LongTREC

Funded by
the European Union

Bioinformatics Summer School

Long-reads Transcriptomics

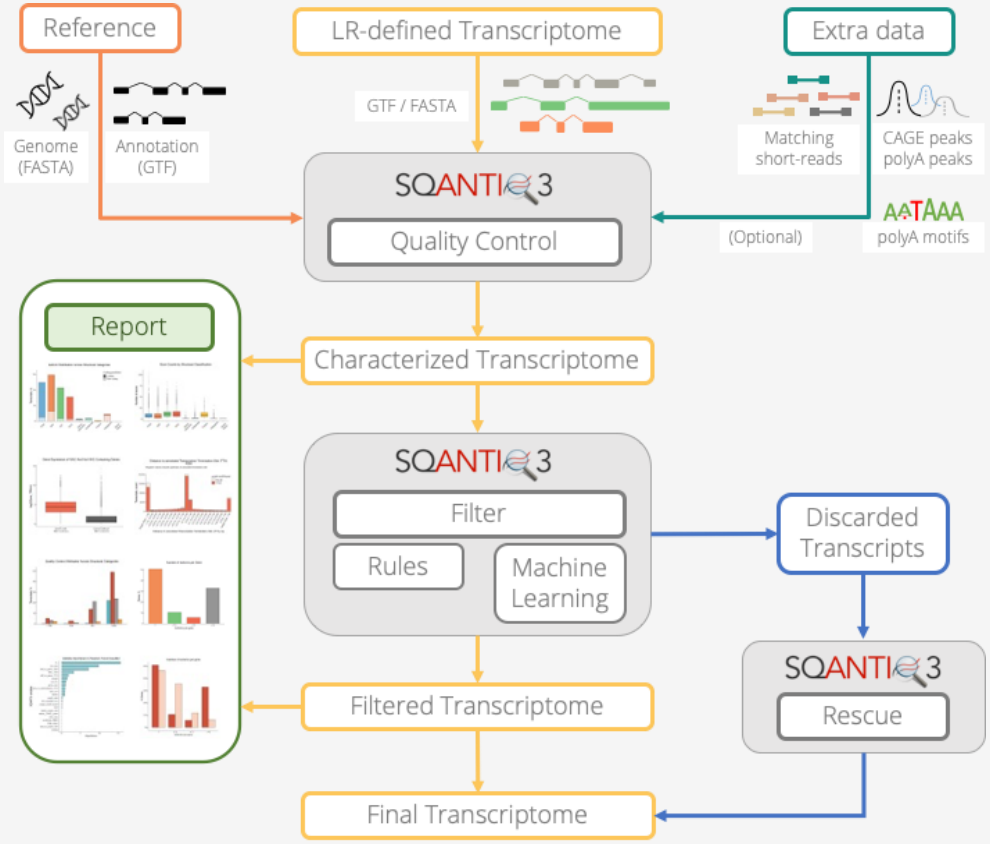
Fabian Jetzinger

BioBam Bioinformatics,
Valencia, Spain

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

Course Contents

1. SQANTI3 Quality Control
2. SQANTI3 Filter
3. SQANTI3 Rescue



Section 1

SQANTI3 Quality Control

Understanding the evaluation of long-read derived transcriptomes

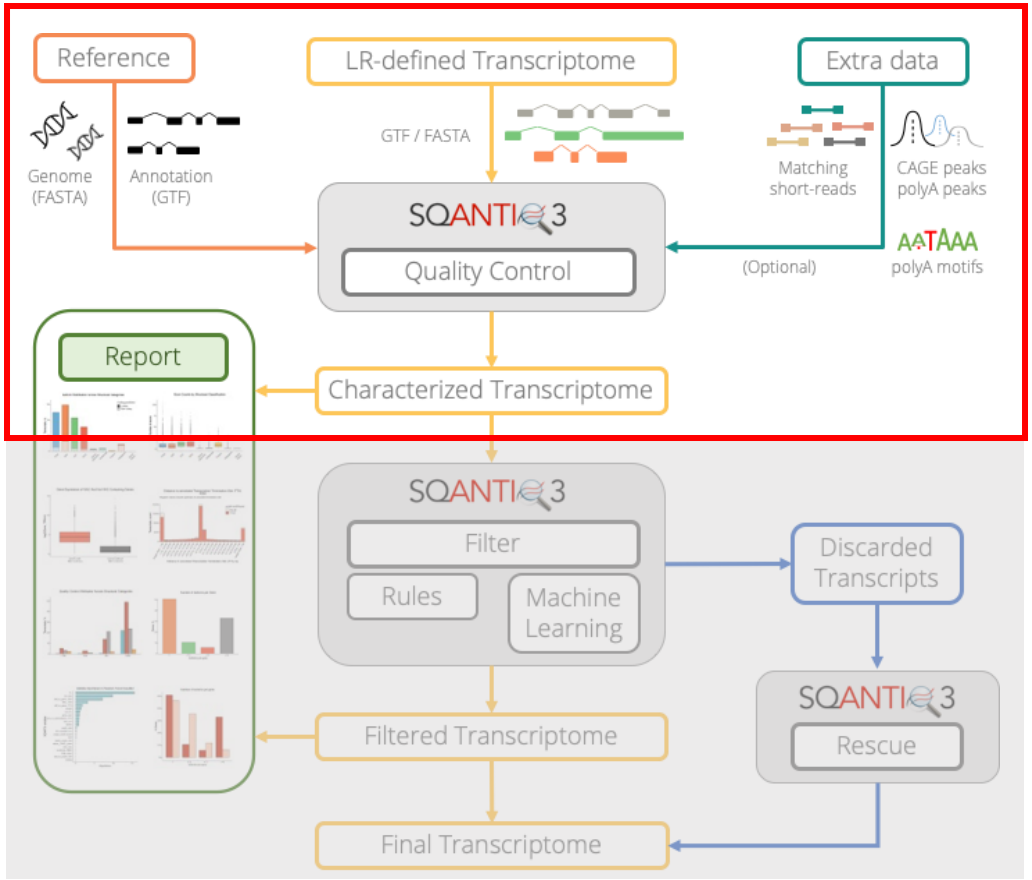
SQANTI3 Quality Control characterizes long-read transcriptomes

SQANTI3 Quality Control for
custom long-read transcriptomes serves to:

- Compare novel transcripts to reference annotations
- Characterize transcripts with orthogonal data

Types of orthogonal data:

- **Short-read RNAseq data**
  (alignments, junction coverage, expression)
- **Transcript Start Sites**
  (CAGE peaks; e.g. refTSS)
- **Transcript Termination Sites**
  (PolyA motifs, peaks; e.g. PolyASite)
- **Long-read expression data**



See:
Running SQANTI3 Quality Control · ConesaLab/SQANTI3 Wiki
Understanding the output of SQANTI3 QC · ConesaLab/SQANTI3 Wiki

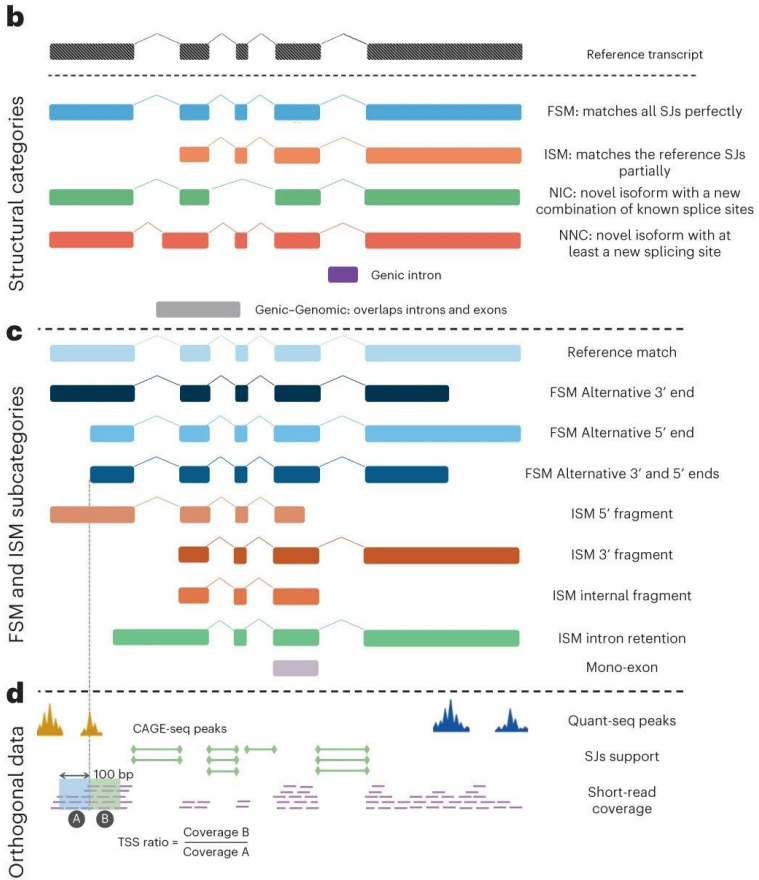
Structural Categories compare a transcript to a reference

SQANTI3 Quality Control for
custom long-read transcriptomes serves to:

- Compare novel transcripts to reference annotations
- Characterize transcripts with orthogonal data

Types of orthogonal data:

- **Short-read RNAseq data**
  (alignments, junction coverage, expression)
- **Transcript Start Sites**
  (CAGE peaks; e.g. refTSS)
- **Transcript Termination Sites**
  (PolyA motifs, peaks; e.g. PolyASite)
- **Long-read expression data**



See:
SQANTI3 isoform classification: categories and subcategories · ConesaLab/SQANTI3 Wiki
Pardo-Palacios, F.J., Arzalluz-Luque, A., Kondratova, L. et al. SQANTI3: curation of long-read transcriptomes for accurate identification of known and novel isoforms. Nat Methods 21, 793–797 (2024). https://doi.org/10.1038/s41592-024-02229-2

Section 2

SQANTI3 Filter

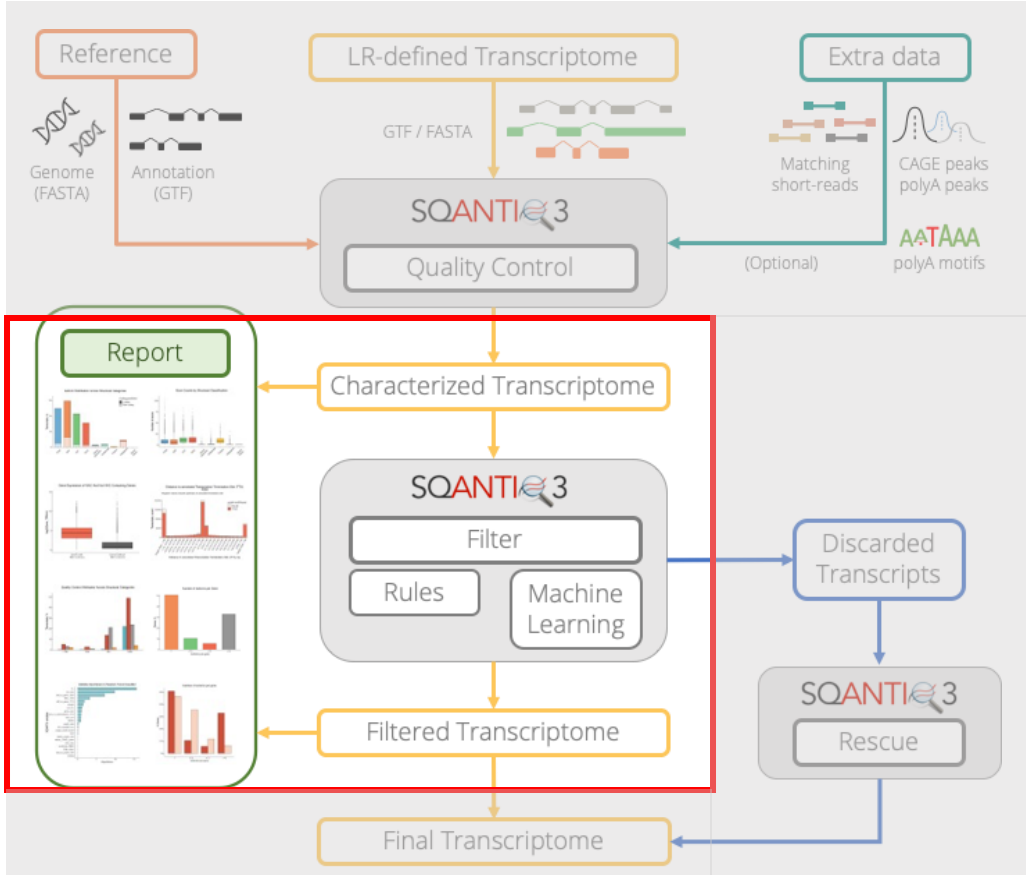
Understanding the *why* and *how* of transcriptome curation

SQANTI3 Filter curates a high-quality transcriptome

SQANTI3 Quality Control has characterized a transcriptome, but how can likely artifacts be identified and removed?

SQANTI3 Filter uses the information added by **QC** to remove artifacts.

- **Rules:** Define an explicit set of rules
- **Machine Learning:** Train a random forest model



See:
[Running SQANTI3 filter · ConesaLab/SQANTI3 Wiki]

SQANTI3 Filter: Rules are customizable and easy to explain, but rigid

Default rules:

No evidence of intra-priming

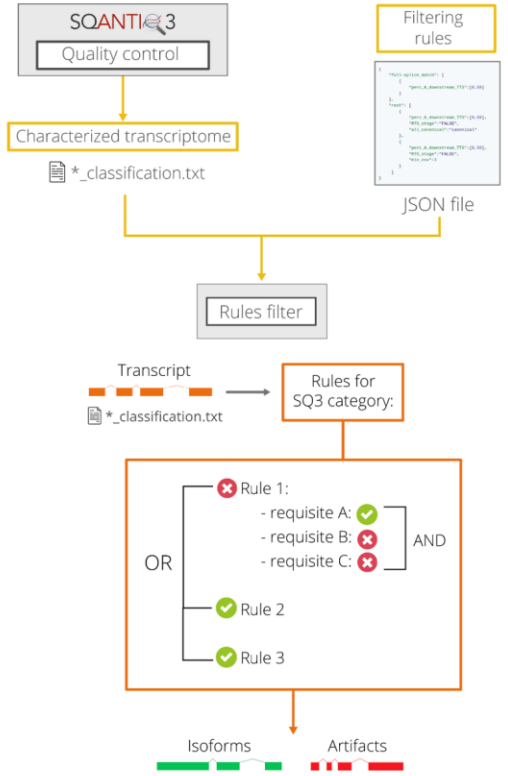
For non-FSM:

- no RT-switching
- only canonical junctions
- OR short-read coverage of junctions

```
{
    "full-splice_match": [
        {
            "perc_A_downstream_TTS":[0,59]
        }
    ],
    "rest": [
        {
            "perc_A_downstream_TTS":[0,59],
            "RTS_stage":"FALSE",
            "all_canonical":"canonical"
        },
        {
            "perc_A_downstream_TTS":[0,59],
            "RTS_stage":"FALSE",
            "min_cov":3
        }
     ]
}
```

Highly customizable

```
{
    "full-splice_match": [
        {
            "perc_A_downstream_TTS":[0,59]
        }
    ],
    "incomplete-splice_match":[
        {
            "length":[2001,14999],
            "subcategory": ["3prime_fragment", "5prime_fragment", "internal_fragment"]
        }
    ],
    "novel_in_catalog":[
        {
            "all_canonical": "canonical"
        },
        {
            "min_cov": 10
        }
    ],
    "novel_not_in_catalog":[
        {
            "all_canonical": "canonical",
            "diff_to_gene_TSS":[-50,50],
            "diff_to_gene_TTS": [-50,50]
        },
        {
            "min_cov": 10,
            "diff_to_gene_TSS":[-50,50],
            "diff_to_gene_TTS": [-50,50]
        }
    ],
    "rest": [
        {
            "RTS_stage":"FALSE",
            "all_canonical":"canonical",
            "coding": "coding",
            "perc_A_downstream_TTS":[0,59],
            "exons": 2
        }
     ]
}
```



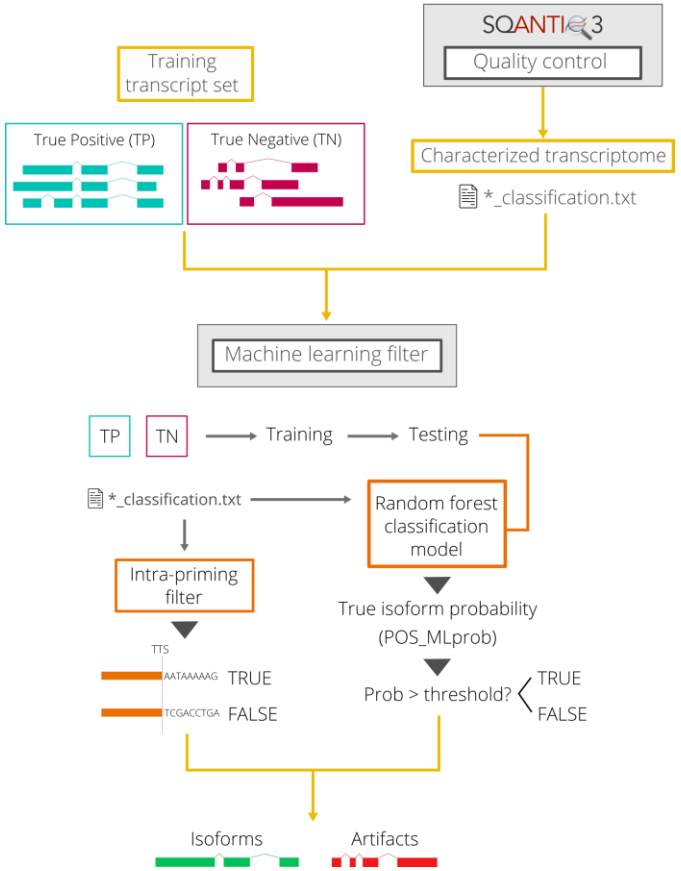
SQANTI3 Filter: Machine Learning is effective but more complex

Training:
Define True Positive / True Negative transcript sets.
Train random forest model to separate these two sets.

Filtering:
Exclude columns used to define training sets.
Filter out transcripts based on model decision.

Explainability:
Feature importance of random forest model.

Recommended to use Machine Learning Filter, BUT evaluate multiple configurations to learn how it affects their particular data set and transcriptome.



Section 3

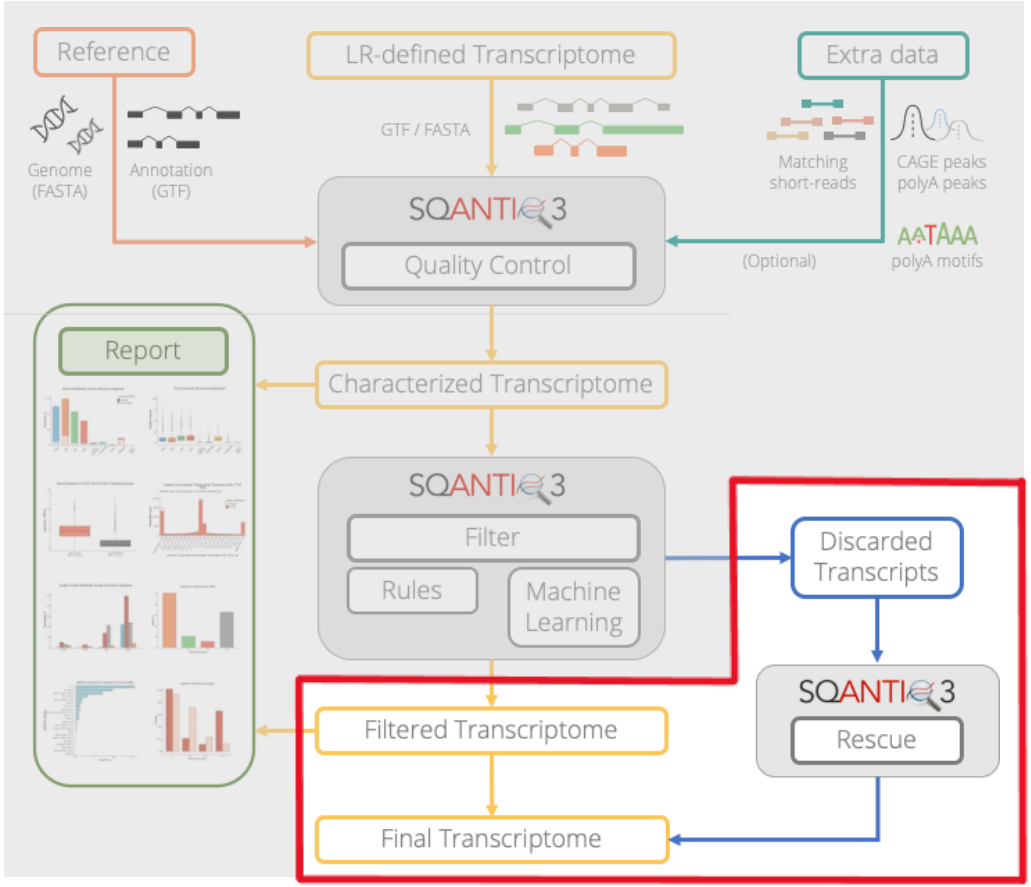
SQANTI3 Rescue

Understanding how to make the most of your data

SQANTI3 Rescue preserves transcriptome diversity lost by the Filter

SQANTI3 Filter has removed likely artifacts from the transcriptome, but this has also caused a loss of transcriptomic diversity.

While the removed artifacts may not have had enough supporting evidence, they can still be represented by FSM or reference transcripts.

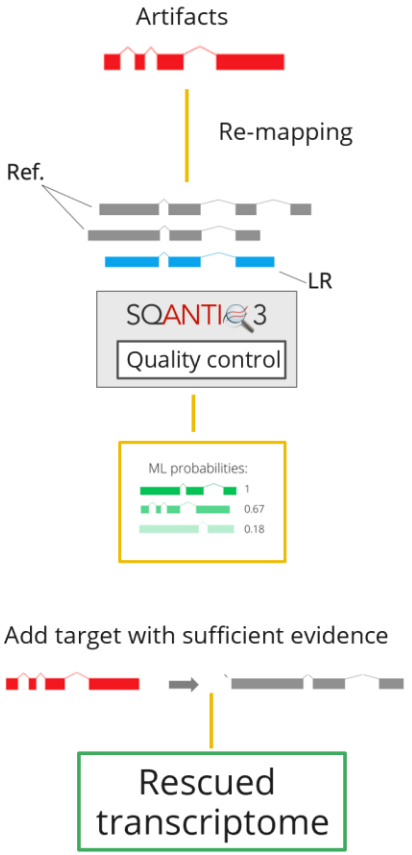


See:
[Running SQANTI3 rescue · ConesaLab/SQANTI3 Wiki]

SQANTI3 Rescue *rescues* transcriptome diversity lost in the Filter

Artifacts are re-mapped to **reference** transcripts and **FSMs** in order to

1. Preserve transcriptome diversity
2. Re-quantify reads



Reflection

- What is the goal of SQANTI3 Quality Control?
- Which types of orthogonal data can SQANTI3 QC use?
- What is the recommended SQANTI3 workflow of transcriptome curation?

Thank You!

For more information about the LongTREC Summer School:

https://longtrec.eu