红系细胞分化发育进程中的转录组时序动态变化及关键调控因子识别

小组成员：周振翼、金宏宇、蔡兴安、孟思含

2025年6月11日

文献介绍

北京大学 PEKING UNIVERSITY

RESEARCH ARTICLE SUMMARY

METABOLISM

A glutamine metabolic switch supports erythropoiesis

Junhua Lyu†, Zhimin Gu†, Yuannyu Zhang†, Hieu S. Vu, Christophe Lechauve, Feng Cai, Hui Cao, Julia Keith, Valentina Brancaleoni, Francesca Granata, Irene Motta, Maria Domenica Cappellini, Lily Jun-Shen Huang, Ralph J. DeBerardinis, Mitchell J. Weiss, Min Ni*, Jian Xu*

徐剑

- 本科和硕士毕业于复旦大学，加州大学洛杉矶分校博士，师从Dr. Stephen Smale研究干细胞分化中的表观遗传调控。
- 2008年加入哈佛医学院，博士后师从Dr. Stuart Orkin, 研究论文以第一或共同第一作者发表在Nature和Science等期刊上。
- 2014年加入西南医学中心建立课颖组，2023年加入圣裘德儿童医院。

方向：基因表达调控，肿瘤基因组学，表观遗传学和肿瘤代谢。

Zhimin Gu(顾志敏)

- 2001年考入南京大学生物技术专业，2005年获得学士学位。
- 2005-2011年在中科院健康科学研究所生物化学与分子生物学专业硕博连读。
- 2011-2022年先后在美国犹他大学、爱荷华大学以及西南医学中心以博士后、助理研究员身份进行科研工作。
- 2022年加入中国中医科学院系统医学研究院，任职研究员。

方向：以血液系统为模型，研究造血系统发育以及血液系统恶性疾病的机制。

目 录

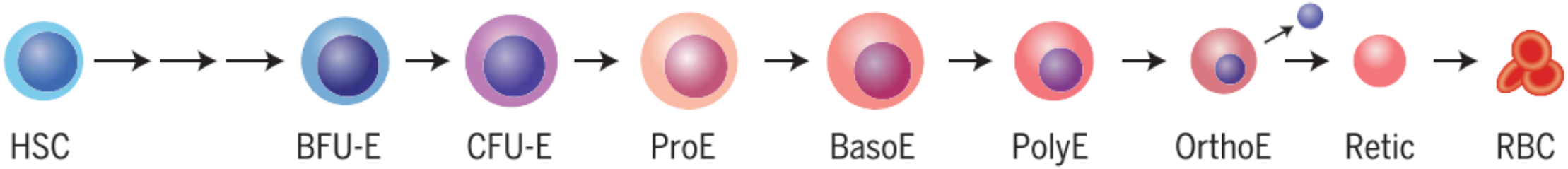
CONTENTS

PART 01

研究背景

Background

红细胞生成



红细胞是人体中数目最多的血细胞，以维持气体交换、代谢平衡等关键生理功能。

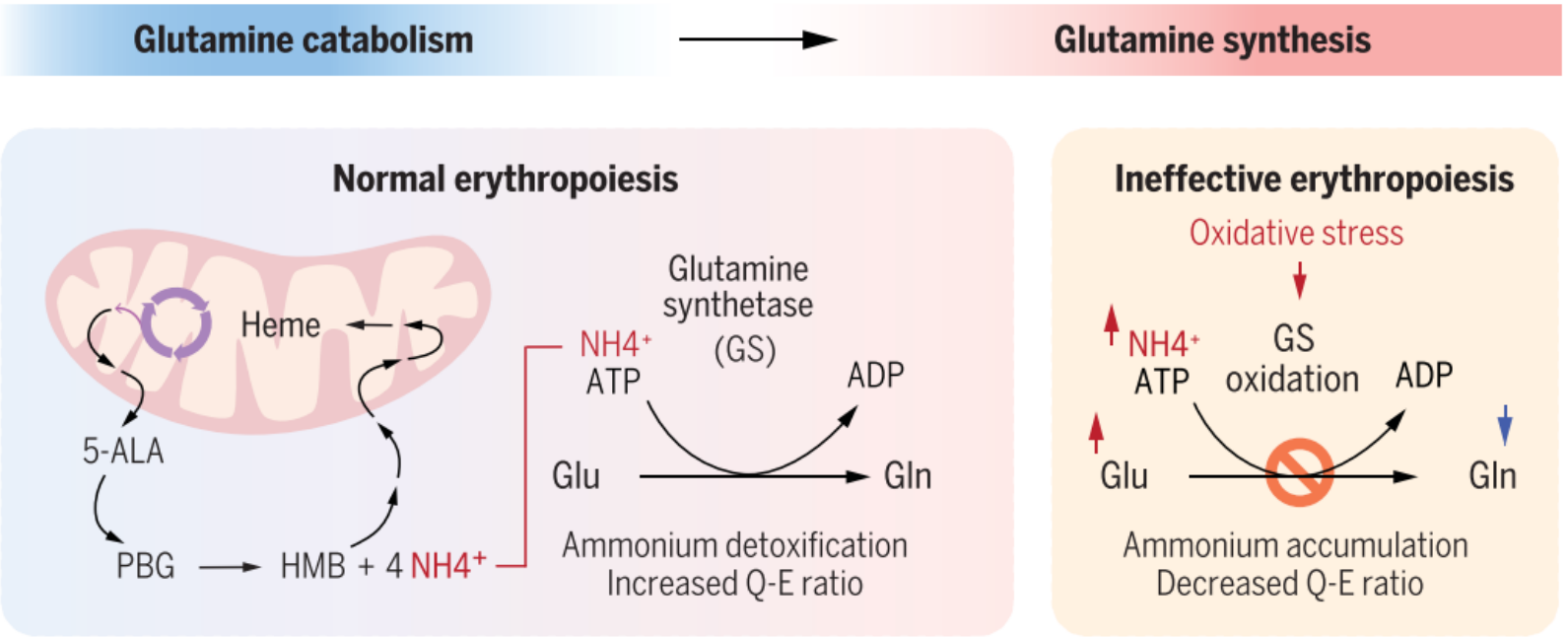
红细胞生成包括红系祖细胞、红系前体细胞和成熟红细胞阶段。

红系祖细胞 (BFU-E) → 红系集落形成单位(CFU-E) → 原始红细胞 → 早幼红细胞 → 中幼红细胞 → 晚幼红细胞 → 网织红细胞 → 成熟红细胞

这一过程经历一系列显著的形态和功能变化，包括形成特化的细胞膜、丧失大部分细胞器、蛋白质的浓缩以及大量血红蛋白的合成。

Lyu et al, 2024, *Science*

理解红细胞发育的动态改变对疾病的研究具有极大意义



研究人员在红细胞定向分化和成熟过程中各个阶段分析了基因转录和代谢水平的变化，并通过整合转录组与代谢组的网络分析，发现谷氨酰胺分解代谢向合成代谢的转换与红细胞分化和成熟高度相关，为疾病治疗提供了新方向。

因此研究红系细胞分化发育进程中的转录组时序动态变化及关键调控因子，对于深入理解红细胞生成的分子机制、揭示贫血等血液疾病的发病机制，以及开发新的治疗策略（如通过调控关键转录因子或信号通路来改善红细胞生成）具有重要的理论和临床应用价值。

Lyu et al, 2024, *Science*

PART 02

实验结果

Results

RNAseq原始测序数据上游分析

利用conda搭建分析环境

```
conda create RNA_seq python=3.10
conda create -n RNA_seq python=3.10
conda install multiqc
conda install samtools
conda install fastp
conda install fastqc
conda install star
conda install salmon
```

获取红细胞的不同发育阶段的八组样本原始测序数据

You selected 8 Items

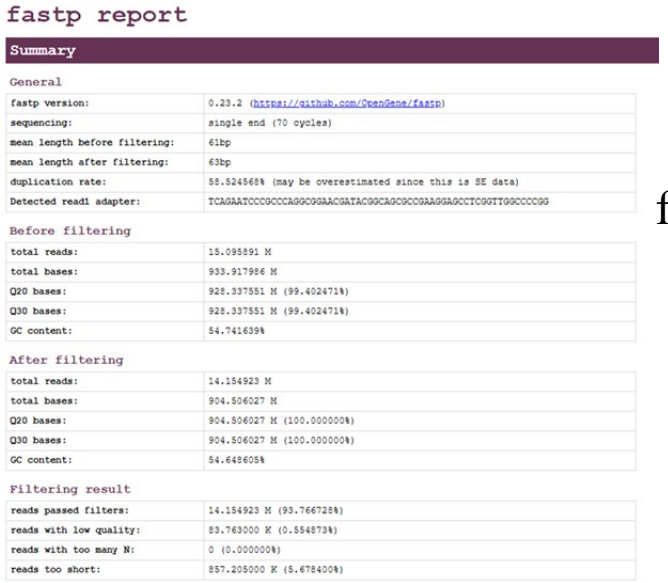
Search within results

	Run	BioSample	Sample Name	source_name
1	SRR23615254	SAMN33439508	GSM7062777	BFU-E
2	SRR23615251	SAMN33439506	GSM7062779	CFU-E
3	SRR23615246	SAMN33439498	GSM7062787	ProE
4	SRR23615240	SAMN33439494	GSM7062791	BasoE
5	SRR23615238	SAMN33439491	GSM7062794	PolyE
6	SRR23615233	SAMN33439486	GSM7062799	OrthoE
7	SRR23615219	SAMN33439481	GSM7062804	Retic
8	SRR23615222	SAMN33439474	GSM7062811	RBC

利用基因组序列和注释创建小鼠参考基因组GRCm38

GRCm38		2025-06-09 …	caixa
GRCm38_transcripts.fa	243 688	2025-06-09 …	caixa
Mus_musculus.GRCm38.99.gtf	788 700	2025-06-08 …	caixa
Mus_musculus.GRCm38.dna.p…	2 711 318	2025-06-08 …	caixa
Mus_musculus.GRCm38.dna.p…	2	2025-06-09 …	caixa

RNAseq原始测序数据上游分析



fastp report

Summary

General

fastp version:	0.23.2 (https://github.com/OpenGene/fastp)
sequencing:	single end (70 cycles)
mean length before filtering:	61bp
mean length after filtering:	63bp
duplication rate:	58.524568% (may be overestimated since this is SE data)
Detected read1 adapter:	TCAGAATCCCGCCCAGGCGGAACGATACGGCAGCGCCGAAGGAGCCTCGGTTGGCCCCGG

Before filtering

total reads:	15.095891 M
total bases:	933.917986 M
Q20 bases:	928.337551 M (99.402471%)
Q30 bases:	928.337551 M (99.402471%)
GC content:	54.741639%

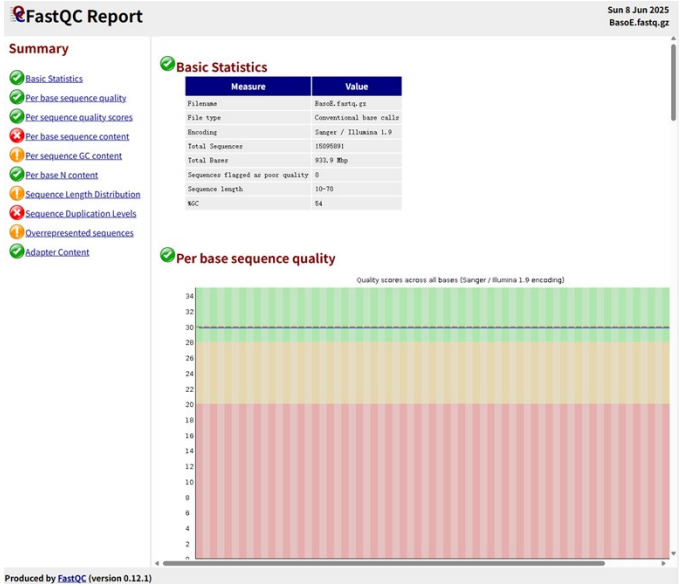
After filtering

total reads:	14.154923 M
total bases:	904.506027 M
Q20 bases:	904.506027 M (100.000000%)
Q30 bases:	904.506027 M (100.000000%)
GC content:	54.648605%

Filtering result

reads passed filters:	14.154923 M (93.766728%)
reads with low quality:	83.763000 K (0.554873%)
reads with too many N:	0 (0.000000%)
reads too short:	857.205000 K (5.678400%)

fastp质控、去接头



FastQC Report

Sun 8 Jun 2025
BasoE.fastq.gz

Summary

- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content

Basic Statistics

Measure	Value
Filename	BasoE.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	15095891
Total Bases	933.9 Mbp
Sequences flagged as poor quality	0
Sequence length	10-70
%GC	54

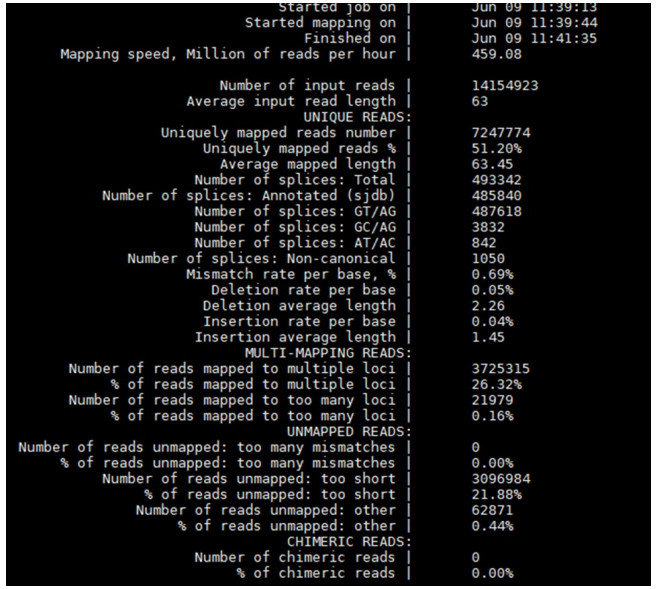
Per base sequence quality

Quality scores across all bases (Sanger / Illumina 1.9 encoding)

Produced by FastQC (version 0.12.1)

fastqc质控数据可视化

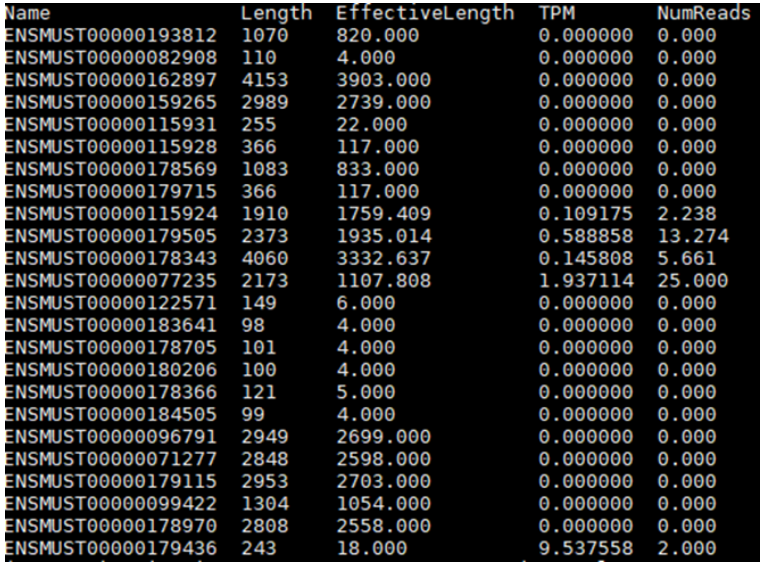
STAR比对



```
                          Started job on |       Jun 09 11:39:13
                      Started mapping on |       Jun 09 11:39:44
                             Finished on |       Jun 09 11:41:35
       Mapping speed, Million of reads per hour |       459.08

                  Number of input reads |       14154923
              Average input read length |       63
                            UNIQUE READS:
           Uniquely mapped reads number |       7247774
                Uniquely mapped reads % |       51.20%
                  Average mapped length |       63.45
               Number of splices: Total |       493342
    Number of splices: Annotated (sjdb) |       485840
               Number of splices: GT/AG |       487618
               Number of splices: GC/AG |       3832
               Number of splices: AT/AC |       842
       Number of splices: Non-canonical |       1050
              Mismatch rate per base, % |       0.69%
                 Deletion rate per base |       0.05%
                Deletion average length |       2.26
                Insertion rate per base |       0.04%
               Insertion average length |       1.45
                         MULTI-MAPPING READS:
      Number of reads mapped to multiple loci |       3725315
           % of reads mapped to multiple loci |       26.32%
      Number of reads mapped to too many loci |       21979
           % of reads mapped to too many loci |       0.16%
                              UNMAPPED READS:
 Number of reads unmapped: too many mismatches |       0
      % of reads unmapped: too many mismatches |       0.00%
           Number of reads unmapped: too short |       3096984
                % of reads unmapped: too short |       21.88%
            Number of reads unmapped: other |       62871
                % of reads unmapped: other |       0.44%
                              CHIMERIC READS:
                   Number of chimeric reads |       0
                        % of chimeric reads |       0.00%
```

salmon定量转录本表达量



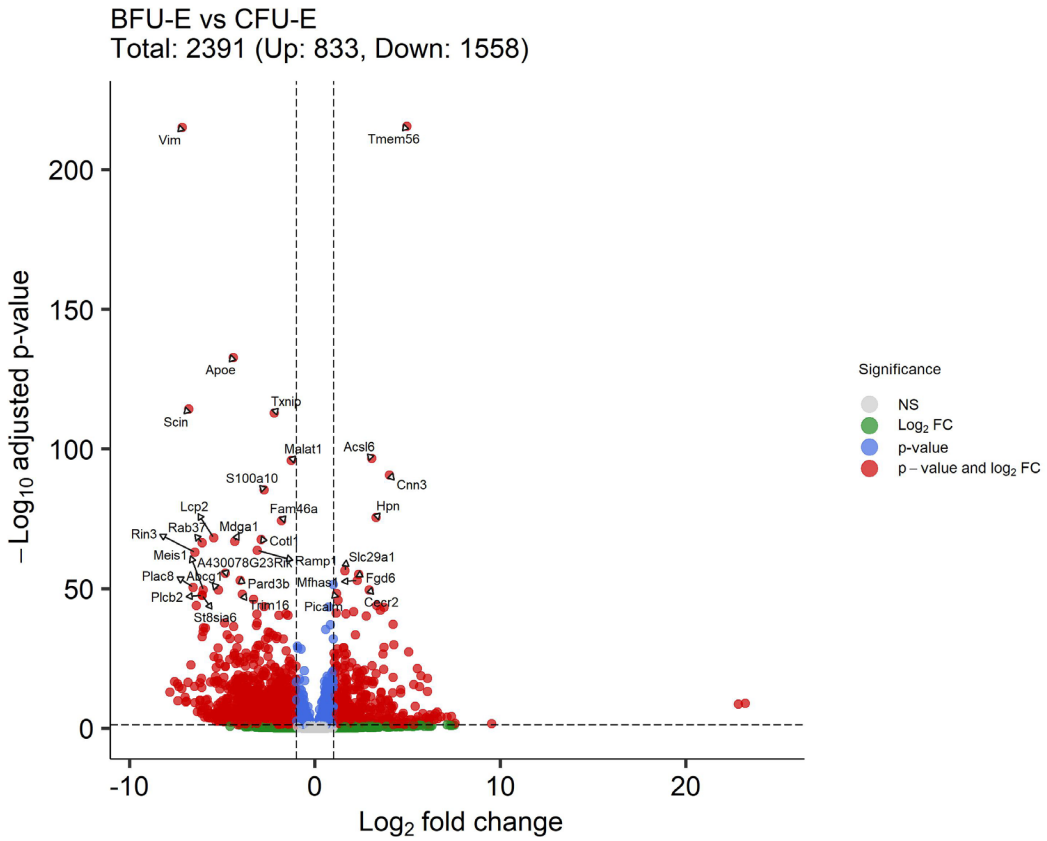
Name	Length	EffectiveLength	TPM	NumReads
ENSMUST00000193812	1070	820.000	0.000000	0.000
ENSMUST00000082908	110	4.000	0.000000	0.000
ENSMUST00000162897	4153	3903.000	0.000000	0.000
ENSMUST00000159265	2989	2739.000	0.000000	0.000
ENSMUST00000115931	255	22.000	0.000000	0.000
ENSMUST00000115928	366	117.000	0.000000	0.000
ENSMUST00000178569	1083	833.000	0.000000	0.000
ENSMUST00000179715	366	117.000	0.000000	0.000
ENSMUST00000115924	1910	1759.409	0.109175	2.238
ENSMUST00000179505	2373	1935.014	0.588858	13.274
ENSMUST00000178343	4060	3332.637	0.145808	5.661
ENSMUST00000077235	2173	1107.808	1.937114	25.000
ENSMUST00000122571	149	6.000	0.000000	0.000
ENSMUST00000183641	98	4.000	0.000000	0.000
ENSMUST00000178705	101	4.000	0.000000	0.000
ENSMUST00000180206	100	4.000	0.000000	0.000
ENSMUST00000178366	121	5.000	0.000000	0.000
ENSMUST00000184505	99	4.000	0.000000	0.000
ENSMUST00000096791	2949	2699.000	0.000000	0.000
ENSMUST00000071277	2848	2598.000	0.000000	0.000
ENSMUST00000179115	2953	2703.000	0.000000	0.000
ENSMUST00000099422	1304	1054.000	0.000000	0.000
ENSMUST00000178970	2808	2558.000	0.000000	0.000
ENSMUST00000179436	243	18.000	9.537558	2.000

红系细胞相邻发育阶段的差异表达分析（1）

北京大学 PEKING UNIVERSITY

红系祖细胞（BFU-E）向红系前体细胞（CFU-E）转变的差异表达基因火山图

- 火山图原理：X轴表示log2倍数变化，Y轴表示-log10（校正P值）
- 颜色编码：灰色（非显著差异基因）绿色（显著但变化幅度小的基因）
  蓝色（变化幅度大但不显著的基因）红色（显著差异表达基因）
- 统计阈值：padj < 0.05，|log2FC| > 1
- 每张图标注最显著的30个差异基因



DESeq2是一种用于分析RNA-Seq数据的强大工具，旨在**识别基因在不同条件或不同样本之间的差异表达**。该方法基于负二项分布模型，考虑了RNA-Seq数据的离散性和高度异质性，以及批次效应和样本之间的技术变异。

关键差异表达基因及功能：

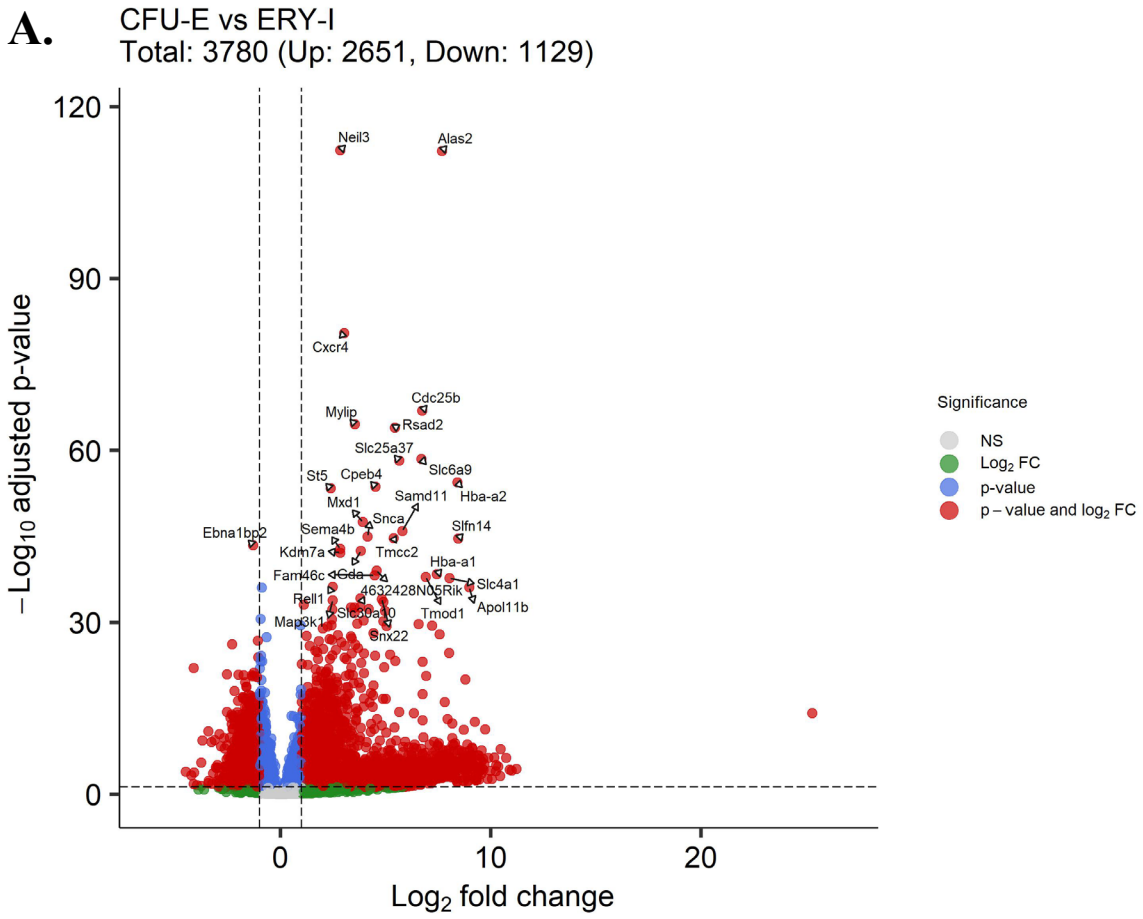
- Tmem56：细胞膜整合蛋白，参与膜组织和细胞内物质转运
- Acsl5：催化长链脂肪酸活化的酶，参与脂质代谢和能量产生
- Cnn3：肌动蛋白结合蛋白，参与细胞骨架组织和钙离子介导的信号传导
- Hpn：跨膜丝氨酸蛋白酶，参与细胞生长和组织重塑
- Slc29a1：膜转运蛋白，负责核苷和核苷类似物的跨膜转运
- Vim：中间丝蛋白，构成细胞骨架的主要成分，维持细胞结构完整性
- Apoe：参与脂质运输和代谢，介导胆固醇稳态和脂蛋白颗粒清除
- Scin：钙调节的肌动蛋白切割蛋白，调控细胞运动和形态变化
- Lcp2：T细胞和B细胞信号转导的接头蛋白，参与免疫响应
- Meis1：转录因子，调控细胞增殖、分化和器官形成的基因表达

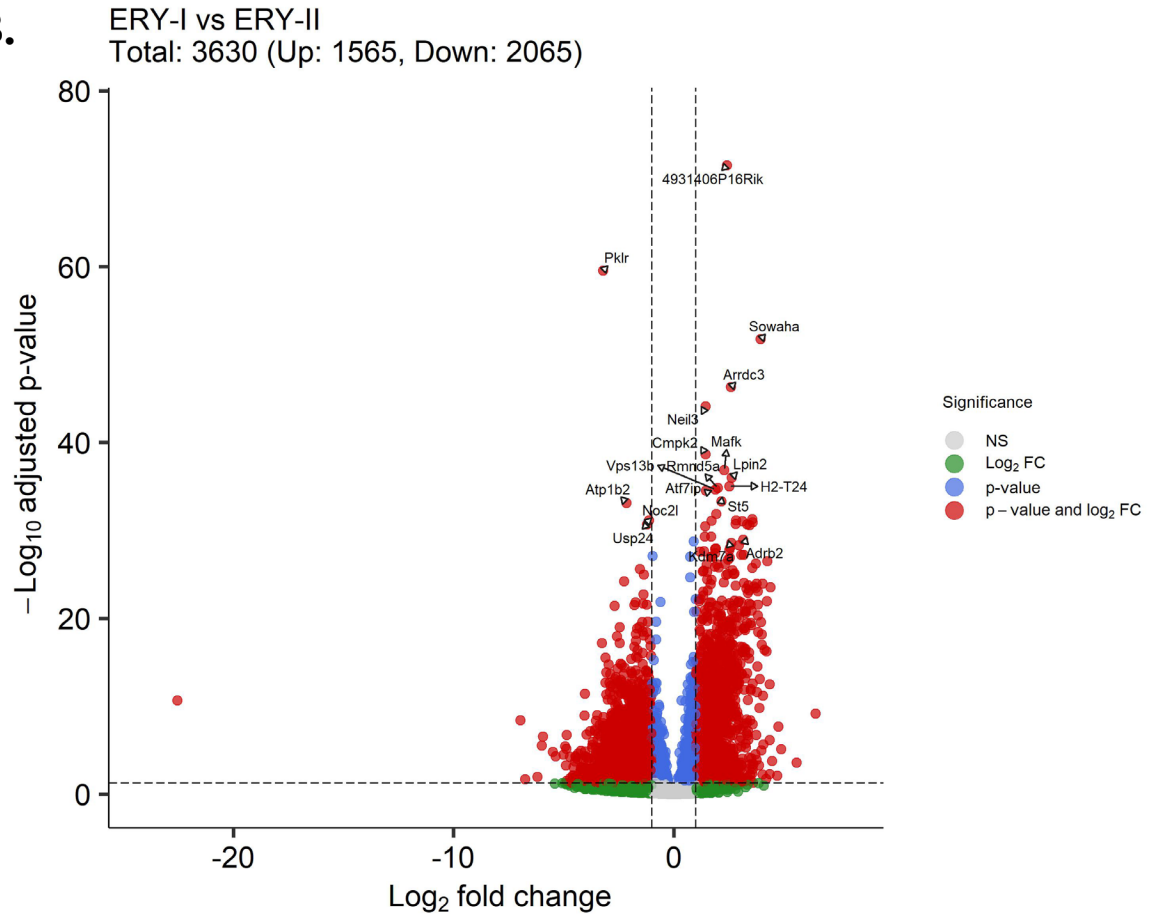
红系细胞相邻发育阶段的差异表达分析（1）

北京大学 PEKING UNIVERSITY

A.红系前体细胞（CFU-E）向前红细胞（ERY-I）分化的差异表达基因火山图

B.前红细胞（ERY-I）向嗜碱性红细胞（ERY-II）转变的差异表达基因火山图





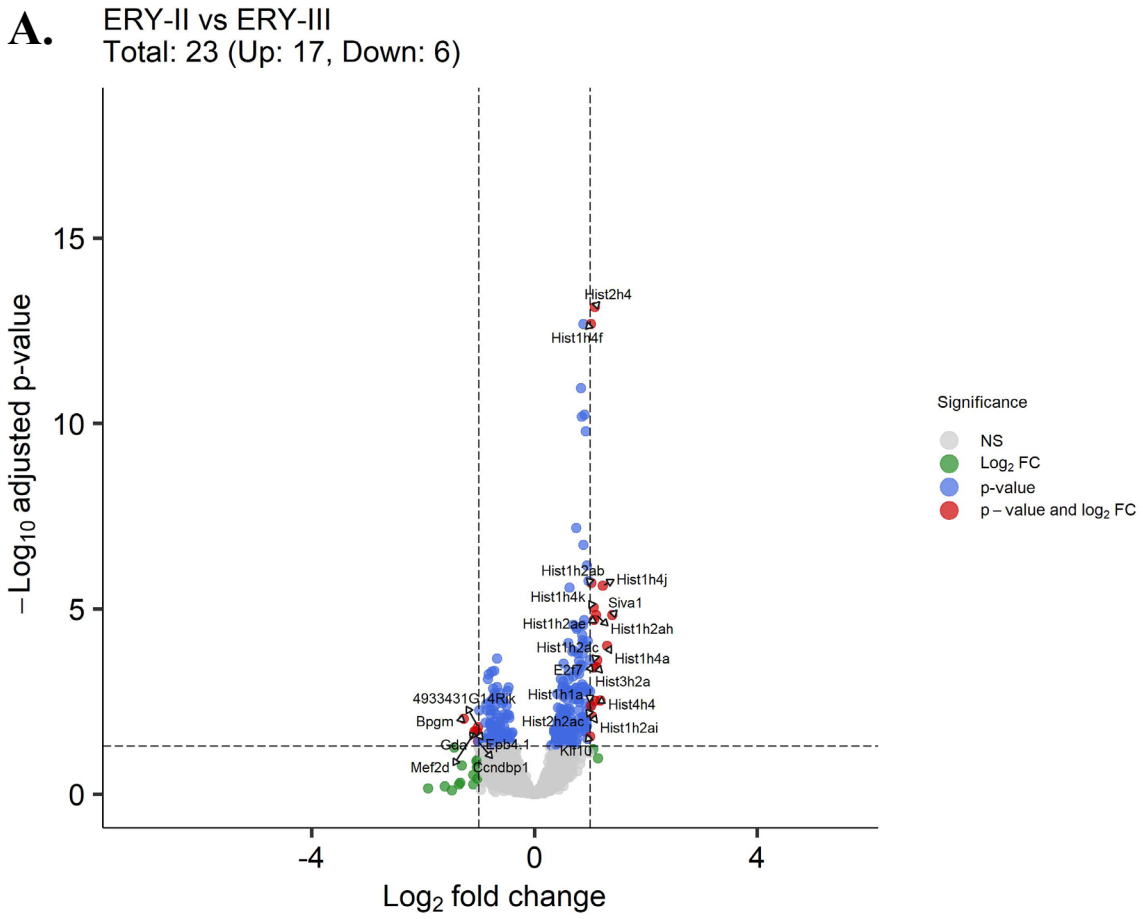
红系细胞相邻发育阶段的差异表达分析（1）

北京大学 PEKING UNIVERSITY

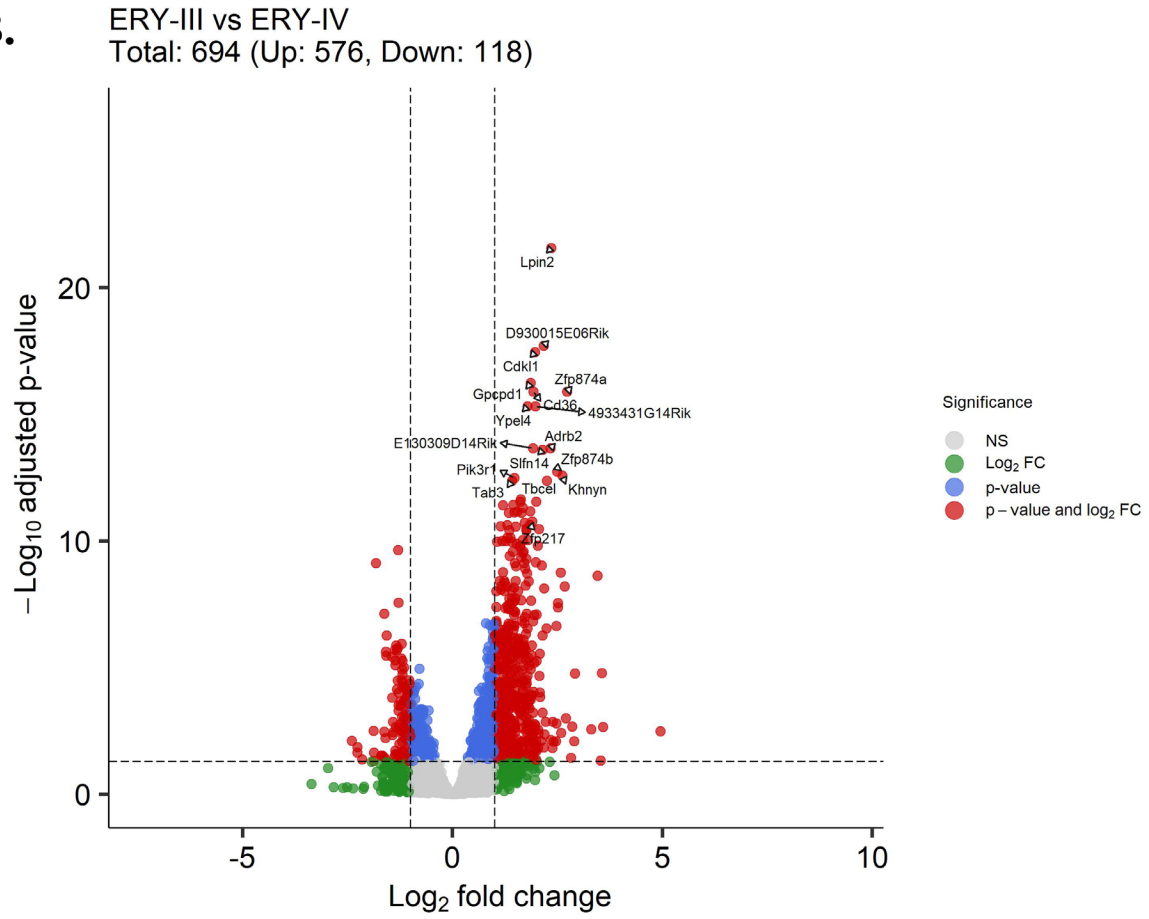
A.嗜碱性红细胞（ERY-II）向多染性红细胞（ERY-III）发育的差异表达基因火山图

B.多染性红细胞（ERY-III）向正染性红细胞（ERY-IV）分化的差异表达基因火山图

A.



B.

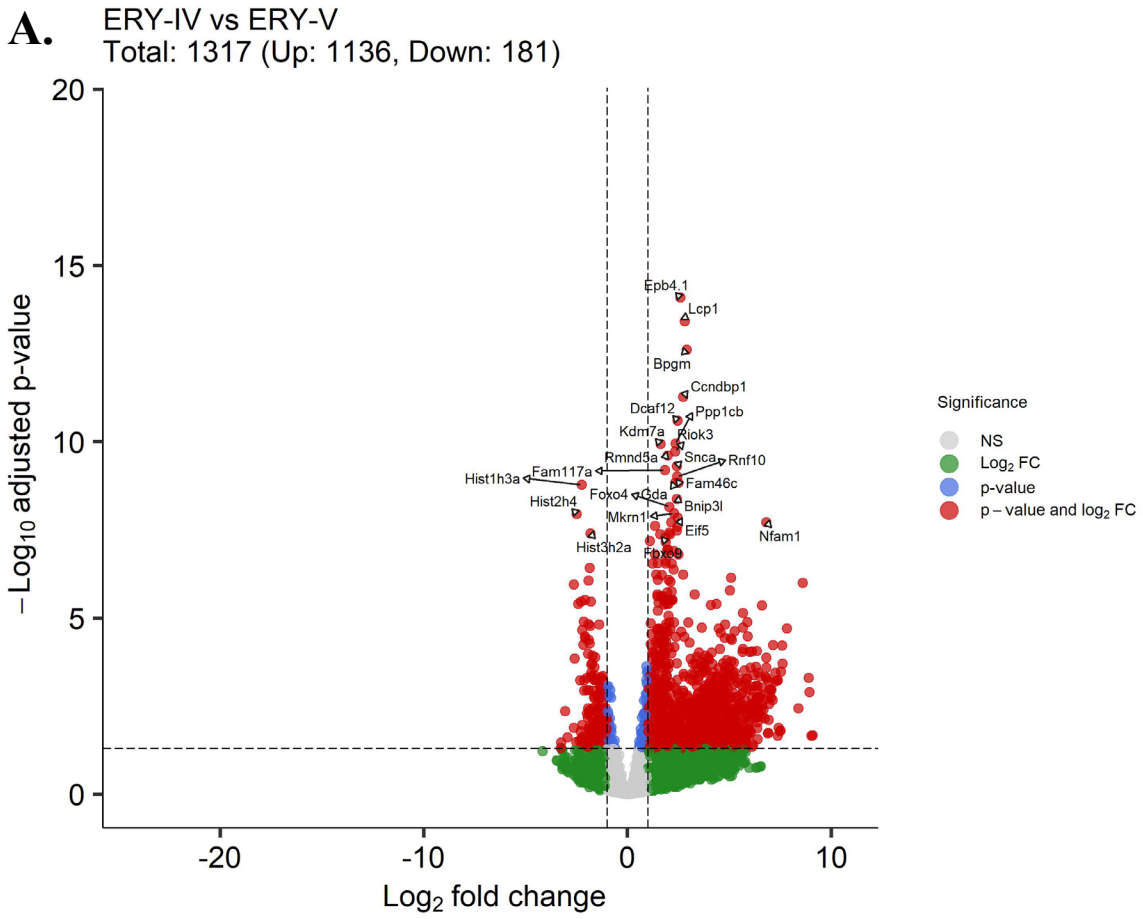


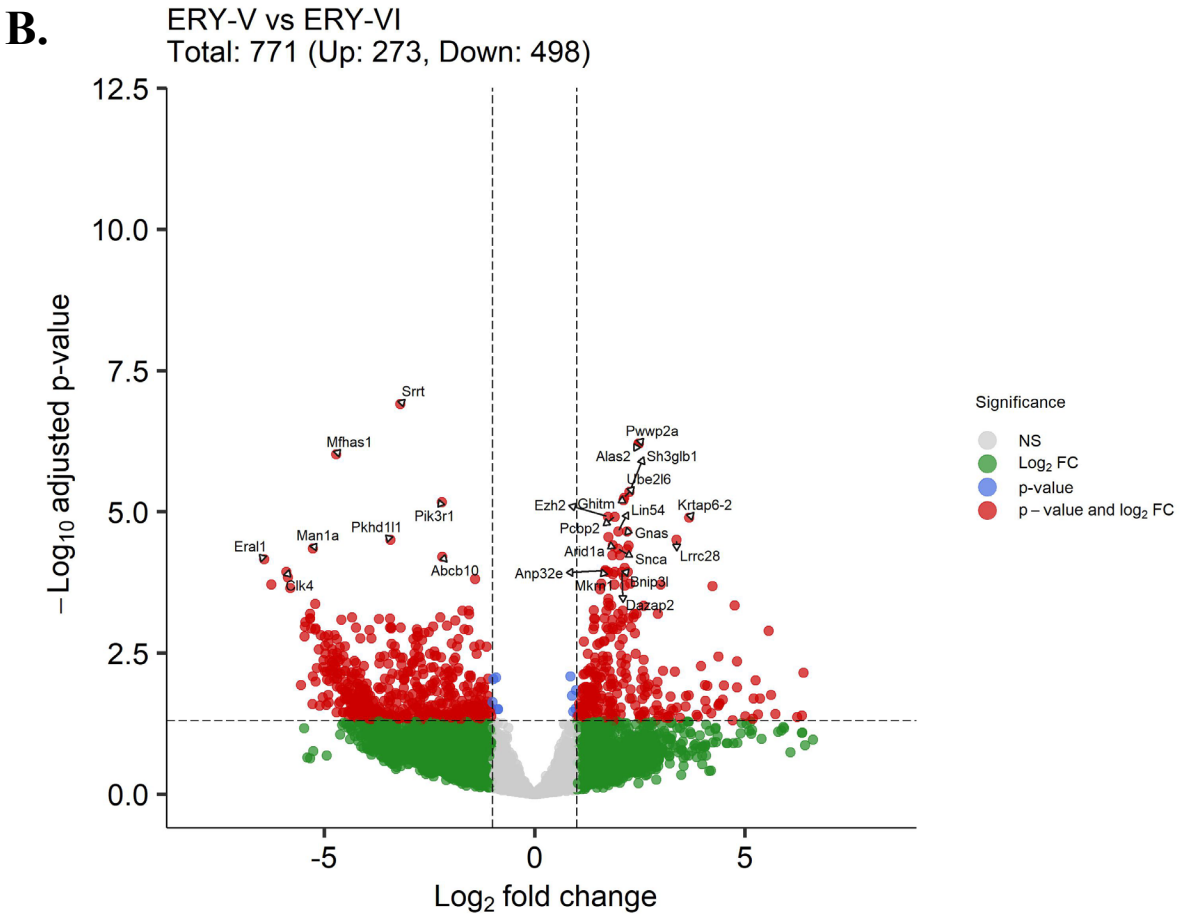
红系细胞相邻发育阶段的差异表达分析（1）

北京大学 PEKING UNIVERSITY

A.正染性红细胞（ERY-IV）向网织红细胞（ERY-V）转变的差异表达基因火山图

B.网织红细胞（ERY-V）向成熟红细胞（ERY-VI）发育的差异表达基因火山图



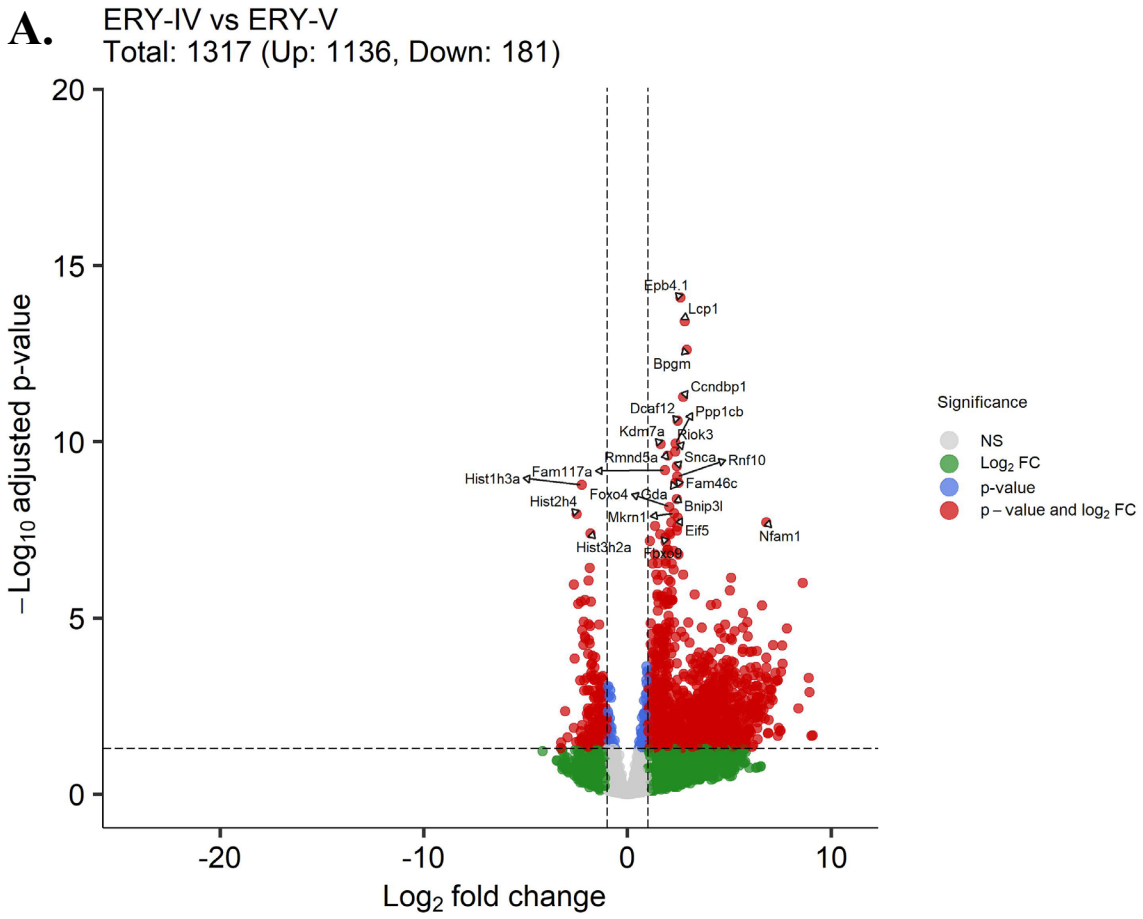


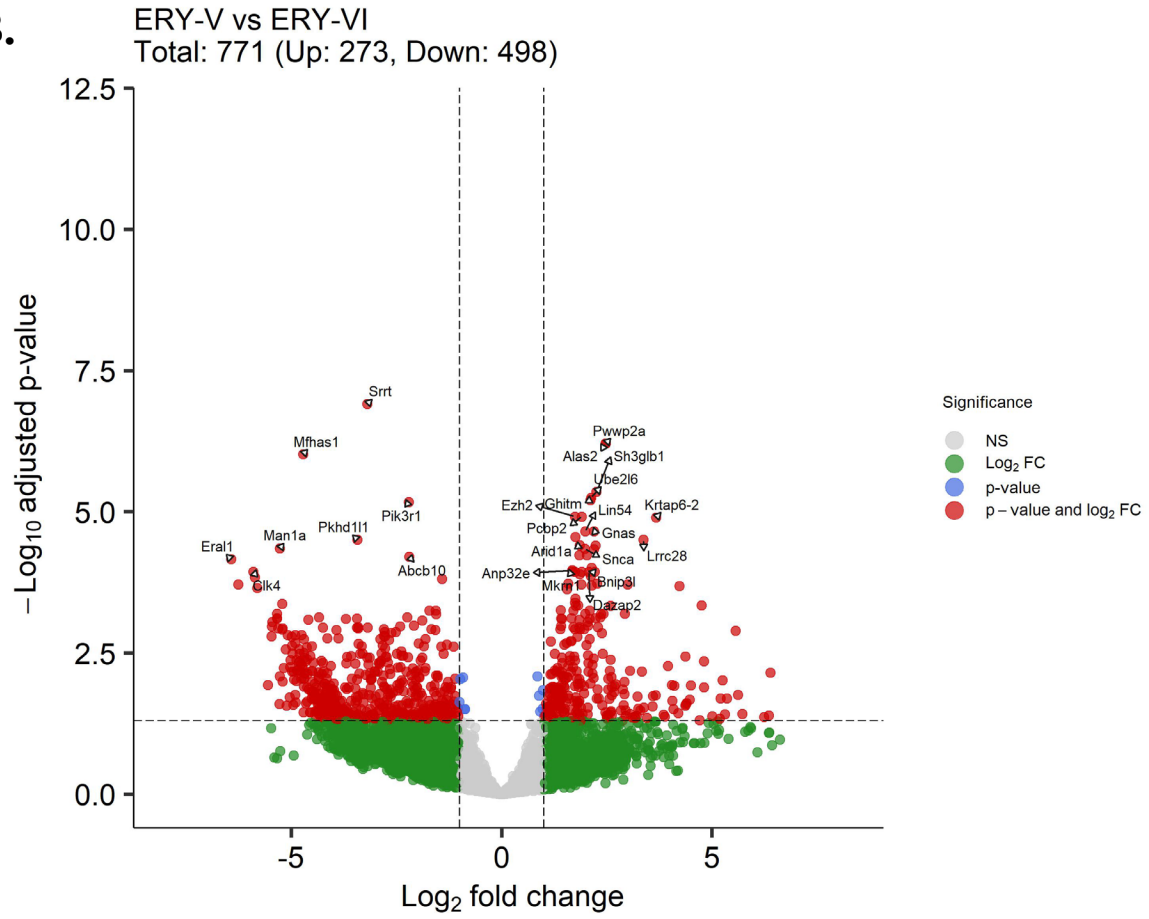
红系细胞相邻发育阶段的差异表达分析（1）

北京大学 PEKING UNIVERSITY

A.正染性红细胞（ERY-IV）向网织红细胞（ERY-V）转变的差异表达基因火山图

B.网织红细胞（ERY-V）向成熟红细胞（ERY-VI）发育的差异表达基因火山图



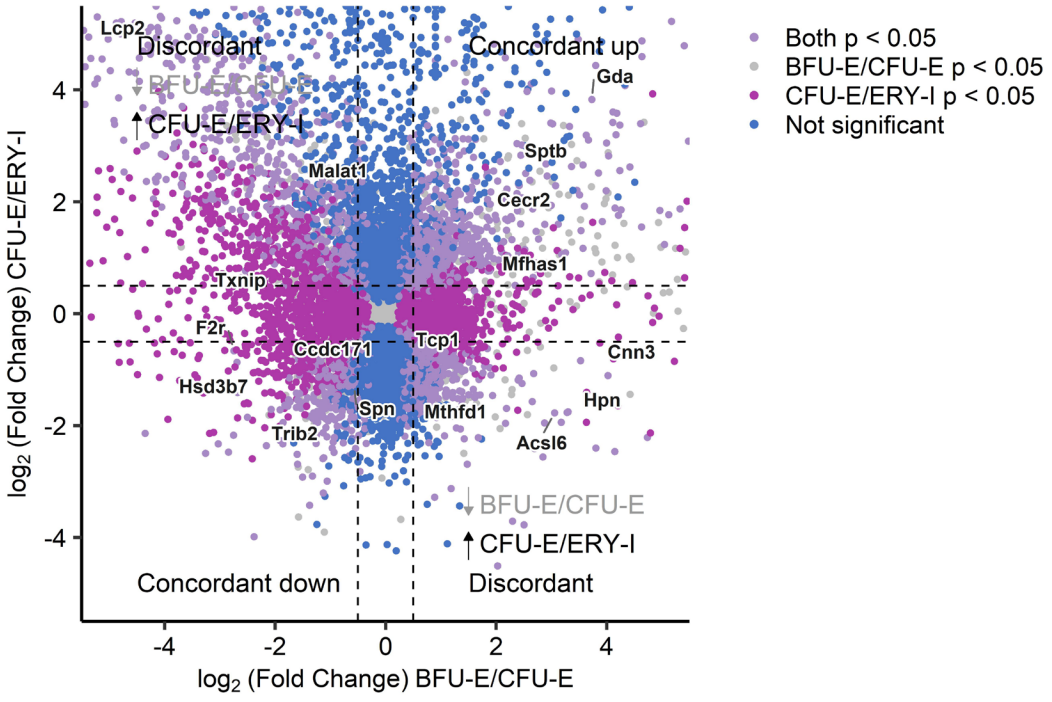


红系细胞相邻发育阶段的差异表达分析（2）

北京大学 PEKING UNIVERSITY

BFU-E向CFU-E与CFU-E向ERY-I转变的基因表达变化协同性散点图

- FC-FC散点图原理：X轴表示第一个转变过程的log2倍数变化，Y轴表示第二个转变过程的log2倍数变化
- 颜色编码：紫色（两个比较中均显著差异表达的基因）红色（仅X轴比较中显著差异表达的基因）
  蓝色（仅Y轴比较中显著差异表达的基因）灰色（两个比较中均无显著差异的基因）
- 统计阈值：padj < 0.05，log2FC = ±0.5
- 每张图标注每个象限中最显著的5个差异基因



FC-FC散点图是一种变异火山图，用于比较基因在不同条件下的表达变化相关性。不同于传统火山图(横轴为log2FC，纵轴为-log10p值)，FC-FC散点图的横轴和纵轴均为log2FC值，可同时展示基因在两个不同发育转变过程中的表达变化。

关键差异表达基因及功能：

- Lcp2：淋巴细胞胞质蛋白2，参与造血细胞信号转导
- Gda：鸟嘌呤脱氨酶，调控嘌呤代谢和核苷酸平衡
- Mthfd1：亚甲基四氢叶酸脱氢酶1，支持DNA合成和细胞增殖
- Sptb：β-谱蛋白，红细胞膜骨架的主要组分
- Cnn3：钙调蛋白3，参与细胞骨架重组和形态变化
- Hpn：肝素酶，丝氨酸蛋白酶，参与细胞生长调控
- Malat1：长链非编码RNA，调控基因表达和细胞分化
- Cdc17：DNA聚合酶α催化亚基，参与DNA复制
- Acsl6：长链脂肪酰辅酶A合成酶6，参与膜脂质合成
- Trpb2：闯缝连接蛋白β2，介导细胞间通讯

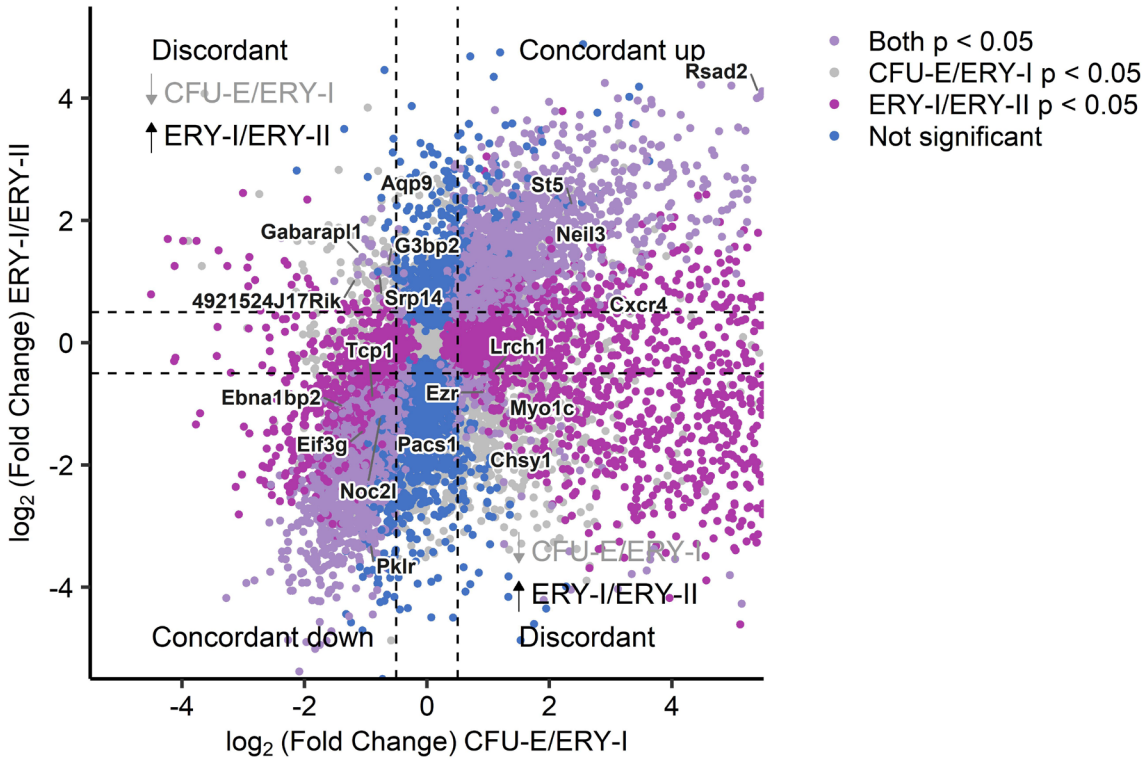
红系细胞相邻发育阶段的差异表达分析（2）

北京大学 PEKING UNIVERSITY

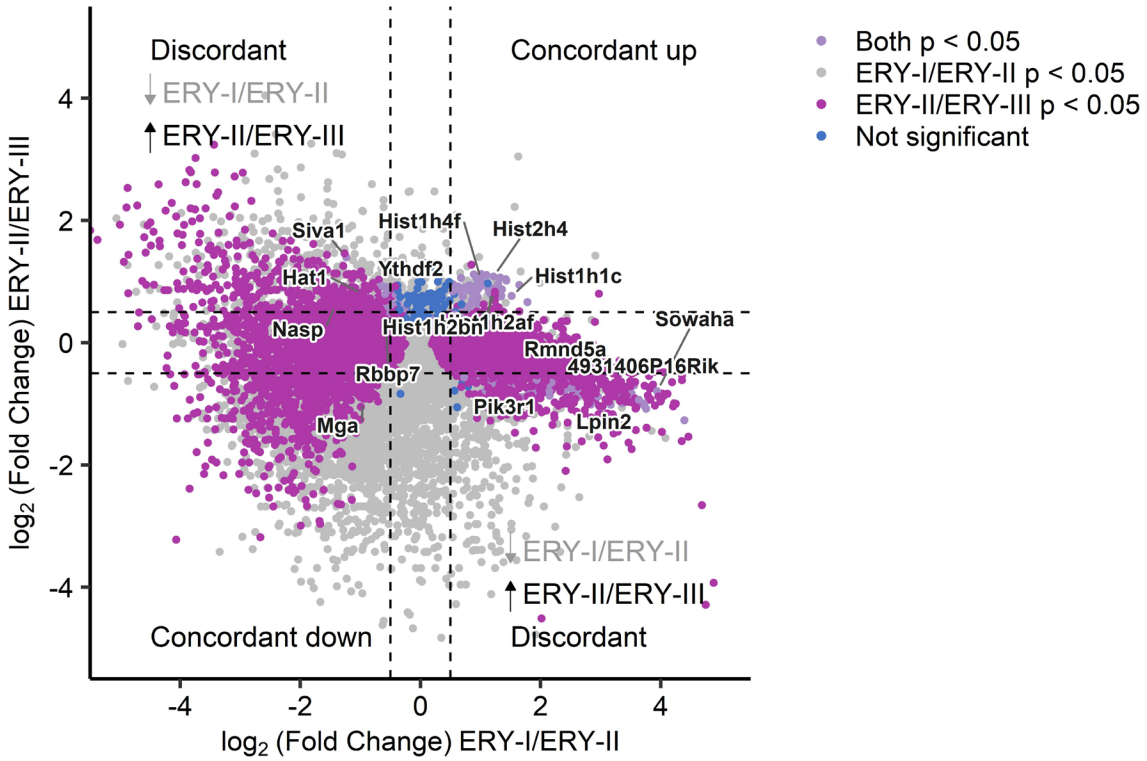
A. CFU-E向ERY-I与ERY-I向ERY-II转变的基因表达变化协同性散点图

B. ERY-I向ERY-II与ERY-II向ERY-III转变的基因表达变化协同性散点图

A.



B.

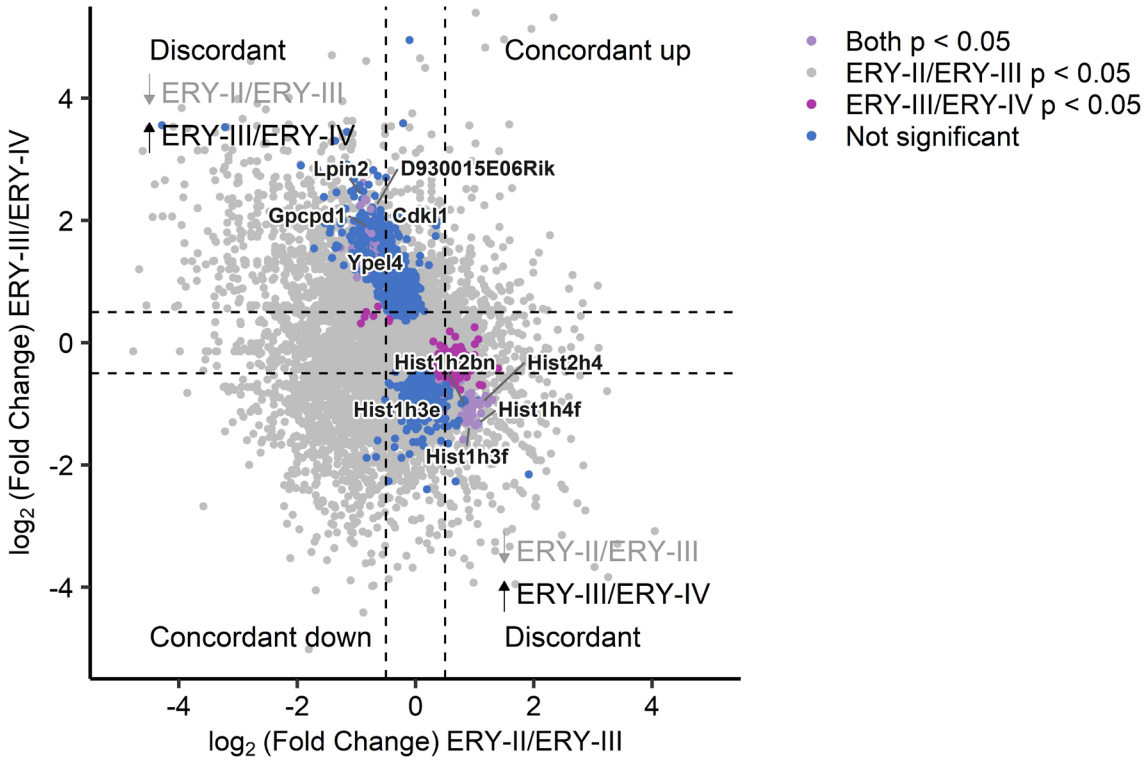


红系细胞相邻发育阶段的差异表达分析（2）

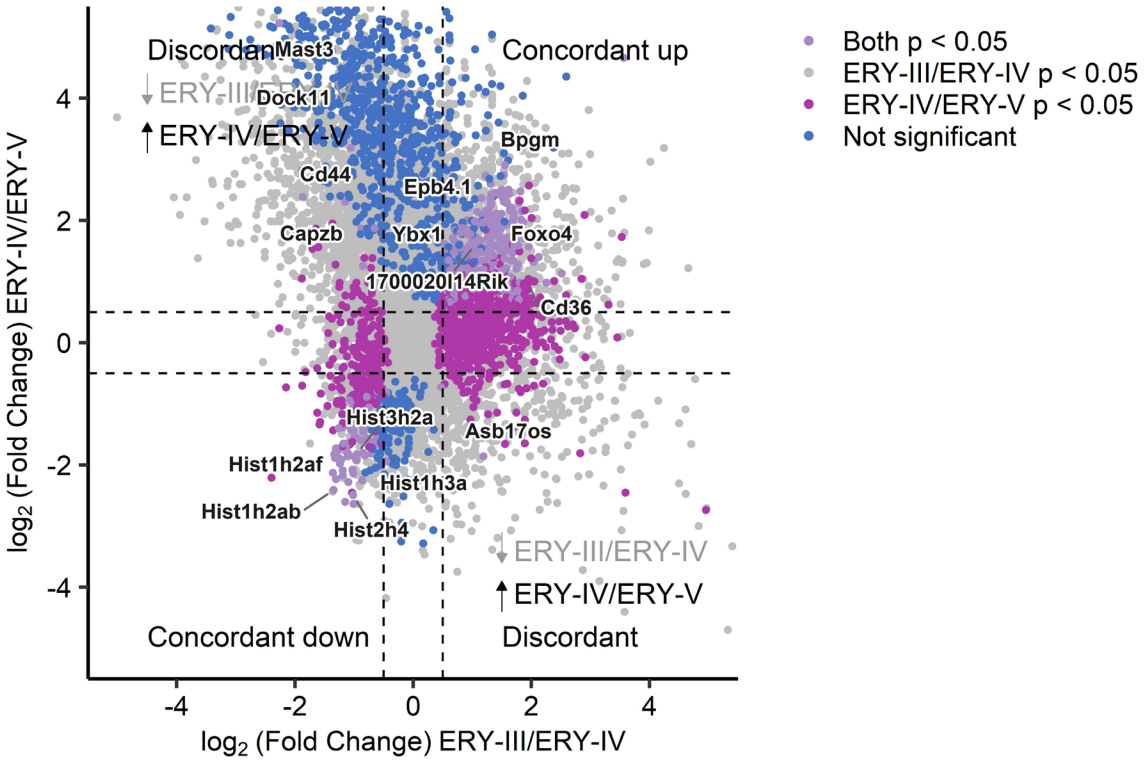
A. ERY-II向ERY-III与ERY-III向ERY-IV转变的基因表达变化协同性散点图

B. ERY-III向ERY-IV与ERY-IV向ERY-V转变的基因表达变化协同性散点图

A.



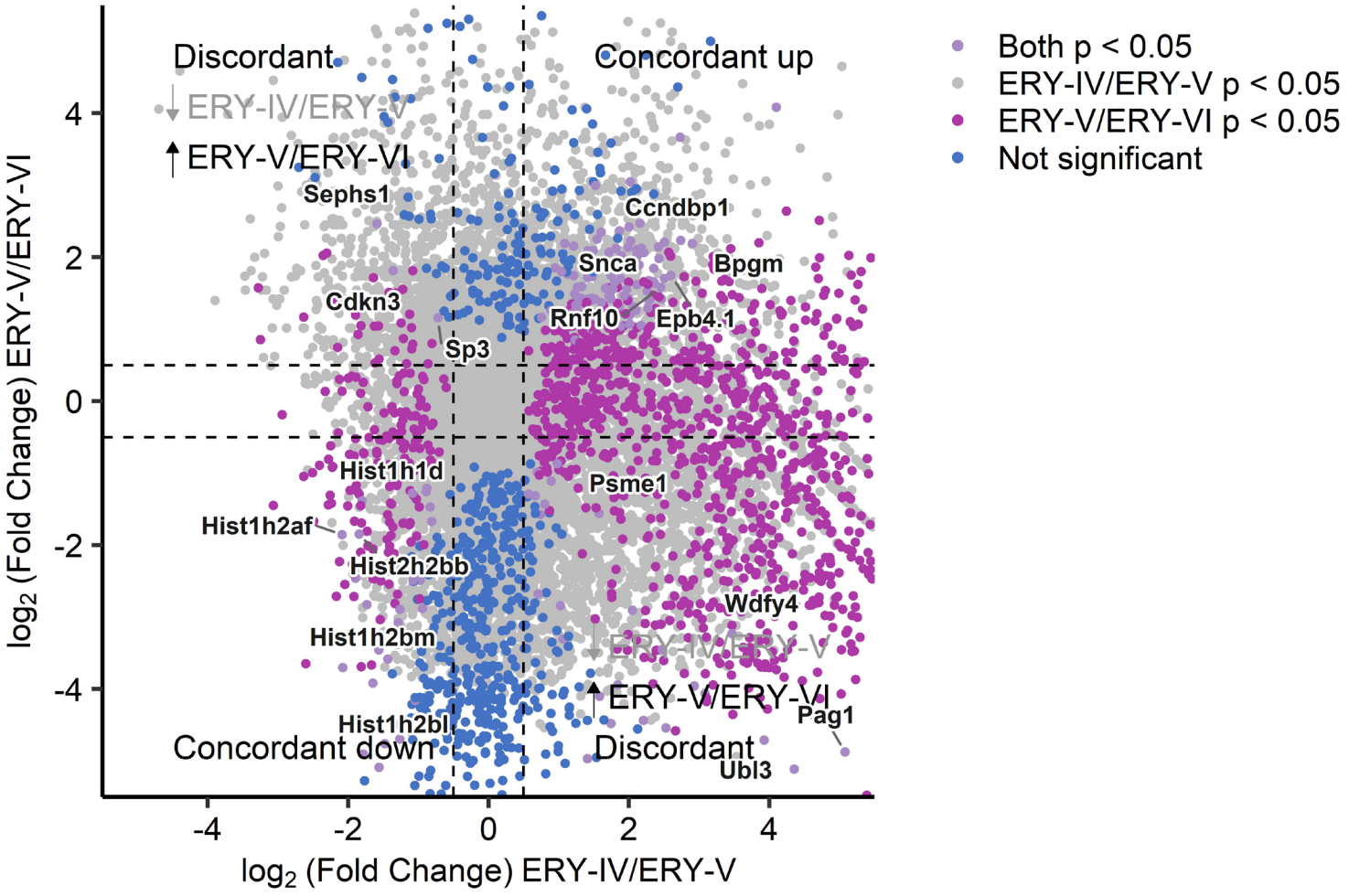
B.



红系细胞相邻发育阶段的差异表达分析（2）

北京大学 PEKING UNIVERSITY

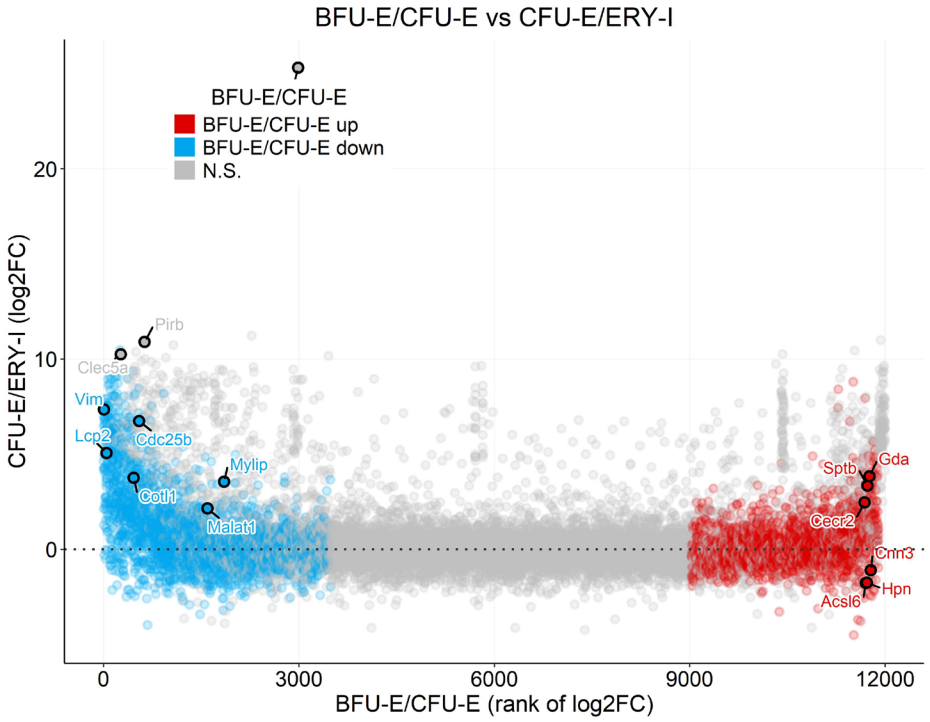
ERY-IV向ERY-V与ERY-V向ERY-VI转变的基因表达变化协同性散点图



红系细胞相邻发育阶段的差异表达分析（3）

BFU-E向CFU-E与CFU-E向ERY-I转变的基因表达变化排序散点图

- 排序散点图原理：X轴表示第一个转变过程中基因log2FC的排序值，Y轴表示第二个转变过程的log2FC值
- 颜色编码：红色（X轴比较中显著差异上调的基因）蓝色（X轴比较中显著差异下调的基因）
  灰色（X轴比较中无显著差异的基因）
- 统计阈值：padj < 0.05，log2FC = ±0.5
- 每张图标注两个转变过程中均显著变化的最显著的差异基因



排序散点图是FC-FC散点图的变种，通过对第一个比较中的基因按表达变化程度排序而非直接显示数值，更清晰地展示两个连续发育阶段中基因表达的整体趋势和相关性模式。这种可视化方法特别适合展示大量基因的表达变化模式，便于识别在连续发育阶段中表达变化相似或相反的基因群组。

- Vim：细胞骨架蛋白，维持细胞形态
- Lcp2：信号转导蛋白，参与造血细胞调控
- Cdc25b：细胞周期调控因子，促进细胞分裂
- Cotl1：肌动蛋白结合蛋白，影响细胞形态
- Mylip：参与蛋白质降解和脂质代谢
- Gda：核苷酸代谢酶，支持细胞增殖
- Sptb：红细胞膜骨架蛋白，维持细胞形态
- Cecr2：染色质重塑因子，调控基因表达
- Cnn3：钙调蛋白，参与细胞骨架组织
- Hpn：丝氨酸蛋白酶，参与细胞生长

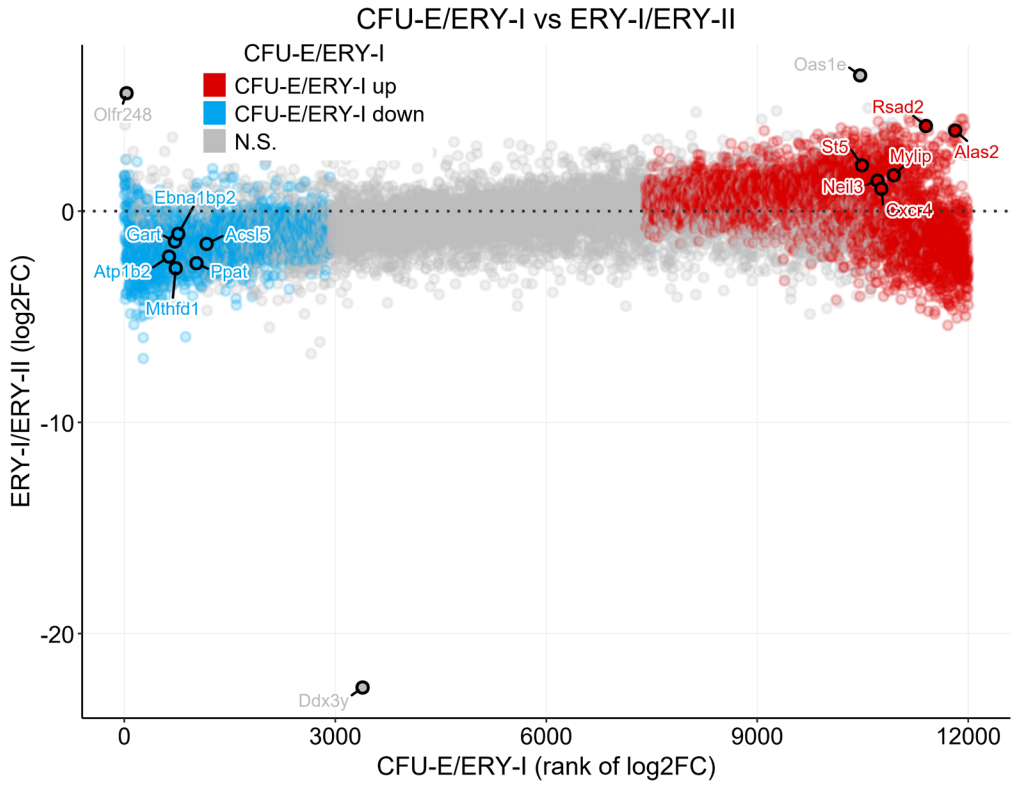
红系细胞相邻发育阶段的差异表达分析（3）

北京大学 PEKING UNIVERSITY

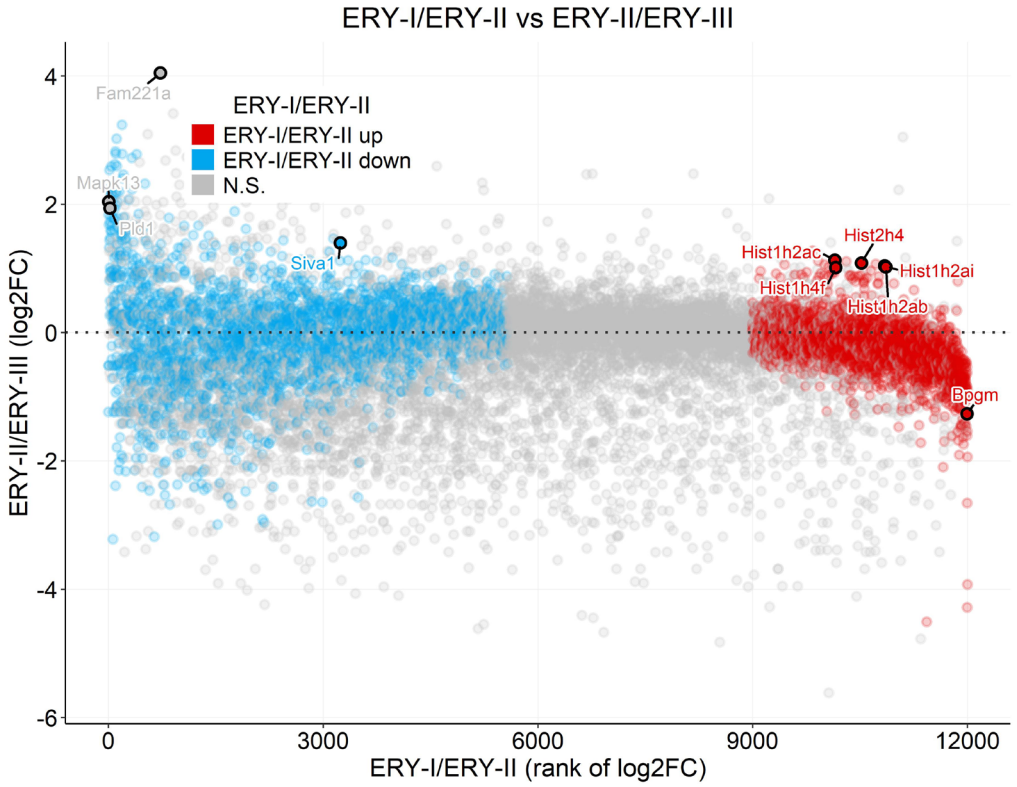
A. CFU-E向ERY-I与ERY-I向ERY-II转变的基因表达变化排序散点图

B. ERY-I向ERY-II与ERY-II向ERY-III转变的基因表达变化排序散点图

A.



B.



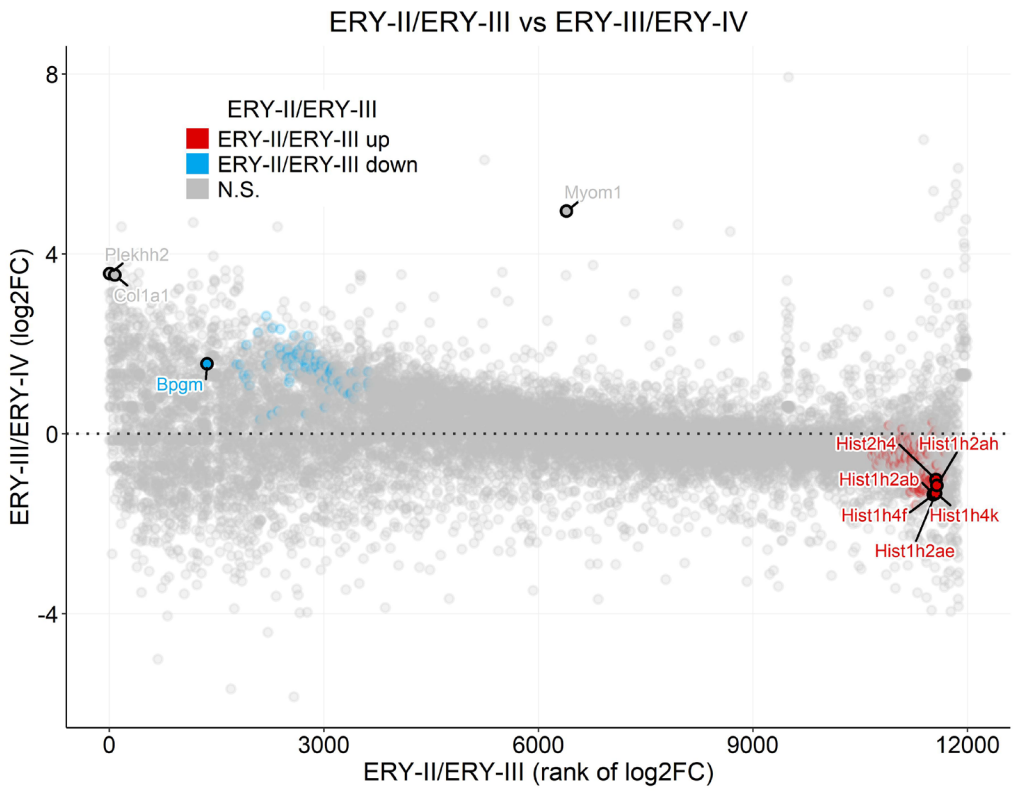
红系细胞相邻发育阶段的差异表达分析（3）

北京大学 PEKING UNIVERSITY

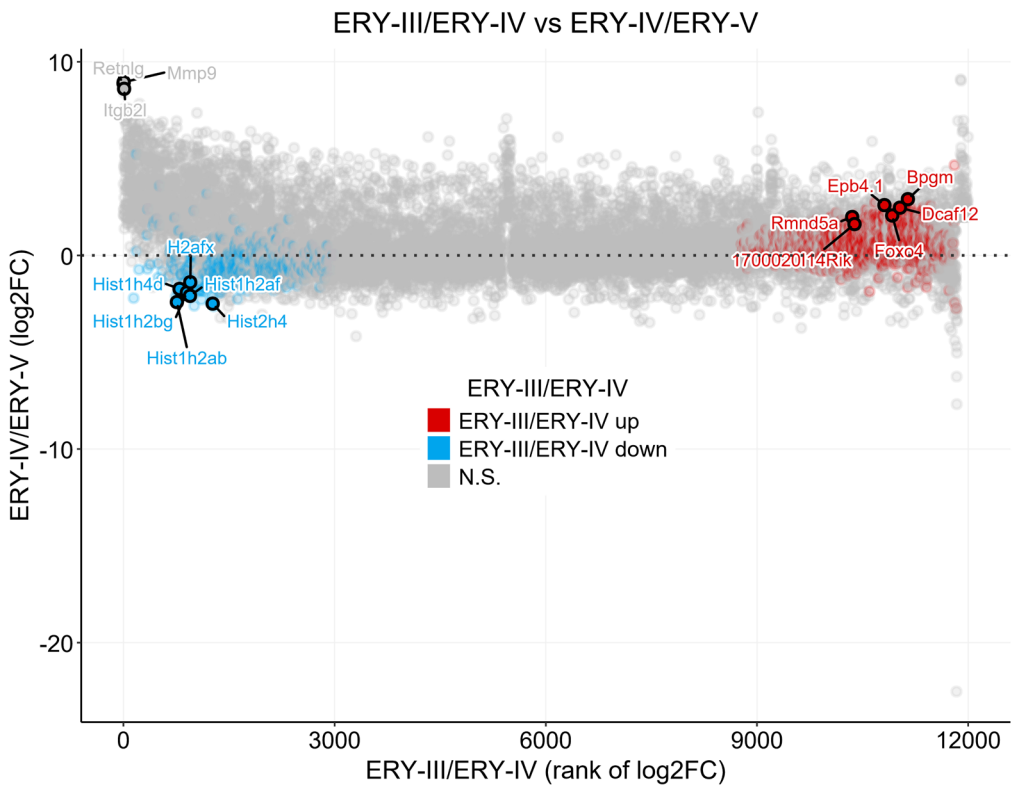
A. ERY-II向ERY-III与ERY-III向ERY-IV转变的基因表达变化排序散点图

B. ERY-III向ERY-IV与ERY-IV向ERY-V转变的基因表达变化排序散点图

A.



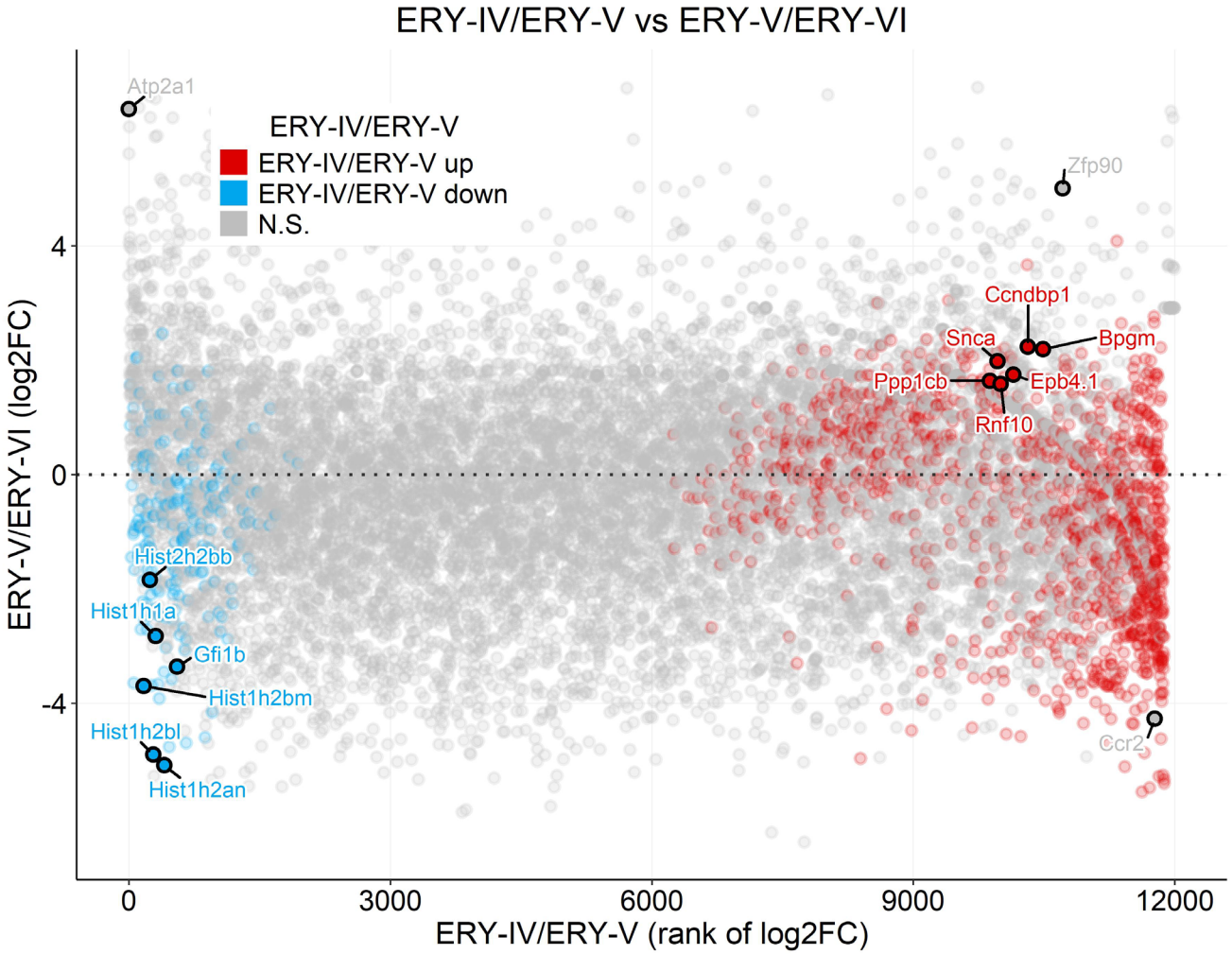
B.



红系细胞相邻发育阶段的差异表达分析（3）

北京大学 PEKING UNIVERSITY

ERY-IV向ERY-V与ERY-V向ERY-VI转变的基因表达变化排序散点图

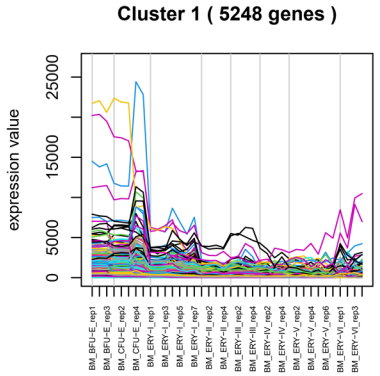


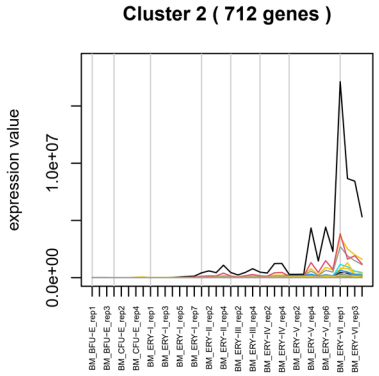
红系细胞分化发育进程的时序变化分析（1）

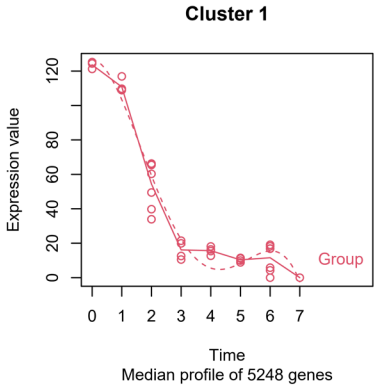
北京大学 PEKING UNIVERSITY

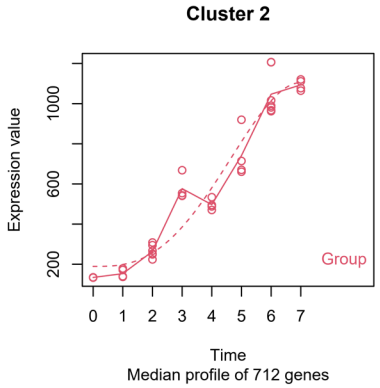
红系细胞发育过程中的基因表达动态聚类模式

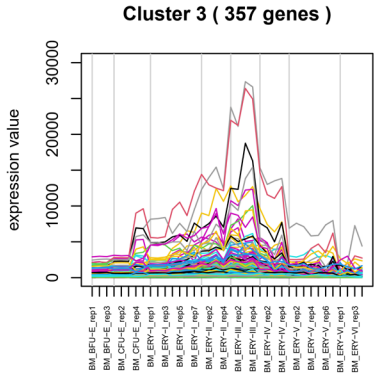
maSigPro全称为Microarray Significant Profiles，最初是应用于分析多时间序列的芯片数据，现在也可用于转录组数据。主要原理为采用两步回归策略，找出具有显著时间表达变化和实验组间显著差异的基因。具体流程为：首先通过最小二乘法确定有差异表达的基因。随后，采用逐步回归作为变量选择策略，研究实验组之间的差异，并找出在统计学上有意义的差异基因。第二步回归是用过选择系数将随时间变化发生显著表达变化的基因进行聚类并且可视化。

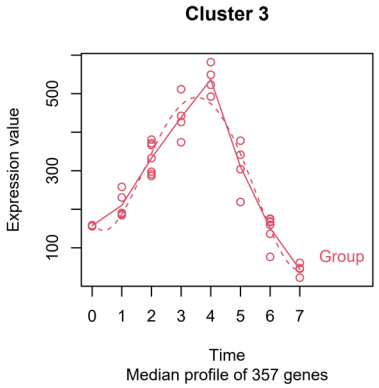










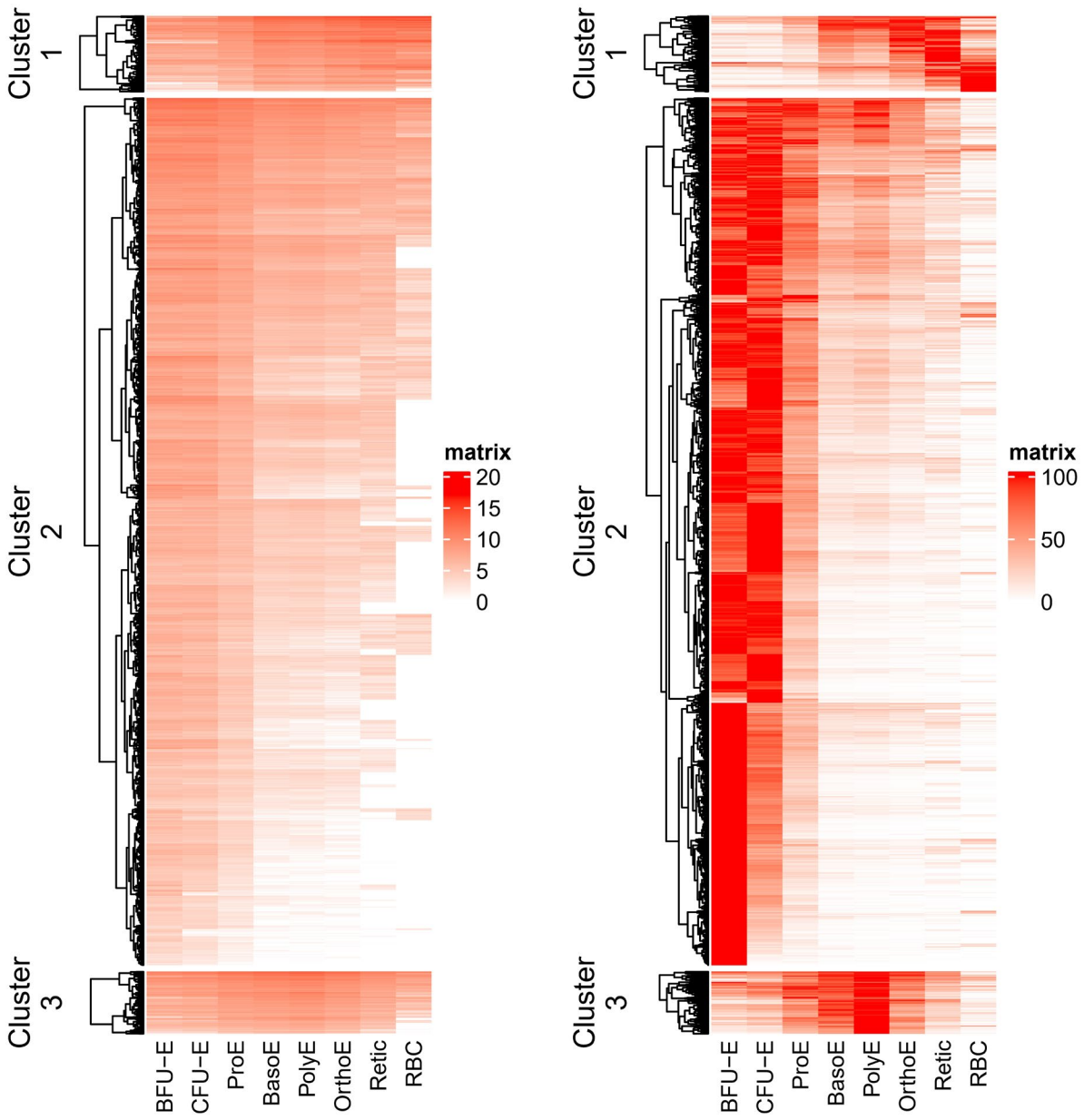


红系细胞分化发育进程的时序变化分析（1）

红系细胞发育过程中的时序基因表达热图

双面板热图，左侧展示log2表达值，右侧展示相对最大表达的百分比。热图按照3个主要聚类组织，展示了每个基因在8个发育阶段的表达模式。

- 左侧热图：使用红色渐变显示绝对表达水平(log2(标准化计数+1))
- 右侧热图：使用红色渐变显示相对最大表达的百分比(0-100%)
- 基因按聚类分组，展示相似表达模式的基因集
- 不显示基因名称，聚焦于整体表达模式的可视化

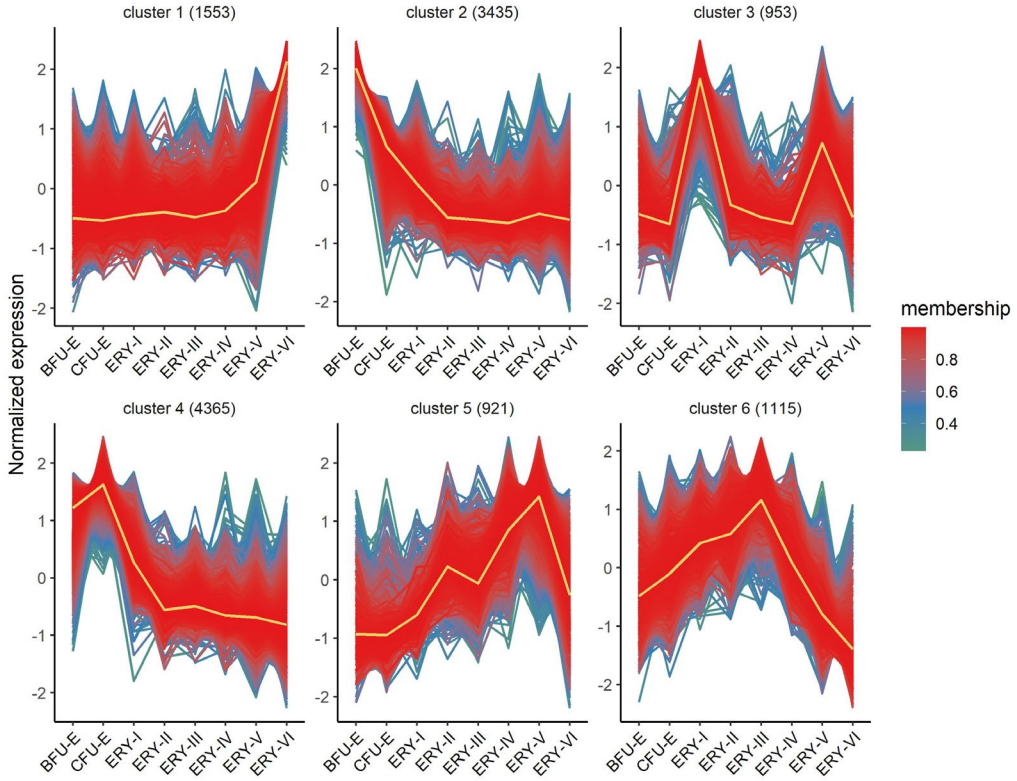


红系细胞分化发育进程的时序变化分析（2）

北京大学 PEKING UNIVERSITY

红系细胞发育过程中的基因表达动态聚类模式

mfuzz是一种基于模糊c均值算法的基因表达聚类方法，专为时间序列转录组数据设计。与传统硬聚类方法不同，mfuzz计算每个基因对各聚类的隶属度，使基因可以同时属于多个聚类，更准确地反映基因表达的复杂动态变化。通过引入模糊化参数(m)控制聚类边界的柔性，mfuzz对表达数据中的噪声和异常值表现出较高的稳健性。在分析流程中，首先对表达数据进行标准化，然后基于内部评估指标确定最优聚类数，最后生成核心基因的表达曲线图，展示各功能模块在发育过程中的表达特征。

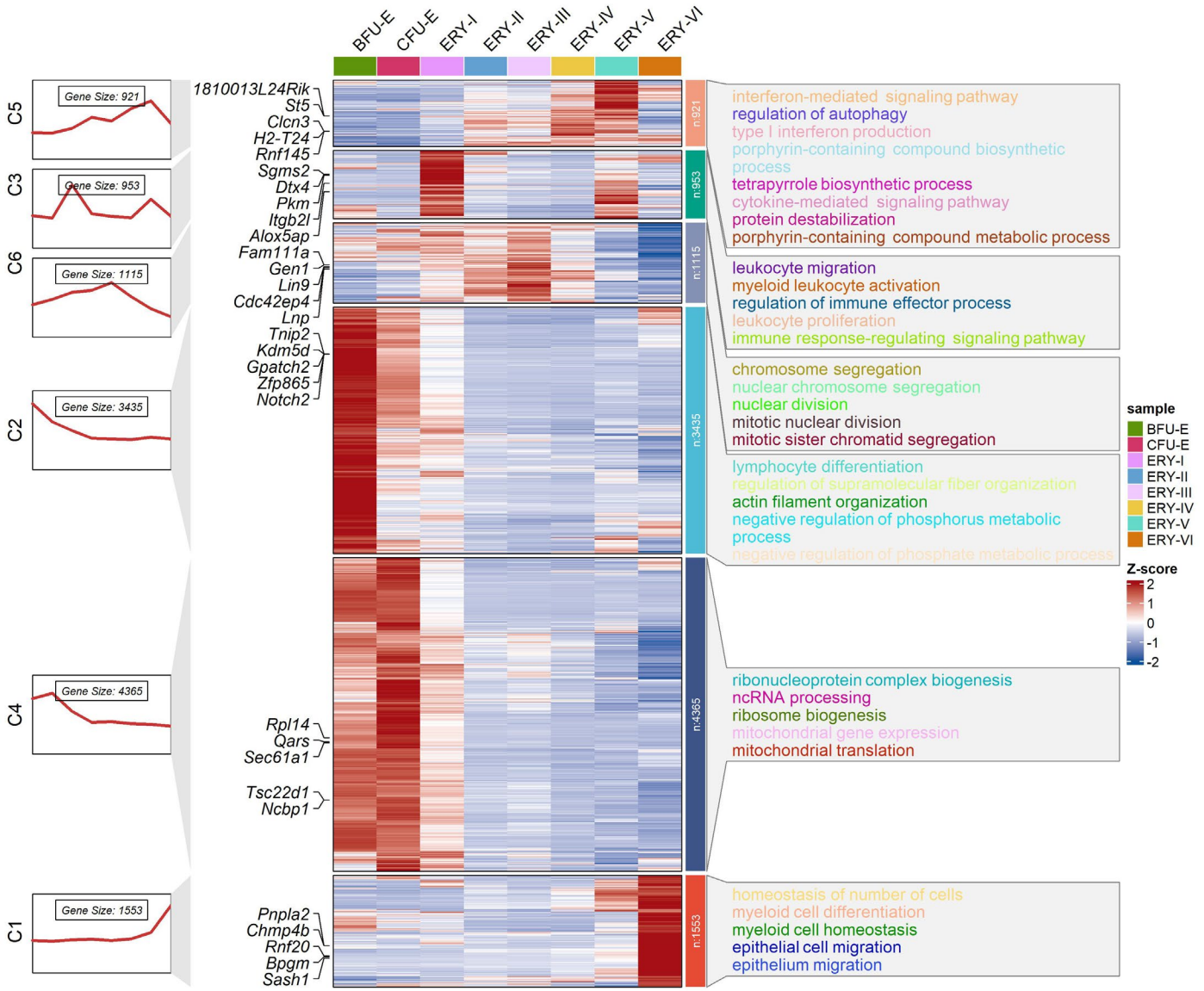


红系细胞分化发育进程的时序变化分析（2）

红系细胞发育过程中的时序基因表达热图

复合热图展示了不同基因模块的表达动态及其功能富集。通过mfuzz模糊聚类，将基因分为6个主要表达模块(C1-C6)，每个模块代表一种独特的时序表达模式。

- 左侧曲线图：显示各聚类的平均表达趋势，并标注每个聚类包含的基因数量
- 中央热图：基于Z-score标准化的表达值，红色表示高表达，蓝色表示低表达，白色表示中等表达水平
- 右侧注释：展示每个聚类富集的主要生物学过程，采用不同颜色标记不同功能类别



感谢倾听，欢迎提问！