Funded by
the European Union

Day 1 Practical Session

Long-read Transcriptome Analysis: Alignment and Quality Control

LongTREC Summer School

Practical Session

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

Practical Session Contents

1. Experiment Design and Dataset Creation
2. Hands-on Part 1: Minimap2 vs uLTRA Aligners
3. Hands-on Part 2: SQANTI-reads Concept
4. Summary

Section 1

Experiment Design and Dataset Creation

LRGASP Challenge 2

This challenge benchmarks long-read transcriptome analysis across platforms and cell types, aiming to reveal transcript diversity and expression. We use a subset of LRGASP Challenge 2.

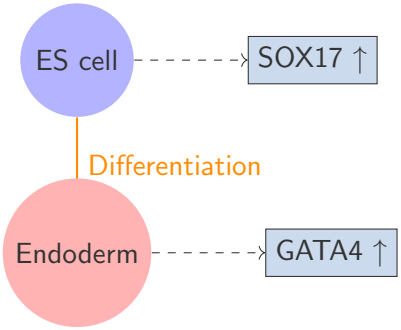
- **2** Cell lines: H1 (human embryonic stem cells) and H1-DE (definitive endodermal (DE))
- **1** Library type: cDNA
- **2** Platforms: Oxford Nanopore Technologies (ONT) and Pacific Biosciences (PacBio)
- **3** Biological replicates per condition
- Total samples: 12 experiments (2 cell lines $\times$ 1 library type $\times$ 2 platforms $\times$ 3 replicates)

Why We Selected Chromosome 8

Chromosome 8 hosts two crucial transcription factors that regulate endoderm differentiation:

- **SOX17** (8q11.23) - Master regulator for definitive endoderm commitment
- **GATA4** (8p23.1) - Key factor in endodermal maturation processes

These factors show dramatic expression changes during H1 to H1-DE transition, making them ideal markers for our analysis.



Section 2

Hands-on Part 1: Minimap2 vs uLTRA Aligners

Comparing Minimap2 and uLTRA Aligners

Minimap2

- General-purpose aligner for short, long, and RNA reads
- Fast and memory efficient
- Versatile across different read lengths and error profiles
- Supports RNA-seq with splicing handling

uLTRA

- Specialized for long RNA sequencing reads
- Uses two-pass collinear chaining for splice junctions
- Guided by exon annotations for higher accuracy
- Can wrap minimap2 for unannotated regions
- Slower and more memory-intensive than minimap2

Hands-on Part 1: Indexing and Alignment Workflow

1 Minimap2: Index & Align

Create index (mmi) file in 1 minute.
Align reads: cDNA ONT (3 min), dRNA ONT (3 min), cDNA PacBio (3 min).
Convert SAM to BAM format efficiently (1 min).

2 Minimap2: Pipeline

Run the full minimap2 alignment in a single shot (4 min).

3 uLTRA: Index & Align

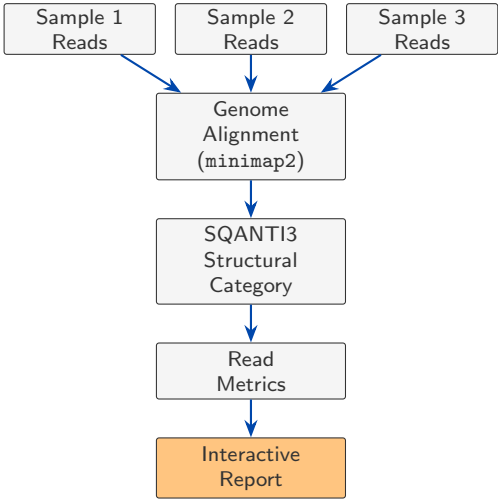
Generate index folder (1 min).
Align ONT reads (13 min), using specialized GTF reference for speed.
Bulk operations recommended due to longer runtime.

Section 3

Hands-on Part 2: SQANTI-reads Concept

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)

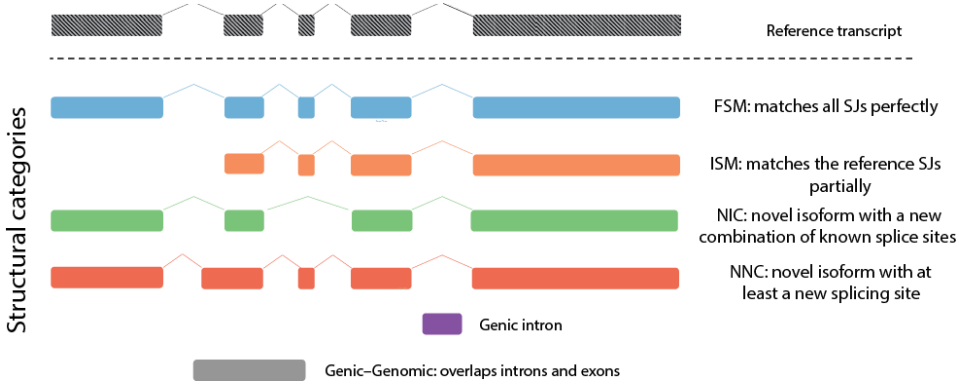
Mode-dependent

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

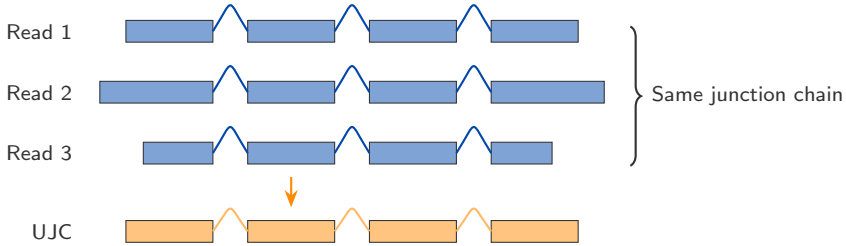
Key Outputs

- 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

Key Features: SQANTI3 Structural Category



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

Practical Exercise Overview

What you'll do:

- Run alignment workflows with both minimap2 and uLTRA
- Compare alignment performance and output quality
- Process alignment results through SQANTI-reads
- Analyze structural categories and junction patterns
- Generate QC reports and visualizations

Key learning outcomes:

- Understand trade-offs between different aligners
- Grasp the importance of read-level quality control
- Learn to interpret SQANTI3 structural categories
- Recognize patterns in junction usage across samples
- Identify potential technical artifacts in lrRNA-seq data

Section 4

Summary

Summary

- **Experiment design**: LRGASP Challenge 2 subset focusing on endoderm differentiation markers on chromosome 8
- **Dataset preparation**: Systematic pipeline to extract, process, and validate chromosome 8 reads
- **Alignment comparison**: Hands-on experience with minimap2 (general-purpose) vs uLTRA (RNA-specialized) aligners
- **Quality assessment**: Introduction to SQANTI-reads for comprehensive read-level QC
- **Structural analysis**: Understanding FSM, ISM, NIC, NNC categories and unique junction chains
- **Practical skills**: End-to-end workflow from raw reads to quality-controlled alignments

Thank You!

Questions about the practical session?

https://longtrec.eu

longtrec@earlham.ac.uk