**生信分析报告**

**项目标题： 再生障碍性贫血 ;**

**单 号： BSCL241113 ;**

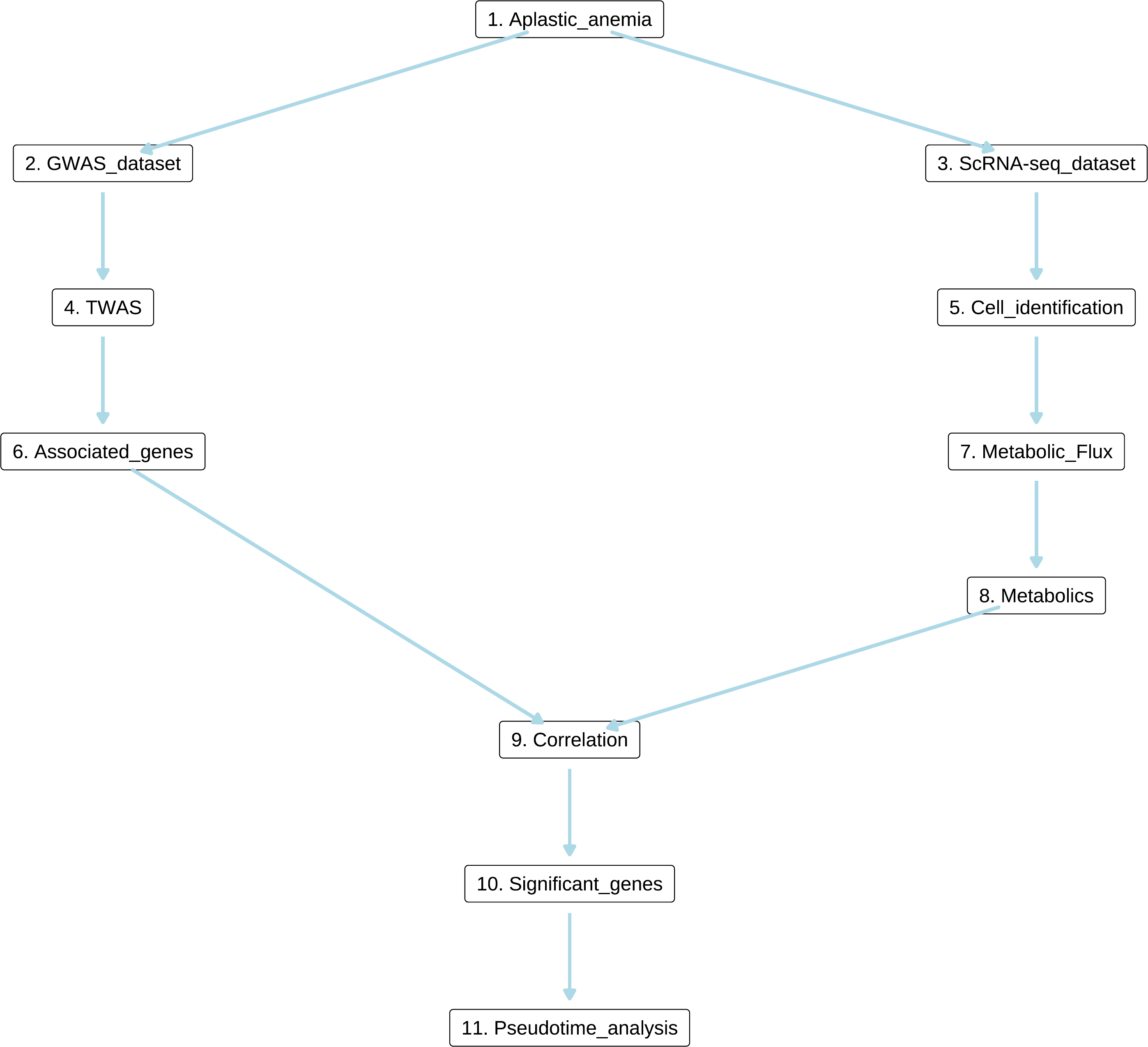
**分析人员： 黄礼闯 ;**

**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 0 个基因，并且基因数量小于 5000。线粒体基因的比例小于 20%。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。

## 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)5。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)6 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：Erythroblast\_AA vs Erythroblast\_Normal, T\_Cell\_AA vs T\_Cell\_Normal, Monocyte\_AA vs Monocyte\_Normal, Natural\_Killer\_Cell\_AA vs Natural\_Killer\_Cell\_Normal, Plasma\_Cell\_AA vs Plasma\_Cell\_Normal, B\_Cell\_AA vs B\_Cell\_Normal, Neutrophil\_AA vs Neutrophil\_Normal, Unknown\_AA vs Unknown\_Normal, Macrophage\_AA vs Macrophage\_Normal, Pre\_B\_Cell\_AA vs Pre\_B\_Cell\_Normal, Platelet\_AA vs Platelet\_Normal, Proliferating\_Cell\_AA vs Proliferating\_Cell\_Normal, Mesenchymal\_Stromal\_Cell\_AA vs Mesenchymal\_Stromal\_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: AA\_ERY)

从 Seurat 数据对象 Cell\_Sample 中提取 Erythroblast 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\_83, Y\_113 (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)** 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)** 为 TWAS 显著统计表 (TWAS.P < 0.05)。

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

| PANEL | FILE | ID | SYMBOL | CHR | P0 | P1 | HSQ | BEST.G......9 | BEST.G......10 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| GTExv8... | /home/... | ENSG00... |  | 7 | 1.29e+08 | 1.29e+08 | 0.2538 | Rs6969930 | -4.33 |
| GTExv8... | /home/... | ENSG00... | NAT10 | 11 | 34110000 | 34110000 | 0.0505 | Rs1176363 | 2.86 |
| GTExv8... | /home/... | ENSG00... |  | 7 | 128900000 | 128900000 | 0.447 | Rs6969930 | -4.33 |
| GTExv8... | /home/... | ENSG00... |  | 4 | 102500000 | 102500000 | 0.04 | Rs4648011 | 3.77 |
| GTExv8... | /home/... | ENSG00... |  | 9 | 133300000 | 133300000 | 0.6183 | Rs1099... | 3.23 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |

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

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

| PANEL | FILE | ID | SYMBOL | CHR | P0 | P1 | HSQ | BEST.G......9 | BEST.G......10 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| GTExv8... | /home/... | ENSG00... |  | 7 | 1.29e+08 | 1.29e+08 | 0.2538 | Rs6969930 | -4.33 |
| GTExv8... | /home/... | ENSG00... | NAT10 | 11 | 34110000 | 34110000 | 0.0505 | Rs1176363 | 2.86 |
| GTExv8... | /home/... | ENSG00... |  | 7 | 128900000 | 128900000 | 0.447 | Rs6969930 | -4.33 |
| GTExv8... | /home/... | ENSG00... |  | 4 | 102500000 | 102500000 | 0.04 | Rs4648011 | 3.77 |
| GTExv8... | /home/... | ENSG00... |  | 9 | 133300000 | 133300000 | 0.6183 | Rs1099... | 3.23 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |

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

## 3.4 GEO 数据获取 (AA\_SCRNA)

以 GEOquery 获取 GSE181989 的数据信息。

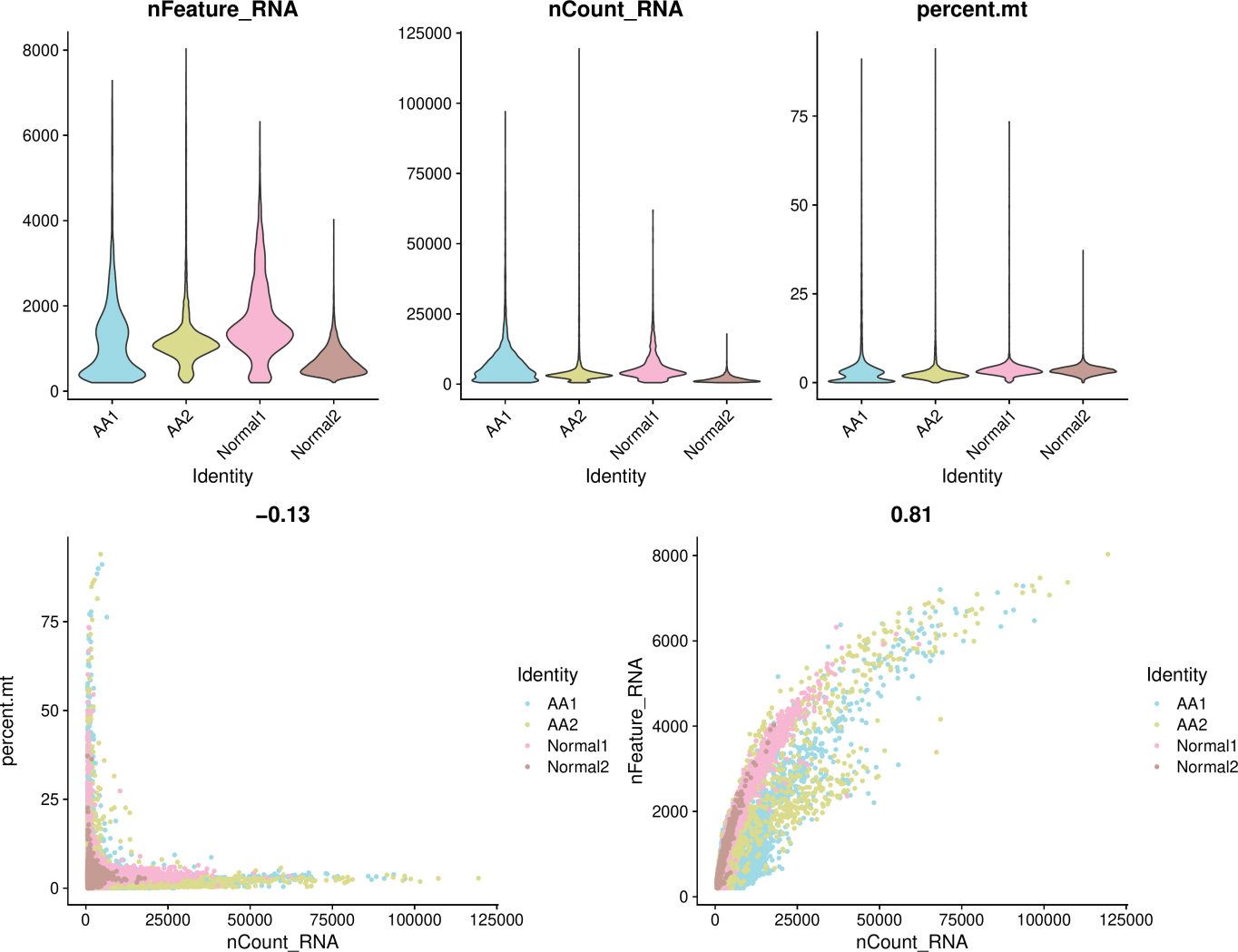
* Data Source ID: GSE181989
* data\_processing: The Cell Ranger software pipeline (version 2.2.0) provided by 10×Genomics was used to demultiplex cellular barcodes, map reads to the genome and transcriptome using the STAR aligner, and down-sample reads as required to generate normalized aggregate data 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)

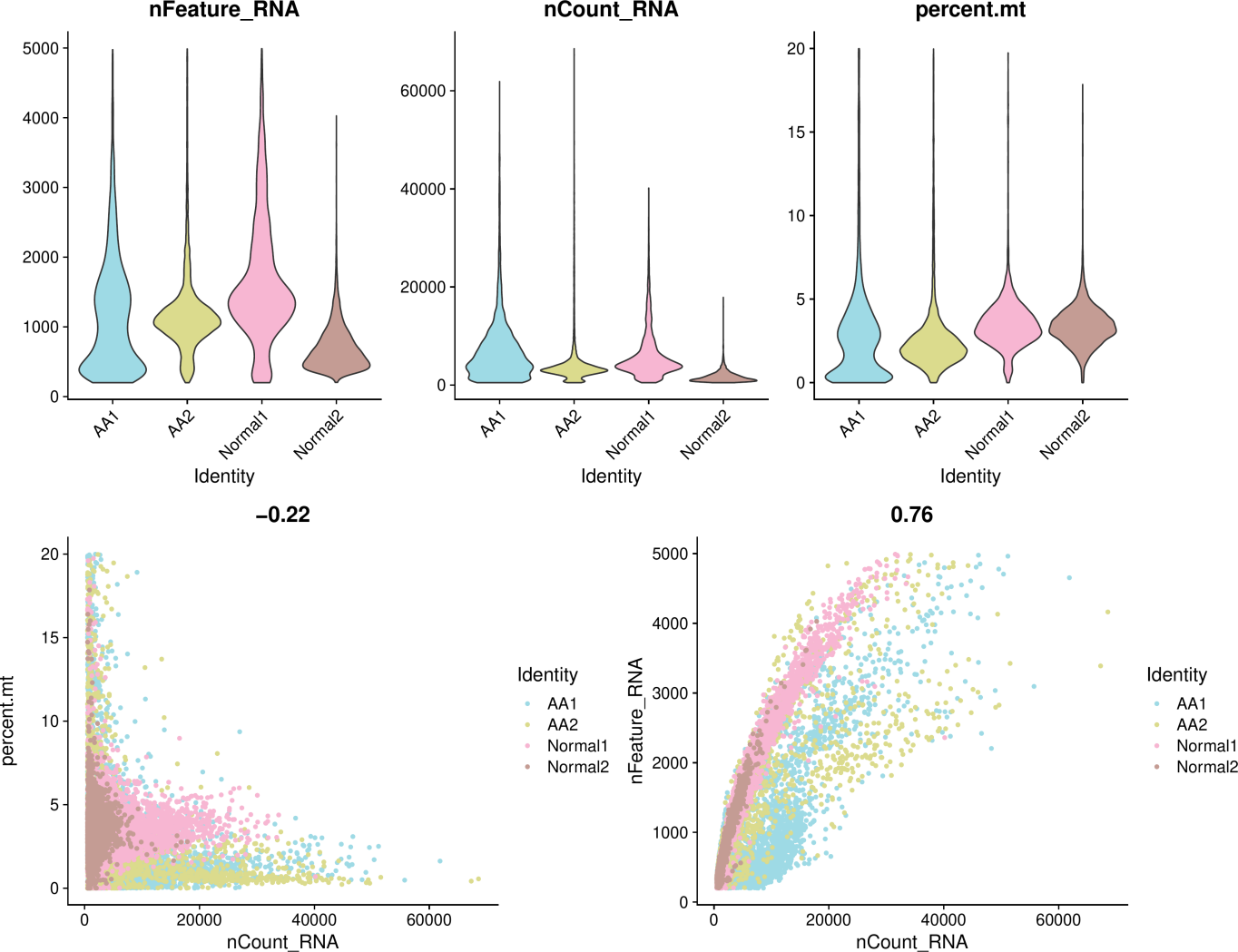
读取 AA1, AA2, Normal1, Normal2 样本的数据集。前期质量控制，一个细胞至少应有 0 个基因，并且基因数量小于 5000。线粒体基因的比例小于 20%。数据归一化，PCA 聚类 (Seurat 标准工作流，见方法章节) 后，绘制 PC standard deviations 图。去除批次效应后 (详见方法章节) ，在 1-15 PC 维度，2 分辨率下，对细胞群 UMAP 聚类。计算所有细胞群的 Marker。根据细胞群 Markers (检出率至少为 0.7，选取 Top 30) ，让 ChatGPT-4 对细胞类型注释。 对细胞群差异分析 (依据 Cell\_Sample)，筛选差异表达基因。

Fig. **[2](#Pre-Quality-control)** 为 QC (质量控制) 图 (数据过滤前) 。 Fig. **[3](#AA-After-Quality-control)** 为数据过滤后的 QC 图。 Fig. **[4](#AA-Standard-deviations-of-PCs)** 为主成分 (PC) 的 Standard deviations。 Fig. **[5](#AA-UMAP-Unintegrated)** 为去除批次效应之前的 UMAP 聚类图。 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)** 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-ChatGPT-Cell-type-annotation)** ChatGPT 对细胞群的注释结果。 Fig. **[8](#AA-Markers-in-cell-types)** 为 ChatGPT 注释细胞群使用的首要 Marker 热图。 Fig. **[9](#AA-ChatGPT4-Cell-Proportions-in-each-sample)** 为 ChatGPT-4 注释的细胞群在各个样本中的占比。 Fig. **[10](#AA-The-Cell-Sample)** 为 Cell\_Sample 的 umap 聚类图。 Tab. **[4](#AA-DEGs-of-the-contrasts)** 细胞群差异表达基因附表 (|log2(FC)| > 1, P-Adjust < 0.5)。



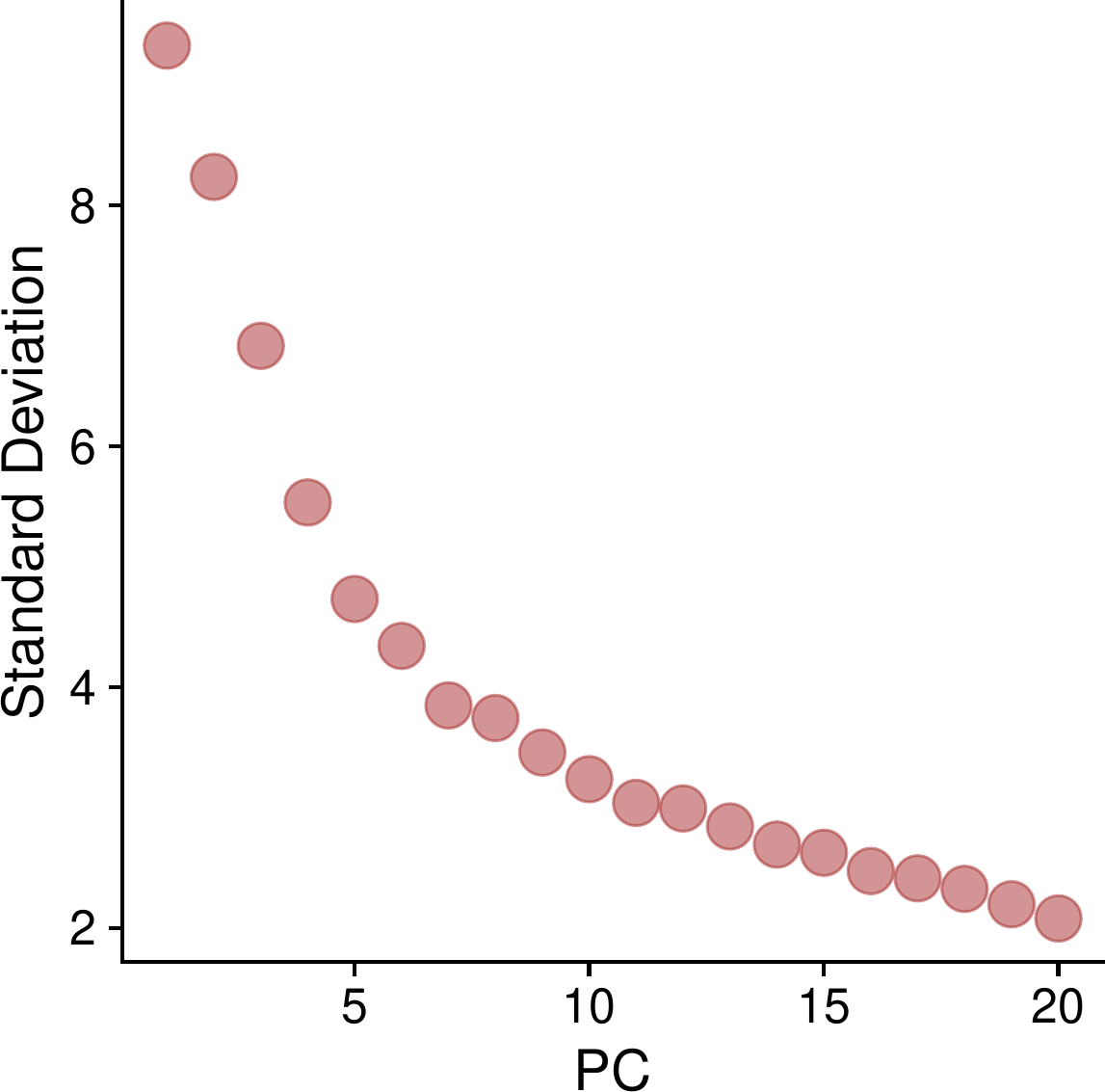
**Fig.** **2** Pre Quality control

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



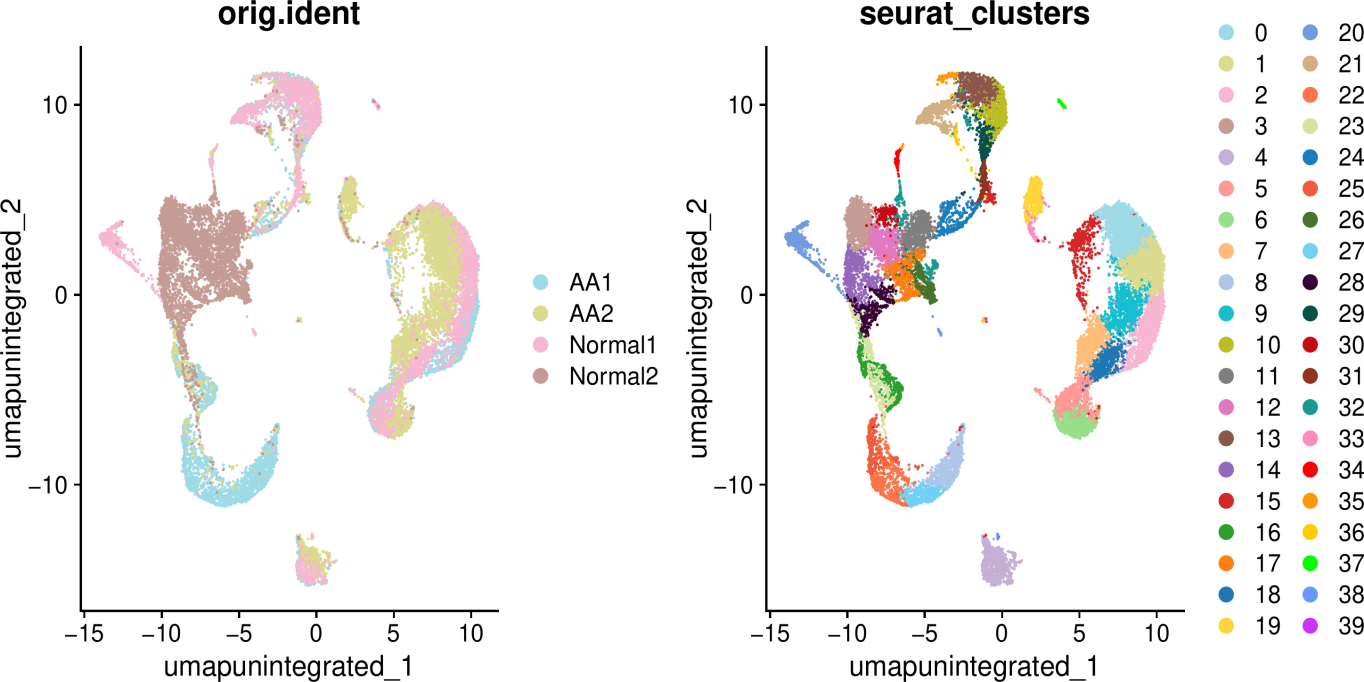
**Fig.** **3** AA After Quality control

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



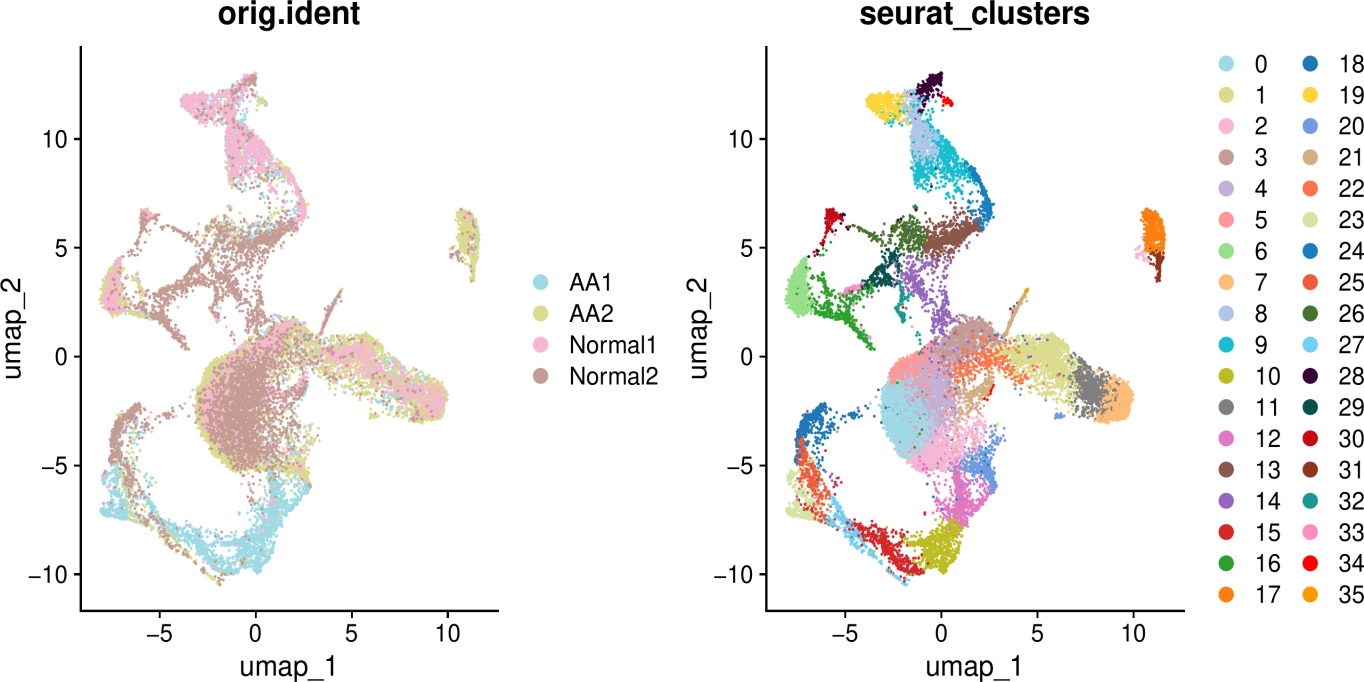
**Fig.** **4** AA Standard deviations of PCs

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



**Fig.** **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)**



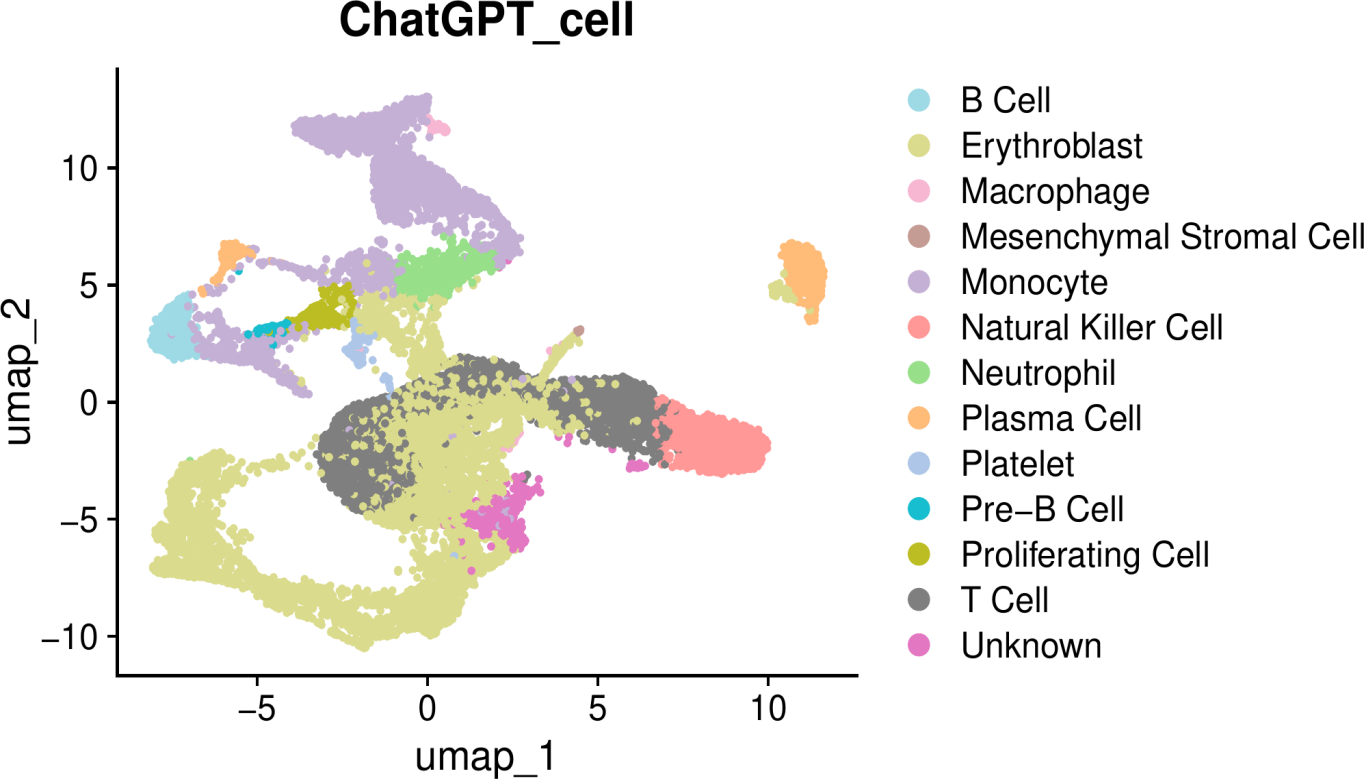
**Fig.** **6** AA UMAP Integrated

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

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

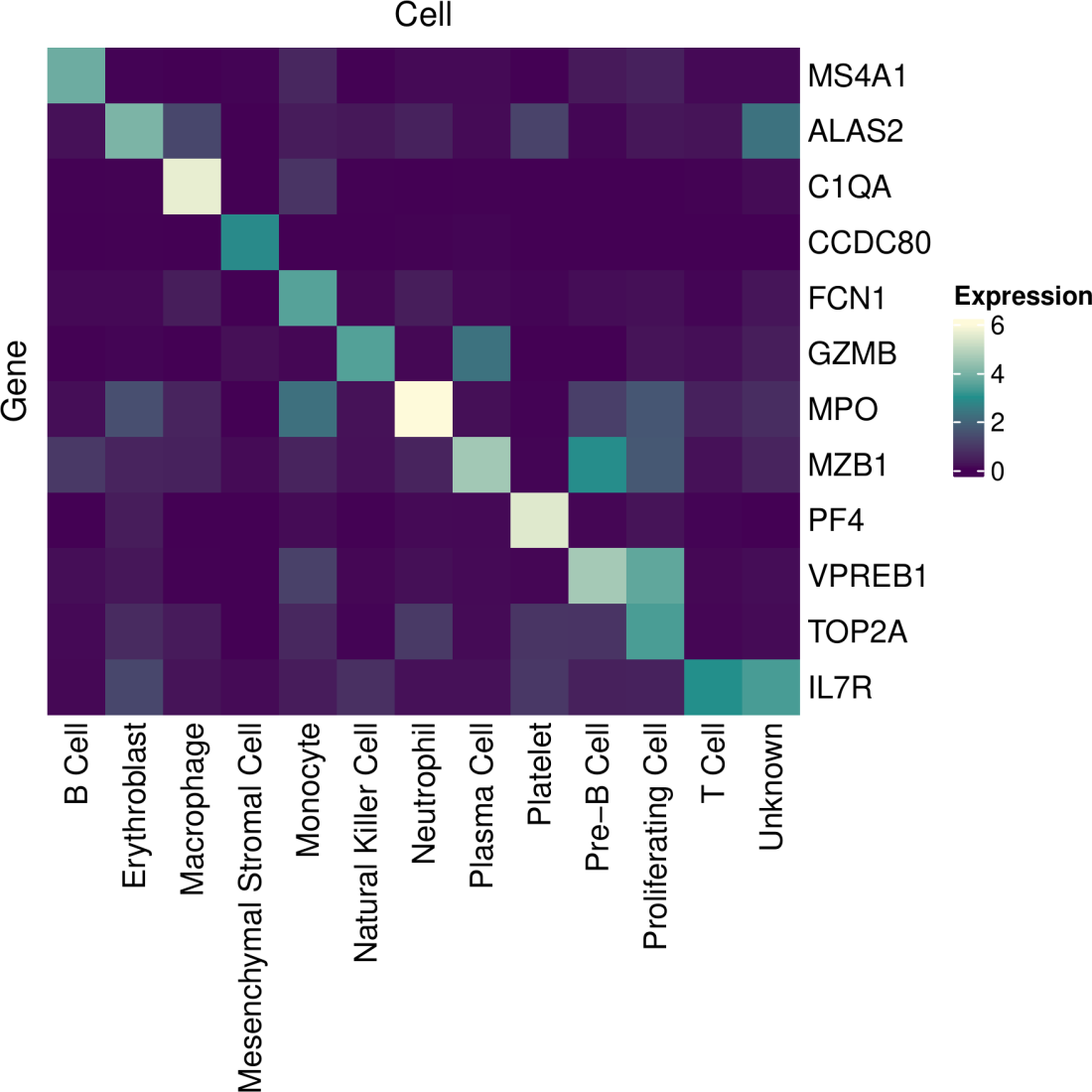
| Rownames | P val | Avg lo... | Pct.1 | Pct.2 | P val adj | Cluster | Gene |
| --- | --- | --- | --- | --- | --- | --- | --- |
| IL7R | 0 | 1.342 | 0.722 | 0.22 | 0 | 0 | IL7R |
| CD3E | 0 | 1.221 | 0.814 | 0.32 | 0 | 0 | CD3E |
| LEF1 | 0 | 2.541 | 0.519 | 0.096 | 0 | 0 | LEF1 |
| LTB | 0 | 1.102 | 0.701 | 0.284 | 0 | 0 | LTB |
| TCF7 | 0 | 2.394 | 0.531 | 0.116 | 0 | 0 | TCF7 |
| ... | ... | ... | ... | ... | ... | ... | ... |

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



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

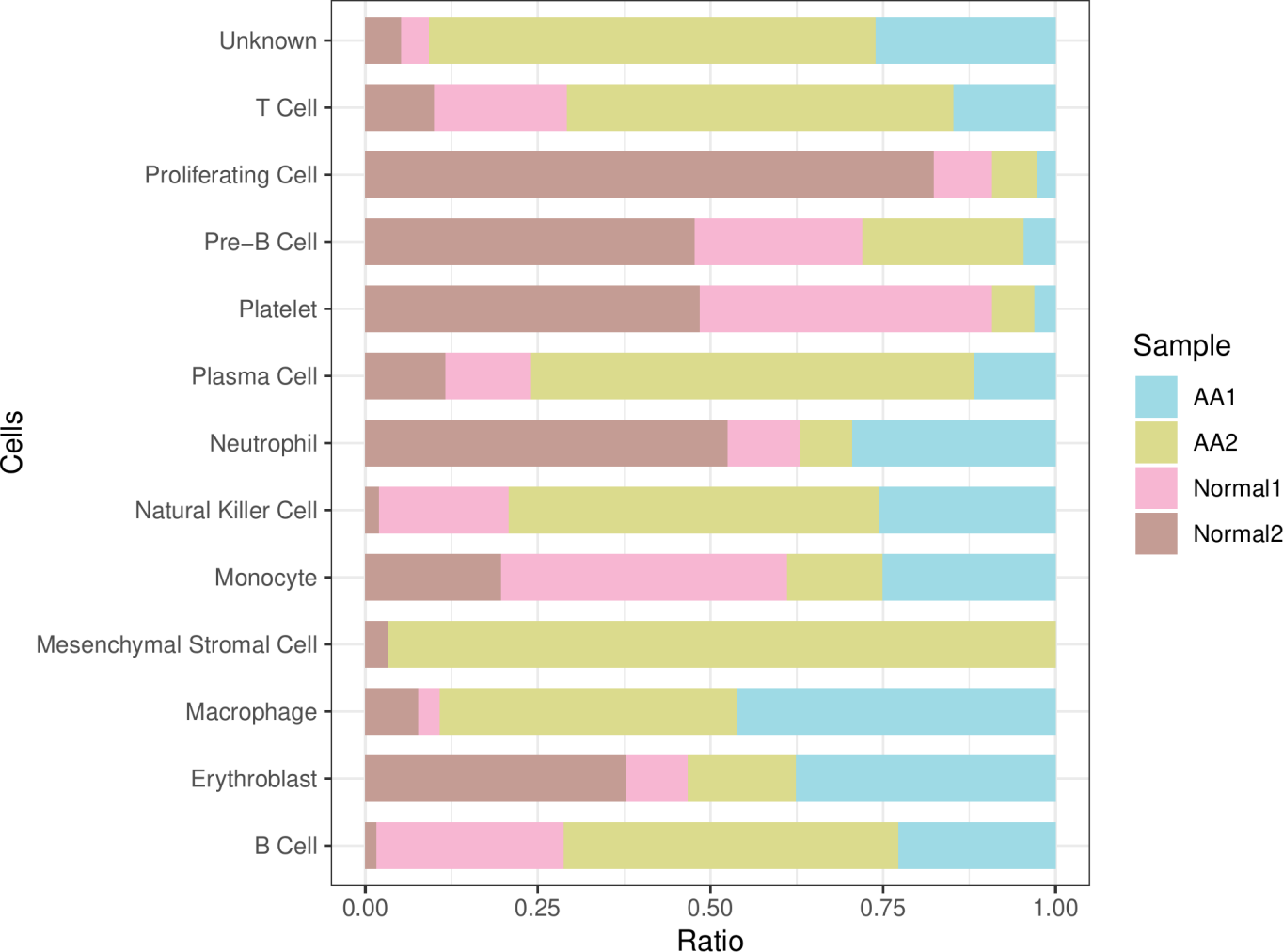
**(File path: Figure+Table/3.5\_Seurat\_集成单细胞数据分析\_(AA)/AA-ChatGPT-Cell-type-annotation.pdf)**



**Fig.** **8** AA Markers in cell types

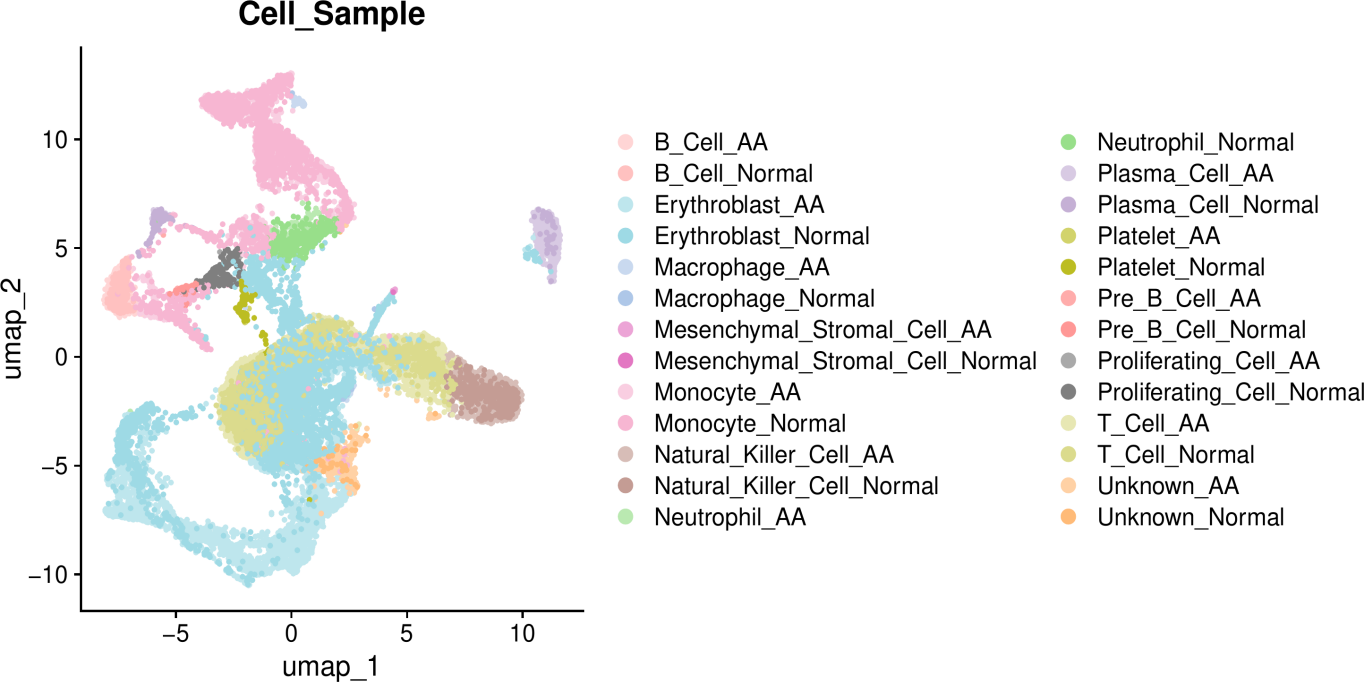
**(File path: Figure+Table/3.5\_Seurat\_集成单细胞数据分析\_(AA)/AA-Markers-in-cell-types.pdf)**

**(See: Figure+Table/3.5\_Seurat\_集成单细胞数据分析\_(AA)/AA-Markers-in-cell-types-content)**



**Fig.** **9** AA ChatGPT4 Cell Proportions in each sample

**(File path: Figure+Table/3.5\_Seurat\_集成单细胞数据分析\_(AA)/AA-ChatGPT4-Cell-Proportions-in-each-sample.pdf)**



**Fig.** **10** AA The Cell Sample

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

**Tab.** **4** AA DEGs of the contrasts

| Contrast | P val | Avg log2FC | Pct.1 | Pct.2 | P val adj | Gene |
| --- | --- | --- | --- | --- | --- | --- |
| Erythrobla... | 0 | 3.338 | 0.756 | 0.108 | 0 | HBA2 |
| Erythrobla... | 0 | 5.391 | 0.698 | 0.053 | 0 | ALAS2 |
| Erythrobla... | 0 | 4.543 | 0.677 | 0.045 | 0 | SLC4A1 |
| Erythrobla... | 0 | 3.468 | 0.752 | 0.126 | 0 | HBA1 |
| Erythrobla... | 0 | 3.447 | 0.691 | 0.066 | 0 | HBM |
| ... | ... | ... | ... | ... | ... | ... |

**(File path: Figure+Table/3.5\_Seurat\_集成单细胞数据分析\_(AA)/AA-DEGs-of-the-contrasts.csv)**

## 3.6 Seurat 细胞群中的 TWAS 风险相关基因 (AA)

分析基因集 (NAT10, LINC02193, DDX10P1, …[n = 116], 来自于FUSION TWAS全转录组关联研究[Section: AA]) 是否在各个细胞群中差异表达。筛选 TWAS 风险相关的 DEGs (筛选至少有检出率 10% 的细胞表达该基因)：Erythroblast\_AA\_vs\_Erythroblast\_Normal (n=12) , Neutrophil\_AA\_vs\_Neutrophil\_Normal (n=3) , Plasma\_Cell\_AA\_vs\_Plasma\_Cell\_Normal (n=4) , Proliferating\_Cell\_AA\_vs\_Proliferating\_Cell\_Normal (n=7) , T\_Cell\_AA\_vs\_T\_Cell\_Normal (n=1) 。 Tab. **[5](#AA-Filtered-TWAS-associated-genes-of-Cell-Cluster-DEGs)** TWAS 风险相关的细胞群 DEGs (筛选至少有检出率 10% 的细胞表达该基因)。

**Tab.** **5** AA Filtered TWAS associated genes of Cell Cluster DEGs

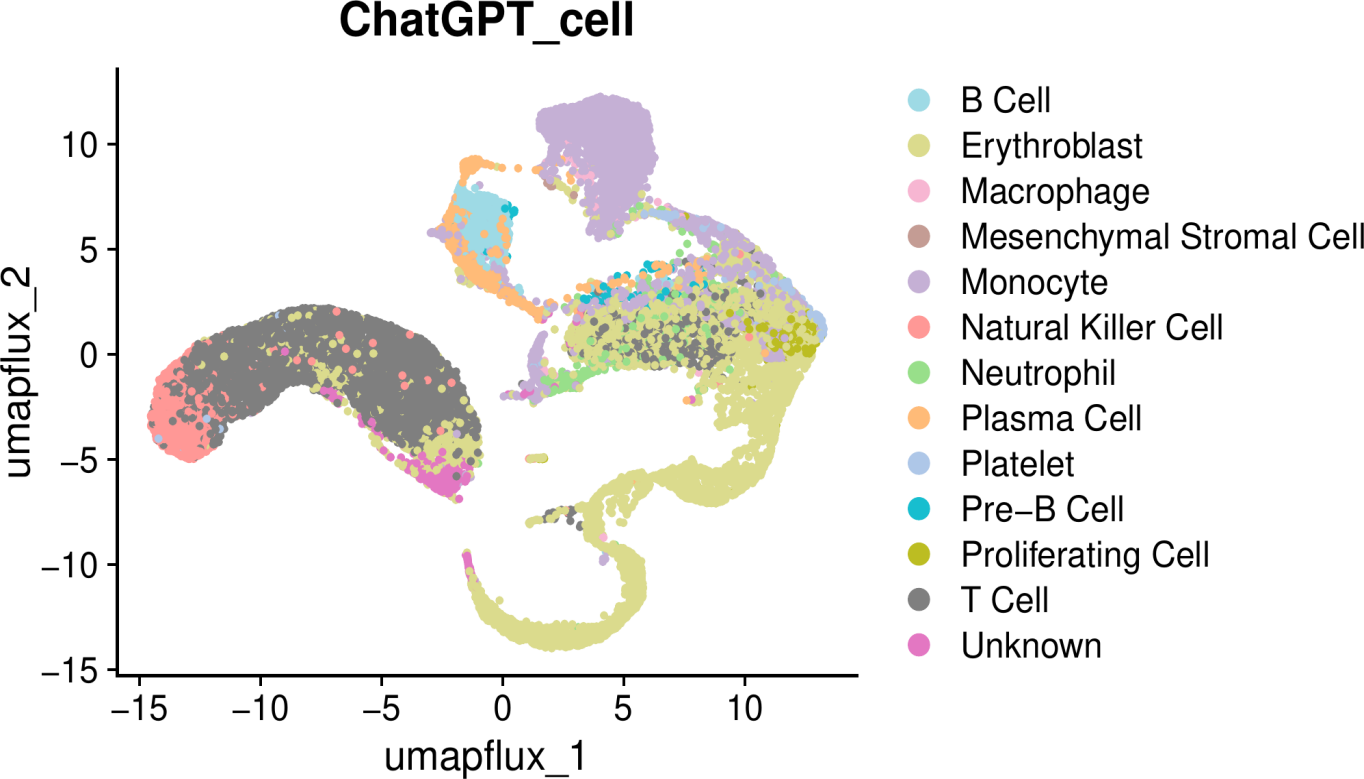
| Gene | Contrast | P val | Avg lo... | Pct.1 | Pct.2 | P val adj | TWAS.Z | TWAS.P |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ABHD10 | Erythr... | 2.778e-16 | -1.739 | 0.055 | 0.106 | 5.307e-12 | 2.877 | 0.00401 |
| ABHD15 | Plasma... | 1.826e-11 | -4.215 | 0.026 | 0.15 | 3.489e-07 | 2.121 | 0.03392 |
| IGKV3D-15 | Plasma... | 4.976e-07 | 3.57 | 0.141 | 0.01 | 0.009506 | 2.013 | 0.04406 |
| IGKV3D-15 | Prolif... | 5.19e-08 | 8.555 | 0.125 | 0 | 0.0009915 | 2.013 | 0.04406 |
| ILF3-DT | Erythr... | 1.538e-08 | -1.49 | 0.092 | 0.13 | 0.0002939 | 2.037 | 0.04167 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |

**(File path: Figure+Table/3.6\_Seurat\_细胞群中的\_TWAS\_风险相关基因\_(AA)/AA-Filtered-TWAS-associated-genes-of-Cell-Cluster-DEGs.csv)**

## 3.7 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. **[11](#AA-cells-metabolic-flux)** 为细胞代谢通量 (scFEA 预测，输入 Seurat) 的 UMAP 聚类。 Tab. **[6](#AA-metabolic-flux-matrix)** 为细胞代谢通量矩阵 (各 M\_ 为代谢模块)。



**Fig.** **11** AA cells metabolic flux

**(File path: Figure+Table/3.7\_scFEA\_单细胞数据的代谢通量预测\_(AA)/AA-cells-metabolic-flux.pdf)**

**Tab.** **6** AA metabolic flux matrix

| V1 | M 1 | M 2 | M 3 | M 4 | M 5 | M 6 | M 7 | M 8 | M 9 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| AAACCT... | -5.593... | 4.424e-09 | 0.007672 | 0.01157 | 9.619e-06 | 0.02032 | -1.66e-06 | 0.004383 | 7.848e-05 |
| AAACCT... | 0.1784 | 0.09102 | 0.3032 | 0.1645 | 1.696e-05 | 0.1946 | -2.91e-11 | 0.00438 | 0.01544 |
| AAACCT... | -5.593... | 0.01488 | 0.0685 | 0.2 | 9.619e-06 | 0.02498 | -1.66e-06 | 0.005351 | 7.848e-05 |
| AAACCT... | -9.518... | 4.424e-09 | 0.002445 | 0.06543 | 9.619e-06 | 0.01855 | -1.66e-06 | 0.004383 | 7.848e-05 |
| AAACCT... | -5.593... | 4.424e-09 | 0.002711 | 0.0618 | 9.619e-06 | 0.01855 | -1.66e-06 | 0.004383 | 7.848e-05 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |

**(File path: Figure+Table/3.7\_scFEA\_单细胞数据的代谢通量预测\_(AA)/AA-metabolic-flux-matrix.csv)**

## 3.8 Limma 代谢通量差异分析 (AA\_FLUX)

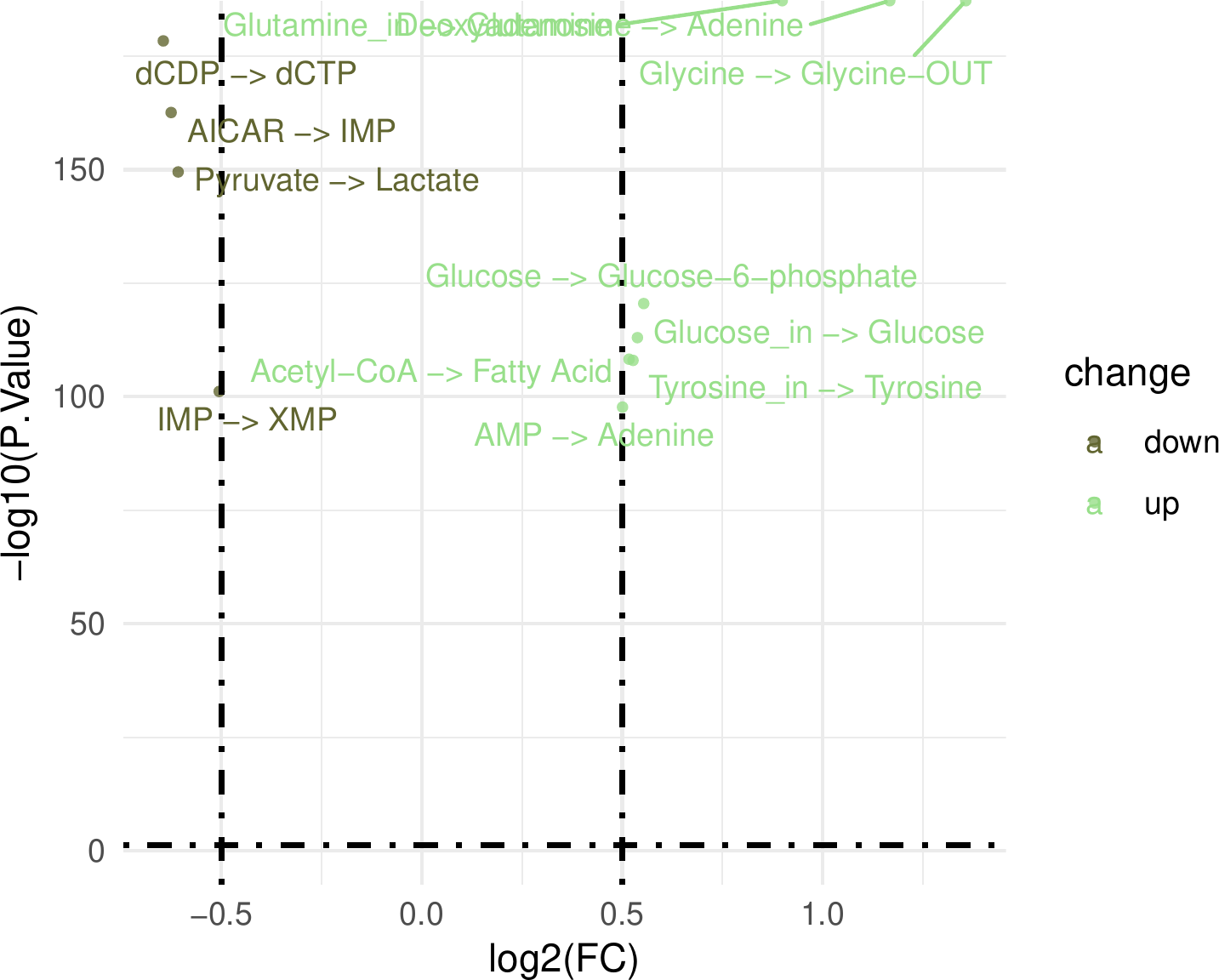
以 limma 的线形分析策略，对细胞的代谢通量差异分析。以 公式 ~ 0 + group 创建设计矩阵 (design matrix) 。差异分析：Erythroblast\_AA vs Erythroblast\_Normal, T\_Cell\_AA vs T\_Cell\_Normal, Monocyte\_AA vs Monocyte\_Normal, Natural\_Killer\_Cell\_AA vs Natural\_Killer\_Cell\_Normal, Plasma\_Cell\_AA vs Plasma\_Cell\_Normal, B\_Cell\_AA vs B\_Cell\_Normal, Neutrophil\_AA vs Neutrophil\_Normal, Unknown\_AA vs Unknown\_Normal, Macrophage\_AA vs Macrophage\_Normal, Pre\_B\_Cell\_AA vs Pre\_B\_Cell\_Normal, Platelet\_AA vs Platelet\_Normal, Proliferating\_Cell\_AA vs Proliferating\_Cell\_Normal, Mesenchymal\_Stromal\_Cell\_AA vs Mesenchymal\_Stromal\_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：

* Erythroblast\_AA vs Erythroblast\_Normal：up (n=8) , down (n=4) 。
* T\_Cell\_AA vs T\_Cell\_Normal：up (n=1) , down (n=0) 。
* Monocyte\_AA vs Monocyte\_Normal：up (n=1) , down (n=0) 。
* Natural\_Killer\_Cell\_AA vs Natural\_Killer\_Cell\_Normal：up (n=0) , down (n=0) 。
* Plasma\_Cell\_AA vs Plasma\_Cell\_Normal：up (n=16) , down (n=4) 。
* B\_Cell\_AA vs B\_Cell\_Normal：up (n=0) , down (n=0) 。
* Neutrophil\_AA vs Neutrophil\_Normal：up (n=1) , down (n=0) 。
* Unknown\_AA vs Unknown\_Normal：up (n=0) , down (n=0) 。
* Macrophage\_AA vs Macrophage\_Normal：up (n=3) , down (n=21) 。
* Pre\_B\_Cell\_AA vs Pre\_B\_Cell\_Normal：up (n=5) , down (n=0) 。
* Platelet\_AA vs Platelet\_Normal：up (n=31) , down (n=5) 。
* Proliferating\_Cell\_AA vs Proliferating\_Cell\_Normal：up (n=65) , down (n=0) 。
* Mesenchymal\_Stromal\_Cell\_AA vs Mesenchymal\_Stromal\_Cell\_Normal：up (n=43) , down (n=0) 。 所有上调 DMFs 共 110 个，所有下调 DMFs 共 28 个。所有非重复 DMFs 共 113 个。

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

Note: The directory 'Figure+Table/AA-FLUX-Differential-Statistic-data' contains 13 files.  
  
1 1\_Erythroblast\_AA - Erythroblast\_Normal.csv  
2 10\_Pre\_B\_Cell\_AA - Pre\_B\_Cell\_Normal.csv  
3 11\_Platelet\_AA - Platelet\_Normal.csv  
4 12\_Proliferating\_Cell\_AA - Proliferating\_Cell\_Normal.csv  
5 13\_Mesenchymal\_Stromal\_Cell\_AA - Mesenchymal\_Stromal\_Cell\_Normal.csv  
6 ...

**(File path: Figure+Table/3.8\_Limma\_代谢通量差异分析\_(AA\_FLUX)/AA-FLUX-Differential-Statistic-data)**



**Fig.** **12** AA FLUX Erythroblast AA vs Erythroblast Normal

**(File path: Figure+Table/3.8\_Limma\_代谢通量差异分析\_(AA\_FLUX)/AA-FLUX-Erythroblast-AA-vs-Erythroblast-Normal.pdf)**

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

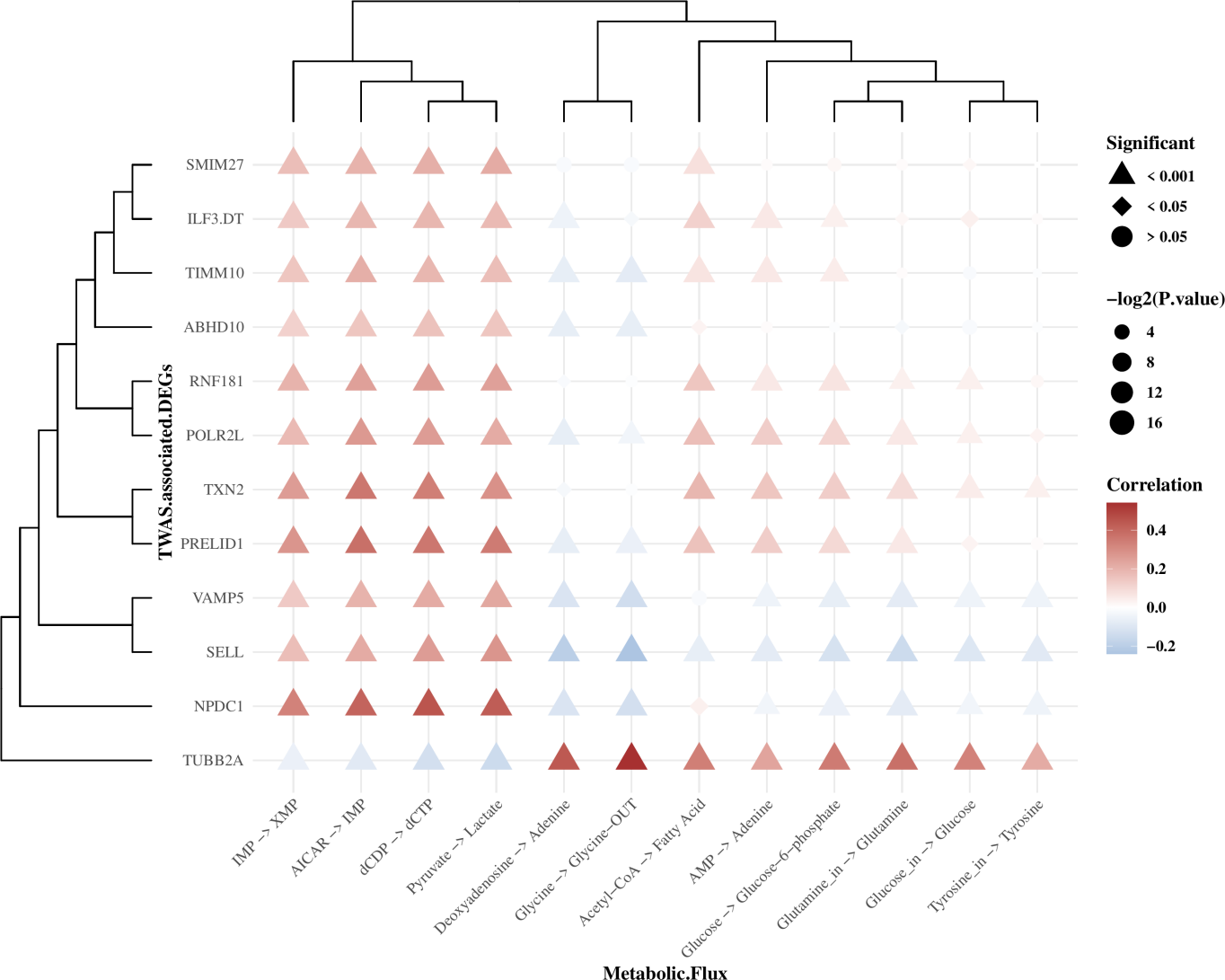
**(See: Figure+Table/3.8\_Limma\_代谢通量差异分析\_(AA\_FLUX)/AA-FLUX-Erythroblast-AA-vs-Erythroblast-Normal-content)**

## 3.9 关联分析

聚焦于 Erythroblast 细胞 (在 TWAS 与 代谢通量分析中，都筛选出了差异变化) ， 将 TWAS 风险基因 (见Tab. **[5](#AA-Filtered-TWAS-associated-genes-of-Cell-Cluster-DEGs)** 中的 Erythroblast) 与差异代谢通量关联分析 (见 Fig. **[12](#AA-FLUX-Erythroblast-AA-vs-Erythroblast-Normal)** )。

Fig. **[13](#Correlation-heatmap)** 为关联分析 (TWAS associated DEGs, Metabolic Flux) 热图。 Tab. **[7](#Significant-correlation)** 为关联分析统计附表 (P-value cutoff: 0.05)。 Fig. **[14](#AA-dimension-plot-of-expression-level-of-the-genes)** 基因 “ABHD10”, “ILF3-DT”, “NPDC1”, …(n = 12) 表达水平的 Dimension reduction plot.

Fig. **[14](#AA-dimension-plot-of-expression-level-of-the-genes)** 中，‘TUBB2A’ 具有差异分布趋势。



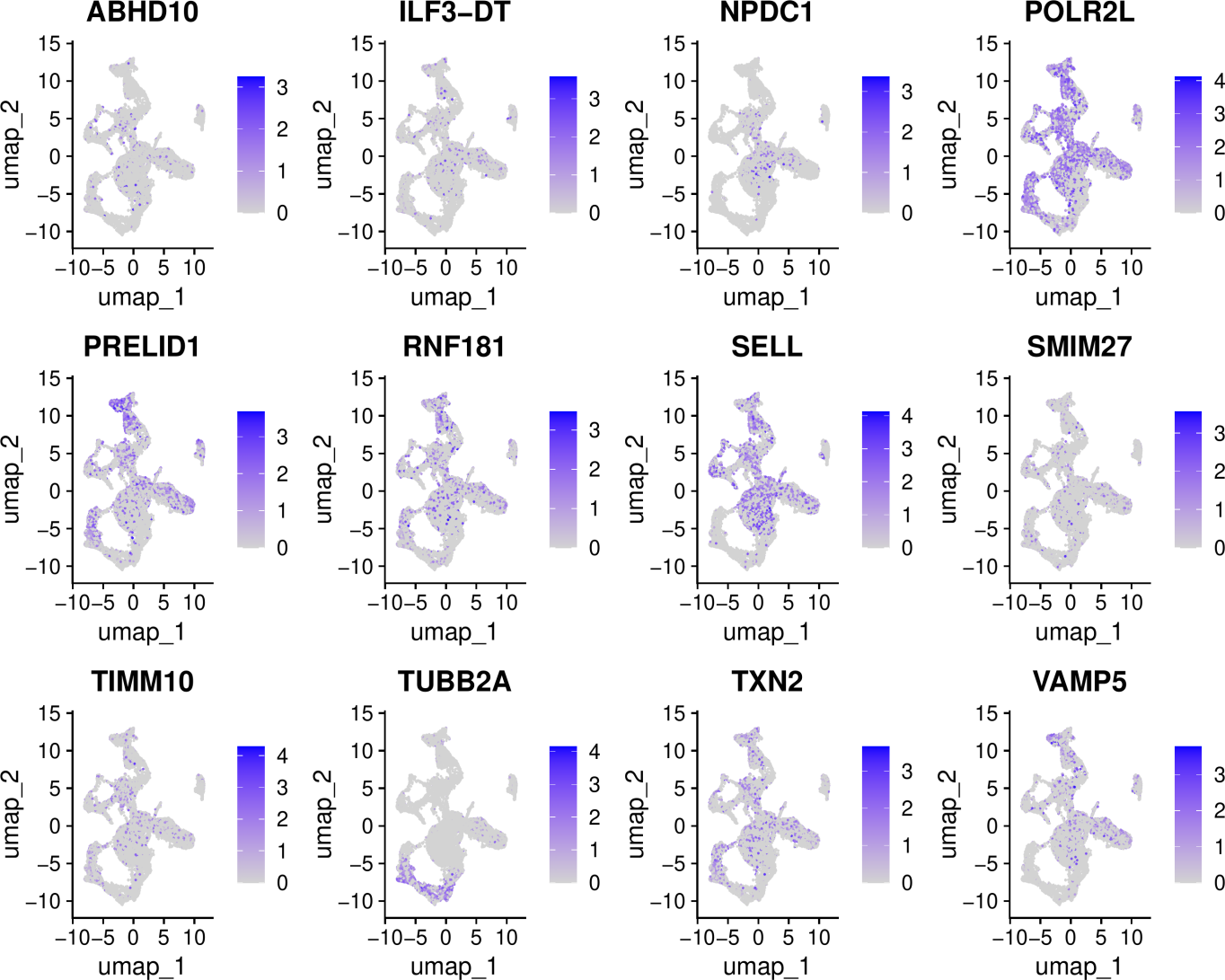
**Fig.** **13** Correlation heatmap

**(File path: Figure+Table/3.9\_关联分析/Correlation-heatmap.pdf)**

**Tab.** **7** Significant correlation

| TWAS.assoc... | Metabolic.... | Cor | Pvalue | -log2(P.va... | Significant | Sign |
| --- | --- | --- | --- | --- | --- | --- |
| SELL | Pyruvate -... | 0.28 | 0 | 16.61 | < 0.001 | \*\* |
| VAMP5 | Pyruvate -... | 0.23 | 0 | 16.61 | < 0.001 | \*\* |
| RNF181 | Pyruvate -... | 0.25 | 0 | 16.61 | < 0.001 | \*\* |
| ABHD10 | Pyruvate -... | 0.15 | 0 | 16.61 | < 0.001 | \*\* |
| PRELID1 | Pyruvate -... | 0.35 | 0 | 16.61 | < 0.001 | \*\* |
| ... | ... | ... | ... | ... | ... | ... |

**(File path: Figure+Table/3.9\_关联分析/Significant-correlation.csv)**



**Fig.** **14** AA dimension plot of expression level of the genes

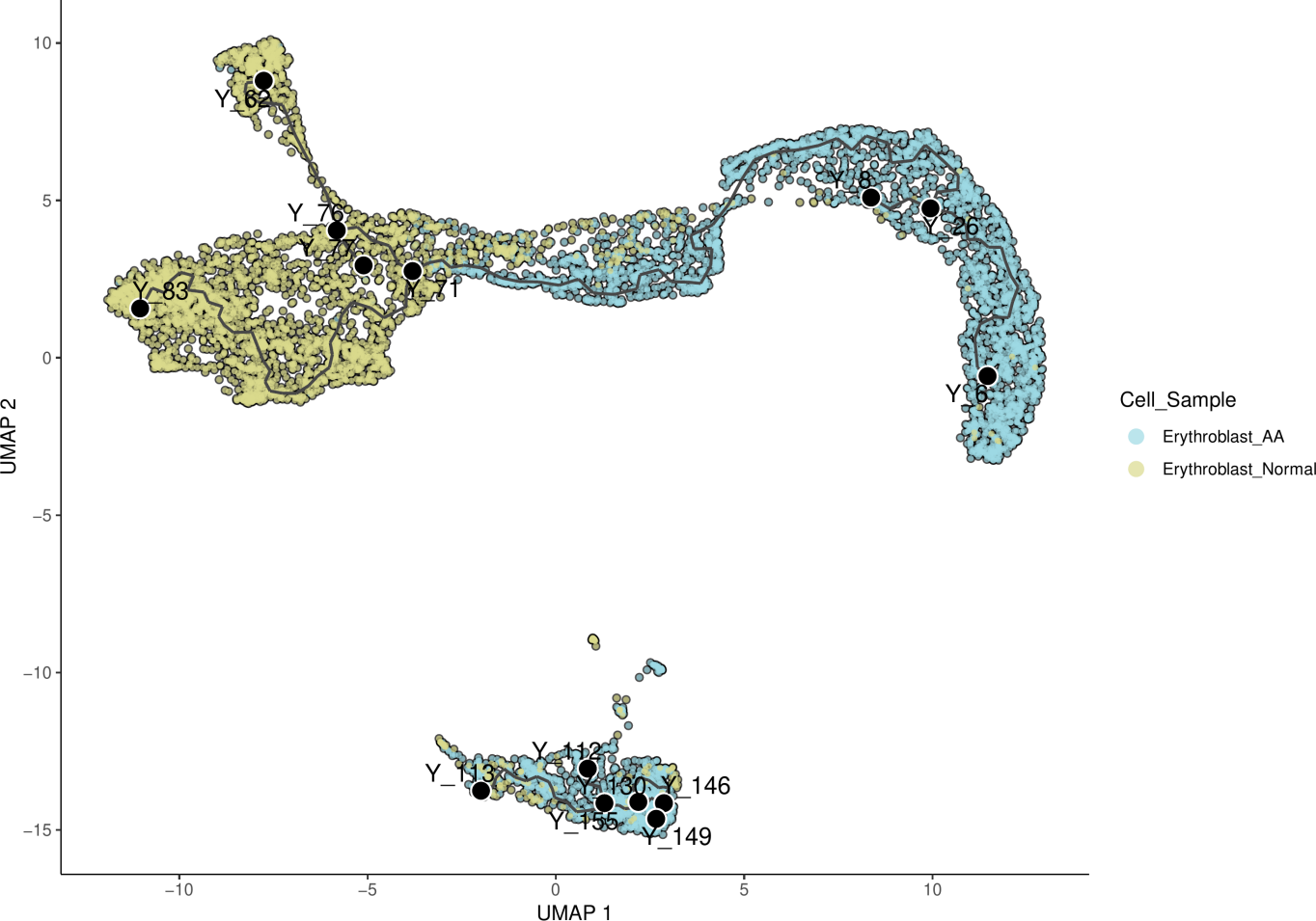
**(File path: Figure+Table/3.9\_关联分析/AA-dimension-plot-of-expression-level-of-the-genes.pdf)**

## 3.10 Monocle3 拟时分析 (AA\_ERY)

从 Seurat 数据对象 Cell\_Sample 中提取 Erythroblast 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\_83, Y\_113 (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。

Fig. **[15](#AA-ERY-principal-points)** 为拟时轨迹与 principal point 示意。 Fig. **[16](#AA-ERY-pseudotime)** 为细胞拟时图。 Tab. **[8](#AA-ERY-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)。 Tab. **[9](#AA-Graph-Test-of-associated-genes)** 为 TWAS 风险相关、细胞群差异表达 (Erythroblast, AA vs Normal) 的 Graph Test 附表。 Fig. **[17](#AA-ERY-Set1-genes-in-pseudotime)** 为 TUBB2A 在拟时 (Pseudotime) 中的表达变化。

Fig. **[17](#AA-ERY-Set1-genes-in-pseudotime)** 表明，TUBB2A 是一种在 AA 中高表达的基因 (Erythroblast)。



**Fig.** **15** AA ERY principal points

**(File path: Figure+Table/3.10\_Monocle3\_拟时分析\_(AA\_ERY)/AA-ERY-principal-points.pdf)**



**Fig.** **16** AA ERY pseudotime

**(File path: Figure+Table/3.10\_Monocle3\_拟时分析\_(AA\_ERY)/AA-ERY-pseudotime.pdf)**

**Tab.** **8** AA ERY Graph Test Significant genes

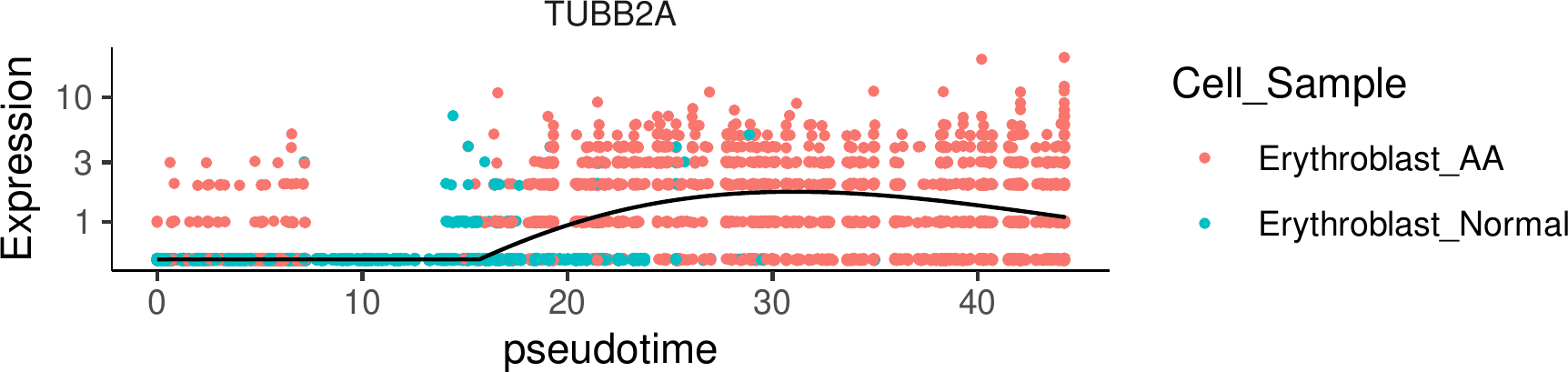
| Gene id | Status | P value | Morans tes... | Morans I | Q value |
| --- | --- | --- | --- | --- | --- |
| HBA1 | OK | 0 | 288.7 | 0.9408 | 0 |
| ALAS2 | OK | 0 | 287.5 | 0.9368 | 0 |
| HBA2 | OK | 0 | 286.8 | 0.9345 | 0 |
| AHSP | OK | 0 | 283.9 | 0.9251 | 0 |
| HBB | OK | 0 | 283.3 | 0.9232 | 0 |
| ... | ... | ... | ... | ... | ... |

**(File path: Figure+Table/3.10\_Monocle3\_拟时分析\_(AA\_ERY)/AA-ERY-Graph-Test-Significant-genes.csv)**

**Tab.** **9** AA Graph Test of associated genes

| Gene id | Status | P value | Morans tes... | Morans I | Q value |
| --- | --- | --- | --- | --- | --- |
| TUBB2A | OK | 0 | 179.9 | 0.5862 | 0 |
| NPDC1 | OK | 0 | 121.9 | 0.397 | 0 |
| SELL | OK | 0 | 113.5 | 0.3697 | 0 |
| POLR2L | OK | 0 | 74.49 | 0.2426 | 0 |
| PRELID1 | OK | 0 | 68.26 | 0.2223 | 0 |
| ... | ... | ... | ... | ... | ... |

**(File path: Figure+Table/3.10\_Monocle3\_拟时分析\_(AA\_ERY)/AA-Graph-Test-of-associated-genes.csv)**



**Fig.** **17** AA ERY Set1 genes in pseudotime

**(File path: Figure+Table/3.10\_Monocle3\_拟时分析\_(AA\_ERY)/AA-ERY-Set1-genes-in-pseudotime.pdf)**

# 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。 在 Erythroblast 细胞中发现了 “Glycine -> Glycine-OUT”, “Deoxyadenosine -> Adenine”, “Glutamine\_in -> Glutamine”, …(n = 12) 等多数代谢通量变化。 在 AA vs Normal 的差异分析中，Erythroblast 细胞中也发现了与 TWAS 风险一致的基因 (见 Tab. **[5](#AA-Filtered-TWAS-associated-genes-of-Cell-Cluster-DEGs)** )。 对差异代谢通量变化与这些风险相关基因的关联分析表明，‘TUBB2A’ 与所有的差异代谢通量改变有显著关联。

‘TUBB2A’ 是一个首要表达于 Erythroblast 细胞中的基因，此外，AA 相较于 Normal 高表达。 (见 Fig. **[14](#AA-dimension-plot-of-expression-level-of-the-genes)** ) 对 Erythroblast 细胞的拟时分析，明确了 ‘TUBB2A’ 在 AA 中的高表达变化。

局限：目前公开可得的 AA 数据集较少，难以多方面验证，该分析可能受限于当前数据集。

# 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. 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).

6. 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).