**生信分析报告**

**项目标题： 再生障碍性贫血 ;**

**单 号： BSCL241113 ;**

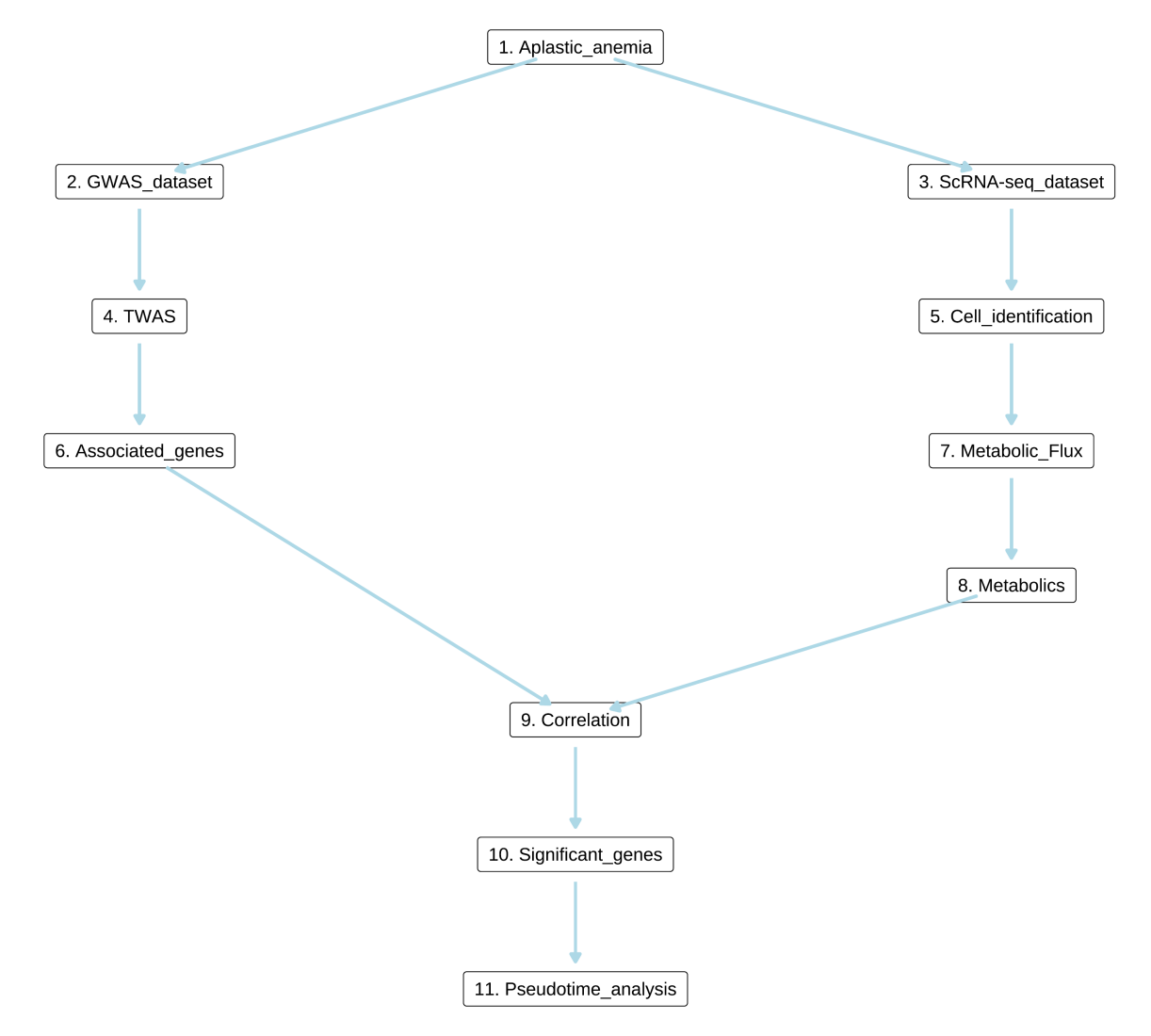
**分析人员： 黄礼闯 ;**

**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: figs/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。 详细见各分析部分。

# 3 分析结果

## 3.1 MungeSumstats 获取 GWAS 数据 (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>)。

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

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 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。

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: tabs/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: tabs/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)。

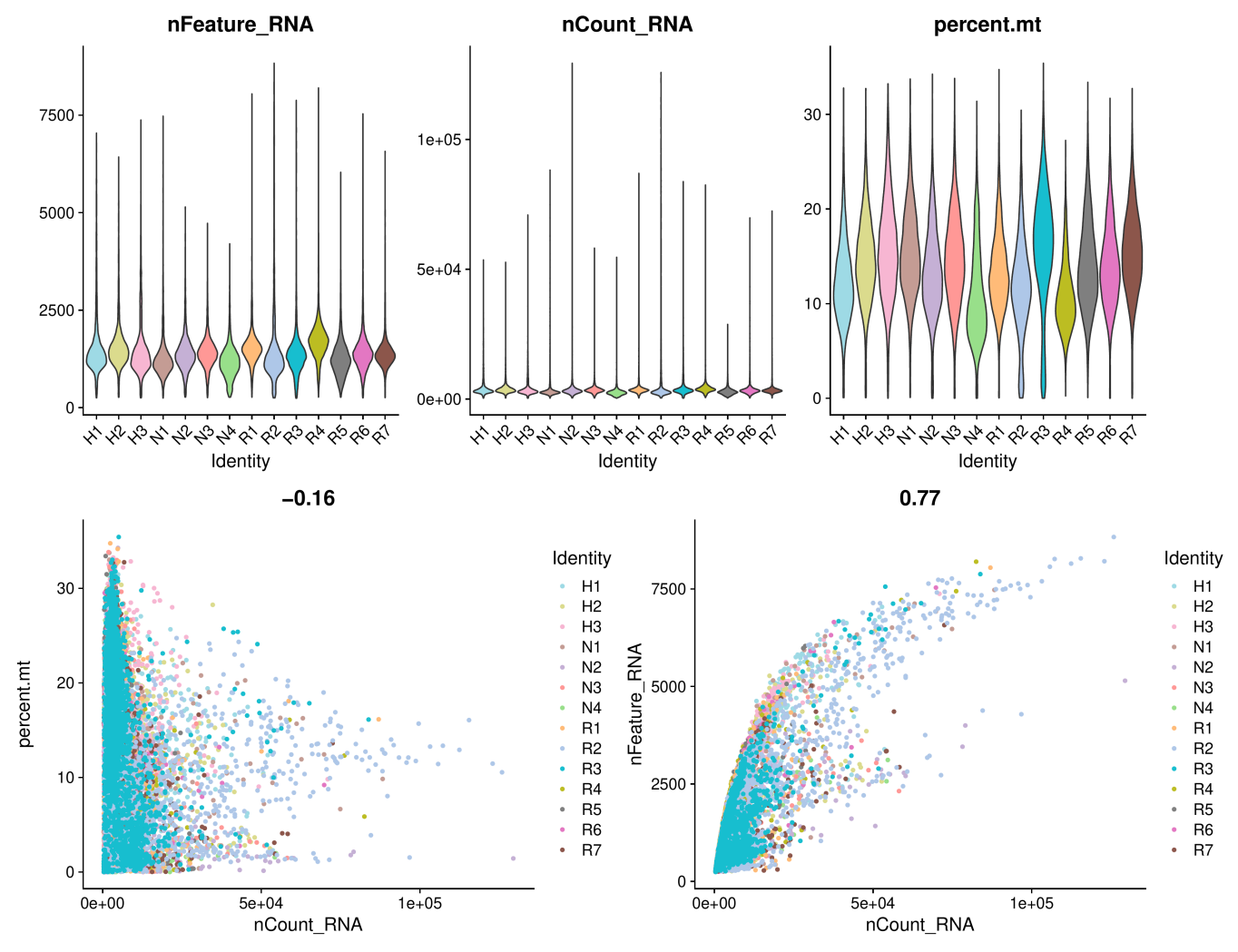
## 3.4 CompBioLab 数据获取 (<PMID:37908861>)

下载 AA 单细胞数据集 (<https://compbio.unist.ac.kr/UNIST-Aplastic_anemia-KRICT-2019-08/download.php?download=Merged_count_matrix.txt.gz>)

## 3.5 Seurat 单细胞数据分析 (AA)

使用 Seurat R 包 (5.2.1) 进行单细胞数据质量控制 (QC) 和下游分析。依据 <Tutorial: <https://satijalab.org/seurat/articles/pbmc3k_tutorial.html>> 为指导对单细胞数据预处理。一个细胞至少应有 1000 个基因，并且基因数量小于 7500。线粒体基因的比例小于 25%。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。

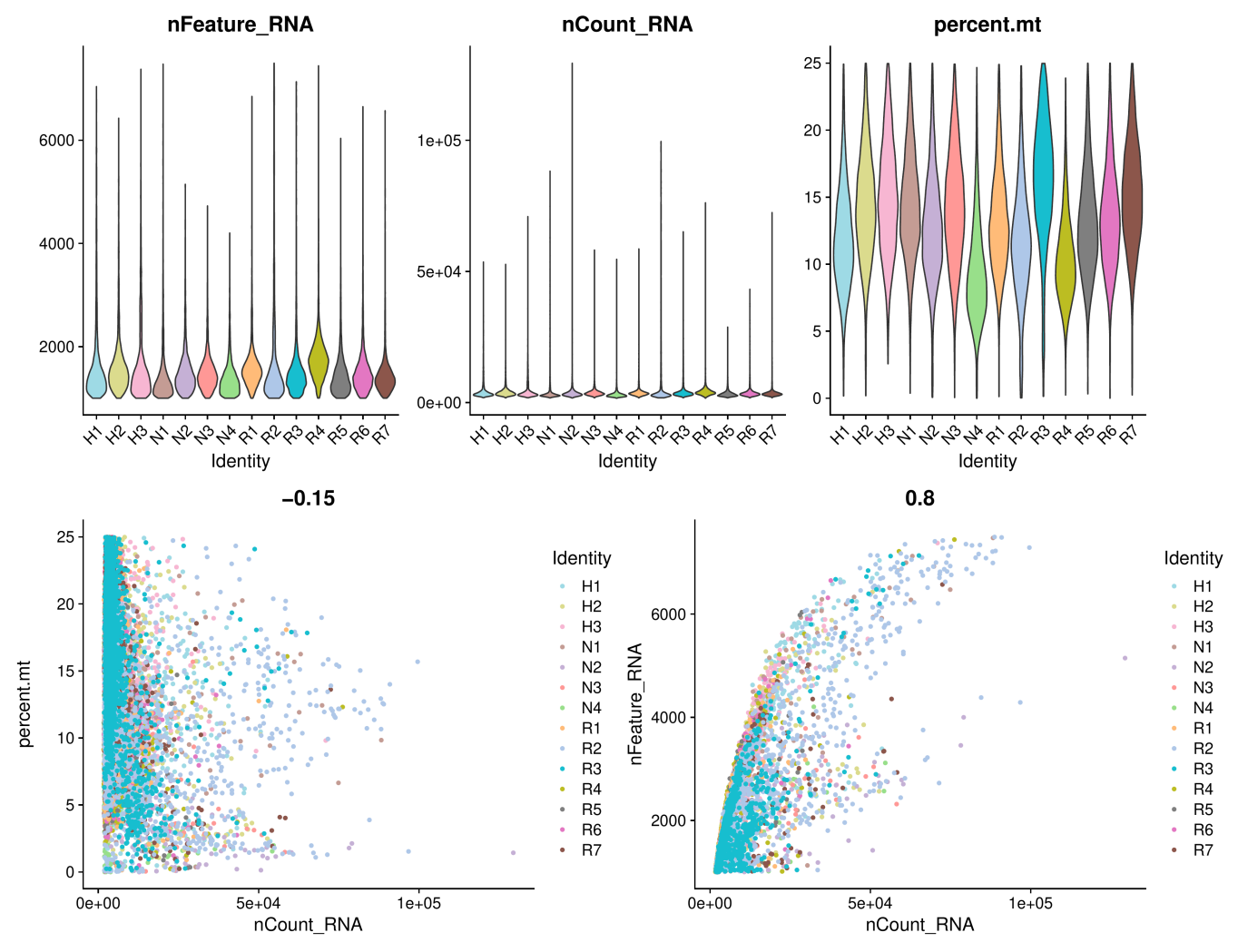
前期质量控制，一个细胞至少应有 1000 个基因，并且基因数量小于 7500。线粒体基因的比例小于 25%。过滤后，数据集共包含 16019 个基因，55006 个细胞。数据归一化，PCA 聚类 (Seurat 标准工作流，见方法章节) 后，绘制 PC standard deviations 图。在 1-15 PC 维度，1.2 分辨率下，对细胞群 UMAP 聚类。计算所有细胞群的 Marker。使用特异性 Marker，以 SCSA 对细胞群注释。



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

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

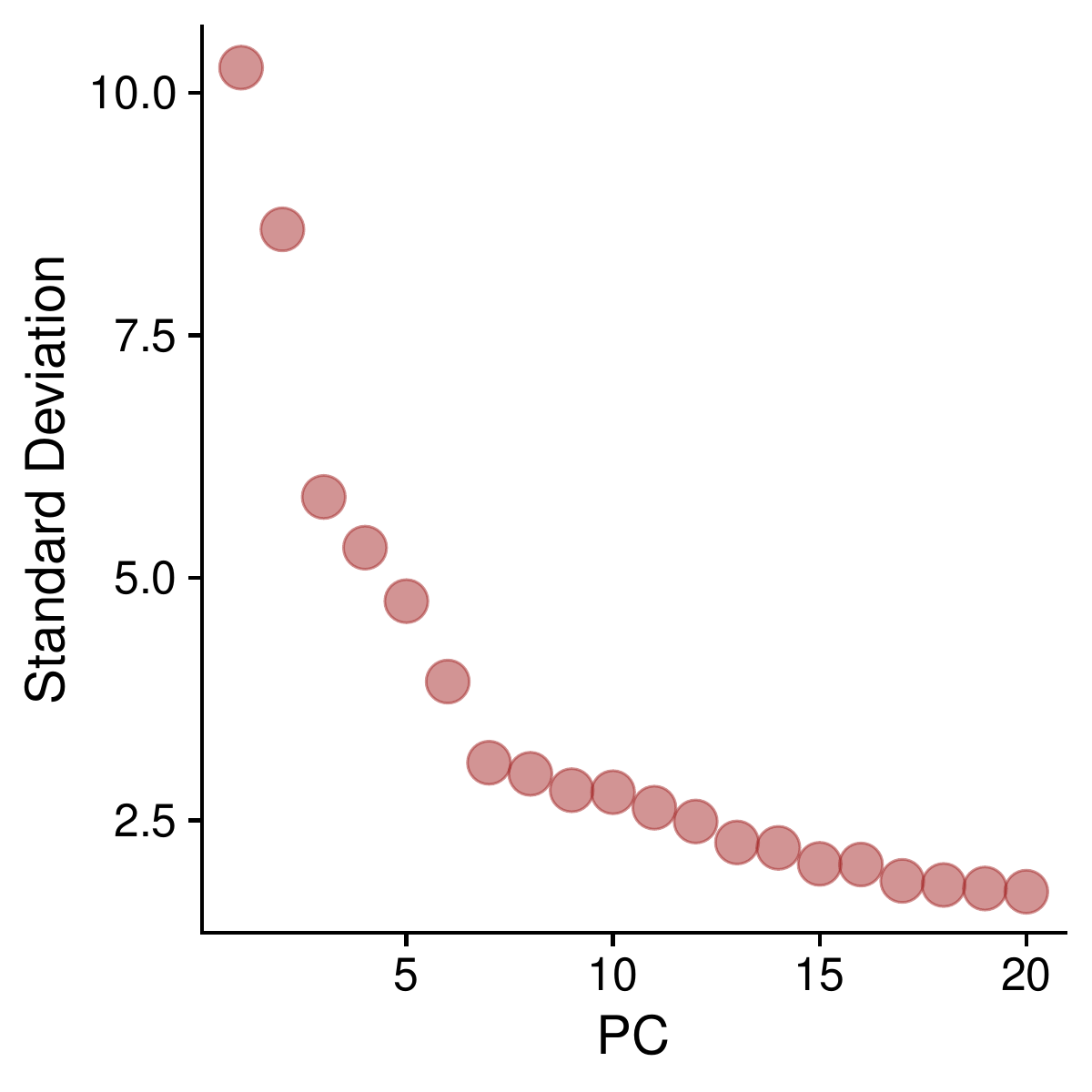
Fig. **[2](#Pre-Quality-control)** 为 QC (质量控制) 图 (数据过滤前) 。



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

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

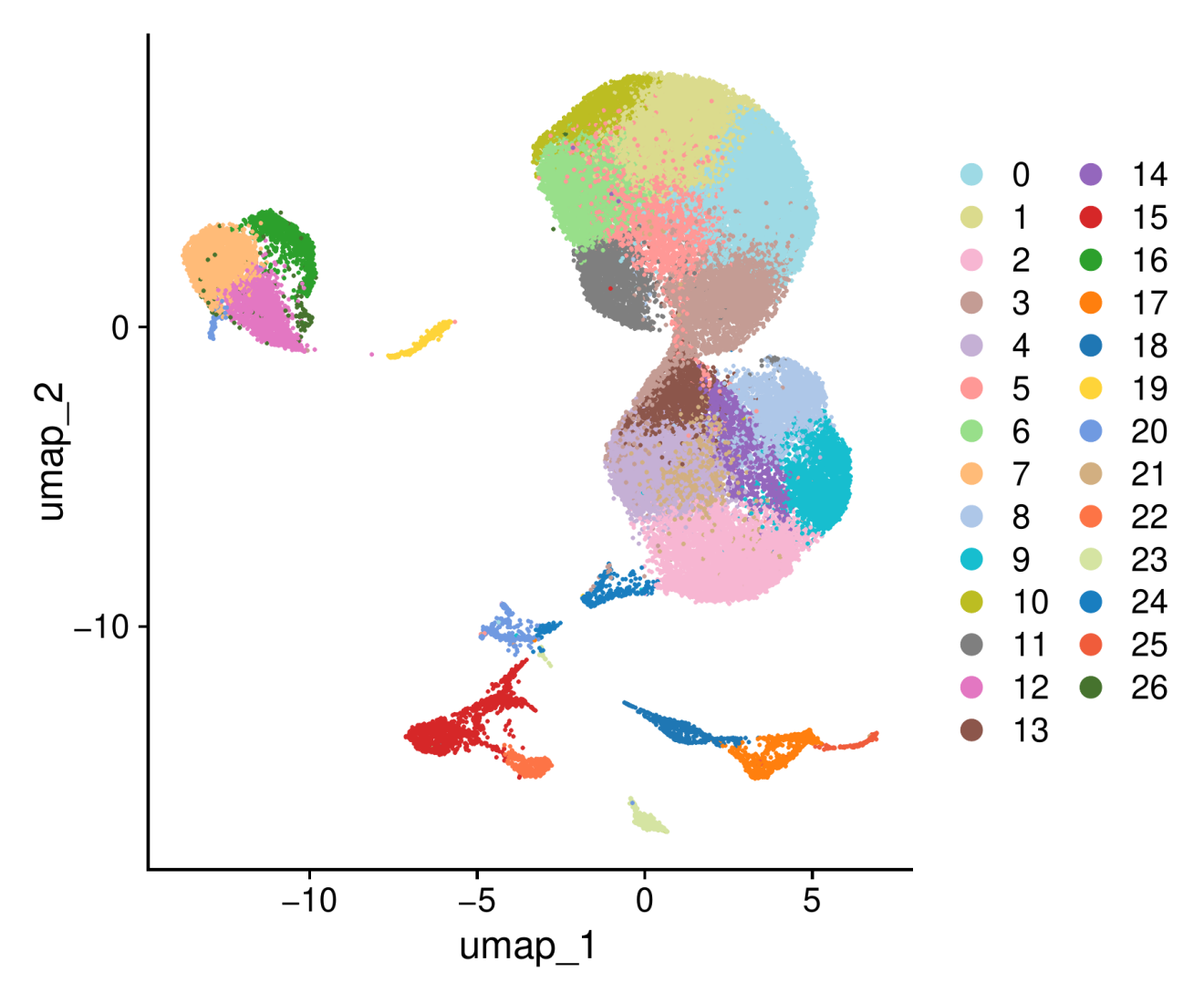
Fig. **[4](#AA-Standard-deviations-of-PCs)** 为主成分 (PC) 的 Standard deviations。

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

| Rownames | P val | Avg log2FC | Pct.1 | Pct.2 |
| --- | --- | --- | --- | --- |
| MAL | 0 | 1.563 | 0.677 | 0.245 |
| CCR7 | 0 | 1.26 | 0.749 | 0.324 |
| IL7R | 0 | 0.6684 | 0.922 | 0.523 |
| NOSIP | 0 | 1.316 | 0.842 | 0.449 |
| TRAC | 0 | 1.082 | 0.903 | 0.511 |
| ... | ... | ... | ... | ... |

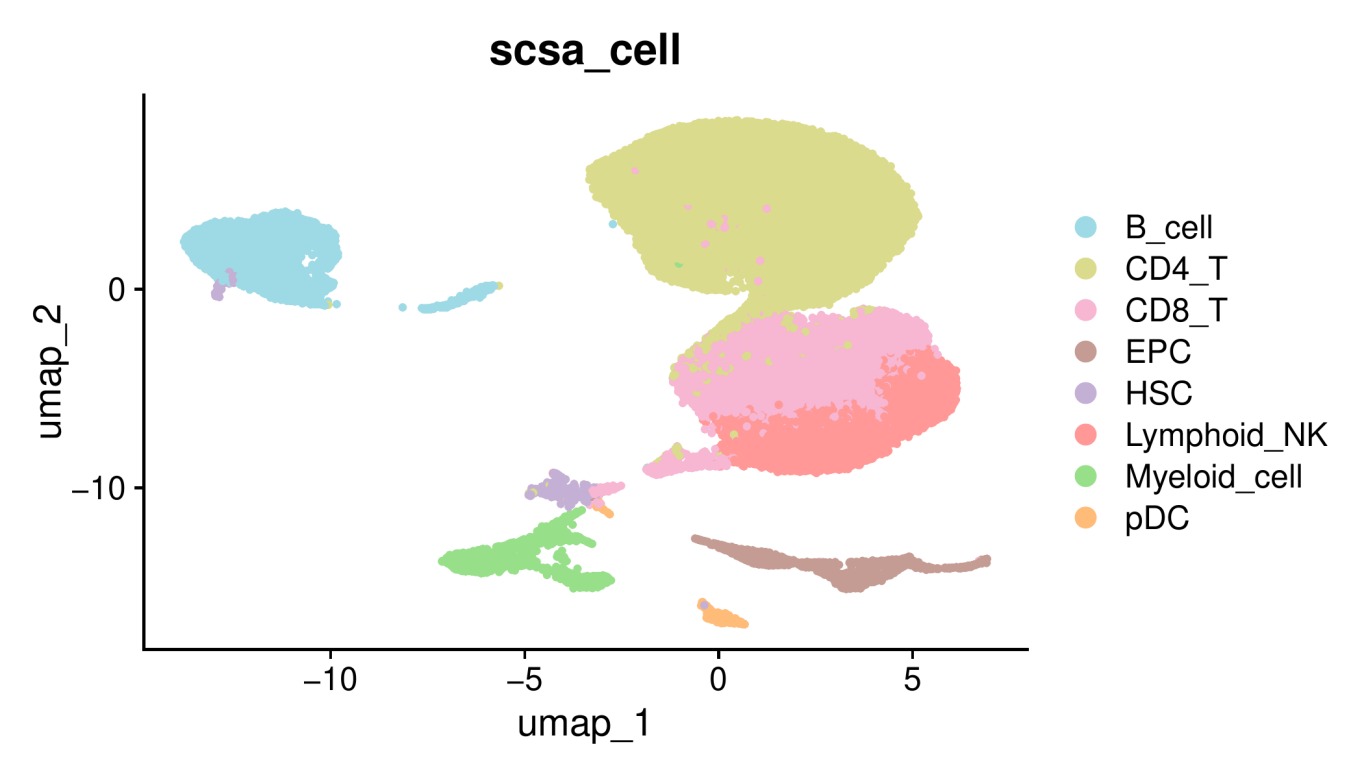
**(File path: tabs/3.5\_Seurat\_单细胞数据分析\_(AA)/AA-significant-markers-of-cell-clusters.csv)**

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.** **5** AA UMAP Clustering

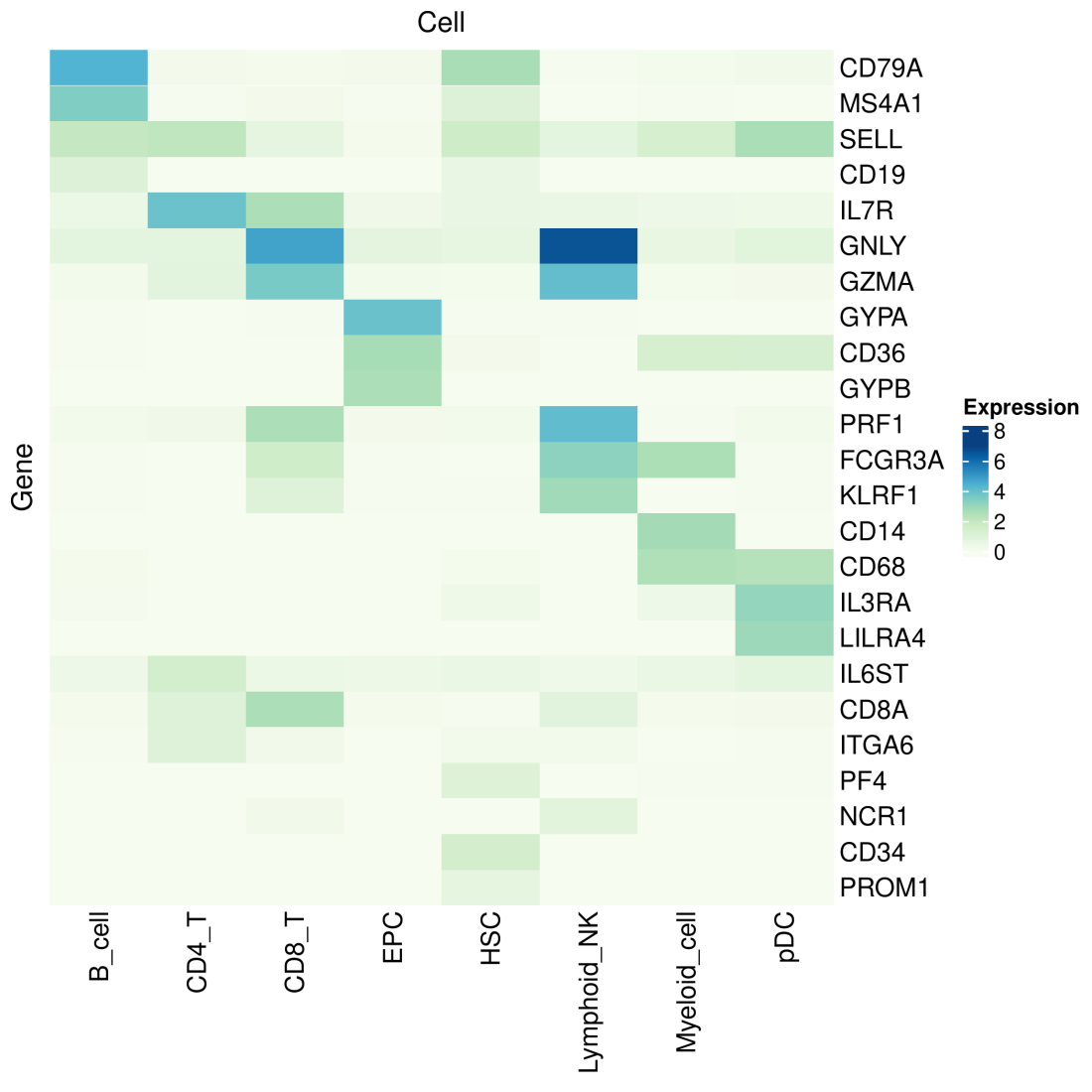
**(File path: figs/3.5\_Seurat\_单细胞数据分析\_(AA)/AA-UMAP-Clustering.pdf)**



**Fig.** **6** AA SCSA Cell type annotation

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

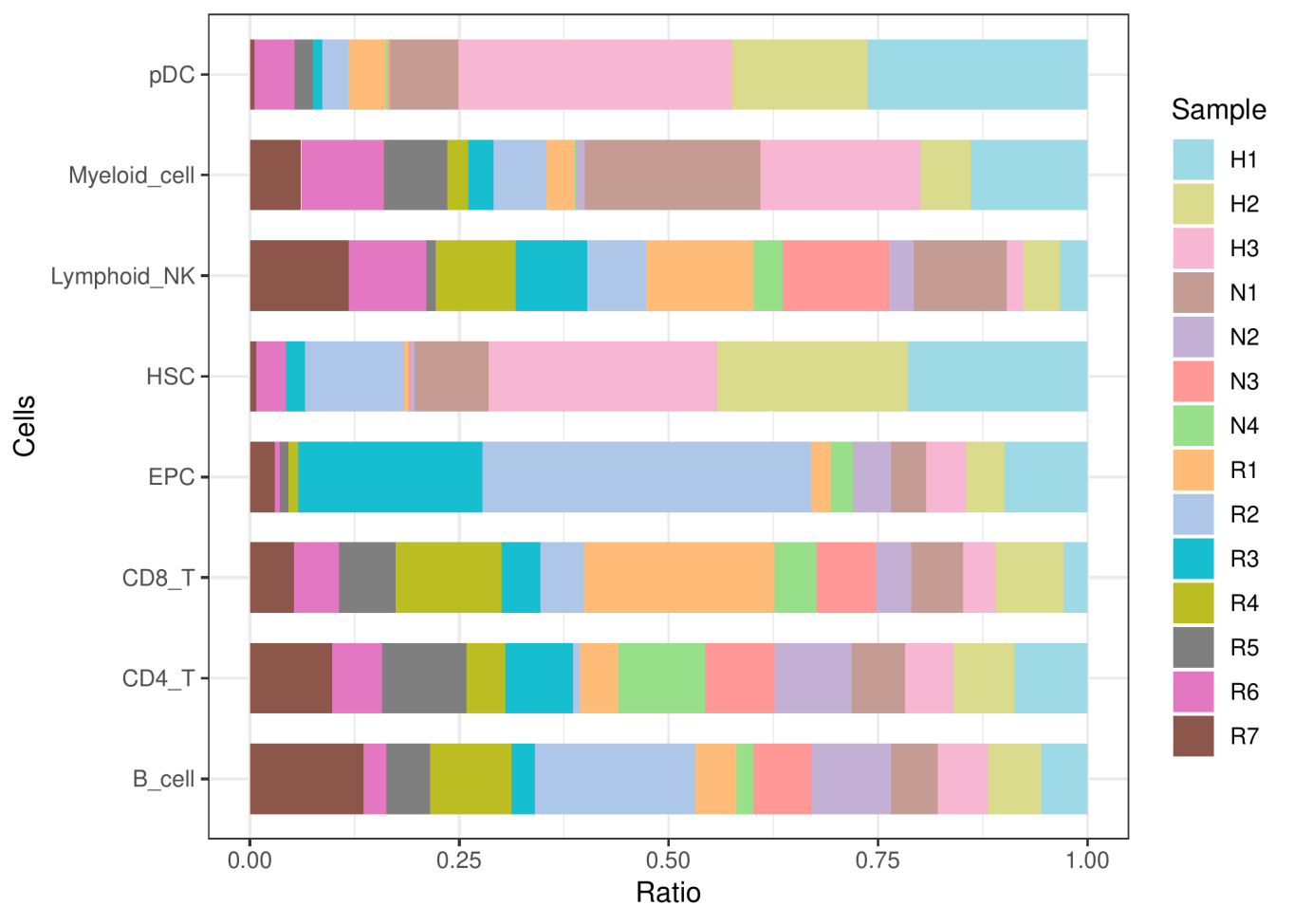
Fig. **[6](#AA-SCSA-Cell-type-annotation)** 为 SCSA 细胞注释结果的 UMAP 图。



**Fig.** **7** AA Marker Validation

**(File path: figs/3.5\_Seurat\_单细胞数据分析\_(AA)/AA-Marker-Validation.pdf)**

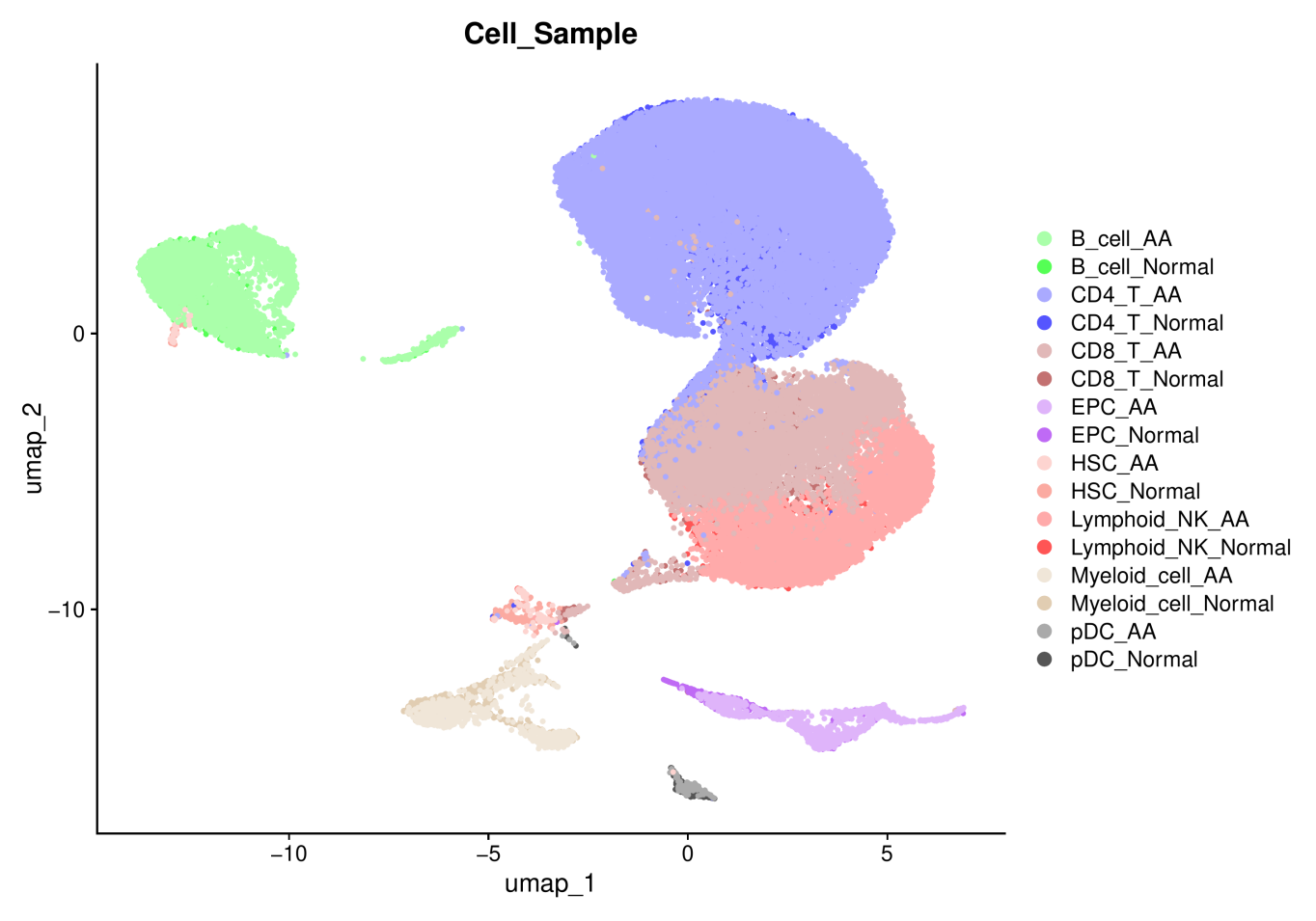
Fig. **[7](#AA-Marker-Validation)** 使用特异性 Marker 对细胞注释结果的验证热图。



**Fig.** **8** AA SCSA Cell Proportions in each sample

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

Fig. **[8](#AA-SCSA-Cell-Proportions-in-each-sample)** 为 SCSA 注释的细胞群在各个样本中的占比。



**Fig.** **9** AA The Cell Sample

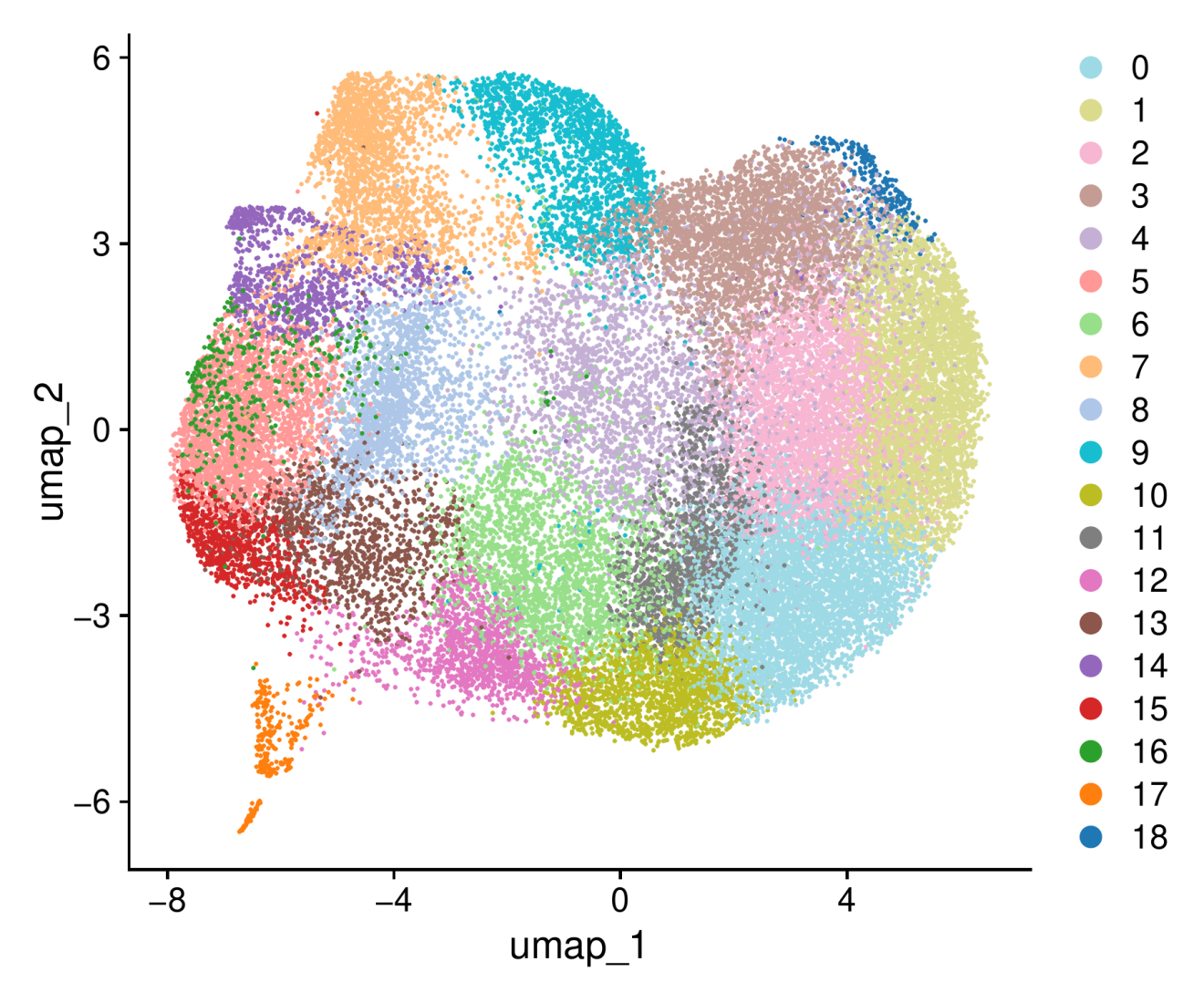
**(File path: figs/3.5\_Seurat\_单细胞数据分析\_(AA)/AA-The-Cell-Sample.pdf)**

Fig. **[9](#AA-The-Cell-Sample)** 为 Cell\_Sample 的 umap 聚类图。

## 3.6 Seurat 细胞亚群分析 (T\_CELL)

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>) 对细胞群注释。

匹配 scsa\_cell 中包含“CD8\_T|CD4\_T”的描述，最终得到 38211 例数据。分析其亚群。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。使用特异性 Marker，以 SCSA 对细胞群注释。



**Fig.** **10** T CELL UMAP Clustering

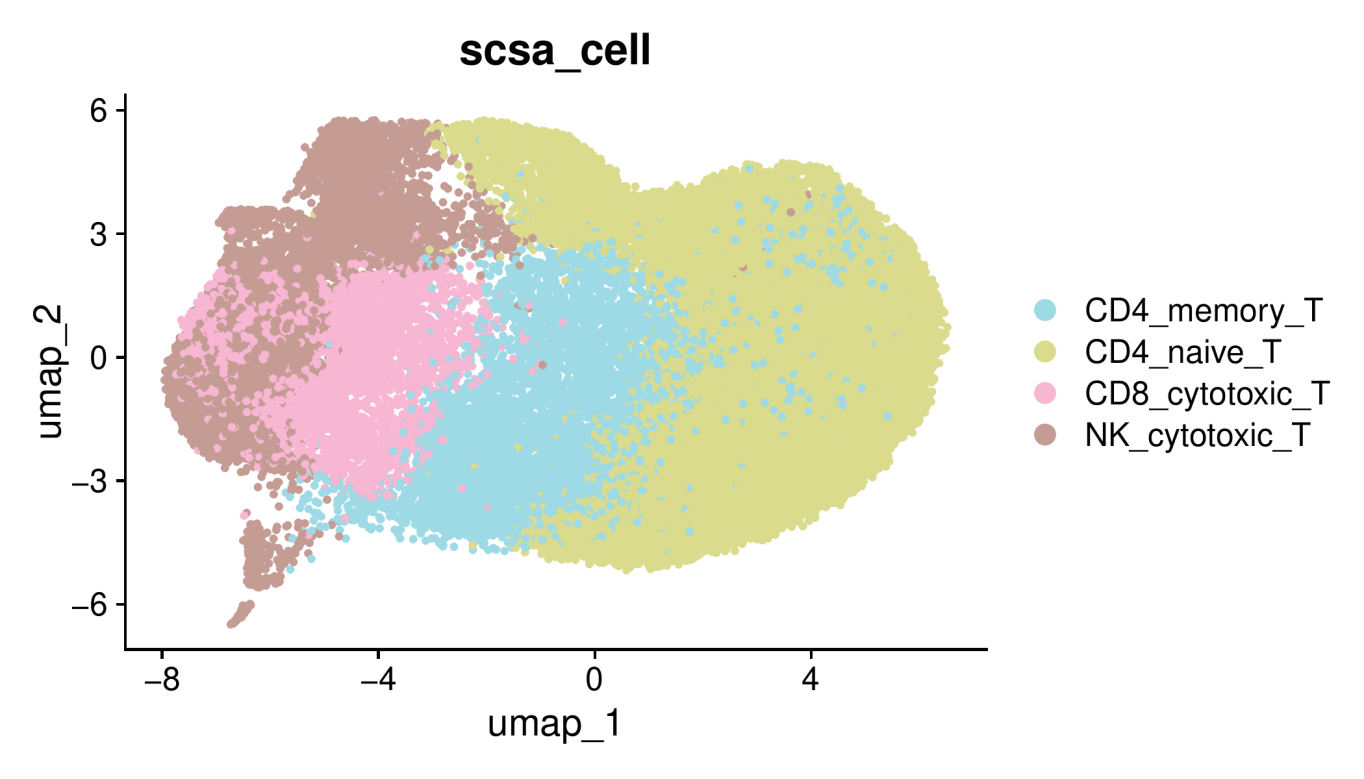
**(File path: figs/3.6\_Seurat\_细胞亚群分析\_(T\_CELL)/T-CELL-UMAP-Clustering.pdf)**

**Tab.** **4** T CELL significant markers of cell clusters

| Rownames | P val | Avg log2FC | Pct.1 | Pct.2 |
| --- | --- | --- | --- | --- |
| AIF1 | 0 | 1.446 | 0.547 | 0.226 |
| CCR7 | 0 | 0.9349 | 0.756 | 0.437 |
| NOSIP | 0 | 0.7736 | 0.837 | 0.583 |
| RGS10 | 0 | 0.7981 | 0.801 | 0.551 |
| LDHB | 0 | 0.7269 | 0.974 | 0.808 |
| ... | ... | ... | ... | ... |

**(File path: tabs/3.6\_Seurat\_细胞亚群分析\_(T\_CELL)/T-CELL-significant-markers-of-cell-clusters.csv)**

Tab. **[4](#T-CELL-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.** **11** T CELL SCSA Cell type annotation

**(File path: figs/3.6\_Seurat\_细胞亚群分析\_(T\_CELL)/T-CELL-SCSA-Cell-type-annotation.pdf)**

Fig. **[11](#T-CELL-SCSA-Cell-type-annotation)** 为 SCSA 细胞注释结果的 UMAP 图。



**Fig.** **12** T CELL Marker Validation

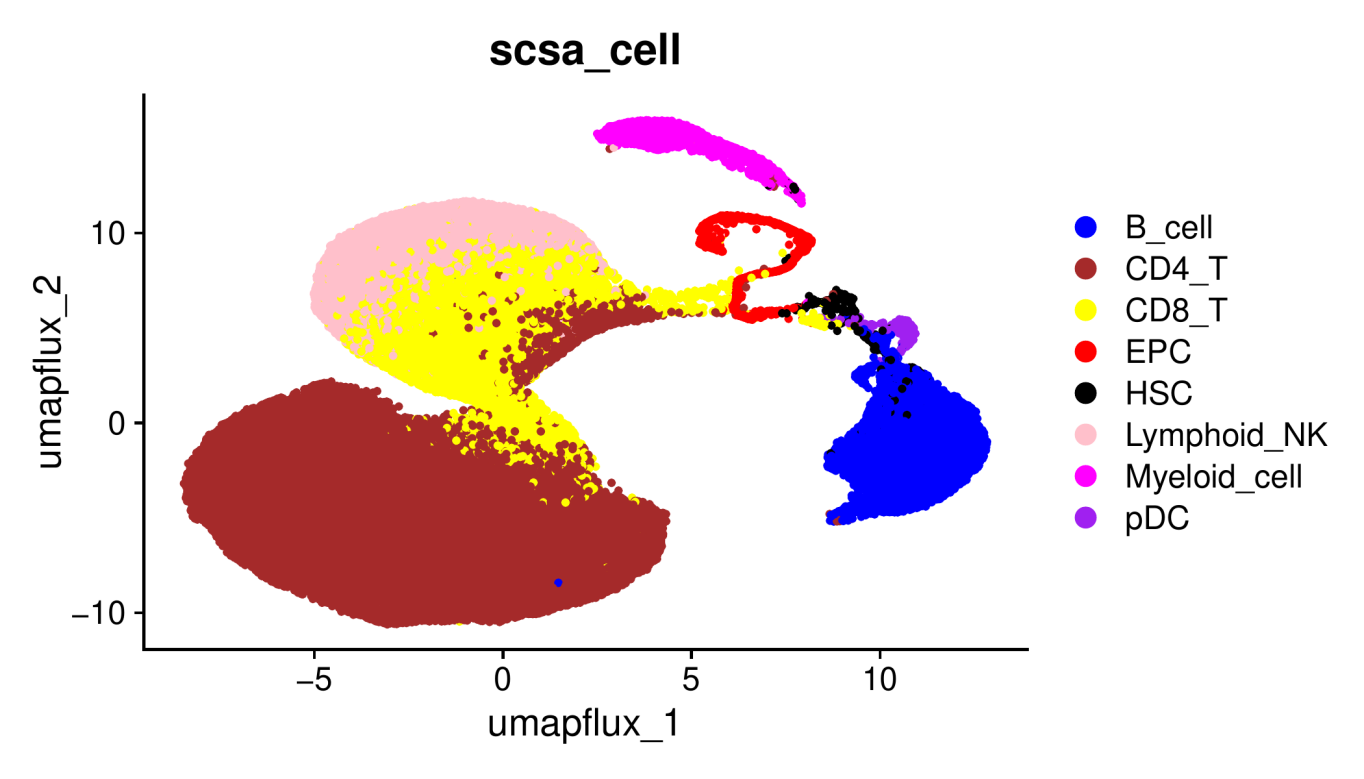
**(File path: figs/3.6\_Seurat\_细胞亚群分析\_(T\_CELL)/T-CELL-Marker-Validation.pdf)**

Fig. **[12](#T-CELL-Marker-Validation)** 使用特异性 Marker 对细胞注释结果的验证热图。

## 3.7 scFEA 单细胞数据的代谢通量预测 (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)6。参考 <https://github.com/changwn/scFEA/blob/master/scFEA_tutorial1.ipynb> 和 <https://github.com/changwn/scFEA/blob/master/scFEA_tutorial2.ipynb>。

将 Seurat (所有细胞) 以 scFEA 预测代谢通量。



**Fig.** **13** AA cells metabolic flux

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

Fig. **[13](#AA-cells-metabolic-flux)** 为细胞代谢通量 (scFEA 预测，输入 Seurat) 的 UMAP 聚类。

**Tab.** **5** AA metabolic flux matrix

| V1 | M 1 | M 2 | M 3 | M 4 |
| --- | --- | --- | --- | --- |
| N4 AAACCCAAGGTCCG... | 0.02322 | 0.01237 | -1.005e-05 | 0.0178 |
| N4 AAACCCAAGTCTTC... | 0.02155 | 0.009927 | 3.288e-07 | 0.01203 |
| N4 AAACCCACACCCAA... | 0.02172 | 0.02053 | 0.008557 | 0.02261 |
| N4 AAACGAACACAGTC... | 0.02155 | 0.009927 | -8.41e-07 | 0.01013 |
| N4 AAACGAACATGAGA... | 0.02155 | 0.009927 | 3.288e-07 | 0.01748 |
| ... | ... | ... | ... | ... |

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

Tab. **[5](#AA-metabolic-flux-matrix)** 为细胞代谢通量矩阵 (各 M\_ 为代谢模块)。

**Tab.** **6** 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: tabs/3.7\_scFEA\_单细胞数据的代谢通量预测\_(AA)/AA-annotation-of-metabolic-flux.xlsx)**

Tab. **[6](#AA-annotation-of-metabolic-flux)** 各代谢模块的注释。

## 3.8 Limma 代谢通量差异分析 (AA\_FLUX)

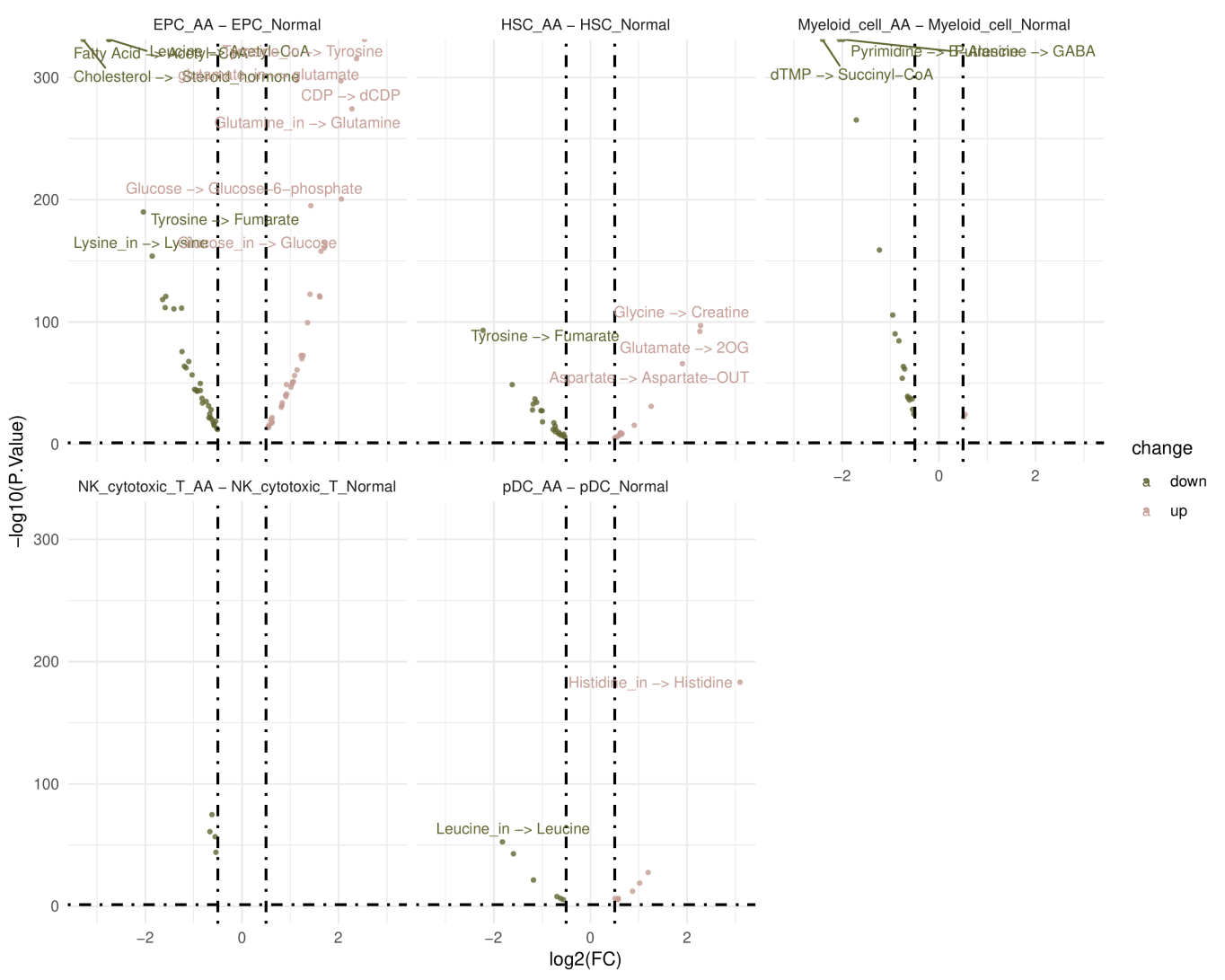
以 limma (3.62.2) (2005)7 差异分析。分析方法参考 <https://bioconductor.org/packages/release/workflows/vignettes/RNAseq123/inst/doc/limmaWorkflow.html>。创建设计矩阵，对比矩阵，差异分析：CD4\_naive\_T\_AA vs CD4\_naive\_T\_Normal, CD8\_cytotoxic\_T\_AA vs CD8\_cytotoxic\_T\_Normal, NK\_cytotoxic\_T\_AA vs NK\_cytotoxic\_T\_Normal, B\_cell\_AA vs B\_cell\_Normal, CD4\_memory\_T\_AA vs CD4\_memory\_T\_Normal, Lymphoid\_NK\_AA vs Lymphoid\_NK\_Normal, pDC\_AA vs pDC\_Normal, EPC\_AA vs EPC\_Normal, Myeloid\_cell\_AA vs Myeloid\_cell\_Normal, HSC\_AA vs HSC\_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。

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：CD4\_naive\_T\_AA vs CD4\_naive\_T\_Normal, CD8\_cytotoxic\_T\_AA vs CD8\_cytotoxic\_T\_Normal, NK\_cytotoxic\_T\_AA vs NK\_cytotoxic\_T\_Normal, B\_cell\_AA vs B\_cell\_Normal, CD4\_memory\_T\_AA vs CD4\_memory\_T\_Normal, Lymphoid\_NK\_AA vs Lymphoid\_NK\_Normal, pDC\_AA vs pDC\_Normal, EPC\_AA vs EPC\_Normal, Myeloid\_cell\_AA vs Myeloid\_cell\_Normal, HSC\_AA vs HSC\_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：

* CD4\_naive\_T\_AA - CD4\_naive\_T\_Normal：up (n=0) , down (n=0) 。
* CD8\_cytotoxic\_T\_AA - CD8\_cytotoxic\_T\_Normal：up (n=0) , down (n=0) 。
* NK\_cytotoxic\_T\_AA - NK\_cytotoxic\_T\_Normal：up (n=0) , down (n=4) 。
* B\_cell\_AA - B\_cell\_Normal：up (n=0) , down (n=0) 。
* CD4\_memory\_T\_AA - CD4\_memory\_T\_Normal：up (n=0) , down (n=0) 。
* Lymphoid\_NK\_AA - Lymphoid\_NK\_Normal：up (n=0) , down (n=0) 。
* pDC\_AA - pDC\_Normal：up (n=8) , down (n=6) 。
* EPC\_AA - EPC\_Normal：up (n=37) , down (n=36) 。
* Myeloid\_cell\_AA - Myeloid\_cell\_Normal：up (n=2) , down (n=18) 。
* HSC\_AA - HSC\_Normal：up (n=13) , down (n=20) 。 所有上调 DMFs 共 56 个，所有下调 DMFs 共 57 个。所有非重复 DMFs 共 104 个。

Note: The directory 'tabs/AA-FLUX-Differential-Statistic-data' contains 10 files.  
  
1 1\_CD4\_naive\_T\_AA - CD4\_naive\_T\_Normal.csv  
2 10\_HSC\_AA - HSC\_Normal.csv  
3 2\_CD8\_cytotoxic\_T\_AA - CD8\_cytotoxic\_T\_Normal.csv  
4 3\_NK\_cytotoxic\_T\_AA - NK\_cytotoxic\_T\_Normal.csv  
5 4\_B\_cell\_AA - B\_cell\_Normal.csv  
6 ...

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



**Fig.** **14** AA FLUX gathered volcano plot

**(File path: figs/3.8\_Limma\_代谢通量差异分析\_(AA\_FLUX)/AA-FLUX-gathered-volcano-plot.pdf)**

Fig. **[14](#AA-FLUX-gathered-volcano-plot)** 为合并的各组差异分析 DMFs 火山图。

## 3.9 细胞群 features 关联分析 (AA)

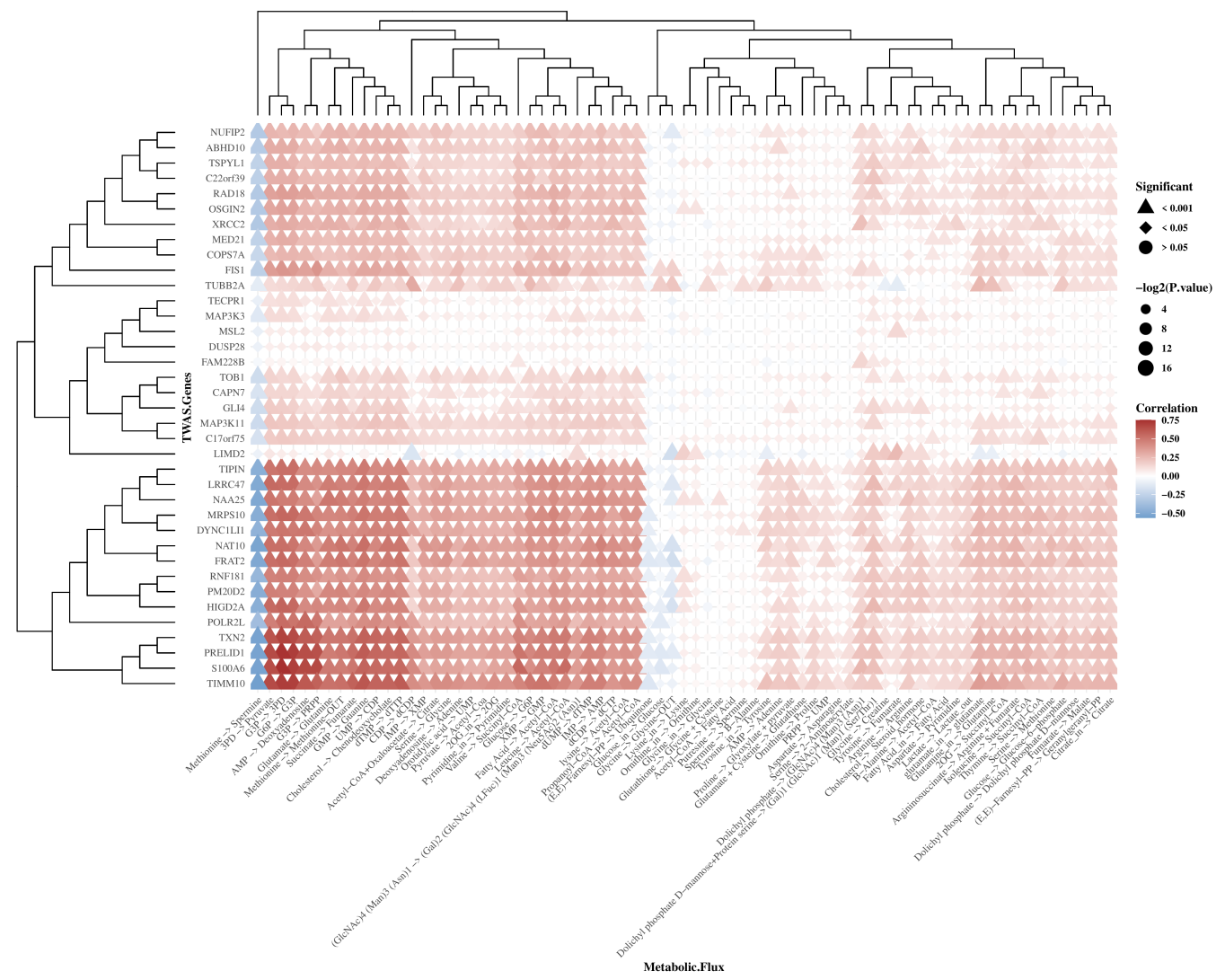
在上述的细胞群中，分析两组 features (即，**基因集** (NAT10, LINC02193, DDX10P1, …[n = 116], 来自于FUSION TWAS全转录组关联研究[Section: AA]) ，与 **代谢通量** (CD4\_naive\_T\_AA - CD4\_naive\_T\_Normal, CD8\_cytoto…, …[n = 10], 来自于Limma 代谢通量差异分析[Section: AA\_FLUX]) )。

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。

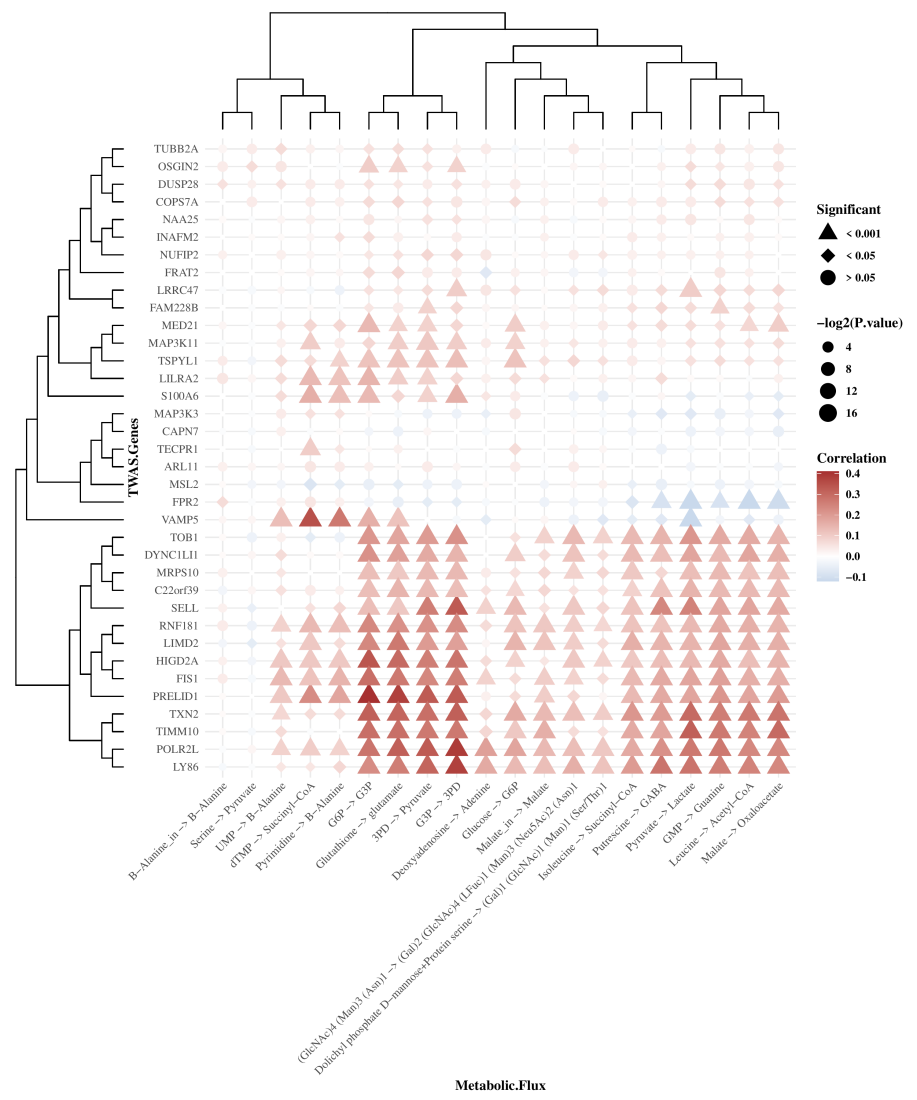
设定 P 阈值 (0.05) 与关联系数 (0.3) 阈值，获取各细胞群 (细胞的筛选算法见方法章节) 中的显著关联组。统计为: EPC (2085), Myeloid\_cell (391), NK\_cytotoxic\_T (69)。设定 P 阈值 (0.05) 与关联系数 (0.3) 阈值，获取各细胞群 (细胞的筛选算法见方法章节) 中的显著关联组。统计为: EPC (2085), Myeloid\_cell (391)。设定 P 阈值 (0.05) 与关联系数 (0.3) 阈值，获取各细胞群 (细胞的筛选算法见方法章节) 中的显著关联组。统计为: EPC (2085), Myeloid\_cell (391), NK\_cytotoxic\_T (69)。



**Fig.** **15** AA EPC correlation heatmap

**(File path: figs/3.9\_细胞群\_features\_关联分析\_(AA)/AA-EPC-correlation-heatmap.pdf)**

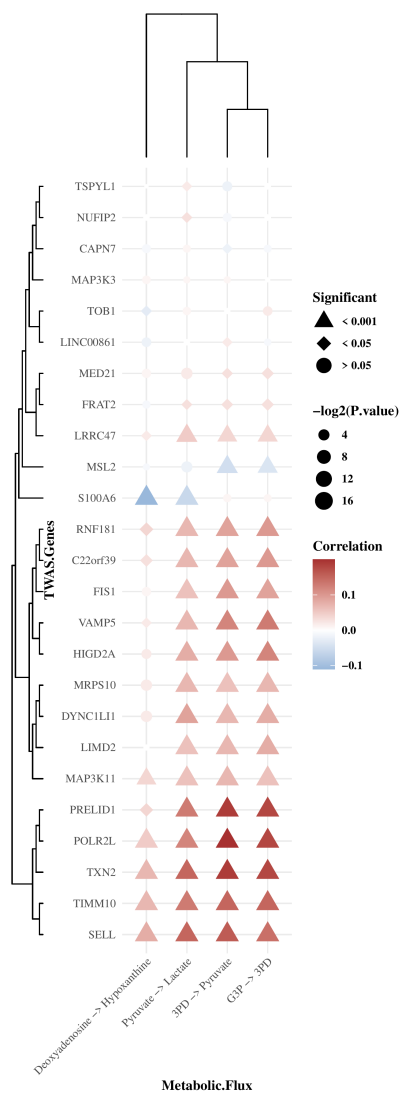
Fig. **[15](#AA-EPC-correlation-heatmap)** 为关联分析 (TWAS Genes, Metabolic Flux) 热图。



**Fig.** **16** AA Myeloid cell correlation heatmap

**(File path: figs/3.9\_细胞群\_features\_关联分析\_(AA)/AA-Myeloid-cell-correlation-heatmap.pdf)**

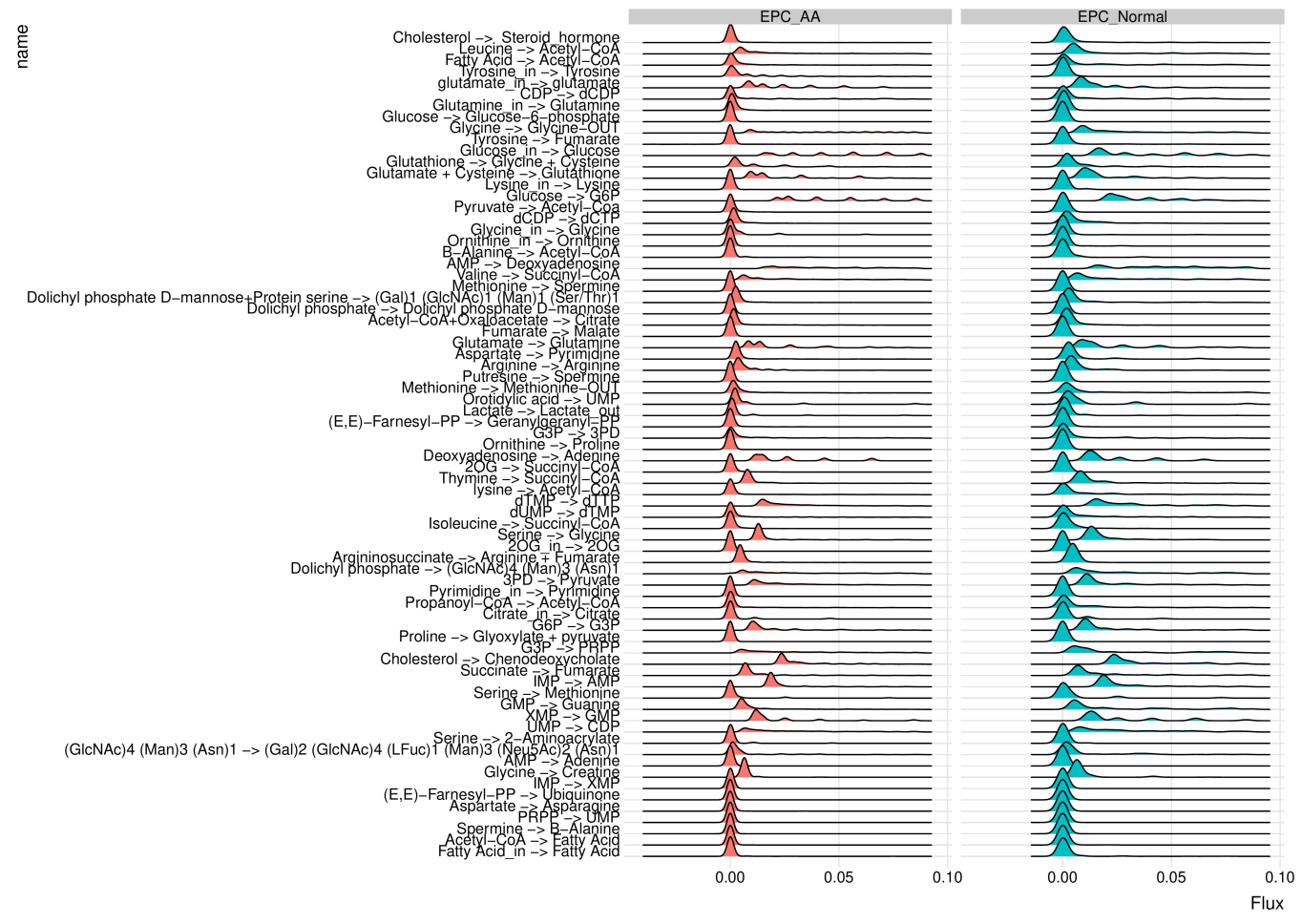
Fig. **[16](#AA-Myeloid-cell-correlation-heatmap)** 为关联分析 (TWAS Genes, Metabolic Flux) 热图。



**Fig.** **17** AA NK cytotoxic T correlation heatmap

**(File path: figs/3.9\_细胞群\_features\_关联分析\_(AA)/AA-NK-cytotoxic-T-correlation-heatmap.pdf)**

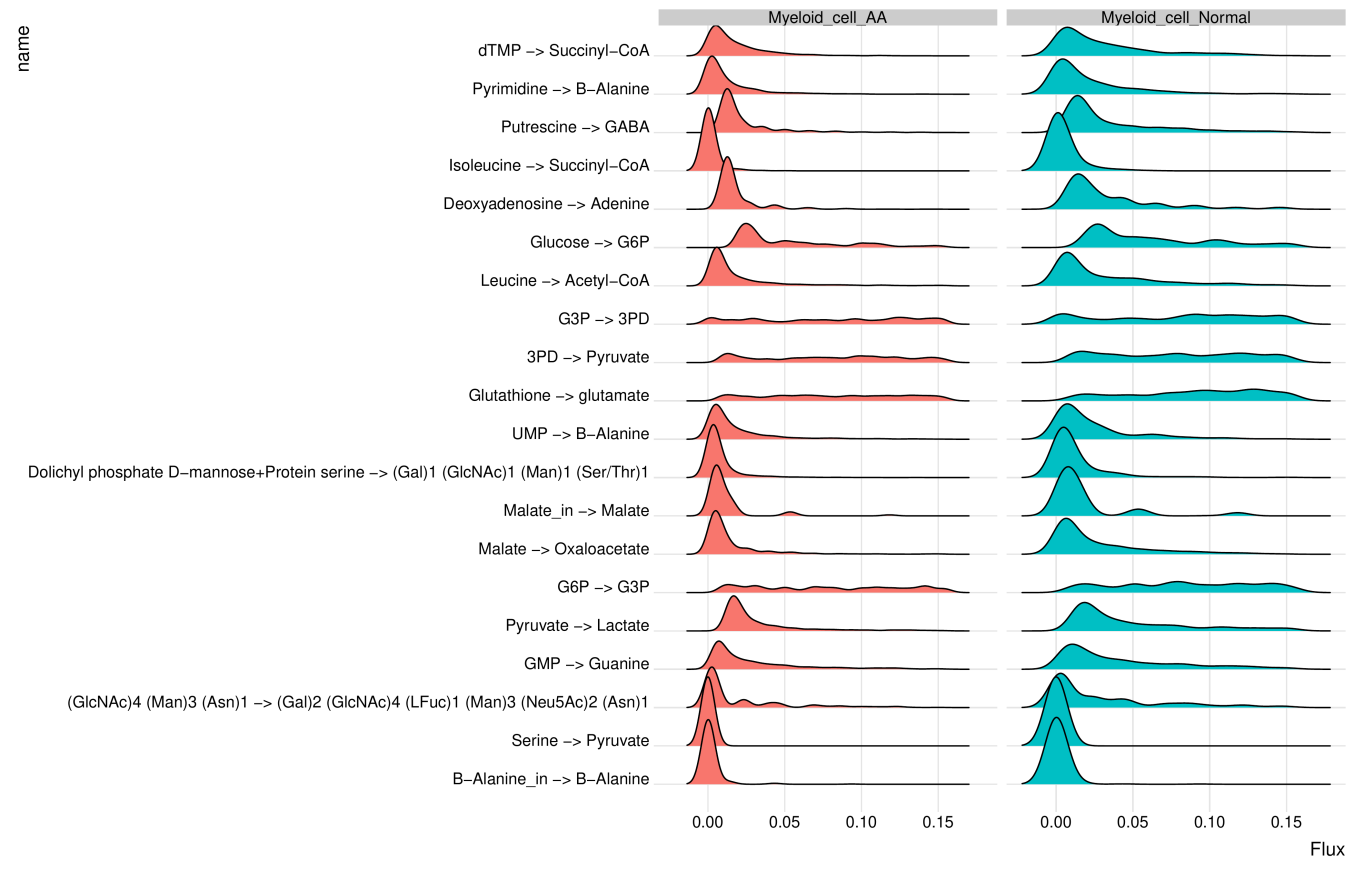
Fig. **[17](#AA-NK-cytotoxic-T-correlation-heatmap)** 为关联分析 (TWAS Genes, Metabolic Flux) 热图。



**Fig.** **18** AA EPC Cell flux ridge plot

**(File path: figs/3.9\_细胞群\_features\_关联分析\_(AA)/AA-EPC-Cell-flux-ridge-plot.pdf)**

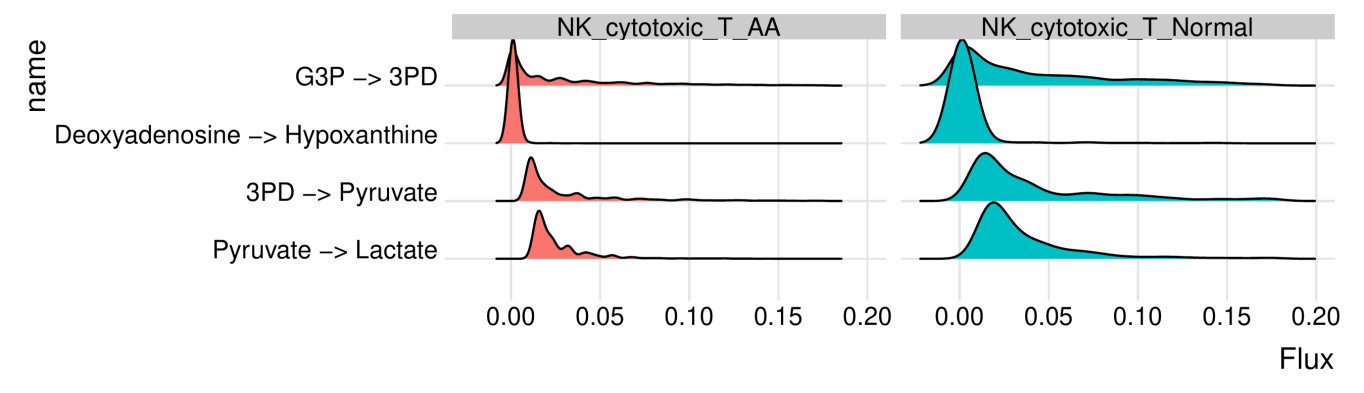
Fig. **[18](#AA-EPC-Cell-flux-ridge-plot)** 为 EPC 的代谢通量山脊图。



**Fig.** **19** AA Myeloid cell Cell flux ridge plot

**(File path: figs/3.9\_细胞群\_features\_关联分析\_(AA)/AA-Myeloid-cell-Cell-flux-ridge-plot.pdf)**

Fig. **[19](#AA-Myeloid-cell-Cell-flux-ridge-plot)** 为 Myeloid\_cell 的代谢通量山脊图。



**Fig.** **20** AA NK cytotoxic T Cell flux ridge plot

**(File path: figs/3.9\_细胞群\_features\_关联分析\_(AA)/AA-NK-cytotoxic-T-Cell-flux-ridge-plot.pdf)**

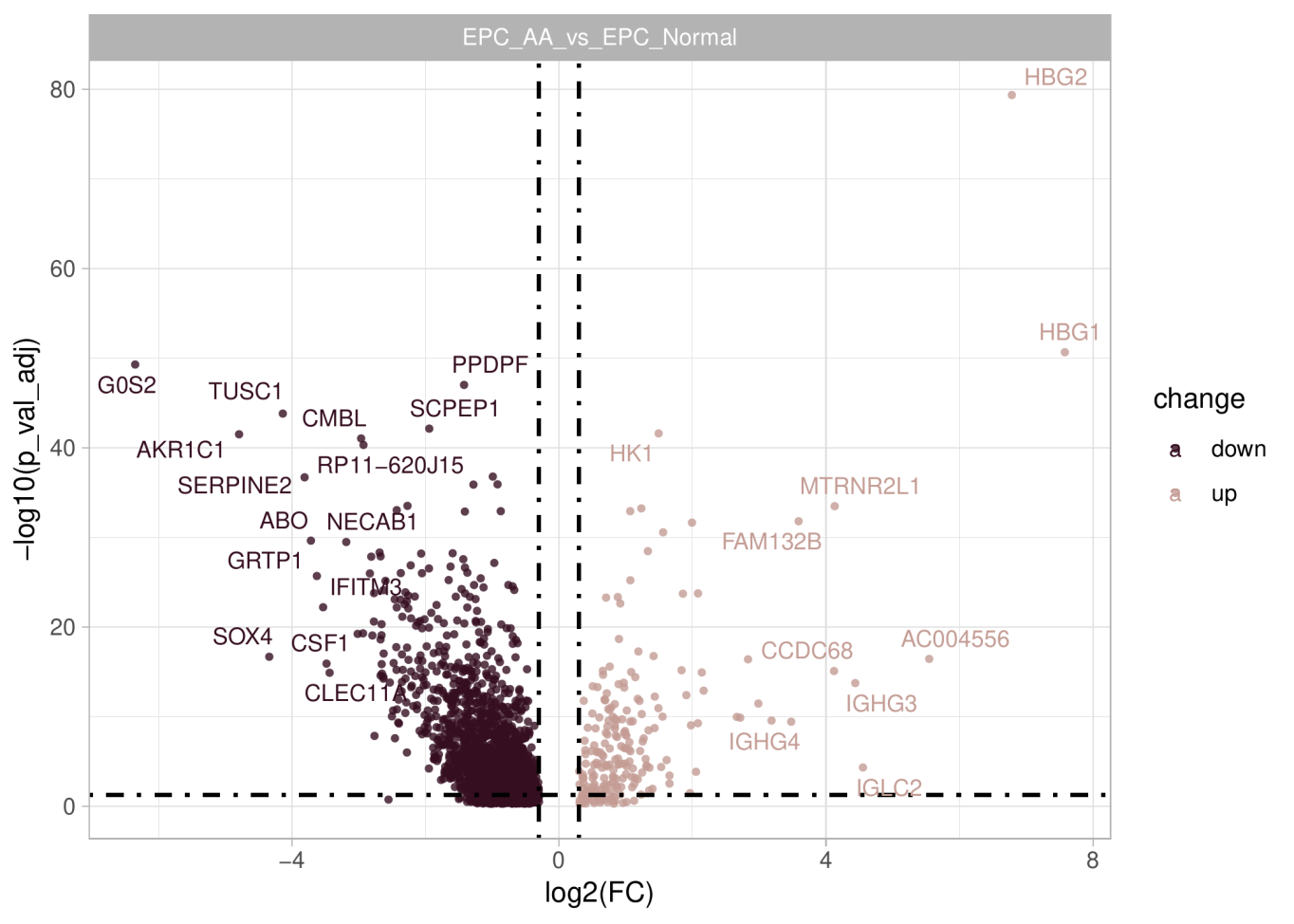
Fig. **[20](#AA-NK-cytotoxic-T-Cell-flux-ridge-plot)** 为 NK\_cytotoxic\_T 的代谢通量山脊图。

## 3.10 细胞差异表达分析 (AA\_EPC)

(EPC, erythroid precursor cells)

对细胞群差异分析 (依据 Cell\_Sample, 分析 “EPC\_AA vs EPC\_Normal”)，筛选差异表达基因。上调或下调统计：

* EPC\_AA\_vs\_EPC\_Normal: up (n=213) , down (n=1928)



**Fig.** **21** AA cell differential expression volcano plot

**(File path: figs/3.10\_细胞差异表达分析\_(AA\_EPC)/AA-cell-differential-expression-volcano-plot.pdf)**

Fig. **[21](#AA-cell-differential-expression-volcano-plot)** 细胞群差异表达基因火山图。

**Tab.** **7** AA DEGs of the contrasts

| Contrast | P val | Avg log2FC | Pct.1 | Pct.2 |
| --- | --- | --- | --- | --- |
| EPC AA vs EPC Normal | 2.73e-84 | 6.785 | 0.831 | 0.102 |
| EPC AA vs EPC Normal | 1.326e-55 | 7.579 | 0.613 | 0.016 |
| EPC AA vs EPC Normal | 3.098e-54 | -6.353 | 0.011 | 0.273 |
| EPC AA vs EPC Normal | 5.861e-52 | -1.424 | 0.757 | 0.906 |
| EPC AA vs EPC Normal | 9.256e-49 | -4.14 | 0.027 | 0.31 |
| ... | ... | ... | ... | ... |

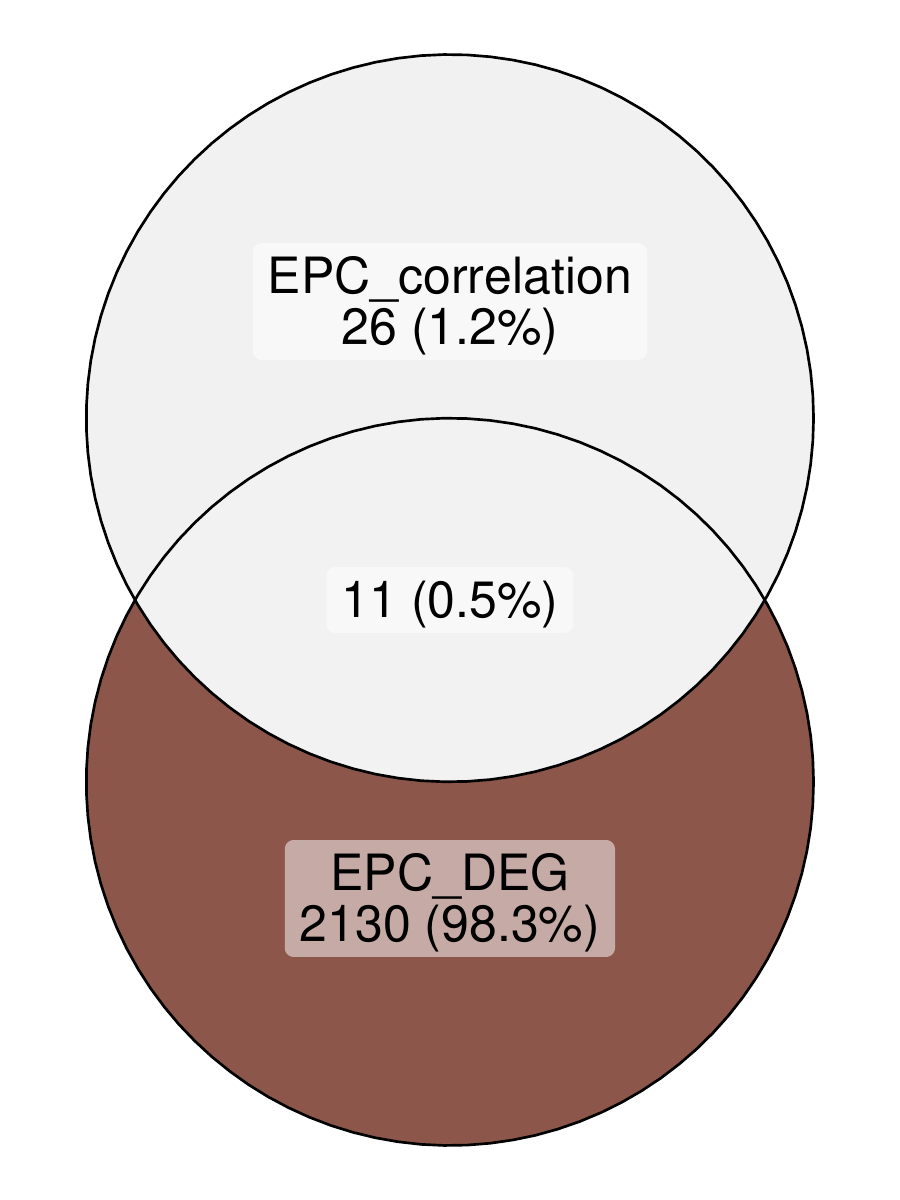
**(File path: tabs/3.10\_细胞差异表达分析\_(AA\_EPC)/AA-DEGs-of-the-contrasts.csv)**

Tab. **[7](#AA-DEGs-of-the-contrasts)** 细胞群差异表达基因附表 (其中 ‘contrast’ 列为比较的两类细胞) (|log2(FC)| > 0.3, P-Adjust < 0.5)。

## 3.11 汇总: EPC\_DEG + EPC\_sig (EPC)

数据集为：

* **基因集** (EPC\_AA\_vs\_EPC\_Normal, 来自于Seurat 细胞群差异表达分析[Section: AA])
* **基因集** (S100A6, PRELID1, TIMM10, …[n = 37], 来自于细胞群 features 关联分析[Section: AA])



**Fig.** **22** EPC Intersection of EPC DEG with EPC correlation

**(File path: figs/3.11\_汇总:\_EPC\_DEG\_+\_EPC\_sig\_(EPC)/EPC-Intersection-of-EPC-DEG-with-EPC-correlation.pdf)**

* All\_intersection: TUBB2A, S100A6, HIGD2A, PM20D2, LIMD2, NAA25, NAT10, FRAT2, TXN2, POLR2L, RNF181

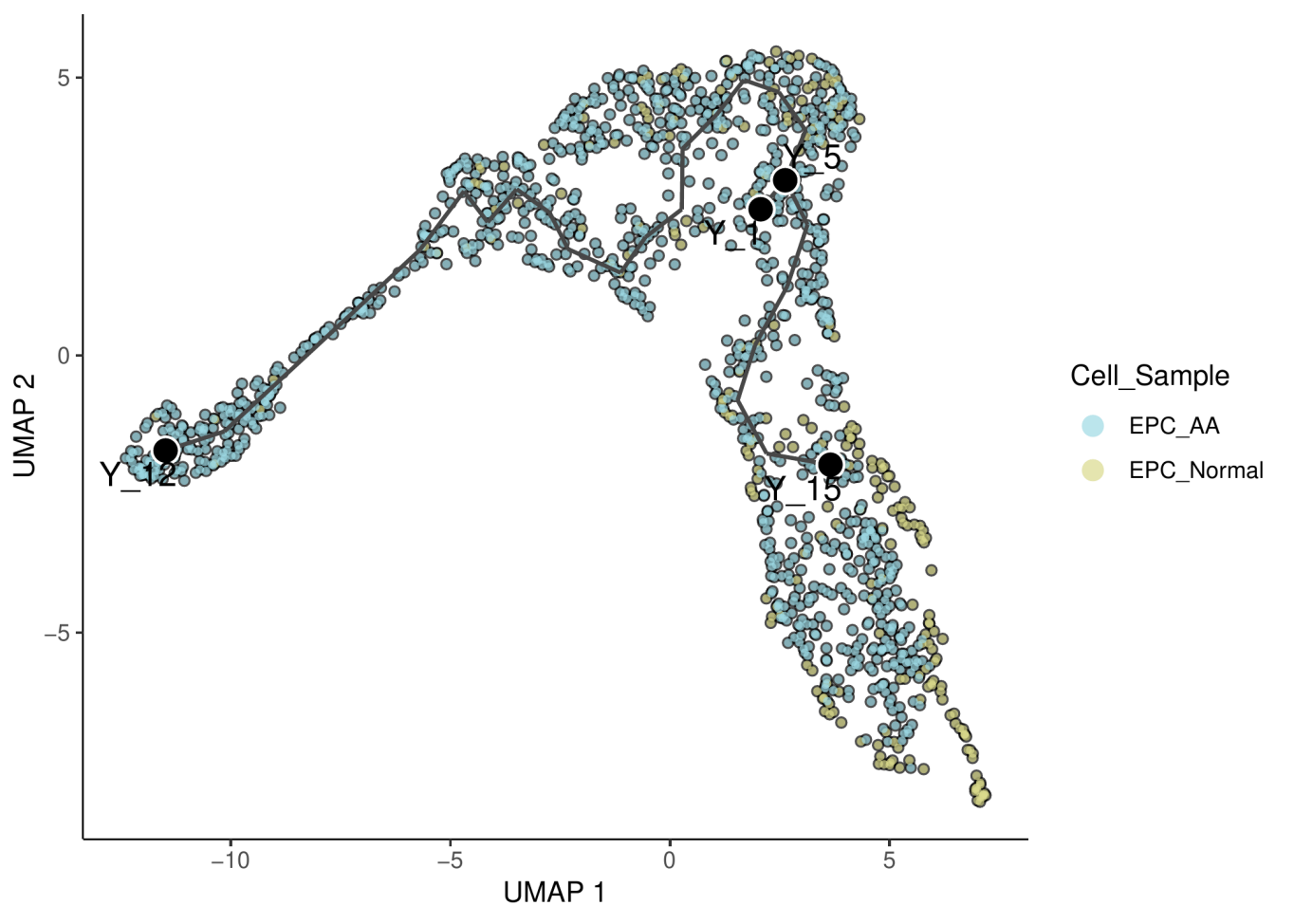
**(See: tabs/3.11\_汇总:\_EPC\_DEG\_+\_EPC\_sig\_(EPC)/EPC-Intersection-of-EPC-DEG-with-EPC-correlation-content)**

Fig. **[22](#EPC-Intersection-of-EPC-DEG-with-EPC-correlation)** 为 EPC\_DEG, EPC\_correlation 各自交集。

## 3.12 Monocle3 拟时分析 (EPC)

从 Seurat 数据对象 Cell\_Sample 中提取 EPC 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\_15 (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。

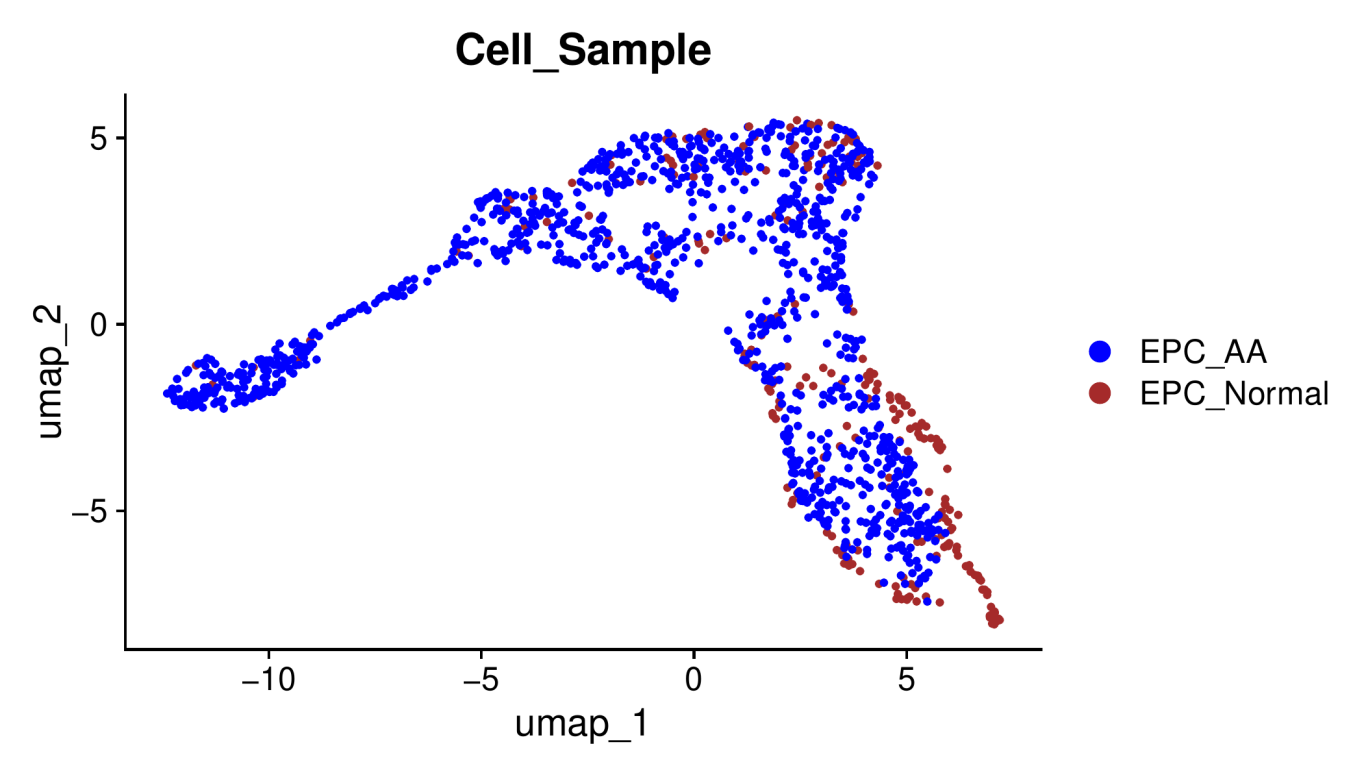
从 Seurat 数据对象 Cell\_Sample 中提取 EPC 类型的细胞，对其重新聚类分析。将 Seurat 数据对象转化为 Monocle3 数据对象 (详见方法章节)。构建轨迹图 (Trajectory)。选择 Y\_15 (principle points) 为拟时起点。构建细胞的拟时变化图 (Pseudotime)。寻找单细胞拟时轨迹 (monocle3::graph\_test) 中差异表达的基因。探究 **基因集** (TUBB2A, S100A6, HIGD2A, …[n = 11], 来自于Venn 交集[Section: EPC]) 在拟时轨迹中的表达变化。这些基因中，共 11 属于 Moran’s I test (Graph Test) 差异表达基因。仅分析这些基因。



**Fig.** **23** EPC principal points

**(File path: figs/3.12\_Monocle3\_拟时分析\_(EPC)/EPC-principal-points.pdf)**

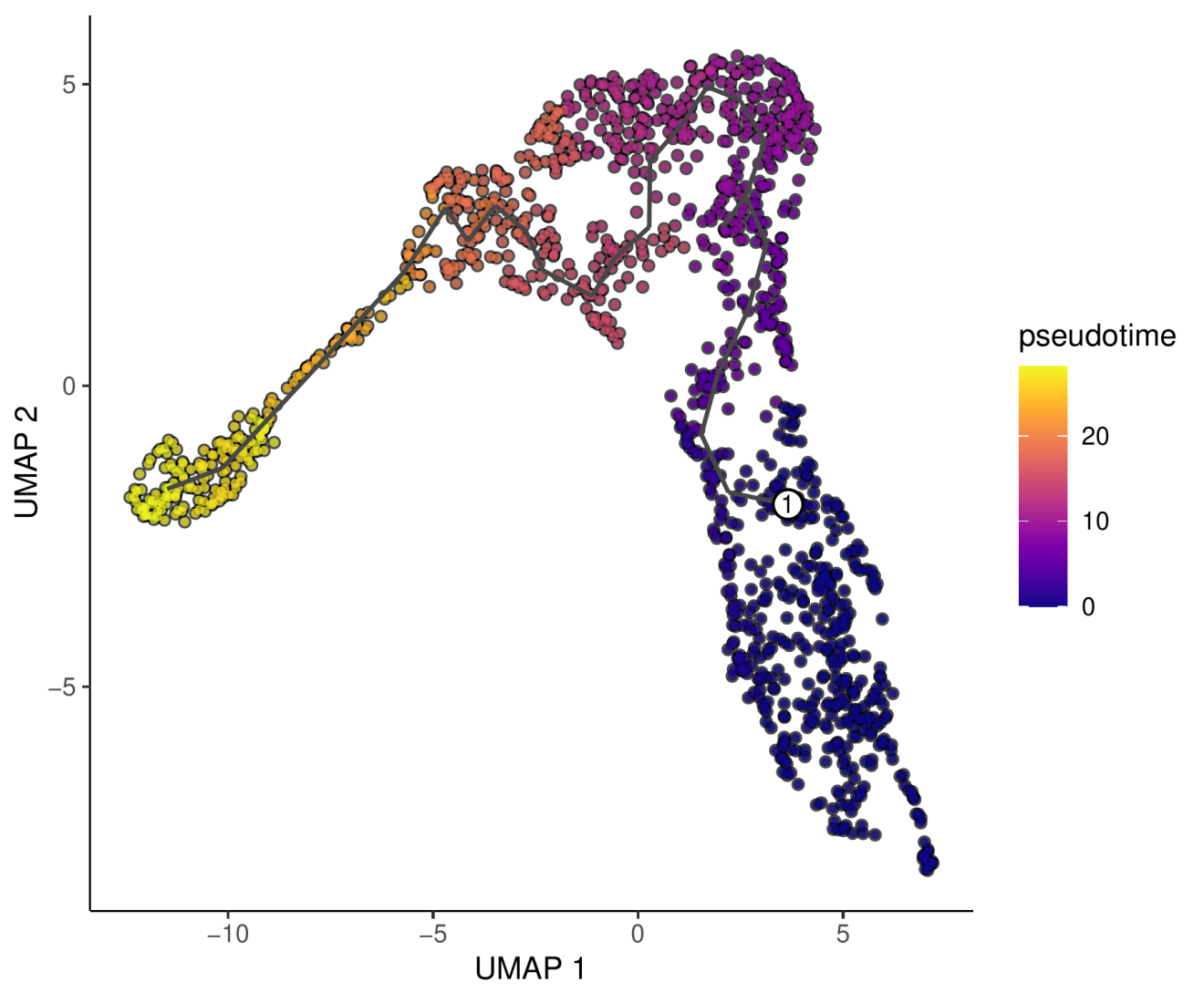
Fig. **[23](#EPC-principal-points)** 为拟时轨迹与 principal point 示意。



**Fig.** **24** AA EPC The Cell Sample

**(File path: figs/3.12\_Monocle3\_拟时分析\_(EPC)/AA-EPC-The-Cell-Sample.pdf)**

Fig. **[24](#AA-EPC-The-Cell-Sample)** 为 Cell\_Sample 的 umap 聚类图。



**Fig.** **25** EPC pseudotime

**(File path: figs/3.12\_Monocle3\_拟时分析\_(EPC)/EPC-pseudotime.pdf)**

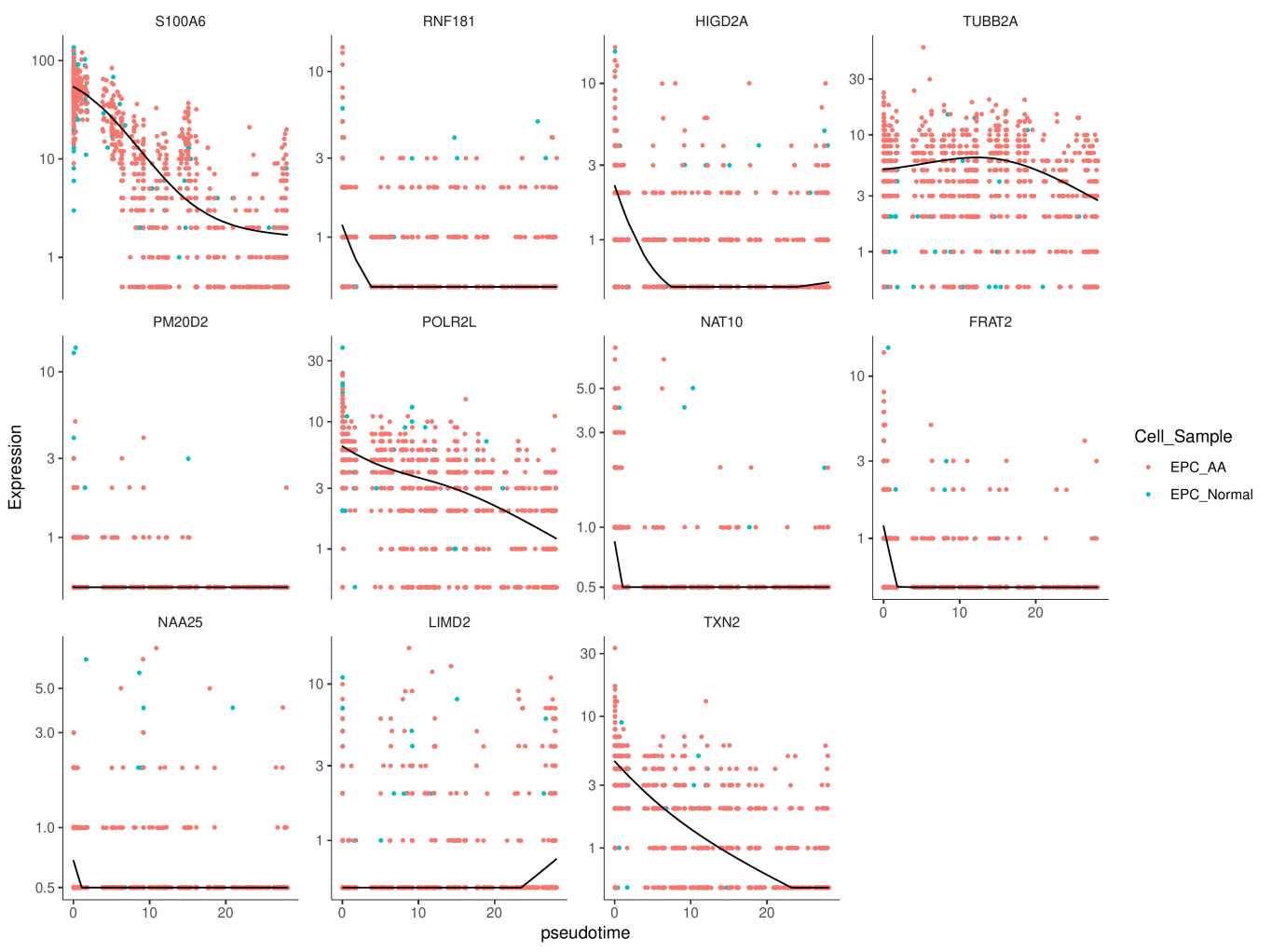
Fig. **[25](#EPC-pseudotime)** 为细胞拟时图。

**Tab.** **8** EPC Graph Test Significant genes

| Gene id | Status | P value | Morans test stati... | Morans I |
| --- | --- | --- | --- | --- |
| SLC4A1 | OK | 0 | 117.9 | 0.8879 |
| FAM178B | OK | 0 | 110.4 | 0.8322 |
| PPP1R14B | OK | 0 | 109.8 | 0.8279 |
| SYNGR1 | OK | 0 | 108.7 | 0.8194 |
| HBA2 | OK | 0 | 108.6 | 0.816 |
| ... | ... | ... | ... | ... |

**(File path: tabs/3.12\_Monocle3\_拟时分析\_(EPC)/EPC-Graph-Test-Significant-genes.csv)**

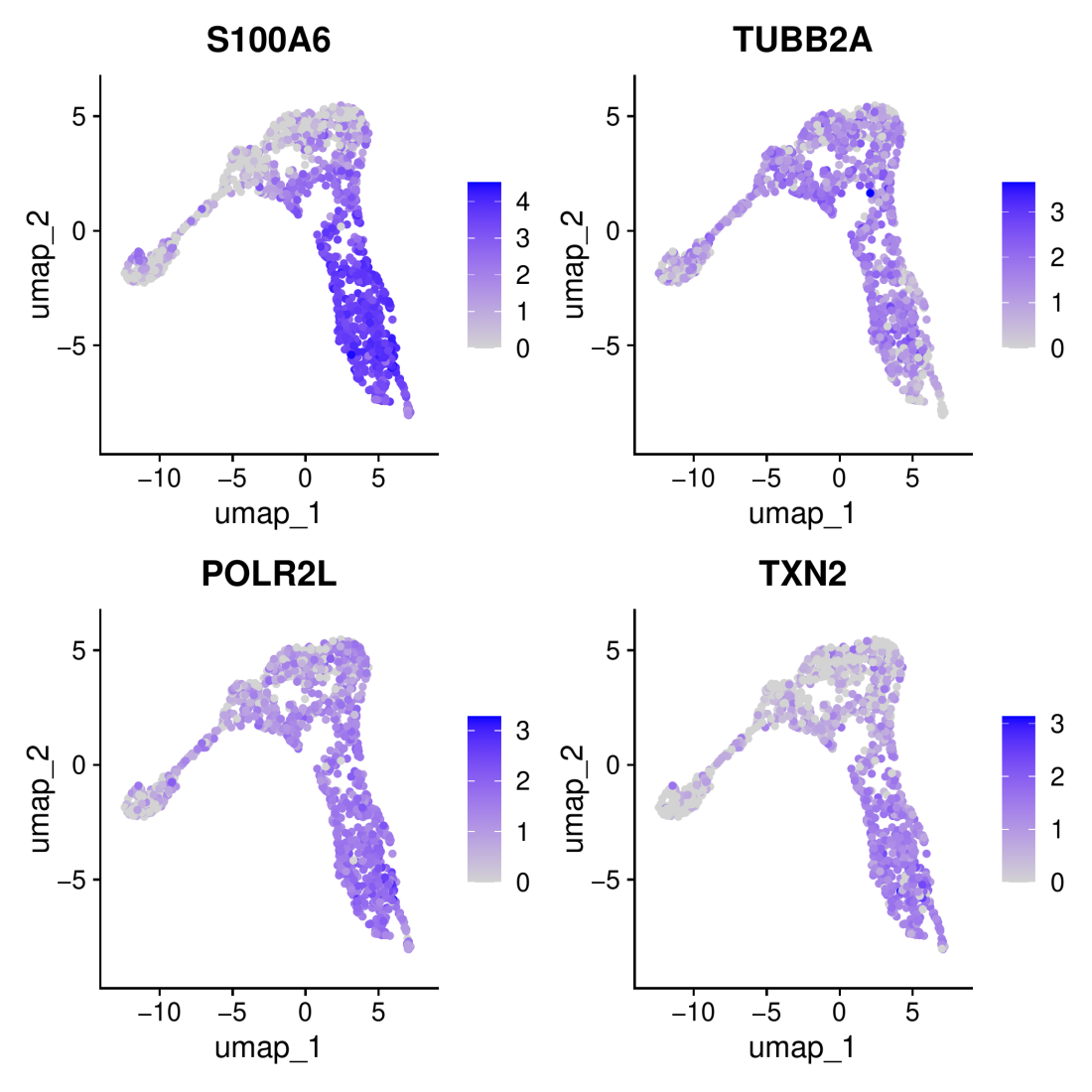
Tab. **[8](#EPC-Graph-Test-Significant-genes)** 为 Moran’s I test (Graph Test) 筛选的差异表达基因 (Q-value cutoff: 0.05)。



**Fig.** **26** EPC Set1 genes in pseudotime

**(File path: figs/3.12\_Monocle3\_拟时分析\_(EPC)/EPC-Set1-genes-in-pseudotime.pdf)**

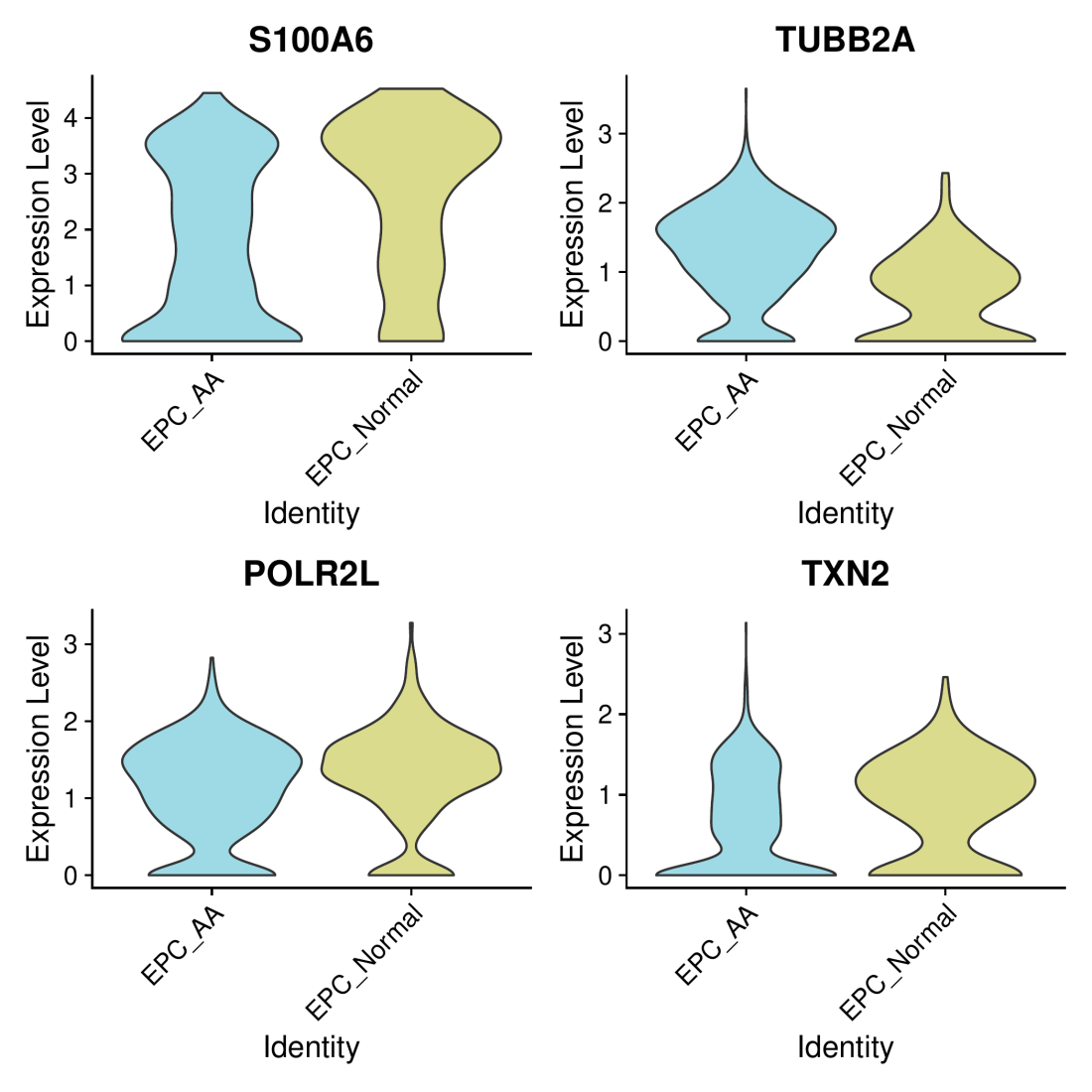
Fig. **[26](#EPC-Set1-genes-in-pseudotime)** 为 TUBB2A, S100A6, HIGD2A, …(n = 11) 在拟时 (Pseudotime) 中的表达变化。



**Fig.** **27** AA dimension plot of expression level genes

**(File path: figs/3.12\_Monocle3\_拟时分析\_(EPC)/AA-dimension-plot-of-expression-level-genes.pdf)**

Fig. **[27](#AA-dimension-plot-of-expression-level-genes)** 基因 **基因集** (S100A6, TUBB2A, POLR2L, …[n = 4], 来自于Monocle3 拟时分析[Section: EPC]) 表达水平的 Dimension reduction plot.



**Fig.** **28** AA violing plot of expression level genes

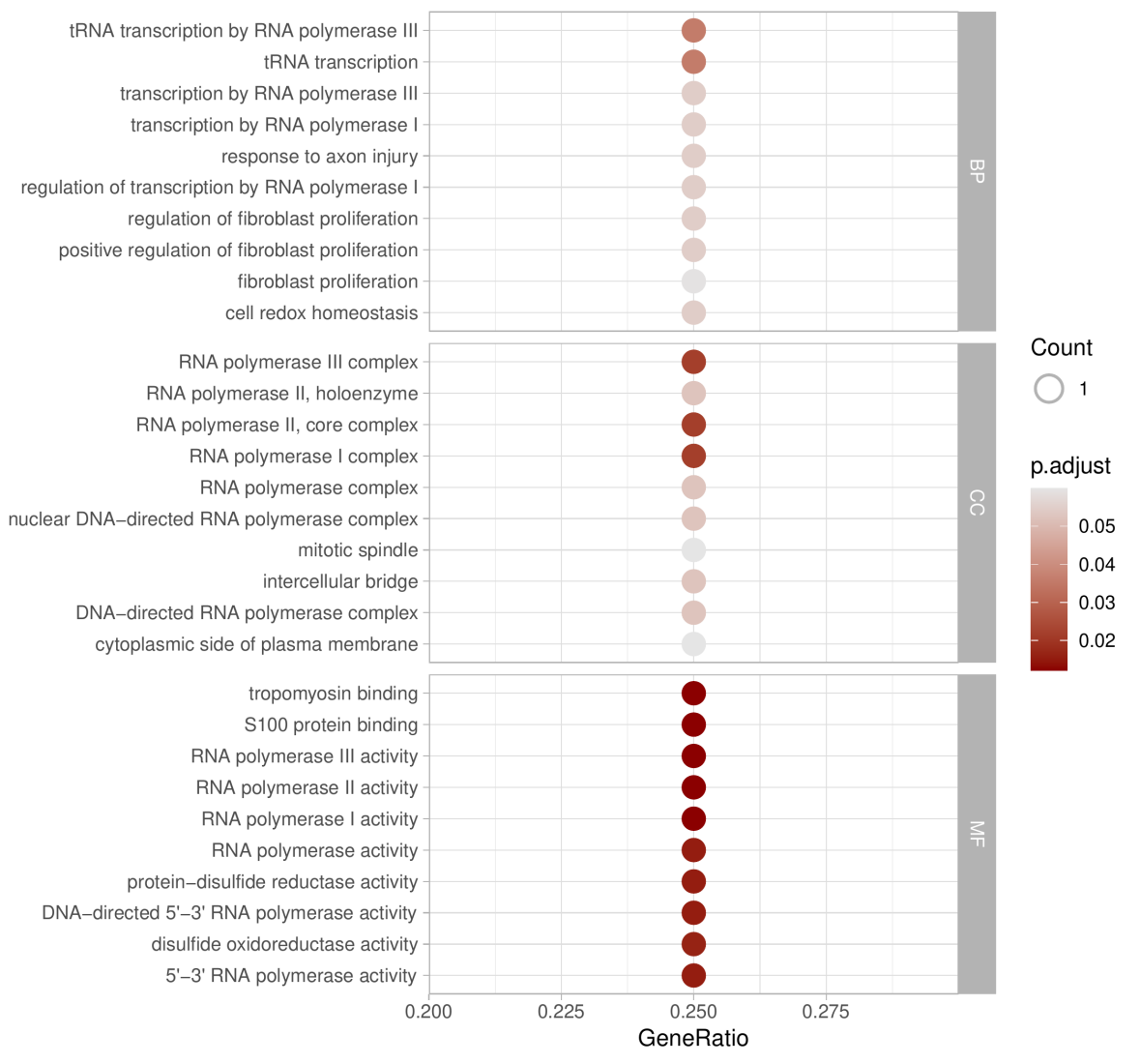
**(File path: figs/3.12\_Monocle3\_拟时分析\_(EPC)/AA-violing-plot-of-expression-level-genes.pdf)**

Fig. **[28](#AA-violing-plot-of-expression-level-genes)** 基因 **基因集** (S100A6, TUBB2A, POLR2L, …[n = 4], 来自于Monocle3 拟时分析[Section: EPC]) 表达水平的小提琴图。

## 3.13 ClusterProfiler 富集分析 (EPC)

以 ClusterProfiler R 包 (4.15.0.2) (2021, **IF:33.2**, Q1, The Innovation)8进行 KEGG 和 GO 富集分析。以 p.adjust 表示显著水平。

对**基因集** (S100A6, TUBB2A, POLR2L, …[n = 4], 来自于Monocle3 拟时分析[Section: EPC]) 进行ClusterProfiler 富集分析。



**Fig.** **29** EPC GO enrichment

**(File path: figs/3.13\_ClusterProfiler\_富集分析\_(EPC)/EPC-GO-enrichment.pdf)**

Fig. **[29](#EPC-GO-enrichment)** 为 GO 富集分析气泡图。

# 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，本研究探究了 TWAS 筛选出的基因与代谢通量改变的关联性。

主要有三种类型的细胞 (EPC, Myeloid\_cell, NK\_cytotoxic\_T)，能检测到较高表达量 (至少包含一部分细胞 Expression > 0) 的 TWAS 相关基因，并且存在代谢通量的差异变化。

其中 EPC (erythroid precursor cells) 发现了大量与代谢通量改变呈正相关的基因 (TWAS) Fig. **[15](#AA-EPC-correlation-heatmap)** 。 进一步分析了 EPC 细胞中的差异表达基因 (DEGs，AA vs Normal)，并将这些与上述筛选的关联基因交集， Fig. **[22](#EPC-Intersection-of-EPC-DEG-with-EPC-correlation)** ， 得到 **基因集** (TUBB2A, S100A6, HIGD2A, …[n = 11], 来自于Venn 交集[Section: EPC]) 。

提取 EPC 细胞构建拟时分析。Fig. **[24](#AA-EPC-The-Cell-Sample)** 可以看出，Normal 类型 EPC 仅表达于一侧的亚群。 将这部分的细胞选作拟时起点。 拟时分析可以发现，**基因集** (S100A6, TUBB2A, POLR2L, …[n = 4], 来自于Monocle3 拟时分析[Section: EPC]) 表达量梯度下降 (Fig. **[26](#EPC-Set1-genes-in-pseudotime)** )。 这些基因仅表达于分化类型接近于 Normal EPC 中，而在拟时末期中几乎不表达。 这些基因在 AA 与 Normal 有显著差异。

# 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. 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).

7. 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).

8. Wu, T. *et al.* ClusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *The Innovation* **2**, (2021).