SQANTI evidence

Long-read transcriptomics producing evidence-driven annotations

Genomics of Gene Expression Lab

Pablo Atienza López[1], Alejandro Paniagua[1], Ana Conesa Cegarra[1]

[1] Genomics of Gene Expression Lab, Institute for Integrative Systems Biology (I2SysBio). Spanish National Research Council (CSIC)

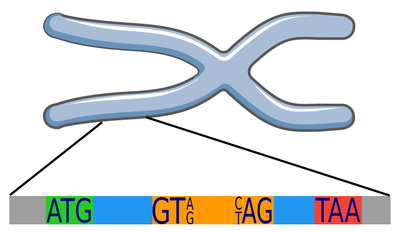
Introduction

The rise of long-read sequencing has unveiled the true complexity of transcriptomes, exposing the limitations of current genome annotation methods in capturing isoform diversity and alternative splicing.

To address this challenge, we developed **SQANTI-evidence**, a novel pipeline that integrates long-read transcriptomes curated with the SQANTI suite into the **AUGUSTUS** framework to guide structural gene prediction.

By combining curated long-read evidence with high-confidence training sets, the method delivers accurate and biologically meaningful annotations, as demonstrated in *Paniagua et al.* (2025).

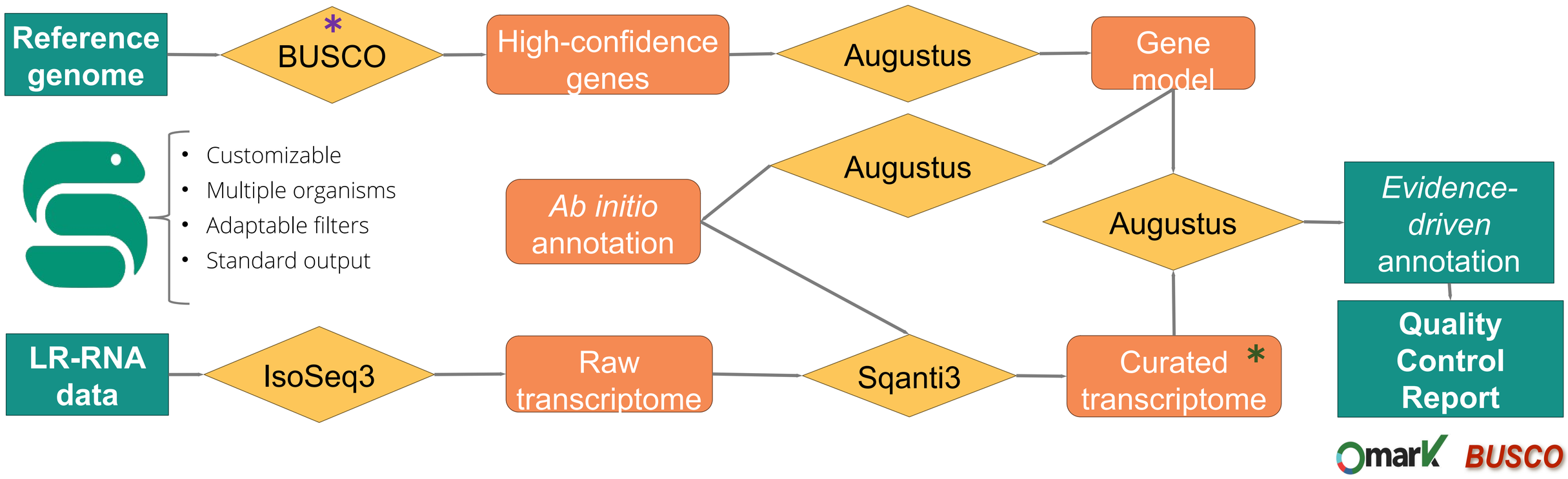
Ab-initio



Evidence-based

Evidence-driven

Methods



- Customizable
- Multiple organisms
- Adaptable filters
- Standard output

Results & Conclusion

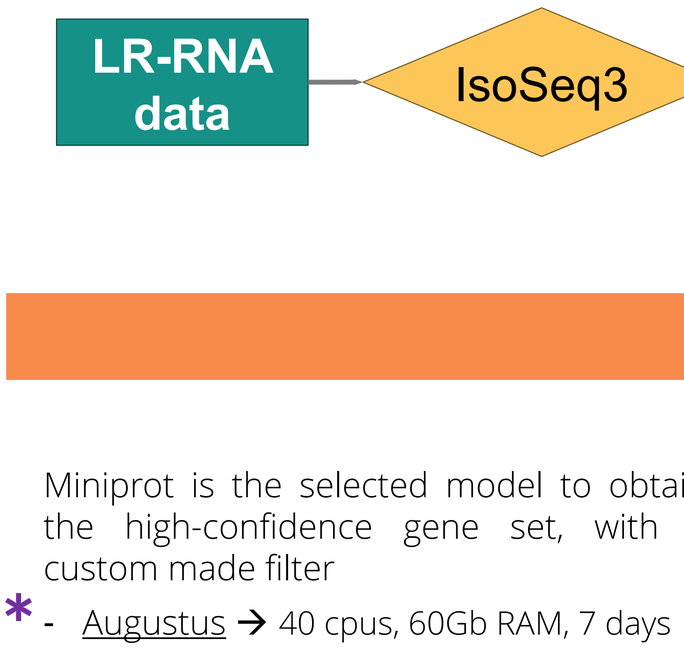
Miniprot is the selected model to obtain the high-confidence gene set, with a custom made filter

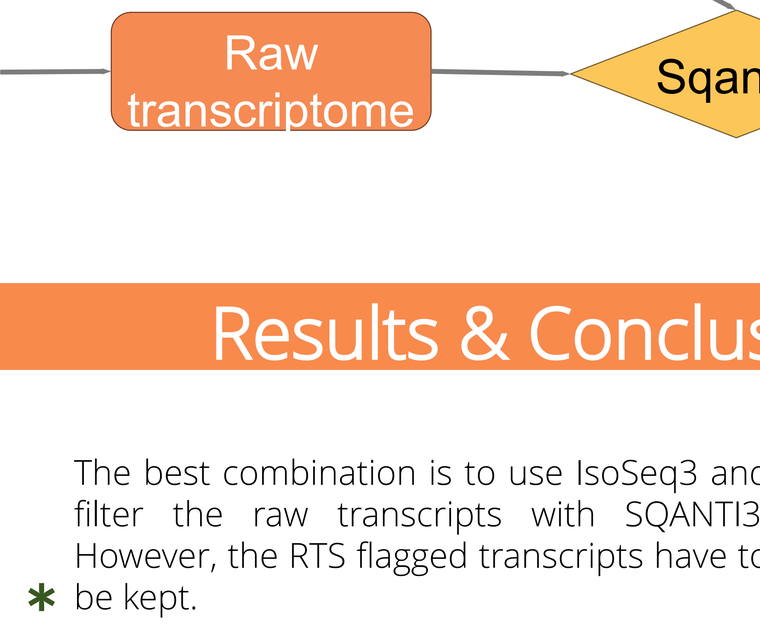
* - Augustus → 40 cpus, 60Gb RAM, 7 days

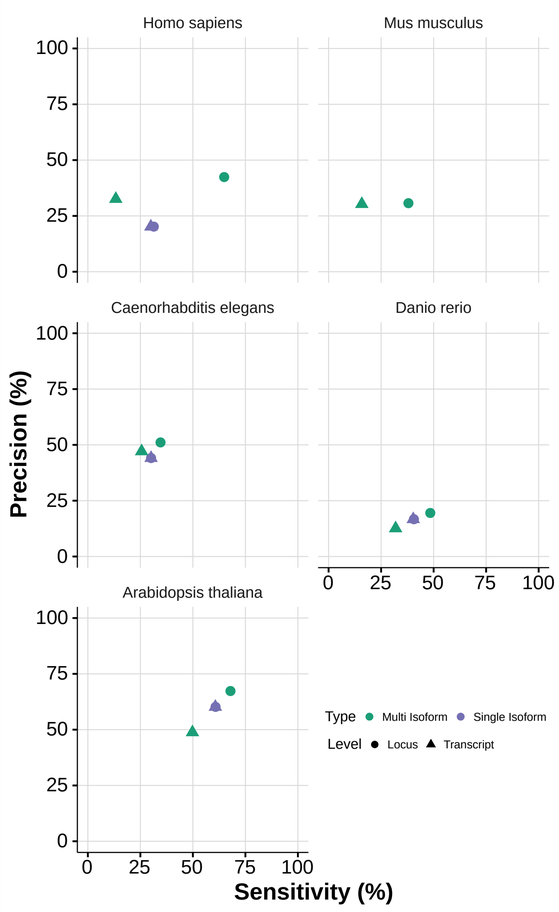
- Miniprot → 40 cpus, 20Gb RAM, 40 mins

The best combination is to use IsoSeq3 and filter the raw transcripts with SQANTI3. However, the RTS flagged transcripts have to be kept.

* IsoQuant does not produce better results. The combination of both is a slight upgrade







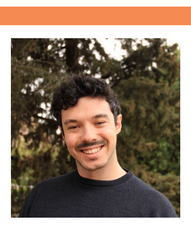
The main challenge is to **compare reliably** between newly assembled transcripts and reference annotation.

With the recent explosion in the number of isoforms per gene, achieving an accurate representation using a single predictive model or data from only one tissue has become virtually impossible.

- **SQANTI-evidence** provides a rapid, efficient, end-to-end solution for genome annotation from long-read RNA-Seq data.
- It can generate **high-quality annotations** for any organism, even in the absence of prior reference annotations.

Extensive benchmark with more organisms and tools will follow after the final datasets have been decided upon

i2sysbio INSTITUTE FOR INTEGRATIVE SYSTEMS BIOLOGY



pablo.atienza@csic.es

eosc

OSCARS Open Science Clusters' Action for Research & Society

CSIC Consejo Superior de Investigaciones Científicas