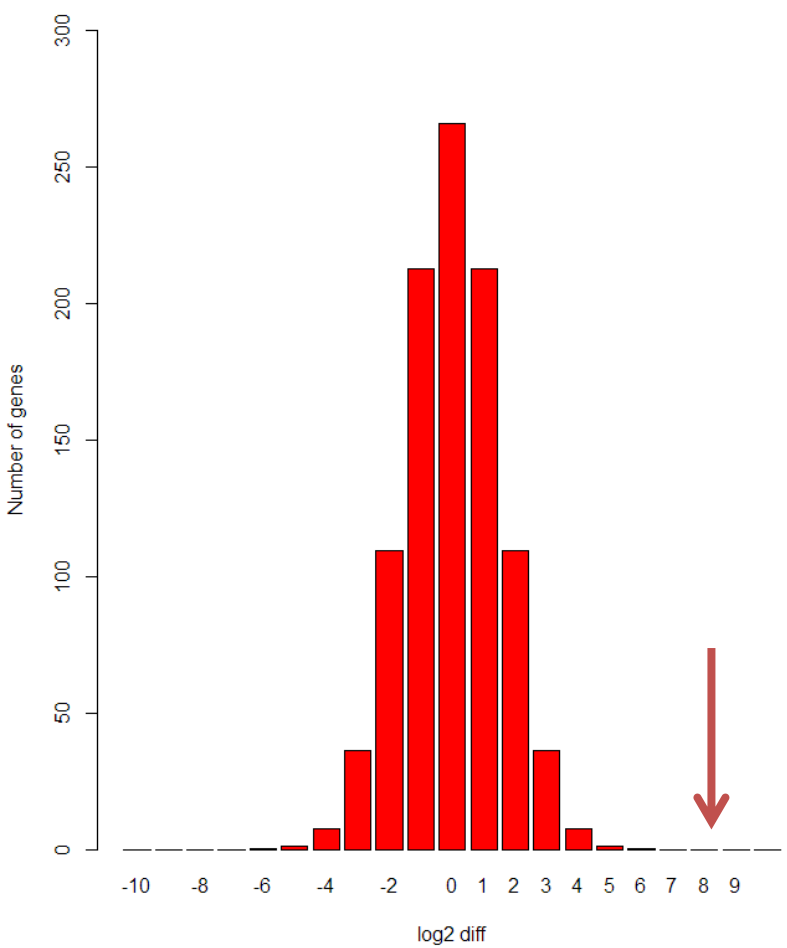
Assumptions

- Noise is related to observation level
  - Similar to DESeq
- Differences between conditions are either
  - A direct response to stimulus
  - Noise, either technical ***or biological***
- Find points whose differences aren't explained by general disruption

Babraham Bioinformatics

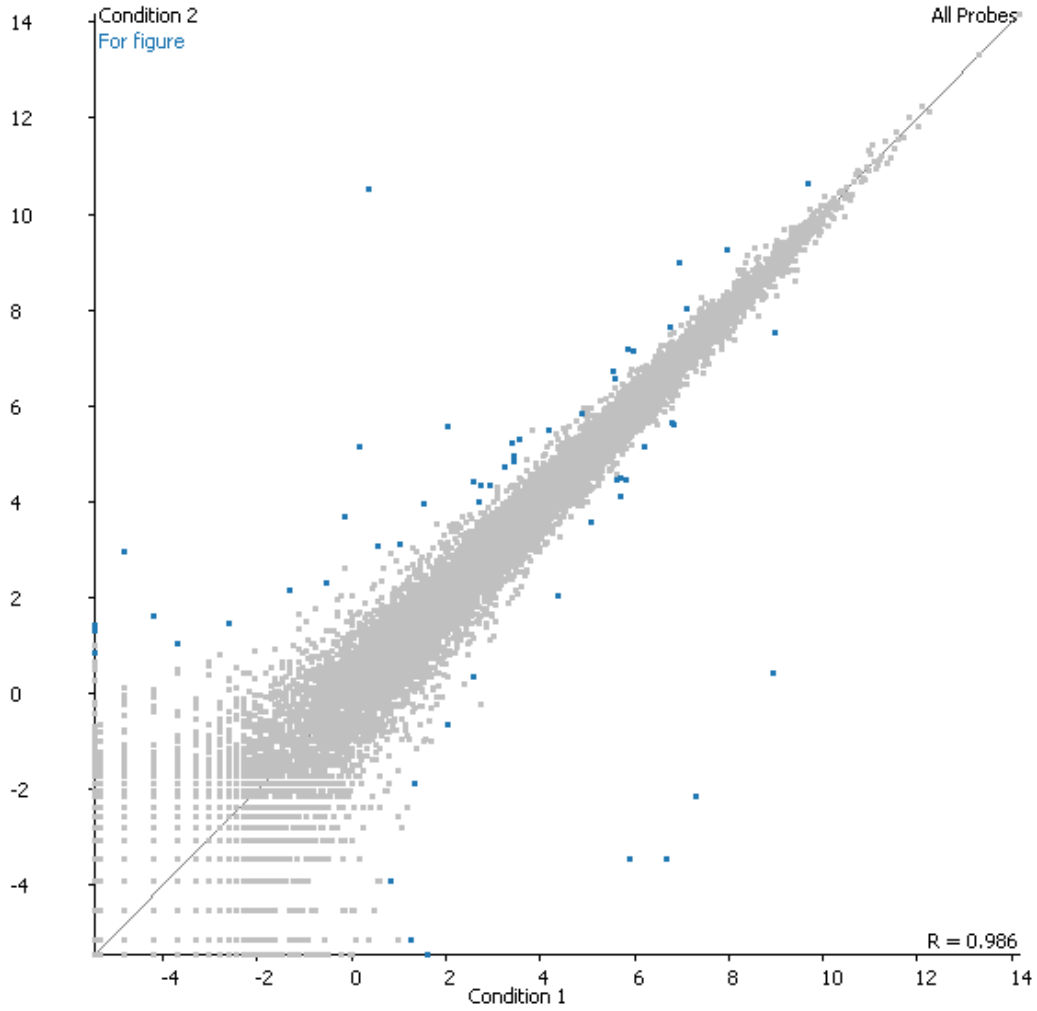
Method





Babraham Bioinformatics

Results



Babraham Bioinformatics

Experimental Design

Babraham Bioinformatics

Practical Experiment Design

- What type of library?
- What type of sequencing?
- How many reads?
- How many replicates?

Babraham Bioinformatics

What type of library?

- Directional libraries if possible
  - Easier to spot contamination
  - No mixed signals from antisense transcription
  - May be difficult for low input samples

- mRNA vs total vs depletion etc.
  - Down to experimental questions
  - Remember LINC RNA may not have polyA tail
  - Active transcription vs standing mRNA pool

Babraham Bioinformatics

What type of sequencing

- Depends on your interest
  - Expression quantitation of known genes
    - 50bp single end is fine
  - Expression plus splice junction usage
    - 100bp (or longer if possible) single end
  - Novel transcript discovery
    - 100bp paired end

Babraham Bioinformatics

How many reads

- Typically aim for 20 million reads for human / mouse sized genome

- More reads:
  - De-novo discovery
  - Low expressed transcripts

- More replicates more useful than more reads

Babraham Bioinformatics

Replicates

- Compared to arrays, RNA-Seq is a very clean technical measure of expression
  - Generally don't run **technical** replicates
  - Must run **biological** replicates
- For clean systems (eg cell lines) 3x3 or 4x4 is common
- Higher numbers required as the system gets more variable
- Always plan for at least one sample to fail
- Randomise across sample groups

Babraham Bioinformatics

Exercises

- Look at raw QC
- Mapping with HiSat2
  - Small test data
- Quantitation and visualisation
  - Use SeqMonk graphical program
  - Larger replicated data
- Differential expression
  - DESeq2
  - Intensity Difference
- Review in SeqMonk

Babraham Bioinformatics

Useful links

- FastQC http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
- HiSat2 https://ccb.jhu.edu/software/hisat2/
- SeqMonk http://www.bioinformatics.babraham.ac.uk/projects/seqmonk/
- Cufflinks http://cufflinks.cbcb.umd.edu/
- DESeq2 https://bioconductor.org/packages/release/bioc/html/**DESeq2**.html
- Bioconductor http://www.bioconductor.org/
- DupRadar http://sourceforge.net/projects/dupradar/

Babraham Bioinformatics