



# *SQANTI: Classification, Curation and Quantification of a PacBio transcriptome*

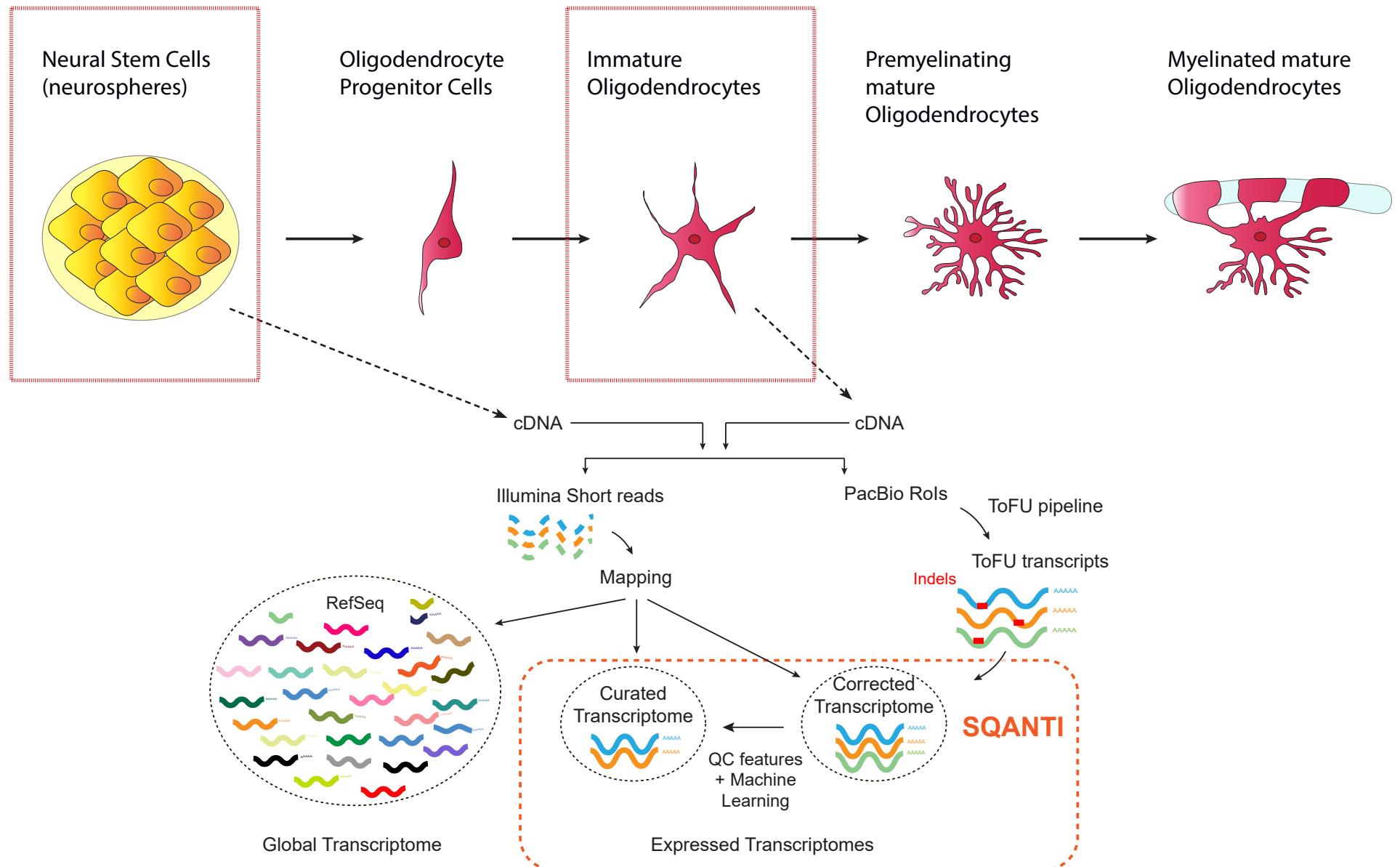
Manuel Tardaguila, PhD

Ana Conesa Lab

**UF** | UNIVERSITY of  
FLORIDA

  
PRINCIPE FELIPE  
CENTRO DE INVESTIGACION

# PacBio in Oligodendrogenesis



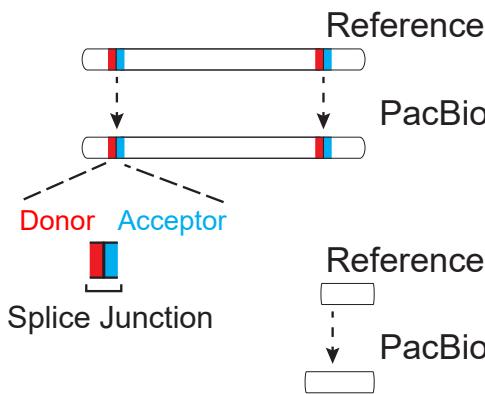
Tardaguila et al **SQANTI**: extensive characterization of long read transcript sequences for quality control in full-length transcriptome identification and quantification Pre-print BioRxiv (2017)



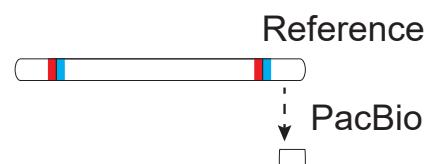
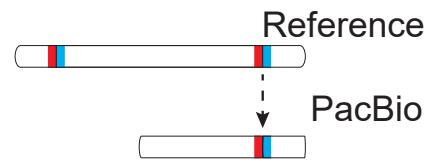
## 1. CLASSIFICATION OF PACBIO TRANSCRIPTS

# Splice-based classification (I): Known Transcripts

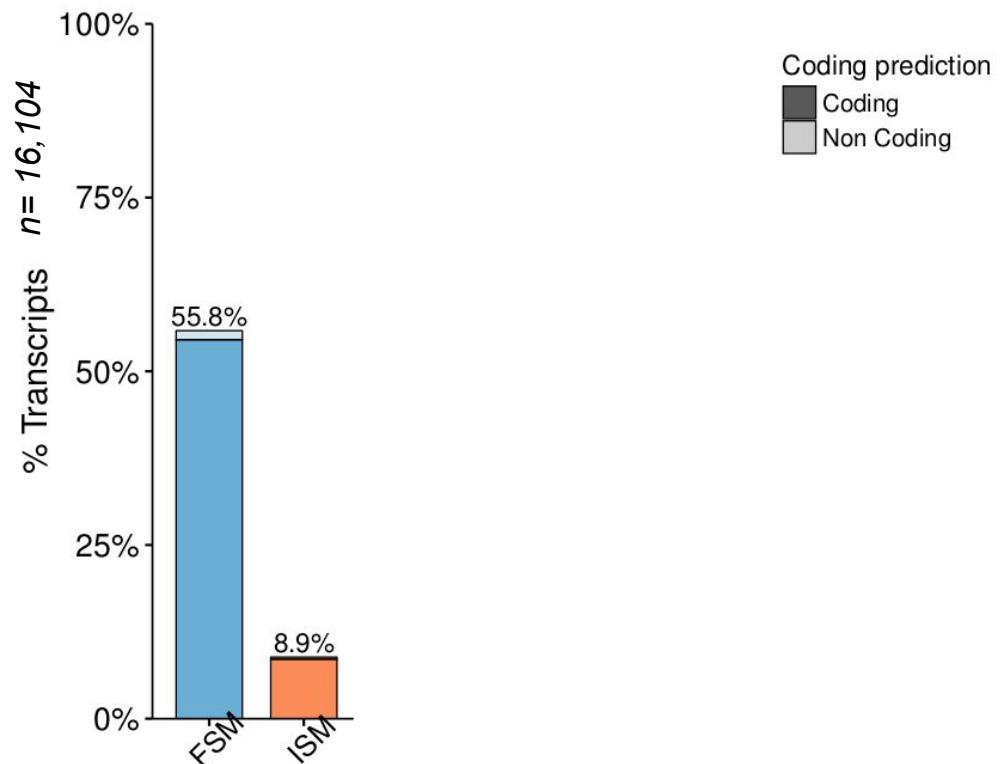
## Full Splice Match (FSM)



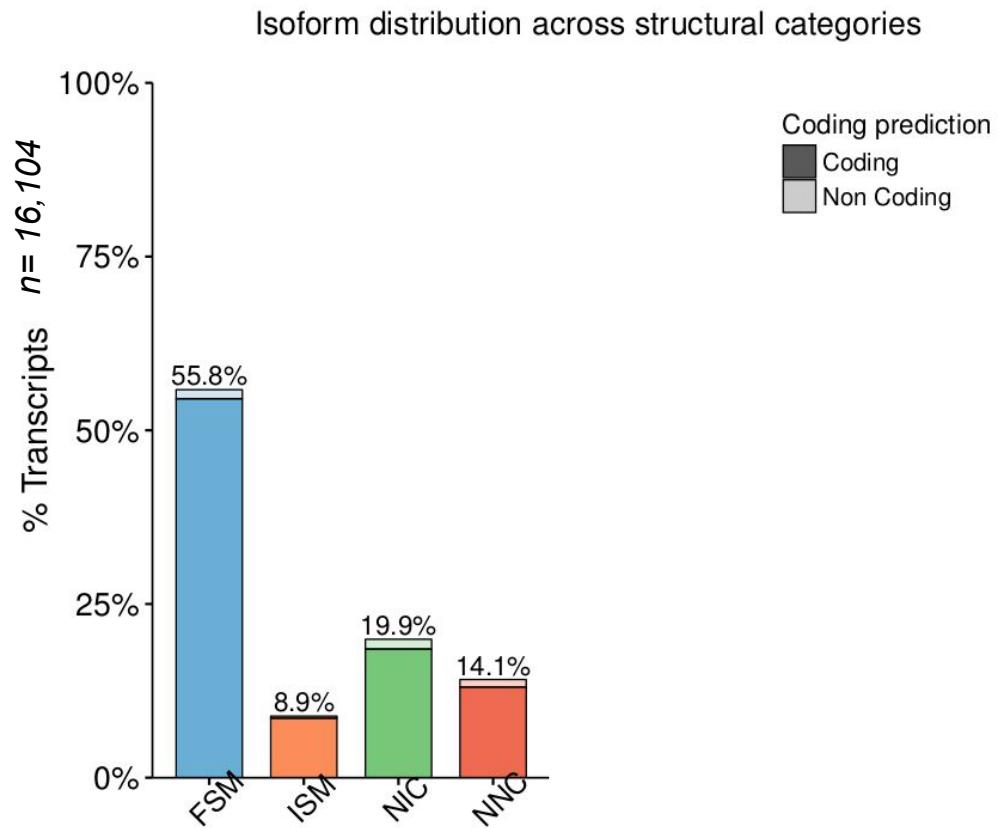
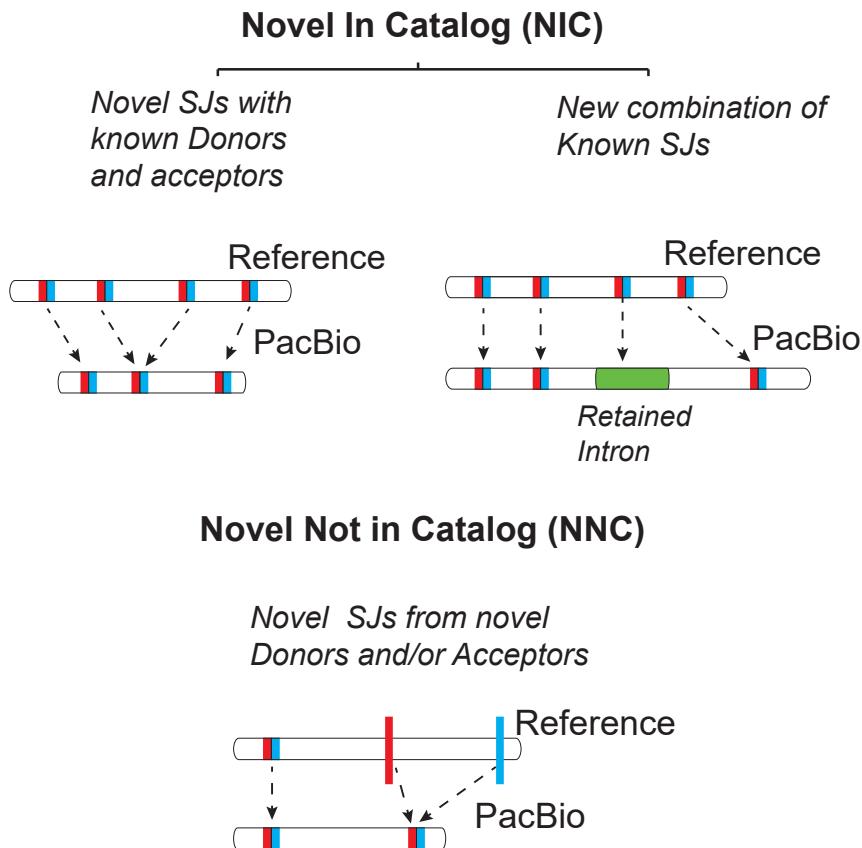
## Incomplete Splice Match (ISM)



Isoform distribution across structural categories

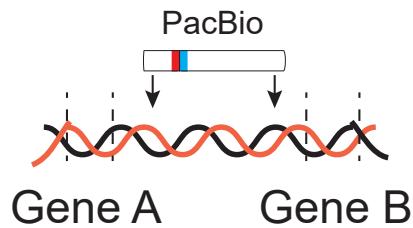


# Splice-based classification (II): Novel Transcripts from known genes

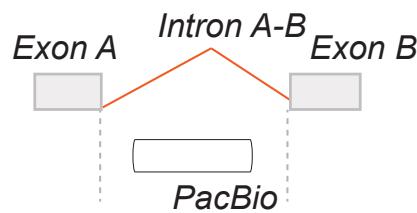


# Splice-based classification (III): Novel Transcripts from novel genes

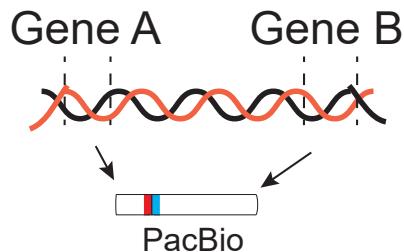
## Intergenic



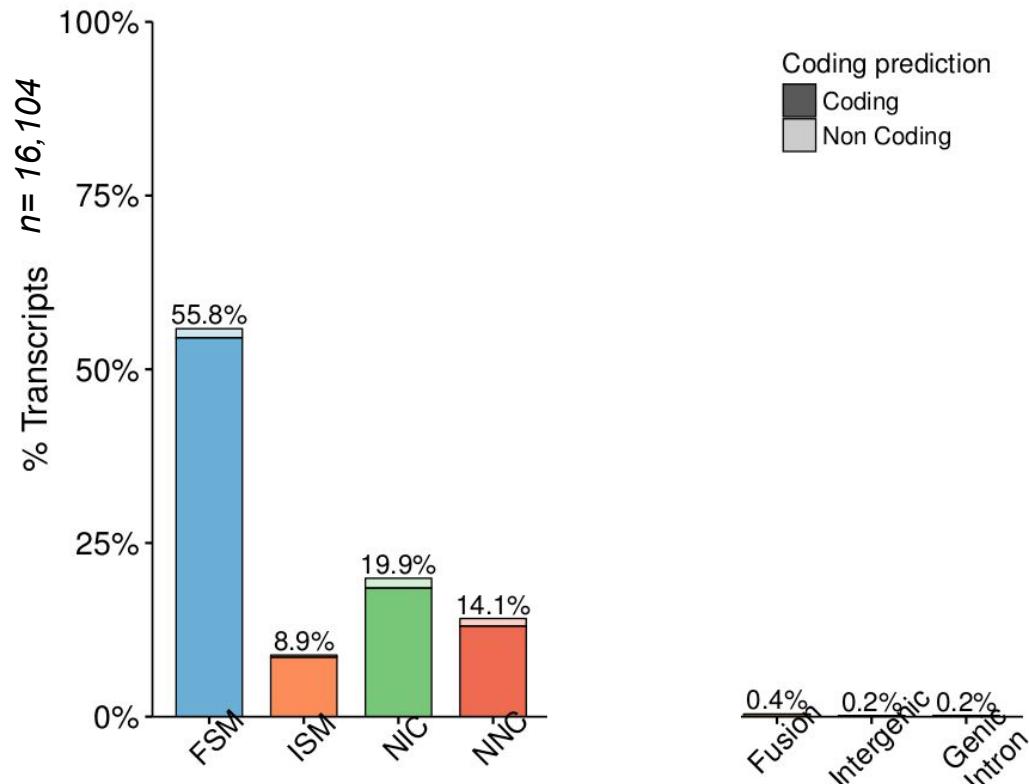
## Genic Intron



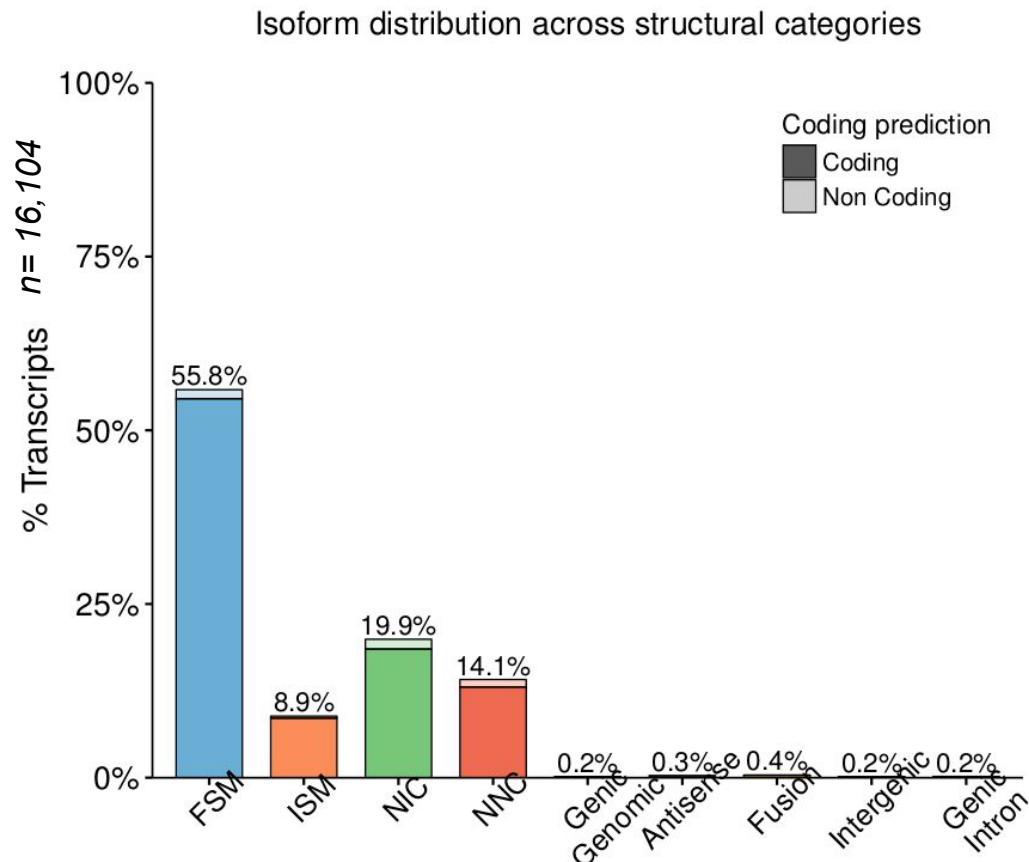
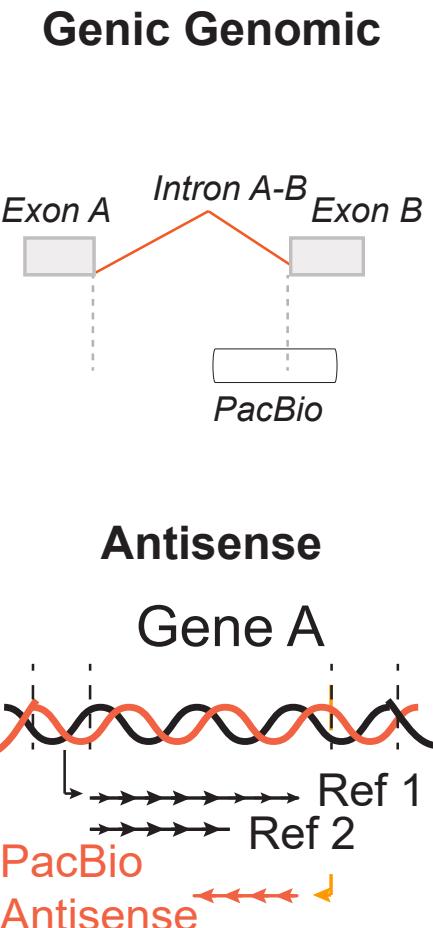
## Fusion Transcripts



Isoform distribution across structural categories



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



# CLASSIFICATION SUMMARY



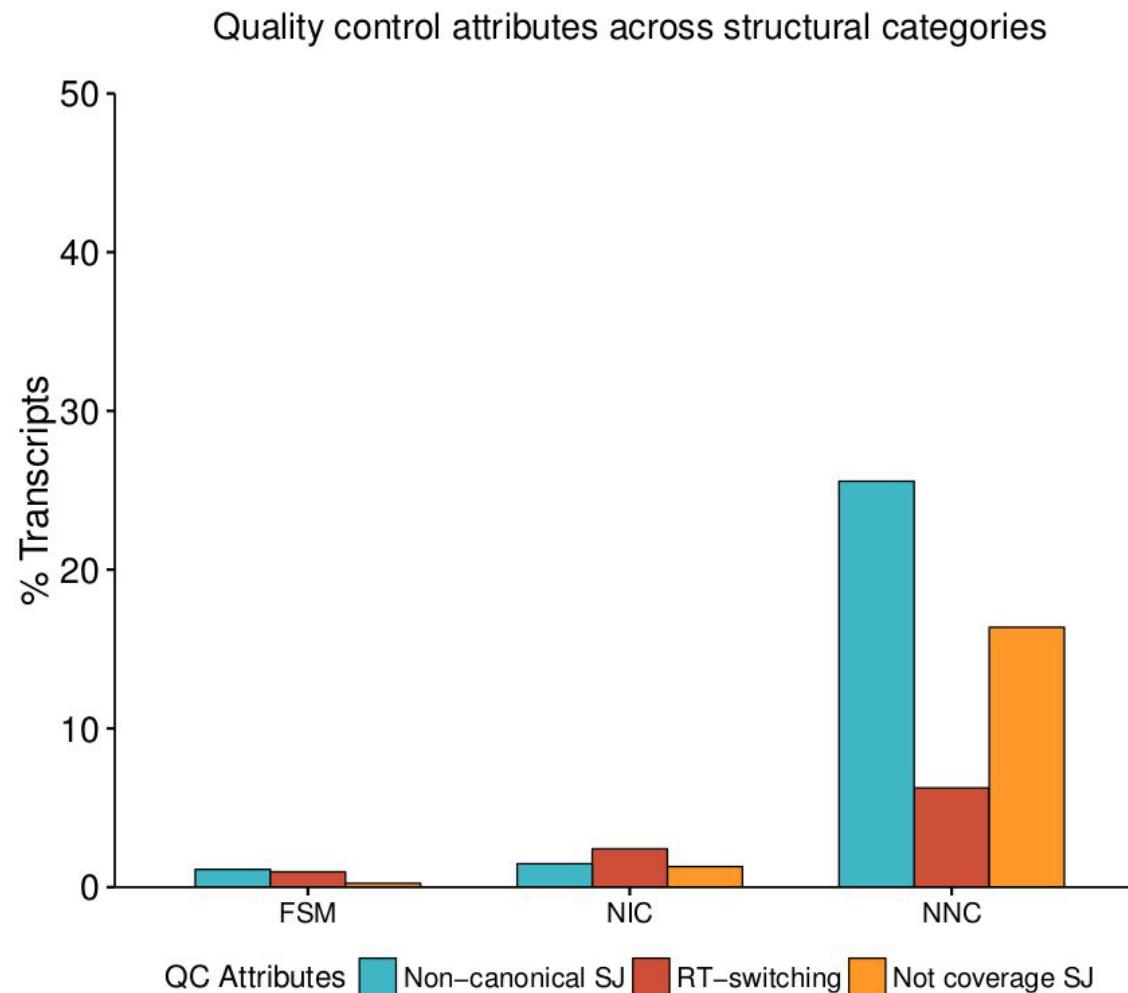
- Splice based classification allows to separate known transcripts (FSM and ISM) from novel transcripts arising from novel splice junctions or new combinations of already known splice junctions (NIC and NNC)
- The predominant categories in our transcriptome are FSM,ISM, NIC and NNC and they are mostly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



## 2. CURATION OF PACBIO TRANSCRIPTS

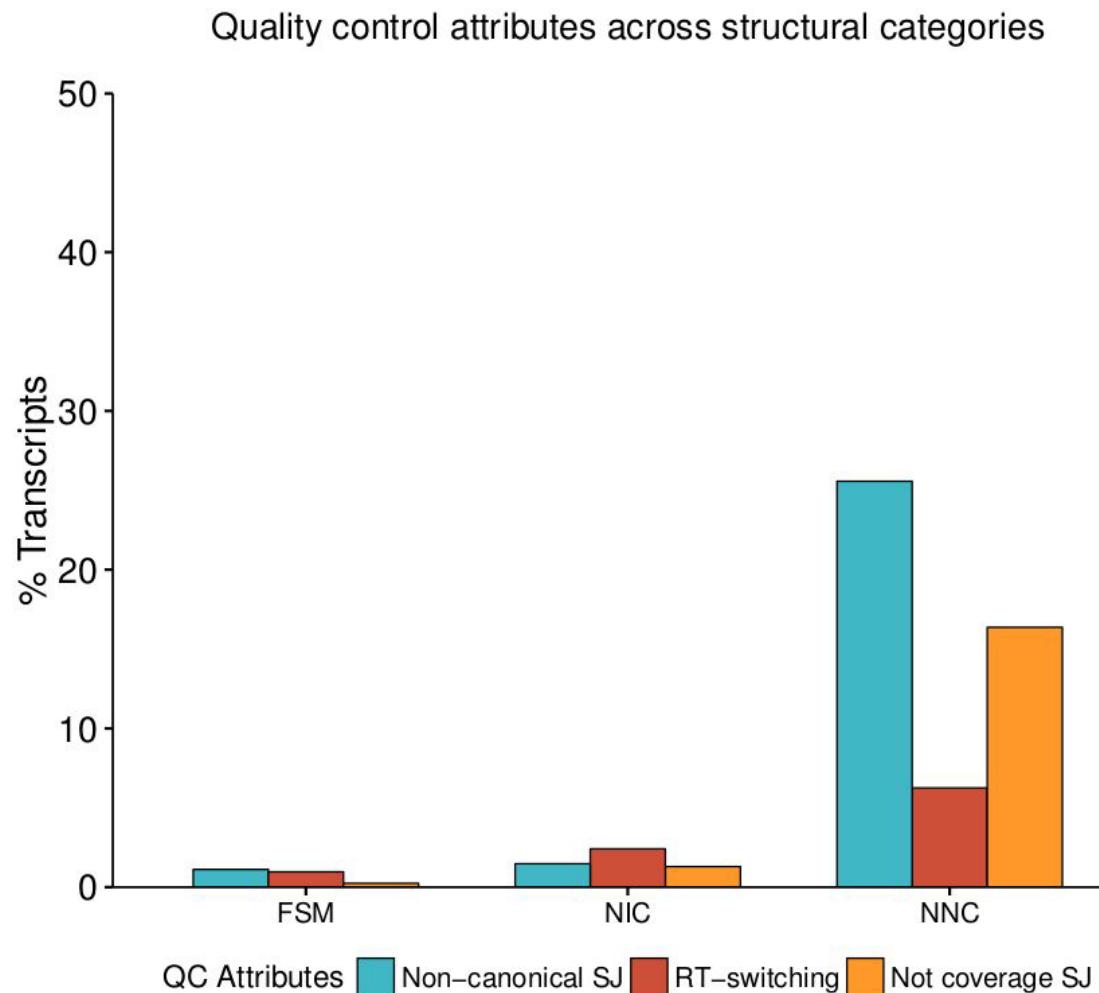
# 40% novel isoforms in mouse... Are all of them real?

---



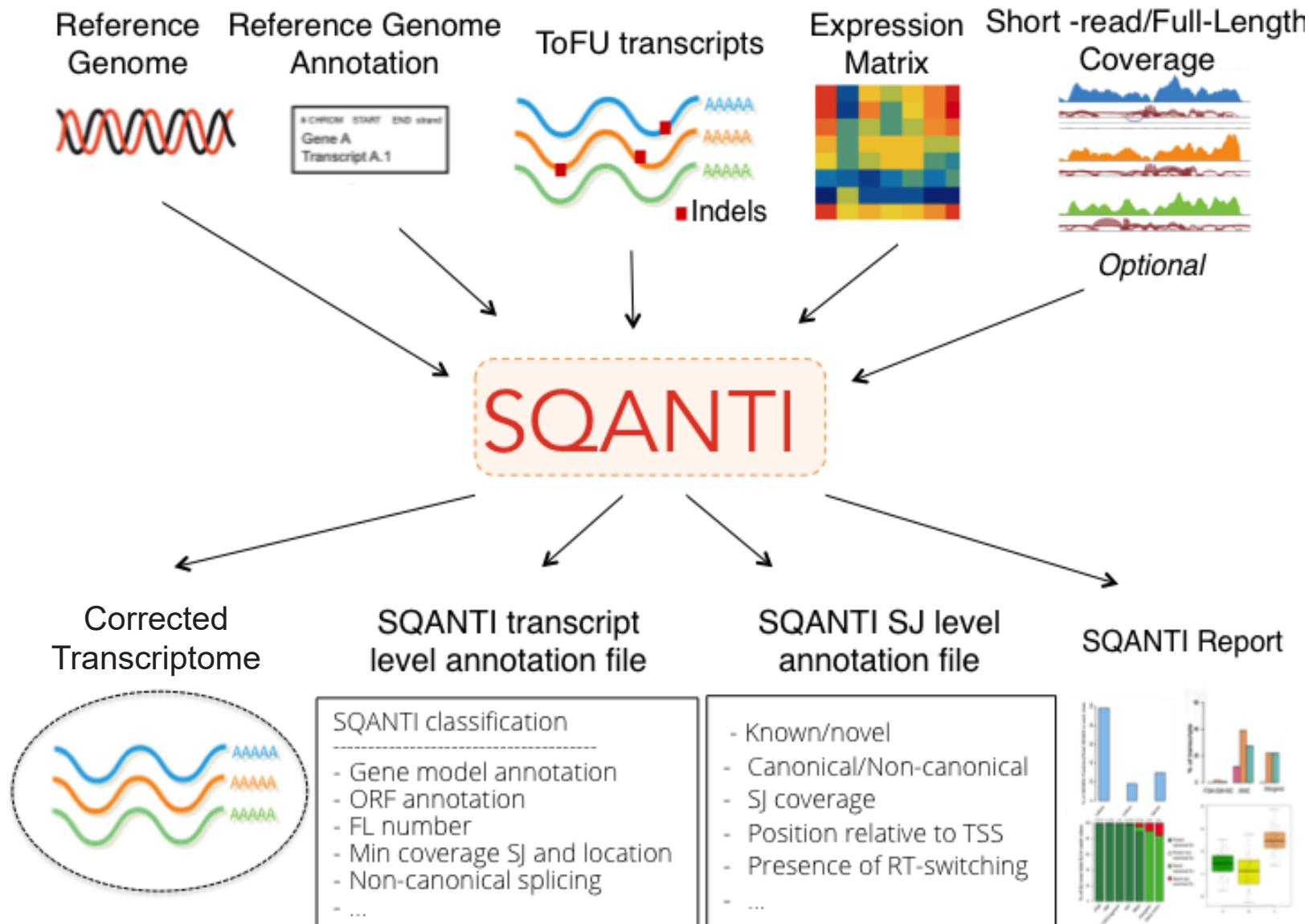
# PacBio output needs curation

---

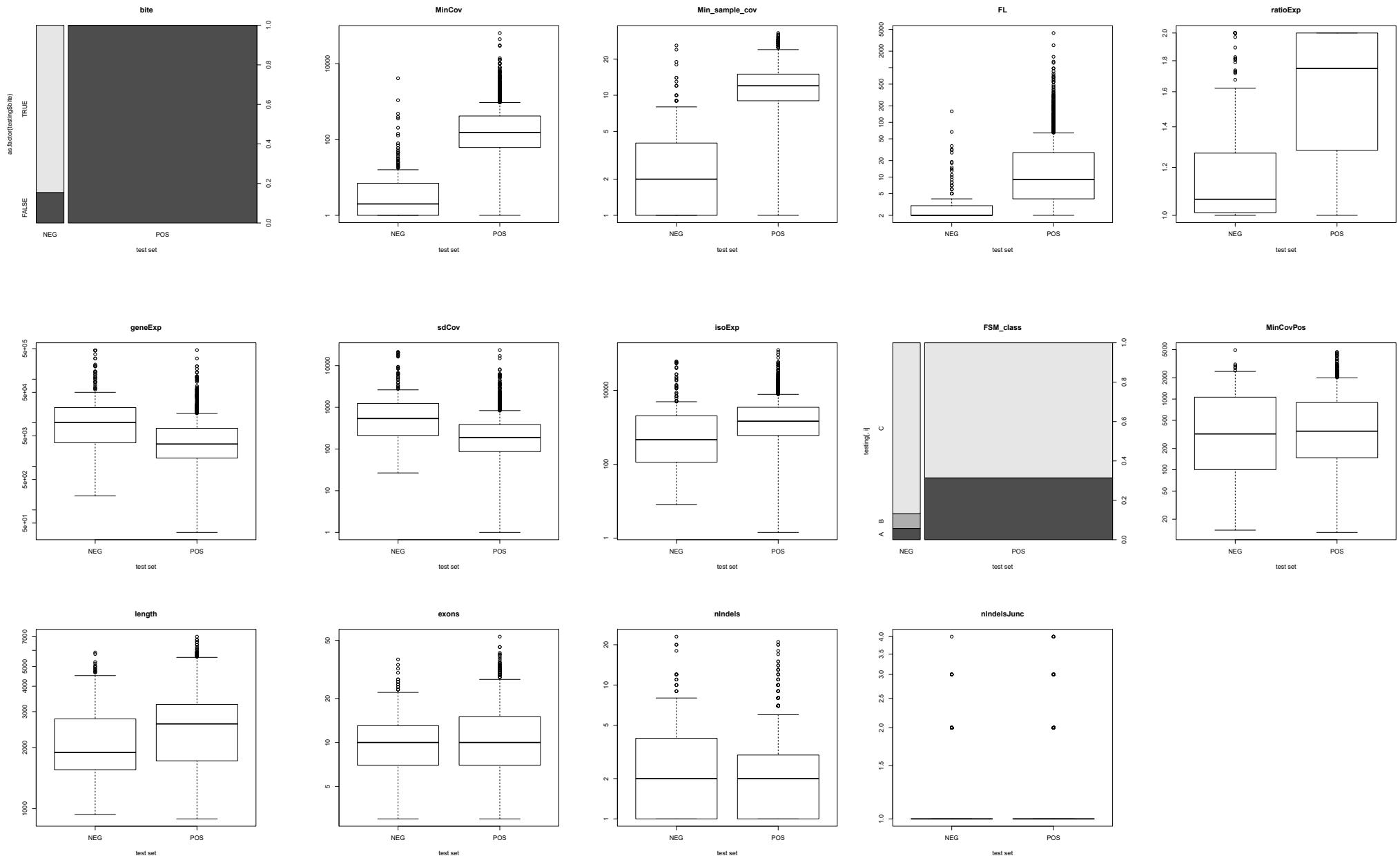


- NNC transcripts characterized by novel Donors and/or Acceptors of splicing concentrate traits of low quality transcripts

# Using Short reads and Long reads to mine QC features



# Features selected for Random Forest Classification

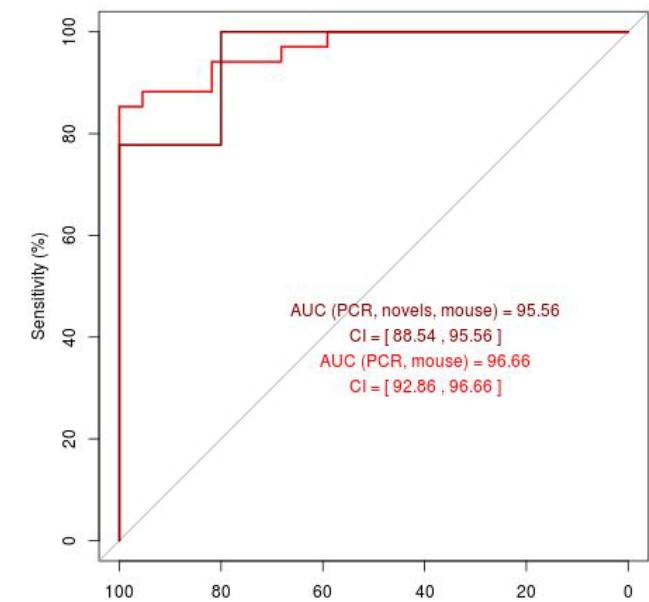
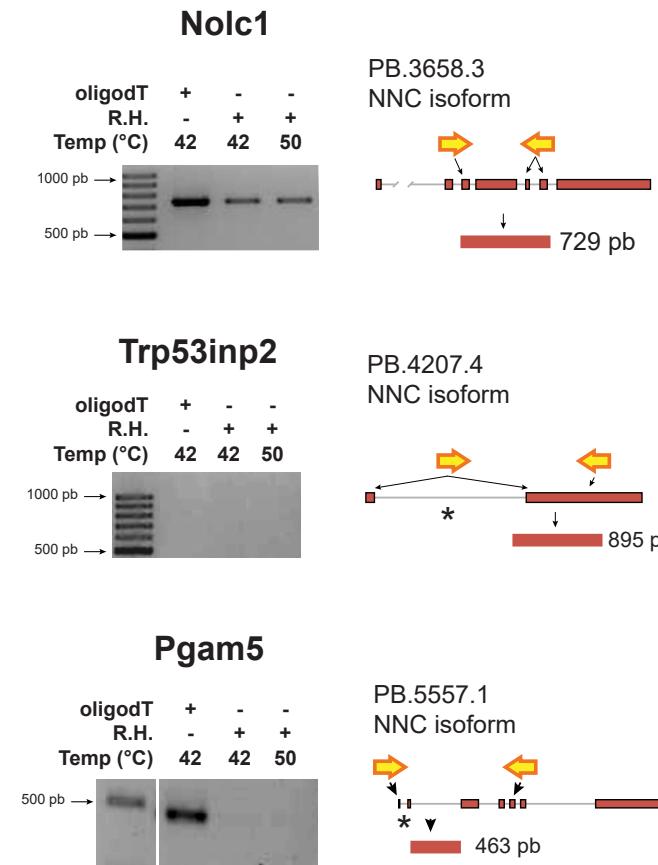


# PCR validation in an independent set of transcripts

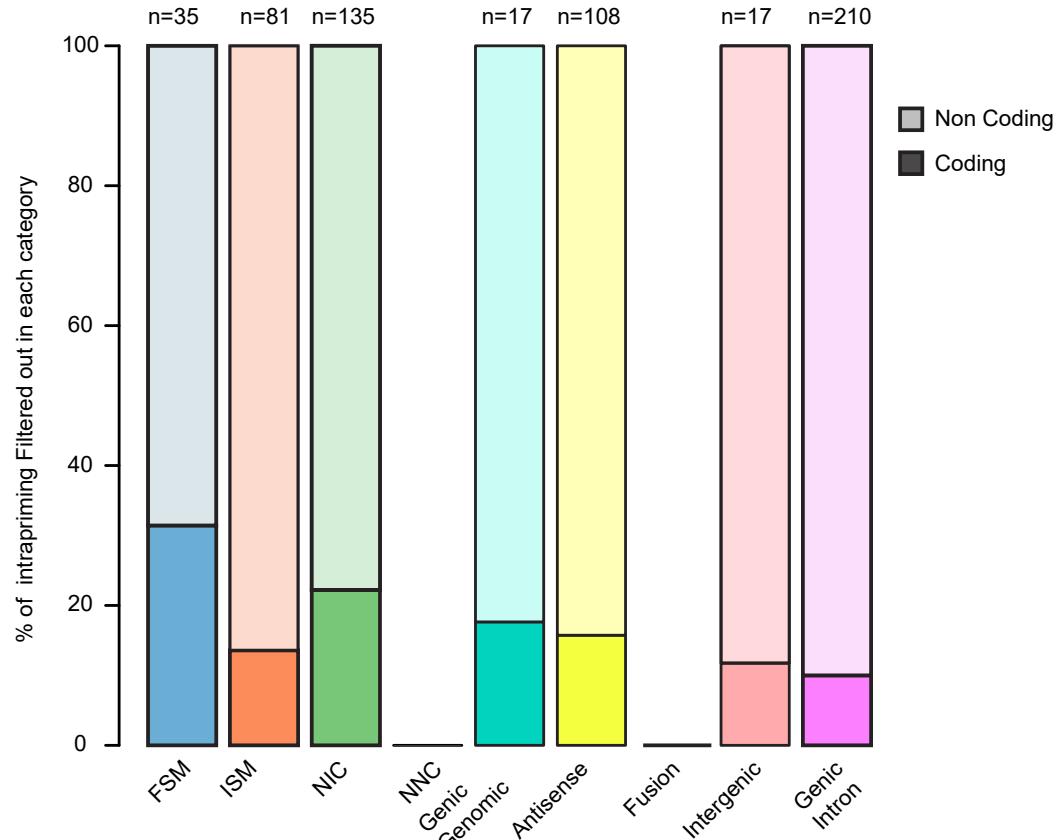
oligo (dT) PCR

	FSM	NIC	NNC	Fusion
Positive	23 (3nc)	7	8 (3 nc)	1
Negative	0	2	12 (8 nc)	2
Total	23	9	20	3
Positive	5 (3nc)	7	2	1
Negative	0	0	6 (3nc)	0
Total	5	7	8	1

R.H. PCR

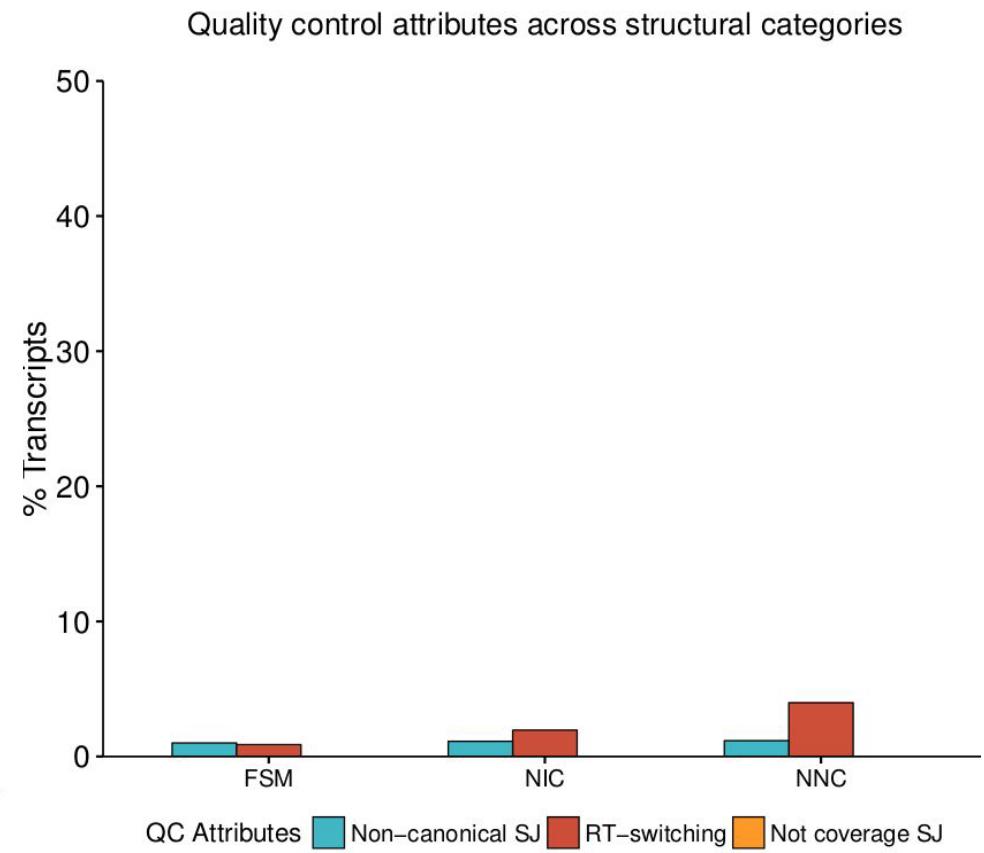
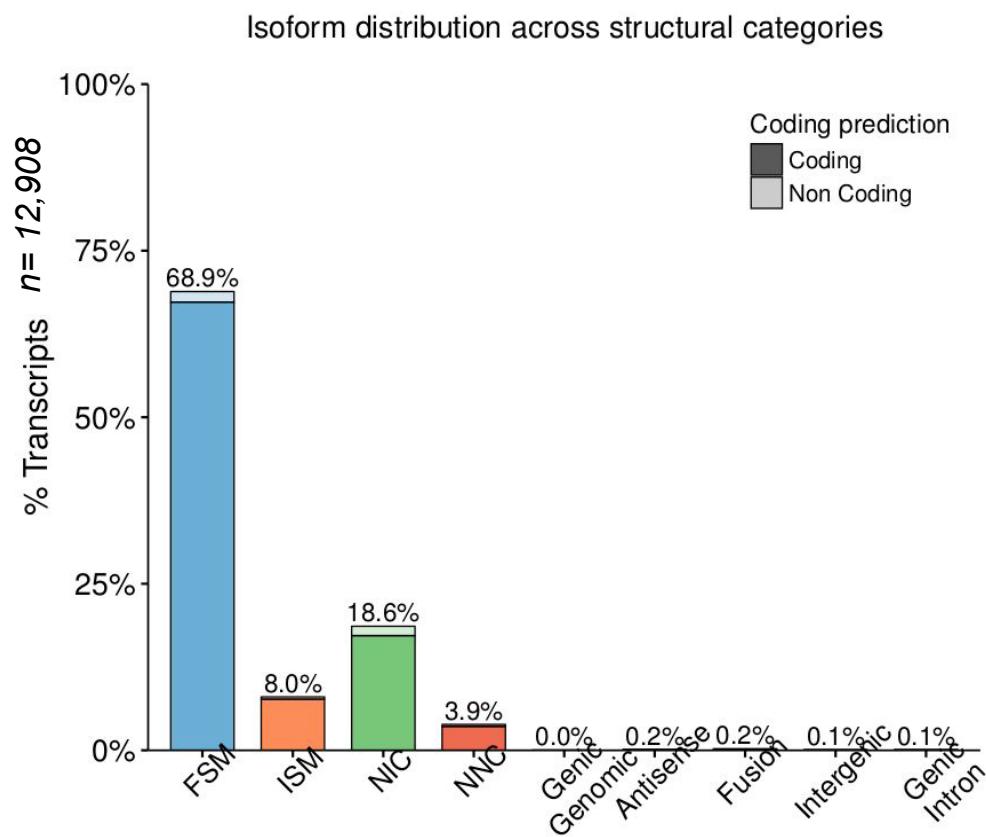


# Splice Junction independent curation: Intraprimer features



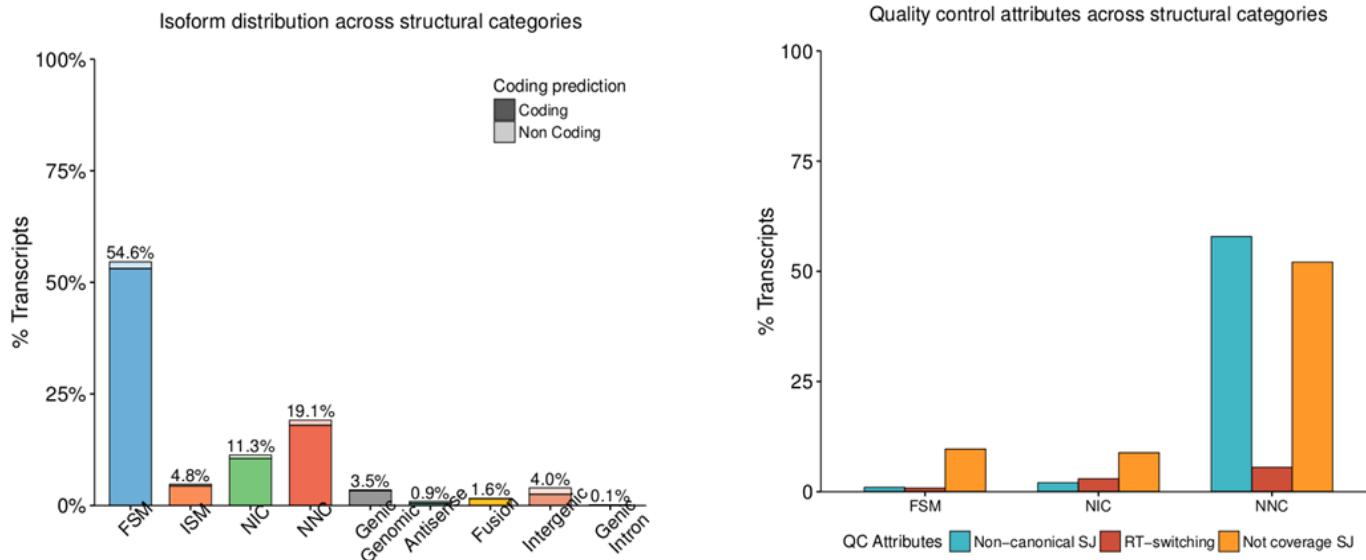
- oligodT can prime outside polyA regions in A rich regions inside transcripts
  - We looked for transcripts showing  $\geq 80\%$  Adenines in the 20 nts downstream (DNA) the end of the transcript
    - If this transcripts lack a consensus poly Adenilation site we filter them out
    - 605 transcripts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

# SQANTI filter eliminates transcript with bad QC features

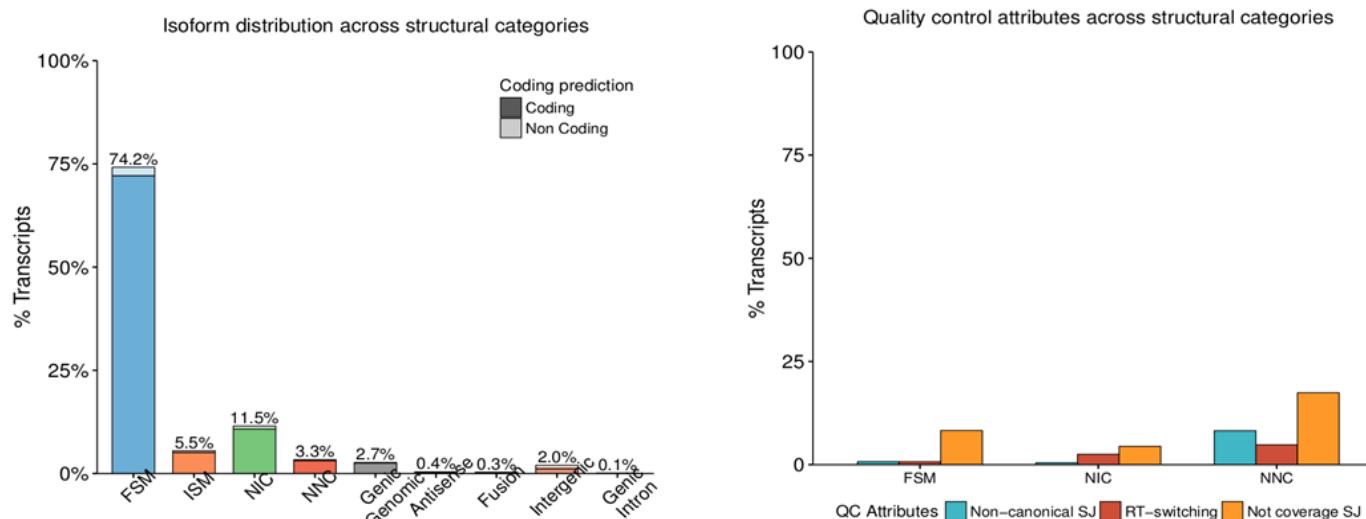


# SQANTI filter works across different PacBio datasets: Maize

- Before SQANTI

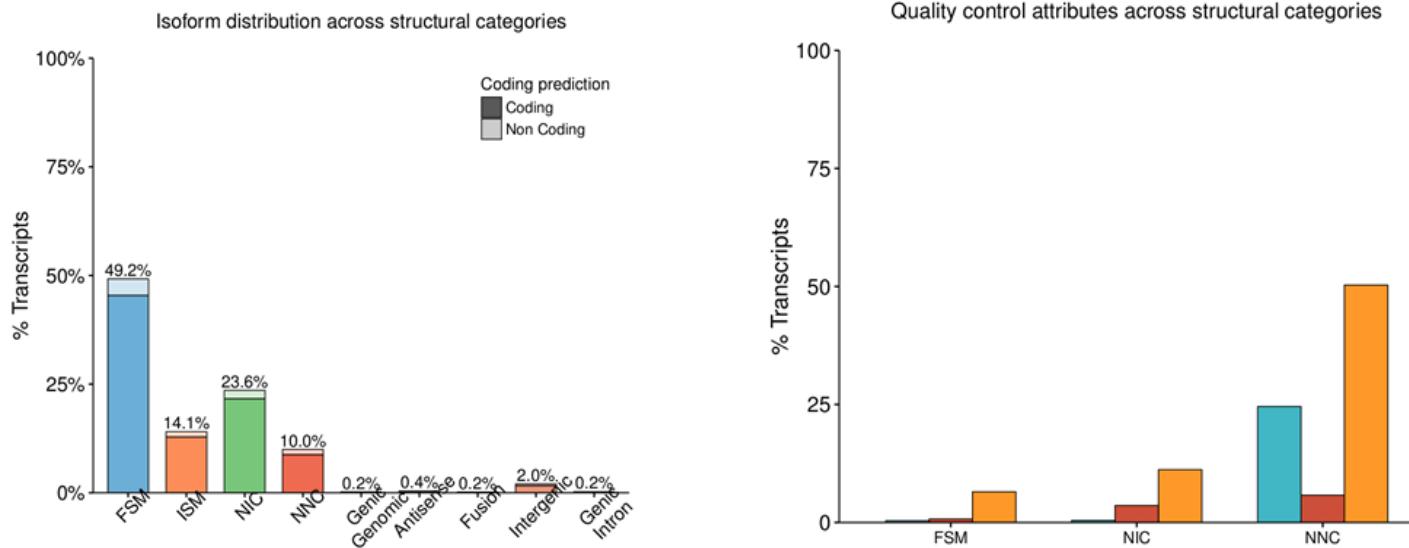


- AFTER SQANTI

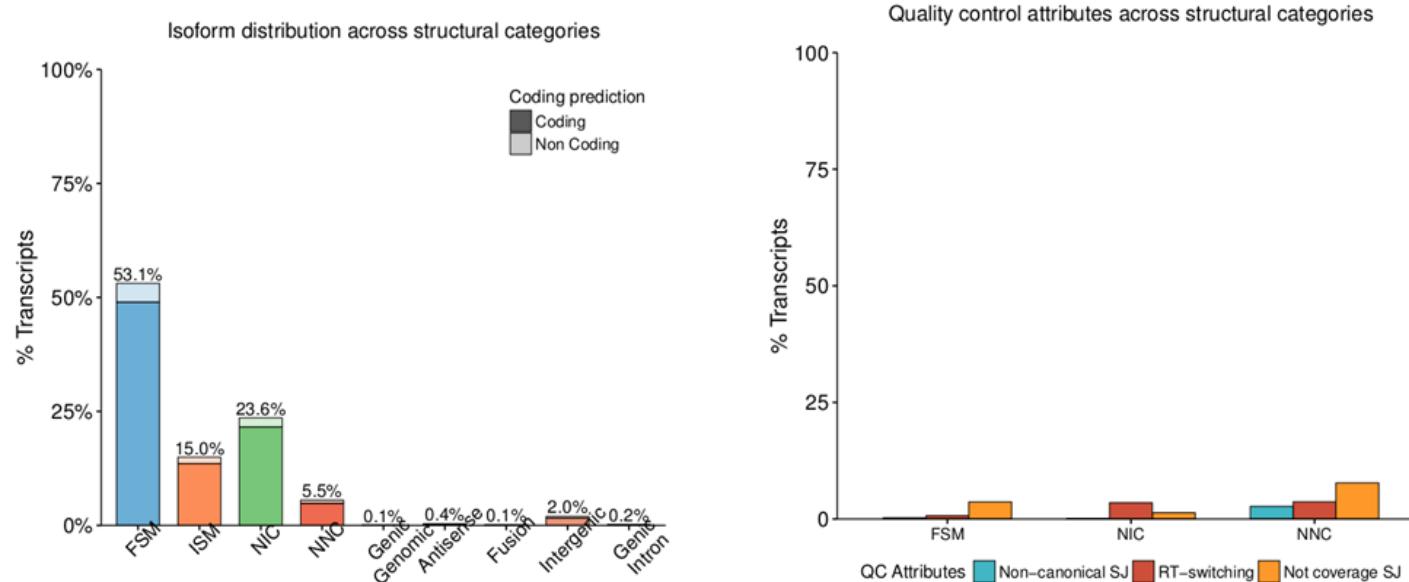


# SQANTI filter works across different PacBio datasets: MCF-7

- Before SQANTI



- AFTER SQANTI



## CURATION SUMMARY

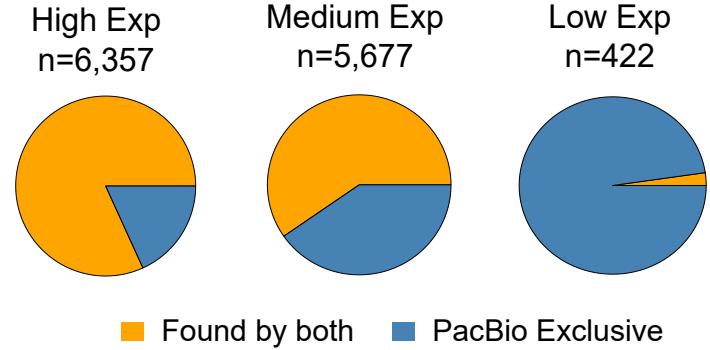
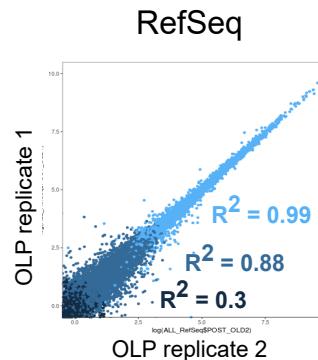
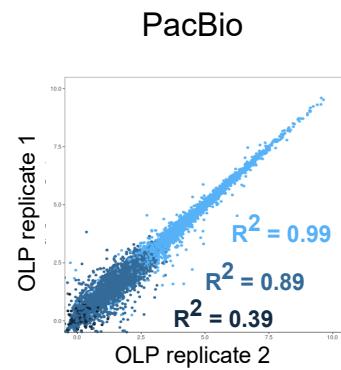
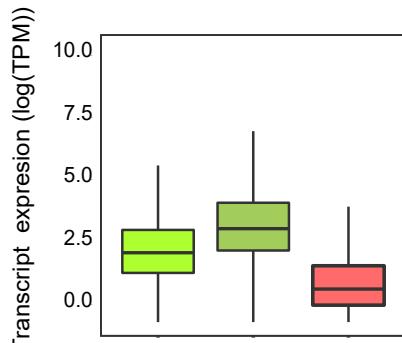
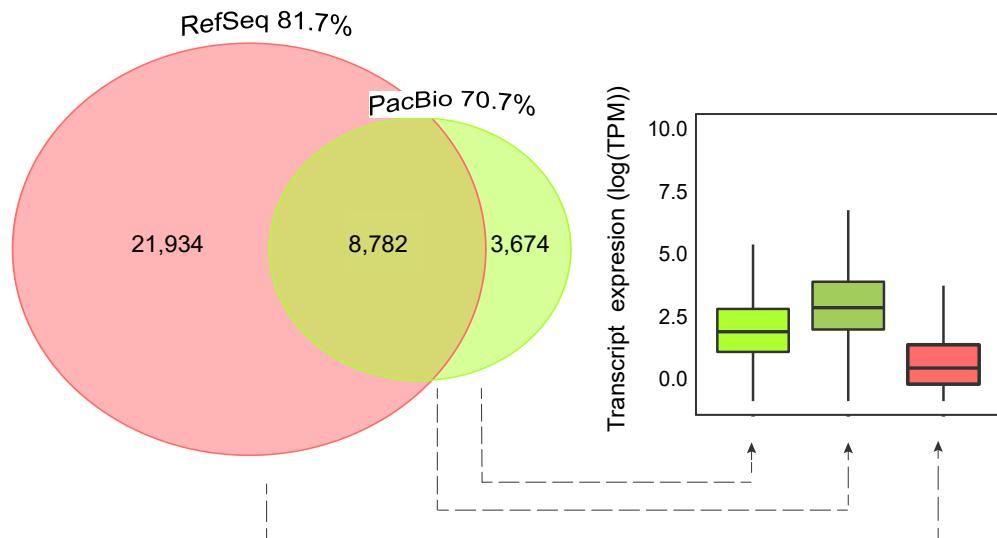


- PacBio output needs curation, specially in NNC novel acceptor/donor transcripts
- SQANTI mines features based in splice junctions, expression and structure to feed a RF classifier that succeeds in filtering out transcripts that fail to validate in an independent PCR set
- SQANTI works across different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

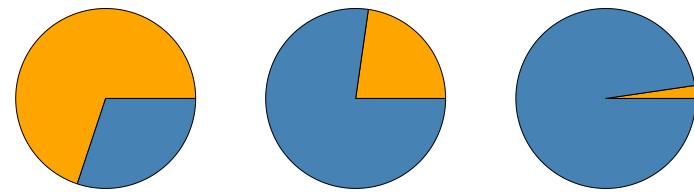


### 3. QUANTIFICATION OF PACBIO TRANSCRIPTOME

# Quantification: PacBio curated vs RefSeq: Transcripts



High Exp n=6,357  
Medium Exp n=5,677  
Low Exp n=422

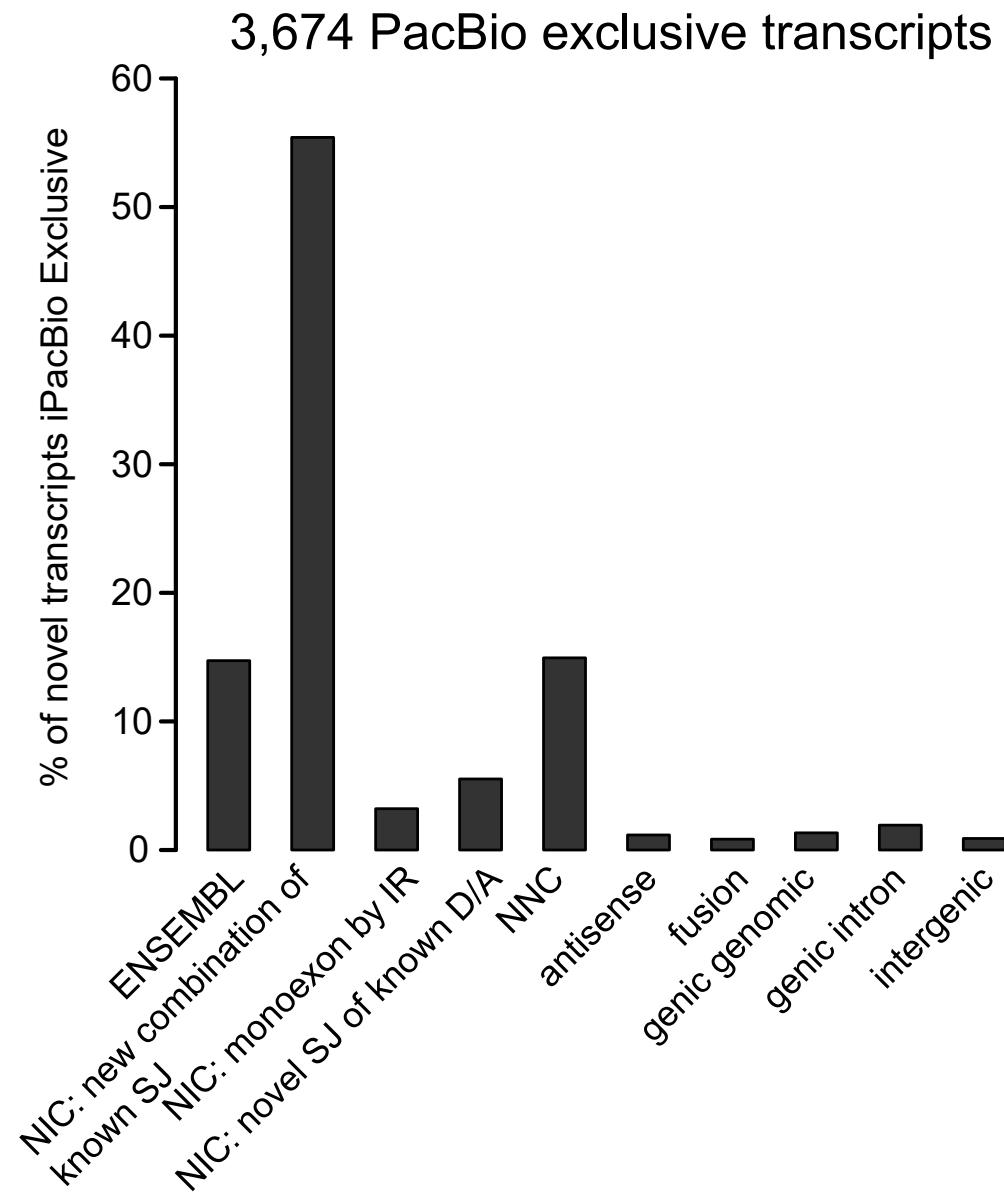


High Exp n=7,440  
Medium Exp n=7,440  
Low Exp n=8,379

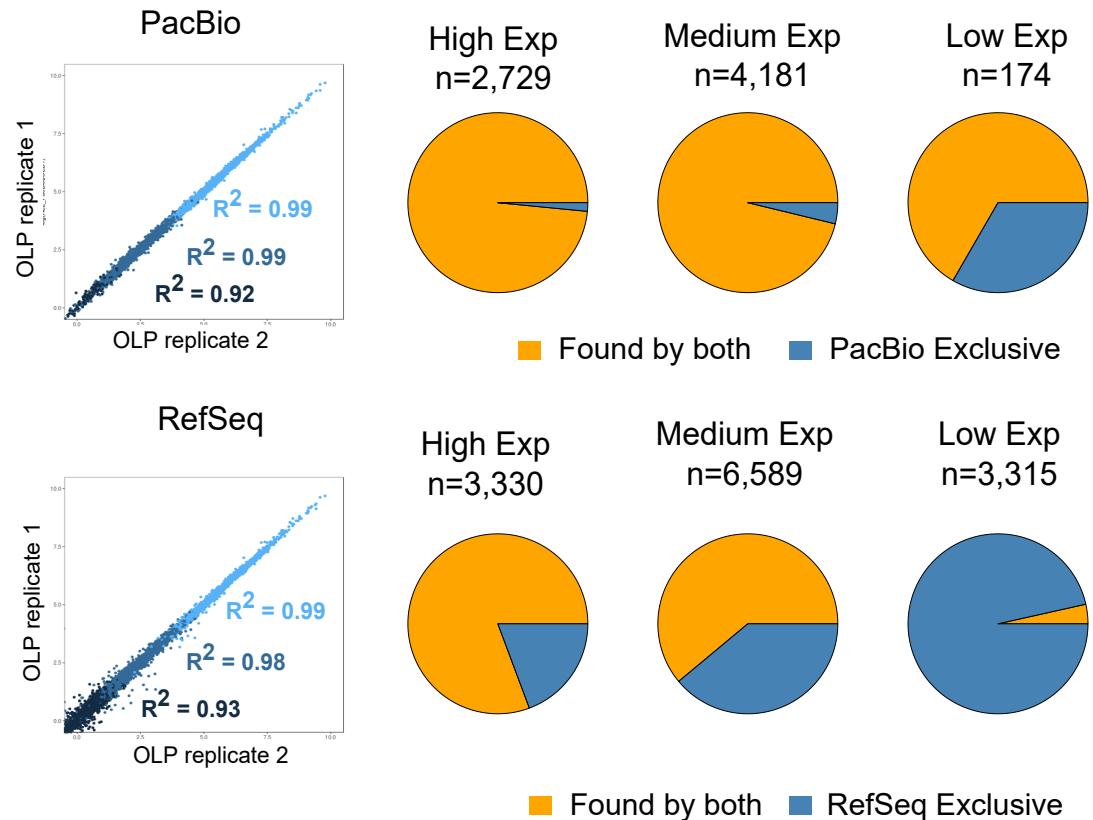
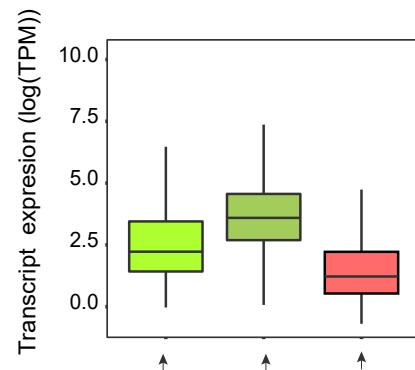
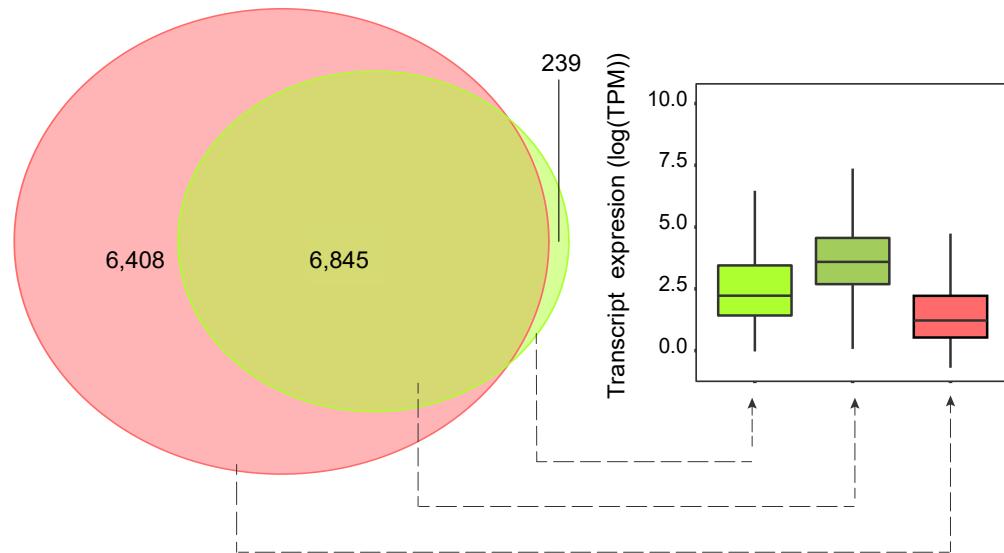
- PacBio finds a set of highly expressed transcripts and reproducible transcripts in common with RefSeq
- RefSeq detects lots of lowly expressed, difficult to reproduce transcripts. However it also captures exclusively 30% of highly expressed ones (PacBio 18%).

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

---

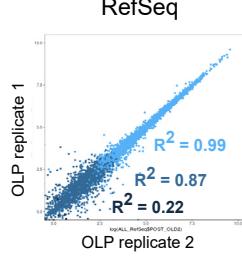
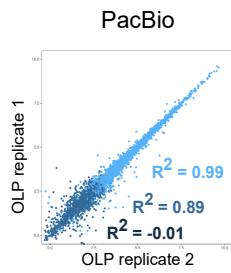
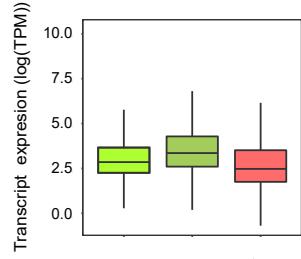
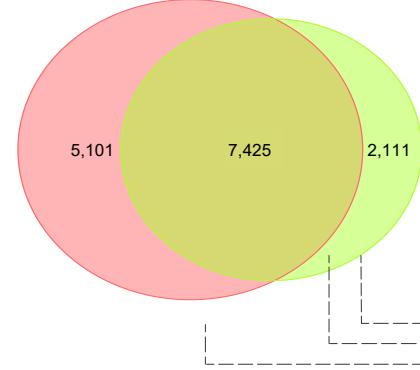


# Quantification: PacBio curated vs RefSeq: Genes



- 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



High Exp  
n=6,350

High Exp  
n=7,339

High Exp  
n=2,698

Medium Exp  
n=3,167

Medium Exp  
n=5,109

Medium Exp  
n=3,725

Low Exp  
n=19

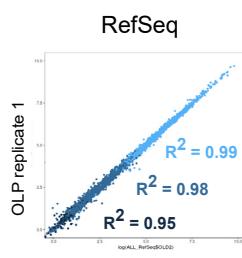
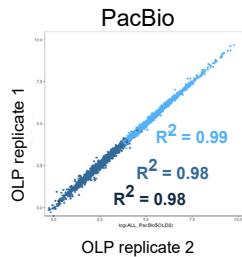
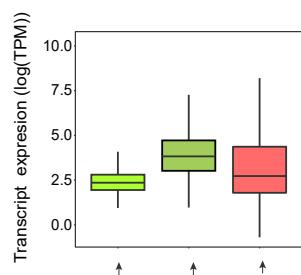
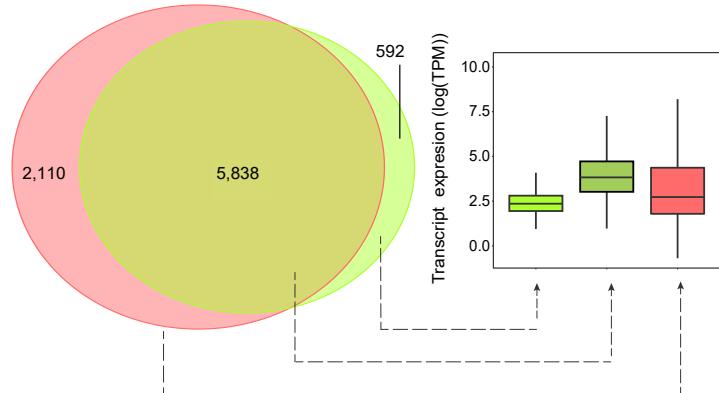
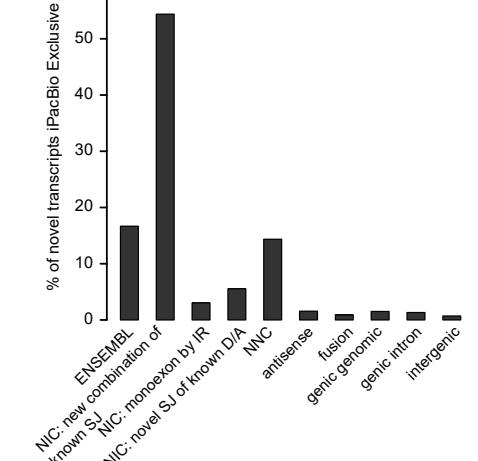
Low Exp  
n=78

Legend: Found by both (orange), PacBio Exclusive (blue)

Legend: Found by both (orange), RefSeq Exclusive (blue)

Legend: Found by both (orange), PacBio Exclusive (blue)

2,111 PacBio exclusive transcripts



High Exp  
n=3,268

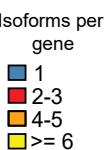
Medium Exp  
n=4,488

Low Exp  
n=186

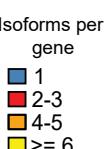
Legend: Found by both (orange), PacBio Exclusive (blue)

Legend: Found by both (orange), RefSeq Exclusive (blue)

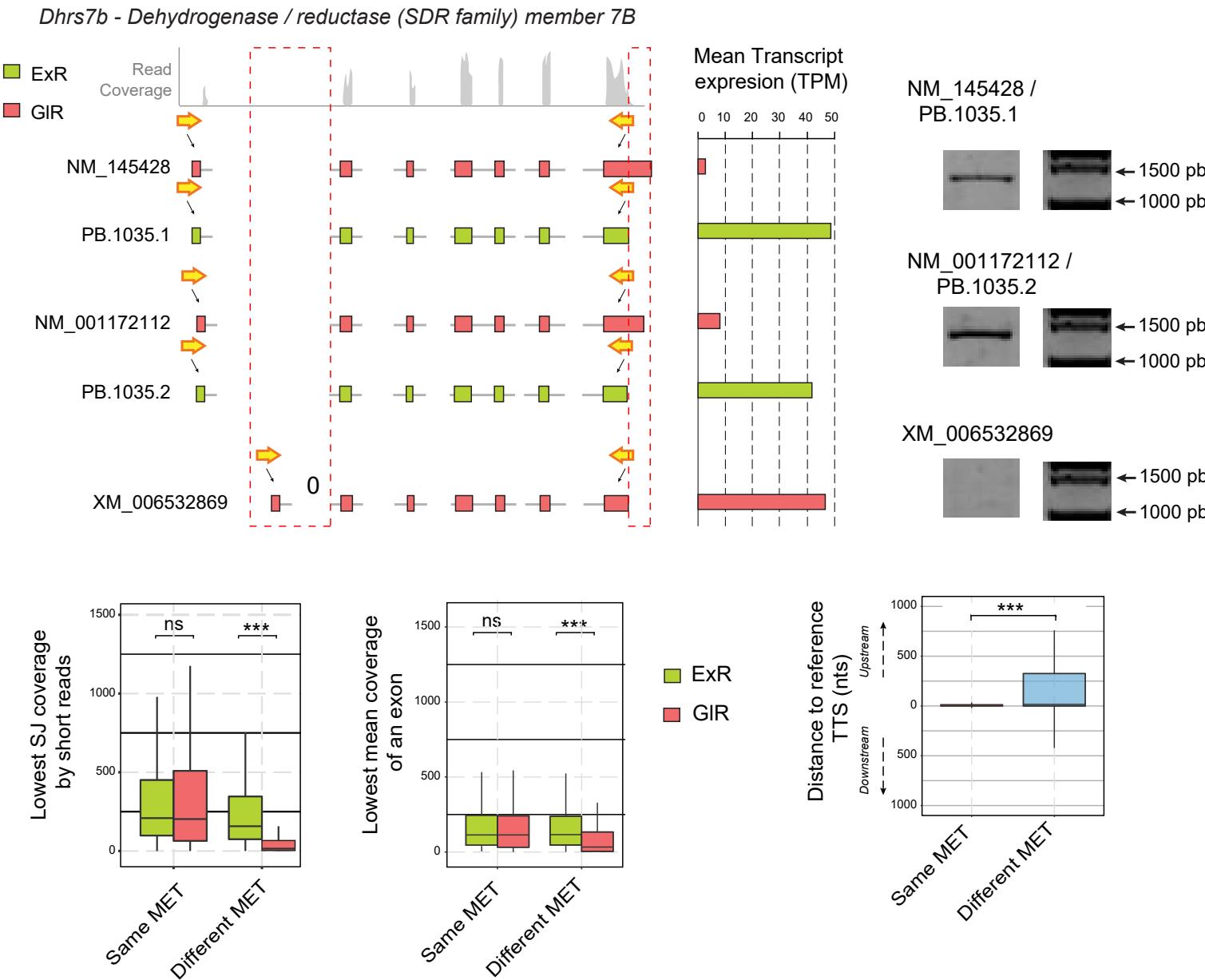
PacBio



RefSeq



# 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.



## 4. Functional outcome of alternative splicing: Transcript2GO

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

**Transcript2GO - aa1**

Projects Data Differential Features Enrichment Graphs Search by ID Name Hide unselected rows

Transcripts +

#	Transcript	Protein	Gene	Gene Description	NSC Meas...	OLD Me...
1	PB.1.1	Q9CRS	Mrp15	mitochondrial ribosomal pr...	4828.38	3987.79
2	PB.10.1	QBR0N6	Aldh1a	alcohol dehydrogenase, iro...	269.63	2371.07
3	PB.10.2	QBR0N6-2	Aldh1a	alcohol dehydrogenase, iro...	323.66	1521.42
4	PB.10.3	ADU00000066	Cox10b	coenzyme Q10 oxidoreduct...	822.00	1665.00
5	PB.1000.1	F9H7Z2	Mef2c	microtubule-associated pro...	906.46	1545.34
6	PB.1000.2	P14824	Anxa6	annexin A6	9470.88	33495.78
7	PB.1000.3	novelProt4205	Anxa6	annexin A6	90.15	703.52
8	PB.1000.4	novelProt1638	Anxa6	annexin A6	384.06	716.32
9	PB.1000.6	novelProt2501	Anxa6	annexin A6	189.36	156.48
10	PB.1000.9	novelProt2315	Anxa6	annexin A6	926.49	2342.39
11	PB.1001.1	Q6E64B	Gm2a	GM2 ganglioside activator ...	17256.57	60655.33
12	PB.1002.1	P07214	Sparc	secreted acidic cysteine ric...	28122.91	48738.97
13	PB.1003.1	Q3bP1	GTPase activating protein ...	5109.39	2911.25	
14	PB.1004.1	F6YNQ1	Gria1	glutamate receptor, ionotr...	226.61	3708.06
15	PB.1004.2	novelProt542	Gria1	glutamate receptor, ionotr...	17.25	429.02
16	PB.1004.3	P23818	Gria1	glutamate receptor, ionotr...	56.47	3564.83

App Info: aa1 Gene: Jmj4d Gene: Shmt1

General Summary +

Data Summary

Transcripts and Proteins per Gene

Transcript Structural Categories

Annotation Sources

Expression Levels PCA Plot

**Transcript2GO - aa1**

Projects Data Differential Features Enrichment Graphs Search by ID Name Hide unselected rows

Transcripts +

#	Transcript	Protein	Gene	Gene Description	NSC Meas...	OLD Me...
73	PB.1030.1	Q80Y17	Lig1	lethal giant larvae homolo...	4414.9	3949.76
74	PB.1030.2	AOA084054	Lig1	lethal giant larvae homolo...	958.84	526.53
75	PB.1031.1	Q9J28	Fili	flightless I homology (Dros...	1785.06	2310.74
76	PB.1032.1	Q3TB95	Mef2	microtubule-associated pro...	129.07	276.23
77	PB.1032.2	Q5NC59	Mef2	microtubule-associated pro...	533.57	1202.21
78	PB.1033.1	Q703a	Top3a	topoisomerase (DNA) II al...	2233.14	1310.98
79	PB.1034.1	P50431	Shmt1	serine hydroxymethyltransfer...	7715.68	1061.76
80	PB.1034.2	novelProt2914	Shmt1	serine hydroxymethyltransf...	545.04	98.41
81	PB.1035.1	Q9947	Dhrs7b	dehydrogenase/reductase ...	1388.27	1308.69
82	PB.1035.2	Q9947-2	Dhrs7b	dehydrogenase/reductase ...	1072.28	1579.63
83	PB.1036.1	Q90BW3	Nat12	N-acetyltransferase domain...	2075.43	1083.05
84	PB.1036.2	novelProt70	Map2k3	mitogen-activated protein ...	1798.61	2193.59
85	PB.1038.1	Q5DU02	Up22	ubiquitin specific peptidac...	4687.77	2753.5
86	PB.1039.2	P47740	Aldh3a2	aldehyde dehydrogenase ...	117.78	223.46
87	PB.1039.4	P47740	Aldh3a2	aldehyde dehydrogenase ...	2117.27	5854.1
88	PB.1039.5	novelProt1606	Aldh3a2	aldehyde dehydrogenase ...	73.75	61.72

App Info: aa1 Gene: Jmj4d Gene: Shmt1

Transcripts + Genomic +

Shmt1 - Serine hydroxymethyltransferase 1 (soluble)

Genomic View - Chromosome 11 for Project 'aa1'

**Transcript2GO - aa1**

Projects Data Differential Features Enrichment Graphs Search by ID Name Hide unselected rows

Transcripts +

#	Transcript	Protein	Gene	Gene Description	NSC Meas...	OLD Me...
46	PB.1015.1	E9Q2M4	Zfp867	zinc finger protein 867	377.92	626.78
47	PB.1015.1	P88078	Zfp867	zinc finger protein 867	274.08	179.63
48	PB.1016.1	novelProt96	Snap47	synaptoosomal-associated p...	532.26	1931.63
49	PB.1016.2	Q8R570	Snap47	synaptoosomal-associated p...	7784.53	1798.02
50	PB.1017.1	novelProt5383	Jmj4d	jumonji domain containing 4	390.85	489.11
51	PB.1018.1	E9Q2M4	Zfp867	zinc finger protein 867	128.58	148.3
52	PB.1018.2	novelProt5295	Zfp867	zinc finger protein 867	1185.7	466.8
53	PB.1018.3	XP_011247286.1	Zfp867	zinc finger protein 867	658.45	5540.7
54	PB.1019.1	novelProt94	Zkscn17	zinc finger with KRAB and ...	520.77	293.72
55	PB.102.1	Q4PF6	Mob4	M63 family member 4, ph...	500.00	50.50
56	PB.102.1	novelProt2314	Mob4	M63 family member 4, ph...	411.48	529.34
57	PB.102.2	novelProt2100	Maria	mavin phosphatase Rho 1...	148.58	108.44
58	PB.102.2	novelProt370	Maria	mavin phosphatase Rho 1...	225.77	245.42
59	PB.102.2.1	Q9QZ53	Flicn	foliculin	470.8	469.15
60	PB.102.2.2	Q9QZ53	Flicn	foliculin	163aa	
61	PB.102.2.3	Q9QZ53	Flicn	foliculin		

App Info: aa1 Gene: Jmj4d Gene: Jmj4d

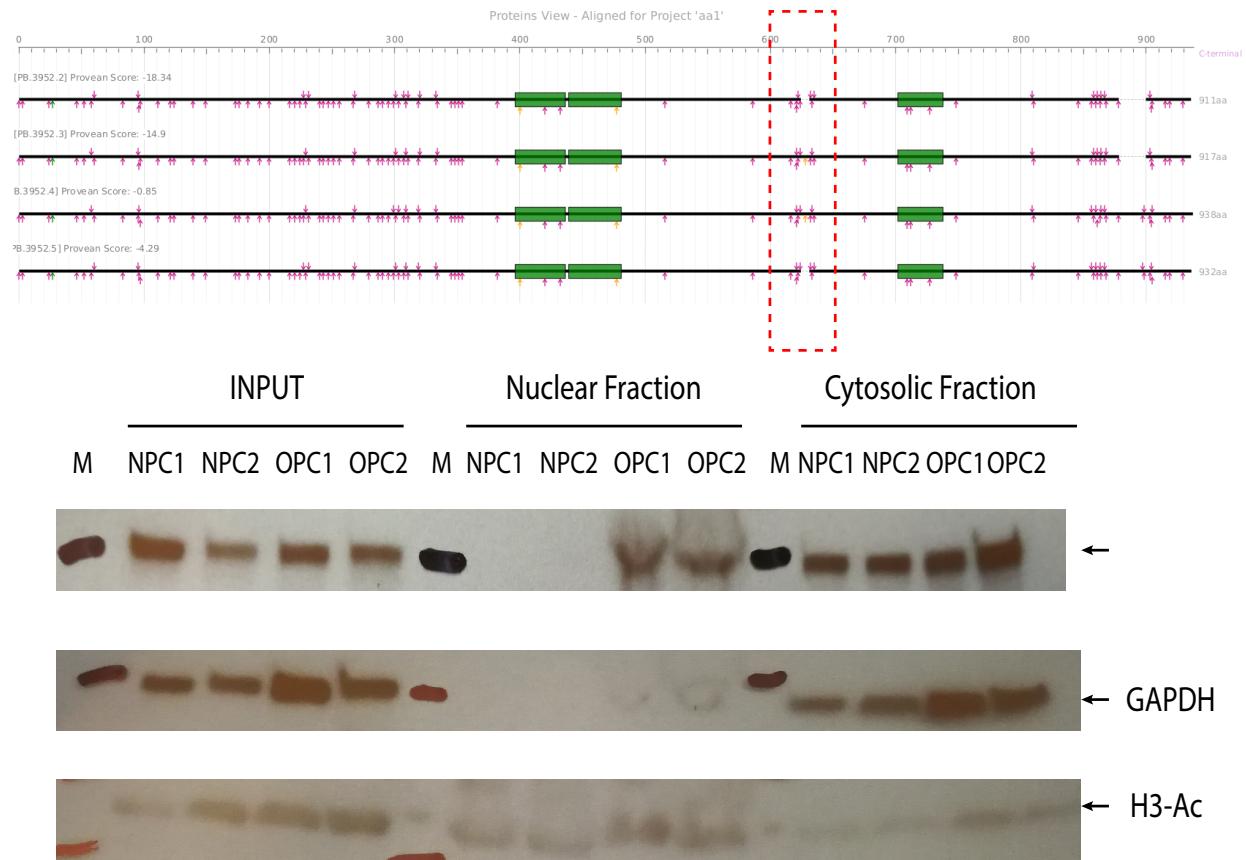
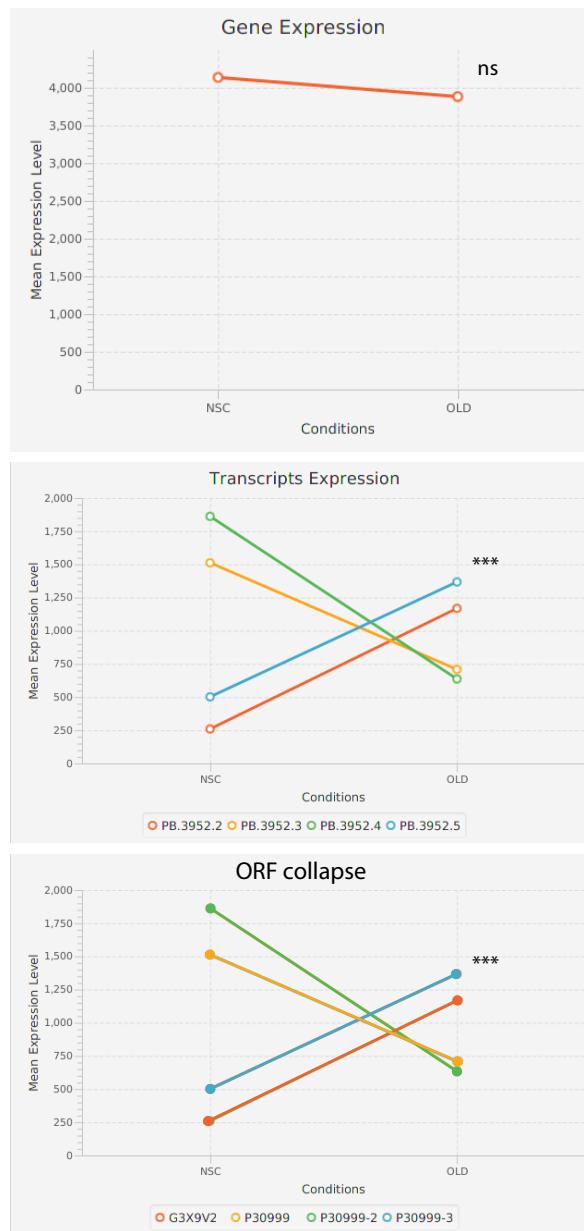
Transcripts + Genomic + Proteins +

Jmj4d - Jumonji domain containing 4

Proteins View - Aligned for Project 'aa1'

**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

