

RNA-Seq Quality Control

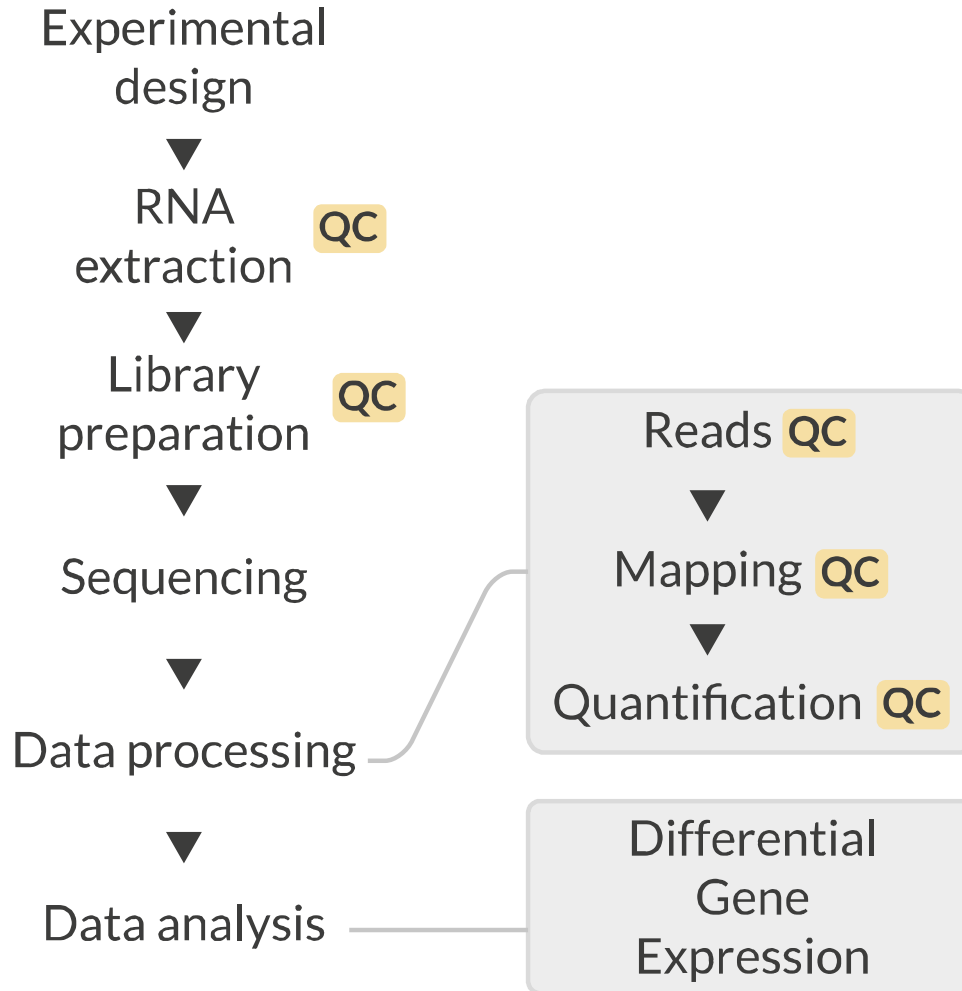
RNA-Seq Analysis Workshop

Roy Francis | 23-Oct-2018

Contents

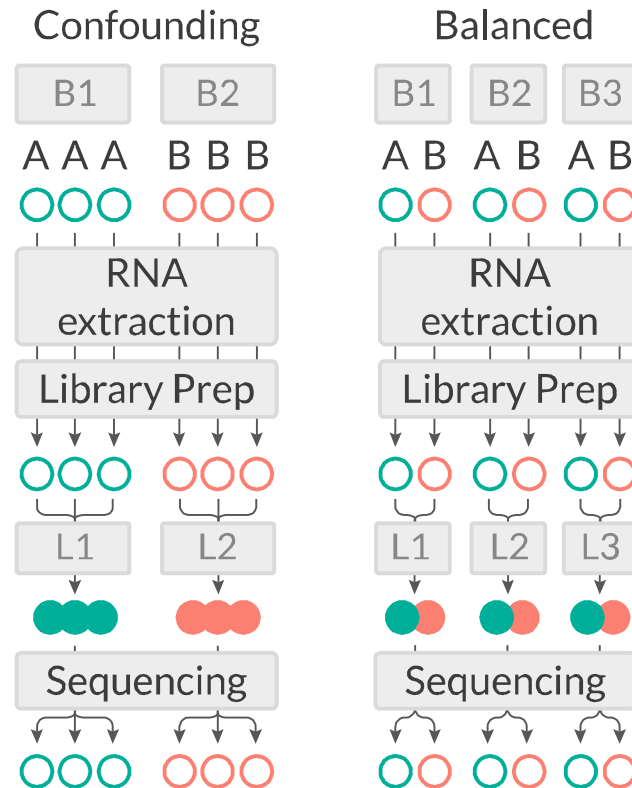
- Workflow
- RNA extraction
- Read QC
- Alignment QC
- Quantification QC
- Exploratory
- Batch correction
- Spike-Ins

Workflow



Experimental design

- Balanced design
- Technical replicates not necessary
(Marioni *et al.*, 2008)
- Biological replicates: 6 - 12 (Schurch *et al.*, 2016)
- ENCODE consortium
- Previous publications
- Power analysis



Busby, Michele A., *et al.* "Scotty: a web tool for designing RNA-Seq experiments to measure differential gene expression." *Bioinformatics* 29.5 (2013): 656-657

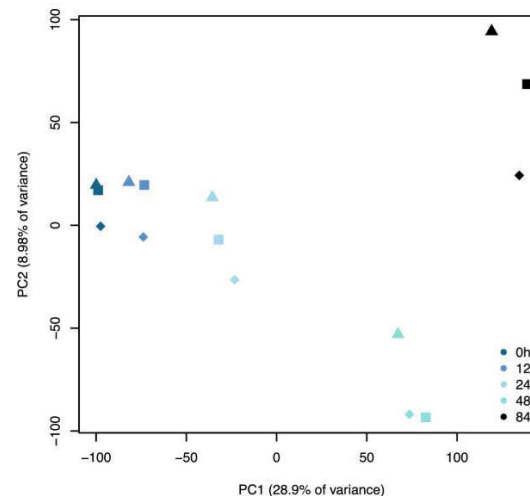
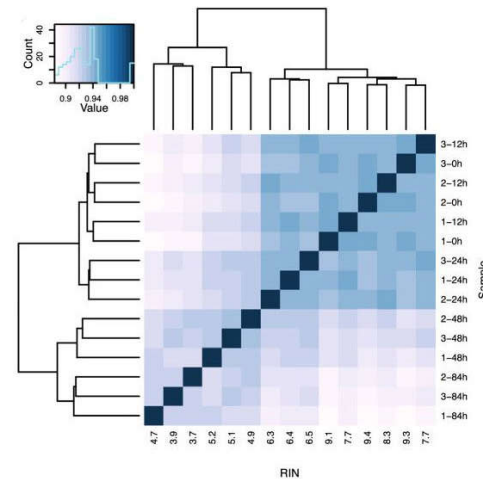
Marioni, John C., *et al.* "RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays." *Genome research* (2008)

Schurch, Nicholas J., *et al.* "How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use?." *Rna* (2016)

Zhao, Shilin, *et al.* "RnaSeqSampleSize: real data based sample size estimation for RNA sequencing." *BMC bioinformatics* 19.1 (2018): 191

RNA extraction

- Sample processing and storage
- RNA quality/quantity
- RIN values (Strong effect)
- DNase treatment
- RNA type
- Contamination/Cross-contamination
- Batch effect
- Extraction method bias (GC bias)



🔗 Romero, Irene Gallego, *et al.* "RNA-seq: impact of RNA degradation on transcript quantification." *BMC biology* 12.1 (2014): 42

🔗 Kim, Young-Kook, *et al.* "Short structured RNAs with low GC content are selectively lost during extraction from a small number of cells." *Molecular cell* 46.6 (2012): 893-89500481-9).

- PolyA selection
- rRNA depletion
- Size selection
- PCR amplification (See section PCR duplicates)
- Stranded (directional) libraries
 - Accurately identify sense/antisense transcript
 - Resolve overlapping genes
- Exome capture
- Library normalisation
- Batch effect

- Number of reads
- Per base sequence quality
- Per sequence quality score
- Per base sequence content
- Per sequence GC content
- Per sequence GC content
- Per base N content
- Sequence length distribution
- Sequence duplication levels
- Overrepresented sequences
- Adapter content
- Kmer content



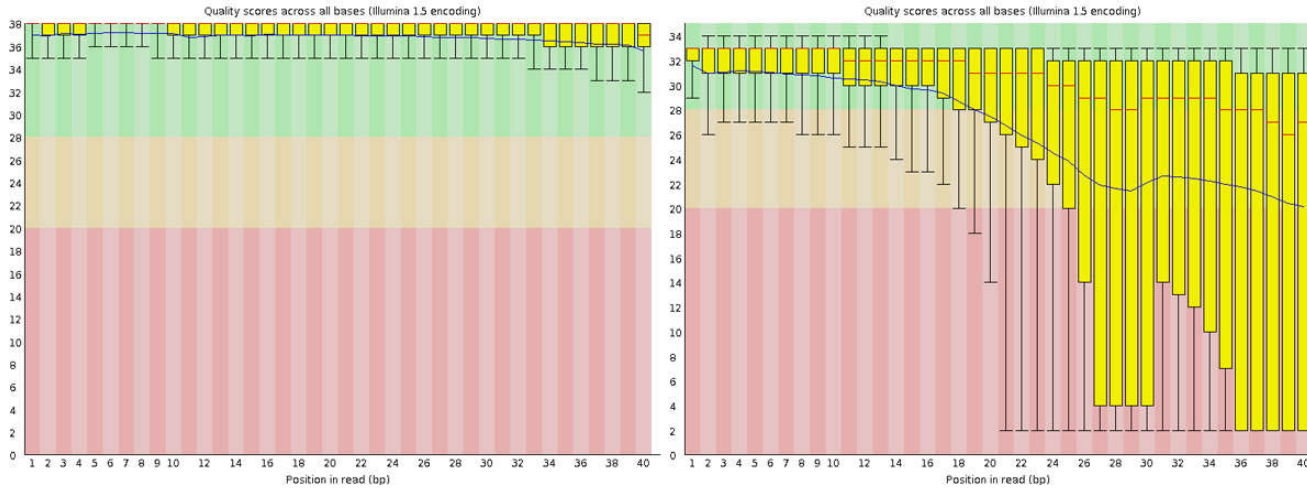
 [FastQC](#), [MultiQC](#)

<https://sequencing.qcfail.com/>

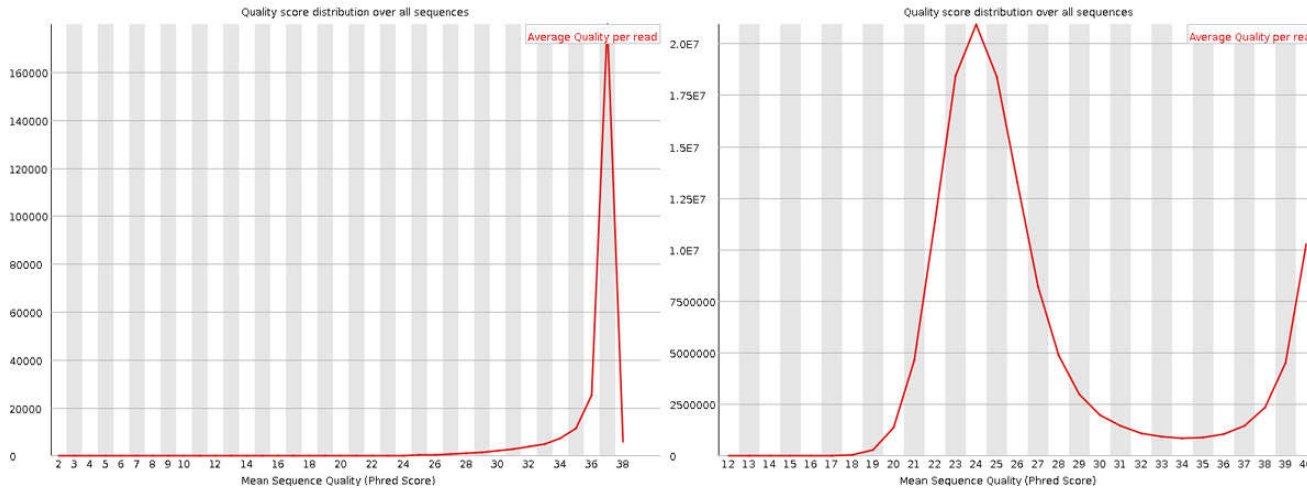
 **QCFAIL.com**

Articles about common next-generation
sequencing problems

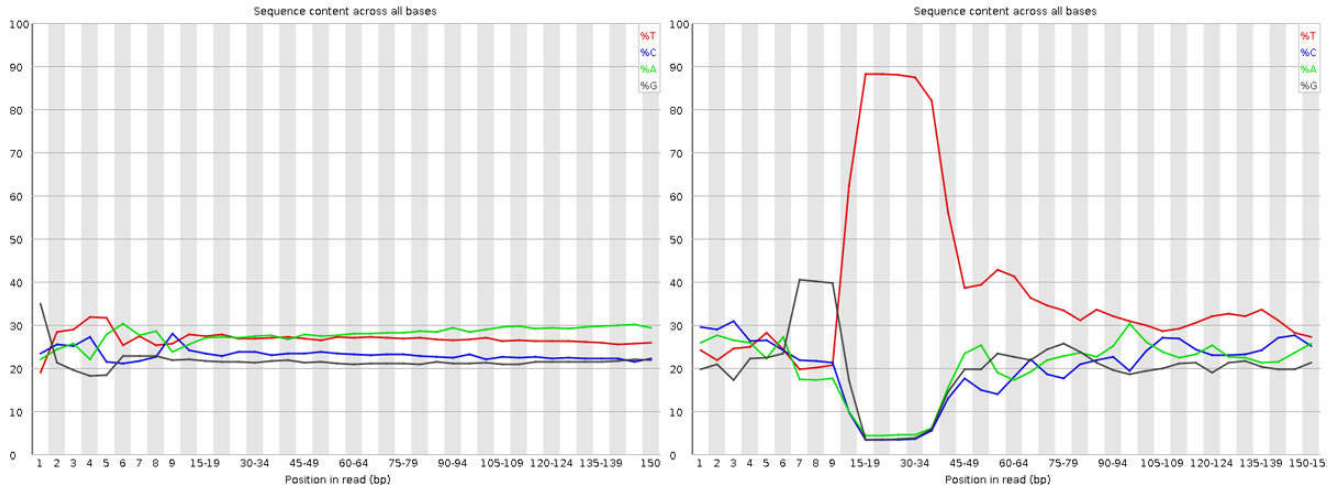
Per base sequence quality



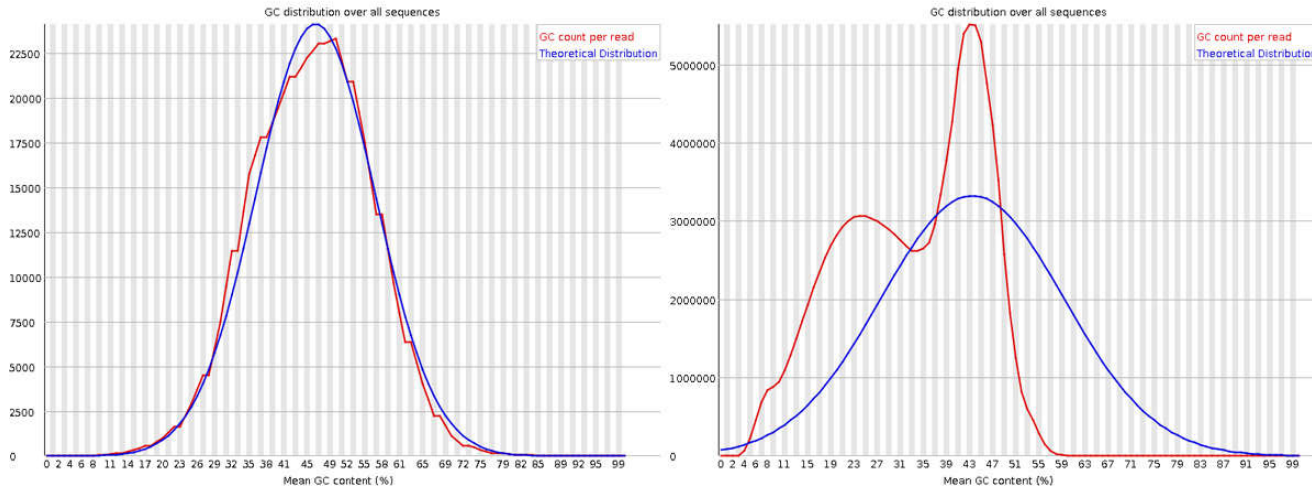
Per sequence quality scores



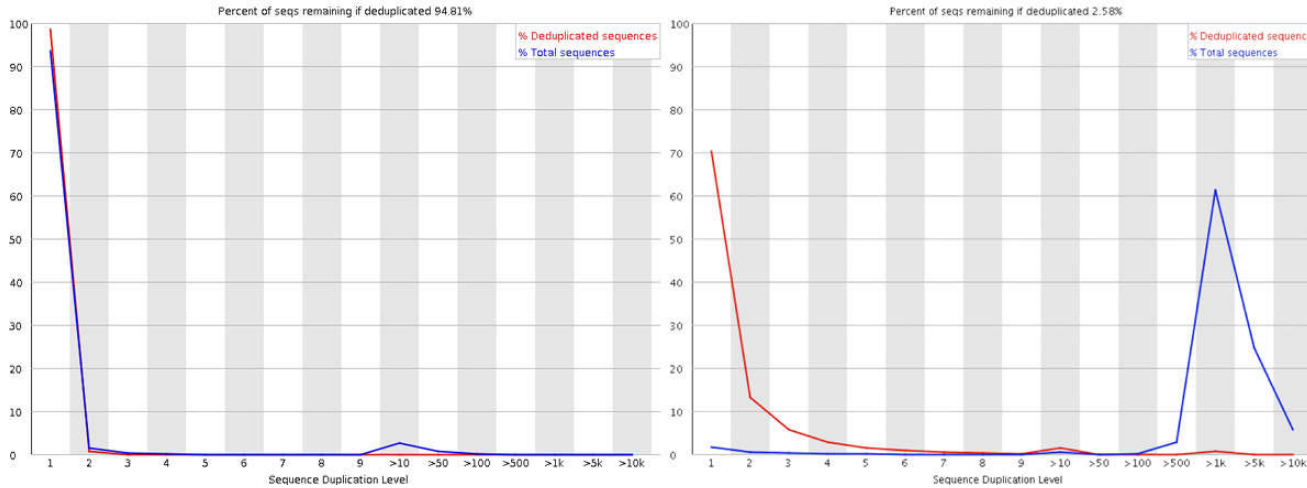
Per base sequence content



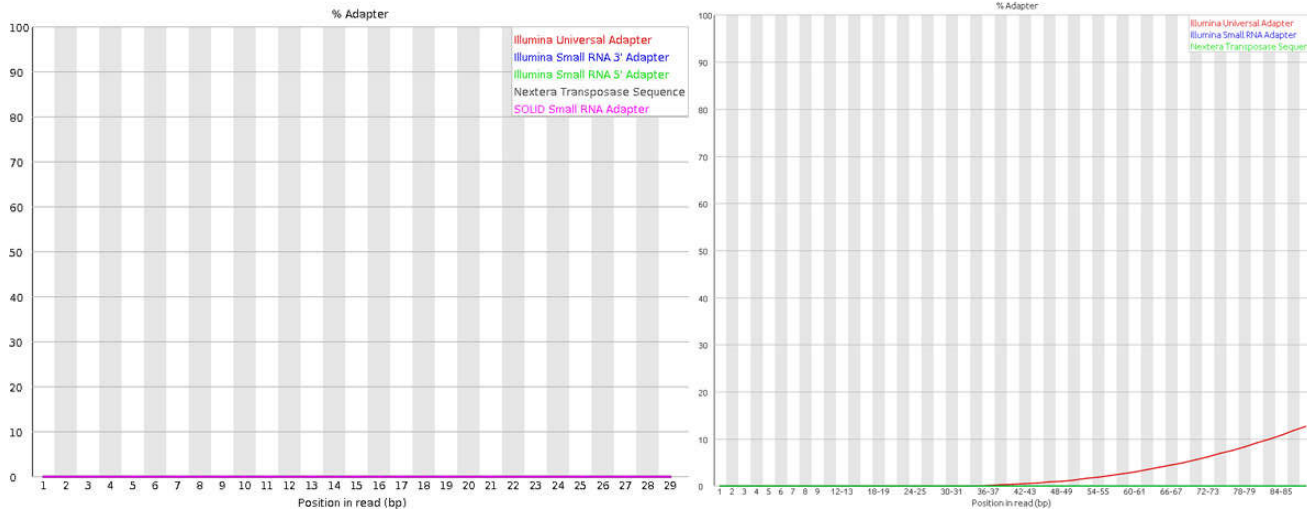
Per sequence GC content



Sequence duplication level



Adapter content



FastQC Report Thu 21 Dec 2017
good_sequence_short.txt

Summary

- [Basic Statistics](#)
- [Per base sequence quality](#)
- [Per tile sequence quality](#)
- [Per sequence quality scores](#)
- [Per base sequence content](#)
- [Per sequence GC content](#)
- [Per base N content](#)
- [Sequence Length Distribution](#)
- [Sequence Duplication Levels](#)
- [Overrepresented sequences](#)
- [Adapter Content](#)

Basic Statistics

Measure	Value
Filename	good_sequence_short.txt
File type	Conventional base calls
Encoding	Illumina 1.5
Total Sequences	250000
Sequences flagged as poor quality	0
Sequence length	40
%GC	45

Per base sequence quality

Quality scores across all bases (Illumina 1.5 encoding)

FastQC Report Thu 21 Dec 2017
bad_sequence.txt

Summary

- [Basic Statistics](#)
- [Per base sequence quality](#)
- [Per tile sequence quality](#)
- [Per sequence quality scores](#)
- [Per base sequence content](#)
- [Per sequence GC content](#)
- [Per base N content](#)
- [Sequence Length Distribution](#)
- [Sequence Duplication Levels](#)
- [Overrepresented sequences](#)
- [Adapter Content](#)

Basic Statistics

Measure	Value
Filename	bad_sequence.txt
File type	Conventional base calls
Encoding	Illumina 1.5
Total Sequences	395288
Sequences flagged as poor quality	0
Sequence length	40
%GC	47

Per base sequence quality

Quality scores across all bases (Illumina 1.5 encoding)

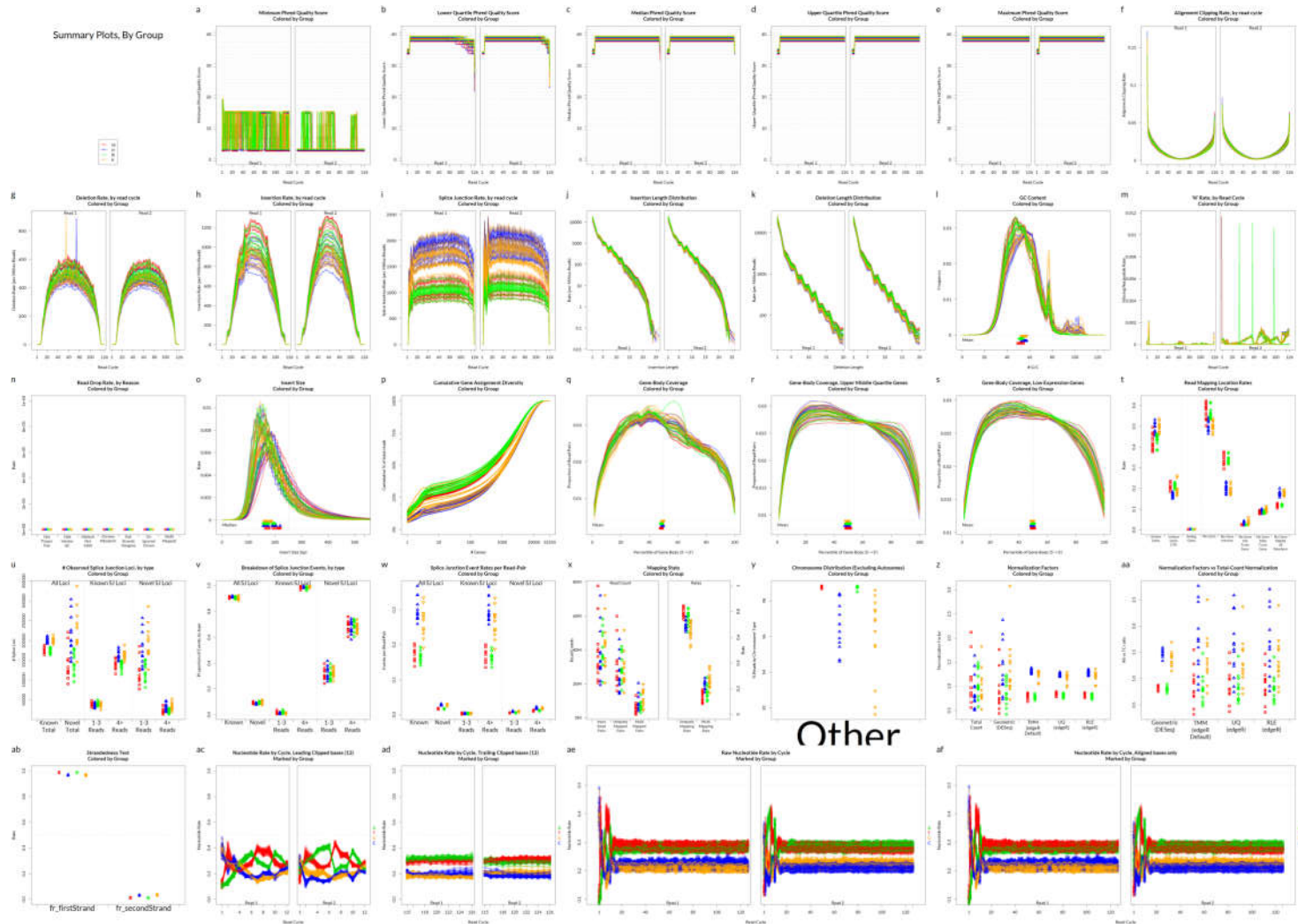
- Trim IF necessary
 - Synthetic bases can be an issue for SNP calling
 - Insert size distribution may be more important for assemblers
- Trim/Clip/Filter reads
- Remove adapter sequences
- Trim reads by quality
- Sliding window trimming
- Filter by min/max read length
 - Remove reads less than ~22nt
- Demultiplexing/Splitting

 [Cutadapt](#), [fastp](#), [Skewer](#), [Prinseq](#)

- Number of reads mapped/unmapped/paired etc
- Uniquely mapped
- Insert size distribution
- Coverage
- Gene body coverage
- Biotype counts / Chromosome counts
- Counts by region: gene/intron/non-genic
- Sequencing saturation
- Strand specificity

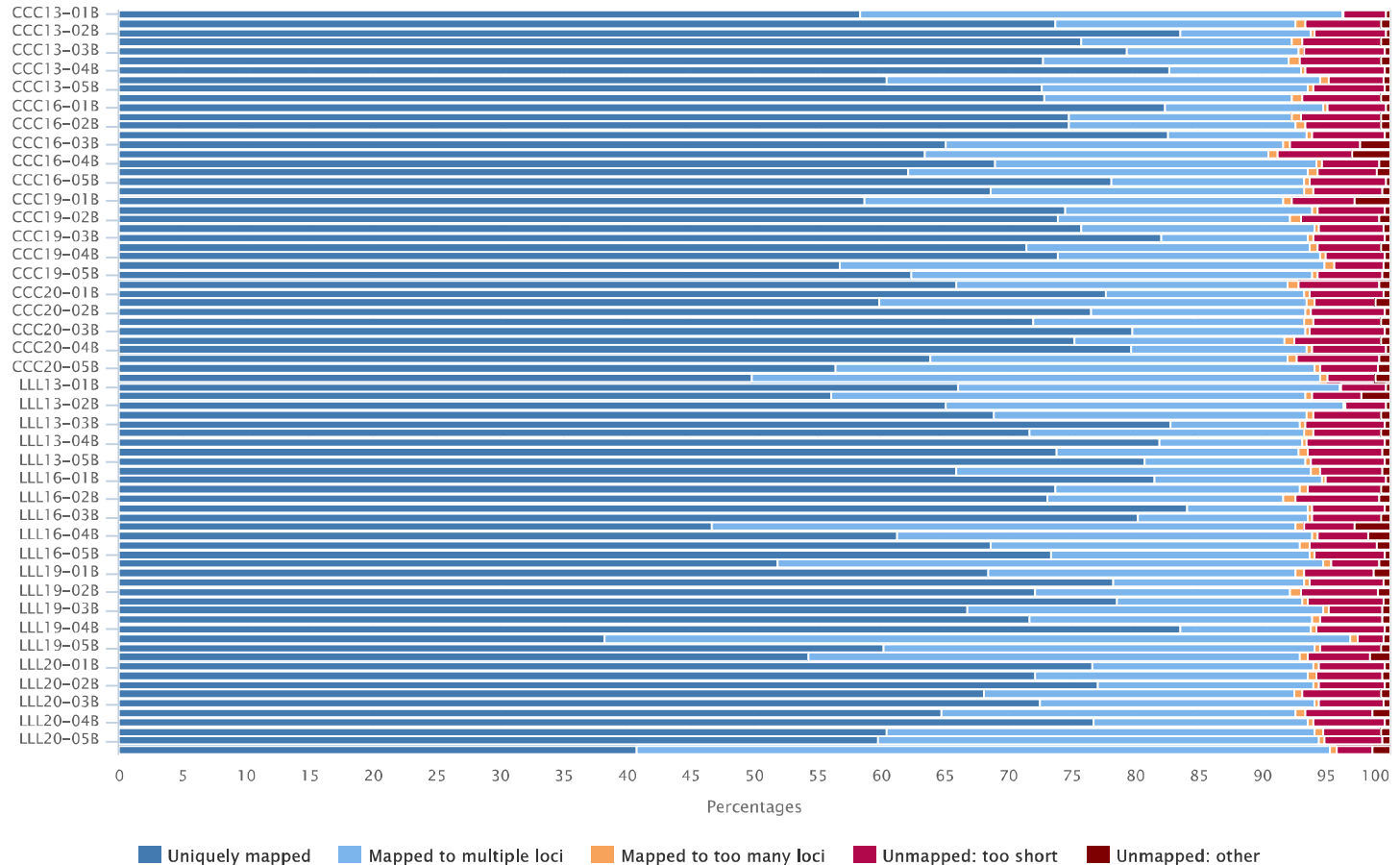
 STAR (final log file), samtools stats, bamtools stats, [QoRTs](#), [RSeQC](#), [Qualimap](#)

Alignment QC | QoRTs



MultiQC can be used to summarise and plot STAR log files.

STAR Alignment Scores



```
samtools stats file.bam
```

```
SN      raw total sequences:    522095280
SN      filtered sequences:    0
SN      sequences:          522095280
SN      is sorted:          1
SN      1st fragments:      261047640
SN      last fragments:    261047640
SN      reads mapped:      514139025
SN      reads mapped and paired: 510035006
SN      reads unmapped:    7956255
SN      reads properly paired: 460249078
SN      reads paired:      522095280
SN      reads duplicated:    60151694
SN      reads MQ0:         54098384
SN      reads QC failed:    0
SN      non-primary alignments: 15023188
SN      total length:      78437013272
SN      bases mapped:      77238941462
SN      bases mapped (cigar): 74139898333
SN      bases trimmed:      0
SN      bases duplicated:    9022025650
SN      mismatches:        1695194781
SN      error rate:        2.286481e-02
SN      average length:    150
SN      maximum length:    151
SN      average quality:    37.6
...

```



```
bamtools stats file.bam
```

```
*****
```

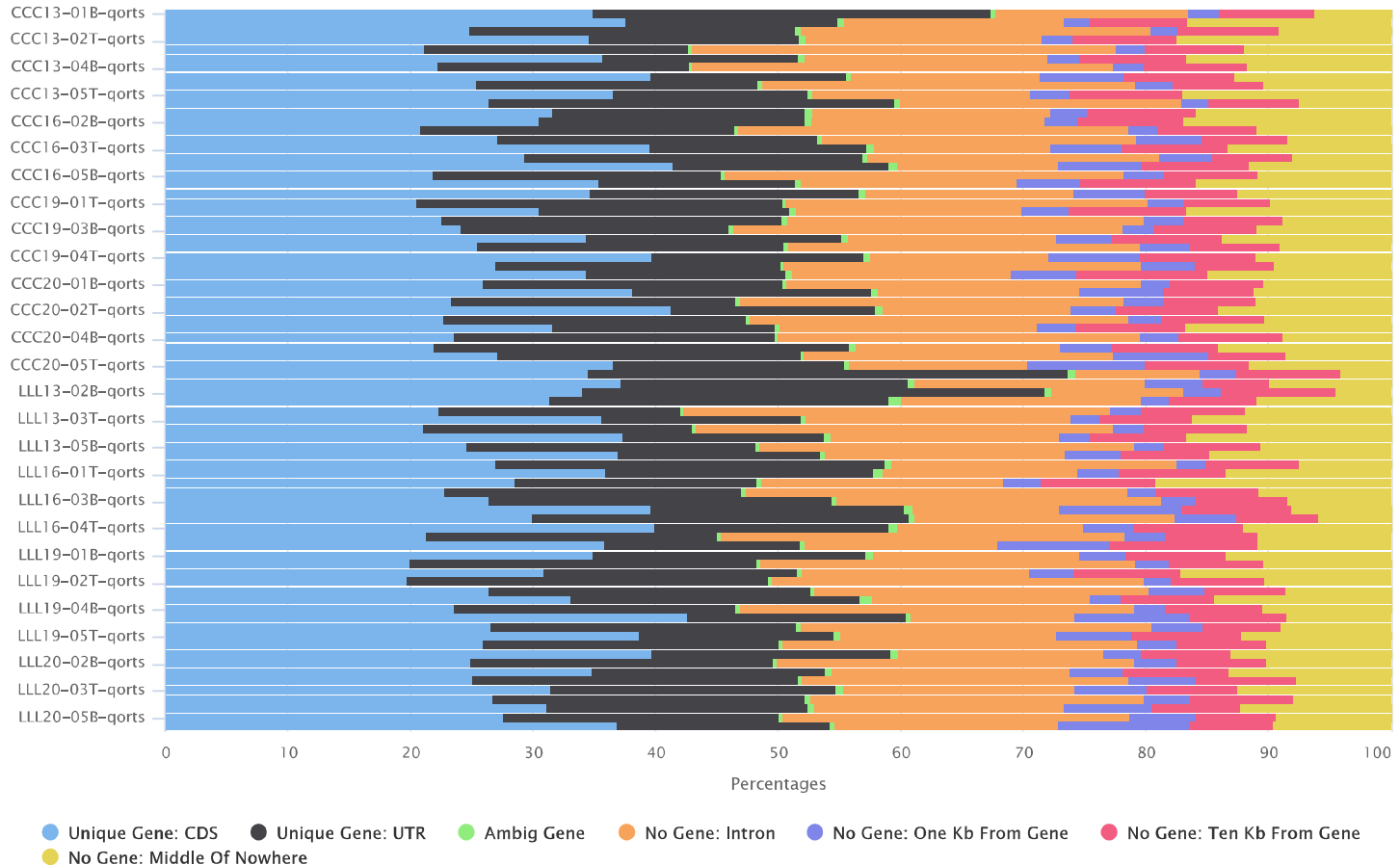
```
Stats for BAM file(s):
```

```
*****
```

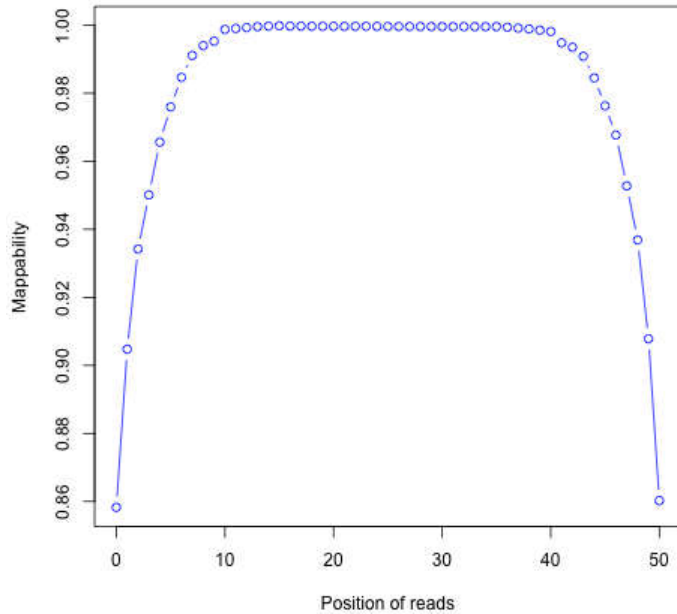
```
Total reads:          537118468
Mapped reads:         529162213   (98.5187%)
Forward strand:      270376825   (50.3384%)
Reverse strand:      266741643   (49.6616%)
Failed QC:           0         (0%)
Duplicates:          61425418   (11.4361%)
Paired-end reads:    537118468   (100%)
'Proper-pairs':      465991264   (86.7576%)
Both pairs mapped:  524501668   (97.651%)
Read 1:              268374707
Read 2:              268743761
Singletons:          4660545    (0.867694%)
```

QoRTs was run on all samples and summarised using MultiQC.

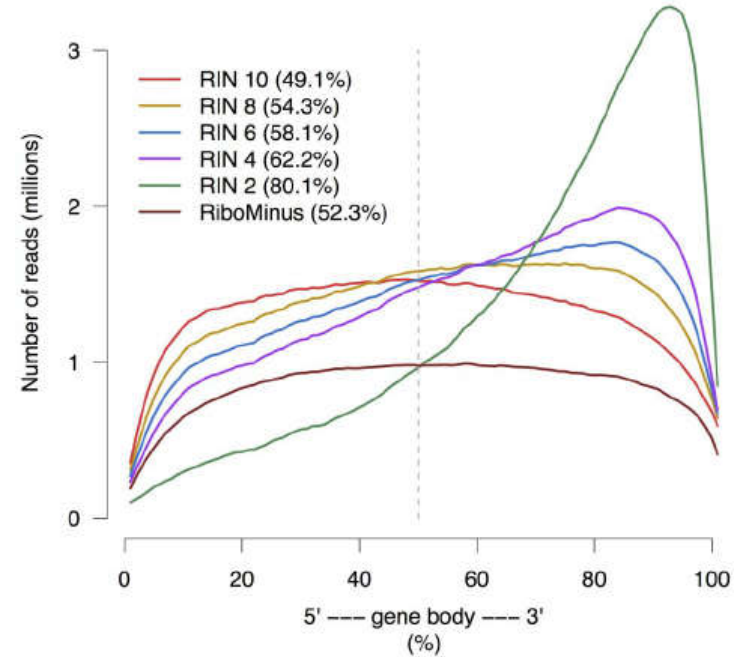
QoRTs: Alignment Locations



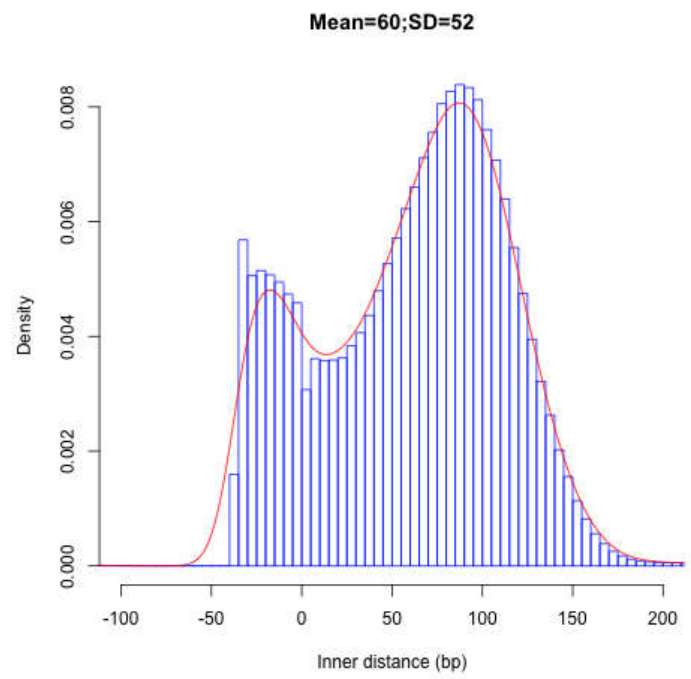
Soft clipping



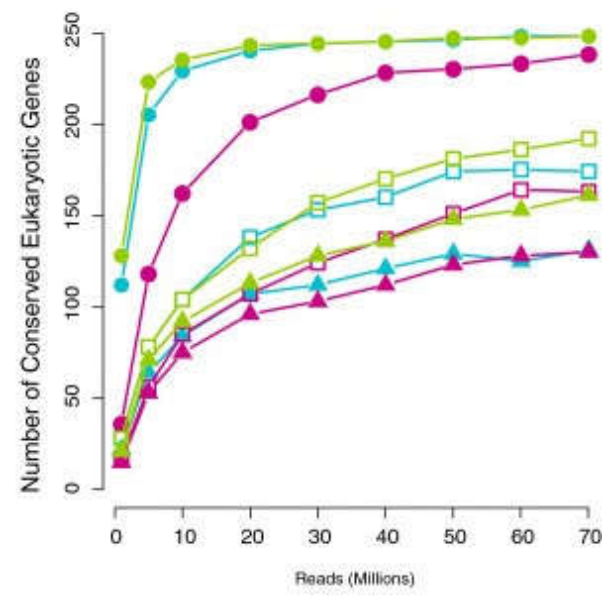
Gene body coverage



Insert size



Saturation curve



MultiQC
v1.6

- General Stats
- featureCounts
- STAR
- Cutadapt
- FastQC
- Sequence Counts
- Sequence Quality Histograms
- Per Sequence Quality Scores
- Per Base Sequence Content
- Per Sequence GC Content
- Per Base N Content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content

MultiQC

A modular tool to aggregate results from bioinformatics analyses across many samples into a single report.

Report generated on 2018-08-04, 01:51 based on data in: `/Users/ewels/GitHub/MultiQC_website/public_html/examples/rna-seq`

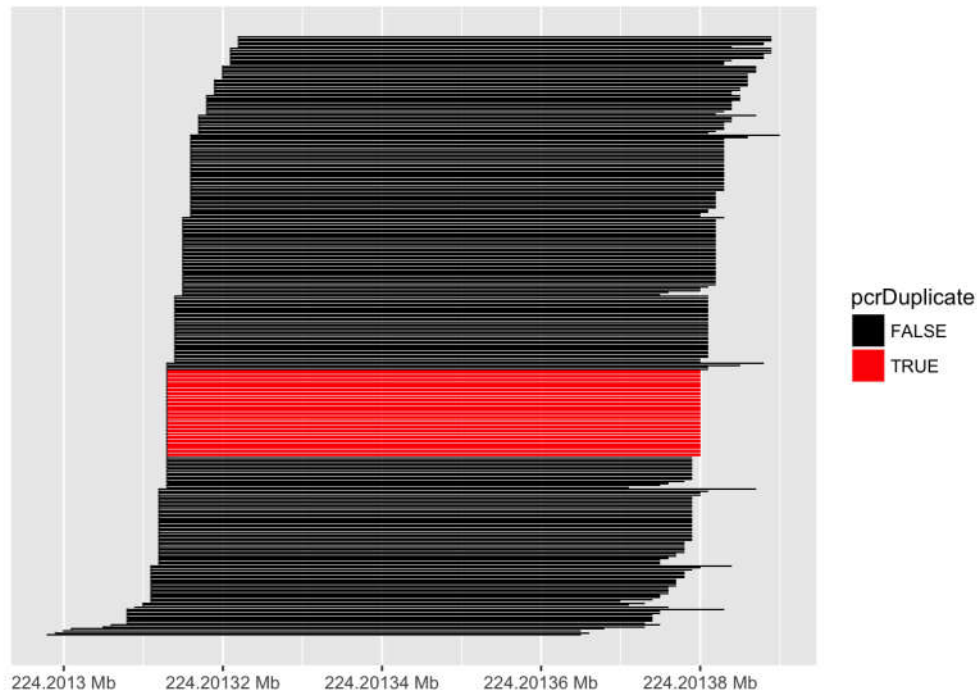
General Statistics

Copy table
Configure Columns
Plot
Showing 8/8 rows and 8/10 columns.

Sample Name	% Assigned	M Assigned	% Aligned	M Aligned	% Trimmed	% Dups	% GC	M Seqs
SRR3192396	67.5%	71.9	93.7%	97.8	4.0%	78.9%	51%	104.4
SRR3192397	66.6%	63.0	94.7%	87.1	3.5%	77.2%	49%	92.0
SRR3192398	50.9%	36.5	88.2%	58.7	5.0%	55.3%	47%	66.6
SRR3192399	52.3%	42.3	88.2%	65.6	5.0%	57.4%	47%	74.3
SRR3192400	70.3%	63.4	77.3%	73.4	7.2%	74.1%	45%	94.9
SRR3192401	71.2%	63.8	76.4%	72.8	6.3%	76.3%	45%	95.2
SRR3192657	73.1%	67.1	91.2%	85.0	3.1%	82.2%	51%	93.1
SRR3192658	71.2%	66.9	89.7%	87.1	3.4%	82.3%	52%	97.1

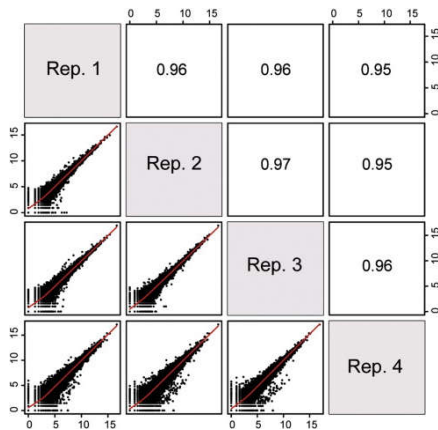
Quantification | PCR duplicates

- Ignore for RNA-Seq data
- Computational deduplication (Don't!)
- Use PCR-free library-prep kits
- Use UMIs

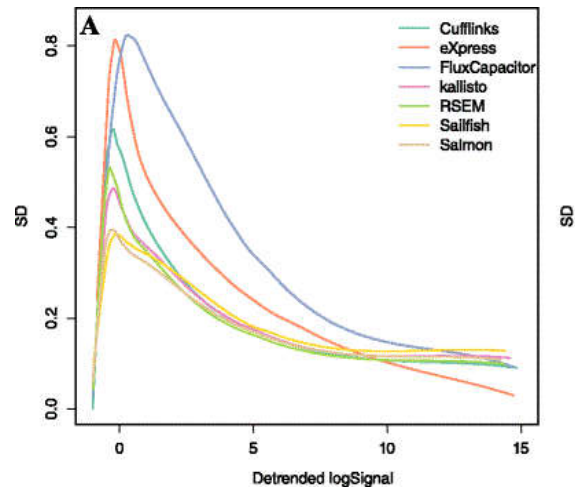


ENSG000000000003	140	242	188	143	287	344	438	280	253
ENSG000000000005	0	0	0	0	0	0	0	0	0
ENSG000000000419	69	98	77	55	52	94	116	79	69
ENSG000000000457	56	75	104	79	157	205	183	178	153
ENSG000000000460	33	27	23	19	27	42	69	44	40
ENSG000000000938	7	38	13	17	35	76	53	37	24
ENSG000000000971	545	878	694	636	647	216	492	798	323
ENSG00000001036	79	154	74	80	128	167	220	147	72

- Pairwise correlation between samples must be high (>0.9)

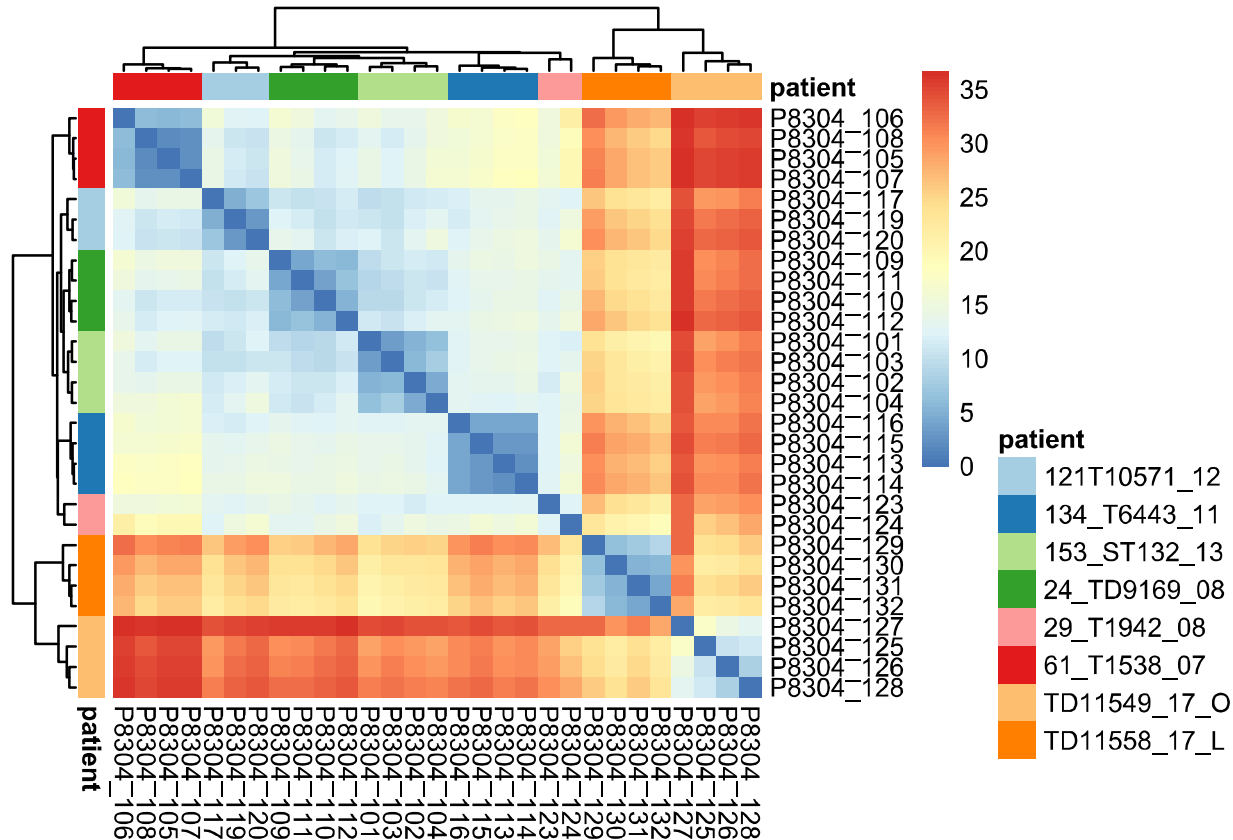


- Count QC using RNASeqComp



Exploratory | Heatmap

- Remove lowly expressed genes
- Transform raw counts to VST, VOOM, RLOG, TPM etc
- Sample-sample distance/correlation clustering heatmap



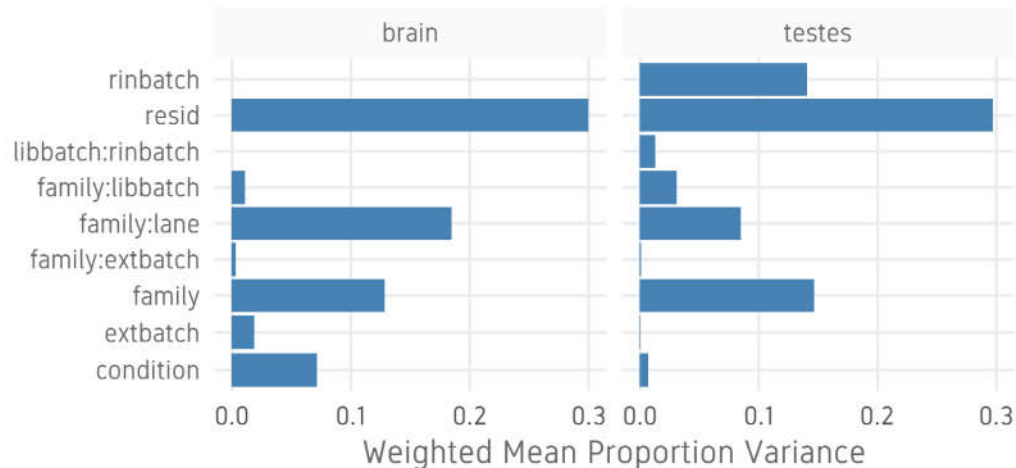
Exploratory | MDS

- 121T10571_12
- 134_T6443_11
- 153_ST132_13
- 24_TD9169_08
- 29_T1942_08
- 61_T1538_07
- TD11549_17_O
- TD11558_17_L

 `cmdscale()`, `plotly`

Batch correction

- Estimate variation explained by variables (PVCA)

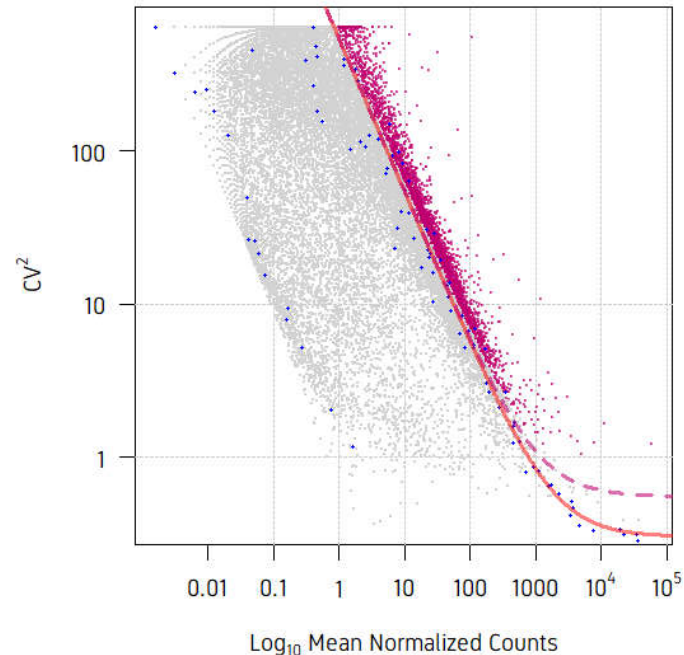


- Find confounding effects as surrogate variables (SVA)
- Model known batches in the LM/GLM model
- Correct known batches (ComBat)(Harsh!)
- Interactively evaluate batch effects and correction (BatchQC)

 SVA, PVCA, BatchQC

Spike-In

- Add synthetic RNA into samples as control
- Usually added before library prep
- Useful for
 - Estimating sensitivity
 - Estimating accuracy
 - Detecting biases
 - Normalisation
 - Absolute quantification
 - Comparing datasets
- ERCC RNA Spike-In Mix/Exiqon
Small RNA Spike-In





- Sound experimental design to avoid confounding
- Plan carefully about lib prep, sequencing etc based on experimental objective
- Biological replicates may be more important than paired-end reads or long reads
- Discard low quality bases, reads, genes and samples
- Verify that tools and methods align with data assumptions
- Experiment with multiple pipelines and tools
- QC! QC everything at every step

 Conesa, Ana, *et al.* "A survey of best practices for RNA-seq data analysis." [Genome biology 17.1 \(2016\): 13](#)



Thank you! Questions?

Built on :  23-Oct-2018 at  18:16:04

2018 Roy Francis | [SciLifeLab](#) | [NBIS](#)