**生信分析报告**

**项目标题： 再生障碍性贫血 ;**

**单 号： BSCL241113 ;**

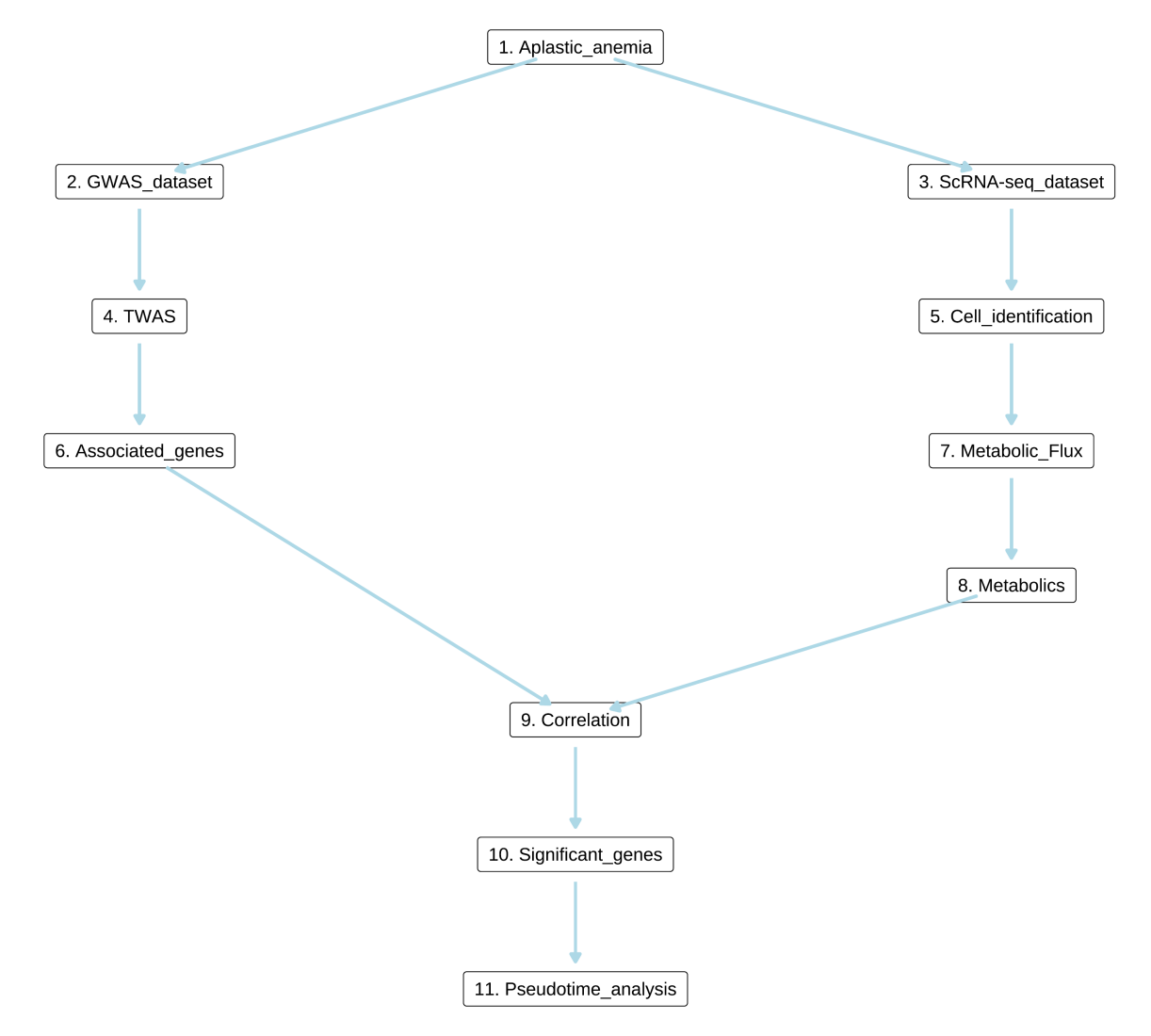
**分析人员： 黄礼闯 ;**

**分析类型： 生信分析 ;**

**委 托 人： 邓姝 ;**

**受 托 人： 杭州铂赛生物科技有限公司 .**

# 1 分析流程



**Fig.** **1** Route

**(File path: Figure+Table/1.0\_分析流程\_{#abstract}/Route.pdf)**

# 2 材料和方法

## 2.1 数据分析平台

在 Linux pop-os x86\_64 (6.9.3-76060903-generic) 上，使用 R version 4.4.2 (2024-10-31) (<https://www.r-project.org/>) 对数据统计分析与整合分析。

## 2.2 MungeSumstats 获取 GWAS 数据 (Dataset: AA)

以 R 包 MungeSumstats (1.15.12) (2021, **IF:4.4**, Q1, Bioinformatics)1 和 R 包 ieugwasr 获取 Open GWAS 的可用数据。 从 GWAS Catalog (<https://www.ebi.ac.uk/>) 下载 GCST90018794 的 Full Summary Statistic 数据 (<https://ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics//GCST90018001-GCST90019000//GCST90018794//GCST90018794_buildGRCh37.tsv.gz>)。

## 2.3 VEP 变异注释 (Dataset: AA)

以 Ensembl-vep 对 SNP 注释 (2016, **IF:10.1**, Q1, Genome Biology)2，获取 rsID。

## 2.4 FUSION TWAS全转录组关联研究 (Dataset: AA)

以 Python 工具 LDSC (munge\_sumstats.py) (<https://github.com/bulik/ldsc>) (2015, **IF:31.7**, Q1, Nature genetics)3 将 GWAS summary 文件检查并格式化为 .sumstats 格式。 获取 Whole Blood 组织的表达权重文件 (Expression Weights) (<https://s3.us-west-1.amazonaws.com/gtex.v8.fusion/ALL/GTExv8.ALL.Whole_Blood.tar.gz>)。 以 FUSION (2016, **IF:31.7**, Q1, Nature Genetics)4 (<http://gusevlab.org/projects/fusion/>) 进行 TWAS 预测，得到基因与疾病之间的关联统计。

## 2.5 GEO 数据获取 (Dataset: AA\_SCRNA)

以 R 包 GEOquery (2.74.0) 获取 GSE181989 数据集。

## 2.6 Seurat 集成单细胞数据分析 (Dataset: AA)

使用 Seurat R 包 (5.1.0) 进行单细胞数据质量控制 (QC) 和下游分析。依据 <https://satijalab.org/seurat/articles/integration_introduction> 为指导对单细胞数据预处理。 一个细胞至少应有 500 个基因，并且基因数量小于 5000。线粒体基因的比例小于 10%。根据上述条件，获得用于下游分析的高质量细胞。 执行标准 Seurat 分析工作流 (NormalizeData, FindVariableFeatures, ScaleData, RunPCA)。以 ElbowPlot 判断后续分析的 PC 维度。 在 1-15 PC 维度下，以 Seurat::FindNeighbors 构建 Nearest-neighbor Graph。随后在 1.2 分辨率下，以 Seurat::FindClusters 函数识别细胞群并以 Seurat::RunUMAP 进行 UMAP 聚类。 以 Seurat::FindAllMarkers (LogFC 阈值 0.25; 最小检出率 0.25) 为所有细胞群寻找 Markers。 以 Python 工具 SCSA ((2020, **IF:2.8**, Q2, Frontiers in genetics)5) (<https://github.com/bioinfo-ibms-pumc/SCSA>)，结合 CellMarker 数据库 ((2019, **IF:16.6**, Q1, Nucleic acids research)6)，对细胞群注释。

## 2.7 scFEA 单细胞数据的代谢通量预测 (Dataset: AA)

将 Seurat 的 RNA Assay (‘counts’) 作为输入数据，以 scFEA 预测细胞的代谢通量 (2021, **IF:6.2**, Q1, Genome research)7。参考 <https://github.com/changwn/scFEA/blob/master/scFEA_tutorial1.ipynb> 和 <https://github.com/changwn/scFEA/blob/master/scFEA_tutorial2.ipynb>。

## 2.8 Limma 代谢通量差异分析 (Dataset: AA\_FLUX)

以 limma (3.62.1) (2005)8 差异分析。分析方法参考 <https://bioconductor.org/packages/release/workflows/vignettes/RNAseq123/inst/doc/limmaWorkflow.html>。创建设计矩阵，对比矩阵，差异分析：Macrophage\_AA vs Macrophage\_Normal, Myeloid\_cell\_AA vs Myeloid\_cell\_Normal, Hematopoietic\_stem\_cell\_AA vs Hematopoietic\_stem\_cell\_Normal, Plasma\_cell\_AA vs Plasma\_cell\_Normal, B\_cell\_AA vs B\_cell\_Normal, Unknown\_AA vs Unknown\_Normal, Plasmacytoid\_dendritic\_cell\_AA vs Plasmacytoid\_dendritic\_cell\_Normal。使用 limma::lmFit, limma::contrasts.fit, limma::eBayes 拟合线形模型。以 limma::topTable 提取所有结果，并过滤得到 P.Value 小于 0.05，|Log2(FC)| 大于 0.5 的统计结果。

## 2.9 Monocle3 拟时分析 (Dataset: HSC)

从 Seurat 数据对象 Cell\_Sample 中提取 Hematopoietic\_stem\_cell 类型的细胞，对其重新聚类分析。在 1-15 PC 维度下，以 Seurat::FindNeighbors 构建 Nearest-neighbor Graph。随后在 1.2 分辨率下，以 Seurat::FindClusters 函数识别细胞群并以 Seurat::RunUMAP 进行 UMAP 聚类。使用 SeuratWrappers (SeuratWrappers::as.cell\_data\_set, 参考 <http://htmlpreview.github.io/?https://github.com/satijalab/seurat-wrappers/blob/master/docs/monocle3.html>) 将 Seurat 转化为 Monocle3 的 cell\_data\_set 数据。该转化将继承 Seurat 前期分析的 PCA、UMAP 等聚类结果，用于 Monocle3 的拟时分析，使前后分析一致。 以 monocle3::cluster\_cells 计算细胞群的 ‘clusters’ 和 ‘partitions’。以 monocle3::learn\_graph 从高维空间 (high-dimensional space) 中构建 ‘trajectory’。 选择 Y\_60 (principle points) 为拟时起点，以 monocle3::order\_cells 将细胞排序，随后构建细胞拟时变化图。 以 monocle3::graph\_test 寻找单细胞拟时轨迹中差异表达的基因。

# 3 分析结果

## 3.1 MungeSumstats 获取 GWAS 数据 (AA)

获取 Open GWAS 的可用数据，匹配 Aplastic anemia (trait)。从 GWAS Catalog 获取 Full Summary Statistic (ID: GCST90018794) 数据。

## 3.2 VEP 变异注释 (AA)

以 VEP (根据 chromosome, position, other allele (REF), effect allele (ALT)) 获取 rsID，

## 3.3 FUSION TWAS全转录组关联研究 (AA)

以 ldsc 将 GWAS summary 转化为 .sumstats 格式。以 FUSION 预测基因与疾病之间的关联 (chromosome: 1-22)。(TWAS 能够提供 SNP 如何通过调控基因表达来影响表型的机制)

**Tab.** **1** AA TWAS statistic

| PANEL | FILE | ID | SYMBOL | CHR |
| --- | --- | --- | --- | --- |
| GTExv8.ALL.Whole ... | /home/echo/disk s... | ENSG00000275106.1 |  | 7 |
| GTExv8.ALL.Whole ... | /home/echo/disk s... | ENSG00000135372.8 | NAT10 | 11 |
| GTExv8.ALL.Whole ... | /home/echo/disk s... | ENSG00000128604.19 |  | 7 |
| GTExv8.ALL.Whole ... | /home/echo/disk s... | ENSG00000109320.11 |  | 4 |
| GTExv8.ALL.Whole ... | /home/echo/disk s... | ENSG00000175164.13 |  | 9 |
| ... | ... | ... | ... | ... |

**(File path: Figure+Table/3.3\_FUSION\_TWAS全转录组关联研究\_(AA)/AA-TWAS-statistic.csv)**

Tab. **[1](#AA-TWAS-statistic)** 为 TWAS 基因与疾病关联性统计结果，显著性 TWAS.P.adjust 由 TWAS.P 以染色体对应的基因数 (Expression Weights) FDR 校正计算。该表格的解释请参考 <http://gusevlab.org/projects/fusion/>。

**Tab.** **2** AA TWAS significant

| PANEL | FILE | ID | SYMBOL | CHR |
| --- | --- | --- | --- | --- |
| GTExv8.ALL.Whole ... | /home/echo/disk s... | ENSG00000275106.1 |  | 7 |
| GTExv8.ALL.Whole ... | /home/echo/disk s... | ENSG00000135372.8 | NAT10 | 11 |
| GTExv8.ALL.Whole ... | /home/echo/disk s... | ENSG00000128604.19 |  | 7 |
| GTExv8.ALL.Whole ... | /home/echo/disk s... | ENSG00000109320.11 |  | 4 |
| GTExv8.ALL.Whole ... | /home/echo/disk s... | ENSG00000175164.13 |  | 9 |
| ... | ... | ... | ... | ... |

**(File path: Figure+Table/3.3\_FUSION\_TWAS全转录组关联研究\_(AA)/AA-TWAS-significant.csv)**

Tab. **[2](#AA-TWAS-significant)** 为 TWAS 显著统计表 (TWAS.P < 0.05)。

## 3.4 GEO 数据获取 (AA\_SCRNA)

以 GEOquery 获取 GSE181989 的数据信息。

* Data Source ID: GSE181989
* data\_processing: The Cell Ranger software pipeline (version 2.2.0) provided by 10×Genomics was used to demultiplex cellular barcodes, map reads to the genome and transcriptome using the STAR aligner, and down-sample reads as required to generate normalized aggregate data across samples, producing a matrix of gene counts versus cells
* data\_processing.1: Genome\_build: mm10
* data\_processing.2: Supplementary\_files\_format\_and\_content: CellRanger output files (barcodes.tsv, features.tsv, matrix.mtx)

**(见Figure+Table/3.4\_GEO\_数据获取\_(AA\_SCRNA)/AA-SCRNA-GSE181989-content)**

## 3.5 Seurat 集成单细胞数据分析 (AA)

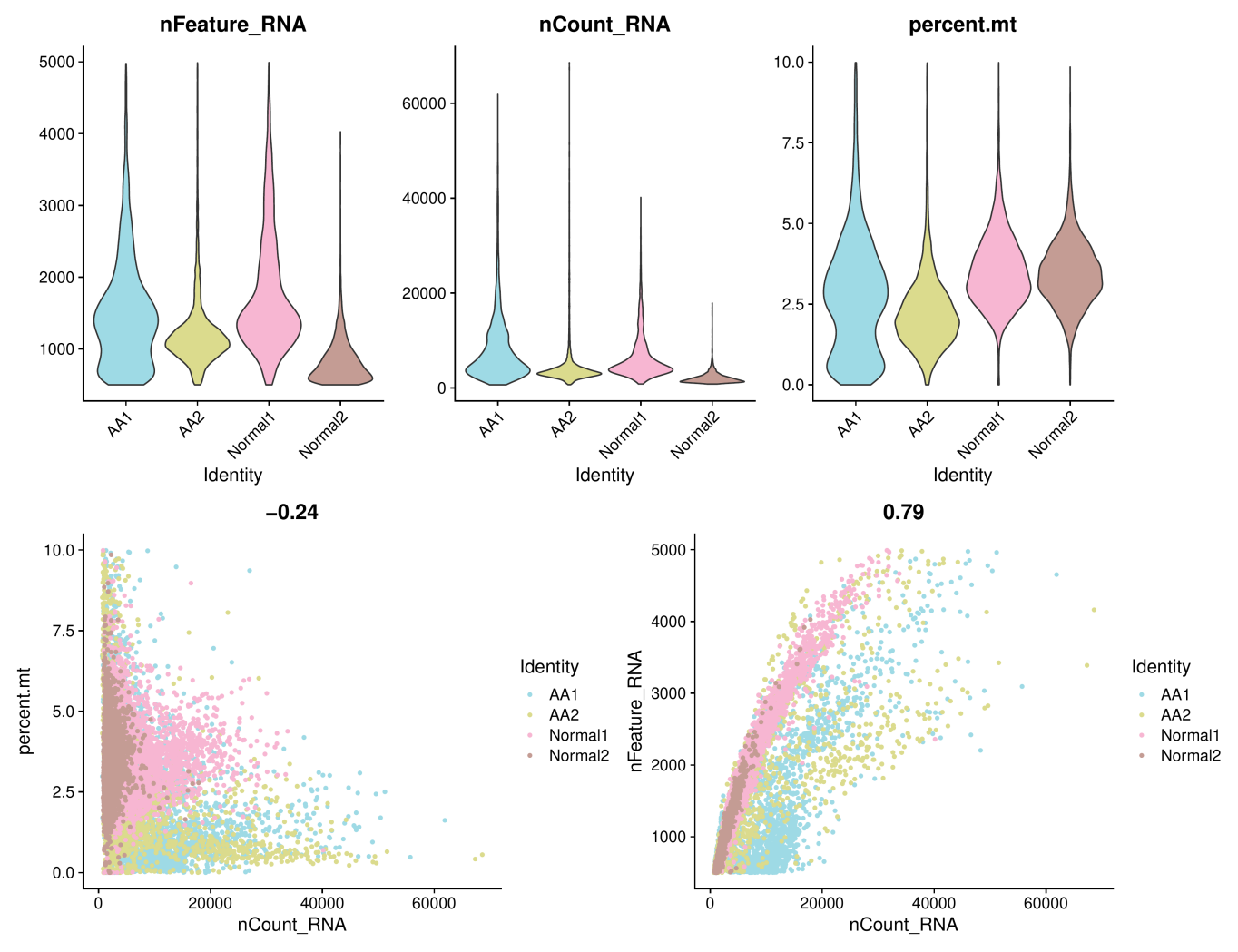
读取 AA1, AA2, Normal1, Normal2 样本的数据集。前期质量控制，一个细胞至少应有 500 个基因，并且基因数量小于 5000。线粒体基因的比例小于 10%。数据归一化，PCA 聚类 (Seurat 标准工作流，见方法章节) 后，绘制 PC standard deviations 图。在 1-15 PC 维度，1.2 分辨率下，对细胞群 UMAP 聚类。计算所有细胞群的 Marker。使用 CellMarker 数据库的 Marker 基因 (Tissue: Bone marrow)，以 SCSA 对细胞群注释。 对细胞群差异分析 (依据 Cell\_Sample, 分析 “Macrophage\_AA vs Macrophage\_Normal”, “Myeloid\_cell\_AA vs Myeloid\_cell\_Normal”, “Hematopoietic\_stem\_cell\_AA vs Hematopoietic\_stem\_cell\_Normal”, …(n = 7))，筛选差异表达基因。



**Fig.** **2** Pre Quality control

**(File path: Figure+Table/3.5\_Seurat\_集成单细胞数据分析\_(AA)/Pre-Quality-control.pdf)**

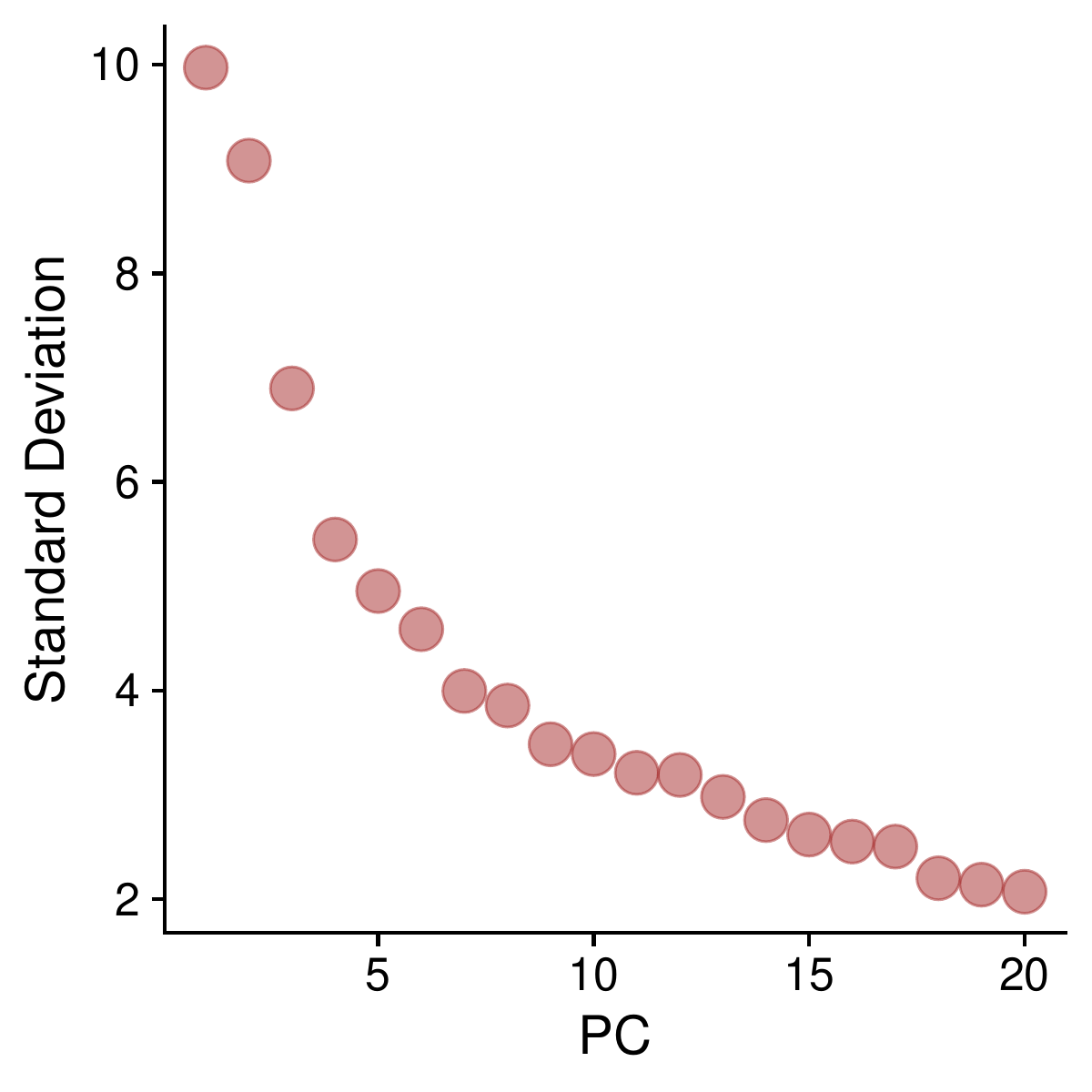
Fig. **[2](#Pre-Quality-control)** 为 QC (质量控制) 图 (数据过滤前) 。



**Fig.** **3** AA After Quality control

**(File path: Figure+Table/3.5\_Seurat\_集成单细胞数据分析\_(AA)/AA-After-Quality-control.pdf)**

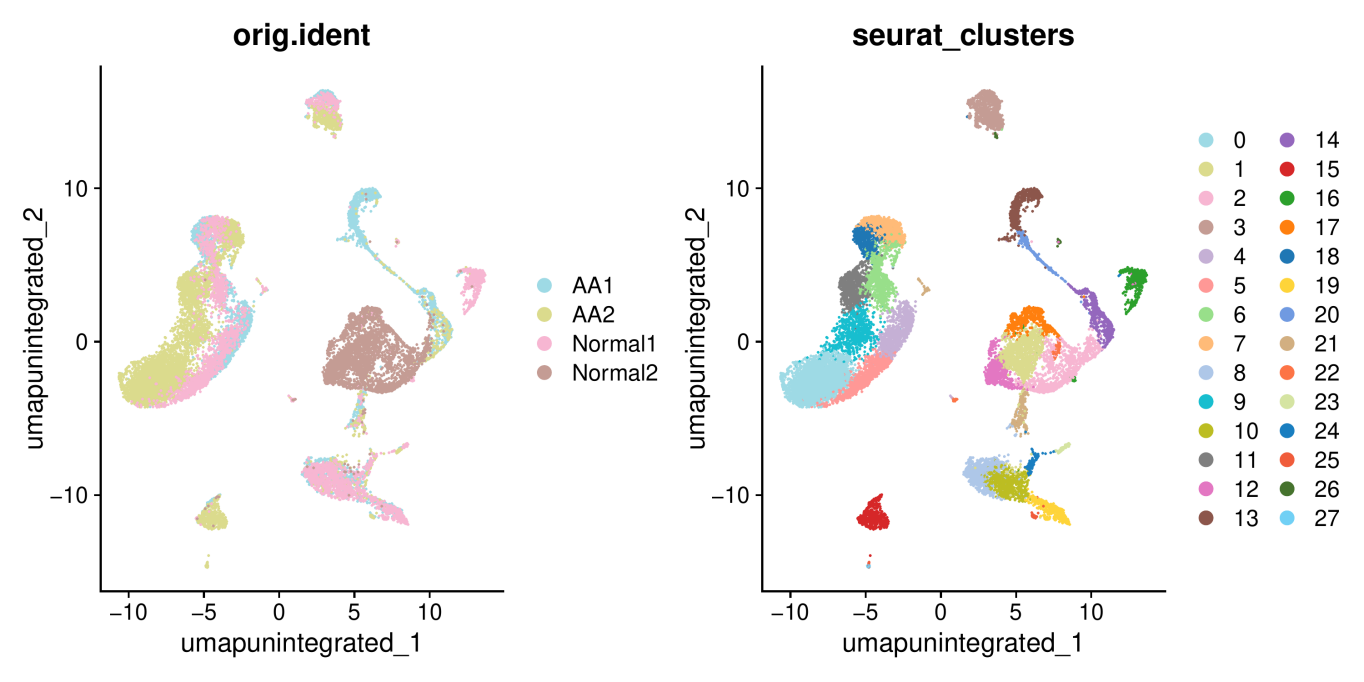
Fig. **[3](#AA-After-Quality-control)** 为数据过滤后的 QC 图。



**Fig.** **4** AA Standard deviations of PCs

**(File path: Figure+Table/3.5\_Seurat\_集成单细胞数据分析\_(AA)/AA-Standard-deviations-of-PCs.pdf)**

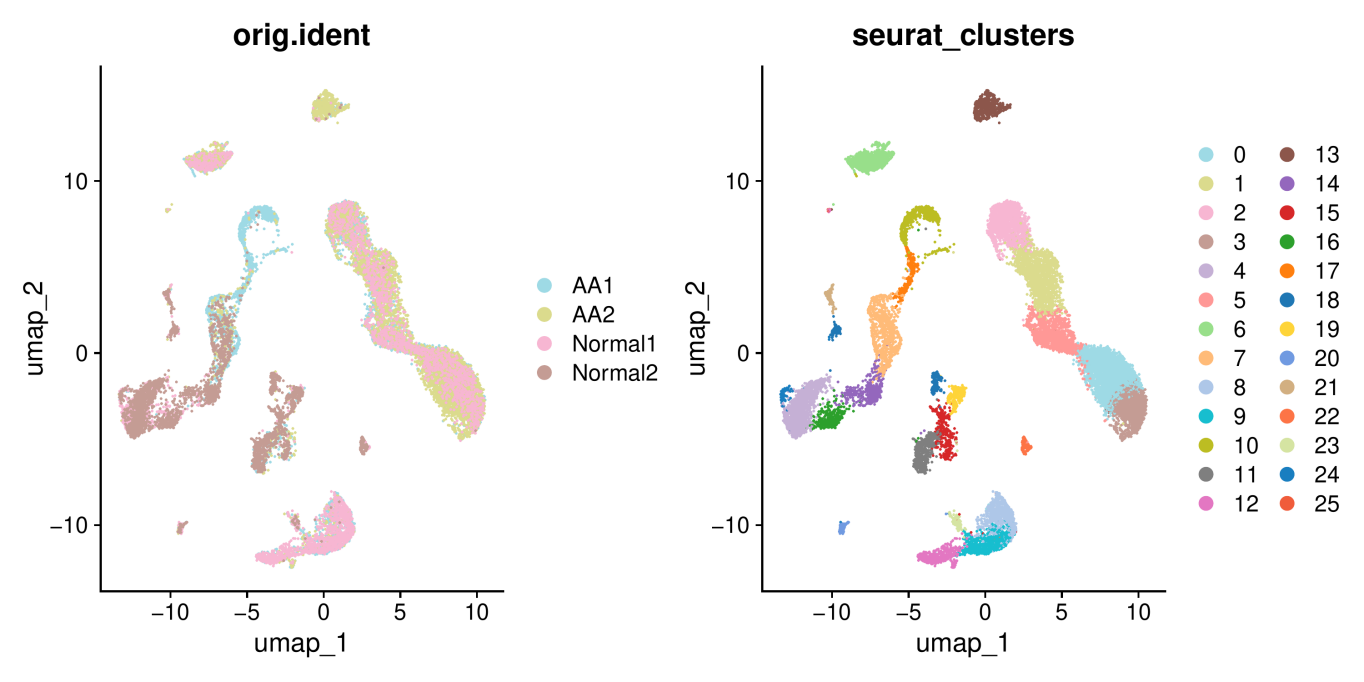
Fig. **[4](#AA-Standard-deviations-of-PCs)** 为主成分 (PC) 的 Standard deviations。



**Fig.** **5** AA UMAP Unintegrated

**(File path: Figure+Table/3.5\_Seurat\_集成单细胞数据分析\_(AA)/AA-UMAP-Unintegrated.pdf)**

Fig. **[5](#AA-UMAP-Unintegrated)** 为去除批次效应之前的 UMAP 聚类图。



**Fig.** **6** AA UMAP Integrated

**(File path: Figure+Table/3.5\_Seurat\_集成单细胞数据分析\_(AA)/AA-UMAP-Integrated.pdf)**

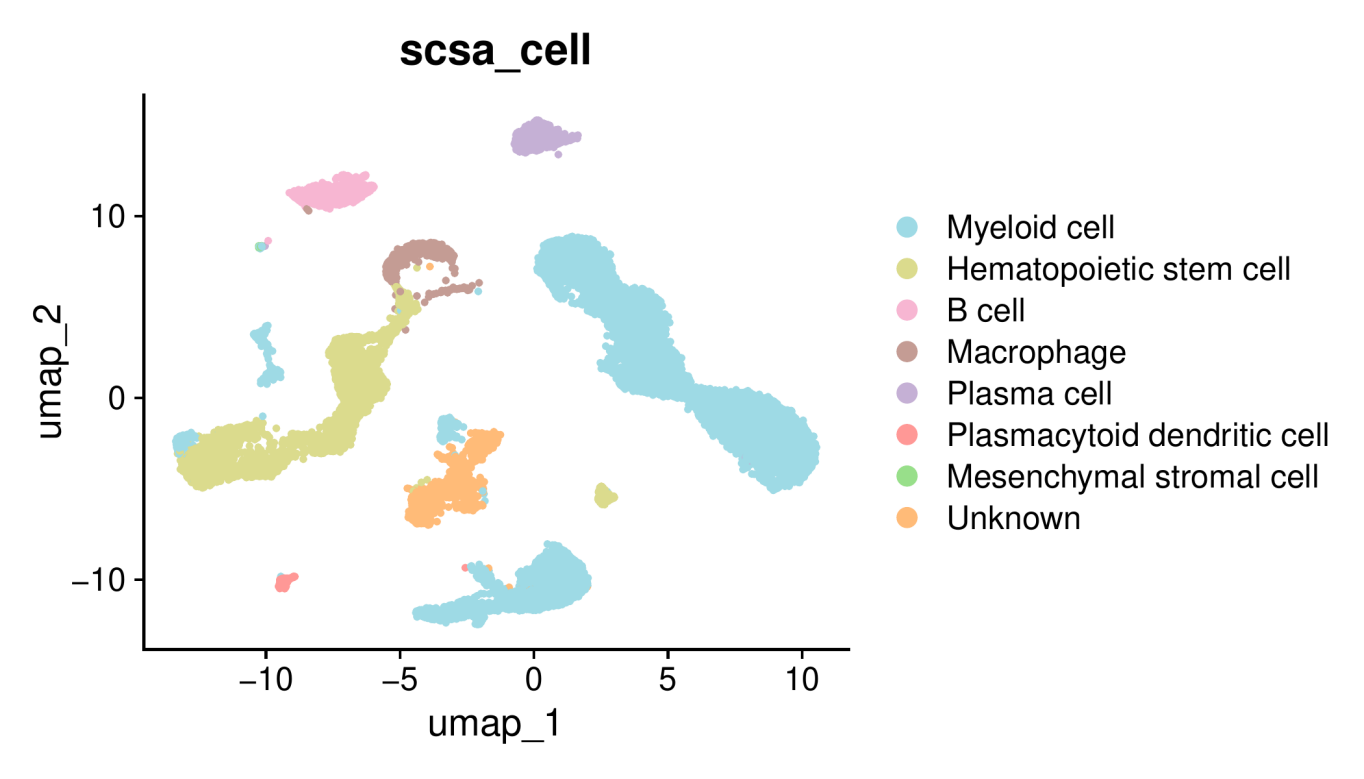
Fig. **[6](#AA-UMAP-Integrated)** 为 去除批次效应之后的 UMAP 聚类图。

**Tab.** **3** AA significant markers of cell clusters

| Rownames | P val | Avg log2FC | Pct.1 | Pct.2 |
| --- | --- | --- | --- | --- |
| IL7R | 0 | 1.591 | 0.878 | 0.242 |
| CD3E | 0 | 1.291 | 0.977 | 0.369 |
| LTB | 0 | 1.267 | 0.865 | 0.321 |
| CAMK4 | 0 | 2.249 | 0.66 | 0.146 |
| TCF7 | 0 | 2.477 | 0.64 | 0.13 |
| ... | ... | ... | ... | ... |

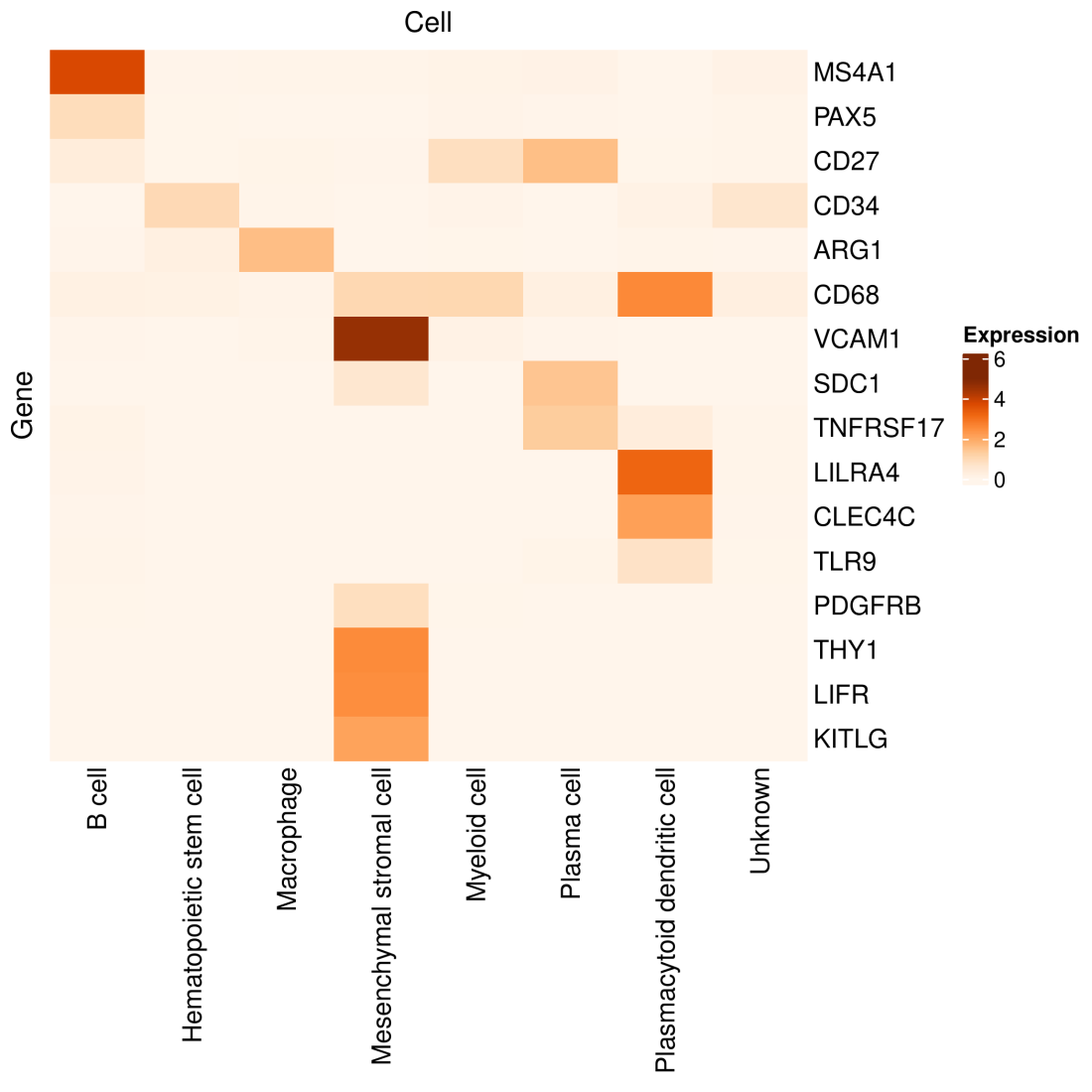
**(File path: Figure+Table/3.5\_Seurat\_集成单细胞数据分析\_(AA)/AA-significant-markers-of-cell-clusters.csv)**

Tab. **[3](#AA-significant-markers-of-cell-clusters)** 为所有细胞群的 Marker (LogFC 阈值 0.25; 最小检出率 0.25; 矫正 P 值阈值 0.05)



**Fig.** **7** AA SCSA Cell type annotation

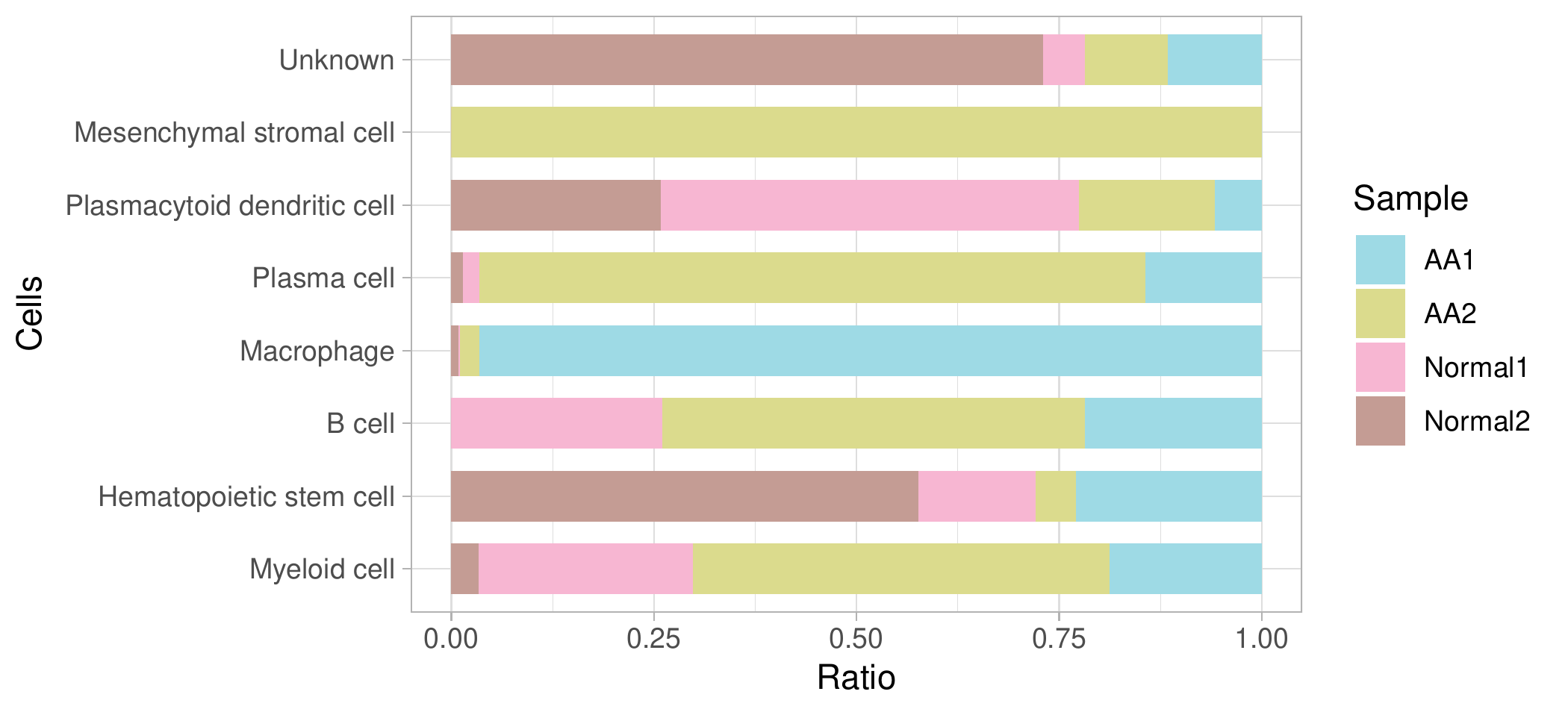
**(File path: Figure+Table/3.5\_Seurat\_集成单细胞数据分析\_(AA)/AA-SCSA-Cell-type-annotation.pdf)**



**Fig.** **8** AA CellMarker Validation

**(File path: Figure+Table/3.5\_Seurat\_集成单细胞数据分析\_(AA)/AA-CellMarker-Validation.pdf)**

Fig. **[8](#AA-CellMarker-Validation)** 为使用 CellMarker 数据库中的特异性 Marker 对单细胞注释结果的验证热图。



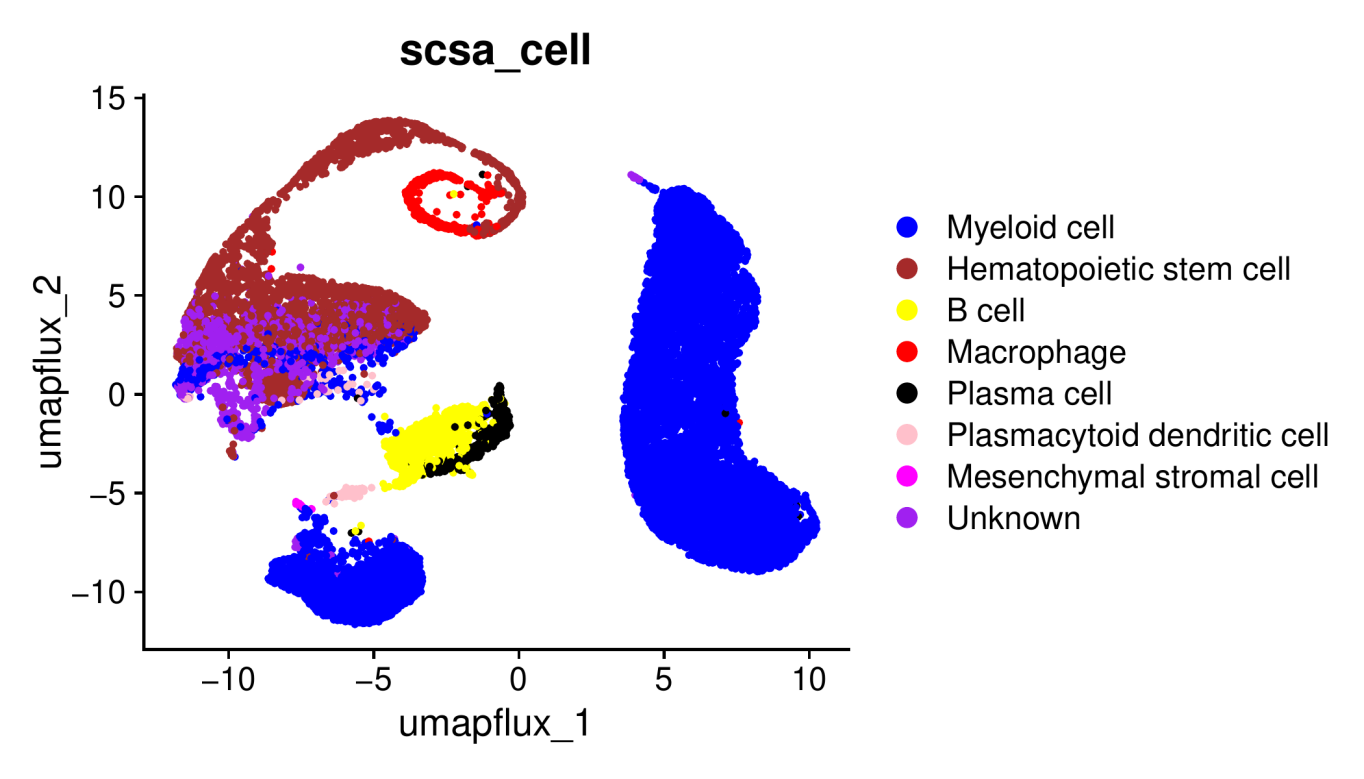
**Fig.** **9** AA SCSA Cell Proportions in each sample

**(File path: Figure+Table/3.5\_Seurat\_集成单细胞数据分析\_(AA)/AA-SCSA-Cell-Proportions-in-each-sample.pdf)**

Fig. **[9](#AA-SCSA-Cell-Proportions-in-each-sample)** 为 SCSA 注释的细胞群在各个样本中的占比。

## 3.6 scFEA 单细胞数据的代谢通量预测 (AA)

将 Seurat (所有细胞) 以 scFEA 预测代谢通量。



**Fig.** **10** AA cells metabolic flux

**(File path: Figure+Table/3.6\_scFEA\_单细胞数据的代谢通量预测\_(AA)/AA-cells-metabolic-flux.pdf)**

Fig. **[10](#AA-cells-metabolic-flux)** 为细胞代谢通量 (scFEA 预测，输入 Seurat) 的 UMAP 聚类。

**Tab.** **4** AA metabolic flux matrix

| V1 | M 1 | M 2 | M 3 | M 4 |
| --- | --- | --- | --- | --- |
| AAACCTGAGGCAGGTT-1 1 | 0.133 | 0.06644 | 0.3119 | 0.1901 |
| AAACCTGCAGCGTCCA-1 1 | -6.328e-05 | 0.009283 | 0.0511 | 0.1628 |
| AAACCTGCAGGCAGTA-1 1 | -6.538e-07 | -1.248e-07 | 0.005326 | 0.04691 |
| AAACCTGCATTAGCCA-1 1 | -6.328e-05 | -1.248e-07 | 0.005277 | 0.04302 |
| AAACCTGGTCTAAACC-1 1 | -6.305e-07 | 0.001004 | 0.01098 | 0.02247 |
| ... | ... | ... | ... | ... |

**(File path: Figure+Table/3.6\_scFEA\_单细胞数据的代谢通量预测\_(AA)/AA-metabolic-flux-matrix.csv)**

Tab. **[4](#AA-metabolic-flux-matrix)** 为细胞代谢通量矩阵 (各 M\_ 为代谢模块)。

**Tab.** **5** AA annotation of metabolic flux

| V1 | Module id | Compound IN name | Compound IN ID | Compound OUT name |
| --- | --- | --- | --- | --- |
| M 1 | 1 | Glucose | C00267 | G6P |
| M 2 | 2 | G6P | C00668 | G3P |
| M 3 | 3 | G3P | C00118 | 3PD |
| M 4 | 4 | 3PD | C00197 | Pyruvate |
| M 5 | 5 | Pyruvate | C00022 | Acetyl-Coa |
| ... | ... | ... | ... | ... |

**(File path: Figure+Table/3.6\_scFEA\_单细胞数据的代谢通量预测\_(AA)/AA-annotation-of-metabolic-flux.xlsx)**

Tab. **[5](#AA-annotation-of-metabolic-flux)** 各代谢模块的注释。

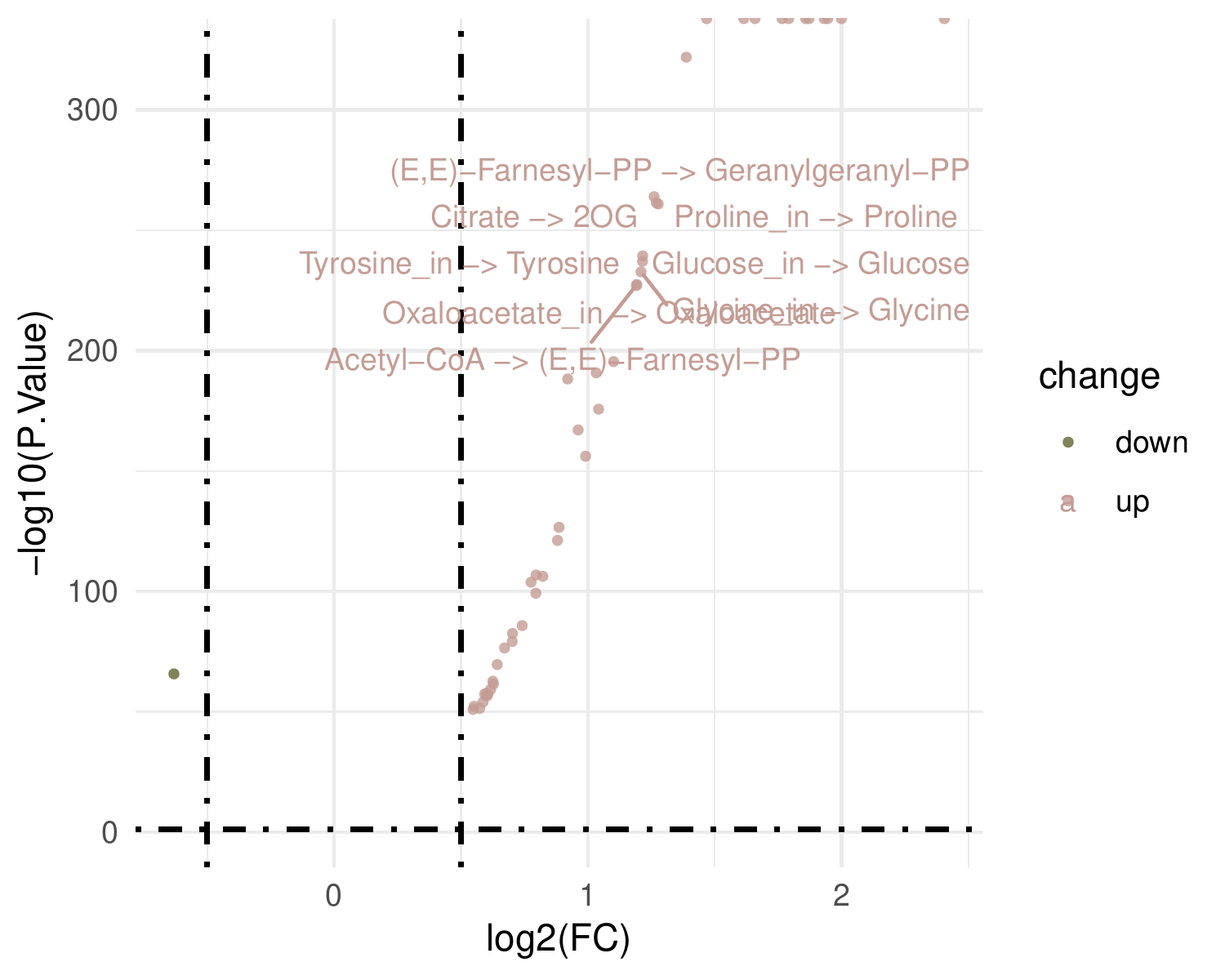
## 3.7 Limma 代谢通量差异分析 (AA\_FLUX)

以 limma 的线形分析策略，对细胞的代谢通量差异分析。以 公式 ~ 0 + group 创建设计矩阵 (design matrix) 。差异分析：Macrophage\_AA vs Macrophage\_Normal, Myeloid\_cell\_AA vs Myeloid\_cell\_Normal, Hematopoietic\_stem\_cell\_AA vs Hematopoietic\_stem\_cell\_Normal, Plasma\_cell\_AA vs Plasma\_cell\_Normal, B\_cell\_AA vs B\_cell\_Normal, Unknown\_AA vs Unknown\_Normal, Plasmacytoid\_dendritic\_cell\_AA vs Plasmacytoid\_dendritic\_cell\_Normal。(若 A vs B，则为前者比后者，LogFC 大于 0 时，A 表达量高于 B)。各组差异分析 DMFs 统计：

* Macrophage\_AA - Macrophage\_Normal：up (n=5) , down (n=0) 。
* Myeloid\_cell\_AA - Myeloid\_cell\_Normal：up (n=0) , down (n=2) 。
* Hematopoietic\_stem\_cell\_AA - Hematopoietic\_stem\_cell\_Normal：up (n=48) , down (n=1) 。
* Plasma\_cell\_AA - Plasma\_cell\_Normal：up (n=2) , down (n=29) 。
* B\_cell\_AA - B\_cell\_Normal：up (n=0) , down (n=0) 。
* Unknown\_AA - Unknown\_Normal：up (n=46) , down (n=0) 。
* Plasmacytoid\_dendritic\_cell\_AA - Plasmacytoid\_dendritic\_cell\_Normal：up (n=27) , down (n=0) 。 所有上调 DMFs 共 87 个，所有下调 DMFs 共 31 个。所有非重复 DMFs 共 97 个。

Note: The directory 'Figure+Table/AA-FLUX-Differential-Statistic-data' contains 7 files.  
  
1 1\_Macrophage\_AA - Macrophage\_Normal.csv  
2 2\_Myeloid\_cell\_AA - Myeloid\_cell\_Normal.csv  
3 3\_Hematopoietic\_stem\_cell\_AA - Hematopoietic\_stem\_cell\_Normal.csv  
4 4\_Plasma\_cell\_AA - Plasma\_cell\_Normal.csv  
5 5\_B\_cell\_AA - B\_cell\_Normal.csv  
6 ...

**(File path: Figure+Table/3.7\_Limma\_代谢通量差异分析\_(AA\_FLUX)/AA-FLUX-Differential-Statistic-data)**



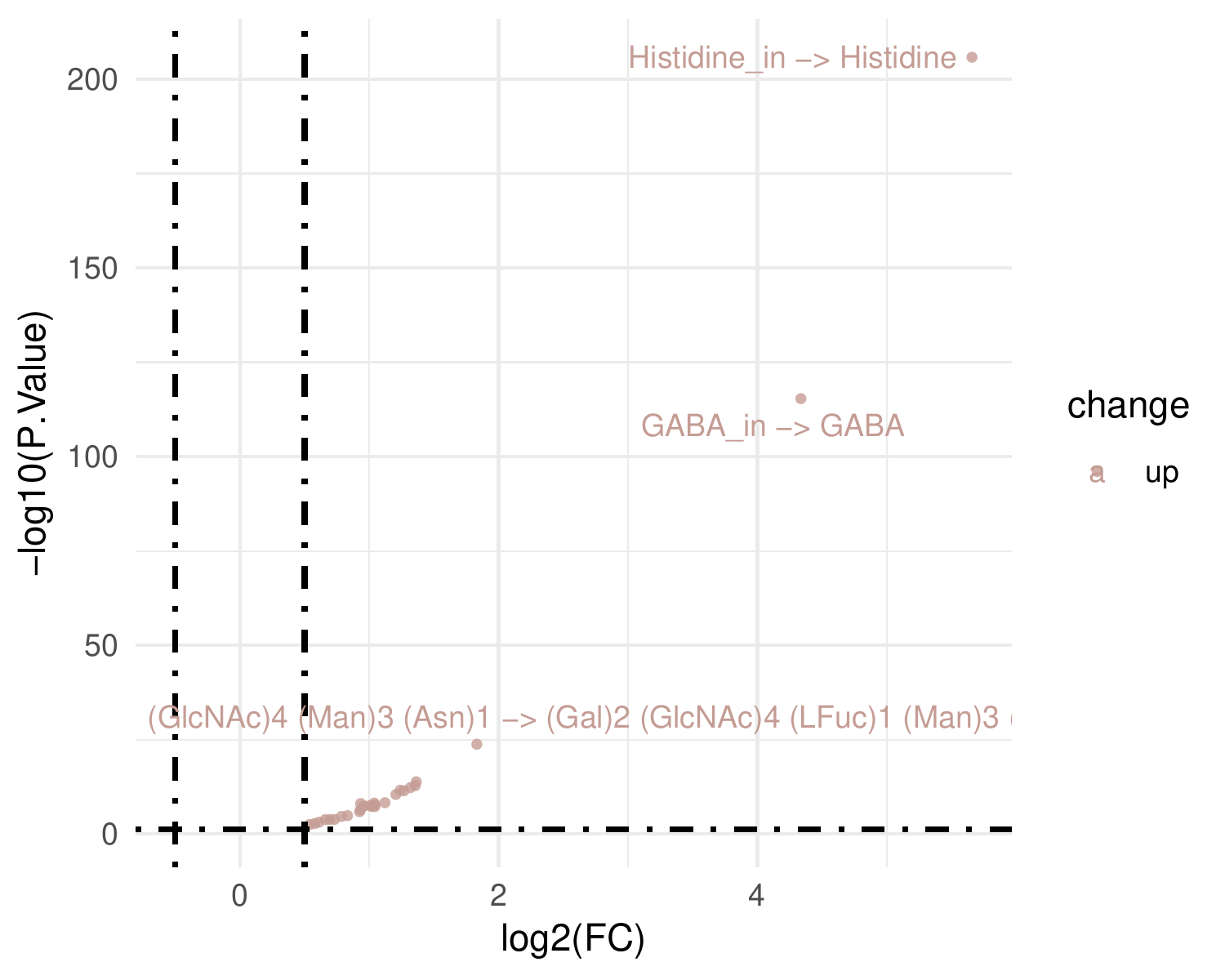
**Fig.** **11** AA FLUX Hematopoietic stem cell AA vs Hematopoietic stem cell Normal

**(File path: Figure+Table/3.7\_Limma\_代谢通量差异分析\_(AA\_FLUX)/AA-FLUX-Hematopoietic-stem-cell-AA-vs-Hematopoietic-stem-cell-Normal.pdf)**

* P.Value cut-off: 0.05
* Log2(FC) cut-off: 0.5

**(See: Figure+Table/3.7\_Limma\_代谢通量差异分析\_(AA\_FLUX)/AA-FLUX-Hematopoietic-stem-cell-AA-vs-Hematopoietic-stem-cell-Normal-content)**

Fig. **[11](#AA-FLUX-Hematopoietic-stem-cell-AA-vs-Hematopoietic-stem-cell-Normal)** 为 Hematopoietic\_stem\_cell\_AA - Hematopoietic\_stem\_cell\_Normal 差异分析火山图。



**Fig.** **12** AA FLUX Plasmacytoid dendritic cell AA vs Plasmacytoid dendritic cell Normal

**(File path: Figure+Table/3.7\_Limma\_代谢通量差异分析\_(AA\_FLUX)/AA-FLUX-Plasmacytoid-dendritic-cell-AA-vs-Plasmacytoid-dendritic-cell-Normal.pdf)**

* P.Value cut-off: 0.05
* Log2(FC) cut-off: 0.5

**(See: Figure+Table/3.7\_Limma\_代谢通量差异分析\_(AA\_FLUX)/AA-FLUX-Plasmacytoid-dendritic-cell-AA-vs-Plasmacytoid-dendritic-cell-Normal-content)**

Fig. **[12](#AA-FLUX-Plasmacytoid-dendritic-cell-AA-vs-Plasmacytoid-dendritic-cell-Normal)** 为 Plasmacytoid\_dendritic\_cell\_AA - Plasmacytoid\_dendritic\_cell\_Normal 差异分析火山图。

## 3.8 细胞群 features 关联分析 (AA)

以标准化熵筛选于样本中相对均衡分布的细胞 (Myeloid cell, Hematopoietic stem cell, Plasmacytoid dendritic cell, Unknown) (cutoff: 0.5)

(给定离散随机变量 ，其取值为 ，对应概率分布为 ，则 **归一化香农熵** 定义为：，取值范围 )。

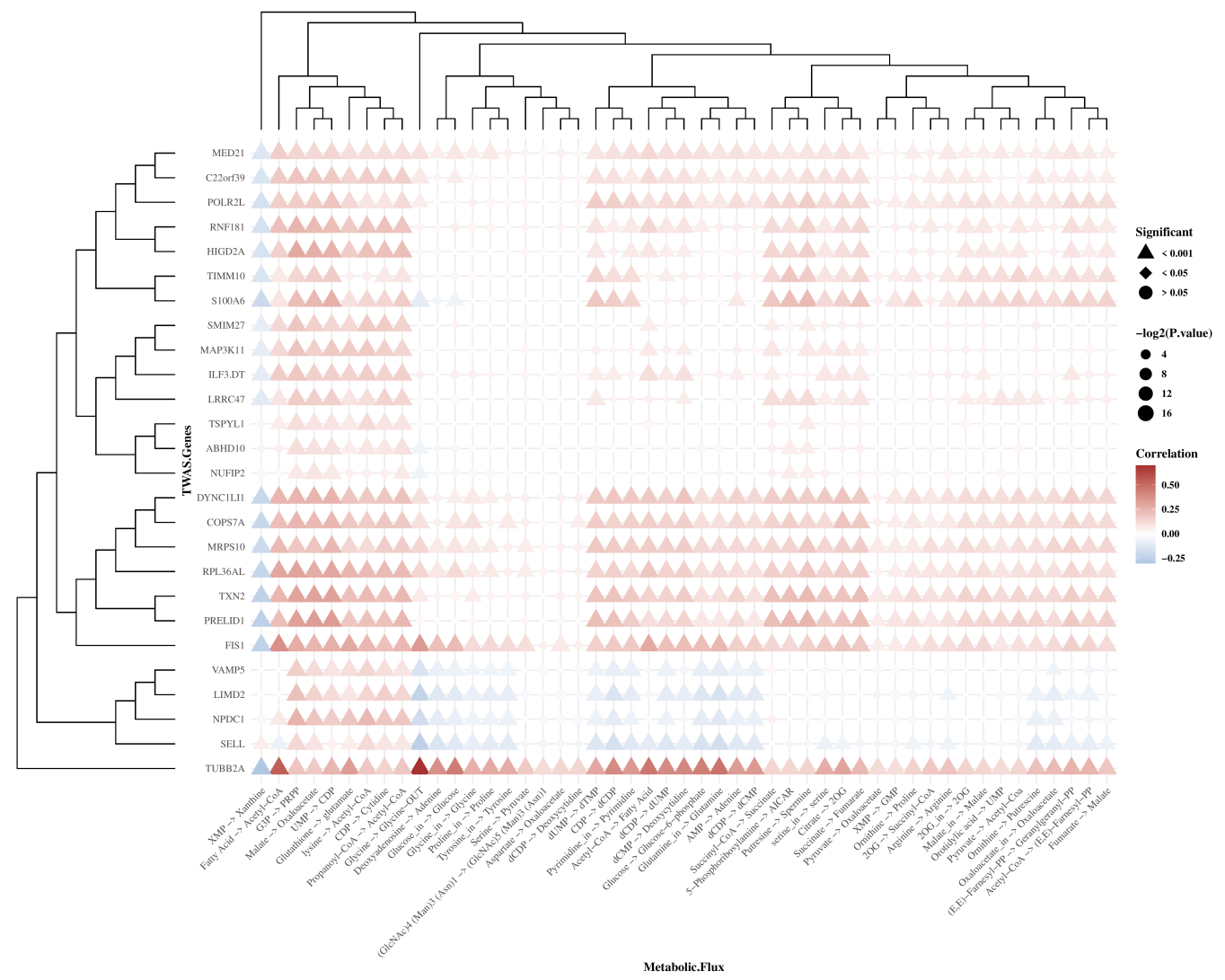
去除未知细胞 (unknown)。 去除 Monocyte 细胞。

在上述的细胞群中，分析两组 features (即，**基因集** (NAT10, LINC02193, DDX10P1, …[n = 116], 来自于FUSION TWAS全转录组关联研究[Section: AA]) ，与 **代谢通量** (Macrophage\_AA - Macrophage\_Normal, Myeloid\_cell…, …[n = 7], 来自于Limma 代谢通量差异分析[Section: AA\_FLUX]) )。

对于基因集，在各组细胞中，以阈值穿透率去除低表达的基因 (例如，去除总体表达为 0 的基因) (阈值：0，穿透率 cutoff：0.1) (设某基因 在细胞群体 中的表达值集合为 ，给定阈值 ，则 **阈值穿透率** 定义为： ( 是指示函数，当 时为 1，否则为 0))。

如果细胞群中，两组 features 均满足数量大于 3，则保留该细胞群，用于后续分析。

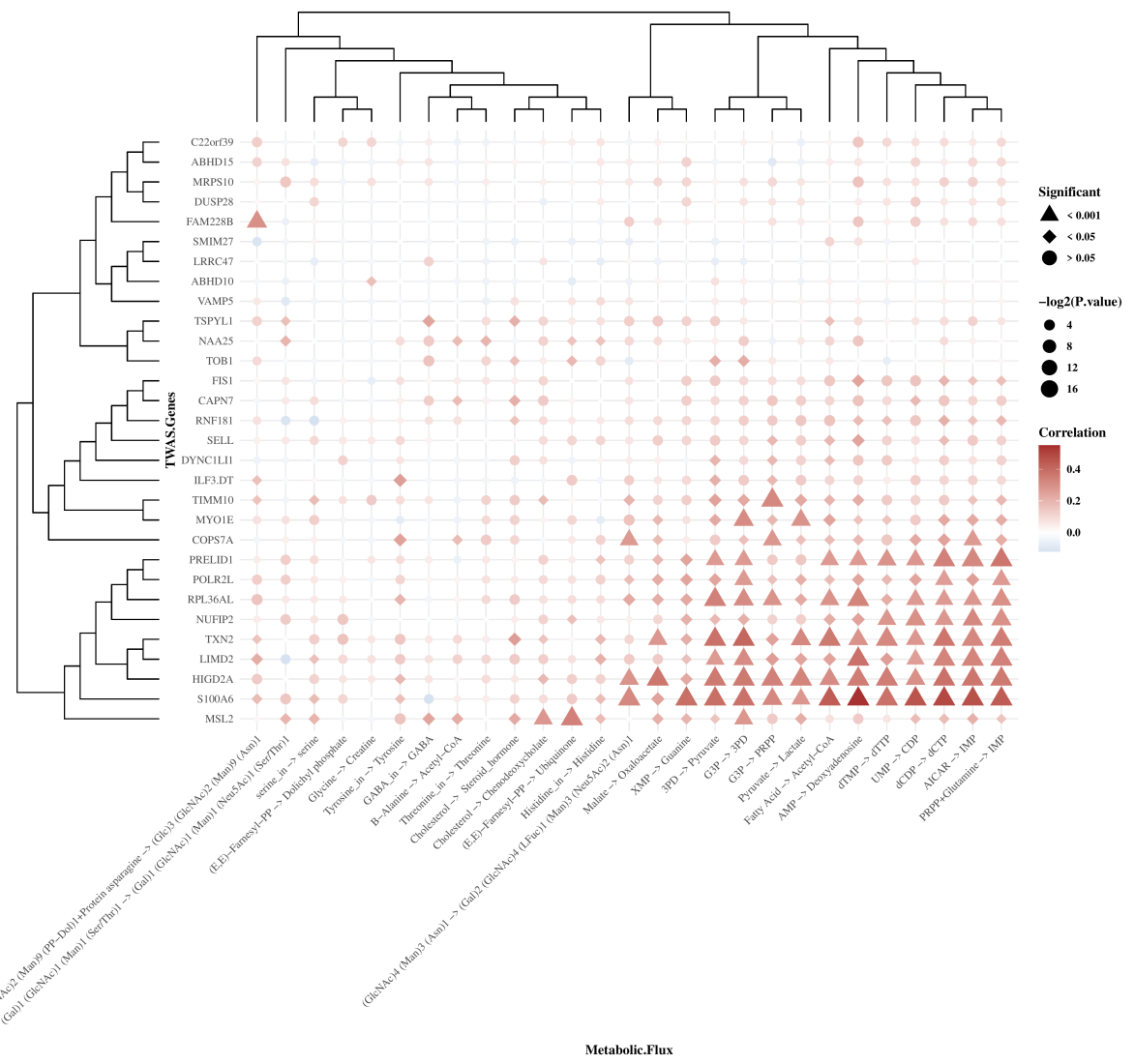
分别对各细胞群的两组 features 关联分析。 设定 P 阈值 (0.05) 与关联系数 (0.3) 阈值，获取各细胞群中的显著关联组。统计为: Hematopoietic stem cell (965), Plasmacytoid dendritic cell (210)。



**Fig.** **13** AA Hematopoietic stem cell correlation heatmap

**(File path: Figure+Table/3.8\_细胞群\_features\_关联分析\_(AA)/AA-Hematopoietic-stem-cell-correlation-heatmap.pdf)**

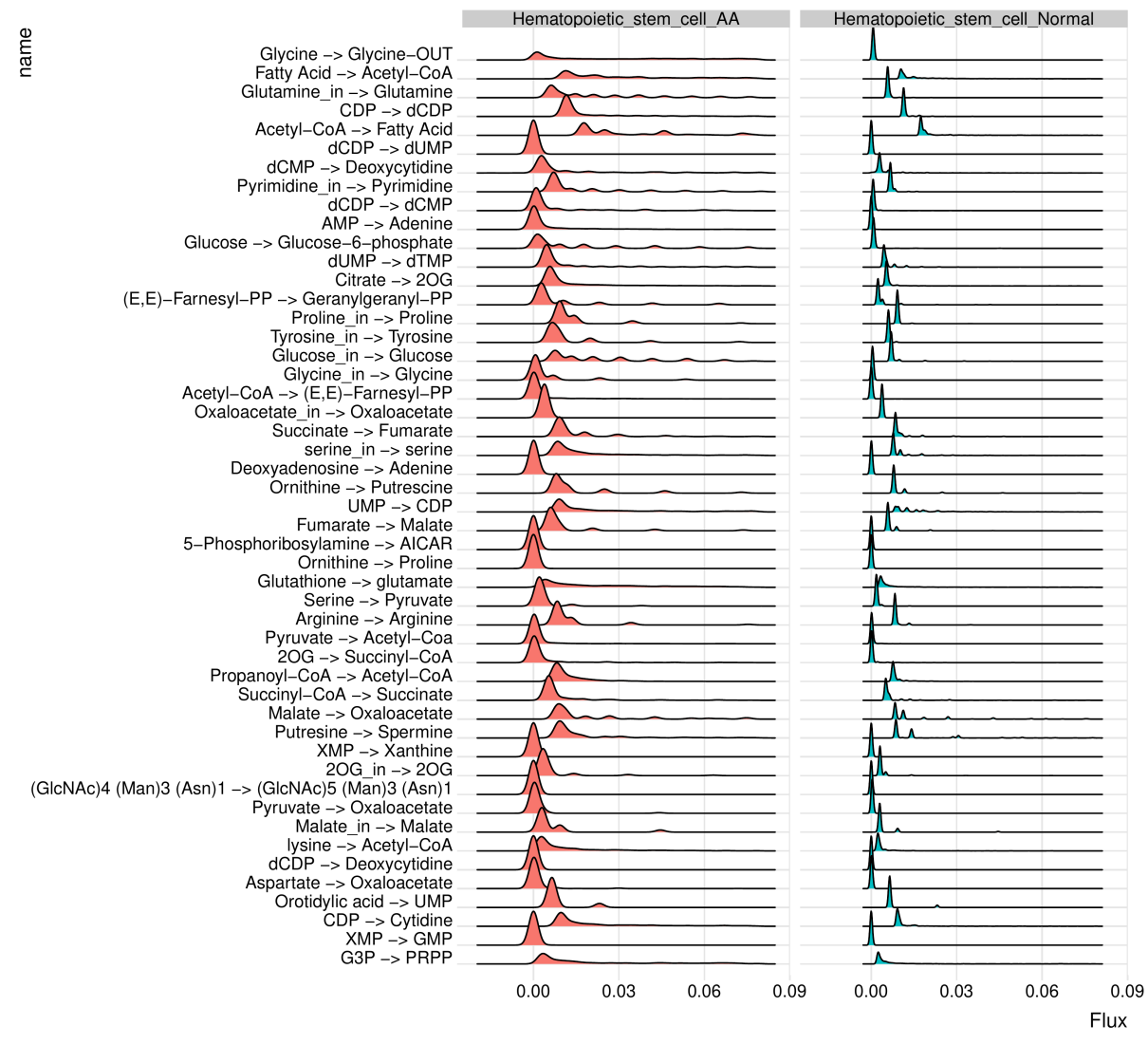
Fig. **[13](#AA-Hematopoietic-stem-cell-correlation-heatmap)** 为关联分析 (TWAS Genes, Metabolic Flux) 热图。



**Fig.** **14** AA Plasmacytoid dendritic cell correlation heatmap

**(File path: Figure+Table/3.8\_细胞群\_features\_关联分析\_(AA)/AA-Plasmacytoid-dendritic-cell-correlation-heatmap.pdf)**

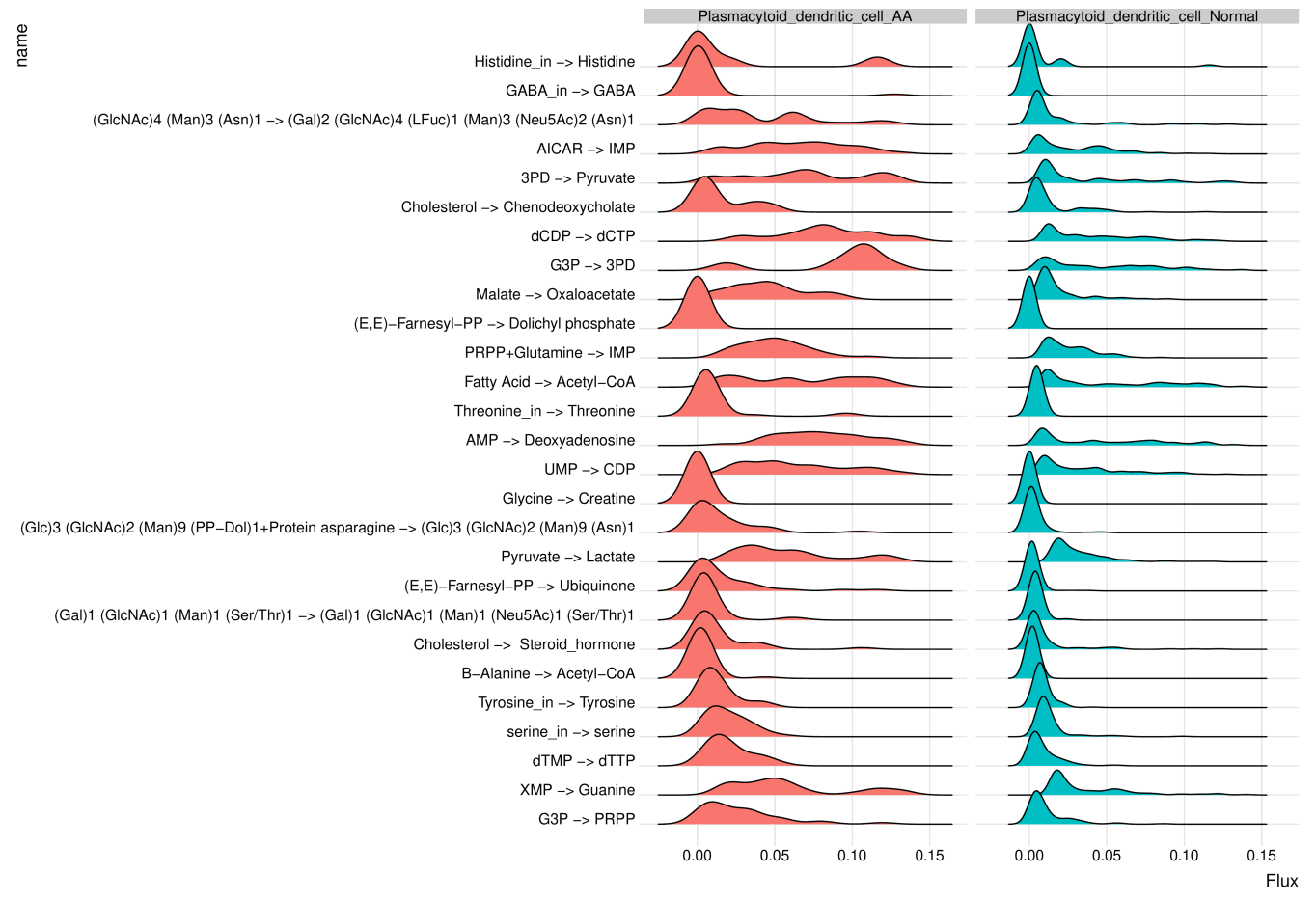
Fig. **[14](#AA-Plasmacytoid-dendritic-cell-correlation-heatmap)** 为关联分析 (TWAS Genes, Metabolic Flux) 热图。



**Fig.** **15** Hematopoietic stem cell Cell flux ridge plot

**(File path: Figure+Table/3.8\_细胞群\_features\_关联分析\_(AA)/Hematopoietic-stem-cell-Cell-flux-ridge-plot.pdf)**

Fig. **[15](#Hematopoietic-stem-cell-Cell-flux-ridge-plot)** 为细胞的代谢通量山脊图。



**Fig.** **16** Plasmacytoid dendritic cell Cell flux ridge plot

**(File path: Figure+Table/3.8\_细胞群\_features\_关联分析\_(AA)/Plasmacytoid-dendritic-cell-Cell-flux-ridge-plot.pdf)**

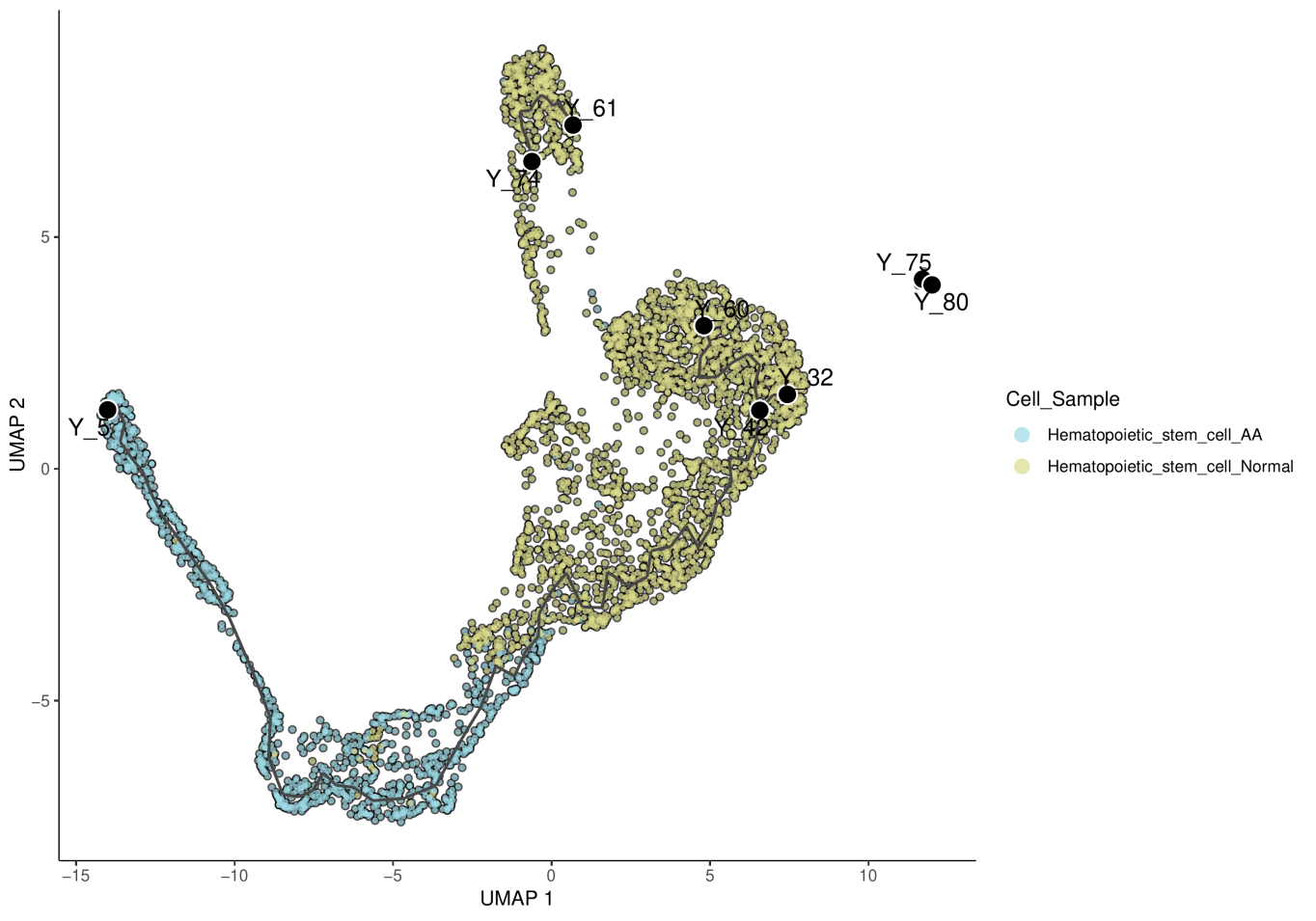
Fig. **[16](#Plasmacytoid-dendritic-cell-Cell-flux-ridge-plot)** 为细胞的代谢通量山脊图。

## 3.9 Monocle3 拟时分析 (HSC)

(Plasmacytoid\_dendritic\_cell 数量较少，未展开分析)

从 Seurat 数据对象 Cell\_Sample 中提取 Hematopoietic\_stem\_cell 类型的细胞，对其重新聚类分析。将 Seurat 数据对象转化为 Monocle3 数据对象 (详见方法章节)。构建轨迹图 (Trajectory)。选择 Y\_60 (principle points) 为拟时起点。构建细胞的拟时变化图 (Pseudotime)。寻找单细胞拟时轨迹 (monocle3::graph\_test) 中差异表达的基因。探究 **基因集** (TUBB2A, FIS1, PRELID1, …[n = 26], 来自于细胞群 features 关联分析[Section: AA]) 在拟时轨迹中的表达变化。这些基因中，共 25 属于 Moran’s I test (Graph Test) 差异表达基因。仅分析这些基因。

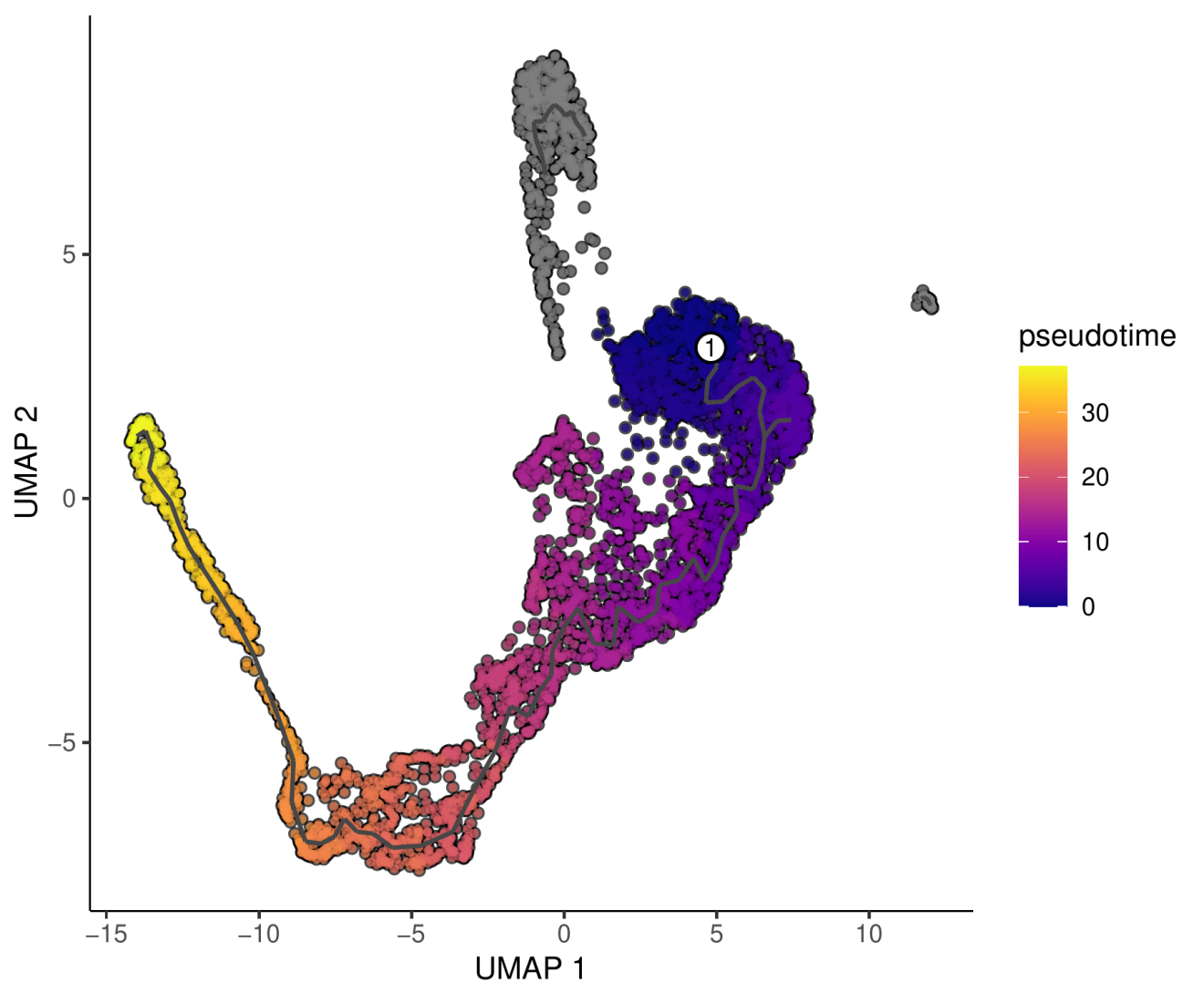
其中，SELL, TUBB2A, LIMD2, HIGD2A, FIS1, POLR2L 在细胞群呈趋势表达变化。



**Fig.** **17** HSC principal points

**(File path: Figure+Table/3.9\_Monocle3\_拟时分析\_(HSC)/HSC-principal-points.pdf)**

Fig. **[17](#HSC-principal-points)** 为拟时轨迹与 principal point 示意。



**Fig.** **18** HSC pseudotime

**(File path: Figure+Table/3.9\_Monocle3\_拟时分析\_(HSC)/HSC-pseudotime.pdf)**

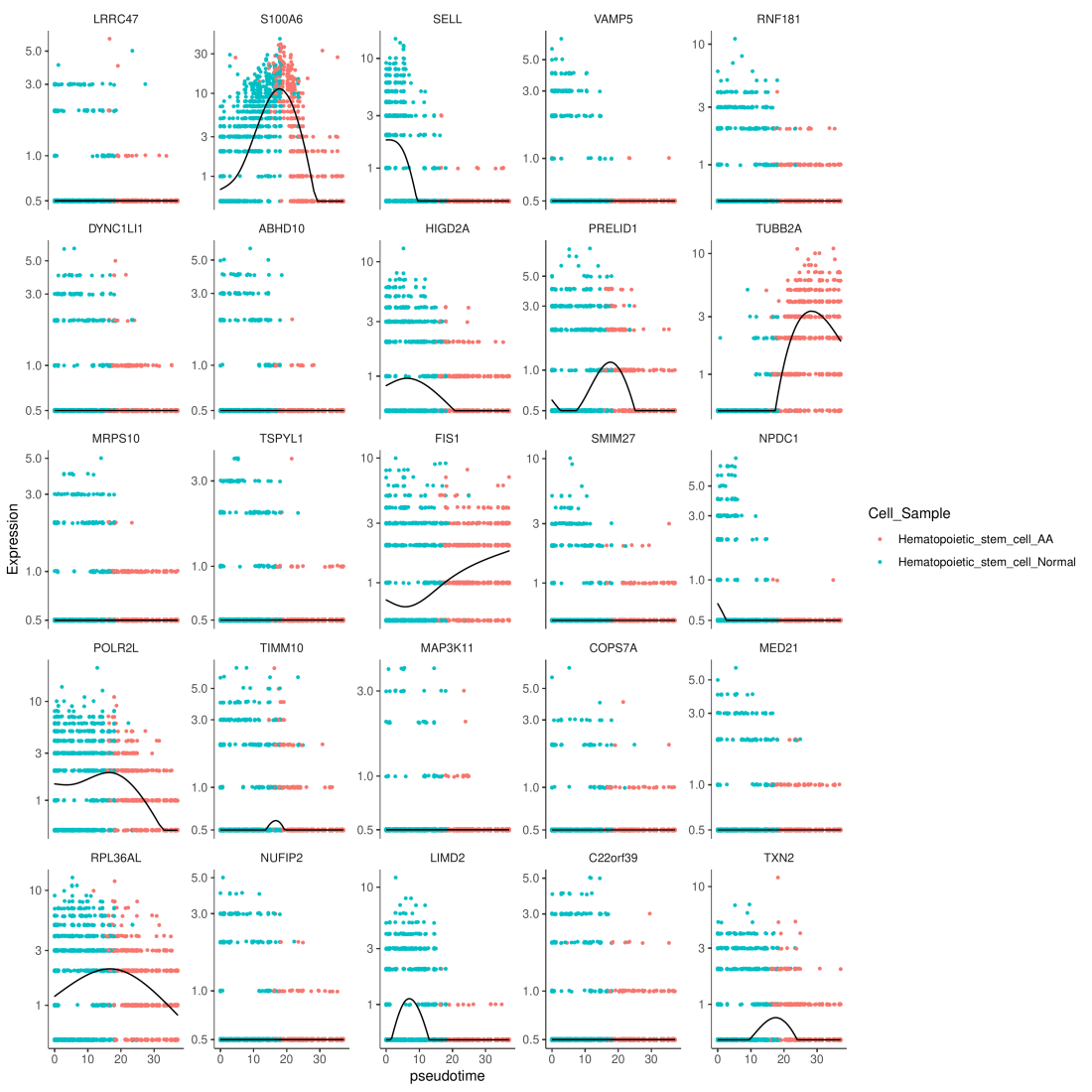
Fig. **[18](#HSC-pseudotime)** 为细胞拟时图。

**Tab.** **6** HSC Graph Test Significant genes

| Gene id | Status | P value | Morans test stati... | Morans I |
| --- | --- | --- | --- | --- |
| HBM | OK | 0 | 206.2 | 0.9518 |
| AHSP | OK | 0 | 203.7 | 0.9404 |
| SLC4A1 | OK | 0 | 202 | 0.9326 |
| ALAS2 | OK | 0 | 201.5 | 0.9302 |
| HBA1 | OK | 0 | 201.3 | 0.9293 |
| ... | ... | ... | ... | ... |

**(File path: Figure+Table/3.9\_Monocle3\_拟时分析\_(HSC)/HSC-Graph-Test-Significant-genes.csv)**

Tab. **[6](#HSC-Graph-Test-Significant-genes)** 为 Moran’s I test (Graph Test) 筛选的差异表达基因 (Q-value cutoff: 0.05)。



**Fig.** **19** HSC Set1 genes in pseudotime

**(File path: Figure+Table/3.9\_Monocle3\_拟时分析\_(HSC)/HSC-Set1-genes-in-pseudotime.pdf)**

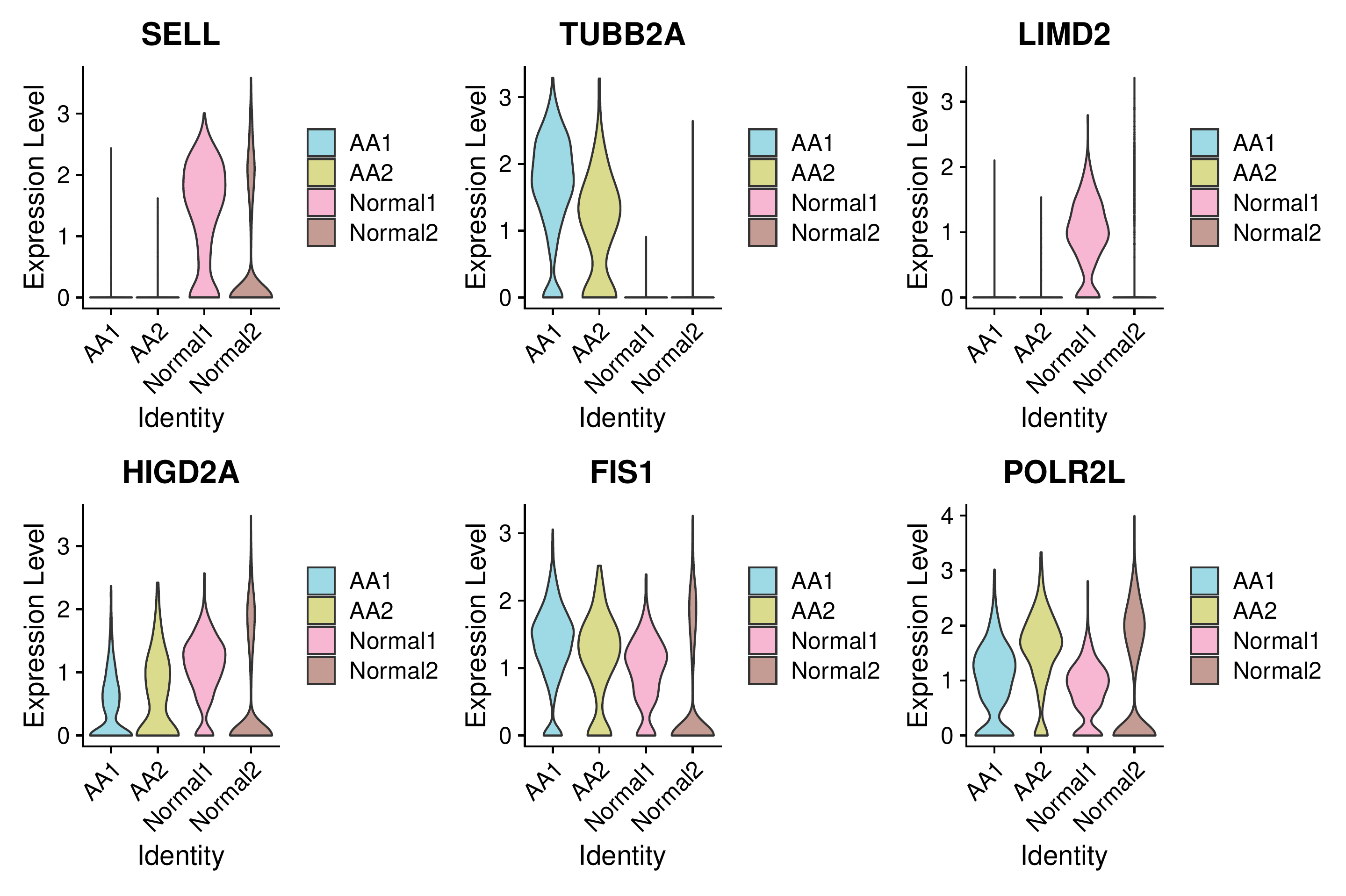
Fig. **[19](#HSC-Set1-genes-in-pseudotime)** 为 TUBB2A, FIS1, PRELID1, …(n = 25) 在拟时 (Pseudotime) 中的表达变化。



**Fig.** **20** AA dimension plot of expression level of the genes

**(File path: Figure+Table/3.9\_Monocle3\_拟时分析\_(HSC)/AA-dimension-plot-of-expression-level-of-the-genes.pdf)**

Fig. **[20](#AA-dimension-plot-of-expression-level-of-the-genes)** 基因 SELL, TUBB2A, LIMD2, …(n = 6) 表达水平的 Dimension reduction plot.

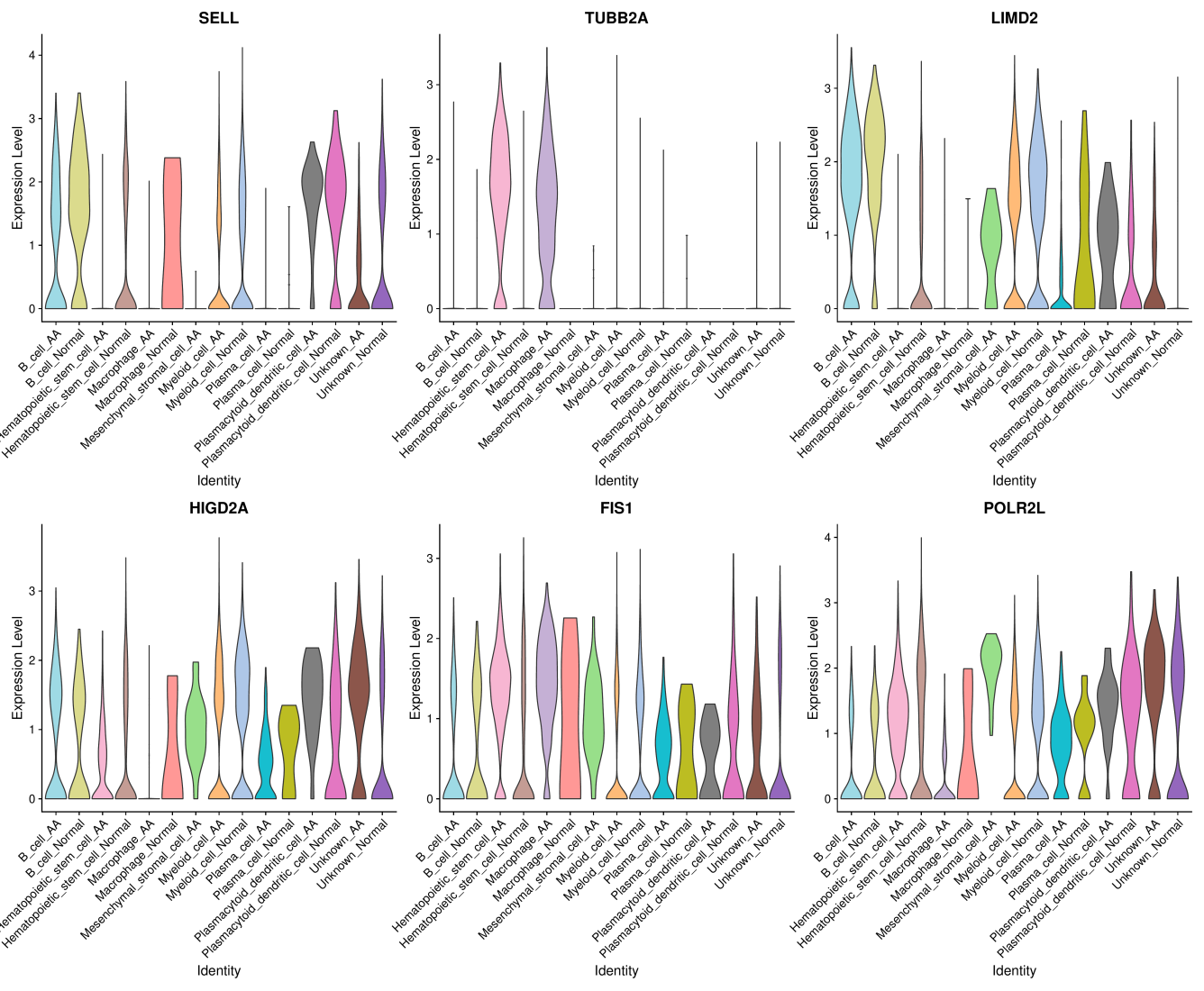


**Fig.** **21** AA violing plot of expression level of the genes

**(File path: Figure+Table/3.9\_Monocle3\_拟时分析\_(HSC)/AA-violing-plot-of-expression-level-of-the-genes.pdf)**

Fig. **[21](#AA-violing-plot-of-expression-level-of-the-genes)** 基因 SELL, TUBB2A, LIMD2, …(n = 6) 表达水平的小提琴图。

## 3.10 TUBB2A 等在其它细胞中的表达



**Fig.** **22** AA violing plot of expression level geneOthers

**(File path: Figure+Table/3.10\_TUBB2A\_等在其它细胞中的表达/AA-violing-plot-of-expression-level-geneOthers.pdf)**

Fig. **[22](#AA-violing-plot-of-expression-level-geneOthers)** 基因 **基因集** (SELL, TUBB2A, LIMD2, …[n = 6], 来自于Monocle3 拟时分析[Section: HSC]) 表达水平的小提琴图。

# 4 总结

本研究从遗传变异 (TWAS)，基因表达变化 (单细胞)，单细胞代谢通量变化三个维度，对 AA 的机制展开的探究。

首先从 GWAS Summary 数据，根据 TWAS 方法，筛选到 AA SNP 影响到表达水平变化的基因。 再通过分析 AA 单细胞数据 (bone marrow)，鉴定主要细胞群，预测这些细胞的代谢通量变化， 并分析各组细胞群之间的差异代谢通量。 推测，基因 (突变) 水平的改变，不仅影响到某些基因表达水平的变化，还会改变细胞的代谢行为。 因此，随后，本研究探究了 TWAS 筛选出的基因与代谢通量改变的关联性。 从 Hematopoietic\_stem\_cell 中发现了大量与代谢通量改变呈正相关的基因 (Fig. **[13](#AA-Hematopoietic-stem-cell-correlation-heatmap)** )。 通过对这些基因的拟时分析 (Fig. **[19](#HSC-Set1-genes-in-pseudotime)** )， 可以发现 TUBB2A、SELL 等基因在 AA 与 Normal 中显著差异表达 (Fig. **[21](#AA-violing-plot-of-expression-level-of-the-genes)** ， Fig. **[20](#AA-dimension-plot-of-expression-level-of-the-genes)** )。 TUBB2A 在 AA 中高表达，而 SELL 在 AA 中低表达。

# Reference

1. Murphy, A. E., Schilder, B. M. & Skene, N. G. MungeSumstats: A bioconductor package for the standardization and quality control of many gwas summary statistics. *Bioinformatics* **37**, 4593–4596 (2021).

2. McLaren, W. *et al.* The ensembl variant effect predictor. *Genome Biology* **17**, (2016).

3. Bulik-Sullivan, B. K. *et al.* LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature genetics* **47**, 291–295 (2015).

4. Gusev, A. *et al.* Integrative approaches for large-scale transcriptome-wide association studies. *Nature Genetics* **48**, 245–252 (2016).

5. Cao, Y., Wang, X. & Peng, G. SCSA: A cell type annotation tool for single-cell rna-seq data. *Frontiers in genetics* **11**, (2020).

6. Zhang, X. *et al.* CellMarker: A manually curated resource of cell markers in human and mouse. *Nucleic acids research* **47**, D721–D728 (2019).

7. Alghamdi, N. *et al.* A graph neural network model to estimate cell-wise metabolic flux using single-cell rna-seq data. *Genome research* **31**, 1867–1884 (2021).

8. Smyth, G. K. Limma: Linear models for microarray data. in *Bioinformatics and Computational Biology Solutions Using R and Bioconductor* (eds. Gentleman, R., Carey, V. J., Huber, W., Irizarry, R. A. & Dudoit, S.) 397–420 (Springer-Verlag, 2005). doi:[10.1007/0-387-29362-0\_23](https://doi.org/10.1007/0-387-29362-0_23).