

SIB  
Swiss Institute of  
Bioinformatics

# Introduction to RNA-Seq – Overview

Wandrille Duchemin

# First some logistics

---

## Internet access :

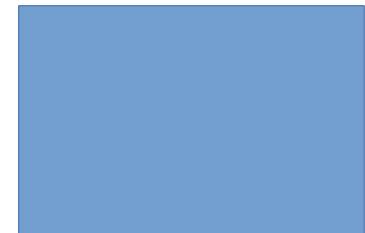
- eduroam
  - unibas-visitor > login page > get code with your phone
- 
- Course from 09:00 to 17:00
  - Lunch break 12:00 → 13:00
  - 15min breaks around 10:30 and 15:00

## Questions are welcome all the time

During practicals: post-its



I have a question/  
problem



I am finished with  
the practical

DNA



Transcription

Nucleic acids  
everywhere

RNA

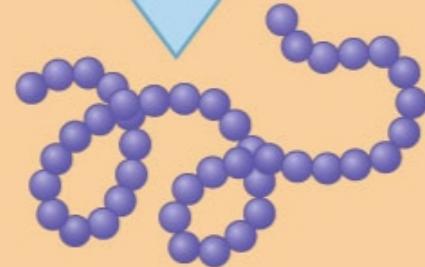


NUCLEUS

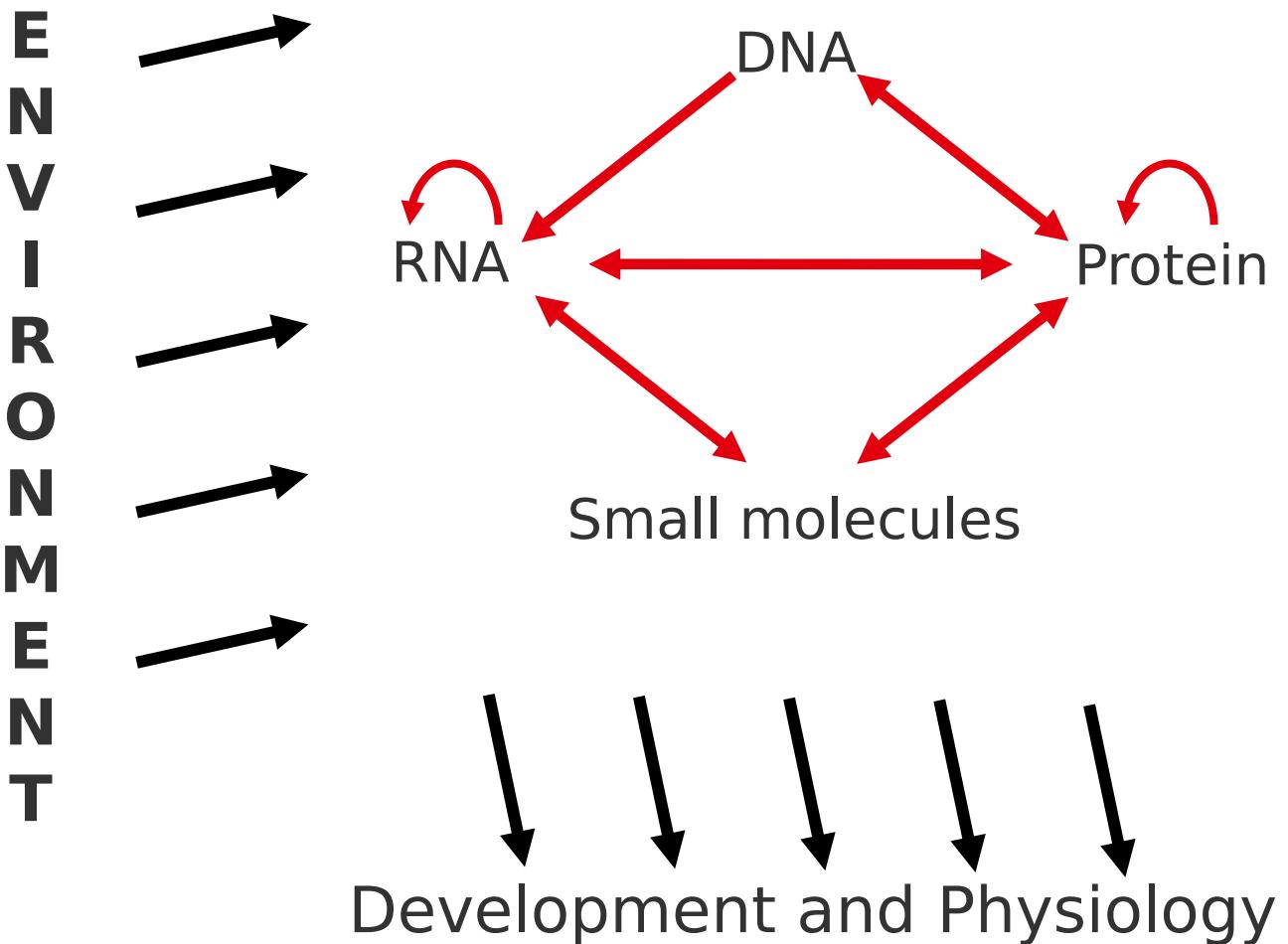
Protein

Translation

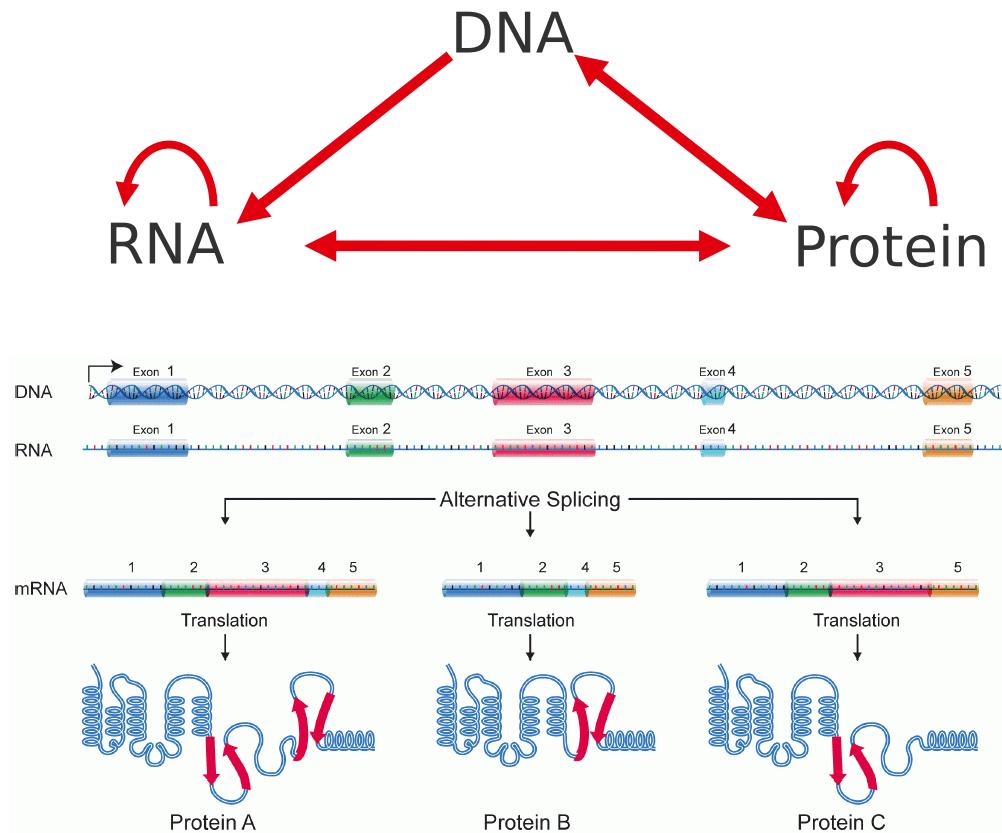
CYTOPLASM



# A more complex model complex interactions and regulations



# A more complex model alternative splicing



[https://en.wikipedia.org/wiki/Alternative\\_splicing#/media/File:DNA\\_alternative\\_splicing.gif](https://en.wikipedia.org/wiki/Alternative_splicing#/media/File:DNA_alternative_splicing.gif)

# A more complex model alternative splicing

~20,000 mammalian genes



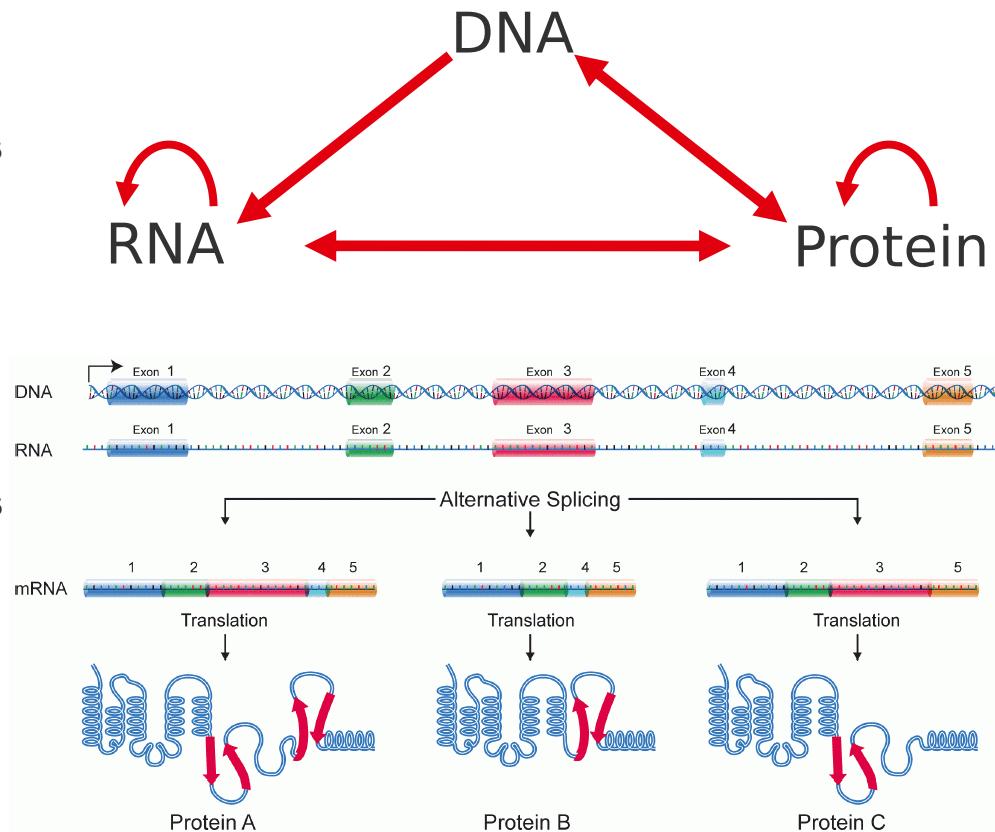
>>100,000 (?) transcripts



>>>1,000,000 (??) proteins



Cellular Diversity

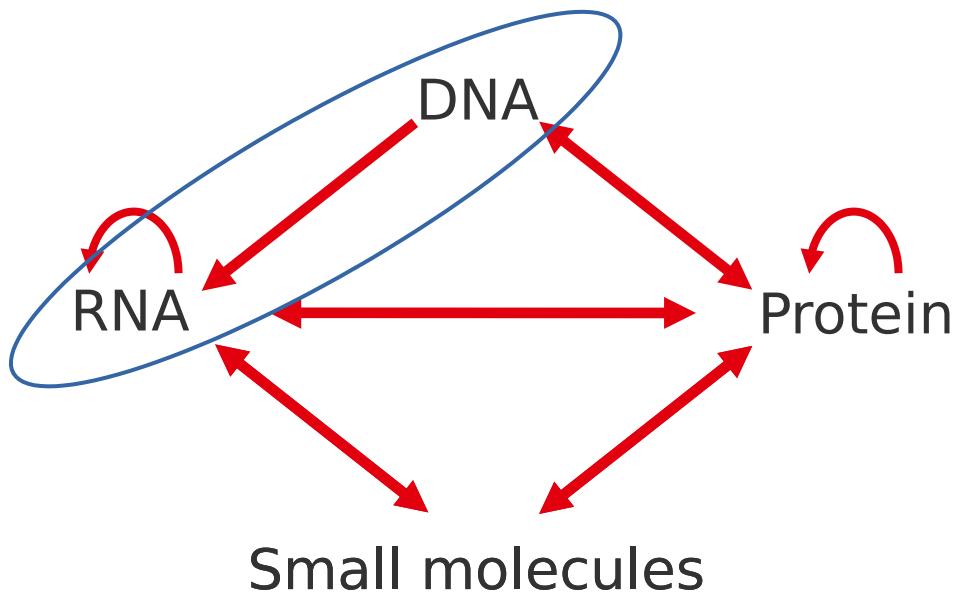


# A more complex model

- sequencing nucleic acids

**describe or  
quantify**

mRNA levels :  
proxy of a proxy  
of cell physiology



Development and Physiology

# What (and why) are we sequencing?

---

## ■ Genomics

- Whole genome/exome sequencing (WGS, WES)
  - SNPs / CNVs / Structural variations

## ■ Epigenomics

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

## ■ Transcriptomics

- Total RNA
- PolyA-tail selection : focus on mRNA
- Ribo depletion : mRNA + non-coding RNAs
- 5'/3' RACE seq : focus on a single gene's isoforms
- Single cell sequencing
- ...

# What (and why) are we sequencing?

## ■ Genomics

- Whole genome/exome sequencing (WGS, WES)
  - SNPs / CNVs / Structural variations

## ■ Epigenomics

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

Imagination is the limit!

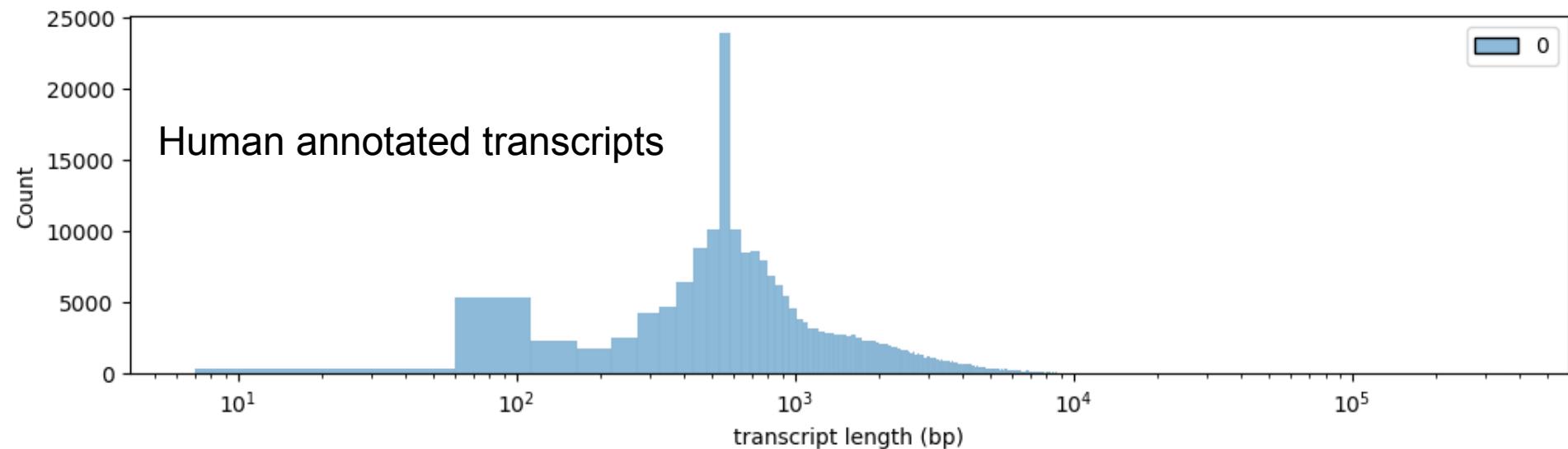
## ■ Transcriptomics

- Total RNA
- PolyA-tail selection : focus on mRNA
- Ribo depletion : mRNA + non-coding RNAs
- 5'/3' RACE seq : focus on a single gene's isoforms
- Single cell sequencing
- ...

A longer but non exhaustive list :  
<https://liorpachter.wordpress.com/seq/>

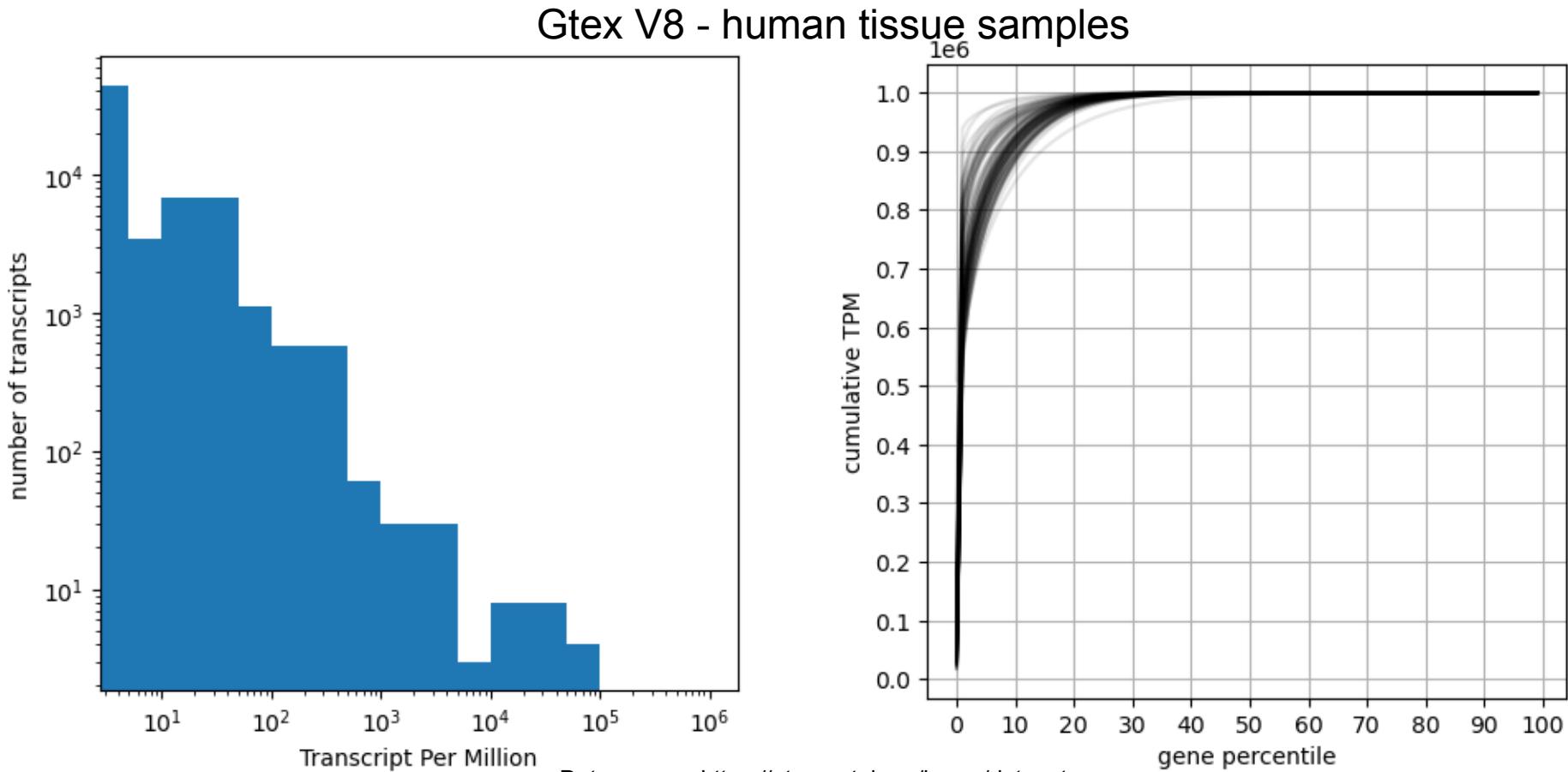
# Some challenges for RNA-Seq

- Transcripts are diverse in size



# Some challenges for RNA-Seq

- Transcripts are diverse in size
- Expression levels have a high dynamic range



# Some challenges for RNA-Seq

---

- Transcripts are diverse in size
- Expression levels have a high dynamic range
- RNA molecules are exposed to degradation enzymes
  - RNA integrity affects results
- Is there a reference genome?
  - Yes : how good?
  - No : transcriptome assembly
- How good is gene annotation?

# Main sequencing technologies

illumina



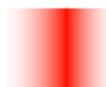
100M to 3B reads

PacBio



4M CCS reads

iontorrent



250M reads

Oxford  
**NANOPORE**  
Technologies



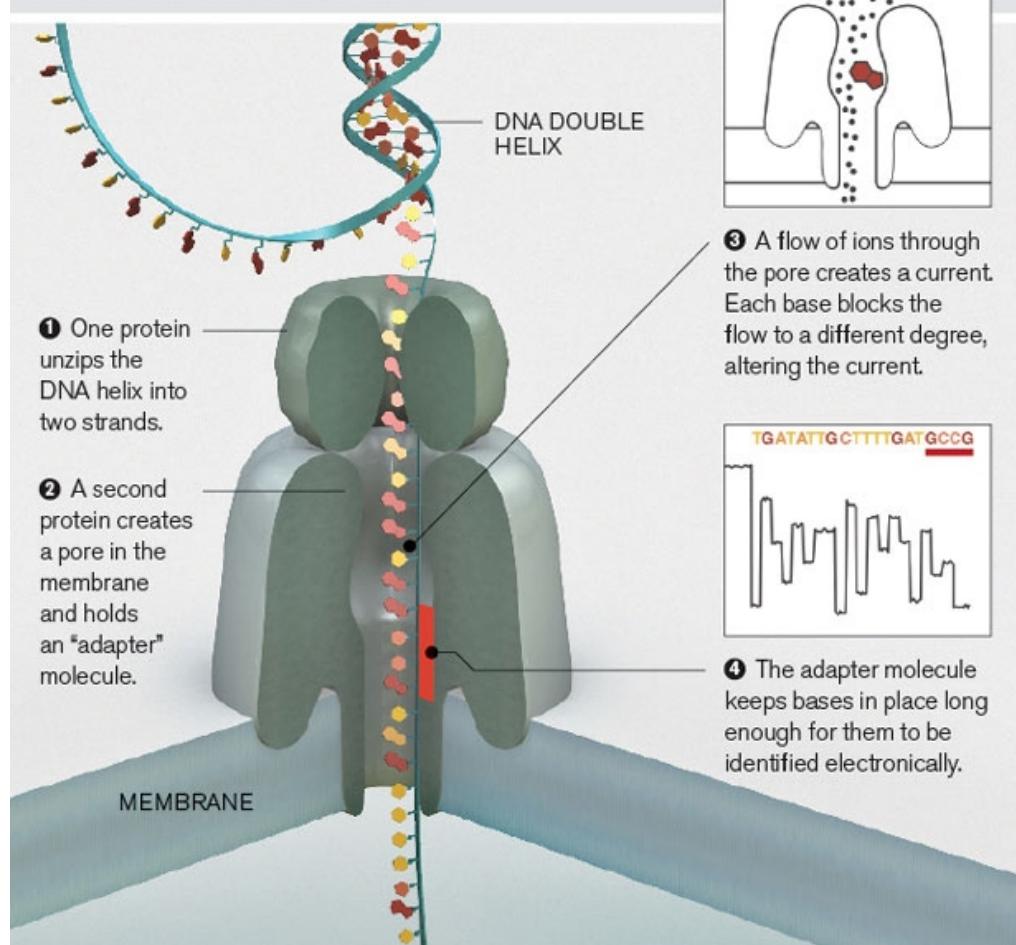
...

enough for  
about 25 Gbases

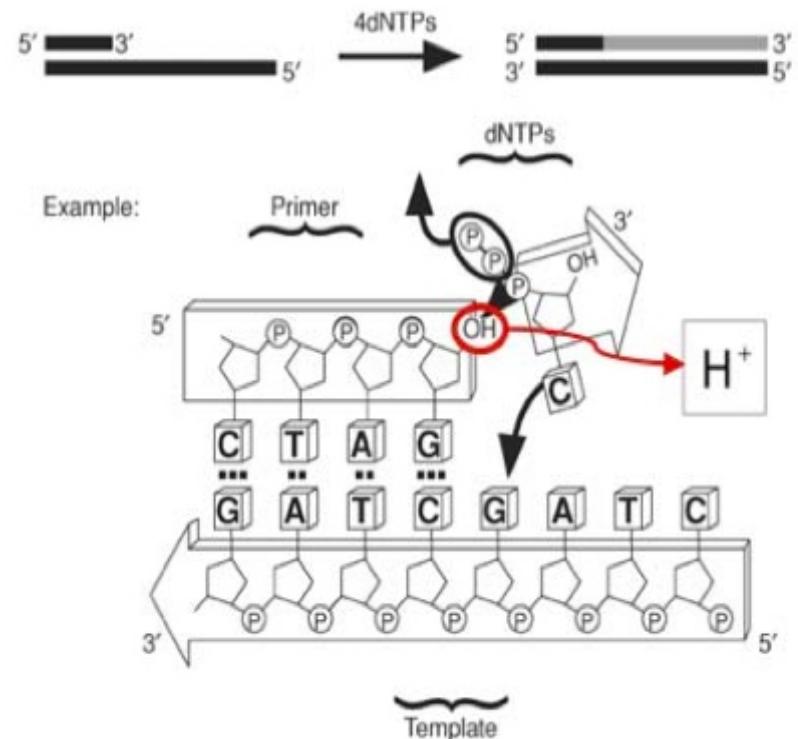
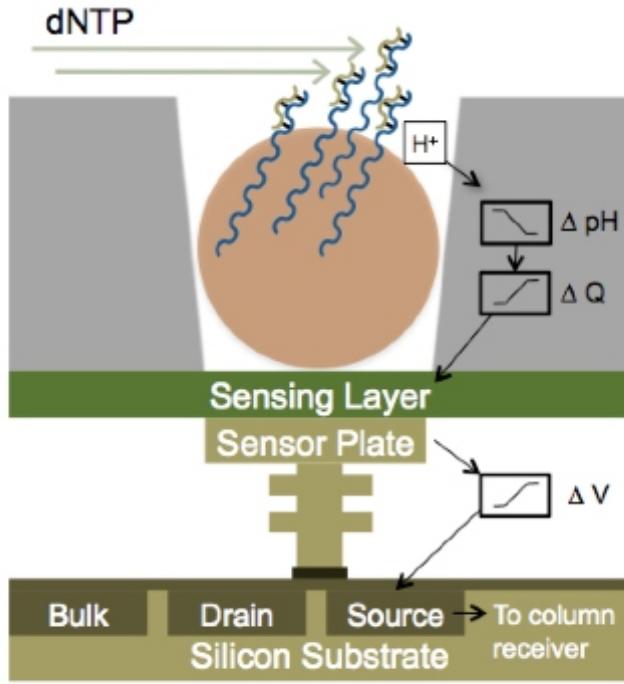


# Oxford Nanopore – direct DNA/RNA sequencing

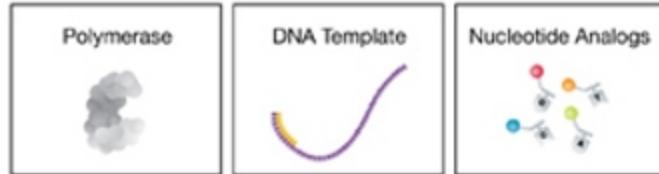
DNA can be sequenced by threading it through a microscopic pore in a membrane. Bases are identified by the way they affect ions flowing through the pore from one side of the membrane to the other.



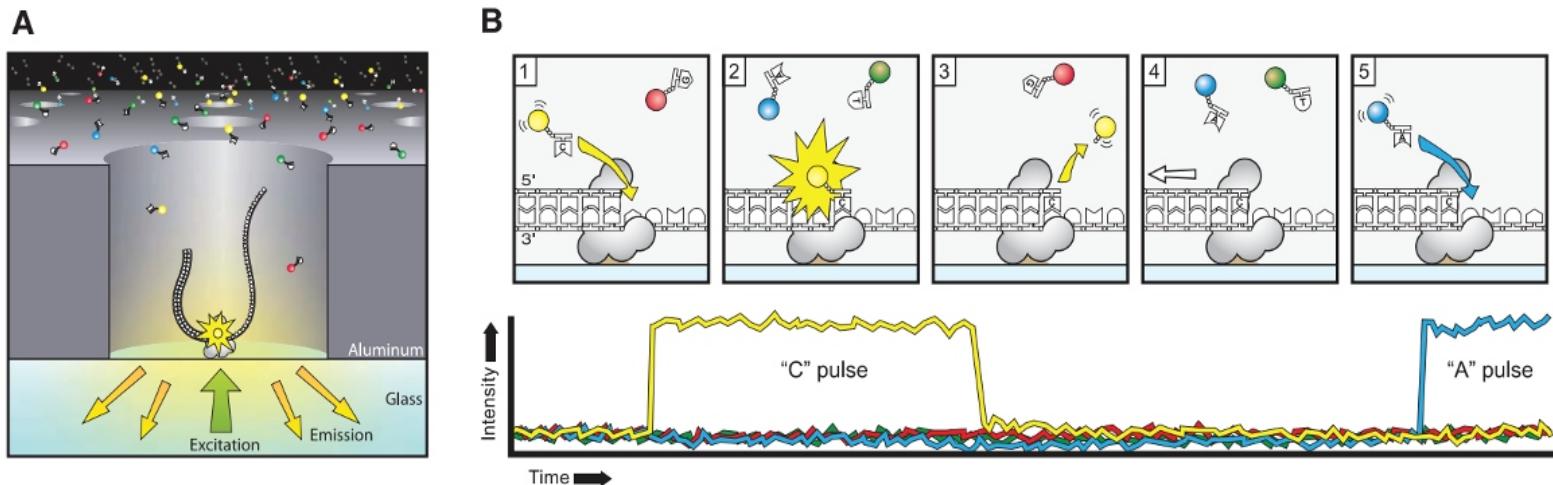
# Ion torrent – reading pH changes



# Pacific Biosciences – Single Molecule Real Time sequencing



Rhoads & Au.  
Genomics Proteomics Bioinformatics 2015



# Pacific Biosciences – Circular Consensus Sequencing

Raw reads : 15% **random** errors

→ CCS strategy

Start with high-quality double stranded DNA



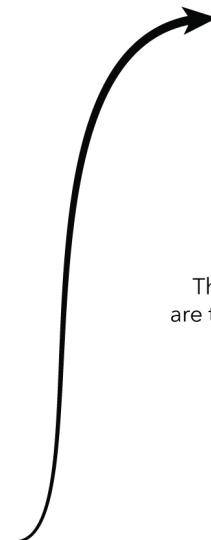
Ligate SMRTbell adapters and size select



Anneal primers and bind DNA polymerase

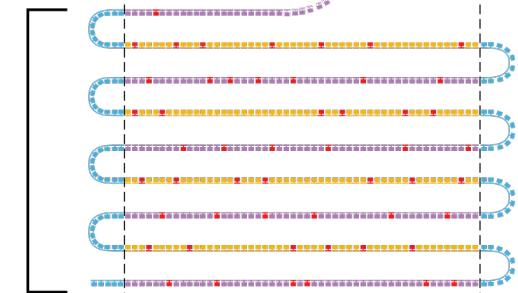
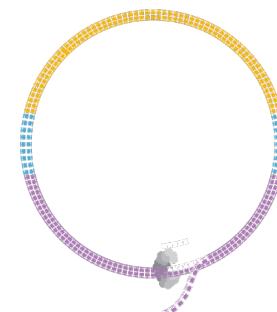


Circularized DNA is sequenced in repeated passes



The polymerase reads are trimmed of adapters to yield subreads

Consensus is called from subreads



**HiFi READ**  
(>99% accuracy)

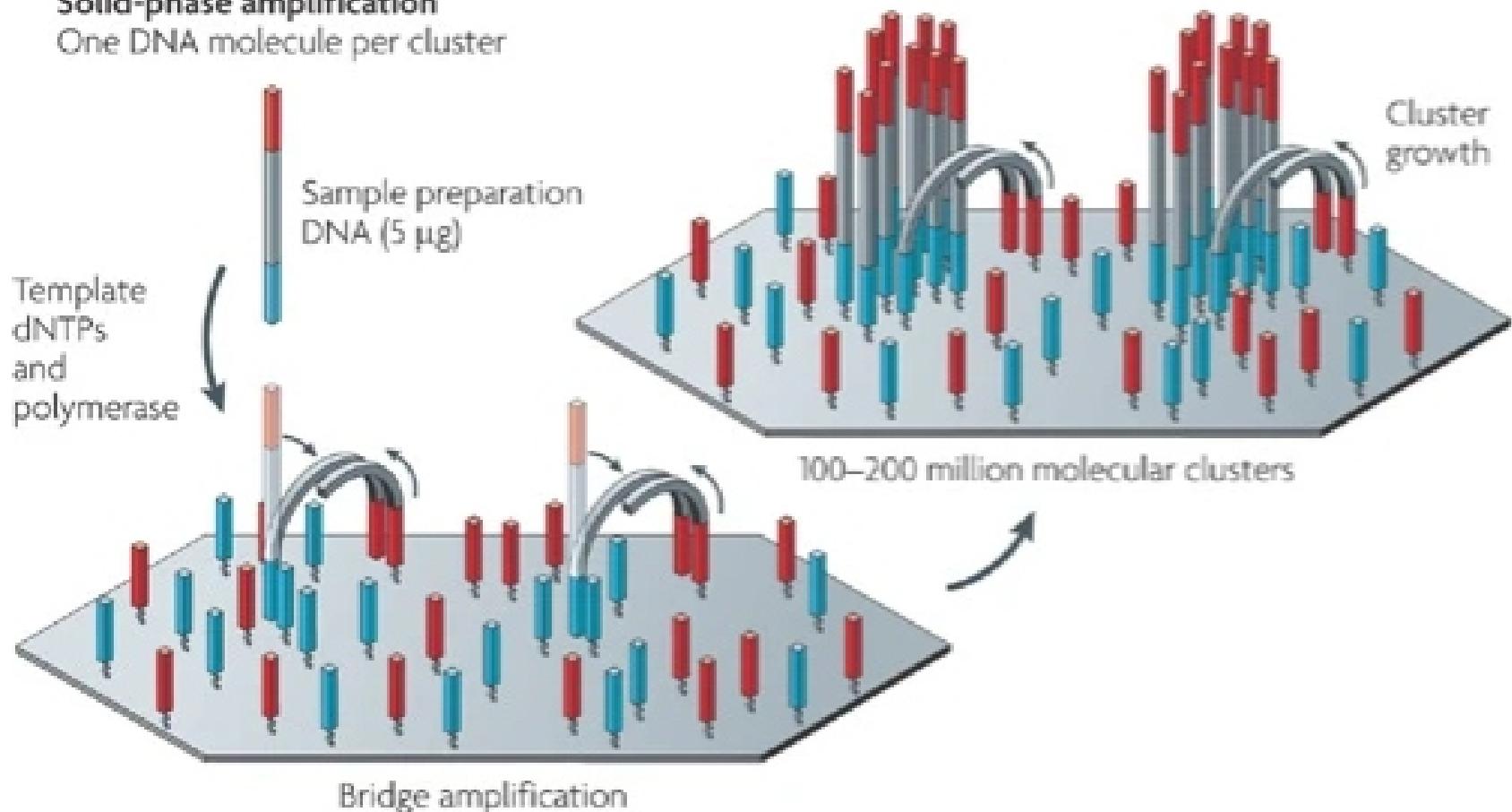
Typical in isoseq

# Illumina sequencing – “cluster formation”

## b Illumina/Solexa

### Solid-phase amplification

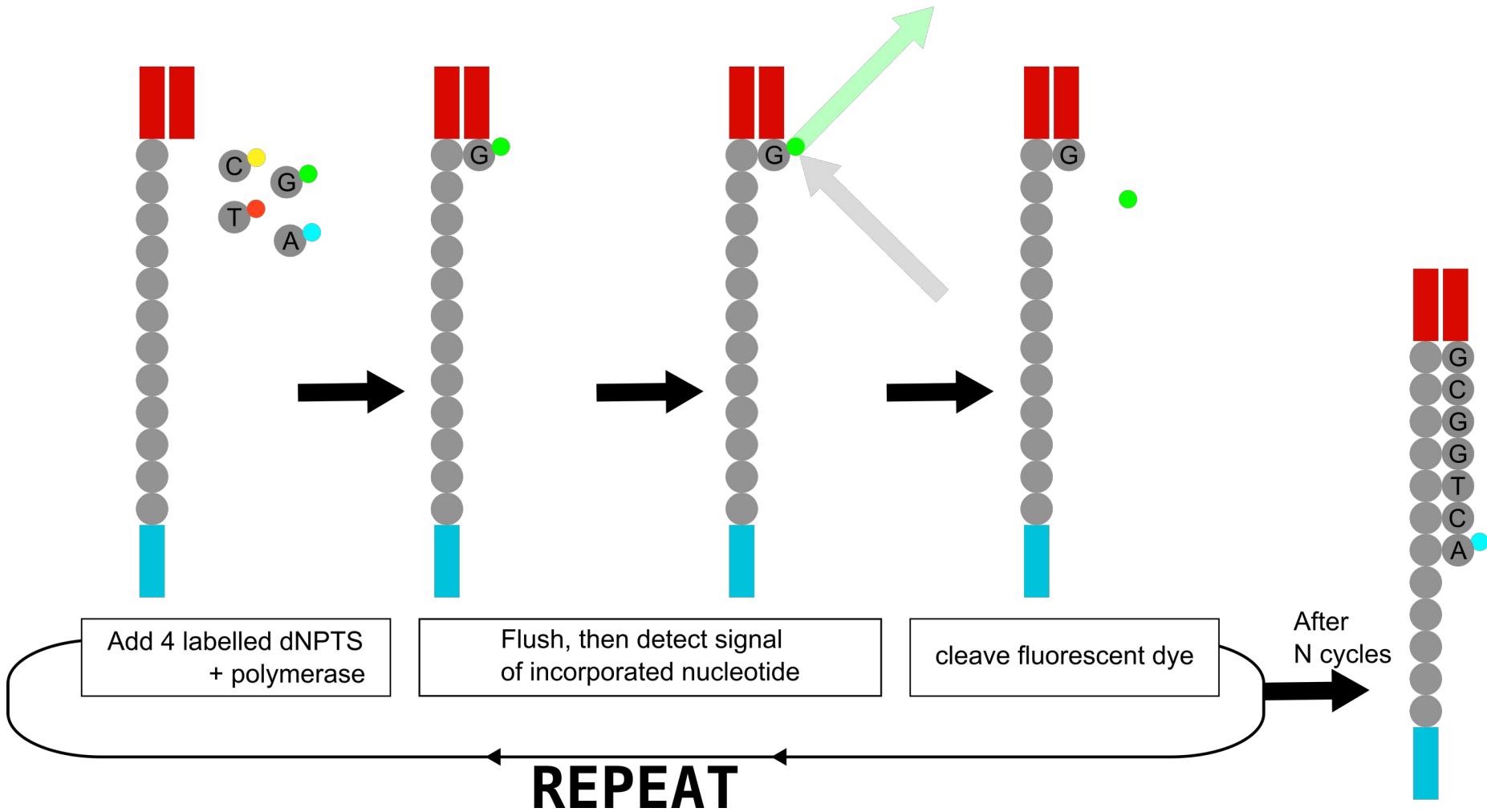
One DNA molecule per cluster



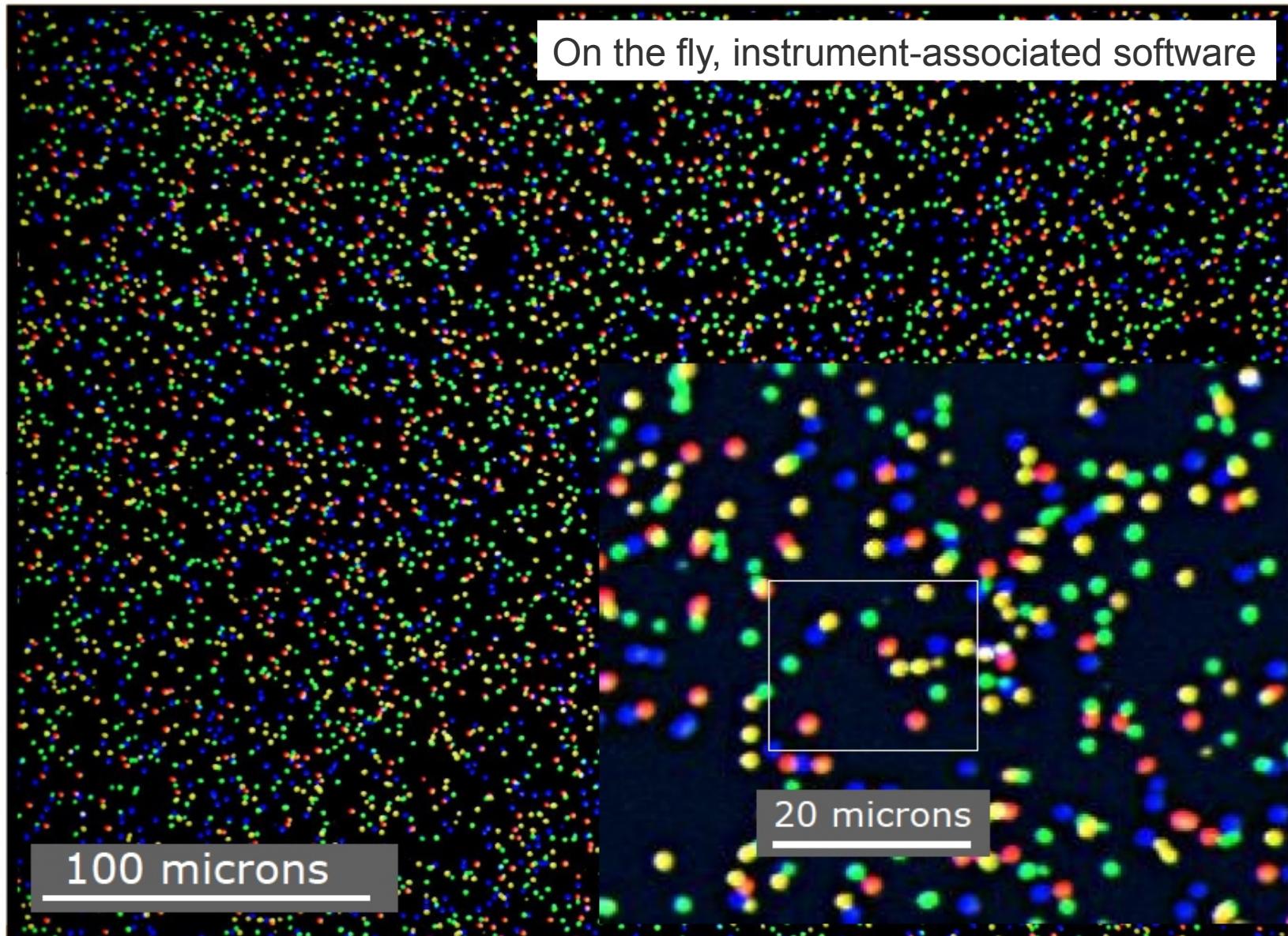
Source: Fig1b of Metzker, M. Sequencing technologies — the next generation.

Nat Rev Genet 11, 31–46 (2010).

# Illumina sequencing – “sequencing by synthesis”

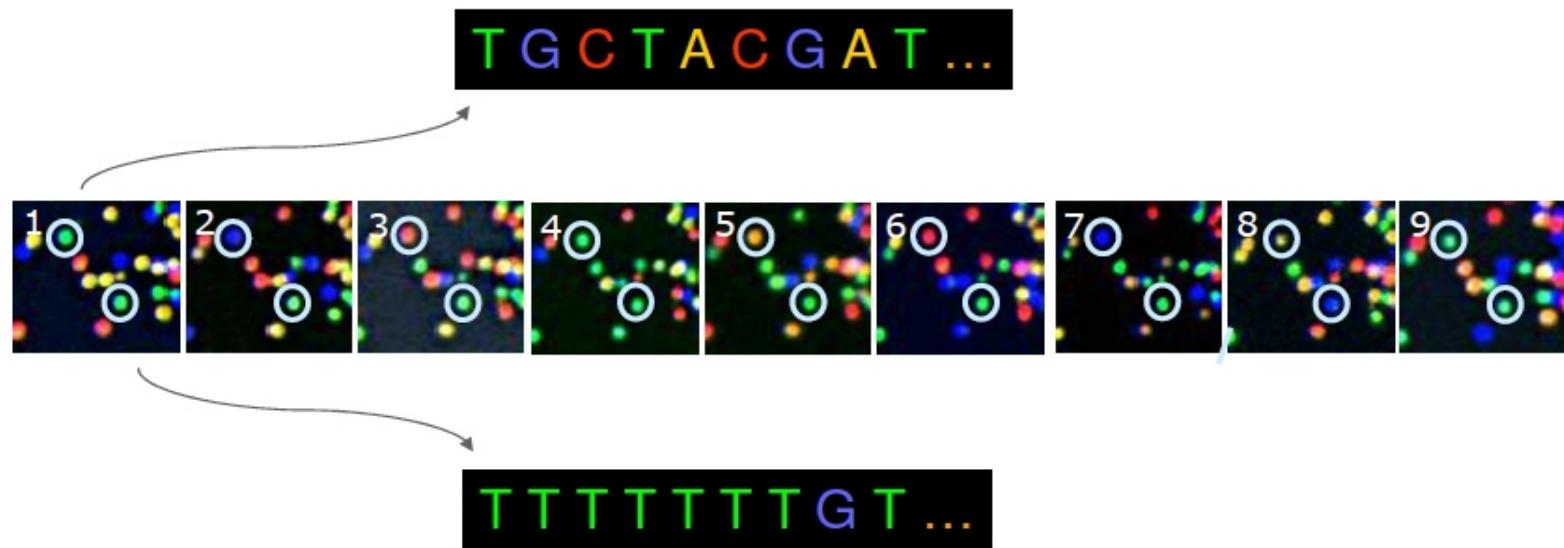


# Illumina sequencing – image analysis



# Illumina sequencing – from image to sequence

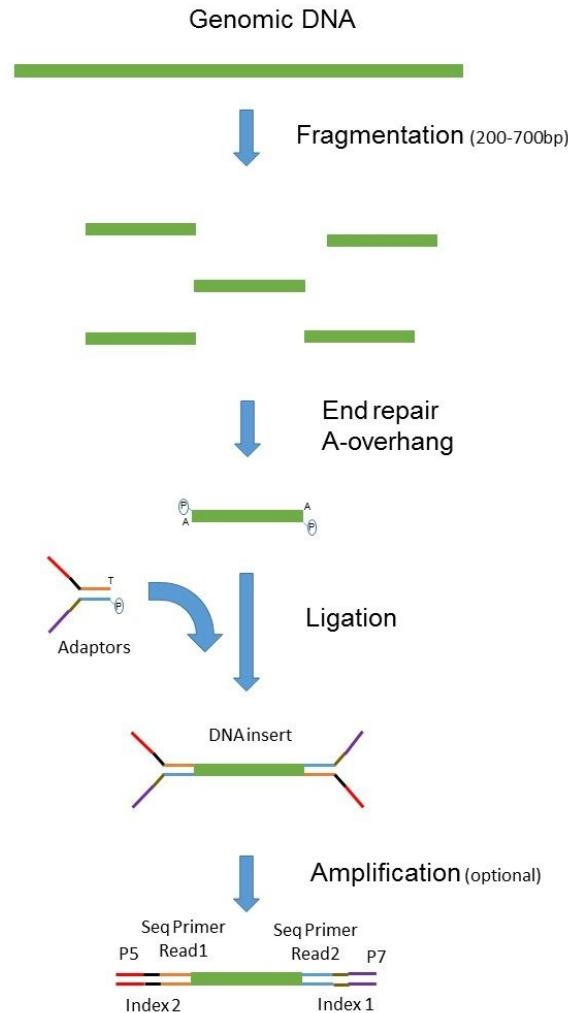
## Base Calling From Raw Data



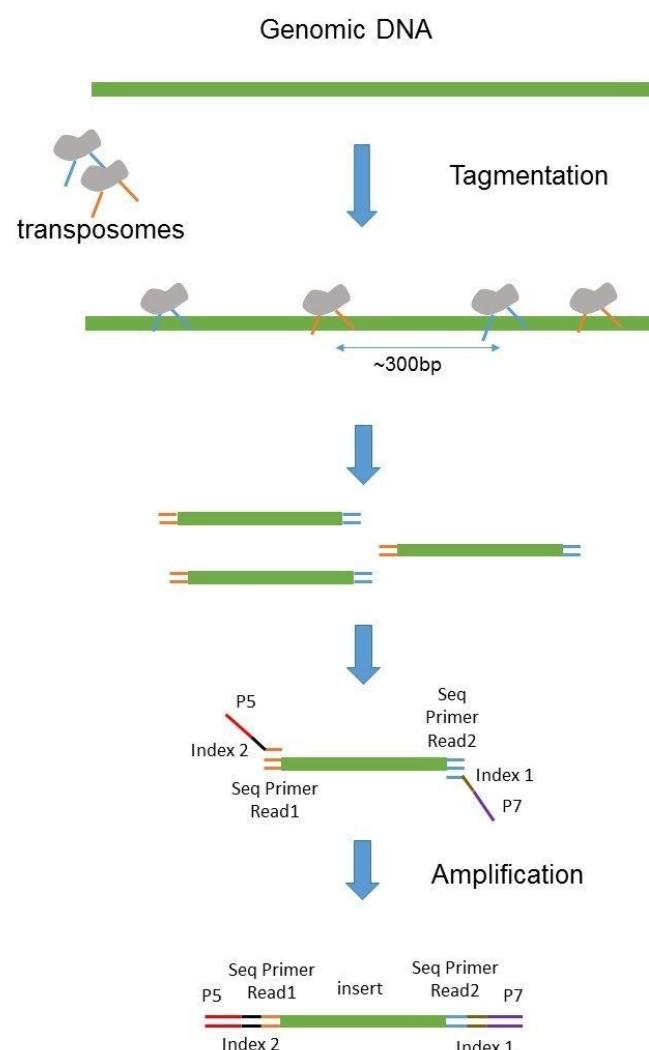
The identity of each base of a cluster is read off from sequential images

# Paired-end sequencing

“Classical” paired end library (illumina)

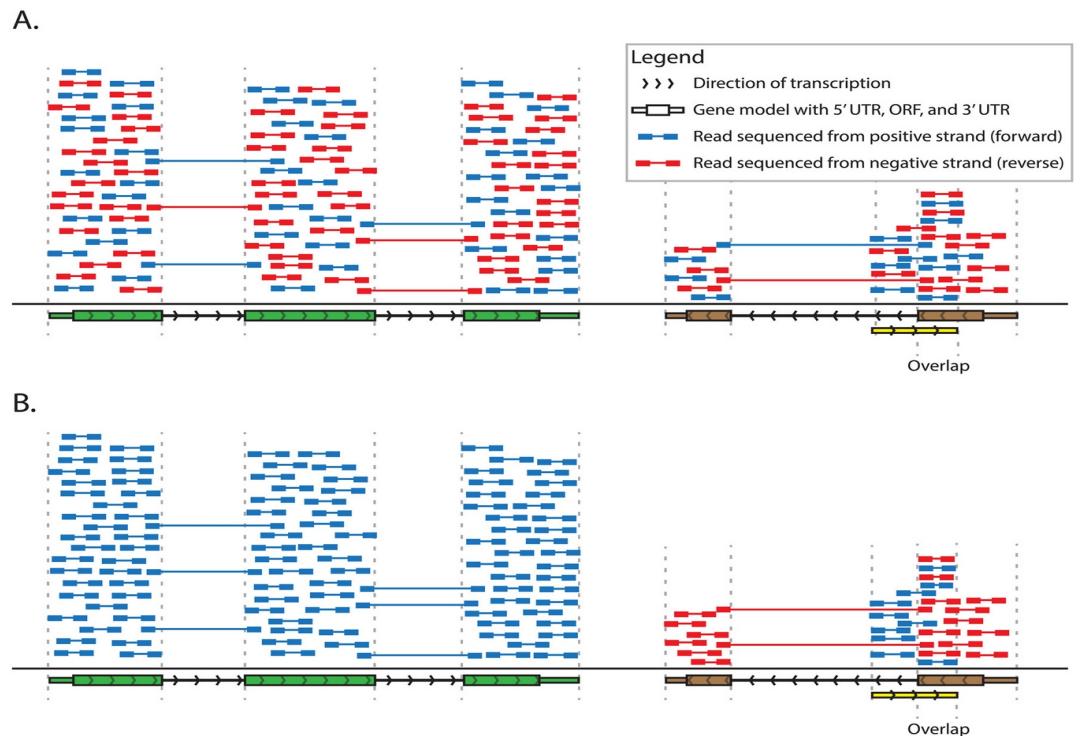


“NexTera” paired end library



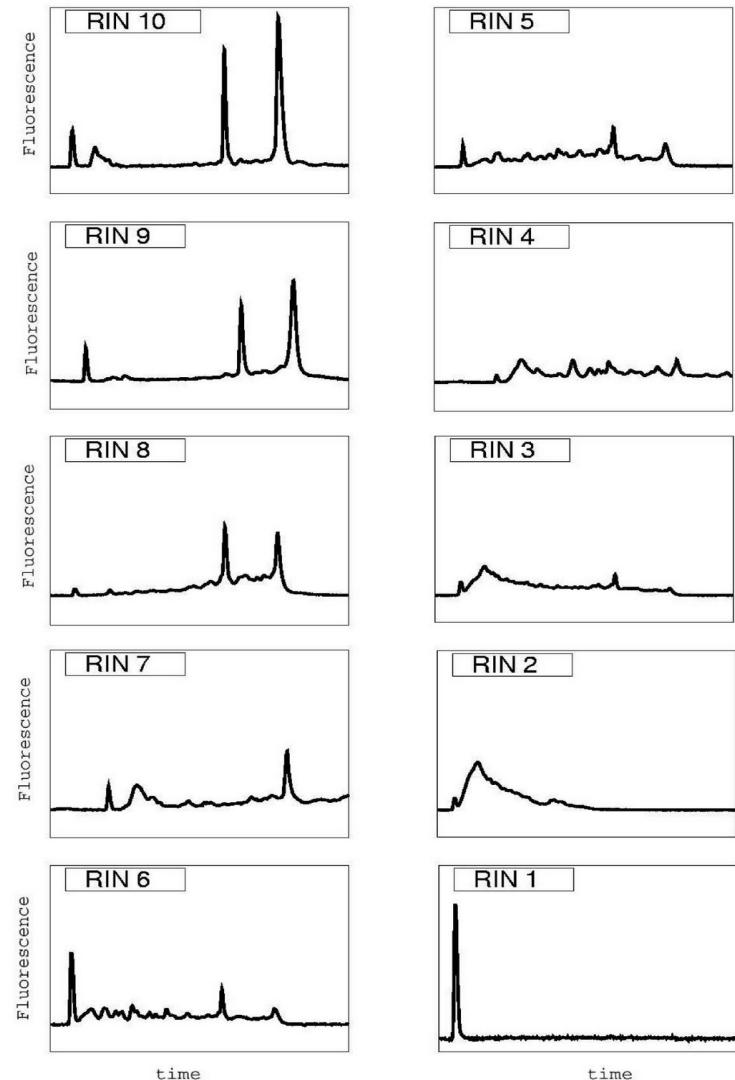
# Stranded vs Unstranded Sequencing

- A substantial proportion of the genome is transcribed in both directions (~8% in *Homo sapiens*)
- Read strand information = quantify expression on overlaps
- Achieved by ligating different adaptors to the 5' and 3' ends



# RNA Sample Preparation – RNA isolation

- Critically important → cannot make up for poor data!
- RNA integrity number (RIN) estimates integrity based on the Agilent profile
- Minimum:
  - 7 or 8 (eukaryot mRNA)
  - 9 (bacterial )



# RNA-Seq Preparation - Purification

---

## ■ PolyA selection

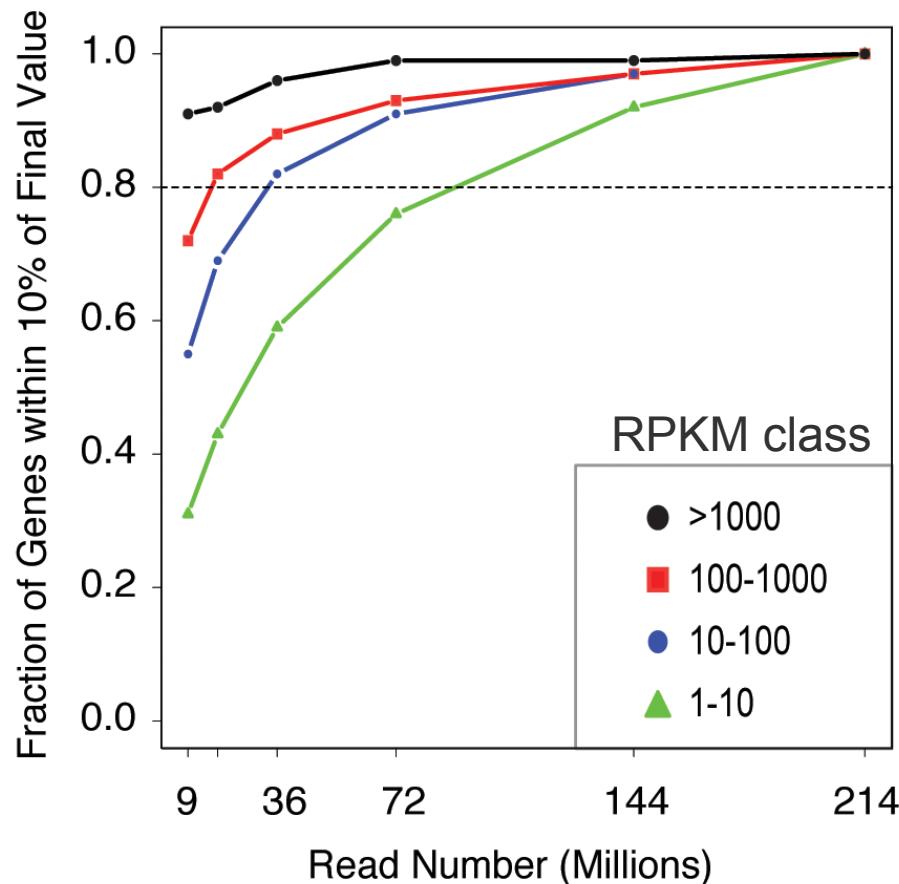
- Commonly used and inexpensive
- 3' end bias when RNA is degraded
- Loses almost all non-polyA transcripts
- Gets rid of vast majority of ribosomal RNAs

## ■ Ribosomal RNA depletion

- Less popular, ~2X more expensive
- Higher proportion of rRNA compared to polyA selection
- Bacterial data (no polyA tails)
- Allows identification of lncRNAs without polyA tails

# Sequencing depth

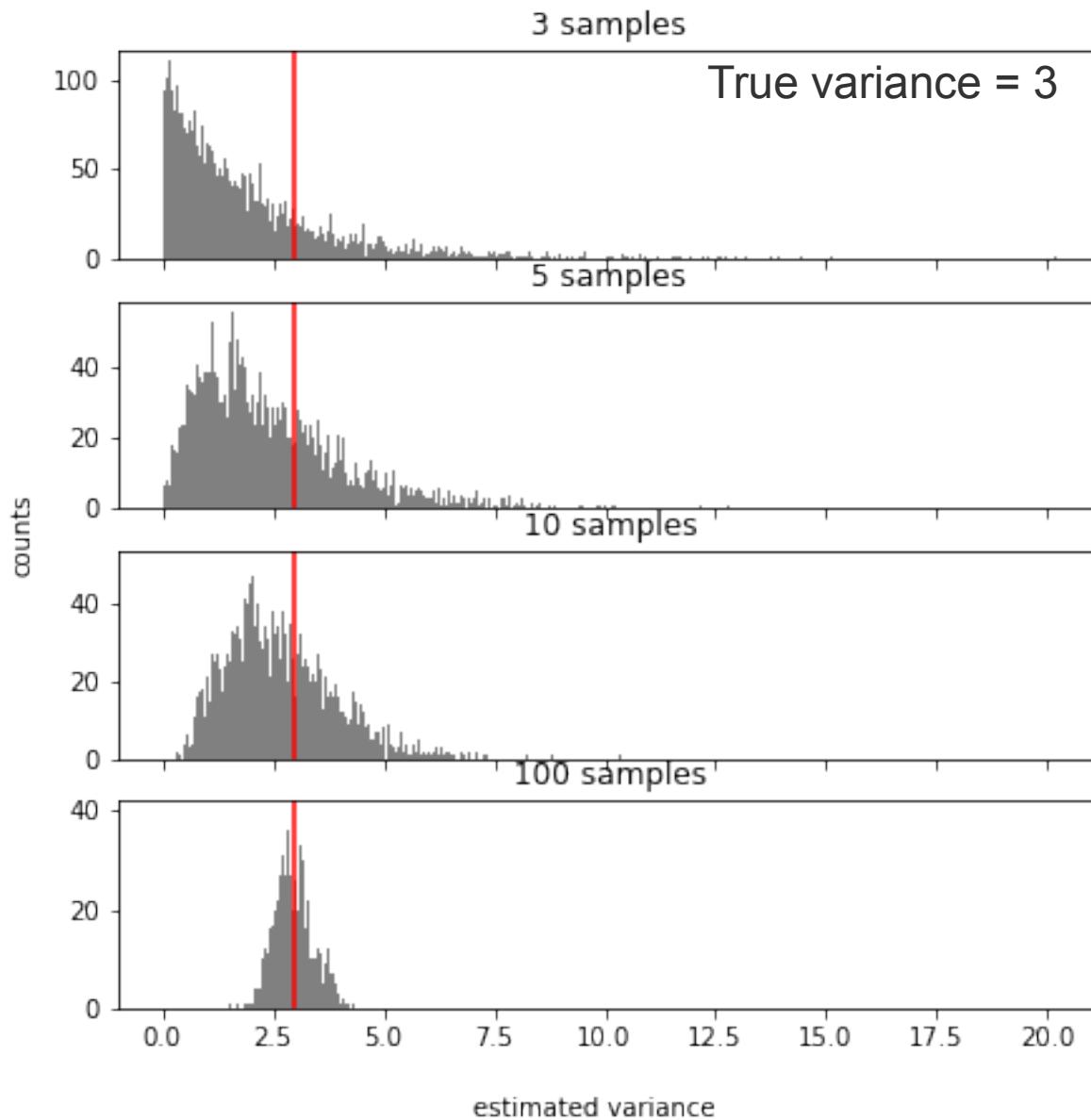
- DE: usually aim for ~30-40 million reads



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

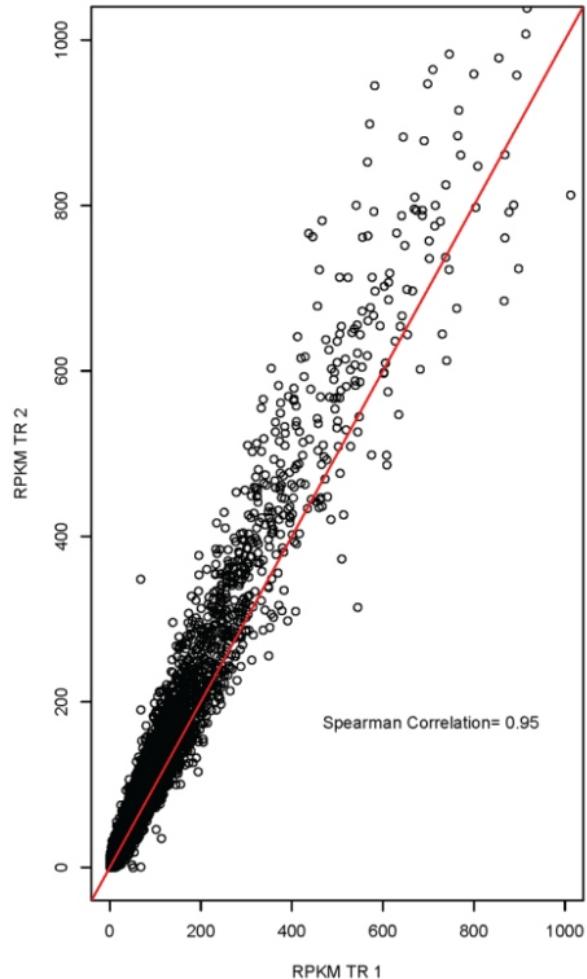
Not easy to know in advance

# Replicates - estimating a biological variance

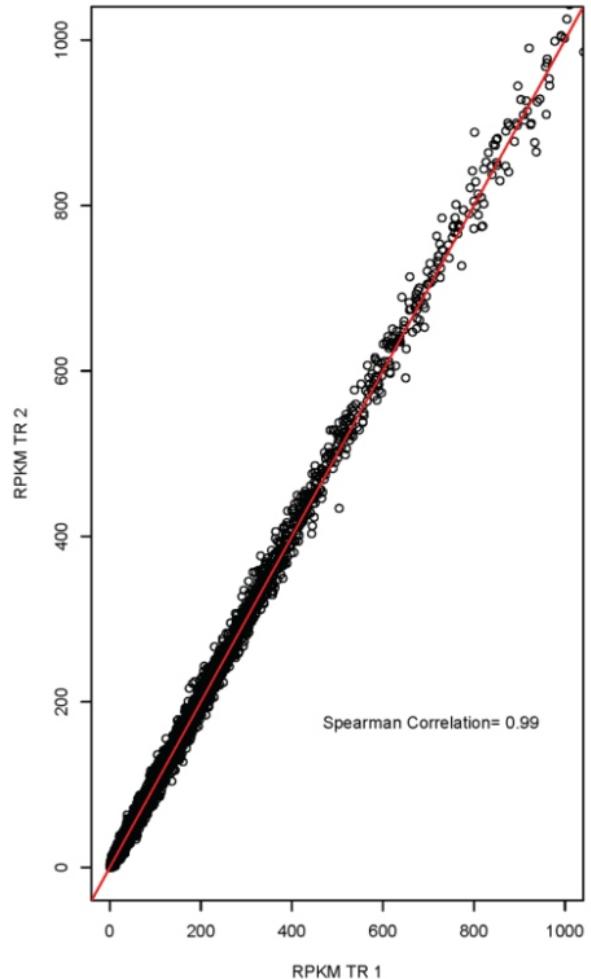


What does this tell you about  
the ideal number of replicates  
for your RNA-Seq experiment?

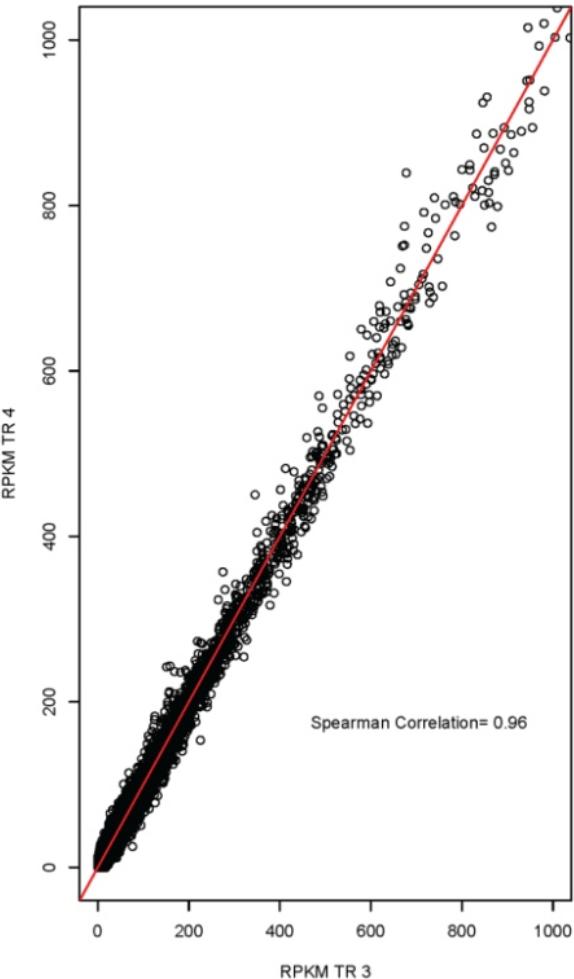
# Technical replicates – generally not required



D. simulans male  
heads BR2



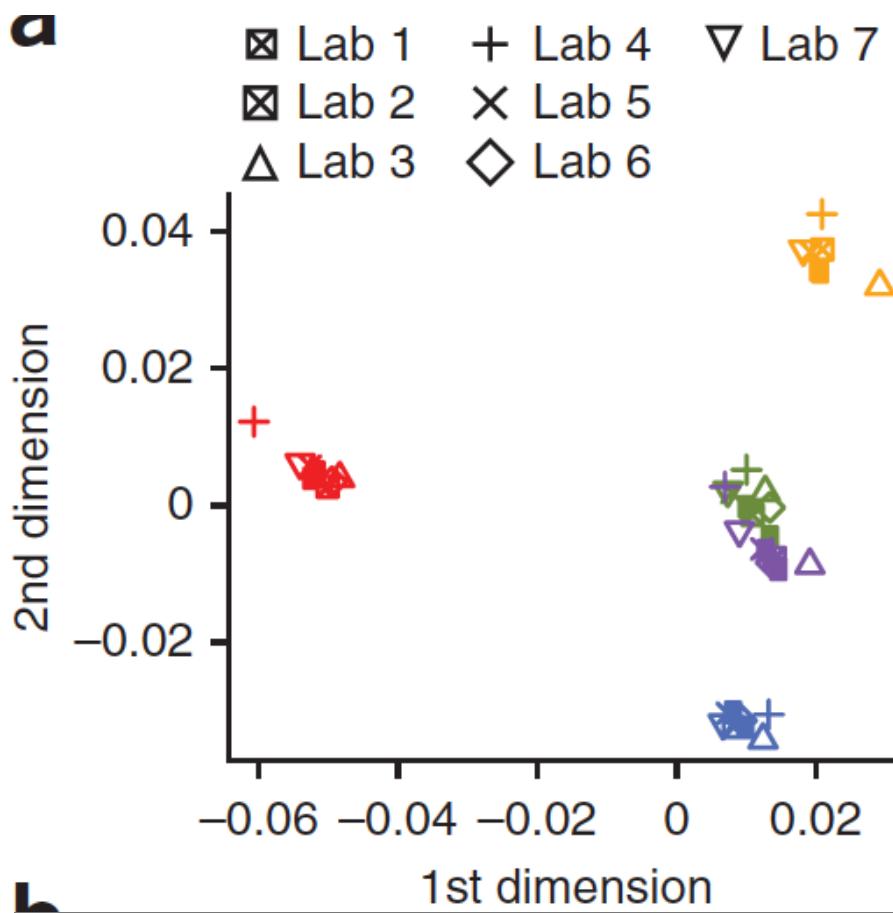
D. melanogaster  
female heads BR2



c167 cell line

# Experimental replicates – interlaboratory variation

- 7 labs, five shared samples, same library kit, paired-end reads of same length



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

[Affiliations](#) | [Contributions](#) | [Corresponding authors](#)

*Nature Biotechnology* 31, 1015–1022 (2013) | doi:10.1038/nbt.2702

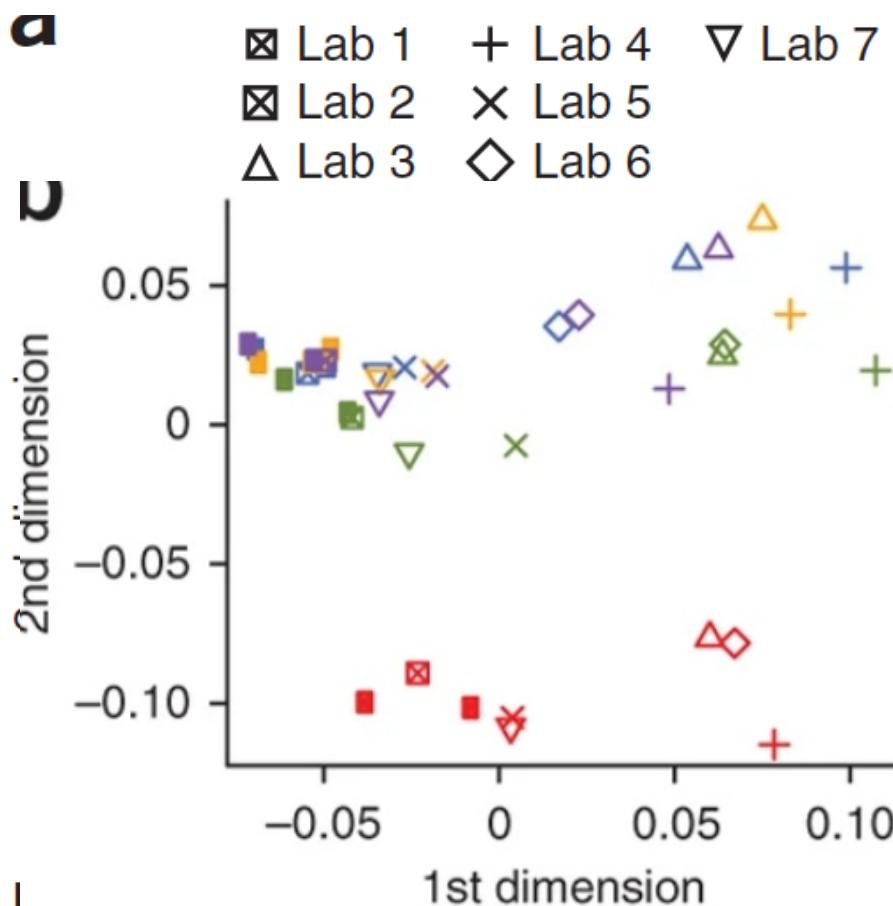
Received 04 January 2013 | Accepted 21 August 2013 | Published online 15 September 2013

| Corrected online 08 November 2013

Exon level : reproducible.

# Experimental replicates – interlaboratory variation

- 7 labs, five shared samples, same library kit, paired-end reads of same length



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

Peter A C 't Hoen, Marc R Friedländer, Jonas Almlöf, Michael Sammeth, Irina Pulyakina, 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

Affiliations | Contributions | Corresponding authors

Nature Biotechnology 31, 1015–1022 (2013) | doi:10.1038/nbt.2702

Received 04 January 2013 | Accepted 21 August 2013 | Published online 15 September 2013

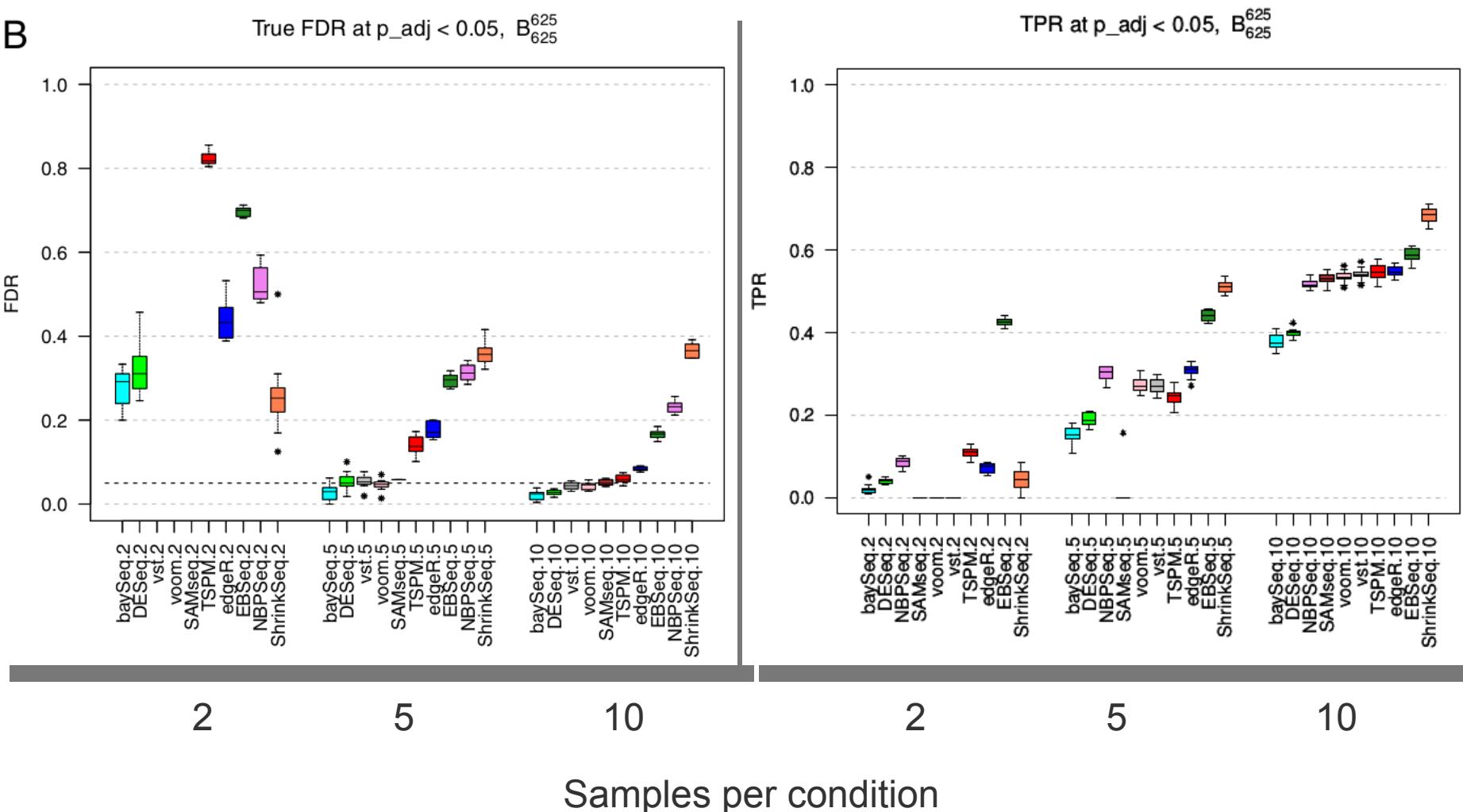
| Corrected online 08 November 2013

Transcript level : reproducible.

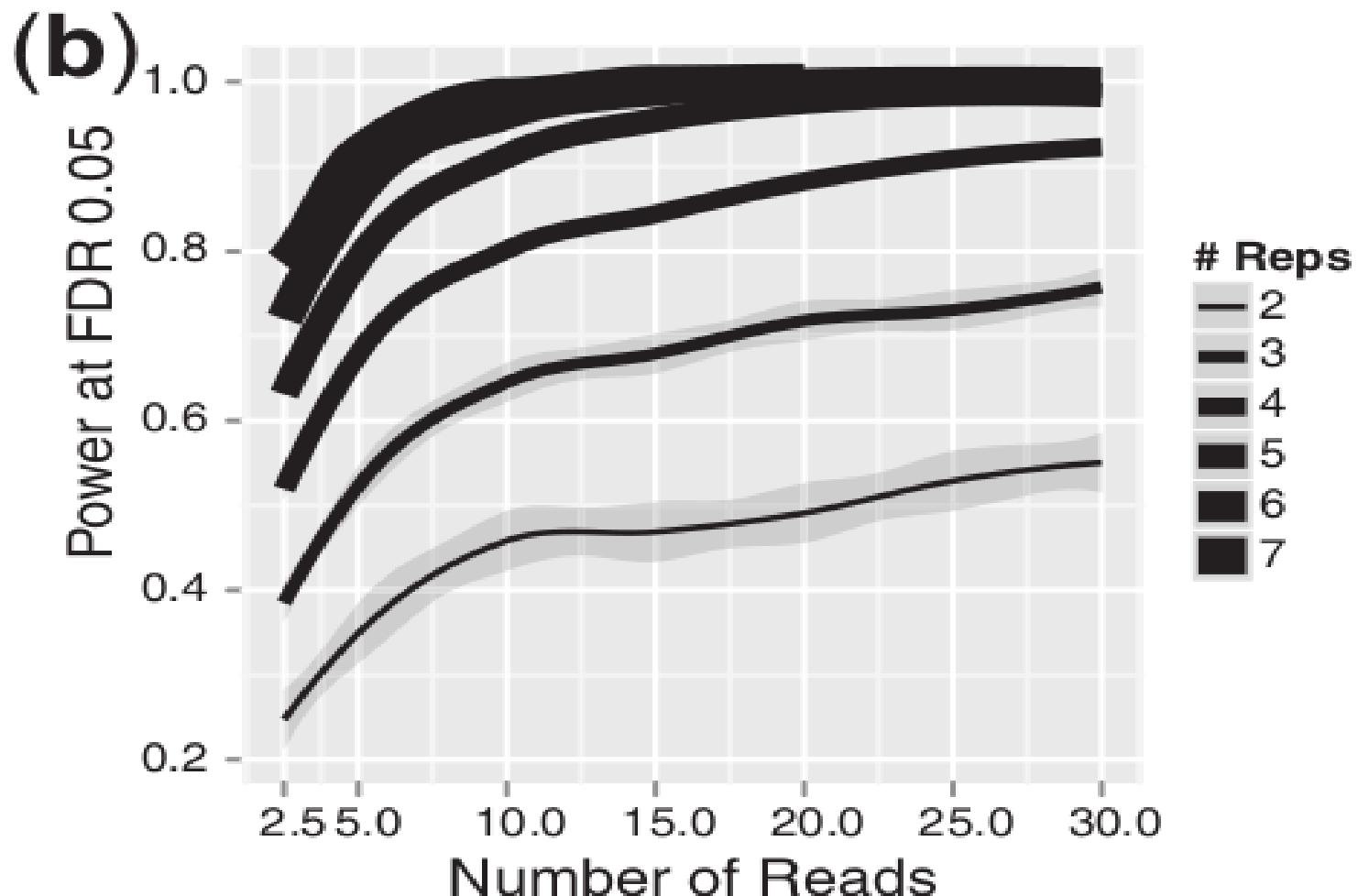
CHECK FOR BATCH EFFECT!

# Biological replicates

B



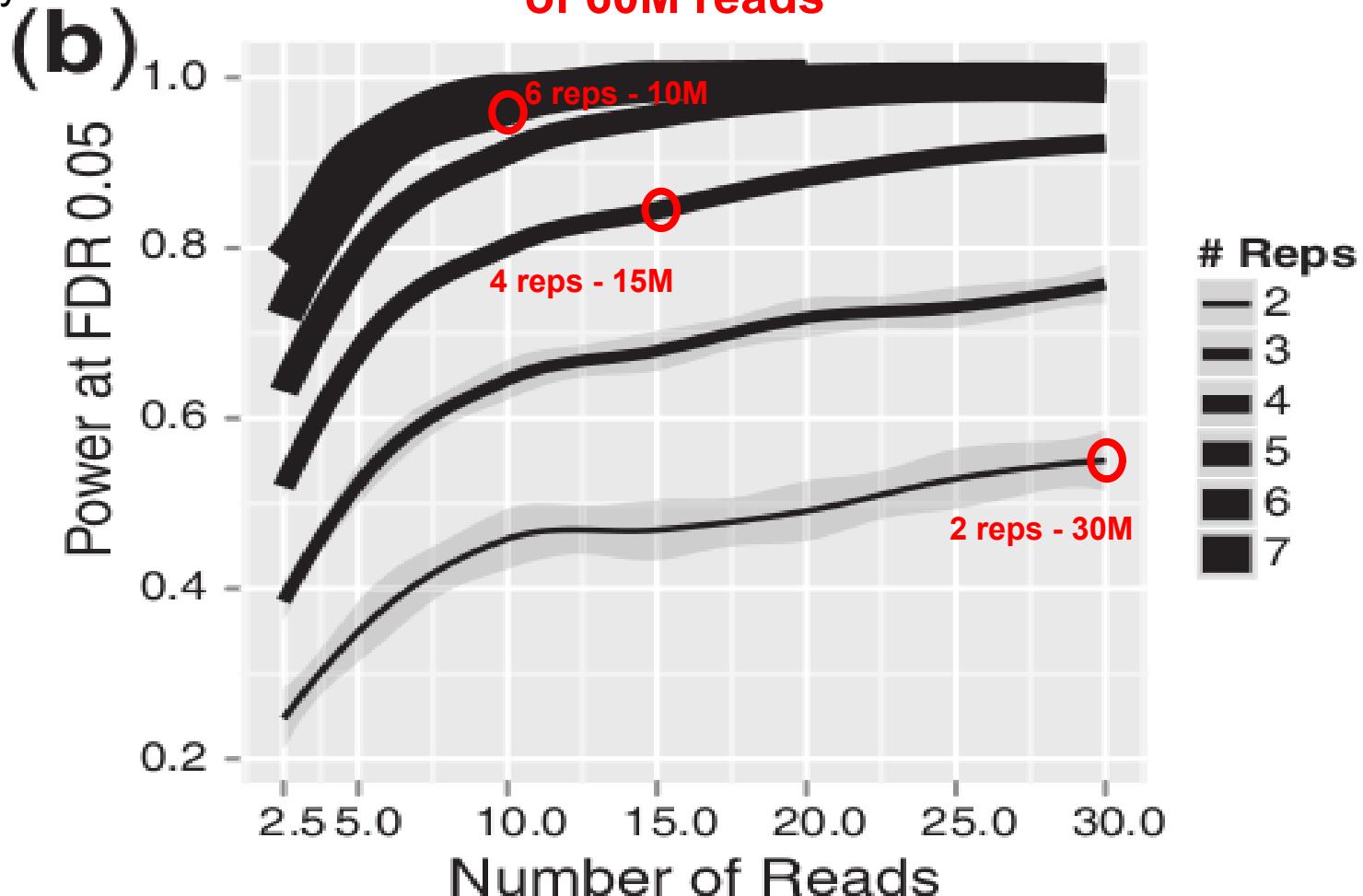
# More reads or more replicates ?



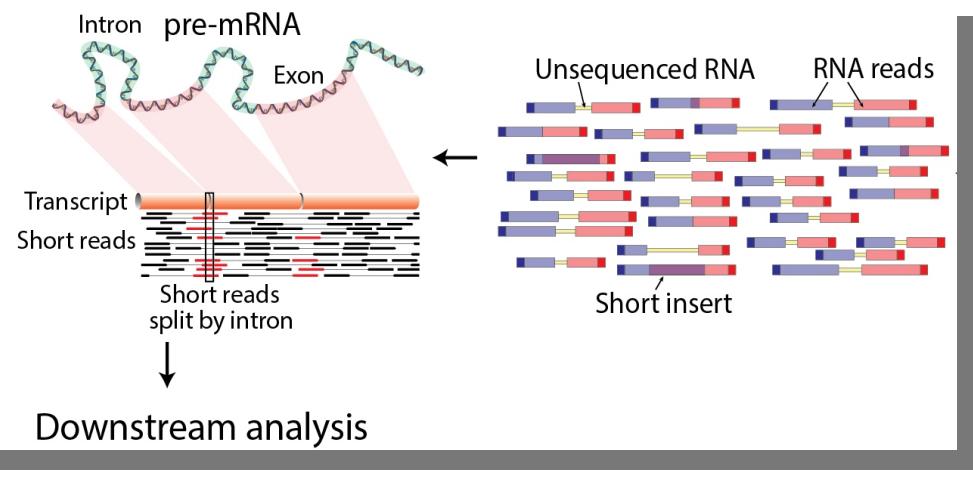
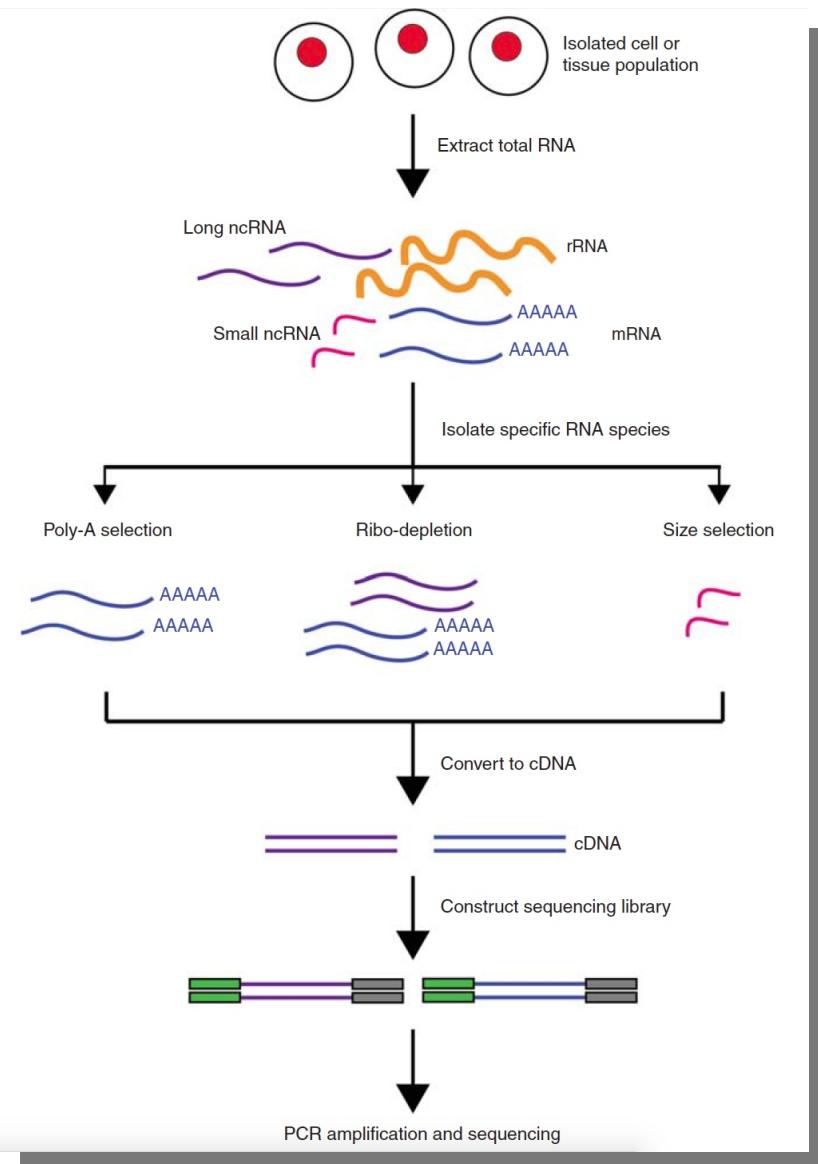
# More reads or more replicates ?

Generally more replicates should be preferred if they can be obtained.

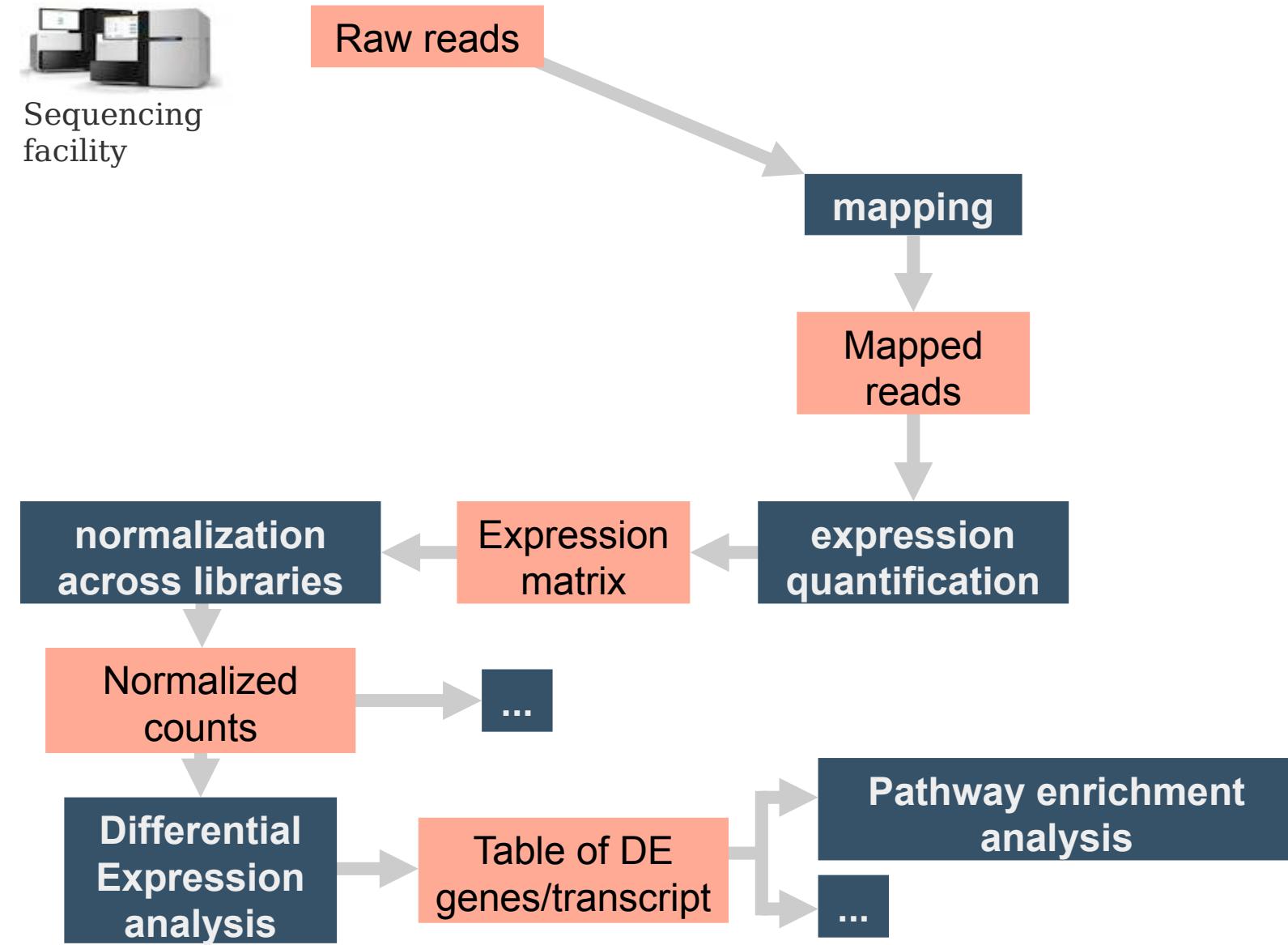
for a fixed total  
of 60M reads



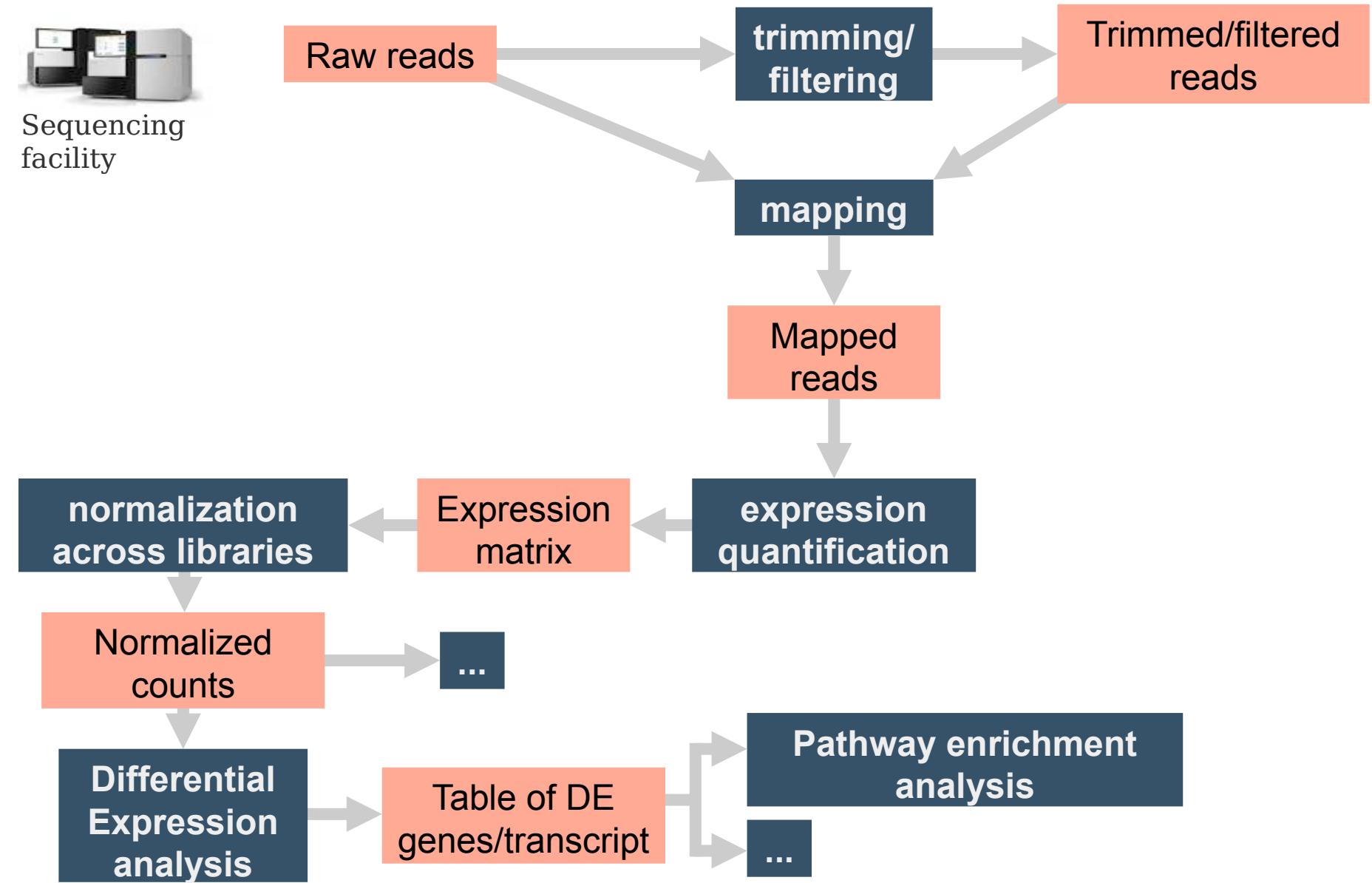
# Basic RNA-Seq Protocol Overview



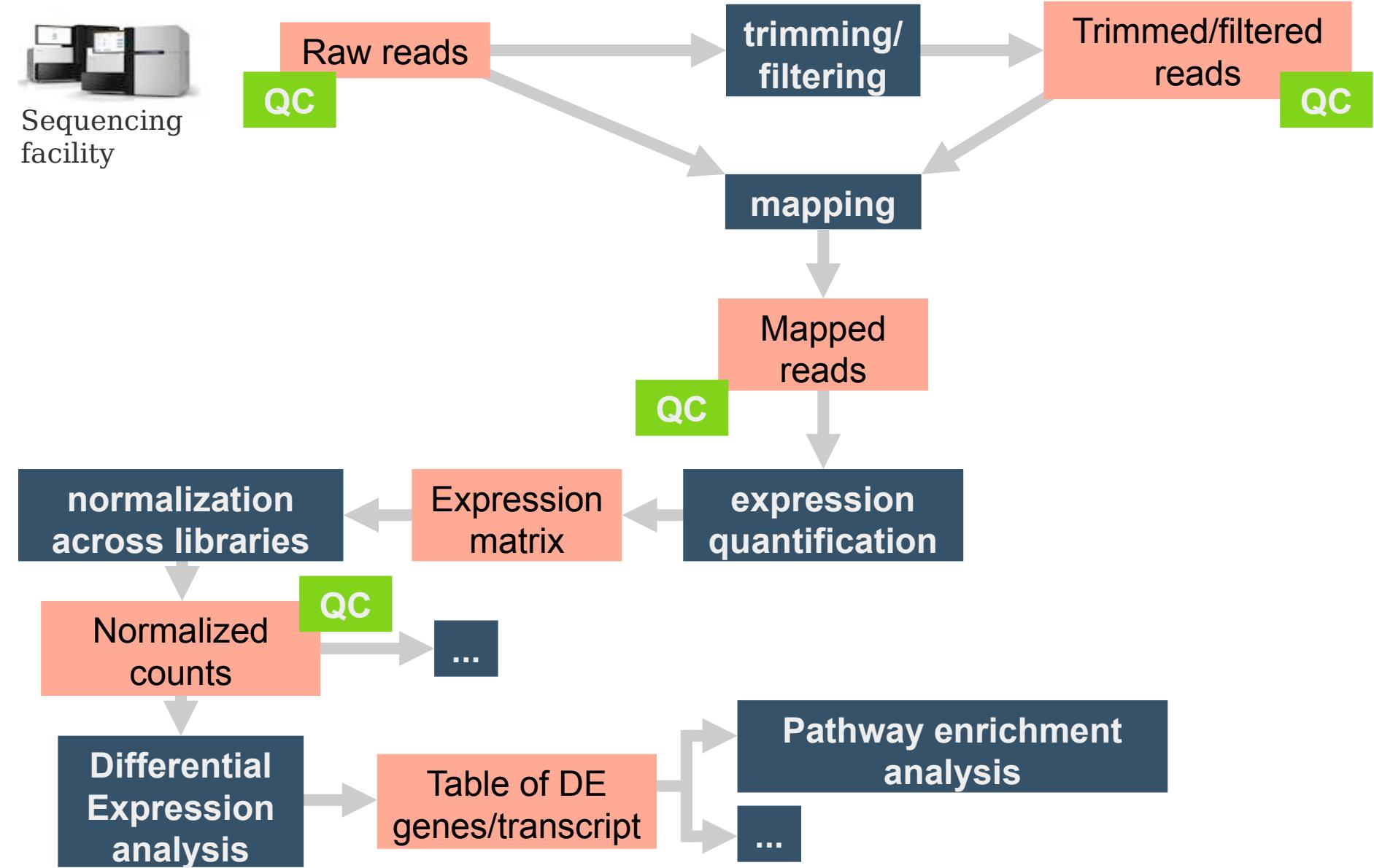
# RNA-seq data analysis – basic analysis



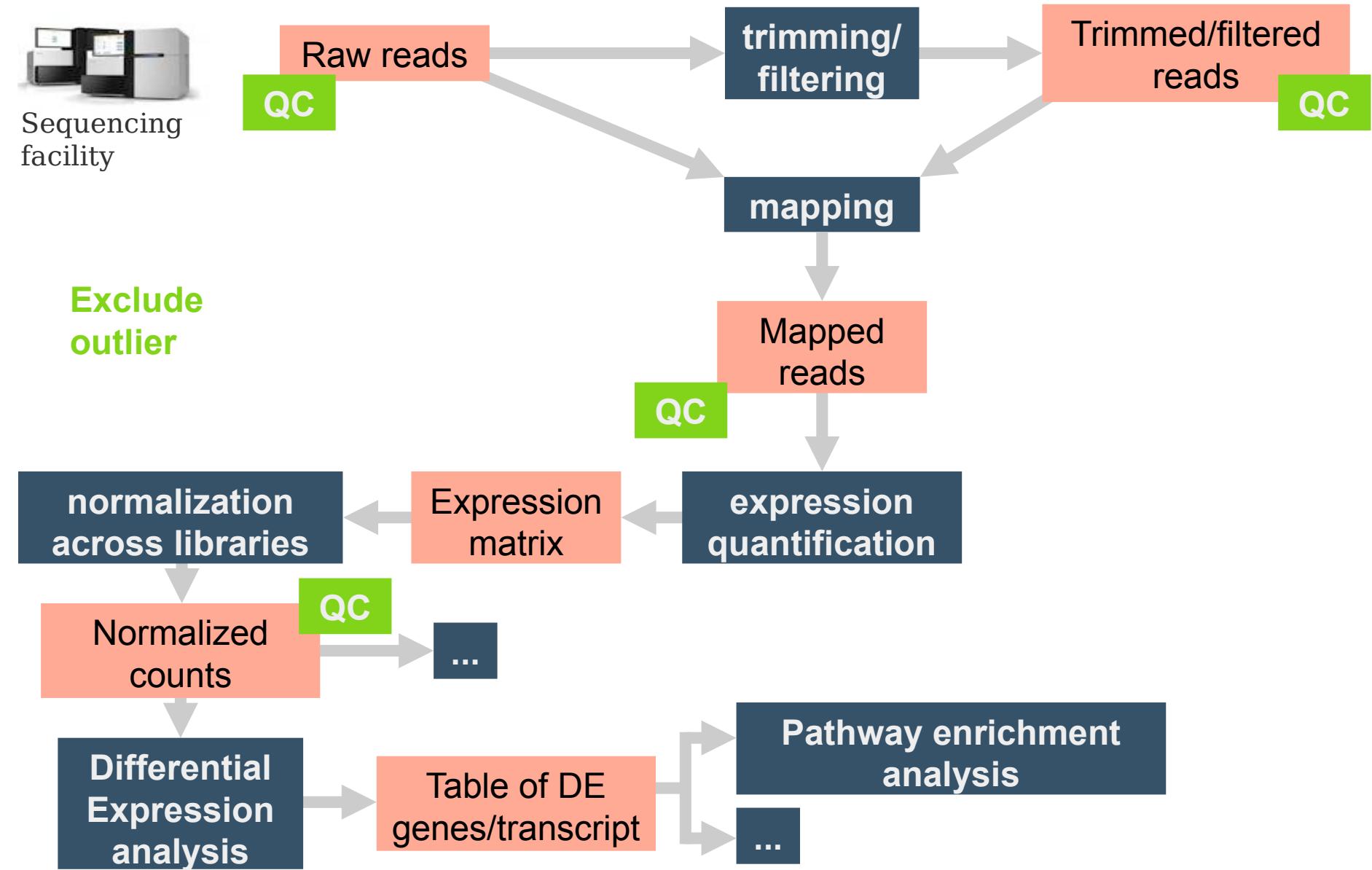
# RNA-seq data analysis – with trimmed reads



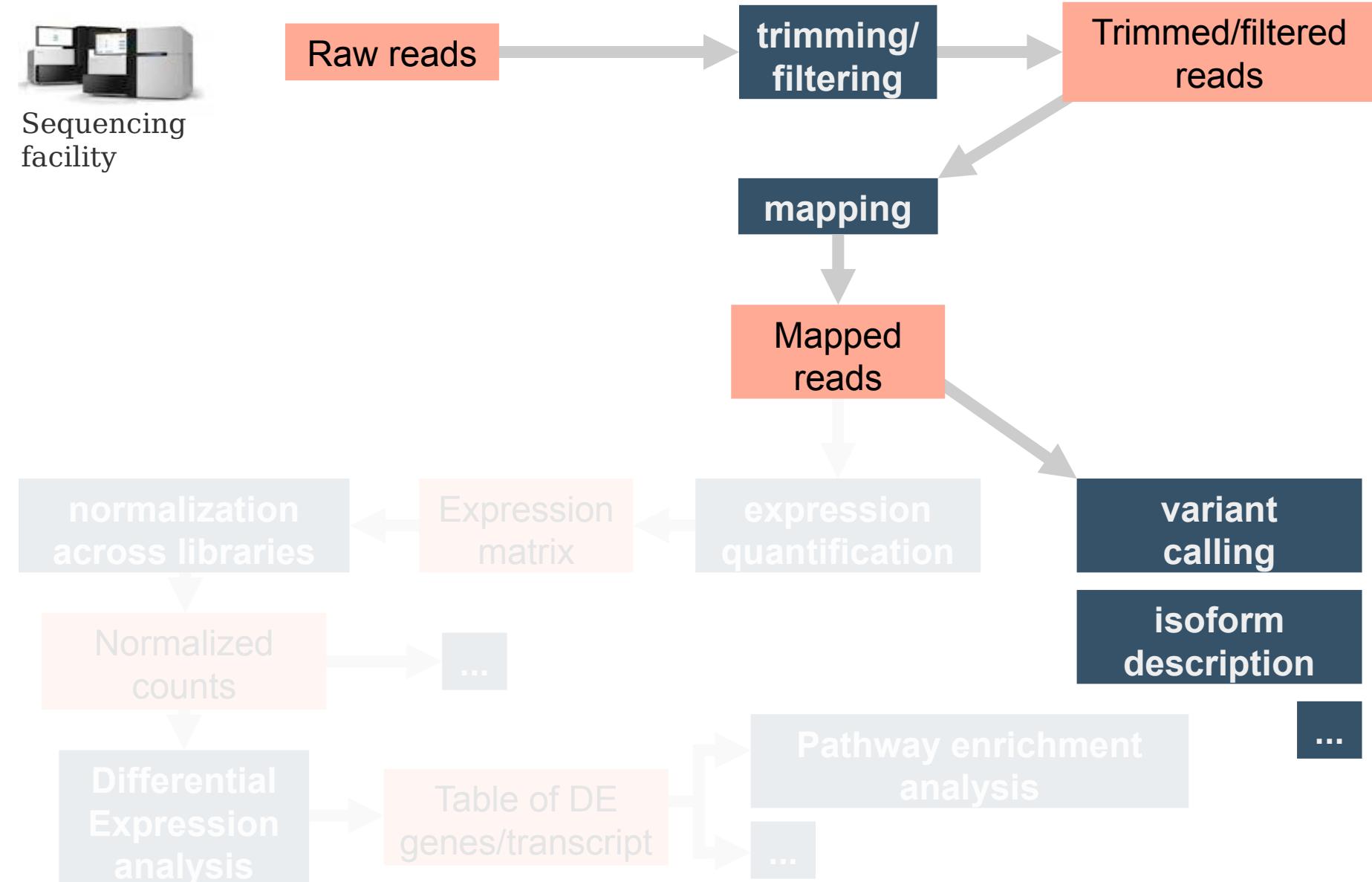
# RNA-seq data analysis – main QC steps



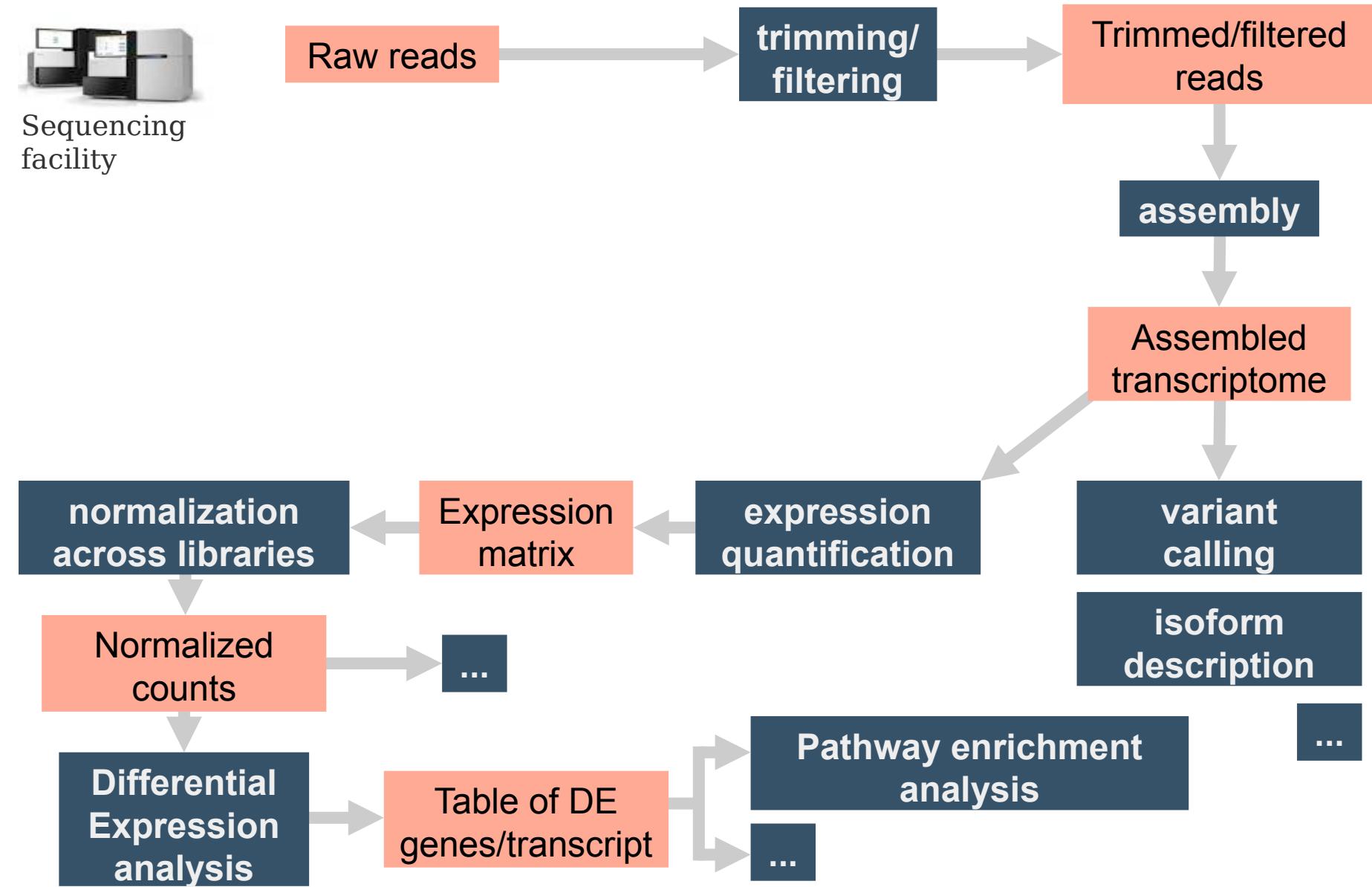
# RNA-seq data analysis – main QC steps



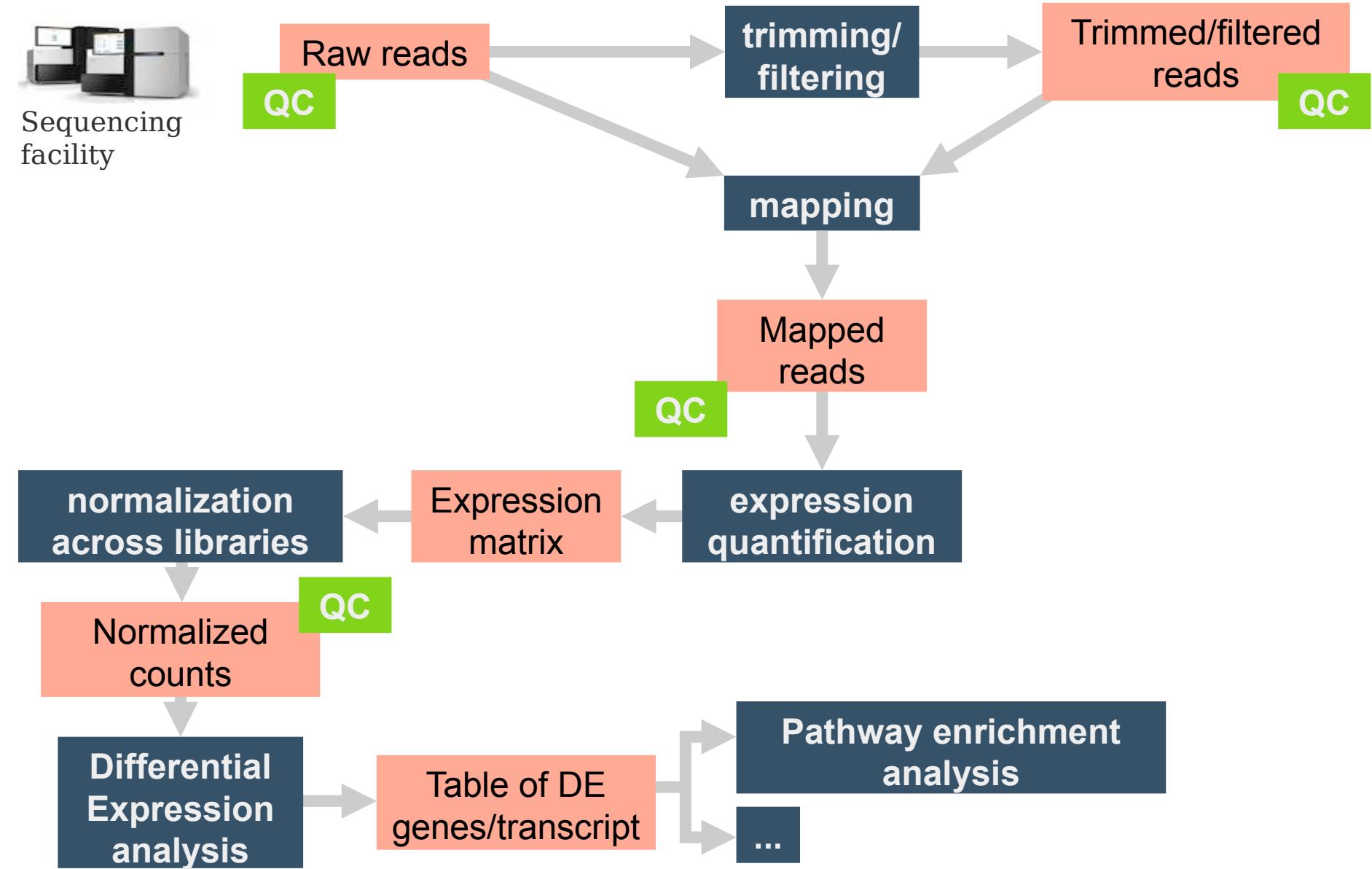
# RNA-seq data analysis – not only differential expression

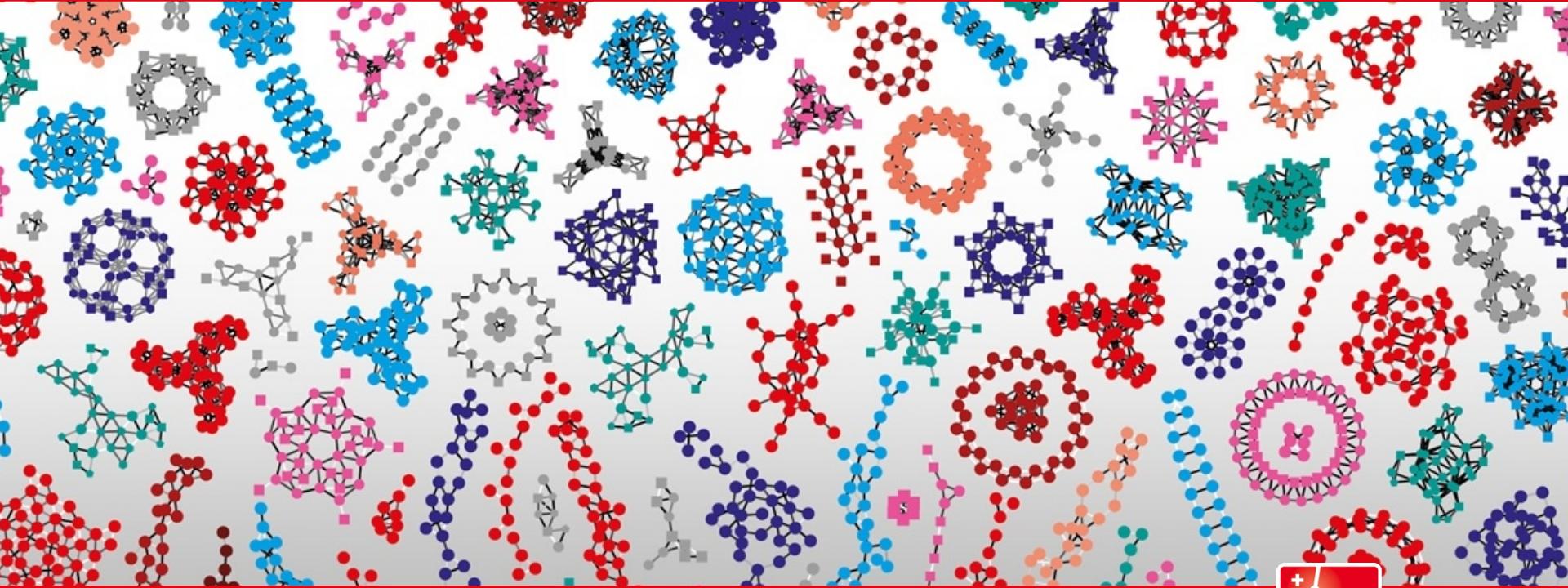


# RNA-seq data analysis – de novo transcriptome assembly



# RNA-seq data analysis – program for this course





SIB  
Swiss Institute of  
Bioinformatics

## Contributors:

**Wandrille Duchemin**  
**Geoffrey Fucile**  
**Walid Gharib**  
**Pablo Escobar Lopez**  
**Charlotte Soneson**  
**Mihaela Zavolan**