

Swiss Institute of
Bioinformatics

Introduction to RNA-Seq – Overview

Wandrille Duchemin

DNA



Transcription

Nucleic acids
everywhere

RNA

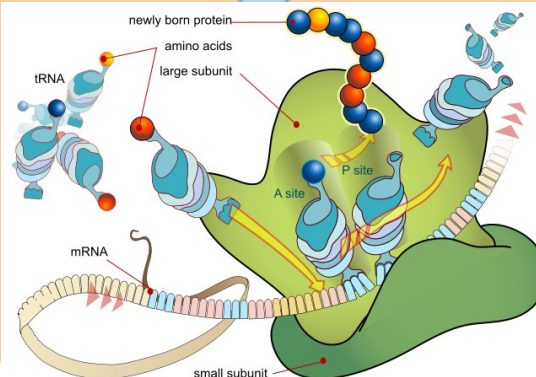


NUCLEUS

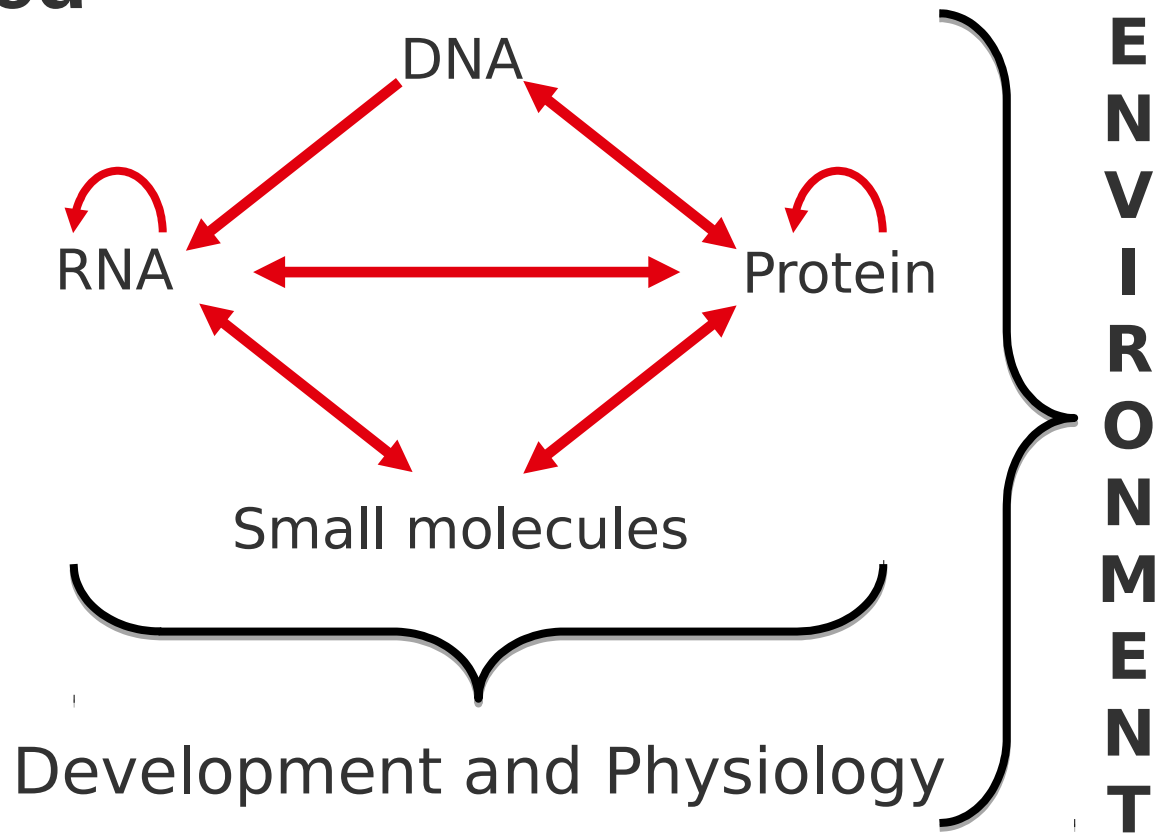
Translation

CYTOPLASM

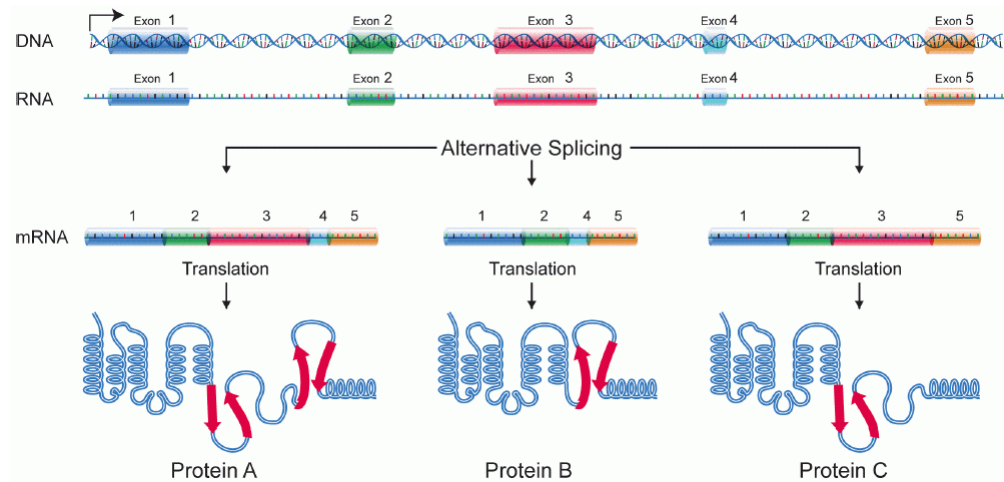
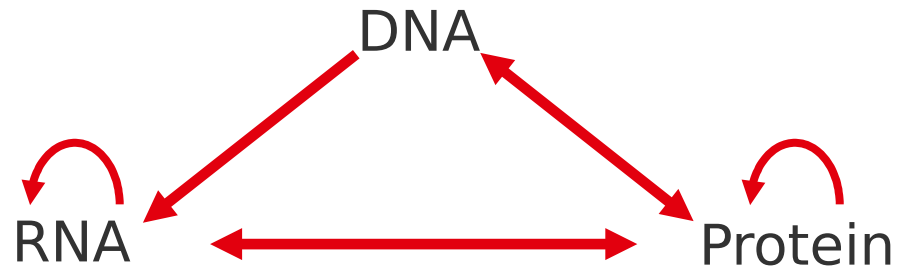
Protein



Dogma Revisited



Dogma Revisited



Dogma Revisited

~20,000 mammalian genes



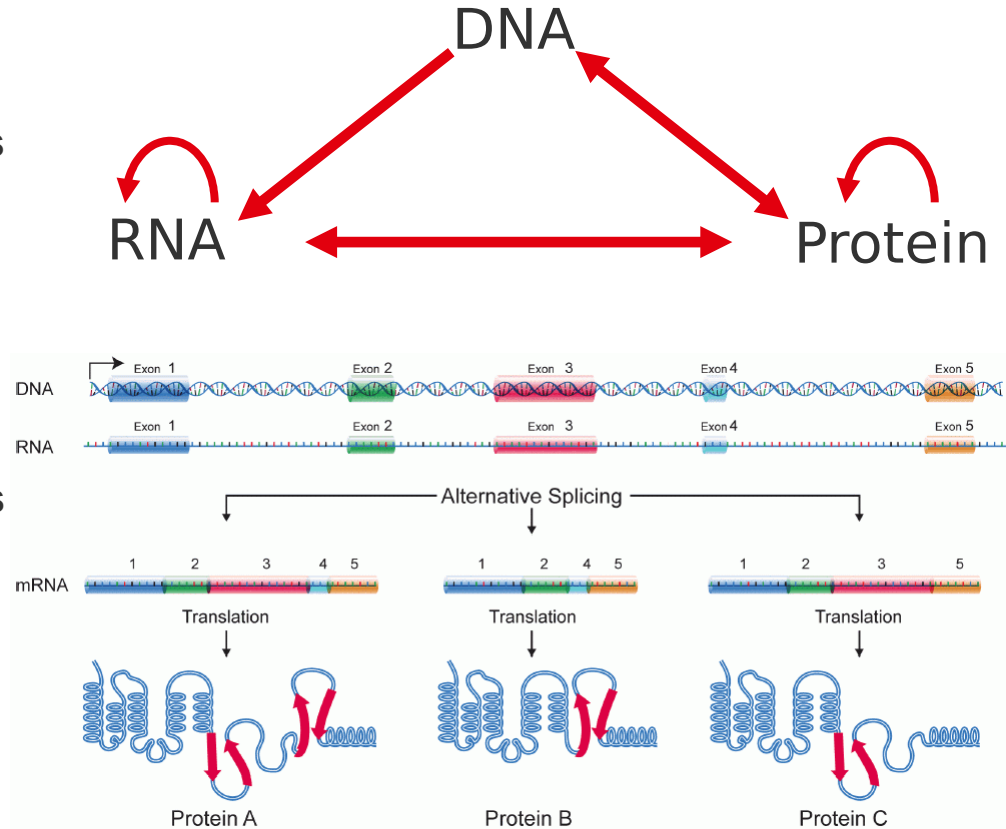
>>100,000 (?) transcripts



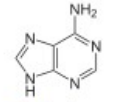
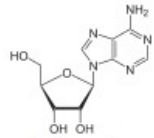
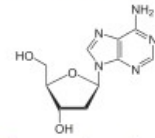
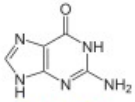
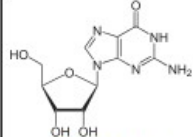
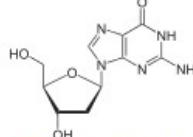
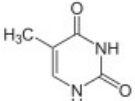
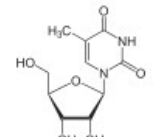
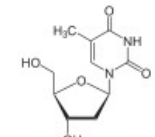
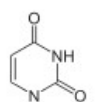
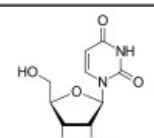
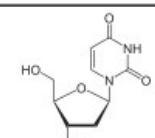
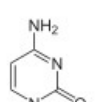
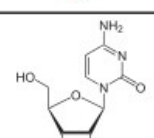
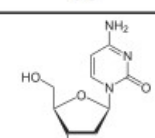
>>>1,000,000 (??) proteins



Cellular Diversity



Nucleobases – building blocks of nucleic acids

Nitrogenous base	Ribonucleoside	Deoxyribonucleoside
 Adenine	 Adenosine A	 Deoxyadenosine dA
 Guanine	 Guanosine G	 Deoxyguanosine dG
 Thymine	 5-Methyluridine m⁵U	 Thymidine dT
 Uracil	 Uridine U	 Deoxyuridine dU
 Cytosine	 Cytidine C	 Deoxycytidine dC

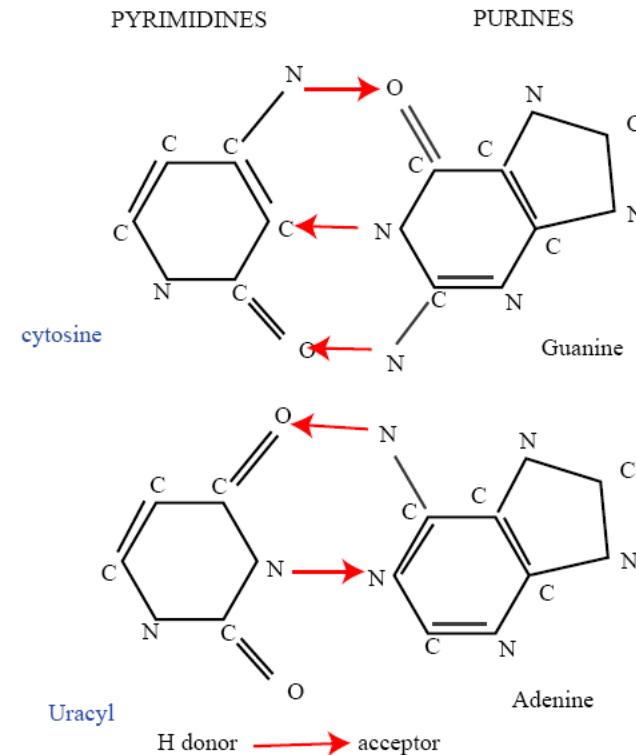


Image source : wikipedia.org

What (and why) are we sequencing?

■ Genomics

- Whole genome/exome sequencing (WGS, WES)
 - Single nucleotide polymorphisms
 - Copy number variations
 - Structural variations
- DNaseI hypersensitivity analysis
 - Chromatin structure, binding of transcription factors
- Hi-C (High throughput Chromosome Conformation Capture)
 - Chromatin structure (looping)

■ Epigenomics

- Bisulphite sequencing (DNA methylation)
- ChIP-seq, ATAC-Seq

What (and why) are we sequencing?

■ Transcriptomics

- Total RNA sequencing
- PolyA+ RNA sequencing
- Ribozero
 - Genome annotation
 - Allele-specific expression
 - Transcript identification (lncRNAs)
 - Alternative splicing
 - Gene Fusion
- 5`/3` tag sequencing
 - Gene expression analysis
- Small RNA sequencing
 - Identification and expression analysis of small RNAs (eg miRNAs)
- Metabolic labeling – RNA Seq
 - Analysis of mRNA decay
- Crosslinking and immunoprecipitation of RNA-binding proteins (CLIP)
 - Binding sites of RBPs
- Ribosome footprinting
 - Analysis of mRNA translation
- Ribometh-seq and similar
 - Detection of RNA modifications
- **Single-cell transcriptomics**

Imagination is the limit!

What (and why) are we sequencing?

■ Transcriptomics

- Total RNA sequencing
- PolyA+ RNA sequencing
- Ribozero
 - Genome annotation
 - Allele-specific expression
 - Transcript identification (lncRNAs)
 - Alternative splicing
 - Gene Fusion
- 5`/3` tag sequencing
 - Gene expression analysis
- Small RNA sequencing
 - Identification and expression analysis of small RNAs (eg miRNAs)
- Metabolic labeling – RNA Seq
 - Analysis of mRNA decay
- Crosslinking and immunoprecipitation of RNA-binding proteins (CLIP)
 - Binding sites of RBPs
- Ribosome footprinting
 - Analysis of mRNA translation
- Ribometh-seq and similar
 - Detection of RNA modifications
- **Single-cell transcriptomics**

Imagination is the limit!

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

Beginning of sequencing: Sanger's method

- **Could be applied to wide range of fragment lengths, used initially in whole genome sequencing**
 - Extract, purify and fragment genomic DNA
 - Make clone library
 - Sequence clone
 - Assemble the genome
 - Fill gaps

- **Computationally intensive for alignment and assembly**



Frederick Sanger



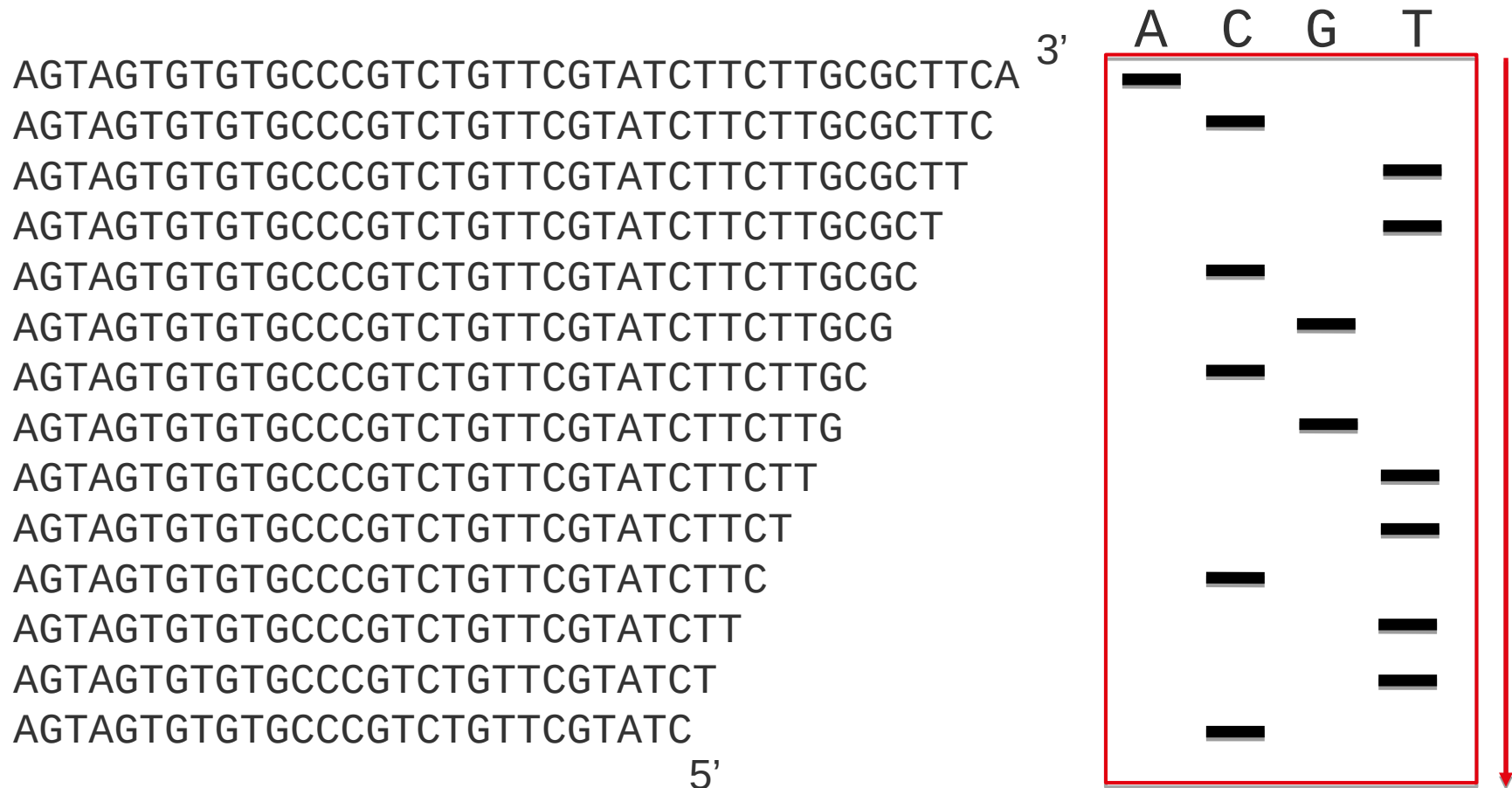
Gene Myers



David Haussler

Sanger sequencing method

- Uses chain terminating nucleotides which are stochastically incorporated
- Gel electrophoresis allows one to 'read' the sequence



Underlying principle

■ Sequencing by synthesis

- Sanger/Dideoxy chain termination – 1st generation
- Pyrosequencing (Roche/454) – 2nd generation
- Reversible terminator (Illumina) – 2nd generation
- Ion torrent (Life Technologies) - 3rd generation
- Zero Mode Waveguide (Pacific Biosciences) - 3rd generation

■ Sequencing by Oligo Ligation Detection

- SOLiD (Applied Biosystems) – 2nd generation

■ Direct reading of DNA/RNA sequence

- Nanopore sequencing – 3rd generation

Major technologies

illumina

Short reads
High throughput
Paired reads

life
technologies™

SOLiD

Short reads
Medium throughput
Short run time
Library prep difficult



PACIFIC
BIOSCIENCES™

Long reads
Individual molecules
Base modifications
High, random error rate
Low throughput

Oxford
NANOPORE
Technologies

Long reads
Individual molecules
Error rate!

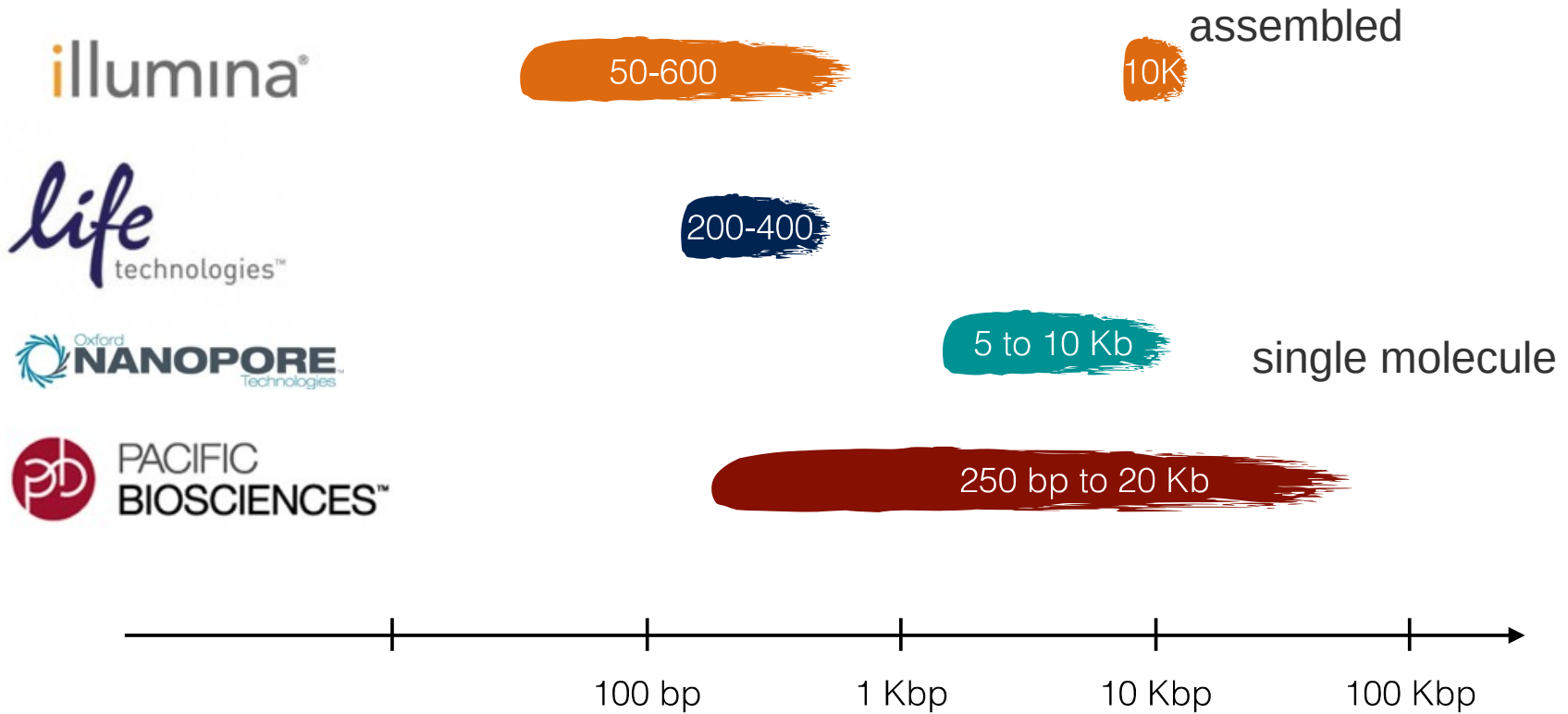
Roche

454

SEQUENCING

retired

Read lengths



Challenges for RNA-Seq

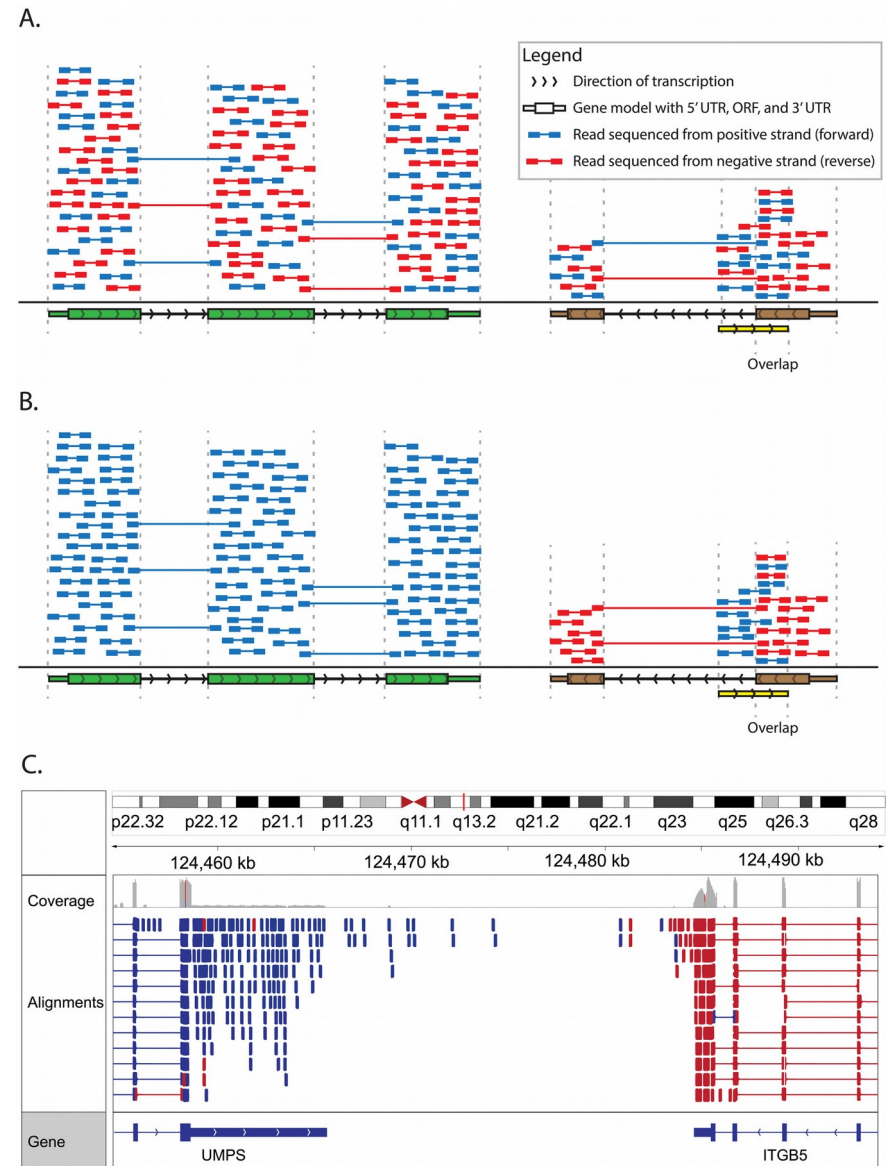
- **RNA is highly labile and exposed to degradation enzymes**
 - Hydroxyl group of ribose - reactive, uracil - less stable than thymine, RNAses
 - How to preserve RNA integrity?

- **RNA molecules are highly diverse**
 - High dynamic range of expression, from hundreds of thousands per cell (rRNAs, small RNAs) to very few (mRNAs)
 - Large differences in lengths, from tens to millions of nucleotides
 - How to capture the species of interest?
 - How to preserve strand information?
 - How to preserve relative abundances?
 - How to normalize inferred abundance to compare between samples?

- **Sequencing reads need to be mapped**
 - To genome or to transcriptome?
 - Which reference to use?
 - How to handle reads that come from exon-exon boundaries?
 - How to handle multi-mapping sequencing reads?

Stranded vs Unstranded Sequencing

- A substantial proportion of the genome is transcribed in both directions (~8% in *Homo sapiens*)
- Use a protocol that preserves the strand information to make it easier to quantify the expression from overlapping loci on opposite strands
- Achieved by ligating different adaptors to the 5' and 3' ends before cDNA synthesis



Paired end Sequencing

- Direct sequencing method: Input genomic DNA is fragmented by methods such as nebulization, hydrodynamic shearing, sonication...



Fragments of <800pb are selected



The ends are repaired, 3' ends are adenylated and paired-end adapters are added



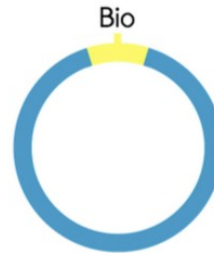
The product is amplified by PCR, purified and sequenced from both ends

Mate pair Sequencing

- For mate pair sequencing, the DNA is fragmented into 2-5kb segments



The labeled fragments are circularized



Fragments with the biotin labels are captured with streptavidin, they are enriched, end-repaired, and ligated with adapters



The ends are repaired with biotin labeled dNTPs



and fragmented again into 400-600bp pieces

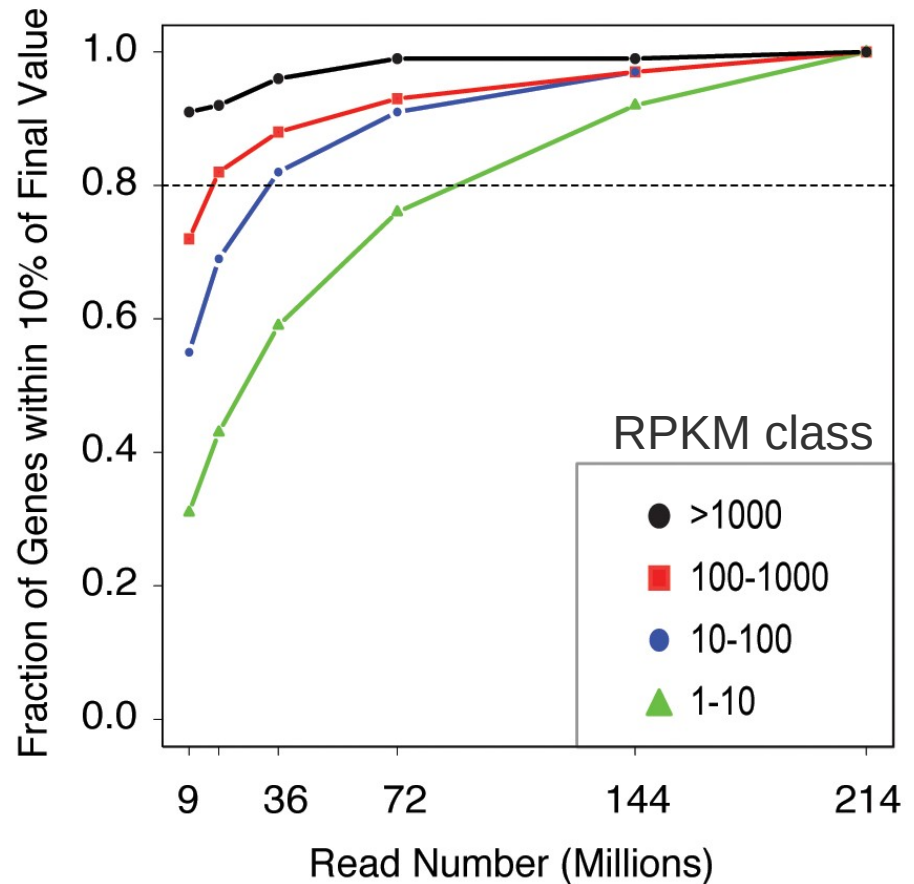


These fragments can now be sequenced from both ends

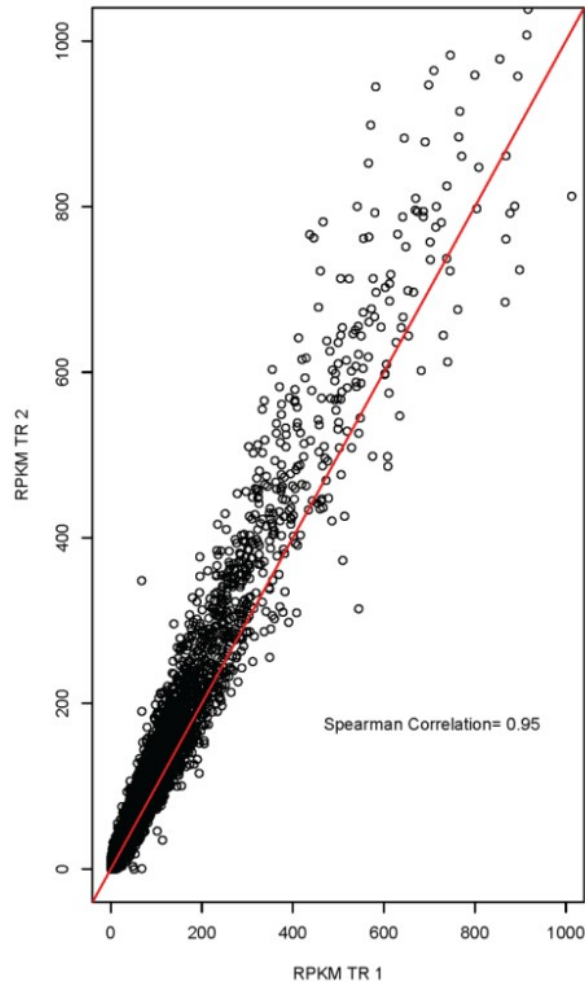


Sequencing depth

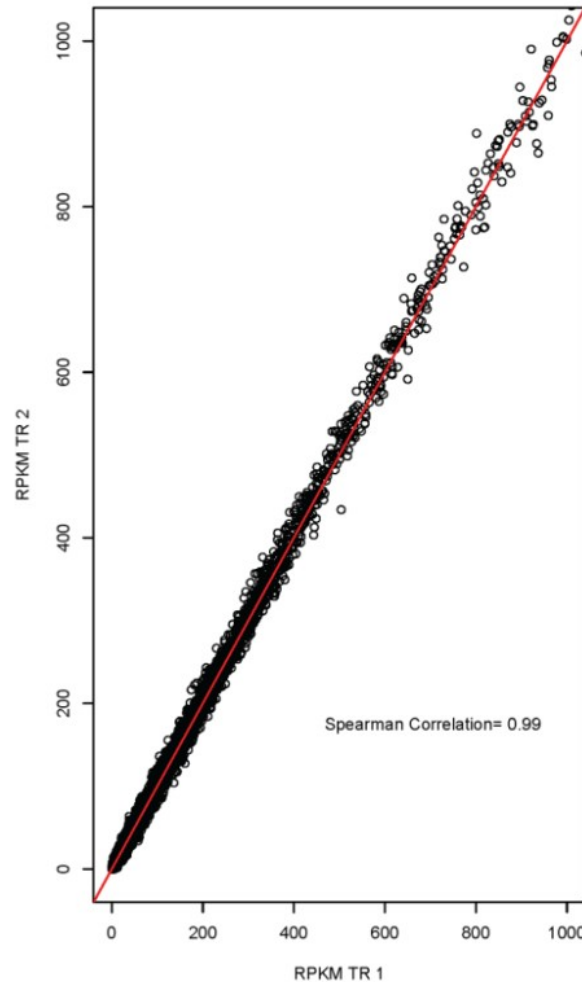
- Quantification of gene expression usually requires ~30-40 million reads
- To identify rare events (isoforms, somatic mutations) much more depth is required
- Not easy to know in advance, some experimentation is typically required



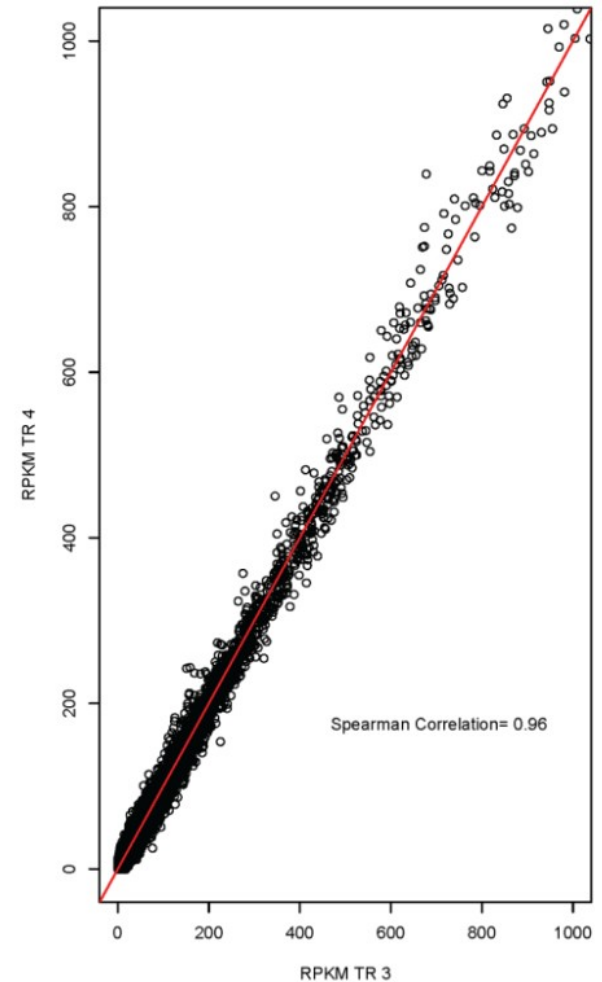
Technical replicates – generally not required



D. simulans male
heads BR2



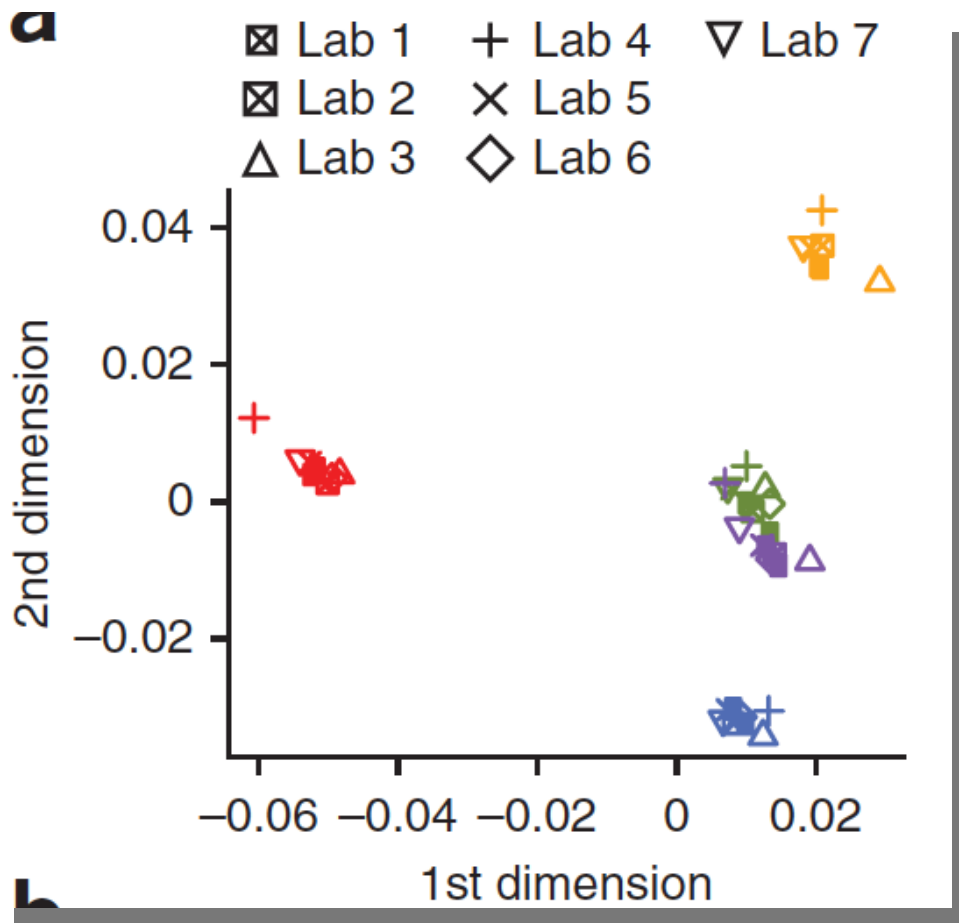
D. melanogaster
female heads BR2



c167 cell line

Experimental replicates – interlaboratory variation

- 7 labs, 48-113 samples, 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 Dermizakis & 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

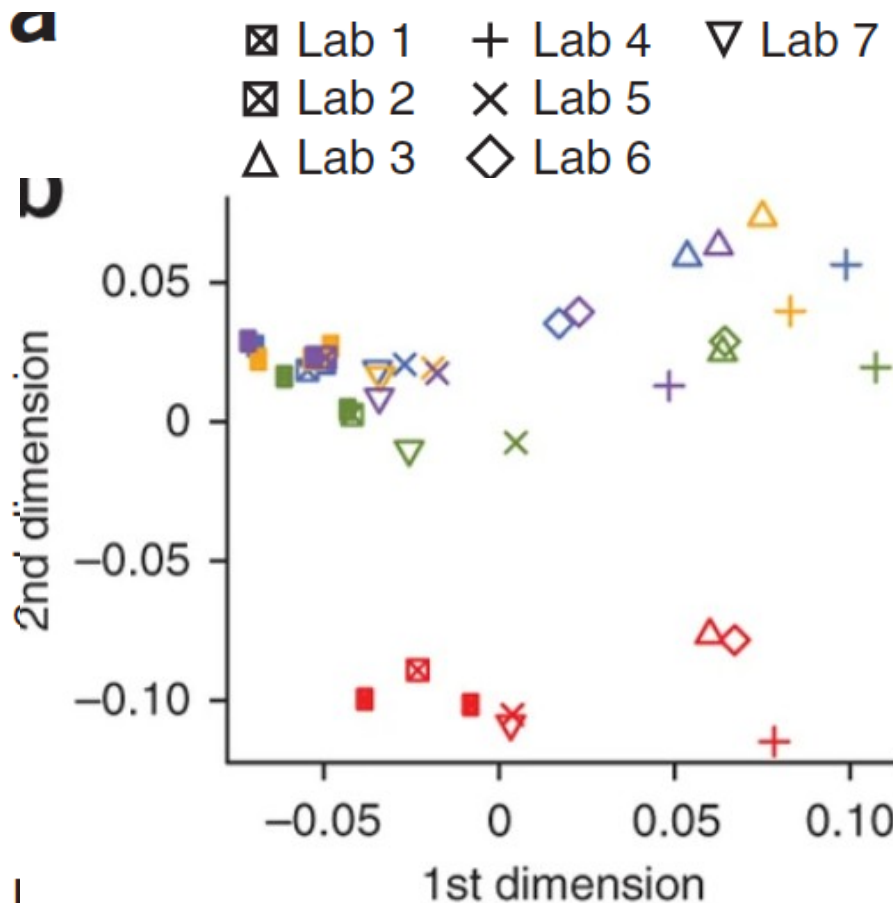
| Corrected online 08 November 2013

Exon level :

Downstream of RNA extraction, sample preparation is quite reproducible.

Experimental replicates – interlaboratory variation

- 7 labs, 48-113 samples, 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 Dermizakis & 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 :

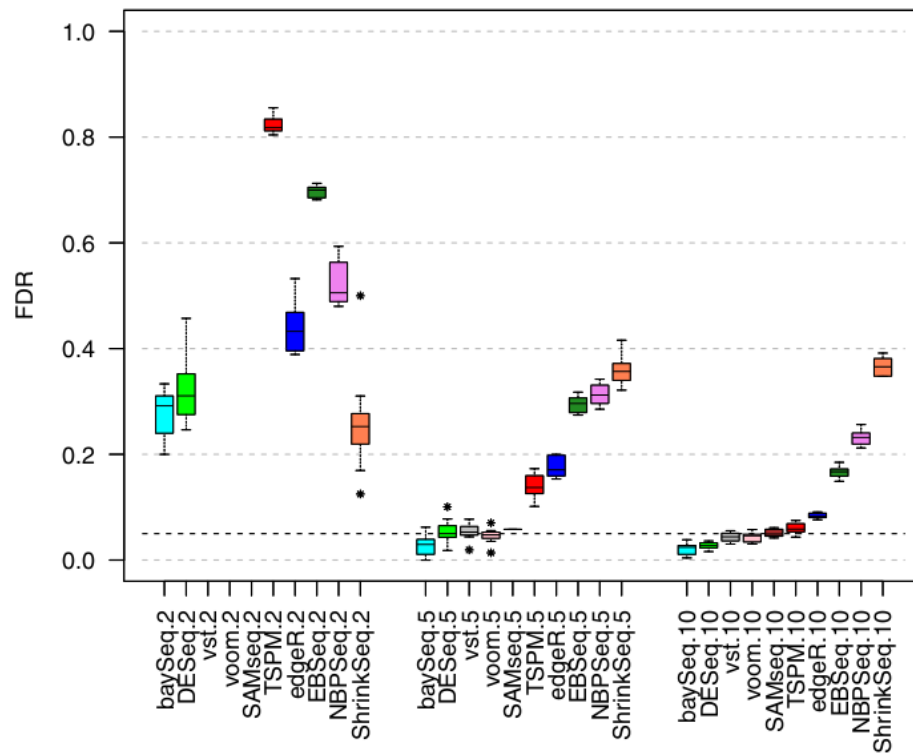
Downstream of RNA extraction, sample preparation is quite reproducible.

CHECK FOR BATCH EFFECT!

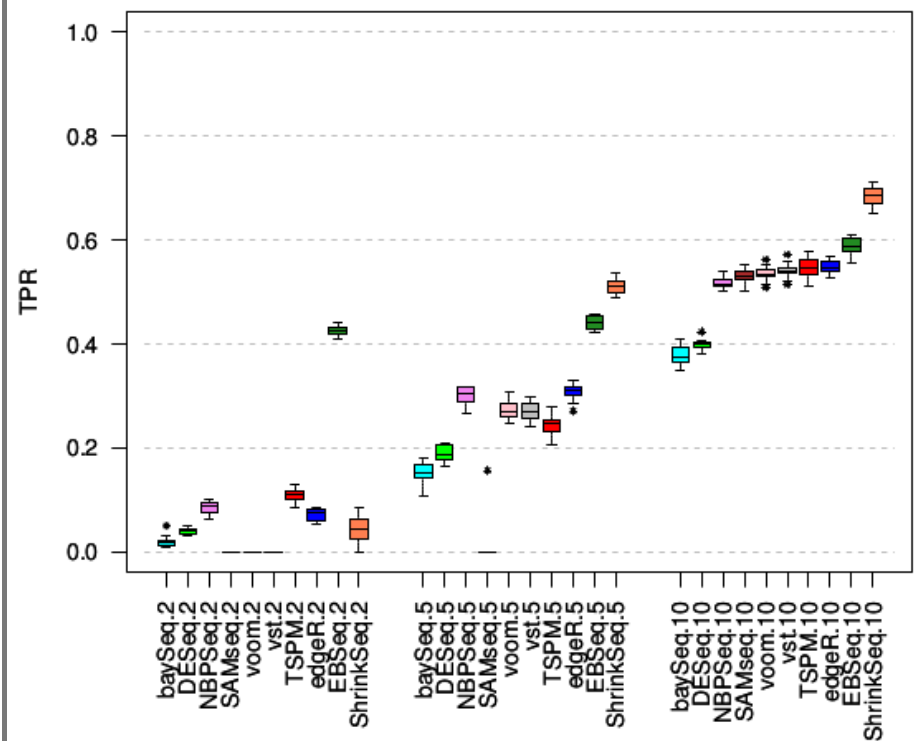
Biological replicates

B

True FDR at $p_{\text{adj}} < 0.05$, B_{625}^{625}

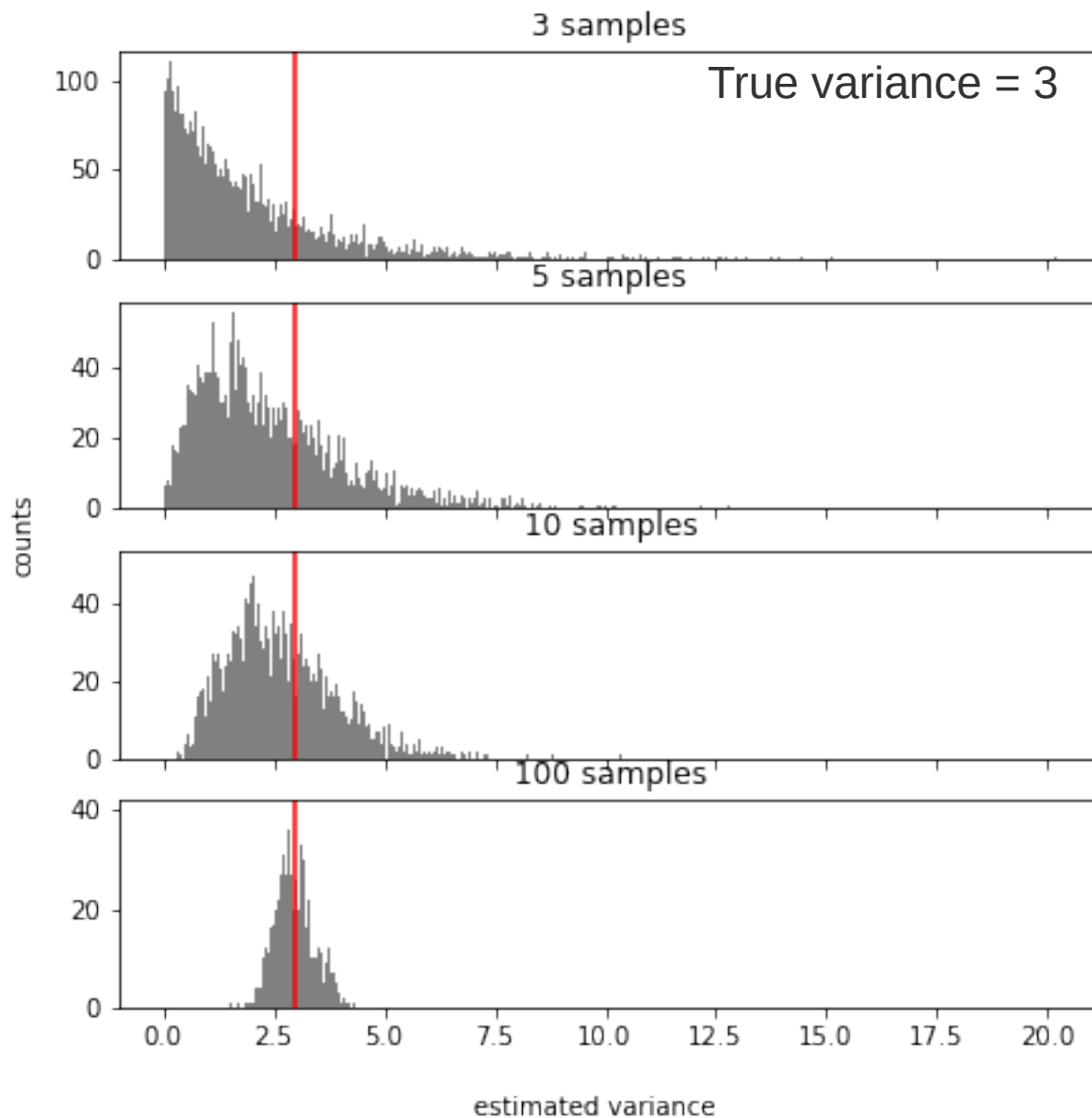


TPR at $p_{\text{adj}} < 0.05$, B_{625}^{625}



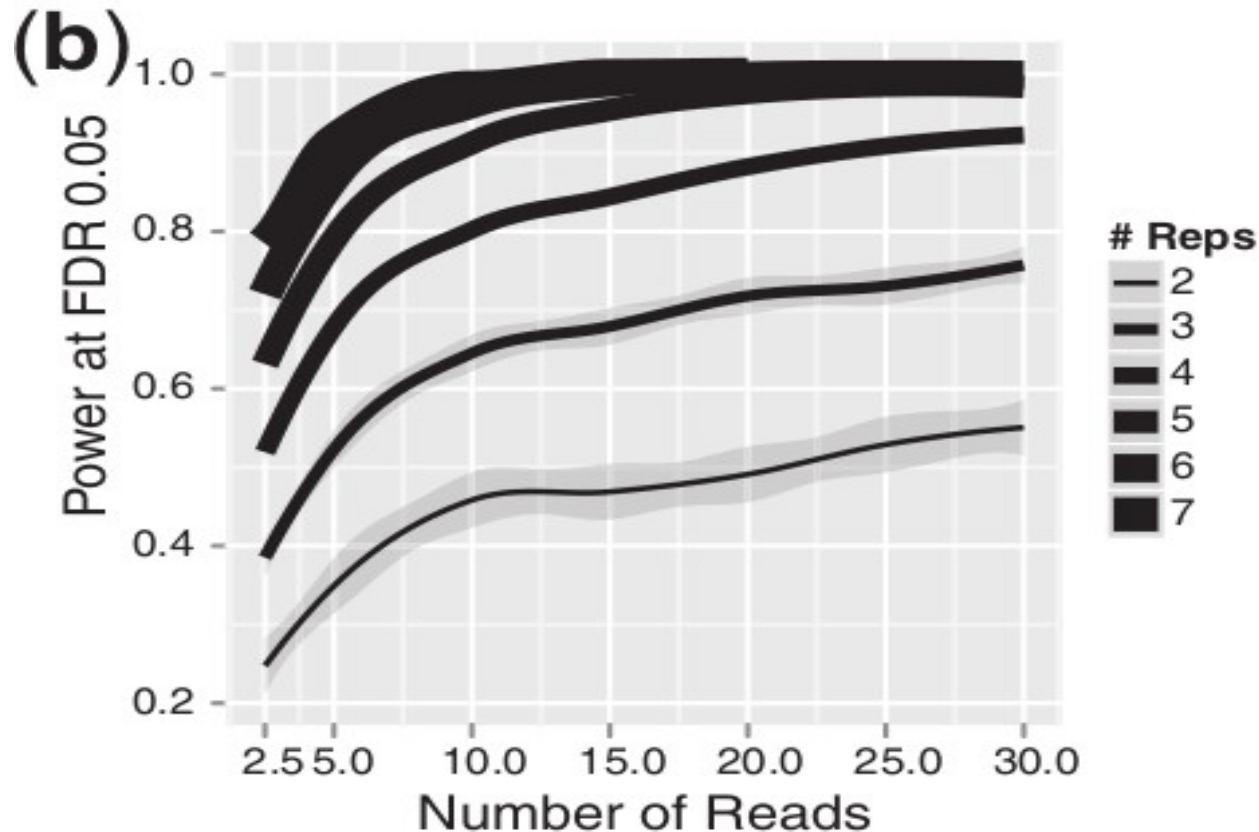
Samples per condition

Estimating variance of a normally distributed variable



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

More reads or more replicates ?



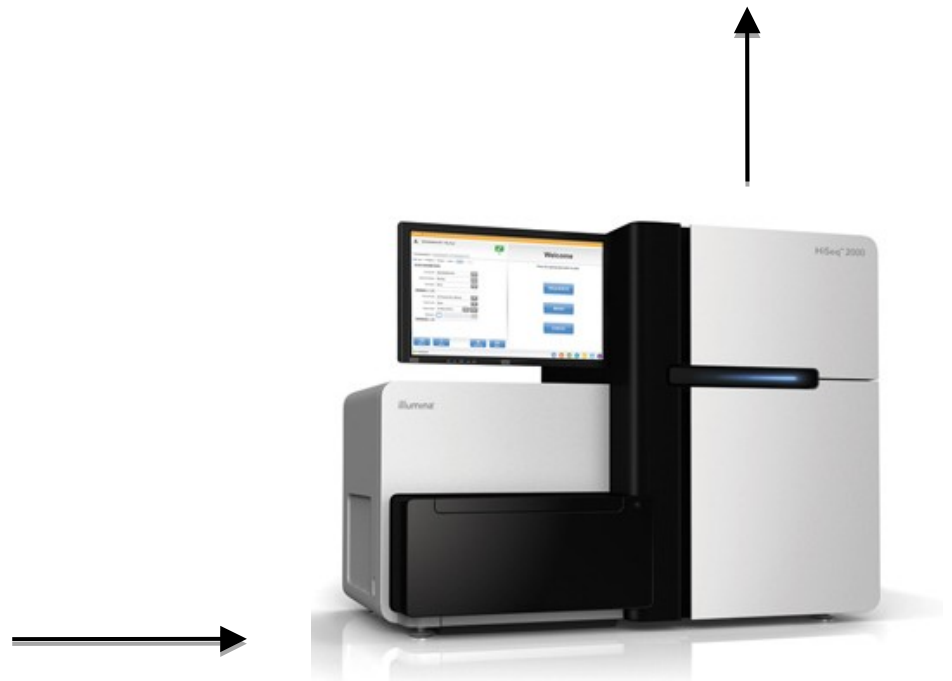
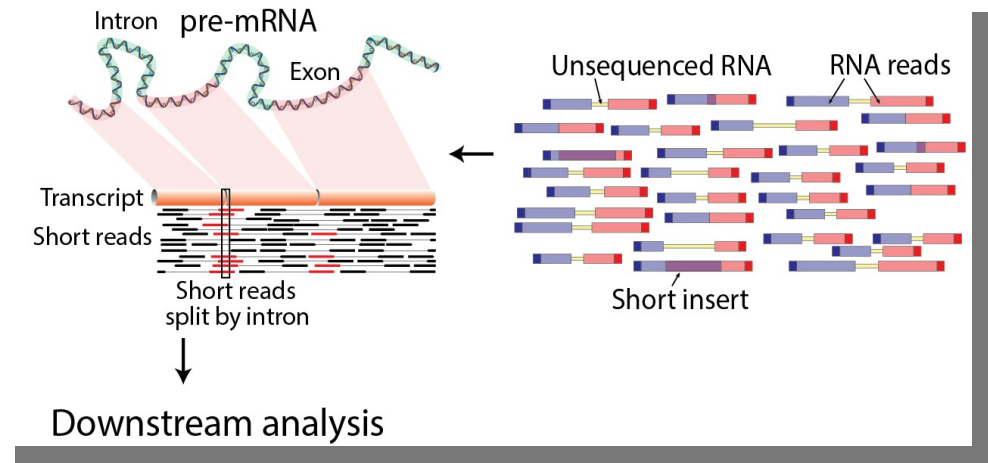
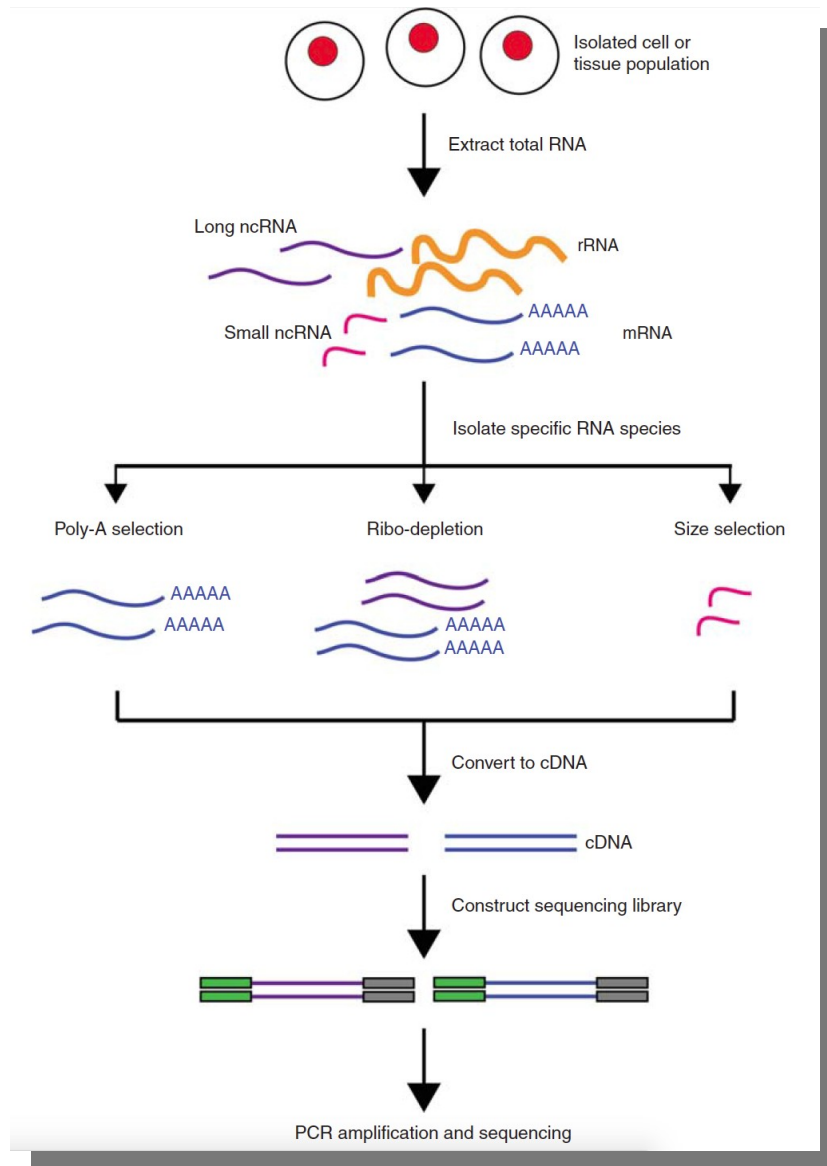
Generally more replicates should be preferred if they can be obtained.

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

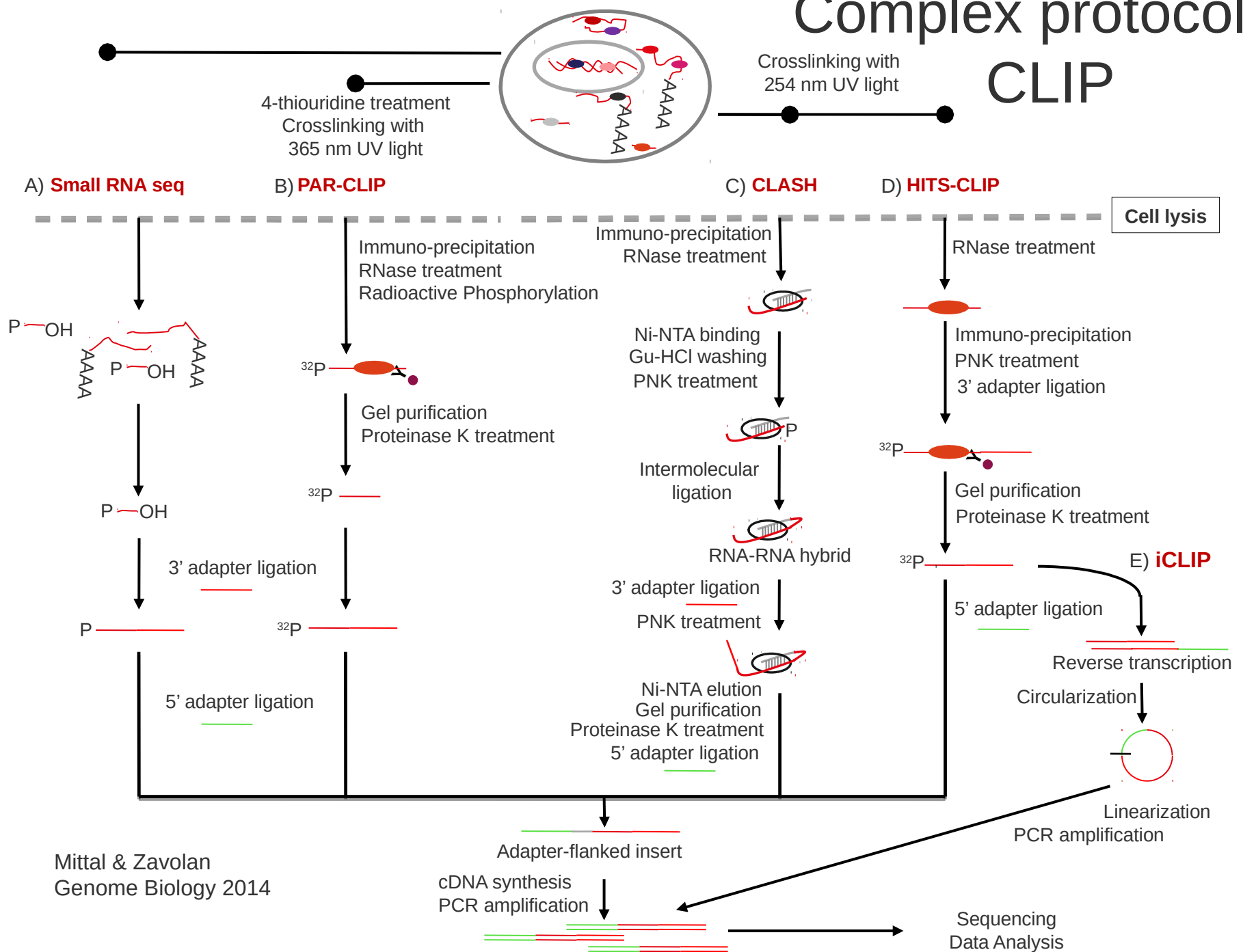
Experimental design resources

- Local sequencing facility and/or bioinformatics unit
- https://genome.ucsc.edu/ENCODE/protocols/dataStandards/ENCODE_RNAseq_Standards_V1.0.pdf
- <http://seqanswers.com/>
- <https://www.biostars.org>

Basic RNA-Seq Protocol Overview

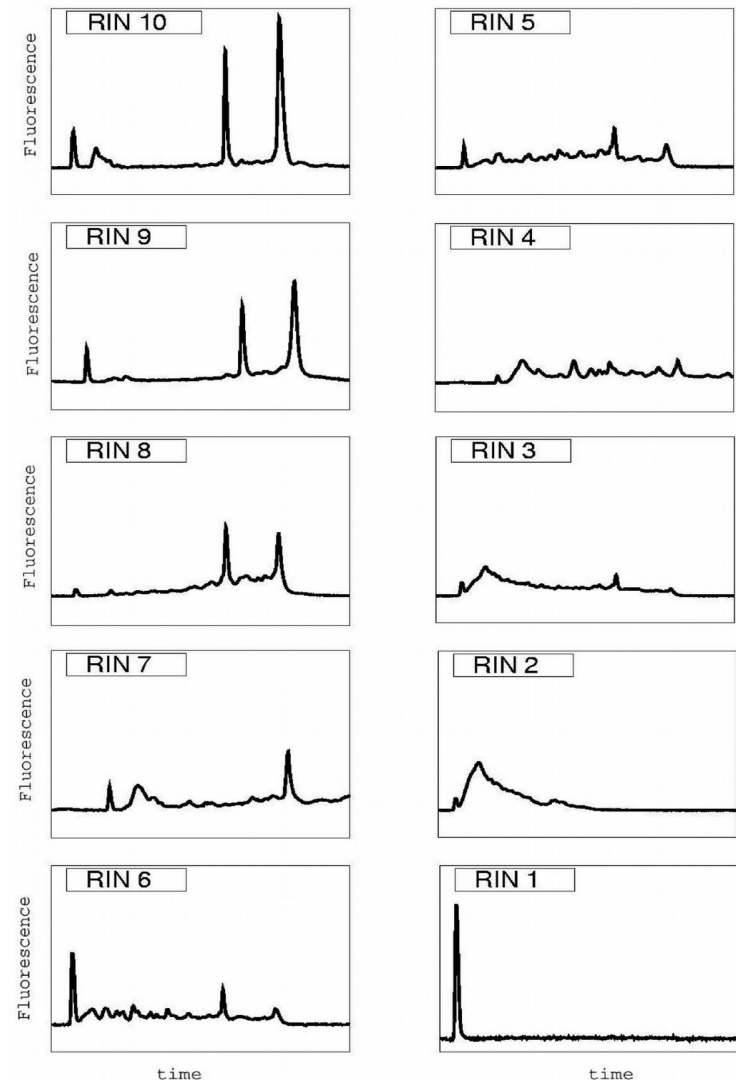


Complex protocols: CLIP



RNA Sample Preparation – RNA isolation

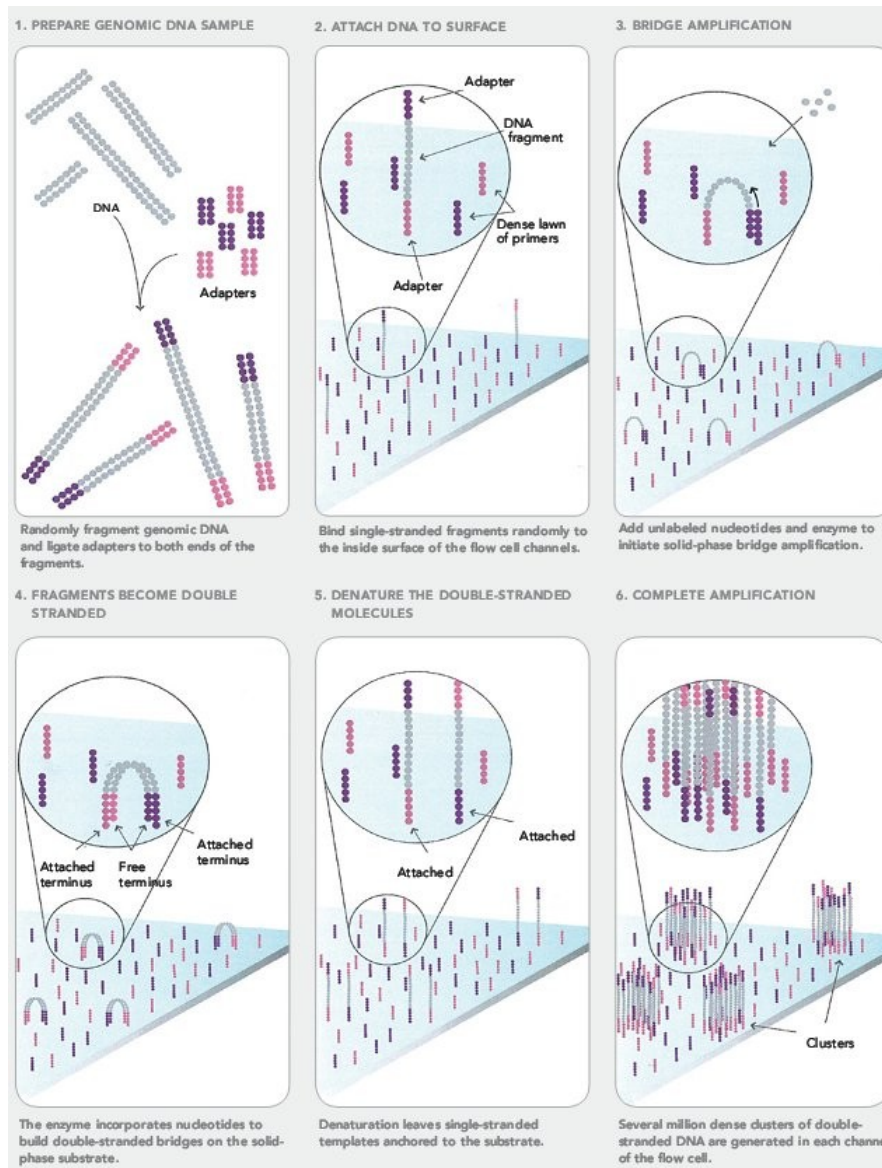
- Critically important, a good analysis cannot make up for poor data!
- RNA integrity number (RIN) estimates integrity based on the Agilent profile



RNA-Seq Preparation - Purification

- **Consider which RNA species you are trying to purify**
- **Consider the tissue source and number of cells**
- **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
 - Allows identification of lncRNAs without polyA tails

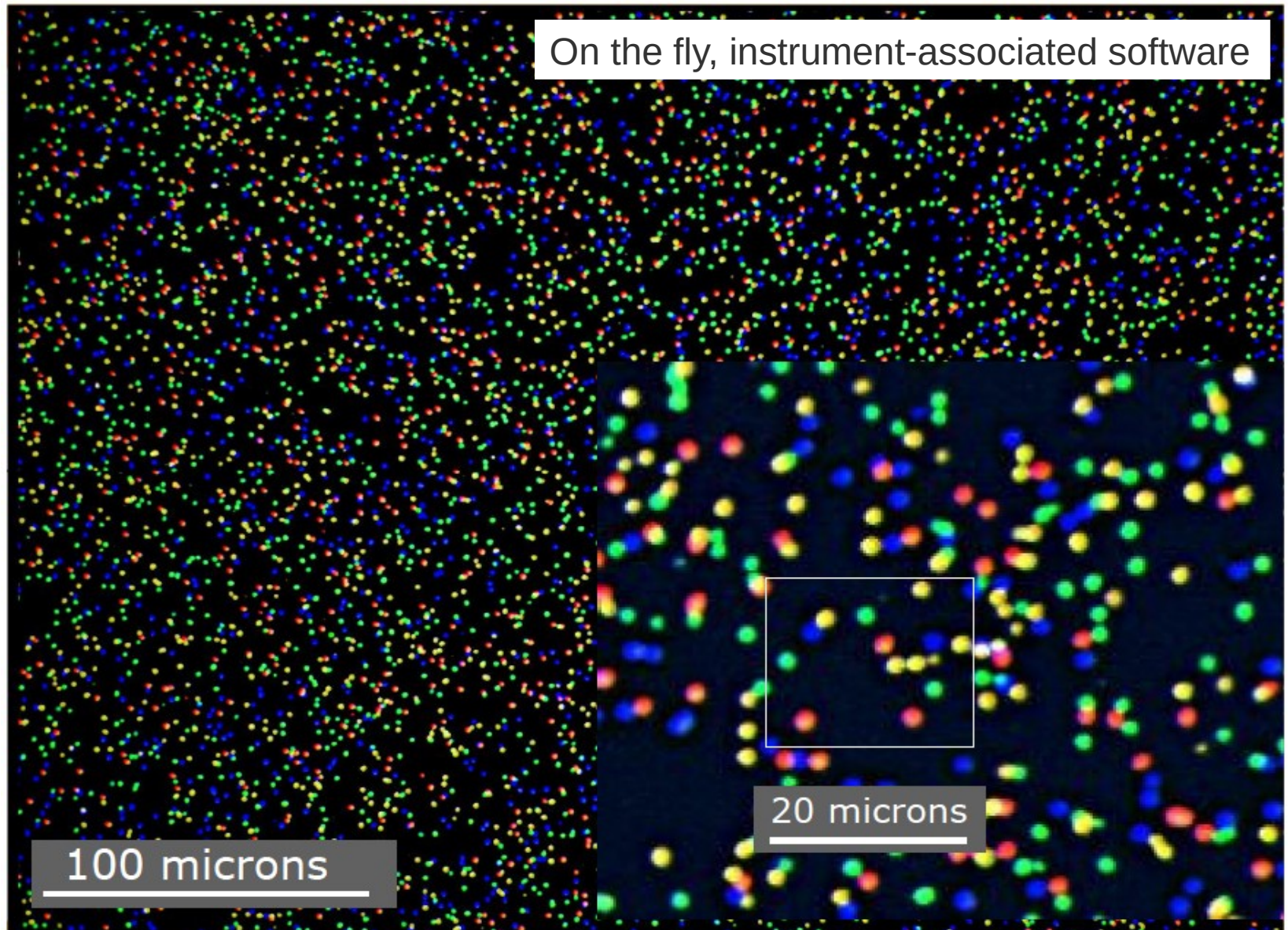
Illumina sequencing – “cluster formation”



Sequencing By Synthesis (SBS)



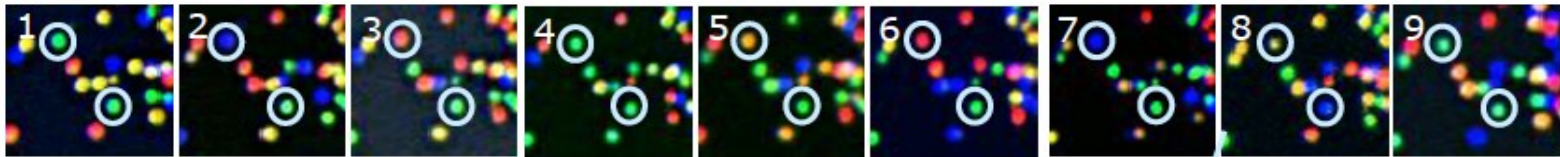
Illumina sequencing – image analysis



Illumina sequencing – from image to sequence

Base Calling From Raw Data

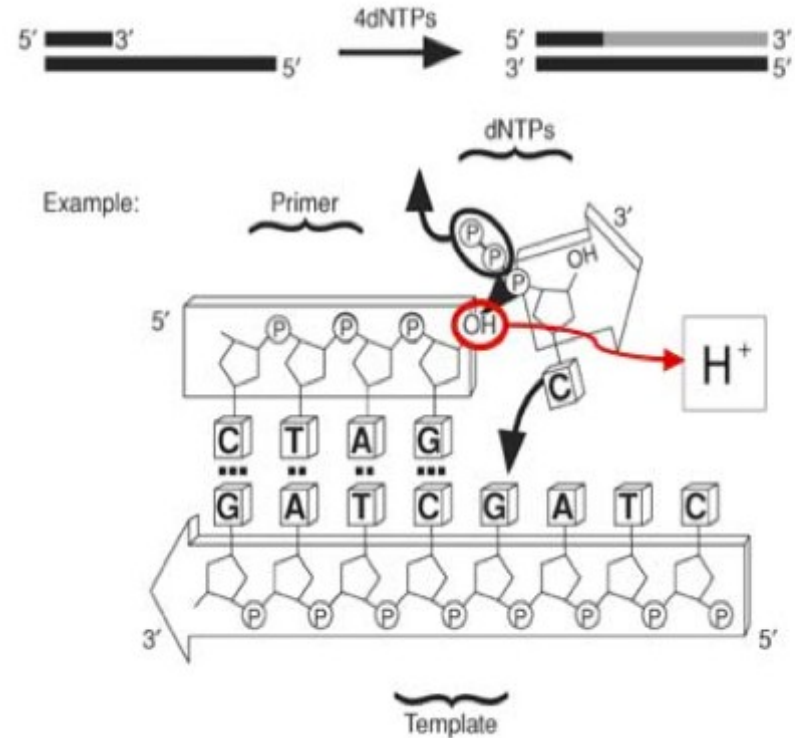
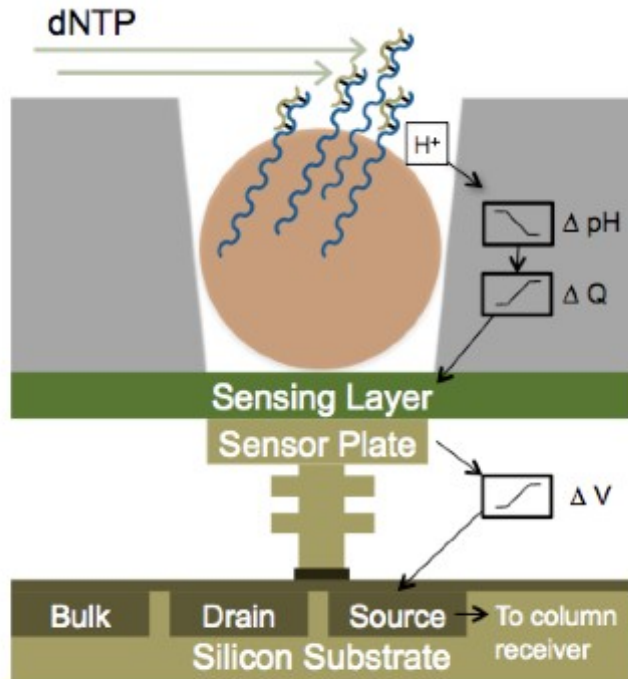
T G C T A C G A T ...



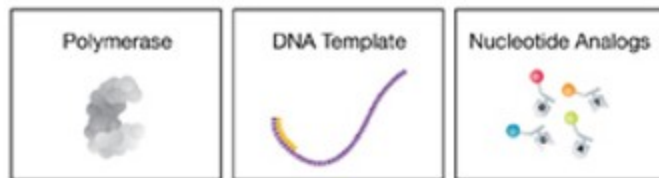
T T T T T T T G T ...

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

Ion torrent – reading pH changes



Pacific Biosciences – single molecule real time sequencing



Rhoads & Au.

Genomics Proteomics Bioinformatics 2015

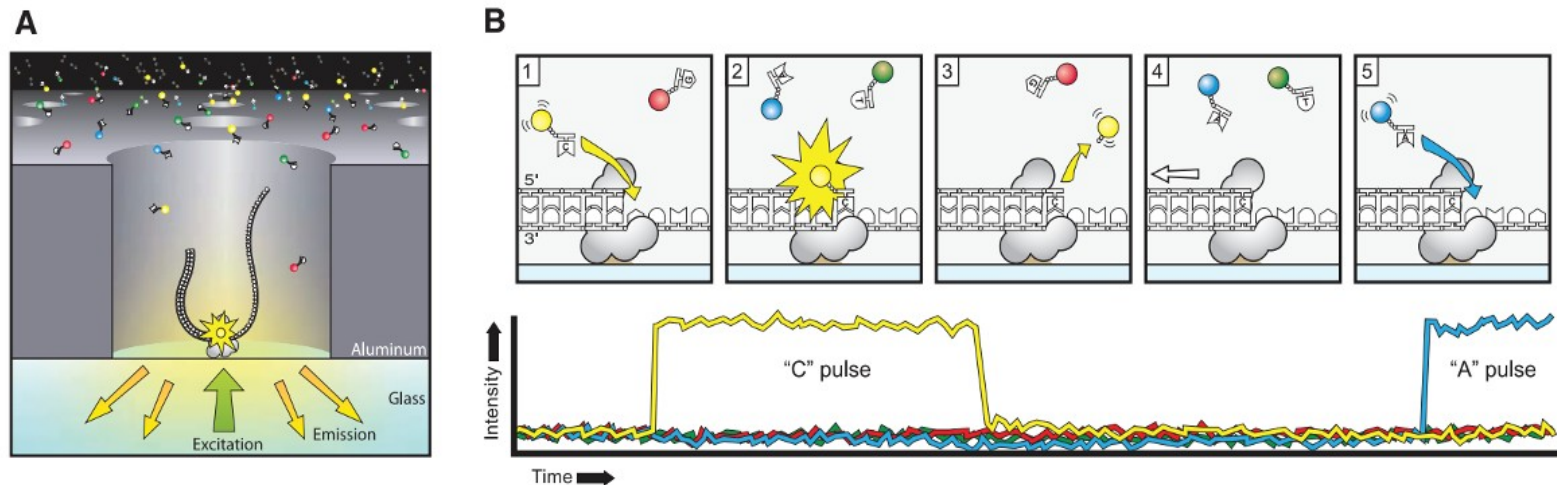
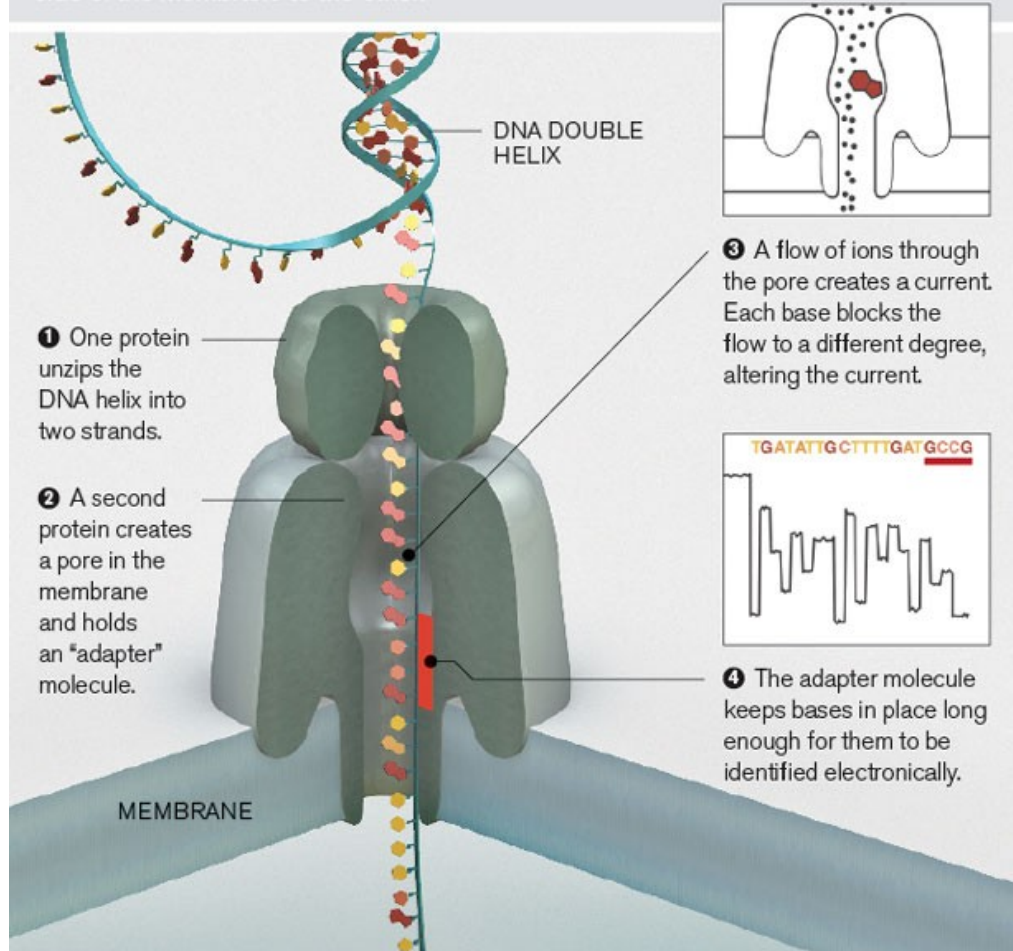


Fig. 1. Principle of single-molecule, real-time DNA sequencing. **(A)** Experimental geometry. A single molecule of DNA template-bound $\Phi 29$ DNA polymerase is immobilized at the bottom of a ZMW, which is illuminated from below by laser light. The ZMW nanostructure provides excitation confinement 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.

Oxford Nanopore – direct DNA/RNA sequencing

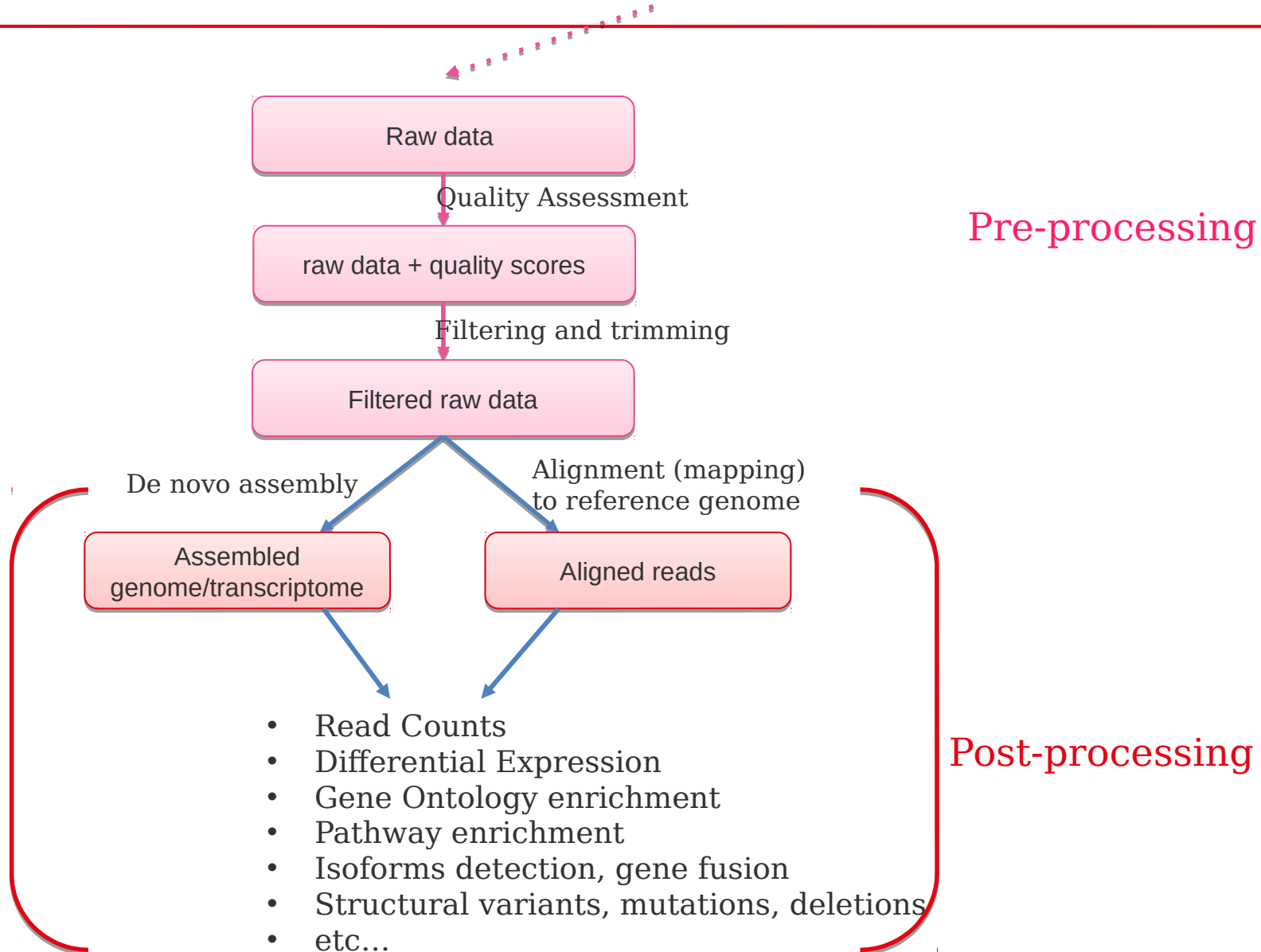
DNA can be sequenced by threading it through a microscopic pore in a membrane. Bases are identified by the way they affect ions flowing through the pore from one side of the membrane to the other.

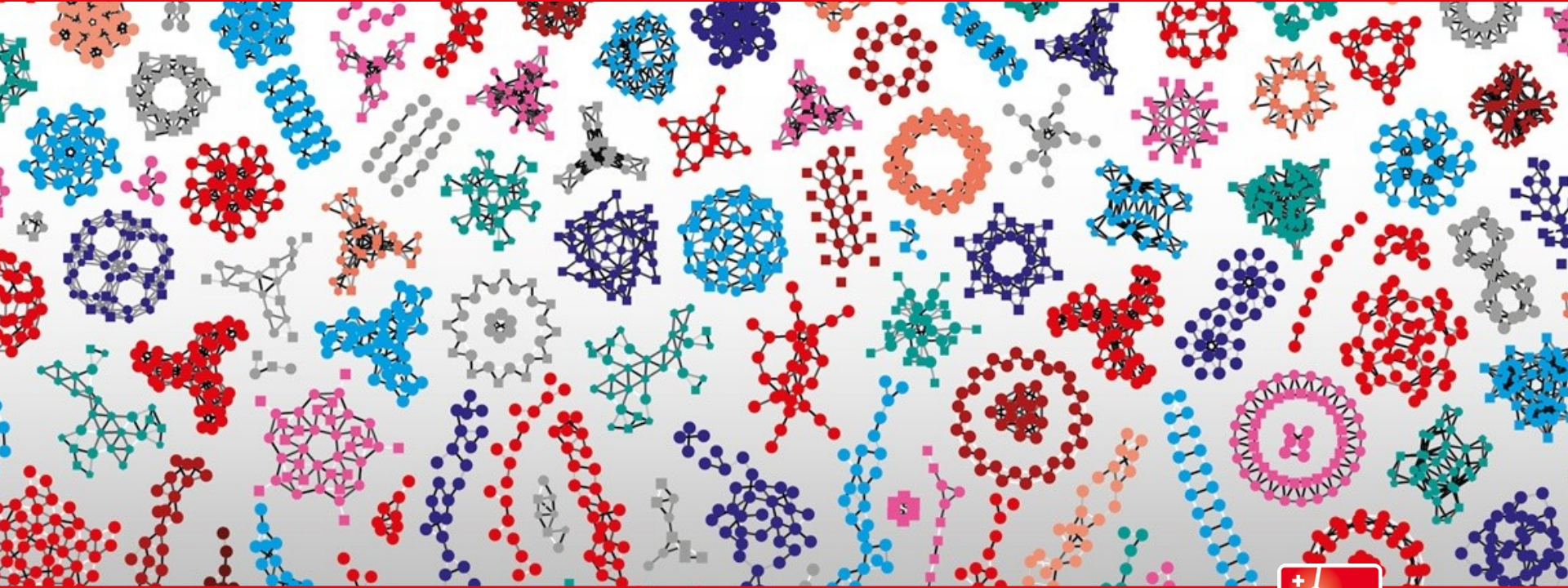


RNA-seq data analysis



Sequencing facility





Swiss Institute of
Bioinformatics

Contributors:

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



www.sib.swiss