Transcriptome analysis with RNA sequencing

Isabelle Dupanloup
(slides taken from Paolo Angelino)

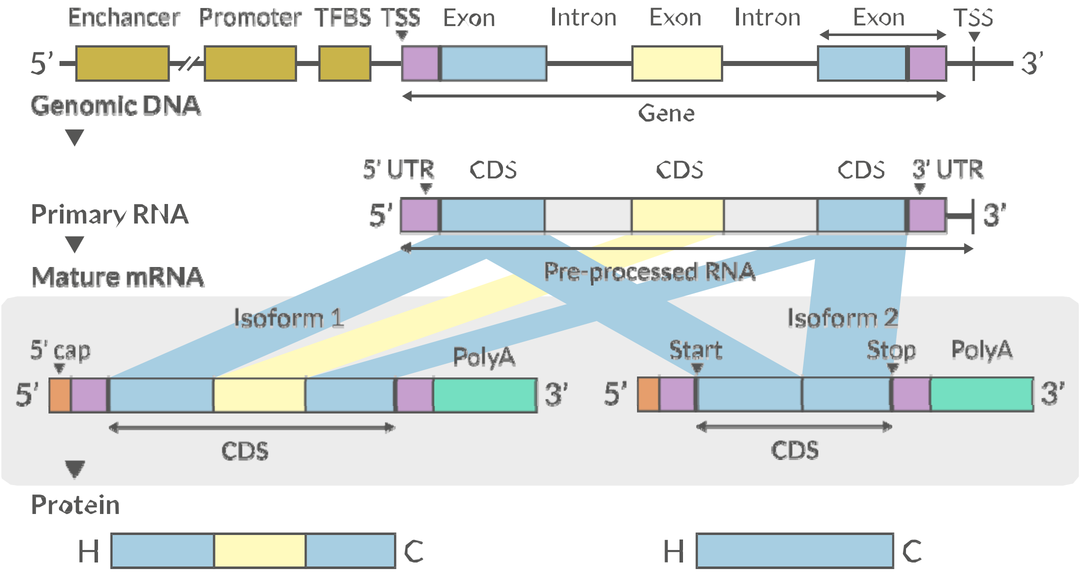
BCF - Bioinformatics Core Facility
SIB - Swiss Institute of Bioinformatics

Day I – Learning objectives

Transcriptomics goals
Experimental design Quality
Check Alignment Quantification

Part of the material presented here is borrowed from
John Garbe, RNAseq tutorial, https://www.msi.umn.edu/sites/default/files/RNA-Seq%20mod1v6.pdf
RNA-seqlopedia https://rnaseq.uoregon.edu/
RNA-seq Bioinformatics https://rnabio.org/
Darya Vanichkina, RNA-seq data analysis, https://sydney-informatics-hub.github.io/training-RNAseq-slides/01_IntroductionToRNASeq/01_IntroductionToRNASeq.html#1

RNAseq Challenges

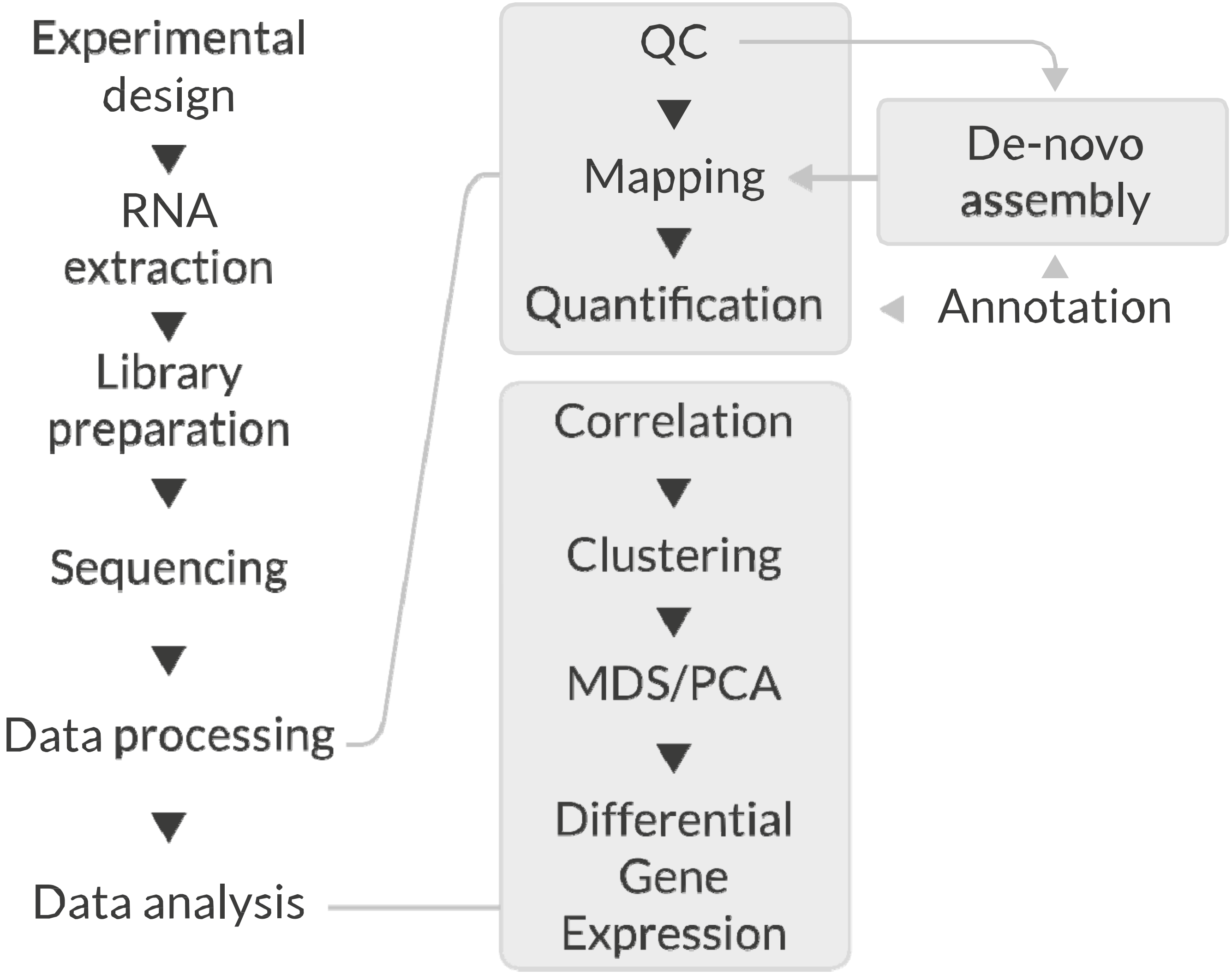


- Sample
  - Purity?, quantity?, quality?
- RNAs consist of small exons that may be separated by large introns
  - Mapping reads to genome is challenging
- The relative abundance of RNAs vary wildly
  - $10^5 – 10^7$ orders of magnitude
  - Since RNA sequencing works by random sampling, a small fraction of highly expressed genes may consume the majority of reads
  - Ribosomal genes
- RNAs come in a wide range of sizes
  - Small RNAs must be captured separately
  - PolyA selection of large RNAs may result in 3' end bias
- RNA is fragile compared to DNA (easily degraded)

Goals

- Gene expression and differential expression
- Alternative expression analysis
- Transcript discovery and annotation
- Allele specific expression
    - Relating to SNPs or mutations
- Mutation discovery
- Fusion detection
- RNA editing

Workflow



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

Experimental design

- Number of samples
- Starting material
- RNA selection
- Library preparation
- Sequencing depth
- Single vs paired-end reads

First question to ask: WHY are you sequencing????

What do you hope to find ? What follow-up experiments do you plan to do **after** the sequencing ?

RNA-seq can be used to carry out accurate analysis of:

- differential gene expression (DGEA)
- whole gene coexpression network analysis (WGCNA)
- alternative splicing (AS)
- novel transcript reconstruction and annotation
- allele-specific expression of variants
- RNA editing and other modifications
- ...

But all of these cannot be accurately analysed at the same time in ONE SINGLE experiment !!!

The main trade-off: replicates vs library depth

The cost of your experiment increases with:

- Number of replicates
- Sequencing depth
- Length of reads

Some analyses are impossible to do without sufficient library depth:

- alternative splicing (AS)
- novel transcript reconstruction and annotation *
- allele-specific expression of variants
- RNA editing and other modifications *

* special protocols have been developed to enrich for rare molecules, thus reducing the need for "brute force" increases in library depth

What is a replicate?

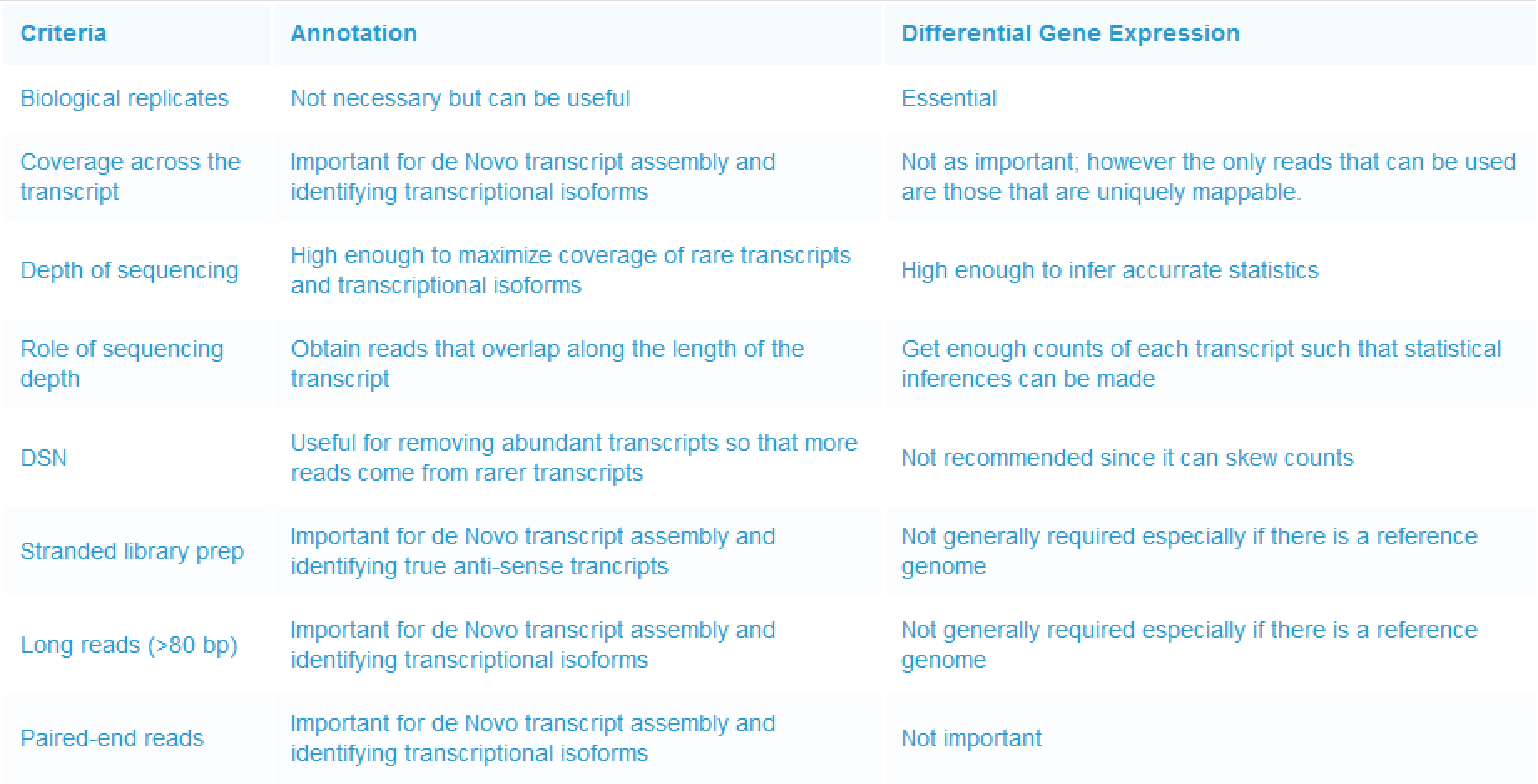
Technical replicate:

- same individual
- same cell line
- same iPSC/ESC clone
- typically, Pearson correlation coefficient > 0.9
- sometimes: same library, different flow cells

Biological replicate:

- different individuals
- different cell lines
- no clear filter for correlation coefficient, but usually < 0.9

Recommendations for RNA-seq options based upon experimental objectives



Criteria	Annotation	Differential Gene Expression
Biological replicates	Not necessary but can be useful	Essential
Coverage across the transcript	Important for de Novo transcript assembly and identifying transcriptional isoforms	Not as important; however the only reads that can be used are those that are uniquely mappable.
Depth of sequencing	High enough to maximize coverage of rare transcripts and transcriptional isoforms	High enough to infer accurrate statistics
Role of sequencing depth	Obtain reads that overlap along the length of the transcript	Get enough counts of each transcript such that statistical inferences can be made
DSN	Useful for removing abundant transcripts so that more reads come from rarer transcripts	Not recommended since it can skew counts
Stranded library prep	Important for de Novo transcript assembly and identifying true anti-sense trancripts	Not generally required especially if there is a reference genome
Long reads (>80 bp)	Important for de Novo transcript assembly and identifying transcriptional isoforms	Not generally required especially if there is a reference genome
Paired-end reads	Important for de Novo transcript assembly and identifying transcriptional isoforms	Not important

https://rnaseq.uoregon.edu/

How much of an effect could library depth vs number of replicates have?

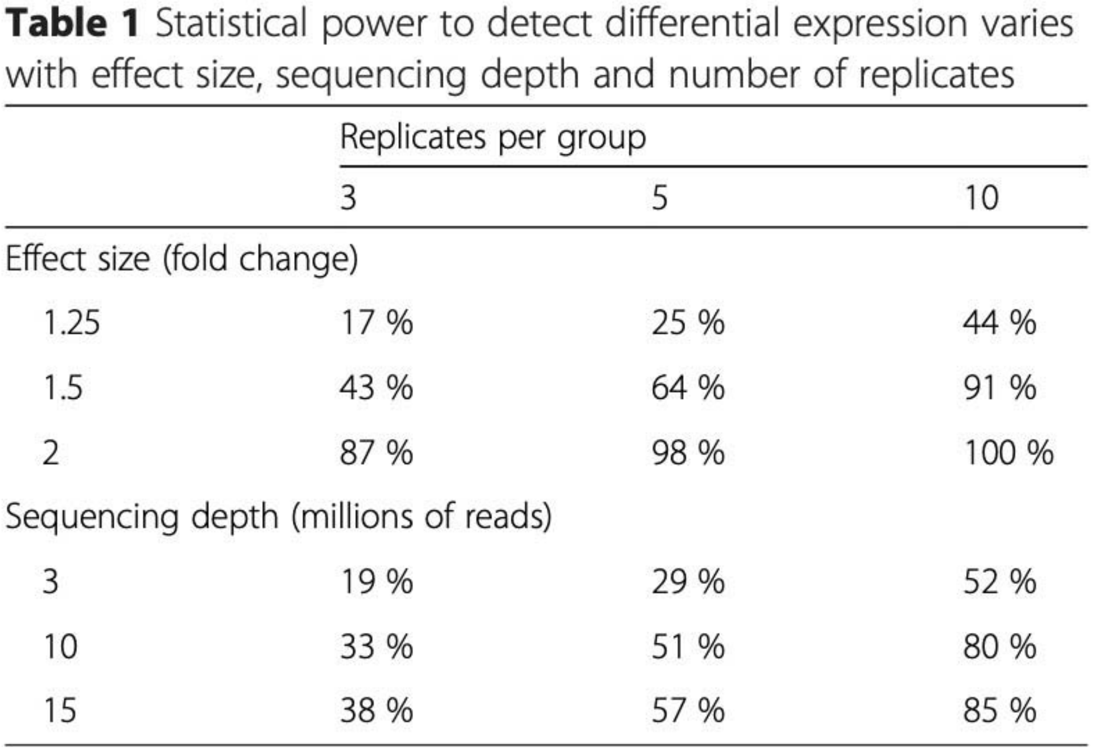
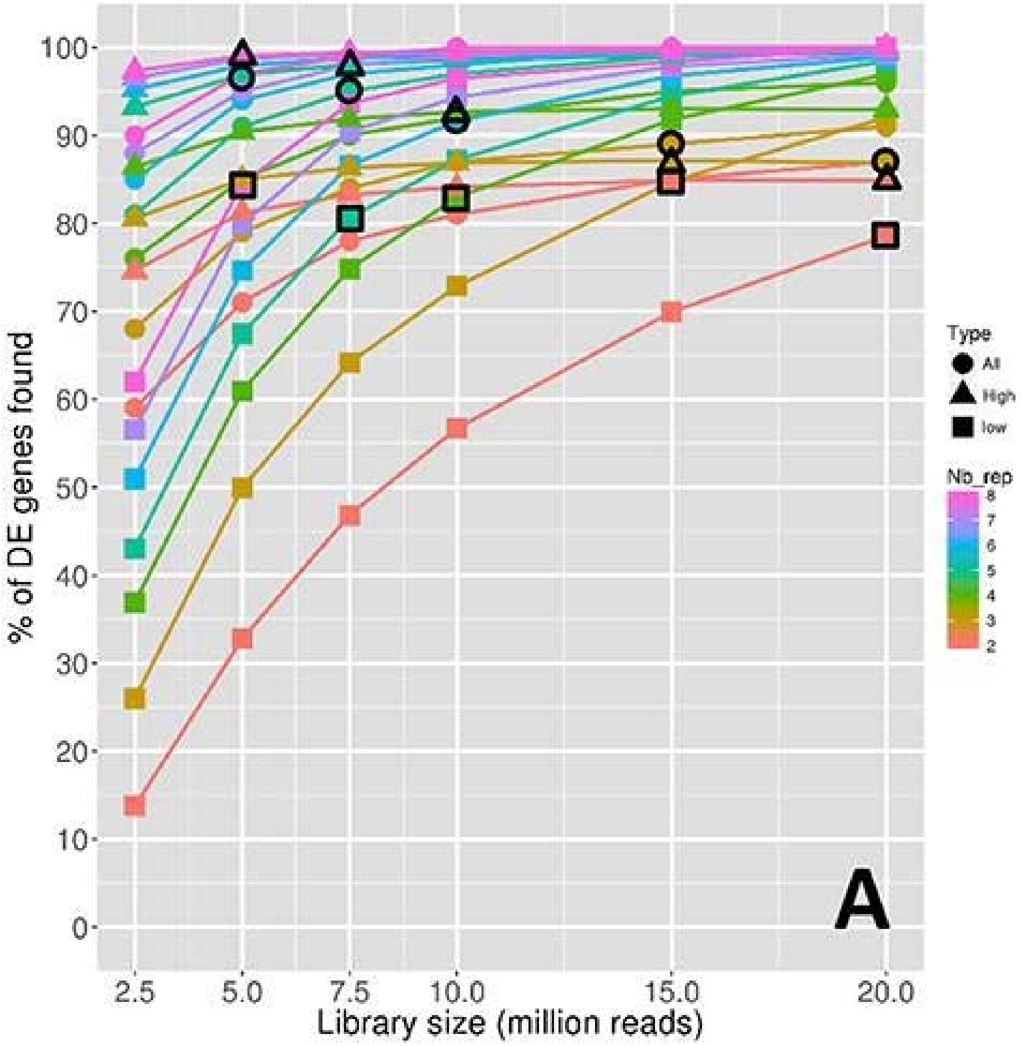


Table 1 Statistical power to detect differential expression varies with effect size, sequencing depth and number of replicates

	Replicates per group		
	3	5	10
Effect size (fold change)			
1.25	17 %	25 %	44 %
1.5	43 %	64 %	91 %
2	87 %	98 %	100 %
Sequencing depth (millions of reads)			
3	19 %	29 %	52 %
10	33 %	51 %	80 %
15	38 %	57 %	85 %

Conesa et al. (2016)
A survey of best practices for RNA-seq data analysis
Genome Biology

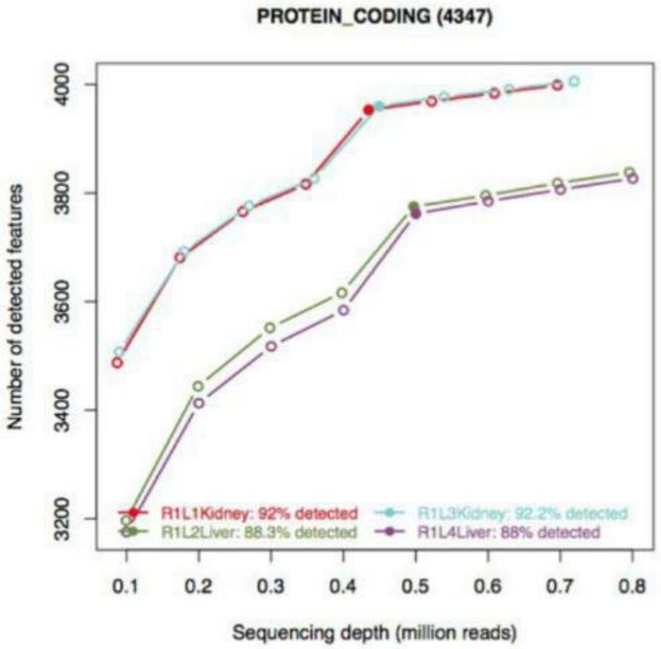
How much of an effect could library depth vs number of replicates have?



Lamarre, S., *et al.* "Optimization of an RNA-Seq Differential Gene Expression Analysis Depending on Biological Replicate Number and Library Size", Front. Plant Sci., 14 February 2018 | https://doi.org/10.3389/fpls.2018.00108

How many reads do we need?

Saturation plots -- gene detection (NOISeq)



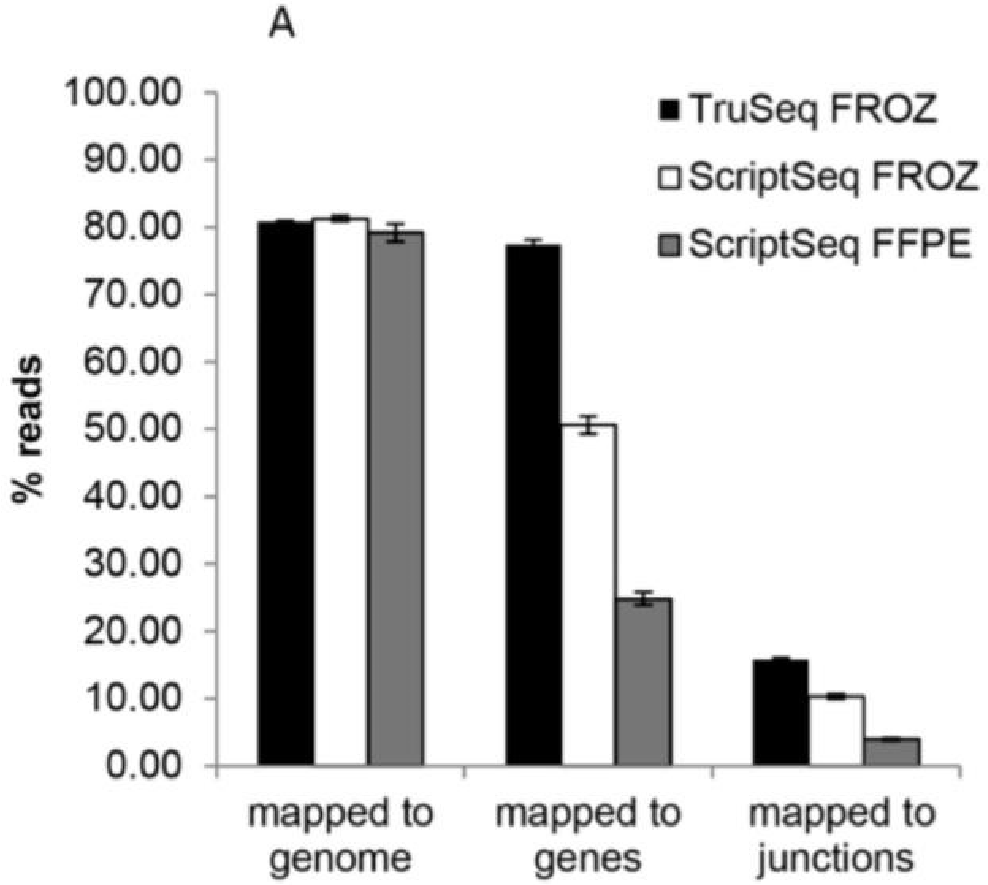
- More genes are detected with larger sequencing depth.
- At a certain point, the curve flattens out.

How many replicates do we need?

- Technical replicates not necessary (Marioni *et al.*, 2008)
- Biological replicates: 6 - 12 (Schurch *et al.*, 2016)
- 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)
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

The input material matters



Norton et al. PLoS One 2013

Assessing RNA quality

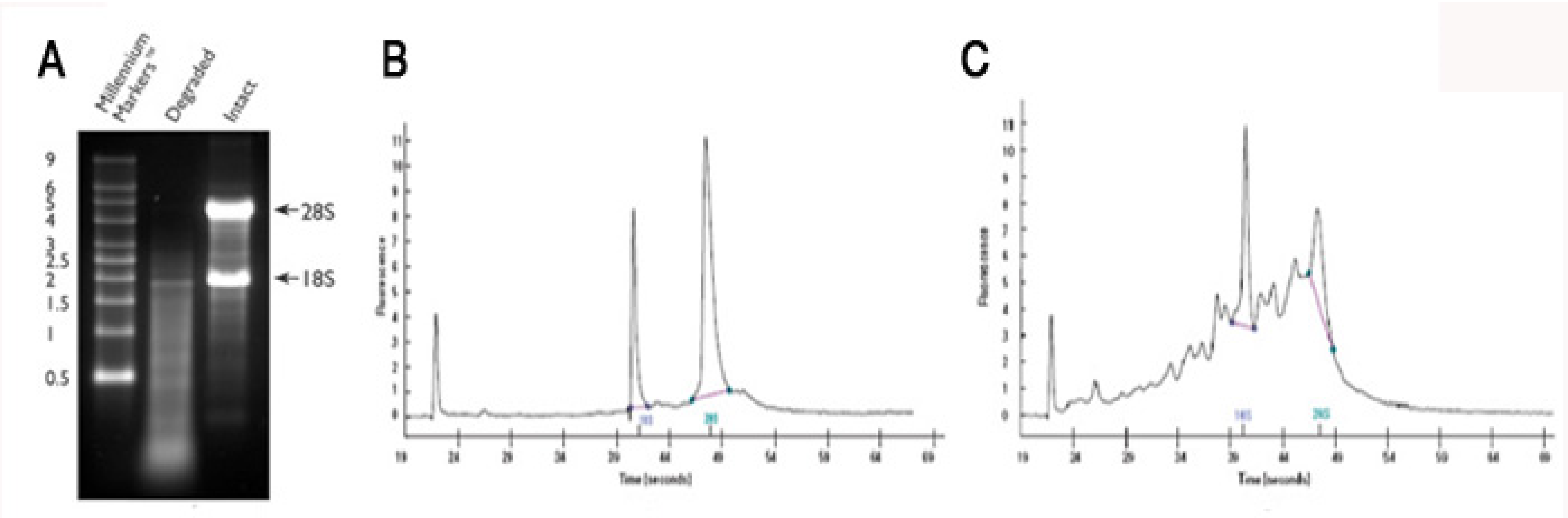
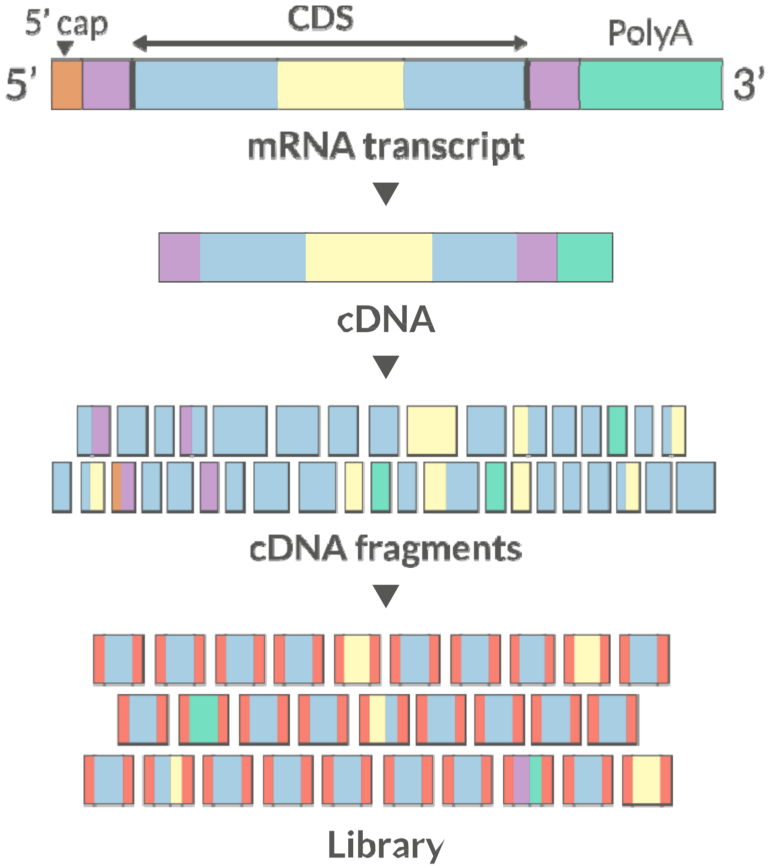


Figure taken from https://rnaseq.uoregon.edu/

Library preparation

- PolyA selection
- rRNA depletion
- Size selection
- PCR amplification
- Stranded (directional) libraries
  - Accurately identify sense/antisense transcript
  - Resolve overlapping genes



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

What kind of RNA do we want to study ?

- Most (90%) of the RNA in a human cell is ribosomal RNA.
- If we are only interested in mRNA, the Ribosomal RNA is just "rubbish ", occupying a lot of the reads.
- But we may also be interested in other types of RNA (miRNA, ncRNA, snoRNA, tRNA, ...).

poly(A) enrichment

- In eukaryotes, polyadenylation (i.e., addition of a poly(A) tail to a transcript) is part of the process that produces mature mRNA that is then translated into proteins.
- Can we use this to extract only the protein-coding mRNA from our RNA pool ?

- Hybridize the RNA to oligo-dT beads.
- Wash away everything that does not hybridize.
- This leaves only the RNA with poly(A) tails.

- Not all protein-coding genes have poly(A) tails.
- With this approach we lose all other types of RNA.
- Can we just get rid of the ribosomal RNA?

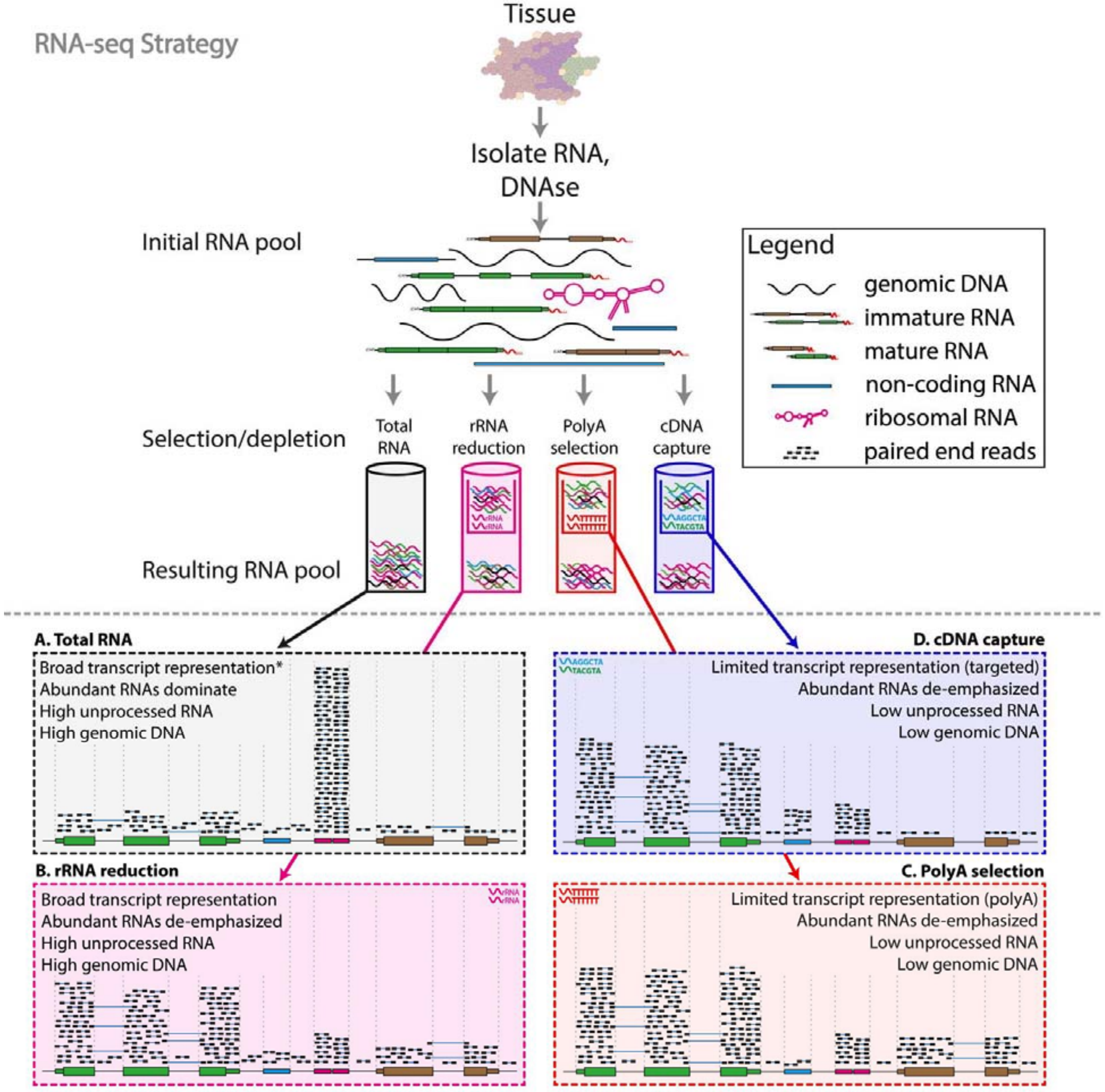
Ribominus/ ribozero protocols

- Aim at selective depletion of ribosomal RNA.
- Hybridization to specific rRNA probes.
- Keep only what doesn't hybridize.
- Degraded rRNA may not be removed.
- Ribozero removes more rRNA than ribominus.

What if we want to focus on small RNA ?

- Specific kits exist for keeping only small RNA.
- Size selection.
- Usually no fragmentation necessary.

RNA sequence selection/depletion



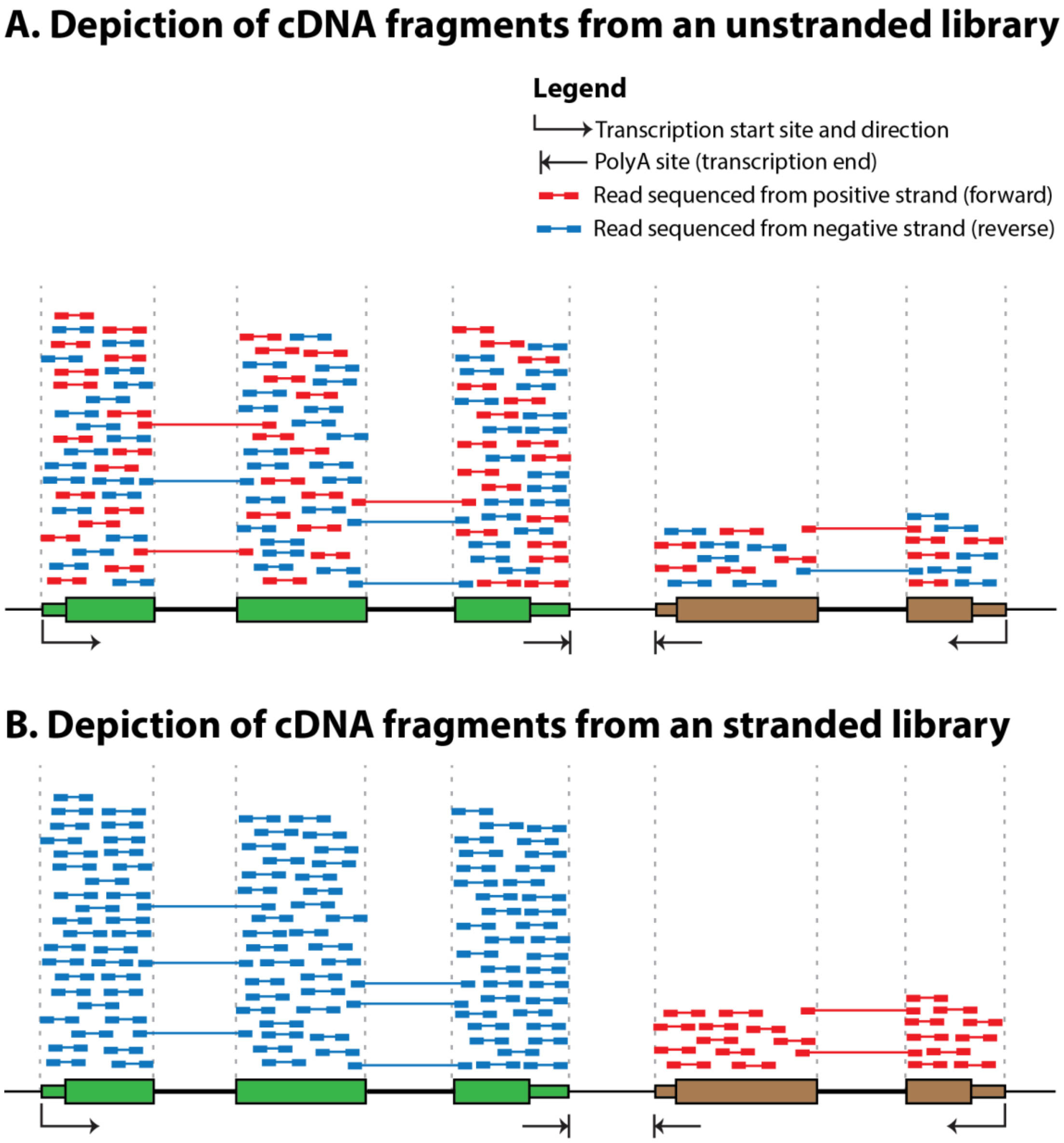
Directional Strand-Specific RNA-seq

- Preserve the strandness information to determine transcript orientation

- Critical for novel transcript discovery and annotation, especially for
  - Non-coding RNA
  - Overlapping genes in lower organisms like bacteria

- Improve alignment of reads to genome or transcriptome

Strand specific Library



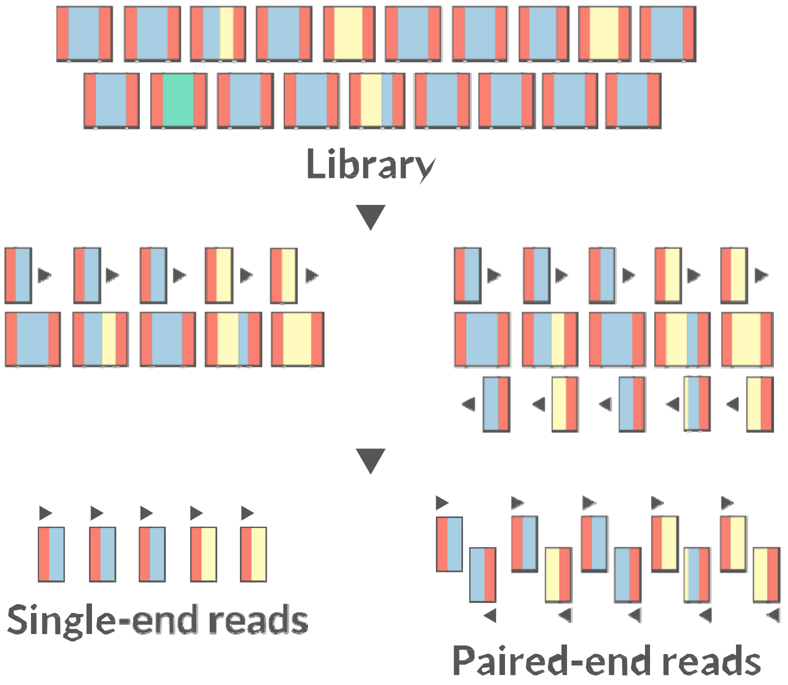
A. Depiction of cDNA fragments from an unstranded library

B. Depiction of cDNA fragments from an stranded library

Legend
Transcription start site and direction
PolyA site (transcription end)
🟥 Read sequenced from positive strand (forward)
🟦 Read sequenced from negative strand (reverse)

Sequencing

- 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

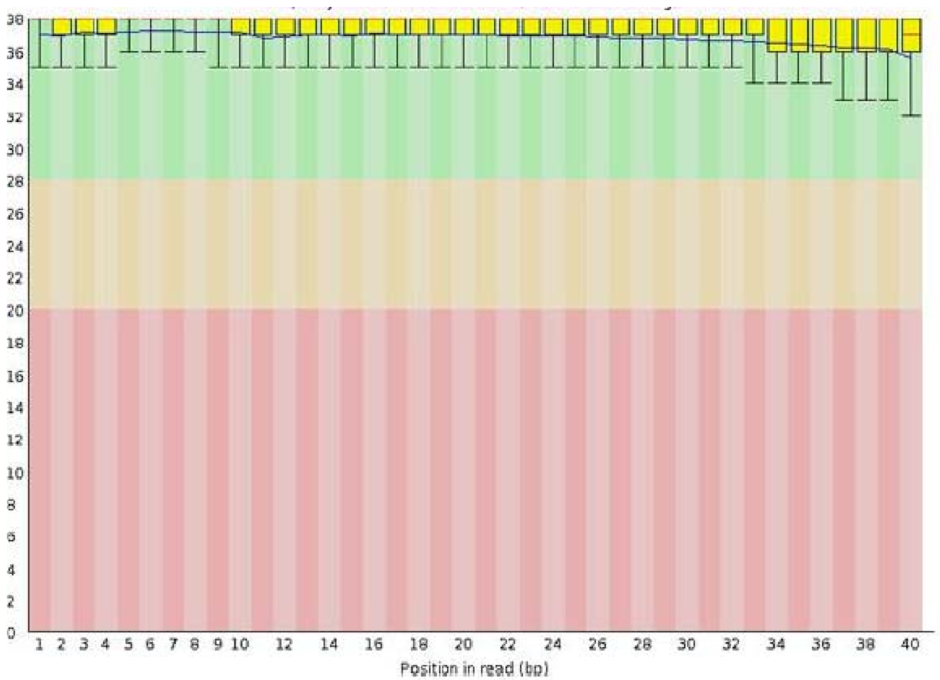
Corley, Susan M., et al. "Differentially expressed genes from RNA-Seq and functional enrichment results are affected by the choice of single-end versus paired-end reads and stranded versus non-stranded protocols." BMC genomics 18.1 (2017): 399

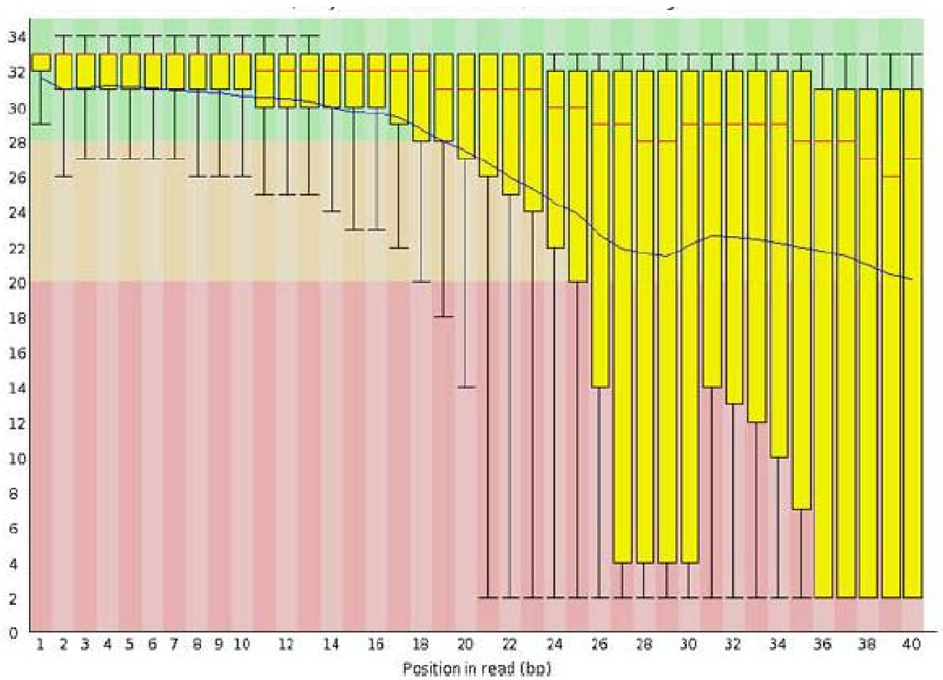
Liu, Yuwen, Jie Zhou, and Kevin P. White. "RNA-seq differential expression studies: more sequence or more replication?." Bioinformatics 30.3 (2013): 301-304 Comparison of PE and SE for RNA-Seq, SciLifeLab

Pre-alignment QC

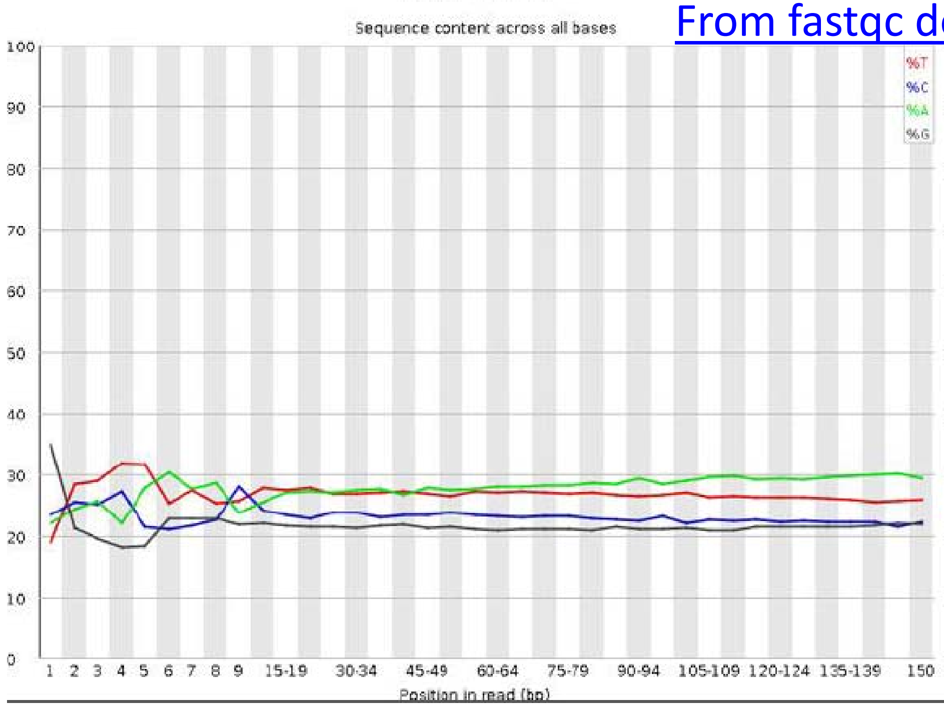
- Number of reads
- Per base sequence quality
- Per sequence quality score
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence length distribution
- Sequence duplication levels
- Overrepresented sequences
- Adapter content
- Kmer content

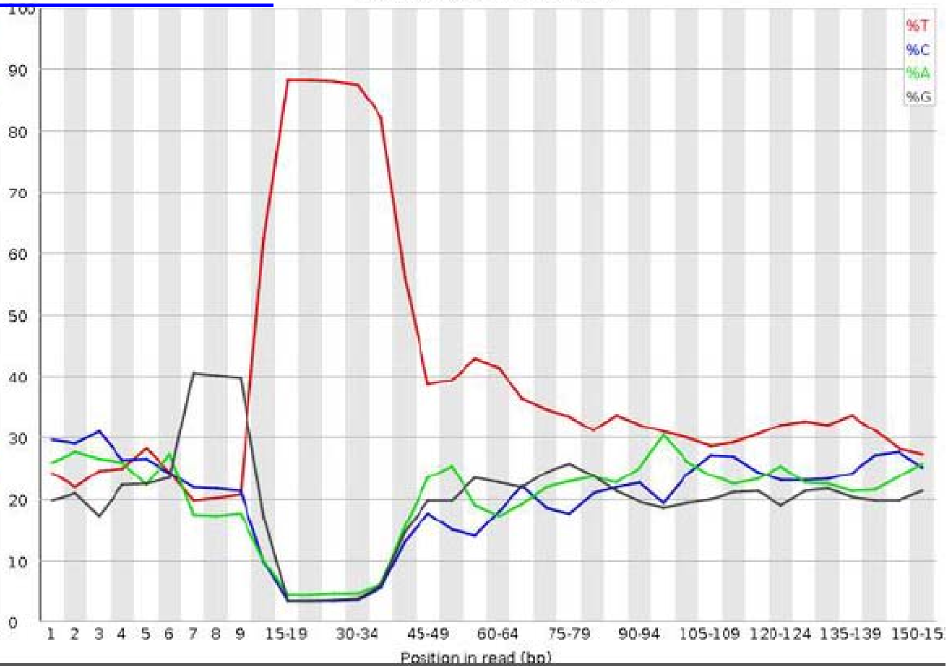
FastQC, MultiQC
https://sequencing.qcfail.com/

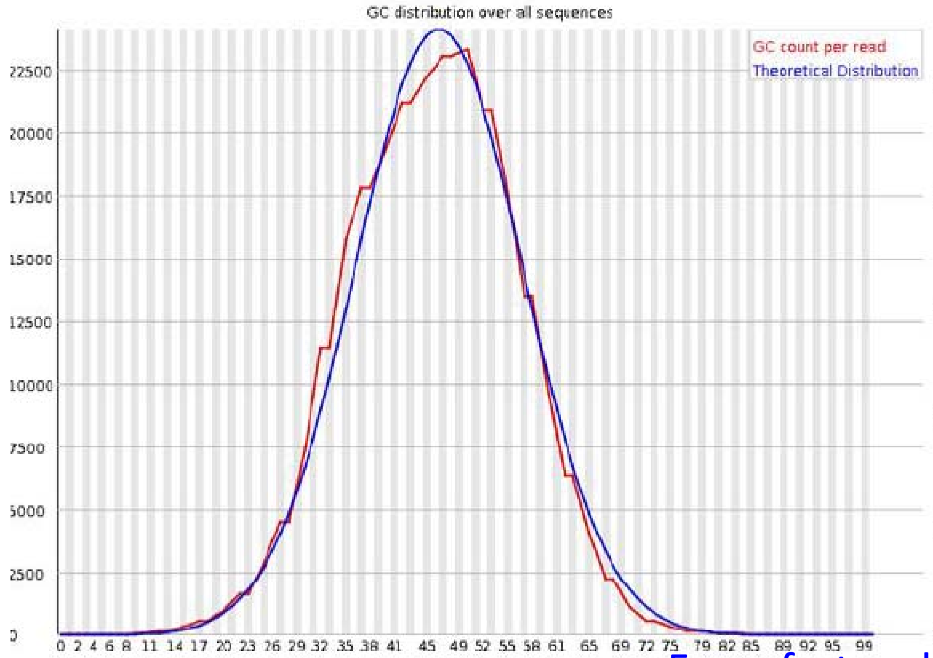


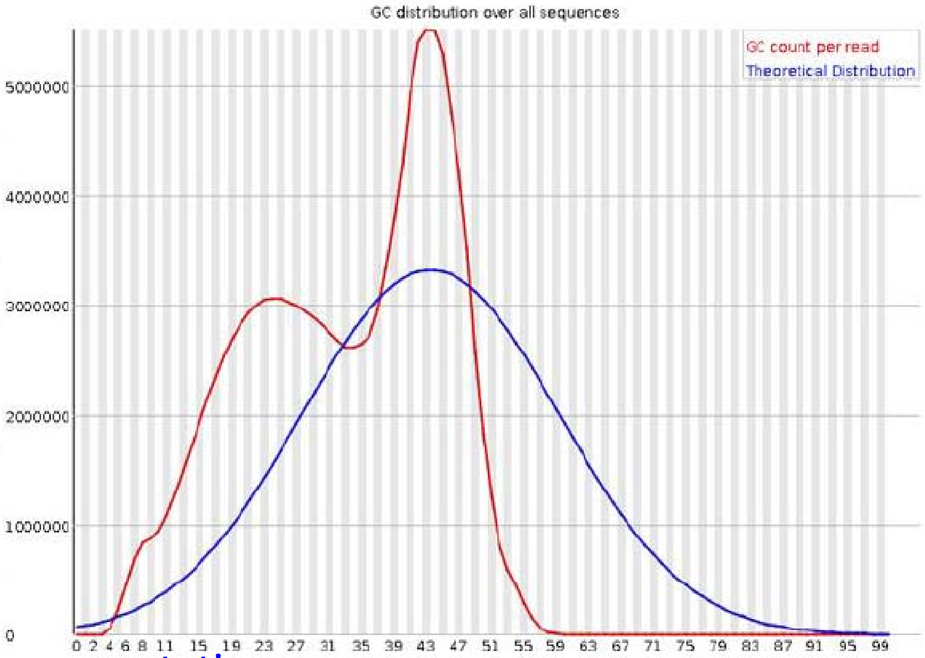


From fastqc documentation

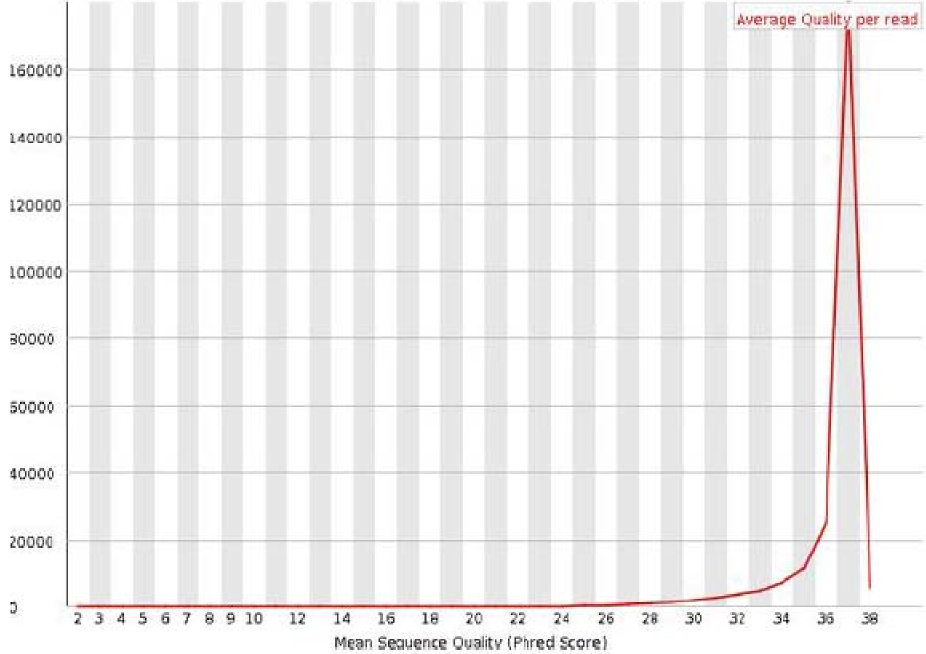


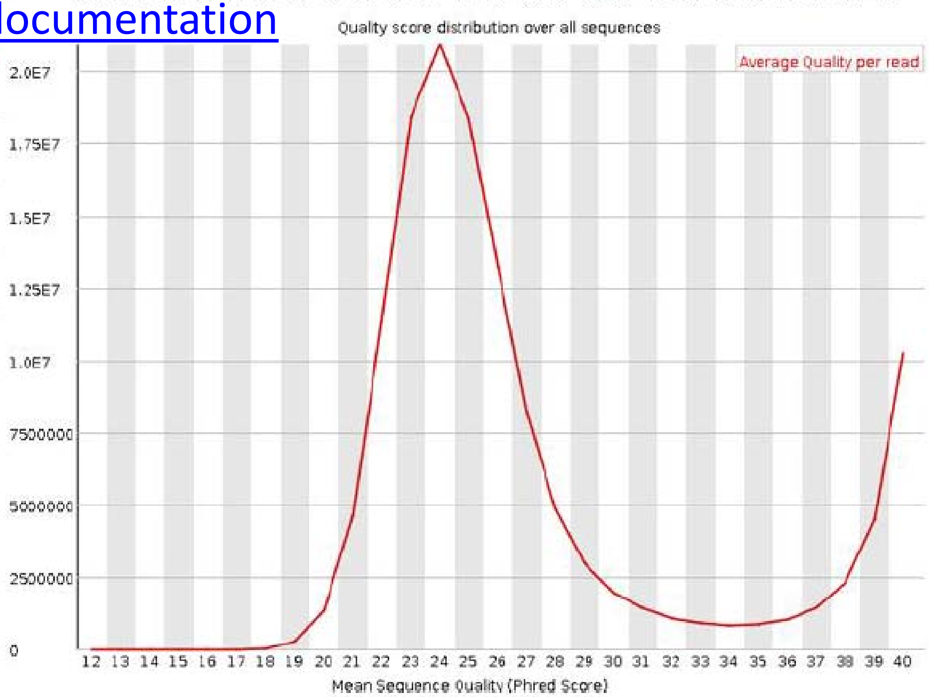


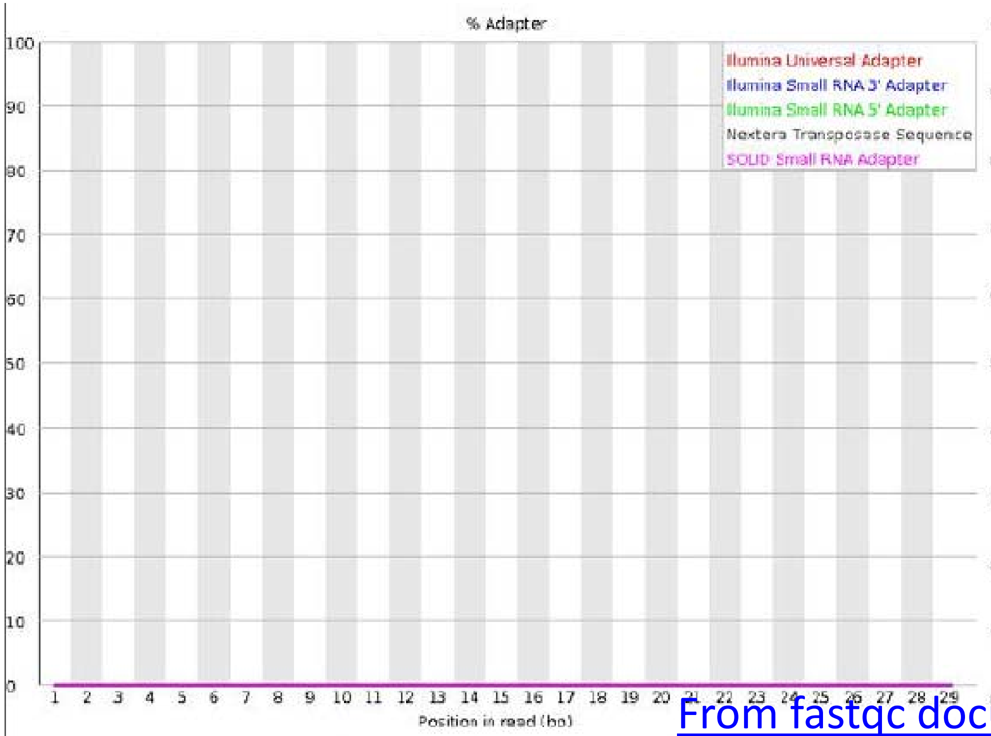


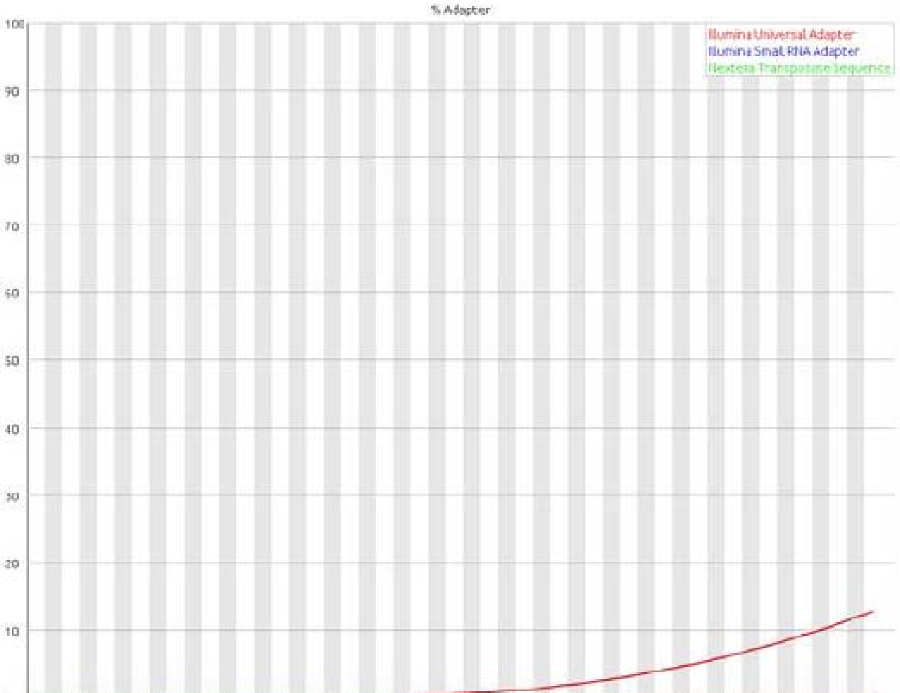


From fastqc documentation

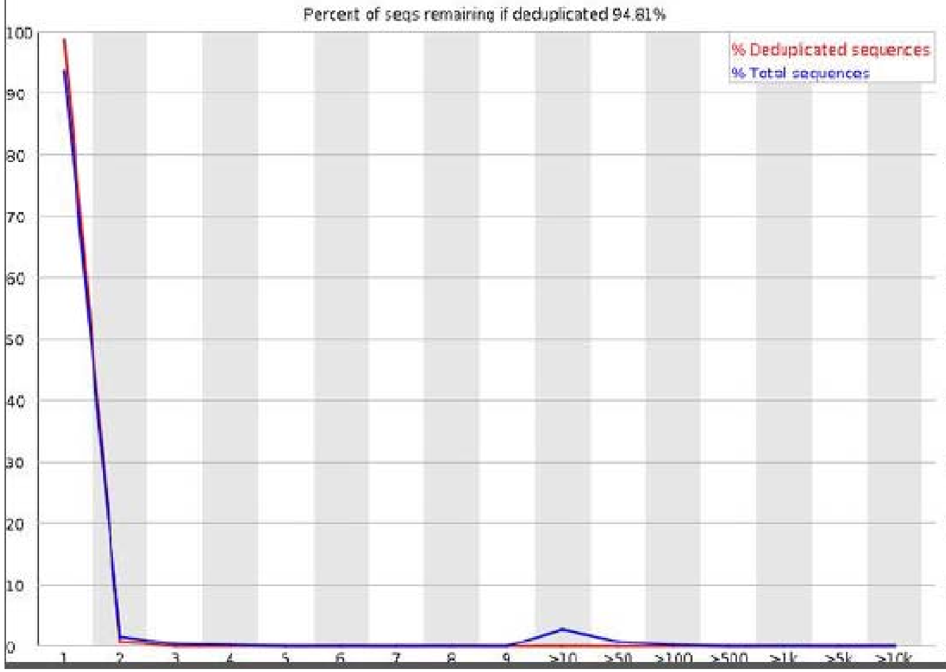


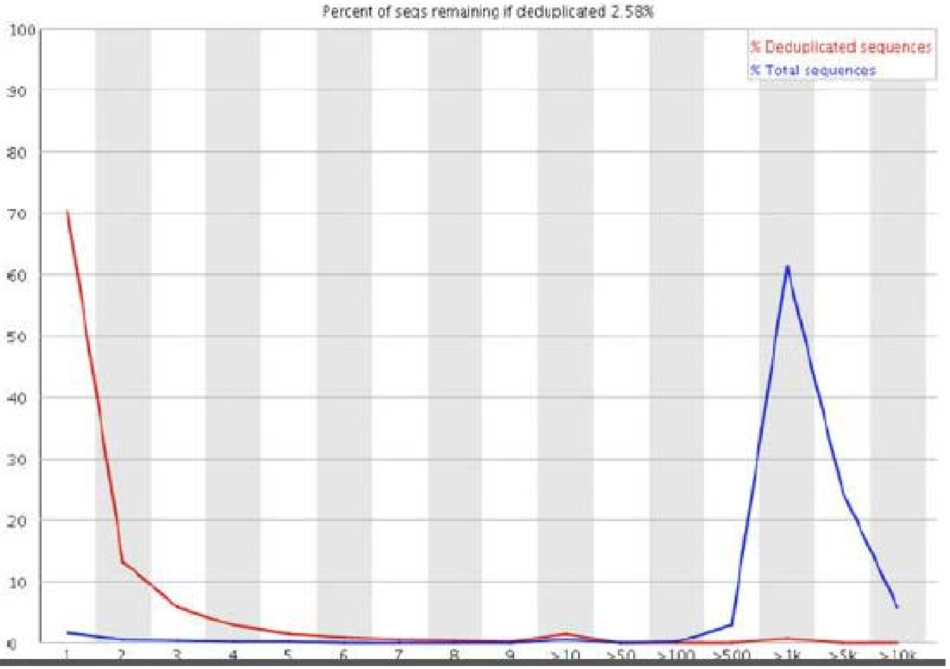






From fastqc documentation

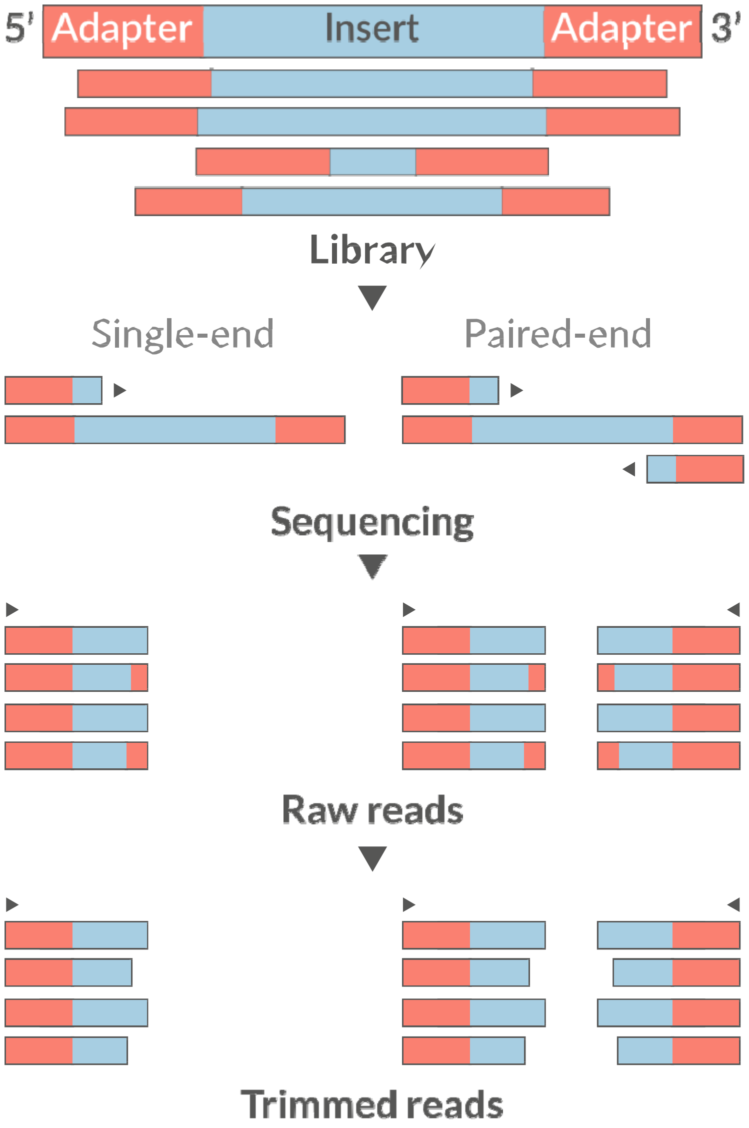




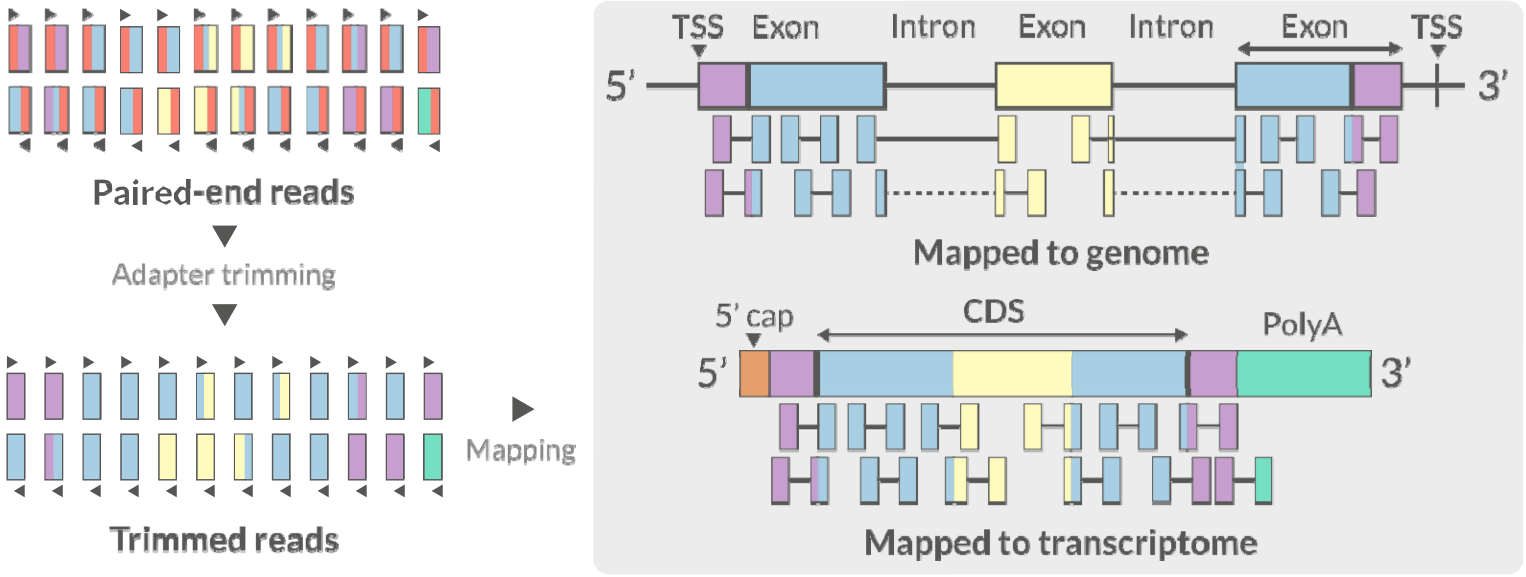
Trim

- Trim/Clip/Filter reads
- Remove adapter sequences
- Trim reads by quality
- Filter by min/max read length
    - Remove reads less than ~18nt

Cutadapt, fastp, Skewer, Prinseq



Mapping

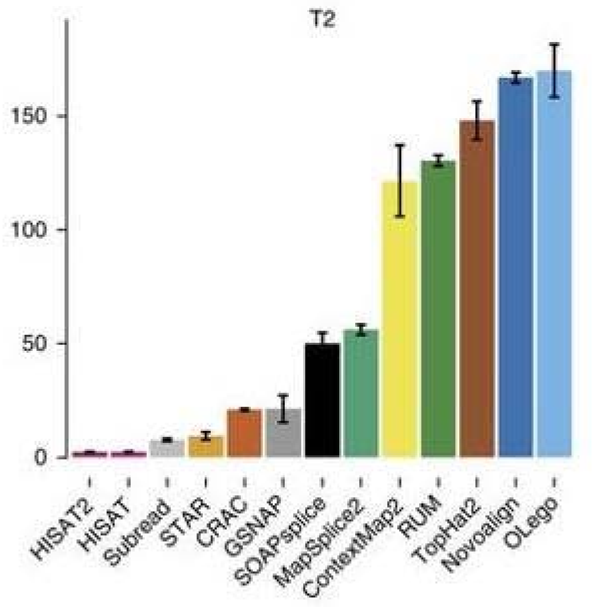


- Aligning reads back to a reference sequence
- Mapping to genome vs transcriptome
- Splice-aware alignment (genome)

STAR, HiSat2, GSNAP, Novoalign (Commercial)

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

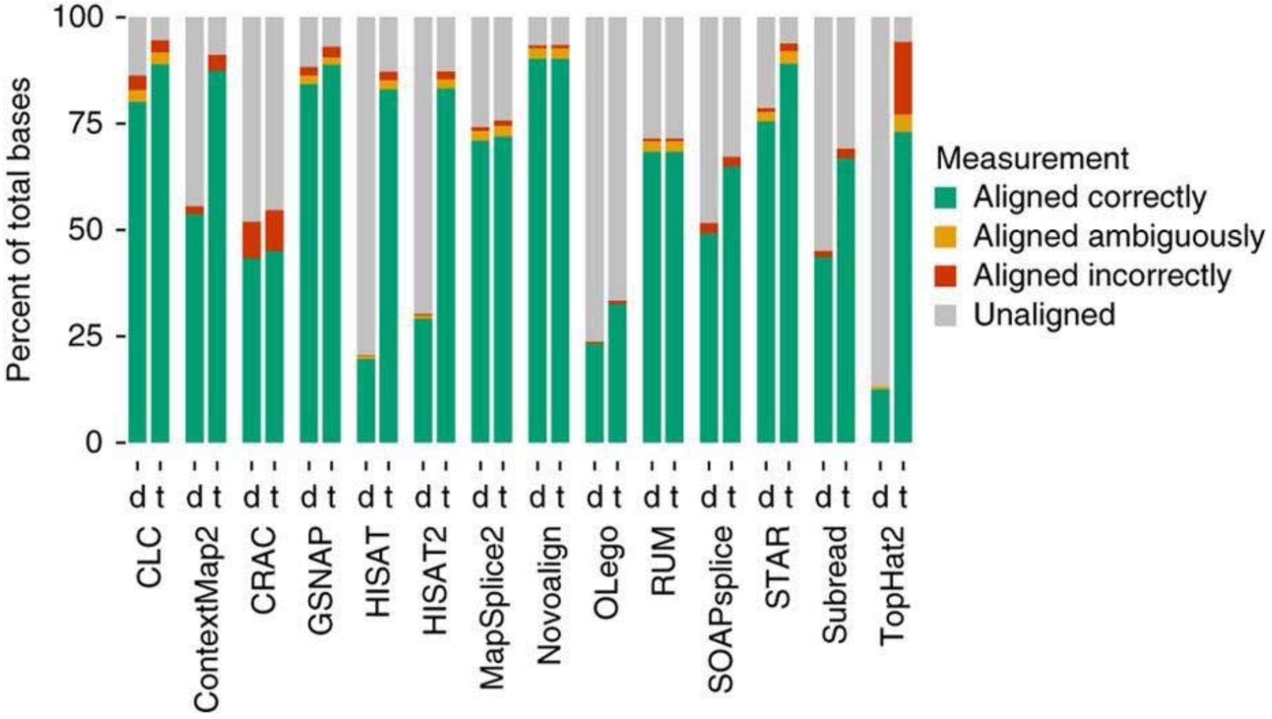
Aligner speed

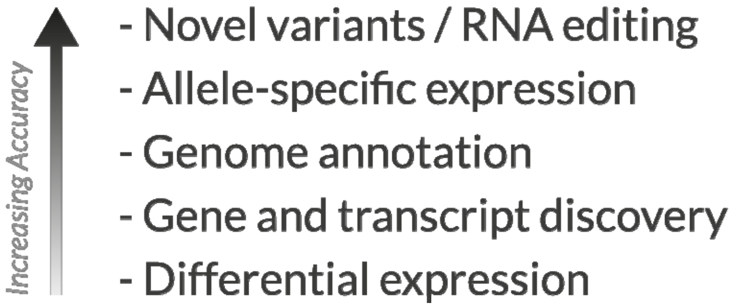


Program	Time_Min	Memory_GB
HISATx1	22.7	4.3
HISATx2	47.7	4.3
HISAT	26.7	4.3
STAR	25	28
STARx2	50.5	28
GSNAP	291.9	20.2
TopHat2	1170	4.3

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

Aligner accuracy





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

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

FASTQ

Reads (FASTQ)

```
@ST-E00274:179:HHYMLALXX:8:1101:1641:1309 1:N:0:NGATGT
NCATCGTGGTATTTGCACATCTTTTCTTATCAAATAAAAAGTTTAACCTACTCAGTTAT
GCGCATACGTTTTTTGATGGCATTTCCATAAACCGATTTTTTTTTTTATGCACGTACCCA
AAACGTGCAGAAAAATACGCTGCTAGAAATGTA
+
#AAAFAFA<-AFFJJJAFA-FFJJJJFFFAJJJJ-<FFJJJ-A-F-7--FA7F7-----
FFFJFA<FFFFJ<AJ--FF-A<A-<JJ-7-7-<FF-FFFJAFFAA--A--7FJ-7----
77-A--7F7)---7F-A----7)7-----7<<-
```

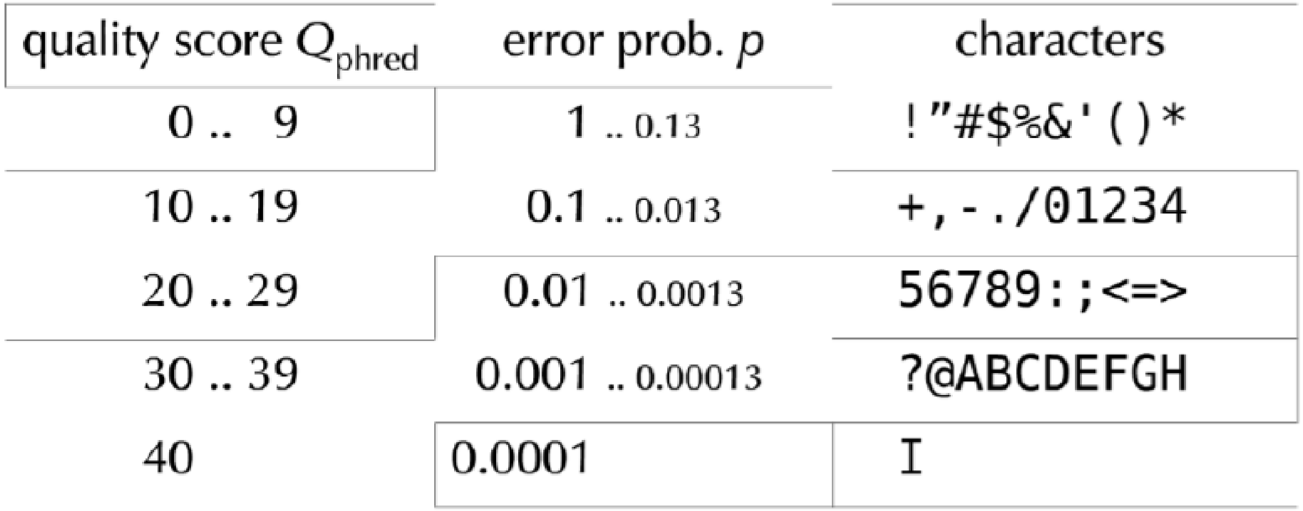
@instrument:runid:flowcellid:lane:tile:xpos:ypos
read:isfiltered:controlnumber:sampleid

FASTQ format: quality string

- If p is the probability that the base call is wrong, the Phred score is:

$$Q = -10 \log_{10} p$$

- The score is written with the character whose ASCII code is Q+ 33 (Sanger Institute standard).



quality score Q_{phred}	error prob. *p*	characters
0 .. 9	1 .. 0.13	!”#$%&'()*
10 .. 19	0.1 .. 0.013	+,-./01234
20 .. 29	0.01 .. 0.0013	56789:;<=>
30 .. 39	0.001 .. 0.00013	?@ABCDEFGH
40	0.0001	I

FASTA + GTF

Reference Genome/Transcriptome (FASTA)

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

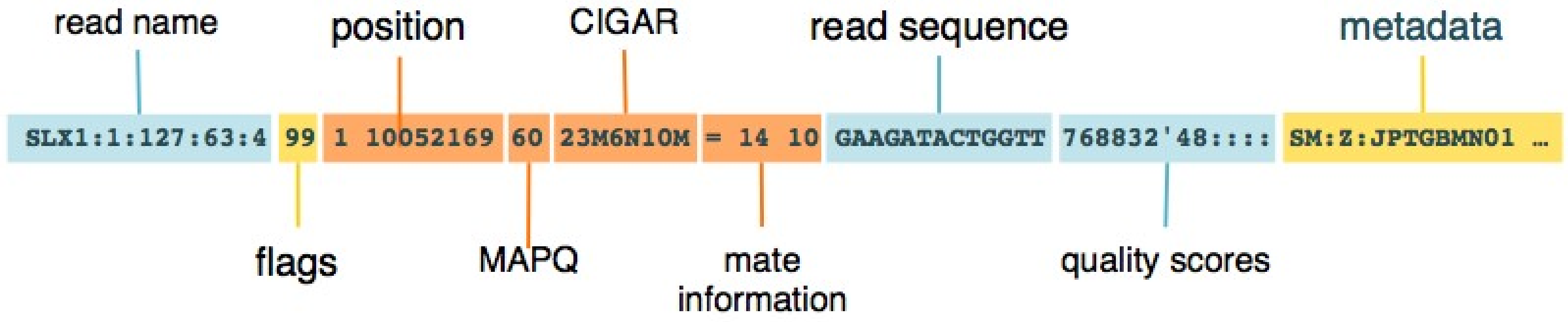
Annotation (GTF/GFF)

```
#!genome-build GRCz10
#!genebuild-last-updated 2016-11
4 ensembl_havana gene 6732 52059 . - . gene_id "ENSDARG00000104632";
gene_version "2"; gene_name "rerg"; gene_source "ensembl_havana"; gene_biotype
"protein_coding"; havana_gene "OTTDARG00000044080"; havana_gene_version "1";
```

seq source feature start end score strand
frame attribute

Alignment: SAM/BAM (Sequence Alignment Map format)

HEADER containing metadata (sequence dictionary, read group definitions etc)
RECORDS containing structured read information (1 line per read record)



read name

position

CIGAR

read sequence

metadata

```
SLX1:1:127:63:4 99 1 10052169 60 23M6N10M = 14 10 GAAGATACTGGTT 768832'48:::: SM:Z:JPTGBMN01 …
```

flags

MAPQ

mate information

quality scores

```
ST-E00274:188:H3JWNCCXY:4:1102:32431:49900 163 1 1 60 8S139M4S = 385 535
TATTTAGAGATCTTAAACATCCATTCCCCCTGCAAACATTTTCAATCATTACATTGTCATTTTCCC
TCCAAATTAAATTTAGCCAGAGGCGCACAACATACGACCTCTAAAAAAGGTGCTGGAACATGT
ACCTATATGCAGCACCACCATC AAAFAFFAFFFFJ7FFFFJ<JAFA7F-<AJ7JJ<FFFJ--
<FAJF<7<7FAFJ-<AFA<-JJJ-AF-AJ-FF<F--A<FF<-7777-7JA-77A---F-7AAFF-FJA--77FJ<--
77)))7<JJA<J77<-------<7--))7)))7- NM:i:4 MD:Z:12T0T40C58T25 AS:i:119 XS:i:102
XA:Z:17,-53287490,4S33M4D114M,11; MQ:i:60 MC:Z:151M RG:Z:ST-
E00274_188_H3JWNCCXY_4
```

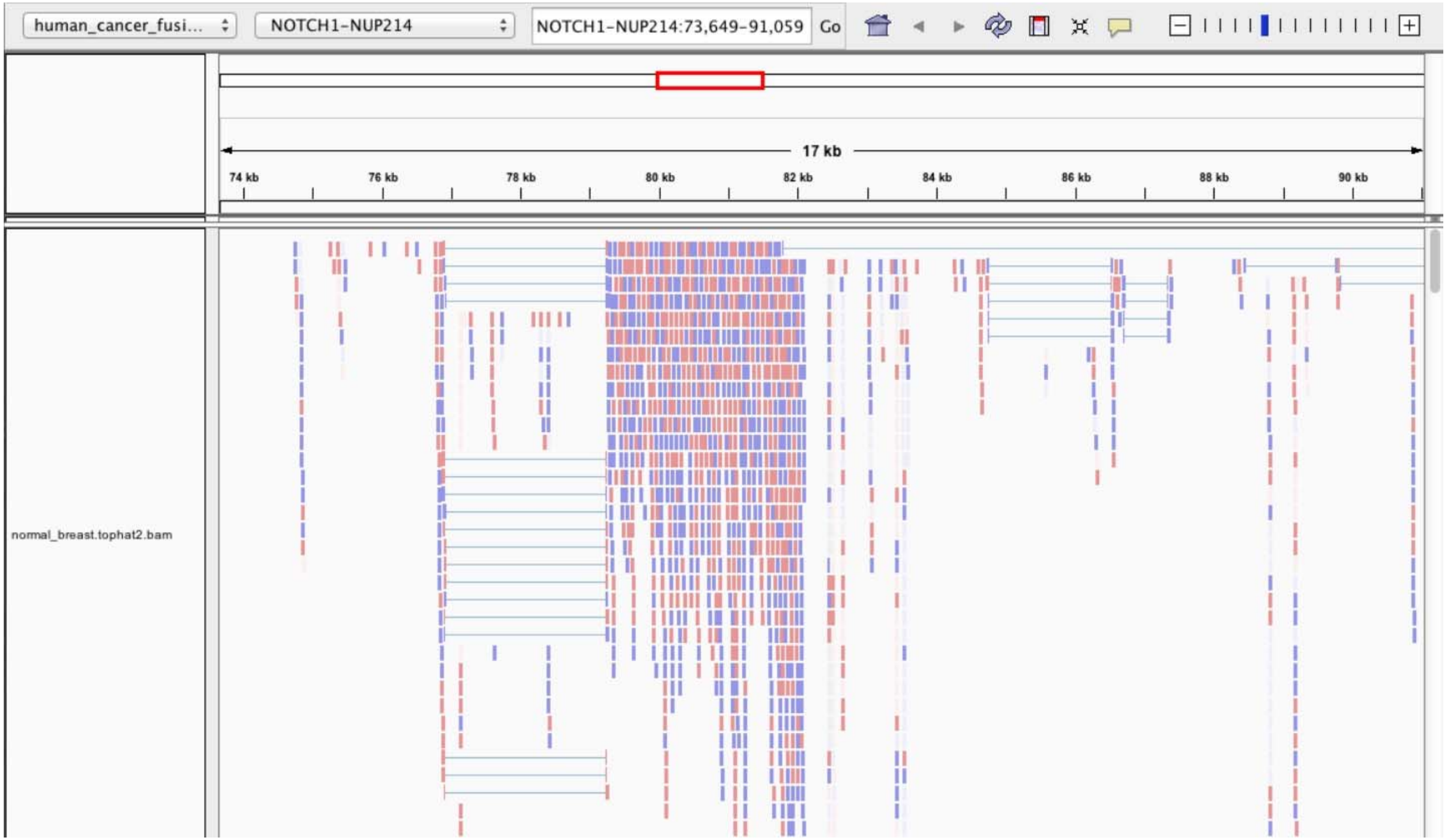
Sam file format: header

- reference chromosomes and their lengths
- which tool was used to generate the data
- what parameters were used (exactly what is reported depends on mapper)

```
samtools view -H filename.bam
```

```
@HD VN:1.4 SO:coordinate
@SQ SN:chr1 LN:248956422
@SQ SN:chr2 LN:242193529
@SQ SN:chr3 LN:198295559
@SQ SN:chr4 LN:190214555
@SQ SN:chr5 LN:181538259
@SQ SN:chr6 LN:170805979
<... more chromosomes and lengths>
@PG ID:STAR PN:STAR VN: # etc etc
@CO user command line:...
```

Visualization



human_cancer_fusi... NOTCH1-NUP214 NOTCH1-NUP214:73,649-91,059 Go

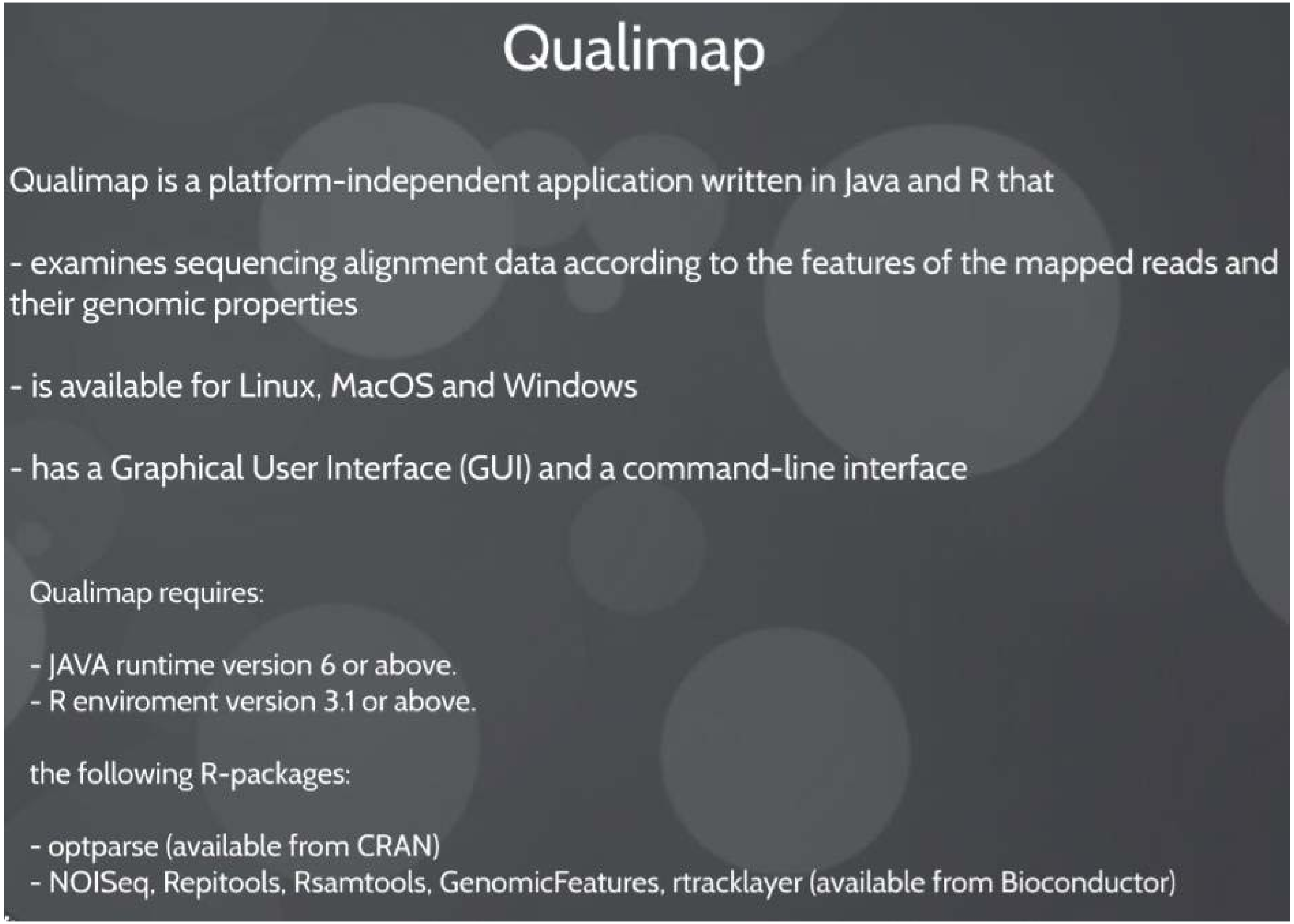
IGV, UCSC Genome Browser

Post-alignment QC

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

QoRTs, RSeQC, Qualimap

Post-alignment QC



Qualimap

Qualimap is a platform-independent application written in Java and R that

- examines sequencing alignment data according to the features of the mapped reads and their genomic properties

- is available for Linux, MacOS and Windows

- has a Graphical User Interface (GUI) and a command-line interface

Qualimap requires:

- JAVA runtime version 6 or above.
- R enviroment version 3.1 or above.

the following R-packages:

- optparse (available from CRAN)
- NOISeq, Repitools, Rsamtools, GenomicFeatures, rtracklayer (available from Bioconductor)

Qualimap examples

RNA-seq QC

Analysis of RNA-seq data (kidney.bam, human.64.gtf): QualiMap HTML report.

Counts QC

Counts QC HTML reports computed from RNA-seq experiment analyzing influence of D-Glucosamine on mice. The analysis was performed for 6 samples in 2 conditions - GlcN positive and negative (mouse_counts_ensembl.txt):

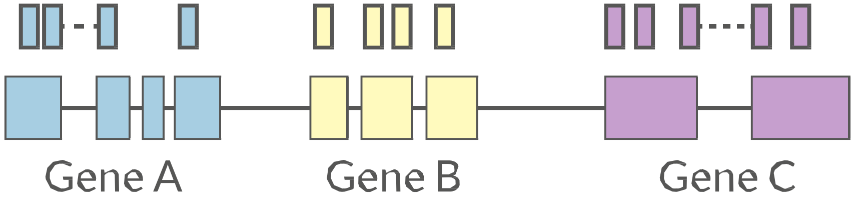
- Global report
- Comparison of conditions

Quantification - Counts

- Read counts = gene expression
- Reads can be quantified on any feature

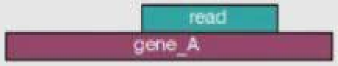
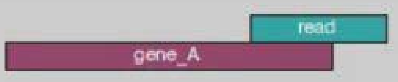
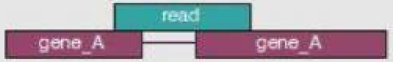
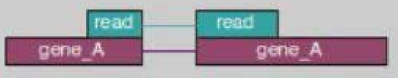
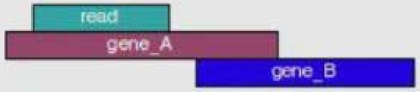
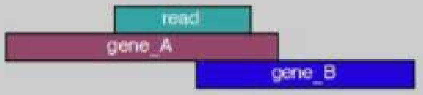
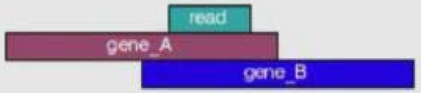
(gene, transcript, exon etc)

- Intersection on gene models
- Gene/Transcript level



featureCounts, HTSeq

How to handle reads overlapping several features

	union	intersection _strict	intersection _nonempty
	gene_A	gene_A	gene_A
	gene_A	no_feature	gene_A
	gene_A	no_feature	gene_A
	gene_A	gene_A	gene_A
	gene_A	gene_A	gene_A
	ambiguous	gene_A	gene_A
	ambiguous	ambiguous	ambiguous

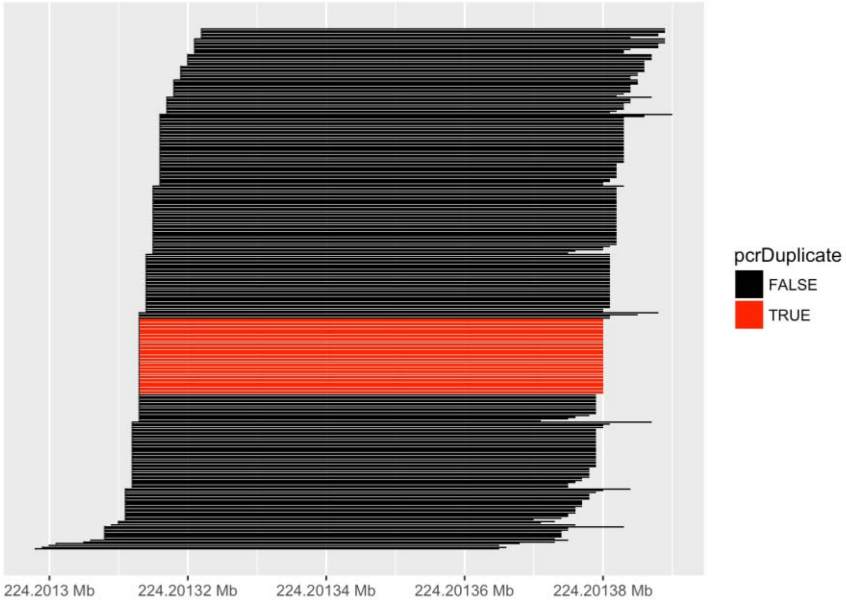
Quantification

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)
- 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

RSEM, Kallisto, Salmon, Cufflinks2

Kallisto, Salmon

- Ultra-fast & alignment-free
- Subsampling & quantification confidence
- Transcript-level estimates improves gene-level estimates
- Kallisto/Salmon > transcript-counts > tximport > gene-counts

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

A count table

very high-dimensional data: few samples,
many "parameters"

Gene	Sample1	Sample2	Sample 3
ENSG00000237613.2	10	12	9
ENSG00000268020.3	0	0	0
ENSG00000240361.2	2	7	7
ENSG00000186092.6	0	0	0
ENSG00000238009.6	0	0	0
ENSG00000239945.1	1092	987	432
ENSG00000233750.3	0	0	0
...	0	0	0
56000+ more rows ...			