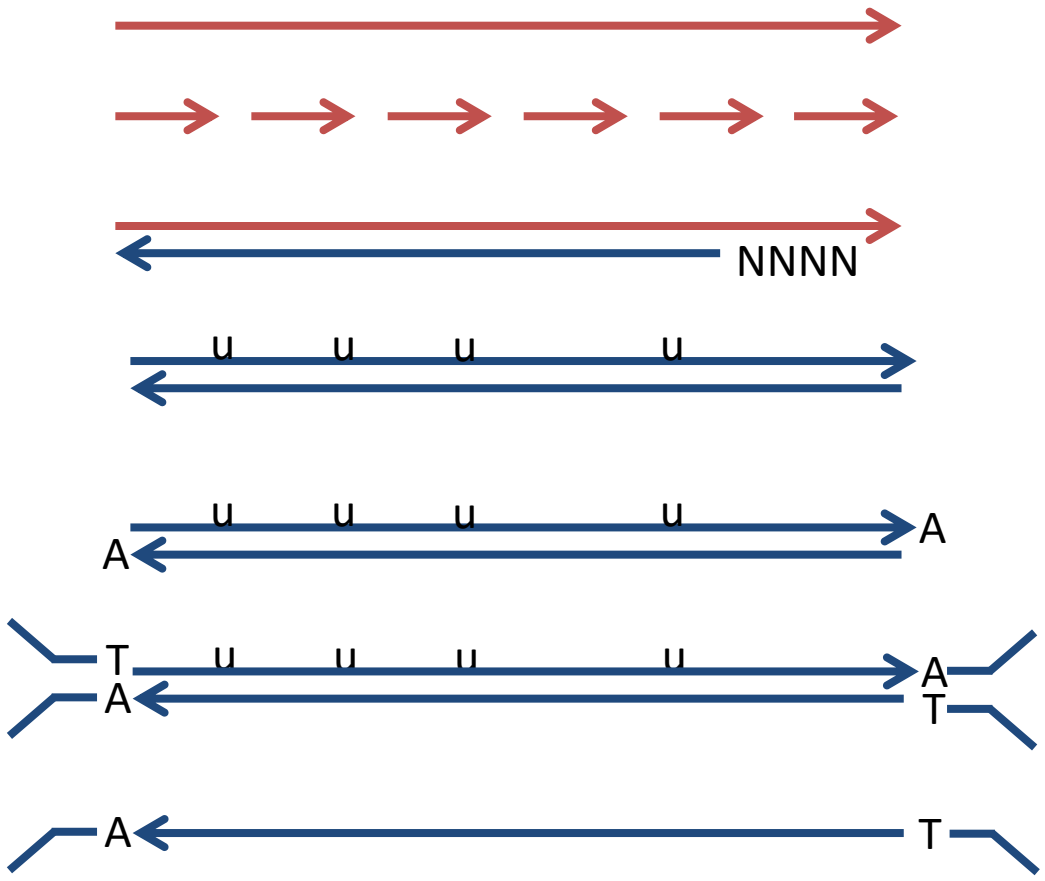
RNA-Seq Analysis

Simon Andrews, Laura Biggins, Sarah Inglesfield
simon.andrews@babraham.ac.uk
v2023-11

Babraham Bioinformatics

RNA-Seq Libraries



rRNA depleted mRNA

Fragment

Random prime + RT

2nd strand synthesis (+ U)

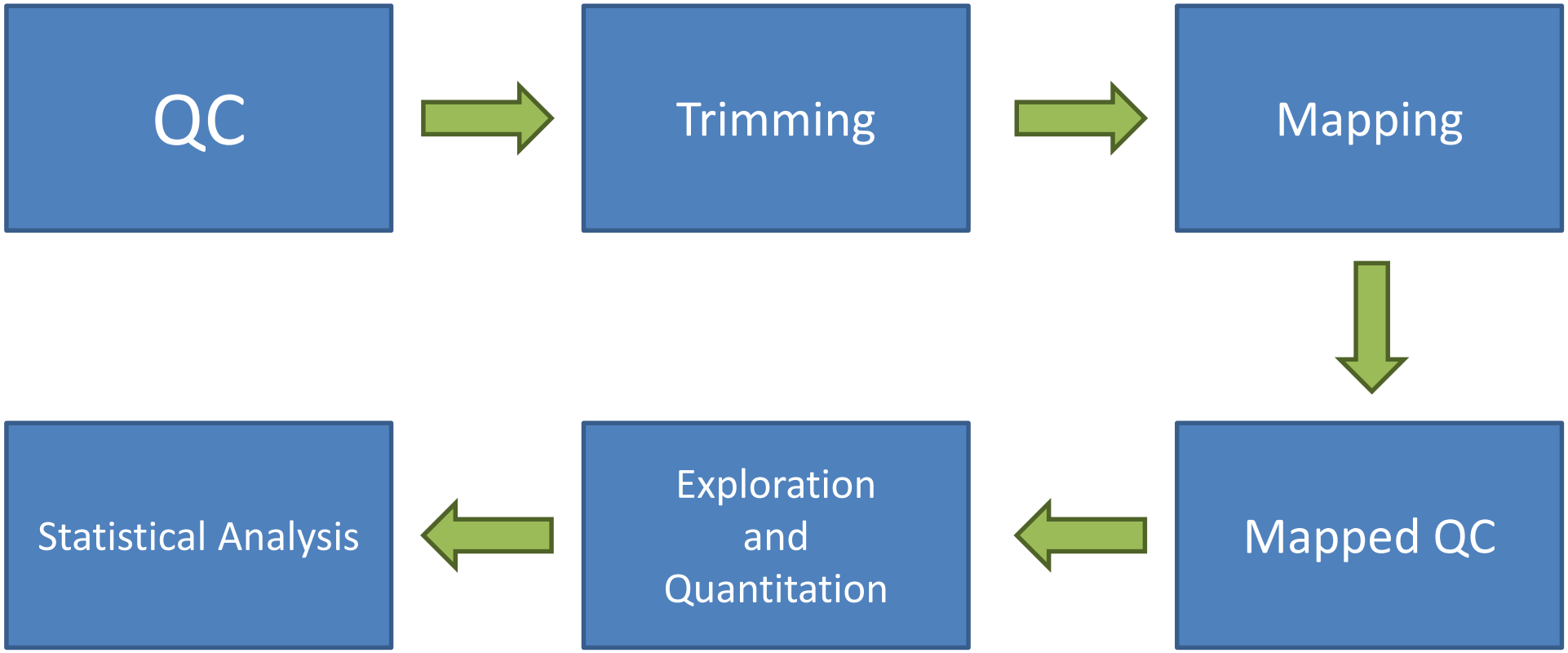
A-tailing

Adapter Ligation

(U strand degradation)

Sequencing

Reference based RNA-Seq Analysis



Sequence Data Processing

Babraham Bioinformatics

Raw Sequence Quality Control

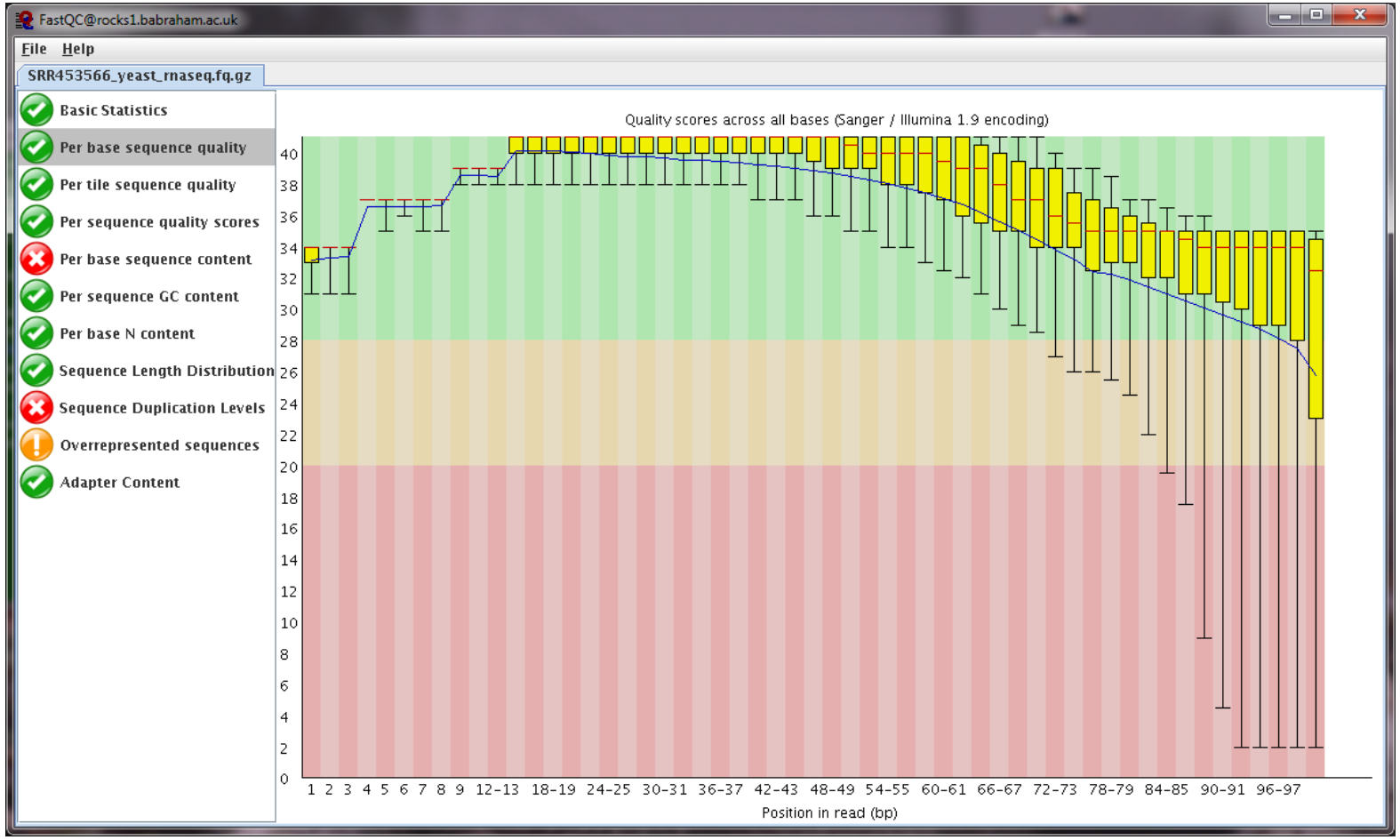
Babraham Bioinformatics

FastQ Format Data

```
@HWUSI-EAS611:34:6669YAAXX:1:1:5069:1159 1:N:0:
TCGATAATACCGTTTTTTTCCGTTTGATGTTGATACCATT
+
IIHIIHIIIIIIIIIIIIIIIIIIIIIIIIIHIIIIHIIIII
@HWUSI-EAS611:34:6669YAAXX:1:1:5243:1158 1:N:0:
TATCTGTAGATTTCACAGACTCAAATGTAAATATGCAGAG
+
DF=DBD<BBFGGGGGGGBD@GGGD4@CA3CGG>DDD:D,B
@HWUSI-EAS611:34:6669YAAXX:1:1:5266:1162 1:N:0:
GGAGGAAGTATCACTTCCTTGCCTGCCTCCTCTGGGGCCT
+
:GBGGGGGGGGGDGGDEDGGDGGGGDHHDHGHHGBGG:GG
```

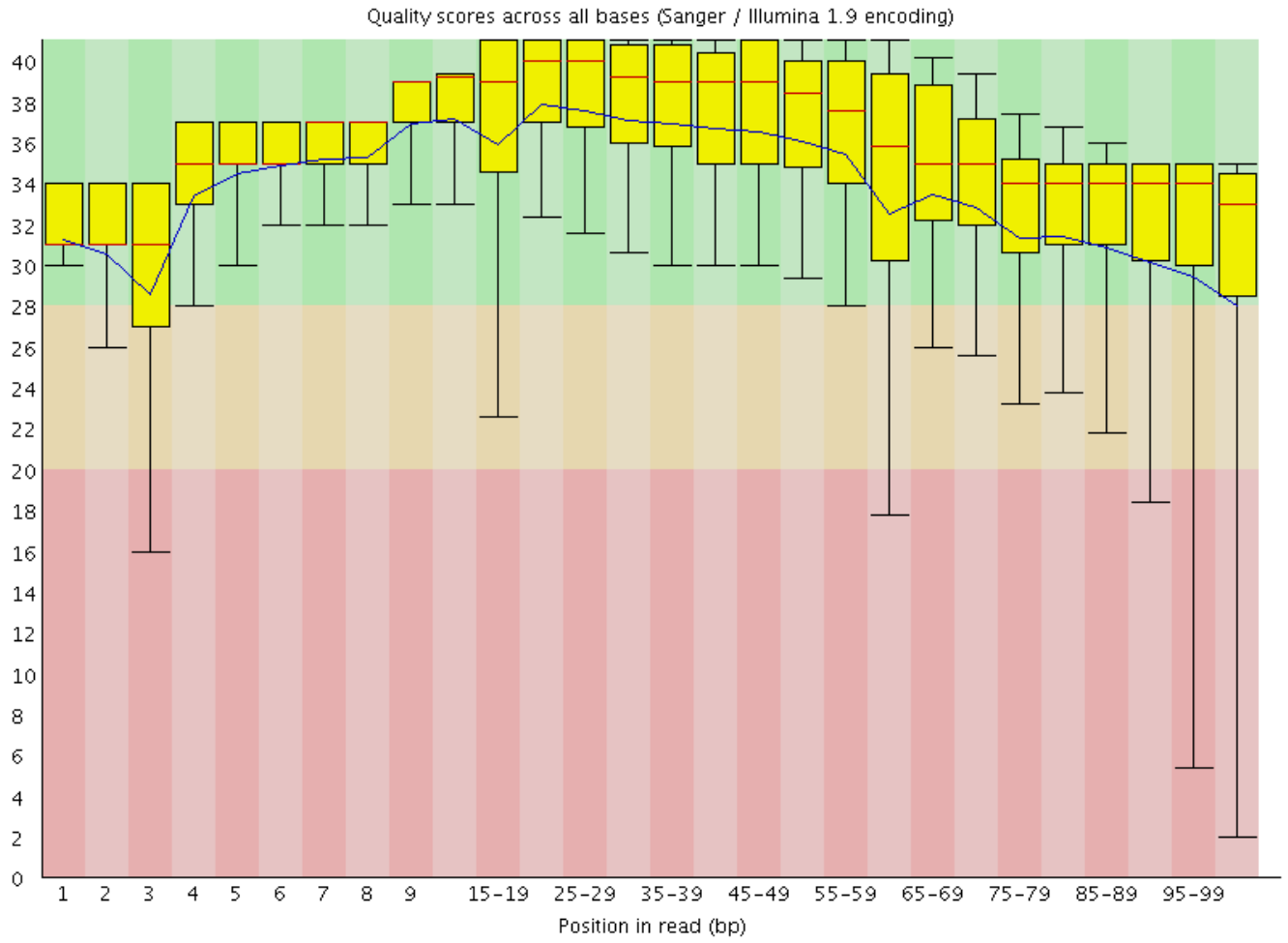
FastQC

- Base call quality
- Composition
- Duplication
- Contamination



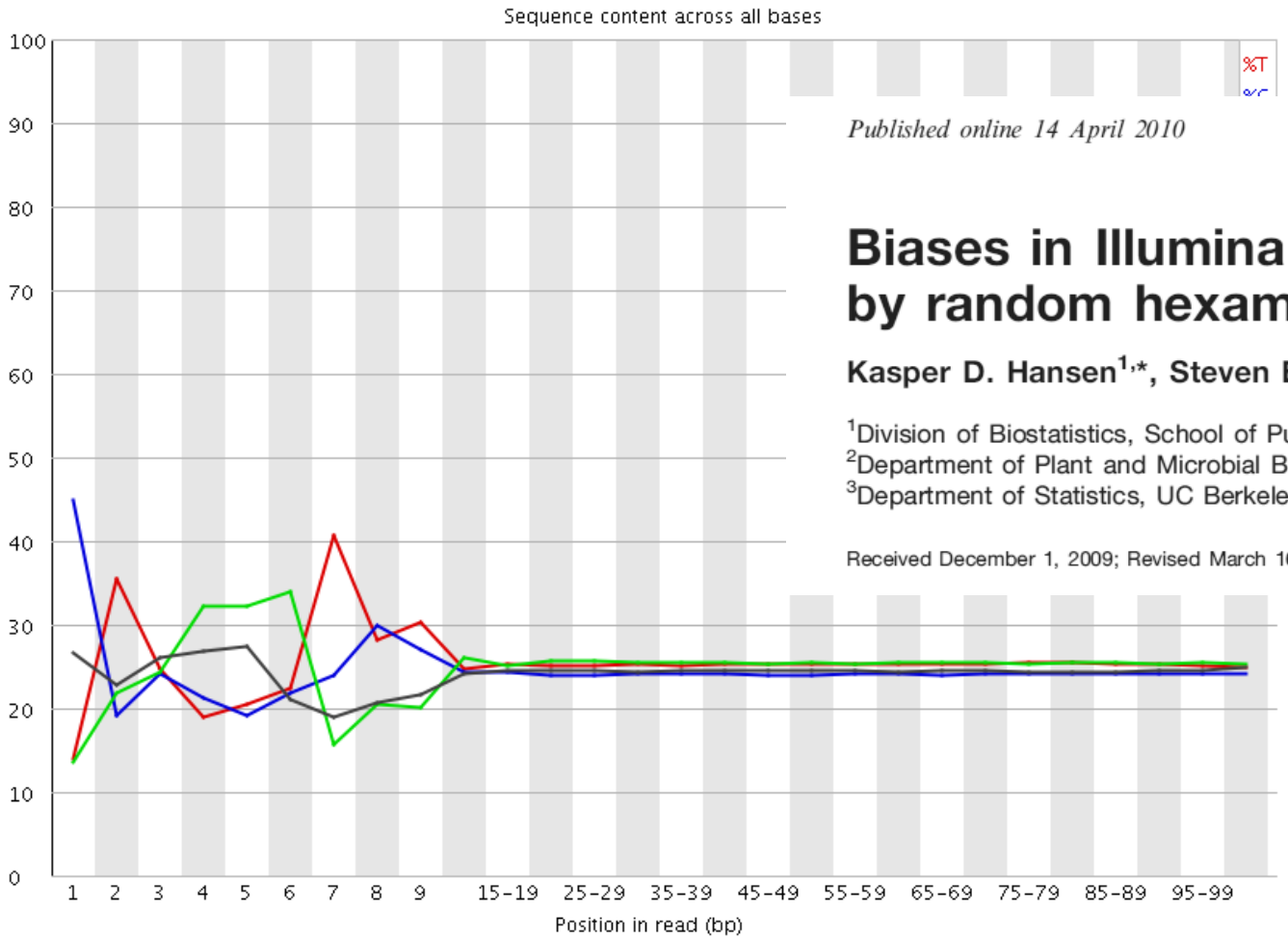
QC: Base Call Quality

Call Quality
(Phred score)



Read Position

QC: Composition



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]

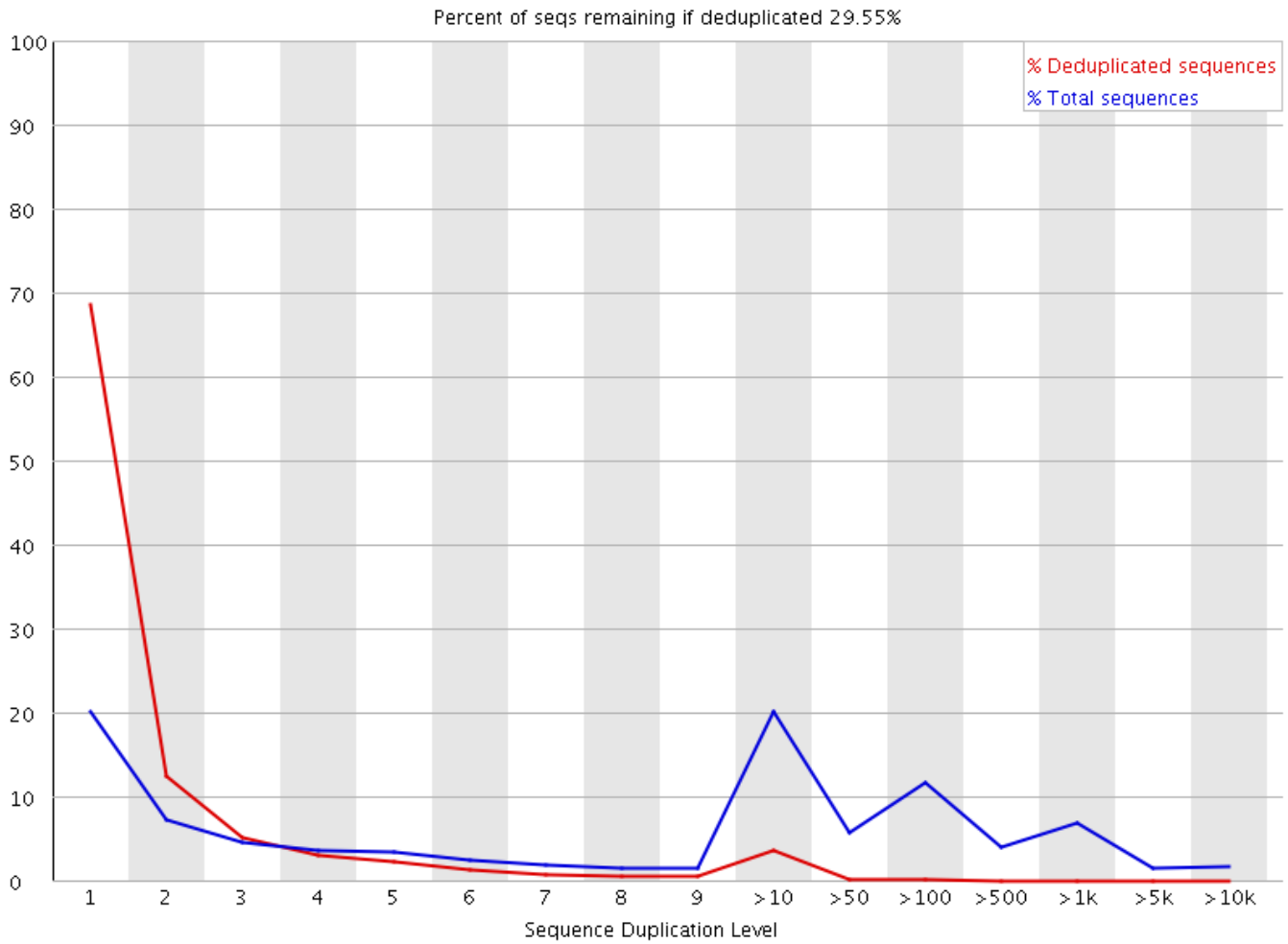
[1]Division of Biostatistics, School of Public Health, UC Berkeley, 101 Haviland Hall, Berkeley, CA 94720-7358, [2]Department of Plant and Microbial Biology, UC Berkeley, 461 Koshland Hall, Berkeley, CA 94720-3102 and [3]Department of Statistics, UC Berkeley, 367 Evans Hall, Berkeley, CA 94720-3860,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

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

Read Position

QC: Duplication (blue trace)

Percentage of library

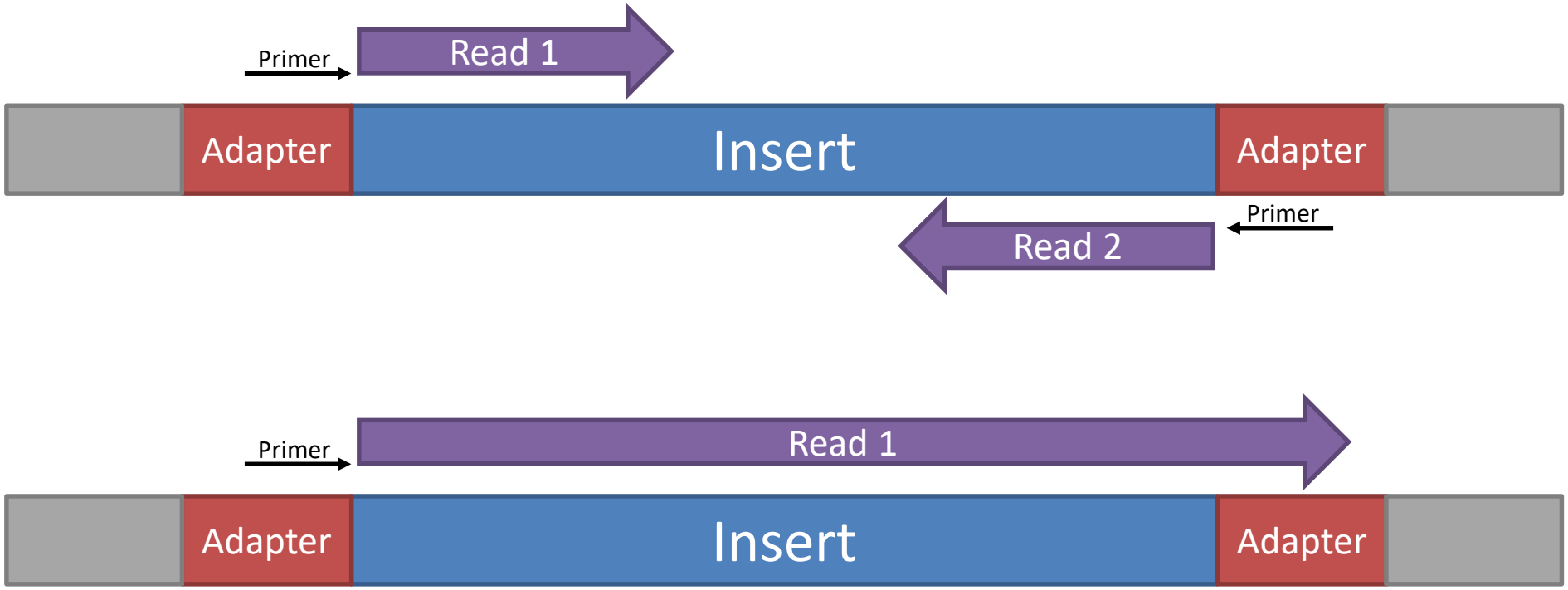


Level of duplication

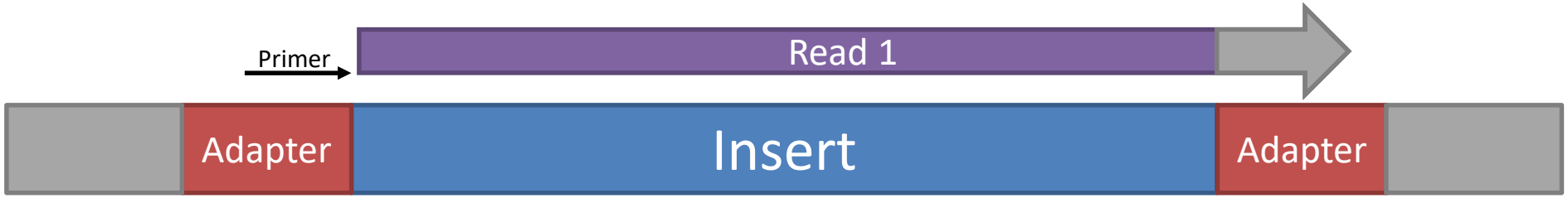
Adapters and Trimming

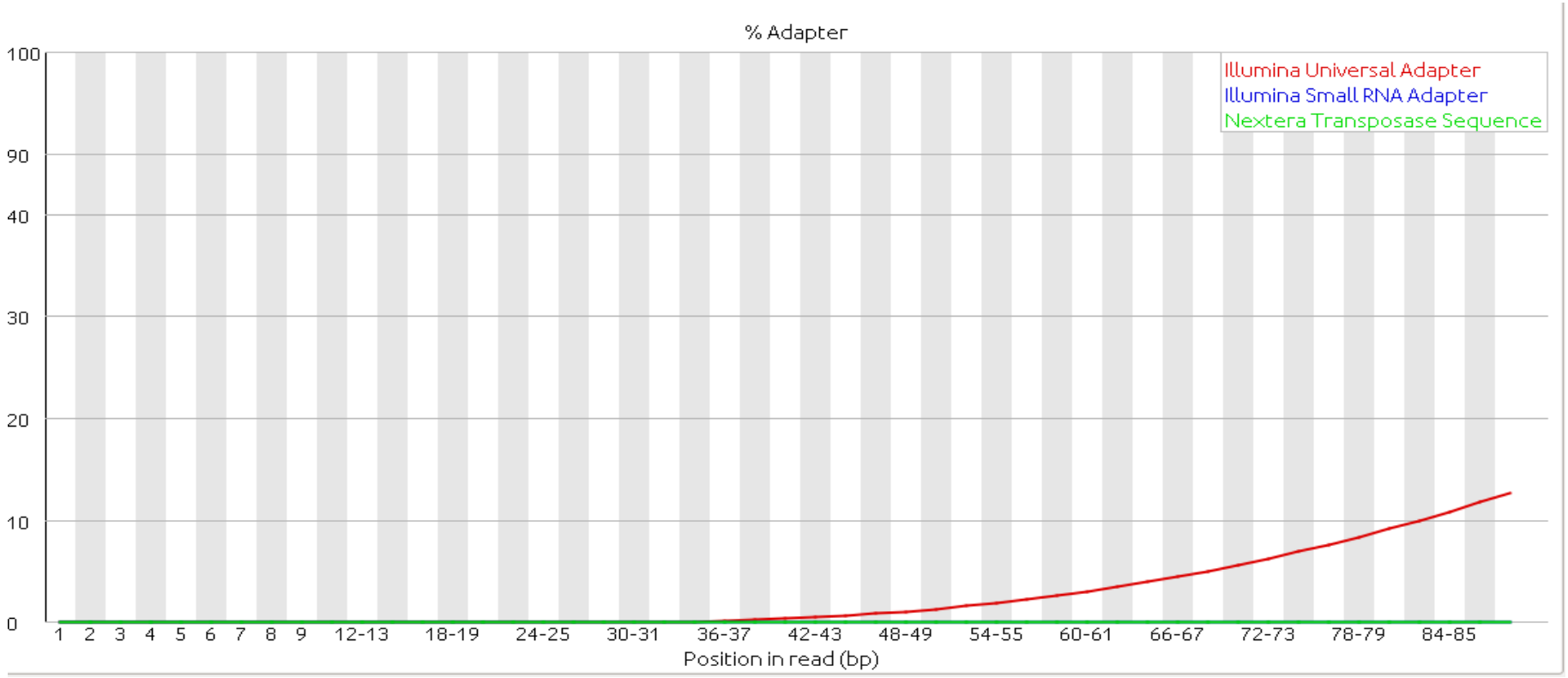
Babraham Bioinformatics

Library Structure

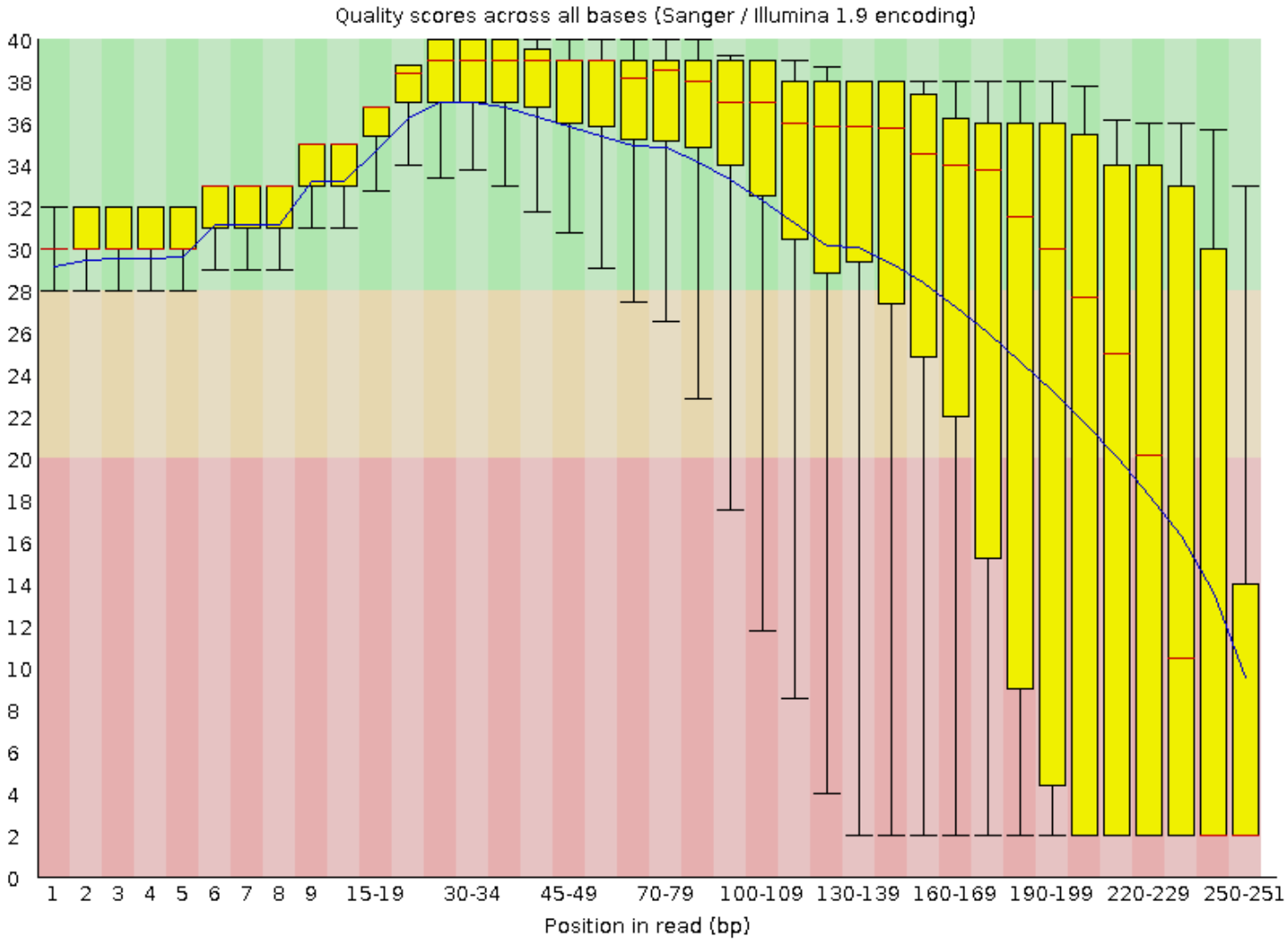


Trimming Adapters





Trimming Quality

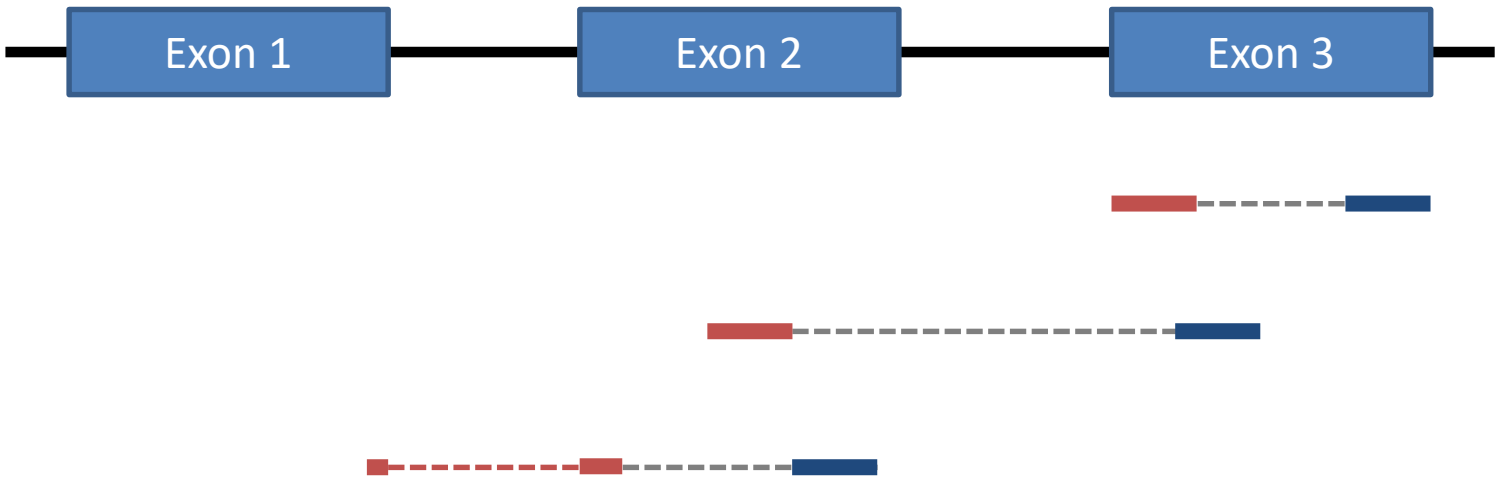


Poor quality data tends to be at the 3’ end

Mapping to a reference

Babraham Bioinformatics

Mapping



Genome

Simple mapping within exons

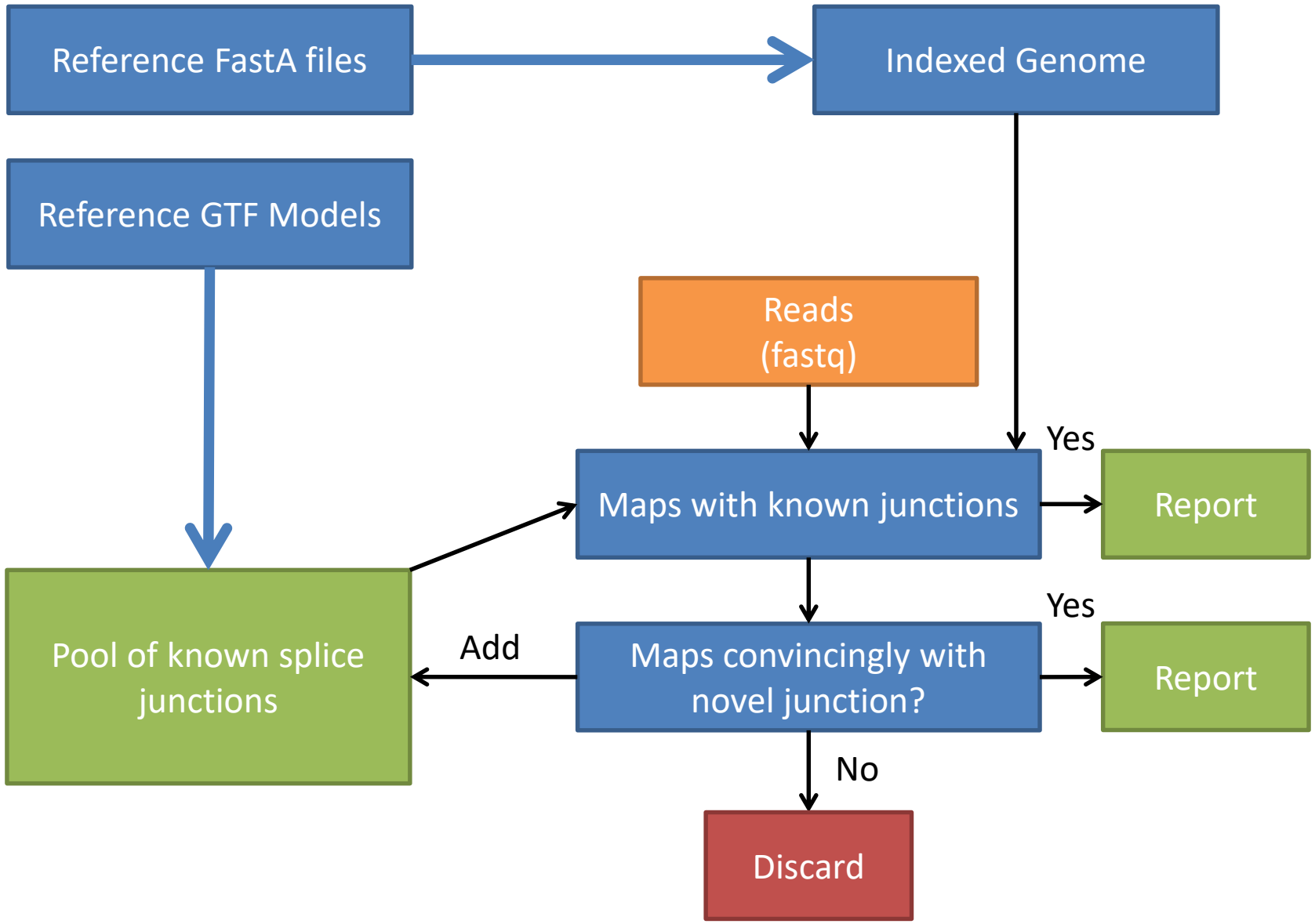
Mapping between exons

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



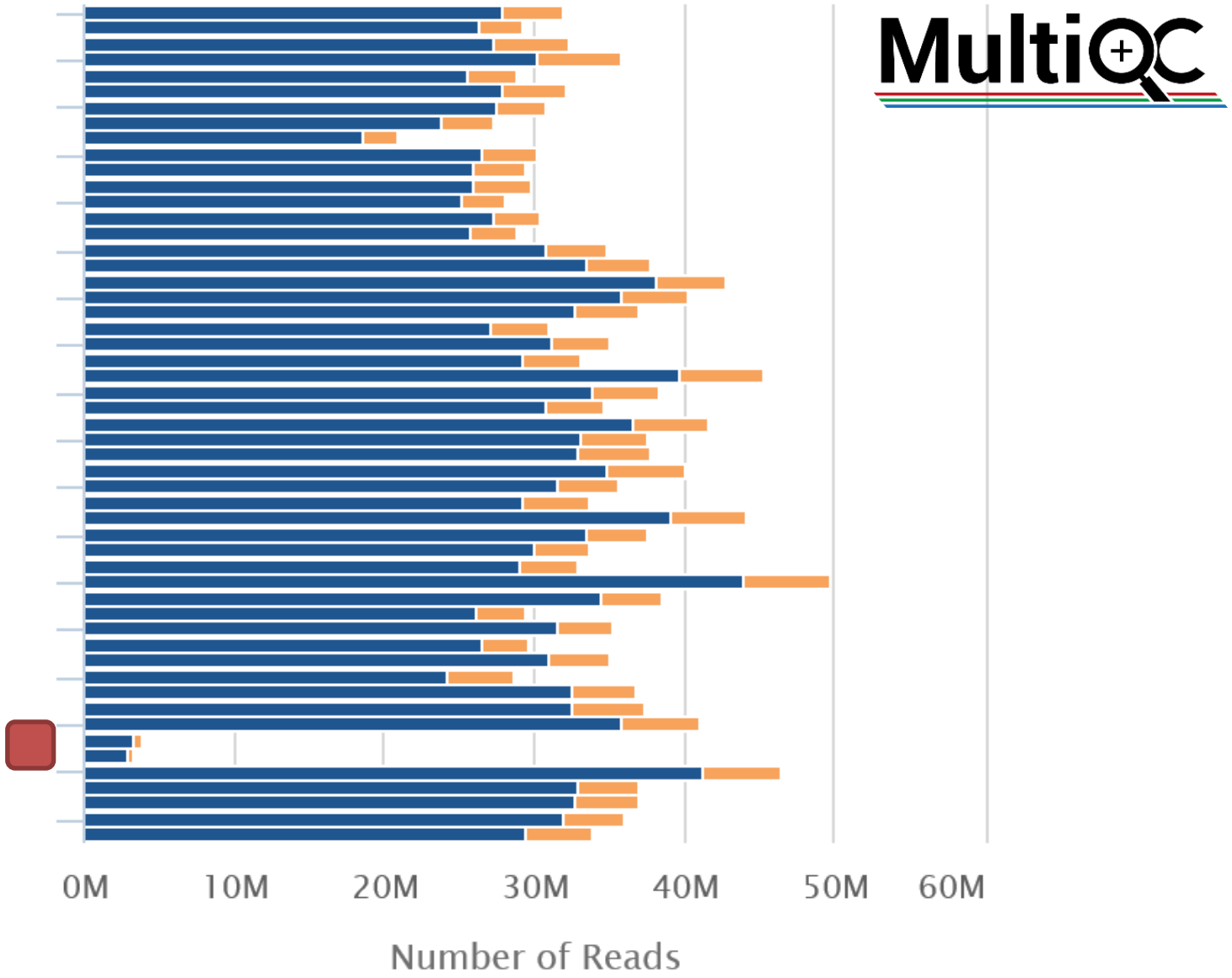
Mapped Data QC

Babraham Bioinformatics

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



Exercise: RNA-Seq QC and Data Processing

Babraham Bioinformatics

Running programs in Linux

- Open a shell (text based OS interface)

- Type the name of the program you want to run
  - Add on any options the program needs
  - Press return - the program will run
  - When the program ends control will return to the shell

- Run the next program!

Running programs

```
user@server:~$ ls
Desktop  Documents  Downloads  examples.desktop
Music  Pictures  Public  Templates  Videos

user@server:~$
```

- Command prompt - you can't enter a command unless you can see this
- The command we're going to run (`ls` in this case, to list files)
- The output of the command - just text in this case

The structure of a unix command

```
ls  -ltd --reverse Downloads/  Desktop/  Documents/
```

Program name

Switches

Data
(normally files)

Each option or section is separated by spaces. Options or files with spaces in must be put in quotes.

Command line switches

- Change the behaviour of the program
- Come in two flavours (each option often has both types available)
  - Minus plus single letter (eg `-x -c -z`)
    - Can be combined (eg `-xcz`)

  - Two minuses plus a word (eg `--extract --gzip`)
    - Can't be combined

- Some take an additional value

  `-f somfile.txt` (specify a filename)

  `--width=30` (specify a value)

Specifying file paths

- Specify names from whichever directory you are currently in
  - If I'm in `/home/simon`
  - `Data/big_data.fq.gz`
    - is the same as `/home/simon/Data/big_data.fq.gz`

- Move to the directory with the data and just use file names
  - `cd Data`
  - `big_data.fq.gz`

Command line completion

- Most errors in commands are typing errors in either program names or file paths
- Shells (ie BASH) can help with this by offering to complete path names for you
- Command line completion is achieved by typing a partial path and then pressing the TAB key (to the left of Q)

Command line completion

List of files / folders:

```
Desktop
Documents
Downloads
Music
Public
Published
Templates
Videos
```

`T` [TAB] → `Templates`

`P` [TAB] → `Publ`

`Do` [TAB] → [beep]

`Do` [TAB] [TAB] → `Documents Downloads`

`Doc` [TAB] → `Documents`

You should ALWAYS use TAB completion to fill in paths for locations which exist so you can't make typing mistakes

(it obviously won't work for output files though)

Debugging Tips

- If anything (except the splice site extraction) completes almost immediately then it didn't work!

- Look for errors before asking for help. They will either be
  - The last piece of text before the program exited
  - The first piece of text produced after it started (followed by the help file)

- To see if a program is running go to another shell and look at the last file produced to see if it's growing

- Programs which are stuck can be cancelled with Control+C

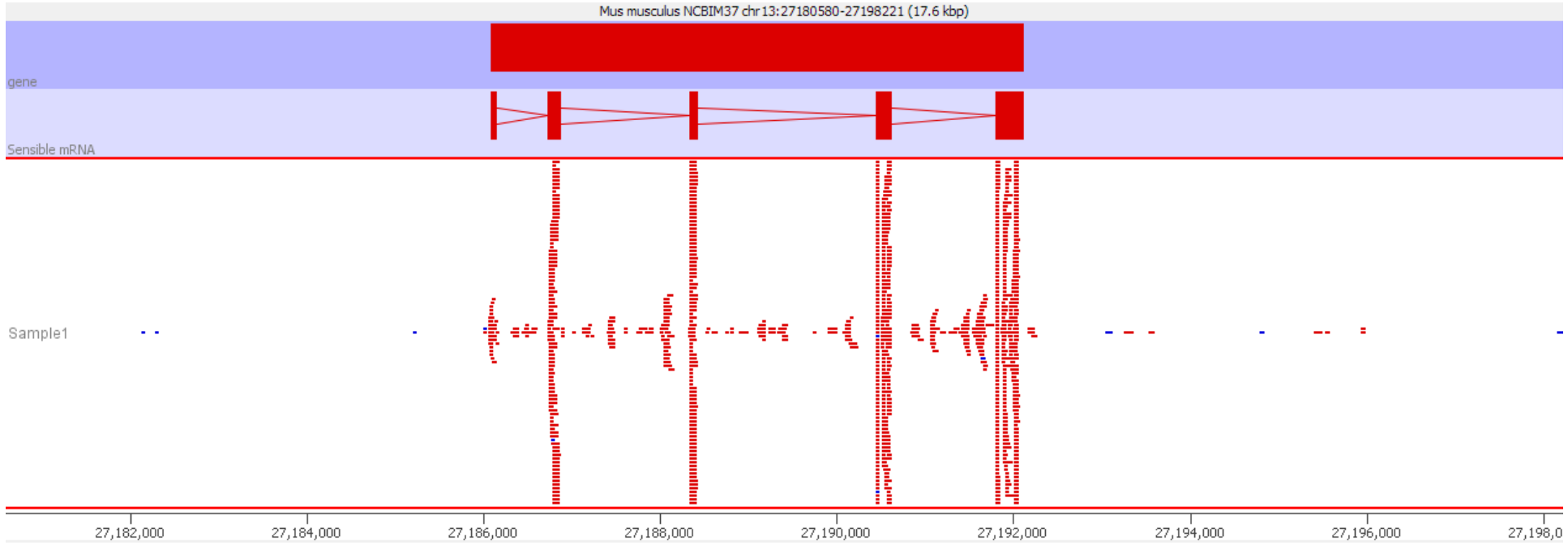
Some useful commands

`cd mydir`	Change directory to `mydir`
`ls -ltrh`	List files in the current directory, show details and put the newest files at the bottom
`less x.txt`	View the `x.txt` text file Return = down one line Space = down one page q = quit

Data Visualisation and Exploration

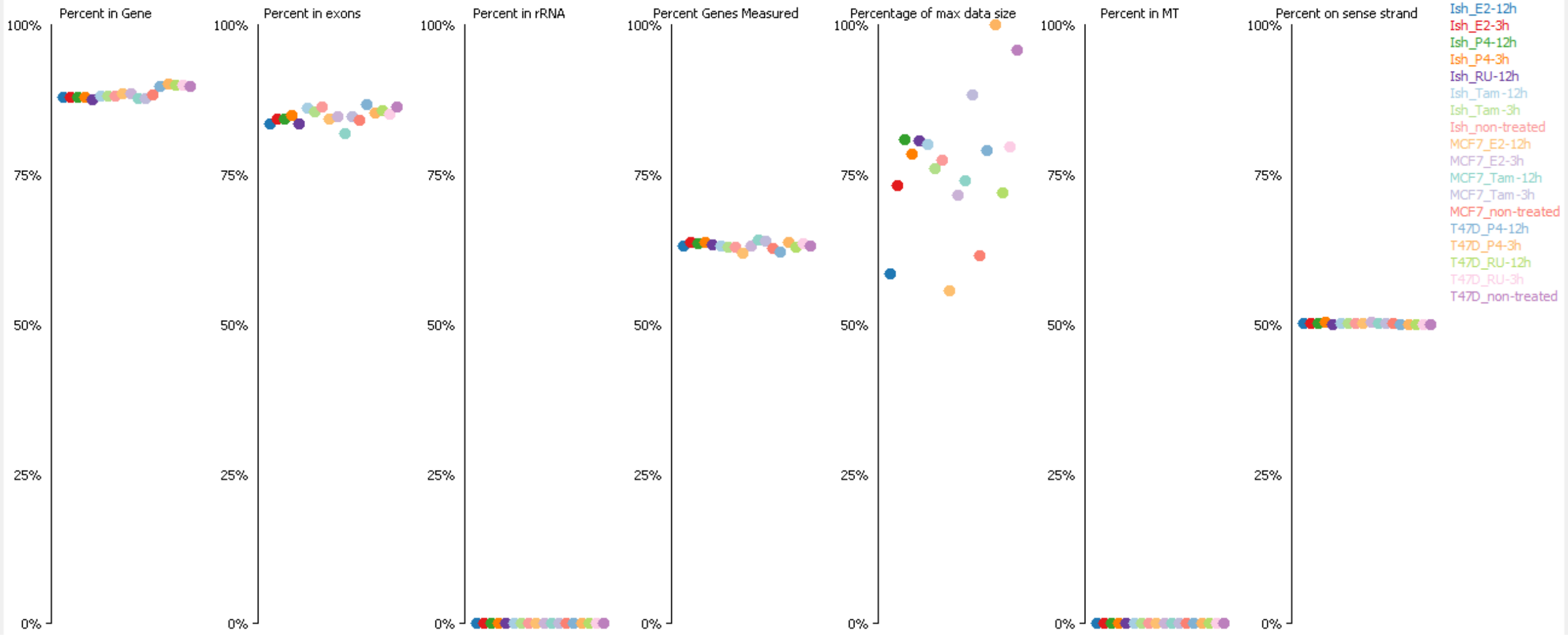
Babraham Bioinformatics

Viewing Mapped Data

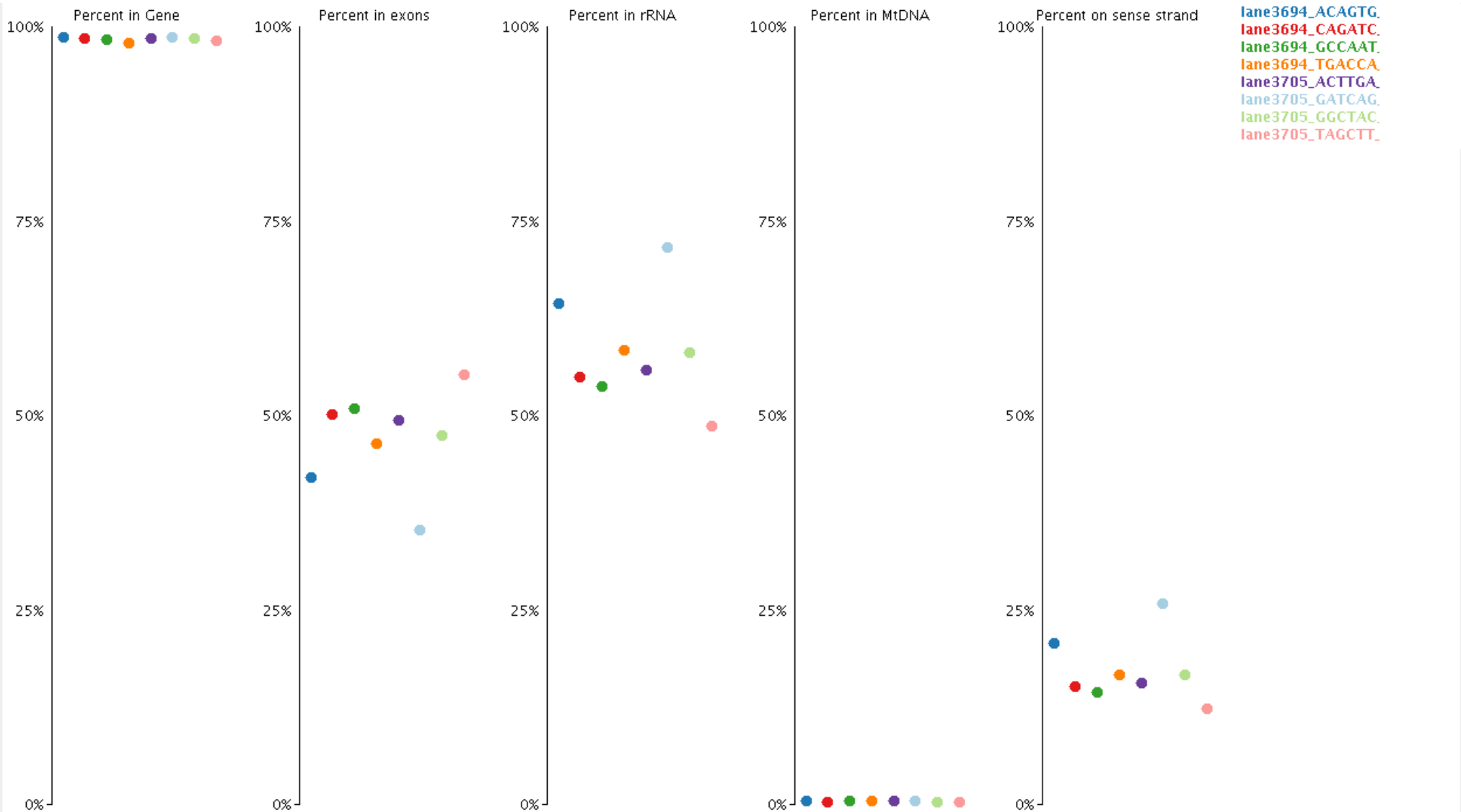


- Reads over exons
- Reads over introns
- Reads in intergenic regions
- Strand specificity

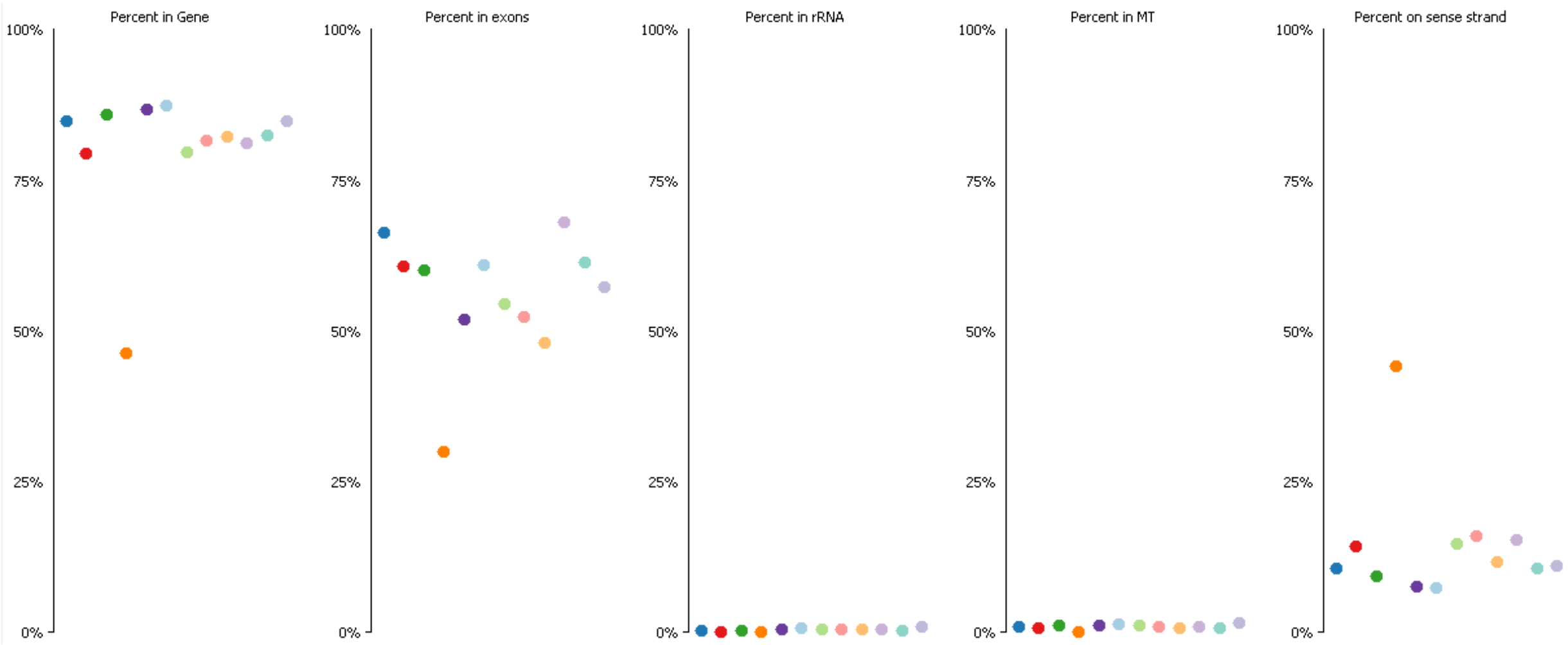
SeqMonk RNA-Seq QC (good)



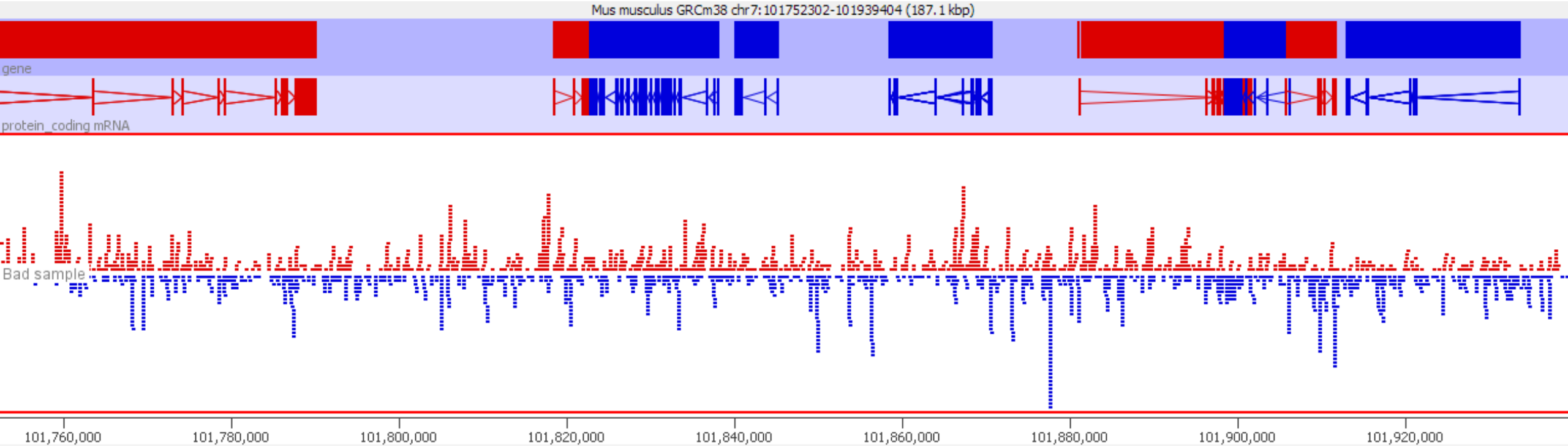
SeqMonk RNA-Seq QC (bad)



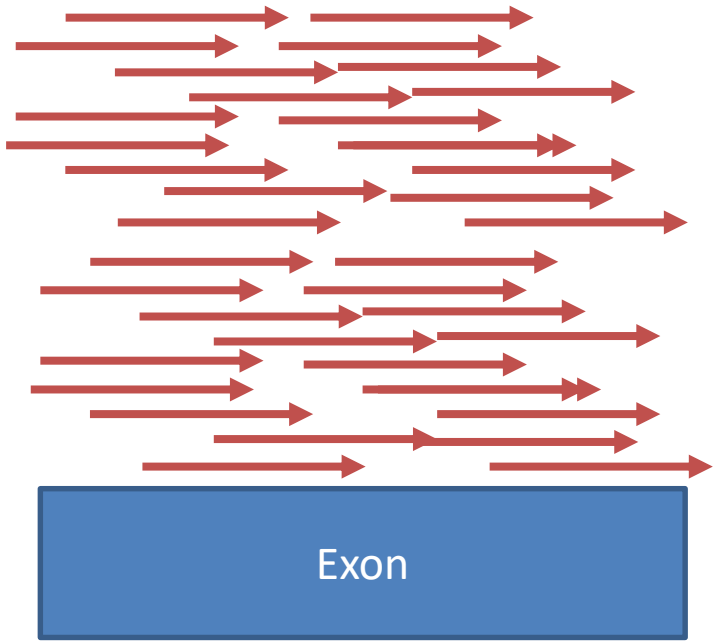
SeqMonk RNA-Seq QC (bad)

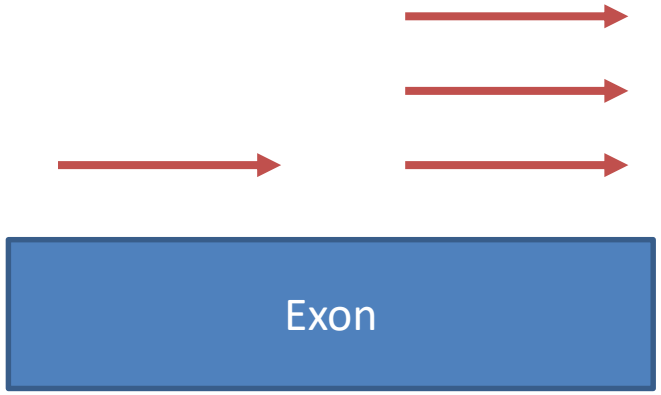


Look at poor QC samples

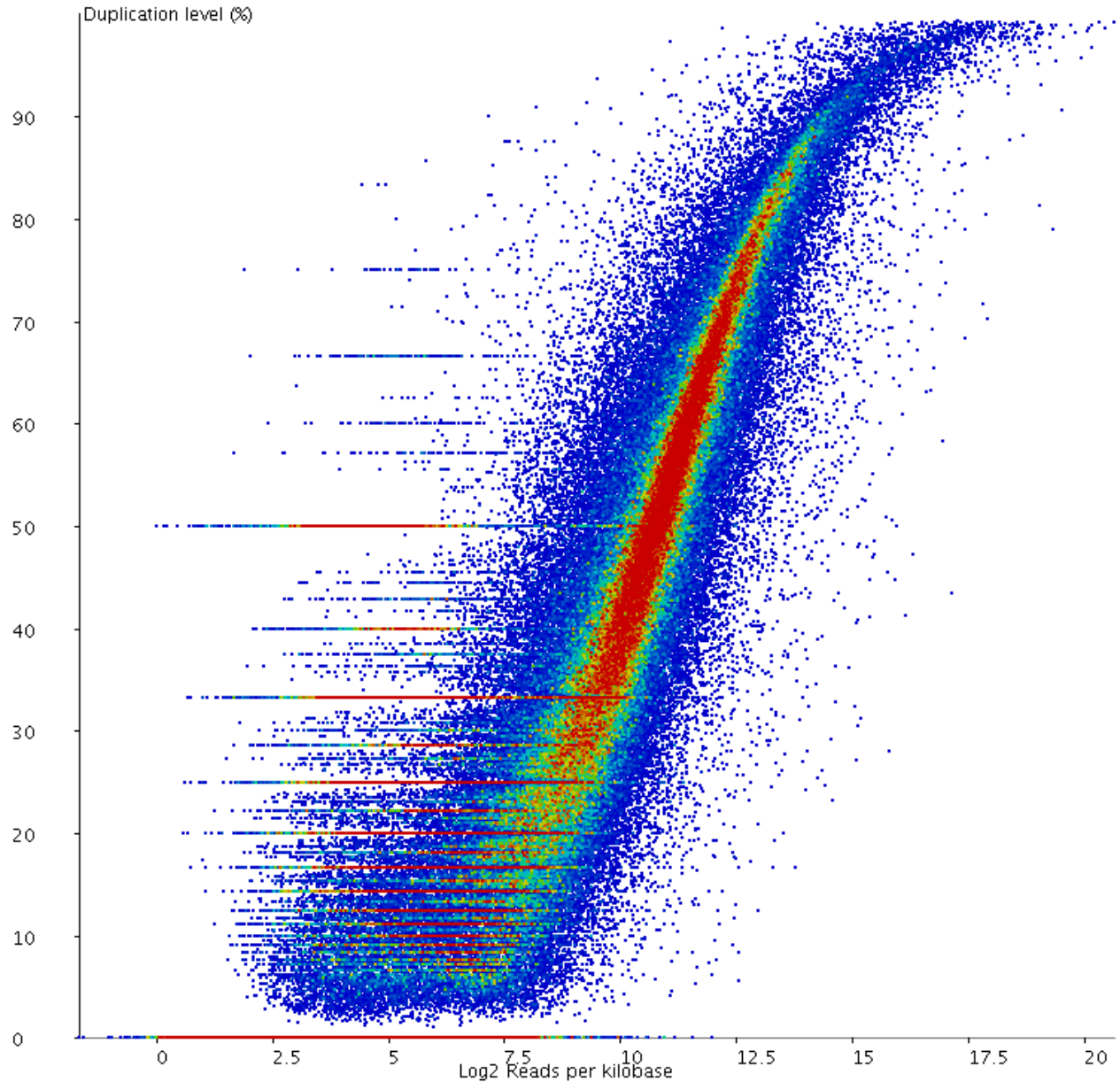


Duplication (again)

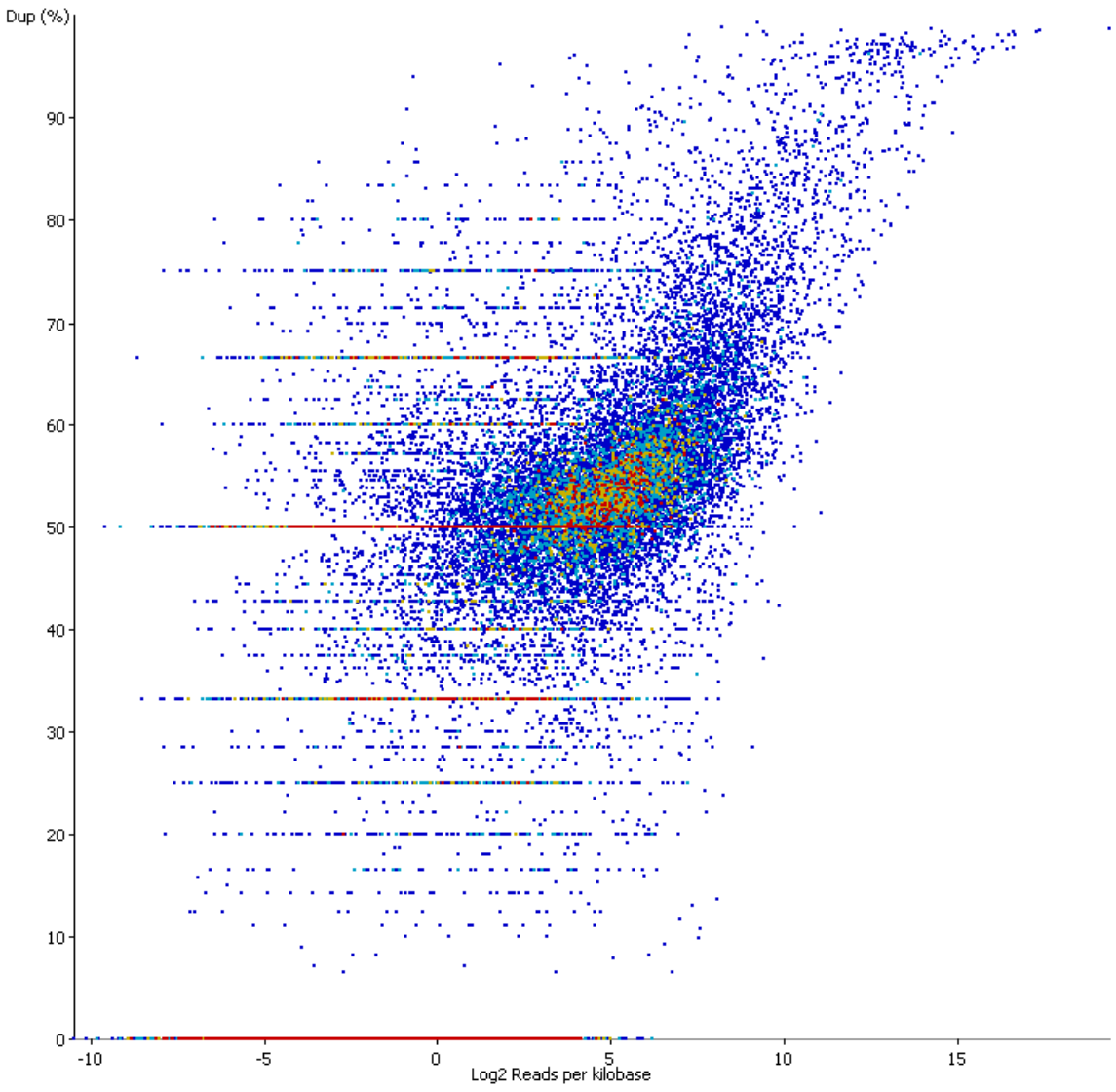




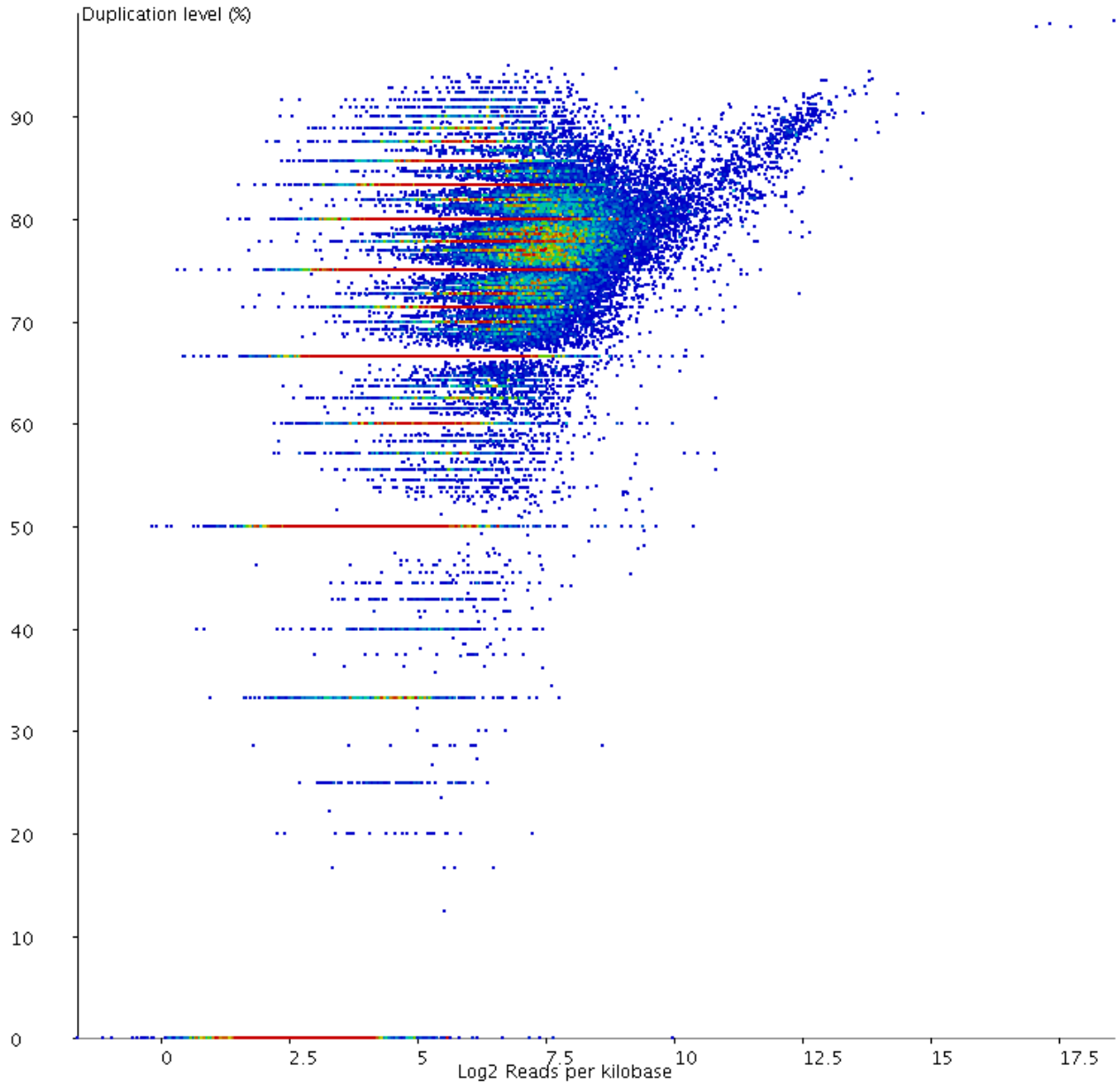
Duplication (good)



Duplication (moderate)



Duplication (bad)



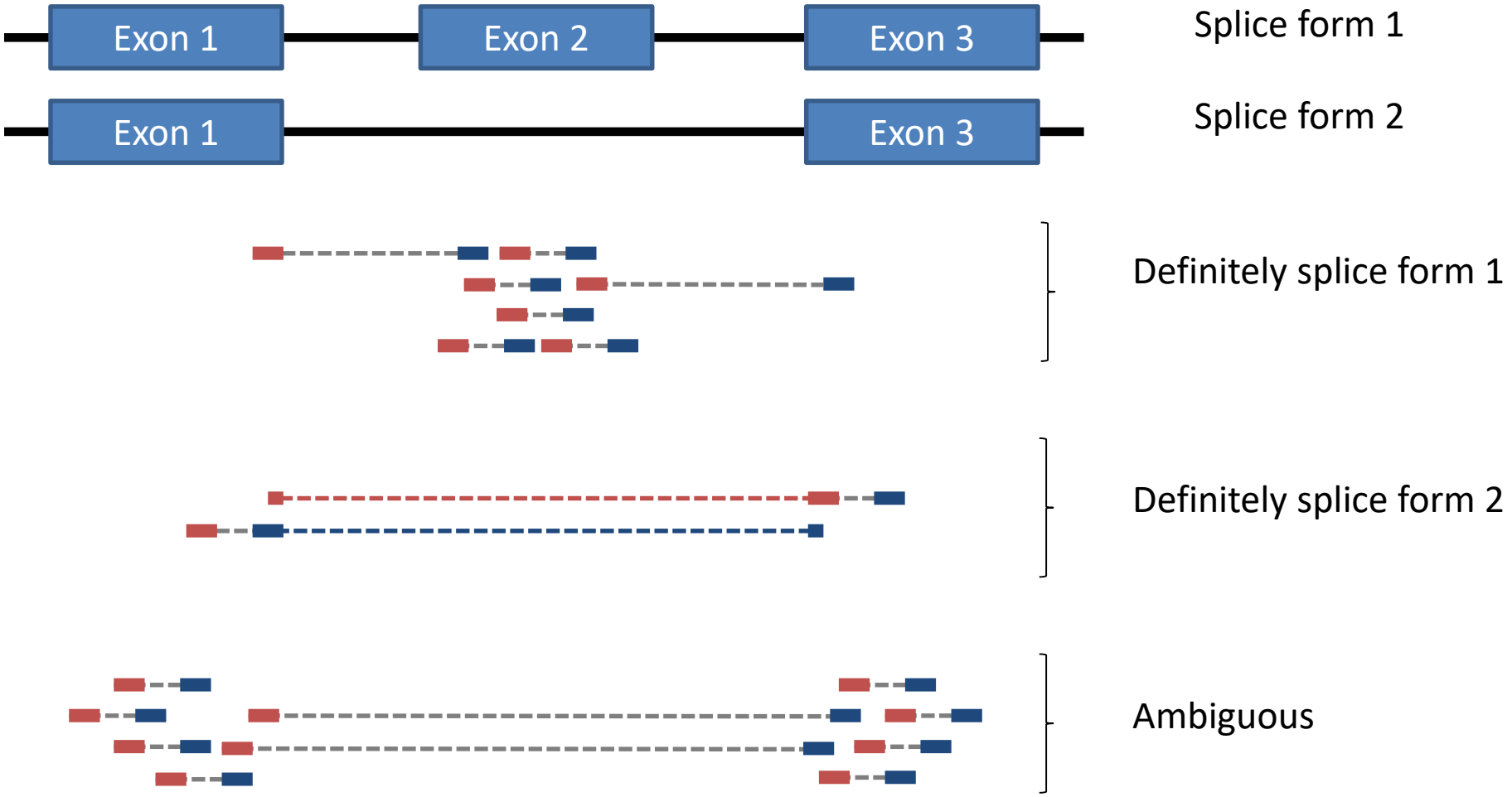
Fixing Duplication?

- If duplication is biased (some genes more than others)
  - Can't be 'fixed' – can still analyse but be cautious

- If it's unbiased (everything is duplicated)
  - Doesn't affect quantitation
  - Will affect statistics

  - Can estimate global level and correct raw counts

Quantitation



Simple Quantitation - Forget splicing

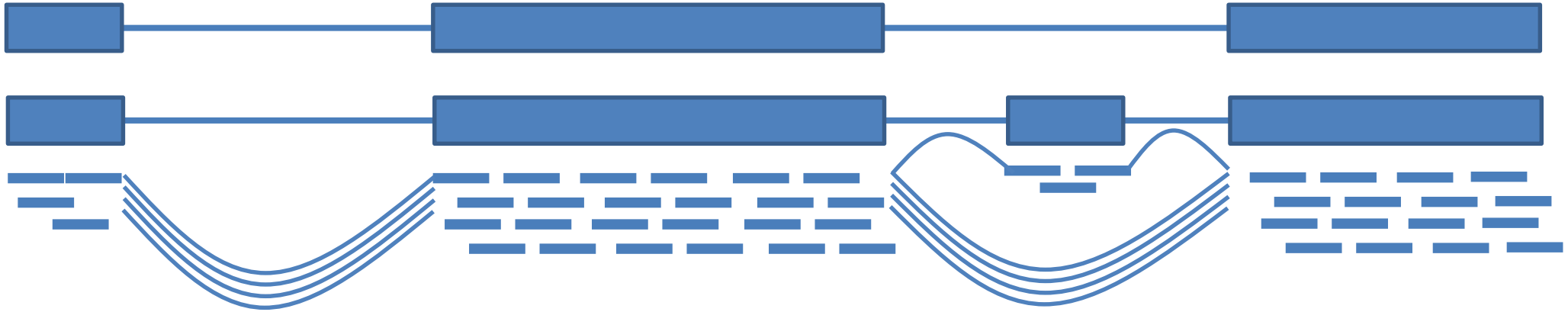
- Count read overlaps with exons of each gene
  - Consider library directionality
  - Simple
  - Gene level quantitation
  - Many programs
    - Seqmonk (graphical)
    - Feature Counts (subread)
    - BEDTools
    - HTSeq

Analysing Splicing

Systematic evaluation of differential splicing tools for RNA-seq studies

Arfa Mehmood, Asta Laiho, Mikko S Venäläinen, Aidan J McGlinchey, Ning Wang, Laura L Elo

Briefings in Bioinformatics, Volume 21, Issue 6, November 2020, Pages 2052–2065,



- Try to quantitate transcripts (cufflinks, RSEM, bitSeq)
- Quantitate exons and compare to gene (EdgeR, DEXSeq)
- Quantitate splicing events (rMATS, MAJIQ)

Normalisation: RPKM / FPKM / TPM

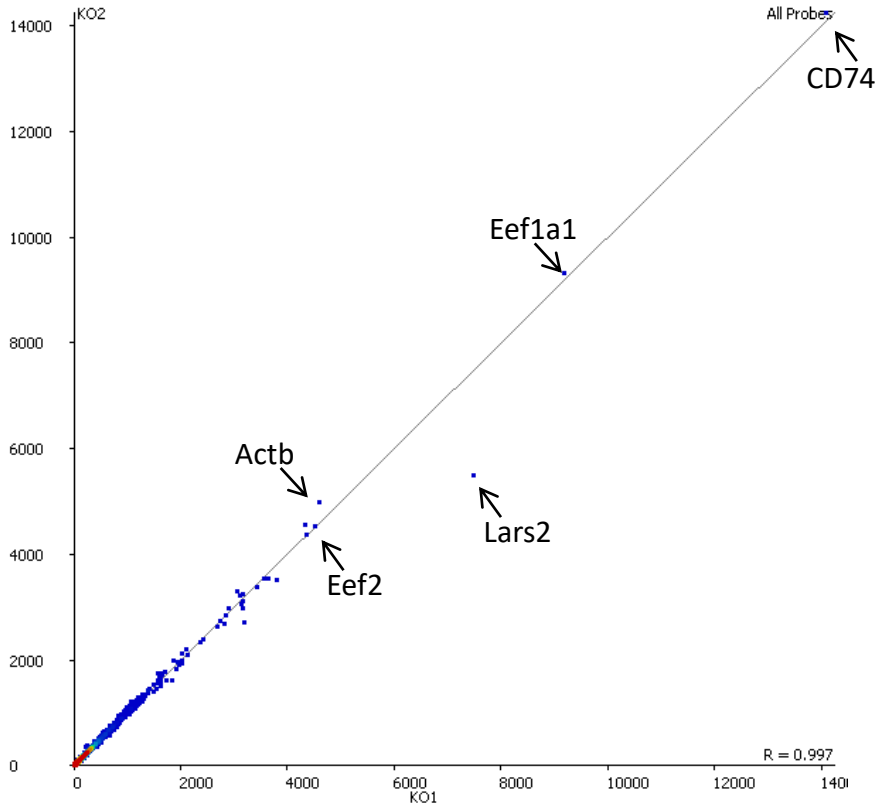
- RPKM (Reads per kilobase of transcript per million reads of library)
  - Corrects for total library coverage
  - Corrects for gene length
  - Comparable between different genes within the same dataset

- 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

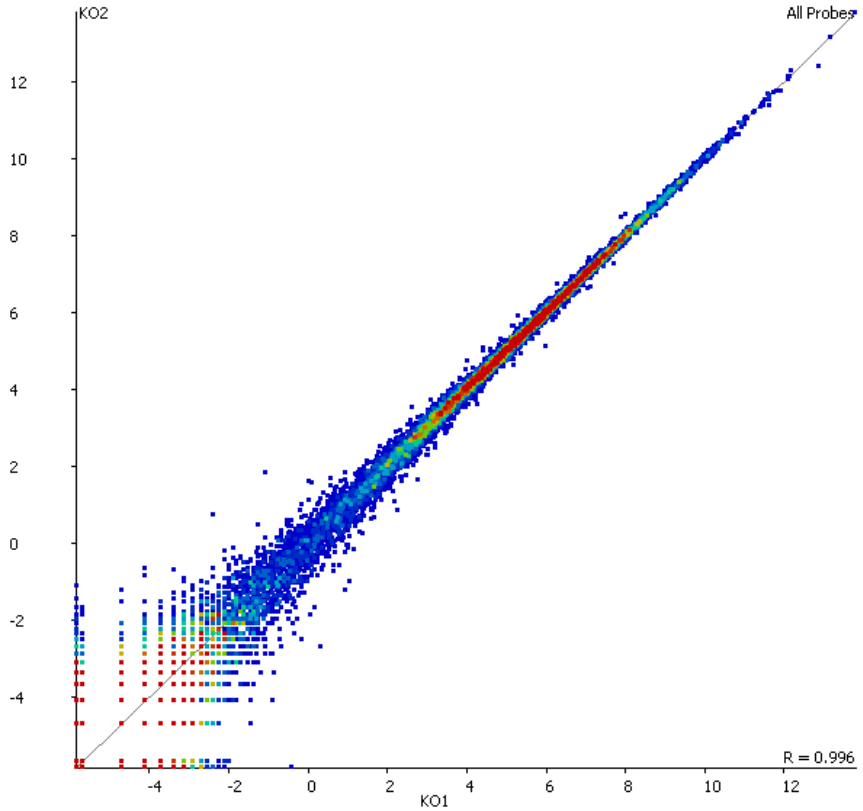
- TPM (transcripts per million)
  - Normalises to transcript copies instead of reads
  - Corrects for cases where the average transcript length differs between samples

Visualising Expression and Normalisation

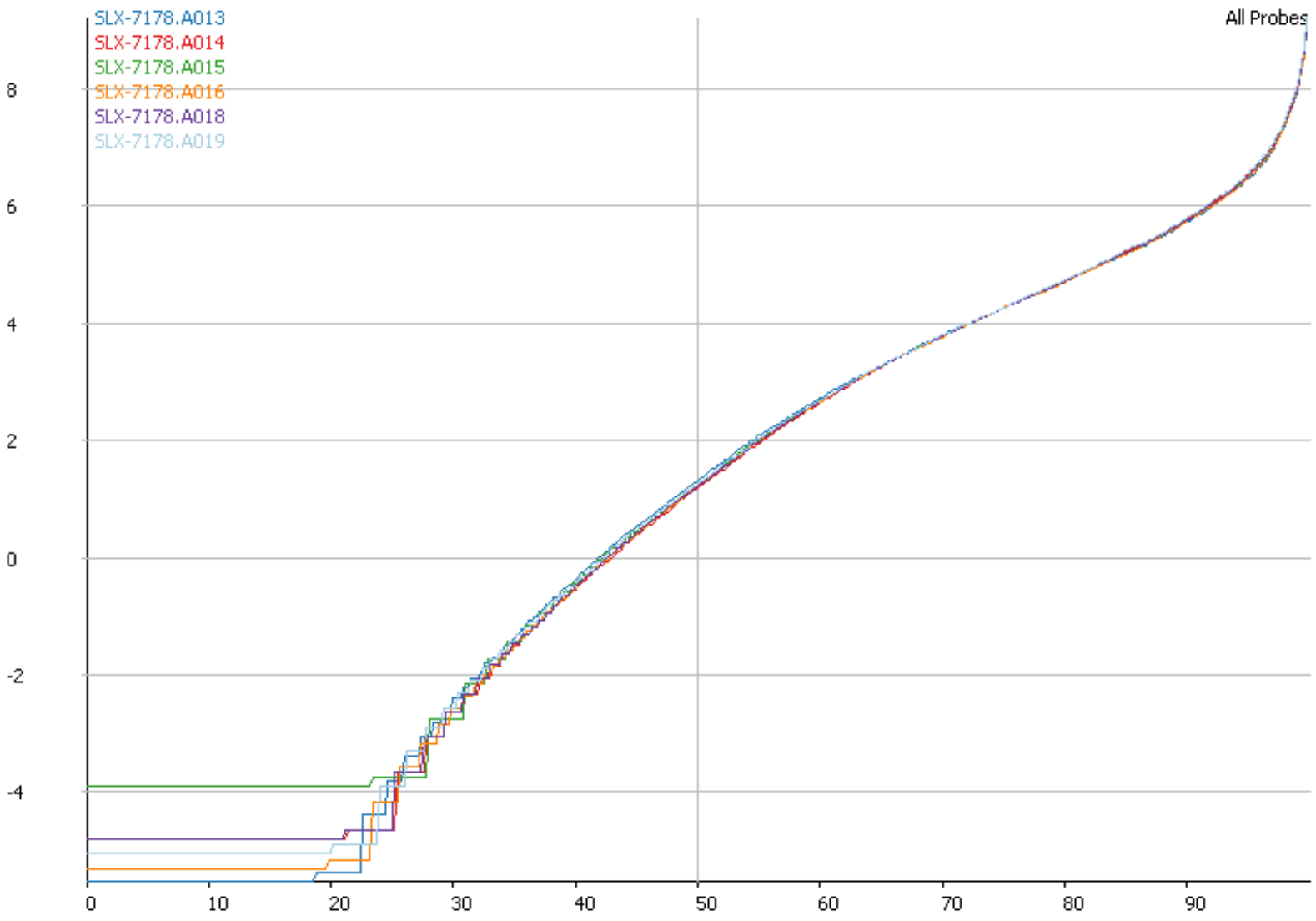
Linear



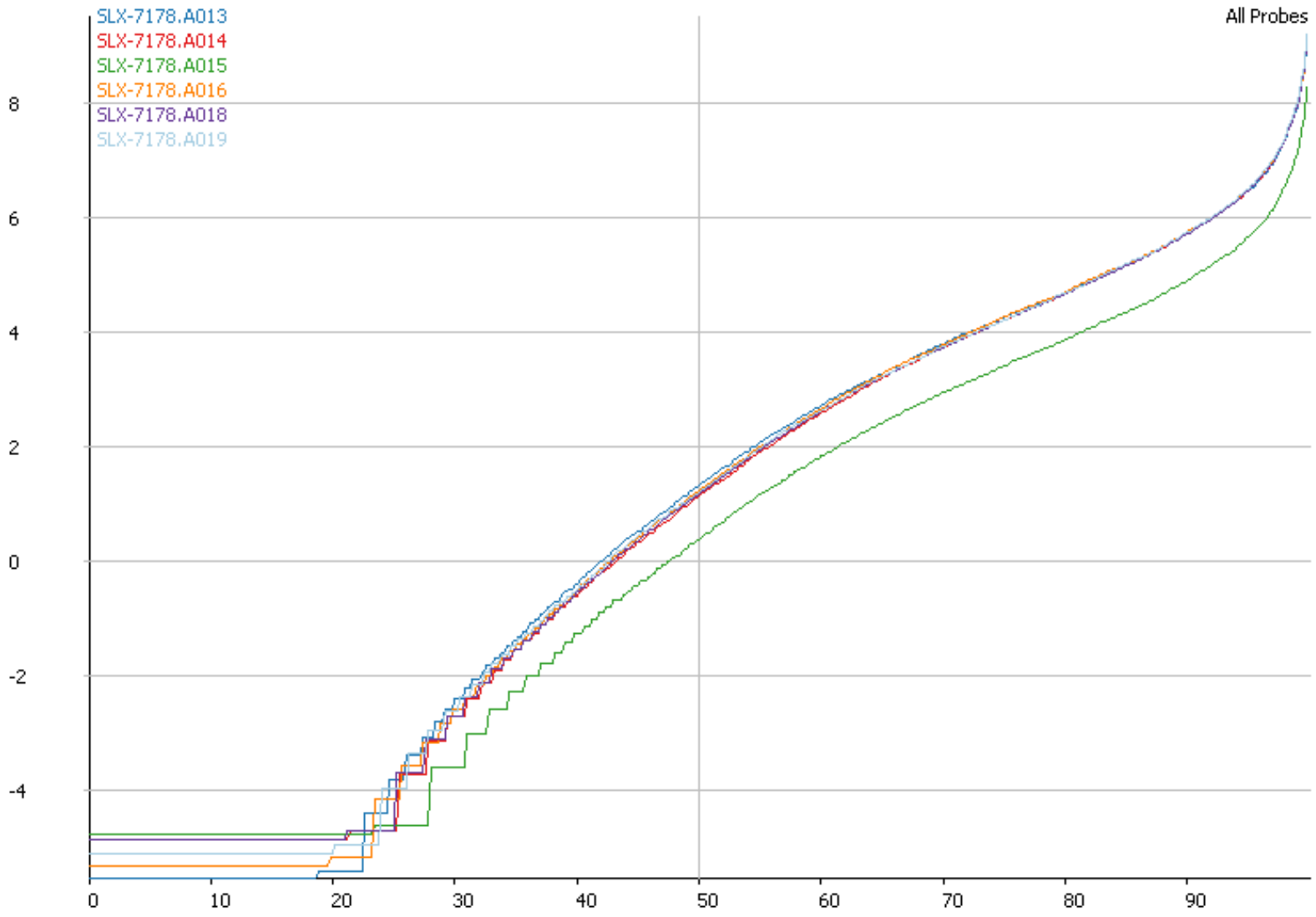
Log2



Visualising Normalisation



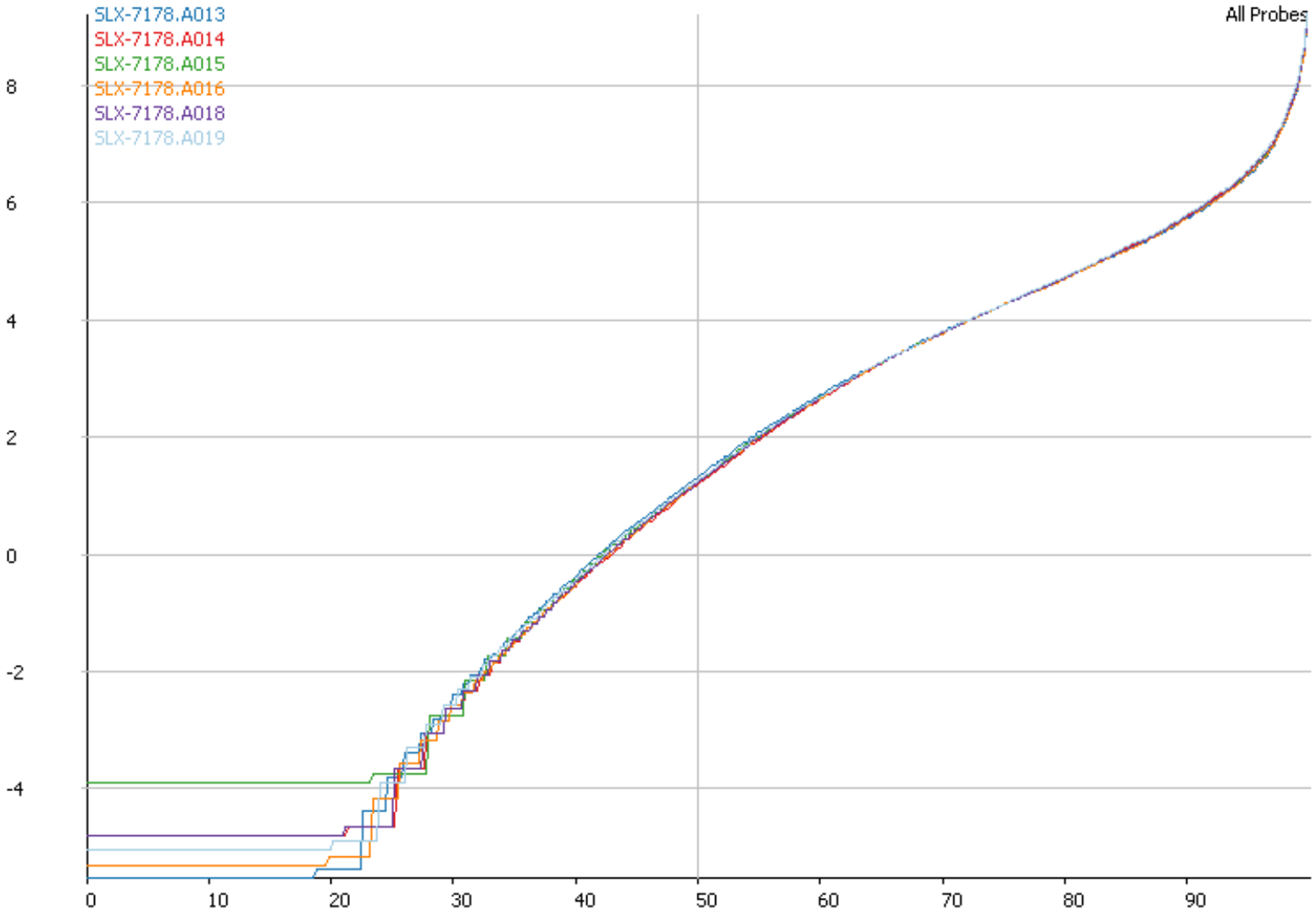
Visualising Normalisation



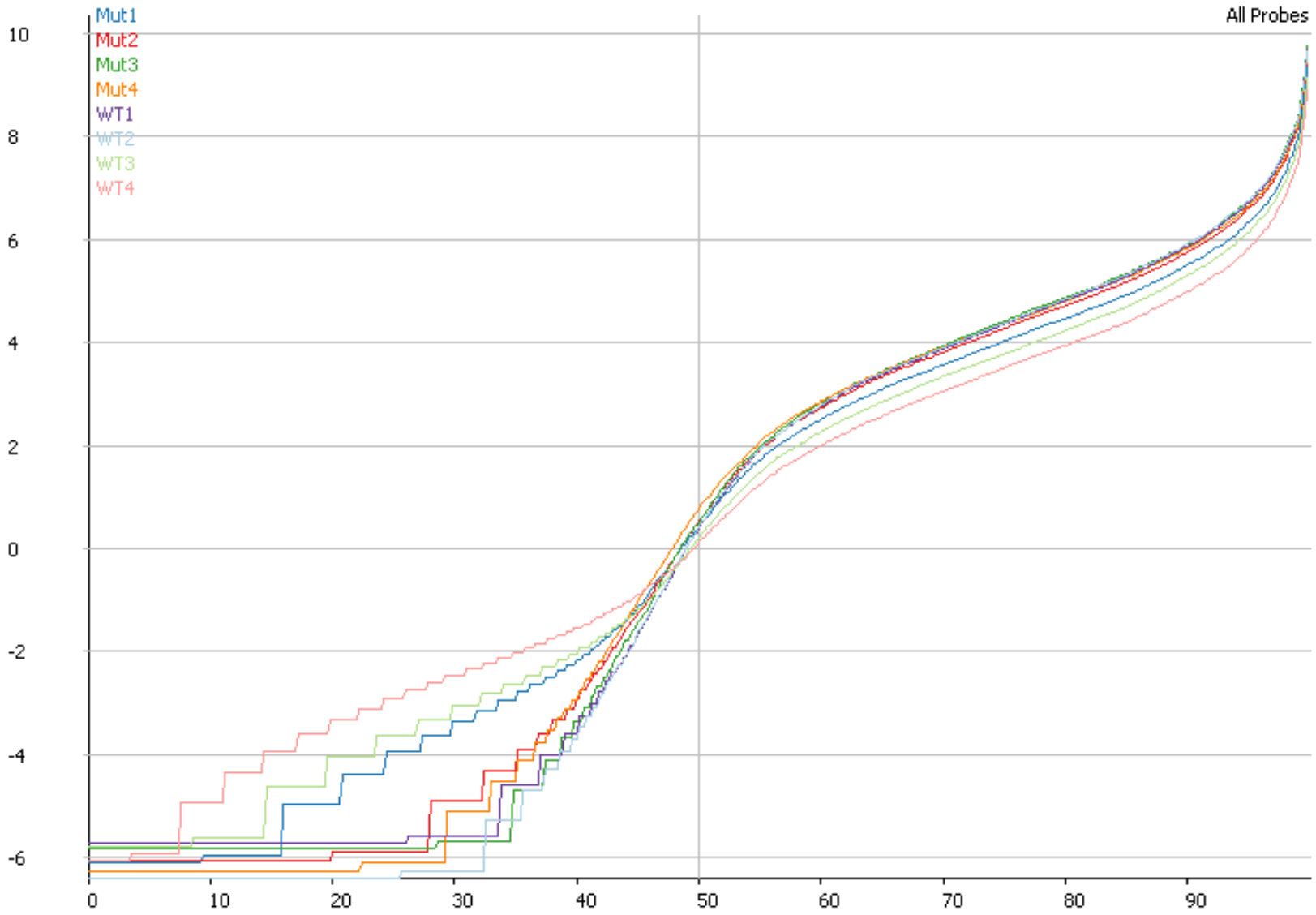
Size Factor Normalisation

- Make an ‘average’ sample from the mean of expression for each gene across all samples
- For each sample calculate the distribution of differences between the data in that sample and the equivalent in the ‘average’ sample
- Use the median of the difference distribution to normalise the data

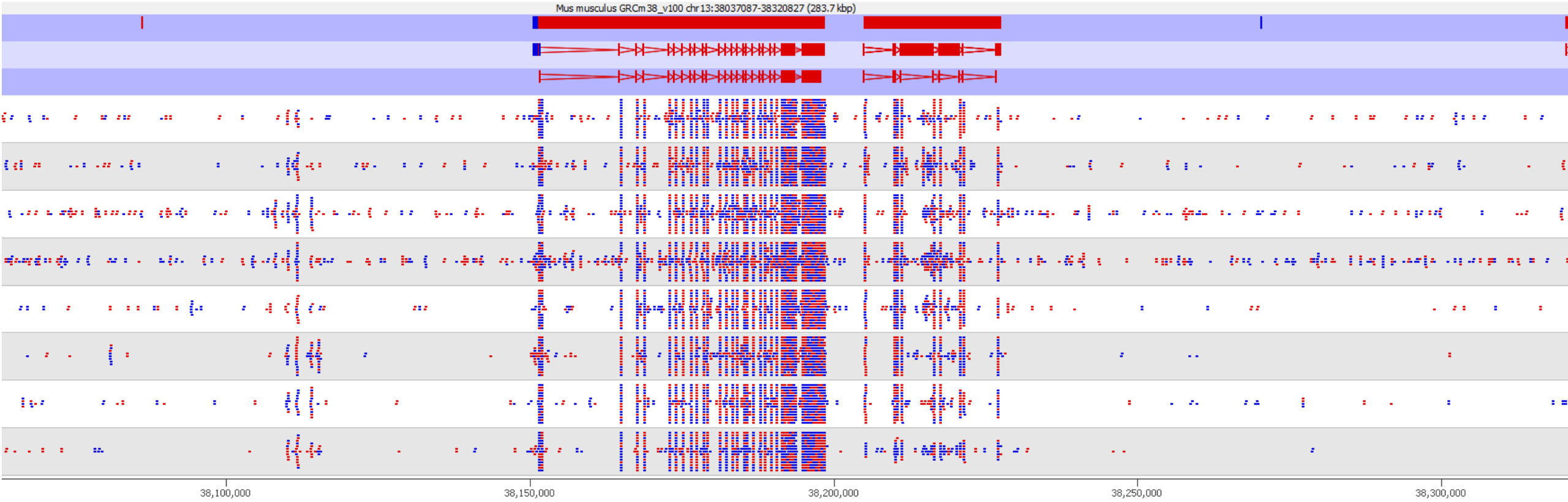
Normalisation – Coverage Outliers



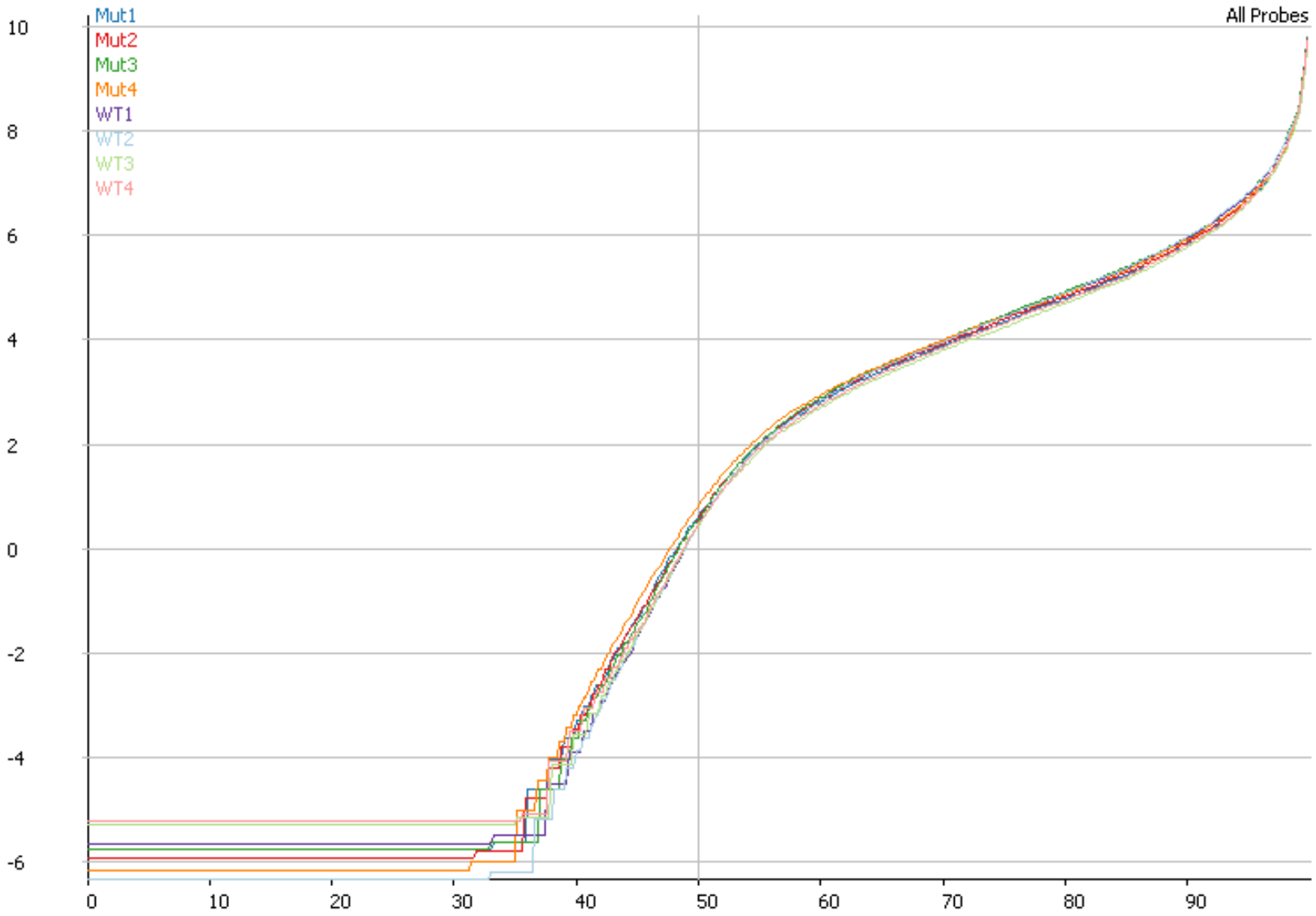
Normalisation – DNA Contamination



Normalisation – DNA Contamination

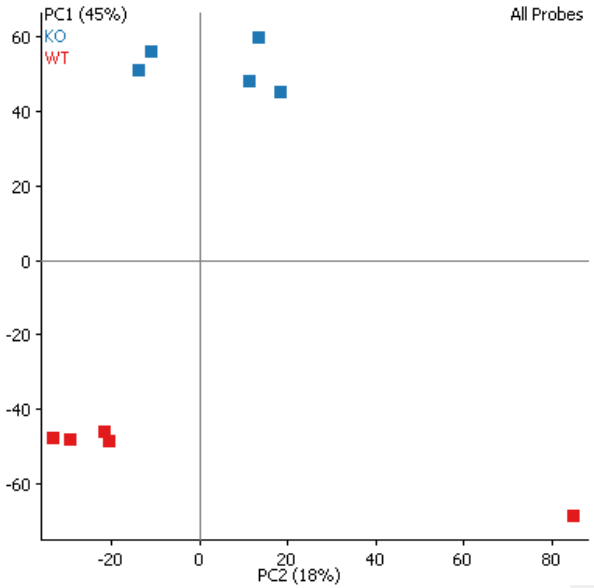


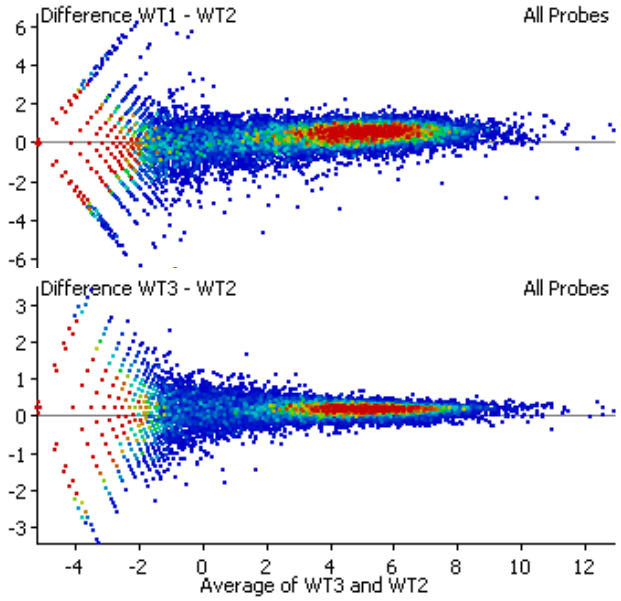
Normalisation – DNA Contamination

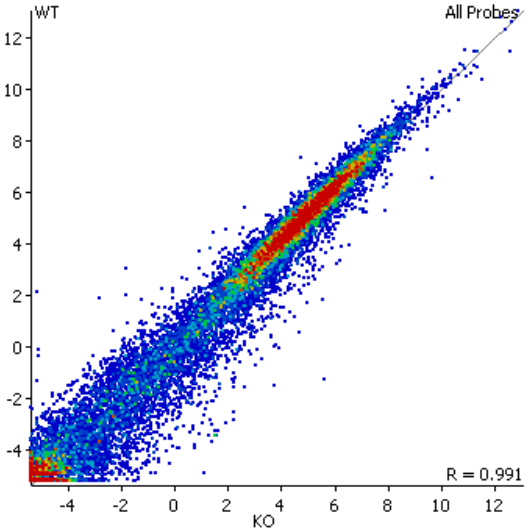


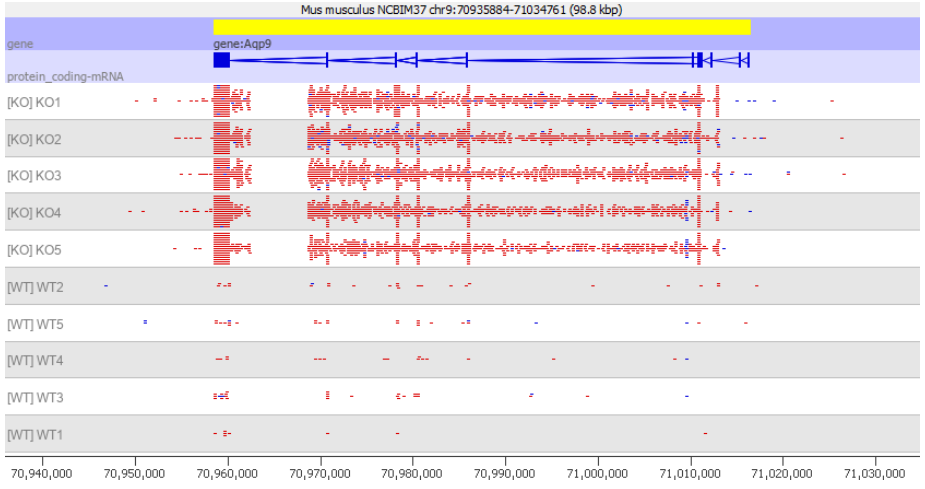
Exploratory Analyses

- Time to **<u>understand</u>** your data
  - Behaviour of raw data and annotation
  - Clustering of samples (PCA / tSNE etc)
  - Pairwise comparisons of samples and groups
  - Are expected effects present (eg KO)?
  - Can I validate other aspects of the samples (eg sex)
  - Can I see obvious changes?
  - Are the changes convincing?





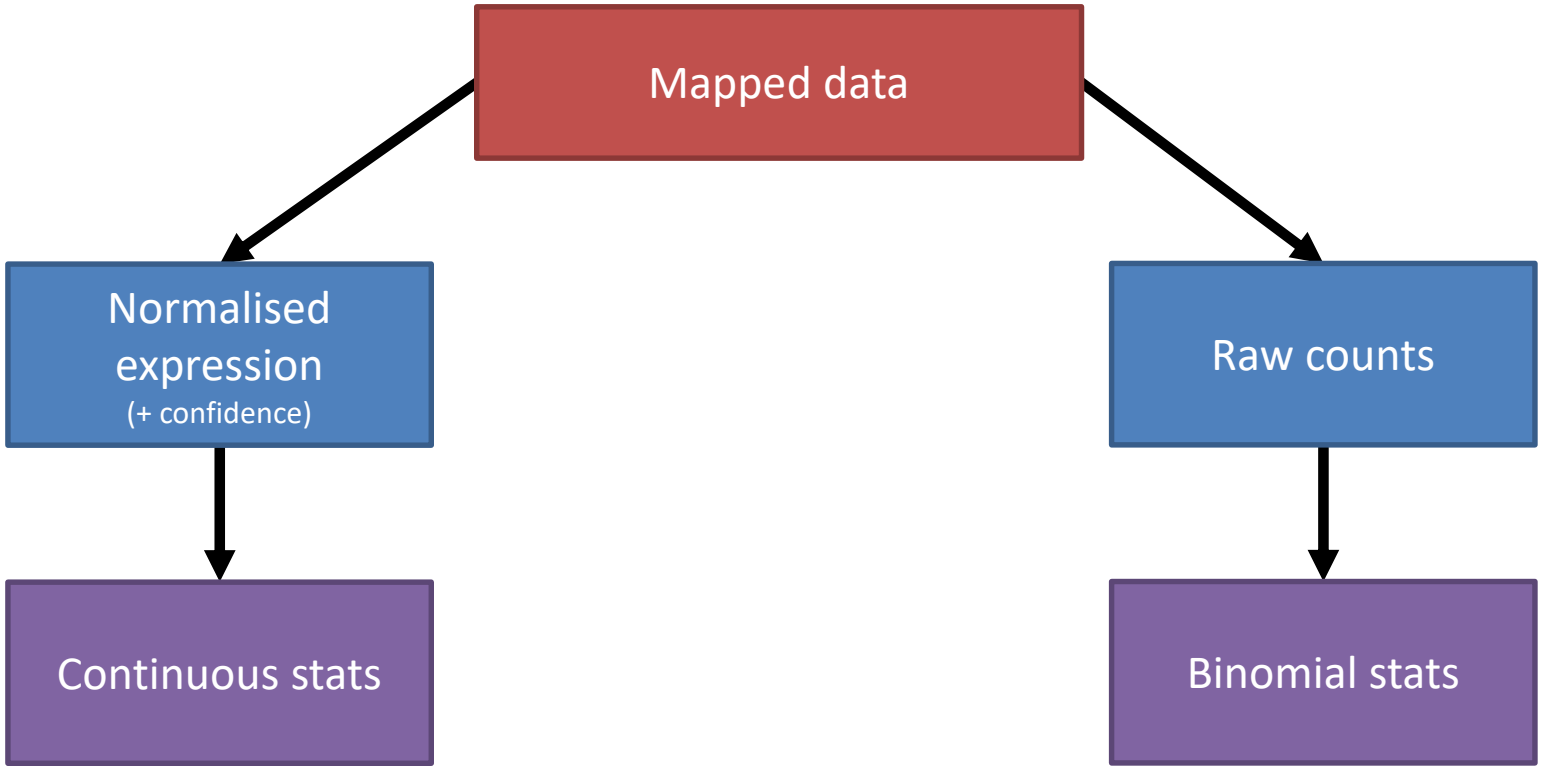




Differential Expression Statistics

Babraham Bioinformatics

Differential Expression



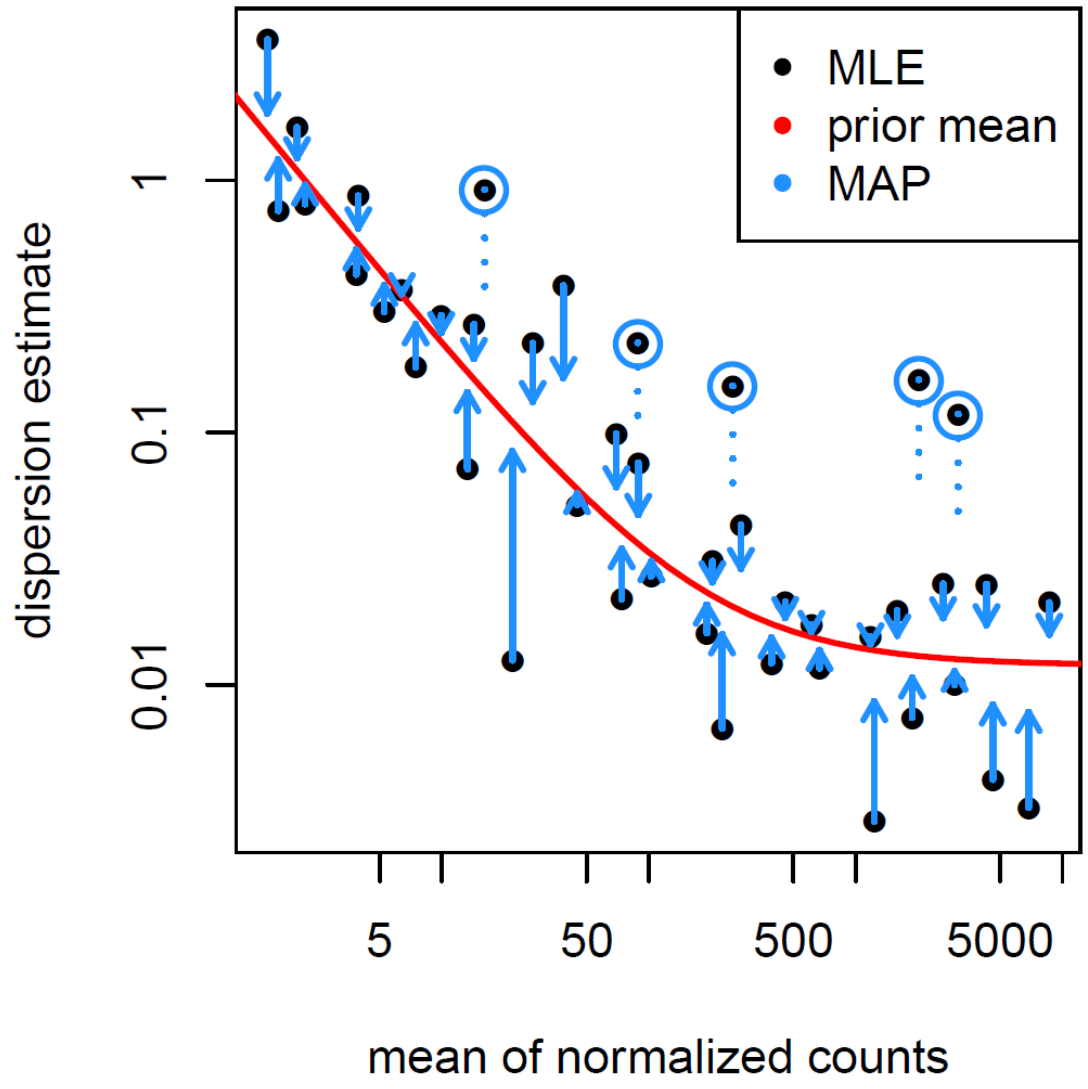
DE-Seq2 binomial Stats

- 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

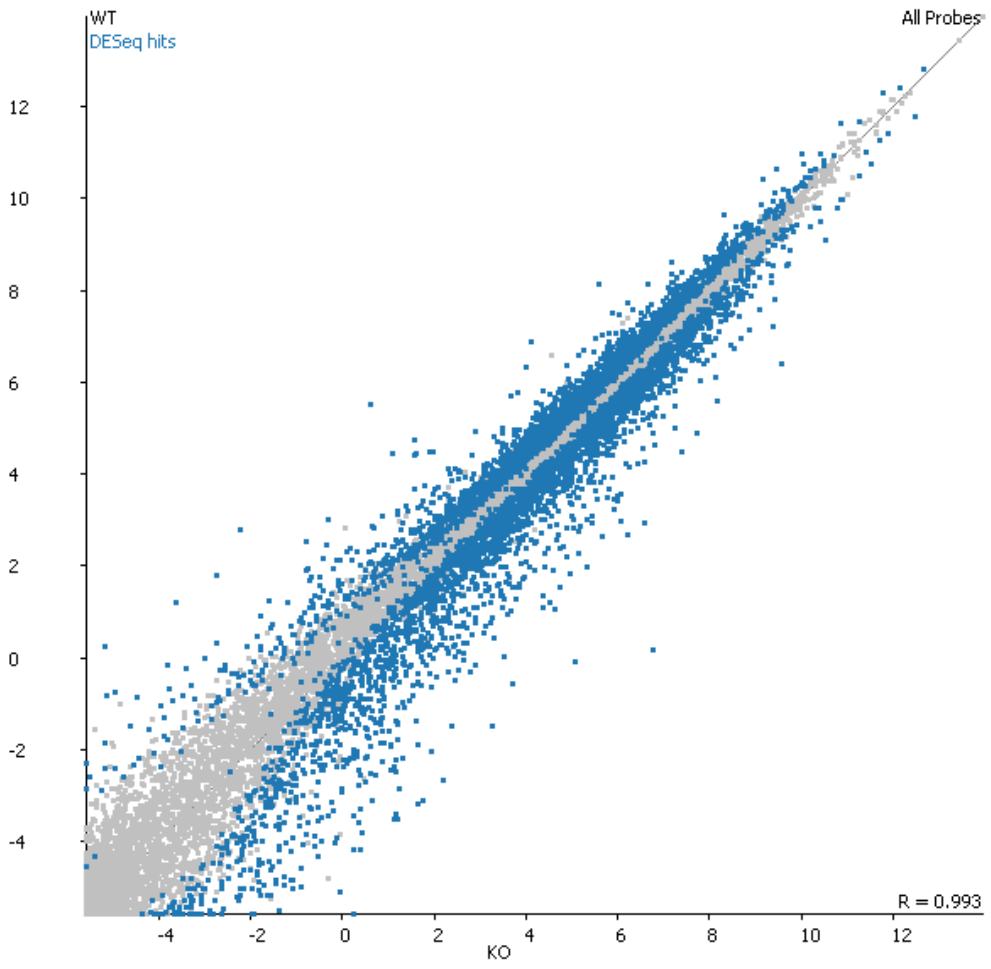
- Variance – required for Negative Binomial distribution
  - Insufficient observations to allow direct measure
  - Custom variance distribution fitted to real data
  - Smooth distribution assumed to allow fitting

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

Visualising Differential Expression Results



5x5 Replicates

8,022 out of 18,570 genes (43%) identified as DE using DESeq ($p<0.05$)

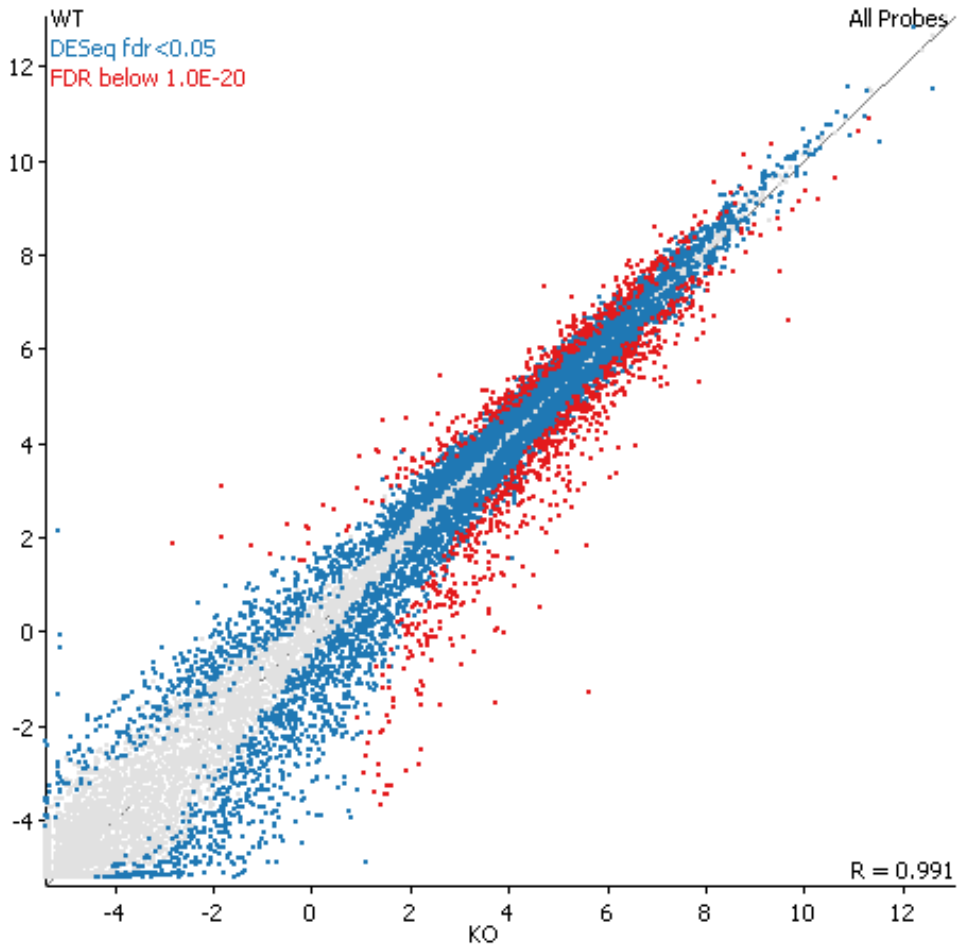
Needs further filtering

Two options:

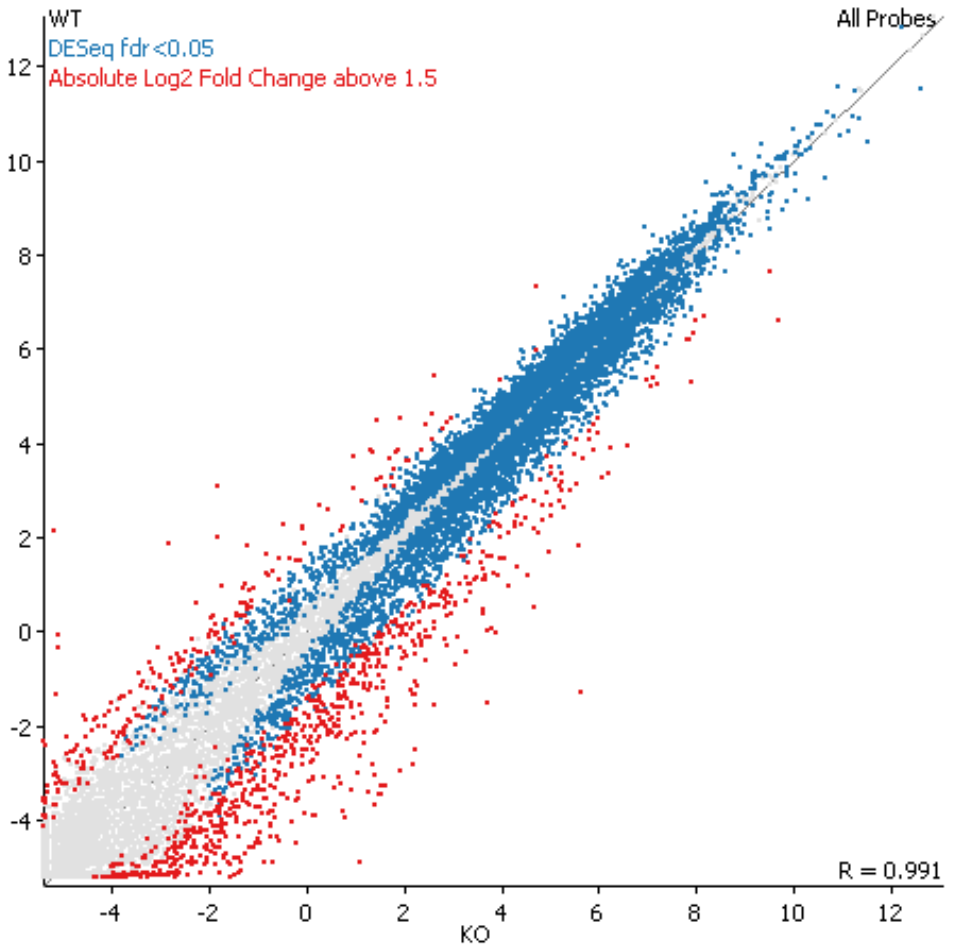
1. Decrease the p-value cutoff
2. Filter on magnitude of change

(both are a bit rubbish)

Visualising Differential Expression Results



Filter by p-value (fdr < 10^{-20})



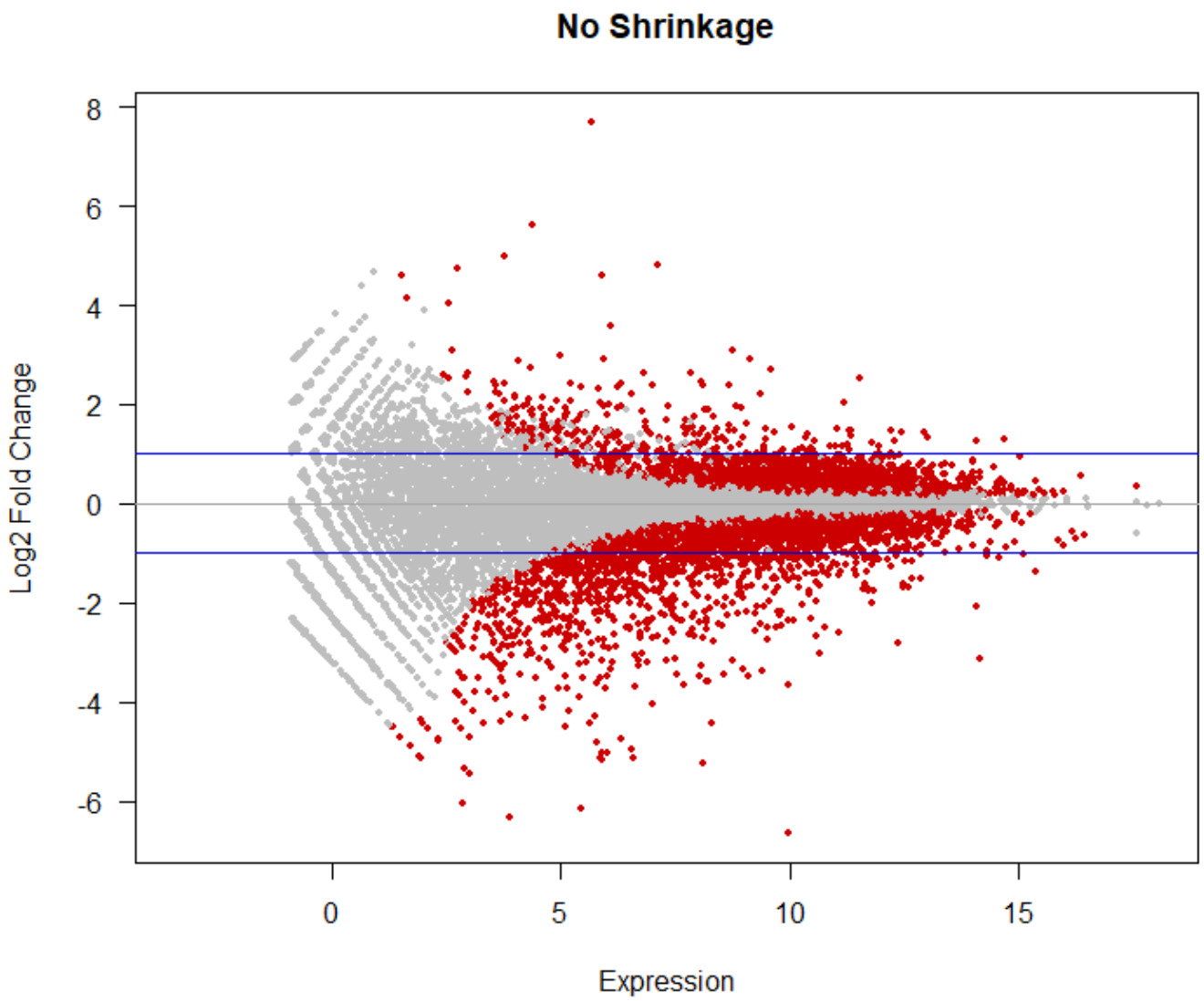
Filter by fold change (abs log2 change > 1.5)

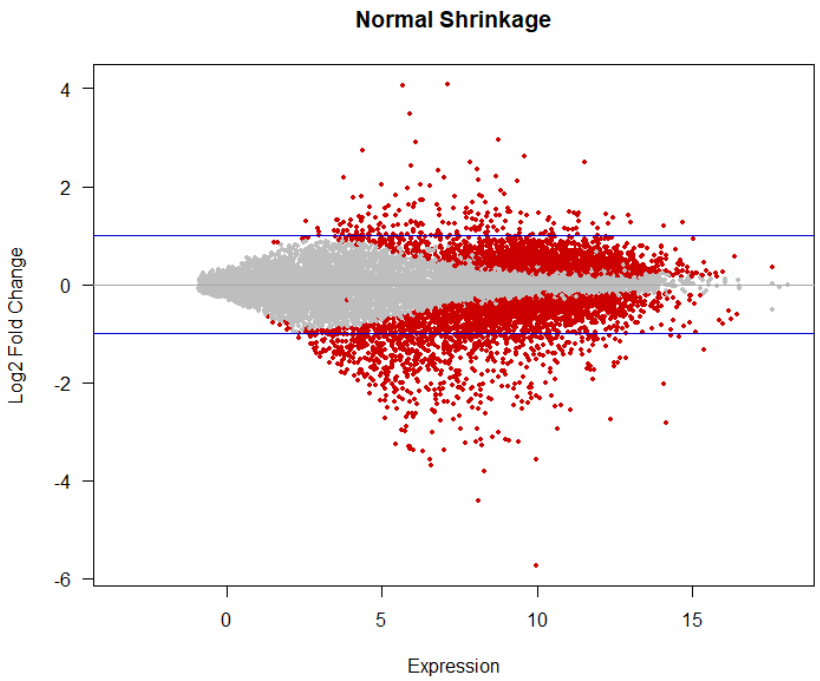
Fold Change Shrinkage

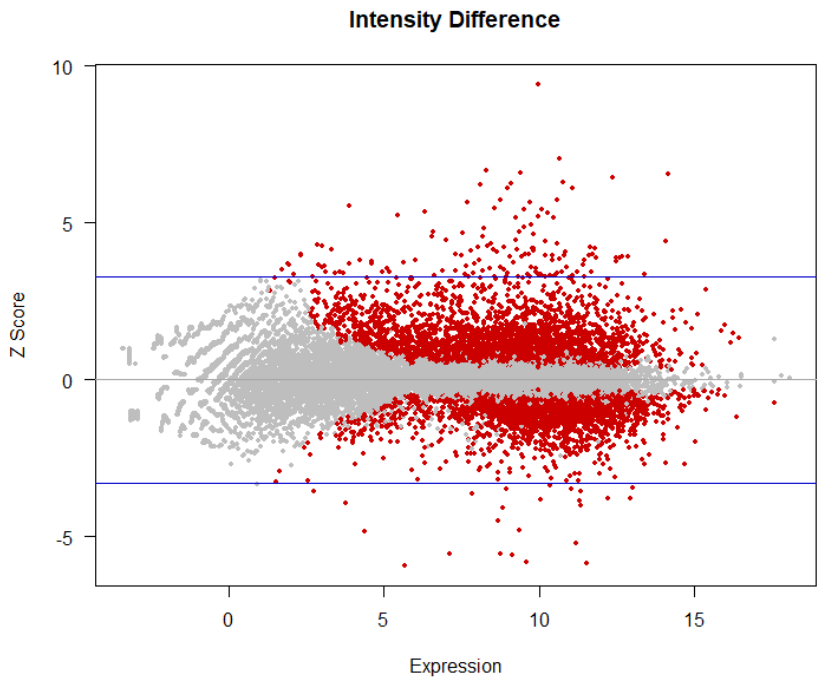
- Aims to make the log2 Fold change a more useful value
- Tries to remove systematic biases
- Two types:
  1. Fold Change Shrinkage – removes bias from both expression level and variance, produces a modified fold change

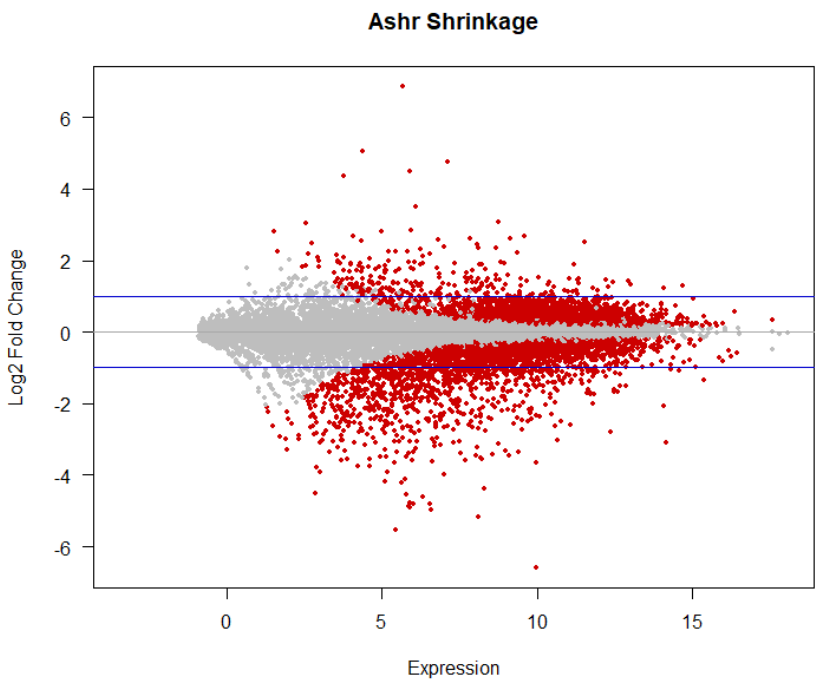
  2. Intensity difference – removes bias from just expression level, produces a p-value

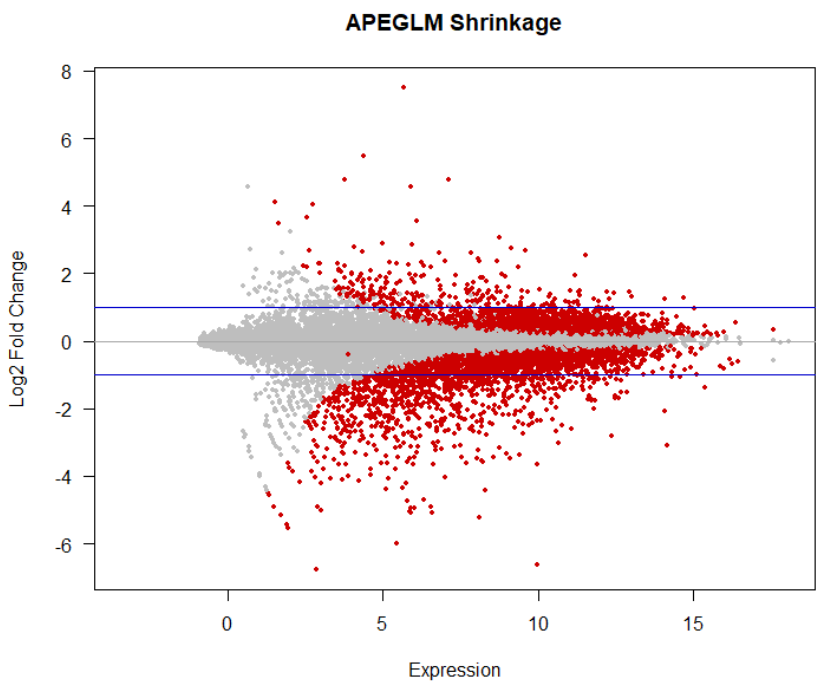
Fold Change Shrinkage




Normal Shrinkage
Log2 Fold Change
Expression
4
2
0
-2
-4
-6
0
5
10
15

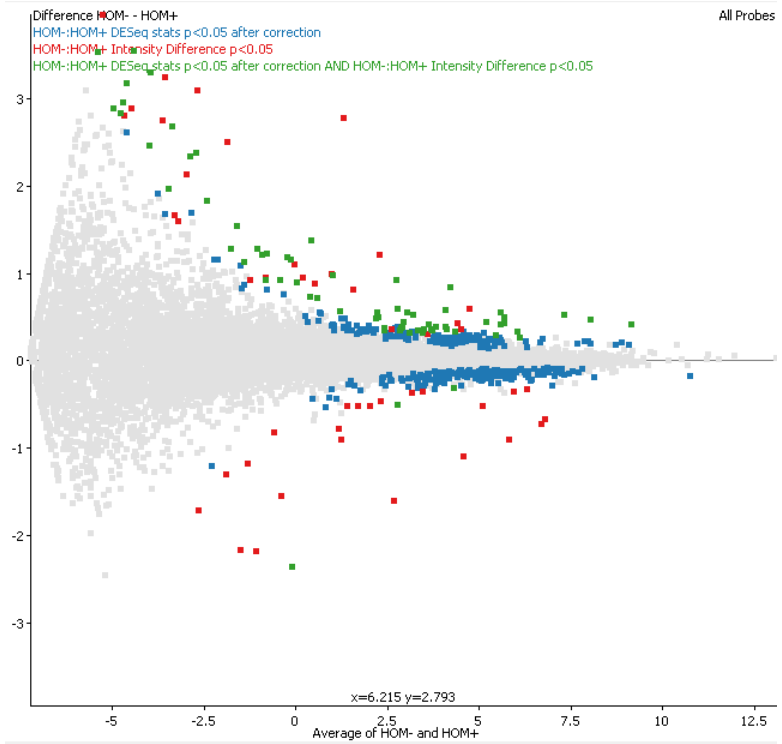

Intensity Difference
Z Score
Expression
10
5
0
-5
0
5
10
15

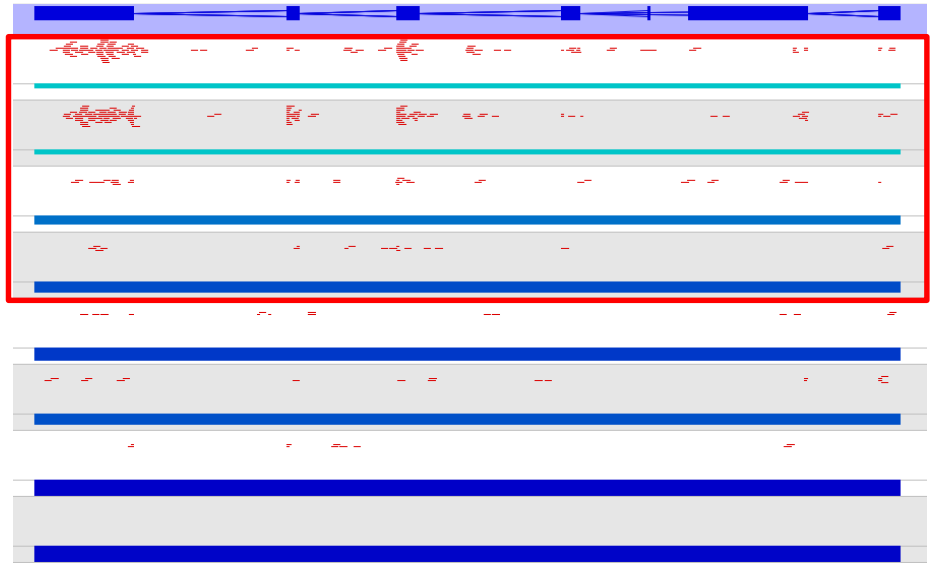

Ashr Shrinkage
Log2 Fold Change
Expression
6
4
2
0
-2
-4
-6
0
5
10
15


APEGLM Shrinkage
Log2 Fold Change
Expression
8
6
4
2
0
-2
-4
-6
0
5
10
15

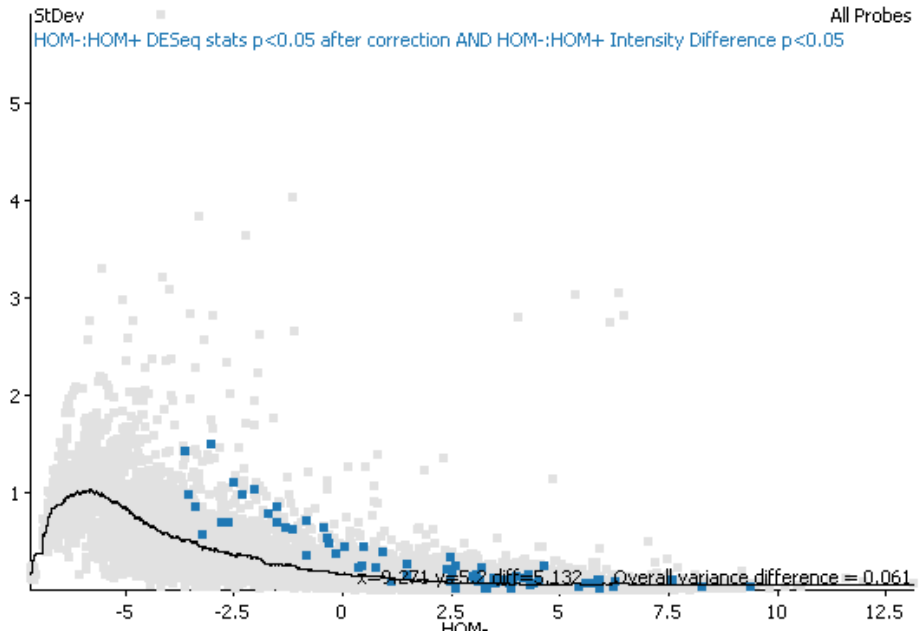
Result Validation
(Can I believe the hits?)

Babraham Bioinformatics





Validation



2900097C17Rik	RIKEN cDNA 2900097C17 gene
Hbb-b1	hemoglobin, beta adult major chain
Rps27a-ps2	ribosomal protein S27A, pseudogene 2
C230073G13Rik	RIKEN cDNA C230073G13 gene
mt-Atp8	mitochondrially encoded ATP synthase 8
mt-Nd4l	mitochondrially encoded NADH dehydrogenase
AC151712.4	erythroid differentiation regulator 1
Gm5641	predicted gene 5641

Data Exploration and Analysis Practical

Babraham Bioinformatics

Experimental Design for RNA-Seq

Babraham Bioinformatics

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, but lots of places don’t offer anything that short!
  - Expression plus splice junction usage
    - 100-150bp single end
  - Novel transcript discovery or per transcript expression
    - 150bp paired end

How many reads

- Typically aim for 20-50 million reads per sample for human or 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

Power Analysis

- Power Analysis is not simple for RNA-Seq data
  - Not a single test – one test per gene
  - Need to apply multiple testing correction
  - Each gene will have different power
    - Power correlates with observation level
    - Variations in variance per gene

- Several tools exist to automate power analysis
  - All require parameters which are difficult to estimate, and have dramatic effects on the outcome

Power Analysis

Yu *et al. BMC Bioinformatics* (2017) 18:234
DOI 10.1186/s12859-017-1648-2

BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

Power analysis for RNA-Seq differential expression studies

Lianbo Yu*, Soledad Fernandez† and Guy Brock†

Tools available

- RnaSeqSampleSize https://cqs-vumc.shinyapps.io/rnaseqsamplesizeweb/
- Scotty http://scotty.genetics.utah.edu/

- All require an estimate of count vs variance
  - Pilot data (if only!)
  - “Similar” studies

We are planning a RNA sequencing experiment to identify differential gene expression between two groups. Prior data indicates that the minimum average read counts among the prognostic genes in the control group is **500**, *the maximum dispersion is* **0.1**, *and the ratio of the geometric mean of normalization factors is* **1**. *Suppose that the total number of genes for testing is* **10000** *and the top* **100** *genes are prognostic. If the desired minimum fold change is* **3**, *we will need to study* **4** *subjects in each group to be able to reject the null hypothesis that the population means of the two groups are equal with probability (power)* **0.8** *using exact test. The FDR associated with this test of this null hypothesis is* **0.05**.

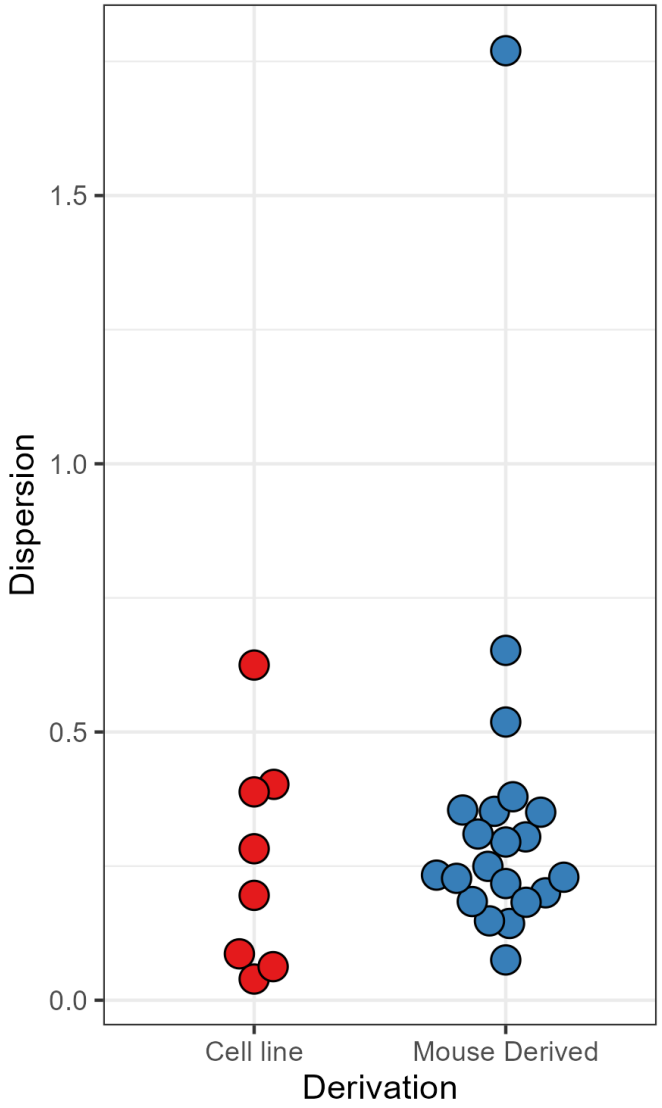
Predicting variability

Simple = clean

Complex = noisy

Umar Hashim

Predicting Variability

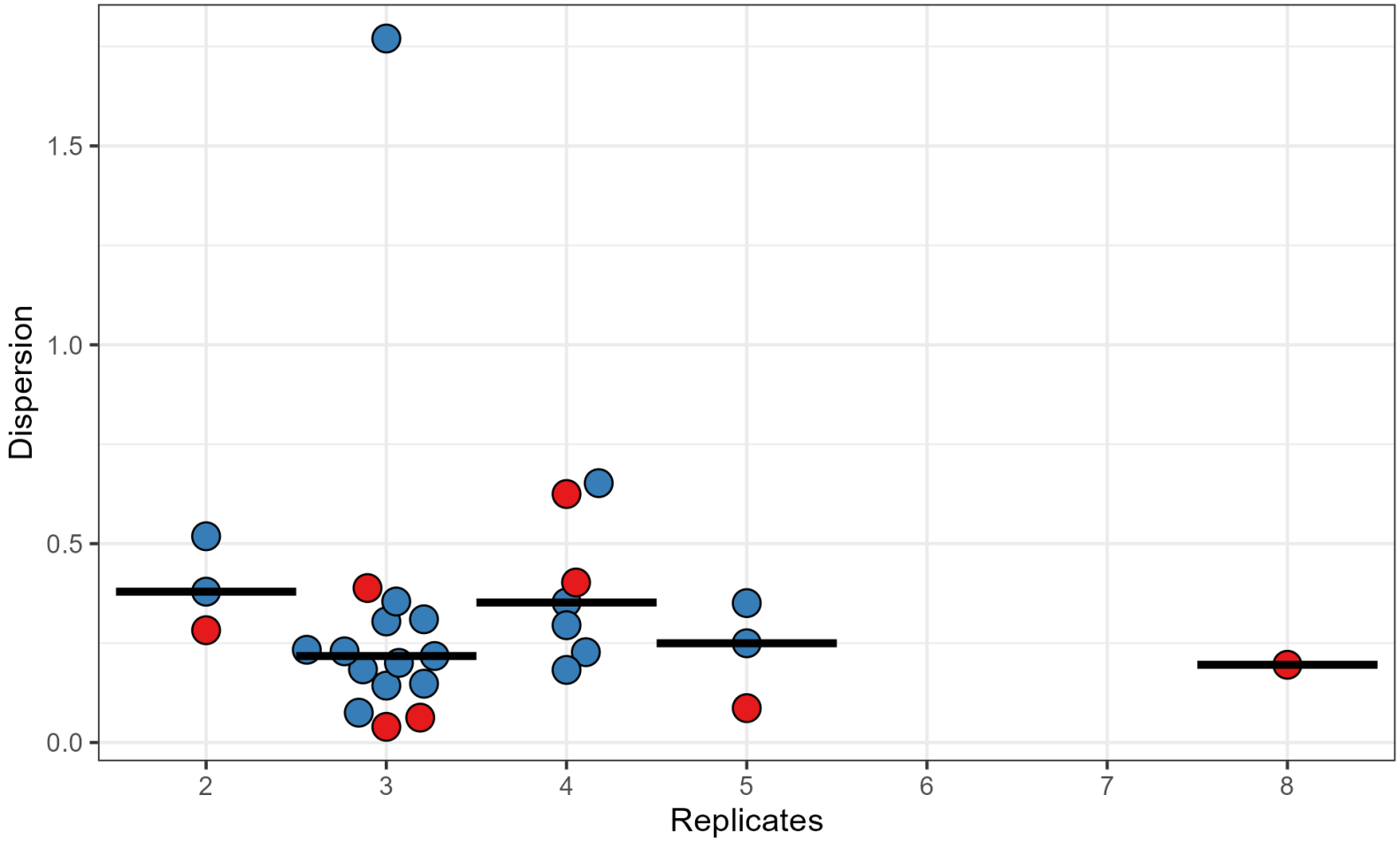


No clear difference in the variability of replicates coming from cell lines vs animal samples

Knowing the sample preparation doesn't help you predict how many replicates you need

Umar Hashim

Predicting Variability



- Most groups just do 3 replicates
- Those that do more, didn't have noisier data
- People are bad at estimating

Umar Hashim

Power Curves

n: Sample Size

10

f: FDR level

0.05

m: Total number of genes for testing

20000

m1: Expected number of prognostic genes

200

rho: Minimum fold changes for prognostic genes between two groups

2

lambda0: Average read counts for prognostic genes

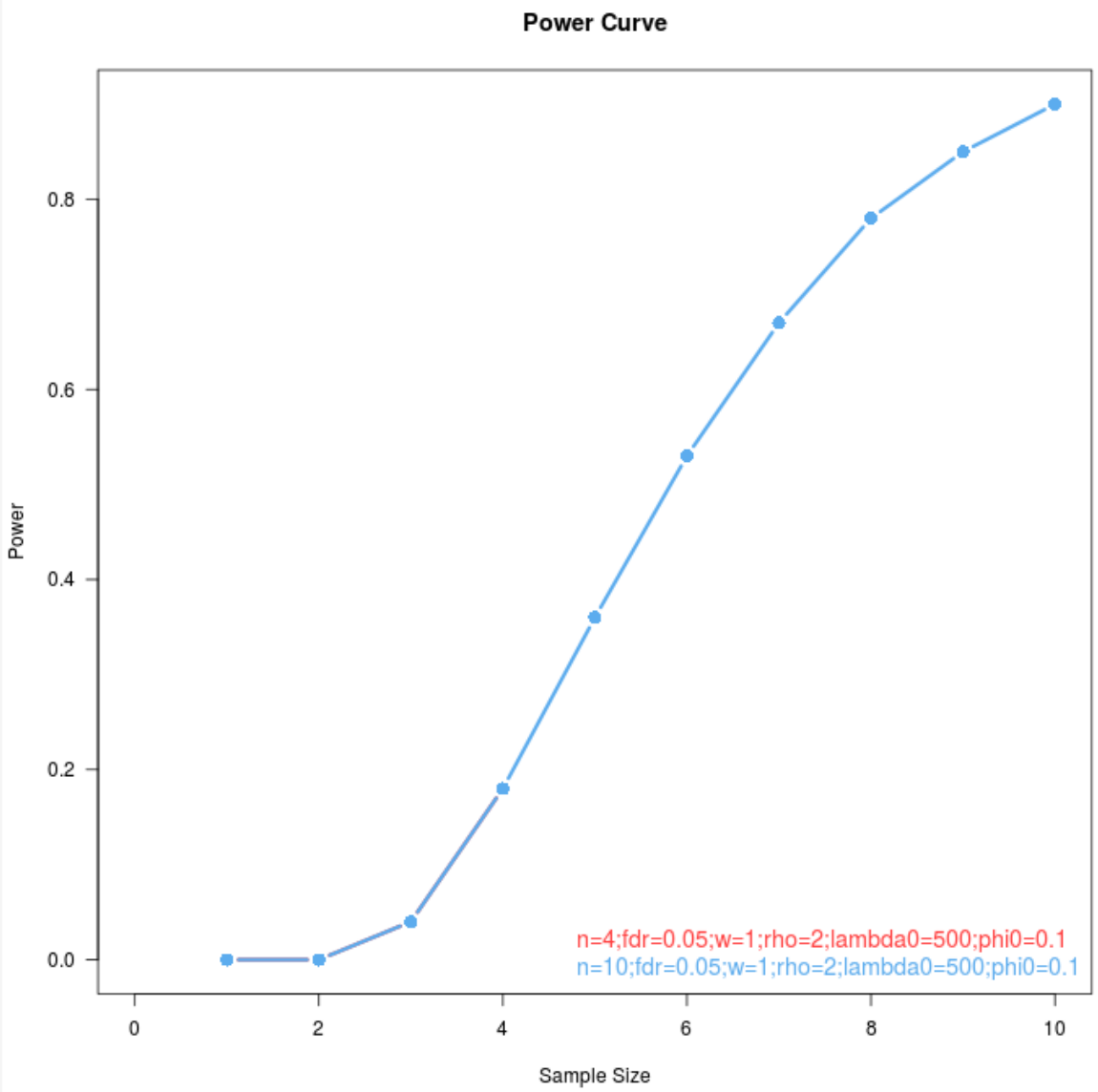
500

phi0: Dispersion for prognostic genes

0.1

w: Ratio of normalization factors between two groups

1



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/