RNA biology II

Genomics Bio5488

Guoyan Zhao

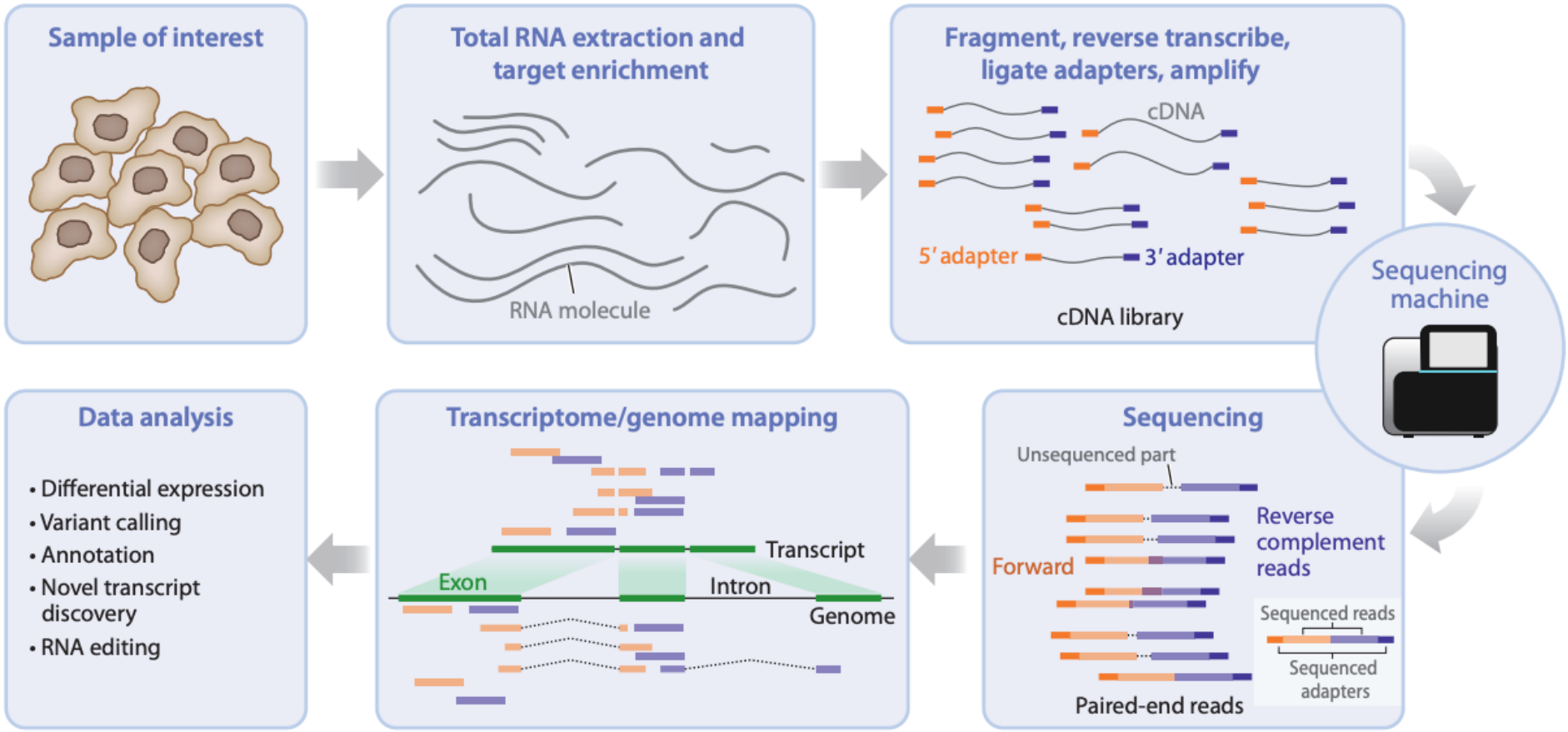
Associate Professor of Genetics & Neurology

Recap …

- RNA biology I
  - Introduction of RNA type and function
  - RNA quantification technology
    - Targeted
      - Northern blotting
      - *In situ* Hybridization
      - RT-qPCR
      - Reporter Assay
    - High-throughput
      - Microarray
      - RNA-seq
      - Single cell/nucleus RNA-seq
  - Experimental design principles
  - General workflow for RNA quantification
- RNA biology II
  - Transcriptome profiling by RNA-seq
    - Experimental design
    - RNA quantification
    - Quality control and normalization
    - Outlier detection
    - Differentially expressed gene detection
    - Result interpretation

RNA-seq and its applications

RNA sequencing (RNA-seq) is a genomic technique that uses next-generation sequencing to analyze the quantity and presence of RNA molecules in a biological sample.



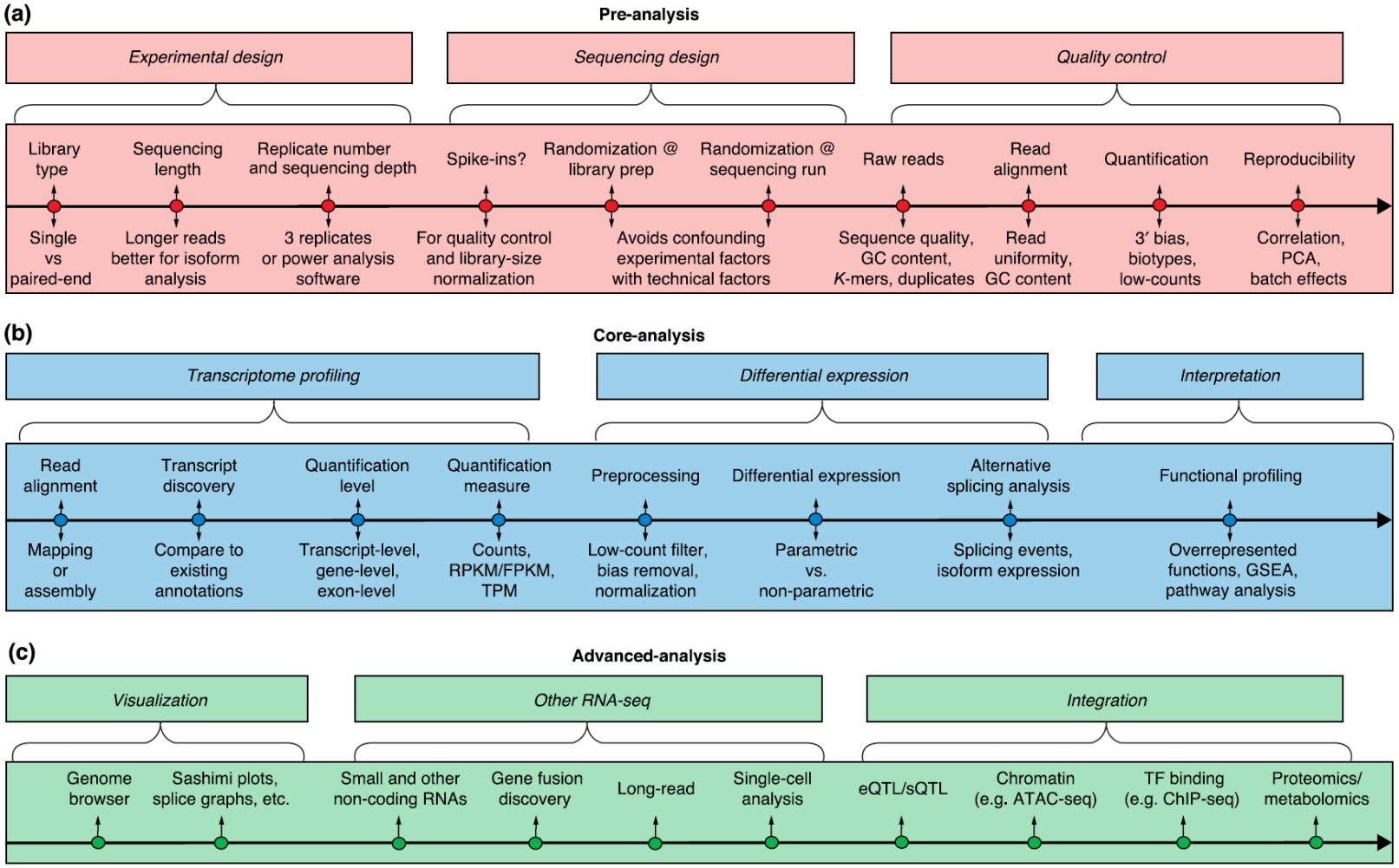
Korf *Nature Methods 10: 1165-1166* (2013)

RNA-seq applications

- differential gene expression between two or more experimental groups
- assessing allele-specific expression
- quantifying alternative transcript usage
- discovering novel transcripts
- detect gene fusions
- ribosomal profiling to measure the translation of each transcript
- RNA structure
- RNA–RNA and RNA–protein interactions
- …

Korf *Nature Methods 10: 1165-1166* (2013)

Roadmap for RNA-seq



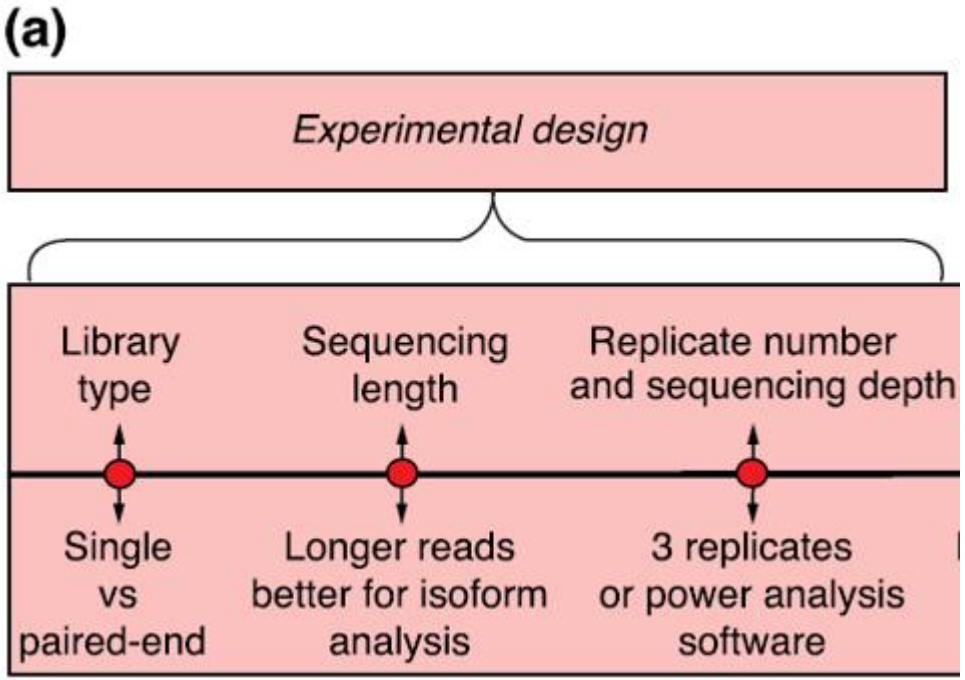
- What's the biological question?
- Do I have a reference genome to map to?
- Do I need to do a special RNA prep?
- Do I have enough samples?
- What kind of sequencing should I do?
- How many reads do I need?

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

Experimental design is critical

"Seventy percent of whether your experiment will work is determined before you touch the first test tube … "

Sun *Nat Rev Mol Cell Biol* 5:577-581 (2004)



What's the biological question?

- Differential gene expression between two or more experimental groups
- Isoform analysis
  - Characterizing transcriptome complexity
- Non-coding RNAs
  - siRNAs, microRNAs, enhancer associated RNAs
- Gene fusion discovery
  - Hybrid genes formed by two previously separate genes
- Novel isoform discovery
  - long-read sequencing
- Single-cell analysis

The library type

- What's my target?
  - whole transcriptome
  - mRNA
  - small RNA
  - miRNA

- strand-preserving libraries (strand-specific or directional RNA-Seq libraries)
  - libraries that preserve the orientation of RNA molecules.
  - allows researchers to distinguish between the sense (coding) and antisense (non-coding) strands of RNA.

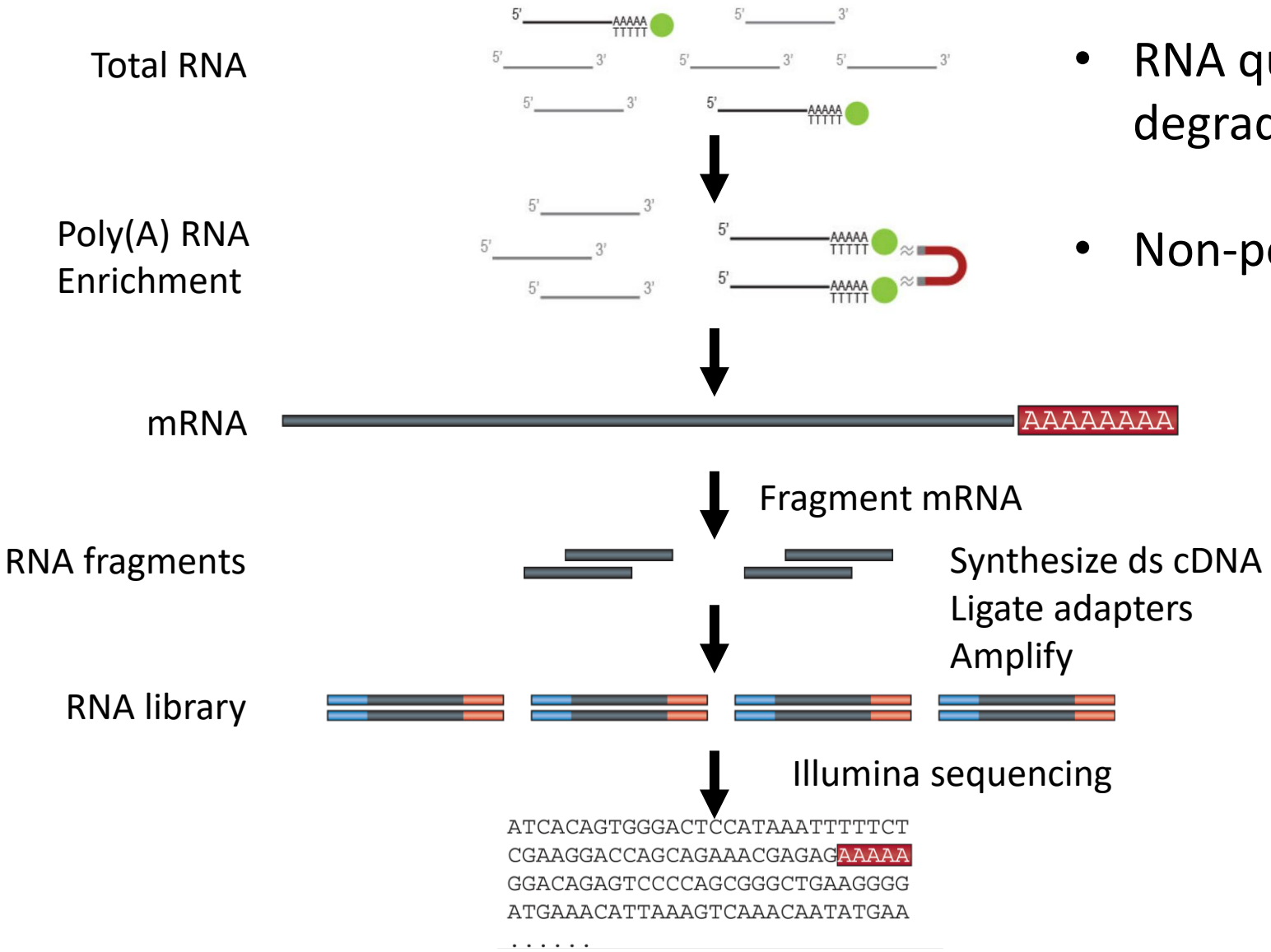
How to remove highly abundant ribosomal RNA (rRNA)?

- Ribosomal RNA typically constitutes > 90% of total RNA
  - Only 1-2% mRNA

- Enrich for mRNA using Poly(A) selection
  - Requires higher amount of starting material
  - Requires minimal degradation (high RNA integrity number)

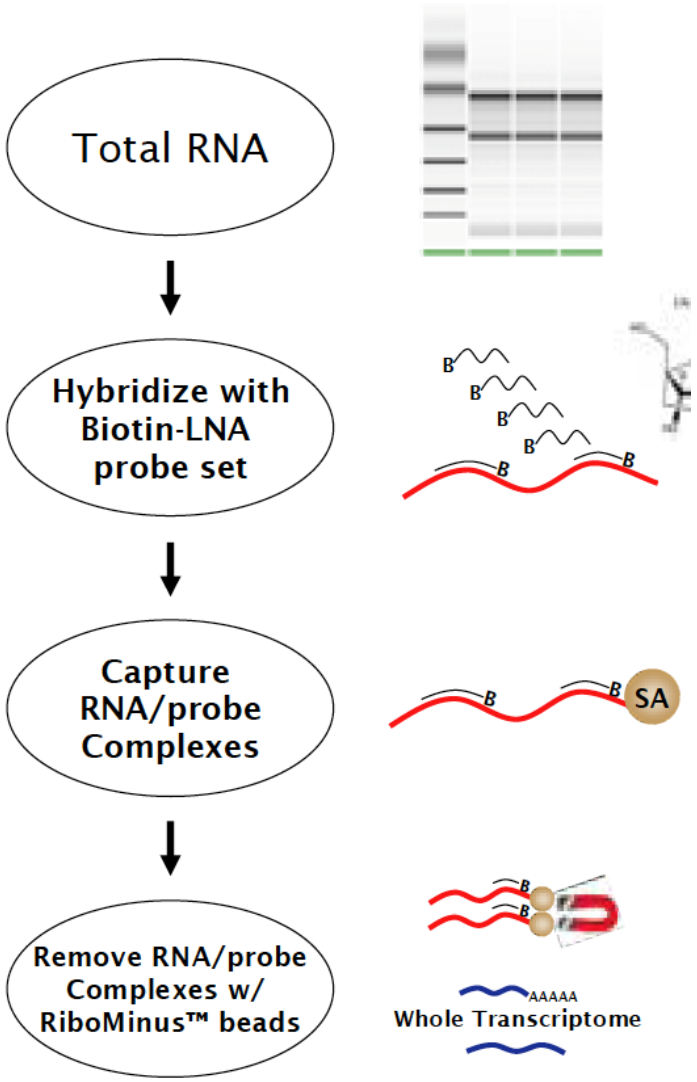
- Deplete rRNA
  - When you cannot obtain great enough quantity
  - When RNA integrity is not high enough
  - Prokaryote can only use rRNA depletion
  - non-polyadenylated RNAs, such as miRNAs and enhancer RNAs

Illumina RNA-seq Library Preparation – polyA selection



- RNA quality must be high – degradation produces 3’ bias
- Non-polyA RNAs are not recovered

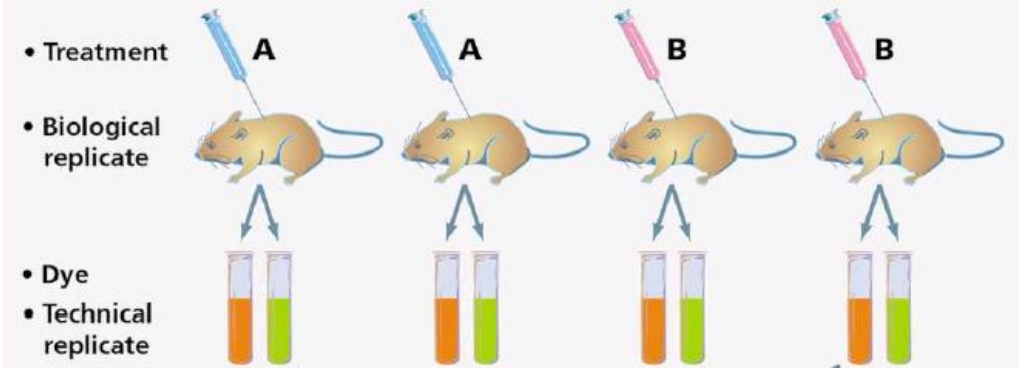
Ribosomal RNA Subtraction

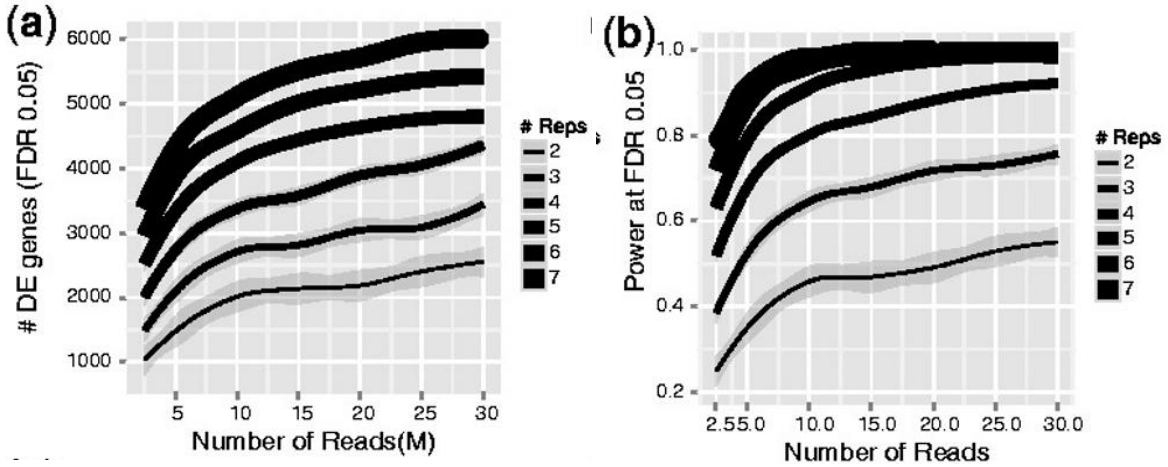


- species-specific probes
- Allow enrichment of non-poly(A) transcripts

Different types of replicates and why we need replicates?

- Biological replicates
- Technical replicates





- Increase in biological replication significantly increases the power and the number of differentially expressed genes identified.

- Outlier detection and removal

RNA-seq differential expression studies: more sequence or more replication? Liu et al., Bioinformatics 2013

How many replicates do I need and power analysis

The statistical power increases with the effect size (difference between the two groups), the sequencing depth, and the number of replicates per group.

	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 %

- Minimum 3-6 biological replicates

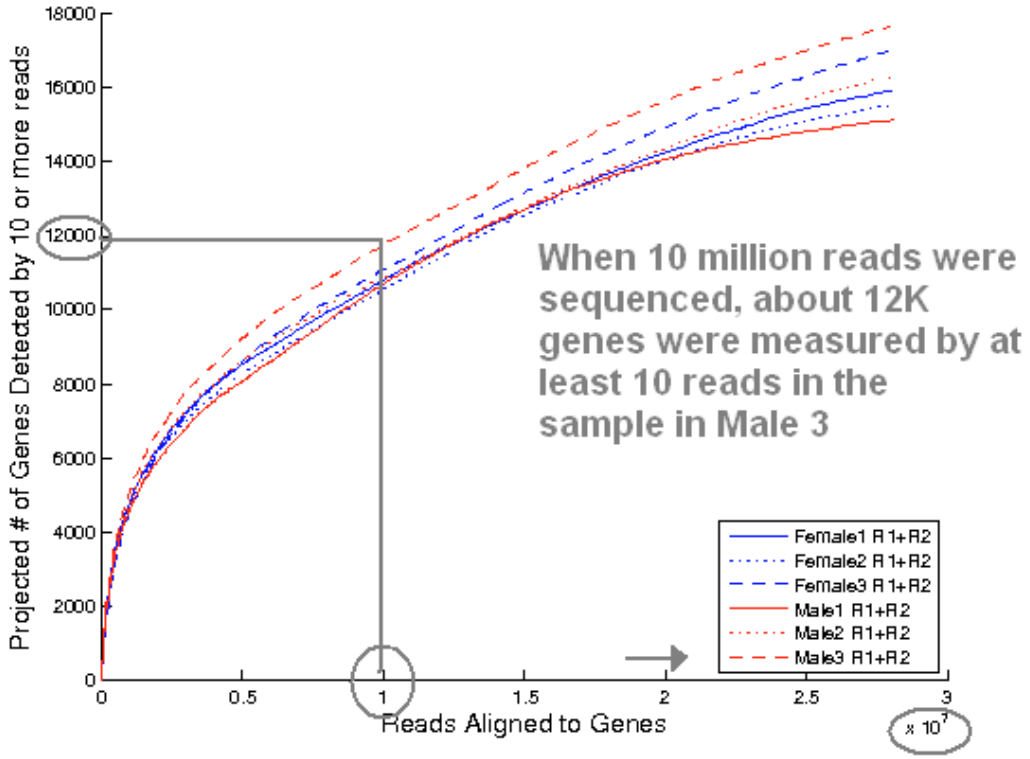
Tools for power calculation:

- PROPER: Wu et al., Bioinformatics (2015).
- EBSeq. Gaye, A., Front. Genet. (2017).

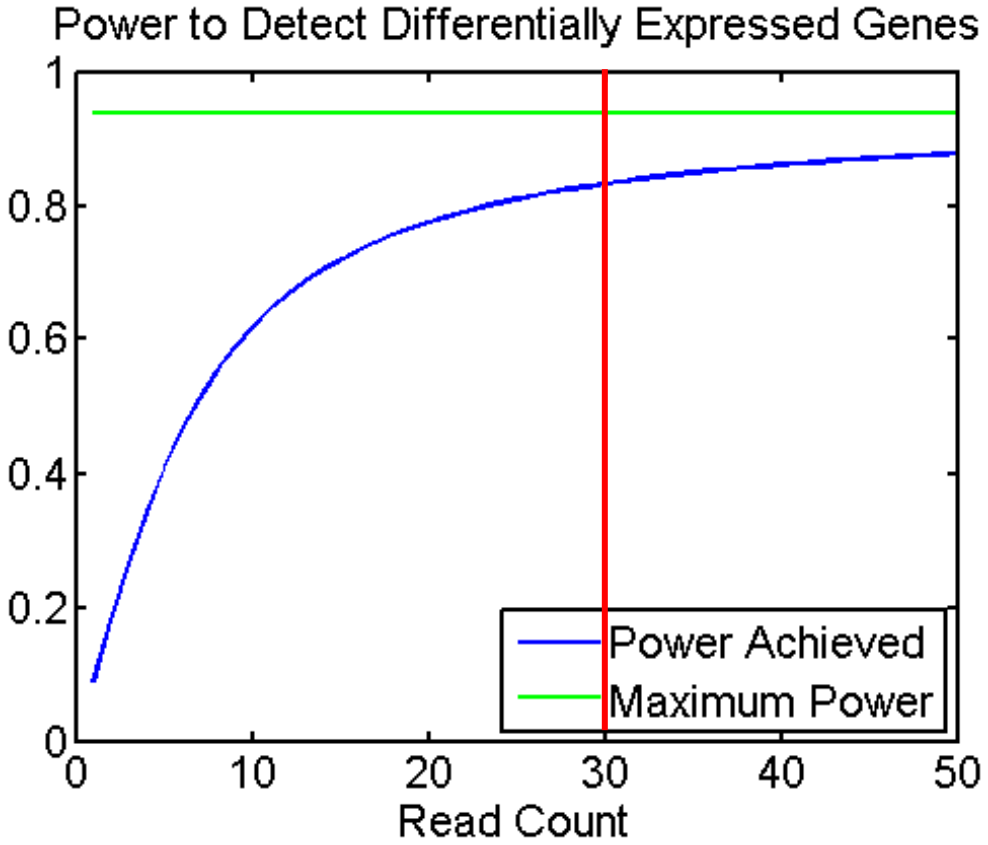
A survey of best practices for RNA-seq data analysis, Conesa et al. Genome Biology (2016)

How many reads do we need?

Genes detected as a function of sequencing depth



- > 10 reads is a standard cutoff



- > 30 million with diminished returns

Different RNA-Seq experiment types require different sequencing read lengths and depth

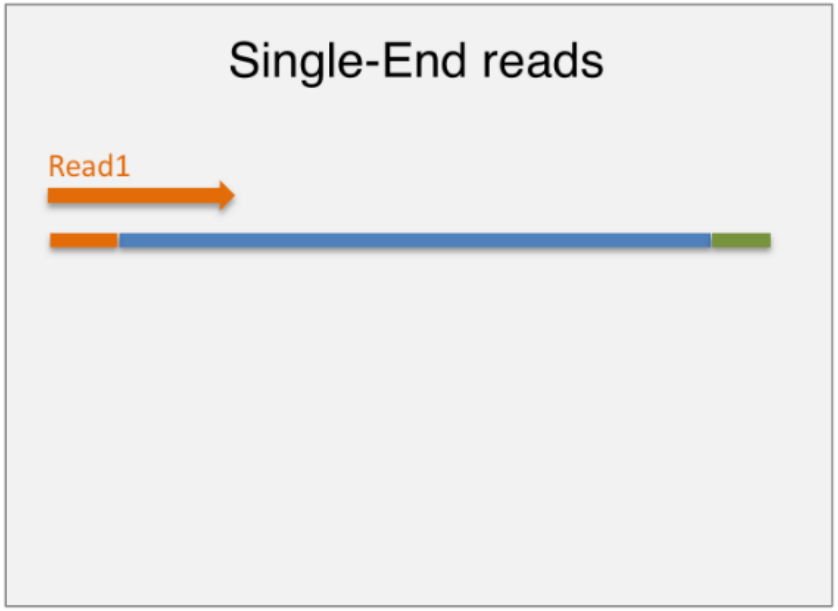
Guideline:

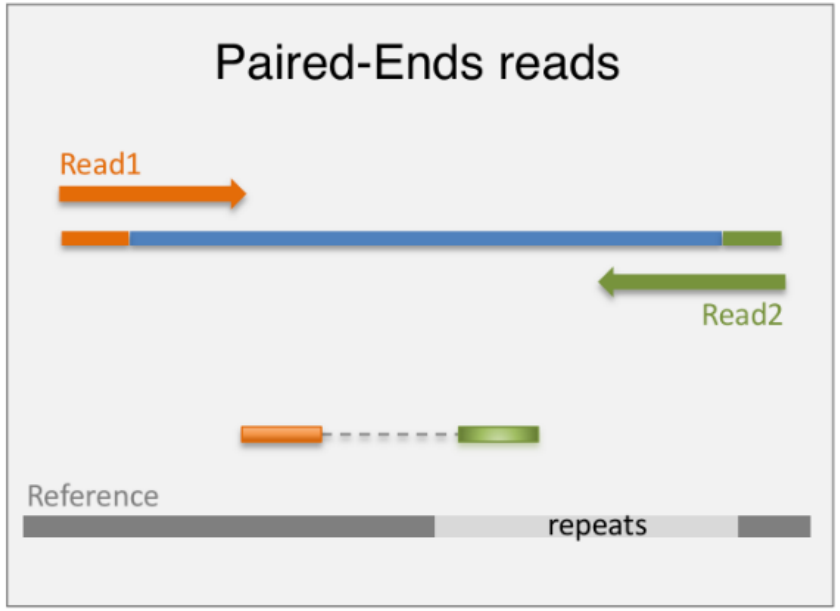
- mRNA gene level: 5 - 30 million reads per sample
- mRNA transcript level (alternative splicing): 30 - 60 million reads per sample
- Transcript discovery: 100 - 200 million short reads, long read data better
- miRNA-Seq or small RNA:
  - Varies significantly depending on the tissue type being sequenced
  - Most applications require 1 - 5 million reads per sample
  - Use primary literature to determine how many reads are needed

Read length

- Affects the ability to determine where each read in the transcriptome came from
- Longer reads do not add much value in a quantification-based analysis but valuable to isoform analysis
- Read length depends on the application and final size of the library.
  - Gene expression / RNA Profiling: 50 - 75 bp
  - Novel transcriptome assembly and annotation: longer, paired-end reads (2 x 75 bp or 2 x 100 bp), or long read sequencing
  - Small RNA: a single read (usually a 50 bp read) typically covers the entire sequence

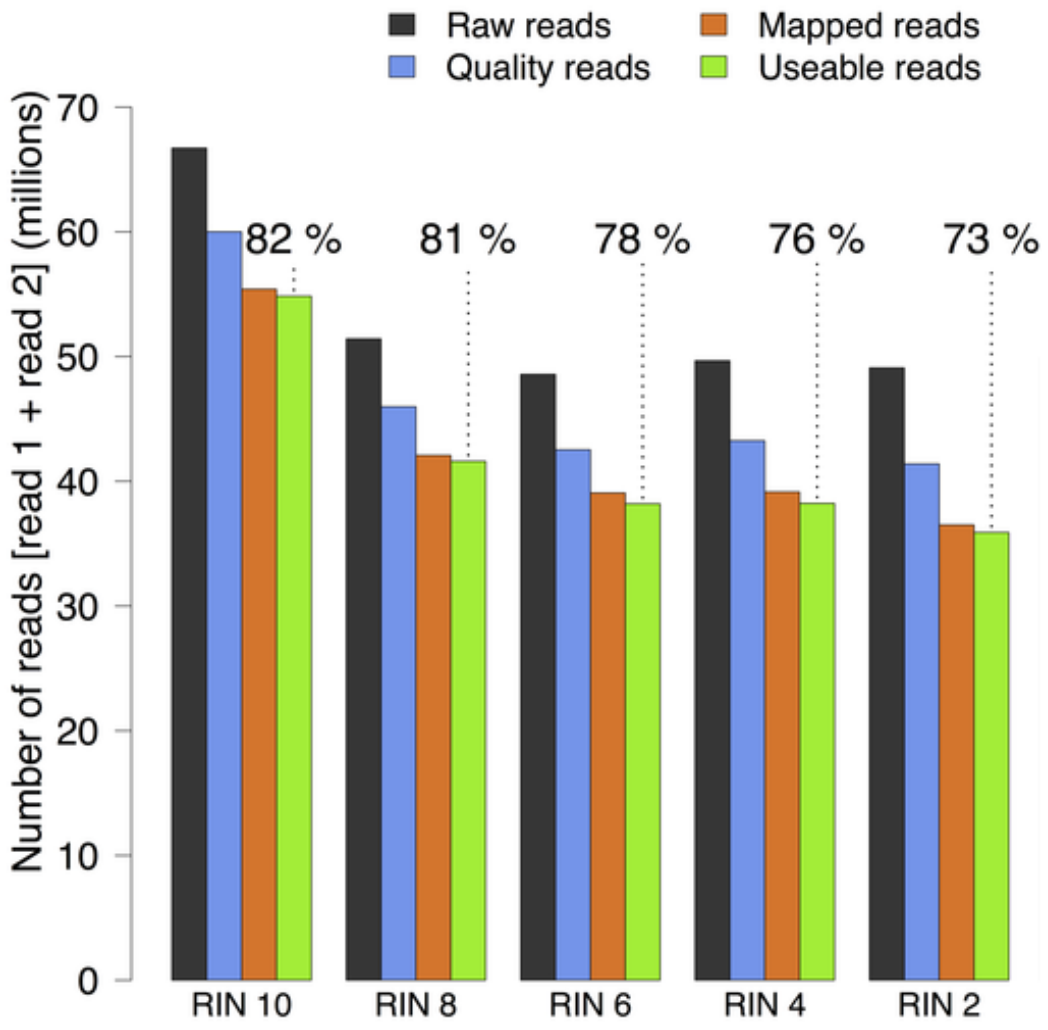
Single-end vs. paired-end sequencing



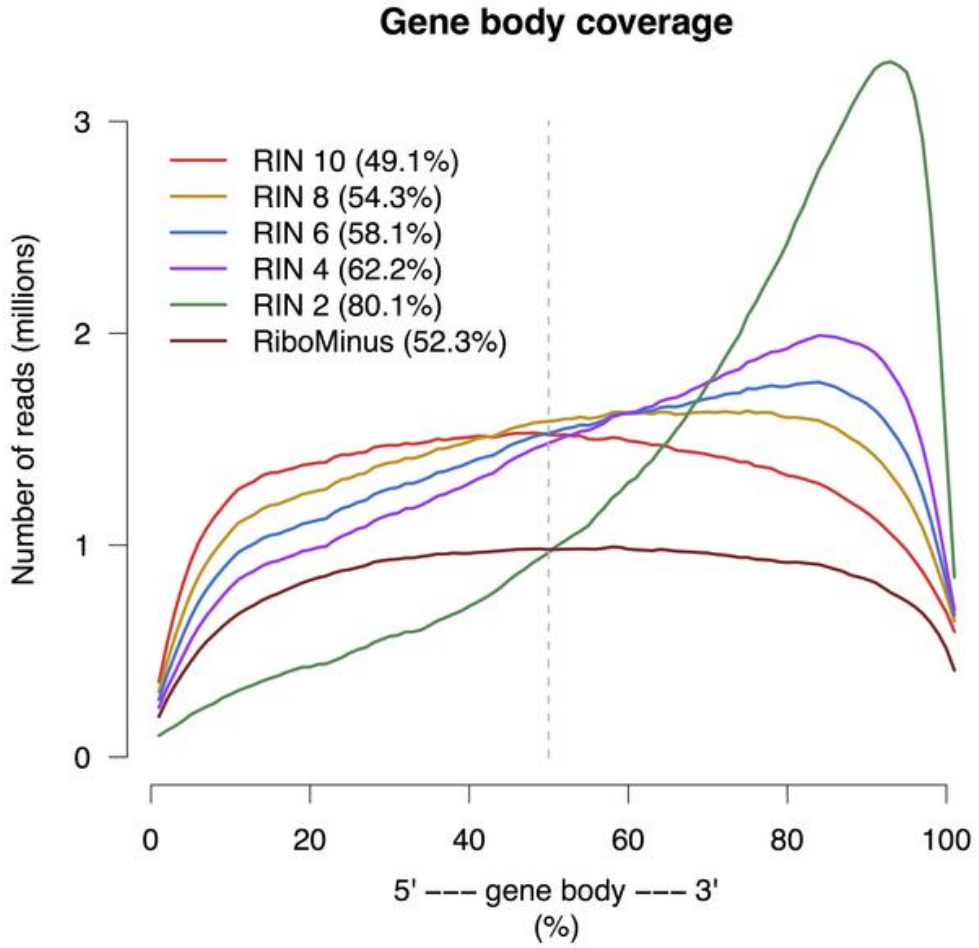


- Paired-end improves read mapping
- Paired-end preferred for alternative-exon quantification, fusion transcript detection and de novo transcript discovery, particularly when working with poorly annotated transcriptomes
- based on cost or on the sequencing technology available to the user.

RNA Isolation: Quality Matters!

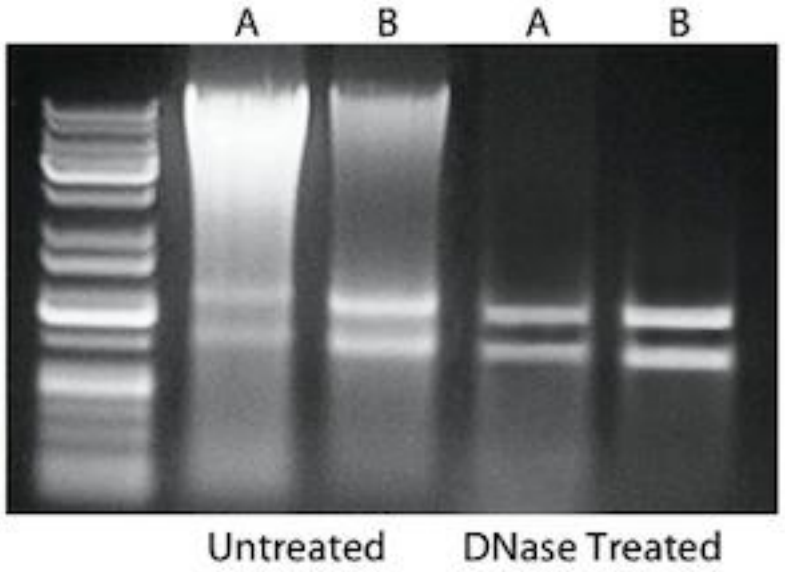


RNA Integrity Number (RIN)



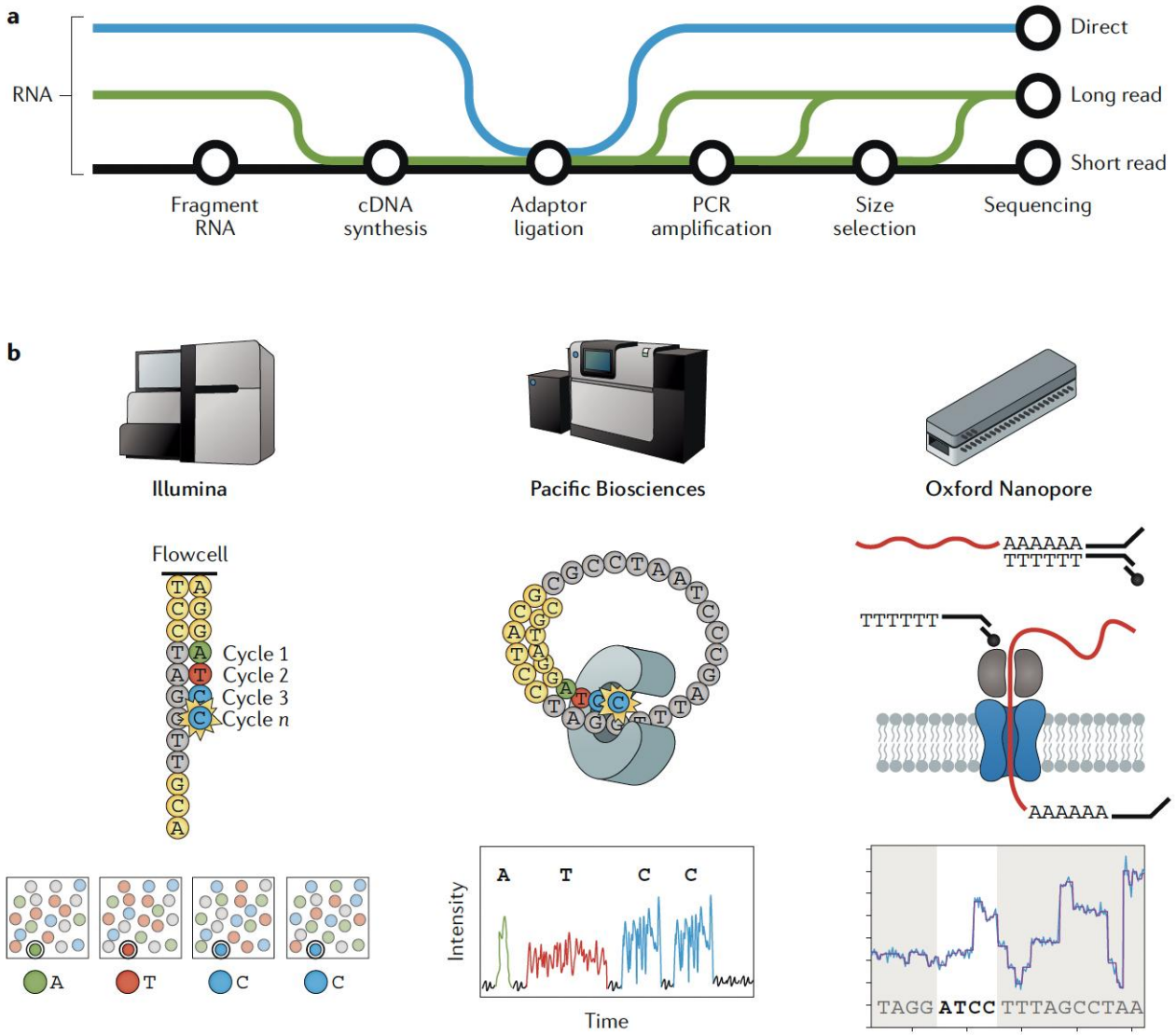
Sigurgeeirsson et al., *PLOS One* (2014)

Remove DNA Contamination



- DNA contamination can be mapped back as intergenic or intronic sequence
- Can not distinguish between contamination vs. alternative splicing, unannotated or noncoding transcripts, or spurious transcription

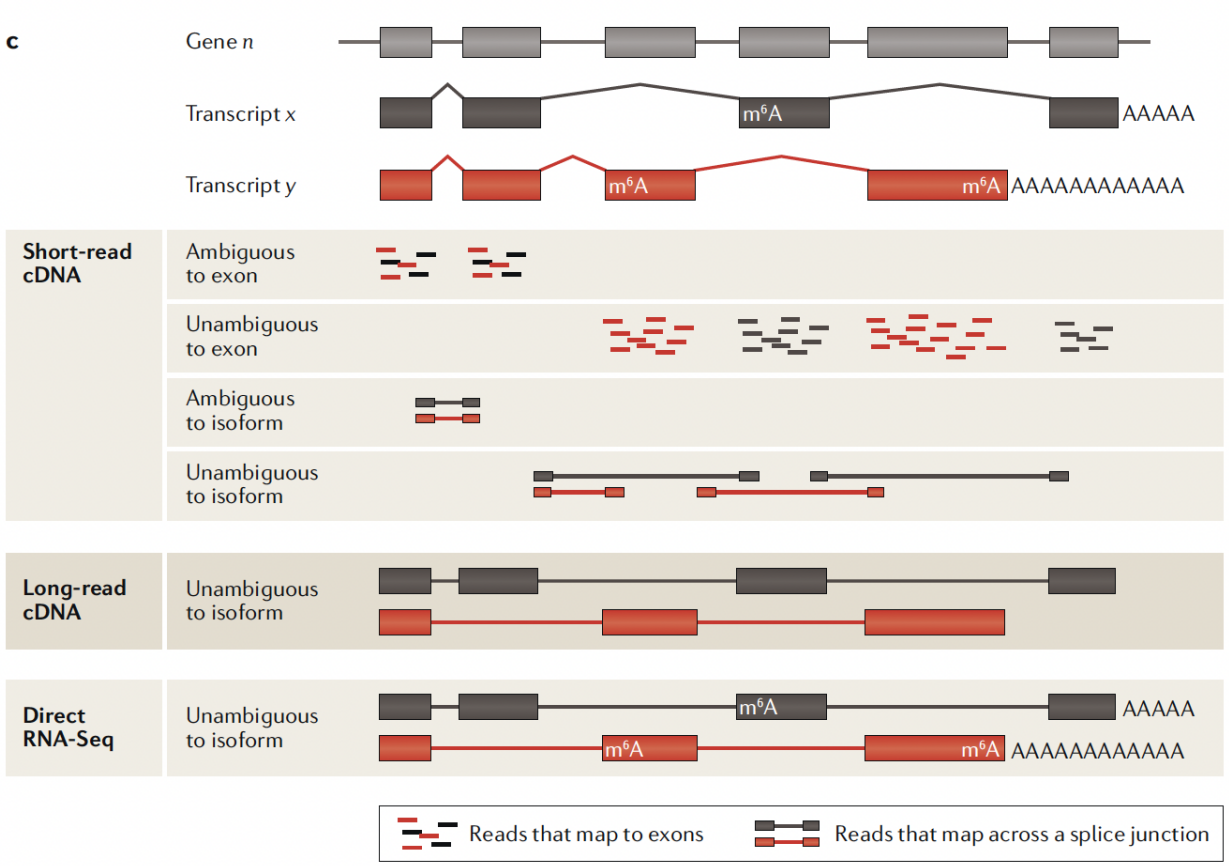
Emerging RNA-seq methods



- Illumina short-read sequencing
  - < 200 bp
- Long-read cDNA sequencing
  - converting mRNA to cDNA before sequencing
  - Pacific Biosciences (PacBio) and Oxford Nanopore (ONT)
  - up to 50 kb
- Long-read direct RNA sequencing (dRNA-seq)
  - No cDNA synthesis and/or PCR amplification during library preparation.
  - Oxford Nanopore
  - 1-10 kb

RNA sequencing: the teenage years, Stark et al., Nature Reviews Genetics, 2019

Comparison of short-read, long-read and direct RNA-seq analysis



Illumina short-read sequencing (< 200 bp):

- the de facto method to detect and quantify transcriptome-wide gene expression
  - cheaper
  - easier to implement
  - comprehensive, high-quality data

Long-read RNA-seq (up to 50 kb):

- generate full-length isoform reads
- Isoform detection
- De novo transcriptome analysis
- Fusion transcript detection

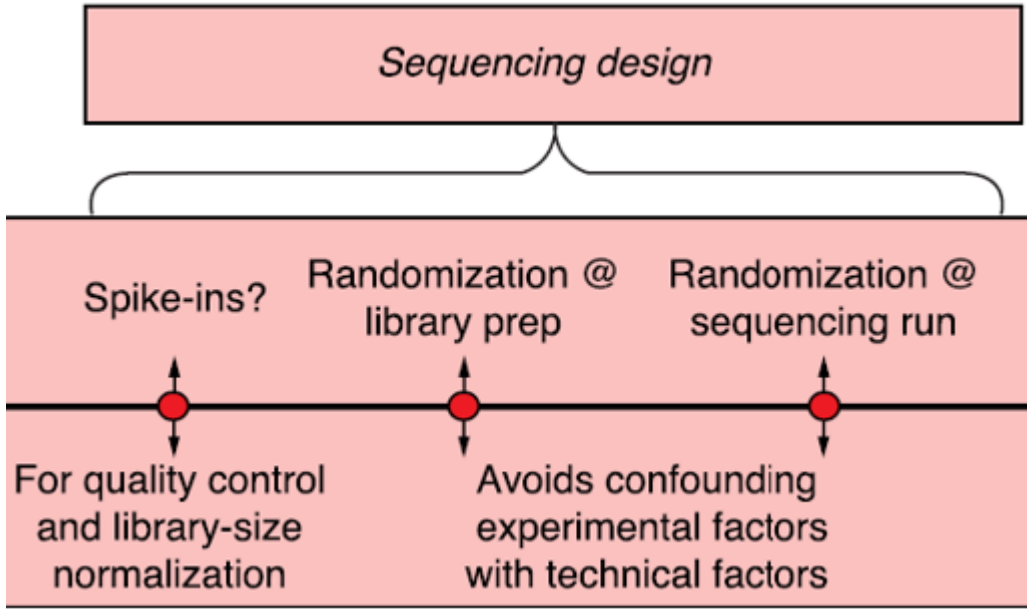
Direct RNA sequencing (dRNA-seq, 1-10 kb):

- All analysis same as Long-read RNA-seq
- Detect base modification (such as N6-methyladenosine (m6A))
- Estimate poly(A) tail length.

Limitation of long-read technologies

- Lower throughput
  - Illumina: up to 16-20 billion paired-end reads per flow cell
  - PacBio platforms: 15–40 million reads per SMRT Cell
  - Oxford Nanopore Technologies (ONT) platforms: up to millions of reads; the exact number of reads will vary based on the sample type, library preparation, and run duration.
- Lower sensitivity
  - depend on RNA integrity (synthesis of degraded mRNA)
  - truncation of cDNA synthesis (need highly processive reverse transcriptase)
- Biases inherent to sequencing platforms
  - low diffusion of long library molecules onto the surface of the sequencing chip can reduce the coverage of longer transcripts.

Sequencing Design



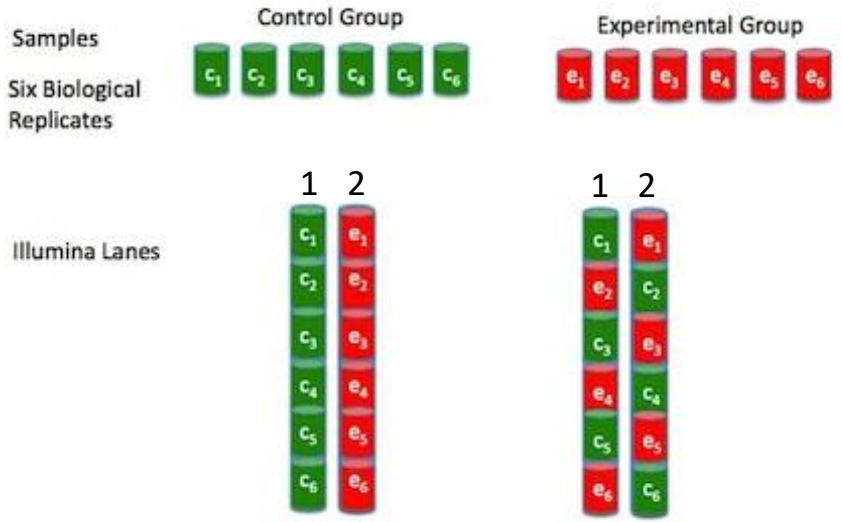
Two main sources of variations:

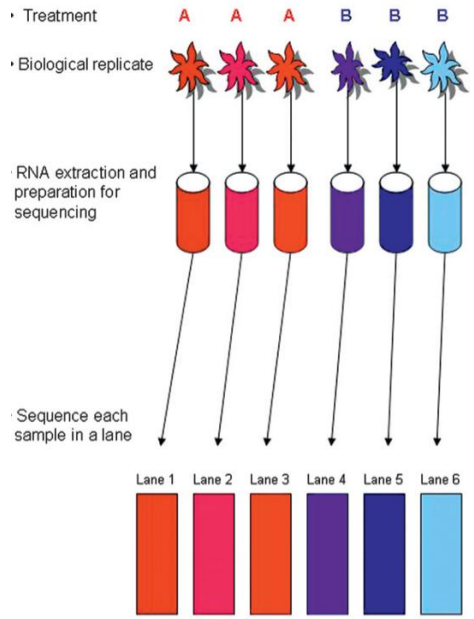
- Batch effects: any errors that occur after random fragmentation of the RNA until it is input to the flow cell (e.g., PCR amplification and reverse transcription artifacts).

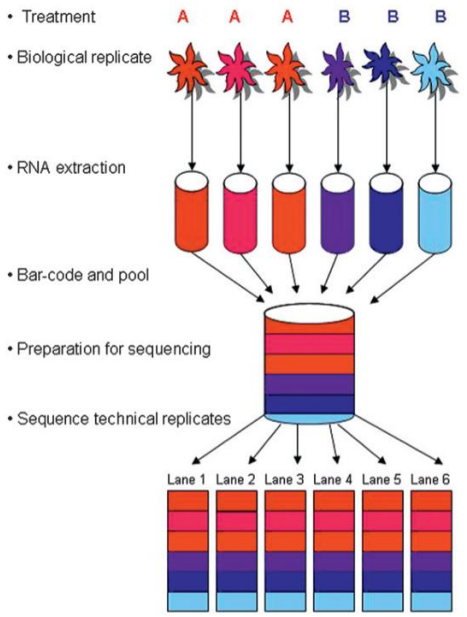
- Lane effects: any errors that occur from the point at which the sample is input to the flow cell until data are output from the sequencing machine (e.g., systematically bad sequencing cycles and errors in base calling).

Randomization:
Randomize samples across library preparation batches and lanes so as to avoid technical factors becoming confounded with experimental factors.

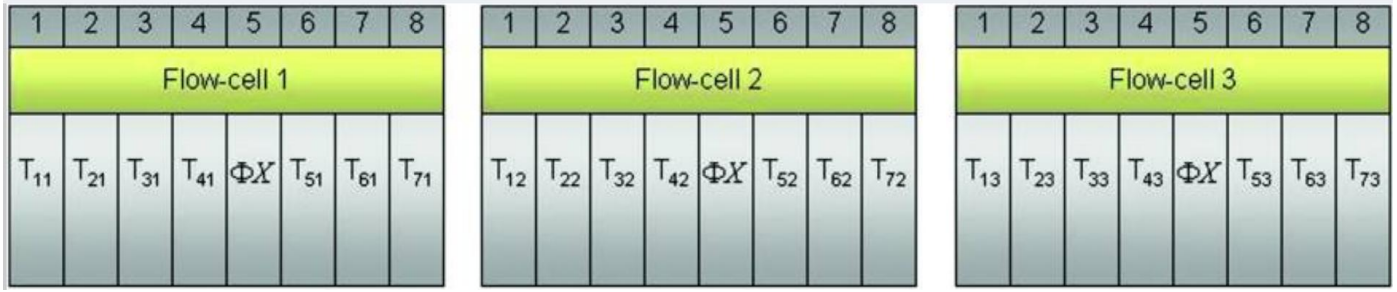
Sequencing Design – good or bad?







Biological rep1 | Biological rep2 | Biological rep3

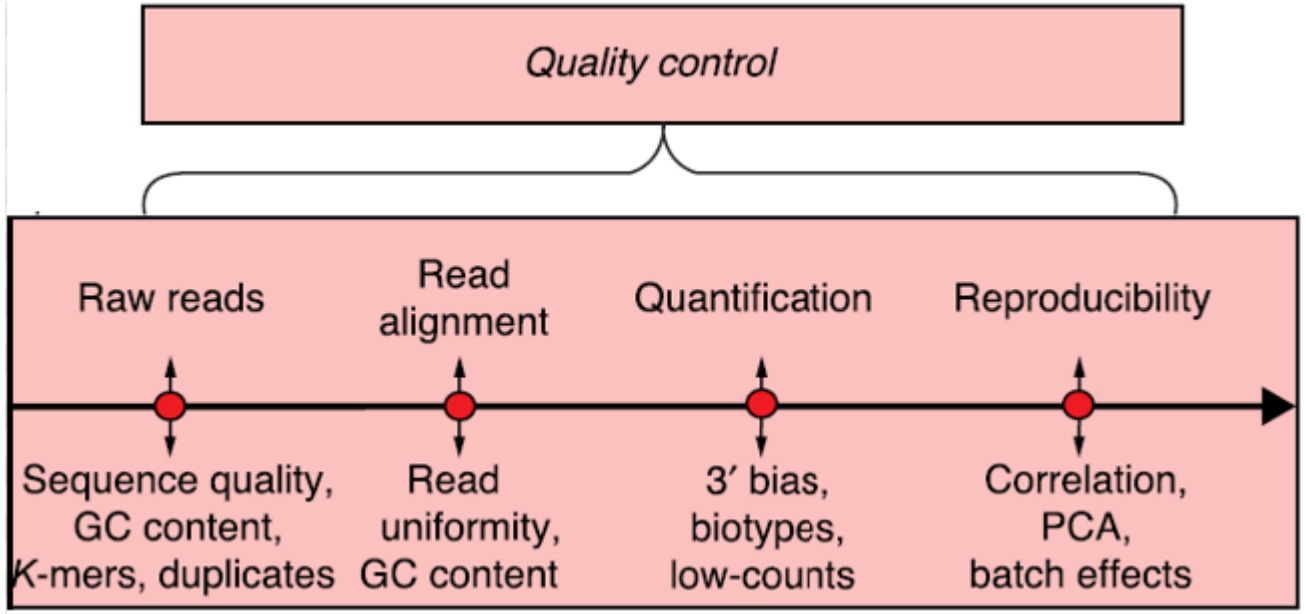


- 7 treatment groups T1, T2, T3, T4, T5, T6, T7

- Ensure the total number of reads evenly distributed among the individual samples
- Pilot sequencing run before committing to an expensive, multi-lane sequencing run.

Statistical Design and Analysis of RNA Sequencing Data
Paul L. Auer and R. W. Doerge, Genetics (2010)

Data quality control



FastQ – The Standard Output, each read has 4 lines

1. @SEQ_ID
2. GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
3. +
4. !“*((((***+))%%%++)(%%%%).1***-+*”))**55CCF>>>>>>CCCCCCC65

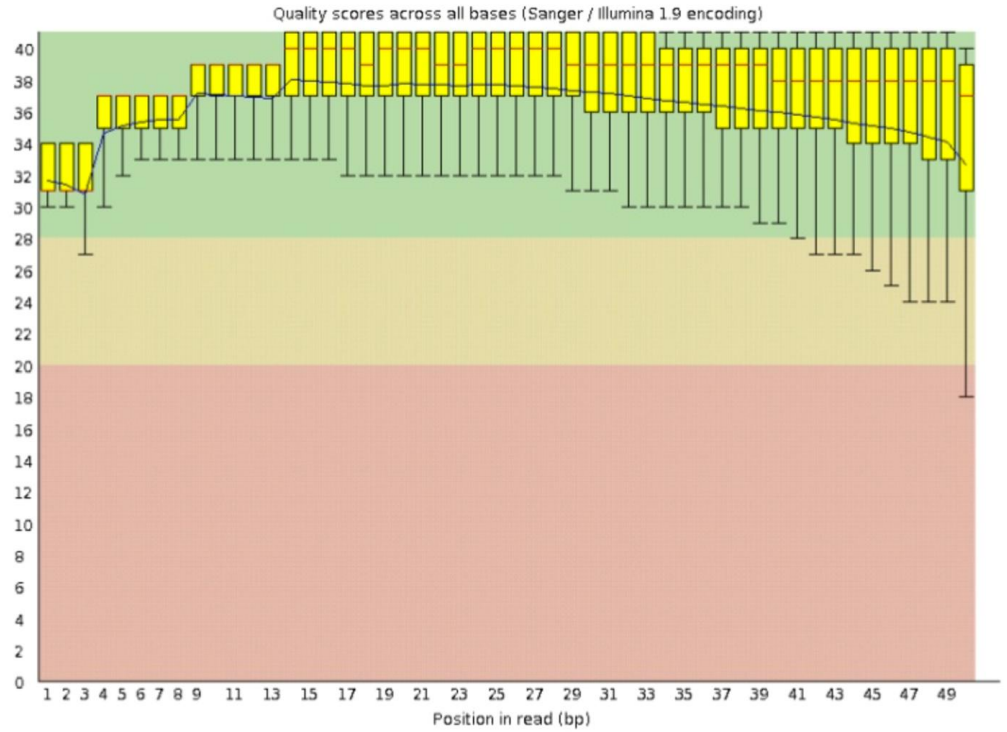
Data quality assessment

FastQC

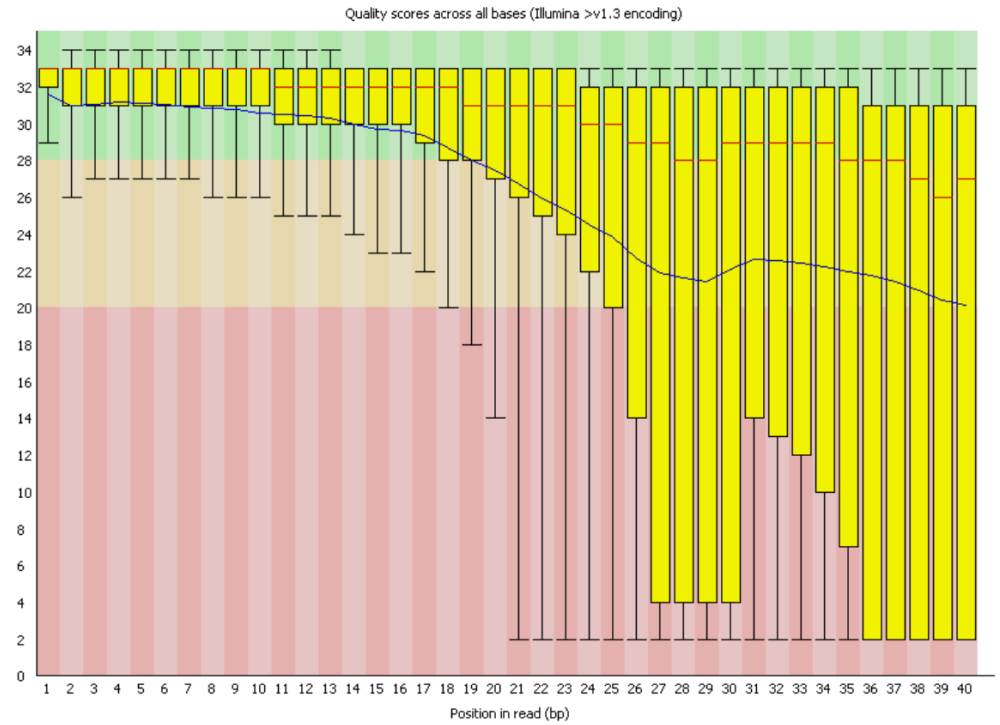
FastQC, MultiQC

- 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 - Per Base Sequence Quality



Good

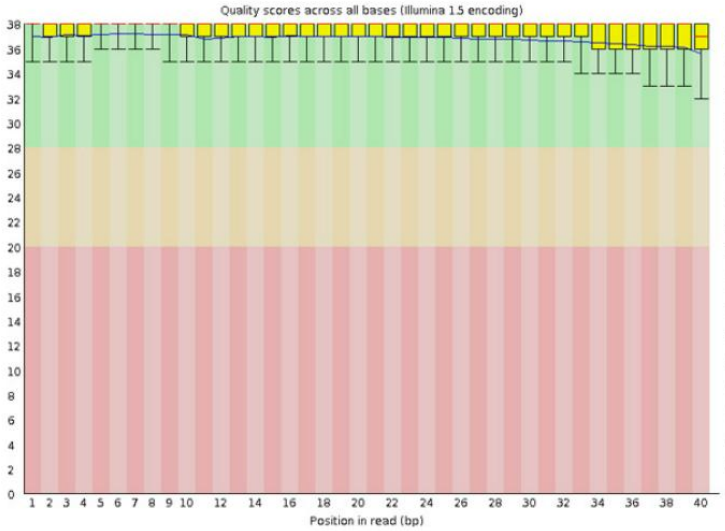


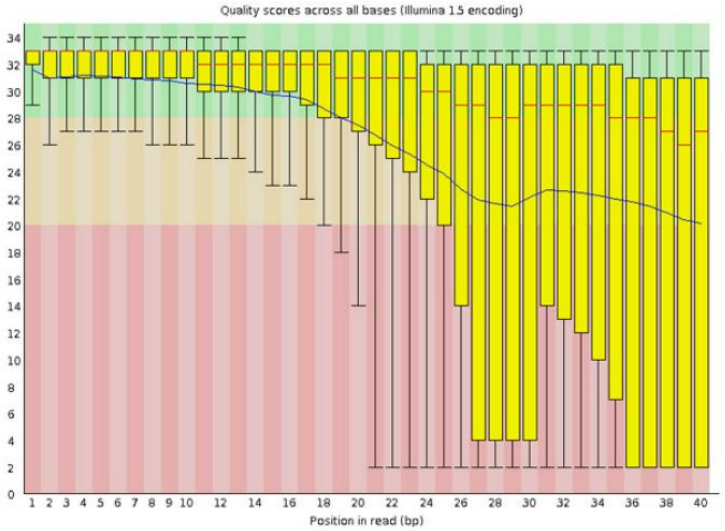
Not so Good

- In general, read quality decreases towards the 3’ end of reads
- To improve read mappability: FASTX-Toolkit, PRINSEQ, and Trimmomatic can be used to discard low-quality reads, trim adaptor sequences, and eliminate poor-quality bases.

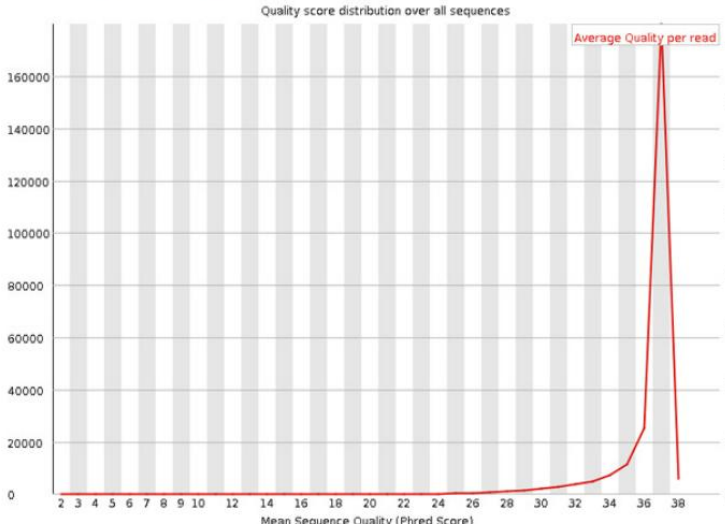
FastQC - Per Sequence Quality Scores

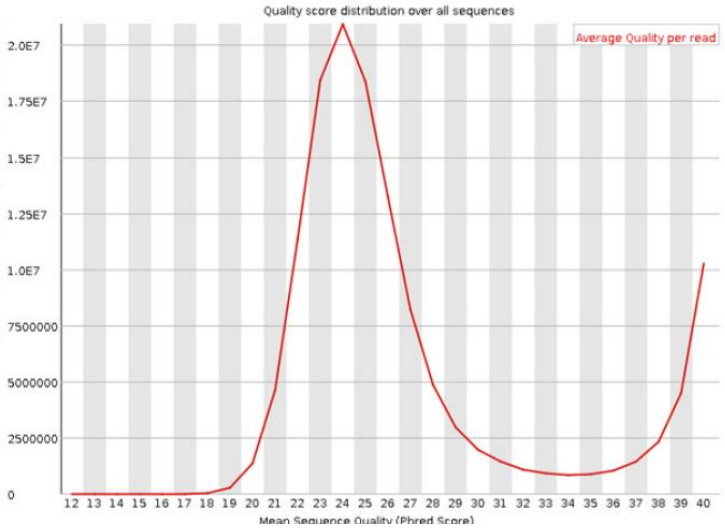
Per base sequence quality



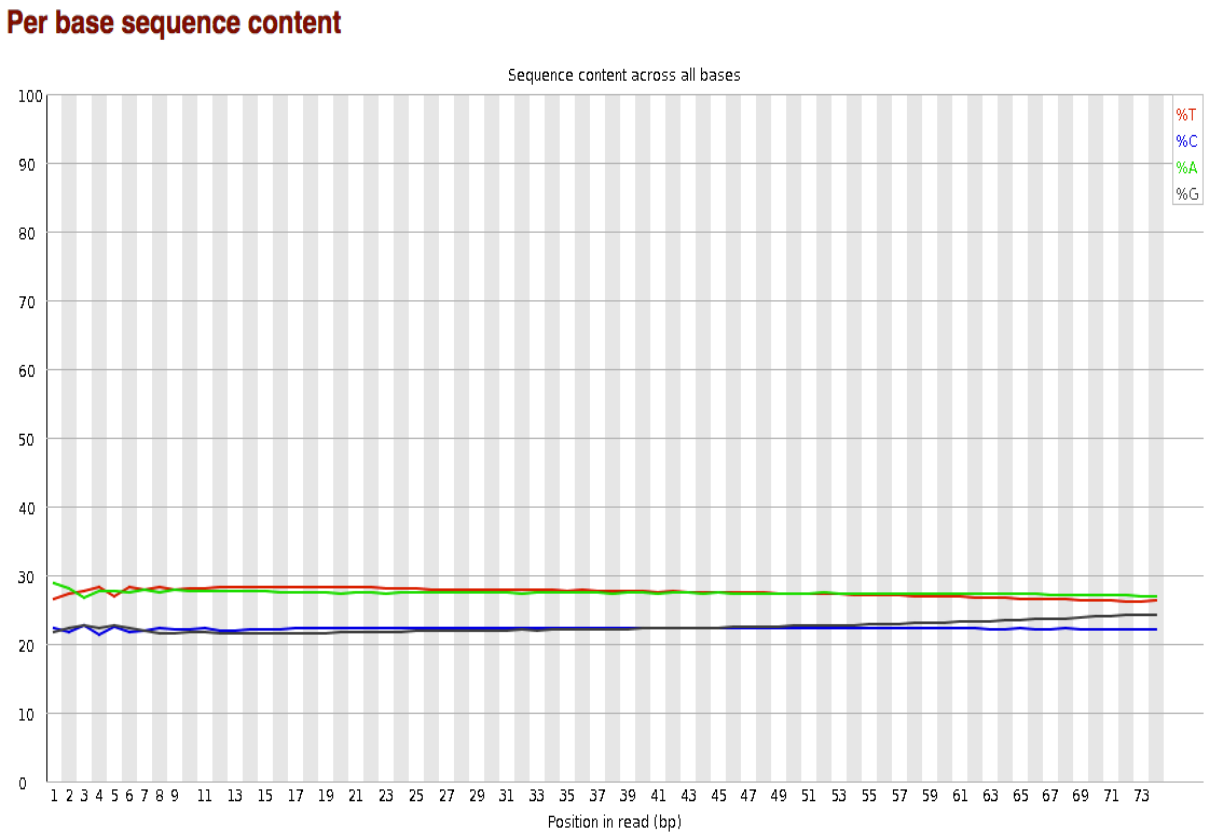


Per sequence quality scores

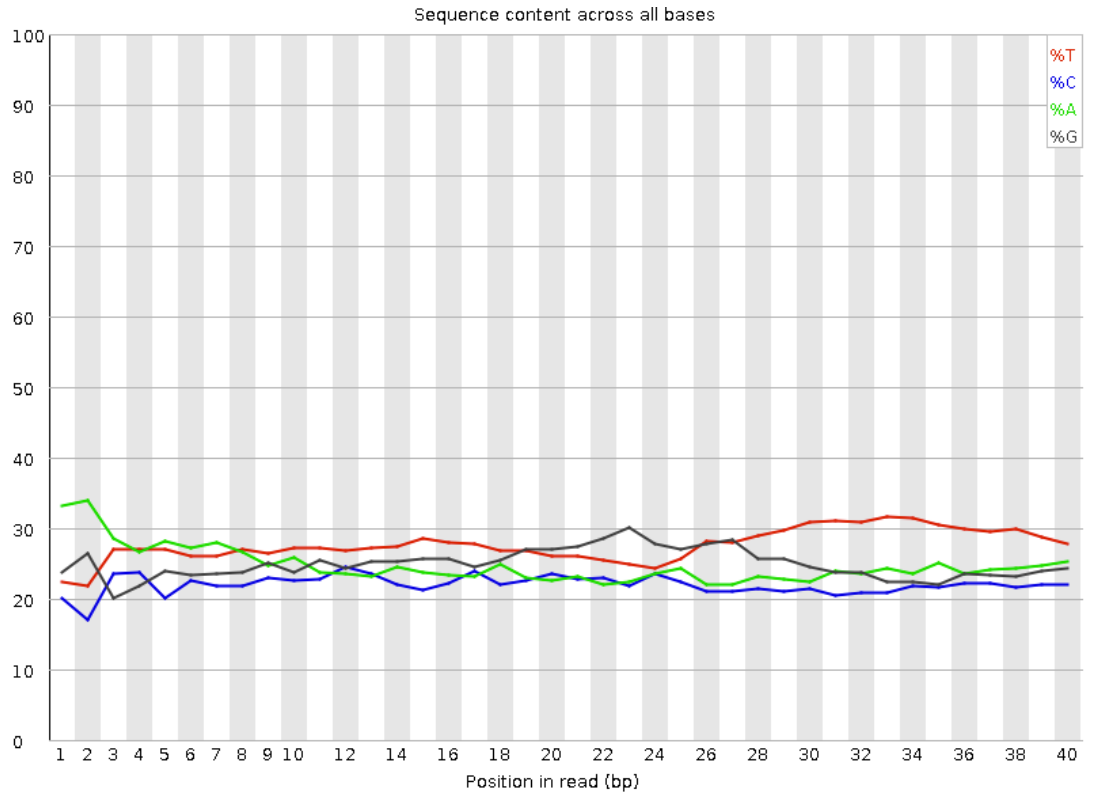




FastQC- Per Base Sequence Content

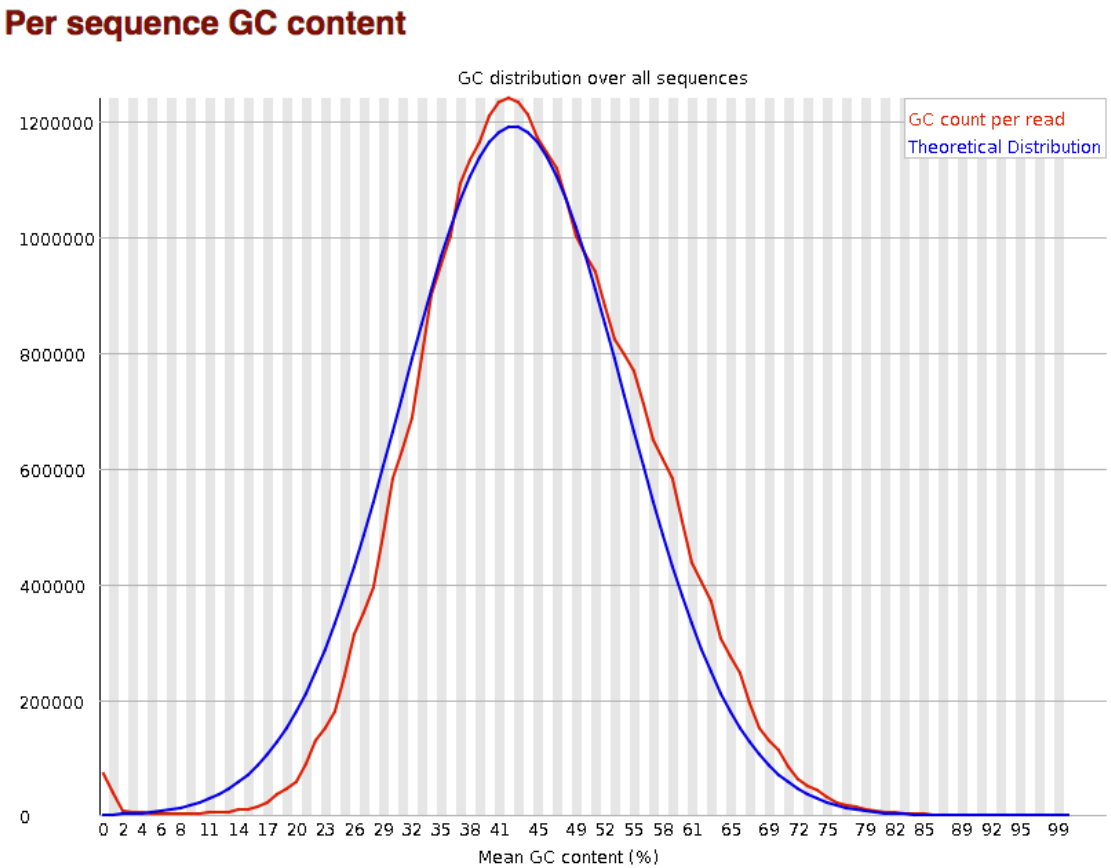


Good

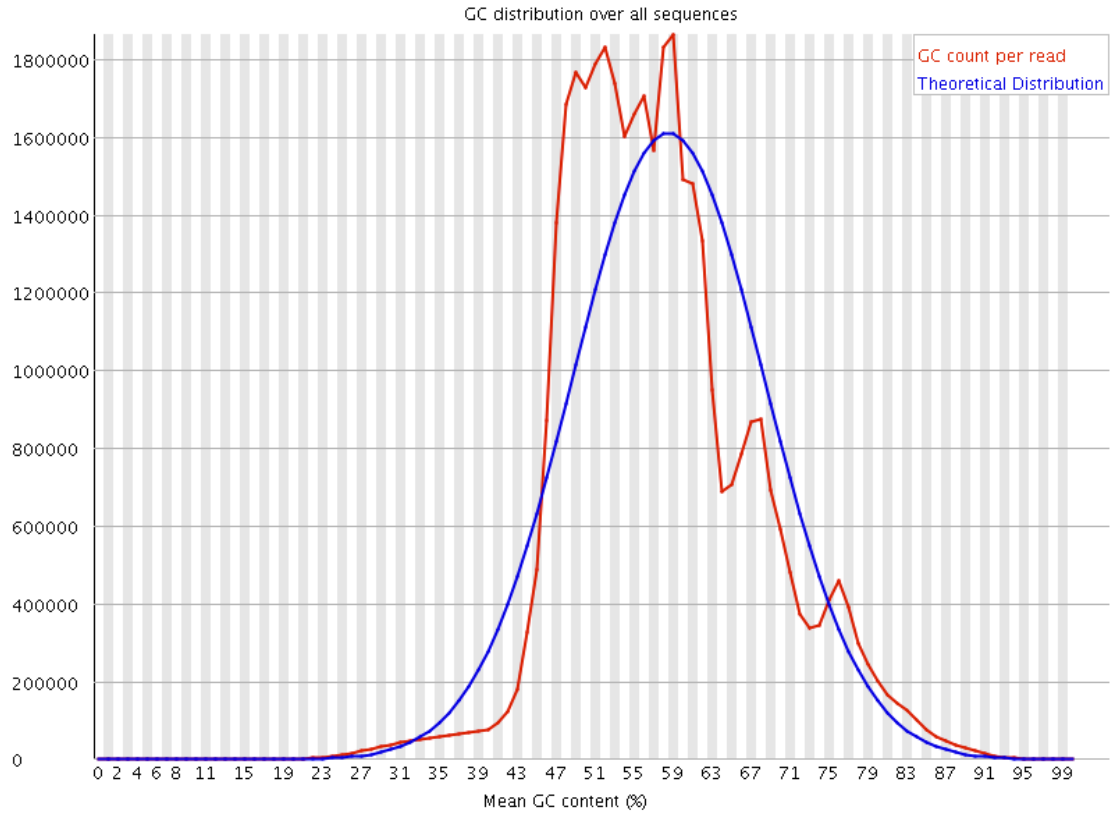


Not so Good

FastQC - Per sequence GC Content

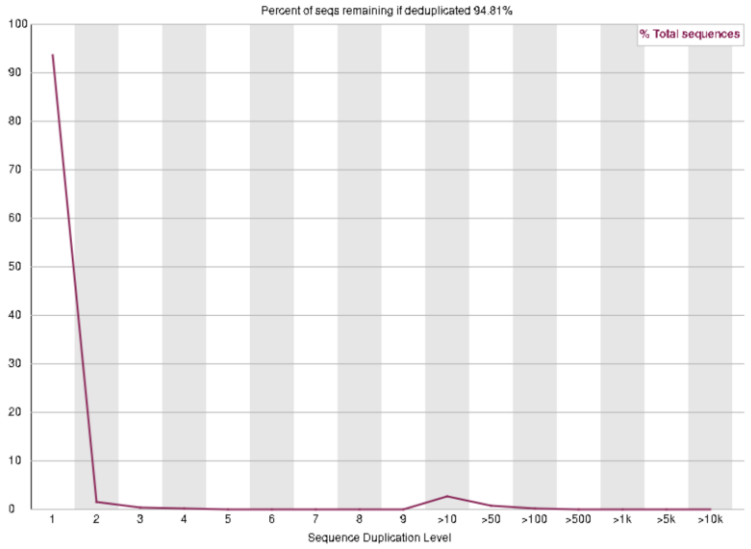


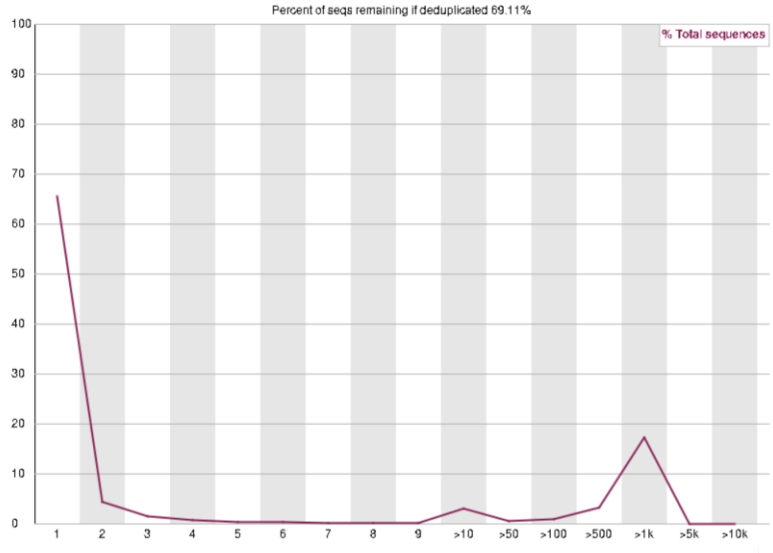
Good

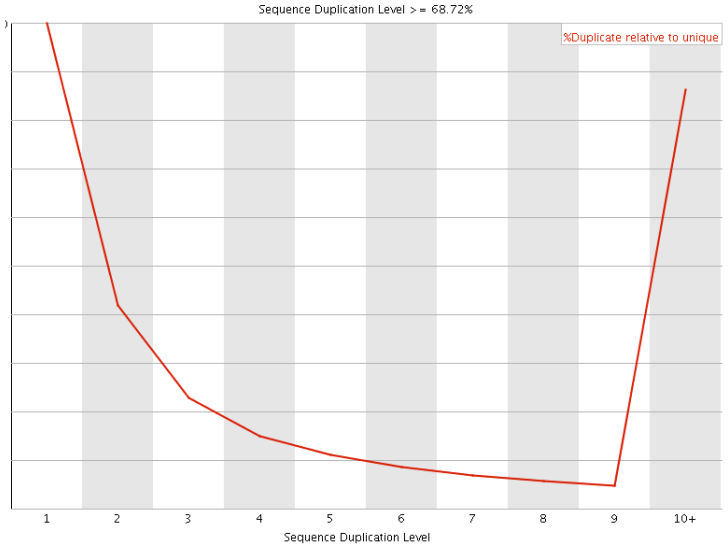


Not so Good

FastQC - Duplicate Sequences

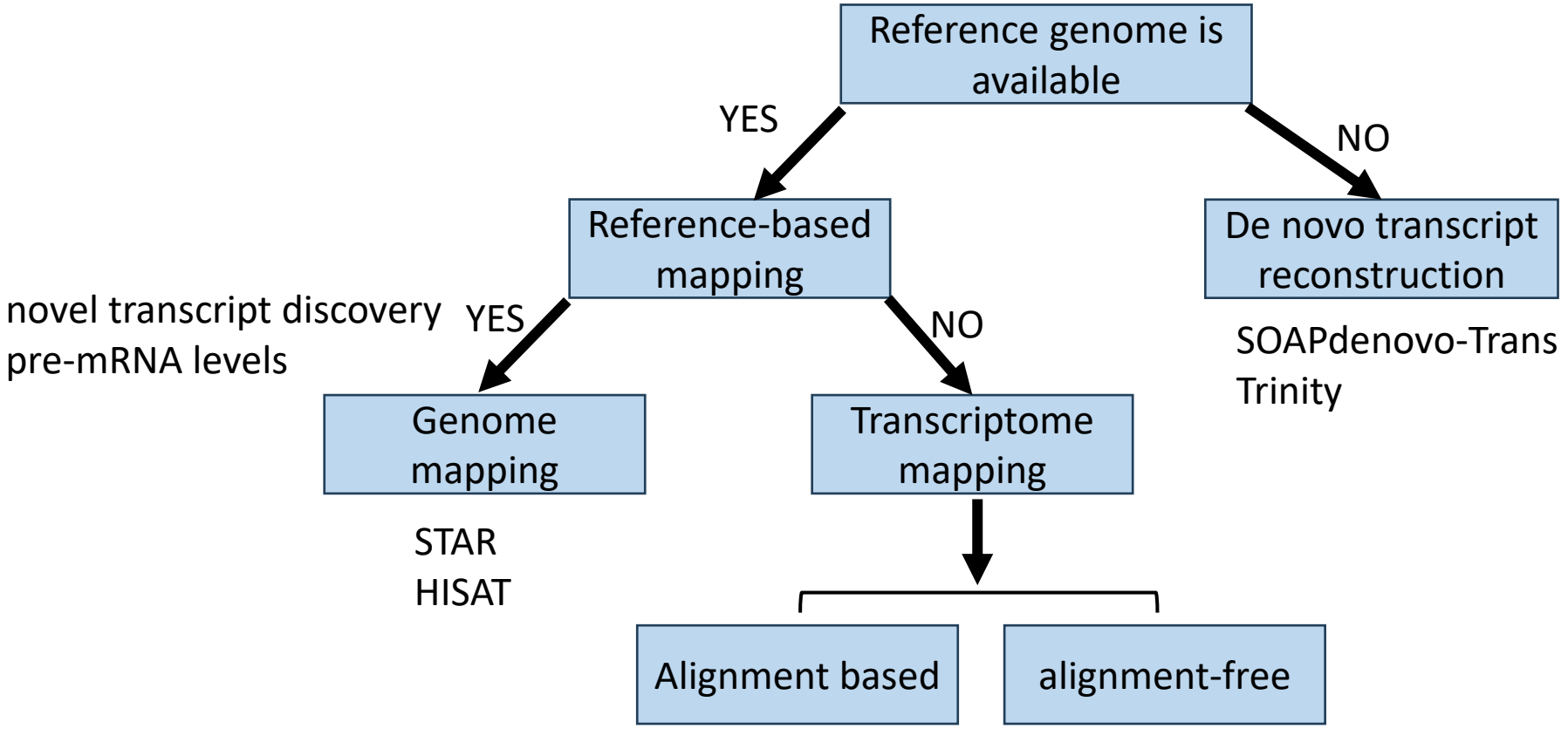






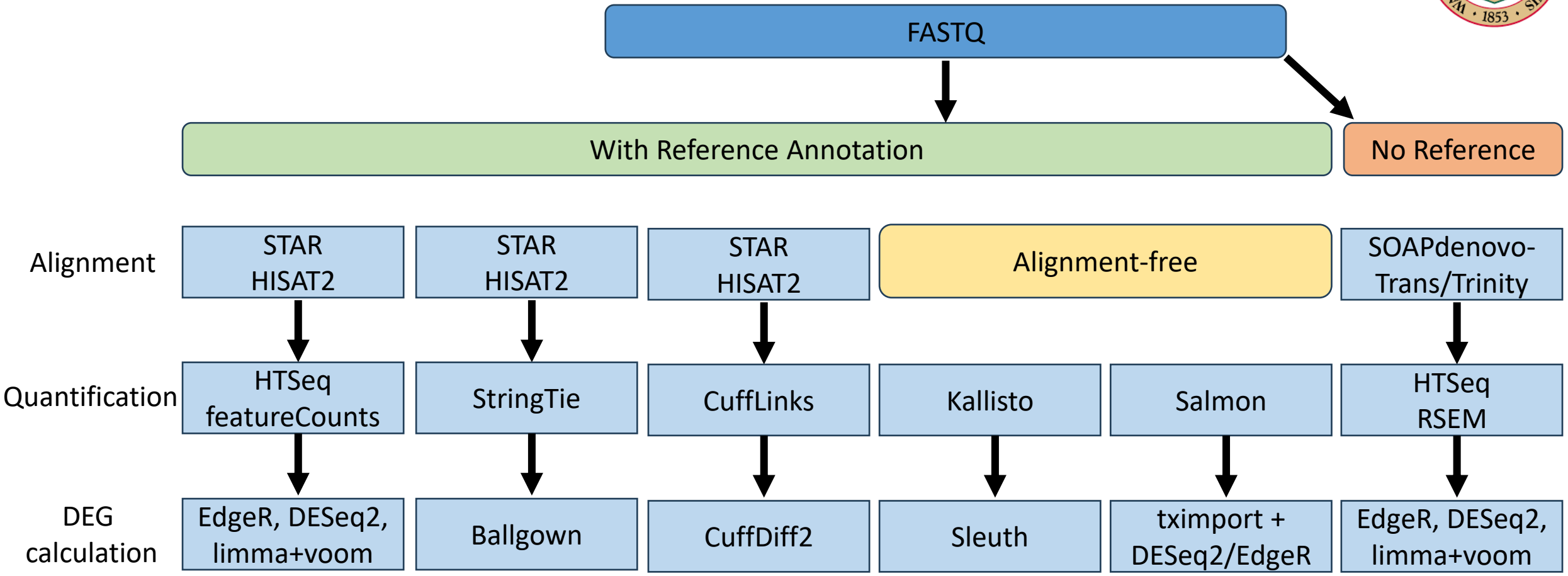
- Some amount of duplication is to be expected in RNAseq
- High-complexity library - low level of duplication may indicate a very high level of coverage of the target sequence
- Highly expressed transcripts can be over-sequenced in order to be able to see lowly expressed transcripts
- A badly PCR duplicated library might have levels above 90%

Read alignment and quantification



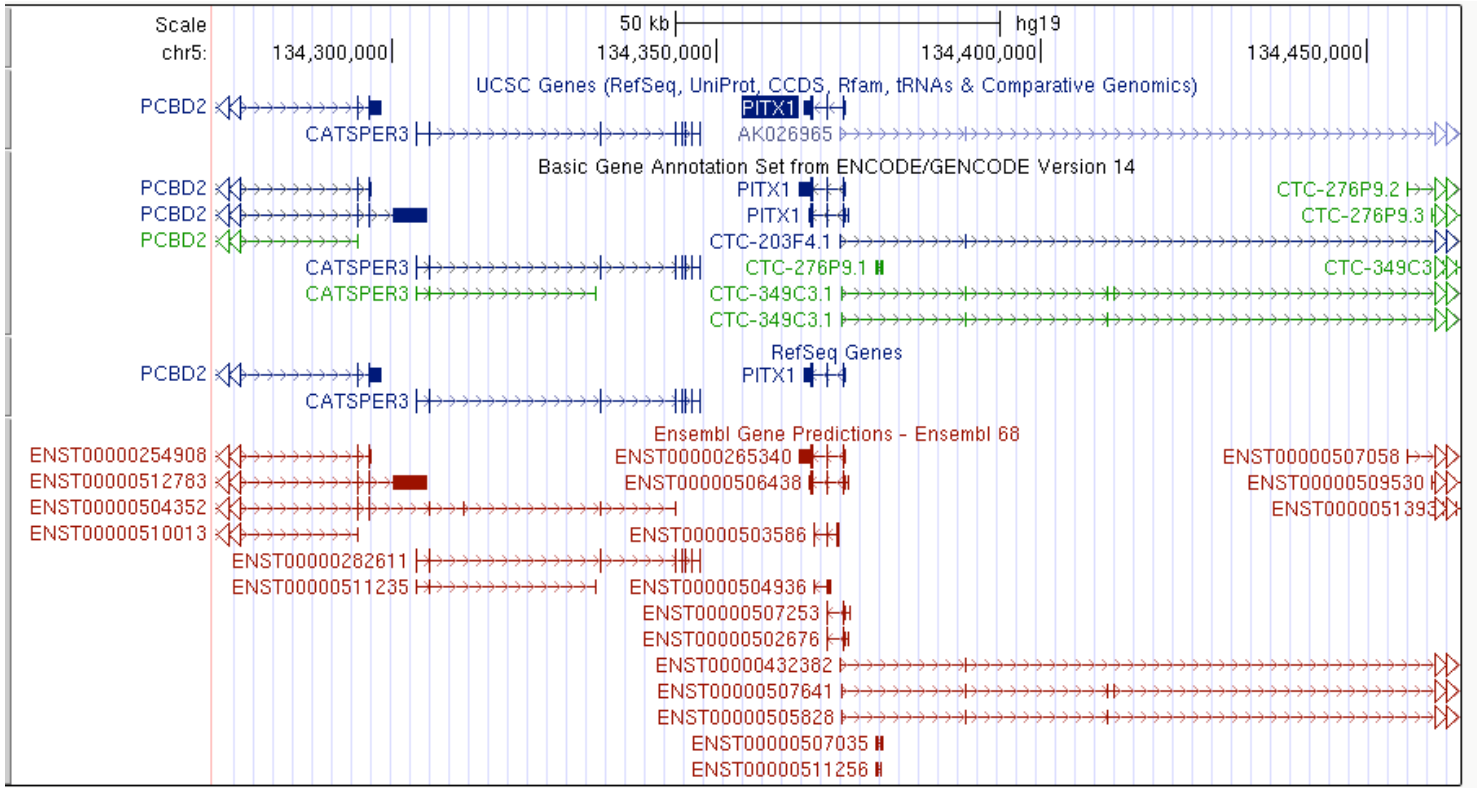
- Salmon, kallisto, Sailfish
- Fast
- systematically poorer performance in quantifying lowly-abundant and small RNAs. (Wu et al., BMC Genomics, 2018)

Popular RNA-seq Workflow



Which Annotation to Use?

The choice of a gene model has a dramatic effect on both gene quantification and differential analysis.



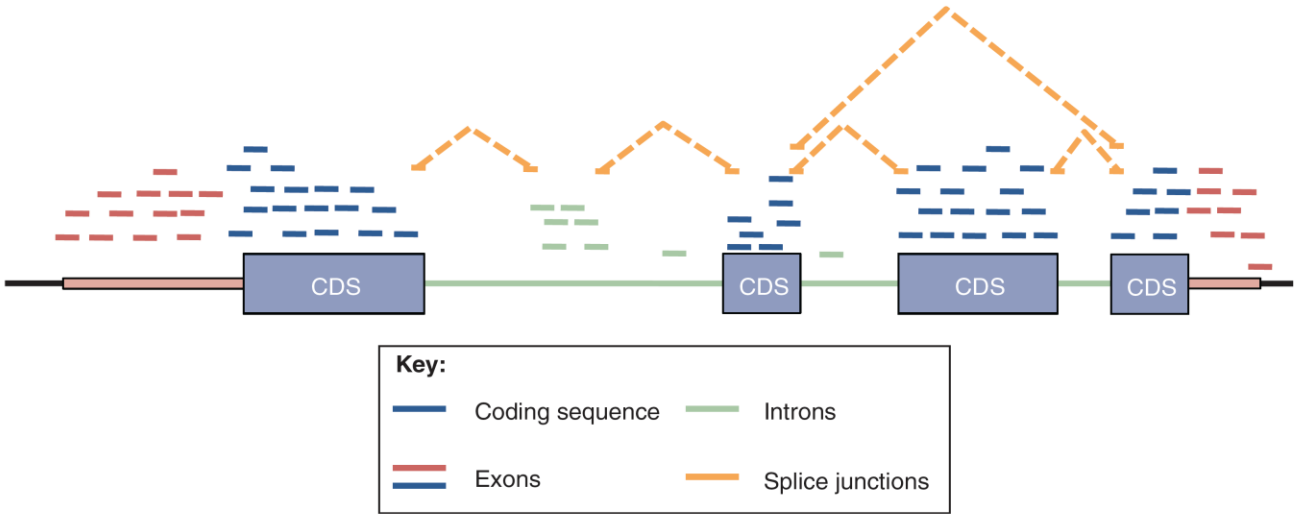
- GENCODE:
  - ENCODE project
- Ensembl project
- RefSeq:
  - the oldest database
  - built by NCBI
- UCSC Known Genes
- Human Pangenome Reference Consortium

A comprehensive evaluation of ensembl, RefSeq, and UCSC annotations in the context of RNA-seq read mapping and gene quantification, Zhao & Zhang, BMC Genomics, 2015
RefSeq gene annotation yields better RNA-seq quantification results than the more comprehensive Ensembl annotation. (Impact of gene annotation choice on the quantification of RNA-seq data, Chisanga et al., BMC Bioinformatics, 2022)

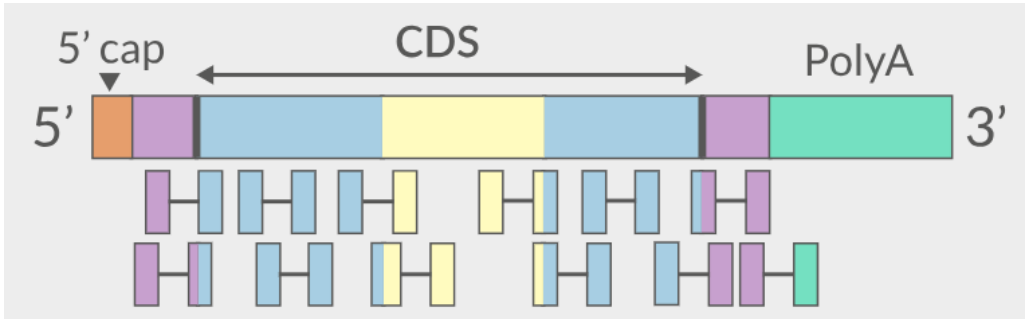
Mapping RNA-seq Reads

- Annotated reference is required (Annotation (GTF/GFF)
- To map junctions the algorithm needs to divide the sequencing reads and map portions independently
- Much more complex algorithms are required to identify alternative transcripts

Genome mapping

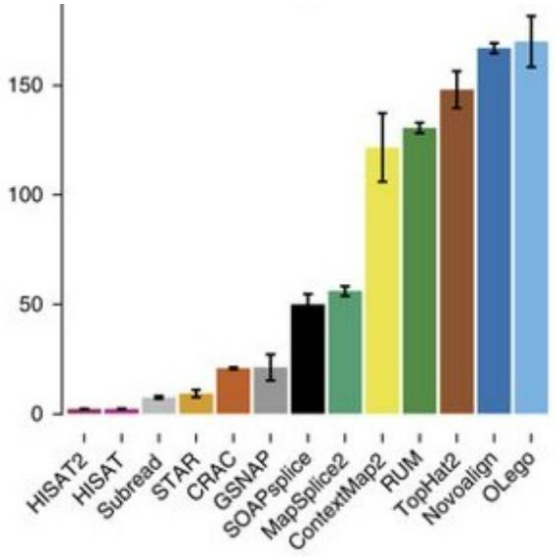


Transcriptome mapping



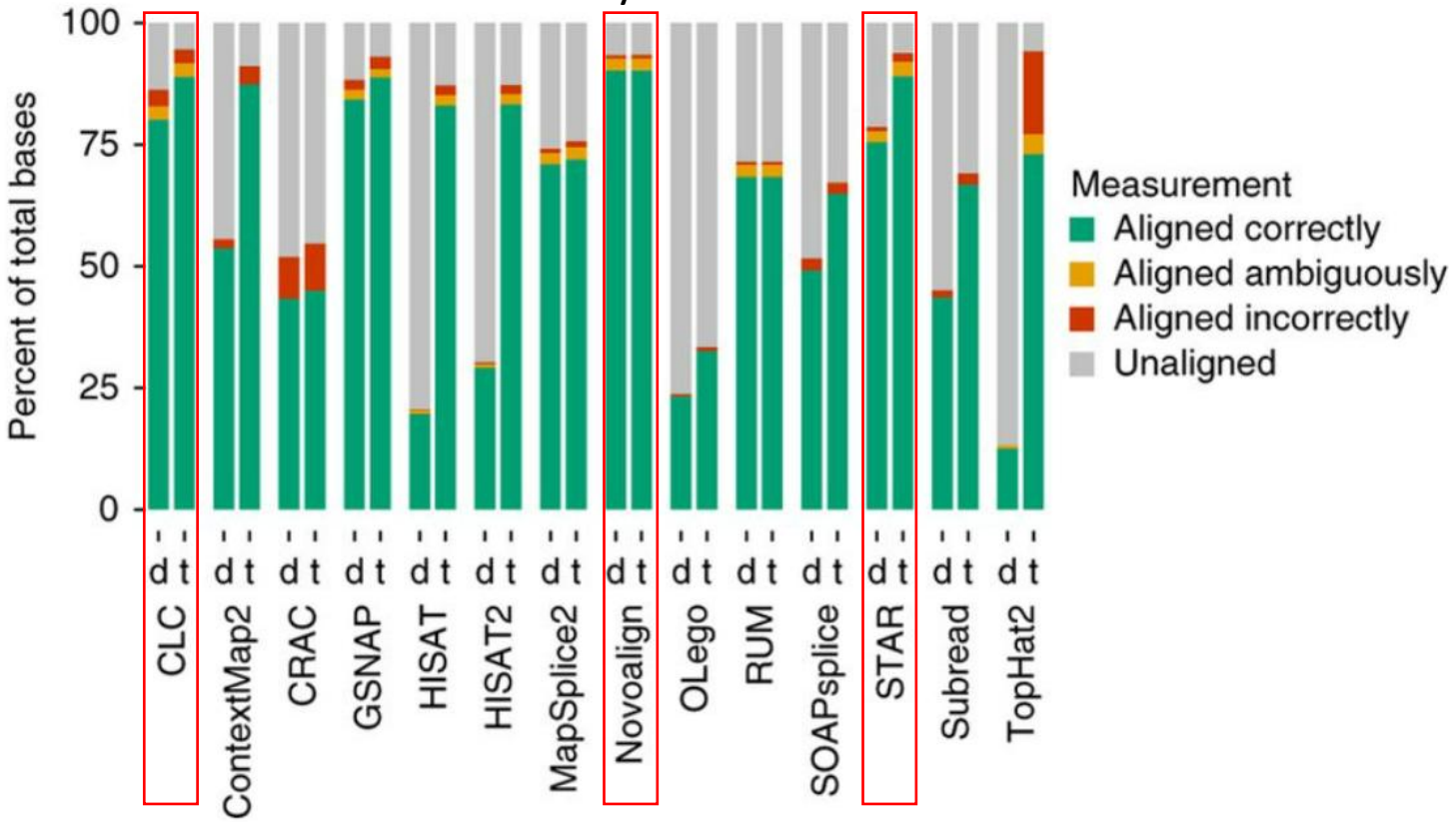
Which aligner to use?

Speed and memory usage



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

Accuracy



- The 'default' (d) vs the 'tuned' (t) alignments
- Different aligner comparison
- Choose algorithms perform well with default settings

Baruzzo, Giacomo, et al. "Simulation-based comprehensive benchmarking of RNA-seq aligners." Nature methods , 2017

Alignment Files

- SAM is the standard alignment file format generated from all mappers
  - **S**equence **A**lignment/ **M**ap format

- Alignments are stored in a BAM file (binary version of SAM)
  - Indexed to be read by other tools and genome browsers

- SAMtools is used to convert between SAM and BAM
  - http://samtools.sourceforge.net/

Example:

```
@HD VN:1.5 SO:coordinate
@SQ SN:ref LN:45
r001   99 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002    0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003    0 ref  9 30 5S6M       *  0   0 GCCTAAGCTAA       * SA:Z:ref,29,-,6H5M,17,0;
r004    0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003 2064 ref 29 17 6H5M       *  0   0 TAGGC             * SA:Z:ref,9,+,5S6M,30,1;
r001  147 ref 37 30 9M         =  7 -39 CAGCGGCAT         * NM:i:1
```

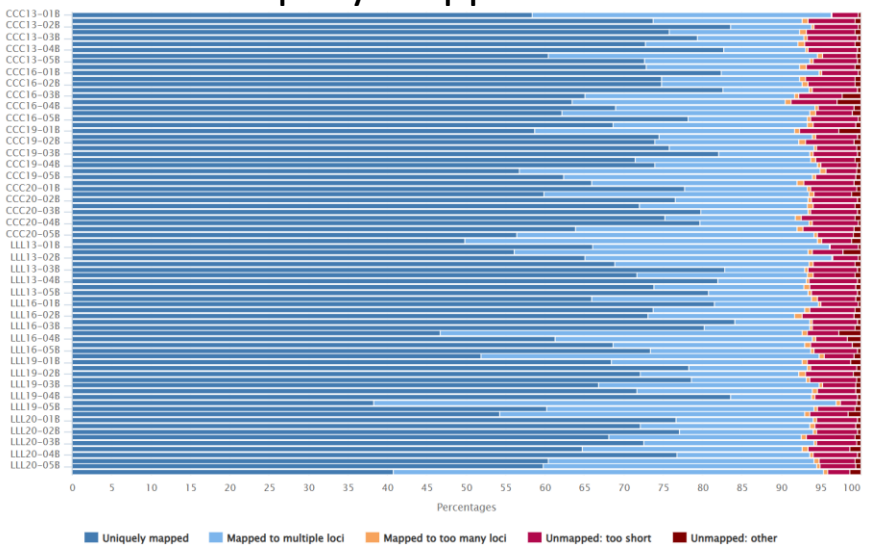
http://samtools.github.io/hts-specs/SAMv1.pdf

Alignment QC

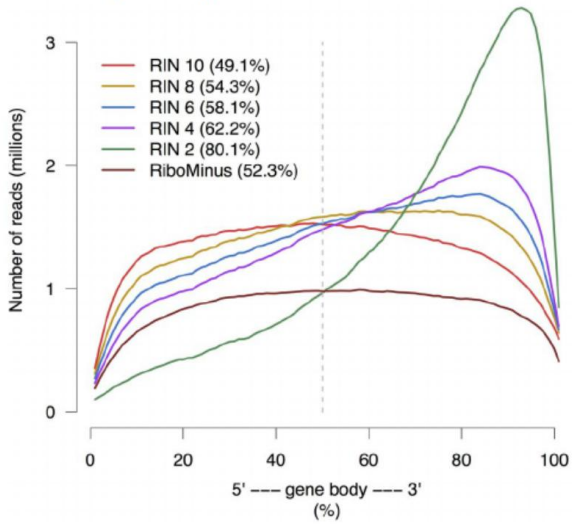
- Number of reads mapped/unmapped/paired etc
- Uniquely mapped
- Insert size distribution
- Coverage
- Gene body coverage
- Biotype counts / Chromosome counts
- Counts by region: gene/intron/non-genic
- Sequencing saturation
- Strand specificity

Tools: STAR (final log file), samtools > stats, bamtools > stats, QoRTs, RSeQC, Qualimap

Uniquely mapped reads



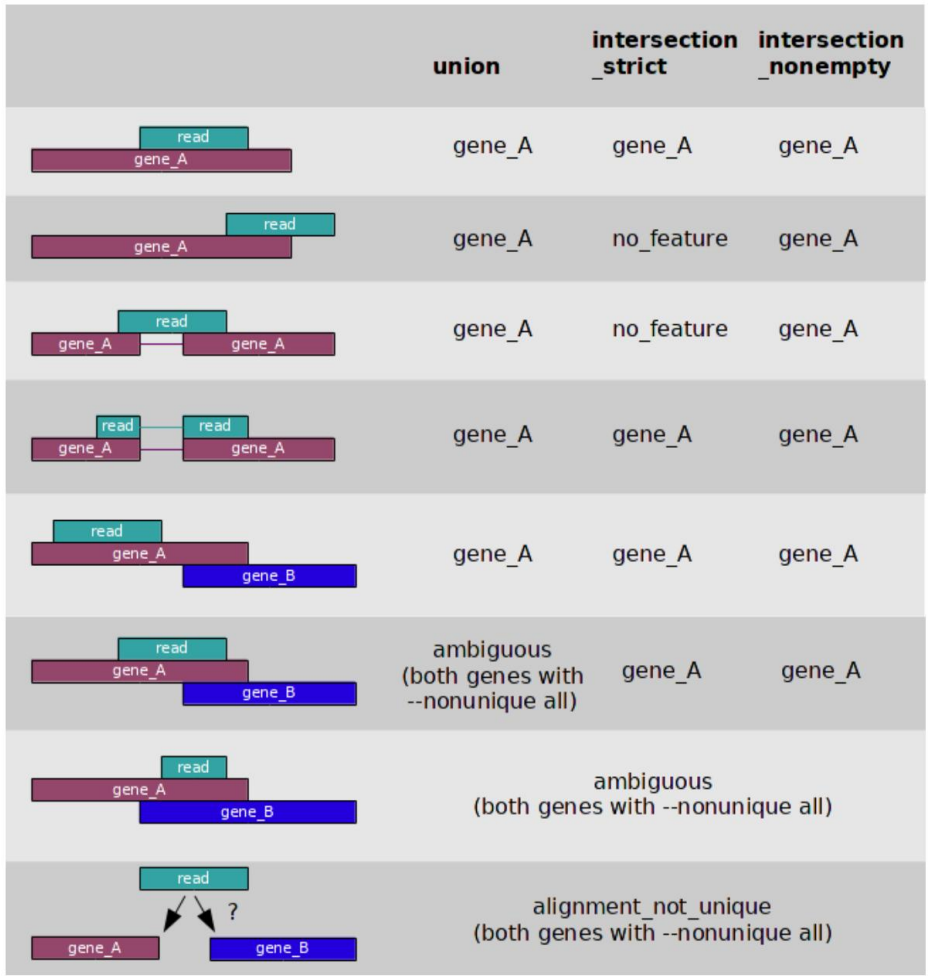
Gene body coverage



Quantification

- Read counts = gene expression
- Quantification at different levels: exon, transcript, gene
- Intersect with gene models

HTSeq



Errors in RNA- Seq quantification affect genes of relevance to human disease. Robert & Watson Genome Biol. 2015

Empirical assessment of analysis workflows for differential expression analysis of human samples using RNA-seq. Williams et al., BMC Bioinformatics 2017

Multi-mapped reads

- How to handle multi-mapped reads
  - Discard (featureCounts, HTSeq – only gene-level counts)
  - Probabilistic assignment (Rcount, Cufflinks)
  - Probabilistic assignment by Expectation Maximization (RSEM)

- Has perhaps the largest impact on the ultimate results

PCR duplicates

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

Gene ID	sample1	sample2	sample3
Gene 1	1	1	0
Gene 2	23	15	4
Gene 3	539	856	576
…	4648	4888	4980
Gene n	2	19	9

Errors in RNA- Seq quantification affect genes of relevance to human disease. Robert & Watson Genome Biol. 2015
Empirical assessment of analysis workflows for differential expression analysis of human samples using RNA-seq. Williams et al., BMC Bioinformatics 2017

Normalization

- Raw read counts alone are not sufficient to compare expression levels among samples
  - transcript length
  - Sequencing depth (total number of reads)
  - sequencing biases

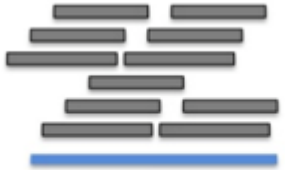
A) Long vs. short transcript

Read Counts = 12, Depth = 3X,

Read Counts = 5, Depth = 3X

B) Deep vs. shallow sequencing depth

Sample A
30 million reads



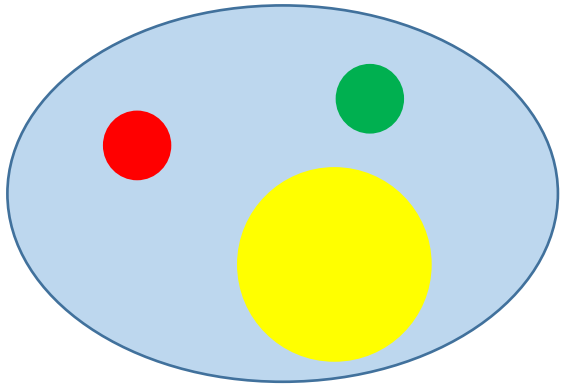
Read Counts = 11, Depth = 5X

Sample B
15 million reads

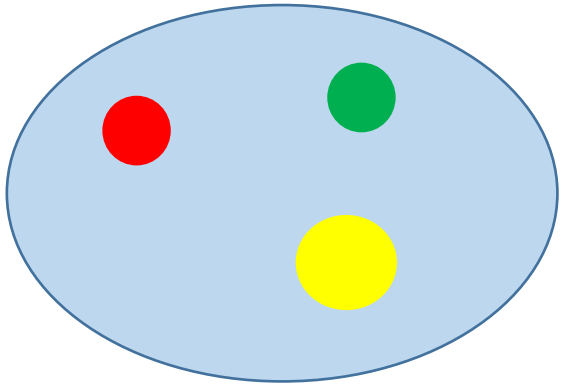
Read Counts = 5, Depth = 3X

C) Difference in RNA composition.

Sample A



Sample B



Within-sample normalization

- CPM = Counts per million
- RPM = Reads per million

$$\text{RPM or CPM} = \frac{\text{Number of reads mapped to gene} \times 10^6}{\text{Total number of mapped reads}}$$

- normalizes only for sequencing depth

- suitable for sequencing protocols that generate reads independent of gene length

- FPKM = fragments per kilobase of transcript per million fragments mapped
- RPKM = reads per kilobase per million mapped reads
- TPM = Transcripts per million

$$\text{FPKM}_i = \frac{X_i}{\left(\frac{\tilde{l}_i}{10^3}\right)\left(\frac{N}{10^6}\right)}$$

- Normalize for feature length and sequencing depth

Where:
X_i = counts in feature of interest
l_i = length of feature
N = total number of reads

Between-sample normalization

- TMM = Trimmed mean of M values (edgeR)
- Trim (remove) the extreme M-values (representing highly differentially expressed genes) from both ends of the distribution.
- accounts for differences in RNA composition between samples
- effective in normalization of samples with diverse RNA repertoires (e.g. samples from different tissues).

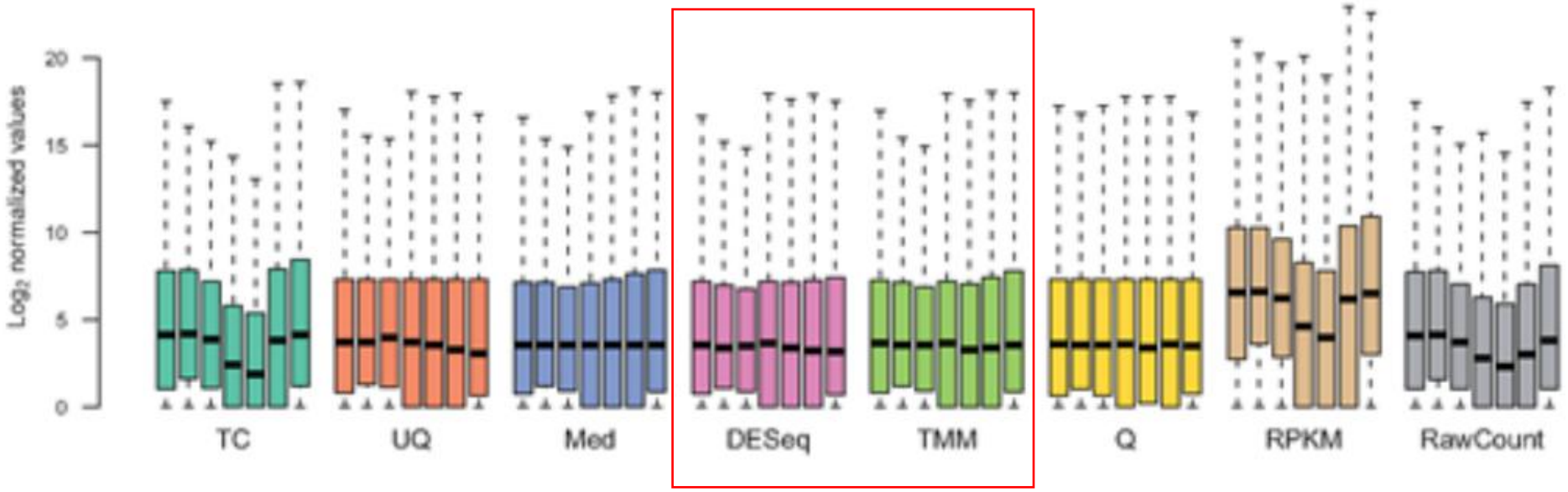
- median-of-ratios method = DESeq2 normalization
- use the median of the ratios of observed counts to a pseudo-reference sample as size factor to scale the counts
- Normalize sequencing depth

Both methods:

- do not consider gene length for normalization as it assumes that the gene length would be constant between the samples
- assume that most of the genes are not differentially expressed

Which Normalization Method?

- Median of Ratios (DESeq2) and TMM (edgeR) perform the best



A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. Dillies, Marie-Agnes, et al., Briefings in bioinformatics (2013)
Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. Evans et al., Briefings in bioinformatics (2017)
Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. Wagneret al., Theory in biosciences (2012)

Reproducibility

Remove lowly expressed genes < 10 reads
Normalization using VST, VOOM, RLOG, TMM etc

- Sample-sample clustering heatmap
- PCA
- Batch effects
- Outlier detection

Explore Your Data: Distances Among Samples

Euclidian Distances

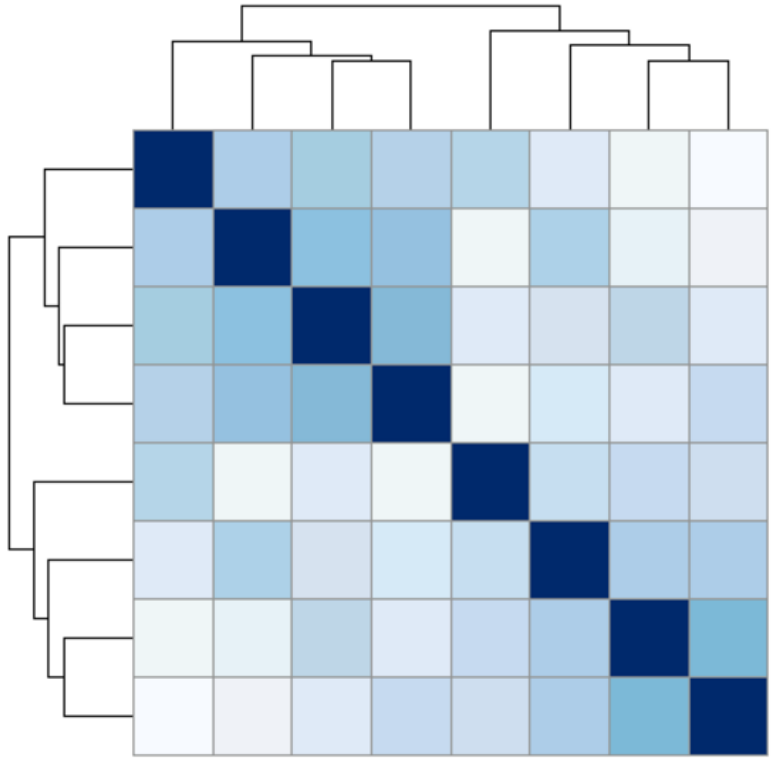
$$d_{(p,q)} = \sqrt{(p_1 - q_1)^2 + \cdots + (p_n - q_n)^2}$$

where:
p_n = normalized count for gene n in first sample of pair
q_n = normalized count for gene n in second sample of pair

Distance Matrix

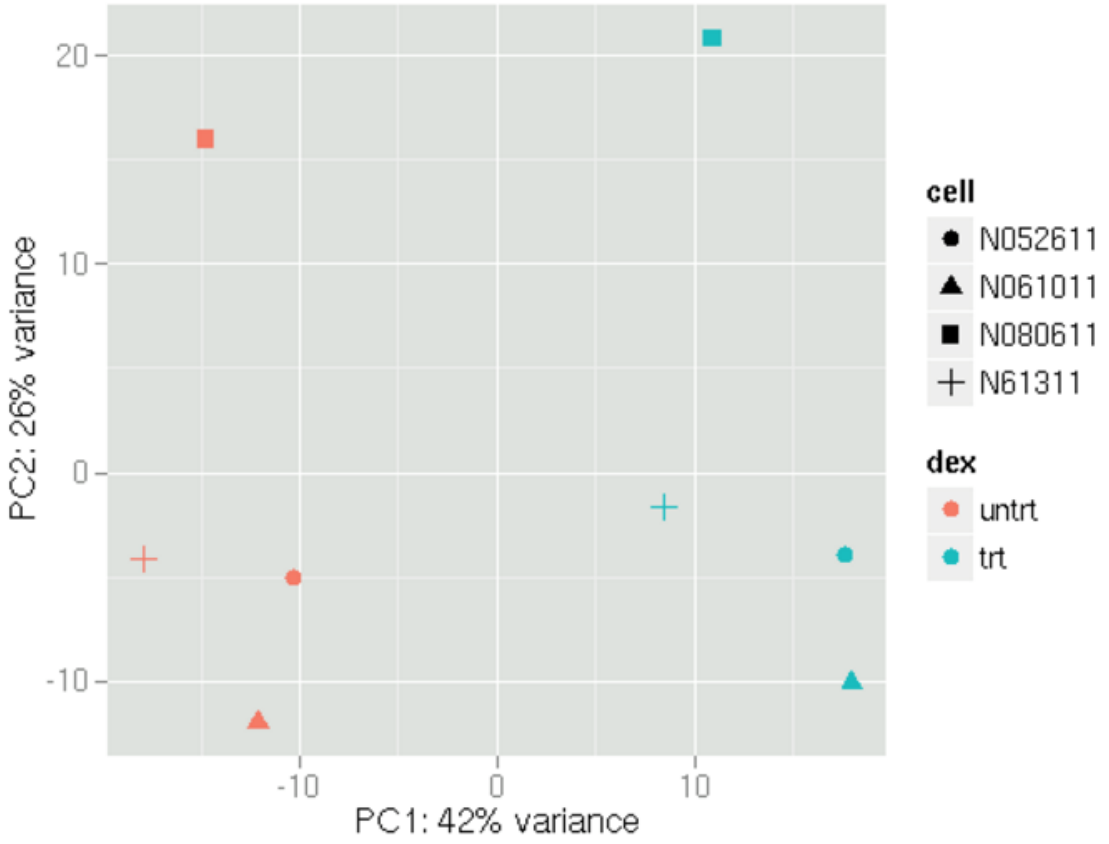
```
sampleDists <- dist( t( assay(rld) ) )
sampleDists

##            SRR1039508 SRR1039509 SRR1039512 SRR1039513 SRR1039516 SRR1039517 SRR1039
## SRR1039509   46.25524
## SRR1039512   39.94490   55.67572
## SRR1039513   63.36642   45.19462   49.30007
## SRR1039516   45.28129   59.89304   44.32383   64.54450
## SRR1039517   65.34730   52.25475   60.05523   50.64861   48.05714
## SRR1039520   40.20215   58.19904   37.35413   59.19401   47.15396   64.44641
## SRR1039521   64.09339   45.70177   58.59277   37.10803   66.36711   53.09669   50.72
```

Tools: pheatmap



Heatmap of Distances

Explore Your Data: Principle Components Analysis

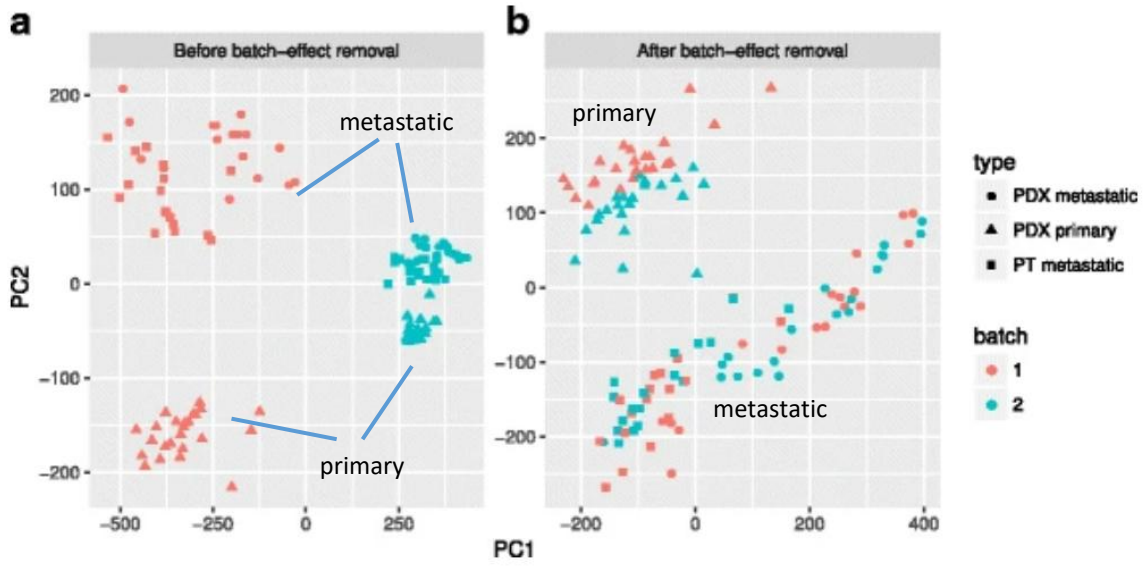


- What pattern(s) do you see?
- What expectations do you have given your experimental design?

Batch effects

- Batch effects (or “unwanted variations”): Variations in reagents, supplies, instruments and operators may introduce random or systematic errors at any step of RNA-seq data generation.

- Batch effects correction vs. Batch effects modeling



Tools: Surrogate Variable Analysis (SVA), PVCA, BatchQC, ComBat, Remove Unwanted Variation (RUV)

Outliers

- Outliers: True biological differences or technical failures during the process of sample preparation could lead to extreme deviation of a sample from samples of the same treatment group (biological replicates).

- Technical outliers
- Biological outliers

Chen *et al. BMC Bioinformatics* (2020) 21:269
https://doi.org/10.1186/s12859-020-03608-0

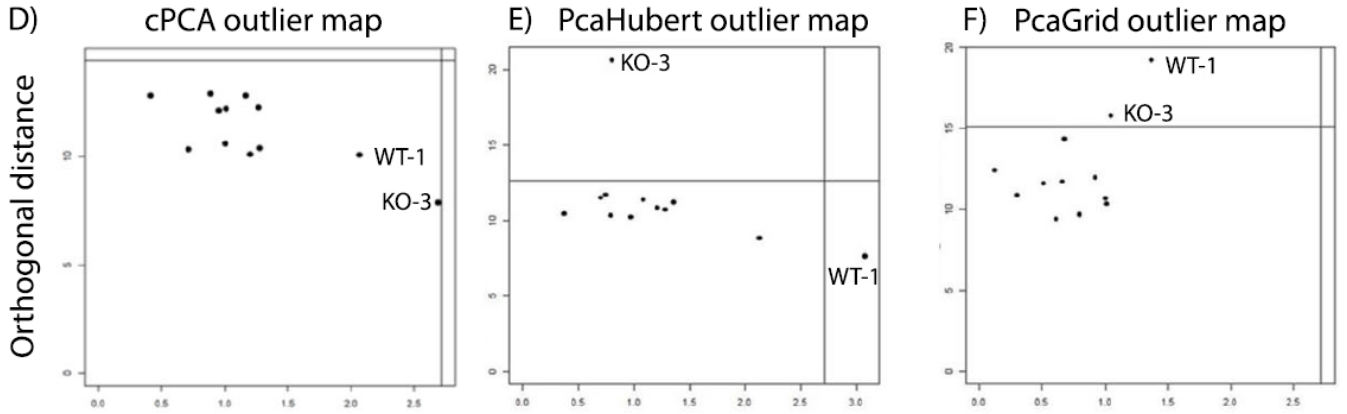
BMC Bioinformatics

RESEARCH ARTICLE Open Access

Robust principal component analysis for accurate outlier sample detection in RNA-Seq data

Xiaoying Chen[1], Bo Zhang[2], Ting Wang[3,4], Azad Bonni[1] and Guoyan Zhao[1*]

Classical principal component analysis	Robust principal component analysis (rPCA)



- PcaGrid achieved 100% sensitivity and 100% specificity in all the tests

Differential Expression Analysis

Method	Normalization	Need Replicates?	Input	DE Statistic	Availability
edgeR	Library size	Yes	Raw counts	Empirical Bayesian estimation based on Negative binomial distribution	R/Bioconductor
DESeq	Library size	No	Raw counts	Negative binomial distribution	R/Bioconductor
baySeq	Library size	Yes	Raw counts	Empirical Bayesian estimation based on Negative binomial distribution	R/Bioconductor
LIMMA	Library size	Yes	Raw counts	Empirical Bayesian estimation	R/Bioconductor
CuffDiff	RPKM	No	RPKM	Log ratio	Standalone

Differential Expression Analysis Results

Gene or Transcript

Mean Expression Values

Significance: use adjusted pvalue rather than raw pvalue

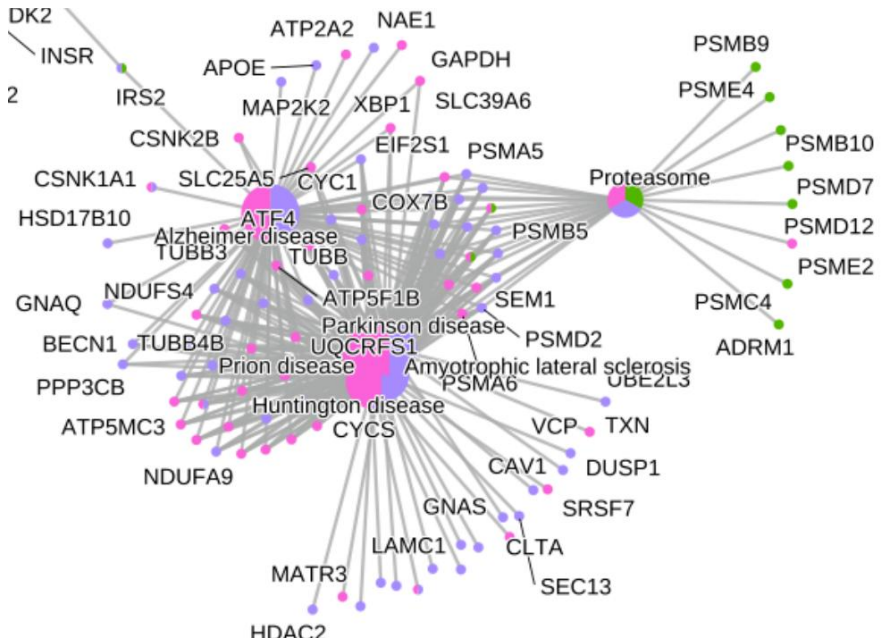
id	baseMean	baseMeanA	baseMeanB	foldChange	log2FoldChange	pval	padj
S100A8	1155.68	313.37	1998.00	6.38	2.67	1.29334E-23	3.38532E-19
S100A2	936.33	273.45	1599.20	5.85	2.55	2.47779E-20	3.2428E-16
NES	151.28	12.97	289.58	22.32	4.48	2.84523E-17	2.48246E-13
PSCA	1032.32	373.25	1691.38	4.53	2.18	2.44806E-16	1.60195E-12
IFI6	9349.03	4582.81	14115.24	3.08	1.62	3.3797E-15	1.76927E-11
IFI44L	1096.73	482.03	1711.42	3.55	1.83	1.73567E-12	7.57187E-09
KRT6A	932.56	404.19	1460.92	3.61	1.85	4.05644E-12	1.51682E-08
SBSN	195.29	50.90	339.68	6.67	2.74	7.55713E-11	2.4726E-07
KLHDC7B	12198.92	6986.00	17411.84	2.49	1.32	9.41463E-11	2.73809E-07

- **Fold Change:** measurement of the changing magnitude (effect size)
  - **FC =** baseMeanB/baseMeanA
  - Typically log2(FC) is reported

- multiple comparison corrections and **padj:** FDR adjusted p-value (or q-value)

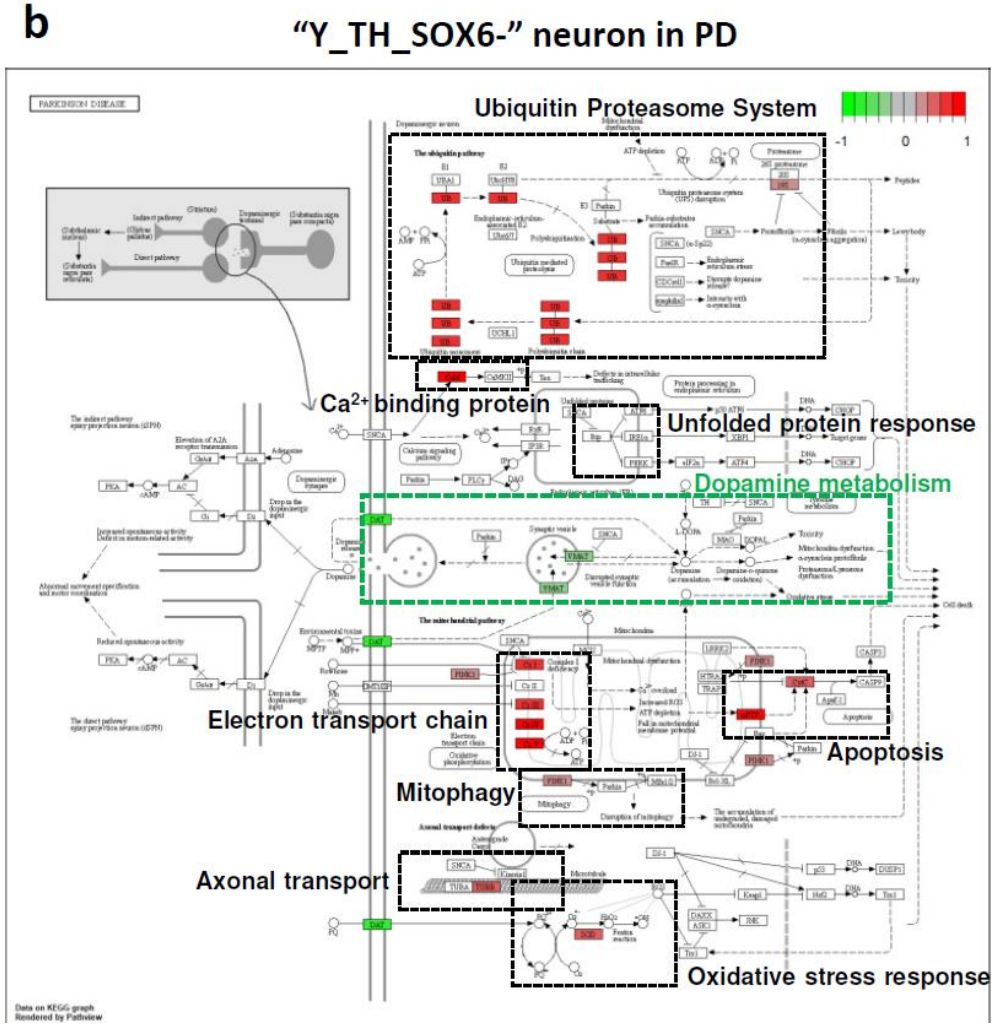
Biological interpretation

- Gene ontology enrichment analysis
- KEGG pathway analysis
- Reactome databases
- Gene set enrichment analysis (GSEA)



Tools: DAVID, clusterProfiler, ClueGO, Enrichr

Mapping DEGs to “Parkinson disease” KEGG pathway



Which analysis pipeline should I use?

- Computing resources: *Cufflinks-Cuffdiff* demands the highest vs. *Kallisto-Sleuth* the least.

- Gene expression values, fold change, *p* and *q* values of differential expression (DE) analysis are highly correlated among procedures (using *HTseq* for quantification).

- Major differences come from genes with particularly high or low expression levels.
  - *HISAT2-StringTie-Ballgown* is more sensitive to genes with low expression levels
  - *Kallisto-Sleuth* may only be useful to evaluate genes with medium to high abundance.

- *Number of DEGs:*
  - *StringTie-Ballgown* the least number of DEGs.
  - *HTseq-DESeq2*, *-edgeR* or *-limma* generally produces more DEGs.
  - *Cufflinks-Cuffdiff* and *Kallisto-Sleuth* varies in different datasets.

- The choice of the method (or even the version of a software package) can markedly affect the outcome of the analysis
- Thoroughly document the parameters and version numbers of programs used

A comparison of transcriptome analysis methods with reference genome, Xu Liu et al., BMC Genomics. 2022

Summing It All Up

- Sound experimental design
  - More biological replicates? How many?
  - What technology is most appropriate?
  - Plan carefully about lib prep, sequencing to avoid confounding factors
- QC! QC everything at every step
  - Are the quality good enough for raw data, mapping, quantification, samples?
  - Are there technical issues?
  - Any outliers?
  - Discard low quality bases, reads, genes and samples
- Normalization
  - Choose the method that is appropriate for your experiment.
- Calculate differential expression
  - Choose the method that is most appropriate for your experiment.
- Interpret the results in terms of biological or physiological knowledge
  - Does it make biological sense?