LongTREC

Funded by
the European Union

SQANTI-reads

Quality assessment of long-read data in multi-sample lrRNA-seq experiments

Tianyuan Liu

LongTREC Summer School 2025

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

Course Contents

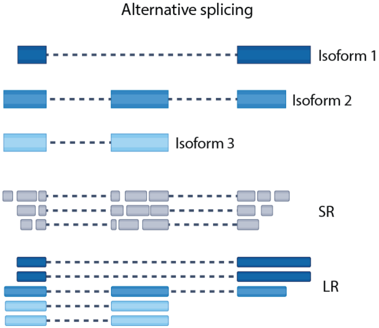
1. Background
2. Introducing SQANTI-reads
3. Case Studies
4. Summary

Section 1

Background

lrRNA-seq captures full-length transcripts

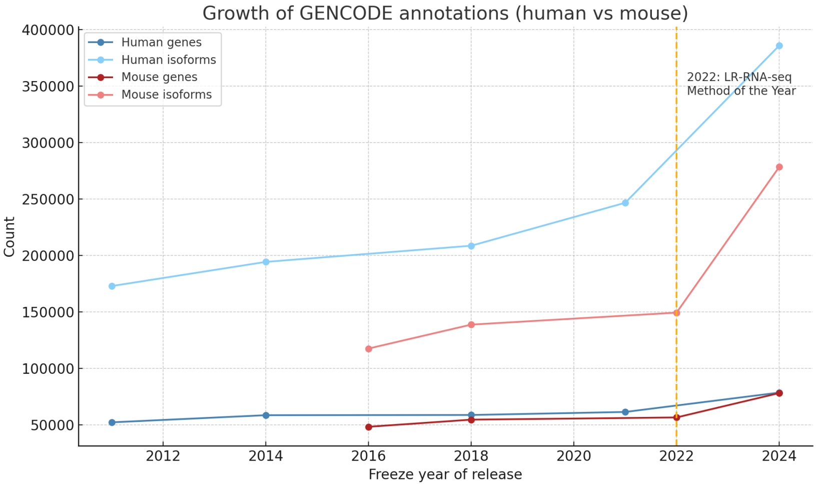
- Long-read RNA sequencing directly sequences **full-length cDNA molecules**, preserving isoform structure [1].



Source: Monzó *et al.*, 2025

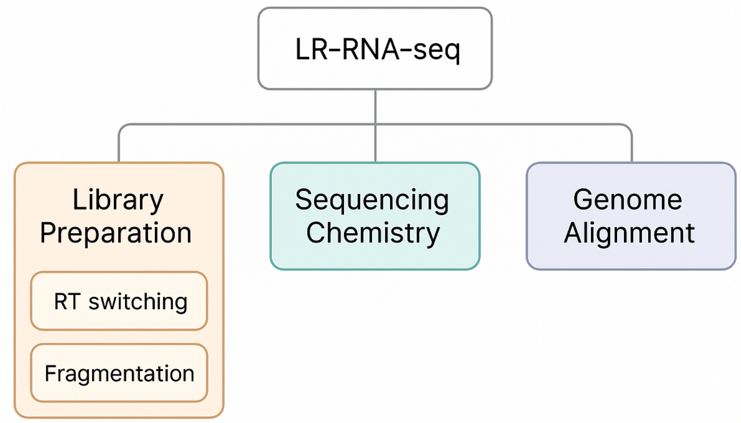
lrRNA-seq reveals transcript diversity

LongTREC



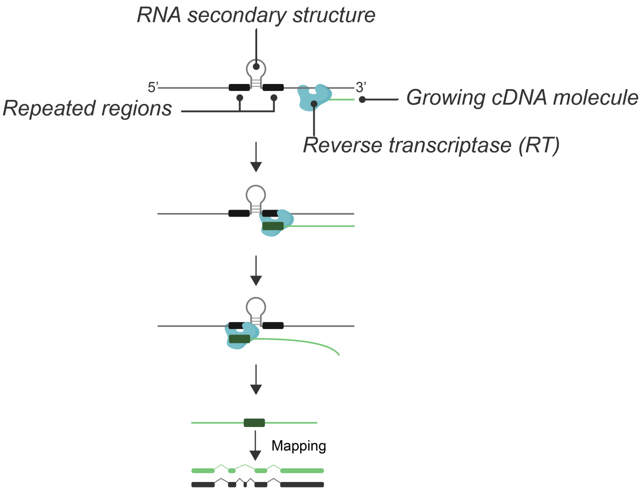
Source: GENCODE (Frankish *et al.*, 2025)

Sources of error in lrRNA-seq



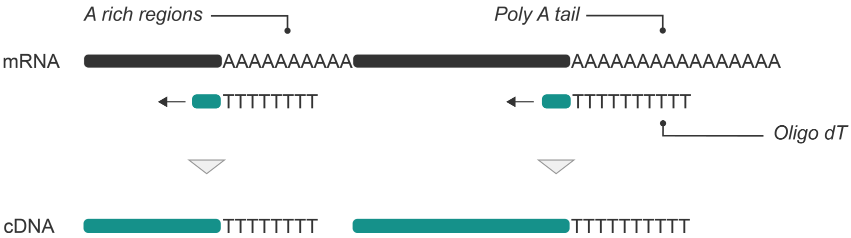
Errors: **library preparation** *(RT switching, degradation)*, **sequencing chemistry**, *or* **genome alignment**.

Reverse Transcriptase Template Switching (RT-switching)



RT enzyme can "jump" between templates at repeated regions, creating chimeric cDNA products that confound transcript analysis.

Intra-priming



Oligo-dT primers can bind to internal A-rich regions, causing truncated cDNA synthesis.

QC in lrRNA-seq: Importance & Current Limitations

Why rigorous QC is essential

- QC filters problematic reads, ensuring reliable downstream **quantification** and **novel isoform discovery** [3].
- Facilitates cross-sample comparison and prevents confounding technical artefacts.

Limitations of traditional QC tools

- Most QC tools evaluate **read-level metrics** (length, Phred quality) but overlook **transcript structure**.
- Structural context is key for identifying mis-spliced or truncated reads [4].

Check-in Questions: Background

1. What are the main advantages of long-read RNA sequencing over short-read sequencing for transcript analysis?

2. Can you name two major sources of error that can occur during lrRNA-seq library preparation?

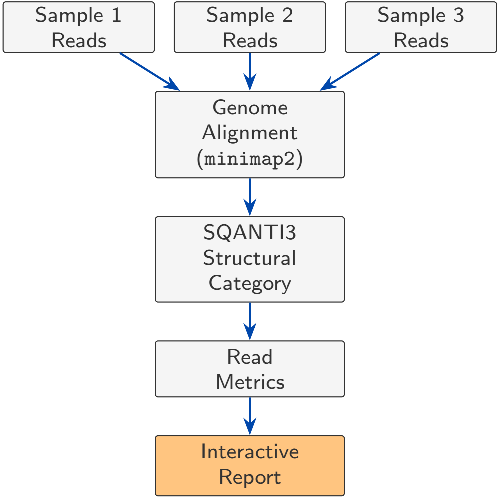
Section 2

Introducing `SQANTI-reads`

What is SQANTI-reads?

A read-centric extension of SQANTI3

- Ports **SQANTI3** structural classification to the **single-read** level.
- Jointly evaluates *raw reads* from **multiple samples** in one run.
- Summarises structural categories, splicing patterns, and junction usage.
- Produces *interactive* visualisations to spot outliers and under-annotated genes.



SQANTI-reads: Inputs & Outputs

Core Inputs

- **Design file (CSV)** – columns `sampleID`, `file_acc`.
- **Reference annotation** (GTF/GFF3).

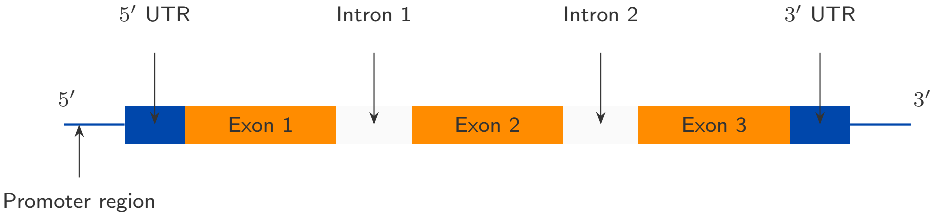
Mode – dependent

- *Fast mode:* pre-computed SQANTI3-QC output directories (given via `--input_dir`).
- *Simple mode:* raw reads (`*.fastq`) or sample GTF/GFF
  + reference genome FASTA.

Key Outputs

- Modified `reads_classification.txt` (adds `jxn_string`, `jxnHash`).
- Updated `design.csv` (adds `classification_file`, `junction_file`).
- Summary CSV tables: gene_counts, ujc_counts, length_summary, cv, etc.
- QC plots PDF (default) & optional HTML report.
- Annotation plots PDF.

Structure of a eukaryotic gene



Schematic of a canonical eukaryotic gene: promoter (blue), untranslated regions, coding exons (orange), introns (grey), and transcriptional orientation from $5'$ to $3'$.

GTF vs GFF3: Key Differences

- Both share the first eight columns: `seqid`, `source`, `type`, `start`, `end`, `score`, `strand`, `phase`.
- **Attribute syntax**
  - **GTF**: `<key> "value";` (semicolon-terminated key–value pairs).
  - **GFF3**: `key=value` (comma-separated if multiple) with `ID` / `Parent` tags enabling feature hierarchies.
- **Specification status**: GTF is legacy (GFF2-derived); GFF3 is the current, more flexible standard.

Example Records: GTF vs GFF3

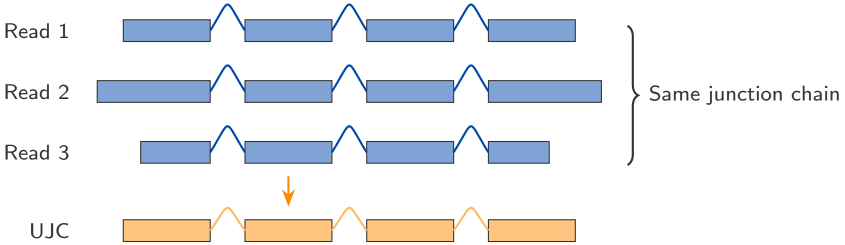
GTF

```
chr1 HAVANA gene            11869 14409 . + . gene_id "ENSG00000223972";
chr1 HAVANA transcript      11869 14409 . + . gene_id "ENSG00000223972"; transcript_id "ENST00000456328";
chr1 HAVANA exon            11869 12227 . + . gene_id "ENSG00000223972"; transcript_id "ENST00000456328"; exon_number "1";
chr1 HAVANA five_prime_UTR  11869 12009 . + . gene_id "ENSG00000223972"; transcript_id "ENST00000456328";
chr1 HAVANA CDS             12010 12057 . + 0 gene_id "ENSG00000223972"; transcript_id "ENST00000456328";
chr1 HAVANA three_prime_UTR 12058 14409 . + . gene_id "ENSG00000223972"; transcript_id "ENST00000456328";
```

GFF3

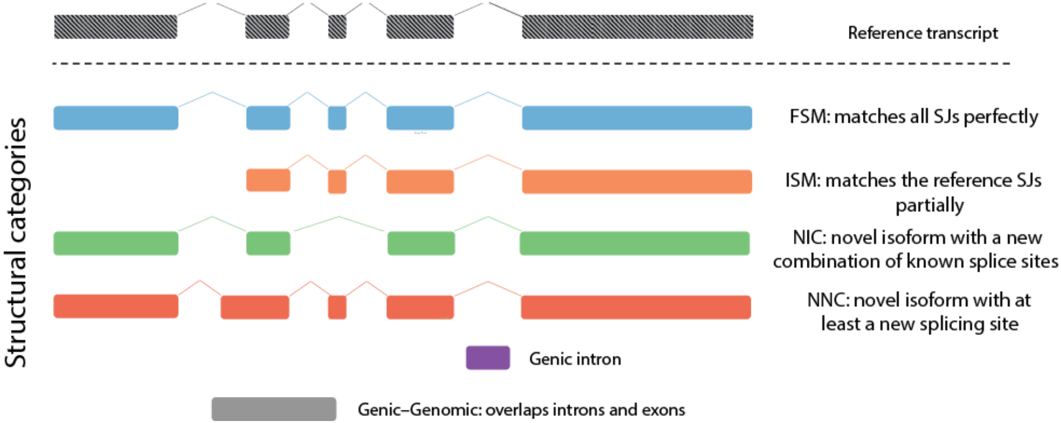
```
chr1 HAVANA gene            11869 14409 . + . ID=gene0;Name=DDX11L1;
chr1 HAVANA transcript      11869 14409 . + . ID=transcript0;Parent=gene0;
chr1 HAVANA exon            11869 12227 . + . ID=exon0;Parent=transcript0;
chr1 HAVANA five_prime_UTR  11869 12009 . + . ID=futr0;Parent=transcript0;
chr1 HAVANA CDS             12010 12057 . + 0 ID=cds0;Parent=transcript0;
chr1 HAVANA three_prime_UTR 12058 14409 . + . ID=tutr0;Parent=transcript0;
```

Key Features: Unique Junction Chain



Reads have variable TSS/TTS but share the same ordered splice junctions.
Such reads collapse into **one** Unique Junction Chain

Key Features: SQANTI3 Structural Category


Reference transcript
Structural categories
FSM: matches all SJs perfectly
ISM: matches the reference SJs partially
NIC: novel isoform with a new combination of known splice sites
NNC: novel isoform with at least a new splicing site
Genic intron
Genic–Genomic: overlaps introns and exons

Source: Pardo-Palacios *et al.*, 2024

QC Feature Block A: Read-length (8 metrics)

No.	Metric
1	Mean read length
2	Median read length
3	Upper-quartile read length
4	Lower-quartile read length
5	% reads <1 kb
6	% reads 1–2 kb
7	% reads 2–3 kb
8	% reads >3 kb

- Captures depth and size bias to reveal degradation or protocol differences.

QC Feature Block B: Read Structural Categories (8 metrics)

No.	Metric
9	% Full-Splice-Match (FSM) reads
10	% Incomplete-Splice-Match (ISM) reads
11	% Novel-In-Catalog (NIC) reads
12	% Novel-Not-in-Catalog (NNC) reads
13	% Antisense reads
14	% Fusion reads
15	% Genic–genomic reads
16	% Intergenic reads

- Describes transcript integrity and novelty at the single-read level.

QC Feature Block C: UJC Structural Categories (8 metrics)

No.	Metric
17	% FSM UJCs
18	% ISM UJCs
19	% NIC UJCs
20	% NNC UJCs
21	% Antisense UJCs
22	% Fusion UJCs
23	% Genic–genomic UJCs
24	% Intergenic UJCs

- Collapses reads with identical junction chains, highlighting transcript-level novelty.

QC Feature Block D: Splice-junction Quality (4 metrics)

No.	Metric
25	% Known–canonical junctions
26	% Known–non-canonical junctions
27	% Novel–canonical junctions
28	% Novel–non-canonical junctions

- Assesses splice-site accuracy—critical for long-read platforms.

QC Feature Block E: Error / Artefact Flags (3 metrics)

No.	Metric
29	% Reads with RT-switching repeat
30	% Reads with intrapriming signature
31	% Reads containing $\geq$1 non-canonical junction

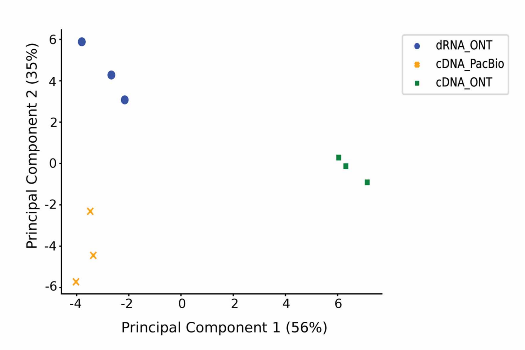
- Flag well-known long-read artefacts to aid filtering.

QC Feature Block F: Coverage & Complexity (4 metrics)

No.	Metric
32	Total mapped read count
33	Mean reads per gene
34	Mean reads per UJC
35	Genes with ≥1 FSM read

- Summarises usable expression breadth and sequencing efficiency.

Example: PCA on QC Features



Key insights from PCA:

- Samples cluster by **sequencing technology**
- PC1 (56%) separates **cDNA ONT** from **dRNA ONT** and **cDNA PacBio**
- PC2 (35%) separates **dRNA ONT** from **cDNA PacBio**
- Clear technology-specific biases
- Enables outlier detection within technology groups

Check-in Questions: SQANTI-reads Features

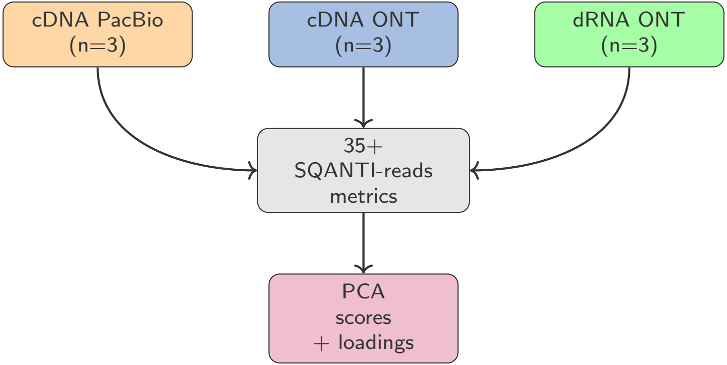
1. What is the difference between analyzing reads versus UJCs (Unique Junction Chains)?

2. Which SQANTI3 structural categories would you expect to see more of in high-quality vs. low-quality samples? Why might the percentage of FSM (Full-Splice-Match) reads be an important QC metric?

Section 3

Case Studies

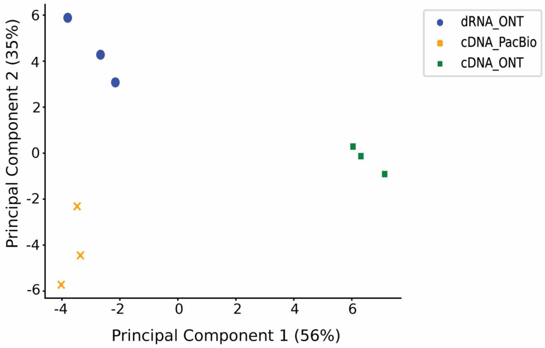
LRGASP WTC11: Experimental Design



Triplicate transcriptome measurements of the WTC11 human cell line. Three long-read sequencing (LRS) library types were compared: cDNA PacBio Sequel II (cDNA PacBio), cDNA Oxford Nanopore MinION (cDNA ONT), Direct-RNA Oxford Nanopore MinION (dRNA ONT).

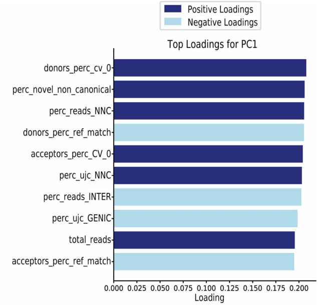
PCA Scores: Technologies Form Distinct Clusters

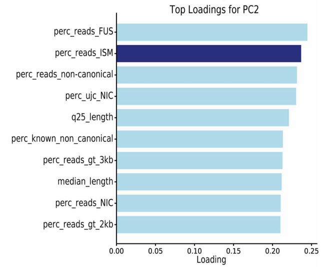
- PC1 (56% variance) separates **cDNA ONT** from PacBio and dRNA ONT.
- PC2 (35%) further discriminates **dRNA ONT** from **cDNA PacBio** libraries.
- Clear clustering indicates technology-specific QC signatures.



PCA Loadings: Drivers of Variance

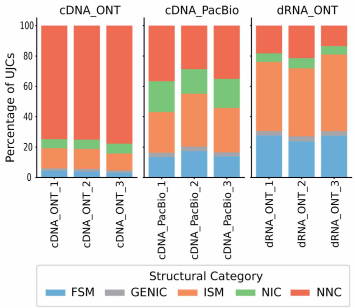
- High **positive** PC1 loadings: total number of reads, % NNC reads, and % UJCs in the NNC category.
- High **negative** PC1 loadings: Intergenic and Genic–Genomic reads.
- PC2 is dominated by read-length metrics (1–2 kb, 2–3 kb, > 3 kb).





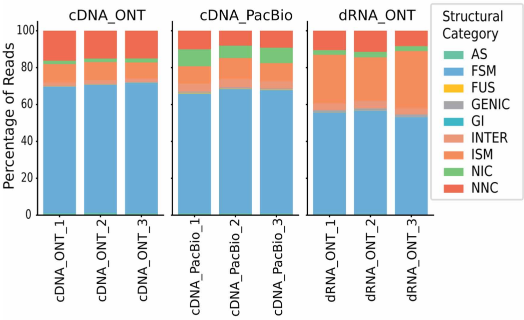
UJC Profiles Differ by Technology

- **cDNA ONT** libraries show the highest proportion of **NNC UJCs**.
- **dRNA ONT** and **cDNA PacBio** display lower UJC novelty, mirroring their lower NNC read fractions.
- Reinforces PCA PC1 loadings that highlight splice-junction novelty as a driver of variance.



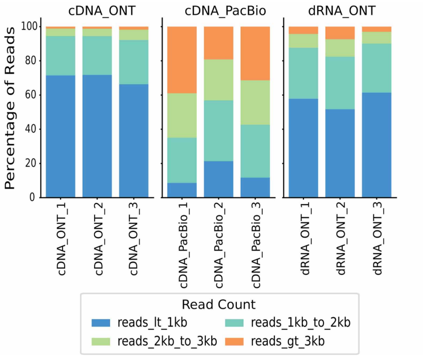
Structural Category Profiles Differ by Technology

- **cDNA ONT** exhibits the highest proportion of NNC reads and UJCs.
- **cDNA ONT** also shows the lowest fraction of intergenic reads.
- Confirms PC1 loadings and underscores library-prep biases.



Read-length Distributions Explain PC2 Separation

- **cDNA PacBio** enriched for longer reads (1–2 kb, 2–3 kb, higher than 3 kb).
- **dRNA ONT** dominated by reads < 1 kb.
- Aligns with strong read-length contributions to PC2.



Check-in Questions: Case Studies

1. What would you conclude if you saw a sample that clustered away from its expected technology group in PCA?

Section 4

Summary

Take-home Messages

- **Multi-sample dashboards**: stacked bars and heatmaps instantly reveal QC metric trends across libraries.
- **PCA explorer**: Scores & loadings plots pinpoint outliers, batch effects and technology biases.
- **Structural maps**: side-by-side visualisation of Unique Junction Chains and SQANTI3 categories clarifies read novelty.
- One-click toggle between *reads*, *UJCs* and *gene*-level views—all colour-coded with the SQANTI3 palette.

References

- Frankish, Adam et al. (2023). "GENCODE reference annotation for human and mouse genomes". In: *Nucleic Acids Research* 51.D1, pp. D100–D110. DOI: 10.1093/nar/gkac836.
- Garalde, Daniel R. et al. (2018). "Highly parallel direct RNA sequencing on an array of nanopores". In: *Nature Methods* 15.3, pp. 201–206. DOI: 10.1038/nmeth.4577.
- Monzó, Cristina, Tianyuan Liu, and Ana Conesa (2025). "Transcriptomics in the era of long-read sequencing". In: *Nature Reviews Genetics* 26.5, ??? DOI: 10.1038/s41576-025-00828-z.
- Soneson, Charlotte et al. (2019). "A comprehensive examination of Nanopore native RNA sequencing for characterization of complex transcriptomes". In: *Nature Communications* 10.1, p. 3359. DOI: 10.1038/s41467-019-11272-z.
- Tardaguila, Manuel et al. (2018). "SQANTI: extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification". In: *Genome Research* 28.3, pp. 396–411. DOI: 10.1101/gr.222976.117.

Thank You!

LongTREC

Questions? Reach out at:

https://longtrec.eu

tianyuan.liu@csic.es