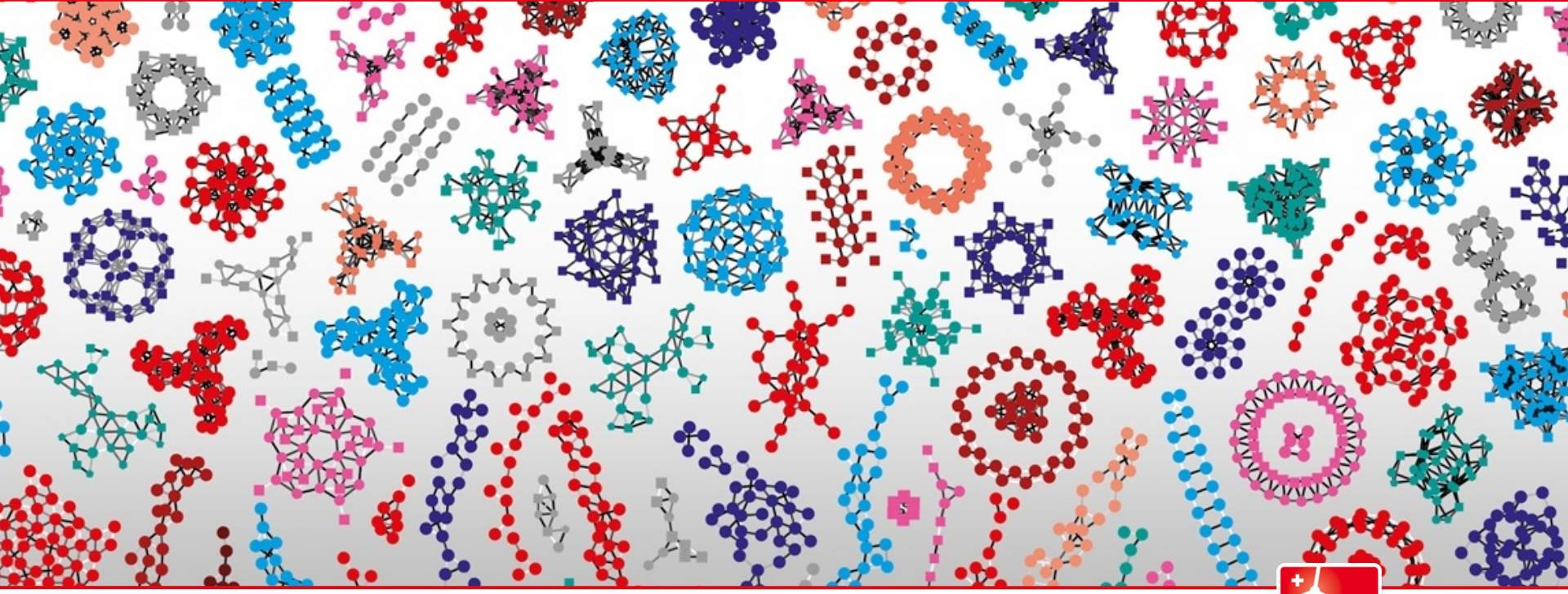


Swiss Institute of Bioinformatics

Introduction to RNA-Seq – Overview

Wandrille Duchemin

First some logistics

Internet access :

- eduroam
- unibas-visitor > login page > get code with your phone

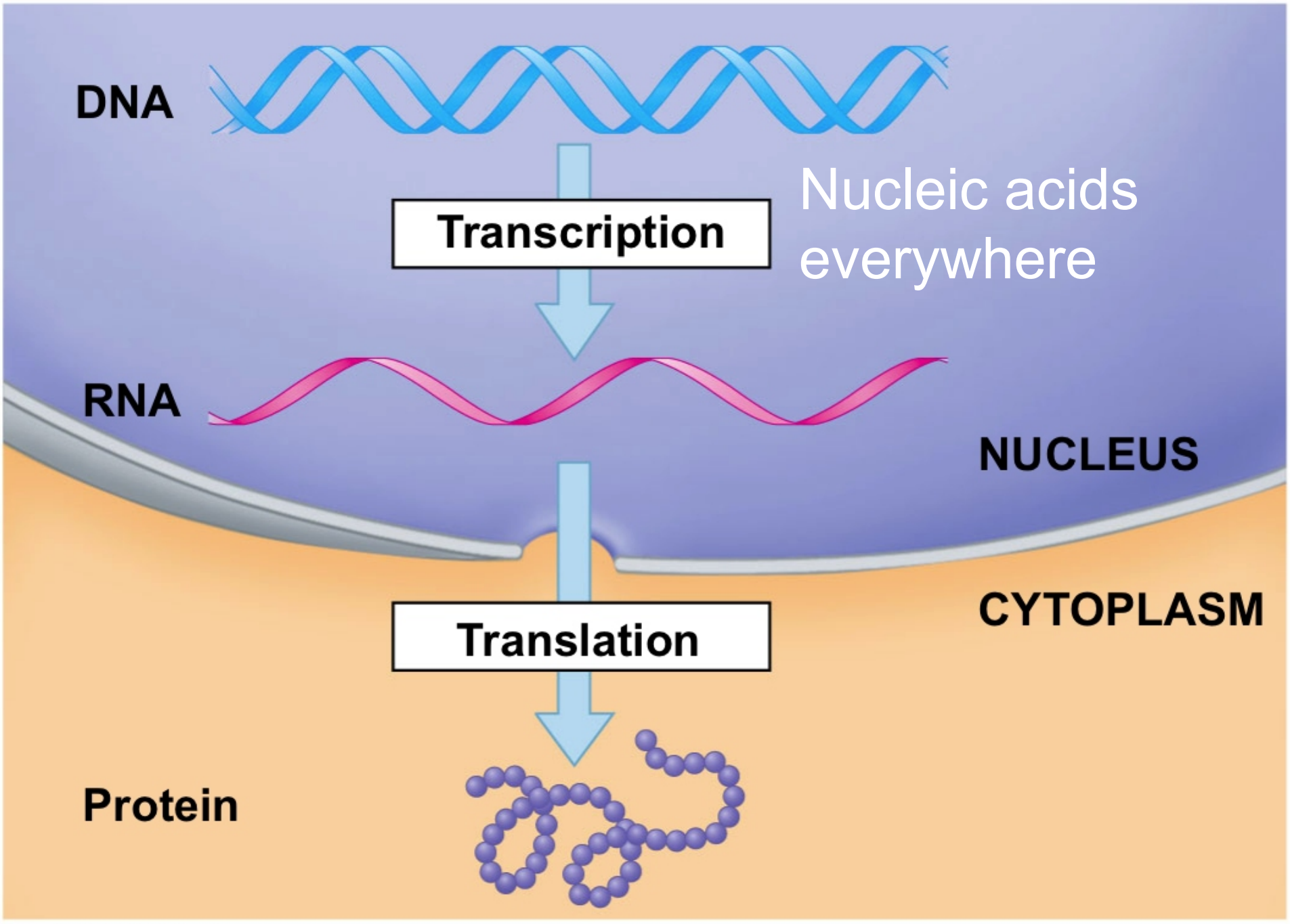
- Course from 09:00 to 17:00
- Lunch break 12:00 → 13:00
- 15min breaks around 10:30 and 15:00

Questions are welcome all the time

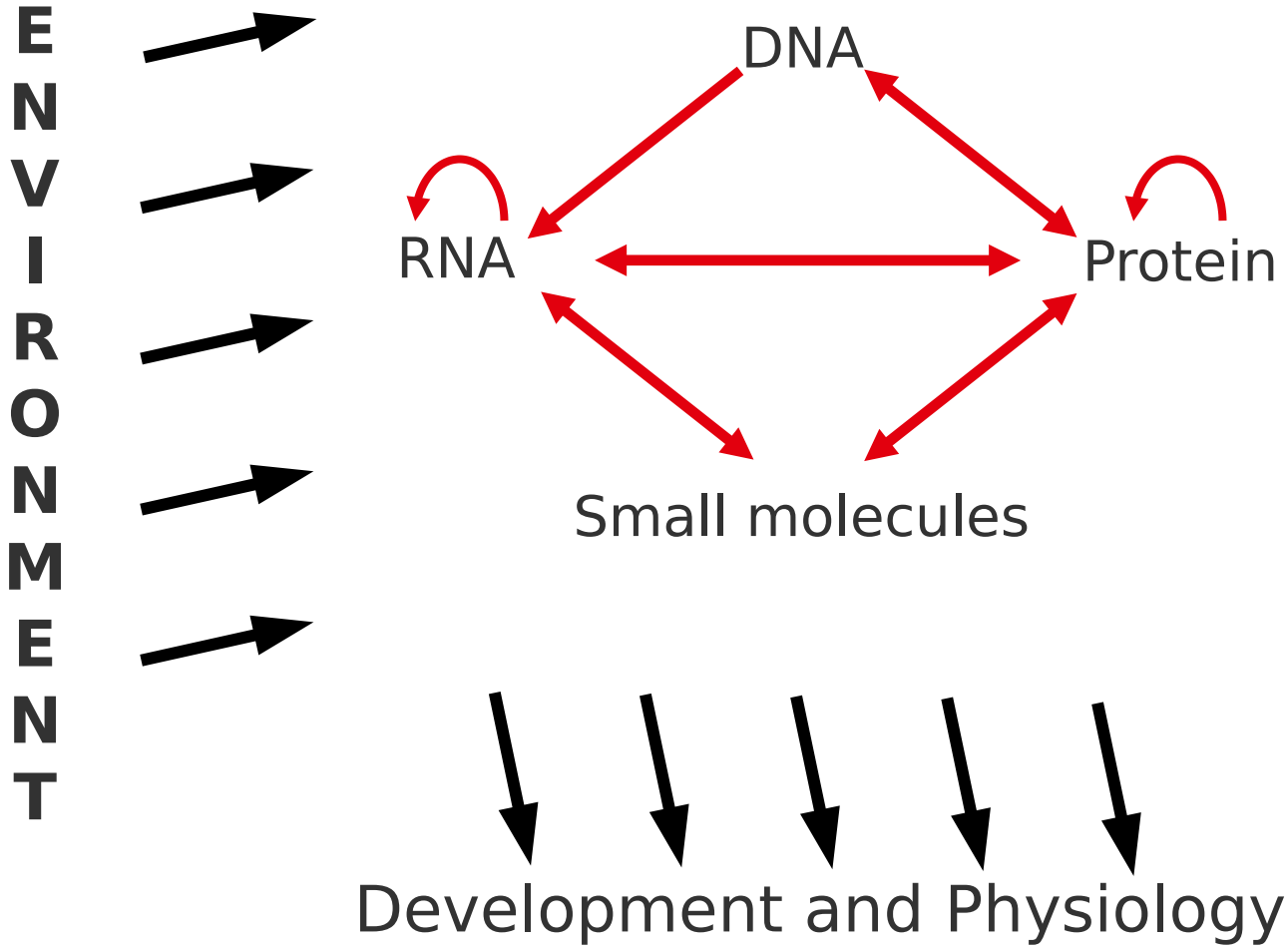
During practicals: post-its

I have a question/ problem

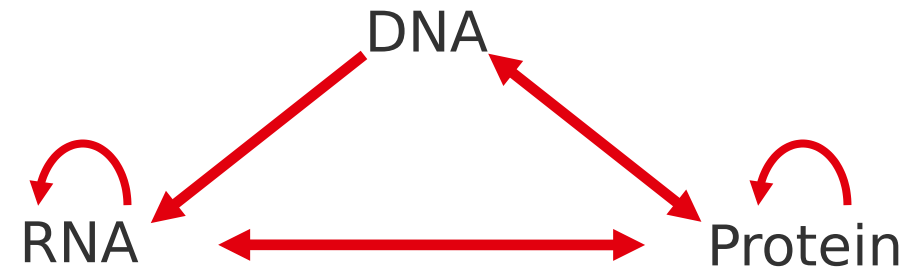
I am finished with the practical

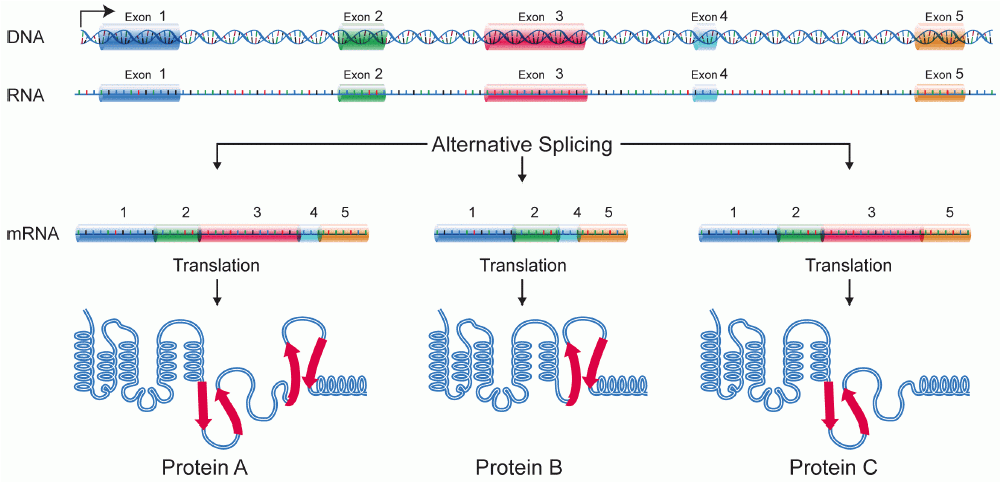


Dogma Nuanced
complex interactions and regulations



Dogma Nuanced
alternative splicing





https://en.wikipedia.org/wiki/Alternative_splicing#/media/File:DNA_alternative_splicing.gif

Dogma Nuanced
alternative splicing

~20,000 mammalian genes

↓

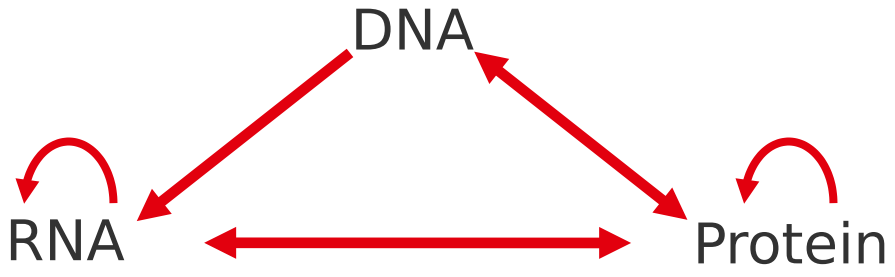
>>100,000 (?) transcripts

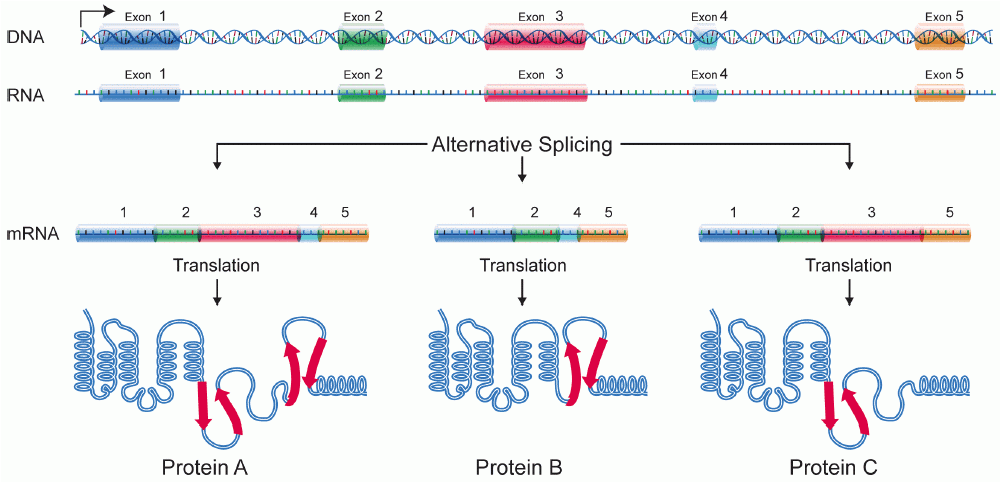
↓

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

↓

Cellular Diversity



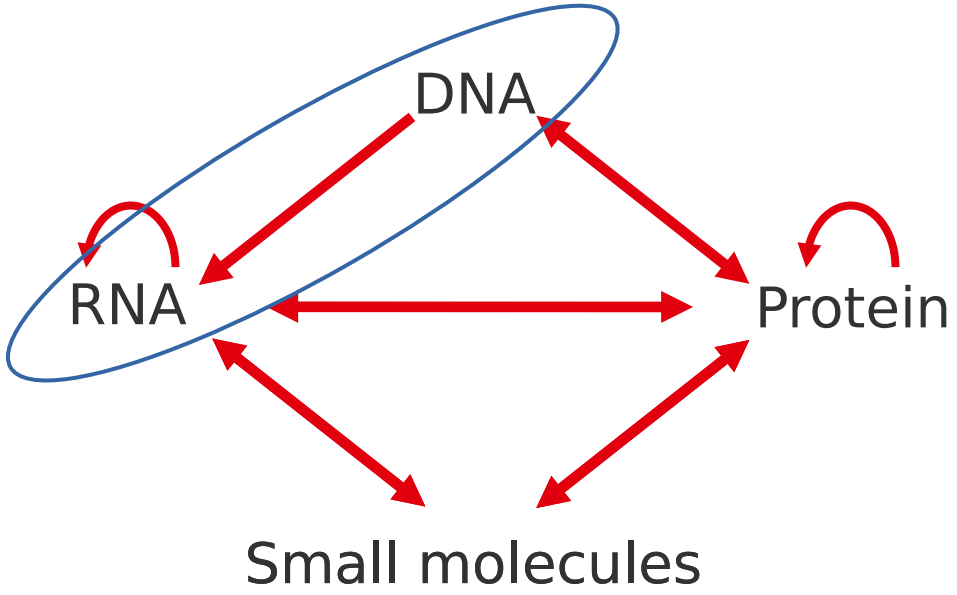


https://en.wikipedia.org/wiki/Alternative_splicing#/media/File:DNA_alternative_splicing.gif

Dogma Nuanced
- 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 : chromatine 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 : chromatine opening

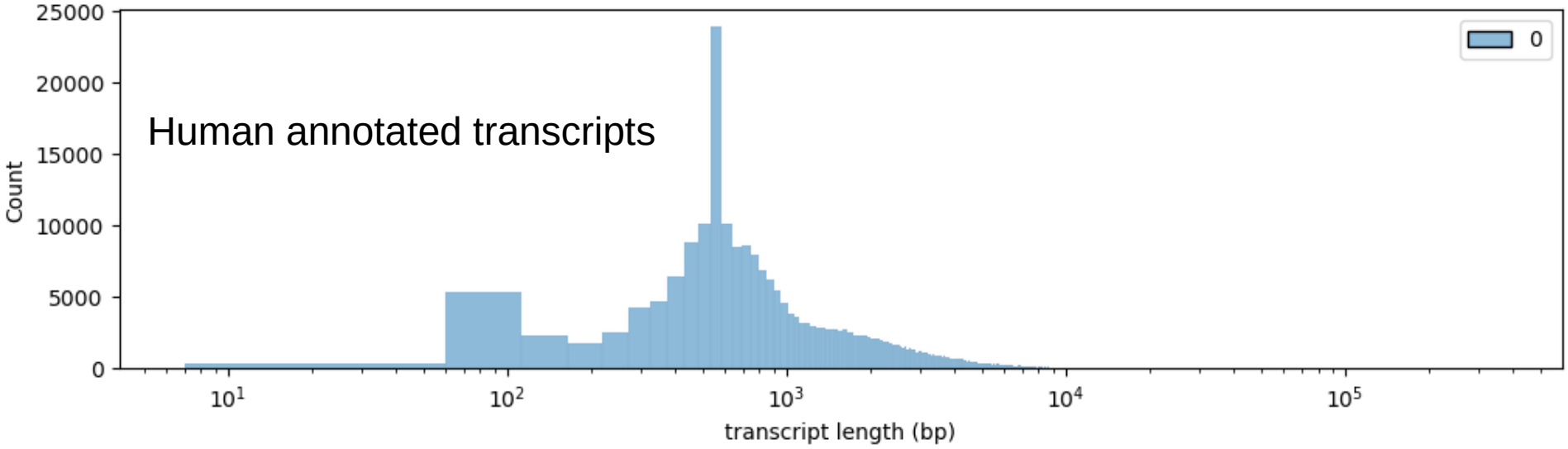
Imagination is the limit!

- **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

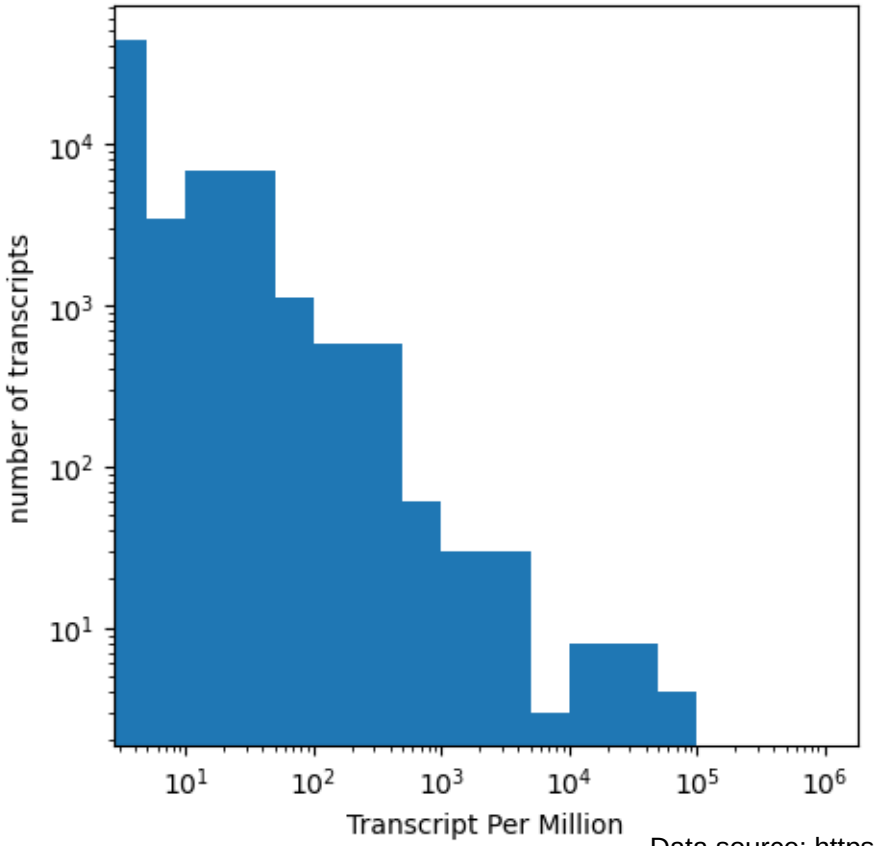
- **Transcripts are diverse in size**

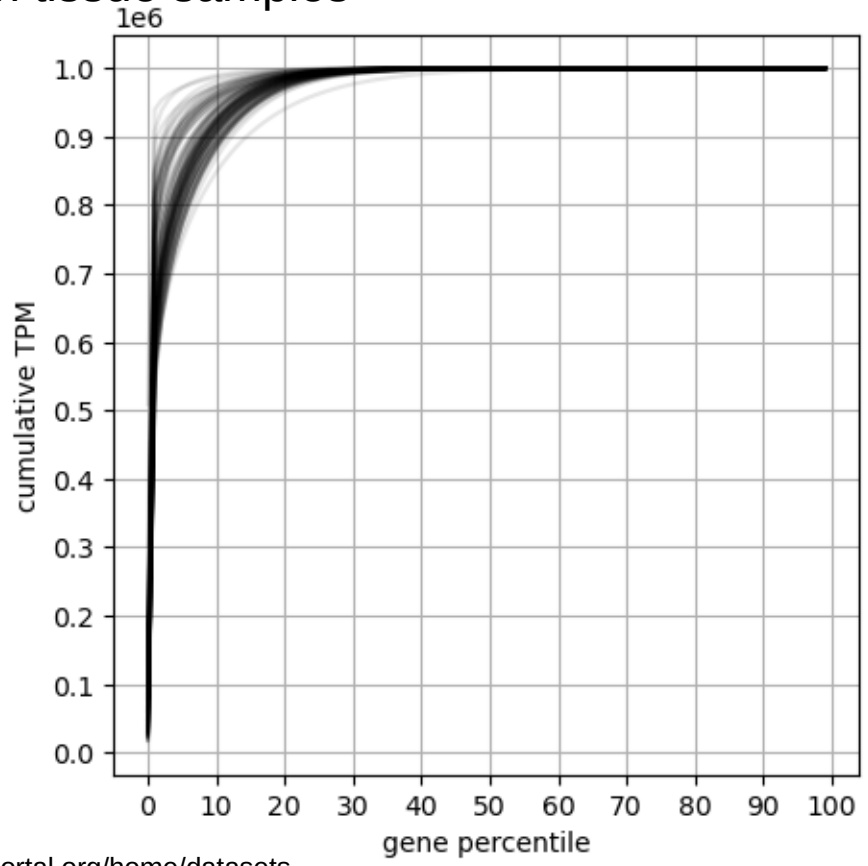


Some challenges for RNA-Seq

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

Gtex V8 - human tissue samples



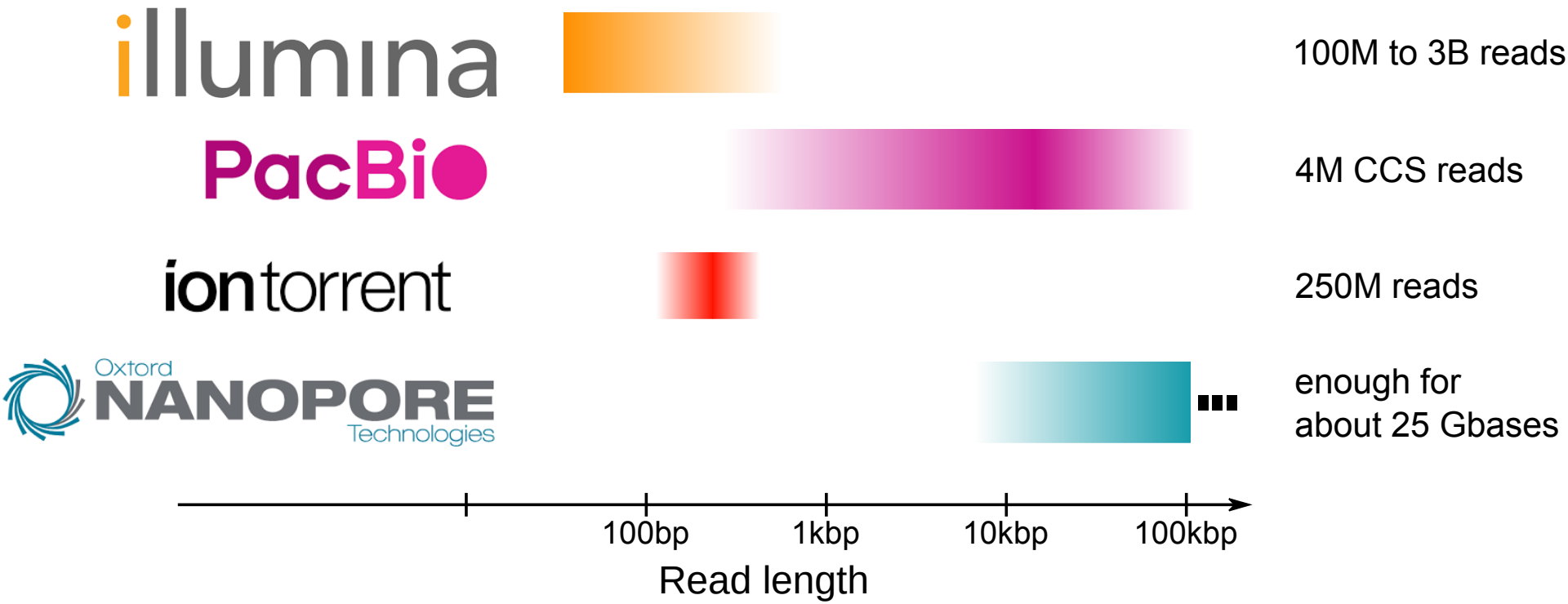


Data source: https://gtexportal.org/home/datasets

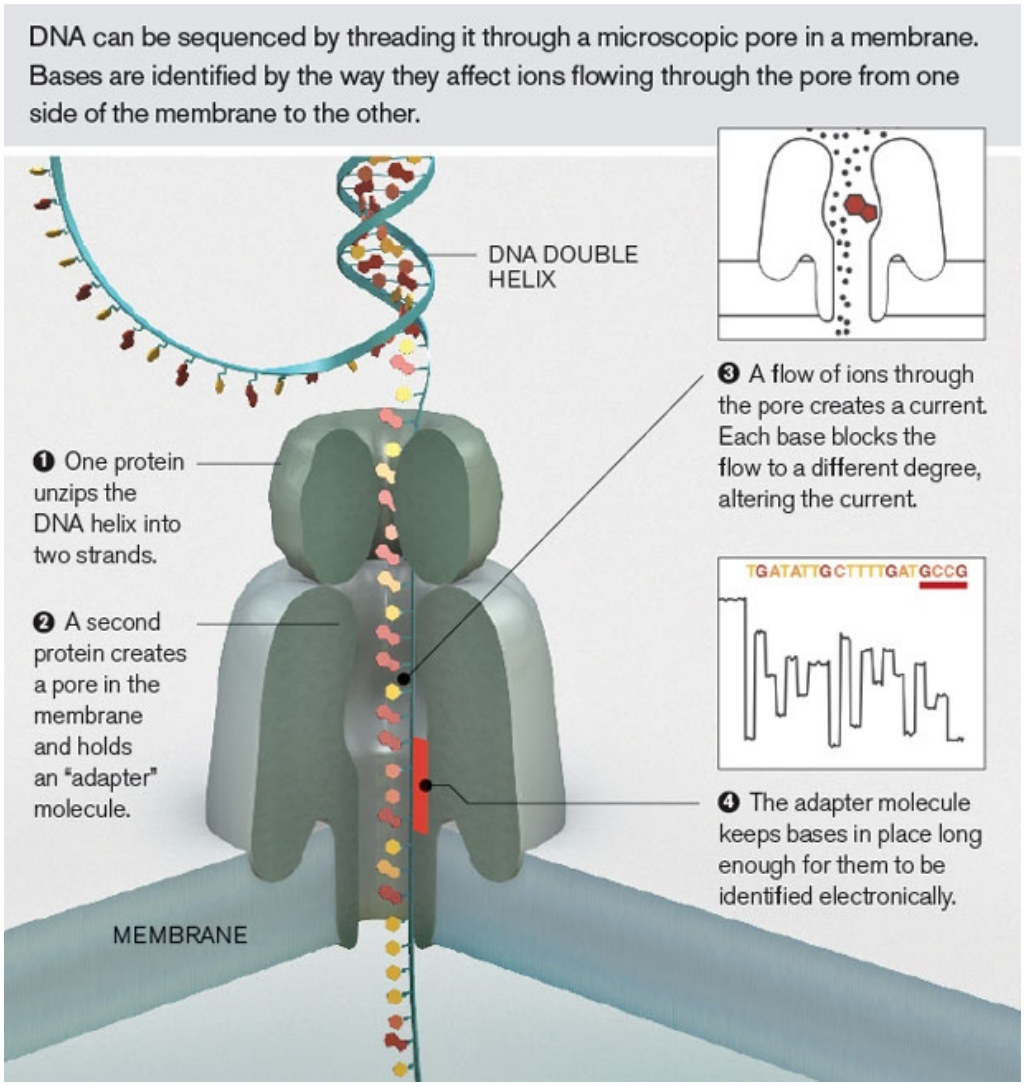
Some challenges for RNA-Seq

- **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?**

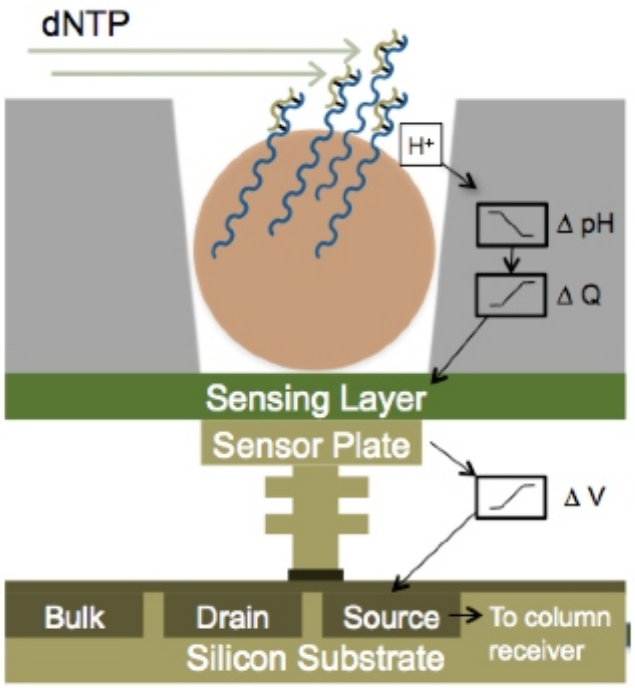
Main sequencing technologies

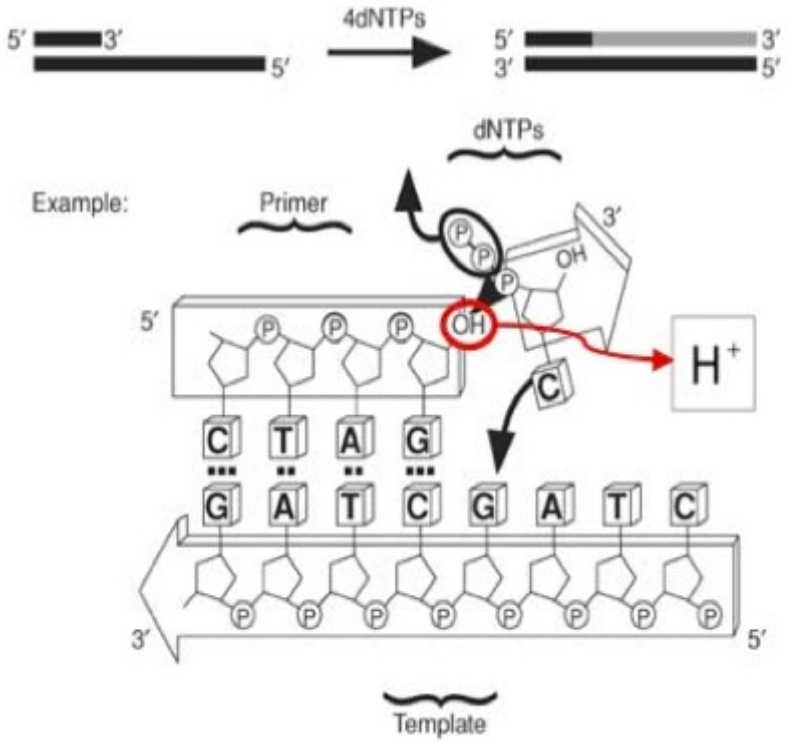


Oxford Nanopore – direct DNA/RNA sequencing

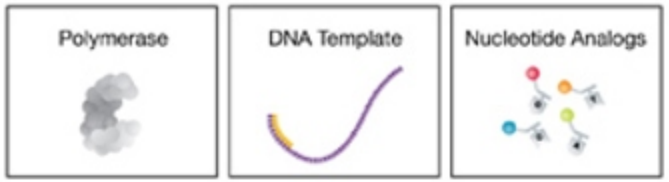


Ion torrent – reading pH changes

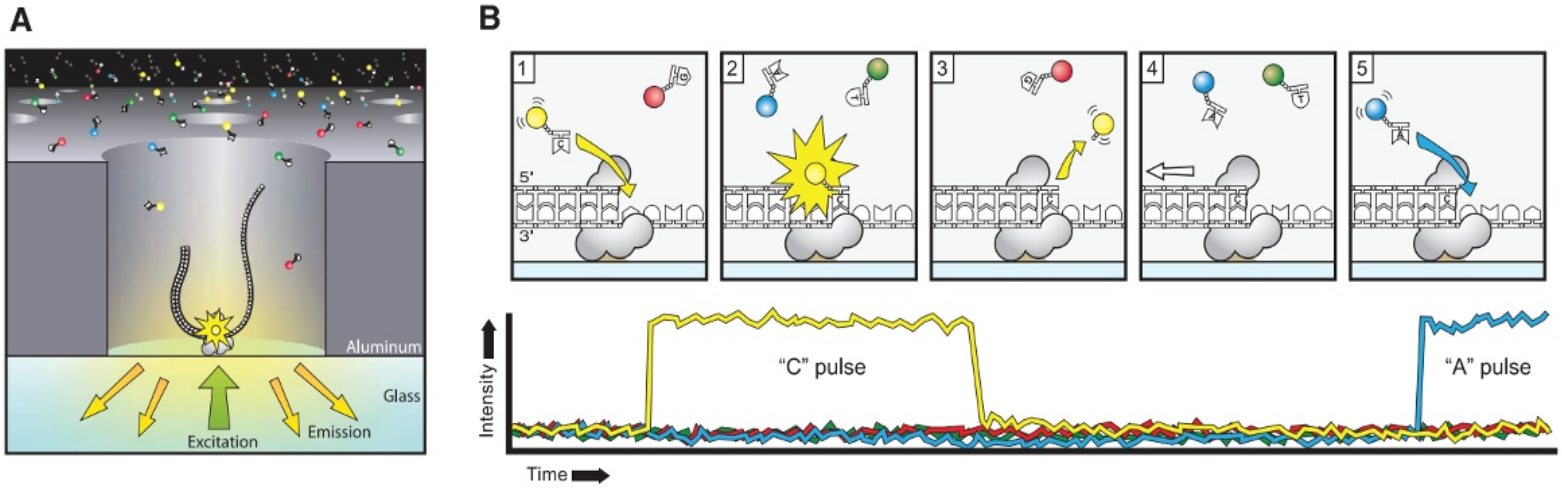




Pacific Biosciences – Single Molecule Real Time sequencing



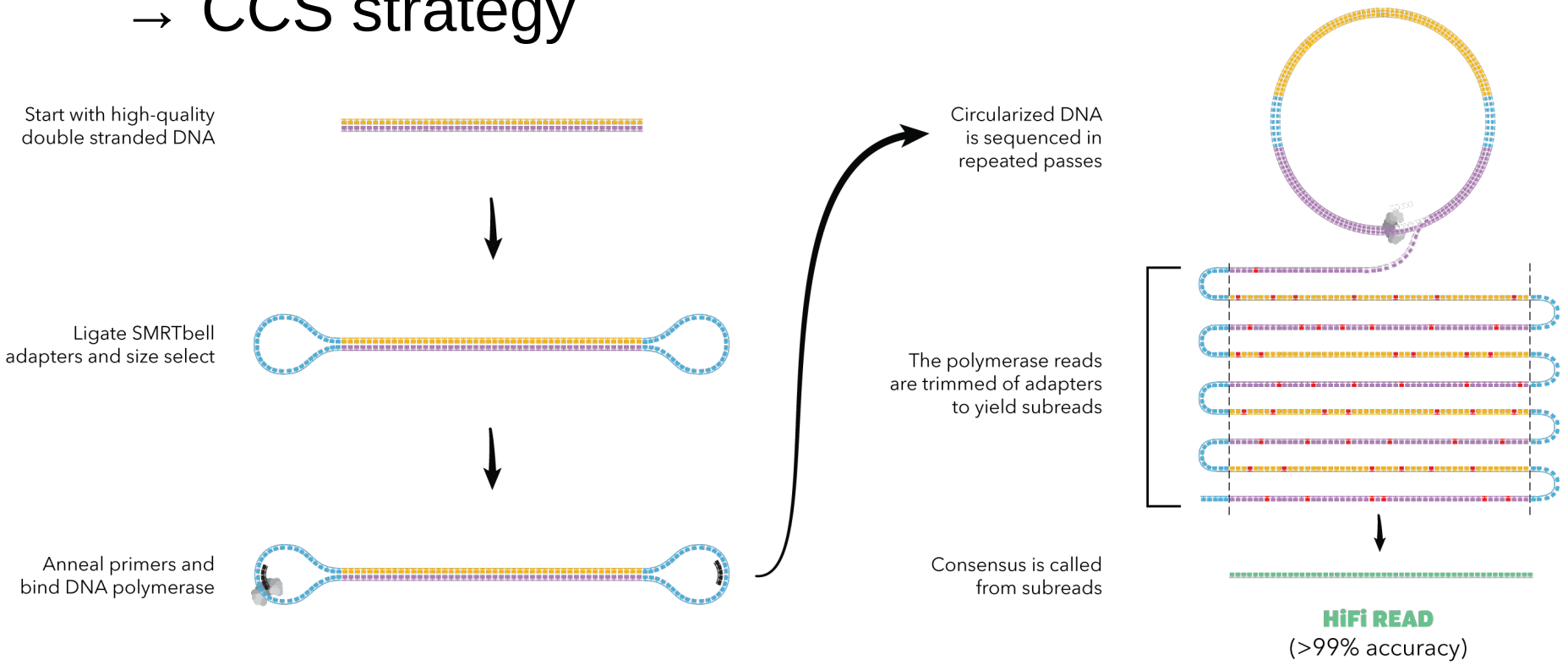
Rhoads & Au.
Genomics Proteomics Bioinformatics 2015



Pacific Biosciences – Circular Consensus Sequencing

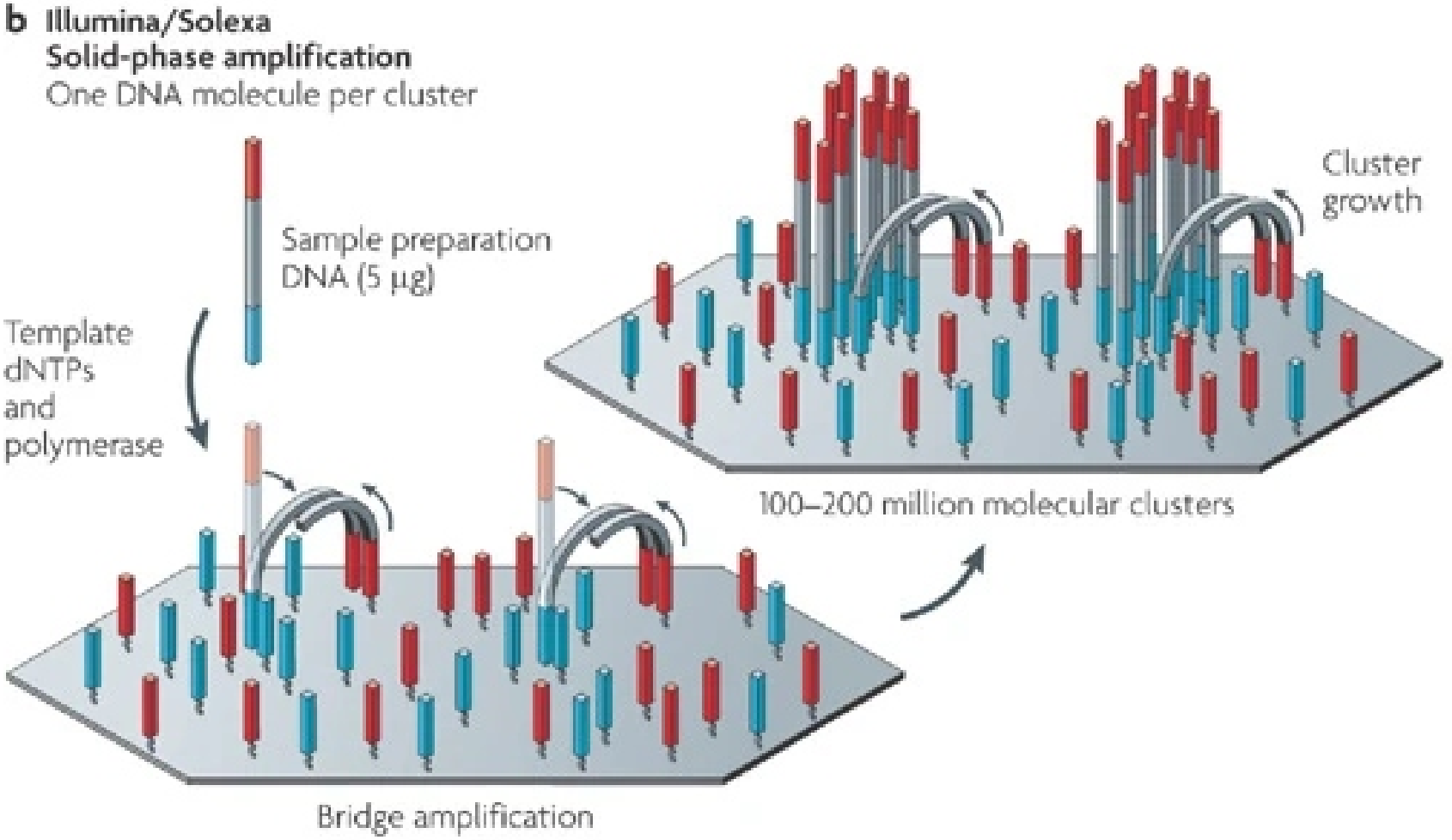
Raw reads : 15% **random** errors

→ CCS strategy



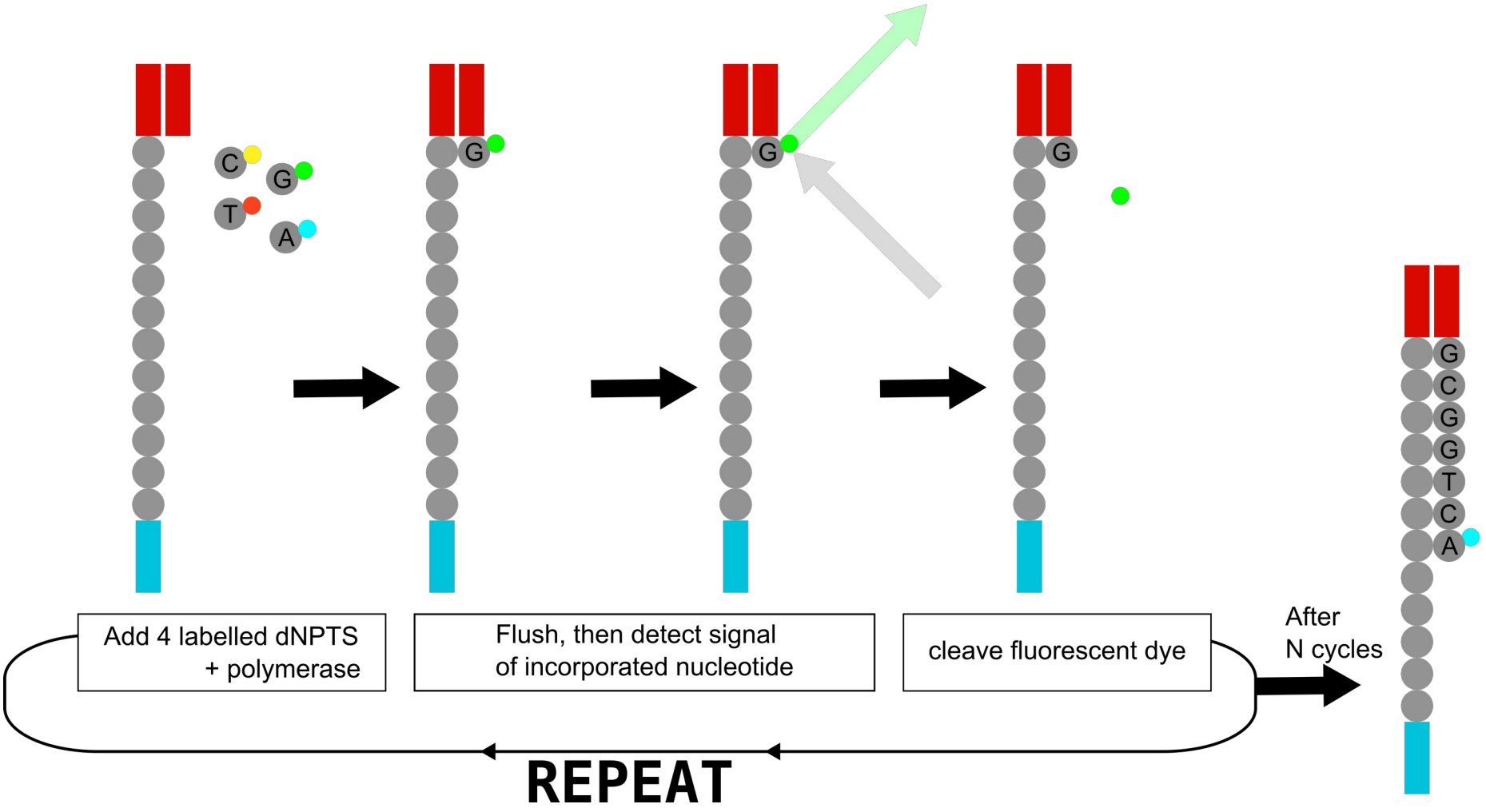
Typical in isoseq

Illumina sequencing – “cluster formation”

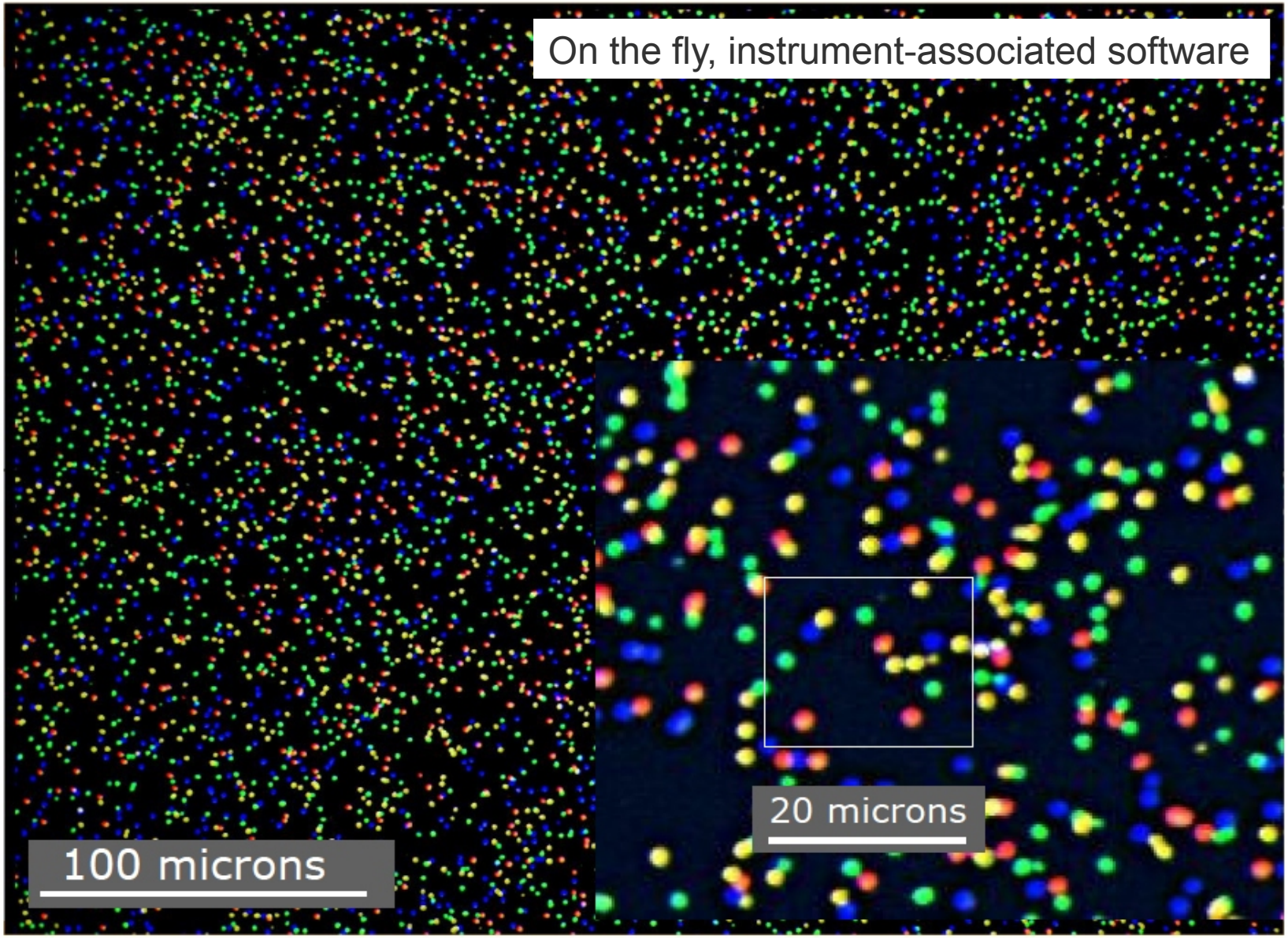


Source: Fig1b of Metzker, M. Sequencing technologies — the next generation. Nat Rev Genet 11, 31–46 (2010). https://doi.org/10.1038/nrg2626

Illumina sequencing – “sequencing by synthesis”



Illumina sequencing – image analysis

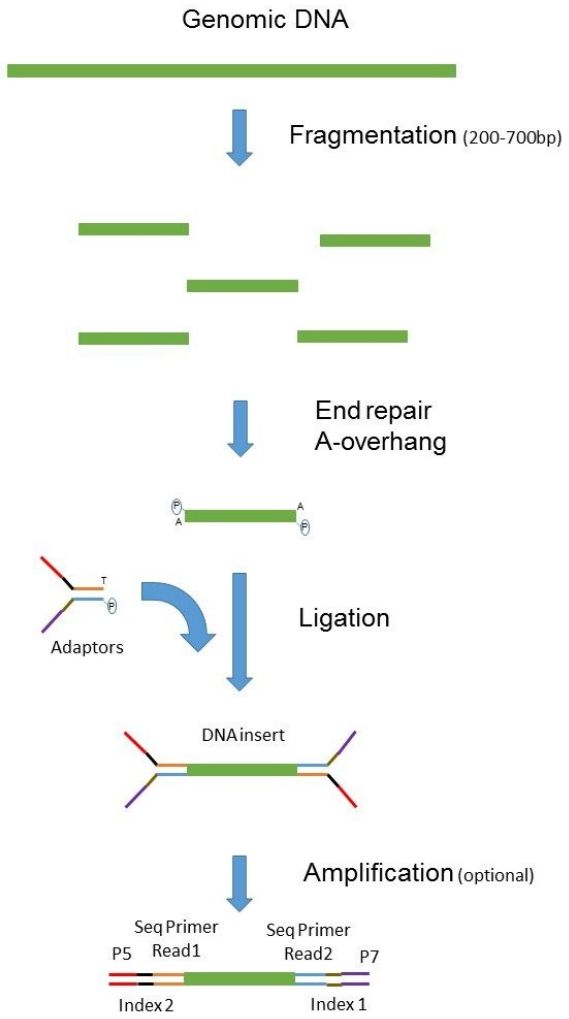


Illumina sequencing – from image to sequence

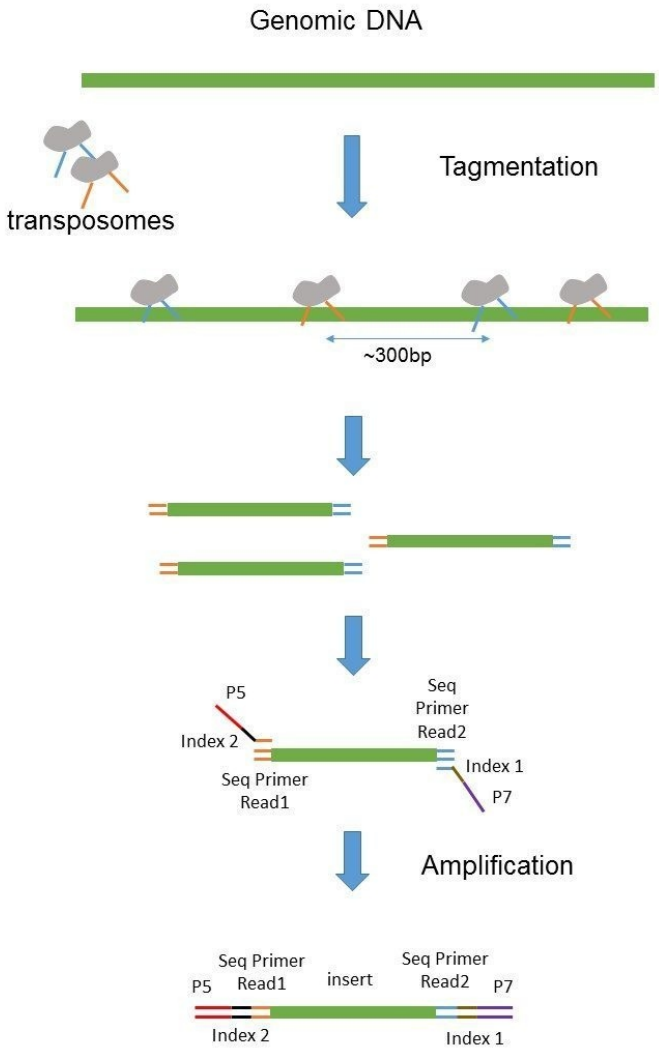
The identity of each base of a cluster is read off from sequential images

Paired-end sequencing

"Classical" paired end library (illumina)



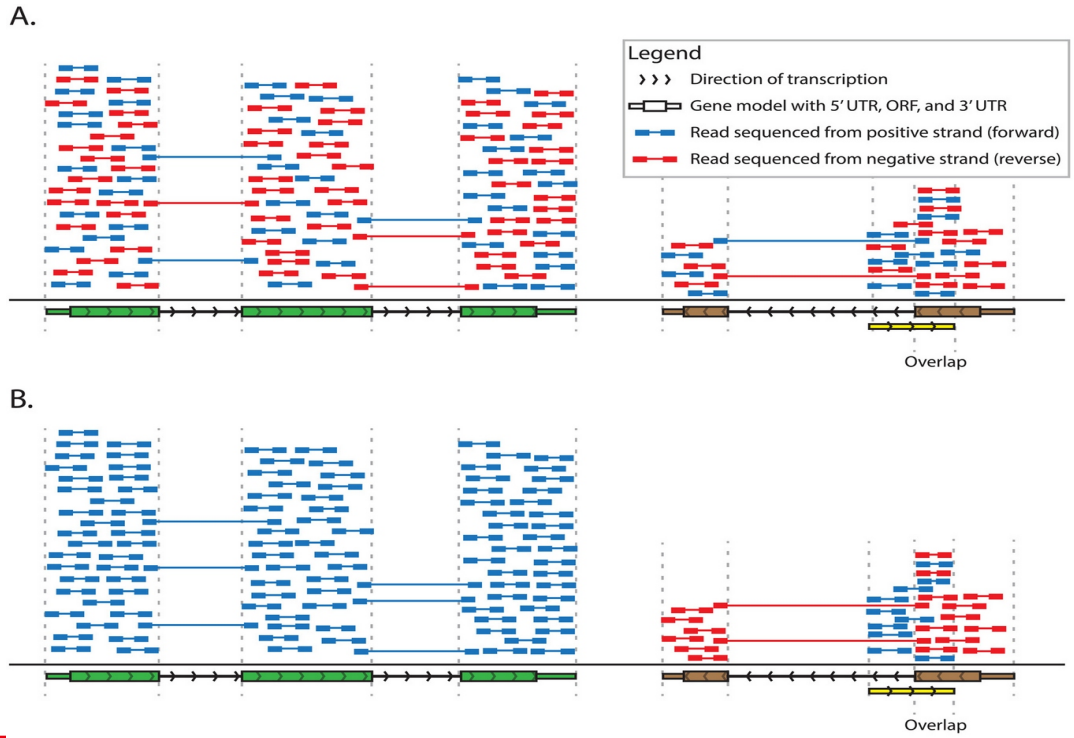
"Nextera" paired end library



Source : France Genomique

Stranded vs Unstranded Sequencing

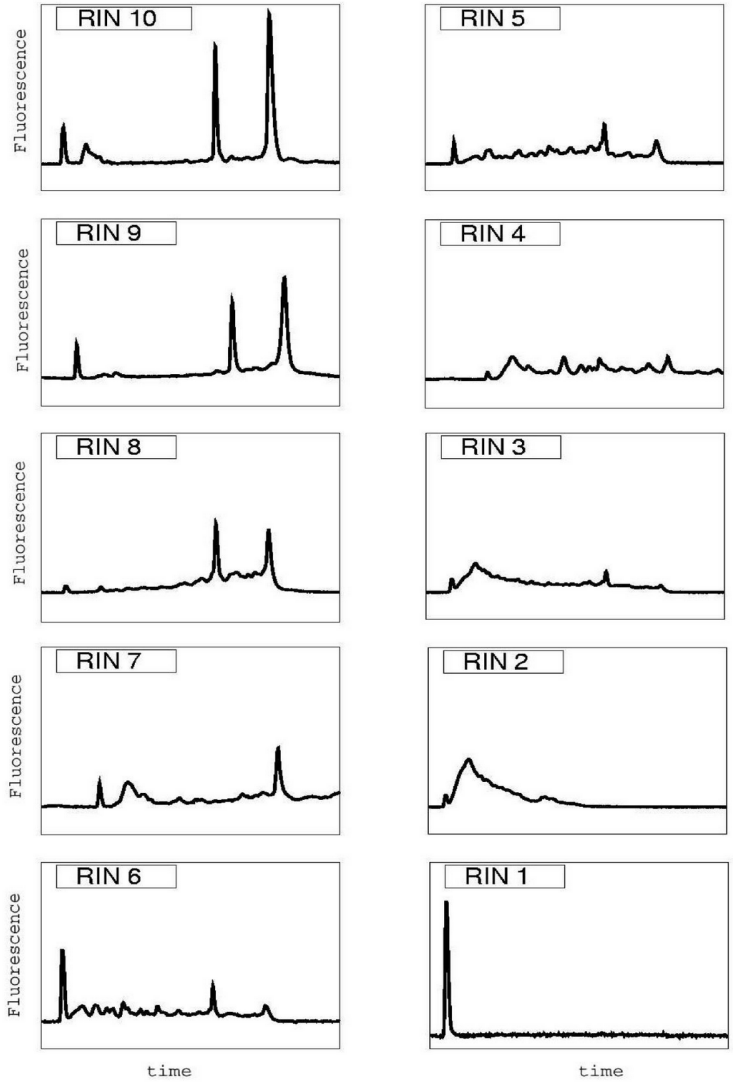
- A substantial proportion of the genome is transcribed in both directions (~8% in *Homo sapiens*)
- Read strand information = quantify expression on overlaps
- Achieved by ligating different adaptors to the 5' and 3' ends



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

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)



Schroeder *et al* BMC Mol Biol 2006

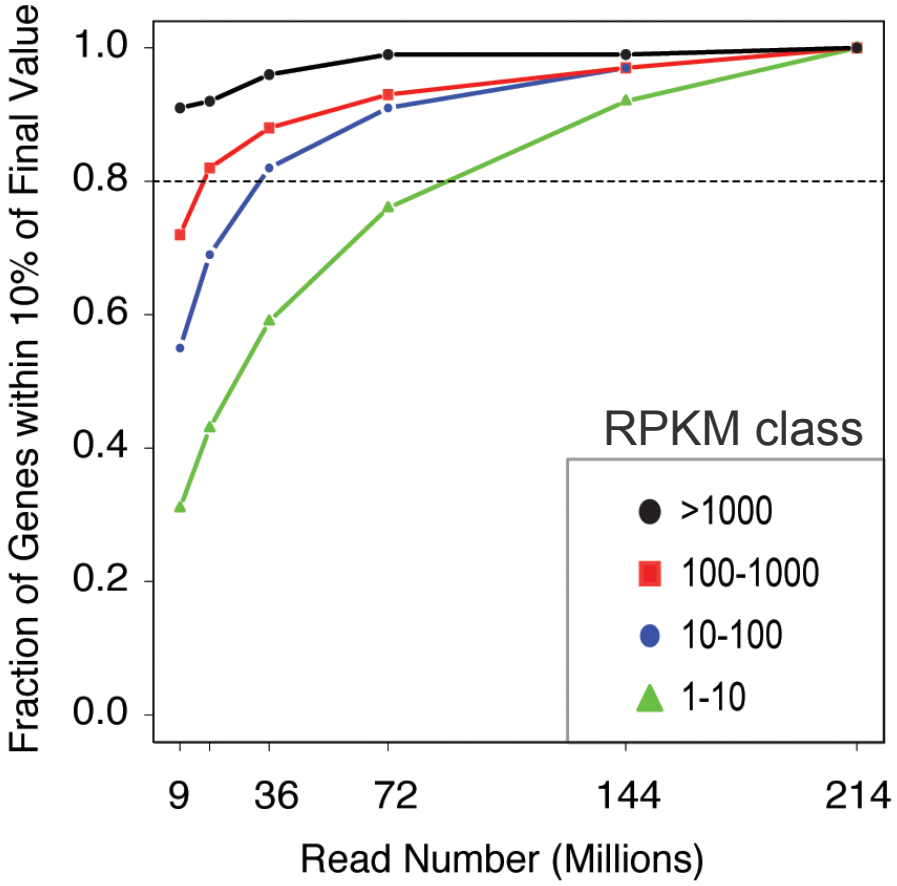
RNA-Seq Preparation - Purification

- **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**

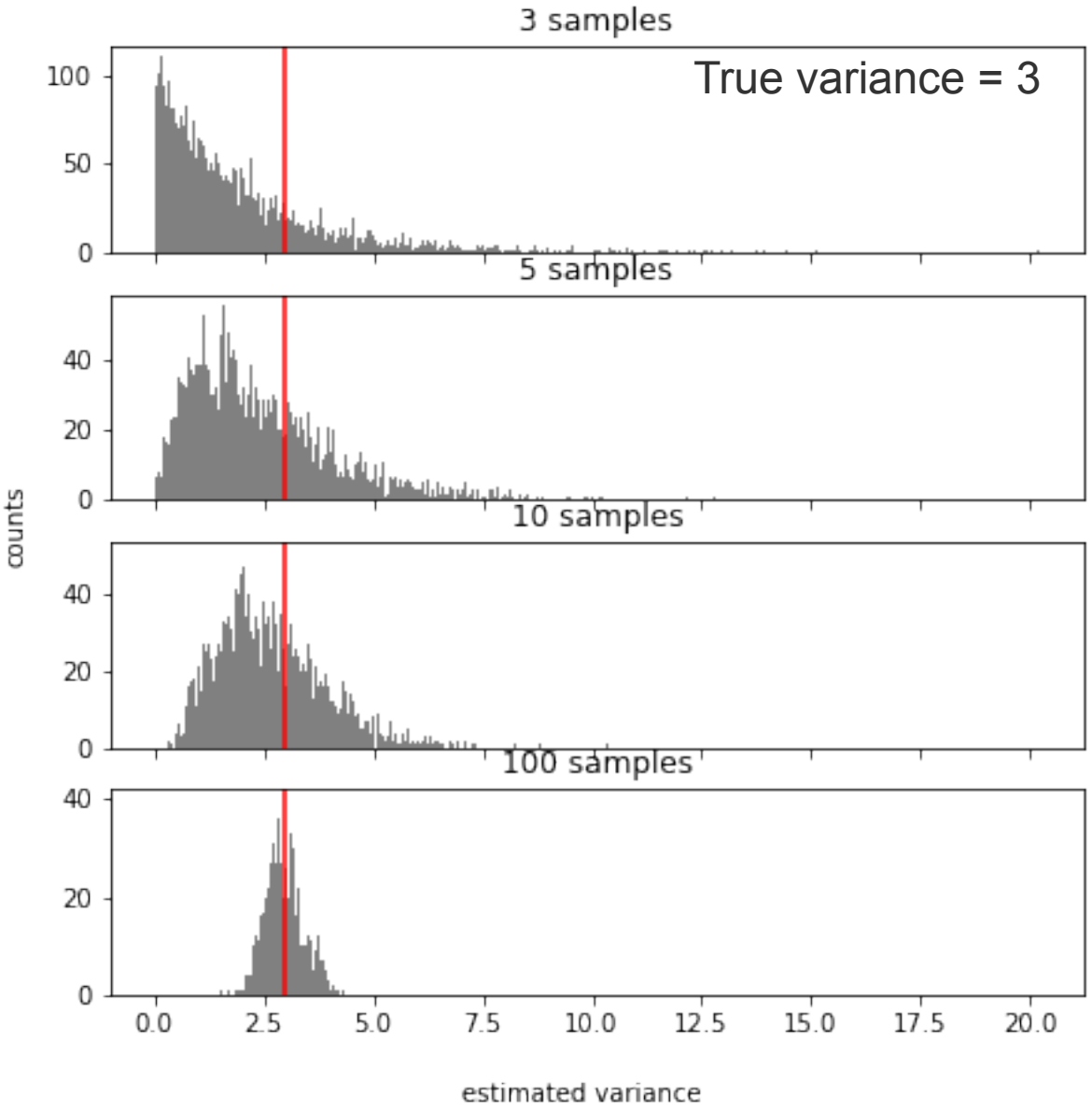


For rare events (isoforms, somatic mutations) much more depth is required

Not easy to know in advance

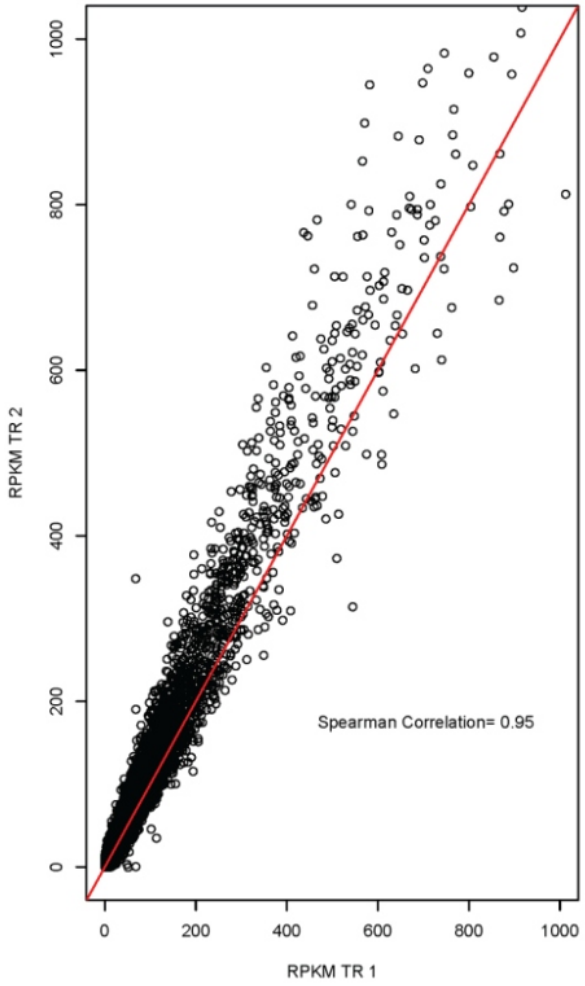
A User's Guide to the Encyclopedia of DNA Elements (ENCODE) PLoS 2011

Replicates - estimating a biological variance

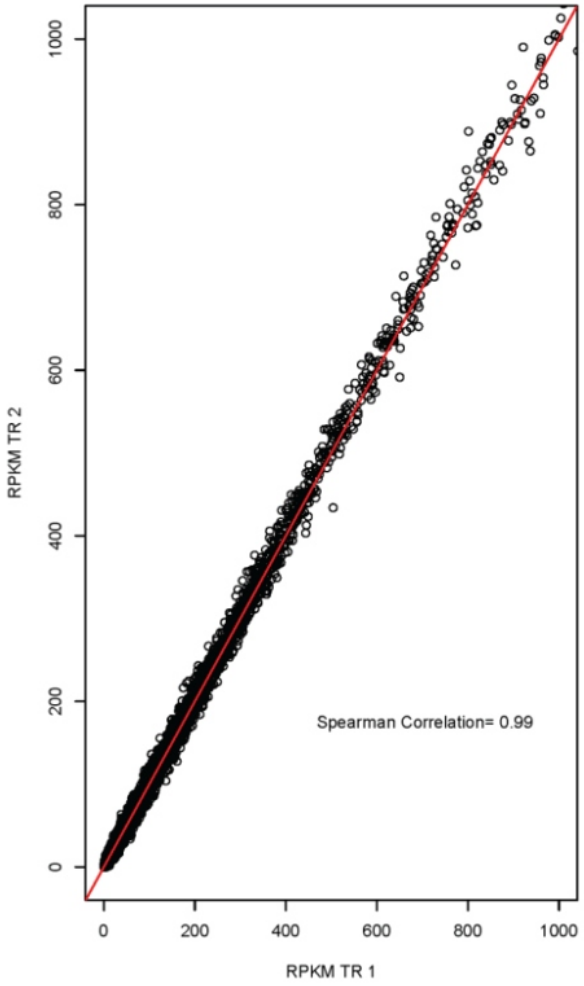


What does this tell you about the ideal number of replicates for your RNA-Seq experiment?

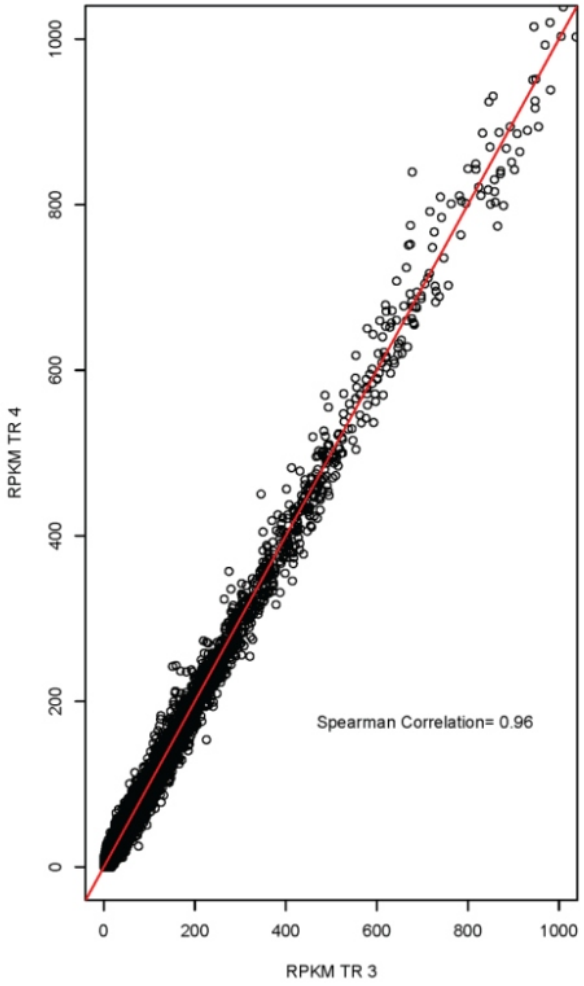
Technical replicates – generally not required



D. simulans male heads BR2



D. melanogaster female heads BR2

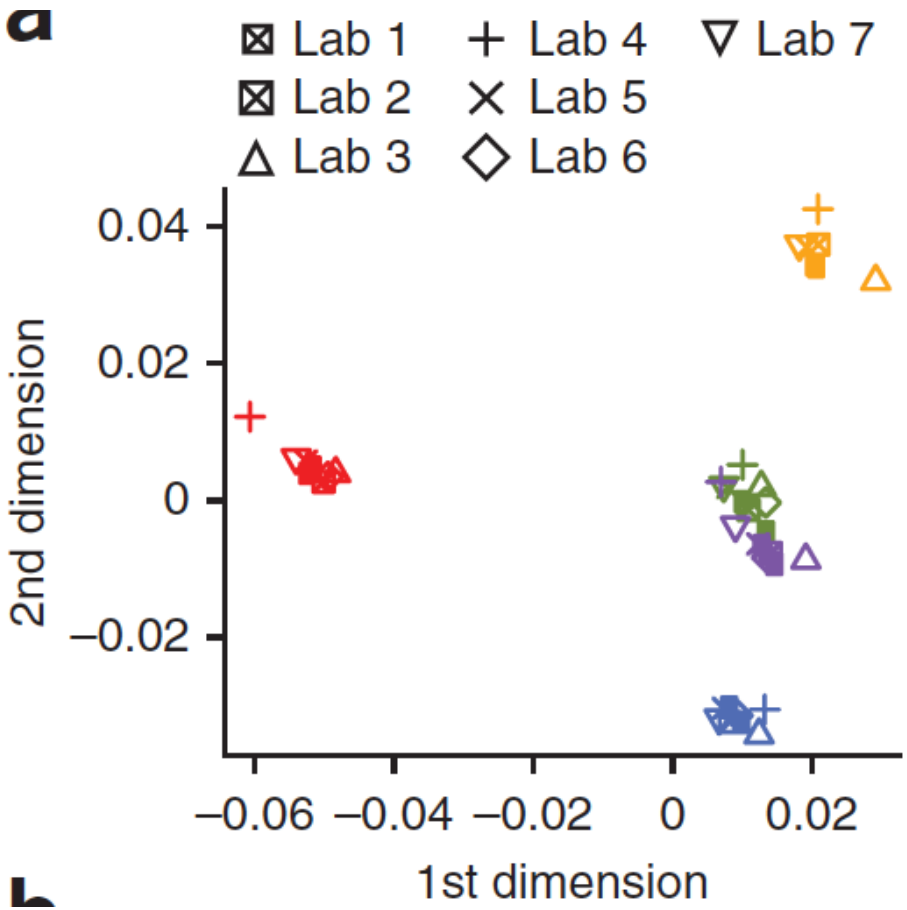


c167 cell line

Lauren M McIntyre et al., BMC Genomics 2011

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

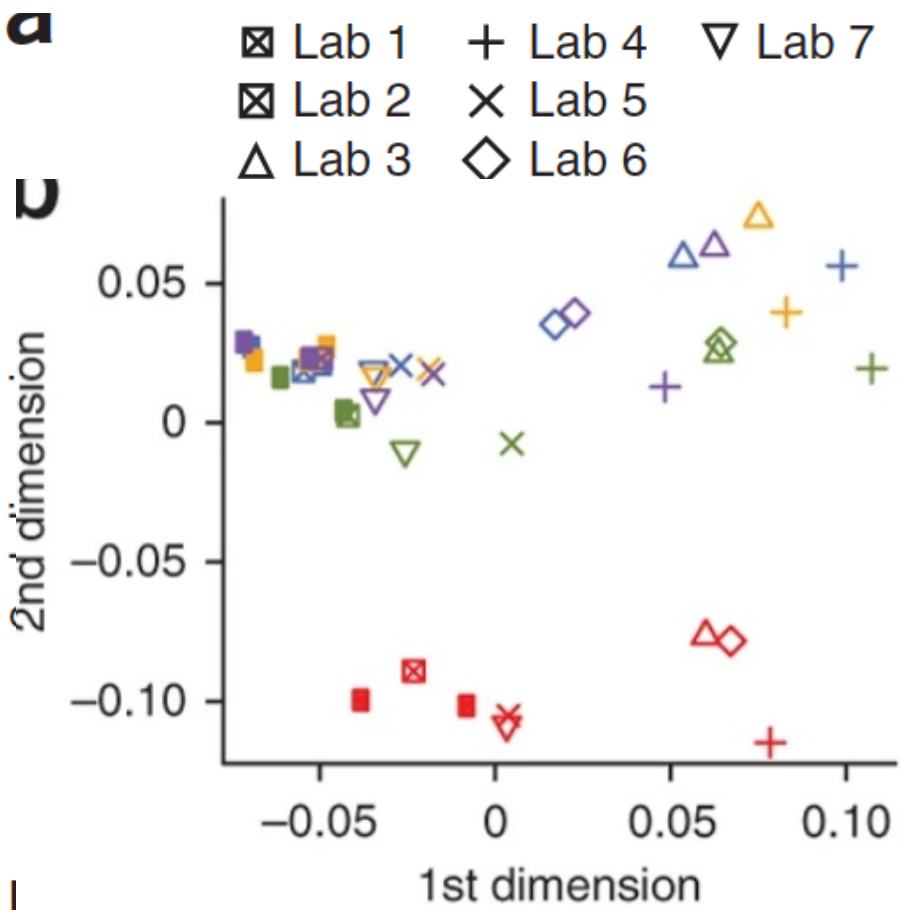
Affiliations | Contributions | Corresponding authors

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 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

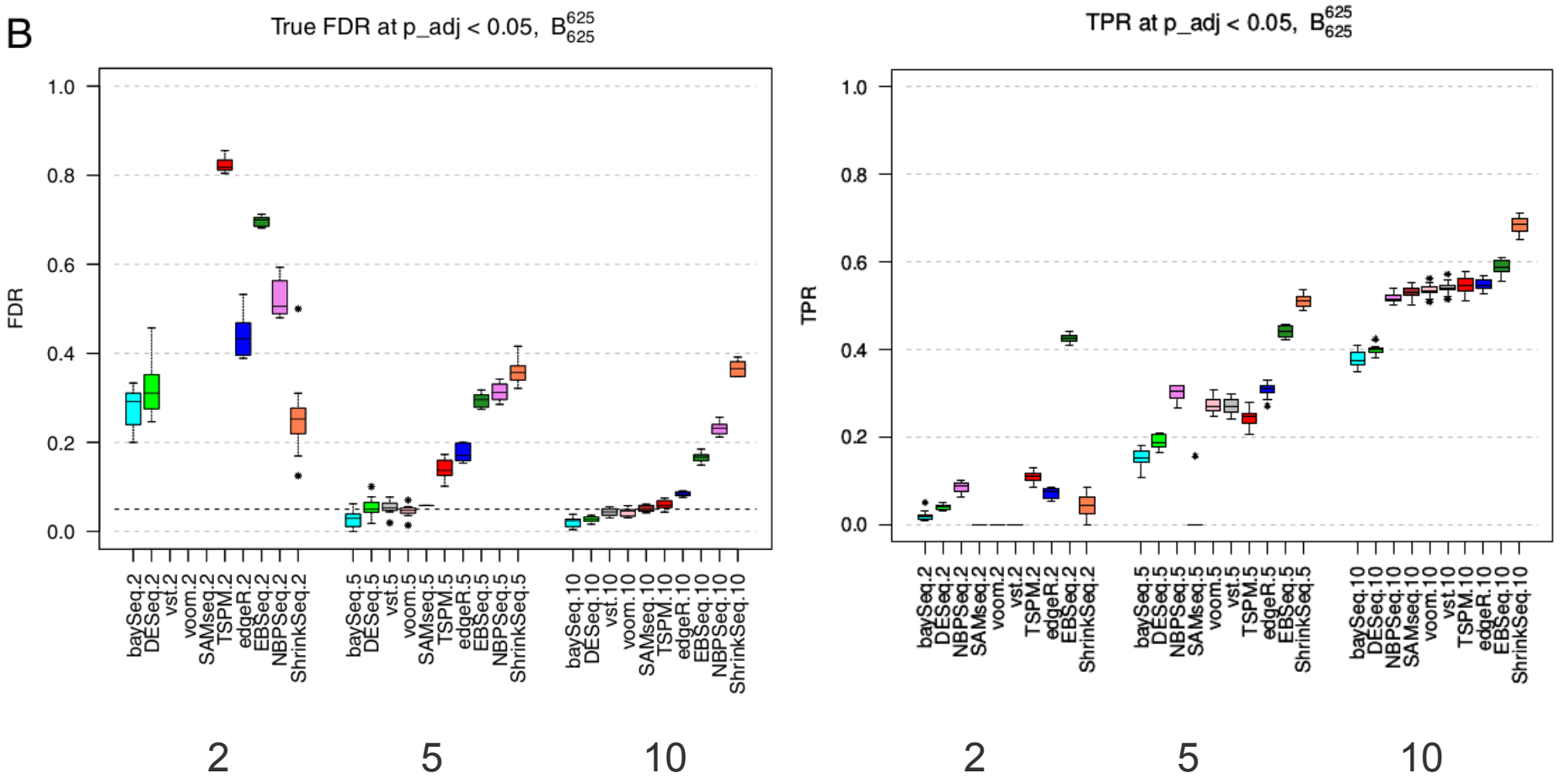
Affiliations | Contributions | Corresponding authors

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**

Transcript level : ~~reproducible~~.

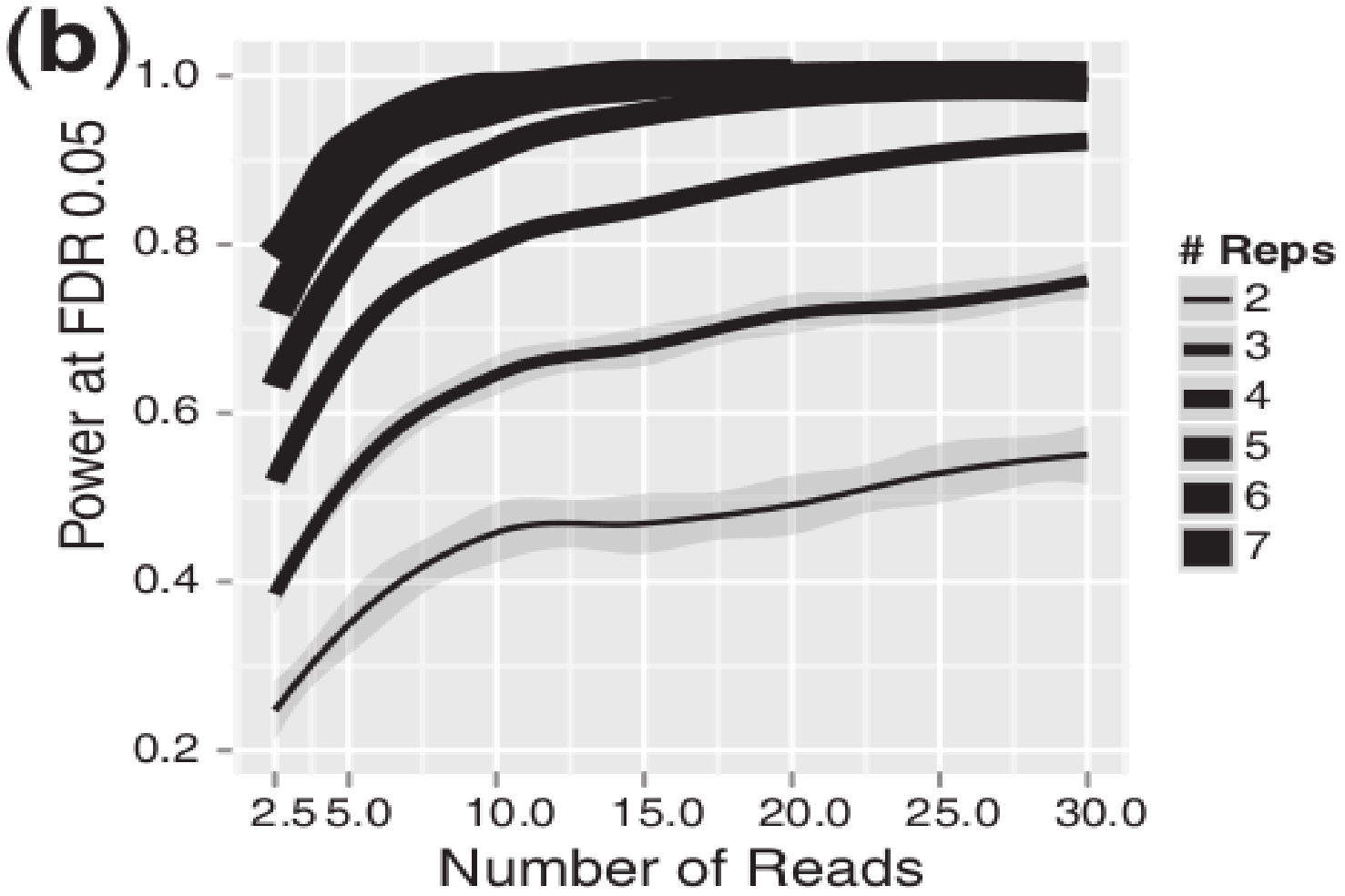
CHECK FOR BATCH EFFECT!

Biological replicates



Soneson & Delorenzi 2013 BMC Bioinformatics

More reads or more replicates ?

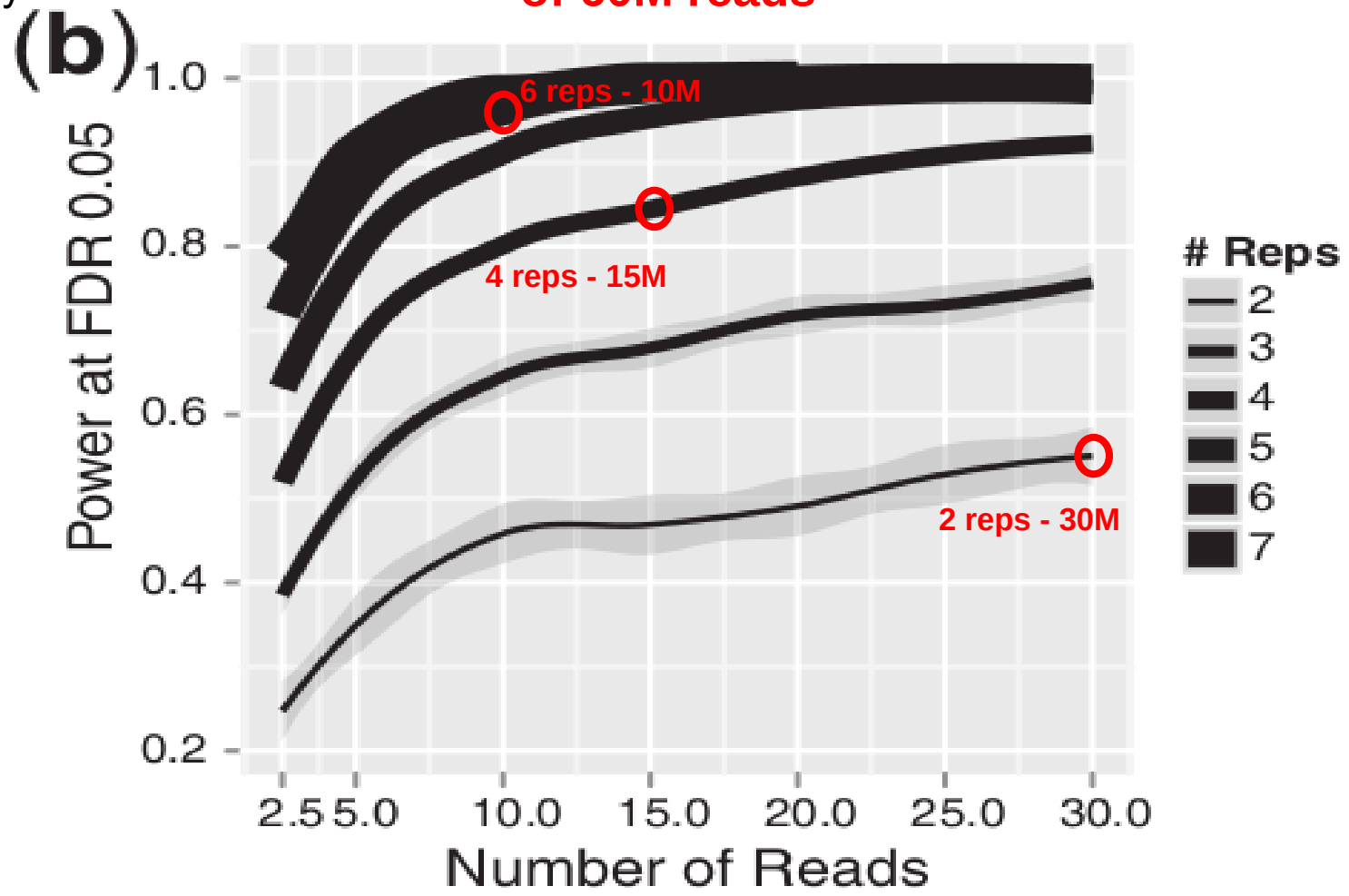


From Liu et al. 2014. RNA-seq differential expression studies: more sequence or more replication?

More reads or more replicates ?

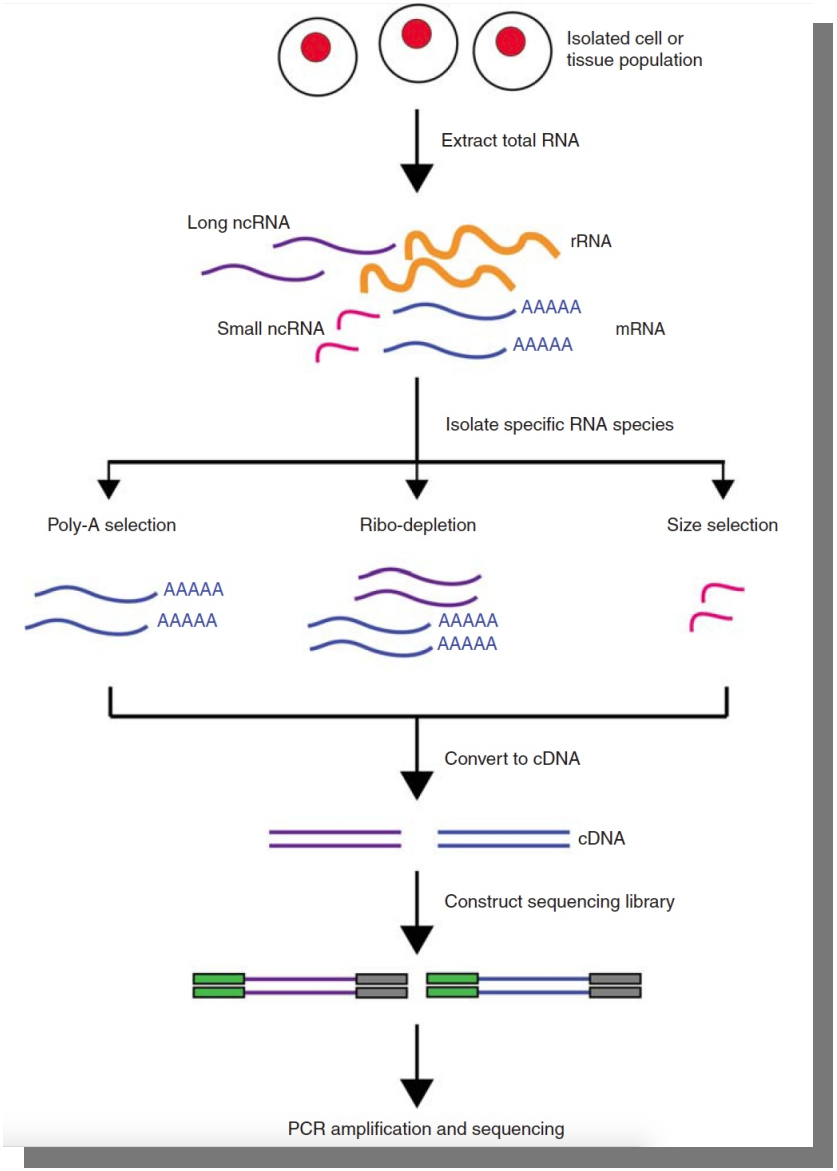
Generally more replicates should be preferred if they can be obtained.

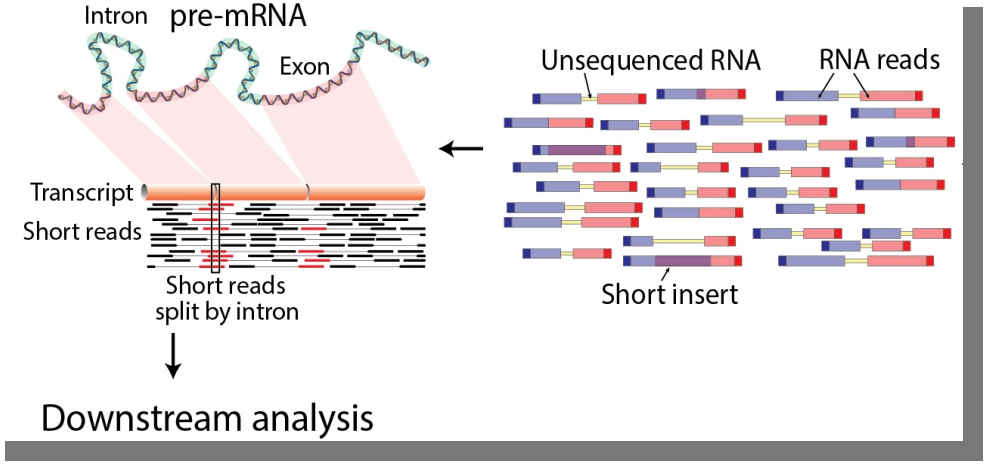
for a fixed total of 60M reads



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

Basic RNA-Seq Protocol Overview





RNA-seq data analysis – basic analysis

Sequencing facility

RNA-seq data analysis – with trimmed reads

Sequencing facility

RNA-seq data analysis – main QC steps

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

...

RNA-seq data analysis – main QC steps

Sequencing facility

Raw reads

QC

trimming/ filtering

Trimmed/filtered reads

QC

mapping

Exclude outlier

Mapped reads

QC

normalization across libraries

Expression matrix

expression quantification

QC

Normalized counts

...

Differential Expression analysis

Table of DE genes/transcript

Pathway enrichment analysis

...

RNA-seq data analysis – not only differential expression

Sequencing facility

Raw reads

trimming/ filtering

Trimmed/filtered reads

mapping

Mapped reads

variant calling

isoform description

...

expression quantification

Expression matrix

normalization across libraries

Normalized counts

...

Differential Expression analysis

Table of DE genes/transcript

Pathway enrichment analysis

...

RNA-seq data analysis – de novo transcriptome assembly

Sequencing facility

Raw reads

trimming/ filtering

Trimmed/filtered reads

assembly

Assembled transcriptome

variant calling

isoform description

...

expression quantification

Expression matrix

normalization across libraries

Normalized counts

...

Differential Expression analysis

Table of DE genes/transcript

Pathway enrichment analysis

...

RNA-seq data analysis – program for this course

Sequencing facility

Raw reads

QC

trimming/ filtering

Trimmed/filtered reads

QC

mapping

Mapped reads

QC

expression quantification

Expression matrix

normalization across libraries

QC

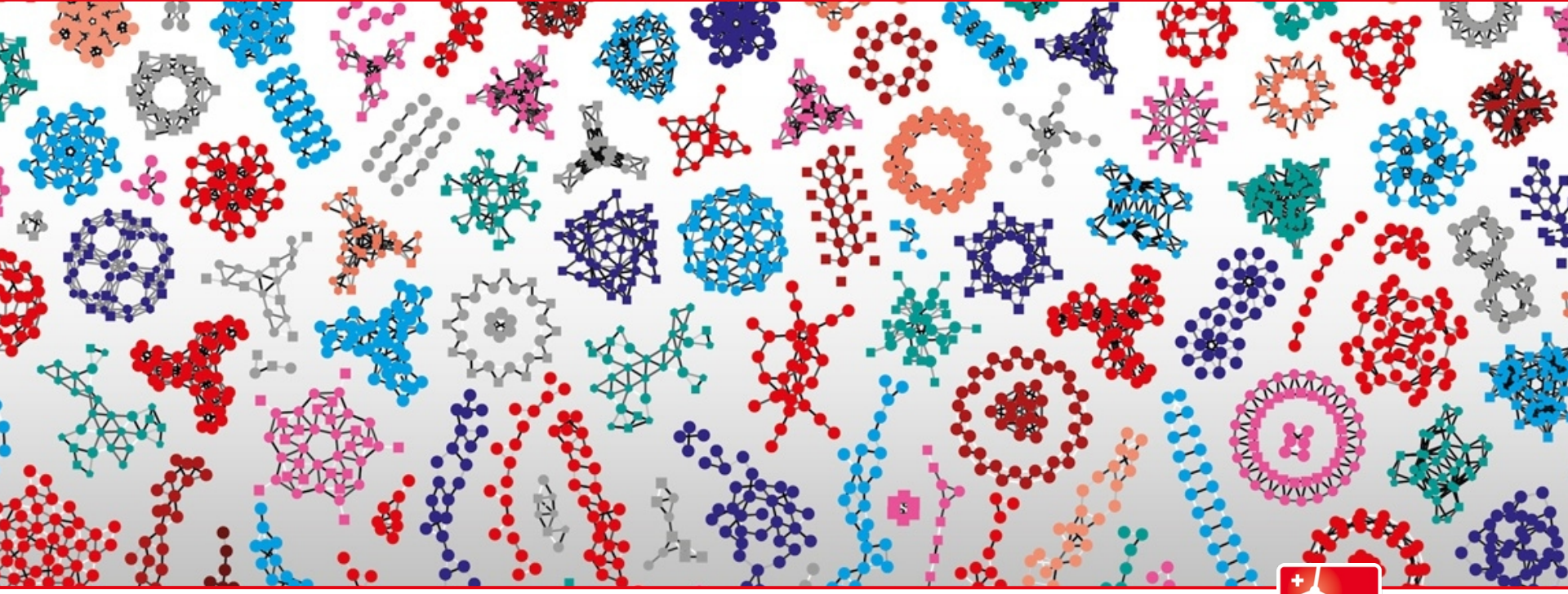
Normalized counts

Differential Expression analysis

Table of DE genes/transcript

Pathway enrichment analysis

...



Swiss Institute of Bioinformatics

Contributors:

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