Funded by
the European Union

Bioinformatics Summer School

Long-reads Transcriptomics

Yalan Bi & Fabian Jetzinger

Max Plank Institute, Berlin, Germany
BioBam Bioinformatics, Valencia, Spain

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

Course Contents

1. Challenges in transcript identification
2. Landscape of bioinformatic tools

Section 1

Challenges in transcript identification

Understanding the complexity of
transcript identification and quantification

Long-read RNAseq data sets contain millions of reads

- Millions of reads, many minor differences
- **Goal:** collapse similar reads to **identify** consistently occurring **transcript models**
- Identify **known** (annotated) and **novel** transcripts
- Many different terms for this process
  - Transcript | Isoform identification | discovery
  - Transcriptome reconstruction
  - Transcript(ome) assembly (holdover from short reads)

Challenges in transcript identification

Long-read RNAseq data poses many challenges

- Sequencing quality and depth
- Biases (e.g. towards sequencing shorter molecules, higher GC content, etc.)
- Mapping inaccuracies (e.g. micro-exons)
- Incomplete reads
- Transcript divergency (minor inconsistent variations in splicing)

Goal: differentiating biological and technical *artifacts* from real *transcripts*

Section 2

Landscape of bioinformatic tools

Understanding strategies for
transcript identification and quantification

Isoform identification and quantification strategies

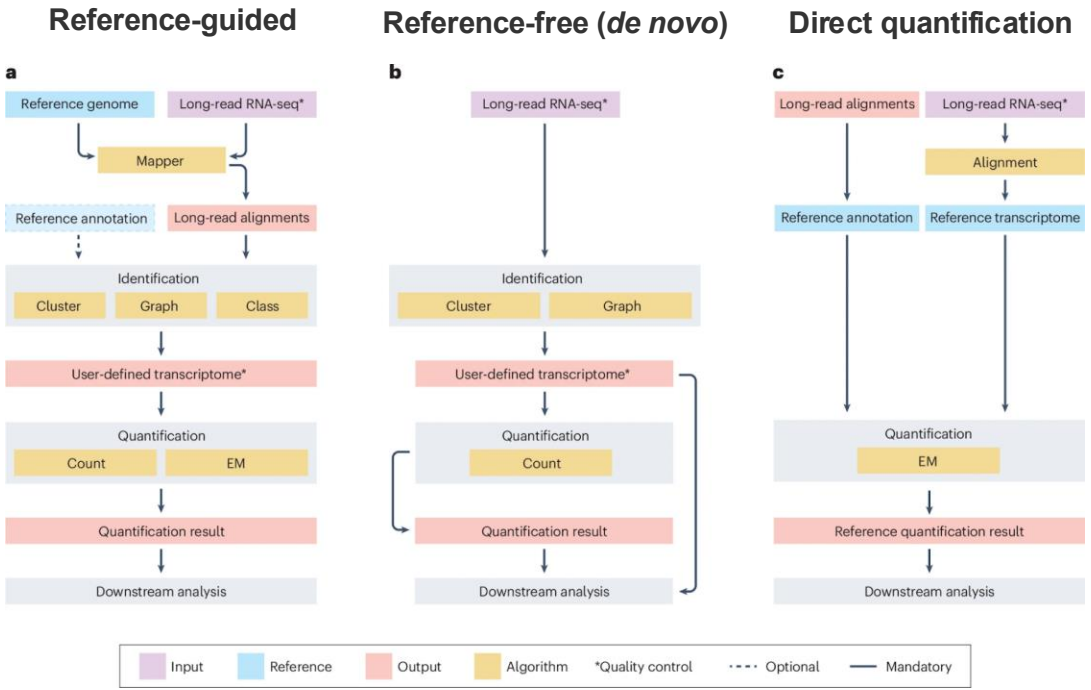

Diagram: *Monzó C, Liu T, Conesa A. Transcriptomics in the era of long-read sequencing. Nat Rev Genet. 2025 Mar 28. doi: 10.1038/s41576-025-00828-z. Epub ahead of print. PMID: 40155769.*

Reference-guided

- Identify isoforms based on:
  - Reference genome
  - Reference annotation
- IsoTools, IsoQuant, FLAIR, Bambu, etc.

Reference-free

- Without any reference information
- isON-pipeline, RNA-Bloom2, RATTLE, etc.

Direct quantification

- No discovery of novel isoforms
- Oarfish, lr-kallisto, LIQA, Nanocount, etc.

LRGASP – one benchmark to rule them all

- **Challenge 1:** Reference-guided isoform identification
- **Challenge 2:** Quantification
- **Challenge 3:** Reference-free isoform identification

Final recommendations

- Prefer sequence quality for isoform identification and sequencing depth for quantification
- To study known isoforms: **Bambu, IsoQuant, FLAIR**
- To study rare, novel isoforms: include orthogonal short reads with **FLAIR, Mandalorion**
- Quantification: **IsoQuant, FLAIR, Bambu**
- To create reference annotations: high-quality data, replicates, orthogonal data, multiple tools

Pardo-Palacios, F.J., Wang, D., Reese, F. et al. Systematic assessment of long-read RNA-seq methods for transcript identification and quantification. Nat Methods 21, 1349–1363 (2024). https://doi.org/10.1038/s41592-024-02298-3

In the following situations, which challenges do you face, and which type of tool would you choose?

Reminder of challenges:
Sequencing quality and depth, biases (e.g. length), mapping inaccuracies, incomplete reads, transcript divergency

Reminder of tool options:
1. Transcriptome reconstruction reliant on reference genome (and annotation)
2. De novo transcriptome reconstruction without any reference information
3. Reference-based quantification only (without transcriptome reconstruction)

- You have lrRNA-seq data of an organism that has not been sequenced before.
- You have lrRNA-seq data of a model organism but are not interested in discovering novel isoforms.
- You have lrRNA-seq data of a model organism and are interested in condition-specific (e.g. disease) differences in alternative splicing, including novel isoforms.
- You have lrRNA-seq data of an organism with a relatively reliable reference genome, but no good reference annotations.

Questions?

For more information about the LongTREC Summer School:

https://longtrec.eu

Thank You!

For more information about the LongTREC Summer School:

https://longtrec.eu