LongTREC Summer School 2026

FOUNDATIONS · LONG-READ RNA-SEQ QC

SQANTI-reads

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

Tianyuan Liu

June 25, 2026

Funded by the European Union

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

LongTREC

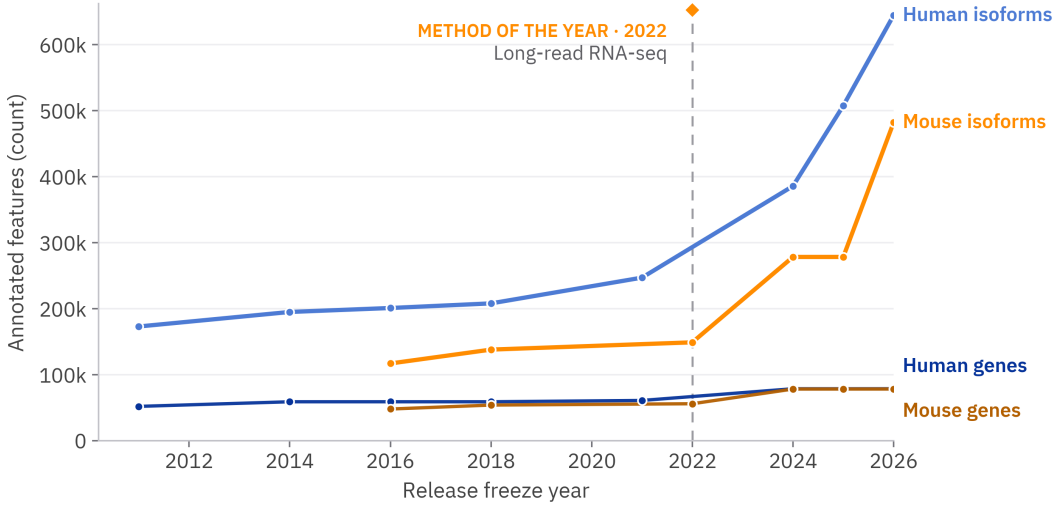
SECTION 1

Background

1

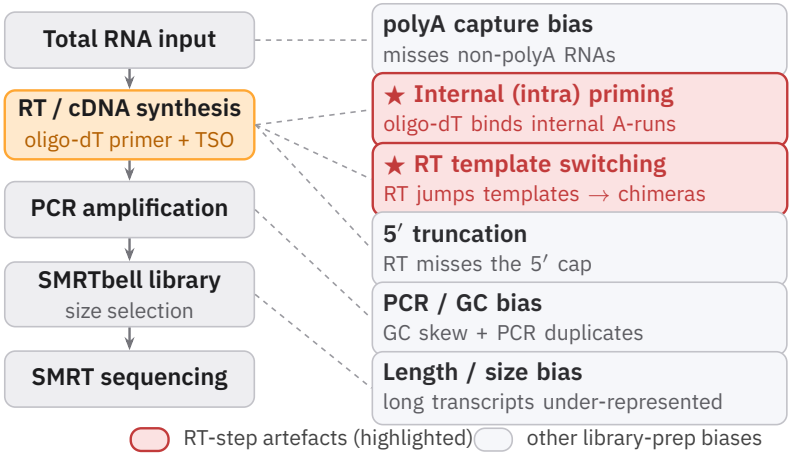
GENCODE ANNOTATION GROWTH

Isoforms surge while gene counts plateau — human vs mouse

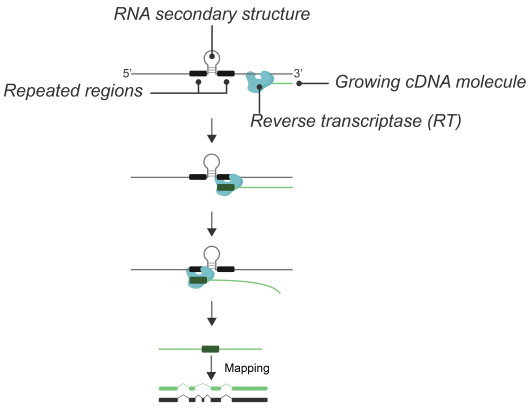


GENCODE total genes & transcripts (reference chromosomes); 2024–26 = human v47/v49/v50, mouse M36/M38/M39

Sources of error in lrRNA-seq

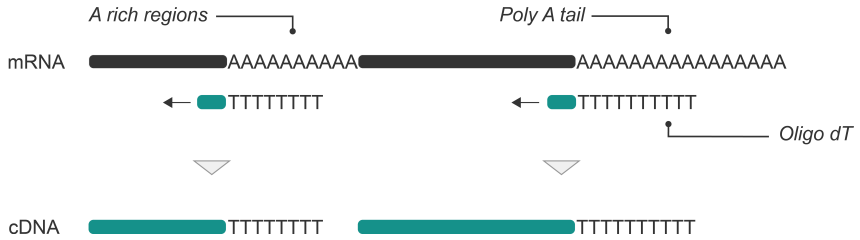


Template switching (RT-switching)



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

Intra-priming



Oligo-dT primers can mis-prime at internal genomically-encoded A-rich stretches, producing 3’-truncated cDNAs that mimic shortened transcript ends (a common source of artefactual novel 3’ ends and spurious ISM calls).

QC: 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 flags mis-spliced or truncated reads [4].

Check-in Questions: Background

1. What can **full-length** long reads resolve that short reads cannot — and name one thing short reads still do *better*.

2. Name two library-prep error sources — and for *each*, which **SQANTI3 category** would the reads fall into?

LongTREC

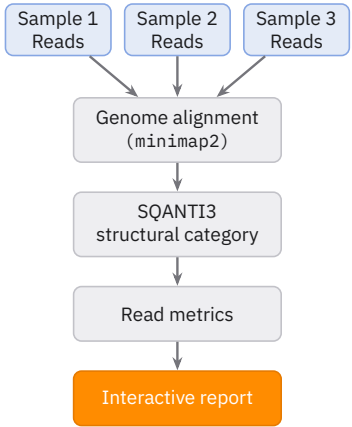
SECTION 2

Introducing SQANTI-reads

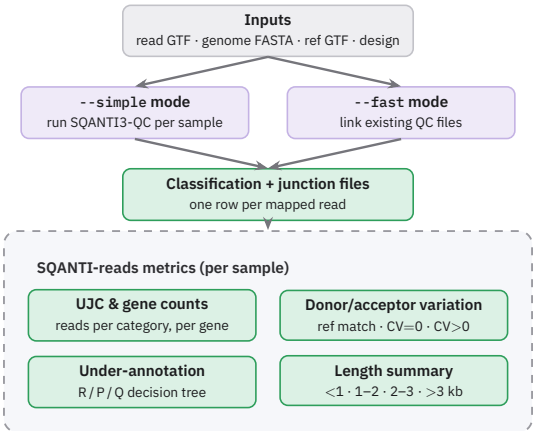
What is SQANTI-reads?

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.



The SQANTI-reads pipeline at a glance

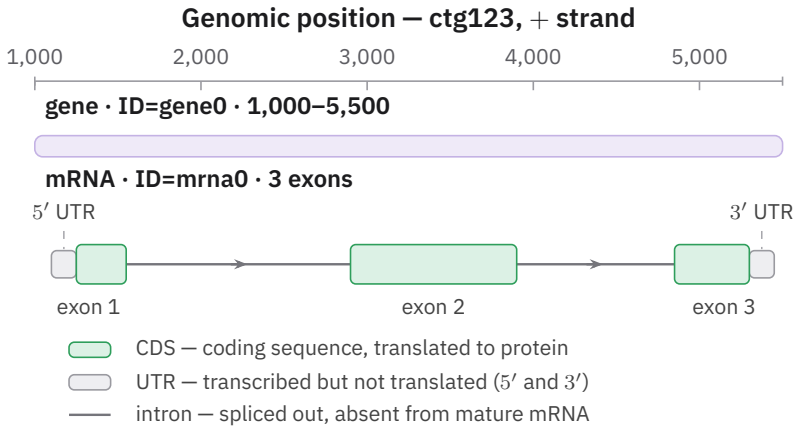


SQANTI-reads: Inputs & Outputs

Inputs	
Design file (CSV)	Sample sheet — columns `sampleID`, `file_acc`
Reference annotation	GTF / GFF3
Fast mode	Pre-computed SQANTI3-QC directories (`--input_dir`)
Simple mode	Raw reads (`*.fastq`) *or* sample GTF/GFF + genome FASTA

Outputs	
`reads_classification.txt`	Adds `jxn_string`, `jxnHash`
`design.csv`	Adds `classification_file`, `junction_file`
Summary CSV tables	`gene_counts`, `ujc_counts`, `length_summary`, `cv`, …
QC plots PDF	Default output; optional HTML report
Annotation plots PDF	Reference-vs-sample annotation summaries

Structure of a eukaryotic gene



GTF vs GFF3: Key Differences

First **eight columns are identical**: `seqid, source, type, start, end, score, strand, phase`.

The formats diverge *only* in **column 9 — the attributes**.

	GTF	GFF3
Attribute syntax	`<key> "value";`	`key=value;`
Pair separator	space-separated, `;`-terminated	`;`-separated (commas only within a multi-value attribute)
Hierarchy	parent IDs **repeated** on every line (flat)	explicit via `ID` / `Parent` tags
Spec status	legacy (GFF2-derived)	current, more flexible standard

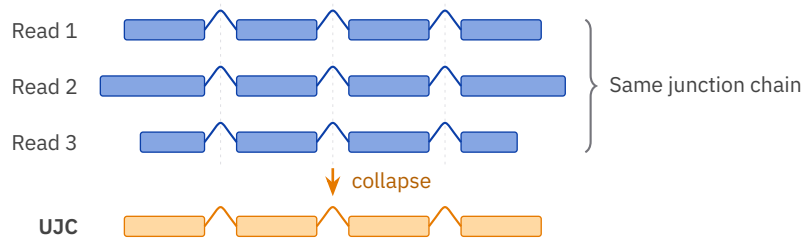
Example Records: GTF vs GFF3

The canonical `EDEN` gene (1 mRNA, 3 exons, spanning 1000–5500). Cols 1–8 match — only **column 9 (attributes)** differs.

Feature	**GTF** — flat, IDs repeated	**GFF3** — `ID` / `Parent` tree
gene	`gene_id "gene0"; gene_name "EDEN";`	`ID=gene0;Name=EDEN`
mRNA	`gene_id "gene0"; transcript_id "mrna0";`	`ID=mrna0;Parent=gene0; Name=EDEN.1`
exon ×3	`gene_id "gene0"; transcript_id "mrna0"; exon_number "1";`	`ID=exon0;Parent=mrna0`
5′ UTR	`gene_id "gene0"; transcript_id "mrna0";`	`Parent=mrna0`
CDS ×3	`gene_id "gene0"; transcript_id "mrna0";`	`ID=cds0;Parent=mrna0`
3′ UTR	`gene_id "gene0"; transcript_id "mrna0";`	`Parent=mrna0`

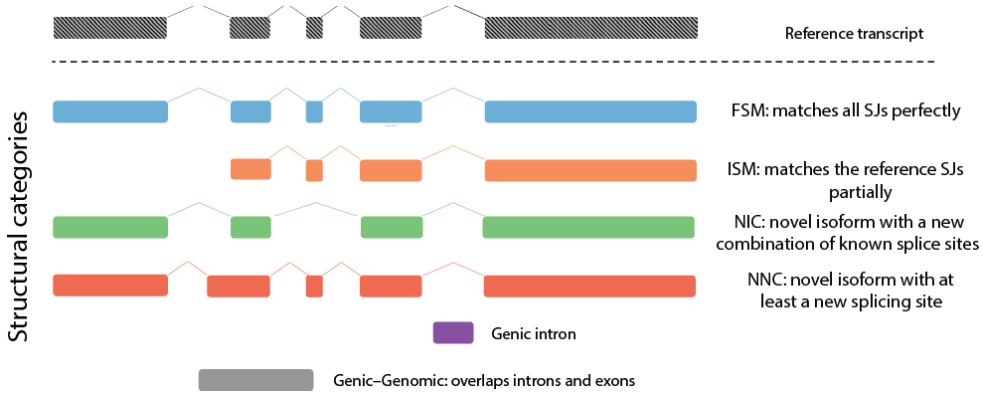
GTF repeats `gene_id` / `transcript_id` on **every** *line (flat); GFF3 names each `ID` once and links children via `Parent` (explicit tree). A few type names also differ, e.g. `five_prime_UTR`→`5UTR`.*

Key Features: Unique Junction Chain



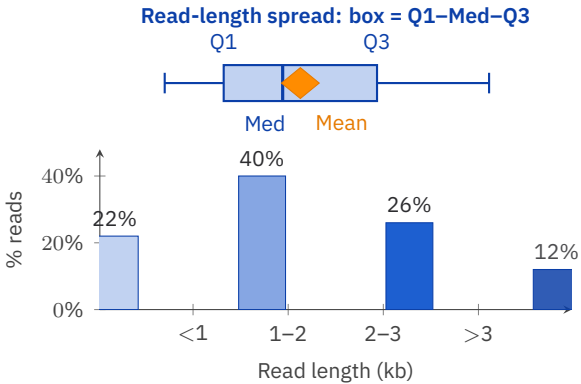
Reads vary in TSS/TTS but share the same ordered splice junctions — they collapse into **one** Unique Junction Chain.

Key Features: SQANTI3 Structural Category



Source: Pardo-Palacios *et al.*, 2024 | colours follow the canonical SQANTI3 palette (...)

Read-length & throughput (8 metrics)

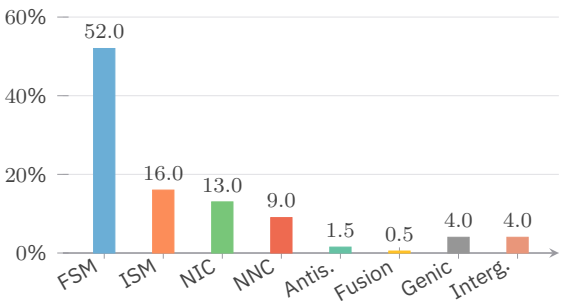


Illustrative read-length profile (WTC11-like PacBio cDNA).

#	Metric	Value
Summary statistics		
1	Q1 length	1.05 kb
2	Median length	1.55 kb
3	Mean length	1.70 kb
4	Q3 length	2.35 kb
Length-bin composition		
5	<1 kb	22%
6	1–2 kb	40%
7	2–3 kb	26%
8	>3 kb	12%

Captures depth & size bias—flags degradation or protocol shifts.

Read structural categories (8 metrics)

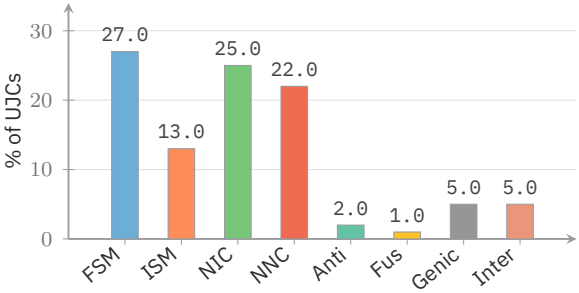


Illustrative single-read distribution (WTC11-like PacBio cDNA); canonical SQANTI3 palette.

No.	Metric	Value (%)
9	FSM	52.0
10	ISM	16.0
11	NIC	13.0
12	NNC	9.0
13	Antisense	1.5
14	Fusion	0.5
15	Genic-genomic	4.0
16	Intergenic	4.0

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

UJC structural categories (8 metrics)

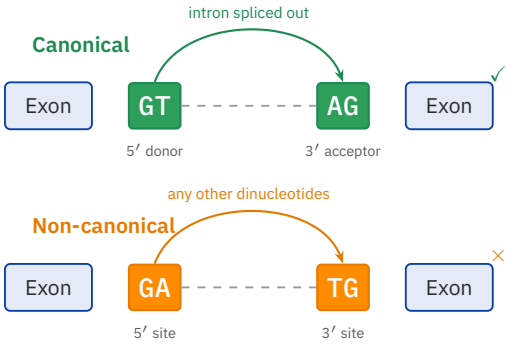


Illustrative UJC distribution (WTC11-like PacBio cDNA); canonical SQANTI3 palette.

No.	Metric	Value
17	% FSM UJCs	27.0
18	% ISM UJCs	13.0
19	% NIC UJCs	25.0
20	% NNC UJCs	22.0
21	% Antisense UJCs	2.0
22	% Fusion UJCs	1.0
23	% Genic-genomic UJCs	5.0
24	% Intergenic UJCs	5.0

Collapsing reads into UJCs lowers FSM's share (many reads, one chain) and raises novel chains (NIC/NNC), each backed by few reads.

Splice-junction quality (4 metrics)



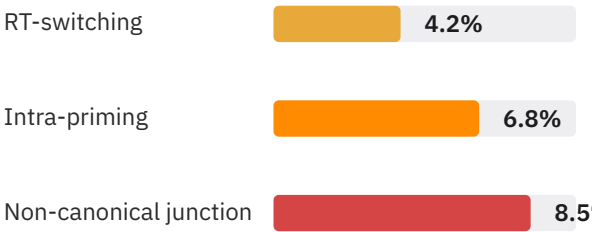
Canonical = GT· · · AG motif
Non-canonical = any other
Known / Novel = in / not in annotation

No.	Metric	Value
25	% Known canonical	91.2%
26	% Known non-canonical	1.8%
27	% Novel canonical	5.6%
28	% Novel non-canonical	1.4%

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

Illustrative junction breakdown (WTC11-like PacBio cDNA).

Error / artefact flags (3 metrics)



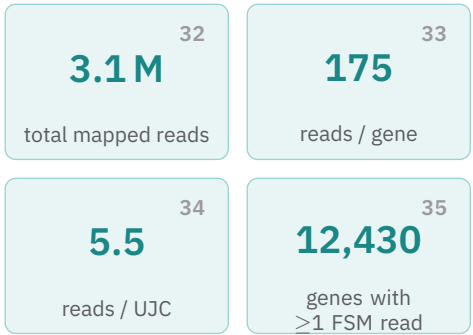
lower is better

No.	Metric	%
29	RT-switching	4.2
30	Intra-priming	6.8
31	≥1 non-canon. junction	8.5

Flags well-known long-read artefacts to aid **filtering**.

Illustrative artefact rates (WTC11-like PacBio cDNA).

Coverage & complexity (4 metrics)

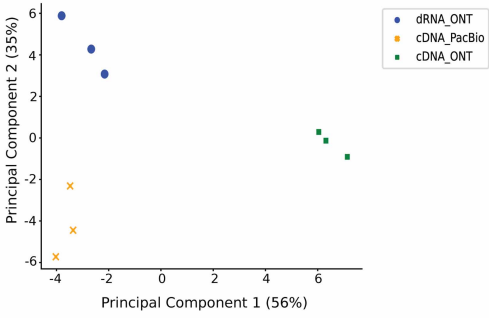


No.	Metric	Value
32	Total mapped reads	3.1 M
33	Mean reads / gene	175
34	Mean reads / UJC	5.5
35	Genes ≥1 FSM read	12,430

Summarises usable expression breadth & sequencing efficiency.

Illustrative coverage summary (WTC11-like PacBio cDNA).

Example: PCA on QC Features



Key insights from PCA:

- Samples cluster by **sequencing technology**
- PC1 (56%) separates **cDNA ONT** from **cDNA PacBio** and **dRNA ONT**
- 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. Reads vs. **UJCs**: what does each count, and when would the two *disagree* about how much novelty a sample has?

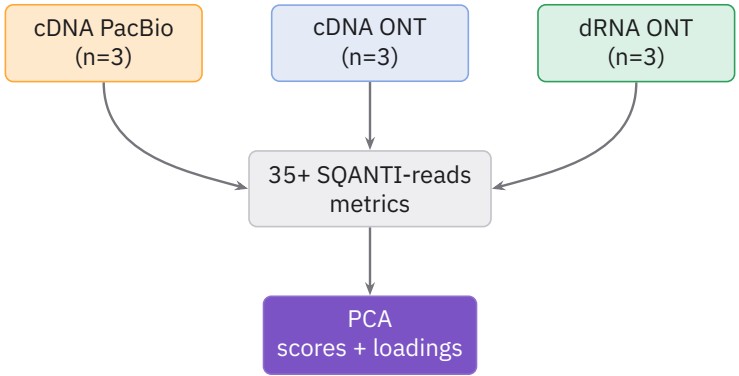
2. Same human reference: A is **70% FSM / 10% ISM**, B is **35% FSM / 45% ISM**. Which has better molecule integrity, and what do FSM and ISM each signal?

LongTREC

SECTION 3

Case Studies

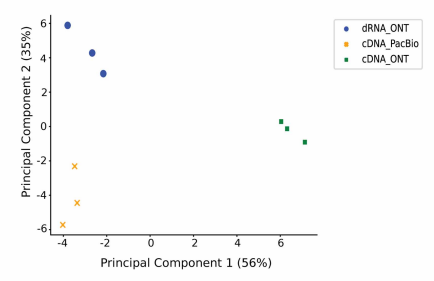
LRGASP WTC11: Experimental Design



Triplicate WTC11 transcriptome measurements across three long-read library types: cDNA PacBio (Sequel II), cDNA ONT (MinION), and direct-RNA ONT (MinION).

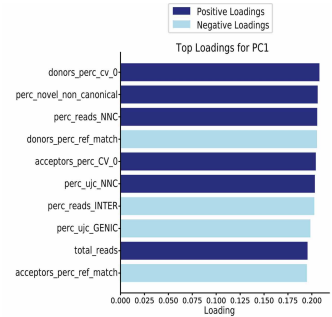
PCA Scores: Technologies Form Distinct Clusters

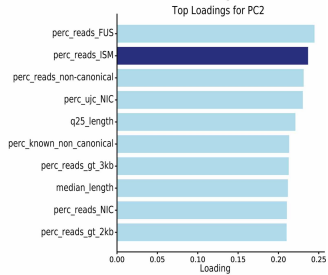
- PC1 (56% variance) separates **cDNA ONT** from cDNA PacBio and dRNA ONT.
- PC2 (35%) further discriminates **dRNA ONT** from **cDNA PacBio** libraries.
- Clear clustering indicates technology-specific QC signatures.



PCA Loadings: Drivers of Variance

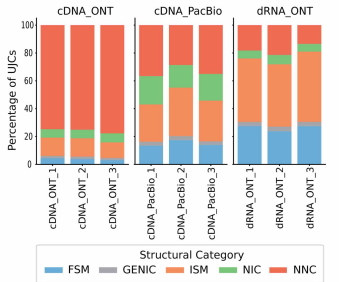
- High **positive** PC1 loadings: total number of reads, % NNC reads, and % UJCs in the NNC category.
- High **negative** PC1 loadings: Intergenic and Genic–Genomic reads.
- PC2 is dominated by read-length metrics (1–2 kb, 2–3 kb, > 3 kb).





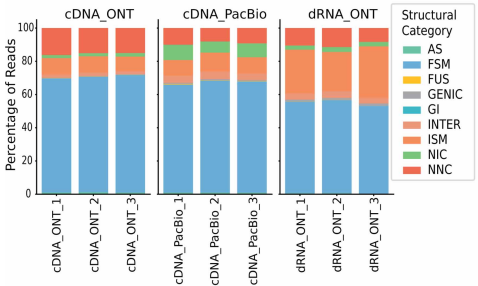
UJC Profiles Differ by Technology

- **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.



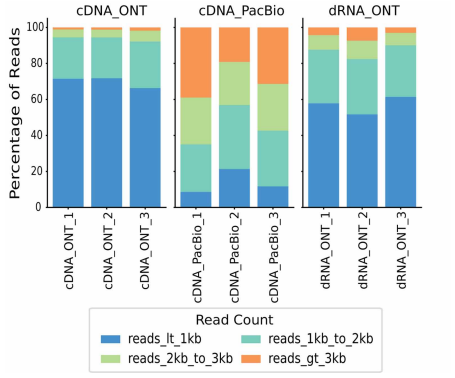
Structural Categories Differ by Technology

- **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 Lengths Drive 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.



Check-in Questions: Structural Categories & QC

1. **Internal priming** (oligo-dT binding an internal A-rich stretch): how does it distort the read's 3′ end, and why does it inflate **ISM** rather than FSM?

2. Why is **canonical-junction %** a useful QC metric — and why does it stay informative for a *non-model species* where FSM% fails?

3. **Going deeper:** Tumour vs. normal (same protocol/reference): the tumour has much *lower* FSM and more NIC, NNC, fusions. Real biology or artefact — and how would you decide?

LongTREC

SECTION 4

Summary

Take-home Messages

- **Multi-sample dashboards**: stacked bars, heatmaps and ridgeplots instantly reveal QC metric trends across libraries.
- **PCA explorer**: interactive 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.
- Garalde, Daniel R. et al. (2018). “Highly parallel direct RNA sequencing on an array of nanopores”. In: *Nature Methods* 15.
- Monzó, Cristina, Tianyuan Liu, and Ana Conesa (2025). “Transcriptomics in the era of long-read sequencing”. In: *Nature Reviews Genetics* 26.
- Soneson, Charlotte et al. (2019). “A comprehensive examination of Nanopore native RNA sequencing for characterization of complex transcriptomes”. In: *Nature Communications* 10.
- 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.
- Tilgner, Hagen et al. (2014). “Defining a personal, allele-specific, and single-molecule long-read transcriptome”. In: *Proceedings of the National Academy of Sciences* 111.

LongTREC

Thank you

Questions? Reach out at:

WEB https://longtrec.eu

EMAIL tianyuan.liu@csic.es

Funded by the European Union