Quality control and characterization of long-read transcriptoms

Francisco J. Pardo-Palacios

Lorena de la Fuente, PhD

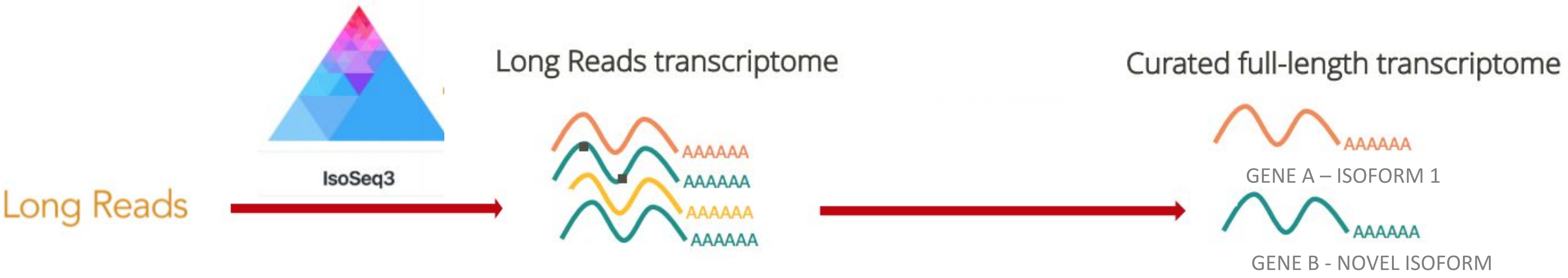
SQANTI 3

Genomics of Gene Expression Lab

Characterising LR transcriptomes

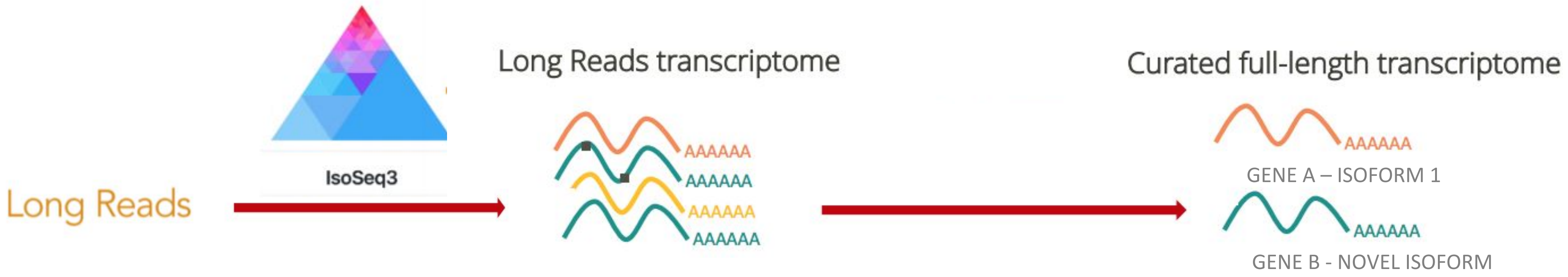
- After running IsoSeq3 pipeline, a *de novo* transcriptome is obtained
  - It will work as **your own reference transcriptome**

- Just like any reference transcriptome, it MUST be curated, compared and annotated.



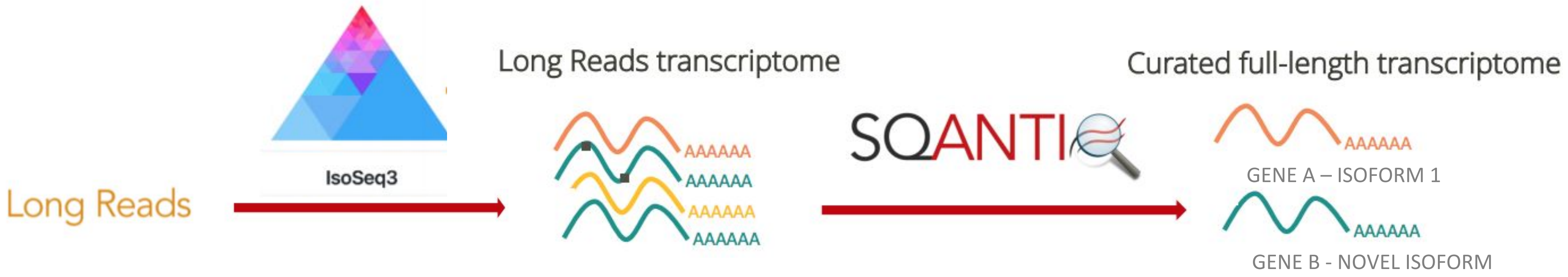
Questions to be solved before using a *de novo* LR transcriptome

- How similar are the isoforms compared to the reference transcriptome?
  - Have we found known…
    - Isoforms?
    - Transcription Starting or Terminating Sites?
    - Splice-junctions?
  - Have we found novel isoforms?
    - How do they look like?
- Are there any artifacts due to library preparation or sequencing issues?
- Can we use complementary data to support novel events in detected isoforms?

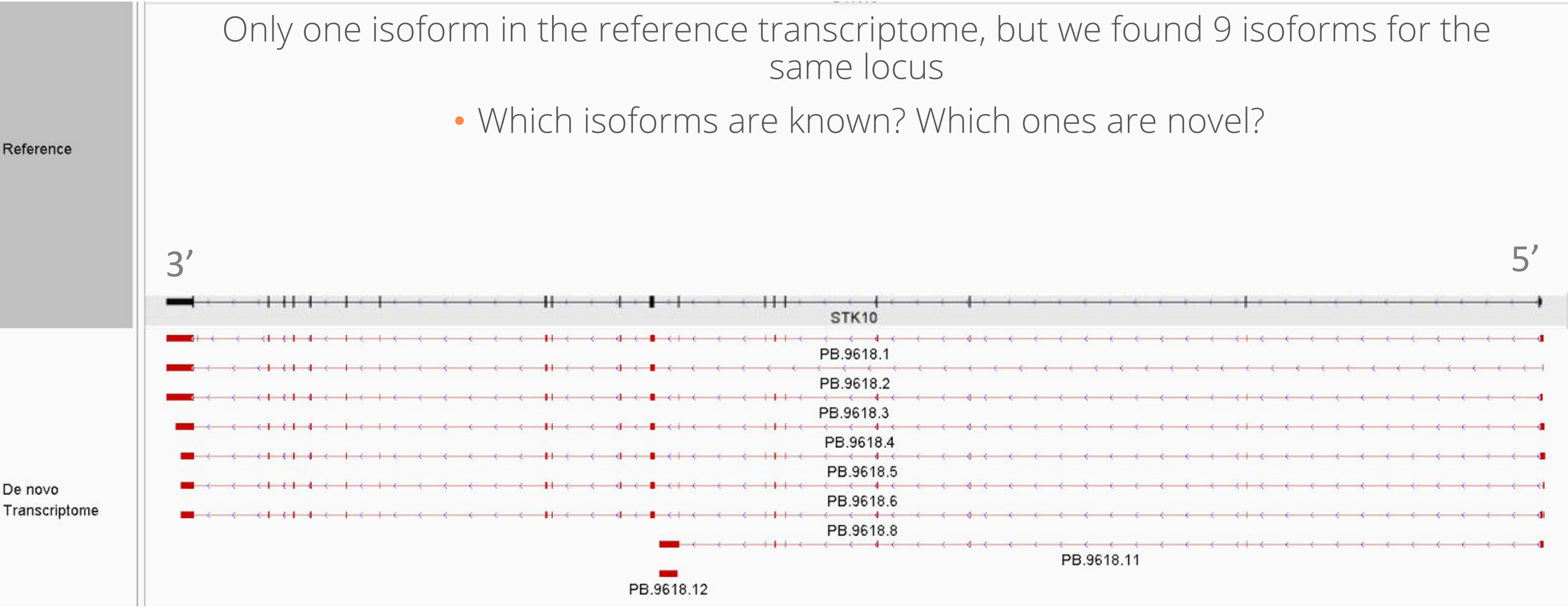


Questions to be solved before using a *de novo* LR transcriptome

- How similar are the isoforms compared to the reference transcriptome?
  - Have we found known…
    - Isoforms?
    - Transcription Starting or Terminating Sites?
    - Splice-junctions?
  - Have we found novel isoforms?
    - How do they look like?
- Are there any artifacts due to library preparation or sequencing issues?
- Can we use complementary data to support novel events in detected isoforms?



IGV example



Only one isoform in the reference transcriptome, but we found 9 isoforms for the same locus

- Which isoforms are known? Which ones are novel?

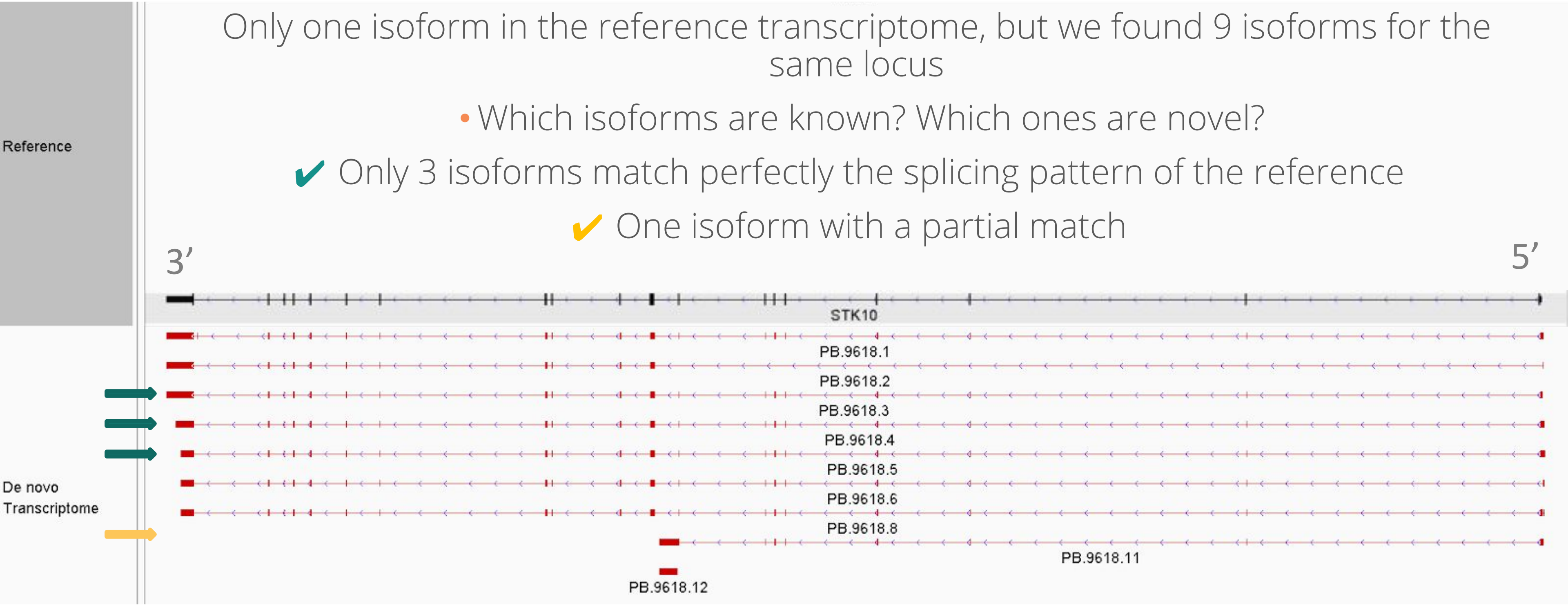
IGV example

Only one isoform in the reference transcriptome, but we found 9 isoforms for the same locus

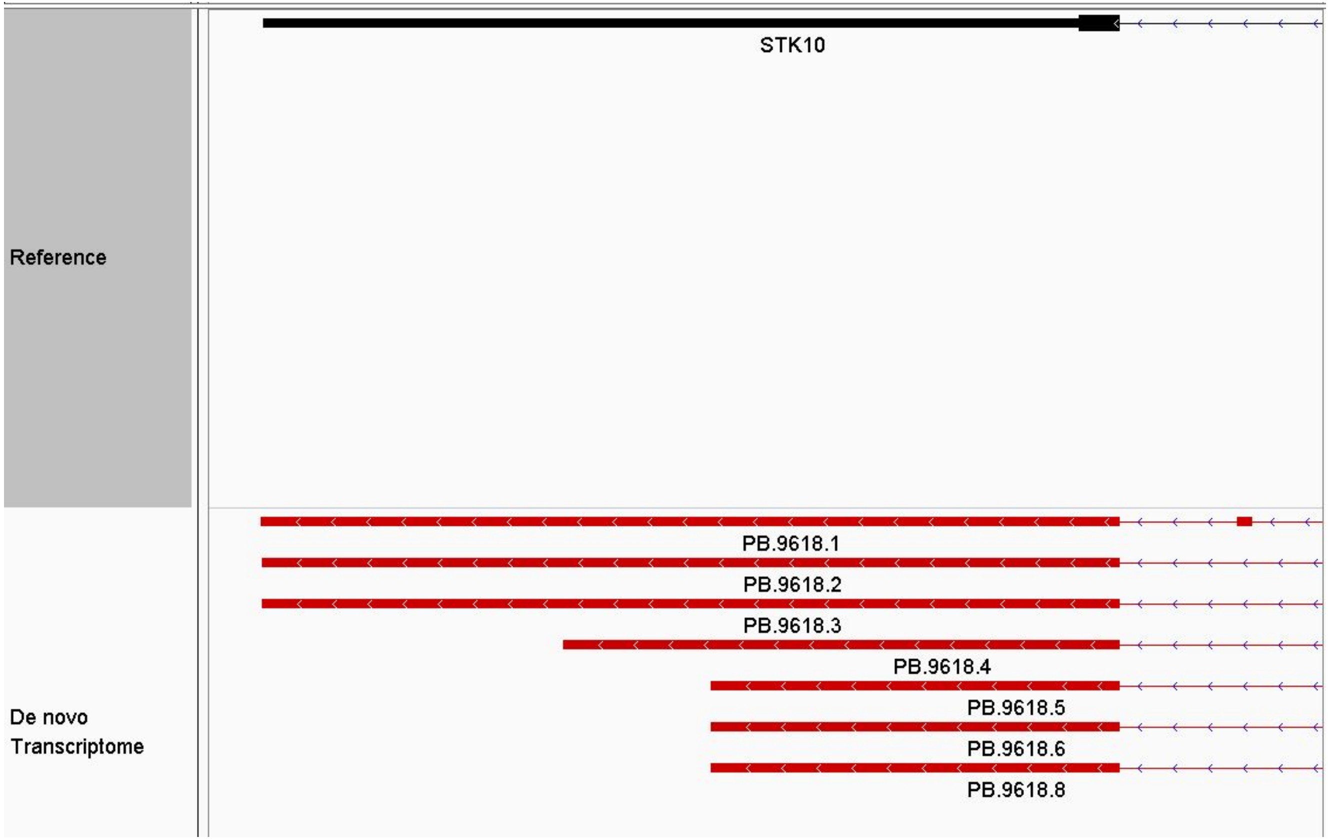
- Which isoforms are known? Which ones are novel?

✓ Only 3 isoforms match perfectly the splicing pattern of the reference

✓ One isoform with a partial match



IGV example

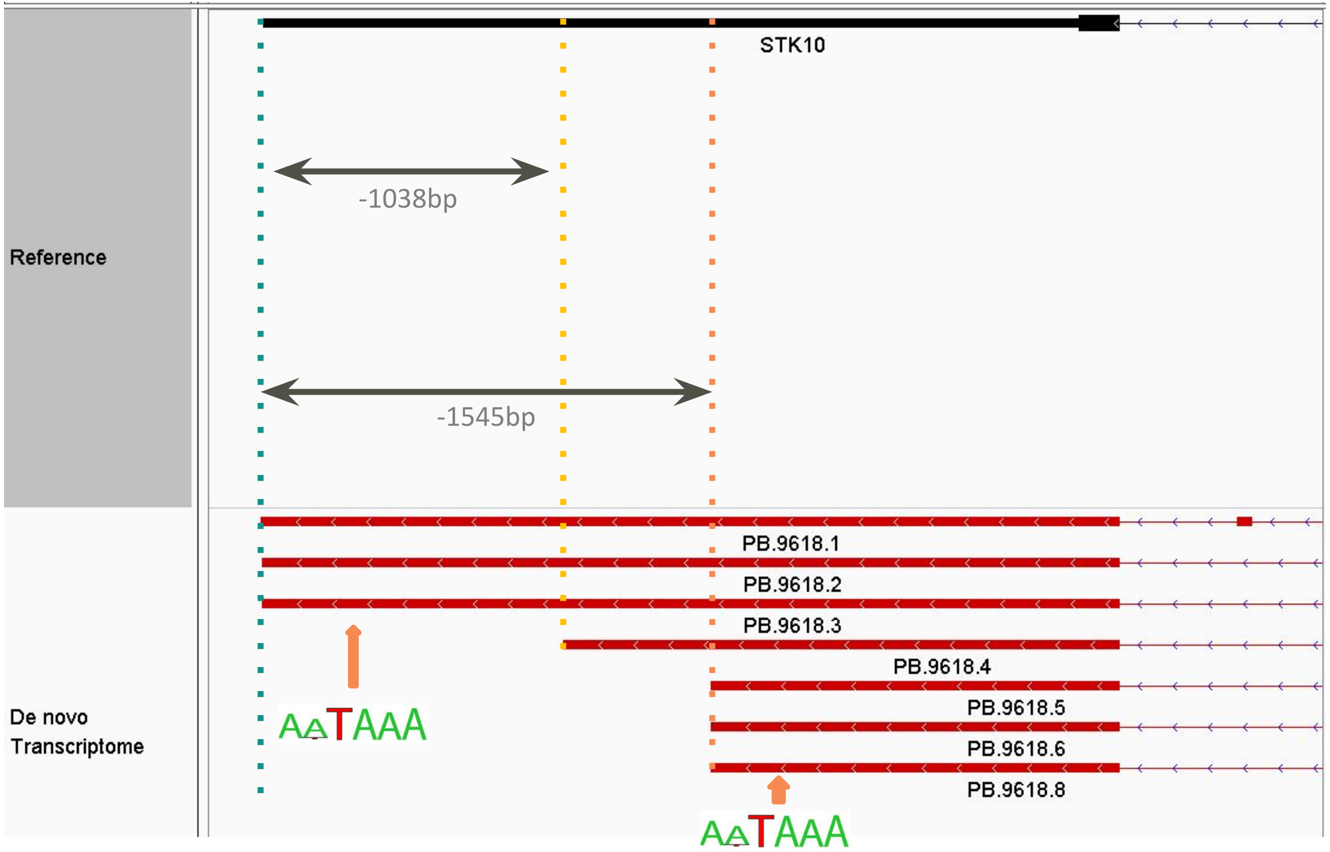


Different TTS sites

- How far a detected TTS falls from the reference TTS?
- Is there any polyA motif found close to the detected TTS?

"polyA motifs tend to be 19 bases upstream of the poly(A) site"

IGV example

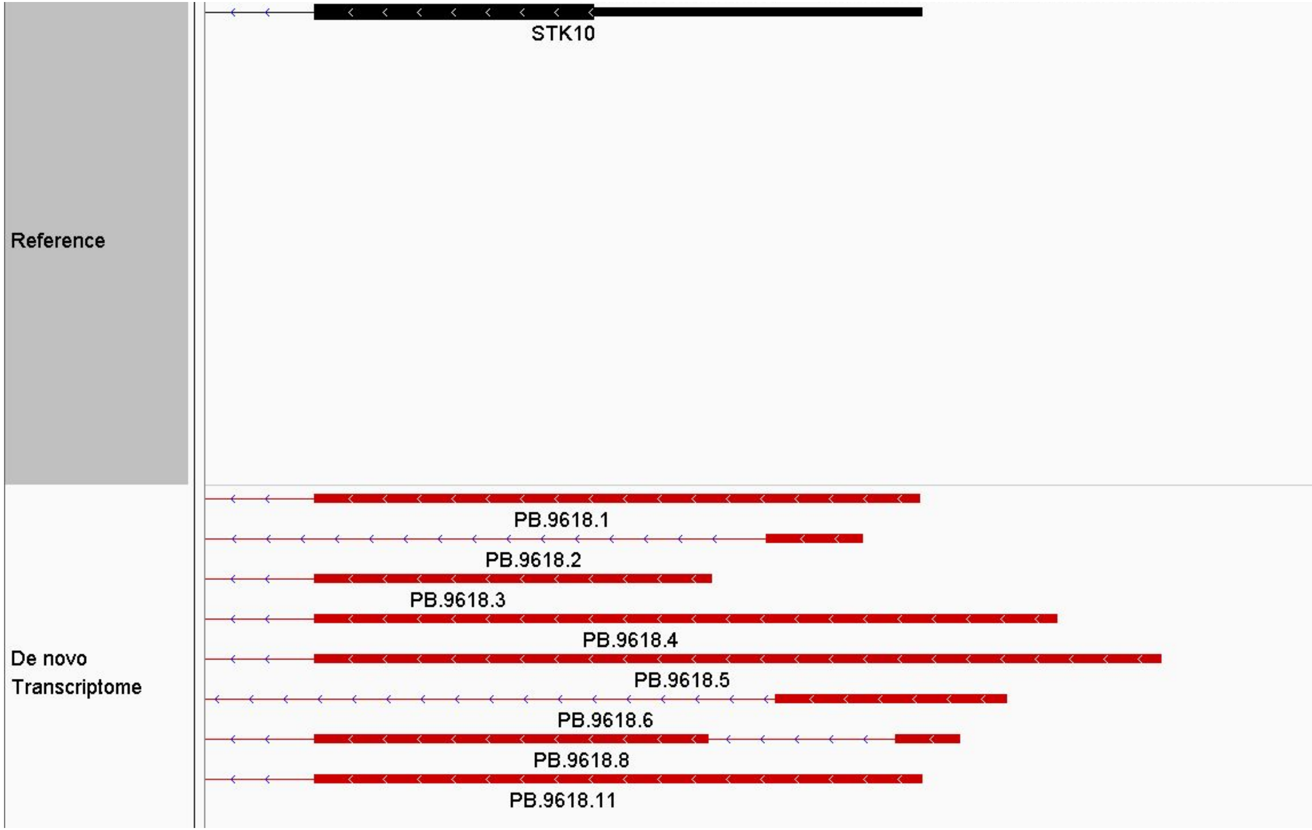


Different TTS sites

- How far a detected TTS falls from the reference TTS?
- Is there any polyA motif found close to the detected TTS?

"polyA motifs tend to be 19 bases upstream of the poly(A) site"

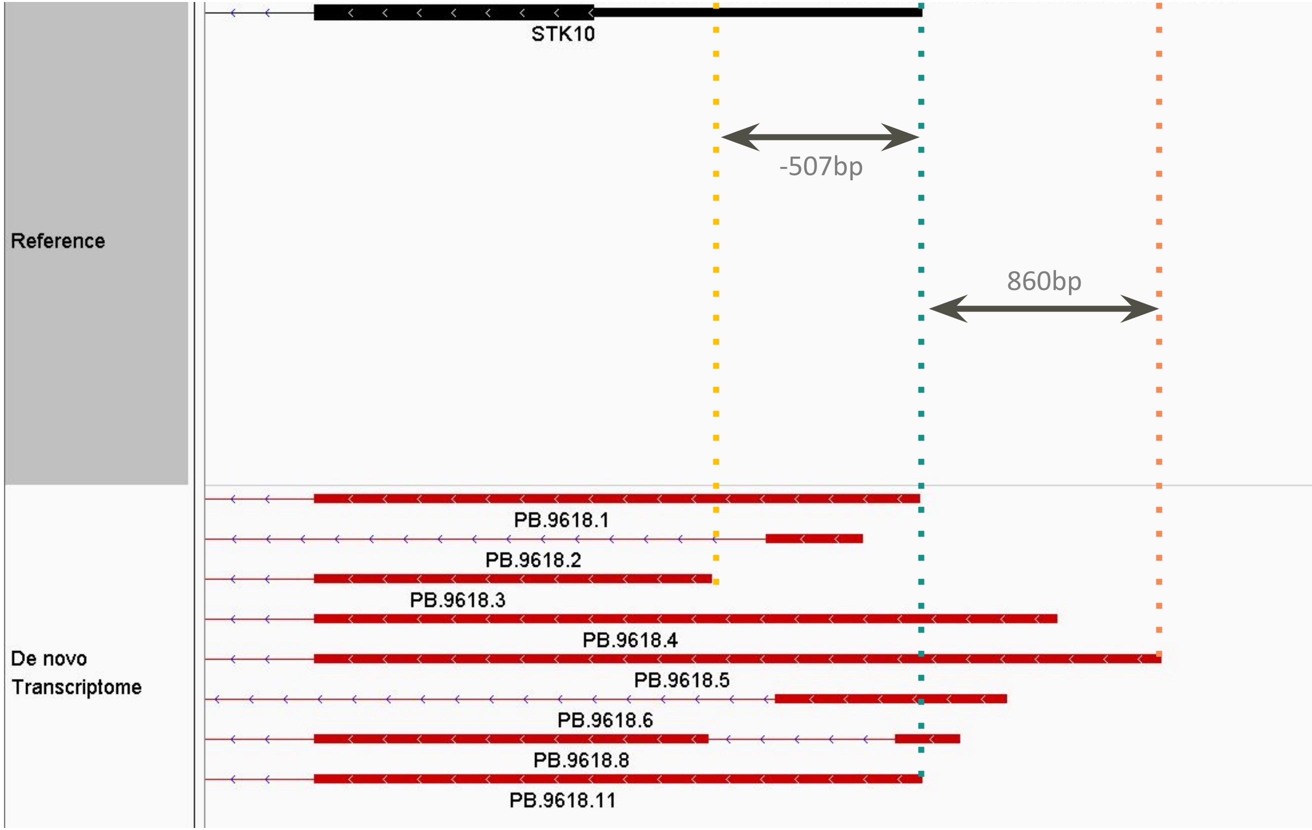
IGV example



Different TSS sites

- How far a detected TSS falls from the reference TSS?
- Can we use complementary data to distinguish true from false TSS?

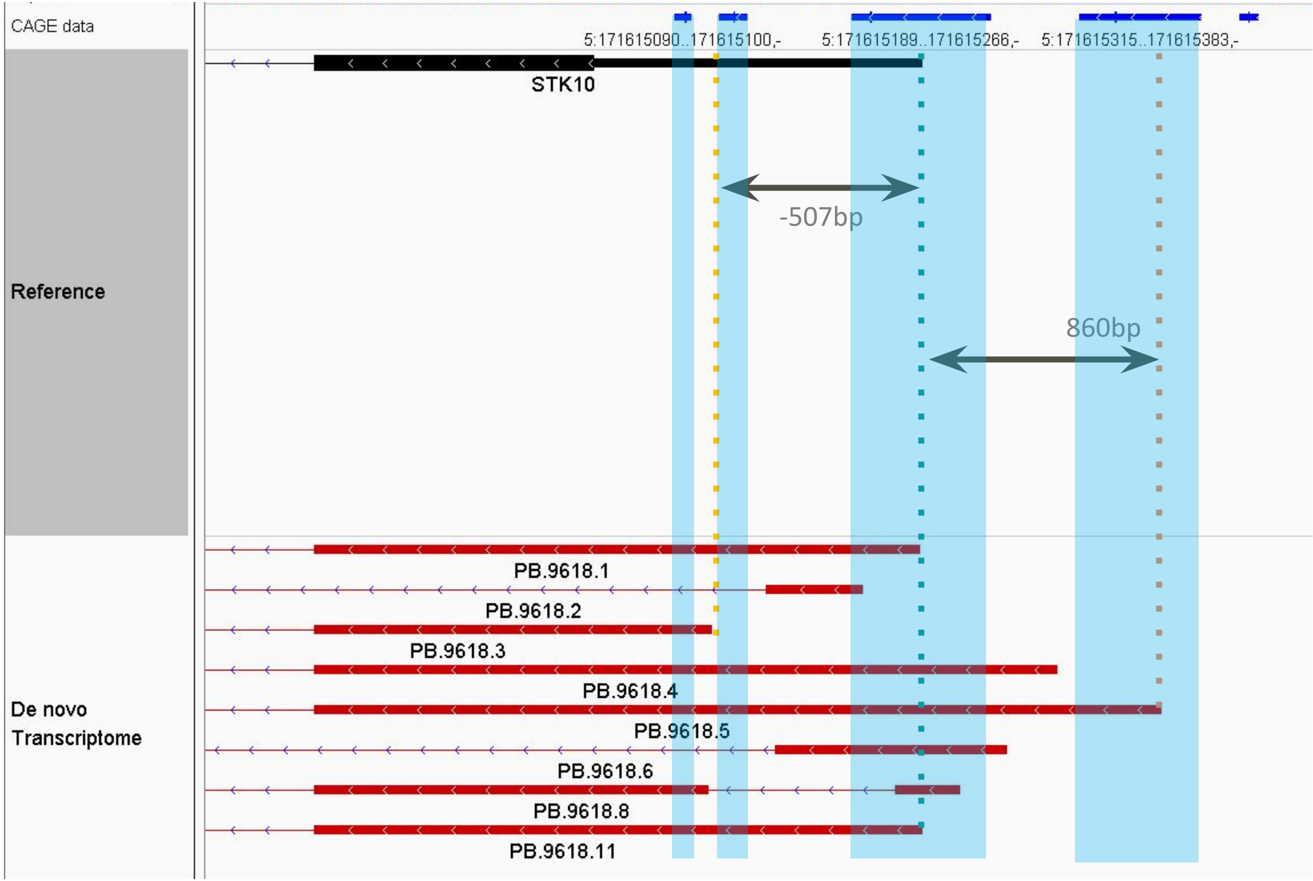
IGV example



Different TSS sites

- How far a detected TSS falls from the reference TSS?
- Can we use complementary data to distinguish true from false TSS?

IGV example

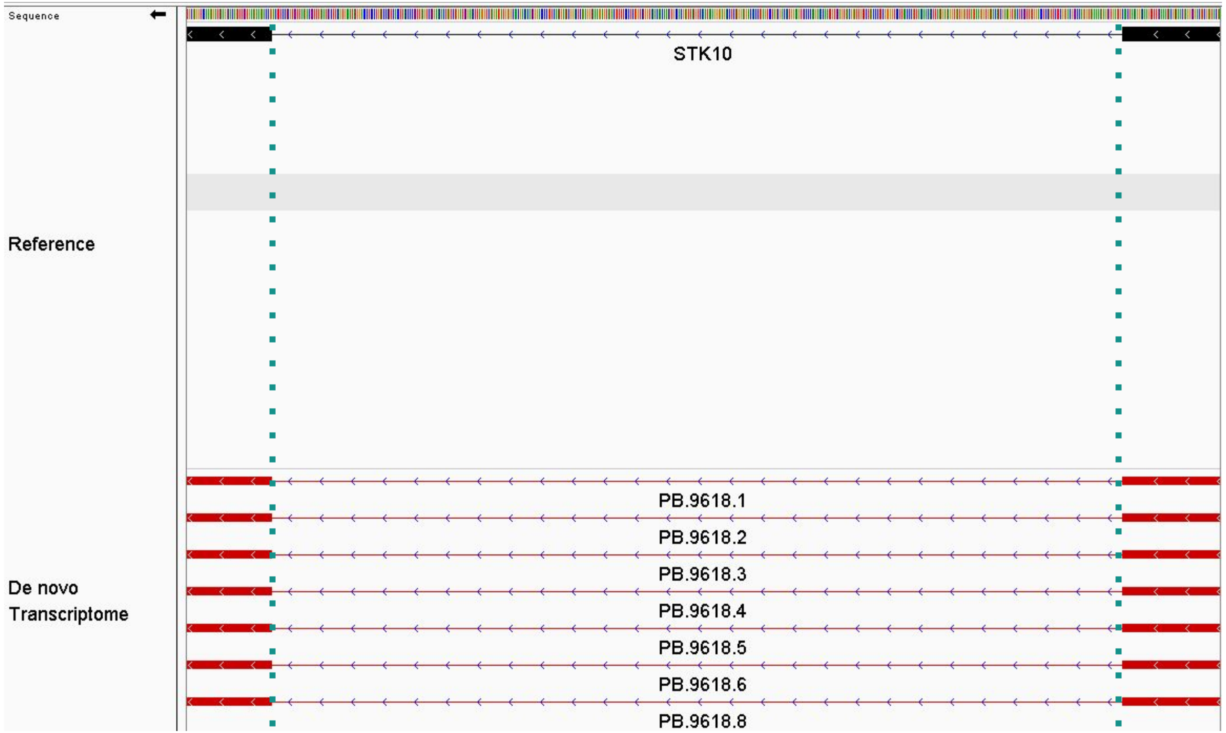


Different TSS sites

- How far a detected TSS falls from the reference TSS?
- Can we use complementary data to distinguish true from false TSS?

✔ CAGE peaks

IGV example



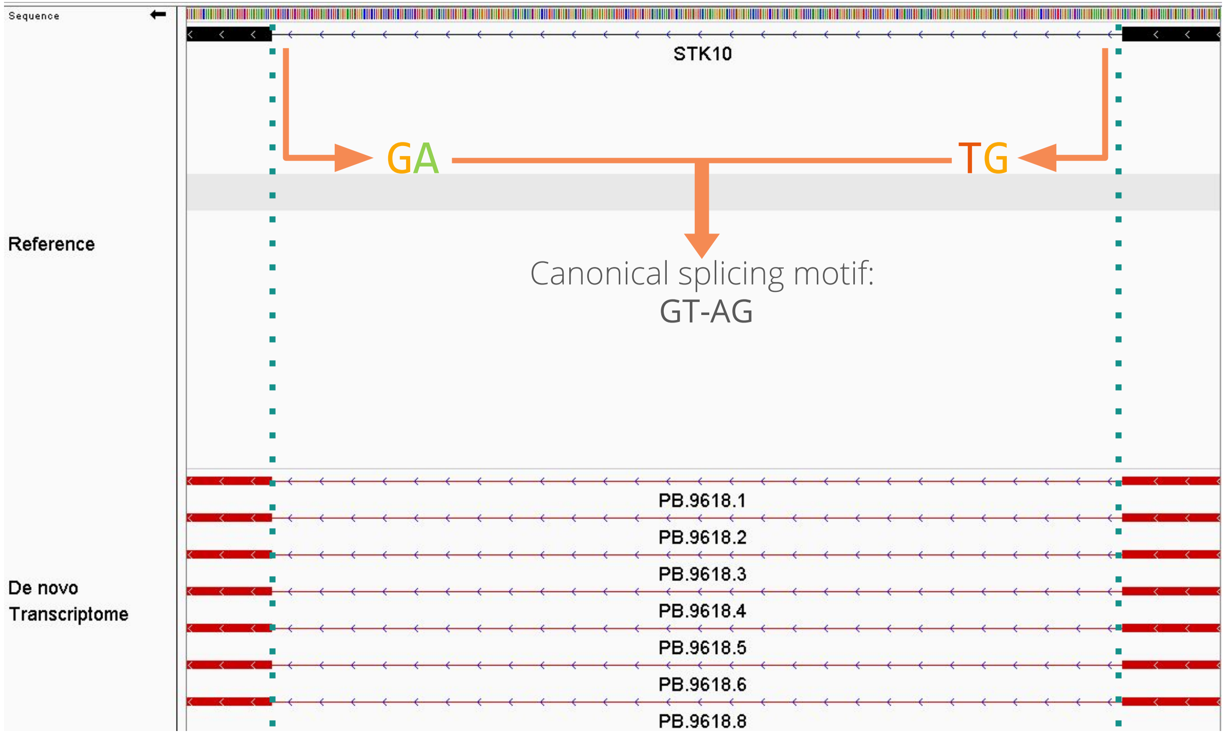
Splice Junctions

- Known or novel?
- Canonical or non-canonical motif?

Canonical motifs represent around 99% of mammalian splice junctions

- Can we use complementary data to distinguish between true and false Splice-Junctions?

IGV example



Canonical splicing motif:
GT-AG

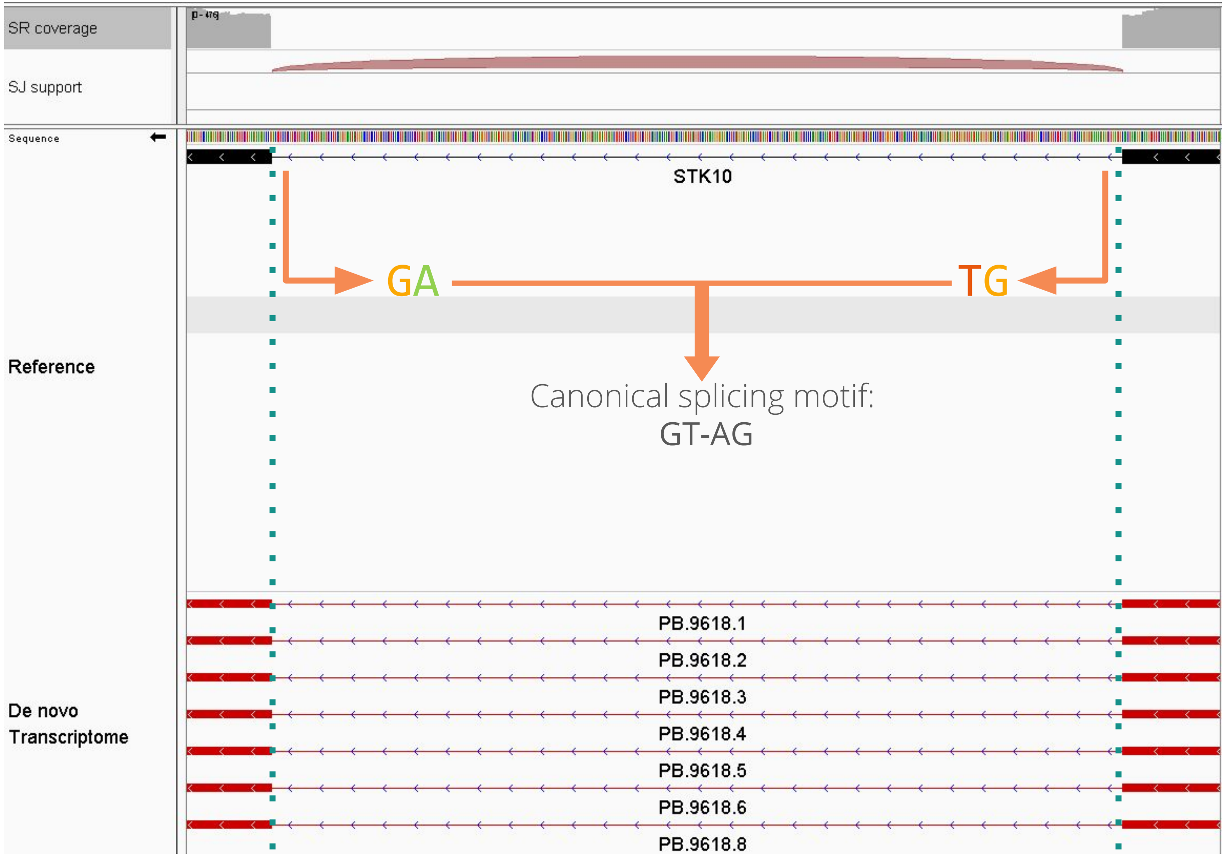
Splice Junctions

- Known or novel?
- Canonical or non-canonical motif?

Canonical motifs represent around 99% of mammalian splice junctions

- Can we use complementary data to distinguish between true and false Splice-Junctions?

IGV example



Canonical splicing motif:
GT-AG

Splice Junctions

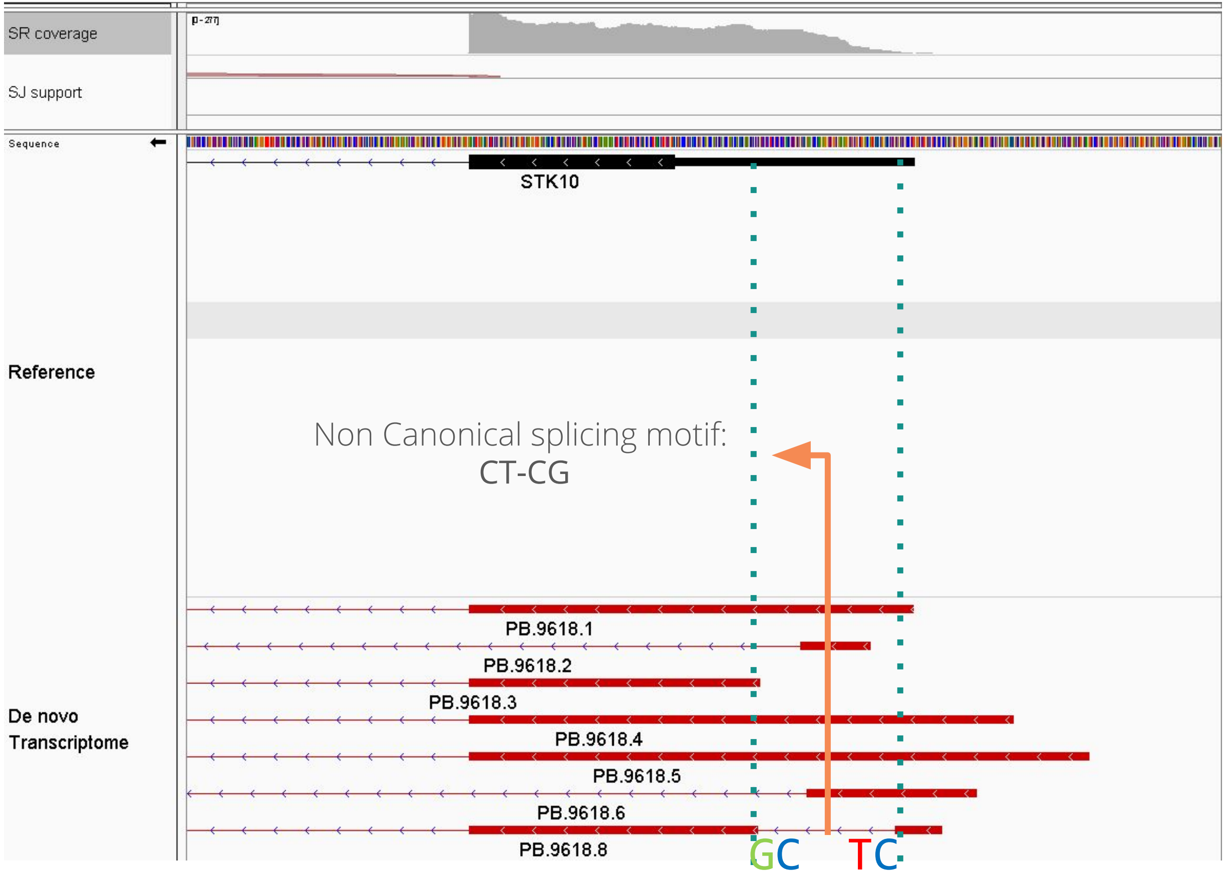
- Known or novel?
- Canonical or non-canonical motif?

Canonical motifs represent around 99% of mammalian splice junctions

- Can we use complementary data to distinguish between true and false Splice-Junctions?

✔ Matching RNA-Seq data

IGV example



Non Canonical splicing motif:
CT-CG

Splice Junctions

- Known or novel?
- Canonical or non-canonical motif?

Canonical motifs represent around 99% of mammalian splice junctions

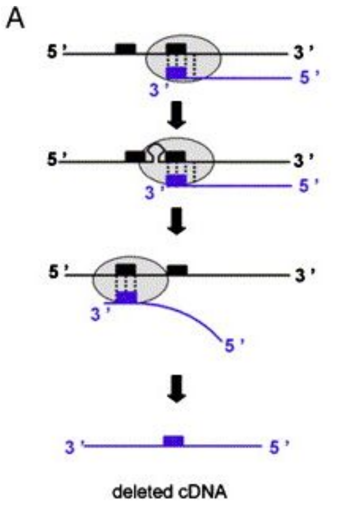
- Can we use complementary data to distinguish between true and false Splice-Junctions?

✗ Not matching RNA-Seq data

IGV example

Possible library preparation artifacts

- RT-Switching
- Intrapriming

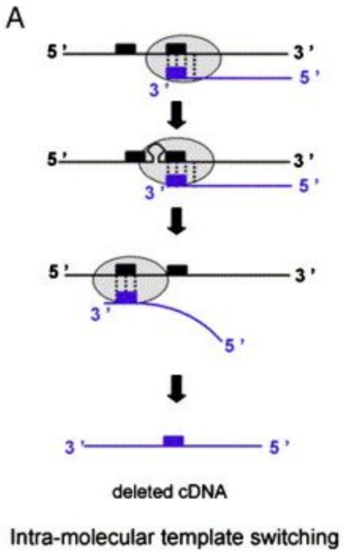


Intra-molecular template switching

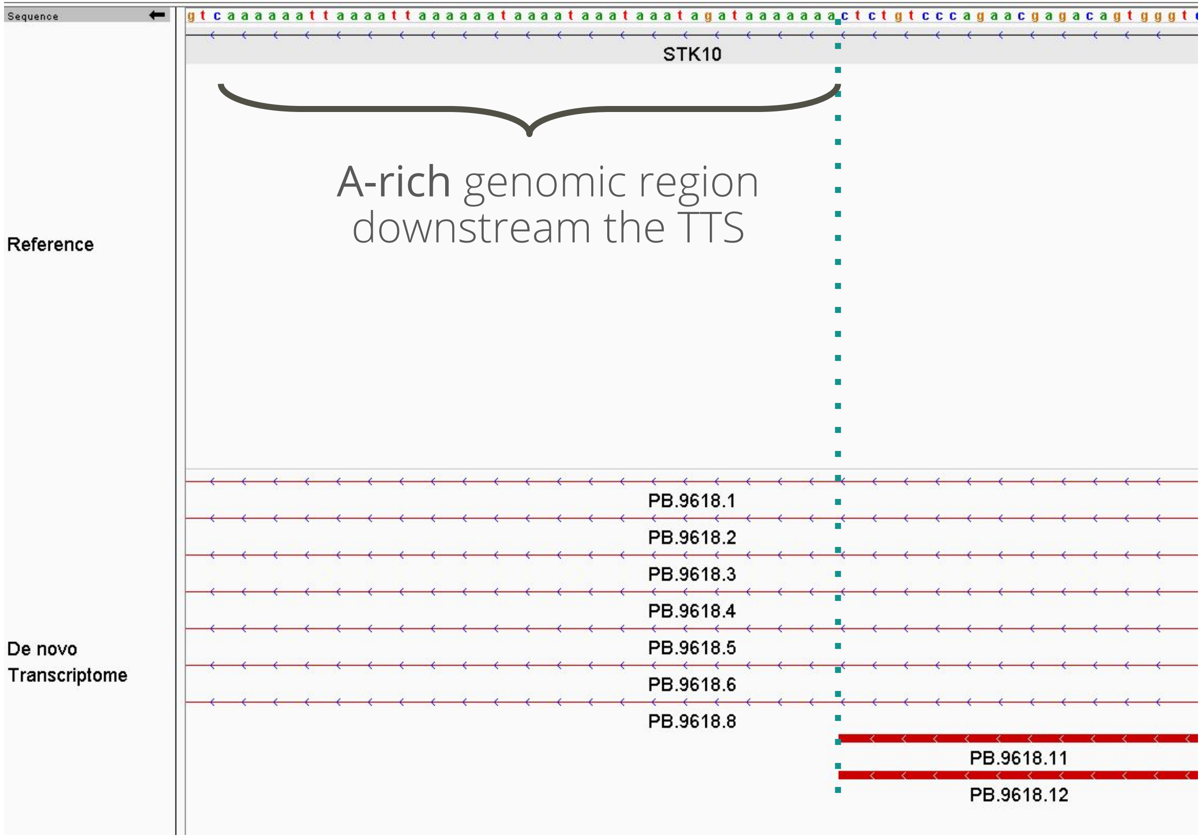
IGV example

Possible library preparation artifacts

- RT-Switching
- Intrapriming



Intra-molecular template switching

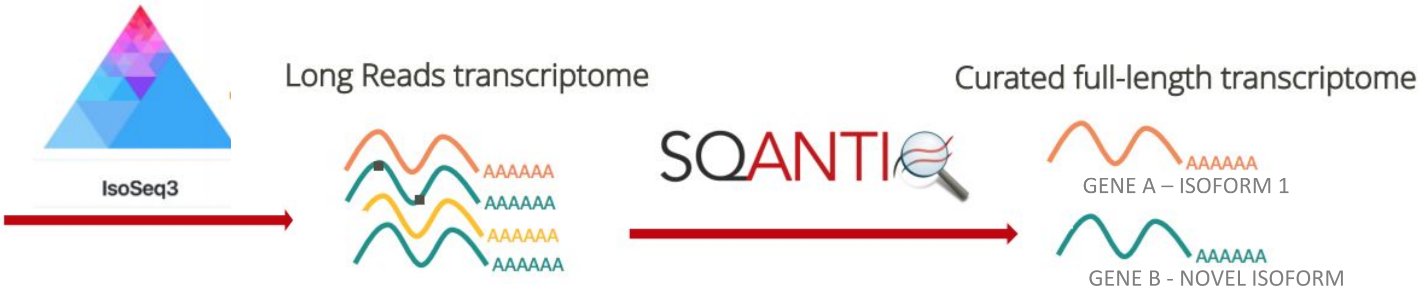

Sequence
g t c a a a a a a t t a a a a t t a a a a a a t a a a a t a a a t a a a t a g a t a a a a a a a c t c t g t c c c a g a a c g a g a c a g t g g g t
STK10
Reference
De novo Transcriptome
PB.9618.1
PB.9618.2
PB.9618.3
PB.9618.4
PB.9618.5
PB.9618.6
PB.9618.8
PB.9618.11
PB.9618.12

A-rich genomic region downstream the TTS

SQANTI aims

We need of a tool that...

- Stablishes a classification system for novel isoforms regarding to the reference transcriptome.
- Describes and addresses quality issues associated to Long-Read Sequencing.
- Gathers supplementary evidence around detected isoforms.
- Helps to filter out all those isoforms suspicious of being an artifact.



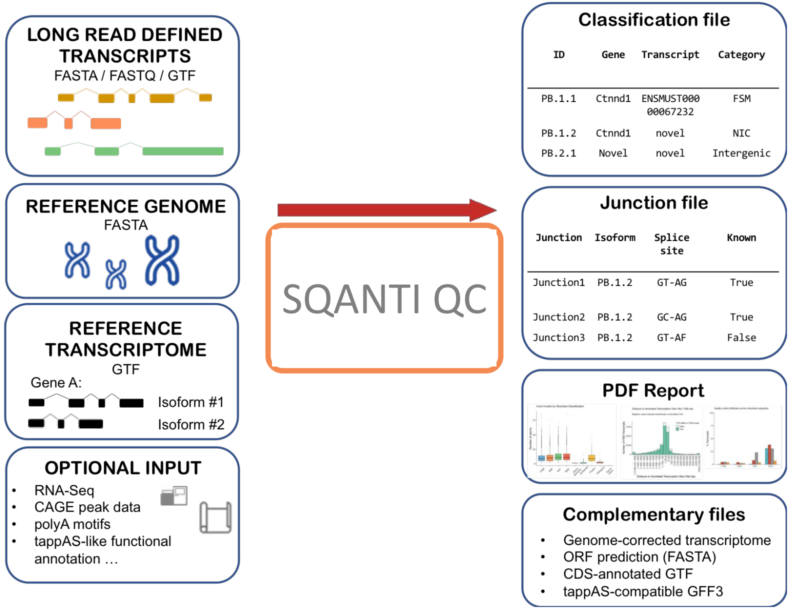
SQANTI3 workflow

USER INPUT

1 TRANSCRIPTOME QC & ANNOTATION

SQANTI3 will map the new transcript models to the reference genome and transcriptome to assess their degree of quality and novelty,

ALL QC ATTRIBUTES BREAKDOWN IN DEFINED CATEGORIES



LONG READ DEFINED TRANSCRIPTS
FASTA / FASTQ / GTF

REFERENCE GENOME
FASTA

REFERENCE TRANSCRIPTOME
GTF
Gene A:
Isoform #1
Isoform #2

OPTIONAL INPUT

- RNA-Seq
- CAGE peak data
- polyA motifs
- tappAS-like functional annotation …

SQANTI QC

Classification file

ID	Gene	Transcript	Category
PB.1.1	Ctnnd1	ENSMUST000 00067232	FSM
PB.1.2	Ctnnd1	novel	NIC
PB.2.1	Novel	novel	Intergenic

Junction file

Junction	Isoform	Splice site	Known
Junction1	PB.1.2	GT-AG	True
Junction2	PB.1.2	GC-AG	True
Junction3	PB.1.2	GT-AF	False

PDF Report

Complementary files

- Genome-corrected transcriptome
- ORF prediction (FASTA)
- CDS-annotated GTF
- tappAS-compatible GFF3

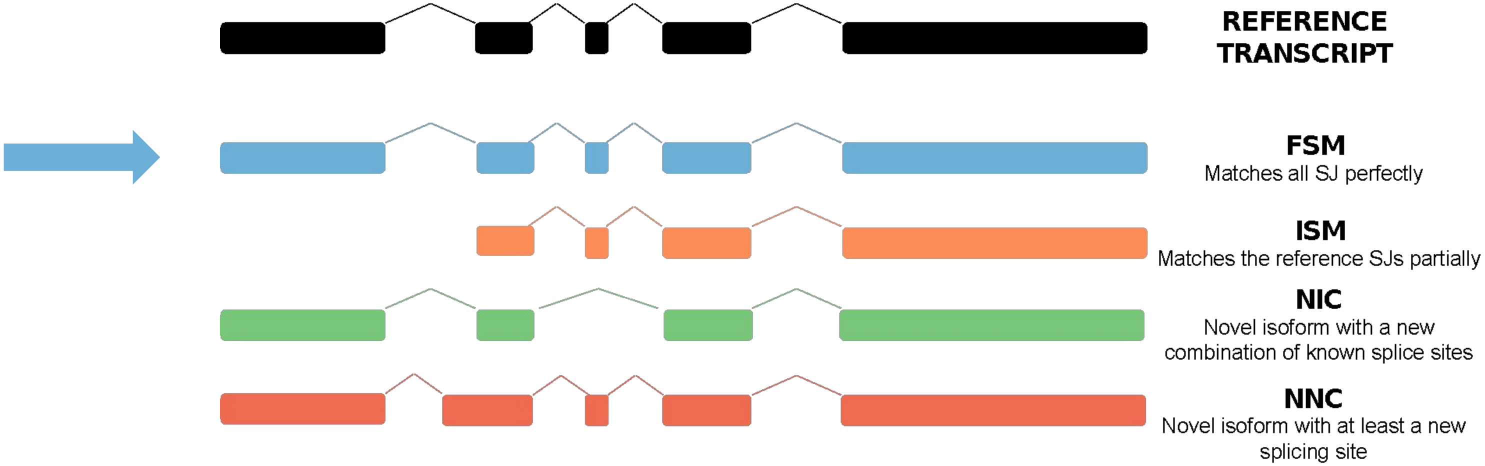
1. **Classification file:** TSV file describing each isoform analyzed by SQANTI3.

2. **Junctions file:** TSV file describing each splice junction found in each isoform.

3. PDF report generated from classification and junctions files:
    - General plots to visualize how is the new transcriptome
    - Specific analyses at the structural category level

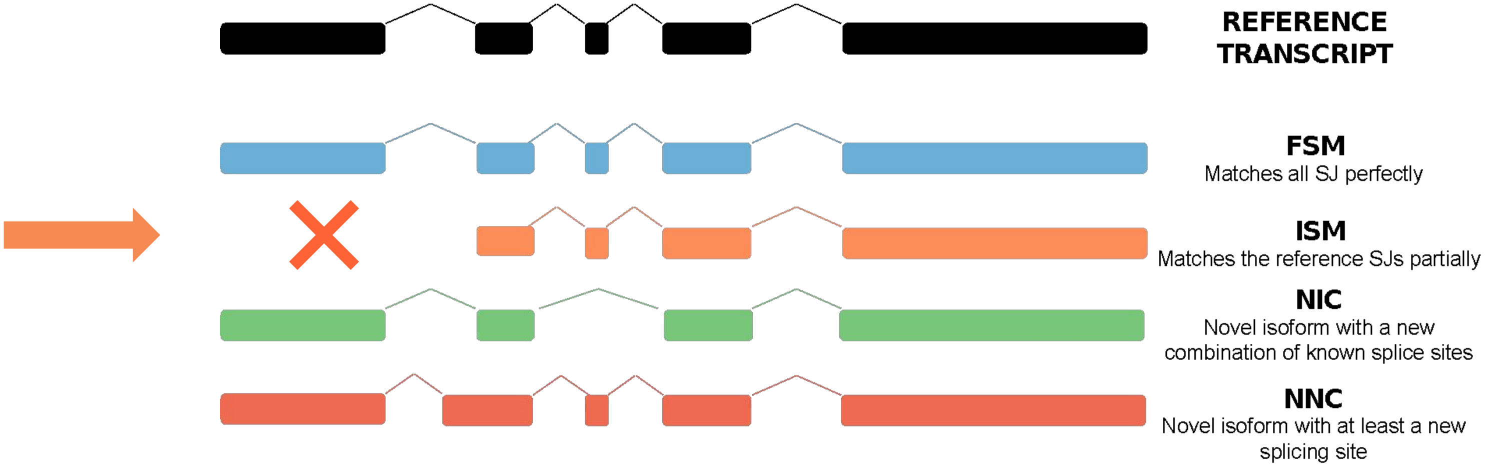
SQANTI3 structural categories

- Transcripts from **known** genes:
  - Full-Splice Match (FSM)
  - Incomplete-Splice Match (ISM)
  - Novel In Catalog (NIC)
  - Novel Not In Catalog (NNC)


REFERENCE TRANSCRIPT
FSM
Matches all SJ perfectly
ISM
Matches the reference SJs partially
NIC
Novel isoform with a new combination of known splice sites
NNC
Novel isoform with at least a new splicing site

SQANTI3 structural categories

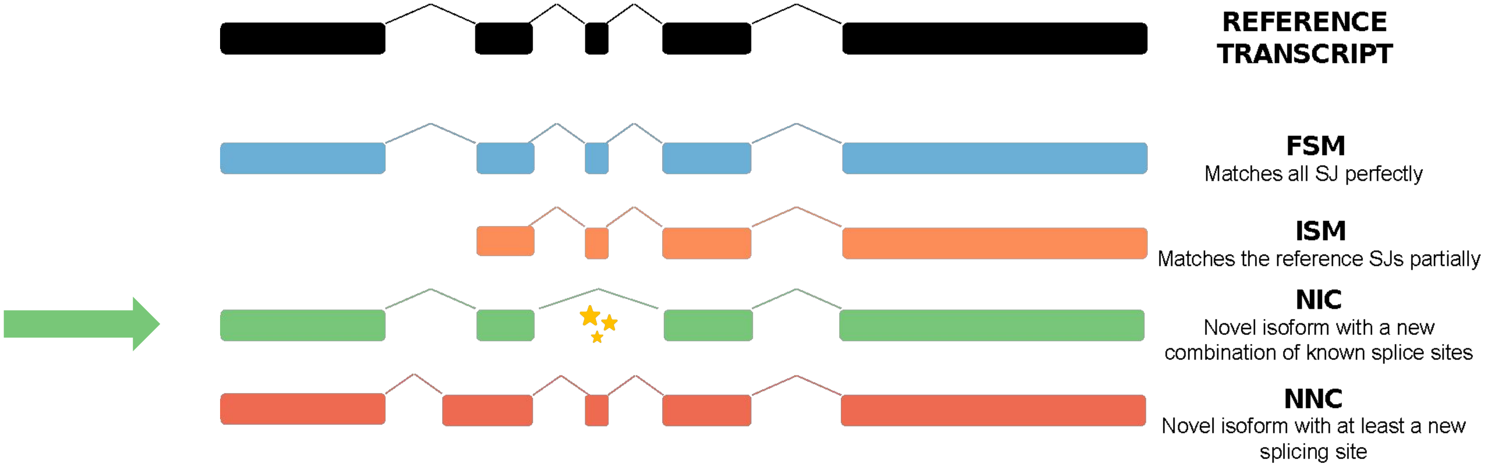
- Transcripts from **known** genes:
  - Full-Splice Match (FSM)
  - Incomplete-Splice Match (ISM)
  - Novel In Catalog (NIC)
  - Novel Not In Catalog (NNC)



REFERENCE TRANSCRIPT
FSM
Matches all SJ perfectly
ISM
Matches the reference SJs partially
NIC
Novel isoform with a new combination of known splice sites
NNC
Novel isoform with at least a new splicing site

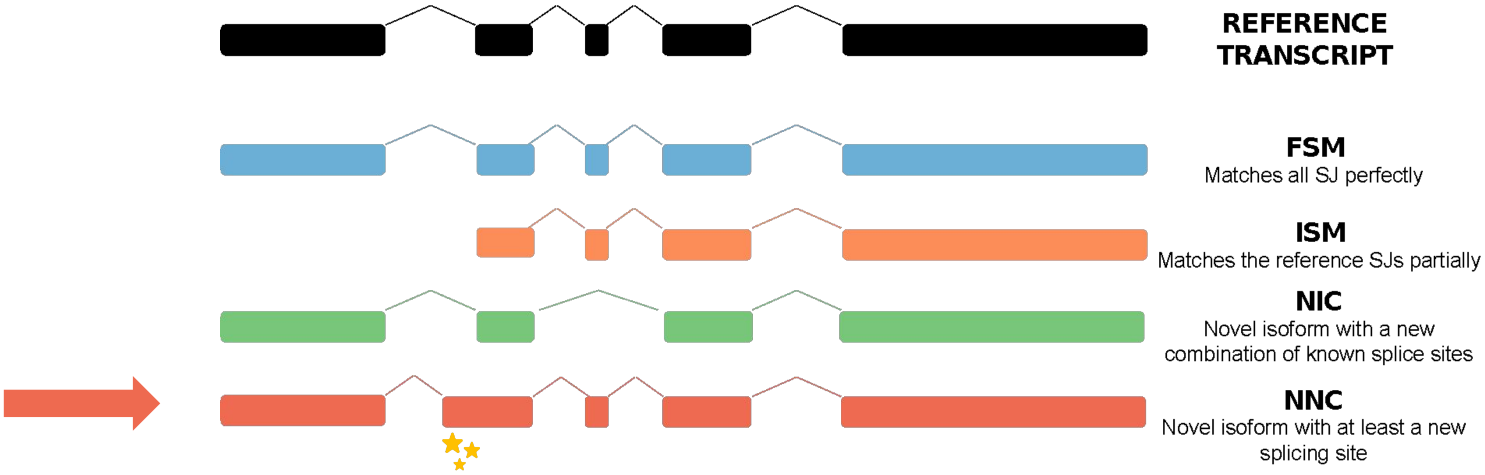
SQANTI3 structural categories

- Transcripts from **known** genes:
  - Full-Splice Match (FSM)
  - Incomplete-Splice Match (ISM)
  - Novel In Catalog (NIC)
  - Novel Not In Catalog (NNC)


REFERENCE TRANSCRIPT
FSM
Matches all SJ perfectly
ISM
Matches the reference SJs partially
NIC
Novel isoform with a new combination of known splice sites
NNC
Novel isoform with at least a new splicing site

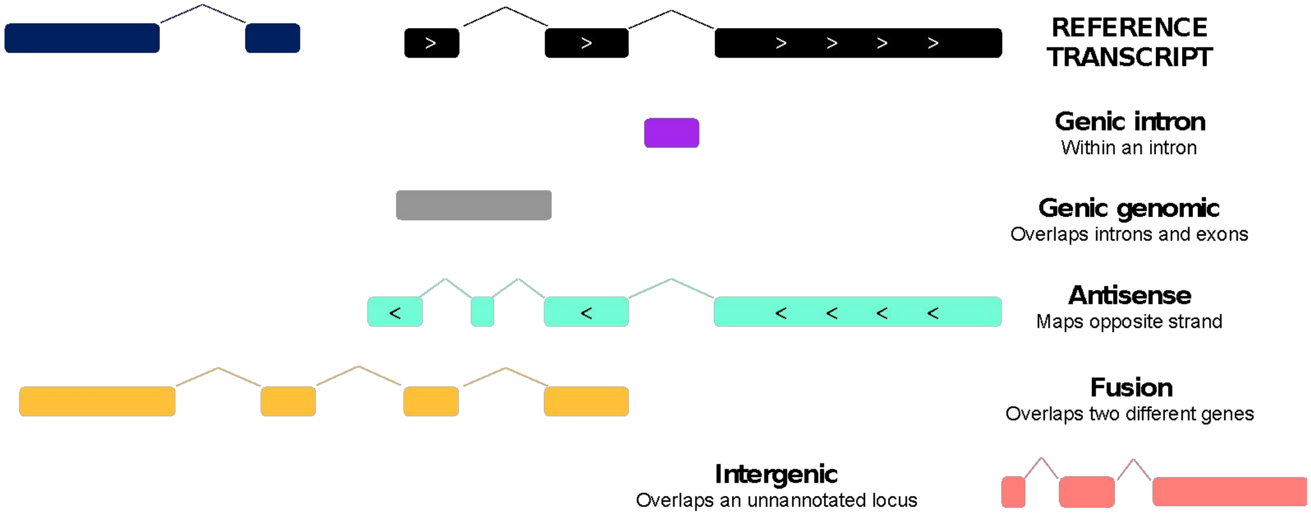
SQANTI3 structural categories

- Transcripts from **known** genes:
  - Full-Splice Match (FSM)
  - Incomplete-Splice Match (ISM)
  - Novel In Catalog (NIC)
  - Novel Not In Catalog (NNC)


REFERENCE TRANSCRIPT
FSM
Matches all SJ perfectly
ISM
Matches the reference SJs partially
NIC
Novel isoform with a new combination of known splice sites
NNC
Novel isoform with at least a new splicing site

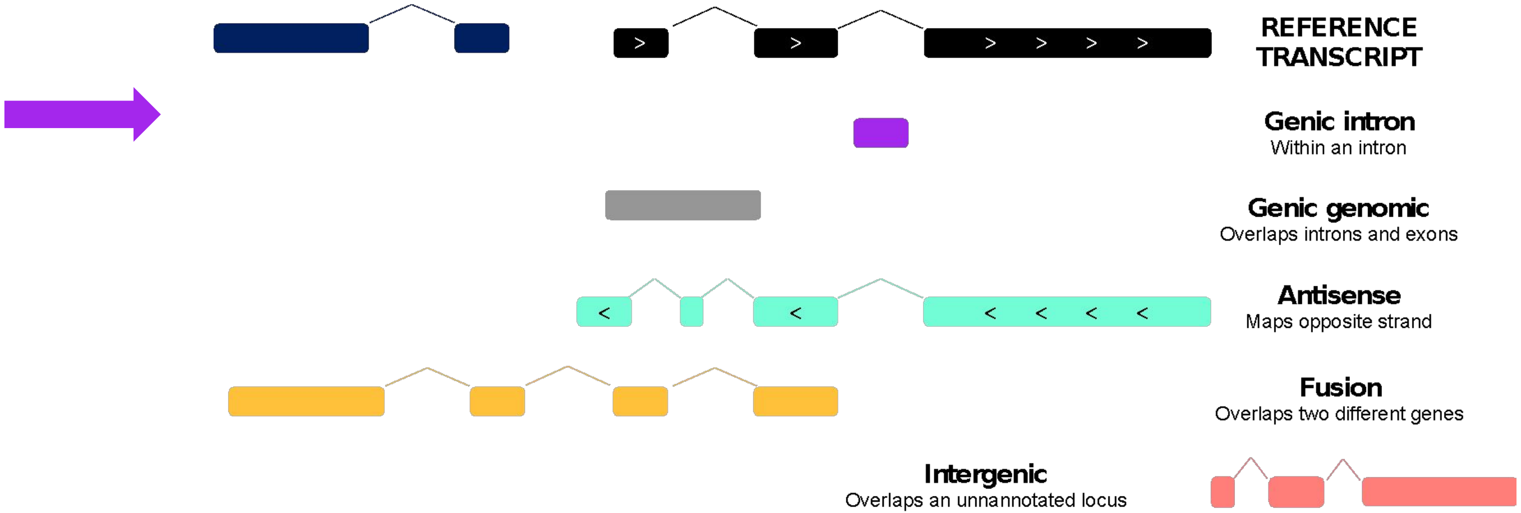
SQANTI3 structural categories

- Transcripts from “**novel**” genes:
  - Genic Intron
  - Genic Genomic
  - Antisense
  - Fusion
  - Intergenic



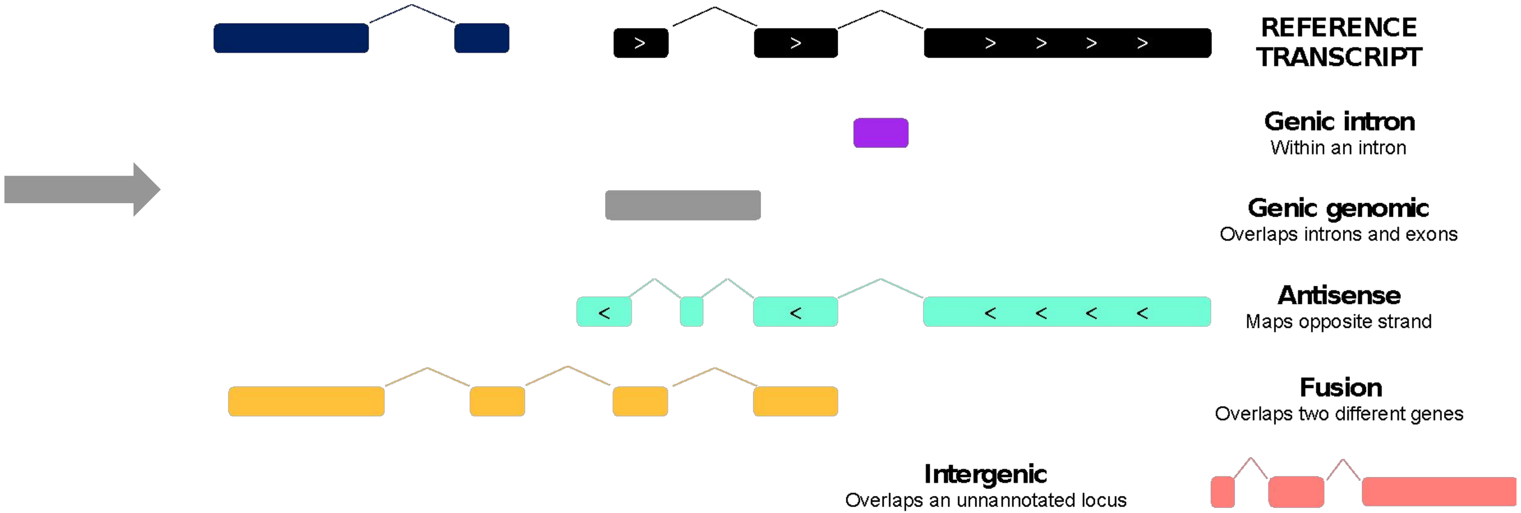
SQANTI3 structural categories

- Transcripts from “**novel**” genes:
  - Genic Intron
  - Genic Genomic
  - Antisense
  - Fusion
  - Intergenic



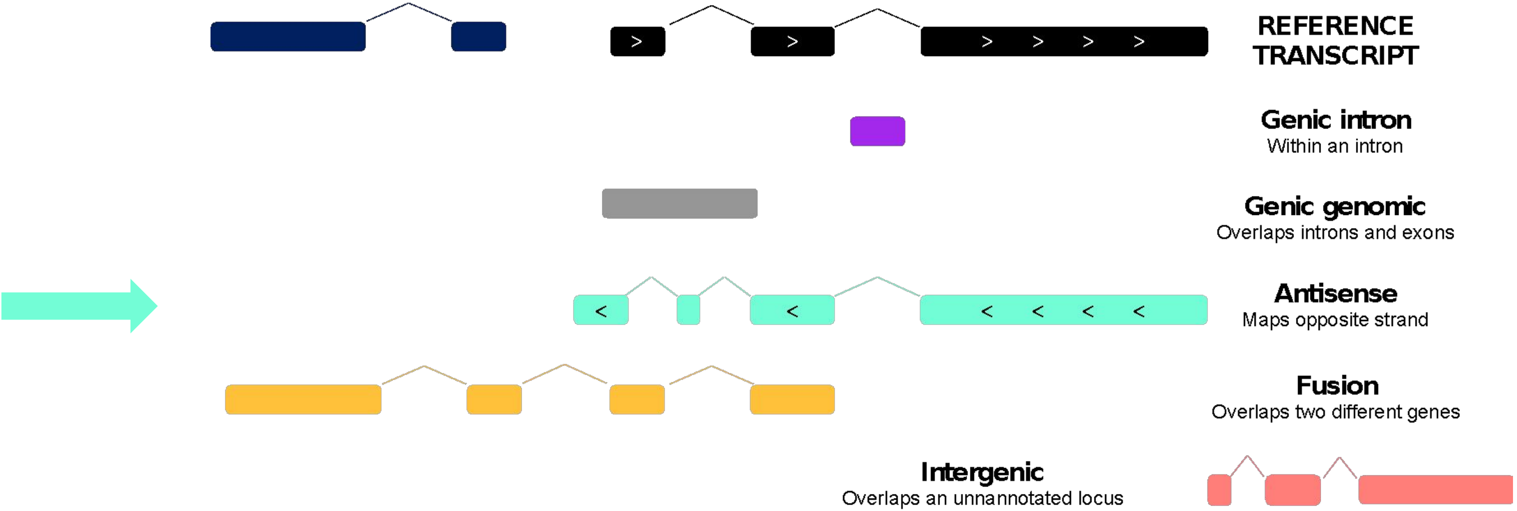
SQANTI3 structural categories

- Transcripts from “**novel**” genes:
  - Genic Intron
  - Genic Genomic
  - Antisense
  - Fusion
  - Intergenic



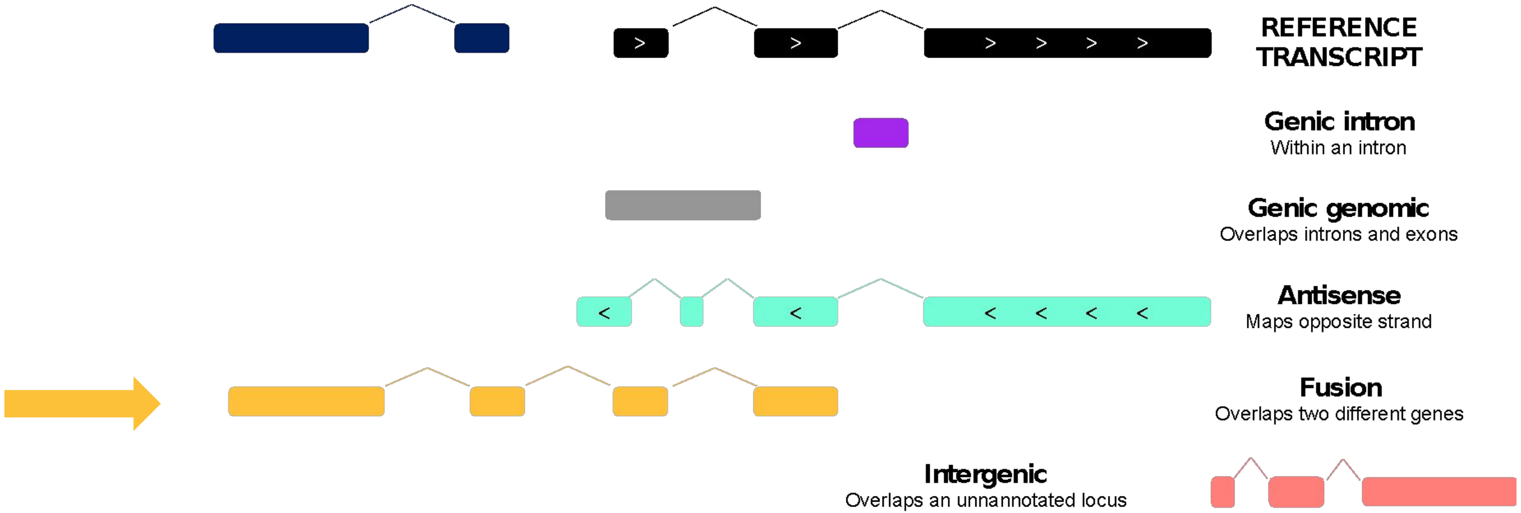
SQANTI3 structural categories

- Transcripts from “**novel**” genes:
  - Genic Intron
  - Genic Genomic
  - Antisense
  - Fusion
  - Intergenic



SQANTI3 structural categories

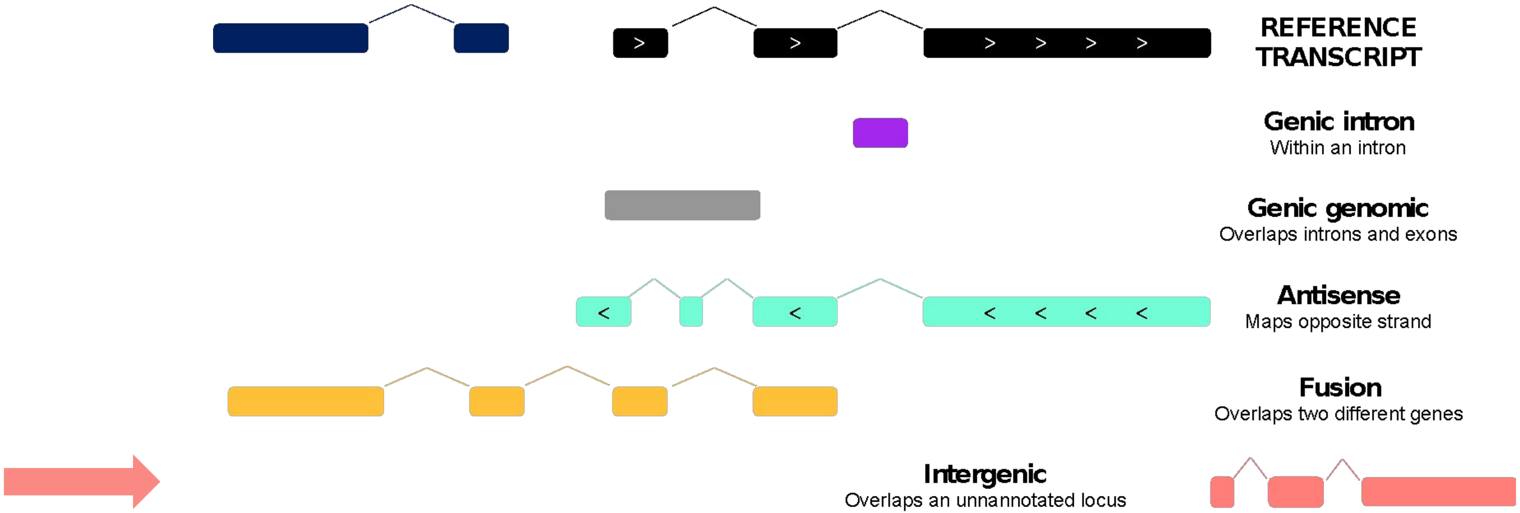
- Transcripts from “**novel**” genes:
  - Genic Intron
  - Genic Genomic
  - Antisense
  - Fusion
  - Intergenic



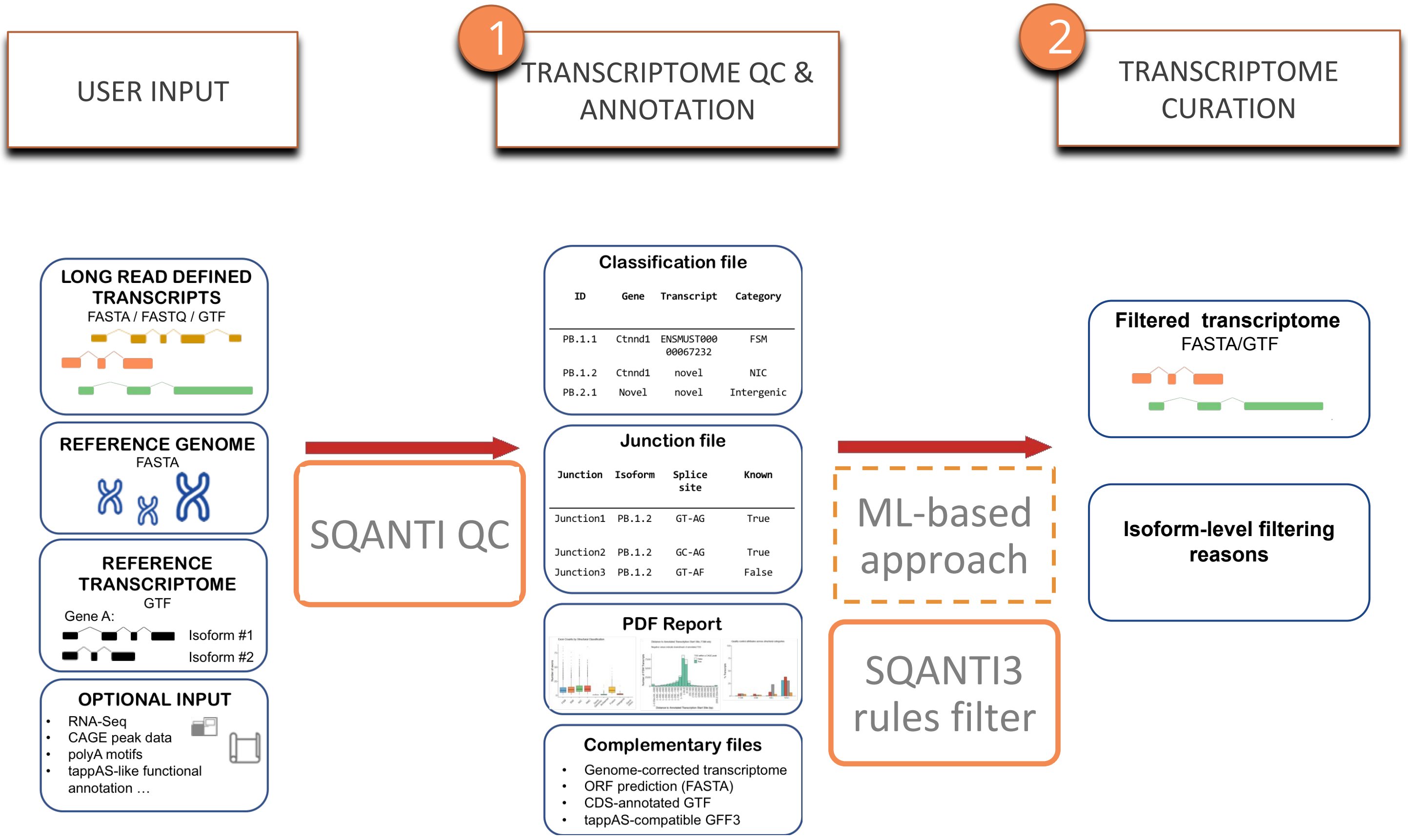
SQANTI3 structural categories

SQANTI 3

- Transcripts from “**novel**” genes:
  - Genic Intron
  - Genic Genomic
  - Antisense
  - Fusion
  - Intergenic



SQANTI3 workflow



USER INPUT

1 TRANSCRIPTOME QC & ANNOTATION

2 TRANSCRIPTOME CURATION

LONG READ DEFINED TRANSCRIPTS
FASTA / FASTQ / GTF

REFERENCE GENOME
FASTA

REFERENCE TRANSCRIPTOME
GTF

OPTIONAL INPUT

- RNA-Seq
- CAGE peak data
- polyA motifs
- tappAS-like functional annotation …

SQANTI QC

Classification file

ID	Gene	Transcript	Category
PB.1.1	Ctnnd1	ENSMUST000 00067232	FSM
PB.1.2	Ctnnd1	novel	NIC
PB.2.1	Novel	novel	Intergenic

Junction file

Junction	Isoform	Splice site	Known
Junction1	PB.1.2	GT-AG	True
Junction2	PB.1.2	GC-AG	True
Junction3	PB.1.2	GT-AF	False

PDF Report

Complementary files

- Genome-corrected transcriptome
- ORF prediction (FASTA)
- CDS-annotated GTF
- tappAS-compatible GFF3

ML-based approach

SQANTI3 rules filter

Filtered transcriptome
FASTA/GTF

Isoform-level filtering reasons

Quality control and characterization of LR transcriptomes

Francisco J. Pardo-Palacios

Lorena de la Fuente

SQANTI 3

Genomics of Gene Expression Lab