Analysis of bulk RNA-Seq data

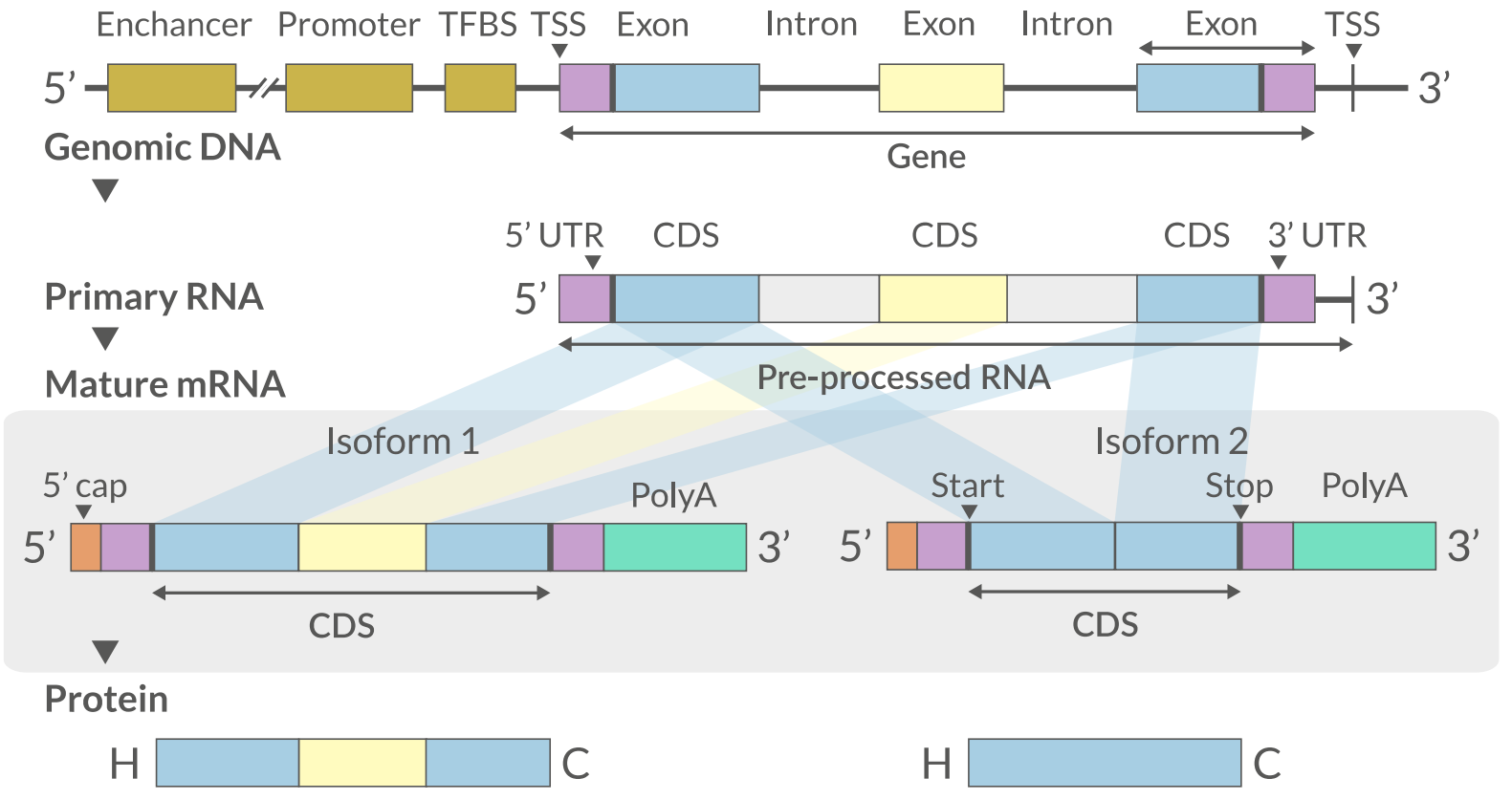
Introduction To Bioinformatics Using NGS Data

25-May-2021

NBIS

Contents

RNA Sequencing



- 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
- 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

Correlation

▼

Clustering

▼

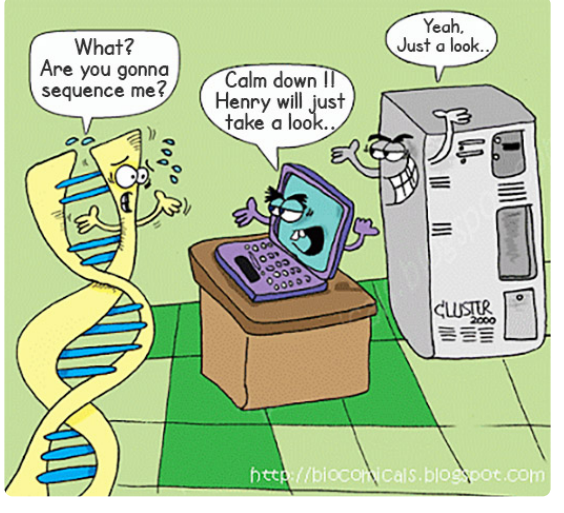
MDS/PCA

▼

Differential Gene Expression

De-novo assembly

Annotation

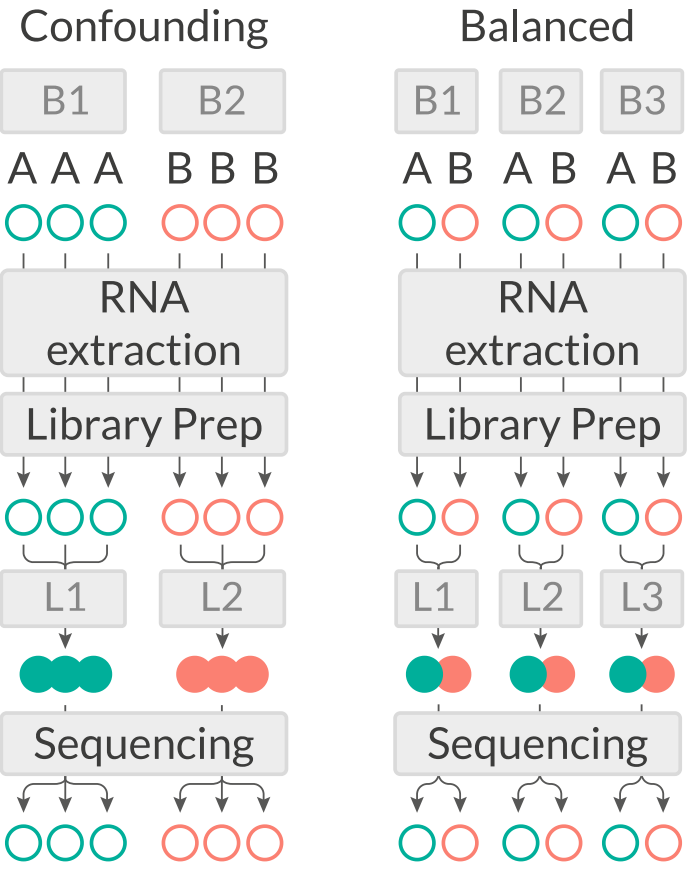


Conesa, Ana, *et al.* "A survey of best practices for RNA-seq data analysis." Genome biology 17.1 (2016): 13

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

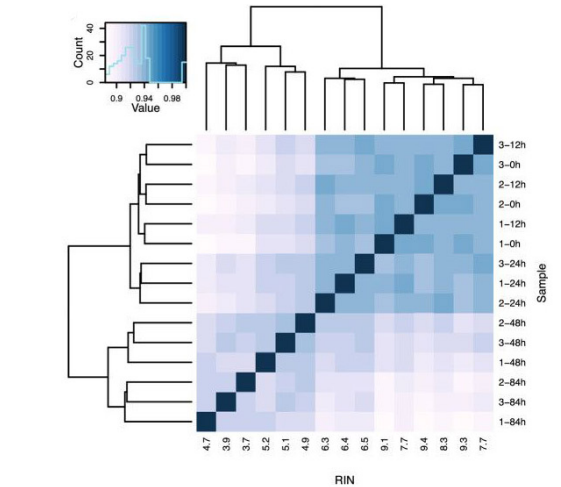
RnaSeqSampleSize (Power analysis), Scotty (Power analysis with cost)

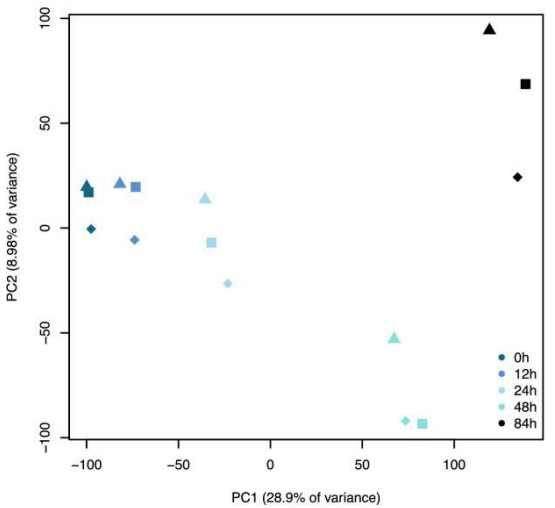


Busby, Michele A., *et al.* "Scotty: a web tool for designing RNA-Seq experiments to measure differential gene expression." Bioinformatics 29.5 (2013): 656-657
Marioni, John C., *et al.* "RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays." Genome research (2008)
Hart, S. N., Therneau, T. M., Zhang, Y., Poland, G. A., & Kocher, J. P. (2013). Calculating sample size estimates for RNA sequencing data. Journal of computational biology, 20(12), 970-978.
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
- Total RNA/mRNA/small RNA
- DNAse treatment
- Quantity & quality
- RIN values (Strong effect)
- Batch effect
- Extraction method bias (GC bias)



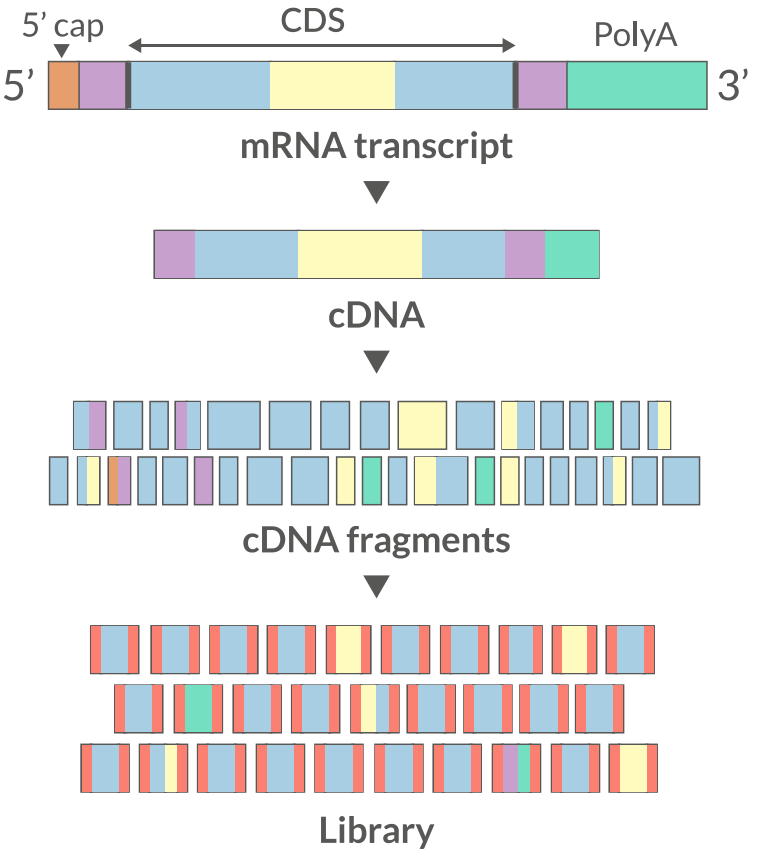


Romero, Irene Gallego, *et al.* "RNA-seq: impact of RNA degradation on transcript quantification." BMC biology 12.1 (2014): 42

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).

Library prep

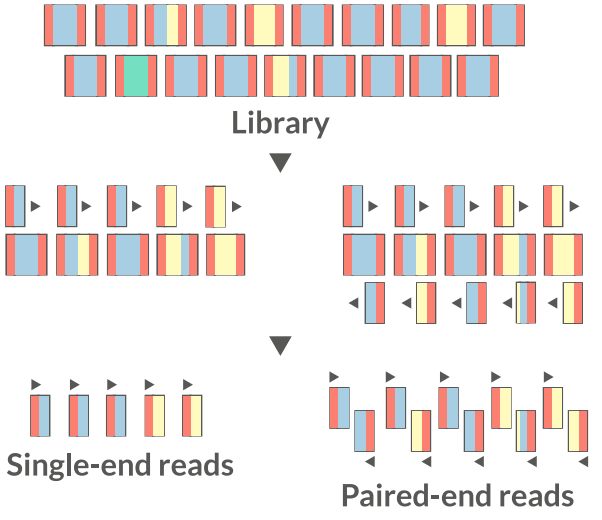
- PolyA selection
- rRNA depletion
- Size selection
- PCR amplification (See section PCR duplicates)
- Stranded (directional) libraries
    - Accurately identify sense/antisense transcript
    - Resolve overlapping genes
- Exome capture
- Library normalisation
- Batch effect



Zhao, Shanrong, et al. "Comparison of stranded and non-stranded RNA-seq transcriptome profiling and investigation of gene overlap." BMC genomics 16.1 (2015): 675
Levin, Joshua Z., et al. "Comprehensive comparative analysis of strand-specific RNA sequencing methods." Nature methods 7.9 (2010): 709

Sequencing

- Sequencer (Illumina/PacBio)
- Read length
  - Greater than 50bp does not improve DGE
  - Longer reads better for isoforms
- Pooling samples
- Sequencing depth (Coverage/Reads per sample)
- Single-end reads (Cheaper)
- Paired-end reads
  - Increased mappable reads
  - Increased power in assemblies
  - Better for structural variation and isoforms
  - Decreased false-positives for DGE



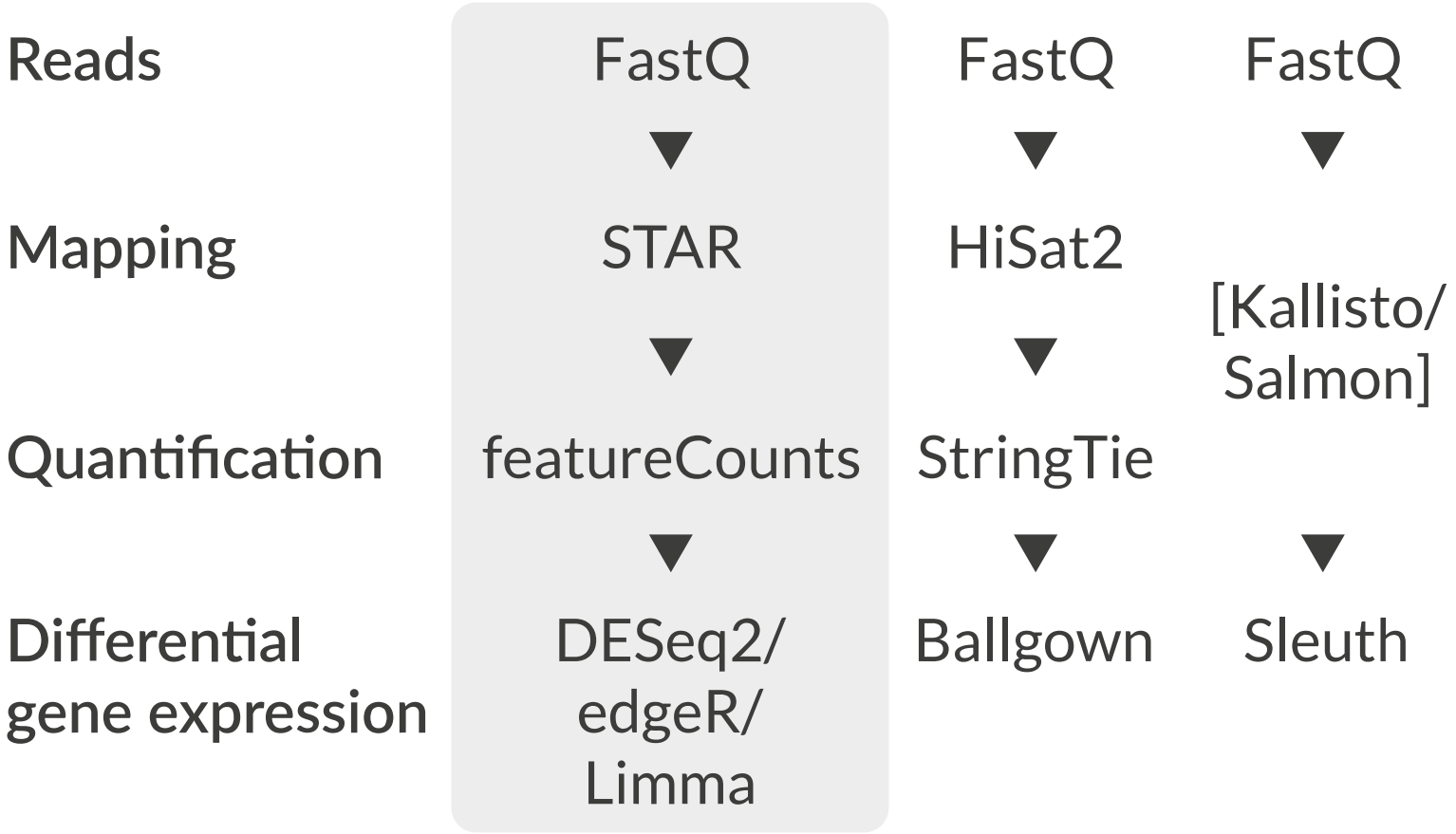
Chhangawala, Sagar, et al. "The impact of read length on quantification of differentially expressed genes and splice junction detection." Genome biology 16.1 (2015): 131

Corley, Susan M., et al. "Differentially expressed genes from RNA-Seq and functional enrichment results are affected by the choice of single-end versus paired-end reads and stranded versus non-stranded protocols." BMC genomics 18.1 (2017): 399

Liu, Yuwen, Jie Zhou, and Kevin P. White. "RNA-seq differential expression studies: more sequence or more replication?." Bioinformatics 30.3 (2013): 301-304

Comparison of PE and SE for RNA-Seq, SciLifeLab

Workflow • DGE



De-Novo assembly

- When no reference genome available
- To identify novel genes/transcripts/isoforms
- Identify fusion genes
- Assemble transcriptome from short reads
- Access quality of assembly and refine
- Map reads back to assembled transcriptome

Trinity, SOAPdenovo-Trans, Oases, rnaSPAdes

Hsieh, Ping-Han *et al.*, "Effect of de novo transcriptome assembly on transcript quantification" 2018 bioRxiv 380998
Wang, Sufang, and Michael Gribskov. "Comprehensive evaluation of de novo transcriptome assembly programs and their effects on differential gene expression analysis." Bioinformatics 33.3 (2017): 327-333

Read QC

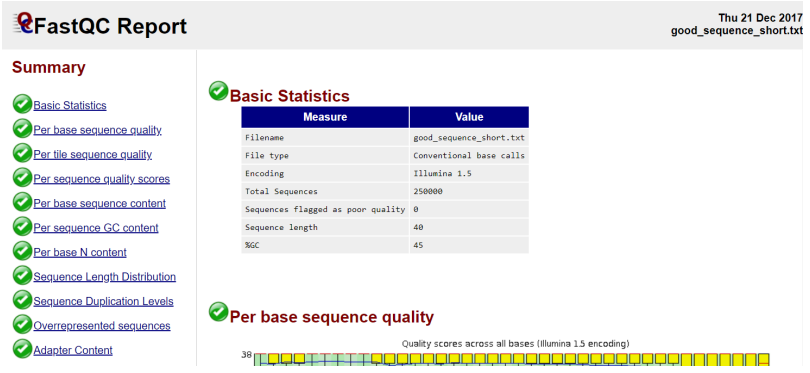
- Number of reads
- Per base sequence quality
- Per sequence quality score
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence length distribution
- Sequence duplication levels
- Overrepresented sequences
- Adapter content
- Kmer content

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

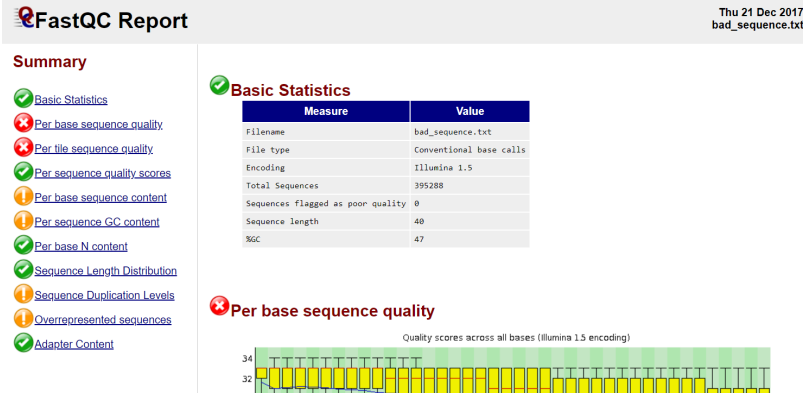
Measure	Value
Filename	good_sequence_short.txt
File type	Conventional base calls
Encoding	Illumina 1.5
Total Sequences	250000
Sequences flagged as poor quality	0
Sequence length	40
%GC	45

Per base sequence quality

Quality scores across all bases (Illumina 1.5 encoding)

38

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

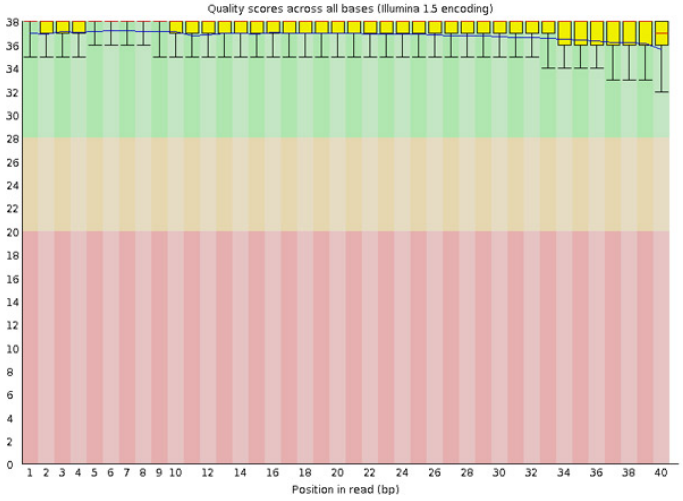
Quality scores across all bases (Illumina 1.5 encoding)

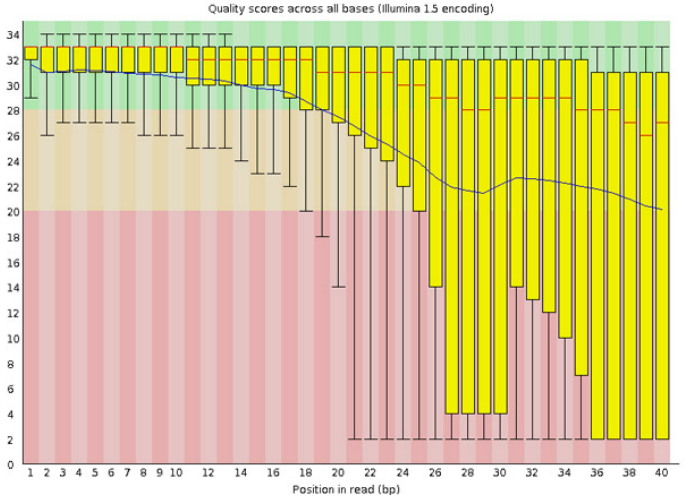
34

32

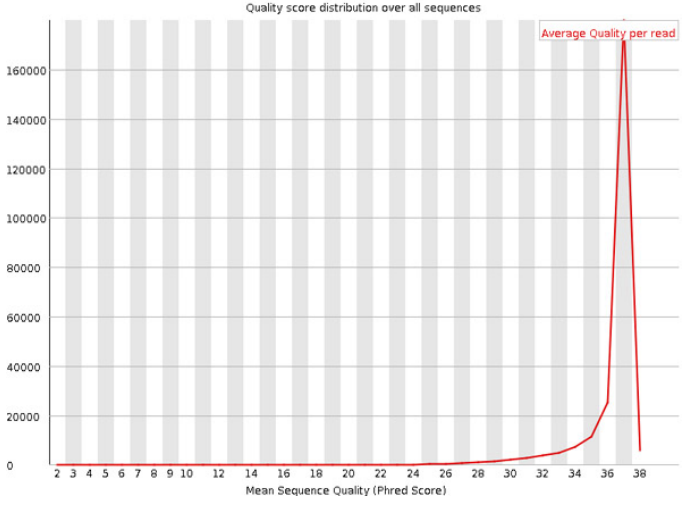
Read QC • PBSQ, PSQS

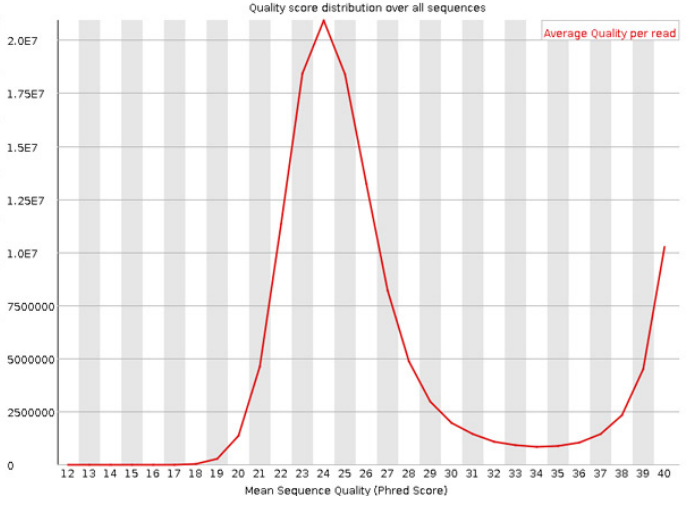
Per base sequence quality





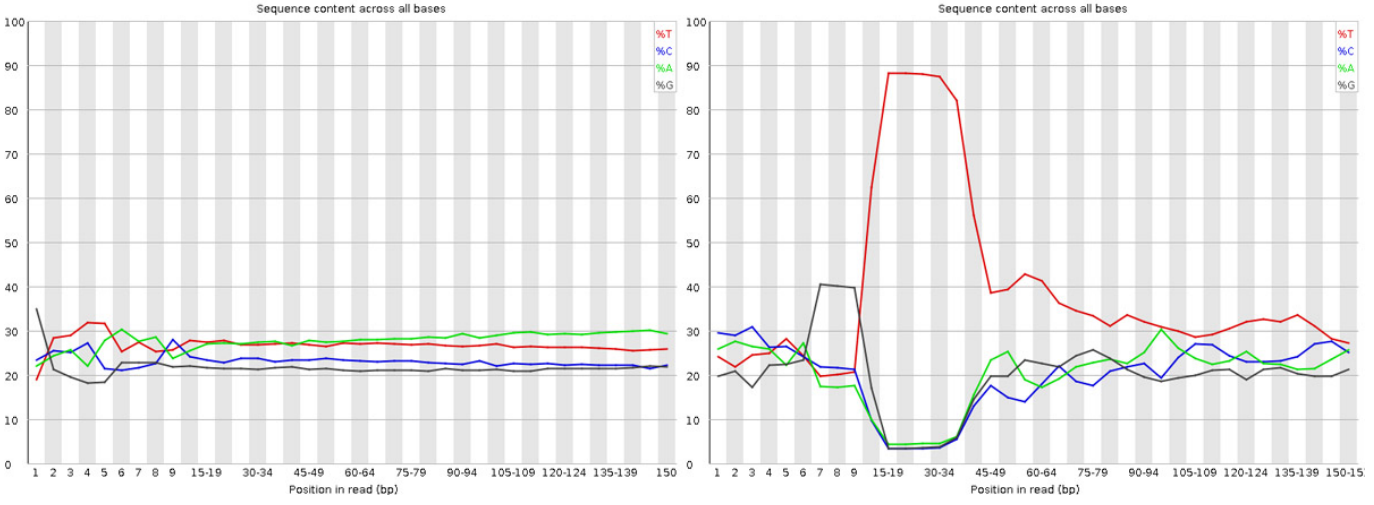
Per sequence quality scores



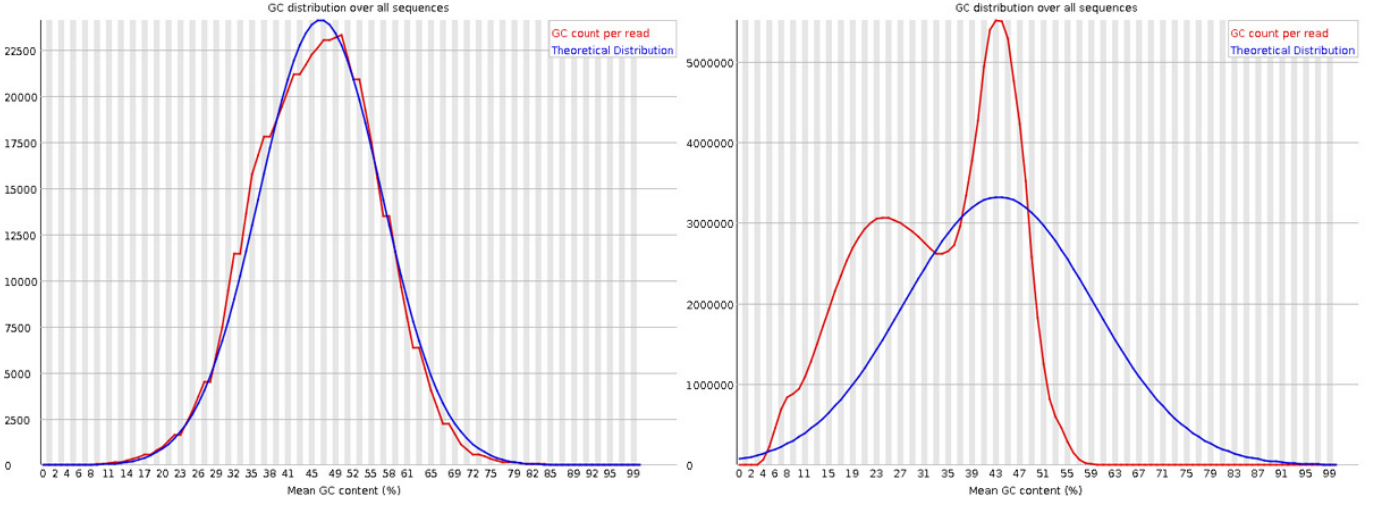


Read QC • PBSC, PSGC

Per base sequence content

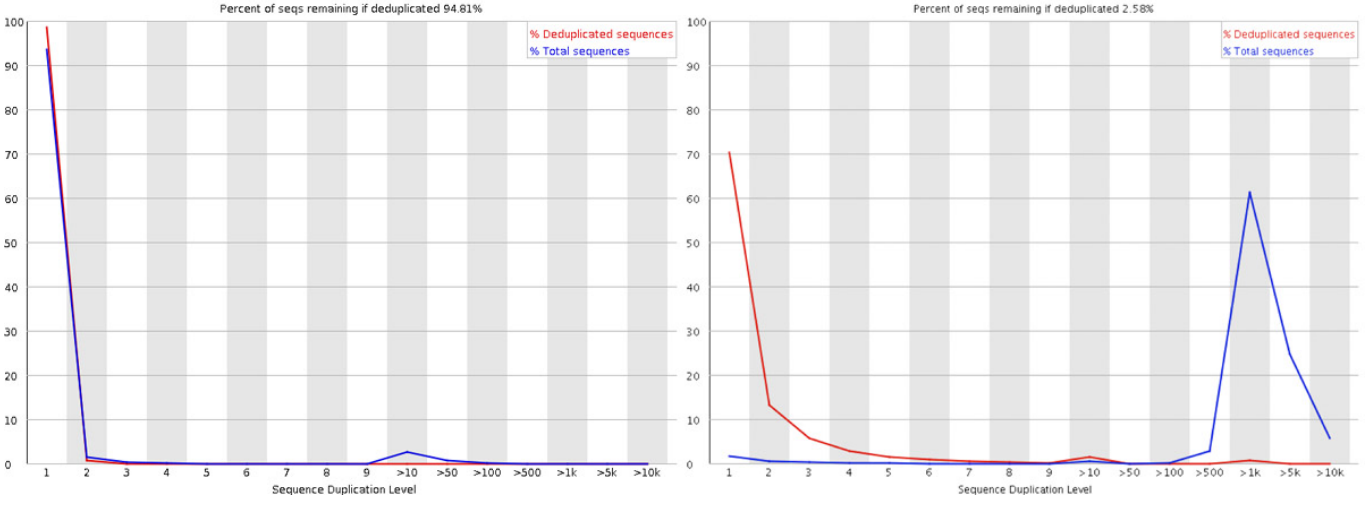


Per sequence GC content



Read QC • SDL, AC

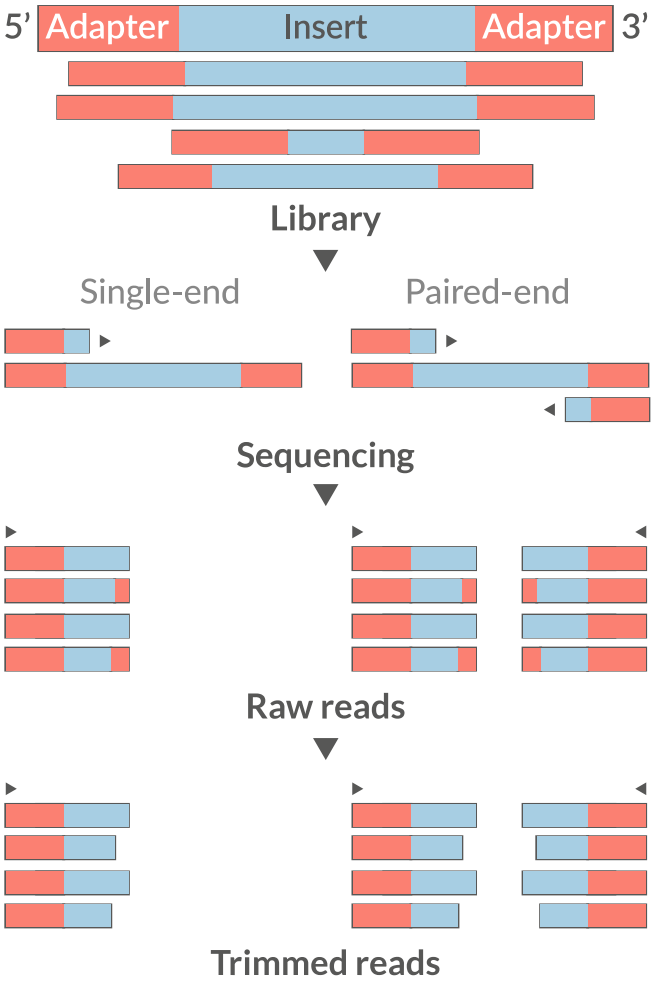
Sequence duplication level



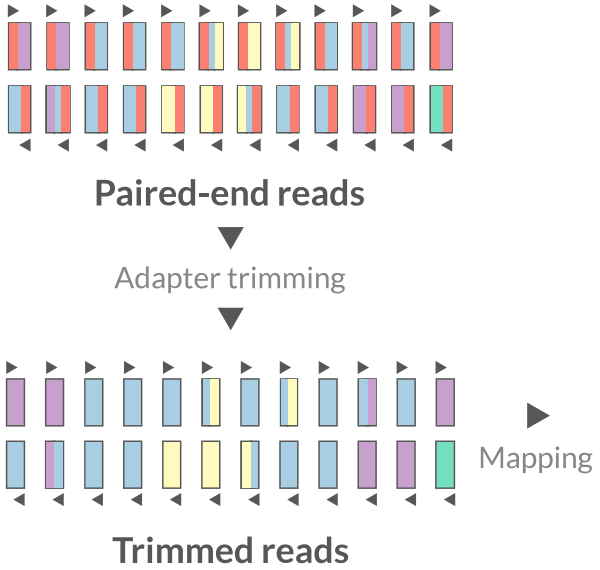
Trimming

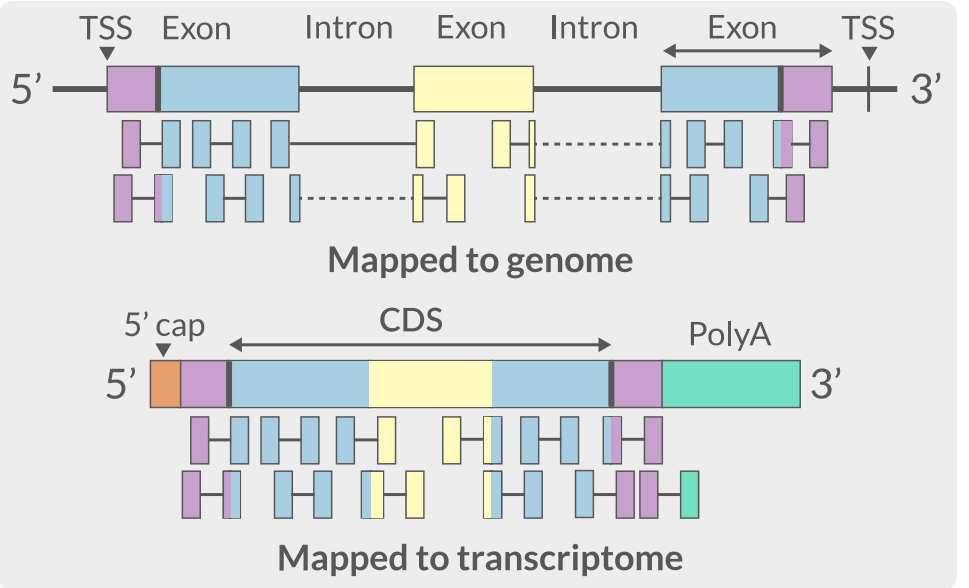
- Trim IF necessary
    - Synthetic bases can be an issue for SNP calling
    - Insert size distribution may be more important for assemblers
- Trim/Clip/Filter reads
- Remove adapter sequences
- Trim reads by quality
- Sliding window trimming
- Filter by min/max read length
    - Remove reads less than ~18nt
- Demultiplexing/Splitting

🧰 Cutadapt, fastp, Skewer, Prinseq



Mapping



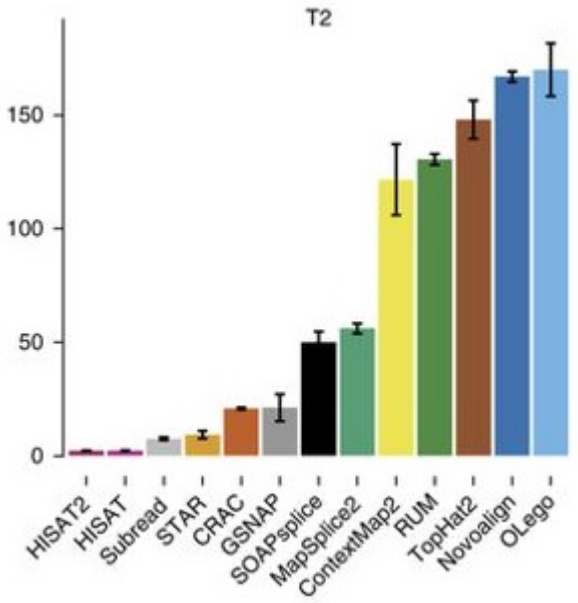


- Aligning reads back to a reference sequence
- Mapping to genome vs transcriptome
- Splice-aware alignment (genome)

STAR, HiSat2, GSNAP, Novoalign (Commercial)

Baruzzo, Giacomo, *et al.* "Simulation-based comprehensive benchmarking of RNA-seq aligners." Nature methods 14.2 (2017): 135

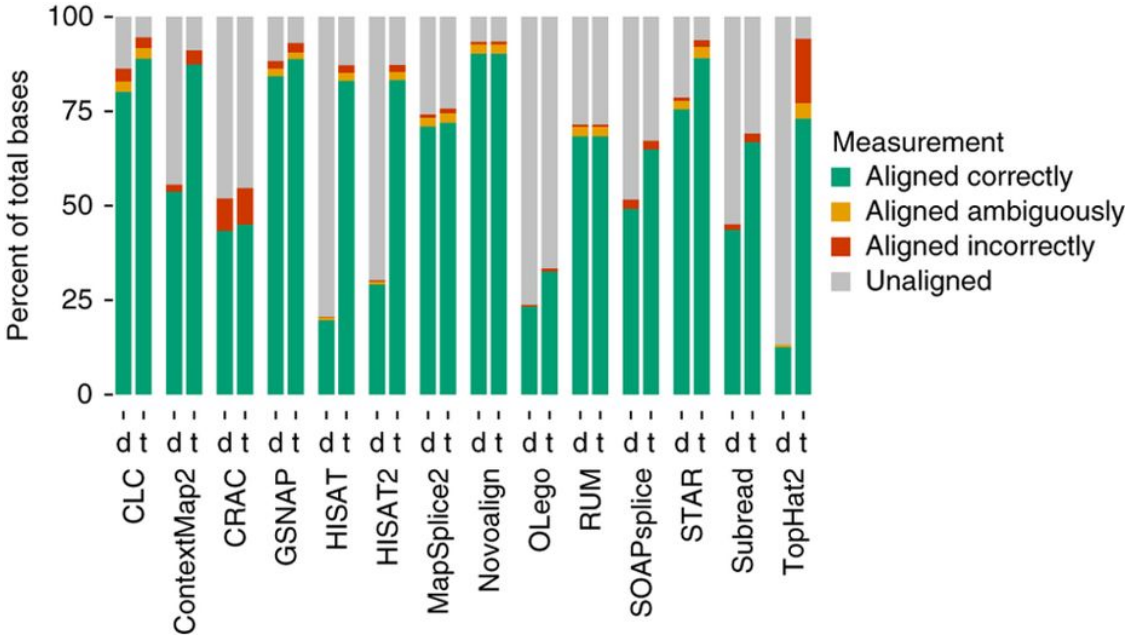
Aligners • Speed



Program	Time_Min	Memory_GB
HISATx1	22.7	4.3
HISATx2	47.7	4.3
HISAT	26.7	4.3
STAR	25	28
STARx2	50.5	28
GSNAP	291.9	20.2
TopHat2	1170	4.3

Baruzzo, Giacomo, *et al.* "Simulation-based comprehensive benchmarking of RNA-seq aligners." Nature methods 14.2 (2017): 135

Aligners • Accuracy



Increasing Accuracy

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

STAR, HiSat2, GSNAP, Novoalign (Commercial)

Baruzzo, Giacomo, *et al.* "Simulation-based comprehensive benchmarking of RNA-seq aligners." Nature methods 14.2 (2017): 135

Mapping

- Reads (FASTQ)

```
@ST-E00274:179:HHYMLALXX:8:1101:1641:1309 1:N:0:NGATGT
NCATCGTGGTATTTGCACATCTTTTCTTATCAAATAAAAAGTTTAACCTACTCAGTTATGCGCATACGTTTTTTGATGGCATTTCCATA
+
#AAAFAFA<-AFFJJJAFA-FFJJJJFFFAJJJJ-<FFJJJ-A-F-7--FA7F7-----FFFJFA<FFFFJ<AJ--FF-A<A-<JJ-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 "ENSDARG0
```

`seq source feature start end score strand frame attribute`

Illumina read name format, GTF format

Alignment

- SAM/BAM (Sequence Alignment Map format)

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

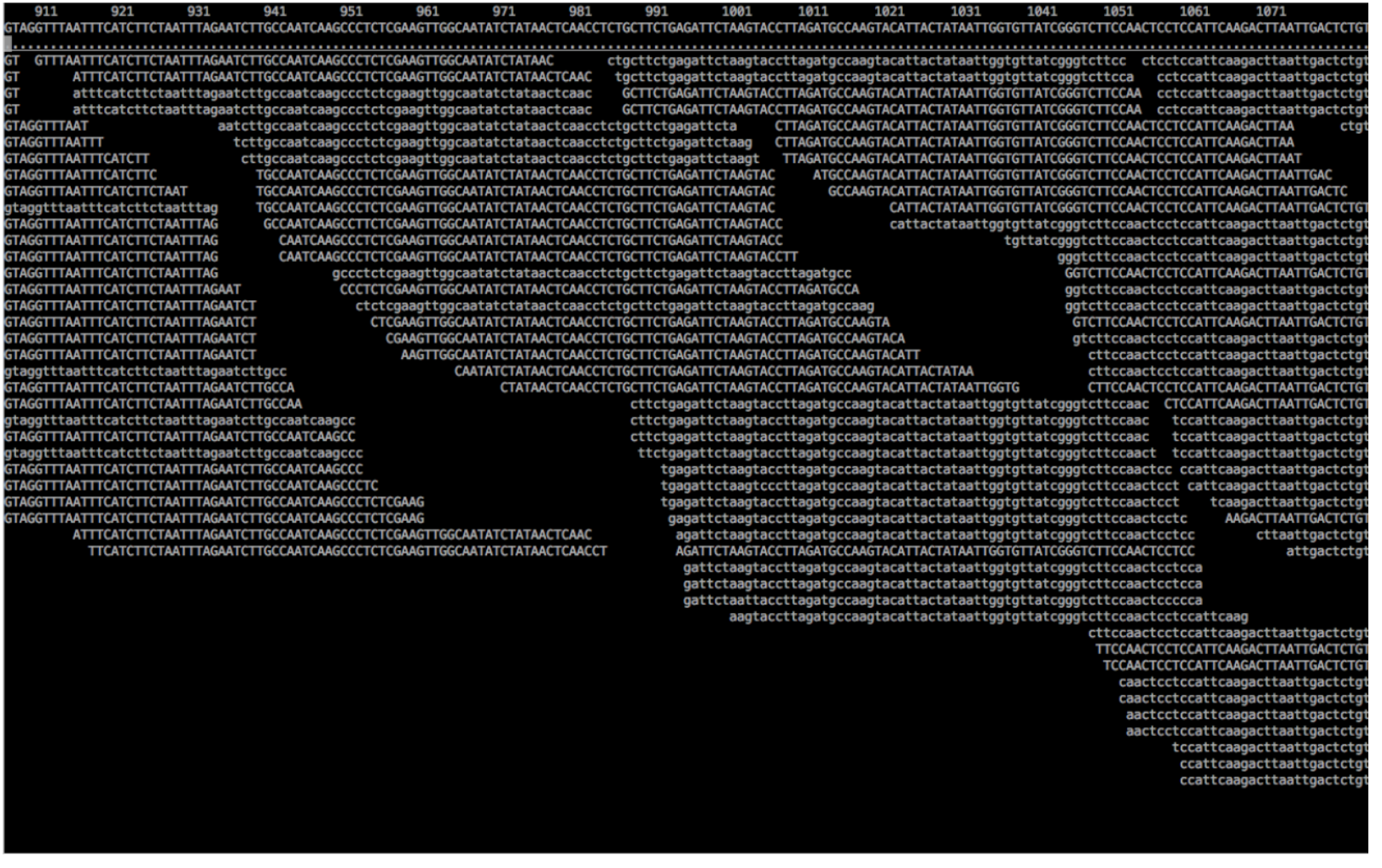
```
query flag ref pos mapq cigar mrnm mpos tlen seq qual opt
```

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

Visualisation • tview

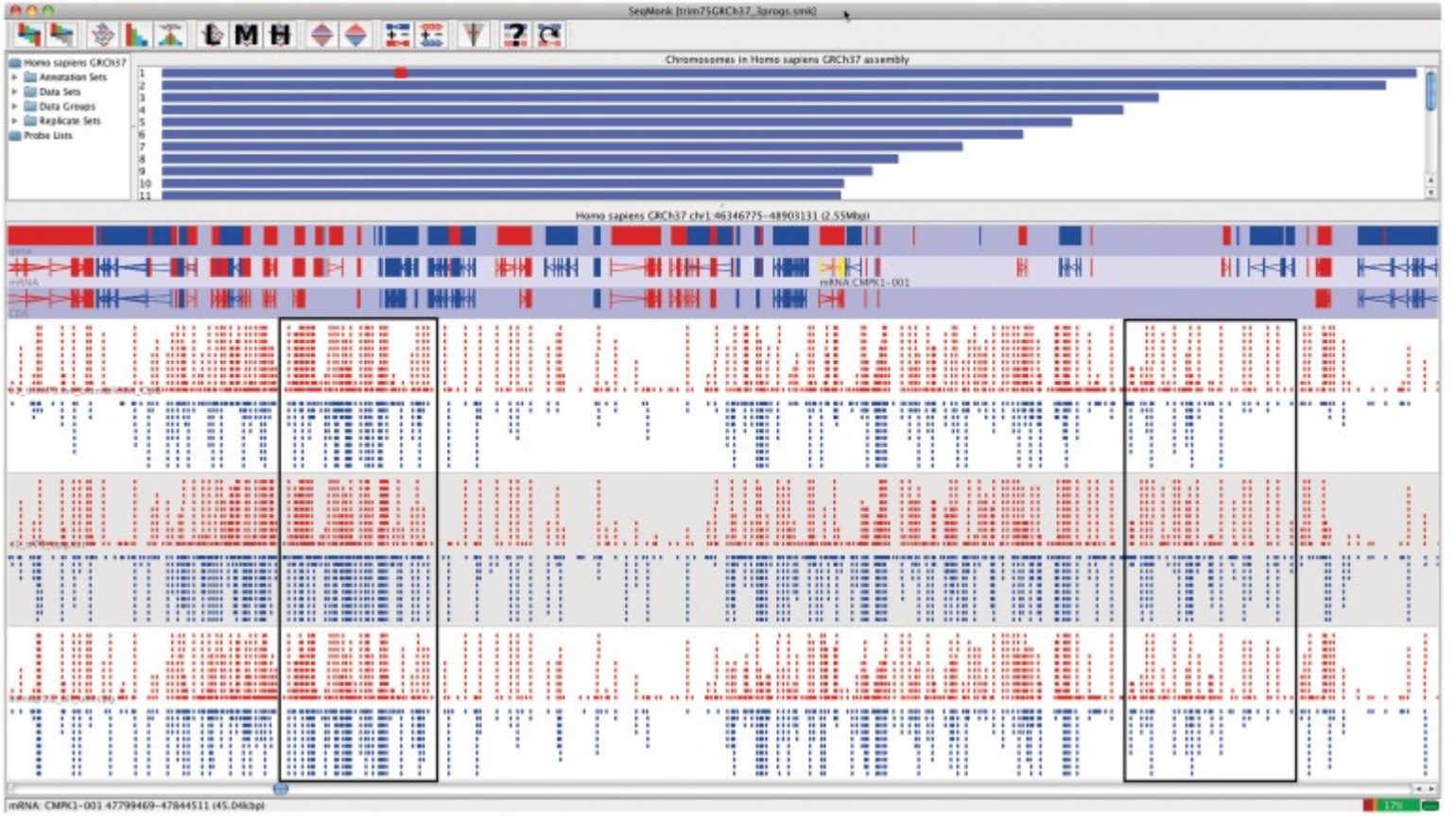
```
samtools tview alignment.bam genome.fasta
```



Visualisation • IGV

IGV, UCSC Genome Browser

Visualisation • SeqMonk



SeqMonk

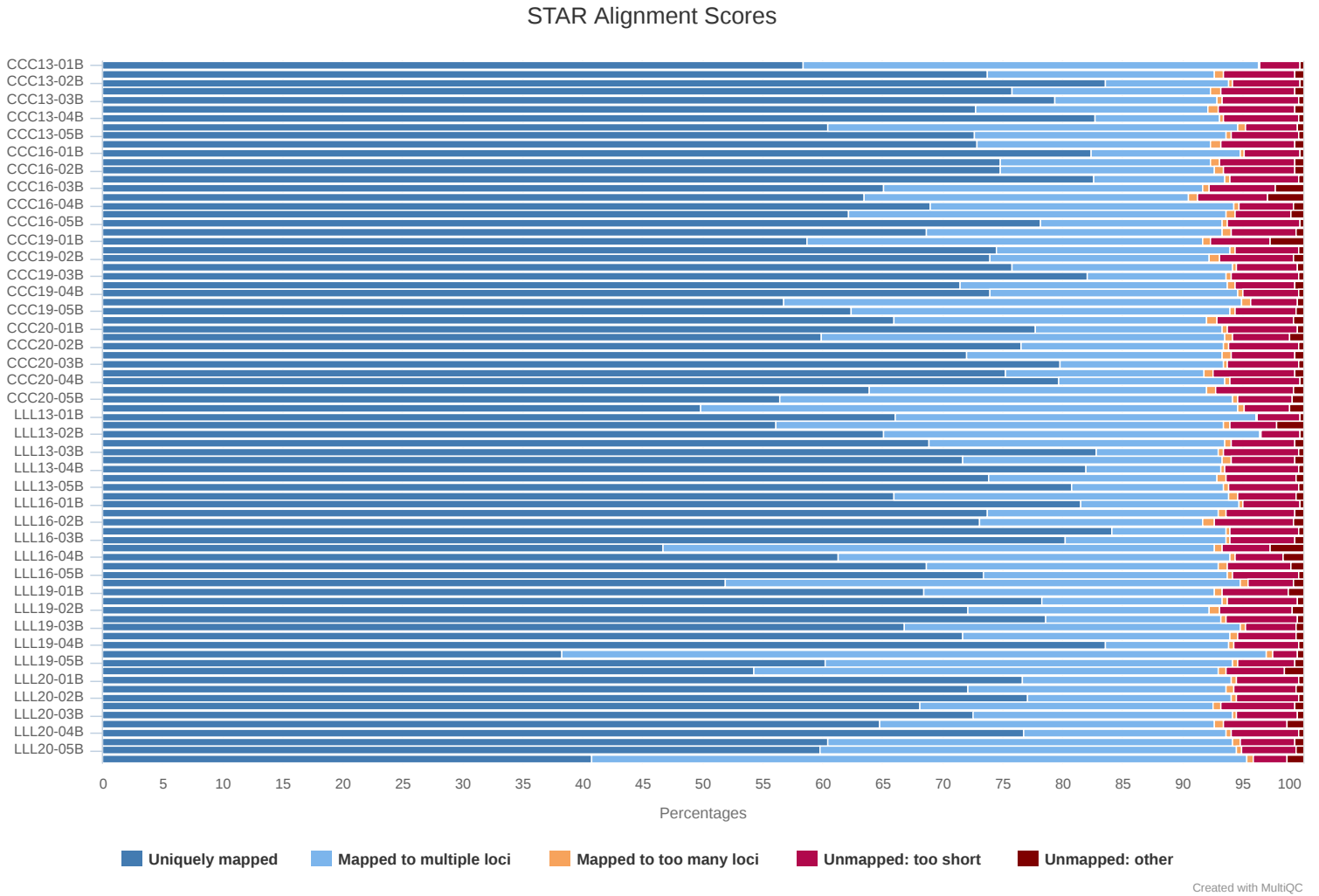
Alignment QC

- 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

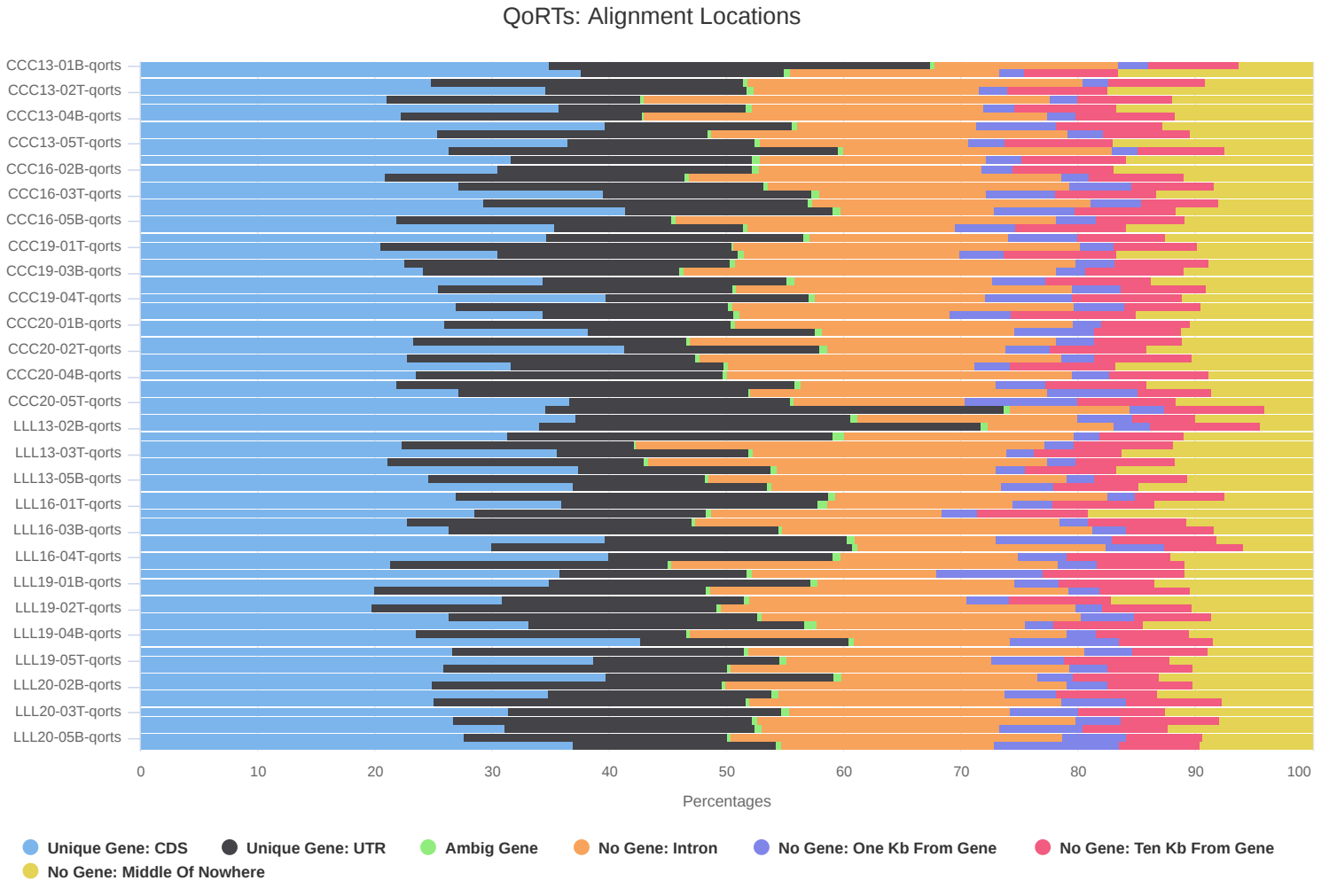
Alignment QC • STAR Log

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

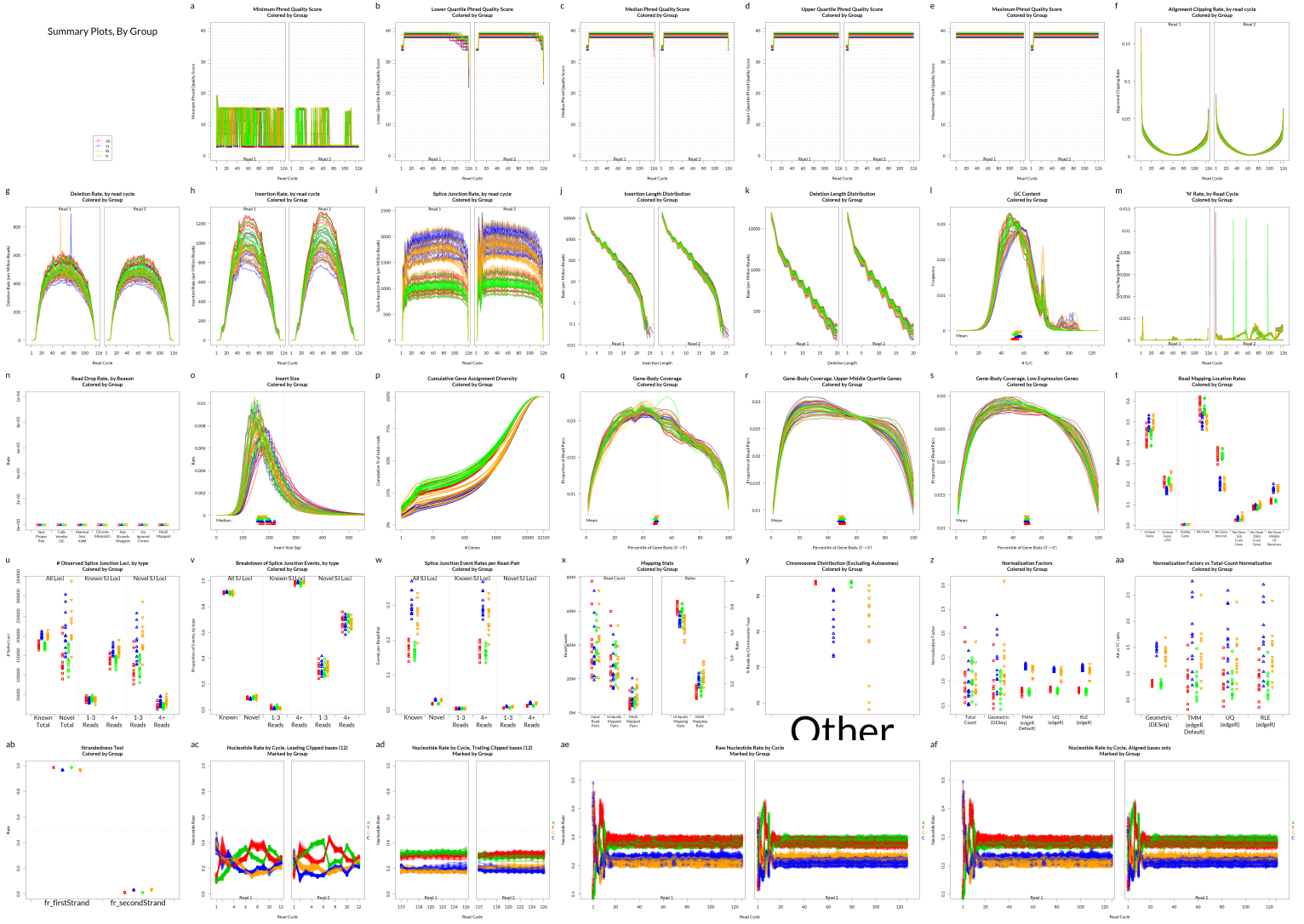


Alignment QC • Features

QoRTs was run on all samples and summarised using MultiQC.

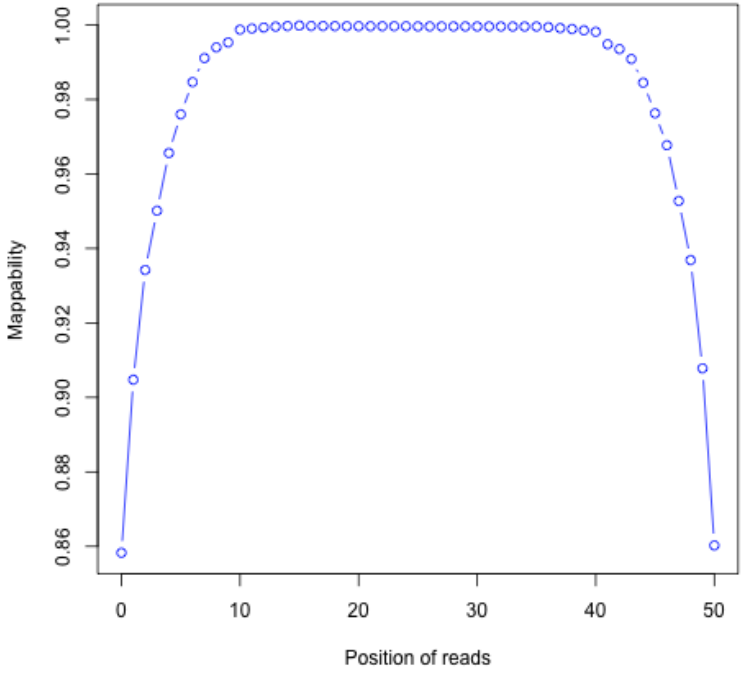


Alignment QC • QoRTs

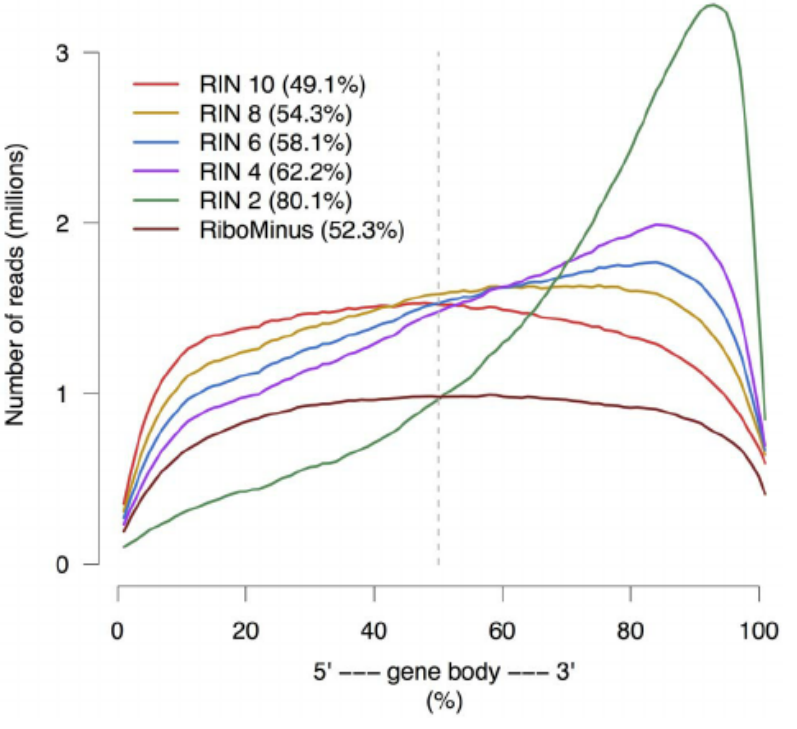


Alignment QC • Examples

Soft clipping

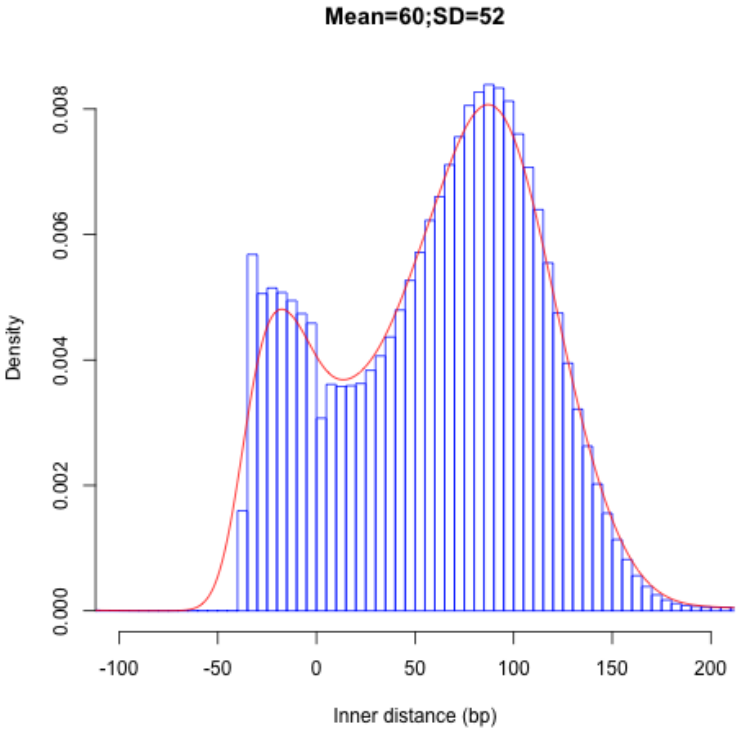


Gene body coverage

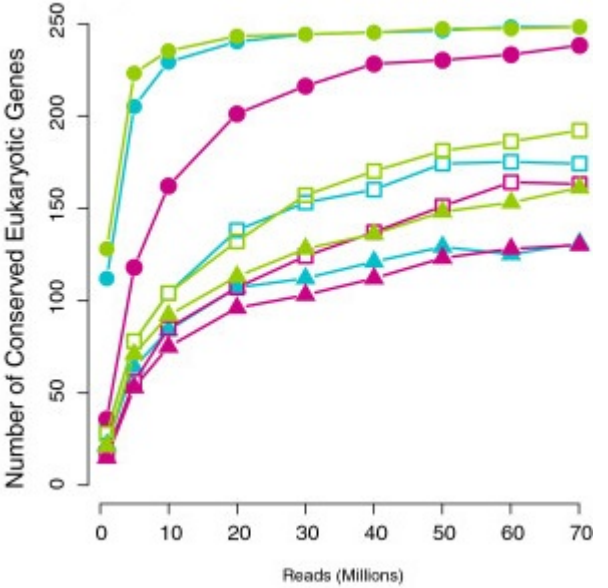


Alignment QC • Examples

Insert size

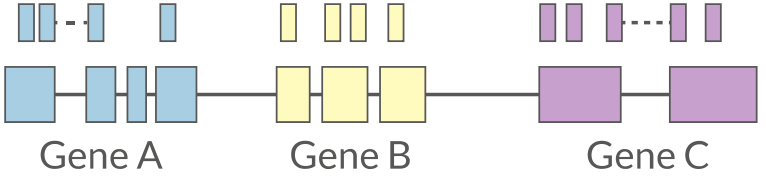


Saturation curve

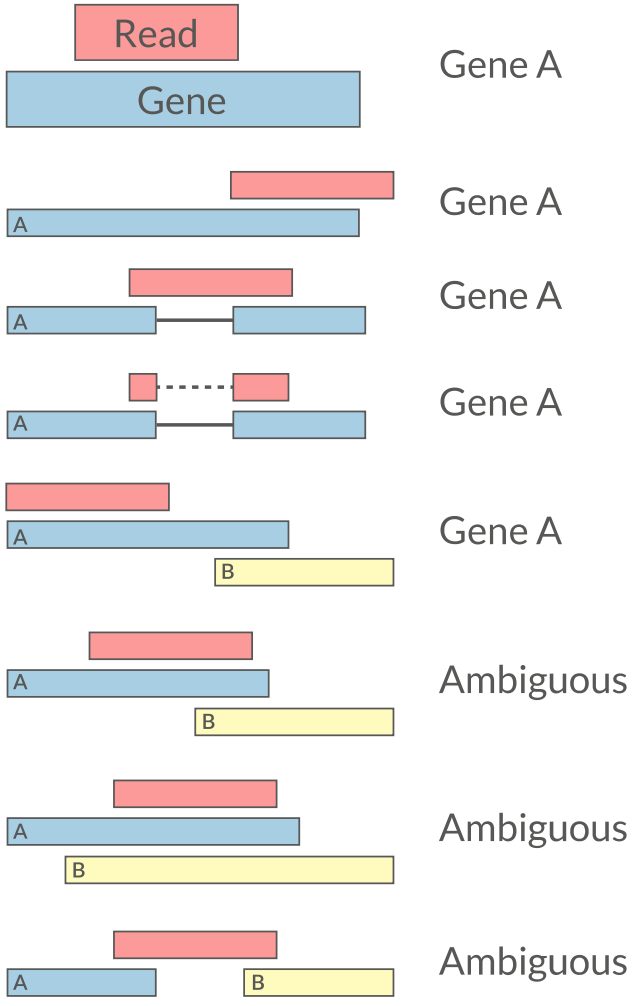


Quantification • Counts

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



featureCounts, HTSeq



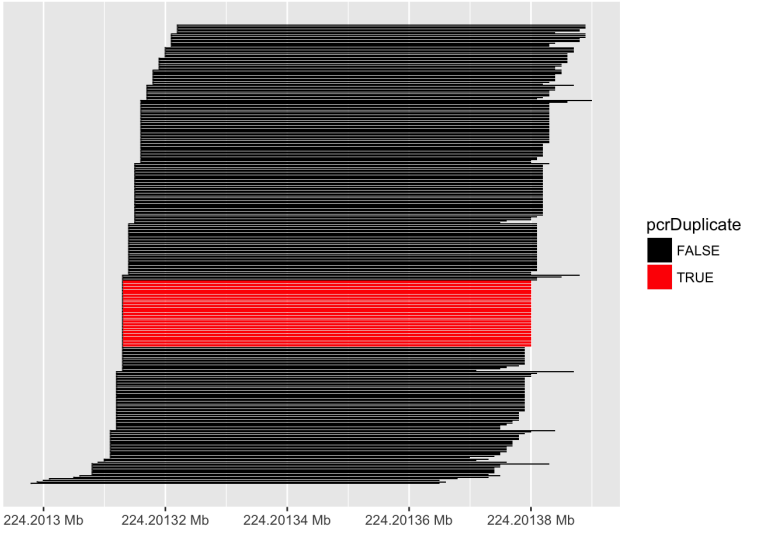
Quantification

PCR duplicates

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

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)



Fu, Yu, *et al*. "Elimination of PCR duplicates in RNA-seq and small RNA-seq using unique molecular identifiers." BMC genomics 19.1 (2018): 531
Parekh, Swati, *et al*. "The impact of amplification on differential expression analyses by RNA-seq." Scientific reports 6 (2016): 25533
Klepikova, Anna V., *et al*. "Effect of method of deduplication on estimation of differential gene expression using RNA-seq." PeerJ 5 (2017): e3091

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
- Subsampling & quantification confidence
- Transcript-level estimates improves gene-level estimates
- Kallisto/Salmon > transcript-counts > `tximport()` > gene-counts

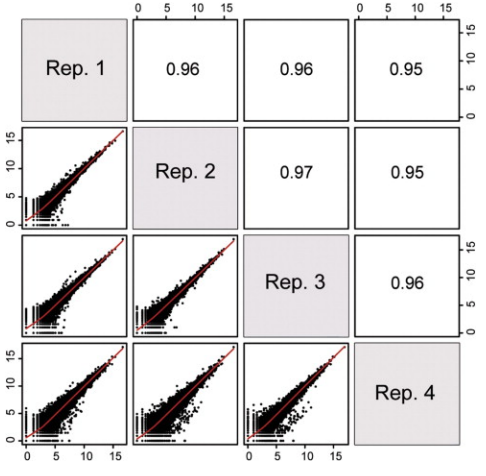
🧰 RSEM, Kallisto, Salmon, Cufflinks2

🔗 Soneson, Charlotte, *et al.* "Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences." F1000Research 4 (2015)
🔗 Zhang, Chi, *et al.* "Evaluation and comparison of computational tools for RNA-seq isoform quantification." BMC genomics 18.1 (2017): 583

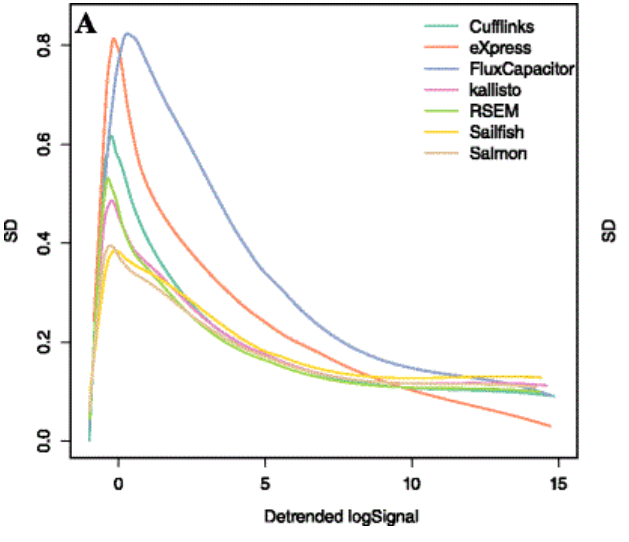
Quantification QC

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)



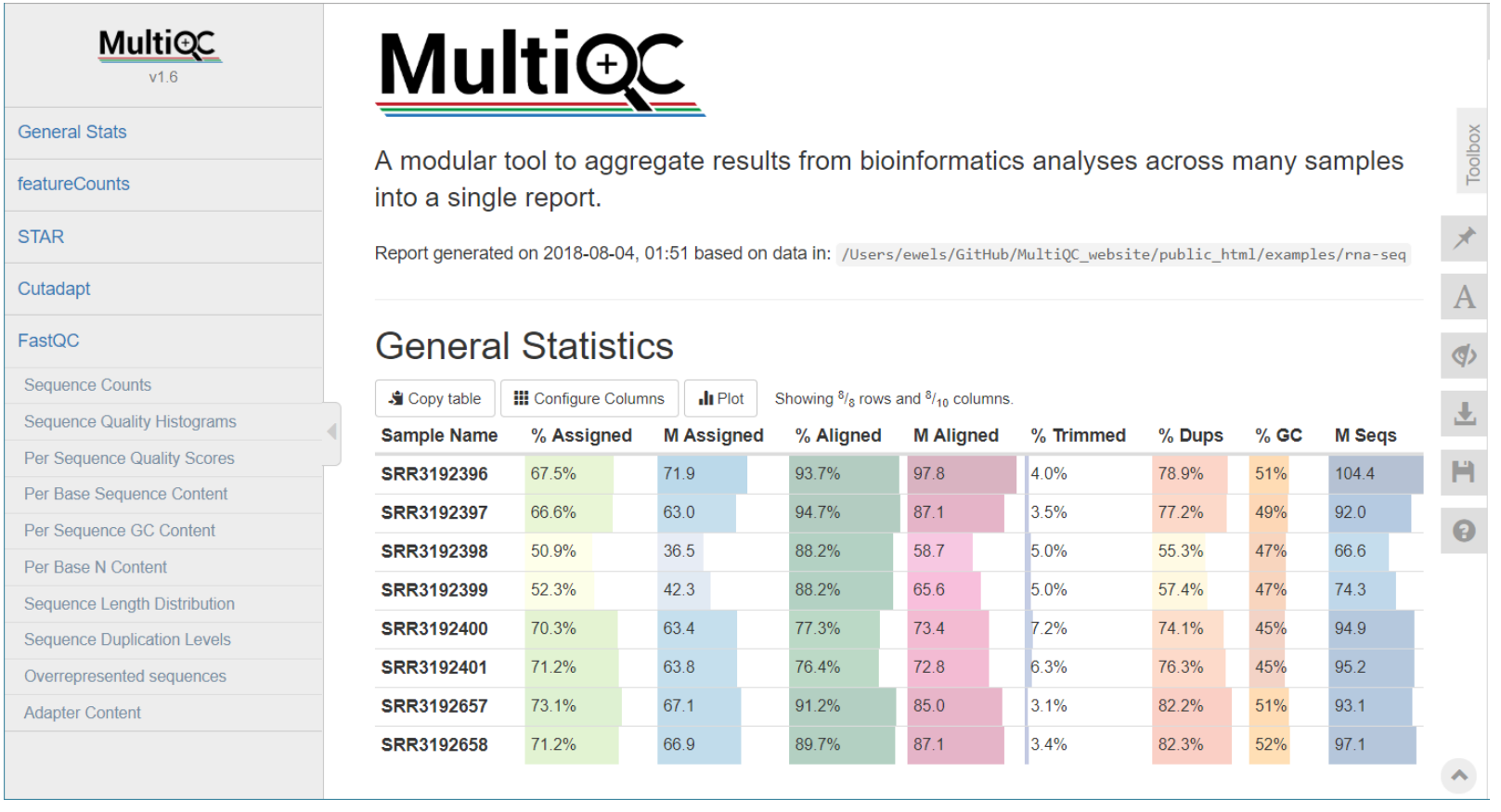
- Count QC using RNASeqComp



RNASeqComp

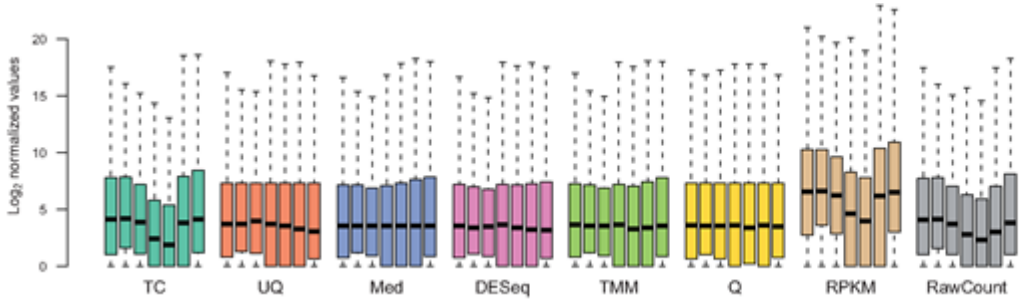
Teng, Mingxiang, *et al.* "A benchmark for RNA-seq quantification pipelines." Genome biology 17.1 (2016): 74

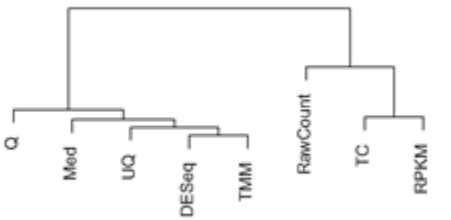
MultiQC


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
Toolbox
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

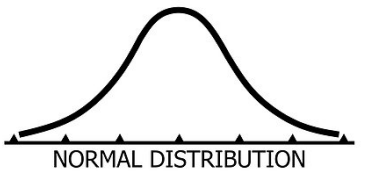
Normalisation

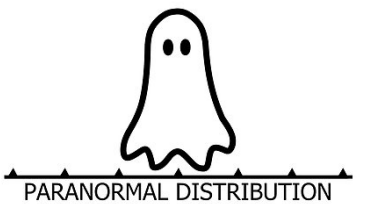
- Control for Sequencing depth & compositional bias
- Median of Ratios (DESeq2) and TMM (edgeR) perform the best





- 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





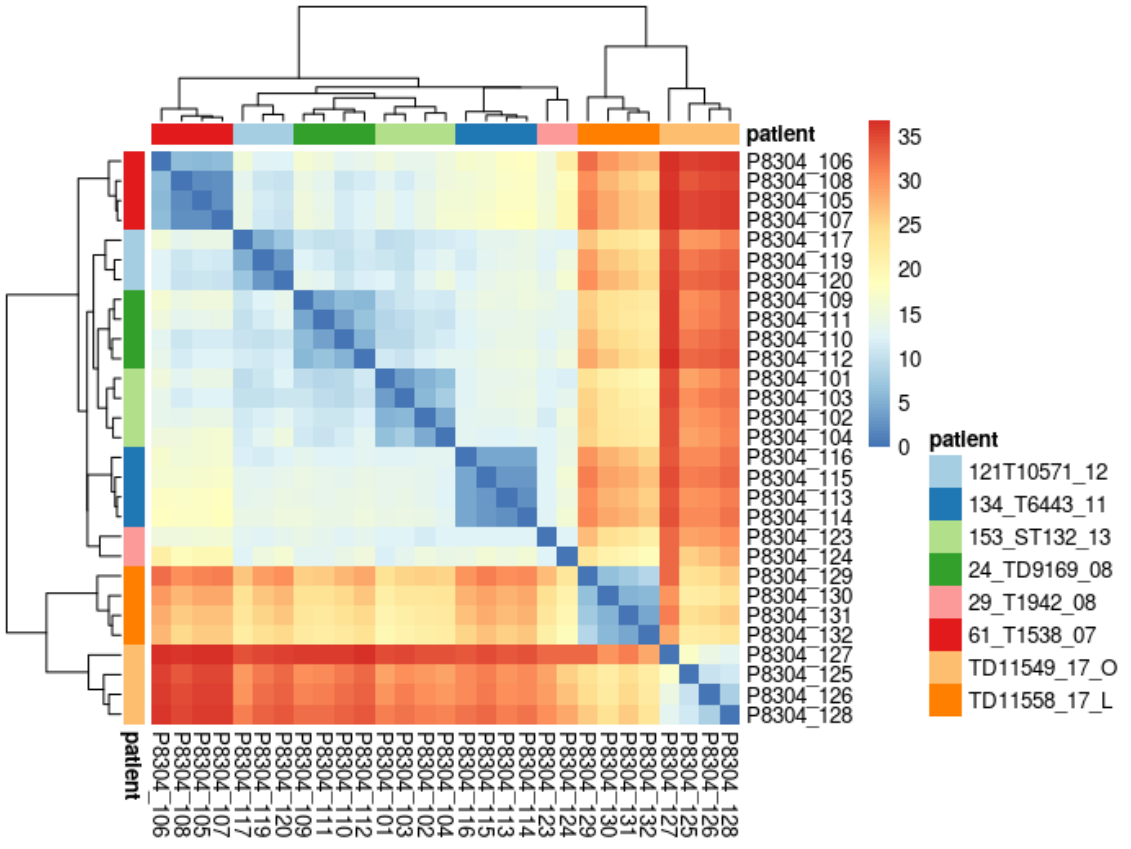
Dillies, Marie-Agnes, *et al.* "A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis." Briefings in bioinformatics 14.6 (2013): 671-683

Evans, Ciaran, Johanna Hardin, and Daniel M. Stoebel. "Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions." Briefings in bioinformatics (2017)

Wagner, Gunter P., Koryu Kin, and Vincent J. Lynch. "Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples." Theory in biosciences 131.4 (2012): 281-285

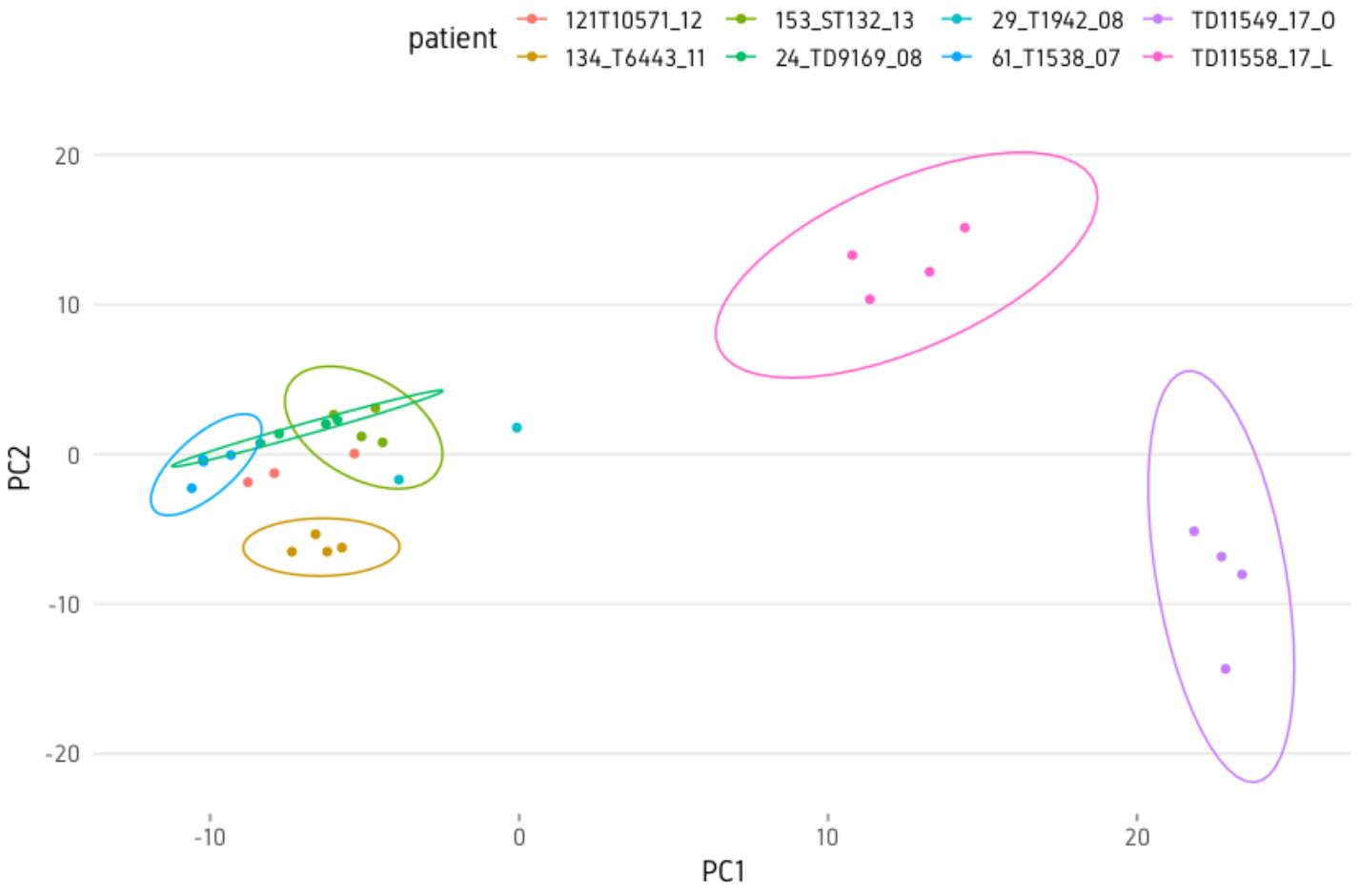
Exploratory • Heatmap

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



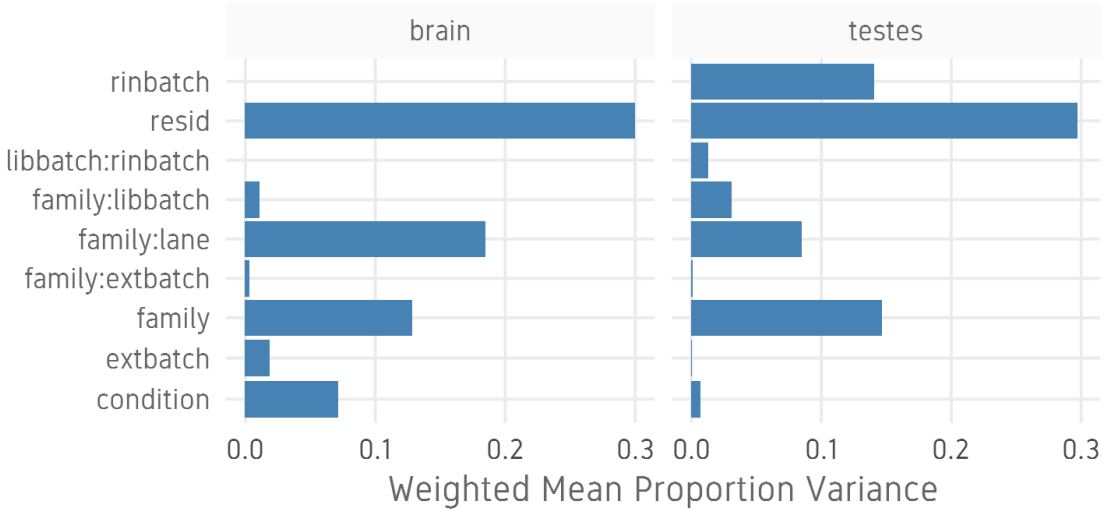
`pheatmap()`

Exploratory • PCA



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

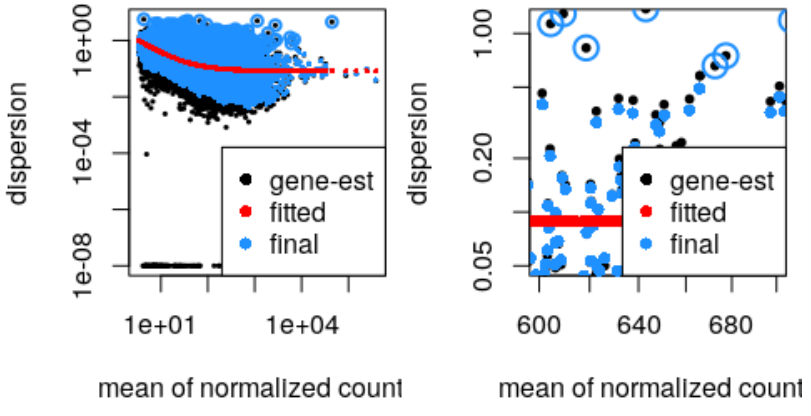
Liu, Qian, and Marianthi Markatou. "Evaluation of methods in removing batch effects on RNA-seq data." Infectious Diseases and Translational Medicine 2.1 (2016): 3-9
Manimaran, Solaiappan, et al. "BatchQC: interactive software for evaluating sample and batch effects in genomic data." Bioinformatics 32.24 (2016): 3836-3838

DGE

		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	...	...

DGE

- DESeq2, edgeR (Neg-binom > GLM > Test), Limma-Voom (Neg-binom > Voom-transform > LM > Test)
- DESeq2 `~age+condition`
    - Estimate size factors `estimateSizeFactors()`
    - Estimate gene-wise dispersion `estimateDispersions()`
    - Fit curve to gene-wise dispersion estimates
    - Shrink gene-wise dispersion estimates
    - GLM fit for each gene
    - Wald test `nbinomWaldTest()`



DESeq2, edgeR, Limma-Voom

Seyednasrollah, Fatemeh, *et al.* "Comparison of software packages for detecting differential expression in RNA-seq studies." Briefings in bioinformatics 16.1 (2013): 59-70

DGE

- Results `results()`

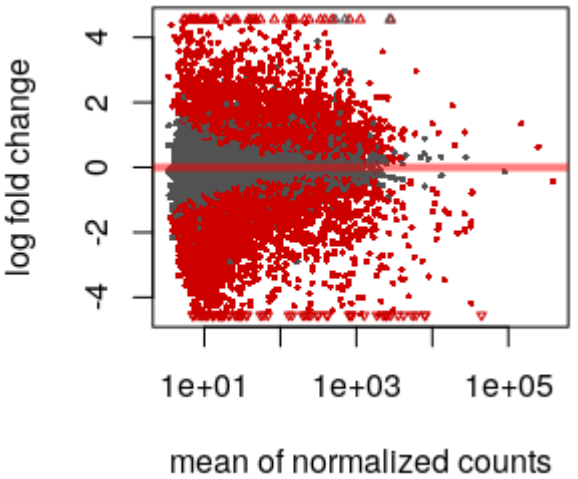
```
log2 fold change (MLE): type type2 vs control
Wald test p-value: type type2 vs control
DataFrame with 1 row and 6 columns
                       baseMean     log2FoldChange             lfcSE
                      <numeric>          <numeric>         <numeric>
ENSG00000000003 242.307796723287 -0.932926089608546 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
```

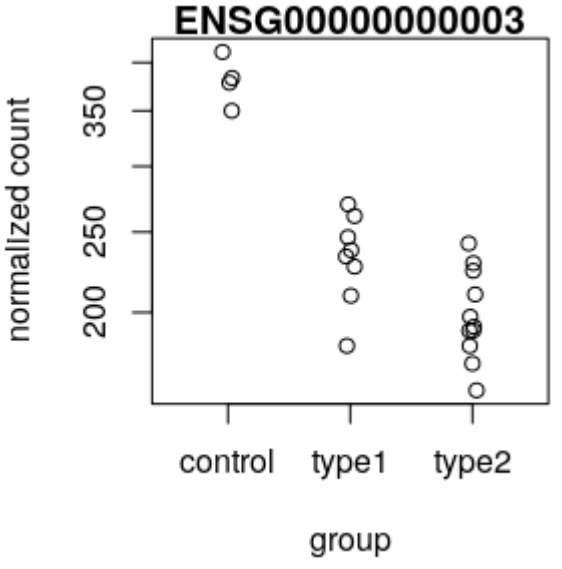
DGE

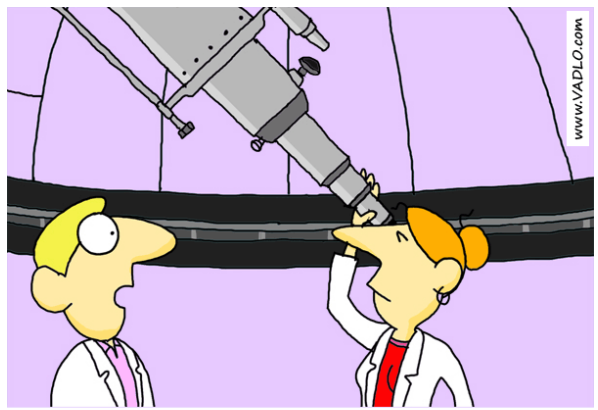
- MA plot `plotMA()`



- Volcano plot

- Normalised counts `plotCounts()`

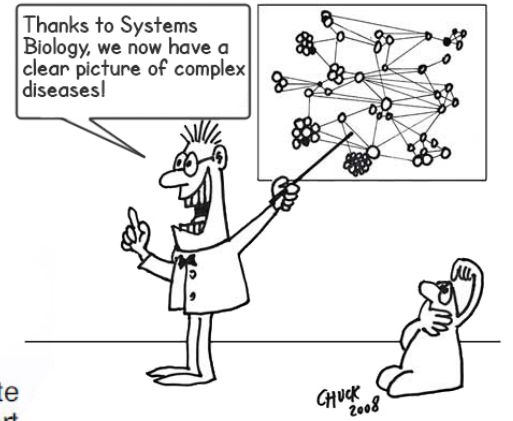


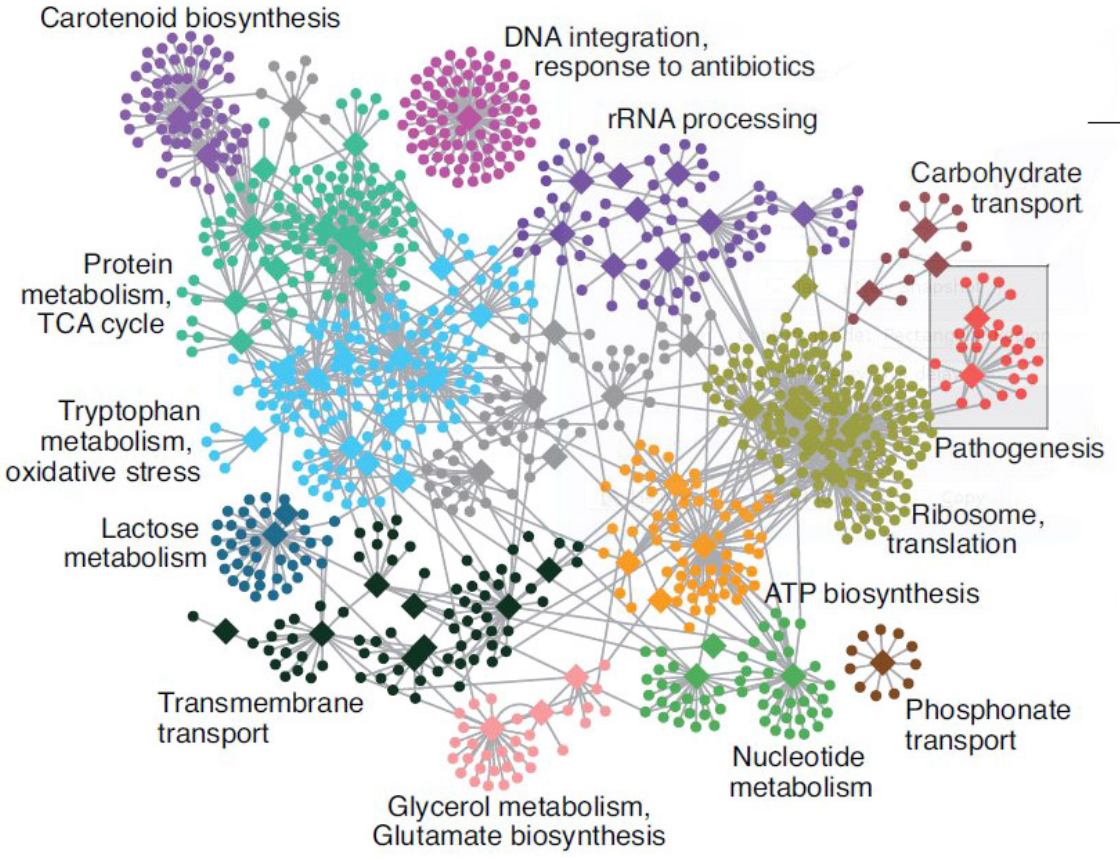


"Can you see the upper points of my scatter plot?"

Functional analysis • GO

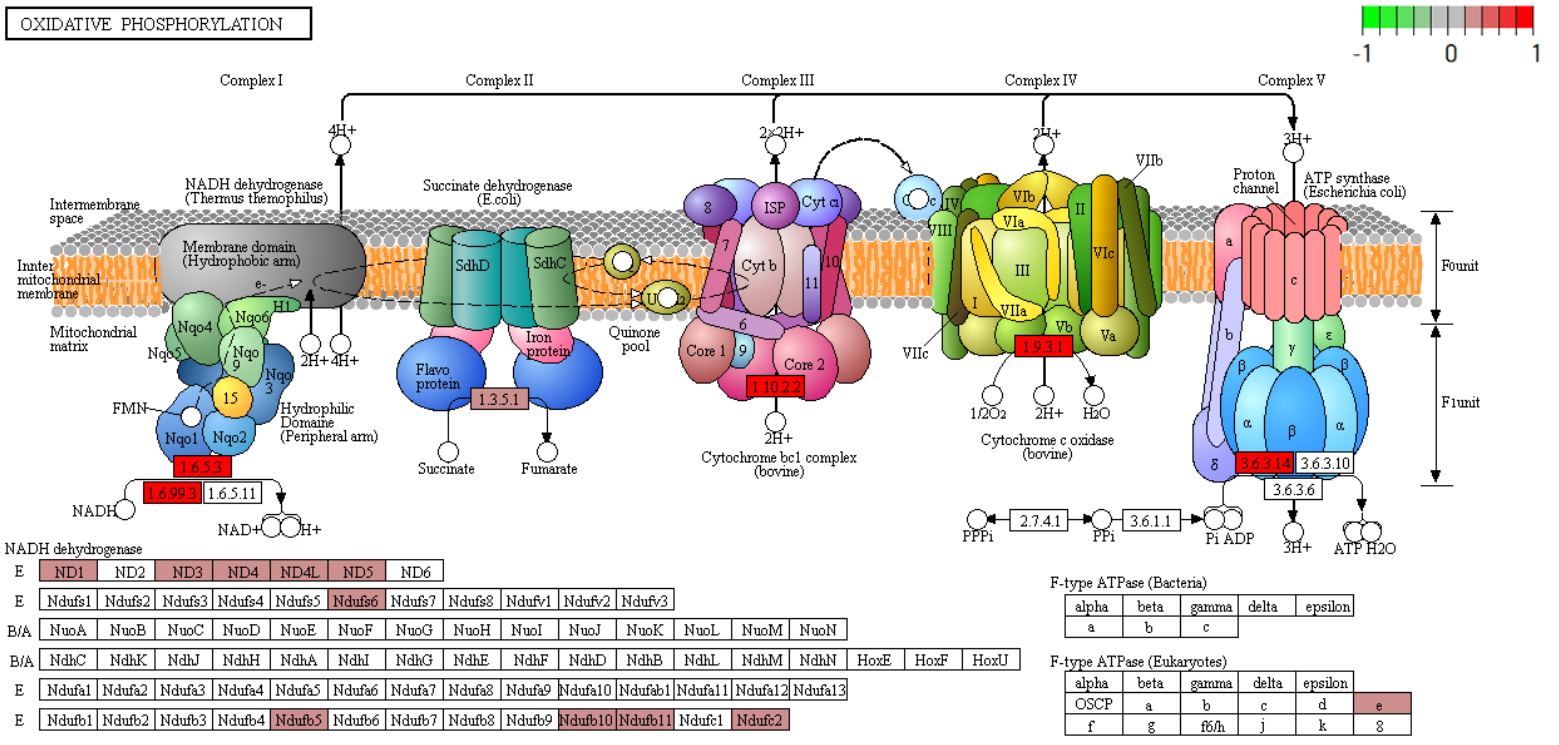
- Gene set analysis (GSA)
- Gene set enrichment analysis (GSEA)
- Gene ontology / Reactome databases





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
- 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?

R version 4.0.5 (2021-03-31)

Platform: x86_64-pc-linux-gnu (64-bit)

OS: Ubuntu 18.04.5 LTS

Built on : 25-May-2021 at 13:35:25

2021 • SciLifeLab • NBIS

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 DE 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/gXXXXXXX/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/gXXXX/nobackup/<user>/`

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