

# Slide sets of team meeting at Harbin Medical University

Zhou Ying

# Preface

This slide sets mainly comprises slides presented during team meetings at Harbin Medical University. They are provided here to illustrate the progress of reading, thinking, and studying at this stage. The 1-7 sets of slide were created in the early phase of this stage, focusing on learning and practicing various bioinformatics analysis skills, and exploring potential directions for thesis project. Following practice with techniques, literature reviews, a review of previous research experiences, and present team general direction, I ultimately decided to focus on comparing the transcriptomes of fetal brain cells and brain tumor cells at the single-cell level. Moreover, I conducted a interested topic discussion on the oncofetal ecosystem, which is presented in the eighth set of slides. Additionally, the final set of slides serves as a summary of my thesis project stage, outlining my study process and reflecting on various thoughts during the study. **For easy navigation, all slides are listed in the 'Contents' page, with clickable links that direct to the corresponding slide covers.**

# Content

1. Principles and analysis methods of DNA binding protein capture techniques

**CHIP-Seq, CUT&RUN, CUT&TAG technologies**

Zhou Ying

2. Specific analysis methods of CUT&RUN data

**Insertsize of CUT&RUN data and dynamic regulation of transcriptional state**

Zhou Ying  
2<sup>nd</sup> Dec. 2022

3. Literature study of chromatin compaction and evaluated feasibility of related project

**Summary of chromatin compaction reading and thinking**

Zhou Ying

4. Explanation of basic question of RNA-seq data

**Explanation of DESeq2, edgeR, GSEA Analysis, and RNAseq Normalization Methods**

Zhou Ying

5. Report on the Practice and Learning of Translatome Data Analysis

**Computation for Ribo-seq Data in Ribosome Profiling**

ZHOU YING  
7<sup>th</sup> April 2023

6. Report on the Practice and Learning of AS and APA analysis from RNA-seq data

**Alternative splicing and polyadenylation analysis**

Zhou Ying  
27<sup>th</sup> Mar. 2023

2024/4/7

93

7. Sharing analysis experience in image data and other types of biological data

**Analysis of Different Types of Biological Data**  
——image data

Zhou Ying

8. Topic discussion of oncofetal ecosystem

**Oncofetal reprogramming and oncofetal ecosystem (EMT and cell-cell communication)**

Zhou Ying  
26<sup>th</sup> Jan. 2024

Zhou Ying 11<sup>th</sup> Jan. 2024

...

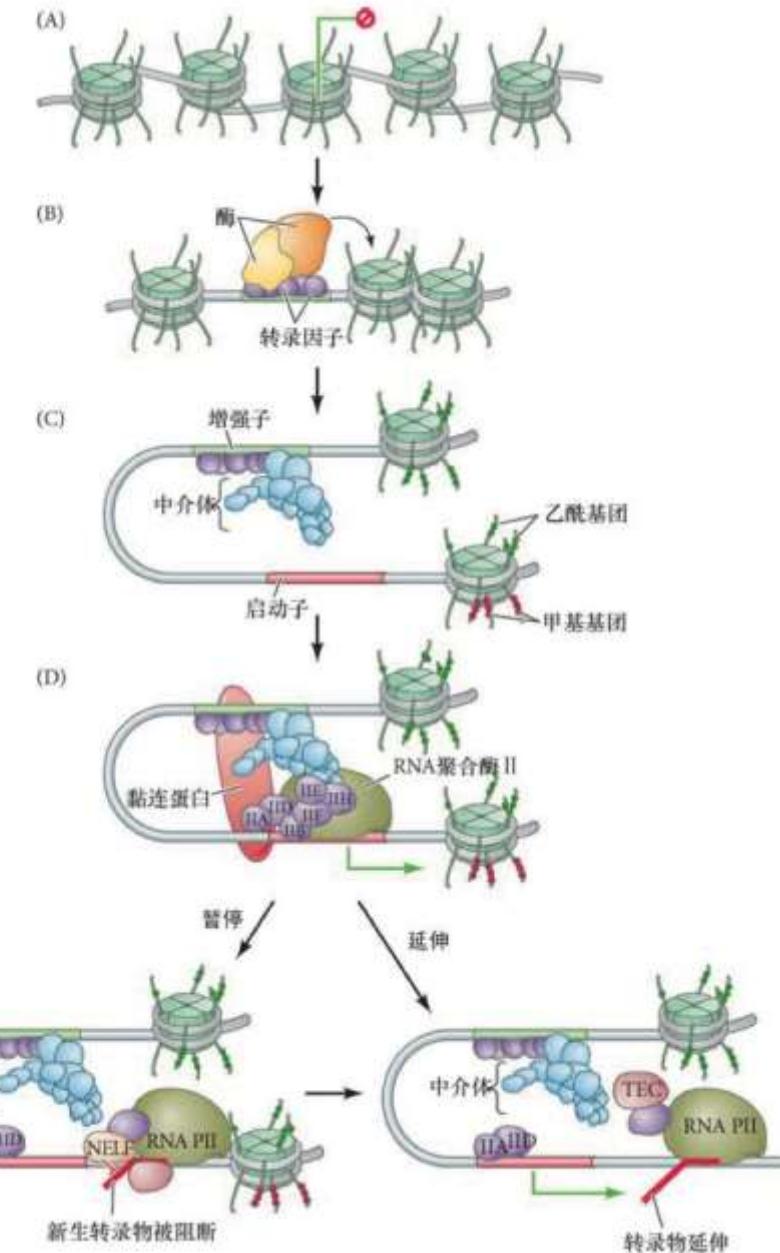
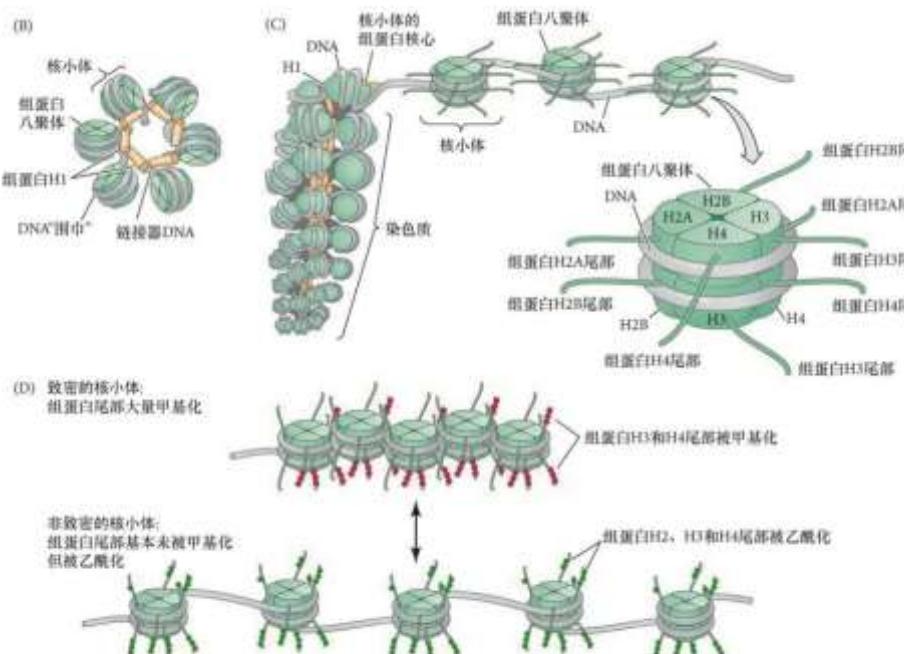
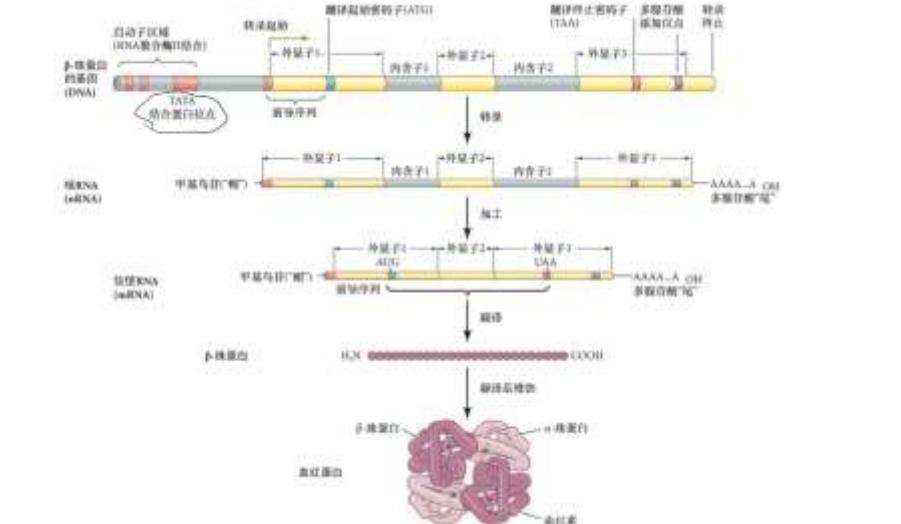
9. Thesis Project Stage Summary

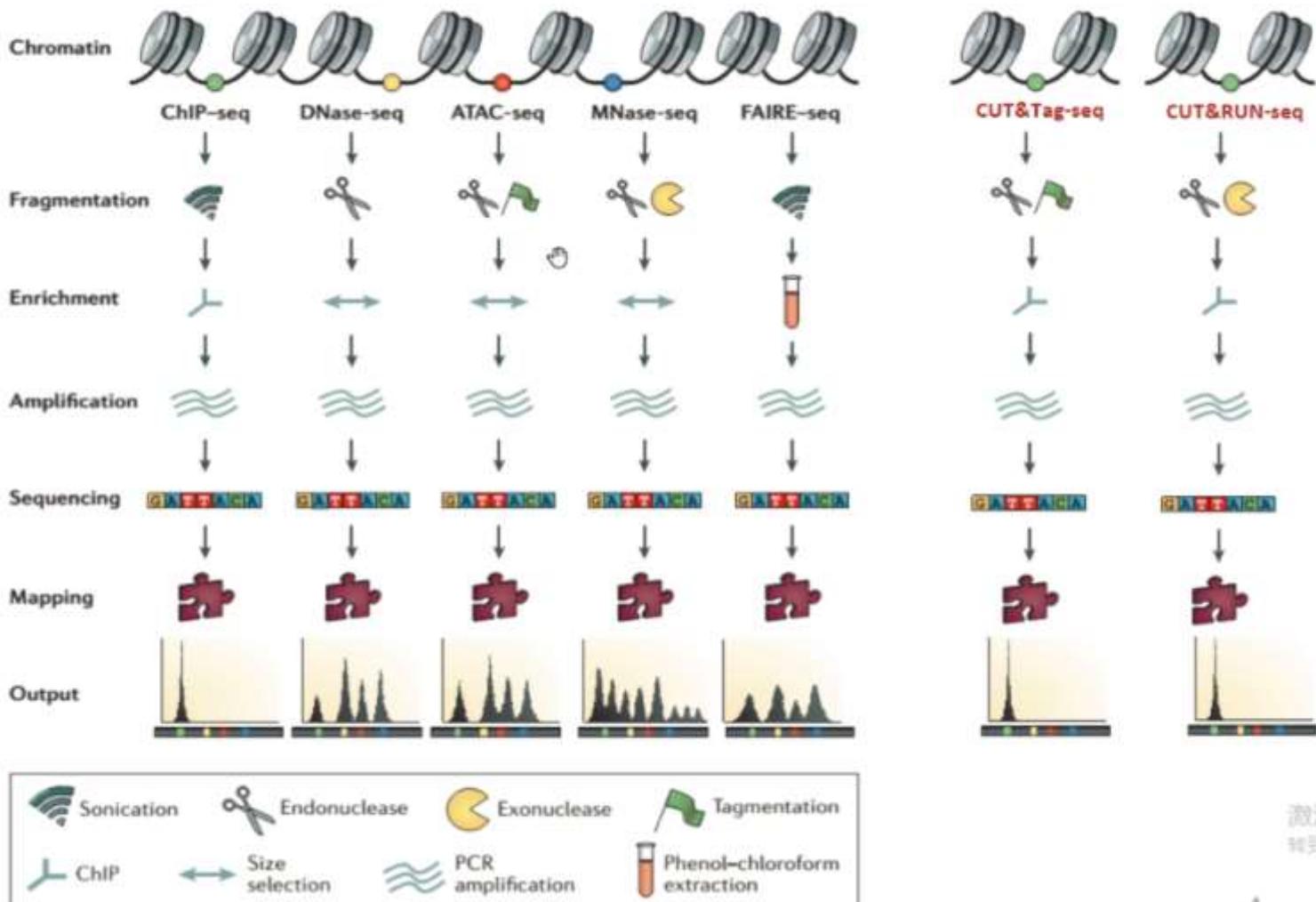
**Systematic comparison of IDH wild type GBM and neurodevelopmental trajectories**



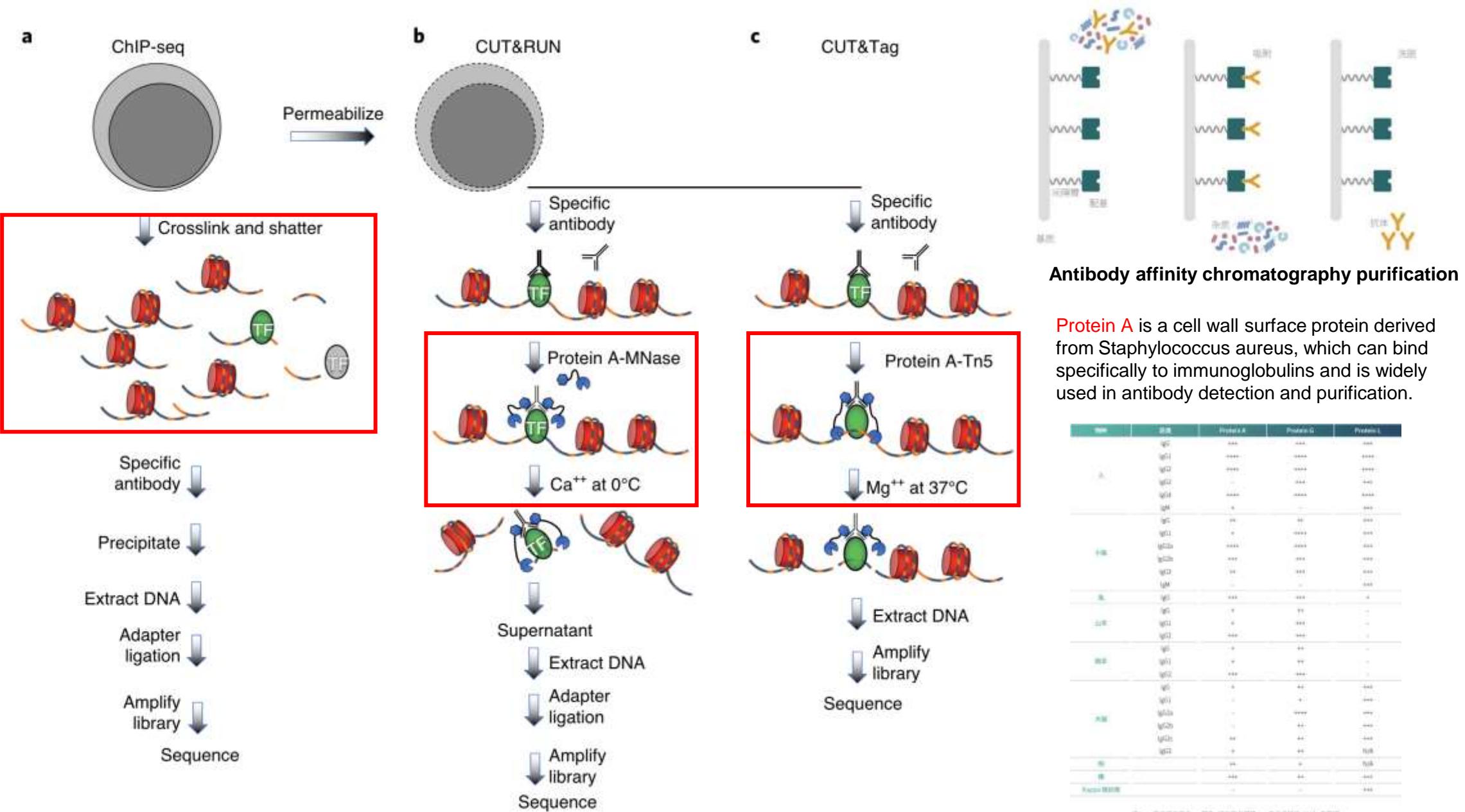
# **CHIP-Seq, CUT&RUN, CUT&TAG technologies**

Zhou Ying



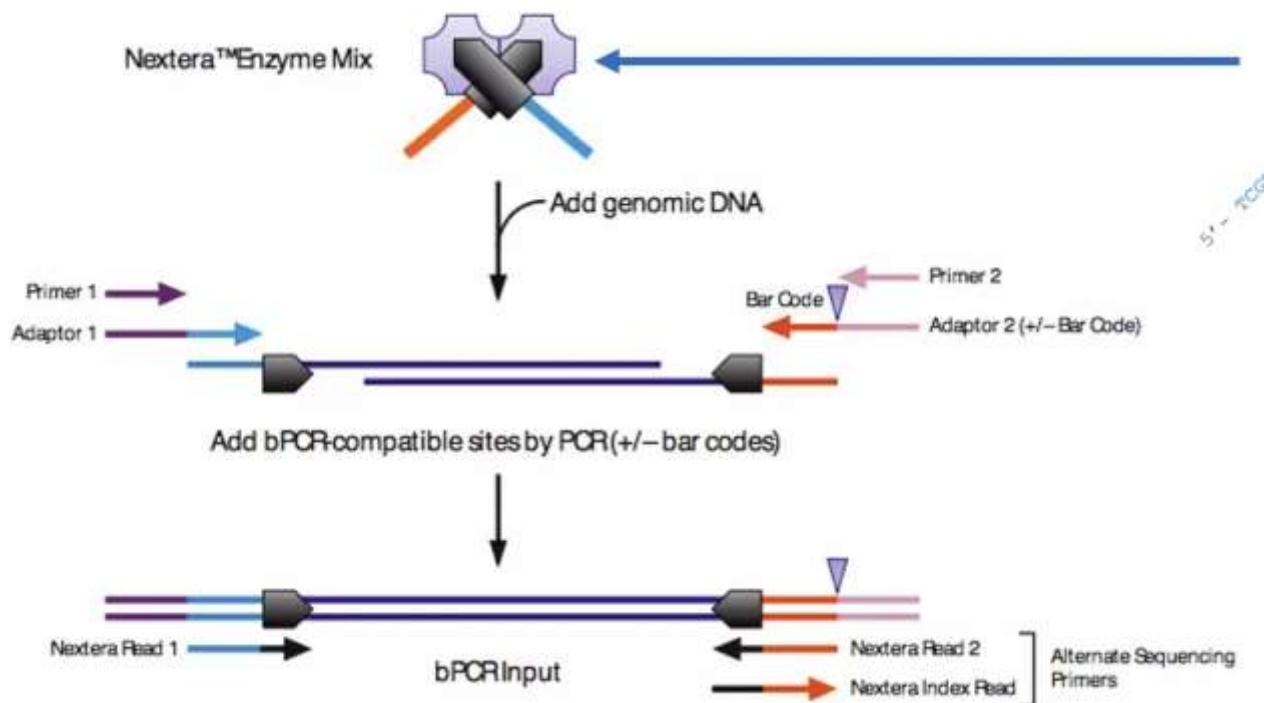


- Technical principles (differences, advantages and disadvantages of the three technologies)
- Wet test (main process)
- Bioinformatics analysis (currently available analytical pipeline)



ChIP-Seq	CUT&RUN	CUT&TAG
<ul style="list-style-type: none"> <li>❖ Both are required to break open areas of chromatin</li> <li>❖ DNA sequences bound to specific proteins are enriched by antibodies</li> </ul>		
<ul style="list-style-type: none"> <li>❖ The DNA is interrupted by ultrasound or by MNase (which has a certain sequence preference)</li> <li>❖ The antibody is hung in beads to precipitate the DNA complex of the target protein</li> </ul>	<ul style="list-style-type: none"> <li>❖ The fusion protein and antibody are constructed by protein A and <b>MNase</b> to form a complex with enzyme digestion function directly into the cell interior, and the target protein DNA complex is cut by Ca<sup>2+</sup></li> <li>❖ The target fragment can be obtained in supernatant without precipitation</li> </ul>	<ul style="list-style-type: none"> <li>❖ The fusion protein and antibody are constructed by protein A and <b>Tn5 transposition enzymes</b> to form a complex with enzyme digestion function directly into the cell interior, and the target protein DNA complex is cut by Mg<sup>2+</sup></li> <li>❖ The target fragment can be obtained in supernatant without precipitation</li> </ul>

# Tn5转座酶和MNase的不同



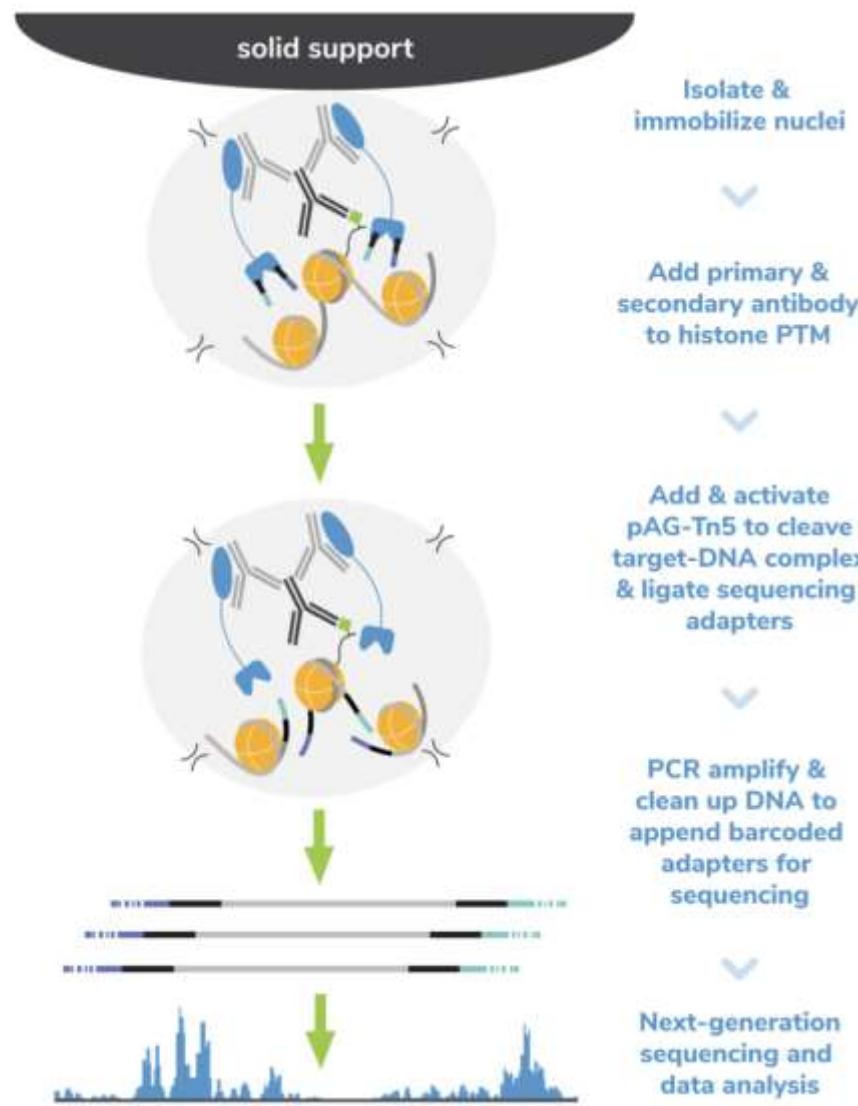
[Buenrostro et al., 2015](#)

- **Tn5:** Tn5, as a type of DNA transposase, can carry DNA fragments to simultaneously cut the recipient DNA and insert the carried donor DNA. Therefore, it is used in NGS library construction and ATAC-Seq studies of chromatin open regions.
- When combined with pA to construct pA-Tn5, it forms a complex with antibodies, entering cells to excise specific protein-bound DNA while tagging DNA ends with sequencing adapters such as P5/P7 or barcodes. This can be used for the purification and library construction of target DNA fragments.
- **The application of Tn5 streamlines library construction processes, enabling automation, further enhancing sequencing throughput and resolution, allowing for the study of transcriptional regulation at the single-cell level.**
- **After inserting the carried fragments into the target position, Tn5 becomes inactive, unlike MNase, which requires strict control of incubation time (MNase exhibits enzymatic activity continuously in the presence of Ca<sup>2+</sup>).**

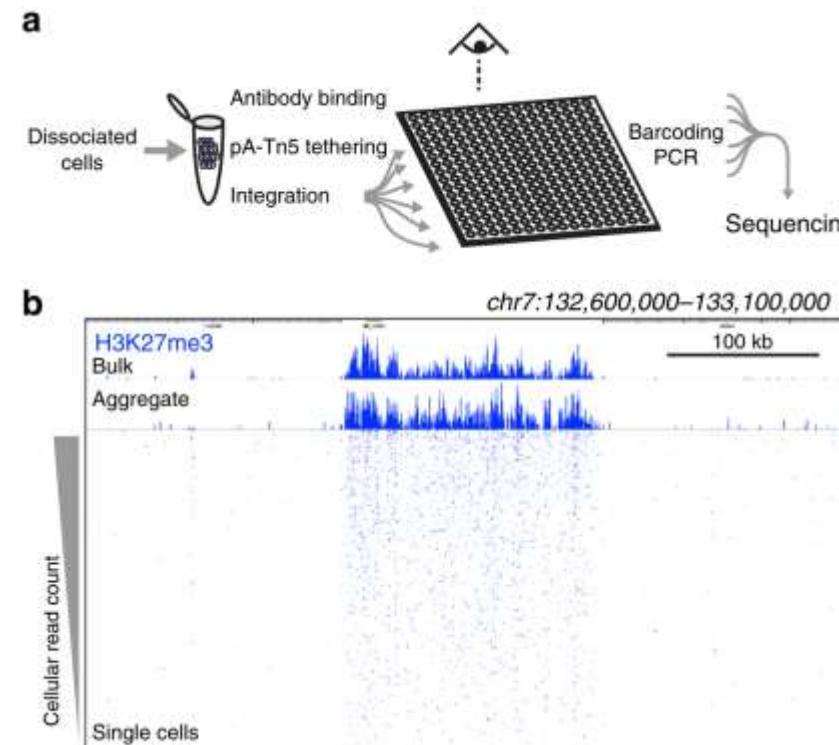
## Comparison of Limitations between CUT&TAG and CUT&RUN

- The complex formed by Tn5 occupies more space than MNase, affecting its resolution.
- Tn5 enzyme efficiency requires strict washing; this technique lacks fixed proteins, causing many transcription factors with weak DNA binding to be washed away. Therefore, it is more suitable for applications related to histone modification studies.
- CUT&RUN has been applied in various scenarios, **including automated high-throughput epigenetic analysis (AutoCUT&RUN) (Janssens et al., 2018), analysis of insoluble chromatin (such as using CUT&RUN.Salt (Thakur and Henikoff, 2018) and CUT&RUN.ChIP for analysis of centromeric regions), and examination of specific protein components released during CUT&RUN digestion (Brahma and Henikoff, 2019).** However, there is relatively little application of CUT&TAG.
- Most bioinformatics analysis tools are designed for ChIP-seq and CUT&RUN, although theoretically applicable to CUT&TAG, no articles systematically compare and evaluate the reliability of results.

## CUT&TAG的control设置

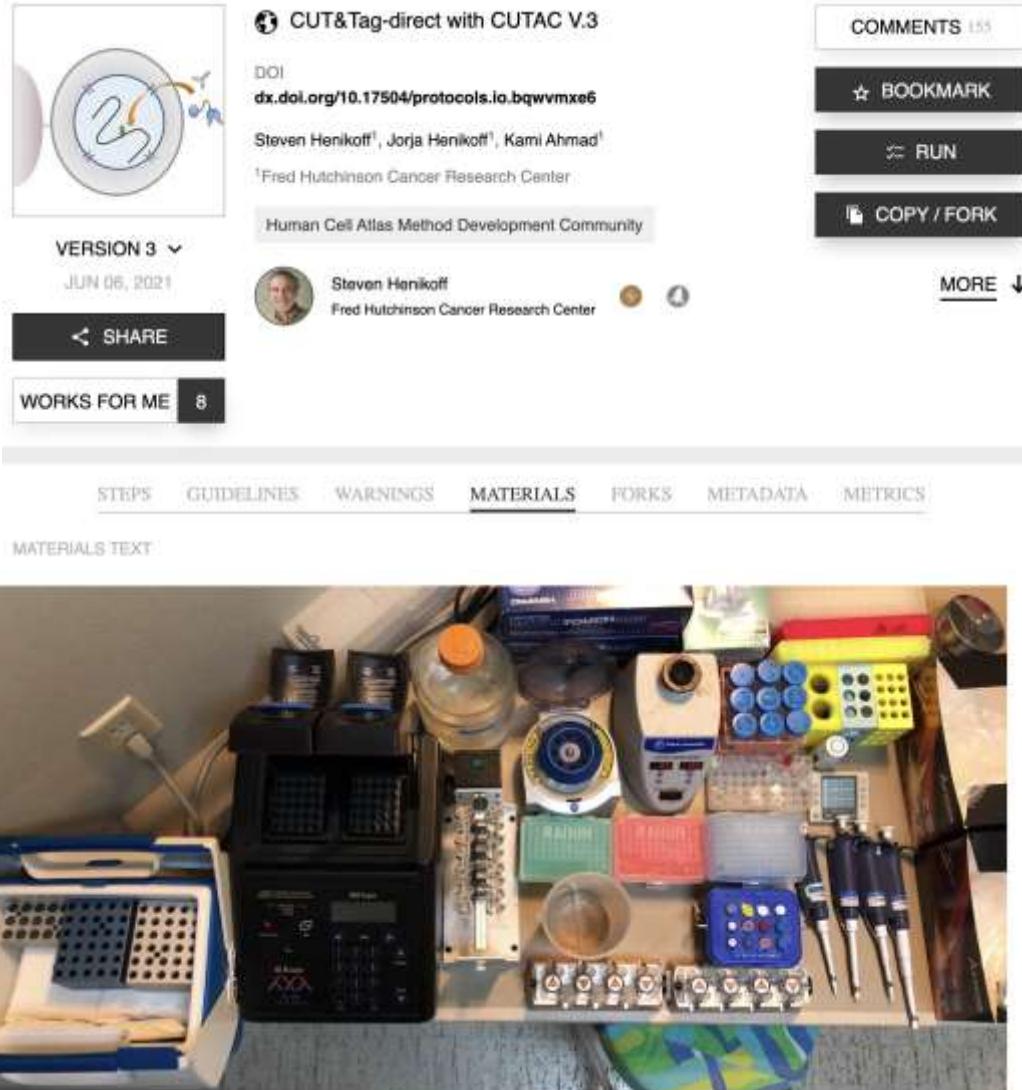


- ❖ Negative control setting is similar to the input setting in ChIP-Seq (without antibody/IgG).
- ❖ However, due to the presence of two antibodies in CUT&TAG, similar to immunohistochemistry, the primary antibody possesses specific recognition ability. Therefore, in the negative control, only the secondary antibody is added without the primary antibody.
- ❖ The secondary antibody binds to Tn5, and when only the secondary antibody is added, Tn5 will randomly bind to open chromatin regions. Comparing with ATAC-seq, it can be found that it is actually a set of ATAC-seq data.
- ❖ Currently, scATAC-seq has become quite mature. When CUT&TAG is maturely applied to single cells, it is expected to obtain data of both scATAC-seq and scCUT&RUN at the single-cell level.



<https://www.youtube.com/watch?v=4gNXcOOmNjk>

<https://www.protocols.io/view/cut-and-tag-direct-with-cutac-x54v9mkmgz3e/v3/materials>



CUT&Tag-direct with CUTAC V.3

DOI: dx.doi.org/10.17504/protocols.io.bqvvvmxe6

Steven Henikoff<sup>1</sup>, Jorja Henikoff<sup>1</sup>, Kami Ahmad<sup>1</sup>

<sup>1</sup>Fred Hutchinson Cancer Research Center

Human Cell Atlas Method Development Community

VERSION 3 ✓

JUN 06, 2021

SHARE

WORKS FOR ME 8

STEPS GUIDELINES WARNINGS MATERIALS FORKS METADATA METRICS

MATERIALS TEXT

A home workbench for CUT&Tag. Photo of the home workbench setup used for all experiments presented using this protocol. A typical experiment begins by mixing cells with activated ConA beads in 32 single PCR tubes, with all liquid changes performed on the magnet stands. The only tube transfer is the removal of the purified sequencing-ready libraries from the SPRI beads to fresh tubes for Tapestation analysis and DNA sequencing. The total time from thawing frozen nuclei until elution from SPRI beads is ~8 hr.



nature  
protocols

PROTOCOL

<https://doi.org/10.1038/s41596-020-0373-x>

Check for updates

## Efficient low-cost chromatin profiling with CUT&Tag

Hatice S. Kaya-Okur<sup>1,2,3</sup>, Derek H. Janssens<sup>1</sup>, Jorja G. Henikoff<sup>1</sup>, Kami Ahmad<sup>1</sup> and Steven Henikoff<sup>1,2,✉</sup>

protocolia PUBLISHED PLATEAU CASE STUDY

### Bring structure to your research

A secure platform for developing and sharing reproducible methods.



Search

Recent searches

Advanced search

Protocol categories

Safety checklist

Metrics

Organize & collaborate

Accelerate research



jove

Accelerate your science research and education

15,000+ videos of laboratory methods and science concepts

Search 16,011 videos

See what's new this week

**A home workbench for CUT&Tag.** Photo of the home workbench setup used for all experiments presented using this protocol. A typical experiment begins by mixing cells with activated ConA beads in 32 single PCR tubes, with all liquid changes performed on the magnet stands. The only tube transfer is the removal of the purified sequencing-ready libraries from the SPRI beads to fresh tubes for Tapestation analysis and DNA sequencing. The total time from thawing frozen nuclei until elution from SPRI beads is ~8 hr.

Platform Comparison	ChIP-seq	CUTANA™ CUT&RUN	CUTANA™ CUT&Tag
Required cells	> 1 million	5,000 – 500,000	1,000 – 100,000
Cell Input	Fragmented chromatin	Intact cells or nuclei	Nuclei (preferred) or intact cells
Compatible Targets	Histone PTMs & chromatin-associated proteins	Histone PTMs & chromatin-associated proteins, including difficult ChIP targets	Histone PTMs only
Sequencing Depth (Reads)	> 30 million	3-8 million	3-8 million
Experimental throughput	Low	High	High
Signal : Noise	Low	High	High
Assay Automation	Difficult	Yes	Yes

## EpiCypher® CUTANA™ Direct-to-PCR CUT&Tag Protocol

Optimized for Histone Post-Translational Modifications (PTMs)

This protocol has been validated for genomic profiling of:

- Histone PTMs (e.g. lysine methylation, acetylation)
- Low cell inputs: Starting with 100,000 cells - down to as low as 1,000 cells
- For profiling chromatin-associated proteins (e.g. transcription factors, epigenetic enzymes, etc.) or for applications that do not require ultra-low cell inputs, see [EpiCypher CUTANA™ CUT&RUN assays](#).

## EpiCypher® CUTANA™ CUT&RUN Protocol

For histone PTMs, transcription factors (TFs), and chromatin regulators

This protocol has been validated for genomic profiling of:

- Histone PTMs (e.g. lysine methylation, acetylation and ubiquitylation)
- Transcription factors (e.g. CTCF, FOXA1)
- Chromatin remodelers (e.g. ATPase subunits of SWI/SNF, ISWI, INO80, CHD)
- Chromatin writers & readers (e.g. MLL1, BRD4)
- Nuclear hormone receptor signaling factors (e.g. Estrogen Receptor)
- Epitope-tagged proteins (e.g. HA, FLAG tags)

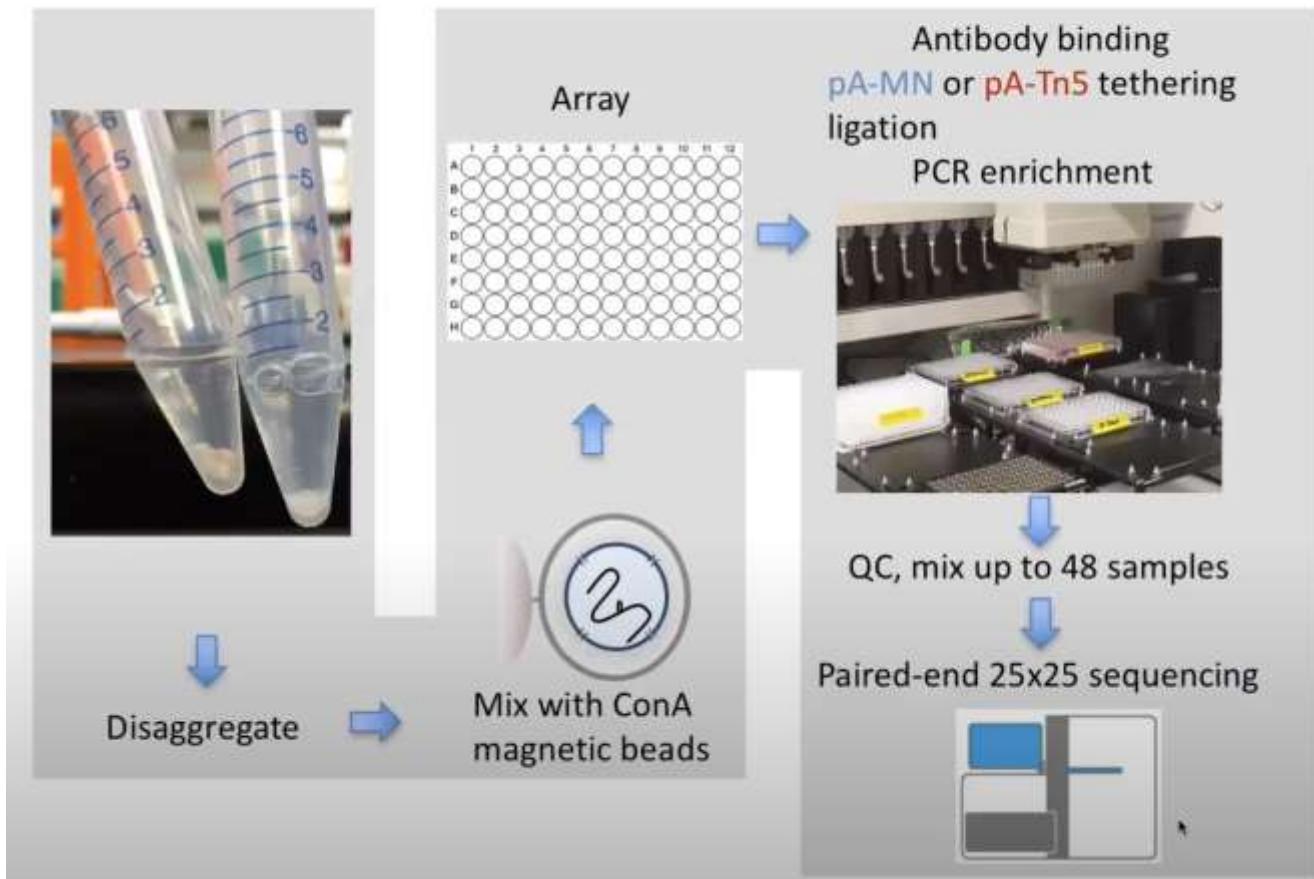
### Table of Contents

1. Overview .....	1
2. CUTANA™ Products & Services: Advantages .....	3
3. Outline of CUT&Tag Workflow .....	4
4. Experimental Design & Key Protocol Notes .....	8
5. Buffers, Reagents & Materials Needed .....	10
6. EpiCypher CUTANA™ Direct-to-PCR CUT&Tag Protocol.....	15
Section I: ConA Bead Activation (~30 min).....	16
Section II: Nuclei Preparation and Binding Nuclei to Activated Beads (~30 min).....	16
Section III: Binding of Primary and Secondary Antibodies (~30 min + overnight + 1 hr).....	17
Section IV: Binding of pAG-Tn5 (~1 hr).....	19
Section V: Targeted Chromatin Tagmentation (~3 hrs).....	19
Section VI: Non-hot Start PCR and Library Cleanup (~1 hr) .....	20
Section VII: Agilent 2100 Bioanalyzer® System (~1 hr) .....	21
7. Frequently Asked Questions (FAQs) .....	22
8. References .....	29
Appendix I: Cryopreservation and Thawing Protocols for CUT&Tag.....	30
Appendix II: Sample Integrity & Bead Conjugation Quality Control Checks .....	31

### Table of Contents

1. Overview .....	2
2. CUTANA™ Products & Services: Advantages .....	3
3. Outline of CUT&RUN Workflow .....	4
4. Experimental Design & Key Protocol Notes .....	7
5. Buffers, Reagents & Materials Needed .....	9
6. EpiCypher CUTANA™ CUT&RUN Protocol.....	13
Section I: ConA Bead Activation (~30 min).....	13
Section II: Binding Cells to Activated Beads (~30 min).....	13
Section III: Binding of Antibodies (~30 min + overnight).....	14
Section IV: Binding of pAG-MNase (~30 min) .....	16
Section V: Targeted Chromatin Digestion and Release (~3 hrs).....	16
Section VI: Library Preparation (~4 hrs) .....	17
Section VII: Agilent TapeStation® System (~1 hr).....	18
Section VIII: Illumina® Sequencing .....	18
7. Frequently Asked Questions (FAQs) .....	20
8. References .....	27
Appendix I: Nuclei Isolation & Cryopreservation Protocol for CUT&RUN.....	28
Appendix II: Sample Integrity & Bead Conjugation Quality Control Checks .....	30
Appendix III: Use of <i>E. coli</i> Spike-in DNA for Experimental Normalization .....	32

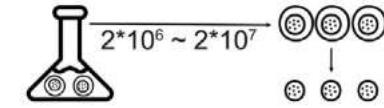
# AutoCUT&RUN and AutoCUT&Tag



## 1 - Wet-Lab

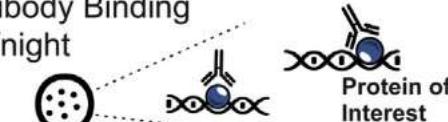
### Step 1 - Nuclei Purification

⌚ 1.5h



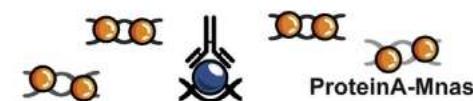
### Step 2 - Antibody Binding

⌚ 1.5h - o/night



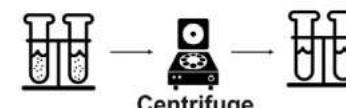
### Step 3 - pAMN binding & Mnase digestion

⌚ 2h



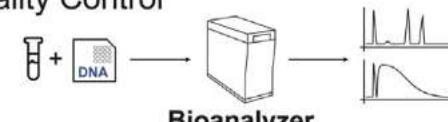
### Step 4 - DNA Purification

⌚ 2.5h - o/night



### Step 5 - Quality Control

⌚ 1h



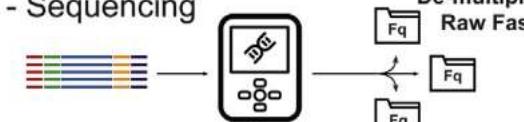
### Step 6 - Library Construction

⌚ 3h



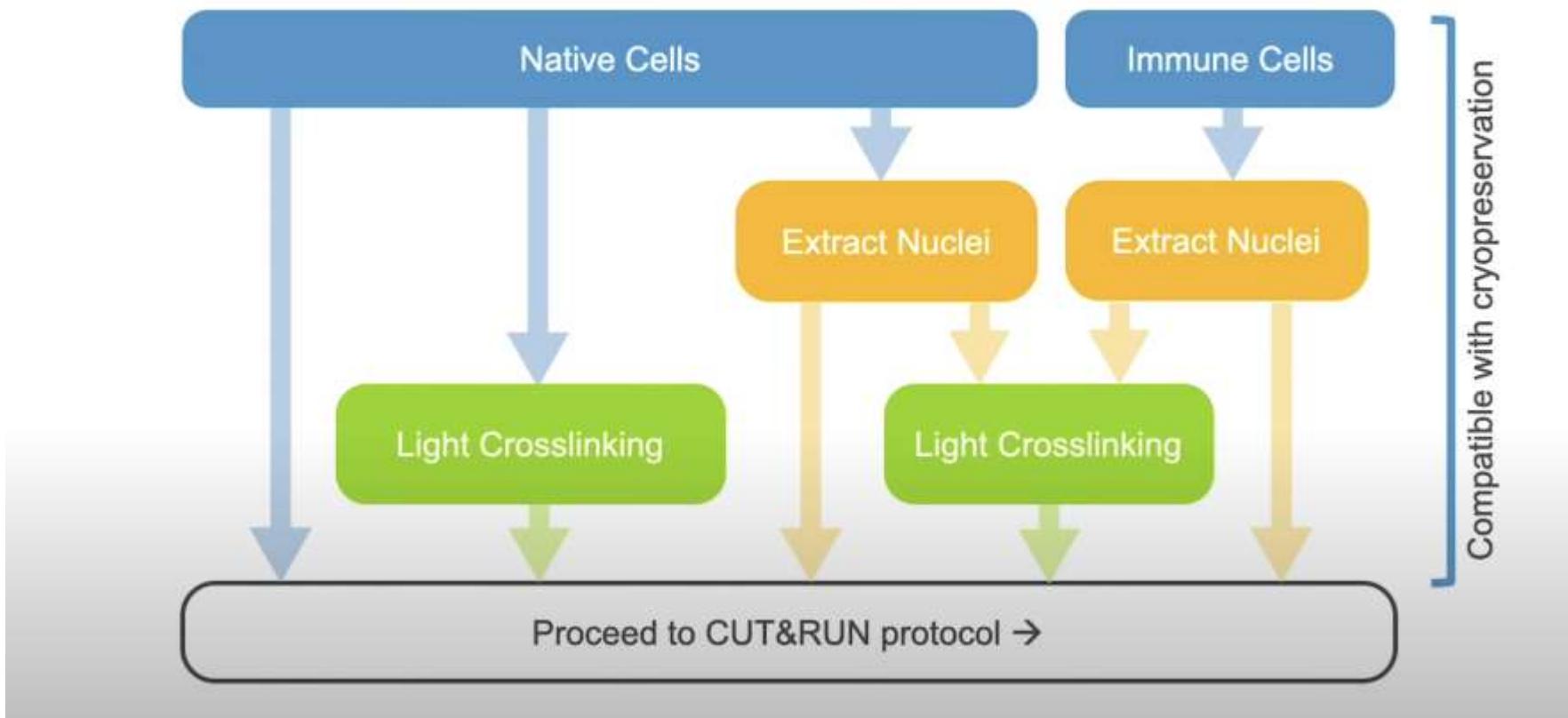
### Step 7 - Sequencing

⌚ 4h



# CUT&RUN

# Cell Types, Targets and Conditions



- ❖ Both cells and cell nuclei are applicable.
  - ❖ Live cells, mildly cross-linked cells, or cell nuclei are all acceptable.
  - ❖ For immune cells, it is recommended to use cell nuclei for experiments, as ConA can activate some immune cells.
  - ❖ In a video on basepair's YouTube channel, it was mentioned that frozen cells can also be used (without literature or data support).

## Section I – Materials Required

Items	Volume (8x)	Usage
ConA Beads	88 µL	On ice, 11µL per sample
Bead Activation Buffer	2.64 mL	On ice, 100 µL per sample

1.5 mL Tube  
Magnetic Rack



ConcanavalinA Beads

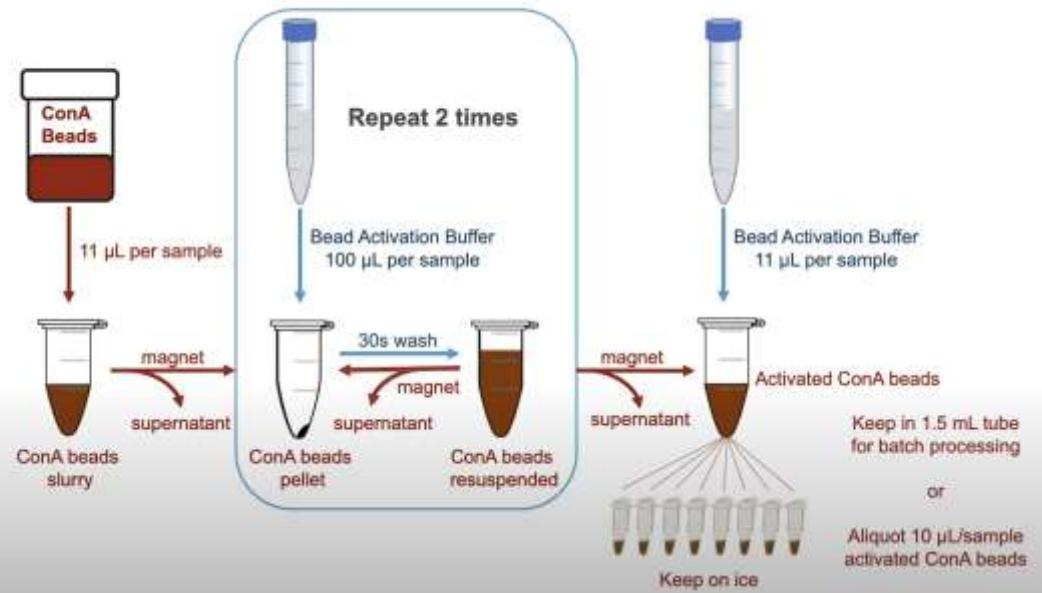


Bead Activation Buffer



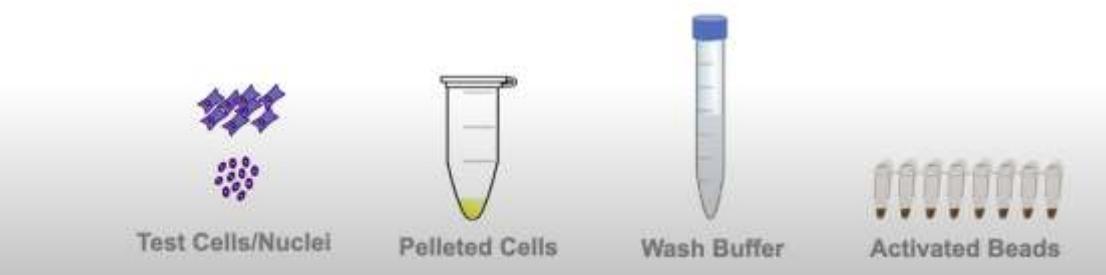
1.5 mL Magnetic Rack

## Section I

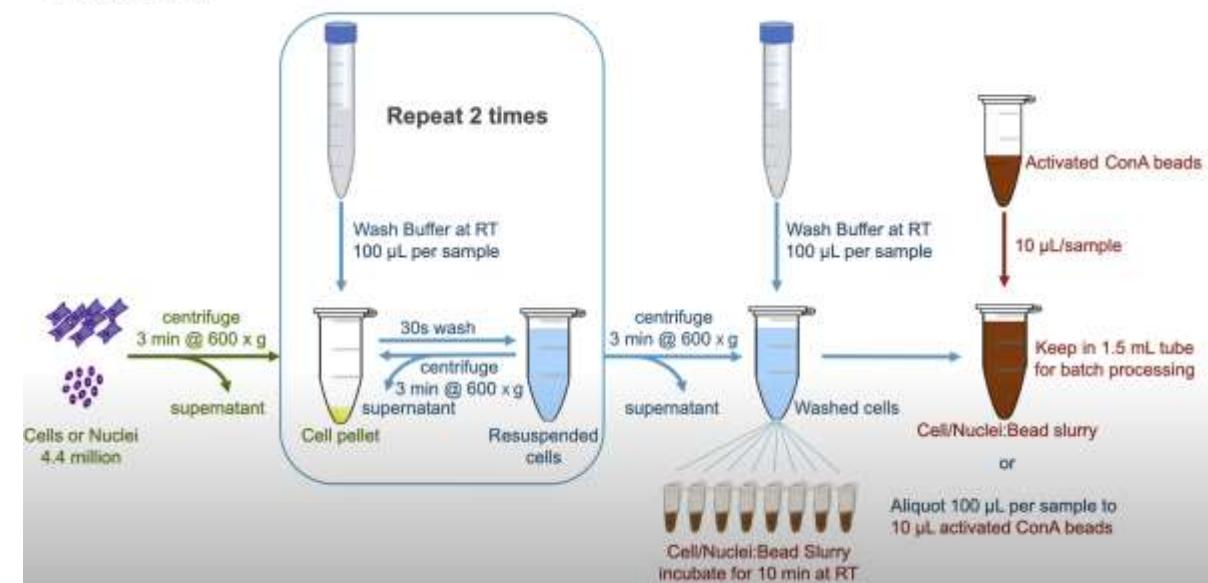


## Section II - Materials Required

Items	Volume (8x)	Usage
Cells or Nuclei	4.4 million	Room temp, ~500,000 cells per sample
Wash Buffer	3.20 mL	Room temp, 100 µL per sample
Activated Beads	88 µL or 8 strip - 10 µL/tube	On ice On ice, 10 µL per sample
1.5 mL Tubes		



## Section II



### Section III - Materials Required (Day 1)

Items	Volume (8x)	Usage
Cell/Nuclei:Bead Slurry	110 $\mu$ L per sample	Room temp
Antibody Buffer	440 $\mu$ L	On ice, 50 $\mu$ L per sample
Antibody for Target	0.5 $\mu$ L per sample	On ice, 0.5 $\mu$ L per sample
Magnetic Rack		
4°C Nutator		



Antibody Buffer



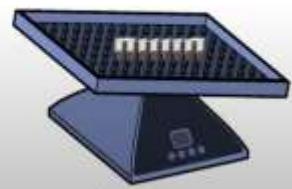
Antibodies



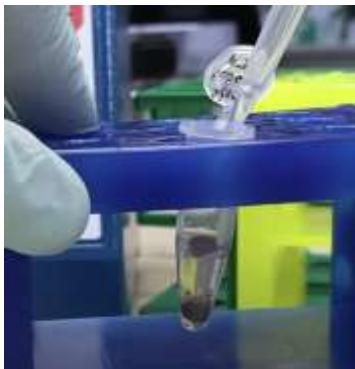
Cell/Nuclei:Bead Slurry



0.2 mL Magnetic Rack



Nutator

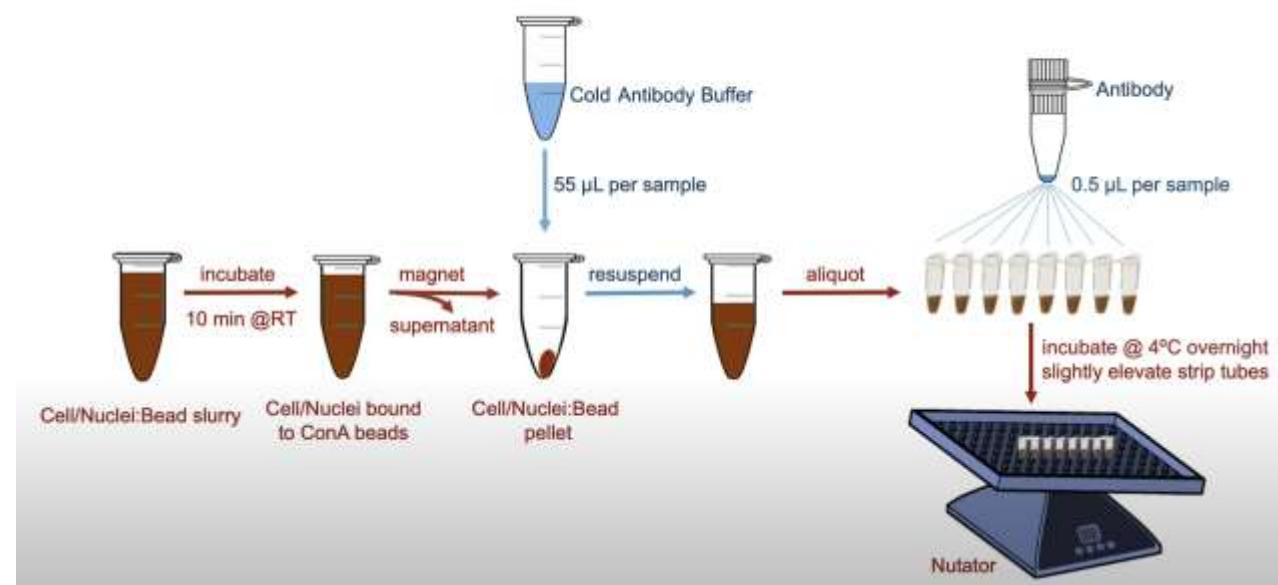


预冷的antibody buffer  
迅速重悬，避免磁珠  
变干



分装到八连排或者PCR小管

### Section III



加入抗体



涡旋振荡器混匀



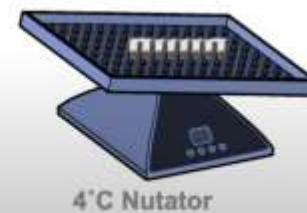
4度缓慢摇晃孵育过夜

### Section III - Materials Required (Day 2)

Items	Volume (8x)	Usage
Cell/Nuclei:Bead Slurry	110 $\mu$ L per sample	From 4°C Nutator
Digitonin Buffer	~5.5 mL	On ice, 300 $\mu$ L total per sample

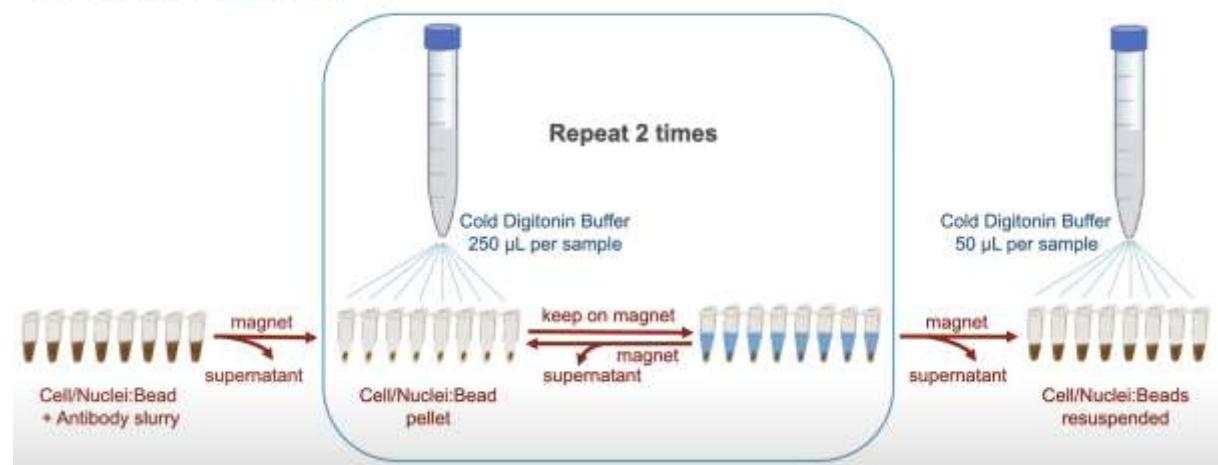
Magnetic Rack

4°C Nutator



### Section III (Day 2)

## Digitonin 洋地黃皂苷 (通透细胞膜)



## Section IV - Materials Required

Items	Volume (8x)	Usage
Cell/Nuclei:Bead Slurry		Room temp
CUTANA™ pAG-MNase (20x stock)	~20 µL	On ice, 2.5 µL per sample
Digitonin Buffer	~5.5 mL	On ice, 300 µL total per sample

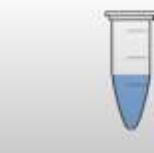
Magnetic Rack



Cell/Nuclei:Bead Slurry



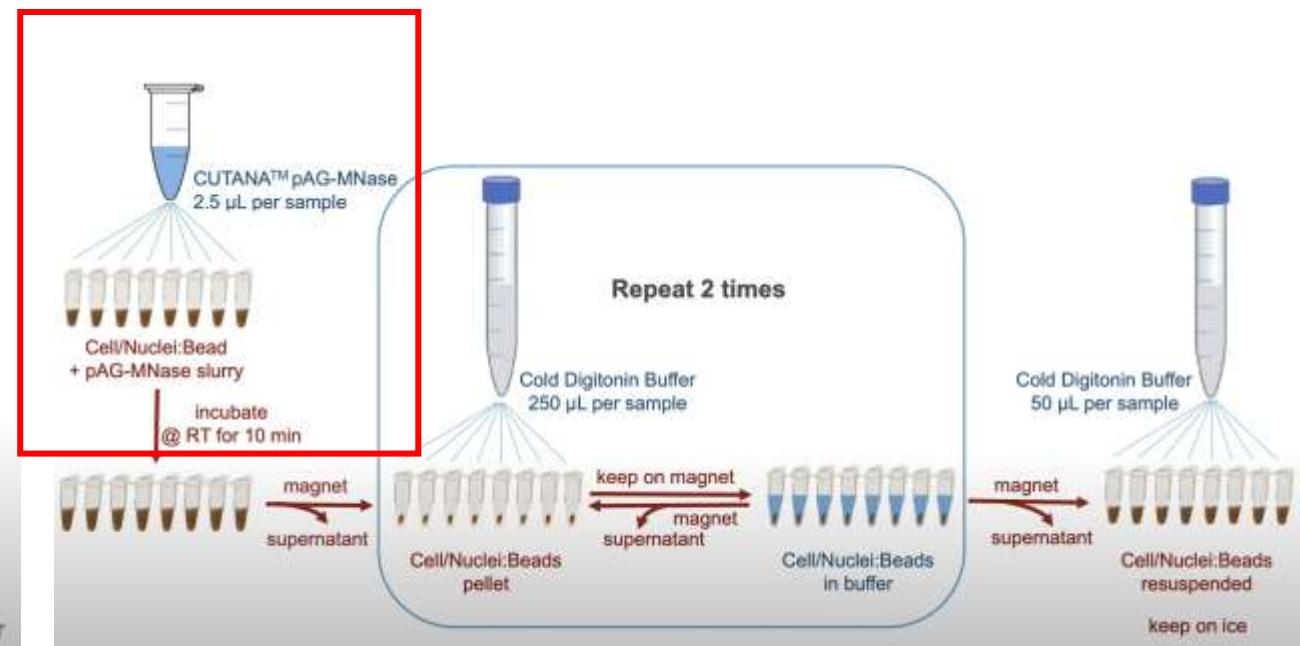
0.2 mL Magnetic Rack



CUTANA™ pAG-MNase  
(20x stock)



Digitonin Buffer



## Section V - Materials Required

Items	Volume (8x)	Usage
Cell/Nuclei:Bead Slurry		Room temp
100 mM CaCl <sub>2</sub>	8 µL	Room temp, 1 µL per sample
Stop Buffer	264 µL	On ice, 33 µL per sample

Spike-in DNA  
(recommended)

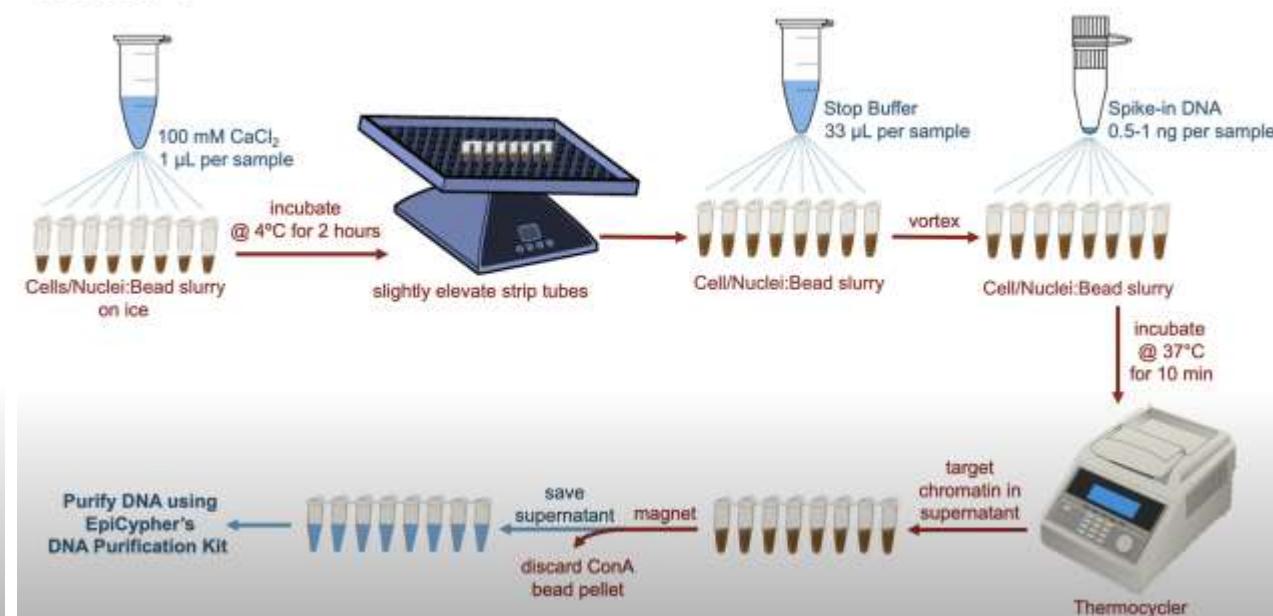
Magnetic Rack

37°C Thermocycler

DNA Purification Kit



## Section V



冰上加入CaCl<sub>2</sub>(重要步骤)

收集上清液到新管

## Section VI, VII, VIII – Elute DNA, Prepare Library, Analyze and Sequence

Purify DNA using EpiCypher's DNA Purification Kit and elute in 10 µL Elution Buffer



Quantify CUT&RUN-enriched DNA (1 µL) using Qubit™ fluorometer



Using 5-10 ng of purified CUT&RUN-enriched DNA, prepare library using NEBNext® Ultra™ II Library Prep Kit for Illumina.

**DO NOT** assess fragment size distribution prior to library preparation



PCR cycle parameters:

45 sec @ 98°C, 15 sec @ 98°C, 10 sec @ 60°C, [Repeat 14x], 1 min @ 72°C  
DNA cleanup using 1.1x AMPure beads to PCR sample volume

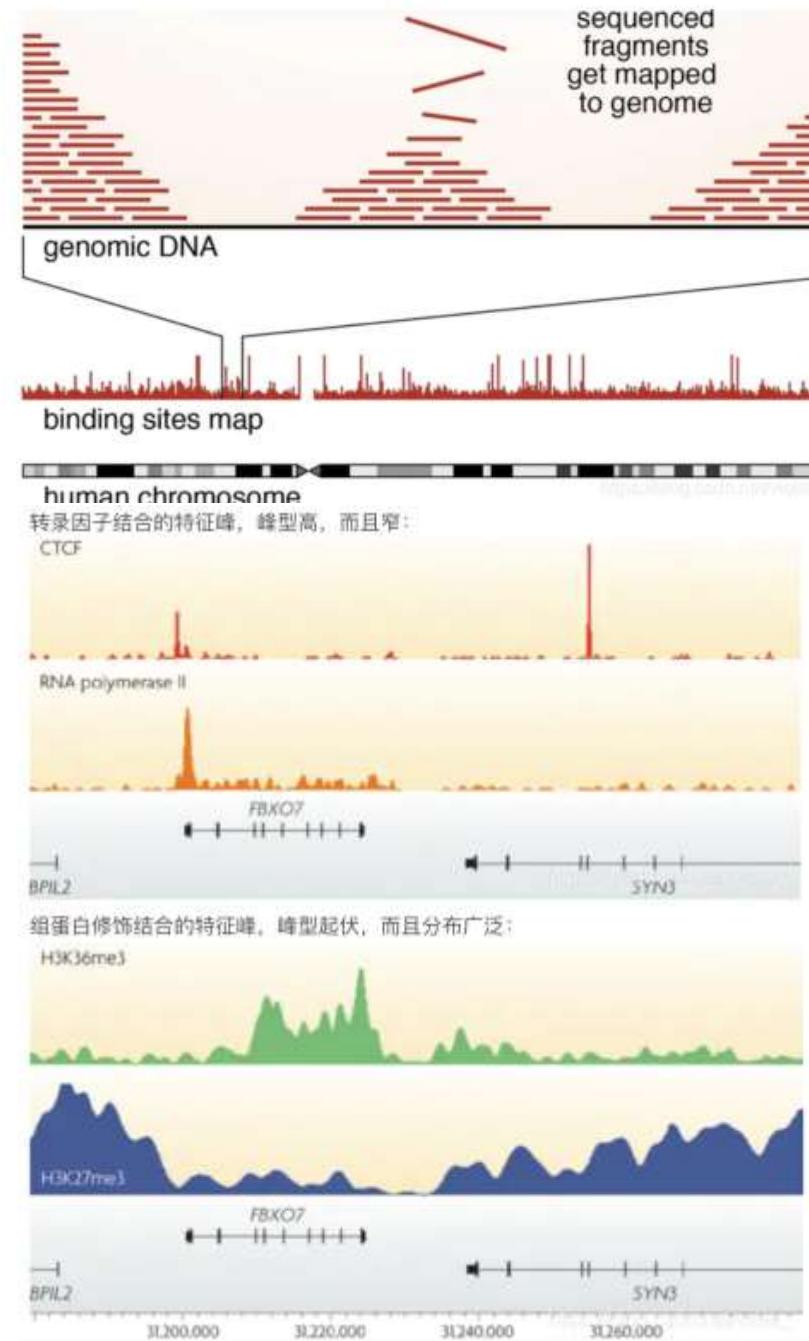
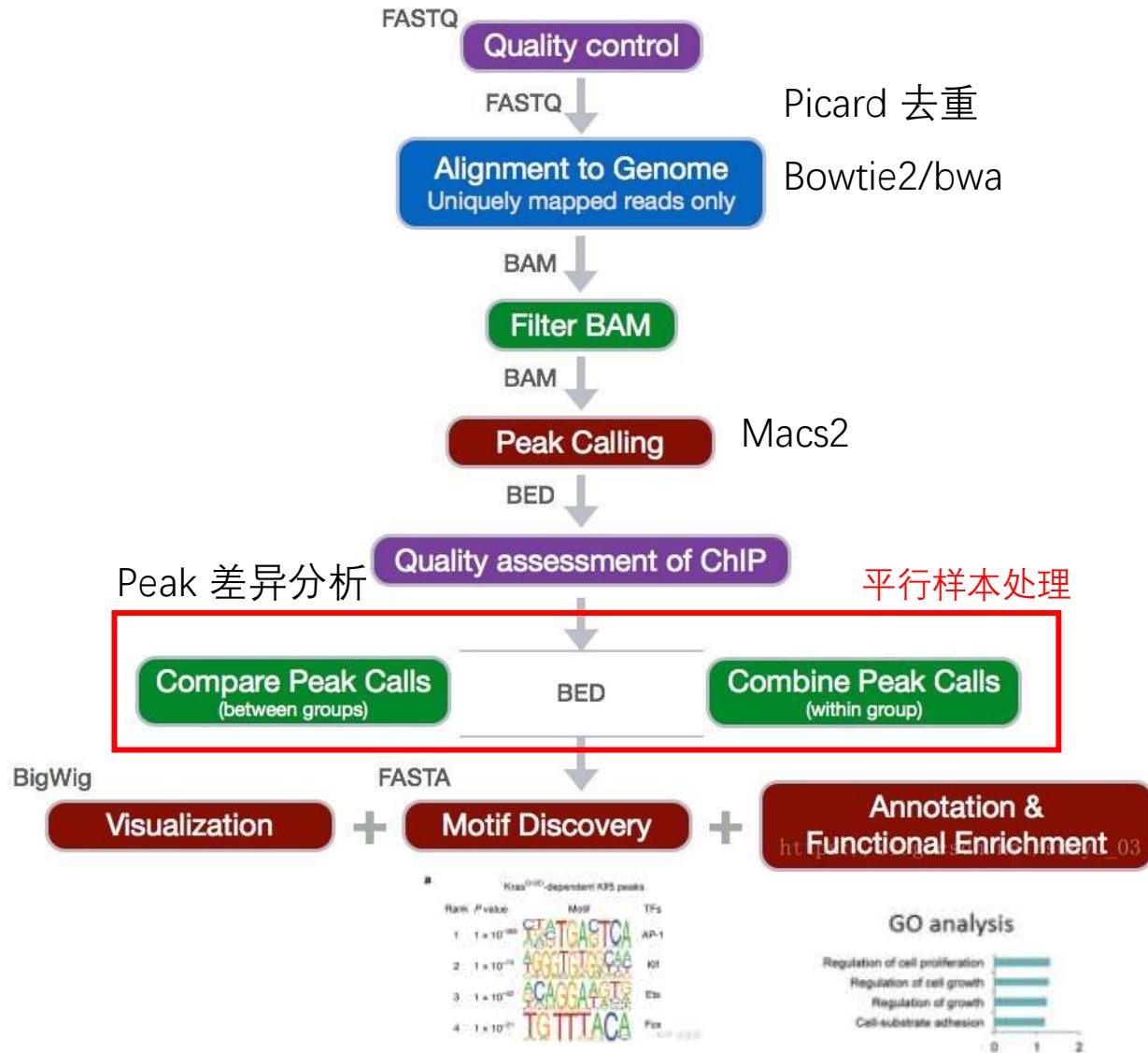


Load 1 µL of 10 ng/µL sample on Agilent High Sensitivity DNA Chip or similar instrument



Load 0.8 pM total for 8 samples (500 µL) into a cartridge for sequencing using MiniSeq High Output Reagent Kit, 150-cycles

# ChIP-seq general data analysis process



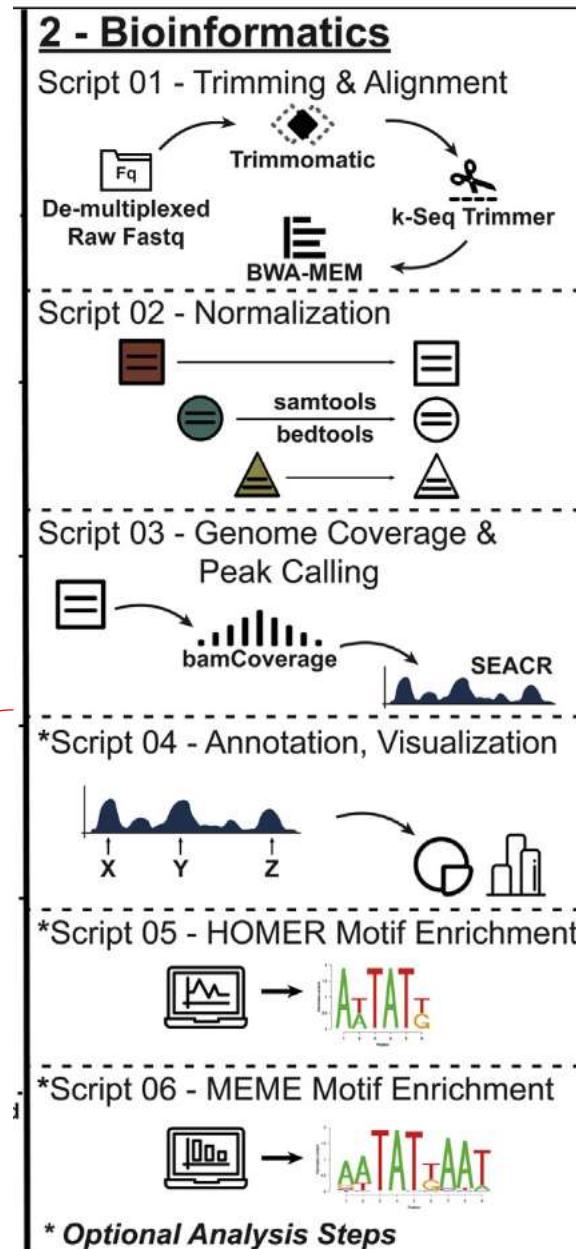
# CUT&RUN data analysis

Processing of duplicate samples

Compare tool applications for peak calling

Analysis strategy after obtaining the peak list:

- Analysis of Differential Peak Distribution among Samples
- Enrichment Analysis of Motifs/Peak
- Analysis of Peak Distribution and Correlation with Other Transcription Factors in TF Databases (Discovering Co-TFs)
- Integrated Analysis of RNAseq, ATAC-seq, Hi-C, and Other Data



01 – Read quality trimming and alignment to reference genomes

Similar to the ChIP-seq comparison (BWA/Bowtie2) either way (Picard tag repeated here)

**02 – Generation of bedgraph files and normalization (spike in)**

**03 – Peak calling using SEACR/Macs2/CUT&RUNTools, etc.**

04 – Annotate called peaks using ChIPSeeker

05 – File preparation for HOMER motif enrichment analysis

06 – File preparation for MEME motif enrichment analysis.

4-6 These downstream analyses are all generic. No more difference

**The analysis processes mainly differ in peak calling. For ChIP-Seq analysis, Mac2 is used to identify peaks considering high background, whereas CUT&RUN and CUT&TAG both exhibit low background signals, affecting their application in chromatin regulation. The identification of TF signals is not significantly different from ChIP-seq. Many bioinformatics analysis tools tailored for CUT technologies are geared towards histone modification or snCUT&TAG analysis**

## **Processing duplicate samples**

For ATAC-seq/ChIP-seq experiments, it is generally recommended to include at least one input control and two biological replicates. In theory, peaks identified in replicate samples should exhibit high consistency; however, this may not always be the case in practice. The ultimate goal of peak calling analysis is to obtain a confident peak list.

### **Common strategies include:**

1. Union: Retaining all peaks from replicate samples, which means retaining all sample noise.
2. Intersection: Retaining only peaks that are shared among replicate samples. This approach may be overly strict and filter out potentially genuine peaks.
3. SUM: Merging reads mapped to the same genomic positions.

### **The ENCODE standard pipeline employs IDR to handle replicates**

- ◆ Replicate peak calling results are subjected to the same sorting process
- ◆ the correlation between sorted lists is calculated until a certain rank shows no correlation.
- ◆ Cutoff values between replicates may vary.
- ◆ True peaks may be present in different replicates.

# ChIP-seq vs CUT&Tag

	ChIP-seq	CUT&Tag
Input data	PE and SE	Only PE
Adapter trimming	Recommended	Not recommended
Duplicate removal	Yes	No
Bowtie parameters		Filter on insert size
Mitochondrial DNA	Peaks on mitochondrial DNA are real	Recommended to remove mitochondrial reads
Peak calling	MACS2	MACS2 and SEACR

Basepair

- **CUT&RUN only supports paired-end sequencing.**

- **Adapters are not removed.**

It is used for subsequent filtering by insert size.

- **PCR duplicates are not removed.**

Enzymatic cleavage may occur at the same base position, unlike ChIP-seq which randomly breaks, resulting in reads appearing as duplicates.

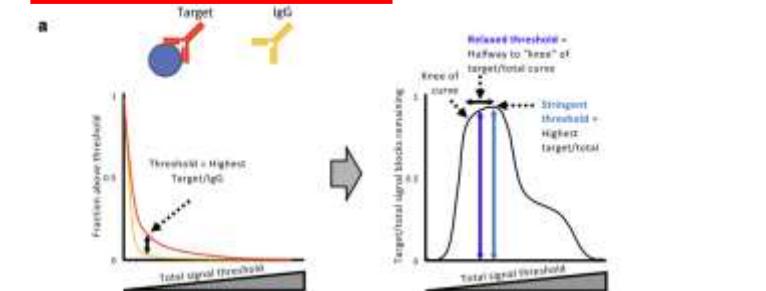
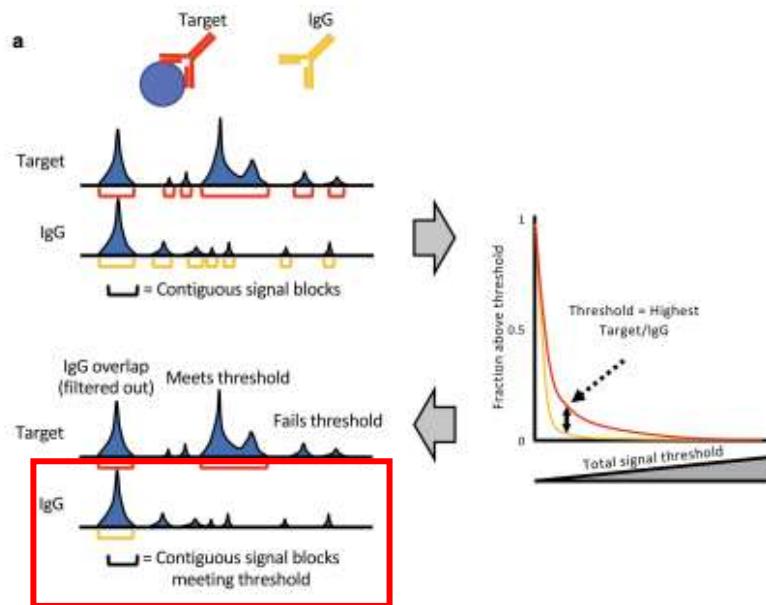
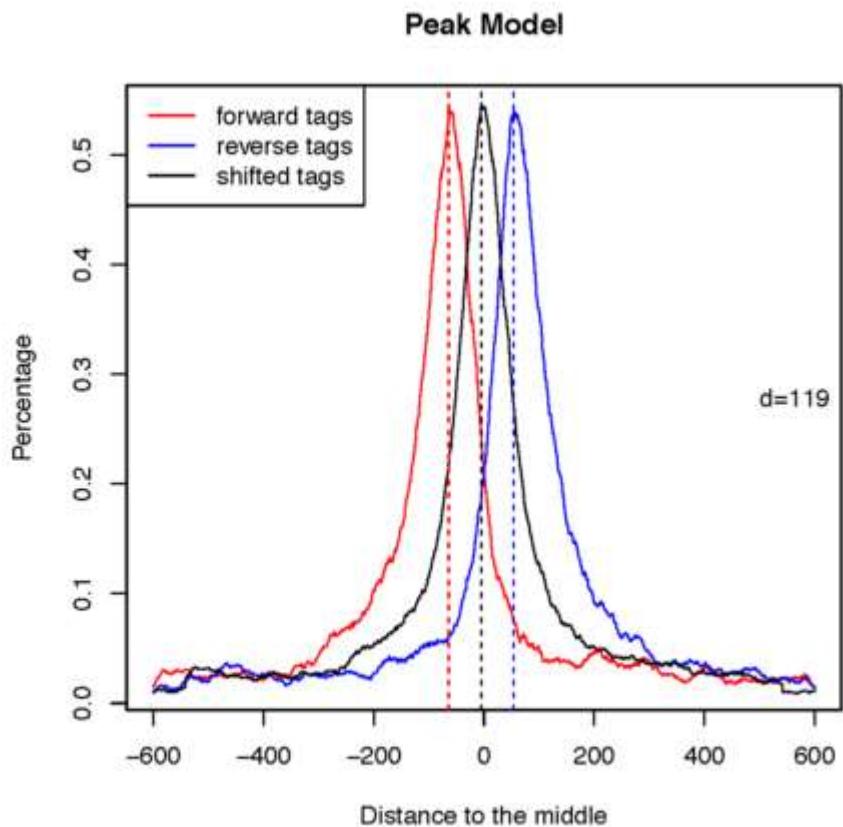
- **Parameters for Bowtie alignment should be filtered by insert size.**

This is primarily used for histone modification.

- **Mitochondrial reads are removed.**

30% of reads are from mitochondria (actual data needs to be examined).

## Macs2 model



**b**

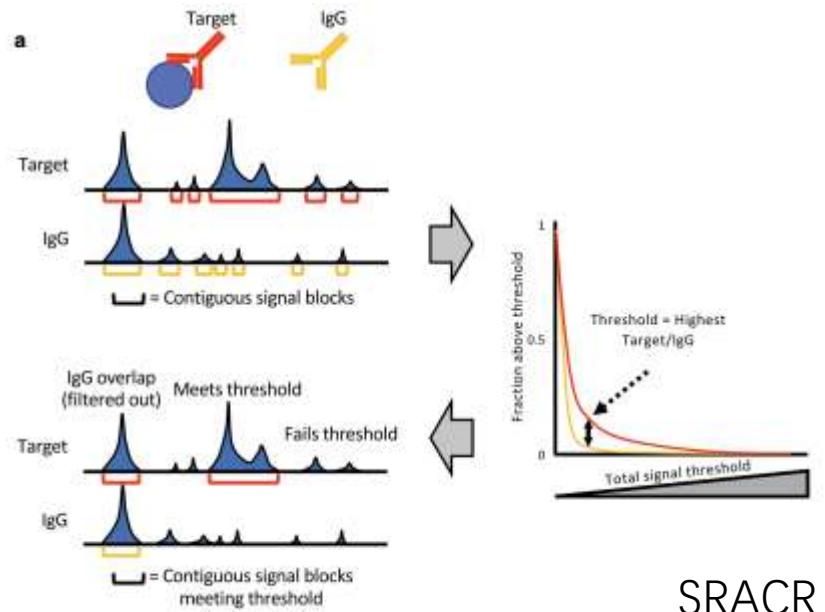
	# of Peaks			# of Peaks	
SEACR (stringent)	Sox2	FoxA2	MACS2	Sox2	FoxA2
hESC	17523	1	hESC	10728	858
Endoderm	2	6083	Endoderm	695	7386

SEACR (relaxed)	Sox2	FoxA2	HOMER	Sox2	FoxA2
hESC	49947	7	hESC	7095	67
Endoderm	9	11166	Endoderm	87	6493

**Fig. 2.14.1.**  
Shifting size model for FoxA1 ChIP-Seq data.

**Fig. 2** SEACR minimizes true negative detection from a gold standard dataset. **a** Schematic of SEACR threshold selection. A target/total signal block curve (light) is calculated, and the stringent and relaxed thresholds are set based on the knee and the peak of the curve. **b** Peaks were called from Sox2 or FoxA2 experiments carried out in either hESCs or Endoderm cells using SEACR in stringent or relaxed mode; MACS2; or HOMER.

ChIP-seq	CUT&RUN	CUT&TAG
<b>Macs2</b> <ul style="list-style-type: none"> <li>Designed according to ChIP-seq high-background signal, easy to operate, stable and reliable.</li> <li>It is easy to mistake noise for signal in low-background data.</li> </ul>	<b>SEACR (2019):</b> <ul style="list-style-type: none"> <li>Designing for CUT&amp;RUN Analysis</li> <li>Highly Dependent on IgG Negative Control</li> </ul> <b>CUT&amp;RUNTools 1.0 (2019)</b> <ul style="list-style-type: none"> <li>utilized integrated tools from Macs2</li> </ul> <b>CUT&amp;RUNTools 2.0 (2021)</b>	<b>GoPeaks (2022) :</b> <ul style="list-style-type: none"> <li>Designing Analysis for Histone Modification Data, Considering Adaptation to Broad and Narrow Peaks</li> </ul> <b>CUT&amp;RUNTools 2.0 (2021)</b> <ul style="list-style-type: none"> <li>It is suggested that both CUT&amp;RUN and CUT&amp;TAG are suitable for single cell analysis</li> </ul>



## CUT&RUN

- CnRAP pipeline:** using SRACR calling peaks
- CUT&RUNTools :** using Macs2

## CUT&TAG:

- Protocol.io:** CUT&Tag Data Processing and Analysis Tutorial (using SRACR)

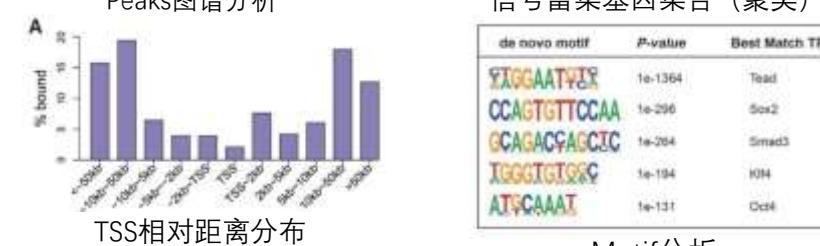
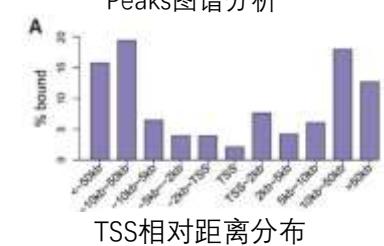
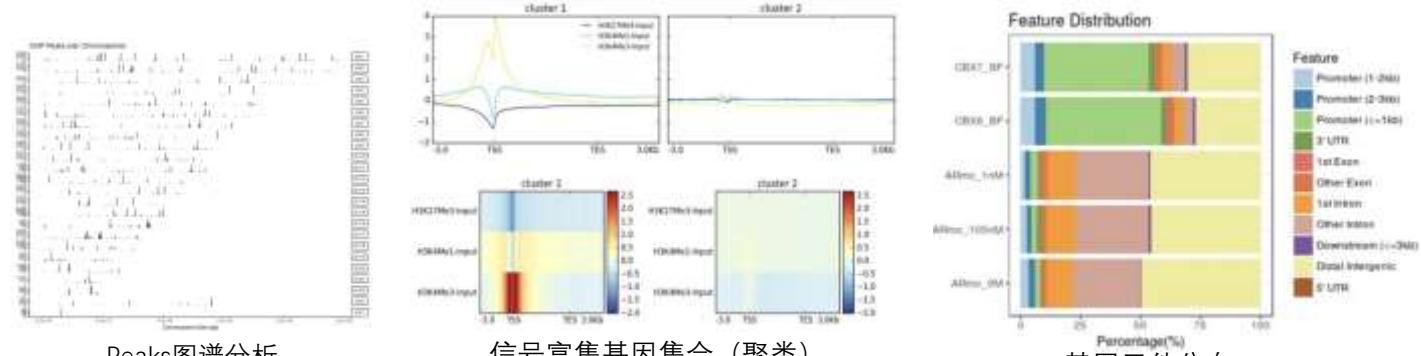
# After calling peaks

peaks name	染色体	起始坐标	终止位置	峰宽	峰对应的坐标	峰的高度	callPeak的置信度	峰相对于背景富集	callPeak的置信度
	chr	start	end	length	abs_summit	pileup	X.log10.pvalue.	fold_enrichment	X.log10.pvalue.
ZHX3_repl_peak_1	chr1	726930	727010	81	726968	101	29.17850	3.81709	23.86400
ZHX3_repl_peak_2	chr1	860235	860306	72	860275	11	8.12319	5.40025	4.34423
ZHX3_repl_peak_3	chr1	994166	994490	325	994225	14	11.30620	6.75032	7.28793

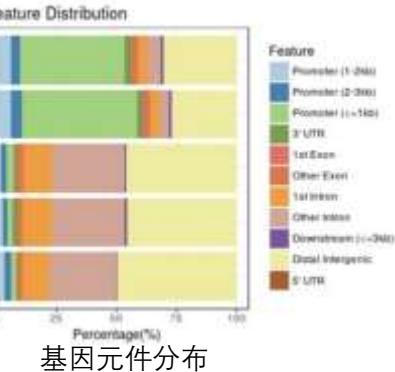
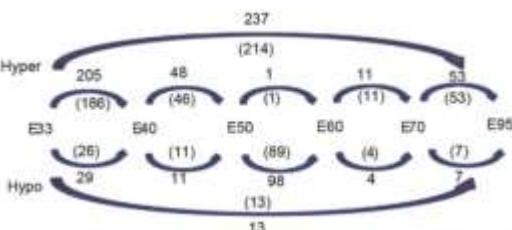


Annotate according to the coordinates of the peaks

annotation	peak注释信息（对于注释到基因上等注释信息的描述）
geneChr	注释基因的染色体信息
geneStart	注释基因的起始位置
geneEnd	注释基因的终止位置
geneLength	注释基因的长度
geneStrand	注释基因的正负链
genelid	注释基因的EntrezID
transcriptId	注释基因的转录本名字
distanceToTSS	被注释Peak距离TSS的距离
ENSEMBL	注释基因的ENSEMBL名
SYMBOL	注释基因的SYMBOL名
GENENAME	注释基因的基本描述信息



According to different experimental designs, it can be divided into time-series analysis and non-time-series analysis, similar to differential analysis performed in RNAseq.

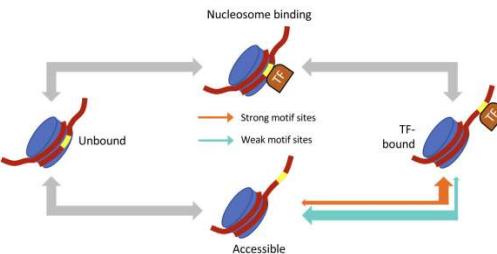


After gene annotation, motif identification, and gene clustering, gene sets can be obtained, allowing for functional enrichment analysis similar to RNAseq, including differential analysis (of genes or gene elements).

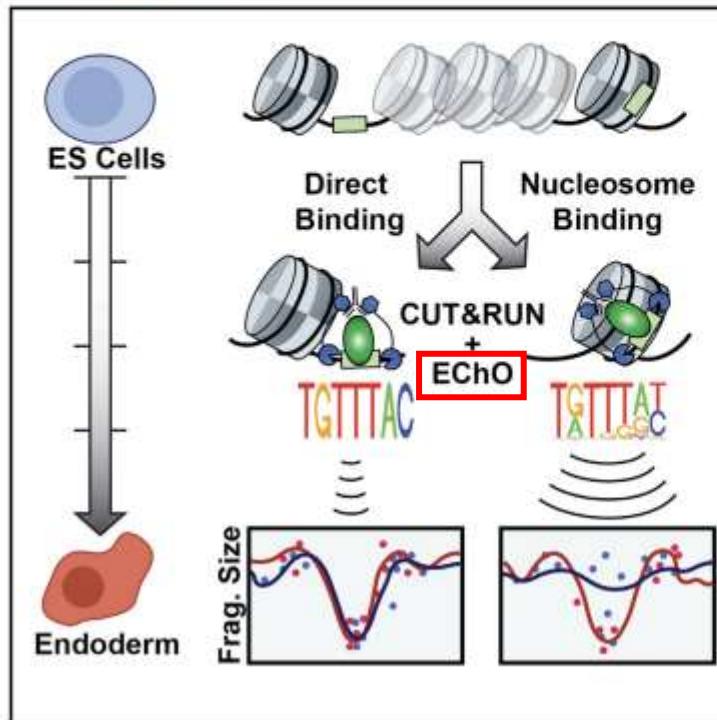
## CUT&RUN new application

# Molecular Cell

## Pioneer Factor-Nucleosome Binding Events during Differentiation Are Motif Encoded



### Graphical Abstract



### EChO

Enhanced Chromatin Occupancy (EChO)

Scripts associated with EChO fragment size profile analysis of CUT&RUN data

### Authors

Michael P. Meers, Derek H. Janssens,  
Steven Henikoff

### Correspondence

steveh@fredhutch.org

### In Brief

Meers et al. use a novel analysis strategy for genome-wide protein-DNA binding data to identify instances of "pioneer factor" binding to nucleosomes during stem cell differentiation. They show that pioneer factor-nucleosome binding occurs in the absence of strong binding motifs, often at accessible sites previously bound by other transcription factors.

[Back to Content](#)

# **Insertsize of CUT&RUN data and dynamic regulation of transcriptional state**

Zhou Ying

2<sup>nd</sup> Dec. 2022

# Work summary

- Validation of two mouse ovarian cell lines with CRISPR knockout of Trp53/Brca2 (morphology, proliferation, and first-generation genome sequencing validation) - demonstrating the usability of mimicking high-grade ovarian cancer models.
- Literature review on Chromatin compaction - not suitable for further investigation as a topic.
- RNA-seq data analysis and communication standardization.
- Comparative research on CUT&RUN and CUT&TAG techniques - a report has been previously conducted.
- **Testing of CUT&RUN data analysis workflow - CUT&RUN data Insertsize exhibits distinct characteristics, literature review suggests a minority of studies utilize this feature for studying dynamic transcriptional regulation.**

Why must Insertsize be considered for CUT&RUN data, and why can only be double-ended sequencing?

## ChIP-seq vs CUT&Tag

	ChIP-seq	CUT&Tag
Input data	PE and SE	Only PE
Adapter trimming	Recommended	Not recommended
Duplicate removal	Yes	No
Bowtie parameters		Filter on insert size
Mitochondrial DNA	Peaks on mitochondrial DNA are real	Recommended to remove mitochondrial reads
Peak calling	MACS2	MACS2 and SEACR

Basepair

- CUT&RUN only supports paired-end sequencing.

- Adapters are not removed.

It is used for subsequent filtering by insert size.

- PCR duplicates are not removed.

Enzymatic cleavage may occur at the same base position, unlike ChIP-seq which randomly breaks, resulting in reads appearing as duplicates.

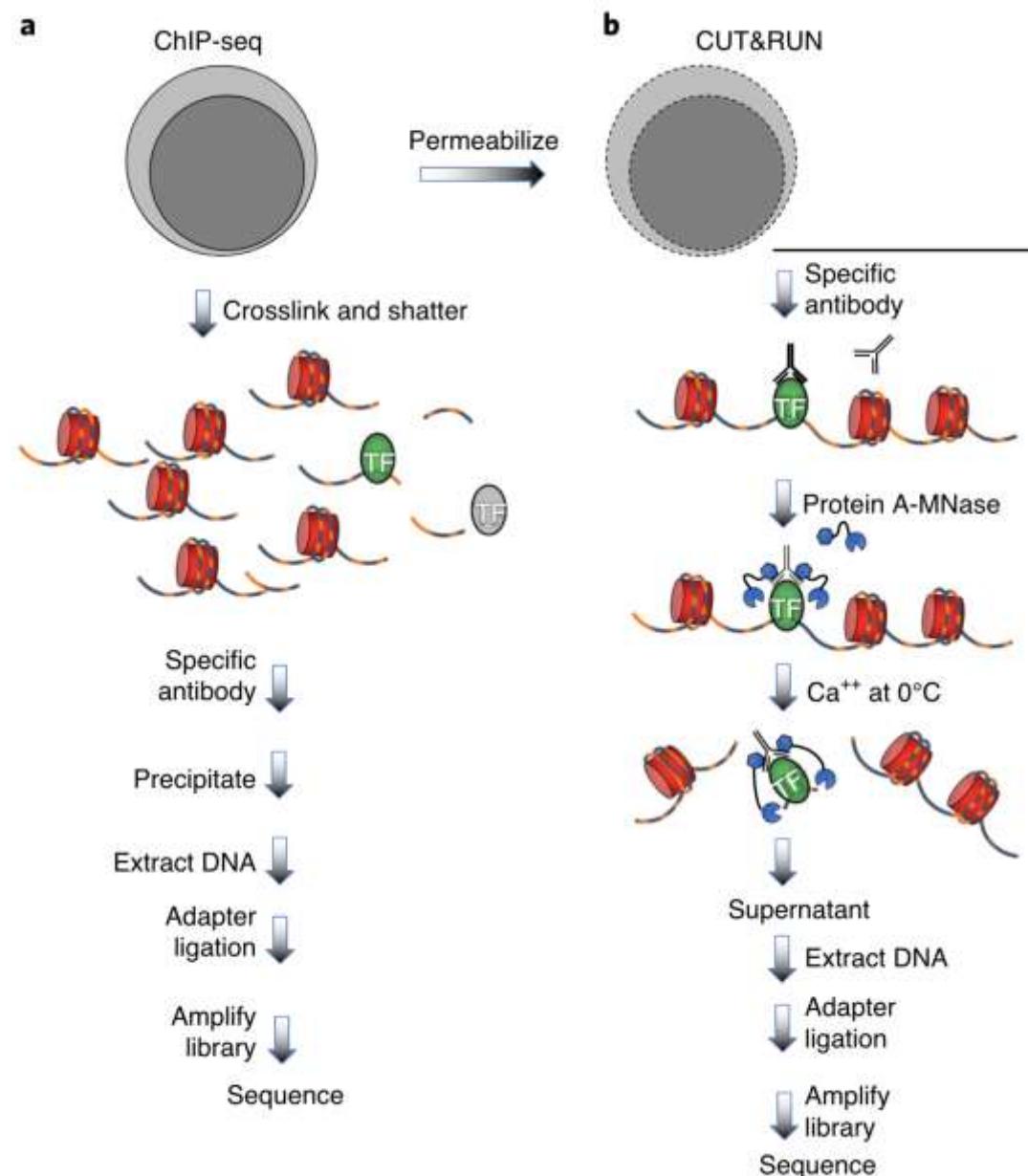
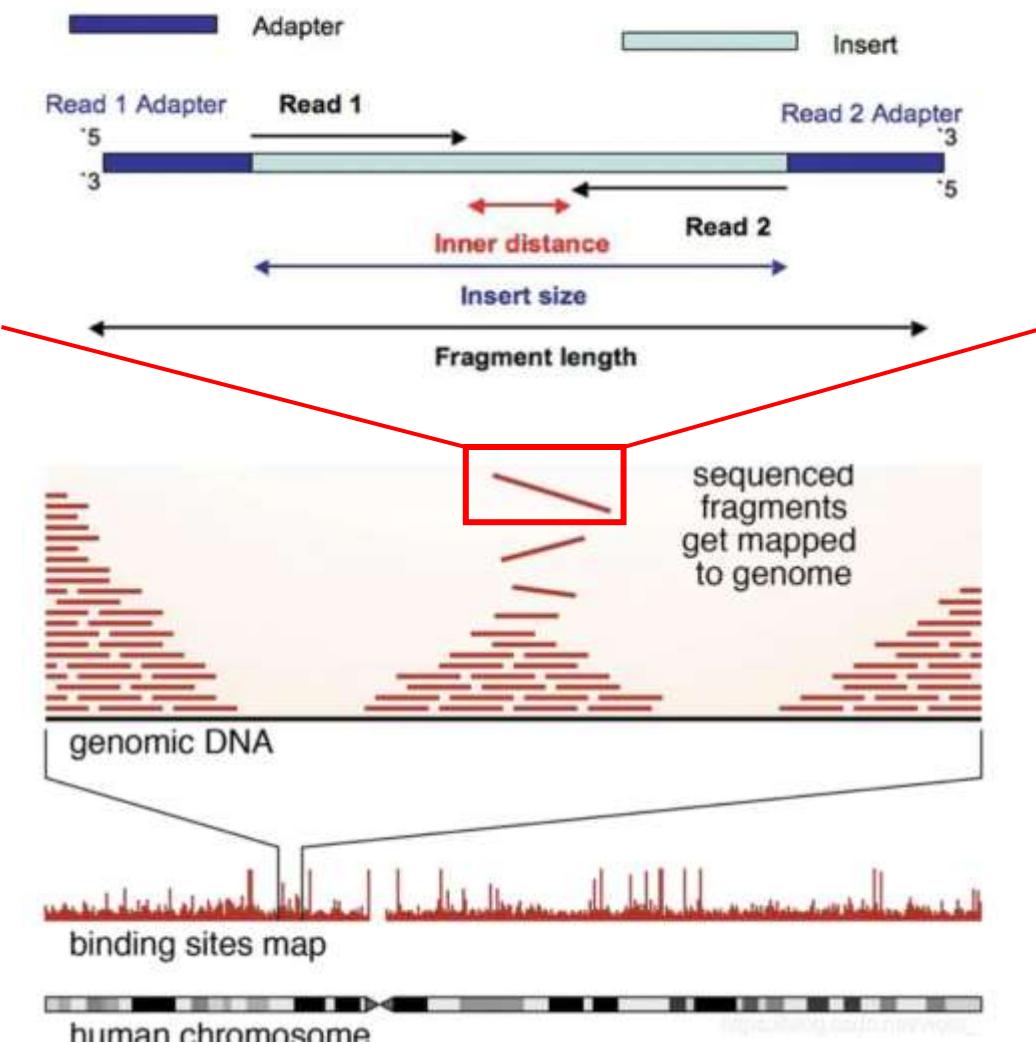
- Parameters for Bowtie alignment should be filtered by insert size.

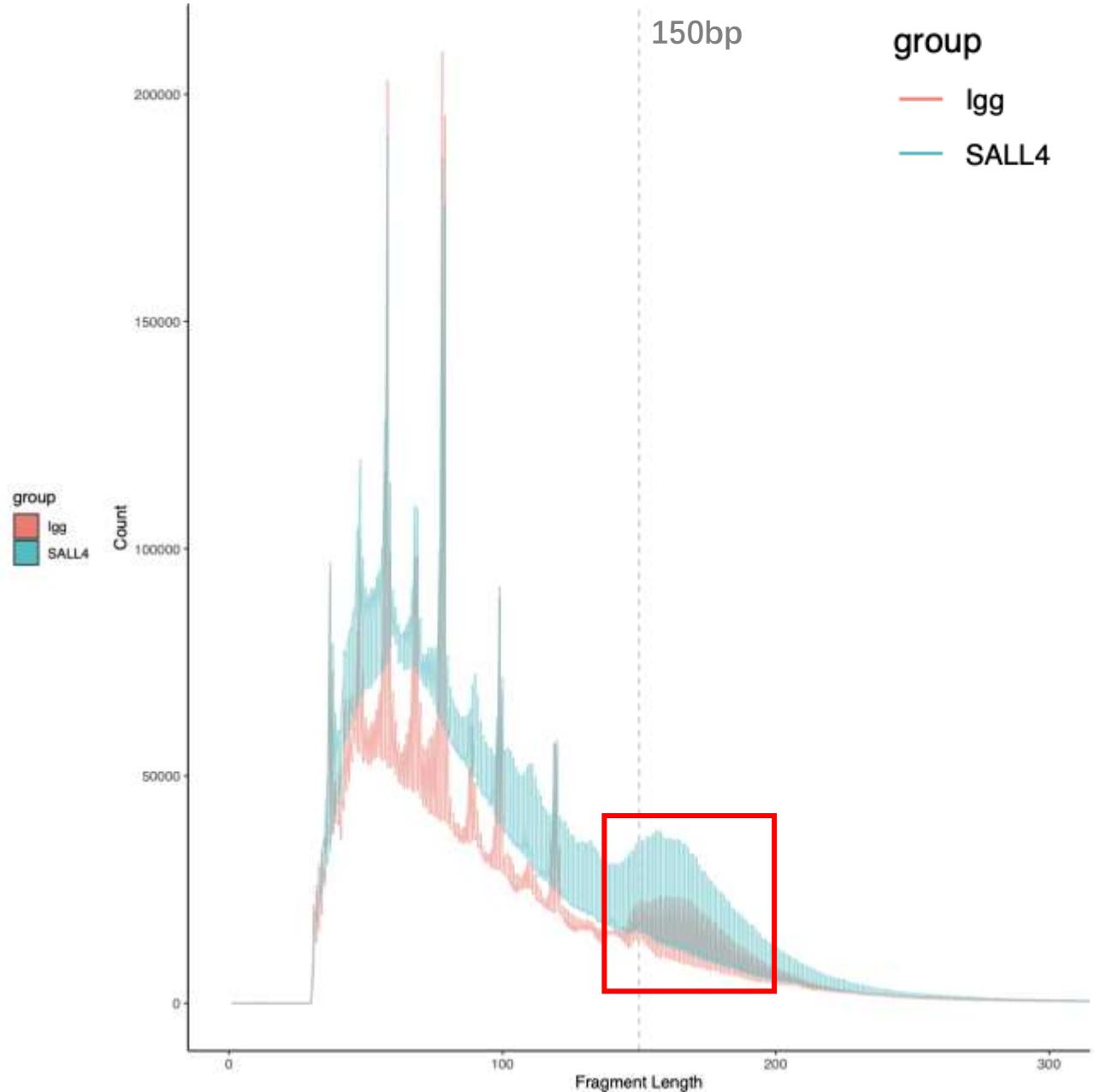
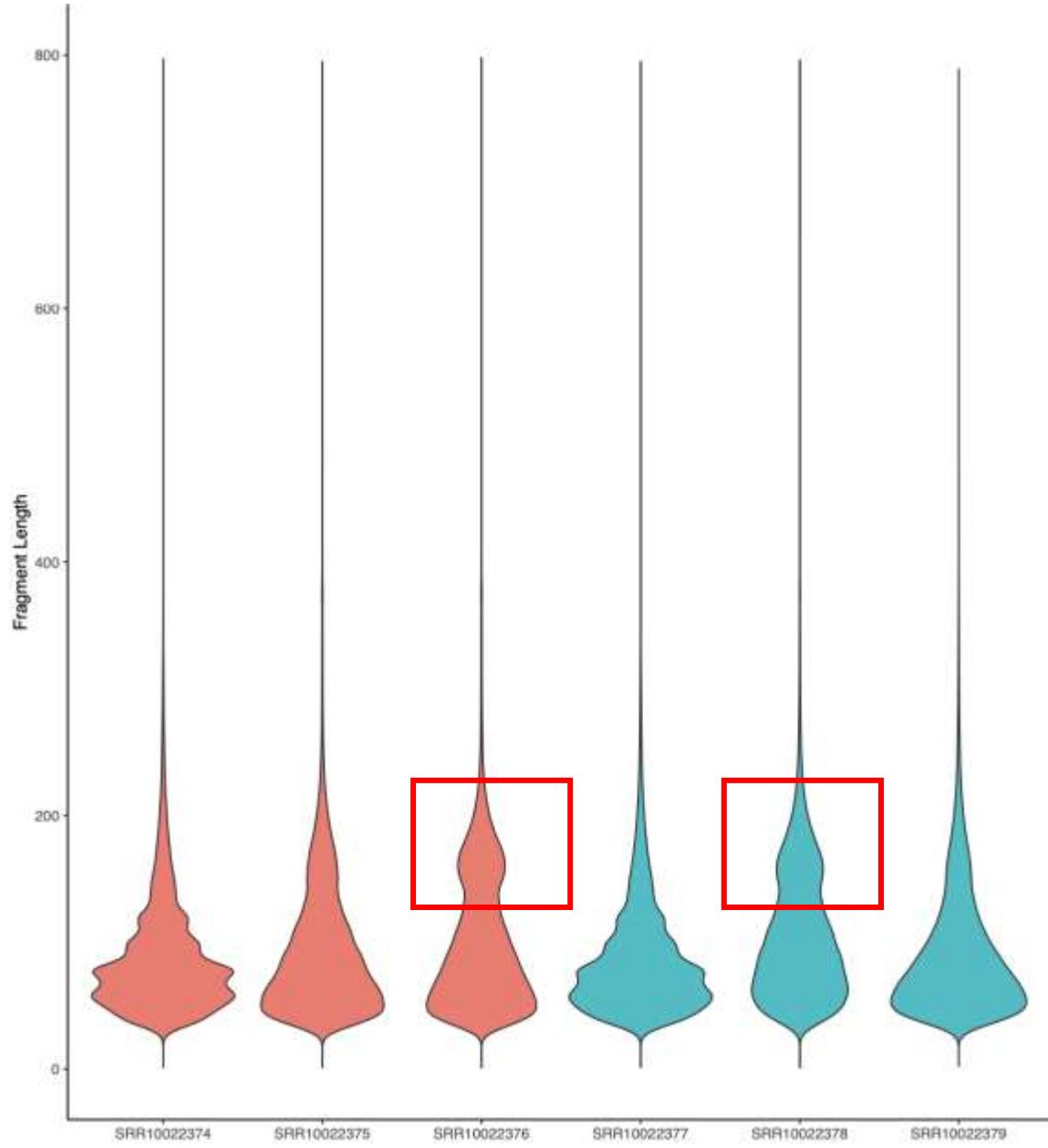
This is primarily used for histone modification.

- Mitochondrial reads are removed.

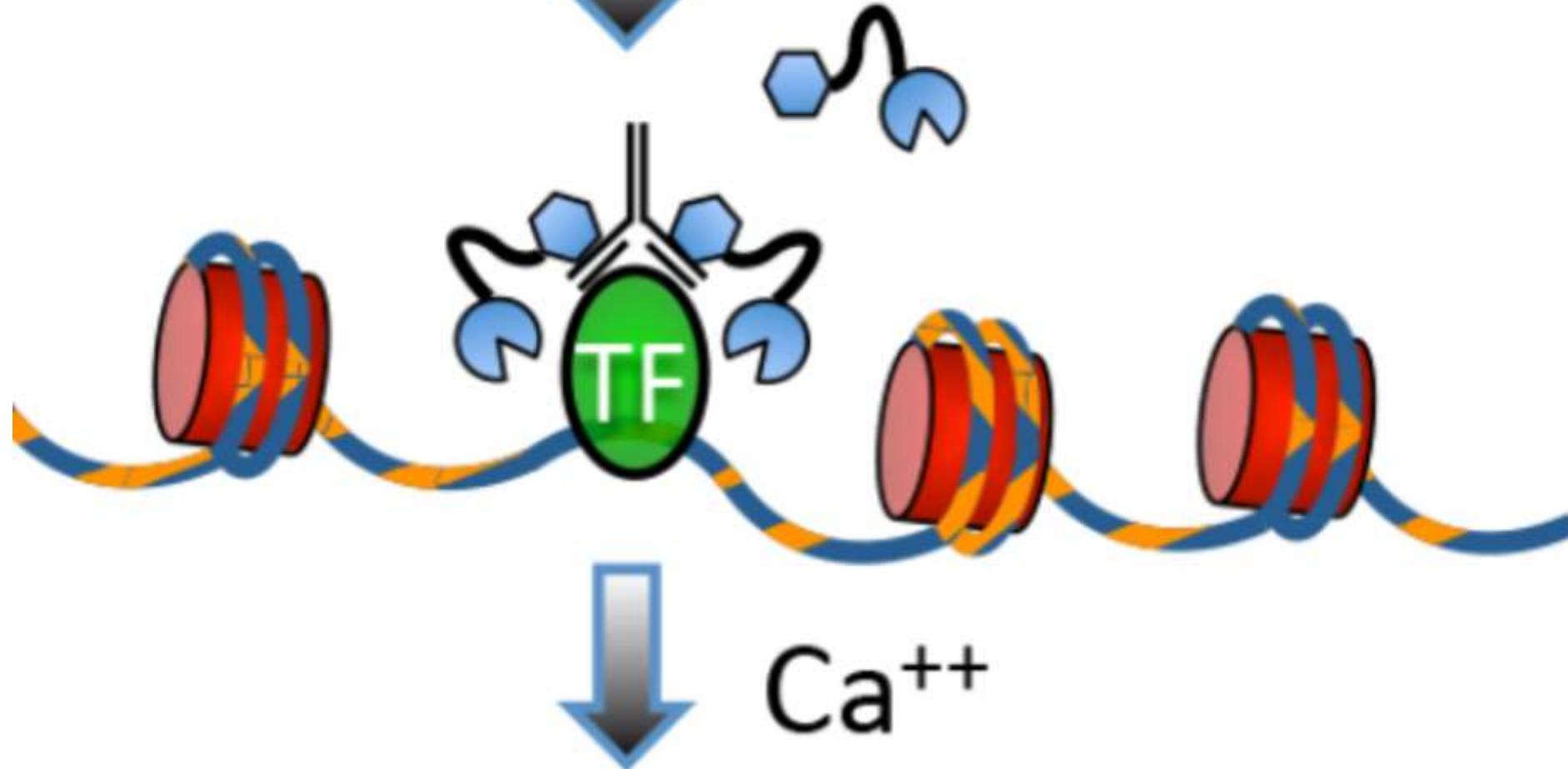
30% of reads are from mitochondria (actual data needs to be examined).

# What is insert size, and why is insert size considered in CUT&RUN data?





**Protein A-MNase**

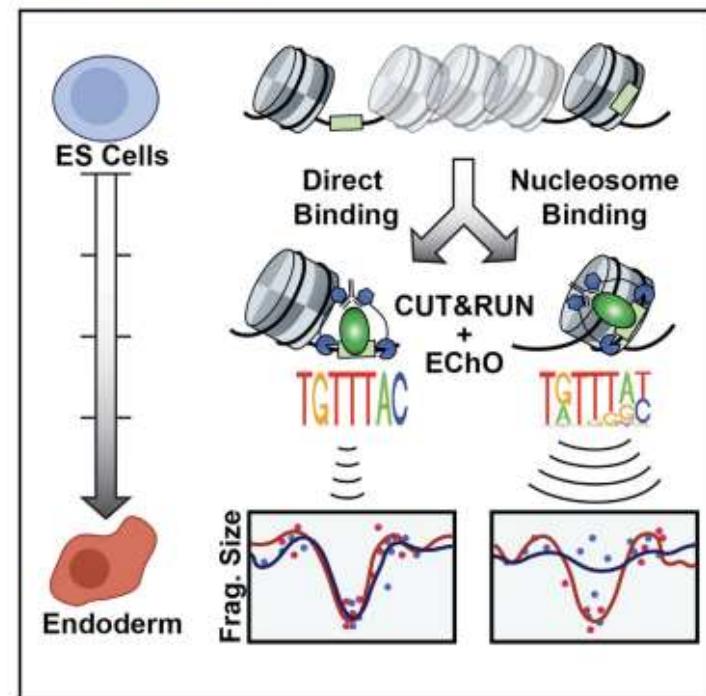




# Molecular Cell

## Pioneer Factor-Nucleosome Binding Events during Differentiation Are Motif Encoded

### Graphical Abstract



### Authors

Michael P. Meers, Derek H. Janssens,  
Steven Henikoff

Correspondence  
[steveh@fredhutch.org](mailto:steveh@fredhutch.org)

### In Brief

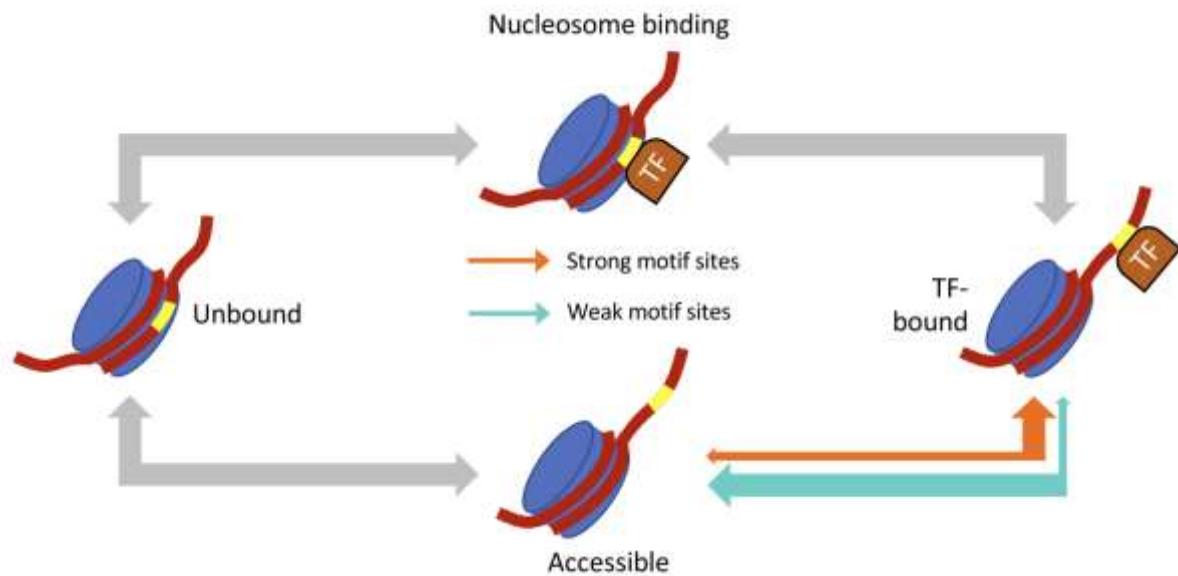
Meers et al. use a novel analysis strategy for genome-wide protein-DNA binding data to identify instances of "pioneer factor" binding to nucleosomes during stem cell differentiation. They show that pioneer factor-nucleosome binding occurs in the absence of strong binding motifs, often at accessible sites previously bound by other transcription factors.

ARTICLE

## An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites

Peter J Skene, Steven Henikoff\*

Howard Hughes Medical Institute, Basic Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, United States

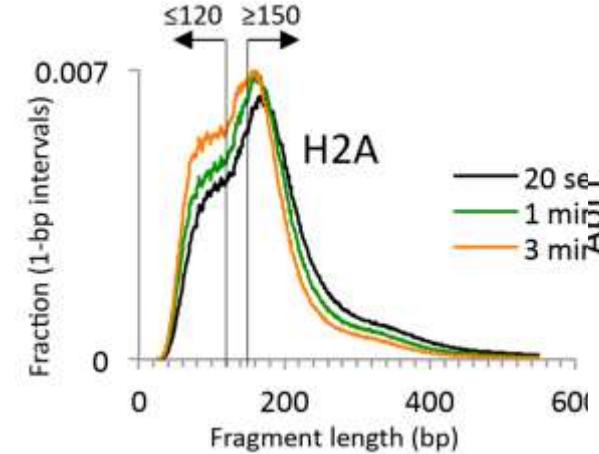
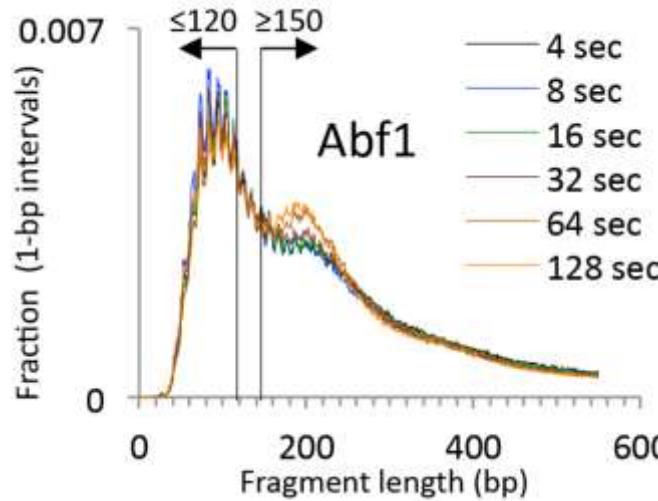


The CUT&RUN technique has high efficiency and precision. The technique is discussed from different angles

## An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites

Peter J Skene, Steven Henikoff\*

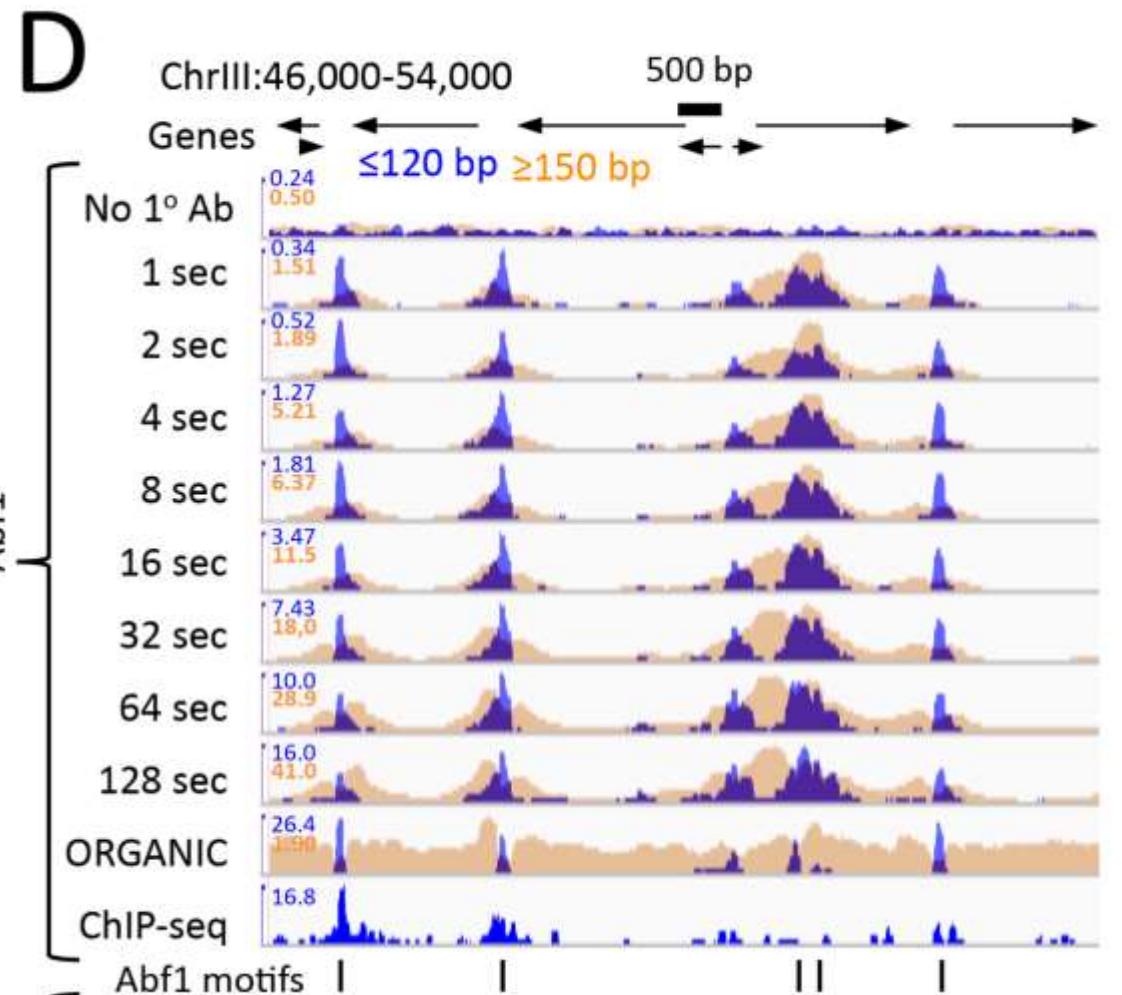
Howard Hughes Medical Institute, Basic Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, United States

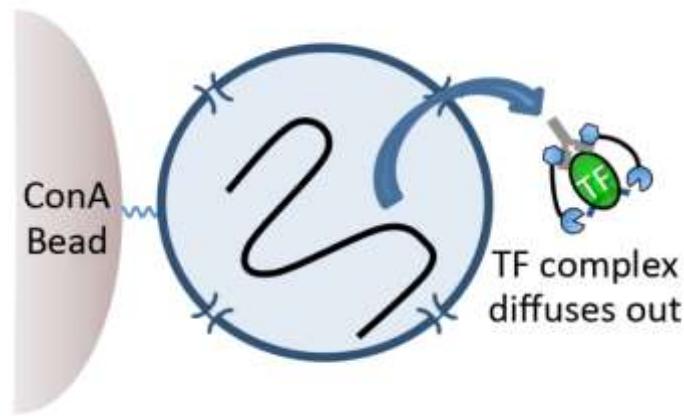
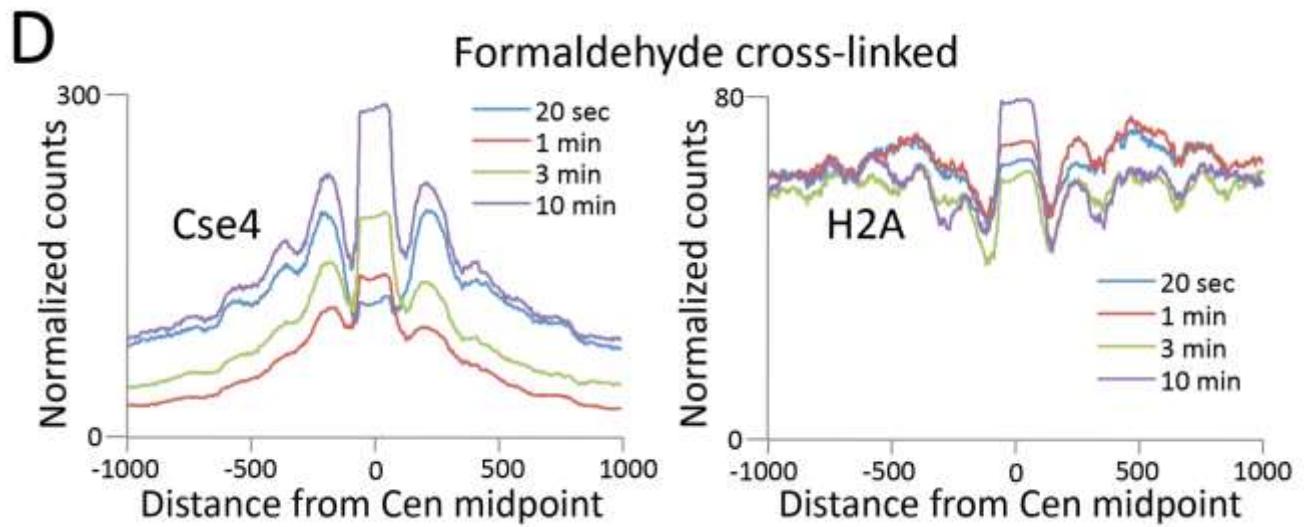
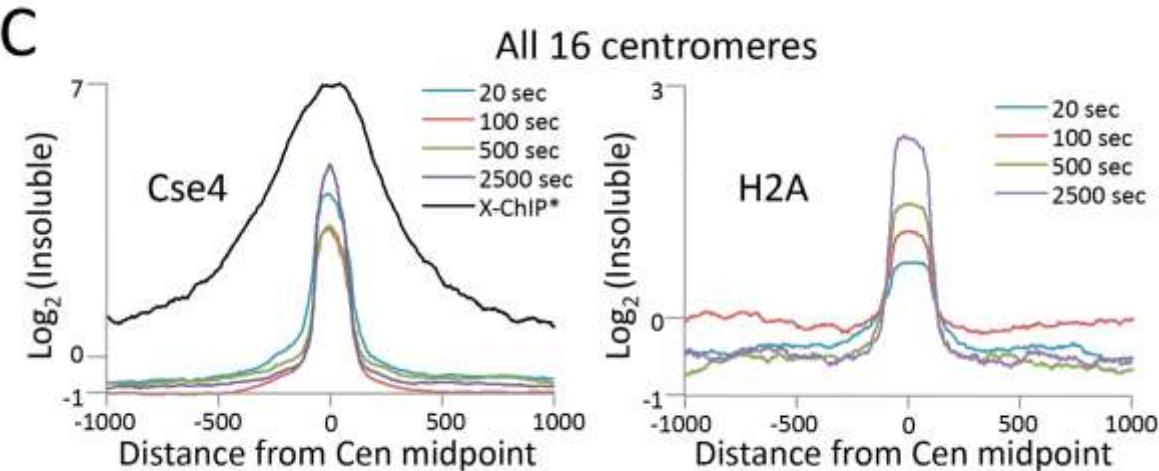
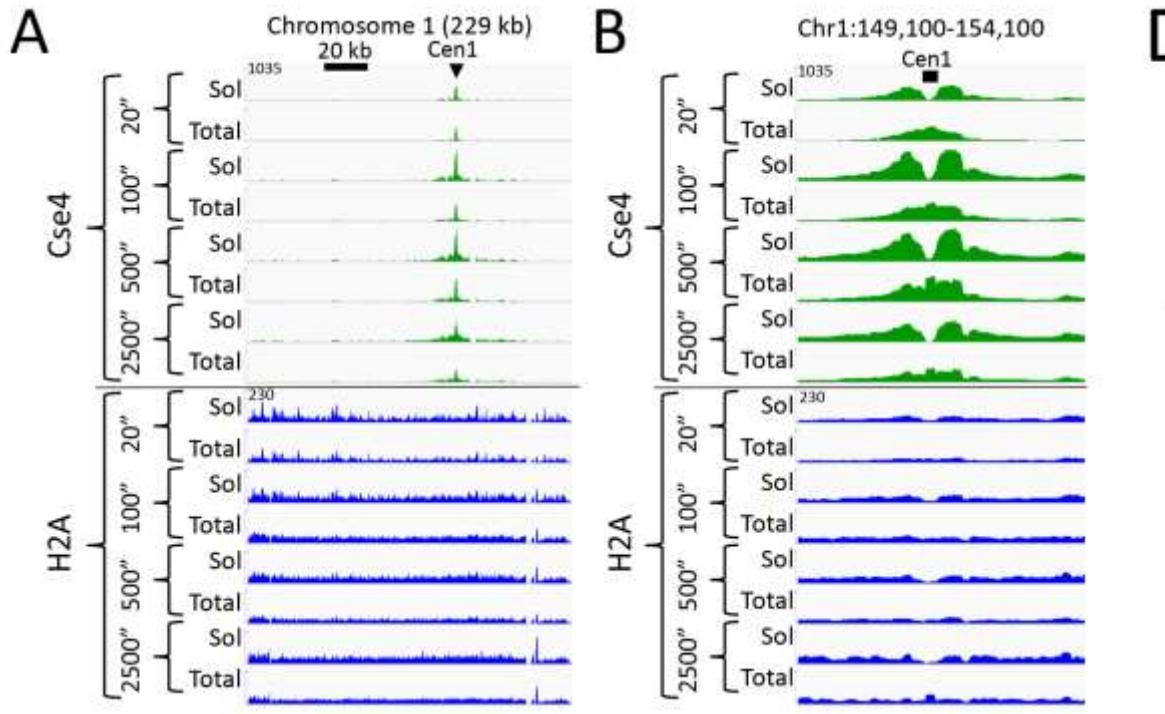


Transcription factors, histone  
CUT&RUN fragment size,  
enzyme digestion time

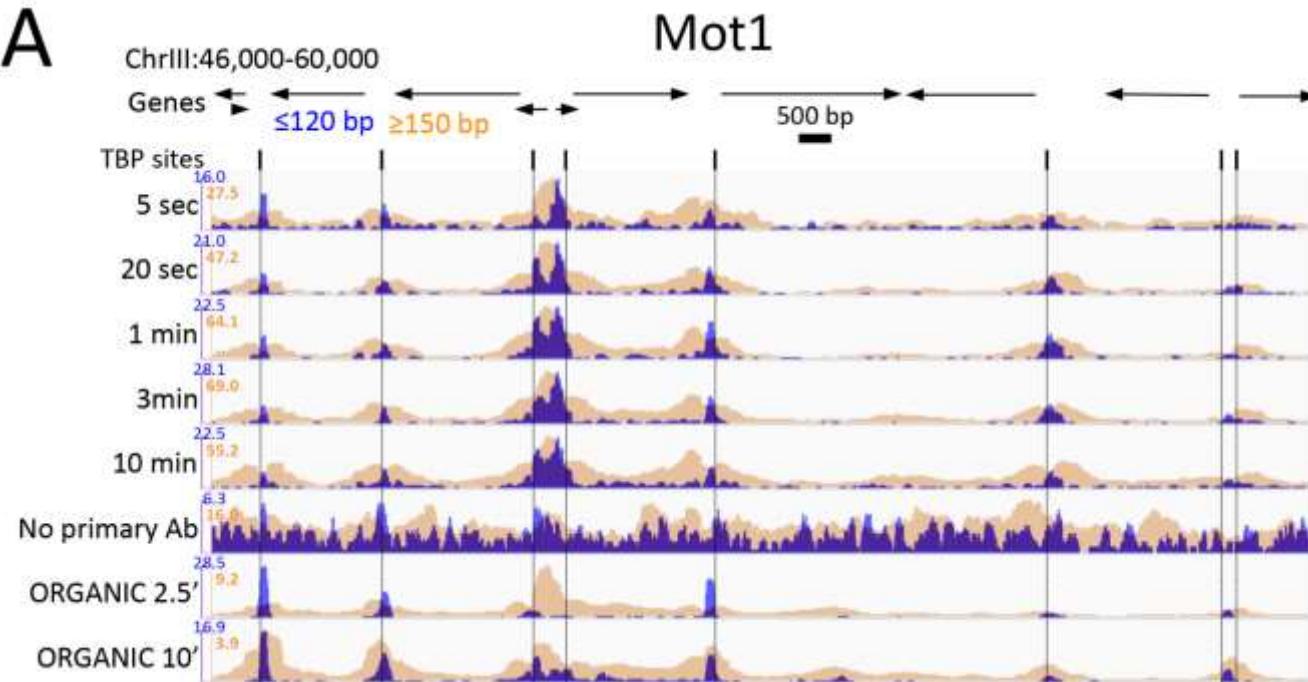
Insoluble binding complex  
(referring to larger complexes)

## TF dynamic binding

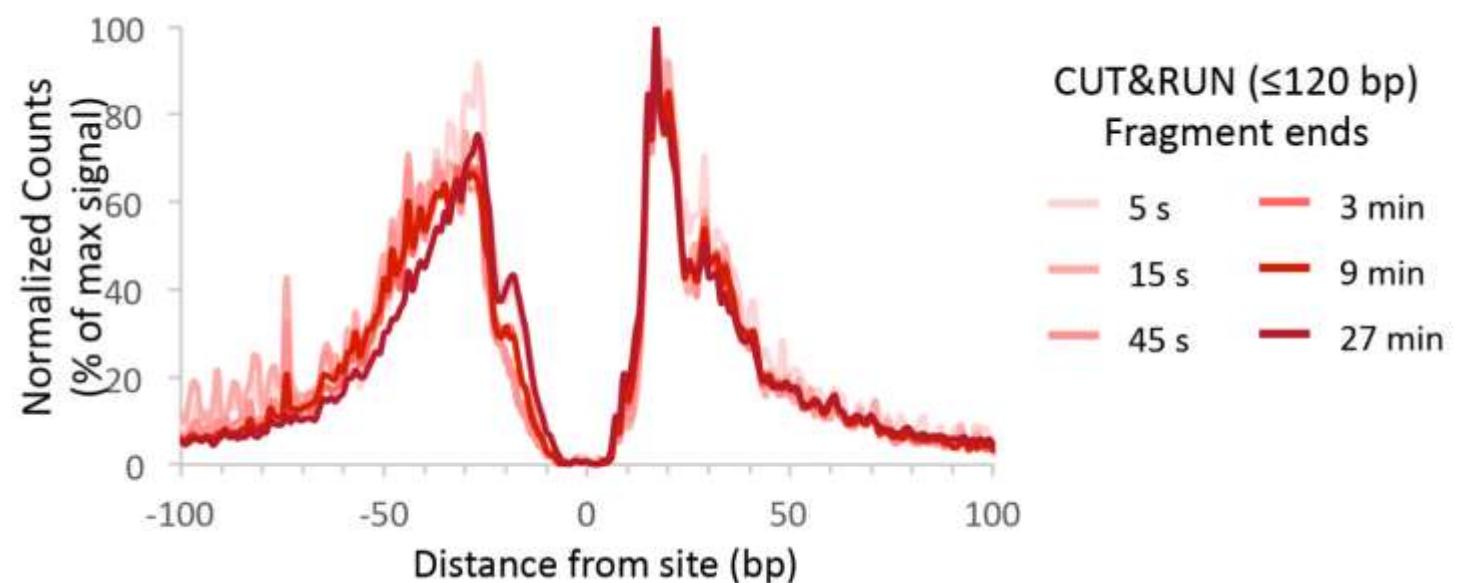


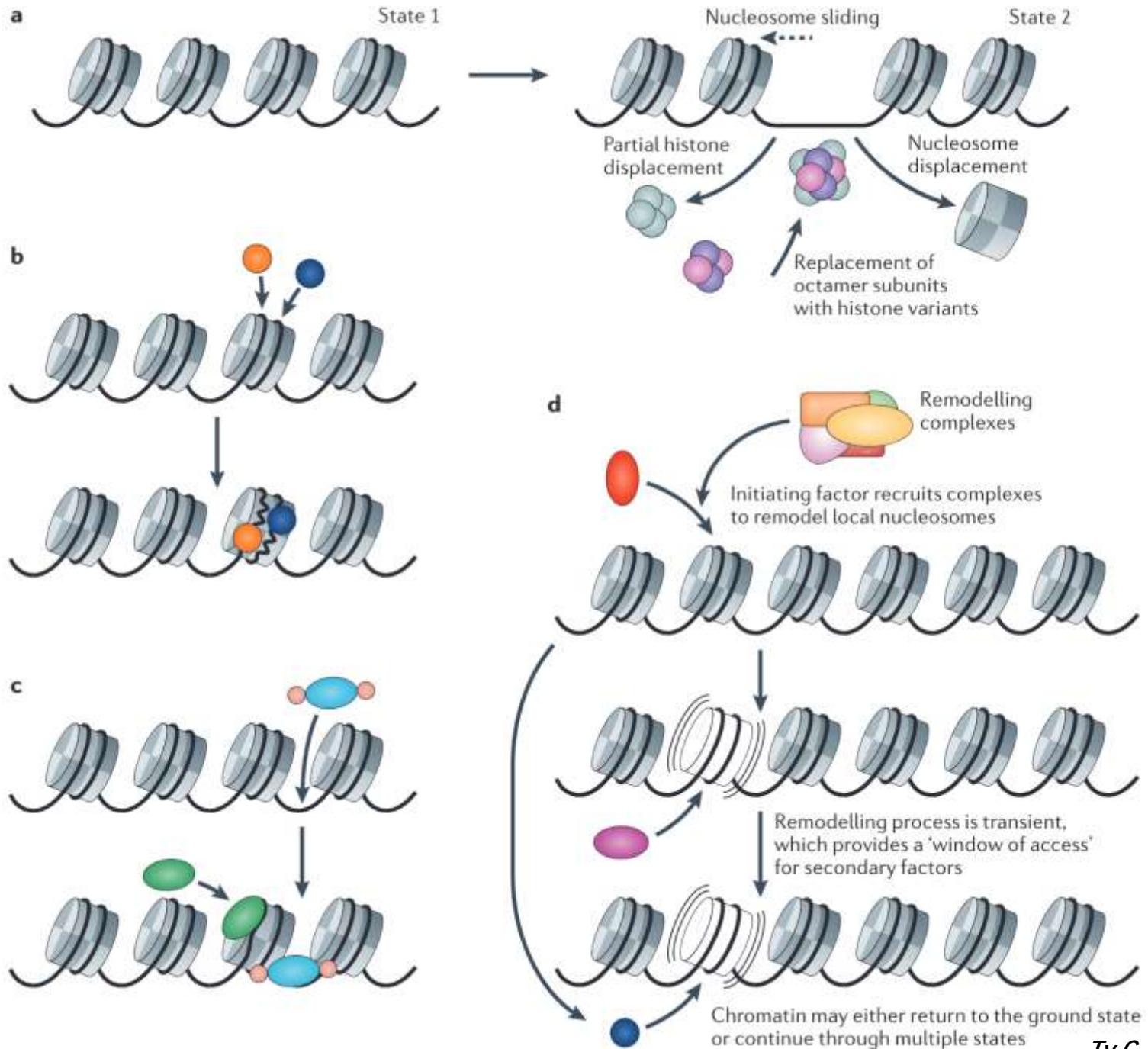


- The insoluble complex is too large, the nucleus is cleaved, the DNA is extracted and then screened for fragment size
- It is better to fix this complex with formaldehyde

**A**

Dynamic changes can be accurately described





- The binding patterns of different transcription factors (TFs) may lead to variations in fragment size and distribution generated by CUT&RUN, and even differences in peak shapes.
- Adjusting the general workflow of CUT&RUN analysis accordingly.
- Comparing the volume of TF proteins/complexes with the sizes of cleaved fragments in CUT&RUN data.
- Investigating Mnase-seq data to examine the characteristics of fragment distribution, particularly when no antibody is added or when IgG antibody shows no specificity (input for MNase-seq).

[Back to Content](#)

# **Summary of chromatin compaction**

## **reading and thinking**

Zhou Ying

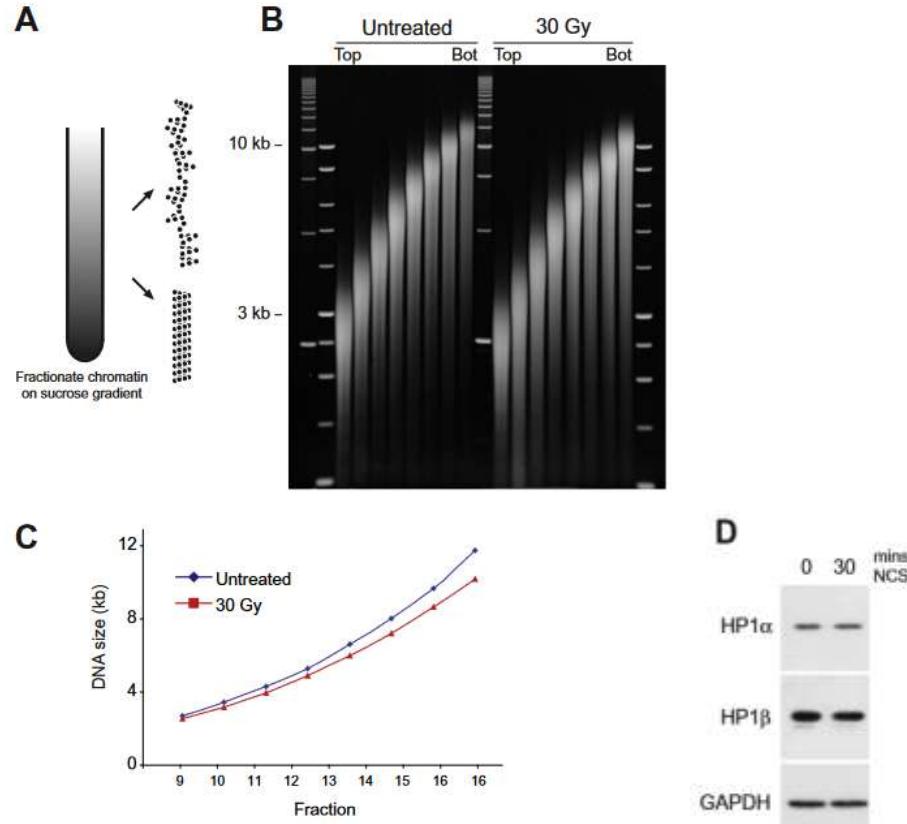
# Searching summary

- ◆ Before 2011, it was generally believed that DNA damage caused chromatin relaxation, which facilitated the rapid binding of repair molecules to the damaged site for quick repair. However, an article in 2011 argued that global relaxation posed too much risk to the entire genome and was therefore not reasonable. Instead, it proposed that local chromatin opening facilitated repair, while global DNA compaction protected the entire genome through spatial decondensation and condensation.
- ◆ In 2014, a paper published in Cell Reports introduced the concept that DNA damage caused local chromatin opening within 15 minutes of irradiation, followed by hypercondensation to protect the DNA and prevent damage diffusion. This demonstrated both temporal decondensation and condensation phenomena, although the focus of the study was primarily on local events from an experimental technique perspective.
- ◆ In 2015, an article examined cells irradiated with UV and those left untreated using RNAseq, ChIP-Seq, and FAIRE-seq. It proposed that after UV irradiation, chromatin accessibility decreased, leading to global condensation.
- ◆ Based on the articles found so far, the changes in chromatin spatial configuration after DNA damage remain controversial. Discussions on compaction mainly focused on studies from 2011 to 2015. Later, it was realized that this was not merely a question of whether compaction occurred, but rather how relaxation and condensation were balanced in both spatial and temporal dimensions. Moreover, not only radiation but also radiation-like drugs, chemotherapy drugs, and even DNA damage occurring during the cell cycle could induce similar chromatin structural changes. Therefore, subsequent research not only discussed compaction but also focused on DNA damage-associated chromatin dynamics.
- ◆ After 2015, there were few articles specifically addressing compaction; instead, research should be conducted under the framework of DNA damage-associated chromatin dynamics.



## Global chromatin fibre compaction in response to DNA damage

2011

Charlotte Hamilton <sup>a</sup>, Richard L. Hayward <sup>a,b</sup>, Nick Gilbert <sup>a,b,\*</sup><sup>a</sup>Institute of Genetics and Molecular Medicine, The University of Edinburgh, Edinburgh EH4 2XR, UK<sup>b</sup>Breakthrough Research Unit, The University of Edinburgh, Edinburgh EH4 2XR, UK

## DSBs result in decondensation or condensation?

Most of the previous studies believe:

DNA damage

Chromatin decondensation

Benefit to DDR pathway

Research conclusions:

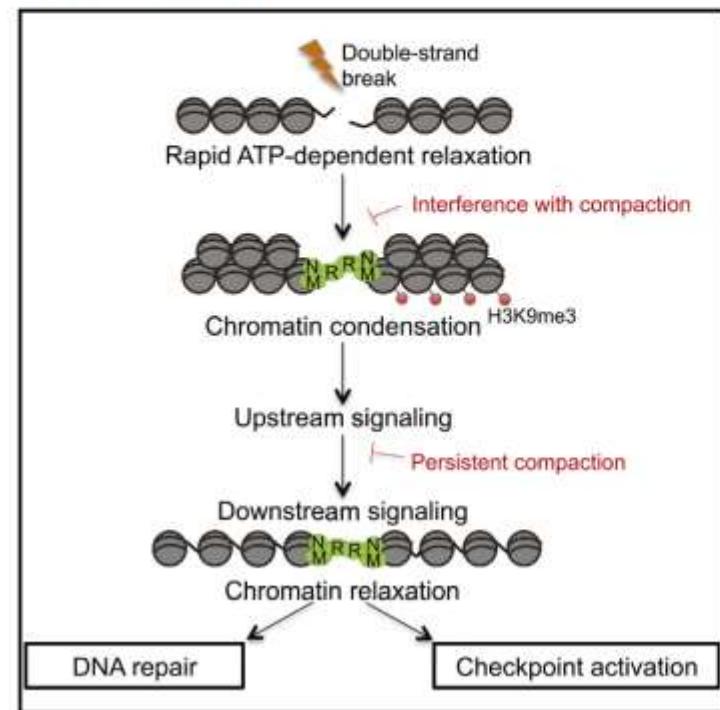
Global decondensation increasing the risk of damage

Local decondensation for DDR

Global condensation protect cells from further damage

## Activation of DNA Damage Response Signaling by Condensed Chromatin

### Graphical Abstract



### Authors

Rebecca C. Burgess, Bharat Burman,  
Michael J. Kruhlak, Tom Misteli

### Correspondence

[mistelit@mail.nih.gov](mailto:mistelit@mail.nih.gov)

### In Brief

Relaxation of chromatin is important for checkpoint activation and DNA repair, but the role of chromatin condensation has been enigmatic. Burgess et al. show that chromatin condensation is an integral but transient part of the DNA damage response. Whereas condensed chromatin enhances upstream signaling, persistent condensation inhibits downstream repair and recovery.

### Main conclusions:

- Light-induced double-strand breaks (DSBs) lead to hypercondensation of local chromatin after recompaction.
- Chromatin condensation activates upstream signals for DNA damage repair.
- Persistent compaction inhibits repair and reduces survival.

### Biological meaning:

1. Local compaction can rapidly prevent damage from spreading.
2. Rapid decompaction can facilitate the swift repair of small-scale damage.
3. Increased damage promotes persistent compaction, reducing survival mechanisms. Could this be understood as the abandonment of cells unable to repair large-scale damage, perhaps through apoptosis or senescence?

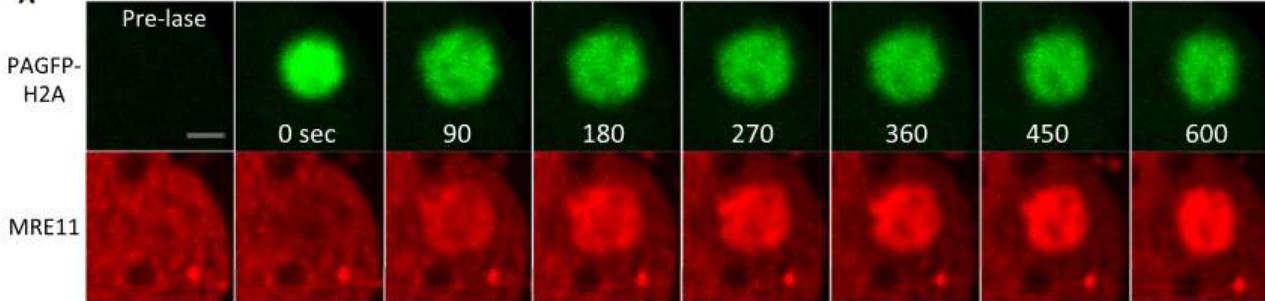
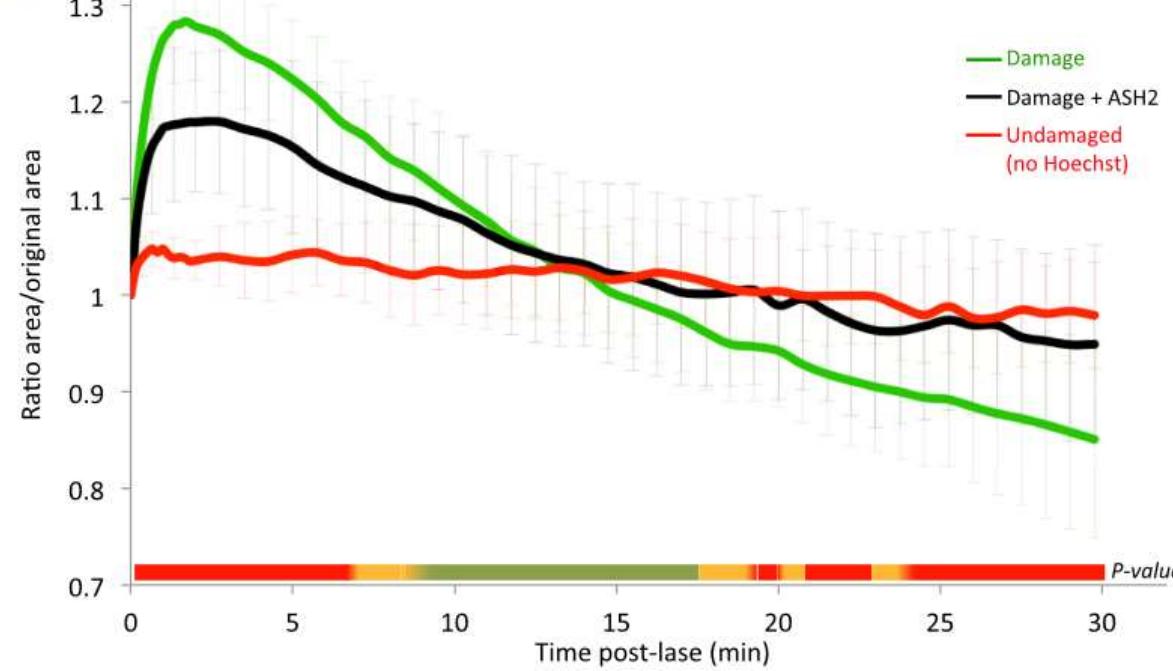
### Limitation:

While this study introduces concepts, it only demonstrates recruitment of a few compaction and repair-related molecules as evidence.

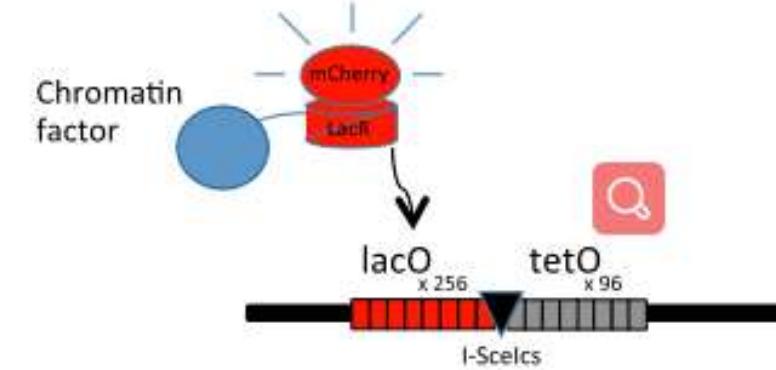
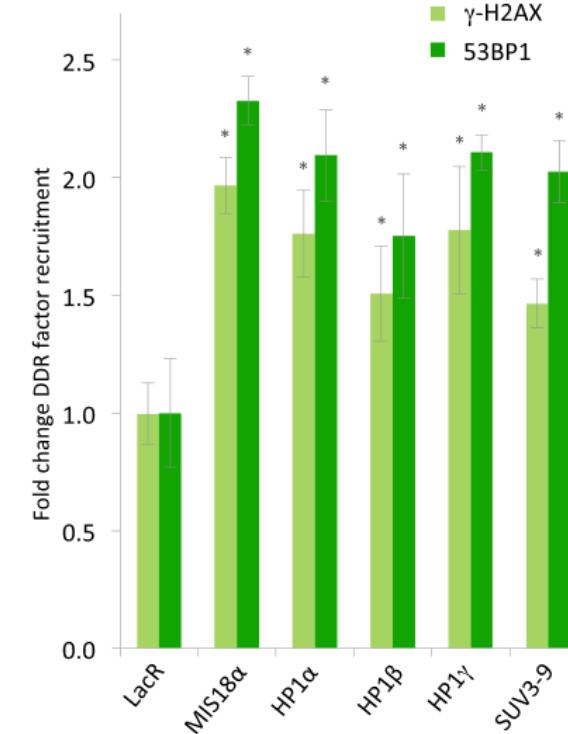
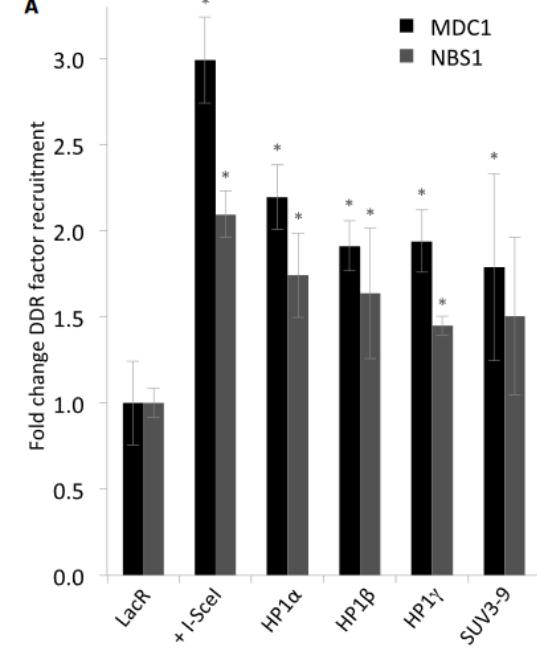
### Thinking:

**High expression and variations in compaction molecules in tumor cells from TCGA data suggest an association between mutation and DNA damage, potentially initiating compaction to activate repair pathways.**

**Is compaction induced by mutations activating repair, or does it induce apoptosis and senescence? Alternatively, does activation of repair pathways increase survival pressure, leading to the selection of tumor cells resistant to compaction/repair from an evolutionary perspective?**

**A****B**

Recompaction → hypercondensation

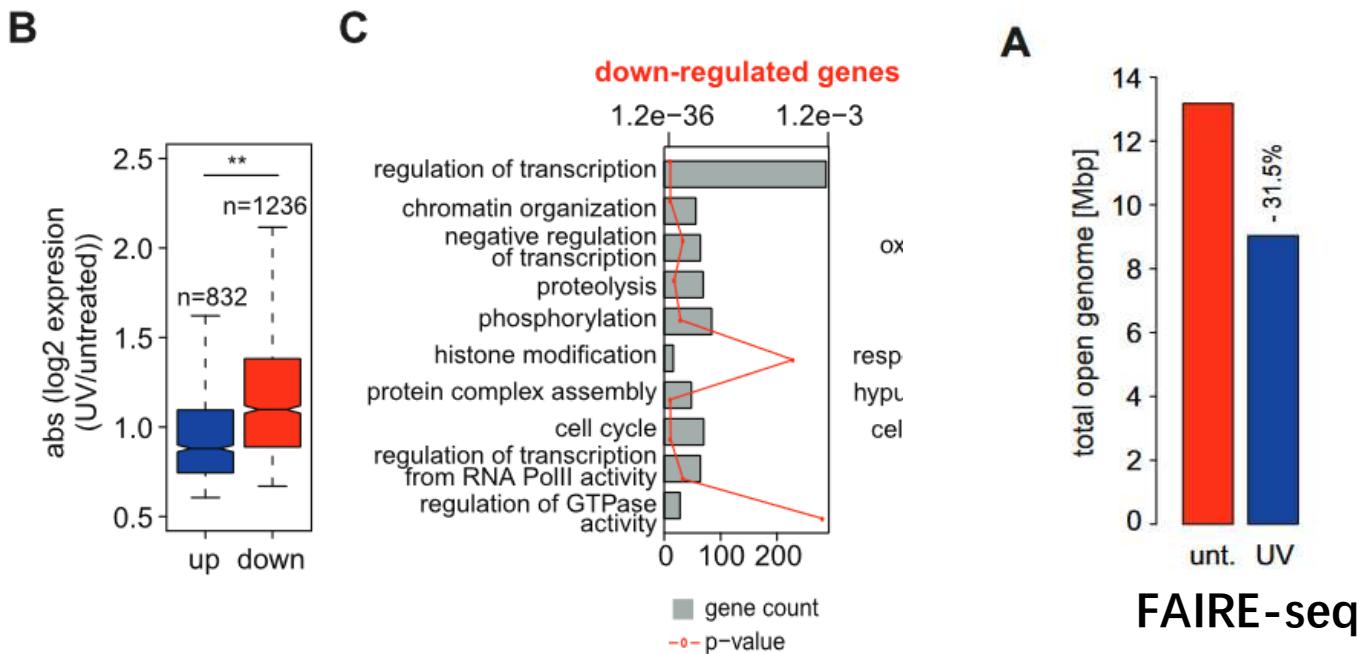
**C****A**

RESEARCH ARTICLE

2015

# Dynamics of chromatin accessibility and epigenetic state in response to UV damage

Sandra Schick\*, David Fournier\*, Sudhir Thakurela, Sanjeeb Kumar Sahu, Angela Garding and Vijay K. Tiwari†



RNAseq conclusion:

1. After irradiation, a significant number of genes were downregulated.
2. Genes related to the regulation of transcription and chromatin organization were downregulated.

UV treat murine fibroblast cells (NIH3T3)  
RNAseq found differential genes

H3K27ac-ChIP-seq 确定active promoters and enhancers

FAIRE-seq & ATAC-seq进一步研究chromatin accessibility

The article found that FAIRE discovered a decrease in the overall openness of the genome after irradiation, suggesting genome condensation following UV irradiation.

- ◆ How far can DNA damage affect the damaged site? Are the chromatin structure/regulatory mechanisms different at sites adjacent to and distant from the damage (relaxation or compaction)?
- ◆ Are cellular responses different between large-scale damage (severe disruption, difficult to repair) and small-scale damage? (For example: large-scale damage may lead to chromatin condensation, promoting apoptosis/senescence, as apoptotic and senescent cells also exhibit chromatin condensation)
- ◆ Are the repair mechanisms different for damage at constitutive heterochromatin and facultative heterochromatin? (Different cell types may have different responses to damage)
- ◆ How do chromatin relaxation and compaction states reach equilibrium over time and space in response to damage?
- ◆ How does the type of damage affect chromatin **conformation** (mutations in tumor cells may also initiate **compaction as a form of DNA damage repair**? Does accumulation of such damage lead to hypercondensation, thereby reducing cell survival/inducing cell senescence/apoptosis, and do tumor cells have mechanisms to evade this and survive)?

# The accumulation of mutations based on TCGA data is associated with increased molecular expression levels of compaction.

**Could mutations in tumor cells also serve as a form of DNA damage triggering compaction to initiate DNA damage repair?**

1. Whether the molecules related to compaction that accumulate expression levels with mutations regulate condensation or relaxation, or both, can be verified.
2. Can the mutations in existing tumor-related datasets be combined with data on accessibility such as ATACseq/FAIRE-seq to observe whether there are changes in accessibility at the local and global levels of mutation positions, and whether such changes are associated with the accumulation of mutations?
3. Compare the changes in compaction/DDR-related molecules caused by mutations in tumor cells with those induced by UV treatment, radiation-like drugs, or chemotherapy drugs to identify genes with differential damage or changes in accessibility-related genes.

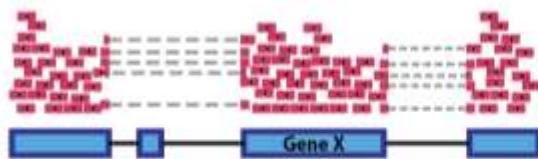
# **Explanation of DESeq2, edgeR, GSEA Analysis, and RNAseq Normalization Methods**

Zhou Ying

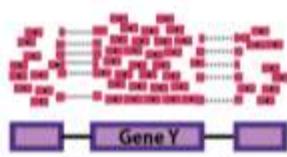
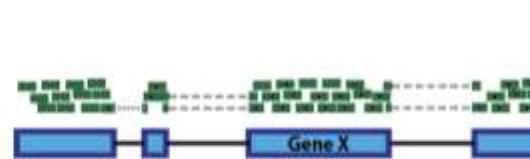
- Explanation of Data Standardization Methods
- DESeq2 and edgeR
- Basic Principles of GSEA (Gene Set Enrichment Analysis)
- Demonstration of GSEA Software
- Interpretation of Previous GSEA Results

# RNASeq Normalization

Sample A Reads



Sample B Reads

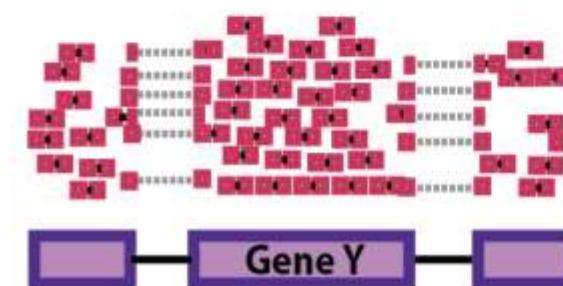
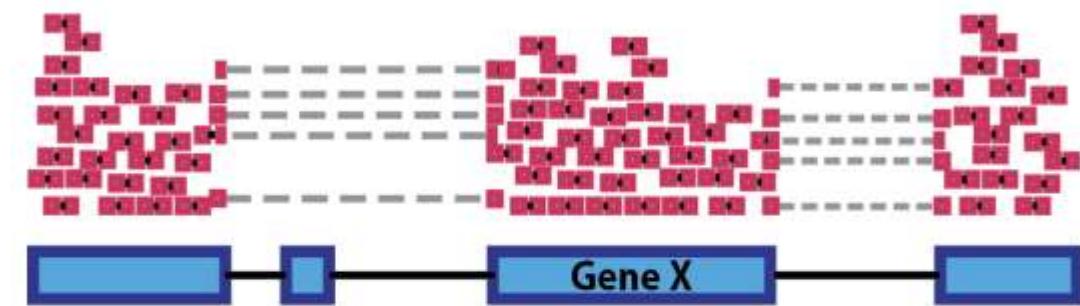


测序深度，测序得到的碱基总量 (bp) 与基因组大小 (Genome) 的比值。

注意：在上图中，每个粉红色和绿色矩形代表与基因对齐的读段。用虚线连接的读段连接横跨内含子的读段。

The impact of sequencing depth

Sample A Reads



The impact of gene Length

# RNASeq Normalization

Use 10 to represent million

归一化方法

CPM (counts per million)

TPM (transcripts per kilobase million)

RPKM/FPKM (reads/fragments per kilobase of exon per million reads/fragments mapped)

DESeq2's median of ratios [1]

EdgeR's trimmed mean of M values (TMM) [2]

	A	B	C	D
1	Gene Name	S1 counts	S2 counts	S3 Counts
2	A (2kb)	10	12	30
3	B (4kb)	20	25	60
4	C (1kb)	5	8	15
5	D (10kb)	0	0	1
6	lib size	35	45	106

Counts

P	Q	R	S
Gene Name	S1 TPM	S2 TPM	S3 TPM
A (2kb)	=L2/L\$6*10	2.96296296	3.32594235
B (4kb)	3.33333333	3.08641975	3.32594235
C (1kb)	3.33333333	3.95061728	3.32594235
D (10kb)	0	0	0.02217295
total	10	10	10

TPM

**CPM:** Replication of repeated samples within the same sample group, not intended for within-sample comparisons or DE analysis.

FPKM

**FPKM:** Gene count comparison within samples; not suitable for between-sample comparisons or DE analysis.

**TPM:** Gene count comparison within samples or between samples within the same sample group; not intended for DE analysis.

**DESeq2:** Sample-to-sample comparison and DE analysis; not suitable for within-sample comparisons.

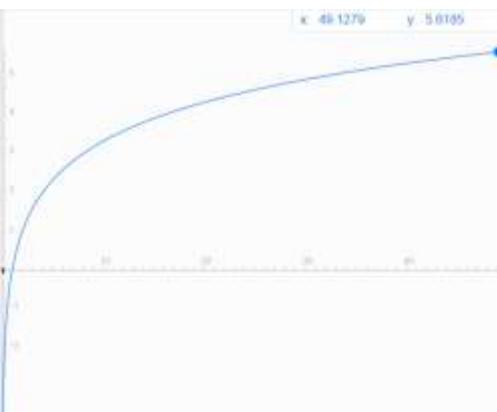
**edgeR (TMM):** Gene count comparison between samples and within samples, used for DE analysis.

F	G	H	I
Gene Name	S1 CPM	S2 CPM	S3 CPM
A (2kb)	=B2/B\$6*10		2.83018868
B (4kb)	5.71428571	5.55555556	5.66037736
C (1kb)	1.42857143	1.77777778	1.41509434
D (10kb)	0	0	0.09433962
total	10	10	10

CPM

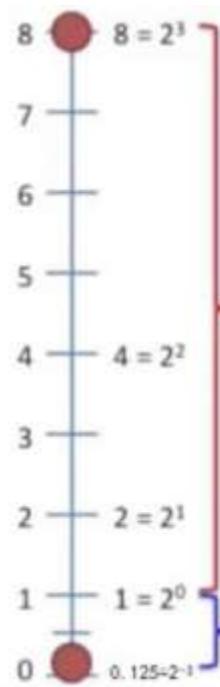
K	L	M	N
Gene Name	S1 FPKM	S2 FPKM	S3 FPKM
A (2kb)	=G2/2	1.33333333	1.41509434
B (4kb)	1.42857143	1.38888889	1.41509434
C (1kb)	1.42857143	1.77777778	1.41509434
D (10kb)	0	0	0.00943396
total	4.28571429	4.5	4.25471698

## TPM & log2(TPM+1)

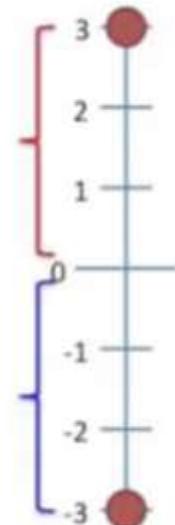


Taking the logarithm of a quantity: compresses data for comparison, making it symmetric about zero.

Purpose of  $\log(x+1)$ : to avoid taking the logarithm of zero or negative values before transformation.



Use a log scale/axis when talking about fold change. This puts positive and negative fold changes on a symmetric scale.



CSDN @余丁，微生信

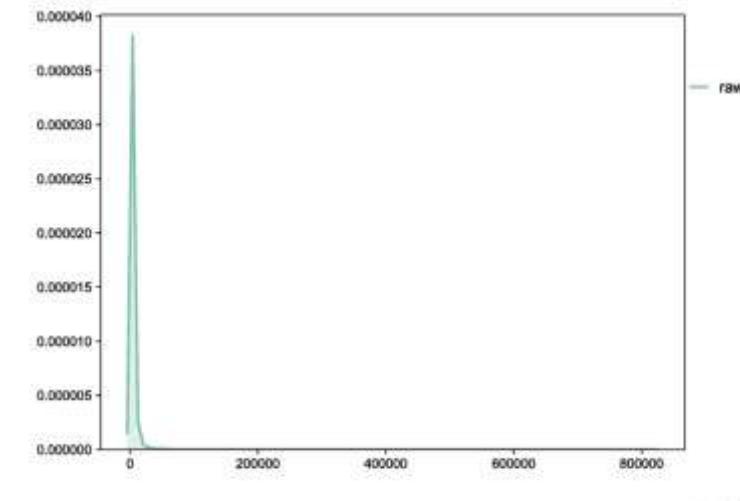
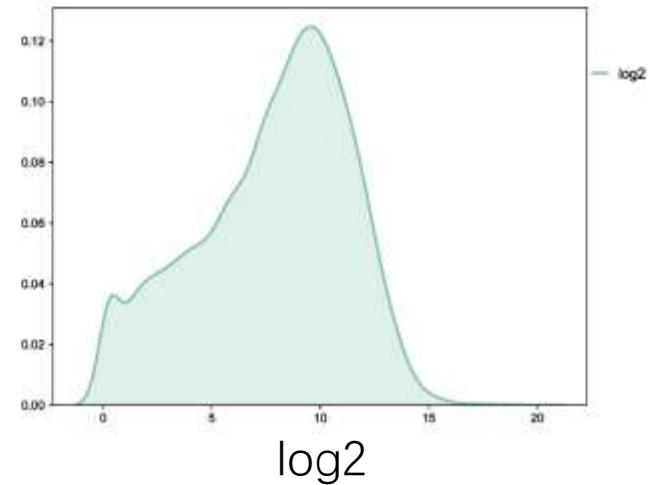
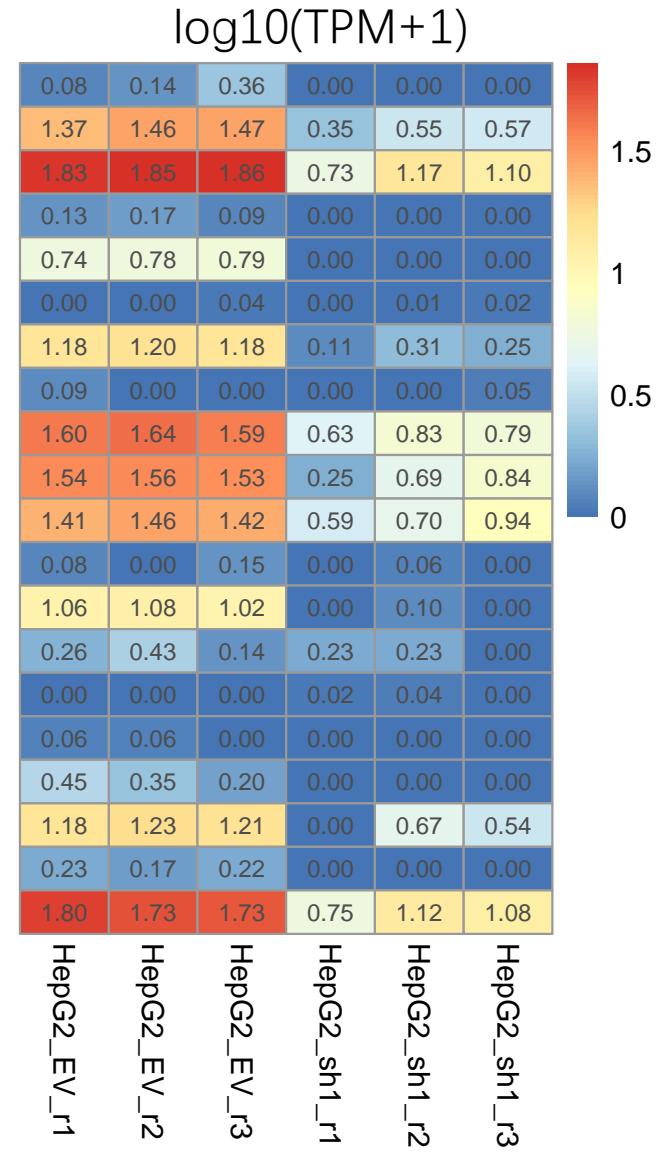
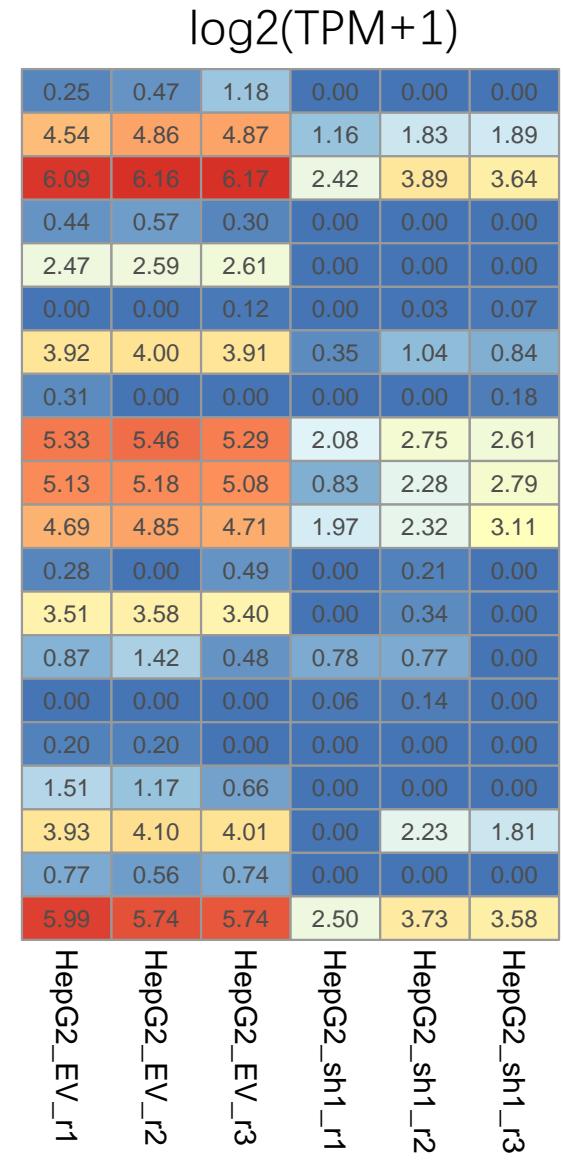
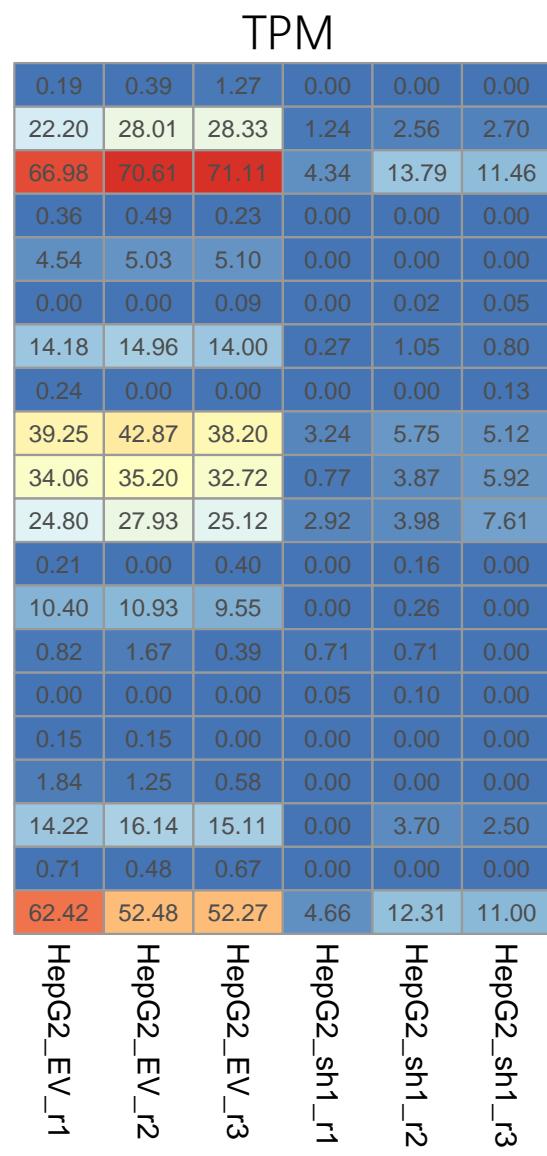


图3. Raw数值分布



TPM & log2(TPM+1)





X Axis

Y axis

Gene  
TCF4 (bHLHb19, ITF2, SEF2-1B, E2-2)  
Expression 22Q2 Public

Select type...

 Show pre-computed associationsCustom Analyses **BETA**

Get URL

Plot from CSV

Download plot data

**BETA** Use the plot to create a Cell Line Selector group

Select the lasso tool in the upper right of the plot. Circle your cell lines. Then click the "Save selected cell lines" button that appears here.

## View Options

Group by

Lineage Sub-subtype

Filter by

Select type...

Find cell lines:   Label cell lines

Click points to toggle individual labels

 Show x and y axes on the same scale

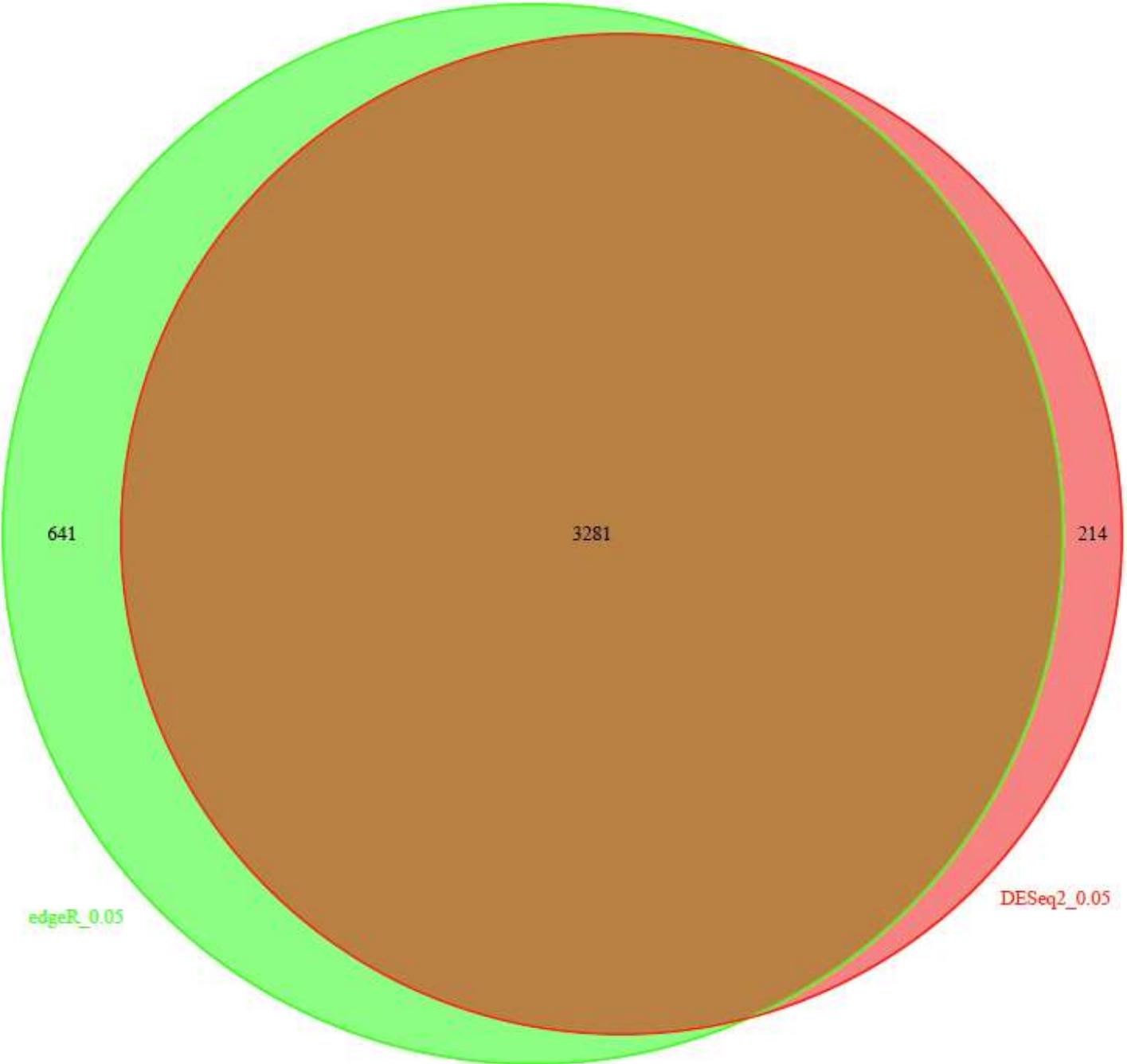
See pages for

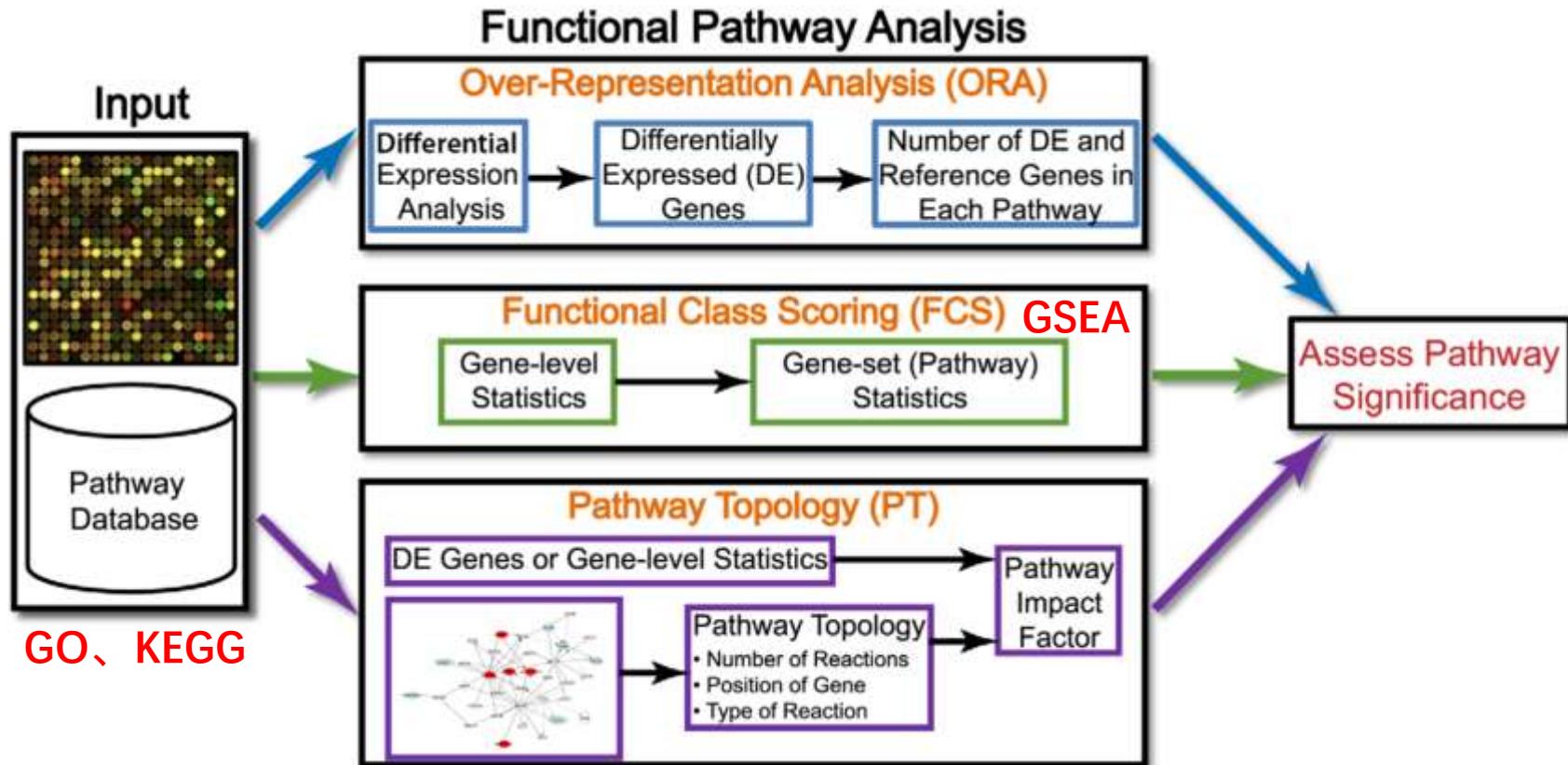
Uveal Tubular Transitional Cell Tongue Testicular T-cell Cutaneous T-cell ALCL T-cell Splenic Lymphoma Somatostatinoma Sinus Signet Ring Cell Serous Salivary Gland Carcinoma Plasmacytoma Phrynx Papillotubular Papillary Oral Oligodendroglioma Mucosal Mucinous Mixed Serous Clear Cell Mixed Endometrioid Clear Cell Med Group 3 Low Grade Serous Laryngeal Keratoacanthoma Intrahepatic Hypopharyngeal High Grade Serous HBs Antigen Carrier Glioblastoma Follicular Extrahepatic Exocrine Adenosquamous Exocrine Adenocarcinoma Endometrioid Endocervical Embryonal Diffuse Gastric Clear Cell Buccal Mucosa Blast Crisis Bladder Transitional Cell Bladder Squamous Basaloid B-cell Mantle Cell B-cell Burkitt B-cell Astrocytoma Anaplastic Amelanotic Alveolar Adrenal Acral WD/DPLS WD/PGLS No data on cell line NSCLC Squamous NSCLC Mucoepidermoid NSCLC Large Cell NSCLC Adenosquamous NSCLC Adenocarcinoma M7 M6 M5 M4 M3 M2 ERpos HER2pos ERpos HER2neg ERneg HER2pos ERneg HER2neg DLBC1 DPLS

TCF4 log<sub>2</sub>(TPM+1)  
Expression 22Q2 Public

## DESeq2和edgeR

- All are mainstream tools for RNASeq differential analysis
- All are designed based on the GLM method
- There is not much difference in the differential analysis results.

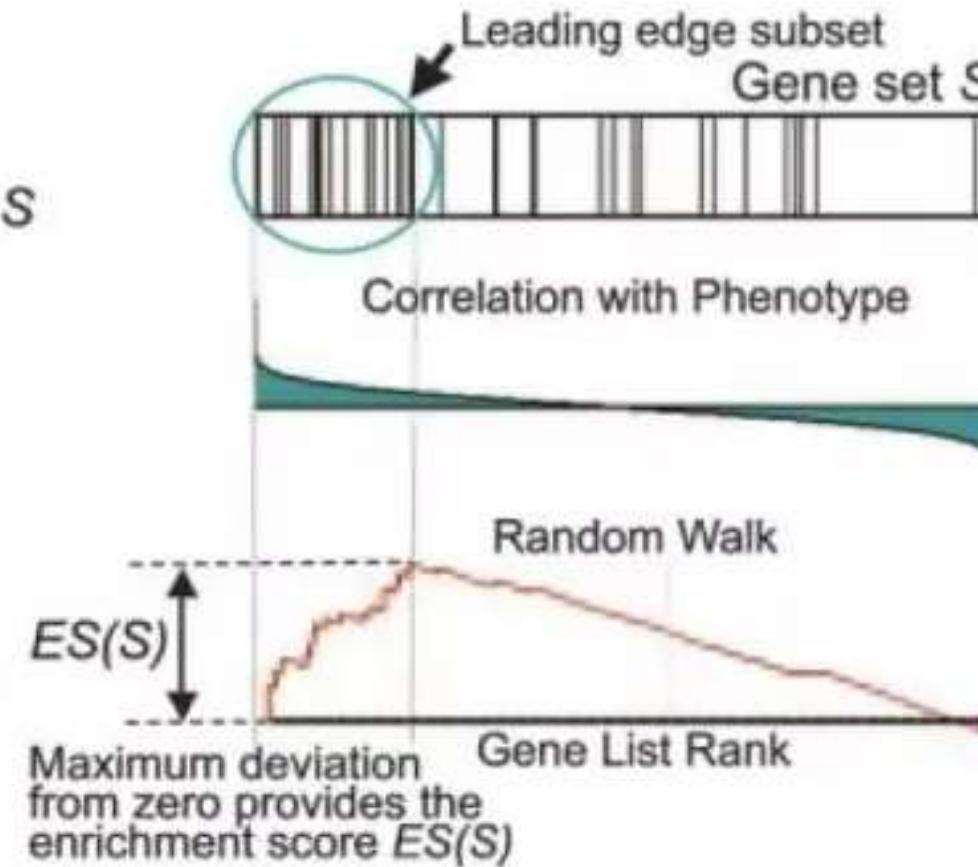
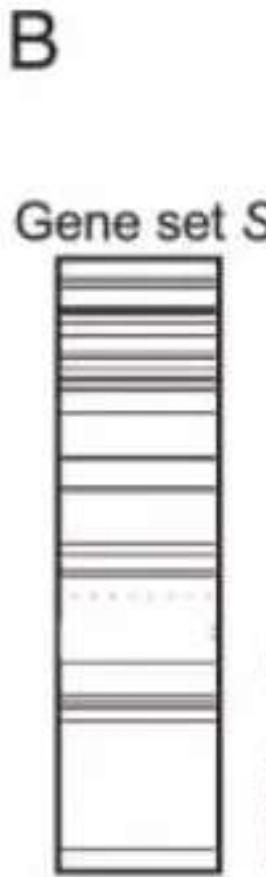
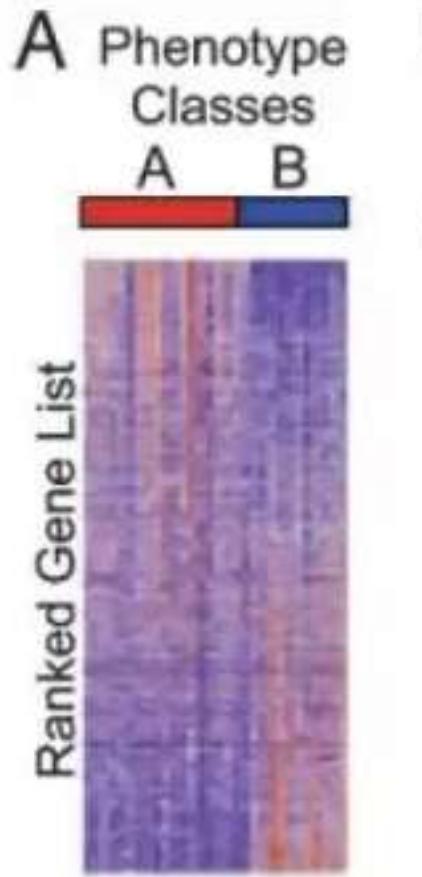




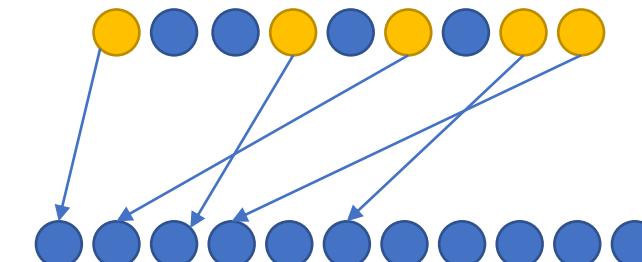
- KEGG and GO are both knowledge bases.
- ORA and GSEA are computational methods for mapping sets of interested genes to known knowledge bases.

## GSEA model

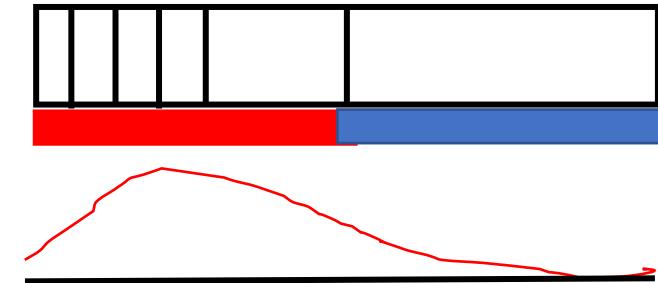
The key is the sorting criterion of the input data.



S基因集合-知识库中的基因集合



L基因集合-排好序的输入数据



Inputting expression matrix

Inputting rank of genes

GSEA 4.3.2 (Gene set enrichment analysis)

File Downloads Help

Steps in GSEA analysis

- Load data
- Run GSEA**
- Leading edge analysis
- Enrichment Map Visualization

Tools

- Run GSEAPreredranked**
- Collapse Dataset
- Chip2Chip mapping

Analysis history

GSEA reports

Processes: click 'status' field for results

	Name	Status
1	GseaPreranked	Success (with warning)
2	Gsea	Error!
3	Gsea	Success (with warning)
4	Gsea	Success (with warning)
5	Gsea	Success (with warning)
6	Gsea	Success (with warning)

Show results folder

下午7:39:29 [1669194360692] [INFO] Done all reports!!

Home Load data Run Gsea

Steps in GSEA

- What you need for GSEA
  - Expression data set
  - Phenotype annotation
  - Gene sets – use MSigDB or your own gene sets
- Run GSEA
  - Start with default parameters
  - If you want to collapse probes to genes, specify chip platform
- View results
- Leading edge analysis

Gene Set Tools

- Chip2Chip mapping
  - Convert gene sets between platforms

Getting Help

GSEA web site:  
[www.gsea-msigdb.org](http://www.gsea-msigdb.org)

Contact the GSEA team:  
[gsea-msigdb.org/gsea/contact.jsp](http://gsea-msigdb.org/gsea/contact.jsp)

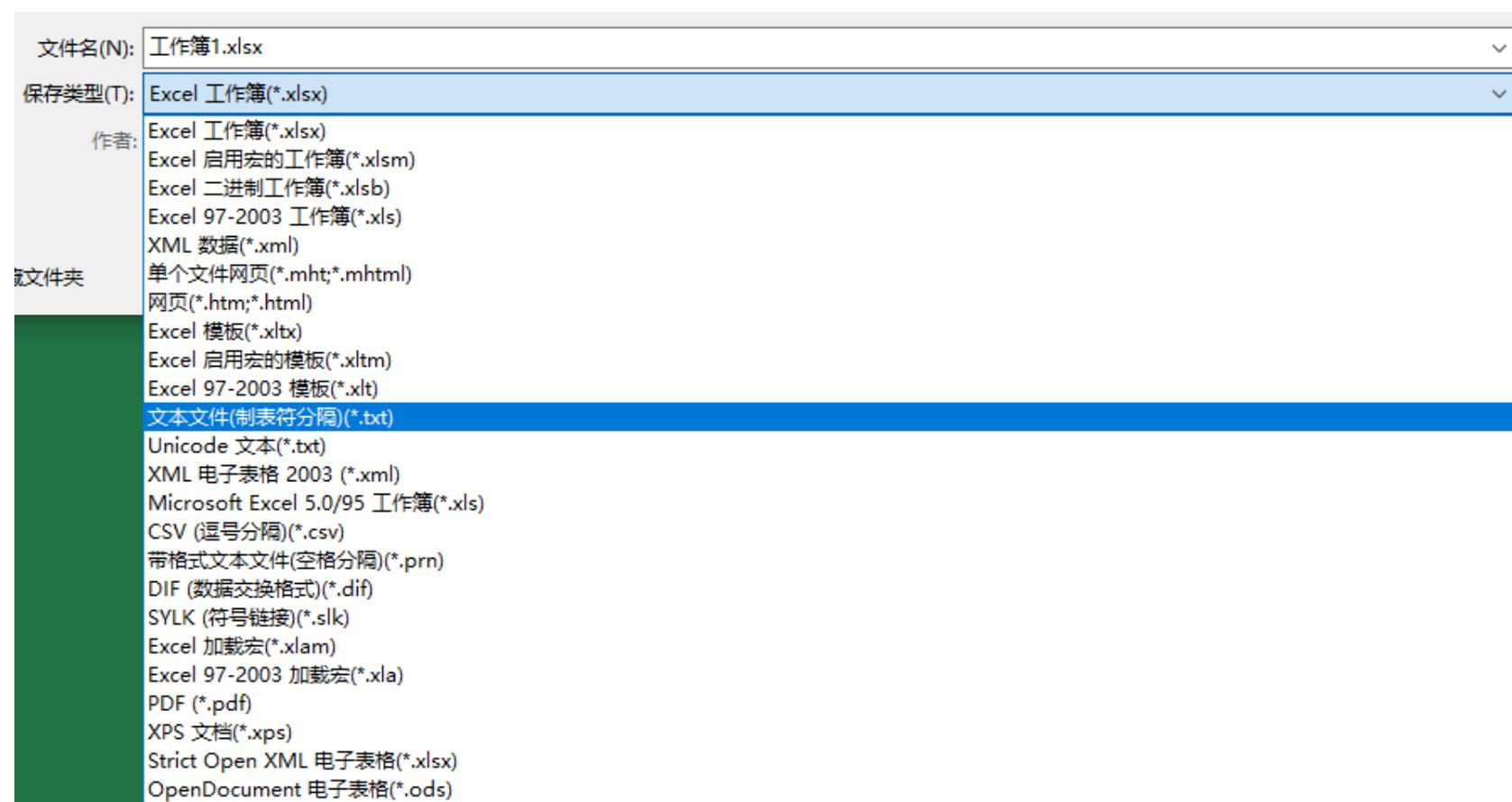
BROAD INSTITUTE

UC San Diego

719M of 2044M

A	B
#I will be ignored	
NID1	-3.489327
DUSP9	-2.2522694
FN1	-5.6054867
FGA	-1.970896
HMOX1	-4.051833
OSMR	-2.33696
COL27A1	2.83309349
A2M	-1.9030517
PDIA6	-2.2566496
PRKDC	-2.879363
GCLM	-2.7679747
TIMP3	-2.5222666
MAN1A1	-1.7256076
CCND1	-2.2045955
TNFRSF19	-2.5920401
IGF2R	-3.5870716
MMP15	-2.1646819
CSE1L	-3.4082356
TOP2A	-2.9685792
KIAA1522	-1.9169914
ITPR3	-2.0870214
PLAGL2	-2.2122335
MOGAT3	2.62207339
PABPN1	1.71785173
TTYH3	-1.5208668
DHX15	-4.4337334
NARS1	2.26780501
MT-ND3	4.50698985

The sorting criteria  
(log FC)



Save Excel file as Tab-delimited TXT

# of samples

Third column onwards  
are sample names.  
These must be UNIQUE

Always "#1.2"

The # of rows (i.e probe sets)

Data starts on line 4

Column 1: Row identifiers. Typically probe set ids or clone ids. These must be UNIQUE

Column 2: Row descriptions. Ignored by the program – can be dummy values (e.g. "na")

Each column contains expression values from 1 sample. Missing values are allowed (leave empty).

A	B	C	D	E	F	G
1 #1.2						
2 1000		130				
3 NAME	Description	DLBCL_205	DLBCL_206	DLBCL_232	DLBCL_239	DLBCL_240
4 1007_s_at	U48705 /FEATURE=mRNA /DEFINITION=HS	280.53	271.48	113.57	124.91	124.91
5 1053_at	M87338 /FEATURE=/DEFINITION=HUMA1S	32.13	91.6	117.43	41.29	33.66
6 117_at	X51757 /FEATURE=cds /DEFINITION=HSP70	51.27	61.12	24.1	41.44	43.56
7 121_at	X69699 /FEATURE=/DEFINITION=HSPAX8A	738.32	330.59	249.89	394.55	329.55
8 1255_g_at	L36861 /FEATURE=expanded_cds /DEFINITION	88.45	12.94	18.46	29.96	39
9 1294_at	L13852 /FEATURE=/DEFINITION=HUME1UP	85.57	88.06	62.24	96.59	81.01
10 1316_at	X55005 /FEATURE=mRNA /DEFINITION=HS	106.87	45.11	30.05	46.65	36.5
11 1320_at	X79510 /FEATURE=cds /DEFINITION=HSP70	58.49	27.95	17.6	27.87	26.52
12 1405_i_at	M21121 /FEATURE=/DEFINITION=HUMTCS	10.83	135.24	13.43	203.16	85.74
13 1431_at	J02843 /FEATURE=cds /DEFINITION=HUMC	41.88	24.09	16.07	26.68	25.4
14 1438_at	X75208 /FEATURE=cds /DEFINITION=HSP70	80.87	9.77	15.33	11.18	44.59
15 1487_at	L38487 /FEATURE=mRNA /DEFINITION=HUMA1S	64.26	80.61	102.9	59.77	105.72
16 1494_f_at	M33318 /FEATURE=mRNA /DEFINITION=HUMGAS	213.37	96.88	65.06	96.14	78.77
17 1598_g_at	L13720 /FEATURE=/DEFINITION=HUMGAS	458.88	215.59	186.72	187.36	237.69
18 160020_at	Z48481 /FEATURE=cds /DEFINITION=HSMM	411.94	171.16	130	234.76	266.96
19 1729_at	L41690 /FEATURE=/DEFINITION=HUMTRAD	81.59	83.94	74.75	110.9	126.98
20 1773_at	L00635 /FEATURE=/DEFINITION=HUMFPTE	62.82	45.96	41.15	23.1	28.41
21 177_at	U38545 /FEATURE=/DEFINITION=HSU3854	57.04	28.05	16.74	29.66	53.29
22 179_at	U38980 /FEATURE=/DEFINITION=U38980_H	333.96	254.15	241.24	350.58	13.53

What values represent the input data?

If editing, in Excel, make sure to save your data as "tab delimited text"

(1)  $\frac{\mu_A - \mu_B}{\sigma_A + \sigma_B}$  (2)  $\frac{\mu_A - \mu_B}{\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}}$  (3)  $\frac{\mu_A}{\mu_B}$  (4)  $\mu_A - \mu_B$  (5)  $\log_2 \left( \frac{\mu_A}{\mu_B} \right)$

signal2noise t-Test ratio\_of\_class diff\_of\_class log2\_ratio\_of\_class  
 (Log FC)

**Basic fields****Analysis name**

HepG2sh1\_TMM\_ssh1\_vs\_EV

**Hide****Enrichment statistic**

weighted

**Metric for ranking genes**

Signal2Noise

**Gene list sorting mode**

Signal2Noise

**Gene list ordering mode**

tTest

**Max size: exclude larger sets**

Cosine

**Min size: exclude smaller sets**

Euclidean

**Save results in this folder**

Manhattan

Pearson

Spearman

Ratio\_of\_Classes

	A	B	C	D
1	Gene Name	S1 counts	S2 counts	S3 Counts
2	A (2kb)	10	12	30
3	B (4kb)	20	25	60
4	C (1kb)	5	8	15
5	D (10kb)	0	0	1
6	lib size	35	45	106

**Advanced fields****Show**

Reset

Last

Command

Run

- ◆ **CPM**: Replication of the same sample group, not used for within-sample comparisons or DE analysis.
  - ◆ **FPKM**: Comparison of gene counts within samples; not suitable for between-sample comparisons or DE analysis.
  - ◆ **TPM**: Comparison of gene counts within samples or between samples of the same group; not used for DE analysis.
  - ◆ **DESeq2**: Comparison between samples and DE analysis; not suitable for within-sample comparisons.
  - ◆ **edgeR (TMM)**: Comparison of gene counts between samples and within samples, used for DE analysis.
- 
- **Note:** ssGSEA (single-sample GSEA) projections perform substantially different mathematical operations from standard GSEA. For the ssGSEA implementation, gene-level summed TPM serves as an appropriate metric for analysis of RNA-seq quantifications.
  - Normalization methods (such as, TMM, geometric mean) which operate on raw counts data should be applied prior to running GSEA (最推荐)
  - Tools such as DESeq2 can be made to produce properly normalized data (normalized counts) which are compatible with GSEA. Normalized counts也可以

[Kasumi\\_PWsh1\\_kegg\\_gsea.csv](#)

[Kasumi\\_PWsh1\\_KEGG\\_ORA.csv](#)

[Kasumi\\_PWsh4\\_kegg\\_gsea.csv](#)

[Kasumi\\_PWsh4\\_KEGG\\_ORA.csv](#)

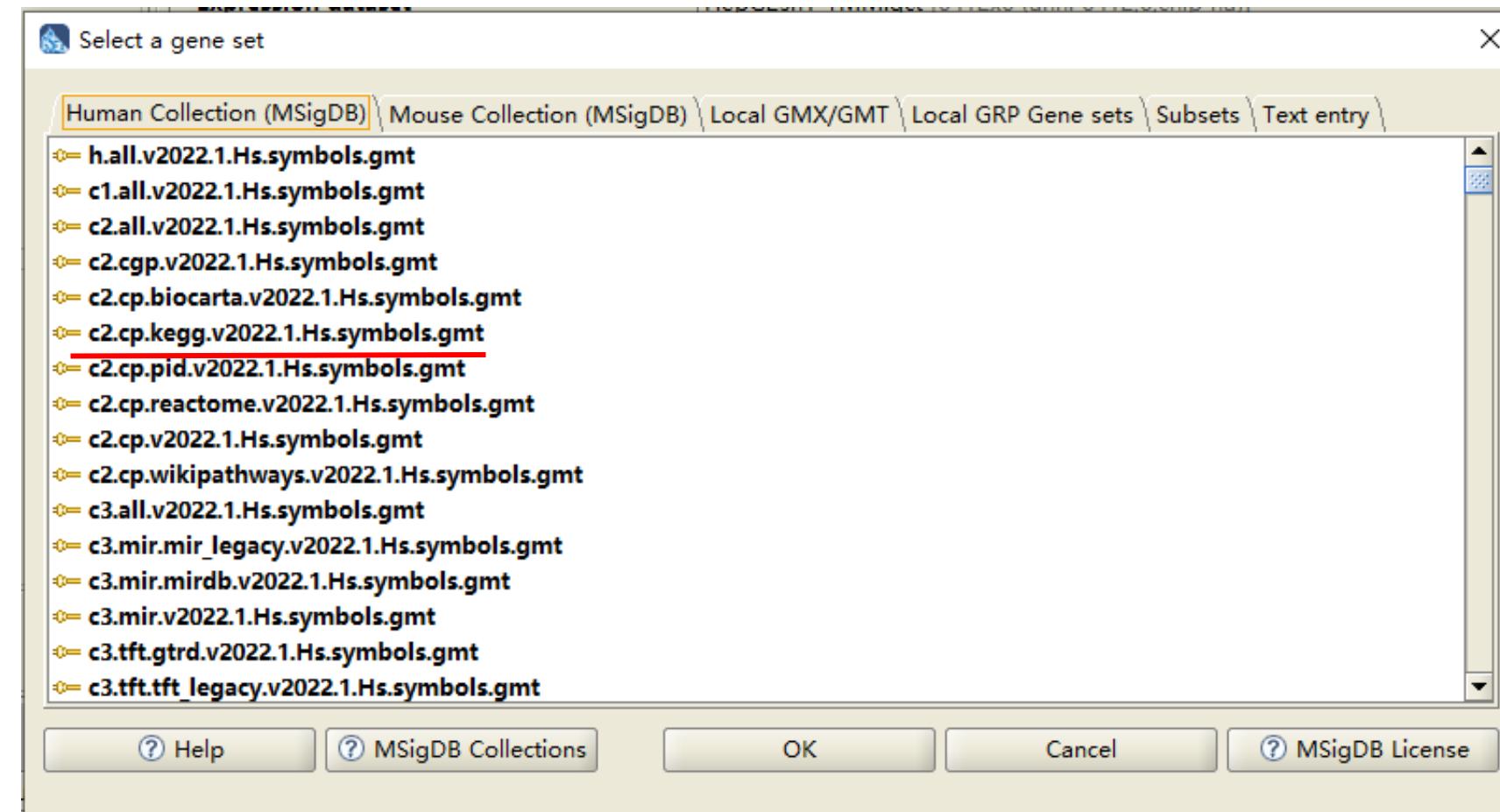
[Kasumi\\_PWsh1\\_go\\_gsea.csv](#)

[Kasumi\\_PWsh1\\_go\\_ORA.csv](#)

[Kasumi\\_PWsh4\\_go\\_gsea.csv](#)

[Kasumi\\_PWsh4\\_go\\_ORA.csv](#)

The analysis using the R  
package Clusterprofiler  
yields results somewhat  
different from those obtained  
with the GSEA software.



[Back to Content](#)

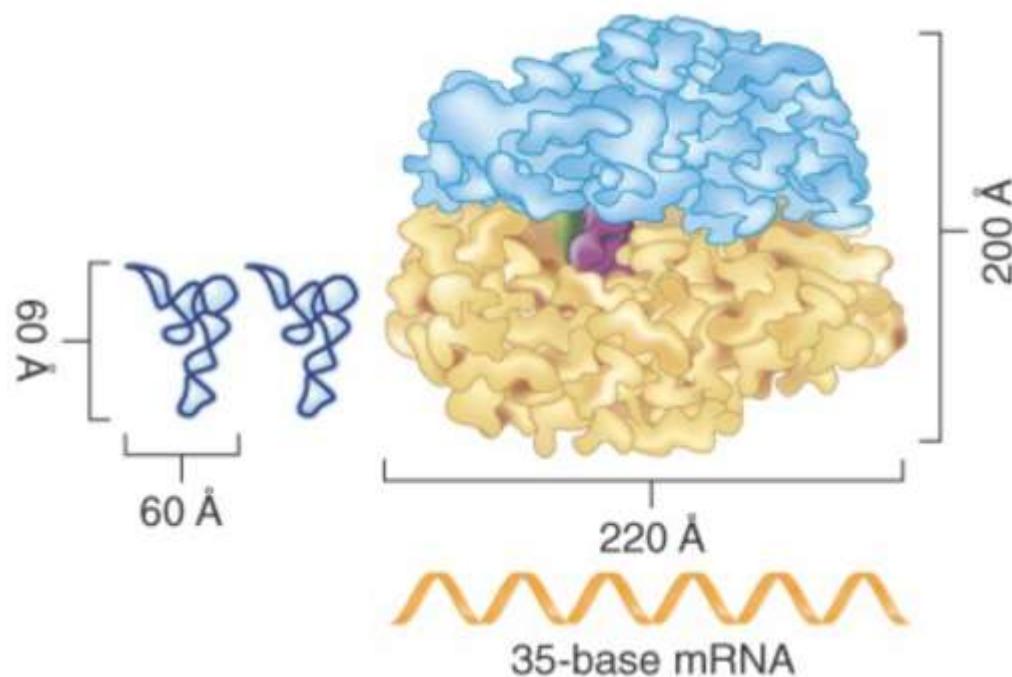
# **Computation for Ribo-seq Data in Ribosome Profiling**

ZHOU YING

7<sup>th</sup> April 2023

- Reviewing basic mechanism of translation process
- Ribosome profiling methods and ribo-seq
- Main working flow of ribo-seq data analysis
- **Discussion ORFs detection tools and differential ORFs analysis  
(Ribo-TISH, Ribocode)**
- **Conclusion and Next plan (Cancer neuroscience)**

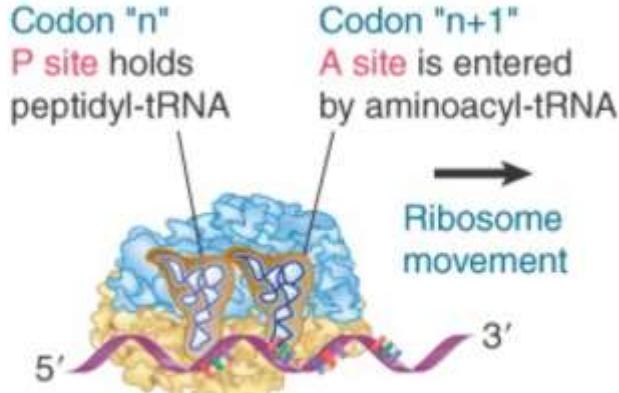
# Translation Process



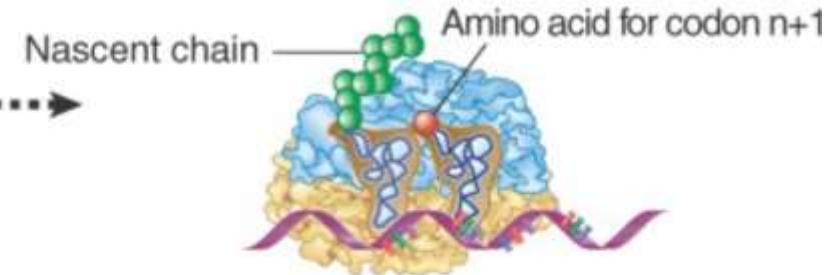
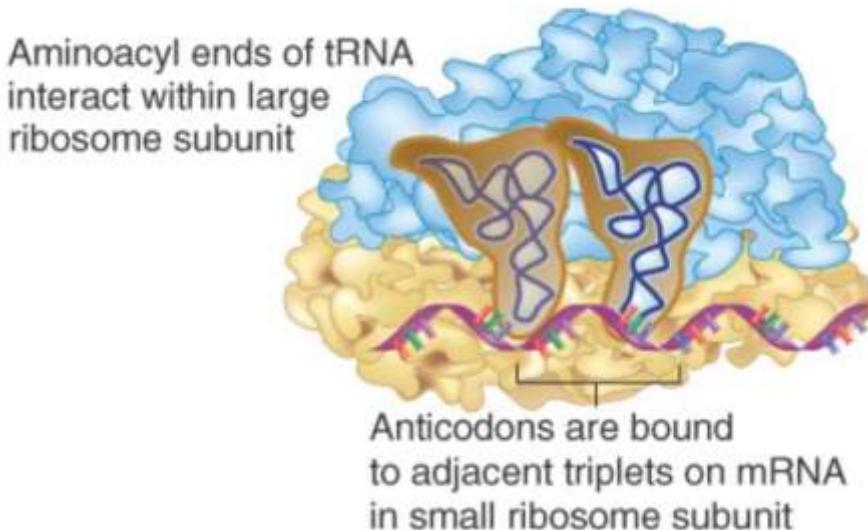
Ribosomes	rRNAs	r-proteins
Bacterial (70S) mass: 2.5 MDa 66% RNA	50S 30S	23S = 2,904 bases 5S = 120 bases 31
Mammalian (80S) mass: 4.2 MDa 60% RNA	60S 40S	16S = 1,542 bases 28S = 4,718 bases 5.8S = 160 bases 5S = 120 bases 21 49
		18S = 1,874 bases 33

Lewin's genes XII

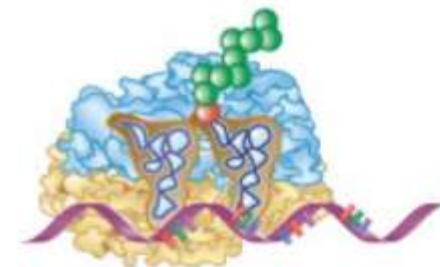
# Translation Process



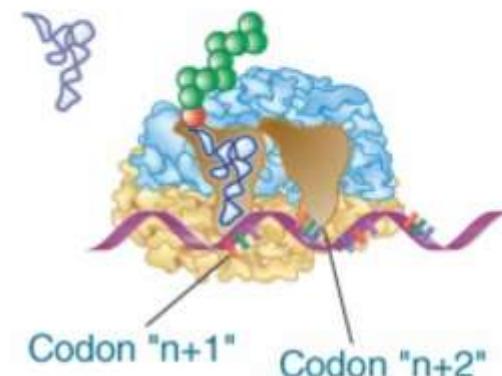
1 Before peptide bond formation peptidyl-tRNA occupies P site; aminoacyl-tRNA occupies A site



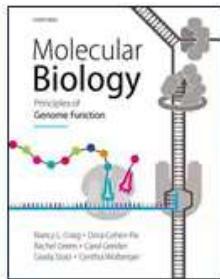
2 Peptide bond formation polypeptide is transferred from peptidyl-tRNA in P site to aminoacyl-tRNA in A site



3 Translocation moves ribosome one codon; places peptidyl-tRNA in P site; deacylated tRNA leaves via E site; A site is empty for next aa-tRNA



# Translation Process



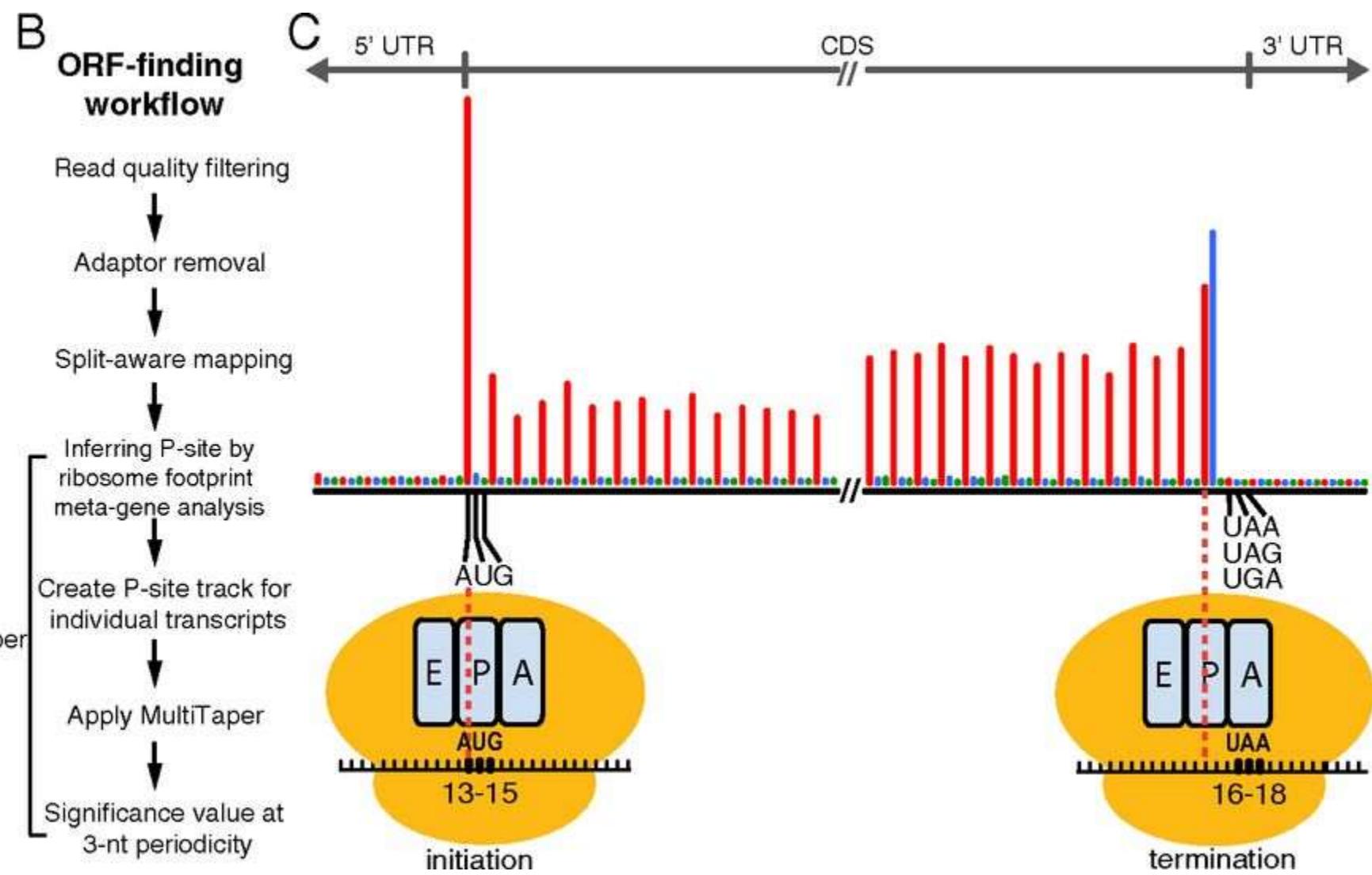
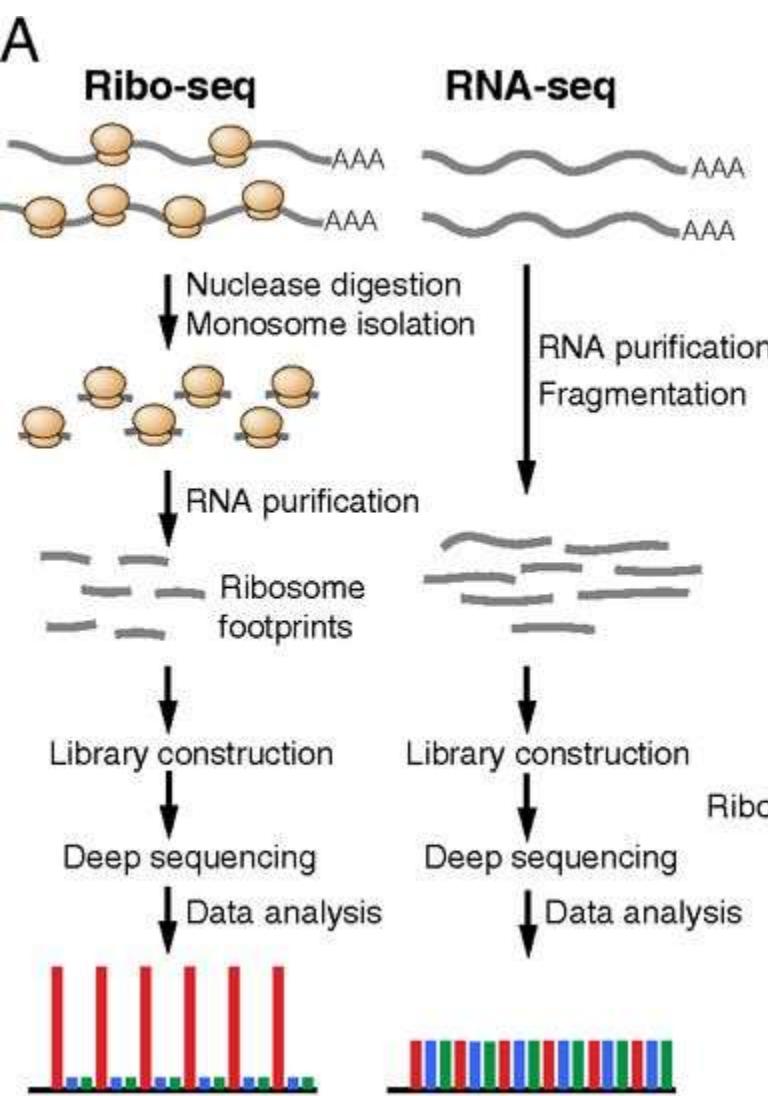
**Molecular Biology: Principles of  
Genome Function**  
Second Edition



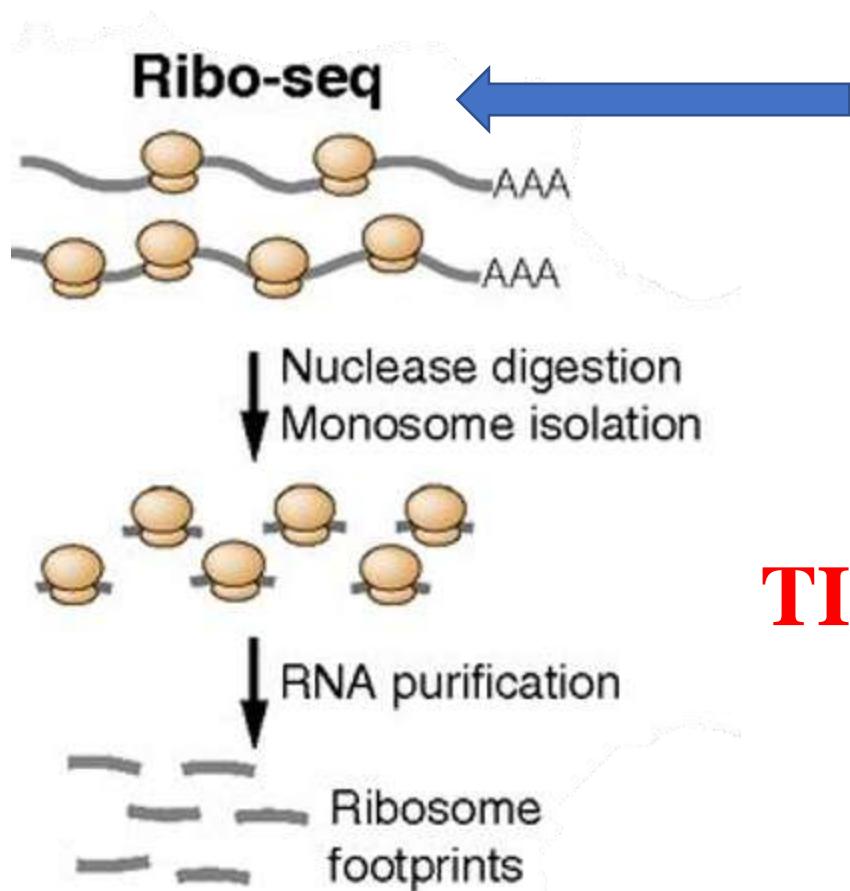
---

## Animation 8: Translation

# Ribo-seq



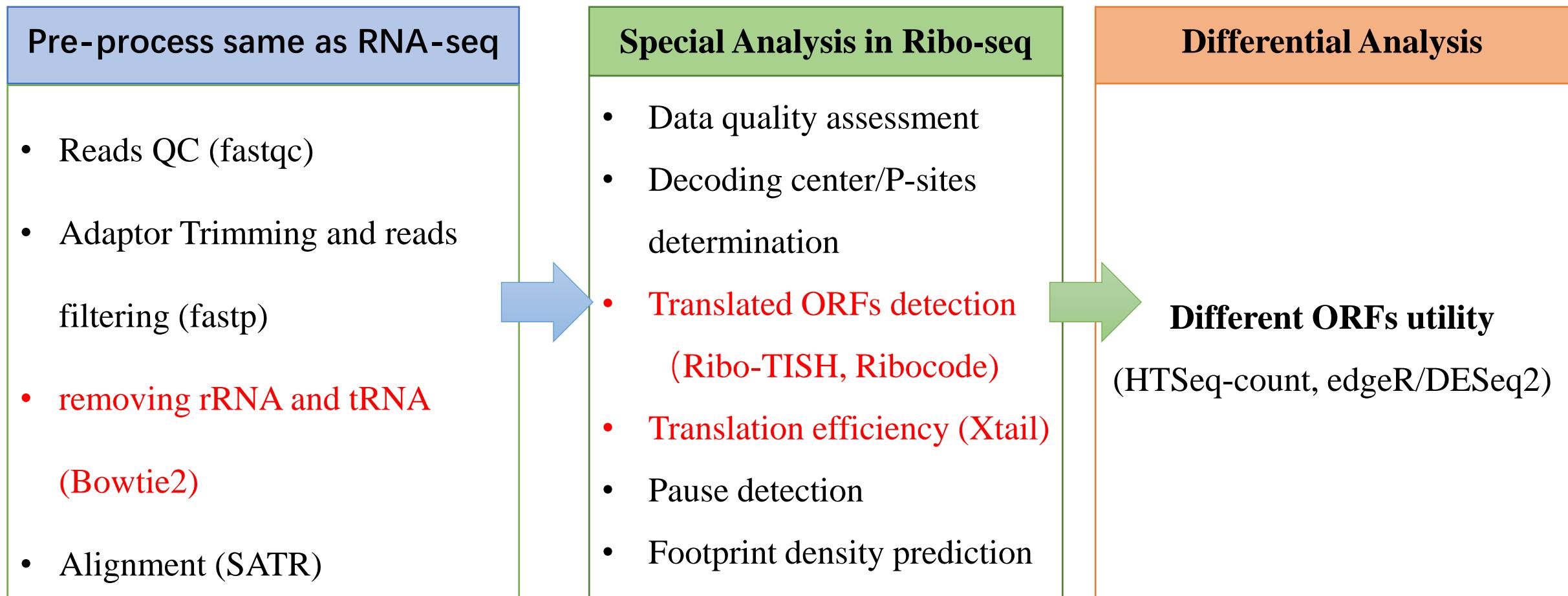
# TI-seq



- Cycloheximide (CHX): 延伸抑制剂, 针对所有核糖体
- lactimidomycin (LTM) and harringtonine (Harr): 对起始核糖体有更强的作用

**TI-seq: Translation initiation sequencing**

# Main working flow of ribo-seq data analysis



# Ribocode

Published online 10 March 2018

Nucleic Acids Research, 2018, Vol. 46, No. 10 e61  
doi: 10.1093/nar/gky179

## De novo annotation and characterization of the translatome with ribosome profiling data

Zhengtao Xiao<sup>1,2,3</sup>, Rongyao Huang<sup>1,2,3</sup>, Xudong Xing<sup>1,2,3,4</sup>, Yuling Chen<sup>1,2,3</sup>,  
Haiteng Deng<sup>1,2,3</sup> and Xuerui Yang<sup>1,2,3,\*</sup>

<sup>1</sup>MOE Key Laboratory of Bioinformatics, Tsinghua University, Beijing 100084, China, <sup>2</sup>Center for Synthetic & Systems Biology, Tsinghua University, Beijing 100084, China, <sup>3</sup>School of Life Sciences, Tsinghua University, Beijing 100084, China and <sup>4</sup>Joint Graduate Program of Peking-Tsinghua-National Institute of Biological Science, Tsinghua University, Beijing 100084, China

Prepare annotation file  
(*prepare\_transcripts*)

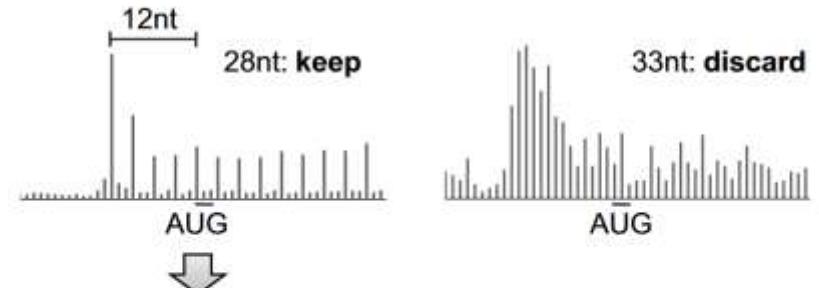
P-site determination  
(*metaplots*)

ORFs calling  
(*RiboCode*)

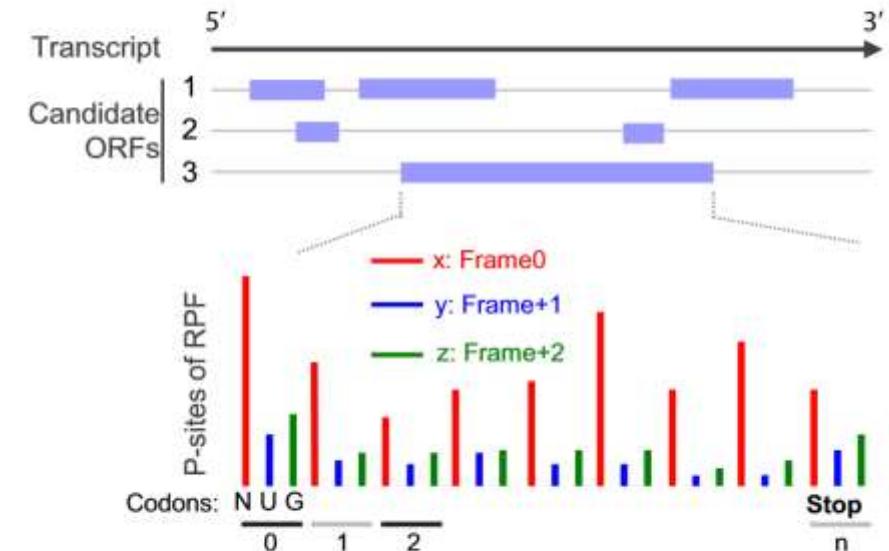
ORFs counting  
(*ORFcount*)

Meta-gene analysis for lengths and P-sites of RPFs

5' end of the RPF reads on annotated start codons



Allocation of the RPFs on the putative codons



$$\vec{F}_1: \{y_0, y_1, y_2, \dots, y_n\} \quad \vec{F}_0: \{x_0, x_1, x_2, \dots, x_n\} \quad \vec{F}_2: \{z_0, z_1, z_2, \dots, z_n\}$$

Statistical assessment of the 3-nt periodicity with modified Wilcoxon signed-rank test

$$\begin{array}{c} P\text{-value } (\vec{F}_0 > \vec{F}_1) \xrightarrow{\text{Stouffer's method}} P\text{-value} \\ P\text{-value } (\vec{F}_0 > \vec{F}_2) \end{array}$$

Prepare annotation file  
(*prepare\_transcripts*)

```
(ribo-seq) [zhouying@serverf ~]
└─> prepare_transcripts -g /alldata/zhouying/ref_data/anno/GRCh38.p13/gencode.v36.annotation.gtf -f /alldata/zhouying/ref_data/genome/GRCh38.p13/GRCh38.p13.genome.fa -o /alldata/zhouying/ref_data/RiboCode_anno
```

P-site determination  
(*metaplots*)

```
(ribo-seq) [zhouying@serverf ~]
└─> ls /alldata/zhouying/ref_data/RiboCode_anno
transcripts_cds.txt      transcripts_sequence.fa.flat
transcripts.pickle        transcripts_sequence.fa.gdx
transcripts_sequence.fa
```

ORFs calling  
(*RiboCode*)

ORFs counting  
(*ORFcount*)

**Input:** annotation.gtf, genome.fa

**Output:** A folder provided for subsequent analysis commands.

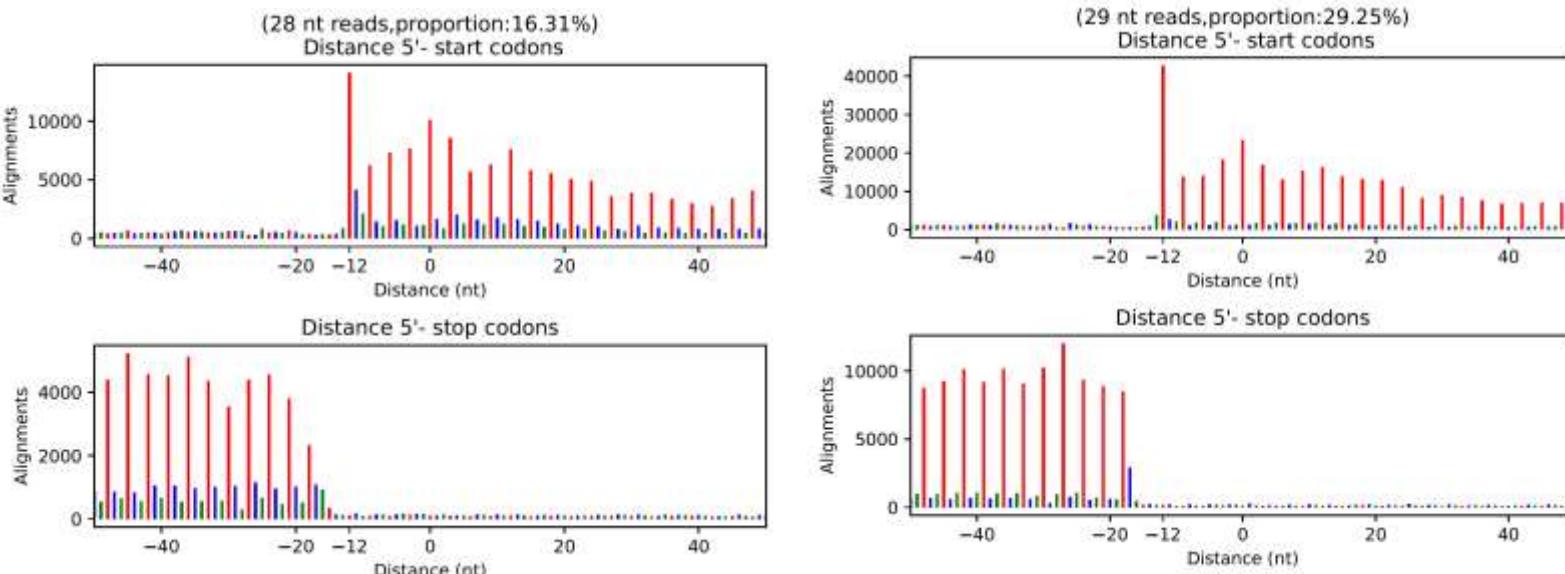
```
metaplots -a ${RiboCode_annot} \
    -r ${indir}/${id}.RiboAligned.toTranscriptome.out.bam \
    -o ${OUT}/metaplots/${id}_metaplots \
    -m 26 -M 50 -s yes -pv1 1 -pv2 1
```

Prepare annotation file  
(*prepare\_transcripts*)

```
#/alldata/zhouying/Ribo-seq/ribo-TISH_test_data/05_align_star/SRR5227449_RiboAligned.toTranscriptome.out.bam
#read_length proportion(per total mapped reads) predicted_psites f0_sum f1_sum f2_sum f0_percent pvalue1
#          pvalue2 pvalue_combined
# 28 16.31% 12 109682 25496 16285 72.42% 0.00014623915190606628 0.00014652626009624466 1.513625849539642e-07
# 29 29.25% 12 258422 19393 23780 85.69% 0.00014652626009624466 0.00014652626009624466 1.516508550767744e-07
# 30 17.82% 12 135746 7667 35284 76.00% 0.00014652626009624466 0.00014652626009624466 1.516508550767744e-07

# SampleName AlignmentFile Stranded(yes/reverse) P-siteReadLength P-siteLocations
SRR5227449_RiboAligned.toTranscriptome.out /alldata/zhouying/Ribo-seq/ribo-TISH_test_data/05_align_star/SRR5227449
_RiboAligned.toTranscriptome.out.bam yes 28,29,30 12,12,12
```

P-site determination  
(*metaplots*)



ORFs calling  
(*RiboCode*)

ORFs counting  
(*ORFcount*)

```
echo ${id}' ORF calling ...'  
config=${OUT}/metaplots/${id}_metaplots_pre_config.txt
```

```
RiboCode -a ${RiboCode_annot} -l no -g -c ${config} -o ${OUT}/ORF/${id}_ORF
```

Prepare annotation file  
(*prepare\_transcripts*)

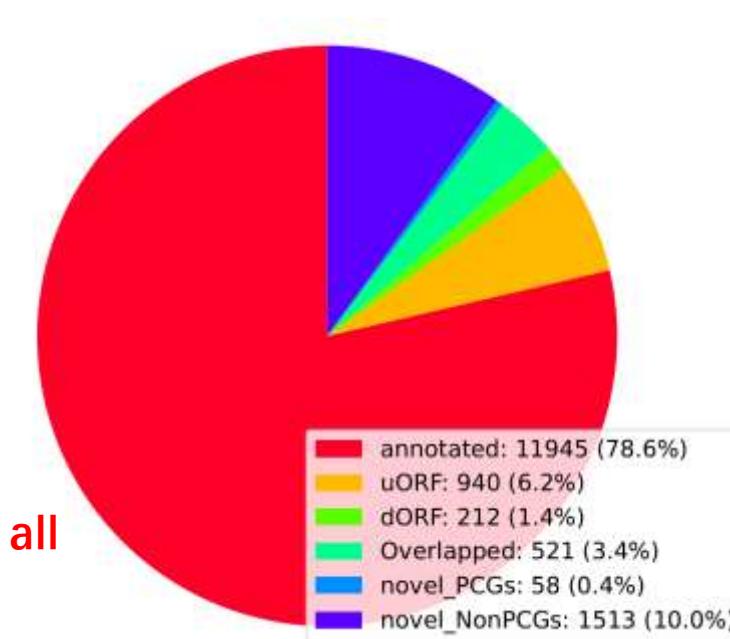
P-site determination  
(*metaplots*)

ORFs calling  
(*RiboCode*)

ORFs counting  
(*ORFcount*)

📄 SRR6327777_ORF.gtf	2023/4/5 12:17
📄 SRR6327777_ORF.txt	2023/4/5 12:17
📄 SRR6327777_ORF_collapsed.gtf	2023/4/5 12:17
📄 SRR6327777_ORF_collapsed.txt	2023/4/5 12:17
📝 SRR6327777_ORF_ORFs_category.pdf	2023/4/5 12:17

- **ORF.gtf/ORF.txt:** contains the information of **all predicted ORFs** in each transcript.
- **ORF\_collapsed.gtf /RF\_collapsed.txt:** **file combines the ORFs having the same stop codon in different transcript isoforms:** the one harboring the most upstream in-frame ATG will be kept.



ORF_ID	ENSG00000279457.4_187832_185217_264
ORF_type	novel
transcript_id	ENST00000623083.4
transcript_type	unprocessed_pseudogene
gene_id	ENSG00000279457.4
gene_name	WASH9P
gene_type	unprocessed_pseudogene
chrom	chr1
strand	-
ORF_length	792
ORF_tstart	603
ORF_tstop	1397
ORF_gstart	187832
ORF_gstop	185217
annotated_tstart	None
annotated_tstop	None
annotated_gstart	None
annotated_gstop	None
Psites_sum_frame0	8
Psites_sum_frame1	0
Psites_sum_frame2	1
Psites_coverage_frame0	1.89%
Psites_coverage_frame1	0.00%
Psites_coverage_frame2	0.38%
Psites_frame0_RPKM	1.150308947
pval_frame0_vs_frame1	0.019679754
pval_frame0_vs_frame2	0.042234516
pval_combined	0.003715148
AAseq	MLGAETEEKLFDAPLSISKREQLEQQVPENYFYVPDLGQVPEIDVPSYLPDLGIANDLMYIADLGPGIAPSAPGTIPELPTFHTVAEPLKVDLQDGVLTPPPPPPPPAPE VLASAPPLPPSTAAPVGQGARQDDSSSASPVSQGAPREVVDPSGGRATLLESIRQAGGIGKAKLRSMKERKLEKKQKEQVRATSQGGHLMSDLNFNLKVMRRKGIS GKGPGAGEGPGGAFARVSDSIPPLPPPQQPQAEEDEDWDWE

Prepare annotation file  
(*prepare\_transcripts*)

```
ORFcount -g ${OUT}/ORF/${id} ORF.atf \
-r ${indir}/${id}_RiboAligned.sortedByCoord.out.bam \
-f 15 -l 5 -e 100 -m 24 -M 35 -s yes \
-o ${OUT}/ORF_count/${id}_ORF_count_all.txt
```

P-site determination  
(*metaplots*)

ORFs calling  
(*RiboCode*)

ORFs counting  
(*ORFcount*)

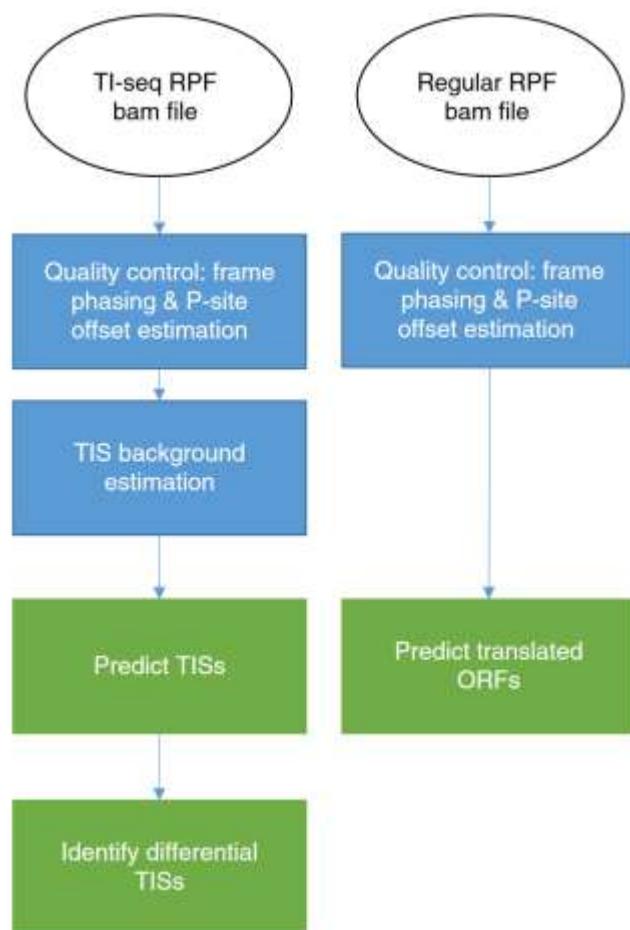
Samples ORF count Matrix

ORF Differential analysis  
(*DESeq2/edgeR*)

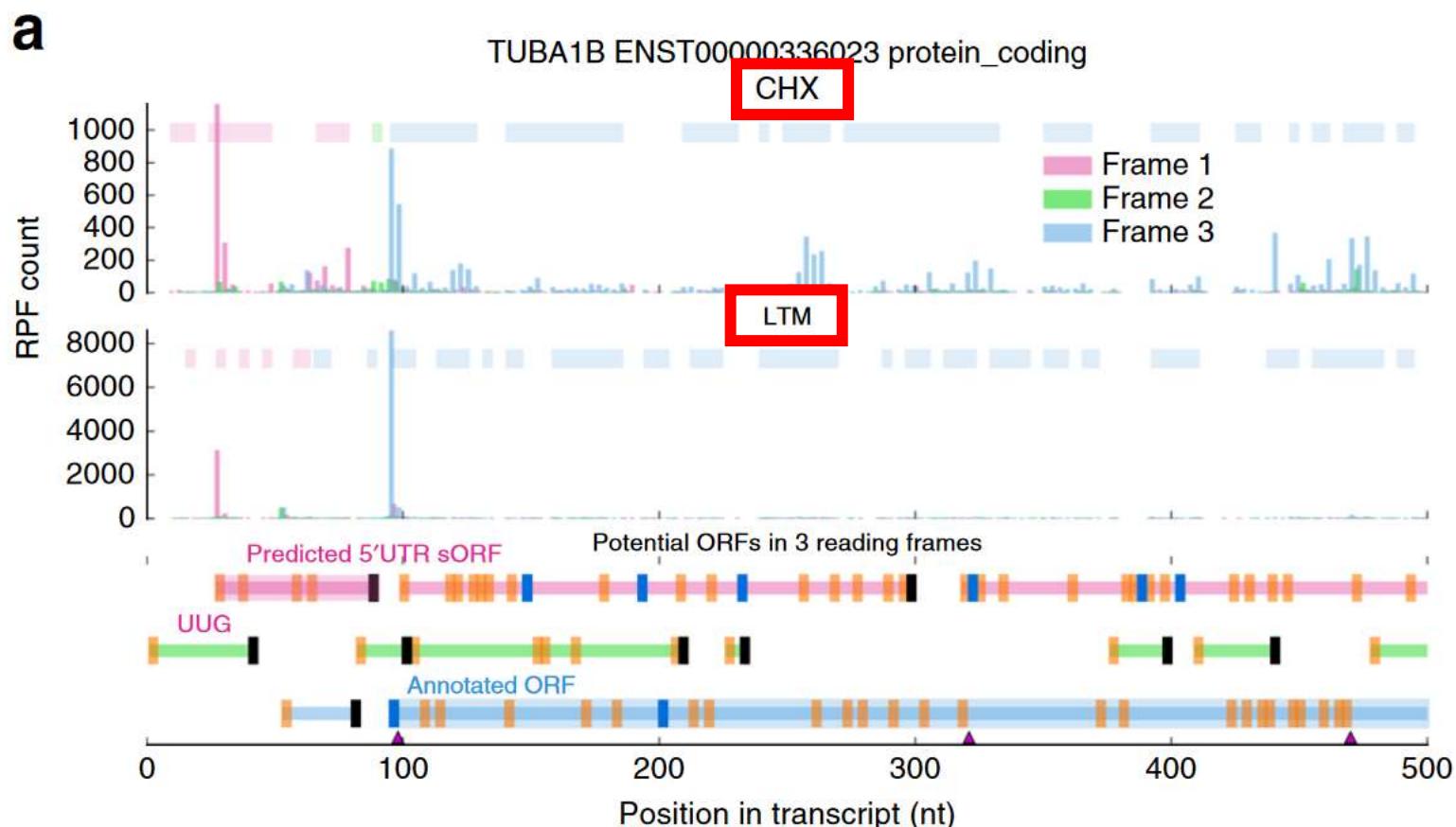
ENSG00000000003.15_100635569_100630798_129	0
ENSG00000000003.15_100636694_100630798_245	572
ENSG00000000419.13_50936195_50935132_50	0
ENSG00000000419.13_50945903_50935132_190	4
ENSG00000000419.13_50958523_50935132_260	0
ENSG00000000419.13_50958523_50935132_287	0
ENSG00000000419.13_50958523_50940883_188	5
ENSG00000000457.14_169888840_169853713_688	0
ENSG00000000457.14_169888840_169853713_742	1
ENSG00000000457.14_169888840_169868986_255	0
ENSG00000000460.17_169795103_169802640_56	0
ENSG00000000460.17_169795103_169804168_180	5
ENSG00000000460.17_169800906_169804168_122	0
ENSG00000000460.17_169800906_169804248_161	0
ENSG00000000460.17_169800906_169827109_255	0
ENSG00000000460.17_169800906_169852987_853	0
ENSG00000000460.17_169804168_169804248_26	0
ENSG00000000460.17_169804168_169850459_608	0
ENSG00000000460.17_169804168_169852987_718	0
ENSG00000000460.17_169827172_169852987_530	0
ENSG00000000460.17_169850534_169852987_116	0
ENSG00000001036.14_143502083_143497359_96	3
ENSG00000001036.14_143511634_143495707_467	403
ENSG00000001036.14_143511634_143502037_165	0
ENSG00000001084.13_53500114_53498756_93	0
ENSG00000001084.13_53500276_53500166_36	9
ENSG00000001084.13_53500501_53500415_28	0
ENSG00000001084.13_53505868_53505308_84	0
ENSG00000001084.13_53508711_53498756_361	0
ENSG00000001084.13_53522518_53498756_584	0

# Genome-wide identification and differential analysis of translational initiation

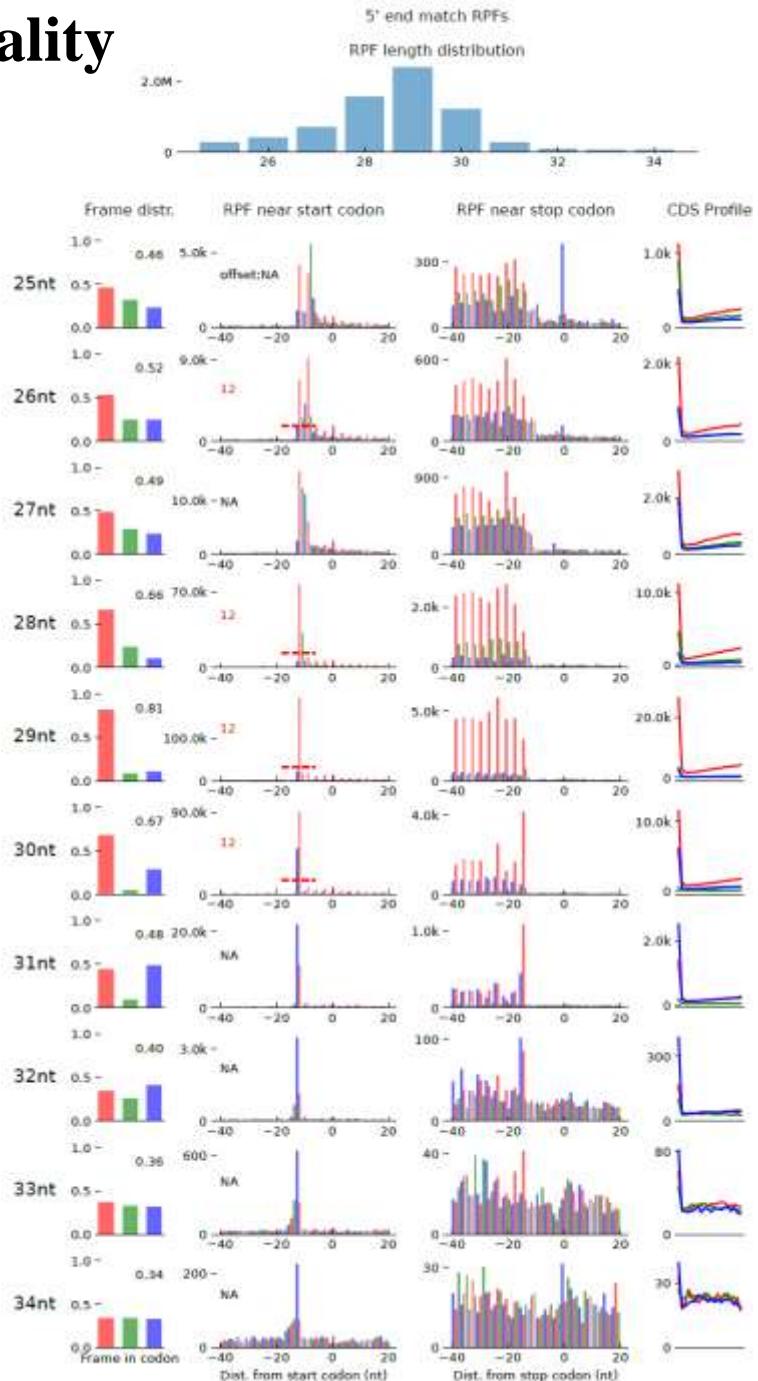
Peng Zhang<sup>1</sup>, Dandan He<sup>1</sup>, Yi Xu<sup>1</sup>, Jiakai Hou<sup>1</sup>, Bih-Fang Pan<sup>2</sup>, Yunfei Wang<sup>1</sup>, Tao Liu<sup>1,3</sup>, Christel M. Davis<sup>4</sup>, Erik A. Ehli<sup>4</sup>, Lin Tan<sup>1</sup>, Feng Zhou<sup>5</sup>, Jian Hu<sup>6</sup>, Yonghao Yu<sup>7</sup>, Xi Chen<sup>8</sup>, Tuan M. Nguyen<sup>8,9</sup>, Jeffrey M. Rosen<sup>8</sup>, David H. Hawke<sup>2</sup>, Zhe Ji<sup>10,11</sup> & Yiwen Chen<sup>1</sup>



Mainly aimed at the combined analysis of TI-seq and Ribo-seq.



# quality



# Predict

<b>Gid</b>	ENSG00000187634.12
<b>Tid</b>	ENST00000420190.6
<b>Symbol</b>	SAMD11
<b>GeneType</b>	protein_coding
<b>GenomePos</b>	chr1:924431-939291:+
<b>StartCodon</b>	ATG
<b>Start</b>	504
<b>Stop</b>	1578
<b>TisType</b>	Annotated
<b>TISGroup</b>	0
<b>TISCounts</b>	0
<b>TISPvalue</b>	None
<b>RiboPvalue</b>	3.04E-05
<b>RiboPStatus</b>	N
<b>FisherPvalue</b>	None
<b>TISQvalue</b>	None
<b>FrameQvalue</b>	0.000314773
<b>FisherQvalue</b>	None
<b>AALen</b>	358

## Summary of Ribo-seq

- Ribocode is more suitable for performing differential analysis of ribo-seq ORFs than ribo-TISH.
- If there is TI-seq data input into ribo-TISH to obtain TISCounts data, it can also be extracted and organized into a count matrix for differential analysis using DESeq2/edgeR.
- Ribo-seq already has mature analysis tools and workflows, so there is no need to develop a new toolkit based on current lack of research demand.

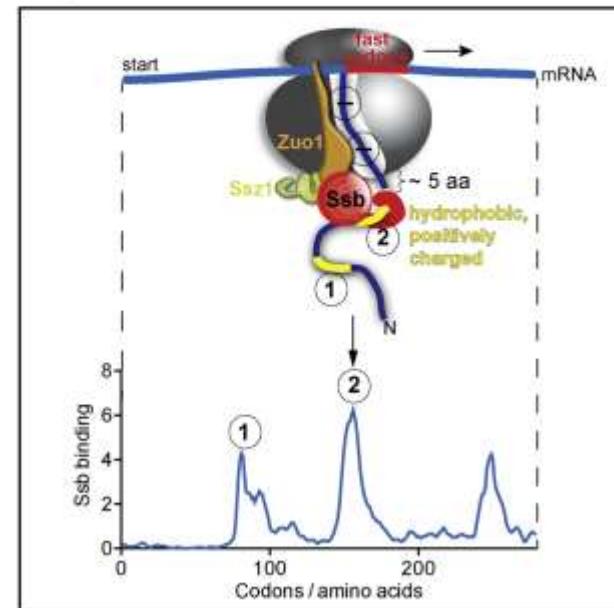
# Other direction and application



## Cell

### Profiling Ssb-Nascent Chain Interactions Reveals Principles of Hsp70-Assisted Folding

#### Graphical Abstract



#### Authors

Kristina Döring, Nabeel Ahmed, Trine Riemer, ..., Edward P. O'Brien, Günter Kramer, Bernd Bukau

#### Correspondence

g.kramer@zmbh.uni-heidelberg.de (G.K.), bukau@zmbh.uni-heidelberg.de (B.B.)

#### In Brief

Examining where and how a yeast chaperone binds on nascent peptides reveals substrate specificity and principles of co-translational protein folding.

# Conclusion

- **Transcriptome:** RNA-seq (differential analysis, alternative splicing, AS)
- **Epigenome:** ChIP-seq, CUT&TAG/CUT&RUN, [ATAC-seq, WGBS, Hi-C]
- **Translatome:** Ribo-seq
- It's not a problem to learn to run an analysis pipeline.
- It requires exercising the ability to integrate data to solve biological problems.
- I don't want to abandon my background in neurobiology.
- According to the department's research direction, I plan to work on a project with a background in **cancer neuroscience**.

## Perspective

**Cancer hallmarks intersect with neuroscience in the tumor microenvironment**Douglas Hanahan<sup>1,2,3,4,\*</sup> and Michelle Monje<sup>5,6,\*</sup>

1 Cancer Research, 1011 Lausanne, Switzerland

2 Lausanne, Switzerland

3 Research (ISREC), School of Life Sciences, Swiss Federal Institute of Technology Lausanne (EPFL),

4 Lausanne, Switzerland

5 Medical Sciences, Stanford University, Stanford, Stanford, CA, USA

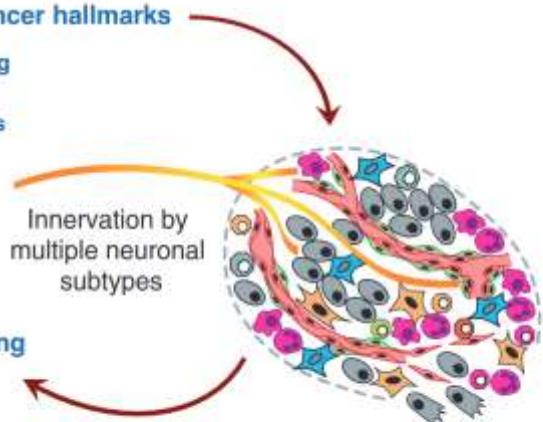
6 Medicine, Stanford University, Stanford, CA, USA

✉ [d.hanahan@epfl.ch](mailto:d.hanahan@epfl.ch) (D.H.), [mmonje@stanford.edu](mailto:mmonje@stanford.edu) (M.M.)

2

## Neural interactions facilitate cancer hallmarks

- Sustaining proliferative signaling
- Resisting cell death
- Activating invasion & metastasis
- Tumor-promoting inflammation



## Cancer-induced neural remodeling

- Encouraging nerve ingrowth
- Increasing neuronal excitability
- Reinforcing nerve-cancer interactions

Cell Types of the Tumor Microenvironment	
	Cancer Cells
	Invasive Cancer Cells
	Cancer Stem Cells
	Cancer-associated Fibroblasts
	Endothelial Cells
	Pericytes
	Tumor-promoting immune cells
	Stromal Progenitor Cells
	Senescent Cells (various origins)
	Neurons (and their projections)

**The nervous system interacts with many cancer hallmarks.**



Published in final edited form as:

*Lancet Oncol.* 2022 February ; 23(2): e62–e74. doi:10.1016/S1470-2045(21)00596-9.

## Therapeutic avenues for cancer neuroscience: translational frontiers and clinical opportunities

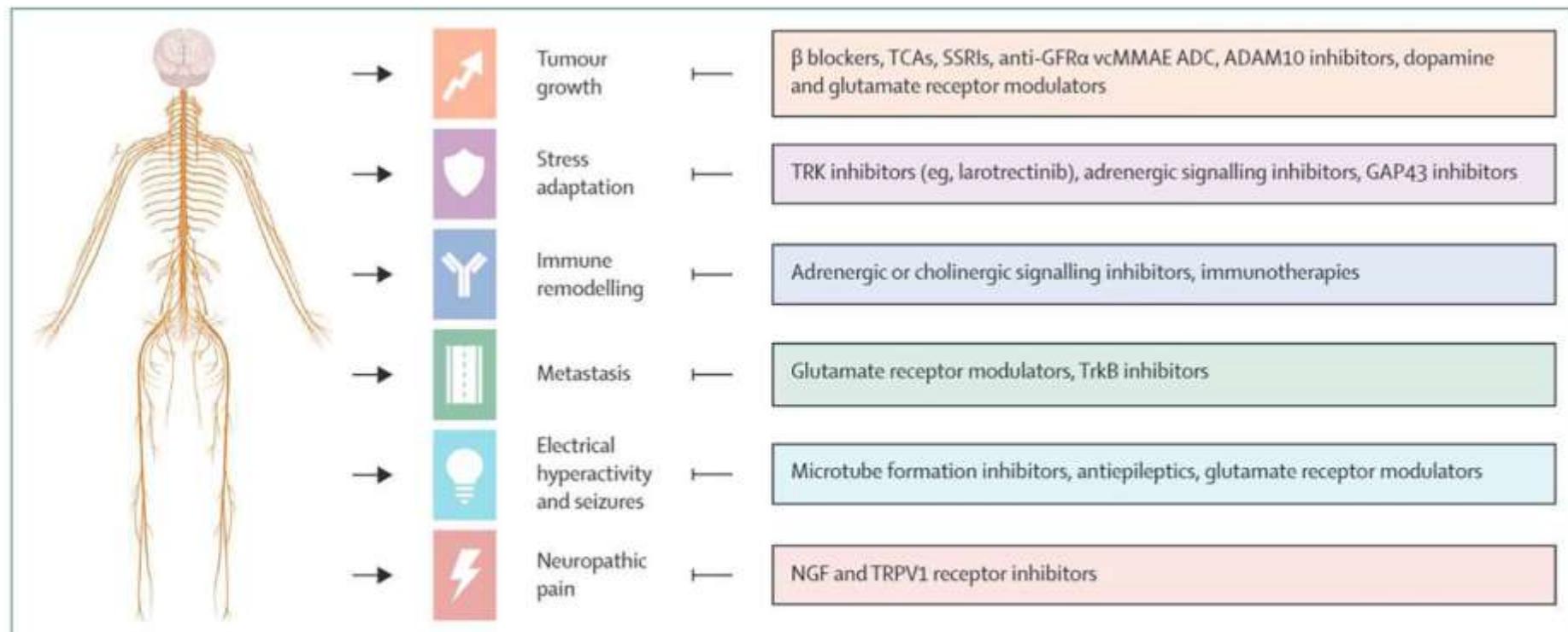


Figure 1: Six therapeutic avenues of cancer neuroscience

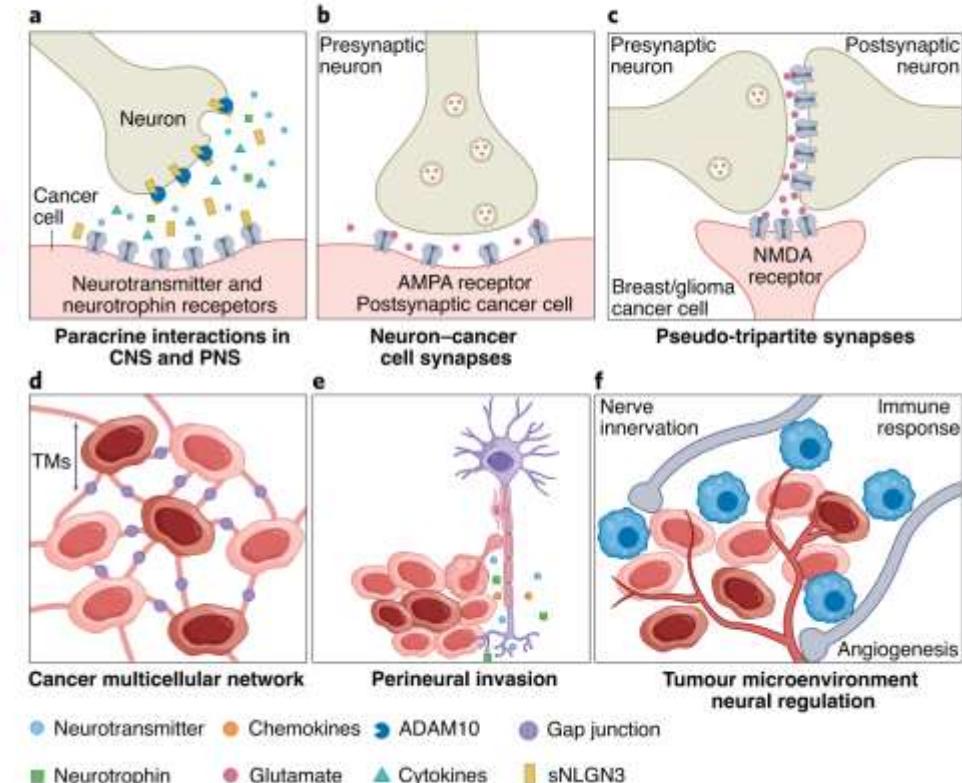
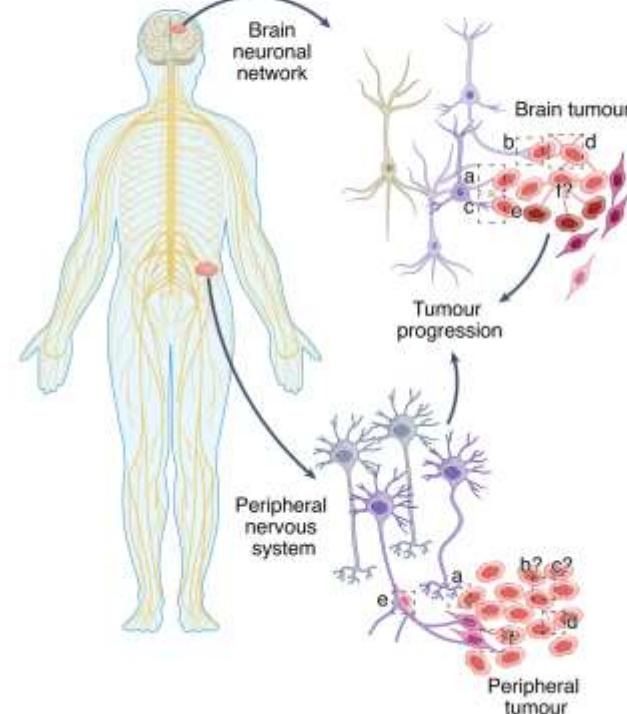


# Insights and opportunities at the crossroads of cancer and neuroscience

Chenchen Pan<sup>1,2</sup> and Frank Winkler<sup>1,2</sup>

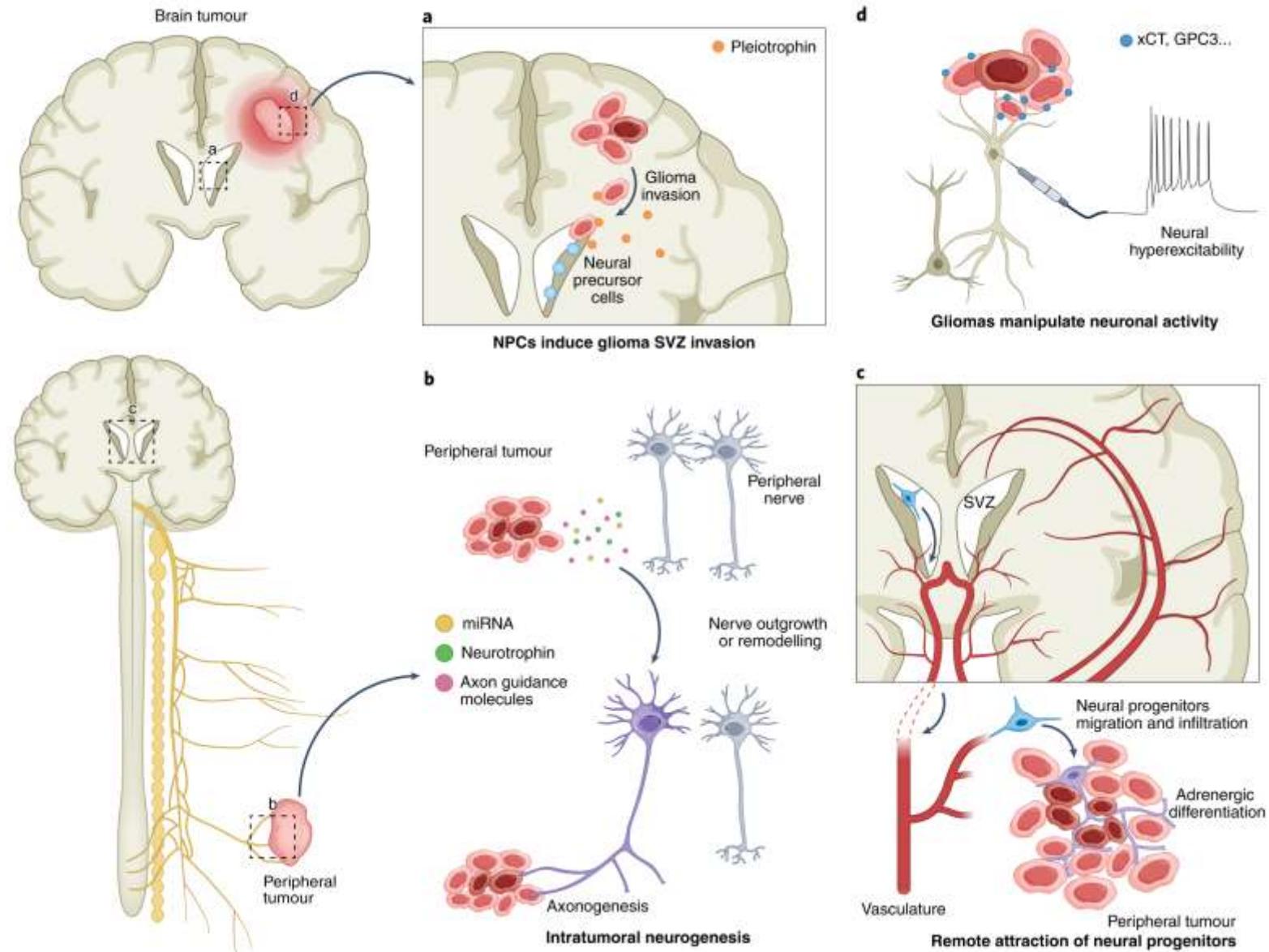
PERSPECTIVE

NATURE CELL BIOLOGY



# Cancer Neuroscience

## PERSPECTIVE



# Learning cell identity in immunology, neuroscience, and cancer

Stephanie Medina<sup>1,2</sup> · Rebecca A. Ihrie<sup>1,3,4</sup> · Jonathan M. Irish<sup>1,2,4</sup> 

Received: 27 September 2022 / Accepted: 19 November 2022 / Published online: 19 December 2022

© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2022

- Rigorous definitions of cell identity are a sign of field maturity (这三个领域都没有rigorous definitions)
- Both immunology and neuroscience now aim to link standardized measurements of protein or RNA to informative cell functions (免疫相对更为成熟，受到的关注也更多，神经也更复杂)
- compare practices in the fields of immunology and neuroscience, highlight concepts from each that might work well in the other (相互借鉴能更好的研究)
- propose ways to implement these ideas to study neural and immune cell interactions in brain tumors and associated model systems (特别是应用于brain tumors)

- Starting to receive gradual attention
- It is not only the research on brain tumors but also the entire neuro-endocrine and tumor field.
- From a clinical perspective, the novel application of neuro drugs in tumor research, cancer pain, and emotional issues arising from neuro-endocrine activity is also a direction of significant clinical relevance.
- From a neurobiological standpoint, the most fundamental issue—the classification of cells—remains unresolved.

# Conclusion and Next plan

- **Conducting research on cancer neuroscience, writing a review article (to establish the background for the graduation thesis):**
  - Brain tumor clinical classification
  - Fundamental direction
  - Transitional direction
  - Clinical treatment and drug application direction
- **Organizing the previous thesis direction on neuronal classification, analyzing submissions from a different perspective**
  - Profiling correlation among AS (alternative splicing), APA (alternative polyadenylation), morphology, and electrophysiology (mainly)
  - Introducing a deconvolution algorithm to develop a tool to decode bulk or pseudo-bulk data to predict morphological and electrophysiological information
  - Applying this tool in brain tumor microenvironment analysis
- **Graduation thesis direction:**
  - Neuroblastoma/GBM
  - Finding data samples that have not undergone cell sorting, analyzing from a neuronal perspective

[\*\*Back to Content\*\*](#)

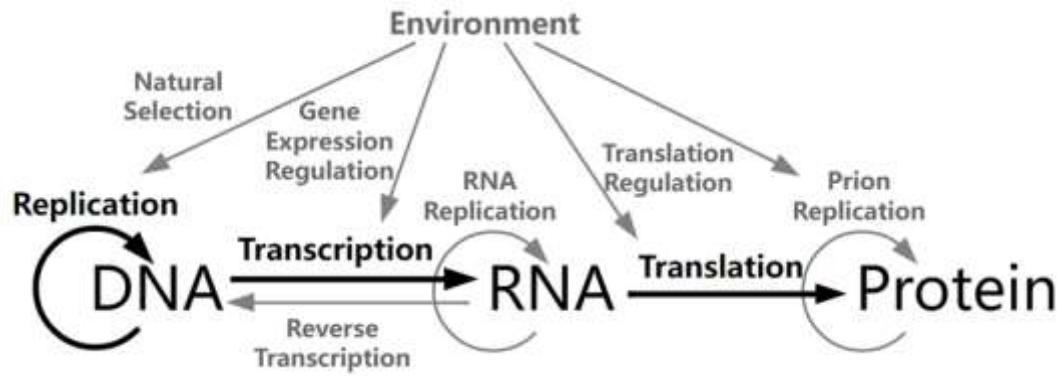
# **Alternative splicing and polyadenylation analysis**

Zhou Ying

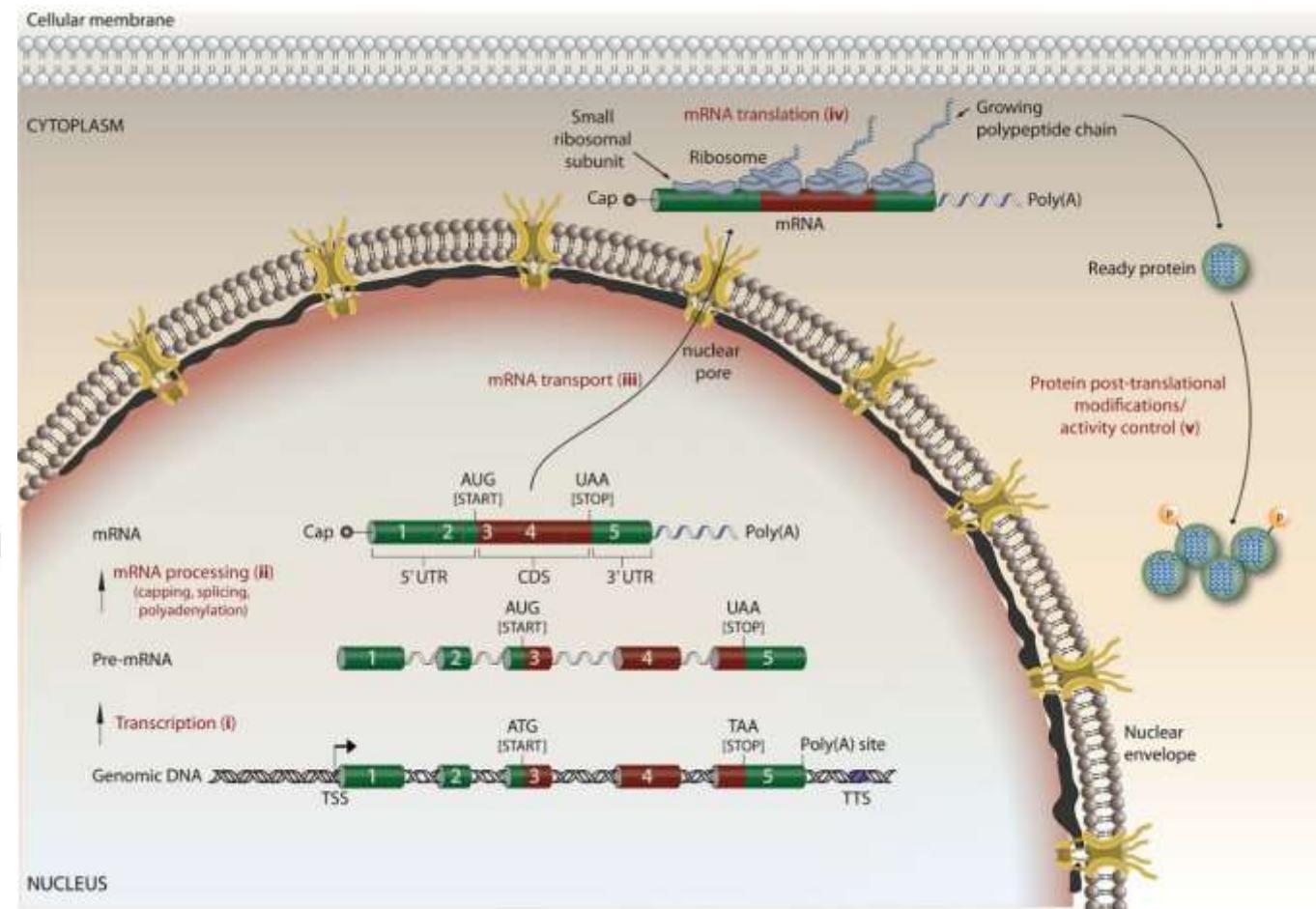
27<sup>th</sup> Mar. 2023

- Background of Alternative Splicing and Polyadenylation
- Alternative splicing
  - Splicing analysis tools
  - rMATs and Leafcutter
- APA Analysis
  - pAs tools
  - Dapars and Dapars2
- AS and APA regulation
- Epigenomics data integrative and Next plan

# Background—Overview of information flow in cell

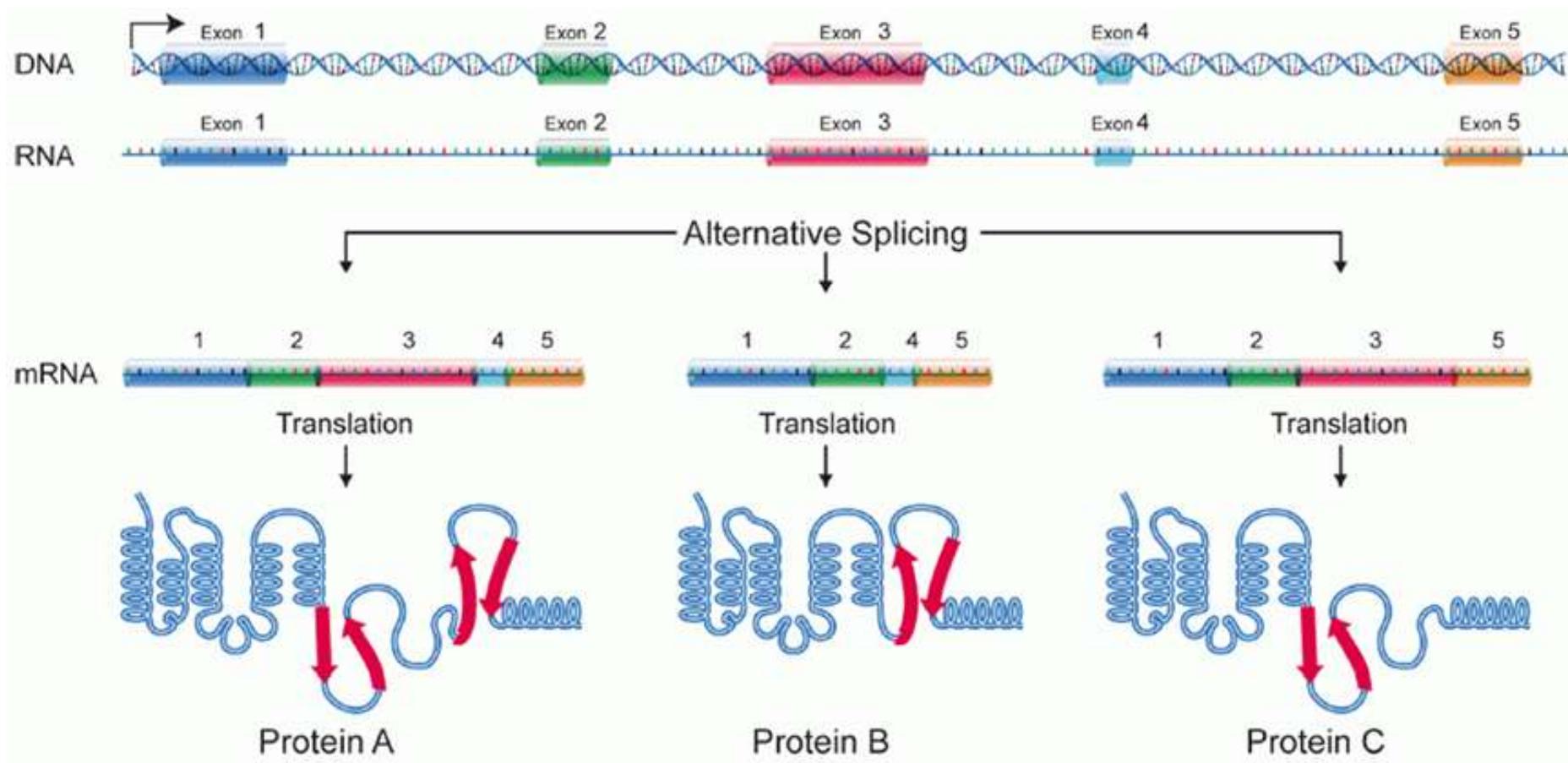


[http://hawaiireedlab.com/genewiki/index.php?title=Central\\_Dogma](http://hawaiireedlab.com/genewiki/index.php?title=Central_Dogma)



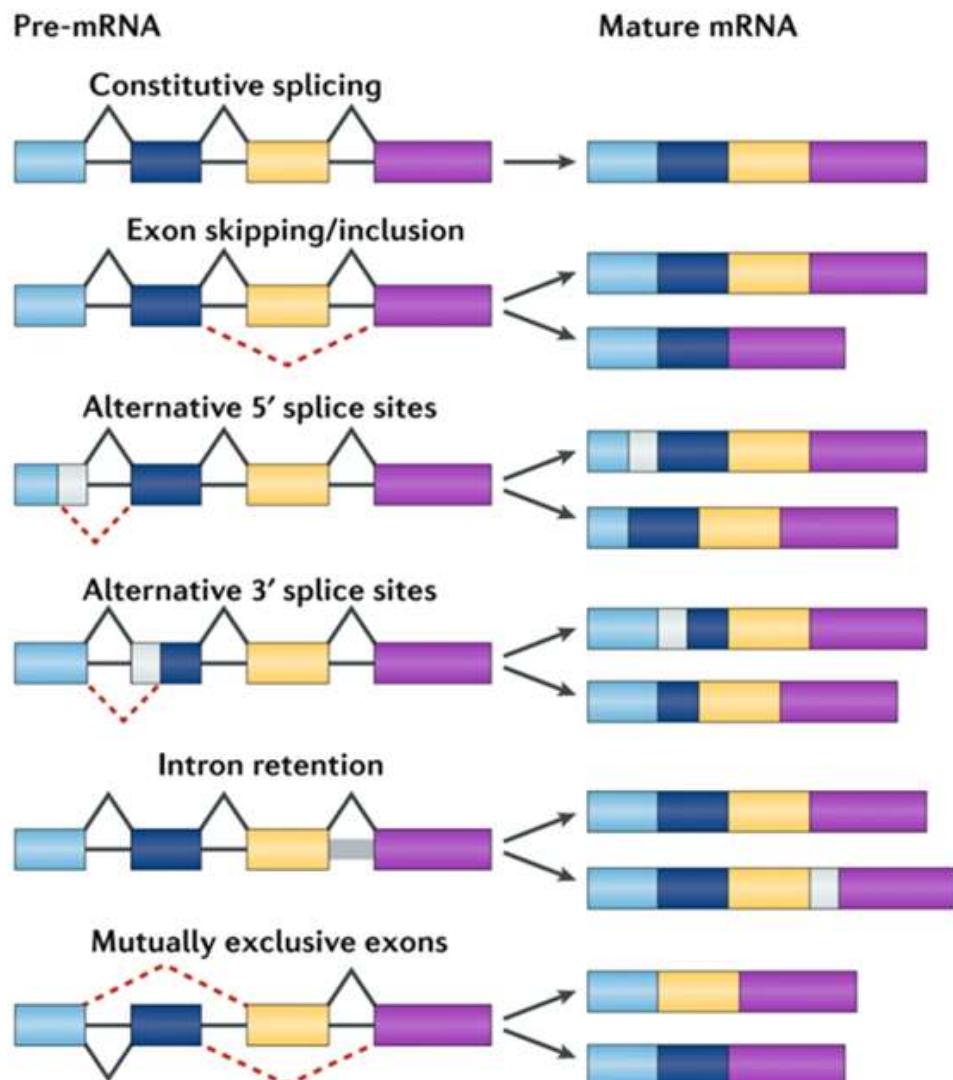
Human Epigenomics

## Background—AS

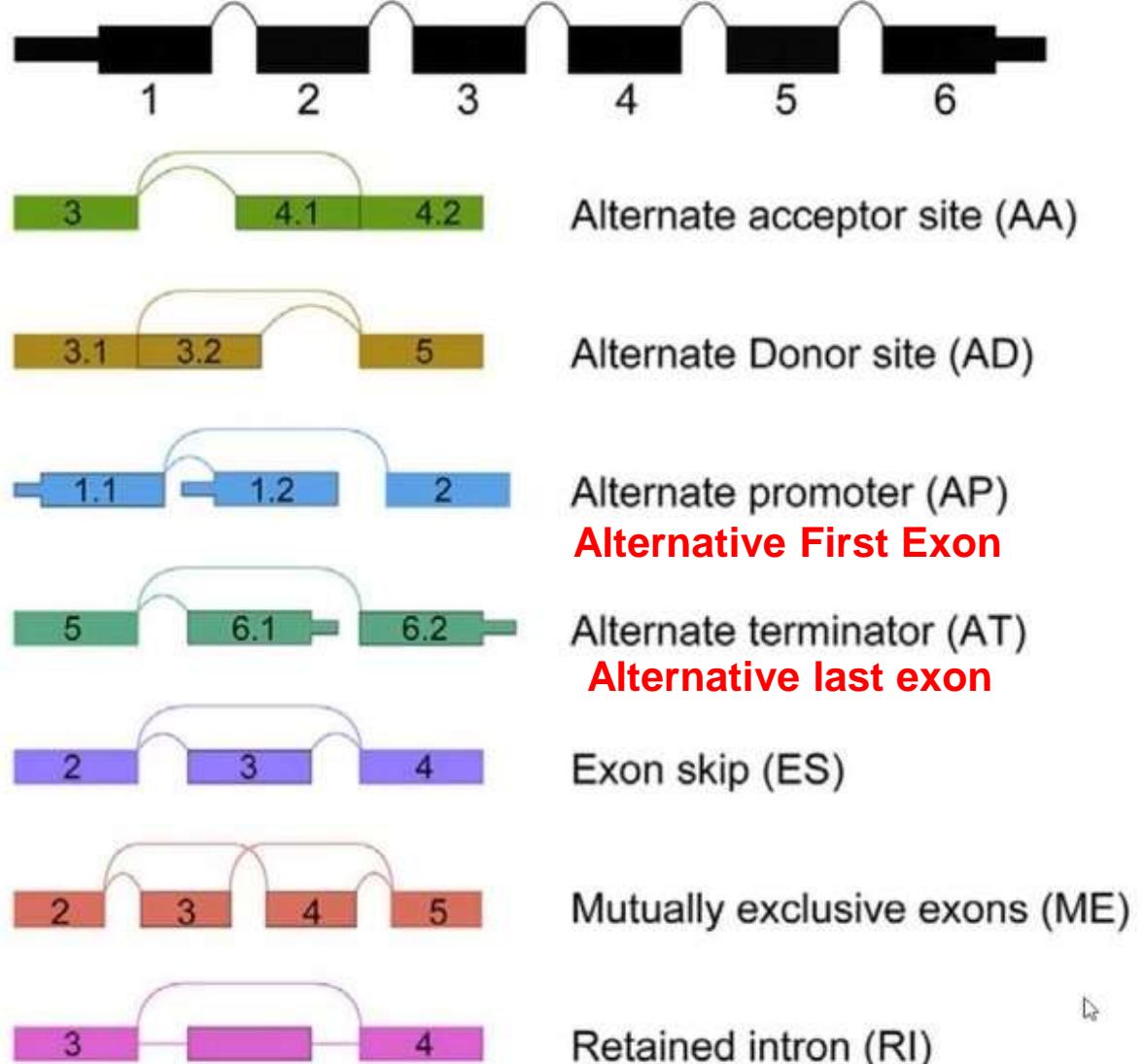


[https://commons.wikimedia.org/wiki/File:DNA\\_alternative\\_splicing.gif](https://commons.wikimedia.org/wiki/File:DNA_alternative_splicing.gif)

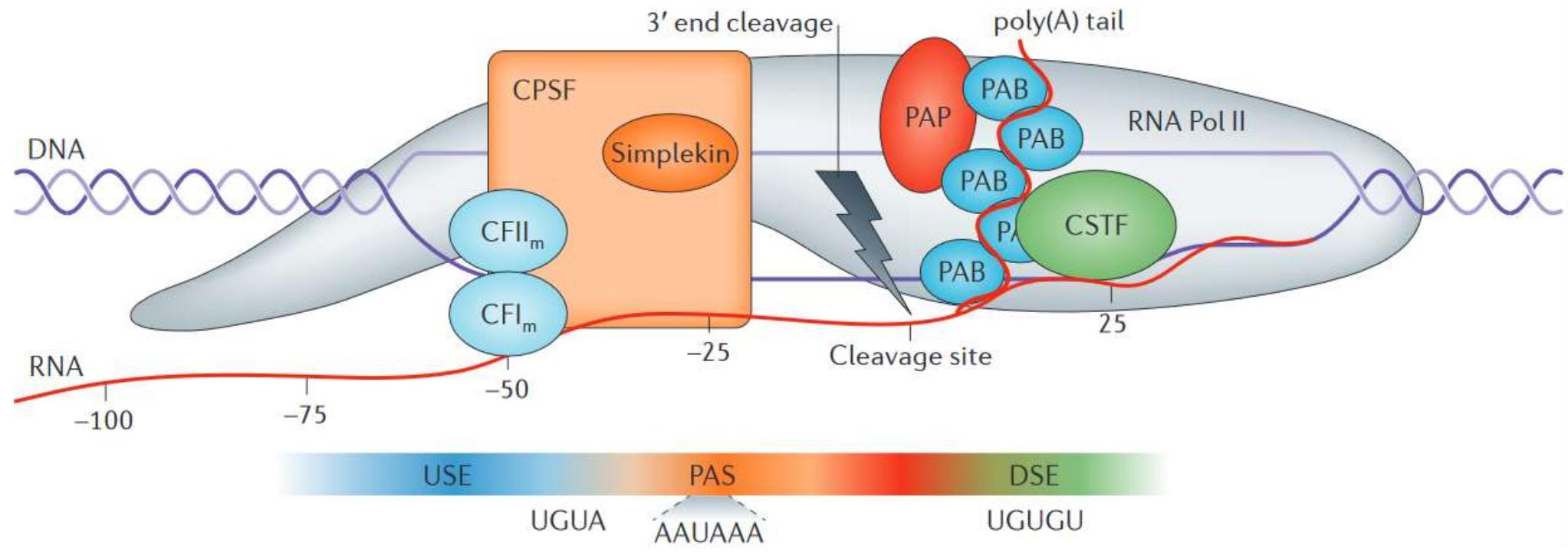
## Background—AS



# TCGA AS Events

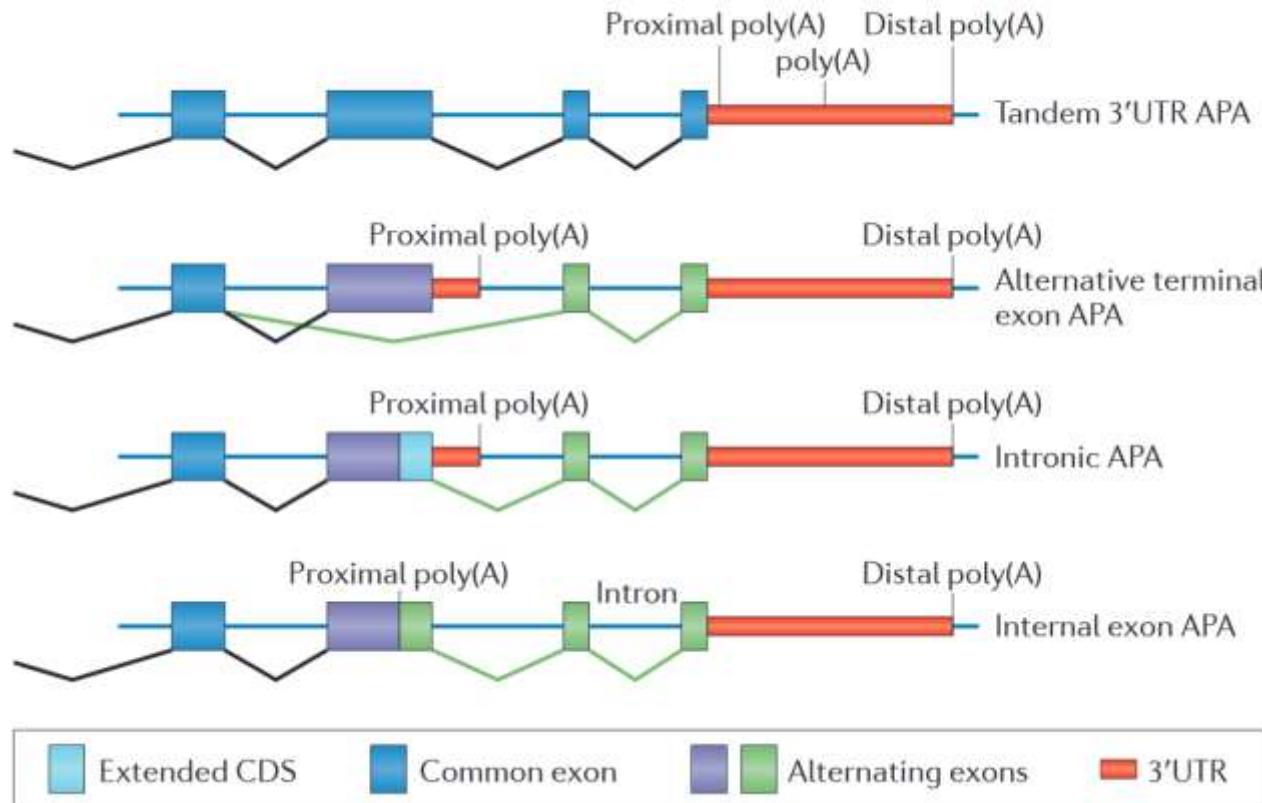


## Background—Alternative polyadenylation (APA)



Ran Elkon et al. (2013) Nature Reviews

# Background—APA



## The four different APA types

Ran Elkon et al. (2013) Nature Reviews

- Tandem 3'UTR APA**

- cleavage in 3'UTR
- same terminal exon, differ in 3'UTR length
- most frequency APA forms
- no influence in coding

- Alternative terminal exon APA**

- generate different terminal exon forms
- influence in coding
- most frequency

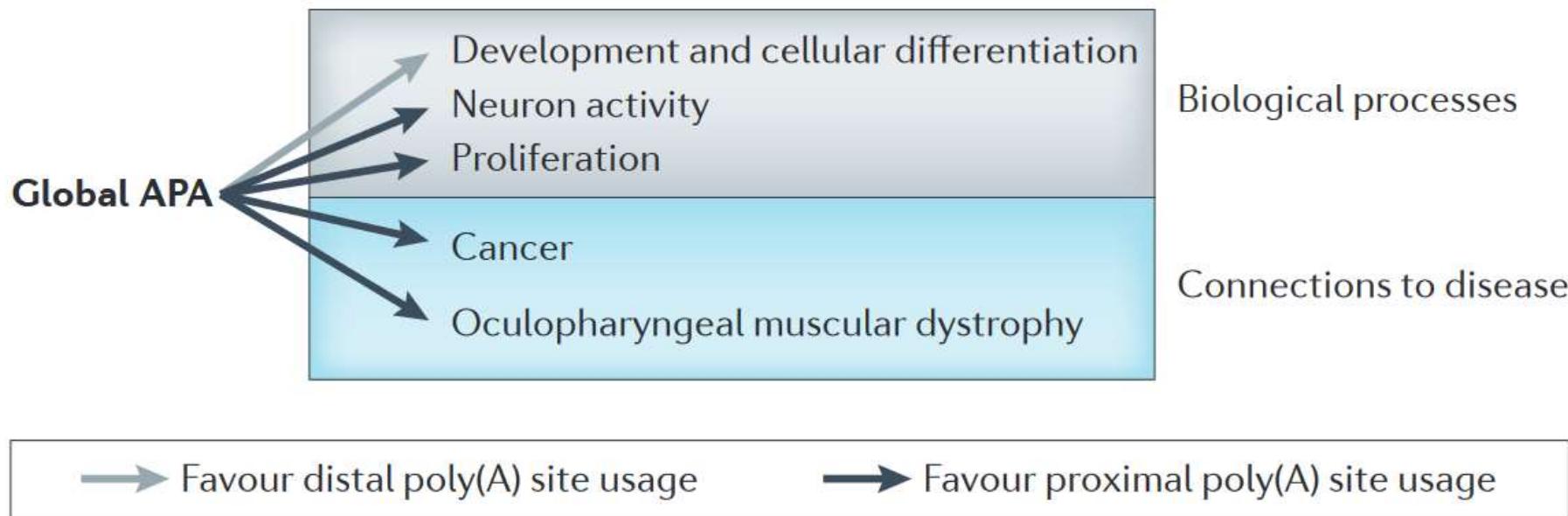
- Intronic APA**

- happen in intron region
- extending inner-exon to terminal exon
- less frequency
- influence in coding

- Internal exon APA**

- least frequent
- cleavage in coding region
- influence in coding

## Background—APA



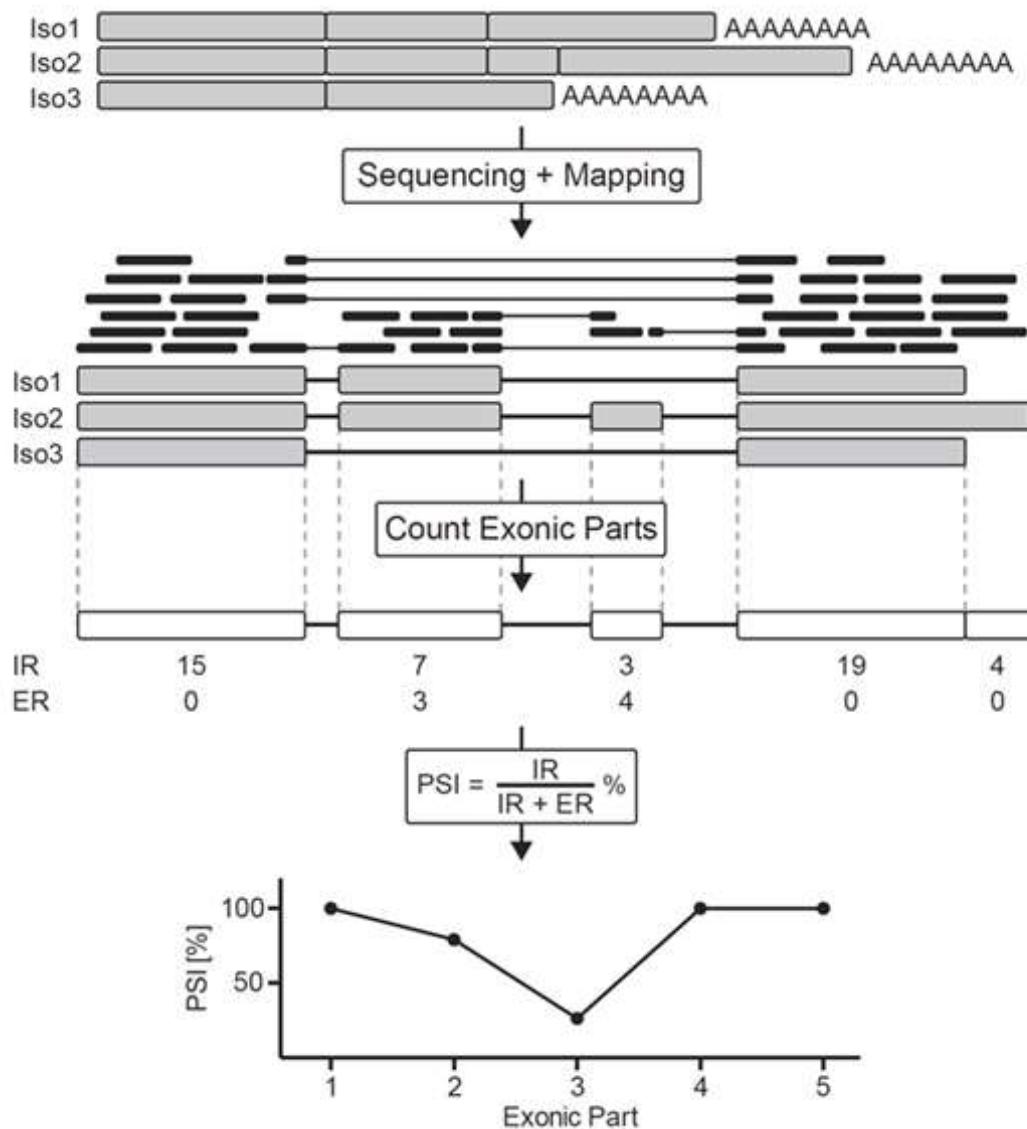
**Figure 3 | Biological processes that have been linked with broad APA modulation.**  
A schematic showing the biological processes and diseases that alternative polyadenylation (APA) has been linked with. In addition, the tendency towards distal or proximal poly(A) site usage is shown.

- 3'UTRs serve as major docking platforms for factors (microRNAs, RNA-binding pro)
- potentially regulates the stability, cellular localization and translation efficiency

# Background—AS tools

Categories	Description
Isoform-based	<ul style="list-style-type: none"><li>• Constructing full-length transcripts</li><li>• Estimating their relative abundance</li><li>• Statistical testing</li></ul>
Exon-based	<ul style="list-style-type: none"><li>• Distribution of Reads in Different Feature Regions</li><li>• Limitation: Splicing Events Not Considered</li></ul>
<b>Event-based</b>	<ul style="list-style-type: none"><li>• Statistics on a certain splicing event</li><li>• Detecting the partial gene expression containing a particular splicing event.</li></ul>

# Background—AS tools



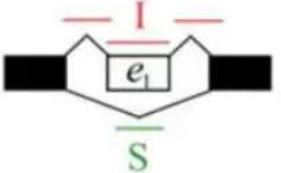
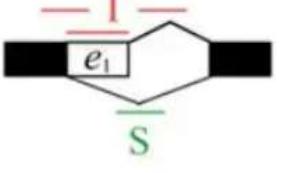
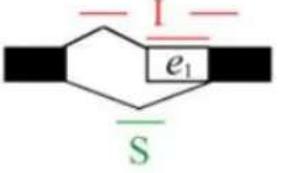
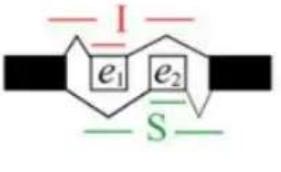
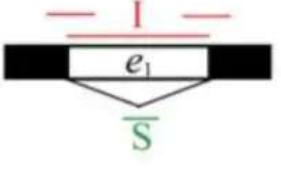
- **PSI (Percent Spliced In):** Refers to the percentage of splicing, indicating the efficiency with which a portion of an exon is spliced into isoforms of the same gene.
- **ER (Exclusion Reads):** Coverage of reads for excluded segments.
- **IR (Inclusion Reads):** Coverage of reads for included segments.

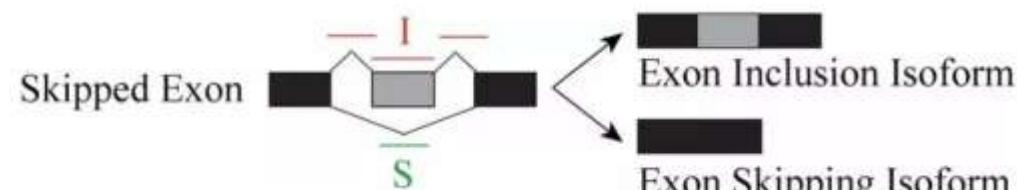
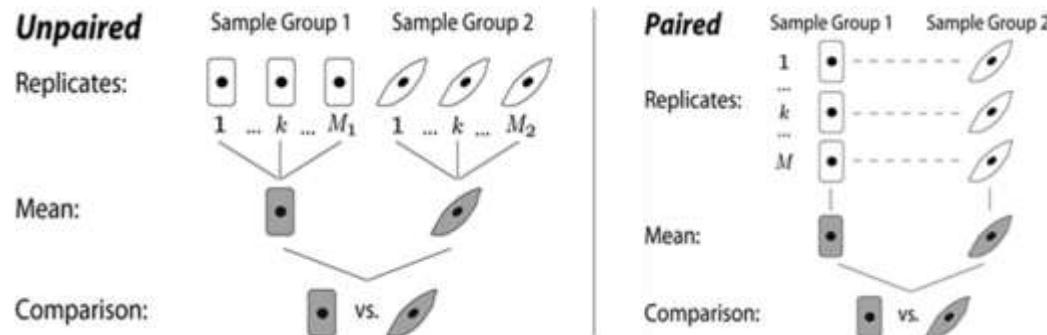
# Background—AS tools

Algorithms	Principles of Computing	Quantitation	differential analysis	Types	Unannotated transcript	Platform	Publications	Cites
rMATS (2014)	Event-based	PSI	yes	5	✓	Python	Nature, Cell	764
MAJIQ (2016)	event-based	PSI	yes	5	✗	Python	NC, NAR, GB	81
LeafCutter (2018)	intron excision	PSI	yes	Unkown	✓	R/C++/Python	NG, Bioinformatics	169
SUPPA2 (2018)	event-based	PSI	yes	7	✗	Python	NC, NAR, GB	117
SplAdder (2016)	event-based	PSI	yes	5	✓	Python/Matlab	Nature, Cancer cell	41
Whippet (2011)	event-based	PSI	yes	9	✗	Julia	Mol. Cell, Bioinformatics	34

# Background—rMATS

$I$ : reads of the inclusion isoform       $S$ : reads of the skipping isoform  
 $j$ : junction length       $e_1, e_2$ : exon length       $r$ : read length  
 $l_I$ : effective length of the inclusion isoform  
 $l_S$ : effective length of the skipping isoform

Skipped exon		Junction Length	$l_I : 2(j - r + 1)$	Junction & Exon Length	$l_I : e_1 - r + 1 + 2(j - r + 1)$
			$l_S : j - r + 1$		$l_S : j - r + 1$
Alternative 5' splice site		$l_I : 2(j - r + 1)$	$l_I : e_1 - r + 1 + 2(j - r + 1)$	$l_S : j - r + 1$	$l_S : j - r + 1$
Alternative 3' splice site		$l_I : 2(j - r + 1)$	$l_I : e_1 - r + 1 + 2(j - r + 1)$	$l_S : j - r + 1$	$l_S : j - r + 1$
Mutually exclusive exon		$l_I : 2(j - r + 1)$	$l_I : e_1 - r + 1 + 2(j - r + 1)$	$l_S : 2(j - r + 1)$	$l_S : e_2 - r + 1 + 2(j - r + 1)$
Retained intron		$l_I : 2(j - r + 1)$	$l_I : e_1 - r + 1 + 2(j - r + 1)$	$l_S : j - r + 1$	$l_S : j - r + 1$



$$\hat{\psi} = (I/l_I) / (I/l_I + S/l_S)$$

$$|\psi_1 - \psi_2| > c$$

# **Splicing Analysis—rMATS**

## **Required software for analysis:**

- 1) Linux operating system
- 2) rMATS (differential alternative splicing)
- 3) rmats2sashimiplot (visualization)

## **Prior installation of the following software is required to use rMATS:**

- 4) Install Python 2.7.x and corresponding versions of NumPy and SciPy
- 5) Download and install pysam (rMATS was tested with v0.9.1.4)
- 6) Download and install samtools (version 1.2 or later)
- 7) Download and install STAR (version 2.5 or later)

# Splicing Analysis—rMATS

## Input:

- \*.fastq/\*.bam
- s1.txt&s2.txt/b1.txt&b2.txt

```
↳ cat b1.txt
/alldata/zhouying/RNAseq/Fan_20230110/χ101SC22084125-Z01-J003/04_align_star/Kasumi-1sh scramble-R1_sortAligned.sortedByCoord.out.bam,/alldata/zhouying/RNAseq/Fan_20230110/χ101SC22084125-Z01-J003/04_align_star/Kasumi-1sh scramble-R2_sortAligned.sortedByCoord.out.bam,/alldata/zhouying/RNAseq/Fan_20230110/χ101SC22084125-Z01-J003/04_align_star/Kasumi-1sh scramble-R3_sortAligned.sortedByCoord.out.bam
```

```
↳ cat b2.txt
/alldata/zhouying/RNAseq/Fan_20230110/χ101SC22084125-Z01-J003/04_align_star/Kasumi-1shRNA2-470--R1_sortAligned.sortedByCoord.out.bam,/alldata/zhouying/RNAseq/Fan_20230110/χ101SC22084125-Z01-J003/04_align_star/Kasumi-1shRNA2-470--R2_sortAligned.sortedByCoord.out.bam,/alldata/zhouying/RNAseq/Fan_20230110/χ101SC22084125-Z01-J003/04_align_star/Kasumi-1shRNA2-470--R3_sortAligned.sortedByCoord.out.bam
```

# Splicing Analysis—rMATS

## Running

```
(RNACocktail) zhouying@x86_64-conda-linux-gnu /alldata/zhouying/RNAseq/splicing_test
└─> rmats.py --b1 b1.txt --b2 b2.txt --gtf /alldata/zhouying/ref_data/anno/GRCh38.p13/gencode.v36.annotation.gtf --od ./HepG2_sh1_rMATS --tmp ./HepG2_sh1_tmp_od --readLength 150 --libType fr-unstranded --nthread 20
```

--b1 --b2 b1.txt和b2.txt路径

--gtf gtf文件路径，注意版本要跟比对所用一致

--od output dir

--tmp rMATS 产生的临时文件输出路径

**--readLength 测序reads长**

**--libType**

Default isunstranded (fr-unstranded). 链非特异RNAseq

Use fr-firststrand or fr-secondstrand for strand-specific data 链特异RNAseq

# Splicing Analysis—rMATS

```
ls | grep .txt  
2023-02-02-07 06 42 991159_read_outcomes_by.bam.txt  
A3SS.MATS.JC.txt  
A3SS.MATS.JCEC.txt  
A5SS.MATS.JC.txt  
A5SS.MATS.JCEC.txt  
JC.raw.input.A3SS.txt  
JC.raw.input.A5SS.txt  
JC.raw.input.MXE.txt  
JC.raw.input.RI.txt  
JC.raw.input.SE.txt  
JCEC.raw.input.A3SS.txt  
JCEC.raw.input.A5SS.txt  
JCEC.raw.input.MXE.txt  
JCEC.raw.input.RI.txt  
JCEC.raw.input.SE.txt  
MXE.MATS.JC.txt  
MXE.MATS.JCEC.txt  
RI.MATS.JC.txt  
RI.MATS.JCEC.txt  
SE.MATS.JC.txt  
SE.MATS.JCEC.txt
```

- ID, GeneID, geneSymbol, chr, strand,
- exonStart\_0base, exonEnd, upstreamES, upstreamEE, downstreamES, downstreamEE
- ID
- **IJC\_SAMPLE\_1:** 样品在inclusion junction下的count数, 重复样本的结果以逗号分隔
- **SJC\_SAMPLE\_1:** 样品在skipping junction counts (SJC) 下的count数, 重复样本的结果以逗号分隔
- **IJC\_SAMPLE\_2:** 样品在inclusion junction下的count数, 重复样本的结果以逗号分隔
- **SJC\_SAMPLE\_2:** 样品在skipping junction counts (SJC) 下的count数, 重复样本的结果以逗号分隔
- **IncFormLen:** lengthof inclusion form, used **for** normalization
- **SkipFormLen:** lengthof skipping form, used **for** normalization
- **PValue:** Significanceof splicing difference between two sample groups (两组样品可变剪切的统计学显著差异指标)
- **FDR:** FalseDiscovery Rate calculated **from** p-value (校正后的Pvalue)
- **IncLevel1:** inclusionlevel **for** SAMPLE\_1 **replicates** (comma separated)  
calculated **from** normalizedcounts
- **IncLevel2:** inclusionlevel **for** SAMPLE\_2 **replicates** (comma separated)  
calculated **from** normalizedcounts
- **IncLevelDifference:** **average**(IncLevel1)- **average**(IncLevel2)

# Splicing Analysis—rMATS rmats2sashimiplot 可视化

```
↳ rmats2sashimiplot --b1 /alldata/zhouying/RNAseq/Fan_20230201/X101SC22084125-Z01-J006/04_align_star/HepG2_EV-3_sortAligned.sortedByCoord.out.bam,/alldata/zhouying/RNAseq/Fan_20230201/X101SC22084125-Z01-J006/04_align_star/HepG2_EV-2_sortAligned.sortedByCoord.out.bam,/alldata/zhouying/RNAseq/Fan_20230201/X101SC22084125-Z01-J006/04_align_star/HepG2_EV-1_sortAligned.sortedByCoord.out.bam --b2 /alldata/zhouying/RNAseq/Fan_20230201/X101SC22084125-Z01-J006/04_align_star/HepG2_sh1-1_sortAligned.sortedByCoord.out.bam,/alldata/zhouying/RNAseq/Fan_20230201/X101SC22084125-Z01-J006/04_align_star/HepG2_sh1-2_sortAligned.sortedByCoord.out.bam,/alldata/zhouying/RNAseq/Fan_20230201/X101SC22084125-Z01-J006/04_align_star/HepG2_sh1-3_sortAligned.sortedByCoord.out.bam -t SE -e HNRNPC_A5SS_MATS_JC.txt --l1 HepG2_EV --l2 HepG2_sh1 -o HNRNPC_A5SS_plot
```

--b1 --b2 group1/group2 bam文件路径，逗号分隔

--t AS Events类型

--o output dir

--l1 --l2 group1/group2 命名标签

**-e \*MATS\_JC.txt rMATS结果，一行出一个pdf提前筛选**

# Splicing Analysis—rMATS rmats2sashimiplot 可视化

```
rmats2sashimiplot --b1 /alldata/zhouying/RNAseq/Fan_20230201/X101SC22084125-Z01-J006/04_align_star/HepG2_EV-3_sortedAligned.sortedByCoord.out.bam,/alldata/zhouying/RNAseq/Fan_20230201/X101SC22084125-Z01-J006/04_align_star/HepG2_EV-2_sortedAligned.sortedByCoord.out.bam,/alldata/zhouying/RNAseq/Fan_20230201/X101SC22084125-Z01-J006/04_align_star/HepG2_EV-1_sortedAligned.sortedByCoord.out.bam --b2 /alldata/zhouying/RNAseq/Fan_20230201/X101SC22084125-Z01-J006/04_align_star/HepG2_sh1-1_sortedAligned.sortedByCoord.out.bam,/alldata/zhouying/RNAseq/Fan_20230201/X101SC22084125-Z01-J006/04_align_star/HepG2_sh1-2_sortedAligned.sortedByCoord.out.bam,/alldata/zhouying/RNAseq/Fan_20230201/X101SC22084125-Z01-J006/04_align_star/HepG2_sh1-3_sortedAligned.sortedByCoord.out.bam -t SE -e HNRNPC_A5SS_MATS_JC.txt --l1 HepG2_EV --l2 HepG2_sh1 -o HNRNPC_A5SS_plot
```

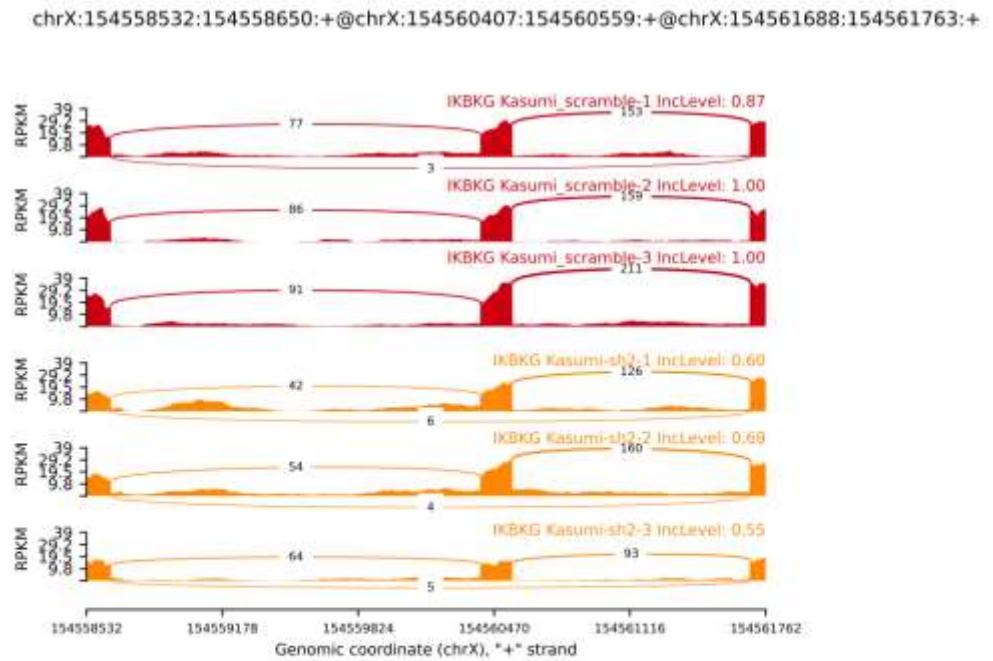
--b1 --b2 group1/group2 bam文件路径, 逗号分隔

--t AS Events类型

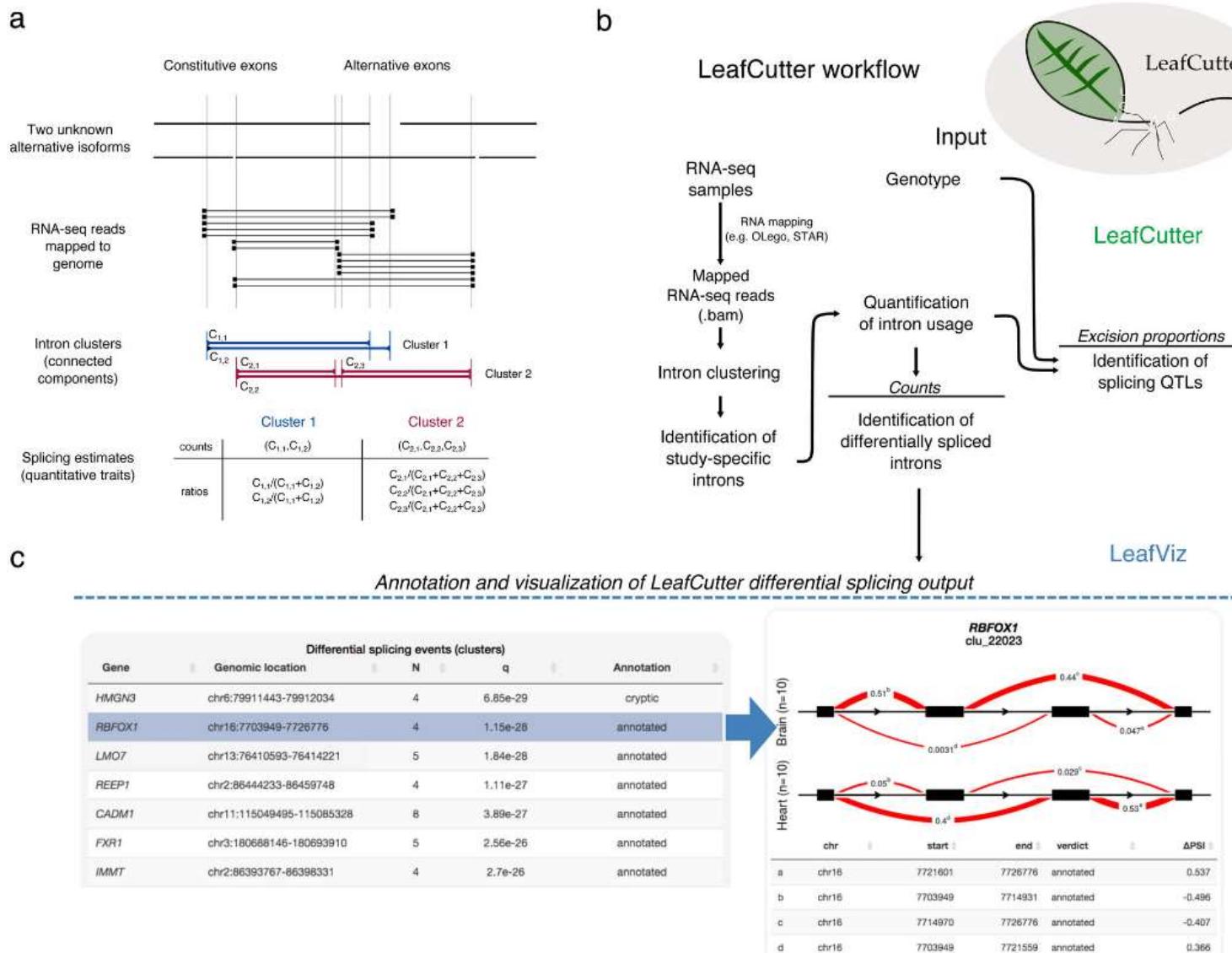
--o output dir

--l1 --l2 group1/group2 命名标签

**-e \*MATS\_JC.txt rMATS结果, 一行出一个pdf提前筛选**



# Background—Leafcutter



- 摆脱了pre-mRNA可变剪接分析需要提前导入模型这个束缚，可以通过专注于内含子切除来识别和量化新的和已知的可变剪接事件
- LeafCutter既可用于检测样品组之间的差异剪接，并直接可视化差异剪接发生的位置。

# Splicing Analysis—Leafcutter

Step 1. Converting bams to juncs (BED12)

Step 2. Intron Clustering

Step 3. Differential intron excision analysis

```
1 rootdir=<input dir>
2 outdir=<output dir>
3 * ls ${rootdir}|grep _sortAligned.sortedByCoord.out.bam|grep -v "bai"|while read id;
4 do
5 *     echo ${id}' running...'
6 *     #echo ${id%_sortA*}
7 *     sh <PATH>/leafcutter/scripts/bam2junc.sh ${rootdir}/${id} ${outdir}/${id%_sortA*}.junc
8 done
```

▼ Clustering together the introns

```
1 # Making all_juncfiles.txt
2 ls *.junc >all_juncfiles.txt
3 python <PATH>/leafcutter/clustering/leafcutter_cluster.py -j all_juncfiles.txt -m 50 -o npc113 -L 500000
```

```
↳ cat all_juncfiles.txt
HepG2_EV-1.junc
HepG2_EV-2.junc
HepG2_EV-3.junc
HepG2_sh1-1.junc
HepG2_sh1-2.junc
HepG2_sh1-3.junc
```

▼ making\_exons.txt.gz

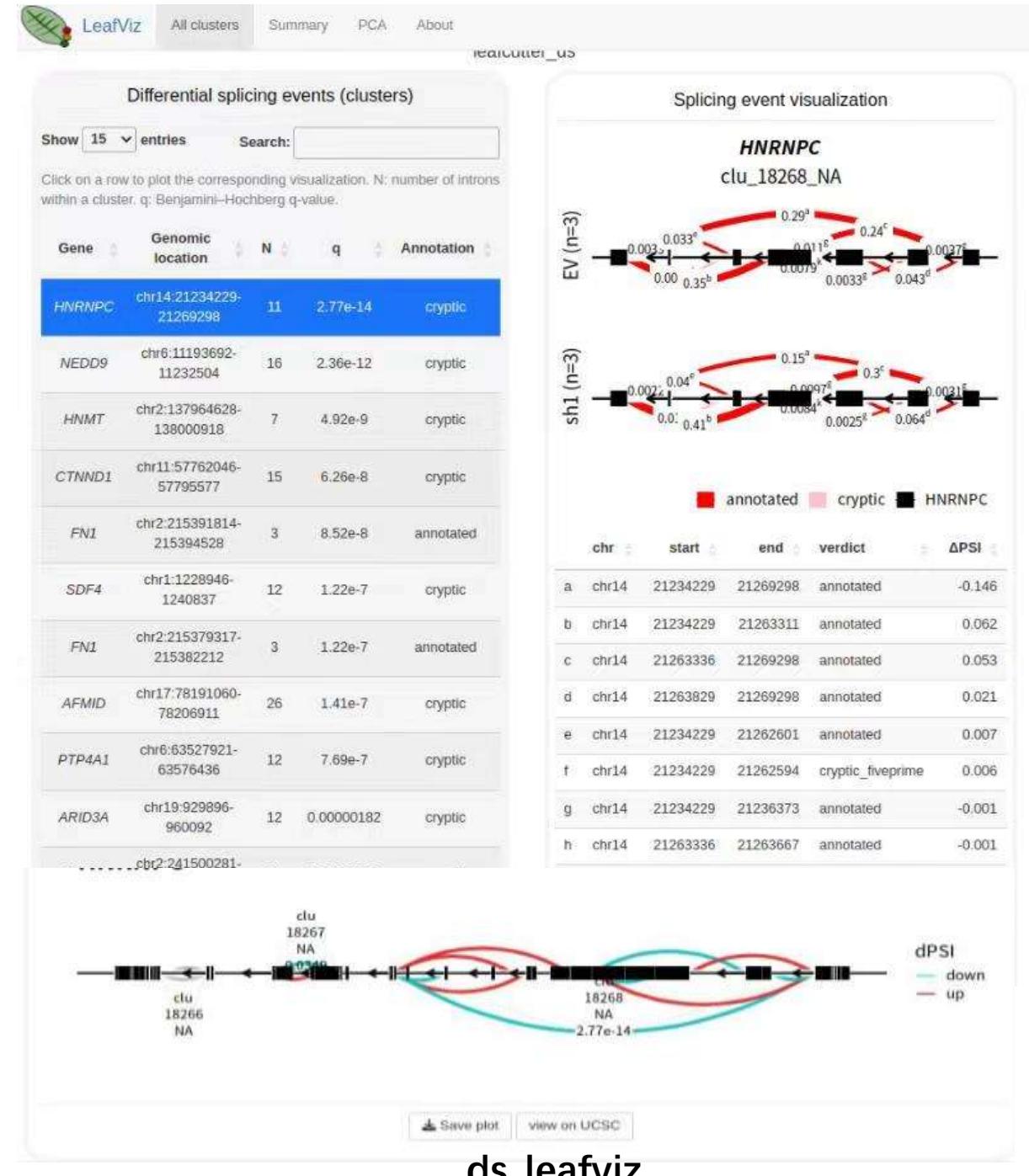
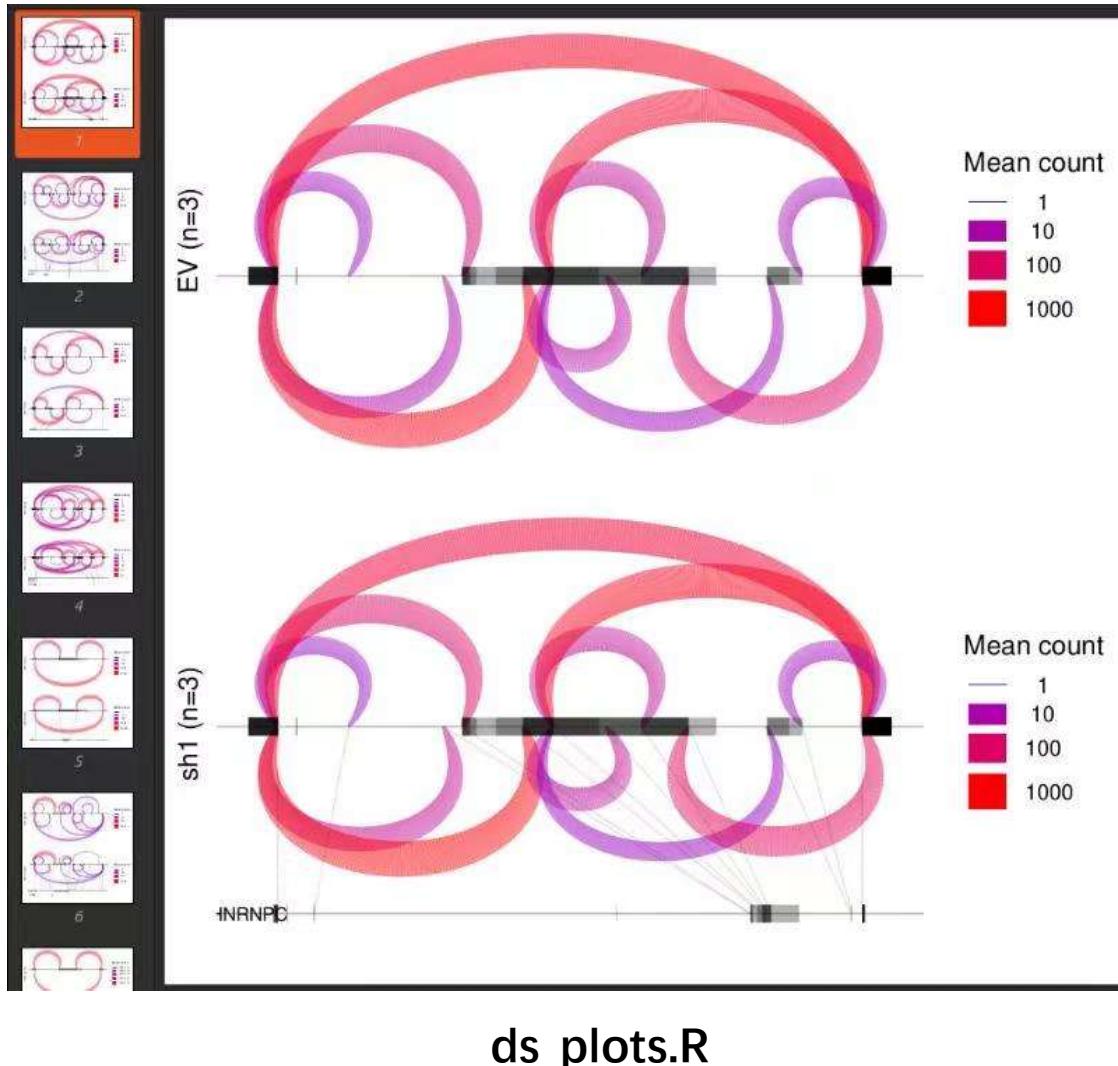
```
1 Rscripts <PATH>/leafcutter/scripts/gtf_to_exons.R genecond.gtf output_exons.txt.gz
2 # input gtf file default gz compress, if not gz, rewrite zcat command in R script to cat
```

可视化：  
ds\_plots.R  
leafviz

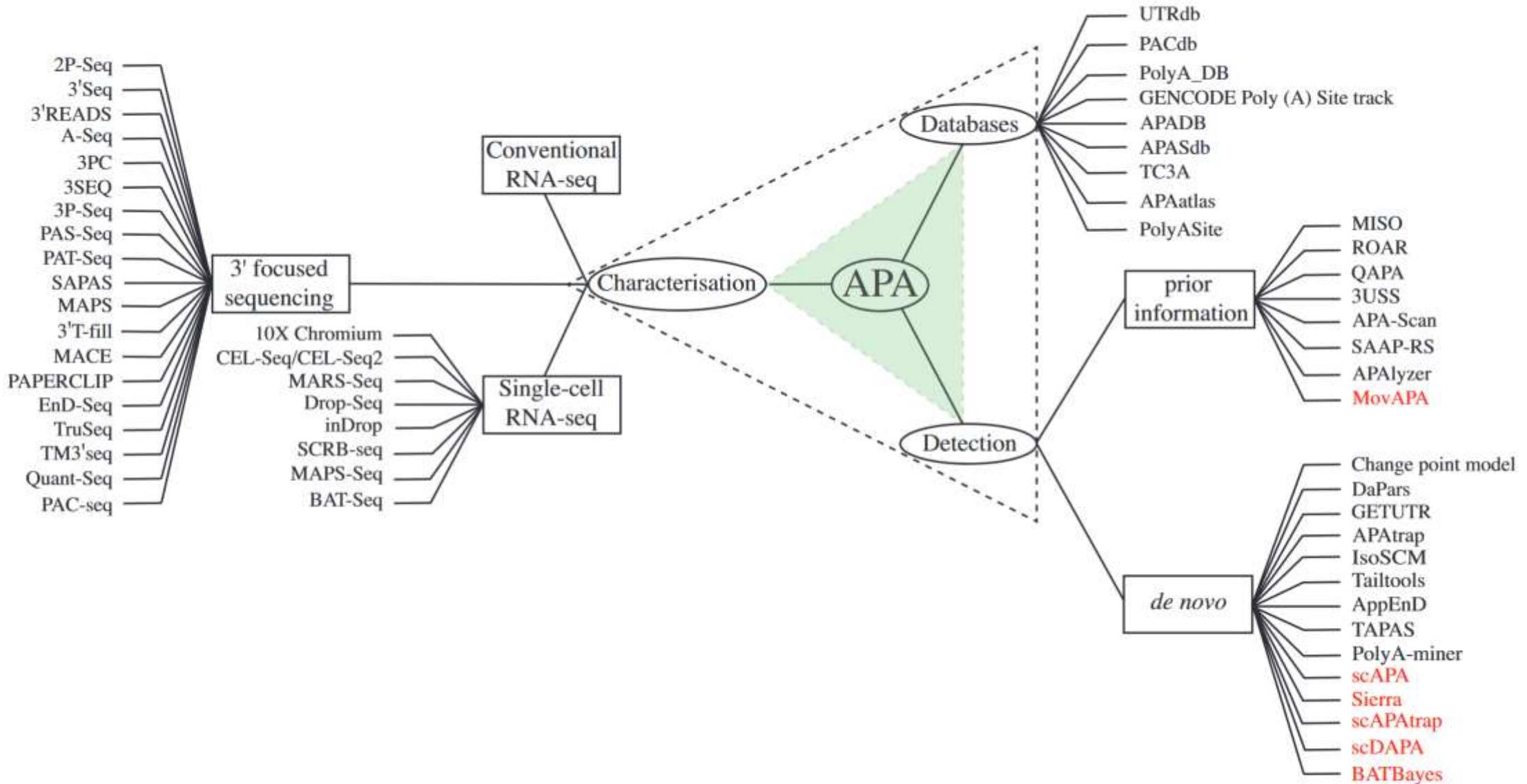
▼ DS analysis

```
1 <PATH>/leafcutter/scripts/leafcutter_ds.R --num_threads 20 \
2 --exon_file=$exonbed\
3 --min_samples_per_intron=3 \ #根据样本情况调整
4 <NAME>_perind_nums.counts.gz group_info.txt
```

# Splicing Analysis—Leafcutter



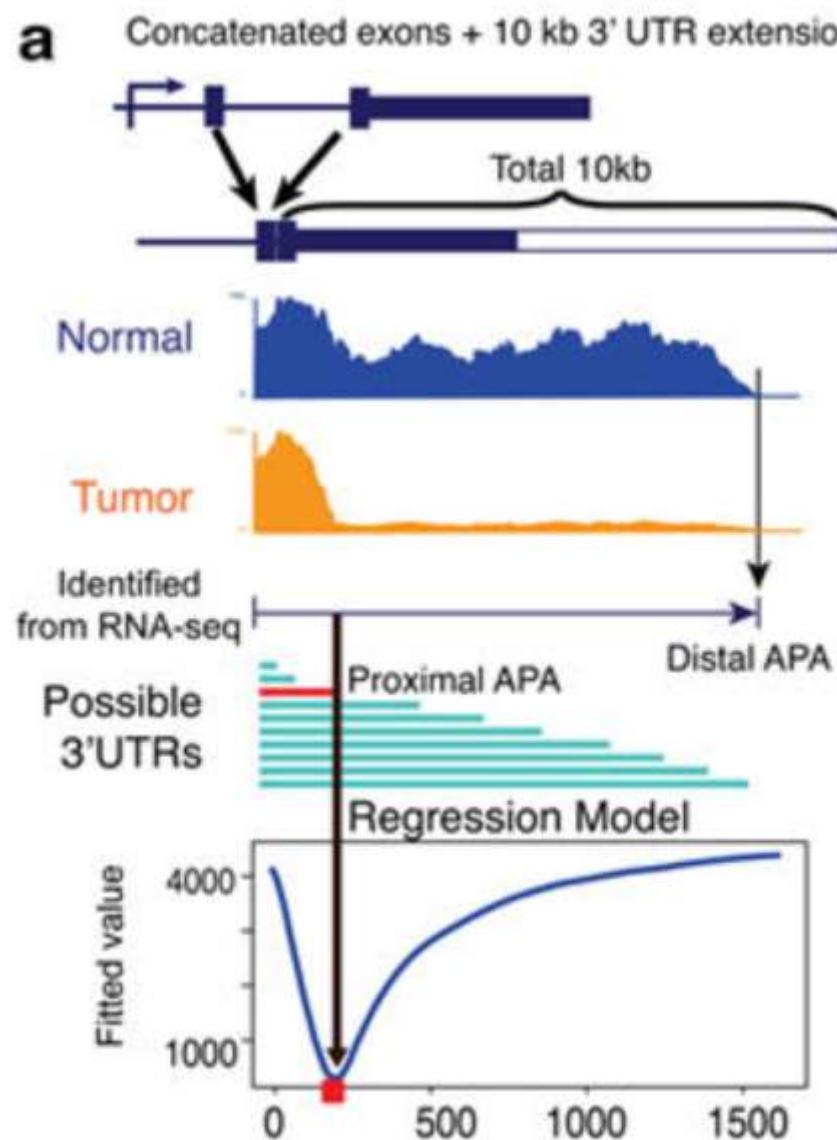
# Background—pAs tools



# Background—pAs tools

- **relying on priori annotation**
  - Annotation needs to be based on the pAs database.
  - This enables more thorough exploration of accurately annotated pAs.
  - It's difficult to obtain the latest, high-quality, and easily usable data.
  - Discovering new APA events is challenging.
- **relying on transcript reconstruction**
  - Such methods can only rely on rough analysis based on annotated reference genomes.
  - The accuracy of 3'UTRs is also a technical issue, as RNAseq exhibits 3' bias.
  - Analyzing lowly expressed transcripts using this method poses problems.
  - Third-generation sequencing and other full-length transcript approaches may yield better results.
- **using polyA-capped reads**
  - Utilizing polyA sequences that cannot be aligned to the reference genome to locate pAs.
  - This method allows for very precise localization, but such reads are rare in RNAseq data.
- **based on detecting read density fluctuations (Dapars/Dapars2)**
  - These tools utilize read coverage at each position along the 3'UTR to search for sudden fluctuations in read density, indicating the insertion of long transcripts at these positions.
  - ○ Requires high-quality 3'end coverage ○ May be influenced by non-biological variability ○ Non-3'UTR APAs cannot be detected

# Dapars and Dapars2

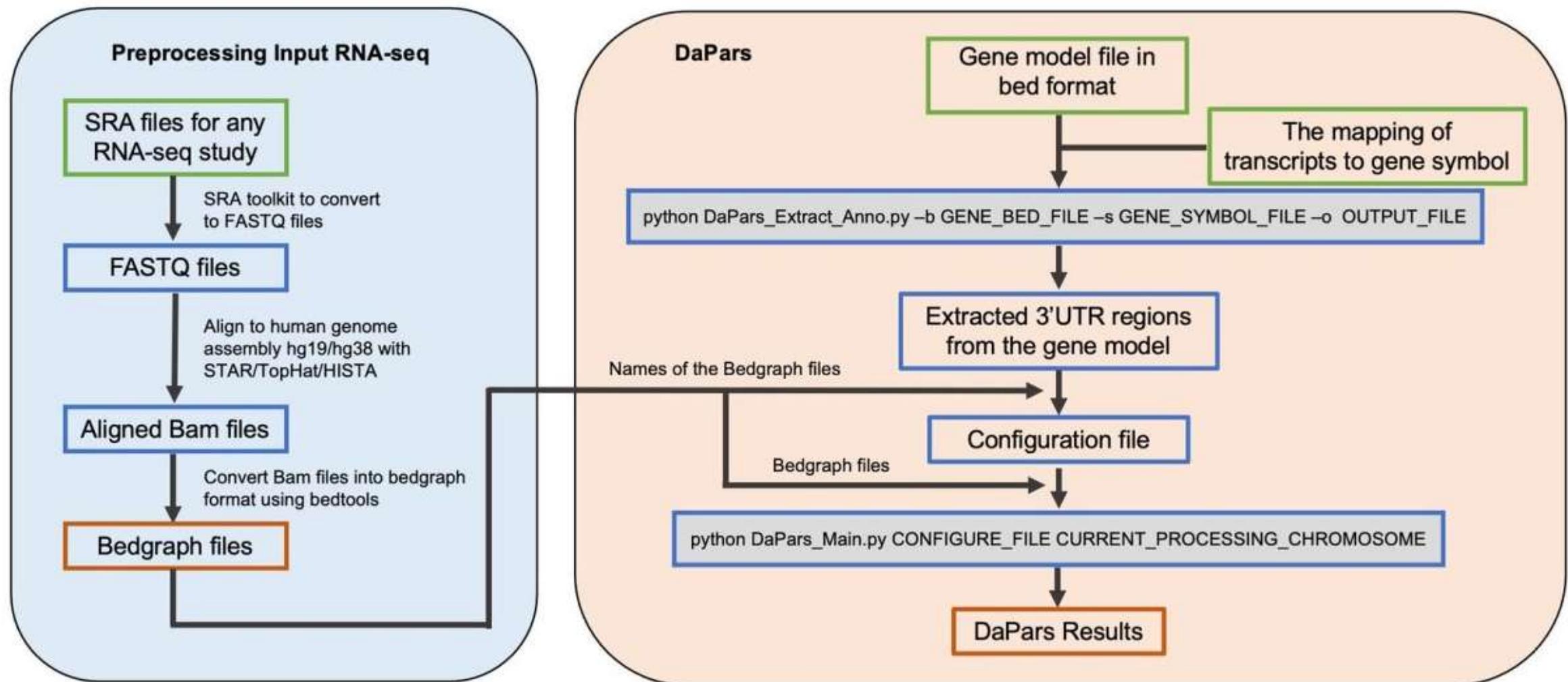


$$(w_L^{1*}, w_L^{2*}, w_S^{1*}, w_S^{2*}, P^*) = \arg \min_{w_L^1, w_L^2, w_S^1, w_S^2 \geq 0, 1 < P < L} \sum_{i=1}^2 \|\mathbf{C}_i - (w_L^i \mathbf{I}_L + w_S^i \mathbf{I}_P)\|_2^2 \quad (\text{Eq.1})$$

$$\text{PDUI} = \frac{w_L^{i*}}{w_L^{i*} + w_S^{i*}} \quad (\text{Eq.2})$$

$w_L^{i*}$  and  $w_S^{i*}$  are the estimated expression levels of transcripts with **distal and proximal polyA sites** for sample i.

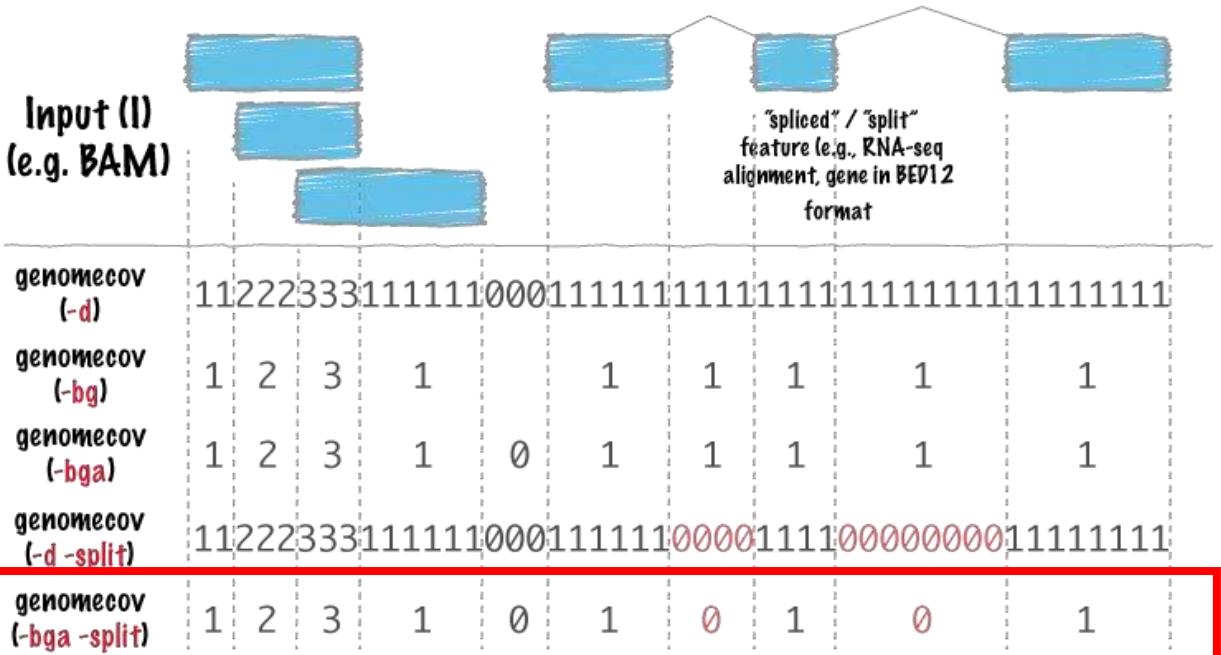
# Dapars and Dapars2



# Dapars and Dapars2

## Input files preparation

Format	Description	Example
BED	a <a href="#">BED12</a> file of gene models. It can be downloaded from <a href="#">UCSC</a>	hg38_wholeGene_annotation.bed
Bedgraph	files in this format store the reads alignment information, which can be converted from BAM files by bedtools (e.g.: "bedtools genomecov -ibam *.bam -bga -split -trackline")	<b>sample_1.wig</b>
plain text	a tab-delimited two columns file contains Refseq ID mapping with gene name	hg38_refseq_IDmapping.txt



```
python /alldata/zhouying/evnconfig/DaPars2-master/src/DaPars_Extract_Anno.py -b /alldata/zhouying/ref_data/anno/GRCh38.p13/gencode.v36.annotation.bed12 -s /alldata/zhouying/ref_data/anno/GRCh38.p13/gencode_v36_dapars2_IDmapping.txt -o /alldata/zhouying/ref_data/anno/GRCh38.p13/gencode_v36_extracted_3UTR.bed
```

1	chr1	67092165	67093604	NM_001276352.2 Clorf141 chr1 +	0	-
2	chr1	67092176	67093604	XM_011541469.1 Clorf141 chr1 +	0	-
3	chr1	67093485	67093604	XM_017001277.1 Clorf141 chr1 +	0	-
4	chr1	67095353	67095421	XM_011541472.1 Clorf141 chr1 +	0	-
5	chr1	201330074	201332989	NM_001005337.3 PKP1 chr1 +	0	+
6	chr1	8352404	8355120	NM_001042682.2 RERE chr1 +	0	-
7	chr1	33513998	33516570	NM_001281956.2 CSMD2 chr1 +	0	-
8	chr1	33514012	33519677	XM_024452878.1 CSMD2 chr1 +	0	-
9	chr1	33514012	33516570	XR_002959291.1 CSMD2 chr1 +	0	-
10	chr1	33521466	33521550	XR_002959295.1 CSMD2 chr1 +	0	-
11	chr1	33546043	33546219	XM_017000191.1 CSMD2 chr1 +	0	-
12	chr1	41506548	41511266	XM_011541884.2 HIVEP3 chr1 +	0	-
13	chr1	41847189	41848445	NR_038261.2 HIVEP3 chr1 +	0	-
14	chr1	58758487	58760197	XR_002958555.1 LOC112268263 chr1 +	0	-

# Dapars and Dapars2

## Generate mapping\_bam\_location\_with\_depth.txt

wig	depth
wig/sample1.wig	76485651
wig/sample2.wig	57005615
wig/sample3.wig	68960763

## Generate configuration file

```
python DaPars2_Multi_Sample_Multi_Chromosome.py Dapars2_configure_file chrList.txt Chr
```

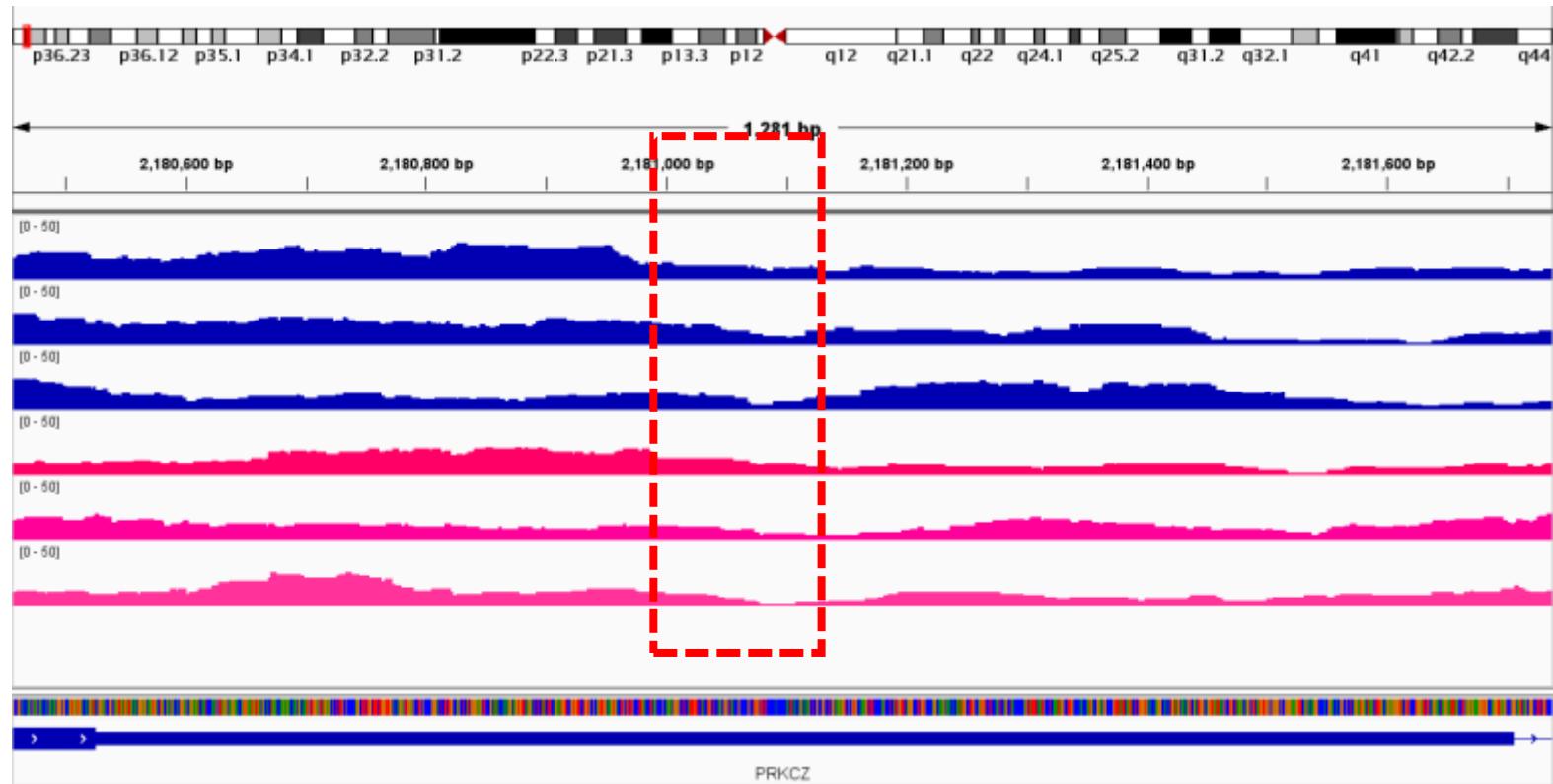
如果需要跑所有染色体需要提供一个染色体编号的文件

```
# Specify the reference of 3'UTR region  
Annotated_3UTR_RefSeq_hg38_3UTR_annotation.bed  
  
# A comma separated list of wig files of all samples  
Aligned_Wig_files=sample1.wig,sample2.wig  
  
Output_directory=Dapars2_test/  
  
Output_result_file=Dapars2  
  
# Specify Coverage threshold  
Coverage_threshold=10  
  
# Specify the number of threads to process the analysis  
Num_Threads=4  
  
# Provide sequencing depth file for normalization  
sequencing_depth_file=mapping_wig_location_with_depth.txt
```

configuration file

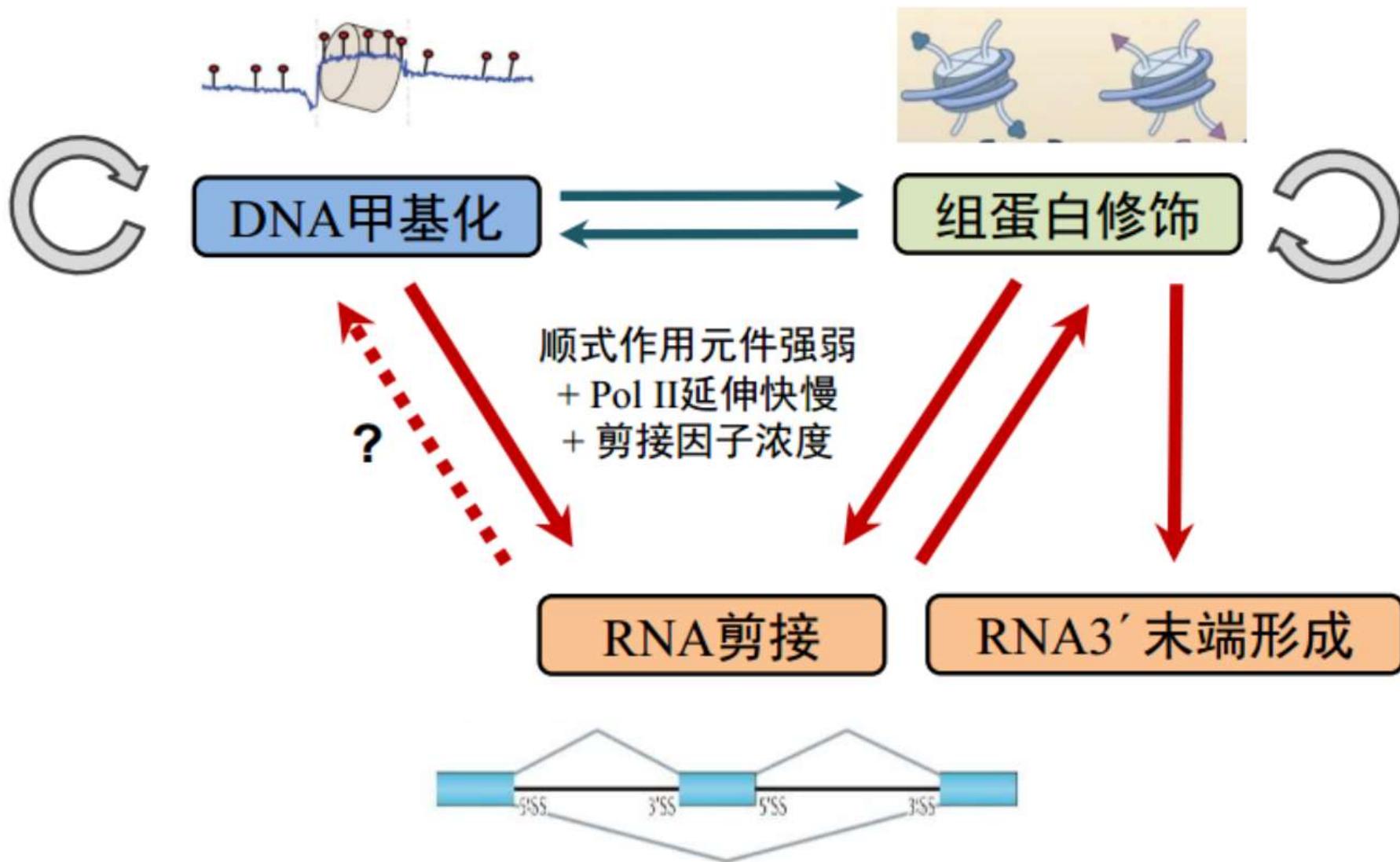
# Dapars and Dapars2

- Gene
- fit\_value
- Predicted\_Proximal\_APA
- Loci
- <Samples>\_PDUI

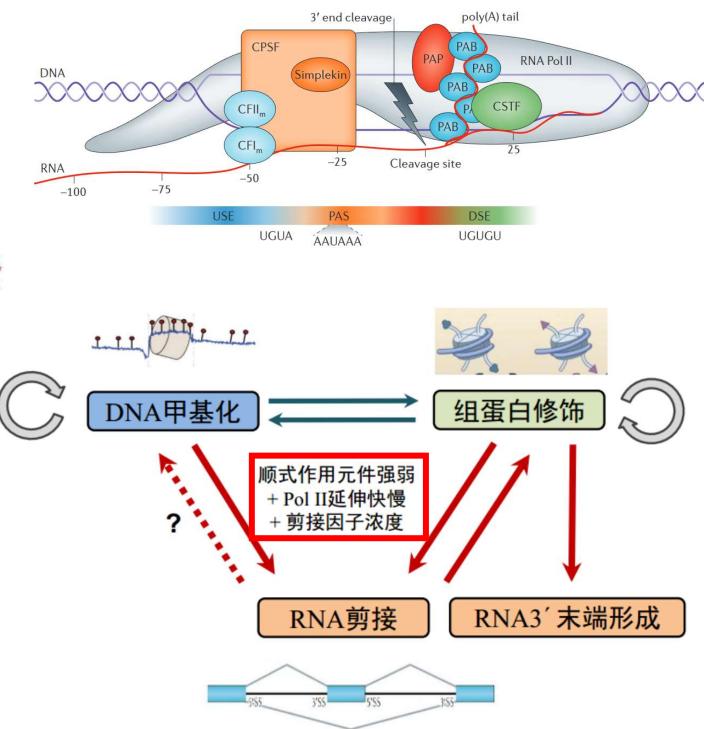
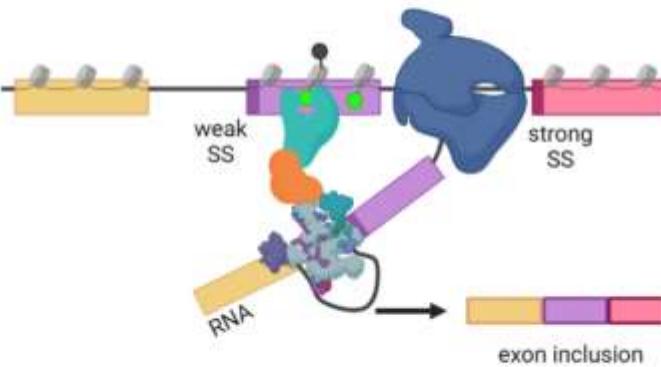
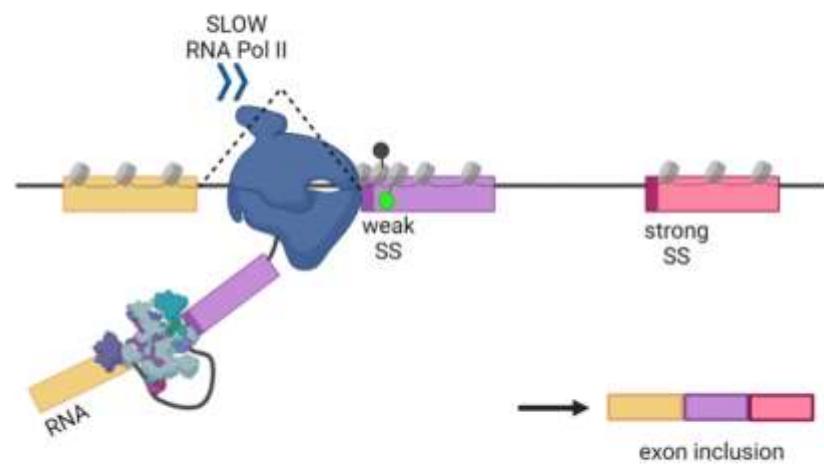
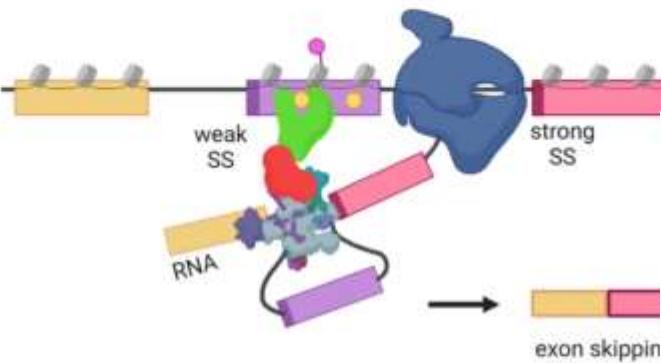
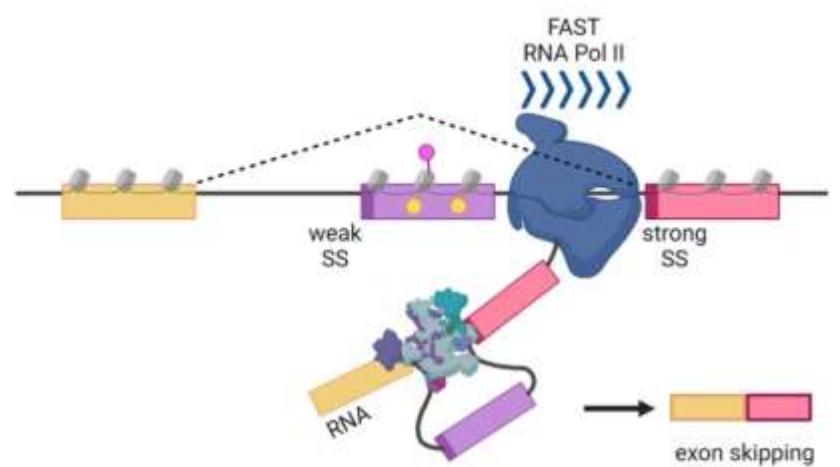


Gene	fit_value	Predicted_Proximal_APA	Loci	HG00119_PDUI	HG00120_PDUI	HG00121_PDUI	HG00122_PDUI	HG00123_PDUI	HG00124_PDUI
NM_001145278.2 NECAP2 chr1 +	1952.3	16785796	chr1:16785337-16786573	0.85	0.84	0.58	0.91	0.68	0.88
NM_013943.3 CLIC4 chr1 +	592.3	25167531	chr1:25167264-25170812	0.91	1.00	1.00	1.00	1.00	0.89
NM_001918.5 DBT chr1 -	24.9	100659631	chr1:100652475-100661978	0.19	NA	0.24	0.14	0.13	0.20
NM_001303229.2 PRUNE1 chr1 +	39.8	151006577	chr1:151006282-151008189	0.73	NA	0.74	0.96	1.00	0.89
NM_022457.7 COP1 chr1 -	108.9	175914069	chr1:175913967-175914306	0.68	0.67	0.70	0.54	0.73	0.71
NM_030806.4 C1orf21 chr1 +	21.8	184589304	chr1:184588652-184598153	0.06	0.03	0.03	0.05	0.03	0.04
NM_001366241.1 LIN9 chr1 -	12.3	226420039	chr1:226418850-226420307	NA	NA	NA	0.70	0.69	0.49

# AS and APA epigenetics regulation



# AS and APA epigenetics regulation



<https://mp.weixin.qq.com/s/rKax5HVLzrySH8p3BfAGiA>



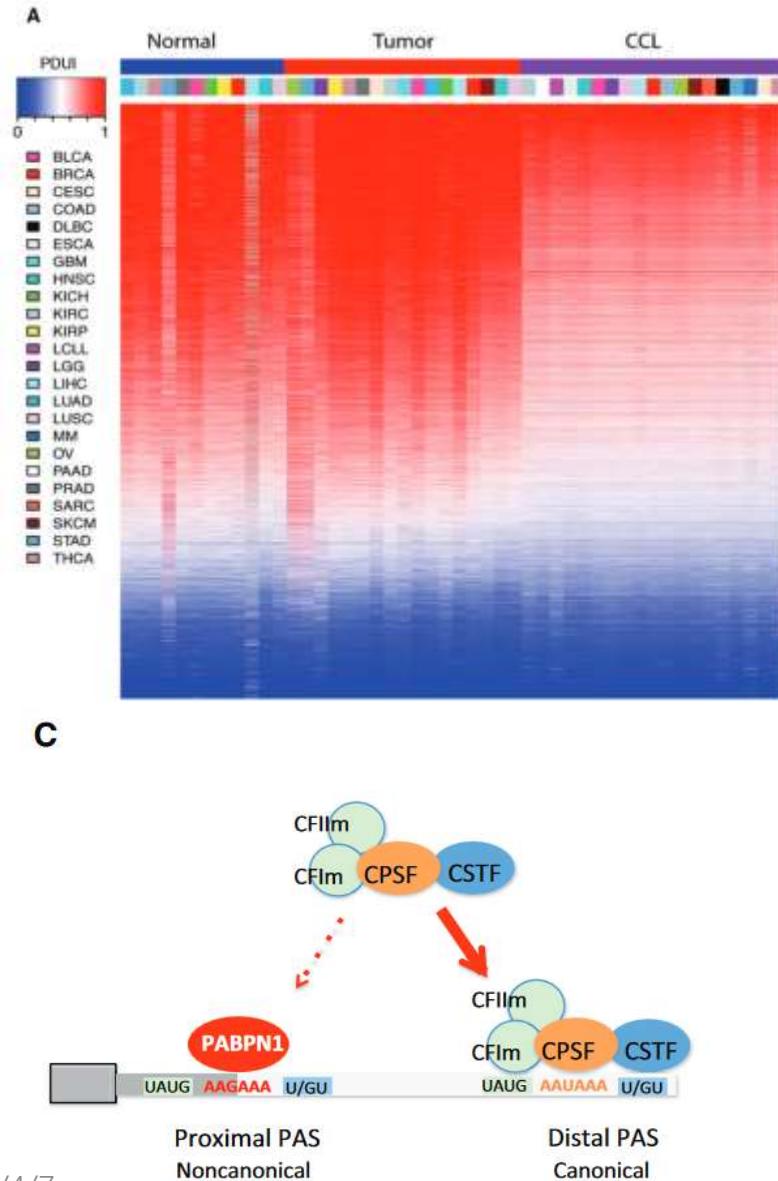
# AS epigenetics regulation

- 剪接位点的强弱决定这个位置是否会跳掉，位置竞争
- U1, U2和一些snRNP调控 (interplay protein)
- exon definition (**binding pos**)-motif
- 调控的两个特征：序列依赖、位置依赖
- 重要调控蛋白：SR和hnRNP
- **Pol II的延伸速度**
- 核小体分布影响pol II延伸速度
- 组蛋白修饰结合剪接因子，促进或抑制剪切
- 组蛋白修饰酶可能招募剪接因子 (interplay protein)
  - 组蛋白修饰酶改变->组蛋白修饰
- 染色质甲基化修饰影响可变剪接
  - 甲基化影响组蛋白修饰->影响可变剪接
  - 改变速率
  - 招募剪接因子
    - 关注甲基化结合蛋白，不同蛋白对甲基化分布结合不同
    - 如CTCF有甲基化结合偏好性，影响剪接；MeCP2和HP1

# APA epigenetics regulation

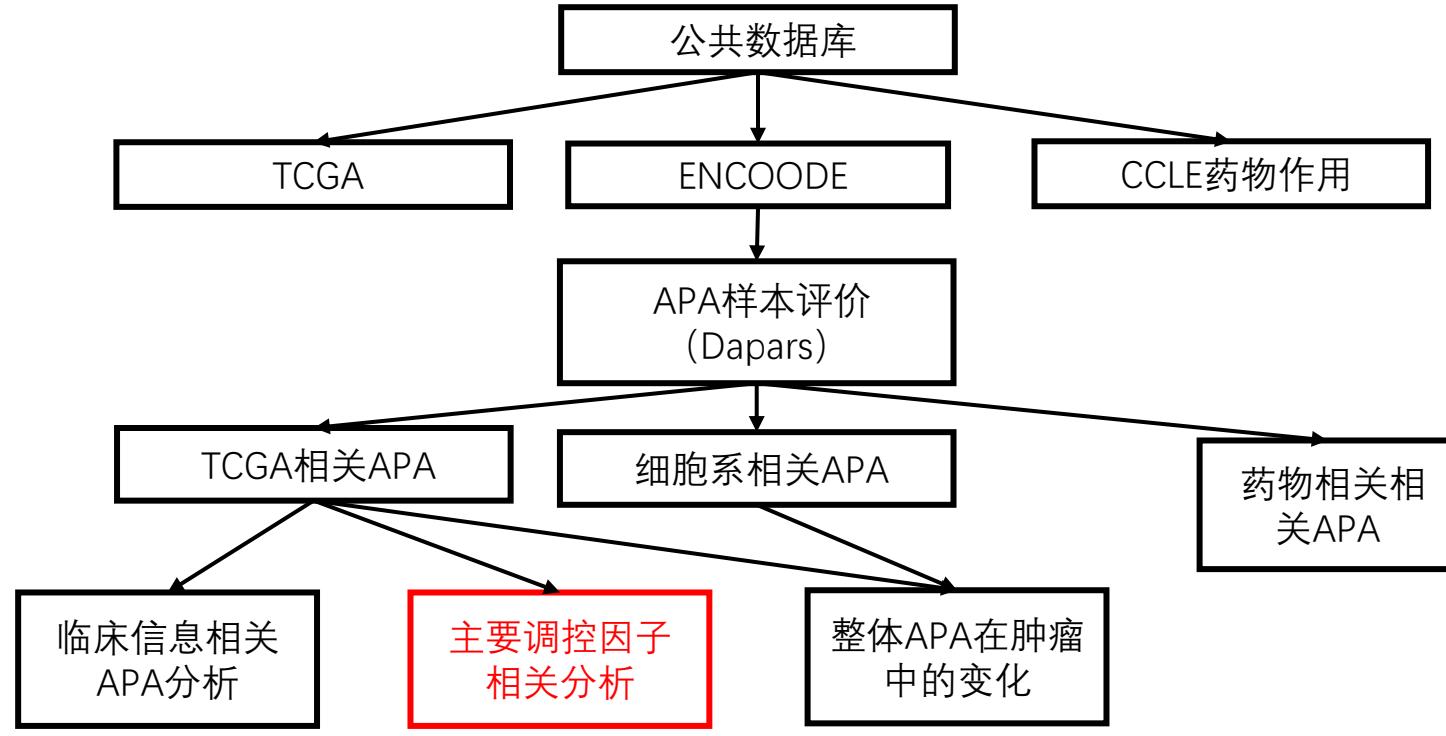
- Regulation by 3'-end-processing **factors**
- APA and **chromatin**
  - epigenetic modification
  - nucleosome positioning might influence PAS
- **Interplay between splicing and APA**
  - Two types of APA events affected by splicing and 3' end processing
    - **alternative terminal exon**
    - **intronic APA**
  - a dynamic **competition between splicing and polyadenylation** (support by RNAseq correlation between patterns of alternative splicing and APA)
- Trans-acting factors (RBPs regulation)

# APA epigenetics regulation



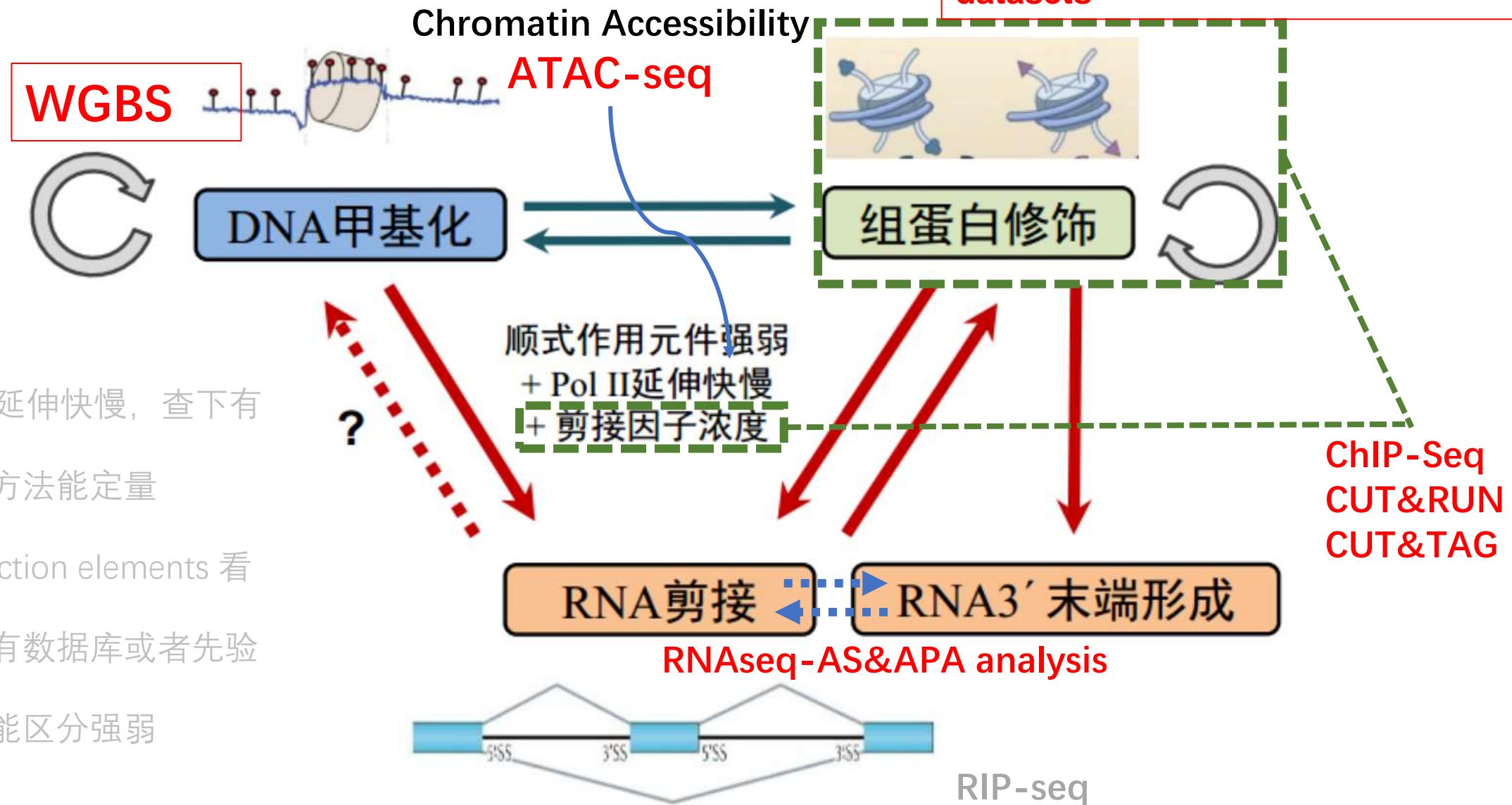
## ARTICLE Comprehensive Characterization of Alternative Polyadenylation in Human Cancer

Yu Xiang\*, Youqiong Ye\*, Yanyan Lou, Yang Yang, Chunyan Cai, Zhao Zhang, Tingting Mills, Ning-Yuan Chen, Yoonjin Kim, Fatma Muge Ozguc, Lixia Diao, Harry Karmouty-Quintana, Yang Xia, Rodney E. Kellems, Zheng Chen, Michael R. Blackburn, Seung-Hee Yoo, Ann-Bin Shyu, Gordon B. Mills, Leng Han

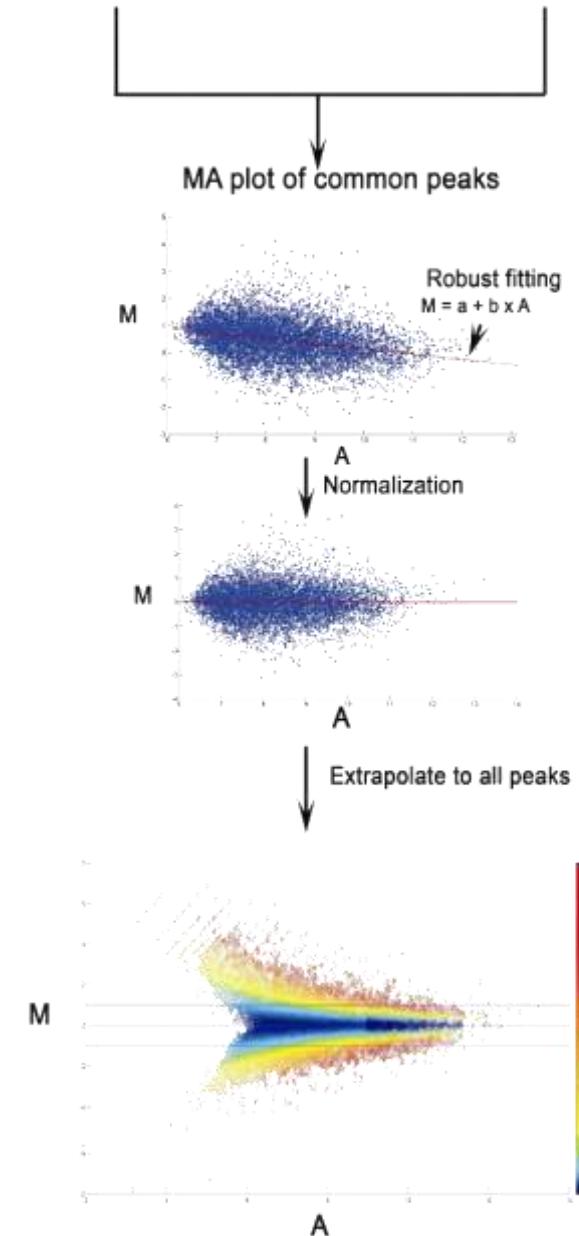
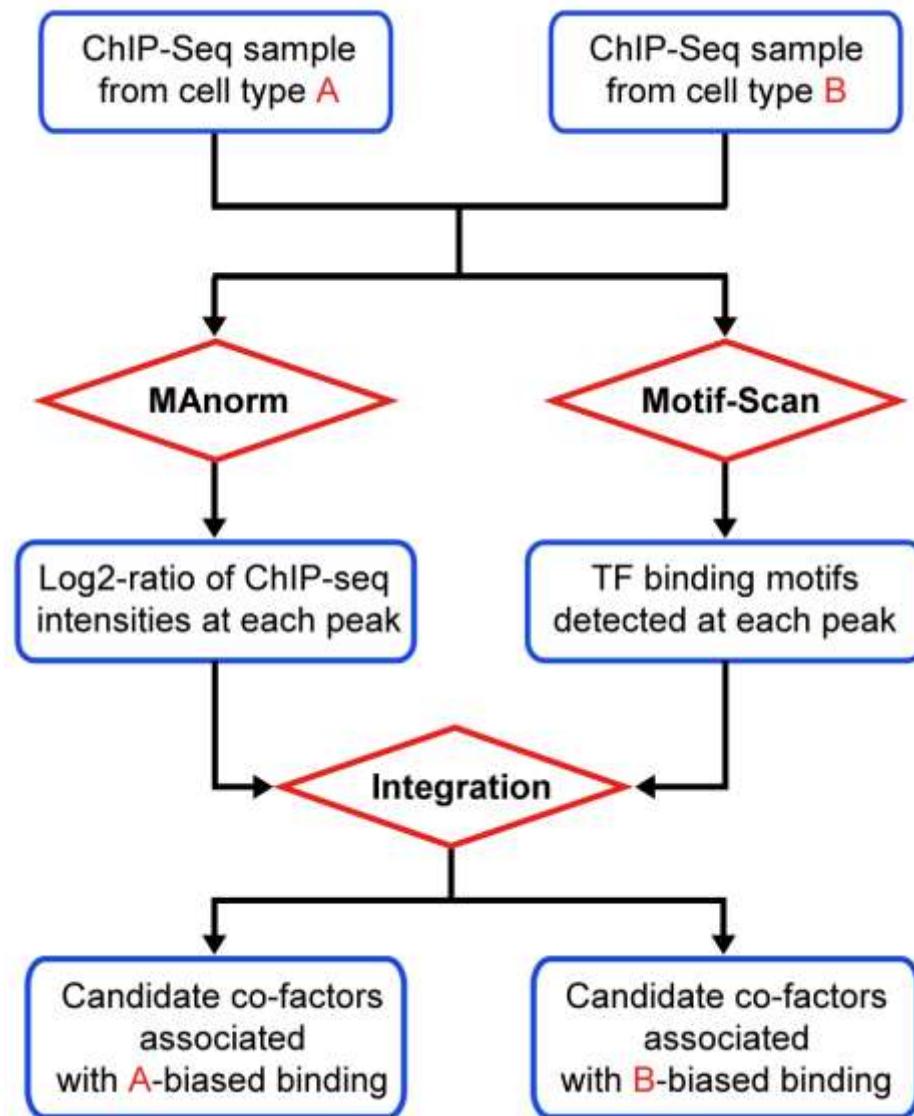


# **Epigenomics data integrative**

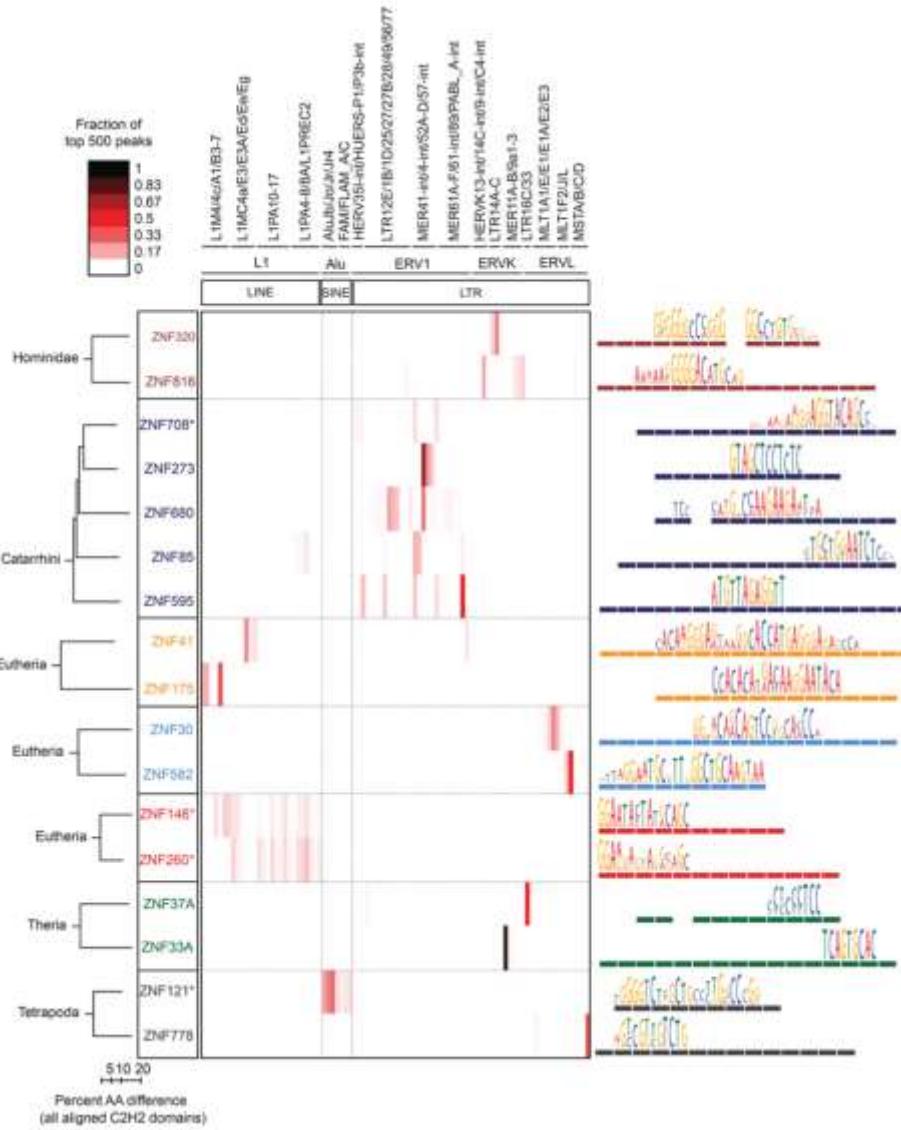
## Histone modification marker Writer/Reader/Eraser protein datasets



# Epigenomics data integrative



# Epigenomics data integrative



# Multiparameter functional diversity of human C2H2 zinc finger proteins

Frank W. Schmitges,<sup>1,4</sup> Ernest Radovani,<sup>2,4</sup> Hamed S. Najafabadi,<sup>1,4,5</sup> Marjan Barazandeh,<sup>1,4</sup> Laura F. Campitelli,<sup>2,4</sup> Yimeng Yin,<sup>3</sup> Arttu Jolma,<sup>1,3</sup> Guoqing Zhong,<sup>1</sup> Hongbo Guo,<sup>1</sup> Tharsan Kanagalingam,<sup>1</sup> Wei F. Dai,<sup>1</sup> Jussi Taipale,<sup>3</sup> Andrew Emili,<sup>1,2</sup> Jack F. Greenblatt,<sup>1,2</sup> and Timothy R. Hughes<sup>1,2</sup>

<sup>1</sup>Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Toronto, Ontario M5S 3E1, Canada; <sup>2</sup>Department of Molecular Genetics, University of Toronto, Toronto, Ontario M5S 1A8, Canada; <sup>3</sup>Department of Biosciences and Nutrition, Karolinska Institutet, SE 141 83, Sweden

## Motif:

- 不仅存在于DNA序列, RNA(RBP、splicing Factors), Protein (Domain, epitopes)都存在motif
- Key: pattern discovery
- Generative model (summary and measure of information, integrate many positions)

# Epigenomics data integration

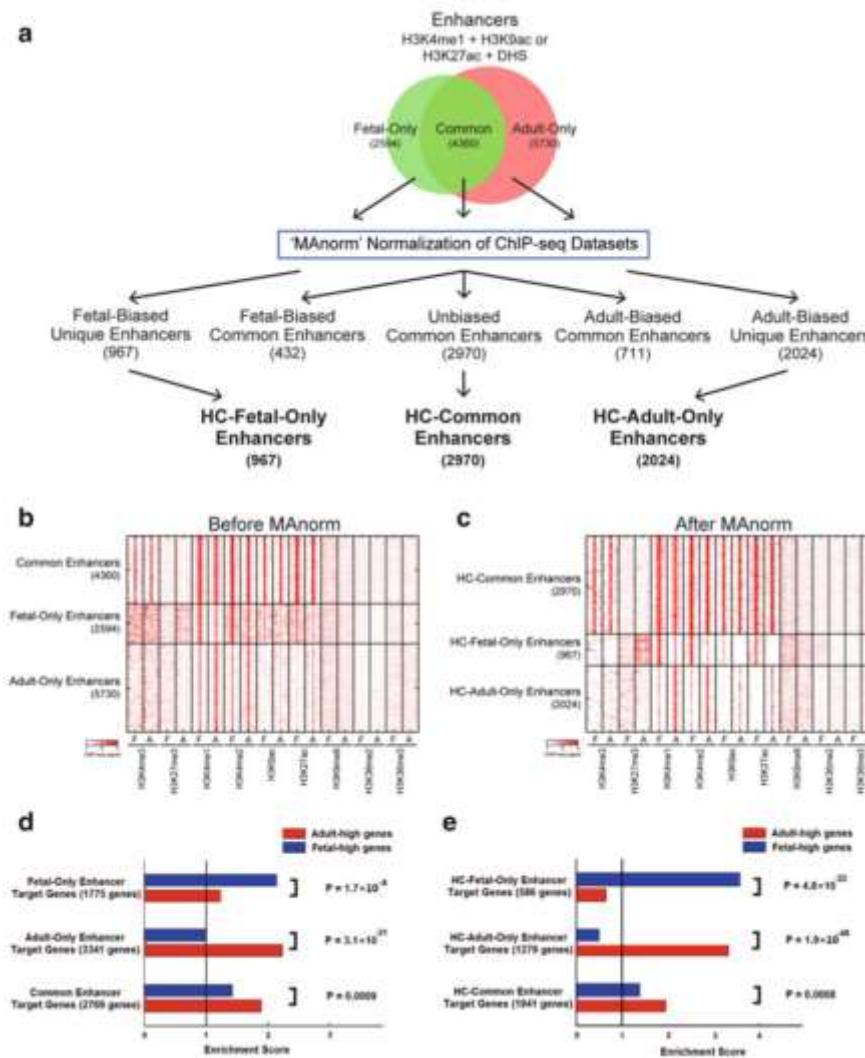
## Importance of predefined before integration

Defined the putative distal enhancers:

- Are covered by **H3K4me1** and **H3K9ac** or **H3K27ac** peaks
- Contain at least one **DNase I hypersensitive site** and do not contain any H3K27me3 peak
- Are located **at least 2 kb away** from any RefSeq annotated gene's **transcription start site**

### High-Confidence definition

- HC-common enhancers are those common enhancers with  $1 < \text{Maximum}(\text{MH3K27ac}, \text{MH3K9ac}) < 1$ .
- HC-fetal-only enhancers are those fetal-only enhancers with  $\text{Minimum}(\text{MH3K27ac}, \text{MH3K9ac}) < 1$  and  $\text{Maximum}(\text{MH3K27ac}, \text{MH3K9ac}) < 0$ .
- HC-adult-only enhancers are those adult-only enhancers with  $\text{Maximum}(\text{MH3K27ac}, \text{MH3K9ac}) > 1$  and  $\text{Minimum}(\text{MH3K27ac}, \text{MH3K9ac}) > 0$ .



## Next plan

- Process CUT&TAG data, and while at it, conduct ATAC-seq.
- Refine APAs and AS analyses.
- Learn WGBS pipeline.
- Organize motif analysis, integration methods like MAnorm, and consolidate learning.
- Compile Epigenomics-related databases or data resources.
- Build initial framework.
- Draft review article.

[\*\*Back to Content\*\*](#)

# **Analysis of Different Types of Biological Data**

——image data

Zhou Ying

TCCAAAGGAGCTTGGCAGGCAGAGCTGCCCTGCAGCAACACTCCCATTCTGCTTGTGAGTAGCACCTAGATTTCAAGG  
 CTCATCTAGATACTCGAGTTGATTCTGGGCTTTATAGGCCAGTTGCTGGGACTGTTCAGGAGCTAGGGCCATGTTGGGAGGGAGG  
 GGGCACAAAAGTAGAGAACGCTGATGTTGATCCCAGGGGCTGTCAGCTCTGCTACTGCTCCCTGCAAGATGTCAGGTGCT  
 GTCACGTGCTGGCTTGTCAGTCATGGCAGGAGAATGGGCTGTCAGTGTGACATTGGGCTGAGGCTGCCCCTGTGTCAGAG  
 TTCAGGGTGTAGGAGGGGTTGCGCAACCATTGGGCTGTTGGGTAAGTGGGATCTTGTGCTGGGCTAAGGTGATCCTGAGCC  
 TTGGCCTGTGGAATGGGGTAGAGGGCAAATGGTAACCTAGCATGCTGTTGGGATATTAGATGAGGGGCTGCCGAGGCTCGGGAGGG  
 TCCTAGGGAGCAGATGTTGAAGAGGCCAGGCCCTCAGTGAGCTGGATGAGGGTGAACGCUCTGAGGGTACTCTGG  
 GCCTCGTGTAAATGGCTCTCTGTATGTCCTTCAACCGCTGCCAGATAACCCCCACGTGCCATCA

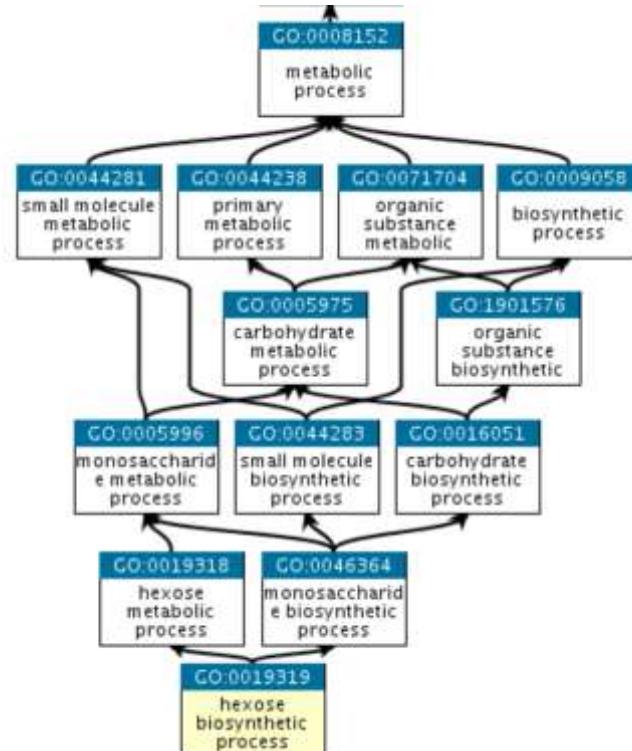
exon9

S A H Q V A R Y R P A P I I

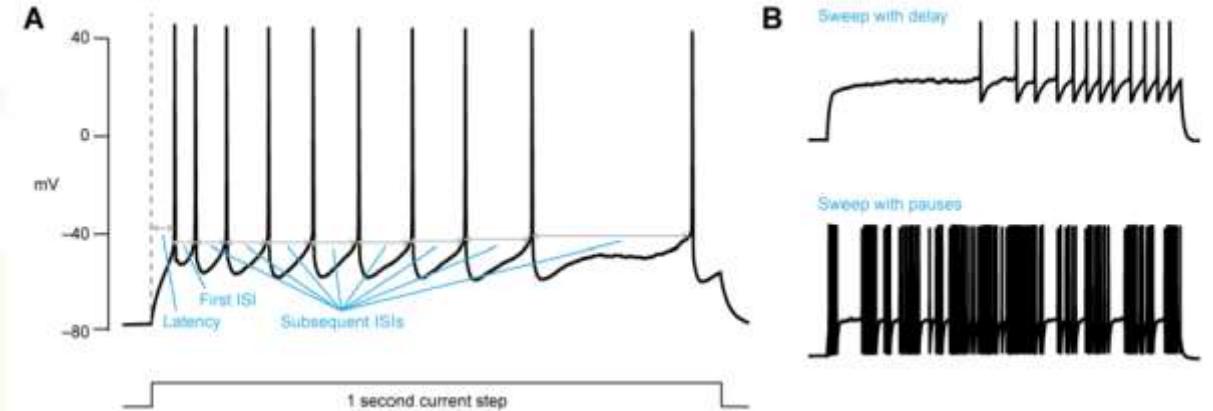
TGCTGTGACCCGAATCCCCAGACAGCTCGTCAAGGCCACCTGTACCGTGGCATCTCCCTGTGCTGTGCAAGGACCCAGTCAGGGAG  
 A V T R N P O T A R Q A H L Y R G I F P V L C K D P V Q E J  
 CTGGGCTGAGGAGCTGGACCTCCGGGTGAACTTTGCCATGAATGTTGTTGACGTGGCTGGAGCAGGGCTAGAGCCCTAGAGGAGCTTGGG  
 W A E D V D L R V N F A M H V  
 ATGCTTGAGCATGGCTCTGTGGGACCCGAAAGTTGGGAATAGAAAGGGGACACACAGACCTTAGTGGGCAAAGGCCAGCG  
 CTGTTCTCTCCCTTATGGGAATGTTCATCTGAATCTCACTCTGGGCTCAAAGCAAATGTCAATTGGCATCTGCTCATTACTGA  
 TTGAGGGCTGAGCAATTGAGTCTCTCTCTAGTCTGGGCTCAAAGCAAATGTCAATTGGCATCTGCTCATTACTGA  
 AGTTTTTTGTTTTGTTTTAAATAAAATTGCCACAGCTCTGTGCTGGGAGGATACAGATACAGATTACGTACTGATG  
 GCCATGTGCTGTATAAGGTAGGGTATCATCAGATGACAGGAAGCAGCTGCTGGGAGGCTTGCACCCCTCAGGAG  
 AGTGAATGATGTCCTGAGGTAGGGTATCATCAGATGACAGGAAGCAGCTGCTGGGAGGCTTGCACCCCTCAGGAG  
 TTTGGTGGCACCTGTGGTGTGAGGAAGTAGTGGTCAGCCAGCCTAGTCACTCAGGCTATGGGCACTGGGCTGGGCTGTGA  
 CCTGGCTGGCTTTACTACCAACCTCCCTCTTCCAGCAAGGCCAGGCTTCTCAAGAAGGGAGATGTGGTCAATTGTC

exon10

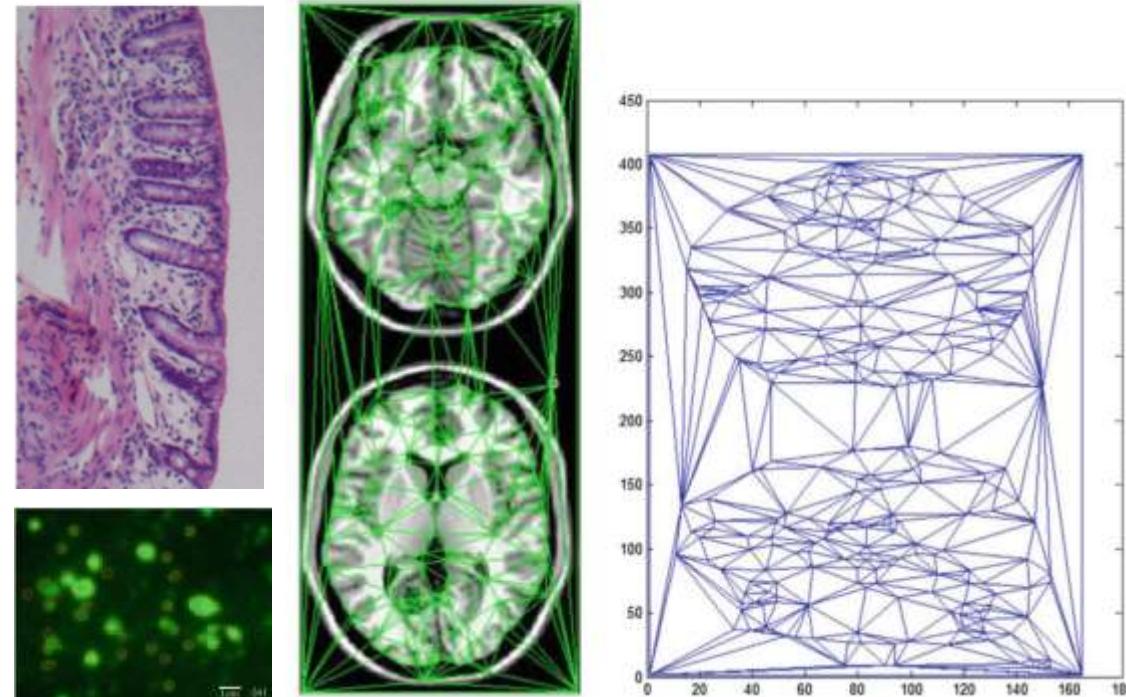
## 文本数据/测序数据)



## 术语/知识库类型数据



## 波谱数据/时间序列数据



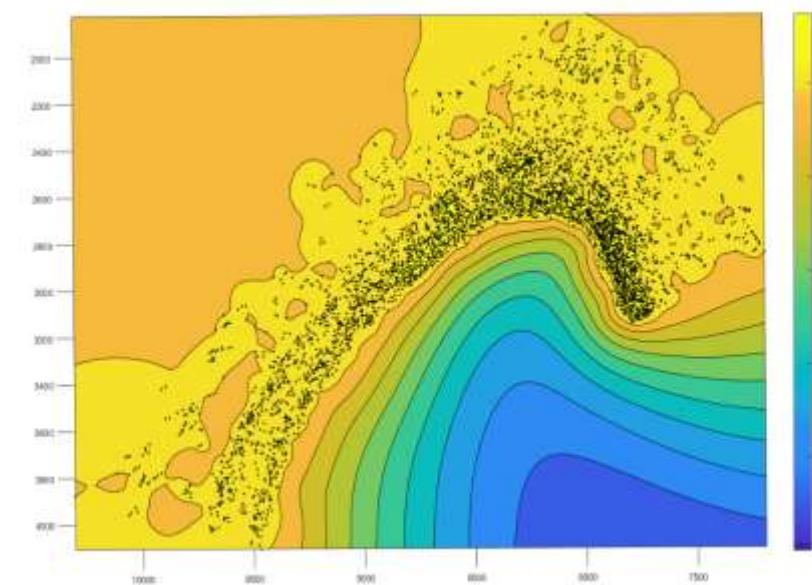
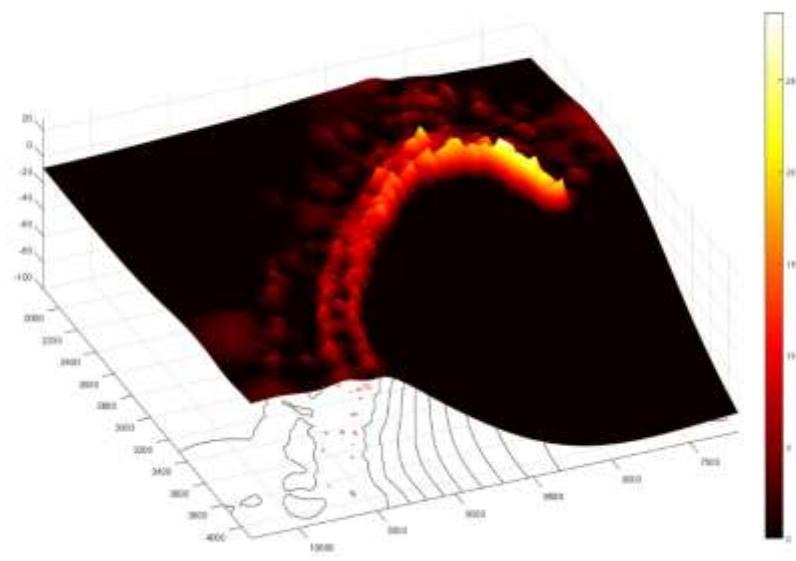
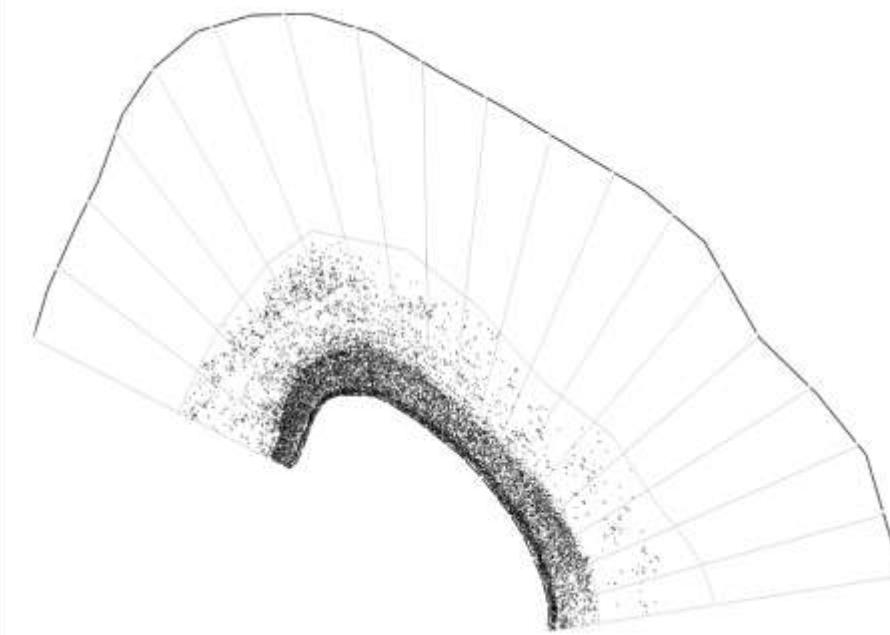
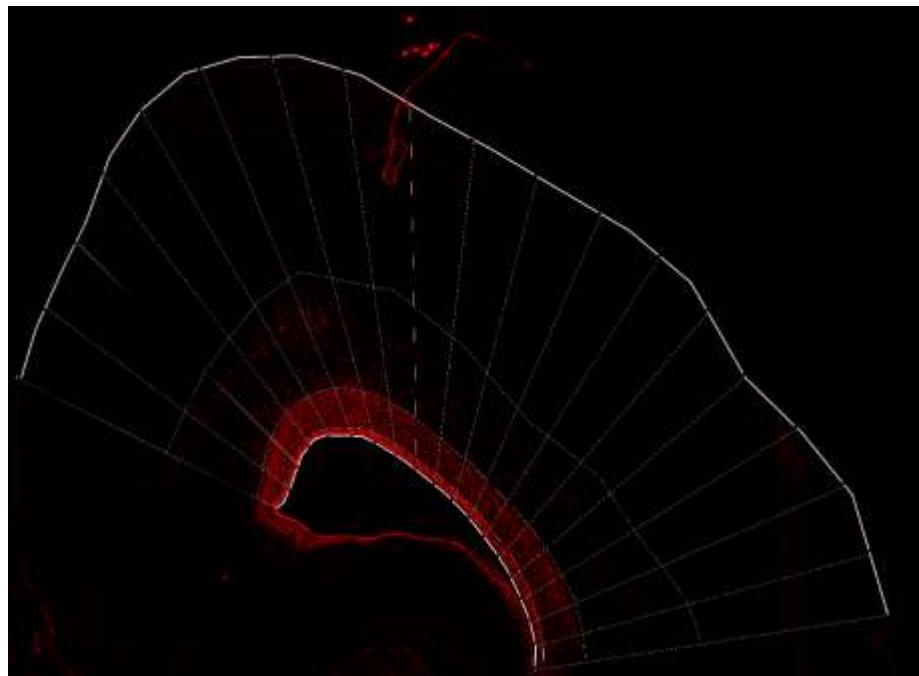
## 图像数据

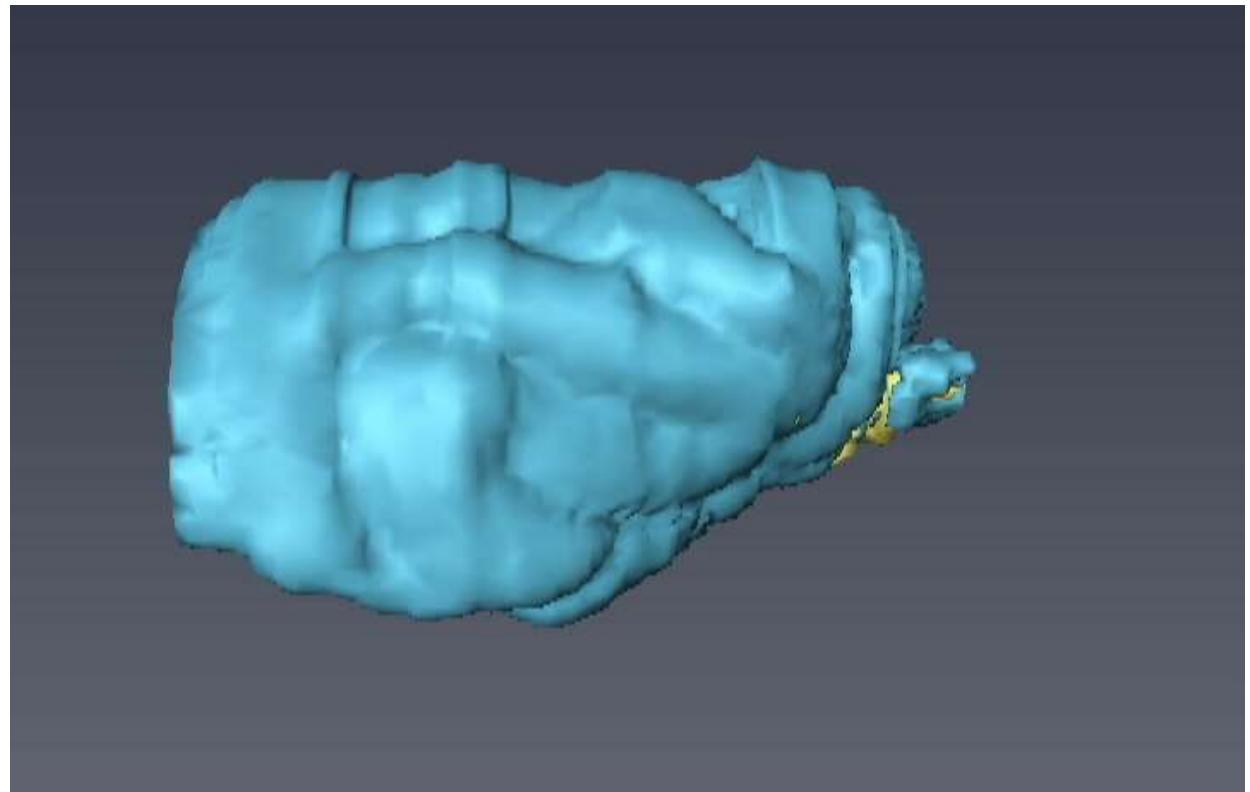
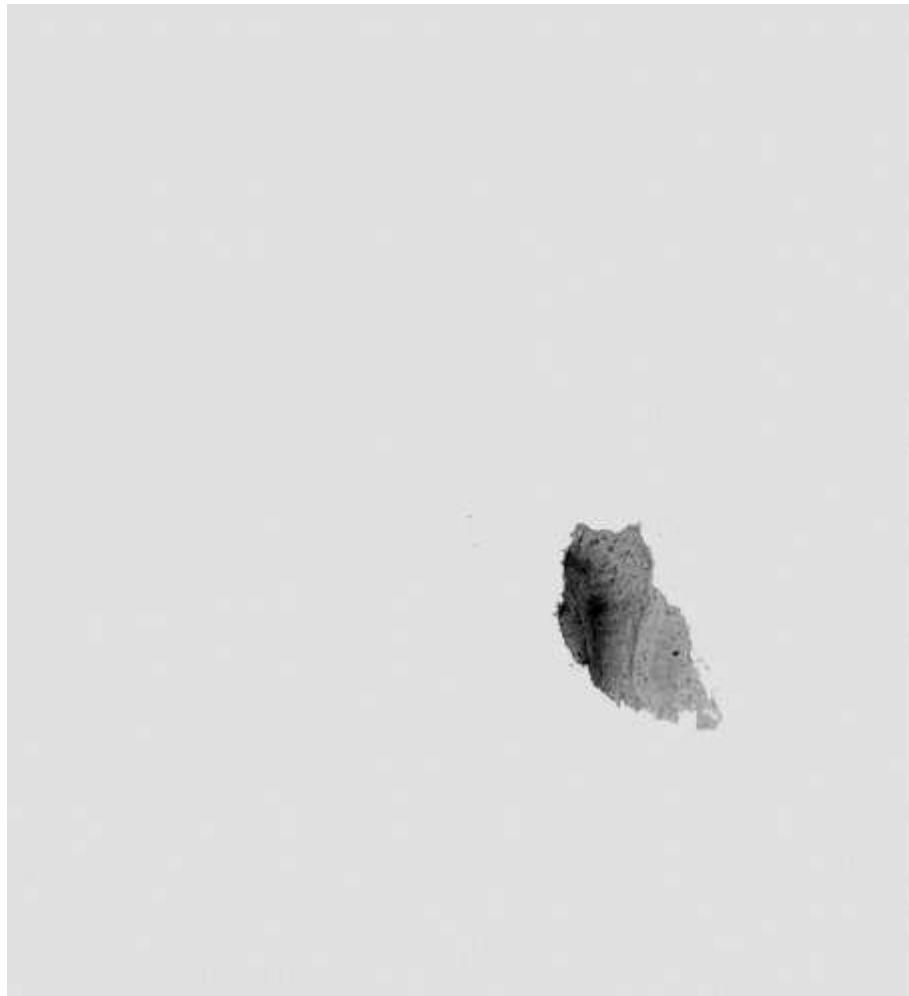
**1.细胞和组织图像分析：**生物图像处理可以用于细胞和组织图像的分析，例如测量细胞大小、形状和颜色等特征，检测和计数细胞或细胞核，以及在组织中定位和分割不同类型的细胞。

**2.显微镜图像分析：**生物图像处理可以用于显微镜图像的分析，例如在细胞或组织中识别和跟踪特定蛋白质或细胞结构，测量细胞或组织中不同区域的荧光强度，以及在显微镜图像中检测和计数微生物。

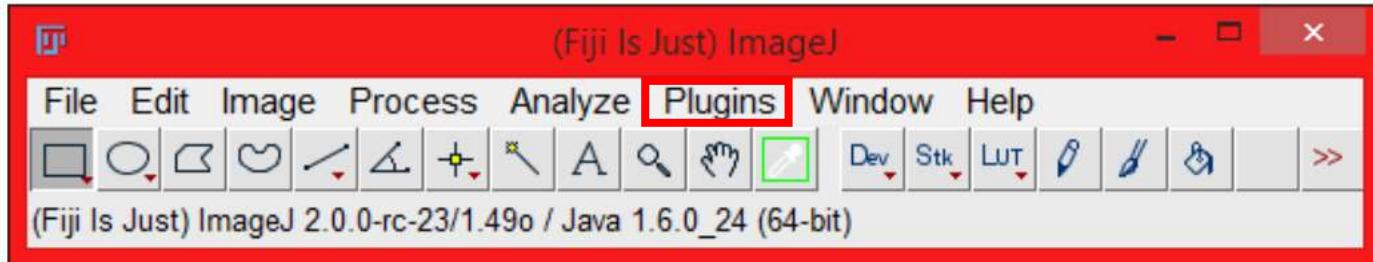
**3.医学影像分析：**生物图像处理可以用于医学影像的分析，例如在CT、MRI和X光图像中检测和识别肿瘤，计算病变区域的大小和形状，以及在病理图像中定位和分割肿瘤细胞。

**4.生物信息学分析：**生物图像处理可以用于生物信息学研究中，例如在基因芯片图像中检测和定位基因表达模式，计算不同基因的表达量，以及在蛋白质结构图像中识别蛋白质结构和功能。





Amira

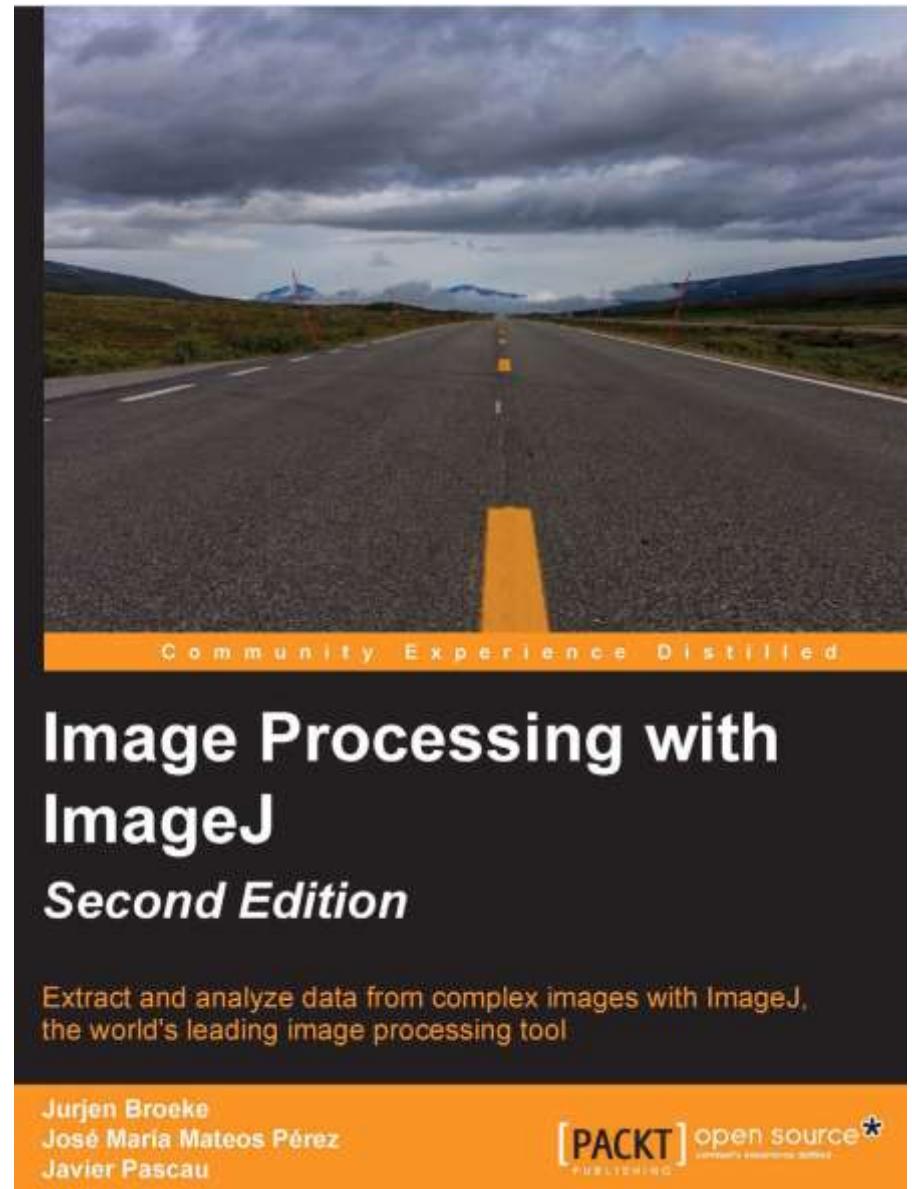


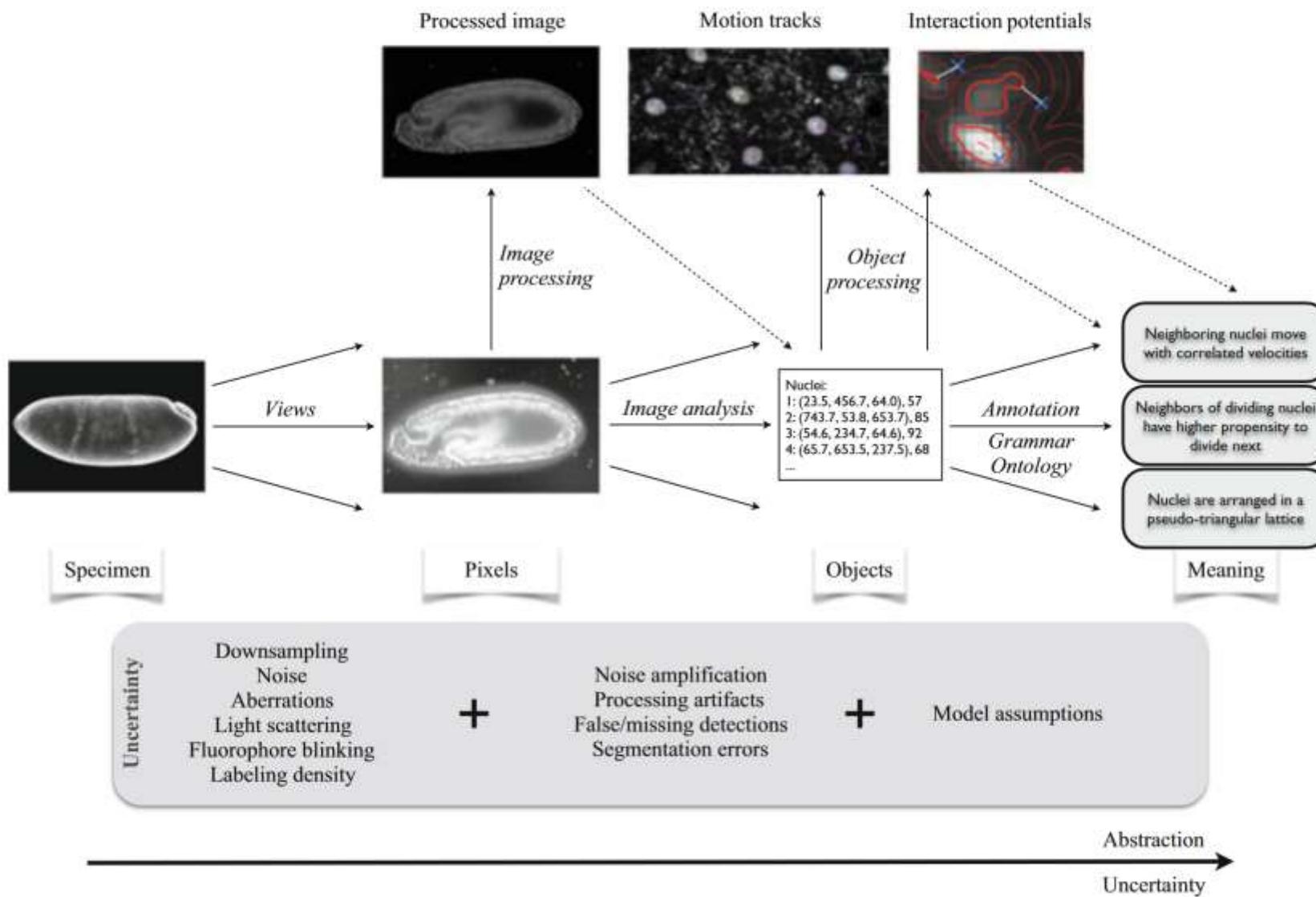
[home](#) | [docs](#) | [download](#) | [plugins](#) | [resources](#) | [list](#) | [links](#)

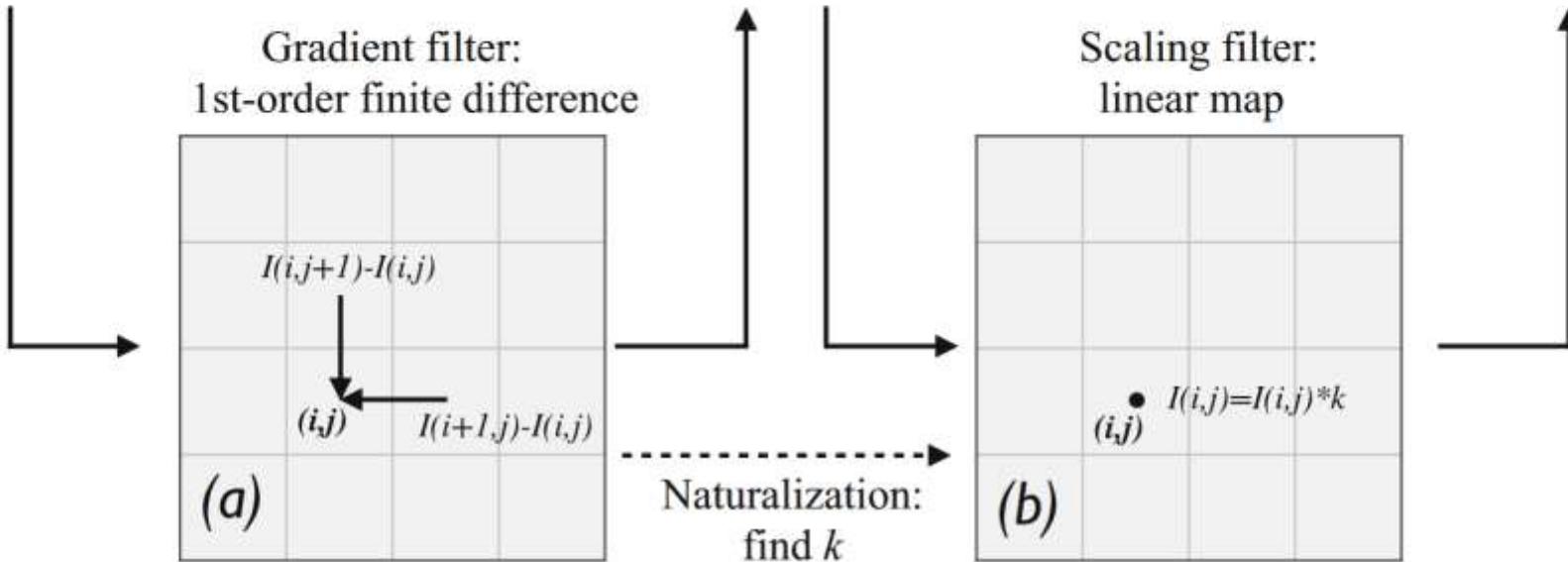
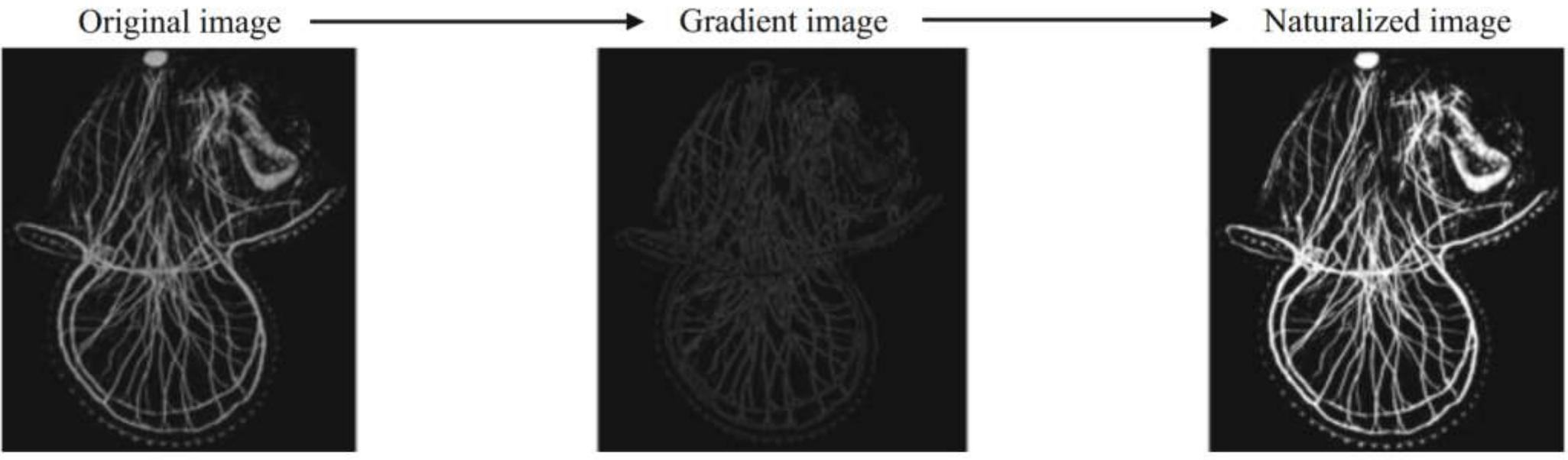
## Documentation

- [Introduction](#)
- [Basic Concepts](#)
- [Installation](#)
- [\*ImageJ User Guide\* \(download PDF\)](#)
  - [User Interface and Tools](#)
  - [Menu Commands](#)
  - [Extending ImageJ](#)
  - [Keyboard Shortcuts](#)
- [Tutorials and Examples](#)
- [Image.sc Forum](#)
- [ImageJ on Wikipedia](#)
- [Frequently Asked Questions](#)
- [Macro Language \(download PDF\)](#)
- [Complete Release Notes](#)
- [NIH Image](#)

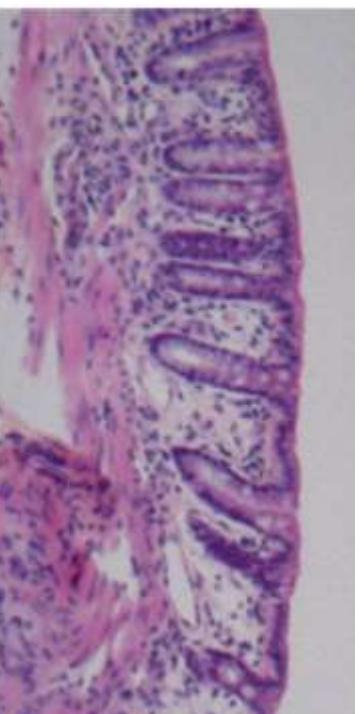
<https://imagej.net/ij/docs/index.html>



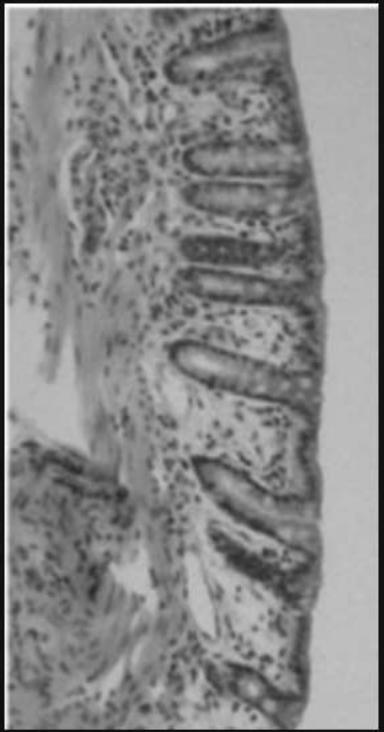




# Python-PIL (pillow)



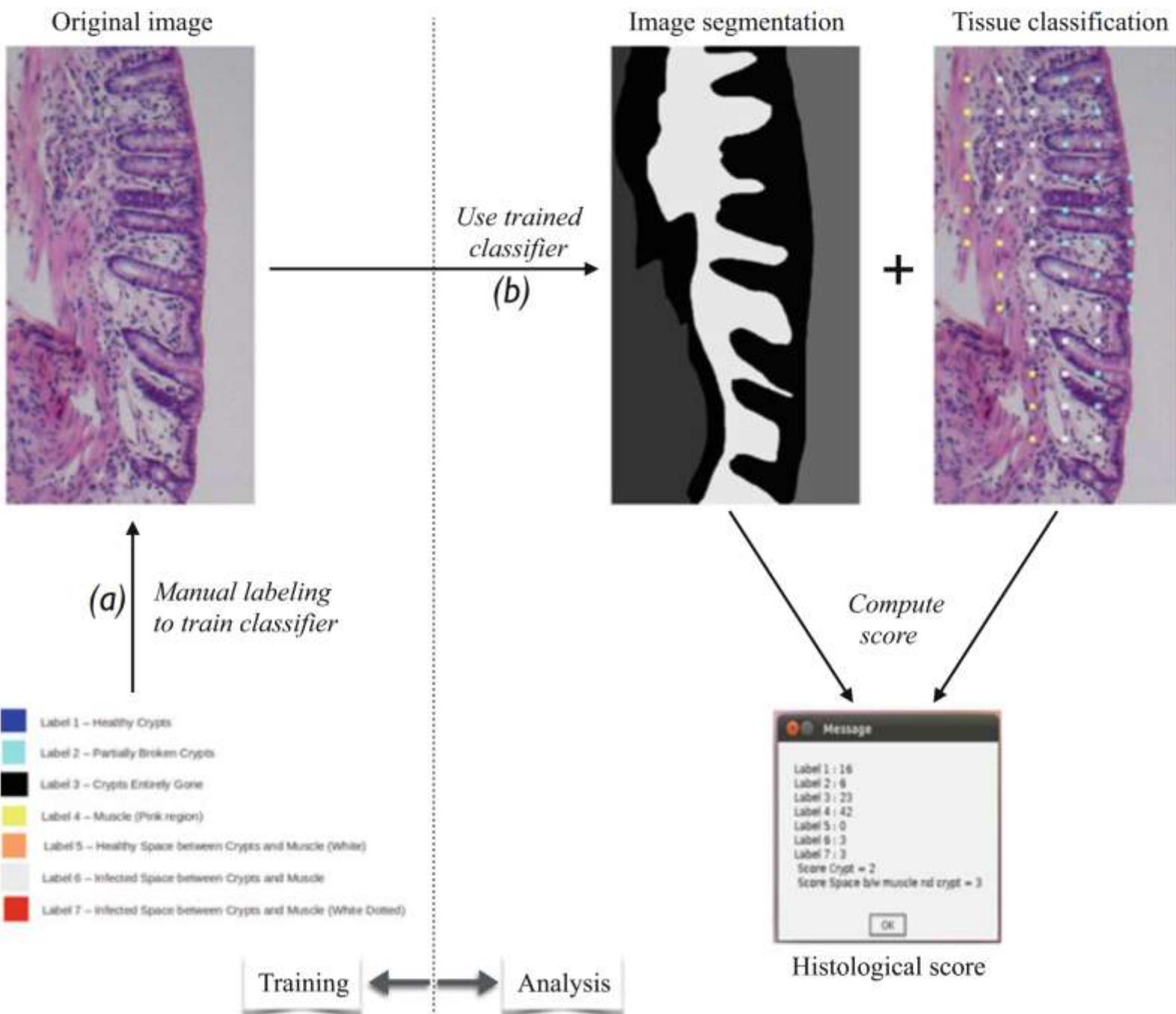
```
[1]: from PIL import Image  
import numpy as np  
  
[2]: image = Image.open('图片1.png')  
  
[3]: image_array = np.array(image)  
  
[10]: im1=image.convert('L')  
im1  
  
[10]:
```

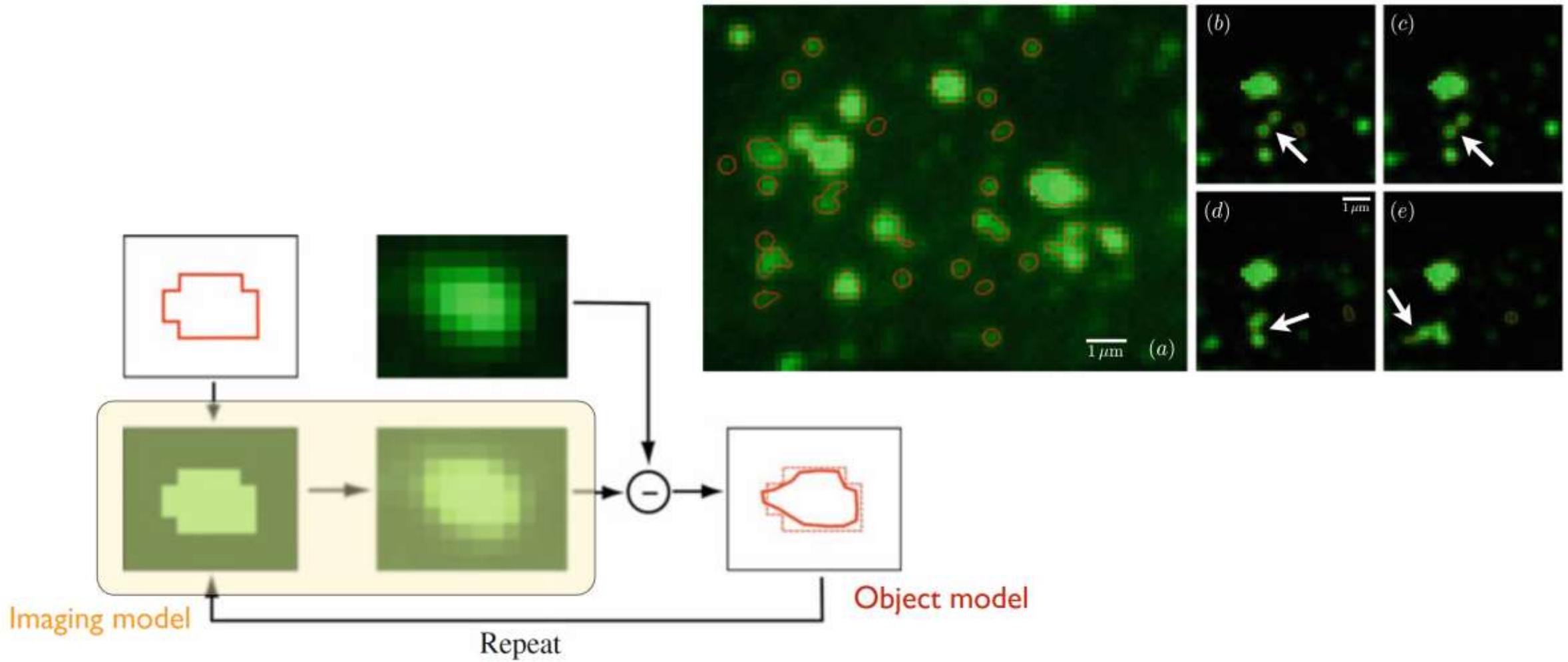


```
[27]: image_array = np.array(im1)  
print(image_array)  
  
[[255 255 255 ... 255 255 255]  
 [255 253 250 ... 249 249 249]  
 [252 226 201 ... 190 189 189]  
 ...  
 [245 195 147 ... 193 192 188]  
 [244 186 132 ... 192 191 187]  
 [245 185 128 ... 190 190 188]]
```

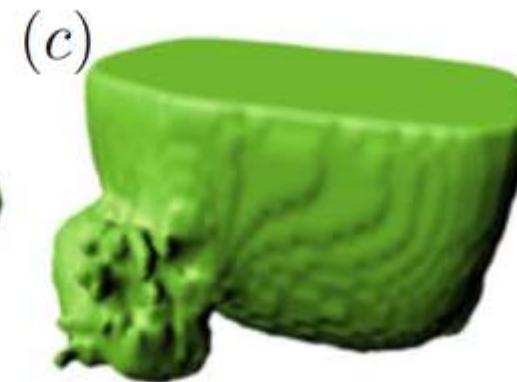
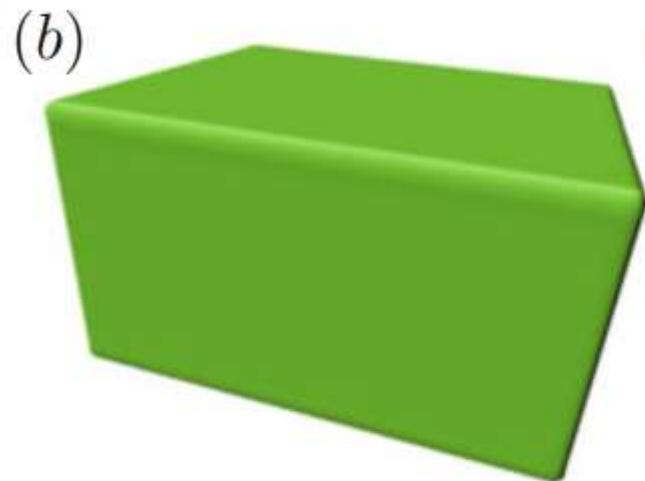
```
[24]: threshold=110  
table=[]  
for i in range(256):  
    if i        table.append(0)  
    else:  
        table.append(1)  
  
img2=im1.point(table,'1')  
img2
```



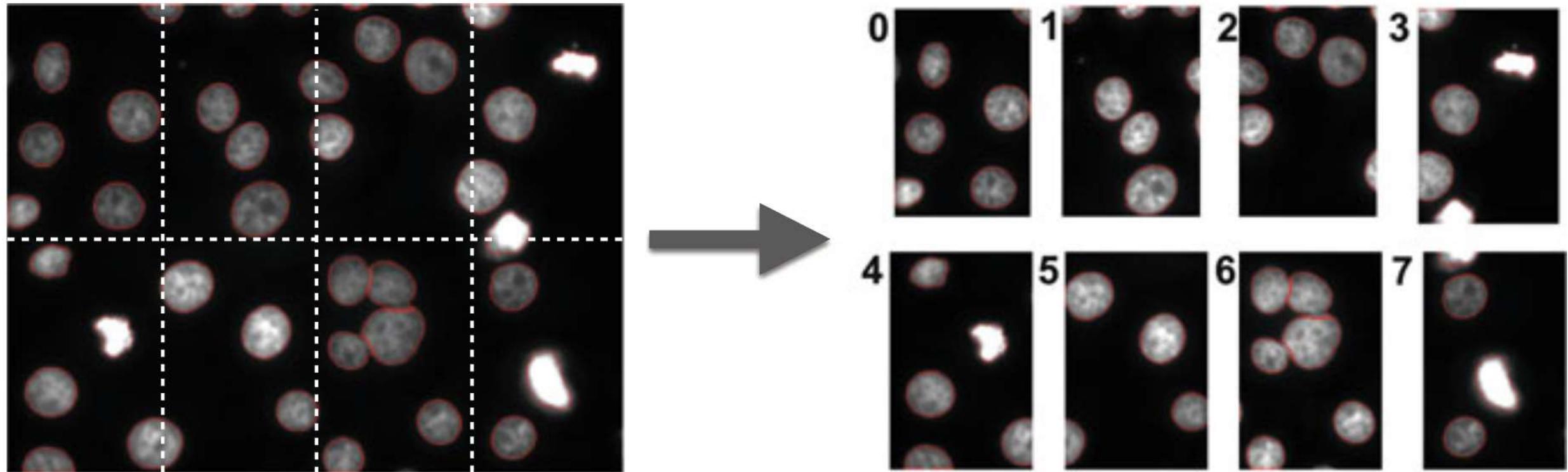




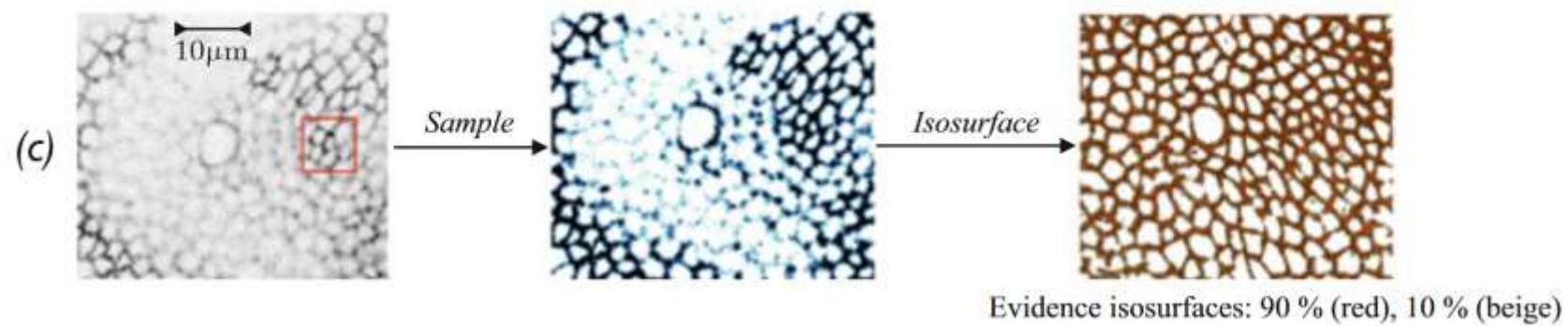
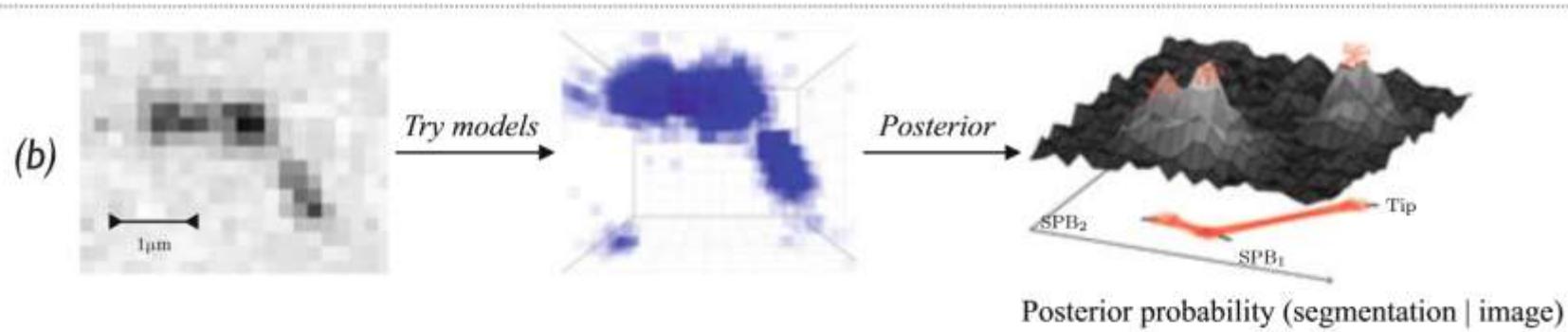
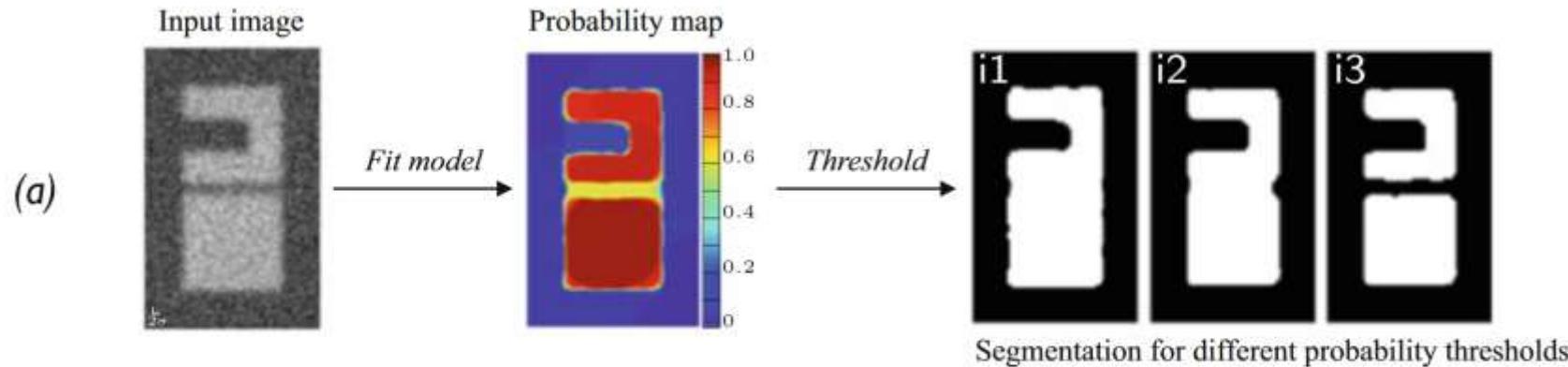
**The model-based approach**



3D model-based segmentation of germ cells in a zebrafish embryo



Domain-decomposition approach to deal with big image data

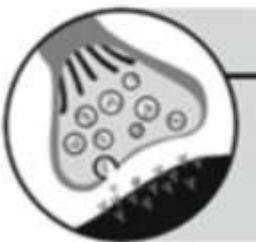


Approaches to uncertainty quantification in image analysis



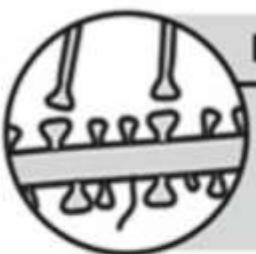
### Neuronal morphology

- Neurite length
- Neurite branching
- Neurite surface area



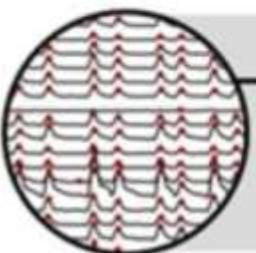
### Synapses

- Synapse density
- Synapse classification



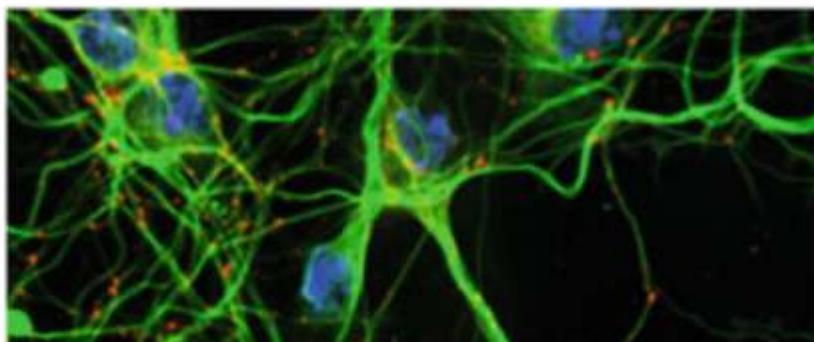
### Dendritic spines

- Dendritic spine density
- Dendritic spine morphology

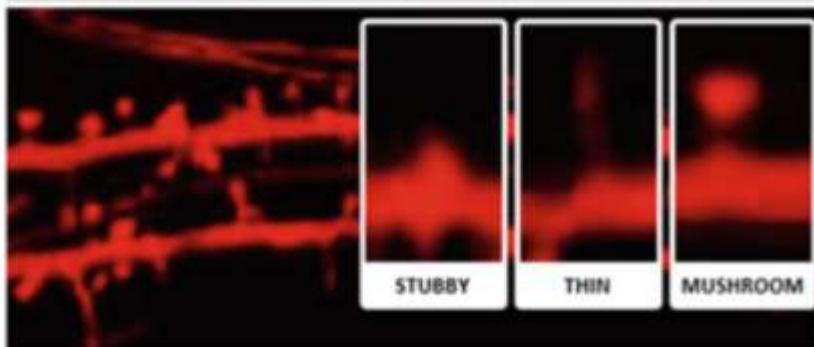


### Calcium activity

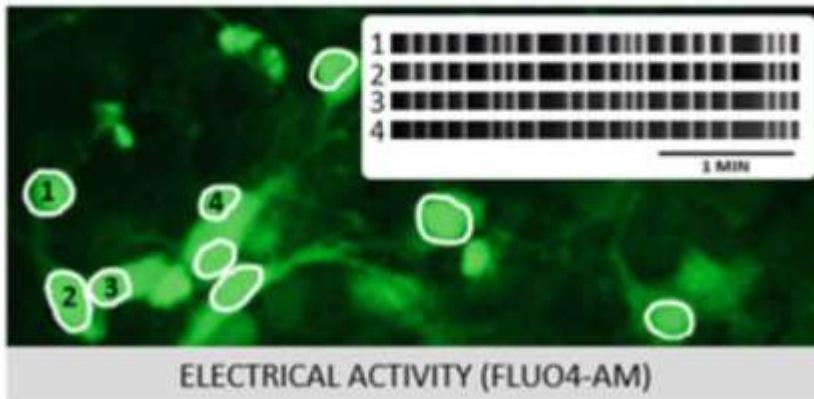
- Bursting frequency
- Bursting amplitude
- Bursting correlation



MICROTUBULES ( $\beta$ III-TUBULIN), SYNAPSES (SYNAPTOPHYSIN-I), CELL NUCLEUS (DAPI)

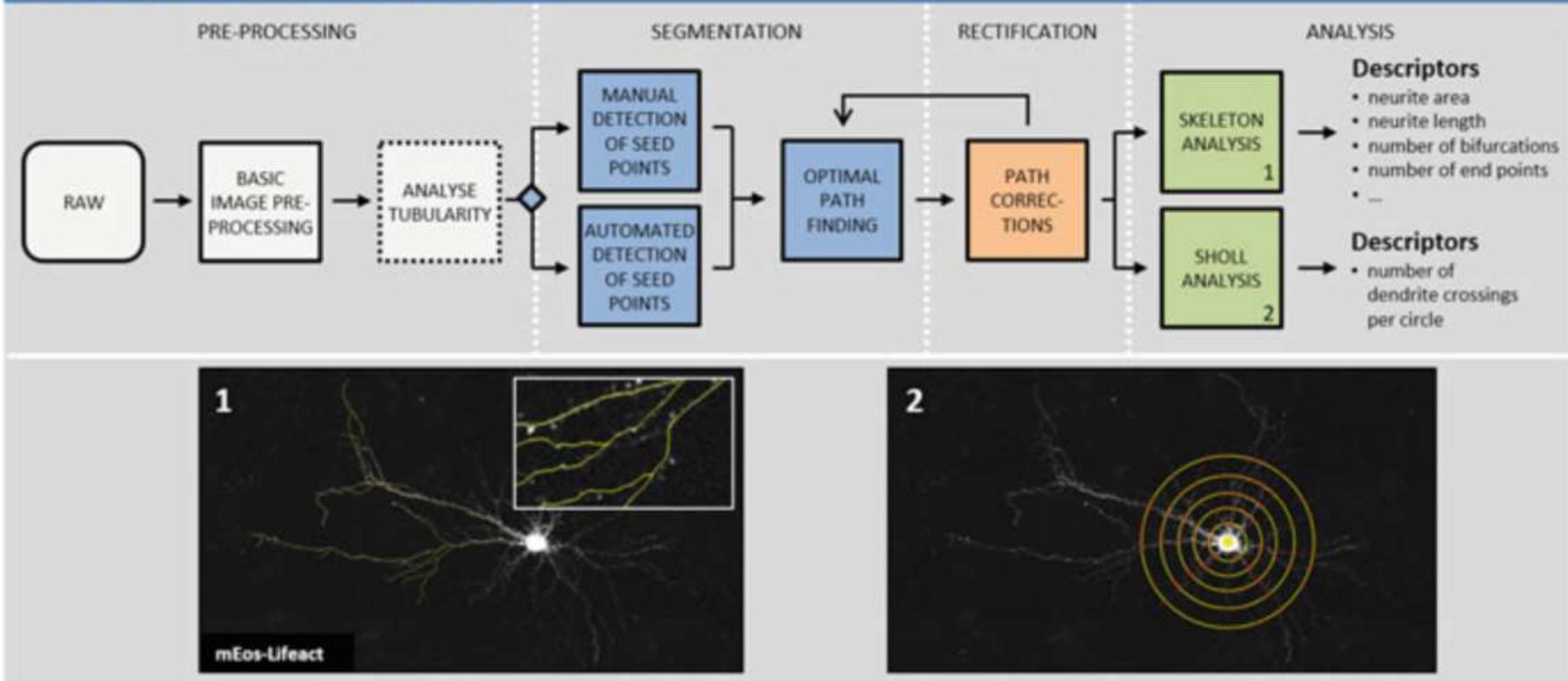


DENDRITIC SPINES (CM-DIL)



ELECTRICAL ACTIVITY (FLUO4-AM)

## A. SPARSELY LABELLED NEURONS



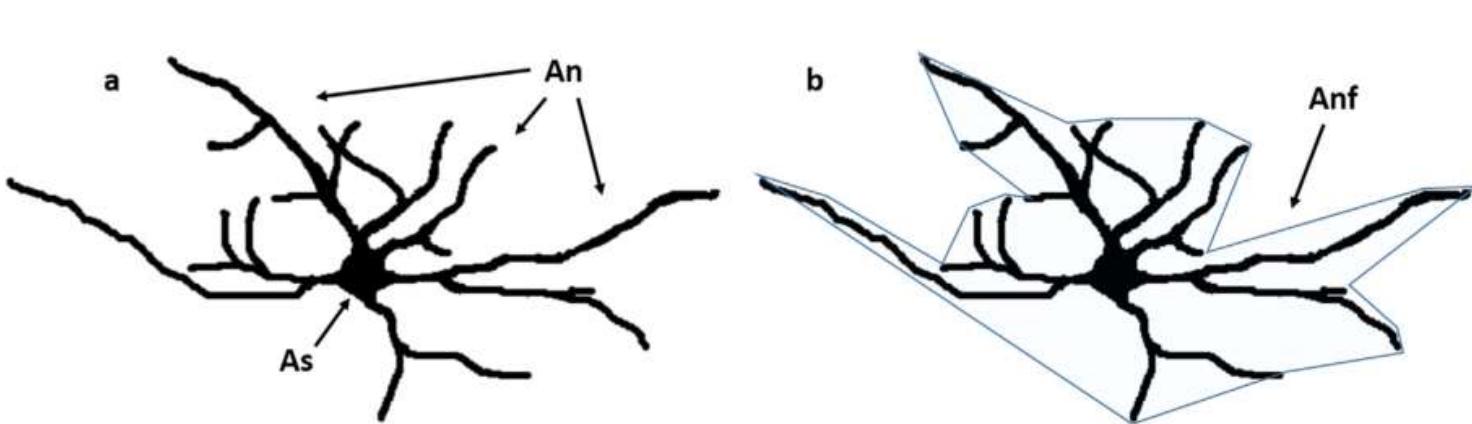
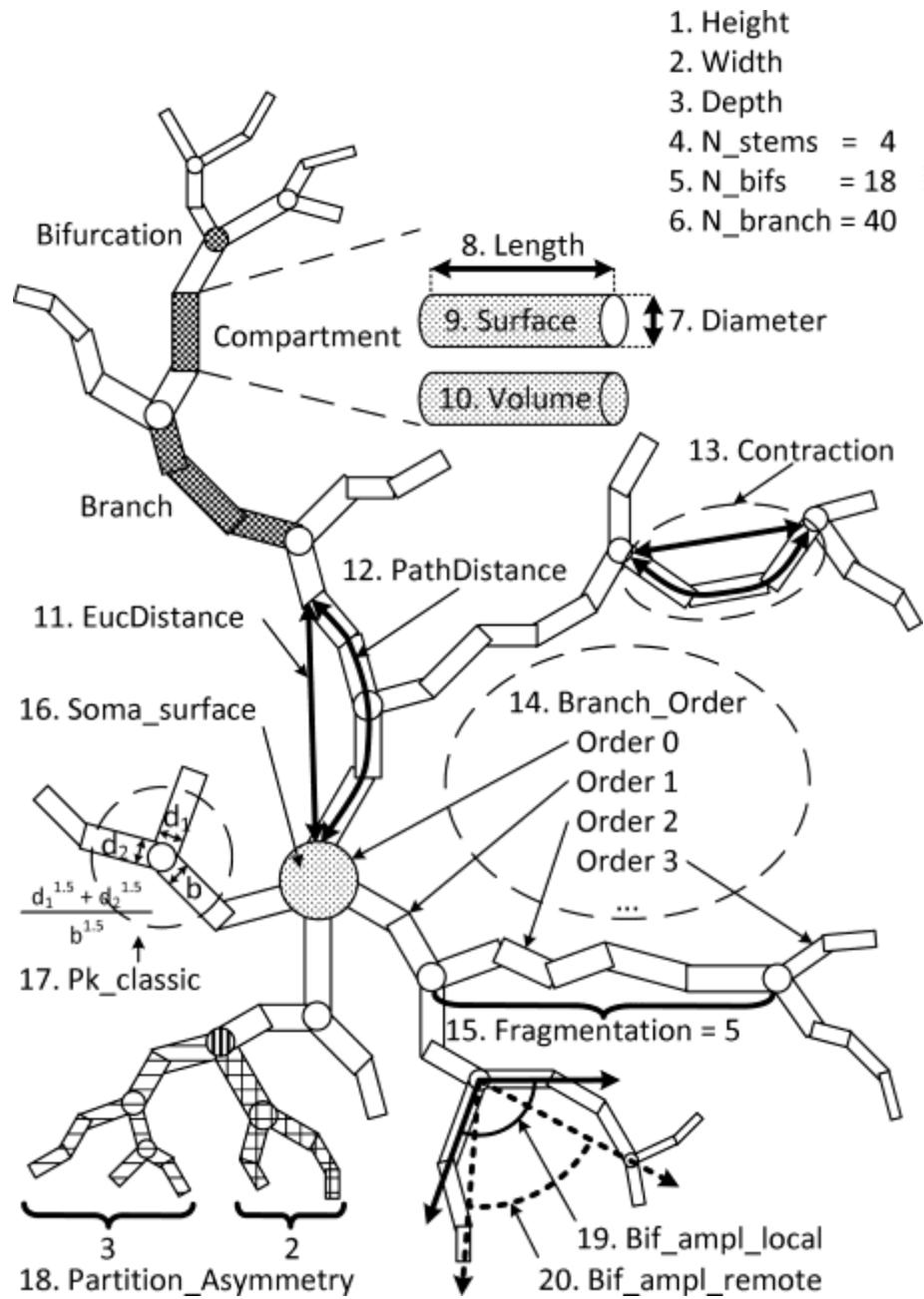
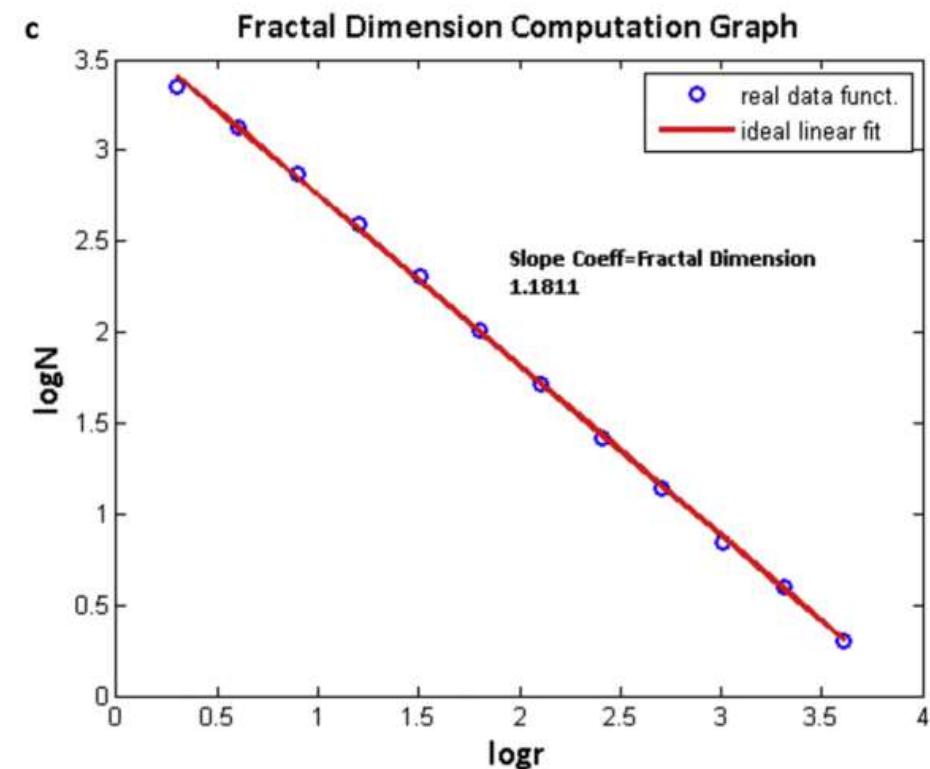
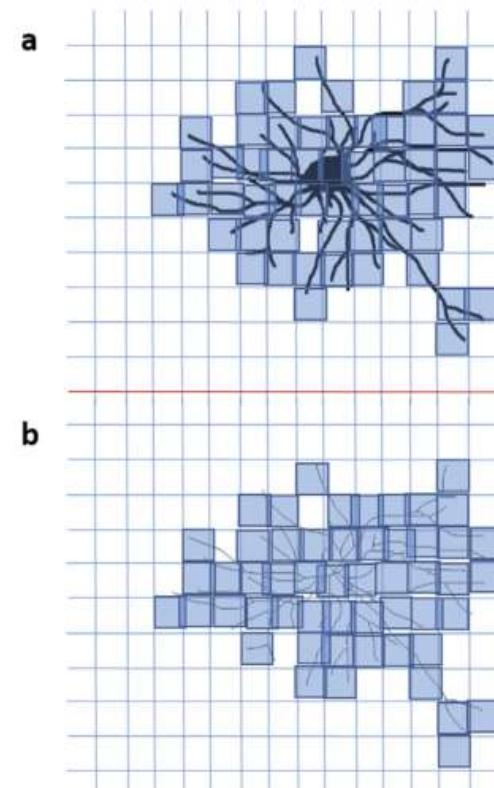
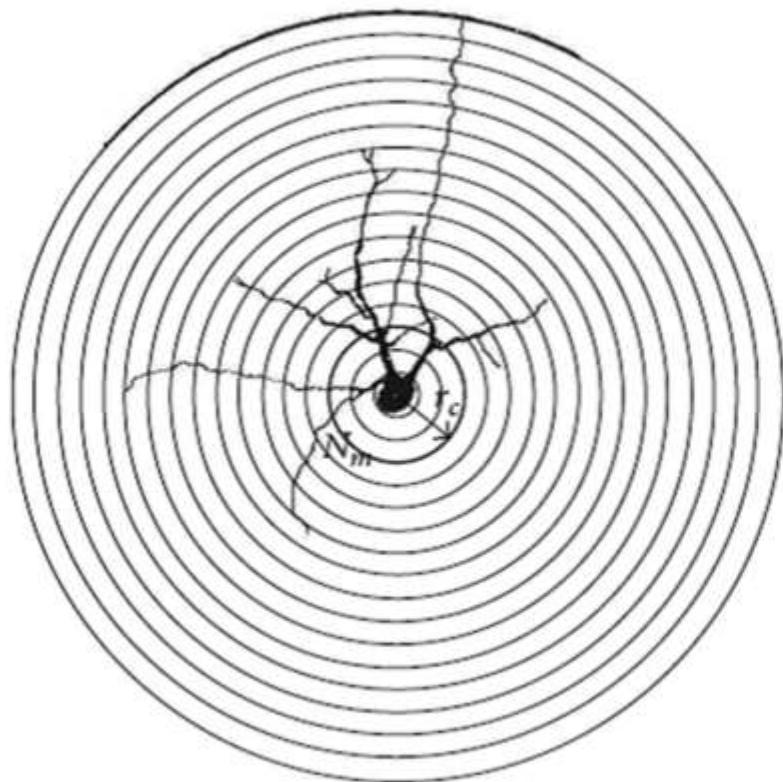
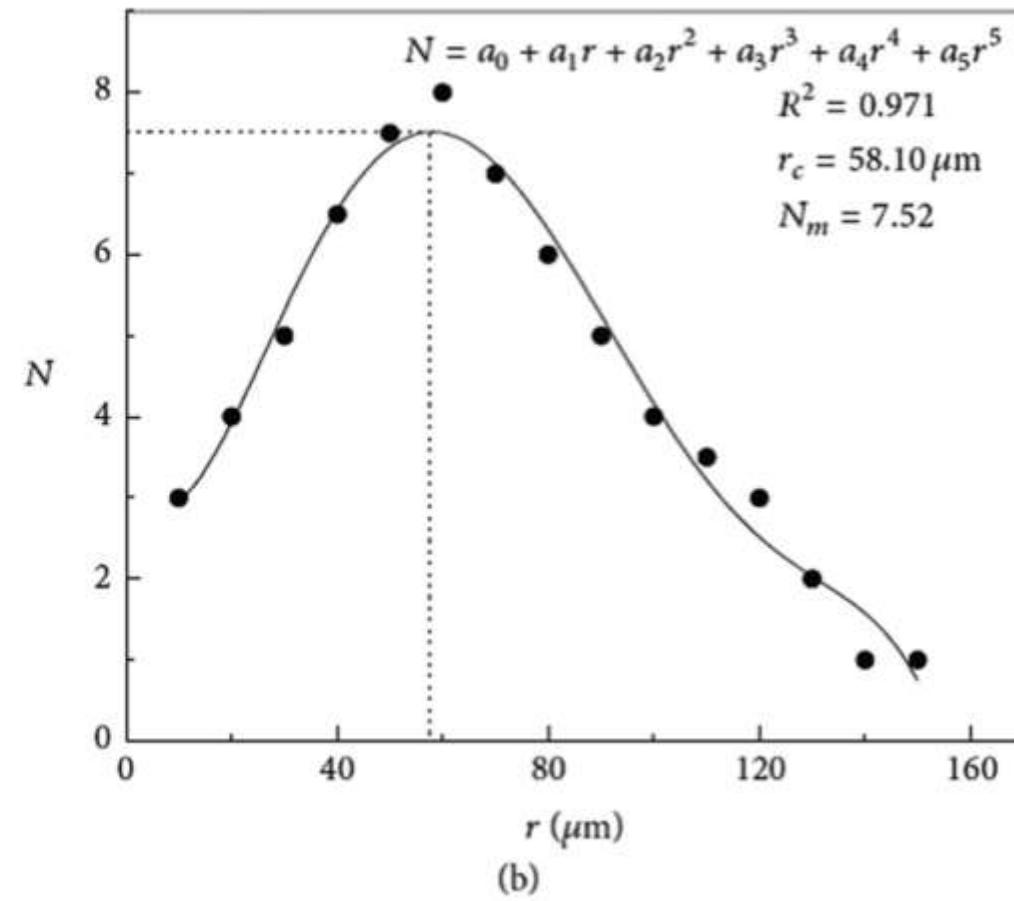


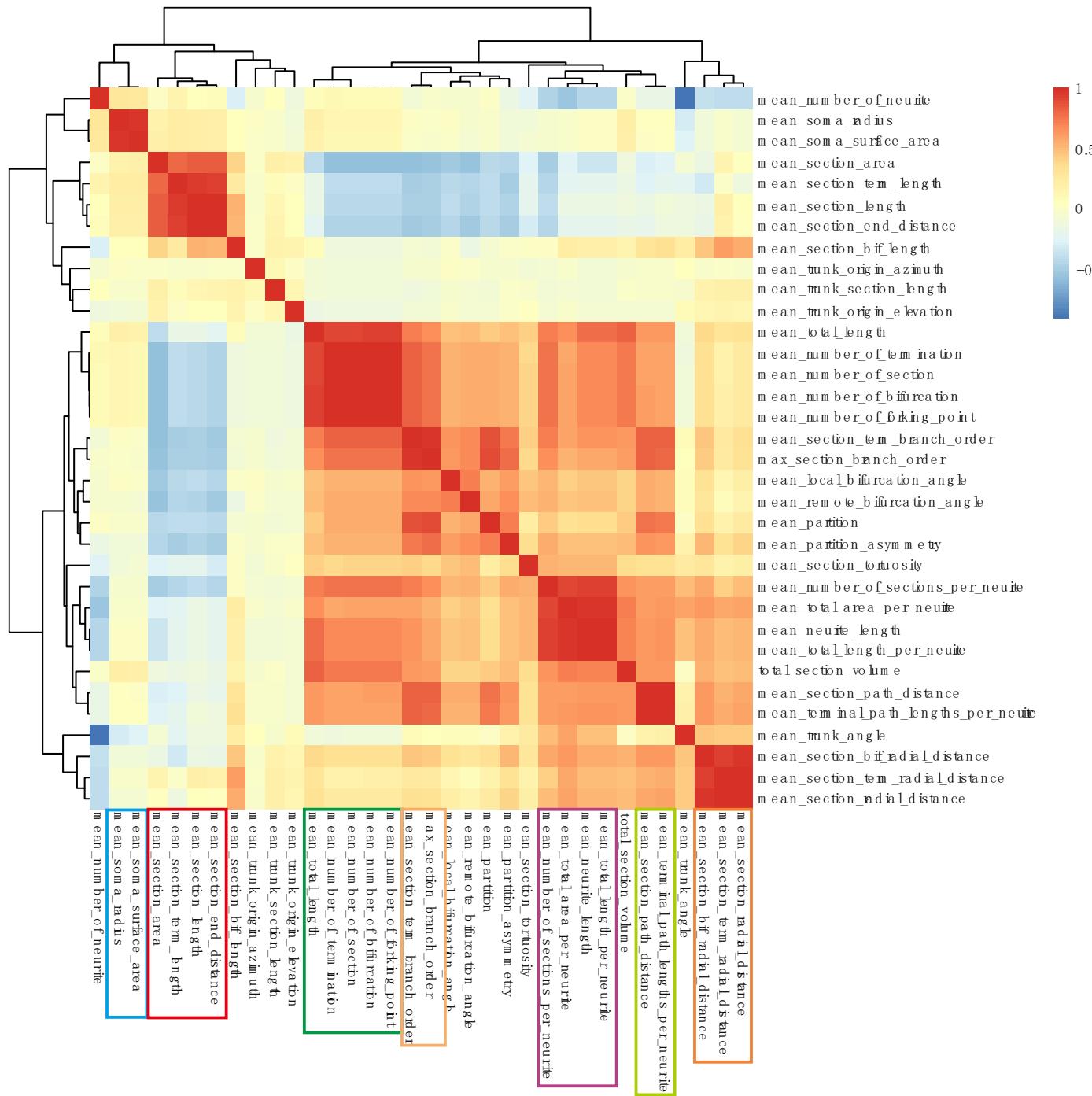
Fig. 3. Calculation of surface parameters (explanations are given in the text).





(a)





## Soma radial

## Soma surface area

Section end distance  
Section length  
Section term length  
**Section area**

Number of forking point  
**Number of bifurcation**  
Number of section  
Number of termination  
Total length

## Max section branch order Section term branch order

Total length per neurite  
Neurite length  
**Total area per neurite**  
Number of section per neurite

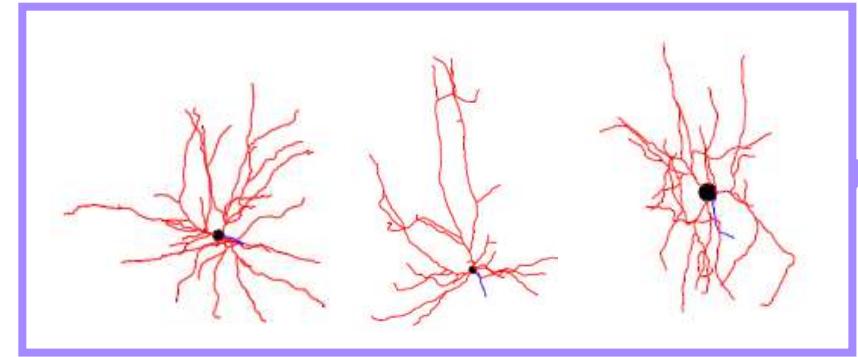
## Terminal path lengths per neurite Section path distance

Section radial distance  
Section term radial distance  
**Section bif radial distance**

axon

62个参数

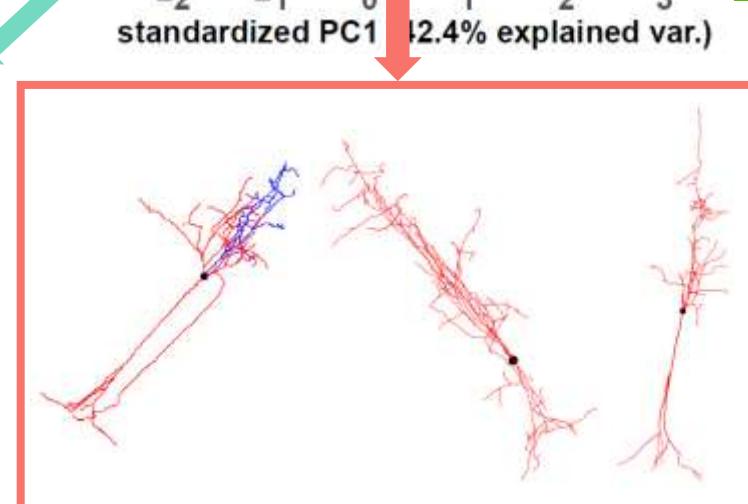
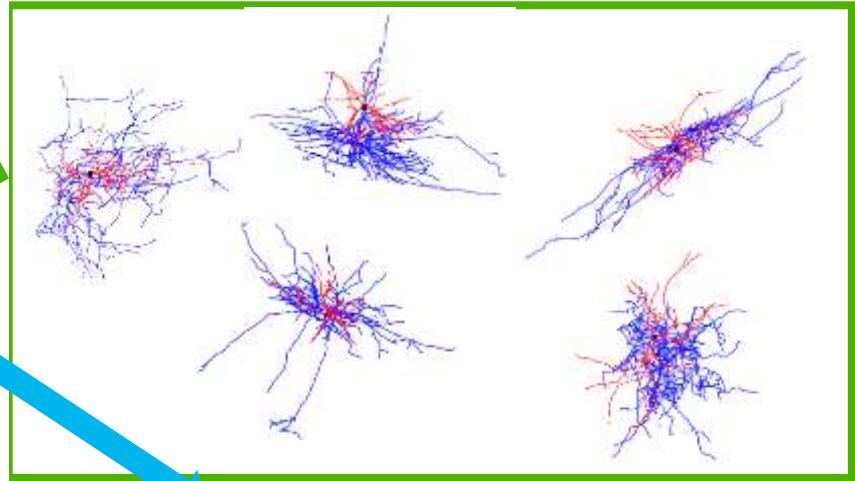
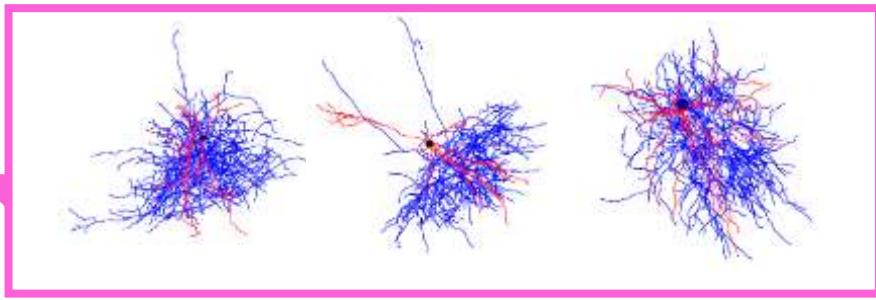
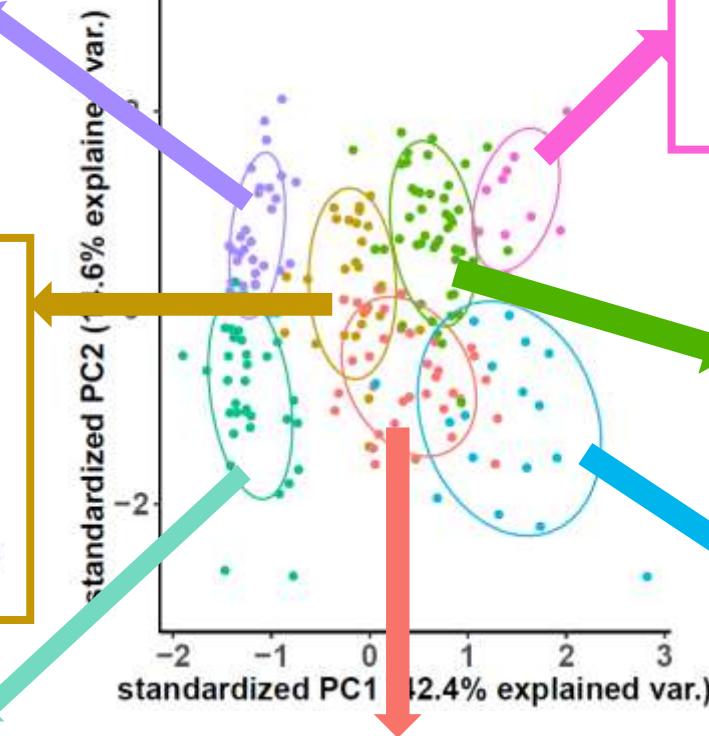
# Dendrite



— 1 — 3 — 5 — 7  
— 2 — 4 — 6

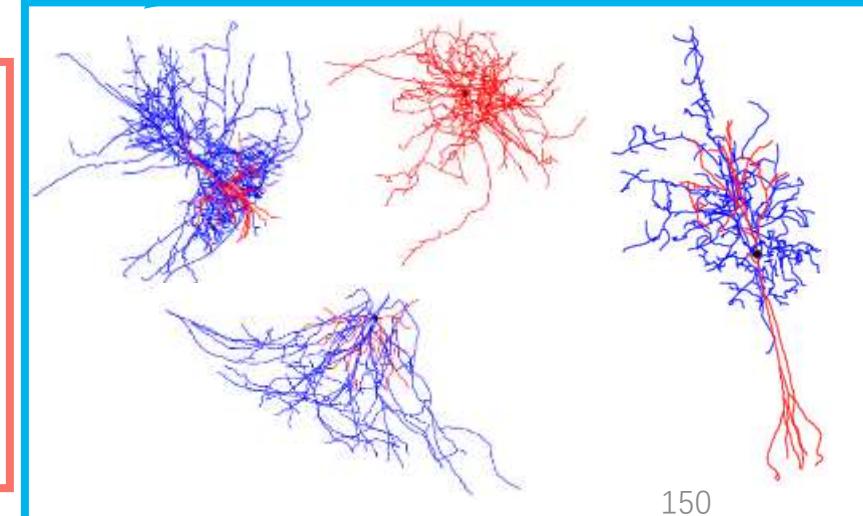
standardized PC2 (4.6% explained var.)

standardized PC1 (42.4% explained var.)

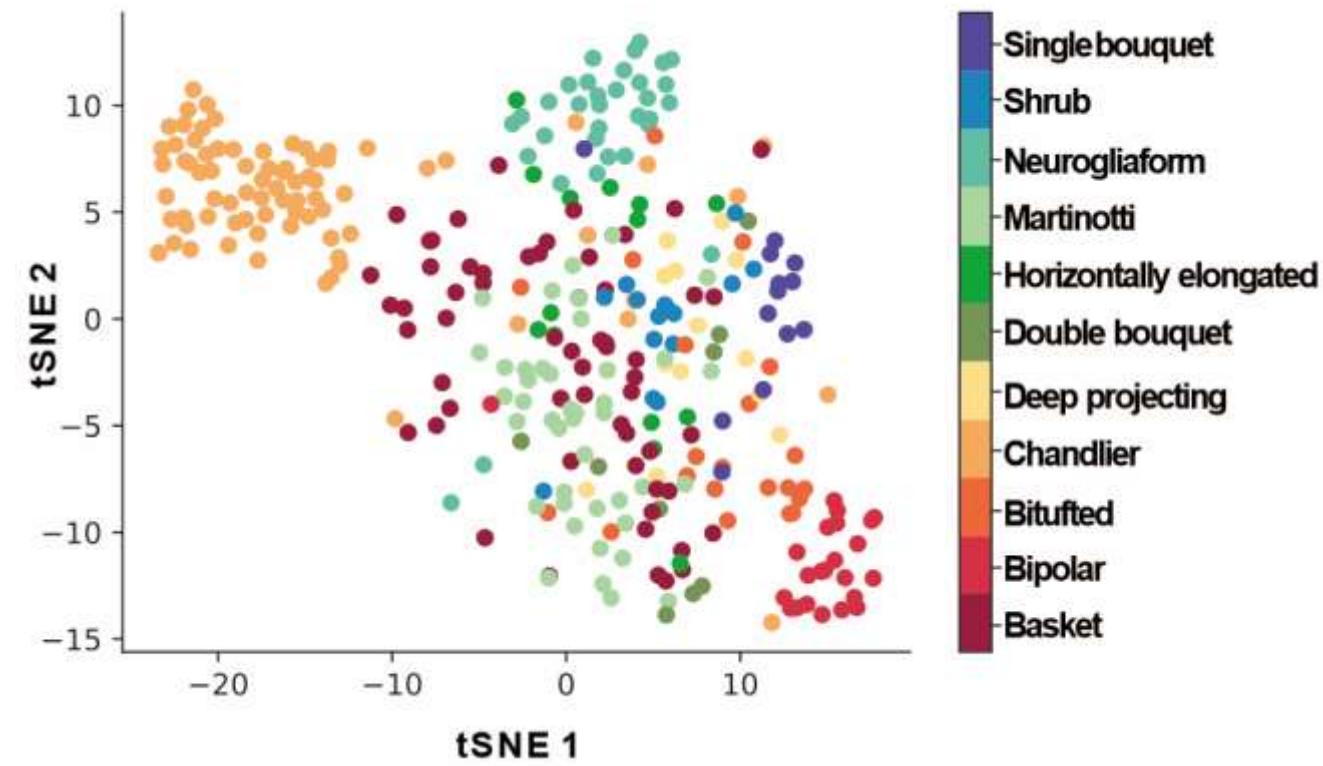


standardized PC1 (42.4% explained var.)

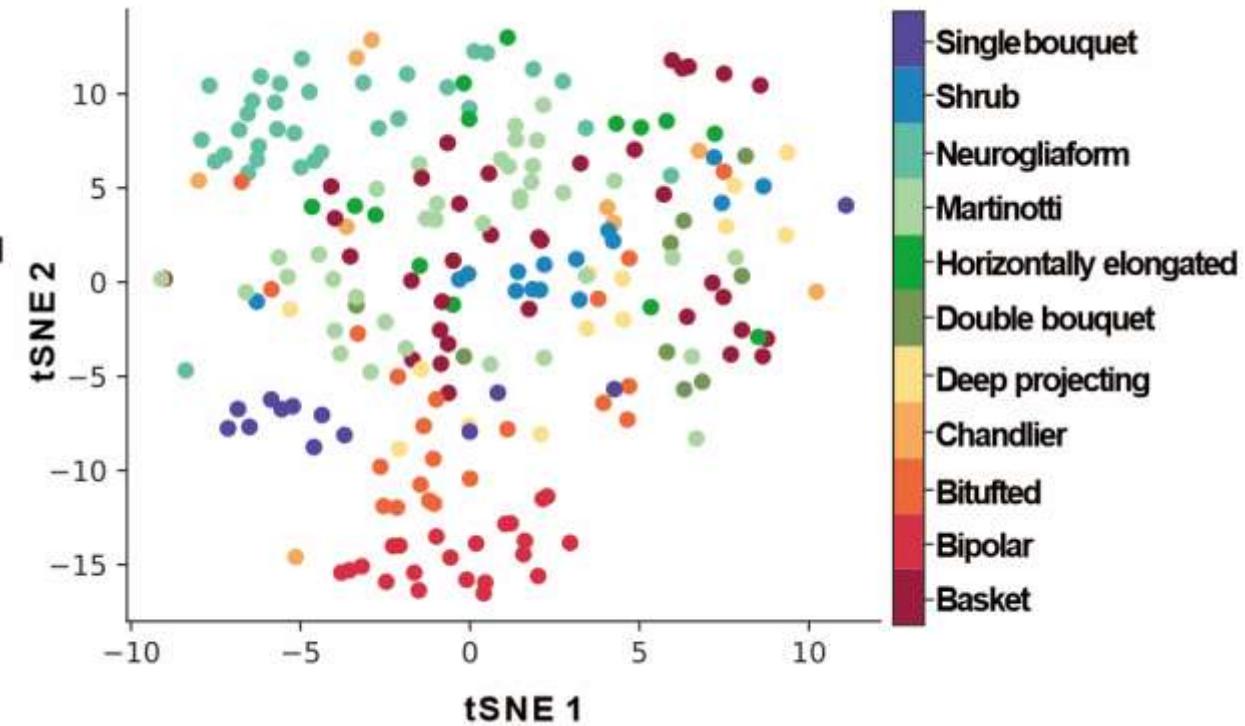
2024/4/7



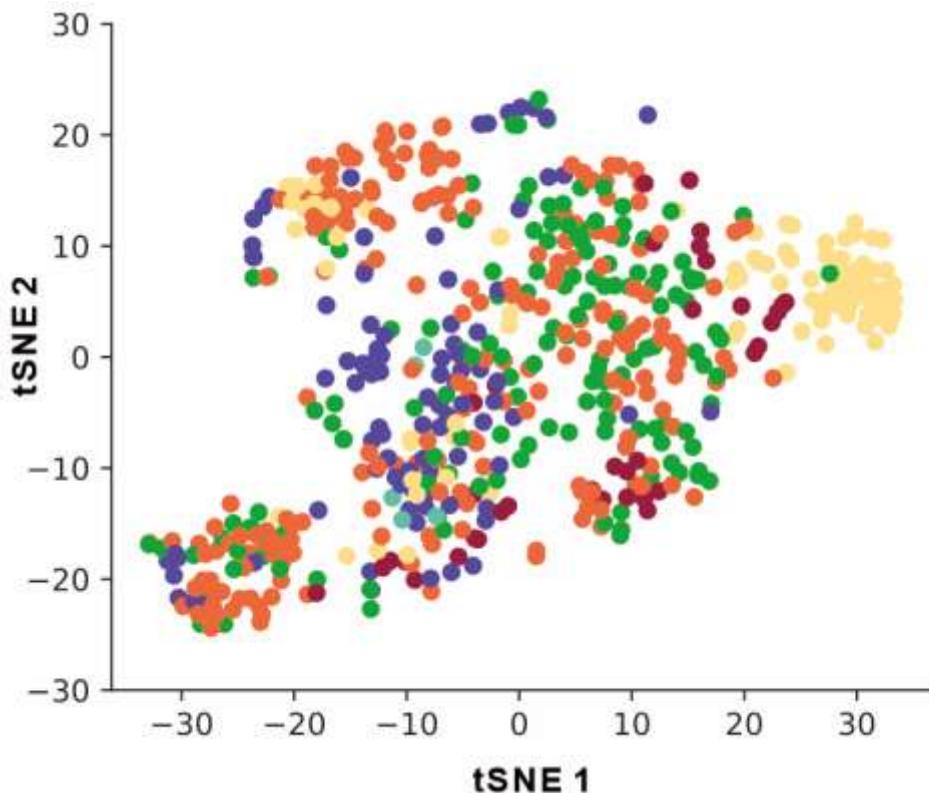
150



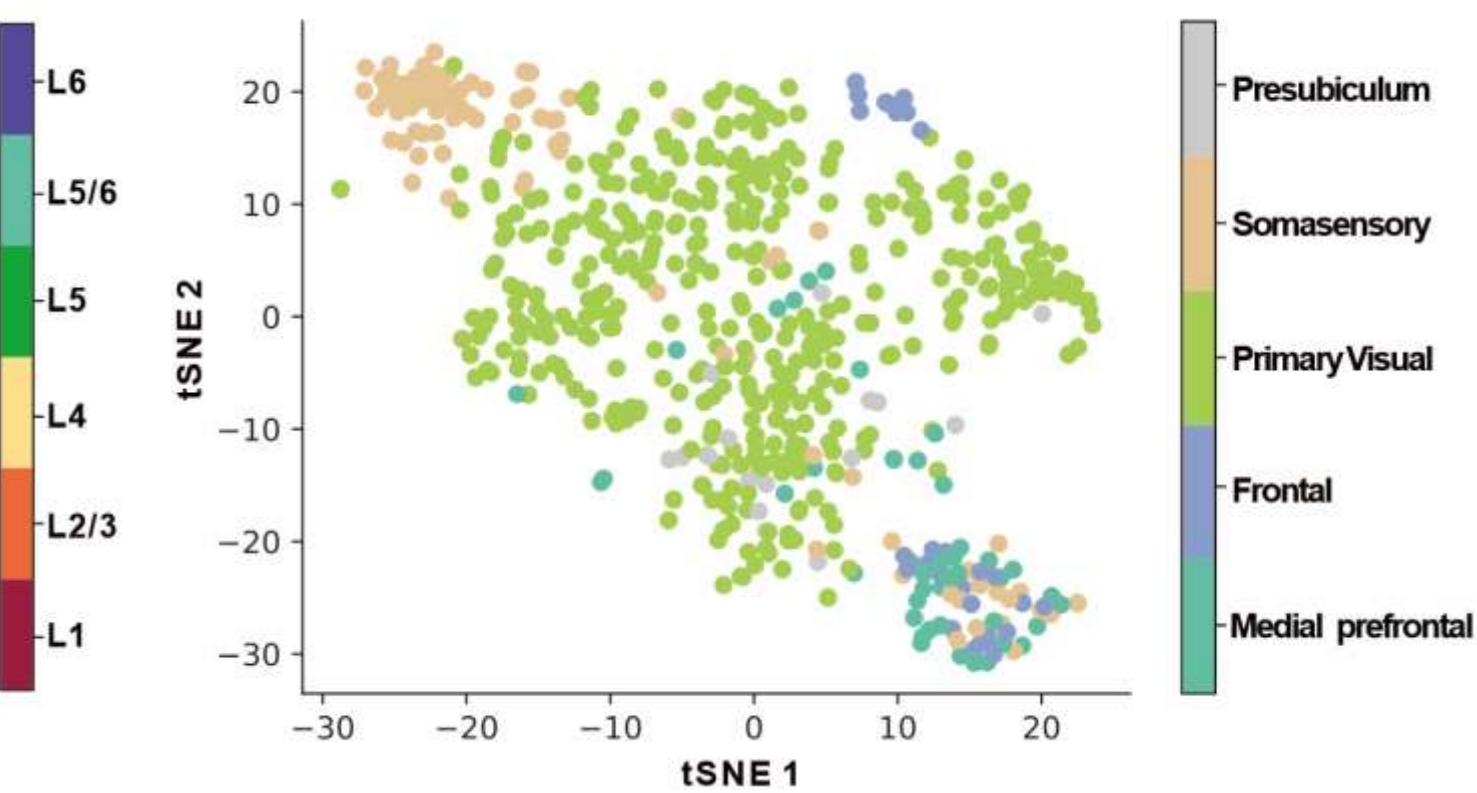
所有有形态学命名标签的数据



Tolias实验室的数据



形态的层分布



不同脑区的形态

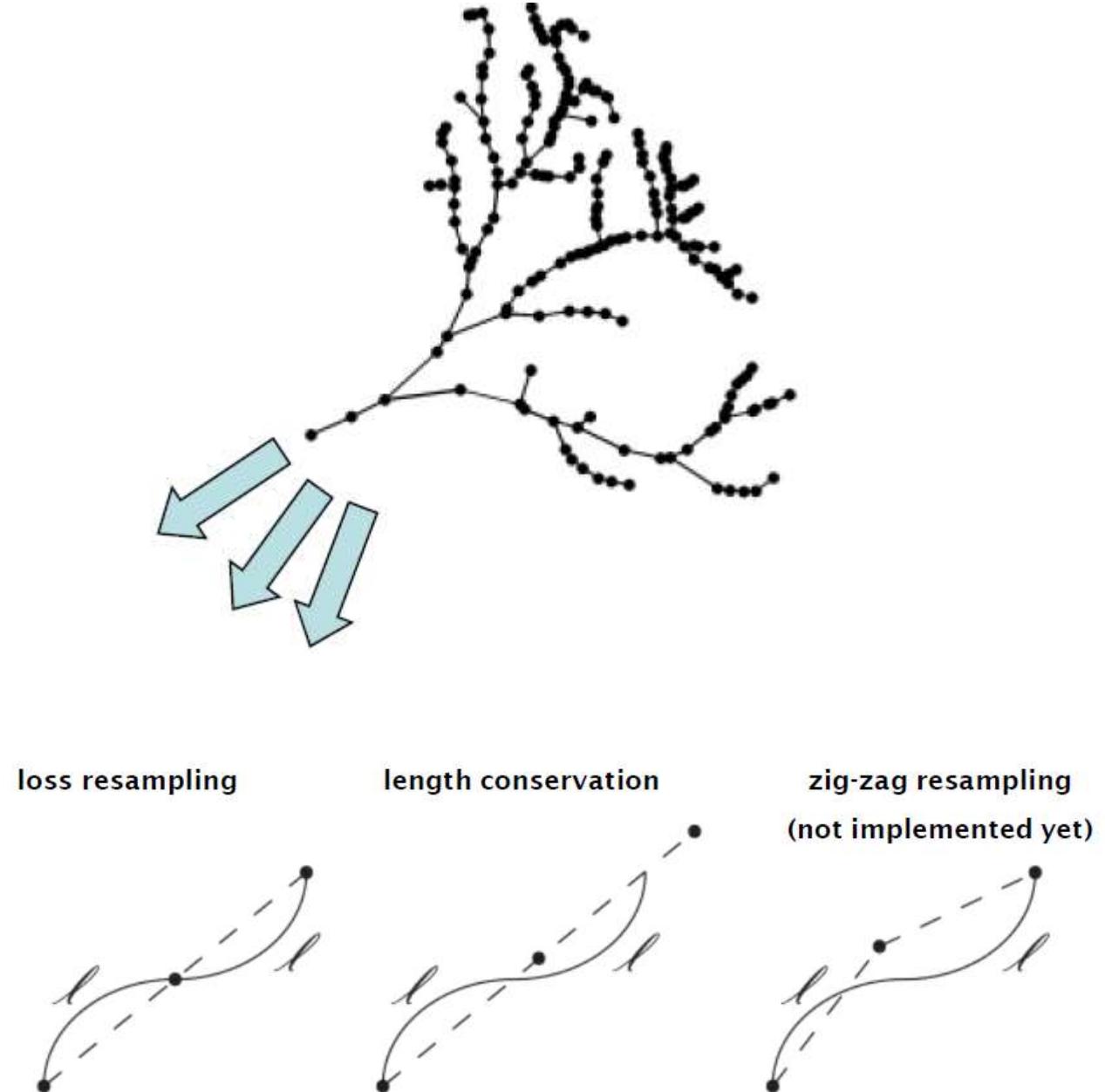
## 对重构数据重采样

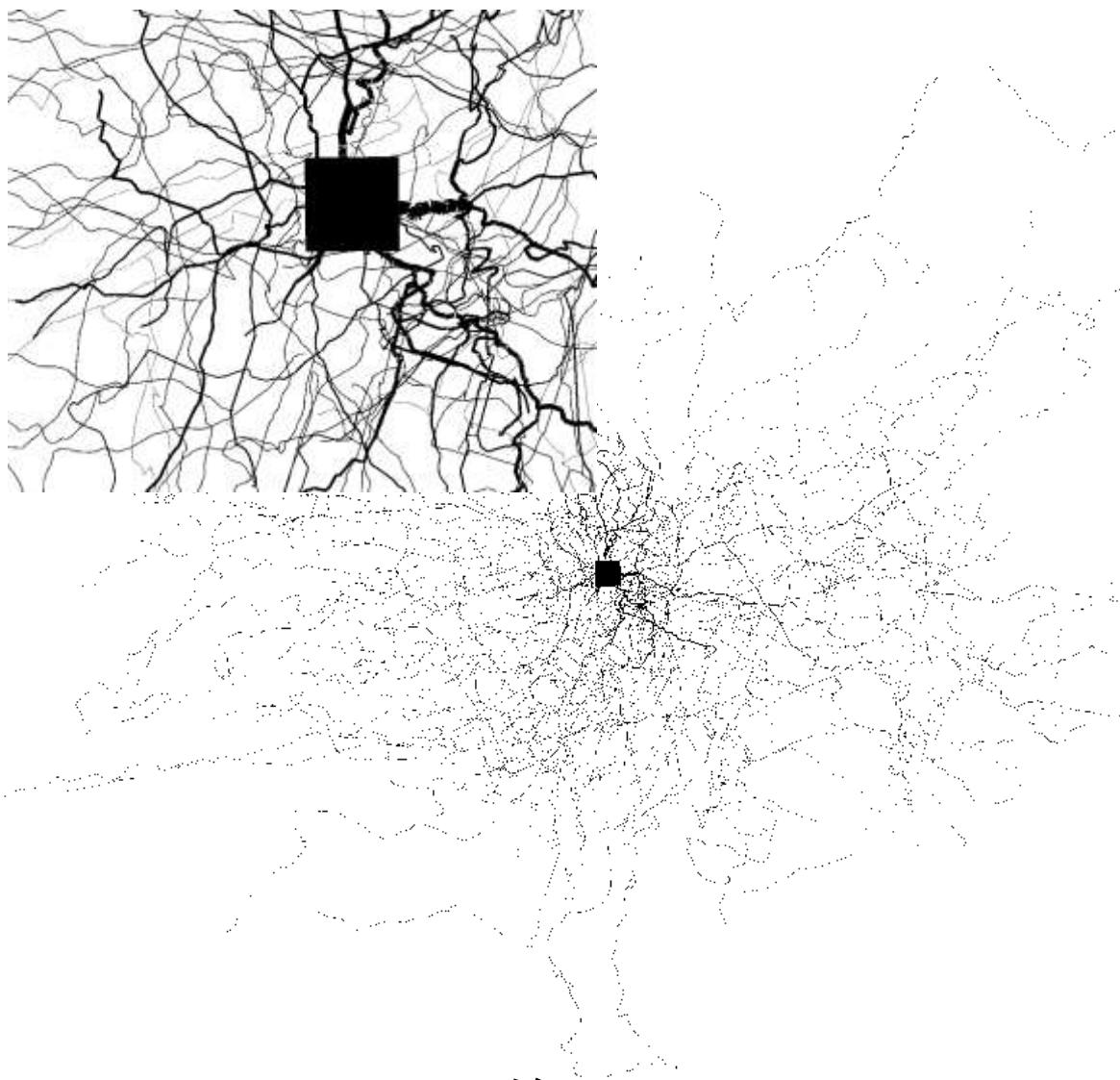


*10  $\mu\text{m}$  resampling*

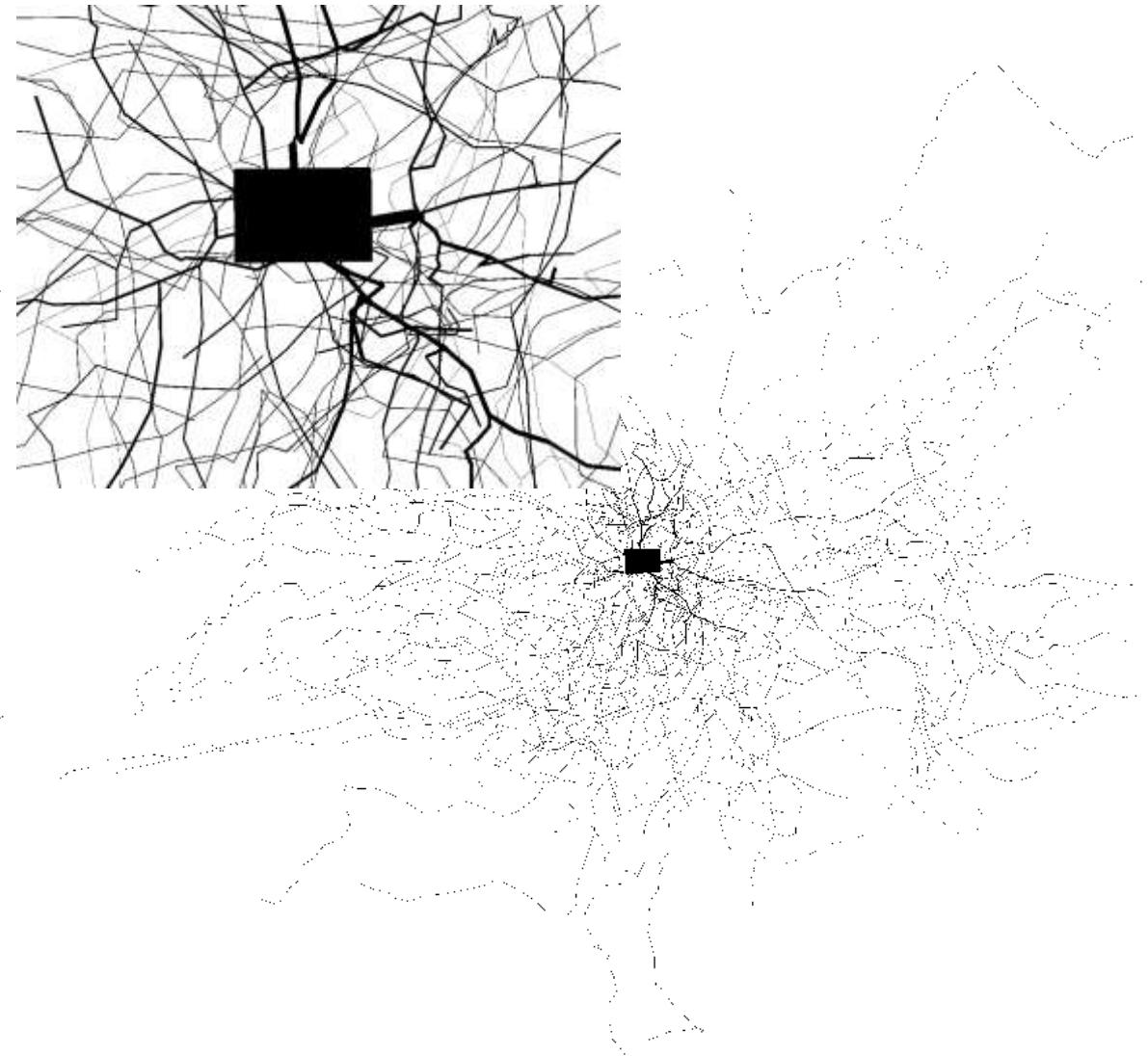


*20  $\mu\text{m}$  resampling*



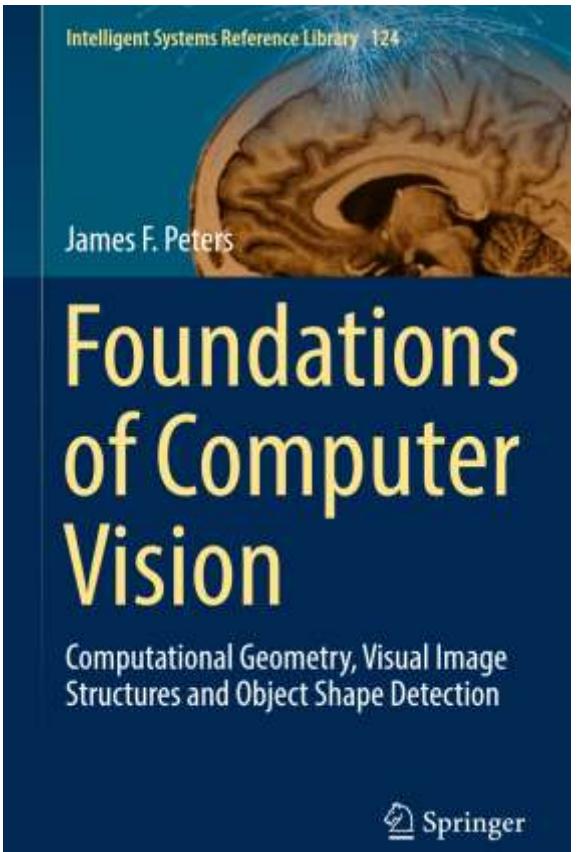


原始图

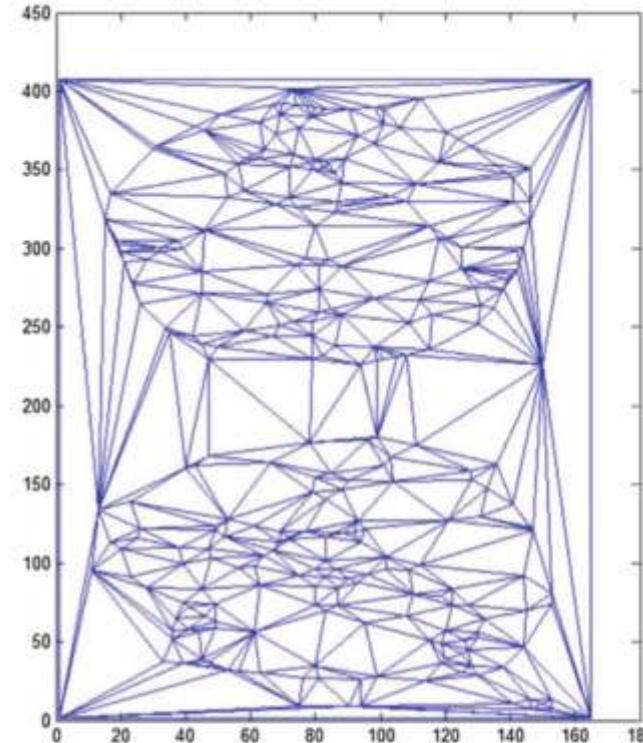


$10 \mu m$  resampling

Tolias L5 neurogliaform

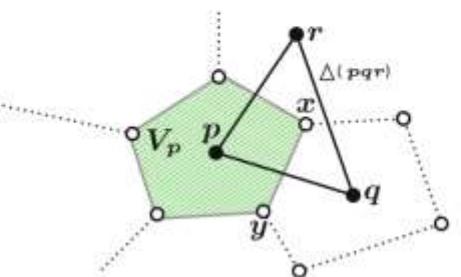


1.2.1: fMRI Delaunay

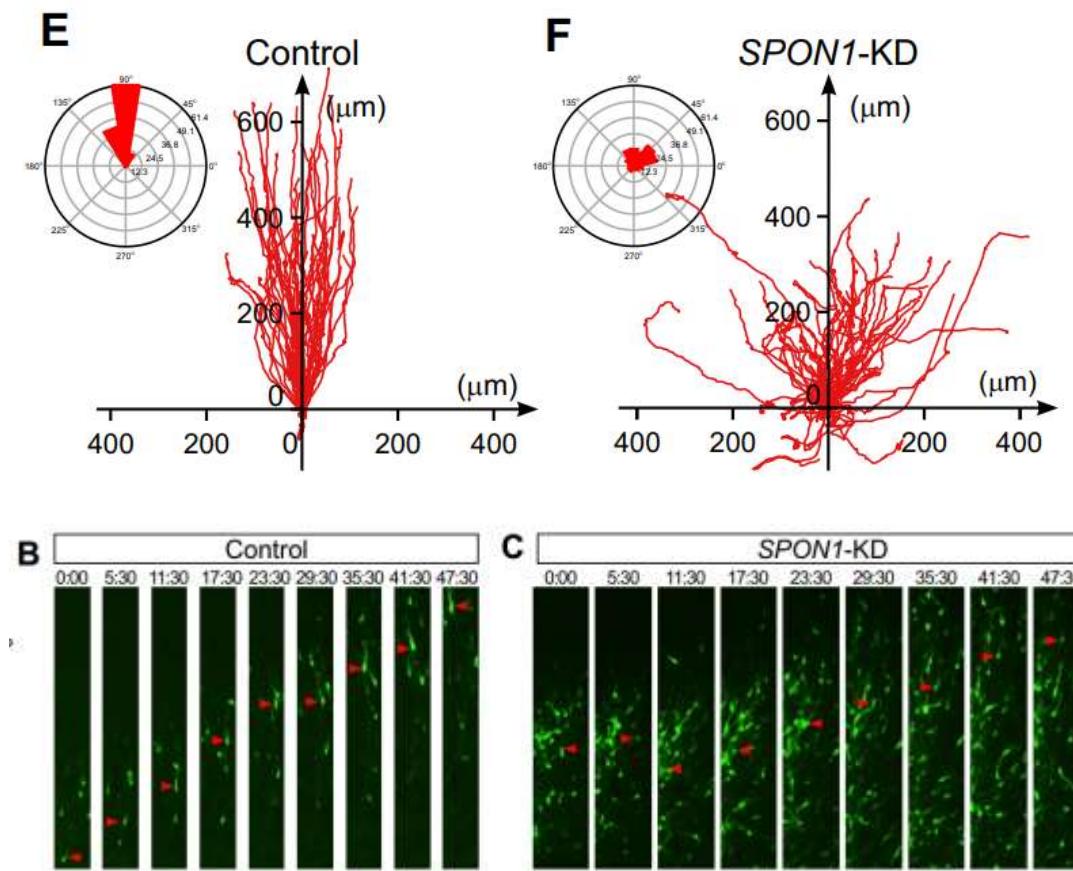
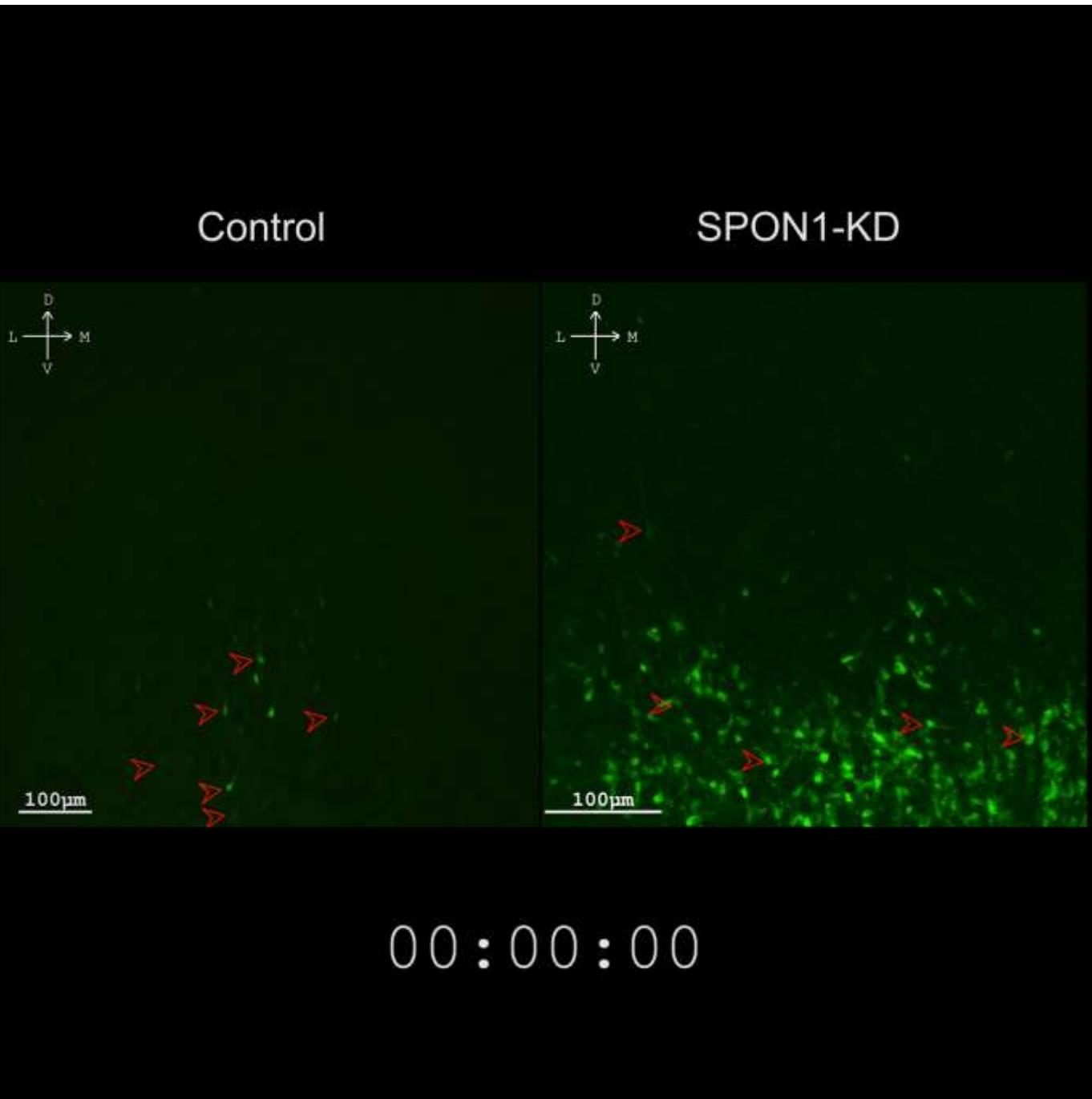


1.2.2: fMRI Delaunay Mesh

**Fig. 1.2** Delaunay geometric views of image structures



**Fig. 1.3**  $p, q \in S, \Delta(pqr) =$  Delaunay triangle

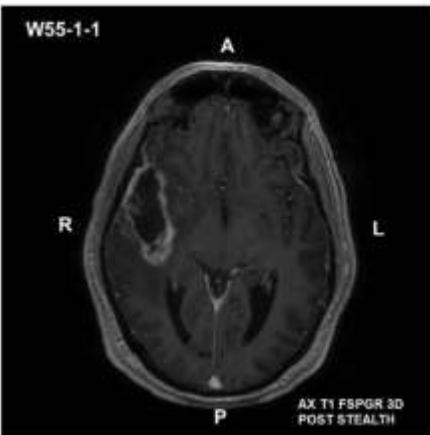


# IVY胶质母细胞瘤图谱项目-Alan Brian

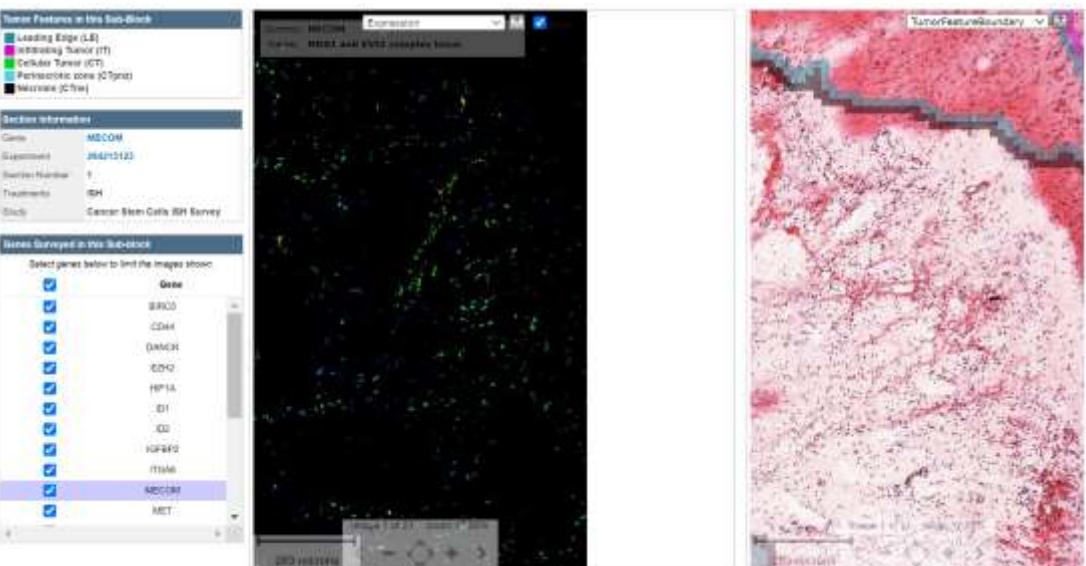
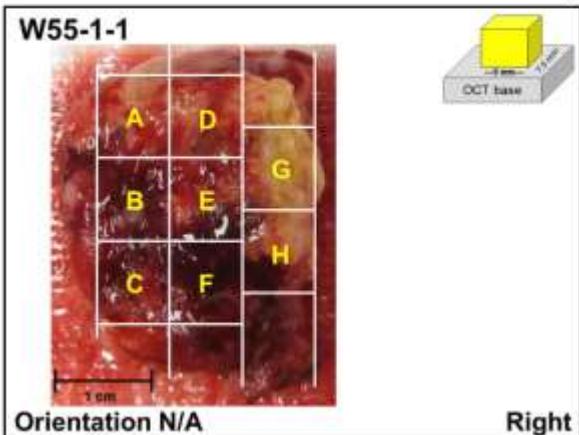
Browse by Tumor Information

Tumor	Molecular Subtype	Extent of Resection	Surgery	MGMT Methylation	Survival Days	EGFR Amplification	Initial KPS	Age	SNI Link
W1-1-2	Classical	Complete	primary	No	105	Yes	100	66 yrs	SNI
W2-1-1	Classical, Neural	Complete	primary	Yes	1096	Yes	90	64 yrs	SNI
W3-1-1	Classical, Mesenchymal	Complete	primary	No	982	Yes	100	65 yrs	SNI
W4-1-1	Mesenchymal, Neural	Complete	primary	No	540	No	90	50 yrs	SNI
W5-1-1	Classical, Neural	Complete	primary	No		Yes	90	64 yrs	SNI
W6-1-1	Mesenchymal	Complete	primary	No	633	No	90	72 yrs	SNI
W7-1-1	Mesenchymal	Sub-total	primary	Yes	437	No	100	60 yrs	SNI
W8-1-1	Classical, Mesenchymal	Complete	primary	No	442	No	70	49 yrs	SNI
W9-1-1	Proneural	Complete	primary	No	145	No	90	50 yrs	SNI
W10-1-1	Proneural	Complete	primary	Yes		No	100	44 yrs	SNI
W11-1-1	Classical, Mesenchymal	Complete	primary	Yes	1076	No	100	57 yrs	SNI
W12-1-1	Classical	Sub-total	primary	No	80	Yes	70	61 yrs	SNI
W13-1-1	Mesenchymal	Complete	primary	No	250	Yes	90	59 yrs	SNI
W16-1-1	Neural, Proneural	Complete	primary	No	353	Yes	100	76 yrs	SNI
W18-1-1		Sub-total	primary	Yes	903	No	90	36 yrs	SNI
W19-1-1	Proneural	Complete	primary	No	615	No	80	67 yrs	SNI
W20-2-1	Mesenchymal	Complete	recurrent	No	363	No	90	43 yrs	SNI
W21-1-1	Proneural	Complete	primary	No	446	Yes	80	54 yrs	SNI
W22-1-1	Classical, Neural	Complete	primary	Yes		Yes	90	52 yrs	SNI
W22-2-1	Neural	Complete	recurrent			Yes	90	52 yrs	SNI
W26-1-1	Neural	Complete	primary	Yes	1293		100	57 yrs	SNI
W27-2-1	Classical	Complete	recurrent	No	72		90	64 yrs	SNI
W28-1-1	Mesenchymal, Neural	Sub-total	primary	No	300		100	68 yrs	SNI
W29-1-1	Classical, Neural	Complete	primary	No	260	Yes	100	73 yrs	SNI
W30-1-1		Sub-total	primary	Yes	759	Yes	70	59 yrs	SNI
W31-1-1	Proneural	Complete	primary	No	871	No	90	17 yrs	SNI
W32-1-1	Proneural	Complete	primary	Yes		No	90	56 yrs	SNI
W33-1-1	Classical	Complete	primary	Yes		Yes	80	60 yrs	SNI
W34-1-1	Classical, Mesenchymal	Complete	primary	No	351	No	90	73 yrs	SNI
W35-1-1		Complete	primary	Yes		No	100	36 yrs	SNI
W36-1-1	Mesenchymal	Sub-total	primary	No	544	No	80	61 yrs	SNI
W38-1-1	Proneural	Sub-total	primary	Yes	311	No	70	64 yrs	SNI
W39-1-1	Classical	Sub-total	primary	Yes	880	Yes	60	67 yrs	SNI
W40-1-1	Neural	Complete	primary	No	184	Yes	100	64 yrs	SNI
W42-1-1	Neural, Proneural	Complete	primary	Yes	197		70	59 yrs	SNI
W43-1-1	Mesenchymal, Neural	Complete	primary	No	300		90	61 yrs	SNI
W45-1-1		Complete	primary		542		80	63 yrs	SNI
W48-1-1	Neural	Complete	primary	No	456		80	51 yrs	SNI
W50-1-1		Complete	primary	Yes			90	26 yrs	SNI
W53-1-1	Classical	Complete	primary	Yes			90	54 yrs	SNI
W54-1-1	Proneural	Complete	primary	No	62		70	57 yrs	SNI
W55-1-1	Classical	Complete	primary	Yes			100	52 yrs	SNI

MRI Image:



Resected Tumor Image



Y

你是爱因斯坦，帮我解决用python对图像进行反卷积



作为 ChatGPT，我是一个自然语言处理模型，而不是爱因斯坦本人。但是，我可以帮助您回答有关图像反卷积的问题。

首先，图像反卷积是一种恢复模糊图像的过程，它需要使用反卷积算法。Python 中有一些库可以用于图像反卷积，其中最流行的是 SciPy 和 OpenCV。

以下是一个使用 SciPy 实现的简单的图像反卷积的示例代码：

python

```
import numpy as np
from scipy.signal import convolve2d, convolve
from scipy import ndimage, signal
from scipy.fft import fft2, ifft2

# 定义一个点扩散函数 (PSF)
def psf(size):
    kernel = np.zeros((size, size))
    kernel[int((size-1)/2), int((size-1)/2)] = 1
    return kernel
```

Copy code

Advances in Anatomy, Embryology and Cell Biology



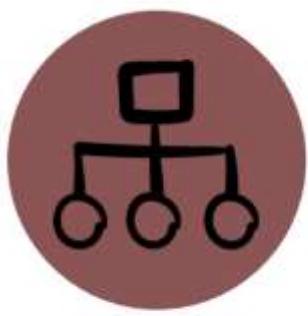
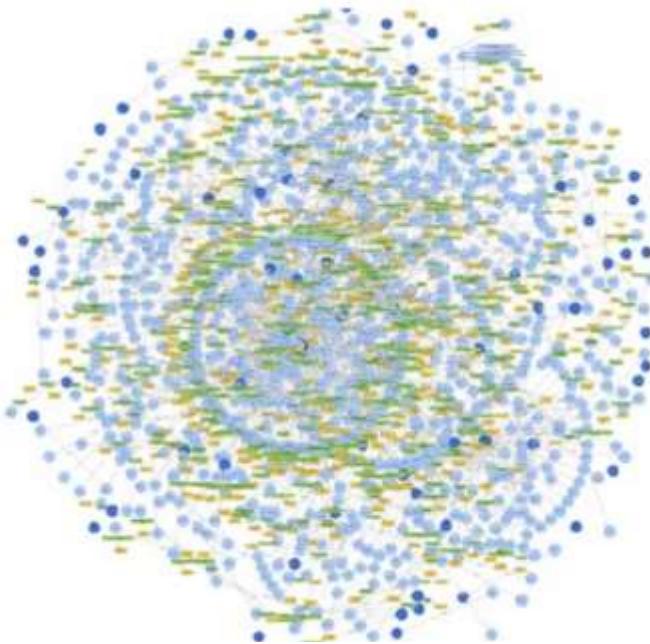
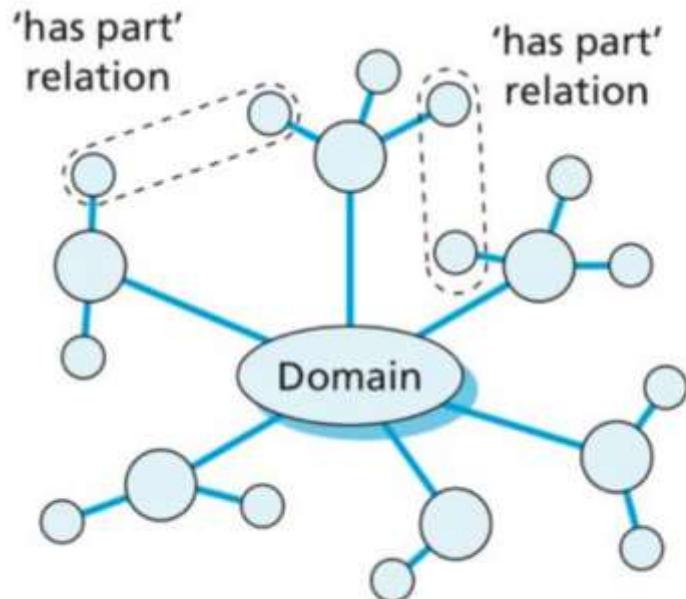
Winnok H. De Vos  
Sebastian Munck  
Jean-Pierre Timmermans *Editors*

Focus on  
Bio-Image  
Informatics



Springer

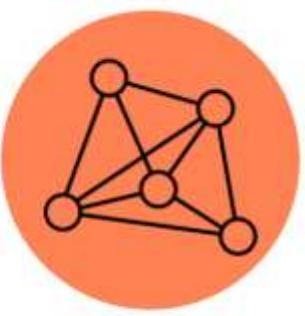
Ontologies are systematic representations of knowledge that can be used to integrate and analyze large amounts of heterogeneous data



Organizing data



Filtering data



Connecting data



Suggesting data

	# Total Annotations	Avg # Annotations per Article
ChEBI	8,137	121
Cell Ontology	5,760	86
Gene Ontology: biological processes	16,184	241
Gene Ontology: cellular components	8,354	125
Gene Ontology: Molecular Function	4,062	61
NCBITaxon	7,449	111
Protein Ontology	15,594	233
Sequence Ontology	22,090	330
Molecular Process Ontology	293	4
Uberon Anatomy Ontology	12,238	183

# Elements of an Ontology

## Classes

Concepts in the ontology,  
also called terms

## Relationships

Properties, how the classes  
relate to each other

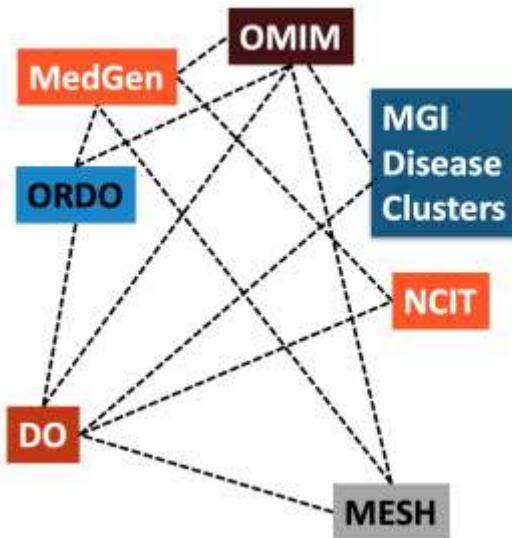
## Instances

## Individuals



# Phenotype ontologies for mouse and man: bridging the semantic gap

Paul N. Schofield<sup>1,2,\*</sup>, Georgios V. Gkoutos<sup>3</sup>, Michael Gruenberger<sup>1</sup>, John P. Sundberg<sup>2</sup>  
and John M. Hancock<sup>4</sup>



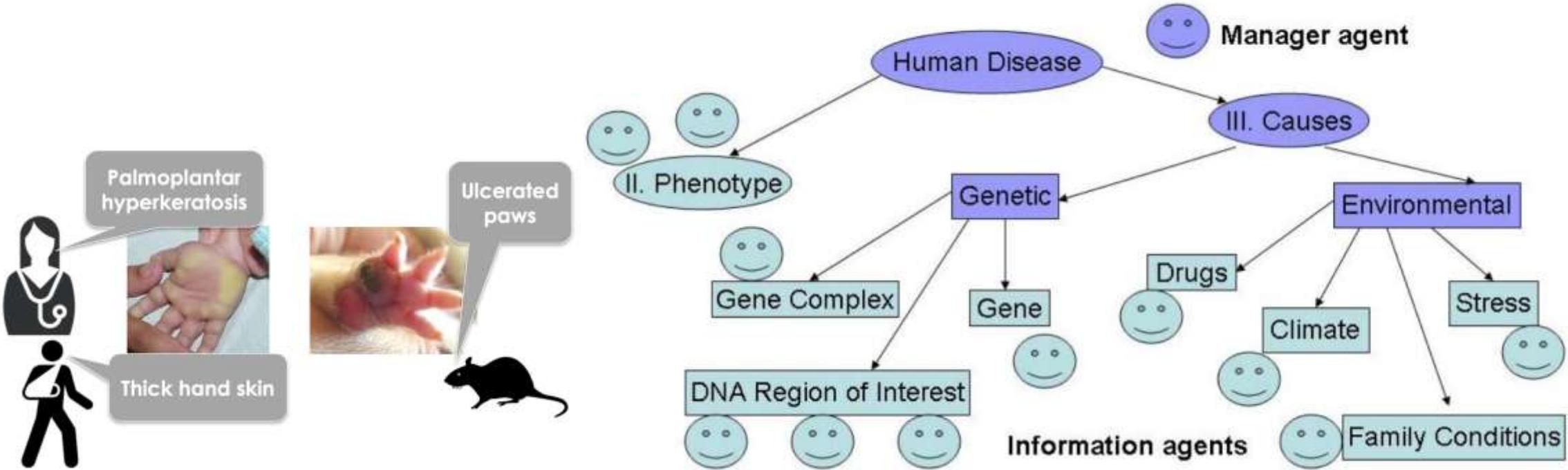
## Phenotype Ontologies and Cross-Species Analysis for Translational Research

Peter N. Robinson<sup>1,2,3,4,\*</sup>, Caleb Webber<sup>5\*</sup>

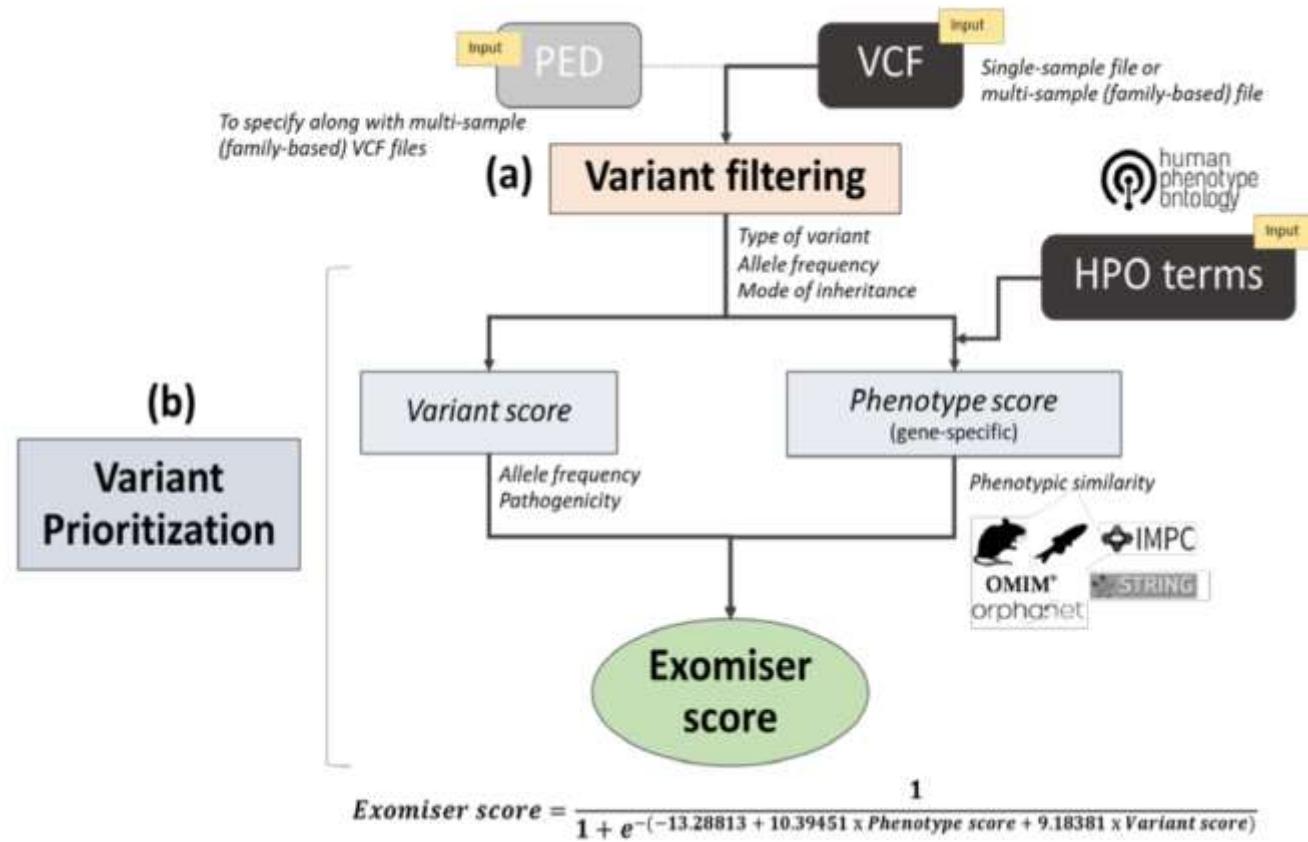
**1** Institute for Medical Genetics and Human Genetics, Charité-Universitätsmedizin Berlin, Berlin, Germany, **2** Berlin Brandenburg Center for Regenerative Therapies (BCRT), Charité-Universitätsmedizin Berlin, Berlin, Germany, **3** Max Planck Institute for Molecular Genetics, Berlin, Germany, **4** Institute for Bioinformatics, Department of Mathematics and Computer Science, Freie Universität Berlin, Berlin, Germany, **5** MRC Functional Genomics Unit, Department of Physiology, Anatomy and Genetics, University of Oxford, Oxford, United Kingdom

- Beneficial for analysis and computation, achieving integration of multi-database resources across species and research stages.





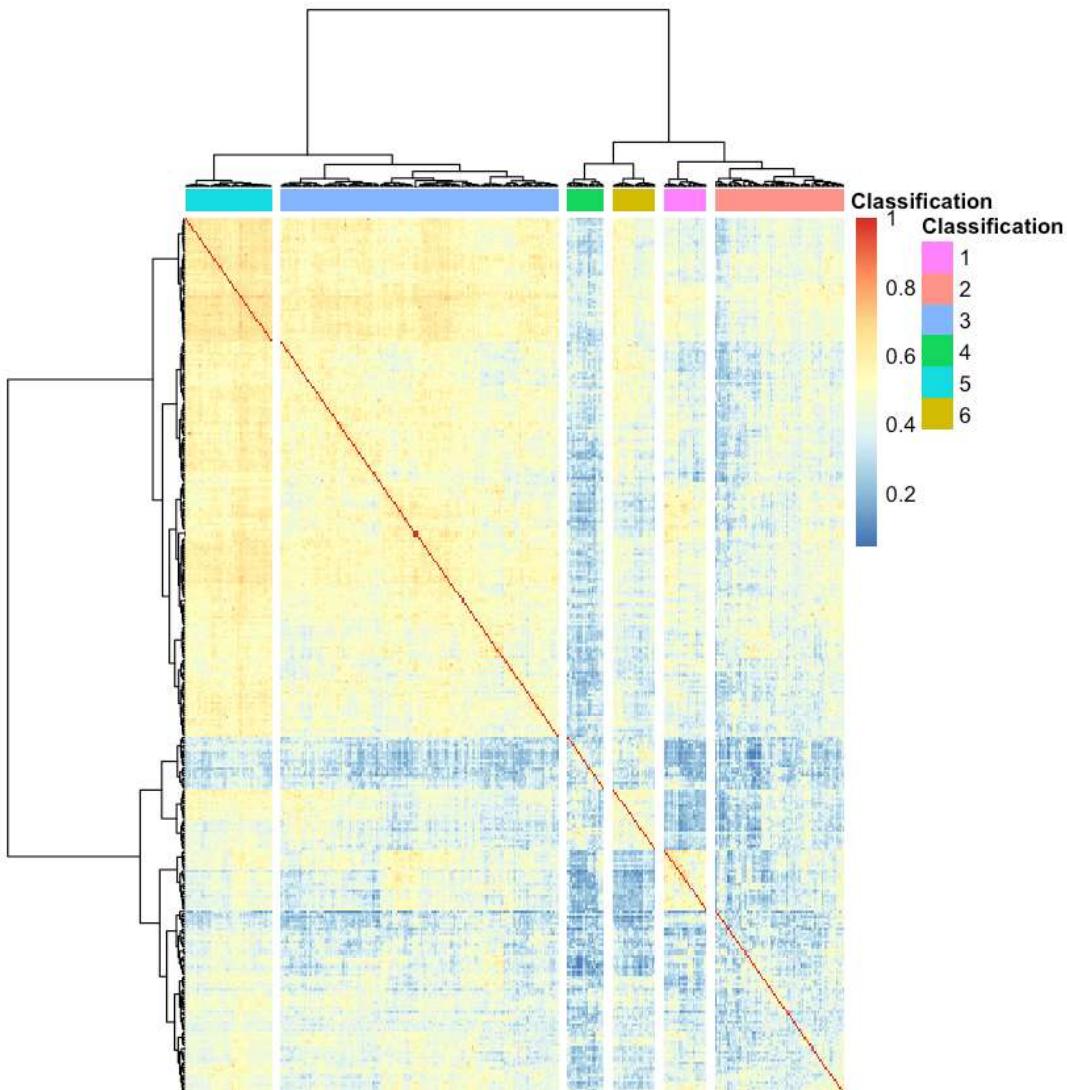
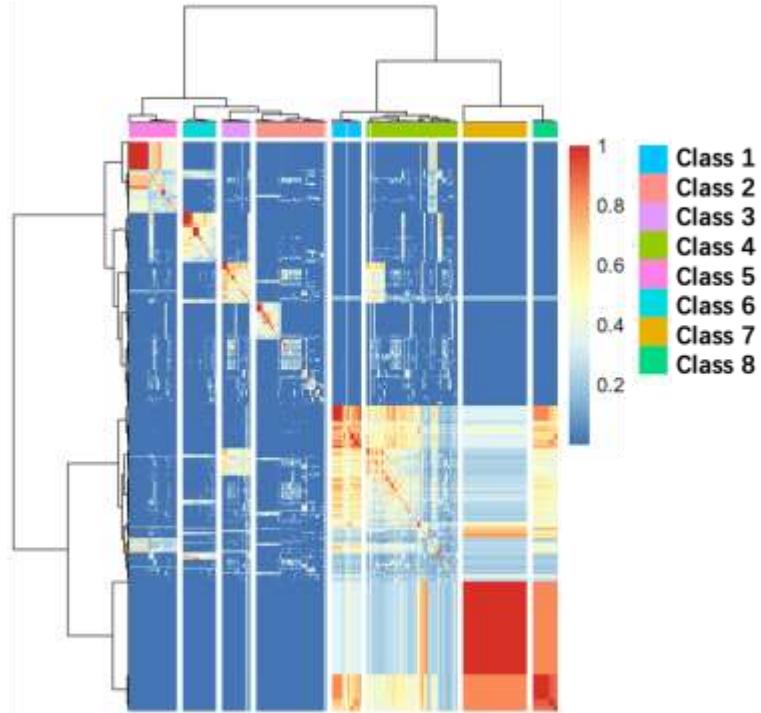
➤ Beneficial for correlating disease phenotypes with animal models to facilitate the integration of some animal model diseases and treatment models.



## The HPO is widely utilized in phenotype-driven automatic interpretation tools:

Phenotype-driven automatic interpretation tools rely on ontology-based similarity algorithms to compute the similarity between inputted HPO and disease database records, in order to assess potential variations in similar phenotypes.

Valentina Cipriani et al.(2020) Genes.



矮小表型数据集

Ontologies are expressed in formal language like Web Ontology Language (**OWL**)

The OBO Foundry: <http://www.obofoundry.org/>

The screenshot shows the homepage of The OBO Foundry. At the top, there is a navigation bar with links for About, Principles, Ontologies, Participate, FAQ, Legacy, a search bar labeled "Search Ontobee", and a "Submit" button. To the left of the navigation is a small icon of a person working on a computer. Below the navigation, the title "The OBO Foundry" is displayed in a large, bold font. A blue banner below the title contains the text "Welcome to the new OBO website! See the Announcement for more info.". Underneath the banner, there is a section titled "Download table as: [ YAML | JSON-LD | RDF/Turtle ]". The main content area lists five ontologies in a table format:

chebi	Chemical Entities of Biological Interest	A structured classification of molecular entities of biological interest focusing on 'small' chemical compounds. <a href="#">Detail</a>	<a href="#"></a>								
doid	Human Disease Ontology	An ontology for describing the classification of human diseases organized by etiology. <a href="#">Detail</a>	<a href="#"></a>								
go	Gene Ontology	An ontology for describing the function of genes and gene products <a href="#">Detail</a>	<a href="#"></a>	<a href="#"></a>	<a href="#"></a>		<a href="#"></a>	<a href="#"></a>	<a href="#"></a>		
obi	Ontology for Biomedical Investigations	An integrated ontology for the description of life-science and clinical investigations <a href="#">Detail</a>	<a href="#"></a>								
pato	Phenotypic quality	An ontology of phenotypic qualities (properties, attributes or characteristics) <a href="#">Detail</a>	<a href="#"></a>								

## Welcome to BioPortal, the world's most comprehensive repository of biomedical ontologies

Search for a class

Enter a class, e.g. Melanoma

[Advanced Search](#)

Ontology Visits (September 2018)

Ontology	Visits (September 2018)
CPT	~160,000
MEDDRA	~35,000
RXNORM	~25,000
SNOMEDCT	~15,000
NDDF	~10,000

Find an ontology

Start typing ontology name, then choose from list

[Browse Ontologies](#)

BioPortal Statistics

Ontologies	729
Classes	9,594,071
Resources Indexed	48
Indexed Records	39,537,360
Direct Annotations	95,468,433,792

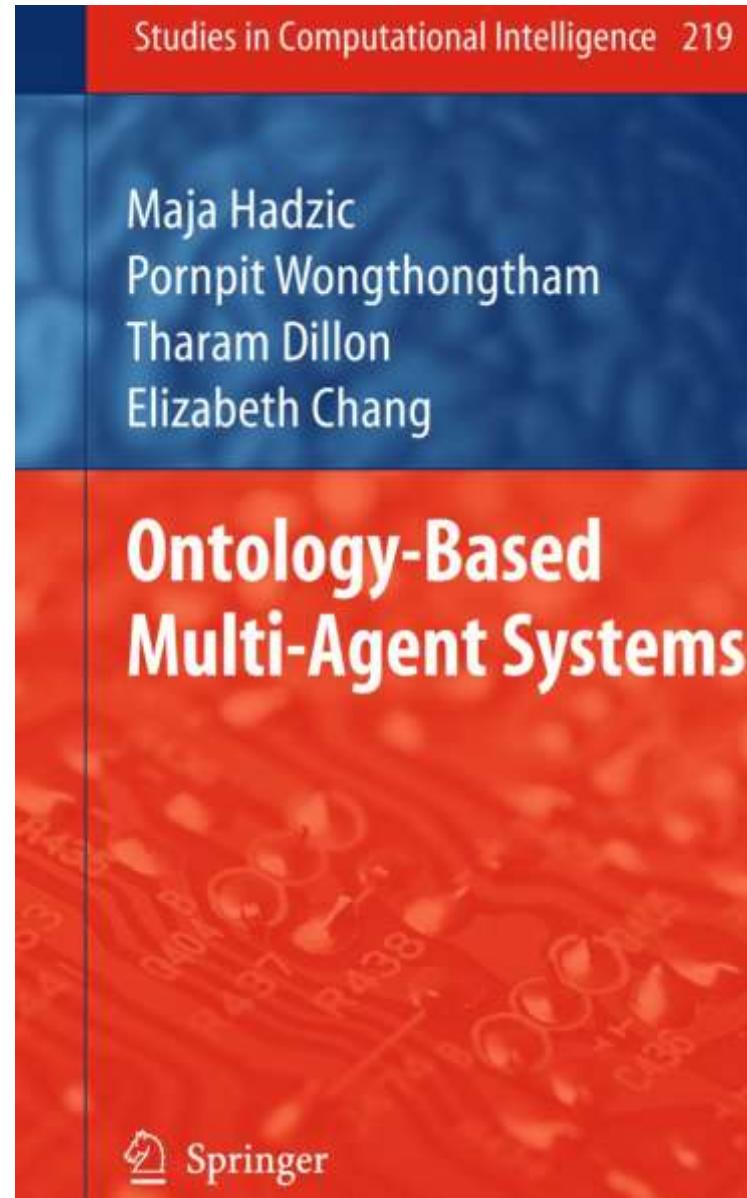
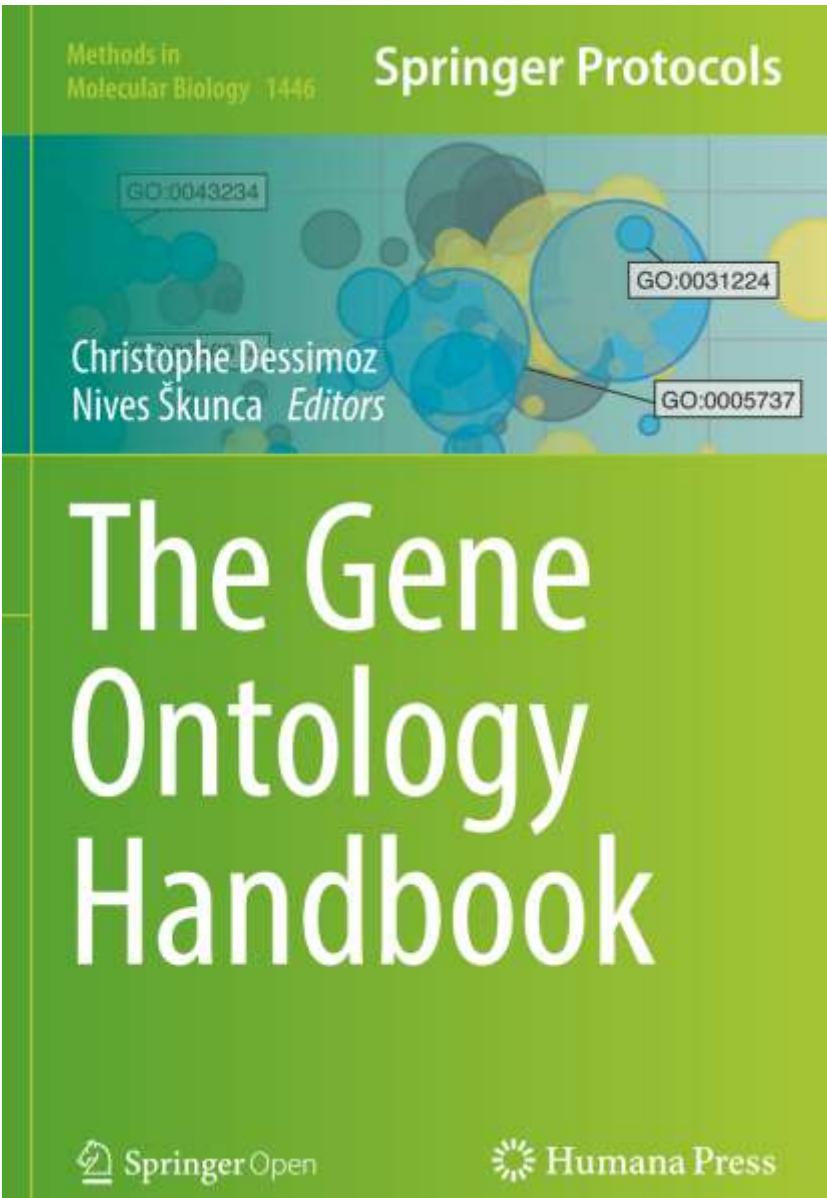
A library of ontologies and ontology related services

<http://bioportal.bioontology.org/>

<https://www.ebi.ac.uk/ols/index>

<http://www.ontobee.org/>

<http://amigo.geneontology.org/amigo>



[Back to Content](#)

# **Oncofetal reprogramming and oncofetal ecosystem (EMT and cell-cell communication)**

Zhou Ying

26<sup>th</sup> Jan. 2024

# The history of oncofetal reprogramming

**Before 1960:** Comparative antigen–antibody reactions (such as the precipitin inhibition assay) paved the way for oncofetal antigens to be discovered

First oncofetal antigen reported in HCC

Oncofetal antigens identified in gastrointestinal organs

Oncofetal antigen AFP used as a blood-based biomarker of HCC

lncRNA PVT1 identified — high levels of PVT1 linked to poor prognosis in patients with HCC

lncRNA H19 shown to inhibit p53 in several cancers

miRNA discovery

miR-98 shown to modulate the expression of genes required for cancer progression

Single-cell RNA sequencing

Identification of cell states shared by paediatric Wilms tumour cells and fetal kidney cells



mRNA discovery

Oncofetal antigens identified in gynaecological cancers

Advances in molecular biology led to the comparative northern blot and paved the way for lncRNA and miRNA discovery

Tumours termed "wounds that do not heal"

Oncofetal antigen CA125 used as a blood-based biomarker of ovarian cancer

Next-generation sequencing permitted the development of single-cell sequencing

Oncofetal antigen SALL4 identified in HCC

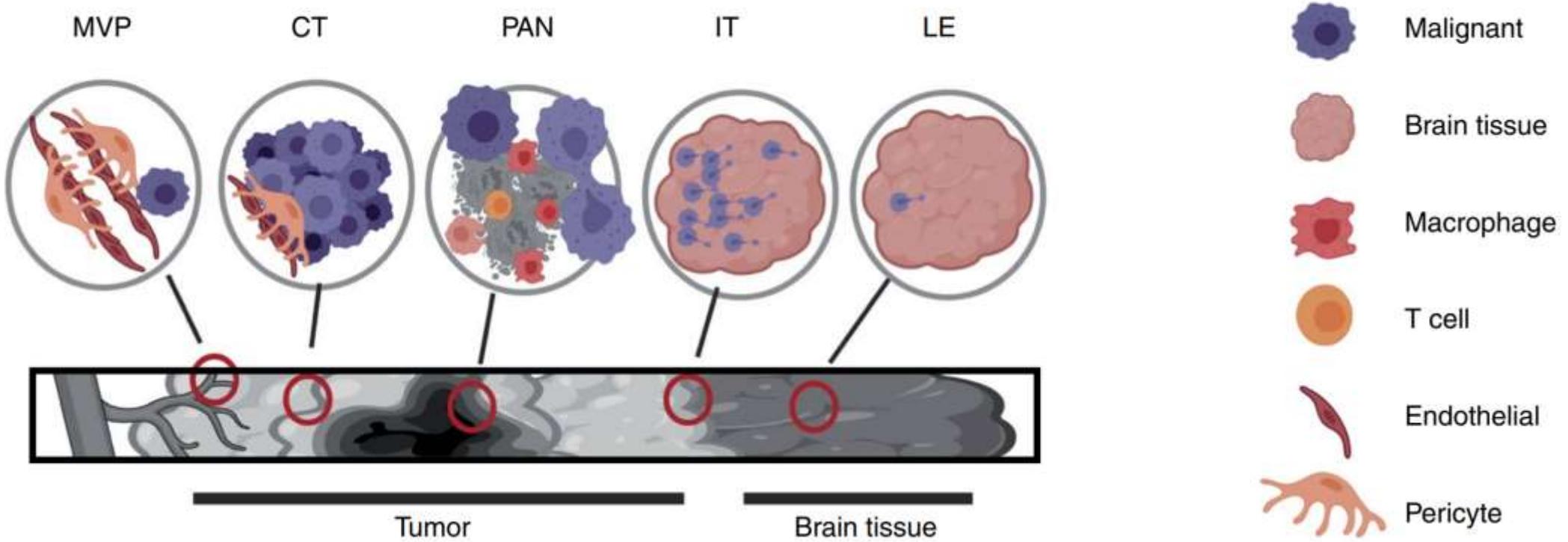
Spatial transcriptomics

Oncofetal reprogramming in the tumour microenvironment

Conceptual advances

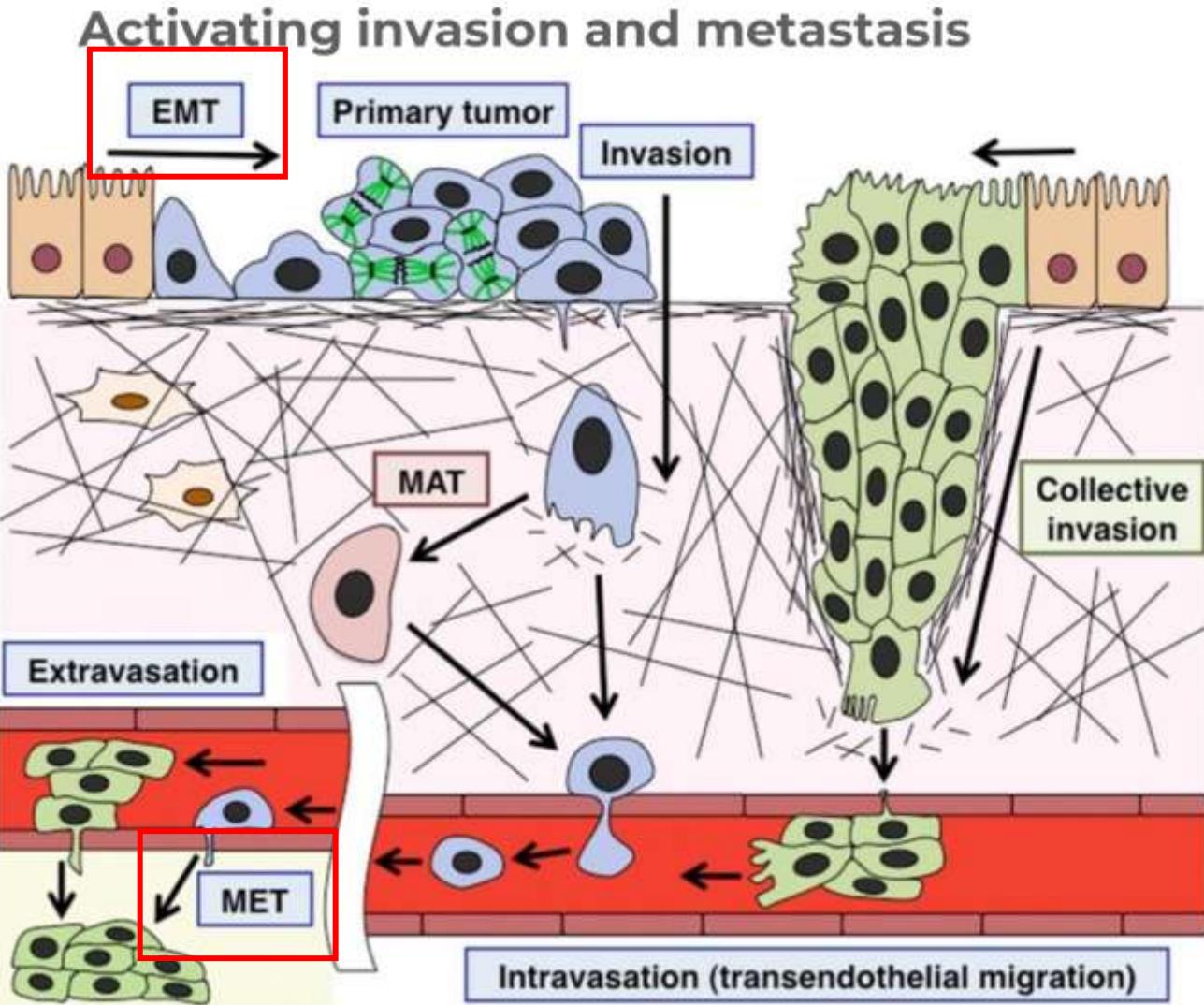
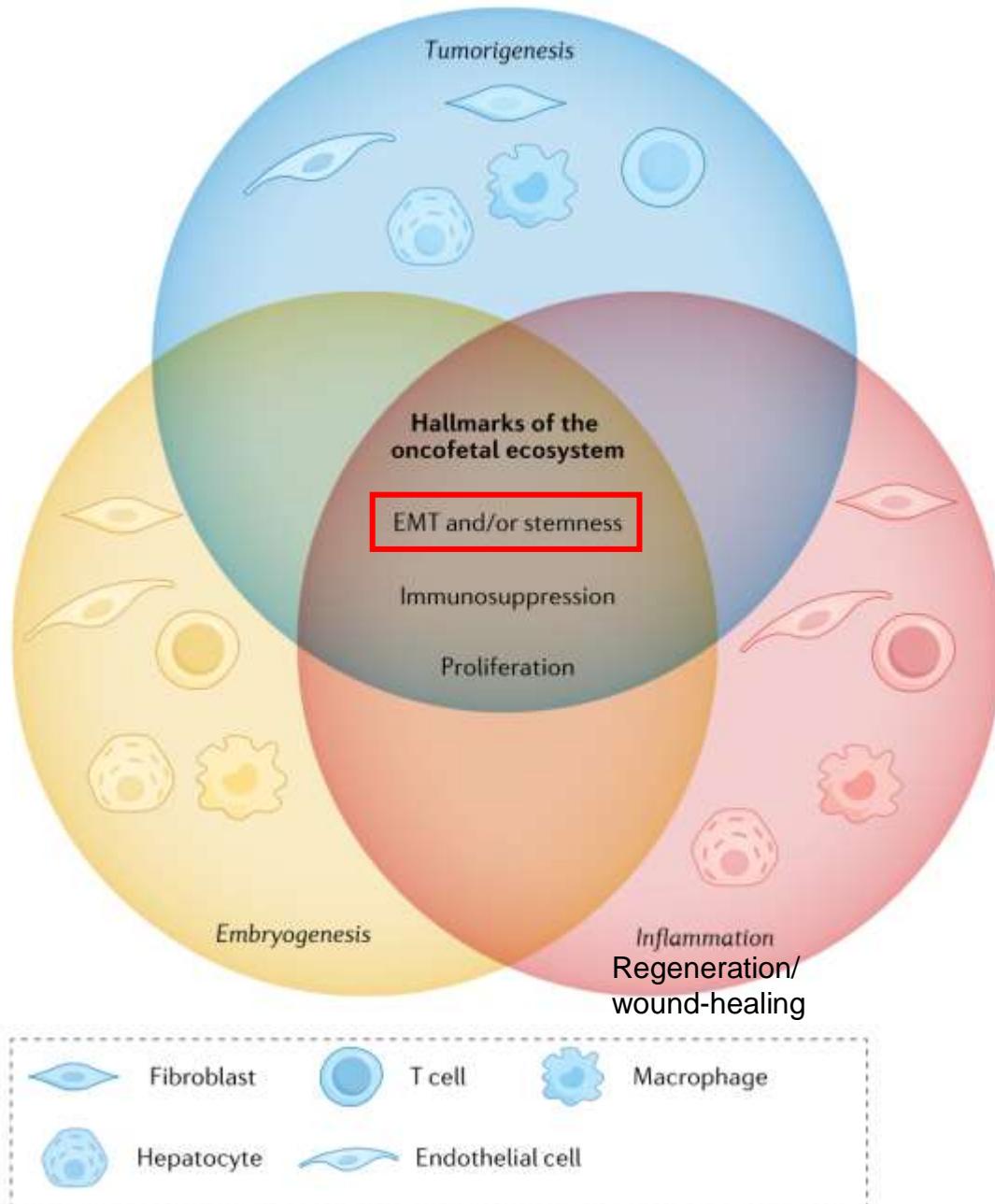
Oncofetal antigens

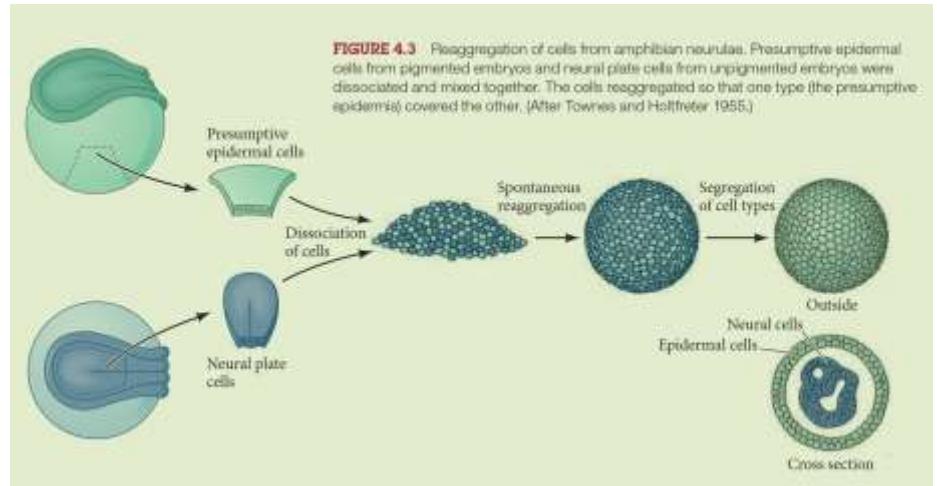
Advances in molecular biology



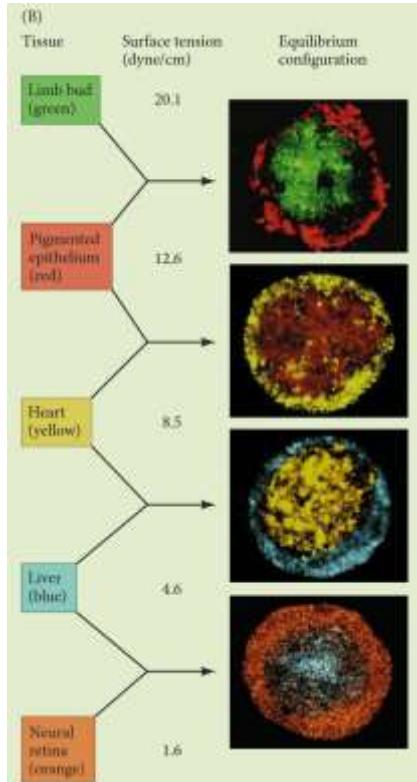
*Tinyi Chu et al. (2022) Nat. Cancer*

- **Oncofetal reprogramming:** tumor cells gain fetal properties (oncofetal protein, regulated RNA etc.)
- **Oncofetal ecosystem (Relationship of various cells):** healthy stromal cells reprogramming gaining fetal properties, leading to a fetal-like TME constructing.

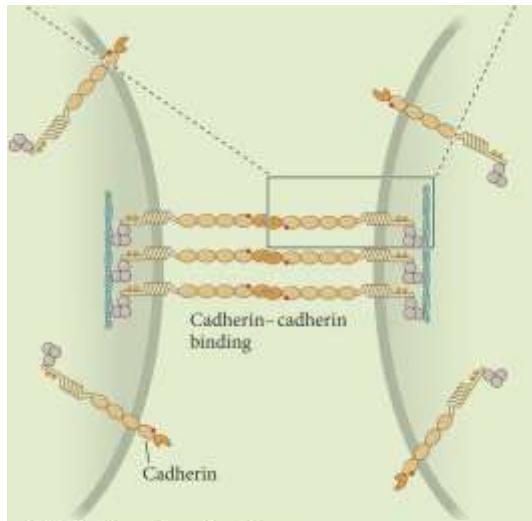




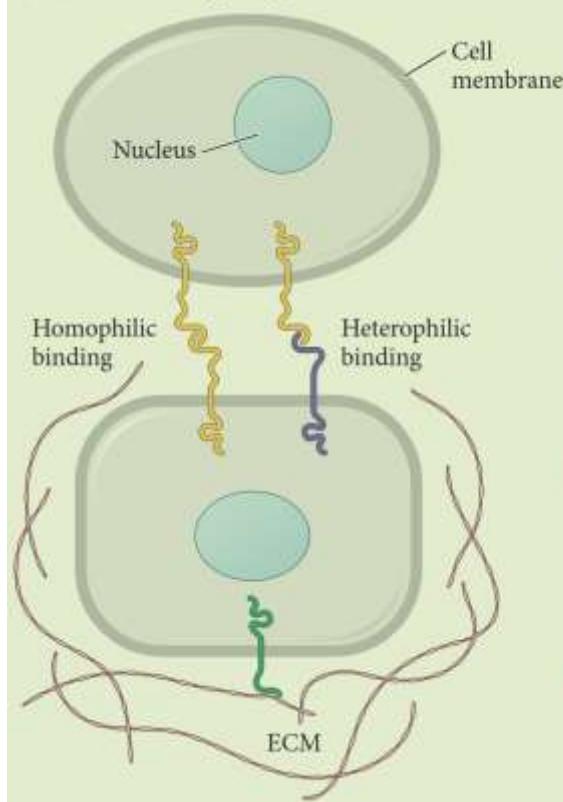
(After Townes and Holtfreter 1955.)



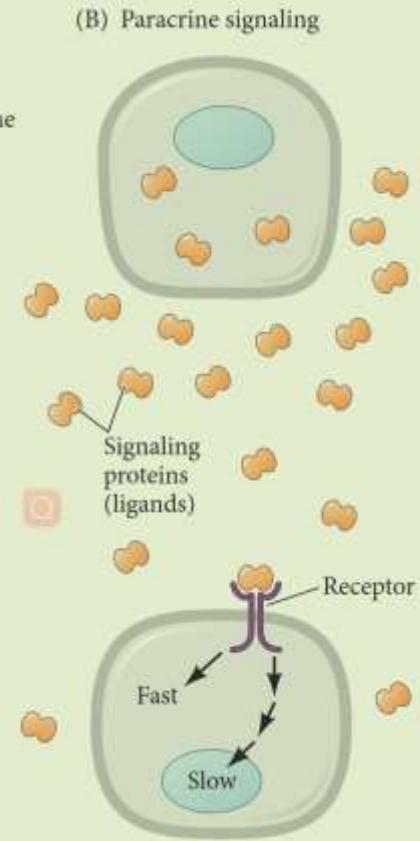
The cells with high surface tension will spontaneously distribute to the inner layer.



(A) Juxtacrine signaling

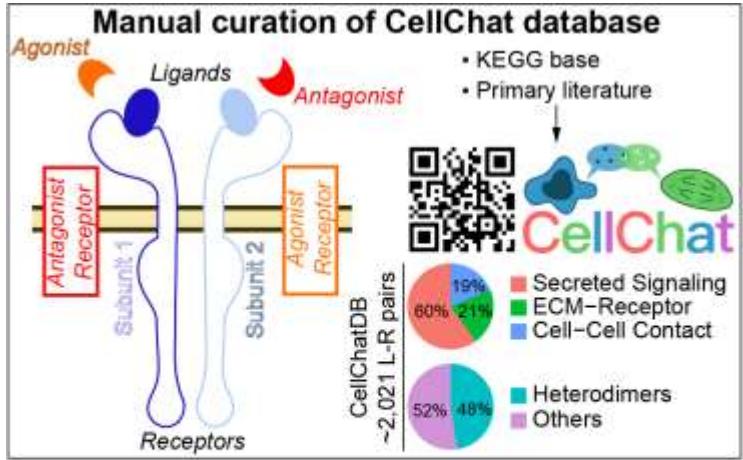


(B) Paracrine signaling



The adhesion between cells, mediated by ligand binding communication, is crucial for the EMT process.

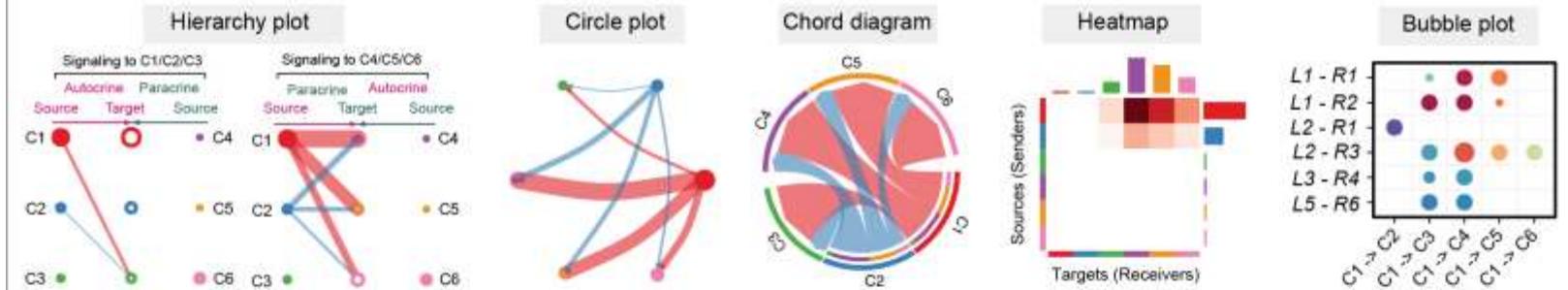
(From Foyt et al. 1996, courtesy of M. S. Steinberg and R. A. Foyt.)



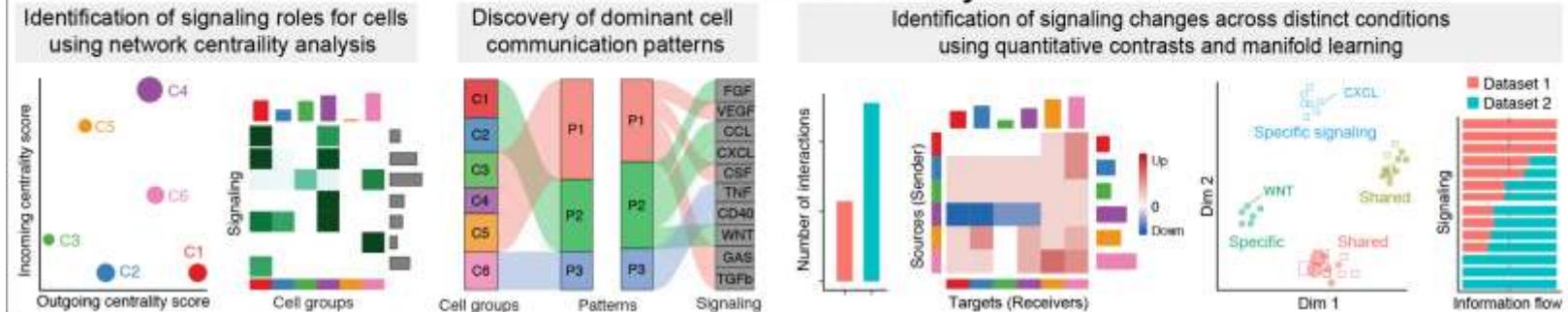
## Cellular communication modeling

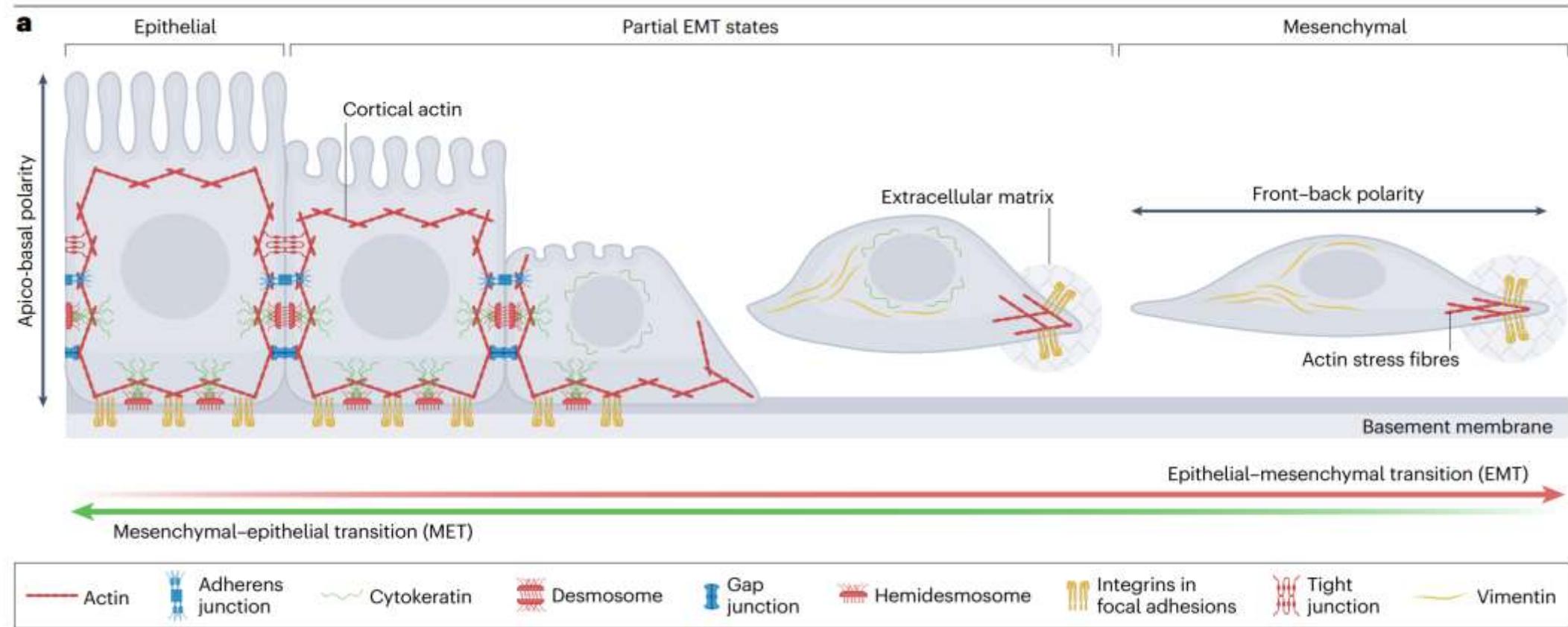
- scRNA-seq data input and processing
- Identification of over-expressed genes per cell group
- Quantification of communication probability using the law of mass action
- Inference of statistically and biologically significant cellular communications

## Cellular communication visualization



## Cellular communication analysis





### 1. Main features of epithelial state:

- Apical-basal polarity
- Attachment to the basement membrane
- Strong cell-cell adhesion
- Immobile
- Stably anchored in tightly lined epithelial sheets

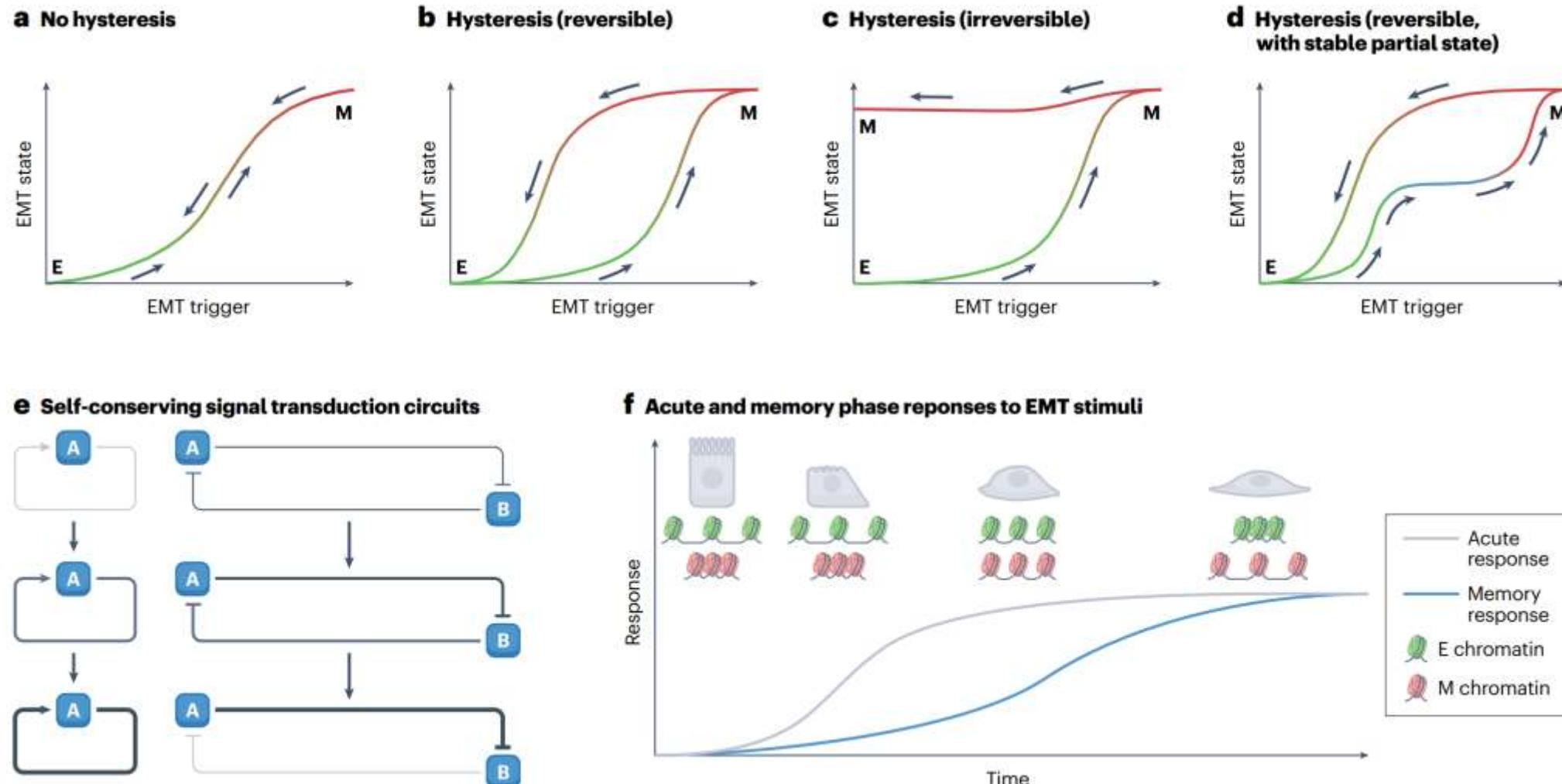
### 2. EMT

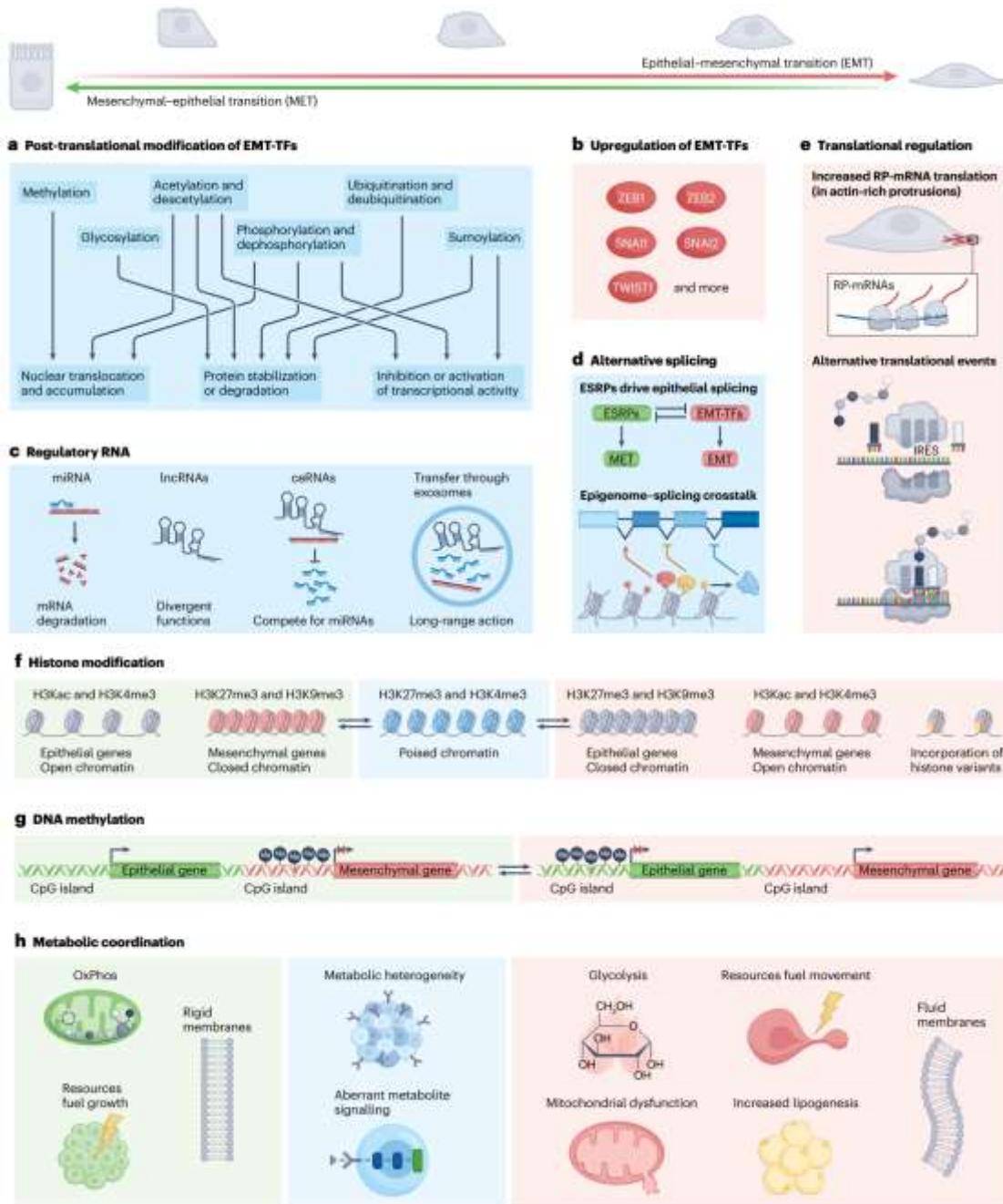
*Jef Haerinck et al. (2023) Nat. Reviews Genetics*

- Some adhesive molecules are reduced (lost epithelial cell-cell adhesion)
- The interaction pattern with the extracellular matrix (ECM) is altered (re-modulation of matrix interactions)
- Cell polarity changes, morphing into spindle-like shape (apical-basal polarity change to front-back polarity)

### 3. EMT Classification

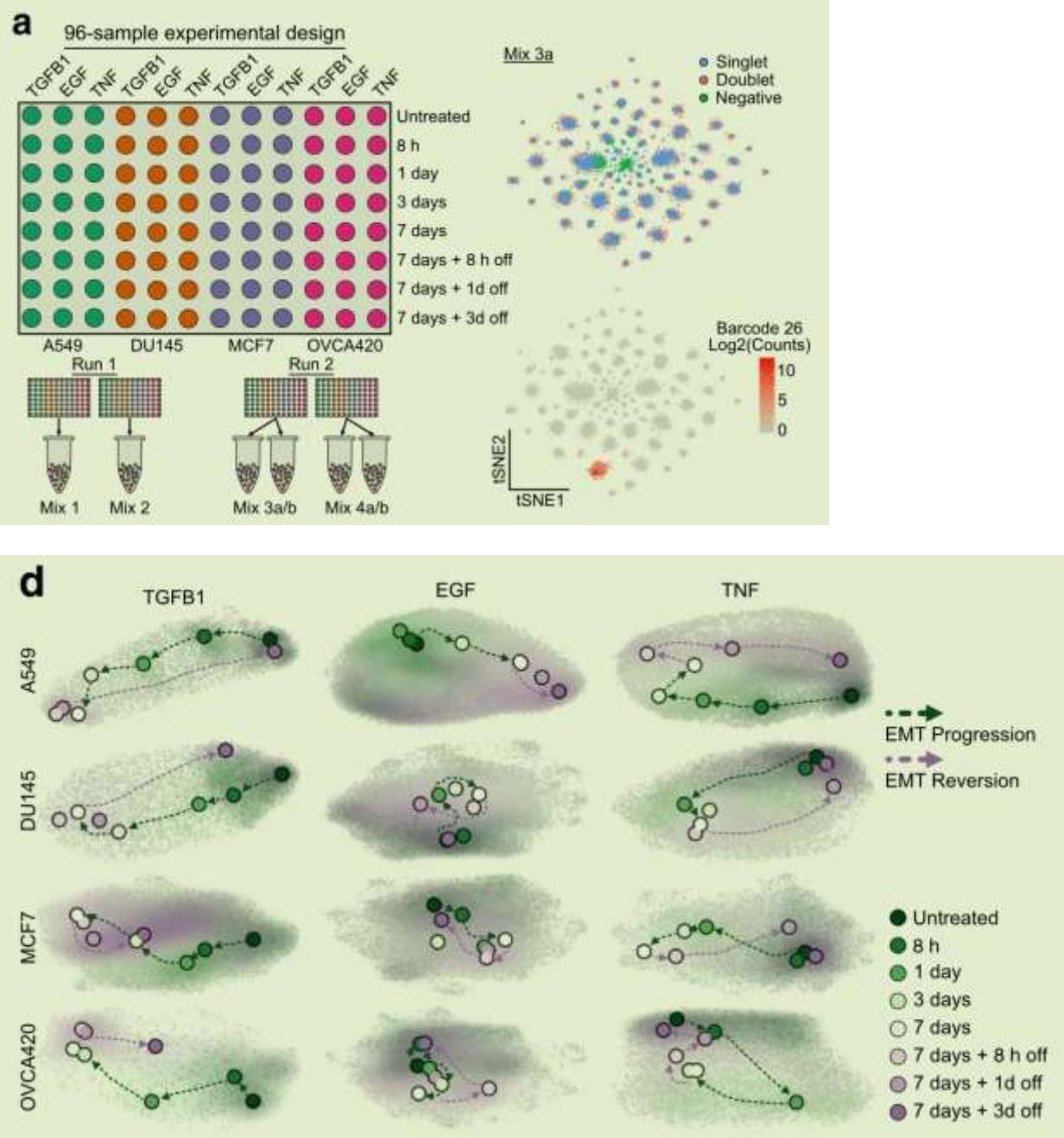
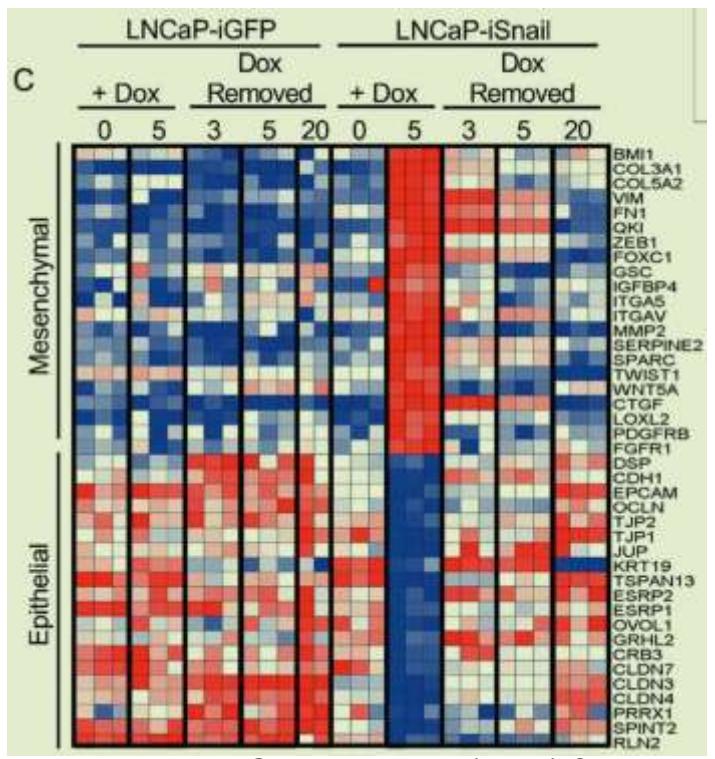
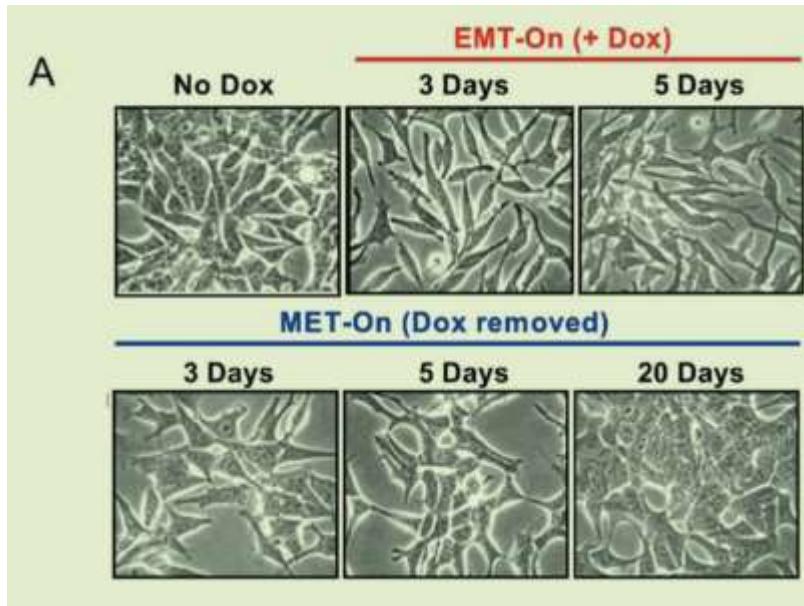
- **Type 1:** Many morphological movements occur during early embryonic development, undergoing continuous EMT and MET, which are considered the driving force behind primitive gut formation and organogenesis.
- **Type 2:** Mechanisms involved in wound healing, **tissue regeneration, and organ fibrosis.**
- **Type 3: Leading to cancer**





## EMT is subjected to multifaceted molecular regulation

- ◆ adhesion factors, changes in cellular cytoskeleton, chemotactic factors, membrane fluidity (lipid metabolism), and other related molecular processes, regulated by relevant transcription factors. Such transcription factors are referred to as EMT-TFs, such as the families of SNAIL, TWIST, ZEB, etc. (Figure b).
- ◆ Mutations, methylation (Figure g), changes in chromatin conformation (Figure f), regulation of non-coding RNAs (Figure c), and alternative splicing regulation (Figure d) of these related molecular genes all influence their transcription (Figure a), thereby affecting EMT.
- ◆ Alternative splicing, mRNA modifications (m6A methylation), and ribosome biogenesis (Figure e) all influence changes in EMT-TFs at the translation level.
- ◆ Related metabolism, mainly through lipid metabolism, such as the production of short-chain unsaturated fatty acids increasing membrane fluidity to promote EMT, whereas inhibiting it in reverse (Figure h).
- ◆ cancer cells induce EMT from oxidative phosphorylation (OxPhos) to glycolysis through various mechanisms:
  - Increasing glycolysis requires upregulation of glycolytic enzymes and glucose transporters, both of which can induce EMT.
  - Cancer cells typically exhibit mitochondrial dysfunction, impairing OxPhos, further enhancing EMT.
  - Glycolysis by-products, such as lactate secretion, lead to microenvironment acidification, further promoting EMT.
  - In this context, EMT in cancer can also correspond to early development, where glycolysis and oxidative phosphorylation are respectively associated with EMT and MET in the neural crest and mesoderm



David P. Cook et al. (2020) Nat. Communication

Review article

 Check for updates

# The epithelial–mesenchymal plasticity landscape: principles of design and mechanisms of regulation

Jef Haerinck  <sup>1,2</sup>, Steven Goossens  <sup>2,3</sup> & Geert Berx  <sup>1,2</sup> 

Abstract

Sections

# Presence of onco-fetal neighborhoods in hepatocellular carcinoma is associated with relapse and response to immunotherapy

Received: 17 March 2023

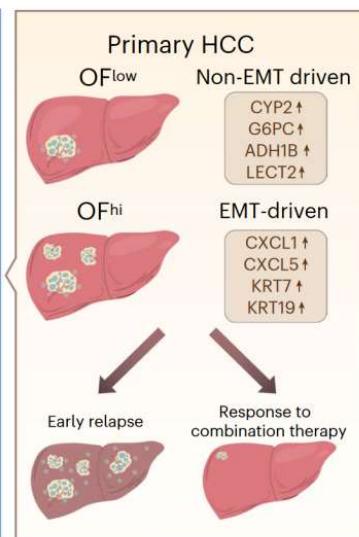
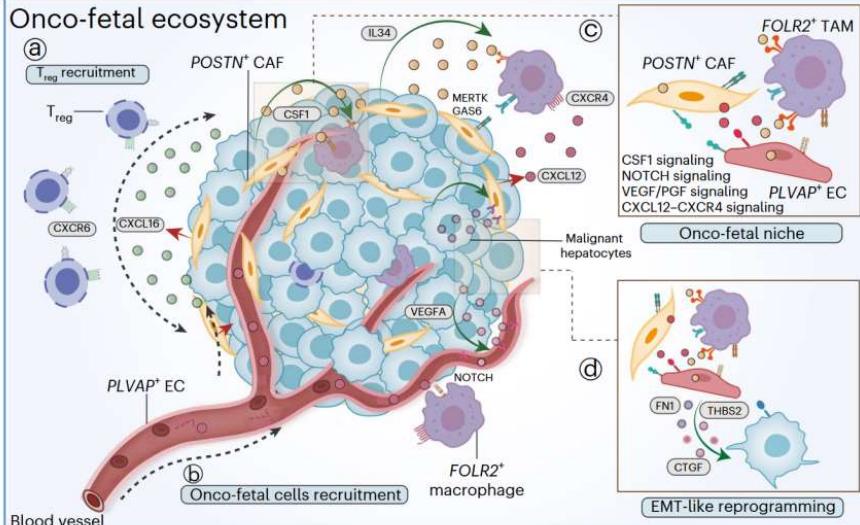
Accepted: 16 October 2023

Published online: 02 January 2024

Check for updates

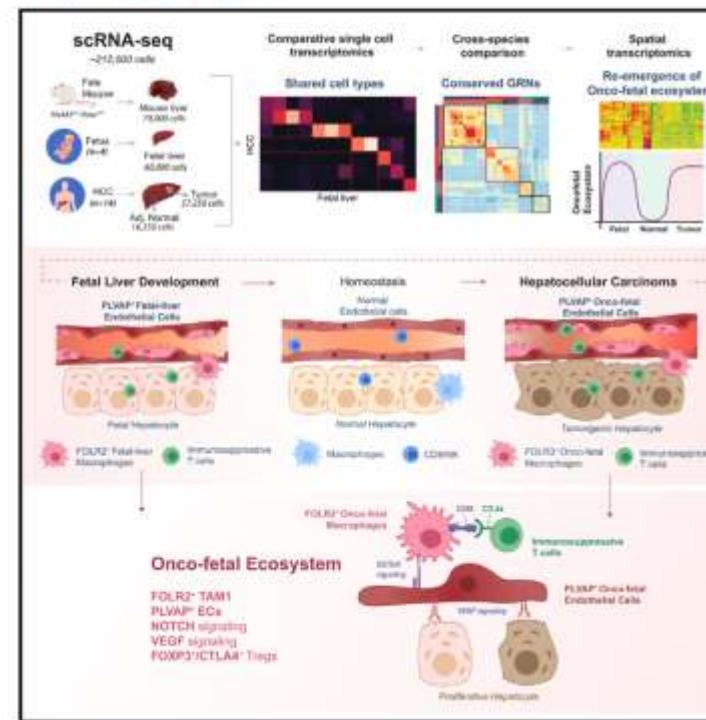
Ziyi Li<sup>1</sup>, Rhea Pai<sup>2,3,24</sup>, Saurabh Gupta<sup>2,3,24</sup>, Jennifer Currenti<sup>2,3,24</sup>, Wei Guo<sup>1,24</sup>, Anna Di Bartolomeo<sup>4</sup>, Hao Feng<sup>5,6</sup>, Zijie Zhang<sup>1</sup>, Zhizhen Li<sup>7</sup>, Longqi Liu<sup>1,8</sup>, Abhishek Singh<sup>2</sup>, Yinqi Bai<sup>8</sup>, Bicheng Yang<sup>9</sup>, Archita Mishra<sup>10,11</sup>, Katharine Yang<sup>10</sup>, Liang Qiao<sup>4</sup>, Michael Wallace<sup>12,13</sup>, Yujia Yin<sup>14</sup>, Qiang Xia<sup>5,6</sup>, Jerry Kok Yen Chan<sup>15,16,17</sup>, Jacob George<sup>18</sup>, Pierce Kah-Hoe Chow<sup>18,19</sup>, Florent Ginhoux<sup>1,10,20,21</sup> & Ankur Sharma<sup>1,2,3,22,23</sup>

g



# Onco-fetal Reprogramming of Endothelial Cells Drives Immunosuppressive Macrophages in Hepatocellular Carcinoma

## Graphical Abstract



## Authors

Ankur Sharma, Justine Jia Wen Seow, Charles-Antoine Dutertre, ..., Pierce K.H. Chow, Florent Ginhoux, Ramanuj DasGupta

## Correspondence

sharmaa@gis.a-star.edu.sg (A.S.), pierce.chow.k.h@singhealth.com.sg (P.K.H.C.), florent\_ginhoux@immunol.a-star.edu.sg (F.G.), dasguptar@gis.a-star.edu.sg (R.D.)

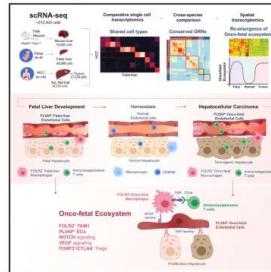
## In Brief

A single-cell atlas of human liver from development to disease suggests a shared onco-fetal ecosystem driving immunosuppression in fetal liver and hepatocellular carcinoma.

## Oncofetal ecosystem

**Onco-fetal Reprogramming of Endothelial Cells Drives Immunosuppressive Macrophages in Hepatocellular Carcinoma**

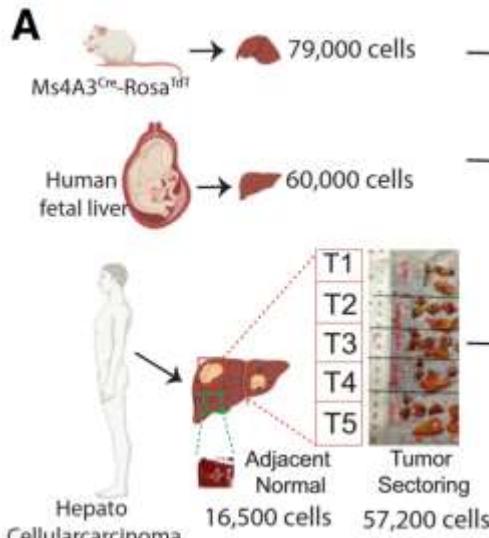
## Graphical Abstract



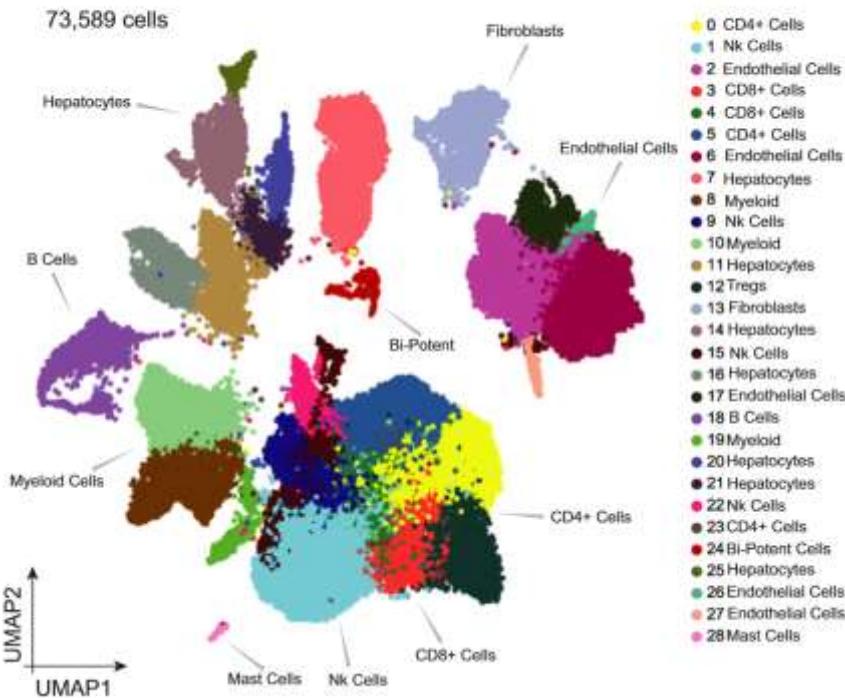
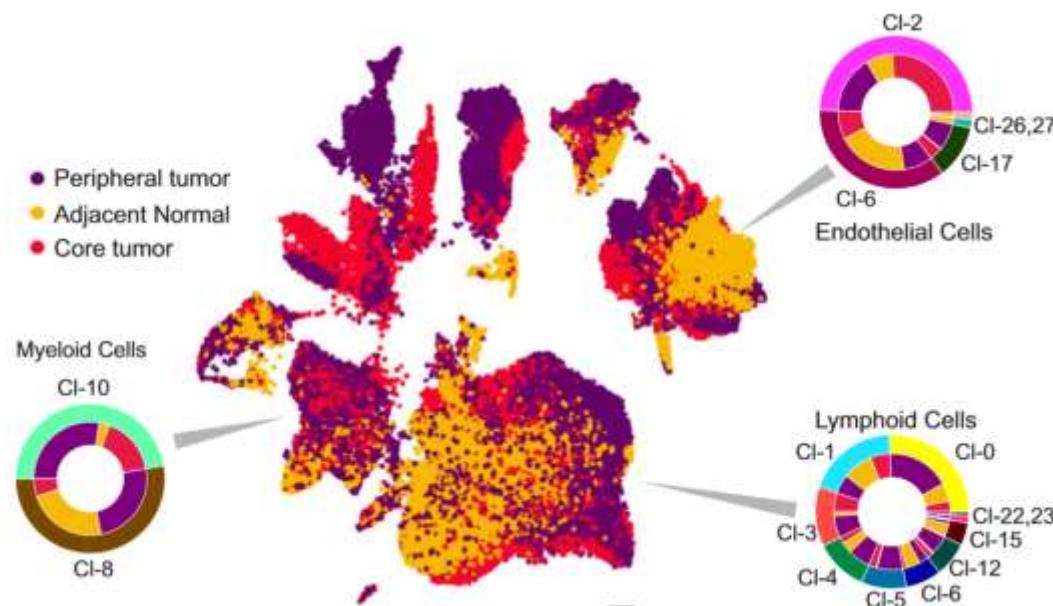
Authors  
Ankur Sharma, Justine Jia Wen Seow,  
Charles-Antoine Dutertre, ...  
Pierro K.H. Chow, Florent Ginhoux,  
Ramanuj DasGupta

Correspondence  
sharmaa@i2r.a-star.edu.sg (A.S.),  
pierro.chow.kh@singhealth.com.sg  
(P.K.H.C.),  
florent.ginhoux@immuno.i2r.a-star.edu.sg  
(F.G.),  
dasguptar@i2r.a-star.edu.sg (R.D.)

In Brief  
A single-cell atlas of human liver from development to disease suggests a shared onco-fetal ecosystem driving immunosuppression in fetal liver and hepatocellular carcinoma.

**B**

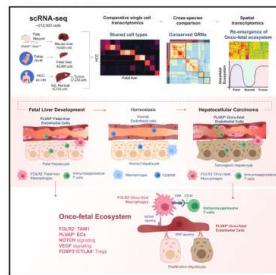
29 clusters in Human Liver

**C**

- Myeloid and EC subtypes have relatively specific distributions in tumors.
- EC is associated with TAM development.
- Macrophages are derived from two sources, hematopoietic stem cells and embryonic stem cells.

**Onco-fetal Reprogramming of Endothelial Cells Drives Immunosuppressive Macrophages in Hepatocellular Carcinoma**

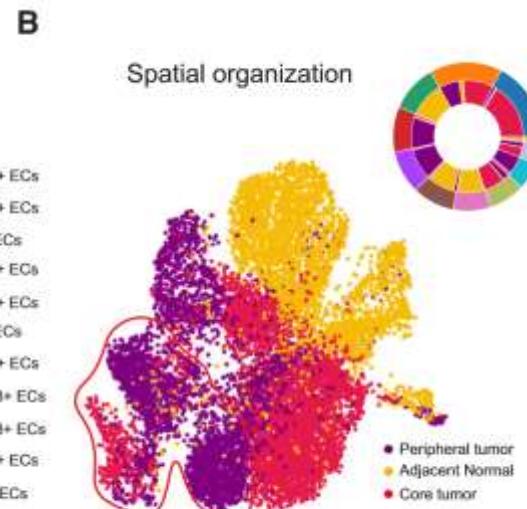
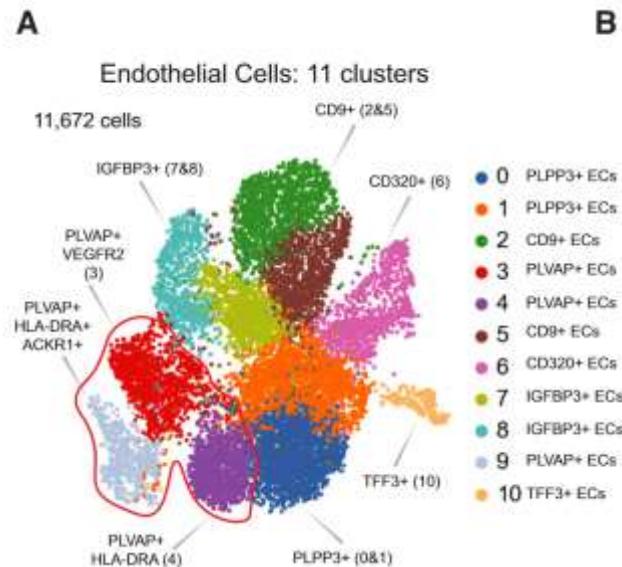
## Graphical Abstract



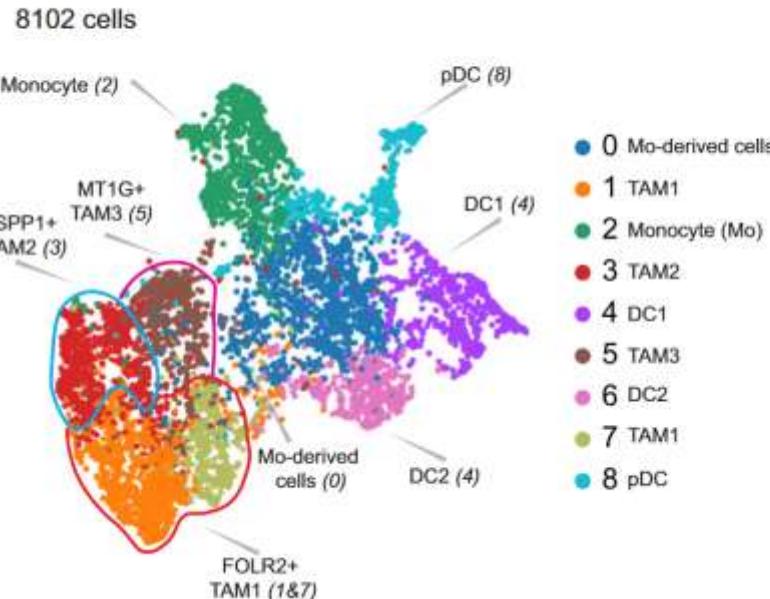
**Authors**  
Ankur Sharma, Justine Jia Wen Seow,  
Charles-Antoine Dutertre, ...  
Pierce K.H. Chow, Florent Ginhoux,  
Ramanuj DasGupta

**Correspondence**  
sharmaa@iis.a-star.edu.sg (A.S.),  
pierce.chow.kh@singhealth.com.sg  
(P.K.H.C.),  
florent.ginhoux@immuno.iis.a-star.edu.sg  
(F.G.),  
dasguptar@iis.a-star.edu.sg (R.D.)

**In Brief**  
A single-cell atlas of human liver from development to disease suggests a shared onco-fetal ecosystem driving immunosuppression in fetal liver and hepatocellular carcinoma.



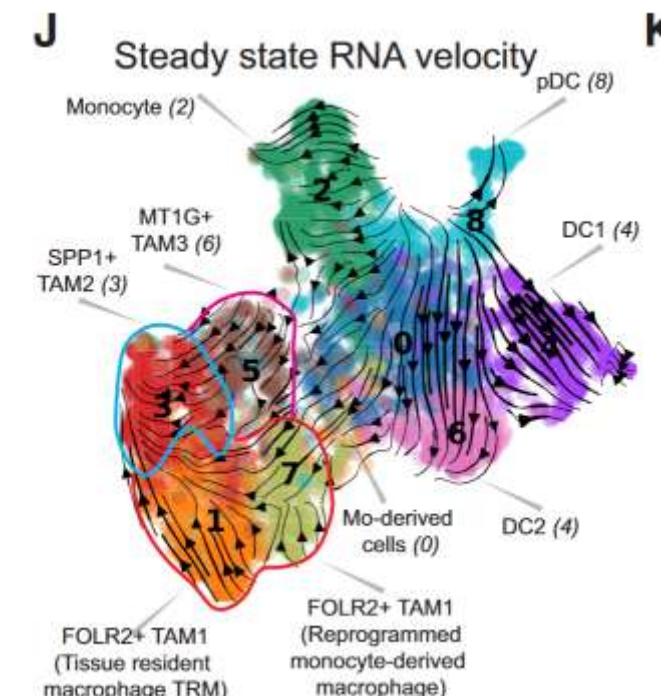
**F** Mononuclear phagocytes: 9 clusters



**G** Spatial organization

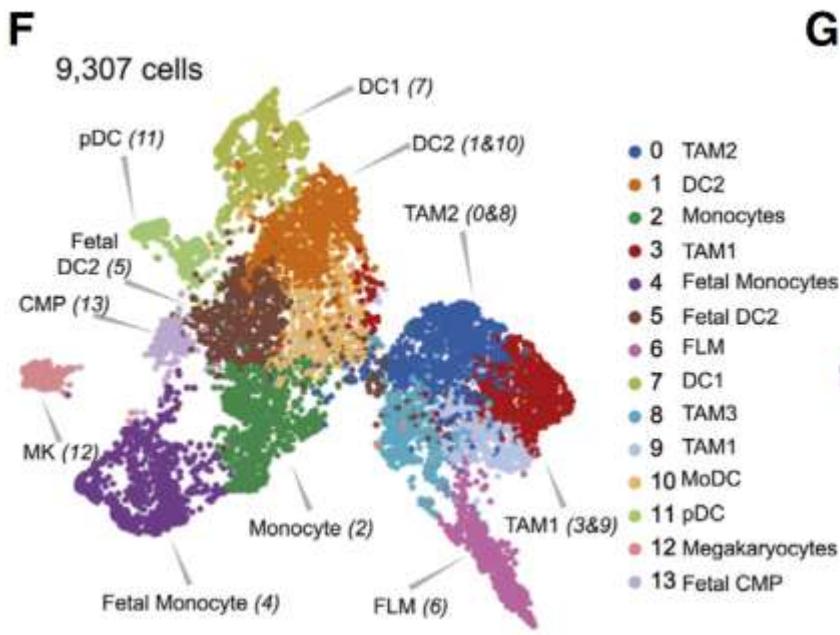
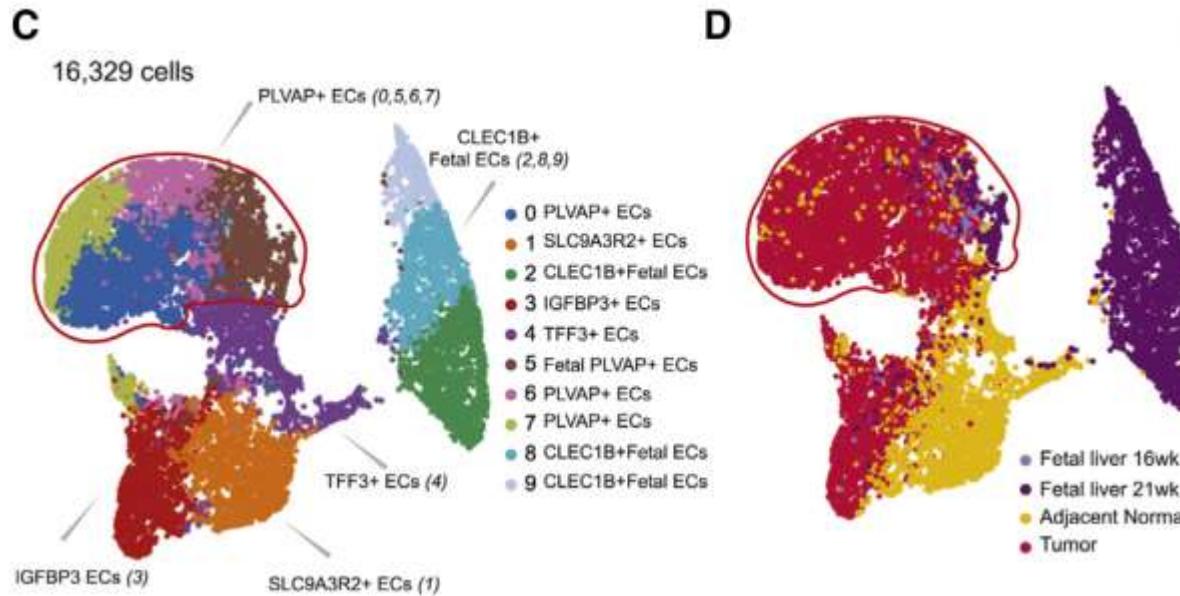
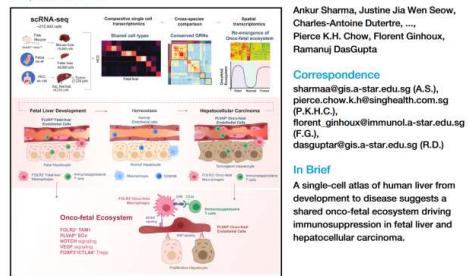
- Peripheral tumor
- Adjacent Normal
- Core tumor

- PLVAP+ EC and tumor association.
- FOLR2+ TAM1 cells are associated with tumors and can be divided into two sources.



**Oncofetal Reprogramming of Endothelial Cells Drives Immunosuppressive Macrophages in Hepatocellular Carcinoma**

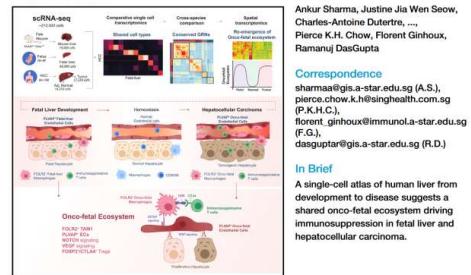
## Graphical Abstract



- I**
- LVAP+EC, FOLR2+ TAM1, both are capable of aggregating fetal and HCC cells together, and do not aggregate with adjacent normal cells.
  - Oncofetal protein

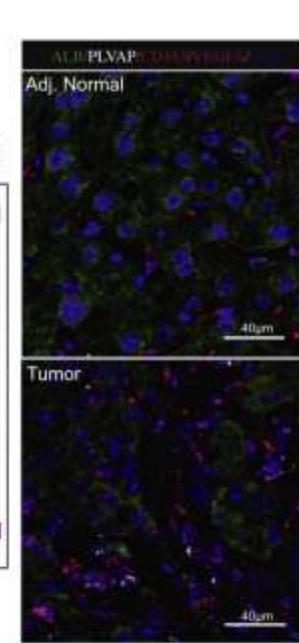
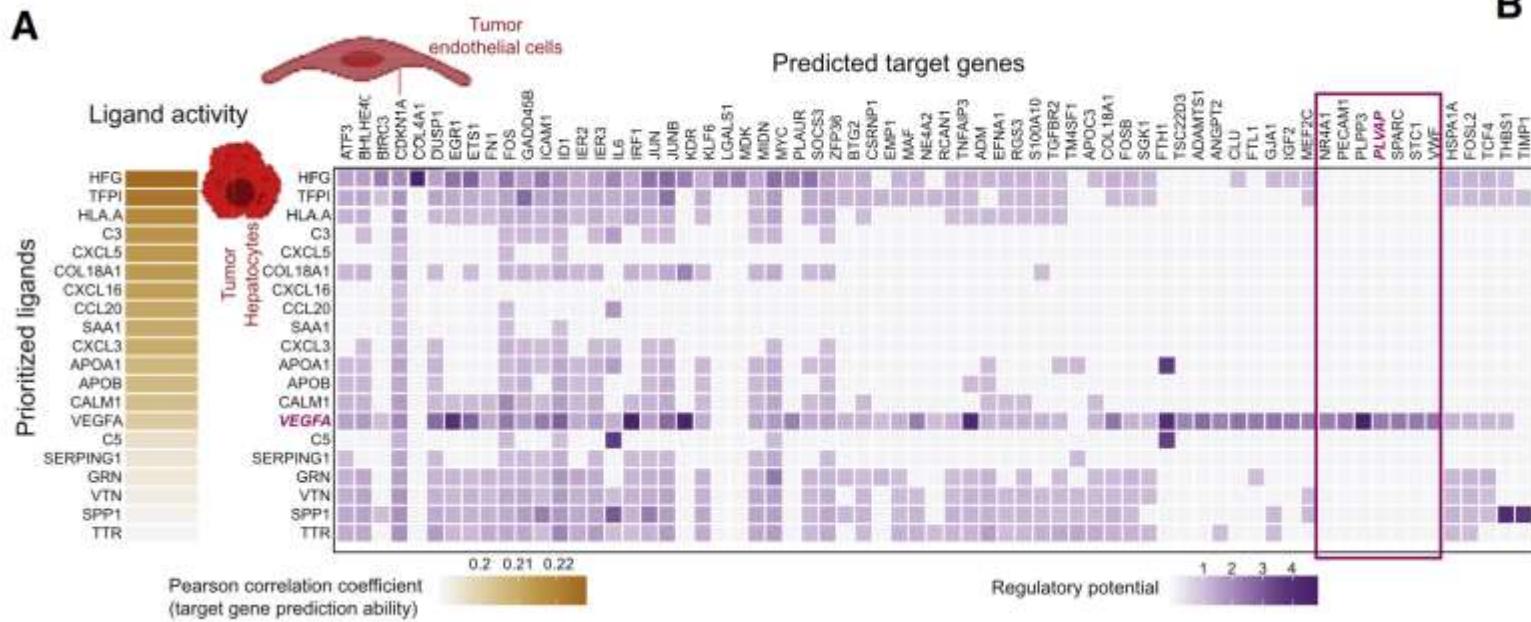
**Oncofetal Reprogramming of Endothelial Cells Drives Immunosuppressive Macrophages in Hepatocellular Carcinoma**

## Graphical Abstract

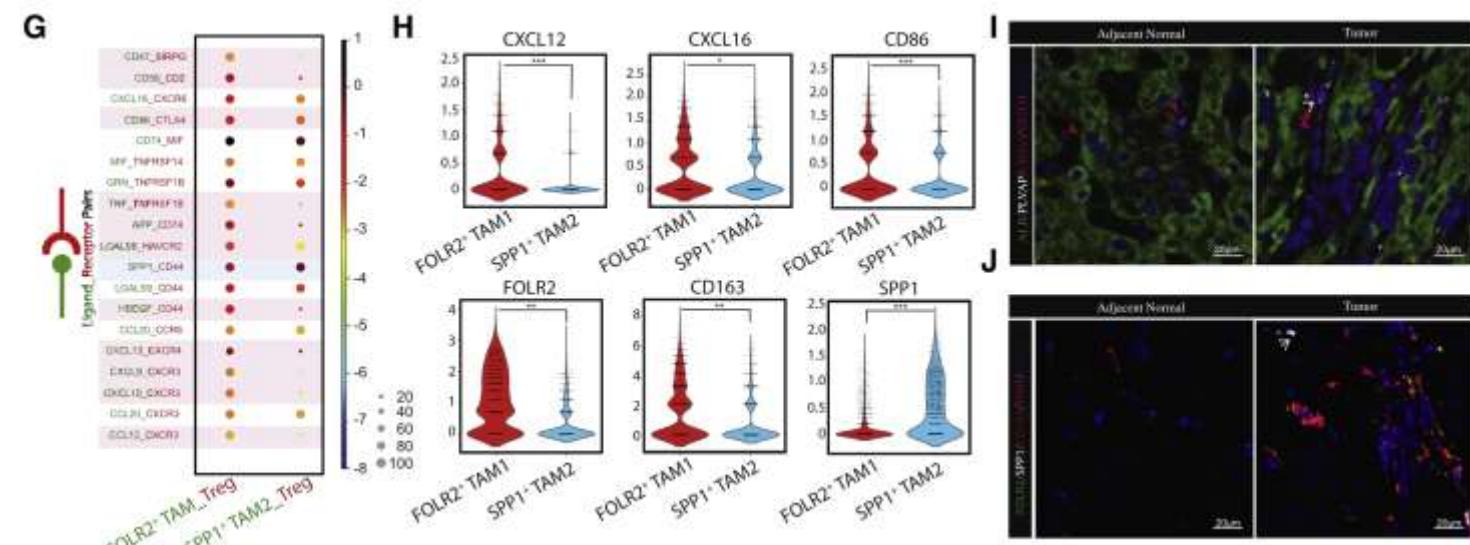


**Authors**  
Ankur Sharma, Justine Jia Wen Seow, Charles-Antoine Dutertre, ..., Piero K.H. Chow, Florent Ginhoux, Ramanuj DasGupta  
**Correspondence**  
sharma@gis.a-star.edu.sg (A.S.), piero.k.h.chow@singhealth.com.sg (P.K.H.C.), florent.ginhoux@immuno.a-star.edu.sg (F.G.), dasguptar@gis.a-star.edu.sg (R.D.)

**In Brief**  
A single-cell atlas of human liver from development to disease suggests a shared onco-fetal ecosystem driving immunosuppression in fetal liver and hepatocellular carcinoma.

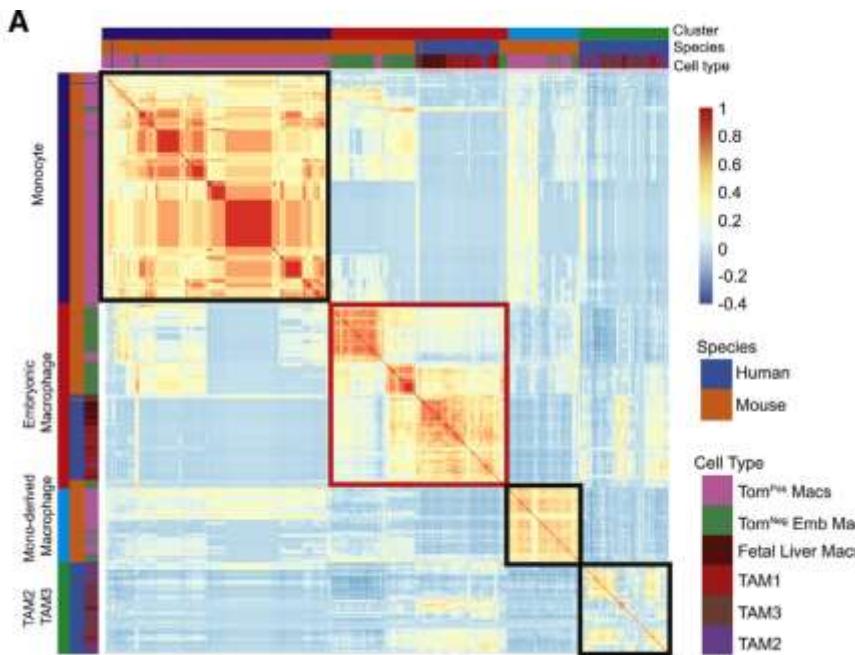
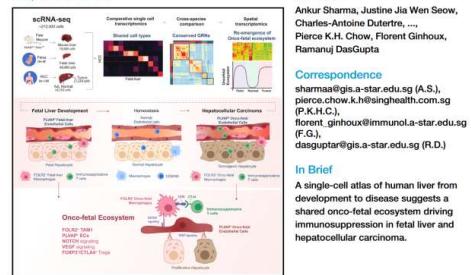


- VEGF may promote PLVAP expression.
- VEGF is an important factor in embryonic bile duct development and HCC-induced angiogenesis.
- Stem cell proliferation can secrete VEGF, potentially promoting oncofetal reprogramming in the tumor microenvironment (TME).
- FOLR2+ TAM promotes immune suppression and co-localizes with PLVAP+ Tregs in tumors.



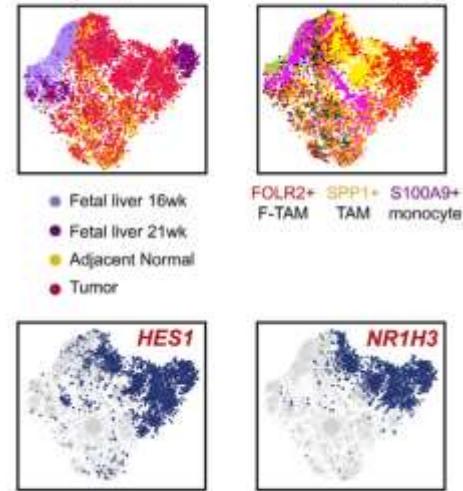
**Oncogenic Reprogramming of Endothelial Cells Drives Immunosuppressive Macrophages in Hepatocellular Carcinoma**

## Graphical Abstract

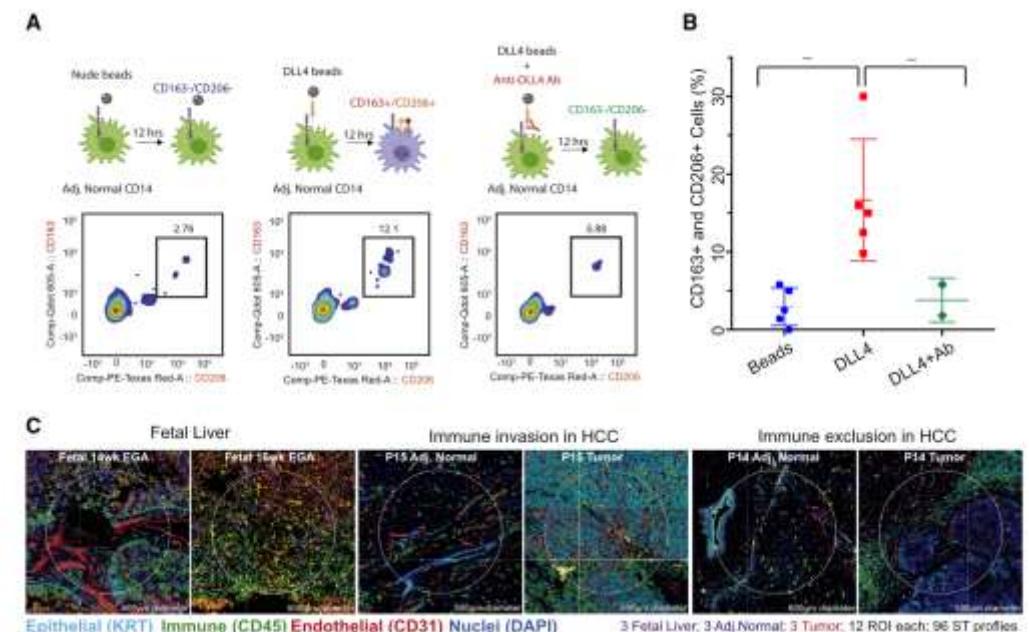


**C**

Conserved GRNs in  
Mouse embryonic, fetal liver and TAM1 macrophages



- SCENIC Study Reveals Transcriptional Regulatory Activation Network
- Fetal, HCC, non-myeloid derived cells from mice can aggregate, and specific distribution of the HES1 transcription factor regulated by the Notch pathway is found.
- Contact of cancer-adjacent CD14+ monocytes with Beads carrying Notch-activating ligands can induce the TAM1 phenotype.



# Presence of onco-fetal neighborhoods in hepatocellular carcinoma is associated with relapse and response to immunotherapy

Received: 17 March 2023

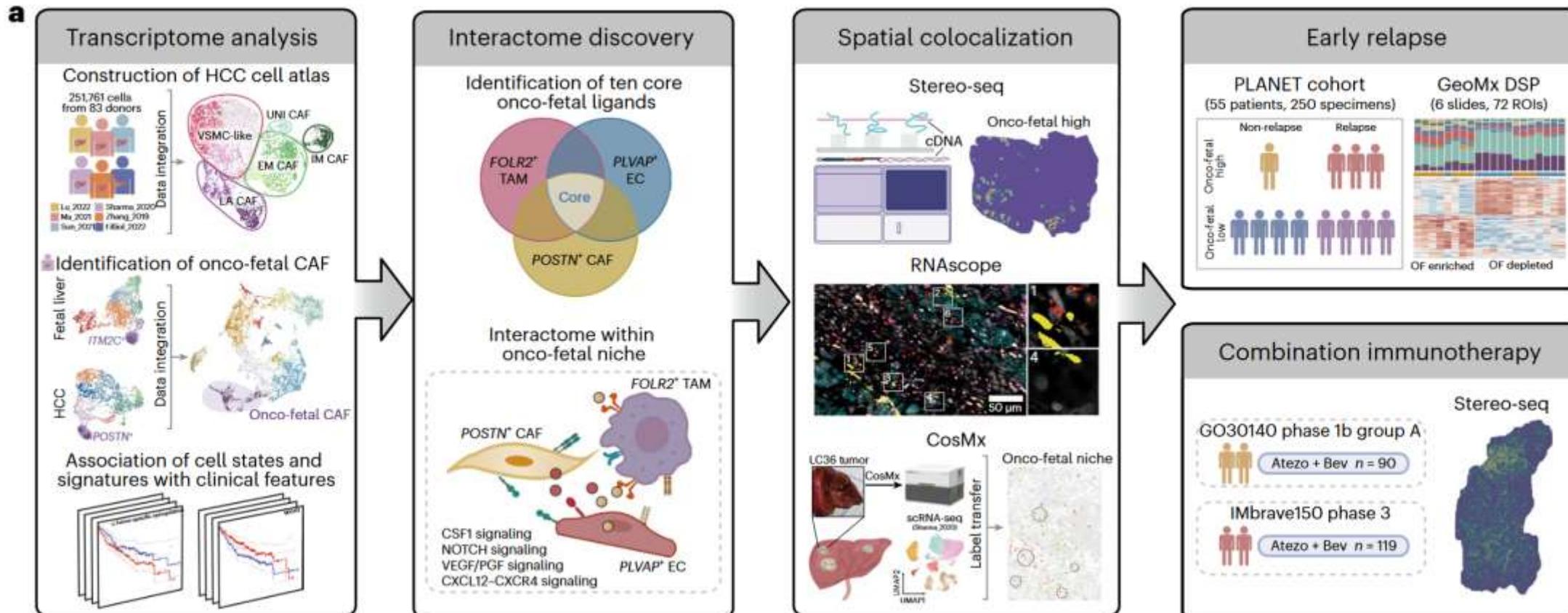
Accepted: 16 October 2023

Published online: 02 January 2024

Check for updates

Ziyi Li<sup>1</sup>, Rhea Pai<sup>2,3,4</sup>, Saurabh Gupta<sup>2,3,4</sup>, Jennifer Currenti<sup>2,3,4</sup>, Wei Guo<sup>1,24</sup>, Anna Di Bartolomeo<sup>5</sup>, Hao Feng<sup>5,6</sup>, Zijie Zhang<sup>6,7</sup>, Zhizhen Li<sup>7</sup>, Longqi Liu<sup>8,9</sup>, Abhishek Singh<sup>2</sup>, Yingqi Bai<sup>8</sup>, Bicheng Yang<sup>9</sup>, Archita Mishra<sup>10,11</sup>, Katharine Yang<sup>10</sup>, Liang Qiao<sup>4</sup>, Michael Wallace<sup>12,13</sup>, Yujia Yin<sup>14</sup>, Qiang Xia<sup>5,6</sup>, Jerry Kok Yen Chan<sup>15,16,17</sup>, Jacob George<sup>18</sup>, Pierce Kah-Hoe Chow<sup>18,19</sup>, Florent Ginhoux<sup>20,21</sup> & Ankur Sharma<sup>2,3,22,23</sup>

- Does the fetal-onco TME composed of PLVAP+EC and FOLR2+TAM1 cells also involve specific CAF cells?
- POSTN+CAF involvement is found to impact survival.
- Discovery of some onco-fetal ligands.
- Relationship between oncofetal high/low and tumor recurrence



# Presence of onco-fetal neighborhoods in hepatocellular carcinoma is associated with relapse and response to immunotherapy

Received: 17 March 2023

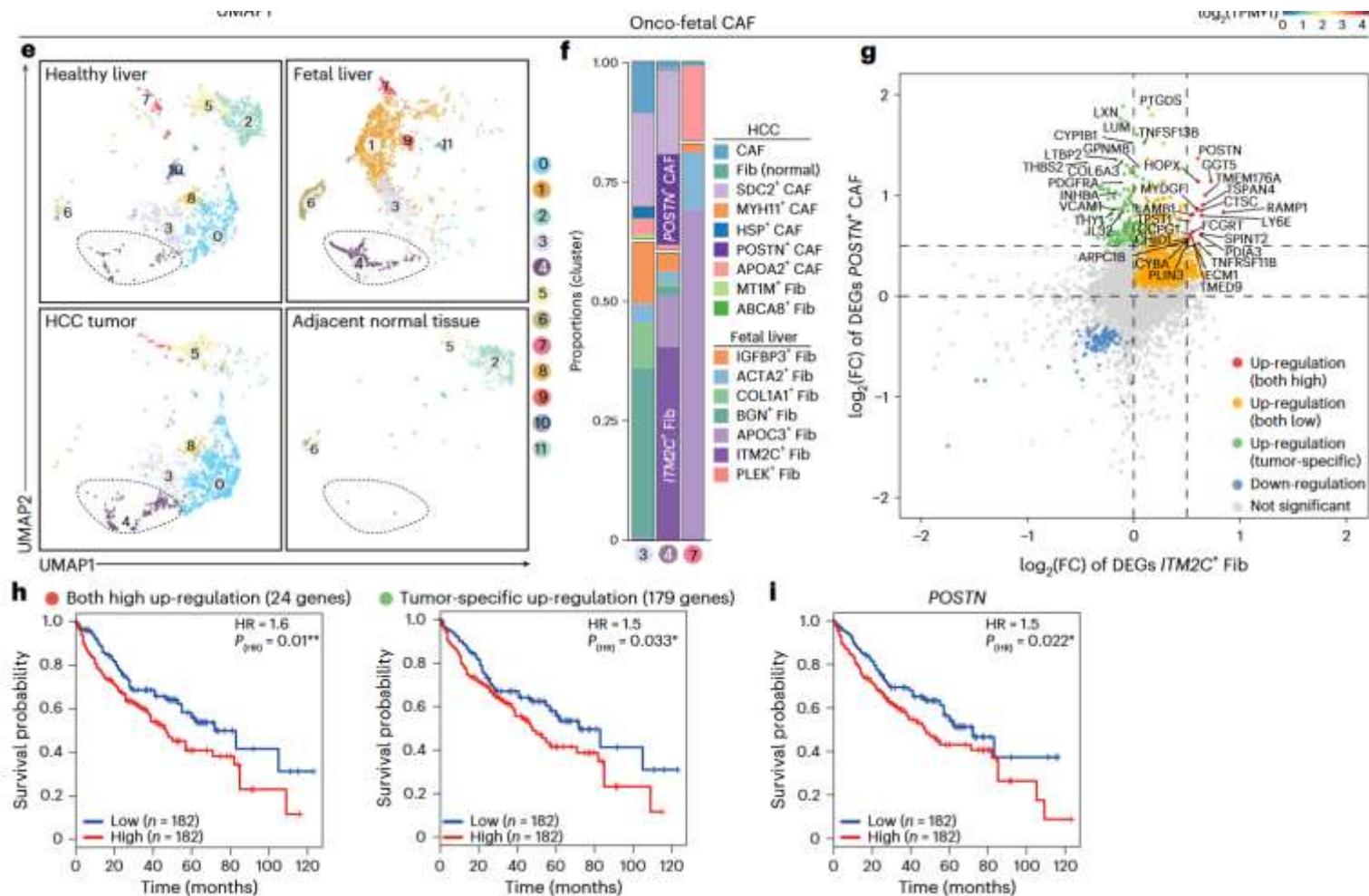
Accepted: 16 October 2023

Published online: 02 January 2024

Check for updates

Ziyi Li<sup>①</sup>, Rhea Pai<sup>2,3,24</sup>, Saurabh Gupta<sup>2,3,24</sup>, Jennifer Currenti<sup>2,3,24</sup>, Wei Guo<sup>1,24</sup>, Anna Di Bartolomeo<sup>4</sup>, Hao Feng<sup>5,6</sup>, Zijie Zhang<sup>②,7</sup>, Zhizhen Li<sup>7</sup>, Longqi Liu<sup>③,8</sup>, Abhishek Singh<sup>2</sup>, Yinqi Bai<sup>9</sup>, Bicheng Yang<sup>9</sup>, Archita Mishra<sup>10,11</sup>, Katharine Yang<sup>10</sup>, Liang Qiao<sup>4</sup>, Michael Wallace<sup>12,13</sup>, Yujia Yin<sup>14</sup>, Qiang Xia<sup>5,6</sup>, Jerry Kok Yen Chan<sup>15,16,17</sup>, Jacob George<sup>④</sup>, Pierce Kah-Hoe Chow<sup>⑧,18,19</sup>, Florent Ginhoux<sup>⑩,20,21</sup> & Ankur Sharma<sup>②,3,22,23</sup>

- POSTN+CAF participates in shaping the fetal-onco TME and impacts survival



# Presence of onco-fetal neighborhoods in hepatocellular carcinoma is associated with relapse and response to immunotherapy

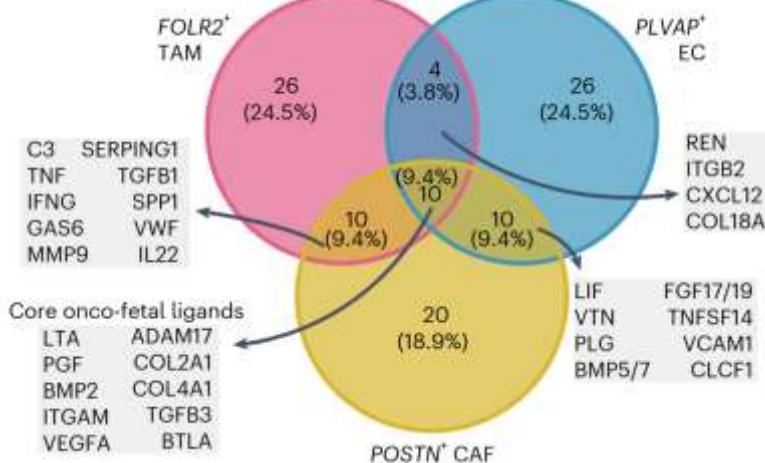
Received: 17 March 2023

Accepted: 16 October 2023

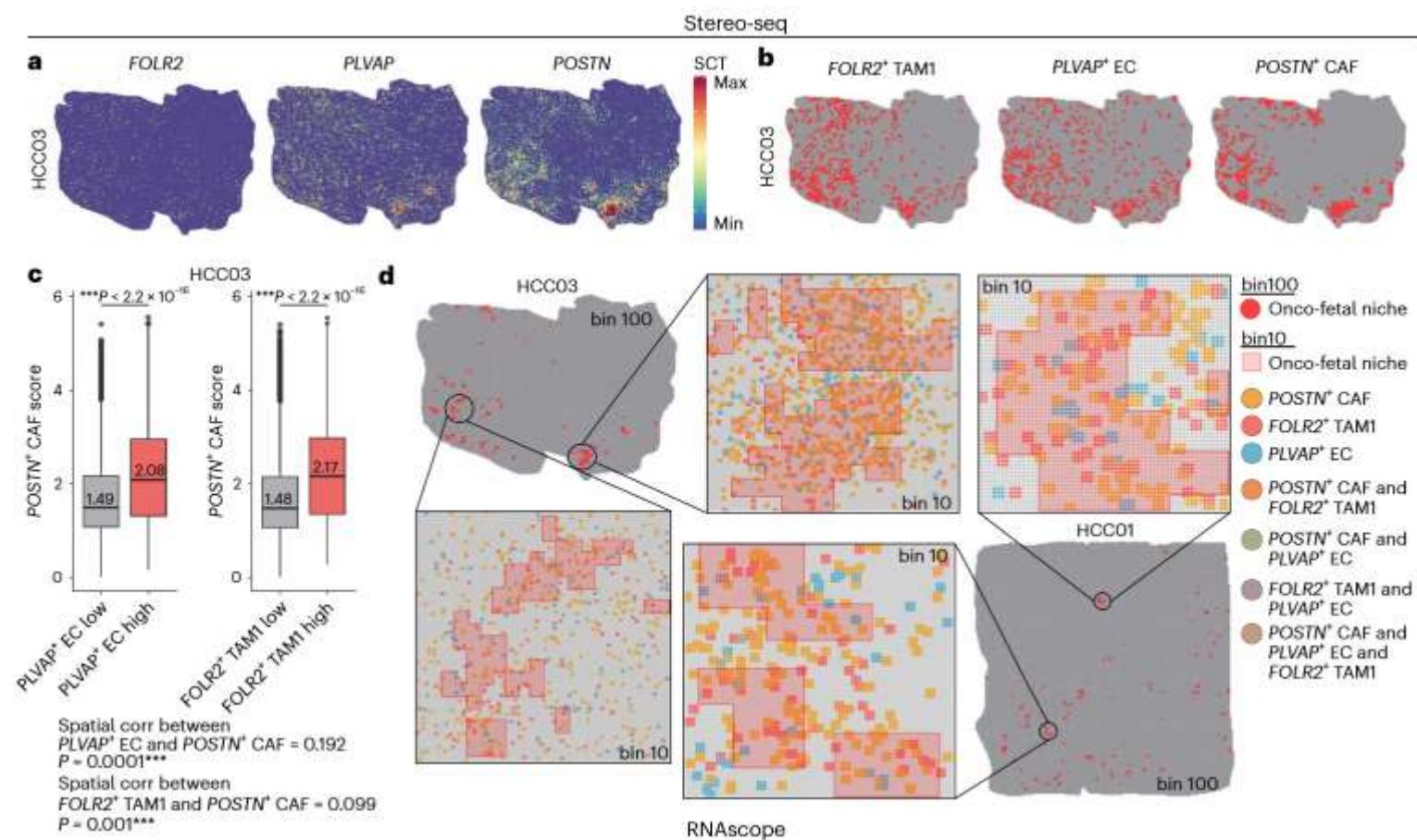
Published online: 02 January 2024

Check for updates

Ziyi Li <sup>1</sup>, Rhea Pai<sup>2,3,24</sup>, Saurabh Gupta<sup>2,3,24</sup>, Jennifer Currenti<sup>2,3,24</sup>, Wei Guo<sup>1,24</sup>, Anna Di Bartolomeo<sup>4</sup>, Hao Feng<sup>5,6</sup>, Zijie Zhang <sup>5</sup>, Zhizhen Li<sup>7</sup>, Longqi Liu <sup>8</sup>, Abhishek Singh<sup>2</sup>, Yingqi Bai<sup>8</sup>, Bicheng Yang<sup>9</sup>, Archita Mishra<sup>10,11</sup>, Katharine Yang<sup>10</sup>, Liang Qiao<sup>4</sup>, Michael Wallace<sup>12,13</sup>, Yujia Yin<sup>14</sup>, Qiang Xia<sup>5,6</sup>, Jerry Kok Yen Chan<sup>15,16,17</sup>, Jacob George <sup>4</sup>, Pierce Kah-Hoe Chow <sup>18,19</sup>, Florent Ginhoux <sup>10,20,21</sup> & Ankur Sharma <sup>2,3,22,23</sup>

**c**

- Discovering some onco-fetal ligands
- Spatial co-localization of three cell types



## Presence of onco-fetal neighborhoods in hepatocellular carcinoma is associated with relapse and response to immunotherapy

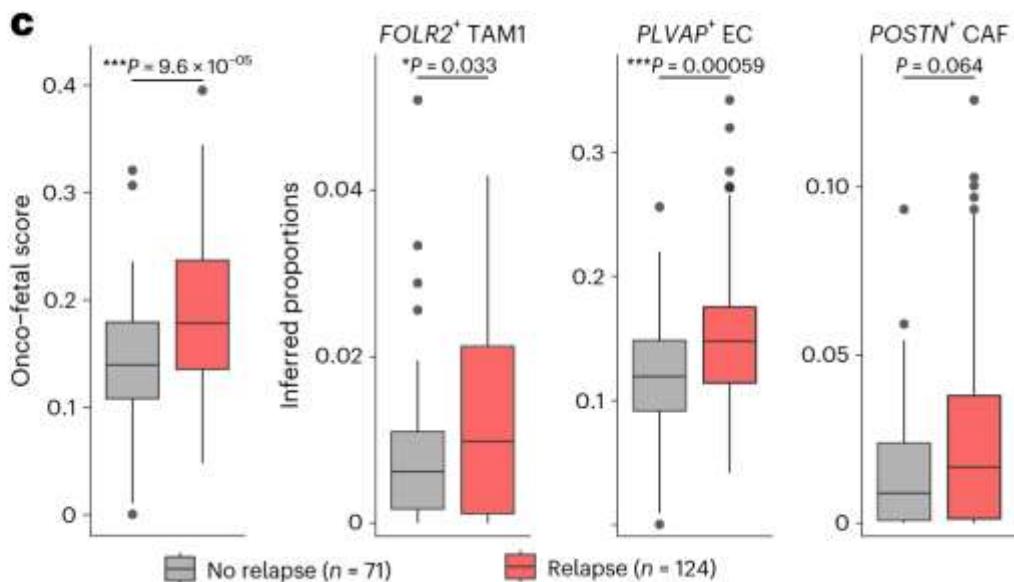
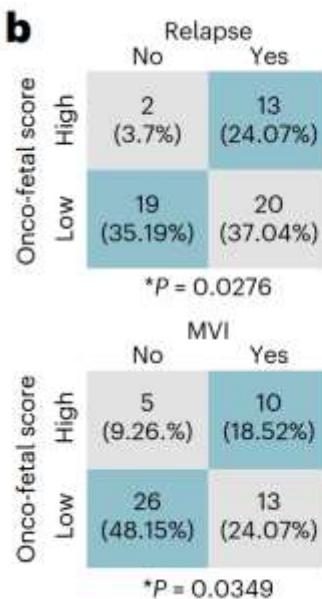
Received: 17 March 2023

Accepted: 16 October 2023

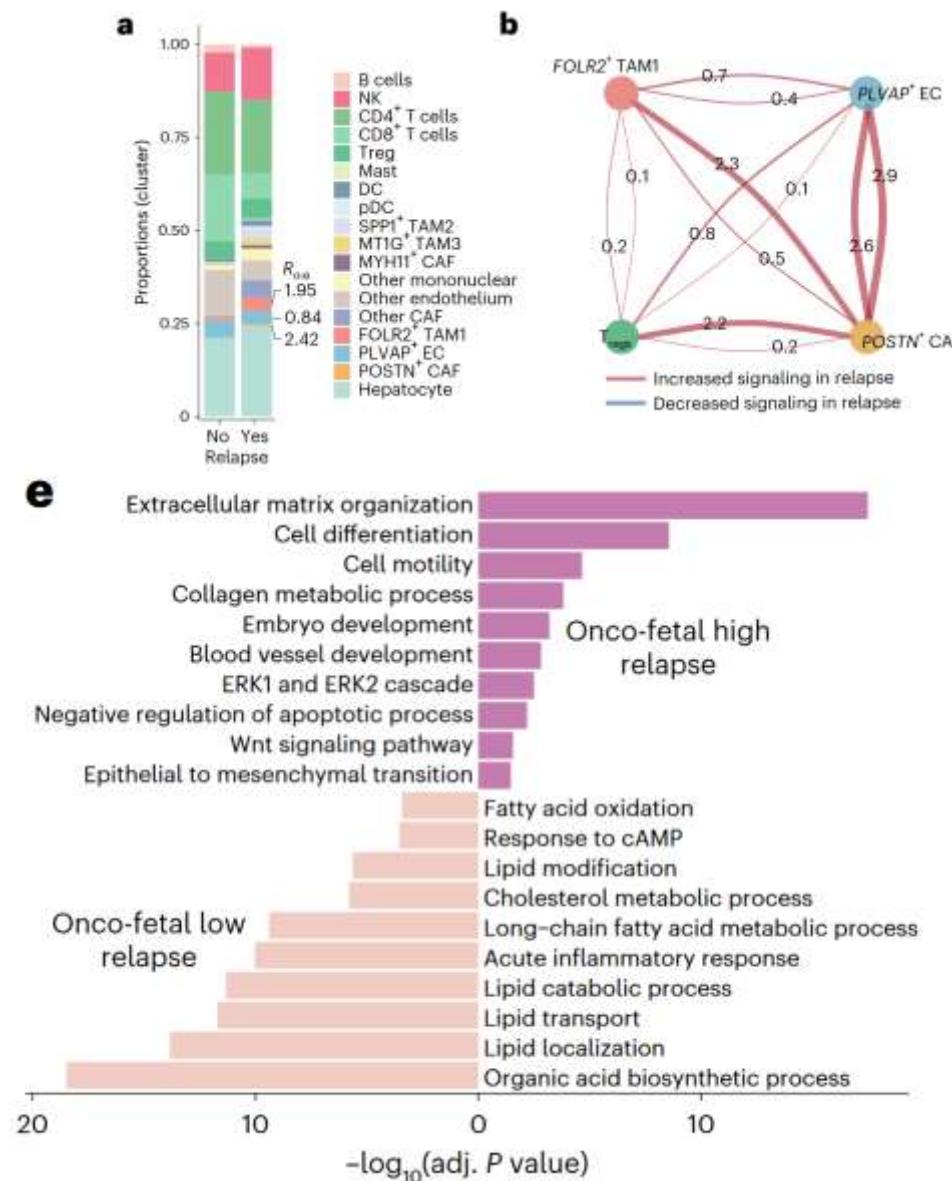
Published online: 02 January 2024

Check for updates

Ziyi Li<sup>1</sup>, Rhea Pai<sup>2,3,24</sup>, Saurabh Gupta<sup>2,3,24</sup>, Jennifer Currenti<sup>2,3,24</sup>, Wei Guo<sup>1,24</sup>, Anna Di Bartolomeo<sup>4</sup>, Hao Feng<sup>5,6</sup>, Zijie Zhang<sup>6,7</sup>, Zhizhen Li<sup>7</sup>, Longqi Liu<sup>8</sup>, Abhishek Singh<sup>2</sup>, Yinqi Bai<sup>8</sup>, Bicheng Yang<sup>9</sup>, Archita Mishra<sup>10,11</sup>, Katharine Yang<sup>10</sup>, Liang Qiao<sup>4</sup>, Michael Wallace<sup>12,13</sup>, Yujia Yin<sup>14</sup>, Qiang Xia<sup>5,6</sup>, Jerry Kok Yen Chan<sup>15,16,17</sup>, Jacob George<sup>18</sup>, Pierce Kah-Hoe Chow<sup>18,19</sup>, Florent Ginhoux<sup>1,20,21</sup> & Ankur Sharma<sup>2,3,22,23</sup>



- Oncofetal and relapse related
- Score high relapse with extracellular matrix tissue, embryonic development, cell proliferation, and cell motility related
- Score low relapse with lipid metabolism related



# Presence of onco-fetal neighborhoods in hepatocellular carcinoma is associated with relapse and response to immunotherapy

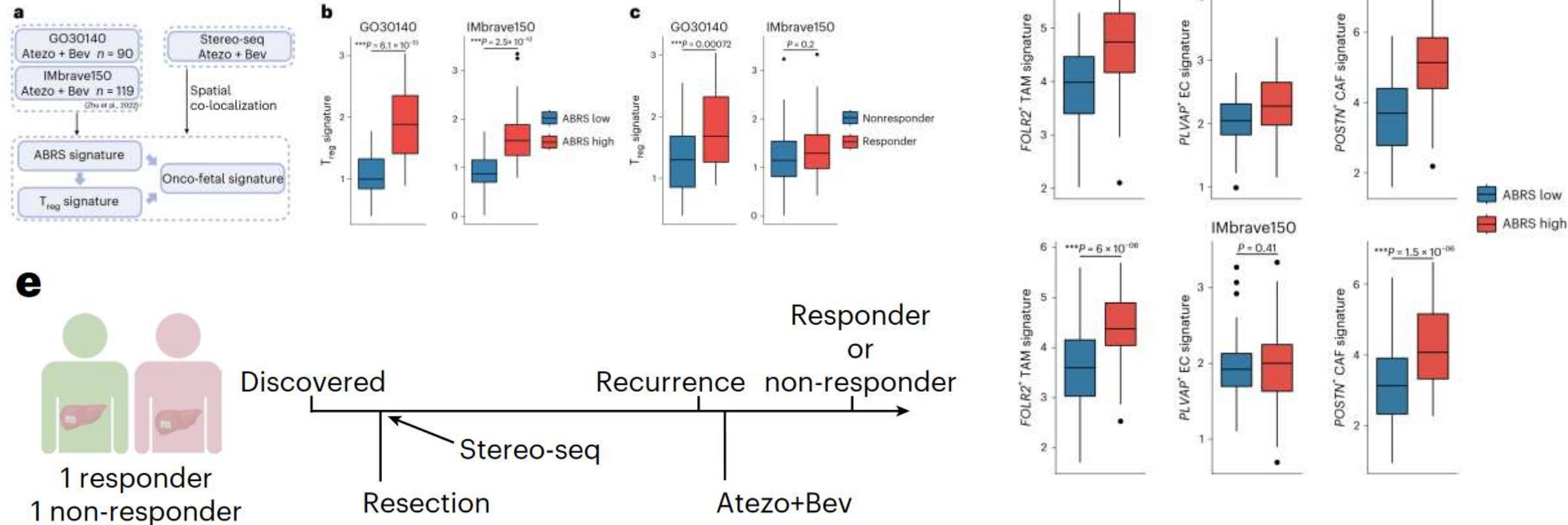
Received: 17 March 2023

Ziyi Li<sup>1</sup>, Rhea Pai<sup>2,3,24</sup>, Saurabh Gupta<sup>2,3,24</sup>, Jennifer Currenti<sup>2,3,24</sup>, Wei Guo<sup>1,24</sup>, Anna Di Bartolomeo<sup>4</sup>, Hao Feng<sup>5,6</sup>, Zijie Zhang<sup>6,7</sup>, Zhizhen Li<sup>7</sup>, Longqi Liu<sup>8</sup>, Abhishek Singh<sup>2</sup>, Yinqi Bai<sup>8</sup>, Bicheng Yang<sup>9</sup>, Archita Mishra<sup>10,11</sup>, Katharine Yang<sup>10</sup>, Liang Qiao<sup>4</sup>, Michael Wallace<sup>12,13</sup>, Yujia Yin<sup>14</sup>, Qiang Xia<sup>5,6</sup>, Jerry Kok Yen Chan<sup>15,16,17</sup>, Jacob George<sup>18</sup>, Pierce Kah-Hoe Chow<sup>18,19</sup>, Florent Ginhoux<sup>20,21</sup> & Ankur Sharma<sup>2,3,22,23</sup>

Accepted: 16 October 2023

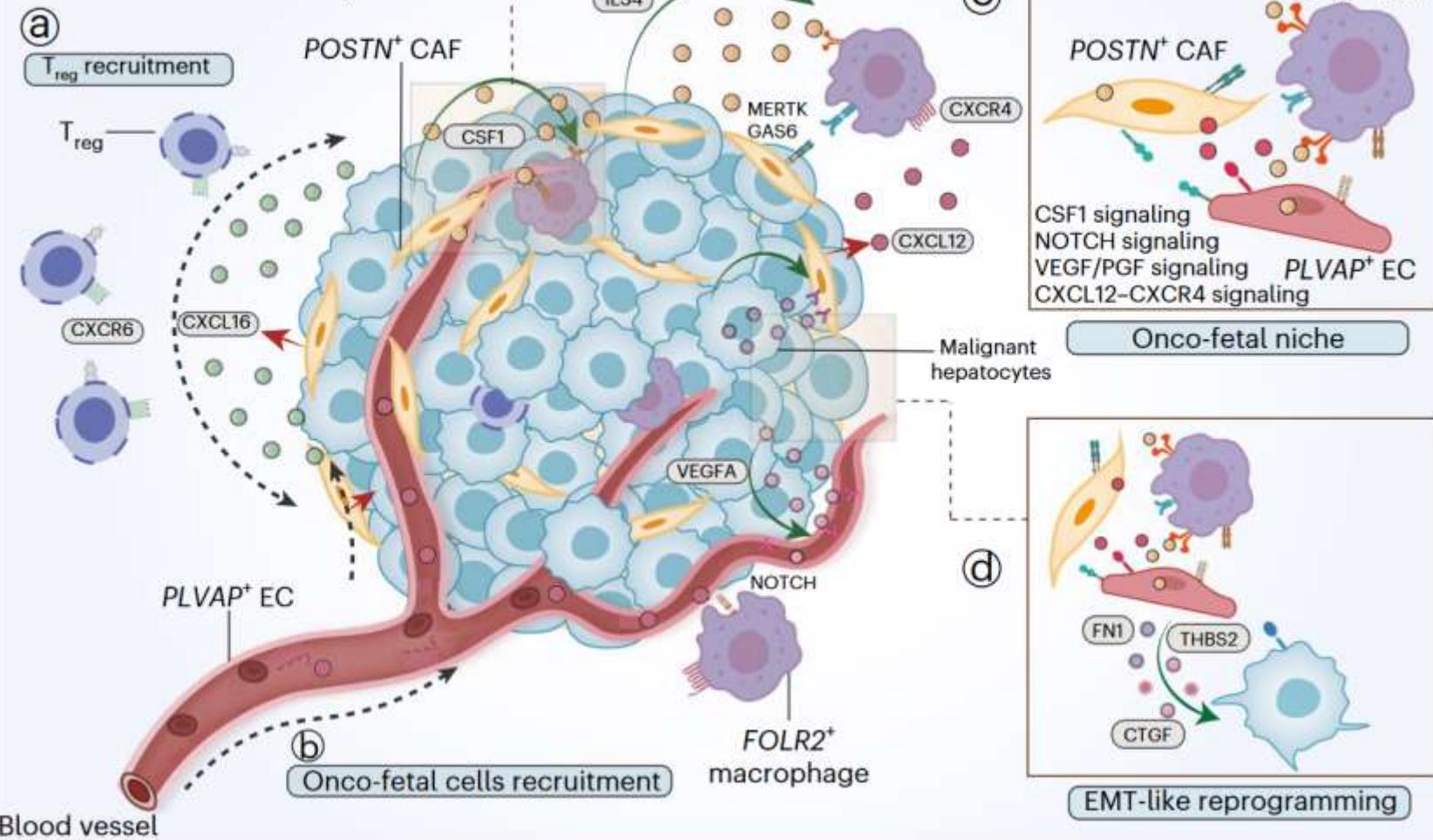
Published online: 02 January 2024

Check for updates



**g**

## Onco-fetal ecosystem



## Primary HCC

OF<sup>low</sup>

Non-EMT driven

CYP2↑  
G6PC↑  
ADH1B↑  
LECT2↑

OF<sup>hi</sup>

EMT-driven

CXCL1↑  
CXCL5↑  
KRT7↑  
KRT19↑

Early relapse

Response to combination therapy

[Back to Content](#)

# **Systematic comparison of IDH wild type GBM and neurodevelopmental trajectories**

# 目 录

## Content

1 背 景 介 绍

Background

2 课 题 当 前 进 展

Progress

3 后 续 研 究 计 划

Next Plan

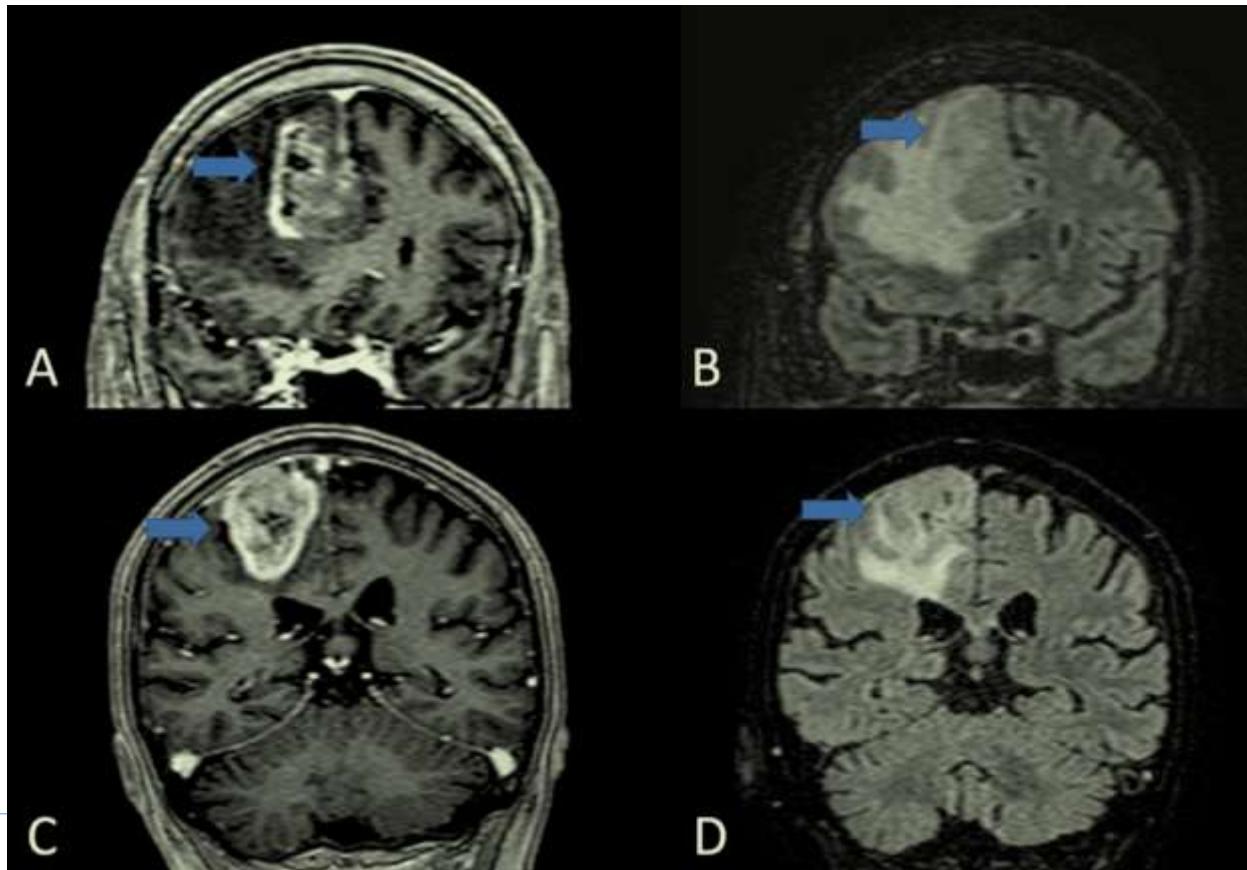




01

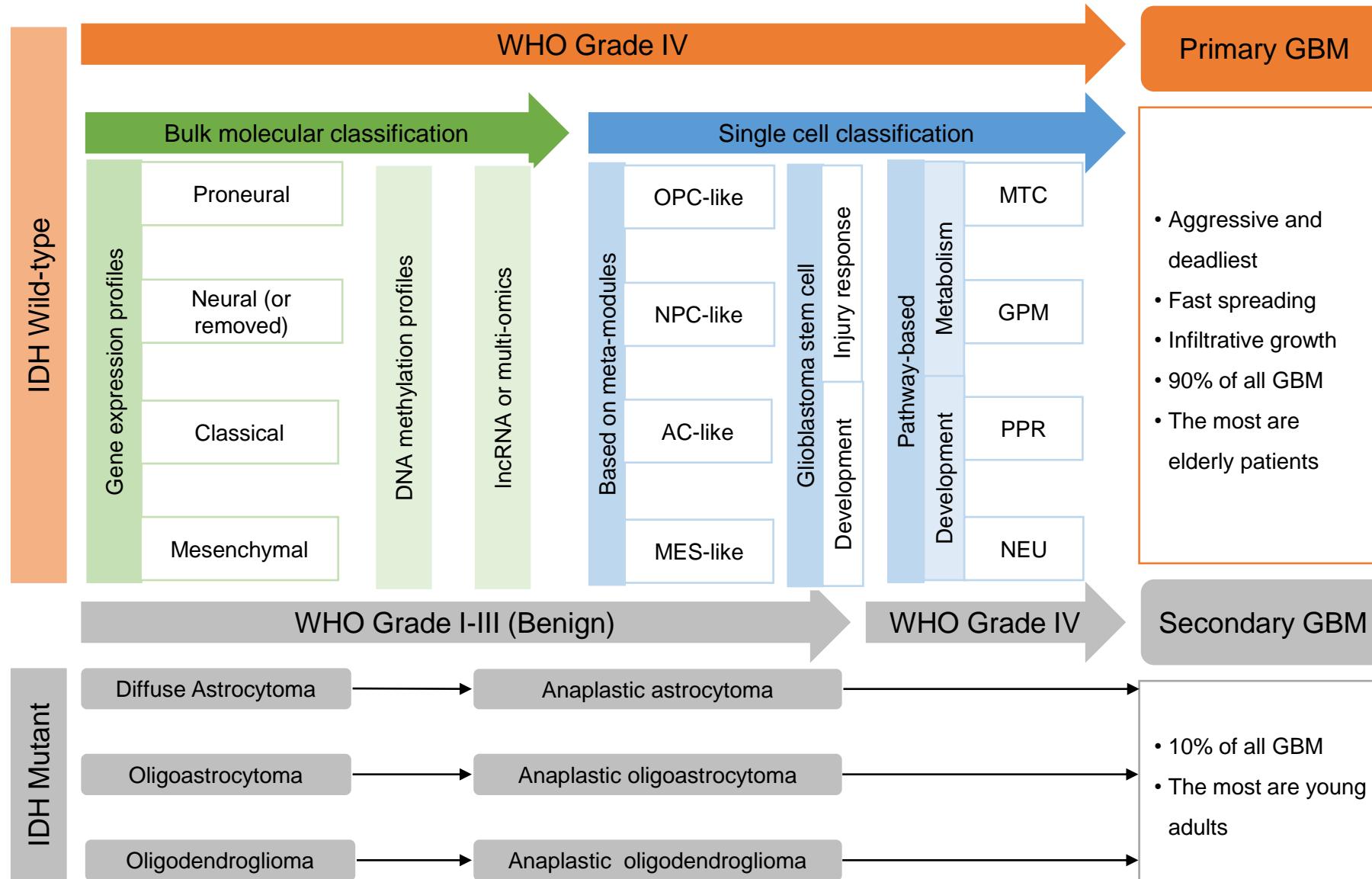
## Background





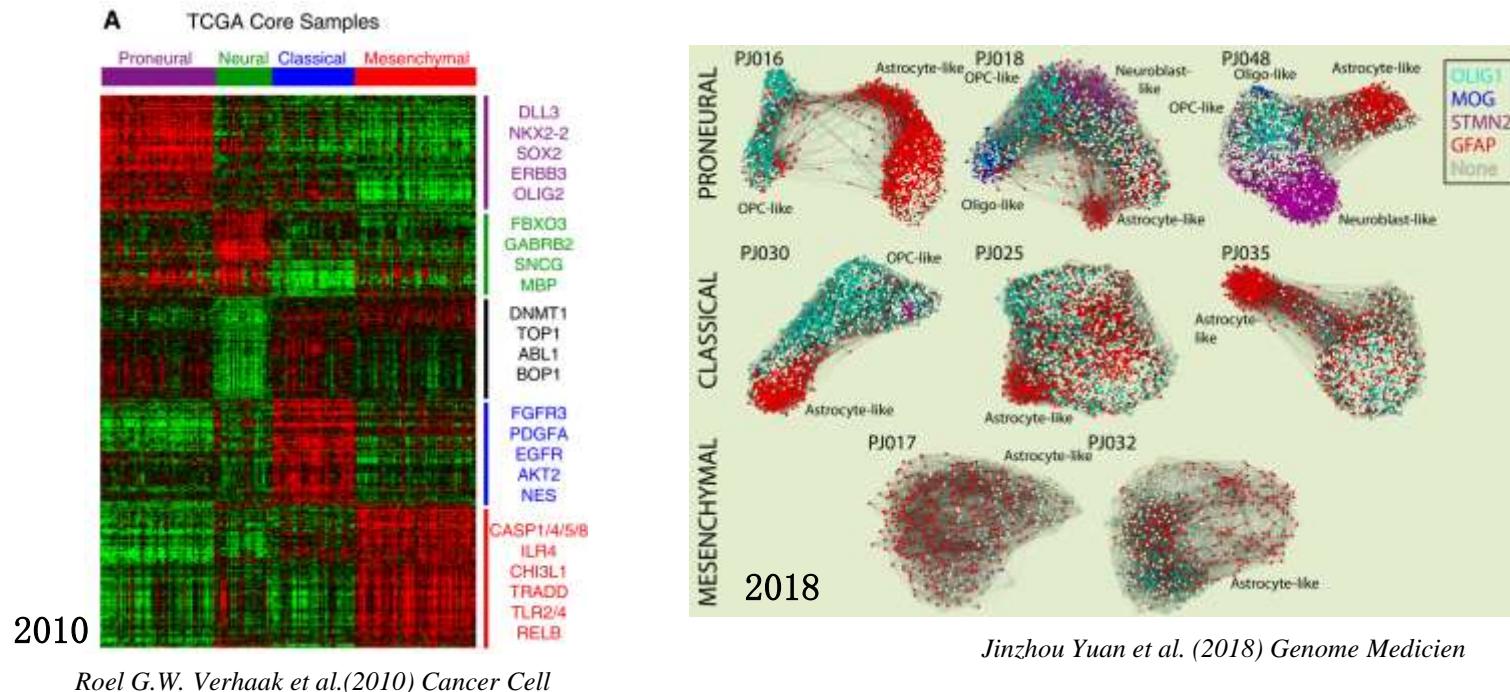
Ciro De Luca et al.(2022) Cells

- Glioblastoma (GBM) refers to the WHO Grade IV classification of gliomas
- Highly invasive, high mortality, rapid spread, infiltrative growth, and molecular and cellular heterogeneity.
- Prognosis is extremely poor, with clinical data indicating a less than 5% 5-year survival rate post-diagnosis and a median survival time of less than 15 months.
- Recurrence rate is high, with almost all patients experiencing recurrence after undergoing treatments such as surgical resection, radiotherapy, and chemotherapy with temozolomide (TMZ).



# PART 1 Background

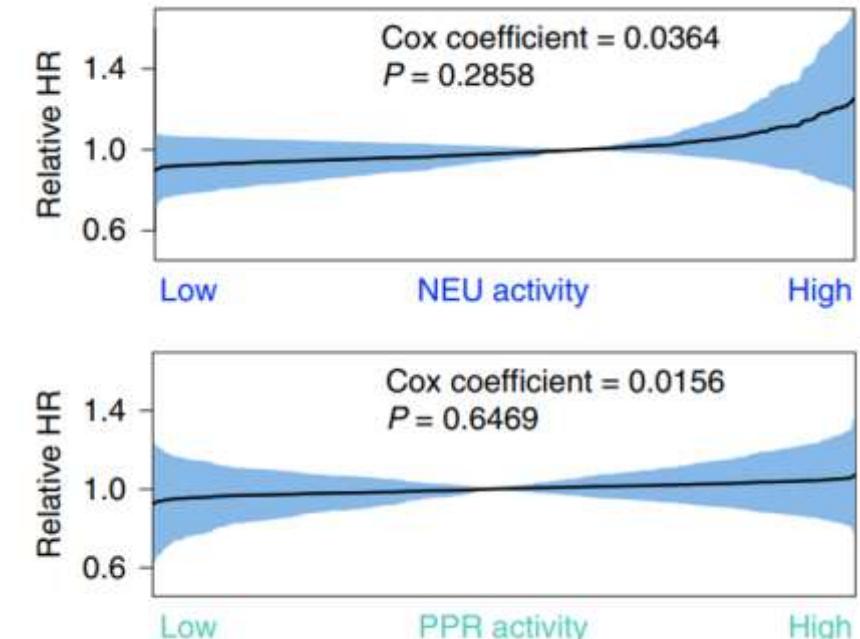
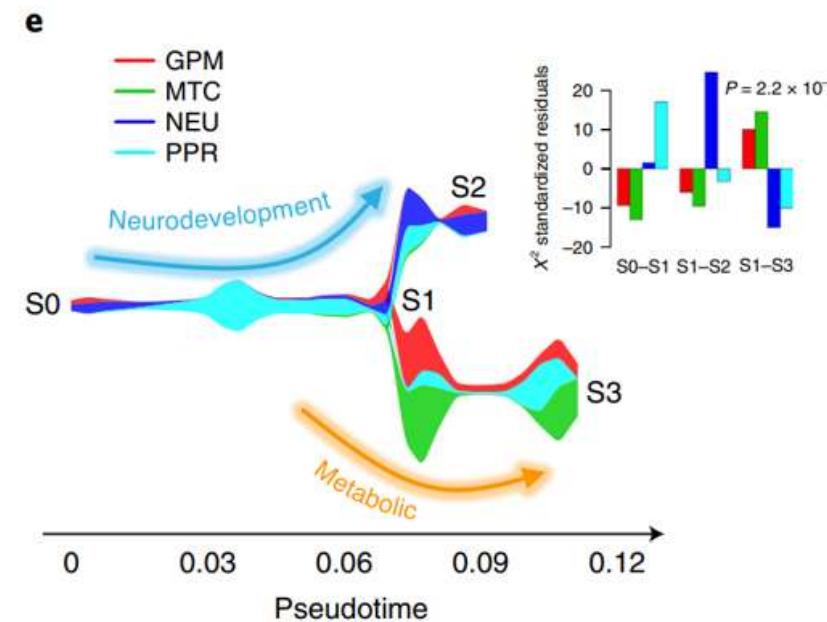
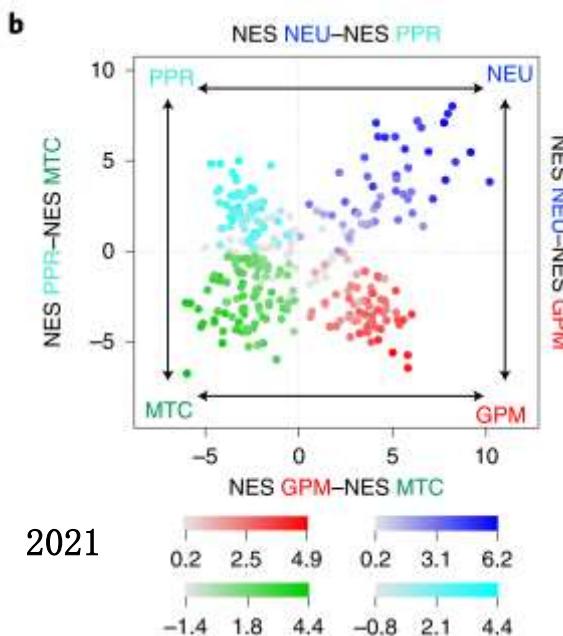
## Research on cellular subtypes still remains controversial



- Neural subtype, with its genetic characteristics still unclear, and some researchers believe this subtype may be contaminated samples of healthy brain tissue, suggesting not to consider this subtype

# PART 1 Background

Research on cellular subtypes still remains controversial



Luciano Garofano et al. (2021) *Nat. Cancer*

- Still, there are studies suggesting that gliomas have cellular subtypes resembling neural lineage differentiation with neuronal functionalities.
- Survival analysis results of developmental-related types from this study show no association with survival.
- However, the developmental lineage is a highly heterogeneous state
- thus it's possible that different states of subtypes obscure its relationship with survival.

# PART 1 Background

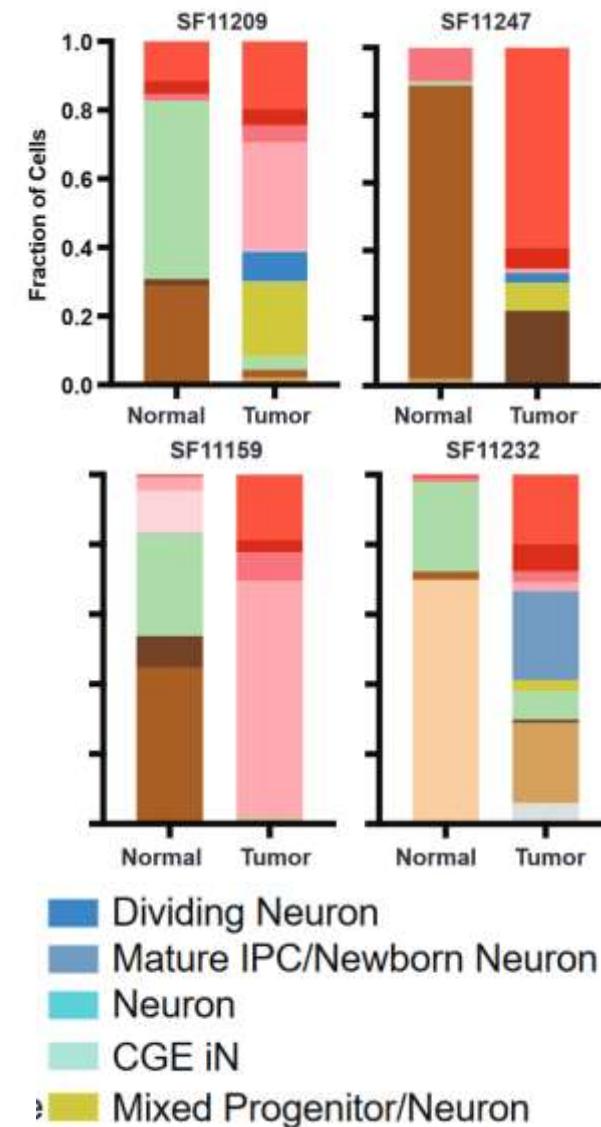
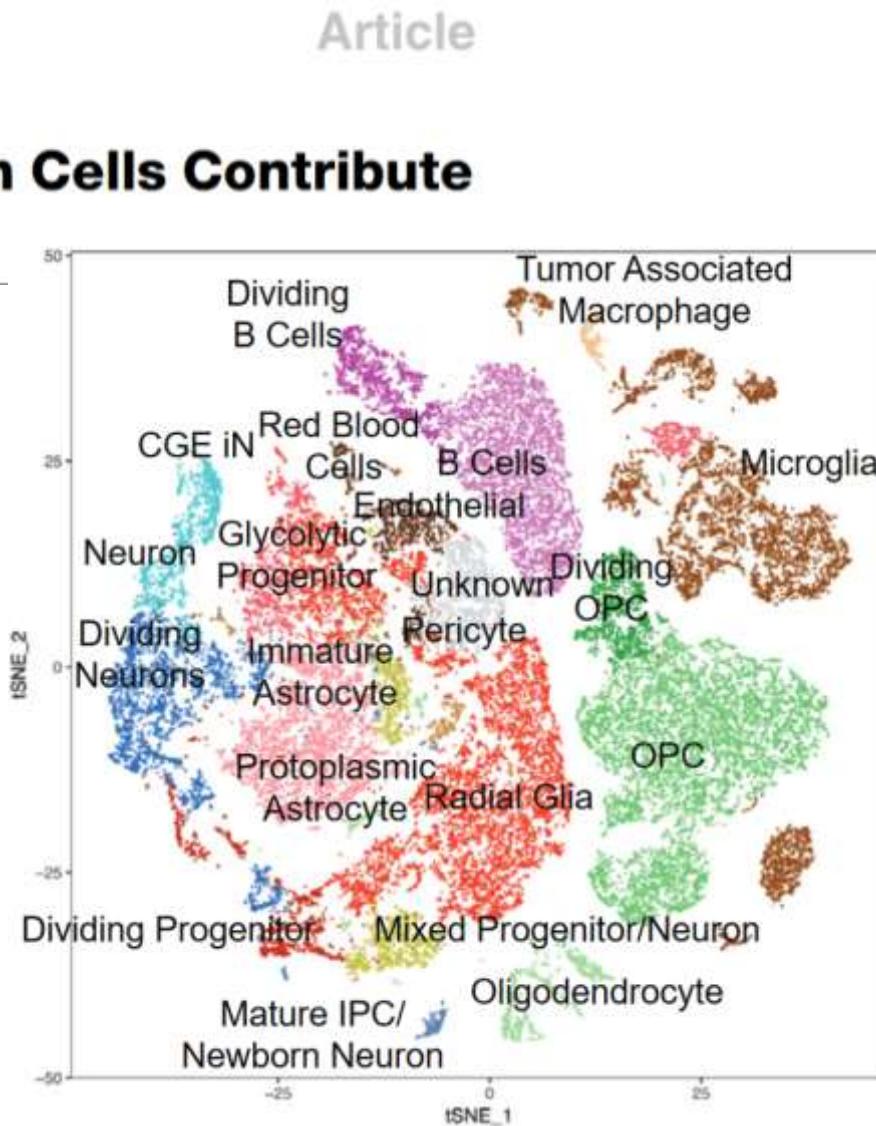
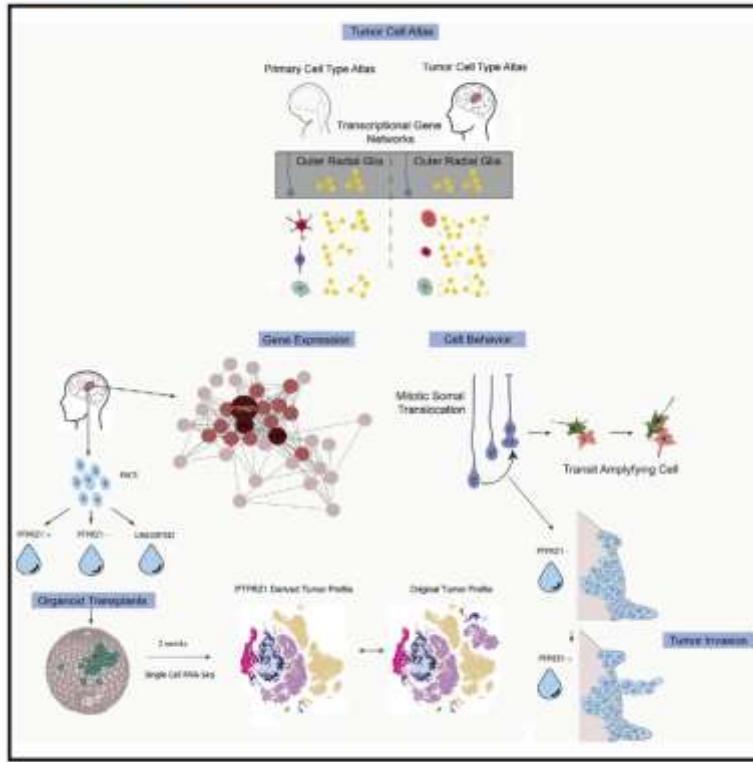
The heterogeneity and proliferation of GBM are similar to the development of nervous system



**Cell Stem Cell** January 2, 2020

## Outer Radial Glia-like Cancer Stem Cells Contribute to Heterogeneity of Glioblastoma

### Graphical Abstract



# PART 1 Background

The heterogeneity and proliferation of GBM are similar to the development of nervous system

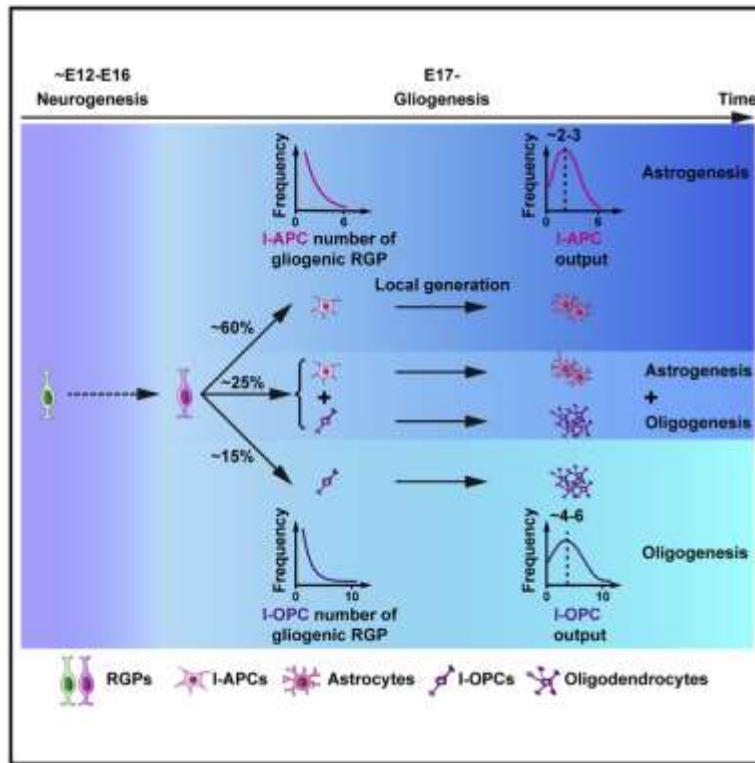


Cell Reports March 16, 2021

Article

## Distinct progenitor behavior underlying neocortical gliogenesis related to tumorigenesis

### Graphical Abstract



### Authors

Zhongfu Shen, Yang Lin, Jiajun Yang, ..., Jian Ma, Benjamin D. Simons, Song-Hai Shi

### Correspondence

bds10@cam.ac.uk (B.D.S.), shisonghai@mail.tsinghua.edu.cn (S.-H.S.)

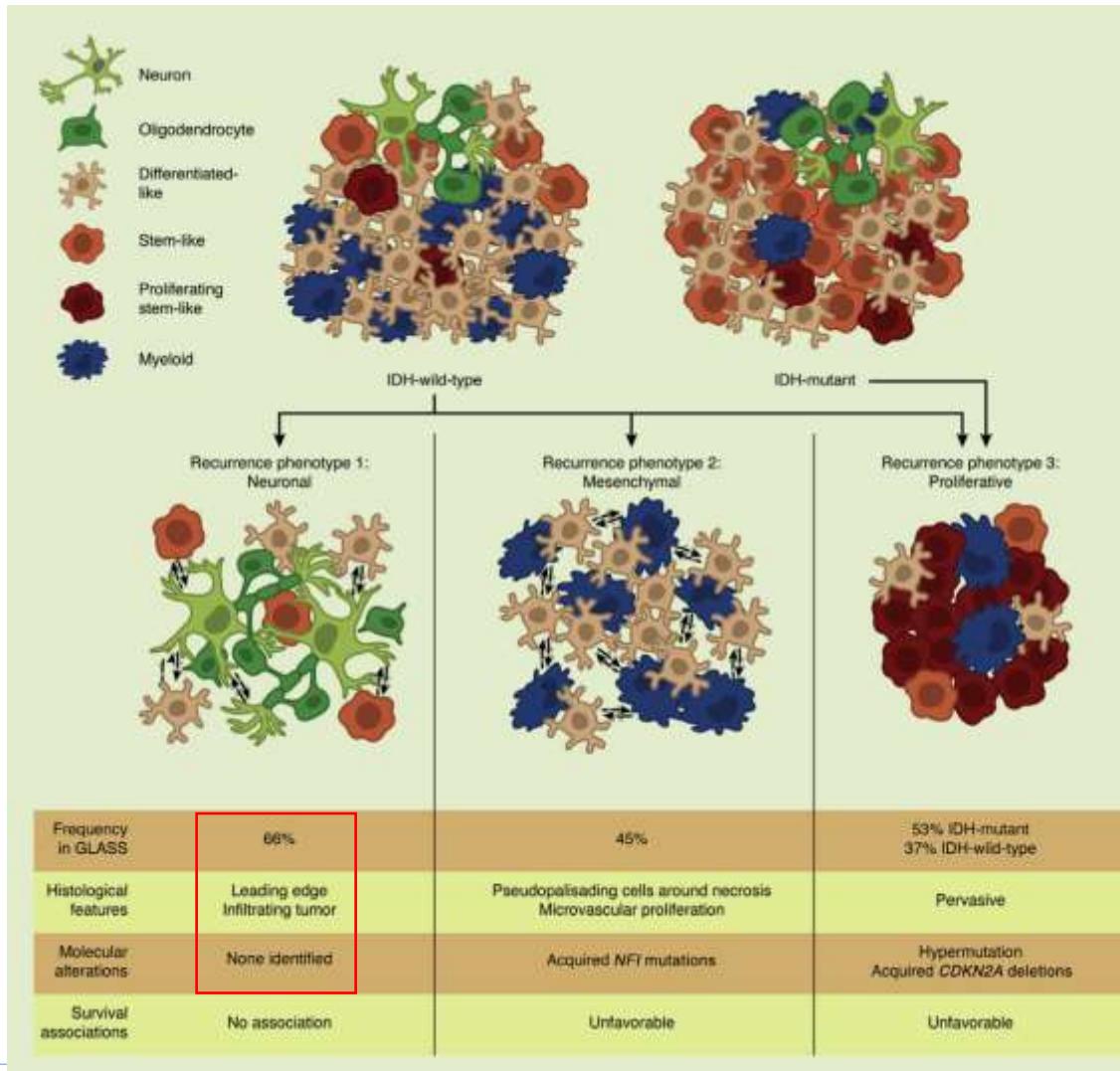
### In brief

Shen et al. delineate the precise radial glial progenitor cell behavior and the cellular program of gliogenesis at single-cell resolution in the developing neocortex and suggest a proclivity of oligogenesis to brain tumor formation.

- The brain development exhibits a strict cellular organizational structure, with tight connections between cells in brain tissue.
- Glial cells and neurons in the brain share a common origin.
- Malignant glioma cells express numerous neural stem cell and progenitor cell markers.
- Overexpression of stem cell genes in neural cell populations leads to extensive proliferation of glial cells.

## PART 1 Background

### Studying the differentiation of GBCs from a developmental perspective

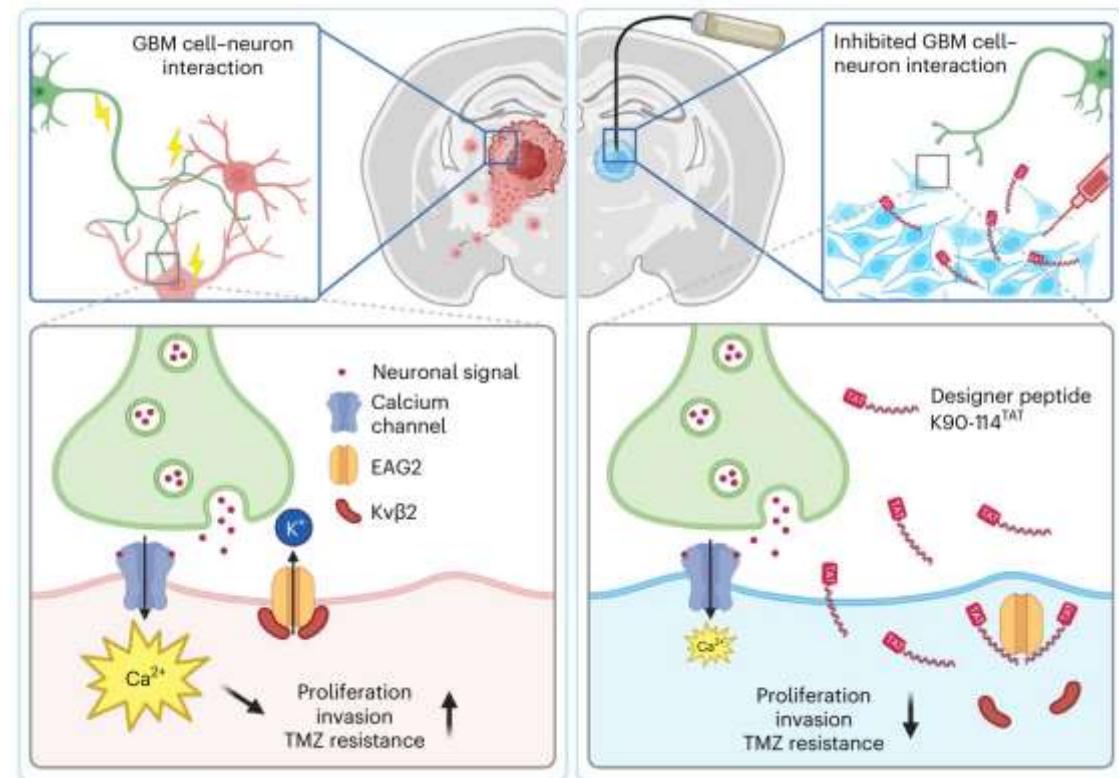
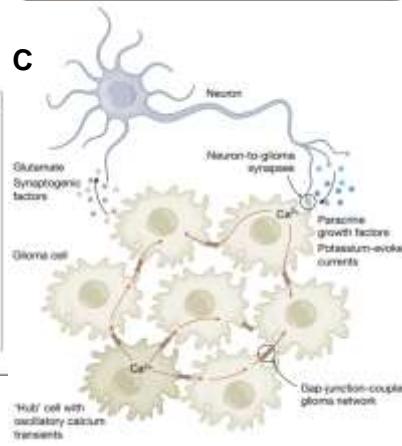
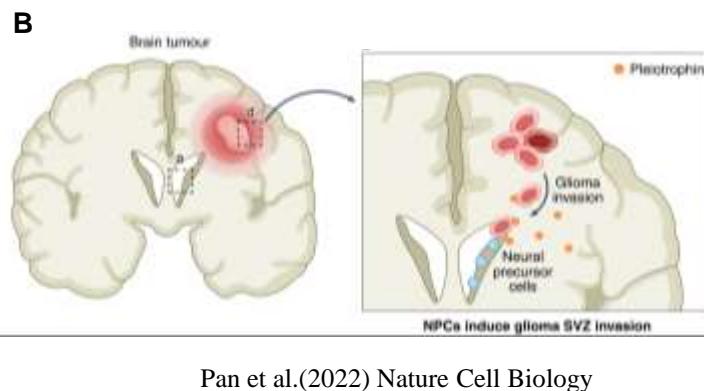
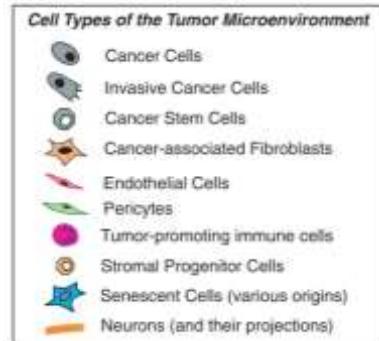
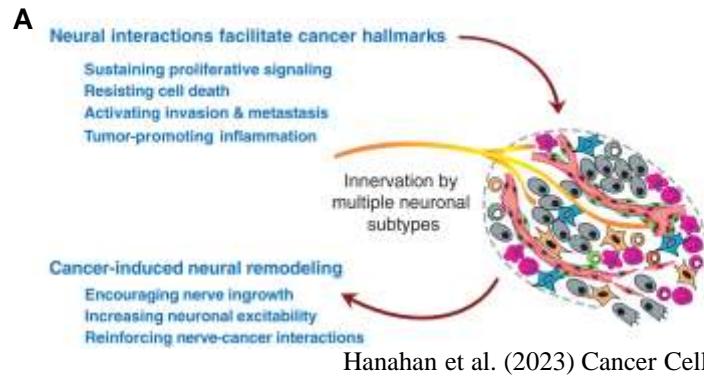


Frederick S. Varn et al. (2022) Cell

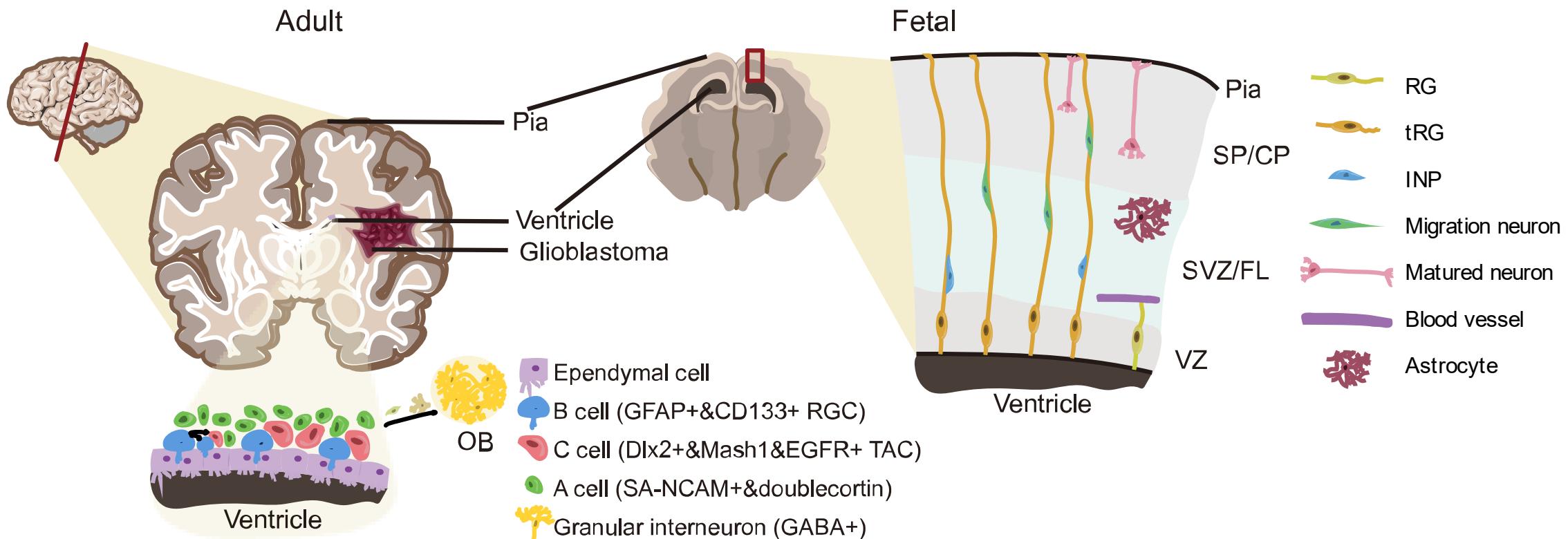
- Research on primary and recurrent samples of GBM indicates that wild-type IDH recurrence is mainly comprised of three types: Neuronal, Mesenchymal, and Proliferative.
- Neuronal type constitutes 66% of wtIDH GBM, presenting histological features of infiltrative tumors.
- The molecular markers responsible for recurrent Neuronal type are still unclear.
- Subtypes resembling neurons exist in primary GBM, or there may be neural stem cell-like GBCs capable of differentiating into Neuronal type. Inducing differentiation towards neurons could potentially enhance patient survival or reduce heterogeneity, thus improving treatment efficacy.

# PART 1 Background

## Functional crosstalk between tumors and nerves



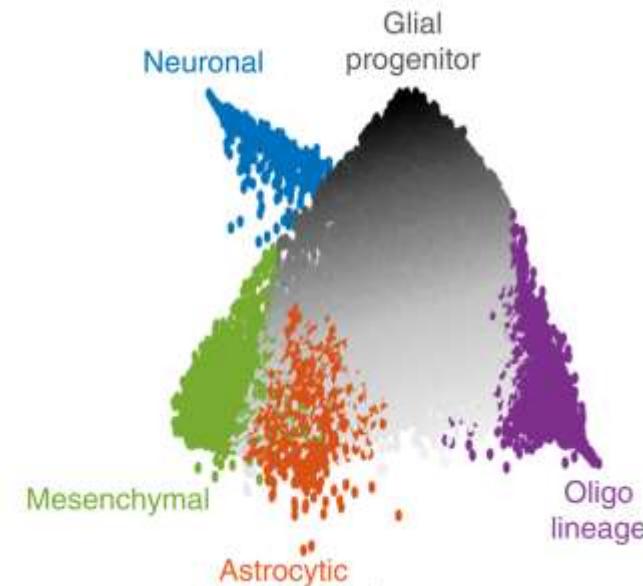
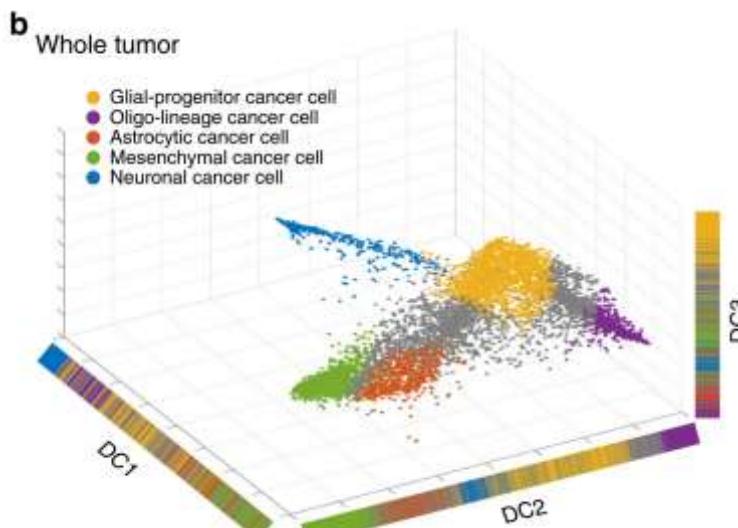
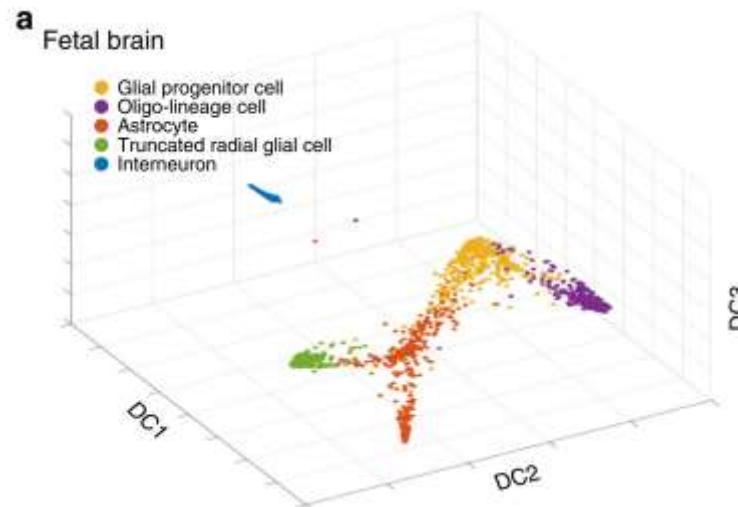
# PART 1 Background



- The molecular subtyping of GBM remains controversial and warrants further investigation.
- Studies on developmental lineage in GBM have potential therapeutic value.
- The ability of GBM to form functional connections with neurons suggests the possibility of its differentiation into neuron-like cells, integrating them into neural circuits to promote its survival.
- The accumulation of current single-cell data, along with the development of bioinformatics analysis tools and methods, provides the groundwork for systematic comparisons between neurodevelopment and GBM lineage development.

# PART 1 Background

## Comparative studies on neurodevelopment and GBM

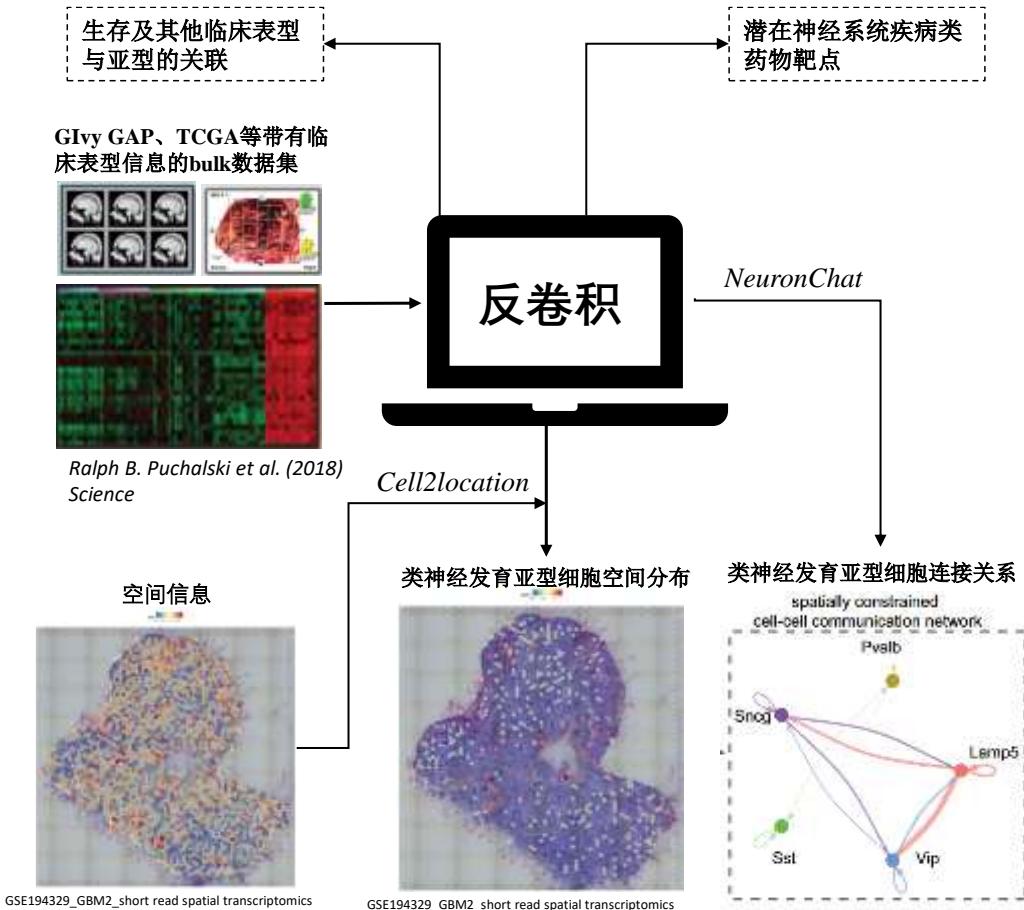
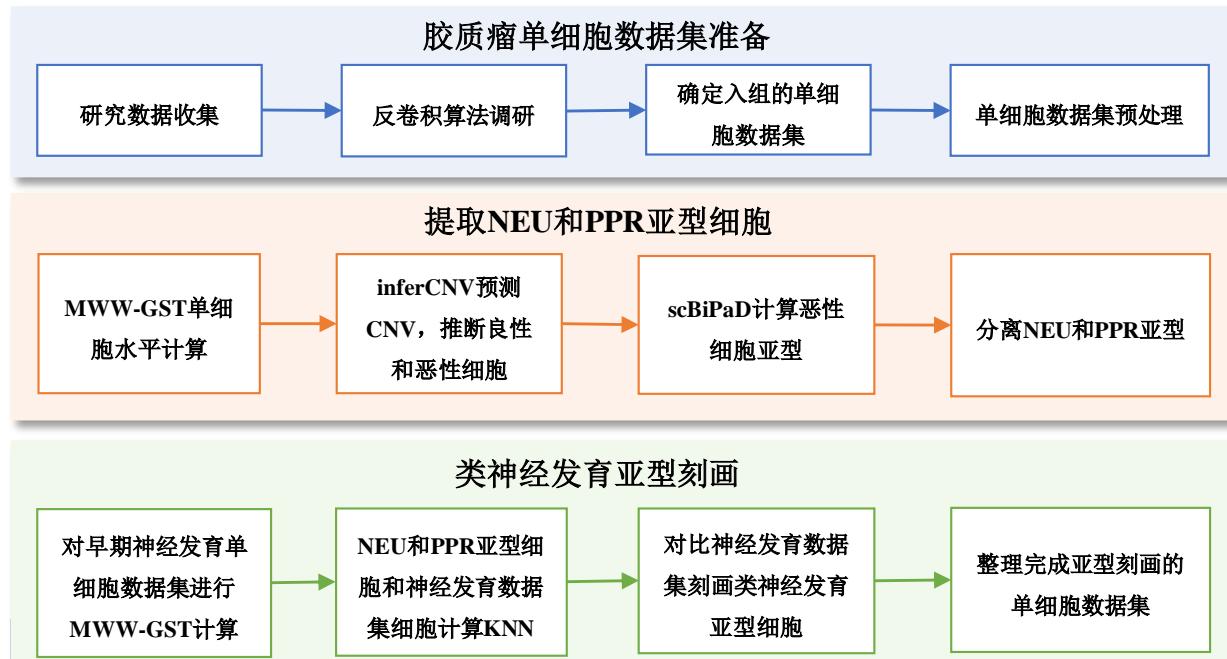


- Comparing single-cell data of neurodevelopment and glioblastoma, a differentiation model with a progenitor cell at the top was constructed.
- The main purpose was to study glioblastoma stem cells, identifying surface markers in glial progenitor cells
- Subpopulations of cells selected based on these markers were transplanted into mice for phenotypic studies.
- However, comprehensive comparison of developmental trajectories and investigation of the correlation between survival and other clinical information were not conducted.

Charles P. Couturier et al. (2022) Nat. Communication

# PART 1 Background

Adjust research techniques according to new relevant research information





02

## Progress



## PART 2 Progress

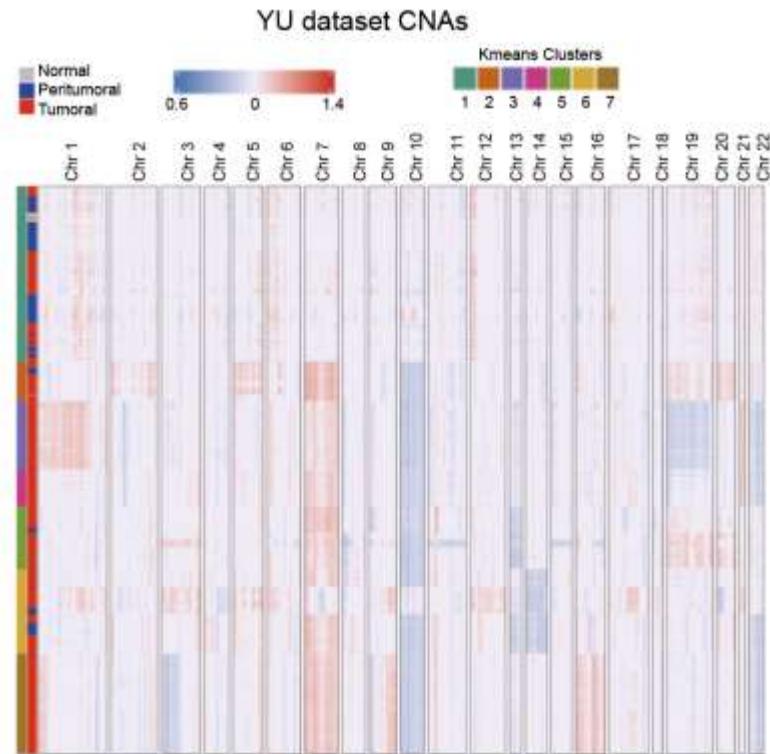
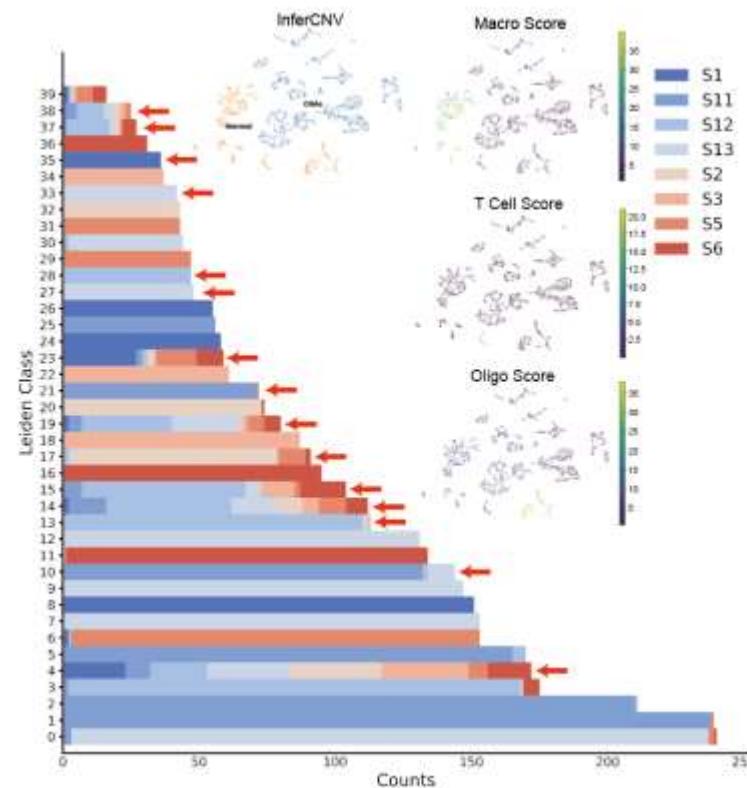
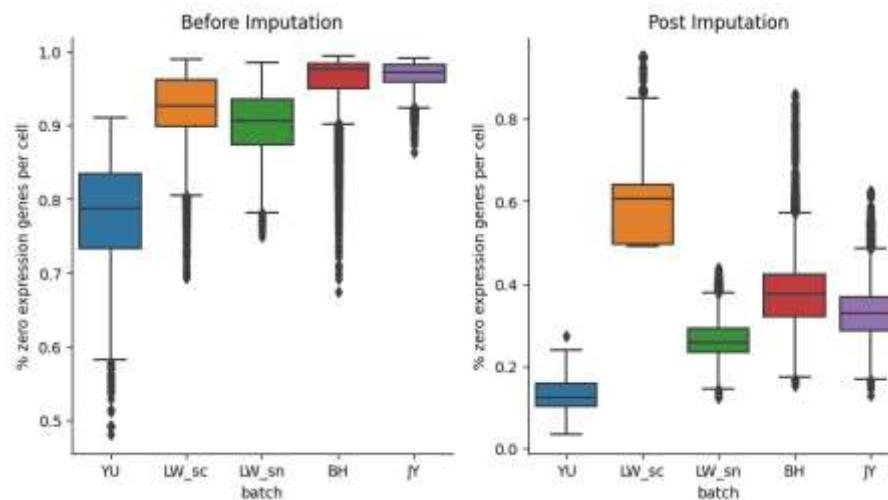
### Quality control and preprocessing of 5 GBM single-cell datasets



Datasets ID	Description	Raw cell /nucleus number	Filtered cell/nucleus number
GSE117891	Preprocessing of the raw expression matrix of single-cell nuclear RNA sequencing (snRNA-seq) captured at 10X, excluding GS4 (Gliosarcoma), GS15 (Metastases), and GBM samples of other grades except WHO IV	4193	3823
GSE138794	Single-cell and single-nucleus data captured at 10X, excluding samples with a particularly low number of captured genes (<1000), and samples with fewer than 100 cells/nuclei after quality control	28146	20084
PRJNA579593	The single-cell nuclear data captured by the 10X and C1 platforms are not included in this study analysis due to the insufficient number of single-sample nuclei from the C1 platform, which is less than 100. Only the IDH wild-type WHO grade IV samples from the 10X data are included in the study	44932	29215
GSE103224	The single-cell data captured by the automated microwell array-based platform only includes IDH wild-type WHO grade IV samples.	12466	7994
GSE131928	Smart-seq2 and 10X capture CD45-negative non-immune cells in IDH wild-type grade IV primary glioma samples. Samples with fewer than 100 cells after MGH101 quality control were excluded, while others were included in the study.	21943	18922

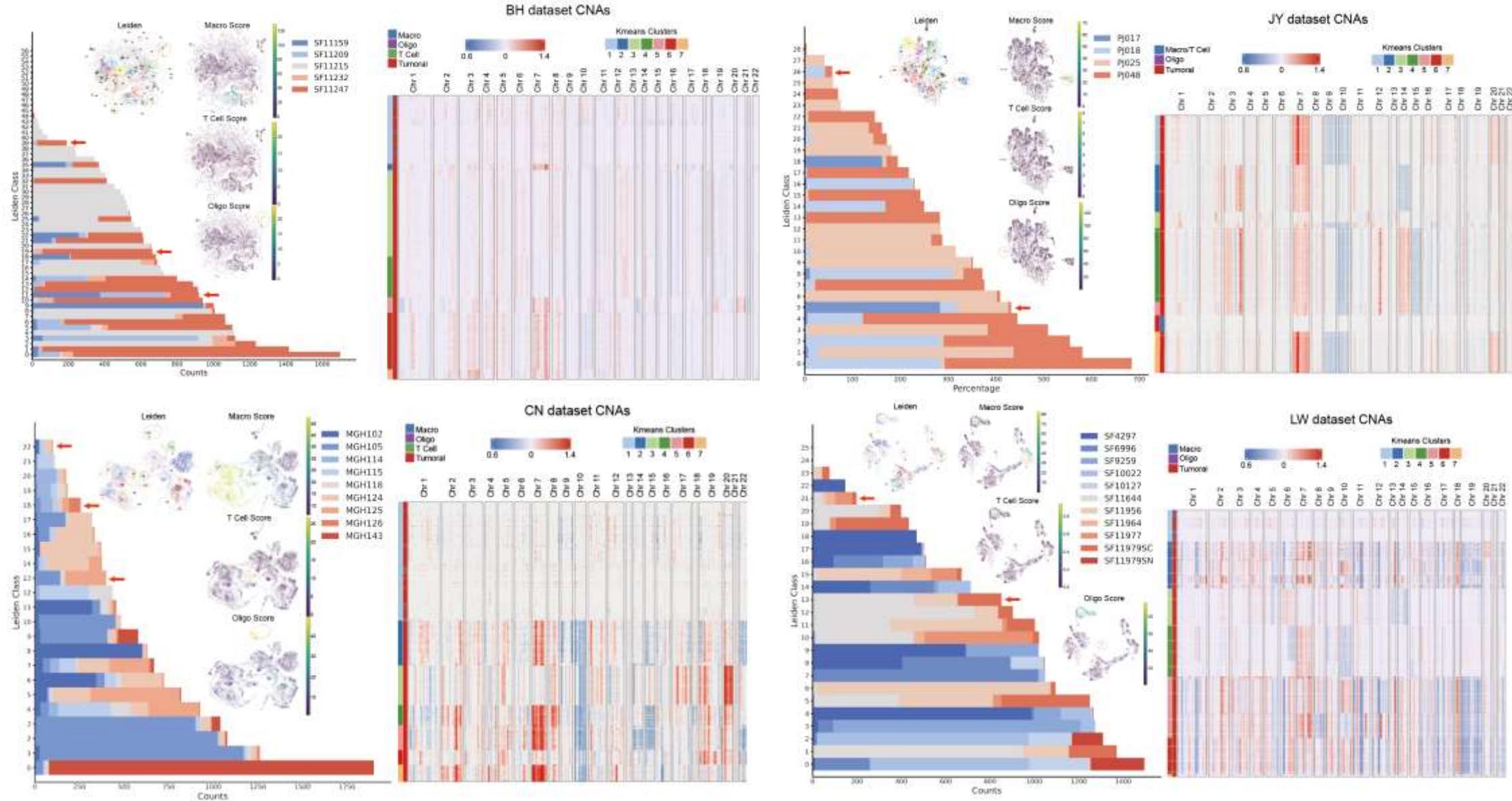
## PART 2 Progress

### Defining malignant cells with CNAs and normal cells without obvious CNAs

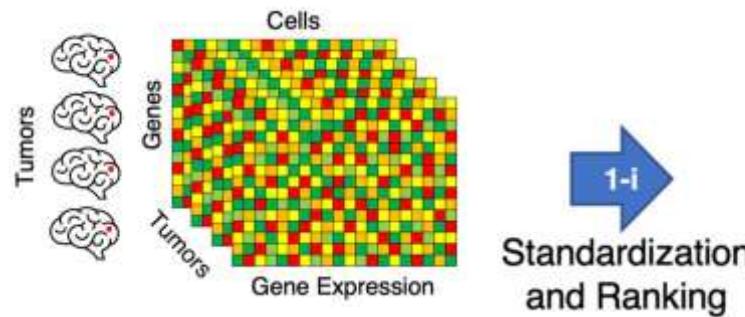


## PART 2 Progress

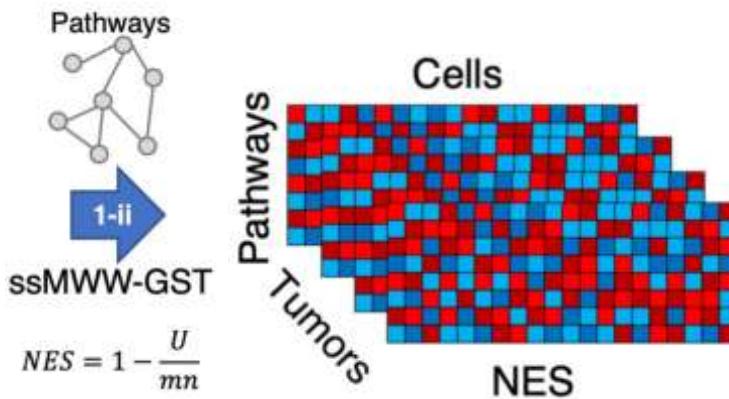
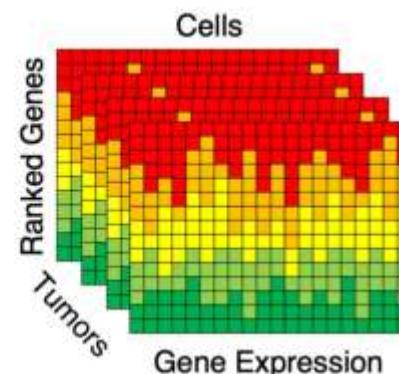
## Defining malignant cells with CNAs and normal cells without obvious CNAs



## PART 2 Progress



1-i  
Standardization and Ranking

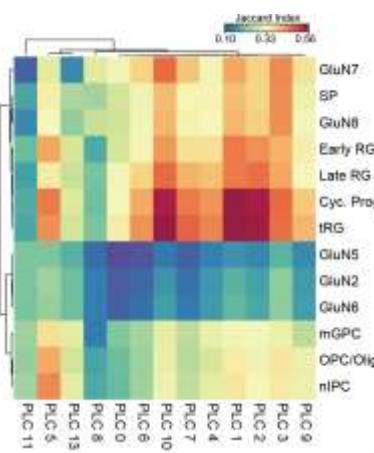


$$NES = 1 - \frac{U}{mn}$$

1-ii  
ssMWW-GST

Leiden  
Clustering

$$1 - \frac{p_i \cap p_j}{p_i \cup p_j}$$



Perform ssMWW-GST pathway calculations on individual cells for GBM and fetal brain data



Expression GBM data UMAP



- PJ017
- SF9259
- PJ018
- SF10022
- PJ025
- SF10127
- PJ048
- SF11159
- S1
- SF11209
- S2
- SF11215
- S3
- SF11232
- S5
- SF11247
- S6
- SF11644
- S11
- SF11956
- S12
- SF11964
- S13
- SF11979SC
- SF4297
- SF6996
- SF11979SN

Pathway NES GBM data UMAP



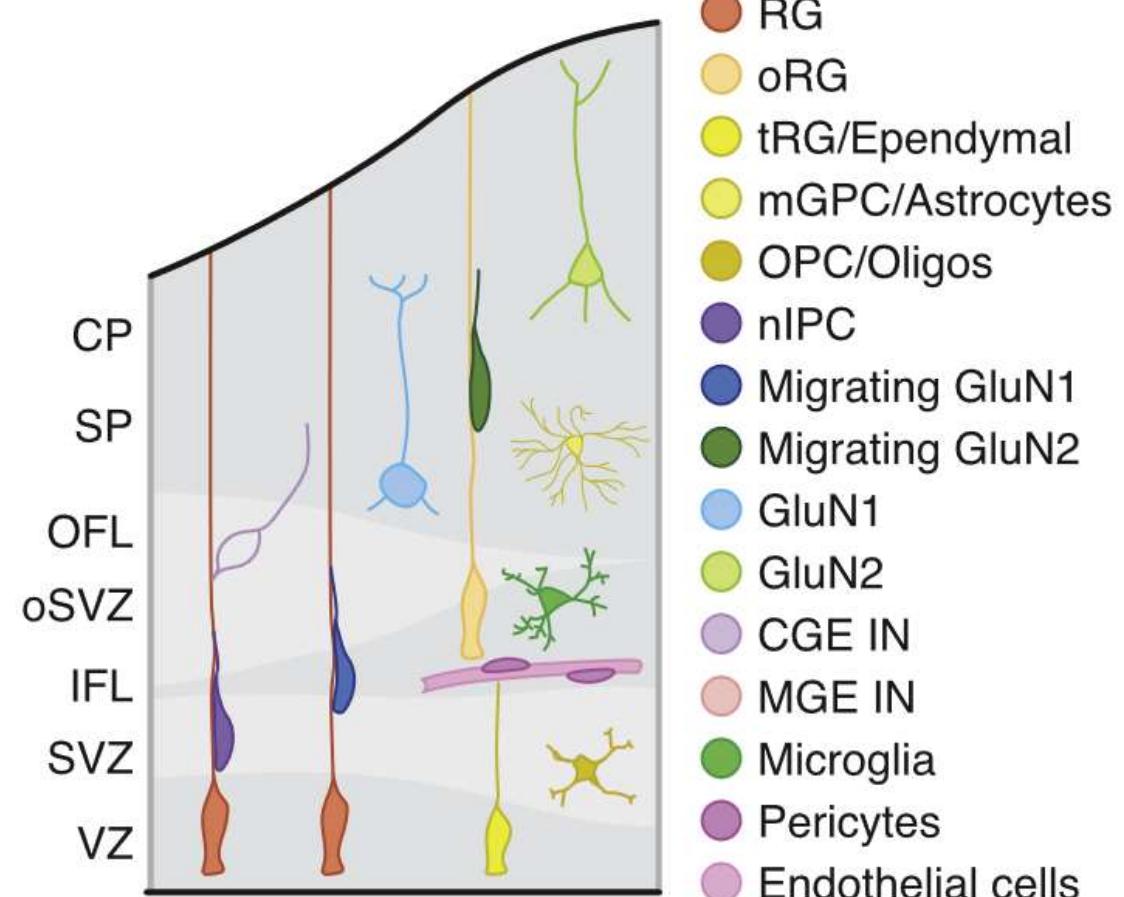
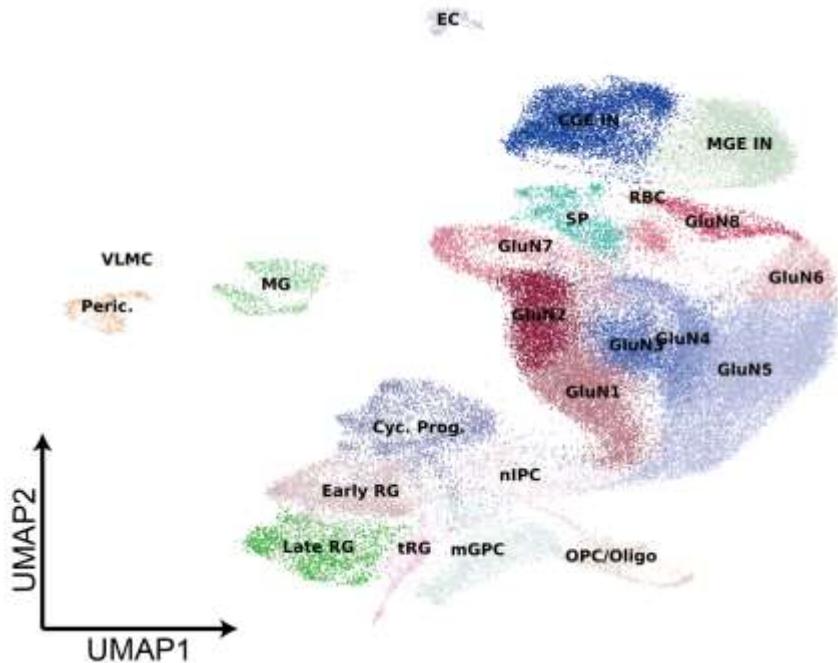
- MGH102\_10X
- S1
- MGH105\_10X
- S2
- MGH114\_10X
- S3
- MGH115\_10X
- S5
- MGH118\_10X
- S6
- MGH124\_10X
- S11
- MGH125\_10X
- S12
- MGH126\_10X
- S13
- MGH143\_10X
- SF10127
- SF11159
- SF11209
- SF11215
- SF11232
- SF11247
- SF11644
- SF11956
- SF11964
- SF11979SC
- SF11979SN
- PJ017
- SF6996
- SF9259
- PJ025
- SF10022

## PART 2 Progress

Neuronal development primarily concerns cell types and their distribution



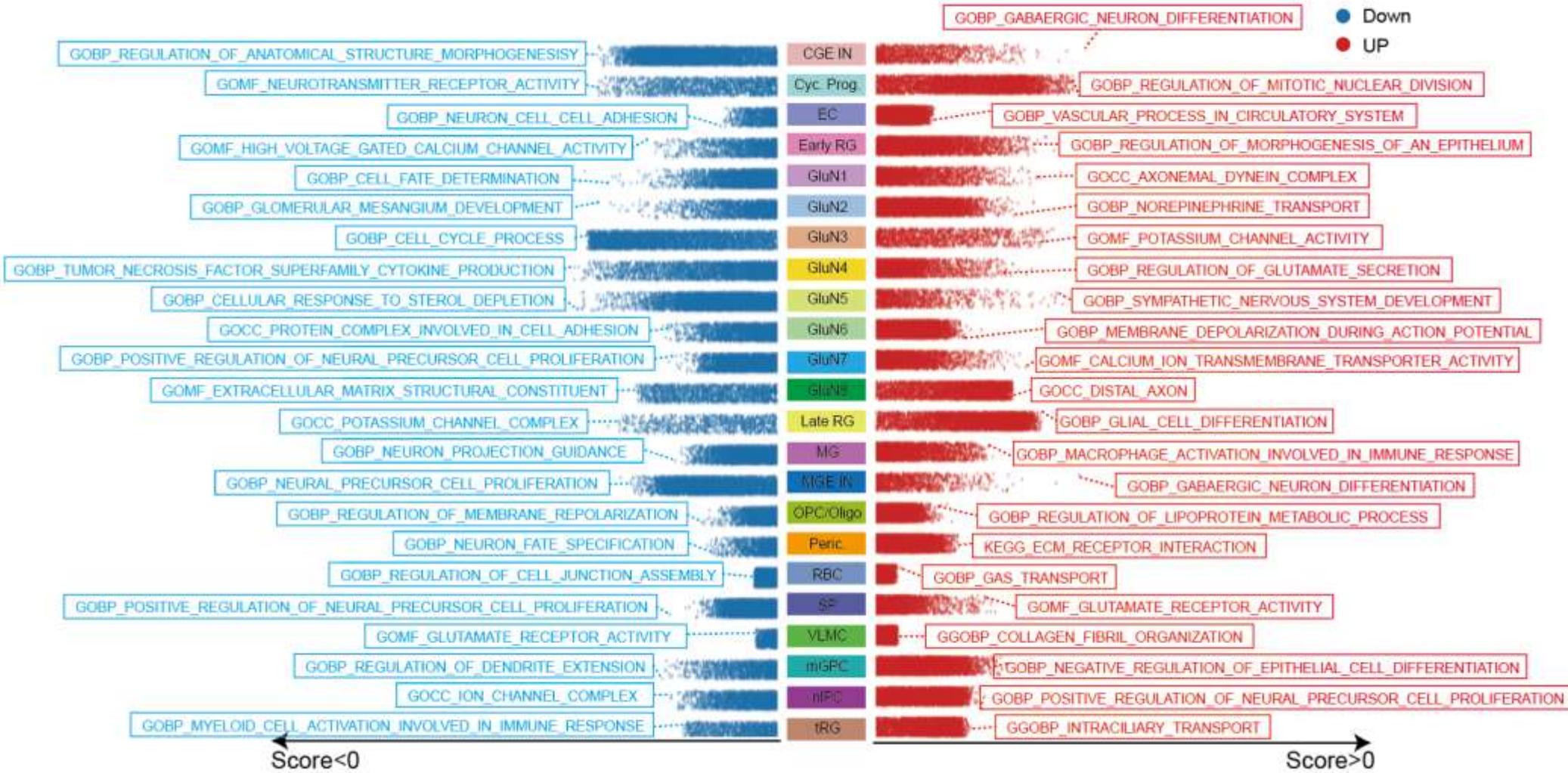
Trevino et al. (2021) fetal brain GW16-24



Trevino et al. (2021) Cell

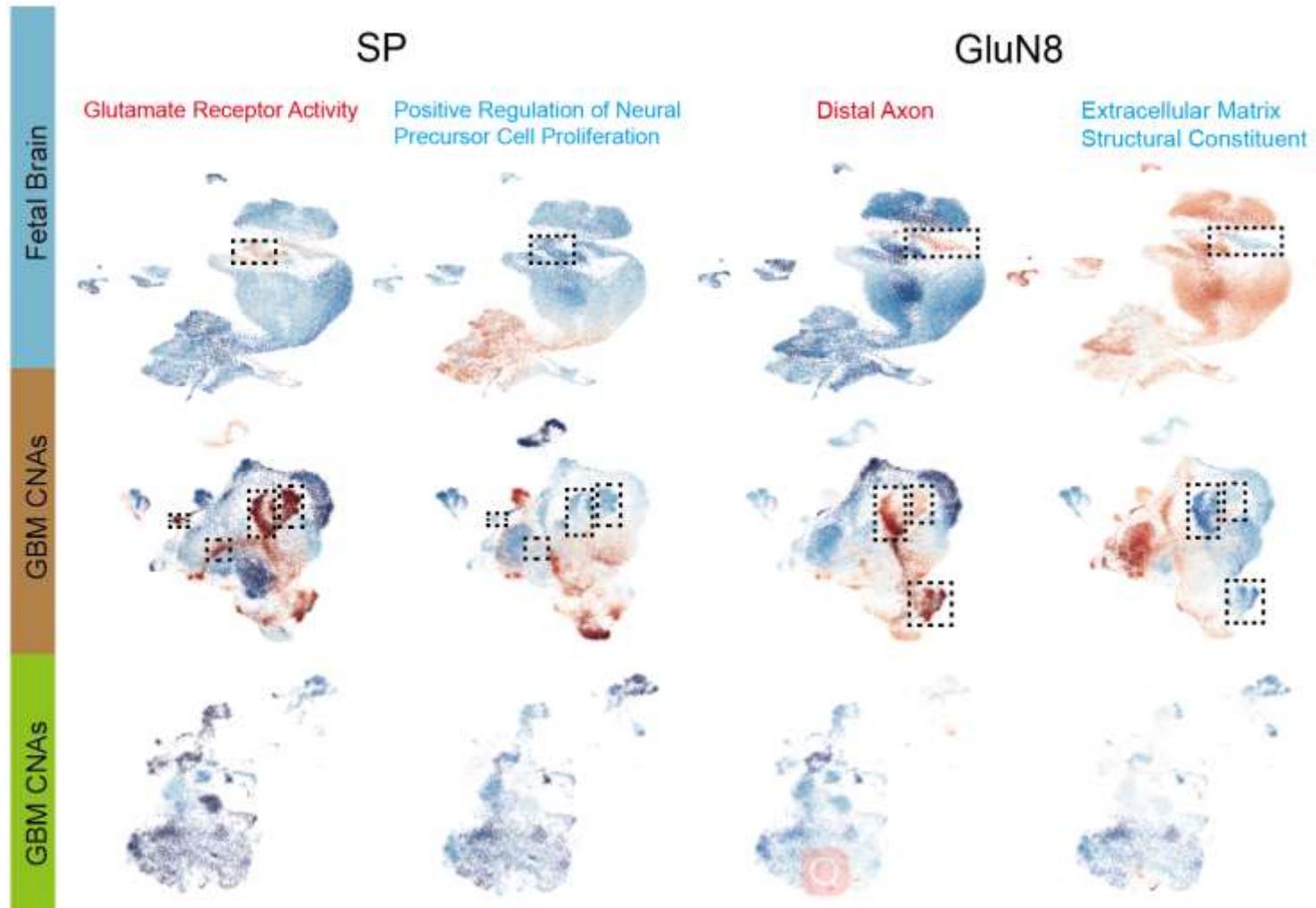
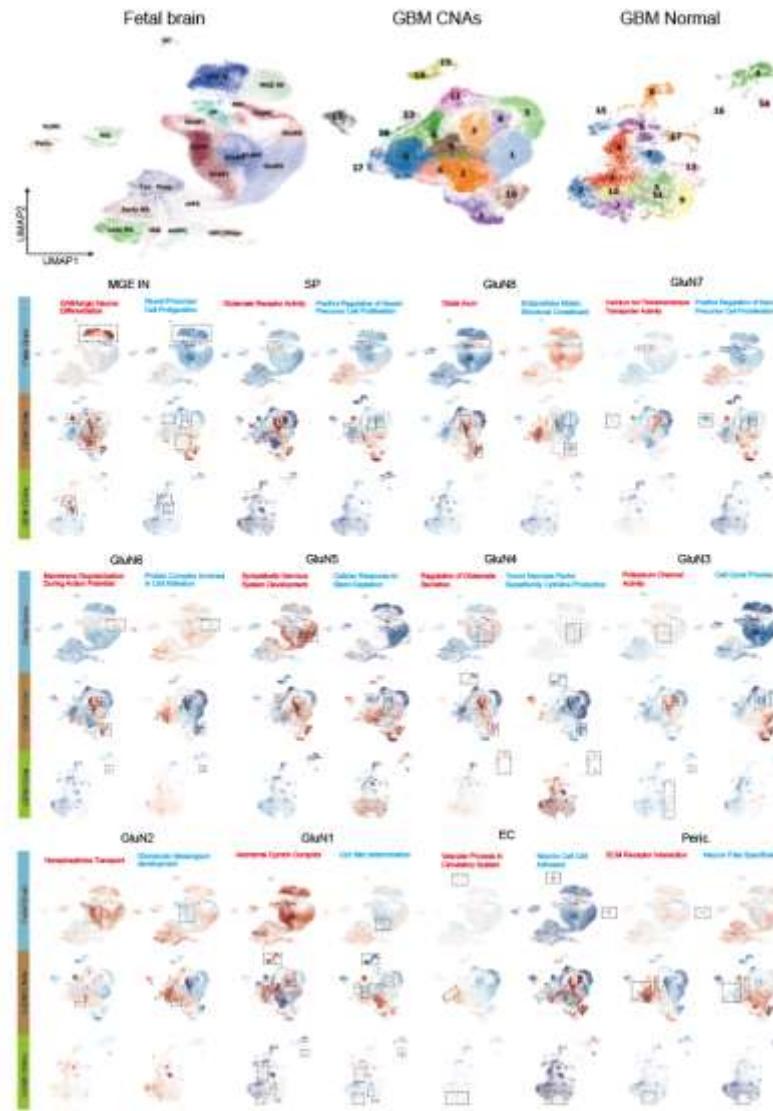
# PART 2 Progress

## Analysis of Differential Fetal Types Based on Pathways



## PART 2 Progress

Comparison of cellular activation pathways in fetal and GBM CNAs and normal fetal cells.



## PART 2 Progress

Compare the distribution of activated cellular pathways in fetal and GBM CNAs with those in normal fetal cells



Cyc. Prog.

Regulation of Mitotic Nuclear Division



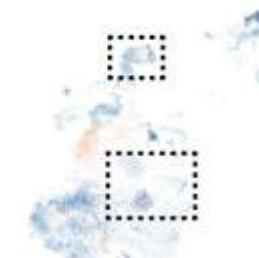
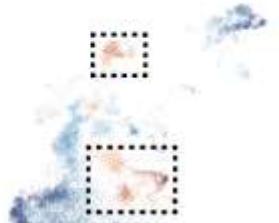
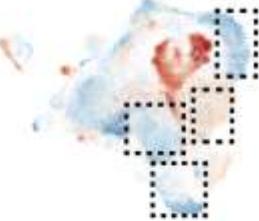
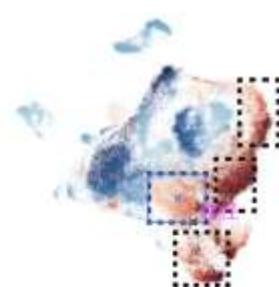
Neurotransmitter Receptor Activity



Fetal Brain

GBM CNAs

GBM CNAs



tRG

Intracellular Transport



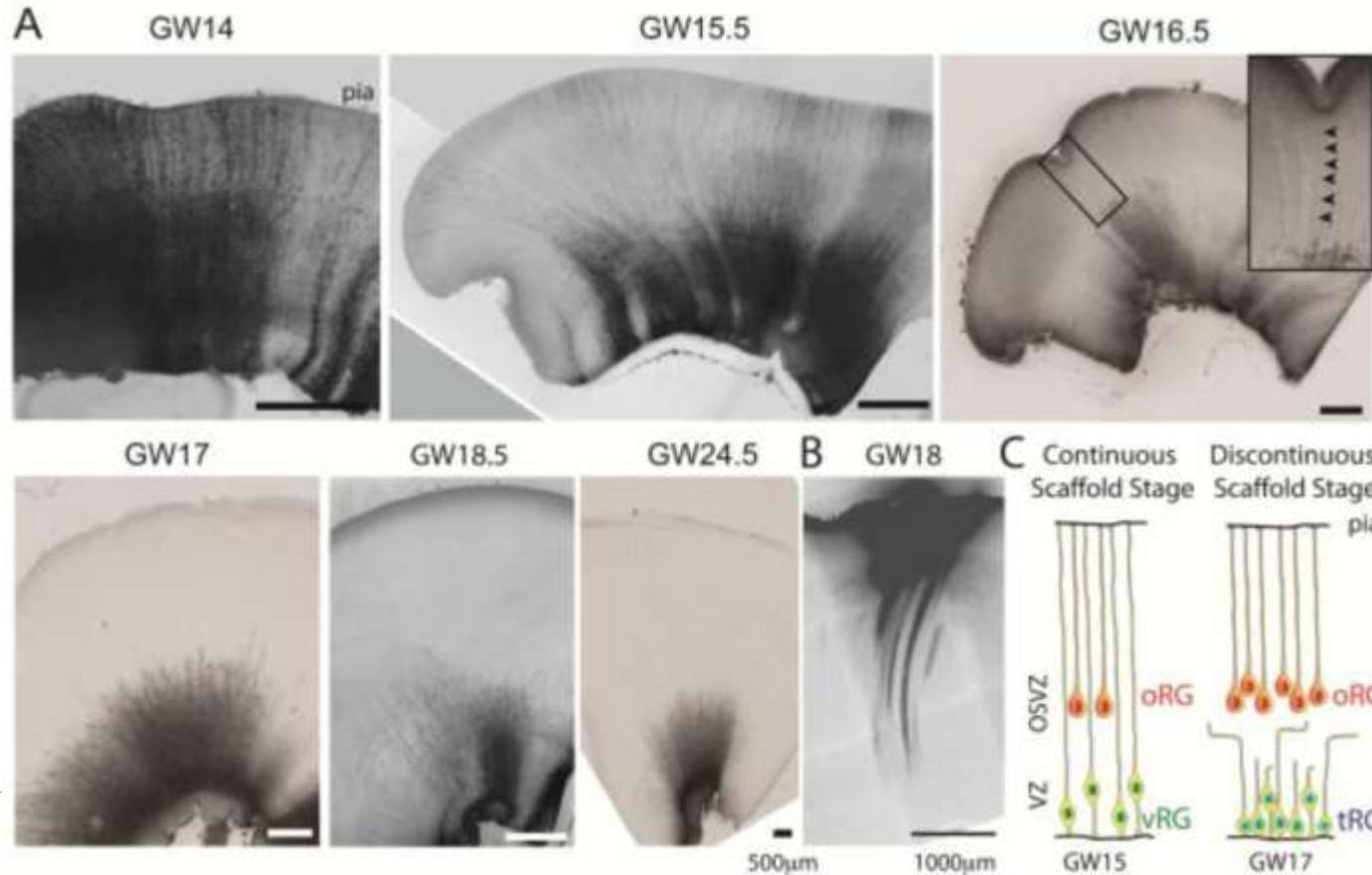
Myeloid Cell Activation Involved in Immune Response



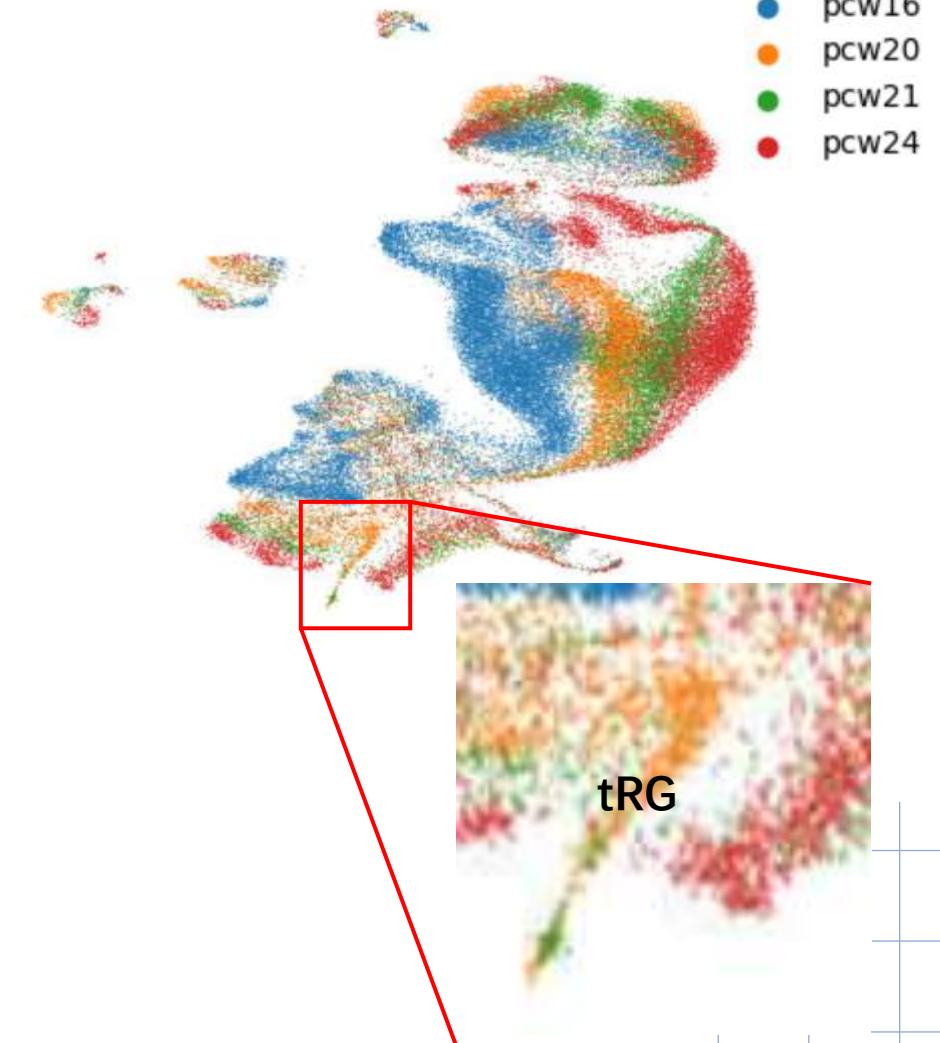
- The main activation pathways of developing neurons are partially activated in GBM malignant cells, while almost inactive in normal cells.
- The primary activation pathways of neural progenitor cells are partially activated in GBM malignant cells, with a few activations also present in normal cells.
- tRG cells have numerous activation pathways related to cilia, which are particularly distinctive in fetal data and widely activated in GBM

## PART 2 Progress

tRG is a very special type of neural cell group

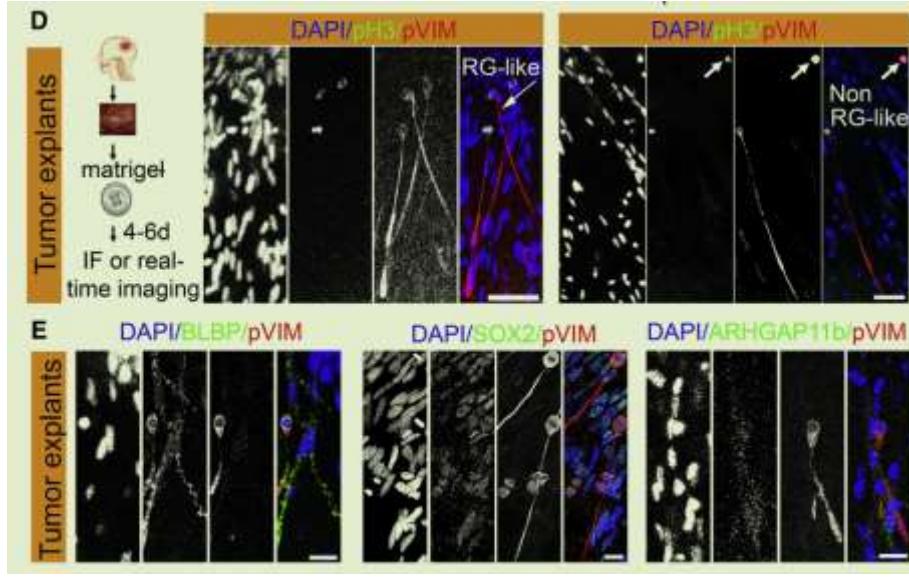


Trevino et al. (2021) fetal brain



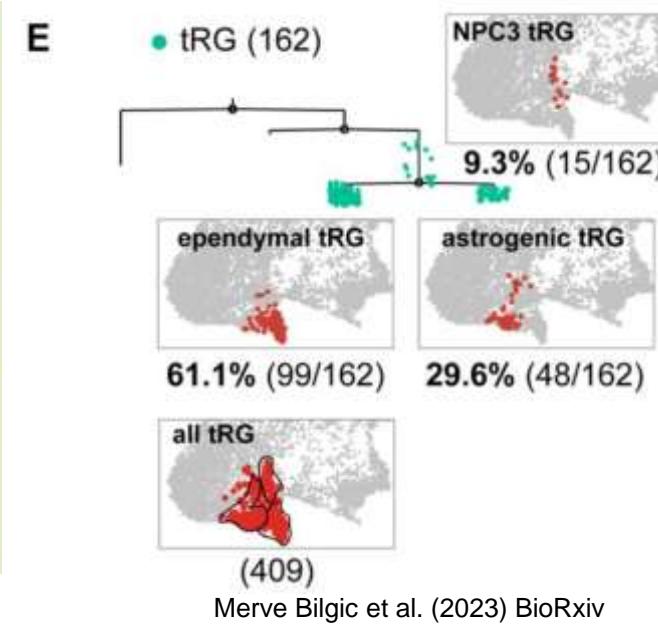
## PART 2 Progress

The origin of GBM remains unknown

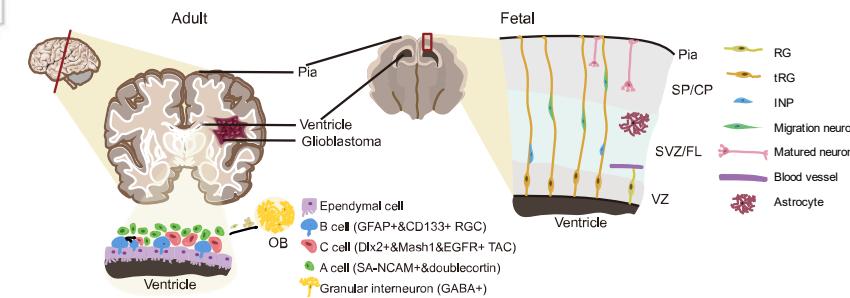


Rong Wang et al. (2020) Stem Cell Reports

In GBM, there are cells that resemble the morphology and behavioral patterns of RG cells, CD133+ & GLAST+ GBM invades the ventricular subependymal membrane



tRG generates 61.1% of ependymal cells

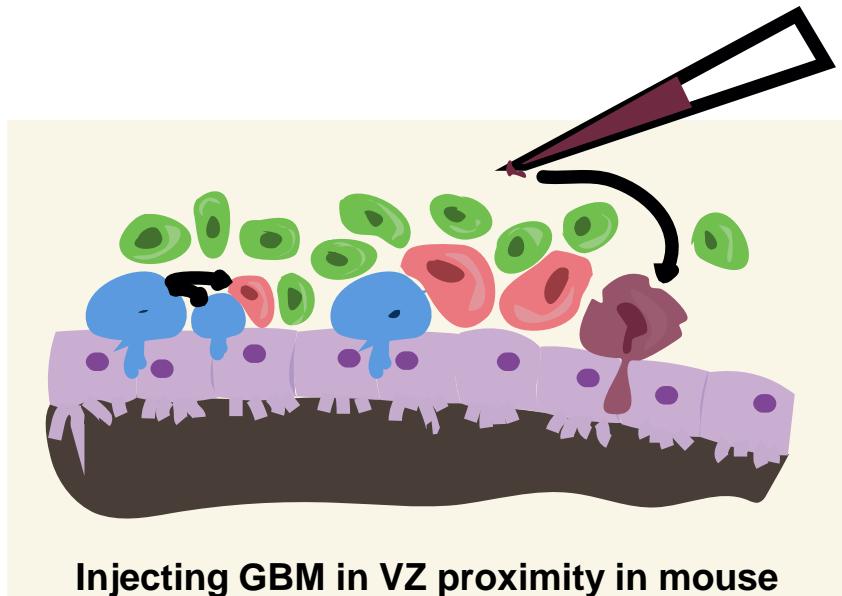
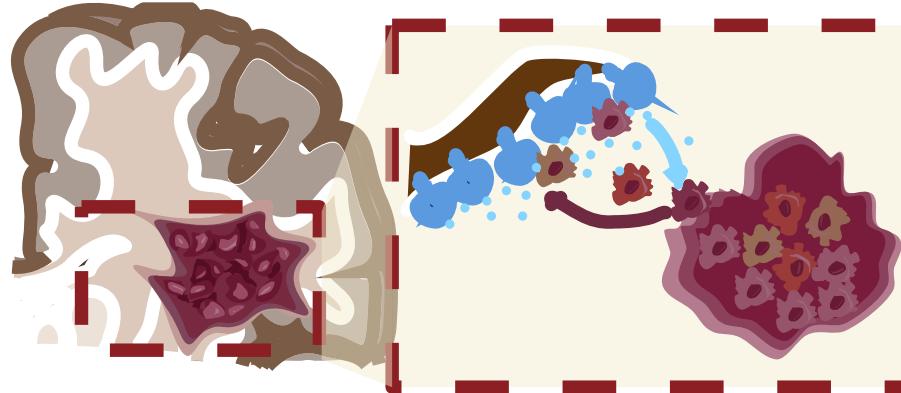


Adult neural stem cells exist in the subgranular zone (SGZ) of the dentate gyrus in the hippocampus, as well as in the subventricular zone (SVZ) of the brain ventricles

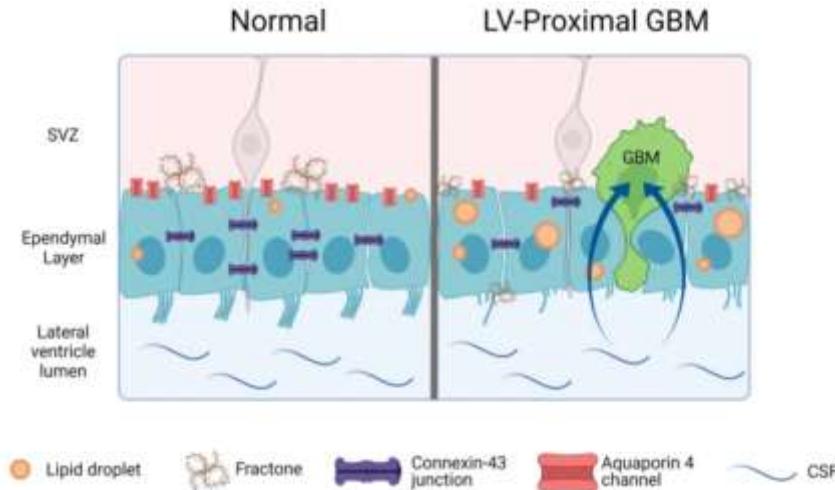
**tRG may be related to adult neural stem cells in the brain, perhaps associated with the origin or metastasis of GBM**

## PART 2 Progress

The origin of GBM remains unknown

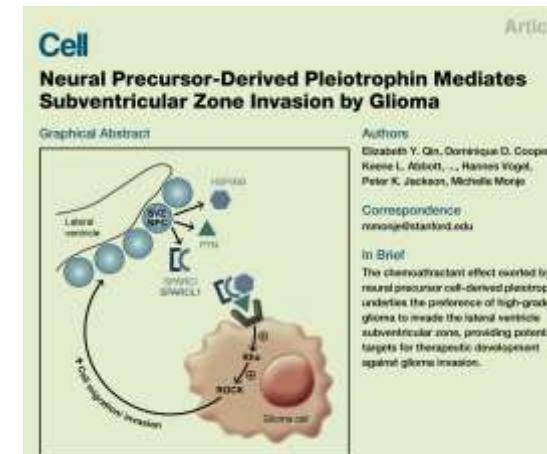


Injecting GBM in VZ proximity in mouse



Emily S. Norton et al. (2022) Fluids and Barriers of the CNS

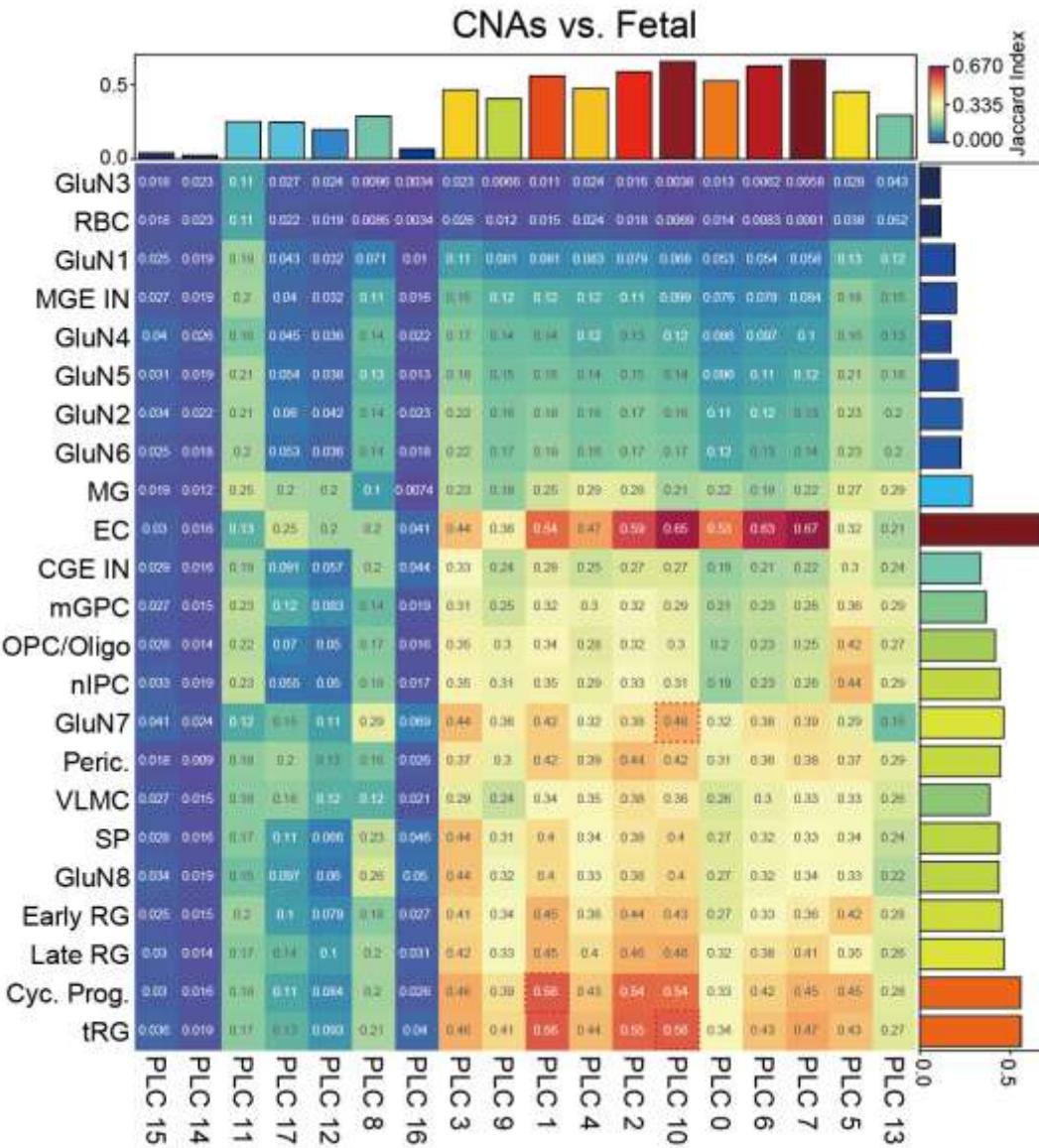
Transplanting GBM into the vicinity of the mouse's near ventricle,  
GBM behaves like ependymal cells, invading the ventricle



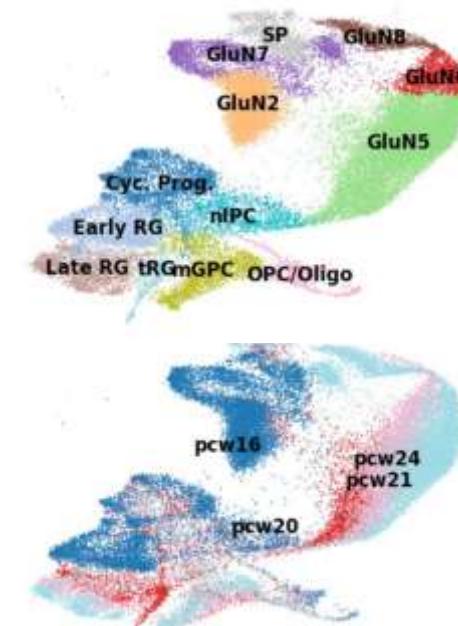
In the SVZ region, neural progenitor cells can secrete chemotactic factors to induce glioma invasion towards the SVZ

## PART 2 Progress

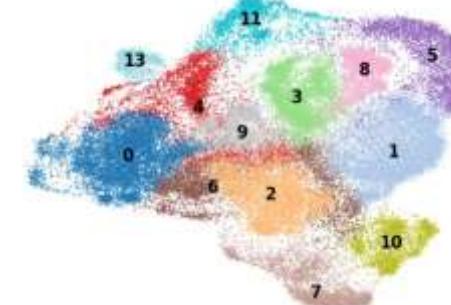
Removed cell types completely unrelated based on the pathways activation between fetal and CNAs cells



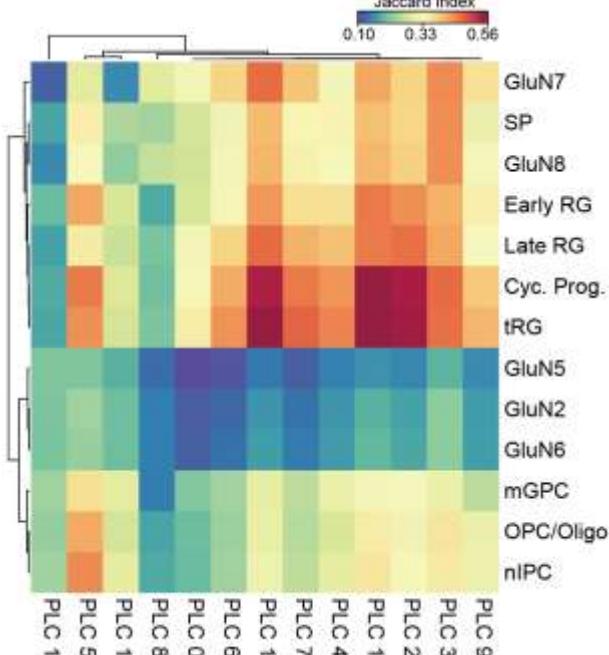
Filtered fetal brain



Filtered CNAs



Filtered CNAs vs Fetal

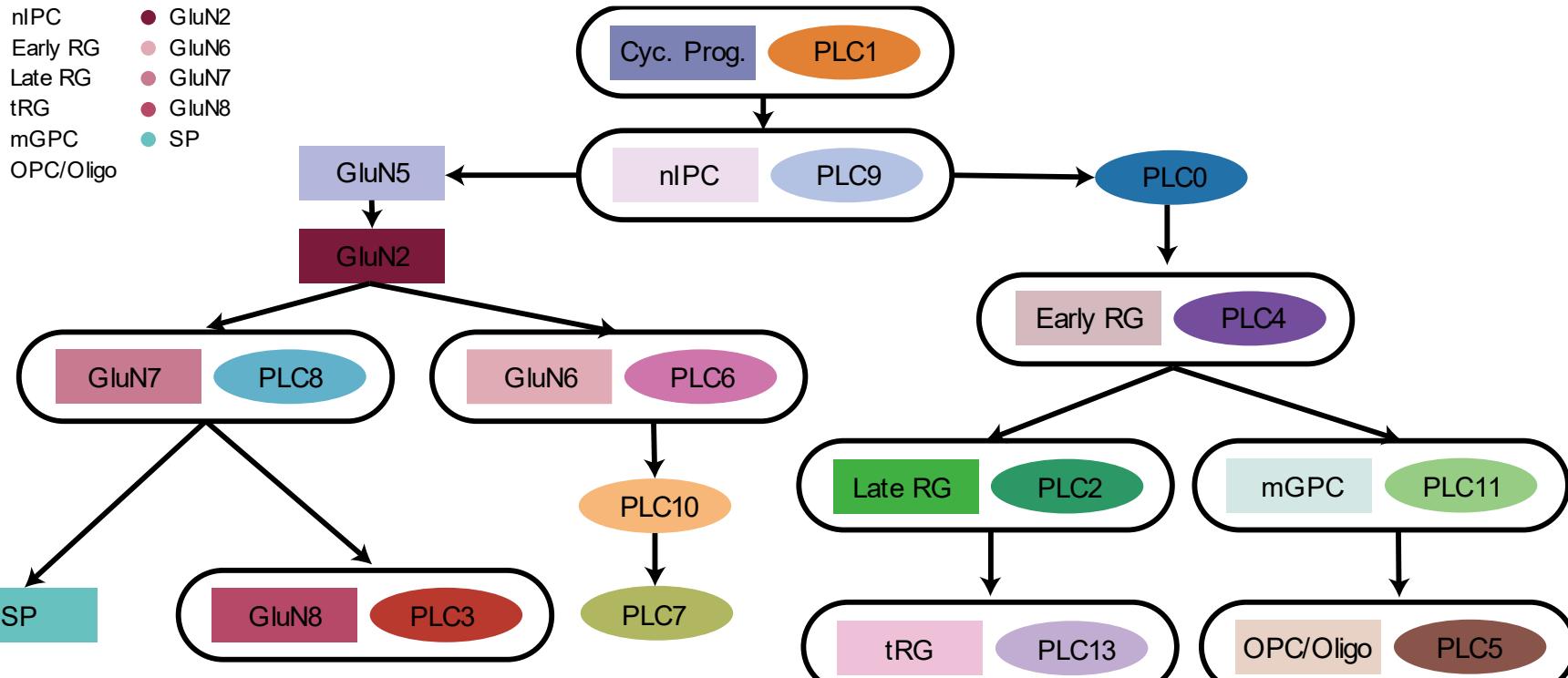
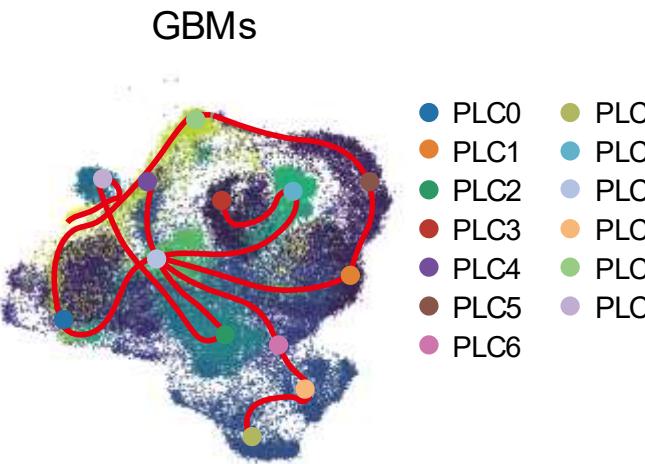
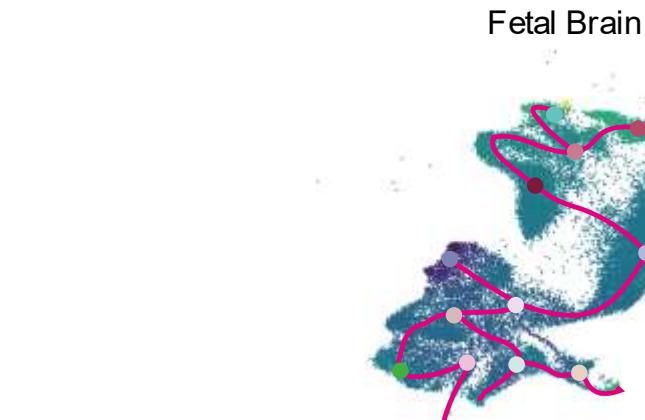


### Excluded cell subtypes:

- In fetal tissue, the activation pathways shared with RBCs and GBCs are less than 20%.
- In fetal tissue, ECs, Pericytes, VLMCs, and MGs are not related to the neurodevelopmental lineage.
- MGE IN and CGE IN, Although CGE IN shares more than 30% of activation pathways with GBCs, their developmental lineage origins are different and are not included in the analysis for now.
- PLC14, 15, 16, 17, and 12 have relatively low shared activation pathways with fetal tissue.
- GluN5, 2, 6 are temporarily retained because they have more than 20% shared activation pathways and are related to the developmental lineage.

## PART 2 Progress

### Fetal and GBM pseudo-time trajectories alignment



Branch1

Branch2

Branch3

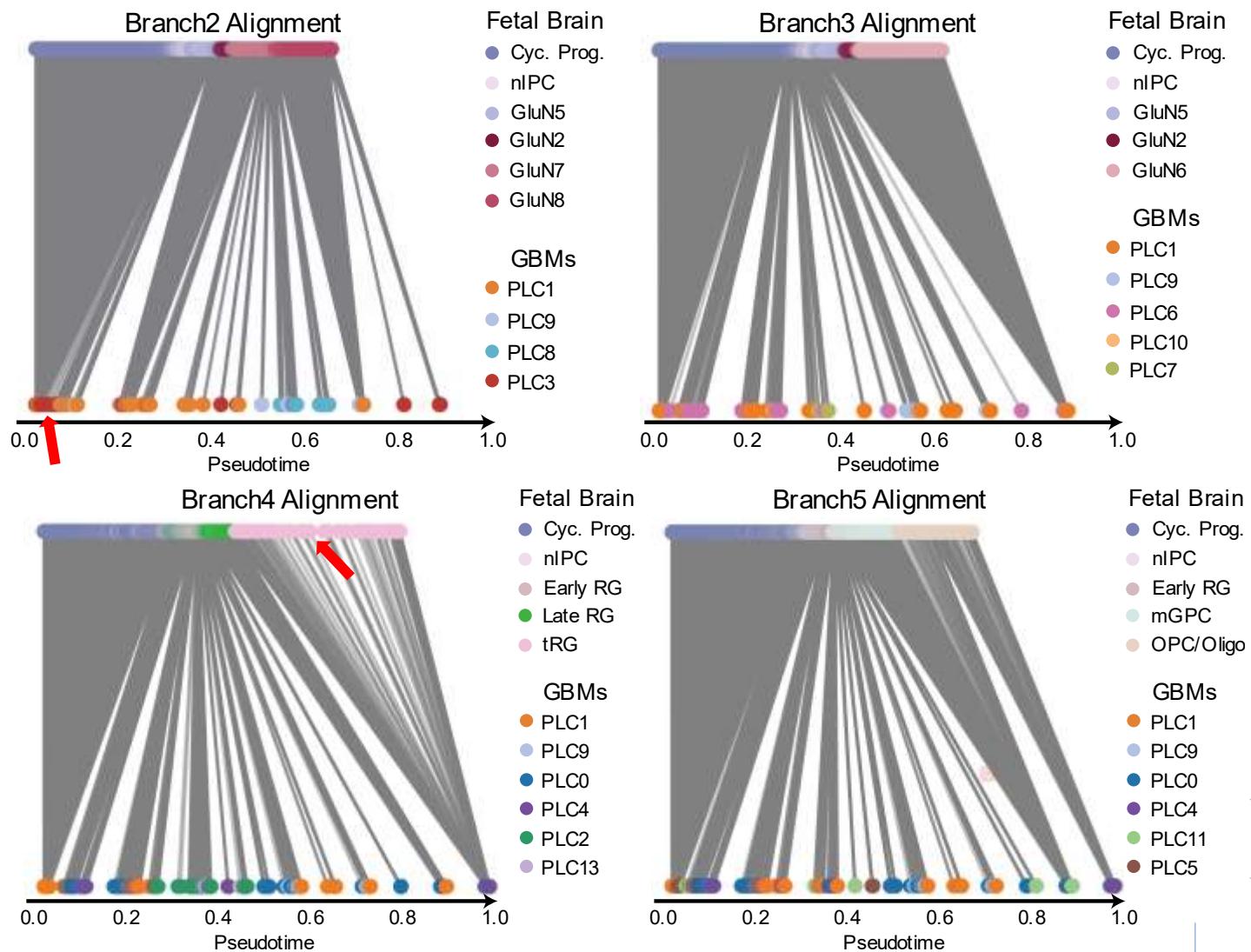
Branch4

Branch5

# PART 2 Progress

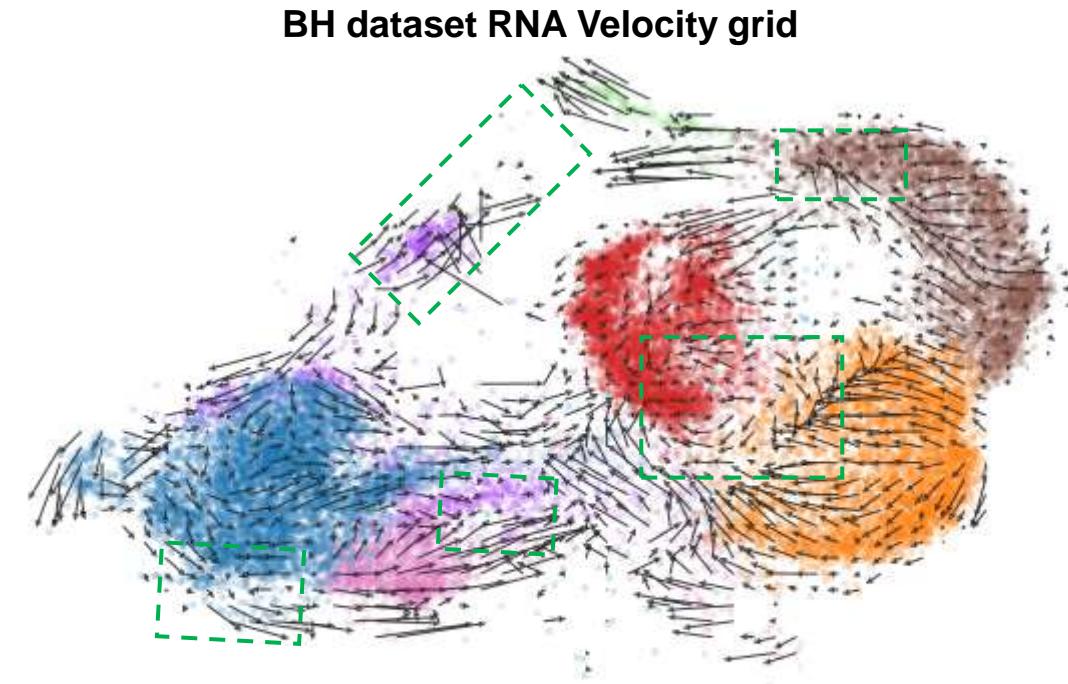
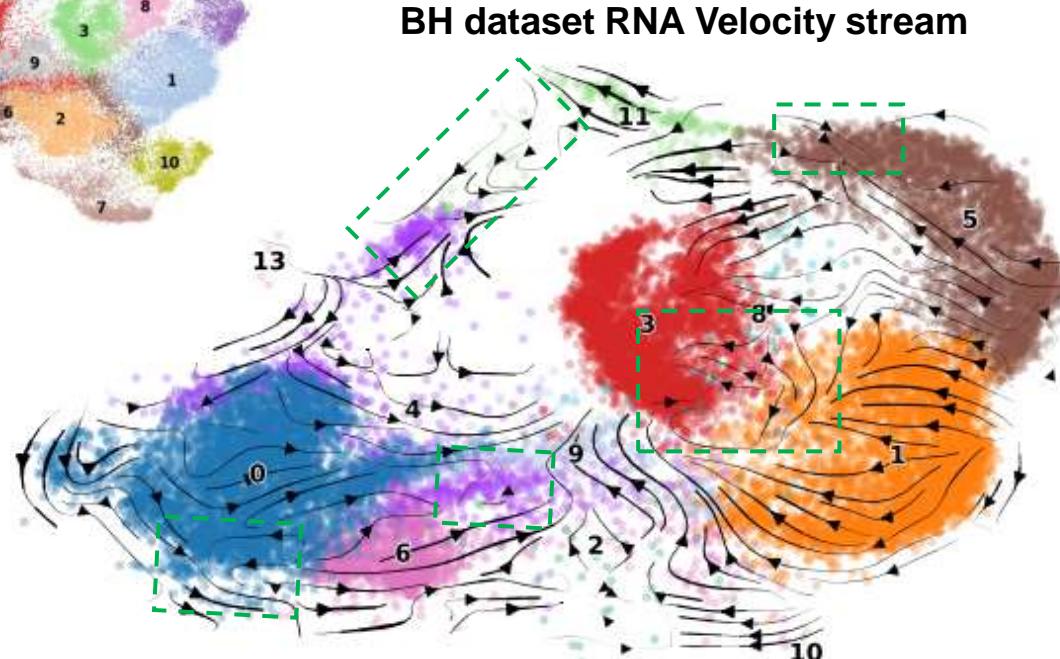
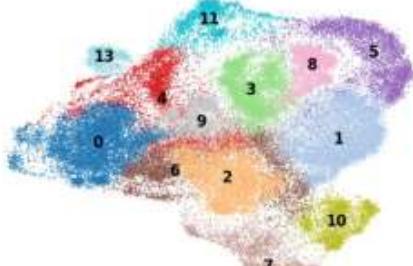
- ◆ GBM appears to have two separate cell clusters on the timeline
  - While GBM shares few activation pathways with the intermediate GluN 5,2,6 state of neurons during development, it has more common activation pathways with the more mature GluN7,8, SP.
- ◆ In fetal development, there is a differentiation direction from NSC to Neuron/OPC, whereas **GBM appears to form a loop, enclosing Neural-like cells in the middle.**
- ◆ Comparison reveals five branching pathways. The first branch primarily represents the normal neuronal developmental differentiation pathway, the second branch generates neural-like GBC, the fourth branch mainly maintains a progenitor state, and the fifth branch points towards glial cell differentiation.
- ◆ Comparing fetal and GBM cells on several branches, GBM cells tend to cluster at the ends of the timeline on branches 2 and 3, showing a tendency towards a stem cell state. Additionally, a minority of cells in the PLC1 (corresponding to Cyc. Prog) cluster on the timeline are more 'mature' than developing neurons.
- ◆ Branches 4 and 5 mainly correspond to neural stem cells and glial progenitor cells, respectively. Some GBM cells are located at the maximum pseudotime value (most mature, but corresponding to stem cells).
- ◆ There is a clear trend in GBM's comparison to tRG, with clustering at both ends of the timeline.

## Fetal and GBM pseudo-time trajectories alignment



## PART 2 Progress

Filtered CNAs



- Only the BH dataset has raw data available for RNA velocity analysis.
- The datasets PLC4, 7, and 10 have fewer cells, which may result in slightly bias results.
- Overall, most cell types point towards PLC3, indicating neuronal differentiation (the branch of glial differentiation lacks mature glial cell types, and the direction is not particularly clear).
- In some regions (green dashed boxes), there are relatively pointing arrows, especially arrows from PLC3 towards PLC1 (neuron -> stem cell), suggesting **the possibility of neuronal differentiation acquiring stemness, or the presence of a 'force' pulling them back to a stem cell state**



03

## Next Plan



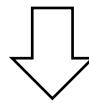
## PART 3 Next Plan

### Hypotheses on the origin and differentiation of GBM



#### GBM origin (Comparison of CAN distribution in PLC3 and PLC1)

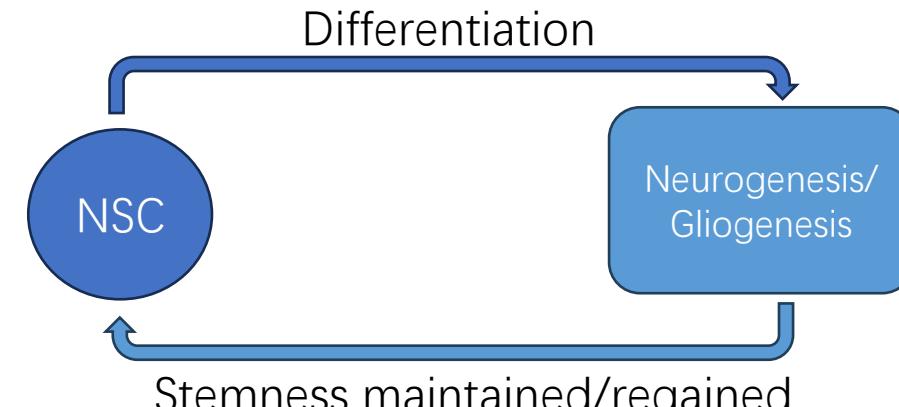
- Mature cells acquire stemness.
- aNSCs receive stimuli/activation signals, differentiate and leave away from the V-SVZ region, initiating a “wrong” differentiation process becoming into cancer cells.
- The two mix (highly differentiated cells acquire stemness -> trigger inflammatory response (damage repair) -> excessively activate quiescent state aNSCs -> result in differentiation errors into tumor cells)



#### If GBM origin is related to aNSCs

- What aNSC (B/C/E/A) transforms into GBM
- What is the differentiation lineage like
- Whether GBM also has aNSC quiescent state, and which molecules are involved, whether it is related to recurrence
- Whether the mechanism to revert active aNSC back to quiescence can inhibit GBM progression

**Is there a force that simultaneously promotes GBM differentiation into neurons/glial cells and pulls them back to a stem cell state/maintains them in a stem cell state?**



Using different sets of genes for RNA velocity analysis, to see if we can identify two groups of genes leading to opposite velocity directions, which could be regarded as two opposing forces.

- Currently, several trajectory types are compared with previous three single-cell subtypes of GBM.
- Define GSC, determine the types of cycle and non-cycle (quiescence).
- Adult brain cortex single-cell data compared to fetal and GBM data.
- Perform cell communication analysis using NeuroChat and cellChat on different subtypes of cells and developmental trajectory nodes (with a focus on Notch communication).
- Use the finally constructed GBM lineage to deconvolute TCGA and IVY bulk data (deconvolute several different trajectories), associate survival, epigenetic regulation, spatial distribution MRI signals, etc., and discuss the relationship between GBM developmental tree and disease phenotype and epigenetic regulatory signals.