SQANTI: Classification, Curation and Quantification of a PacBio transcriptome

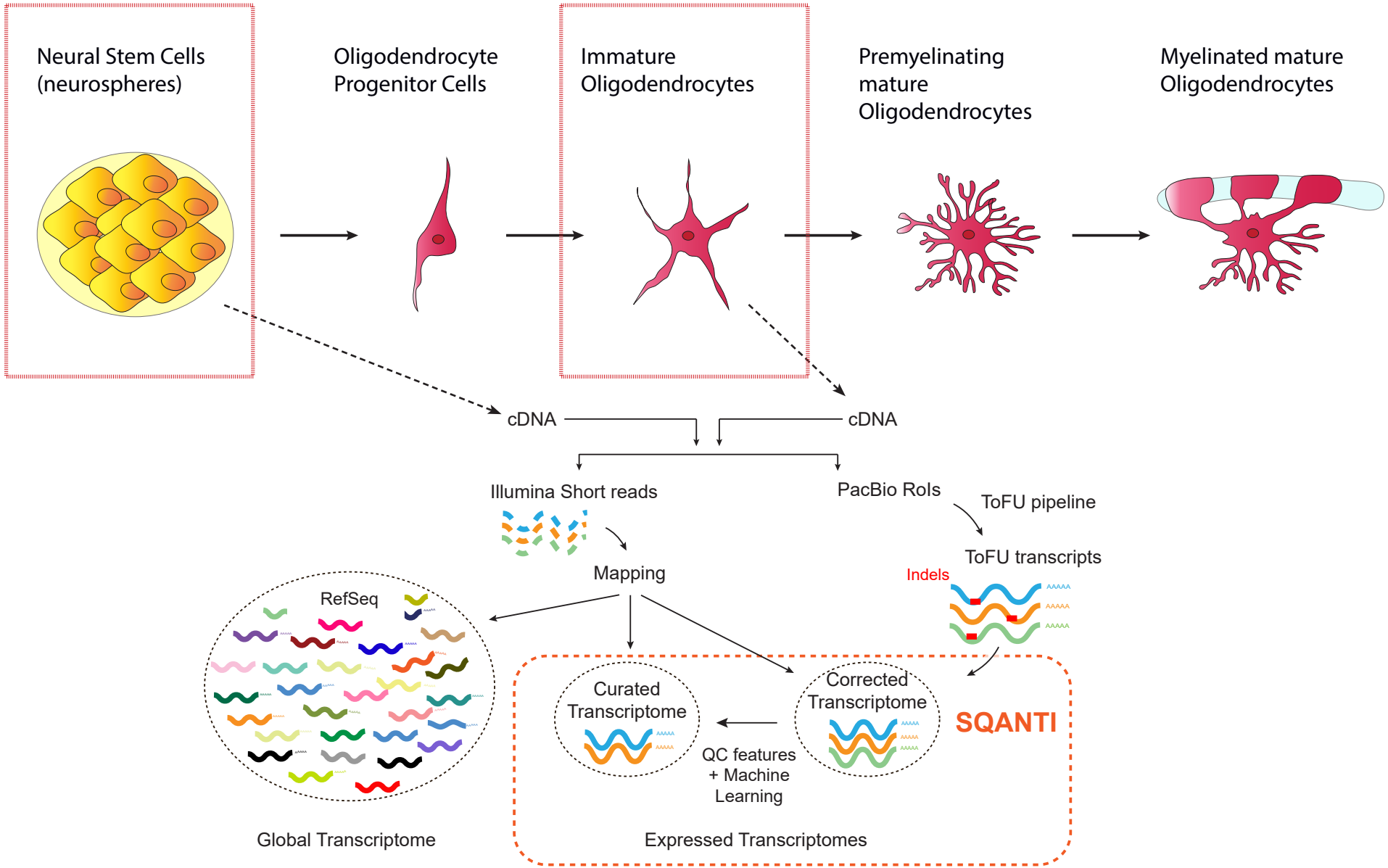
Manuel Tardaguila, PhD

Ana Conesa Lab

UF | UNIVERSITY of FLORIDA

PRINCIPE FELIPE
CENTRO DE INVESTIGACION

PacBio in Oligodendrogenesis

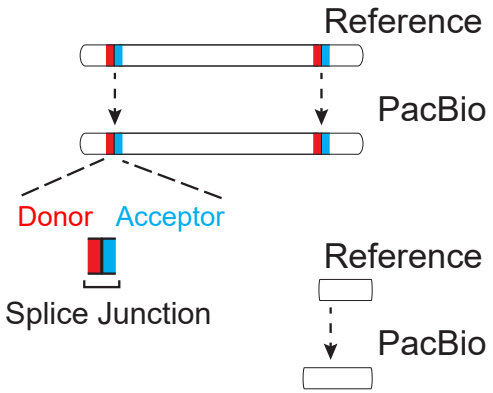


Tardaguila et al ***SQANTI: extensive characterization of long read transcript sequences for quality control in full-length transcriptome identification and quantification*** *Pre-print BioRxiv (2017)*

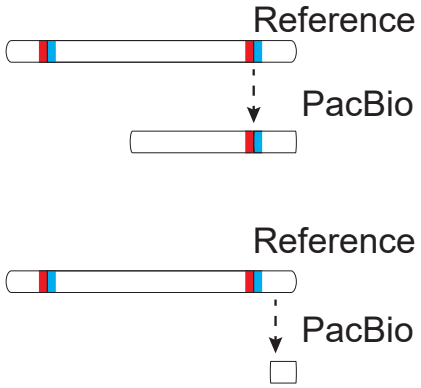
1. CLASSIFICATION OF PACBIO TRANSCRIPTS

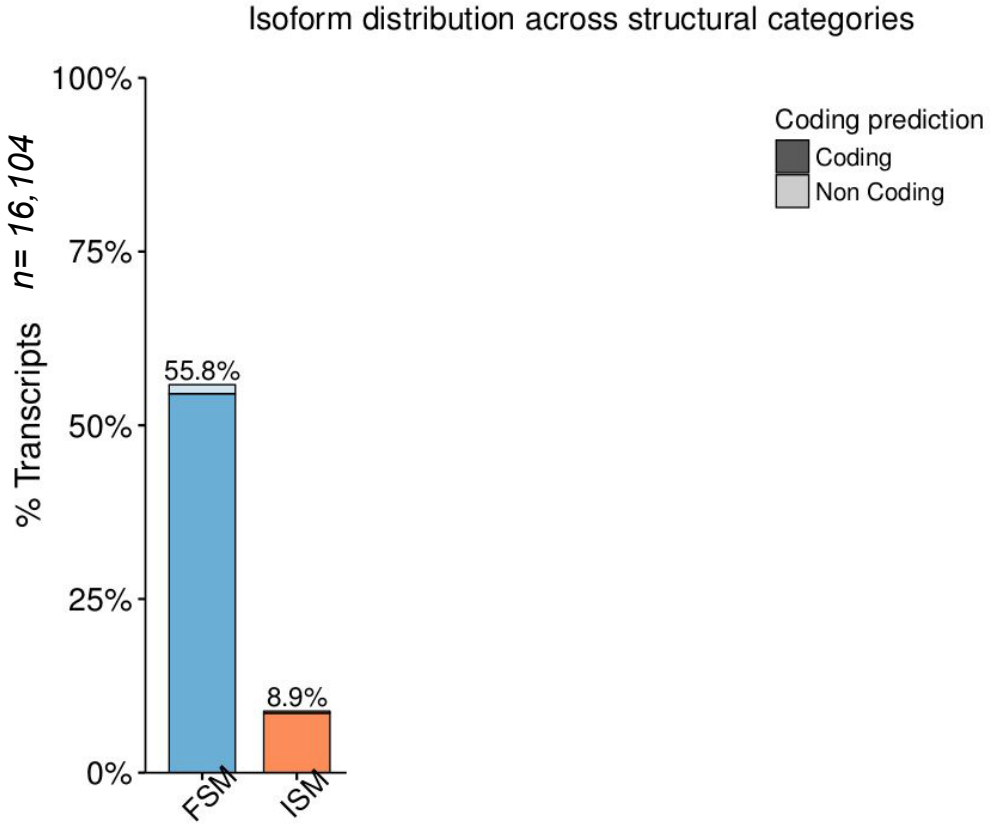
Splice-based classification (I): Known Transcripts

Full Splice Match (FSM)

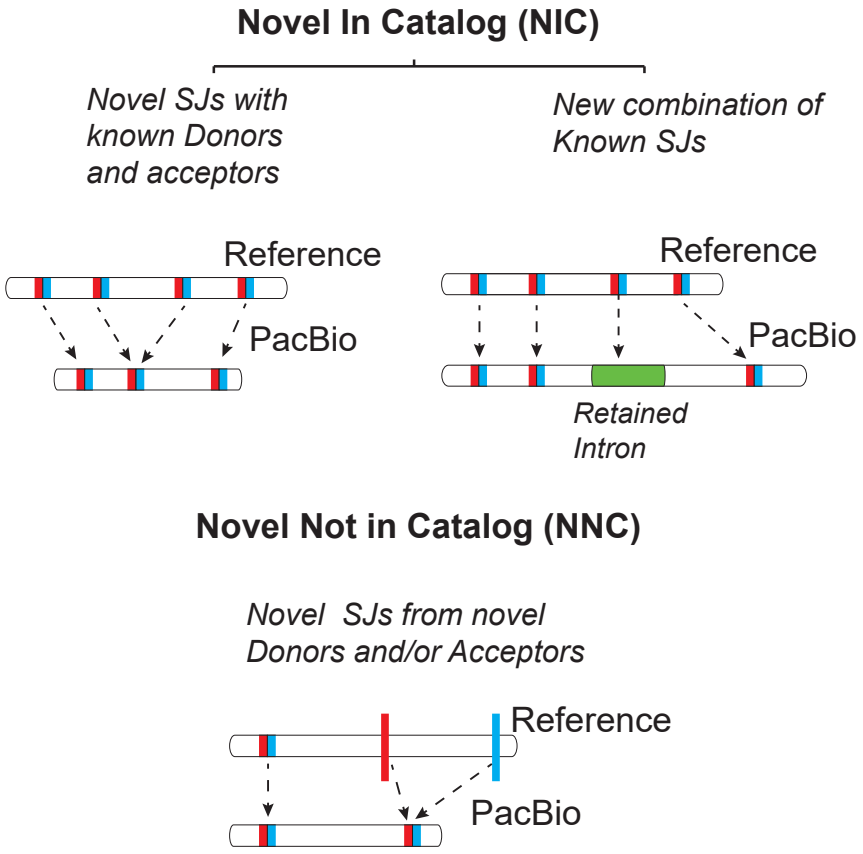


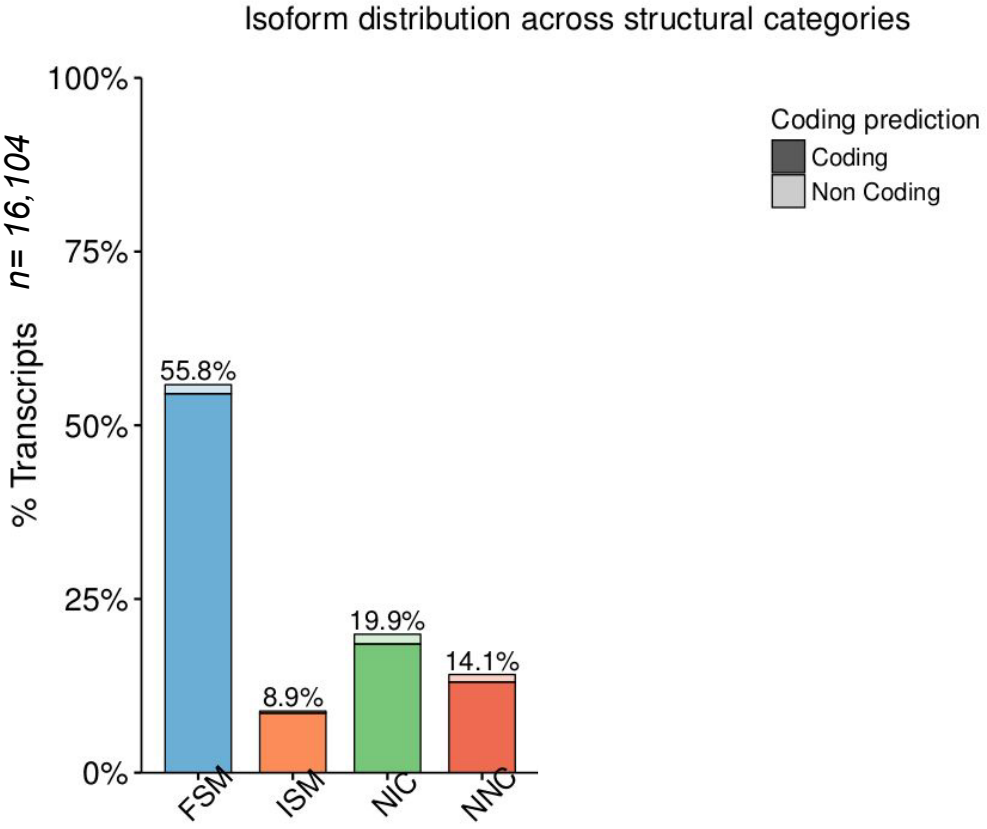
Incomplete Splice Match (ISM)





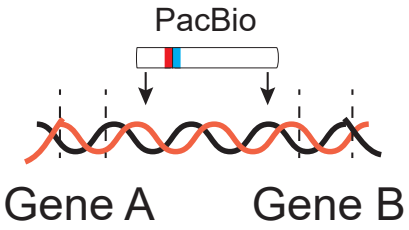
Splice-based classification (II): Novel Transcripts from known genes



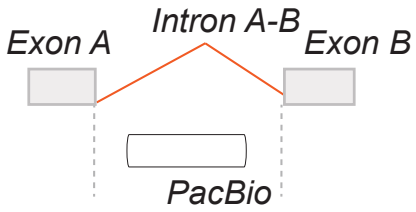


Splice-based classification (III): Novel Transcripts from novel genes

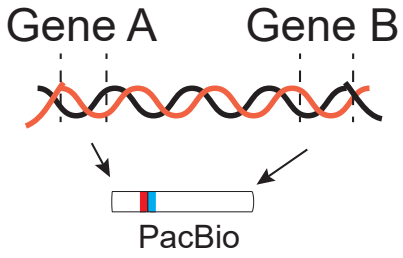
Intergenic

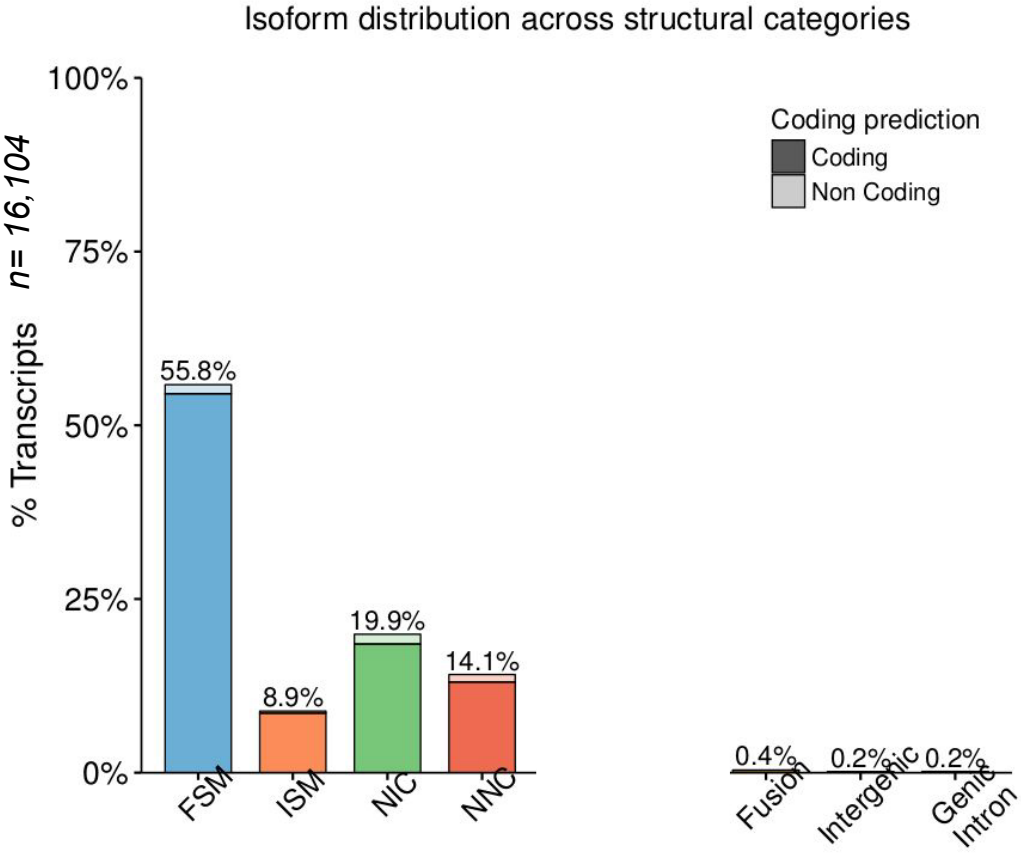


Genic Intron



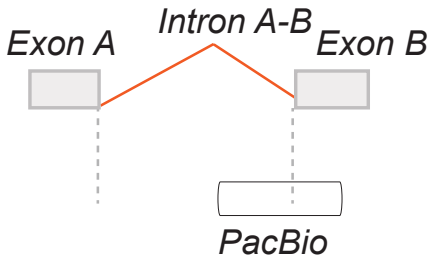
Fusion Transcripts



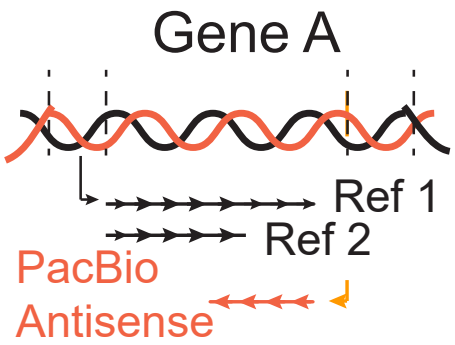


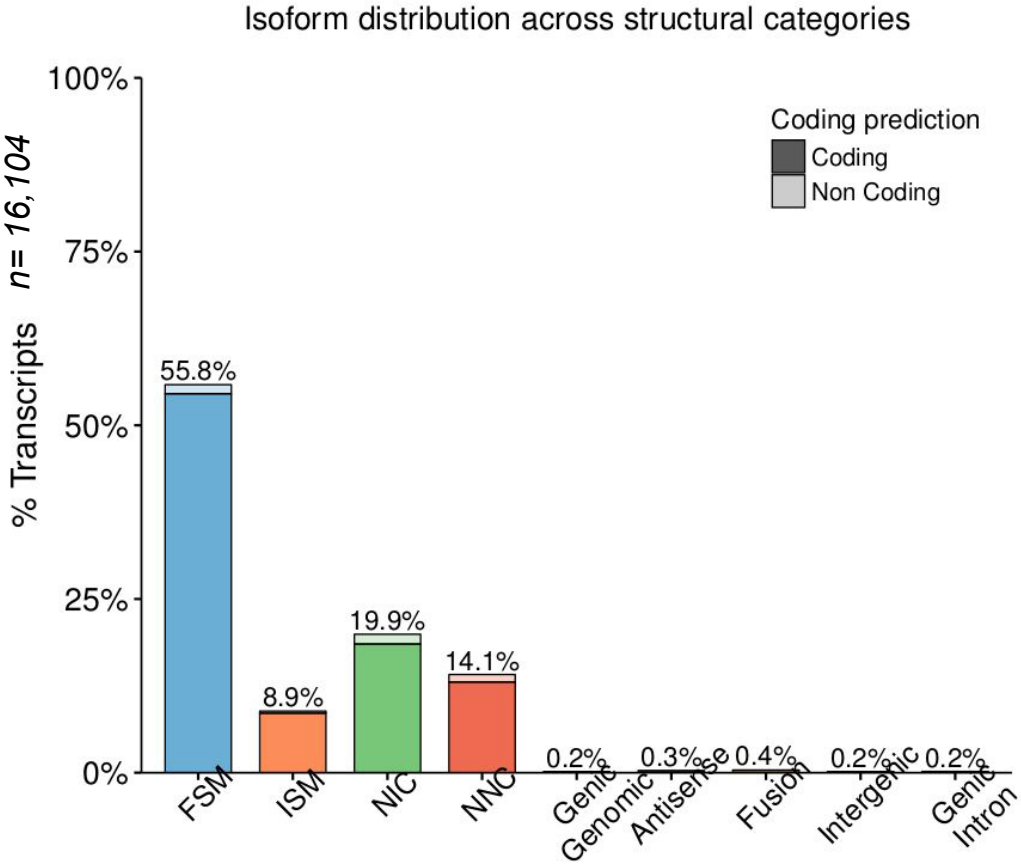
Splice-based classification (IV): Antisense and Genic Genomic transcripts

Genic Genomic



Antisense



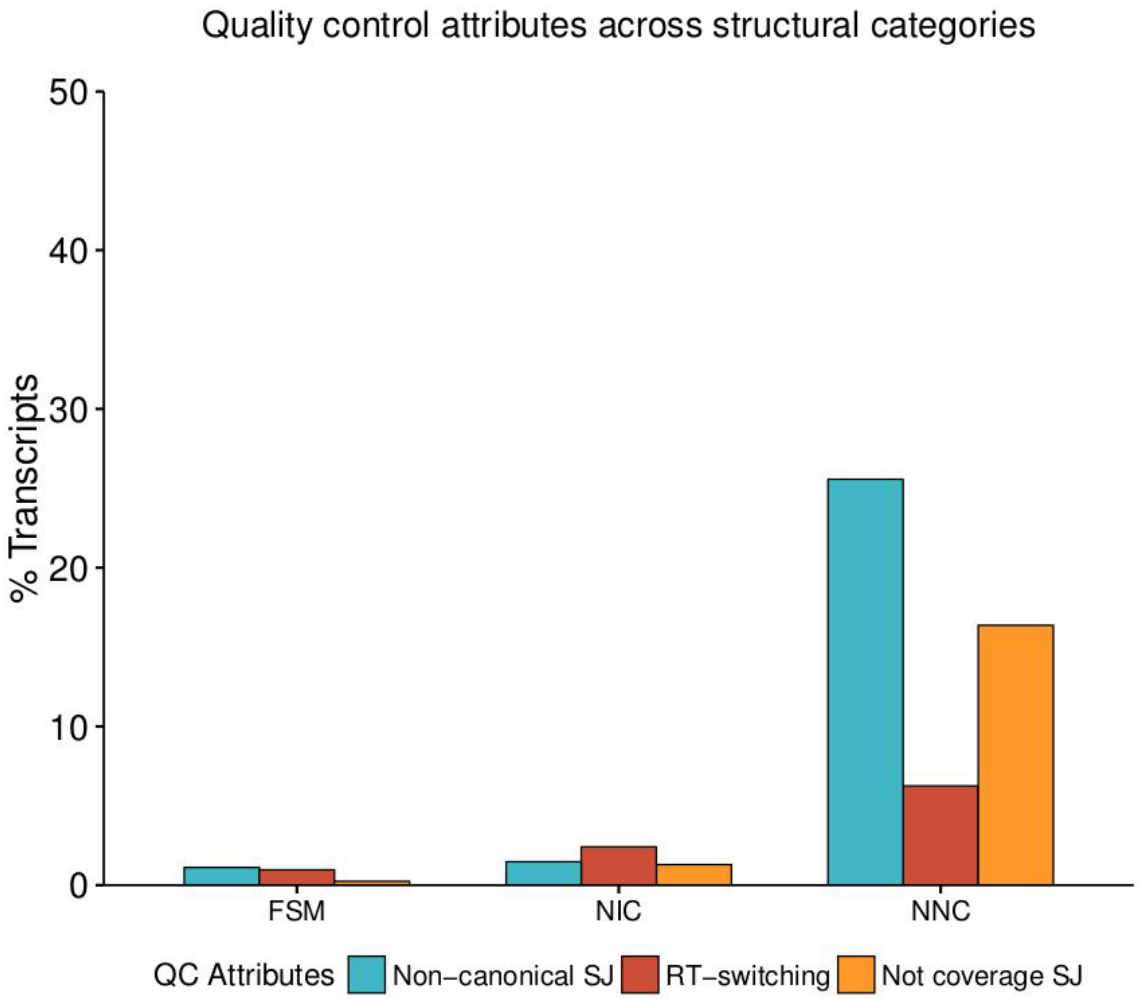


CLASSIFICATION SUMMARY

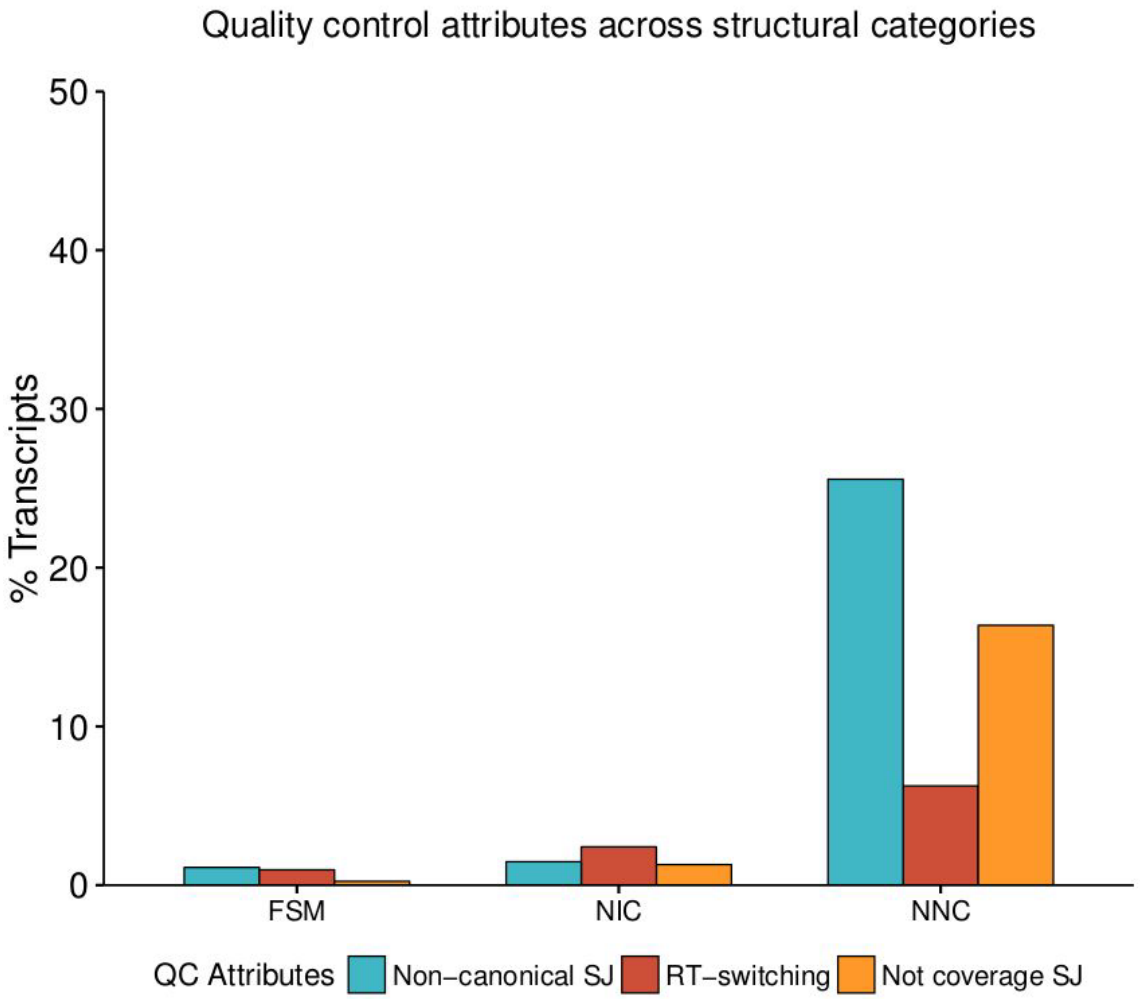
- Splice based classification allows to separate known transcripts (FSM and ISM) from novel transcripts arising from novel splice junctions or new combinations of already known splice junctions (NIC and NNC)
- The predominant categories in our transcriptome are FSM,ISM, NIC and NNC and they are mostly coding (oligodT library preparation)

2. CURATION OF PACBIO TRANSCRIPTS

40% novel isoforms in mouse... Are all of them real?

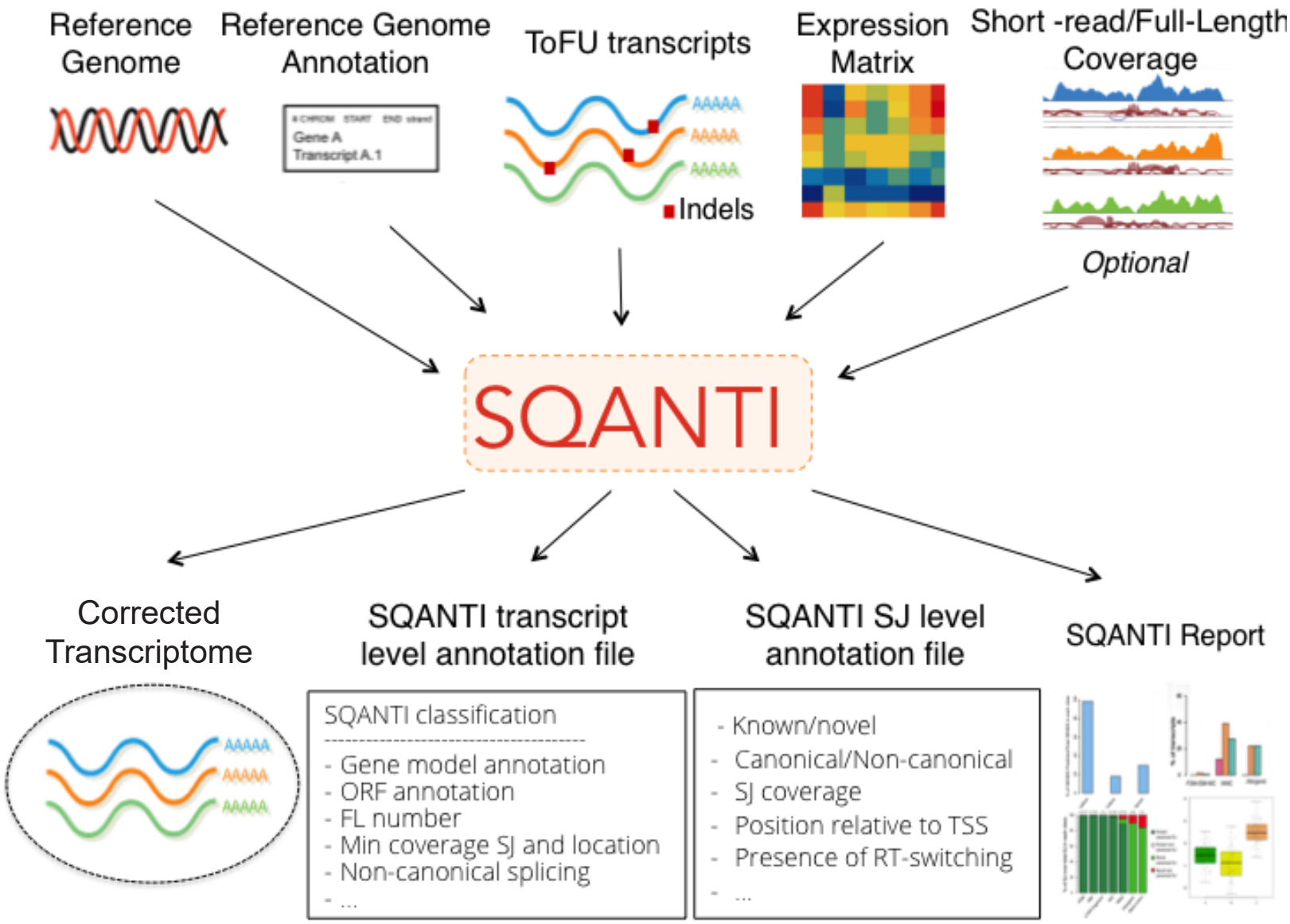


PacBio output needs curation

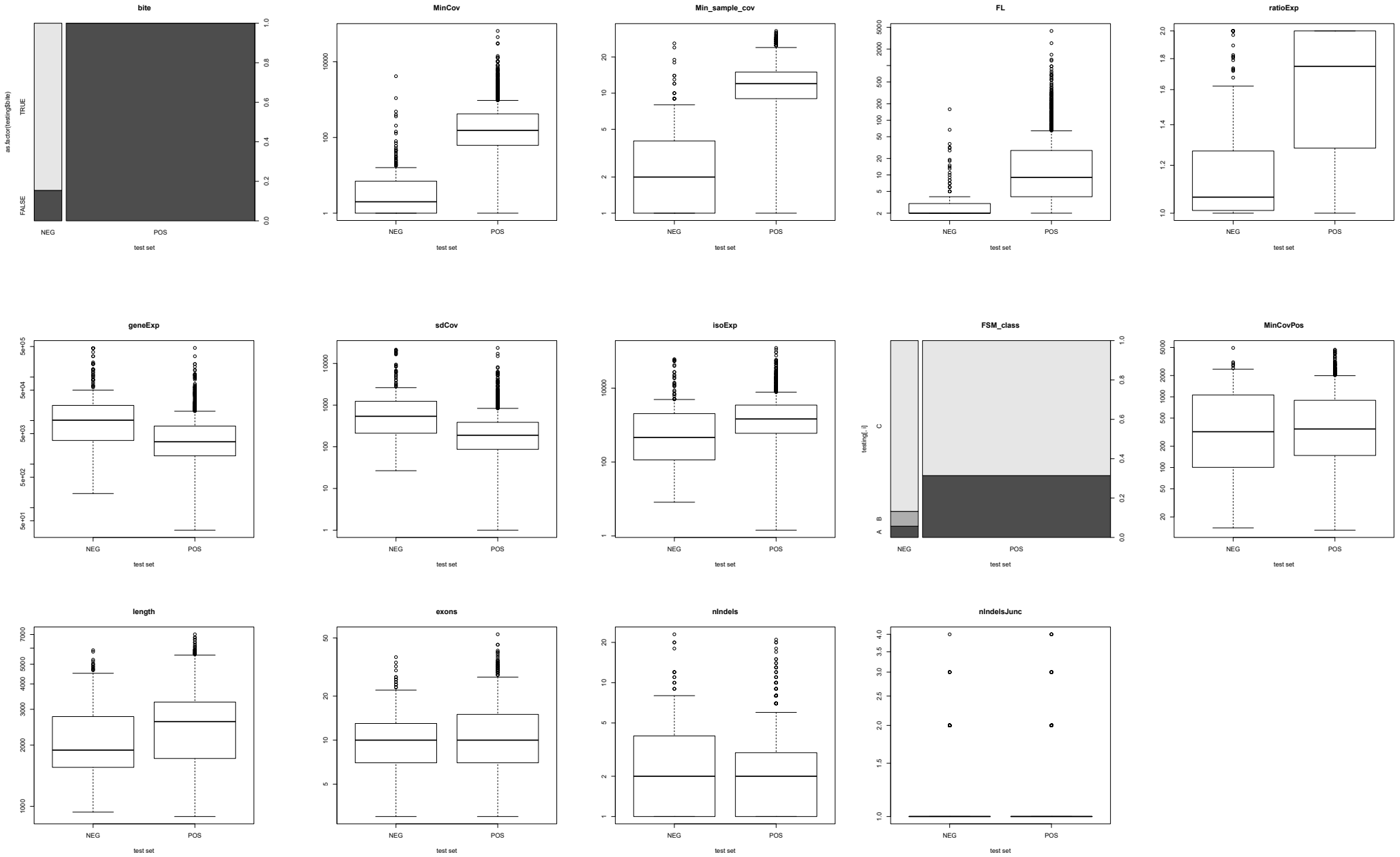


- NNC transcripts characterized by novel Donors and/or Acceptors of splicing concentrate traits of low quality transcripts

Using Short reads and Long reads to mine QC features

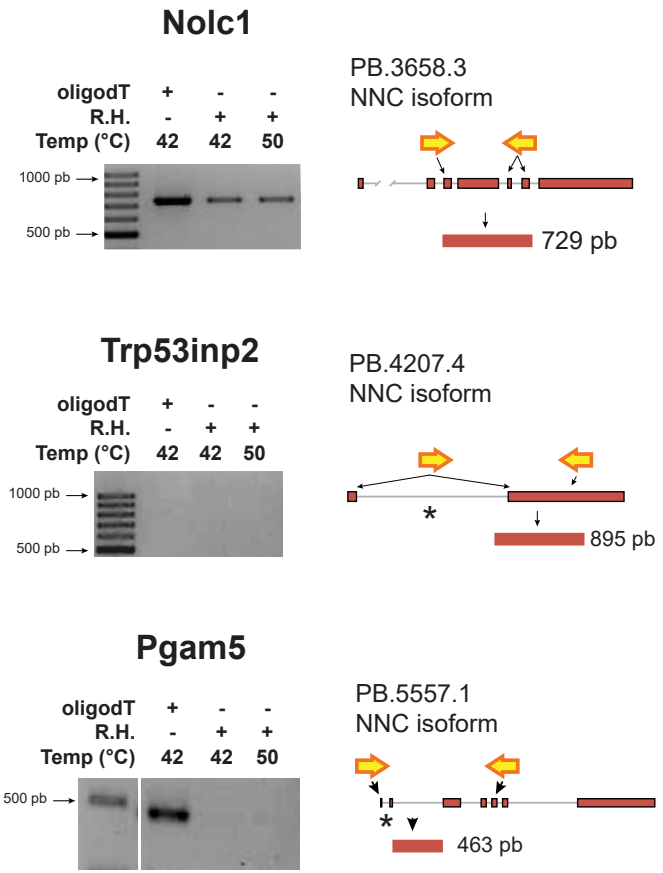


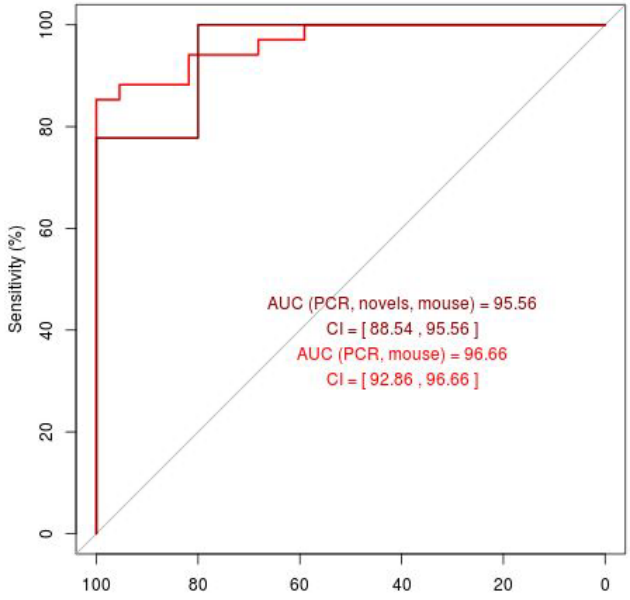
Features selected for Random Forest Classification



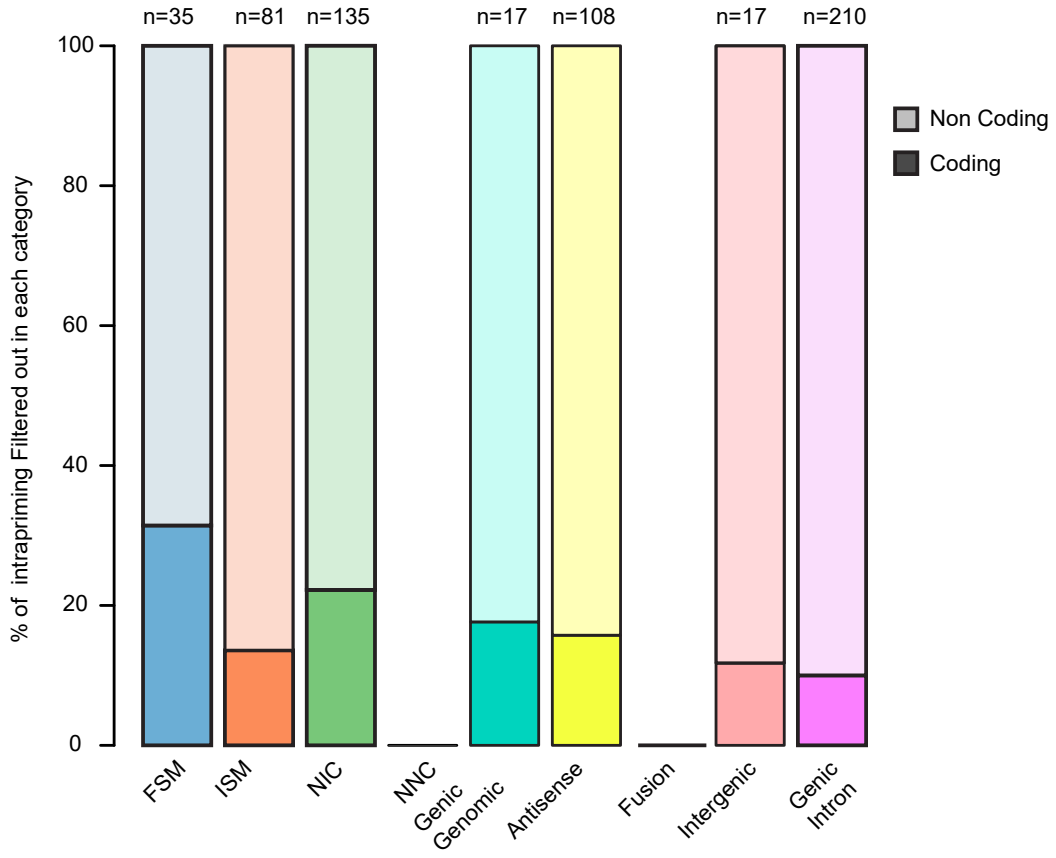
PCR validation in an independent set of transcripts

		FSM	NIC	NNC	Fusion
oligo (dT) PCR	Positive	23 (3nc)	7	8 (3 nc)	1
	Negative	0	2	12 (8 nc)	2
	Total	23	9	20	3
R.H. PCR	Positive	5 (3nc)	7	2	1
	Negative	0	0	6 (3nc)	0
	Total	5	7	8	1



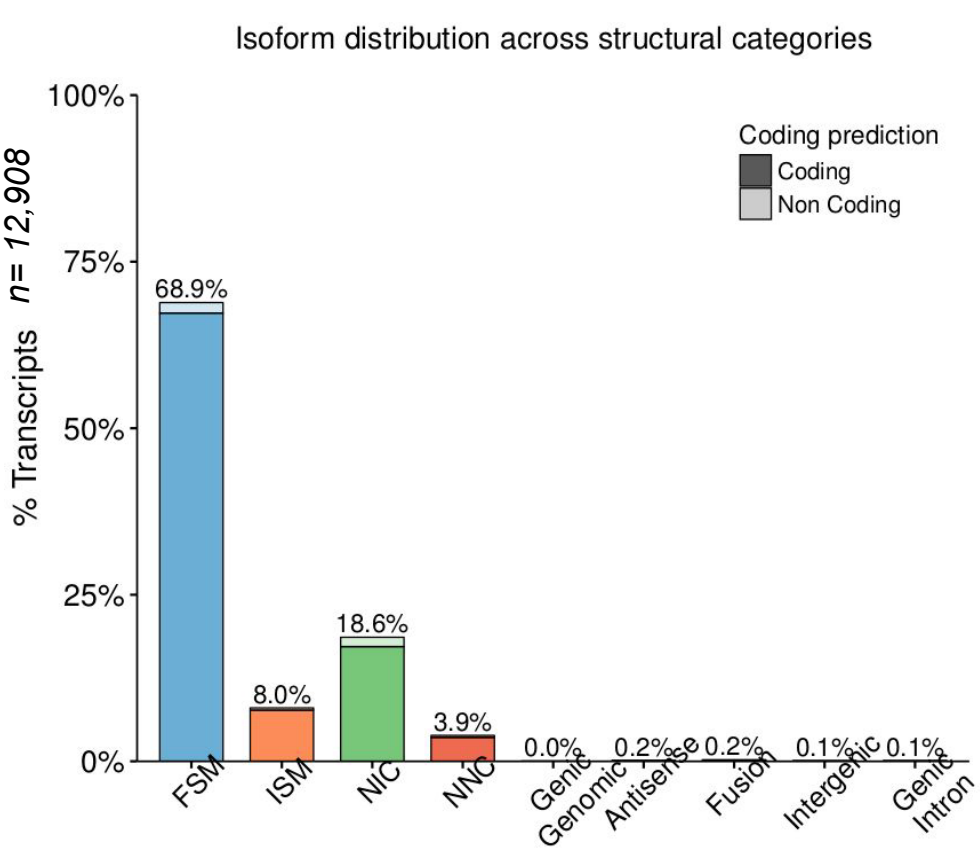


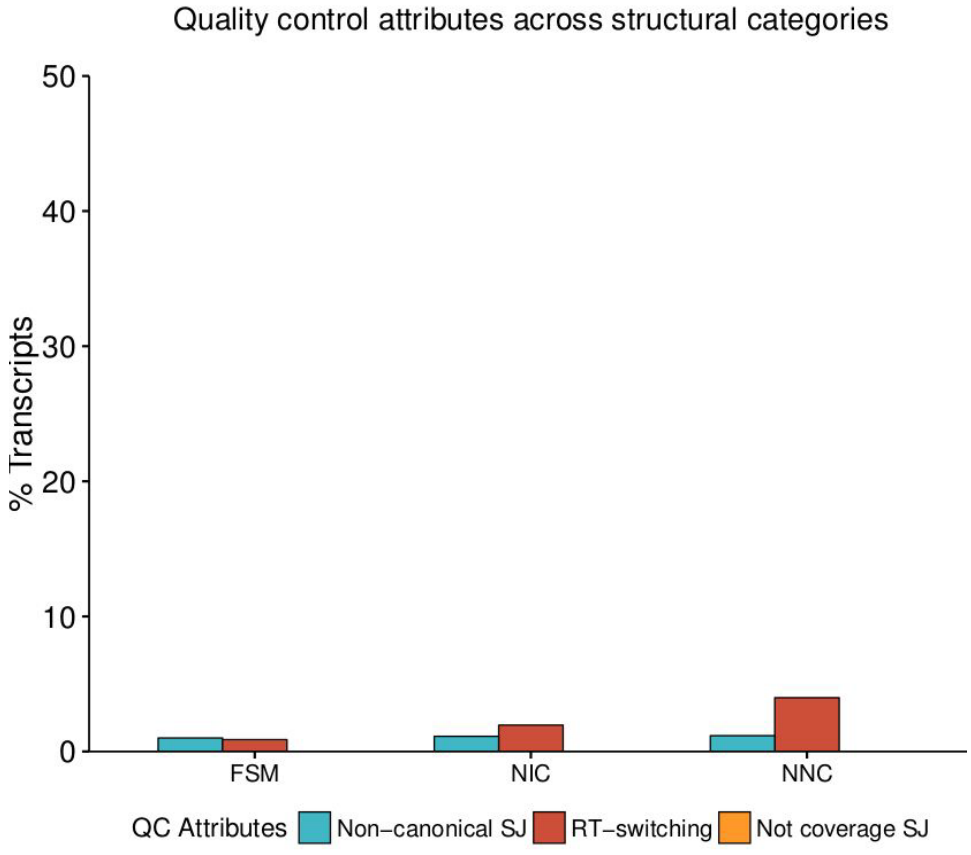
Splice Junction independent curation: Intrapriming features



- oligodT can prime outside polyA regions in A rich regions inside transcripts
- We looked for transcripts showing >= 80% Adenines in the 20 nts downstream (DNA) the end of thetranscript
- If this transcripts lack a consensus poly Adenilation site we filter them out
- 605 transcripts filtered out

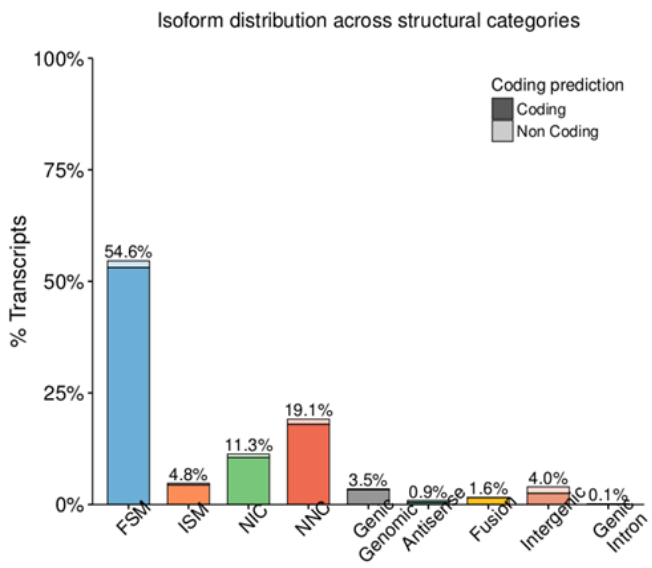
SQANTI filter eliminates transcript with bad QC features

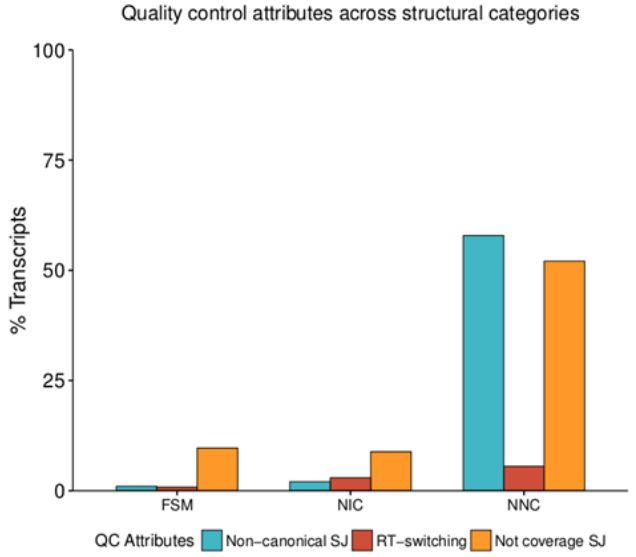




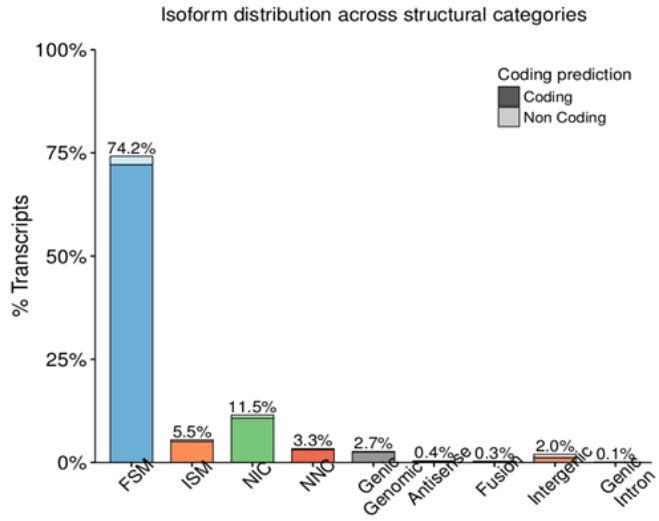
SQANTI filter works across different PacBio datasets: Maize

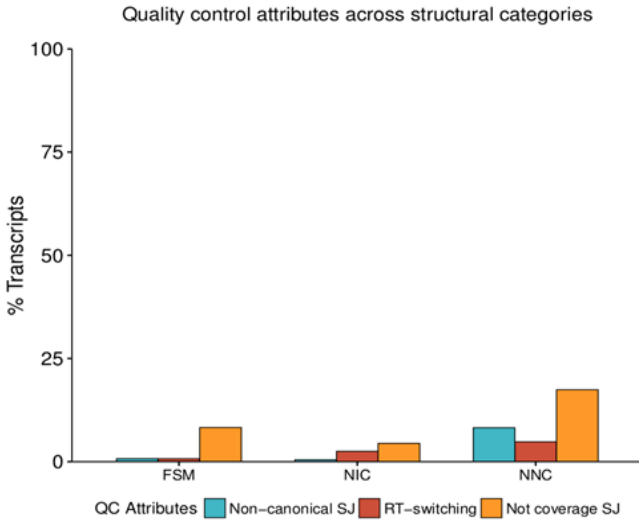
- Before SQANTI





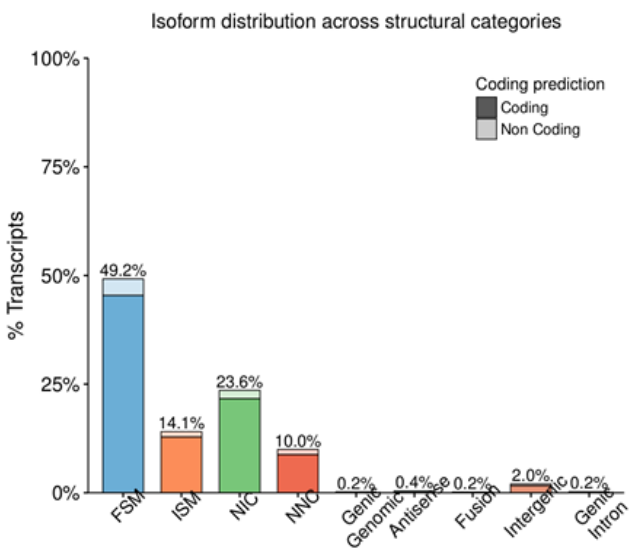
- AFTER SQANTI

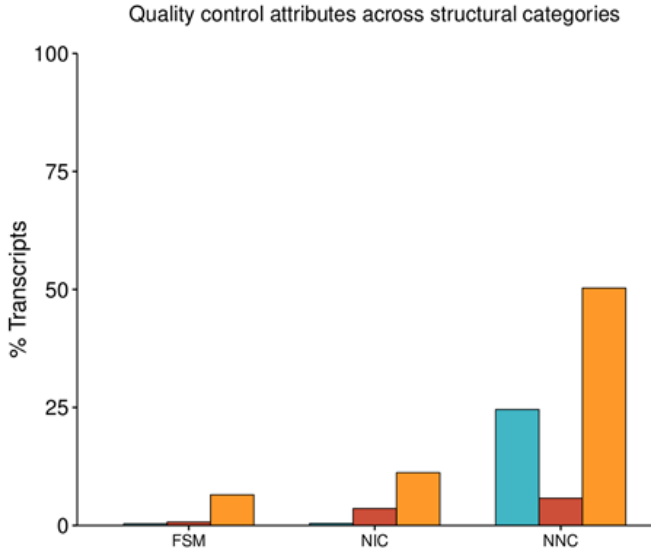




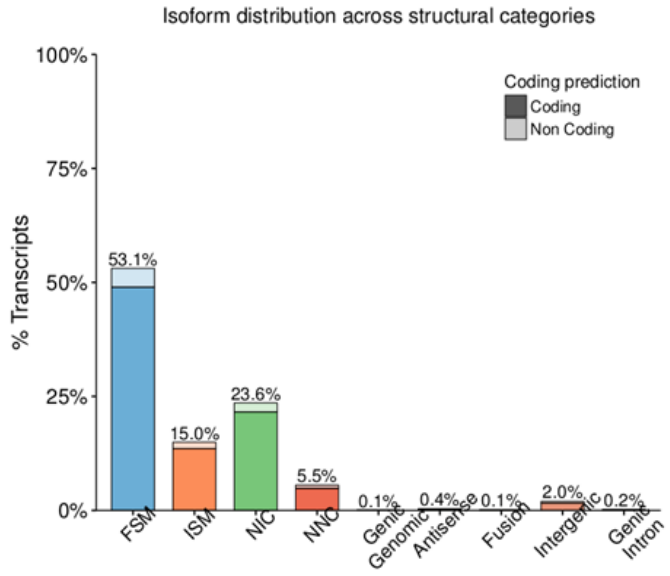
SQANTI filter works across different PacBio datasets: MCF-7

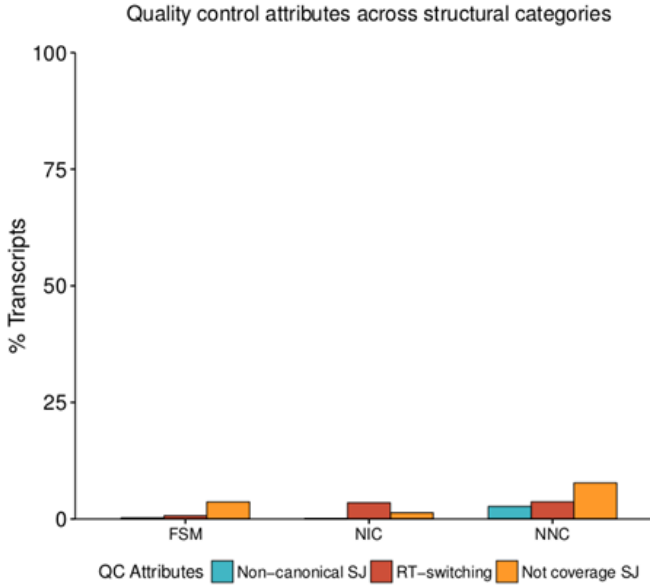
- Before SQANTI





- AFTER SQANTI



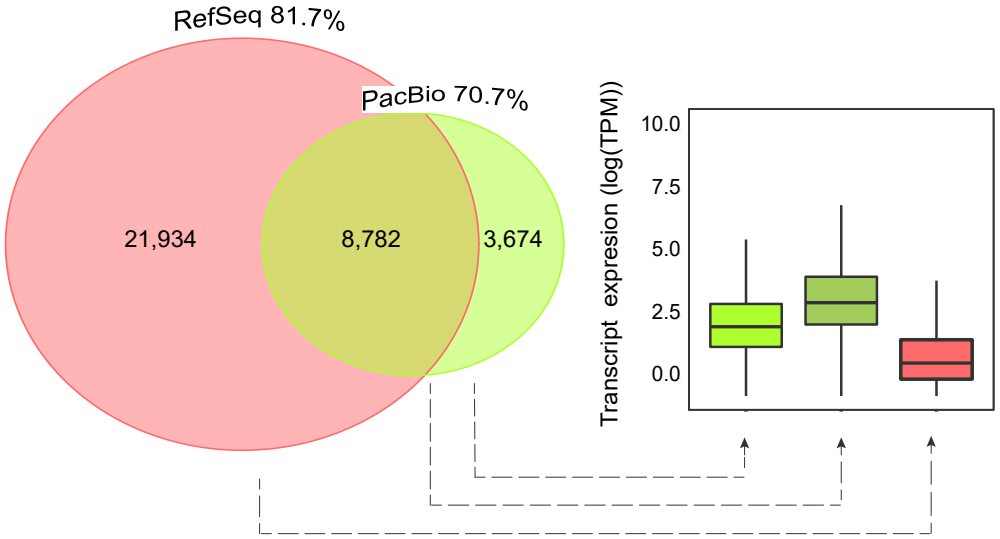


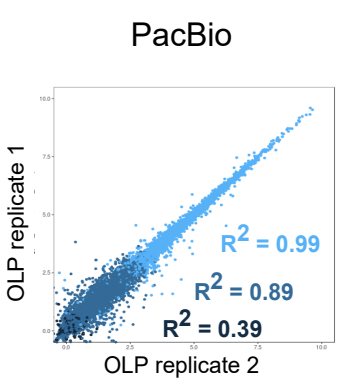
CURATION SUMMARY

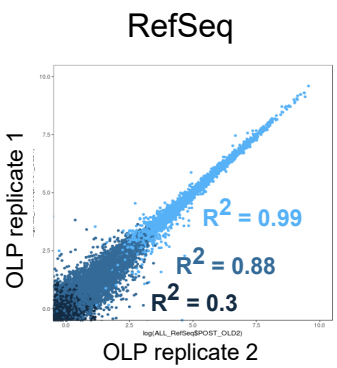
- PacBio output needs curation, specially in NNC novel acceptor/ donnor transcripts
- SQANTI mines features based in splice junctions, expression and structure to feed a RF classifier that succeeds in filtering out transcripts that fail to validate in an independent PCR set
- SQANTI works across different datasets

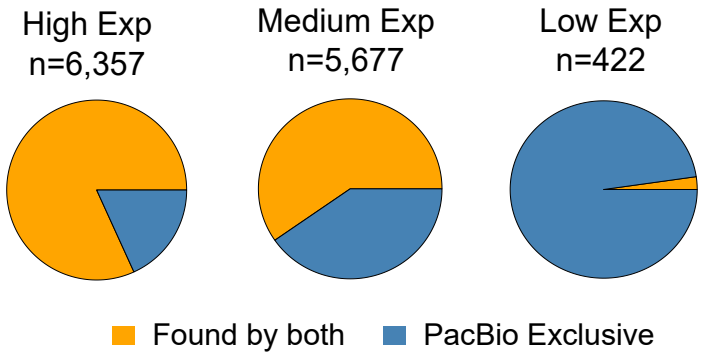
3. QUANTIFICATION OF PACBIO TRANSCRIPTOME

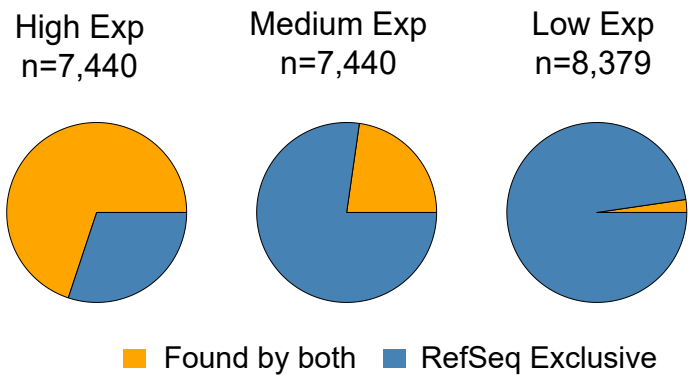
Quantification: PacBio curated vs RefSeq: Transcripts





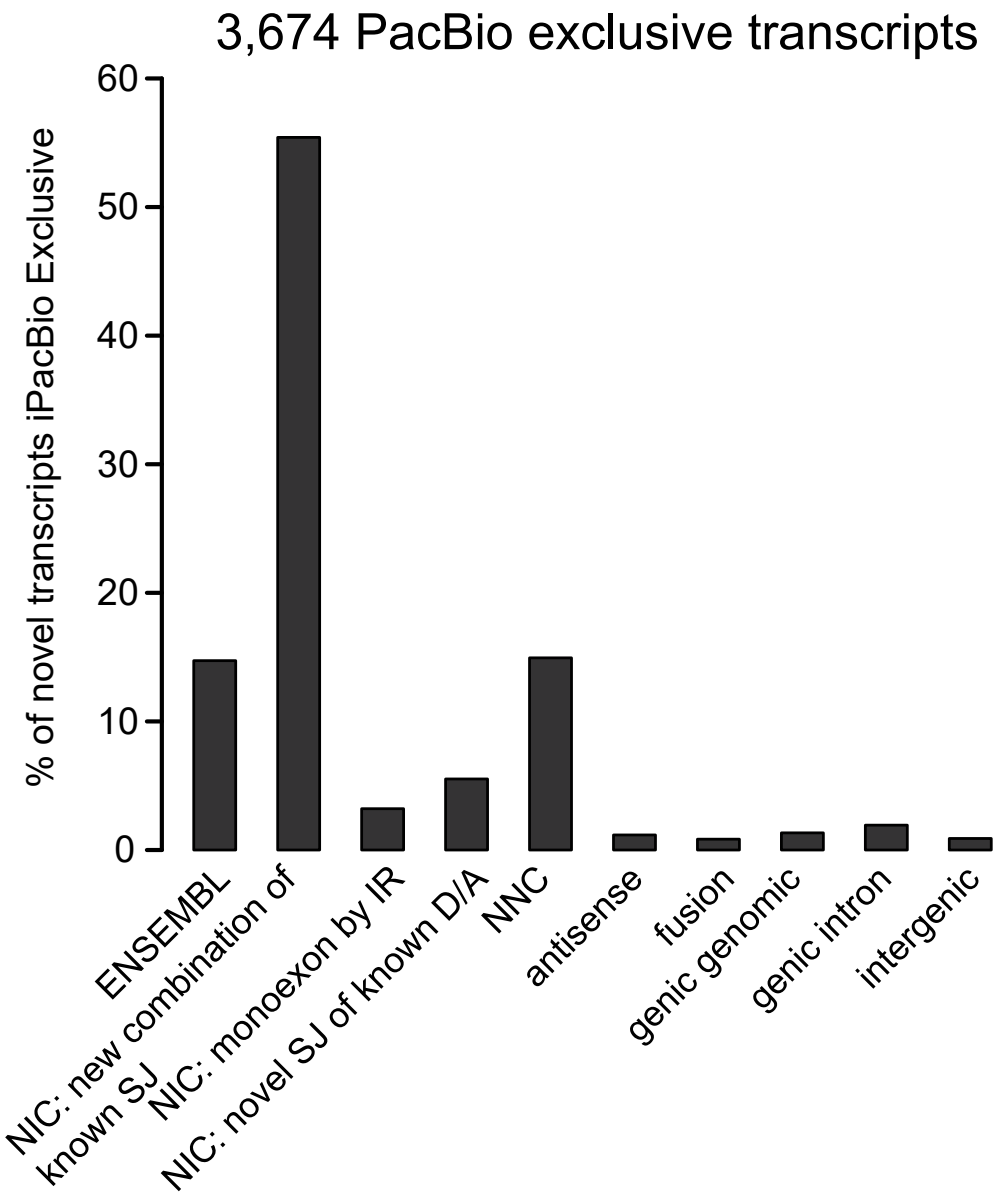




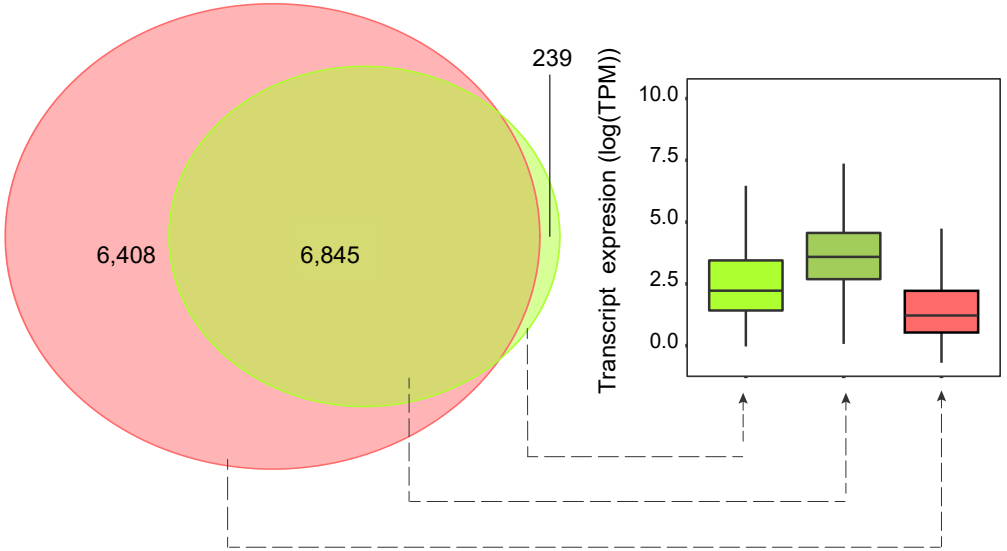


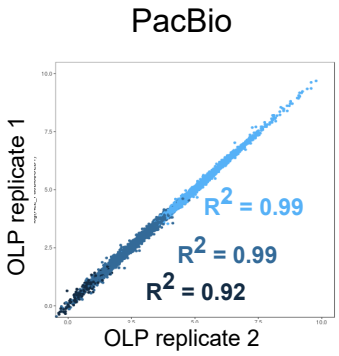
- PacBio finds a set of highly expressed transcripts and reproducible transcripts in common with RefSeq
- RefSeq detects lots of lowly expressed, difficult to reproduce transcripts. However it also captures exclusively 30% of highly expressed ones (PacBio 18%).

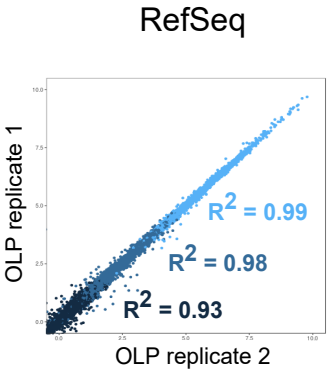
Most of the PacBio exclusive transcripts are new combinations of already known Splice Junction

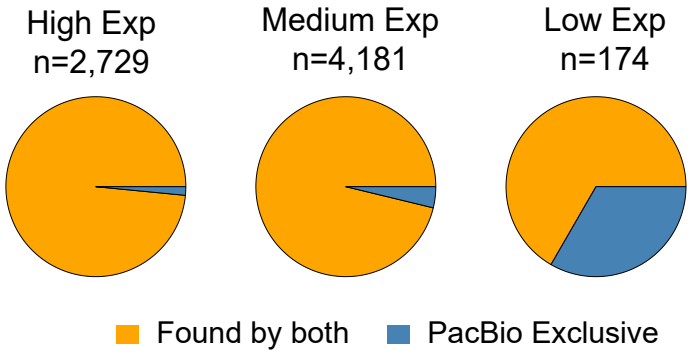


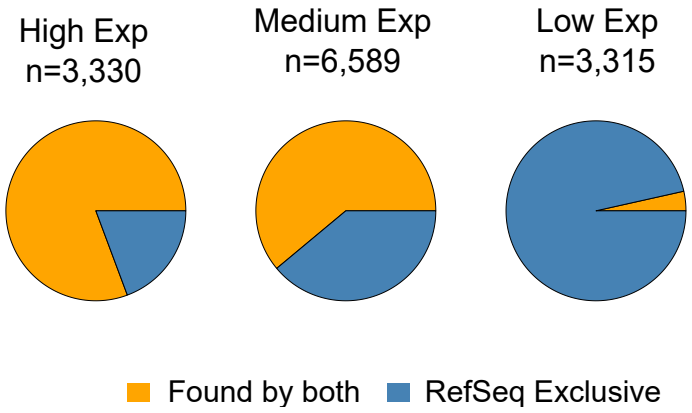
Quantification: PacBio curated vs RefSeq: Genes





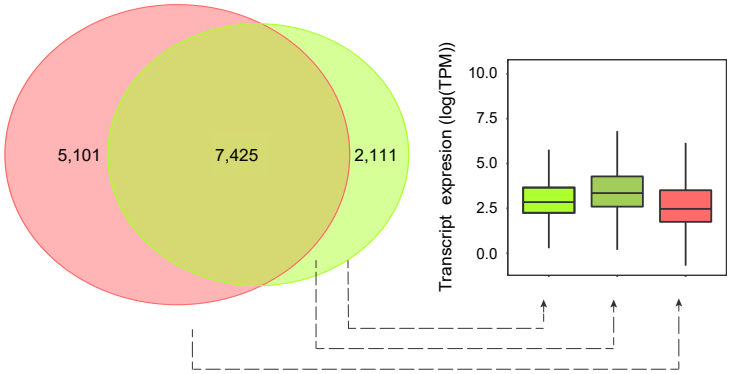


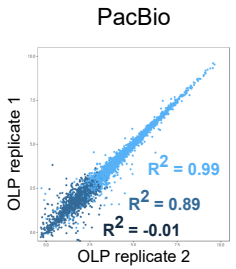


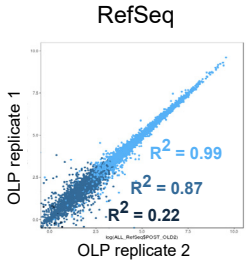


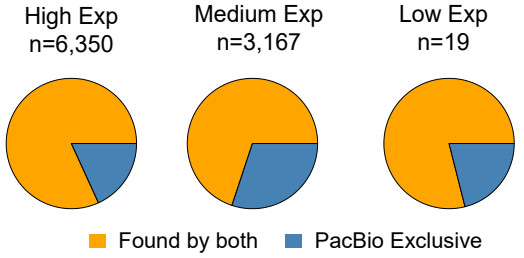
- RefSeq detects lots of lowly expressed genes. However it captures an important fraction of highly expressed ones 19% (PacBio 1.5%)

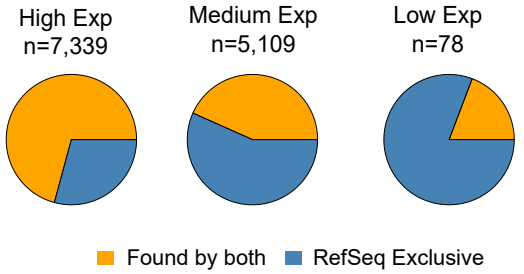
Imposing a higher EXP threshold highlights advantages and disadvantages of PacBio quantification

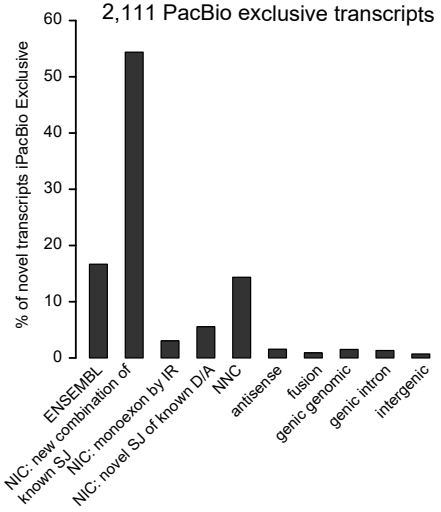


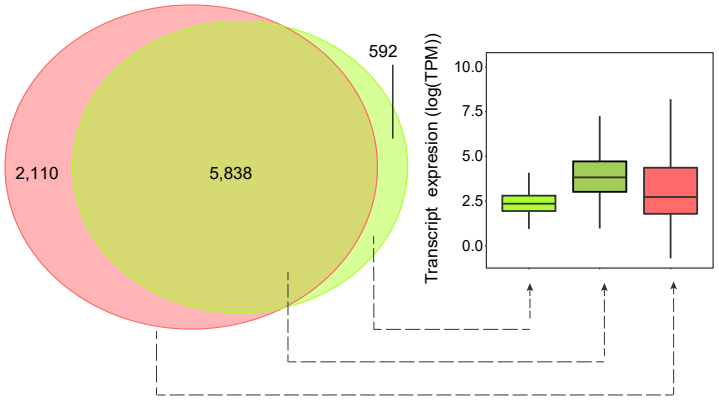


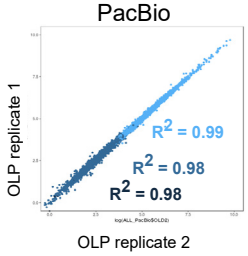


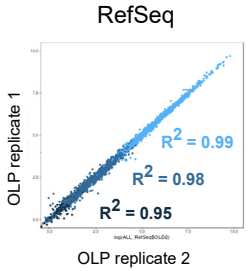


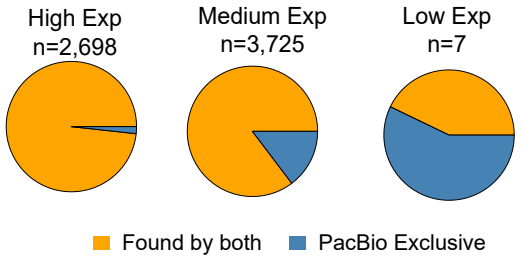


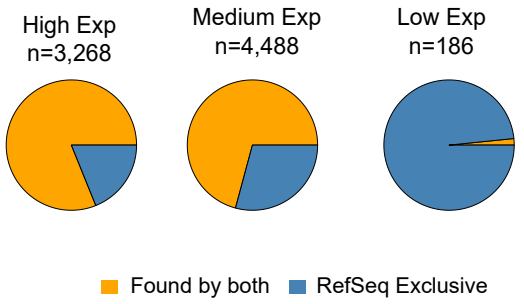


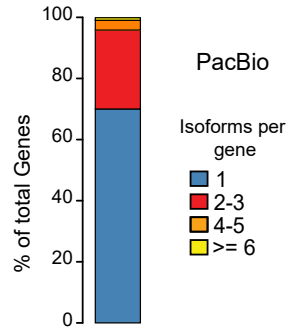


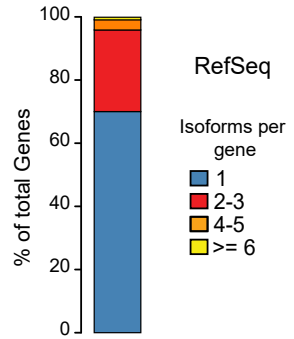




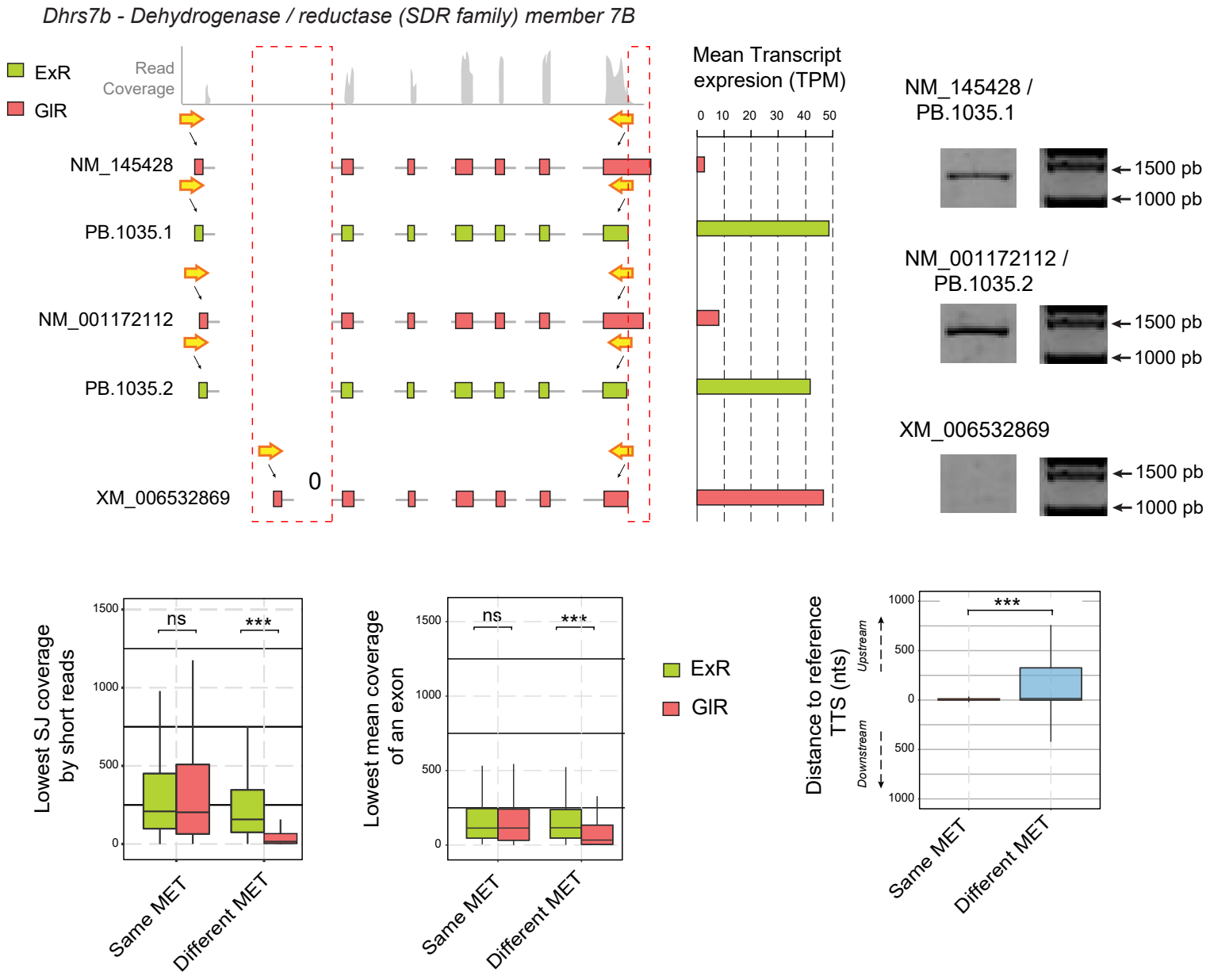








Analysis of Most Expressed Transcript (MET) reveal unaccounted 3' end variability that is captured by PacBio



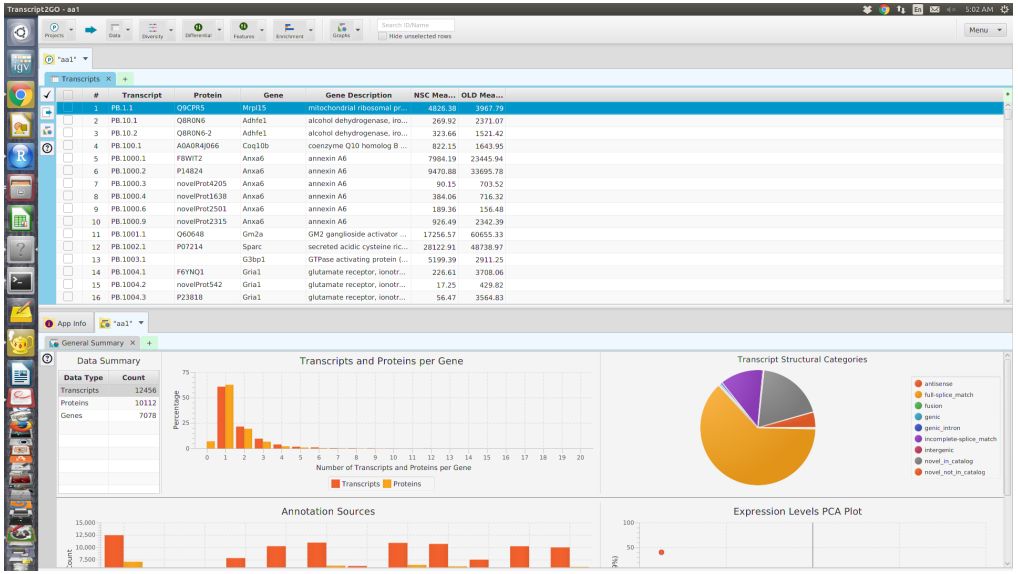
CUANTIFICATION SUMMARY

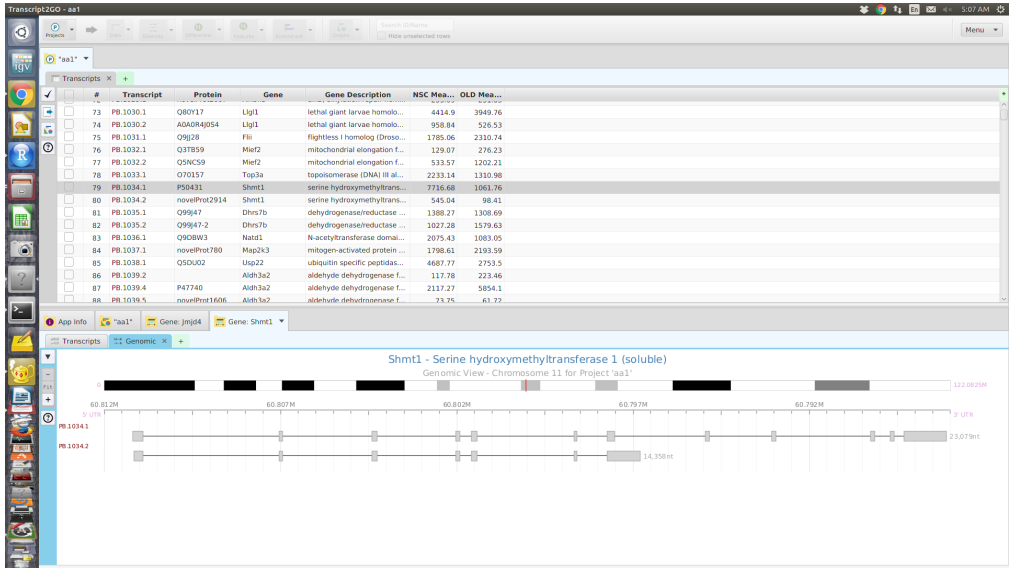
- PacBio captures a robust fraction of transcripts and genes being expressed and at the same time allow for novel discovery.
- RefSeq captures more transcripts and genes, many lowly expressed and hardly reproducible ones
- However, PacBio TOFU fails to capture many highly expressed genes detected by RefSeq.
- PacBio transcriptome reveals unaccounted for 3' end variability in known transcripts that hamper RefSeq quantification.

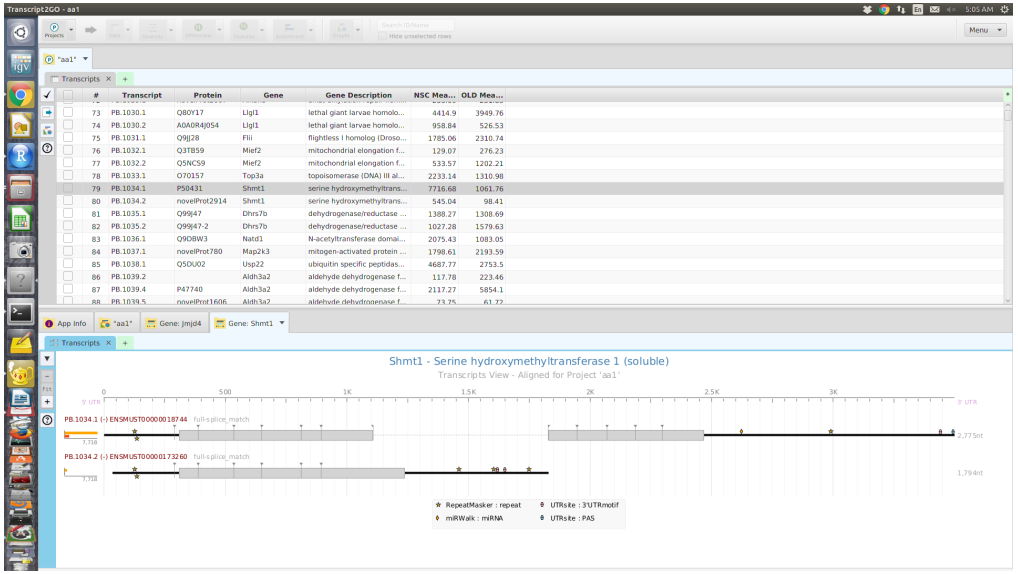
Tardaguila et al ***SQANTI: extensive characterization of long read transcript sequences for quality control in full-length transcriptome identification and quantification*** *Pre-print BioRxiv (2017)*

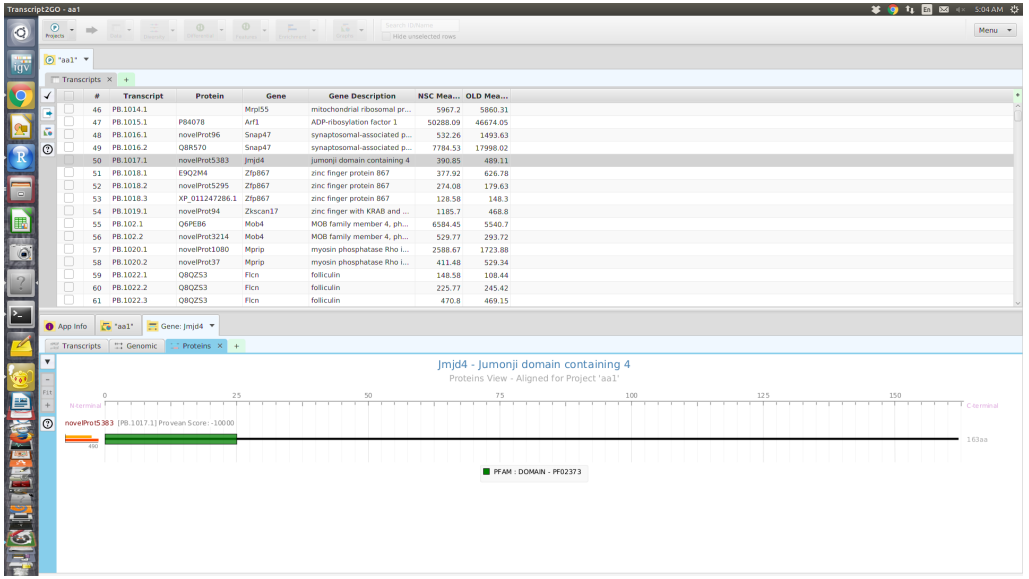
4. Functional outcome of alternative splicing: Transcript2GO

T2GO combines SQANTI classification with Genomic, transcript and protein annotation to maximize analytical possibilities



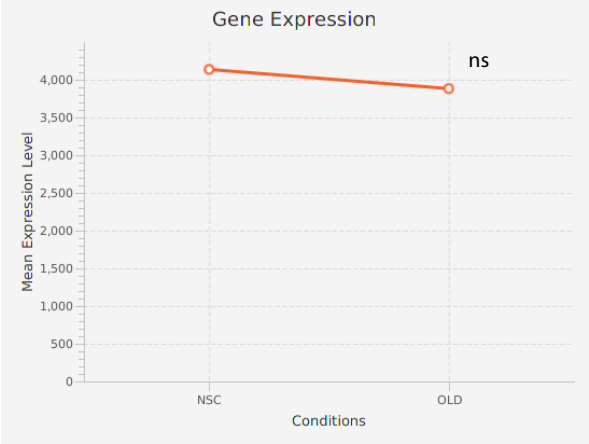


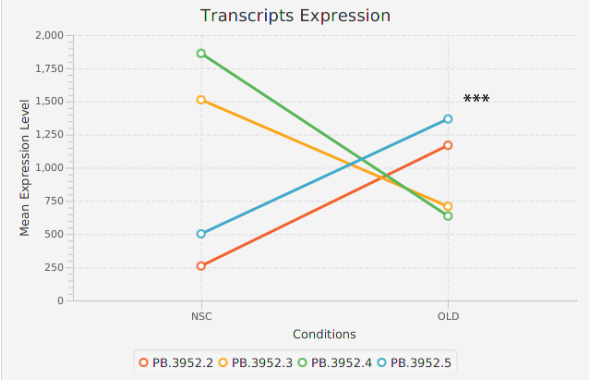


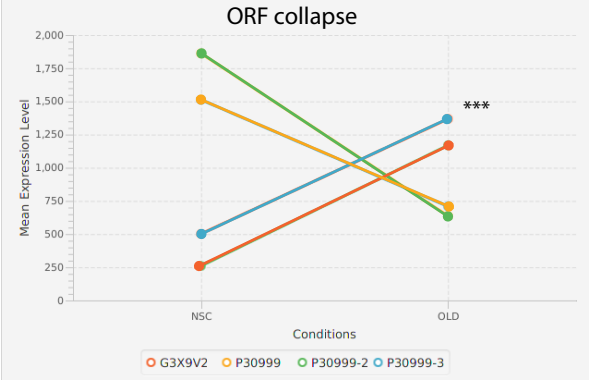


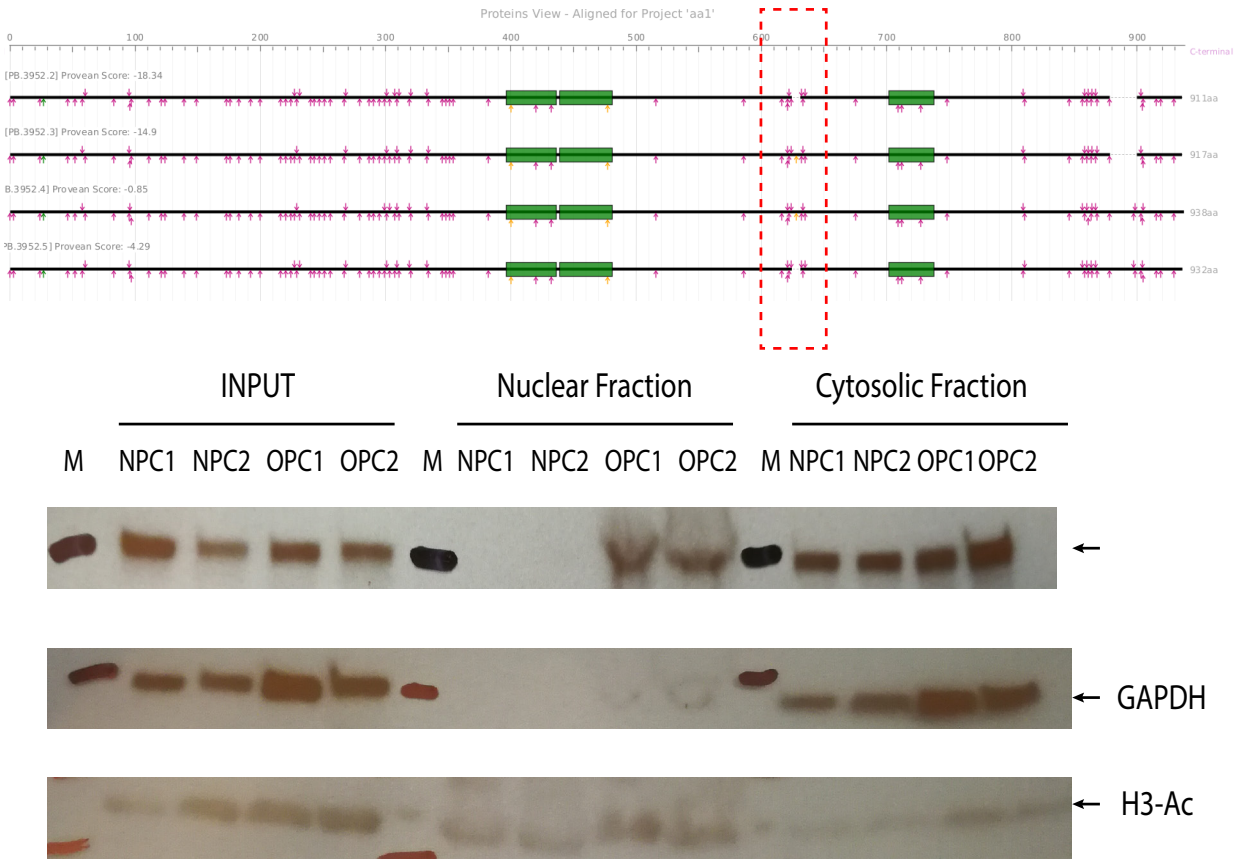
Transcript2GO: assessing the functional outcome of alternative splicing
manuscript in preparation

Diferential splicing linked to the appearance of sequence motifs on a transcriptome wide scale









Transcript2GO: assessing the functional outcome of alternative splicing
manuscript in preparation

Thanks!

UF
William Farmerie
Eric Triplett
Lauren McIntyre

UCI
Ali Mortazavi

Pacbio
Liz Tseng

CIPF
Victoria Moreno
Susana Rodriguez