Unlocking the Full Transcriptome: Advances in Long-read Transcriptomics

Current Trends in Long Read Sequencing and Bioinformatics Analysis
03/10/2024

Júlia Faura, PhD – VIB-UAntwerp Center for Molecular Neurology

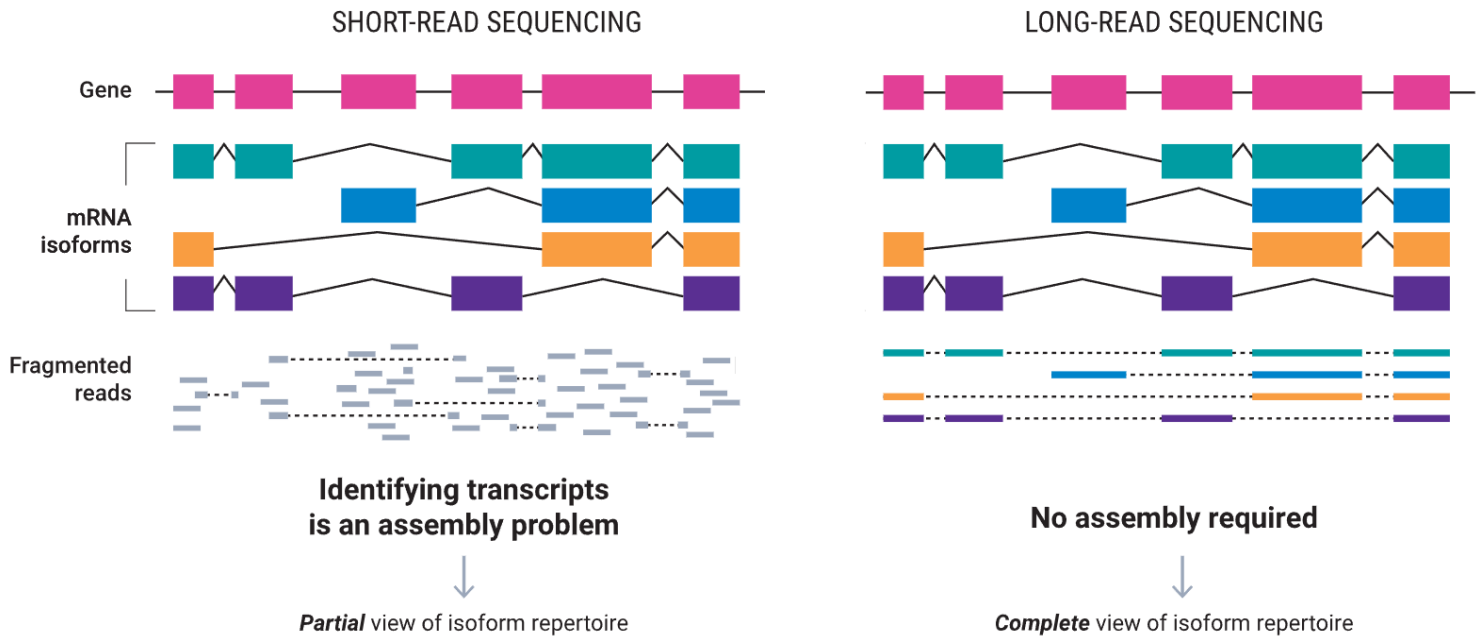
VIB-UANTWERP CENTER FOR MOLECULAR NEUROLOGY

University of Antwerp

SCIENCE MEETS LIFE

INTRODUCTION

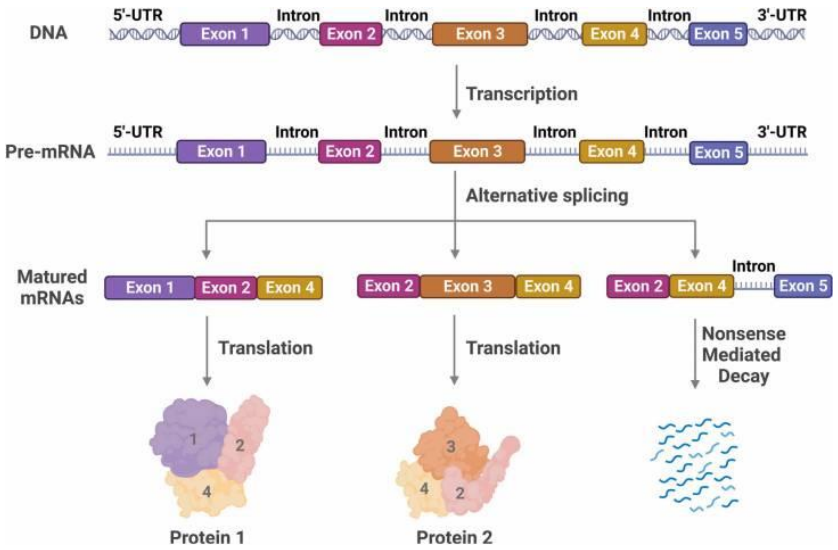
- Traditional short-read sequencing technologies have played a key role in advancing transcriptomics but are not able to fully capturing the structures of RNA molecules.
- This can lead to difficulties in identifying full-length transcripts, detecting isoforms, or resolving complex splicing events.
- Long-read sequencing allows identification of full-length transcripts, capturing all splice junctions with no assembly required.



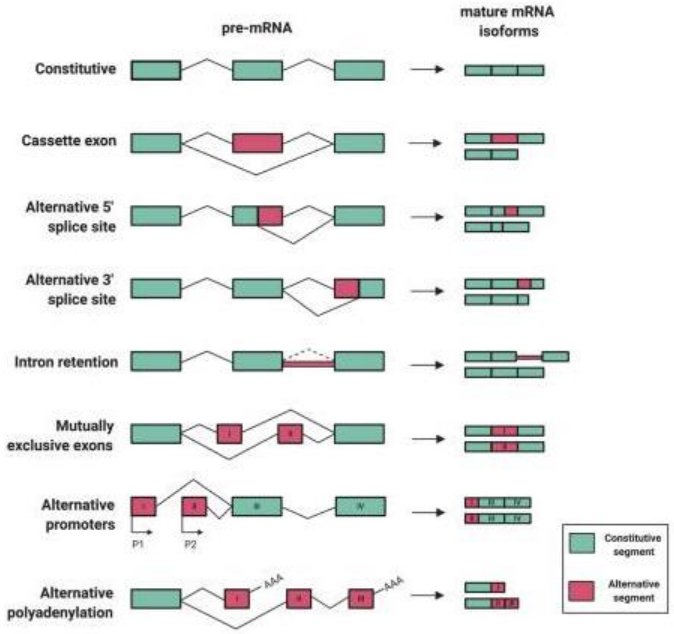
https://www.pacb.com/products-and-services/applications/rna-sequencing/

Why is it relevant to study the gene isoforms and splicing?

- Alternative splicing of mRNA transcripts is a mechanism by which several isoforms can be generated from individual genomic loci, enabling significant increases in transcriptomic and proteomic complexity.
- Abnormal expressions of specific splicing isoforms can lead to disease but also can serve as a biomarkers and therapeutic targets, e.g. in neurological diseases or cancer.



Wu et al. (2023)



Liu et al. (2022)

Available technologies for LR-RNAseq

- Both PacBio and ONT have kits and have adapted their technologies for transcriptomics.

PacBio

Iso-Seq method:

- Sequencing of full-length cDNA using PacBio SMRT seq technology.

ONT

3 different types of sequencing:

- Direct RNA → it can also allow the study of RNA modifications (m6A)
- Direct cDNA
- PCR-cDNA

Analysis pipeline of long-read transcriptomics data

1. Alignment

Analysis pipeline of long-read transcriptomics data

1. Alignment

2. Transcript identification and quantification – Also at a gene level

FLAIR

Article | Open access | Published: 18 March 2020

Full-length transcript characterization of *SF3B1* mutation in chronic lymphocytic leukemia reveals downregulation of retained introns

Alison D. Tang, Cameron M. Soulette, Marijke J. van Baren, Kevyn Hart, Eva Hrabeta-Robinson, Catherine J. Wu & Angela N. Brooks

Nature Communications **11**, Article number: 1438 (2020) | Cite this article

26k Accesses | **200** Citations | **35** Altmetric | Metrics

Method | Open access | Published: 16 December 2019

Transcriptome assembly from long-read RNA-seq alignments with StringTie2

Sam Kovaka, Aleksey V. Zimin, Geo M. Pertea, Roham Razaghi, Steven L. Salzberg & Mihaela Pertea

Genome Biology **20**, Article number: 278 (2019) | Cite this article

41k Accesses | **818** Citations | **136** Altmetric | Metrics

StringTie2

Bambu

Article | Published: 12 June 2023

Context-aware transcript quantification from long-read RNA-seq data with Bambu

Ying Chen, Andre Sim, Yuk Kei Wan, Keith Yeo, Joseph Jing Xian Lee, Min Hao Ling, Michael I. Love & Jonathan Göke

Nature Methods **20**, 1187–1195 (2023) | Cite this article

9449 Accesses | **24** Citations | **103** Altmetric | Metrics

Brief Communication | Open access | Published: 02 January 2023

Accurate isoform discovery with IsoQuant using long reads

Andrey D. Prjibelski, Alla Mikheenko, Anoushka Joglekar, Alexander Smetanin, Julien Jarroux, Alla L. Lapidus & Hagen U. Tilgner

Nature Biotechnology **41**, 915–918 (2023) | Cite this article

23k Accesses | **34** Citations | **82** Altmetric | Metrics

IsoQuant

Analysis pipeline of long-read transcriptomics data

1. Alignment

2. Transcript identification and quantification – Also at a gene level

FLAIR

Article | Open access | Published: 18 March 2020

Full-length transcript characterizati
mutation in chronic lymphocytic leu
downregulation of retained introns

Alison D. Tang, Cameron M. Soulette, Marijke J. van Baren, Kevyn Hart, Wu & Angela N. Brooks

Nature Communications **11**, Article number: 1438 (2020) | Cite this a

26k Accesses | 200 Citations | 35 Altmetric | Metrics

Registered Report | Open access | Published: 07 June 2024

Systematic assessment of long-read RNA-seq methods for transcript identification and quantification

Francisco J. Pardo-Palacios, Dingjie Wang, Fairlie Reese, Mark Diekhans, Sílvia Carbonell-Sala, Brian Williams, Jane E. Loveland, Maite De María, Matthew S. Adams, Gabriela Balderrama-Gutierrez, Amit K. Behera, Jose M. Gonzalez Martinez, Toby Hunt, Julien Lagarde, Cindy E. Liang, Haoran Li, Marcus Jerryd Meade, David A. Moraga Amador, Andrey D. Prjibelski, Inanc Birol, Hamed Bostan, Ashley M. Brooks, Muhammed Hasan Çelik, Ying Chen, ... Angela N. Brooks + Show authors

Nature Methods **21**, 1349–1363 (2024) | Cite this article

ification from long-

an Lee, Min Hao Ling, Michael I. Love &

Method | Open access | Published: 16 December 2019

Transcriptome assembly from l
alignments with StringTie2

Sam Kovaka, Aleksey V. Zimin, Geo M. Pertea, Roham Razagh

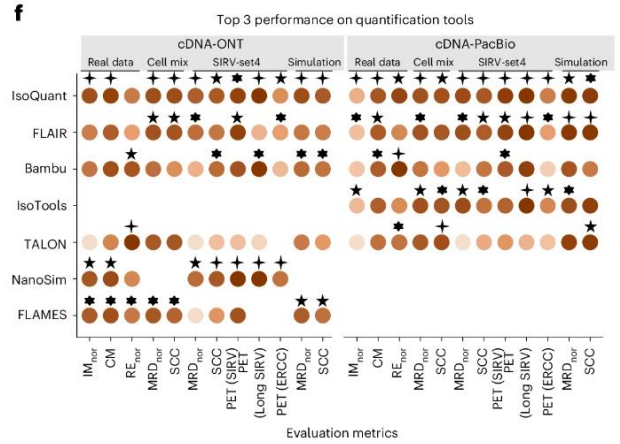
Genome Biology **20**, Article number: 278 (2019) | Cite this a

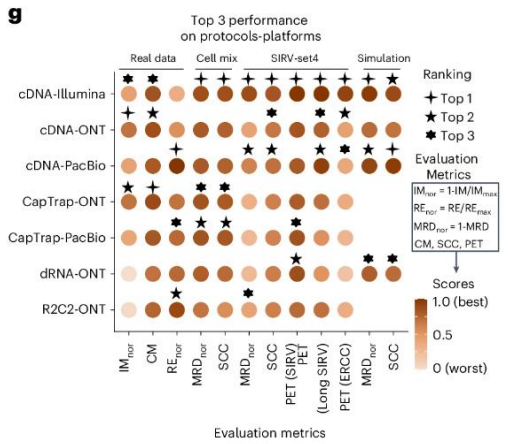
41k Accesses | 818 Citations | 136 Altmetric | Metrics

StringTie2

Quant using long

er Smetanin, Julien Jarroux, Alla L.





VIB-UANTWERP CENTER FOR MOLECULAR NEUROLOGY

University of Antwerp

SCIENCE MEETS LIFE

Analysis pipeline of long-read transcriptomics data

1. Alignment
2. Transcript identification and quantification – Also at a gene level
3. Quality control and curation of known and novel transcripts

Brief Communication | Open access | Published: 20 March 2024

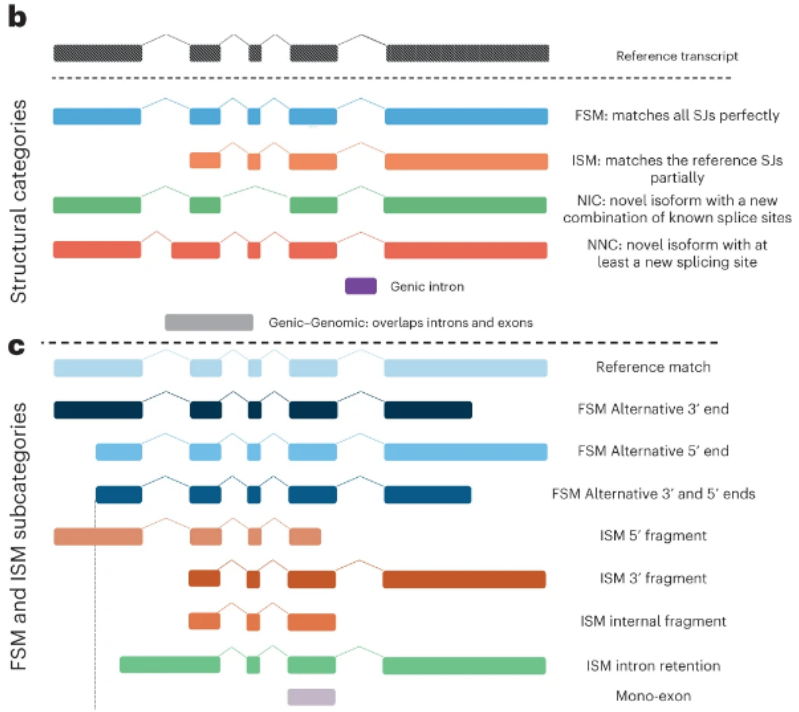
SQANTI3: curation of long-read transcriptomes for accurate identification of known and novel isoforms

Francisco J. Pardo-Palacios, Angeles Arzalluz-Luque, Liudmyla Kondratova, Pedro Salguero, Jorge Mestre-Tomás, Rocío Amorín, Eva Estevan-Morió, Tianyuan Liu, Adalena Nanni, Lauren McIntyre, Elizabeth Tseng & Ana Conesa

Nature Methods **21**, 793–797 (2024) | Cite this article

10k Accesses | **11** Citations | **85** Altmetric | Metrics

SQANTI3 tool performs quality control and filtering of long read-defined transcriptomes, which are often rich in artifacts and false-positive isoforms



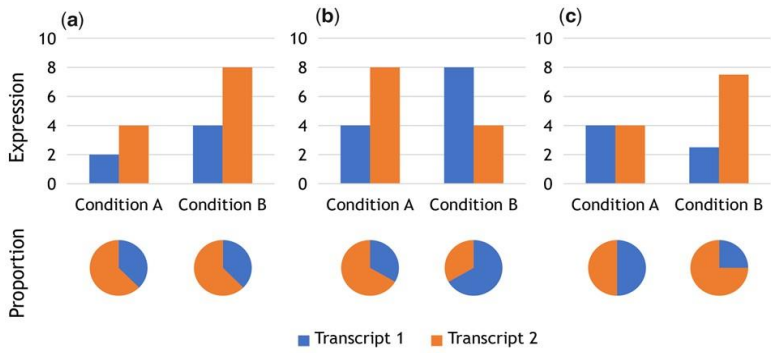
VIB-UANTWERP CENTER FOR MOLECULAR NEUROLOGY

University of Antwerp

SCIENCE MEETS LIFE

Analysis pipeline of long-read transcriptomics data

1. Alignment
2. Transcript identification and quantification – Also at a gene level
3. Quality control and curation of known and novel transcripts
4. Differential gene expression
   Differential transcript expression
   Differential transcript usage



(d)

Gene identified by:	a)	b)	c)
DGE	✓	✗	✗
DTE	✓	✓	✗
DTU	✗	✓	✓

Tekah et al. (2021)

Analysis pipeline of long-read transcriptomics data

1. Alignment
2. Transcript identification and quantification – Also at a gene level
3. Quality control and curation of known and novel transcripts
4. Differential analysis

Differential gene expression	Differential transcript expression	Differential transcript usage
DESeq2 **edgeR** **limma-voom**	**DESeq2** **edgeR** **limma-voom**	**DEXSeq** **DRIMSeq** **DTUrtle**

Illustrative example: Transcriptome variation in human tissues revealed by long-read sequencing (Glinos et al, *Nature*, 2022)

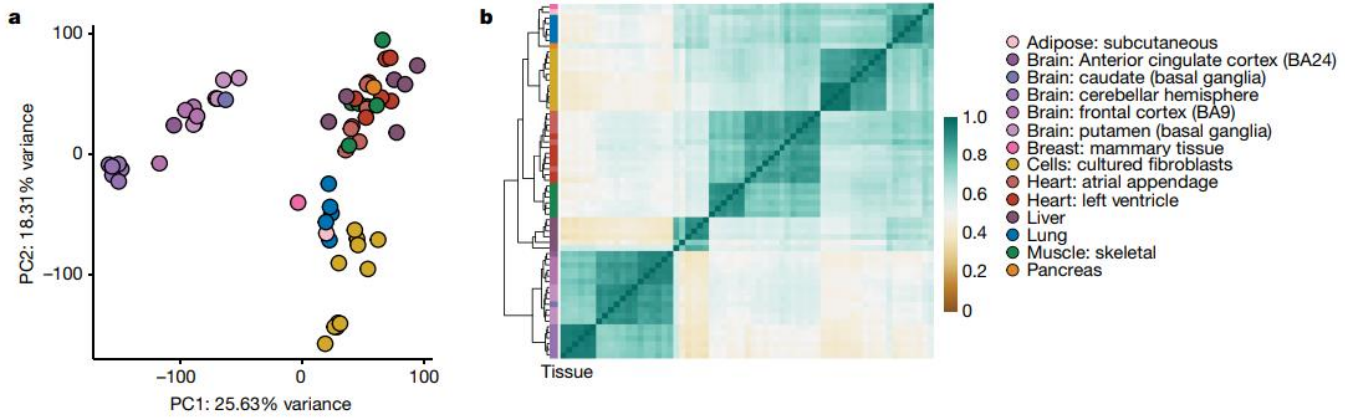
Article | Published: 03 August 2022

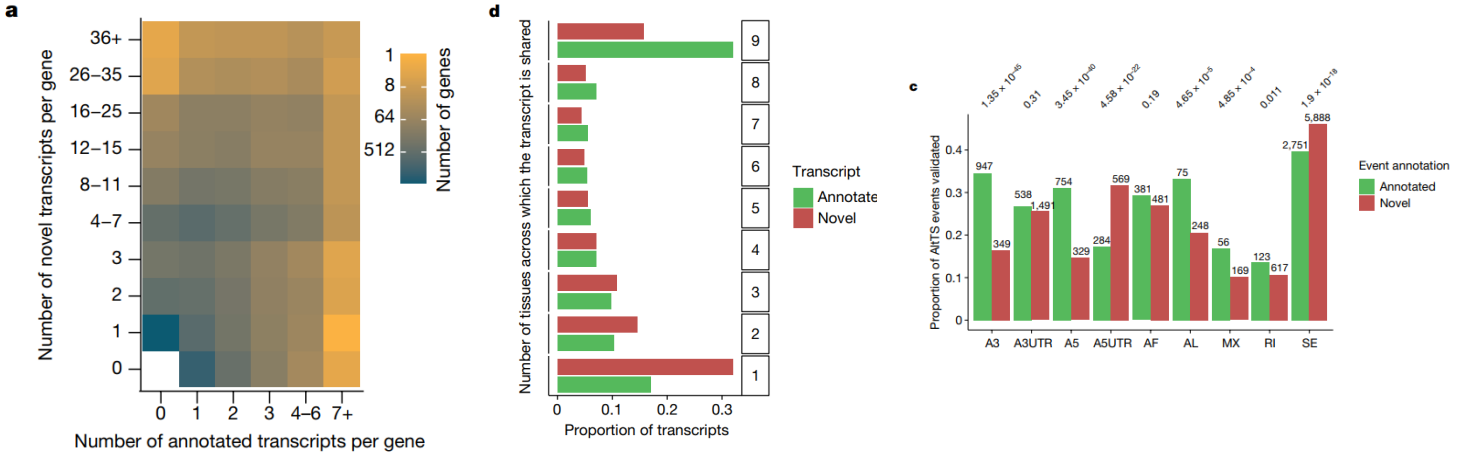
Transcriptome variation in human tissues revealed by long-read sequencing

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

- 90 samples from 56 donors across 16 tissues and 4 K562 cell line samples, coming from GTEx.
- ONT sequencing





VIB-UANTWERP CENTER FOR MOLECULAR NEUROLOGY

University of Antwerp

SCIENCE MEETS LIFE

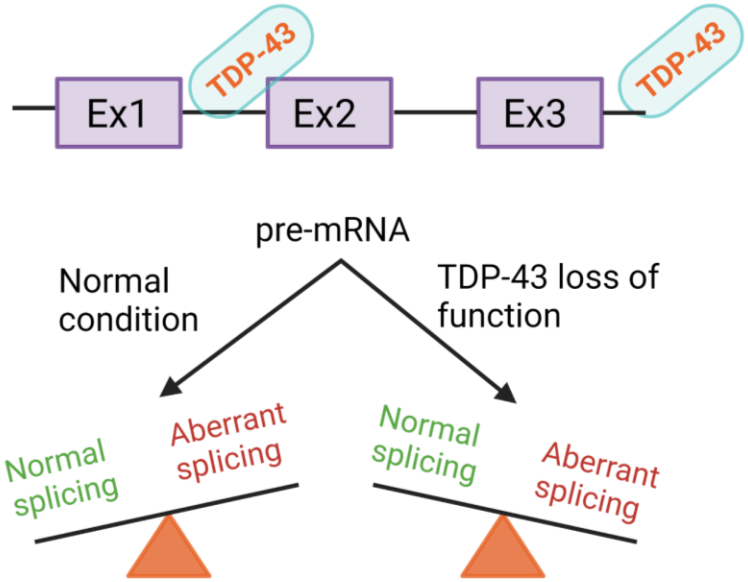
INTEGRATION OF SHORT- AND LONG-READ TRANSCRIPTOMICS TO IDENTIFY NOVEL SPLICING EVENTS DRIVEN BY TDP-43 LOSS OF FUNCTION

Frontotemporal lobar degeneration (FTLD)

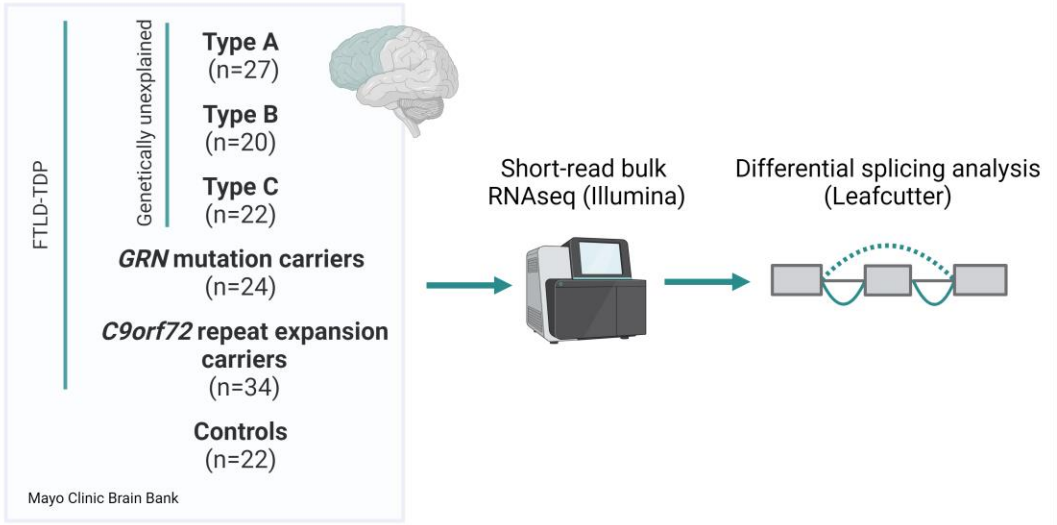
TDP-43

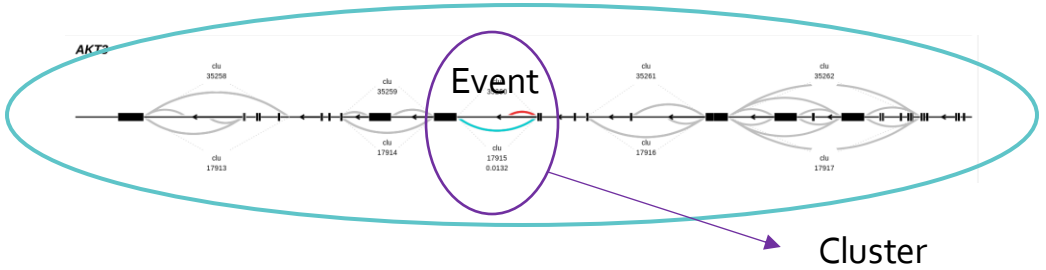
Tau

FUS



Discovery of novel TDP-43-driven splicing events using short-read RNAseq data

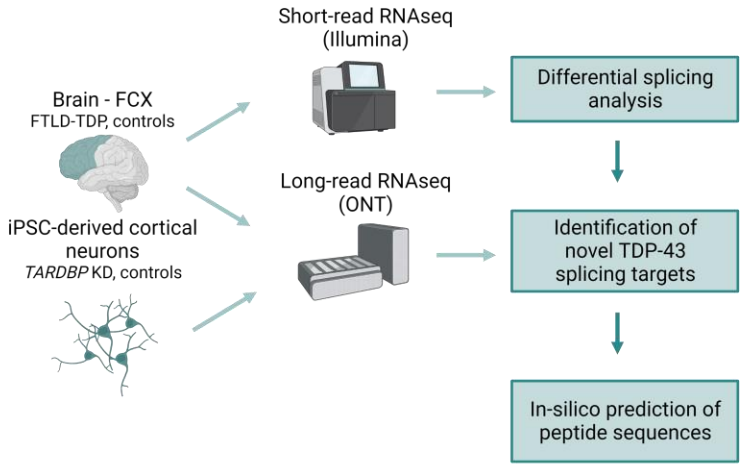




Quantification → marginal percent selected index (PSI) for each event in the cluster
Comparison of experimental conditions → relative changes in PSI (**dPSI**)

- Identification of 1881 differentially spliced events (FDR<0.05, | ΔPSI | > 0.1) in FTLD-TDP patients vs controls

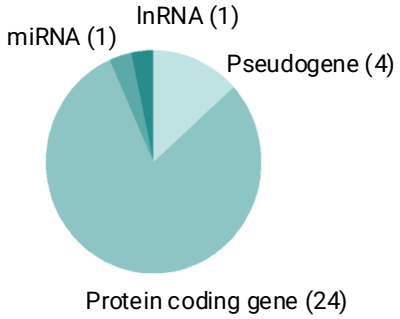
Discovery of novel TDP-43-driven splicing events using short-read RNAseq data



Events with more potential:

- FDR<0.05
- ΔPSI > 0.1 → more present in FTLD-TDP
- Not annotated in the transcriptome

30 cryptic events in 30 unique genes

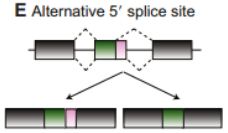


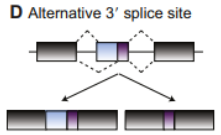
Alternative 5′ splice site: 13

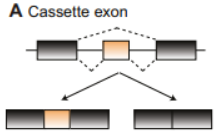
Alternative 3′ splice site: 6

Exon skipping: 7

Exon inclusion: 4

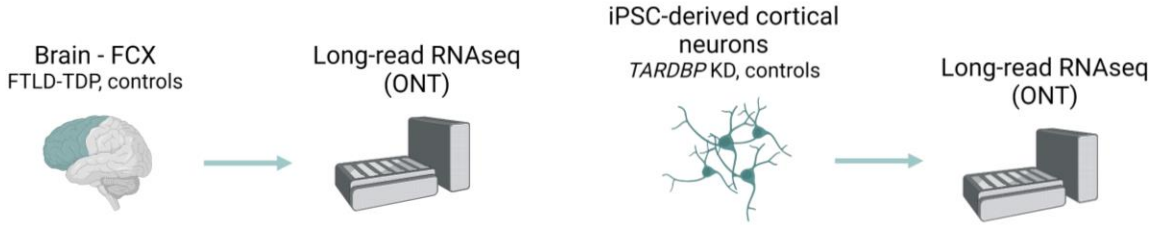






Validation using publicly available data and newly generated long-read direct cDNA data

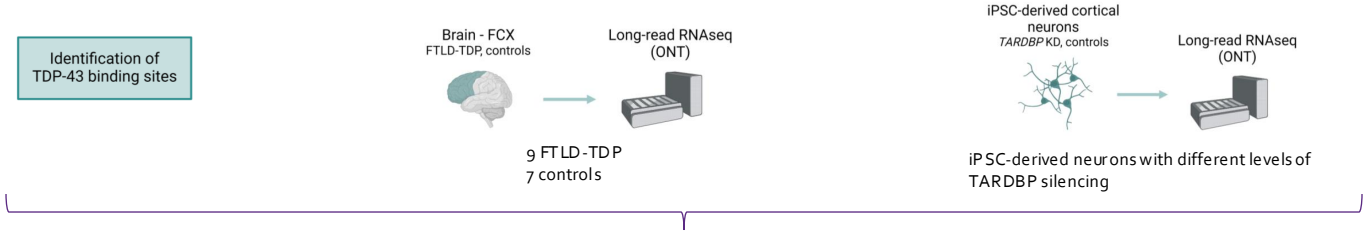
Identification of TDP-43 binding sites



9 FTLD-TDP
7 controls

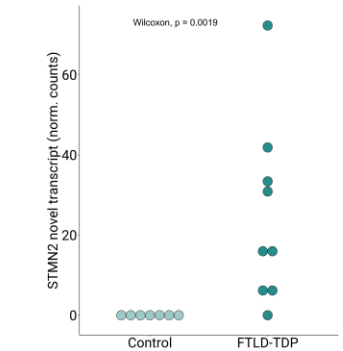
iPSC-derived neurons with different levels of *TARDBP* silencing

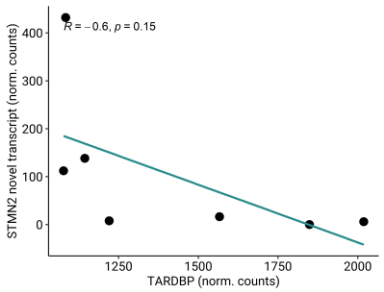
Validation using publicly available data and newly generated long-read direct cDNA data



STMN2

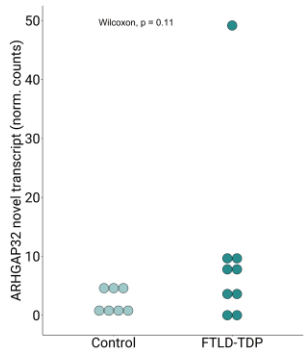
Exon inclusion

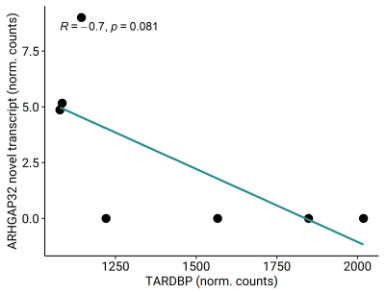




ARHGAP32

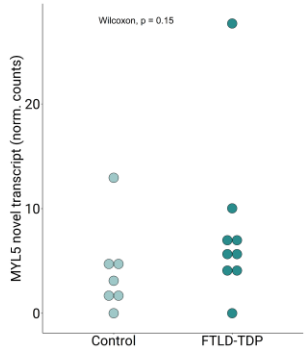
Exon inclusion

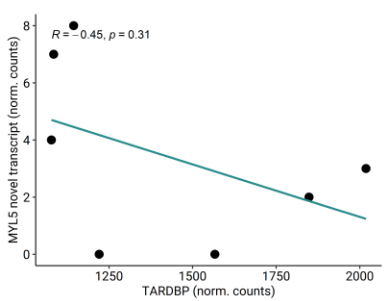




MYL5

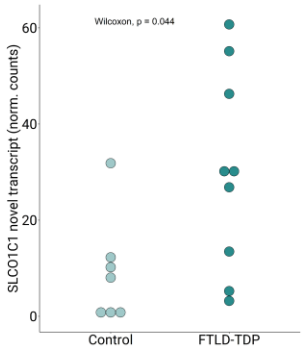
Exon inclusion

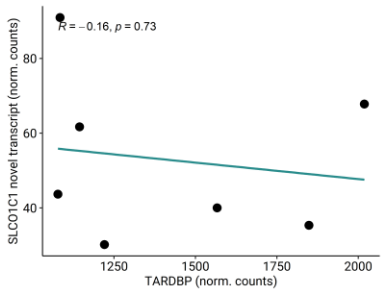




SLCO1C1

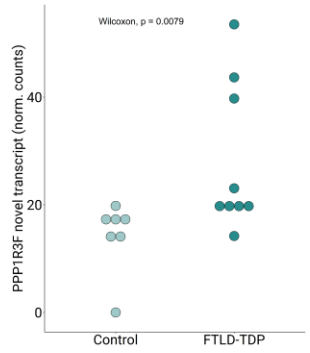
Alternative 5′ splice site

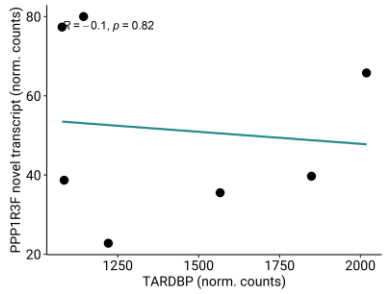




PPP1R3F

Exon skipping





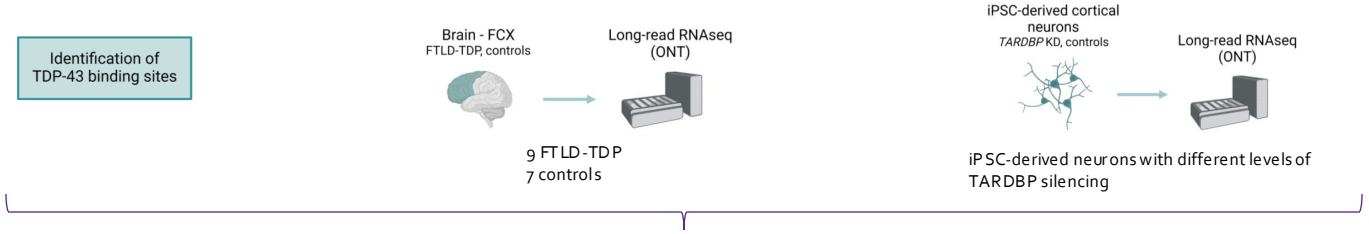
VIB-UANTWERP CENTER FOR MOLECULAR NEUROLOGY

University of Antwerp

Faura et al., in preparation

SCIENCE MEETS LIFE

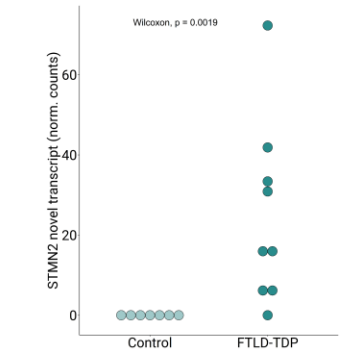
Validation using publicly available data and newly generated long-read direct cDNA data

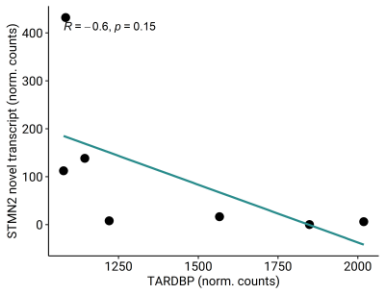


In-silico prediction of peptide sequences

STMN2

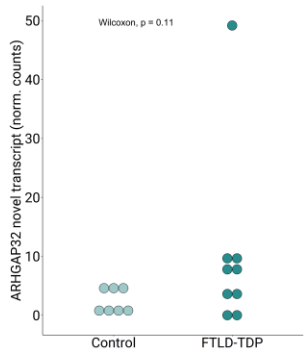
Exon inclusion

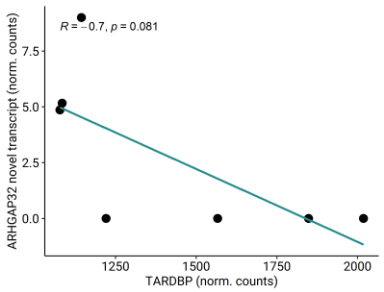




ARHGAP32

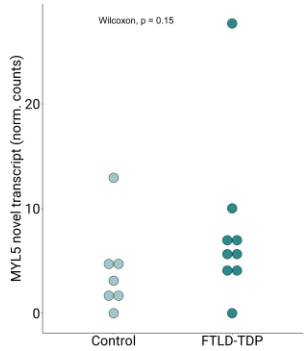
Exon inclusion

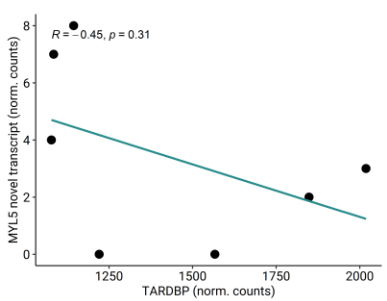




MYL5

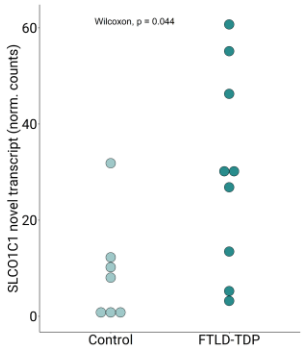
Exon inclusion

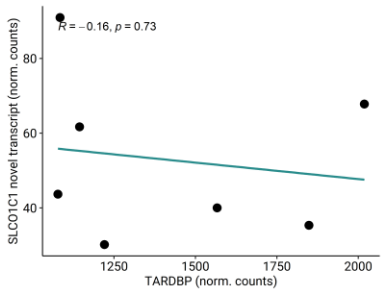




SLCO1C1

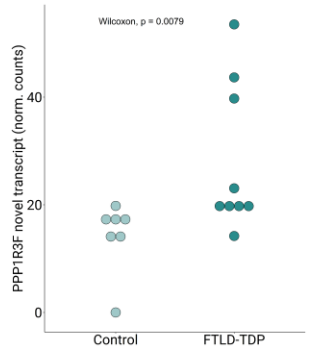
Alternative 5′ splice site

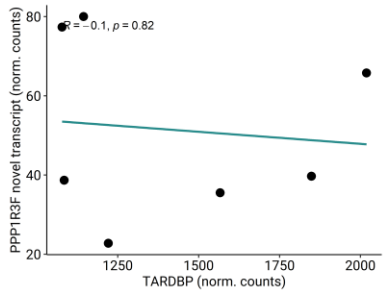




PPP1R3F

Exon skipping





VIB-UANTWERP CENTER FOR MOLECULAR NEUROLOGY

University of Antwerp

Faura et al., in preparation

SCIENCE MEETS LIFE

In-silico peptide prediction of candidate genes

Events supported with long-read data:

Get the exon sequences of each transcript

↓

Combine all the exon sequences to have the cDNA sequence

↓

Check for ORF with ORFFinder

↓

Blast the longest peptide sequence, also get the other possible ORFs

Events NOT supported with long-read data:

Check event in leafviz, UCSC or IGV

↓

Get the MANE selected transcript and select the affected exons

↓

Check each exon coordinate in ENSEMBL

↓

If needed, extract with the new sequences of non-annotated exons

↓

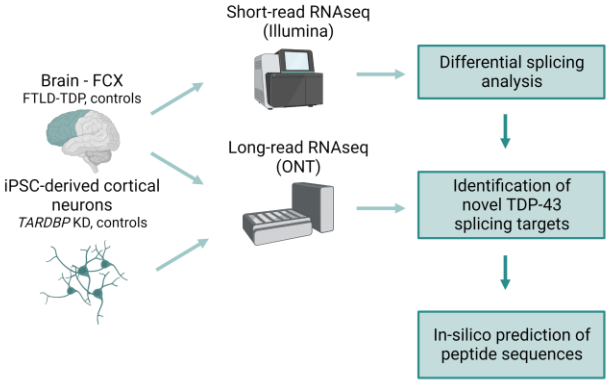
Combine all the exon sequences to have the cDNA sequence

↓

Check for ORF with ORFFinder

↓

Blast the longest peptide sequence

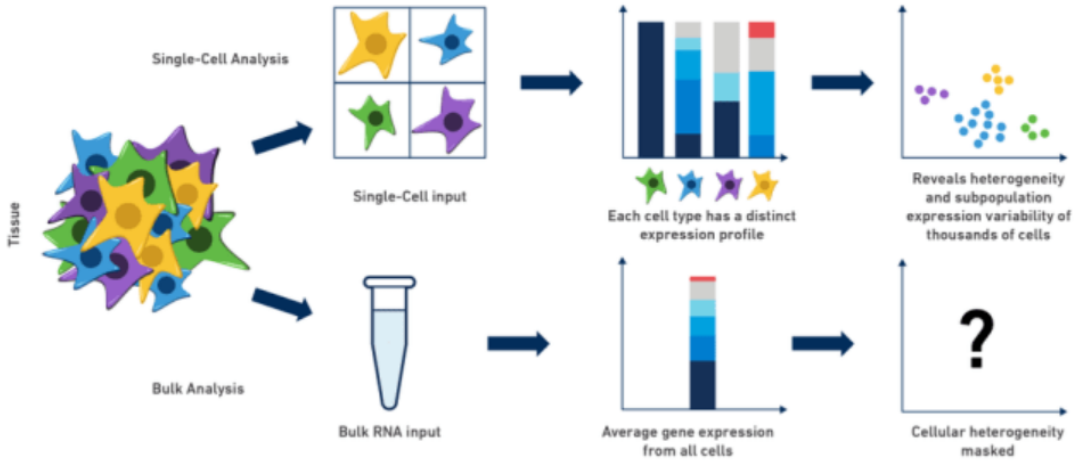


RESULT:

List of in-silico predicted peptide sequences

SINGLE-CELL LONG-READ TRANSCRIPTOMICS – GAINING SINGLE CELL RESOLUTION AT A TRANSCRIPT LEVEL

- A new frontier in the development of third-generation sequencing technologies is the implementation and data analysis of long-read sequencing at the single-cell level.

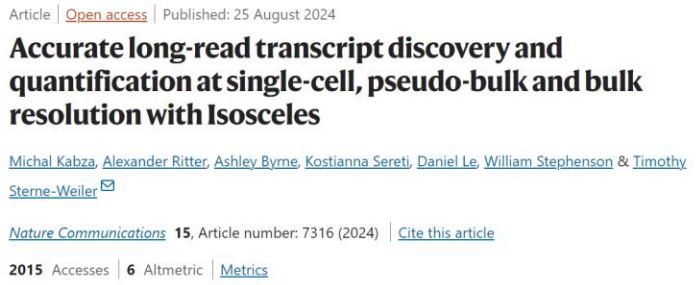


https://www.10xgenomics.com/blog/single-cell-rna-seq-an-introductory-overview-and-tools-for-getting-started

SINGLE-CELL LONG-READ TRANSCRIPTOMICS – GAINING SINGLE CELL RESOLUTION AT A TRANSCRIPT LEVEL

- A new frontier in the development of third-generation sequencing technologies is the implementation and data analysis of long-read sequencing at the single-cell level.

Iscosceles


Article | Open access | Published: 25 August 2024

Accurate long-read transcript discovery and quantification at single-cell, pseudo-bulk and bulk resolution with Isosceles

Michal Kabza, Alexander Ritter, Ashley Byrne, Kostianna Sereti, Daniel Le, William Stephenson & Timothy Sterne-Weiler

Nature Communications **15**, Article number: 7316 (2024) | Cite this article

2015 Accesses | **6** Altmetric | Metrics

FLAMES


Method | Open access | Published: 11 November 2021

Comprehensive characterization of single-cell full-length isoforms in human and mouse with long-read sequencing

Luyi Tian, Jafar S. Jabbari, Rachel Thijssen, Quentin Gouil, Shanika L. Amarasinghe, Oliver Voogd, Hasaru Kariyawasam, Mei R. M. Du, Jakob Schuster, Changqing Wang, Shian Su, Xueyi Dong, Charity W. Law, Alexis Lucattini, Yair David Joseph Prawer, Coralina Collar-Fernández, Jin D. Chung, Timur Naim, Audrey Chan, Chi Hai Ly, Gordon S. Lynch, James G. Ryall, Casey J. A. Anttila, Hongke Peng, ... Matthew E. Ritchie + Show authors

Genome Biology **22**, Article number: 310 (2021) | Cite this article

27k Accesses | **70** Citations | **49** Altmetric | Metrics

BLAZE


Software | Open access | Published: 06 April 2023

Identification of cell barcodes from long-read single-cell RNA-seq with BLAZE

Yupei You, Yair D. J. Prawer, Ricardo De Paoli-Iseppi, Cameron P. J. Hunt, Clare L. Parish, Heejung Shim & Michael B. Clark

Genome Biology **24**, Article number: 66 (2023) | Cite this article

9139 Accesses | **14** Citations | **30** Altmetric | Metrics

- Analysis tools have an extra challenge: cell barcode and UMI extraction and correction.
  - Examples: BLAZE-FLAMES, wf-single-cell (epi2me labs), Isoceles, Scywalker

Scywalker: scalable end-to-end data analysis workflow for long-read single-cell transcriptome sequencing

- Existing nanopore single-cell data analysis tools showed severe limitations in handling current data sizes.
- Isoform discovery in scywalker is based on IsoQuant, with the initial discovery performed without taking cell barcodes into account, i.e. on bulk cDNA.

Gene expression

Scywalker: scalable end-to-end data analysis workflow for long-read single-cell transcriptome sequencing

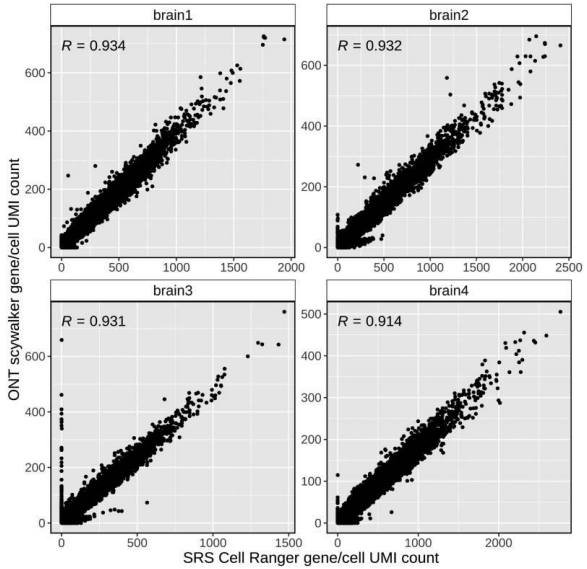
Peter De Rijk[1,2], Tijs Watzeels[2,3], Fahri Küçükali[2,3], Jasper Van Dongen[2,3], Júlia Faura[2,4], Patrick Willems[5,6,7,8], Lara De Deyn[2,3], Lena Duchateau[2,3], Carolin Grones[5,6], Thomas Eekhout[5,6,9], Tim De Pooter[1,2], Geert Joris[1,2], Stephane Rombauts[5,6], Bert De Rybel[5,6], Rosa Rademakers[2,4], Frank Van Breusegem[5,6], Mojca Strazisar[1,2], Kristel Sleegers[2,3], Wouter De Coster[2,4,*]

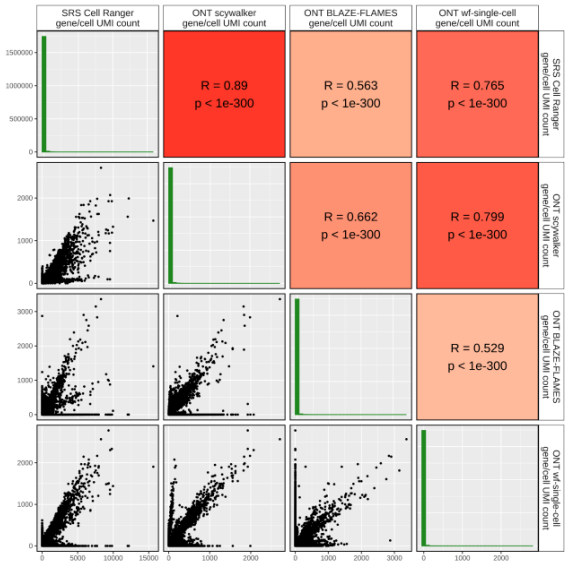
[1] Neuromics Support Facility, VIB Center for Molecular Neurology, VIB, Antwerp, Belgium; [2] Department of Biomedical Sciences, University of Antwerp, Antwerp, Belgium, [3] Complex Genetics of Alzheimer's Disease Group, VIB Center for Molecular Neurology, VIB, Antwerp, Belgium, [4] Applied and Translational Neurogenomics Group, VIB Center for Molecular Neurology, VIB, Antwerp, Belgium, [5] Department of Plant Biotechnology and Bioinformatics, Ghent University, Ghent, Belgium, [6] VIB Center for Plant Systems Biology, VIB, Ghent, Belgium, [7] Department of Biomolecular Medicine, Ghent University, Ghent, Belgium, [8] VIB Center for Medical Biotechnology, VIB, Ghent, Belgium, [9] VIB Single Cell Core, VIB, Ghent/Leuven, Belgium.

Bioinformatics (2024), btae549

https://github.com/derijkp/scywalker

SCAN ME





- In a smaller public dataset, BLAZE-FLAMES was the fastest (3h19), scywalker second (3h37) and wf-single-cell the slowest (5h41). But the analysis of 1 brain sample took 14h33 with scywalker, 172h06 with BLAZE-FLAMES.

VIB-UANTWERP CENTER FOR MOLECULAR NEUROLOGY

University of Antwerp

SCIENCE MEETS LIFE

Scywalker: scalable end-to-end data analysis workflow for long-read single-cell transcriptome sequencing

- Scywalker generates one interactive html report per sample, just like you get with CellRanger (10x Genomics) in short-read scRNAseq.

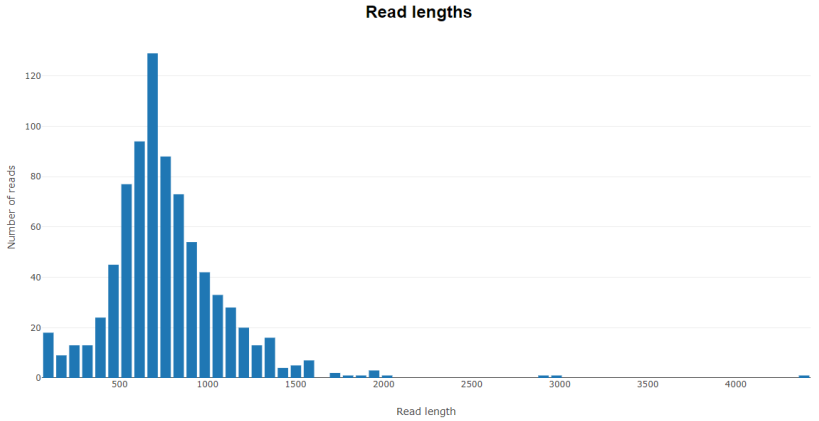
Scywalker QC report

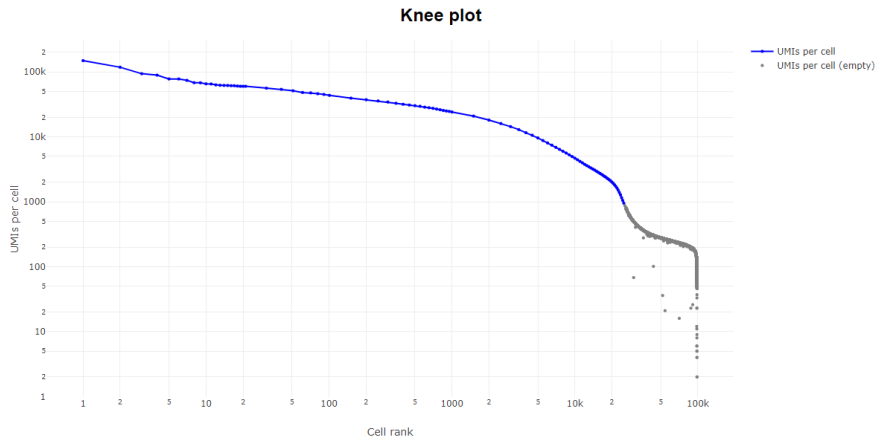
Alignment summary

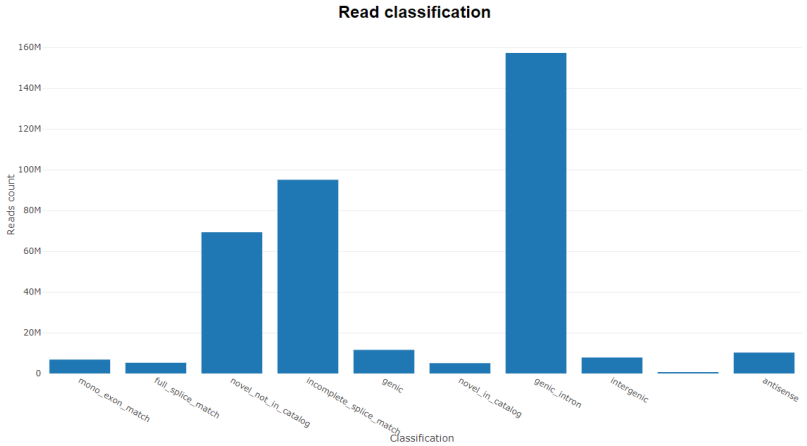
File name	map-sminimap2_splice-merged_rr_CVI007_200134m139690d89_hg38s.cram
Number of alignments	247369462
% from total reads	81.43
Yield [Gb]	134.80
N75	459
Median length	495.00
Mean length	544
Median identity	98.84
Mean identity	97.67

Cells and genes metrics

Total cells	98587
Cells passing filter	26021
% umis in cells passing filter	89.03%
Median UMI count	3457
Median genes per cell	1733







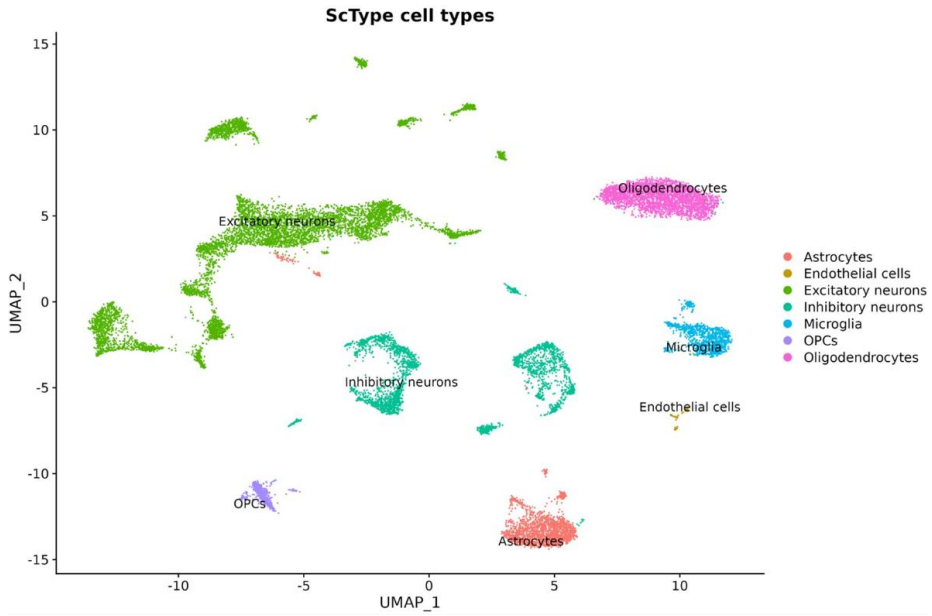
VIB-UANTWERP CENTER FOR MOLECULAR NEUROLOGY

University of Antwerp

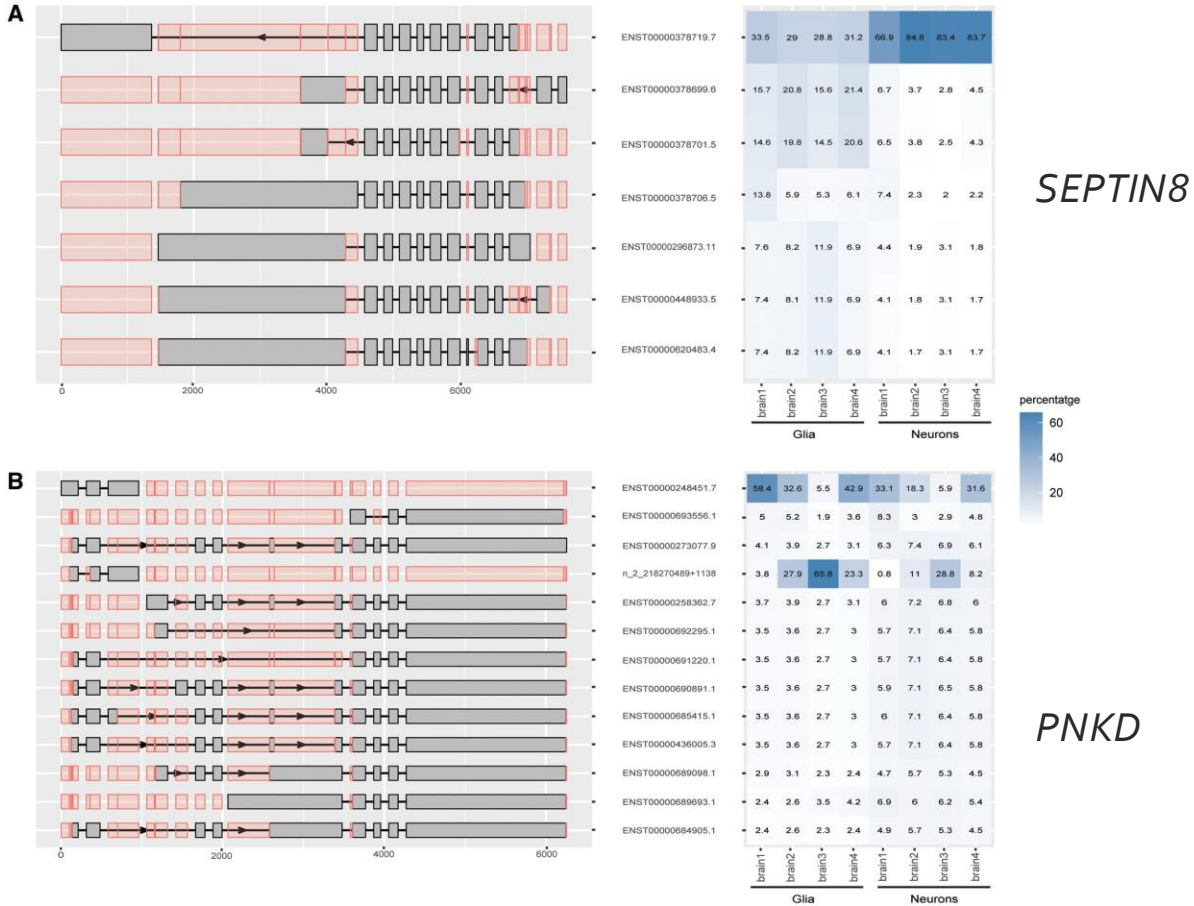
SCIENCE MEETS LIFE

Scywalker: scalable end-to-end data analysis workflow for long-read single-cell transcriptome sequencing

- Scywalker is able to successfully assign different cell types and provides pseudobulk data (per cell type), creating a multi-sample pseudobulk count matrix.



- Also tested in *Arabidopsis thaliana* and publicly available PacBio data



SEPTIN8

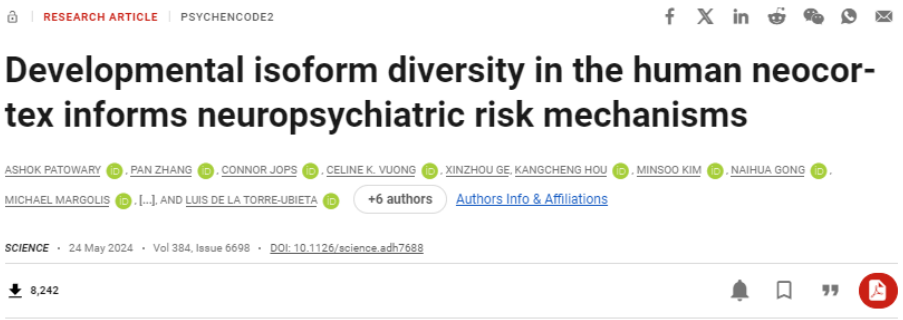
PNKD

VIB-UANTWERP CENTER FOR MOLECULAR NEUROLOGY

University of Antwerp

SCIENCE MEETS LIFE

Illustrative example: Developmental isoform diversity in the human neocortex informs neuropsychiatric risk mechanisms (Patowary et al, *Science*, 2024)

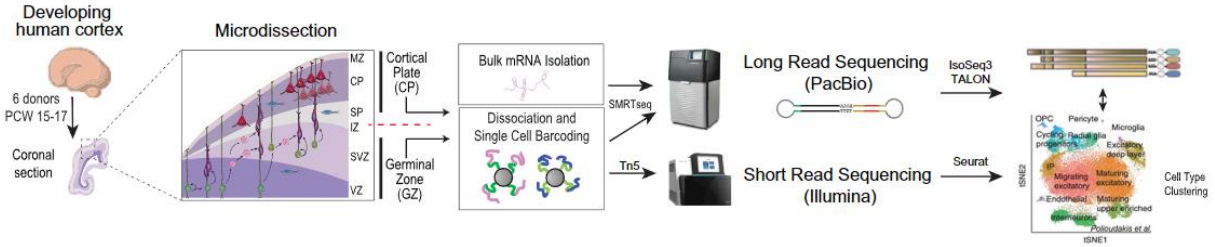

RESEARCH ARTICLE | PSYCHENCODE2

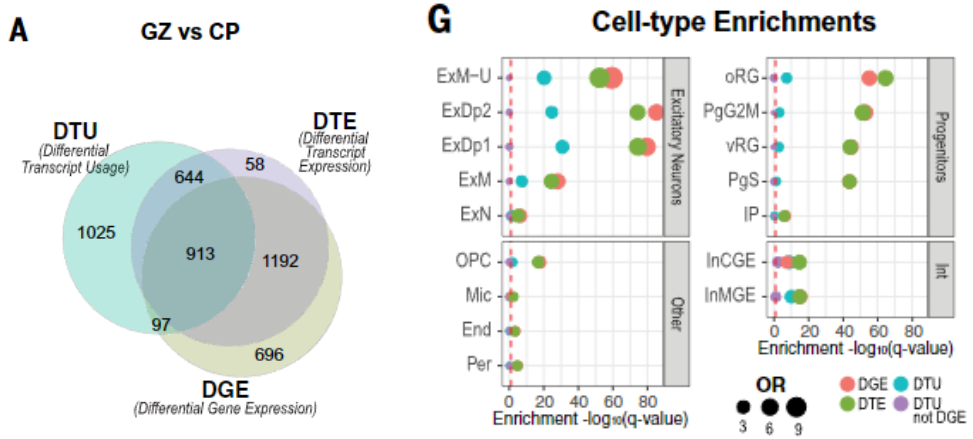
Developmental isoform diversity in the human neocortex informs neuropsychiatric risk mechanisms

ASHOK PATOWARY, PAN ZHANG, CONNOR JOPS, CELINE K. VUONG, XINZHOU GE, KANGCHENG HOU, MINSOO KIM, NAIHUA GONG, MICHAEL MARGOLIS, [...], AND LUIS DE LA TORRE-UBIETA +6 authors Authors Info & Affiliations

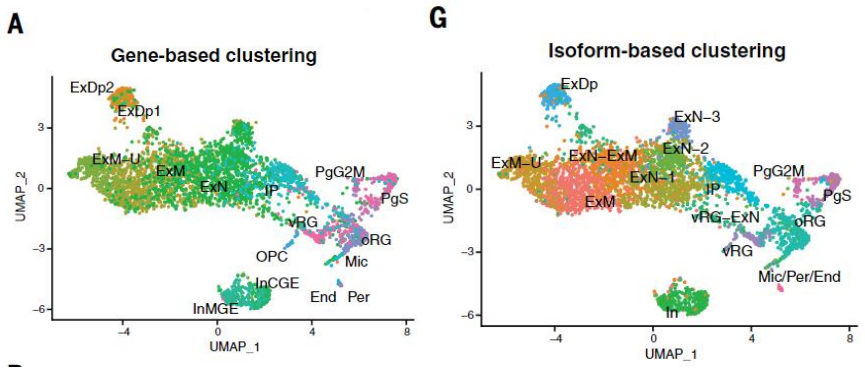
SCIENCE · 24 May 2024 · Vol 384, Issue 6698 · DOI: 10.1126/science.adh7688

8,242





- Single-nuclei LR-RNAseq allows the differential expression usage analyses at a cell type level



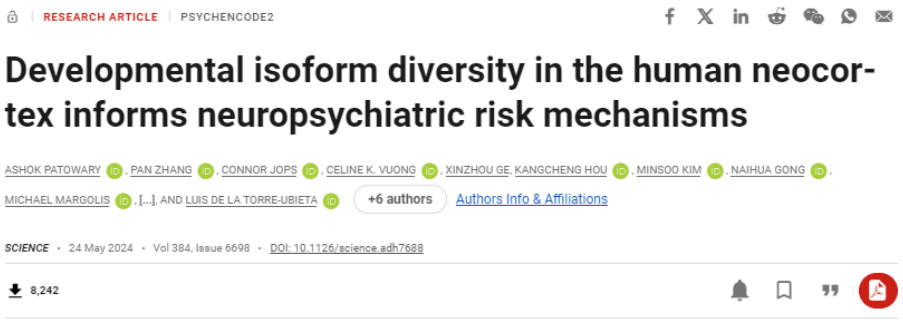
- Additional stages of excitatory neuron maturation can be defined using isoform-level data.

VIB-UANTWERP CENTER FOR MOLECULAR NEUROLOGY

University of Antwerp

SCIENCE MEETS LIFE

Illustrative example: Developmental isoform diversity in the human neocortex informs neuropsychiatric risk mechanisms (Patowary et al, *Science*, 2024)

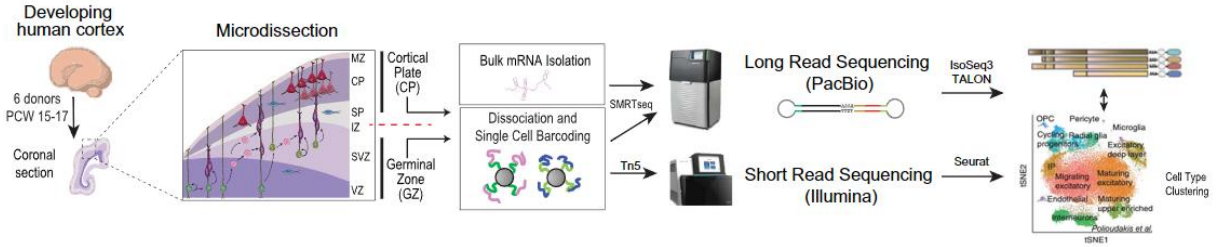

RESEARCH ARTICLE | PSYCHENCODE2

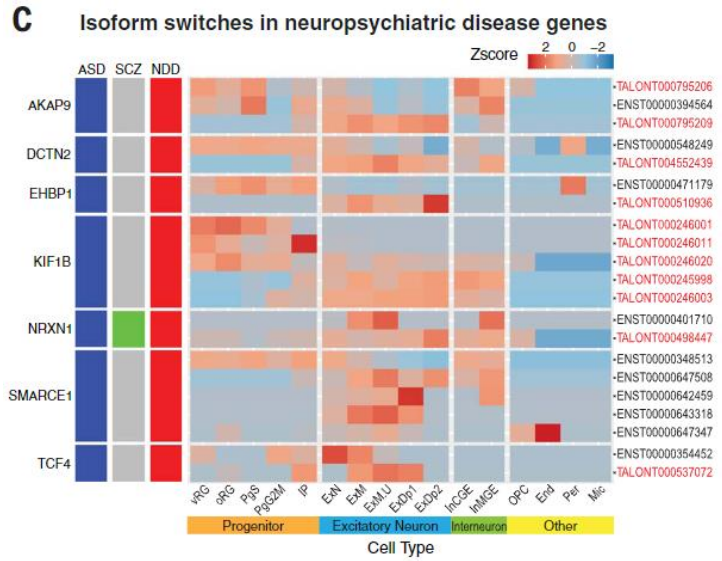
Developmental isoform diversity in the human neocortex informs neuropsychiatric risk mechanisms

ASHOK PATOWARY, PAN ZHANG, CONNOR JOPS, CELINE K. VUONG, XINZHOU GE, KANGCHENG HOU, MINSOO KIM, NAIHUA GONG, MICHAEL MARGOLIS, [...], AND LUIS DE LA TORRE-UBIETA +6 authors Authors Info & Affiliations

SCIENCE · 24 May 2024 · Vol 384, Issue 6698 · DOI: 10.1126/science.adh7688

8,242





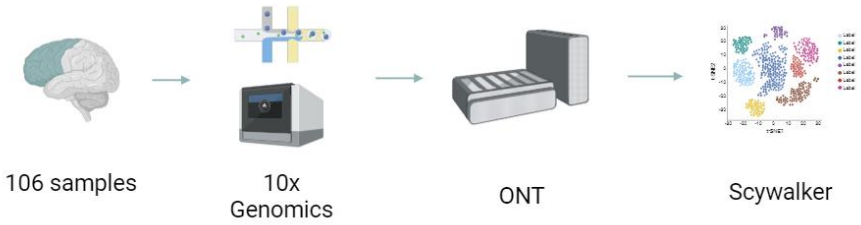
- Identification of isoform switches in several NDD and ASD risk genes across the cell types ofthe developing cortex

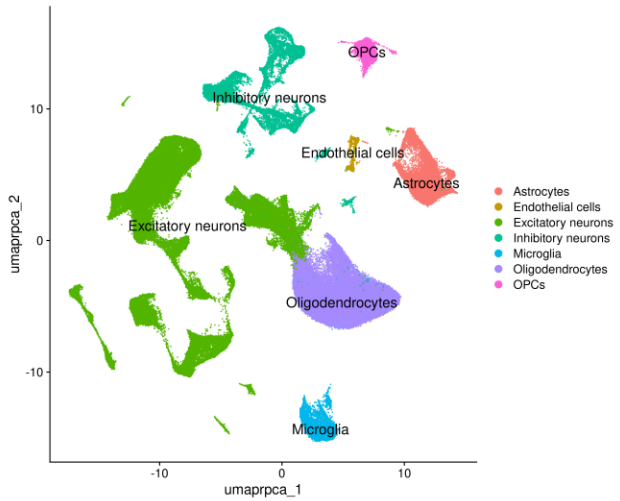
VIB-UANTWERP CENTER FOR MOLECULAR NEUROLOGY

University of Antwerp

SCIENCE MEETS LIFE

GENERATION OF LONG-READ SINGLE NUCLEI TRANSCRIPTOMICS DATA IN THE CONTEXT OF FTLD

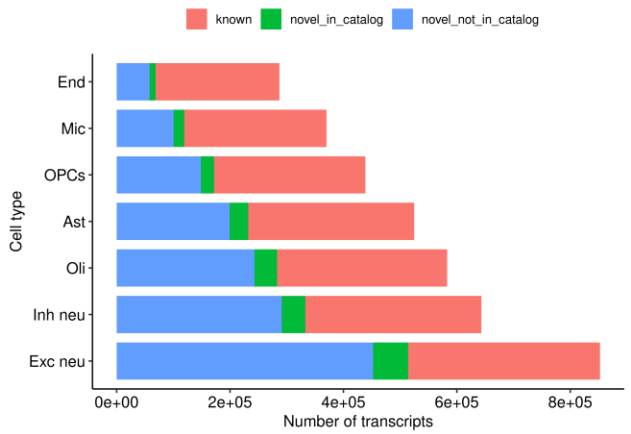


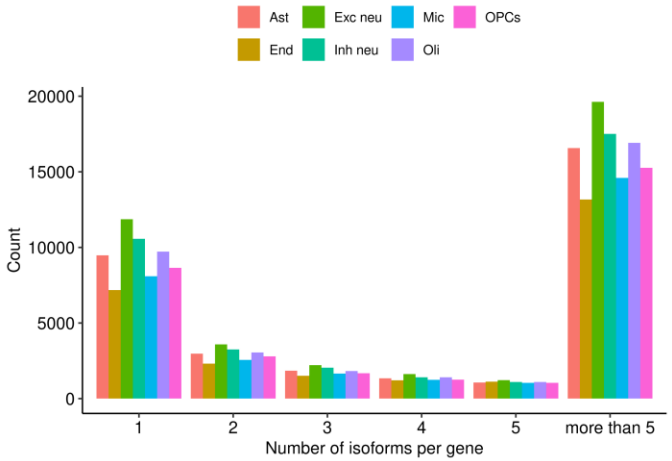


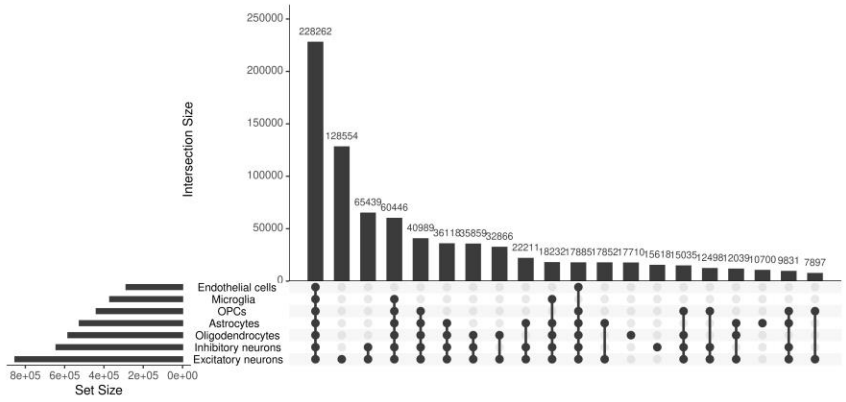
20 neurologically healthy controls

- Annotation at the major cell type level

- Characterization of the isoform diversity







VIB-UANTWERP CENTER FOR MOLECULAR NEUROLOGY

University of Antwerp

SCIENCE MEETS LIFE

TAKE-HOME MESSAGES

- Gene isoform and alternative splicing analysis can provide biological insights that gene expression alone cannot reveal.
- Long-read transcriptomics is the optimal tool for analyzing gene isoforms and splicing events.
- A variety of methods, technologies, and analysis tools are available, depending on your specific research questions, needs, and preferences.

Rosa Rademakers lab
Rosa Rademakers
Bavo Heeman
Sarah Wynants
Marleen Van den Broeck
Wouter de Coster
Cristina Vicente
Rafaela Policarpo
Marijne Vandebergh
Sara Alidadiani
Jolien Perneel
Miranda Lastra Osua
Elise Coopman
Vanshika Bidhan
Linus De Witte
Nele Peeters
Lorraine Murphy
Lars Mohren
Laura Heiß

Neuromics Support Facility
Mojca Strazisar
Tim De Pooter
Geert Joris
Peter De Rijk

VIB Single Cell Core
Niels Vandamme
Ria Roelandt
Melanie Verspeeten
Robin Boiy

VIB-UANTWERP CENTER FOR MOLECULAR NEUROLOGY

University of Antwerp

Contact
julia.faurallorens@uantwerpen.be

SCIENCE MEETS LIFE

Unlocking the Full Transcriptome: Advances in Long-read Transcriptomics

Current Trends in Long Read Sequencing and Bioinformatics Analysis
Leuven, 03/10/2024

Júlia Faura, PhD – VIB-UAntwerp Center for Molecular Neurology

VIB-UANTWERP CENTER FOR MOLECULAR NEUROLOGY

University of Antwerp

SCIENCE MEETS LIFE