LongTREC

Funded by
the European Union

Bioinformatics Summer School

Long-reads Transcriptomics

Carolina Monzó

I2SysBio, Valencia, Spain

LongTREC - The Long-reads TRanscriptome European Consortium
Marie Skłodowska-Curie grant agreement No 101072892

Course Contents

1. Introduction to Long-read Technologies
2. Experimental design
3. Use cases and wrap up
4

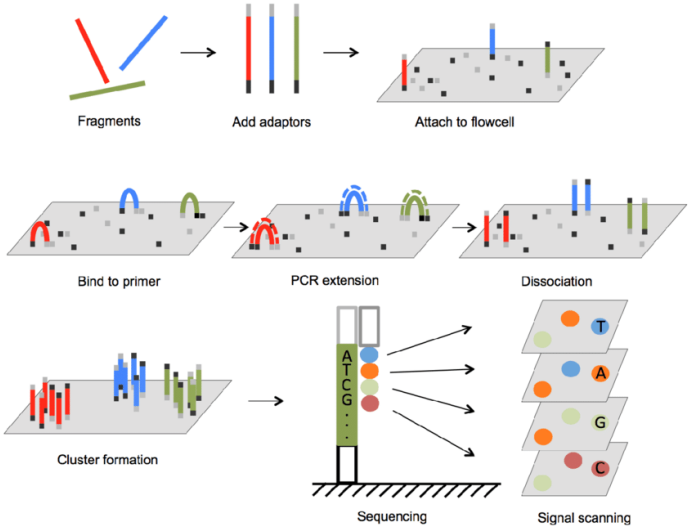
Section 1

Introduction to Long-read Technologies

Understanding the fundamentals of third-generation sequencing

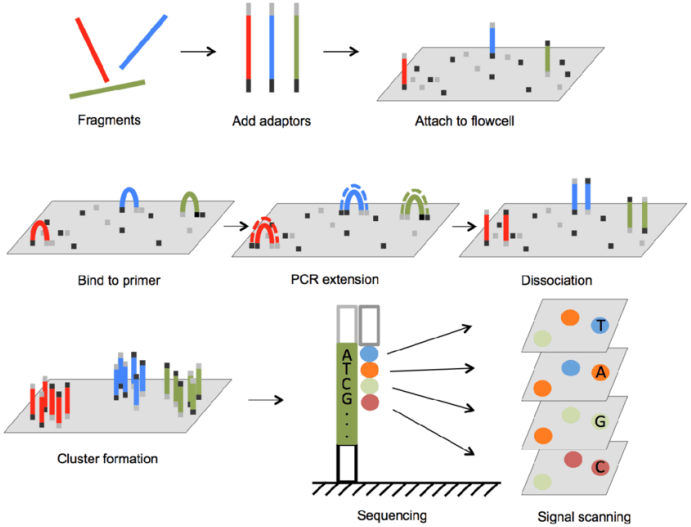
Classical sequencing, 2nd generation

illumina®



Classical sequencing, 2nd generation

illumina®



Advantages:

- Higher accuracy
- Higher throughput
- Cost-effective

Disadvantages:

- Repetitive sequences
- Structural variants
- Isoform resolution – Alternative splicing

Solving the issues: long-read sequencing

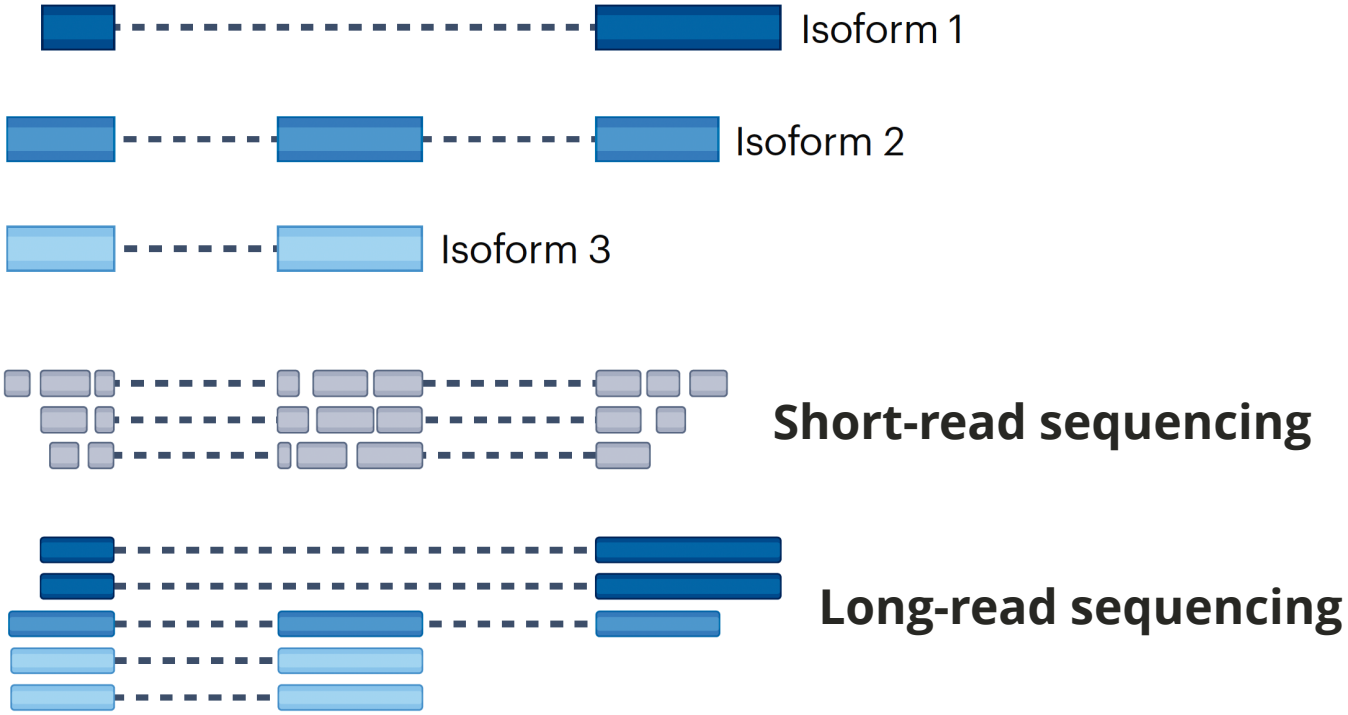


Figure source: Monzó, Liu, Conesa. 2025. 10.1038/s41576-025-00828-z

PacBio

Oxford NANOPORE Technologies

Advantages:

- Single-molecule
- Isoform resolution – Alternative splicing
- Assembly simplification

Disadvantages:

- Lower accuracy
- Lower throughput
- Higher costs
- Bioinformatic challenges

Introduction to Long-read Sequencing, 3rd generation

PacBio

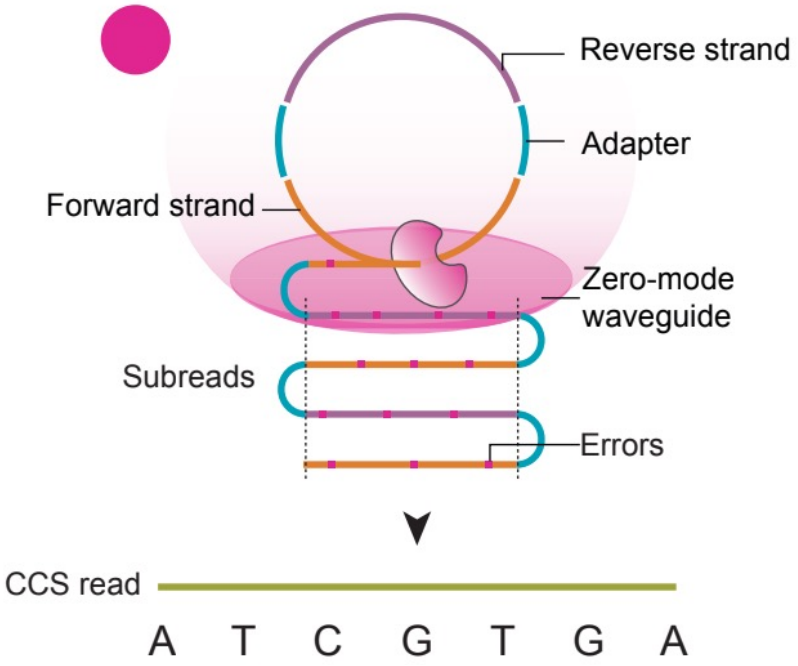
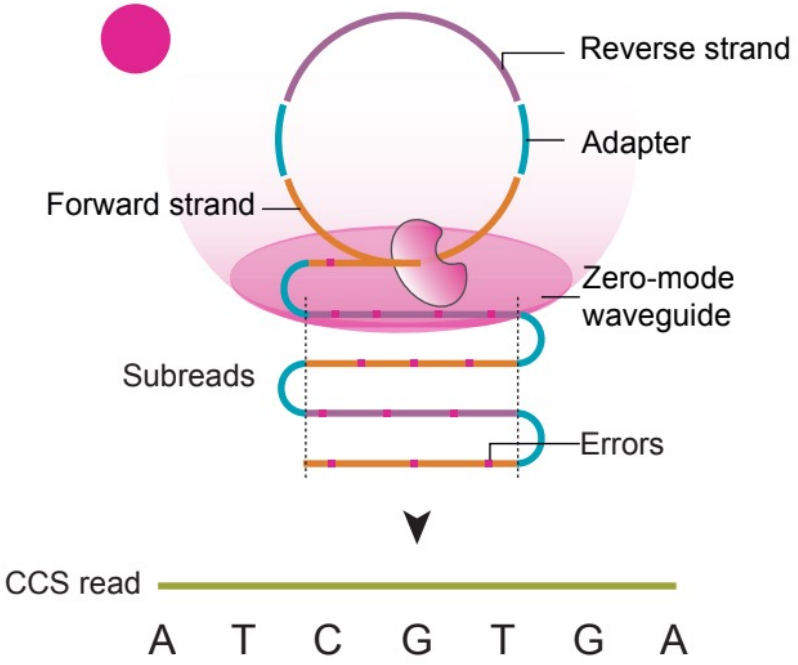


Figure source: Monzó, Liu, Conesa. 2025. 10.1038/s41576-025-00828-z

Introduction to Long-read Sequencing, 3rd generation

PacBio



Oxford NANOPORE Technologies

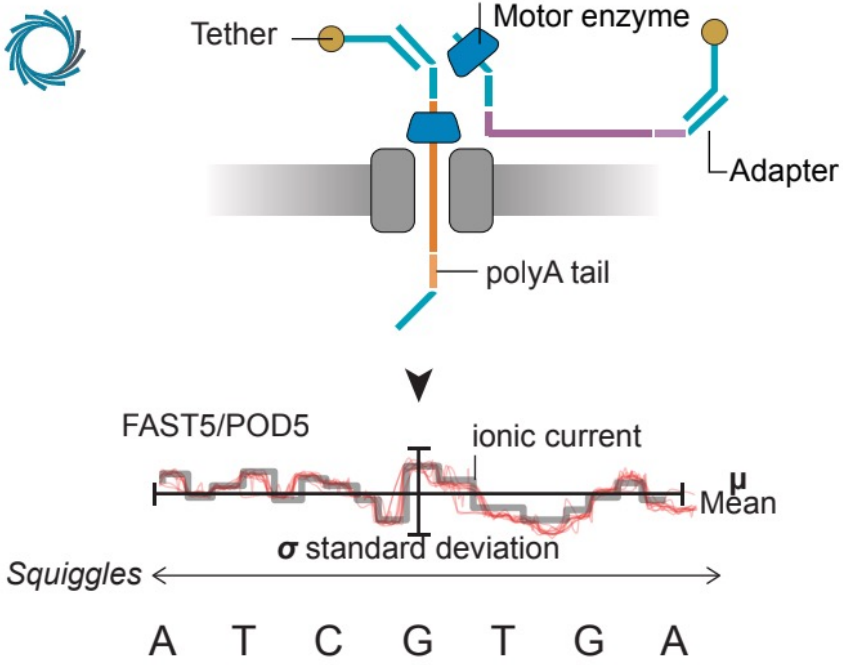


Figure source: Monzó, Liu, Conesa. 2025. 10.1038/s41576-025-00828-z

The Long-read transcriptomics sequencing roadmap

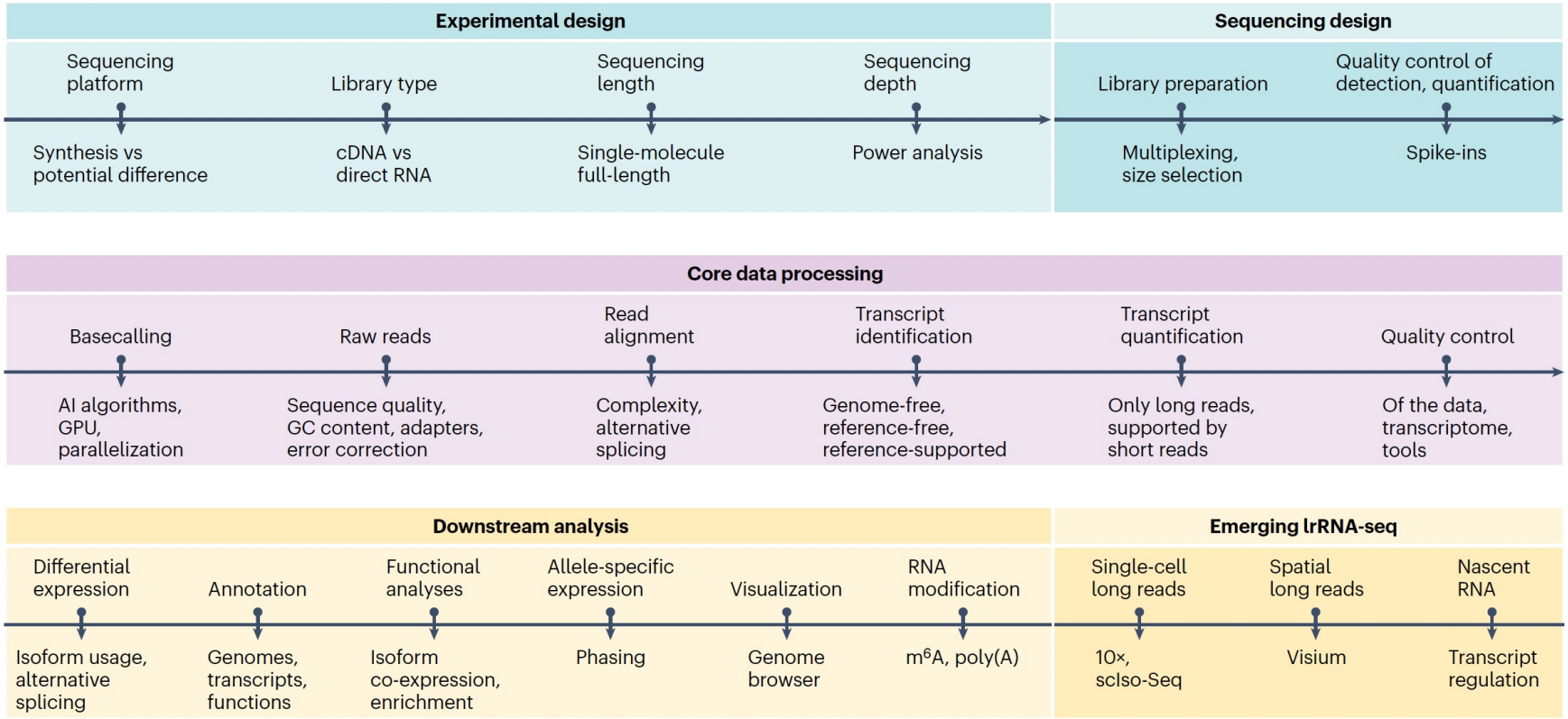


Figure source: Monzó, Liu, Conesa. 2025. 10.1038/s41576-025-00828-z

The Long-read transcriptomics sequencing roadmap

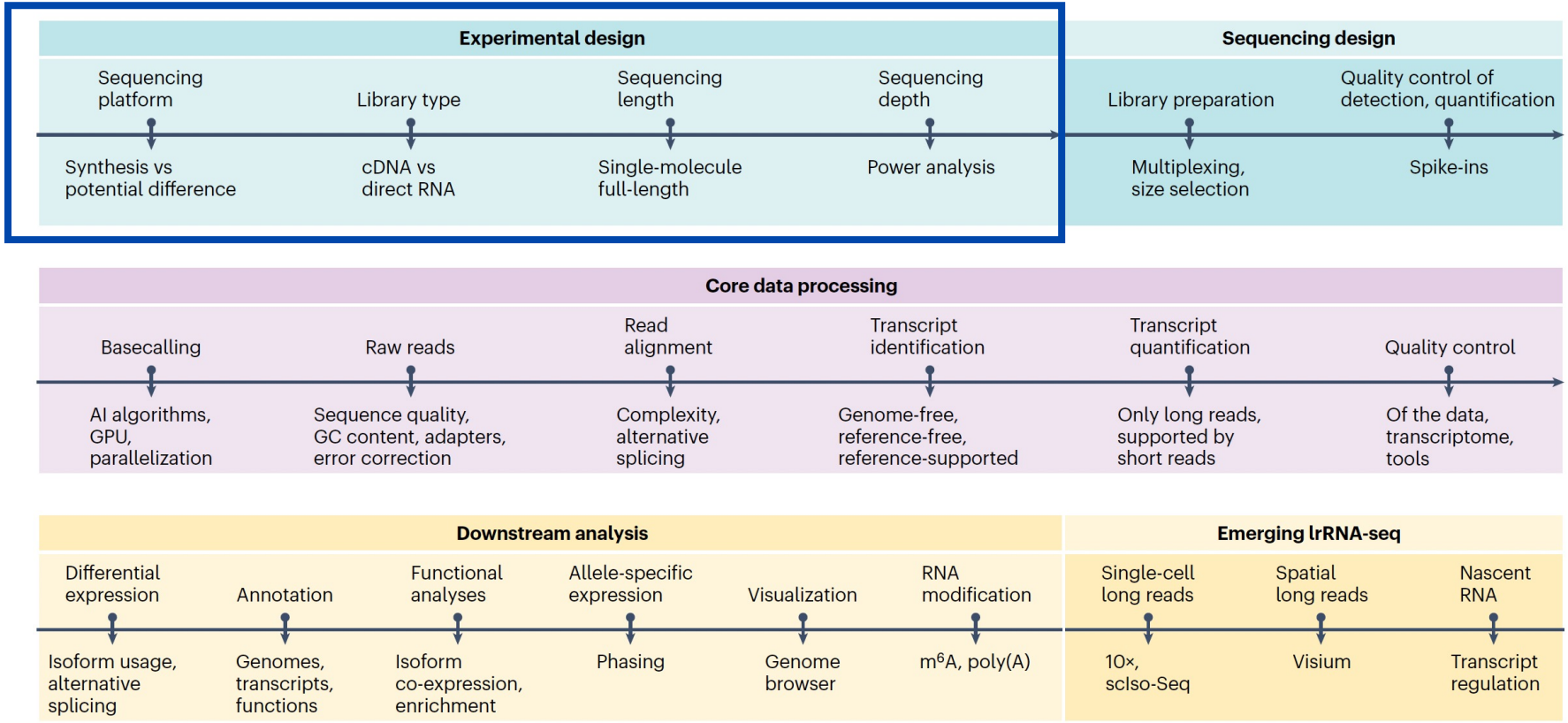


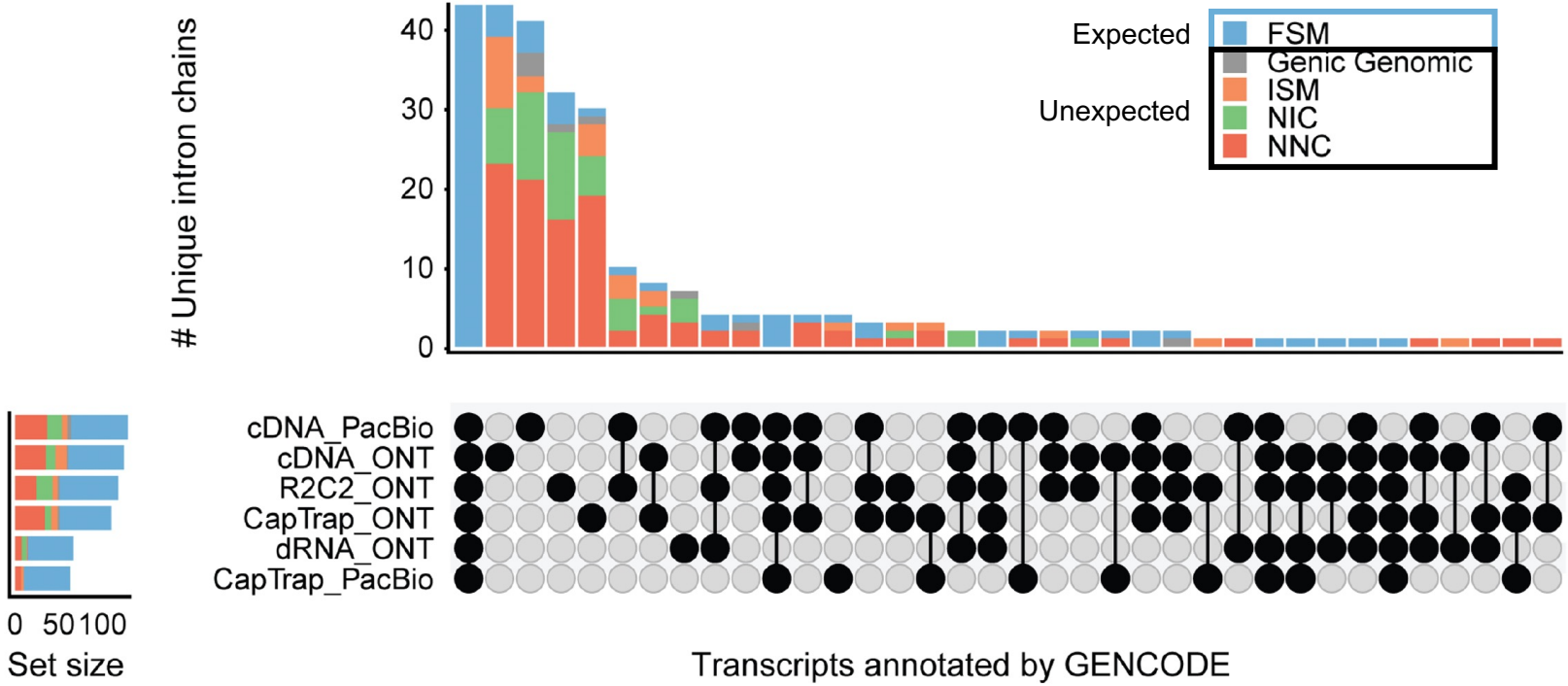
Figure source: Monzó, Liu, Conesa. 2025. 10.1038/s41576-025-00828-z

Section 2

Experimental design

Understanding the fundamentals of third-generation sequencing

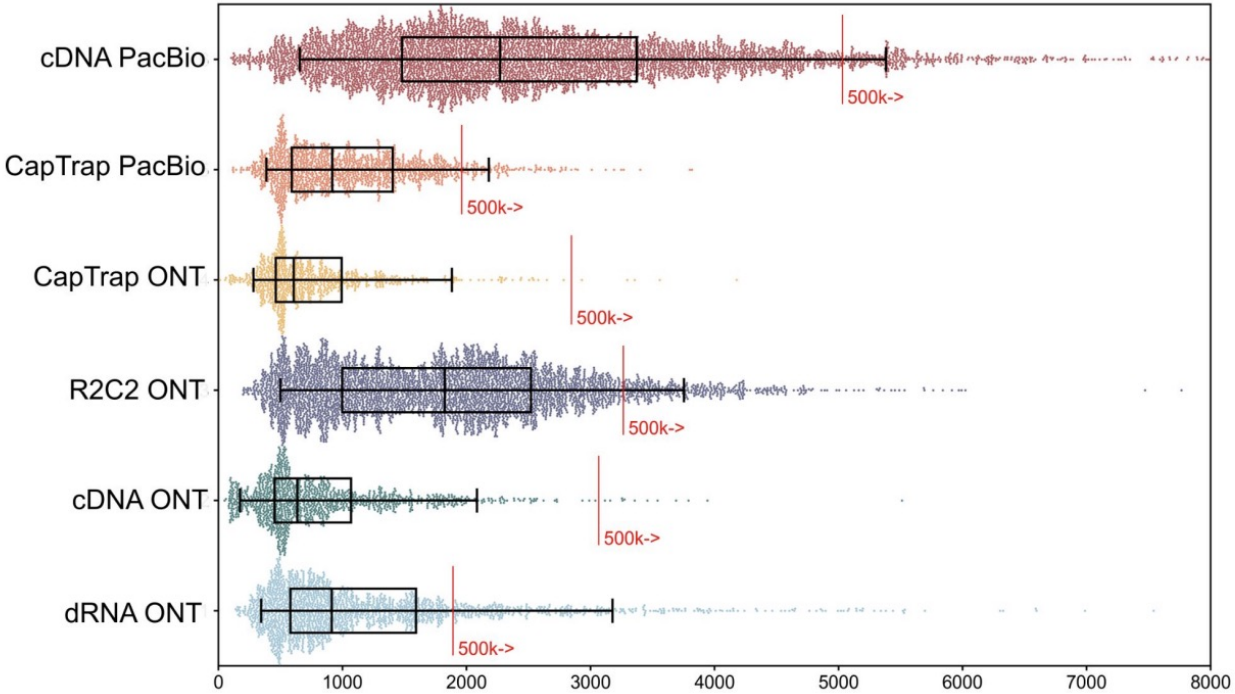
Impact of sequencing platform and library preparation choice



WARNING!! Combination of sequencing platform and library preparation find **different transcripts**

Figure source: Monzó, Frankish, Conesa. 2025. 10.1101/gr.279865.124

Impact of sequencing platform and library preparation choice



WARNING!! Combination of sequencing platform and library preparation find **different transcripts**

Figure source: Pardo-Palacios. 2024. 10.1038/s41592-024-02298-3

What is your biological question?

RNA quality

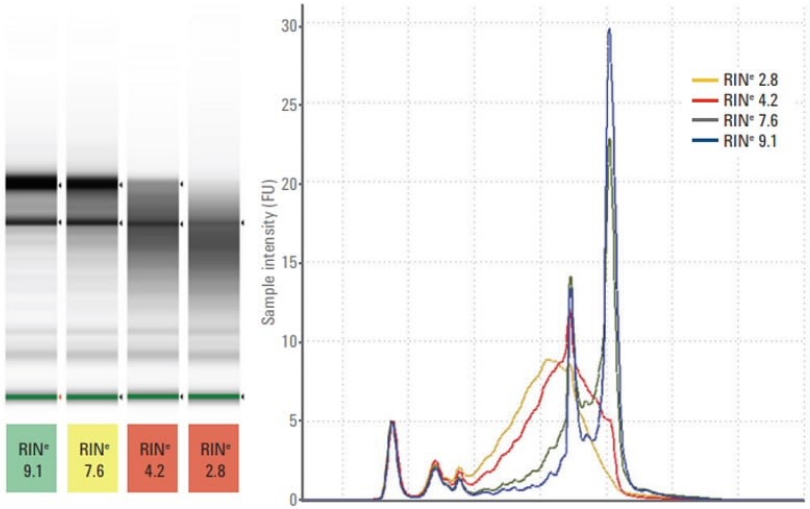
Degraded RNA:

- Incomplete transcripts (difficulty to differentiate TSS and TTS)
- Biased coverage, tendency to over-represent 3' ends
- Compromised measurements of polyA tail lengths

RNA Integrity Number: ratio of intact ribosomal RNA peaks to degraded RNA fragments

- Get RIN as high as possible
- RIN > 7 is the minimum, but **RIN > 8 recommended** by ONT and PacBio
- Validate TTS and TSS with orthogonal data (CAGE-Seq or Quant-Seq)
- Incorporate RIN as a covariate in the Differential Expression analysis

Checking the RNA quality is not only important to get good quality full-length transcripts, but also to see the distribution of transcripts that we have in the transcriptome of our sample



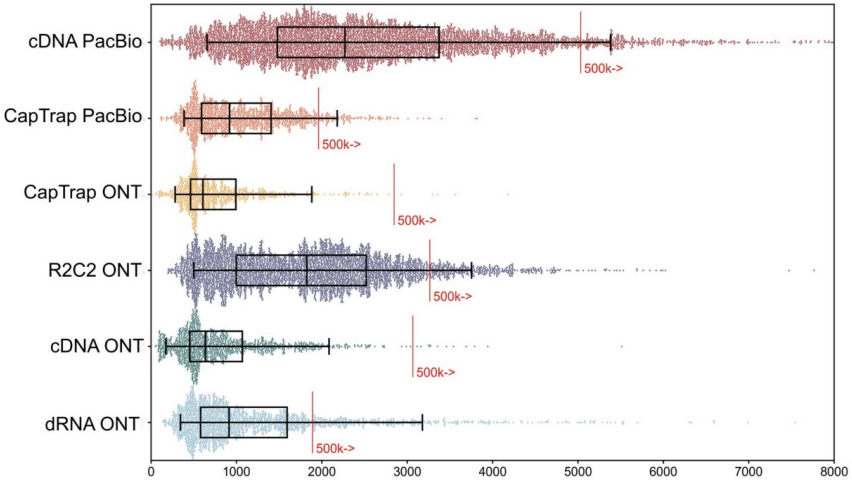
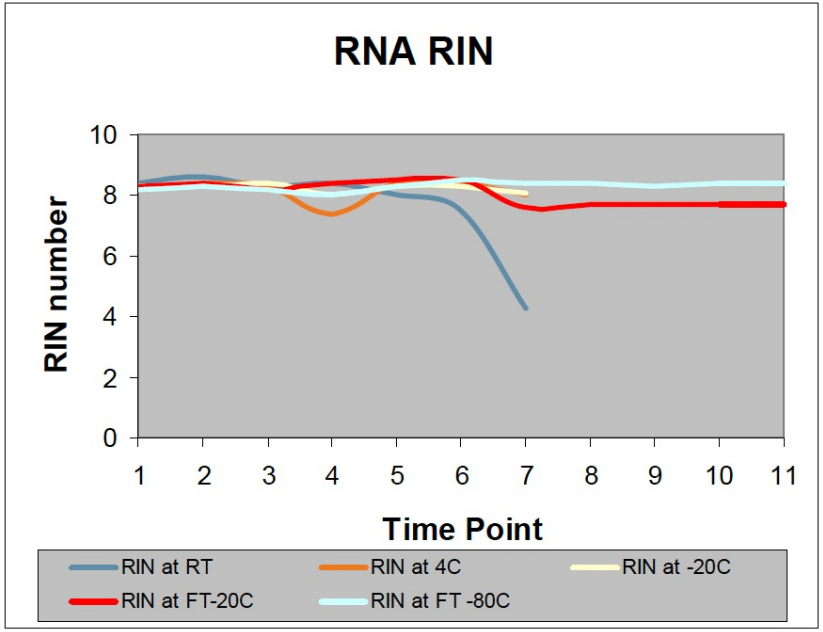


Figure source: Pardo-Palacios. 2024. 10.1038/s41592-024-02298-3

RNA quality



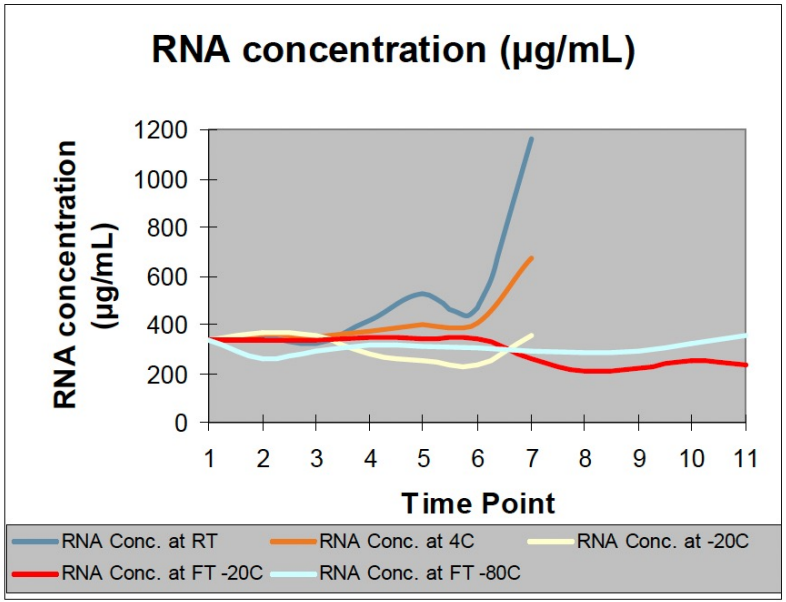
Tissue extracted RNA integrity is not affected by the storage conditions in 14 days, and not affected by freeze-thaws up to 10 cycles.

Extract the RNA from your samples quickly!

Don’t store the whole tissue, store the RNA

Figure source: Kim, Wu, Huang et al. 2011

RNA concentration



The RNA concentration increased over time at both room temperature and 4°C, showing that evaporation occurred at room temperature from day 3 and at 4°C from day 7.

Figure source: Kim, Wu, Huang et al. 2011

Choosing a sequencing platform

Similarities between PacBio and ONT:

- Throughput (high number of reads per run)
- Capacity to sequence full-length transcripts

Technology	Platform	Median read length (kb)	Median throughput per run (10^6 transcript reads)
PacBio	cDNA+IsoSeq+Sequel II	~2.1	~2.6
	cDNA+Kinnex+Sequel II	~1.7	~40
	cDNA+Kinnex+Revio	~1.7	~100
ONT	cDNA+MinION (R10.4)	~0.939	~20
	cDNA+PromethION (R9.4)	~1	~130
	dRNA+MinION (R9.4)	~0.8	~1.1
	dRNA+PromethION (R9.4)	~0.6	~20

Figure source: Monzó, Liu, Conesa. 2025. 10.1038/s41576-025-00828-z

PacBio

- **Higher read accuracy**
- Commercial protocols have broad read length distribution

- Can't sequence dRNA or direct modifications

Oxford NANOPORE Technologies

- Can sequence cDNA or dRNA
- Can simultaneously sequence dRNA and epitranscriptomic modifications

- Specific protocols can sequence ultra-long reads. **But standard protocols are biased towards short reads**
- Lower read accuracy
- Difficulty differentiating modifications

Sequencing depth

Number of reads per sample:

- Higher depth discovers more novel transcripts
- More complex organisms require more sequencing depth

$$\text{Depth}_{\text{sample}} = \frac{\Sigma \text{Depth}_{\text{isoform}}}{N_{\text{expressed isoforms in sample}}}$$

Recommendations (human):

- 10 million reads for 80% of known isoforms
- > 20 million reads for detection of rare isoforms

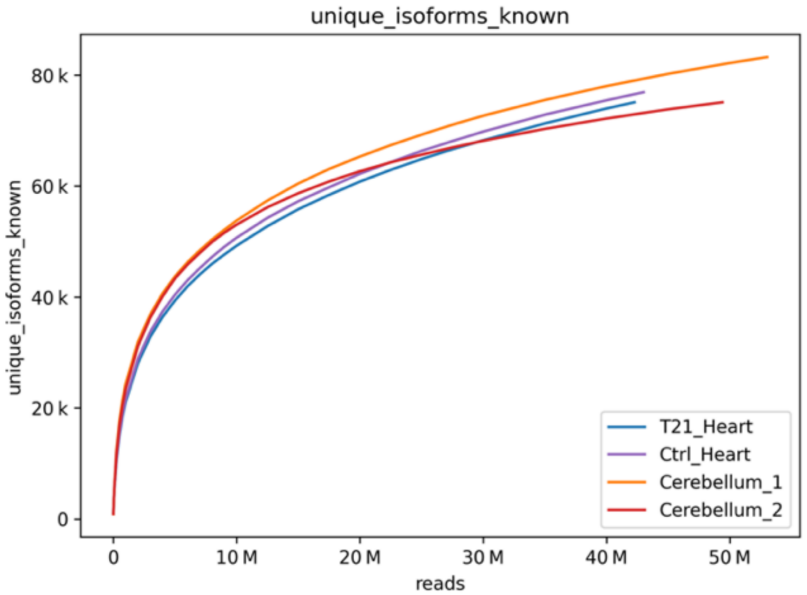
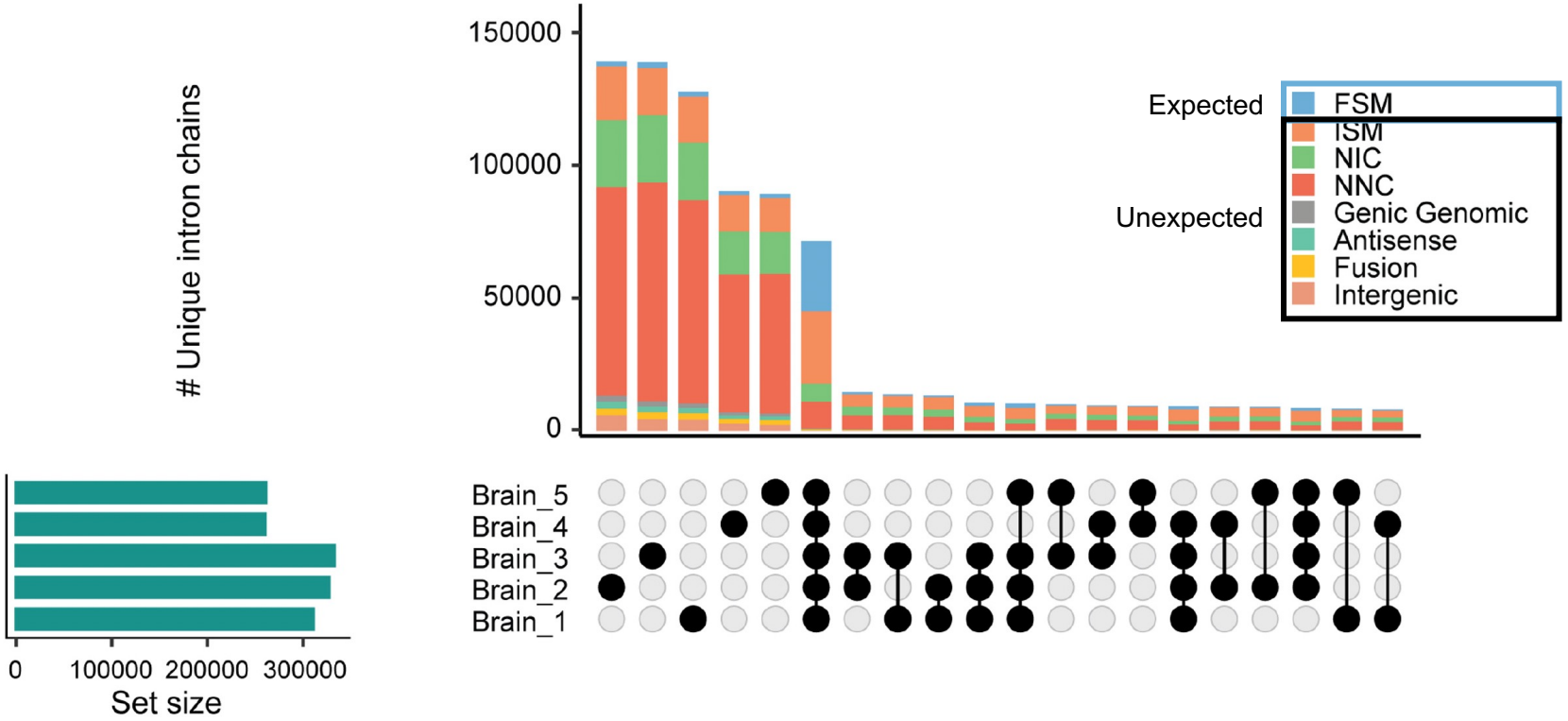


Figure source: https://www.pacb.com/wp-content/uploads/Application-note-Kinnex-full-length-RNA-kit-for-isoform-sequencing.pdf

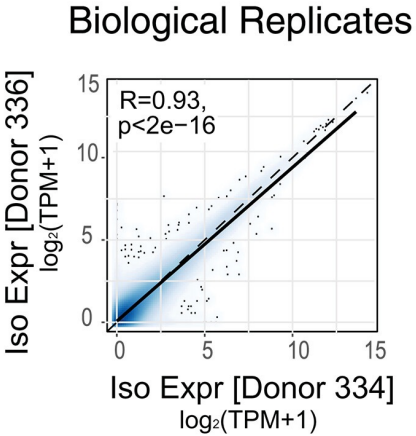
Impact of replication



Bimodal distribution, **most** transcripts are present in **all** samples **or** are **sample-specific**, few are found in two or more samples

Figure source: Monzó, Frankish, Conesa. 2025. 10.1101/gr.279865.124

Replication and multiplexing



Replication is necessary for statistical power and to reduce the inclusion of spurious transcripts in downstream analyses.

High expression correlation between replicates

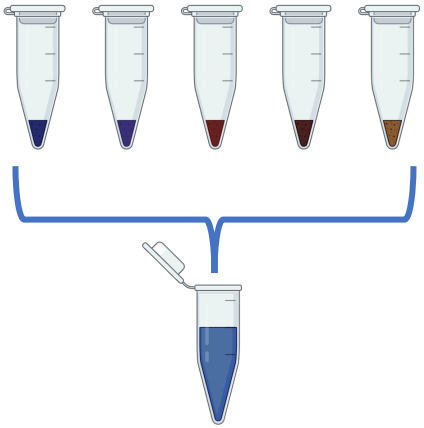


Figure source: Patowary, Zhang et al. 2024. 10.1126/science.adh7688

Replication and multiplexing

PacBio

bc = 4 different barcodes for library prep

cx = 12 different barcodes for cDNA synthesis

Possible to have 48 samples in one run

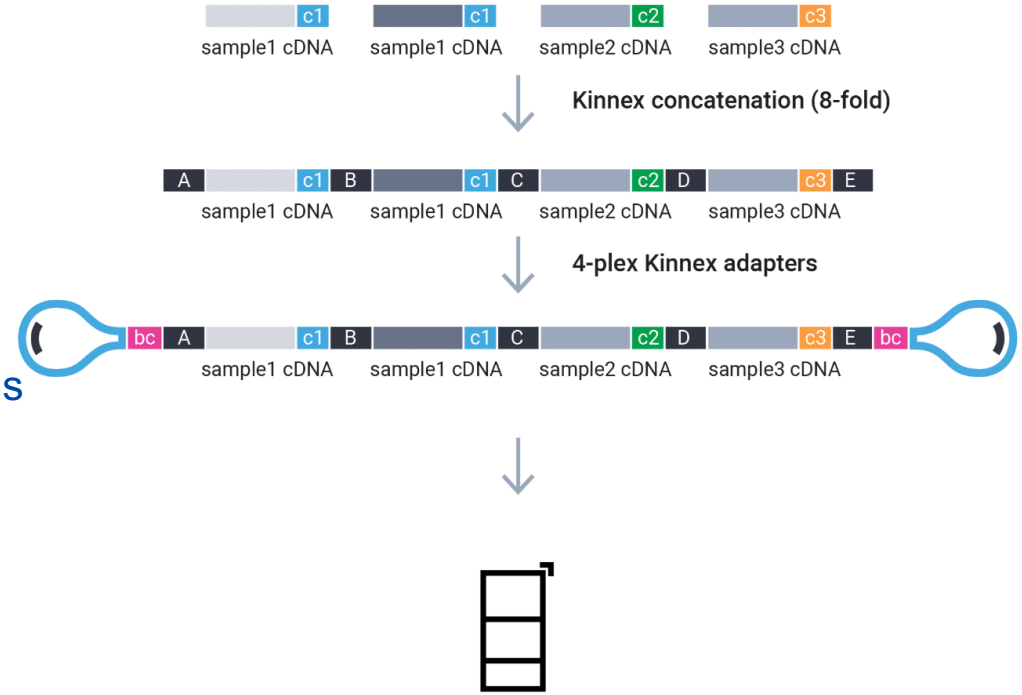
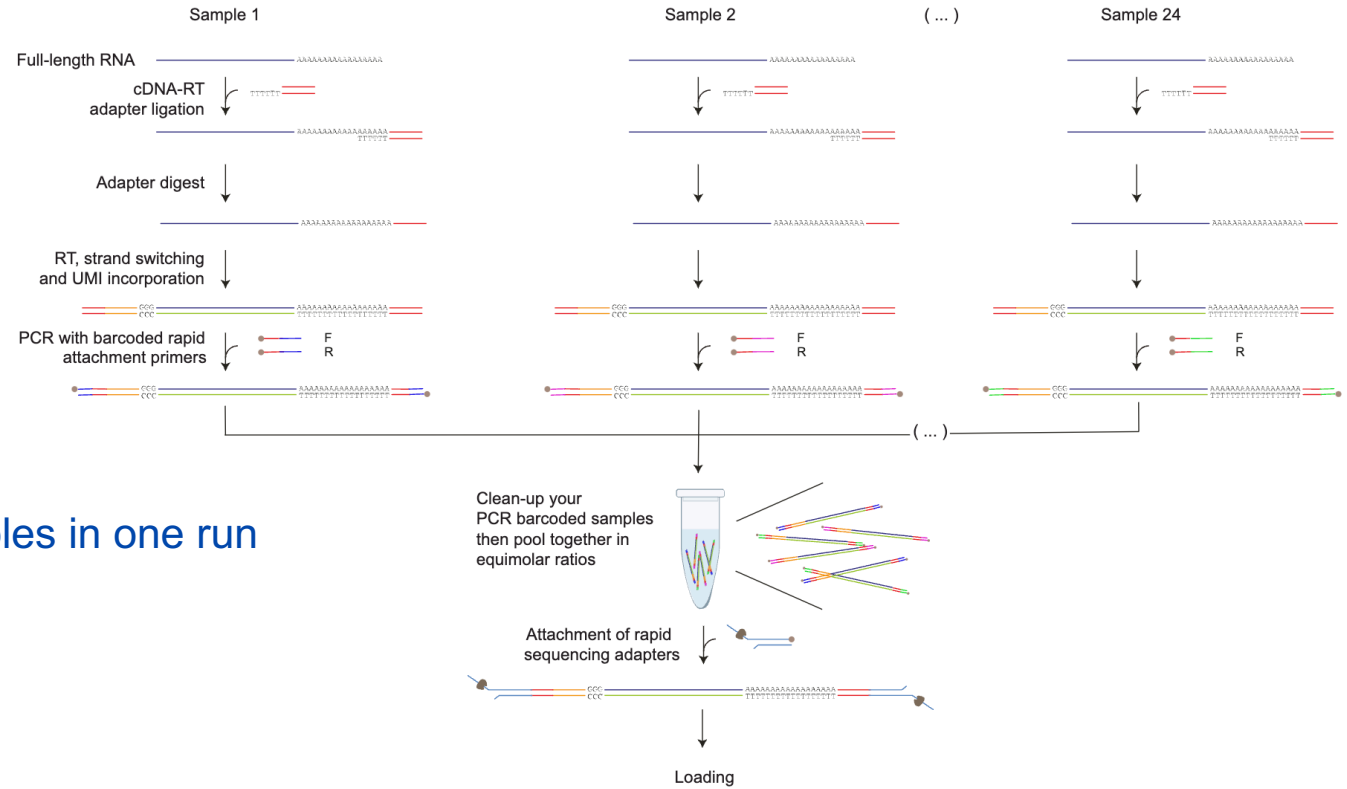


Figure source: https://www.pacb.com/multiplexing/

For cDNA!!

Possible to have 24 samples in one run



Synthetic RNA molecules that mimic: isoforms, abundance, and transcript length

SIRVs and sequins

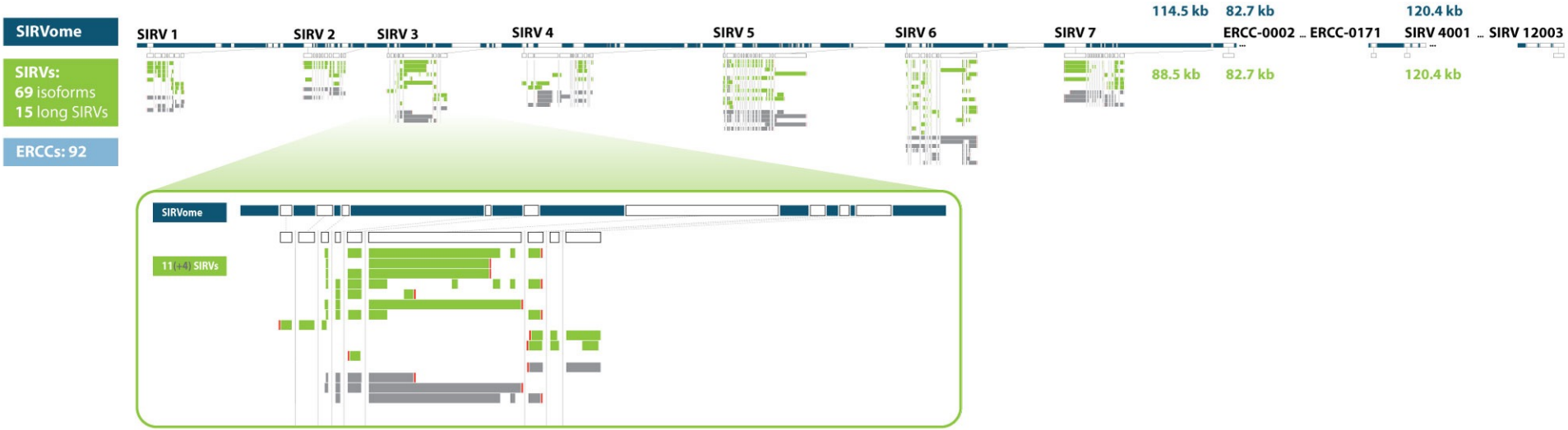
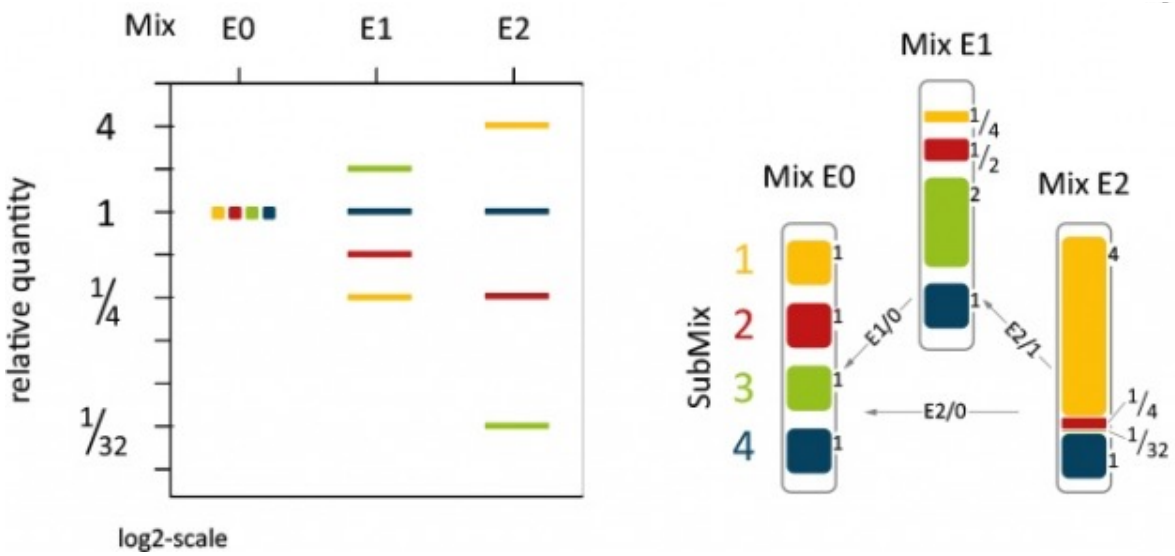


Figure source: https://www.lexogen.com/sirvs/design/

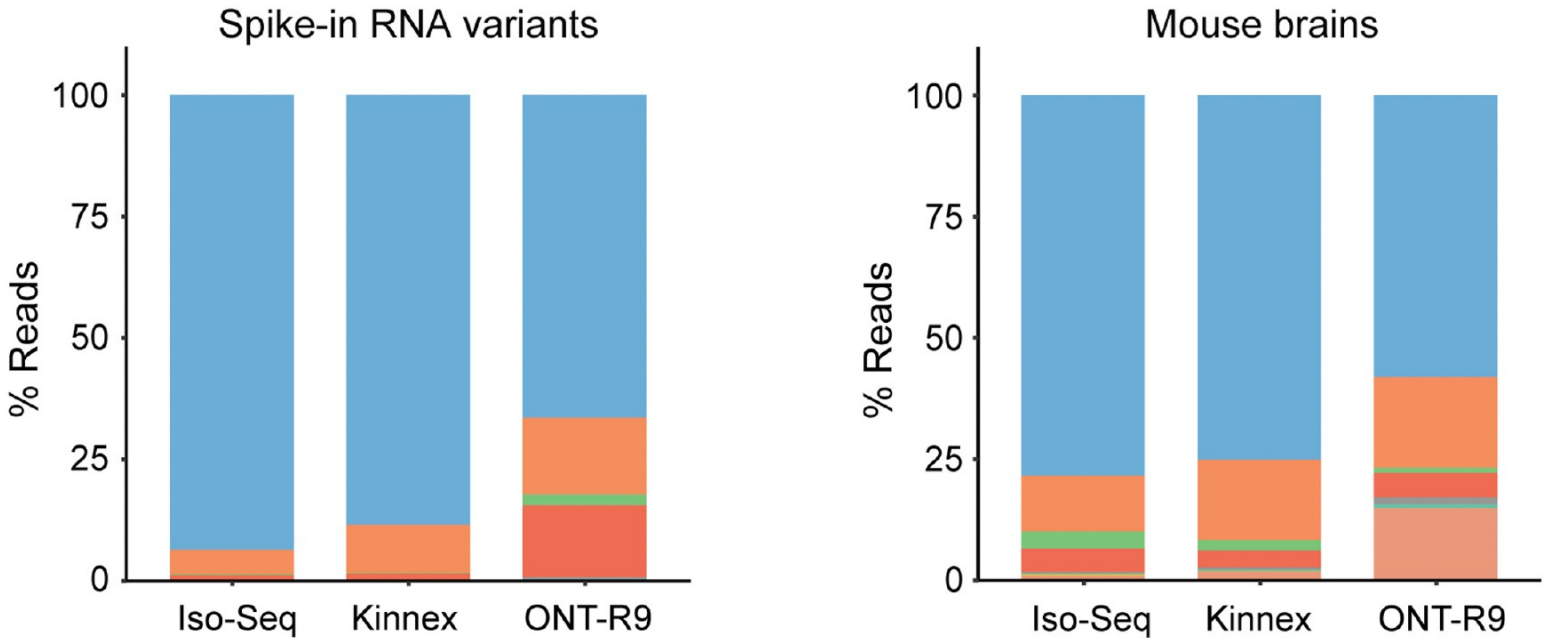
Ground truth spike-ins



- Check if all SIRVs in the mix were detected in my sample (read length or depth biases)
- Check that quantification of SIRVs in the mix is as expected

Figure source: https://www.lexogen.com/sirvs/design/

Ground truth spike-ins



- Non-blue reads in **spike-ins** represent the technical **noise level of the technology**
- Non-blue reads in the **mouse brains** are a combination of **technical and biological noise**

Figure source: Monzó, Frankish, Conesa. 2025. 10.1101/gr.279865.124

Use cases and wrap up

What's the most important thing to take into account when:

1. Studying drought effect on polyA length in Arabidopsis over time

2. Reconstructing the transcriptome of a non-model organism

3. Benchmarking bioinformatic tools for transcript quantification

What is the most important aspect in the experimental design to answer your biological question?

Take home messages

- **Figure out your biological question before deciding on sequencing platform and library preparation method**
- **Use good quality RNA**
- **Sequence with enough depth**
- **Use biological replicates in your experiment**
- **Use synthetic controls**

Thank You!

For more information about the LongTREC Summer School:

https://longtrec.eu