Analysis of bulk RNA-seq data

Analysis of Next-Generation Sequencing Data

An introduction to transcriptome or gene expression

05 June 2023
Phuc Loi Luu, PhD
Loi.lp@pacificinformatics.com.vn

Analysis of bulk RNA-seq data

Analysis of Next-Generation Sequencing Data

Friederike Dündar

Applied Bioinformatics Core

Slides at https://bit.ly/2T3sjRg[1]

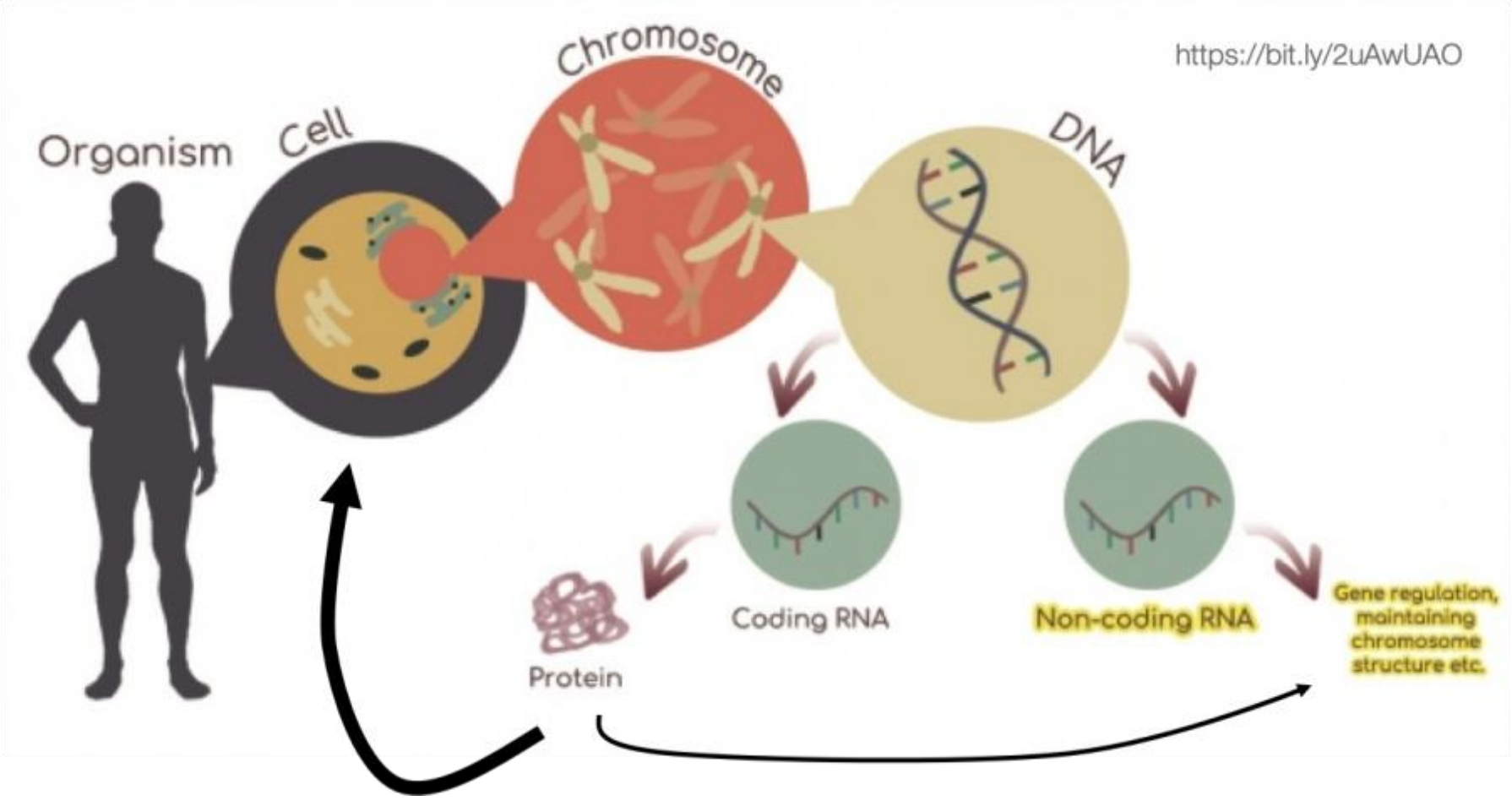
February 18, 2020

[1] https://physiology.med.cornell.edu/faculty/skrabanek/lab/angsd/schedule_2020/

1. Why study RNA?
2. Different types of RNA – different library preps
3. Gene expression : Microarray vs RNA-seq

5	**1. An introduction to transcriptomes/gene expression 2. Comparison of Expression Array vs RNA-seq**	06/05/2023	Loi, Thien, Thong
6	**Design a Bulk RNA-seq experiment**	06/12/2023	Loi, Phu
7	**Upstream analysis of Bulk RNA-seq with: 1) Mapping to reference genome 2) De novo assembly**	06/19/2023	Duy, Minh
8	**Downstream analysis of Bulk RNA-seq**	06/26/2023	Minh, Duy
9	**Use cases of poly-A Bulk RNA-seq**	07/03/2023	Hoang, Thien
10	**Use cases of microRNA Bulk RNA-seq**	07/10/2023	Lan, Thanh
11	**Use cases of poly-A bulk RNA-seq for variant calling (SNV), Copy number variant (CNV) and gene fusion**	07/17/2023	Bac, Quyen and Xuan

Why study RNA?



DNA is just the blueprint, it is not an effector molecule.

DNA is just the blueprint, it is not an effector molecule

GENOMICS

- DNA sequence of an organism
- genetic basis of phenotypic differences
- sites of DNA-protein or DNA-RNA interactions
- sites of open vs. closed chromatin

TRANSCRIPTOMICS

- = characterization of gene products
- identification of specific RNAs
- quantification of RNAs
- RNA-protein interactions
- RNA structure

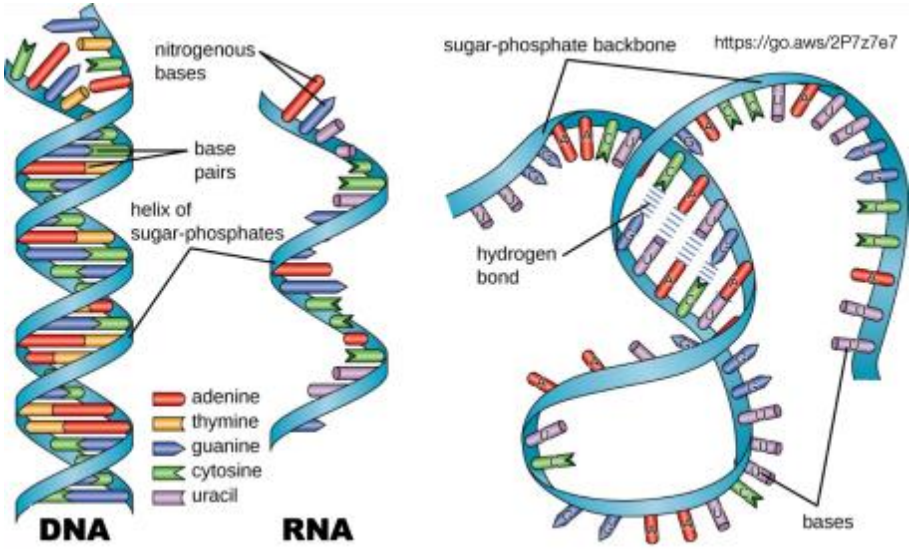
In order to understand the functional consequences (capacity) of a DNA sequence, we need to study its products, i.e. RNA and proteins.

Different types of RNA – different library preps

DNA and RNA have different properties

DNA

- usually double-stranded
- very stable
- mutations are heritable
- same amount in (almost) all cells
- same sequence in every cell of an organism

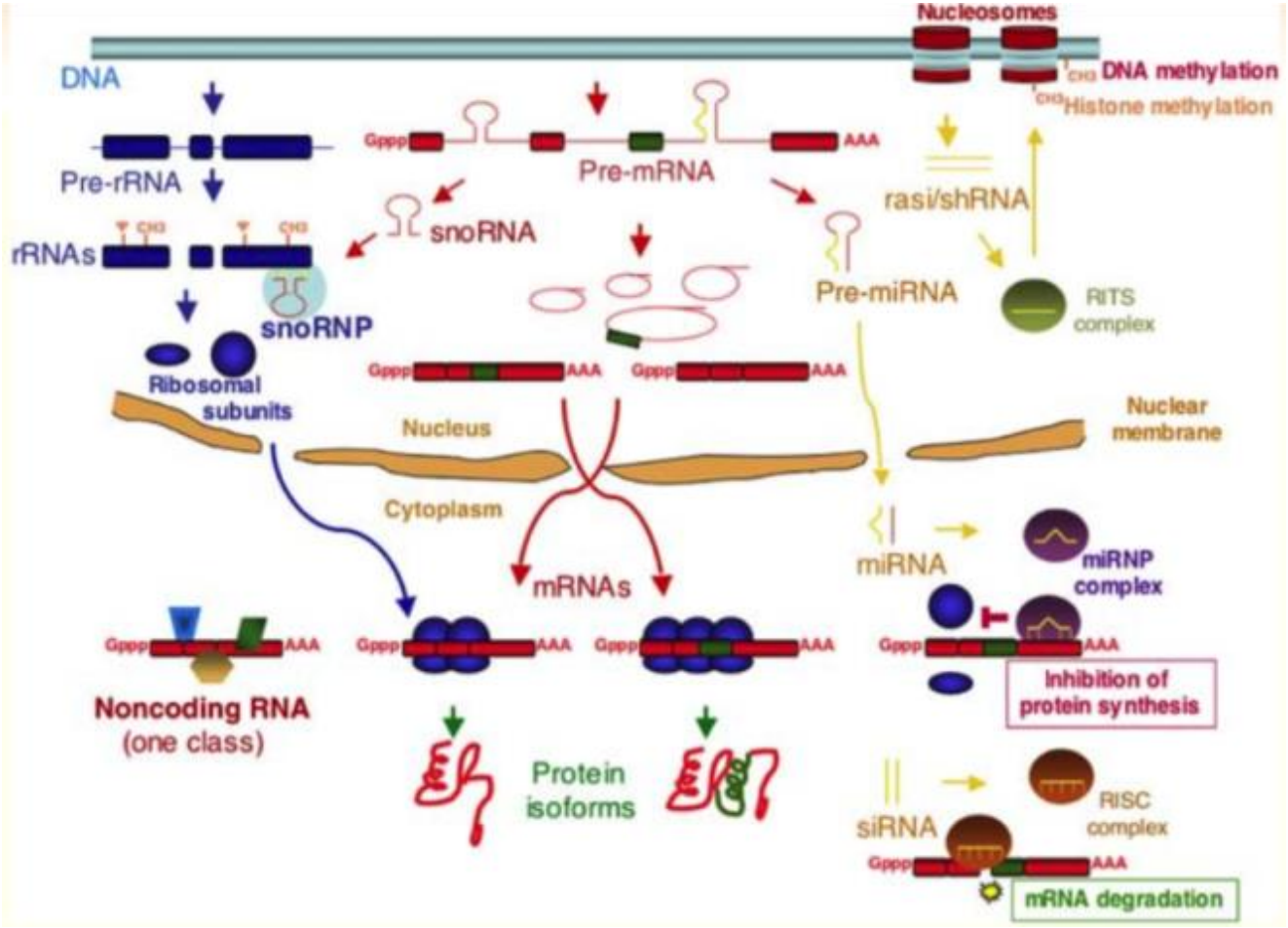


RNA

- generally single-stranded, but with the capacity for complementary base pairing ⇒ ability to form myriad different shapes
- usually fairly short-lived (minutes to hours)
- easily degraded/damaged without protection
- mutations are not passed on
- individual transcript *amounts* differ greatly depending on the gene, the cell type, the developmental status, the environment etc.
- transcript *sizes* range from 10-20bp to several kb

Different types of RNA

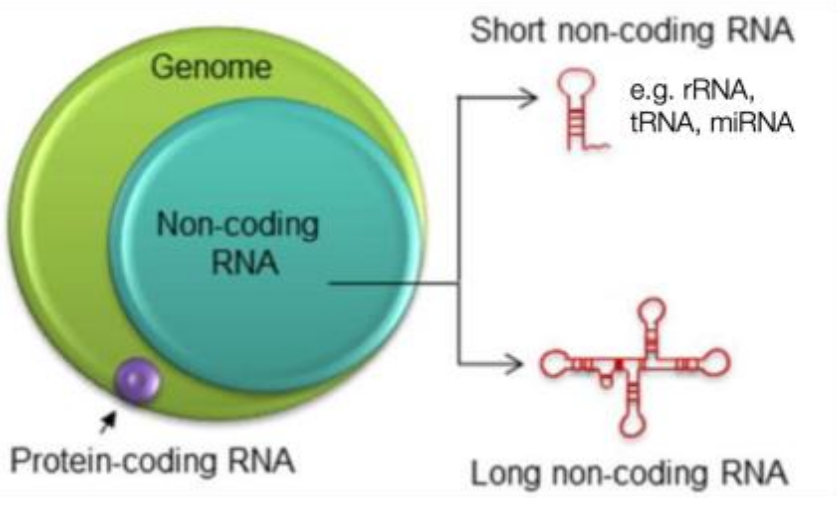
There are numerous different types of ***functional*** RNA molecules *in addition* to messenger RNA, which does *not* carry out a function of its own except transporting the DNA code (genetic information) into the cytoplasm where it can be translated into proteins.



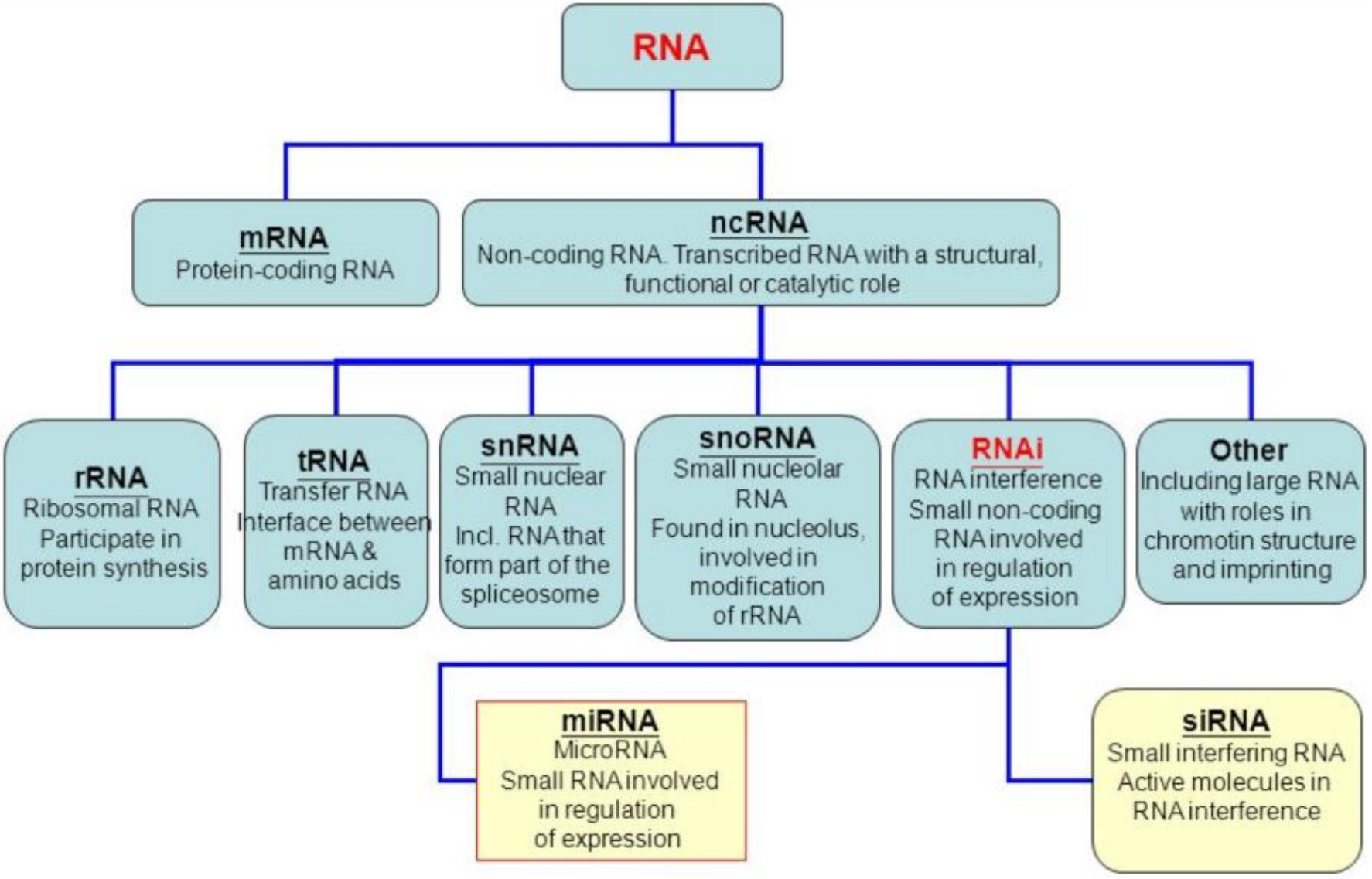
Different types of RNA

- Ca. 75% of the human genome can be transcribed (= copied into RNA) but <3% of the genome is subsequently translated into proteins genes can therefore be
  - coding (⇒ final product: protein) or
  - non-coding (⇒ final product: RNA)
- Non-coding RNAs cover a wide range of functions including protein assembly (⇒ ribosomal RNA, transfer RNA) and gene expression regulation

See Wilkes et al. [2017], Bartoszewski and Sikorski [2018] and Dai et al. [2020] for an introduction into the diverse RNA families and their functions.



Different types of RNA (there are more!)



Typical applications of RNA-seq

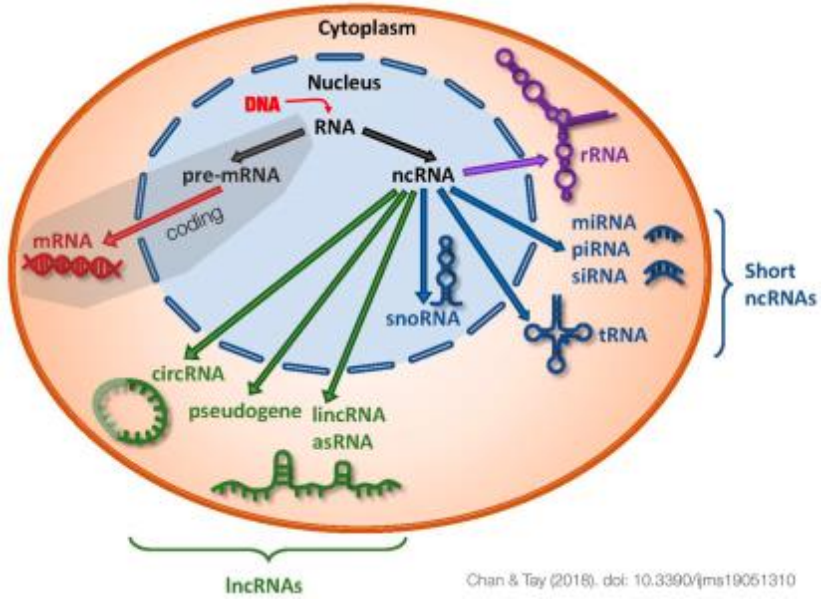
- **Identification** of transcripts – *which portions of the genome are expressed?*
  - identification of splice variants
  - transcriptome assembly
  - detection of gene fusion events
- **Quantification** of transcripts
  - comparison of different cell types/conditions/diseases and their effect on individual mRNA quantities
  - allele-specific expression

Illumina technology is best suited for the **quantification of known** transcripts; its short reads are not a good match for the identification of novel transcripts in very complex transcriptomes such as the ones found in mammals.

Sequencing prep protocol depends on the RNA properties

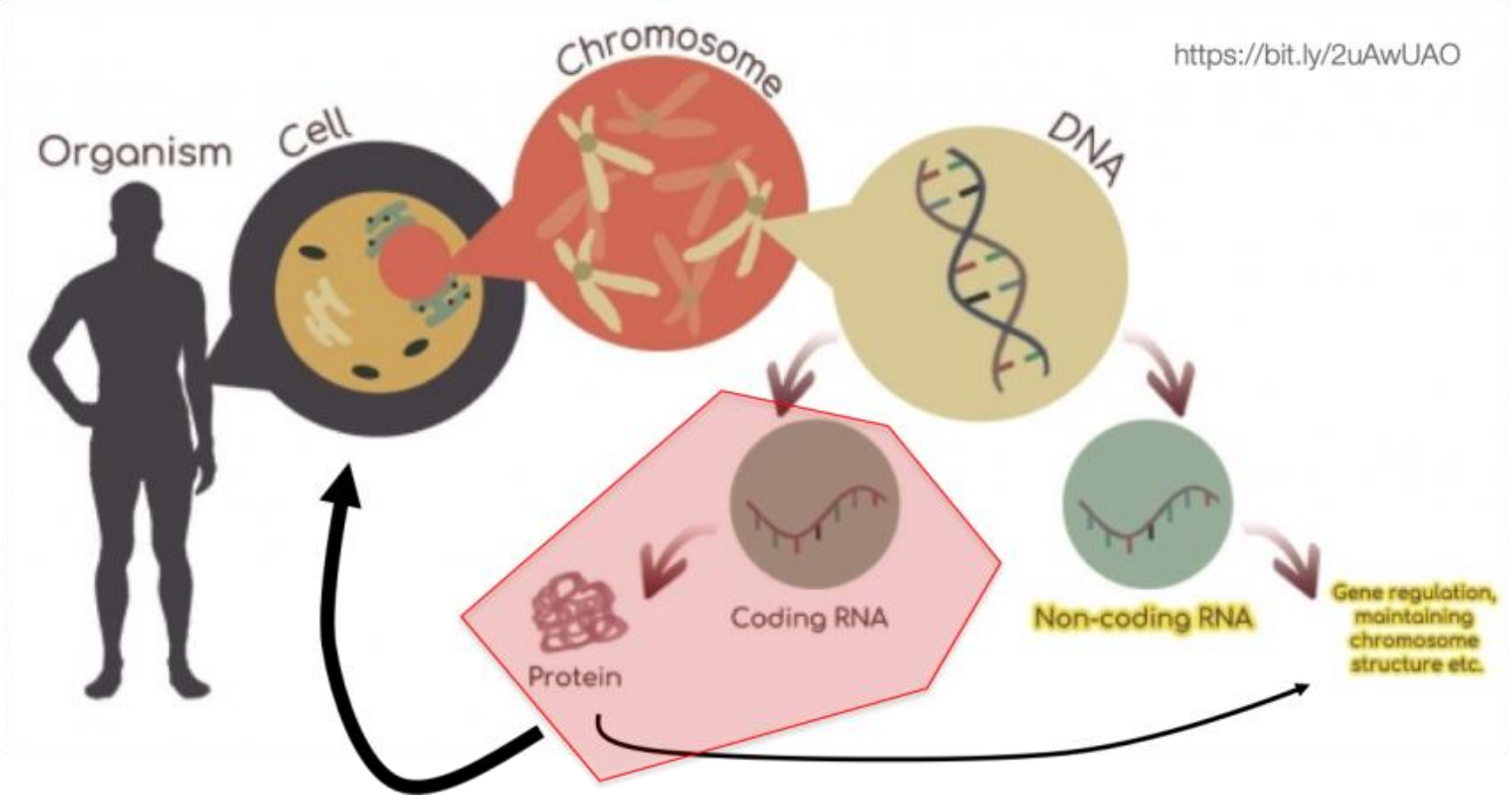
It is not a one-size-fits-all situation!

- abundance and stability
  - rRNA: 90-95% (!)
  - tRNA: 3-5%
  - mRNA: 2%
  - all other non-coding RNAs: *well* below 1%
- cellular location
  - most are in the cytoplasm
- size
  - miRNAs: 18-23bp
  - mRNA: several 100 to 1000 bp
- specific sequences/modifications
  - poly(A) tails of mRNA
  - 2D structure
  - antisense transcripts

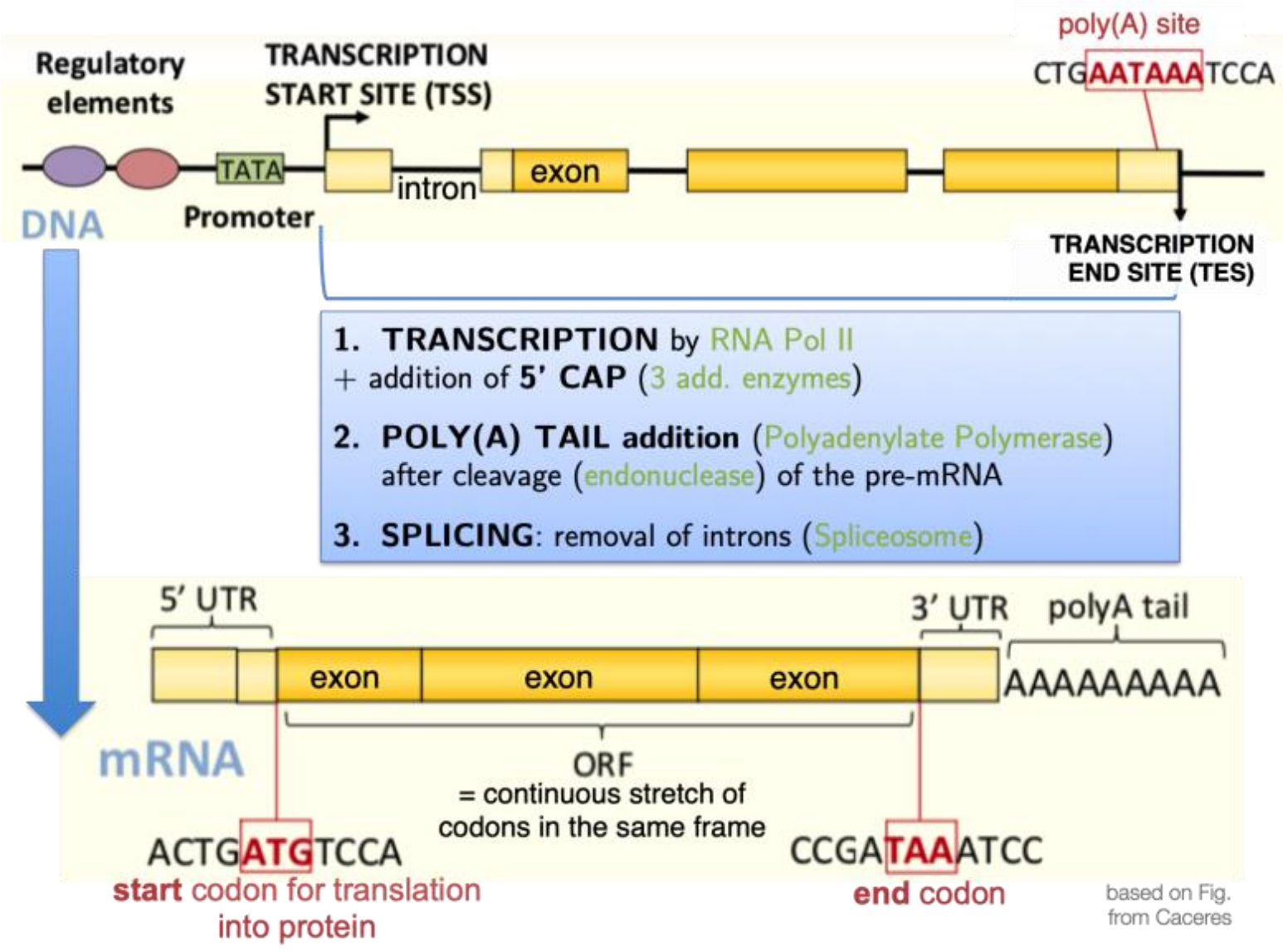


Focus today: messenger RNA

mRNA amounts are used as a proxy for the amounts of their corresponding proteins within a given tissue.



Focus today: messenger RNA



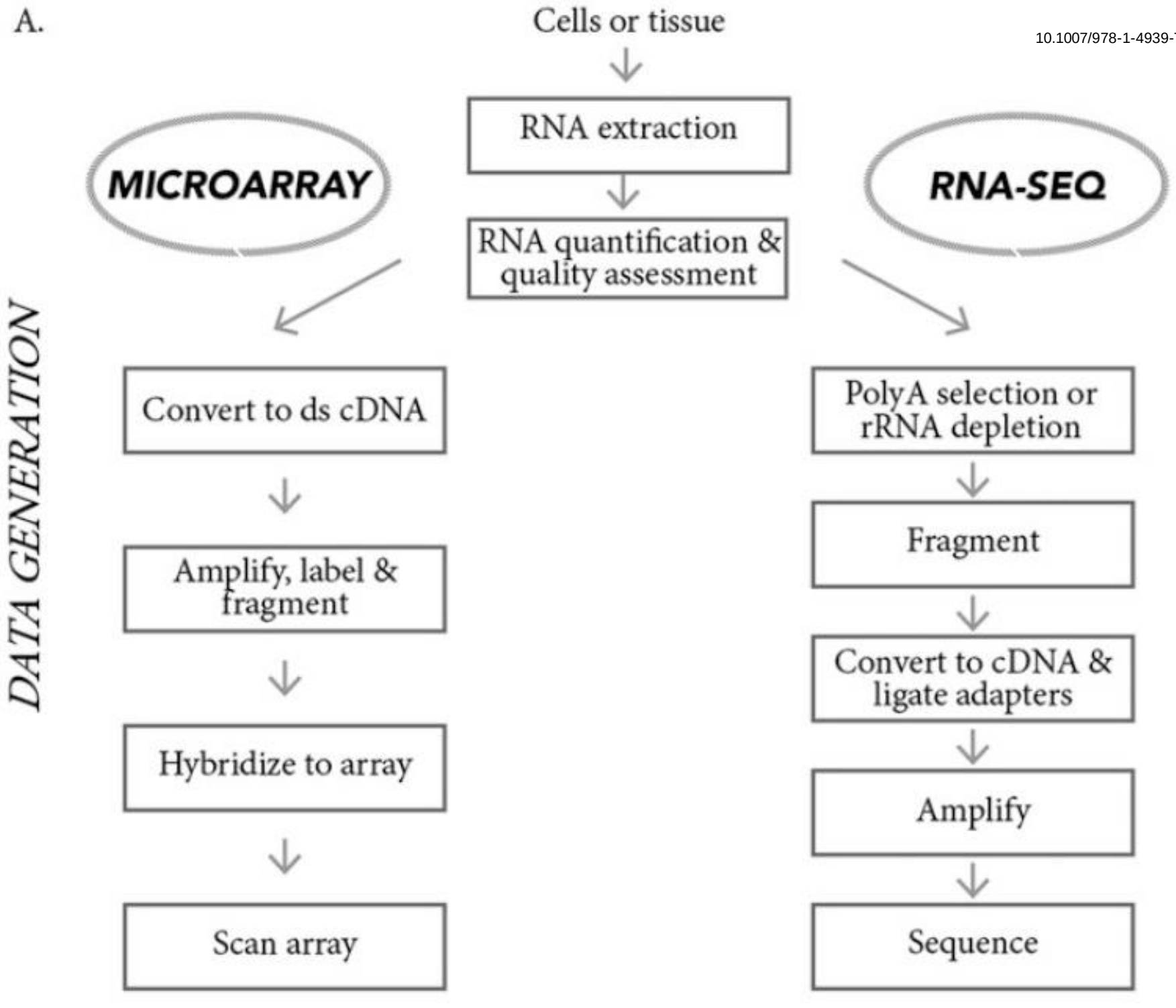
How to profile gene expression?

How to profile gene expression?

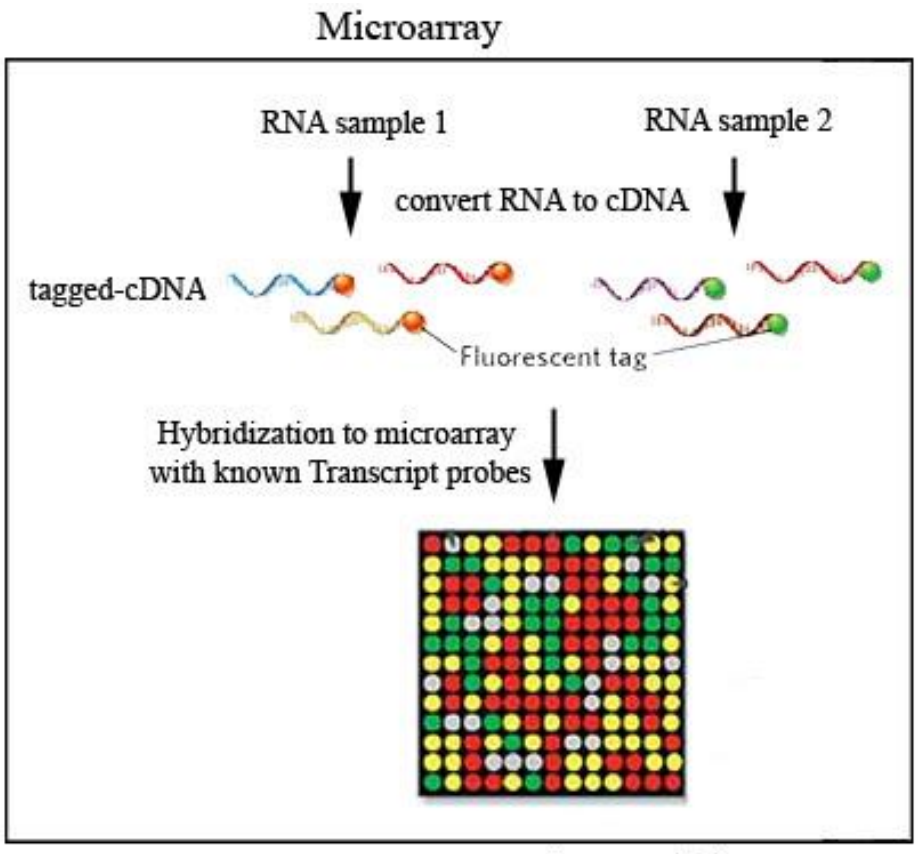
1. Quantitative Reverse Transcription and Polymerase Chain Reaction (qRT/PCR)
2. Northern Blotting
3. Fluorescence *In situ* Hybridization (FISH)
4. Microarrays
5. RNA-seq

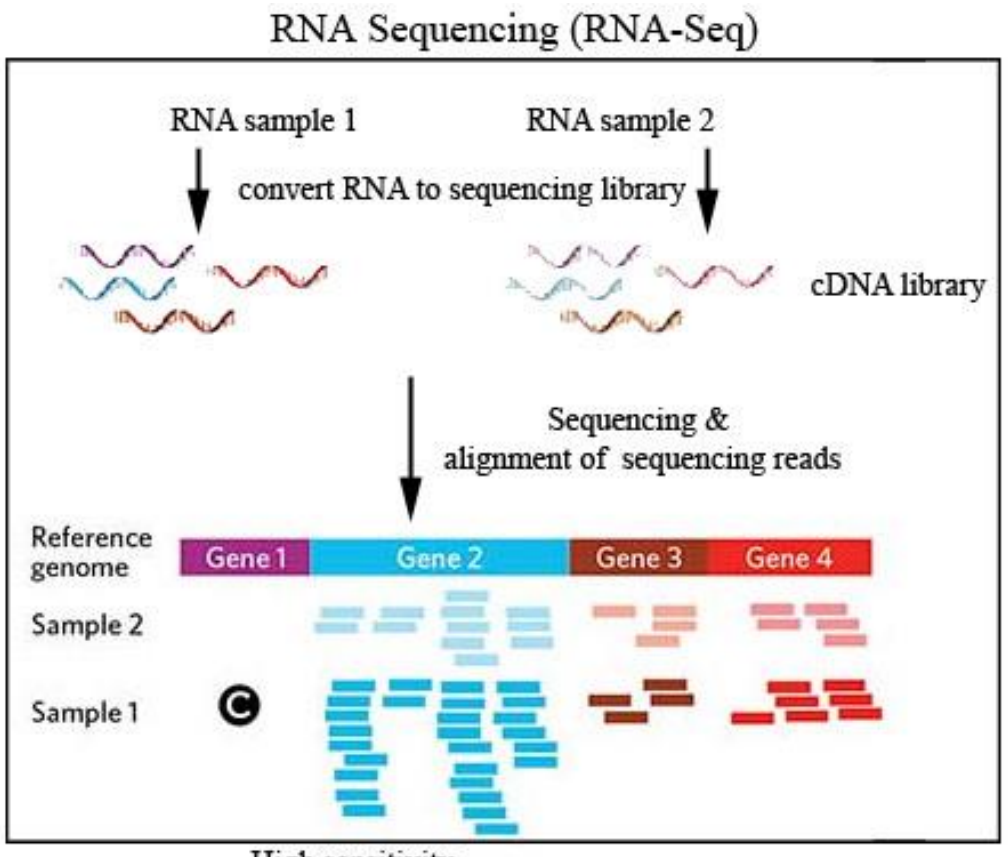
10.1007/978-1-4939-7834-2_2

A.



Microarray vs RNA-seq





relative intensity
=
expression levels

Low sensitivity
Low dynamic range
known transcript only
No alternative splicing information
lower cost

High sensitivity
High dynamic range
Novel transcripts sequences identified
structural variation & alternative splicing revealed
unlimited sample comparisons

Sequencing Reads
=
expression levels

https://www.otogenetics.com/rna-sequencing-vs-microarray/

10.1007/978-1-4939-7834-2_2

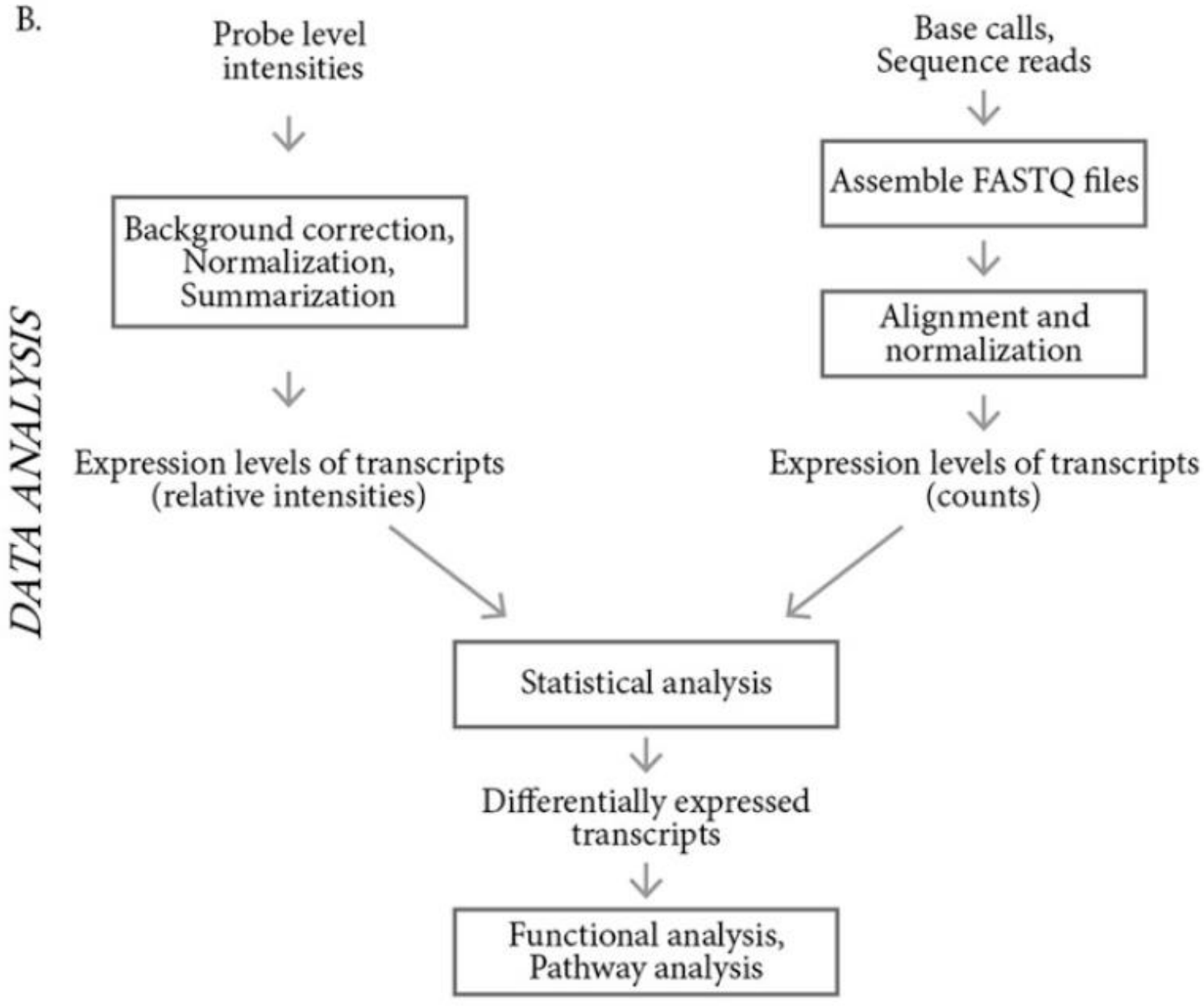


Fig. 1 Workflow of typical gene expression profiling with expression array or RNA-Seq

Table 1
Comparisons of RNA-Seq and microarray in the literature—technology and design details

Reference	Array type[a]	RNA-Seq platform	Library prep	Sequencing details[b]	Sample	Experimental design
Marioni et al. [41]	Affymetrix HG-U133 Plus 2	Illumina Genome Analyzer	polyA+ selection	32 bp, ~8–15mil	One human male sample from liver and kidney	Seven technical replicates, two cDNA conc., two sequencing runs, three arrays
Bottomly et al. [43]	Affymetrix MOE 2.0, Illumina MouseRef-8	Illumina GAIIx	polyA+ selection, Illumina mRNA-Seq Sample Preparation kit	300 bp, 21 samples in 21 lanes	Adult B6 and D2 strain mice; variation in prep dates, sex of samples among platforms	10+ strain replicates, all male for RNA-Seq and Illumina array, equal male/female Affymetrix array
Raghavachari et al. [42]	Affymetrix Human Exon 1.0 ST	Illumina GAIIx	polyA+ selection (Illumina recommendations)	Single lane, 36 cycles, sequencing depth ~10mil	Human Whole Blood (PAXGene), GLOBINclear depletion	Six patients sickle cell disease, four controls
Zwemer et al. [44]	Affymetrix HG-U133 Plus 2	Illumina HiSeq 2000	NuGEN Ovation RNA-Seq V2 (Husdson Alpha Institute)	Paired end, 50 bp, ~17–100 mil pairs; low alignment stats (5–35% for all but 1)	Human amniotic fluid	Three male, two female fetuses

10.1007/978-1-4939-7834-2_2

SEQC Consortium [45]	Affymetrix HG-U133 Plus 2, Affymetrix HuGene2.0, Affymetrix PrimeView, Illumina Bead array	Illumina HiSeq 2000, Life Technologies SOLiD 5500, Roche 454 GS FLX	Varies with site	Paired end; Illumina: 100 bp, ~110 mil pairs; Solid: 51/36 bp, ~50mil; Roche: ~1mil	Human, rat, multiple tissue types	Many samples in consortium
Zhao et al. [26]	Affymetrix HG-U133 Plus 2	Illumina HiSeq 2000 (BGI)	polyA+ selection	Paired end, 90 bp	Human CCR6+ CD4+ T cells	6 time points, +/− stimulation in duplicate
Yu et al. [27]	Agilent Hum Genome 4x44K, Illumina Hum HT12v4, Affymetrix Hum Gene 1.0 and HTA2.0	Illumina HiSeq 2500	SMARTer UltrLow RNA kit, Epicentre RiboZero	50 bp, ~36mil reads	One pool of human bone marrow RNAs, Universal Human Ref RNA (Agilent)	Two samples plus ratio-varied mix of the two samples (1–3 reps)
Nazarov et al. [46]	Affymetrix HTA 2.0	Illumina HiSeq 2000	Ribosomal RNA depletion	Paired end, 77 bp, 120–280 mil reads	Matched primary human lung tumor and adjacent tissue	Nine matched samples

[a]Total RNA
[b]Unstranded and single end unless otherwise stated

10.1007/978-1-4939-7834-2_2

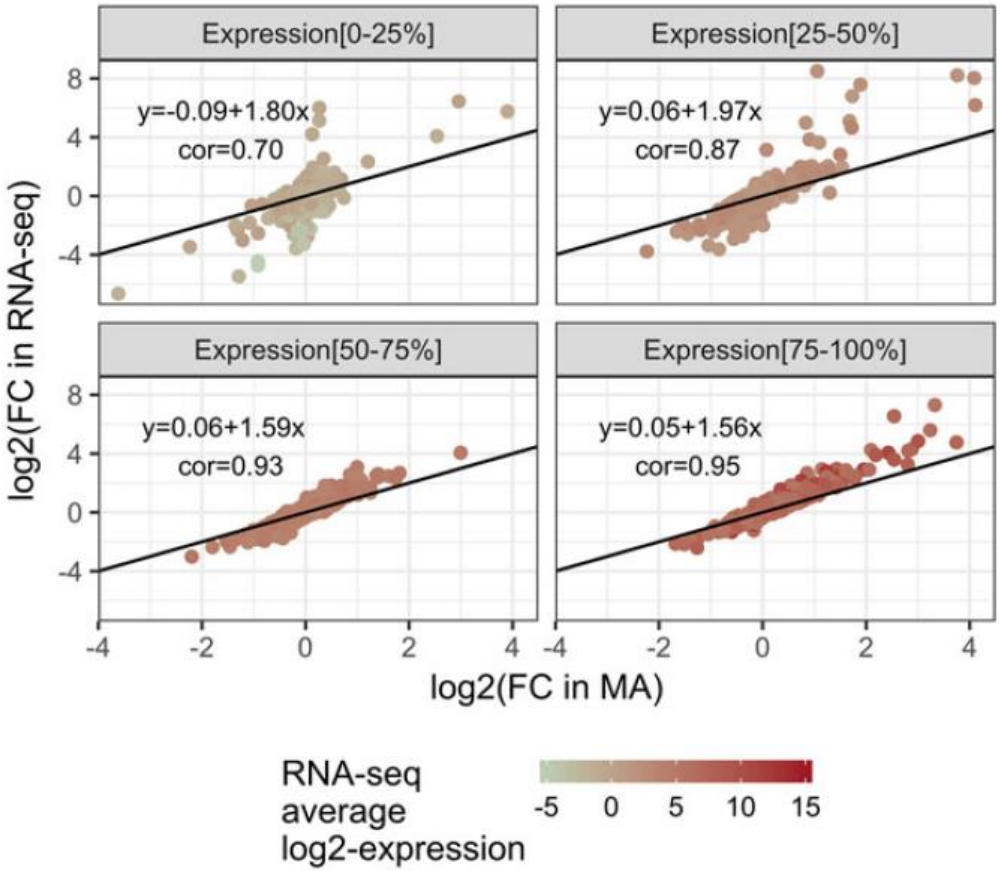


Fig. 2 Log2-Fold change (Condition 2 vs 1) from microarray and RNA-Seq data. Each point represents a gene that is represented on both platforms (8520 genes total). Genes were divided into four panels based on quantiles of average RNA-Seq log2-expression. Darker shades represent larger average abundance. Pearson's correlation as well as the intercept and slope from the univariate regression model of log2-FC in RNA-Seq regressed on log2-FC in array are displayed. Black lines represent the diagonal slope of 1, which would occur when fold changes in the platforms are identical. A slope larger than 1 from the univariate regression denote a trend of larger fold change magnitudes measured in RNA-Seq

10.1007/978-1-4939-7834-2_2

Table 2

Number of genes and gene ontology terms detected by each platform separately as well as with both platforms

Number	Microarray	RNA-Seq	Platform overlap
Genes[a] quantified: Protein-coding (All RNA) genes	26,596 (65,956)	22,001 (45,706)	22,001 (37,921)
Genes[a] after low expression filtering[b]	9262 (13230)	14,554 (16696)	*8254 (8520)*
DE genes[a] FDR 5%; *As subset of overlap*[c]	671 (962); *601 (604)*	1918 (2095); *1241 (1268)*	*550 (553)*
DE genes[a] FDR 5% and log2FC > 0.5; *As subset of overlap*[c]	484 (745); *431 (434)*	1595 (1770); *956 (983)*	*411 (414)*
GO terms identified	11,365	12,964	11,298
GO terms $p < 0.05$	889	1682	662

[a]Number of protein coding genes (number of total genes)

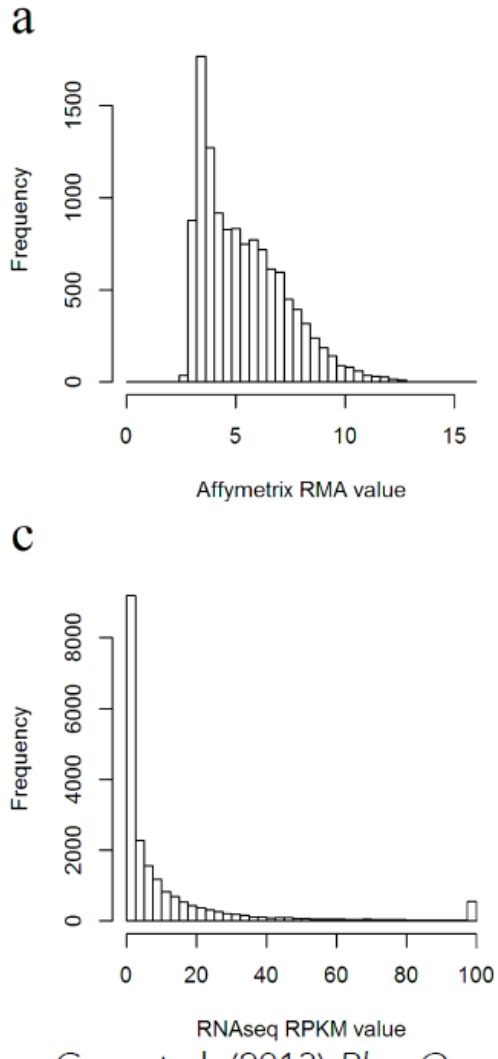
[b]Filtering: For microarray data, the 95th percentile of the normalized intensity distribution of antigenomic background control transcript clusters was calculated as a background cutoff, and transcript clusters were removed from further analysis if fewer than two samples exhibited normalized expression levels above this cutoff. RNA-Seq genes were removed from further analysis if fewer than two samples exhibited CPM greater than 0.2. This constitutes the set of genes tested for differential expression

[c]Numbers in italics are a subset of the 8520 genes found on both platforms. For each platform, these correspond to significant detection in one platform regardless of whether it was significant for the other platform

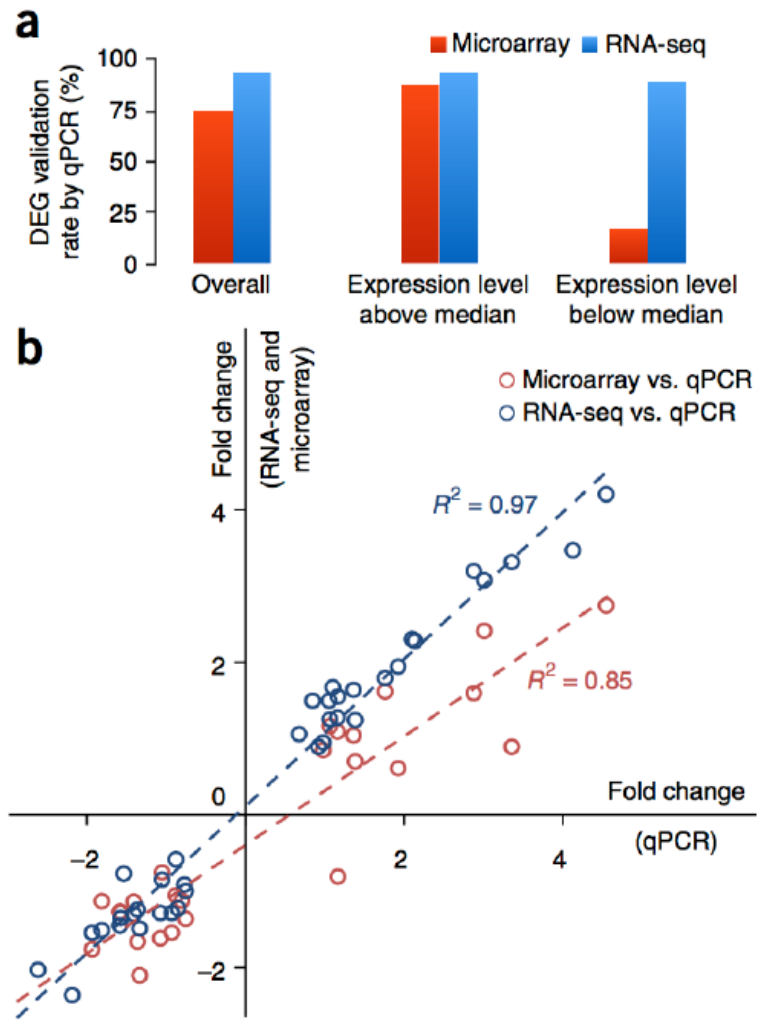
[d]DE denotes differentially expressed genes, GO denotes Genome Ontology terms belonging to the Biological Process ontology

10.1007/978-1-4939-7834-2_2

Microarray vs RNA-seq



Guo et al. (2013) *Plos One*



Wang et al (2014) *Nature Biotech.*

How to choose gene expression assay?

	Value	*Pros*	*Cons*
Arrays	- Accurate and reproducible expression data - Established, reliable platform	- Standardized workflows and user-friendly software available for data processing and analysis - Fast turnaround times, esp. for small studies	- Limited dynamic range for fold change measurements - Require known sequence info for array fabrication - Limited detection of very low abundance transcripts
RNA-Seq	- Accurate and reproducible expression data - Transcript sequence information included	- Extended dynamic range - Ability to detect rare transcripts with deep sequencing - Ability to identify novel transcripts and sequence variation	- Complex data analysis w/ no 'gold-standard' pipeline - Requires mRNA selection or abundant transcript removal to avoid deep sequencing costs

Fig. 3 Performance and use considerations for gene expression profiling with microarrays versus RNA-Seq

10.1007/978-1-4939-7834-2_2

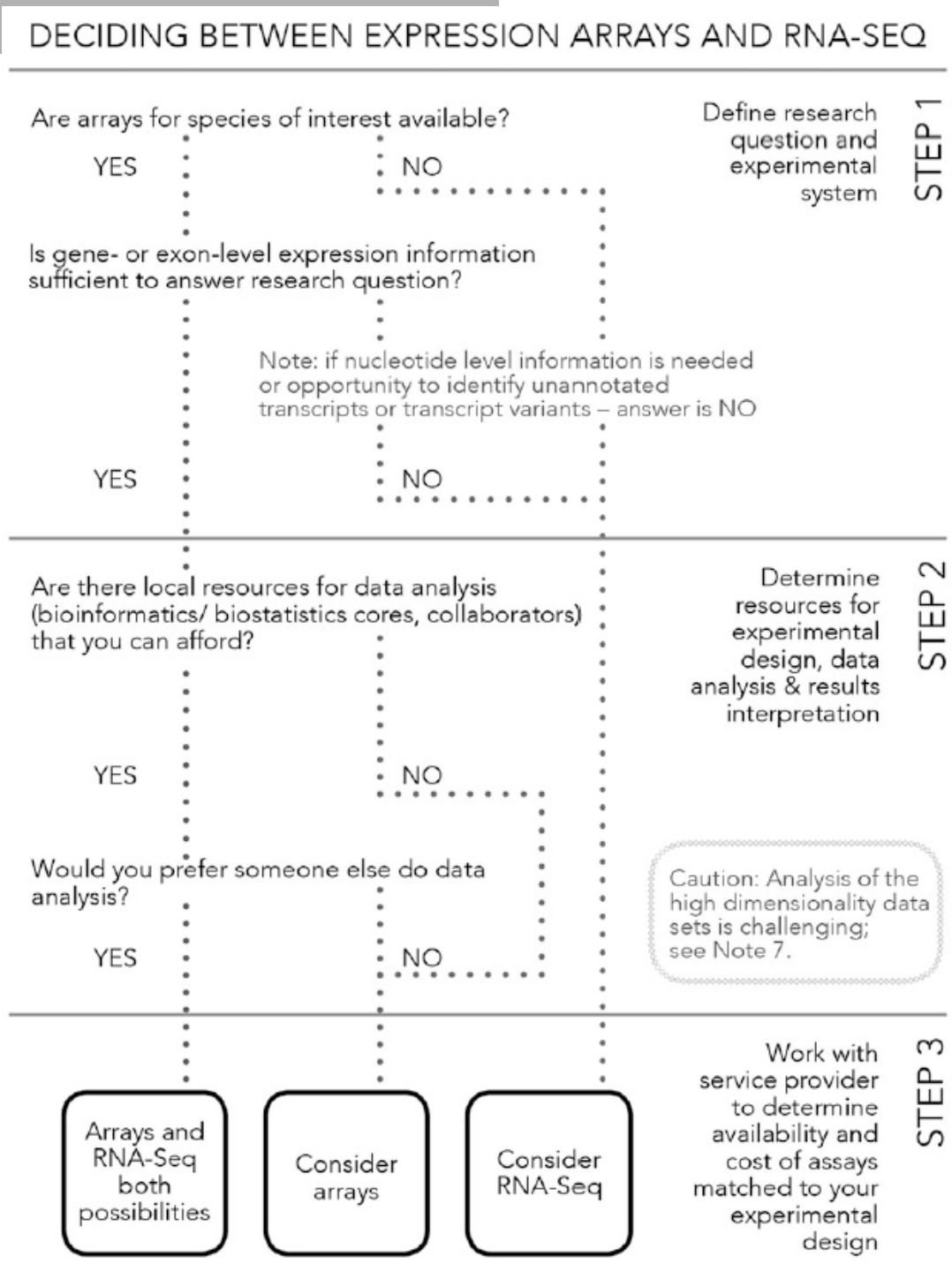


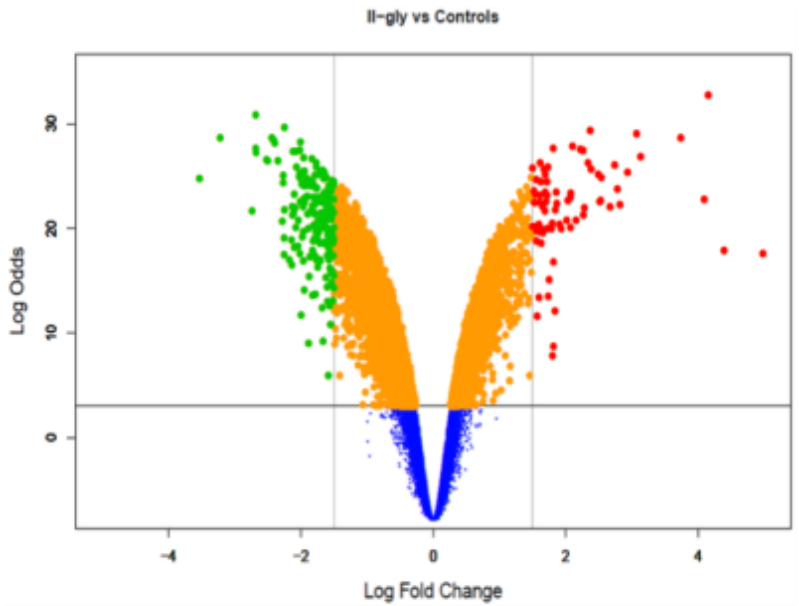
Fig. 4 Technology platform selection process for gene expression profiling

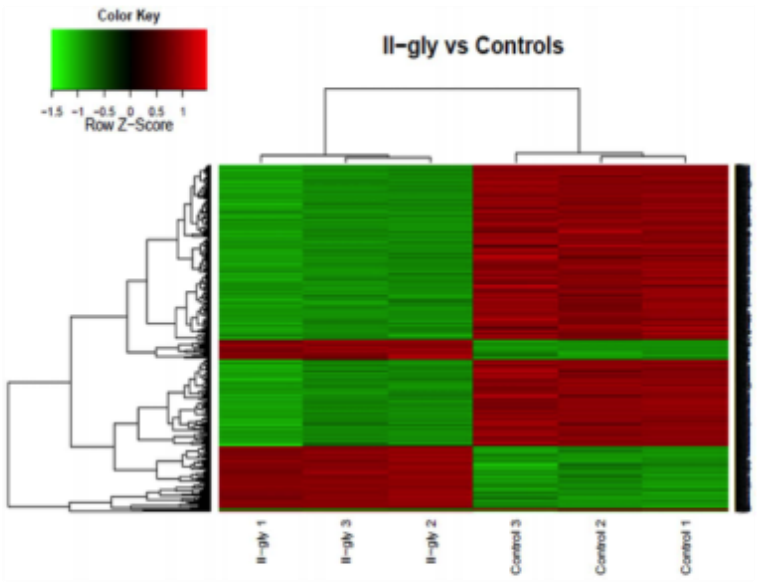
10.1007/978-1-4939-7834-2_2

Focus today: messenger RNA

Bulk RNA-seq of mRNA

- expression quantification of (mostly) mRNA transcripts
- extracted from populations of cells
- and tested for gene-specific differences between distinct conditions





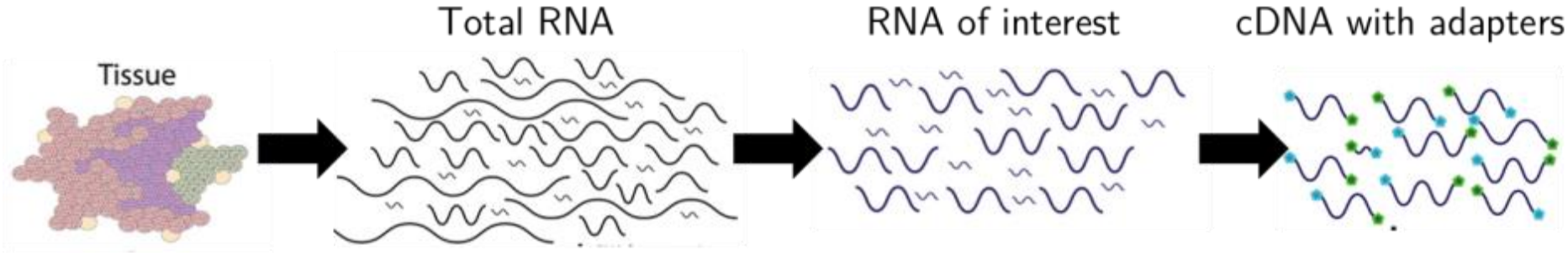
Valencia-Cruz et al. (2013). doi: 10.1371/journal.pone.0054664

Typical questions addressed with bulk RNA-seq

- Does a certain treatment induce gene expression changes? And if it does, which genes are most strongly affected?
- How does the gene expression profile of a cancer cell differ from a healthy cell?
- Which genes are turned on/off during the course of embryonic development?
- Which genes differ in mice that have been engineered to lack a certain gene? E.g., which genes – in addition to the one that's been "knocked-out" – may be depleted or overcompensating for the loss?
- Which genes are activated in response to an environmental stimulus, e.g. heat shock or alcohol poisoining?
- How does the gene expression profile change in the same tissue in an aging individual?
- . . .

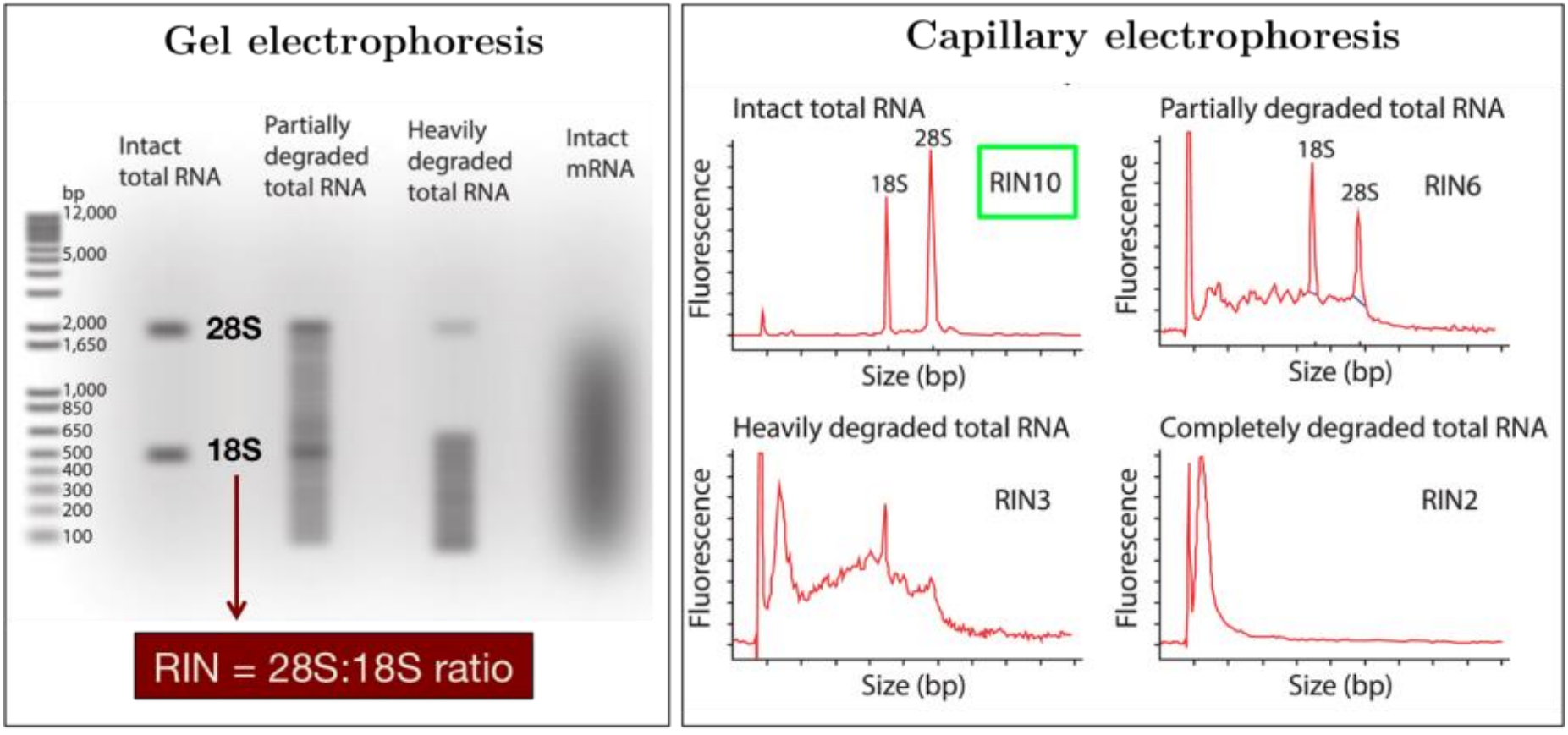
General steps of RNA-seq preparation

1. RNA **extraction**[2] (cell lysis, RNA purification)
2. **enrichment** of the RNA of interest
3. **fragmentation** (ca. 200 bp)
4. **cDNA** synthesis
5. library prep to obtain cDNA with **adapters** for sequencing



[2]Most standard extraction methods will lose RNA <100 bp!

QC of RNA extraction

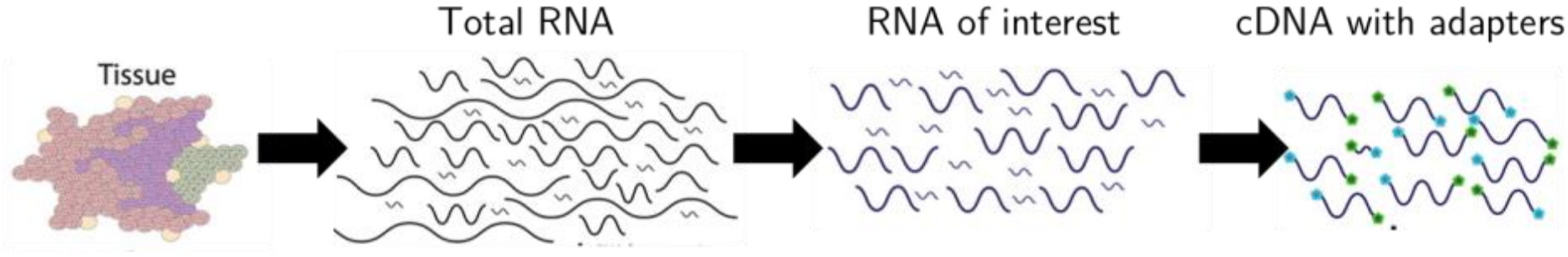


Griffith et al. (2015). doi: 10.1371/journal.pcbi.1004393

Avoid degraded RNA! Optimum: RNA Integraty Score (RIN) of 10.

General steps of RNA-seq preparation

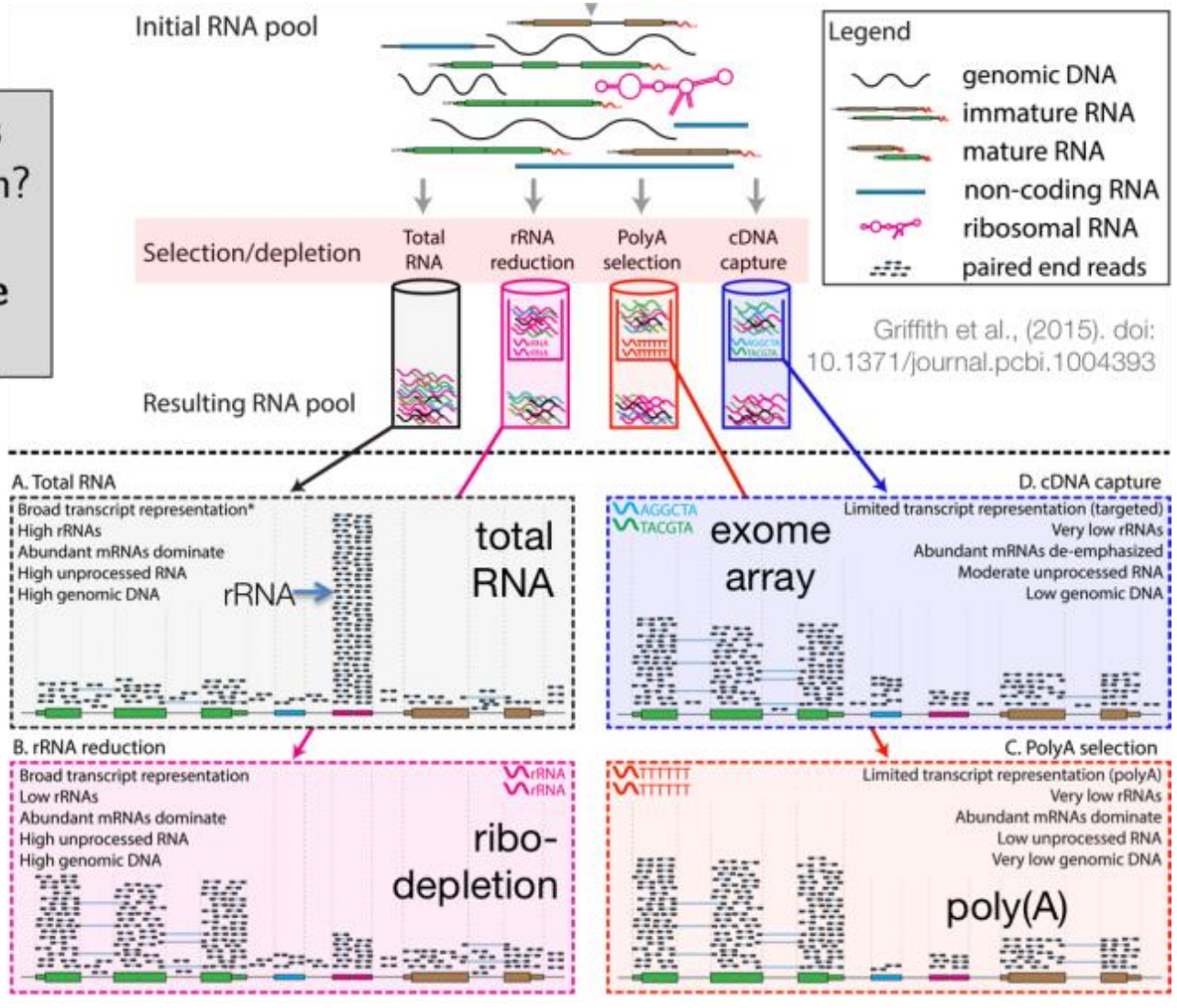
1. RNA extraction (cell lysis, RNA purification)
2. enrichment of the RNA of interest
   - mRNA: poly(A) enrichment vs. ribosomal-depletion
   - small RNAs: size-based enrichment
3. fragmentation (ca. 200 bp)
4. cDNA synthesis
5. library prep to obtain cDNA with adapters for sequencing



Every step has consequences – example: mRNA enrichment strategies

which **transcripts** are you interested in?

what type of **noise** can you tolerate?

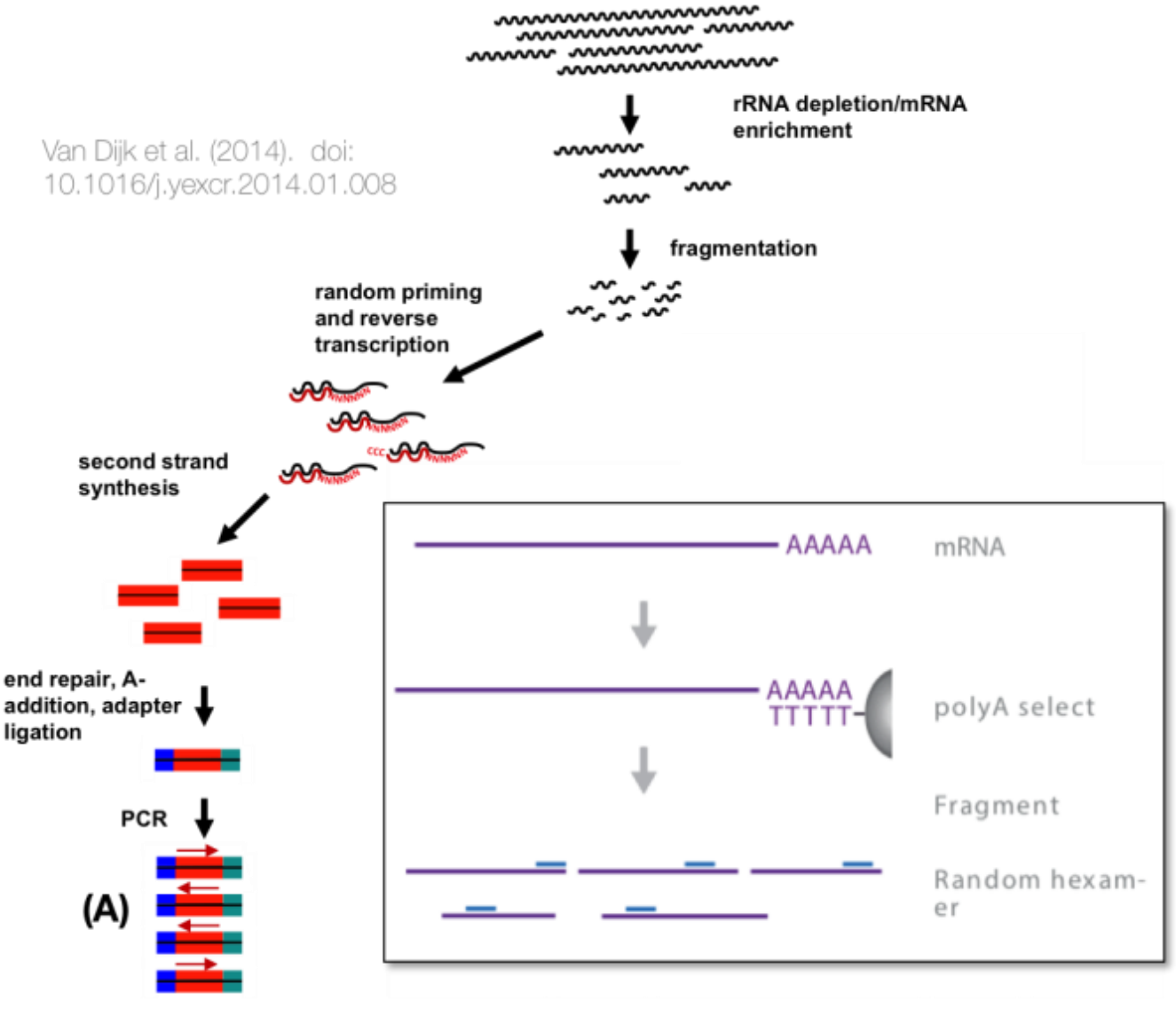


rRNA

protein coding (strongly expressed)

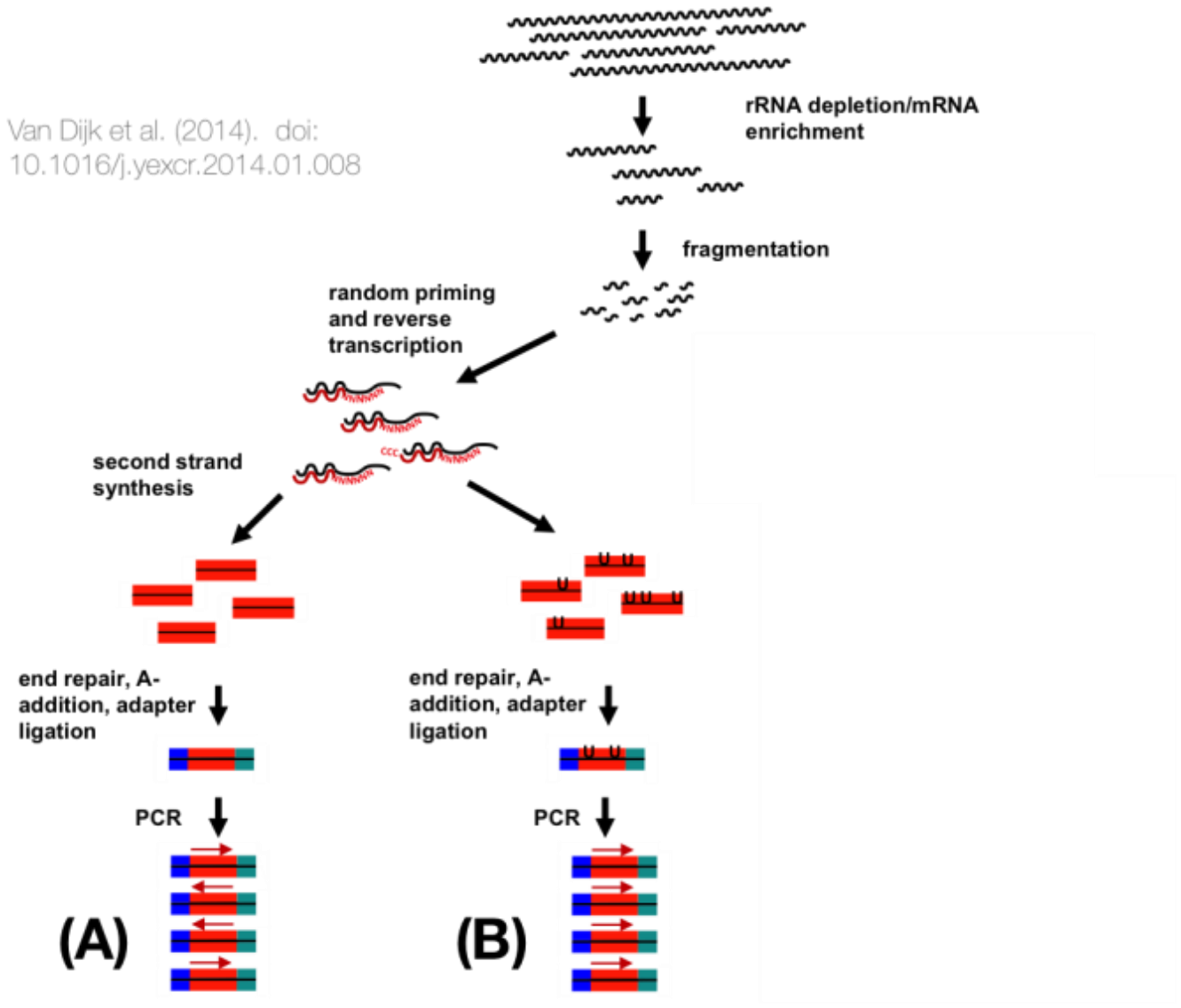
protein coding (lowly expressed)

The most common library preparation methods



- **(A)** classical unstranded mRNA library prep

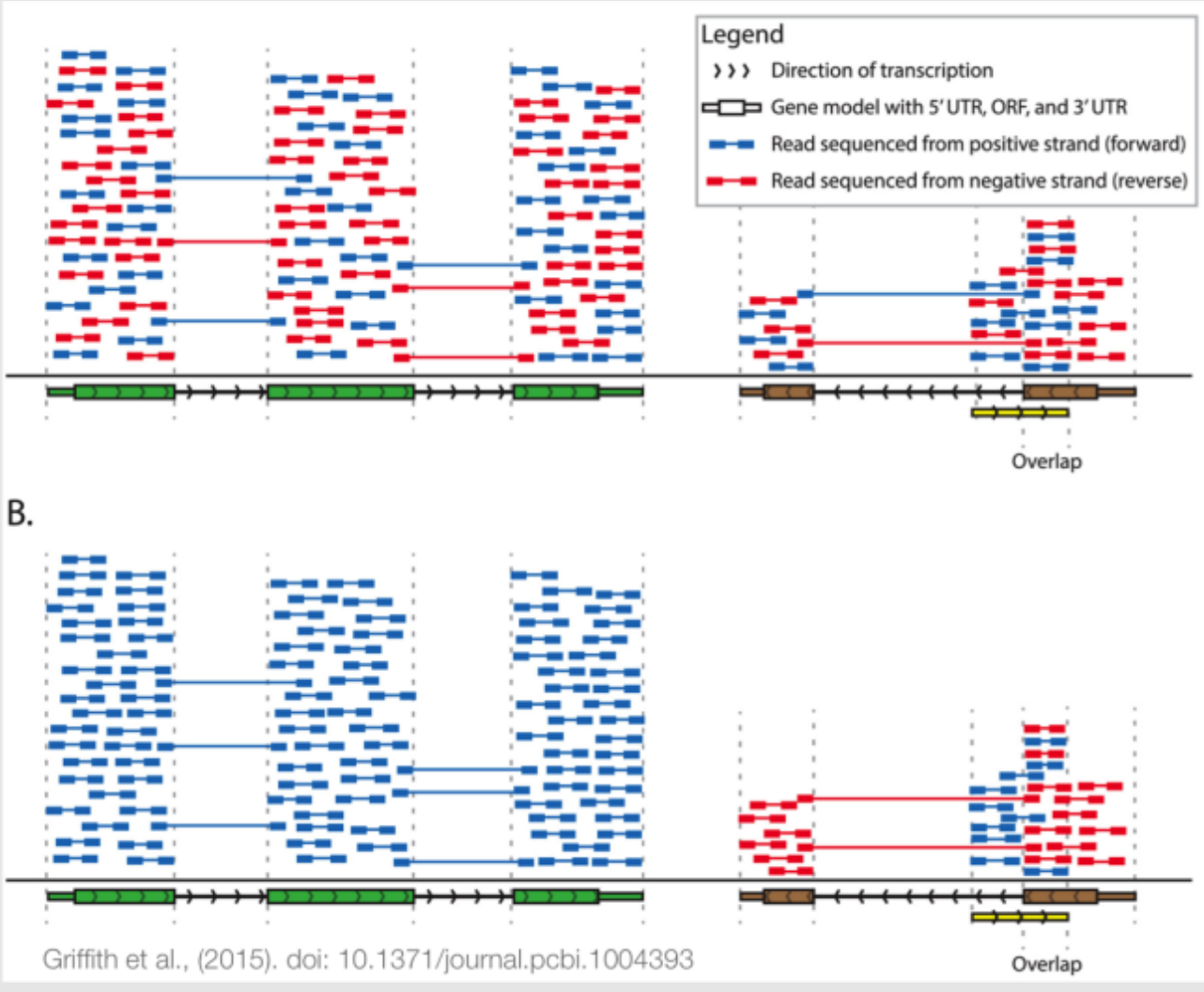
The most common library preparation methods



- (A) classical unstranded mRNA library prep
- (B) stranded mRNA (dUTP-based) (see Levin et al. [2010] and Zhao et al. [2015] for details)

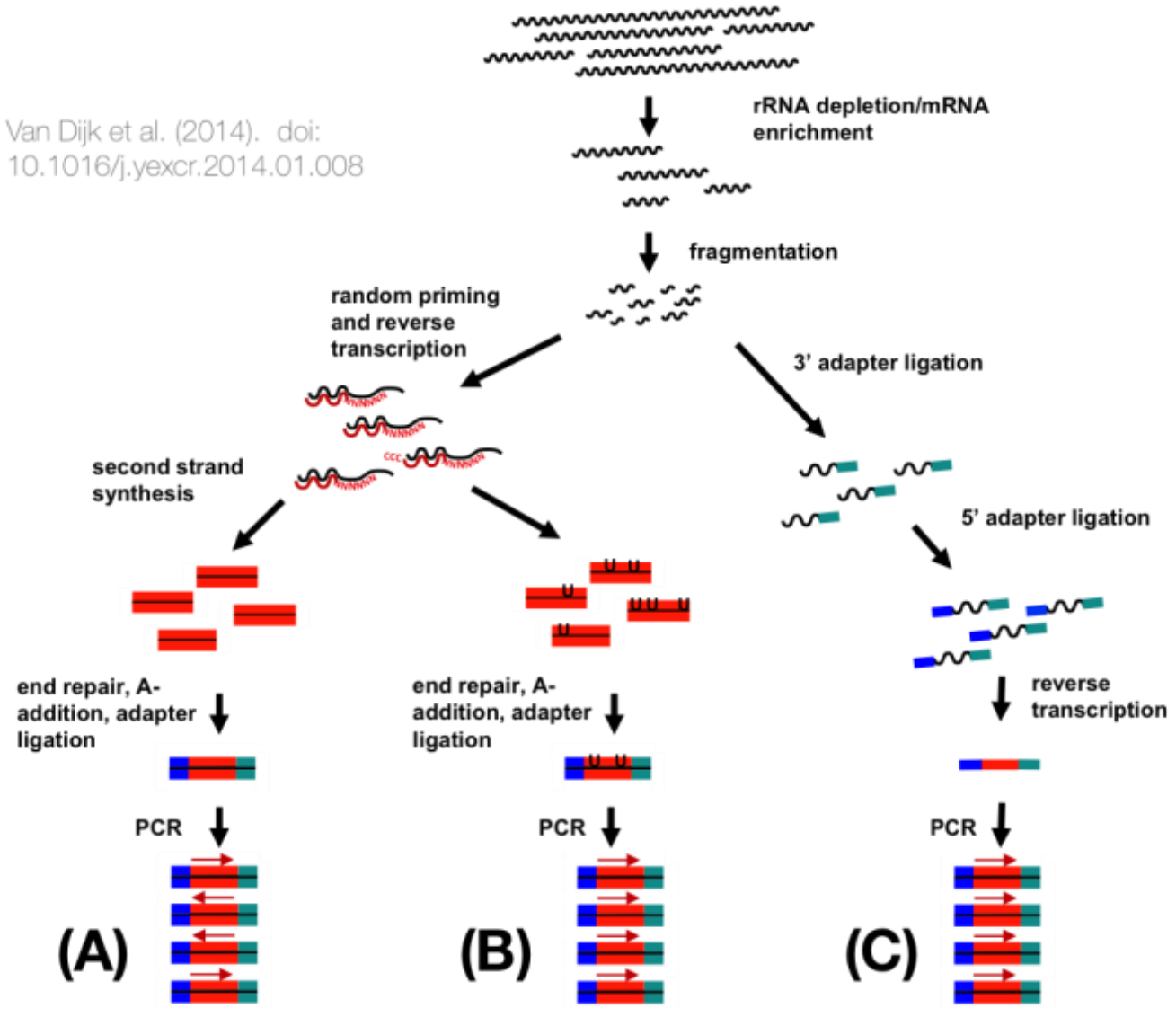
The most common library preparation methods

Unstranded vs. stranded



Griffith et al., (2015). doi: 10.1371/journal.pcbi.1004393

The most common library preparation methods



- (A) classical unstranded mRNA library prep
- (B) stranded mRNA (dUTP-based) (see Levin et al. [2010] and Zhao et al. [2015] for details)
- (C) small RNAs (miRNA, piRNA, tRNA, ... <100 bp) using 2 adapters – not optimal for differential expression analyses!

Every step has consequences

- Do not mix different strategies for samples that are to be compared to each other!
  - extraction, enrichment, library prep

There are many papers comparing different aspects of different RNA-seq approaches, e.g.

- *Library preparation methods for next-generation sequencing: Tone down the bias* [van Dijk et al., 2014]
- *Systematic comparison of small RNA library preparation protocols for next-generation sequencing* [Dard-Dascot et al., 2018]
- *A comprehensive assessment of RNA-seq protocols for degraded and low-quantity samples.* [Schuierer et al., 2017]
- many more – PubMed is your friend!

Make an informed decision!