Long read single-cell and spatial transcriptomics

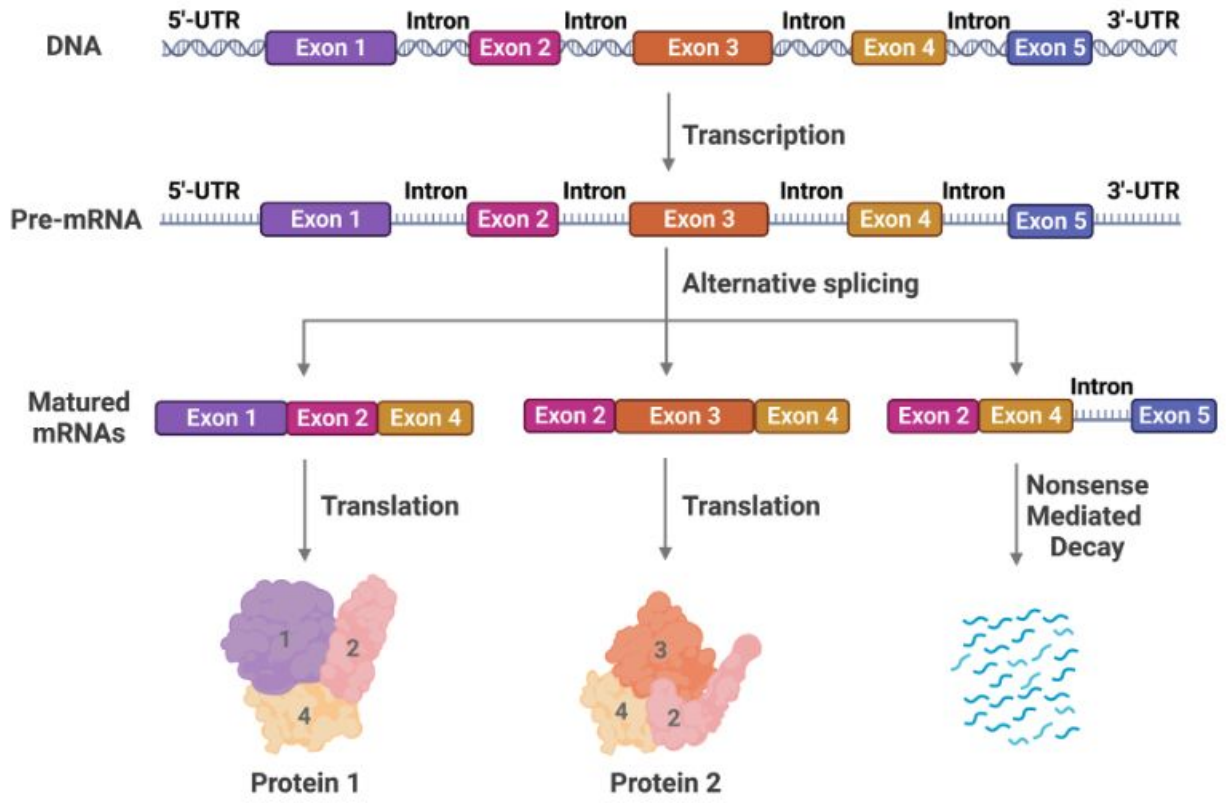
Ali Hamraoui & Morgane Fierville

Thrusday 24 October 2025

Alternative splicing: the main mechanism for generating transcriptome complexity

Alternative splicing plays a critical roles in:

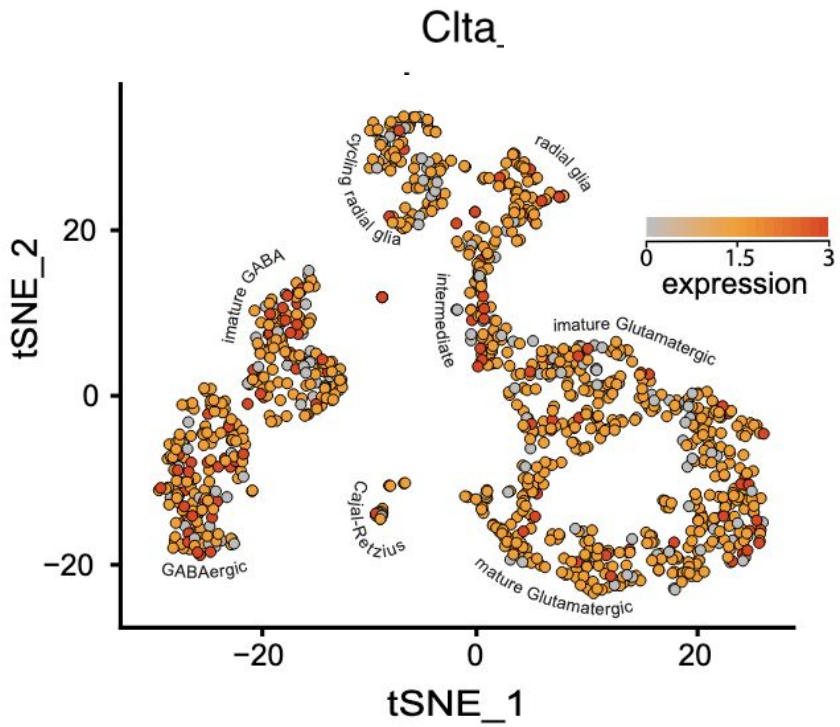
- Cell differentiation
- Speciation
- Human diseases such as cancer, diabetes and neurological disorders.



Wu and Schmitz et al. Computational and Structural Biotechnology (2023)

Single-cell analysis at isoform level is biologically more informative than at gene level

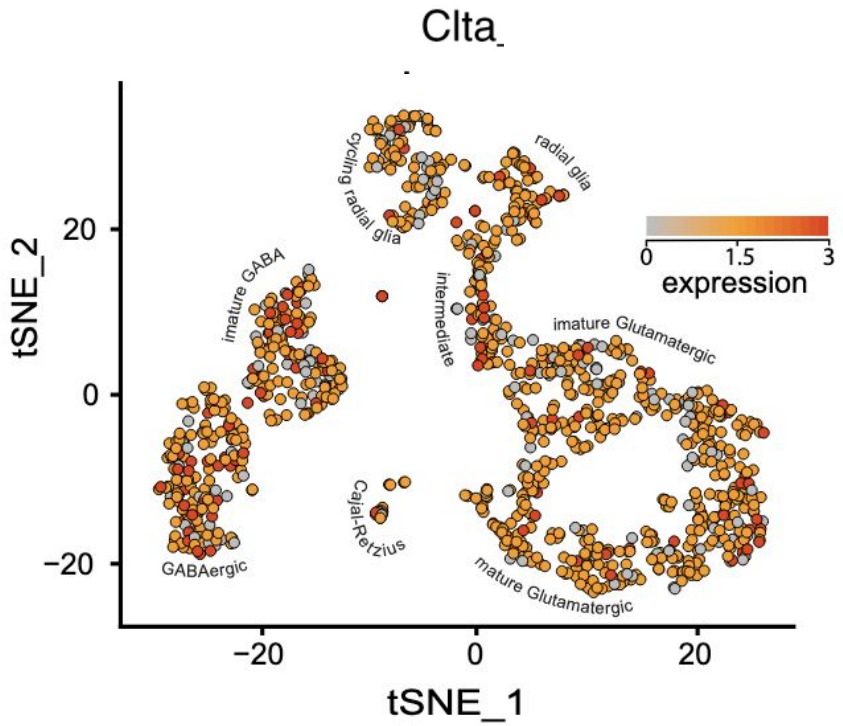
Clta isoform expression switch during **neuronal maturation** visualized on the t-SNE plot.

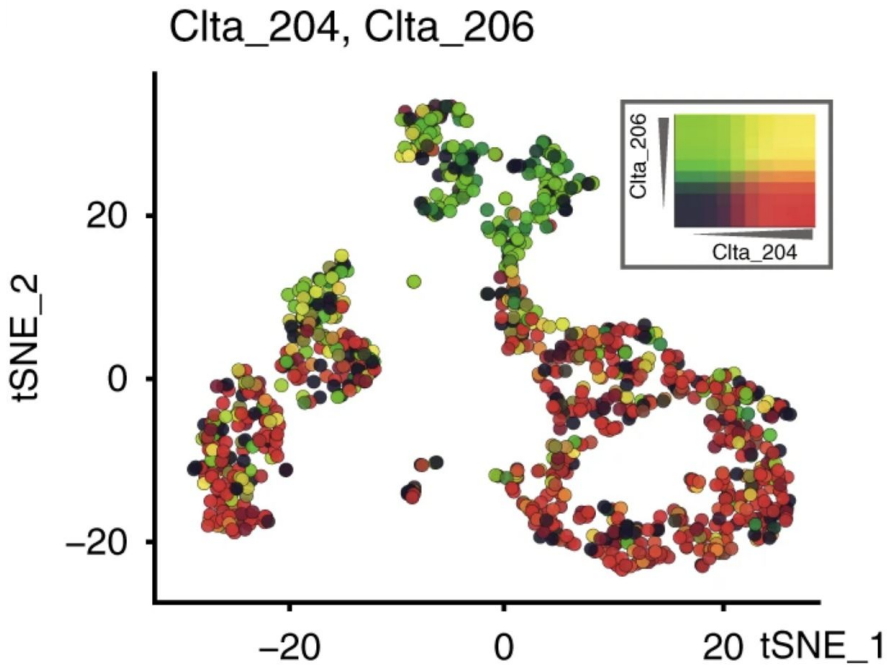


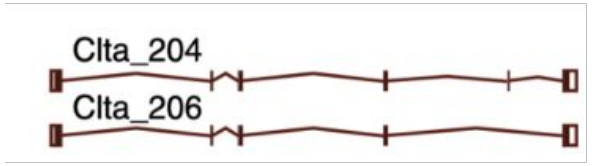
Lebrigand et al., Nature comm (2020)

Single-cell analysis at isoform level is biologically more informative than at gene level

Clta isoform expression switch during **neuronal maturation** visualized on the t-SNE plot.





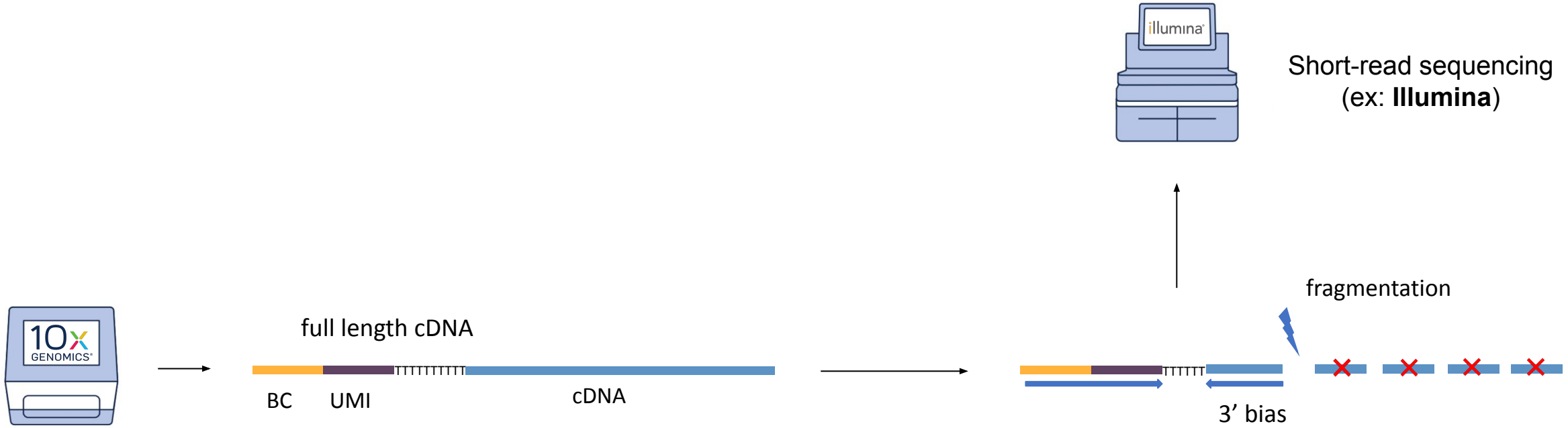


Lebrigand et al., Nature comm (2020)

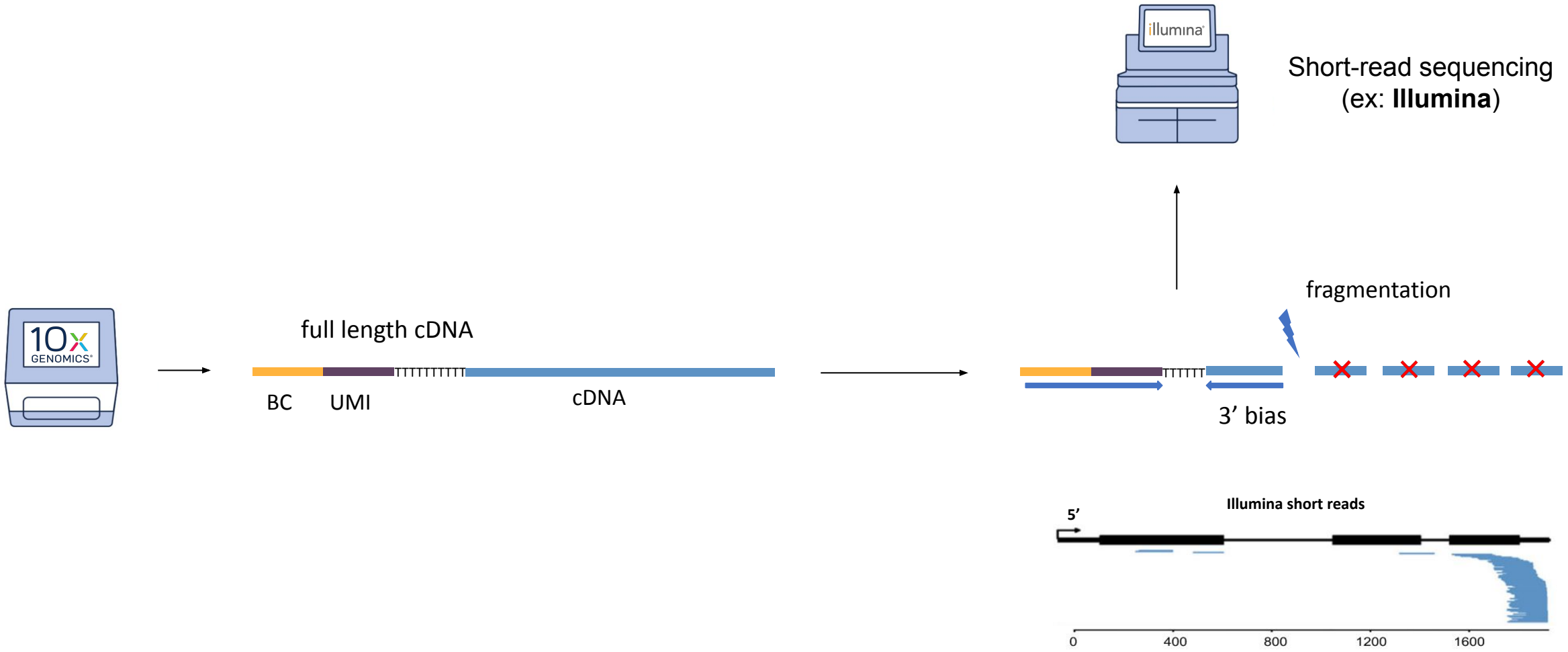
Morning program

- **Why single cell long read is usefull?**
- Preprocessing and comparison of tools.
- Spatial isoform transcriptomics.
- Demo: scRNAseq long read data pre-processing.

Single-cell transcriptomics with 10X Genomics technology

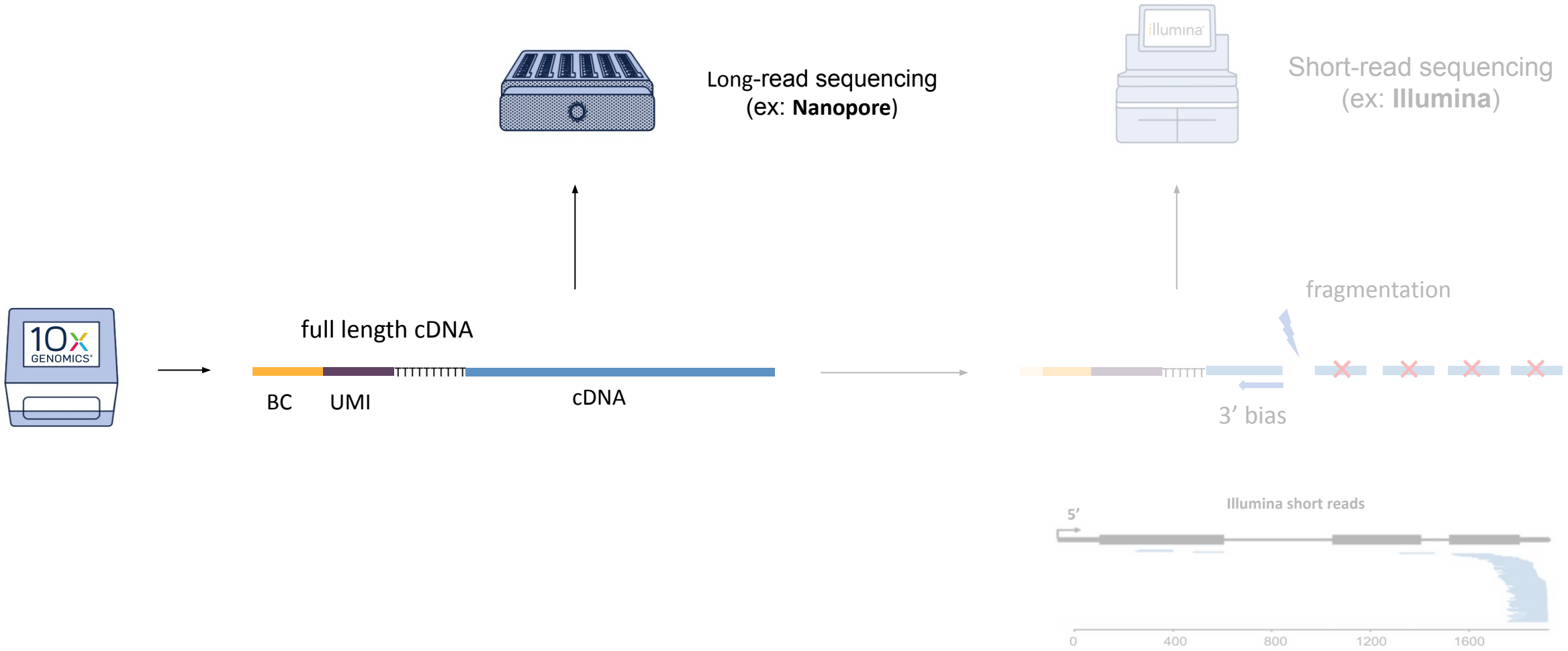


Single-cell transcriptomics with 10X Genomics technology



Long et al. Genome Biology (2021)

Single-cell transcriptomics with 10X Genomics technology



Long et al. Genome Biology (2021)

Single-cell transcriptomics with 10X Genomics technology

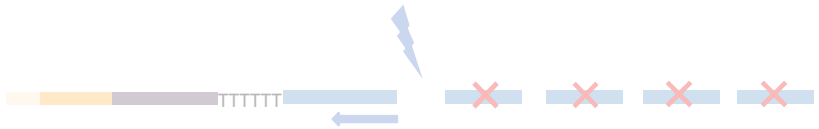
Long-read sequencing
(ex: **Nanopore**)

Short-read sequencing
(ex: **Illumina**)

full length cDNA

BC UMI cDNA

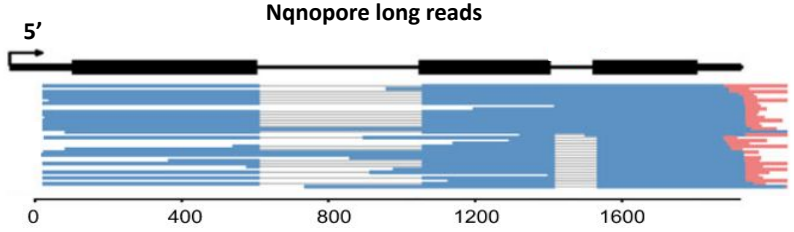
fragmentation



3’ bias

Information on splicing, SNP, editing are preserved

Nqnopore long reads



Illumina short reads

Long et al. Genome Biology (2021)

Single-cell transcriptomics with 10X Genomics technology

Long-read sequencing
(ex: **Nanopore**)

Short-read sequencing
(ex: **Illumina**)

full length cDNA

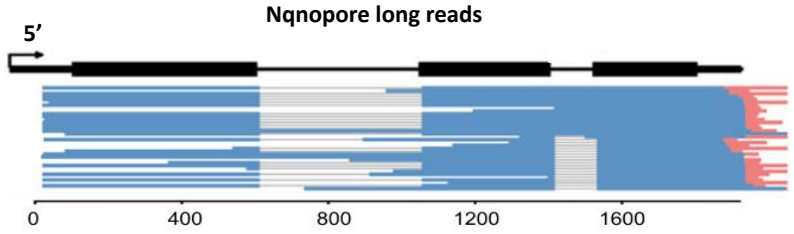
BC UMI cDNA

fragmentation

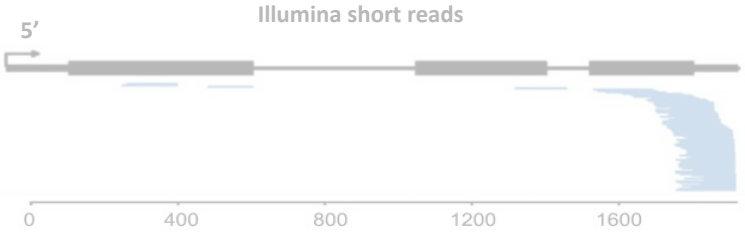
3’ bias

Information on splicing, SNP, editing are preserved

Nqnopore long reads



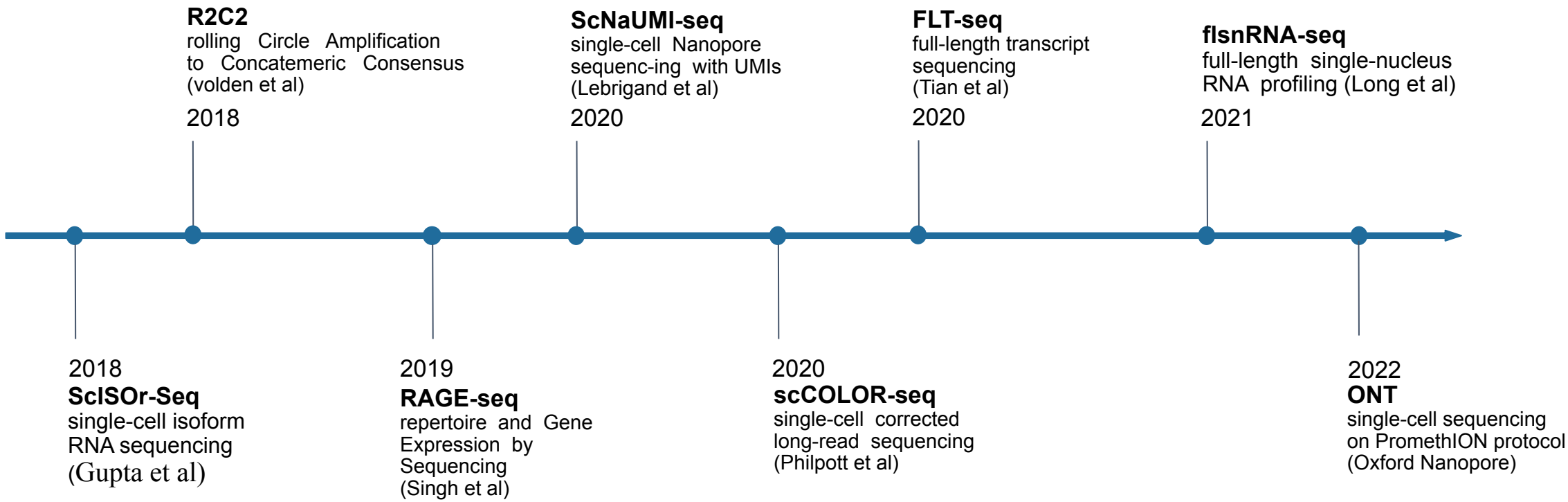
Illumina short reads



~95% accuracy on R9.4.1 flowCell

Long et al. Genome Biology (2021)

State of the art of the single cell long read with 10x Genomics

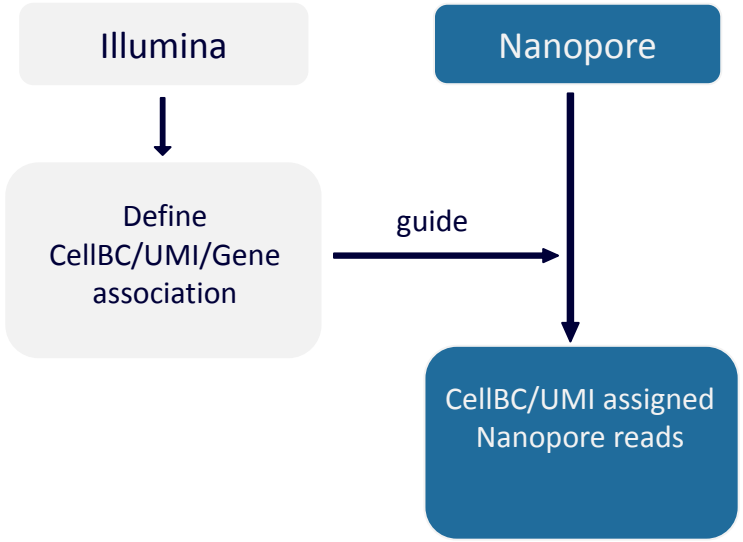


Bioinformatics preprocessing approaches for single cell long read

Several bioinformatics pre-processing tools have been developed to analyse nanopore-sequenced 10x transcriptome libraries.

Year	Authors	Journal	Protocol	Approach	Tool
2020	Lebrigand et al.	Nature comm.	scNaUMI-seq	Hybrid	**SiCeLoRe**
2021	Long et al.	Genome Biology	FIsnRNA-seq	Hybrid	**Snuupy**
	Wang et al.	RNA		Hybrid	**scNapBar**
	Tian et al.	Genome Biology	FLT-seq	Nanopore-only	**FLAMES**
2022	ONT		ONT	Nanopore-only	**Sockeye**
	Lebrigand et al.			Nanopore-only	**SiCeLoRe 2.1**
2023	Shiau et al.	Nature comm.		Nanopore-only	**ScNanoGPS**

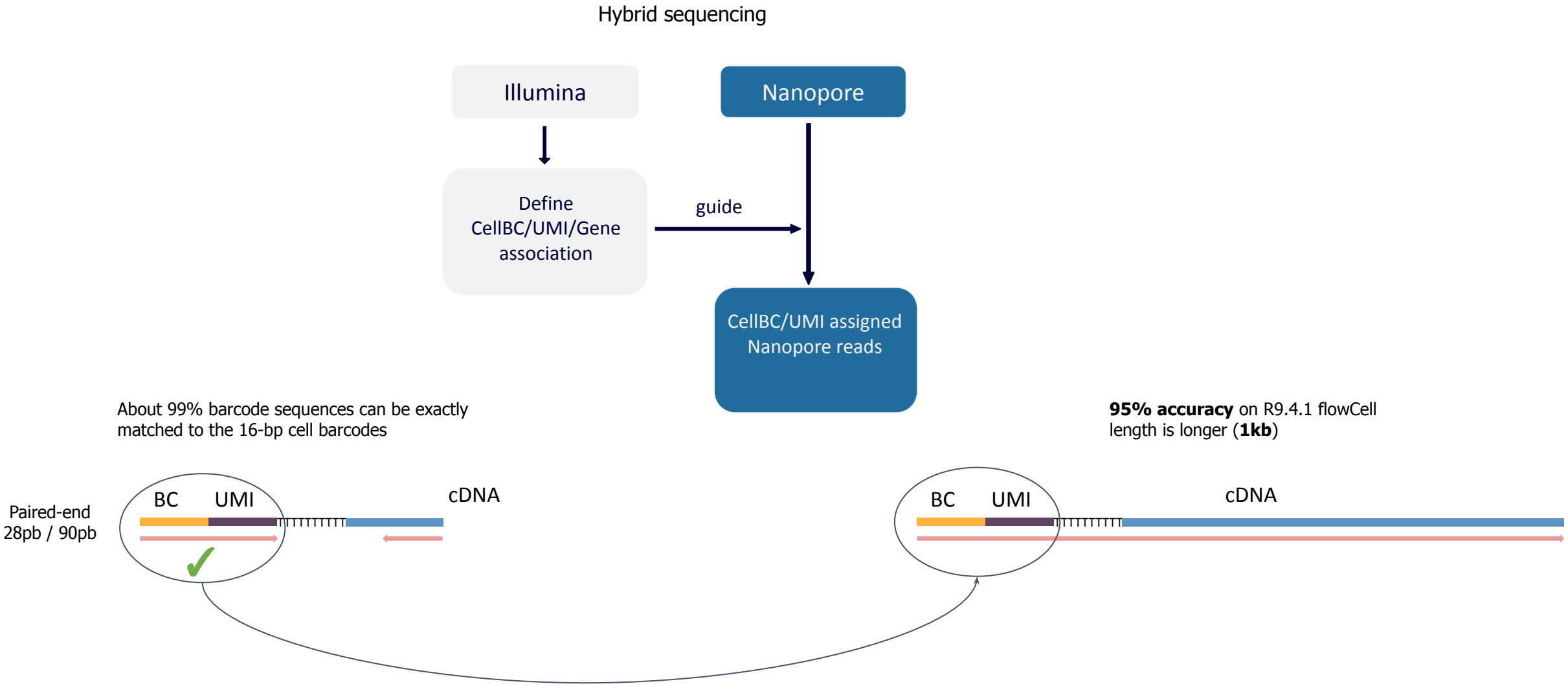
Hybrid approaches



ScNaUMI-seq (Single Cell Long Read) - Lebrigand et al. 2020

Bioinformatics preprocessing approaches for single cell long read

Several bioinformatics pre-processing tools have been developed to analyse nanopore-sequenced 10x transcriptome libraries.



Bioinformatics preprocessing approaches for single cell long read

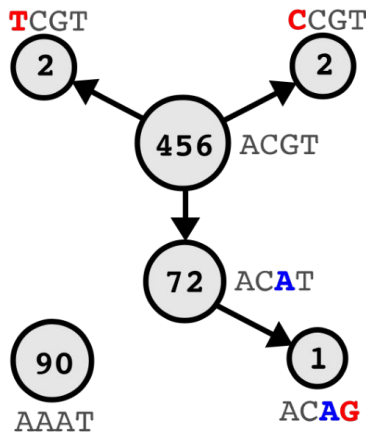
Several bioinformatics pre-processing tools have been developed to analyse nanopore-sequenced 10x transcriptome libraries.

Year	Authors	Journal	Protocol	Approach	Tool
2020	Lebrigand et al.	Nature comm.	scNaUMI-seq	Hybrid	**SiCeLoRe**
2021	Long et al.	Genome Biology	FlsnRNA-seq	Hybrid	**Snuupy**
	Wang et al.	RNA		Hybrid	**scNapBar**
	Tian et al.	Genome Biology	FLT-seq	Nanopore-only	**FLAMES**
2022	ONT		ONT	Nanopore-only	**Sockeye**
	Lebrigand et al.			Nanopore-only	**SiCeLoRe 2.1**
2023	Shiau et al.	Nature comm.		Nanopore-only	**ScNanoGPS**

Nanopore-only approaches

Correction of BC & UMI similar to **umi tools**

Tom Sean Smith et al, Genome Research (2017)



scRNA-seq Nanopore data preprocessing involves multiple steps

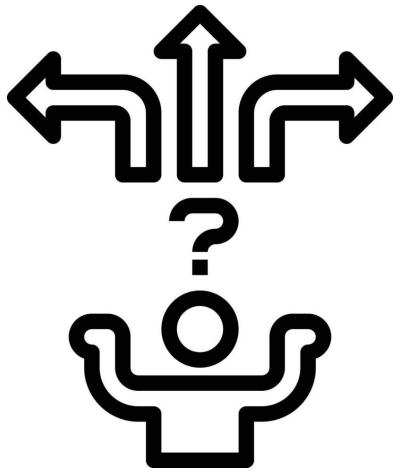
Comparative overview of scRNA-seq Nanopore preprocessing workflows : Diverse approaches for each step

scRNA-seq Nanopore data preprocessing involves multiple steps

Comparative overview of scRNA-seq Nanopore preprocessing workflows : Diverse approaches for each step

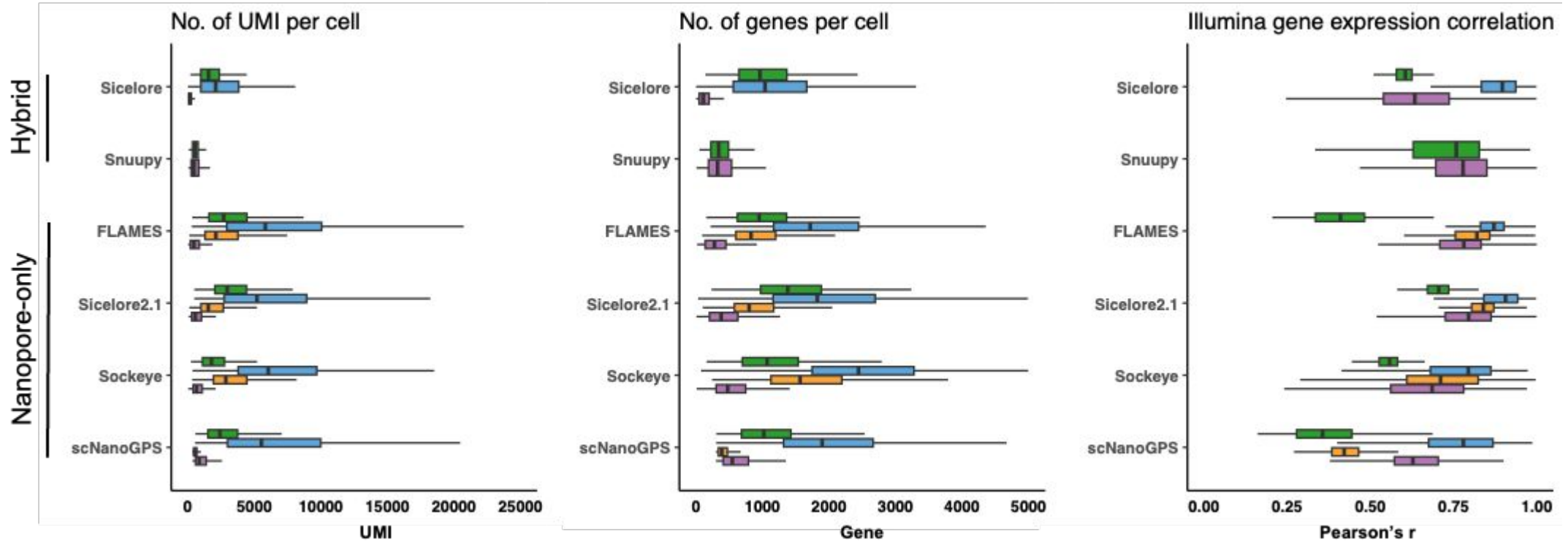
FASTQ	Cell barcod demulteplexing	Filter for chimeric reads	UMI deduplication	Consensus error correction	Isoform analysis
Sicelore	SR data	✓	SR data	✓	✓
Snuupy	SR data	✓	SR data	✓	PolyA based
scNapBar	SR data	✗	SR data	✗	✓
FLAMES	SR data or **BLAZE** 10X Shortlist (count over threeshold)	✗	Merge UMIs within a fixed edit distance.	✗	✓
Sockeye	10X Shortlist (count over threeshold)	✓	**UMItools** clustering algorithm	✗	**StringTie**
Sicelore 2.1	10X Shortlist	✓	Cluster UMIs based on edit distance.	✓	✓
scNanoGPS	No guidance (count over threeshold)	✓	Merge UMIs within a fixed edit distance.	✓	**LIQA**

LIQA *(Yu Hu et al., Genome Biology 2021)*
BLAZE *(Youpei You., Genome Biology 2023)*

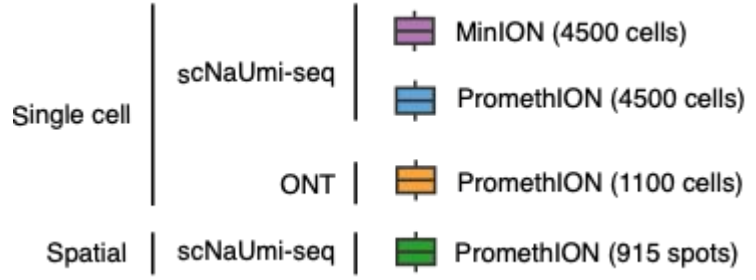


Which is **the most appropriate bioinformatics approach(es)** to analyse scRNAseq data to detect and quantify isoforms at the single-cell level?

Comparing the methods at different protocols and variable sequencing depth



- Gene expression correlation between Nanopore and Illumina improves with sequencing depth
- Sockeye reveals a subtle increase in the number of genes
- Sicelore2.1 shows the best gene expression correlation with short read data across all data types



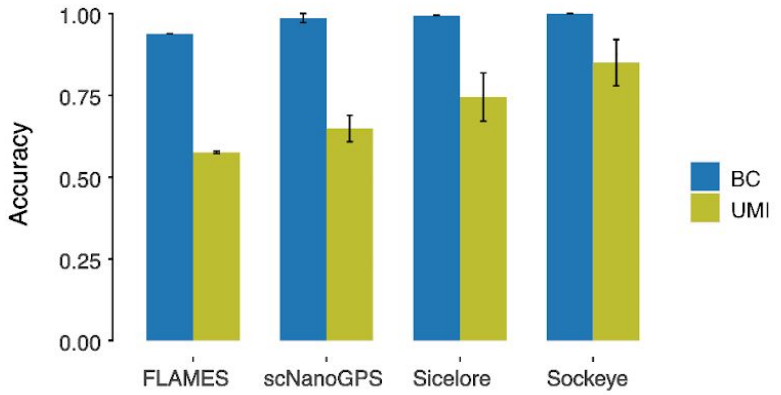
hamraoui et al., (in preparation)

Comparing the performances of long-read methods on simulated data

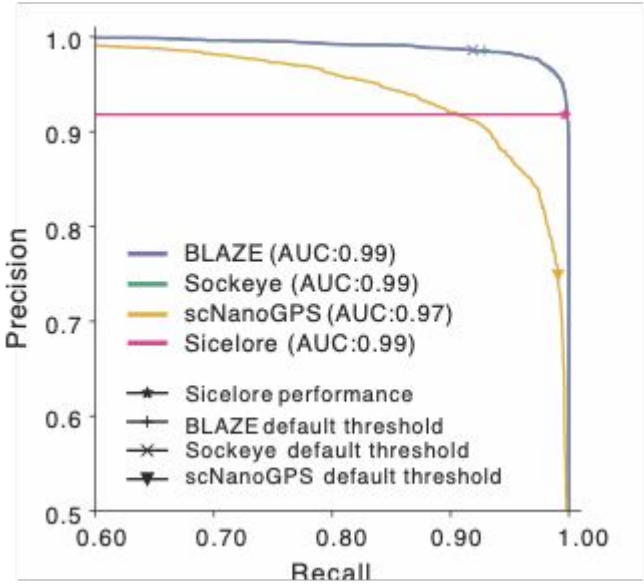
We simulated **5 dataset** (each comprising 5 million reads) with an **amplification rate** varying between **1** and **5**.

- Sockeye and BLAZE accurately identify single cells in nanopore data
- Sockeye shows the best UMI error correction strategy
- Accurate UMI error correction leads to best gene expresseion quatification performances

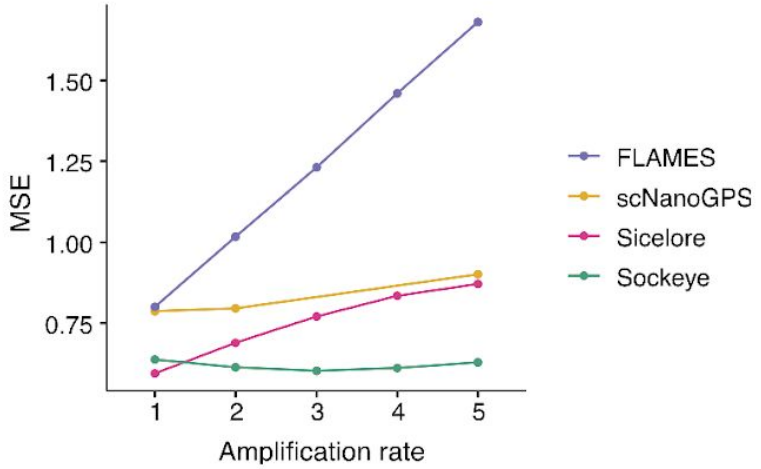
BC & UMI errors correction



Single cell identification



Gene expression estimation

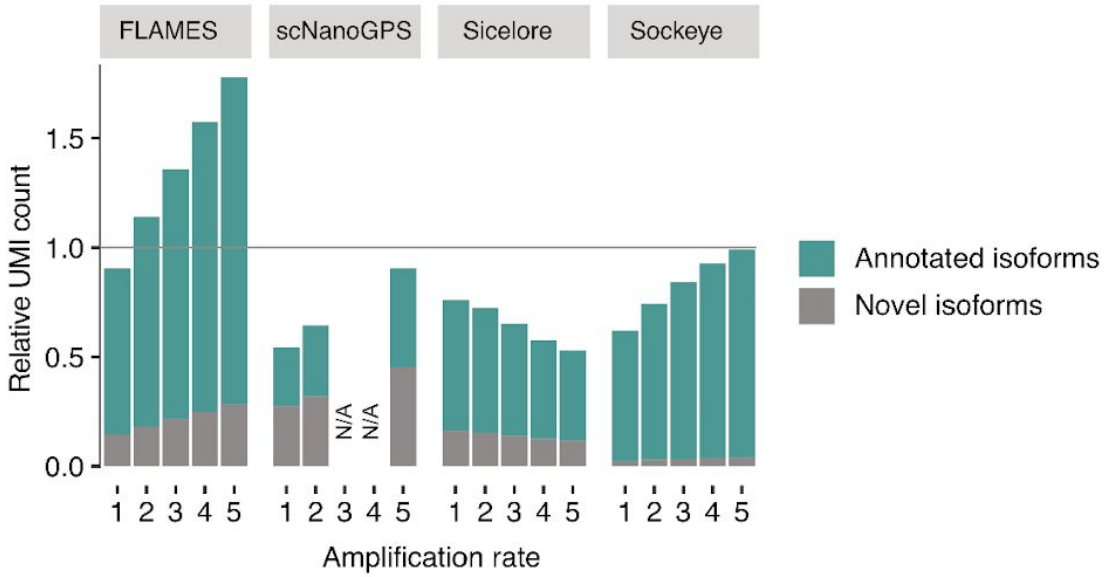


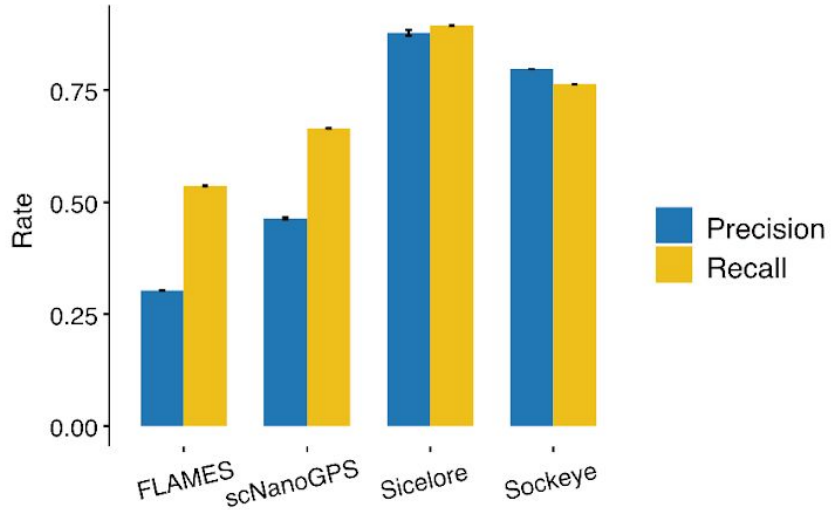
hamraoui et al., (in preparation)

Comparing the performances of long-read methods on simulated data

We simulated **5 dataset** (each comprising 5 million reads) with an **amplification rate** varying between **1** and **5**.

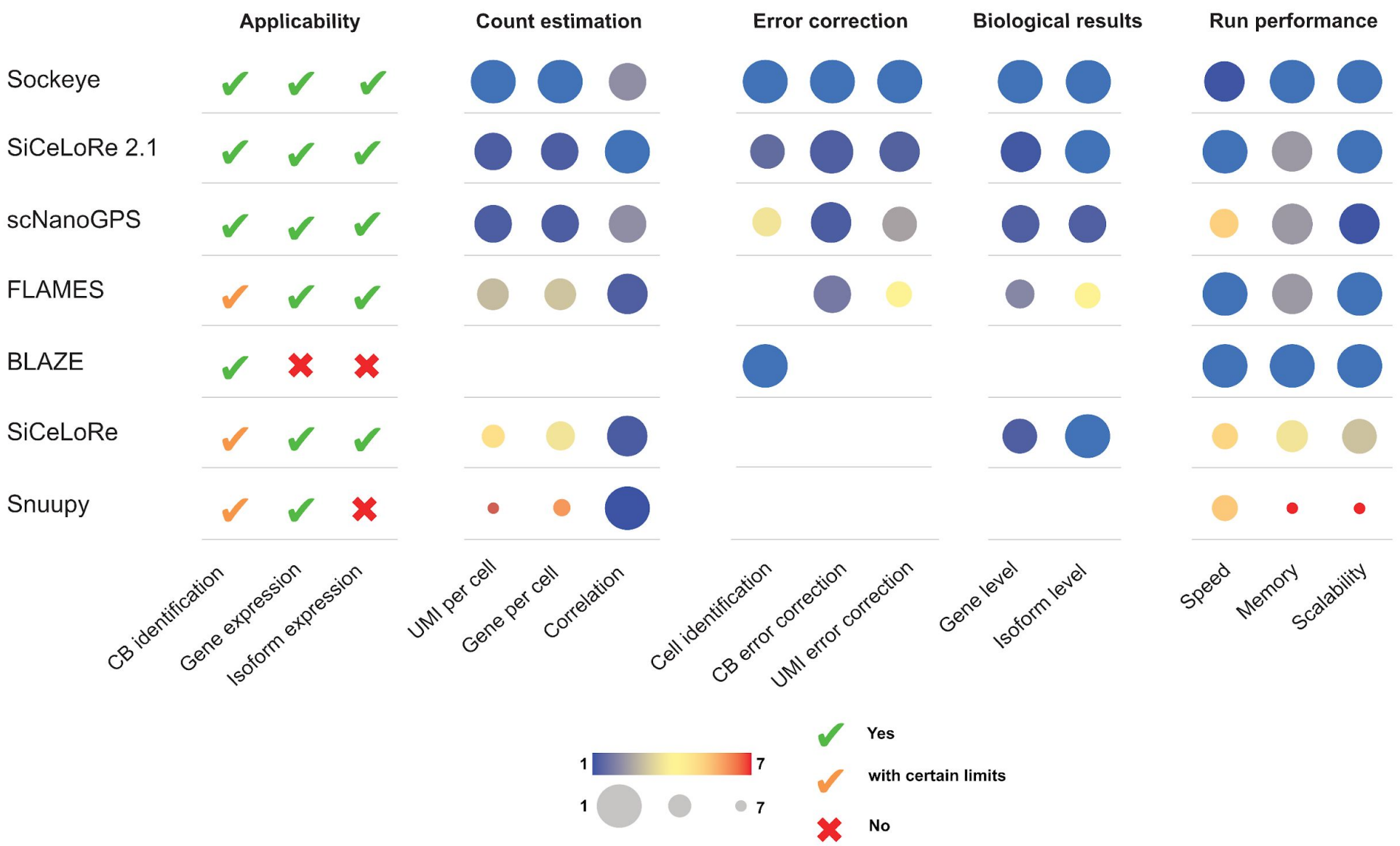
- Sockeye (StringTie) shows the best isoforms quantification accuracy
- Sicelore 2.1 outperforms other methods in isoforms identification





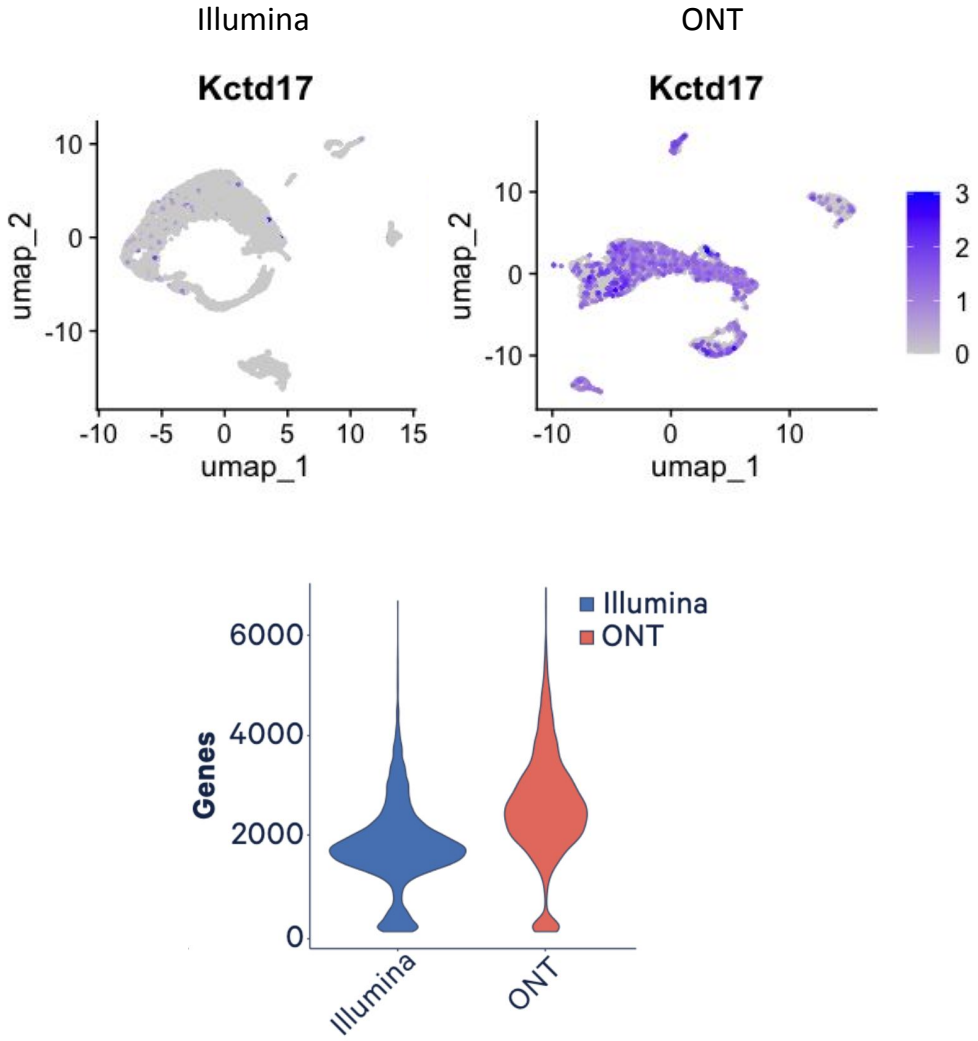
hamraoui et al., (in preparation)

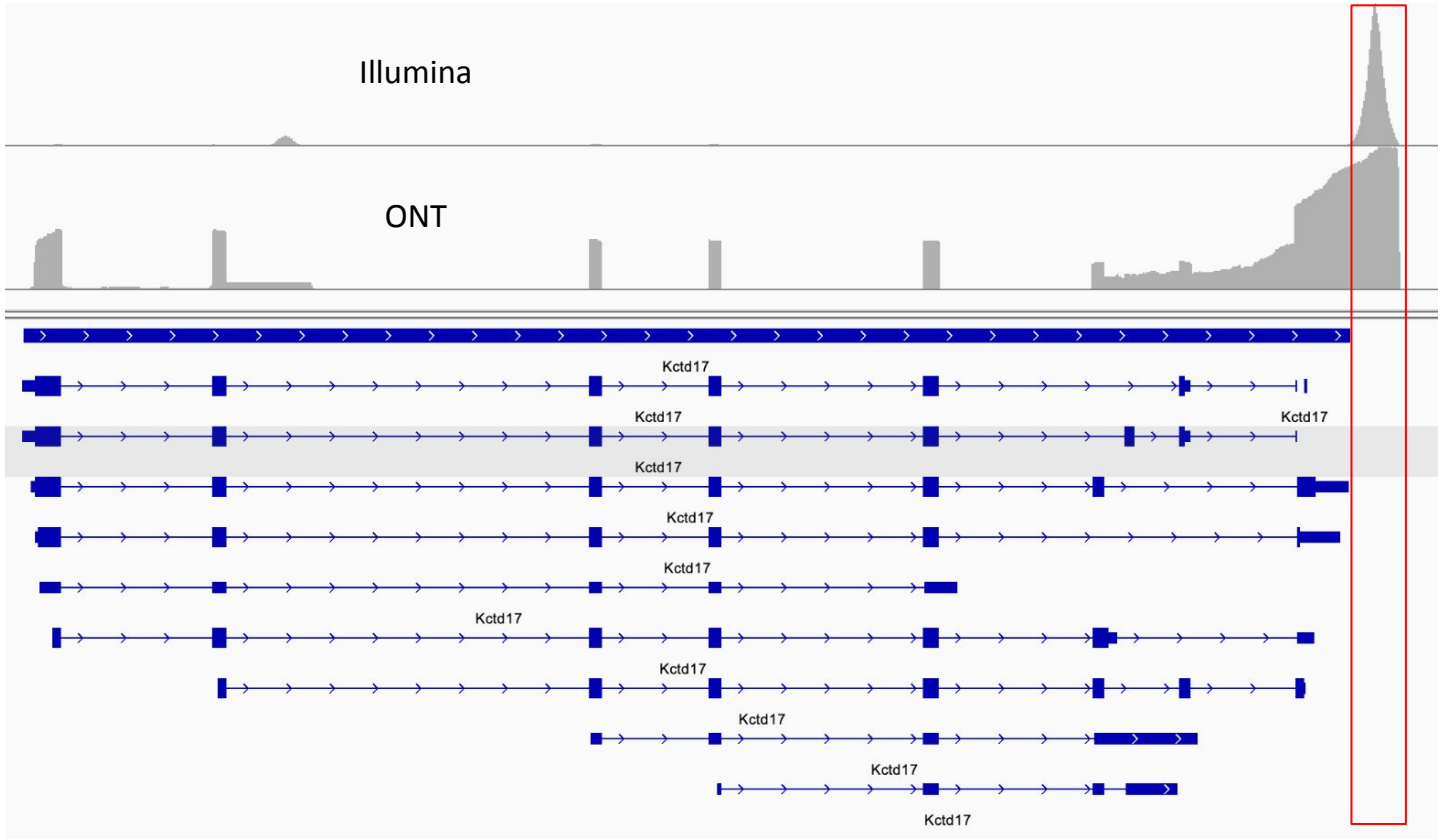
Ranking of methods across key aspects of evaluation criteria



hamraoui et al., (in preparation)

Short read 3’ biased with incomplete 3’UTR annotation lead to loss of gene expression





IGV visualisation of Kctd17 mouse gene

10x Genomics., Alternative transcript isoform detection with single cell and spatial resolution *App Note* (2022)

Single cell long read: Key points

- Nanopore sequencing preserve informations on splicing, SNP …
- Single-cell transcriptomics in long read enables a finer detection of isoforms.
- Long-read scRNA-seq improves gene detection in poorly annotated species.
- Sockeye *(wf-single-cell)* & Sicelore are efficient methods that provide a good compromise between the number of reads assigned, accuracy and computing cost.

Questions?

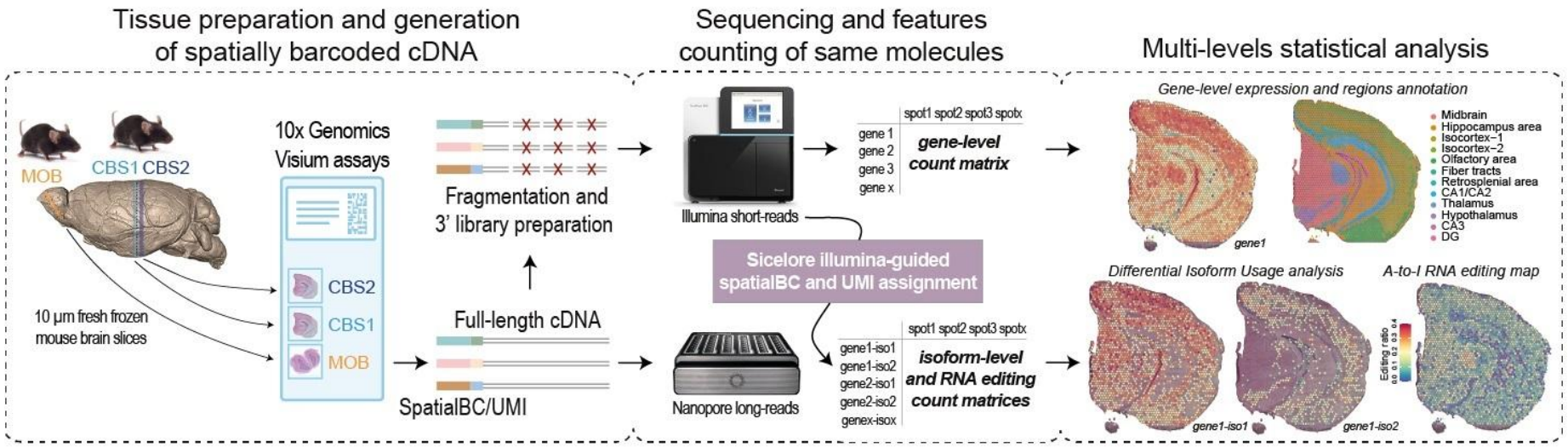
Morning program

- Why single cell long read is usefull?
- Preprocessing and comparison of tools.
- **Spatial isoform transcriptomics.**
- Demo: scRNAseq long read data pre-processing.

Spatial Isoform Transcriptomics (SiT)

Explorative method for characterizing spatial isoform variation and sequence heterogeneity

3 samples :
- MOB
- CBS1
- CBS2



(A) Experimental and computational steps for SiT analysis. Right side shows unsupervised gene expression clustering and gene- (short-read) and isoform-level (SiT) expression of Snap25 in a mouse coronal brain section (CBS2).

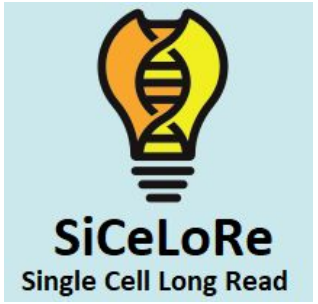
Data exploration online (https://www.isomics.eu)

- Different steps :
  - Sequenced by Nanopore long-read and analyzed with ***SiCeLoRe***
  - Visium data analysed with ***SpaceRanger***
  - Spatial analyses with ***Seurat***
  - Isoform switch with ***Isoswitch***
  - A-to-I RNA editing with ***Seurat***
  - Deconvolution of the spot to have an estimation of the cell type proportion with ***SPOTlight***

Lebrigand K, Bergenstråhle J, Thrane K, Mollbrink A, Meletis K, Barbry P, Waldmann R, Lundeberg J. The spatial landscape of gene expression isoforms in tissue sections. Nucleic Acids Res. 2023 May 8;51(8):e47. https://doi.org/10.1093/nar/gkad169

Single Cell Long Read (SiCeLoRe) pipeline

- A suite of tools dedicated to cell barcode / UMI (unique molecular identifier) assignment and bioinformatics analysis of highly multiplexed single cell Nanopore or PacBIo long read sequencing data.



The workflow integrates several sequential steps :

- Cell barcode and UMI assignment to long reads guided by a list of valid barcodes
- Transcript isoform identification
- Generation of molecules consensus sequences (UMI-guided error-correction)
- Production of count matrices (isoforms / junctions / SNPs x cells) for new modalities integration into standard single cell RNA-seq statistical analysis.

```
.
├── 01.readscan
│   ├── BarcodesValidated.csv
│   ├── fastq_pass.fastq.gz
│   └── passed
│       ├── BarcodesAssigned.tsv
│       └── ReadScanner.html
├── 02.mapping
│   ├── passed.bam -> /data/analysis/data_fierville/
│   └── passed.bam.bai -> /data/analysis/data_fiervi
├── 03.umis
│   ├── passedParsed.bai
│   ├── passedParsed.bam
│   ├── passedParsed.bam.genecounts.tsv
│   ├── passedParsed.bam.html
│   └── passedParsed.bam.UMIdepths.tsv
├── 04a.matrices
│   ├── histogram.png
│   ├── sicelore_cellmetrics.txt
│   ├── sicelore_genematrix.txt
│   ├── sicelore_genemetrics.txt
│   ├── sicelore.html
│   ├── sicelore_isobam.bam
│   ├── sicelore_isomatrix.txt
│   ├── sicelore_isometrics.txt
│   ├── sicelore_juncmatrix.txt
│   ├── sicelore_juncmetrics.txt
│   ├── sicelore.log
│   └── sicelore_molinfos.txt
├── 04b.matrices
│   ├── histogram.png
│   ├── molecules.fastq
│   ├── molecules.tags.GE.bam
│   ├── molecules.tags.GE.bam.bai
│   ├── sicelore_cellmetrics.txt
│   ├── sicelore_genematrix.txt
│   ├── sicelore_genemetrics.txt
│   ├── sicelore.html
│   ├── sicelore_isobam.bam
│   ├── sicelore_isomatrix.txt
│   ├── sicelore_isometrics.txt
│   ├── sicelore_juncmatrix.txt
│   ├── sicelore_juncmetrics.txt
│   ├── sicelore.log
│   └── sicelore_molinfos.txt
```

Lebrigand K, Waldmann R et al. (2020). High throughput error corrected Nanopore single cell transcriptome sequencing. Nature Communication 11, 4025. https://doi.org/10.1038/s41467-020-17800-6

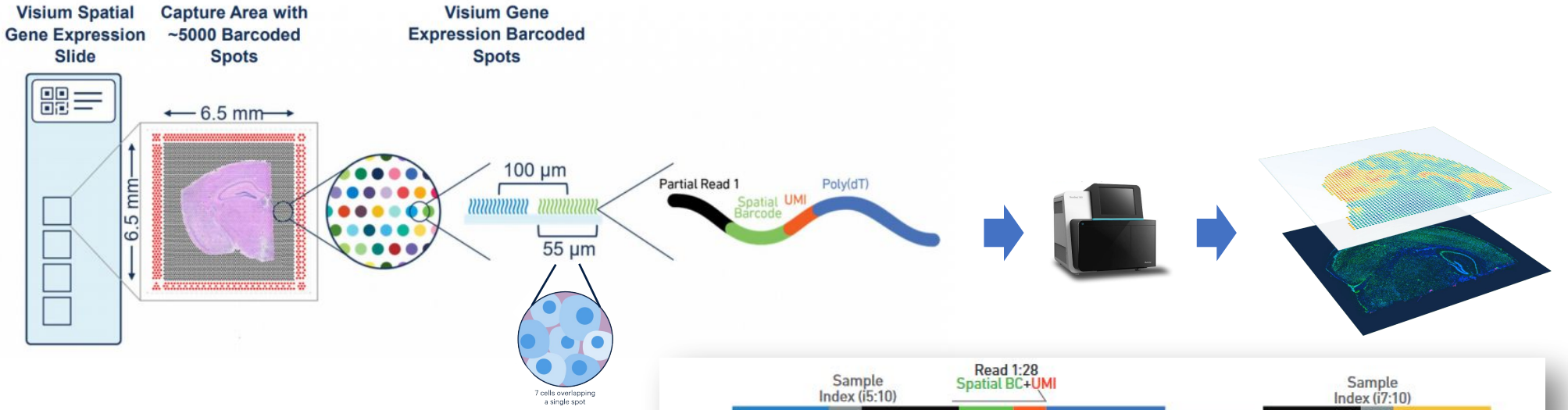
Visium - *in situ* capture spatial transcriptomics

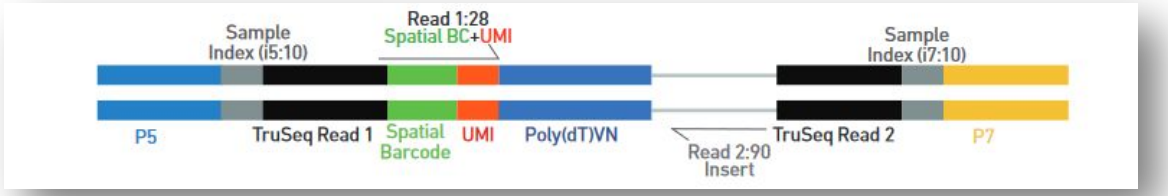
Whole-transcriptome spatial method developed by 10X Genomics

- Slide with 4 captures areas :
  - 6.5 mm x 6.5 mm
  - ~5 000 barcoded spots
- Individual spot is 55 microns
- Capture ~1-10 cells per spot depending on tissue type

- Tissue sections cut and placed onto the 10X Visium slide
- H&E stained and imaged with a brightfield microscope
- Spot's spatial barcode is retained throughout library preparation
- Sequenced by NGS to visualize gene expression across the tissue

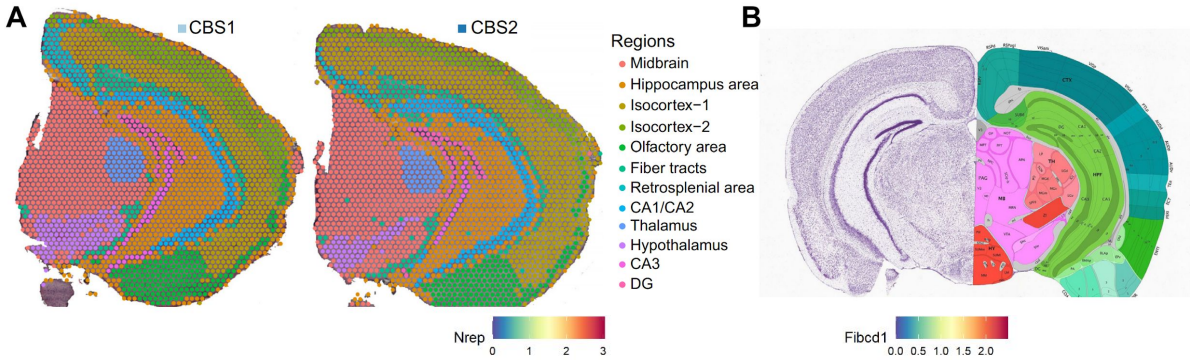
Spatial barcode / UMI assignment strategy identical to single cell transcriptomics

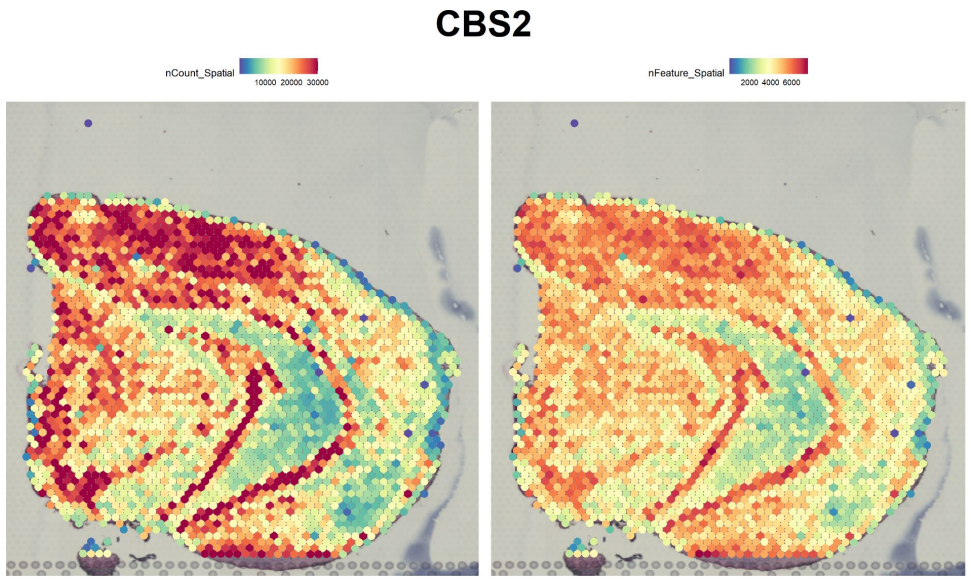


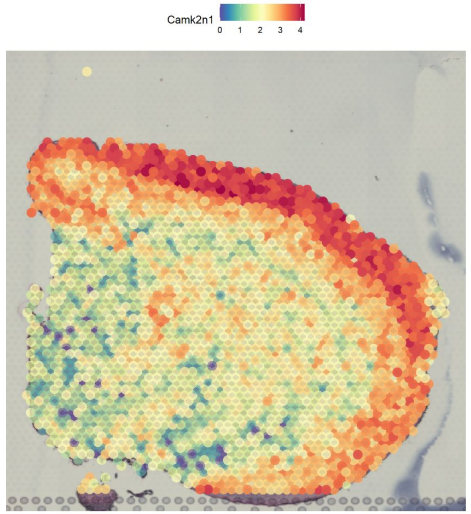


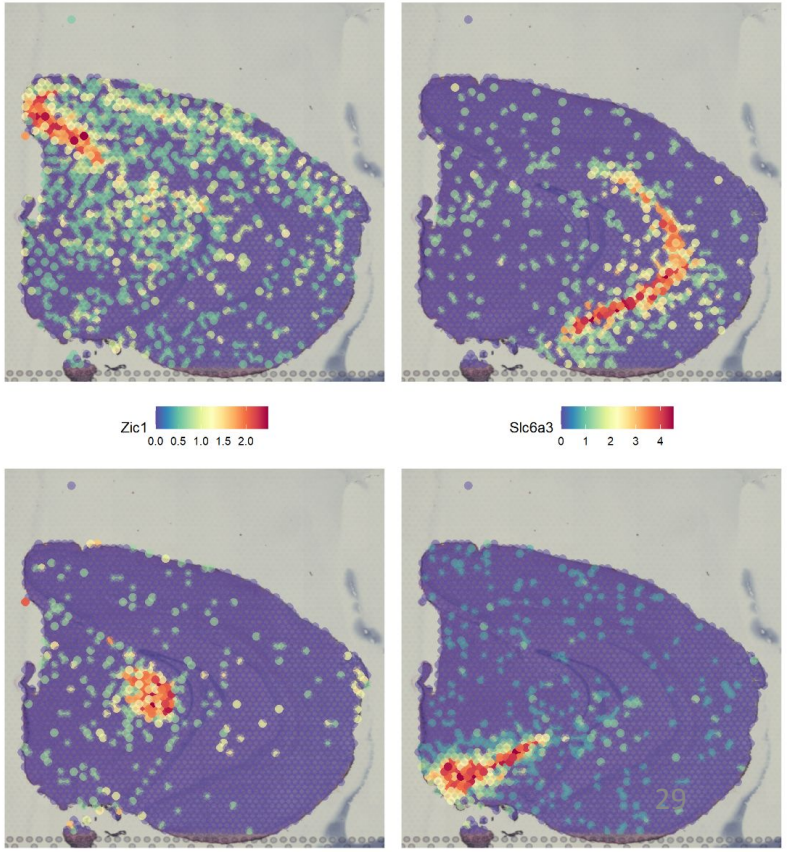
Spatial analyses with Seurat

```
CBS2 <- Load10X_Spatial(data.dir = paste0(data_dir, "data/spatial/CBS2"))

CBS2 <- SCTransform(CBS2, assay = "Spatial", verbose = FALSE, vst.flavor = 'v1')
CBS2 <- RunPCA(CBS2, assay = "SCT", verbose = TRUE)
CBS2 <- RunUMAP(CBS2, reduction = "pca", dims = 1:30)
CBS2 <- FindNeighbors(CBS2, reduction = "pca", dims = 1:30)
CBS2 <- FindClusters(CBS2, verbose = FALSE, resolution = 0.4)
```









More explored this afternoon

Seurat object with spatial and long-read data

```
## An object of class Seurat
## 157917 features across 918 samples within 6 assays
## Active assay: Spatial (31053 features, 0 variable features)
##  3 layers present: counts, data, scale.data
##  5 other assays present: SCT, ISOG, ISO, JUNC, AtoI
##  2 dimensional reductions calculated: pca, umap
##  1 image present: slice1
```

```
An object of class Seurat
27998 features across 190 samples within 1 assay
Active assay: RNA (27998 features, 0 variable features)
 1 layer present: counts
```

5 different assays :

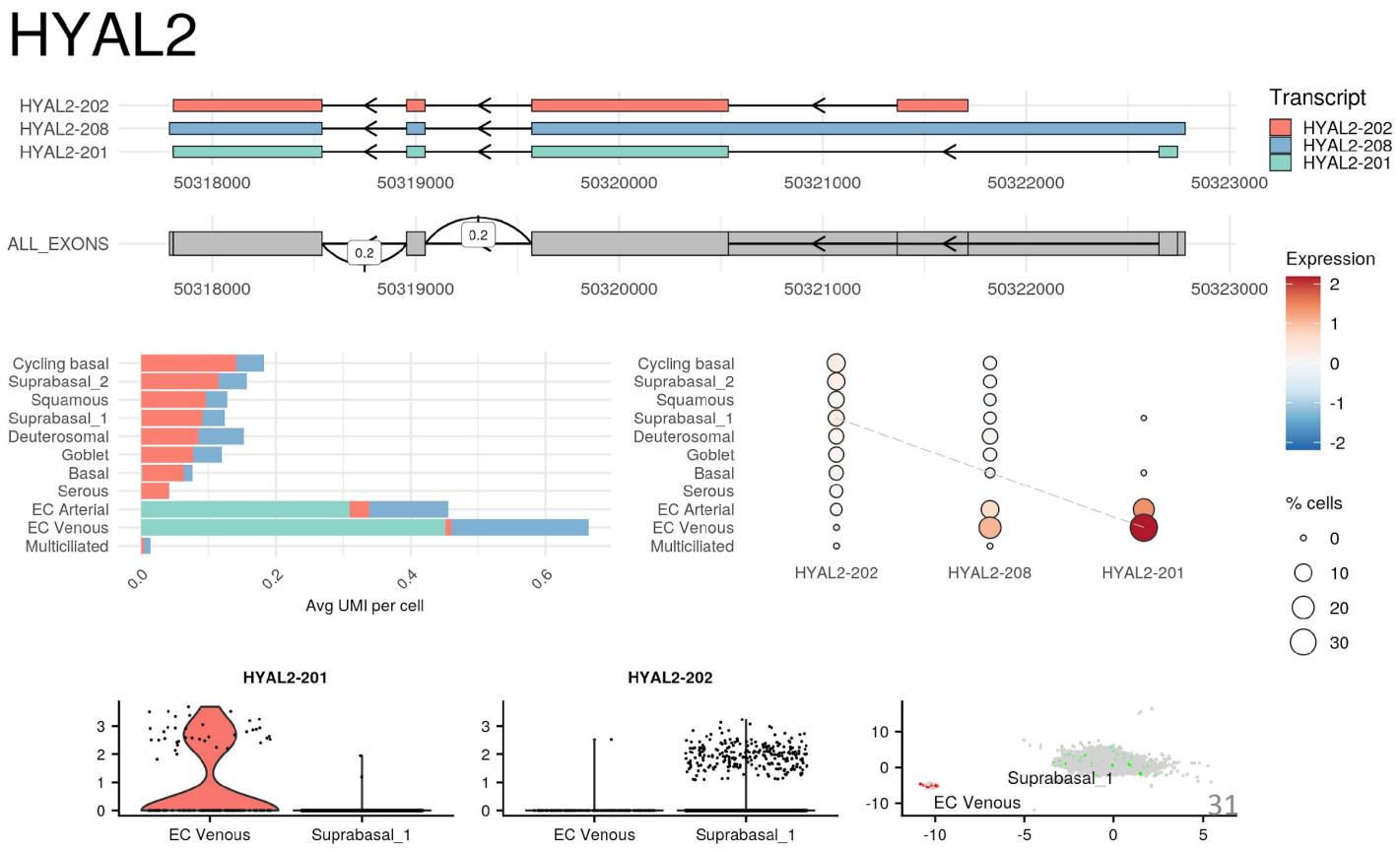
- **'Spatial'** containing gene-level raw short-read data from the Space Ranger output
- **'ISOG'** containing the gene-level Nanopore long-read data
- **'ISO'** containing isoform-level transcript information where only the molecules where all exons are observed are kept
- **'JUNC'** containing each individual exon-exon junction observation per isoform
- **'AtoI'** cotnaining exonic editing sites from the RADAR database and from the Licht et al., 2019, study, for which we observed at least one UMI in our dataset

Isoswitch – R package

- Facilates the characterization of isoform expression in long-read single-cell datasets.
- Includes a set of functions and reports built on top of *Seurat*, *ggplot* and *rmarkdown* that can be used
  - to search,
  - to visualize and
  - to document isoform expression patterns, and particularly isoform switches between cell identities.

General Workflow :

1. Input data setup & pre-processing
2. Isoform characterization
3. Isoform switch detection
4. Gene reports



Ignacio Atienza (previous student)

Isoswitch – Input data setup & pre-processing

1. Input data setup & pre-processing
   - Work with Seurat objects with gene- and isoform-level counts generated by ScNaUmi-seq protocol (Lebrigand et al 2020)
     - A gene-level [gene x cell] matrix count
       - stored in assay “RNA” in Seurat object
     - An isoform-level [isoform x cell] matrix count
       - generated by SiCeLoRe pipeline
       - stored in a separate “isoform” assay
   - The method ***iso_preprocess()*** removes low-expression transcripts and single-isoform genes which are irrelevant for the isoform switch analysis
   - New “multi” matrix stored in a new assay

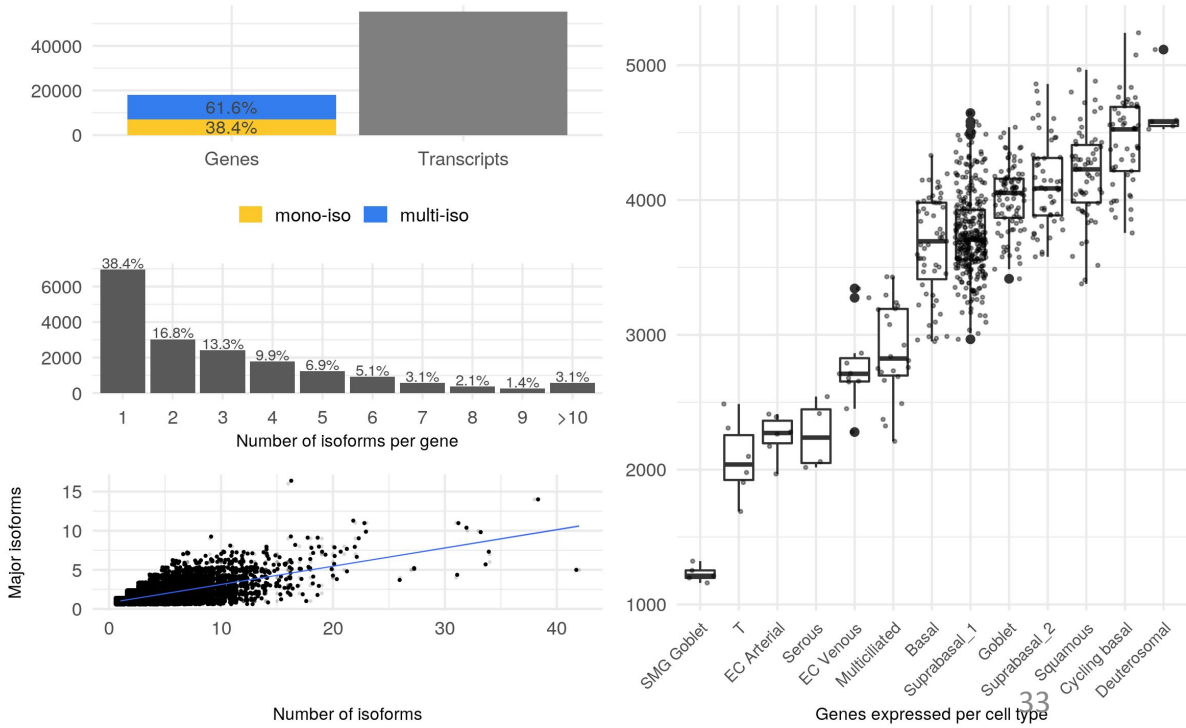
```
head(rownames(seurat@assays$multi@counts))
#> [1] "BCS1L..ENST00000359273"   "PPP1R10..ENST00000461593"
#> [3] "OSBPL9..ENST00000428468"  "TEF..ENST00000406644"
#> [5] "CBX5..ENST00000209875"    "TRAP1..ENST00000246957"

seurat <- iso_preprocess(seurat, assay="ISO", new_assay="multi", filter_threshold=5)
```

feature <chr>	gene_id <chr>	transcript_id <chr>	sum <dbl>	total_gene <dbl>	n_isofs <int>	max_sum <dbl>	perc <dbl>	is_major <lgl>	is_top <lgl>
Klc2..ENSMUST00000156717	Klc2	ENSMUST00000156717	21	175	5	128	12.00000000	FALSE	FALSE
Cyfip1..ENSMUST00000163845	Cyfip1	ENSMUST00000163845	1	38	3	26	2.63157895	FALSE	FALSE
Capn15..ENSMUST00000212520	Capn15	ENSMUST00000212520	15	15	1	15	100.00000000	TRUE	TRUE
Klc2..ENSMUST00000025798	Klc2	ENSMUST00000025798	22	175	5	128	12.57142857	TRUE	FALSE
Osbpl1a..ENSMUST00000132594	Osbpl1a	ENSMUST00000132594	4	98	6	55	4.08163265	FALSE	FALSE
Trabd..ENSMUST00000169891	Trabd	ENSMUST00000169891	1	175	5	153	0.57142857	FALSE	FALSE
Mocs1..ENSMUST00000173033	Mocs1	ENSMUST00000173033	1	4	2	3	25.00000000	FALSE	FALSE
Afmid..ENSMUST00000131268	Afmid	ENSMUST00000131268	1	1	1	1	100.00000000	TRUE	TRUE

Isoswitch – Isoform characterization

1. Input data setup & pre-processing
2. Isoform characterization
   - The method ***iso_compute_stats()*** parses the isoform raw count matrix
     - returning a data frame with stats on the expression of each transcript
   - The method ***plot_assay_stats()*** plots a summary to describe succintly the isoform distribution in the dataset
     - number of genes,
     - number of transcripts,
     - distribution of isoforms and
     - number of genes per cell type

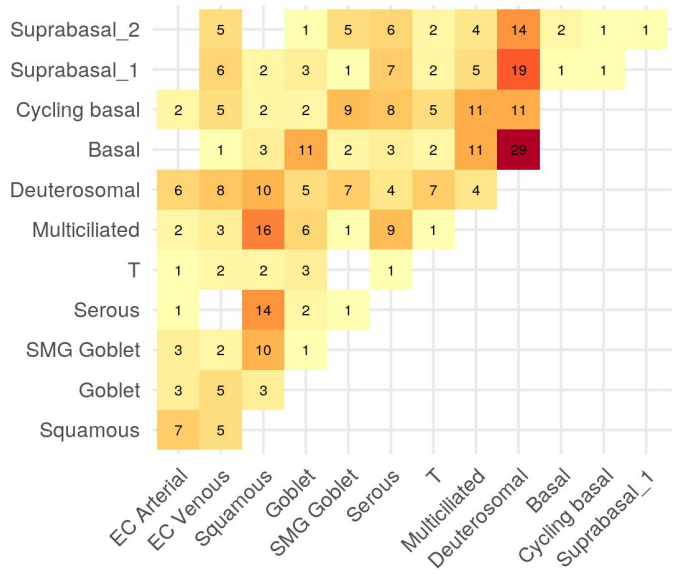
```
stats <- iso_compute_stats(seurat@assays$multi@counts) %>% arrange(gene_id)
head(stats, n=4)
#>                 feature gene_id   transcript_id sum total_gene n_isofs max_sum
#> 1 A1BG..ENST00000596924    A1BG ENST00000596924   5          8       2       5
#> 2 A1BG..ENST00000598345    A1BG ENST00000598345   3          8       2       5
#> 3  A2M..ENST00000495709     A2M ENST00000495709  10         14       2      10
#> 4  A2M..ENST00000318602     A2M ENST00000318602   4         14       2      10
#>       perc is_major is_top
#> 1 62.50000     TRUE   TRUE
#> 2 37.50000     TRUE  FALSE
#> 3 71.42857     TRUE   TRUE
#> 4 28.57143    FALSE  FALSE

plot_assay_stats(seurat, "isoform")
```

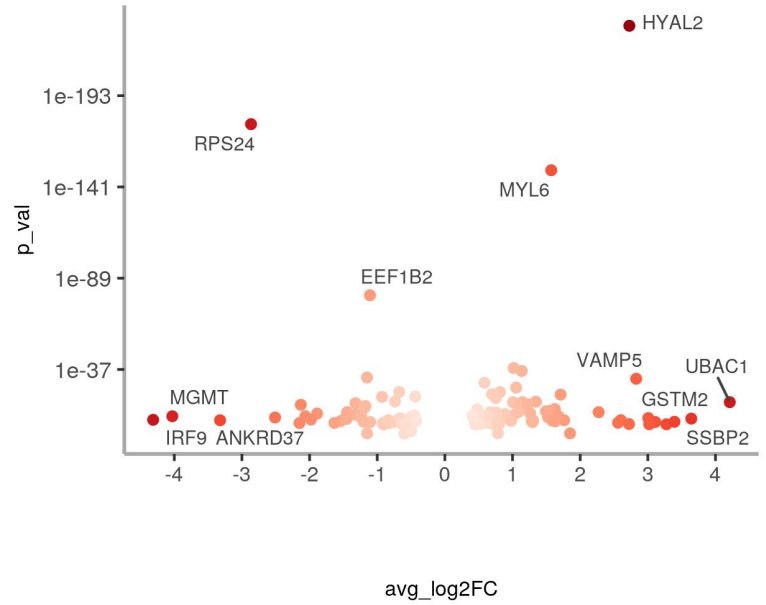


Isoswitch – Isoform switch detection

1. Input data setup & pre-processing
2. Isoform characterization
3. Isoform switch detection
   - **“Isoform switch”** refers to an event where two isoforms of the same gene are considered markers of different clusters.
   - The marker search is implemented on the method ***ISO_SWITCH_ALL()*** and are based on Seurat’s FindMarkers function.
   - Returns a data frame of transcripts identified as markers of a given cluster for a given gene, one transcript per row.

```
clusters <- levels(seurat@active.ident)
switch_markers <- ISO_SWITCH_ALL(seurat, clusters, assay="isoform",
                                 min.pct=0, logfc.threshold=0.40)

pl1 <- plot_marker_matrix(seurat, switch_markers)
pl2 <- plot_marker_score(adult, switch_markers, facet=FALSE, overlaps=16)
pl1 | pl2
```





Isoswitch – Gene reports

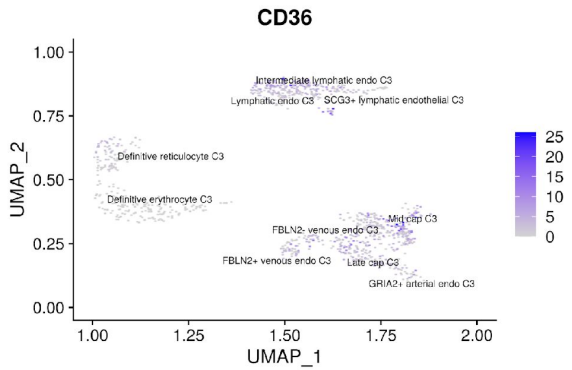
1. Input data setup & pre-processing
2. Isoform characterization
3. Isoform switch detection
4. Gene reports
   - After identifying genes of interest, two ways to produce gene-level reports:
     - The ***isoswitch_report()*** method produces a compact plot of the gene.
     - The ***render_html_gene_report()*** method renders an html version of the report

CD36 gene report

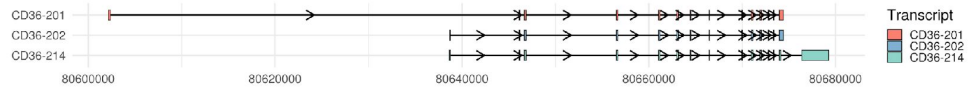
Code

Description

The protein encoded by this gene is the fourth major glycoprotein of the platelet surface and serves as a receptor for thrombospondin in platelets and various cell lines. Since thrombospondins are widely distributed proteins involved in a variety of adhesive processes, this protein may have important functions as a cell adhesion molecule. It binds to collagen, thrombospondin, anionic phospholipids and oxidized LDL. It directly mediates cytoadherence of Plasmodium falciparum parasitized erythrocytes and it binds long chain fatty acids and may function in the transport and/or as a regulator of fatty acid transport. Mutations in this gene cause platelet glycoprotein deficiency. Multiple alternatively spliced transcript variants have been found for this gene. [provided by RefSeq, Feb 2014]



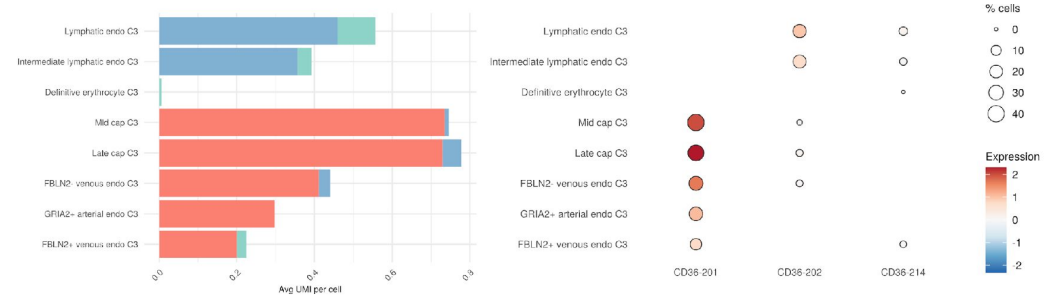
Isoforms

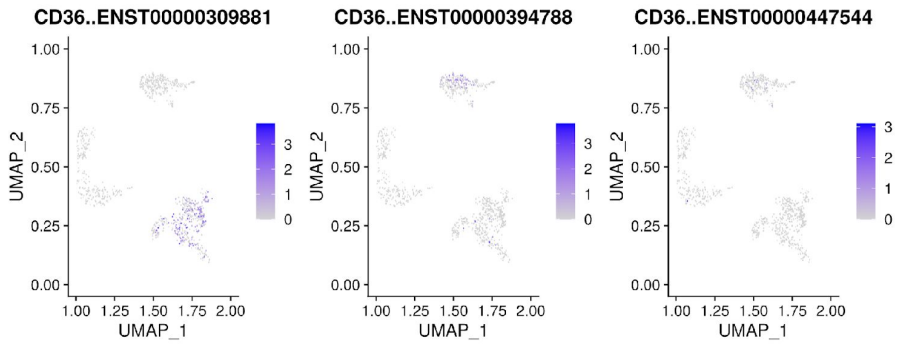


Stats on raw UMI counts:

id	transcript_id	biotype	cds_length	sum	total_gene	perc	is_major	is_top
CD36-201	ENST00000309881	protein_coding	1419	327	483	67.701863	TRUE	TRUE
CD36-202	ENST00000394788	protein_coding	1419	130	483	26.915114	TRUE	FALSE
CD36-214	ENST00000447544	protein_coding	1419	26	483	5.383023	FALSE	FALSE

Cell-type expression:





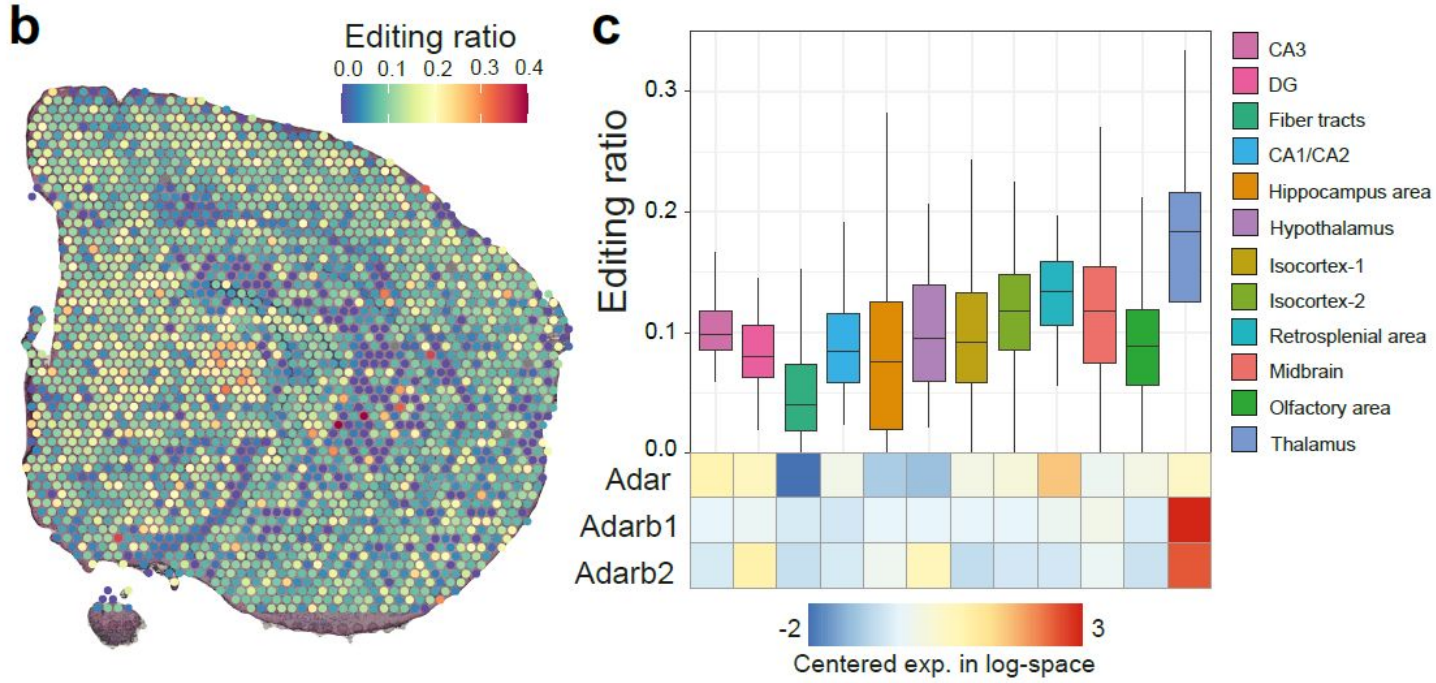
Switches

1 switches found for CD36

geneId	t1	c1	p1	log2fc1	t2	c2	p2	log2fc2
CD36	ENST00000309881	Mid cap C3	1.55e-23	-1.8	ENST00000394788	Lymphatic endo C3	5.25e-10	1.1

A-to-I RNA editing explored thanks to long-read

- SiCeLoRe pipeline allows to explore Single Nucleotide Variation (SNV)
- Exploration of 5 817 adenosine-to-inosine (A-to-I) RNA editing sites (Ramaswami et al., 2013 (RADAR), Licht et al., 2019)
  - Principal source of transcript sequence heterogeneity in the mammalian transcriptome
  - Involved in proper neuronal function and reported in neurological and neurodegenerative diseases (epilepsy, developmental disorders…)
- New assay 'AtoI' in Seurat object, containing exonic editing sites (_snpmatrix.csv) :
  - count = non edited UMI count = nbrA
  - data = edited UMI count = nbrG (equal to I)
  - scale.data = editing ratio = nbrG / (nbrG + nbrA)



Cell type deconvolution tools

Article | Published: 02 May 2022

Spatially informed cell-type deconvolution for spatial transcriptomics

Ying Ma & Xiang Zhou

Nature Biotechnology **40**, 1349–1359 (2022) | Cite this article

36k Accesses | **137** Citations | **60** Altmetric | Metrics

Deep learning and alignment of spatially resolved single-cell transcriptomes with Tangram

Tommaso Biancalani, Gabriele Scalia, Lorenzo Buffoni, Raghav Avasthi, Ziqing Lu, Aman Sanger, Neriman Tokcan, Charles R. Vanderburg, Åsa Segerstolpe, Meng Zhang, Inbal Avraham-Davidi, Sanja Vickovic, Mor Nitzan, Sai Ma, Ayshwarya Subramanian, Michal Lipinski, Jason Buenrostro, Nik Bear Brown, Duccio Fanelli, Xiaowei Zhuang, Evan Z. Macosko & Aviv Regev

Article | Published: 18 February 2021

Robust decomposition of cell type mixtures in spatial transcriptomics

Dylan M. Cable, Evan Murray, Luli S. Zou, Aleksandrina Goeva, Evan Z. Macosko, Fei Chen & Rafael A. Irizarry

CellDART: Cell type inference by domain adaptation of single-cell and spatial transcriptomic data

Sungwoo Bae, Kwon Joong Na, Jaemoon Koh, Dong Soo Lee, Hongyoon Choi, Young Tae Kim

doi: https://doi.org/10.1101/2021.04.26.441459

SPOTlight: seeded NMF regression to deconvolute spatial transcriptomics spots with single-cell transcriptomes

Marc Elosua-Bayes, Paula Nieto, Elisabetta Mereu, Ivo Gut, Holger Heyn

Single-cell and spatial transcriptomics enables probabilistic inference of cell type topography

Alma Andersson, Joseph Bergenstråhle, Michaela Asp, Ludvig Bergenstråhle, Aleksandra Jurek, José Fernández Navarro & Joakim Lundeberg

Comprehensive mapping of tissue cell architecture via integrated single cell and spatial transcriptomics

Vitalii Kleshchevnikov, Artem Shmatko, Emma Dann, Alexander Aivazidis, Hamish W King, Tong Li, Artem Lomakin, Veronika Kedlian, Mika Sarkin Jain, Jun Sung Park, Lauma Ramona, Elizabeth Tuck, Anna Arutyunyan, Roser Vento-Tormo, Moritz Gerstung, Louisa James, Oliver Stegle, Omer Ali Bayraktar

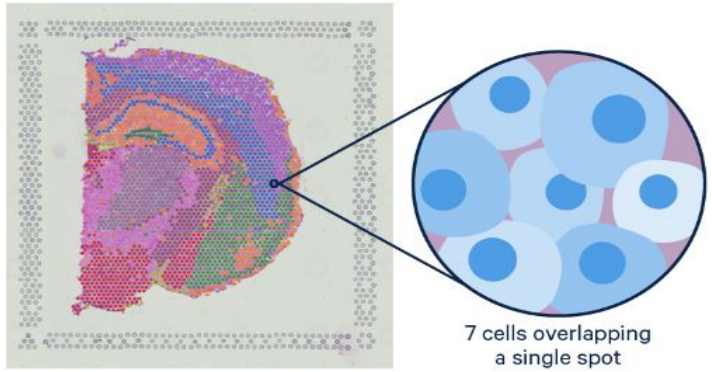
Article | Open access | Published: 29 April 2022

Reference-free cell type deconvolution of multi-cellular pixel-resolution spatially resolved transcriptomics data

Brendan F. Miller, Feiyang Huang, Lyla Atta, Arpan Sahoo & Jean Fan

Nature Communications **13**, Article number: 2339 (2022) | Cite this article

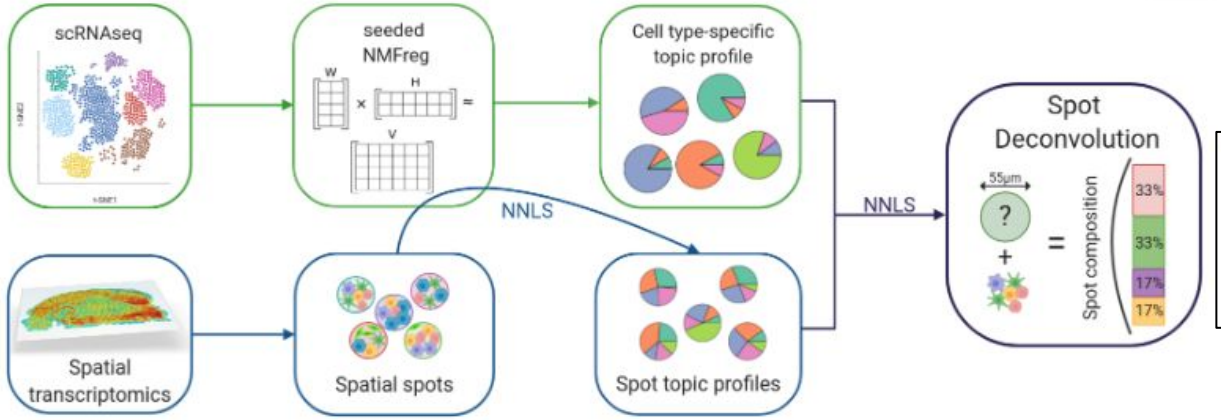
Spots deconvolution with SPOTlight



- Tool that enables the deconvolution of cell types and cell type proportions present within each capture location comprising mixtures of cells.
- Based on learning topic profile signatures for each cell type and finding which combination of cell types fits best the spot we want to deconvolute.

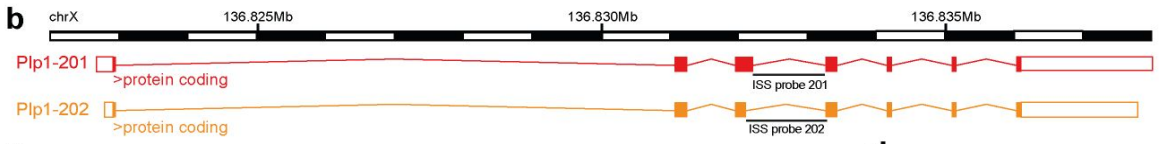
The minimal unit of data required to run SPOTlight are:

- ST matrix with the expression, raw or normalized (rows = genes and columns = capture locations)
- Single cell matrix with the expression, raw or normalized (rows = genes and columns = cells)
- Vector indicating the cell identity for each column in the single cell expression matrix

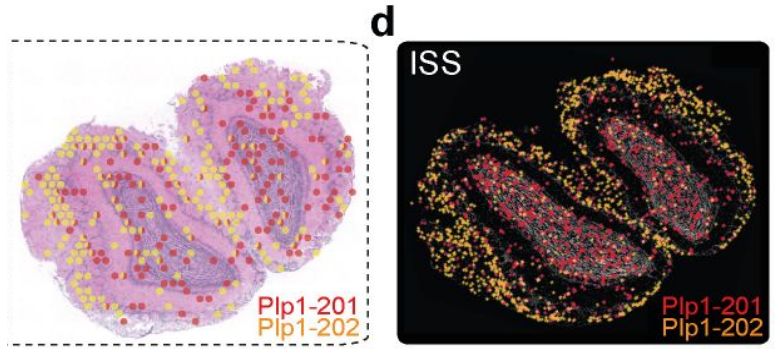


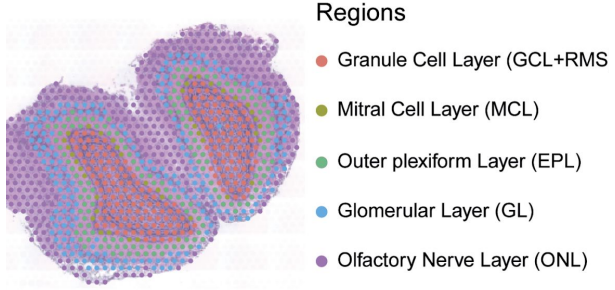

```
> res$mat[100:105,] # deconvoluted matrix
                    MyOligo         N2       OEC1       OEC2       OEC3       OEC4       OEC5       OPC
ACGTTAATGTCGAAGA-1 0.5971502 0.06315422 0.00000000 0.00000000 0.00000000 0.25386656 0.05045814 0.03537091
ACGTTAGATTTGCCCG-1 0.6460356 0.06911981 0.00000000 0.00000000 0.00000000 0.14160721 0.03132053 0.11191686
ACTACGCGTTAGAATT-1 0.3171349 0.26365579 0.00000000 0.00000000 0.00000000 0.34649135 0.00000000 0.07271800
ACTCCCGAATTCGTTT-1 0.3401568 0.12717052 0.00000000 0.06963253 0.06858932 0.10704909 0.06585297 0.22154874
ACTCCCTAATGCTAAA-1 0.2854633 0.14452655 0.01057627 0.31511973 0.01401977 0.08444532 0.12775068 0.01809836
ACTCCCTAGAATAGTA-1 0.0000000 0.28400478 0.00000000 0.13748356 0.01161386 0.15171681 0.38633667 0.02884432
```

Spots deconvolution with SPOTlight

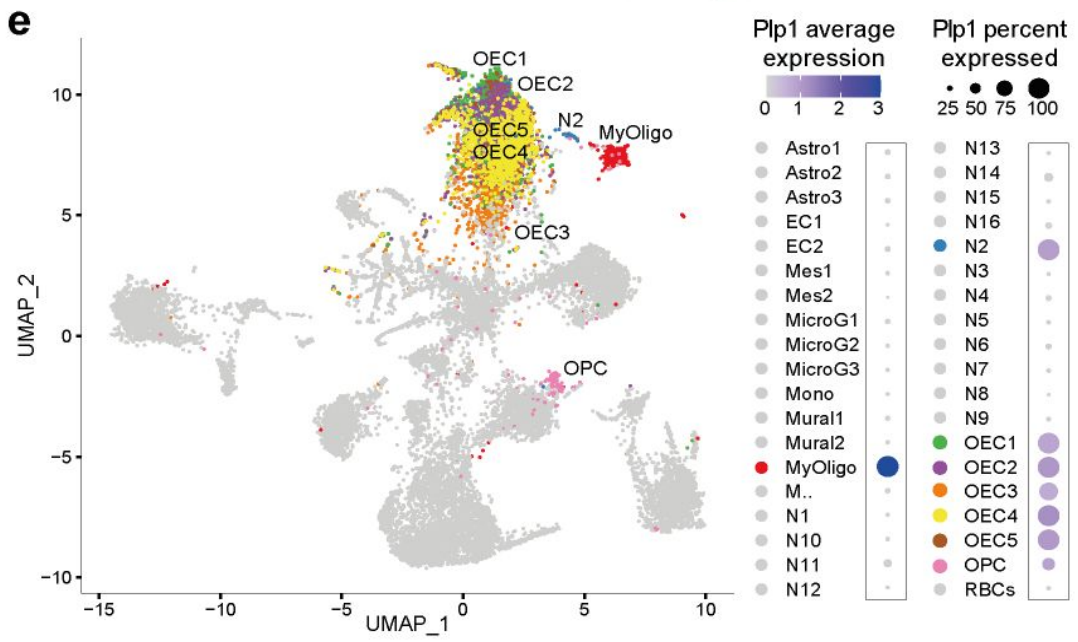


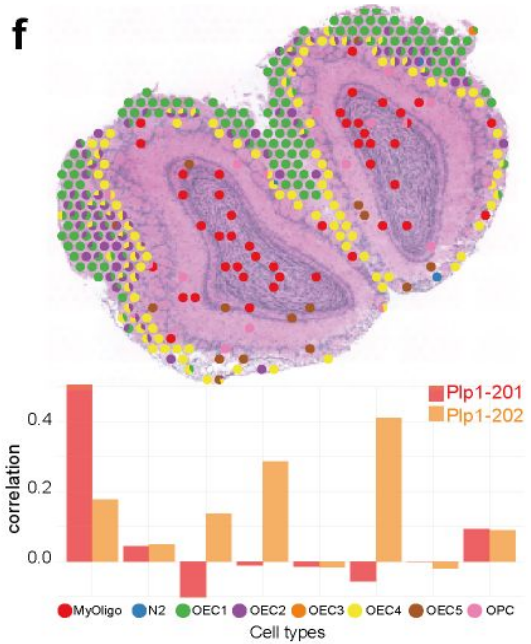
Proteolipid Protein 1 (Plp1) is a gene involved in severe pathologies associated with CNS dysmyelination






```
# read the single-cell dataset
tepe <- readRDS("../LONG_READ/data/spatial/tepe.WT.rds")
Idents(tepe) <- "ClusterName"
DimPlot(tepe, label = FALSE)

# Expression of Plp1 in Tepe
plp1_sum <- aggregate(tepe@assays$RNA@counts["Plp1",], list(tepe$ClusterName), sum)
plp1_sum$nb <- as.data.frame(table(tepe$ClusterName))$Freq
plp1_sum$avg <- plp1_sum$x / plp1_sum$nb
plp1_cell_types <- as.character(plp1_sum[plp1_sum$avg > 1,]$Group.1)

tepe[['ClusterNamePlp1']] <- tepe$ClusterName
tepe <- SetIdent(tepe, value=tepe$ClusterNamePlp1)

# find the marker genes for each cell types
Idents(tepe) <- "ClusterName"
cluster_markers_all <- FindAllMarkers(object = tepe,
                                      assay = "SCT",
                                      verbose = TRUE,
                                      only.pos = TRUE)

# keep only the cells that expressed PLP1
Idents(tepe) <- "ClusterNamePlp1"
sub_tepe <- subset(x = tepe, idents = plp1_cell_types,downsample = 100,  seed = 1)
sub_tepe@meta.data <- droplevels(sub_tepe@meta.data)
table(sub_tepe@active.ident)
cluster_markers_all <- cluster_markers_all[cluster_markers_all$cluster %in% plp1_cell_types,]
cluster_markers_all$cluster <- droplevels(cluster_markers_all$cluster)

# run spotlight
set.seed(123)
spotlight_ls <- SPOTlight(x = sub_tepe, y = MOB, n_top = 10,
                          groups = sub_tepe@active.ident, mgs = cluster_markers_all,
                          weight_id = "avg_log2FC", group_id = "cluster",
                          verbose = TRUE, gene_id = "gene")
saveRDS(spotlight_ls, "output/spotlight_tepe_ClusterName_MOB_100.rds")
```





Used a single-cell dataset : Tepe et al., Cell report, 2018 (GSE121891)

Spatial Isoform Transcriptomics (SiT)

- Accurate single-cell and spatial transcriptomics using Nanopore long-read sequencing is feasible
- Long reads sequencing reveals transcript diversity that is missed with standard short reads workflows
- Single Nucleotide Variation calls (SNV, editing) in single-cell and in a spatial context can be achieve
- More analysis on this dataset this afternoon…

The spatial landscape of gene expression isoforms in tissue sections

Kevin Lebrigand, Joseph Bergenstråhle, Kim Thrane, Annelie Mollbrink, Konstantinos Meletis, Pascal Barbry ✉, Rainer Waldmann, Joakim Lundeberg Author Notes

Nucleic Acids Research, Volume 51, Issue 8, 8 May 2023, Page e47, https://doi.org/10.1093/nar/gkad169

Published: 17 March 2023 **Article history** ▾

https://github.com/ucagenomix/SiT

Morning program

- Why single cell long read is usefull?
- Preprocessing and comparison of tools.
- Spatial isoform transcriptomics.
- **Demo: scRNAseq long read data pre-processing.**

ARTICLE

https://doi.org/10.1038/s41467-020-17800-6 OPEN

High throughput error corrected Nanopore single cell transcriptome sequencing

Kevin Lebrigand [1✉], Virginie Magnone[1], Pascal Barbry [1✉] & Rainer Waldmann [1✉]

Practical session (Afternoon)

Identification of transcript isoforms in full-length E18 mouse brain transcriptome

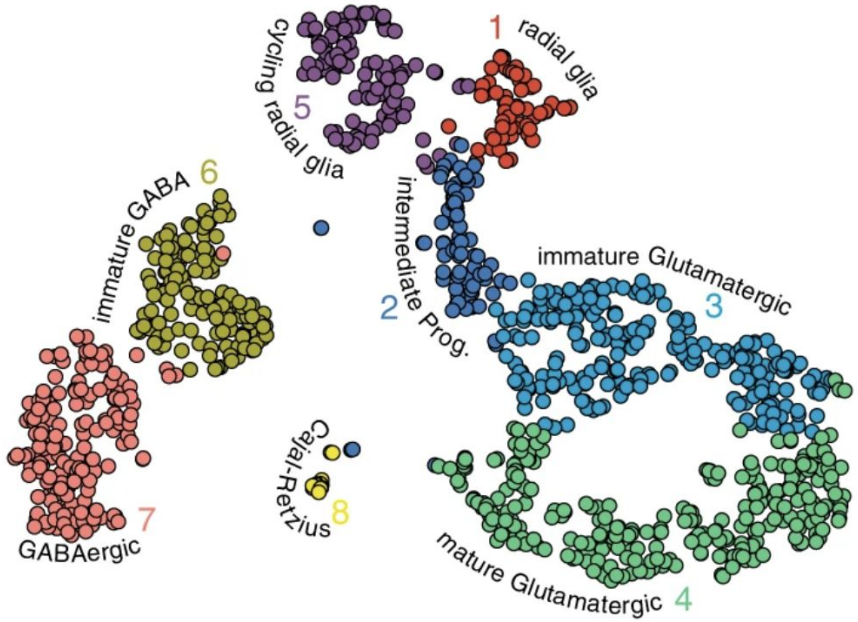
2 technical replicates

- 190 cells
- 951 cells

Long read sequencing
ONT PromethION

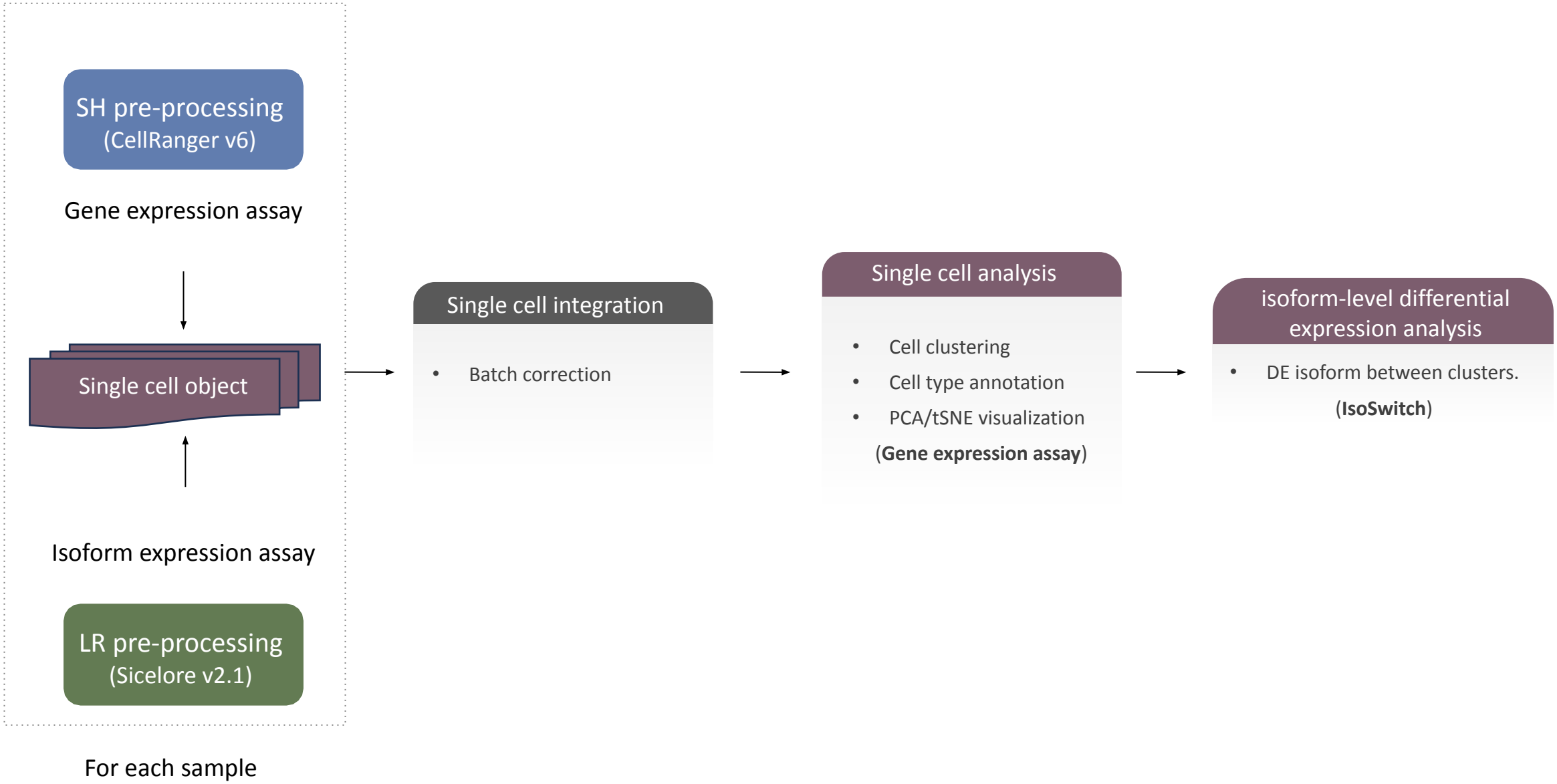
Pre-processing
Sicelore

SiCeLoRe
Single Cell Long Read



Aim 1: Searching for genes with a transcript isoform expression that differed between clusters (**Isoform switch**).

Downstream analysis for differential expressed Isoform



Demo: pre-processing raw ONT reads using Sicelore 2.1 workflow

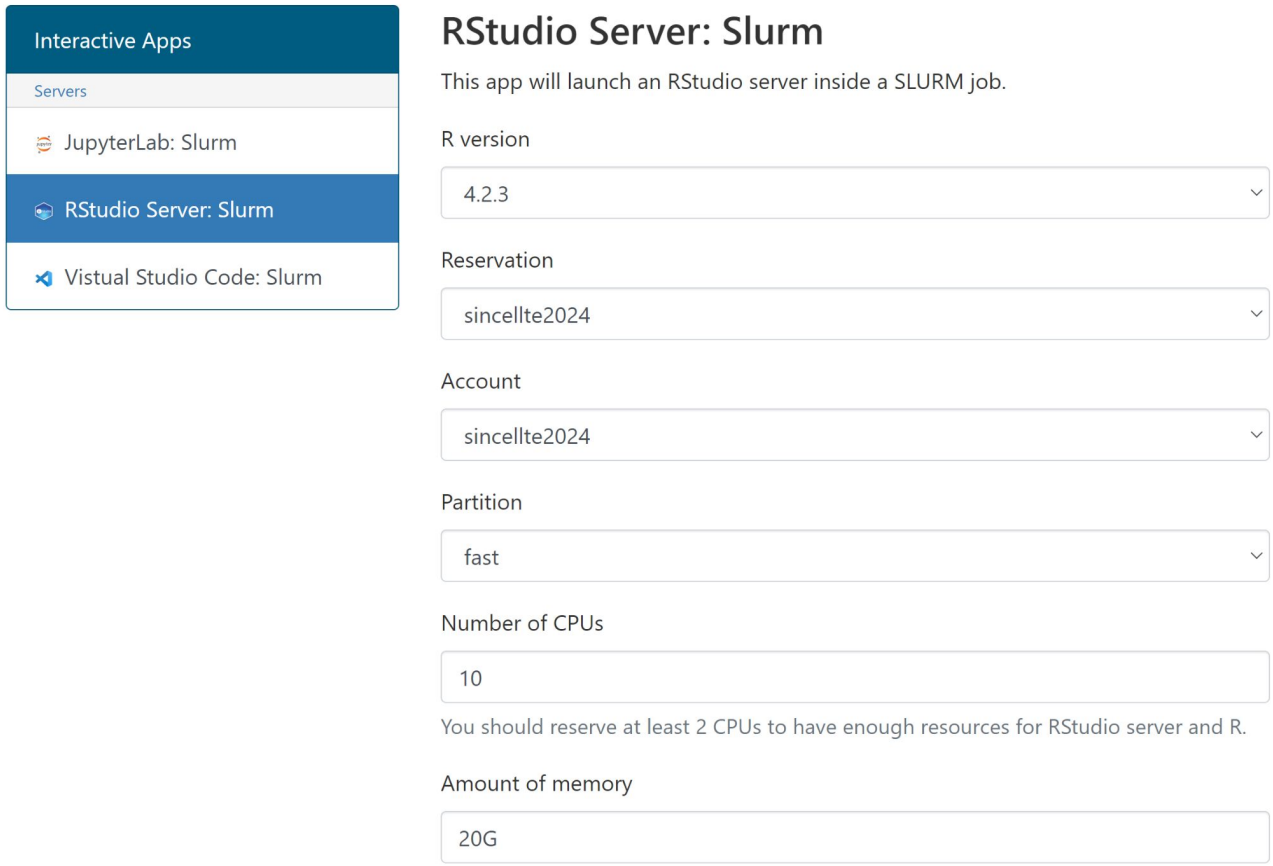
- With Nextflow on Roscoff server :

```
> git clone https://github.com/ucagenomix/sicelore-2.1.git
> cd sicelore-2.1
> module purge
> module load nextflow/24.04.1 samtools/1.21 minimap2/2.28 spoa/4.1.4
> nextflow run sicelore-nf/main.nf
```

- With Nextflow and conda :

```
> git clone https://github.com/ucagenomix/sicelore-2.1.git
> cd sicelore-2.1
> nextflow run sicelore-nf/main.nf -profile conda
```

Workshop session



R version: 4.2.3
Nb of CPUs: 10
Memory: 40G
Nb of hours: 2

1- cp notebook (3 files):

cp -r /shared/projects/sincellte2024/04_thursday/scripts_long_read/ $HOME

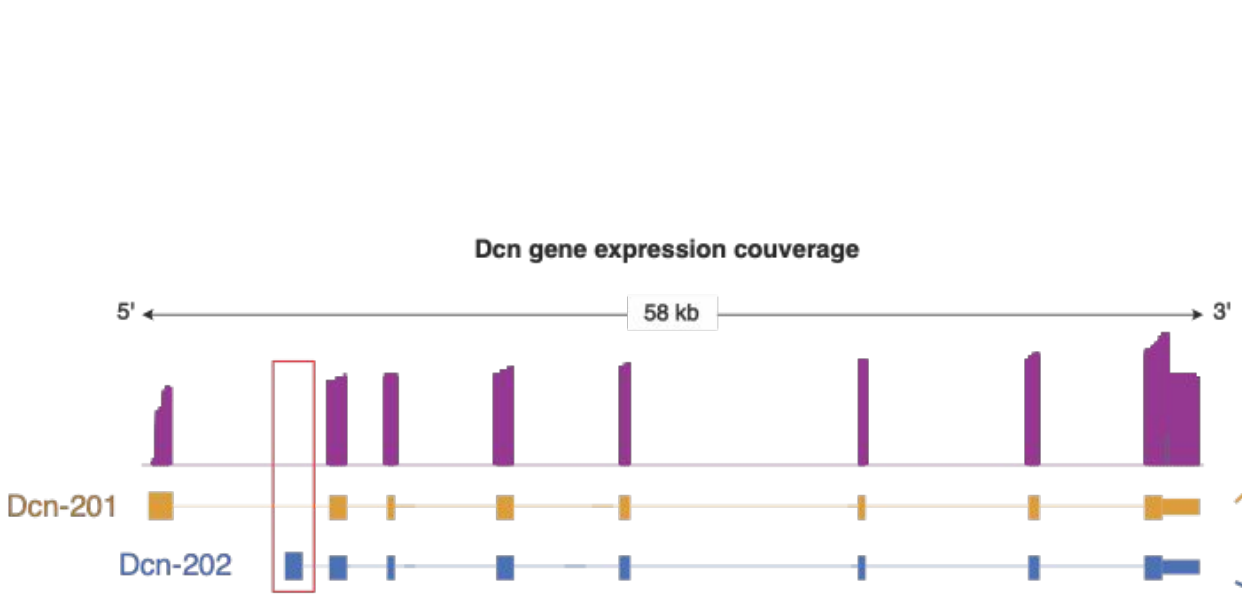
PS Don't copy all the data!

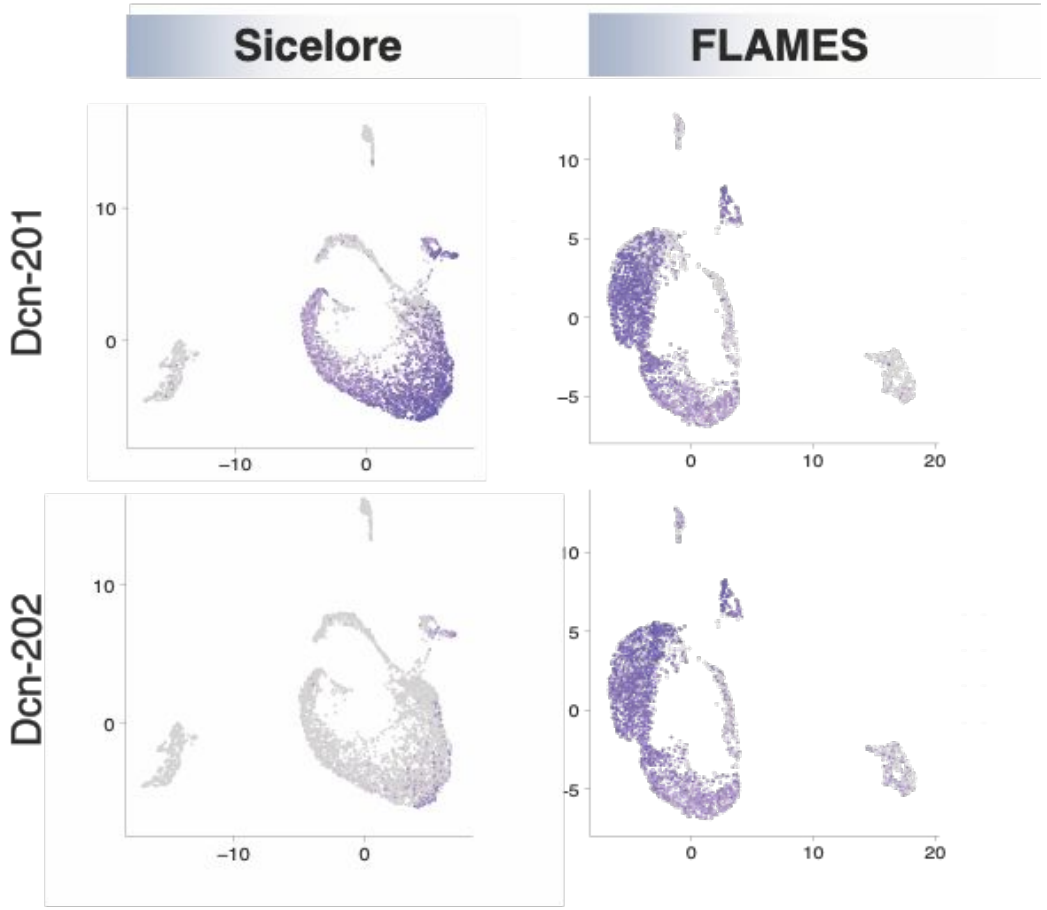
Ressources:

- https://www.isomics.eu/index.html
- https://github.com/ucagenomix/sicelore/tree/master
- https://github.com/ucagenomix/sicelore-2.1
- https://github.com/ucagenomix/isoswitch
- https://github.com/ucagenomix/SiT

Biological results at the isoform level

Comparison of isoform identification and quantification between Sicelore and FLAMES.





- Sicelore accurately detects and quantifies isoforms.