Isoform detection and genome annotation using long-reads transcriptome sequencing

Ana Conesa

Institute for Integrative Systems Biology , CSIC, Spain

@anaconesa.bsky.social

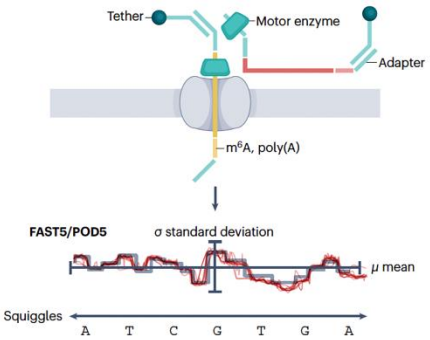
@conesalab.bsky.social

ana.conesa@csic.es

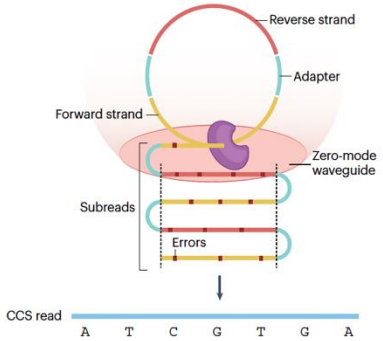
My cats, being Spanish...

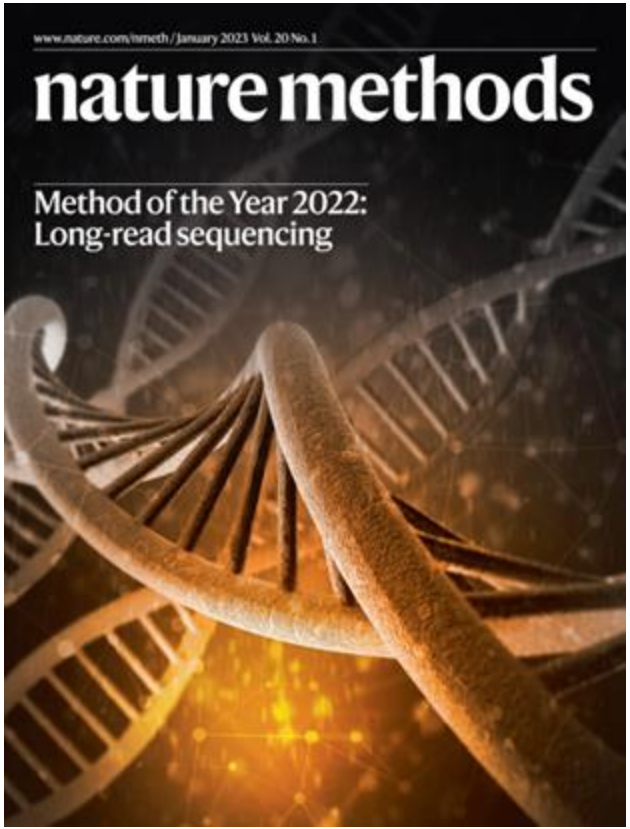
Long-read sequencing of the transcriptome (lrRNA–seq)

Oxford NANOPORE Technologies



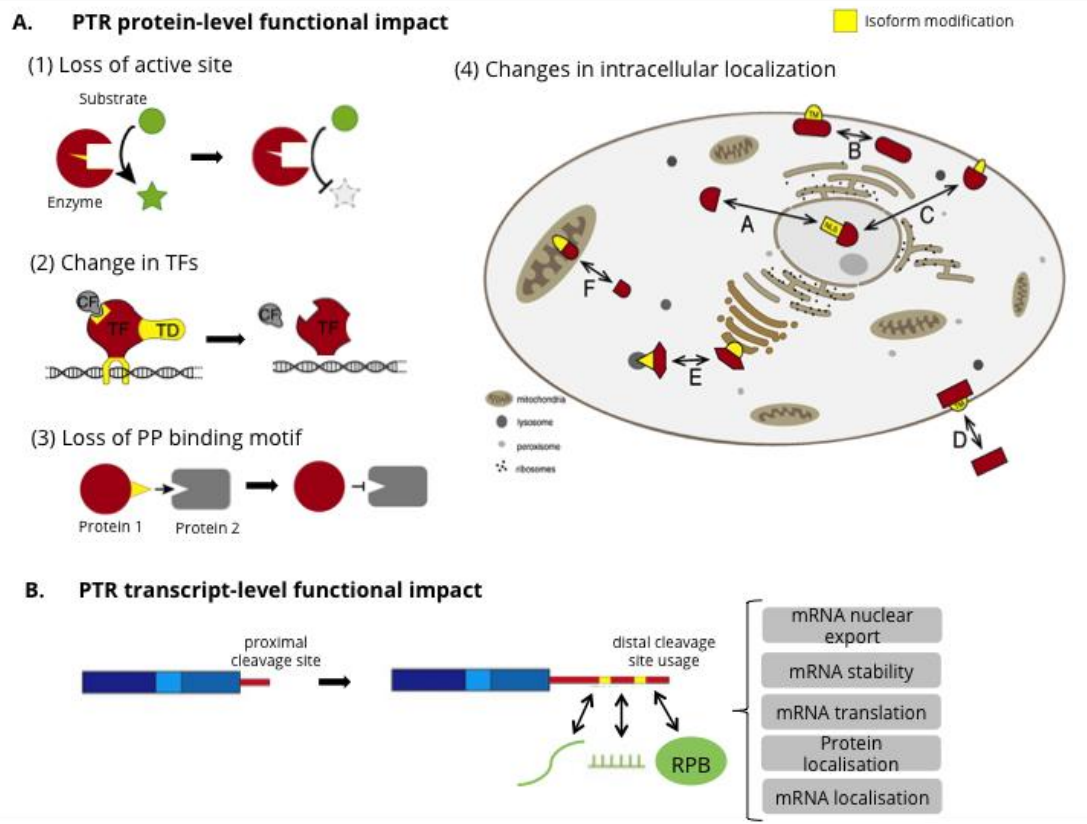
PacBio





Monzó *et al*. **Nat Rev Genetics**, *2025*

Isoforms are multifunctional….



lrRNA-seq reveals MANY novel transcripts

Genomics of Gene Expression Lab

nature

Explore content | About the journal | Publish with us

nature > articles > article

Article | Published: 03 August 2022

Transcriptome variation in human tissues revealed by long-read sequencing

Dafni A. Glinos, Garrett Garborcauskas, Paul Hoffman, Nava Ehsan, Lihua Jiang, Alper Gokden, Xiaoguang Dai, François Aguet, Kathleen L. Brown, Kiran Garimella, Tera Bowers, Maura Costello, Kristin Ardlie, Ruiqi Jian, Nathan R. Tucker, Patrick T. Ellinor, Eoghan D. Harrington, Hua Tang, Michael Snyder, Sissel Juul, Pejman Mohammadi, Daniel G. MacArthur, Tuuli Lappalainen & Beryl B. Cummings

Nature **608**, 353–359 (2022) | Cite this article

24k Accesses | **16** Citations | **290** Altmetric | Metrics

Abstract

Regulation of transcript structure generates transcript diversity and plays an important role in human disease[1,2,3,4,5,6,7]. The advent of long-read sequencing technologies offers the opportunity to study the role of genetic variation in transcript structure[8,9,10,11,12,13,14,15,16]. In this Article, we present a large human long-read RNA-seq dataset using the Oxford Nanopore Technologies platform from 88 samples from Genotype-Tissue Expression (GTEx) tissues and cell lines, complementing the GTEx resource. We identified just over 70,000 novel transcripts for annotated genes, and validated the protein expression of 10% of novel transcripts. We

Genome Biology

Home About Articles Submission Guidelines

Research | Open Access | Published: 07 July 2022

A high-resolution single-molecule sequencing-based Arabidopsis transcriptome using novel methods of Iso-seq analysis

Runxuan Zhang, Richard Kuo, Max Coulter, Cristiane P. G. Calixto, Juan Carlos Entizne, Wenbin Guo, Yamile Marquez, Linda Milne, Stefan Riegler, Akihiro Matsui, Maho Tanaka, Sarah Harvey, Yubang Gao, Theresa Wießner-Kroh, Alejandro Paniagua, Martin Crespi, Katherine Denby, Asa ben Hur, Enamul Huq, Michael Jantsch, Artur Jarmolowski, Tino Koester, Sascha Laubinger, Qingshun Quinn Li, ... John W. S. Brown + Show authors

Genome Biology **23**, Article number: 149 (2022) | Cite this article

4946 Accesses | **9** Citations | **69** Altmetric | Metrics

Abstract

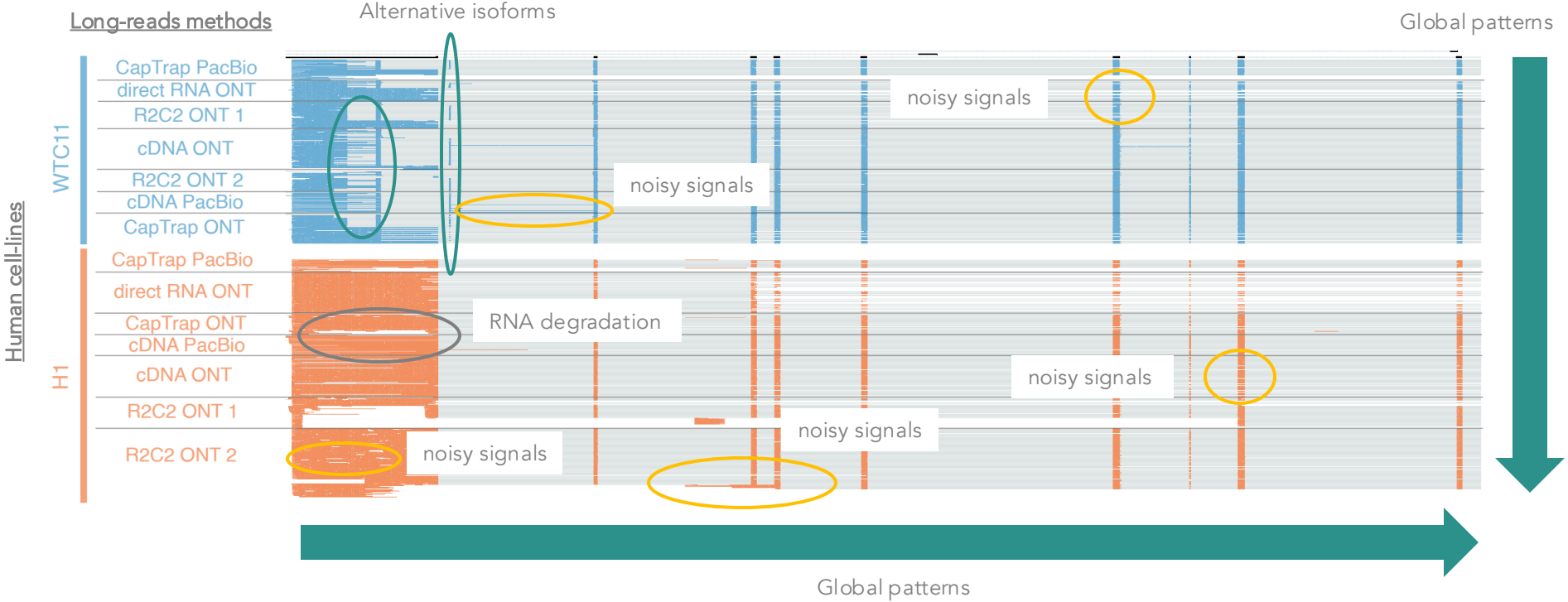
Background

Accurate and comprehensive annotation of transcript sequences is essential for transcript quantification and differential gene and transcript expression analysis. Single-molecule long-read sequencing technologies provide improved integrity of transcript structures including alternative splicing, and transcription start and polyadenylation sites. However, accuracy is significantly affected by sequencing errors, mRNA degradation, or incomplete cDNA synthesis.

Results

We present a new and comprehensive *Arabidopsis thaliana* Reference Transcript Dataset 3 (AtRTD3). AtRTD3 contains over 169,000 transcripts—twice that of the best current Arabidopsis transcriptome and including over 1500 novel genes. Seventy-eight percent of

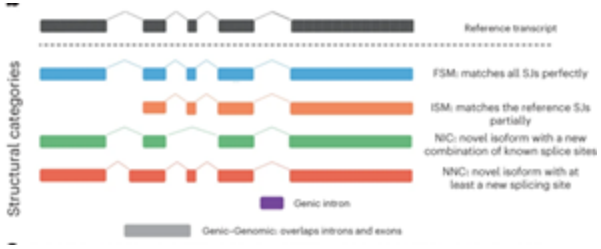
But.... Long reads are noisy....



SQANTI verse

Genomics of Gene Expression Lab

Quality control



SQANTI3

SQANTI curation

SQANTI reads

SQANTI proteins

SQANTI sim

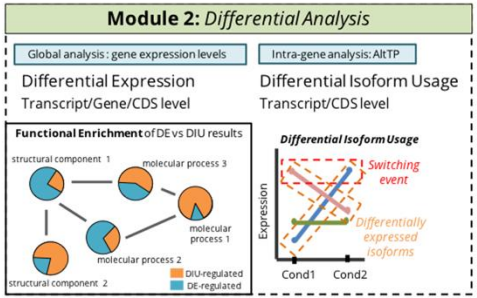
SQANTI tusco

Annotation

SQANTI isoannot

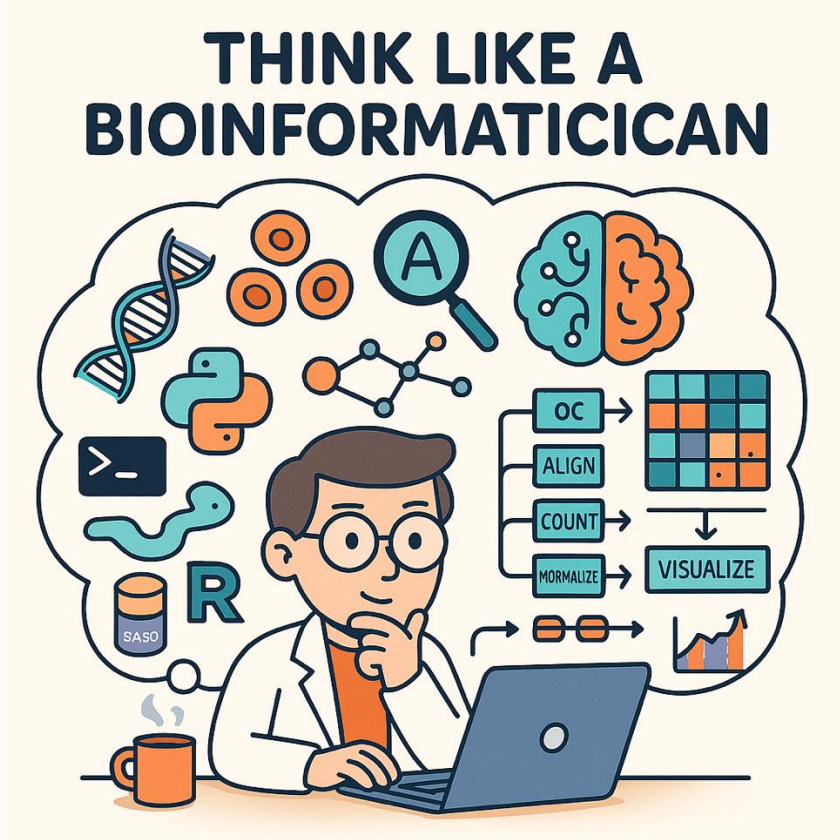
SQANTI evidence

Analysis

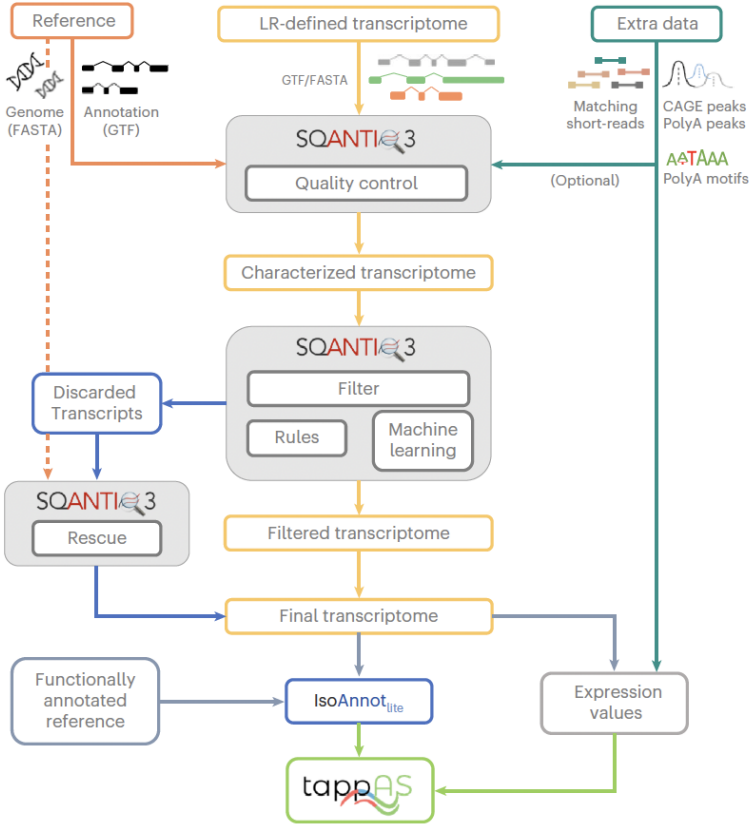


tappAS

What do I expect from this class?

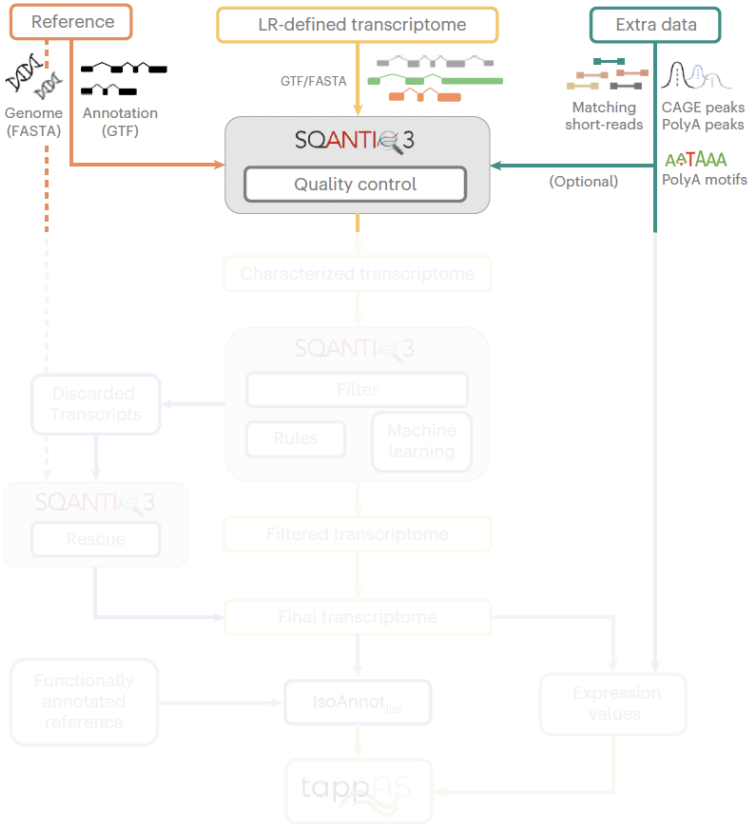


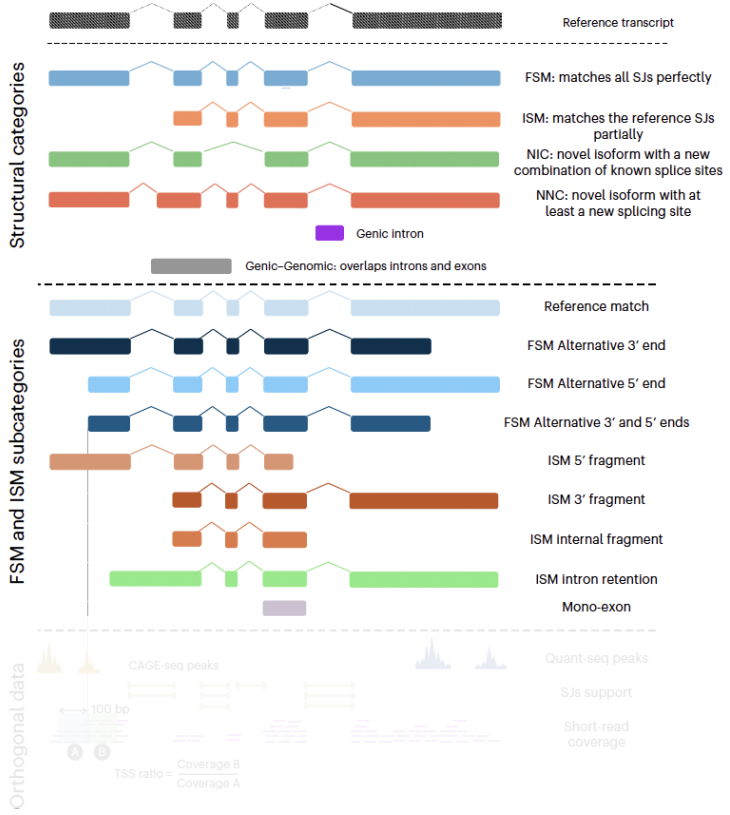
SQANTI3 for quality control of transcript models



Pardo-Palacios *et al.* **Nat Methods**, *2024*

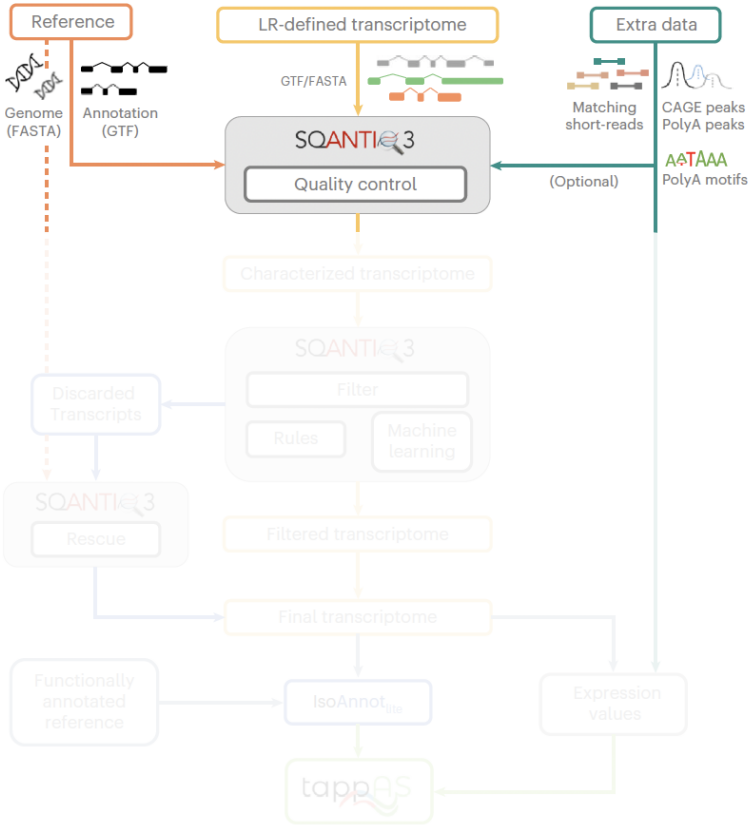
SQANTI3 for quality control of transcript models





Pardo-Palacios *et al.* **Nat Methods**, *2024*

SQANTI3 for quality control of transcript models



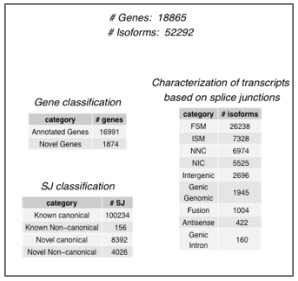
11. `diff_to_TSS` : distanc
downstream of TSS, s
12. `diff_to_TTS` : distanc
upstream of TTS, so th
13. `diff_to_gene_TSS` : d
This field is different fr
starts downstream of
14. `diff_to_gene_TTS` : d
Negative value means
15. `subcategory` : additio
`exon` . Intron rententio
16. `RTS_stage` : TRUE if o
17. `all_canonical` : TRUE
18. `min_sample_cov` : sam
19. `min_cov` : minimum ju
20. `min_cov_pos` : the jun
21. `sd_cov` : standard dev
22. `FL` or `FL.<sample>` :
23. `n_indels` : total numb
24. `n_indels_junc` : numl
potentially unreliable j
25. `bite` : TRUE if contair
26. `iso_exp` : short read e
27. `gene_exp` : short read
`expression` is provide
28. `ratio_exp` : ratio of `i`
29. `FSM_class` : This featu

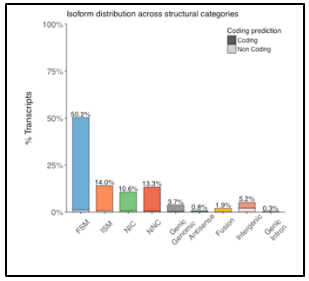
30. `coding` : Coding potential capacity acc
31. `ORF_length` : predicted ORF length.
32. `CDS_length` : predicted CDS length. It
33. `CDS_start` : CDS start.
34. `CDS_end` : CDS end.
35. `CDS_genomic_start` : genomic coordin
36. `CDS_genomic_end` : genomic coordinat
37. `predicted_NMD` : TRUE if there's a prec
otherwise. NA if non-coding.
38. `perc_A_downstreamTTS` : percent of ge
the 3' end site of this isoform is probab
39. `seq_A_downstream_TTS` : sequence of t
40. `dist_to_CAGE_peak` : distance to close
means downstream of TSS. Strand-spe
PacBio transcript start site. Will be `NA`
41. `within_CAGE_peak` : TRUE if the transc
42. `dist_to_polyA_site` : distance to the
43. `within_polyA_site` : TRUE if the trans
seq).
44. `polyA_motif` : if `--polyA_motif_list`
45. `polyA_dist` : if `--polyA_motif_list` i
putative poly(A) site. This distance is h
46. `polyA_motif_found` : TRUE if a polyA m
47. `ORF_seq` : Predicted ORF sequence. Th
48. `ratio_TSS` : Using Short-Read data, w

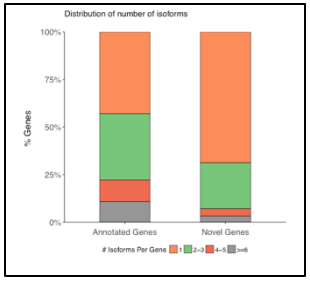
8. `junction_category` : `known` if the
that it is possible to have a `novel` j
combination might be novel.
9. `start_site_category` : `known` if th
10. `end_site_category` : `known` if the
11. `diff_to_Ref_start_site` : distance
site.
12. `diff_to_Ref_end_site` : distance t
site.
13. `bite_junction` : Applies only to no
value is TRUE, otherwise it is FALSE
14. `splice_site` : Splice motif.
15. `RTS_junction` : TRUE if junction is
16. `indel_near_junct` : TRUE if there i
incorrectness.
17. `sample_with_cov` : If `--coverage`
files) that have uniquely mapped sh
18. `total_coverage_unique/multi` : To
cover this junction.

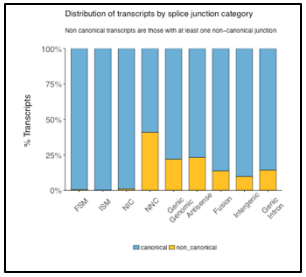
Pardo-Palacios *et al.* **Nat Methods**, *2024*

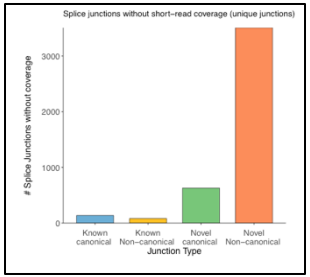
SQANTI3 QC plots

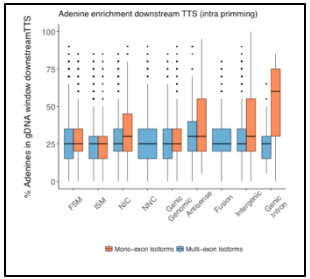


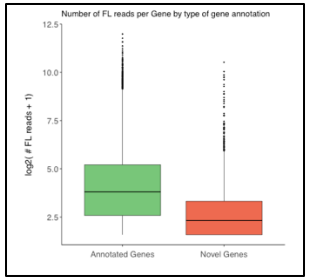


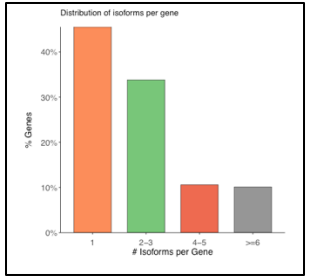


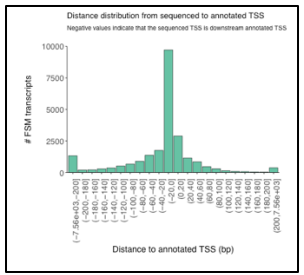


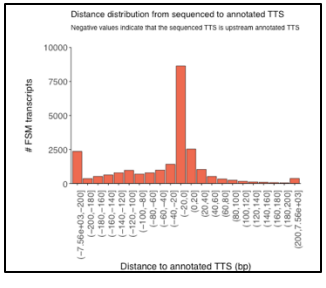


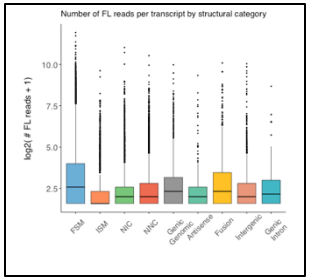


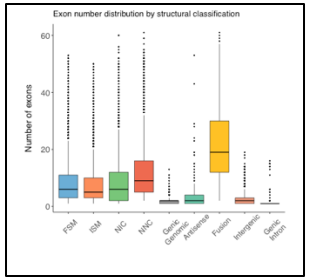


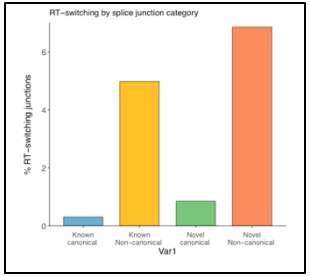


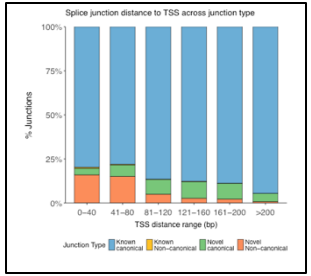


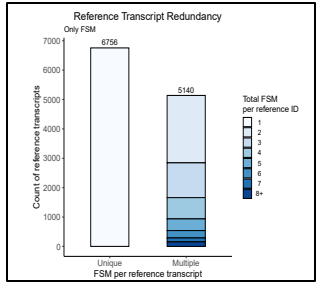






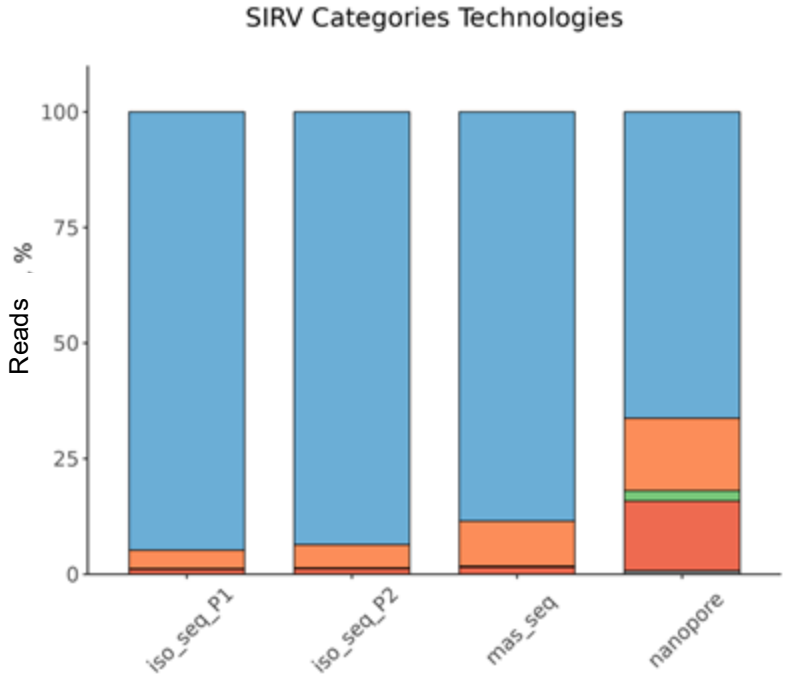






SQANTI3-reads to assess library preparation

Mouse brain samples: Isoseq, MAS-seq (Kinnex) and Nanopore, with Spike-ins (SIRVs)

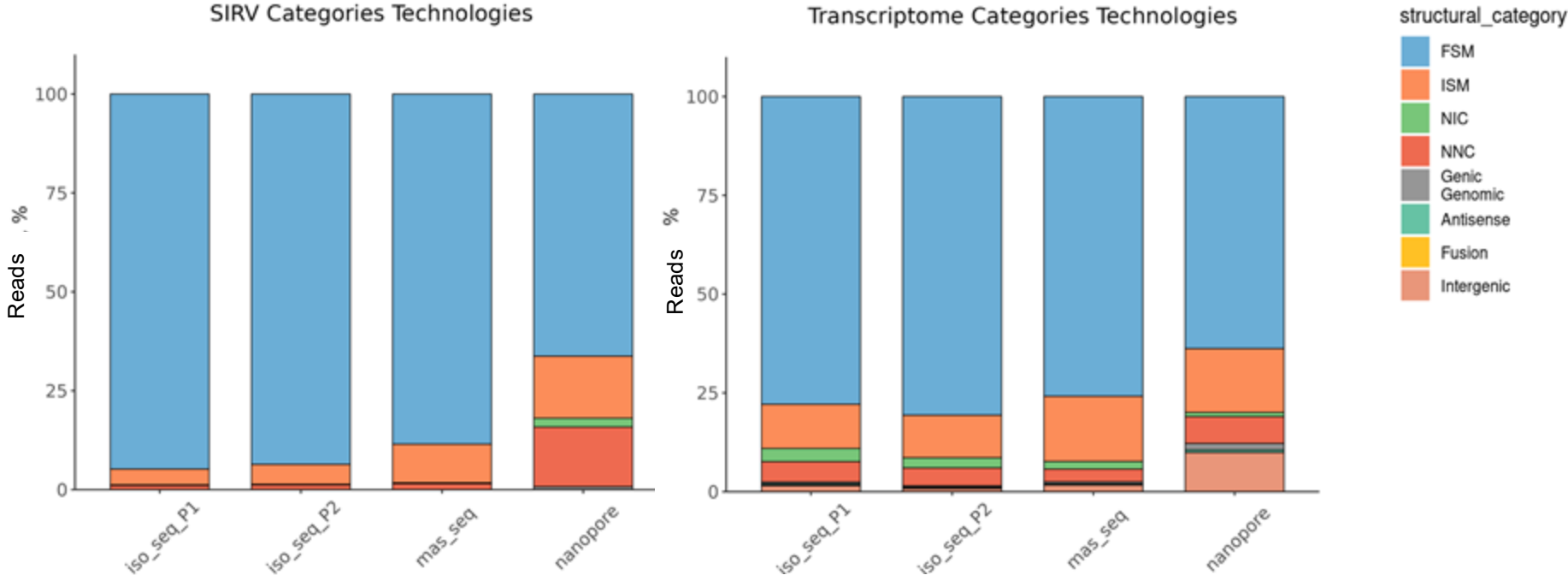


- >90% PacBio reads capture existing transcript models.
- Up to 10% reads might not be full-length. This is more accentuated with the MAS-seq protocol.
- < %1 PB reads contain errors leading to wrong transcript models
- Nanopore has more error reads

Keil *et al.* **Genome Research**, 2025

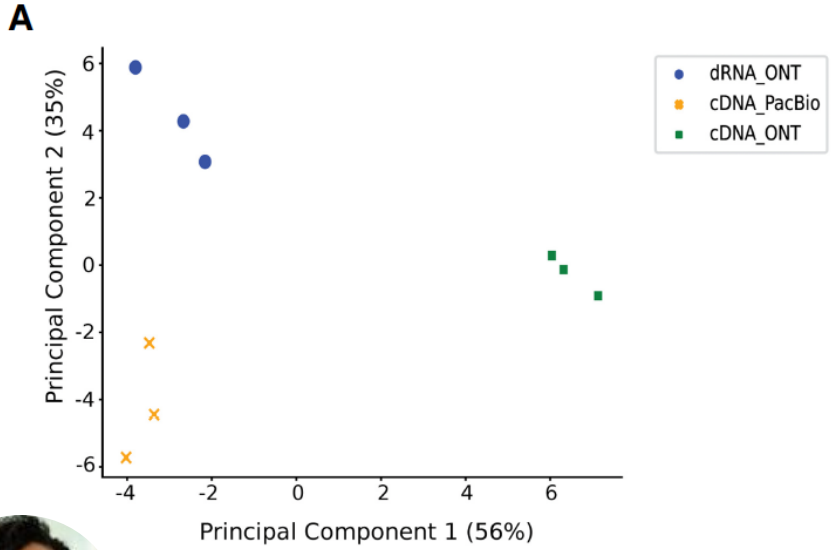
SQANTI3-reads to assess library preparation

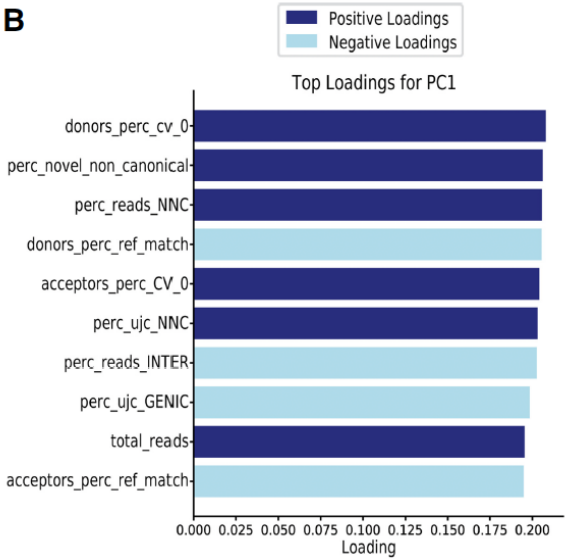
Mouse brain samples: Isoseq, MAS-seq (Kinnex() and Nanopore, with Spike-ins (SIRVs)



SQANTI3-reads to compare samples and technologies

Comparison of sequencing technologies from LRGASP WTC11 samples





Keil *et al.* **Genome Research**, 2025

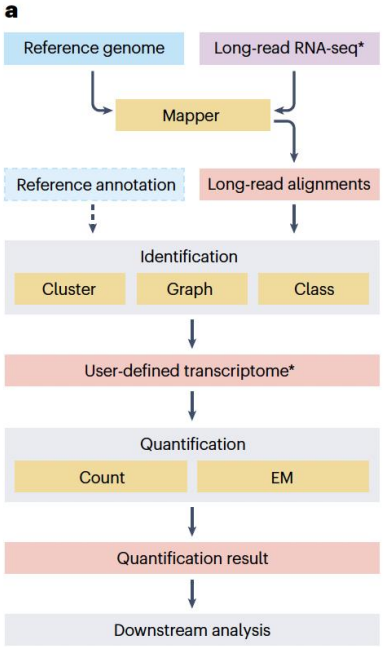
Strategies to go from reads to transcripts

Genomics of Gene Expression Lab

Bambu
Flames
FLAIR
IsoQuant

Iso_IB
IsoTools
Mandalorion
Spectra
TALON

LyRic



Lr-Kallisto
Oarfish

Monzó *et al*. **Nat Rev Genetics**, *2025*

LRGASP benchmarking of lrRNA-seq

Genomics of Gene Expression Lab

3 species

human

mouse

manatee

4 sample types

one cell type

cell mixtures

spike-ins

synthetic data

4 library protocols

cDNA

direct RNA

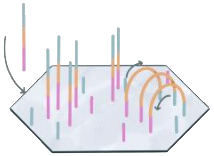
R2C2

CapTrap

3 sequencing platforms

Sequel II

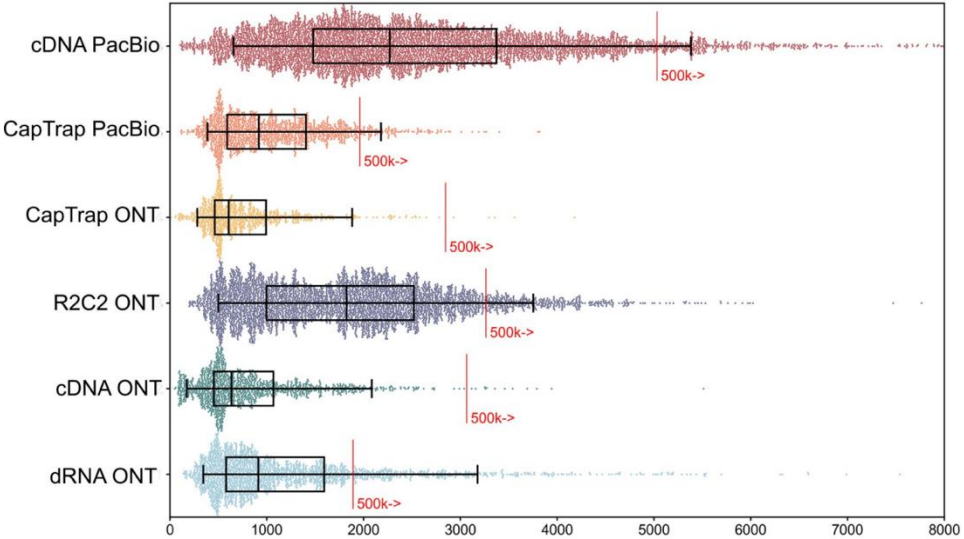
MinION

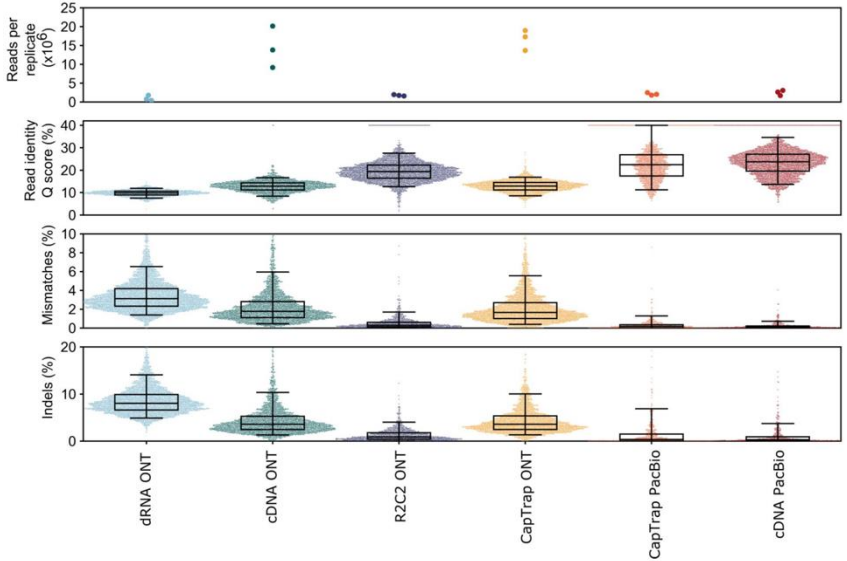


HiSeq

Pardo-Palacios *et al.* **Nat Methods**, *2024*[b]

LRGASP: the data





The magnitude of the challenge

Genomics of Gene Expression Lab

6 library prep + sequencing combinations
X
3 samples (WTC11, H1mix, ES_mouse)
X
2 analysis options (Long Only / Long & Short)
=
36 data analysis possibilities

11 different analysis tools participated

Heavily based on the reference annotation

Ba: Bambu
FM: Flames
FR: FLAIR
IQ: IsoQuant

Supported by the reference annotation

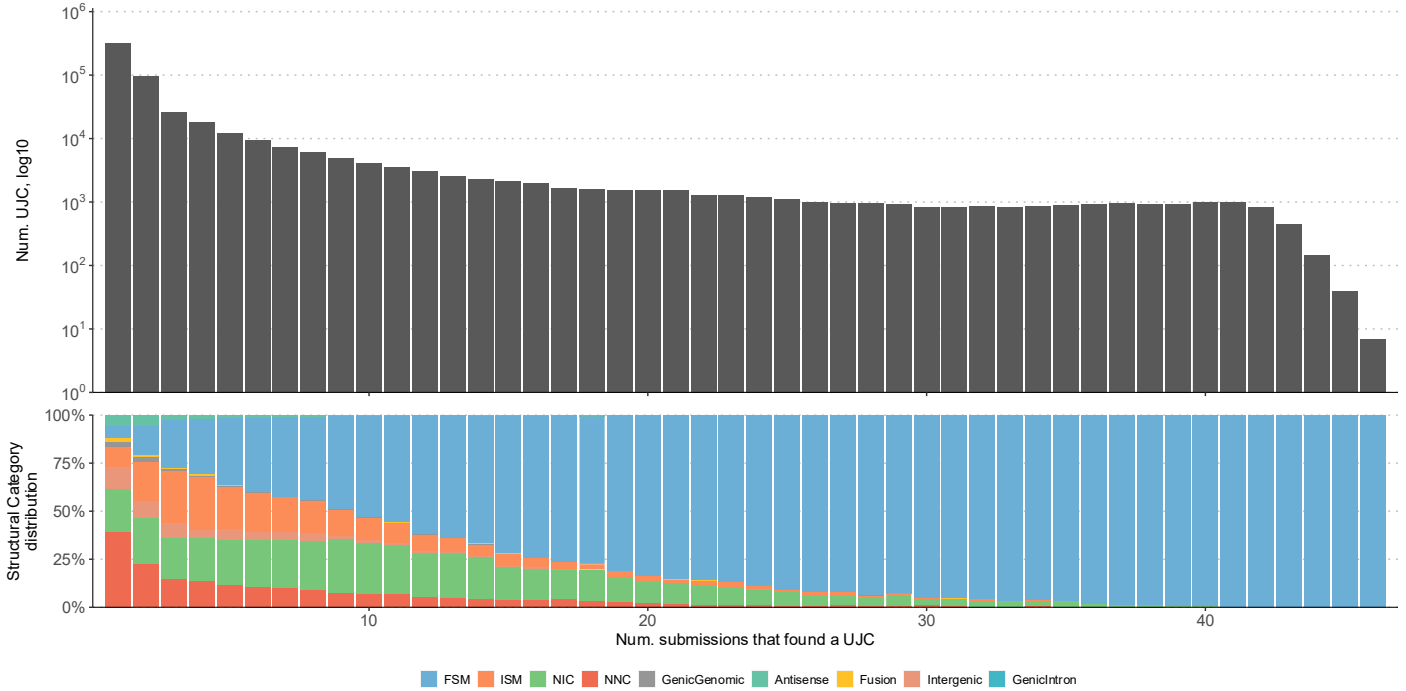
IB: Iso_IB
IT: IsoTools
Ma: Mandalorion
Sp: Spectra
TL: TALON

Agnostic to the reference annotation

Ly: LyRic

47 Pipelines = library preparation + sequencing platform + analysis tool

Limited overlap in detection across pipelines



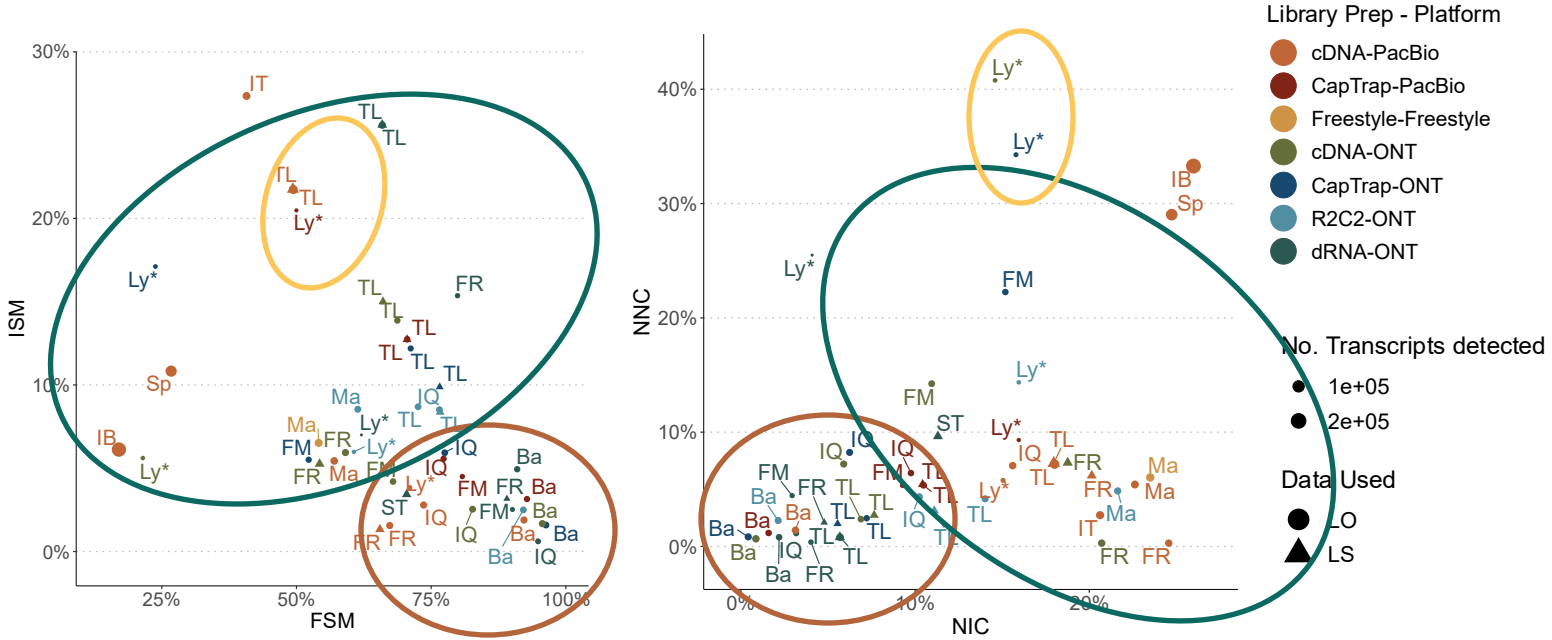
Pipelines agree on known transcripts, novel transcripts tend to be pipeline-specific

Results depend on algorithmic choices

Ba: Bambu
FM: Flames
FR: FLAIR
IQ: IsoQuant

IB: Iso_IB
IT: IsoTools
Ma: Mandalorion
Sp: Spectra
TL: TALON

Ly: LyRic



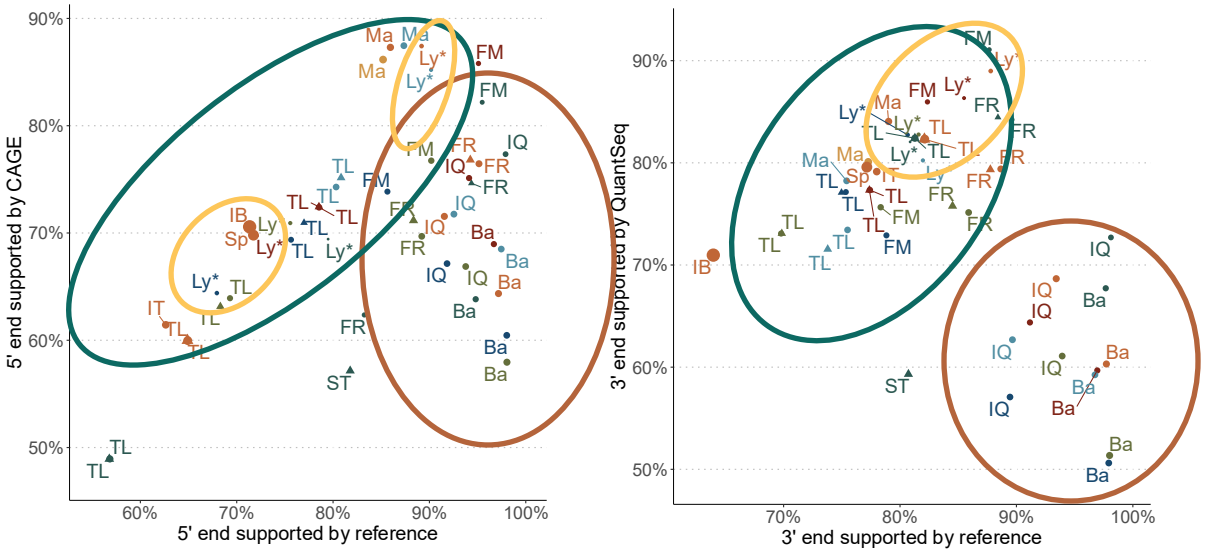
In general, tools that provide a high number of FSM strongly base predictions on genome annotation

Results depend on algorithmic choices

Ba: Bambu
FM: Flames
FR: FLAIR
IQ: IsoQuant

IB: Iso_IB
IT: IsoTools
Ma: Mandalorion
Sp: Spades
TL: TALON

Ly: LyRic



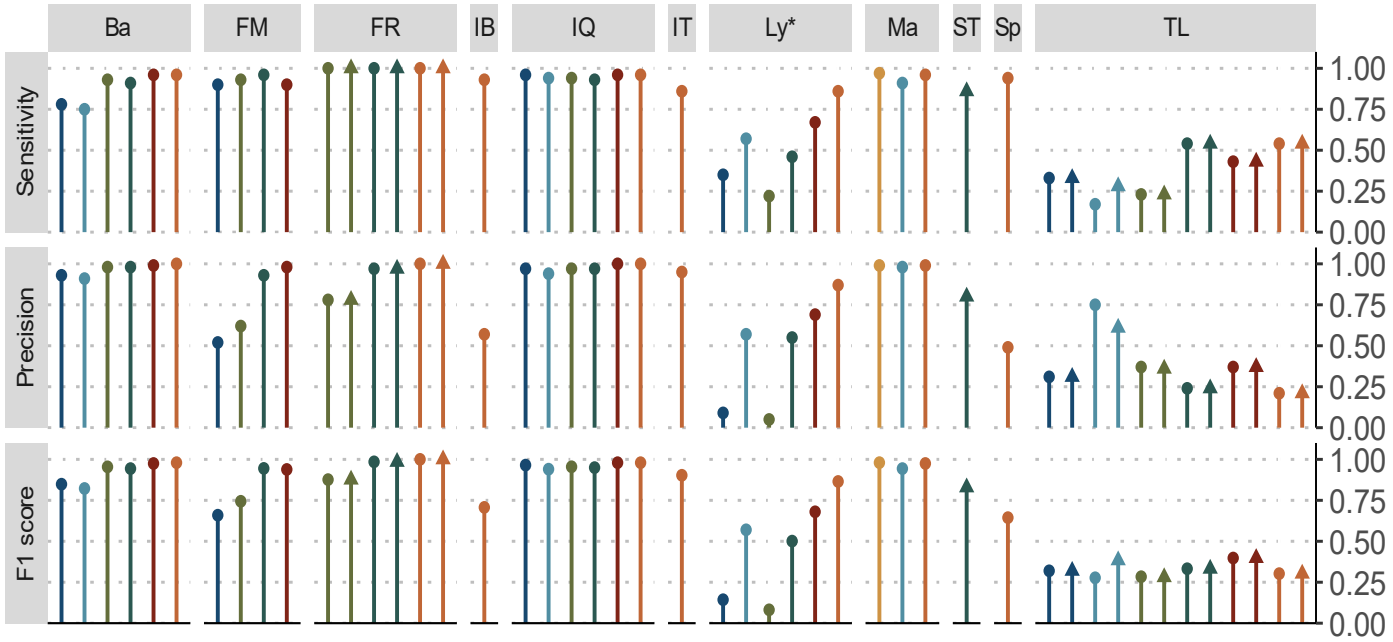
Library Prep - Platform
- cDNA-PacBio
- CapTrap-PacBio
- Freestyle-Freestyle
- cDNA-ONT
- CapTrap-ONT
- R2C2-ONT
- dRNA-ONT

No. Transcripts detected
- 1e+05
- 2e+05

Data Used
- LO
- LS

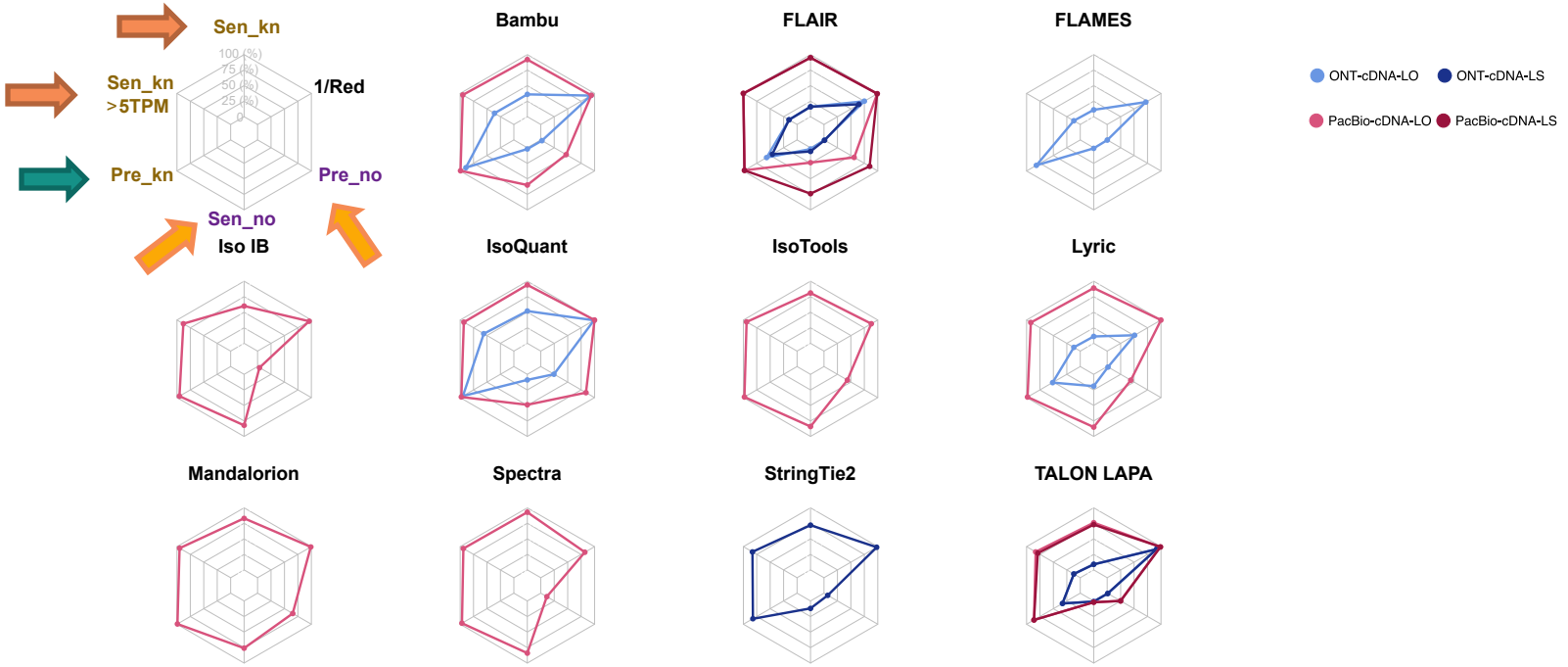
Tools that use the reference annotation are better supported by the annotation than by the experimental data

Performance based on spikes (SIRVs)



Note: SIRVs do not allow to assess capacity for detection of novel transcripts

Performance based on PacBio simulation data

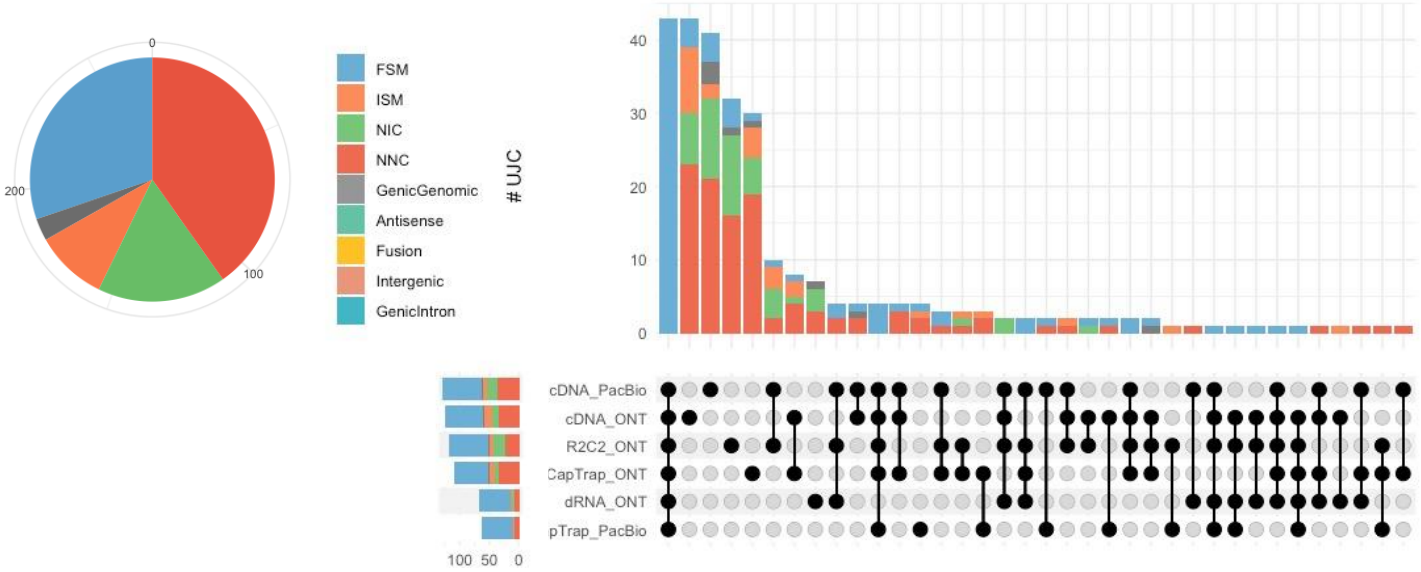


Performance for Novel Transcripts is substantially worse than for known transcripts

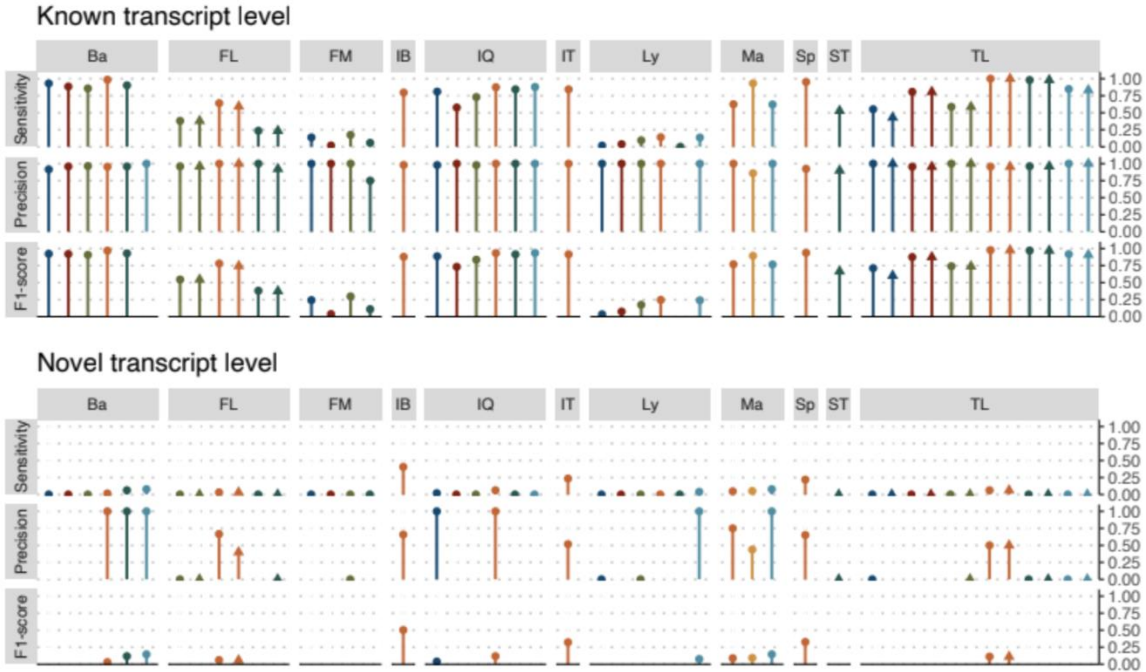
Note: Simulated data simulates sequencing errors, but does not simulate library preparation errors

GENCODE Manual annotation

- 50 loci selected by GENCODE
- Manual annotation for each of the 6 experimental datasets (library prep + sequencing platforms) INDEPENDENTLY
- Most transcripts were NOVEL and found in only ONE experimental dataset !!!



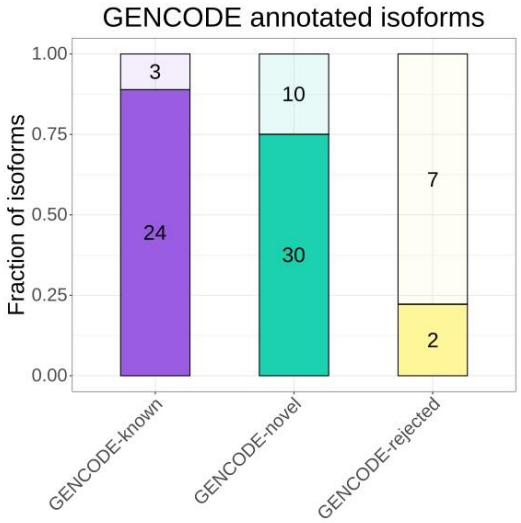
Evaluation on GENCODE manually annotated loci



Sensitivity for novel transcripts is low

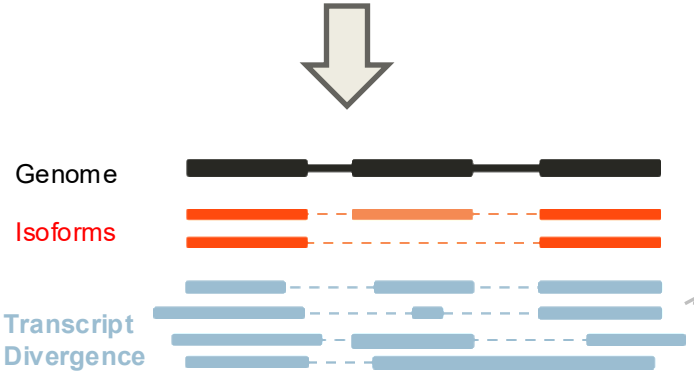
Note: Most tools did report a low number of novel transcripts in these loci

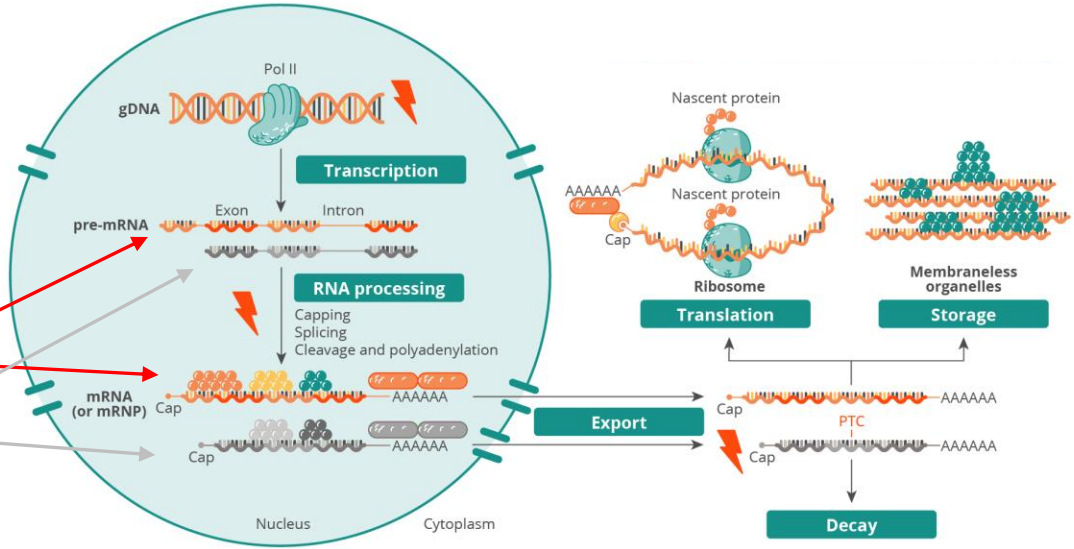
Many novel transcripts were experimentally validated



High validation of reproducible novel transcripts and 50% validation rare NNC

mRNA biogenesis can go wrong leading to *diverging* RNAs





Monzó, Frankish, Conesa, **Genome Research**, *2025*

Summary so-far

- lr-RNA-seq is a **powerful technique** to sequence full-length transcripts, depending on the technology.
- The tech olecules that we hypothe ng
- Transcri night be worth to explo

How can we help?

- However, we see that most tools struggle to detect faithfully novel transcripts

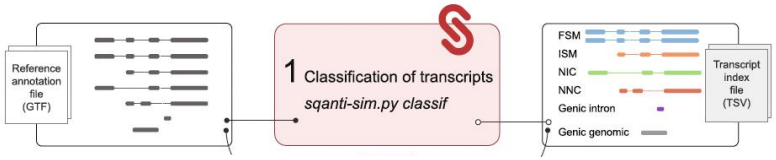
Summary so-far

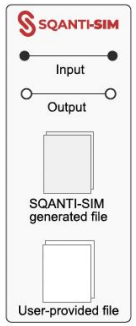
- lr-RNA-seq is a **powerful technique** to sequence full-length transcripts, depending on the technology.
- The tech olecules that we hypothe ng
- Transcri night be worth to explo

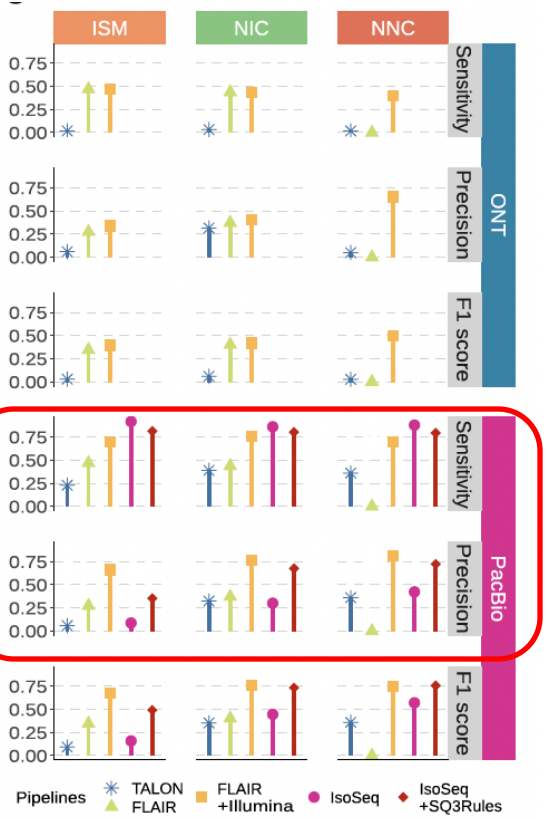
Benchmarking,
Quality Control and Curation

- However, we see that most tools struggle to detect faithfully novel transcripts

SQANTI-SIM: simulating the transcript novelty



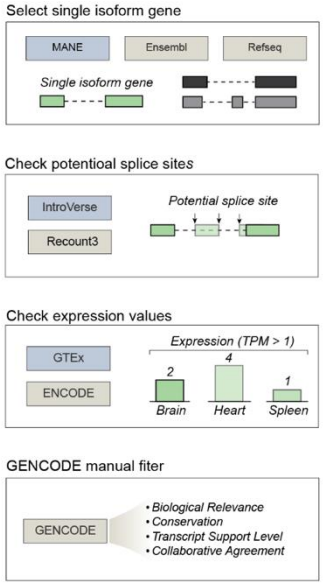




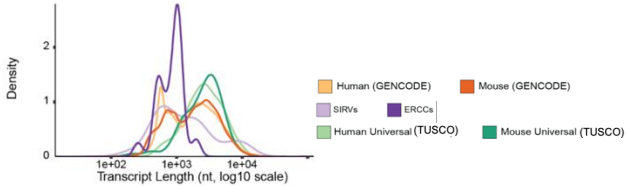
Mestre *et al.* **Genome Biology**, *2023*

Development of an internal control: TUSCO (I)

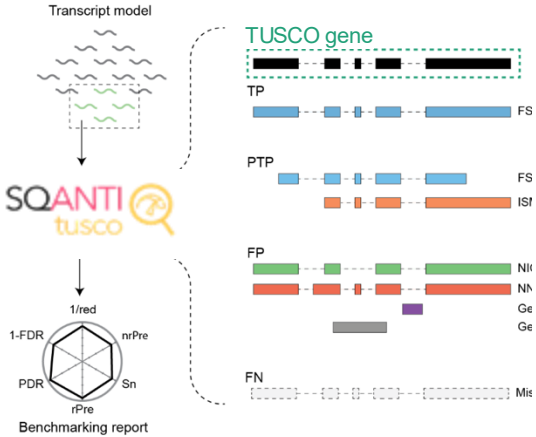
High-confidence one-isoform genes



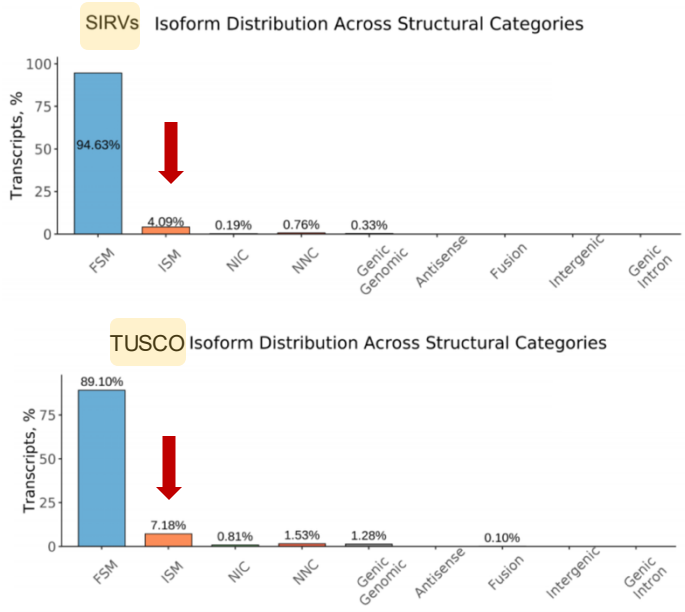
More realistic transcript length distribution



Ground truth parameters

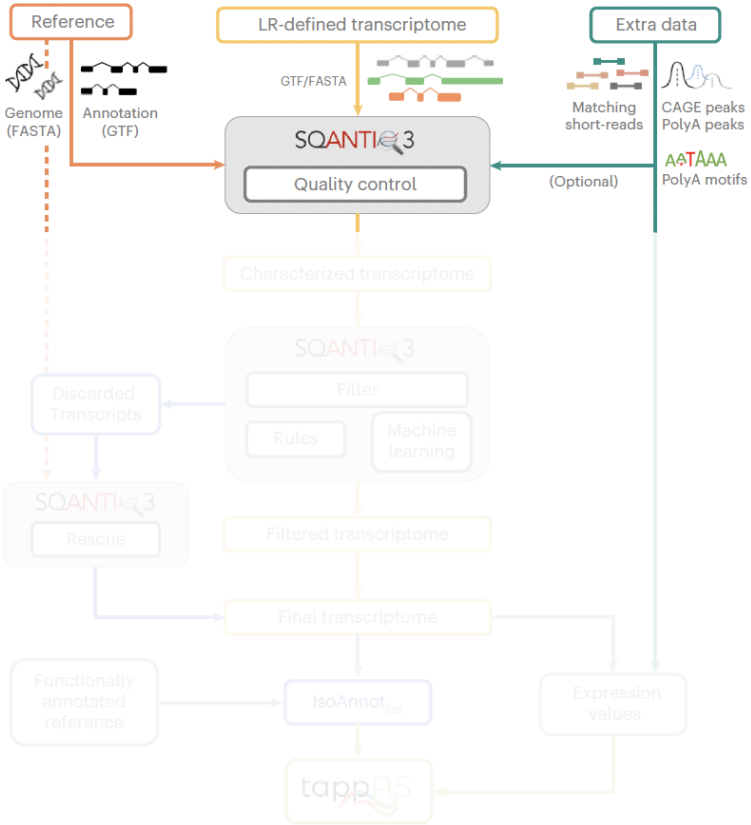


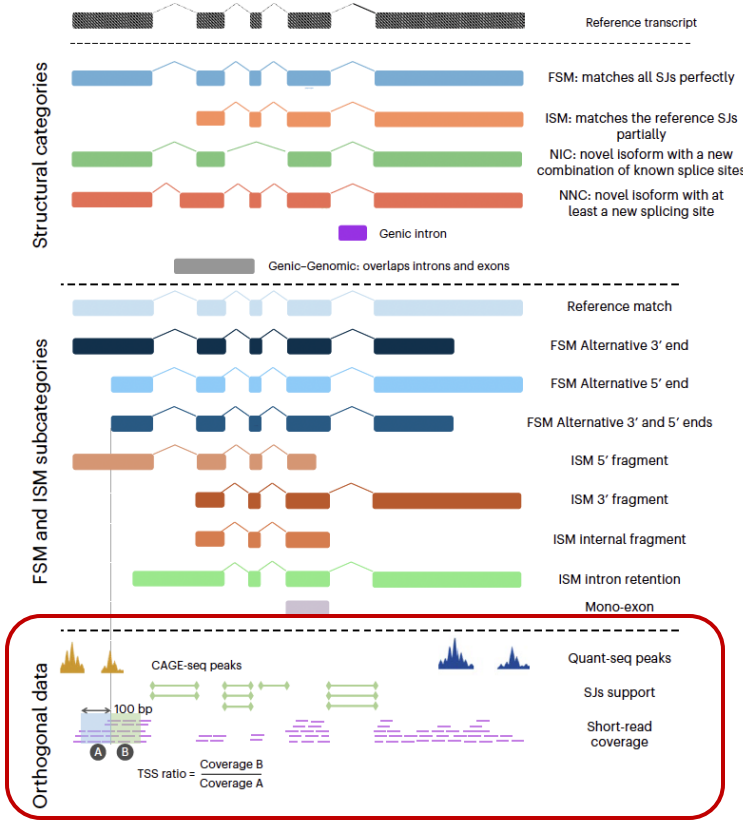
TUSCO reads are similarly annotated as SIRVs



Liu *et al.* **Under review**

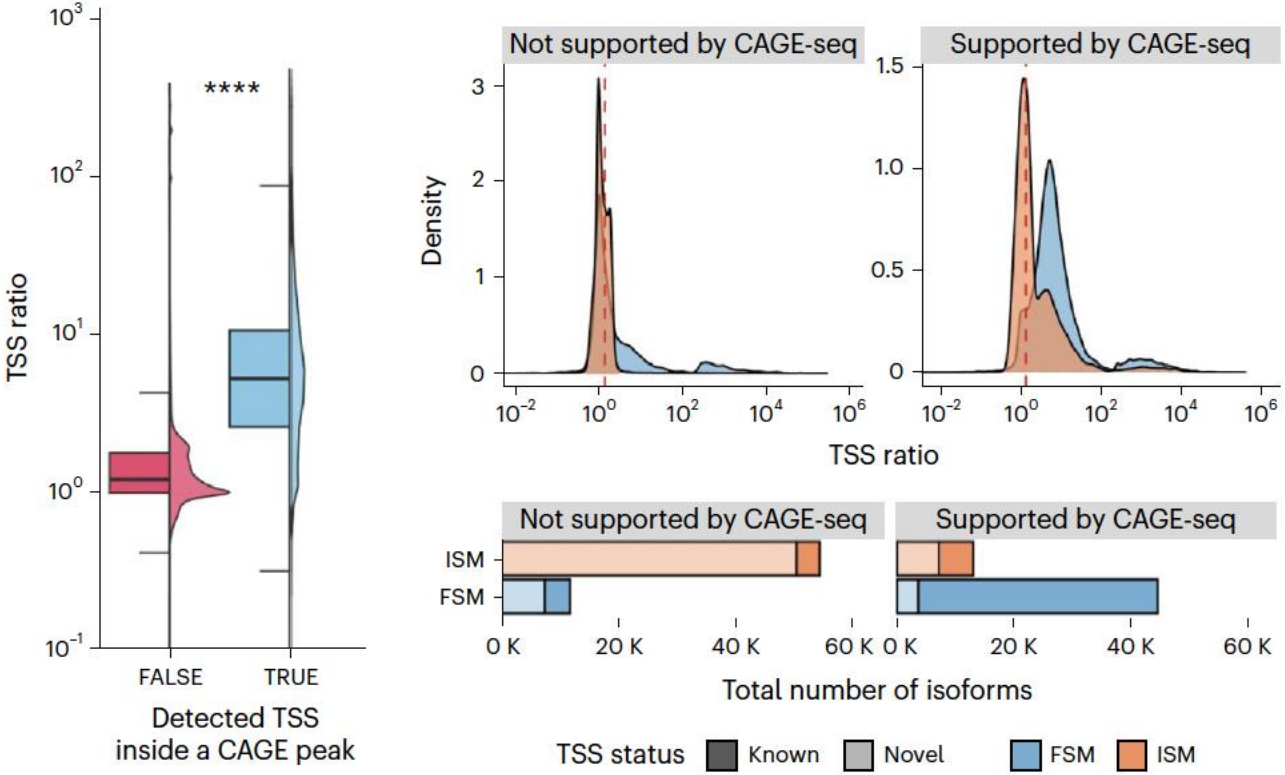
The SQANTI3 strategy for transcript curation





Pardo-Palacios *et al.* **Nat Methods**, *2024*

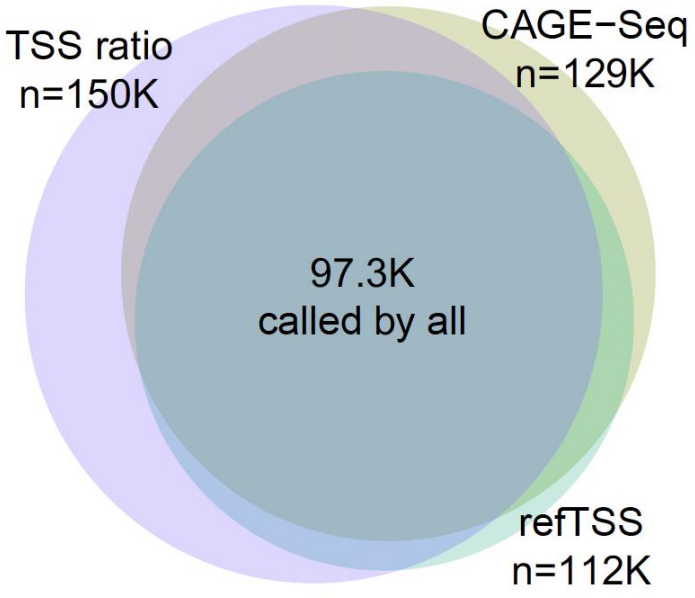
TSS ratio and CAGE help to discriminate correct TSS



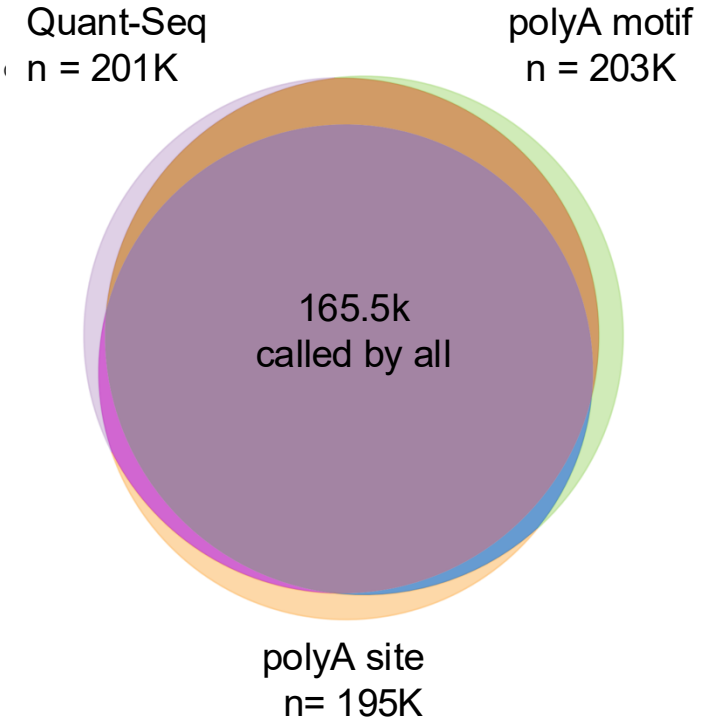
Pardo-Palacios *et al*. **Nat Methods**, *2024*

Curation of TSS and TTS with ortogonal data

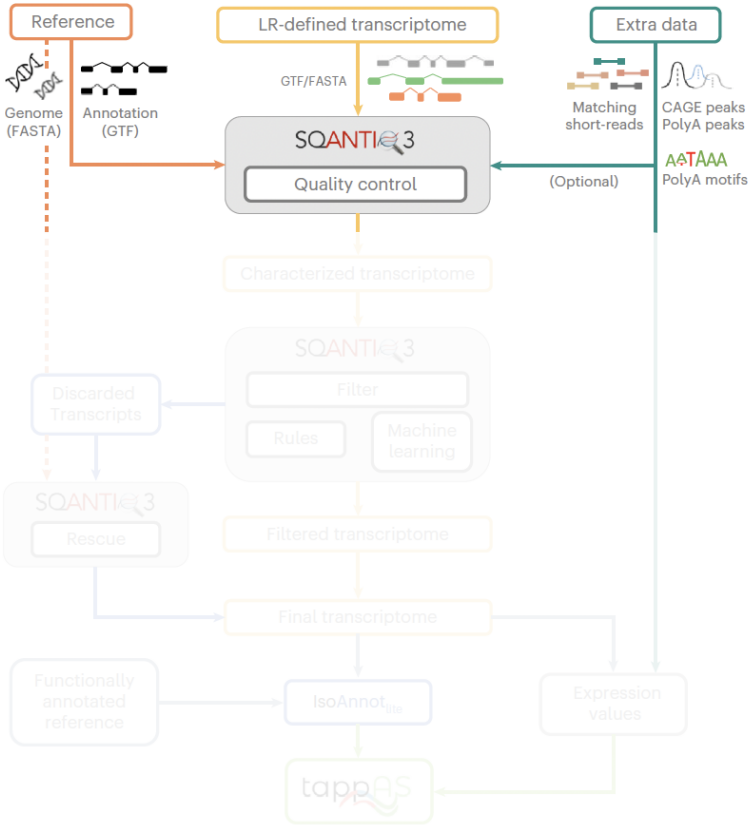
TSS



TTS



We use QC features to curate transcripts



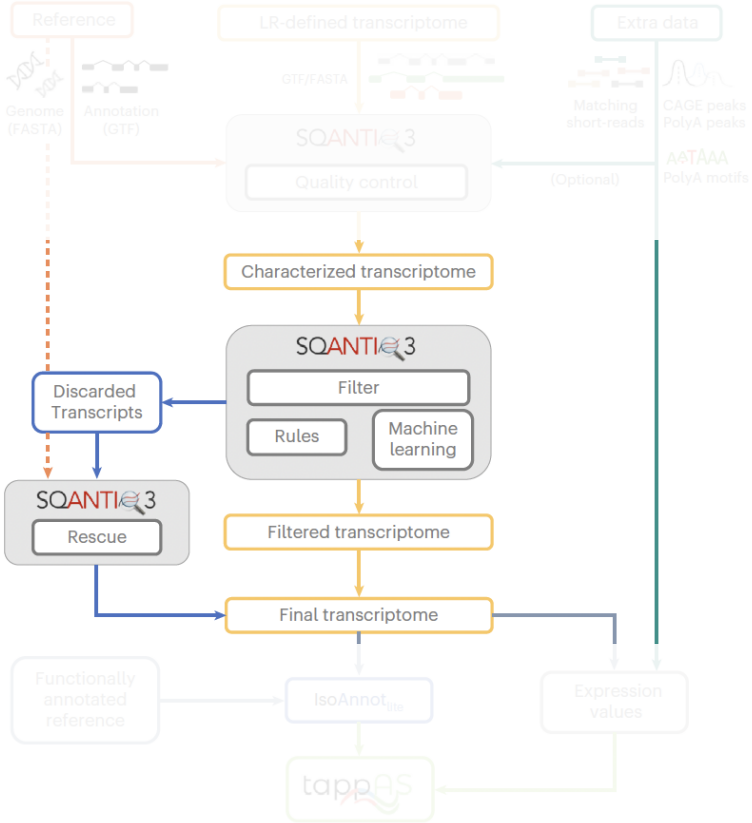
11. `diff_to_TSS` : distanc
downstream of TSS, s
12. `diff_to_TTS` : distanc
upstream of TTS, so th
13. `diff_to_gene_TSS` : d
This field is different fr
starts downstream of
14. `diff_to_gene_TTS` : d
Negative value means
15. `subcategory` : additio
`exon` . Intron rententio
16. `RTS_stage` : TRUE if o
17. `all_canonical` : TRUE
18. `min_sample_cov` : sam
19. `min_cov` : minimum ju
20. `min_cov_pos` : the jun
21. `sd_cov` : standard dev
22. `FL` or `FL.<sample>` :
23. `n_indels` : total numb
24. `n_indels_junc` : numl
potentially unreliable j
25. `bite` : TRUE if contair
26. `iso_exp` : short read e
27. `gene_exp` : short read
`expression` is provide
28. `ratio_exp` : ratio of `i`
29. `FSM_class` : This featu

30. `coding` : Coding potential capacity acc
31. `ORF_length` : predicted ORF length.
32. `CDS_length` : predicted CDS length. It
33. `CDS_start` : CDS start.
34. `CDS_end` : CDS end.
35. `CDS_genomic_start` : genomic coordin
36. `CDS_genomic_end` : genomic coordinat
37. `predicted_NMD` : TRUE if there's a prec
otherwise. NA if non-coding.
38. `perc_A_downstreamTTS` : percent of ge
the 3' end site of this isoform is probab
39. `seq_A_downstream_TTS` : sequence of t
40. `dist_to_CAGE_peak` : distance to close
means downstream of TSS. Strand-spe
PacBio transcript start site. Will be `NA`
41. `within_CAGE_peak` : TRUE if the transc
42. `dist_to_polyA_site` : distance to the
43. `within_polyA_site` : TRUE if the trans
seq).
44. `polyA_motif` : if `--polyA_motif_list`
45. `polyA_dist` : if `--polyA_motif_list` i
putative poly(A) site. This distance is h
46. `polyA_motif_found` : TRUE if a polyA m
47. `ORF_seq` : Predicted ORF sequence. Th
48. `ratio_TSS` : Using Short-Read data, w

8. `junction_category` : `known` if the
that it is possible to have a `novel` j
combination might be novel.
9. `start_site_category` : `known` if th
10. `end_site_category` : `known` if the
11. `diff_to_Ref_start_site` : distance
site.
12. `diff_to_Ref_end_site` : distance t
site.
13. `bite_junction` : Applies only to no
value is TRUE, otherwise it is FALSE
14. `splice_site` : Splice motif.
15. `RTS_junction` : TRUE if junction is
16. `indel_near_junct` : TRUE if there i
incorrectness.
17. `sample_with_cov` : If `--coverage`
files) that have uniquely mapped sh
18. `total_coverage_unique/multi` : To
cover this junction.

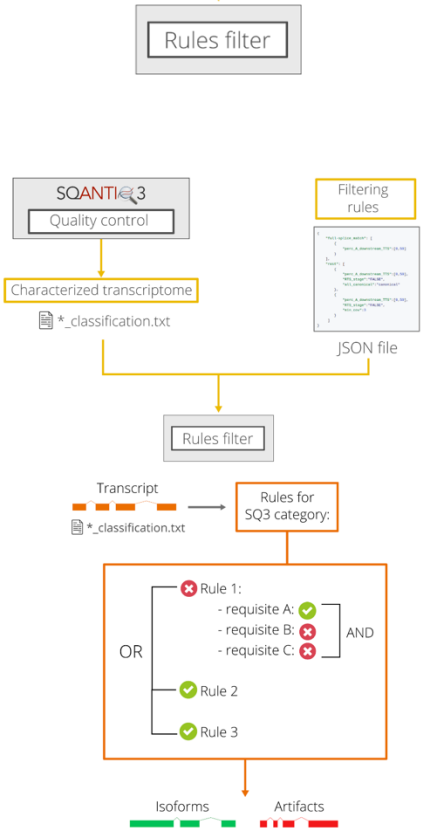
Pardo-Palacios *et al.* **Nat Methods**, *2024*

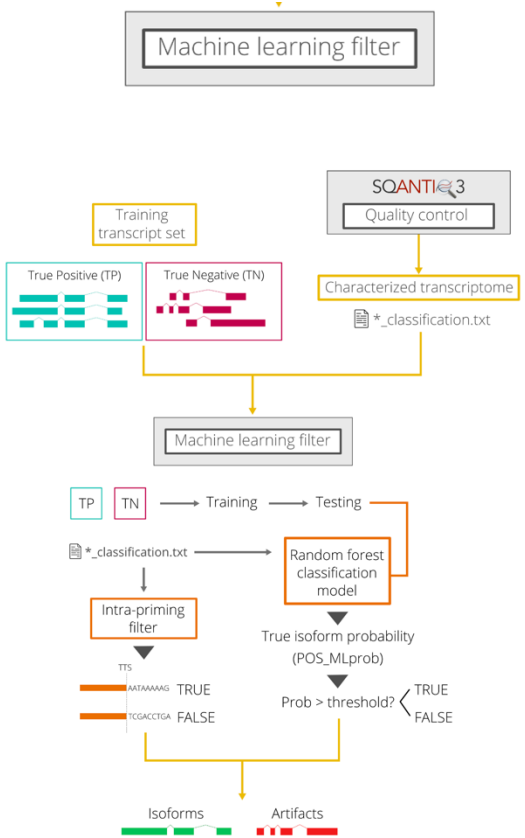
SQANTI3 for quality control of transcript models

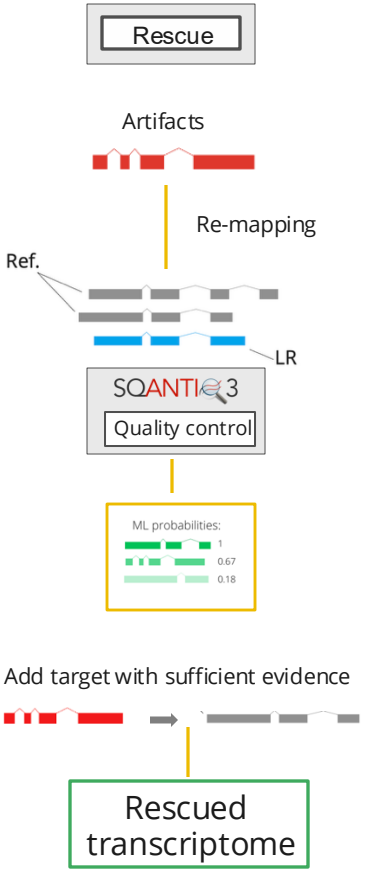


Arzalluz-Luque *et al.* **Nat Methods**, *2024*

SQANTI curation to improve the transcriptome



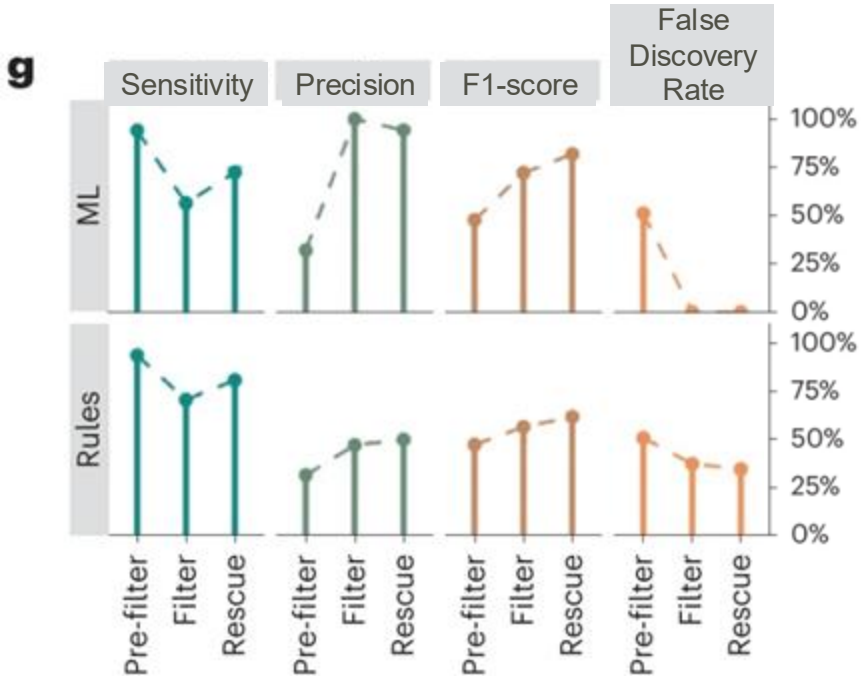




Evaluation of the SQANTI curation approach

Spike-ins

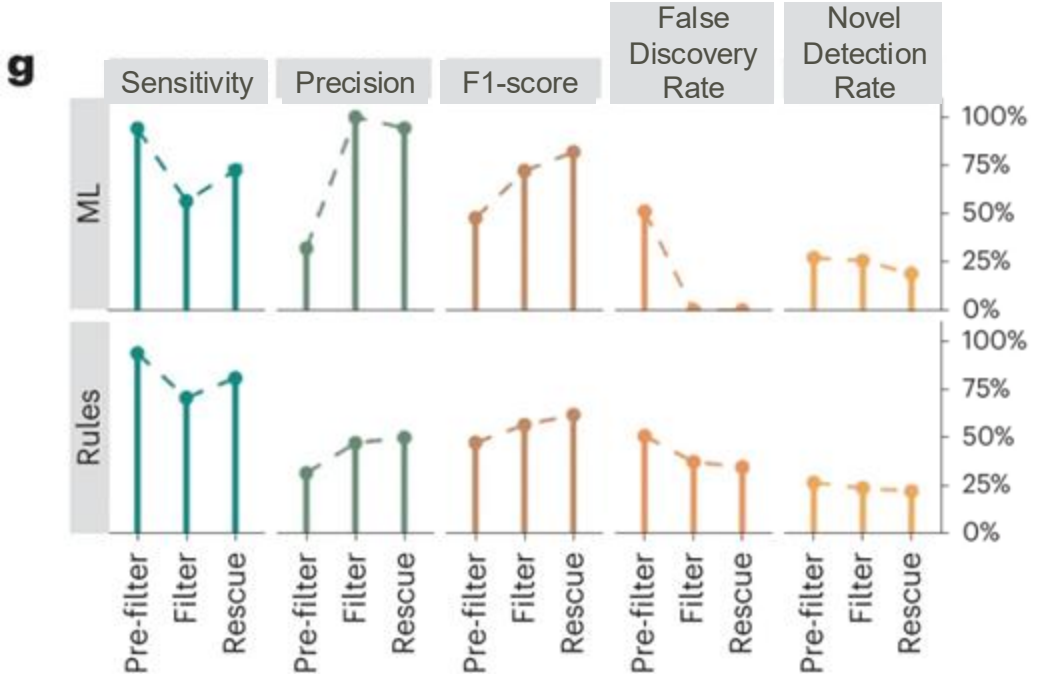
Ground truth



Evaluation of the SQANTI curation approach

Spike-ins

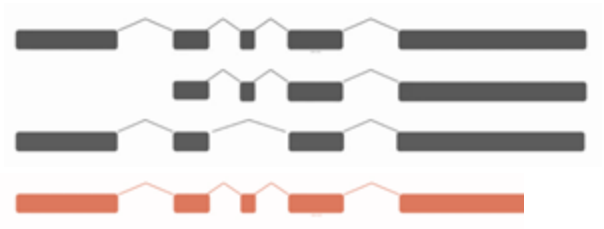
Ground truth

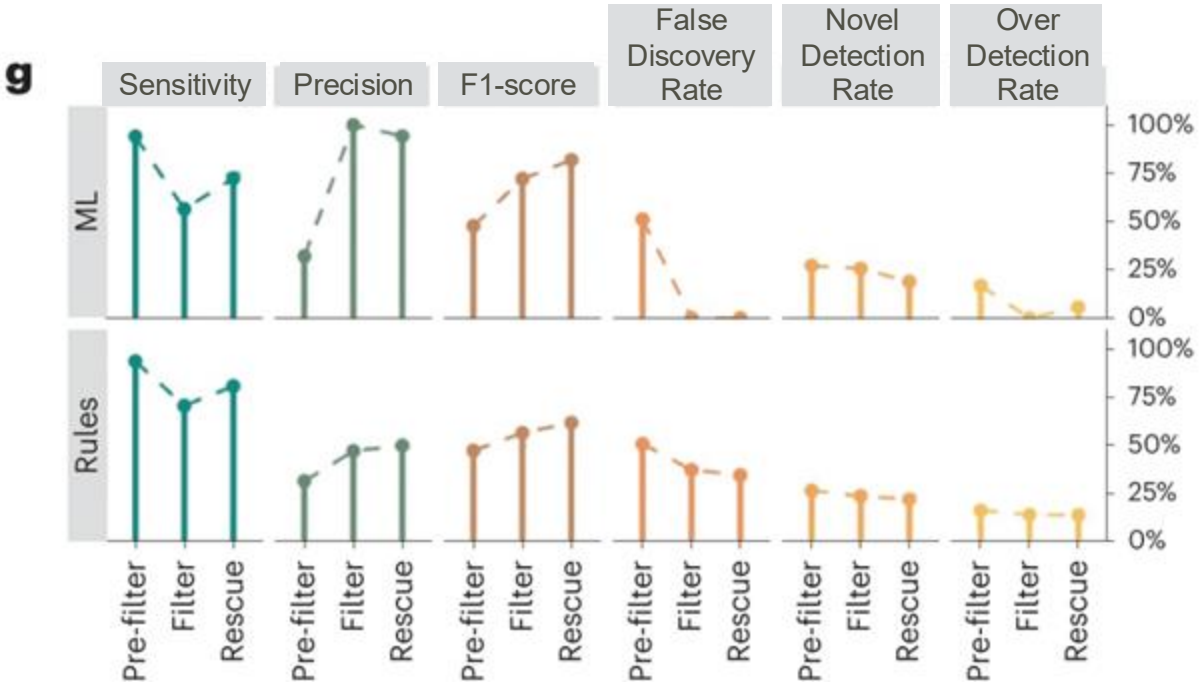


Evaluation of the SQANTI curation approach

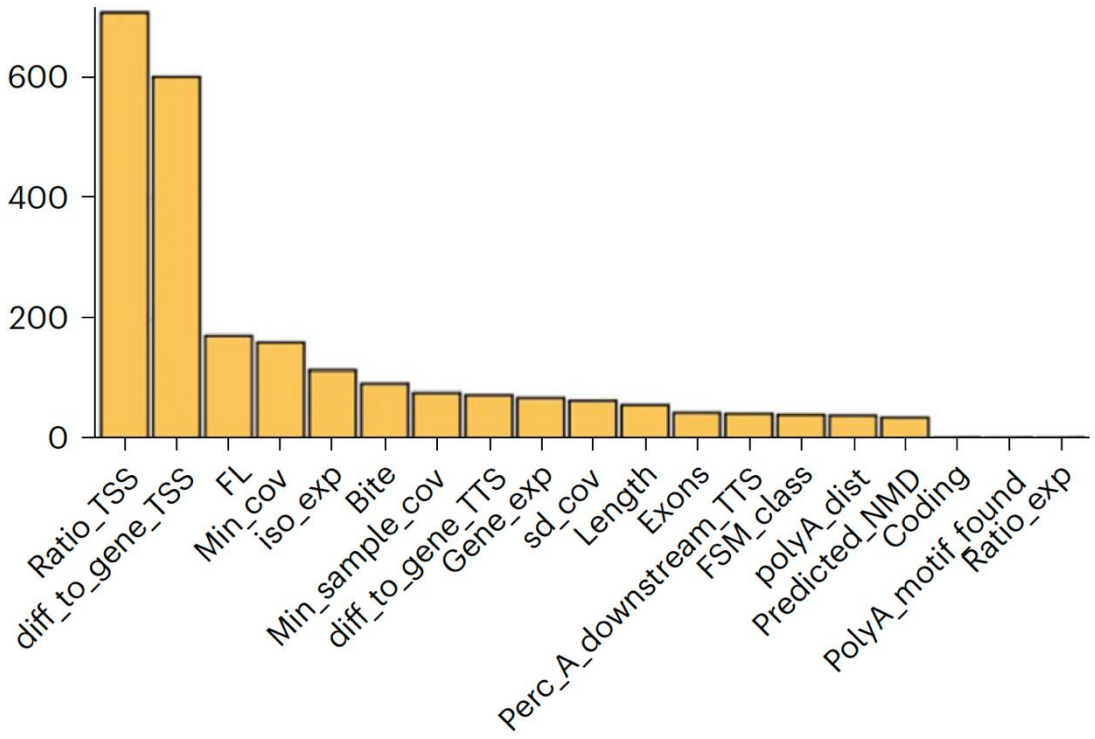
Spike-ins

Ground truth





Which variables are important in filtering?



Pardo-Palacios et al. **Nat Methods**, *2024*

Follow up- summary

- SQANTI-SIM and SQANTI-TUSCO are useful benchmarking tools for lrRNA-seq transcript reconstruction methods.
- SQANTI-curation provides an effective strategy to use complementary data to curate lr-RNAseq transcript models
- ISM are a mix of transcript degradation and true new transcripts and require careful consideration.

SQANTI verse

Genomics of Gene Expression Lab

Quality control

SQANTI3

SQANTI curation

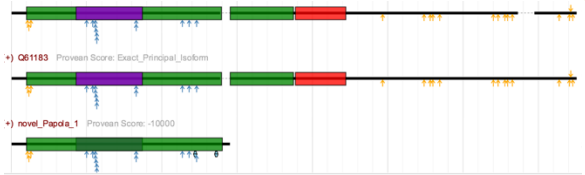
SQANTI reads

SQANTI proteins

SQANTI sim

SQANTI tusco

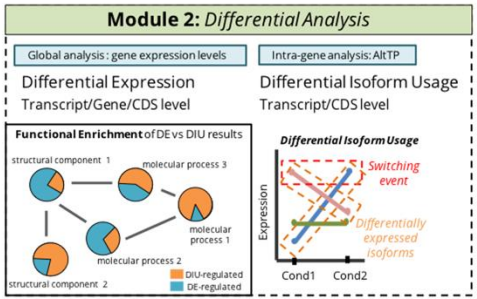
Annotation



SQANTI isoannot

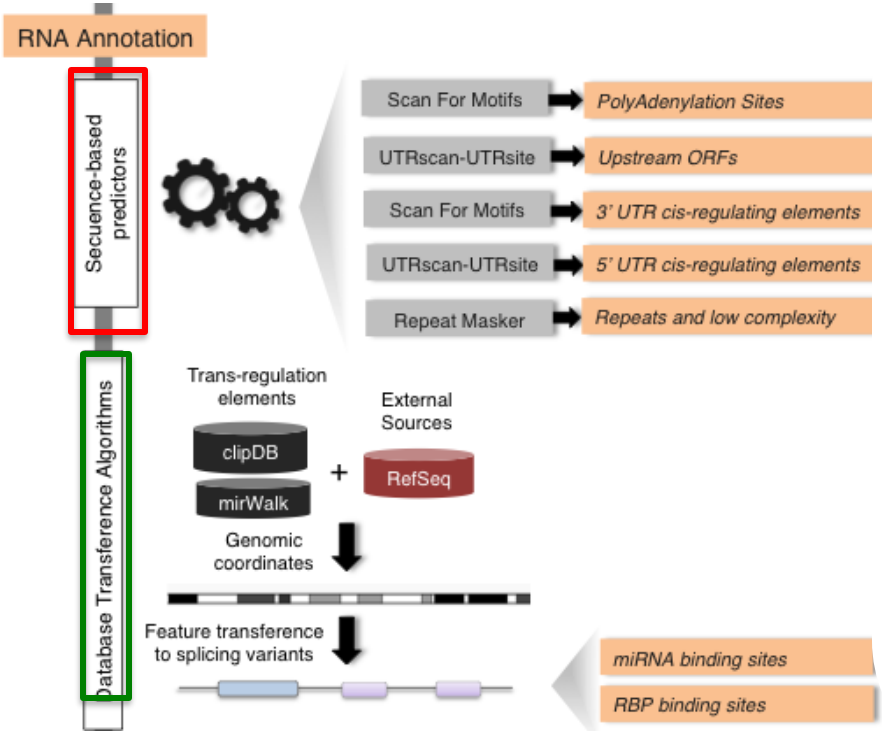
SQANTI evidence

Analysis



tappAS

IsoAnnot: Functional annotation at isoform resolution

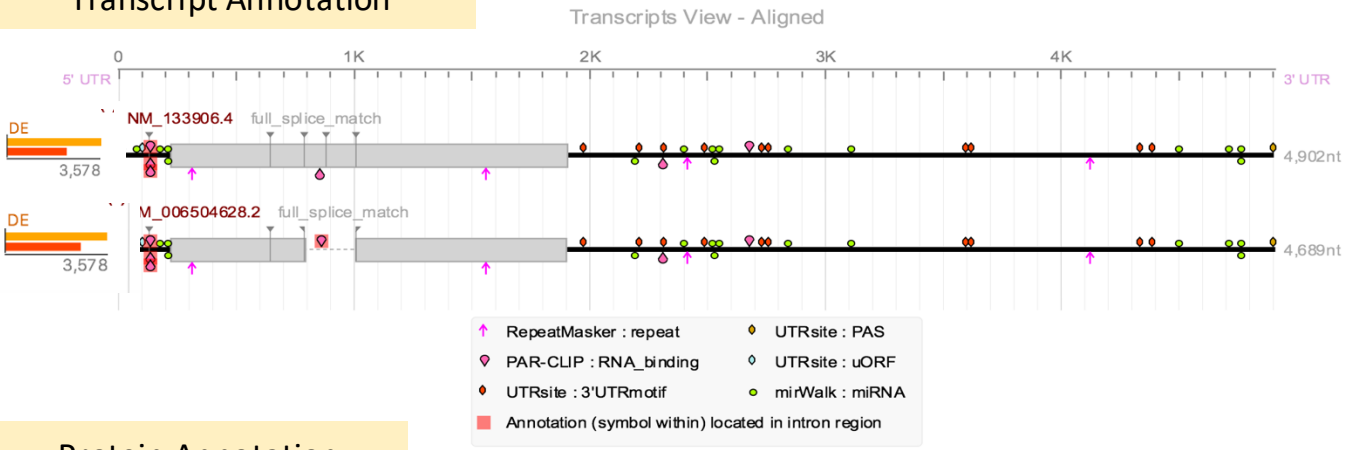


IsoAnnot-Lite is a SQANTI parameter. Allows functional annotation for human, mouse, ATH and *Drosophila*

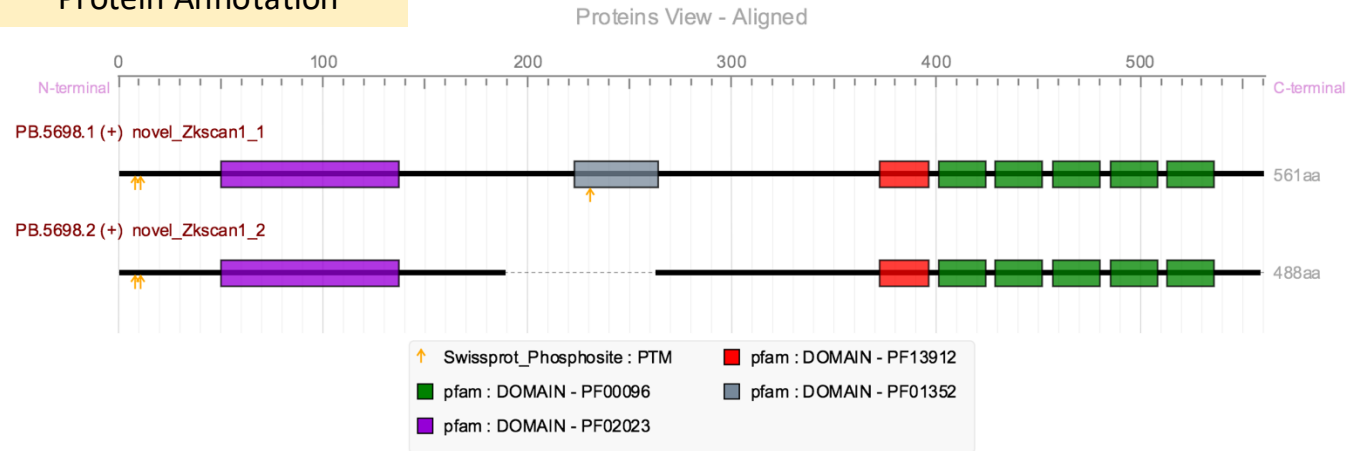
Functional annotation at isoform resolution

SQANTI isoannot

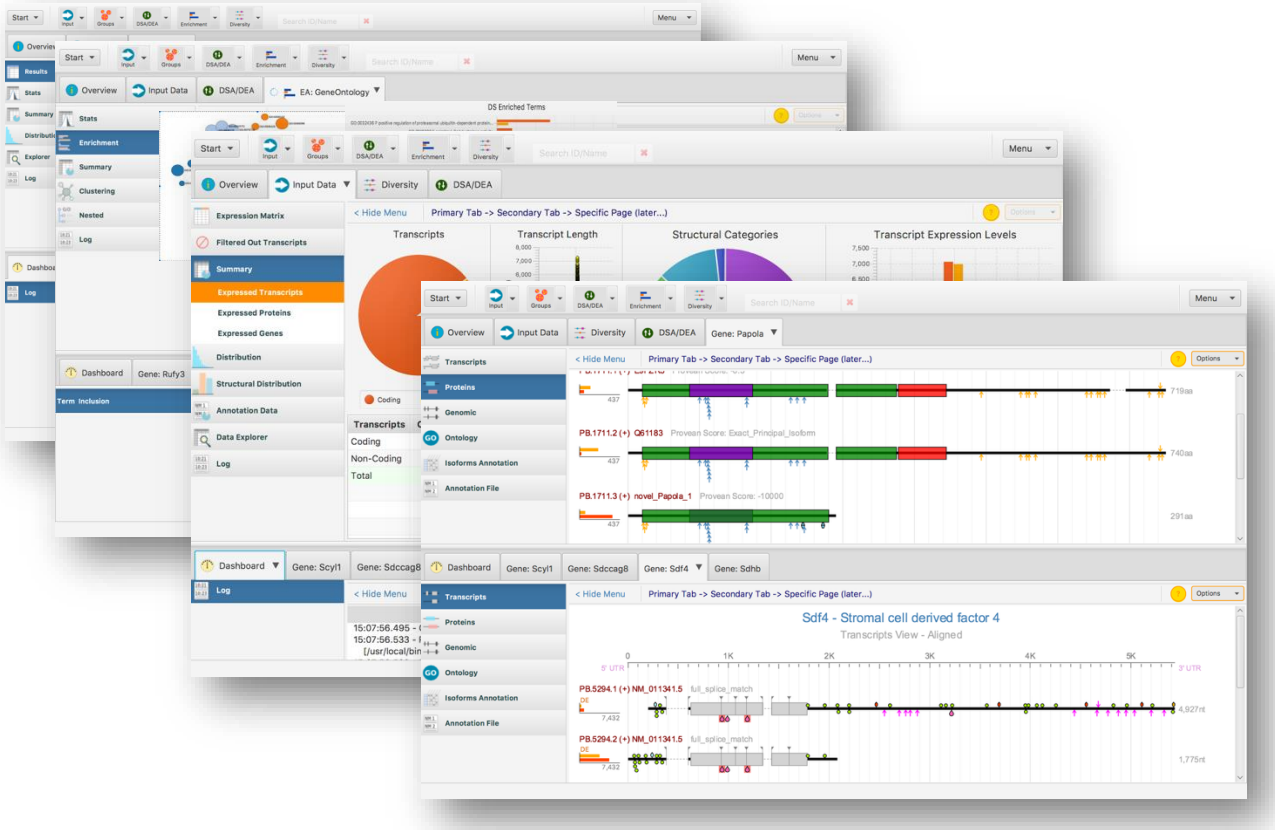
Transcript Annotation



Protein Annotation



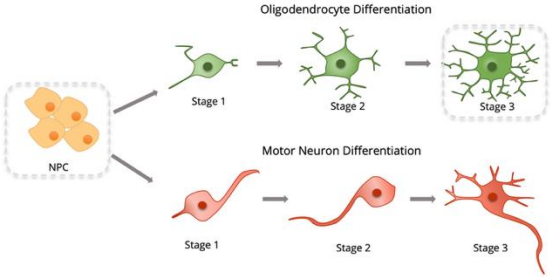
tappAS for functional analysis of isoform function

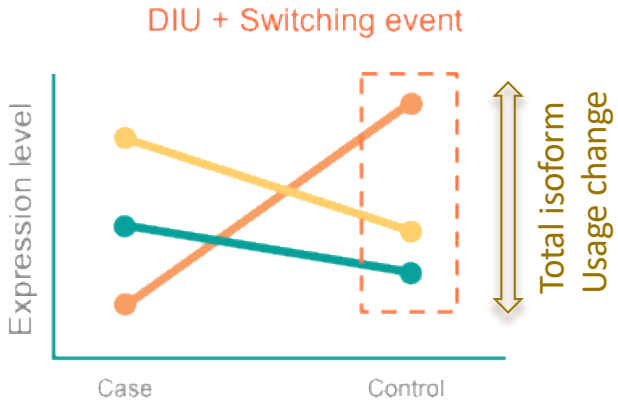


de La Fuente *et al.* **Genome Biology**, 2021

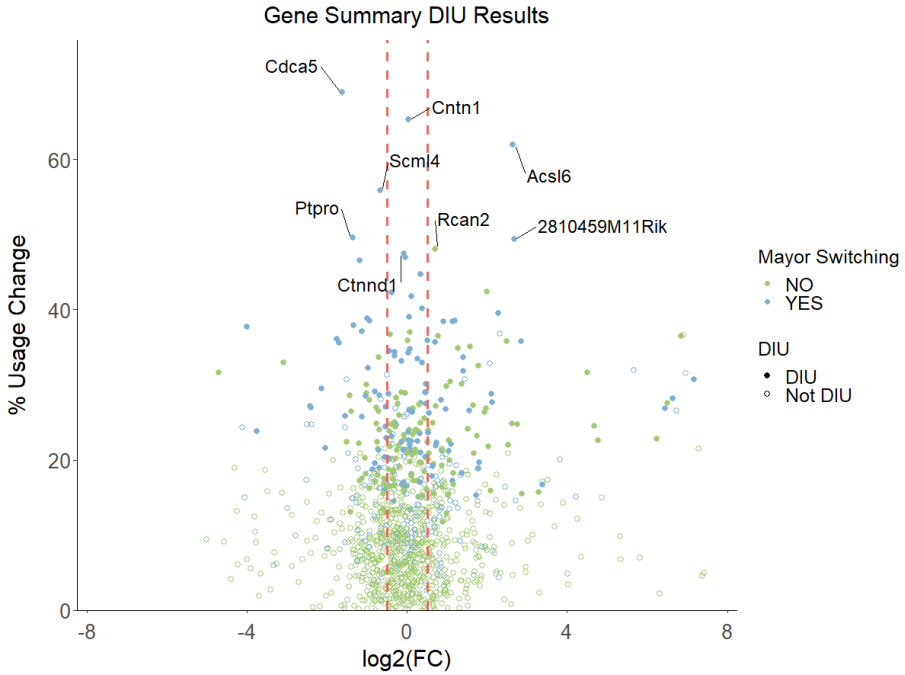
Evaluating the magnitude of changes in isoform usage

Genomics of Gene Expression Lab

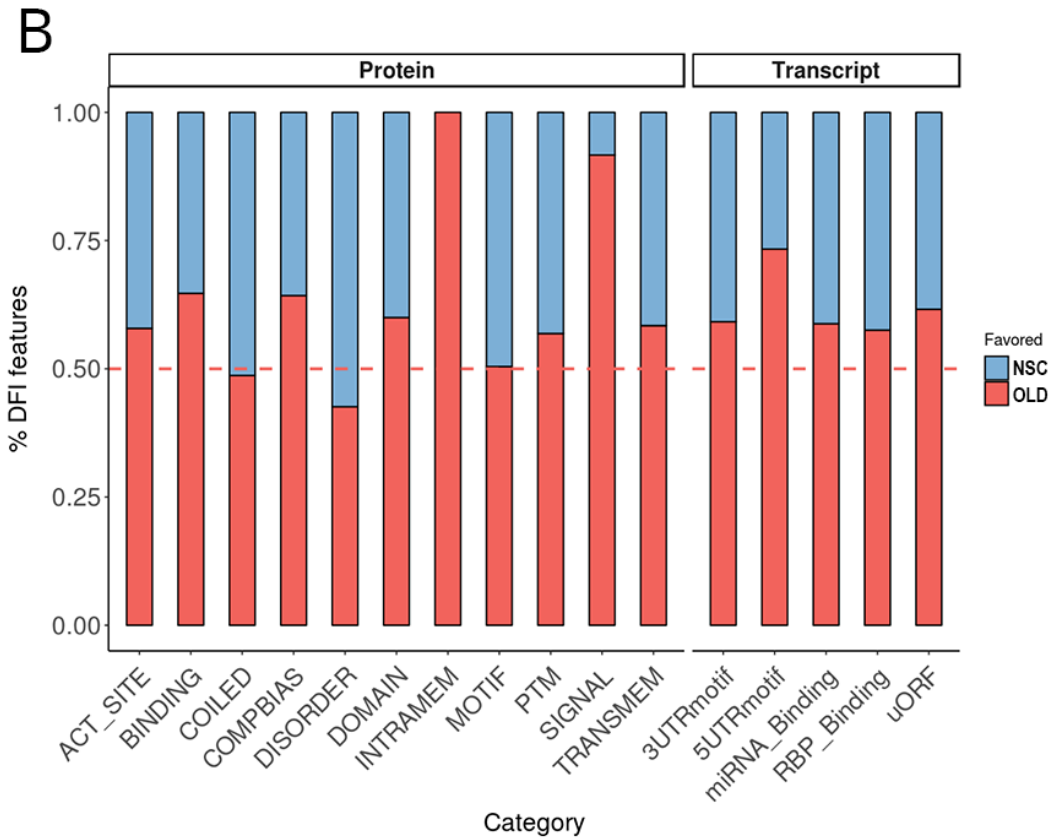




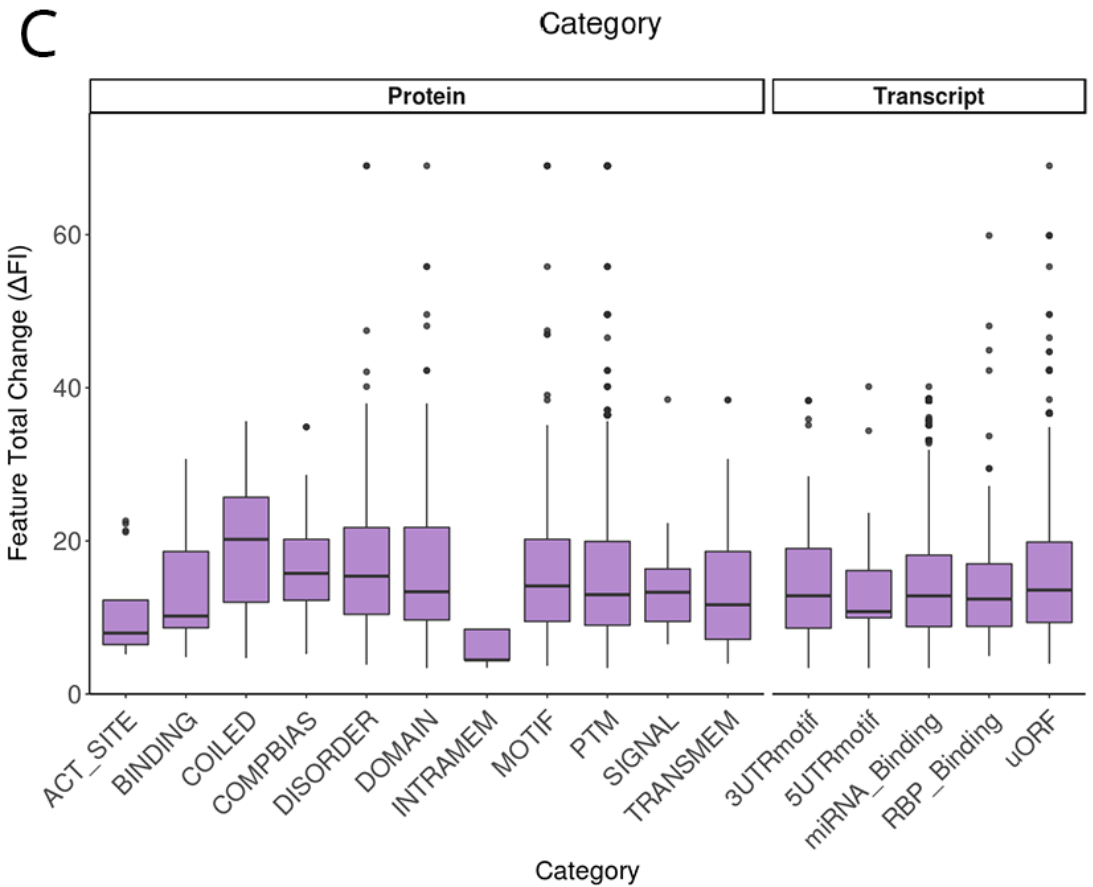
Many genes with AS are not differentially expressed



Mbnl1 accumulates in cytosol in OLIGOs



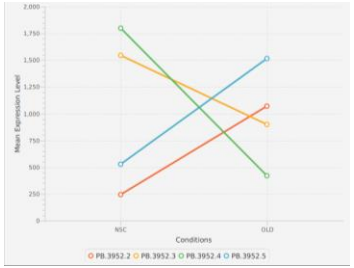
Generally moderate magnitude of AS-driven functional changes



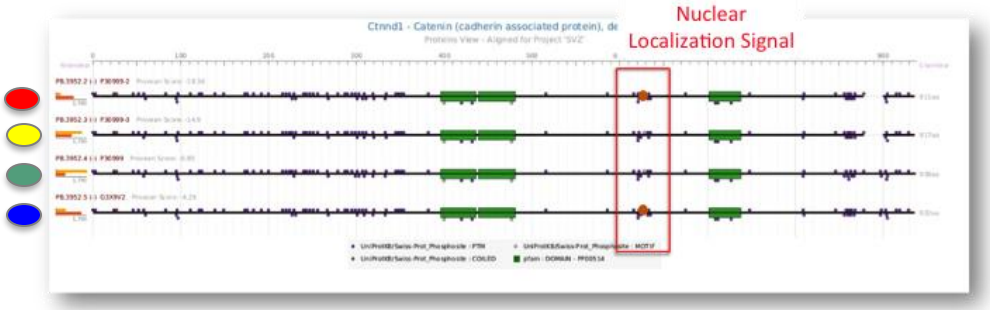
Ctnnd1: Isoform switch regulating protein localization

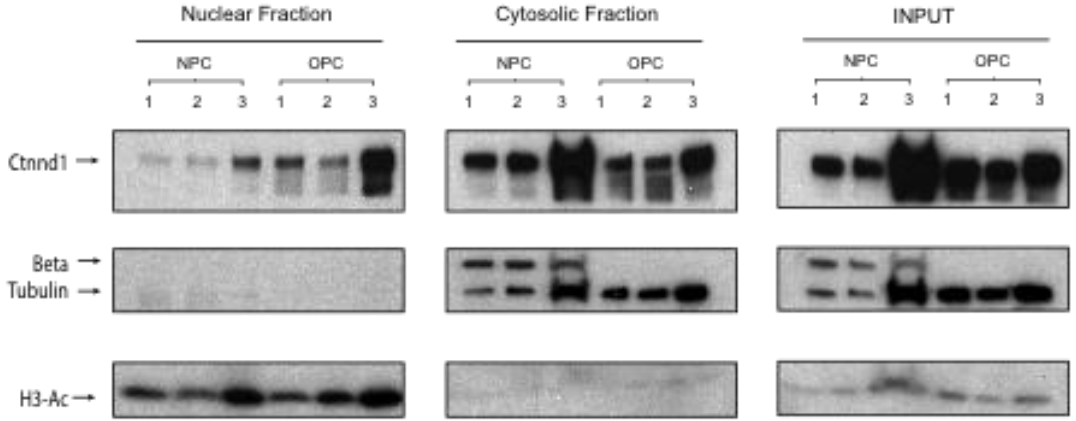
Genomics of Gene Expression Lab

Significant isoform switch

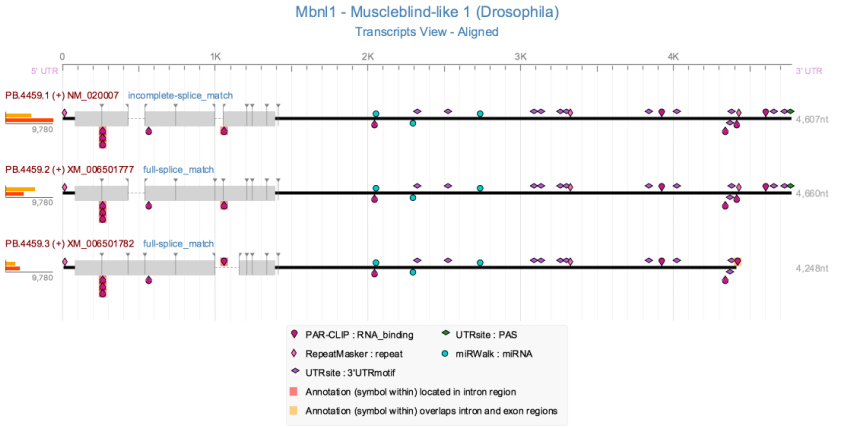


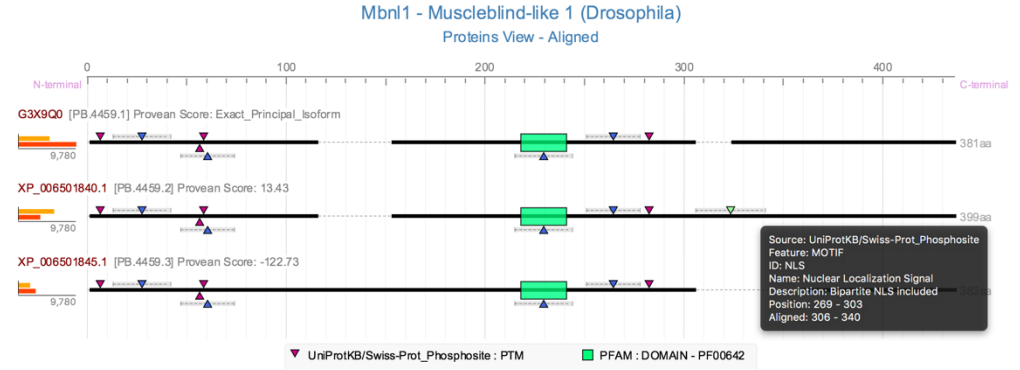
NPC major isoform is excluded from nucleus

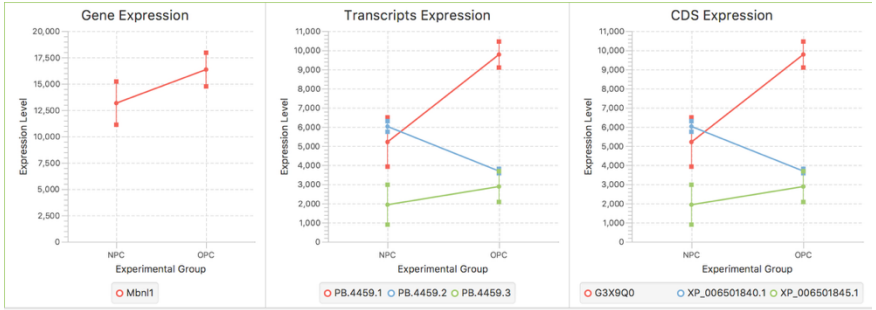


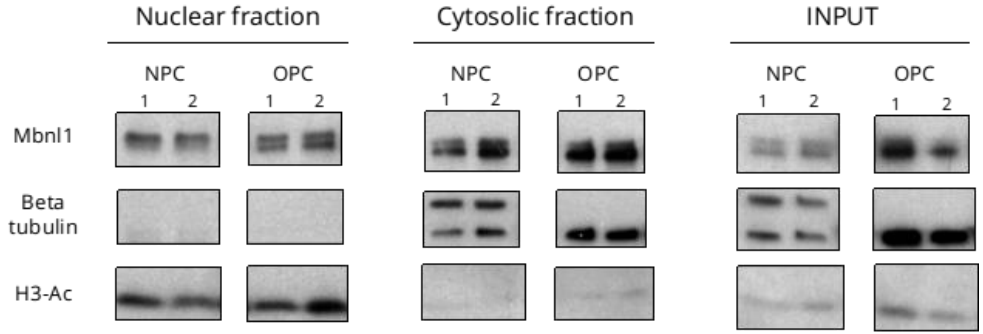


Mbnl1 accumulates in cytosol in OLIGOs

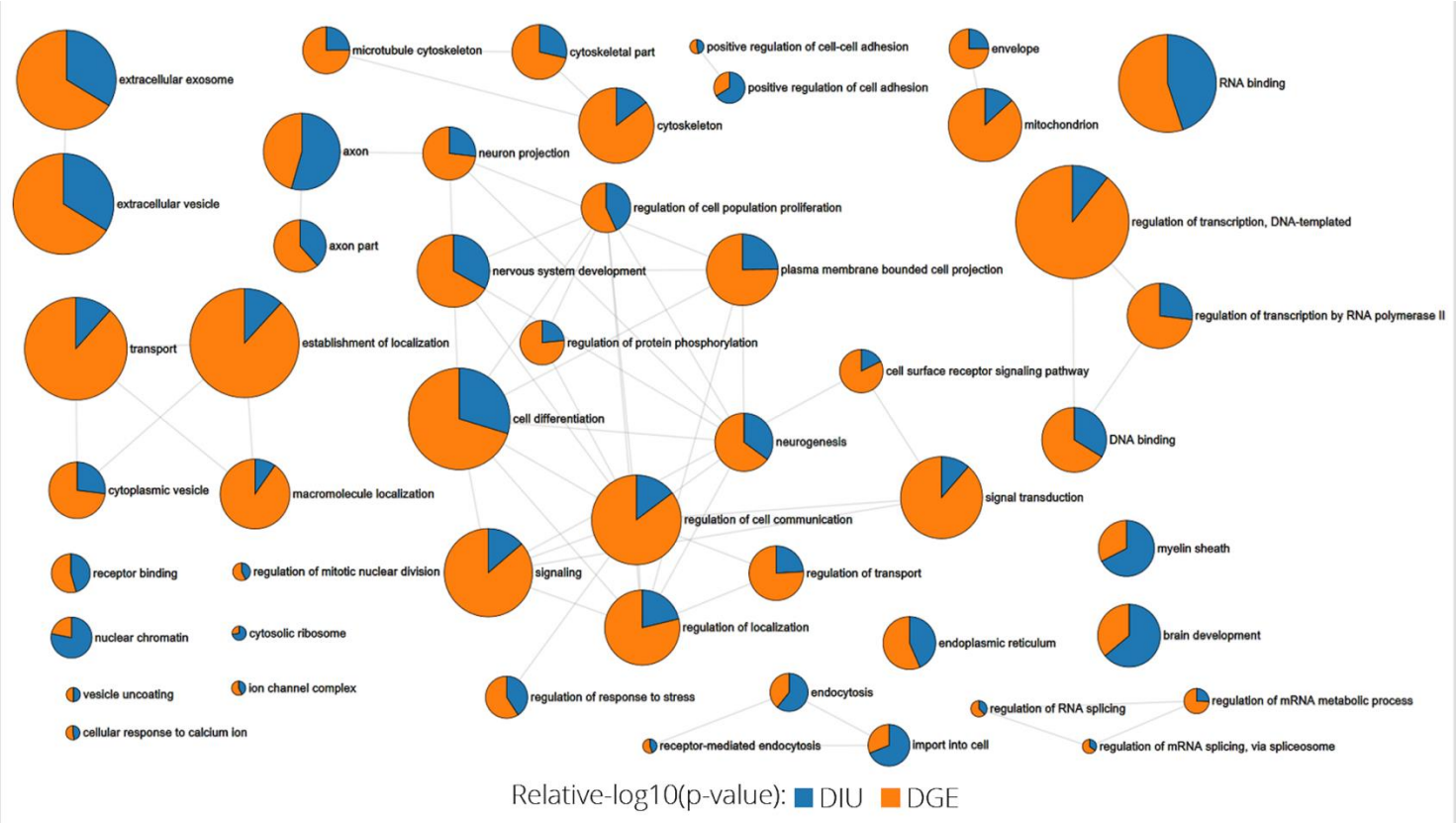




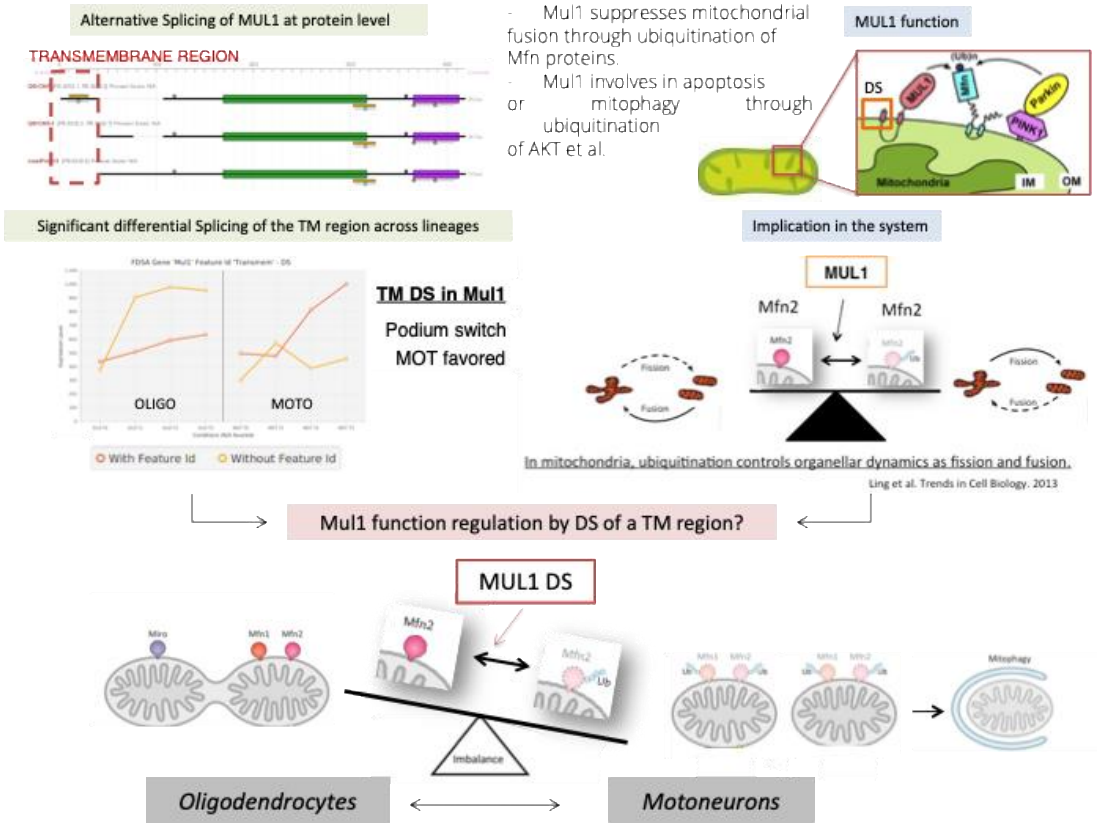




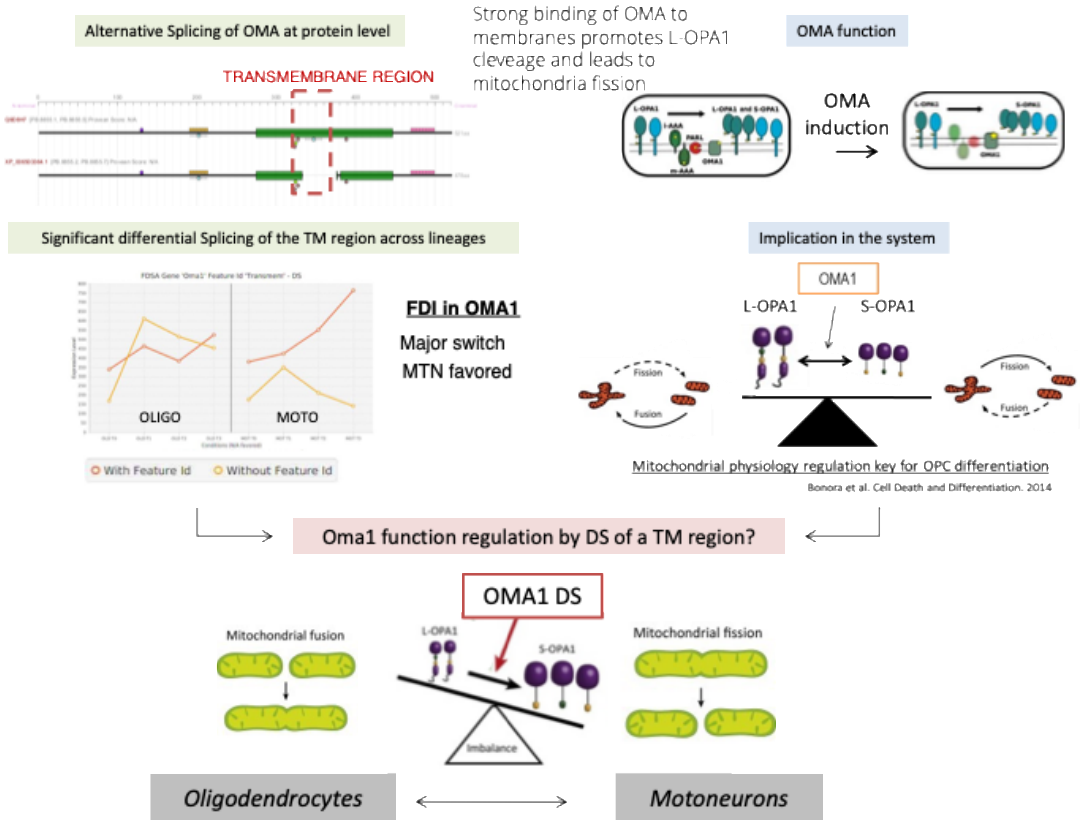
Differential Isoform Usage (DIU) and Gene Expression (DGE)



New hypothesis in the role of AS for mitochondrial function

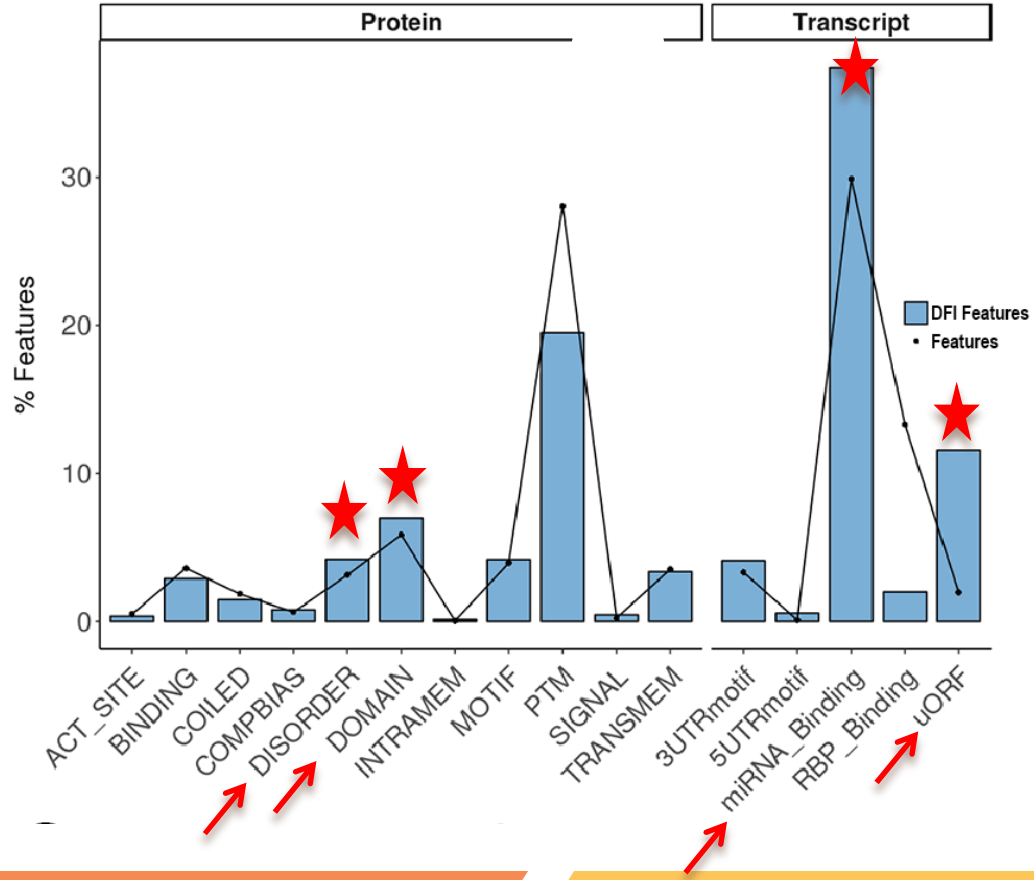


New hypothesis in the role of AS for mitochondrial function



Distinct motifs are regulated post-transcriptionally

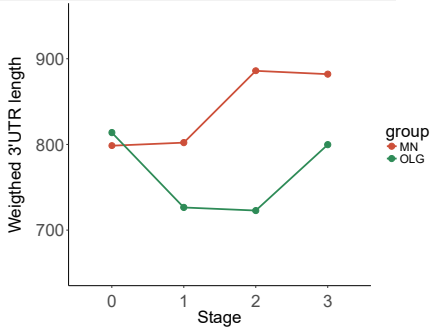
Genomics of Gene Expression Lab



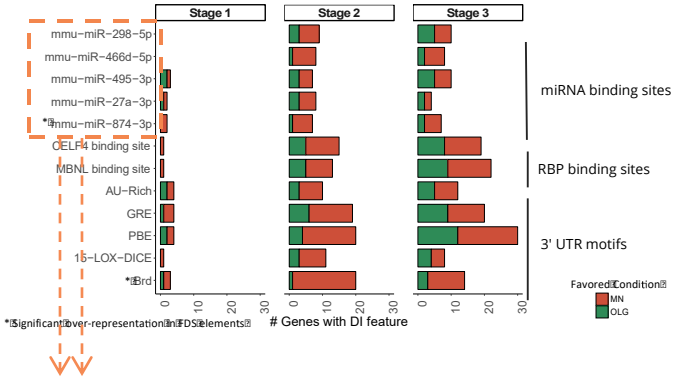
Differential PolyAdenylation Analysis (DPA)

Genomics of Gene Expression Lab

Analysis of UTR length

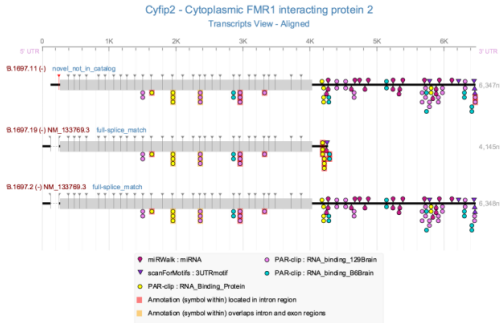


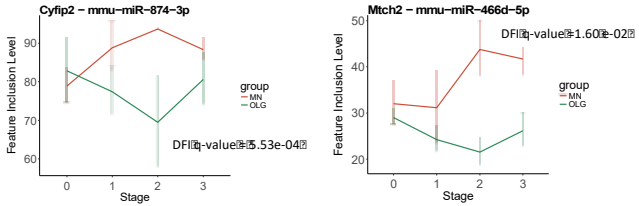
DFI of UTR functional features



miRNAs involved in axon regeneration

Motti et al. 2017

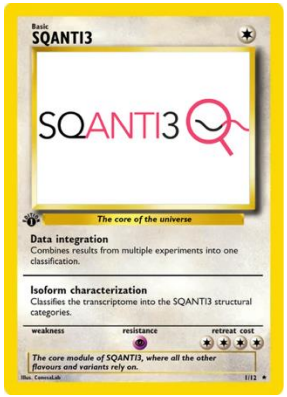


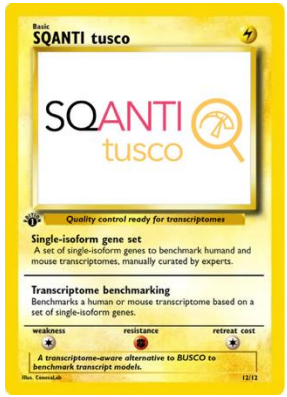


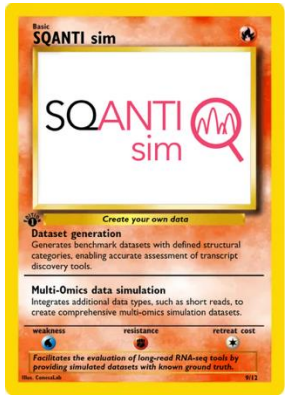
Final wrap-up

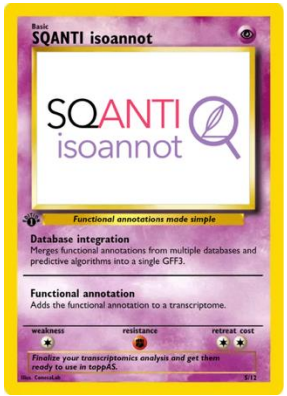
- lrRNA-seq can boost the study of the functional impact of **differential isoform usage**
- We find that changes in expression levels and AS **jointly** control cellular processes
- **Tools to annotate and visualize** the data can help to postulate hypothesis about alternative isoform function
- lrRNA-seq rocks! You just need to play the right cards!!

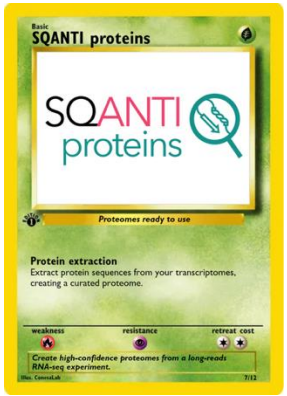
And now... SQANTI cards!!!

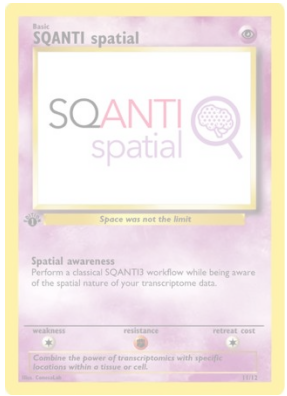


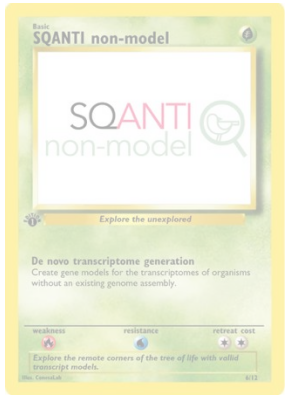


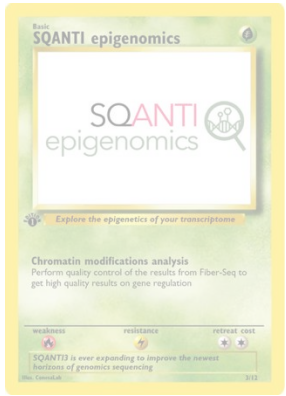












Acknowledgements

Collaborators:

Sonia Tarazona: Polytechnical University of Valencia

Victoria Moreno and Maria Jesus Vicent: Prince Felipe Research Center

Angela Brooks & LRGASP: University of California at Santa Cruz

Elizabeth Tseng: Pacific Biosciences

Lauren McIntyre: University of Florida

Lab members

Carol Monzó
Alejandro Paniagua
Tian Liu
Quique Vidal
Clara Rodríguez
Virtudes Robledo
Carlos Blanco
Fabian Robledo
Víctor Gaya
Julen Santiago
Pablo Atienza
Julia Liénard
Maite Benlloch
Priyansh Srivastava
Fabian Jetzinger

Past members

Sonia Tarazona
Manuel Ugidos,
Pedro Salguero
Lorena de la Fuente
Angeles Arzalluz
Carlos Martínez,
Héctor Carmona,
Alberto Lerma,
Tatyana Zamkovaya,
Jorge Mestre
Myla Kondratova
Leandro Balzano,
Salva Casaní
Fran Pardo
Rocío Amorín
Cristina Martí,
Víctor Sánchez,
Raymond Scott,
Paco Huerta
Roció Amorín
Wouter Maessen