SQANTI-single cell: Quality control and curation of long-read sequencing data and transcriptome assemblies at single cell resolution

Juan Francisco Cervilla Martínez[1], **Carlos Blanco**[2], Carolina Monzo[2], Wilfried Haerty[1], Ana Conesa[2]

Earlham Institute, Norwich Research Park, Norwich, UK
Genomics of Gene Expression Lab, Institute for Integrative Systems Biology (I2SysBio). Spanish National Research Council (CSIC).

Introduction

- **Long read** RNA-seq enables full-length transcripts characterization.
- **Single cell** resolution allows studying this full length molecules within a complex mixture of cells which cannot be performed using bulk.
- **Technical artifacts** and **transcriptome reconstruction** algorithms can introduce false positive isoforms, affecting downstream operations.
- **SQANTI3** curates transcriptomes using references and orthogonal data, but this tool was developed for bulk data.
  - To solve this, we introduce **SQANTI single-cell (SQANTI-sc)** to QC long read, single-cell data by adapting and expanding SQANTI3 and SQANTI-reads functions.
  - **SQANTI-sc** provides comprehensive reports to facilitate user decisions and downstream processing.

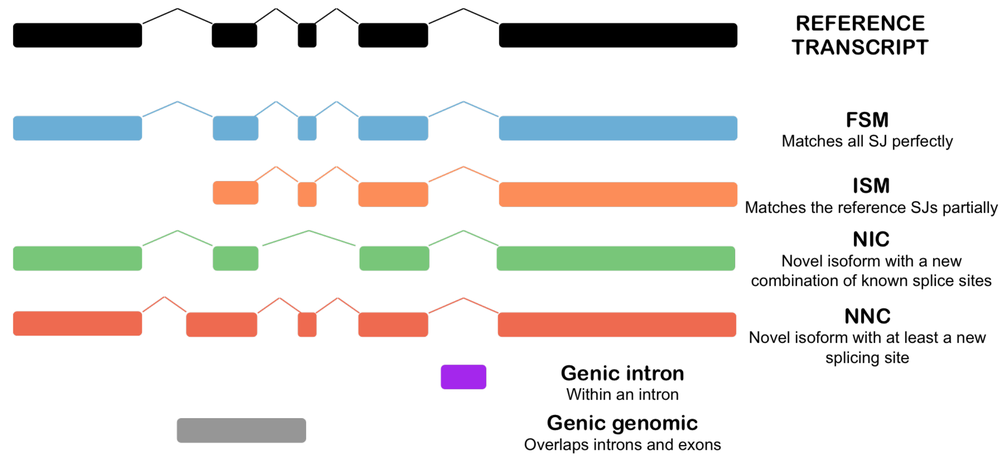

Figure 1. Overview of structural categories defined by SQANTI3.

SQANTI-sc workflow

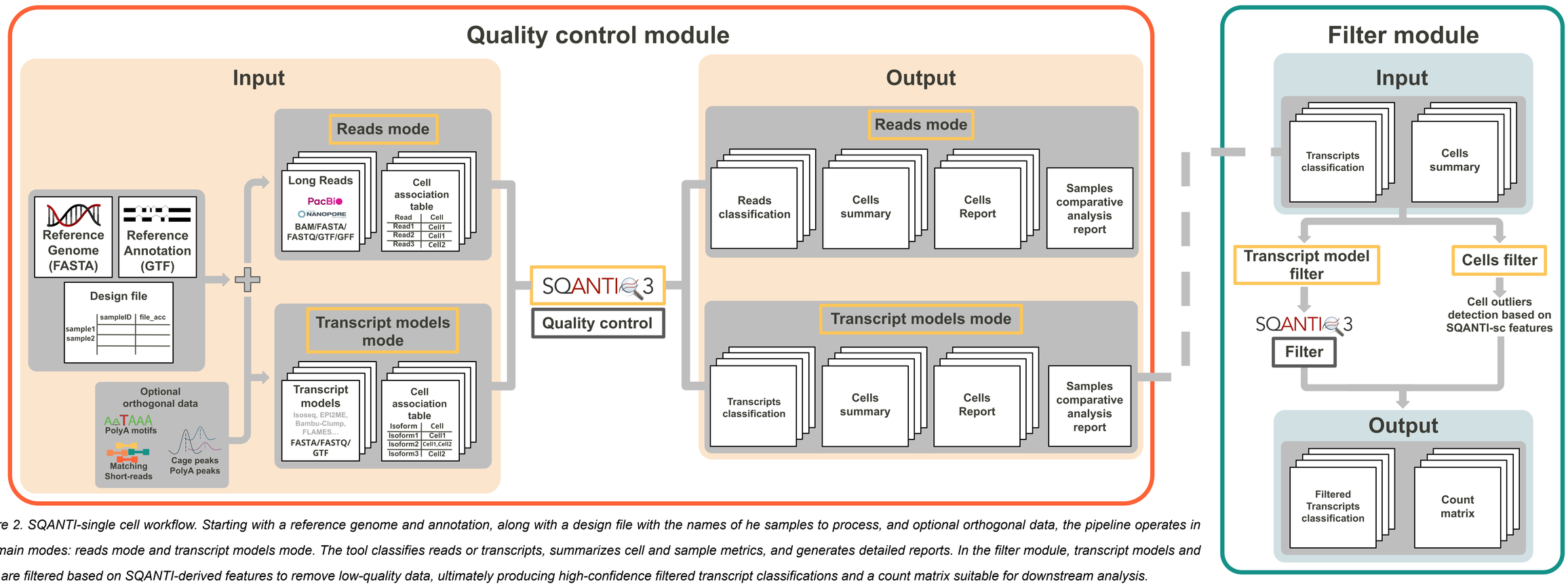

Figure 2. SQANTI-single cell workflow. Starting with a reference genome and annotation, along with a design file with the names of he samples to process, and optional orthogonal data, the pipeline operates in two main modes: reads mode and transcript models mode. The tool classifies reads or transcripts, summarizes cell and sample metrics, and generates detailed reports. In the filter module, transcript models and cells are filtered based on SQANTI-derived features to remove low-quality data, ultimately producing high-confidence filtered transcript classifications and a count matrix suitable for downstream analysis.

Results

SQANTI-sc QC helps understanding the quality of the transcriptome at the single cell level

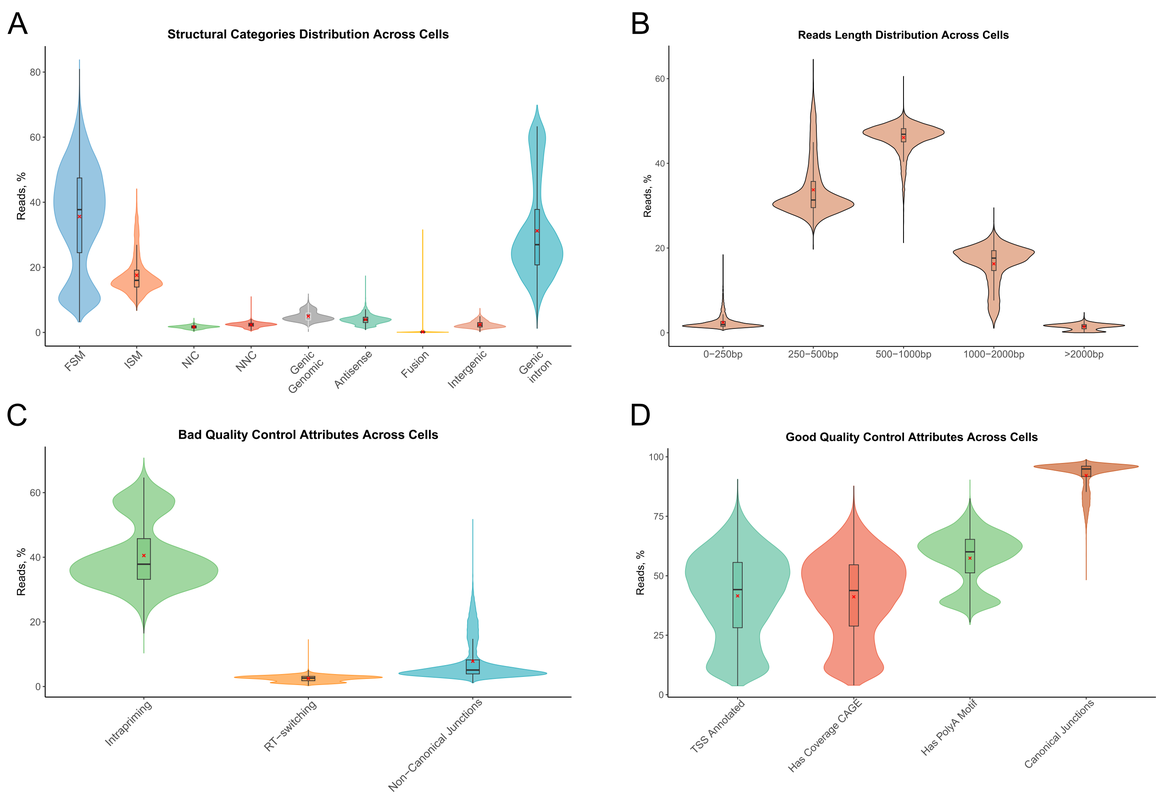

Figure 3. Examples of SQANTI-sc visualizations using PacBio's publicly available MAS-Seq 10x 3' PBMCs dataset sequenced on PacBio's Sequel II. Distributions of reads across cells by: (A) SQANTI structural category; (B) read length bin; (C) feature of bad quality; (D) feature of good quality.

Transcriptomic profile varies between cell types

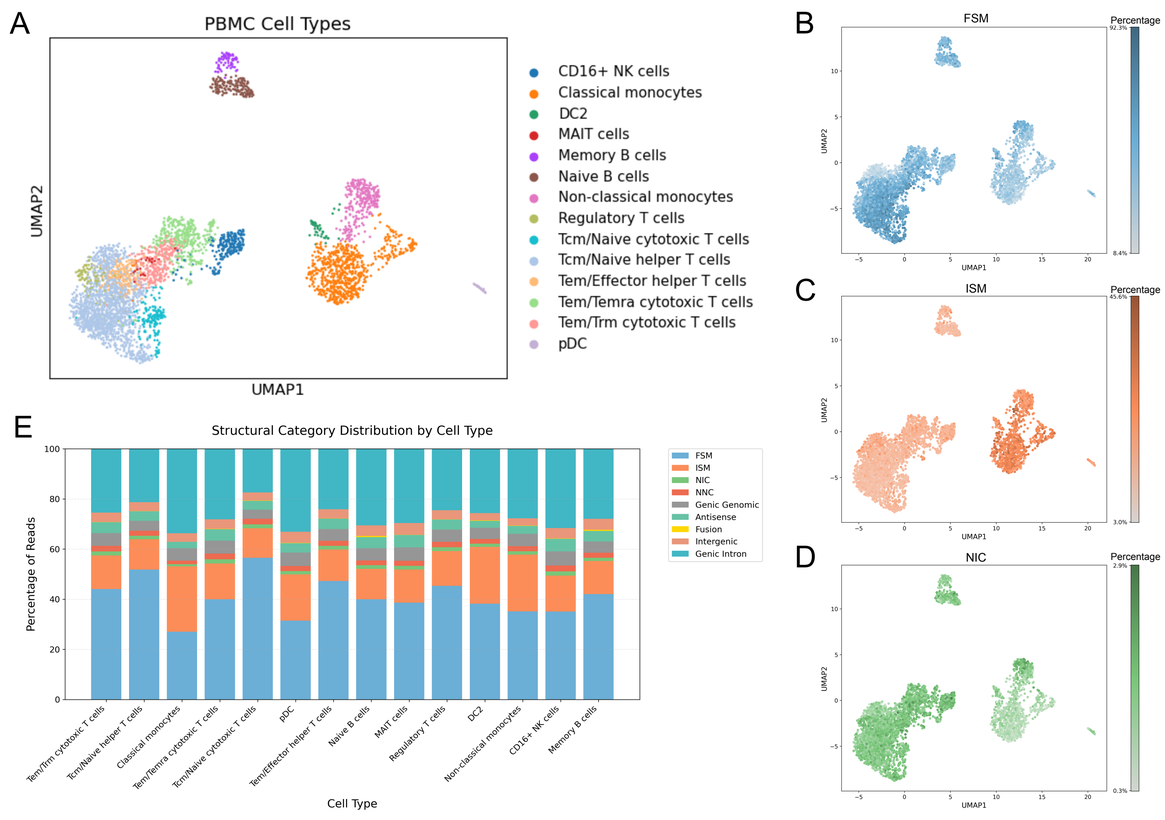

Figure 4. Examples of SQANTI-sc cell type analysis visualizations using PacBio's publicly available MAS-Seq 10x 3' PBMCs dataset sequenced on PacBio's Sequel II. (A) UMAP visualization using scanpy. Cell annotation was performed using Celltypist; (B-D) visualization of FSM, ISM and NIC using UMAP embedding; (E) stacked bar plots illustrating structural categories composition of each cell type identified using Celltypist.

Next steps

- Finish transcript models mode and filter module.
- Further testing the tool using additional datasets.
- Implement Nextflow pipeline.
- Containerization of SQANTI-sc.

References

Pardo-Palacios, F.J., Arzalluz-Luque, A. et al. SQANTI3: curation of long-read transcriptomes for accurate identification of known and novel isoforms. Nat Methods (2024). https://doi.org/10.1038/s41592-024-02229-2
Keil, N., Monzó, C., McIntyre, L., Conesa, A. Quality assessment of long read data in multisample lrRNA-seq experiments with SQANTI-reads. Genome Res (2025). https://doi.org/10.1101/gr.280021.124
Wolf, F. Alexander, Philipp Angerer, and Fabian J. Theis. "SCANPY: large-scale single-cell gene expression data analysis." Genome biology 19 (2018): 1-5.
Domínguez Conde, C., et al. "Cross-tissue immune cell analysis reveals tissue-specific features in humans." Science 376.6594 (2022): eabl5197.

Earlham Institute
norwich research park
Biotechnology and Biological Sciences Research Council
Genomics of Gene Expression Lab
CSIC Consejo Superior de Investigaciones Científicas
Universitat de València
sys bio
LongTREC
Engineering and Physical Sciences Research Council
Generalitat Valenciana
Conselleria d'Educació, Universitats i Ocupació
CIACIF/2023/124
Cofinanciado por la Unión Europea
Fondos Europeos
CSIC Consejo Superior de Investigaciones Científicas
Biología Computacional y Bioinformática (BCB)
AYUDAS PARA LA PARTICIPACIÓN A CONGRESOS INTERNACIONALES DE LOS GRUPOS DE LA CONEXIÓN BCB