NBIS

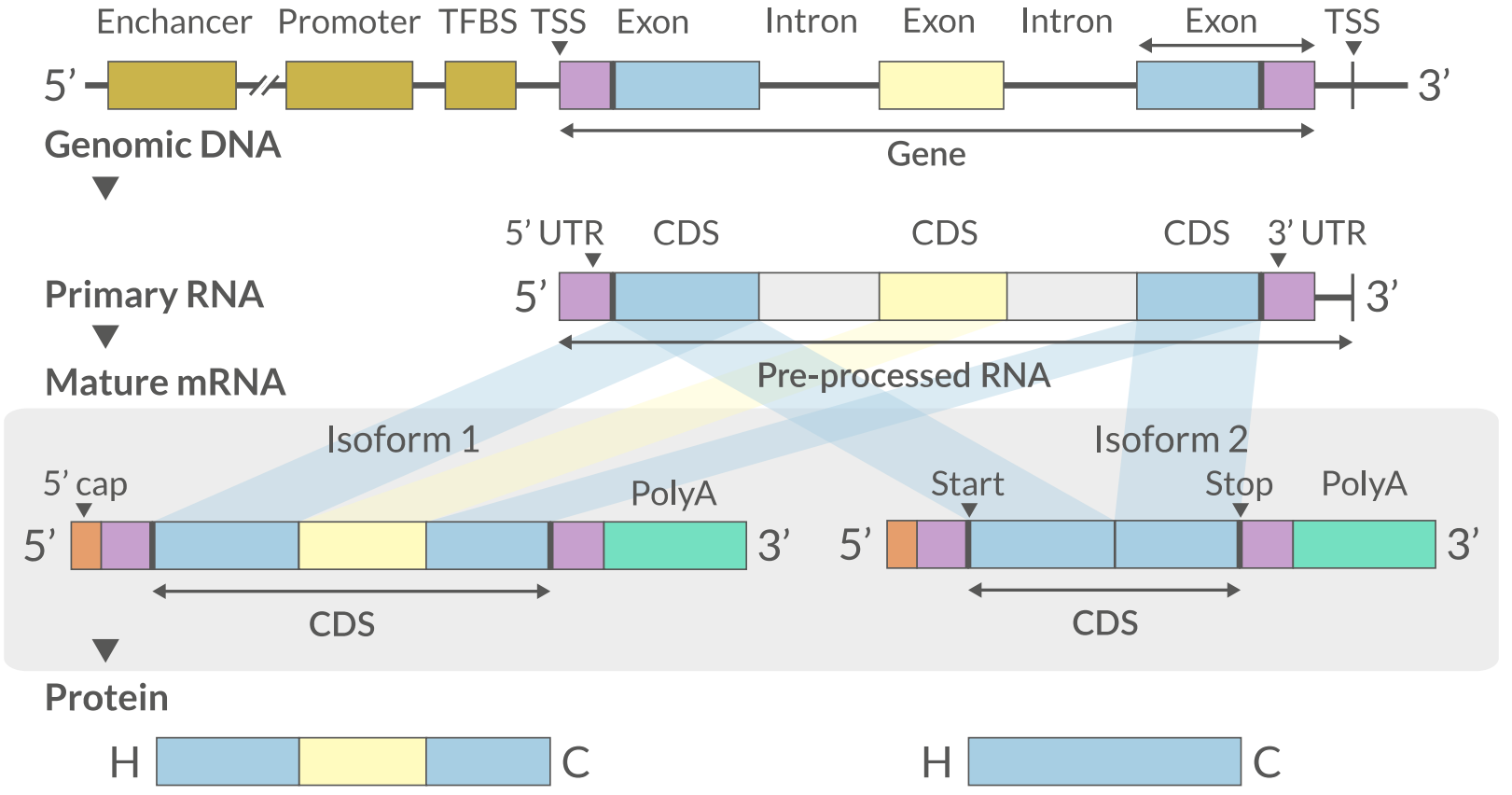
Analysis of bulk RNA-Seq data

Introduction To Bioinformatics Using NGS Data

26-Nov-2021

Contents

What is RNA?



- The transcriptome is spatially and temporally dynamic
- Data comes from functional units (coding regions)
- Only a tiny fraction of the genome

Applications

- Identify gene sequences in genomes (annotation)
- Learn about gene function
- Differential gene expression
- Explore isoform and allelic expression
- Understand co-expression, pathways and networks
- Gene fusion
- RNA editing

Workflow

Experimental design

▼

RNA extraction

▼

Library preparation

▼

Sequencing

▼

Data processing

▼

Data analysis

QC

▼

Mapping

▼

Quantification

De-novo assembly

Annotation

Correlation

▼

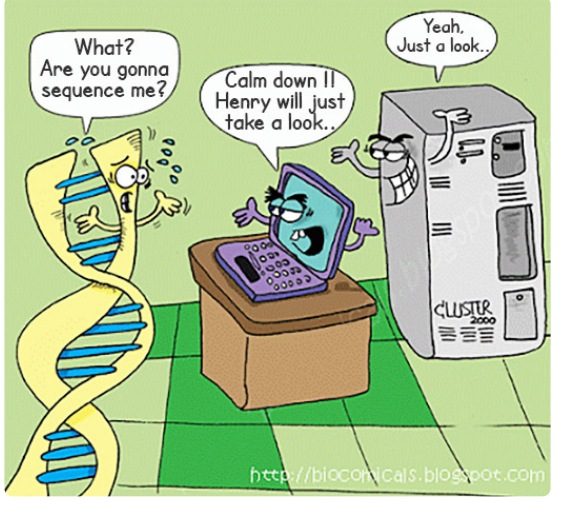
Clustering

▼

MDS/PCA

▼

Differential Gene Expression

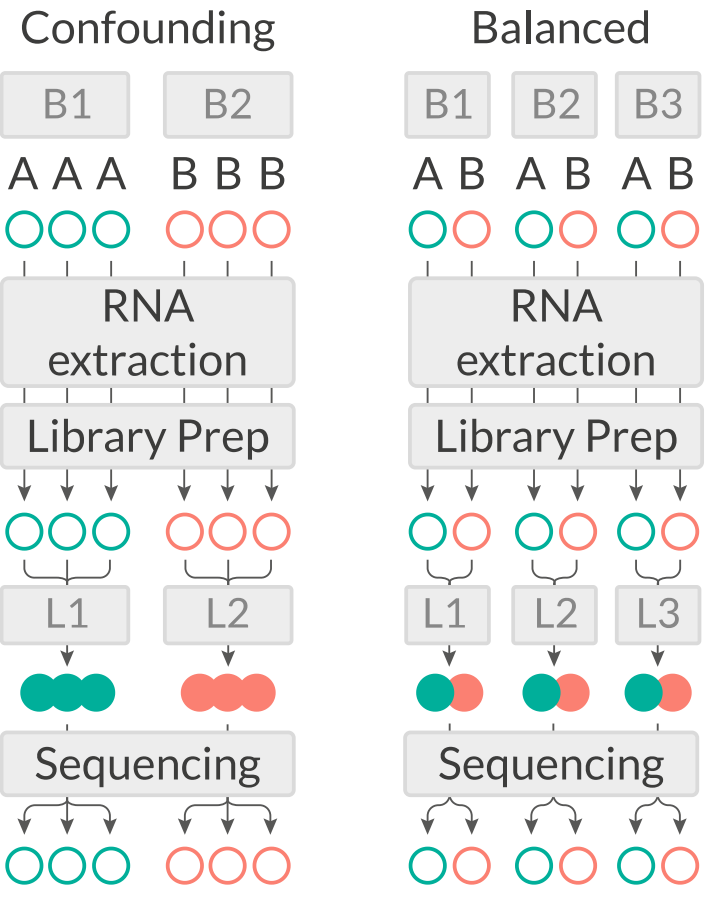


Conesa et. al, 2016

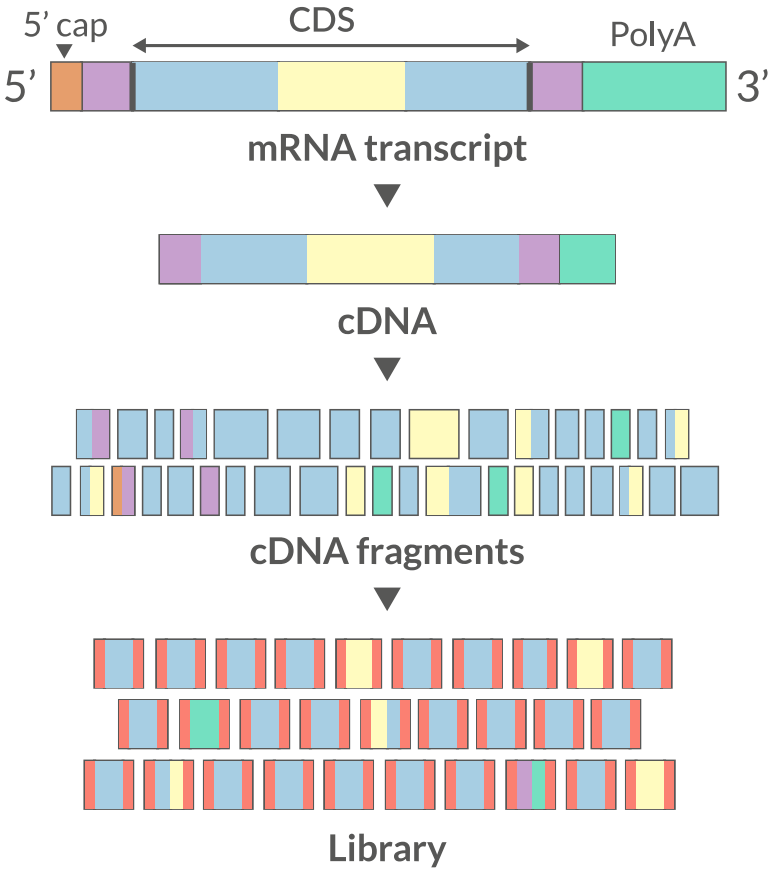
Experimental design

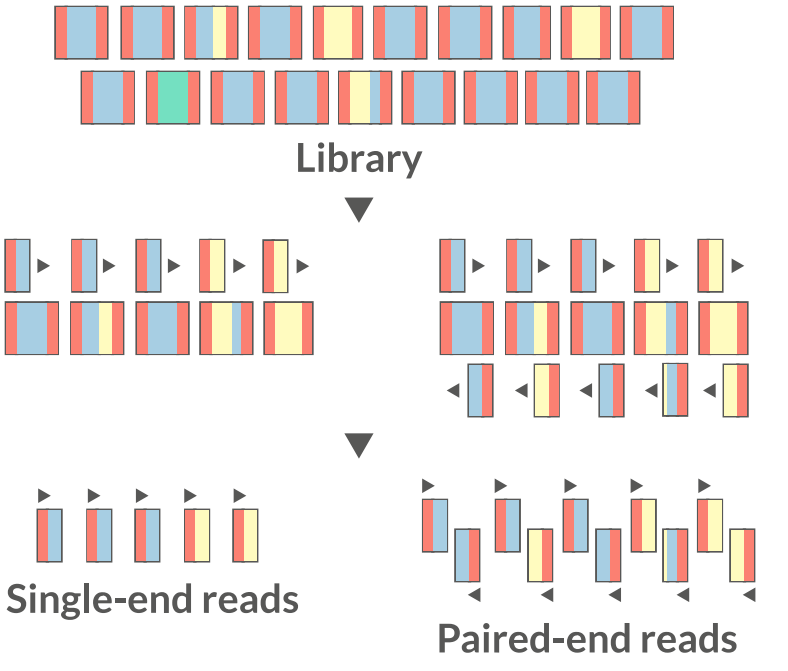
- Biological replicates: 6 - 12 Schurch et al, 2016
- Sample size estimation Hart et al, 2013
- Power analysis RNASeqPower RNASeqPower web app
- Balanced design to avoid batch effects

  experDesign DeclareDesign
- RIN values have strong effect Romero et al, 2014



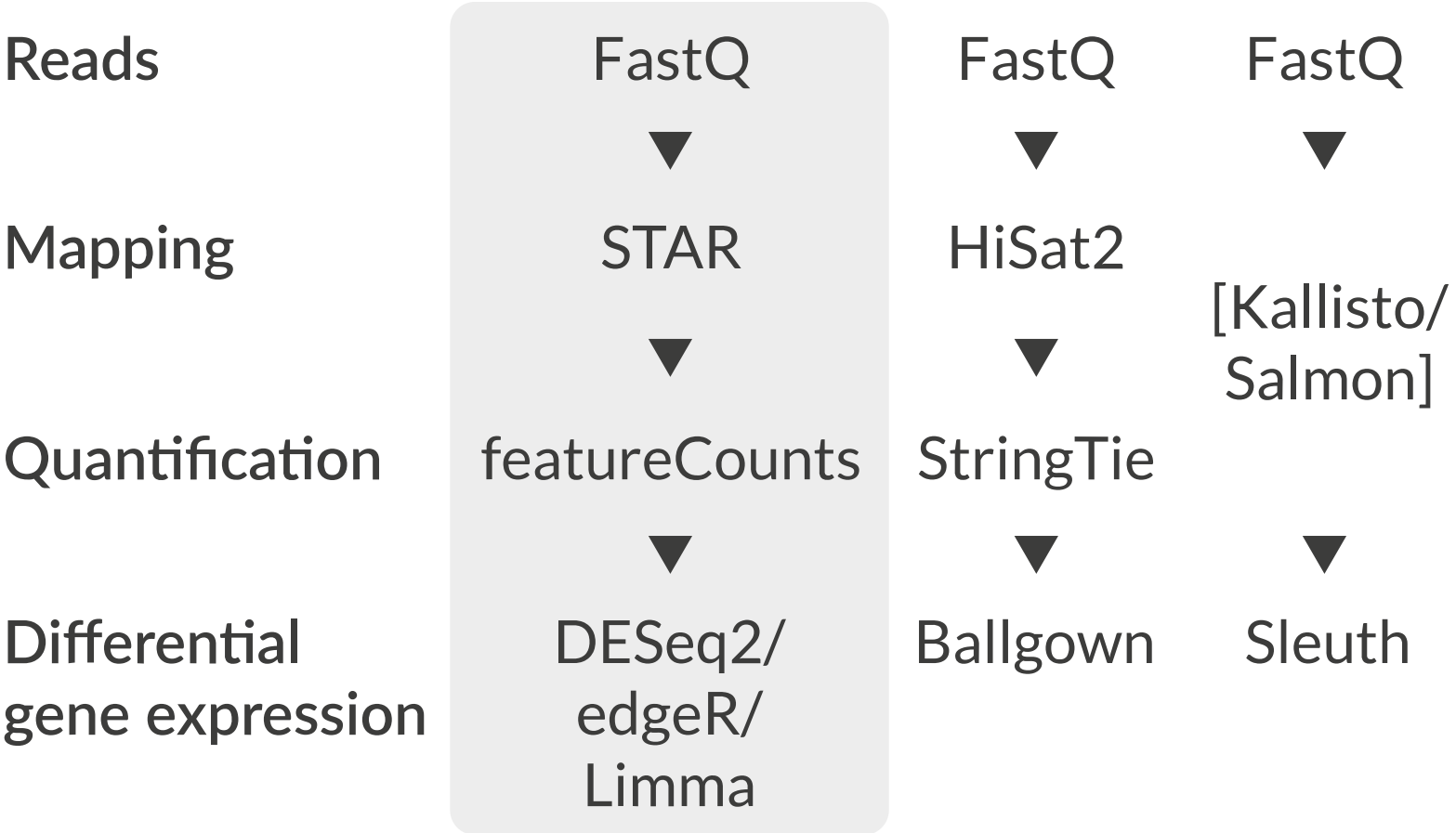
Library & Sequencing





- polyA selection/Ribosomal RNA depletion
- single-end/Paired-end

Workflow • DGE



Read QC

- Number of reads
- Per base sequence quality
- Per sequence quality score
- Per base sequence 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/

FastQC

Good quality

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

Poor quality

FastQC Report — Thu 21 Dec 2017 — bad_sequence.txt

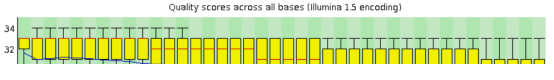
Summary

- 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

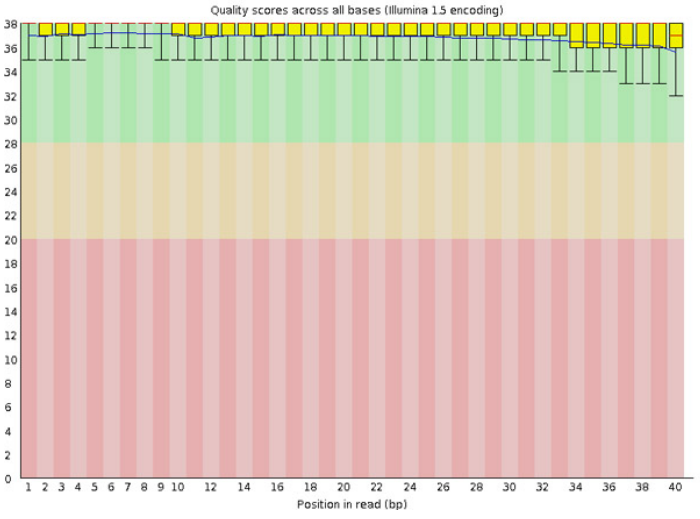
Measure	Value
Filename	bad_sequence.txt
File type	Conventional base calls
Encoding	Illumina 1.5
Total Sequences	395288
Sequences flagged as poor quality	0
Sequence length	40
%GC	47

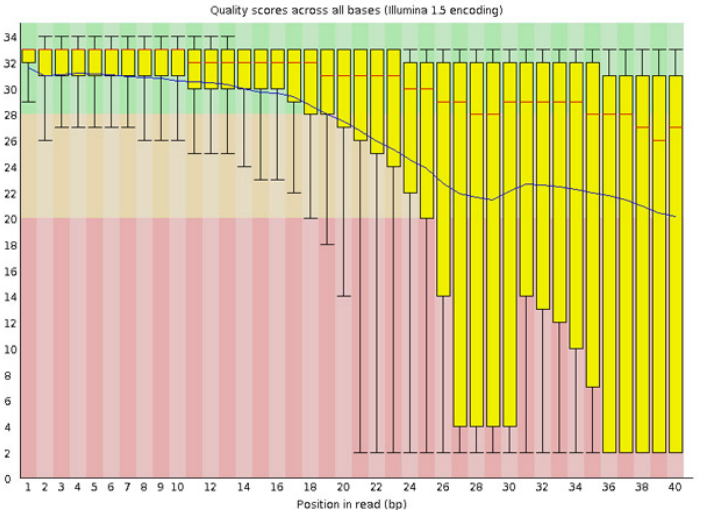
Per base sequence quality



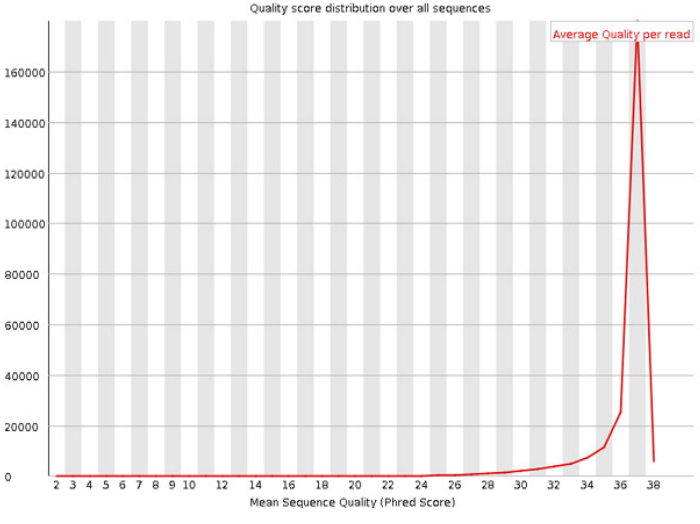
Read QC • PBSQ, PSQS

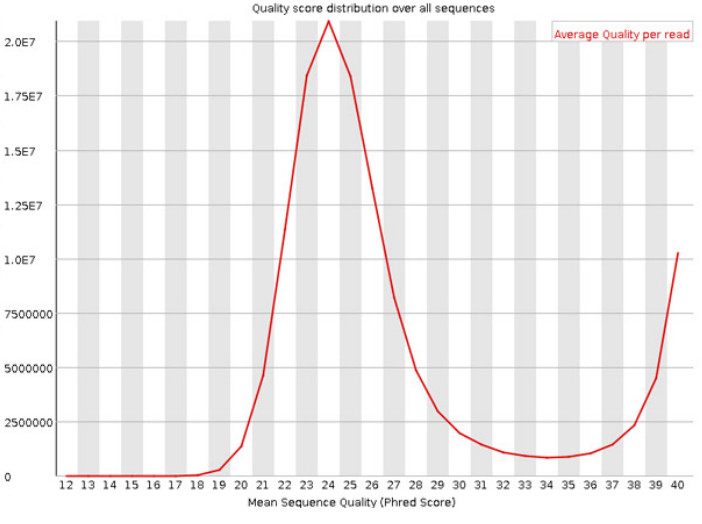
Per base sequence quality





Per sequence quality scores





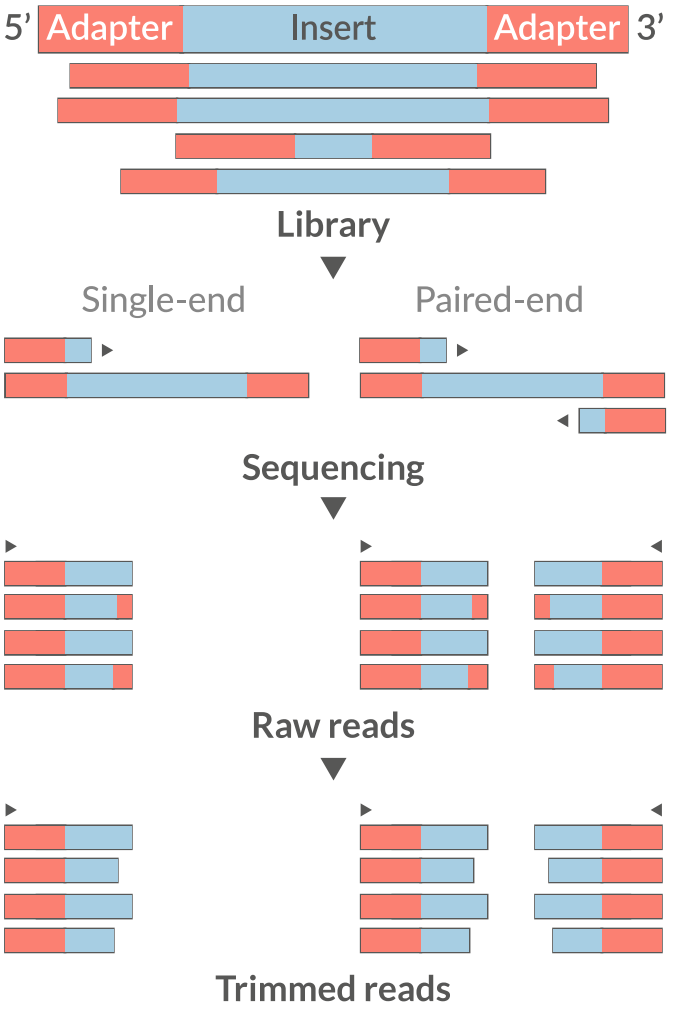
Trimming

- Trimming reads to remove adapter/readthrough or low quality bases
- Related options are hard clipping, filtering reads
- Sliding window trimming
- Filter by min/max read length
    - Remove reads less than ~18nt
- Demultiplexing/Splitting

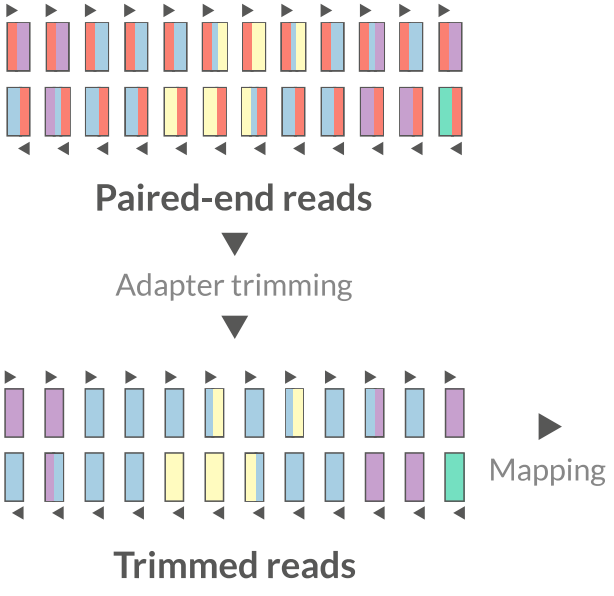
When to avoid trimming?

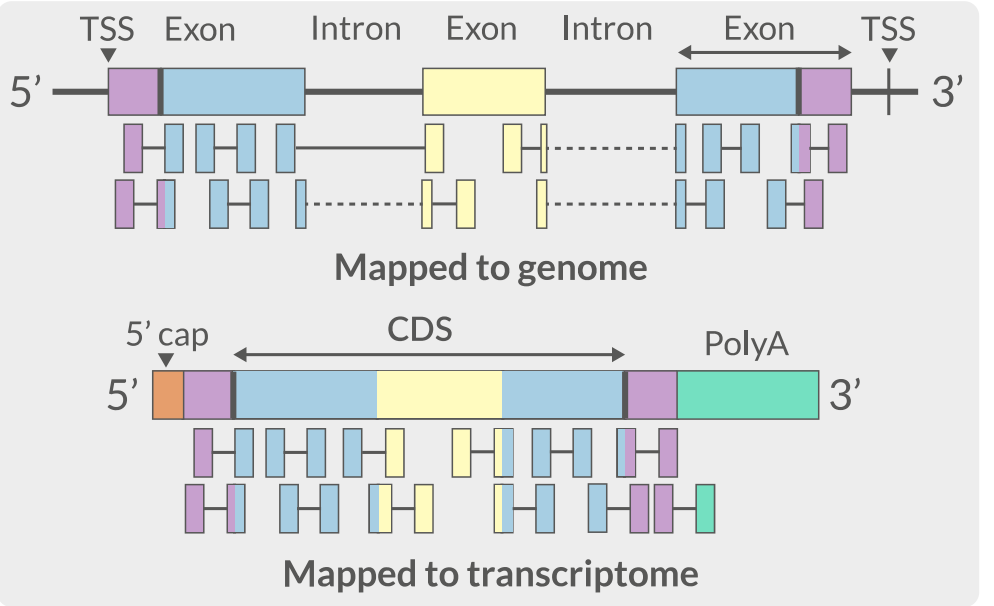
- Read trimming may not always be necessary Liao et al, 2020
- Fixed read length may sometimes be more important
- Expected insert size distribution may be more important for assemblers

Cutadapt fastp Prinseq



Mapping

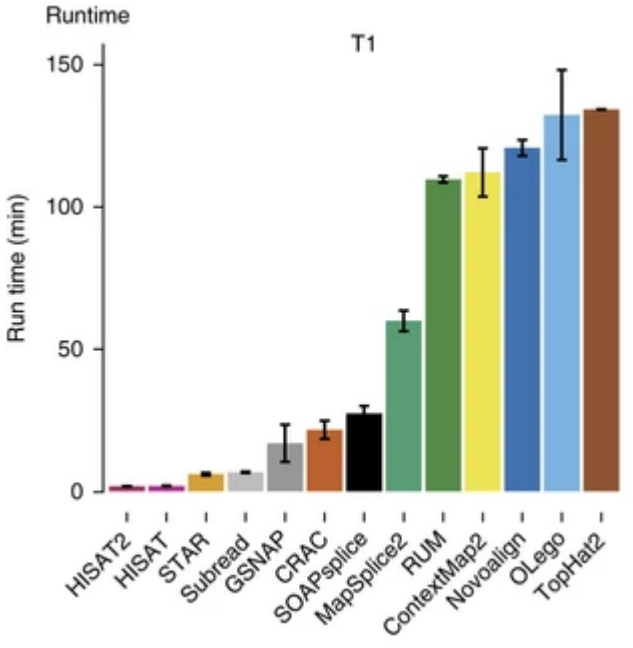




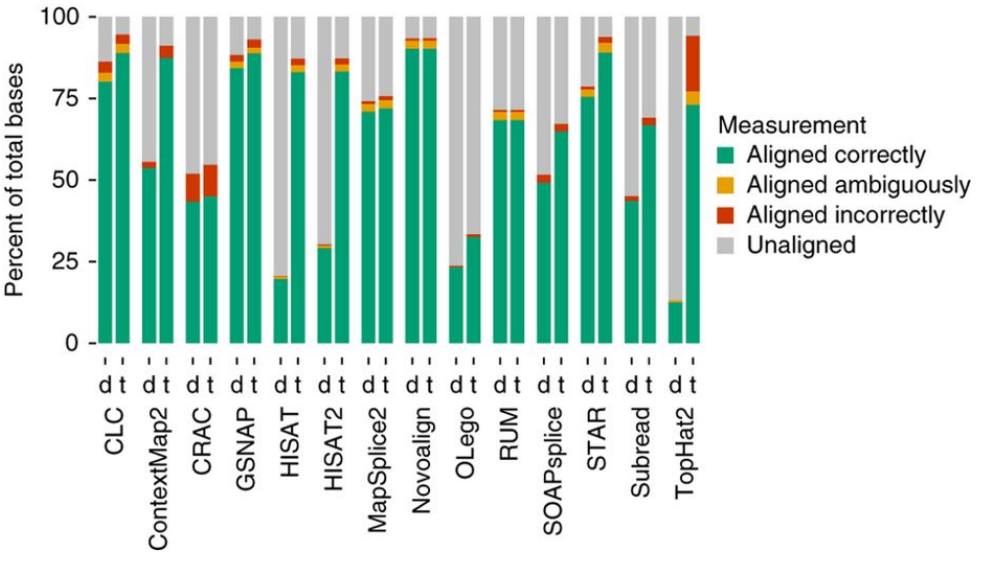
- Aligning reads back to a reference sequence
- Mapping to genome vs transcriptome
- Splice-aware alignment (genome) (STAR, HISAT2 etc)

STAR HiSat2 Baruzzo et al, 2017

Aligners • Metrics



Baruzzo et al, 2017



Increasing Accuracy

- Novel variants / RNA editing
- Allele-specific expression
- Genome annotation
- Gene and transcript discovery
- Differential expression

Mapping

- Reads (FASTQ)

```
@ST-E00274:179:HHYMLALXX:8:1101:1641:1309 1:N:0:NGATGT
NCATCGTGGTATTTGCACATCTTTTCTTATCAAATAAAAAGTTTAACCTACTCAGTTATGCGCATACGTTTTTTGATGGCATTTCCATAAA
+
#AAAFAFA<-AFFJJJAFA-FFJJJJFFFAJJJJ-<FFJJJ-A-F-7--FA7F7-----FFFJFA<FFFFJ<AJ--FF-A<A-<JJ-7-7-
```

```
@instrument:runid:flowcellid:lane:tile:xpos:ypos read:isfiltered:controlnumber:sampleid
```

- Reference Genome/Transcriptome (FASTA)

```
>1 dna:chromosome chromosome:GRCz10:1:1:58871917:1 REF
GATCTTAAACATTTATTCCCCCTGCAAACATTTTCAATCATTACATTGTCATTTCCCCTC
CAAATTAAATTTAGCCAGAGGCGCACAACATACGACCTCTAAAAAAGGTGCTGTAACATG
```

- Annotation (GTF/GFF)

```
#!genome-build GRCz10
#!genebuild-last-updated 2016-11
4       ensembl_havana  gene    6732    52059   .       -       .       gene_id "ENSDARG000
```

```
seq source feature start end score strand frame attribute
```

Illumina FASTQ format GTF format

Alignment

- SAM/BAM (Sequence Alignment Map format)

```
ST-E00274:188:H3JWNCCXY:4:1102:32431:49900      163     1       1       60      8S139M4S
```

```
query flag ref pos mapq cigar mrnm mpos tlen seq qual opt
```

Never store alignment files in raw SAM format. Always compress it!

Format	Size_GB
SAM	7.4
BAM	1.9
CRAM lossless Q	1.4
CRAM 8 bins Q	0.8
CRAM no Q	0.26

SAM format

Visualisation • IGV

IGV UCSC Genome Browser SeqMonk

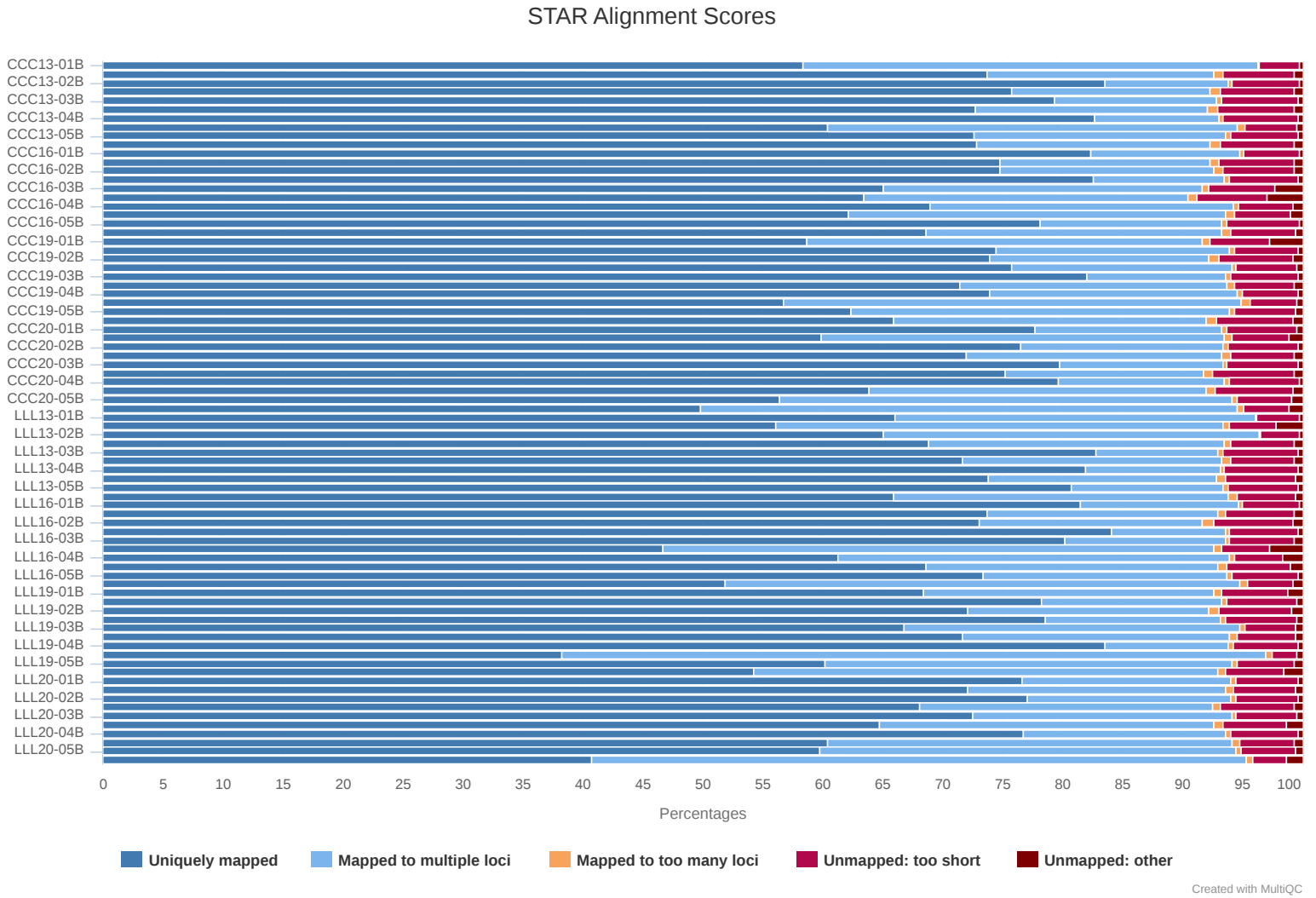
Alignment QC

- 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

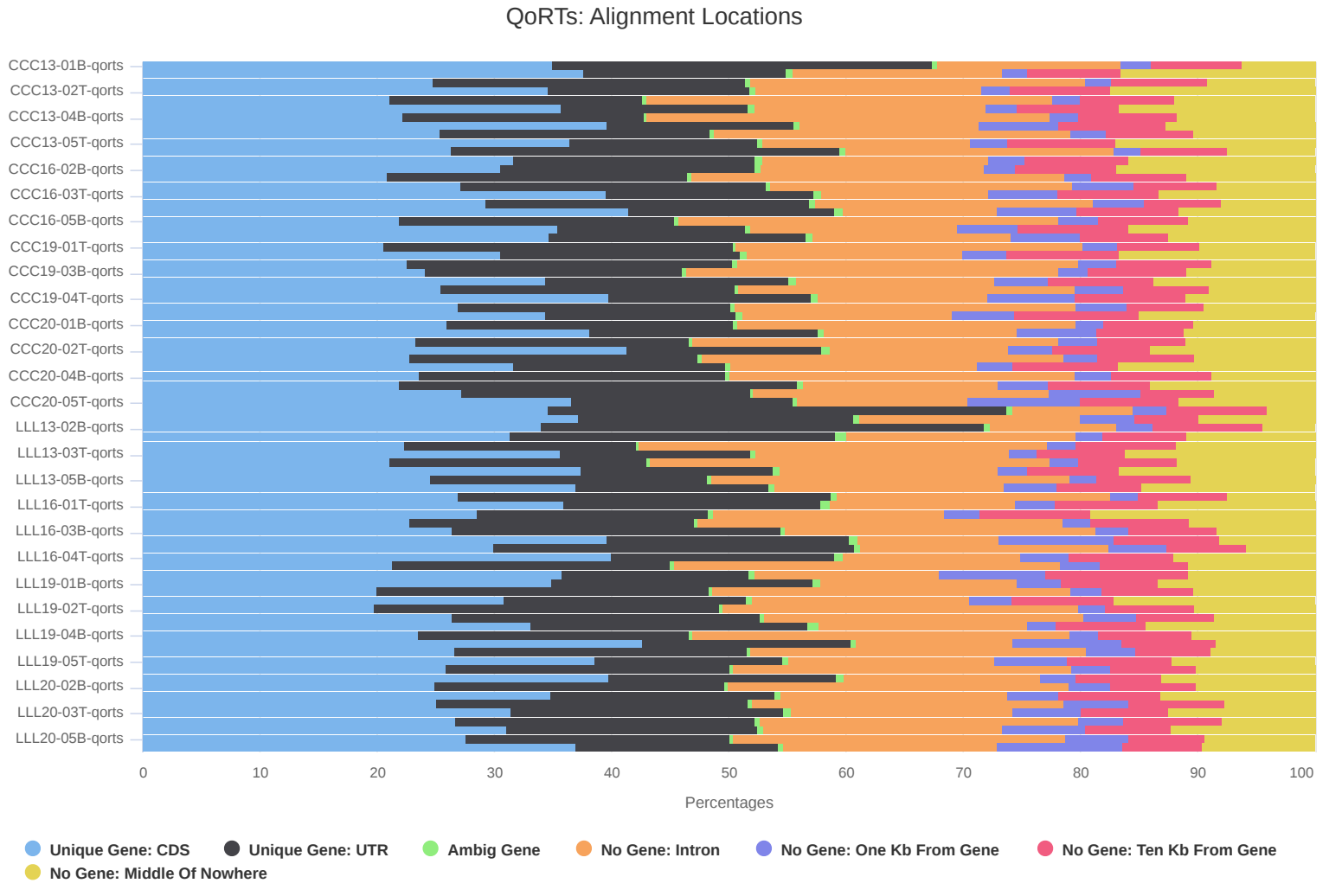
Alignment QC • STAR Log

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

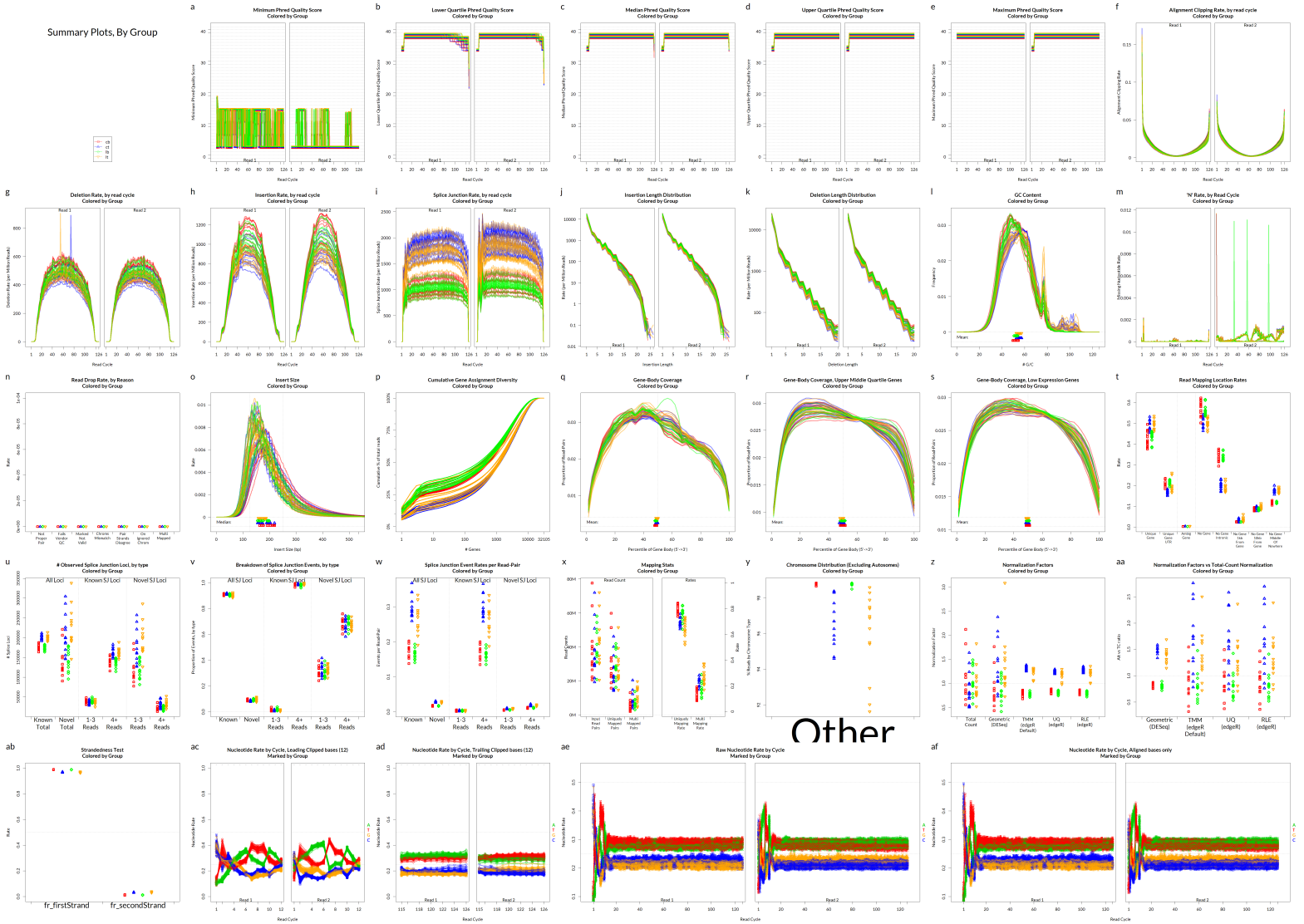


Alignment QC • Features

QoRTs was run on all samples and summarised using MultiQC.

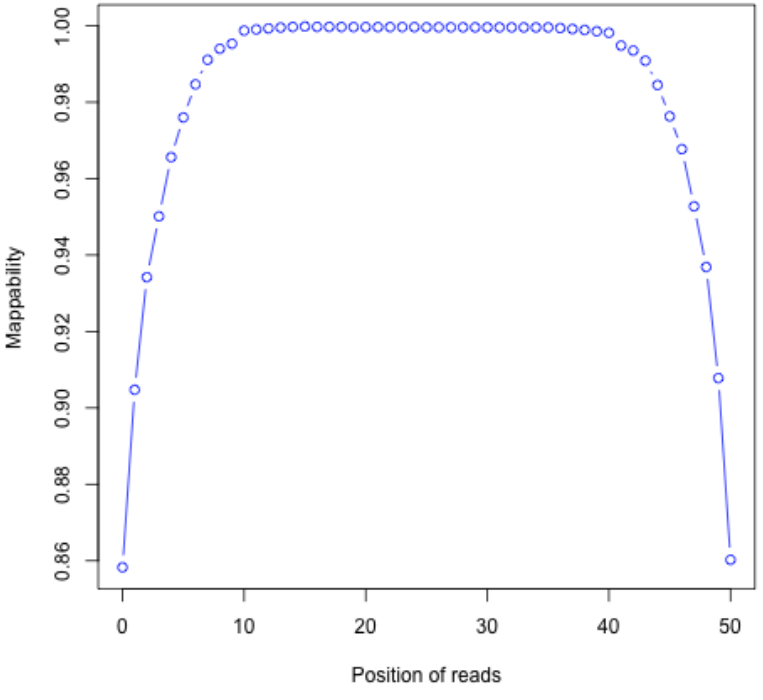


Alignment QC • QoRTs

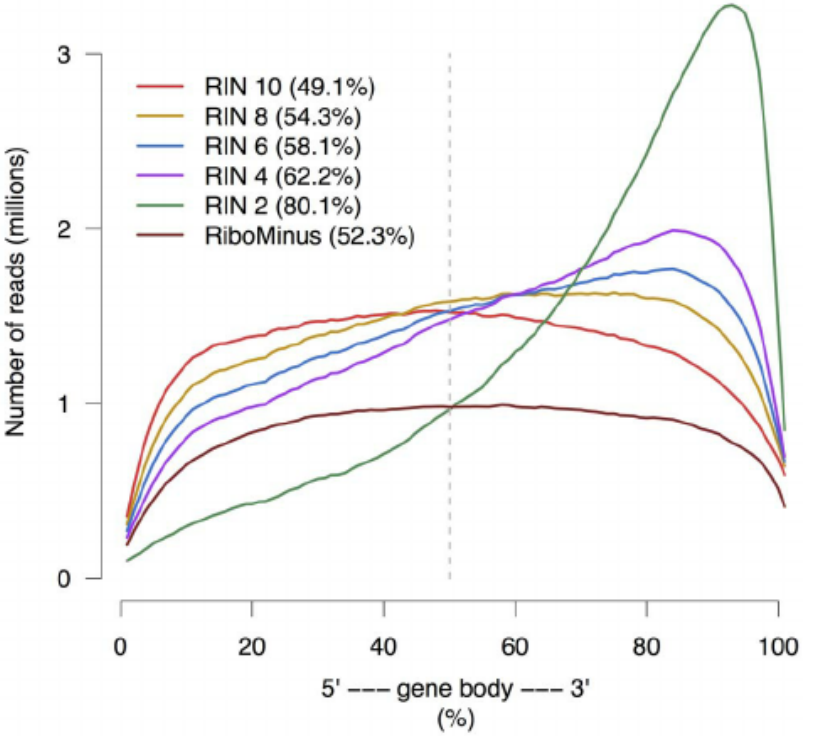


Alignment QC • Examples

Read mapping profile



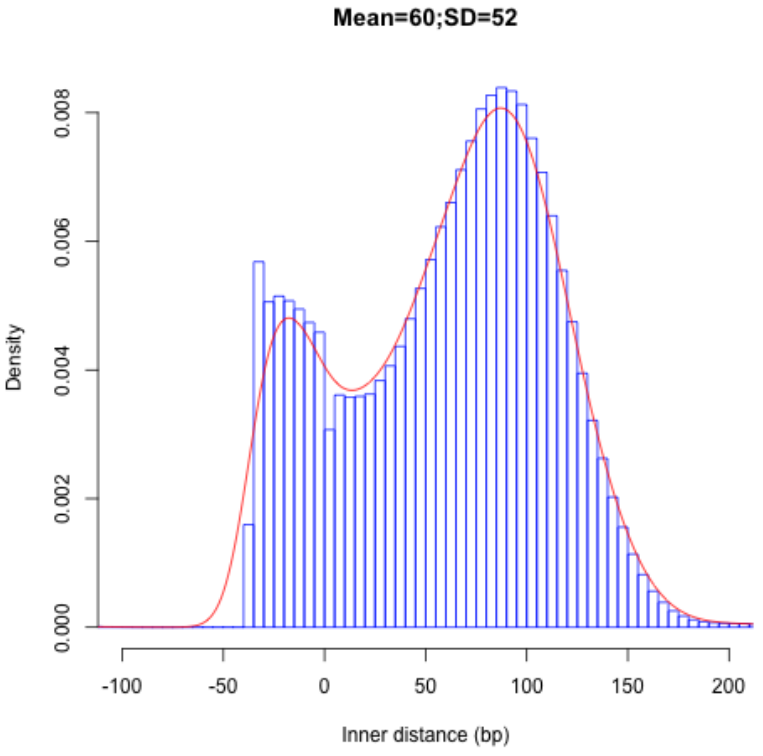
Gene body coverage



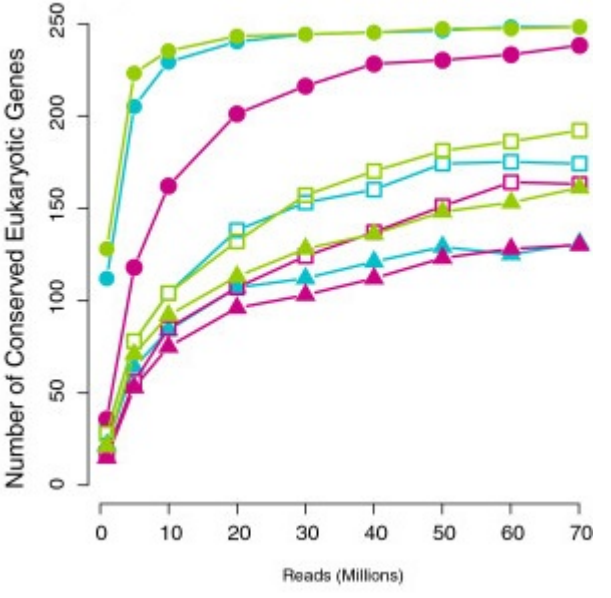
Sigurgeirsson et al, 2014

Alignment QC • Examples

Insert size



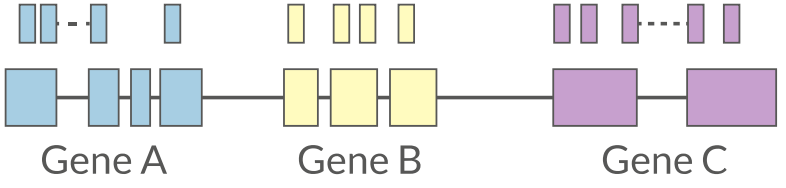
Saturation curve



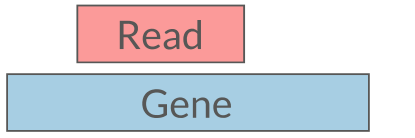
Francis et al, 2013

Quantification • Counts

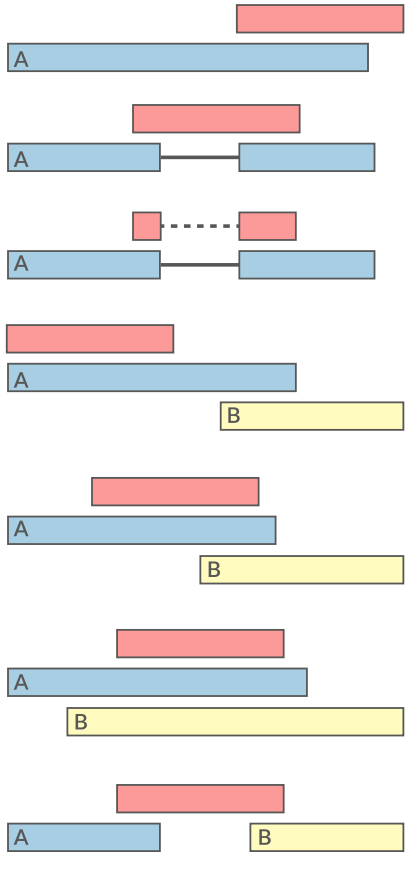
- Read counts = gene expression
- Reads can be quantified on any feature (gene, transcript, exon etc)
- Intersection on gene models
- Gene/Transcript level



featureCounts HTSeq



Union



Gene A

Gene A

Gene A

Gene A

Ambiguous

Ambiguous

Ambiguous

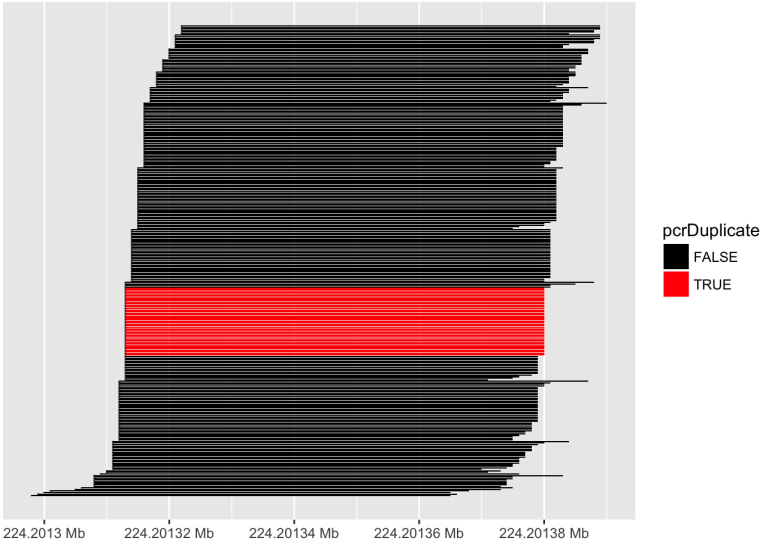
Quantification

PCR duplicates

- Computational deduplication not recommended Klepikova et al, 2017 Parekh et al, 2016
- Use PCR-free library-prep kits
- Use UMIs during library-prep Fu et al, 2018

Multi-mapping

- Added (BEDTools multicov)
- Discard (featureCounts, HTSeq)
- Distribute counts (Cufflinks, featureCounts)
- Rescue
    - Probabilistic assignment (Rcount, Cufflinks)
    - Prioritise features (Rcount)
    - Probabilistic assignment with EM (RSEM)



Quantification • Abundance

- Count methods
  - Provide no inference on isoforms
  - Cannot accurately measure fold change
- Probabilistic assignment
  - Deconvolute ambiguous mappings
  - Transcript-level
  - cDNA reference

Kallisto, Salmon

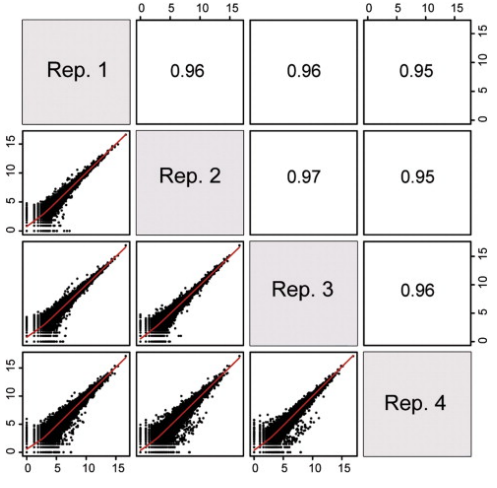
- Ultra-fast & alignment-free
- Bootstrapping & quantification confidence
- Transcript-level counts
- Transcript-level estimates improves gene-level estimates Soneson et al, 2015 tximport
- Evaluation and comparison of isoform quantification tools Zhang et al, 2017

RSEM Kallisto Salmon

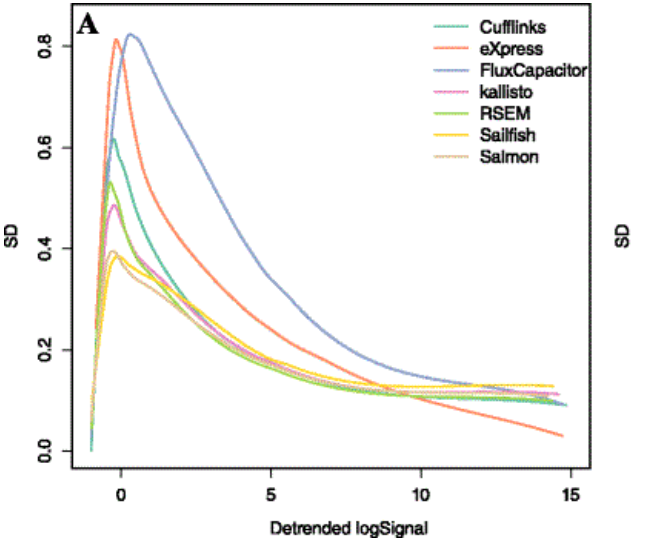
Quantification QC

ENSG00000000003	140	242	188	143	287	344	438	280	253
ENSG00000000005	0	0	0	0	0	0	0	0	0
ENSG00000000419	69	98	77	55	52	94	116	79	69
ENSG00000000457	56	75	104	79	157	205	183	178	153
ENSG00000000460	33	27	23	19	27	42	69	44	40
ENSG00000000938	7	38	13	17	35	76	53	37	24
ENSG00000000971	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)

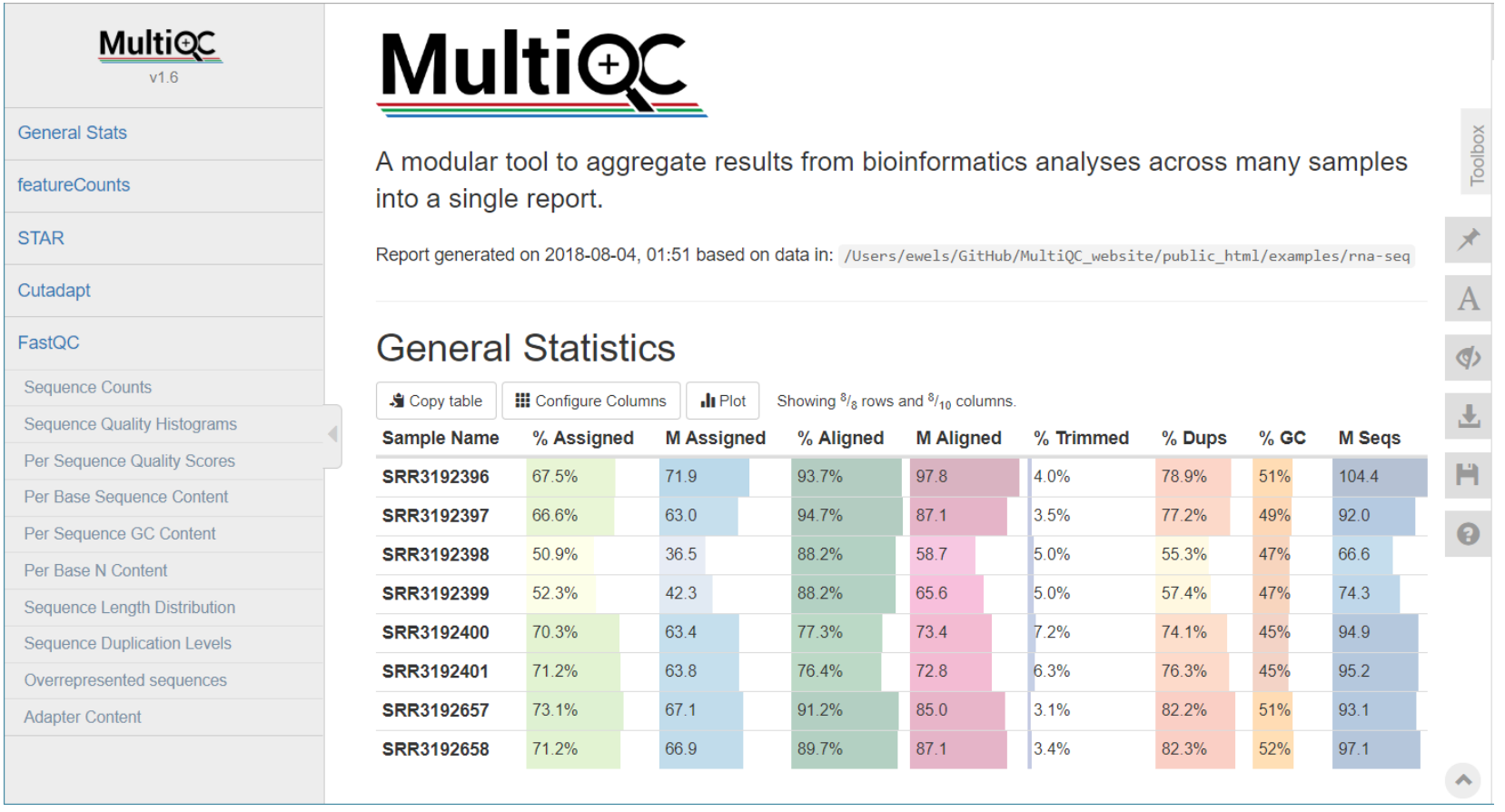


- Count QC using RNASeqComp



RNASeqComp | Teng et al, 2016

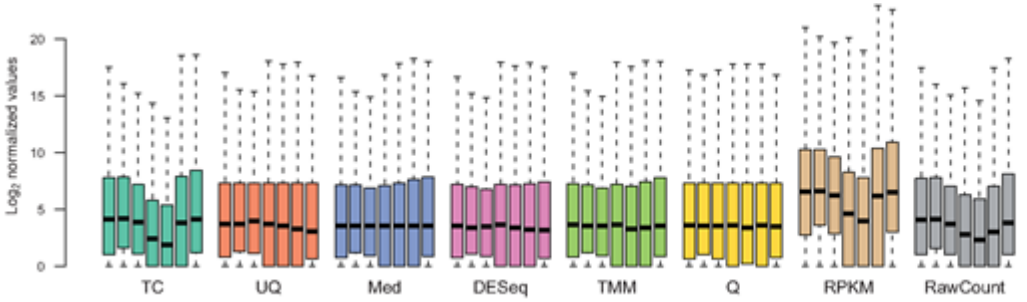
MultiQC

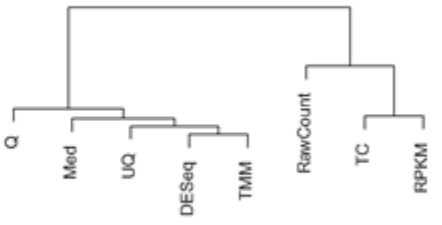


Sample Name	% Assigned	M Assigned	% Aligned	M Aligned	% Trimmed	% Dups	% GC	M Seqs
SRR3192396	67.5%	71.9	93.7%	97.8	4.0%	78.9%	51%	104.4
SRR3192397	66.6%	63.0	94.7%	87.1	3.5%	77.2%	49%	92.0
SRR3192398	50.9%	36.5	88.2%	58.7	5.0%	55.3%	47%	66.6
SRR3192399	52.3%	42.3	88.2%	65.6	5.0%	57.4%	47%	74.3
SRR3192400	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

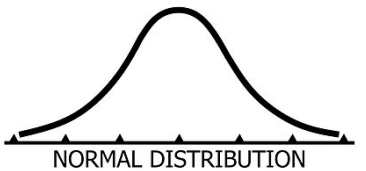
Normalisation

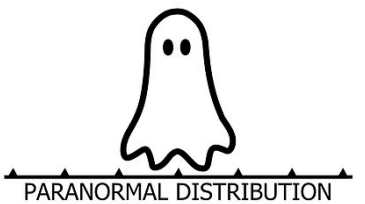
- Control for Sequencing depth, compositional bias and more
- Median of Ratios (DESeq2) and TMM (edgeR) perform the best





- For DGE using DGE packages, use raw counts
- For clustering, heatmaps etc use VST, VOOM or RLOG
- For own analysis, plots etc, use TPM
- Other solutions: spike-ins/house-keeping genes

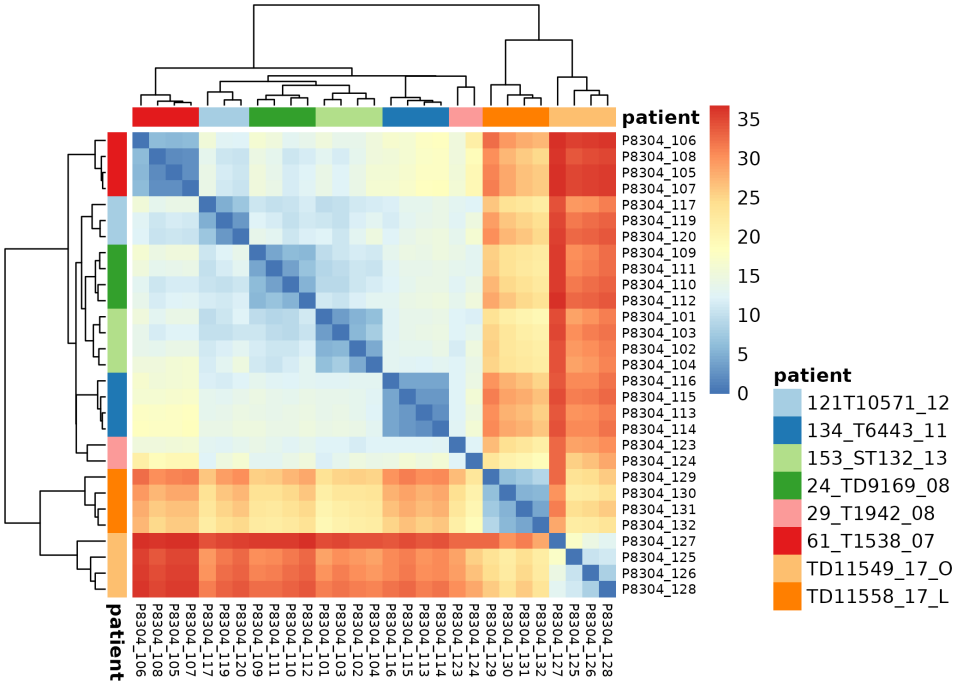


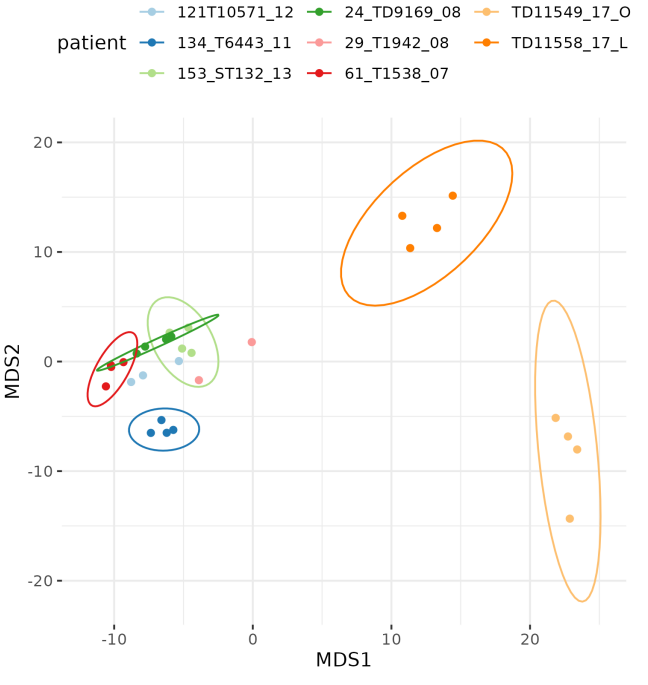


Dillies et al, 2013 | Evans et al, 2017 | Wagner et al, 2012

Exploratory

- Remove lowly expressed genes
- Transform raw counts to VST, VOOM, RLOG, TPM etc
- Heatmaps, MDS, PCA etc.

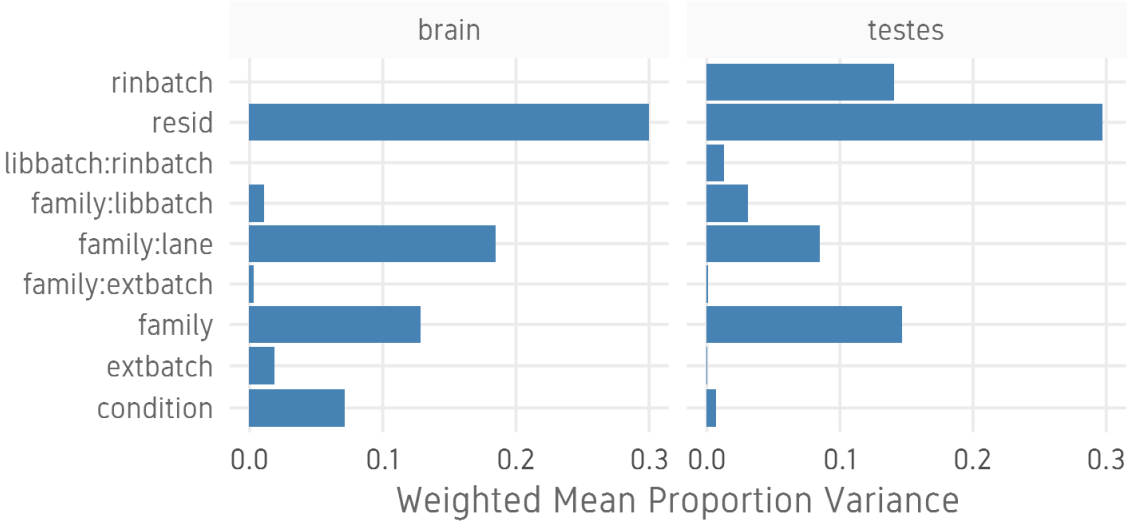




pheatmap

Batch correction

- Estimate variation explained by variables (PVCA) PVCA



- Find confounding effects as surrogate variables (SVA) SVA
- Model known batches in the LM/GLM model
- Correct known batches (ComBat from SVA)(Only if you are desperate! Zindler et al, 2020)
- Interactively evaluate batch effects and correction (BatchQC) Manimaran et al, 2016 BatchQC

Differential expression

		Gene A	Gene B	...	Gene N
Group 1	Sample 1	12	54	...	...
	Sample 2	8	47	...	...
	Sample 3	13	48	...	...
Group 2	Sample 1	22	50	...	...
	Sample 2	18	48	...	...
	Sample 3	25	41	...	...

- Univariate testing gene-by-gene
- More descriptive, less predictive

DGE

- Results `results()`

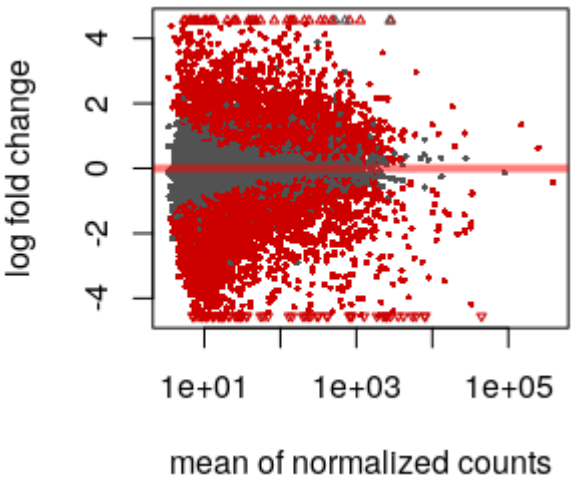
```
log2 fold change (MLE): type type2 vs control
Wald test p-value: type type2 vs control
DataFrame with 1 row and 6 columns
                       baseMean     log2FoldChange             lfcSE
                      <numeric>          <numeric>         <numeric>
ENSG00000000003 242.307796723287 -0.932926089608546 0.114285150312285
                           stat               pvalue                 padj
                      <numeric>            <numeric>             <numeric>
ENSG00000000003 -8.16314356729037 3.26416150242775e-16 1.36240609998527e-14
```

- Summary `summary()`

```
out of 17889 with nonzero total read count
adjusted p-value < 0.1
LFC > 0 (up)       : 4526, 25%
LFC < 0 (down)     : 5062, 28%
outliers [1]       : 25, 0.14%
low counts [2]     : 0, 0%
(mean count < 3)
[1] see 'cooksCutoff' argument of ?results
[2] see 'independentFiltering' argument of ?results
```

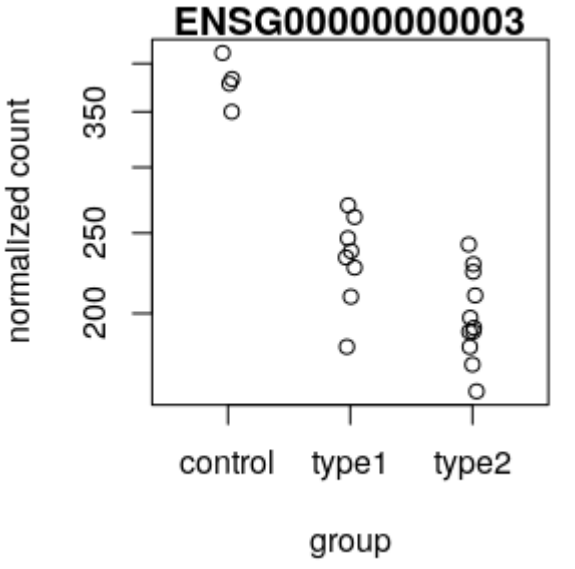
DGE

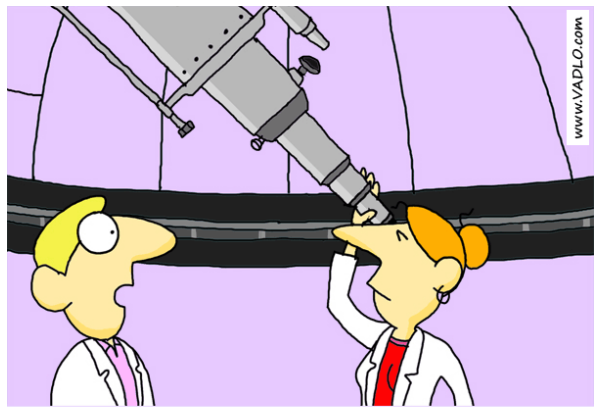
- MA plot `plotMA()`



- Volcano plot

- Normalised counts `plotCounts()`

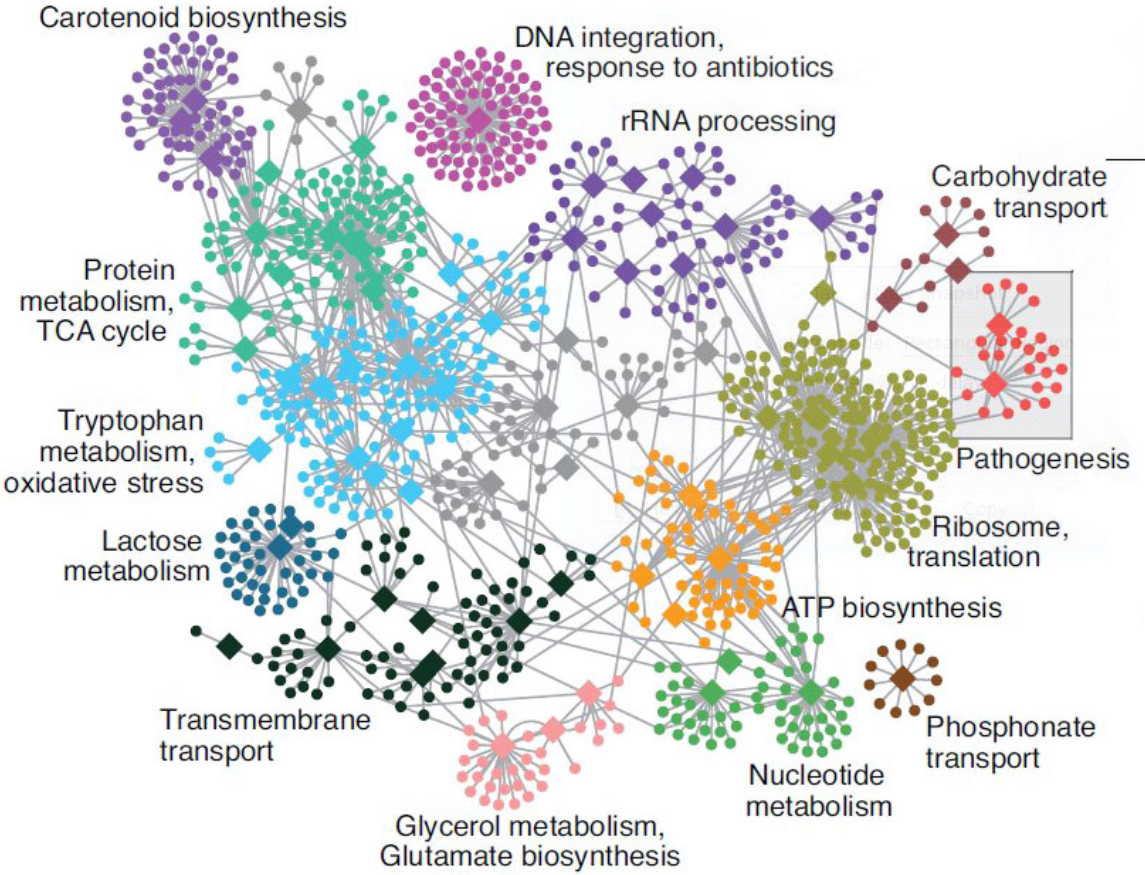


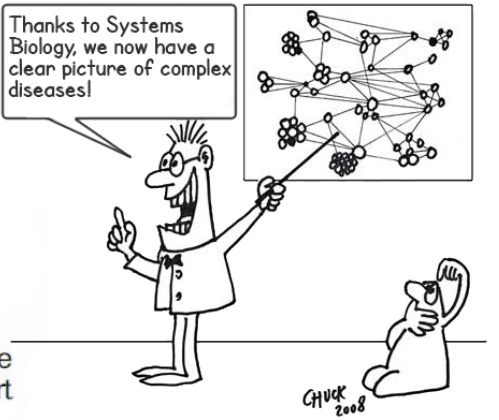


"Can you see the upper points of my scatter plot?"

Functional analysis • GO

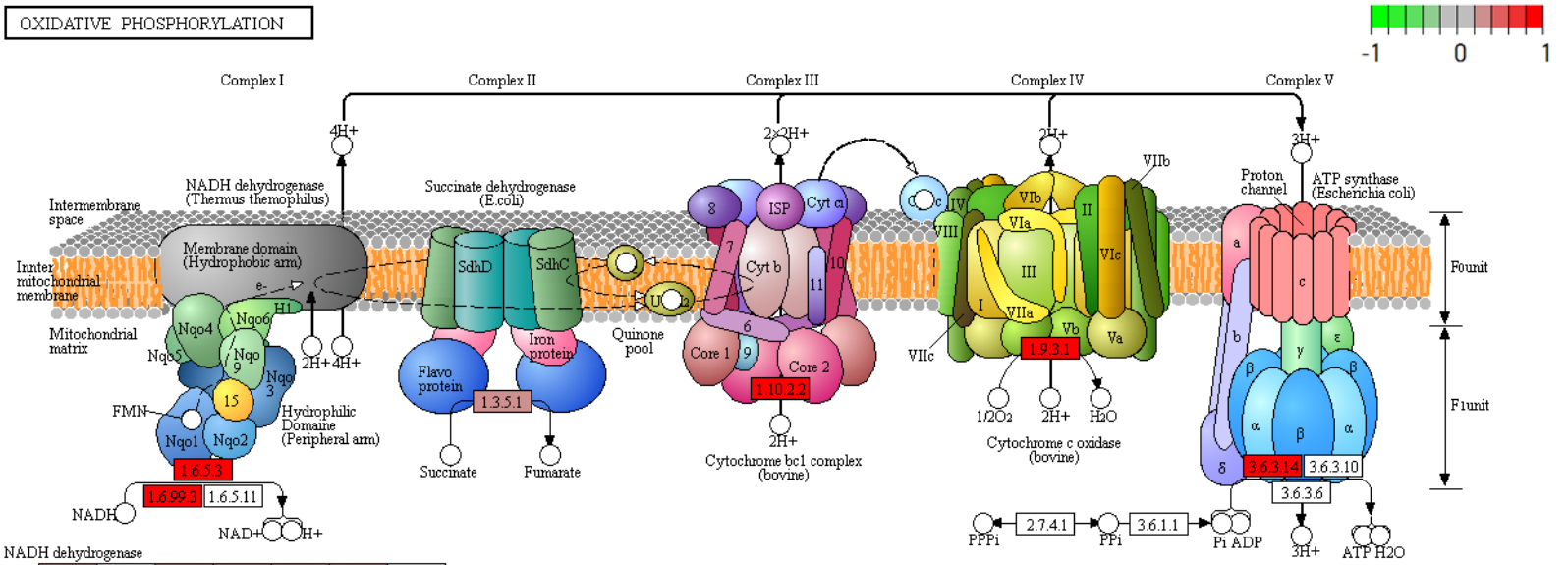
- Gene set analysis (GSA)
- Gene set enrichment analysis (GSEA)
- Gene ontology / Reactome databases





Functional analysis • Kegg

- Pathway analysis (Kegg)



DAVID clusterProfiler ClueGO ErmineJ pathview

Summary

- Sound experimental design to avoid confounding
- Plan carefully about lib prep, sequencing etc based on experimental objective
- For DGE, biological replicates may be more important than other considerations (paired-end, sequencing depth, long reads etc)
- 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, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., ... & Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. Genome biology, 17(1), 1-19.

Thank you. Questions?

Graphics from freepik.com

Hands-On tutorial

NBIS

Main exercise

- 01 Check the quality of the raw reads with **FastQC**
- 02 Map the reads to the reference genome using **HISAT2**
- 03 Assess the post-alignment quality using **QualiMap**
- 04 Count the reads overlapping with genes using **featureCounts**
- 05 Find differentially expressed genes using **DESeq2** in R

Bonus exercises

- 01 Functional annotation of DE genes using **GO/Reactome/Kegg** databases
- 02 RNA-Seq figures and plots using **R**
- 03 Visualisation of RNA-seq BAM files using **IGV** genome browser

Data: `/sw/courses/ngsintro/rnaseq/`
Work: `/proj/snic2021-22-644/nobackup/user/rnaseq/`

Hands-On tutorial

- Course data directory

`/sw/courses/ngsintro/rnaseq/`

```
rnaseq/
+-- bonus/
|   +-- assembly/
|   +-- exon/
|   +-- funannot/
|   +-- plots/
+-- documents/
+-- main/
    +-- 1_raw/
    +-- 2_fastqc/
    +-- 3_mapping/
    +-- 4_qualimap/
    +-- 5_dge/
    +-- 6_multiqc/
    +-- reference/
    |   +-- mouse_chr19_hisat2/
    +-- scripts/
```

- Your work directory

`/proj/snic2021-22-644/nobackup/user/rnaseq/`

```
[user]/
rnaseq/
  +-- 1_raw/
  +-- 2_fastqc/
  +-- 3_mapping/
  +-- 4_qualimap/
  +-- 5_dge/
  +-- 6_multiqc/
  +-- reference/
  |   +-- mouse_chr19_hisat2/
  +-- scripts/
  +-- funannot/
  +-- plots/
```