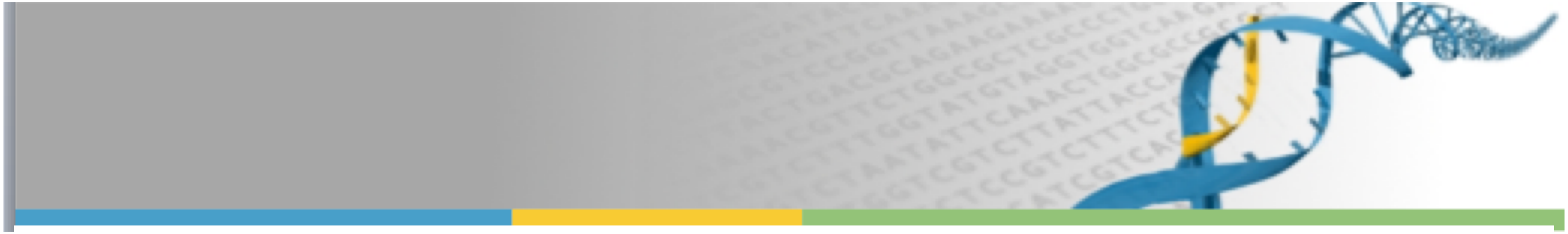


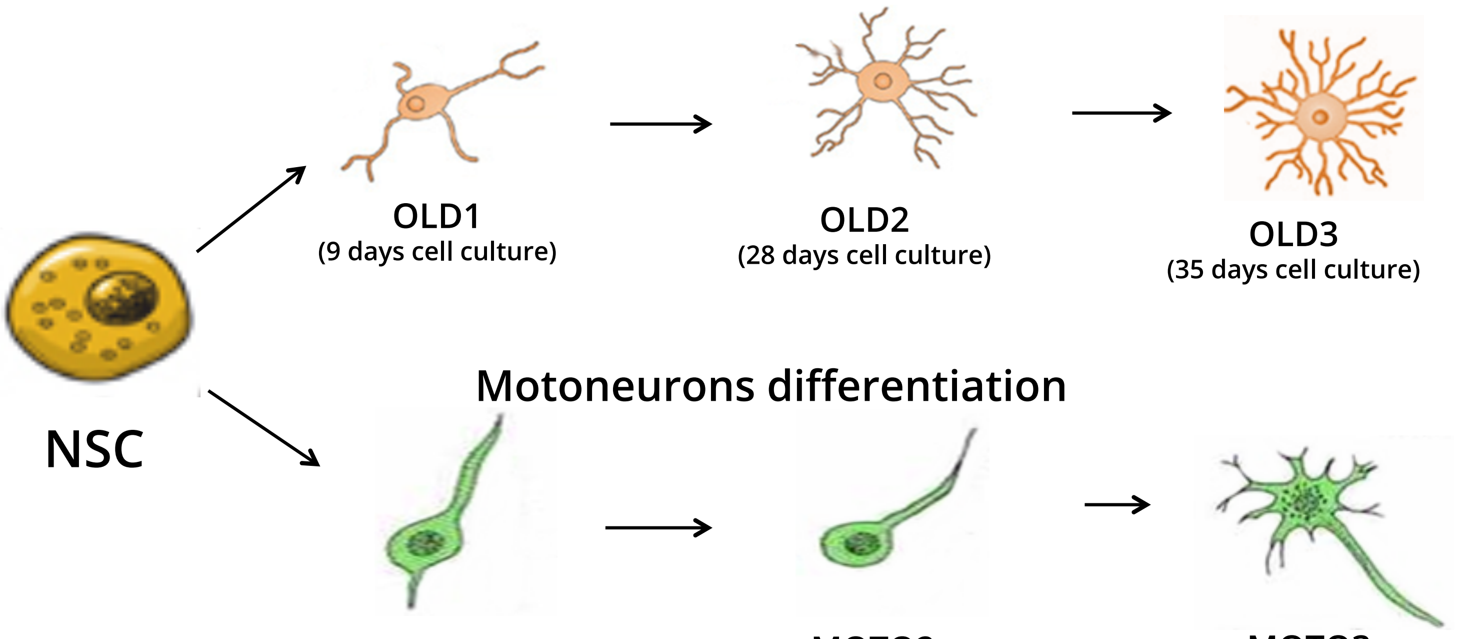
SQANTI and TAPP**AS**:
Making Sense of Iso-Seq Data

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

PRINCIPE FELIPE
CENTRO DE INVESTIGACION

UF
UNIVERSITY of FLORIDA

Functional Implications of Differential Splicing?



Oligodendrocytes differentiation

NSC

OLD1
(9 days cell culture)

OLD2
(28 days cell culture)

OLD3
(35 days cell culture)

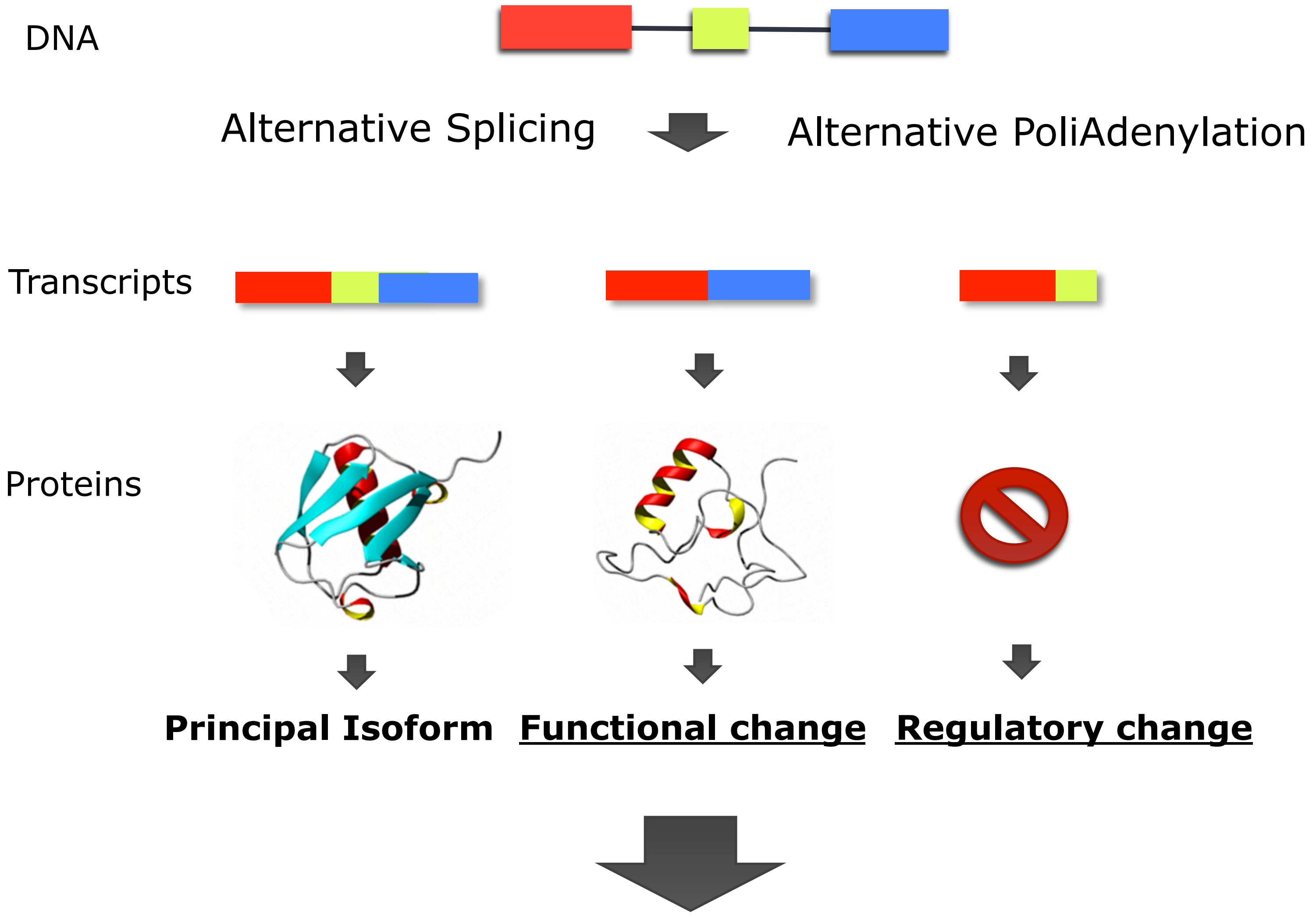
Motoneurons differentiation

MOTO1
(9 days cell culture)

MOTO2
(15 days cell culture)

MOTO3
(35 days cell culture)

Functional Implications of Differential Splicing



DNA

Alternative Splicing

Alternative PoliAdenylation

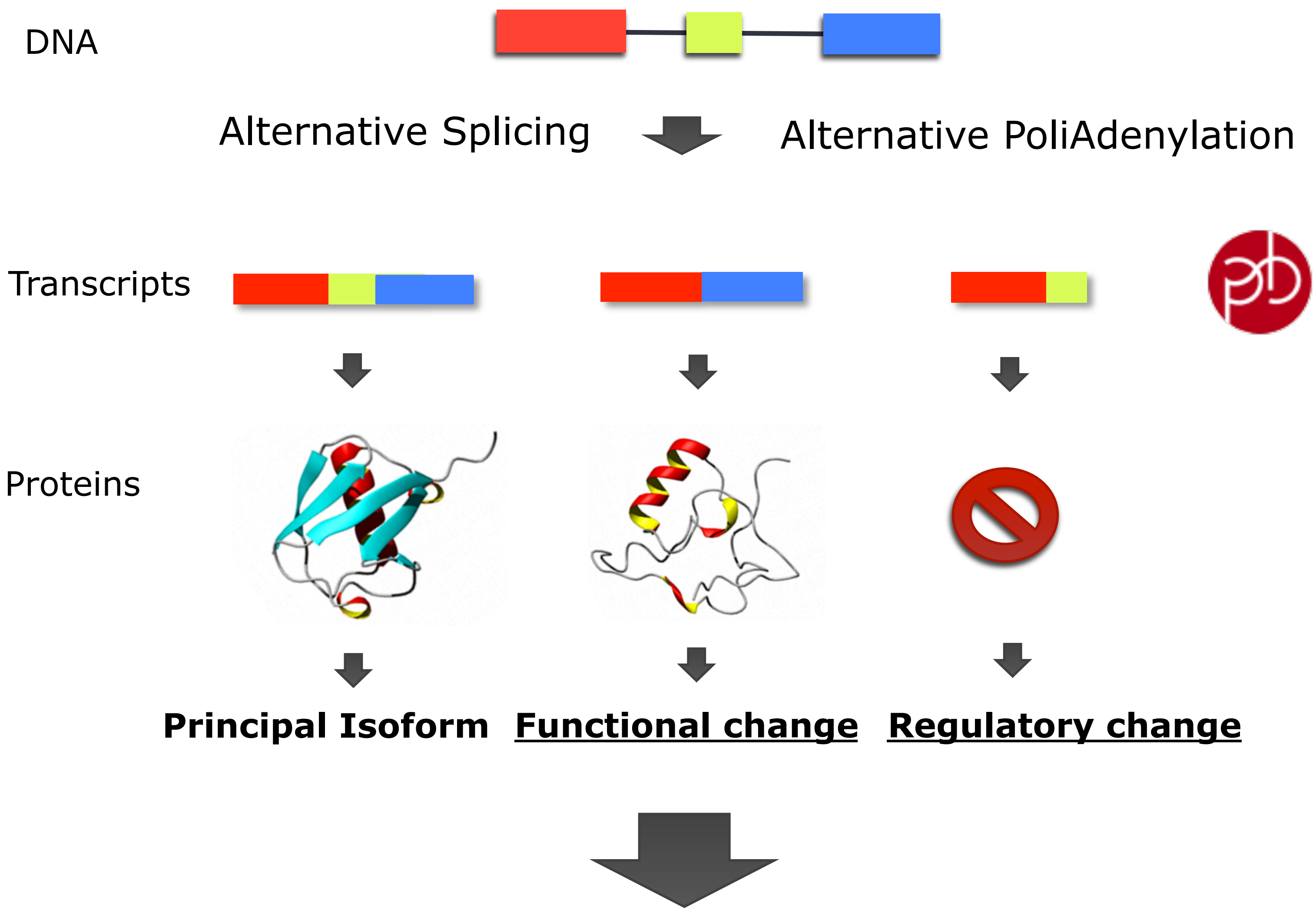
Transcripts

Proteins

Principal Isoform **<u>Functional change</u>** **<u>Regulatory change</u>**

Functional Impact of Differential Splicing

Functional Implications of Differential Splicing



DNA

Alternative Splicing

Alternative PoliAdenylation

Transcripts

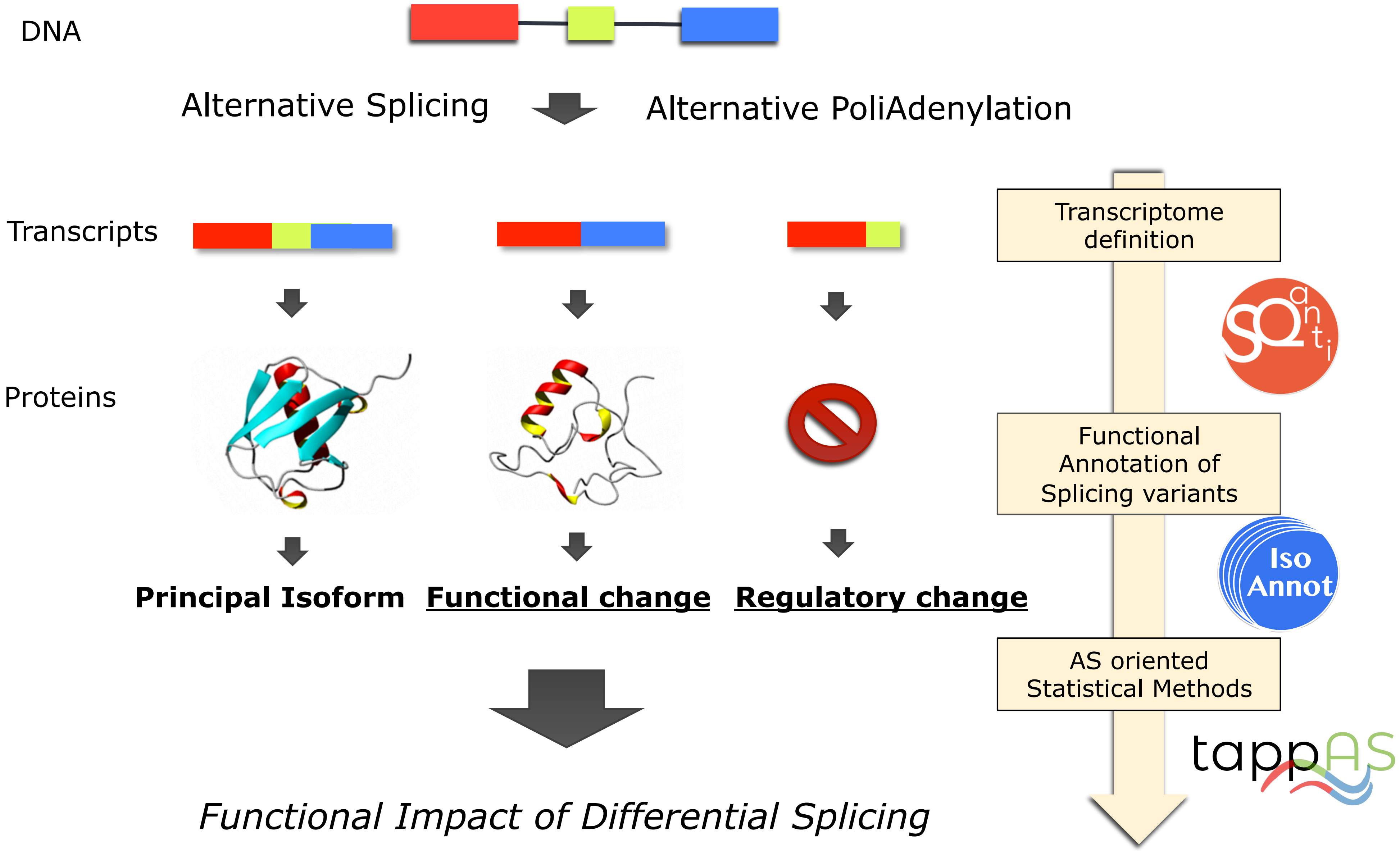
Proteins

Principal Isoform **<u>Functional change</u>** **<u>Regulatory change</u>**

PACIFIC BIOSCIENCES™

Functional Impact of Differential Splicing

Functional Implications of Differential Splicing



DNA

Alternative Splicing

Alternative PoliAdenylation

Transcripts

Proteins

Principal Isoform

<u>Functional change</u>

<u>Regulatory change</u>

Functional Impact of Differential Splicing

Transcriptome definition

SQanti

Functional Annotation of Splicing variants

Iso Annot

AS oriented Statistical Methods

tappAS

Structural and **Q**uality **An**notation of **T**ranscript Isoforms

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 Research, in press.***

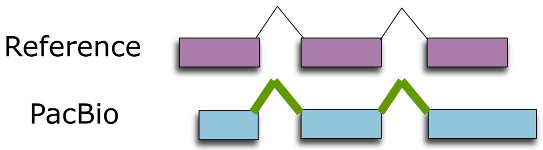
Transcriptome characterization

1. Classification

Known Isoforms

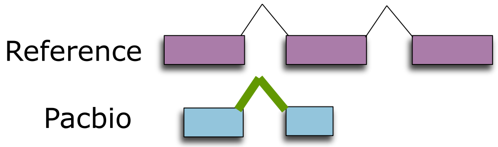
Full-Splice Match

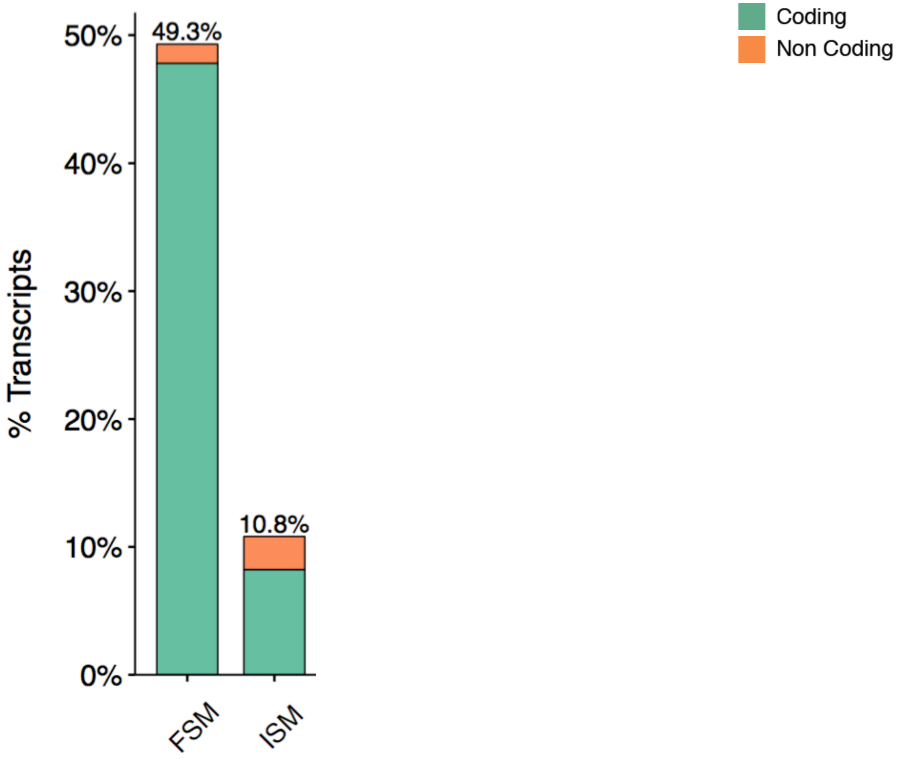
FSM



Incomplete-Splice Match

ISM





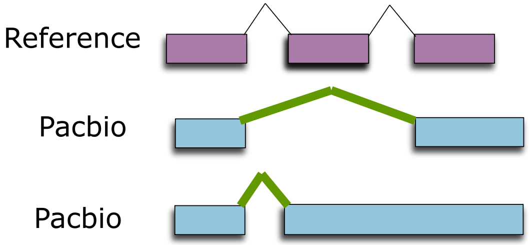
Transcriptome characterization

1. Classification

Novel Isoforms – Known genes

Novel In catalog

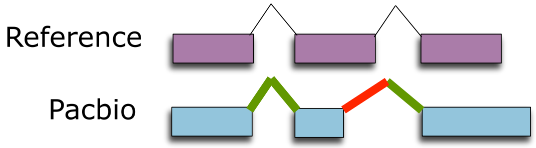
NIC



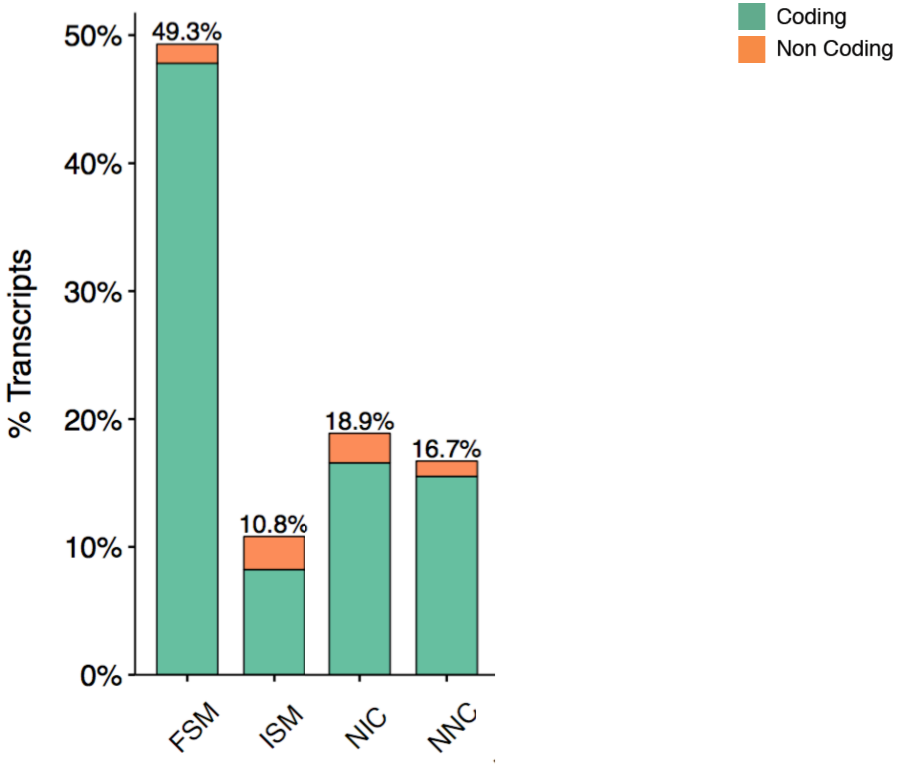
Known donors and acceptors

Novel Not in Catalog

NNC

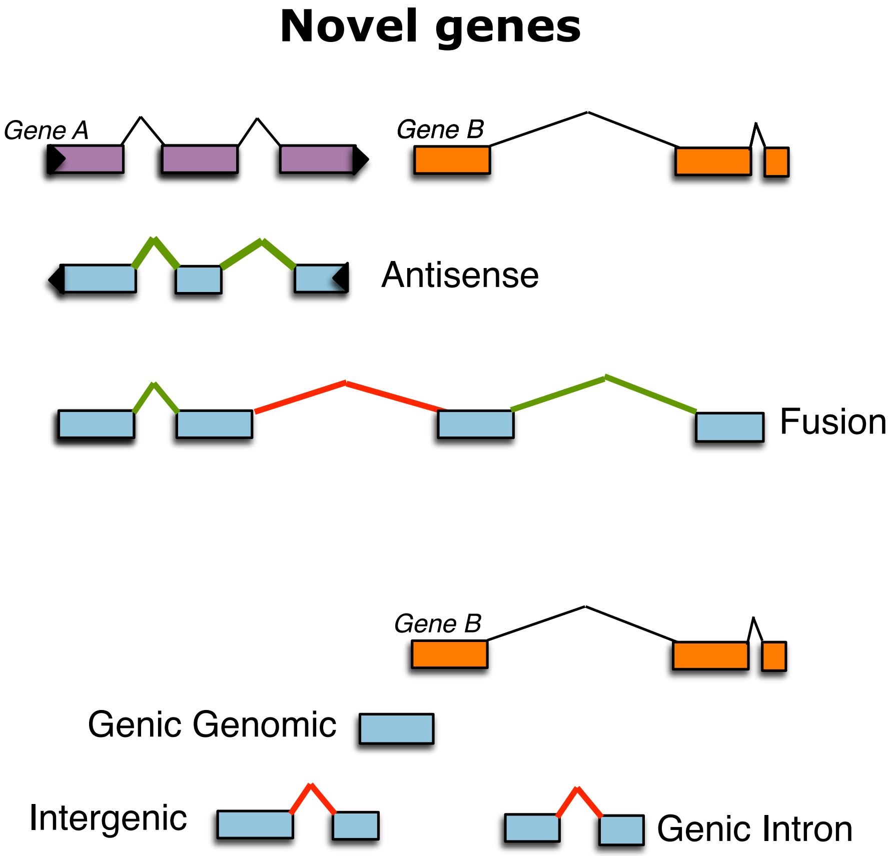


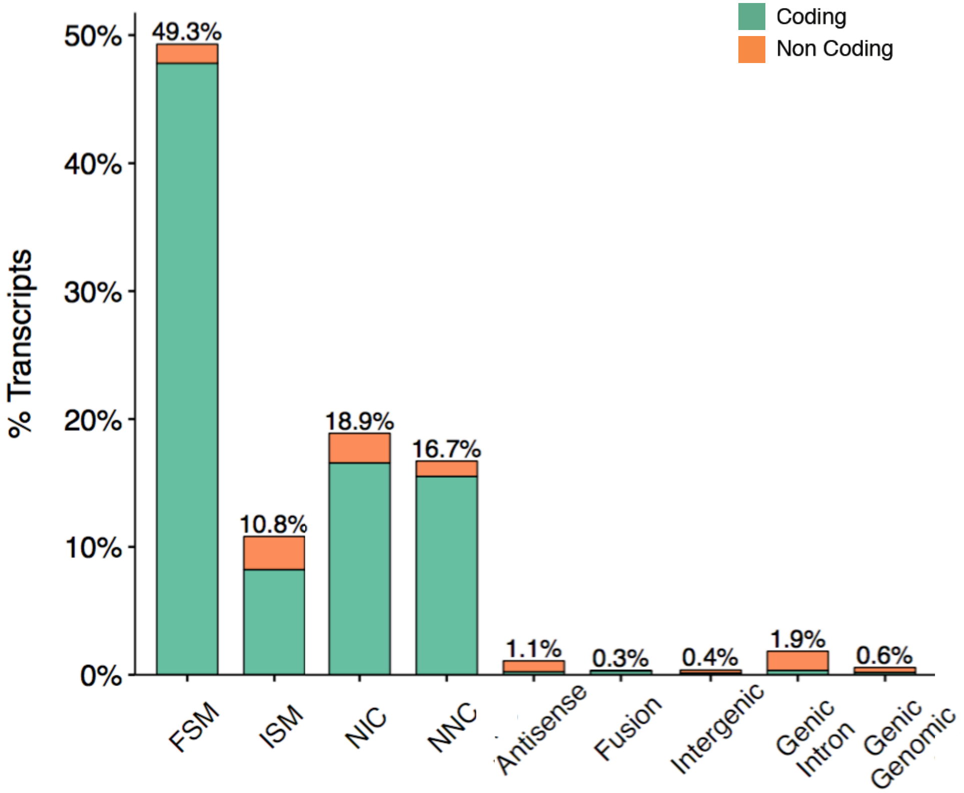
Novel donor or acceptor



Transcriptome characterization

1. Classification



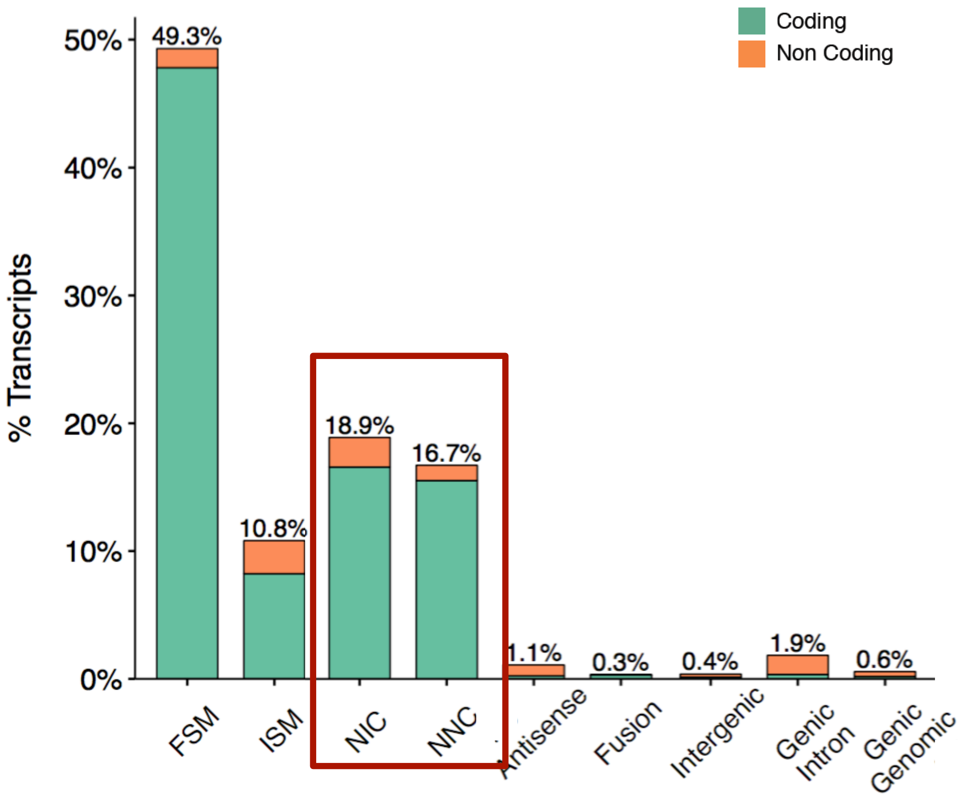


Transcriptome characterization

1. Classification

35 % of novel isoforms in mouse…

Are all of them real?



Transcriptome characterization

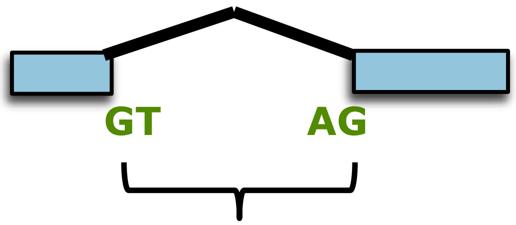
1. Classification

2. **QC descriptors**

Transcriptome characterization

1. Classification

2. **QC descriptors: SJ canonical status**



Canonical splice junction

≈ 98,7 % of canoncal SJ in mammalian*

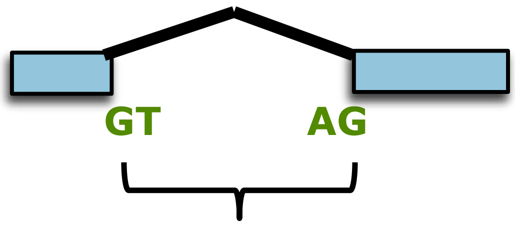
97,7 % of total splice junctions in our neural transcriptome are canonical

*Burset et al, 2000

Transcriptome characterization

1. Classification

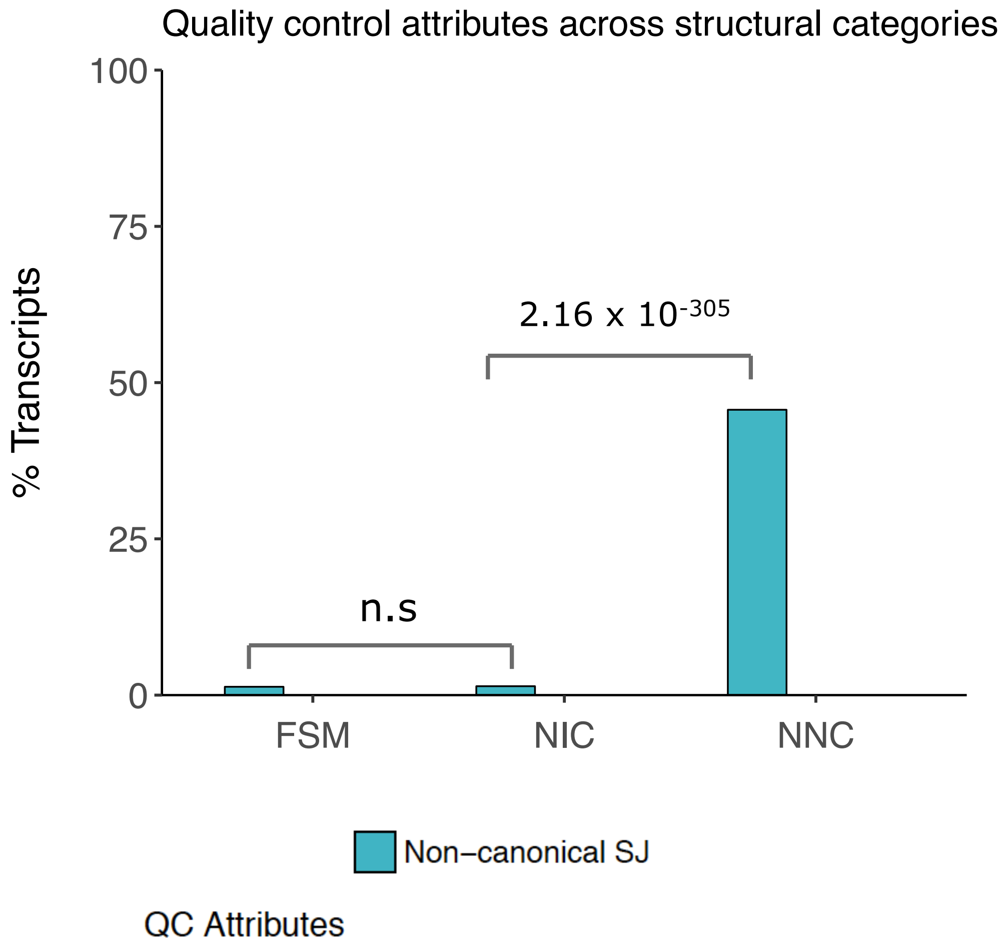
2. **QC descriptors: SJ canonical status**



Canonical splice junction

≈ 98,7 % of canoncal SJ in mammalian*

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

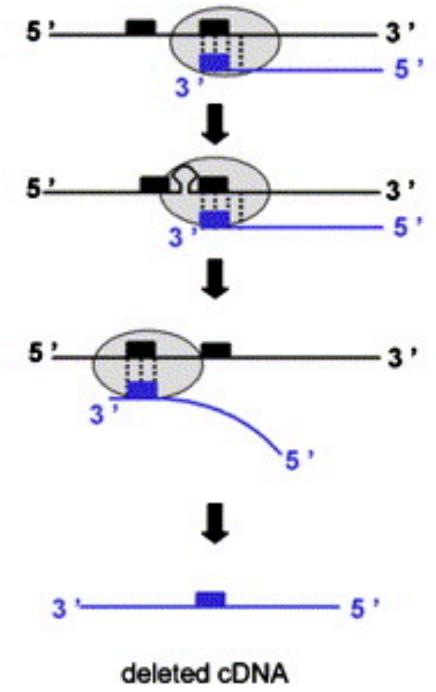


*Burset et al, 2000

Transcriptome characterization

1. Classification

2. QC descriptors: RT-switching

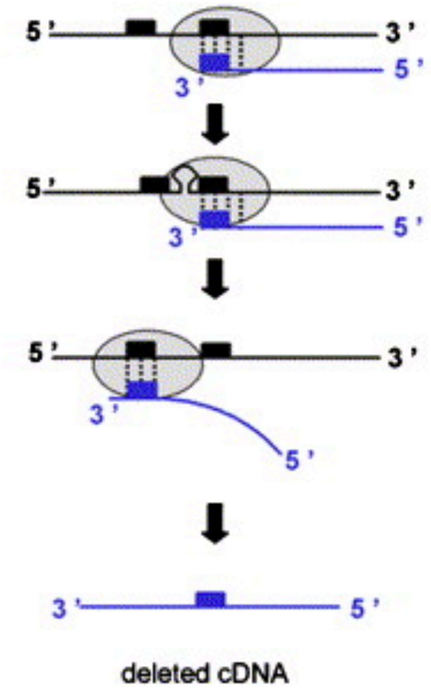


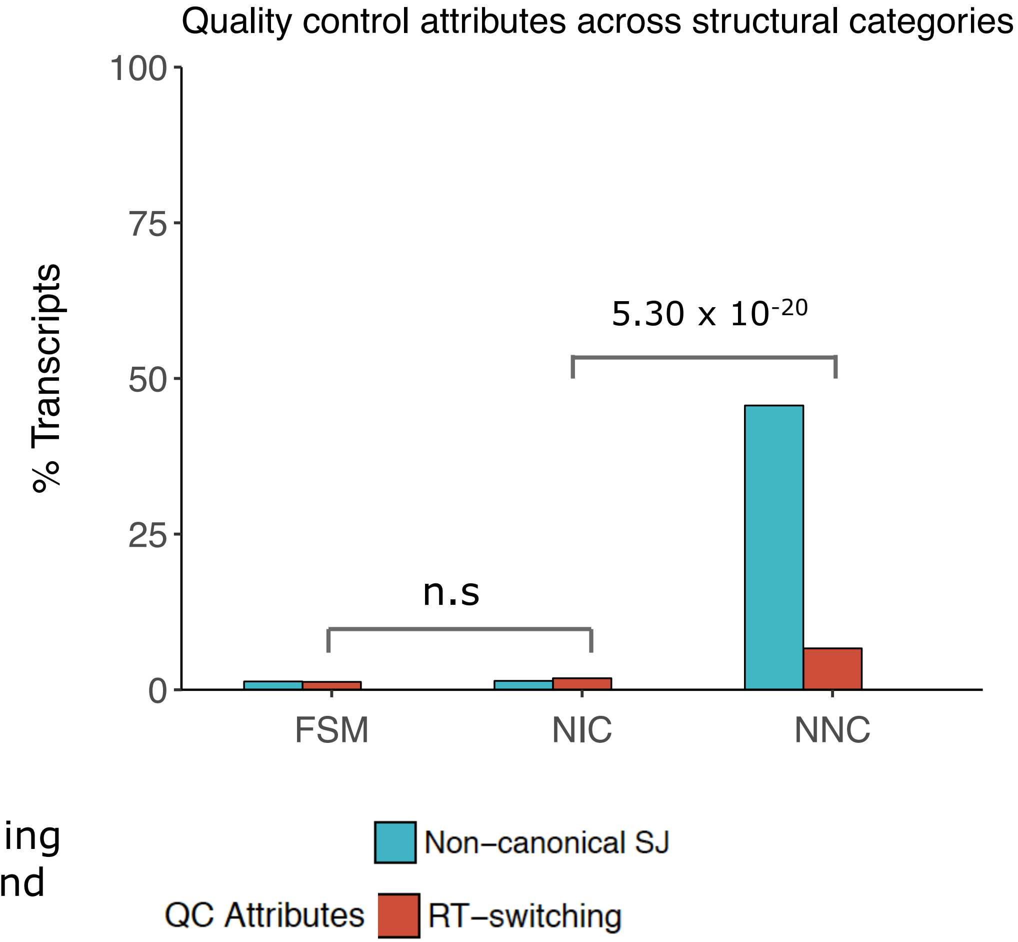
- Reverse transcriptase template switching
- Caused by RNA secondary structure and repeated regions.
- Appears as novel splice junctions

Transcriptome characterization

1. Classification

2. QC descriptors: RT-switching



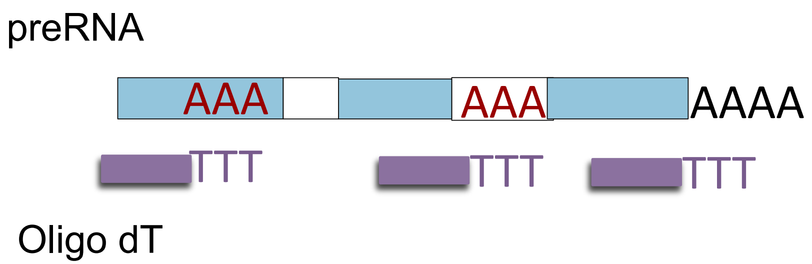


- Reverse transcriptase template switching
- Caused by RNA secondary structure and repeated regions.
- Appears as novel splice junctions

Transcriptome characterization

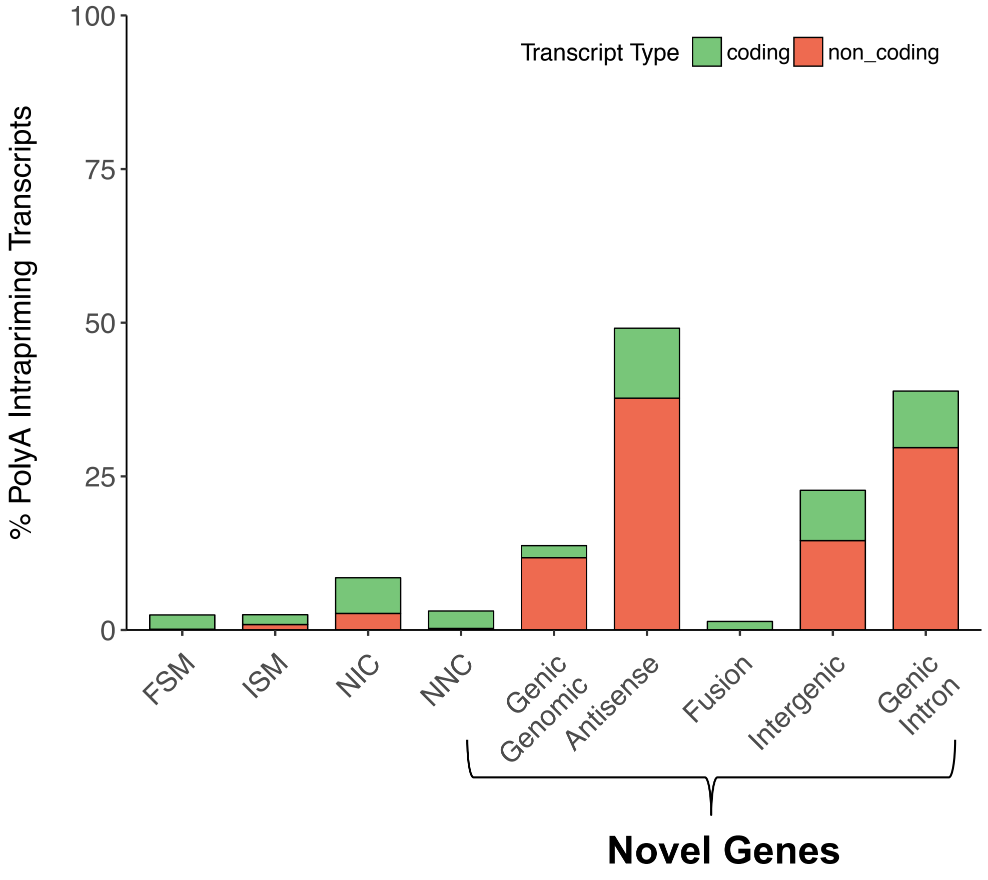
1. Classification

2. **QC descriptors: PolyA intra-priming**



- oligodT can prime outside polyA tail in A rich regions inside transcribed regions.

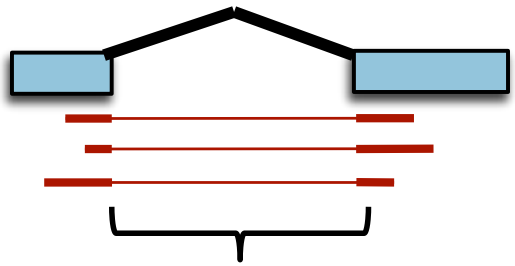
- We looked for transcripts showing >= 80% Adenines in the 20 nts downtream "detected" 3' end



Transcriptome characterization

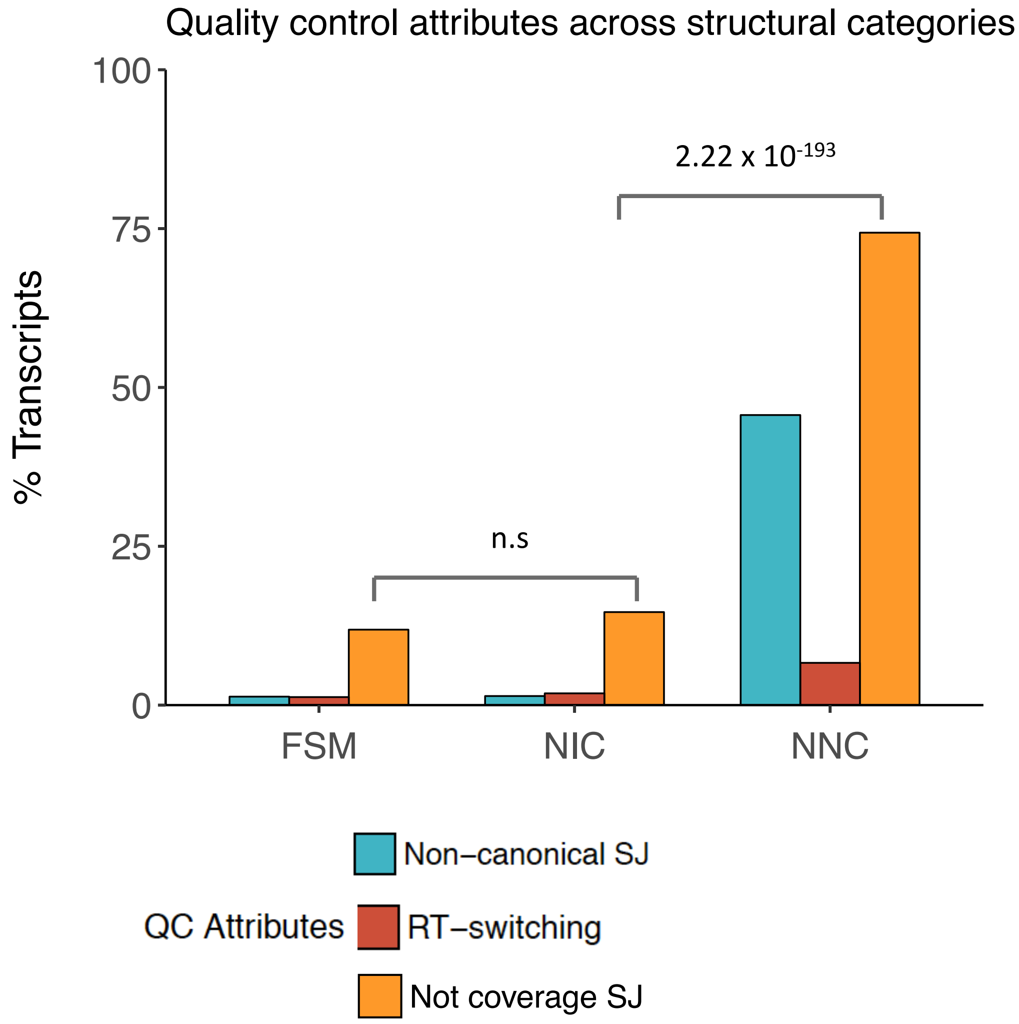
1. Classification

2. **QC descriptors: SJ support**



Supported splice junction

Illumina Reads from same cDNA sequenced by PacBio



SQANTI

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

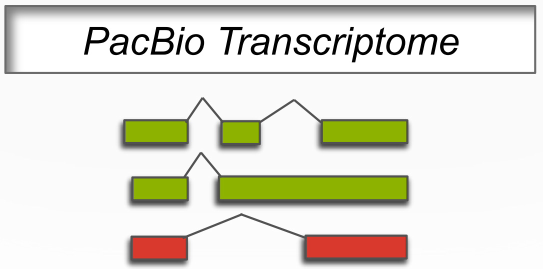
1. Junction Classification
    1. Novel/Known
    2. Splice site motif

2. Structural characteristics
    1. Diffterence to nearest ref. donor
    2. Diffterence to nearest ref. acceptor
    3. Bite

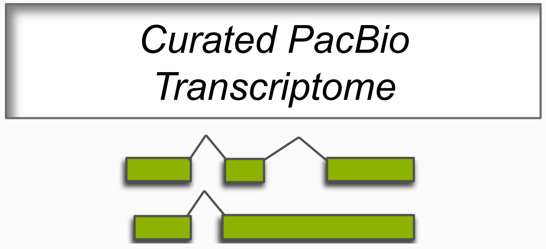
3. Quality Control attributes
    1. Canonical
    2. Rts_junction
    3. Indel near junc

4. Support
    1. Samples with cov
    2. Total coverage
    3. Coverage per sample

Filtering out artifact isoforms



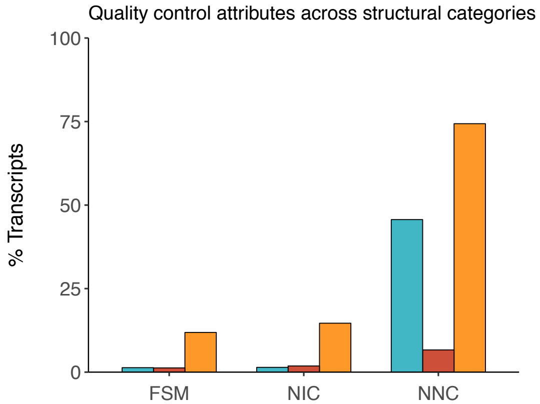
Machine learning
with SQANTI descriptors

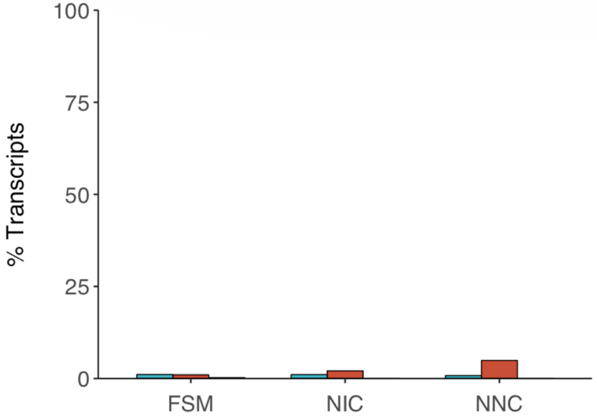


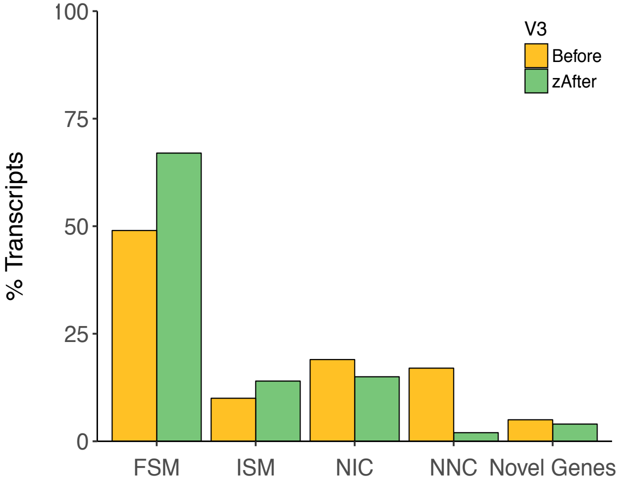
🟦 Non−canonical SJ

🟥 RT−switching

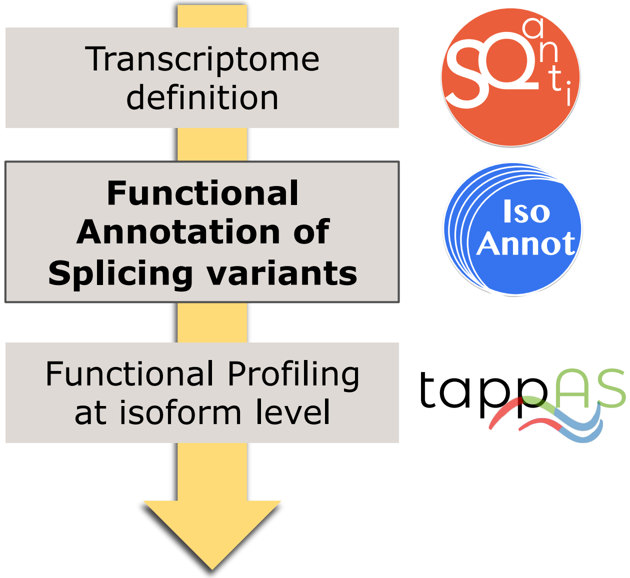
🟧 Not coverage SJ







Functional Annotation of Splicing Variants



Transcriptome definition

Functional Annotation of Splicing variants

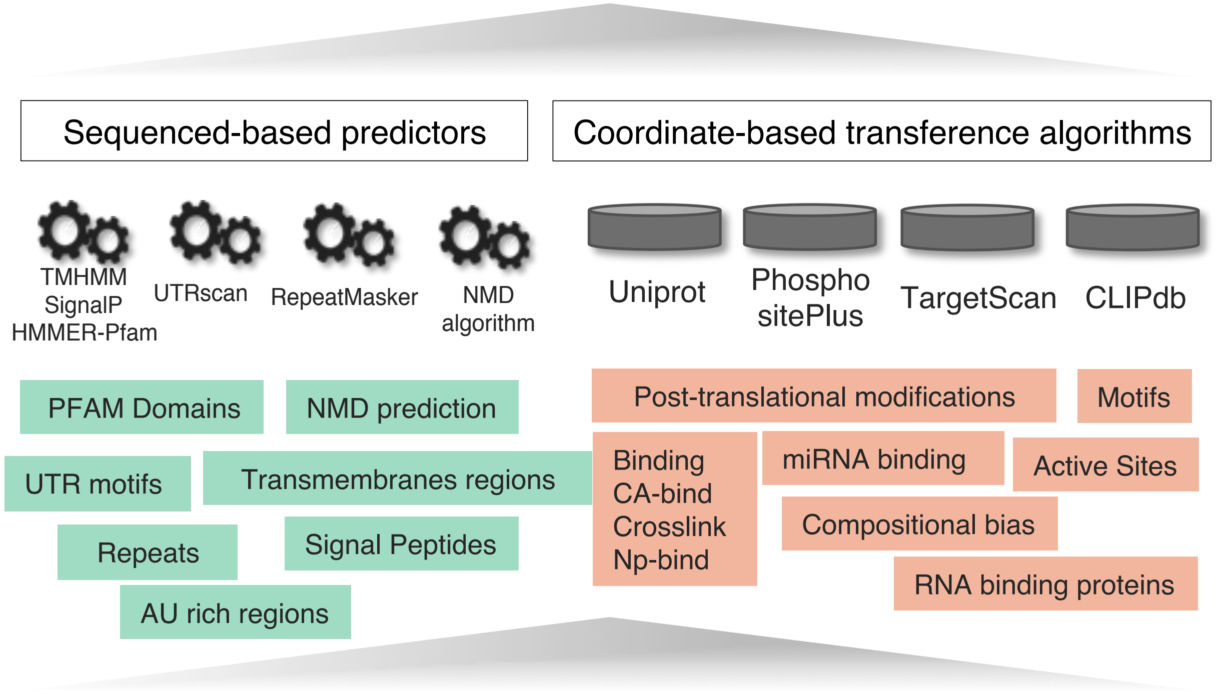
Functional Profiling at isoform level

SQANTI

Iso Annot

tappAS

Splicing Variants (known and novel)



Sequenced-based predictors

Coordinate-based transference algorithms

TMHMM SignalP HMMER-Pfam

UTRscan

RepeatMasker

NMD algorithm

Uniprot

Phospho sitePlus

TargetScan

CLIPdb

PFAM Domains

NMD prediction

UTR motifs

Transmembranes regions

Repeats

Signal Peptides

AU rich regions

Post-translational modifications

Motifs

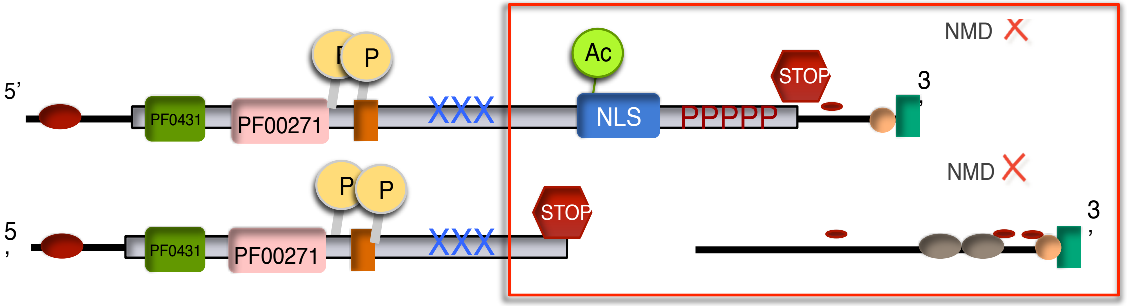
Binding CA-bind Crosslink Np-bind

miRNA binding

Active Sites

Compositional bias

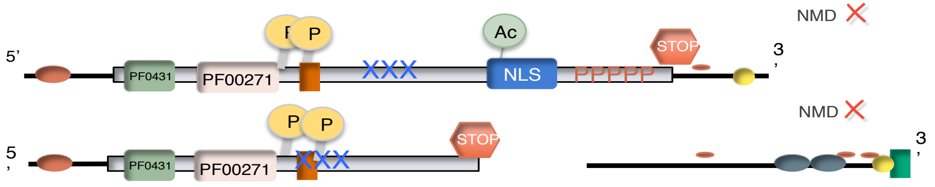
RNA binding proteins



Domain and Motif annotation at Isoform Resolution

Functional Profiling at Isoform Level

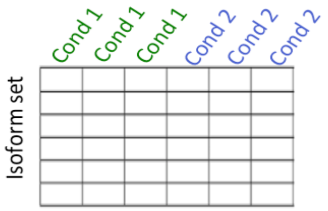
Structural Annotation and Functional Annotation



Reference Annotation provided

OPTIONAL: User-defined

Isoform quantification



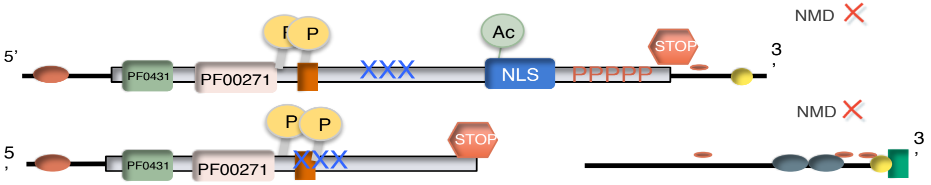
User-defined

INPUT

tappAS

Functional Profiling at Isoform Level

Structural Annotation and Functional Annotation

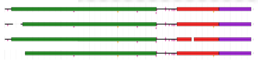


Reference Annotation provided

OPTIONAL: User-defined

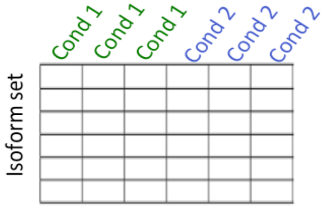
Module 1

Visualization Interface



tappAS

Isoform quantification

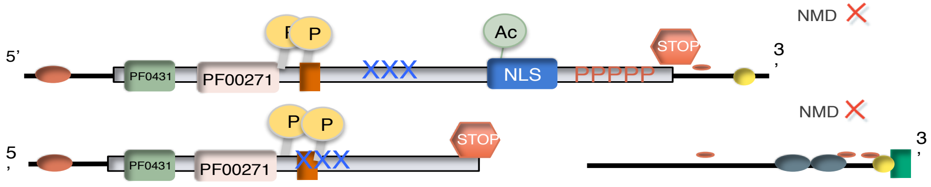


User-defined

INPUT

Functional Profiling at Isoform Level

Structural Annotation and Functional Annotation

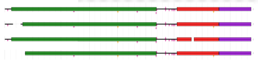


Reference Annotation provided

OPTIONAL: User-defined

Module 1

Visualization Interface

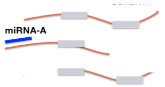


Module 2

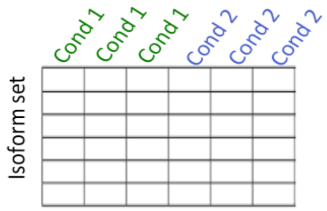
Functional Diversity Analysis (FDA)

Motif Diversity

CDS/UTRs Diversity



Isoform quantification

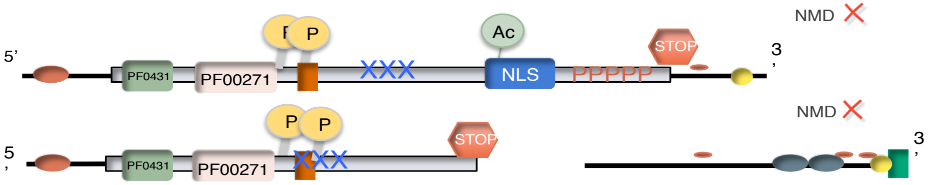


User-defined

INPUT

Functional Profiling at Isoform Level

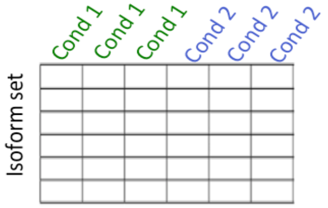
Structural Annotation and Functional Annotation



Reference Annotation provided

OPTIONAL: User-defined

Isoform quantification

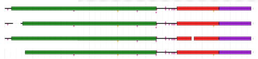


User-defined

INPUT

Module 1

Visualization Interface

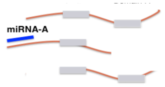


Module 2

Motif Diversity

CDS/UTRs Diversity

Functional Diversity Analysis (FDA)



tappAS

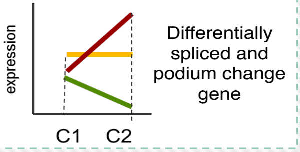
Module 3

Differentially Spliced genes

Mayor Isoform Switching

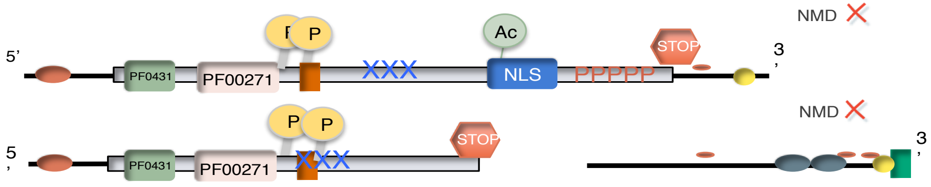
Differentially expressed genes/transcripts/ORFs

Differential Analysis (DSA/DEA)



Functional Profiling at Isoform Level

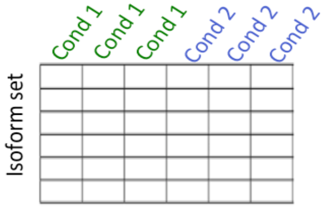
Structural Annotation and Functional Annotation



Reference Annotation provided

OPTIONAL: User-defined

Isoform quantification

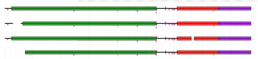


User-defined

INPUT

Module 1

Visualization Interface

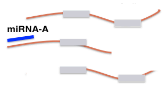


Module 2

Motif Diversity

CDS/UTRs Diversity

Functional Diversity Analysis (FDA)



tappAS

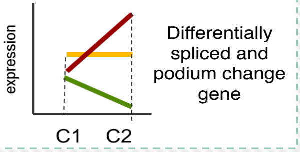
Module 3

Differentially Spliced genes

Mayor Isoform Switching

Differentially expressed genes/transcripts/ORFs

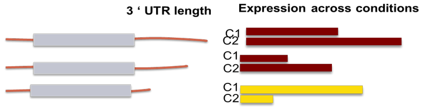
Differential Analysis (DSA/DEA)



Integrative methods

Module 4

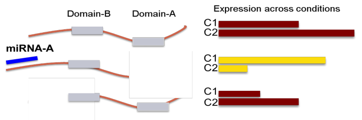
UTR length regulation



UTR Differential Splicing (UtrDS)

Regulation of annotated elements

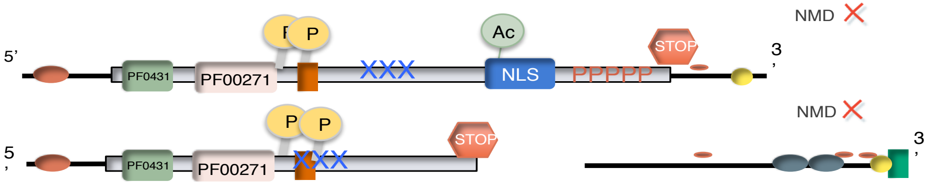
Motifs,Domains, miRNA binding sites



Motif Differential Splicing (mDS)

Functional Profiling at Isoform Level

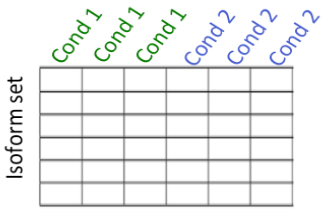
Structural Annotation and Functional Annotation



Reference Annotation provided

OPTIONAL: User-defined

Isoform quantification

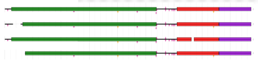


INPUT

User-defined

Module 1

Visualization Interface

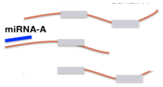


Module 2

Motif Diversity

CDS/UTRs Diversity

Functional Diversity Analysis (FDA)



tappAS

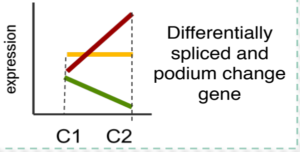
Module 3

Differentially Spliced genes

Mayor Isoform Switching

Differentially expressed genes/transcripts/ORFs

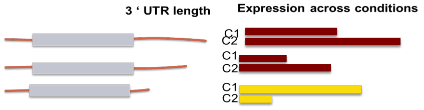
Differential Analysis (DSA/DEA)



Integrative methods

Module 4

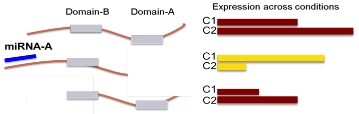
UTR length regulation



UTR Differential Splicing (UtrDS)

Regulation of annotated elements

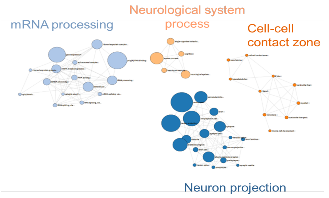
Motifs,Domains, miRNA binding sites



Motif Differential Splicing (mDS)

Module 5

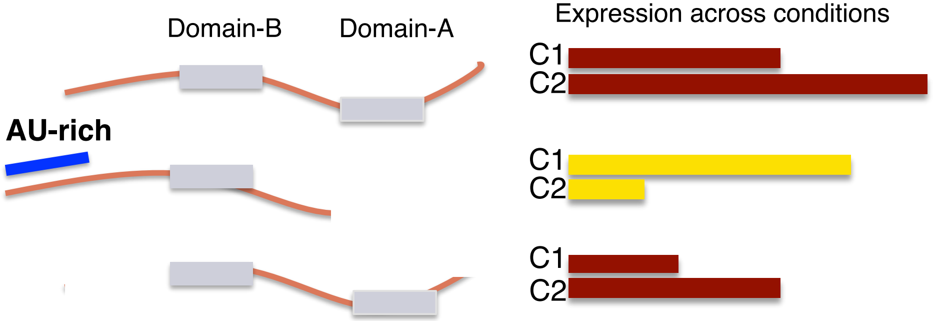
Functional enrichment over any annotated category



Functional Enrichment and Gene Set Analysis (FEA/GSA)

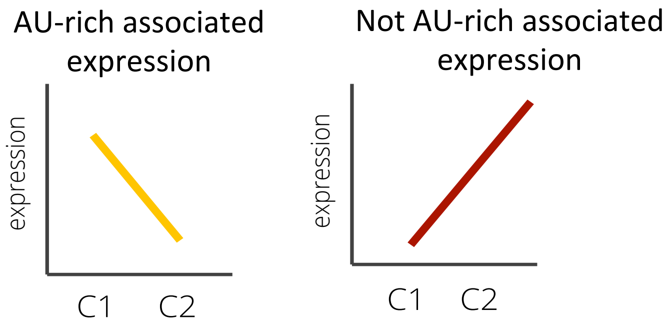
Motif and Feature Differential Splicing

Gene A



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

GLM



DS AU-rich region

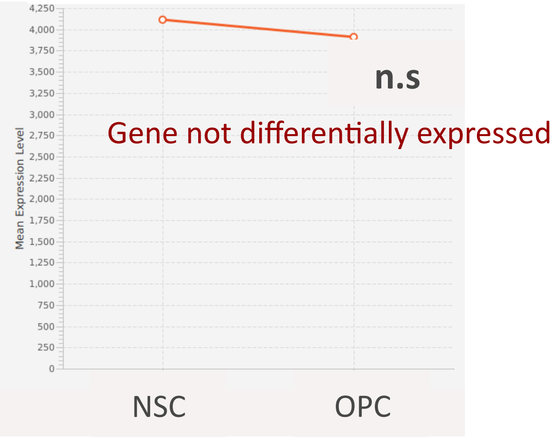
AU-rich element favored in condition 1 by DS

Ctnnd1 – p120 – catenin delta 1

Involved in cell-cell adhesion and signal transduction

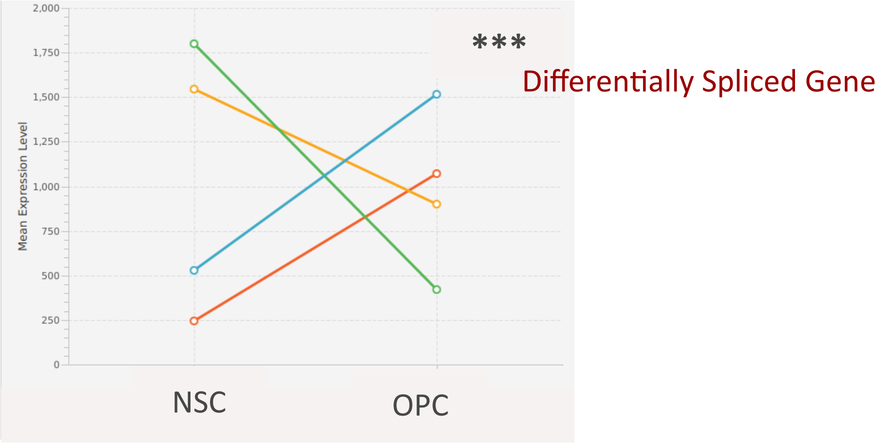
Regulation of protein motifs by differential splicing

Gene Expression



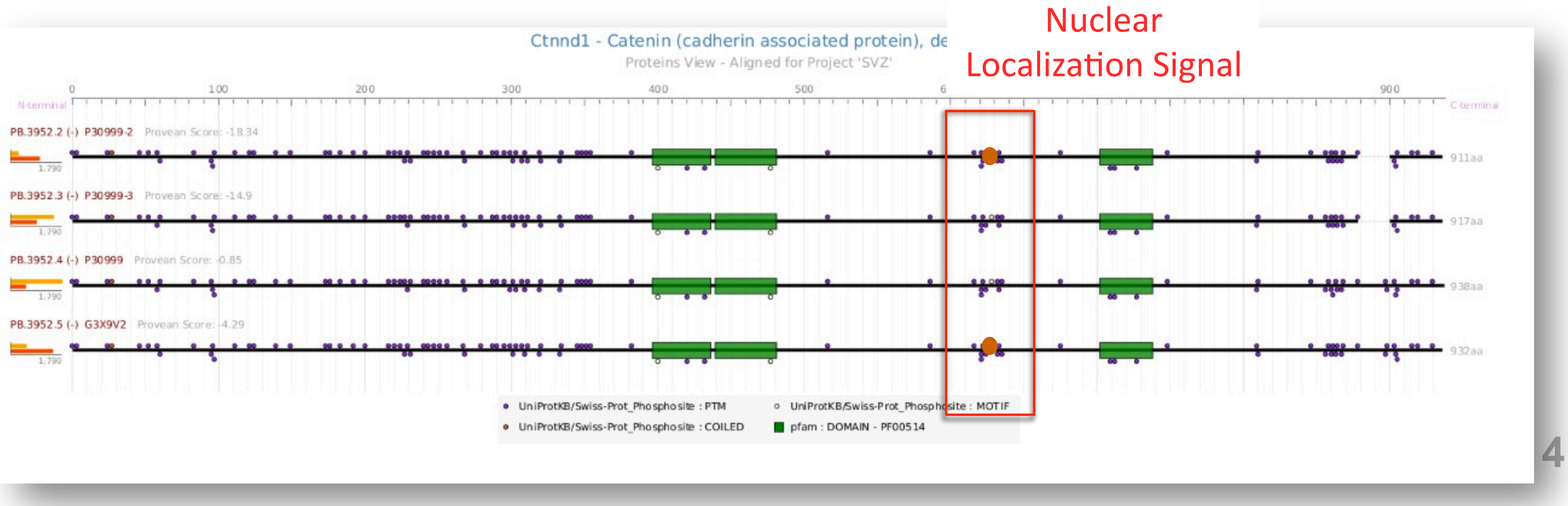
Gene not differentially expressed

Transcripts Expression



Differentially Spliced Gene

Functional impact?


Nuclear Localization Signal

Ctnnd1 - Catenin (cadherin associated protein), de

Proteins View - Aligned for Project 'SVZ'

PB.3952.2 (-) P30999-2 Provean Score: -18.34

PB.3952.3 (-) P30999-3 Provean Score: -14.9

PB.3952.4 (-) P30999 Provean Score: -0.85

PB.3952.5 (-) G3X9V2 Provean Score: -4.29

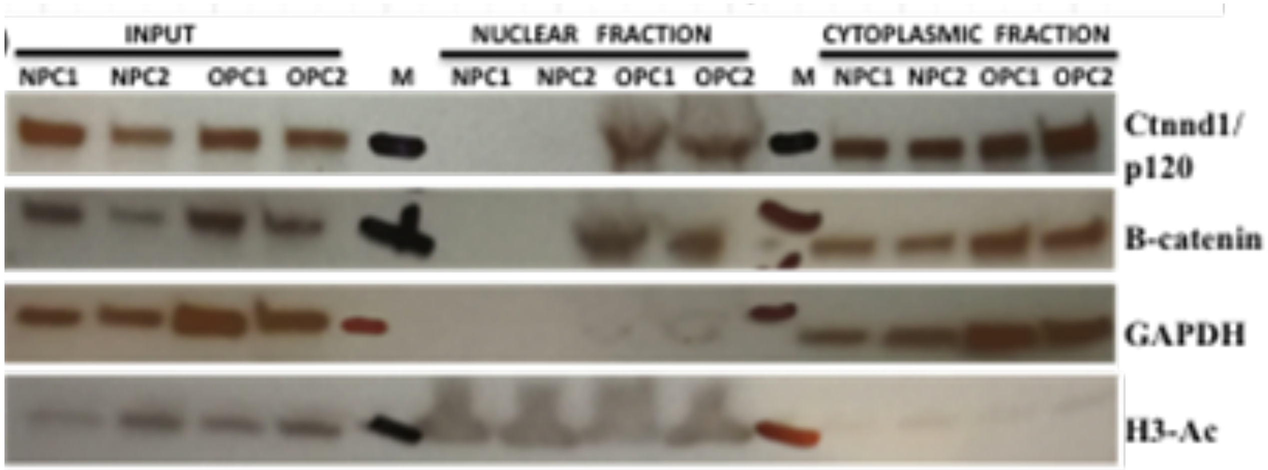
● UniProtKB/Swiss-Prot_Phosphosite : PTM
○ UniProtKB/Swiss-Prot_Phosphosite : MOTIF
● UniProtKB/Swiss-Prot_Phosphosite : COILED
🟩 pfam : DOMAIN - PF00514

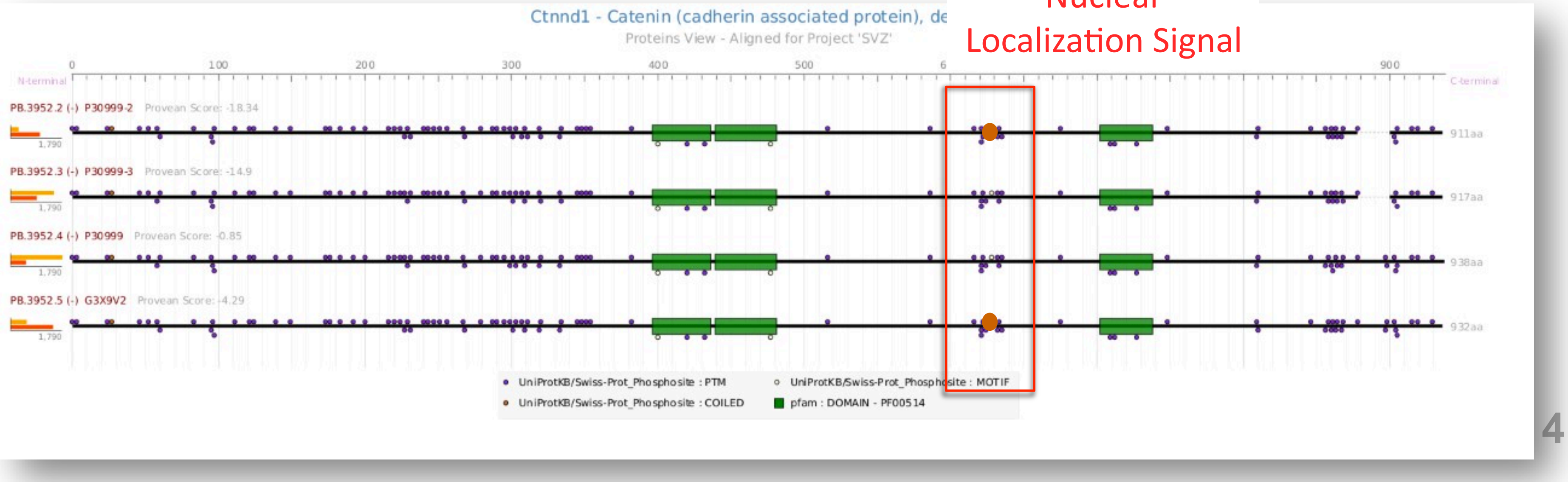
Ctnnd1 – p120 – catenin delta 1

Involved in cell-cell adhesion and signal transduction

Regulation of protein motifs by differential splicing

Experimental validation

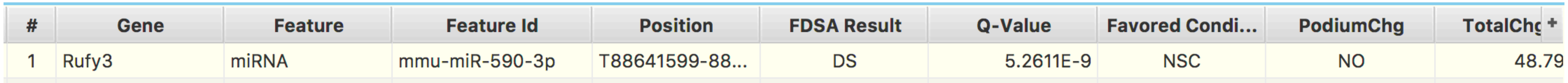



Ctnnd1 - Catenin (cadherin associated protein), de
Proteins View - Aligned for Project 'SVZ'
0
100
200
300
400
500
6
900
N-terminal
C-terminal
PB.3952.2 (-) P30999-2 Provean Score: -18.34
1,790
911aa
PB.3952.3 (-) P30999-3 Provean Score: -14.9
1,790
917aa
PB.3952.4 (-) P30999 Provean Score: -0.85
1,790
938aa
PB.3952.5 (-) G3X9V2 Provean Score: -4.29
1,790
932aa
● UniProtKB/Swiss-Prot_Phosphosite : PTM
○ UniProtKB/Swiss-Prot_Phosphosite : MOTIF
● UniProtKB/Swiss-Prot_Phosphosite : COILED
🟩 pfam : DOMAIN - PF00514

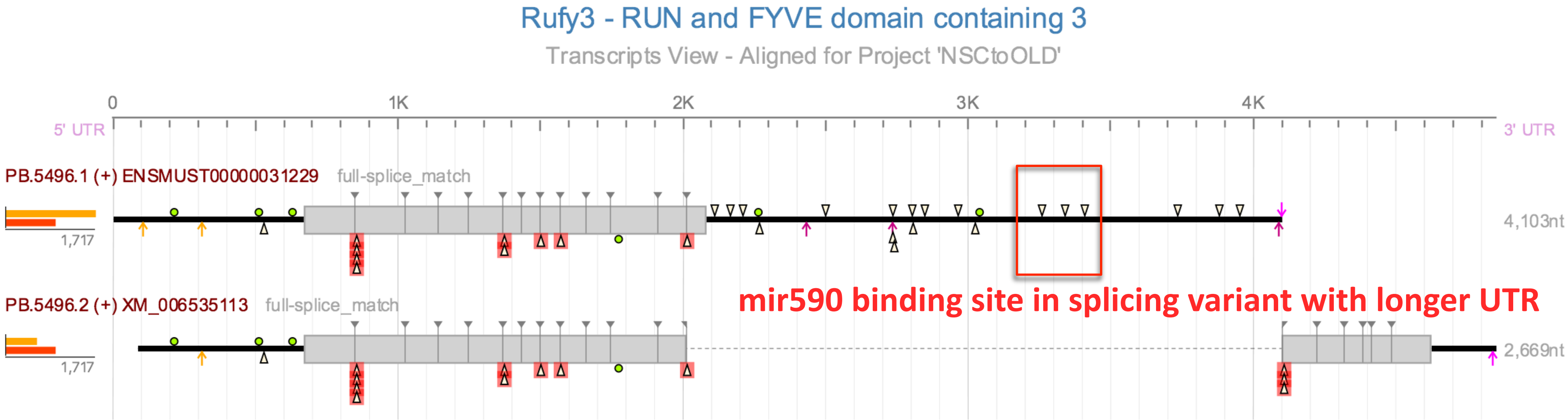
Nuclear Localization Signal

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



Rufy3 - RUN and FYVE domain containing 3

Transcripts View - Aligned for Project 'NSCtoOLD'

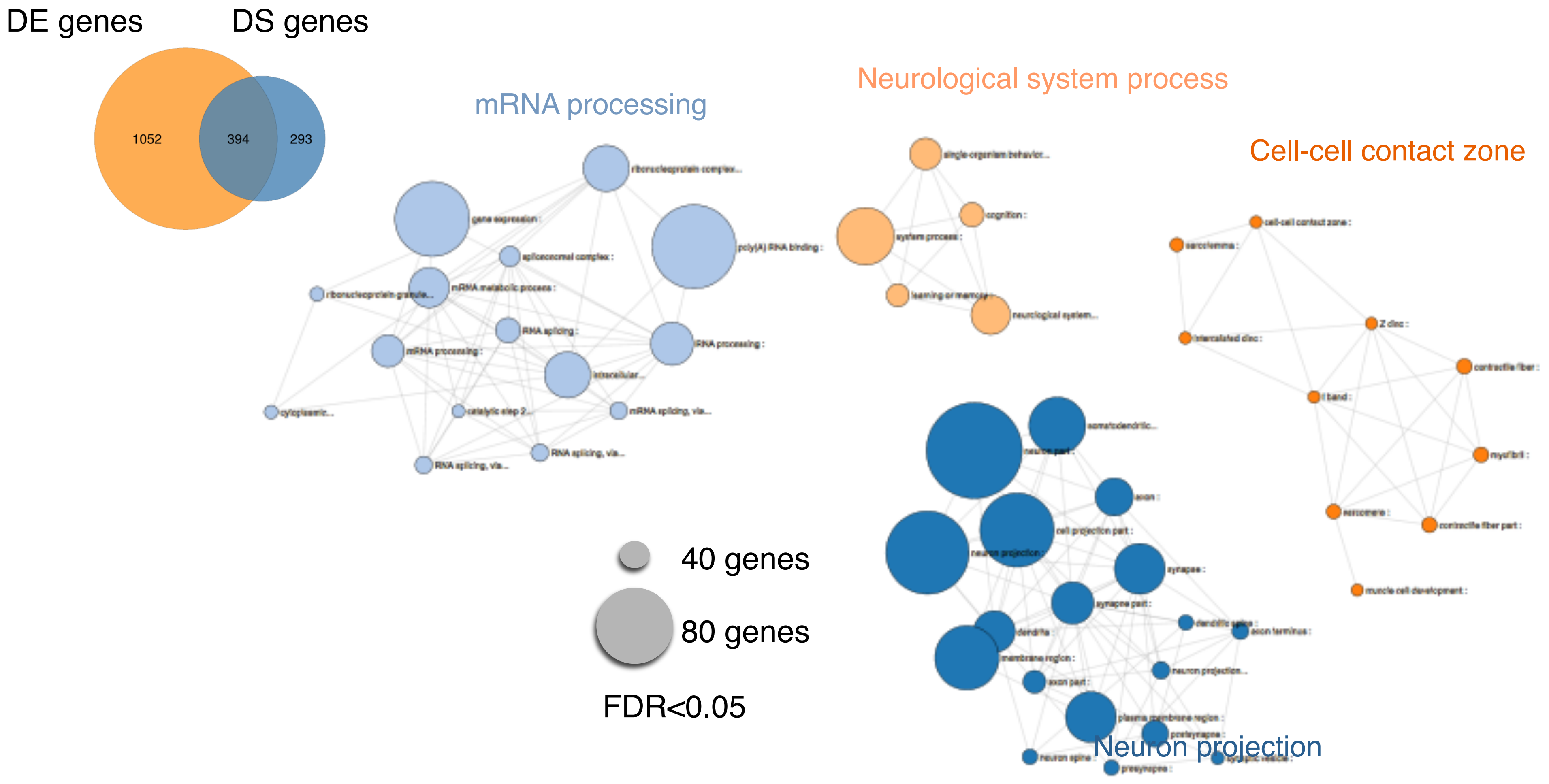
mir590 binding site in splicing variant with longer UTR

<u>*Experimental validation **ongoing**:*</u>

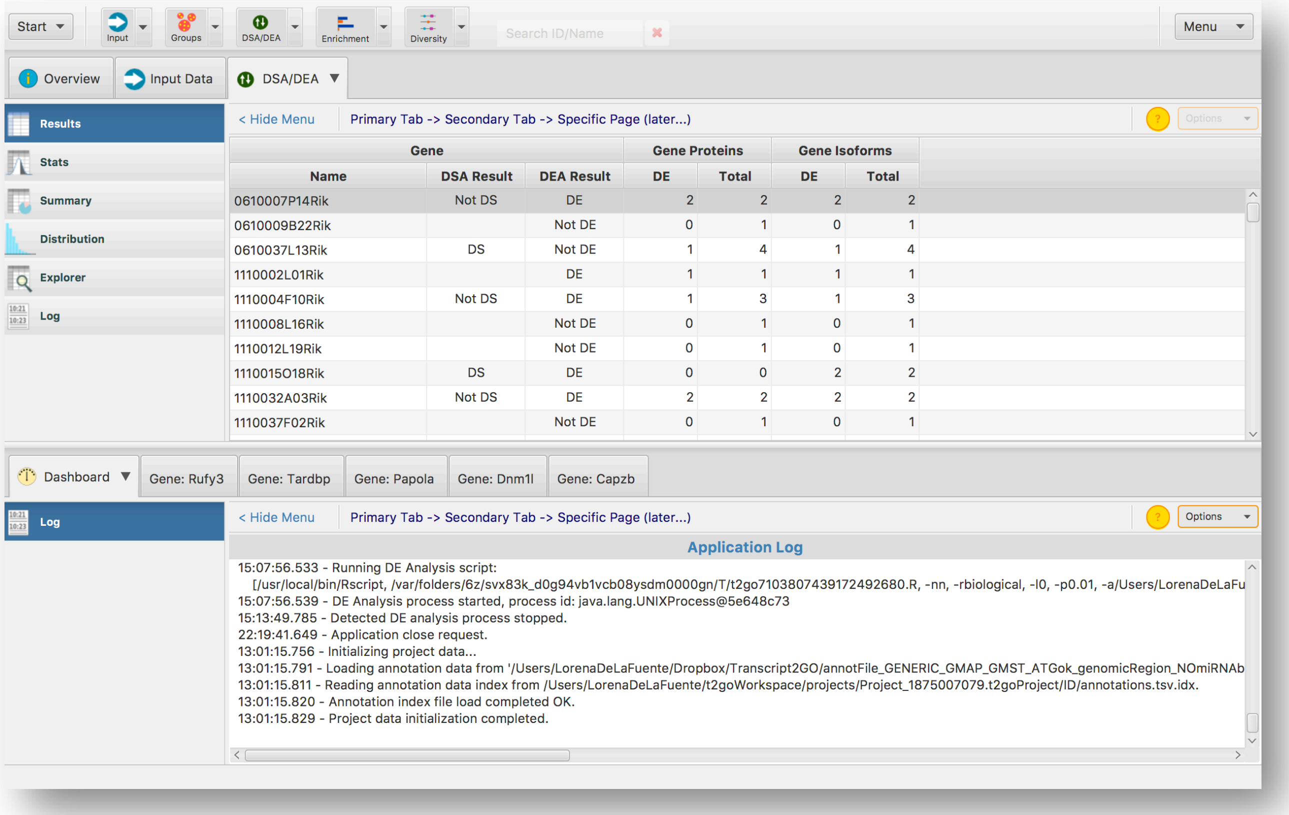
- Analysis of miRNA 590 expression.

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

Functions enriched in Differential Spliced genes



http://tappas.org

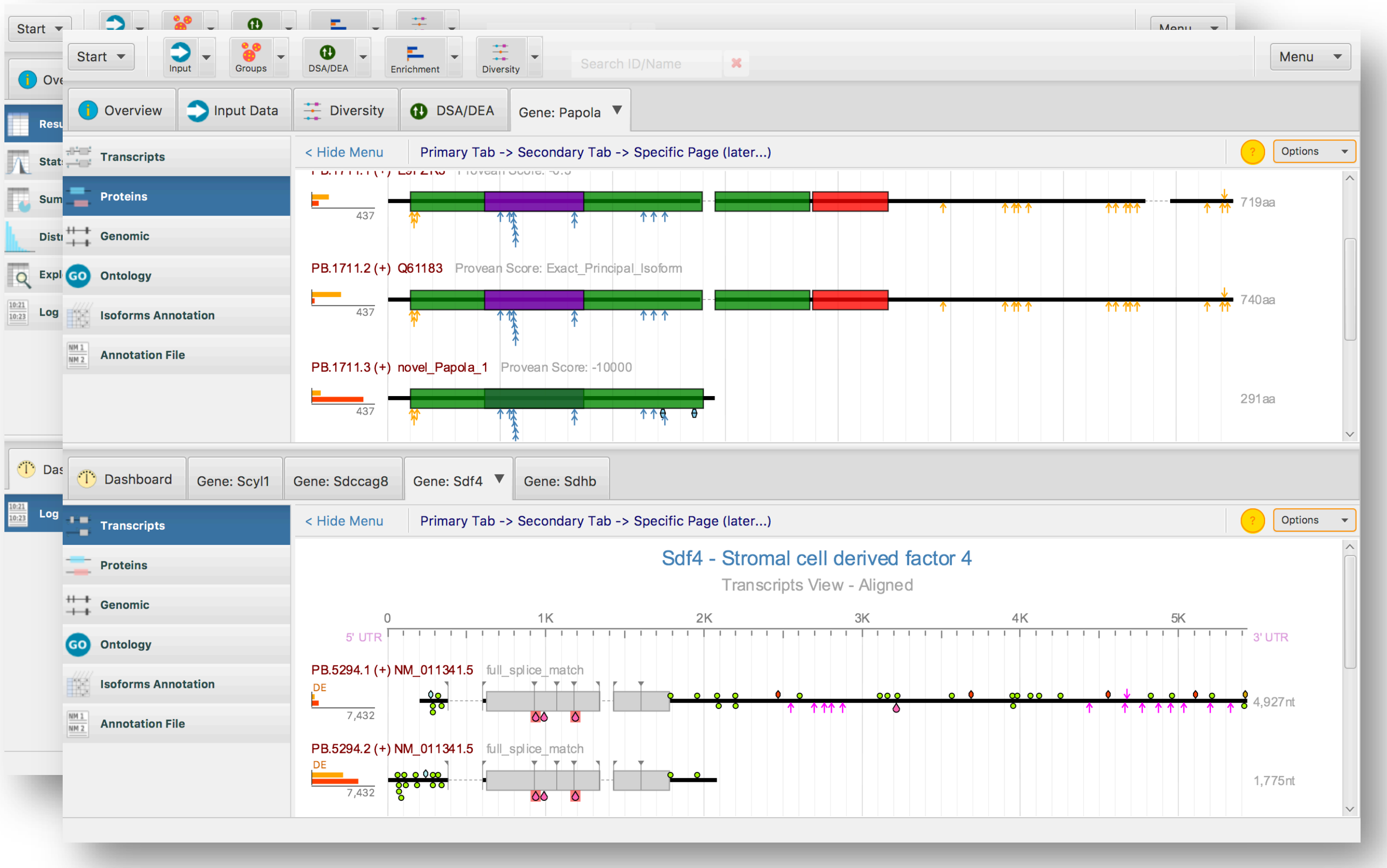


Gene			Gene Proteins		Gene Isoforms	
Name	DSA Result	DEA Result	DE	Total	DE	Total
0610007P14Rik	Not DS	DE	2	2	2	2
0610009B22Rik		Not DE	0	1	0	1
0610037L13Rik	DS	Not DE	1	4	1	4
1110002L01Rik		DE	1	1	1	1
1110004F10Rik	Not DS	DE	1	3	1	3
1110008L16Rik		Not DE	0	1	0	1
1110012L19Rik		Not DE	0	1	0	1
1110015O18Rik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

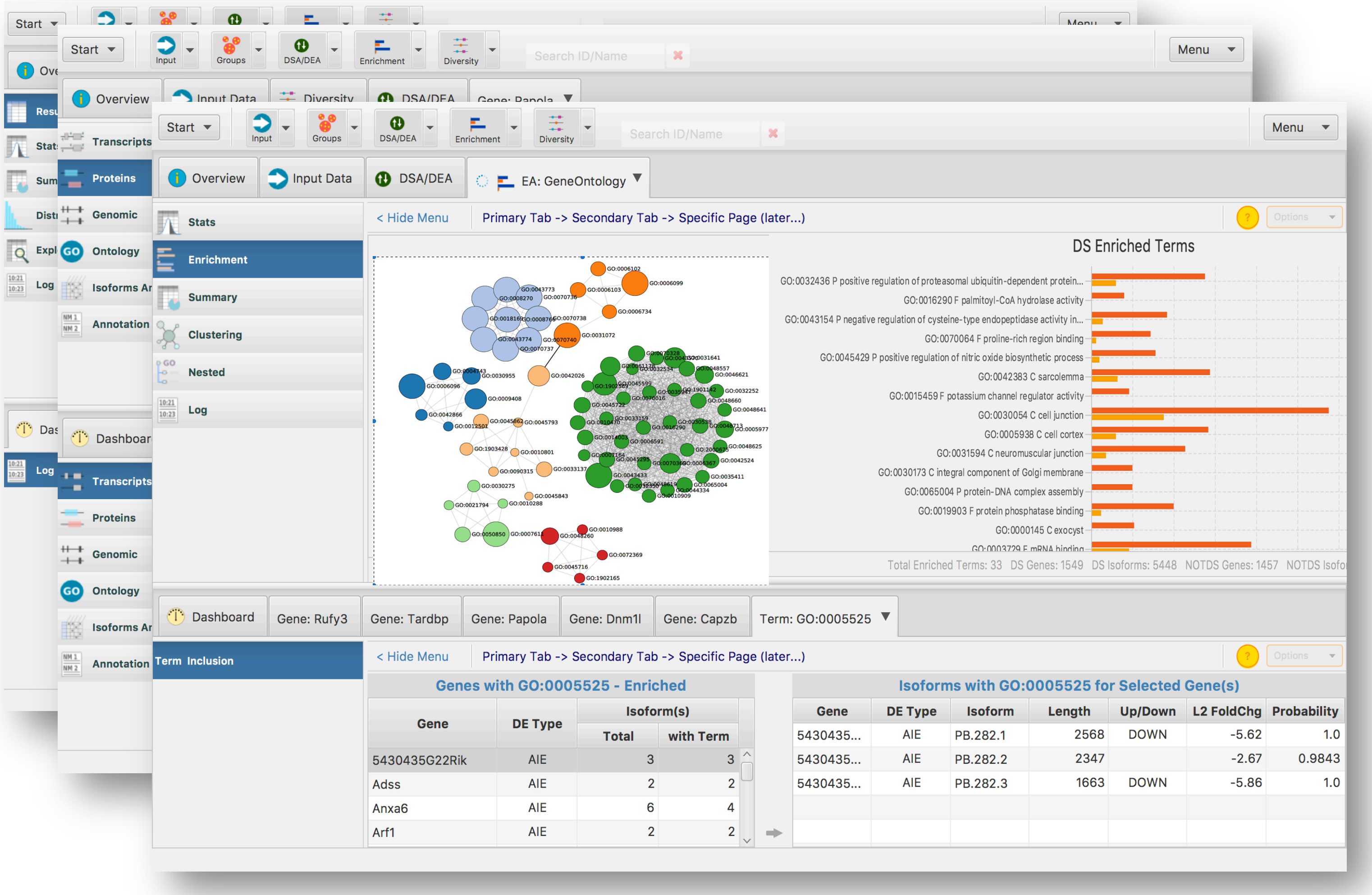
Application Log

```
15:07:56.533 - Running DE Analysis script:
   [/usr/local/bin/Rscript, /var/folders/6z/svx83k_d0g94vb1vcb08ysdm0000gn/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.
```

http://tappas.org


Start
Input
Groups
DSA/DEA
Enrichment
Diversity
Search ID/Name
Menu
Overview
Input Data
Diversity
DSA/DEA
Gene: Papola
Transcripts
Proteins
Genomic
Ontology
Isoforms Annotation
Annotation File
< Hide Menu
Primary Tab -> Secondary Tab -> Specific Page (later...)
Options
PB.1711.2 (+) Q61183 Provean Score: Exact_Principal_Isoform
PB.1711.3 (+) novel_Papola_1 Provean Score: -10000
437
719aa
740aa
291aa
Dashboard
Gene: Scyl1
Gene: Sdccag8
Gene: Sdf4
Gene: Sdhb
Sdf4 - Stromal cell derived factor 4
Transcripts View - Aligned
0
1K
2K
3K
4K
5K
5' UTR
3' UTR
PB.5294.1 (+) NM_011341.5 full_splice_match
PB.5294.2 (+) NM_011341.5 full_splice_match
DE
7,432
4,927nt
1,775nt

http://tappas.org


Start
Input
Groups
DSA/DEA
Enrichment
Diversity
Search ID/Name
Menu
Overview
Input Data
DSA/DEA
EA: GeneOntology
Diversity
Gene: Papola
Stats
Enrichment
Summary
Clustering
Nested
Log
Transcripts
Proteins
Genomic
Ontology
Isoforms Ar
Annotation
Dashboard
Gene: Rufy3
Gene: Tardbp
Gene: Papola
Gene: Dnm1l
Gene: Capzb
Term: GO:0005525
Term Inclusion
< Hide Menu
Primary Tab -> Secondary Tab -> Specific Page (later...)
Options
DS Enriched Terms
GO:0032436 P positive regulation of proteasomal ubiquitin-dependent protein...
GO:0016290 F palmitoyl-CoA hydrolase activity
GO:0043154 P negative regulation of cysteine-type endopeptidase activity in...
GO:0070064 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 C exocyst
GO:0003729 F mRNA binding

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

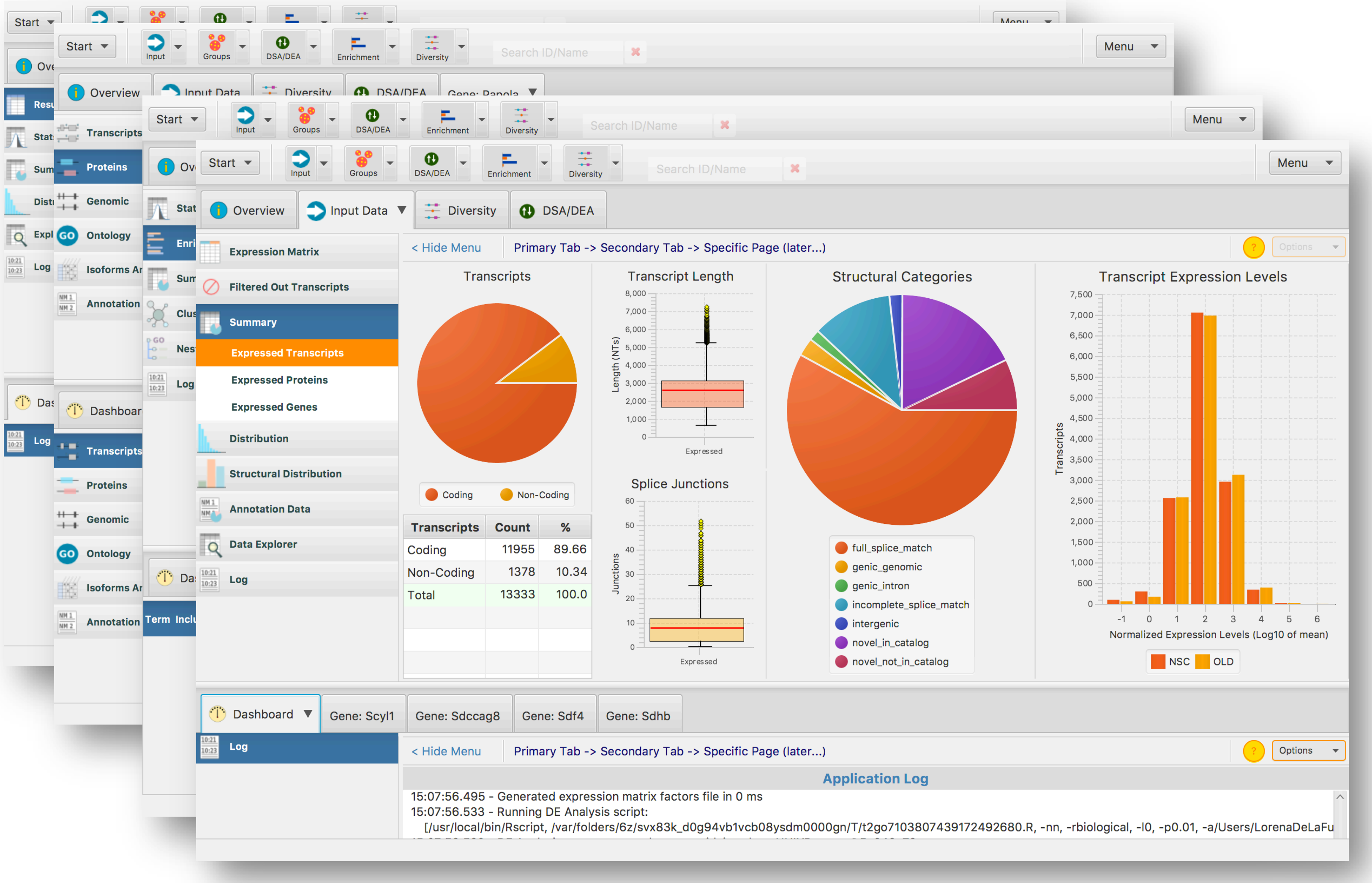
Genes with GO:0005525 - Enriched

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

http://tappas.org



Transcripts	Count	%
Coding	11955	89.66
Non-Coding	1378	10.34
Total	13333	100.0

Application Log

```
15:07:56.495 - Generated expression matrix factors file in 0 ms
15:07:56.533 - Running DE Analysis script:
   [/usr/local/bin/Rscript, /var/folders/6z/svx83k_d0g94vb1vcb08ysdm0000gn/T/t2go7103807439172492680.R, -nn, -rbiological, -l0, -p0.01, -a/Users/LorenaDeLaFu
```

Acknowledgements

UF	UCI	Pacbio	CIPF
William Farmerie	Ali Mortazavi	Liz Tseng	Victoria Moreno
Eric Triplett			Susana Rodriguez
Lauren McIntyre			

SEVENTH FRAMEWORK PROGRAMME

MINISTERIO DE CIENCIA E INNOVACIÓN

MARIE CURIE ACTIONS

cost

BECAS DE PRÁCTICAS EN PYMES Santander UNIVERSIDADES

USDA

Ministerio de Educación