Research Institute for Medicines

UNIVERSIDADE DE LISBOA

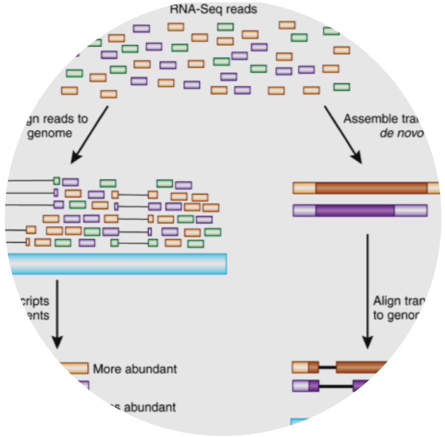
FACULDADE DE FARMÁCIA
Universidade de Lisboa

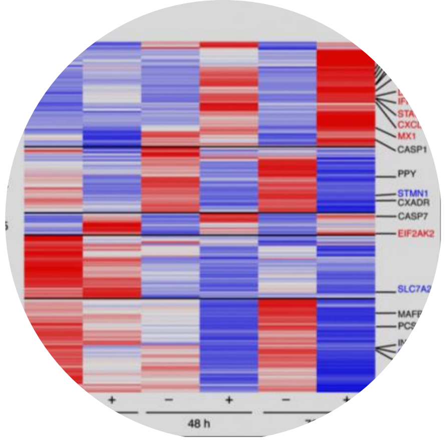
PMB2023

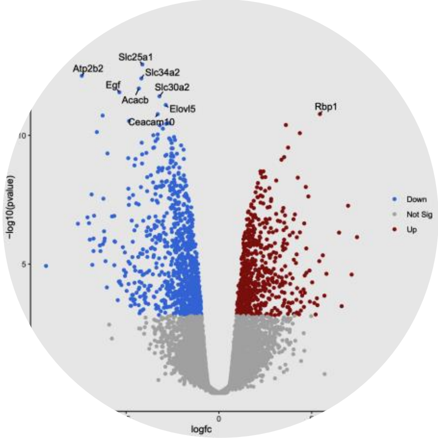
PATHOGEN MULTIOMICS AND BIOINFORMATICS

Rio Grande/RS 2023

Module 6: RNA-Seq and Transcriptomics

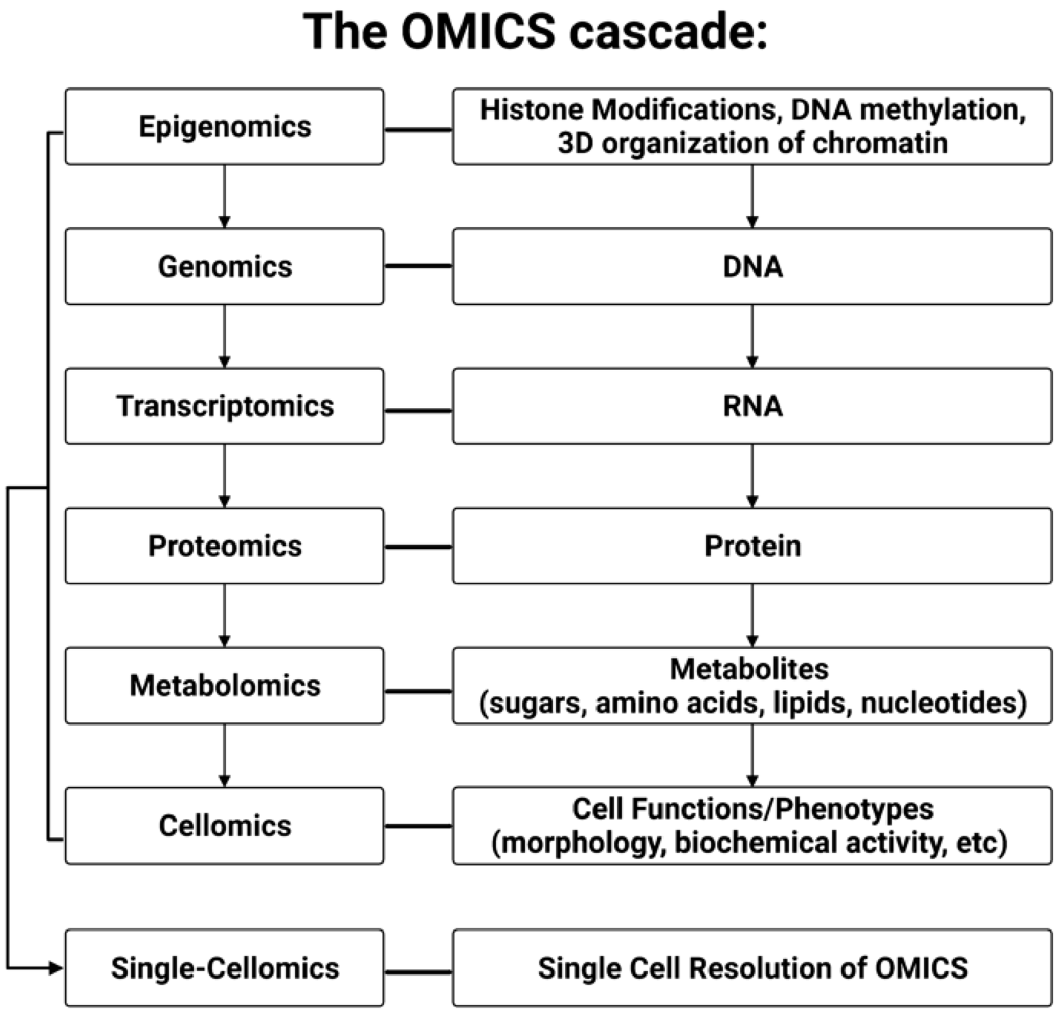






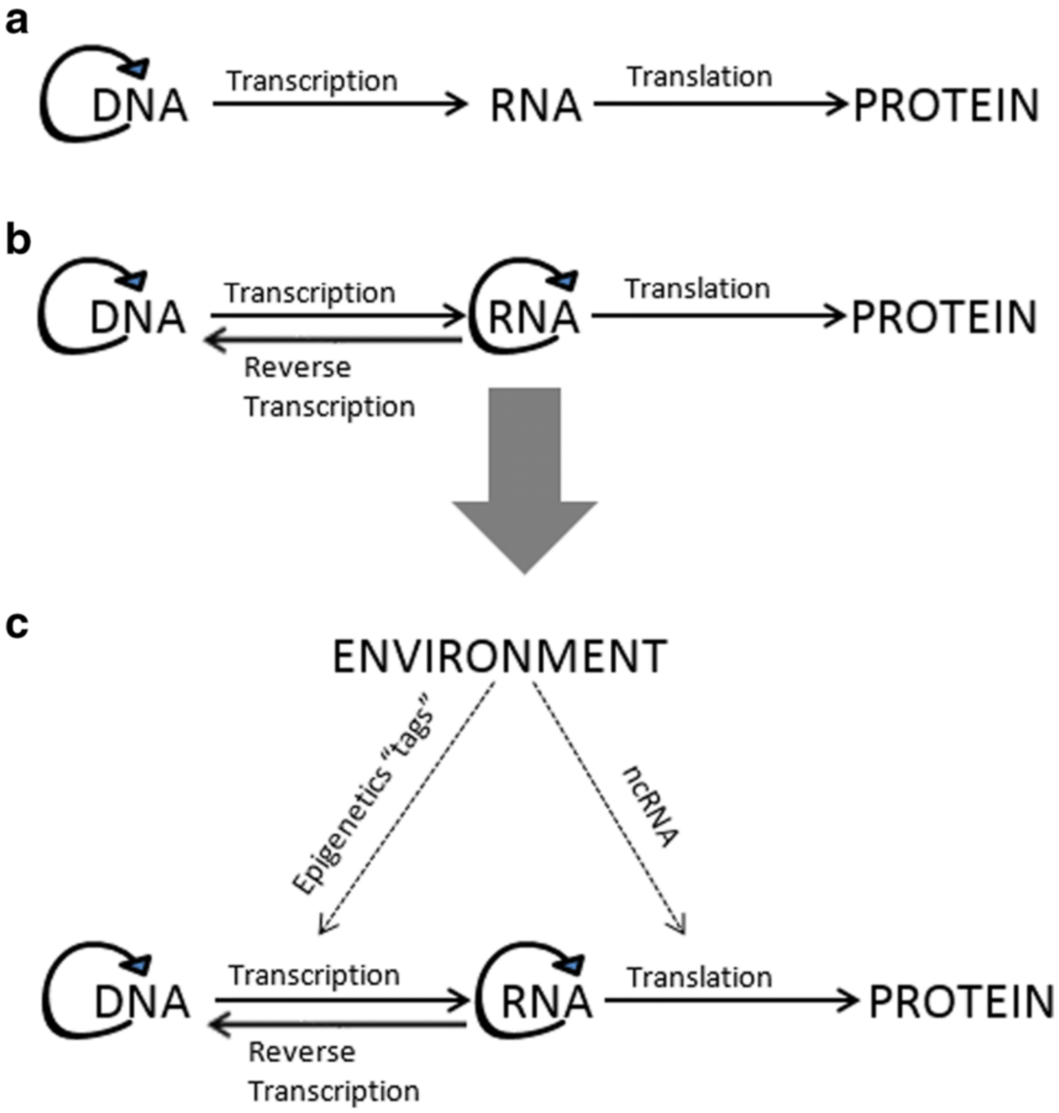
João Perdigão

Omics Cascade



The OMICS cascade:

Epigenomics	Histone Modifications, DNA methylation, 3D organization of chromatin
Genomics	DNA
Transcriptomics	RNA
Proteomics	Protein
Metabolomics	Metabolites (sugars, amino acids, lipids, nucleotides)
Cellomics	Cell Functions/Phenotypes (morphology, biochemical activity, etc)
Single-Cellomics	Single Cell Resolution of OMICS



Why using RNA-Seq?

Use of new sequencing technologies to capture and study the transcriptome

- **Identify novel transcripts**
- **Exon/transcript boundaries**
- **Splice junctions/alternative splicing**
- **Measure transcript abundance**
- **Gene expression differences across multiple samples (i.e. differential expression)**

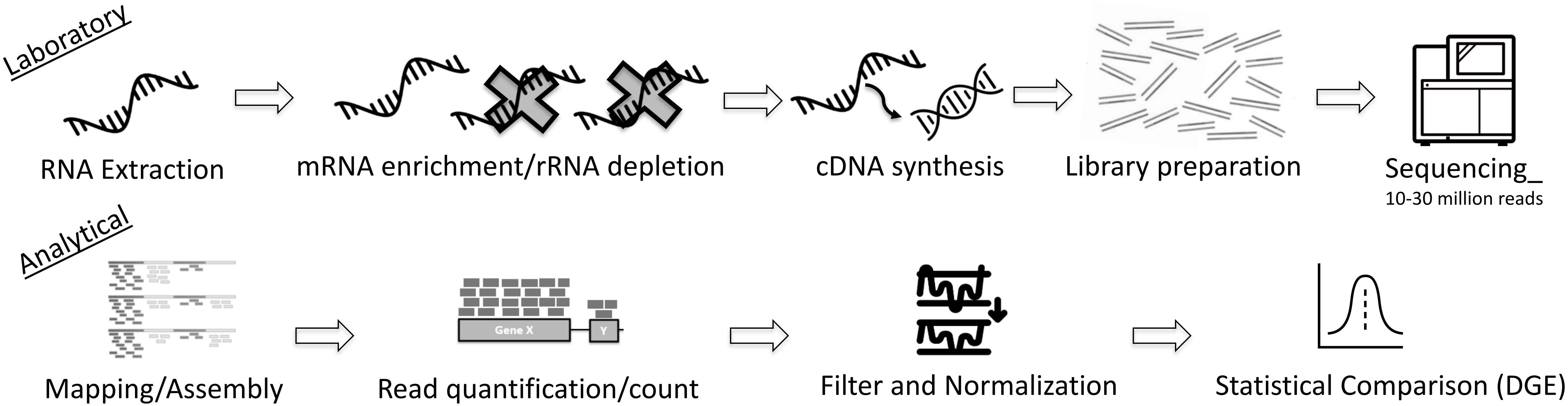
imed
Research Institute
for Medicines

RNA-Seq: General Workflow

What is RNA-Seq?

RNA-Seq consists of a method to analyze the transcriptomics of thousands of features in a single assay and, hence, evaluate and compare gene expression in a genome-wide manner.

Two main stages:



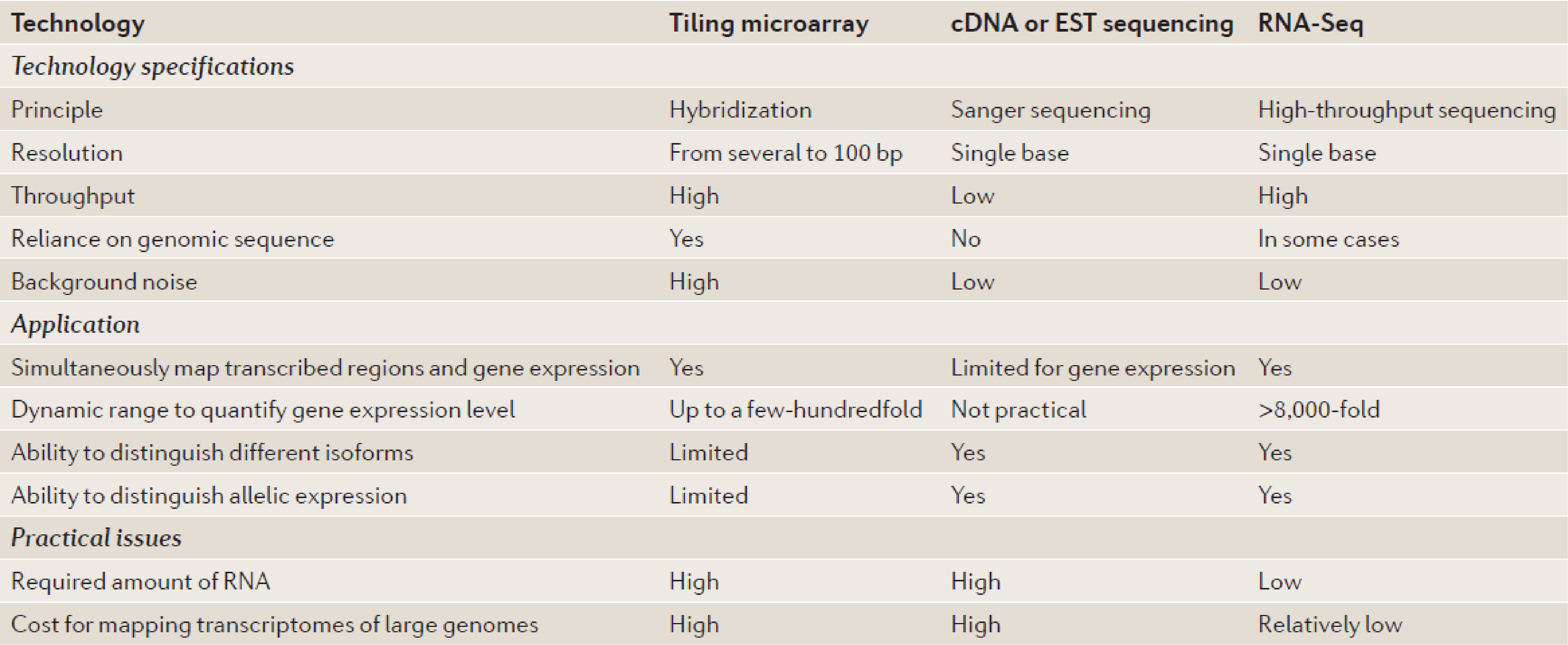
Laboratory

RNA Extraction → mRNA enrichment/rRNA depletion → cDNA synthesis → Library preparation → Sequencing_
10-30 million reads

Analytical

Mapping/Assembly → Read quantification/count → Filter and Normalization → Statistical Comparison (DGE)

RNA-Seq *vs* cDNA/EST *Seq vs* Microarrays



Technology	Tiling microarray	cDNA or EST sequencing	RNA-Seq
Technology specifications			
Principle	Hybridization	Sanger sequencing	High-throughput sequencing
Resolution	From several to 100 bp	Single base	Single base
Throughput	High	Low	High
Reliance on genomic sequence	Yes	No	In some cases
Background noise	High	Low	Low
Application			
Simultaneously map transcribed regions and gene expression	Yes	Limited for gene expression	Yes
Dynamic range to quantify gene expression level	Up to a few-hundredfold	Not practical	>8,000-fold
Ability to distinguish different isoforms	Limited	Yes	Yes
Ability to distinguish allelic expression	Limited	Yes	Yes
Practical issues			
Required amount of RNA	High	High	Low
Cost for mapping transcriptomes of large genomes	High	High	Relatively low

Wang *et al* 2009

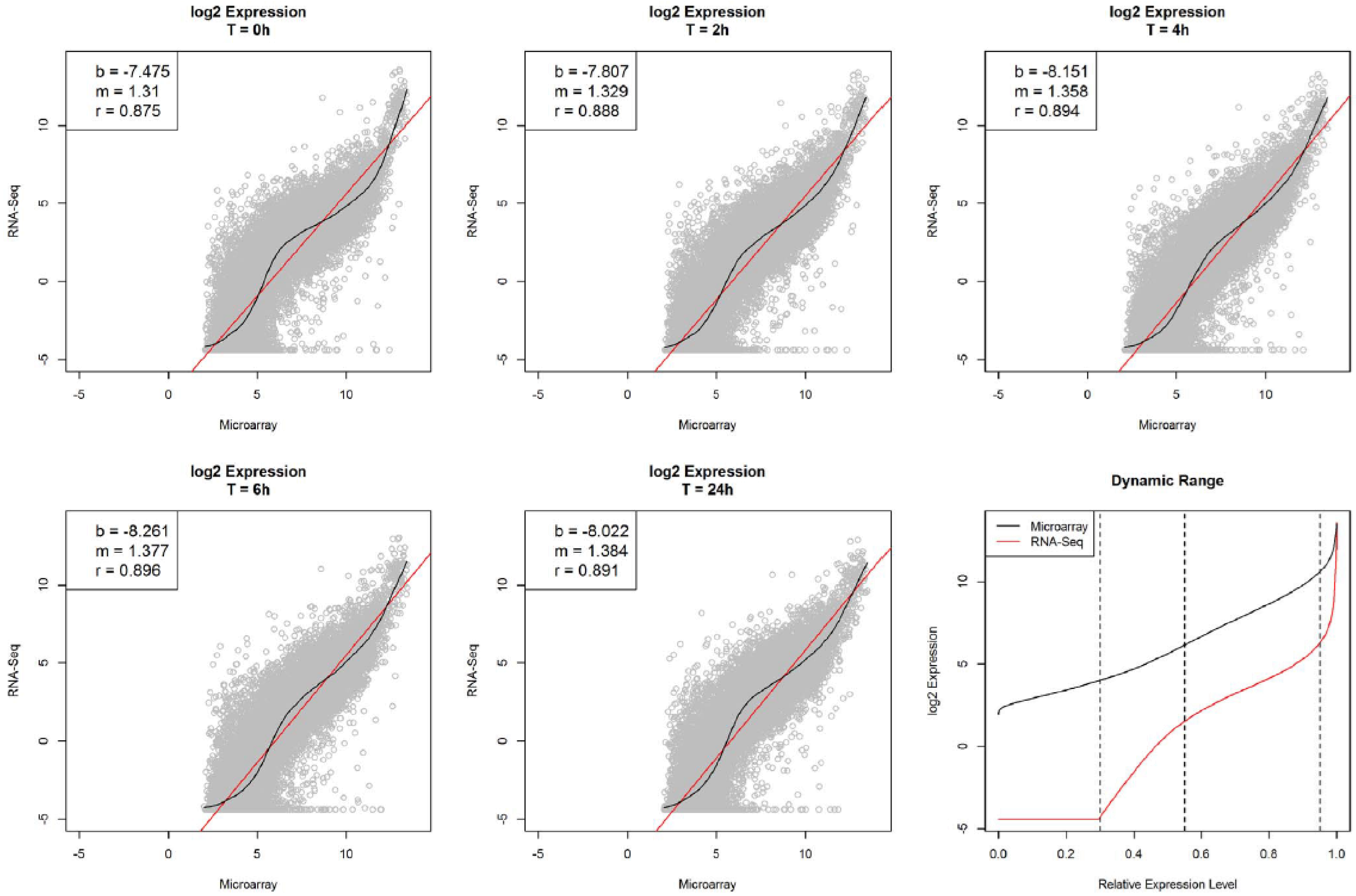
Ability to detect novel transcripts

Wider dynamic range

Higher specificity and sensitivity

Simple detection of rare and low-abundance transcripts

RNA-Seq *vs* Microarrays

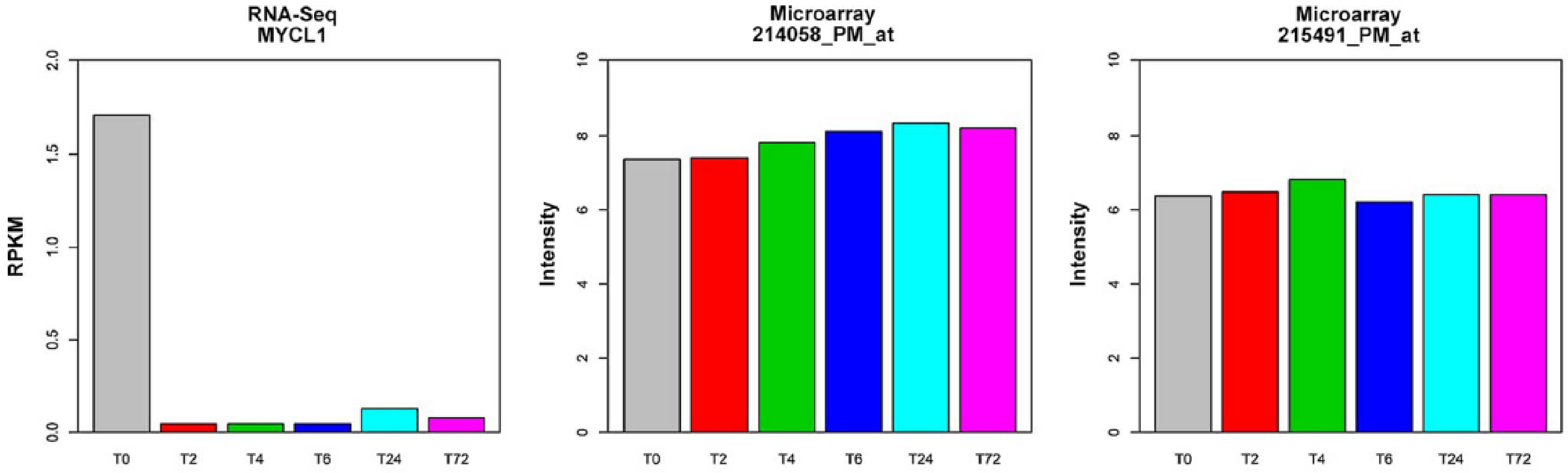


The plots show that the overall dynamic range of the 18,306 common genes generated by the two platforms is much broader in RNA-Seq (2.66105) than in microarray (3.66103).

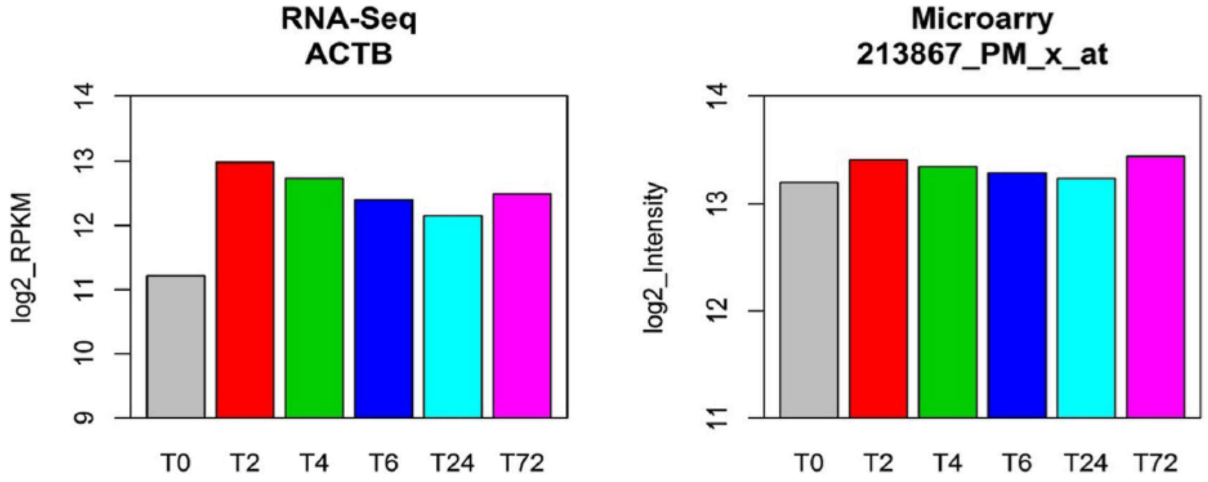
Zhao *et al* 2014

RNA-Seq *vs* Microarrays

RNA-Seq is able to detect subtle changes to the level of genes with low expression levels whereas microarrays are not



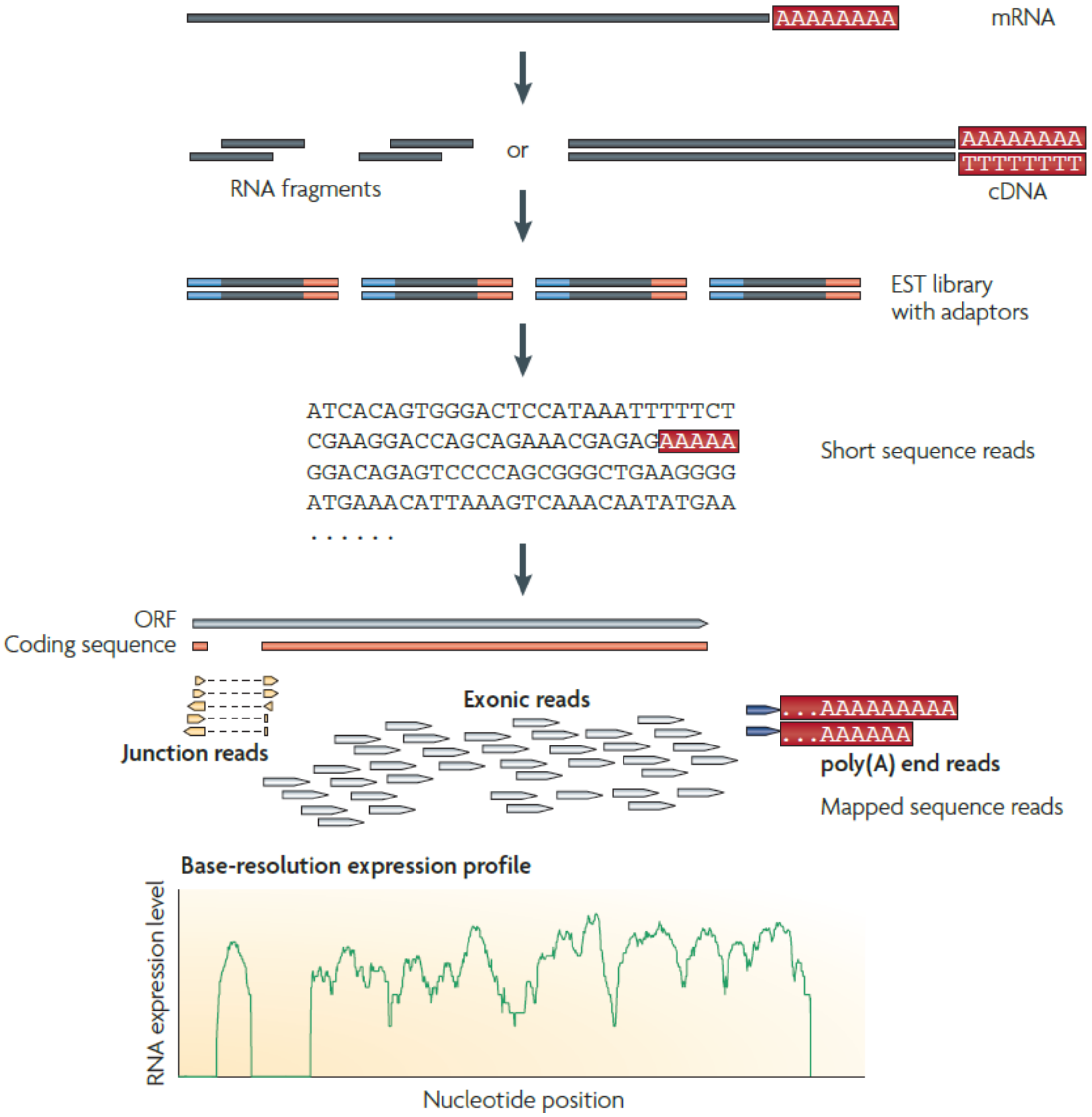
Zhao *et al* 2014



… Similarly RNA-Seq is able to detect expression level changes to highly expressed genes and microarrays are not (staturation).

Zhao *et al* 2014

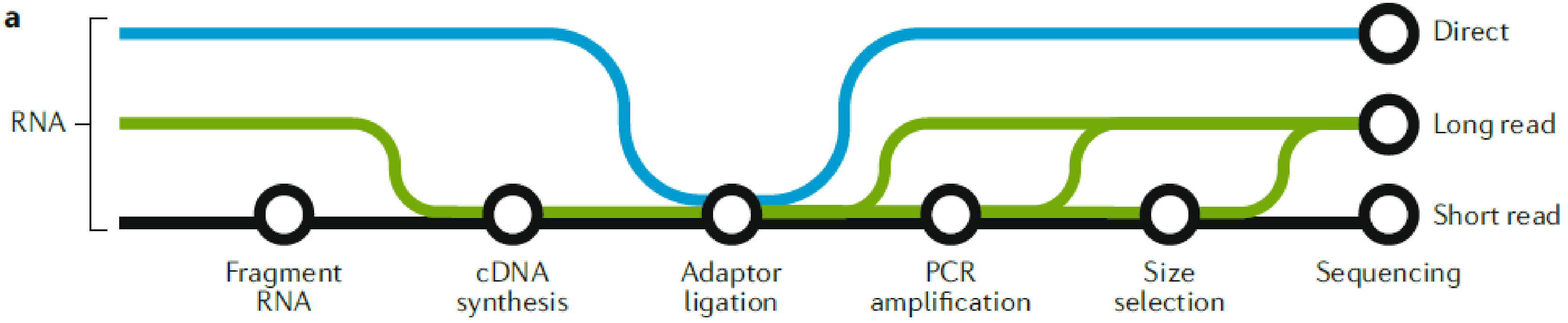
Library Preparation



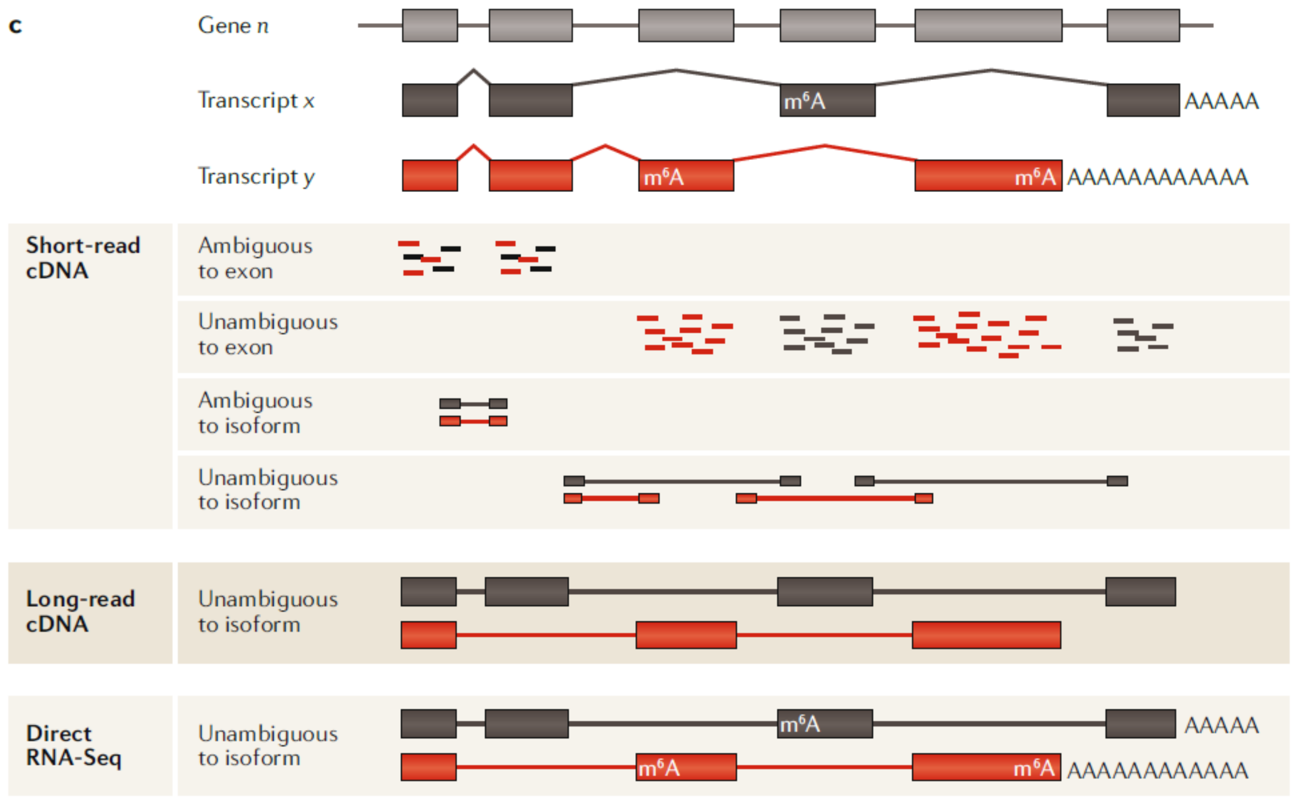
Aspects and factors to consider:

- RNA Source: Total RNA, mRNA, depletion of rRNA?
- Strand specific?
- Replicates?
  - – Technical (multiple libraries from the same sample)
  - – Biological (multiple samples from the same condition)
- Which platform?
- Multiple samples/multiplexing

Library Preparation: comparison between technologies and limitations



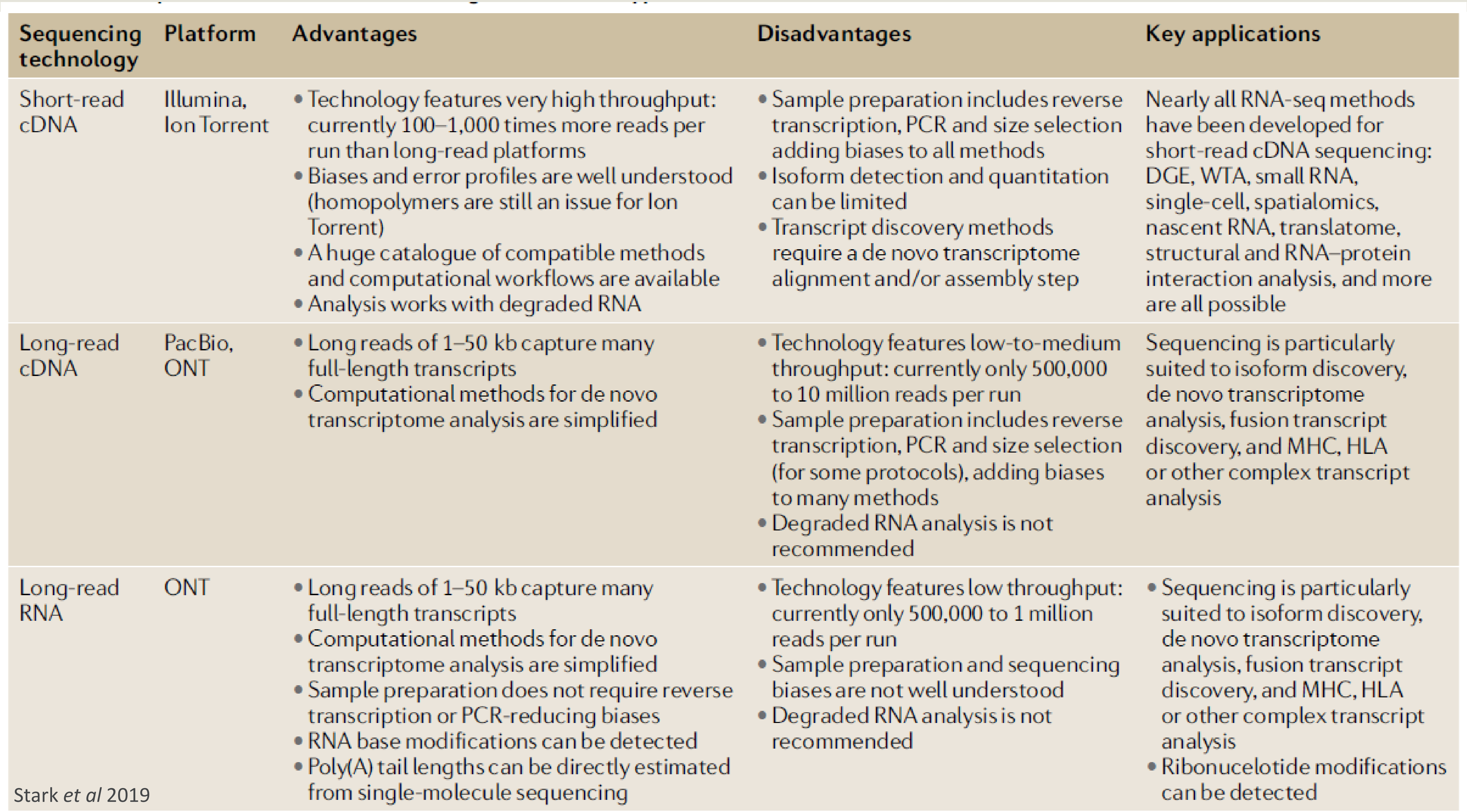
All long-read and short-read approaches require adpter ligation



Short-reads can be ambiguously mapped to diferente isoforms

Stark *et al* 2019

Library Preparation: comparison between technologies and limitations

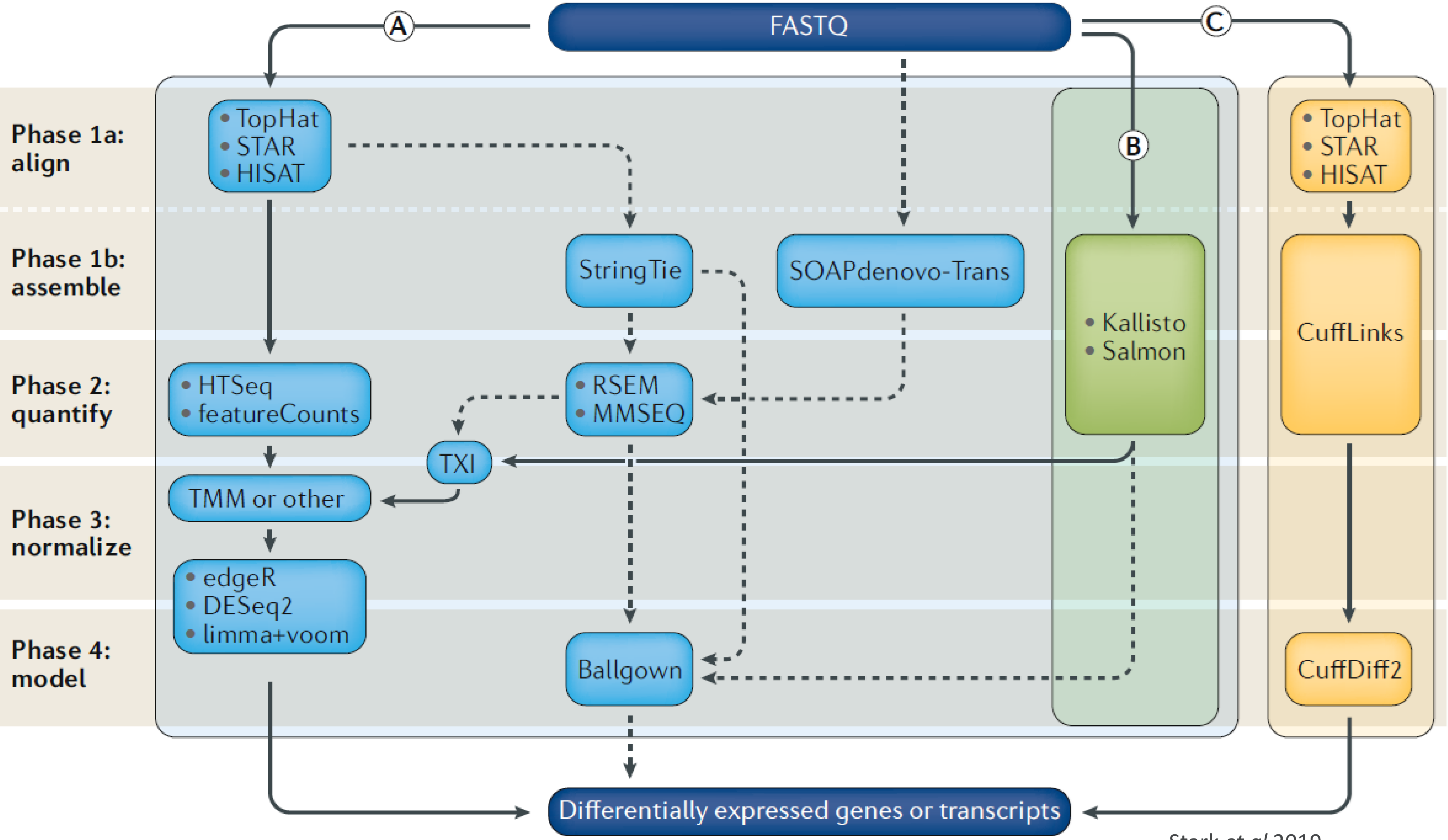


Sequencing technology	Platform	Advantages	Disadvantages	Key applications
Short-read cDNA	Illumina, Ion Torrent	• Technology features very high throughput: currently 100–1,000 times more reads per run than long-read platforms • Biases and error profiles are well understood (homopolymers are still an issue for Ion Torrent) • A huge catalogue of compatible methods and computational workflows are available • Analysis works with degraded RNA	• Sample preparation includes reverse transcription, PCR and size selection adding biases to all methods • Isoform detection and quantitation can be limited • Transcript discovery methods require a de novo transcriptome alignment and/or assembly step	Nearly all RNA-seq methods have been developed for short-read cDNA sequencing: DGE, WTA, small RNA, single-cell, spatialomics, nascent RNA, translatome, structural and RNA–protein interaction analysis, and more are all possible
Long-read cDNA	PacBio, ONT	• Long reads of 1–50 kb capture many full-length transcripts • Computational methods for de novo transcriptome analysis are simplified	• Technology features low-to-medium throughput: currently only 500,000 to 10 million reads per run • Sample preparation includes reverse transcription, PCR and size selection (for some protocols), adding biases to many methods • Degraded RNA analysis is not recommended	Sequencing is particularly suited to isoform discovery, de novo transcriptome analysis, fusion transcript discovery, and MHC, HLA or other complex transcript analysis
Long-read RNA	ONT	• Long reads of 1–50 kb capture many full-length transcripts • Computational methods for de novo transcriptome analysis are simplified • Sample preparation does not require reverse transcription or PCR-reducing biases • RNA base modifications can be detected • Poly(A) tail lengths can be directly estimated from single-molecule sequencing	• Technology features low throughput: currently only 500,000 to 1 million reads per run • Sample preparation and sequencing biases are not well understood • Degraded RNA analysis is not recommended	• Sequencing is particularly suited to isoform discovery, de novo transcriptome analysis, fusion transcript discovery, and MHC, HLA or other complex transcript analysis • Ribonucelotide modifications can be detected

Stark *et al* 2019

imed
Research Institute
for Medicines

Analytical Pipeline Overview

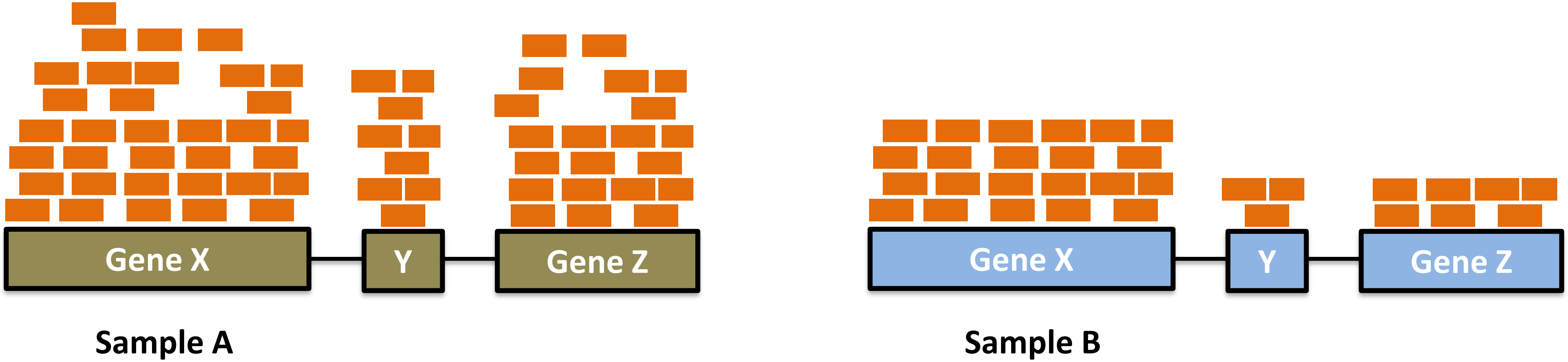


Stark *et al* 2019

Searching for Differentially Expressed Genes

For differential expression the number of reads mapping to each gene (read count) is used to evaluate expression levels…

In which sample is Gene X overexpressed?



Raw counts cannot be used to evaluate or compare expression beween samples! And within a sample?

Factors to consider:

- *Sequencing Depth*
- *Gene Length*
- *RNA Composition*

imed Research Institute for Medicines

Normalizing Raw Read Counts: CPM/RPM, RPKM/FKPM and TPM

The answer: Normalization - this is required for differential expression analysis vizualization, etc.

Some Normalization Methods:

CPM/RPM – Counts/Reads per million

$$CPM = \frac{No.reads\ mapped\ to\ gene\ x\ 10^6}{Total\ number\ of\ mapped\ reads}$$

RPKM/FPKM – Reads/fragments per kilobase million

$$Scaling\ Factor\ (SF) = \frac{Total\ number\ of\ mapped\ reads}{10^6}$$

$$RPM = \frac{No.reads\ mapped\ to\ gene}{SF}$$

$$RPMK = \frac{RPM}{gene\ length(Kbp)}$$

$$\boxed{RPMK = \frac{No.reads\ mapped\ to\ gene\ x\ 10^3\ x\ 10^6}{Total\ number\ of\ mapped\ reads\ x\ gene\ length(bp)}}$$

TPM – Transcripts per kilobase million

$$RPK = \frac{No.reads\ mapped\ to\ gene}{gene\ length(Kbp)}$$

$$Scaling\ Factor\ (SF) = \frac{\sum RPK}{10^6}$$

$$\boxed{TPM = \frac{RPK}{SF}}$$

imed Research Institute for Medicines

Normalizing Raw Read Counts: Median of Ratios (DESeq2)

Anders and Huber *Genome Biology* 2010, **11**:R106
http://genomebiology.com/2010/11/10/R106

Genome **Biology**

METHOD Open Access

Differential expression analysis for sequence count data

Simon Anders*, Wolfgang Huber

DESeq2 – Median of Ratios Method Normalization

Accounts for <u>sequencing depth</u> and <u>*RNA composition*</u>*… but not gene length*

1. Starting on raw counts, calculate the geometric mean for each gene across all sample – pseudo-reference;

Gene	Sample A	Sample B	Pseudo-reference
rpoB	1100	750	$\sqrt[2]{1100 \ x \ 750}$=908,30
eis	15	10	$\sqrt[2]{15 \ x \ 10}$=12,25

2. Calculate the ratio of each sample to the pseudo-reference;

Gene	Ratio Sample A	Ratio Sample B
rpoB	1100/908,30=1,21	750/908,30=0,83
eis	15/12,25=1,22	10/12,25=0,82

3. Calculate the normalization factor for each sample (**size factor**) by taking the median of all ratios;

<u>Normalization Factors:</u>
Sample A – Median(1,21; 1,22)= 1,215
Sample B - Median(0,83; 0,82)= 0,825

4. Normalized counts are obtained by dividing the raw count of each gene by the normalization factor;

Gene	Normalized Count A	Normalized Count B
rpoB	1100/1,215=905,35	750/0,825=909,09
eis	15/1,215=12,35	10/0,825=12,12

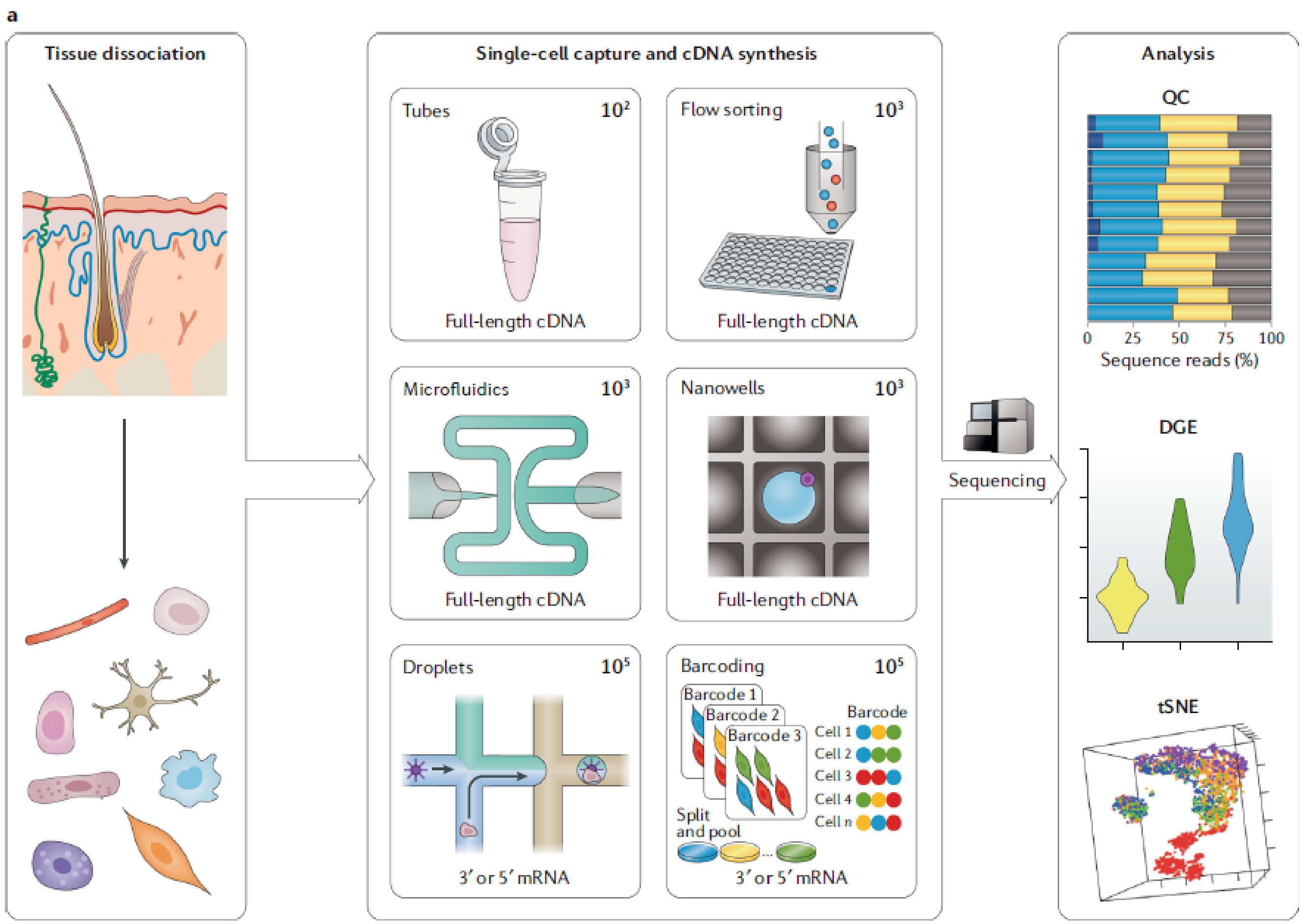
Comparison of Normalization Methods

Method	Factors Accounted			Applications		
	Sequencing Depth	Gene Length	RNA Composition	Within sample	Comparisons between samples	DE Analysis
CPM	✓	✗	✗	✗	✓	✗
RPKM/FPKM	✓	✓	✗	✓	✗	✗
TPM	✓	✓	✗	✓	✓	✗
Median of Ratios (*DESeq2*)	✓	✗	✓	✗	✓	✓
Trimmed Mean of M Values (*EdgeR*)	✓	✓	✓	✓	✓	✓

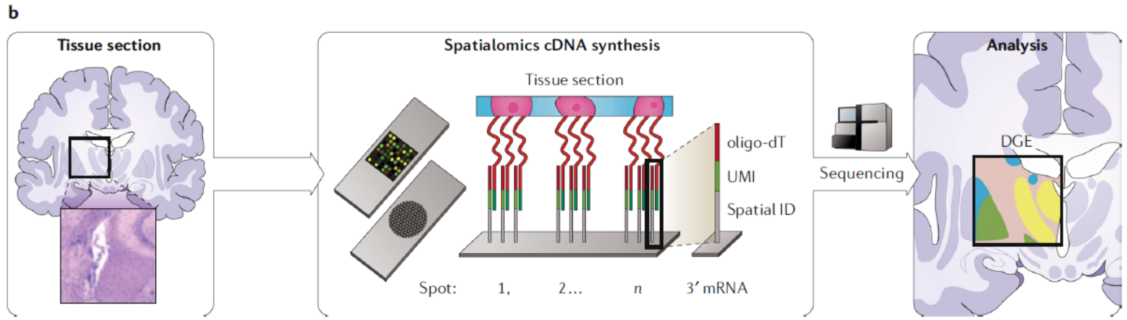
Other than bulk RNA-Seq …

Single-cell RNA-Seq

Spatialomics



a
Tissue dissociation
Single-cell capture and cDNA synthesis
Tubes
10^2
Full-length cDNA
Flow sorting
10^3
Full-length cDNA
Microfluidics
10^3
Full-length cDNA
Nanowells
10^3
Full-length cDNA
Droplets
10^5
3′ or 5′ mRNA
Barcoding
10^5
Barcode 1
Barcode 2
Barcode 3
Barcode
Cell 1
Cell 2
Cell 3
Cell 4
Cell n
Split and pool
3′ or 5′ mRNA
Sequencing
Analysis
QC
0
25
50
75
100
Sequence reads (%)
DGE
tSNE



b
Tissue section
Spatialomics cDNA synthesis
Tissue section
oligo-dT
UMI
Spatial ID
Spot:
1,
2…
n
3′ mRNA
Sequencing
Analysis
DGE