单细胞RNA修饰数据解构及多组学数据整合分析实践

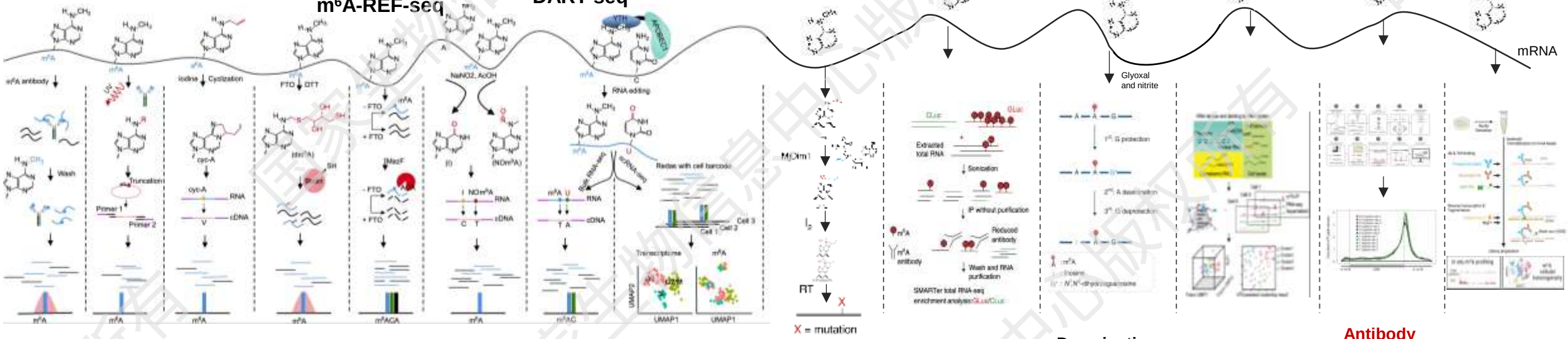
高纯纯、凡秀

国家生物信息中心

China National Center for Bioinformation

m⁶A测序技术的发展

Bulk（> 1 μg）	Low-input （100 ng~μg）	Ultra-low input（< 100 ng）	Single cell （~ 5-300 pg）



Antibody
Dominissini *et al.*, *Nature*, 2012; Meyer et al., *Cell*, 2012

Chemical label
Shu *et al.*, *Nat Chem Biol*, 2020

Demethylation
Endonuclease
Zhang *et al.*, *Sci Adv*, 2019

Deamination
Nucleotide mutation
Meyer *Nat Methods*, 2019

m⁶A-selective allyl chemical labeling
Hu et al., *Nat Biotechnol*, 2022

Deamination
Nucleotide mutation
Liu *et al.*, *Nat Biotechnol*, 2022

Antibody
Li et al., *Nat Biotechnol*, 2023

Antibody
Crosslinking
Linder et al., *Nat Methods*, 2015

Demethylation
Chemical label
Wang *et al.*, *Nat Chem Biol*, 2020

Chemical transition
Nucleotide mutation
Werner *et al.*, *Nucleic Acids Res*, 2021

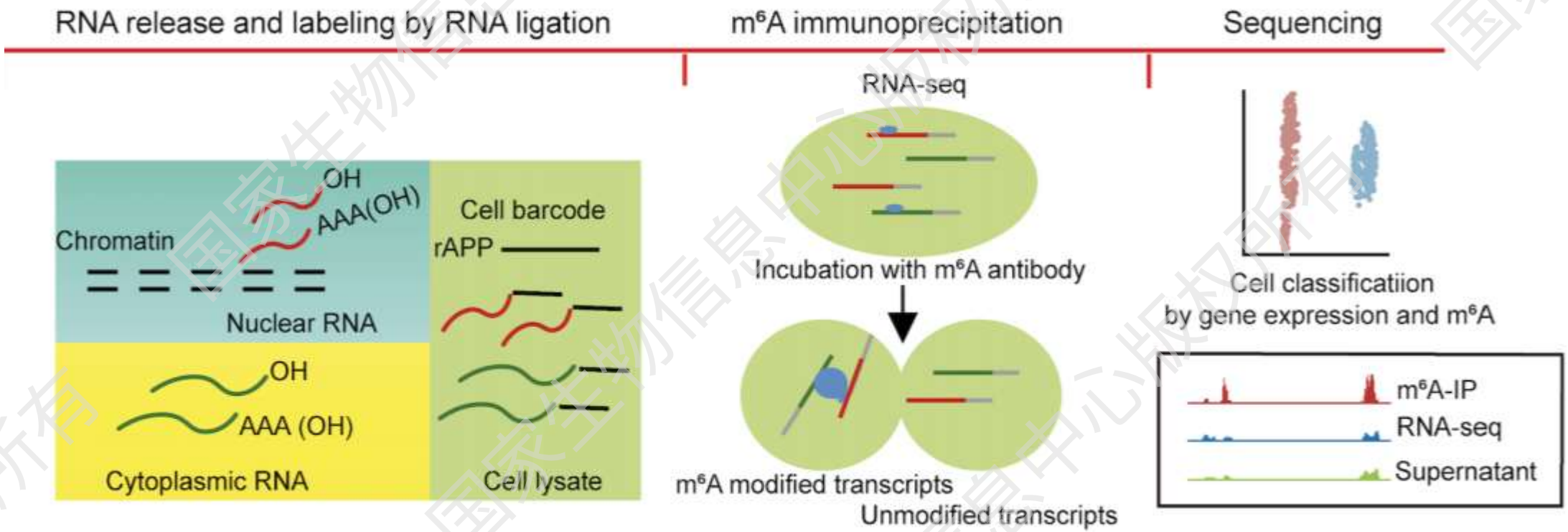
Deamination
Nucleotide mutation
Tegowski *et al.*, *Mol cell*, 2022

Antibody
External RNA Controls Consortium (ERCC)
Wu et al., *Nat Cell Biol*, 2022

Antibody & Multi-plex labeling
Yao *et al.*, *Nat Commun*, 2023

Antibody & CUT&Tag
Hamashima *et al.*, *Mol cell*, 2023

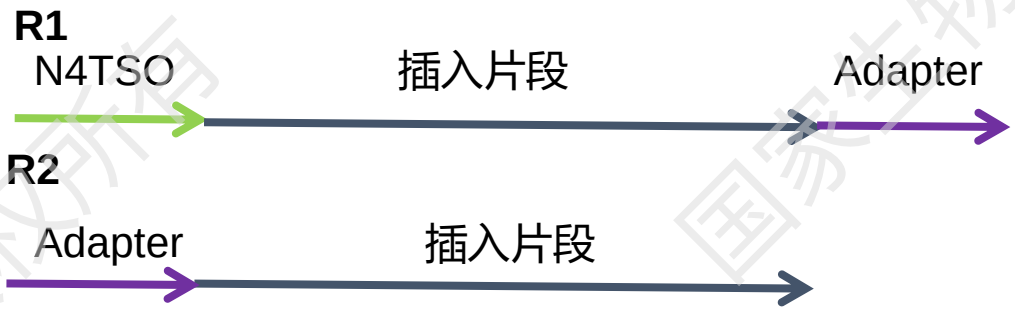
低通量单细胞RNA m⁶A测序技术流程



scm^{6}A-seq文库结构解析

- **文库结构示意图：**

- **数据结构示意图（由于公司会去掉P5端引物和Index的部分序列，所有部分序列是缺失）：**



scm^{6}A-seq单文库数据拆分

############### The tso-mix index information. Index ID　　Index sequence　(5' of 1st read in pair-end)
N4　　^NNNNGGG
index1　^ATCACGNNNNGGG
index2　^CGATGTNNNNGGG
index3　^NNNNTTAGGCNGGG

################## If there are hunderds and more cells are in scm^{6}A-seq data, tso-mix separation is needed.

sh ./scm6A-tsomix-separation.sh ${dir}/Test.fastq.gz

Expcted output file:
*index1.fastq
*index2.fastq
*index3.fastq
*N4.fastq

scm^{6}A-seq单细胞数据拆分和比对

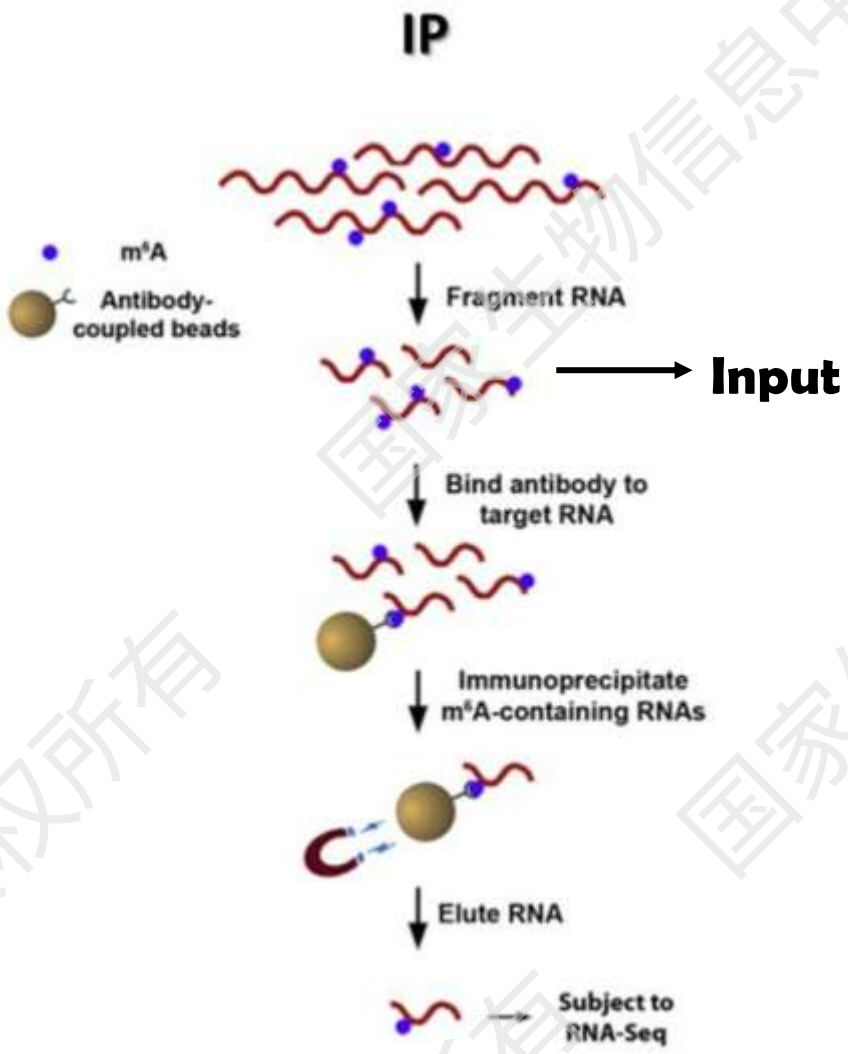
################ After tso-mix separation,cellular barcodes is needed for single cell reads isolation. Or if less than 20 cells in one library, single cell reads isolation is performed directlty.
Test 10 thousand reads for single cell isolation with 6 cores, cost 1m24.507s

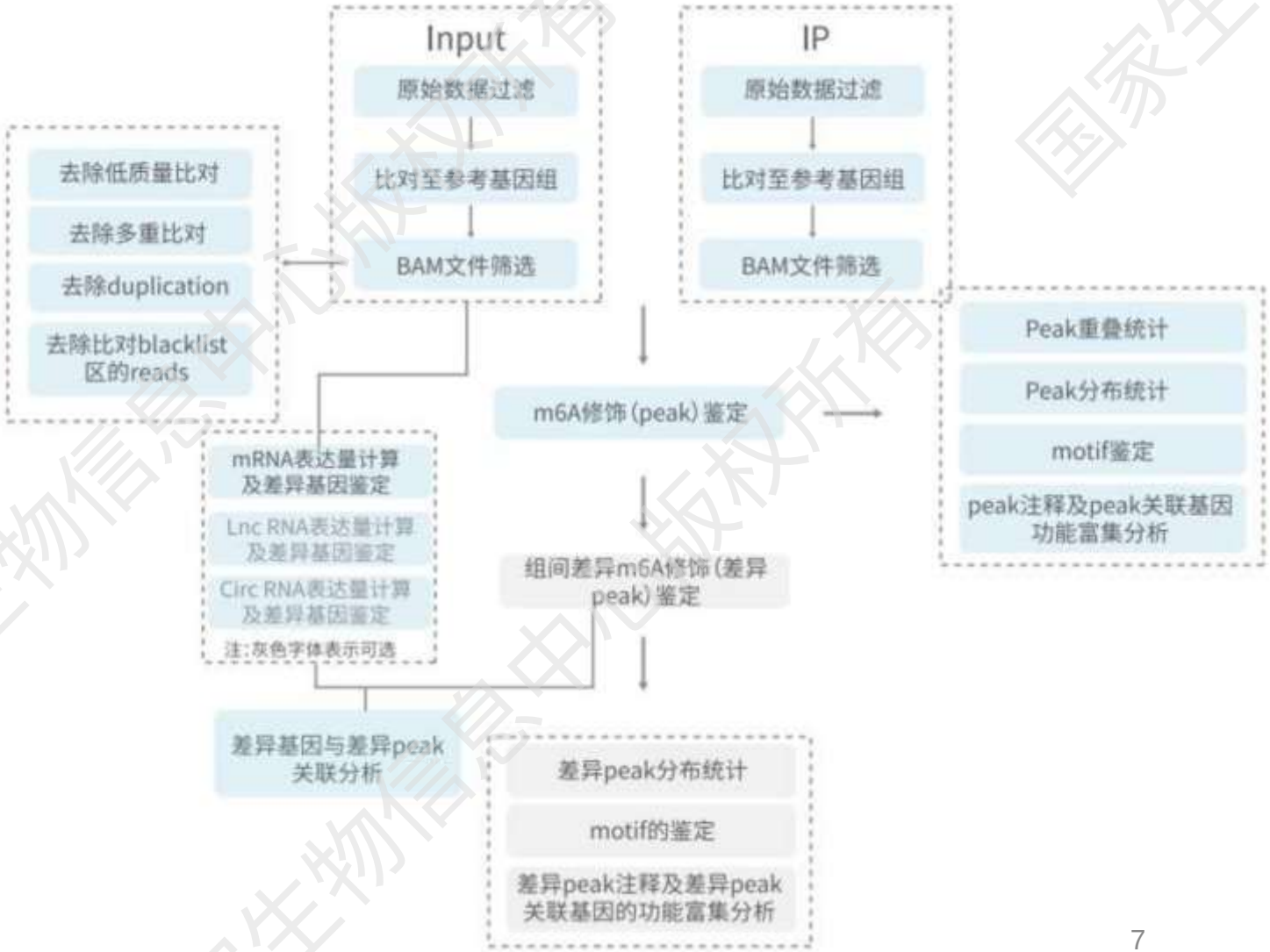
sh ./scm6A-SCbarcodes-isolation.sh

Expected output files:
*_${barcode}.fastq.gz # (single cell reads)

################ Mapping
```
for i in AAGCTA ATTGGC CTCTAC CTGATC GATCTG GCGGAC GGAACT GTAGCC TTGACT;
        do
                cd $Dir/$Filename
                mkdir ${Filename}_${i}
                hisat2 -p 6 --dta -x
/asnas/yangyg_group/gaochch/REFERENCE/Mouse/ENSEMBL_GRCm38/HISAT2-INDEX/genome_trans -
1 ./1-clean_data/${Filename}_N4_${i}_1.fq.gz -2 ./1-clean_data/${Filename}_N4_${i}_2.fq.gz -
S ./${Filename}_${i}/hisat2.sam
done
```

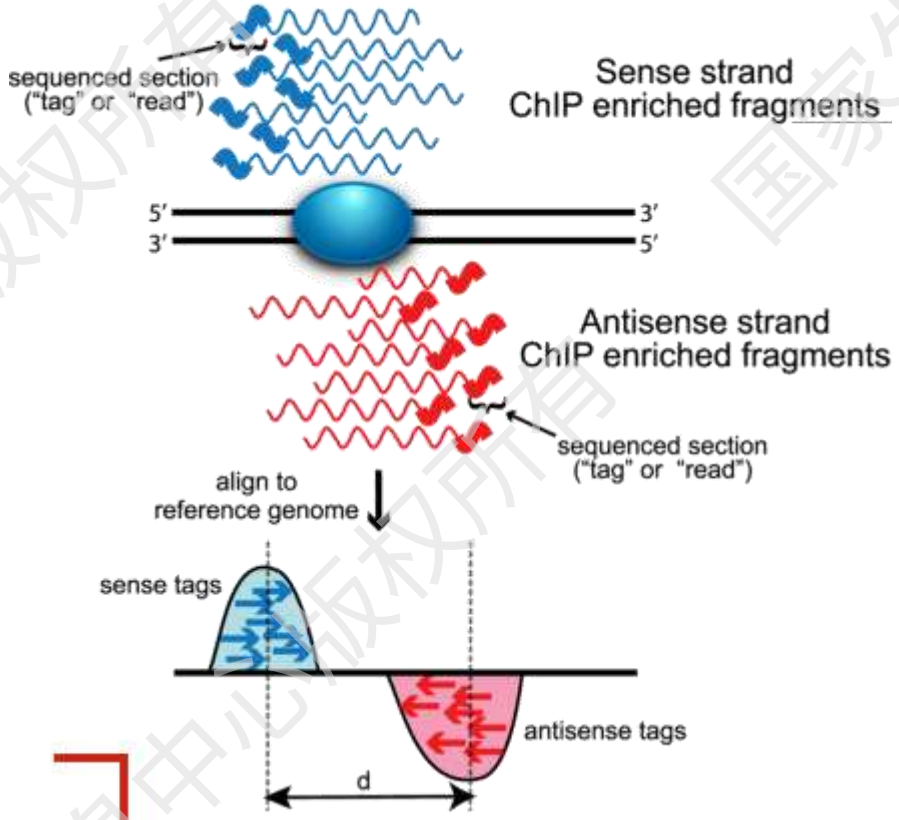
MeRIP-seq技术原理和数据处理流程





修饰区域鉴定

```
$ macs2 callpeak -t IP_aln.bam \
-c Input_aln.bam \
-f BAM -g 1.3e+8 \
-n Nanog-rep1 \
--outdir macs2
```



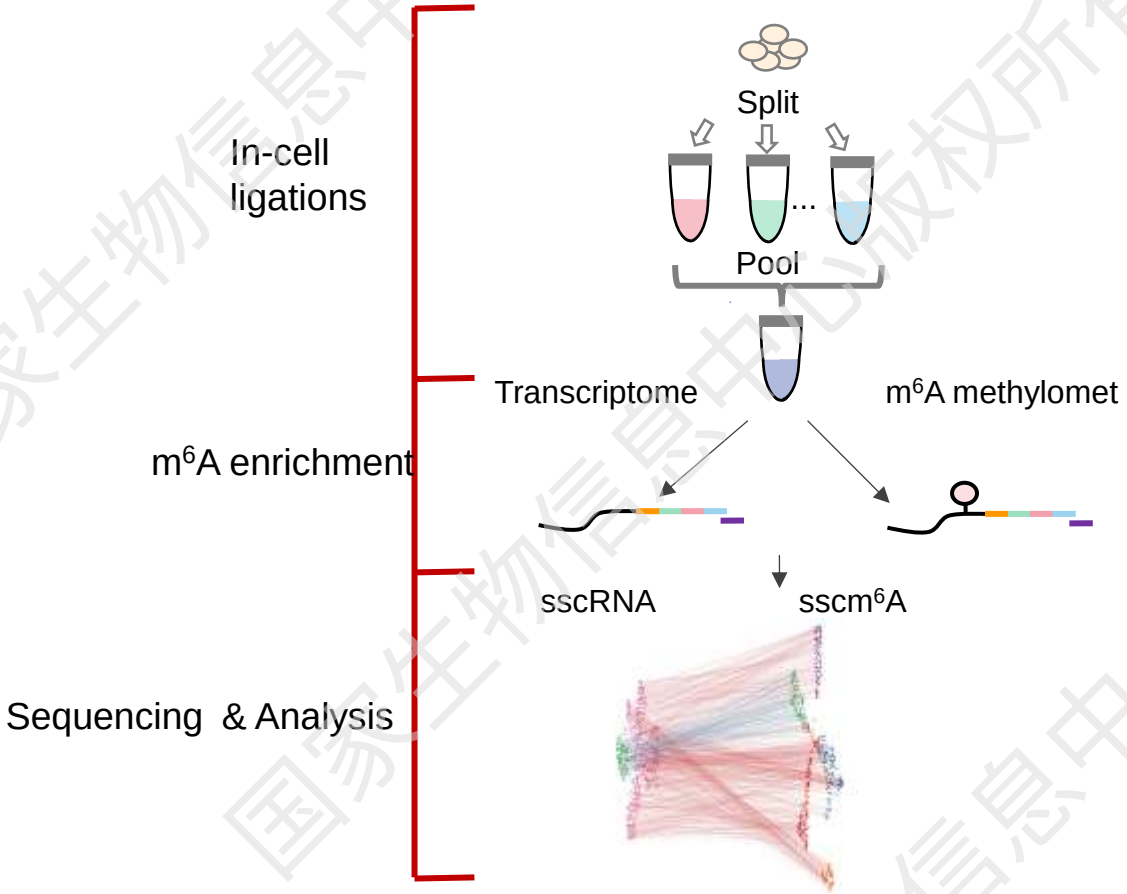
1. chromosome
2. start coordinate
3. end coordinate
4. name
5. score
6. strand

Standard BED file fields

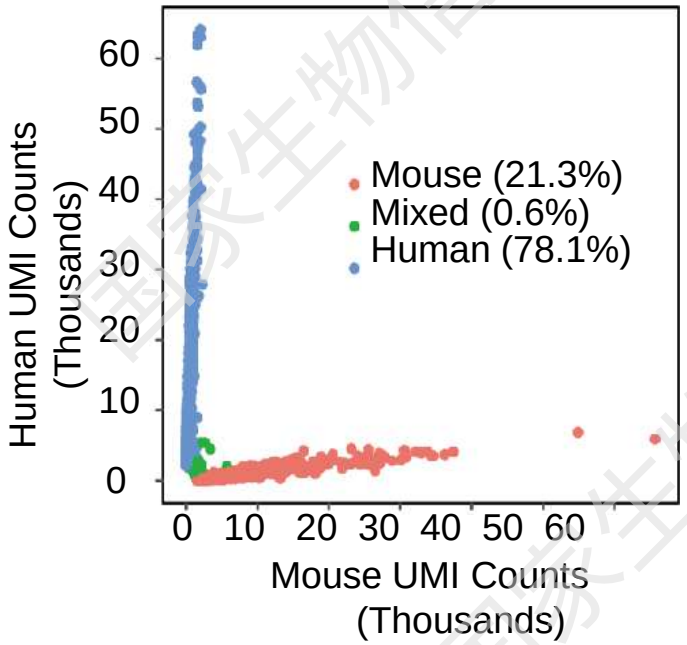
7. **signalValue - Measurement of overall enrichment for the region**
8. **pValue - Statistical significance (-log10)**
9. **qValue - Statistical significance using false discovery rate (-log10)**
10. **peak - Point-source called for this peak; 0-based offset from chromStart**

narrowPeak specific fields

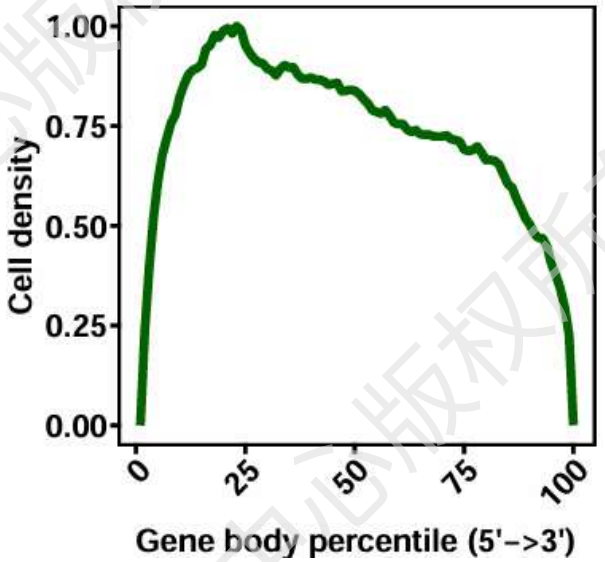
高通量单细胞修饰组和转录组测序技术



单细胞标记的纯净度和覆盖度分析

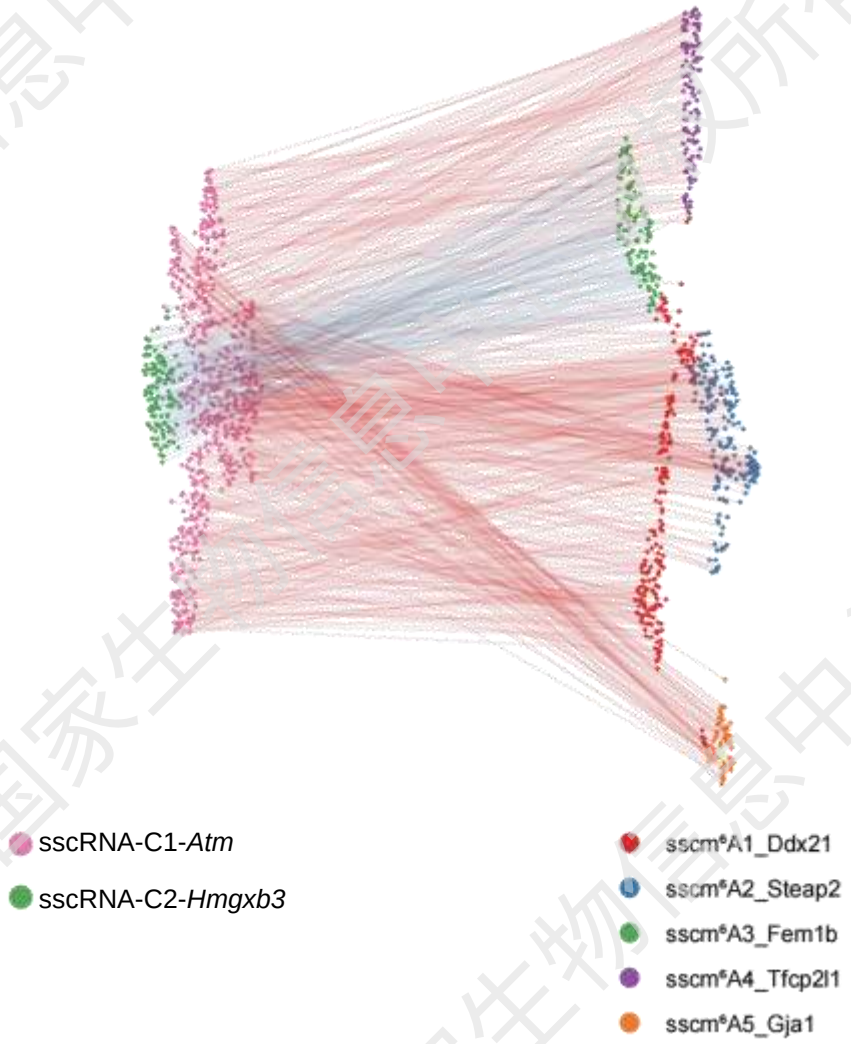


纯净度

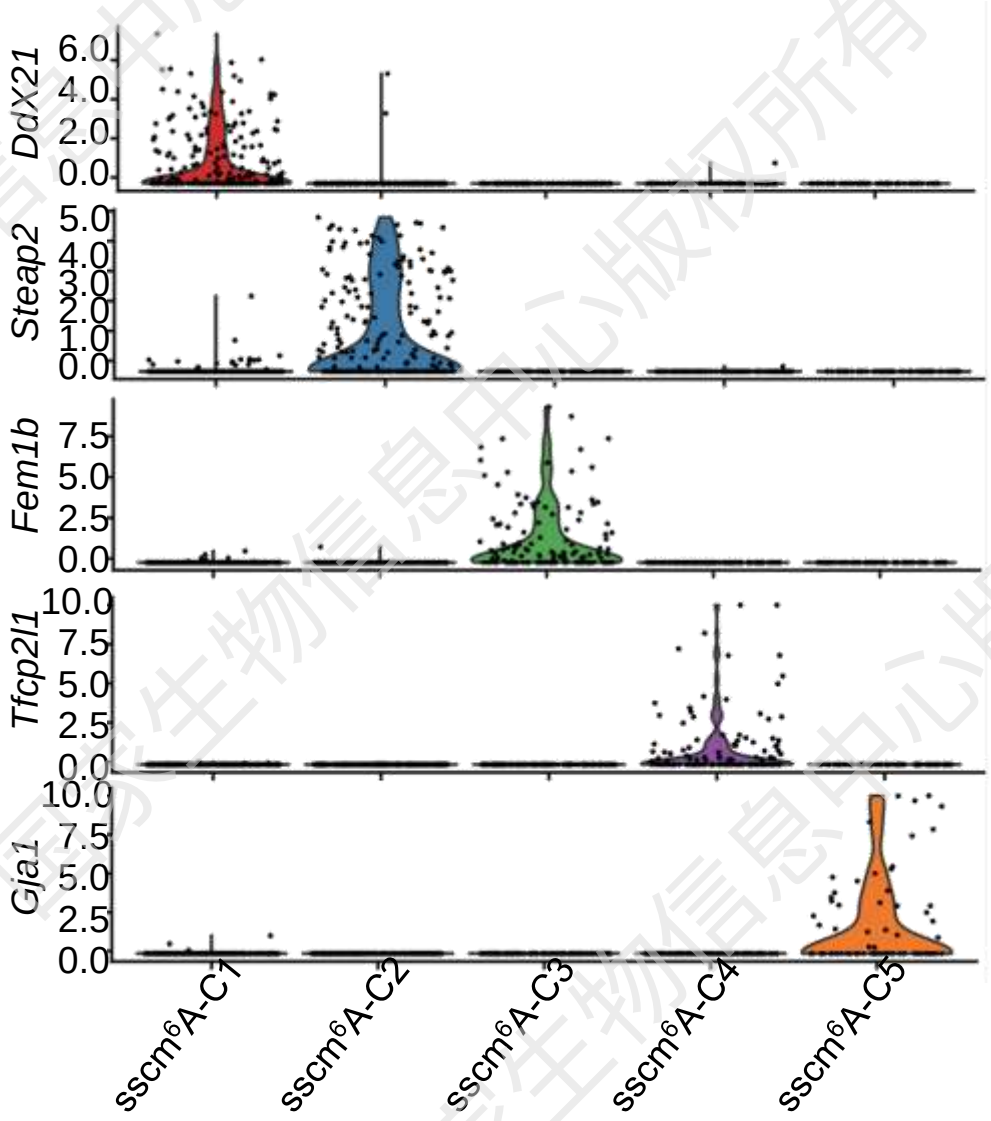


覆盖度

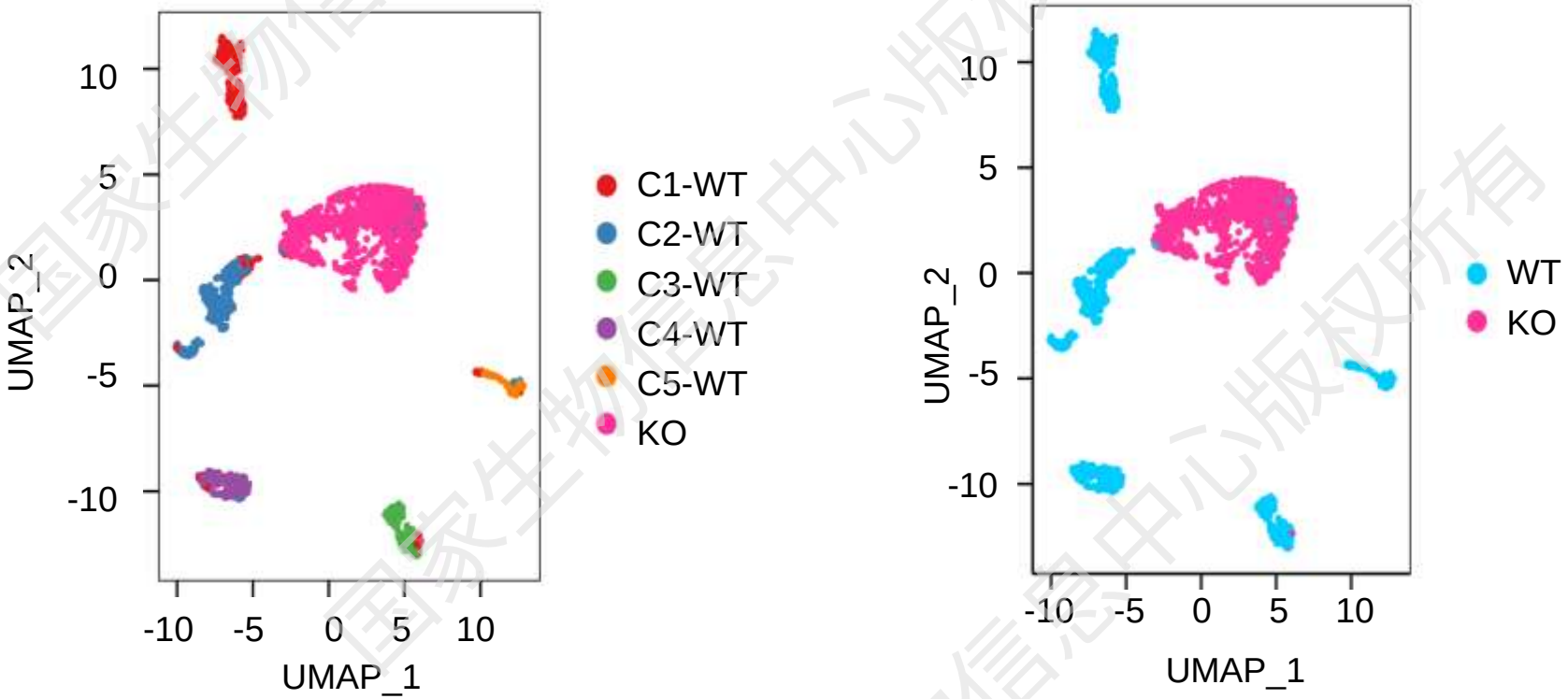
单细胞多组学组学数据解构和对齐



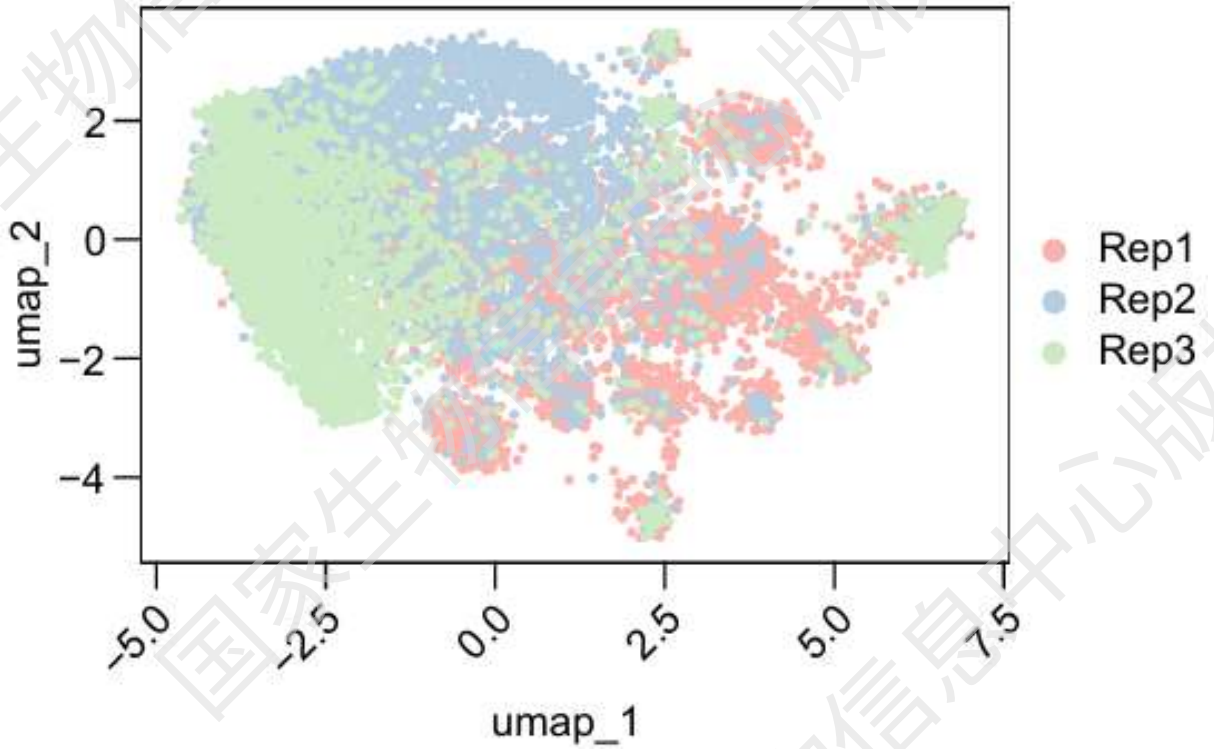
细胞类型特异的修饰基因鉴定



多样本整合降维分析



多批次样本整合去批次降维- Harmony



单细胞聚类质量显著性评估

1. Parametric model of gene expression

Define the G × N UMI counts matrix as Y.

$$\boldsymbol{Y}_{J_0,n} \sim \prod_{g \in J_0} \text{Poisson}(\mu_g).$$

This implies that unexpressed genes are described by independent Poisson distributions with means μ_g. The expressed genes were assumed to follow a Poisson log-MVN distribution:

$$\boldsymbol{Y}_{J_1,n} \sim \text{Poisson log-MVN}(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

Given observed Y， fit $\mathcal{F}$， a joint distribution for the cell population

For each expressed gene g, we set

$$\hat{\Sigma}_{g,g} = \log\left(\frac{v_g - m_g}{m_g^2} + 1\right)$$

and

$$\hat{\mu}_g = \log m_g - \frac{1}{2}\hat{\Sigma}_{g,g},$$

where m_g and v_g are the sample mean and variance, respectively, for gene g. Then for any two genes g_1, g_2 such that $g_1 \neq g_2$, we set

$$\hat{\Sigma}_{g_1,g_2} = \log\left(\frac{\sigma_{g_1,g_2}}{\exp\left(\hat{\mu}_{g_1} + \hat{\mu}_{g_2} + \frac{1}{2}\left(\hat{\Sigma}_{g_1,g_1} + \hat{\Sigma}_{g_2,g_2}\right)\right)} + 1\right),$$

2. Quantifying clustering quality (Wald linkage)

$$\text{ESS}(C_1) = \sum_{i \in C_1} \left\| \boldsymbol{x}_i - \frac{1}{|C_1|} \sum_{i \in C_1} \boldsymbol{x}_i \right\|^2$$

$$\text{ESS}(C_2) = \sum_{i \in C_2} \left\| \boldsymbol{x}_i - \frac{1}{|C_2|} \sum_{i \in C_2} \boldsymbol{x}_i \right\|^2$$

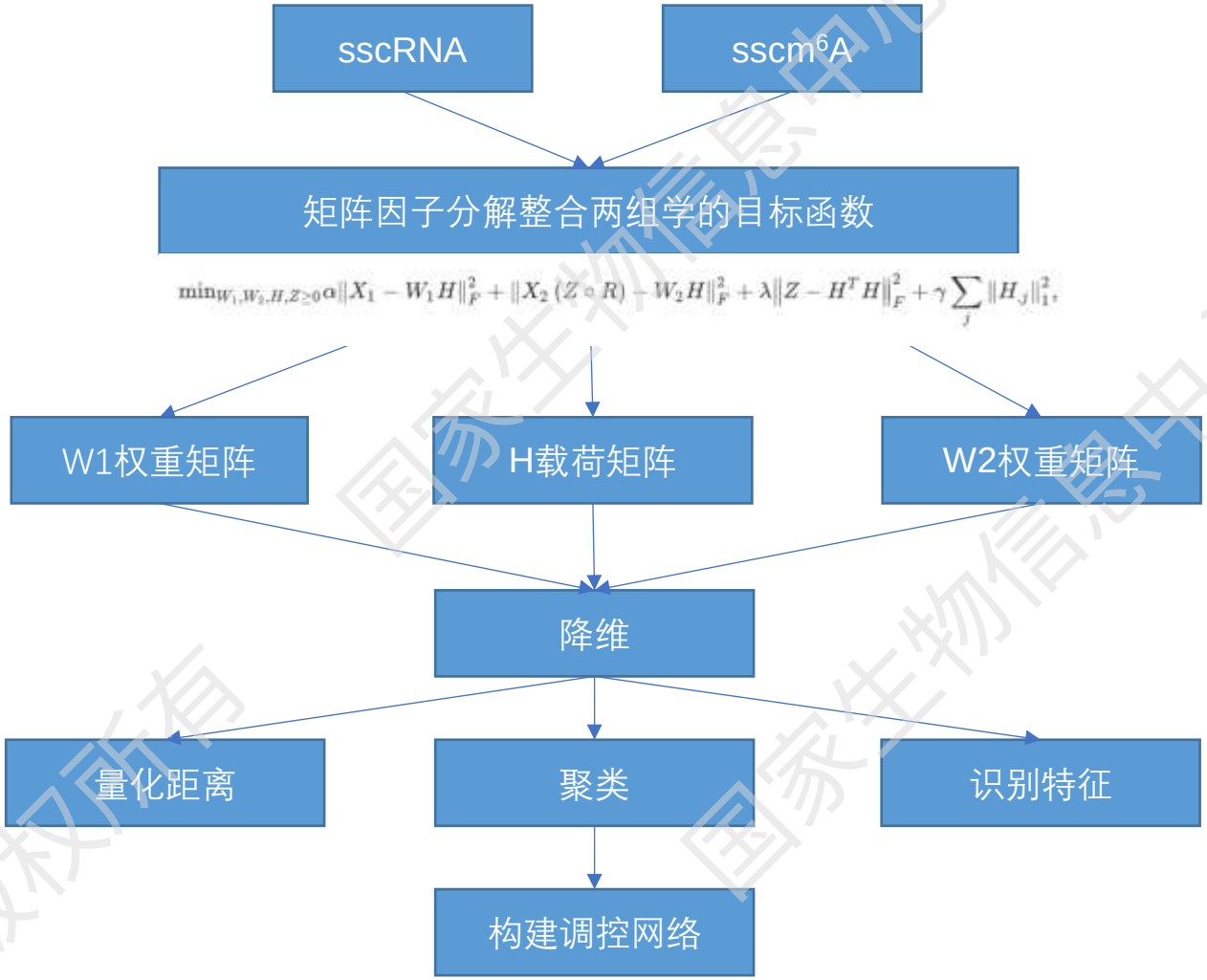
$$\text{ESS}(C_1, C_2) = \sum_{i \in C_1 \cup C_2} \left\| \boldsymbol{x}_i - \frac{1}{|C_1|+|C_2|} \sum_{i \in C_1 \cup C_2} \boldsymbol{x}_i \right\|^2$$

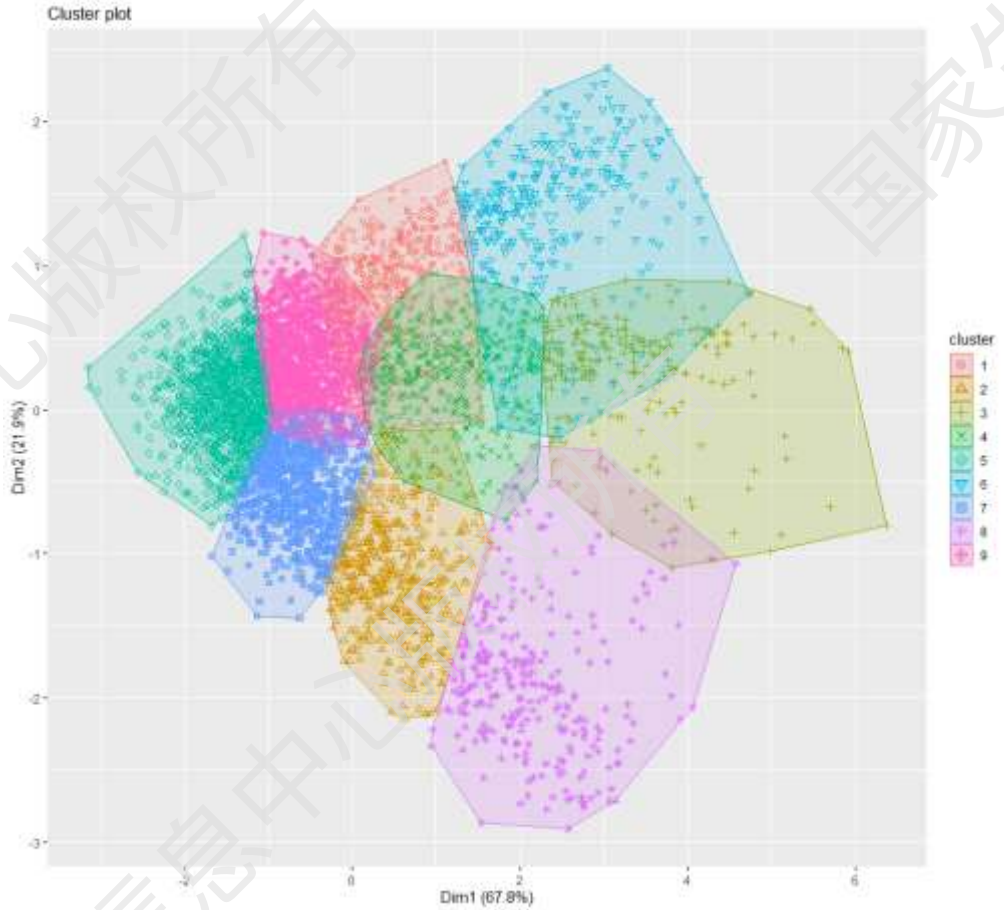
Ward linkage is then found as ESS(C1, C2) − (ESS(C1) + ESS(C2))

3. Quantifying uncertainty (FWER)

achieve significance at any α belonging to the set $\left\{\alpha : \alpha > \frac{N-1}{N_k - 1} p_k\right\}$,

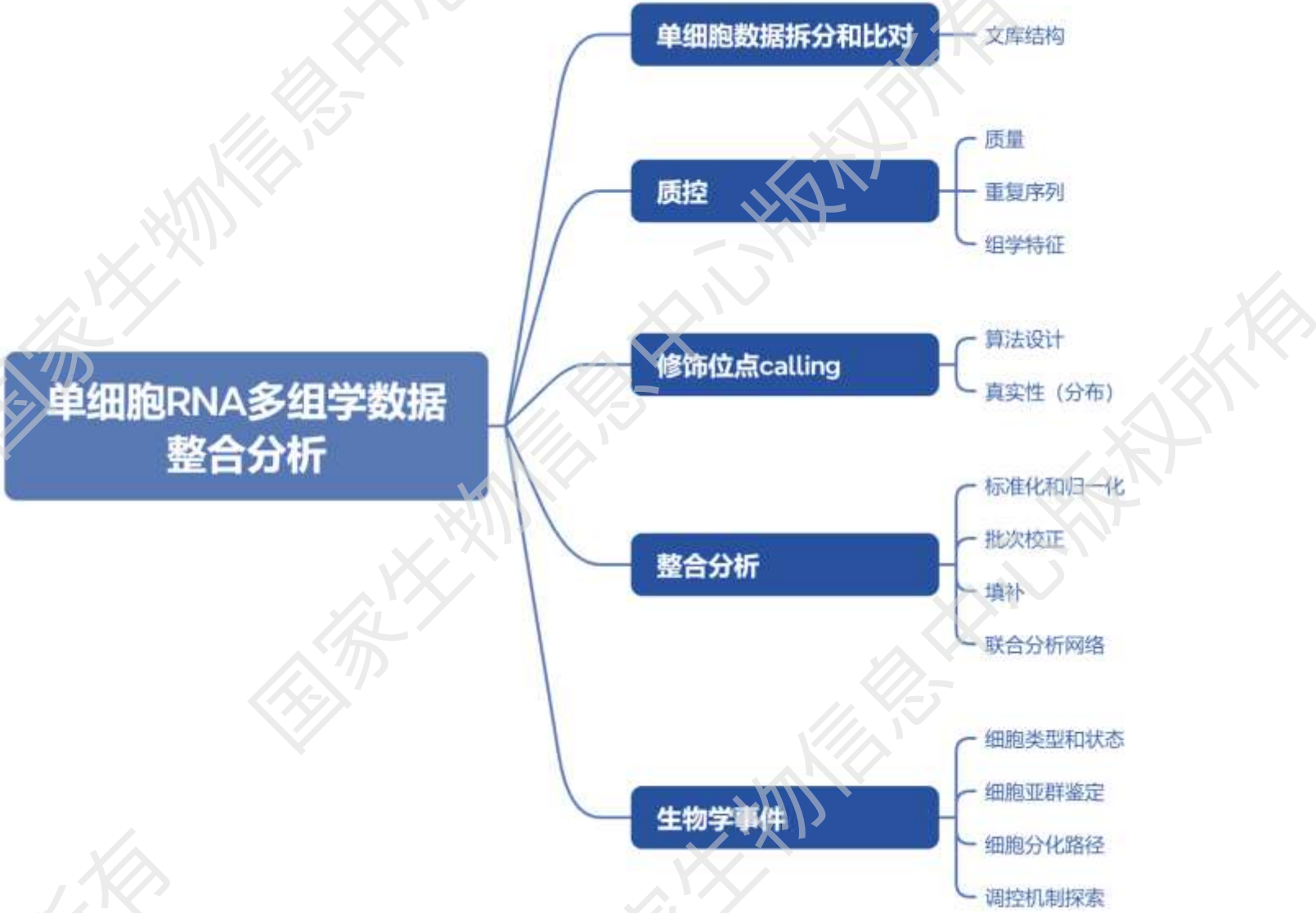
矩阵分解（NMF）整合多组学分析实践





配对多组学数据整合单细胞转录组和修饰组整合结果

单细胞RNA修饰与转录多组学测序数据处理思路



Acknowledgements

Lab

Yang Yun-Gui

Yang Ying

Yao Huan

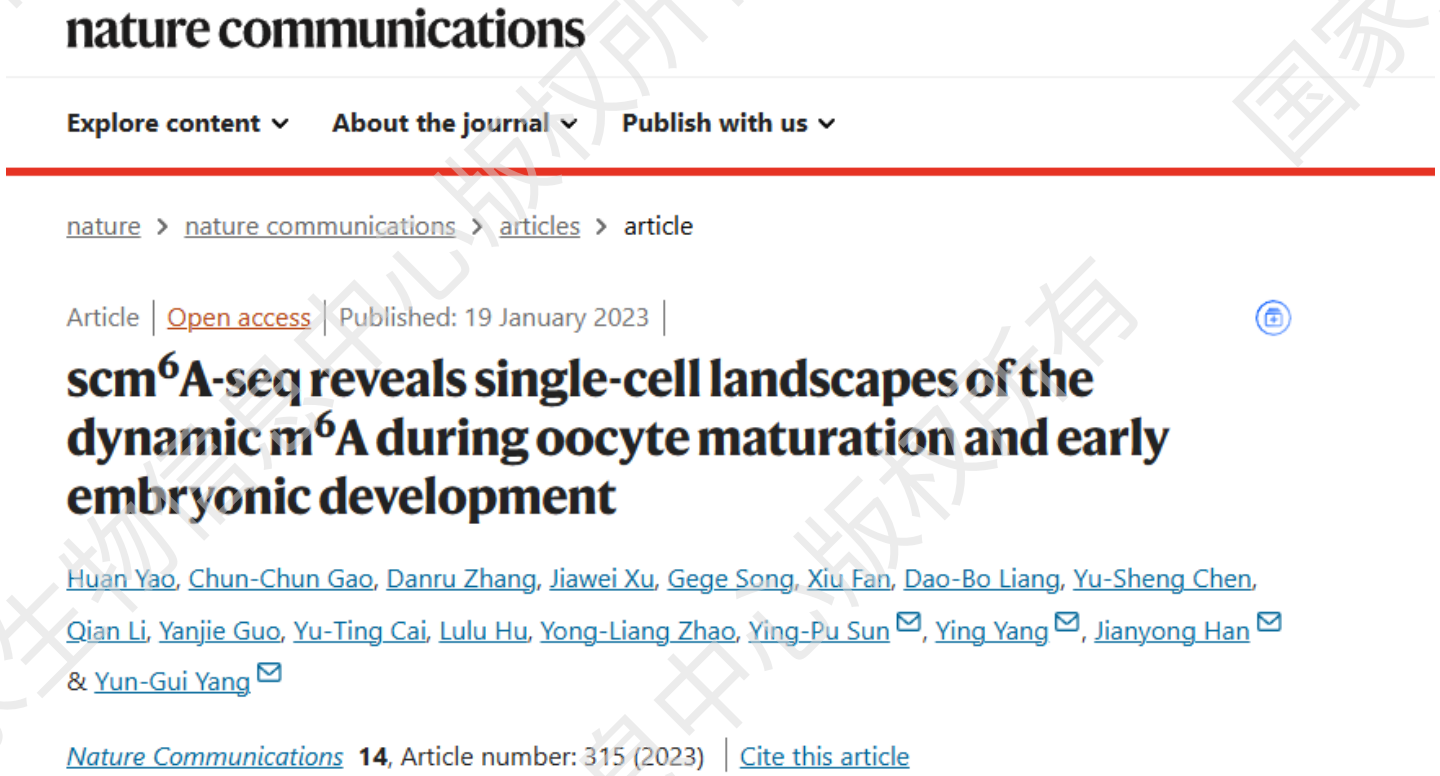
Song Gege

Zhang Danru

Collaborators

Han Jianyong (CAU)

Sun Yingpu (ZZU)



nature communications

Explore content ∨ About the journal ∨ Publish with us ∨

nature › nature communications › articles › article

Article | Open access | Published: 19 January 2023

scm^{6}A-seq reveals single-cell landscapes of the dynamic m^{6}A during oocyte maturation and early embryonic development

Huan Yao, Chun-Chun Gao, Danru Zhang, Jiawei Xu, Gege Song, Xiu Fan, Dao-Bo Liang, Yu-Sheng Chen, Qian Li, Yanjie Guo, Yu-Ting Cai, Lulu Hu, Yong-Liang Zhao, Ying-Pu Sun ✉, Ying Yang ✉, Jianyong Han ✉ & Yun-Gui Yang ✉

Nature Communications **14**, Article number: 315 (2023) | Cite this article