

# RNA-Seq Analysis

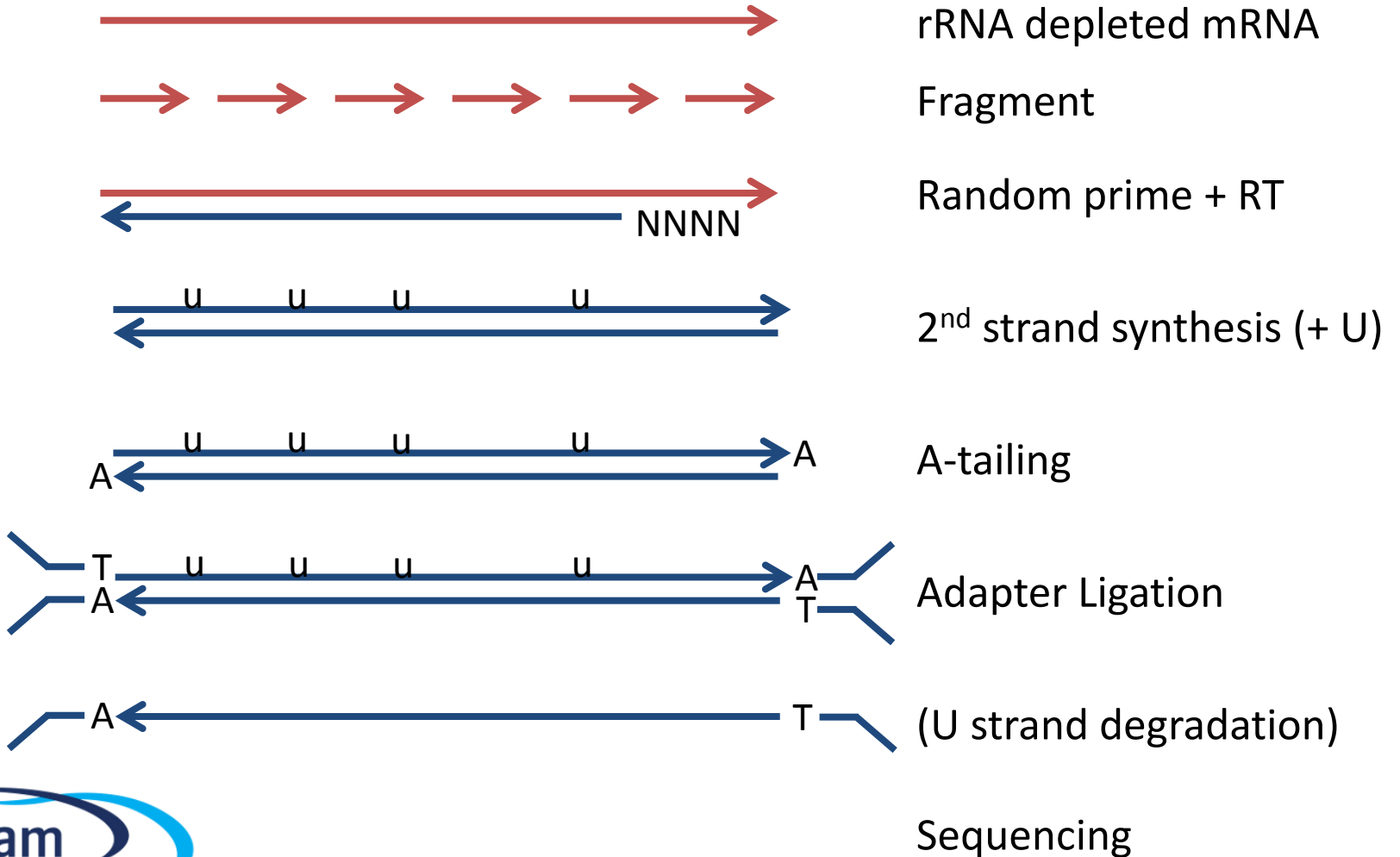
Simon Andrews

[simon.andrews@babraham.ac.uk](mailto:simon.andrews@babraham.ac.uk)

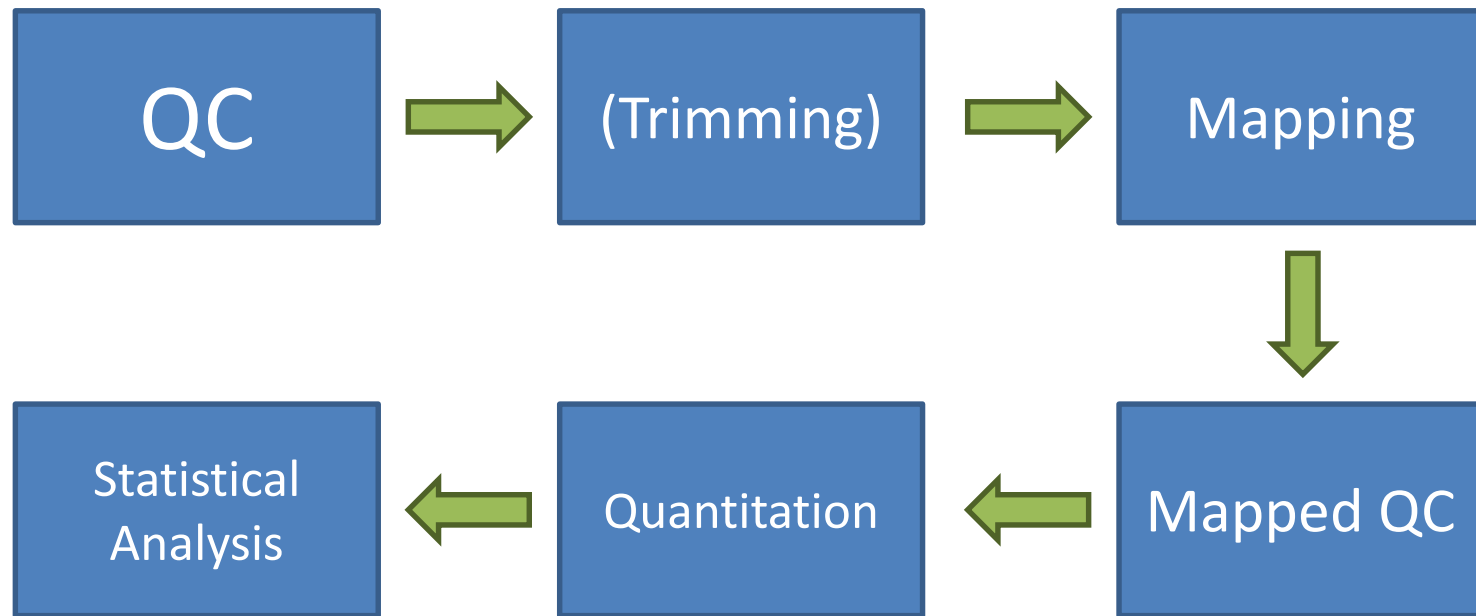
[@simon\\_andrews](#)

v2017-10

# RNA-Seq Libraries



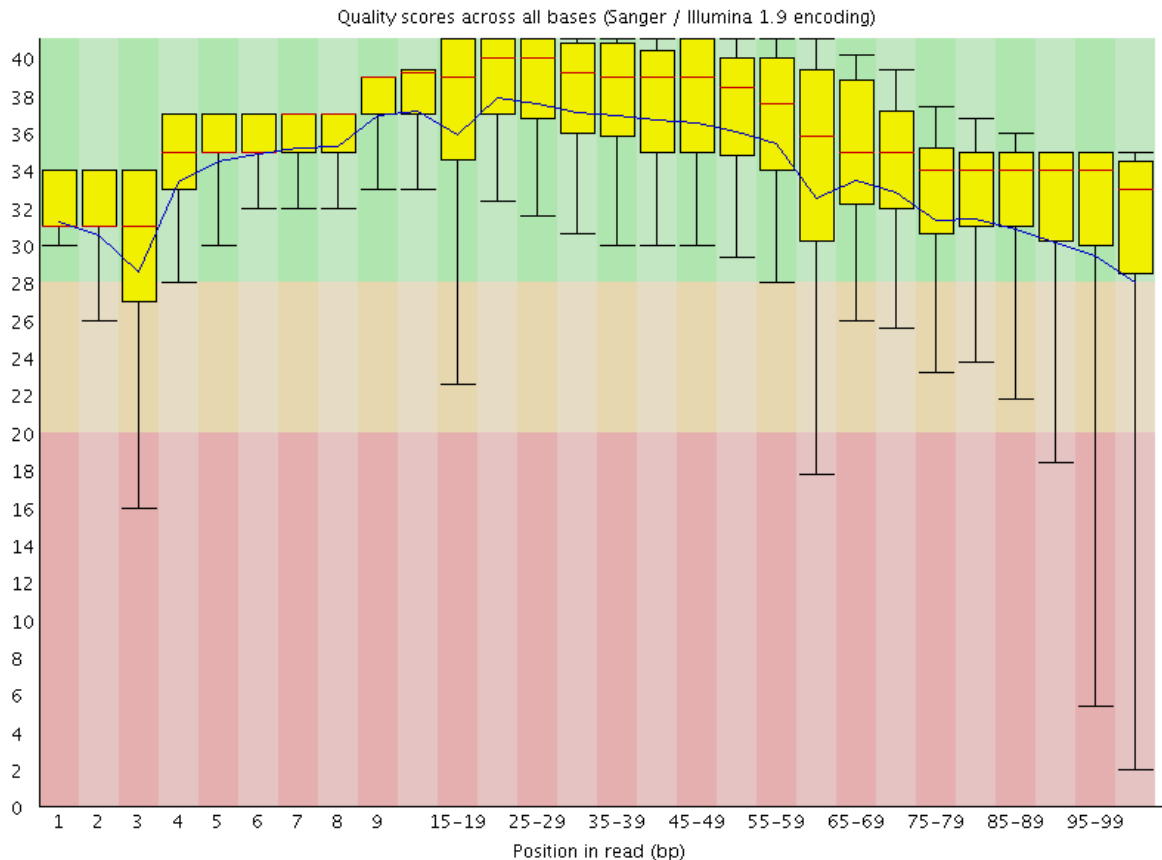
# Reference based RNA-Seq Analysis



# Quality Control

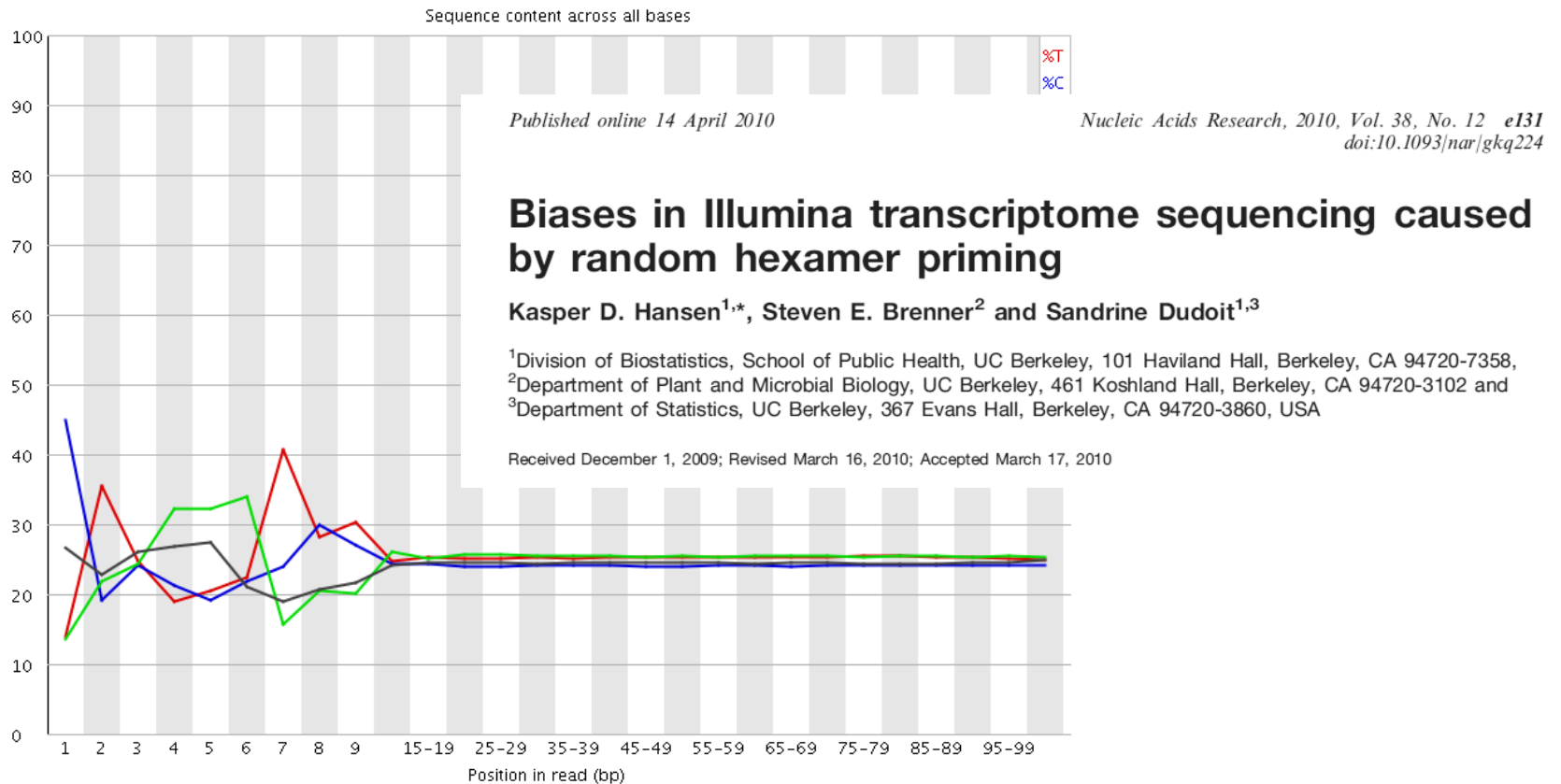
# QC: Raw Data

- Sequence call quality



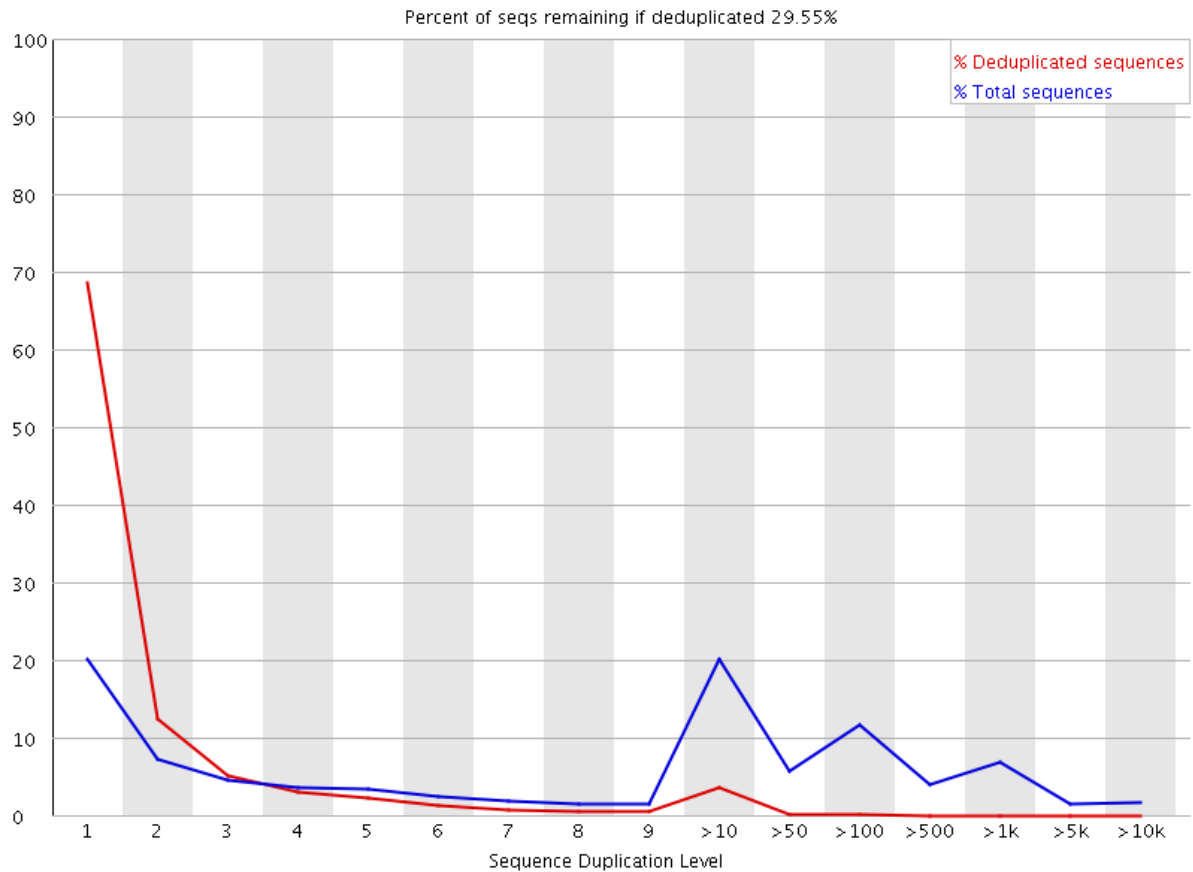
# QC: Raw Data

- Sequence bias



# QC: Raw Data

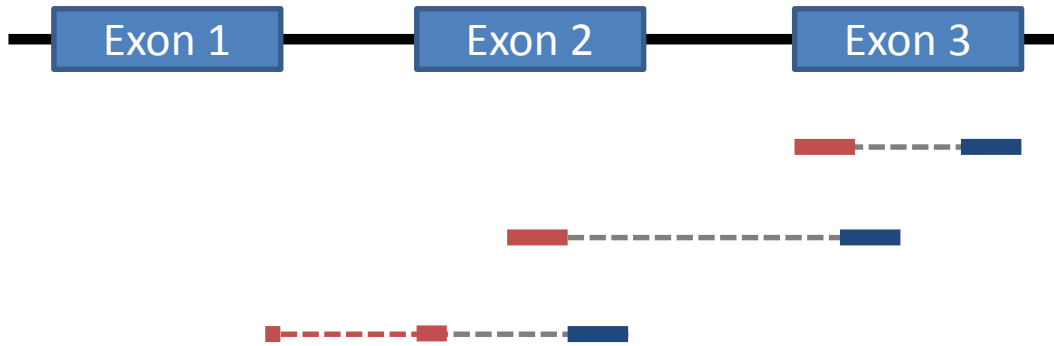
- Duplication level



# Mapping to a reference



# Mapping



Genome

Simple mapping within exons

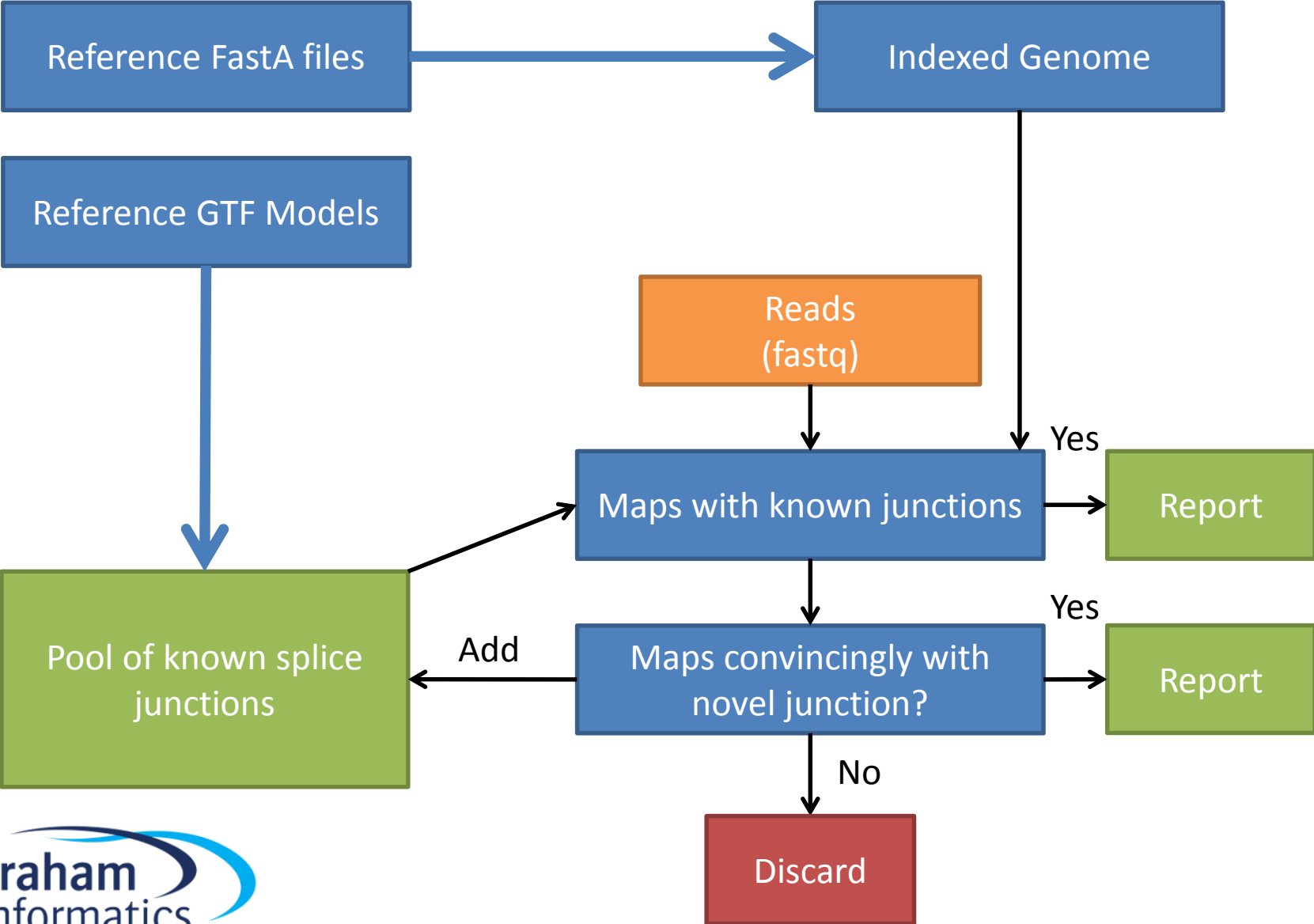
Mapping between exons

Spliced mapping

# 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/>)

# HiSat2 pipeline



# Quality Control (mapped data)

# Mapping Statistics

Time loading forward index: 00:01:10

Time loading reference: 00:00:05

Multiseed full-index search: 00:20:47

24548251 reads; of these:

24548251 (100.00%) were paired; of these:

1472534 (6.00%) aligned concordantly 0 times

21491188 (87.55%) aligned concordantly exactly 1 time

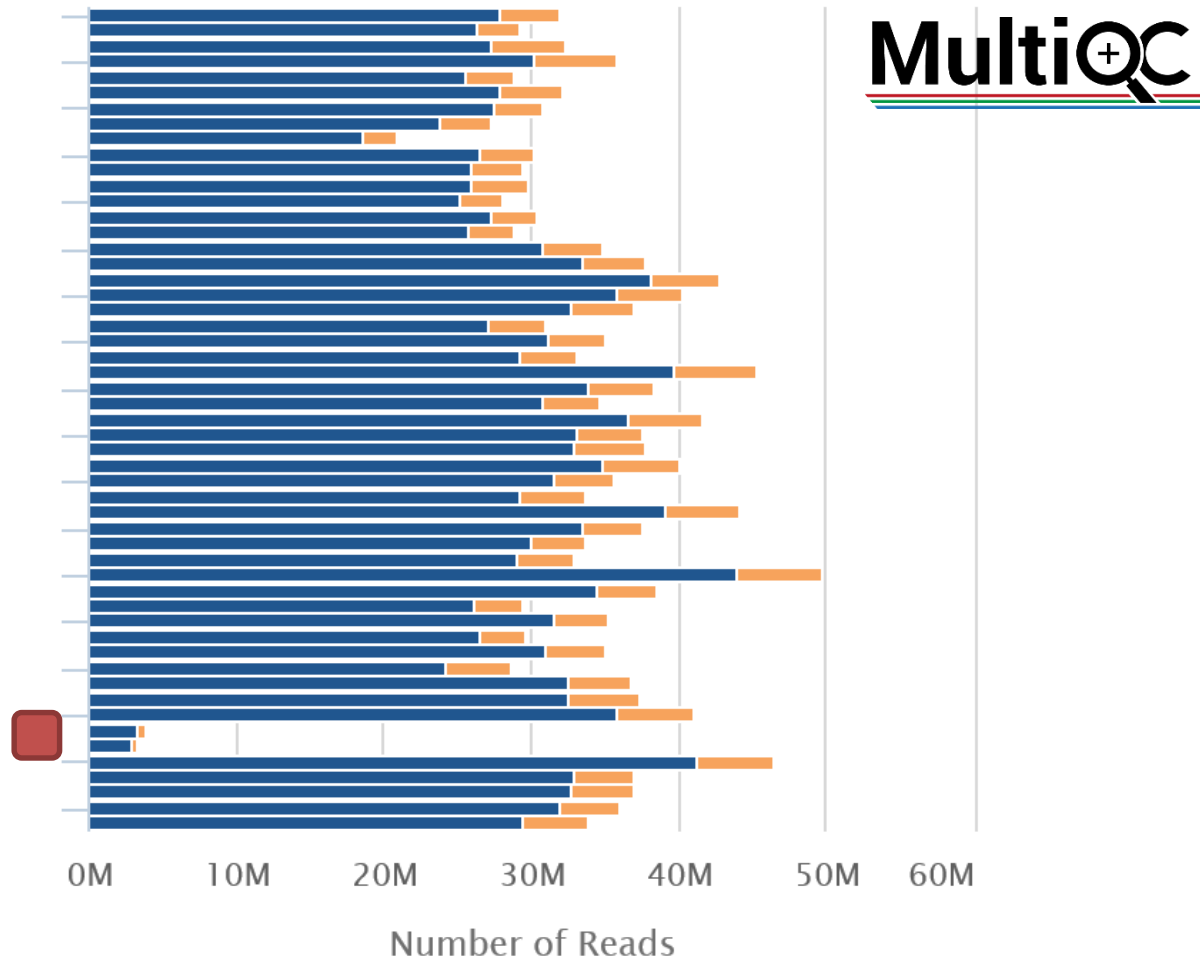
1584529 (6.45%) aligned concordantly >1 times

94.00% overall alignment rate

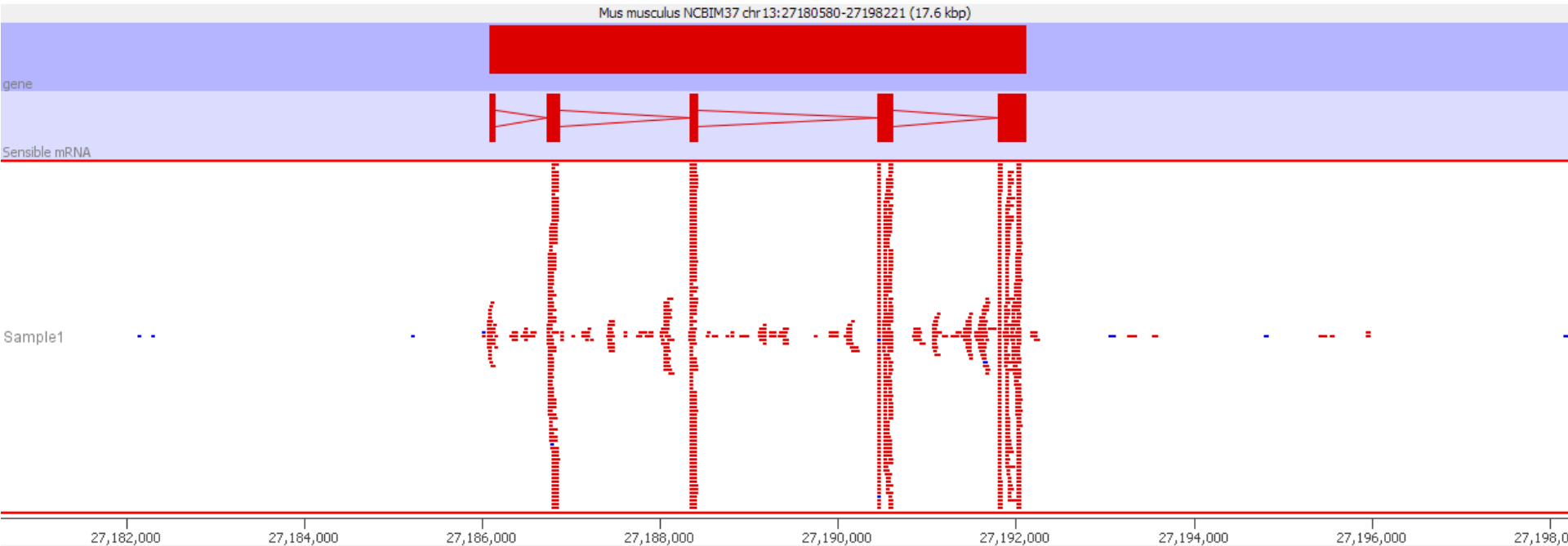
Time searching: 00:20:52

Overall time: 00:22:02

# Mapping Statistics

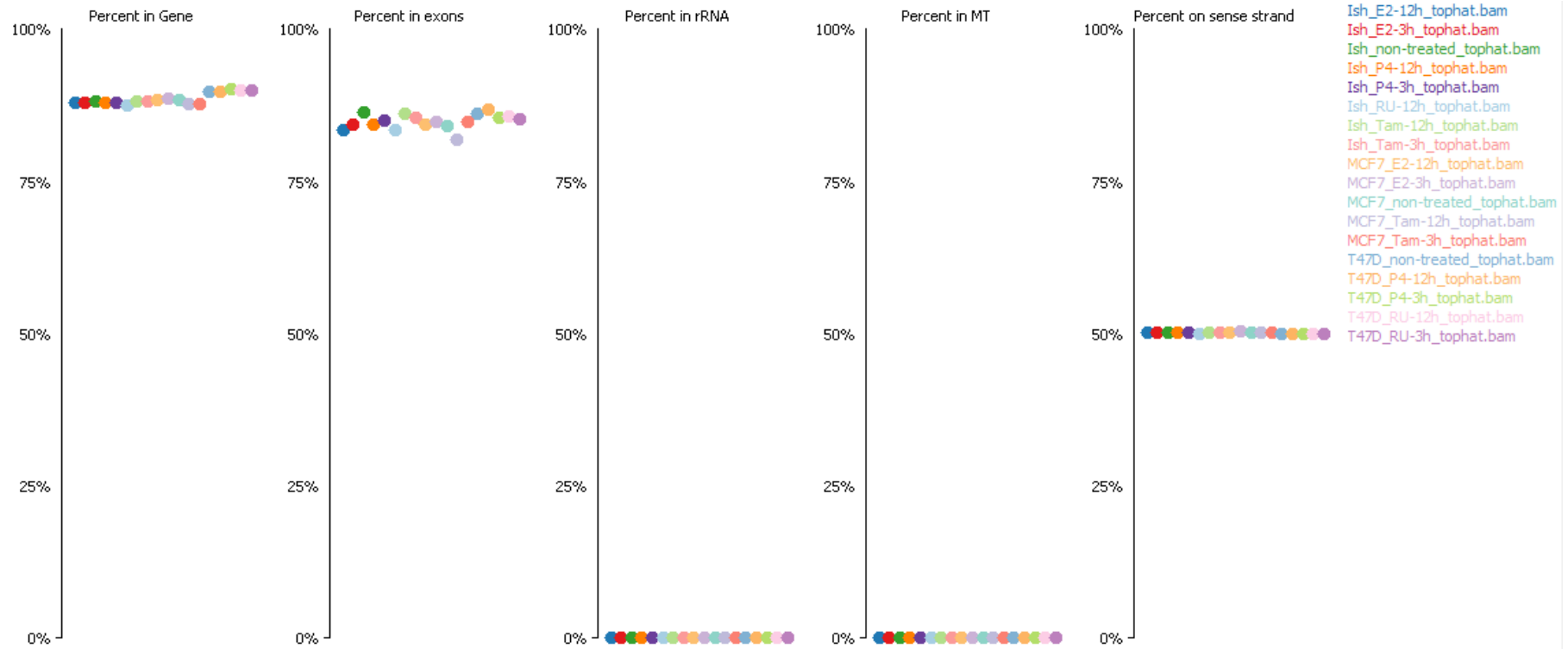


# Viewing Mapped Data



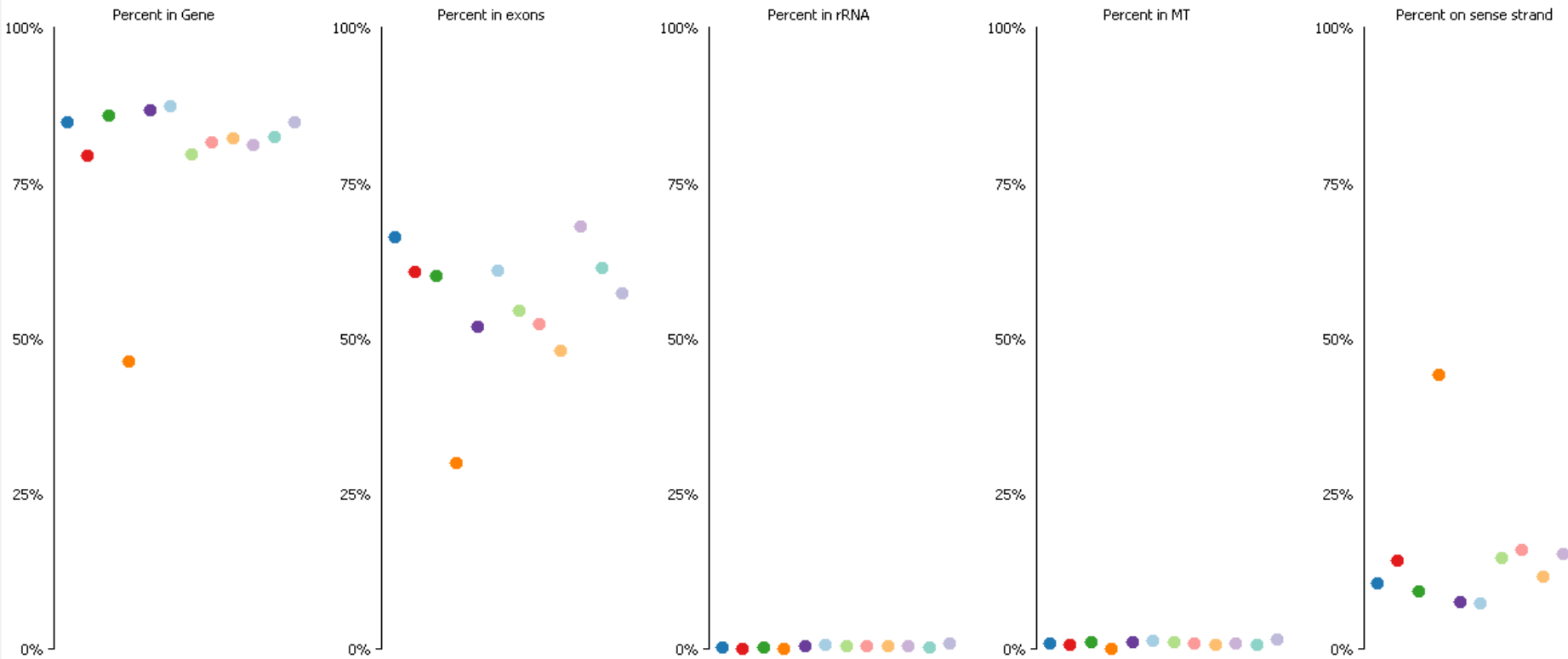
- Reads over exons
- Reads over introns
- Reads in intergenic
- Strand specificity

# SeqMonk RNA-Seq QC (good)

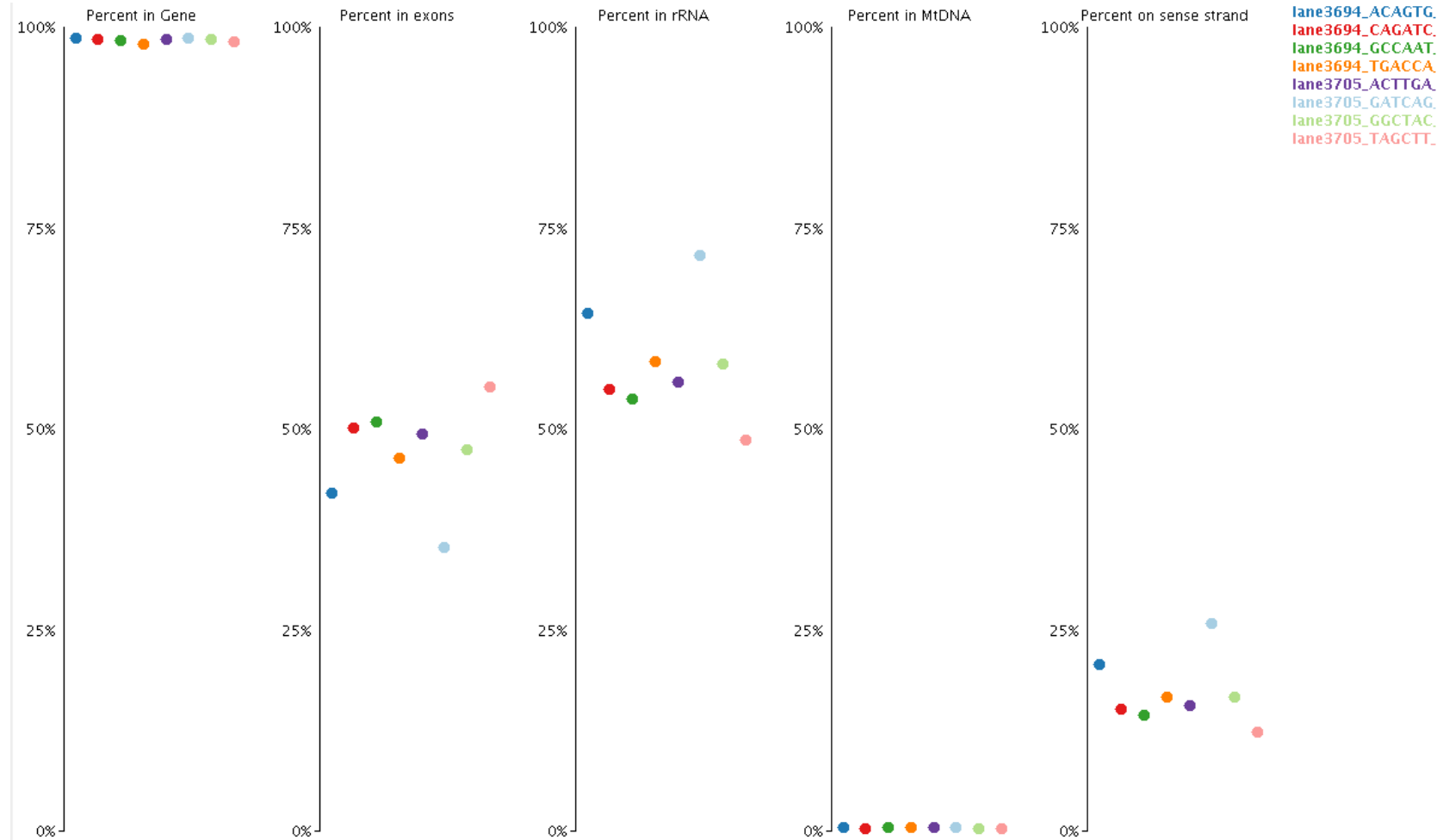




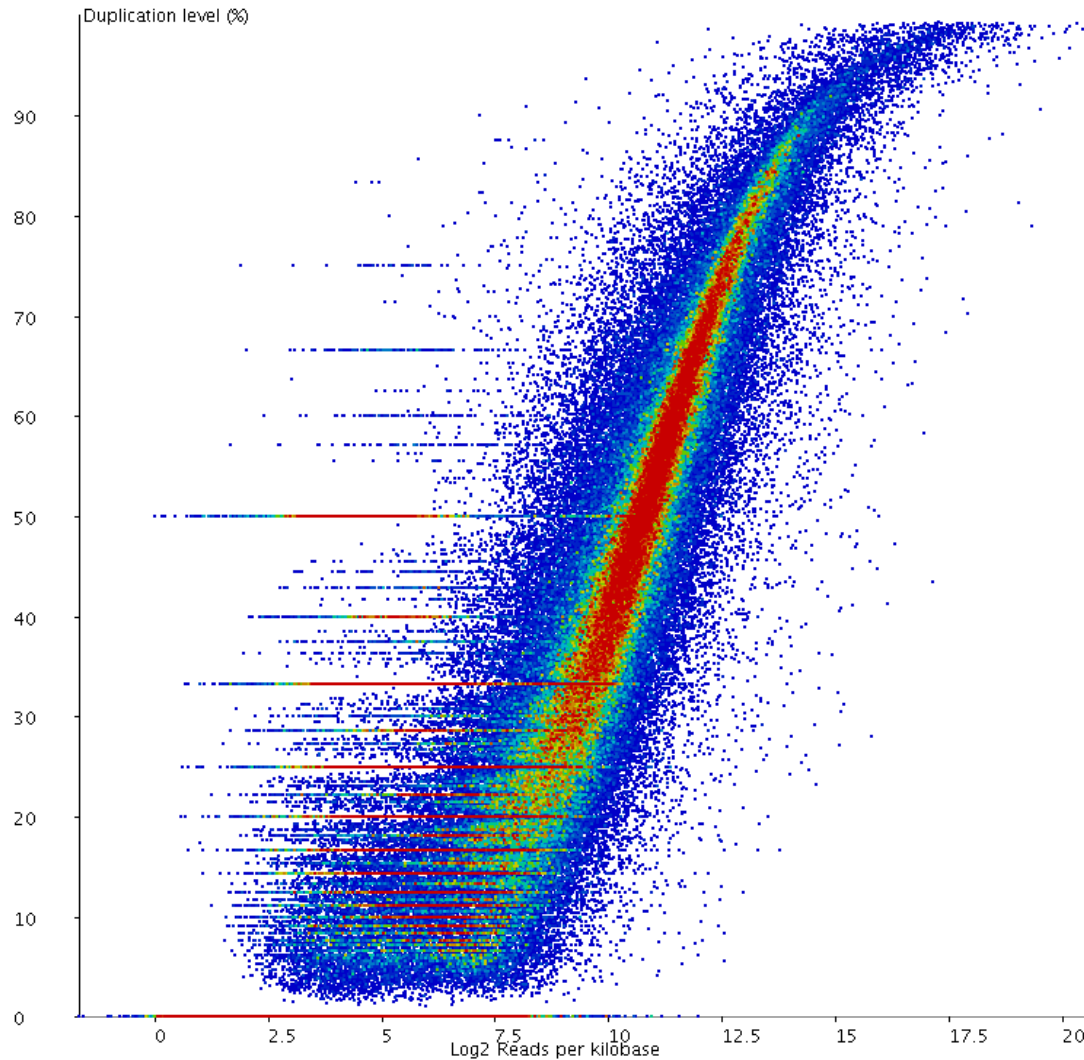
# SeqMonk RNA-Seq QC (bad)



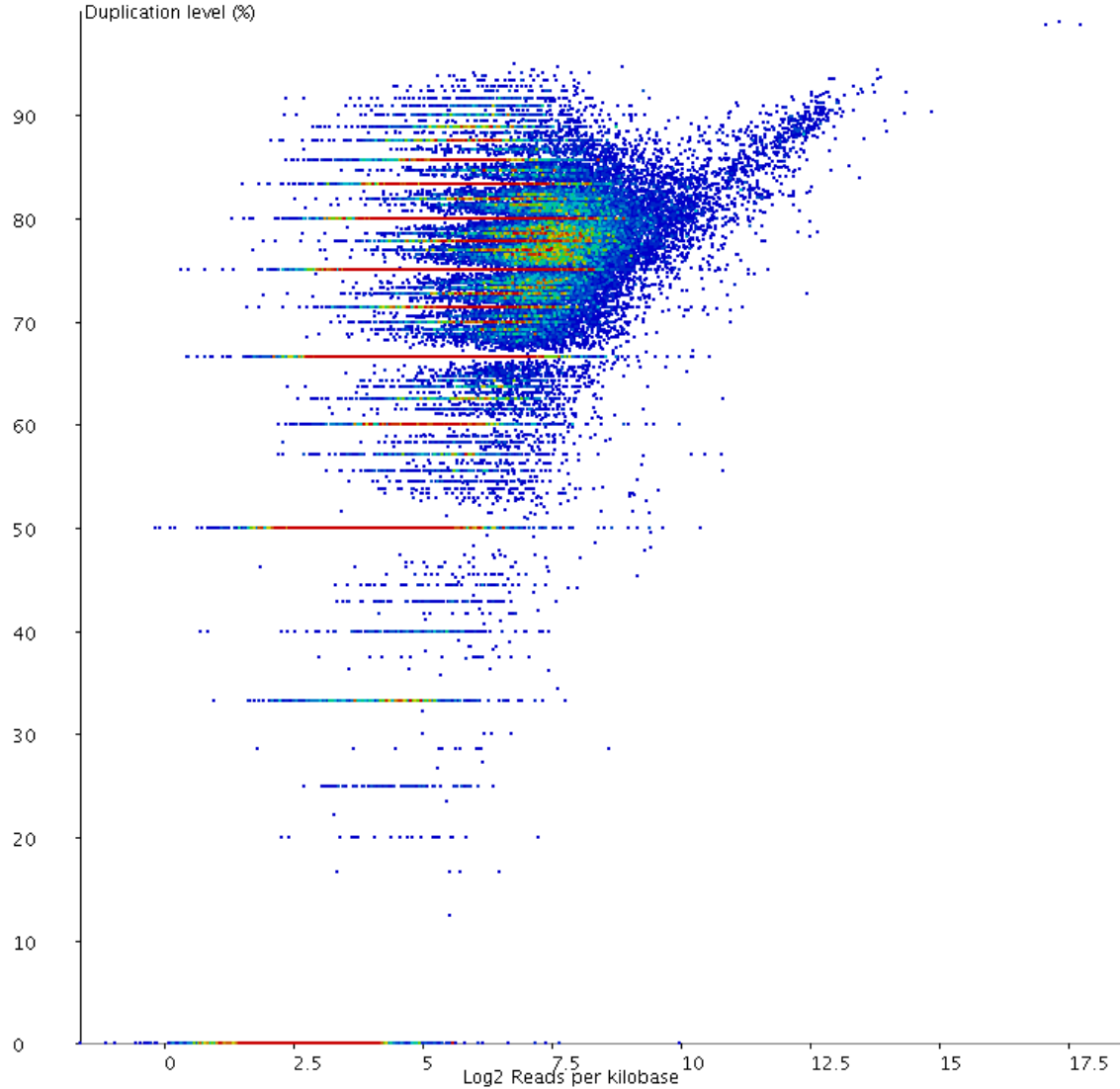
# SeqMonk RNA-Seq QC (bad)



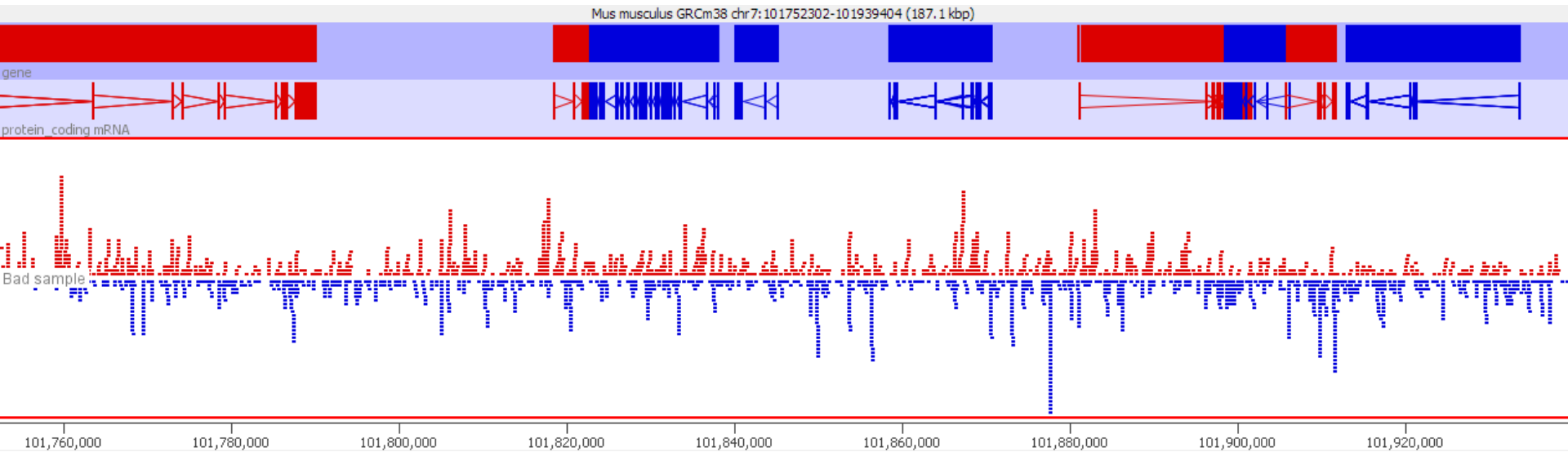
# Duplication (good)



# Duplication (bad)

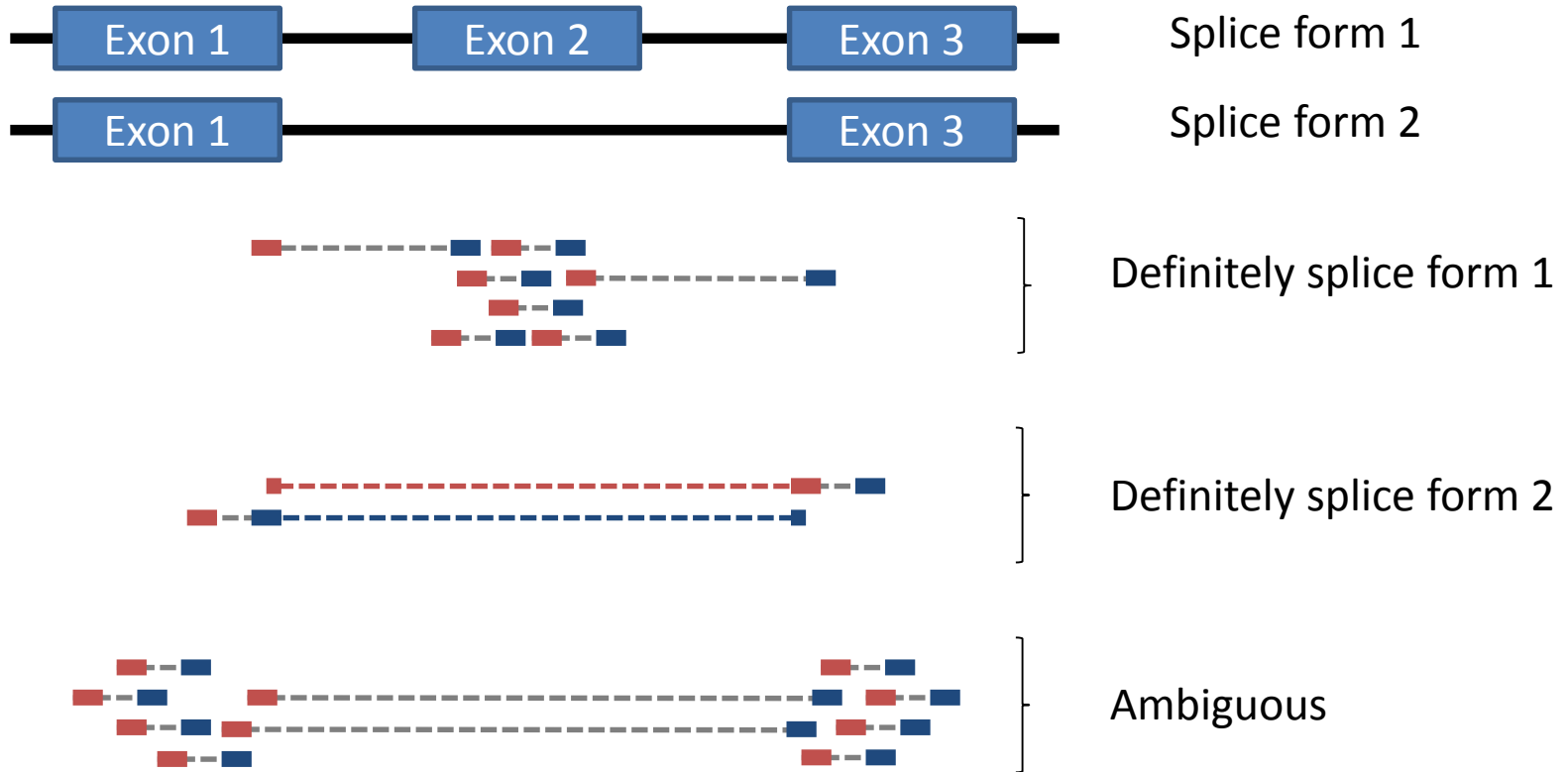


# Look at poor QC samples



# Quantitation

# Quantitation



# 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



# Analysing Splicing

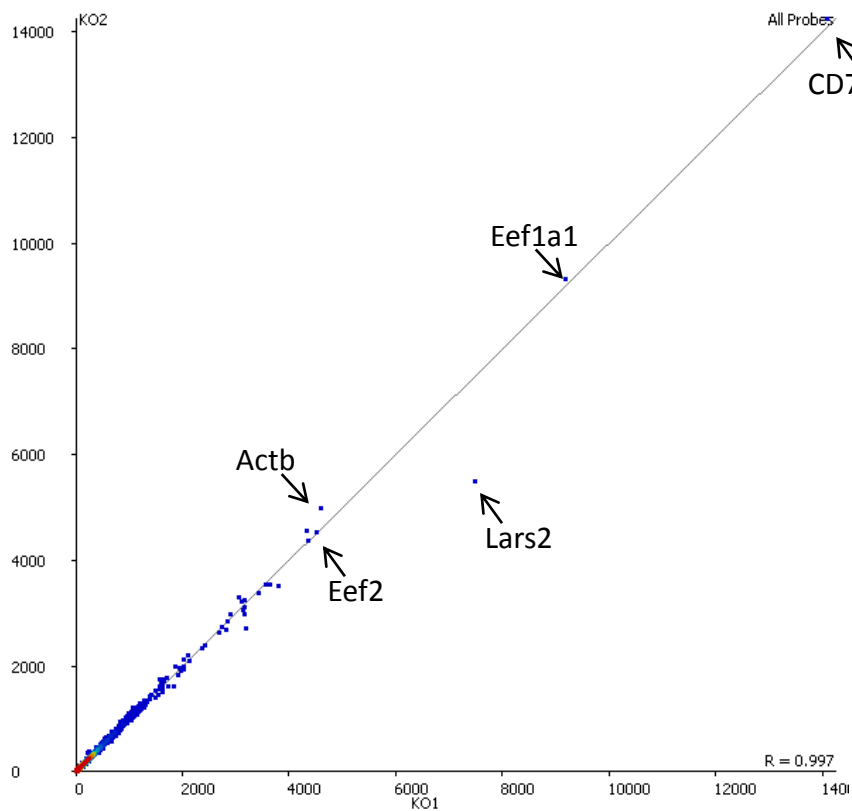
- Estimate transcript level expression
  - Cufflinks, RSEM, bitSeq etc.
- Analyse junction usage directly
  - Counts + DESeq/EdgeR
- Compare splicing decision ratios
  - rMATS, logistic regression

# RPKM / FPKM / TPM

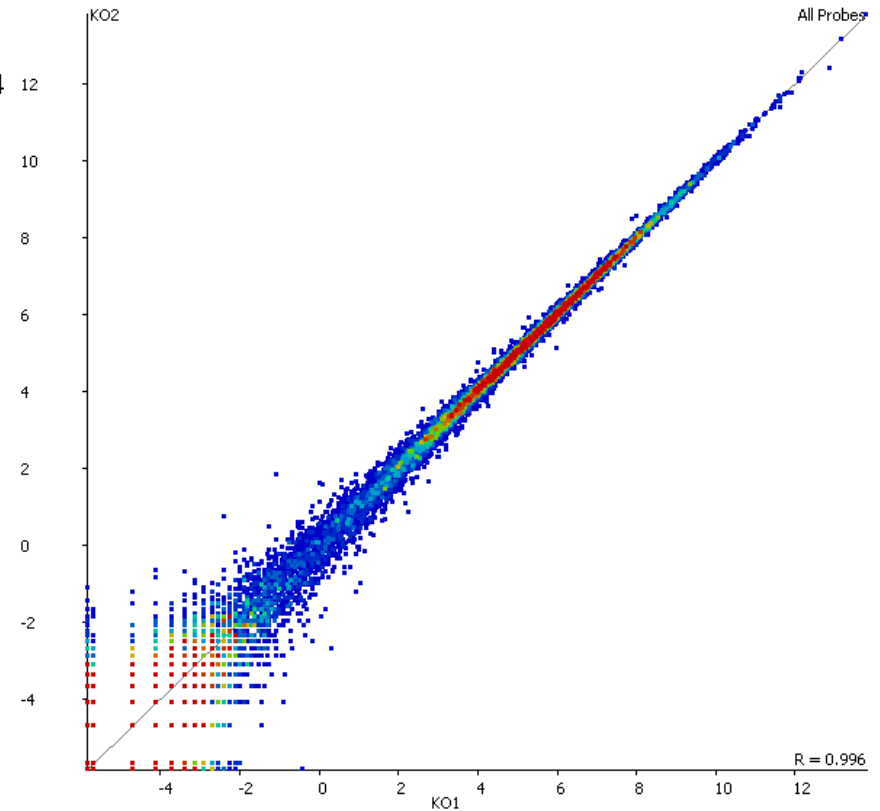
- **RPKM** (Reads per kilobase of transcript per million reads of library)
  - Corrects for total library coverage
  - Corrects for gene length
  - Comparable between different genes within the same dataset
- **FPKM** (Fragments per kilobase of transcript per million fragments of library)
  - Only relevant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

# Visualising Expression

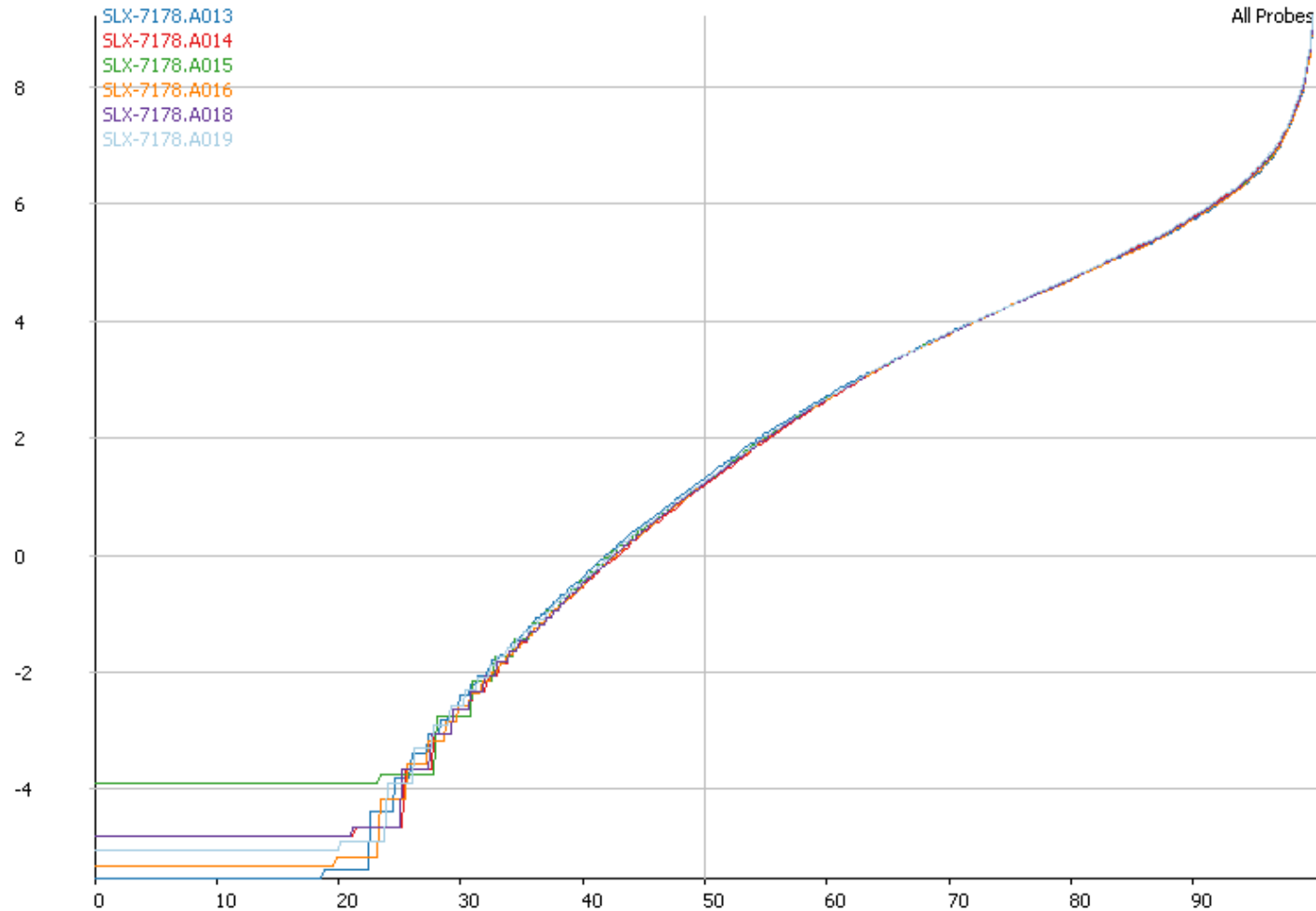
Linear



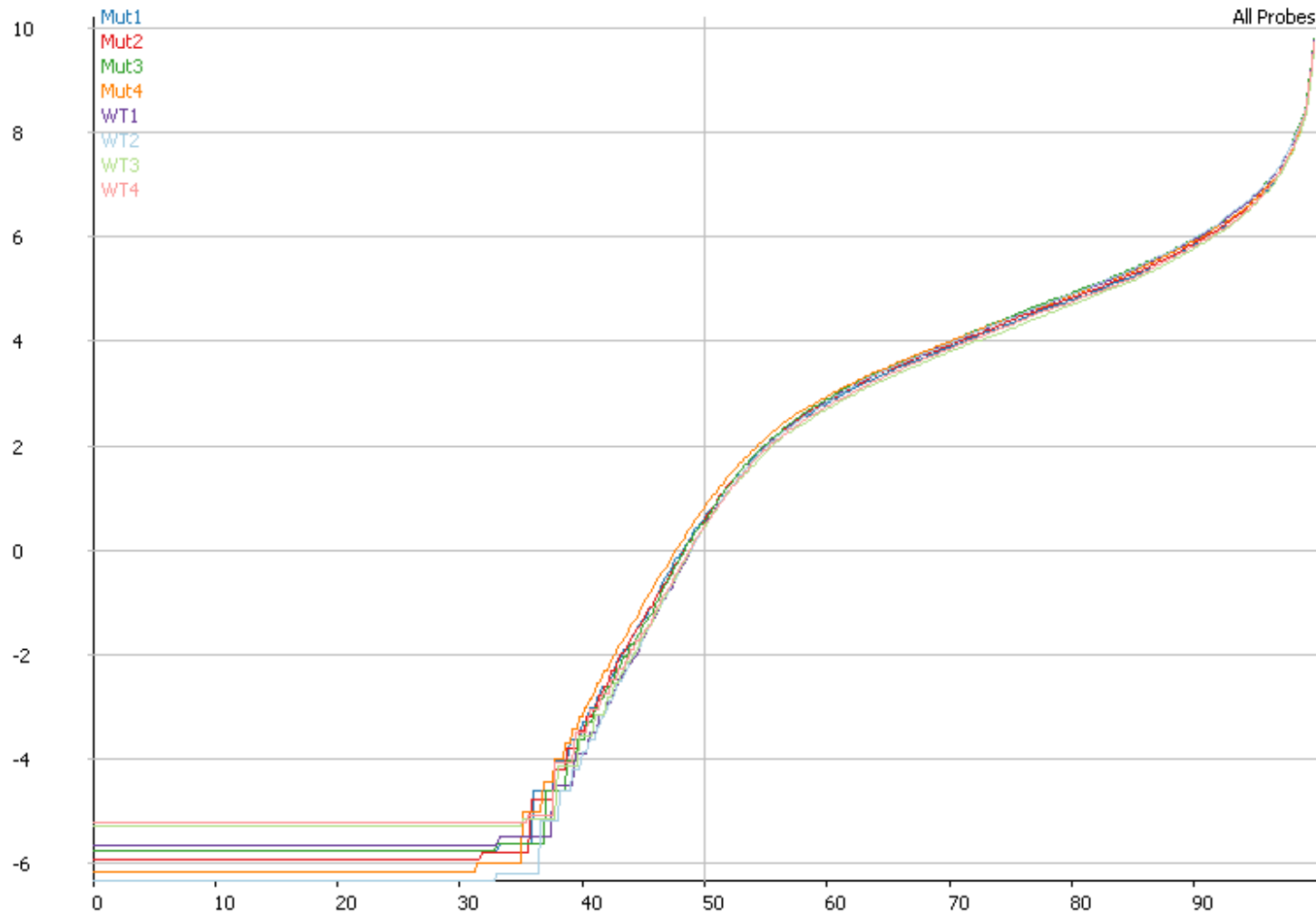
Log2



# Normalisation – Coverage Outliers



# Normalisation – DNA Contamination

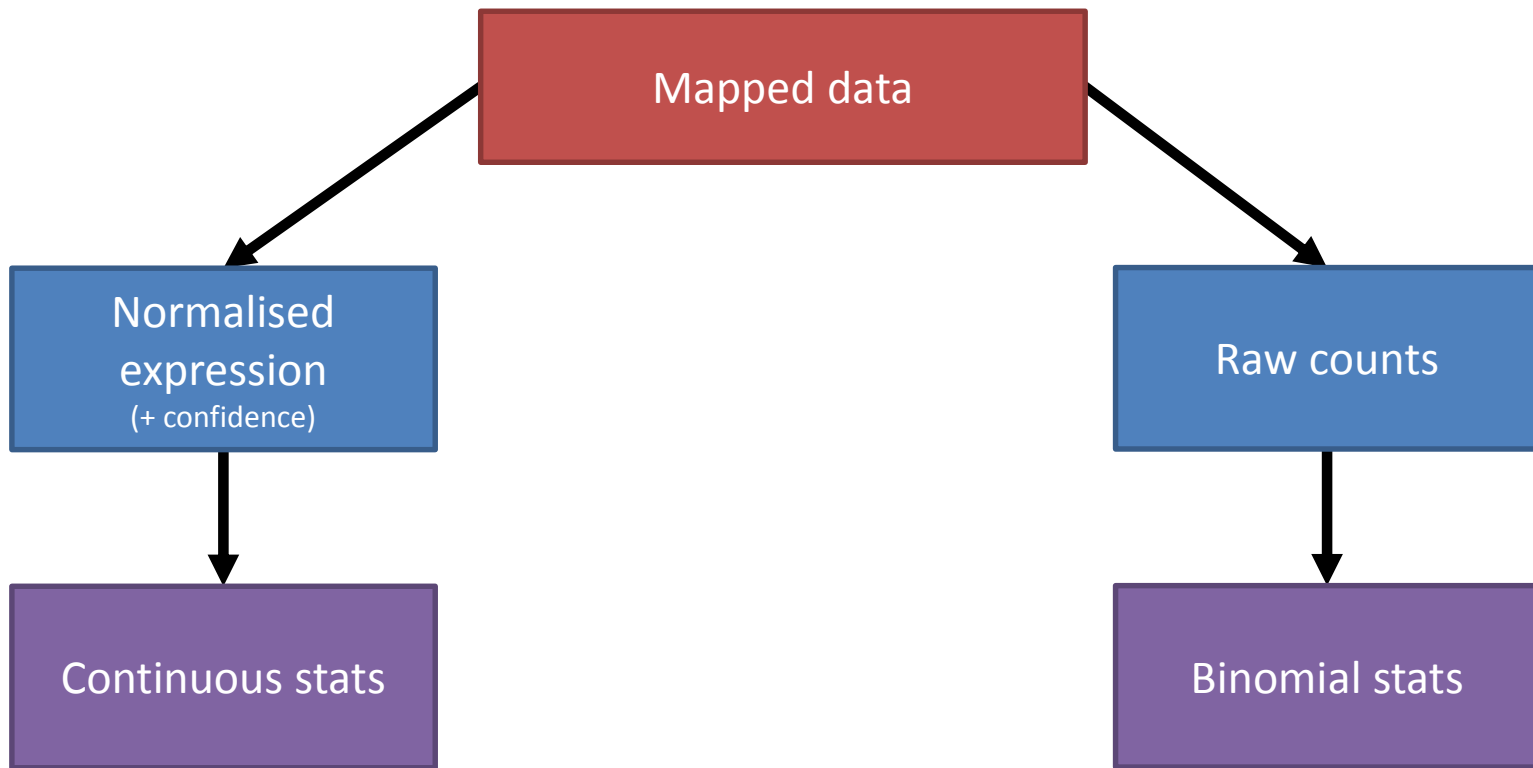


# Filtering Genes

- 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

# Differential Expression

# Differential Expression

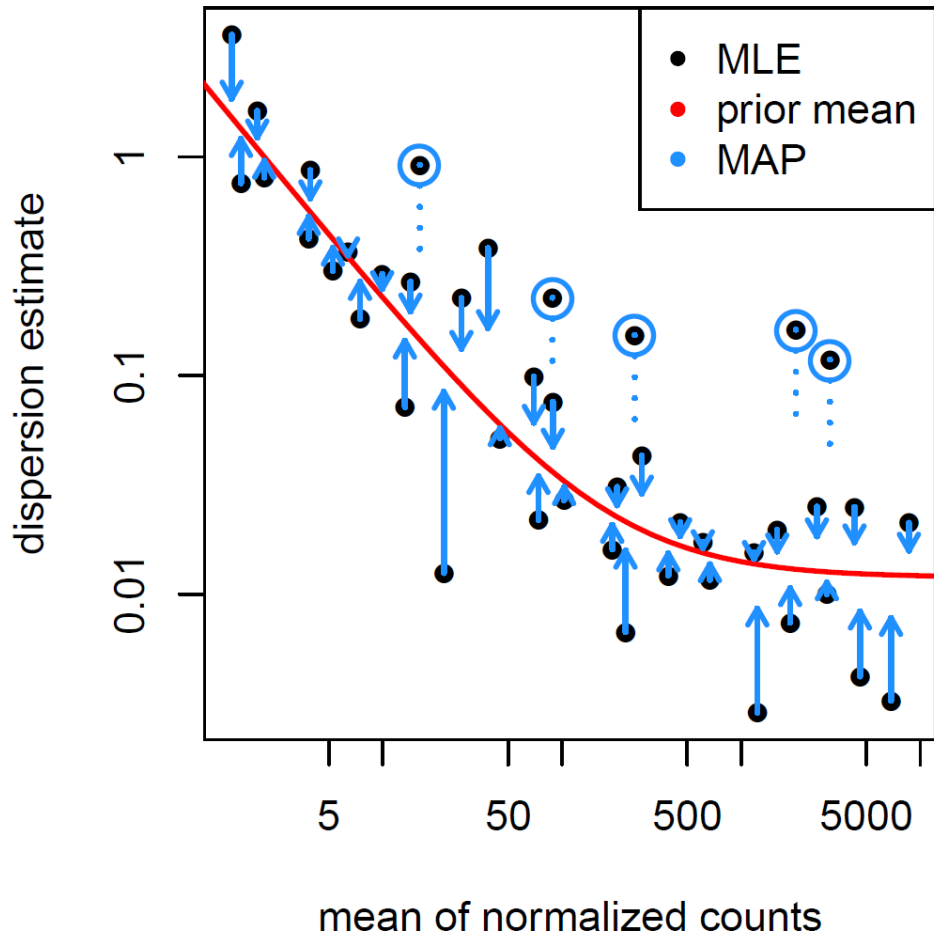




# DE-Seq binomial Stats

- 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

# Dispersion shrinkage

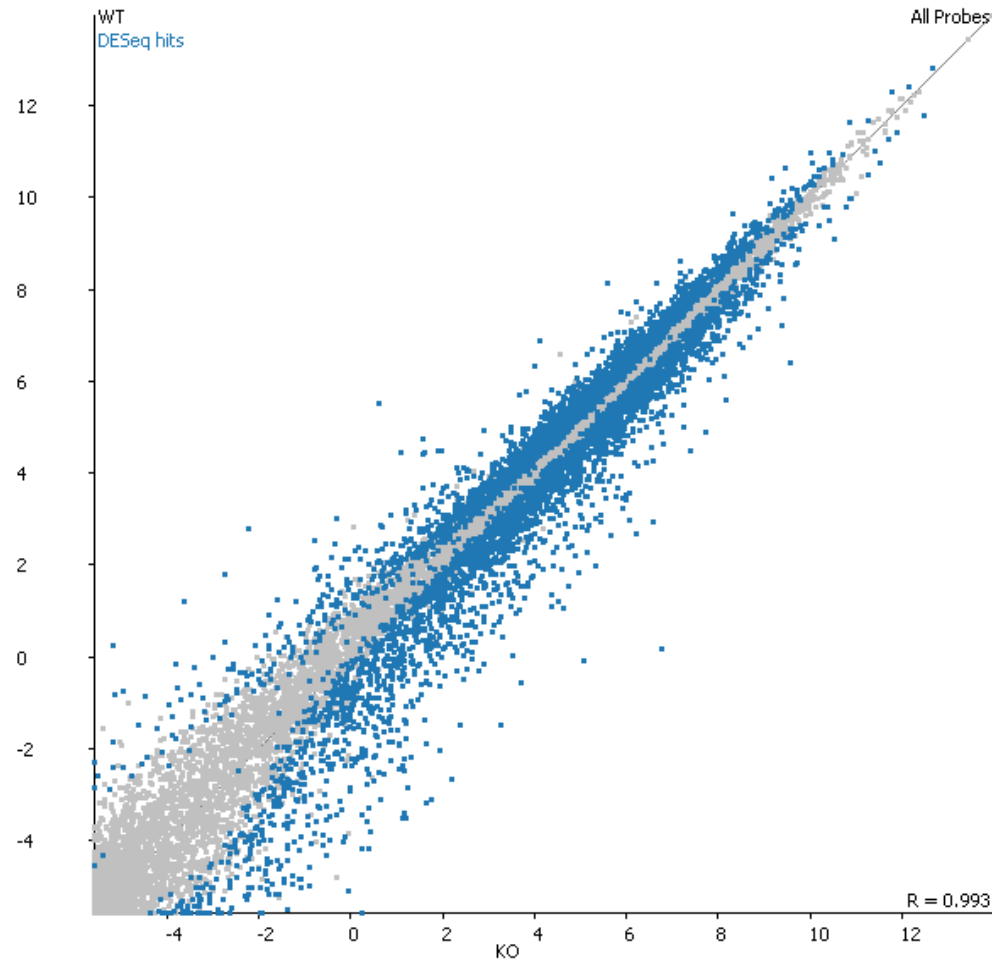


- Plot observed per gene dispersion
- Calculate average dispersion for genes with similar observation
- Individual dispersions regressed towards the mean. Weighted by
  - Distance from mean
  - Number of observations
- Points more than 2SD above the mean are not regressed

# 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`

# 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!

# Magnitude of effect filtering

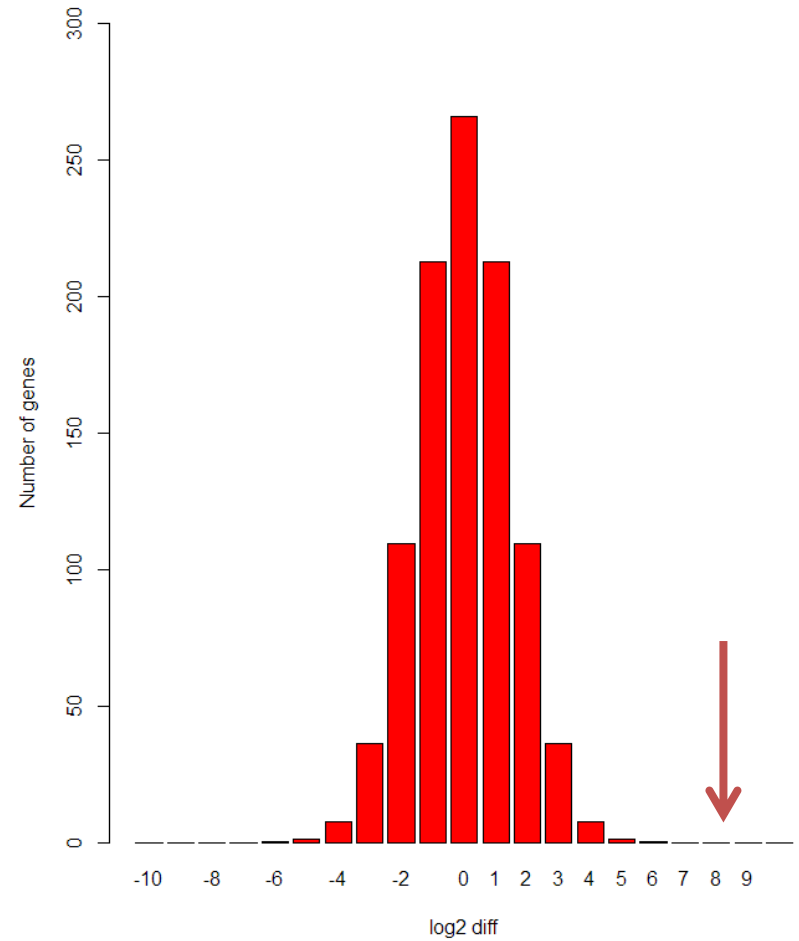
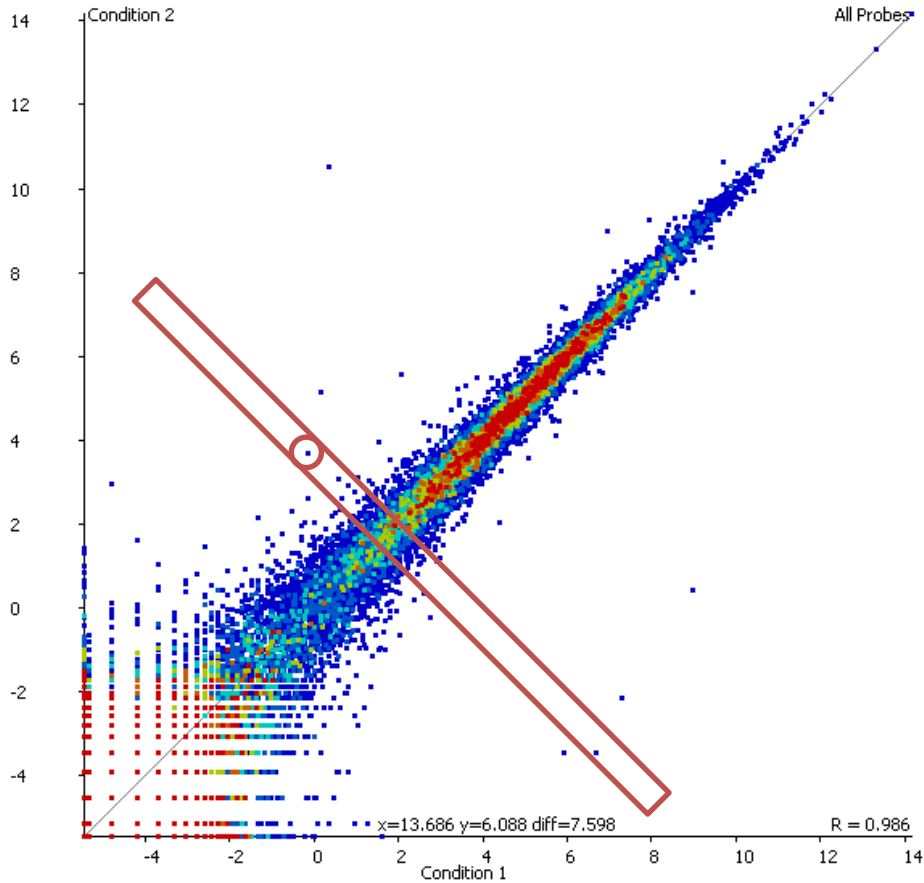
## Intensity difference test

- 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

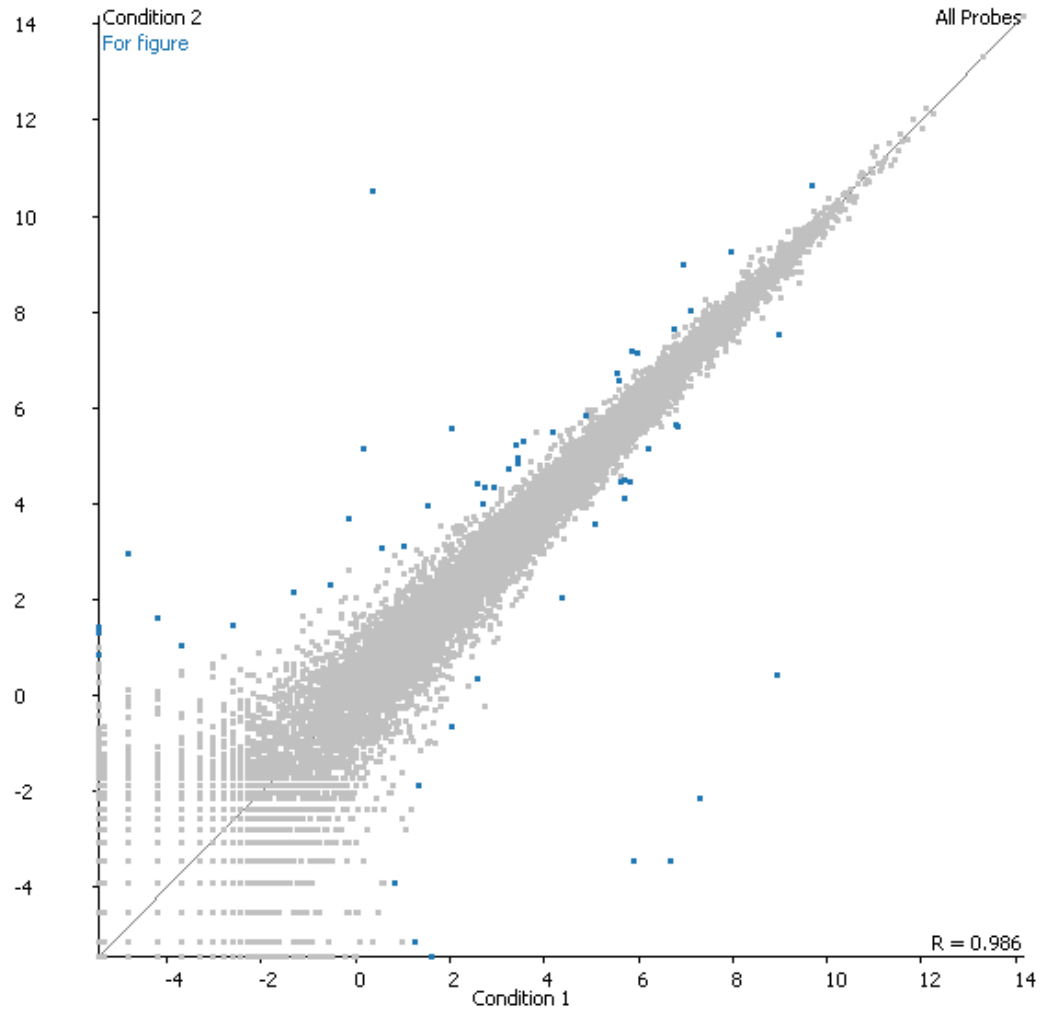
# Assumptions

- 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

# Method



# Results





# Experimental Design

# Practical Experiment Design

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

# 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

# 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

# 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

# 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

# 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

# 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.cbc.umd.edu/>
- DESeq2 <https://bioconductor.org/packages/release/bioc/html/DESeq2.html>
- Bioconductor <http://www.bioconductor.org/>
- DupRadar <http://sourceforge.net/projects/dupradar/>