



# SQANTI and TAPPAS: Making Sense of Iso-Seq Data

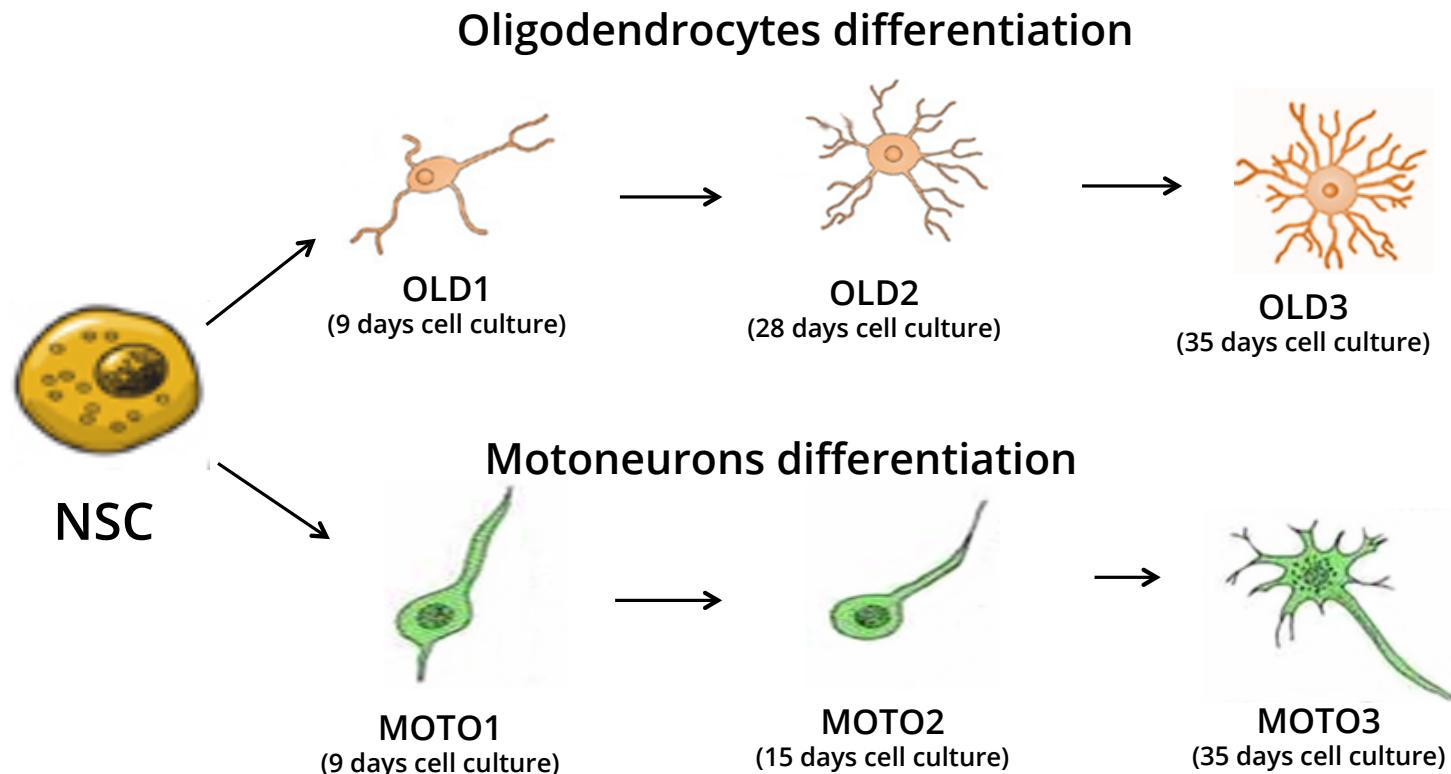
Ana Conesa, PhD  
Genomics of Gene Expression Lab  
CIPF/UF



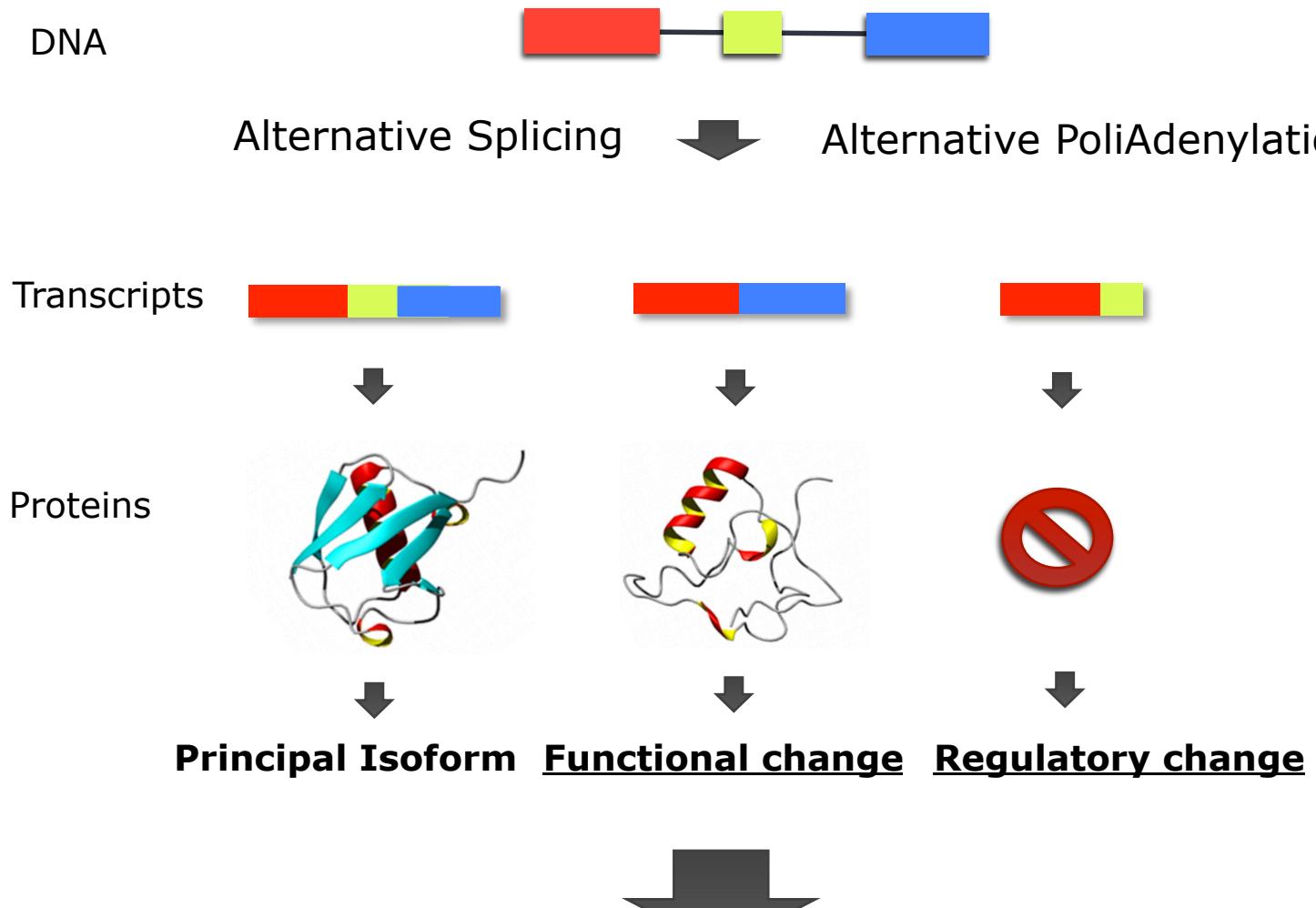
PRINCIPE FELIPE  
CENTRO DE INVESTIGACION



# Functional Implications of Differential Splicing?

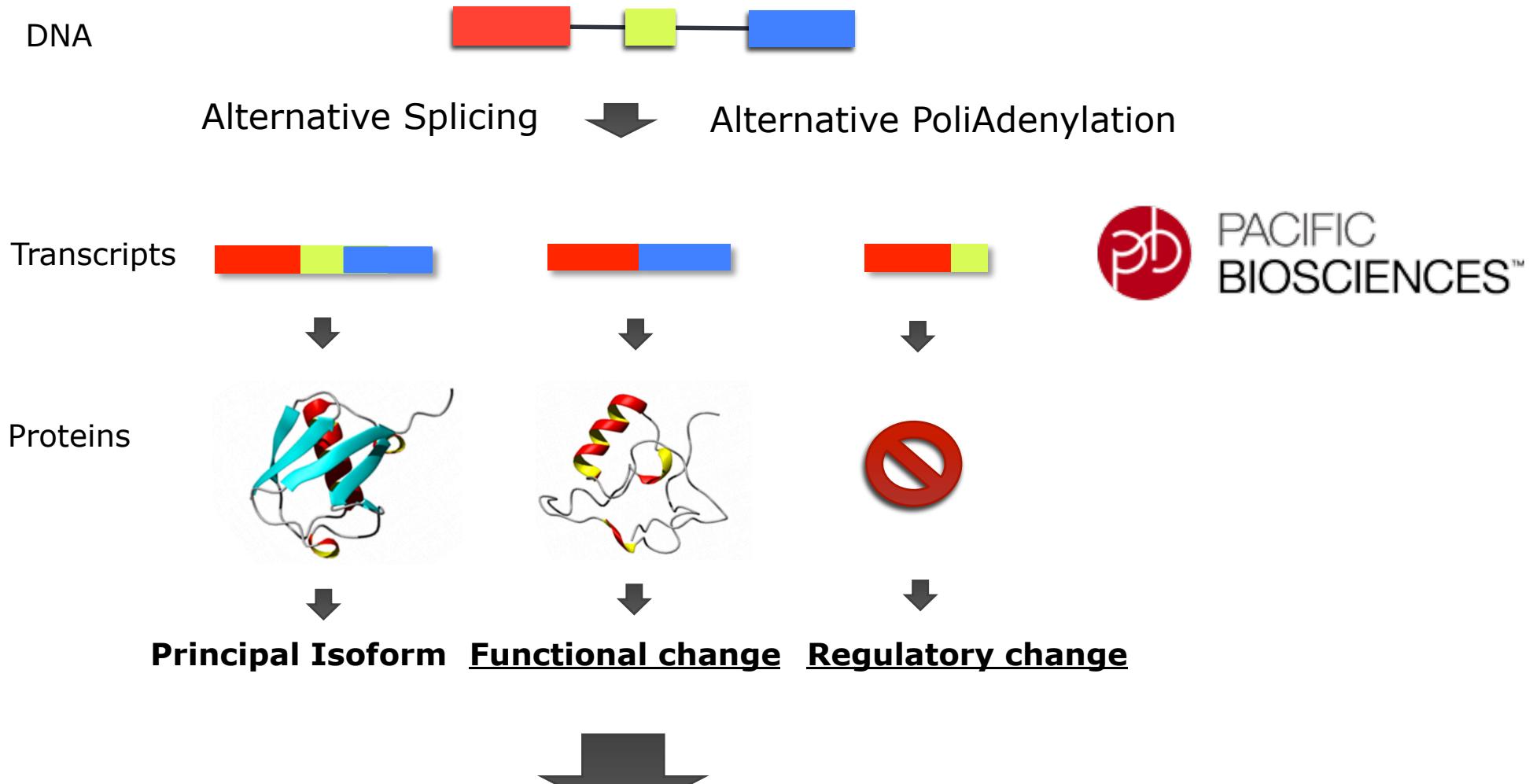


# Functional Implications of Differential Splicing



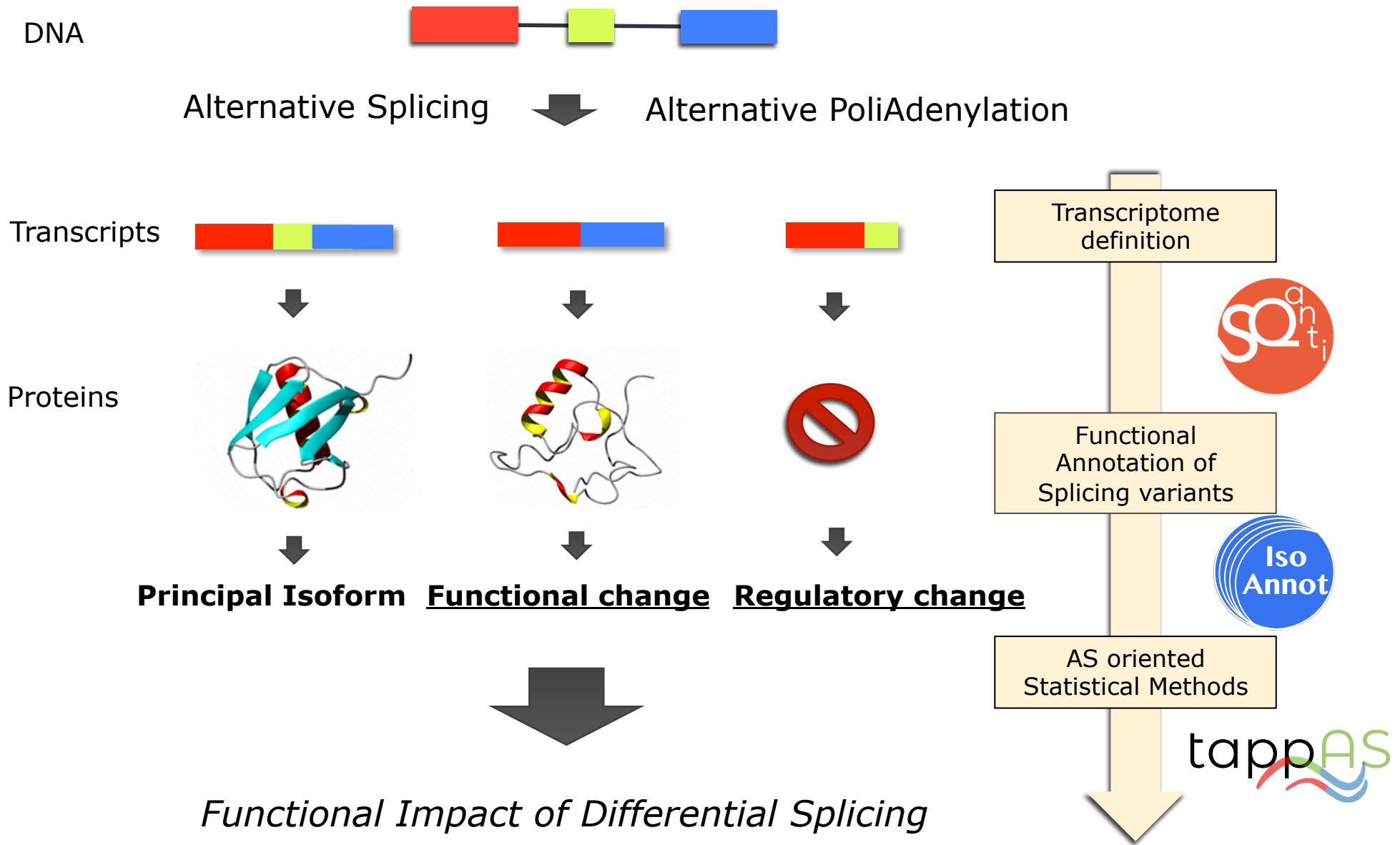
*Functional Impact of Differential Splicing*

# Functional Implications of Differential Splicing



*Functional Impact of Differential Splicing*

# Functional Implications of Differential Splicing





## Structural and Quality Annotation of Transcript Isoforms

6

<https://bitbucket.org/ConesaLab/sqanti>

Tardaguila et al. *SQANTI: extensive characterization of long read transcript sequences for quality control in full-length transcriptome identification and quantification*. **Preprint at BiorXiv. 2017**  
**Genome Biology, in press.**

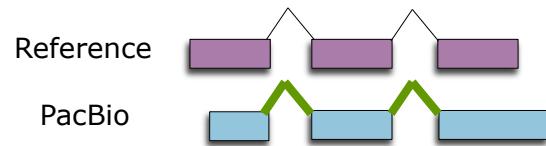
# Transcriptome characterization



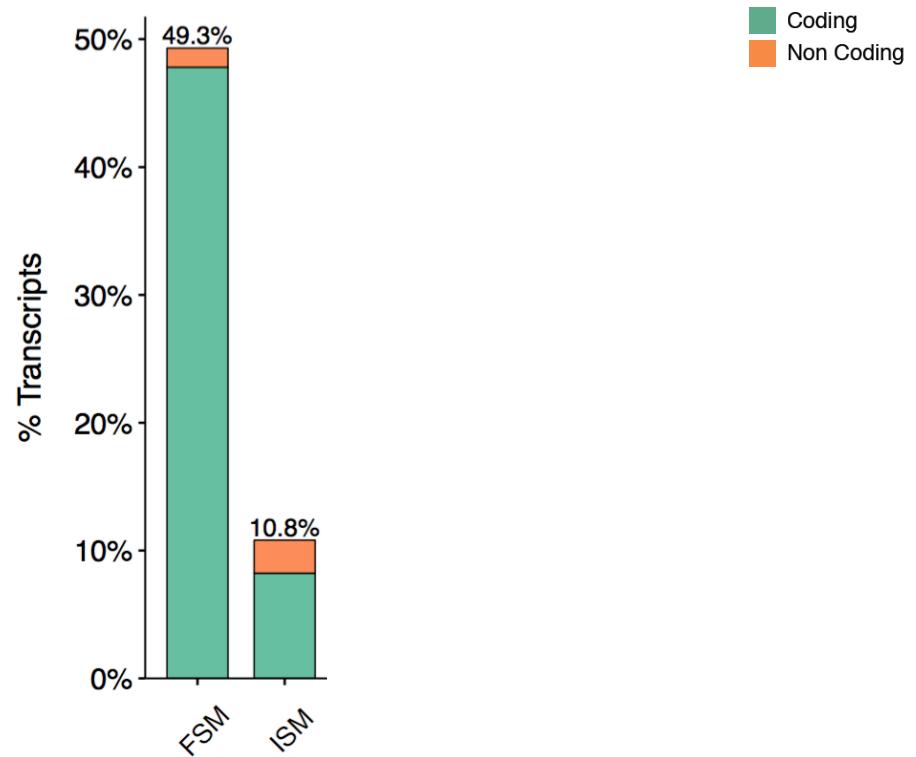
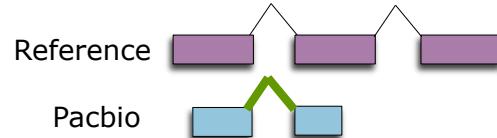
## 1. Classification

### Known Isoforms

#### Full-Splice Match FSM



#### Incomplete-Splice Match ISM



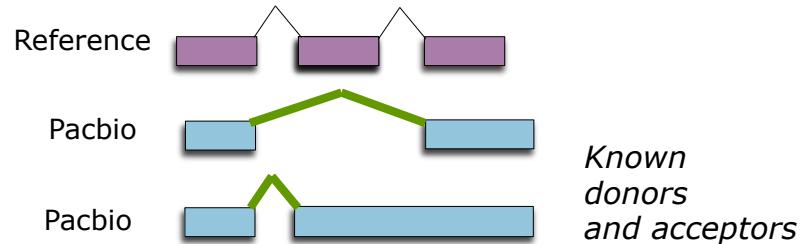
# Transcriptome characterization



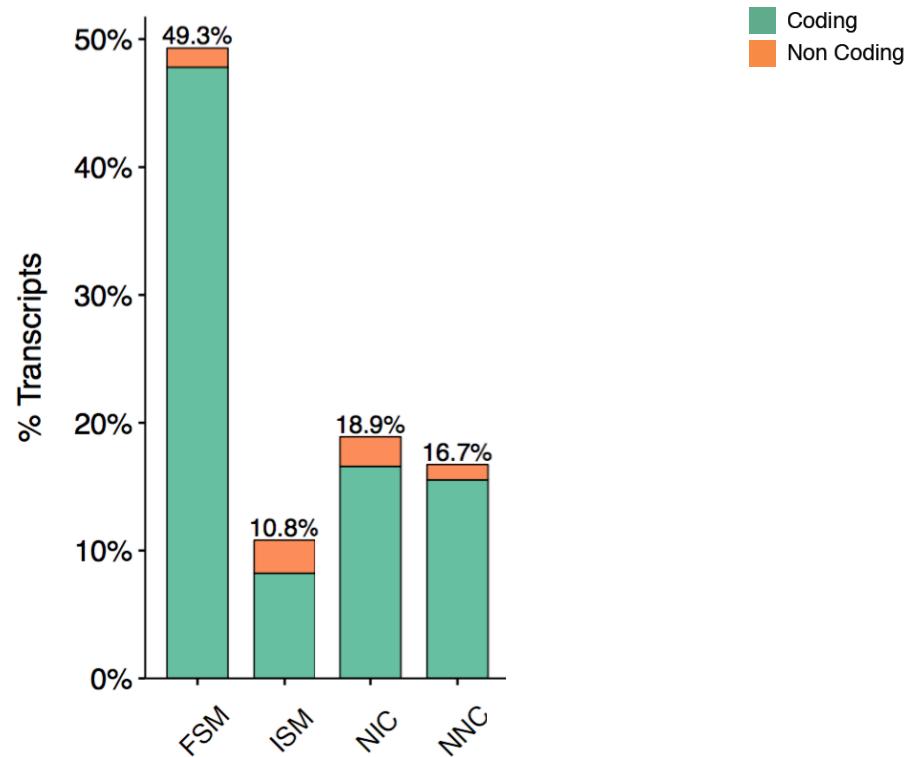
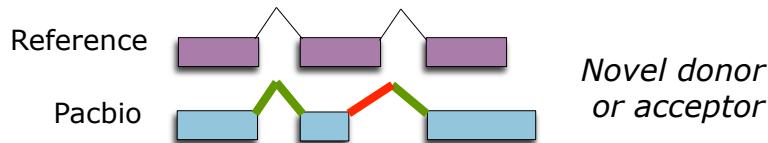
## 1. Classification

### Novel Isoforms – Known genes

#### Novel In catalog **NIC**



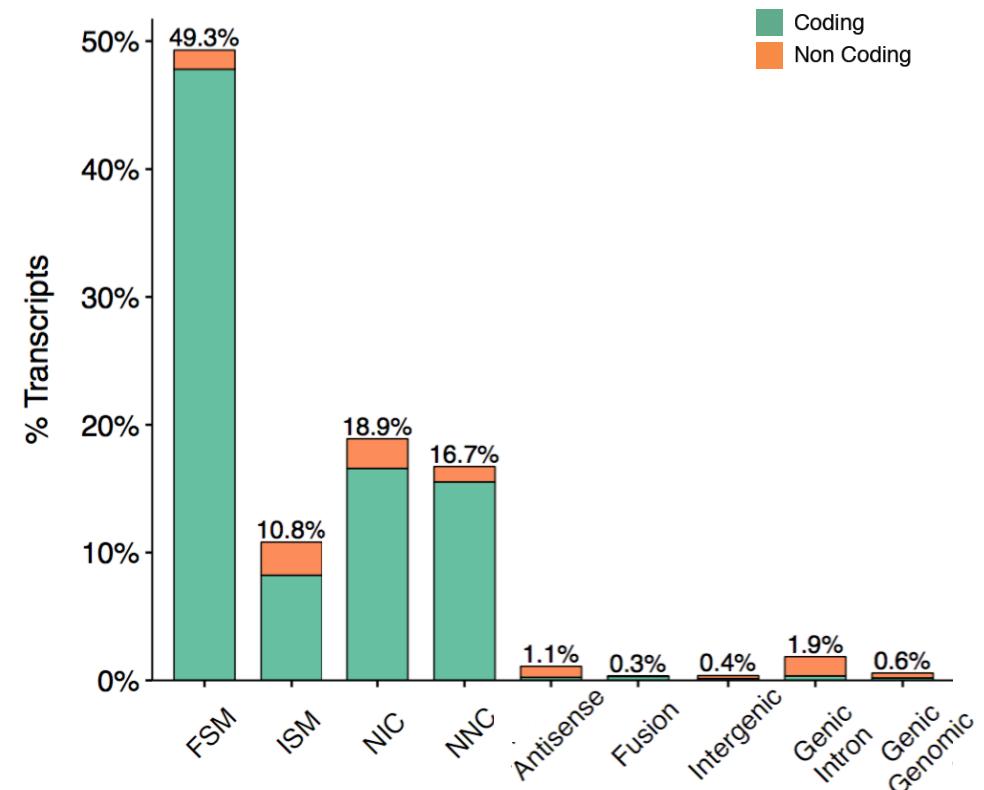
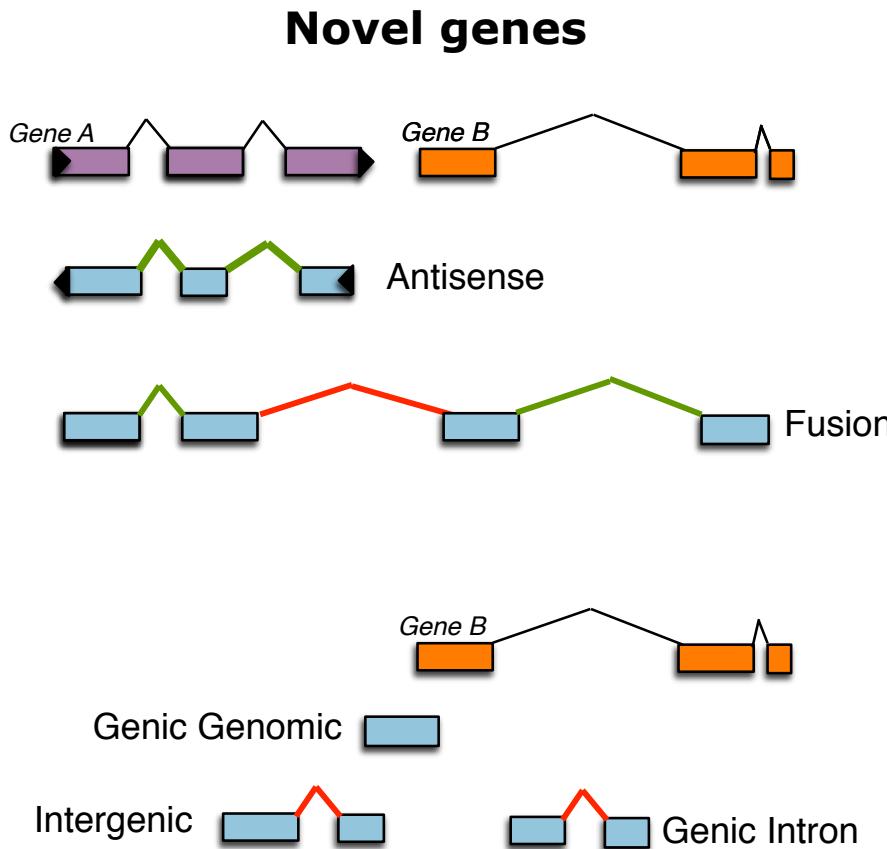
#### Novel Not in Catalog **NNC**



# Transcriptome characterization



## 1. Classification

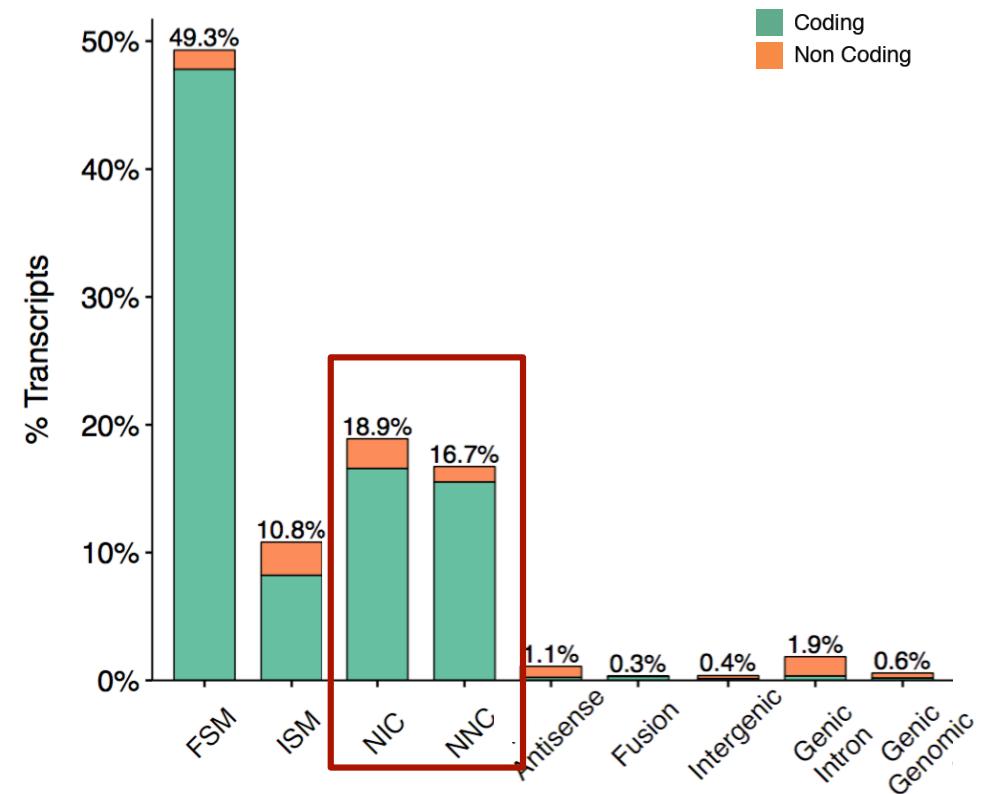




## 1. Classification

35 % of novel isoforms in mouse...

Are all of them real?



# Transcriptome characterization

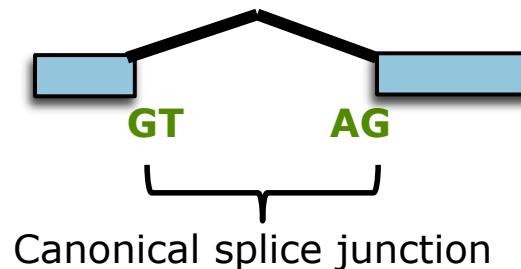


1. Classification
- 2. QC descriptors**



## 1. Classification

## 2. QC descriptors: SJ canonical status



≈ 98,7 % of canonical SJ in mammalian\*

**97,7 % of total splice junctions in our neural transcriptome are canonical**

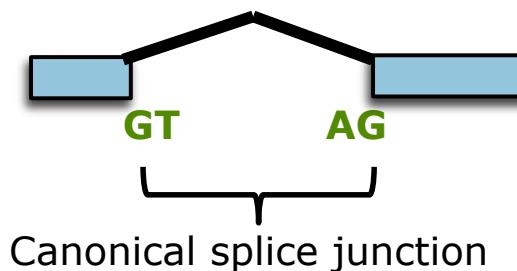
\*Burset et al, 2000

# Transcriptome characterization



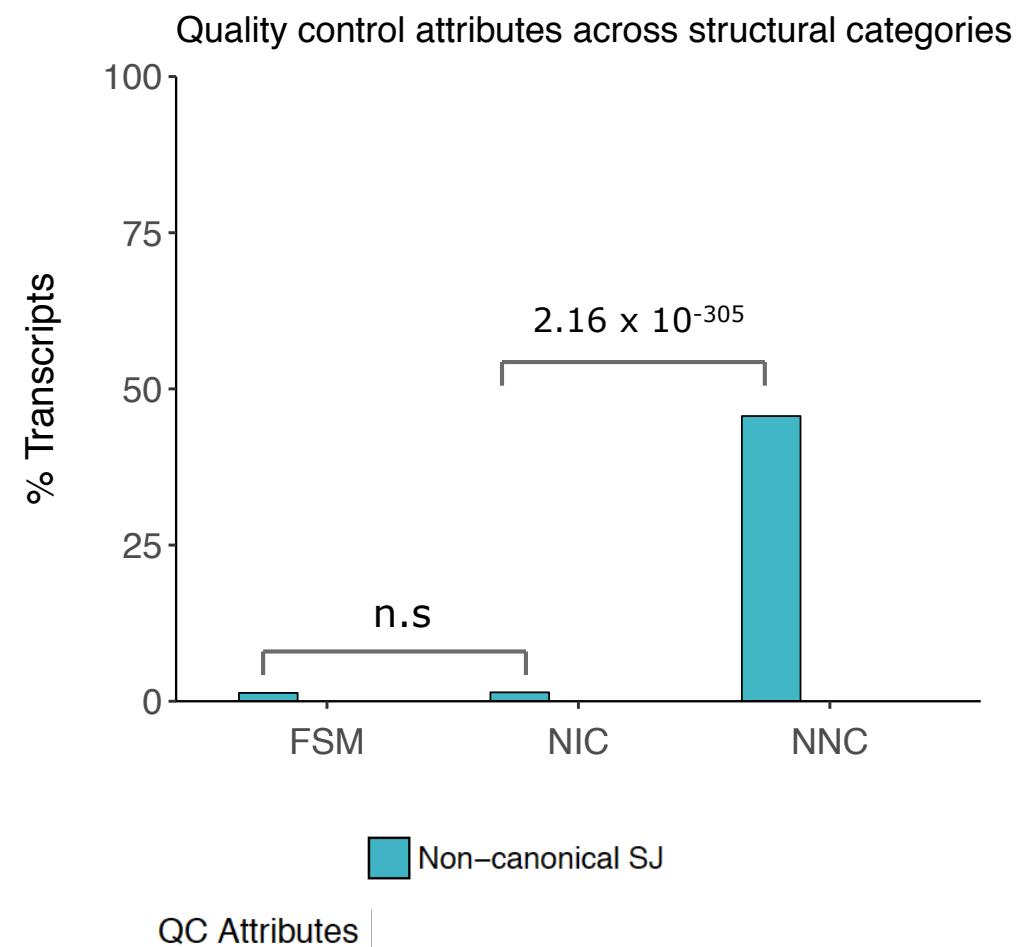
## 1. Classification

## 2. QC descriptors: SJ canonical status



≈ 98,7 % of canonical SJ in mammalian\*

**97,7 % of total splice junctions in our neural transcriptome are canonical**



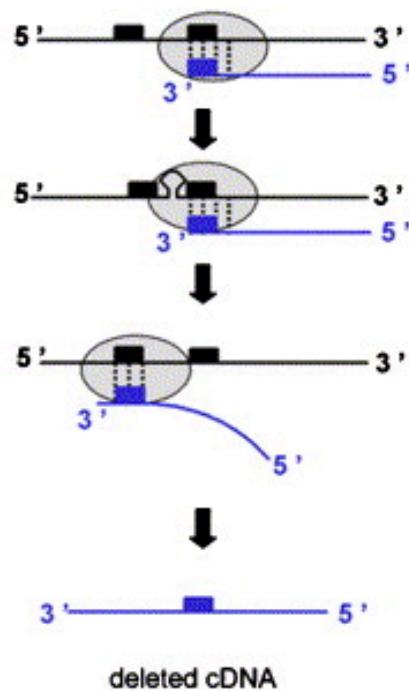
\*Burset et al, 2000

# Transcriptome characterization



## 1. Classification

## 2. QC descriptors: **RT-switching**



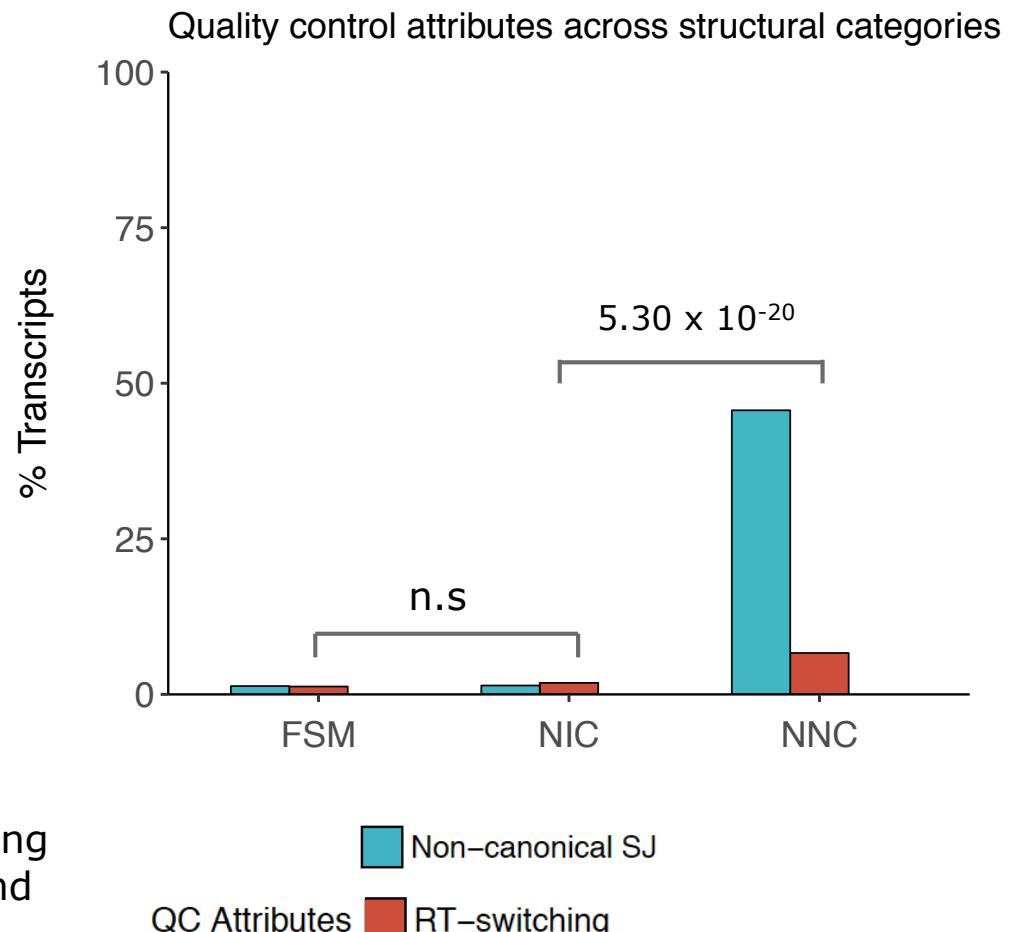
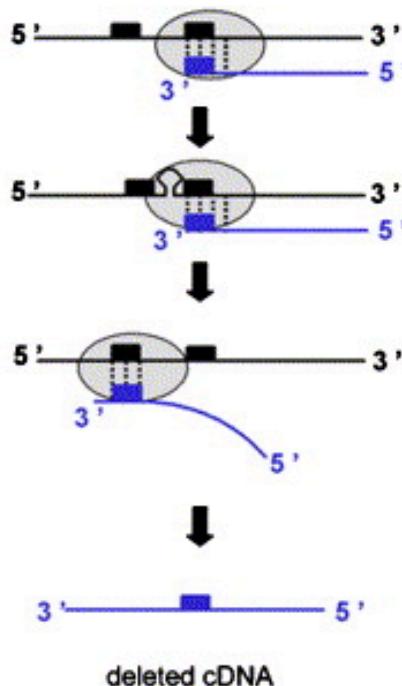
- Reverse transcriptase template switching
- Caused by RNA secondary structure and repeated regions.
- Appears as novel splice junctions

# Transcriptome characterization



## 1. Classification

## 2. QC descriptors: **RT-switching**



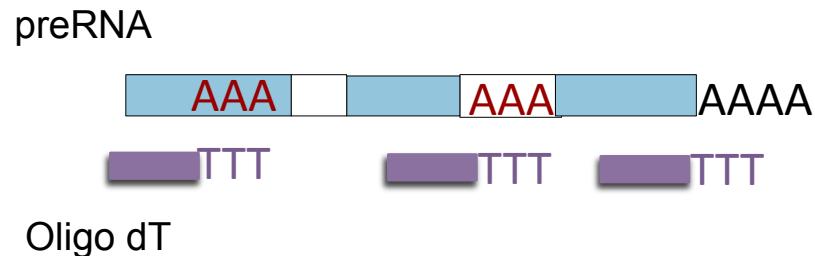
- Reverse transcriptase template switching
- Caused by RNA secondary structure and repeated regions.
- Appears as novel splice junctions

# Transcriptome characterization

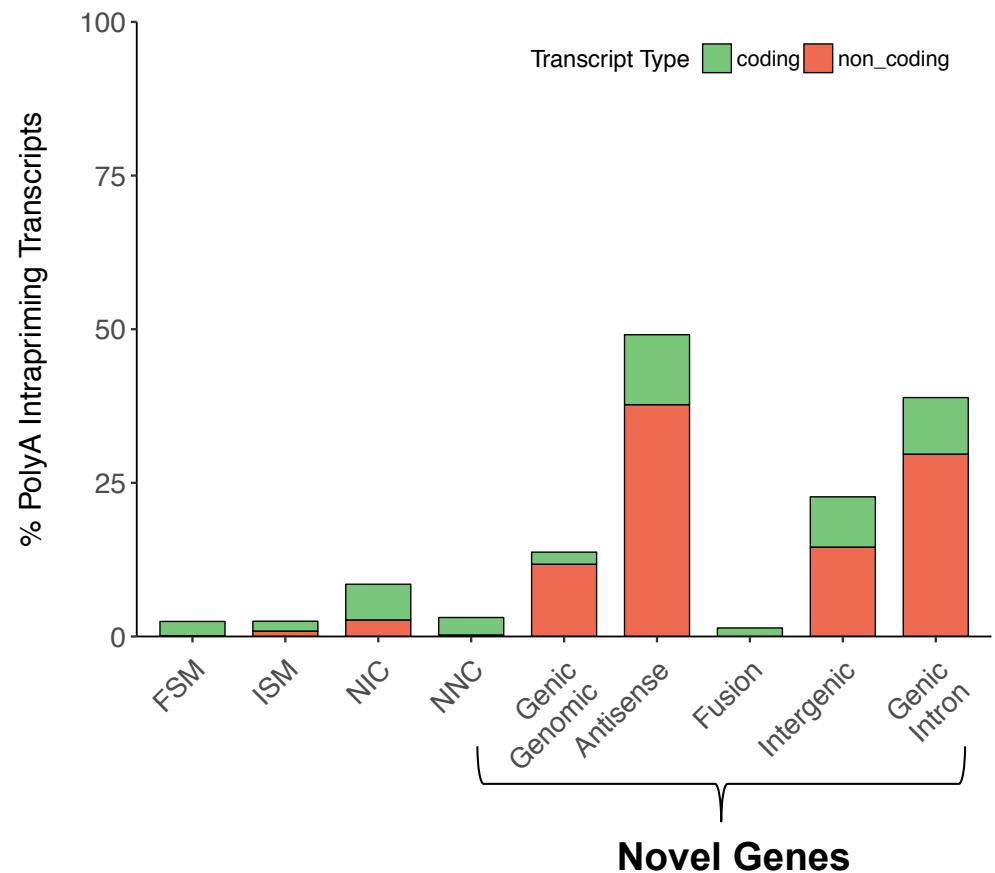


## 1. Classification

## 2. QC descriptors: PolyA intra-priming



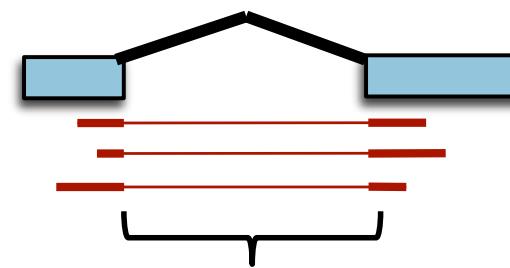
- oligodT can prime outside polyA tail in A rich regions inside transcribed regions.
- We looked for transcripts showing  $\geq 80\%$  Adenines in the 20 nts downstream “detected” 3’ end



# Transcriptome characterization

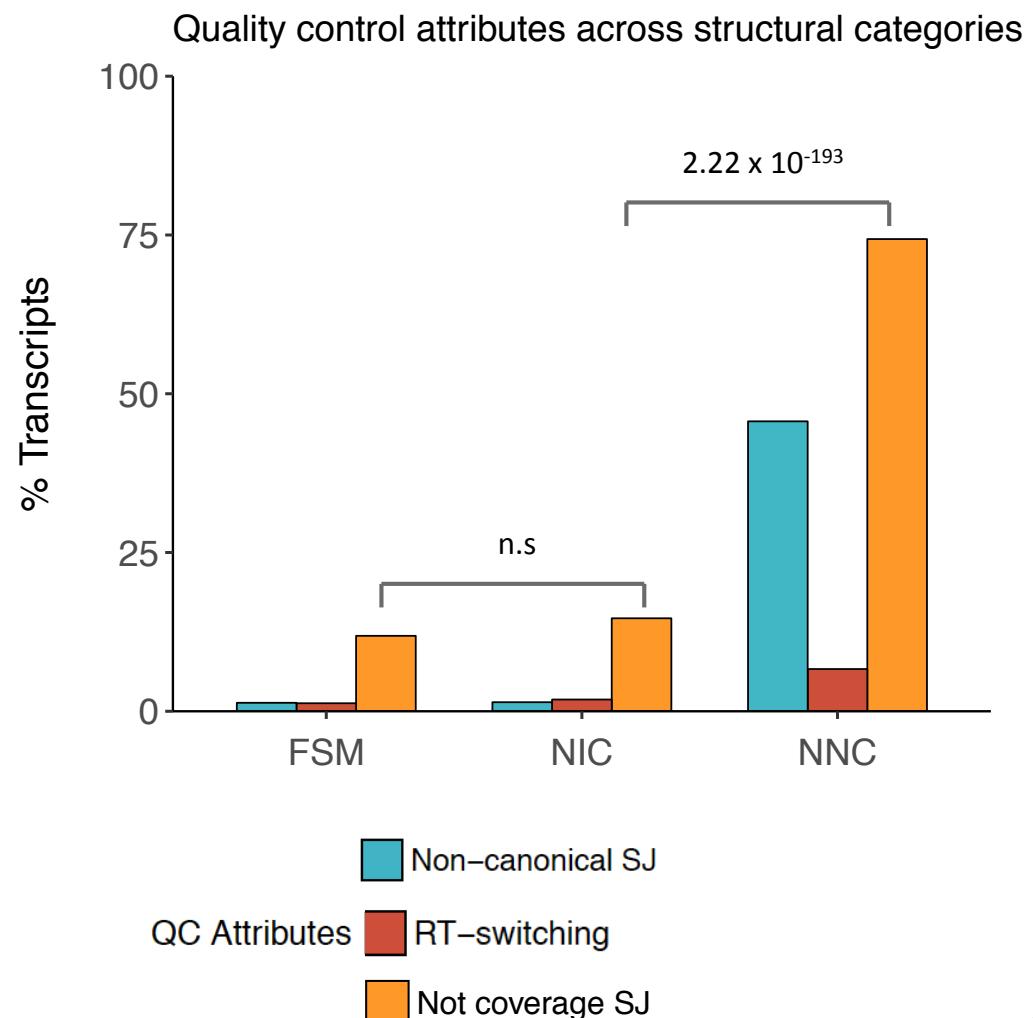


1. Classification
2. QC descriptors: **SJ support**



Supported splice junction

*Illumina Reads from same  
cDNA sequenced by PacBio*





## Transcript level attributes

1. Transcript Classification
  1. Reference Gene match
  2. Reference Transcript match
  3. Structural Category
2. Structural characteristics
  1. Detected/Reference Length
  2. Detected/Reference number of exons
  3. Distance to nearest annotated TSS
  4. Distance to nearest annotated TTS
  5. Bite
3. Quality Control attributes
  1. RT-switching
  2. PolyA Intrapriming
  3. Canonical status
  4. Indels near SJ
4. Support
  1. Minimum splice junction coverage
  2. Minimum sample coverage
  3. Minimum coverage position
  4. Number of Full-length reads supporting the transcript
5. Expression levels:
  1. Transcript level
  2. Gene level
6. Coding potential
  1. Coding/non coding
  2. ORF/CDS length
  3. CDS start and end positions

...

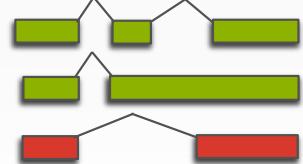
## Junction level attributes

1. Junction Classification
  1. Novel/Known
  2. Splice site motif
2. Structural characteristics
  1. Difference to nearest ref. donor
  2. Difference to nearest ref. acceptor
  3. Bite
3. Quality Control attributes
  1. Canonical
  2. Rts\_junction
  3. Indel near junc
4. Support
  1. Samples with cov
  2. Total coverage
  3. Coverage per sample

# Filtering out artifact isoforms

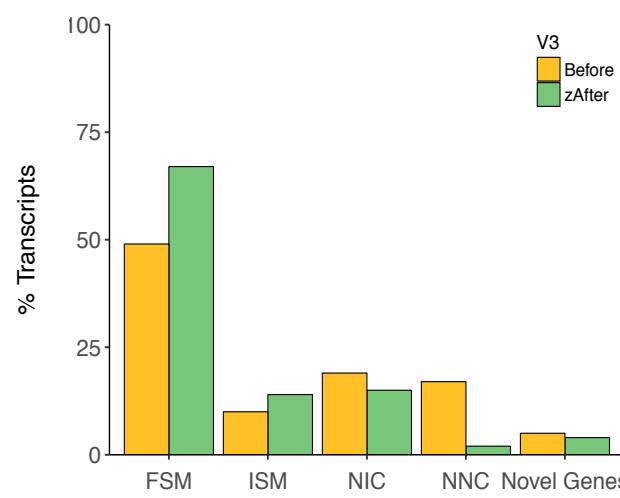
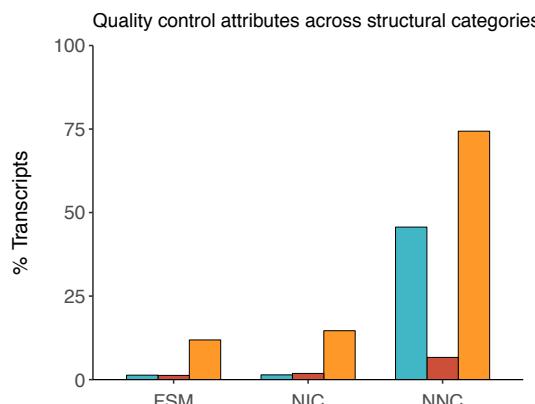


PacBio Transcriptome

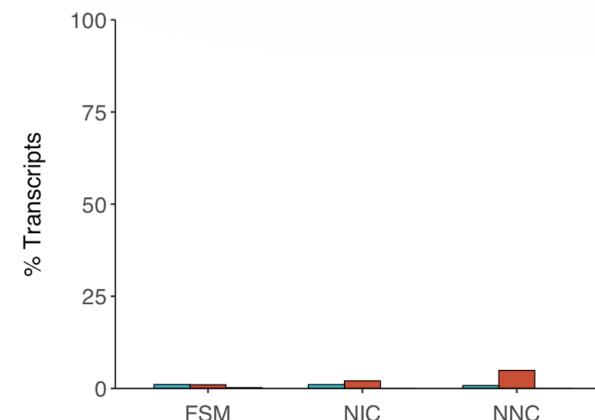


Machine learning  
with SQANTI descriptors

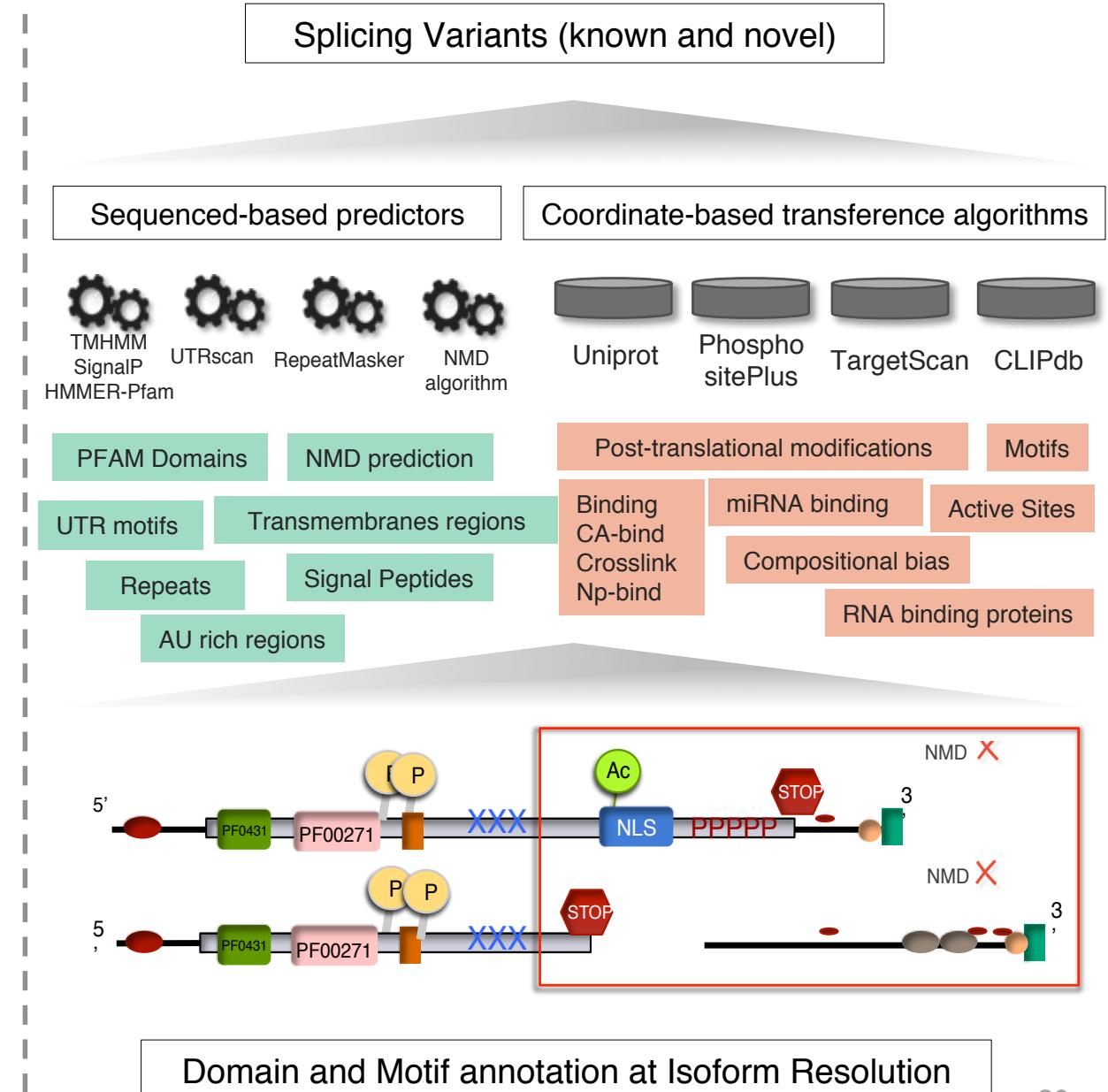
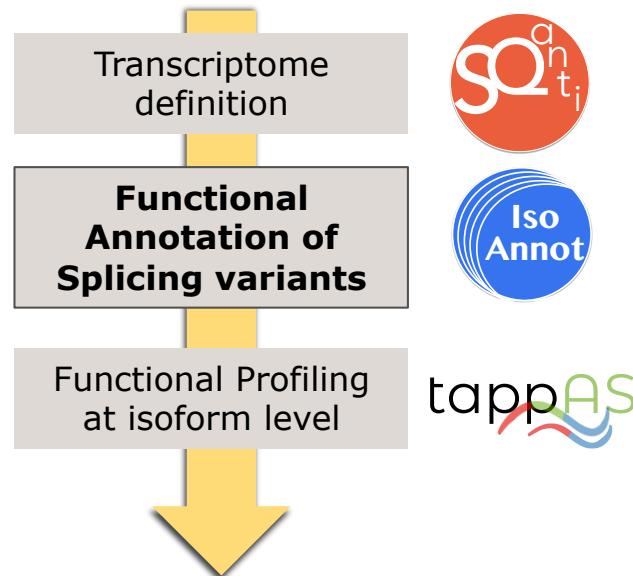
- Non-canonical SJ
- RT-switching
- Not coverage SJ



Curated PacBio  
Transcriptome



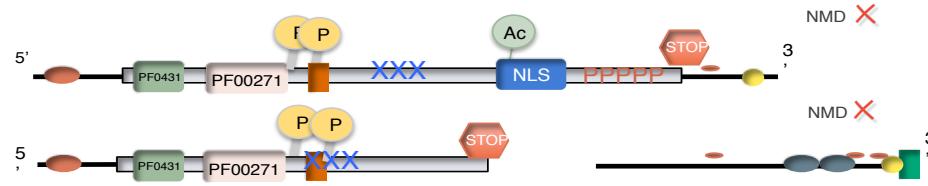
# Functional Annotation of Splicing Variants



# Functional Profiling at Isoform Level



## Structural Annotation and Functional Annotation



Reference Annotation provided  
OPTIONAL: User-defined

I

## Isoform quantification

Isoform set	Cond 1	Cond 1	Cond 1	Cond 2	Cond 2	Cond 2

User-defined

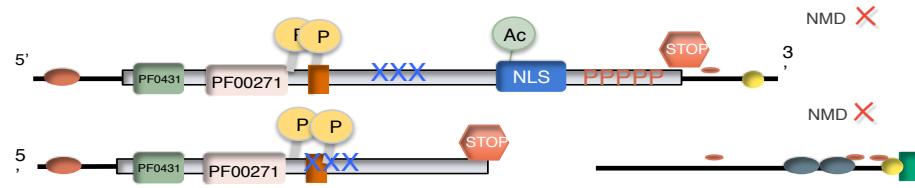
INPUT

tappAS

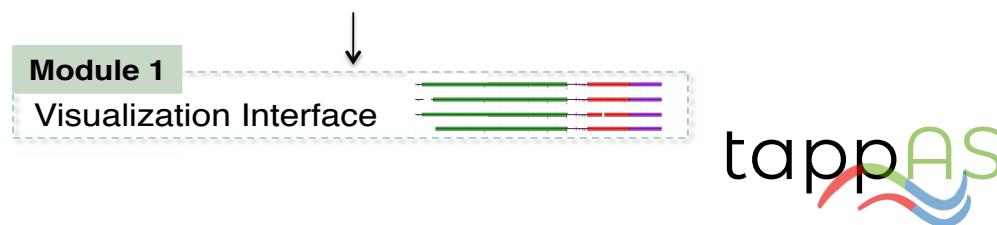
# Functional Profiling at Isoform Level



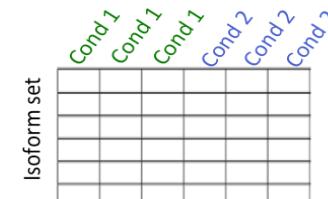
## Structural Annotation and Functional Annotation



Reference Annotation provided  
OPTIONAL: User-defined



## Isoform quantification



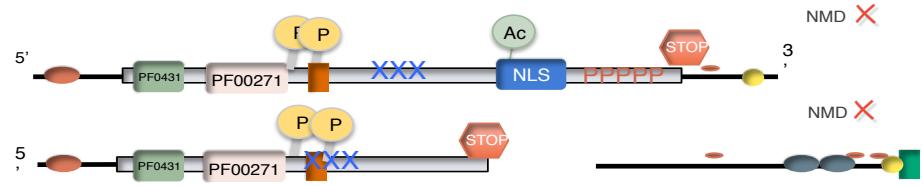
## User-defined

## INPUT

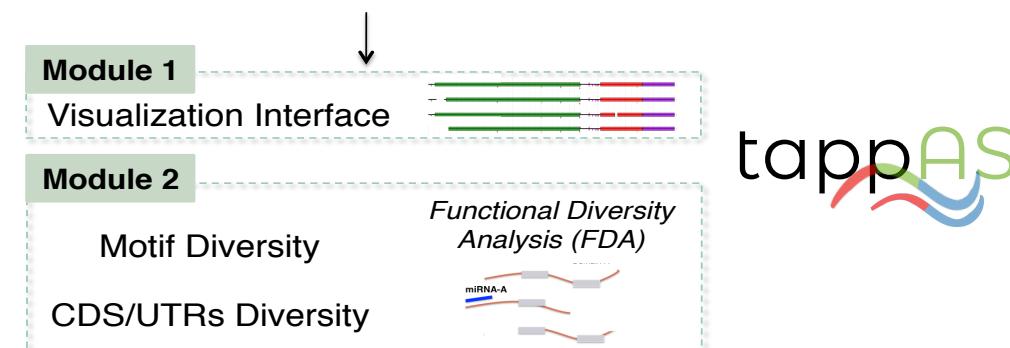
# Functional Profiling at Isoform Level



## Structural Annotation and Functional Annotation



Reference Annotation provided  
OPTIONAL: User-defined



## Isoform quantification

Isoform set	Cond 1	Cond 1	Cond 1	Cond 2	Cond 2	Cond 2

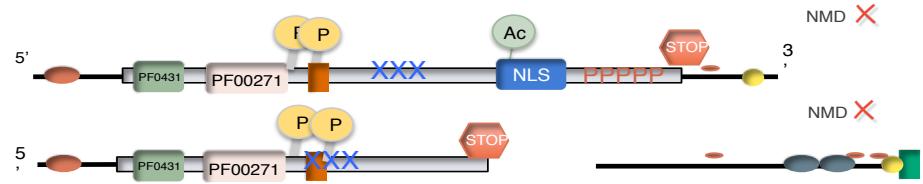
User-defined

INPUT

# Functional Profiling at Isoform Level



## Structural Annotation and Functional Annotation



Reference Annotation provided  
OPTIONAL: User-defined

### Module 1

Visualization Interface

### Module 2

Motif Diversity

CDS/UTRs Diversity

### Functional Diversity Analysis (FDA)



tappAS

## Isoform quantification

Isoform set	Cond 1	Cond 1	Cond 1	Cond 2	Cond 2	Cond 2

User-defined

INPUT

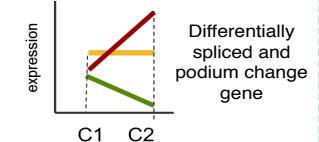
### Module 3

Differentially Spliced genes

Major Isoform Switching

Differentially expressed genes/transcripts/ORFs

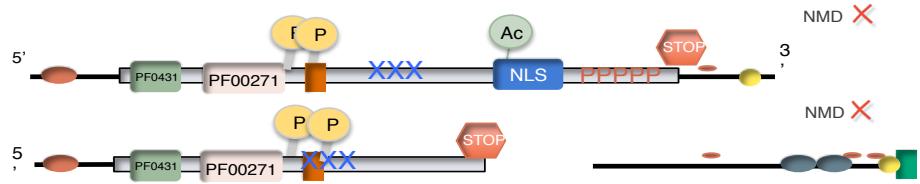
### Differential Analysis (DSA/DEA)



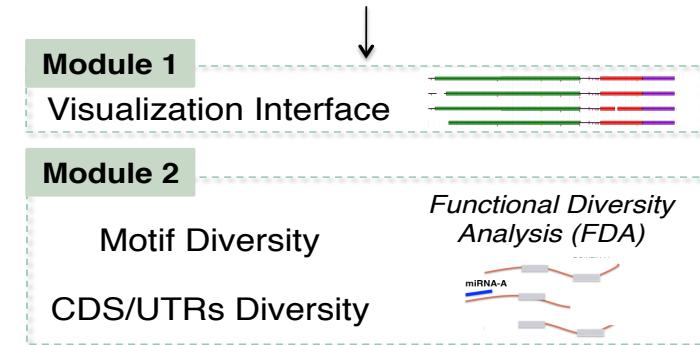
# Functional Profiling at Isoform Level



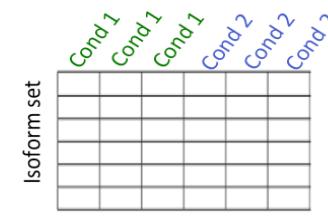
## Structural Annotation and Functional Annotation



Reference Annotation provided  
OPTIONAL: User-defined



## Isoform quantification



User-defined

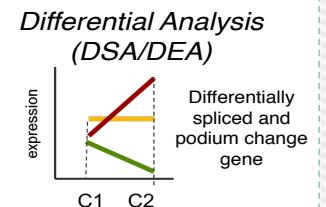
INPUT

**Module 3**

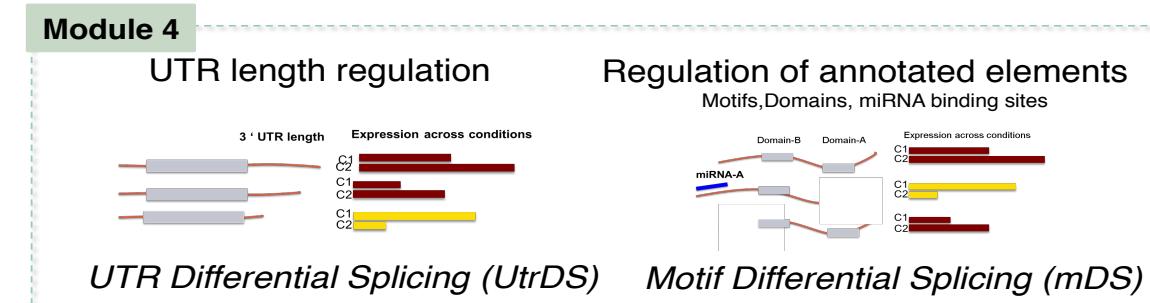
Differentially Spliced genes

Major Isoform Switching

Differentially expressed genes/transcripts/ORFs



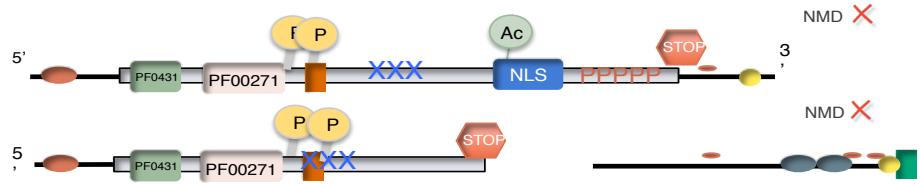
Integrative methods



# Functional Profiling at Isoform Level



## Structural Annotation and Functional Annotation



Reference Annotation provided  
OPTIONAL: User-defined

### Module 1

Visualization Interface

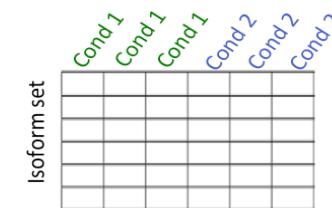
### Module 2

Motif Diversity

CDS/UTRs Diversity



## Isoform quantification



User-defined

INPUT

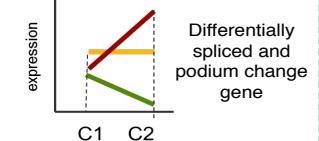
### Module 3

Differentially Spliced genes

Major Isoform Switching

Differentially expressed genes/transcripts/ORFs

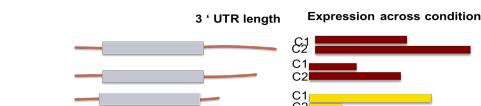
### Differential Analysis (DSA/DEA)



## Integrative methods

### Module 4

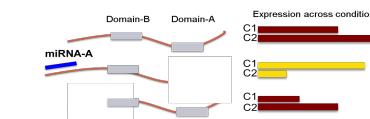
UTR length regulation



*UTR Differential Splicing (UtrDS)*

Regulation of annotated elements

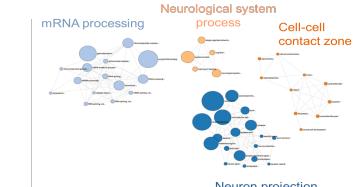
Motifs, Domains, miRNA binding sites



*Motif Differential Splicing (mDS)*

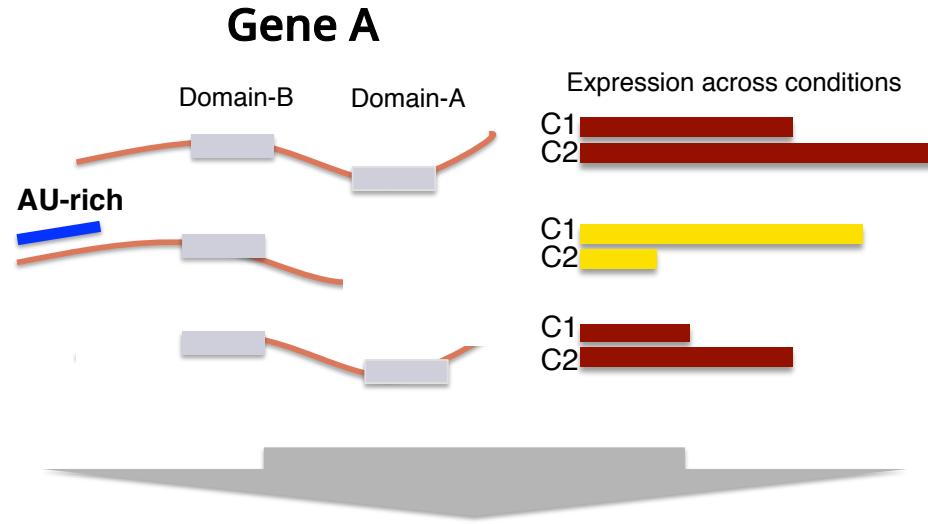
### Module 5

Functional enrichment over any annotated category

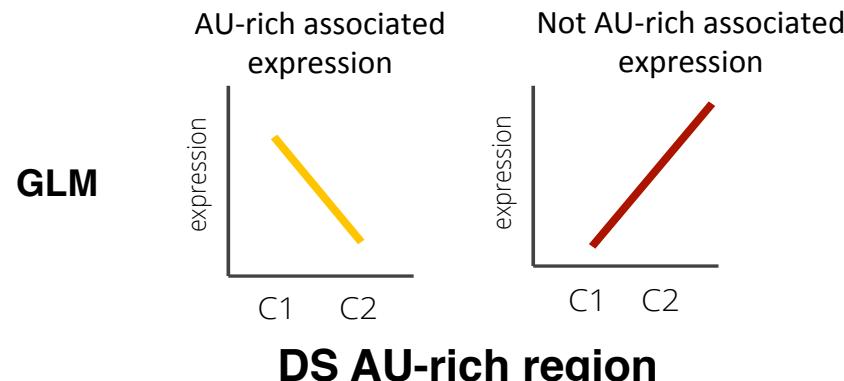


*Functional Enrichment and Gene Set Analysis (FEA/GSA)*

# Motif and Feature Differential Splicing



Significant differential usage of AU-rich motif in Gene A?

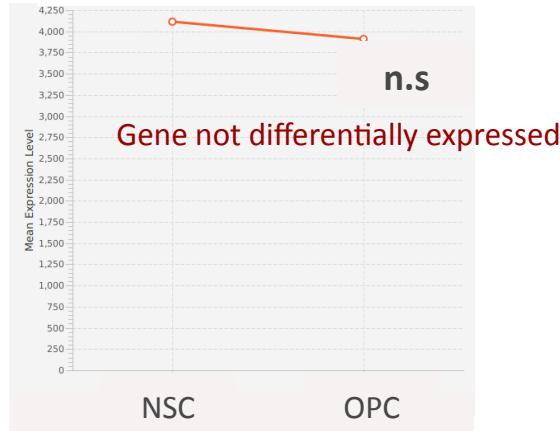


**AU-rich element favored in condition 1 by DS**

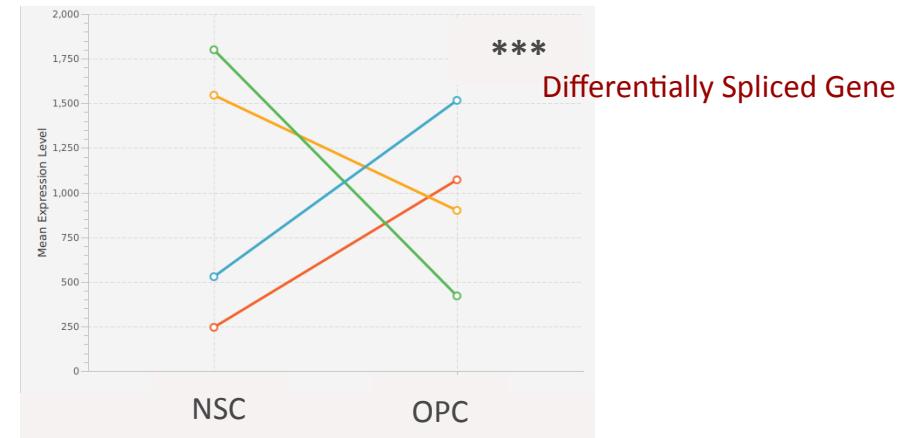


## *Regulation of protein motifs by differential splicing*

Gene Expression

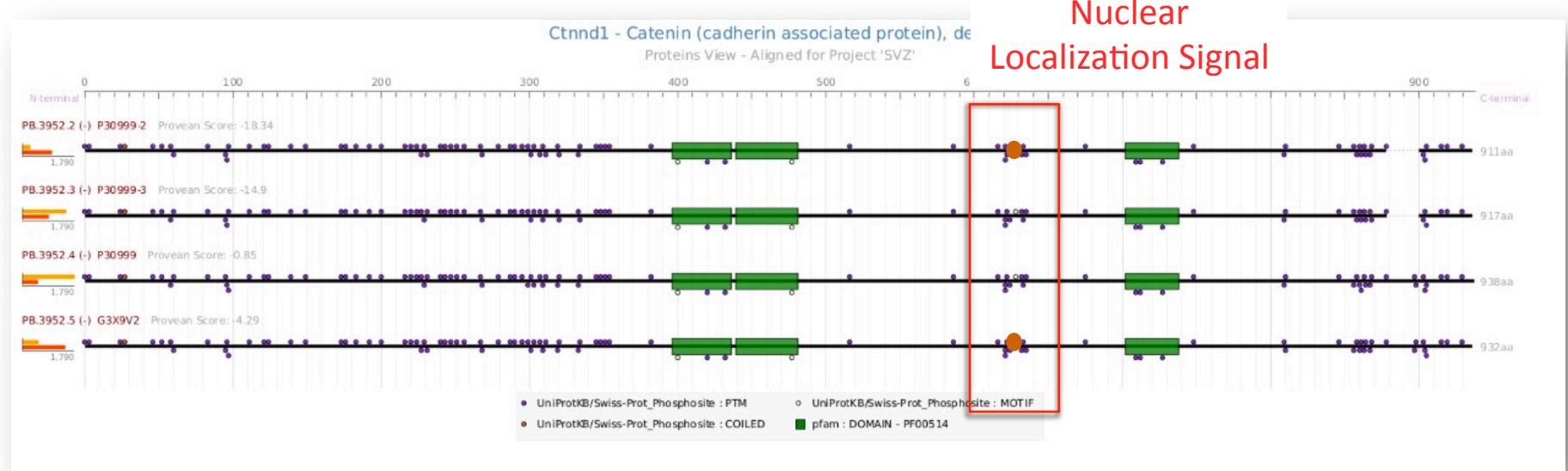


Transcripts Expression



Functional impact?

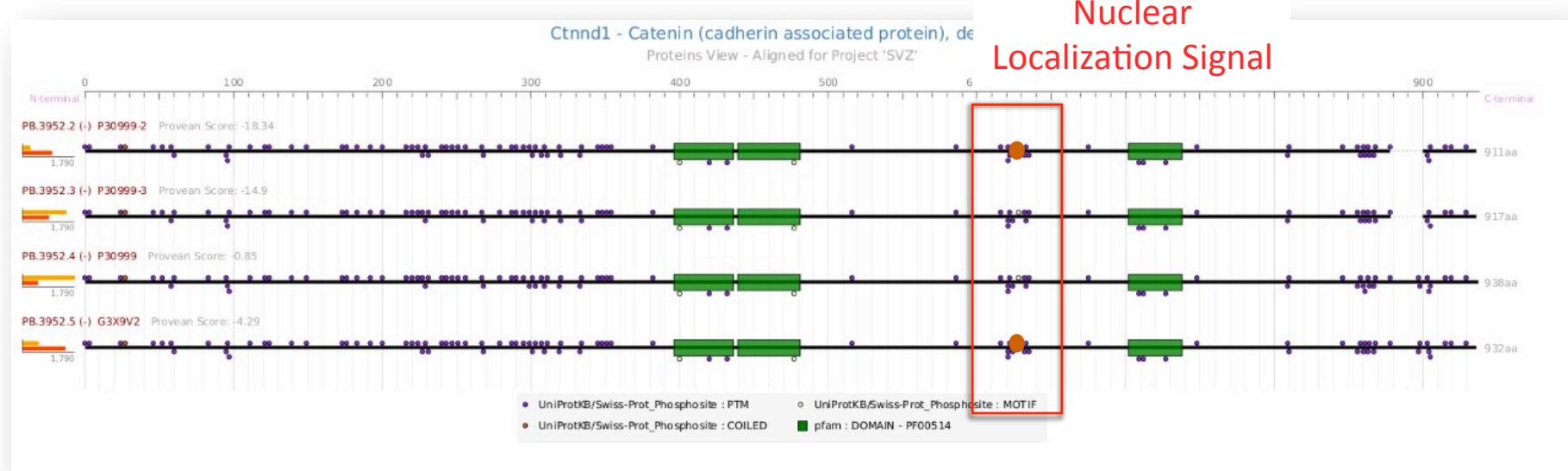
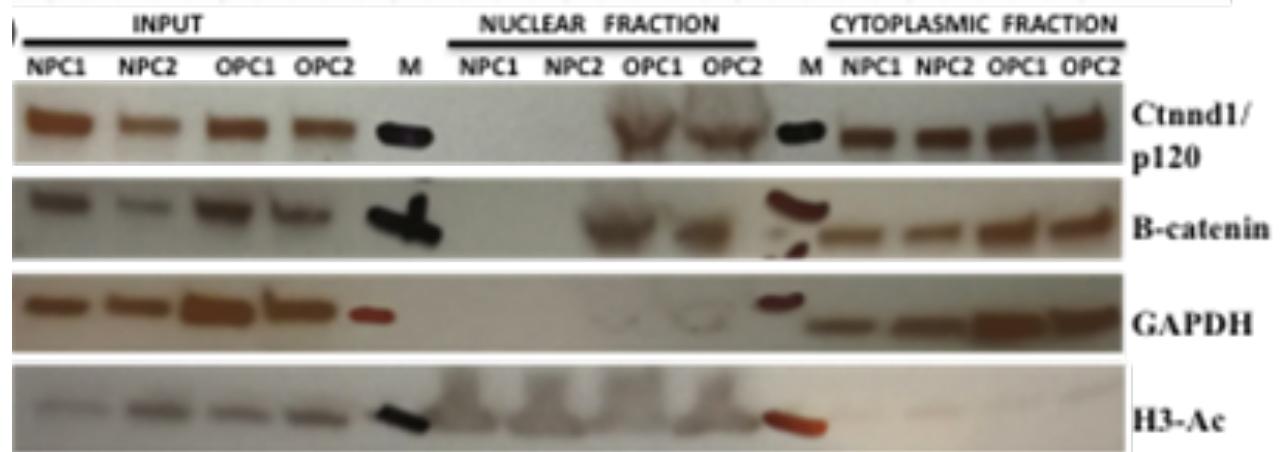
Nuclear  
Localization Signal





## *Regulation of protein motifs by differential splicing*

### Experimental validation



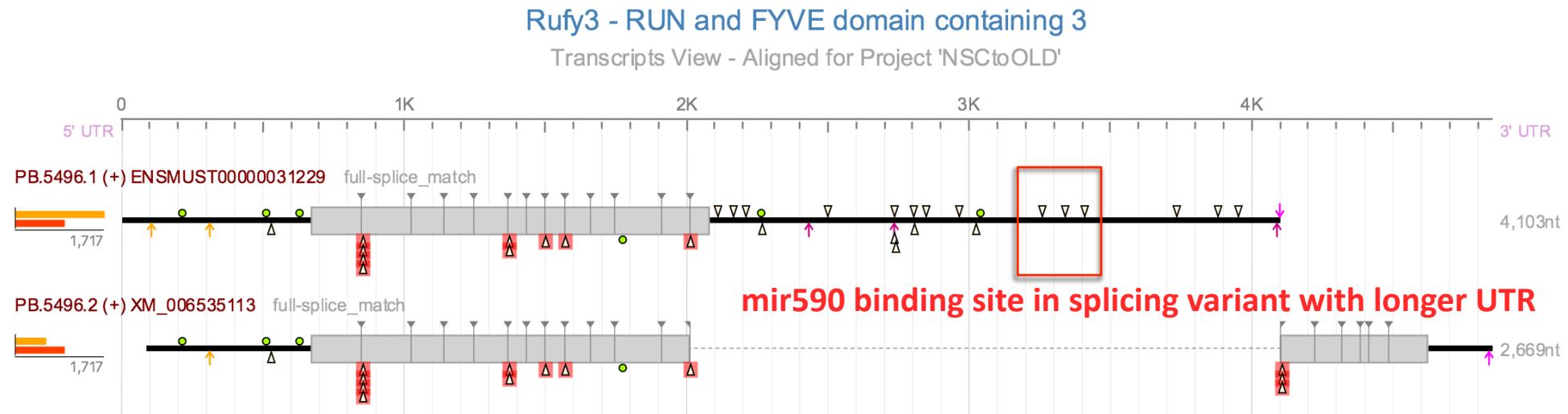
# Rufy3

Generation of neuronal polarity formation and axon growth



## Regulation of UTR motifs by differential splicing

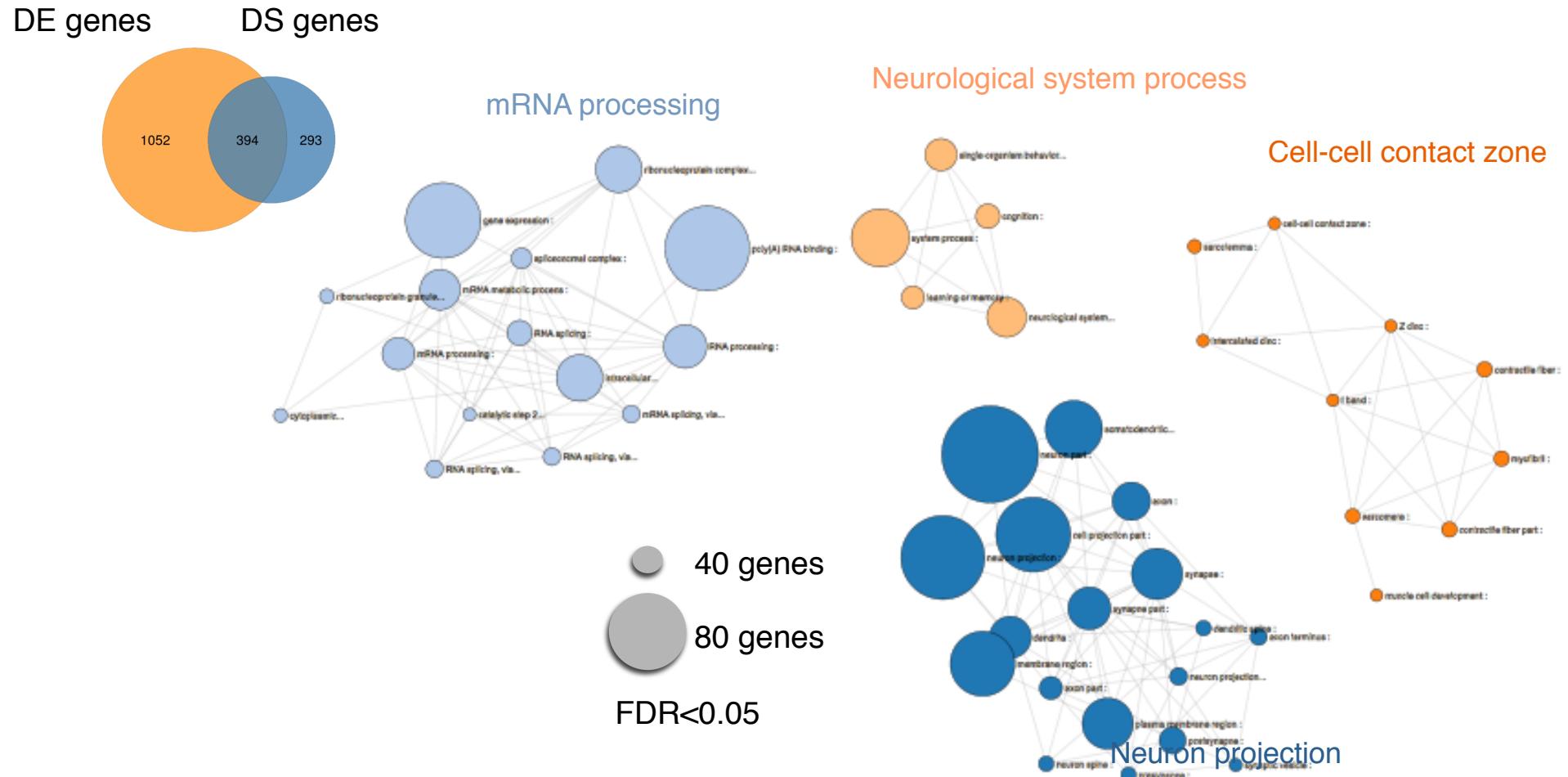
#	Gene	Feature	Feature Id	Position	FDSA Result	Q-Value	Favored Condi...	PodiumChg	TotalChg +
1	Rufy3	miRNA	mmu-miR-590-3p	T88641599-88...	DS	5.2611E-9	NSC	NO	48.79



Experimental validation ongoing:

- Analysis of miRNA 590 expression.
- Validation of the mirna binding site in Isoform 1 by miRNA pull-down assays.

# Functions enriched in Differential Spliced genes





Screenshot of the TAPPAS software interface showing a DE Analysis script log.

The interface includes a top navigation bar with buttons for Start, Input, Groups, DSA/DEA, Enrichment, Diversity, and a search bar. A sidebar on the left provides access to Overview, Input Data, Results, Stats, Summary, Distribution, Explorer, and Log.

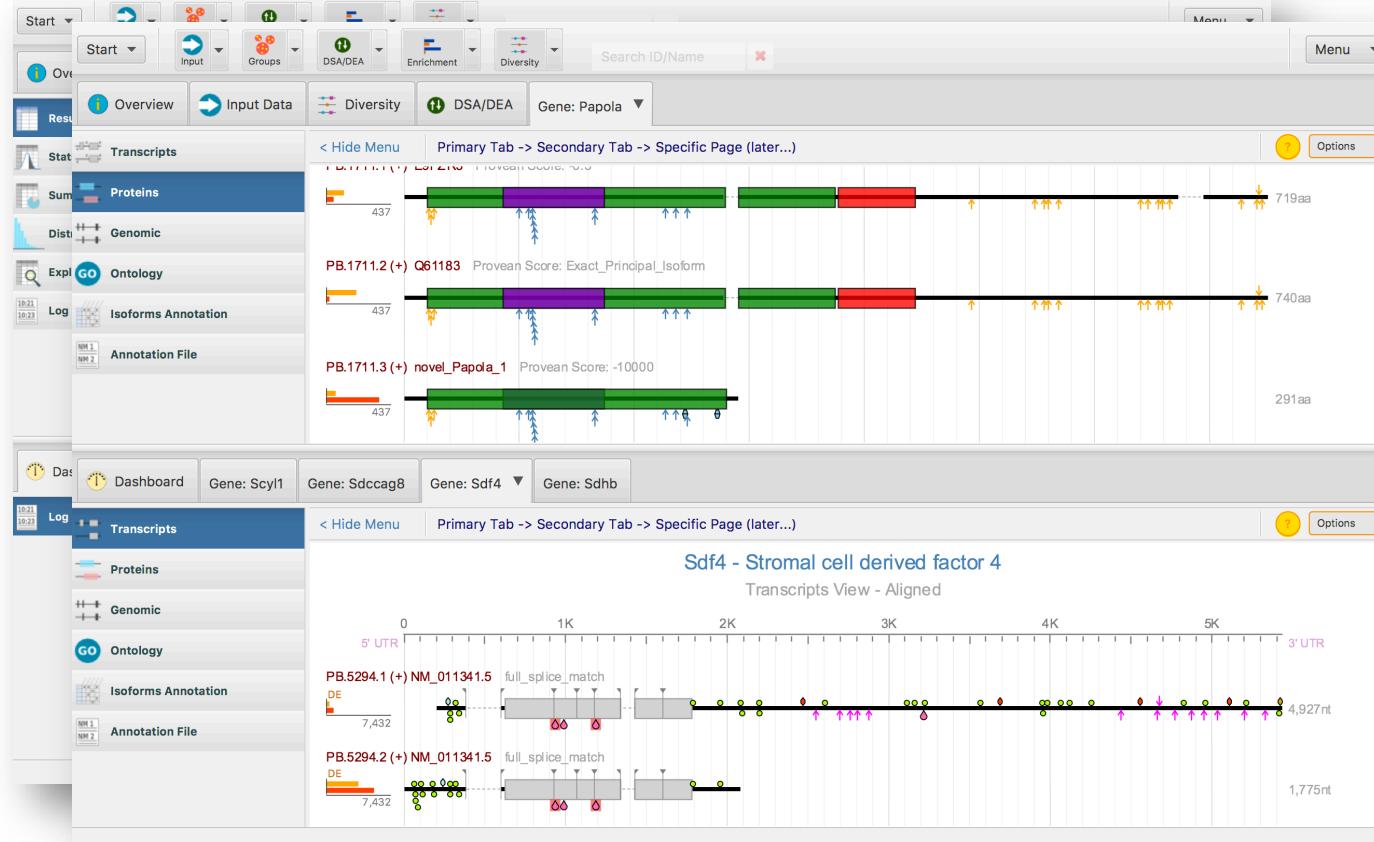
The main content area displays two tabs: "DSA/DEA" (selected) and "Log".

**DSA/DEA Tab:** This tab shows a table of gene analysis results. The columns include Name, DSA Result, DEA Result, DE, Total, DE, and Total. The data is as follows:

Name	DSA Result	DEA Result	DE	Total	DE	Total
0610007P14Rik	Not DS	DE	2	2	2	2
0610009B22Rik		Not DE	0	1	0	1
0610037L13Rik	DS	Not DE	1	4	1	4
1110002L01Rik		DE	1	1	1	1
1110004F10Rik	Not DS	DE	1	3	1	3
1110008L16Rik		Not DE	0	1	0	1
1110012L19Rik		Not DE	0	1	0	1
1110015O18Rik	DS	DE	0	0	2	2
1110032A03Rik	Not DS	DE	2	2	2	2
1110037F02Rik		Not DE	0	1	0	1

**Log Tab:** This tab displays the "Application Log" with the following entries:

```
15:07:56.533 - Running DE Analysis script:  
[usr/local/bin/Rscript, /var/folders/6z/svx83k_d0g94vb1vcb08ysdm000gn/T/t2go7103807439172492680.R, -nn, -rbiological, -l0, -p0.01, -a/Users/LorenaDeLaFu  
15:07:56.539 - DE Analysis process started, process id: java.lang.UNIXProcess@5e648c73  
15:13:49.785 - Detected DE analysis process stopped.  
22:19:41.649 - Application close request.  
13:01:15.756 - Initializing project data..  
13:01:15.791 - Loading annotation data from '/Users/LorenaDeLaFuente/Dropbox/Transcript2GO/annotFile_GENERIC_GMAP_GMST_ATGok_genomicRegion_NOmiRNAb  
13:01:15.811 - Reading annotation data index from /Users/LorenaDeLaFuente/t2goWorkspace/projects/Project_1875007079.t2goProject/ID/annotations.tsv.idx.  
13:01:15.820 - Annotation index file load completed OK.  
13:01:15.829 - Project data initialization completed.
```





Start Overview Input Groups DSA/DEA Enrichment Diversity Search ID/Name Menu

Start Overview Input Data Diversity DSA/DEA Gene: Papola Search ID/Name Menu

Start Overview Input Data DSA/DEA EA: GeneOntology Search ID/Name Menu

Start Overview Input Data DSA/DEA EA: GeneOntology Search ID/Name Menu

Overview Input Data DSA/DEA EA: GeneOntology Search ID/Name Menu

Results Transcripts Proteins Genomic Ontology Isoforms Annotation Log Nested Log Data Transcripts Proteins Genomic GO Ontology Isoforms Annotation Term Inclusion

DS Enriched Terms

GO:0032436 P positive regulation of proteasomal ubiquitin-dependent protein...  
 GO:0016290 F palmitoyl-CoA hydrolyase activity  
 GO:0043154 P negative regulation of cysteine-type endopeptidase activity in...  
 GO:0070084 F proline-rich region binding  
 GO:0045429 P positive regulation of nitric oxide biosynthetic process  
 GO:0042383 C sarcolemma  
 GO:0015459 F potassium channel regulator activity  
 GO:0030054 C cell junction  
 GO:0005938 C cell cortex  
 GO:0031594 C neuromuscular junction  
 GO:0030173 C integral component of Golgi membrane  
 GO:0065004 P protein-DNA complex assembly  
 GO:0019903 F protein phosphatase binding  
 GO:0000145 Exocyst  
 GO:0003779 F mRNA binding

Total Enriched Terms: 33 DS Genes: 1549 DS Isoforms: 5448 NOTDS Genes: 1457 NOTDS Isoforms: 1022

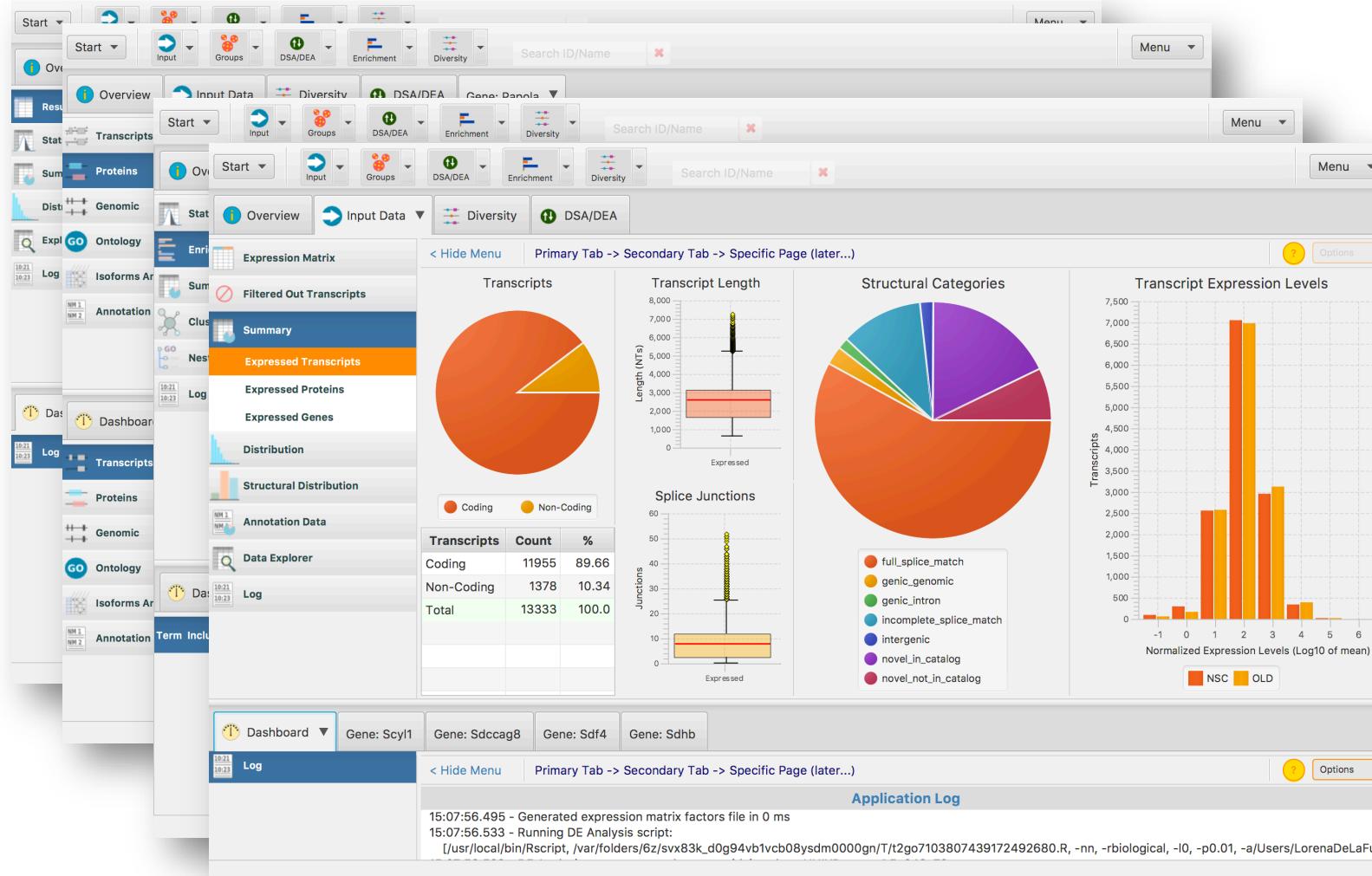
Gene: Tardbp Gene: Papola Gene: Dnm1 Gene: Capzb Term: GO:0005525

Genes with GO:0005525 - Enriched

Gene	DE Type	Isoform(s)	
		Total	with Term
5430435G22Rik	AIE	3	3
Adss	AIE	2	2
Anxa6	AIE	6	4
Arf1	AIE	2	2

Isoforms with GO:0005525 for Selected Gene(s)

Gene	DE Type	Isoform	Length	Up/Down	L2 FoldChg	Probability
5430435...	AIE	PB.282.1	2568	DOWN	-5.62	1.0
5430435...	AIE	PB.282.2	2347		-2.67	0.9843
5430435...	AIE	PB.282.3	1663	DOWN	-5.86	1.0



# Acknowledgements



**UF**

William Farmerie  
Eric Triplett  
Lauren McIntyre

**UCI**

Ali Mortazavi

**Pacbio**

Liz Tseng

**CIPF**

Victoria Moreno  
Susana Rodriguez

