Swiss Institute of Bioinformatics

Introduction to RNA-Seq: Overview

Wandrille Duchemin

elixir SWITZERLAND

General Information

Course page: https://sib-swiss.github.io/RNAseq-introduction-training/

- Slides, Data sets, Exercises, Solutions

Optional exam, 0.5 ECTS value

- Course from 09:00 to 17:00
- Lunch break 12:00 to 13:00
- 15min breaks around 10:30 and 15:00

Asking questions - Communication

- Raise your hand anytime ✋

- Done with an exercise?

Course Outline

Day 1

1. **Overview** of RNAseq
2. Getting started with the **cluster**
3. **Quality Control** of the raw data
4. Sequence **trimming**

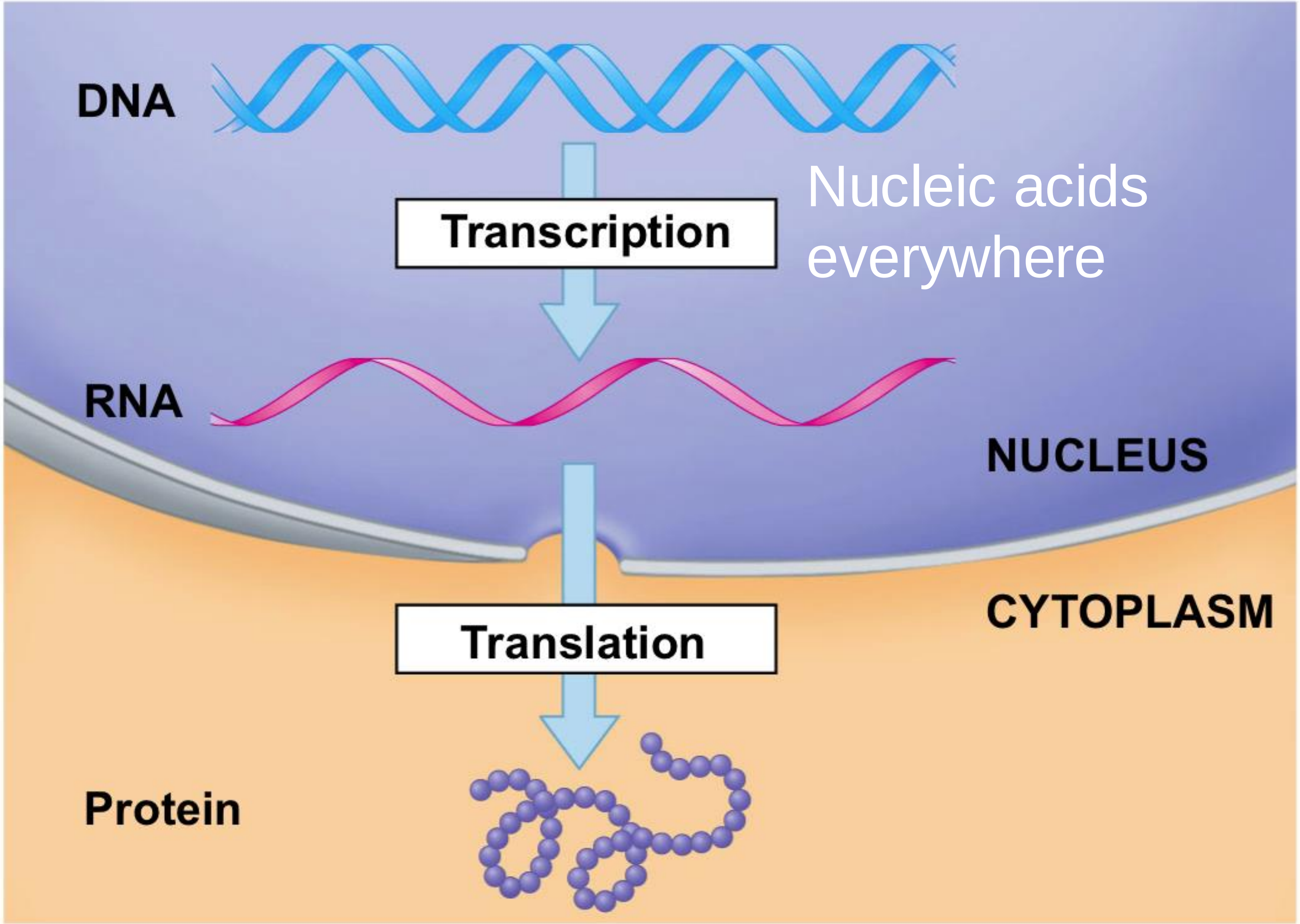
Day 2

1. Reas **mapping**
2. **Differential Expression** Inference
3. Enrichment Analysis

slides Outline

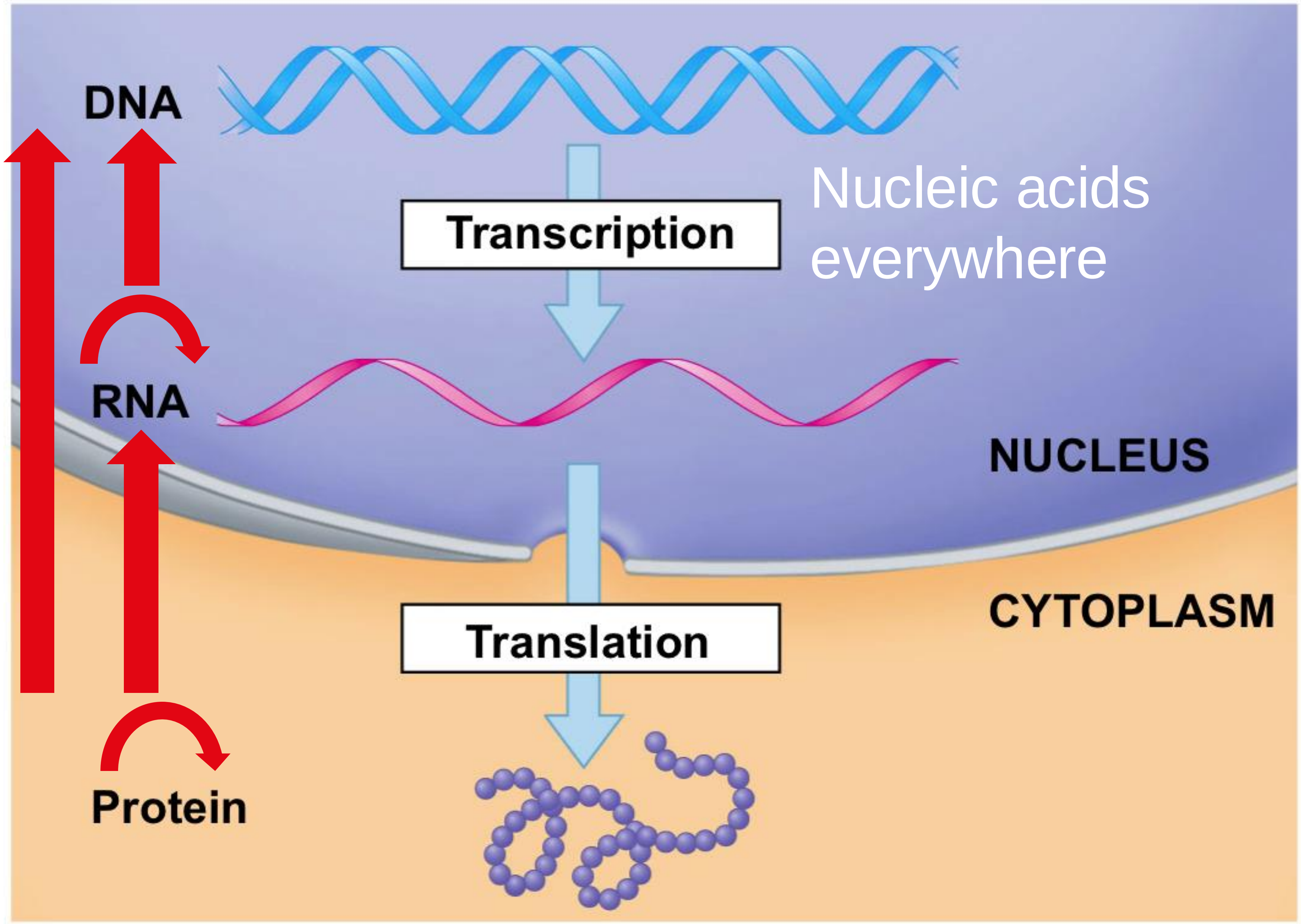
- RNA and molecular biology
- Main challenges for RNAseq
- Major Sequencing technologies
- Planning your sequencing : choices, number of samples, …
- Bioinformatics analysis overview

Introducing Ourselves



Nucleic acids everywhere

© 2012 Pearson Education, Inc.



Nucleic acids everywhere

alternative splicing adds a layer of complexity

~20'000 mammalian genes

>>100'000 (?) transcripts

>>1'000'000 (?) proteins

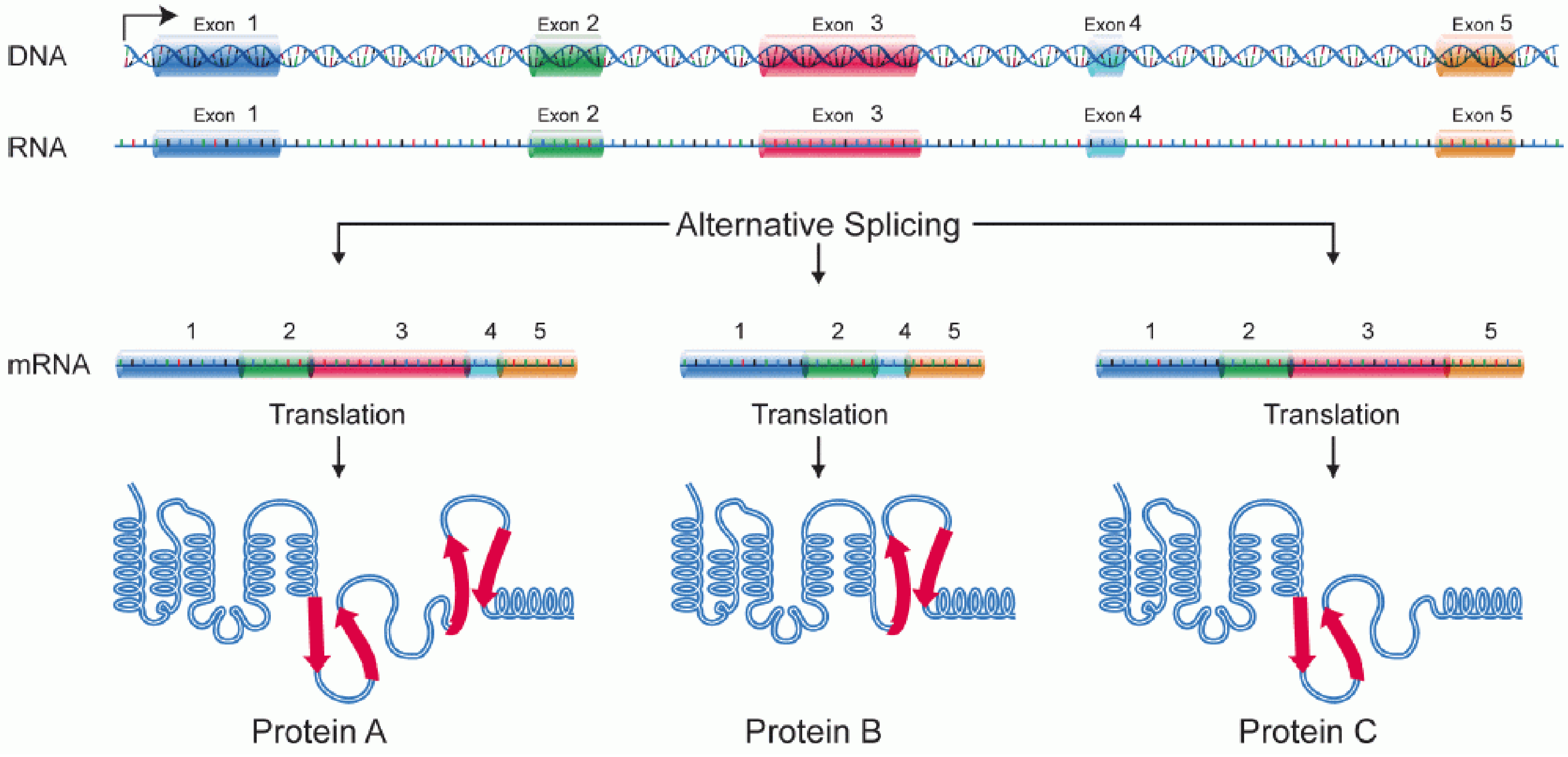
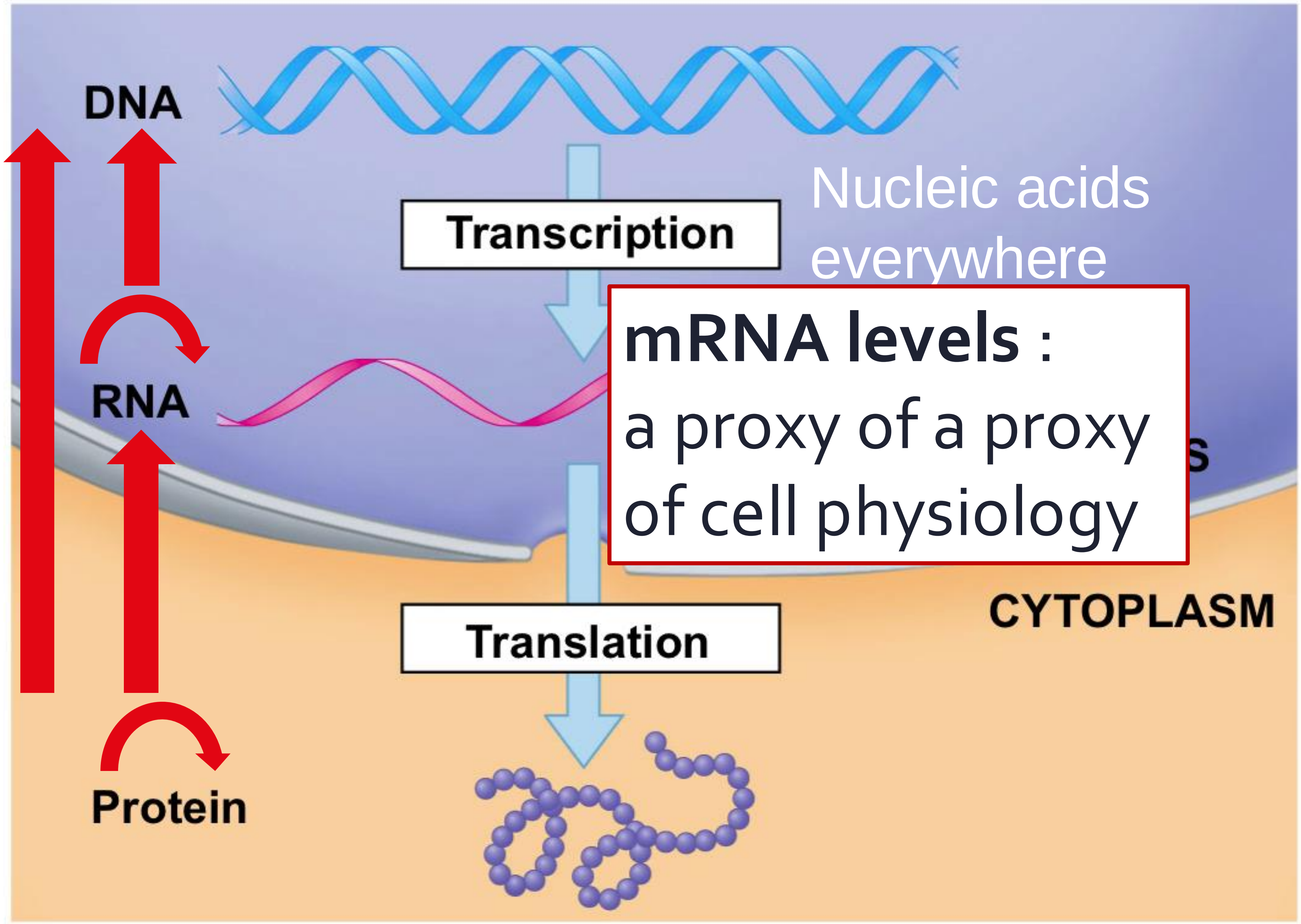


Image credit: National Human Genome Research Institute – public domain



© 2012 Pearson Education, Inc.

Nucleic acids everywhere

mRNA levels :
a proxy of a proxy
of cell physiology

What (and why) are we sequencing

Genomics

- Whole genome/exome sequencing (WGS/WES)
- Variant calling (SNPs, CNVs, structural variations)

Epigenomics

- Bisulphite sequencing : DNA methylation
- ATAC-Seq : chromatine opening
- ChIP-seq : TF binding sites

Transcriptomics

- Total RNA
- Poly-A tail selection : focus on mRNA
- Ribo depletion: mRNA + ncRNA
- 5'/3' RACE seq : isoform characterization for one gene
- scRNAseq
- Long read RNA sequencing
- ...

What (and why) are we sequencing

Genomics

- Whole genome/exome sequencing (WGS/WES)
- Variant calling (SNPs, CNVs, structural variations)

Epigenomics

- Bisulphite sequencing : DNA methylation
- ATAC-Seq : chromatine opening
- ChIP-seq : TF binding sites

Transcriptomics

- Total RNA
- Poly-A tail selection : focus on mRNA
- Ribo depletion: mRNA + ncRNA
- 5'/3' RACE seq : isoform characterization for one gene
- scRNAseq
- Long read RNA sequencing
- ...

Imagination is the limit

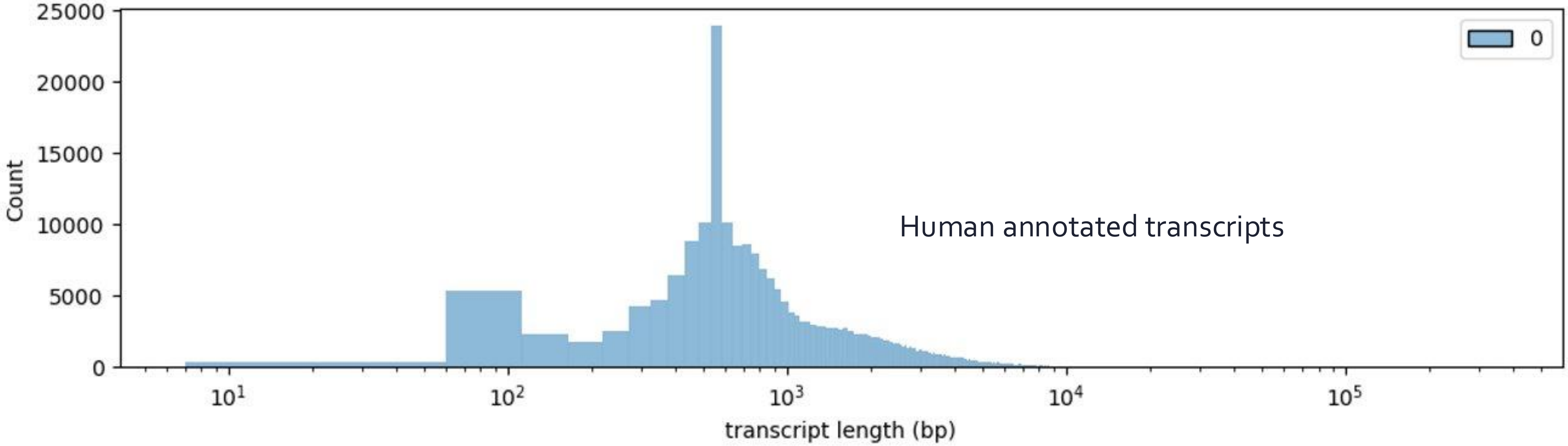
See : https://liorpachter.wordpress.com/seq/

slides Outline

- RNA and molecular biology
- **Main challenges for RNAseq**
- Major Sequencing technologies
- Planning your sequencing : choices, number of samples, ...
- Bioinformatics analysis overview

Main challenges of RNAseq

Transcripts are diverse in size

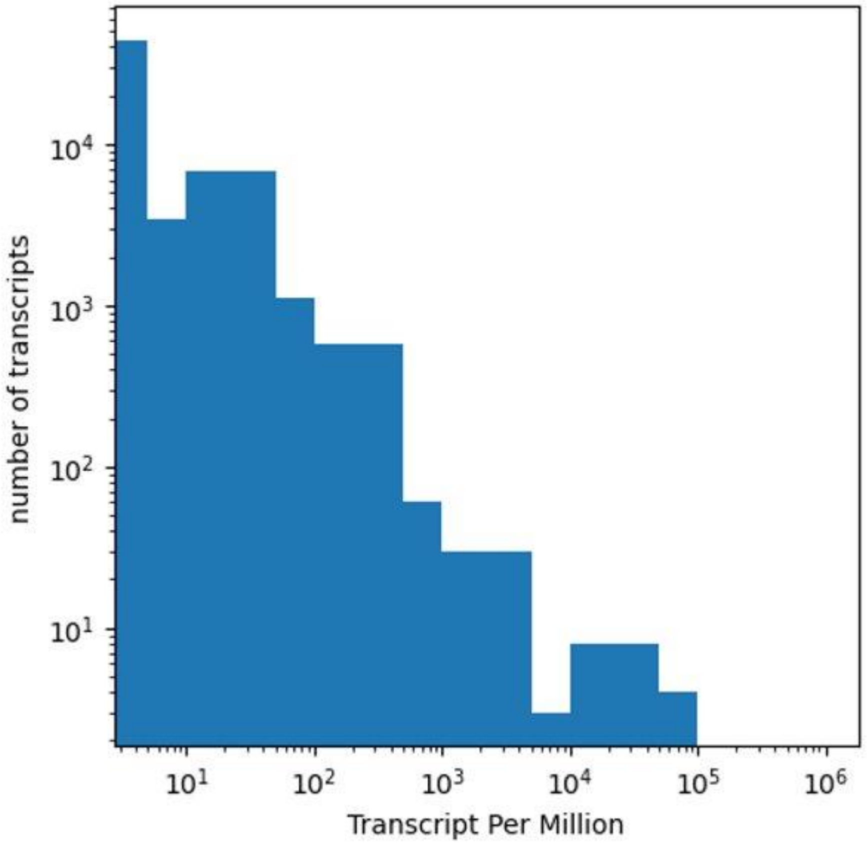


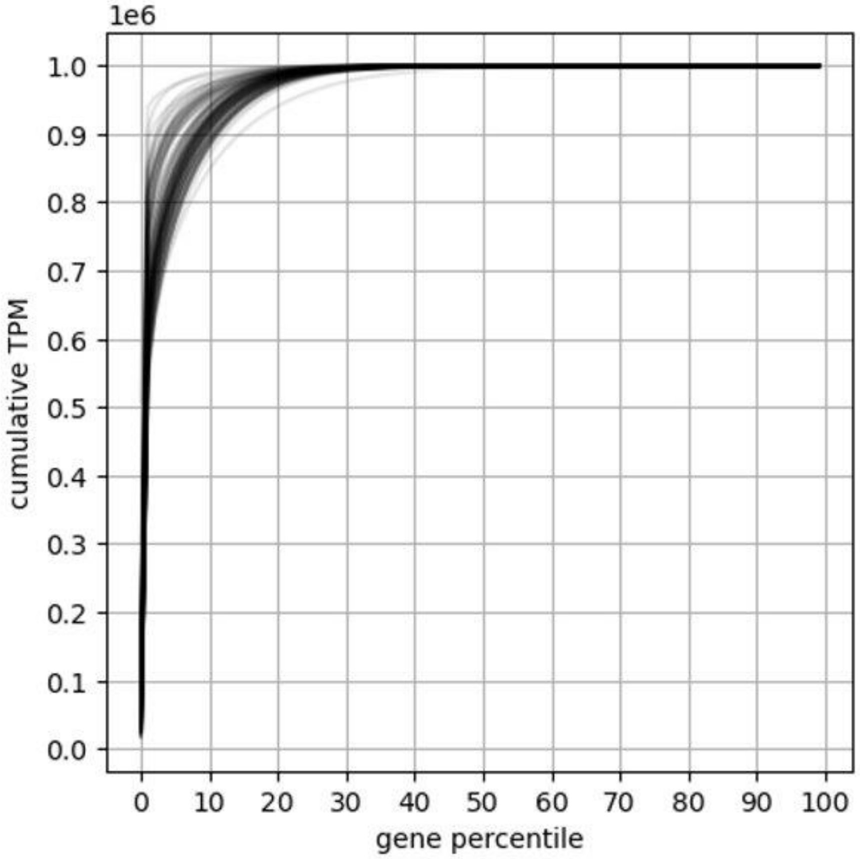
Human annotated transcripts

Main challenges of RNAseq

Transcripts are diverse in size

Expression levels have a *high dynamic range*





From Gtex V8 – human tissue samples

Data source : https://gtexportal.org/home/downloads/adult-gtex/bulk_tissue_expression

Main challenges of RNAseq

Transcripts are diverse in size

Expression levels have a *high dynamic range*

RNA molecules are exposed to degradation enzyme:

- RNA integrity affects results

Is there a reference genome.

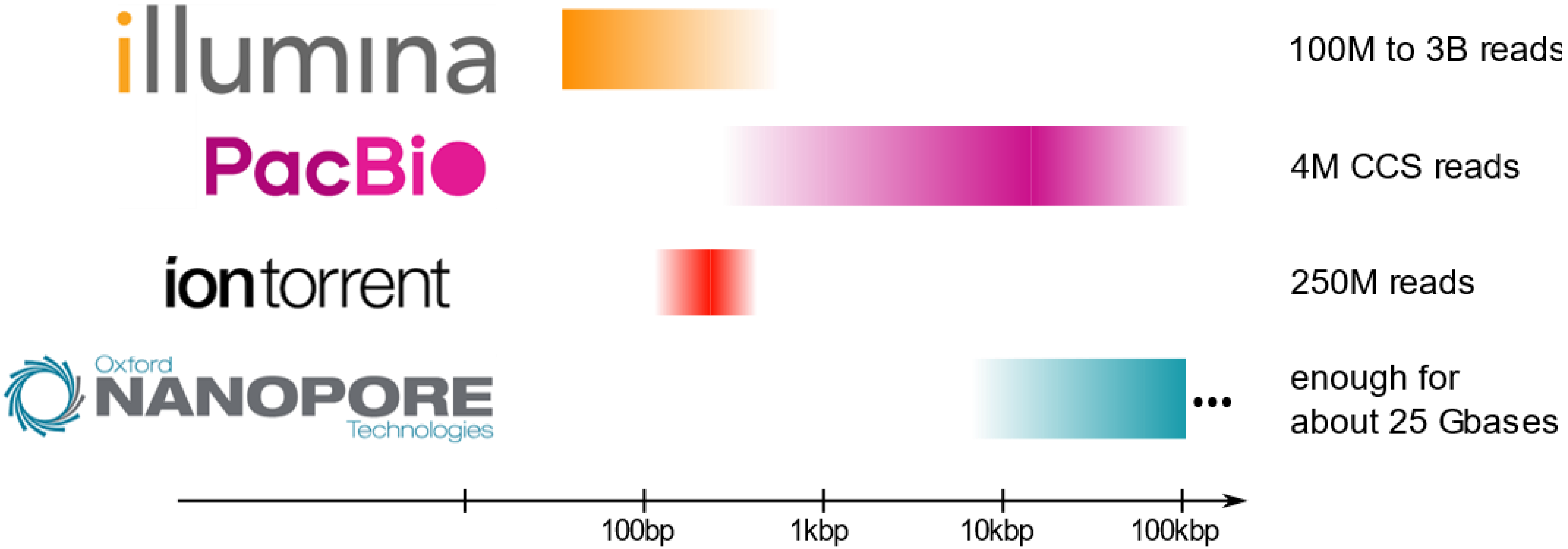
If yes,

- How good is it?
- How good is the gene annotation ?

slides Outline

- RNA and molecular biology
- Main challenges for RNAseq
- **Major Sequencing technologies**
- Planning your sequencing : choices, number of samples, ...
- Bioinformatics analysis overview

Main sequencing technologies



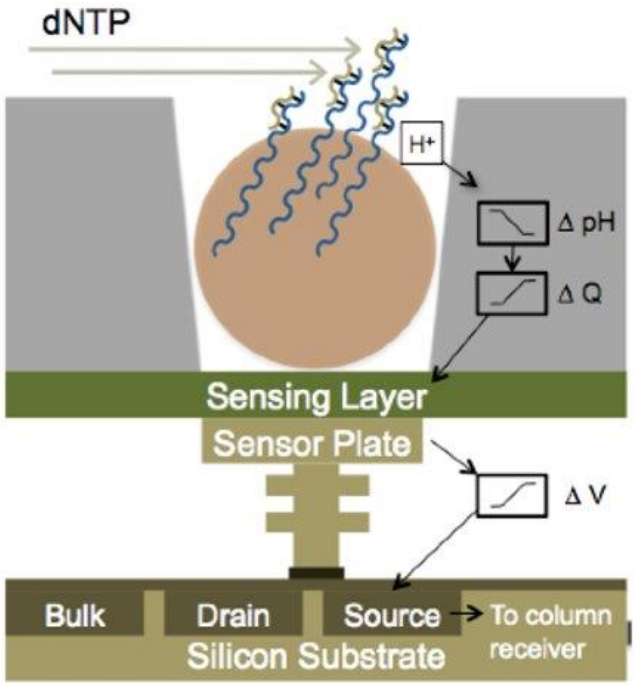
100M to 3B reads

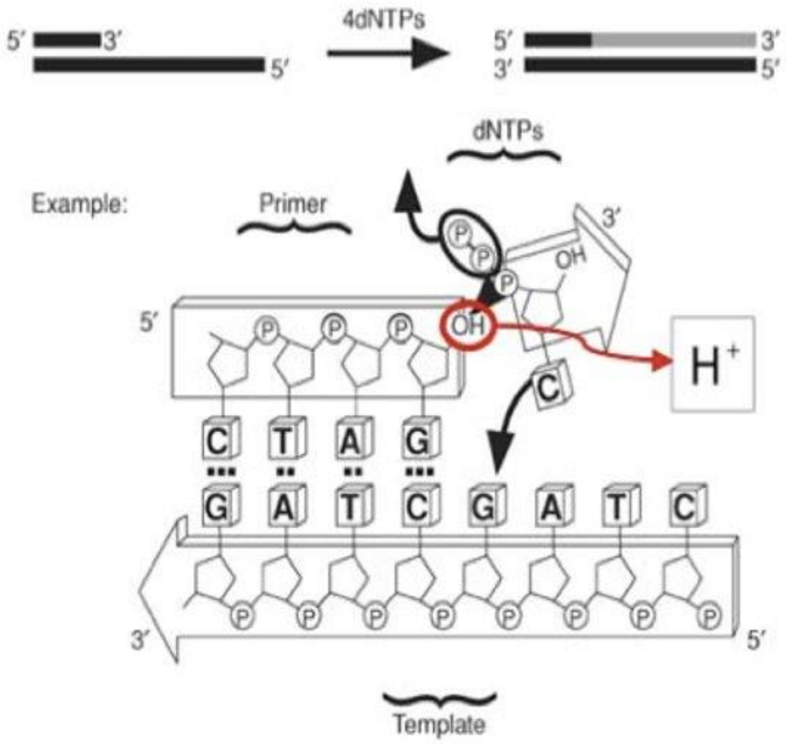
4M CCS reads

250M reads

enough for
about 25 Gbases

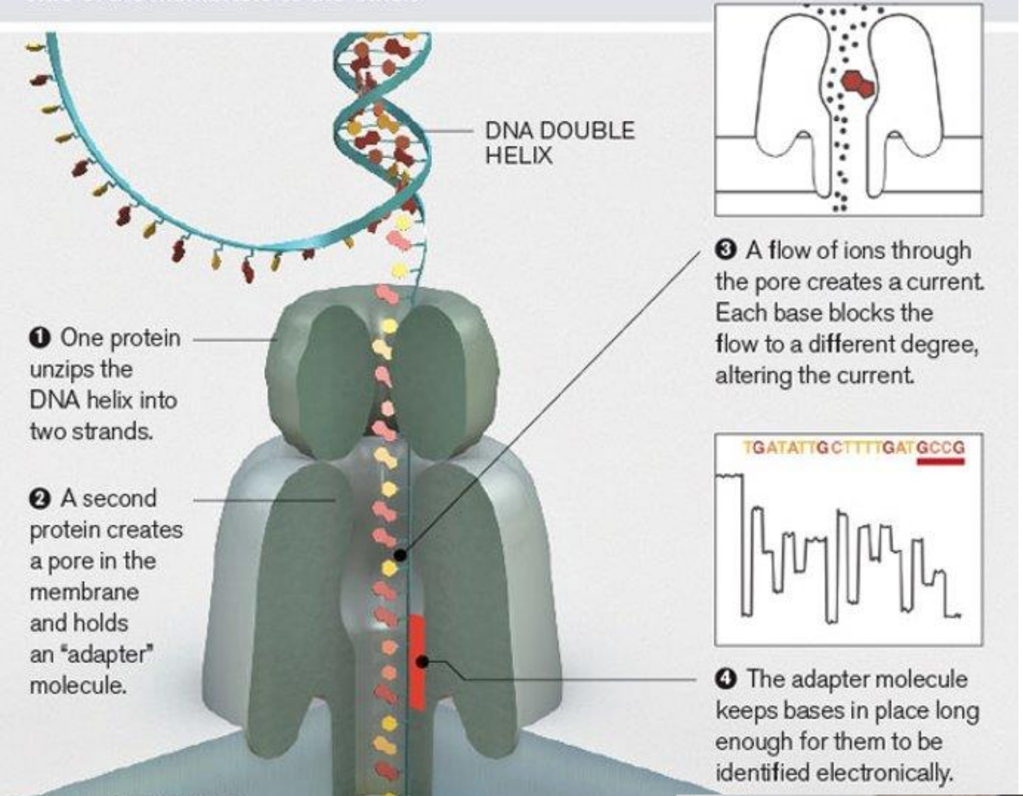
Ion torrent - reading pH changes

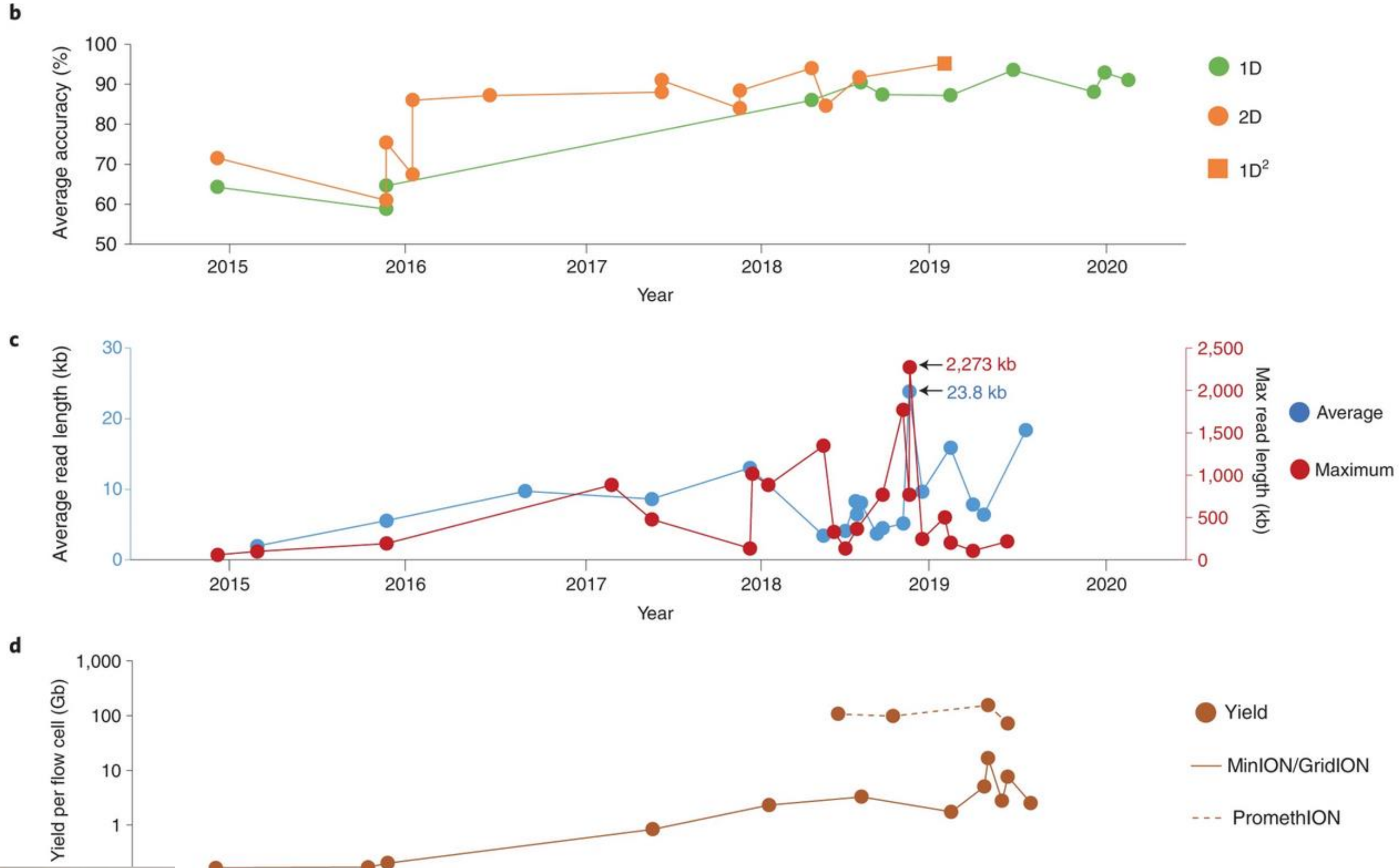




Oxford Nanopore - direct sequencing

DNA can be sequenced by threading it through a microscopic pore in a membrane. Bases are identified by the way they affect ions flowing through the pore from one side of the membrane to the other.


DNA DOUBLE HELIX
❶ One protein unzips the DNA helix into two strands.
❷ A second protein creates a pore in the membrane and holds an "adapter" molecule.
❸ A flow of ions through the pore creates a current. Each base blocks the flow to a different degree, altering the current.
TGATATTGCTTTTGATGCCG
❹ The adapter molecule keeps bases in place long enough for them to be identified electronically.
MEMBRANE



From Wang, Y., *et al.* Nanopore sequencing technology, bioinformatics and applications. *Nat Biotechnol* **39**, 1348–1365 (2021). https://doi.org/10.1038/s41587-021-01108-x

Pacific Biosciences - Single Molecule Real Time

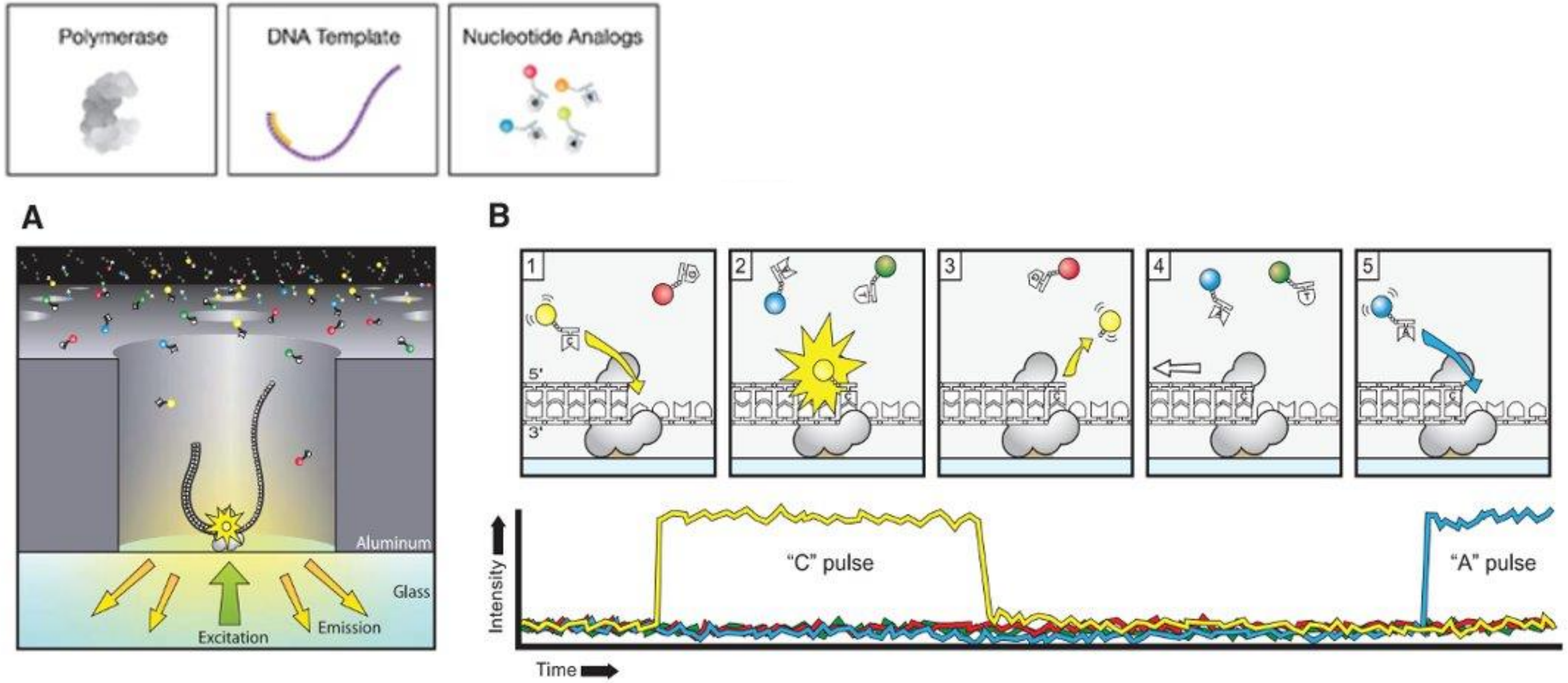
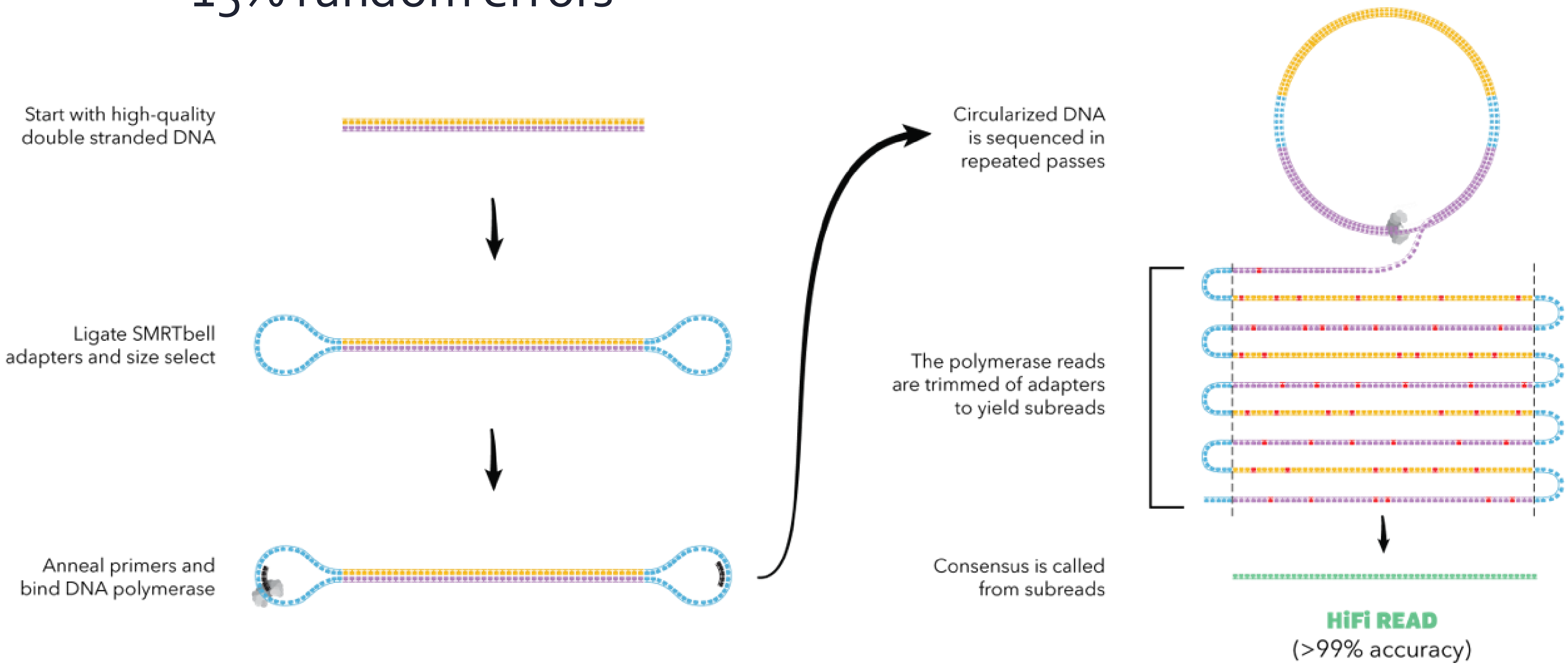


Fig. 1. Principle of single-molecule, real-time DNA sequencing. (**A**) Experimental geometry. A single molecule of DNA template-bound Φ29 DNA polymerase is immobilized at the bottom of a ZMW, which is illuminated from below by laser light. The ZMW nanostructure provides excitation confinement in the zeptoliter (10^{-21} liter) regime, enabling detection of individual phospholinked nucleotide substrates against the bulk solution background as they are incorporated into the DNA strand by the polymerase. (**B**) Schematic event sequence of the phospholinked dNTP incorporation cycle, with a corresponding expected time trace of detected fluorescence intensity from the ZMW. (1) A phospholinked nucleotide forms a cognate association with the template in the polymerase active site, (2) causing an elevation of the fluorescence output on the corresponding color channel. (3) Phosphodiester bond formation liberates the dye-linker-pyrophosphate product, which diffuses out of the ZMW, thus ending the fluorescence pulse. (4) The polymerase translocates to the next position, and (5) the next cognate nucleotide binds the active site beginning the subsequent pulse.

From Rhoads & Au. Genomics Proteomics Bioinformatics 2015

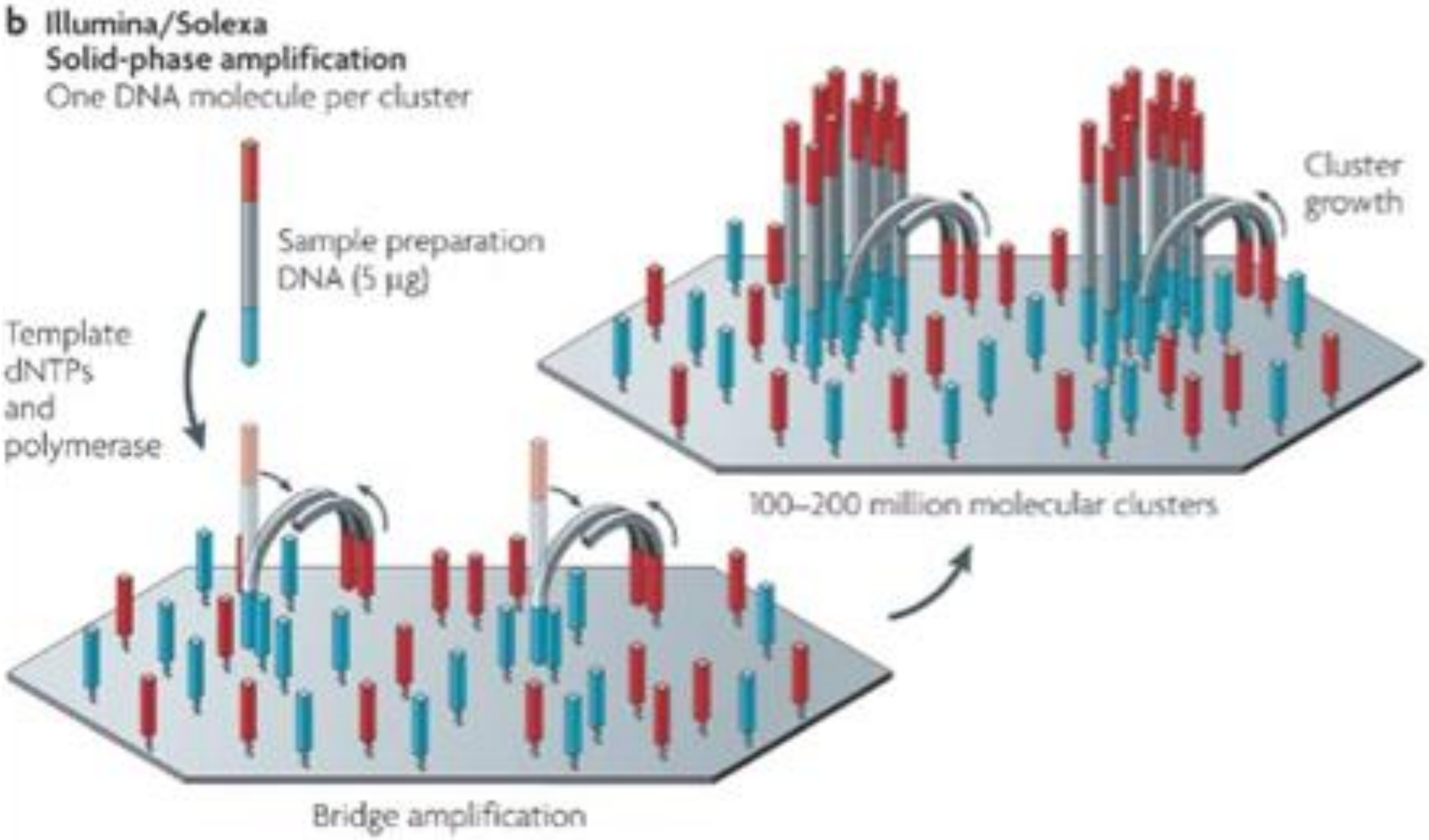
Pacific Biosciences - Circular Consensus Sequencing

Raw reads
~15% random errors


Start with high-quality double stranded DNA
Ligate SMRTbell adapters and size select
Anneal primers and bind DNA polymerase
Circularized DNA is sequenced in repeated passes
The polymerase reads are trimmed of adapters to yield subreads
Consensus is called from subreads
HiFi READ
(>99% accuracy)

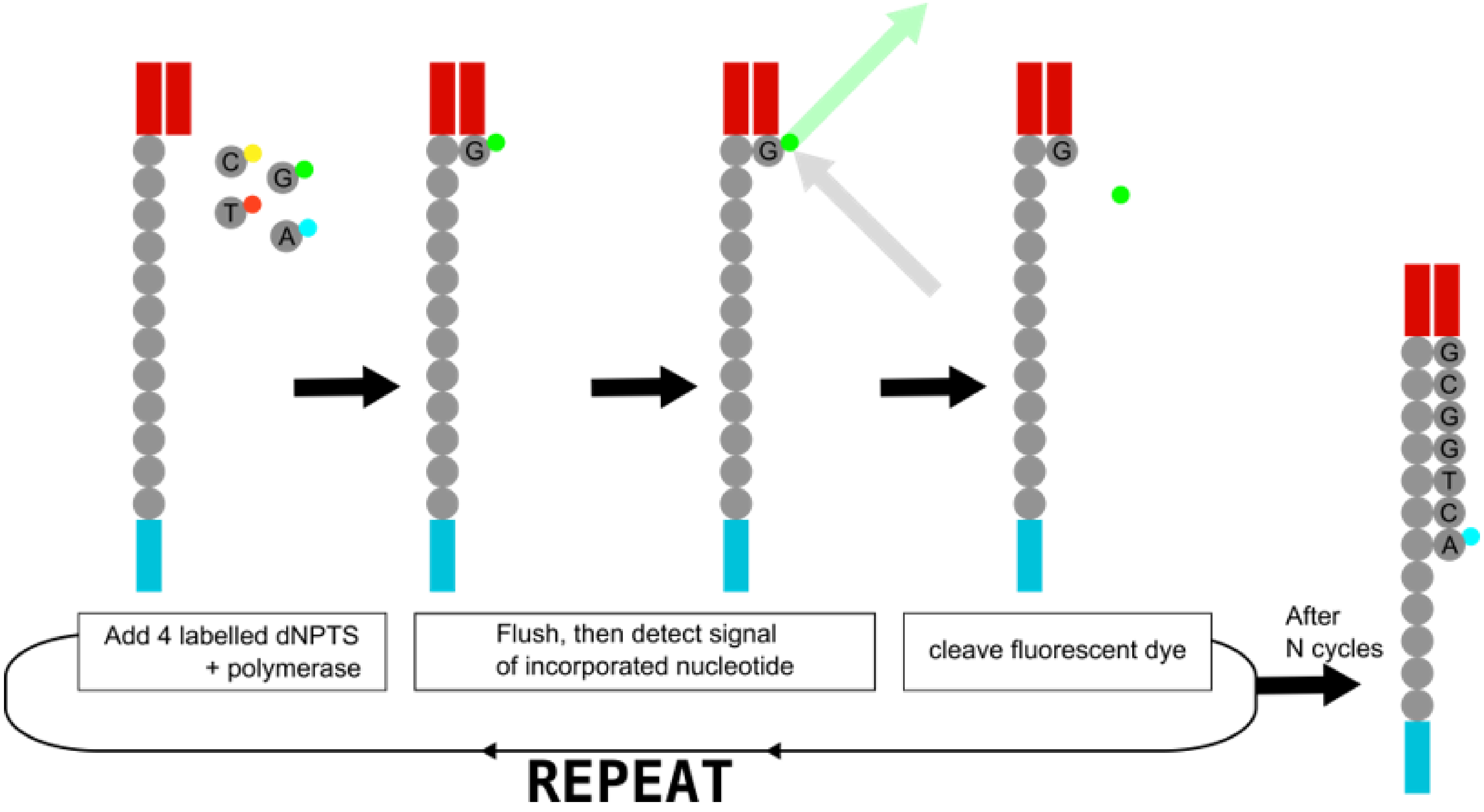
Typical in isoseq

Illumina sequencing - cluster formation


b Illumina/Solexa
Solid-phase amplification
One DNA molecule per cluster
Sample preparation DNA (5 μg)
Template dNTPs and polymerase
Cluster growth
100–200 million molecular clusters
Bridge amplification

Source: Fig1b of Metzker, M. Sequencing technologies — the next generation. Nat Rev Genet 11, 31–46 (2010). https://doi.org/10.1038/nrg2626

Illumina sequencing - sequencing by synthesis



Add 4 labelled dNPTS + polymerase

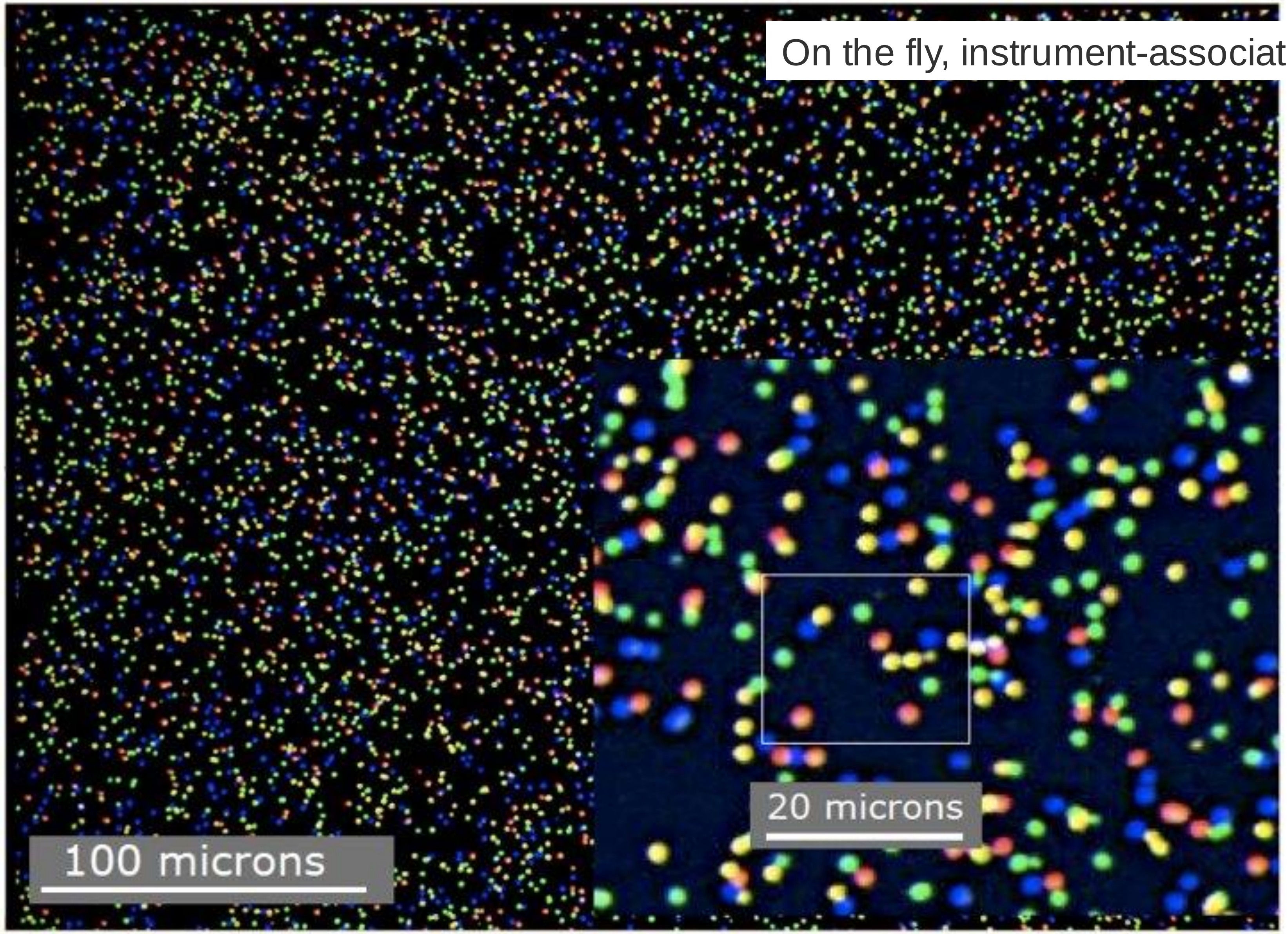
Flush, then detect signal of incorporated nucleotide

cleave fluorescent dye

After N cycles

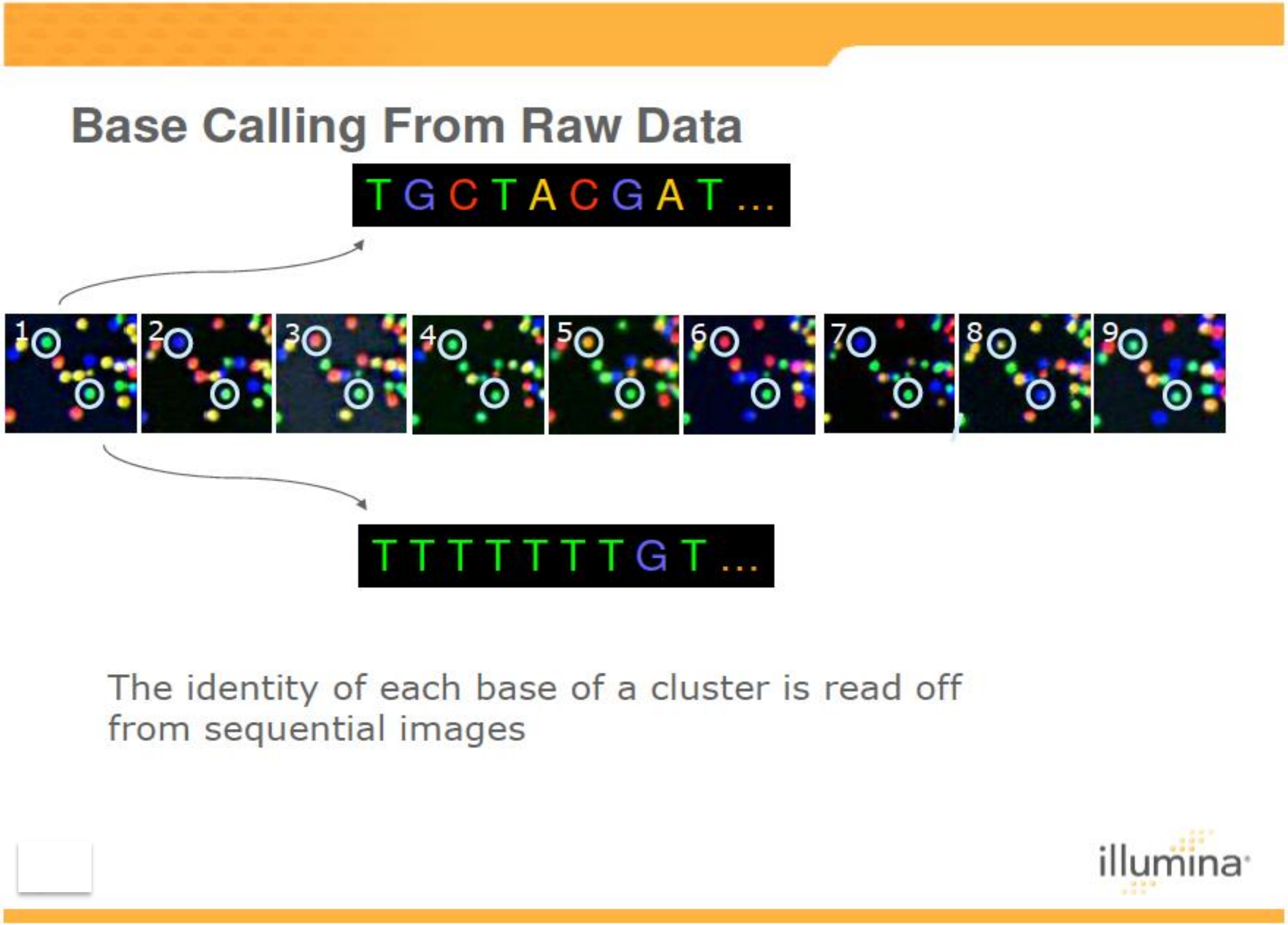
REPEAT

Illumina sequencing - image analysis



On the fly, instrument-associated software

Illumina sequencing - from image to sequence



Base Calling From Raw Data

T G C T A C G A T …

T T T T T T T G T …

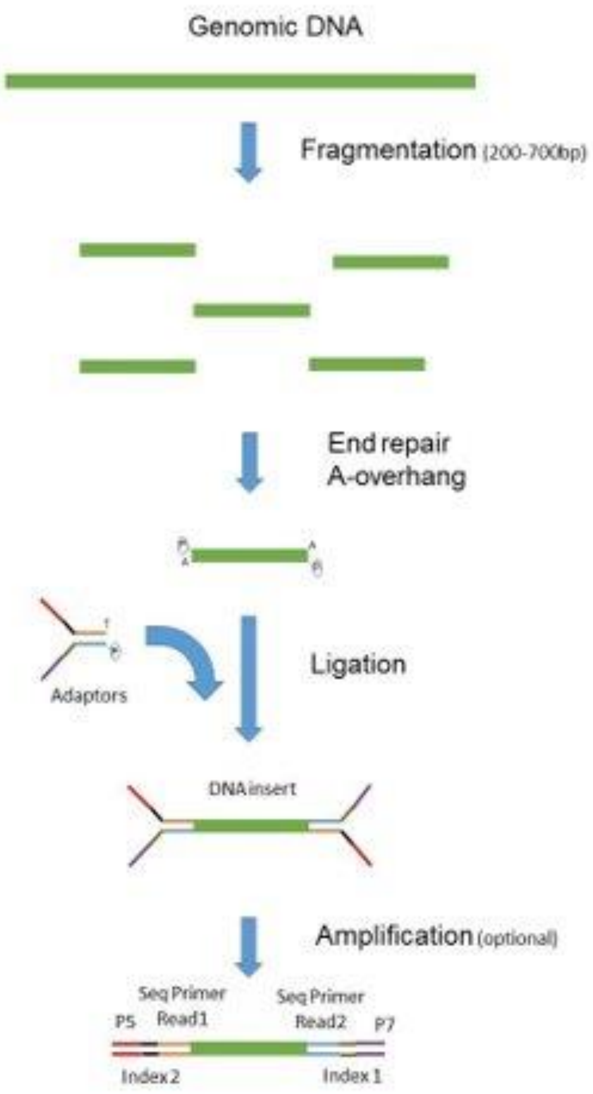
The identity of each base of a cluster is read off from sequential images

slides Outline

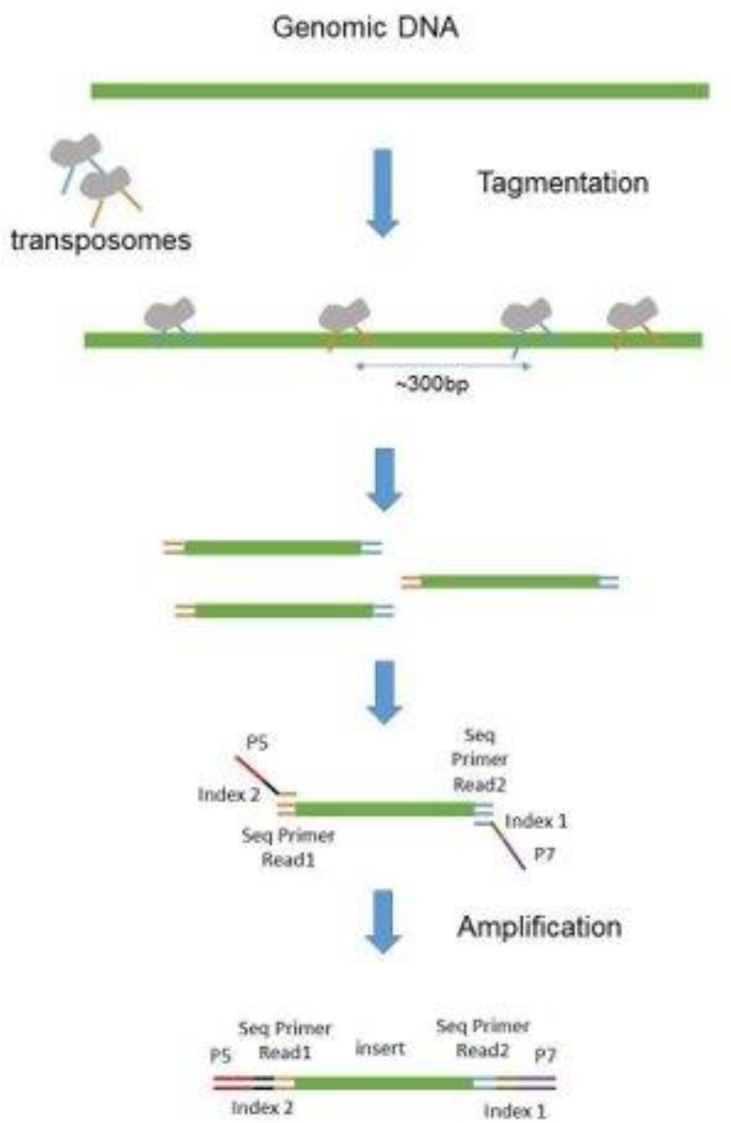
- RNA and molecular biology
- Main challenges for RNAseq
- Major Sequencing technologies
- **Planning your sequencing : choices, number of samples, …**
- Bioinformatics analysis overview

Paired-end sequencing

“Classical” paired end library (illumina)



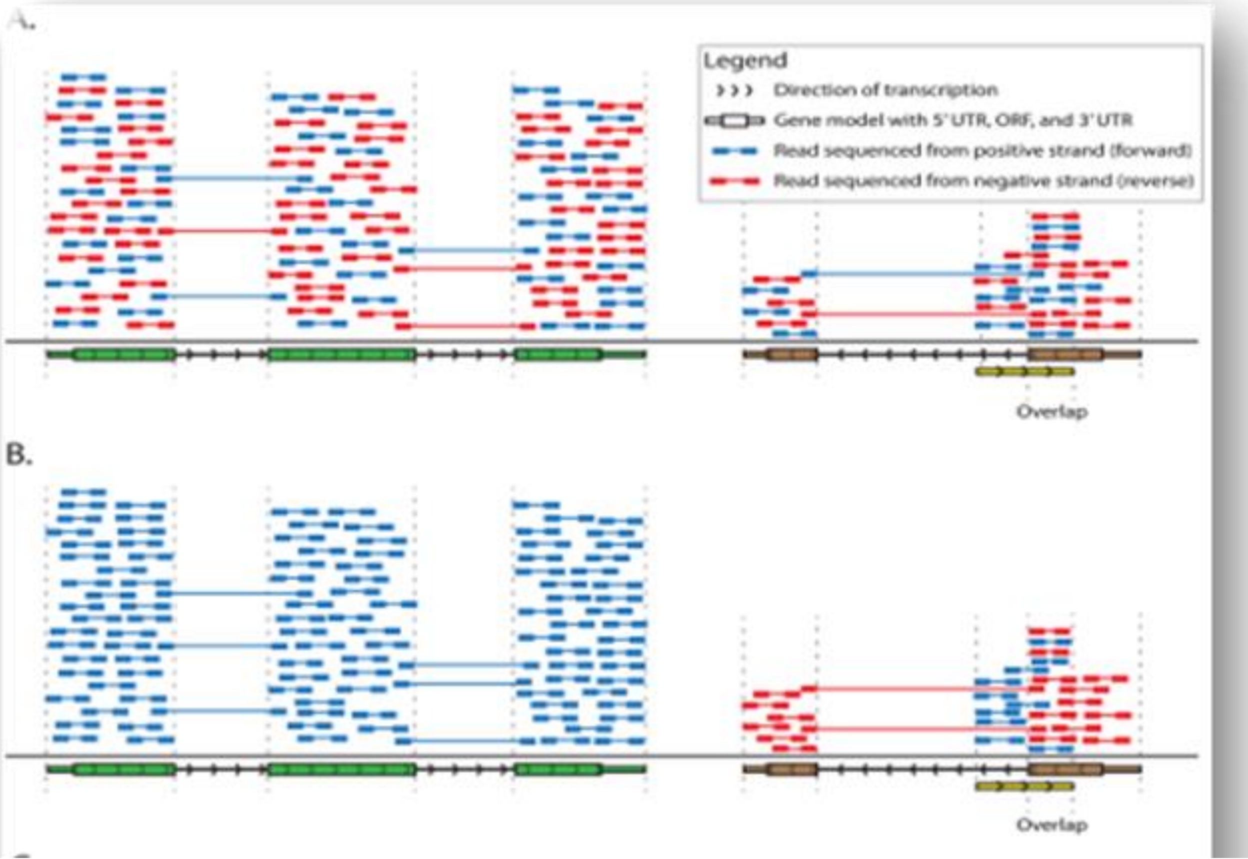
“Nextera” paired end library



Source : France Genomique

Stranded vs Unstranded Sequencing

- Overlapping genes regions are substantial (~8% in *Homo sapiens*)
- Stranded sequencing allows us to quantify expression in these overlapping regions
- Achieved by ligating different adapters to 5' and 3' ends


A.
Legend
››› Direction of transcription
Gene model with 5' UTR, ORF, and 3' UTR
Read sequenced from positive strand (forward)
Read sequenced from negative strand (reverse)
Overlap
B.
Overlap

M.Griffith et al.
PLoS Comp Biol 2015
doi:10.1371/journal.pcbi.1004393

RNA purification

PolyA selection

- Commonly used and inexpensive
- 3' end bias when RNA is degraded
- Loses almost all non-polyA transcripts
- Gets rid of vast majority of ribosomal RNAs, but ncRNA too

Ribosomal RNA depletion

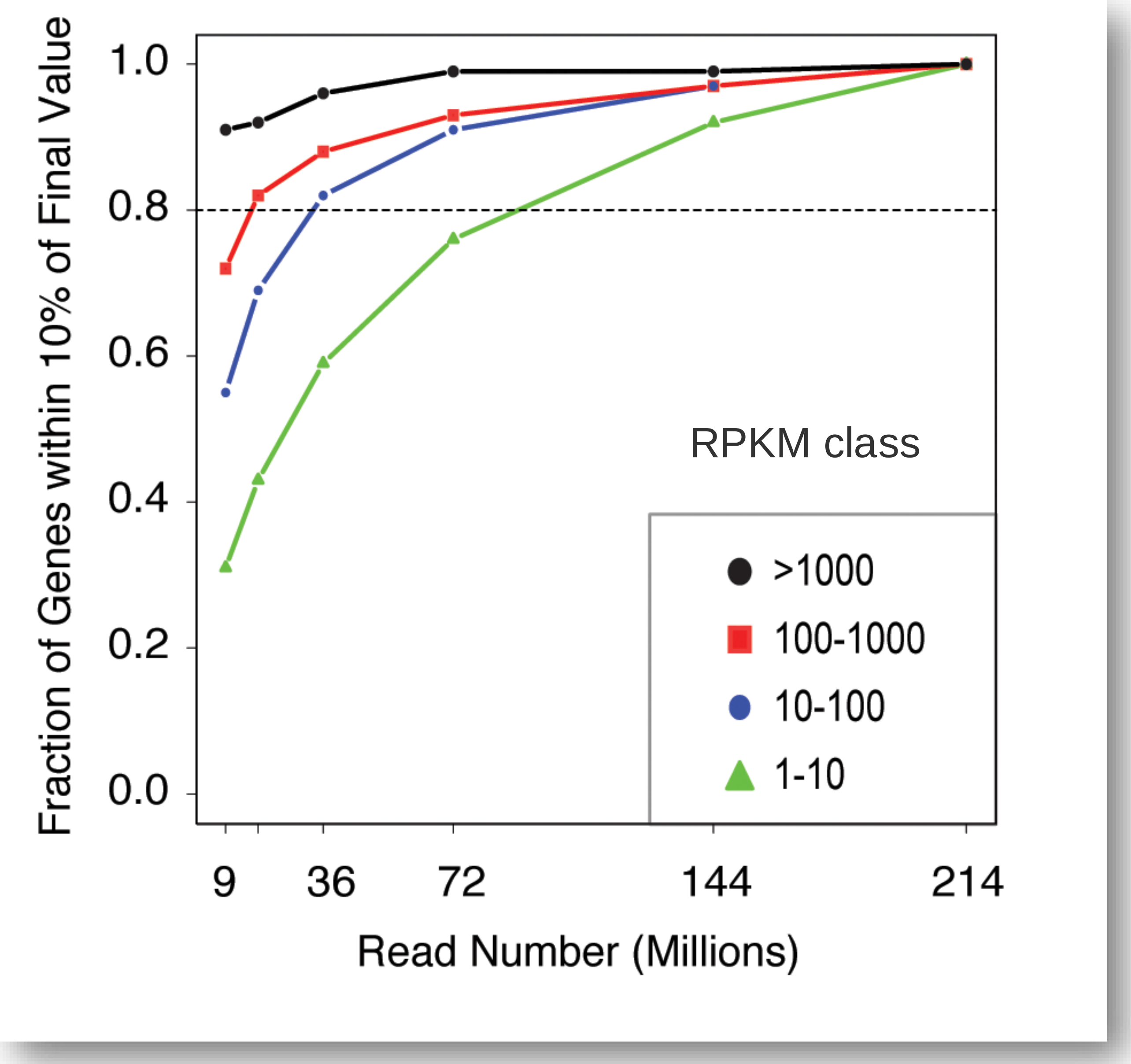
- Less popular, ~2x more expensive
- Higher proportion of rRNA than in polyA selection
- Bacterial data
- Allows identification of lncRNAs without polyA tails
- Retains more immature mRNAs (bad for gene expression quantification)

Sequencing depth

DE : usually aim for ~30-40 million reads

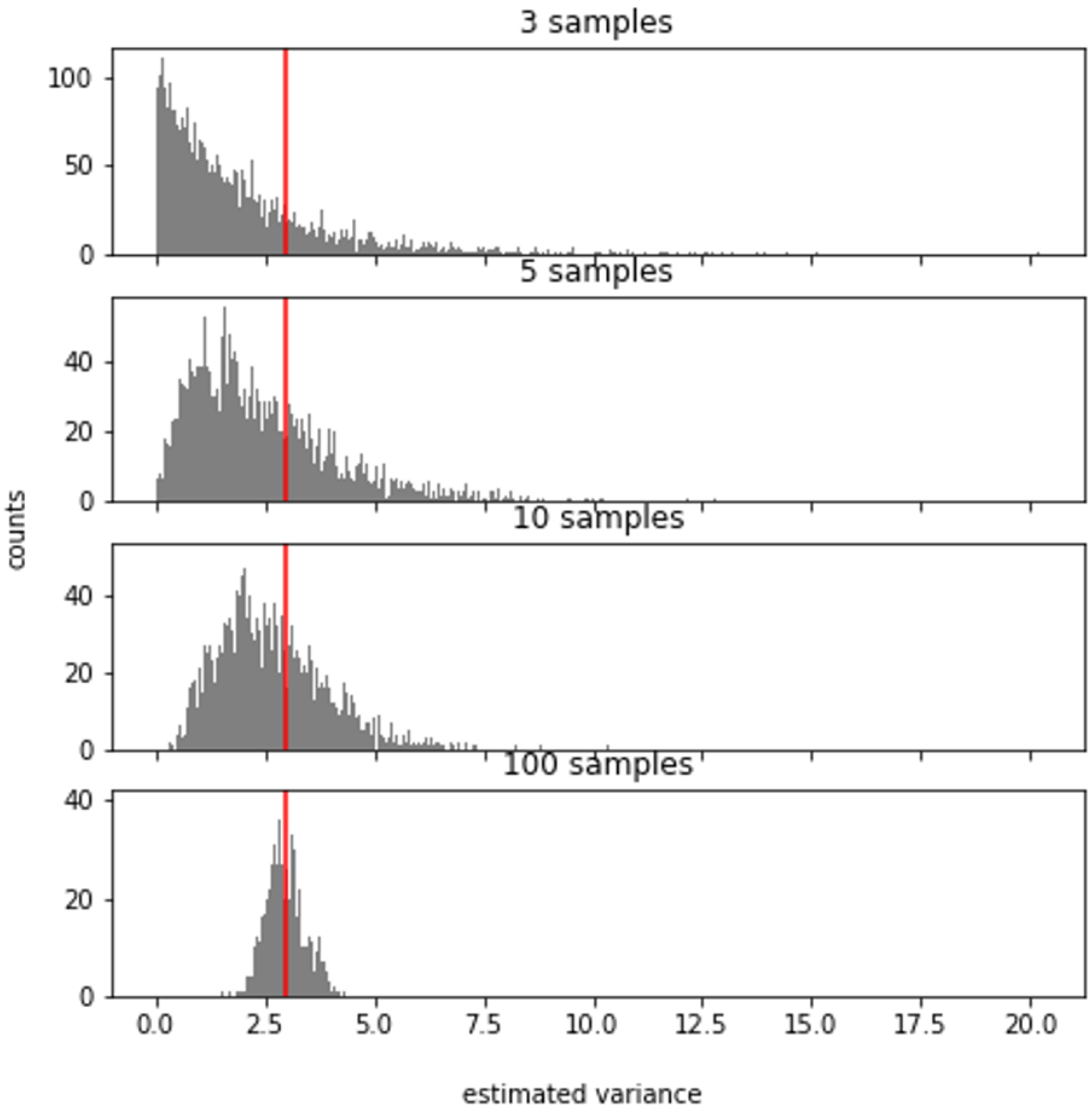
For rare events (isoforms, somatic mutations) much more depth is required

Not easy to know in advance



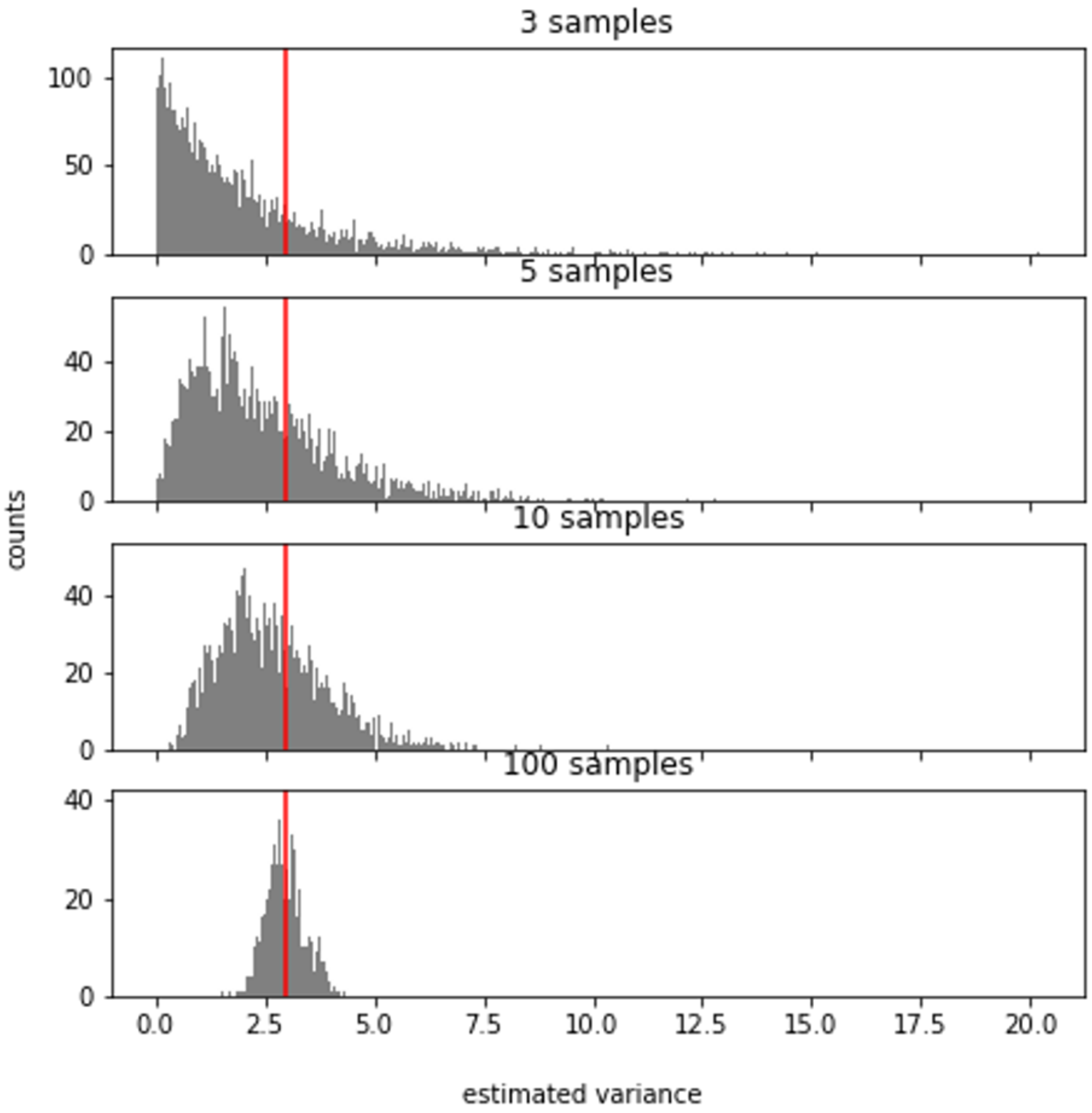
A User's Guide to the Encyclopedia of DNA Elements (ENCODE) PLoS 2011

Replicates - estimating a biological variance



What does this tell you about the number of replicates needed ?

Replicates - estimating a biological variance

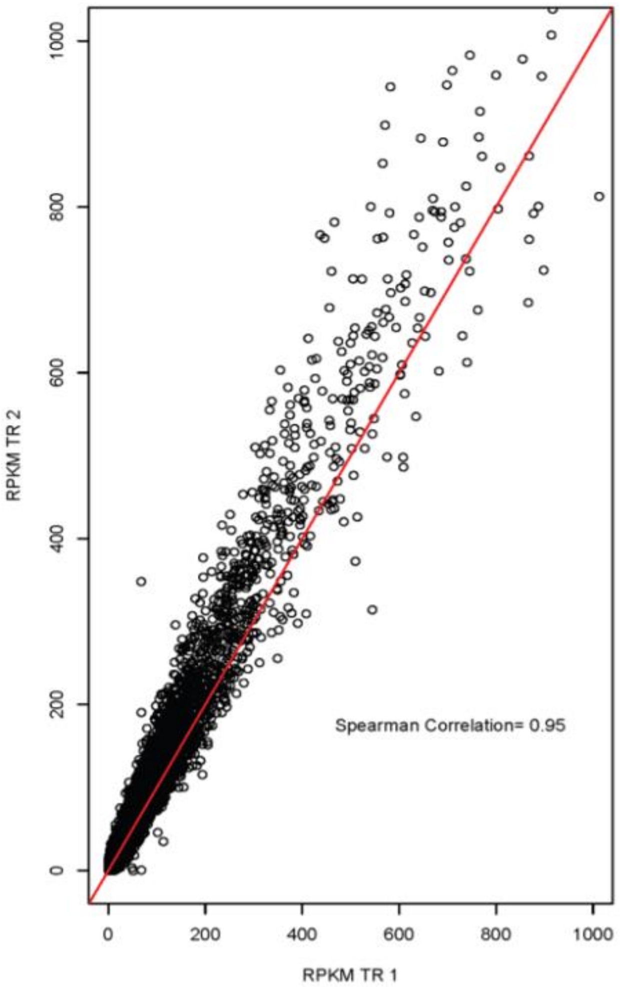


What does this tell you about the number of replicates needed ?

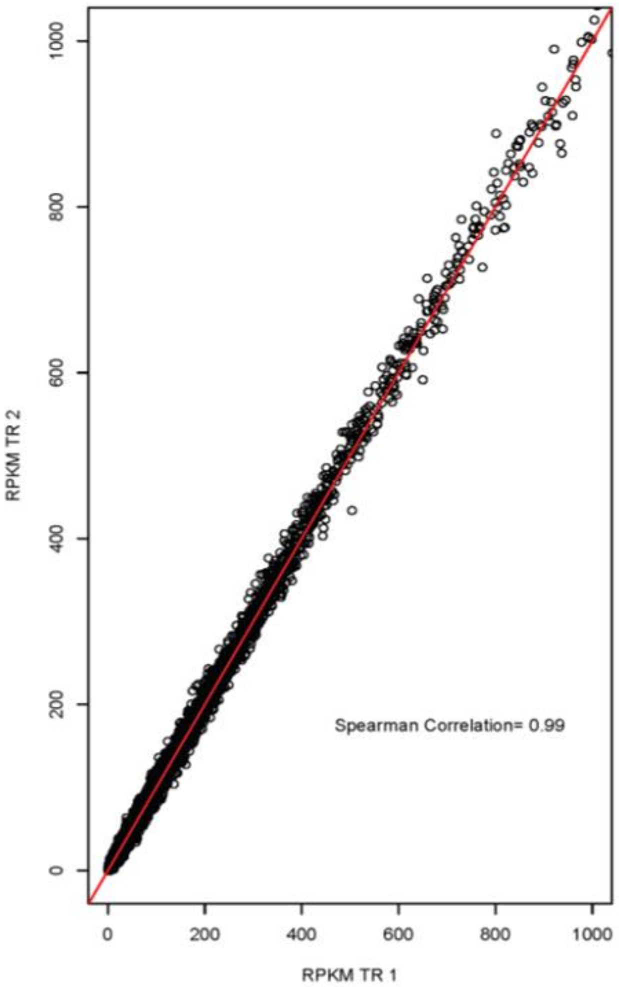
2 types of replicates:

- **Technical**: same RNA extract
- **Biological**: same biological condition

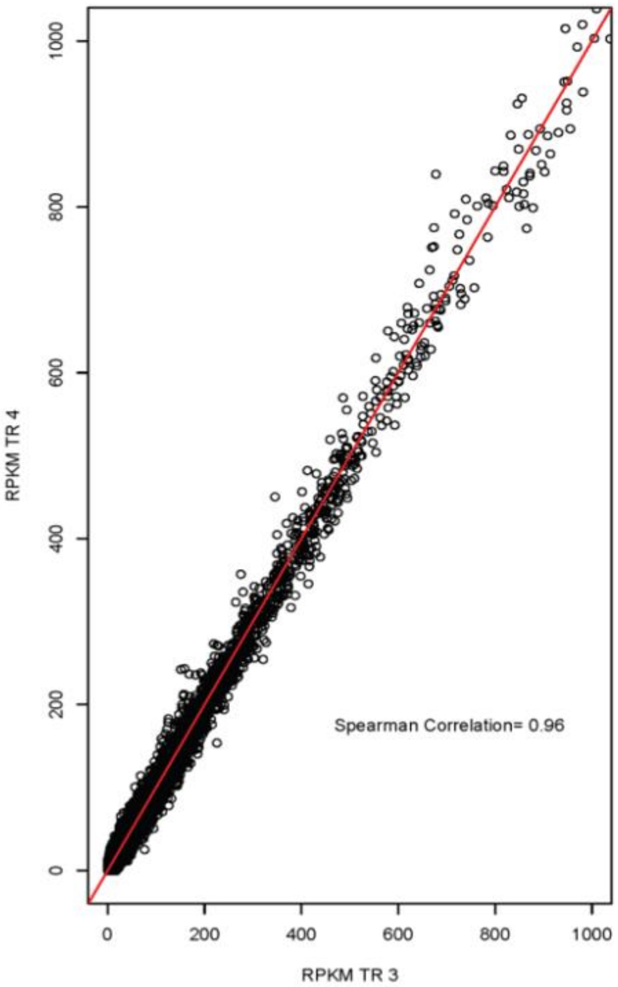
Technical replicates



D. simulans
Male heads



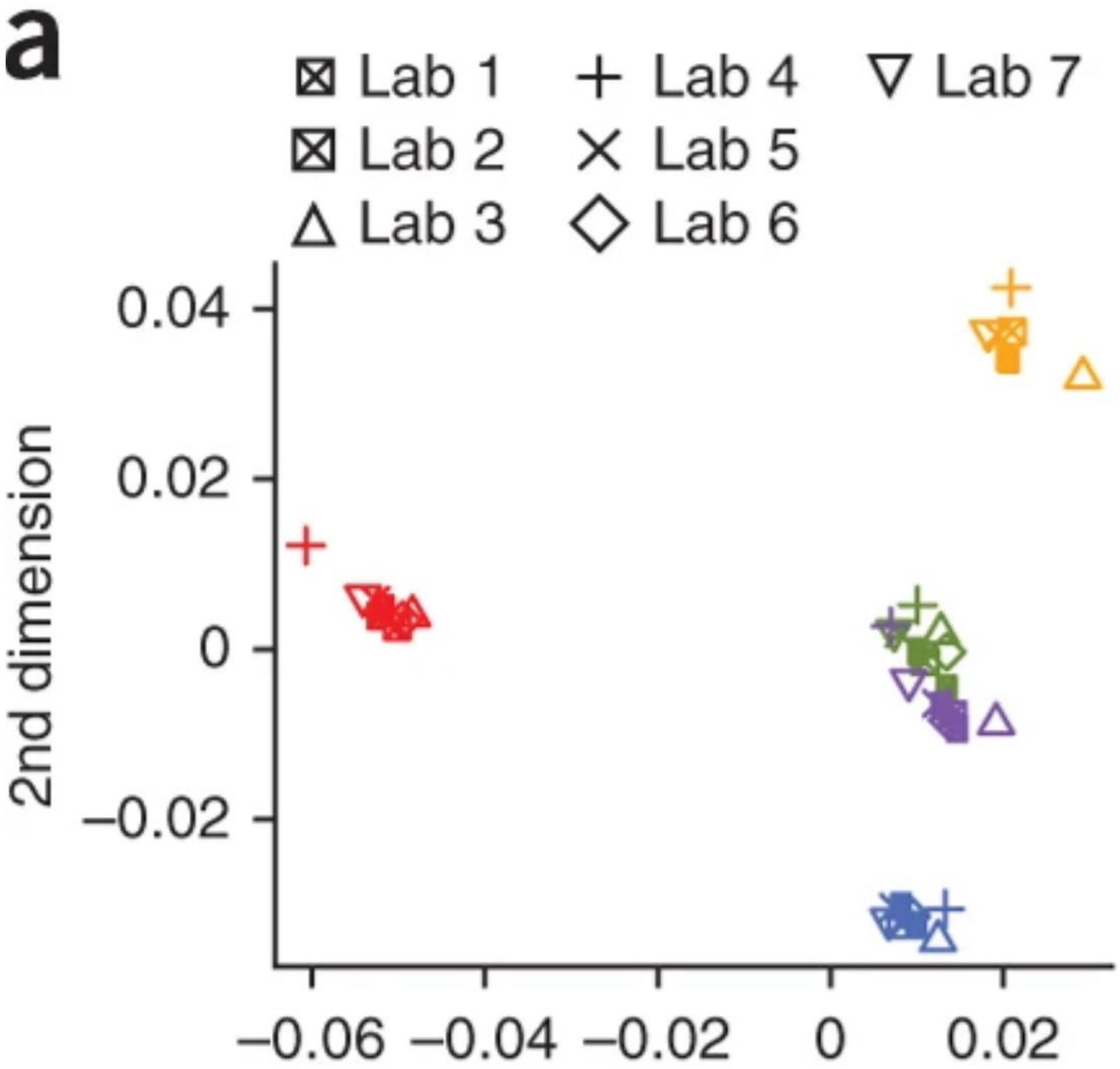
D. melanogaster
Female heads

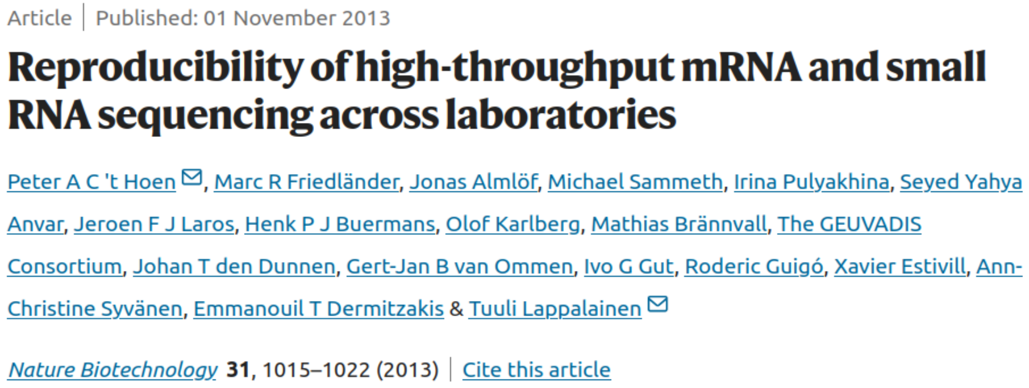


C167 cell line

Lauren M McIntyre et al., BMC Genomics 2011

Technical replicates




Article | Published: 01 November 2013

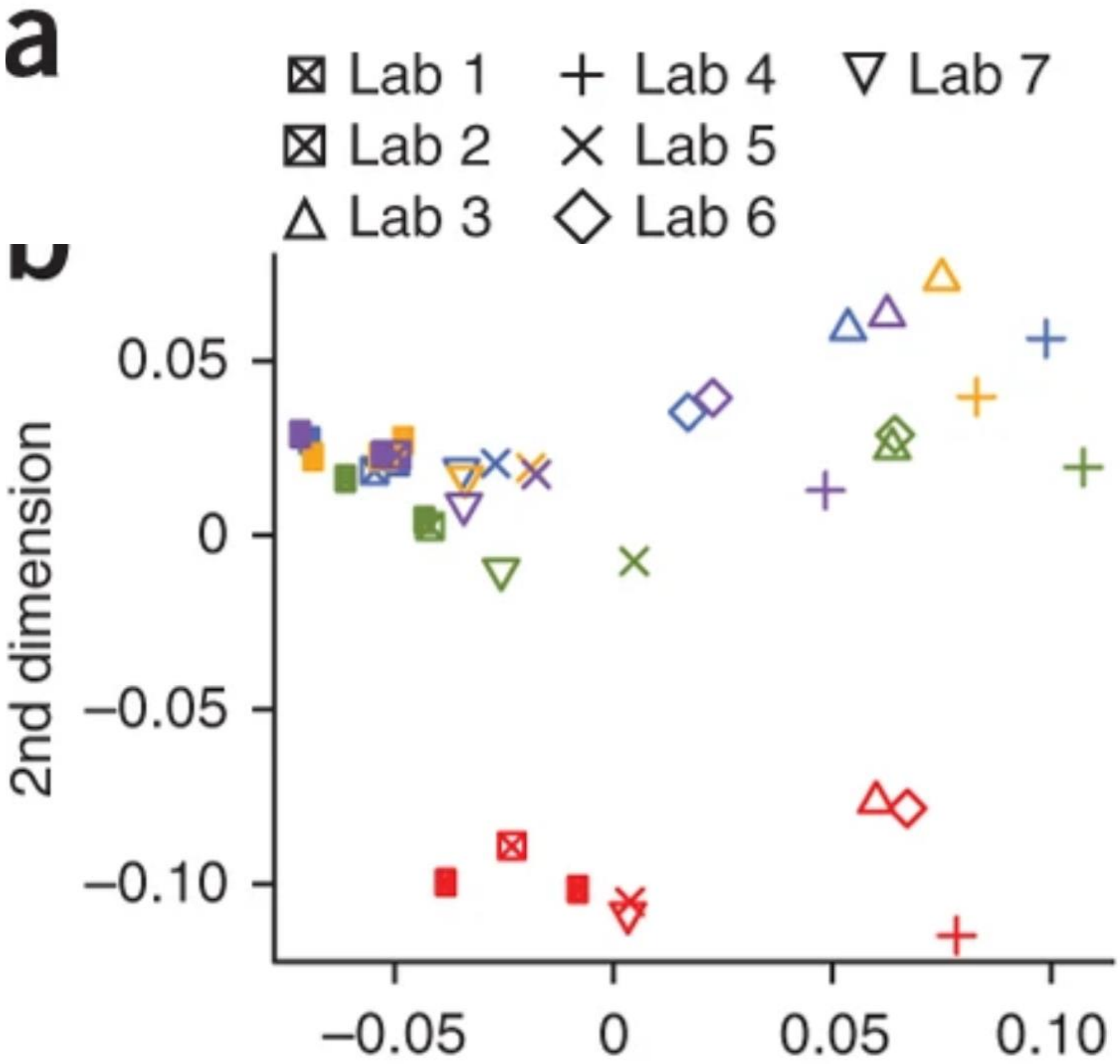
Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories

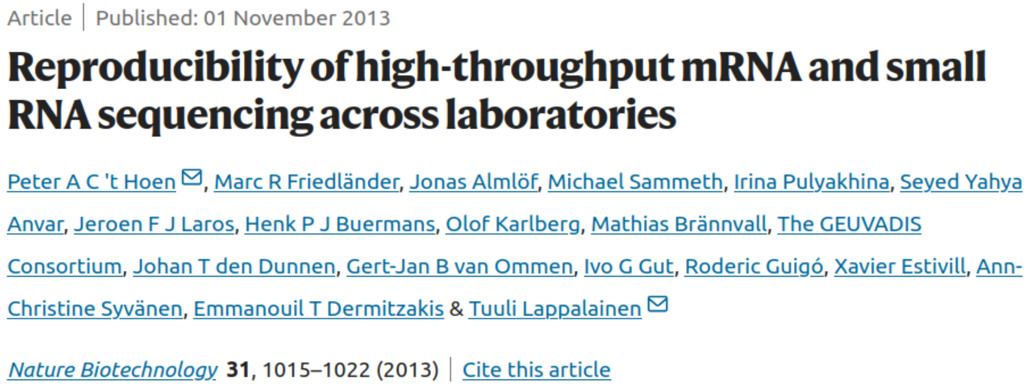
Peter A C 't Hoen, Marc R Friedländer, Jonas Almlöf, Michael Sammeth, Irina Pulyakhina, Seyed Yahya Anvar, Jeroen F J Laros, Henk P J Buermans, Olof Karlberg, Mathias Brännvall, The GEUVADIS Consortium, Johan T den Dunnen, Gert-Jan B van Ommen, Ivo G Gut, Roderic Guigó, Xavier Estivill, Ann-Christine Syvänen, Emmanouil T Dermitzakis & Tuuli Lappalainen

Nature Biotechnology **31**, 1015–1022 (2013) | Cite this article

Exon level: reproducible

Technical replicates




Article | Published: 01 November 2013

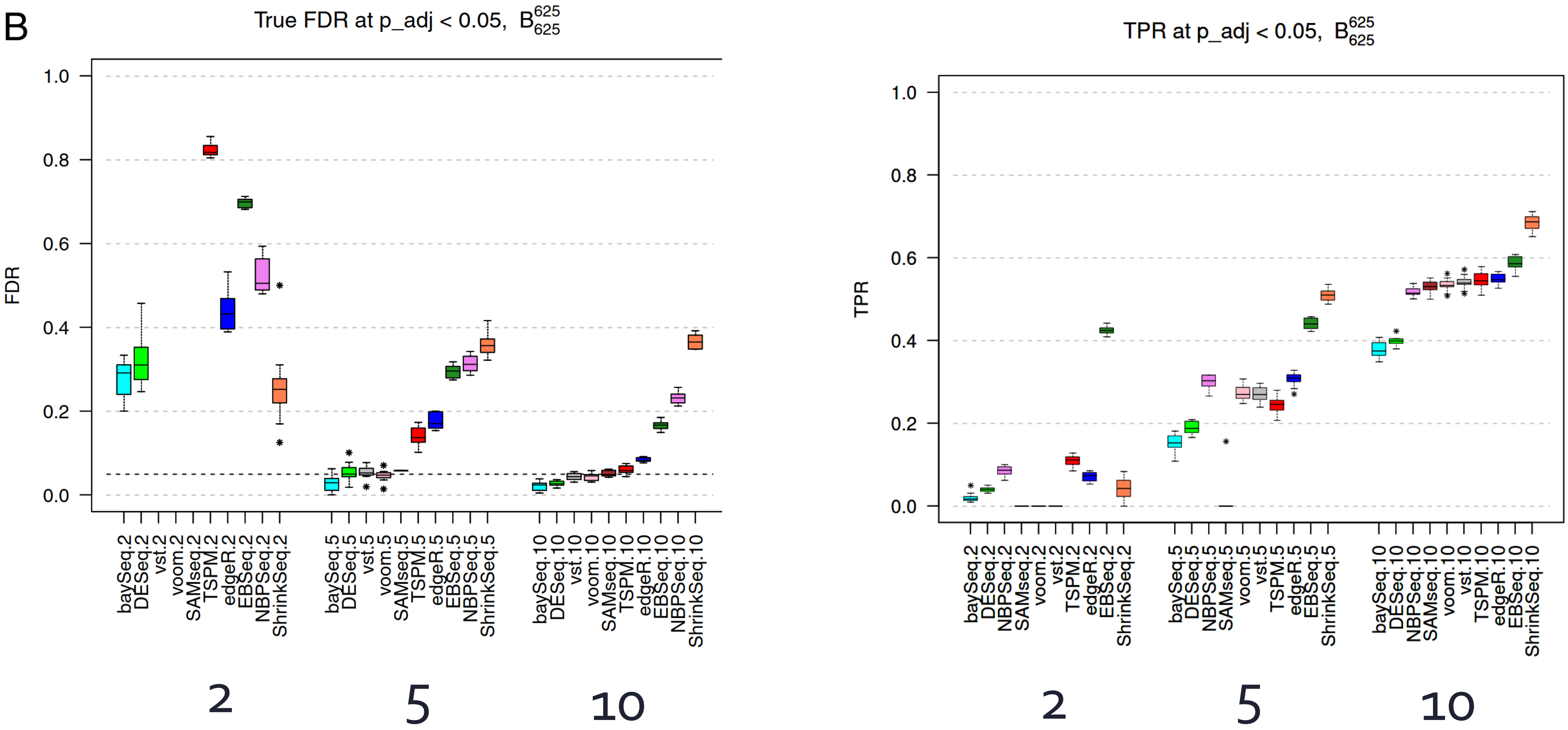
Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories

Peter A C 't Hoen, Marc R Friedländer, Jonas Almlöf, Michael Sammeth, Irina Pulyakhina, Seyed Yahya Anvar, Jeroen F J Laros, Henk P J Buermans, Olof Karlberg, Mathias Brännvall, The GEUVADIS Consortium, Johan T den Dunnen, Gert-Jan B van Ommen, Ivo G Gut, Roderic Guigó, Xavier Estivill, Ann-Christine Syvänen, Emmanouil T Dermitzakis & Tuuli Lappalainen

Nature Biotechnology **31**, 1015–1022 (2013) | Cite this article

Transcript level: not reproducible

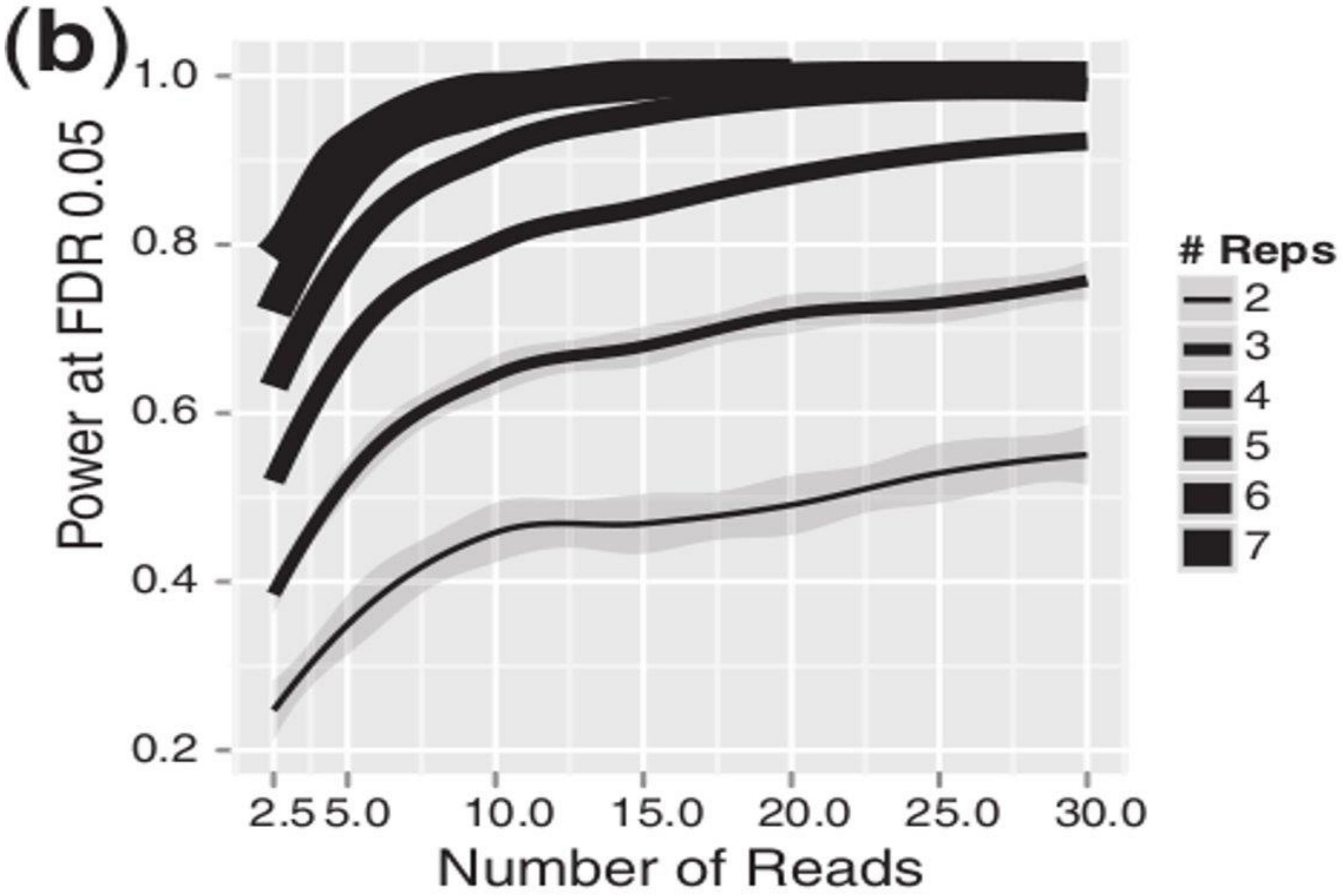
Biological replicates



Samples per condition

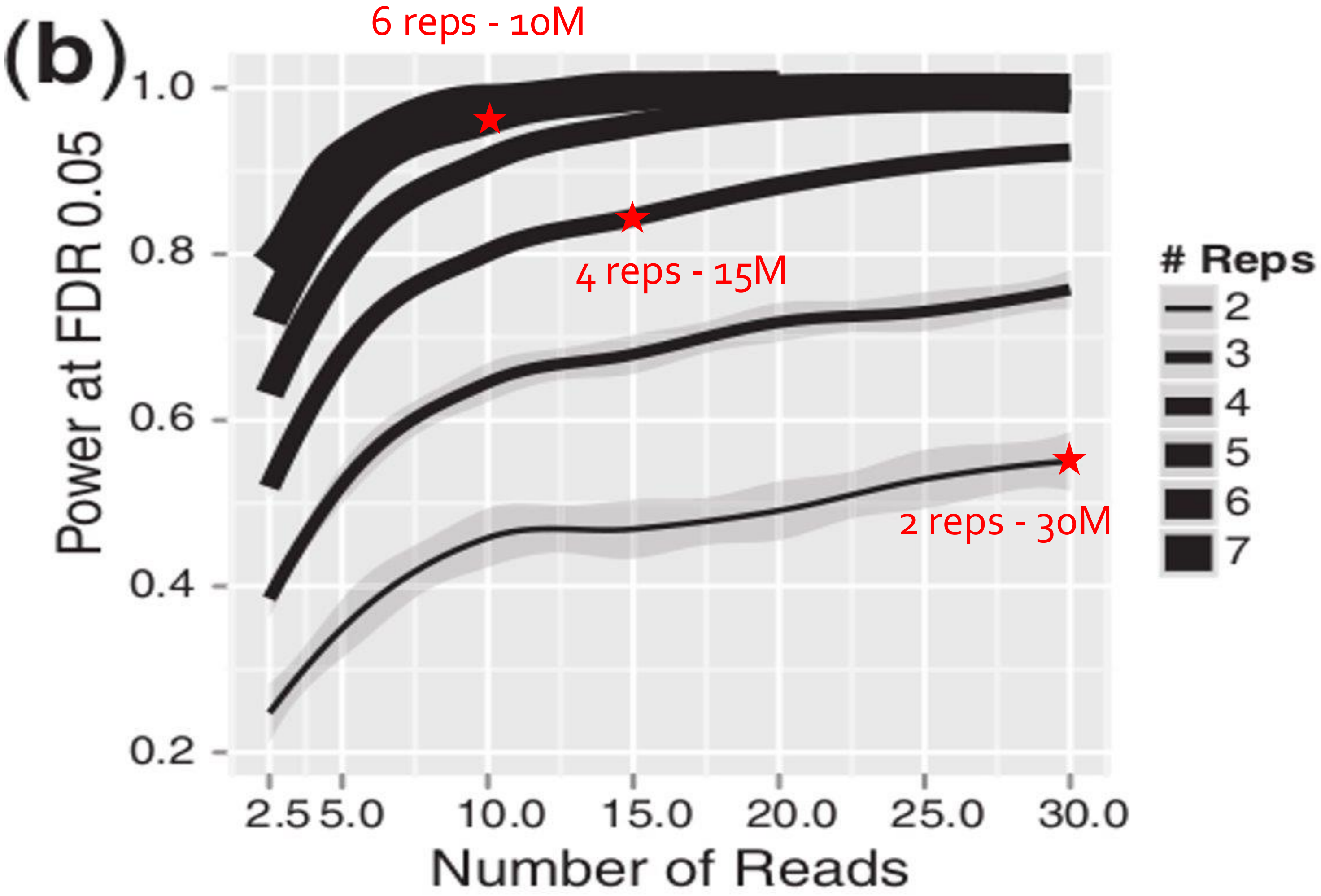
Soneson, C., Delorenzi, M. A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics* **14**, 91 (2013).

More reads or more replicates?



From Liu et al. 2014. RNA-seq differential expression studies: more sequence or more replication?

More reads or more replicates?

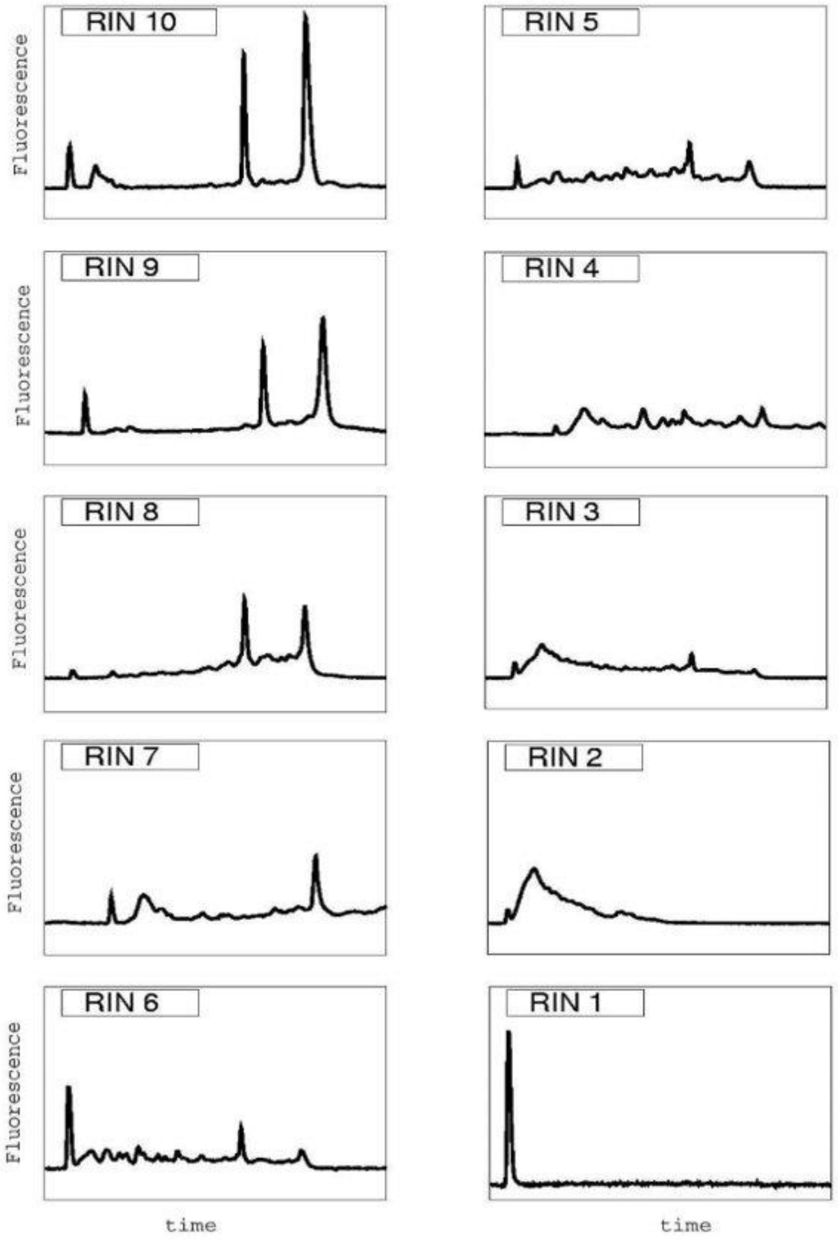


From Liu et al. 2014. RNA-seq differential expression studies: more sequence or more replication?

RNA sample preparation - RIN

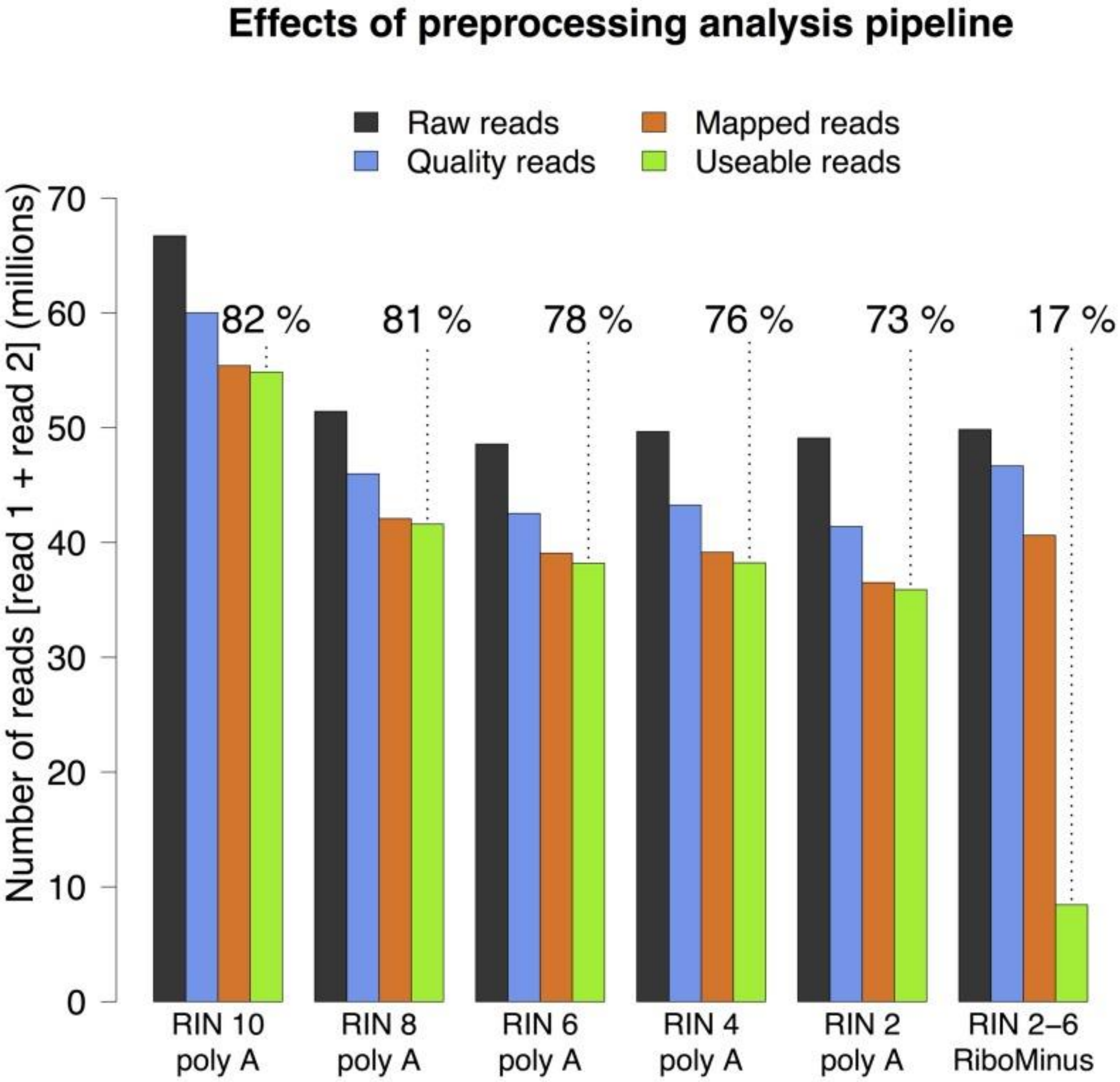
- Sample quality is critically important: we cannot make up for poor data
- RNA Integrity Number (RIN)

- Minimums:
  - 7-8 : eukaryot mRNA
  - 9 : bacterial

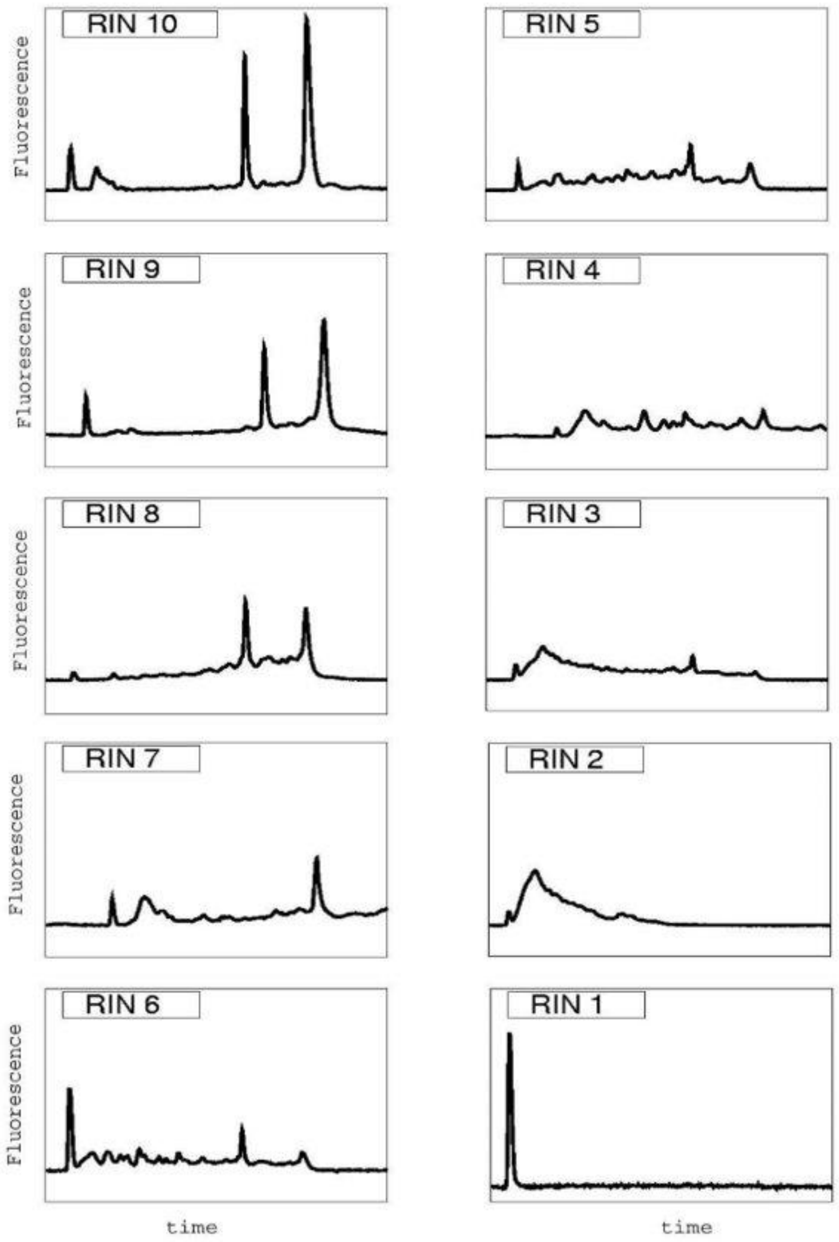


Schroeder *et al* BMC Mol Biol 2006

RNA sample preparation - RIN



Schroeder *et al* BMC Mol Biol 2006

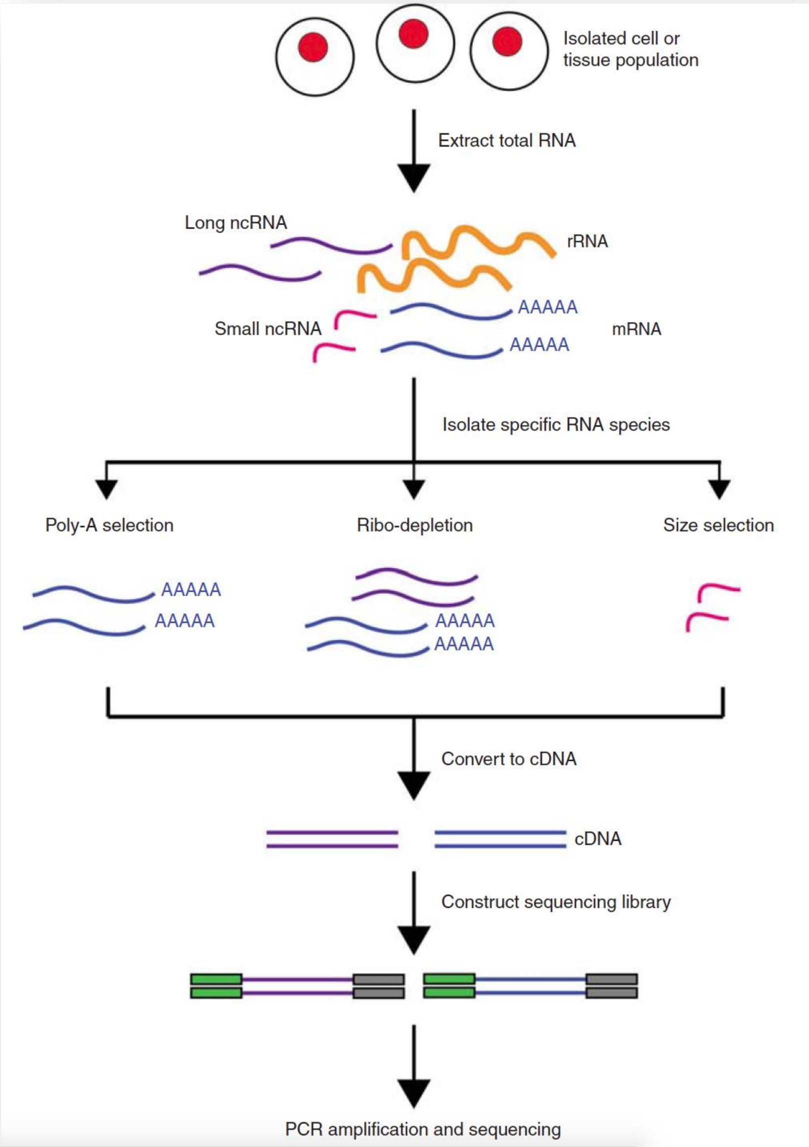


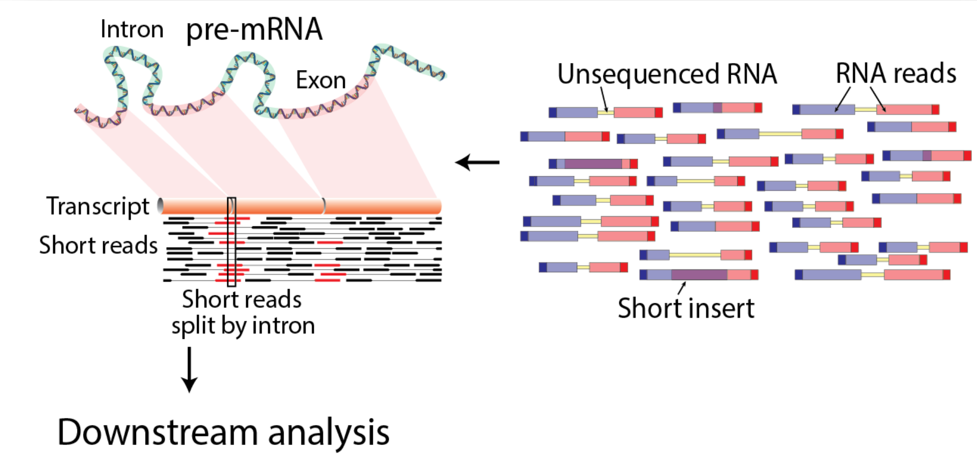
Sigurgeirsson B, Emanuelsson O, Lundeberg J. Sequencing degraded RNA addressed by 3' tag counting. PLoS One. 2014 Mar 14;9(3):e91851.

slides Outline

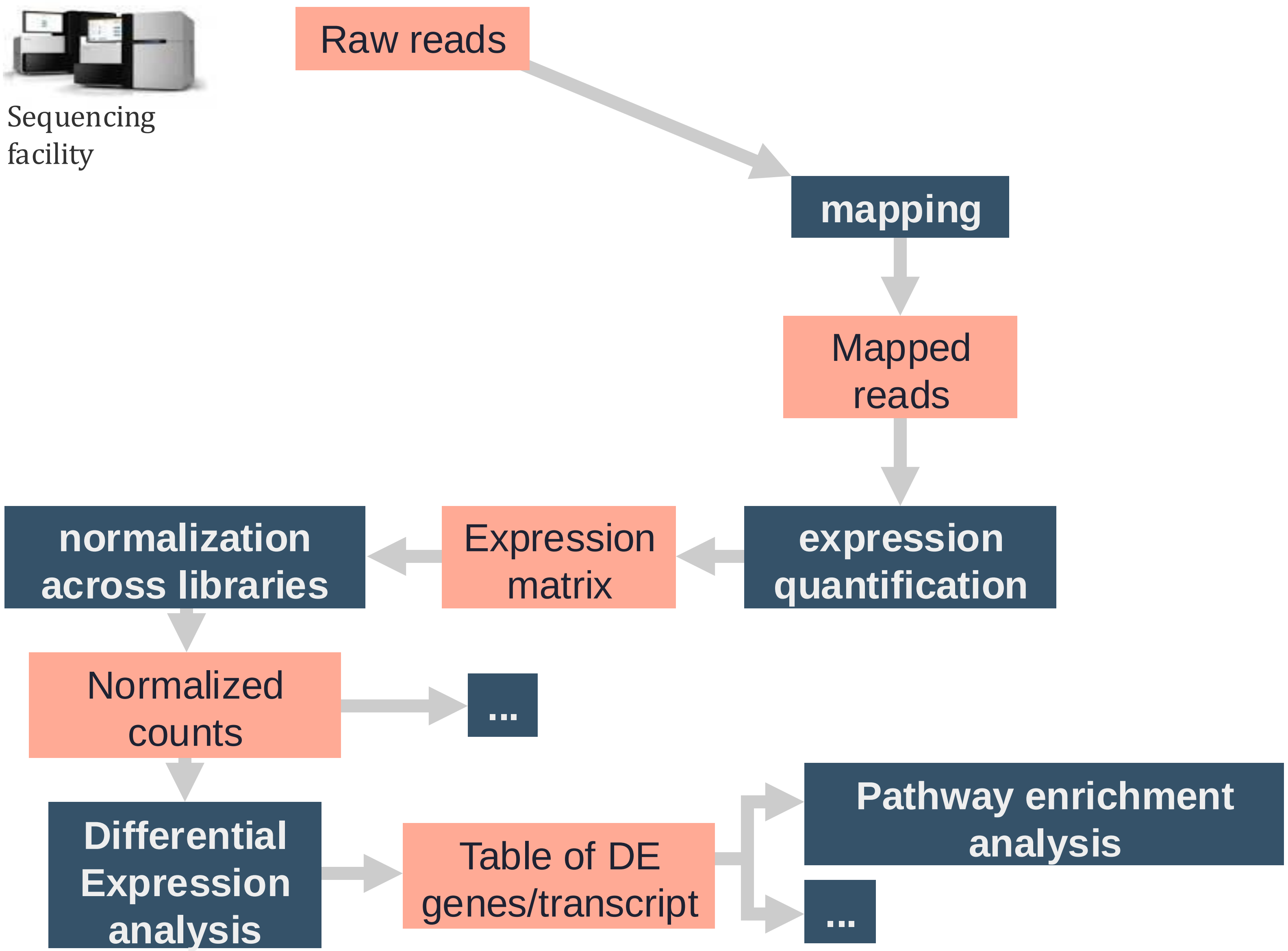
- RNA and molecular biology
- Main challenges for RNAseq
- Major Sequencing technologies
- Planning your sequencing : choices, number of samples, ...
- **Bioinformatics analysis overview**

Basic RNAseq protocol overview





RNAseq data analysis - basic pipeline



Sequencing facility

Raw reads

mapping

Mapped reads

expression quantification

Expression matrix

normalization across libraries

Normalized counts

...

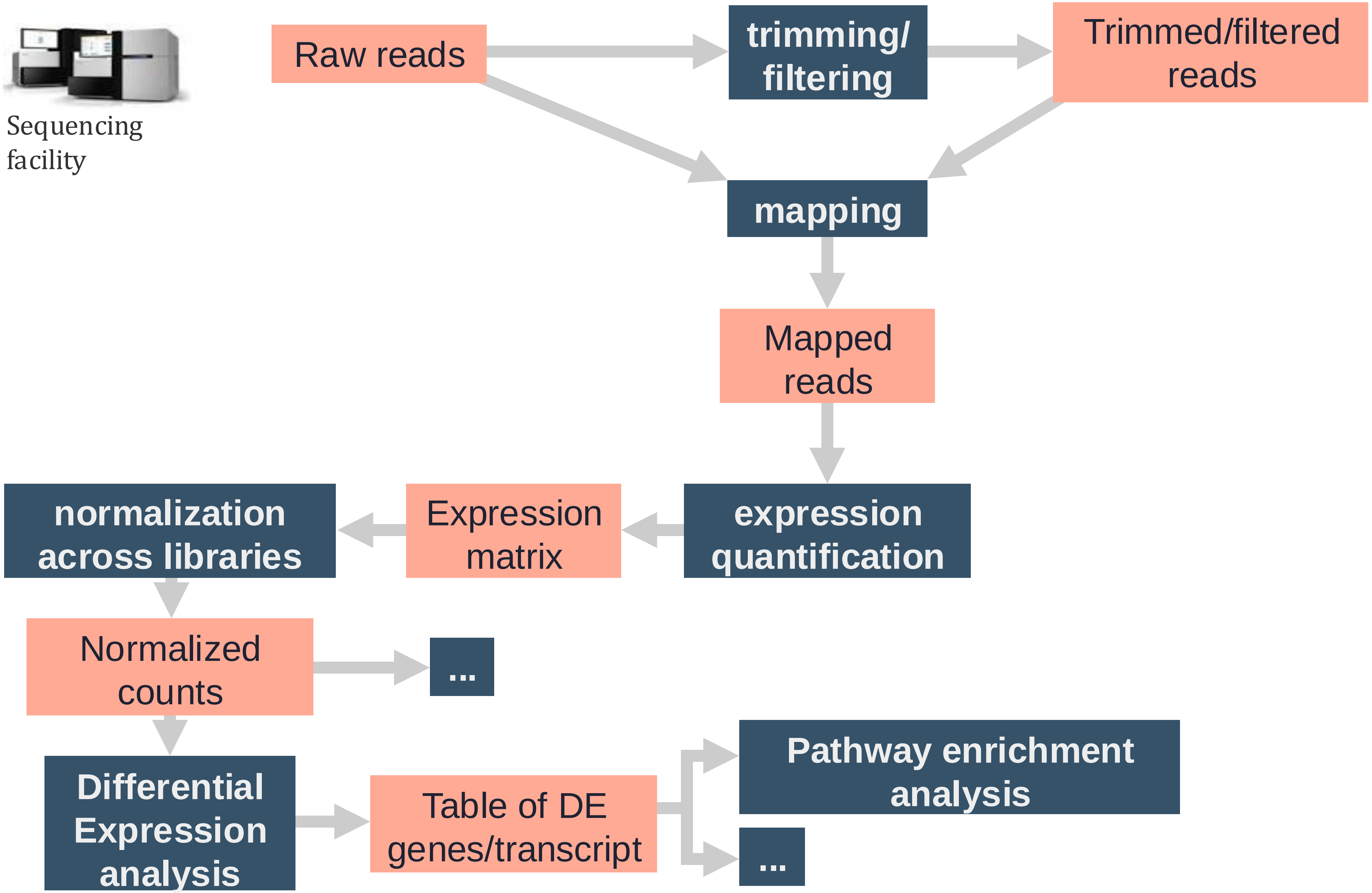
Differential Expression analysis

Table of DE genes/transcript

Pathway enrichment analysis

...

RNAseq data analysis - basic pipeline



Sequencing facility

Raw reads

trimming/ filtering

Trimmed/filtered reads

mapping

Mapped reads

expression quantification

Expression matrix

normalization across libraries

Normalized counts

...

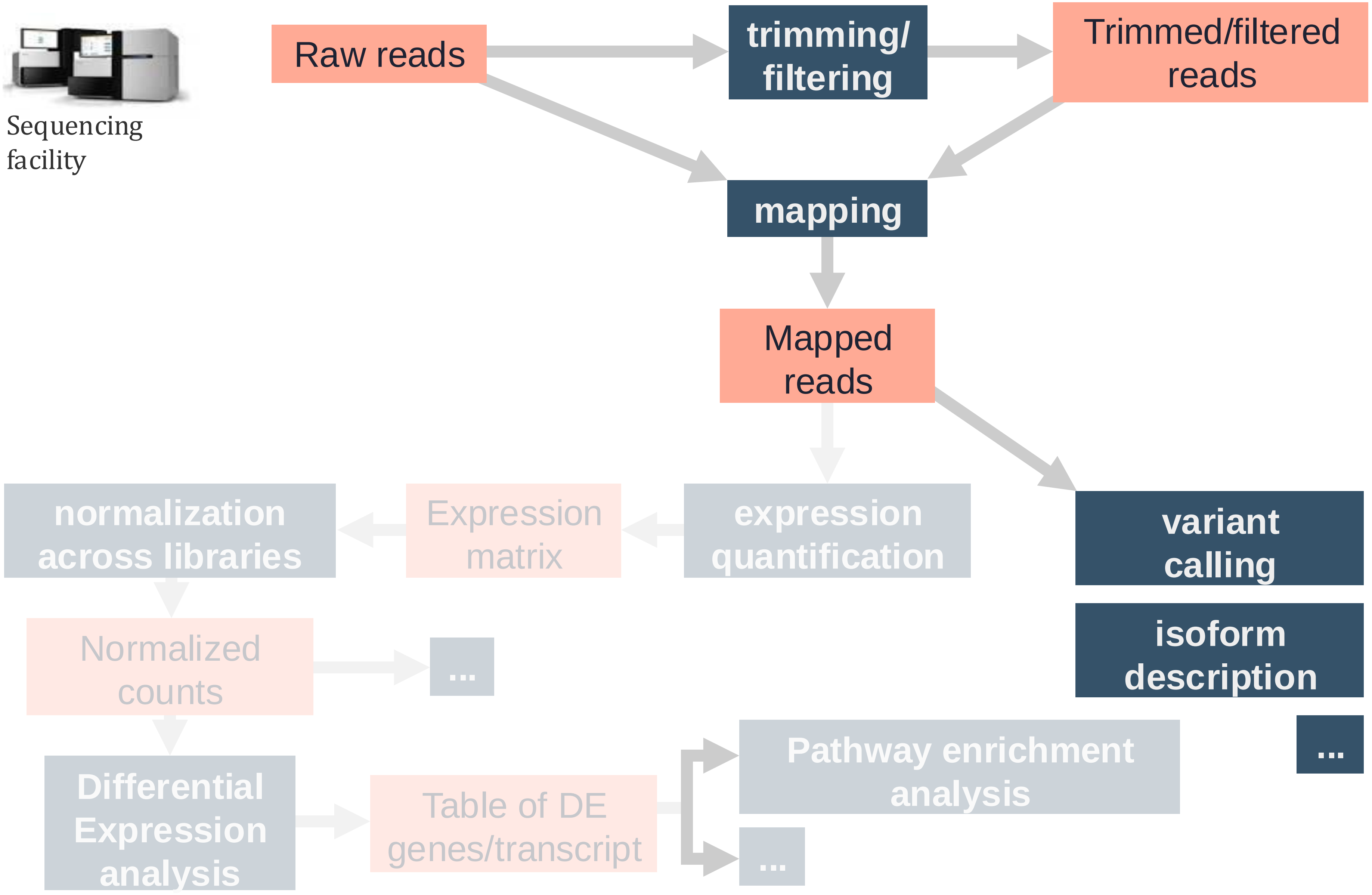
Differential Expression analysis

Table of DE genes/transcript

Pathway enrichment analysis

...

RNAseq data analysis - basic pipeline



Sequencing facility

Raw reads

trimming/ filtering

Trimmed/filtered reads

mapping

Mapped reads

expression quantification

Expression matrix

normalization across libraries

Normalized counts

...

Differential Expression analysis

Table of DE genes/transcript

Pathway enrichment analysis

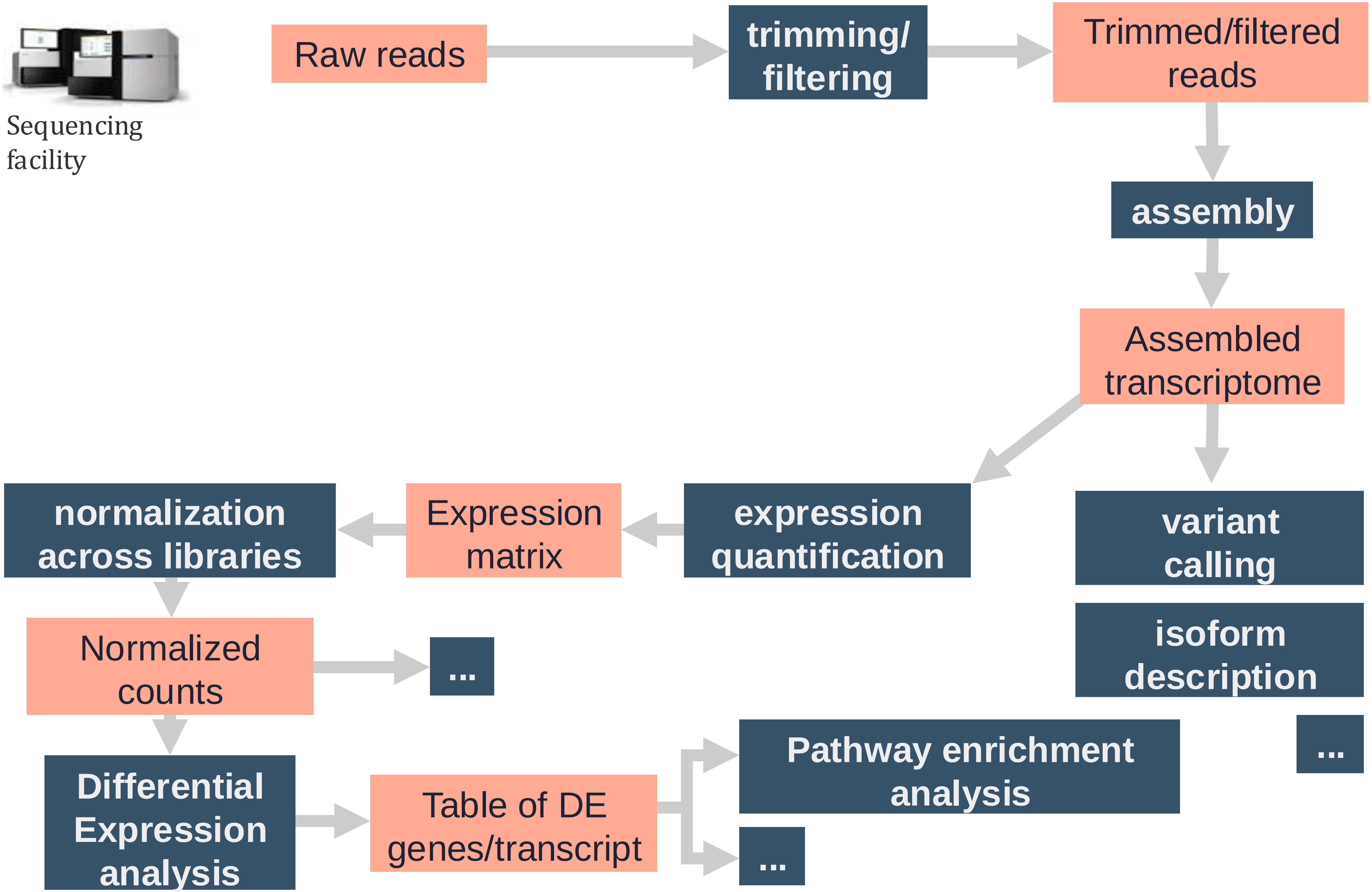
...

variant calling

isoform description

...

RNAseq data analysis - basic pipeline



Sequencing facility

Raw reads

trimming/ filtering

Trimmed/filtered reads

assembly

Assembled transcriptome

expression quantification

variant calling

isoform description

...

Expression matrix

normalization across libraries

Normalized counts

...

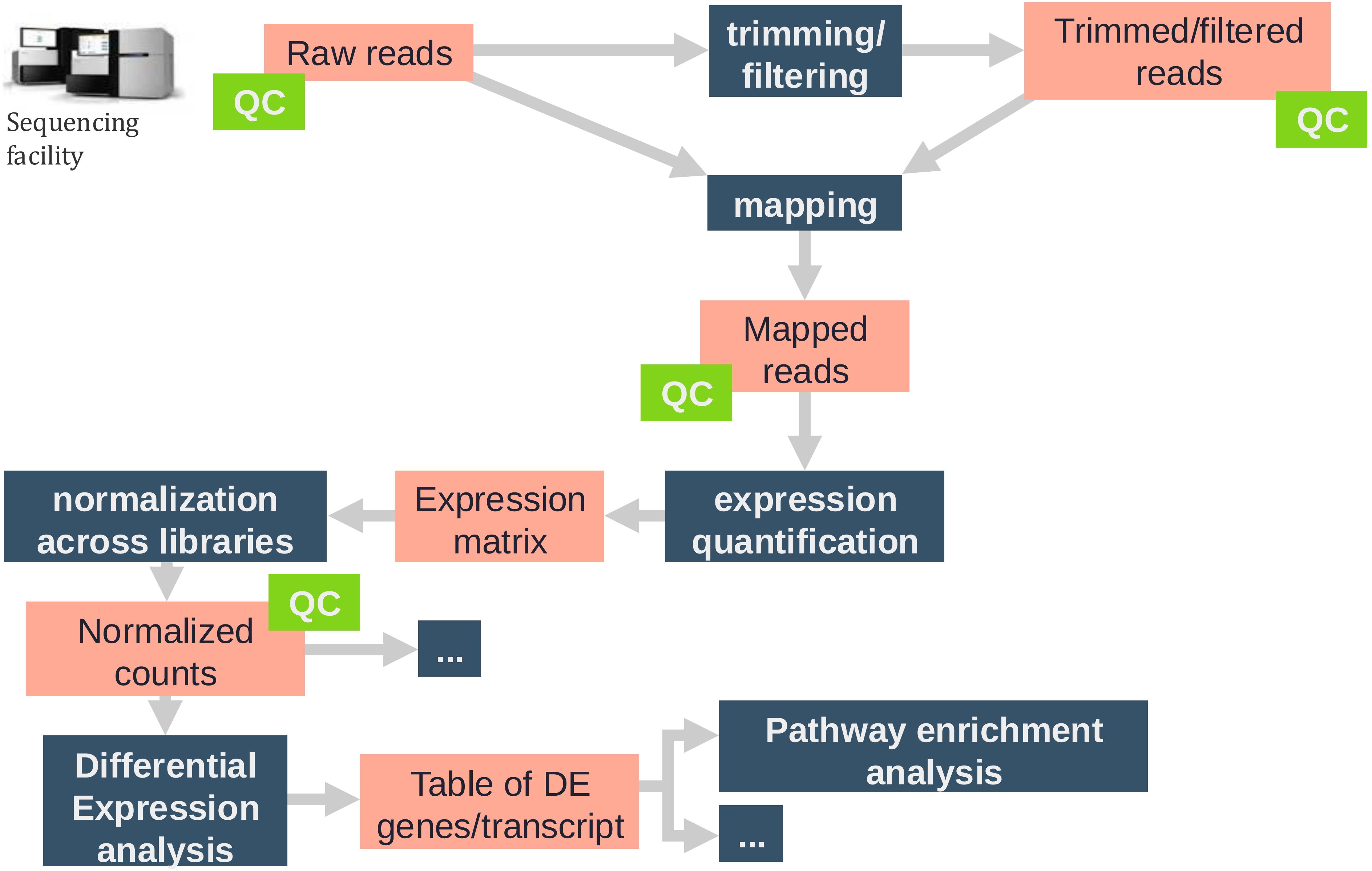
Differential Expression analysis

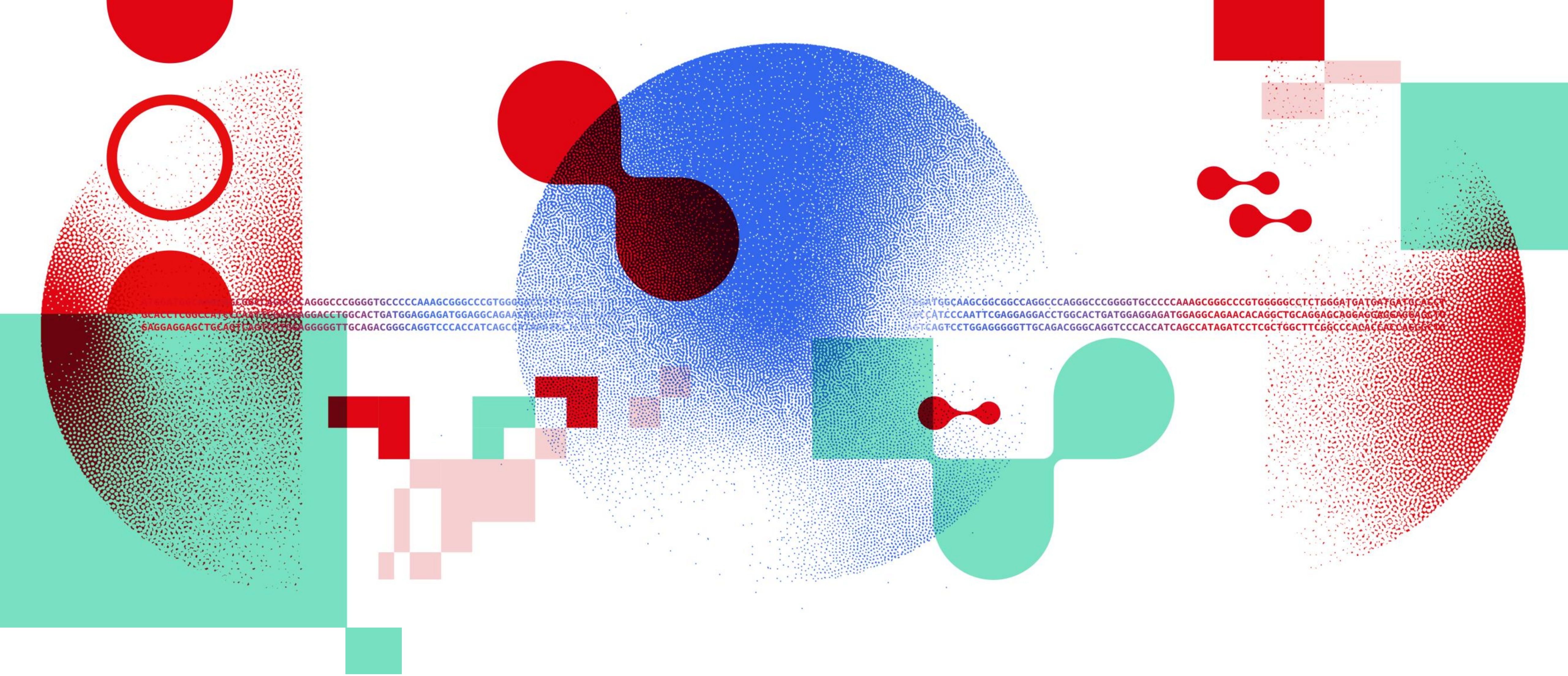
Table of DE genes/transcript

Pathway enrichment analysis

...

RNAseq data analysis - basic pipeline


Sequencing facility
Raw reads
QC
trimming/ filtering
Trimmed/filtered reads
QC
mapping
Mapped reads
QC
expression quantification
Expression matrix
normalization across libraries
Normalized counts
QC
...
Differential Expression analysis
Table of DE genes/transcript
Pathway enrichment analysis
...



Thank you