

结 题 报 告

项目名称：_____

客户单位：_____

客户姓名：_____

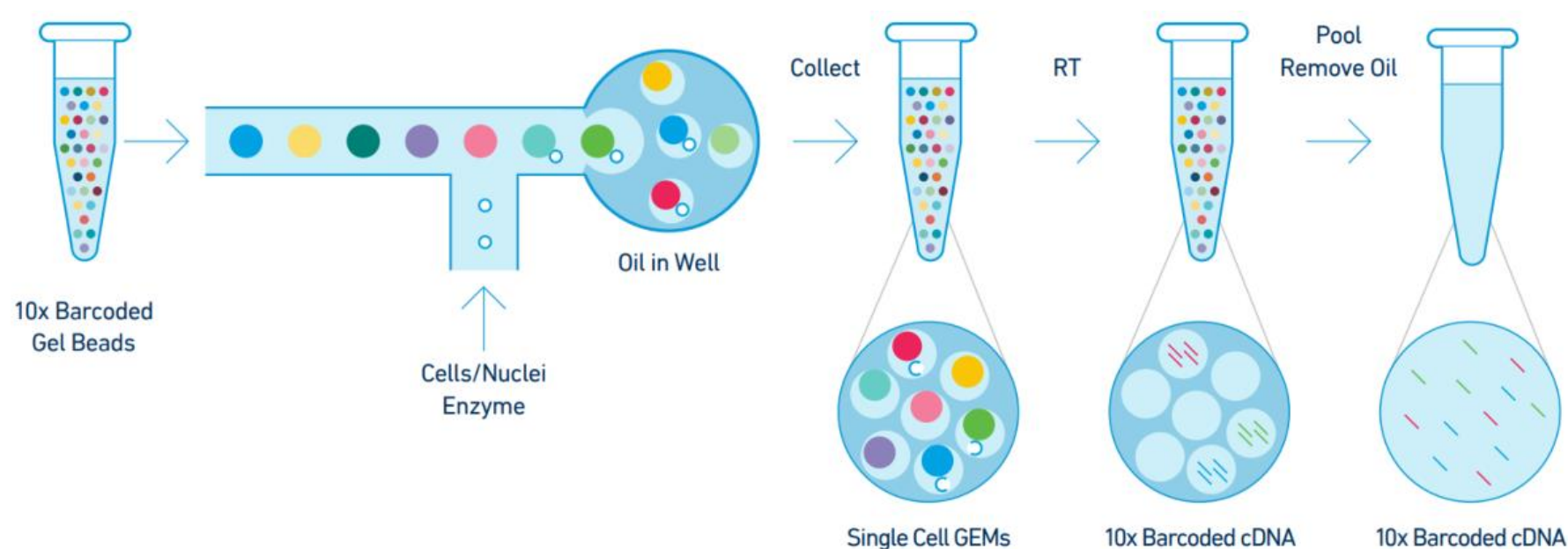
完成时间：_____

一、工作流程

1.1 技术简介

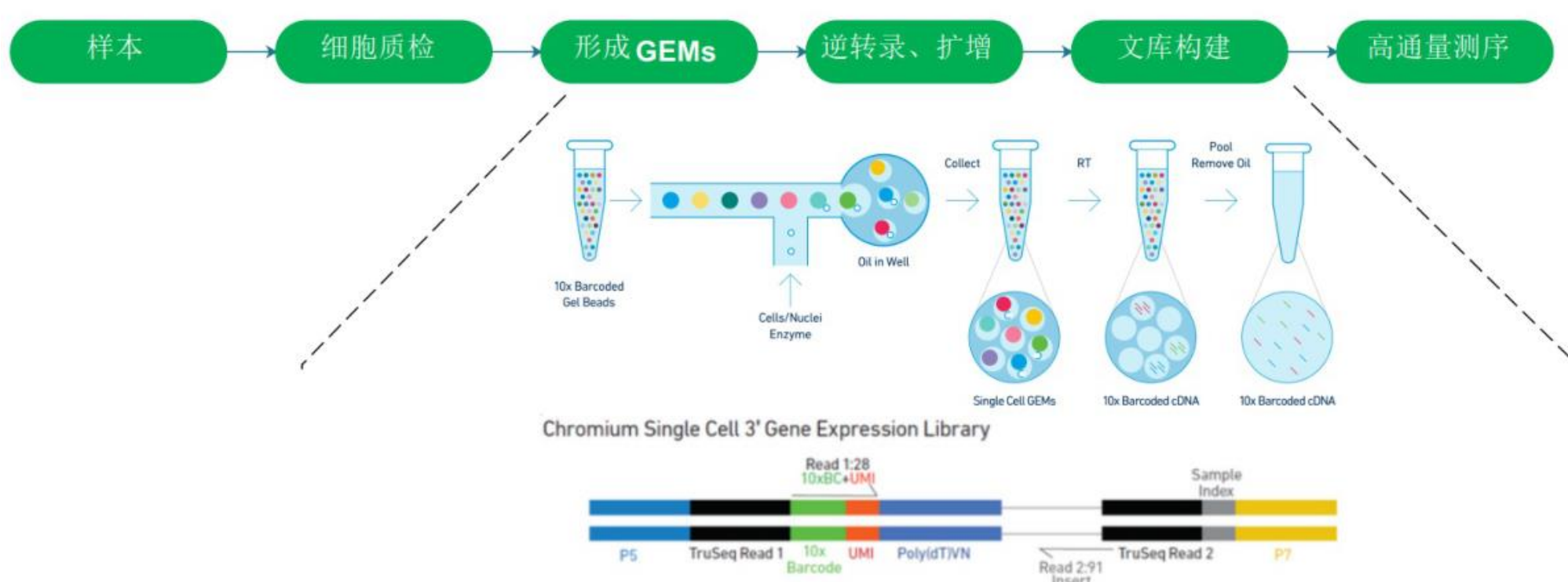
该技术利用 10x Genomics Chromium 系统在单细胞水平进行高通量基因表达谱检测，对复杂细胞群进行深入分析，表征单个细胞的表达谱，避免单个细胞的异质性生物学信息被大量细胞的均质化掩盖。

其核心技术是利用微流控芯片对细胞/细胞核进行精确区分，使每个小的反应体系（Gel Bead in Emulsion, GEM）中仅有一个细胞/细胞核。



1.2 实验流程

利用 10x Genomics Chromium TM 系统，将带有序列标签的 Gel Bead、样本和试剂预混液、油加载到各自的进样通道，通过微流体通道网络形成的“双十字”交叉系统，最终形成由油滴包裹的单细胞微反应体系 GEMs。



（1）细胞准备

对单细胞悬液进行质检和计数，一般要求细胞存活率在 80%以上，将检测合格的细胞经洗涤、重悬制备成合适的细胞浓度 700~1200 cells/ul，以备进行 10x Genomics Chromium™系统上机操作。

（2）GEM 生成和标签化生成和标签化

根据预期获得的目的细胞数目上样，构建 GEMs（Gel Bead in Emulsion）进行

单细胞分离，GEMs 正常形成后，收集 GEMs 在 PCR 仪中进行反转录实现标签化。

(3) Post GEM-RT Post GEM-RT 纯化和纯化和 cDNA cDNA 扩增扩增

对 GEMs 破油处理，用磁珠纯化富集一链 cDNA，然后进行 cDNA 扩增和质检。

(4) 文库构建和定量

将质检合格的 cDNA 进行二代测序文库构建，经片段化、连接测序接头、样本 Index PCR 等实验过程，最终对文库定量质检。

(5) 上机测序

对构建完成的文库测序，使用 Illumina Novaseq/DNBSEQ（此处根据实际项目情况选择性区分）高通量测序平台，采用 PE150 测序模式，建议测序量达到 20 k reads/cell 及以上。

实验步骤	所需仪器
GEM生成	10X Genomics
Post GEM-RT纯化和cDNA扩增	PCR仪
文库定量	Qubit
簇生成	CBOT簇生成仪
上机测序	Illumina Novaseq /DNBSEQ测序平台

二、分析报告

2.1 项目结果速览

2.1.1 CellRanger 测序数据统计

参考基因来源：Homo_sapiens Homo_sapiens;参考基因组版本：GRCh38 GRCh38; 参 考 基 因 组 来 源 :

https://apr2022.archive.ensembl.org/Homo_sapiens/Info/Index

https://apr2022.archive.ensembl.org/Homo_sapiens/Info/Index 完成 22 个样品的单细胞转录组分析，共获得 156.03 156.03Gb RawData;

各样品捕获细胞数分布在 10850 到 1431614316 个，共捕获细胞数 25166 个，每个细胞中位值分布在 1863 到 30473047 个，有效 Barcodes 范围分布在 98.2%到 98.4%

2.1.2 细胞质量过滤与评估

本次过滤标准如下：线粒体（已过滤）、核糖体、红细胞（仅限动物）、双细胞统计与过滤（已过滤）。

过滤标准以及过滤前后结果见下表：

sample	ncount_rna	nfeature_rna	percent_mito	percent_rb	percent_redcell	percent_double_cell	num_orig	num_filtered
DTC_Melanoma_10k	22163	200-4376	18.552%	29.014%	62.734%	10.262%	14213	12077
SI_PBMC_10k	24808	200-5626	16.419%	40.973%	0.02%	6.812%	10784	9234

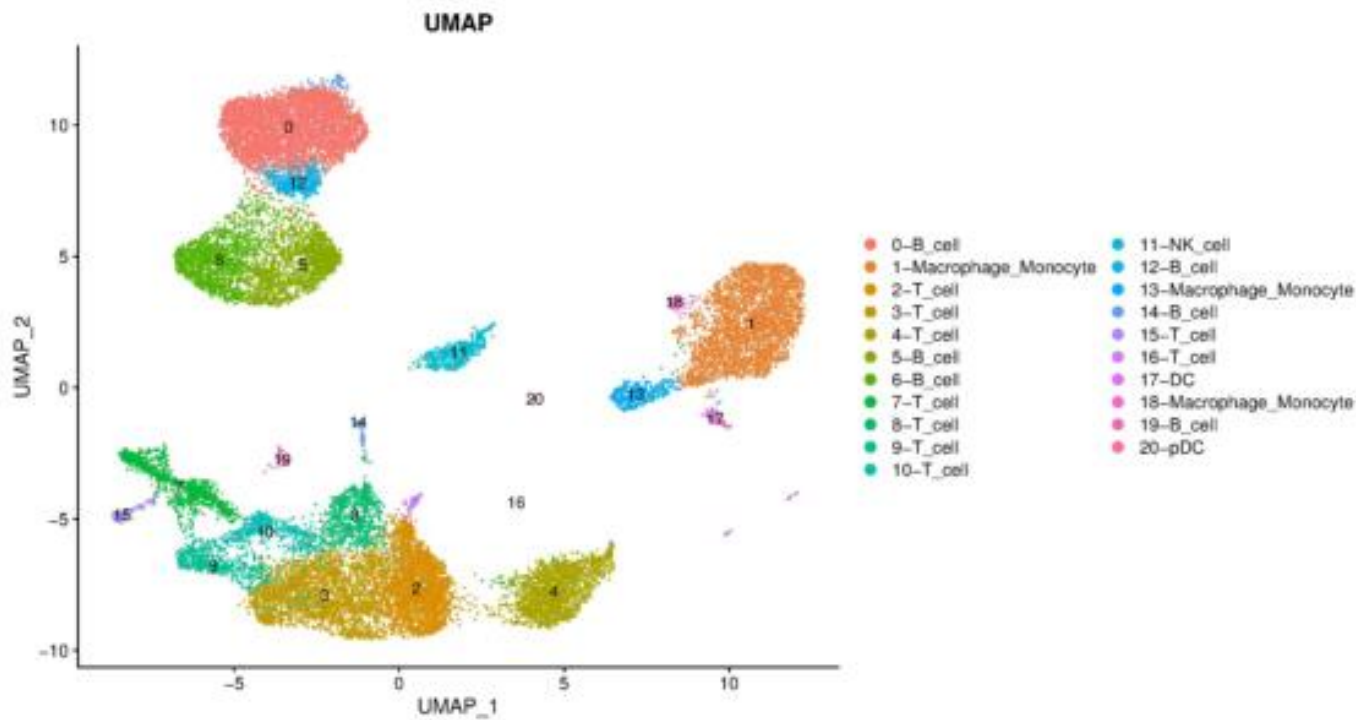
2.1.3 降维与聚类分析

样本整体降维聚类后共为 17 群细胞。

2.1.4 细胞类型注释

初步鉴定方法为：SingleR 所用数据库为：SingleR 默认共鉴定到 66 种细胞。

整体 UMAP 可视化结果见下：



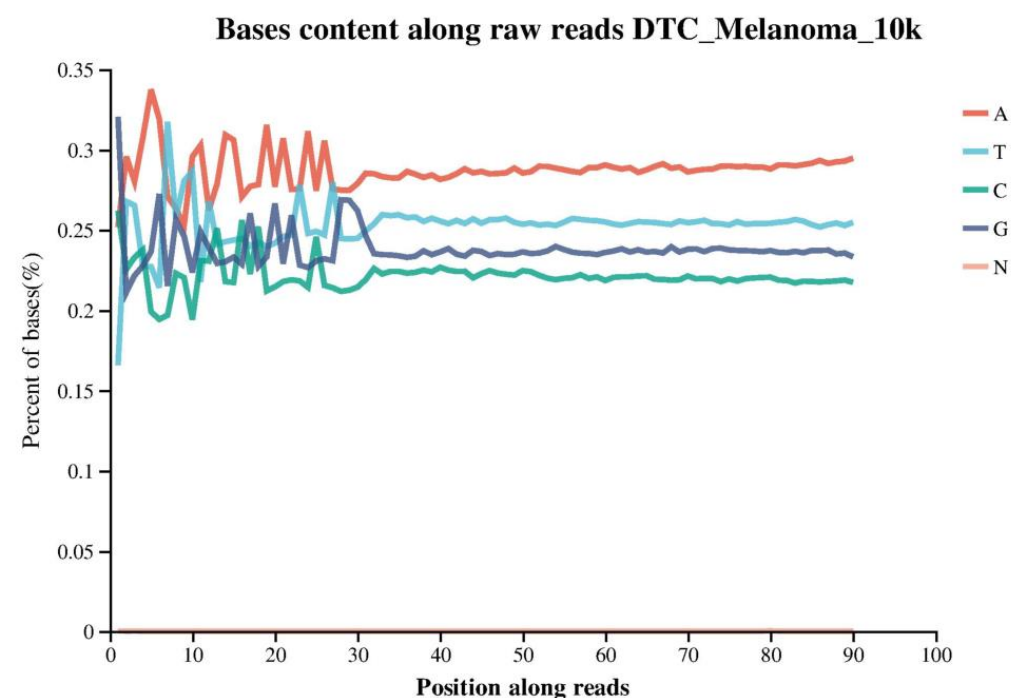
2.2 基础分析

2.2.1 原始数据统计与质量评估

10x Genomics 单细胞转录组高通量测序属于第二代测序技术，单次运行能产生数十亿级的 reads，如此海量的数据无法逐个展示每条 read 的质量情况；运用统计学的方法，对所测序列进行

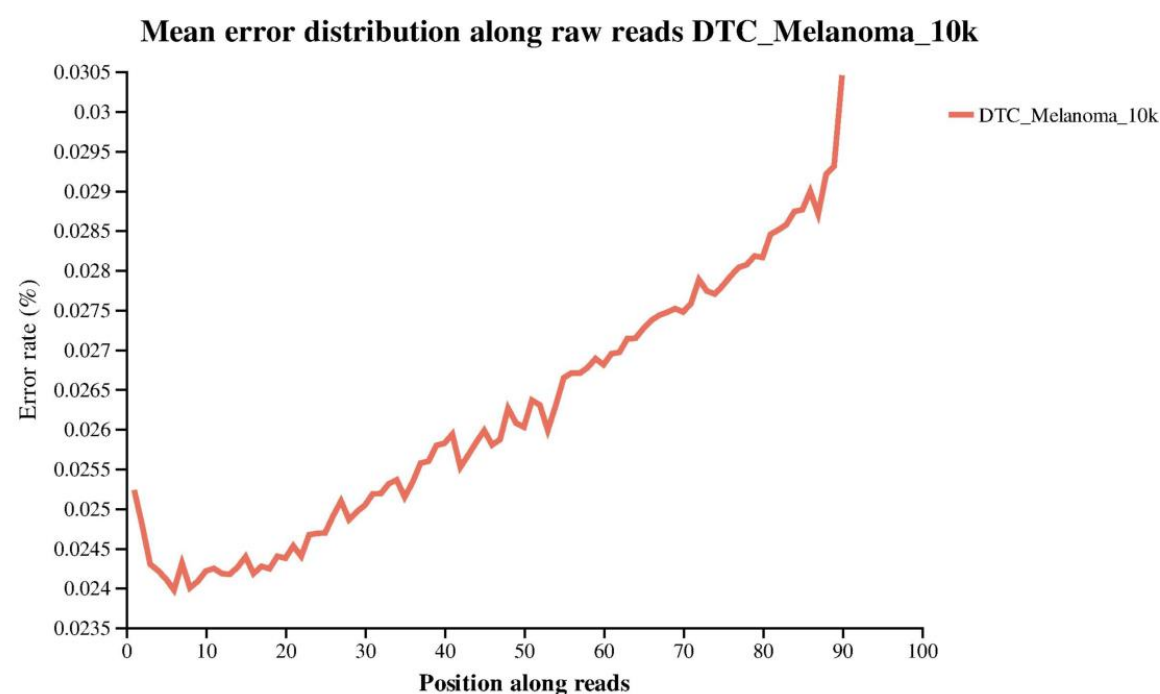
统计和质控，可以从宏观上直观地反映出样本的文库构建质量和测序质量。

原始数据碱基含量分布图



注：横坐标是 reads 的碱基坐标，纵坐标是所有 reads 在该测序位置（如第一个测序碱基）A、C、G、T、N 碱基分别占的百分比，其中 N 指的是模糊碱基，即不能准确判断为 A、T、G、C 的碱基，不同碱基用不同颜色表示。

原始数据碱基错误率分布图



注：展示所有测序 reads 的各碱基测序错误率均值分布，一般在 0.1% 以下。横坐标是 reads 的碱基坐标，表示 reads 上从 5'到 3'端依次碱基的排列；纵坐标是所有 reads 在该位点处碱基的平均错误率。

原始测序数据统计表

Sample	Raw reads	Raw bases	Q20(%)	Q30(%)	GC content(%)
DTC_Melanoma_10k	427455023	76941904140	97.54	93.15	45.97
SI_PBMK_10k	434556032	79089197824	97.54	93.31	46.65

注：（1）Sample：样品名称；（2）Raw reads：原始测序数据的总条目数（reads，代表测序读段，一个 reads 即为一行）；（3）Raw bases：原始测序总数据量（即 Raw reads 数目乘以 reads 读长）；（4）、（5）Q20（%）、Q30（%）：对质控后测序数据进行质量评估，Q20、Q30 分别指测序质量在 99%和 99.9%以上的碱基

占总碱基的百分比，一般Q20在85%以上，Q30在80%以上；（6）GC content（%）：原始数据对应的G和C碱基总和占总碱基的百分比。

2.2.2 CellRanger 数据统计

2.2.2.1 CellRanger 数据统计

得到测序数据之后，需要对这部分数据进行质量评估。我们使用 10x Genomics 官方软件 Cell Ranger 对原始数据进行质控。Cell Ranger 是 10x genomics 公司官方提供的专门用于其

单细胞转录组数据分析的软件包。Cell Ranger 使用比对软件 star，将测序获得的 reads 比对到参考基因组上，可以获得样本数据的高质量细胞数、基因数、基因组比对信息等质控结果。Cell Ranger 也可以通过 barcode 序列标记区分细胞，通过每个细胞内不同 mRNA 的 UMI 标记对单细胞转录组数据进行定量。

cellranger 测序数据统计表

Sample	Estimated Nu...	Mean Reads p...	Median Gene...	Number of Re...	Valid Barcodes	Fraction Read...	Sequencing S...	Reads Mappe...	Reads Mappe...
DTC_Melanom...	14,316	29,859	1,863	427,455,023	98.4%	87.3%	66.0%	97.7%	56.7%
SI_PBMK_10k	10,850	40,051	3,047	434,556,032	98.2%	95.1%	62.7%	97.4%	61.8%

表注：

Sample: 样本名称

Estimated Number of Cells: 检测到的细胞数量

Mean reads per Cell: 每个细胞的平均 reads 数

Median genes per Cell: 细胞中的中位基因数，比对到序列大于 0 被定义为检测到的基因

Valid Barcodes: 带有正确标记的序列数百分比，每个标记对应到每个细胞

Fraction Reads in Cells: 落于高质量细胞的序列数百分比

Sequencing Saturation: 测序饱和度

Reads Mapped to Genome: 比对到数据库参考基因组的序列数百分比

Reads Mapped Confidently to Exonic Regions: 比对到数据库的外显子序列数百分比

Number of Reads: 测序得到的总 read 数目

Q30 Bases in Barcode: 测得的 Cell barcode 序列质量大于 Q30 的序列占比

Q30 Bases in RNA Read: 测得的 R2 reads（单细胞文库的非接头序列）中质量

大于 Q30 的序列占比

Q30 Bases in UMI UMI: 测得的 UMI 序列中序列质量大于 Q30 的序列占比

Reads Mapped Confidently to Transcriptome: 比对到数据库已知参考基因的序列数百分比

Reads Mapped Confidently to Intronic Regions: 比对到数据库的内含子区域的序列数百分比

Reads Mapped Confidently to Intergenic Regions: 比对到数据库的基因间隔区域的序列数百分比

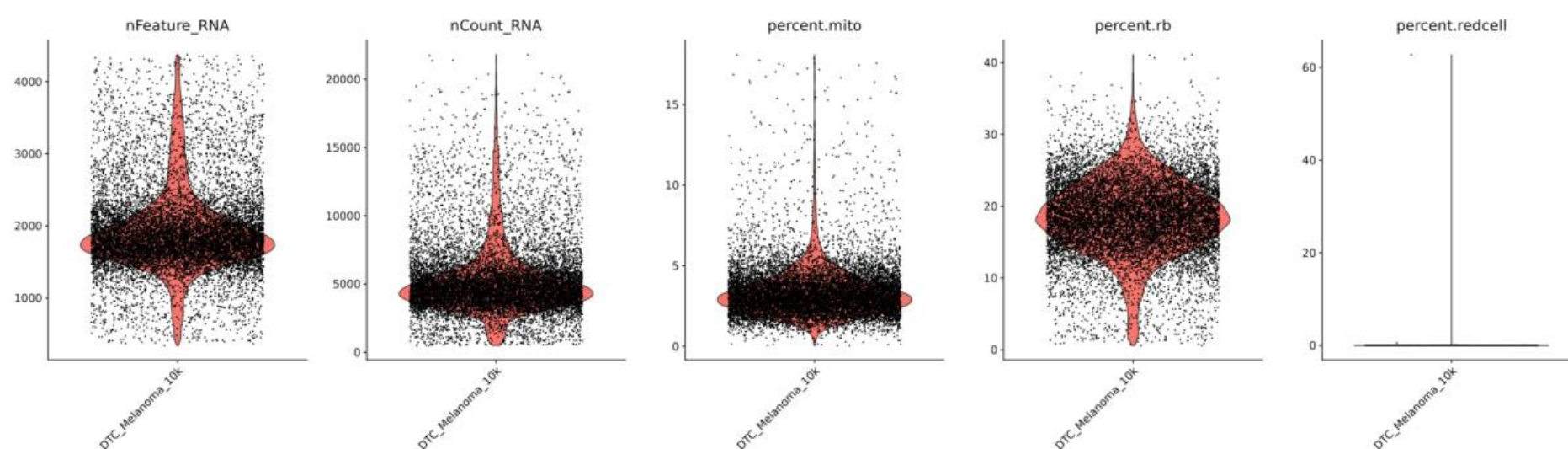
Total Genes Detected: 所有细胞检测到的基因总数

Median UMI Counts per Cell: 每个细胞中的中位转录本数

2.2.3 细胞质量评估与过滤

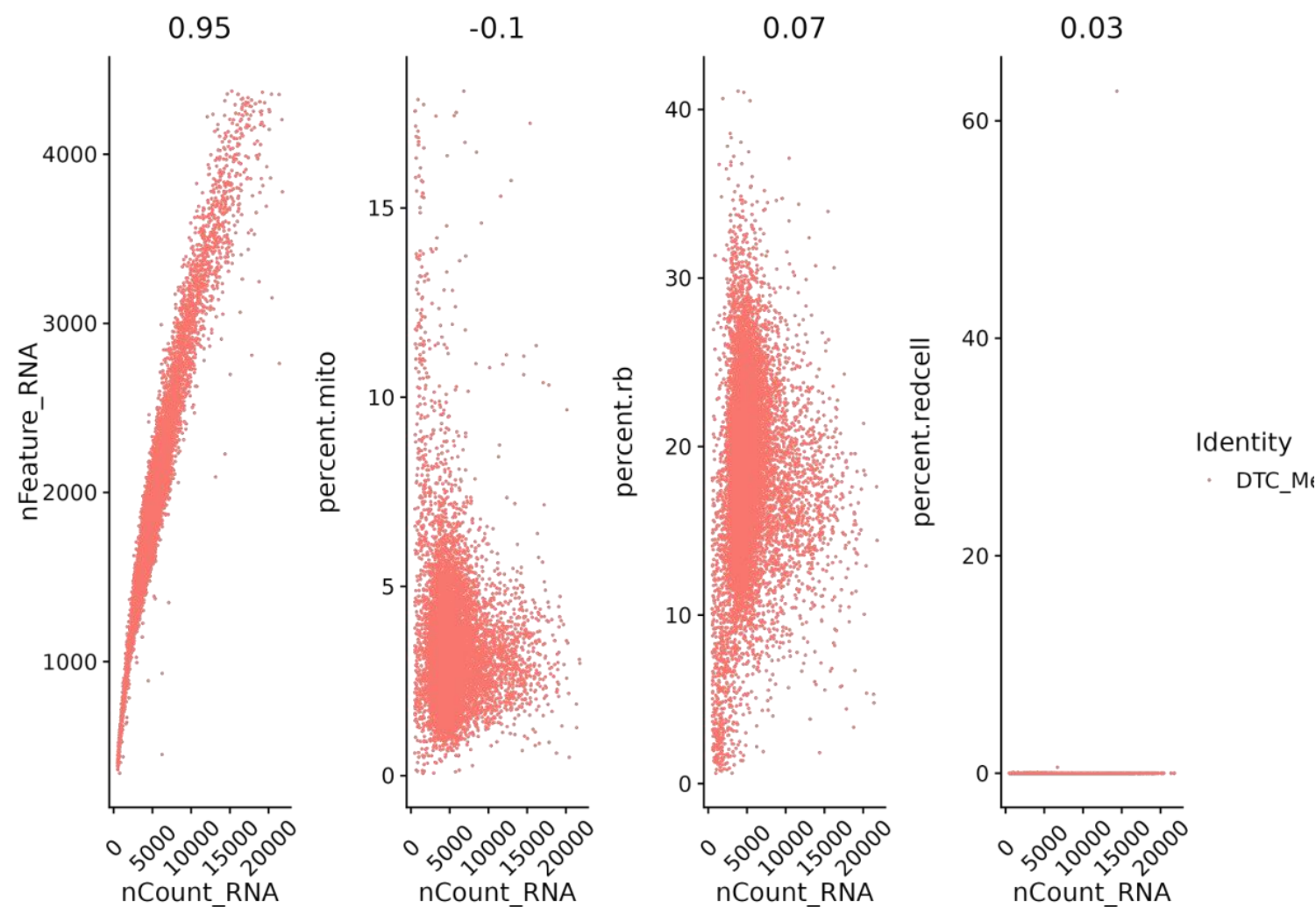
虽然单细胞转录组测序采用的是 barcode 以及 UMI 序列信息去标记细胞和转录本信息，但也有可能会出现带有这些标记信息的凝胶珠中没有细胞或者细胞数目不止一个的情况，且细胞死亡或者破碎时，线粒体基因比例会上升。理论上每个细胞中表达的基因数（nGene）、每个细胞的 UMI 数（nUMI）和每个细胞的线粒体基因的表达量会集中分布在某一区域内，根据它们的分布特征可以拟合分布模型，使用该模型找到其中的离域值。故而 Cell Ranger 生成表达矩阵之后，我们还需要在其初步质控的基础上，利用 Seurat 软件进一步质控过滤，根据每个细胞中表达的基因数(nGene)、UMI 数(nUMI)和线粒体基因表达量、核糖体基因表达量、双细胞过滤（DoubletFinder 软件）等信息，去除异常值，以保证后续分析结果的可靠性以及准确性，一般情况默认采用两倍标准差法。

细胞质量评估小提琴图



注：横坐标表示样本编号，纵坐标表示此细胞中的基因数、UMI 序列数以及来自线粒体/核糖体基因的 UMI 序列比例

细胞质量评估散点



注：横坐标表示细胞中的 count，纵坐标表示此细胞中的基因数、UMI 序列数以及来自线粒体/核糖体基因的 UMI 序列比例

细胞质量评估与过滤统计表

Sample	nCount_RNA	nFeature_RNA	Percent_mito	Percent_rb	Percent_redcell	Percent_double...	Num_Orig	Num_filterd
DTC_Melanoma...	22163	200-4376	18.552%	29.014%	62.734%	10.262%	14213	12077
SI_PBMC_10k	24808	200-5626	16.419%	40.973%	0.02%	6.812%	10784	9234

注：

- (1) Sample: 样本名称
- (2) nCount_RNA: 过滤后样本中单个细胞 UMI 数量最大值
- (3) nFeature_RNA: 过滤后样本中每个细胞检测到的基因数量的上下阈值
- (4) Percent.mito: 过滤后样本中细胞的线粒体基因表达占比最大值
- (5) Percent.rb: 过滤后样本中细胞的核糖体基因表达占比最大值
- (6) Percent.redcell: 过滤后样本中细胞的红细胞基因表达占比最大值
- (7) Percent.double_cell: 样本中判定为双细胞的细胞读数百分比
- (8) Percent.cp: 叶绿体基因表达占比(仅在植物中存在)
- (9) Num_Orig: 过滤前细胞数目
- (10) Num_filterd: 过滤后细胞数目

备注：点击【QC_Graph】的按钮，可切换查看过滤前和过滤后数据的细胞质量评估小提琴图和细胞质量评估散点图！

2.3 降维与聚类分析

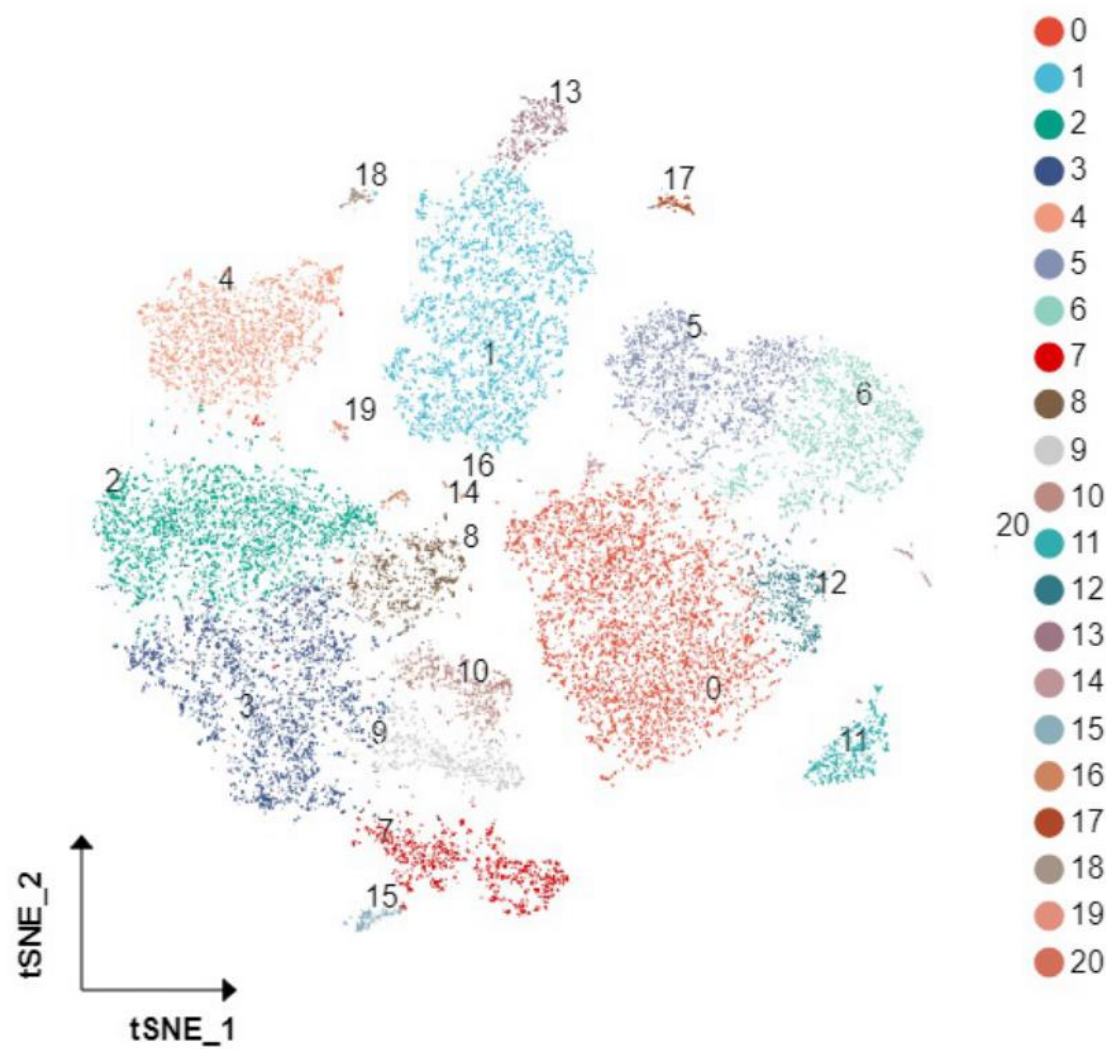
降维处理之后，需要进一步去除批次效应，之后会进行进一步的聚类分析。

为了达到聚类之后可视化目的，非线性降维应运而生。本模块采用两种非线性降维可视化方式：t-SNE 和 UMAP，同时也会展示 cluster 之间的相关性。

2.3.1 聚类分析与可视化

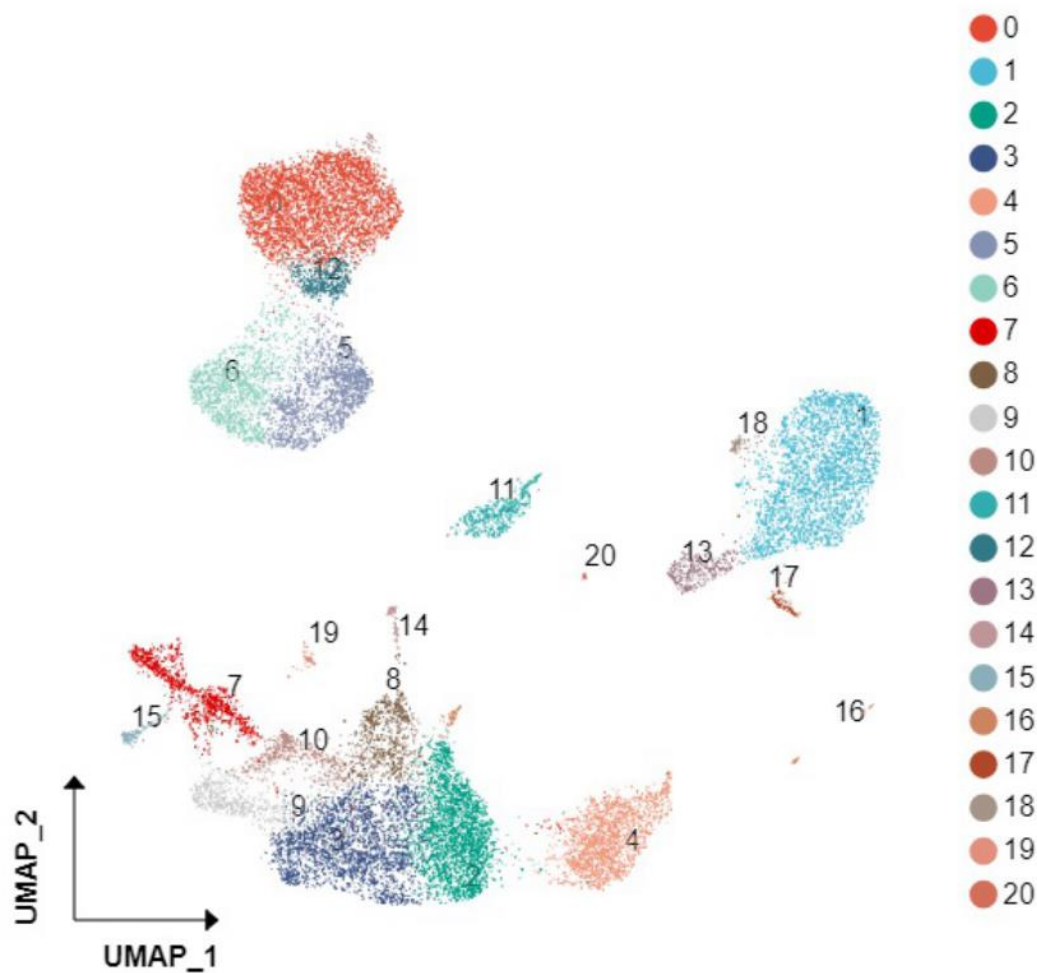
降维处理之后，需要进一步去除批次效应，之后会进行进一步的聚类分析。为了达到聚类之后可视化目的，非线性降维应运而生。本模块采用两种非线性降维可视化方式：t-SNE 和 UMAP，同时也会展示 cluster 之间的相关性。

降维可视化结果图



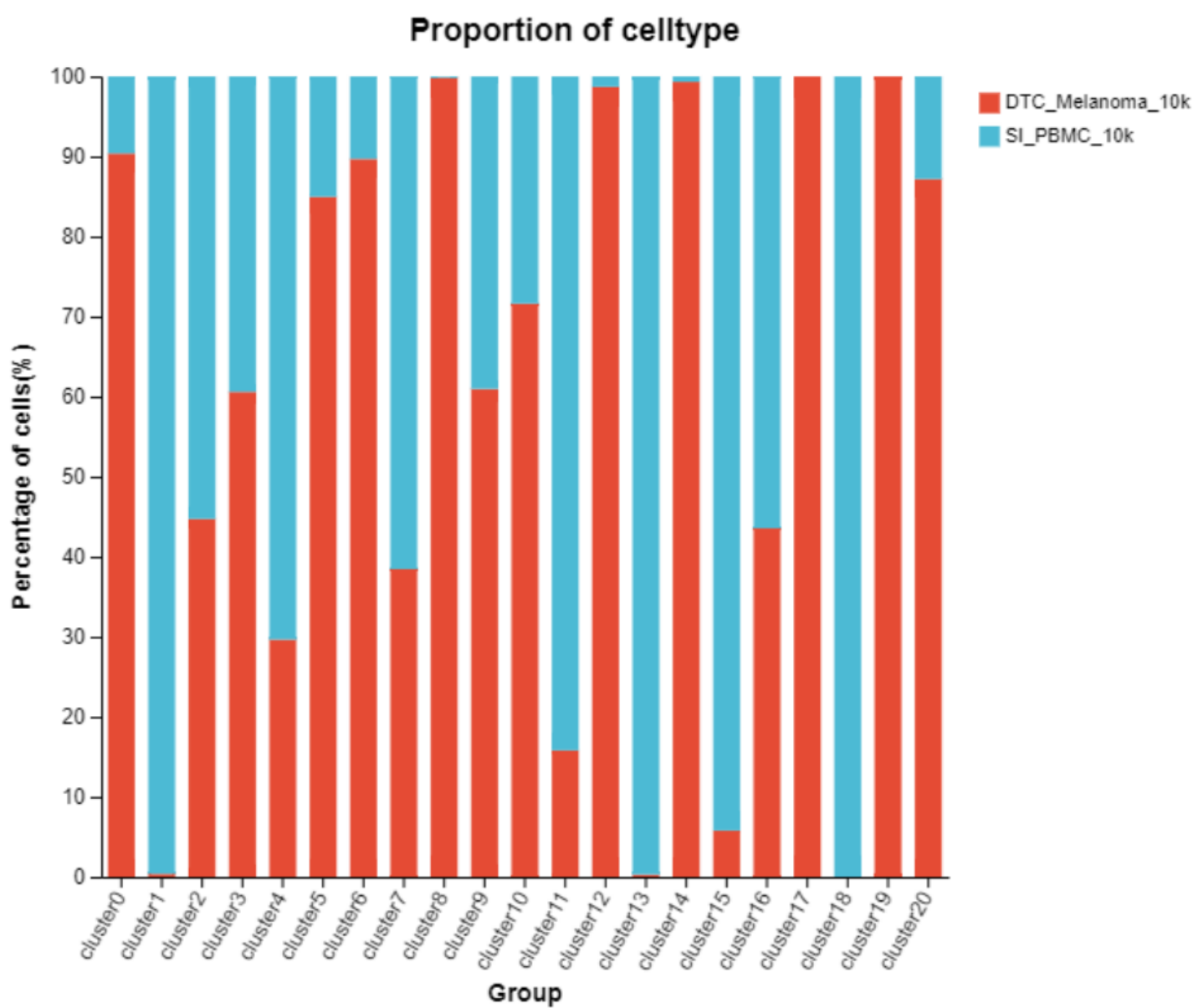
注：横纵坐标分别代表降维后的成分，图中的每个点代表一个细胞，距离相近的细胞认为是同一类型细胞，不同群细胞以不同的颜色区分。【可切换不同可视化形式进行展示】

降维可视化结果图



注：横纵坐标分别代表降维后的成分，图中的每个点代表一个细胞，距离相近的细胞认为是同一类型细胞，不同群细胞以不同的颜色区分。【可切换不同可视化形式进行展示】

各 cluster 统计结果堆积图



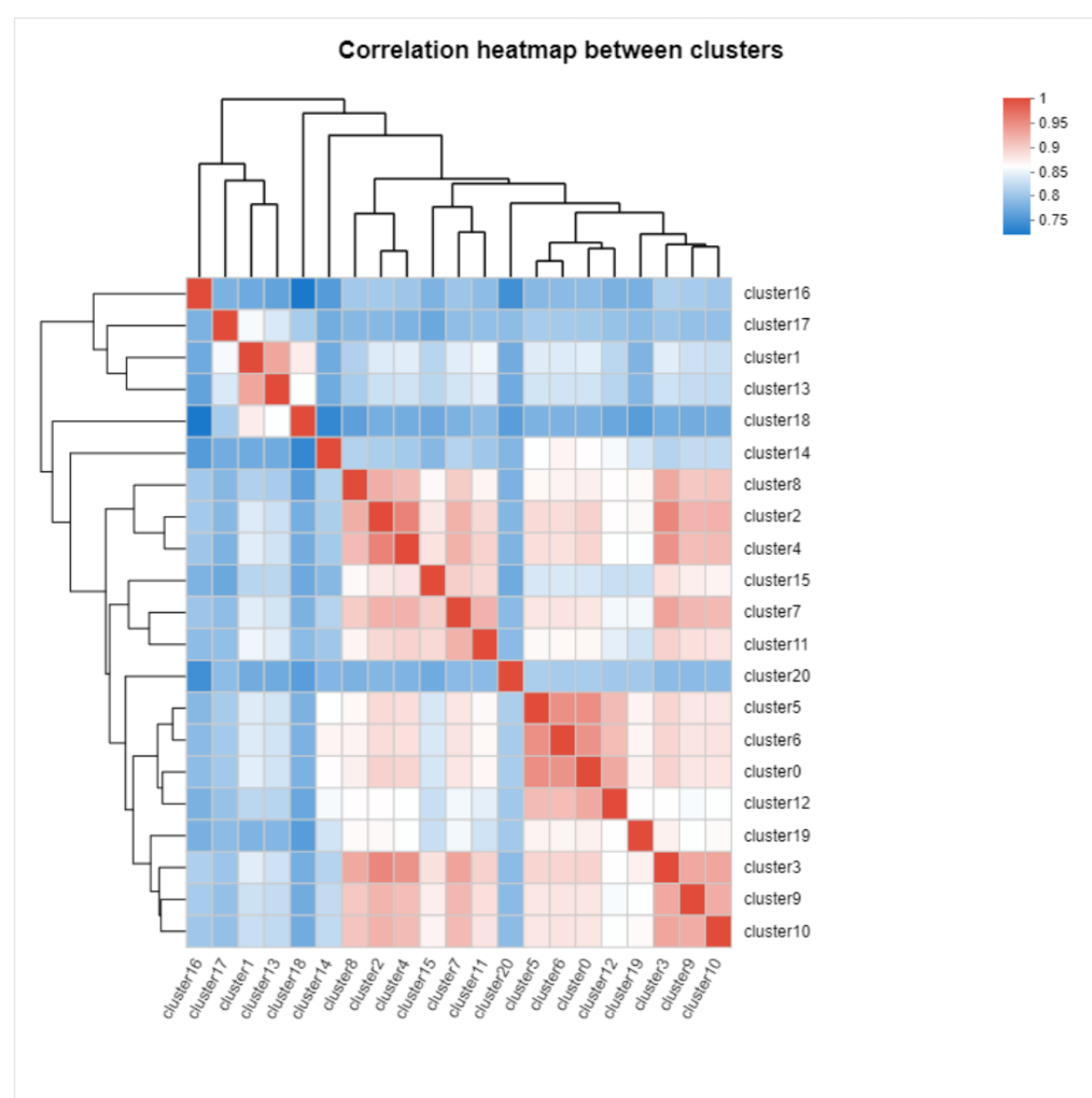
注：每个样本各 cluster 比例/数量堆积图。不同颜色代表不同的样本/组别 /cluster 等信息。【图表工具可以切换不同的展示形式】

各 cluster 细胞个数统计表

cluster	0	1	2	3	4	5	6	7	8
DTC_Melanom...	3647	11	1081	1443	536	1367	1245	353	686
SI_PBMC_10k	388	2910	1336	938	1271	242	143	565	1

注：第一列为样本名称；第一行表示按照当前结果运行出的 cluster 名称。数值表示细胞个数。

cluster 间相关性热图



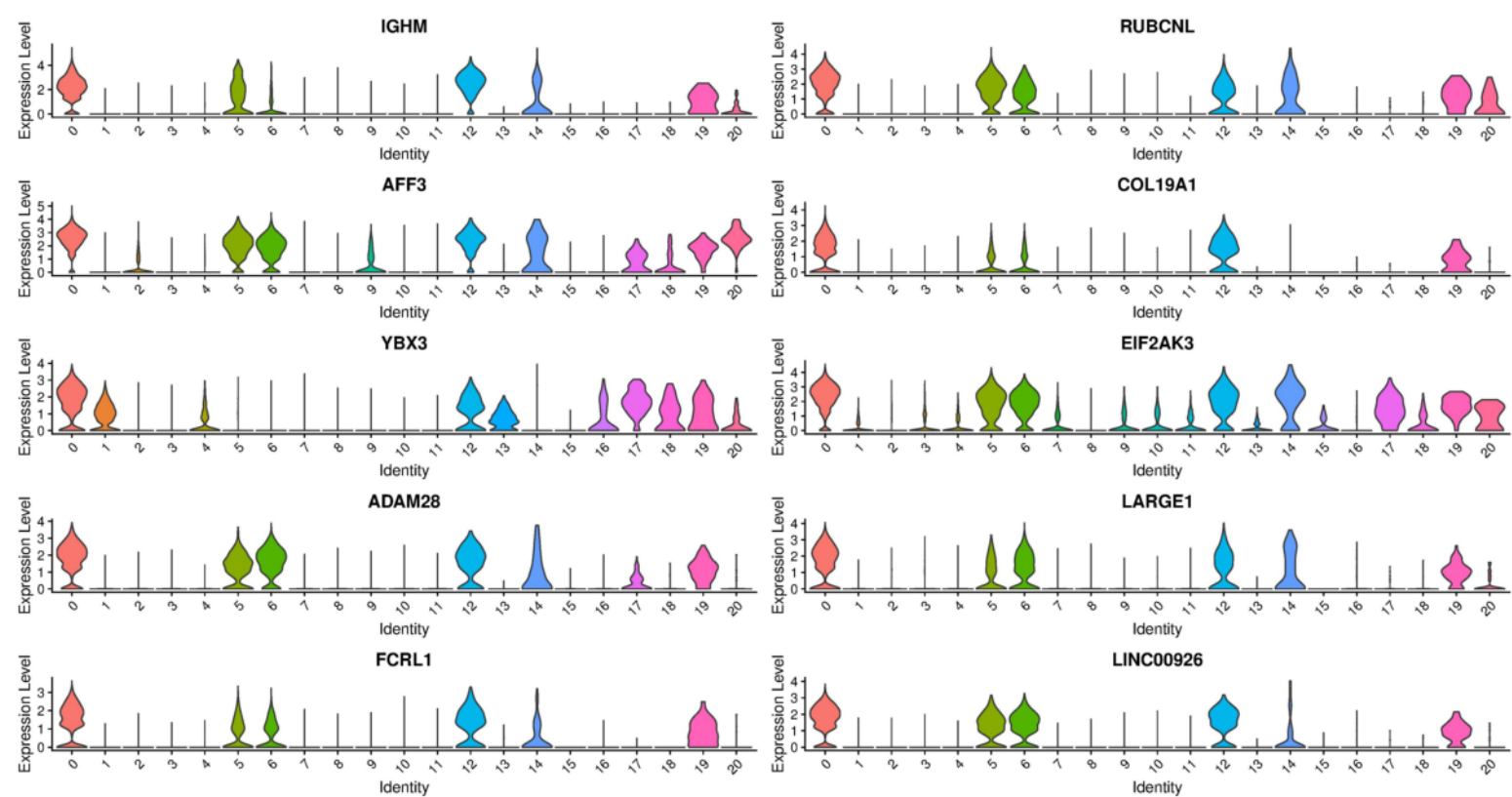
注：图中右侧和下侧为本项目的 cluster 名称，不同颜色的方块代表两个 cluster 的相关性高低，数值越大，表示两个 cluster 间的相关性越大，越相近，说明 cluster 间的相似性越高。

2.3.2 cluster 特征基因分析

通过上述聚类分析，已经将细胞区分成了不同的 cluster。之后 Seurat 可以进行一个亚群与剩余其他亚群之间的比较分析，以获得不同细胞亚群之间的差异基因。Seurat 计算差异的方法是 FindAllMarkers，通常采用默认参数运行软件，同时根据实际情况，如测序数据量，基因组情况对参数做适当的调整。利用差异分析算法寻找每个 cluster 的差异高表达基因，为进一步筛选 marker 基因提供依据。针对各 cluster 中差异高表达的基因以多种形式进行展示。

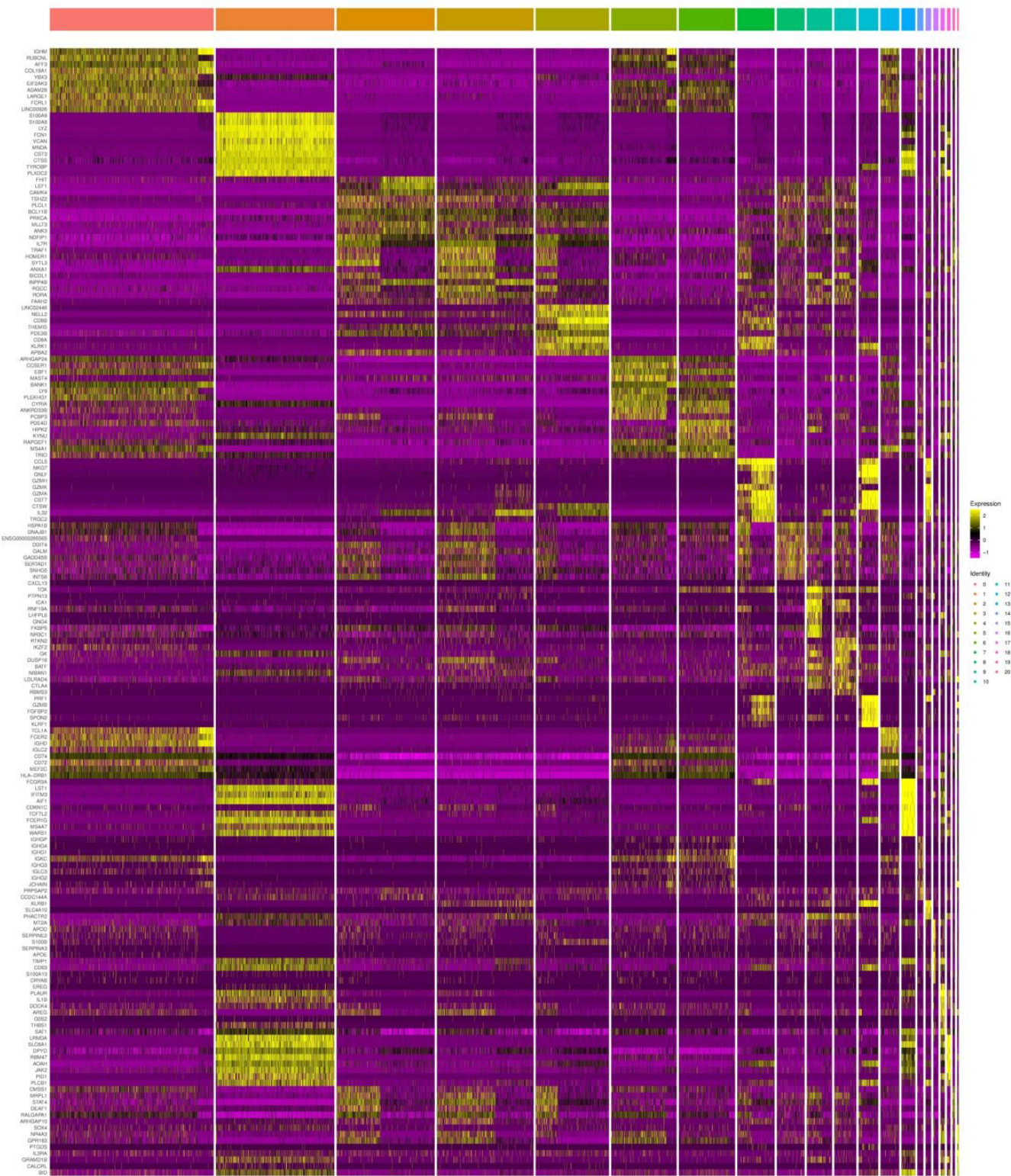
针对各个 cluster 中差异表达基因的展示方法有很多，包括小提琴图、t-SNE 图、UMAP 图以及聚类热图。

各 cluster-Marker-top 基因分别展示-小提琴图



注：横坐标为细胞群编号，纵坐标为标准化后的基因表达值，图中每个点表示每个细胞中差异基因的表达值。从左到右，从上到下，依次展示每个 cluster 中差异变化 top10 的基因。

cluster 差异分析热图



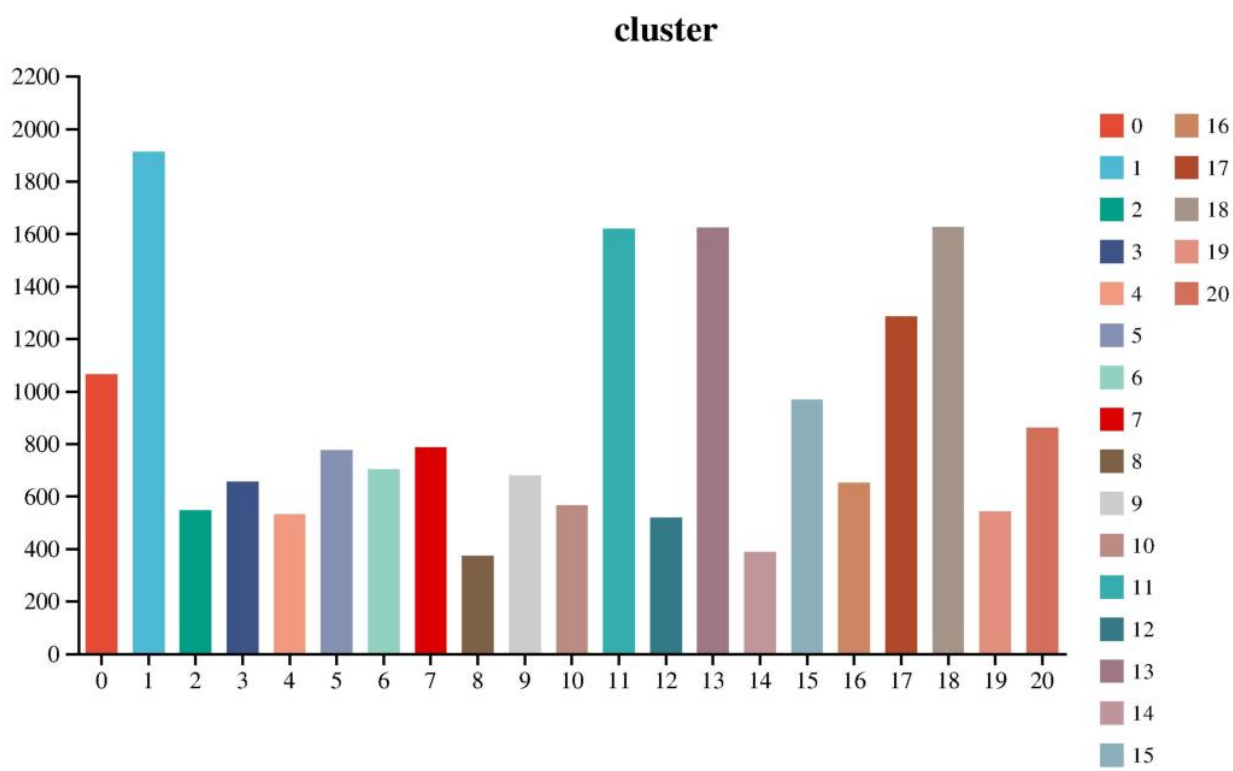
注：细胞群特异表达基因聚类热图，横坐标表示亚群，一列就是一个细胞，纵坐标表示基因，图中黄色表示高表达，紫色表示低表达。

表达量差异统计表

Gene id	Gene name	Gene descript...	cluster0	cluster1	cluster2	cluster3	cluster4	cluster5	cluster6
ENSG000000000...	FKBP4	FKBP prolyl is...	yes up	None	yes up	yes down	yes down	None	yes down
ENSG000000000...	SKAP2	src kinase ass...	yes up	yes up	yes up	yes down	yes down	yes up	yes down
ENSG000000000...	ACSM3	acyl-CoA synt...	yes up	yes up	None	yes down	yes down	yes down	yes down
ENSG000000000...	SPPL2B	signal peptide...	yes up	None	None	yes down	yes down	None	None
ENSG000000000...	ABCB4	ATP binding c...	yes up	yes up	None	yes down	yes down	None	yes down
ENSG000000000...	KMT2E	lysine methyl...	yes up	None	None	None	None	None	yes down
ENSG000000000...	ZFX	zinc finger pro...	yes up	yes up	None	yes down	yes down	None	yes down
ENSG000000000...	FARP2	FERM, ARH/R...	yes up	yes up	yes up	yes down	yes down	yes up	yes down
ENSG000000000...	PAFAH1B1	platelet activa...	yes up	yes up	None	yes down	None	None	yes down
ENSG000000000...	CD79B	CD79b molecu...	yes up	None	None	yes down	yes down	None	yes down

注：（1）Gene id：基因编号；（2）Gene name：基因名称；（3）Gene description：基因描述信息；（4）其余几列为该基因在 cluster 中是否显著差异，“yes”表示显著性差异，“up”为上调，“down”为下调；（5）Sum：该基因在不同 cluster 中显著差异表达的次数。

cluster 上调差异基因统计图



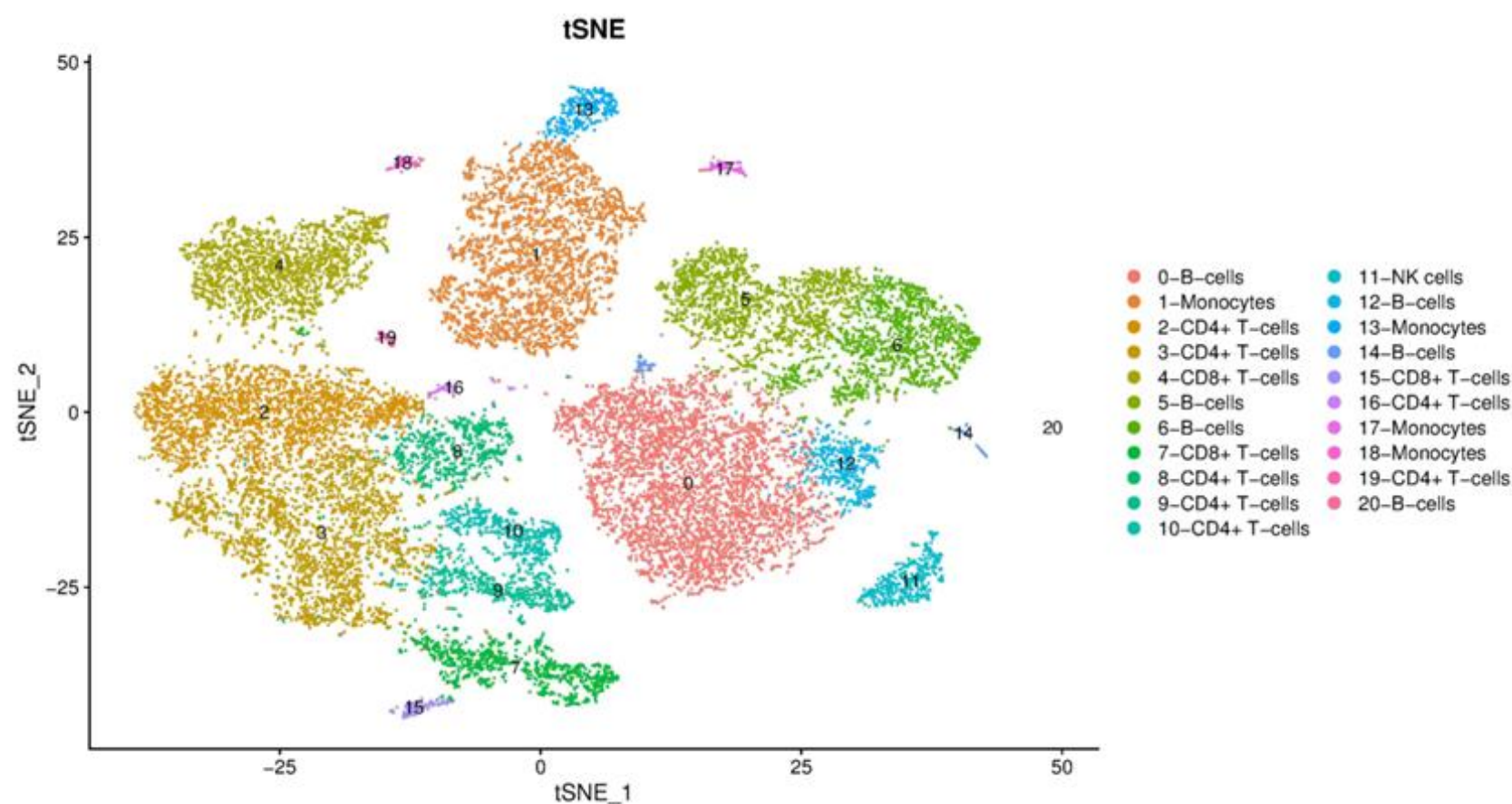
注：横坐标代表不同的 cluster，用不同颜色标识。纵坐标代表对应的上调基因数目。

2.4 细胞类型注释

2.4.1 数据集注释

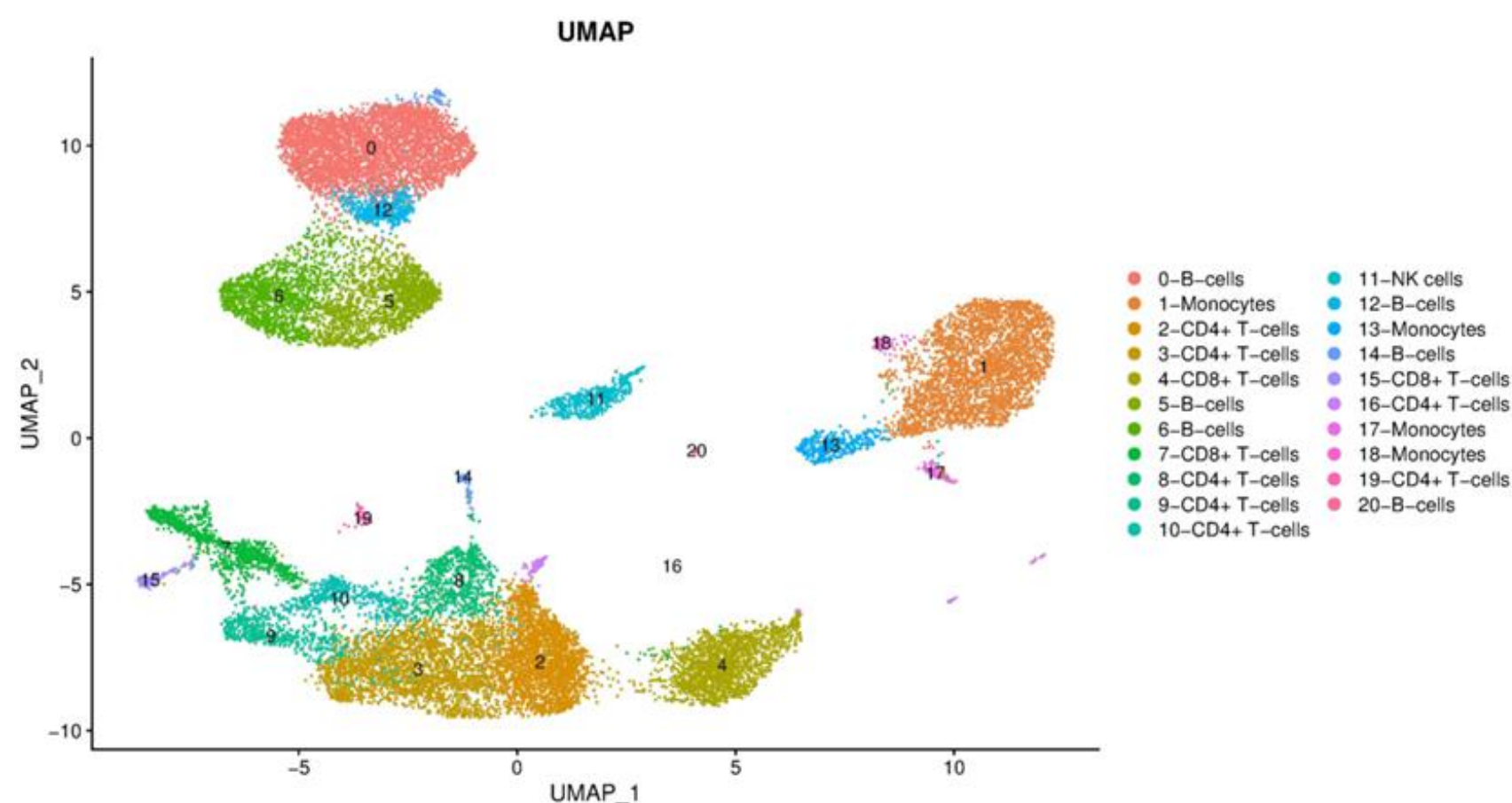
本模块采用软件 SingleR（Single-cell Recognition of cell types）包进行细胞类型注释。

按 celltype 展示 tSNE 可视化结果



注：横纵坐标分别代表降维后的成分，图中的每个点代表一个细胞，距离相近的细胞认为是同一类型细胞，不同群细胞以不同的颜色区分。

按 celltype 展示 UMAP 可视化结果



注：横纵坐标分别代表降维后的成分，图中的每个点代表一个细胞，距离相近的细胞认为是同一类型细胞，不同群细胞以不同的颜色区分。

细胞类型鉴定统计表-按 Cluster 展示

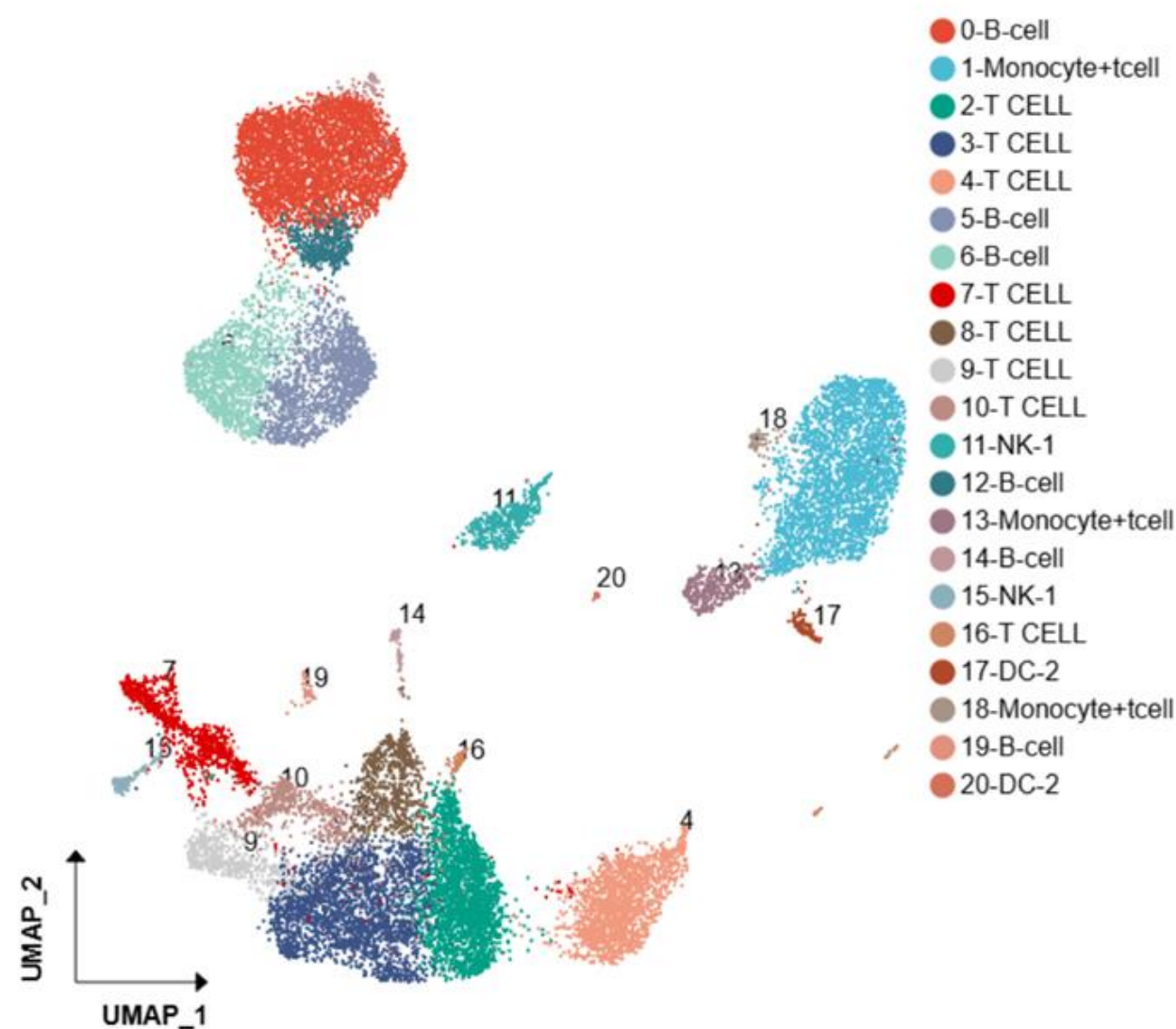
Cluster	Celltype	Cluster_Num
cluster0	B_cell	4035
cluster1	Macrophage_Monocyte	2921
cluster2	T_cell	2417
cluster3	T_cell	2381
cluster4	T_cell	1807
cluster5	B_cell	1609
cluster6	B_cell	1388
cluster7	T_cell	918
cluster8	T_cell	687
cluster9	T_cell	620

注：（1）Cluster：表示某个 cluster，为聚类得到的细胞群；（2）Celltype：注释得到的细胞类型；（3）Cluster_Num：该 Cluster 包含的细胞数目。【如需要修改本次注释结果，可在右侧进行手动编辑】

2.4.2 Marker 软件注释

SCINA（半监督类别识别和分配）等软件是一种半监督模型，可以将输入的 Marker 基因列表，利用软件进行细胞类型判定。其中，SCINA 软件利用期望最大化 (EM) 算法利用先前输入的 Marker 基因。

按 celltype 展示 UMAP 可视化结果

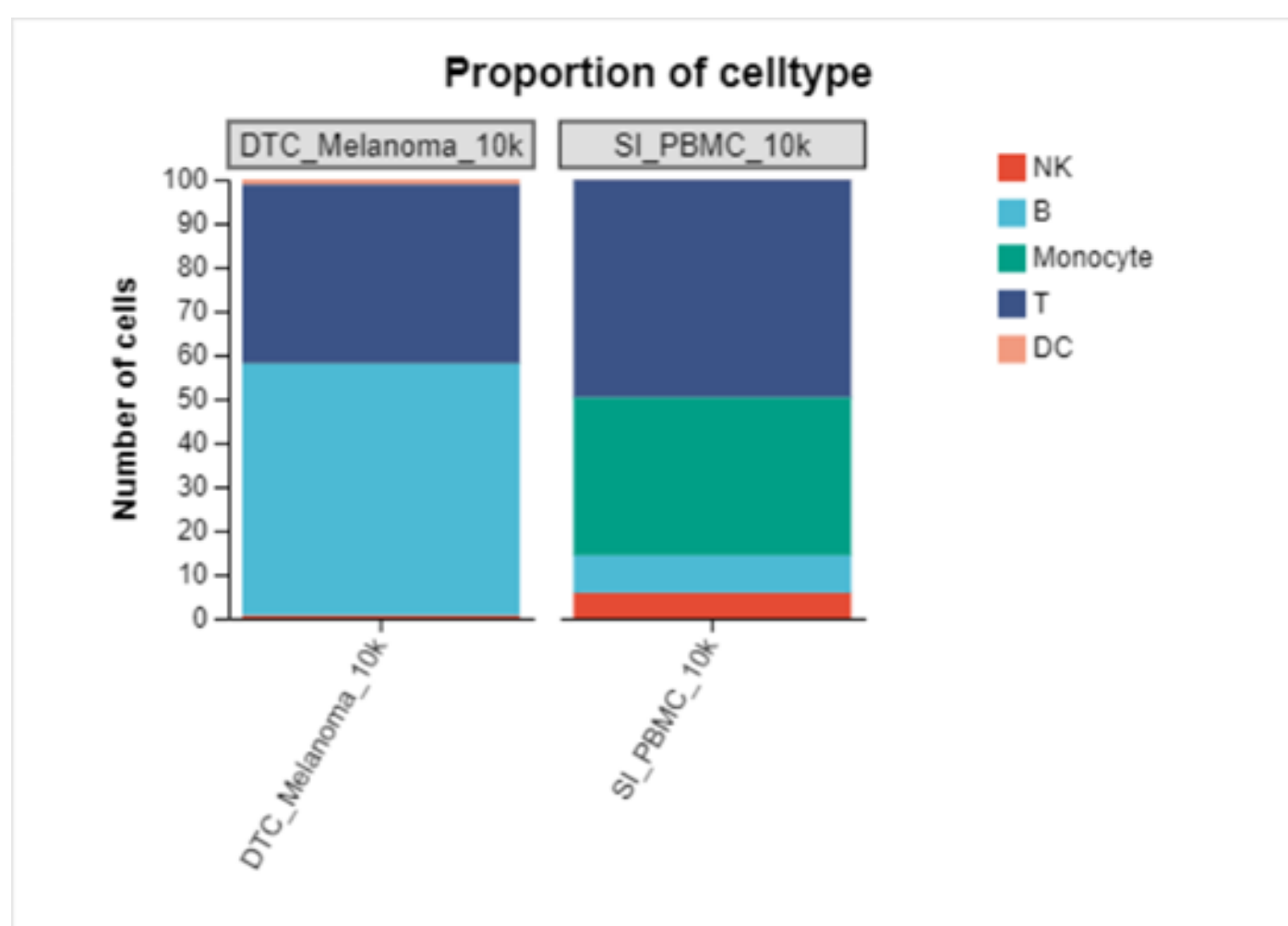


注：横纵坐标分别代表降维后的成分，图中的每个点代表一个细胞，距离相近的细胞认为是同一类型细胞，不同群细胞以不同的颜色区分。

2.4.3 细胞类型可视化

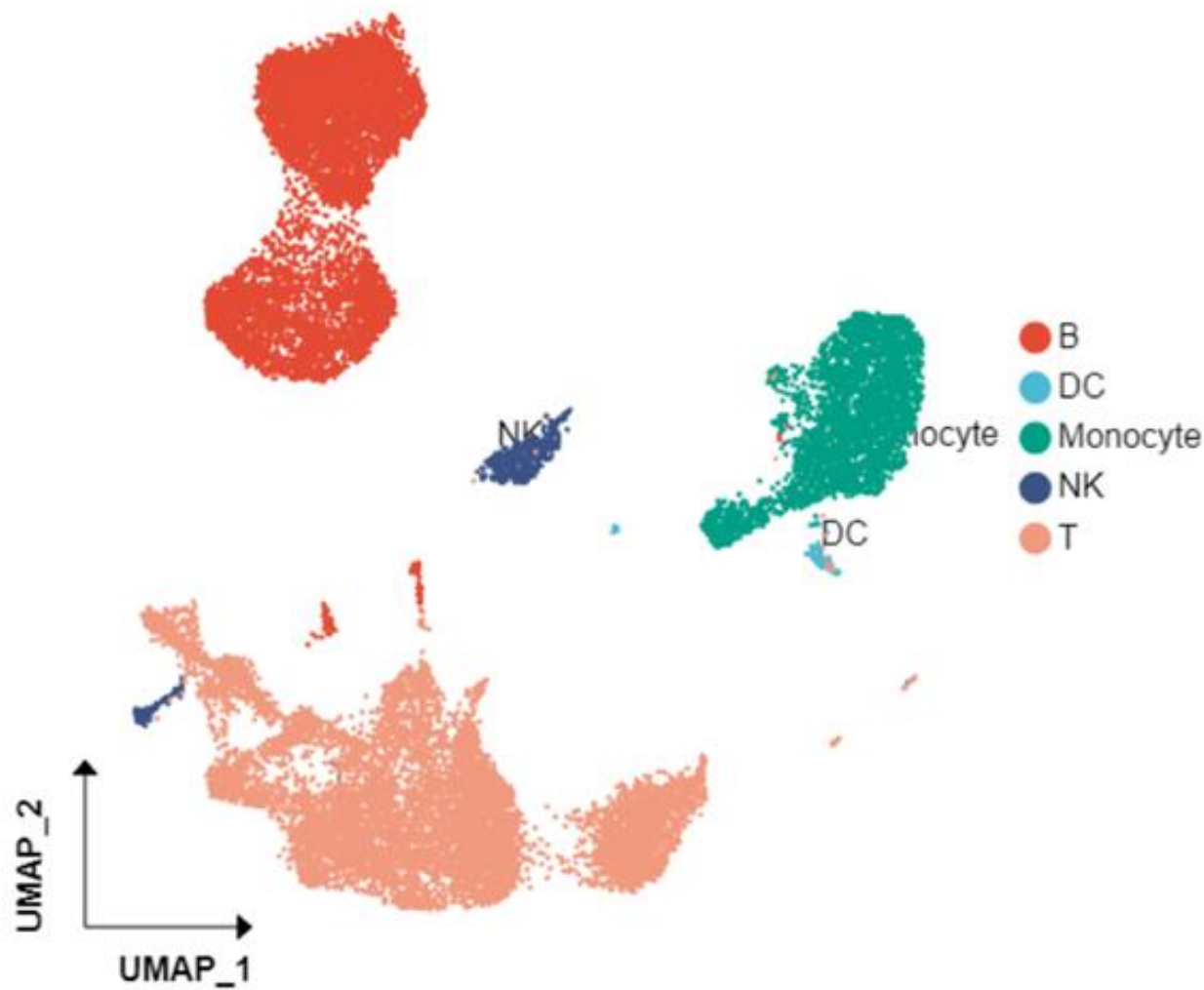
本模块，可以将不论哪种手段获得的细胞类型，进行可视化展示，包括人工判定后的可视化展示结果和人工判定后的细胞类型统计与可视化展示。

样本-celltype 比较分析统计图-堆积图



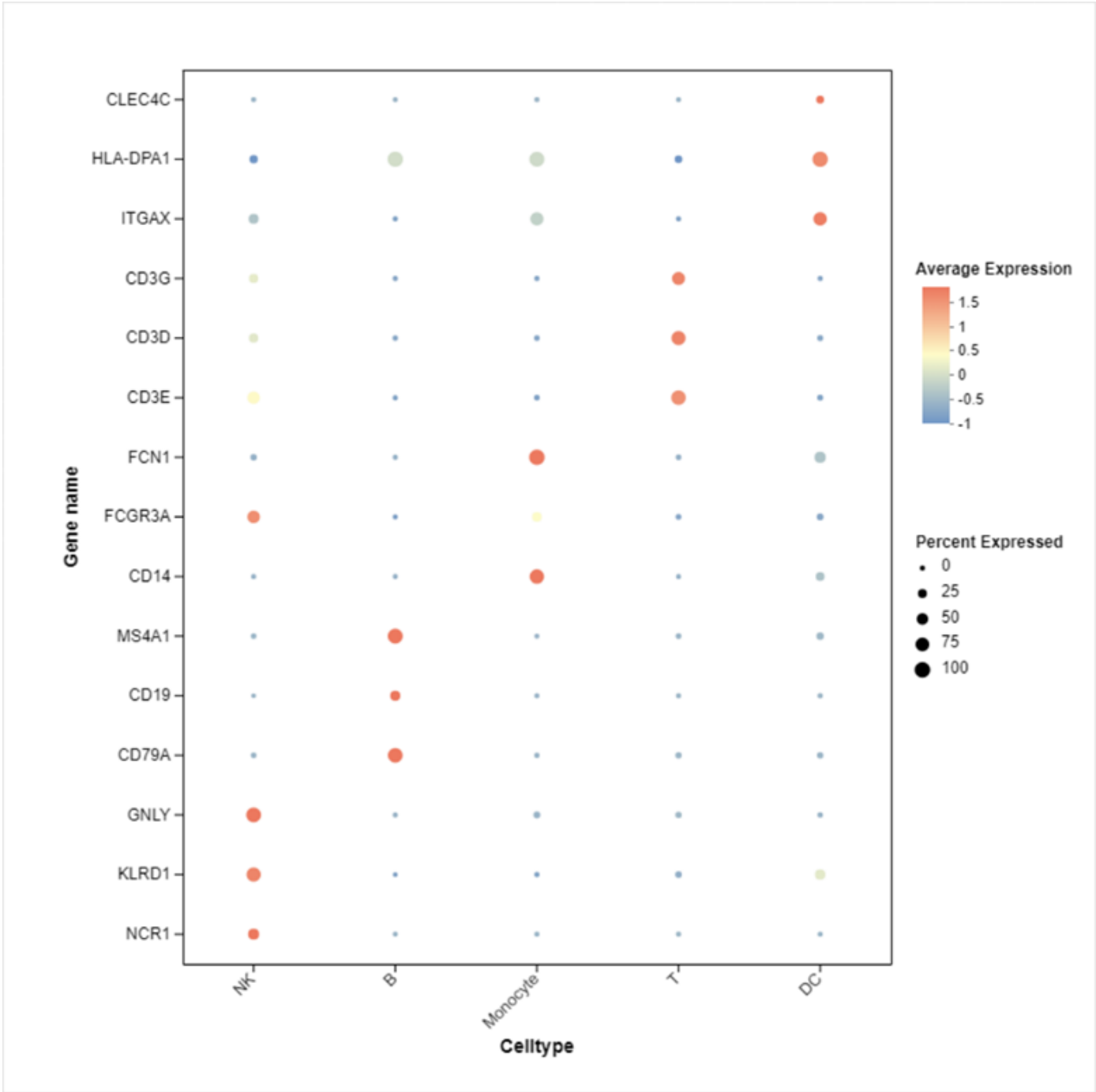
注：每个样本细胞类型比例柱状堆叠图。不同颜色代表不同的样本/组别/celltype 等信息。

按 celltype 展示 UMAP 可视化结果



注：横纵坐标分别代表降维后的成分，图中的每个点代表一个细胞，距离相近的细胞认为是同一类型细胞，不同群细胞以不同的颜色区分。

气泡图



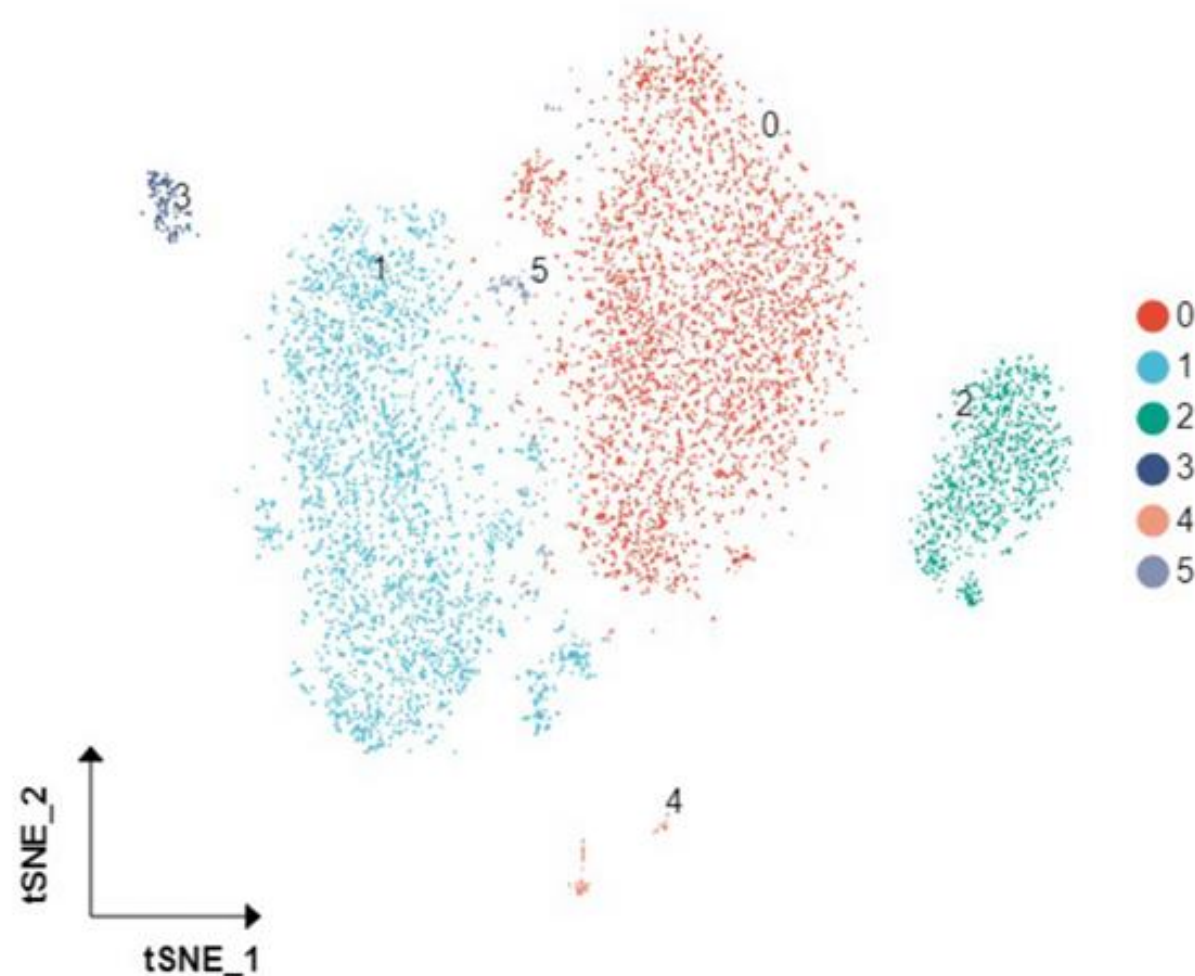
注：横坐标表示差异基因，纵坐标表示细胞类型【图表工具可切换按照样本/组别进行绘图】，图中颜色越深表达量越高，气泡越大表达占比越高。

2.5 细胞亚群细分

2.5.1 细胞亚群降维聚类分析

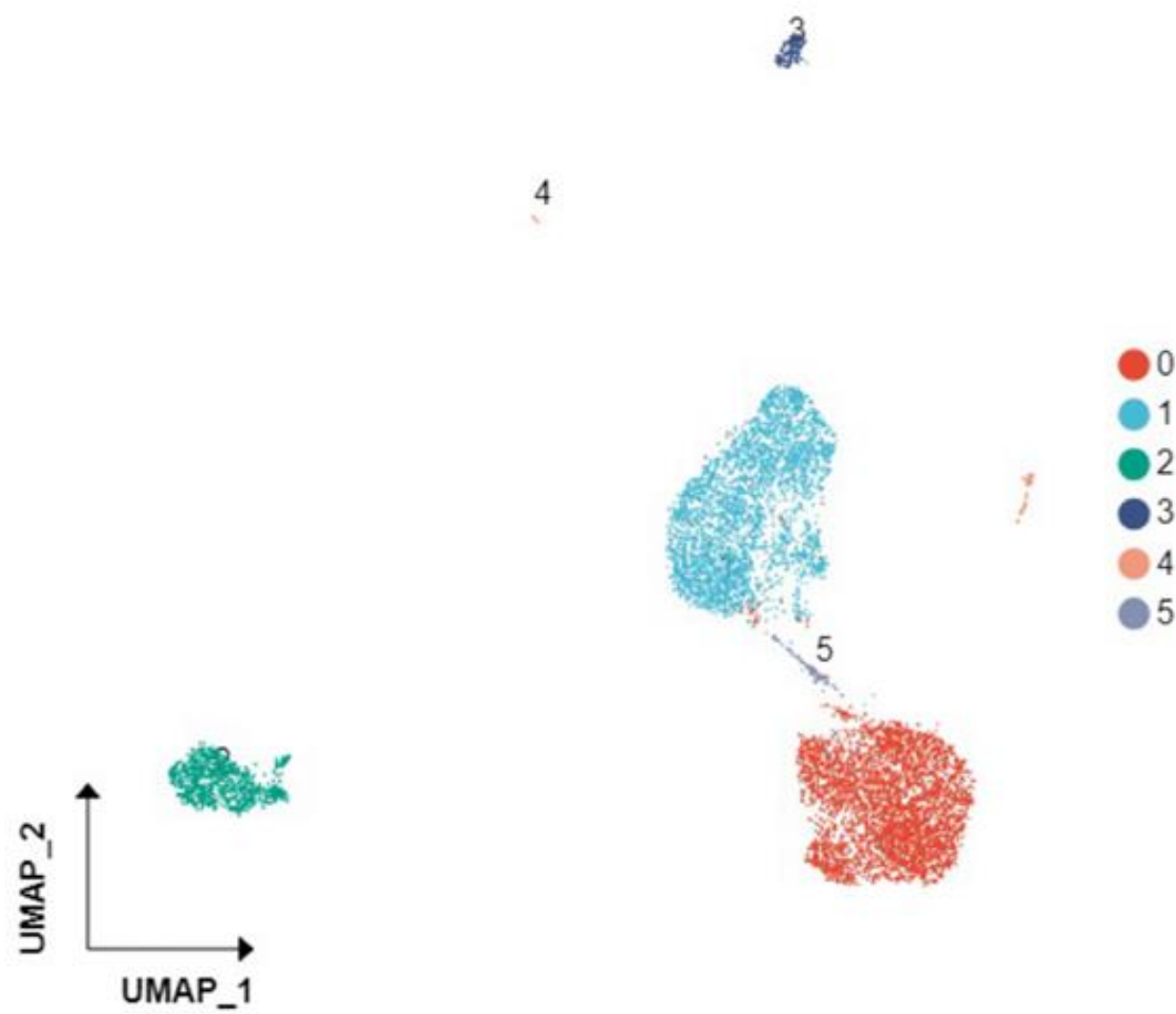
本模块的内容是对感兴趣的细胞 **cluster** 进行再分群的分析，无需重新上传数据，只需要选择对应的细胞亚群，就可以完成感兴趣的 **cluster** 重新聚类。所谓重新聚类，其本质，与整体样本的降维聚类分析原理完全一致。

降维可视化结果图



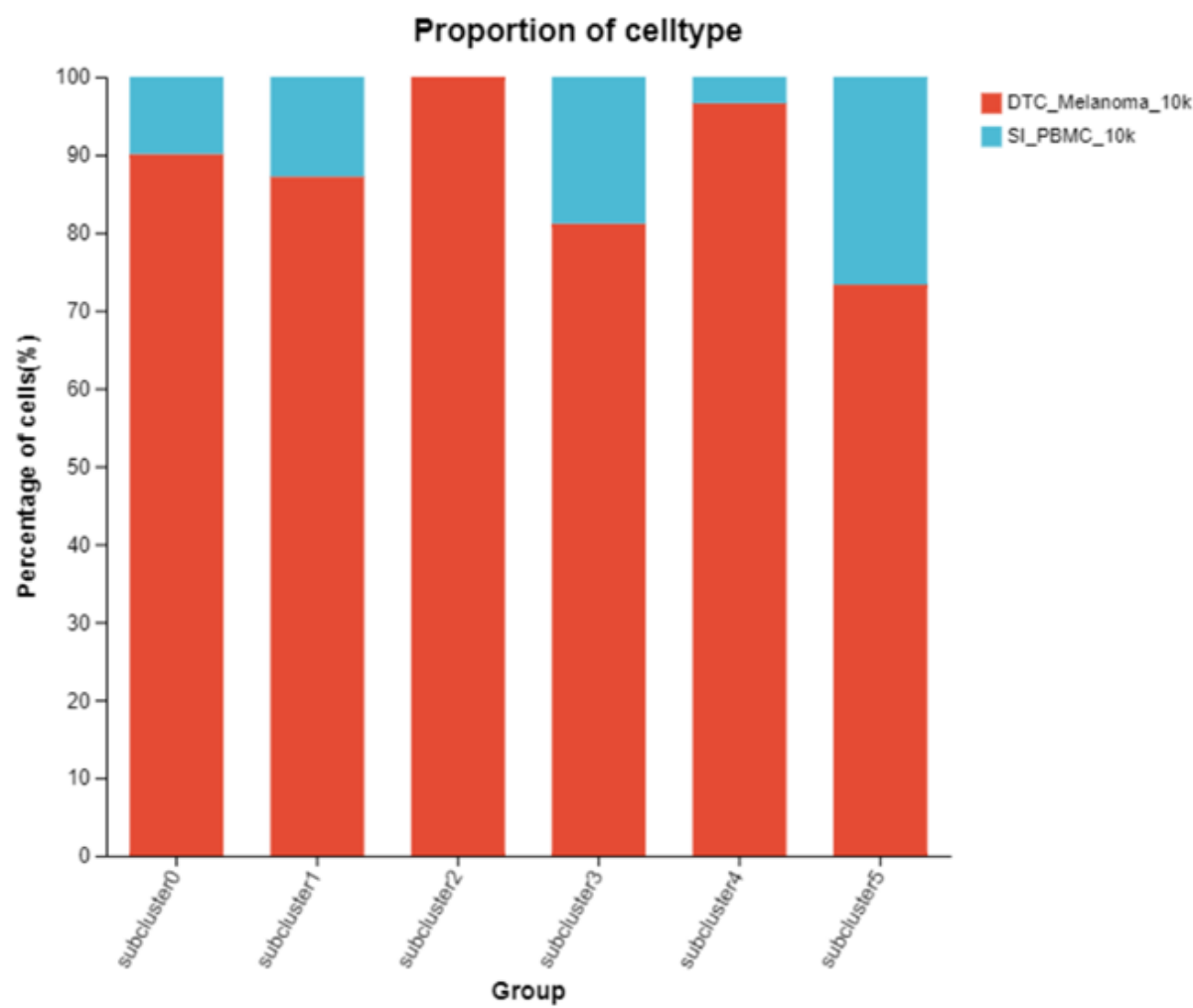
注：横纵坐标分别代表降维后的成分，图中的每个点代表一个细胞，距离相近的细胞认为是同一类型细胞，不同群细胞以不同的颜色区分。【可切换不同可视化形式进行展示】

降维可视化结果图



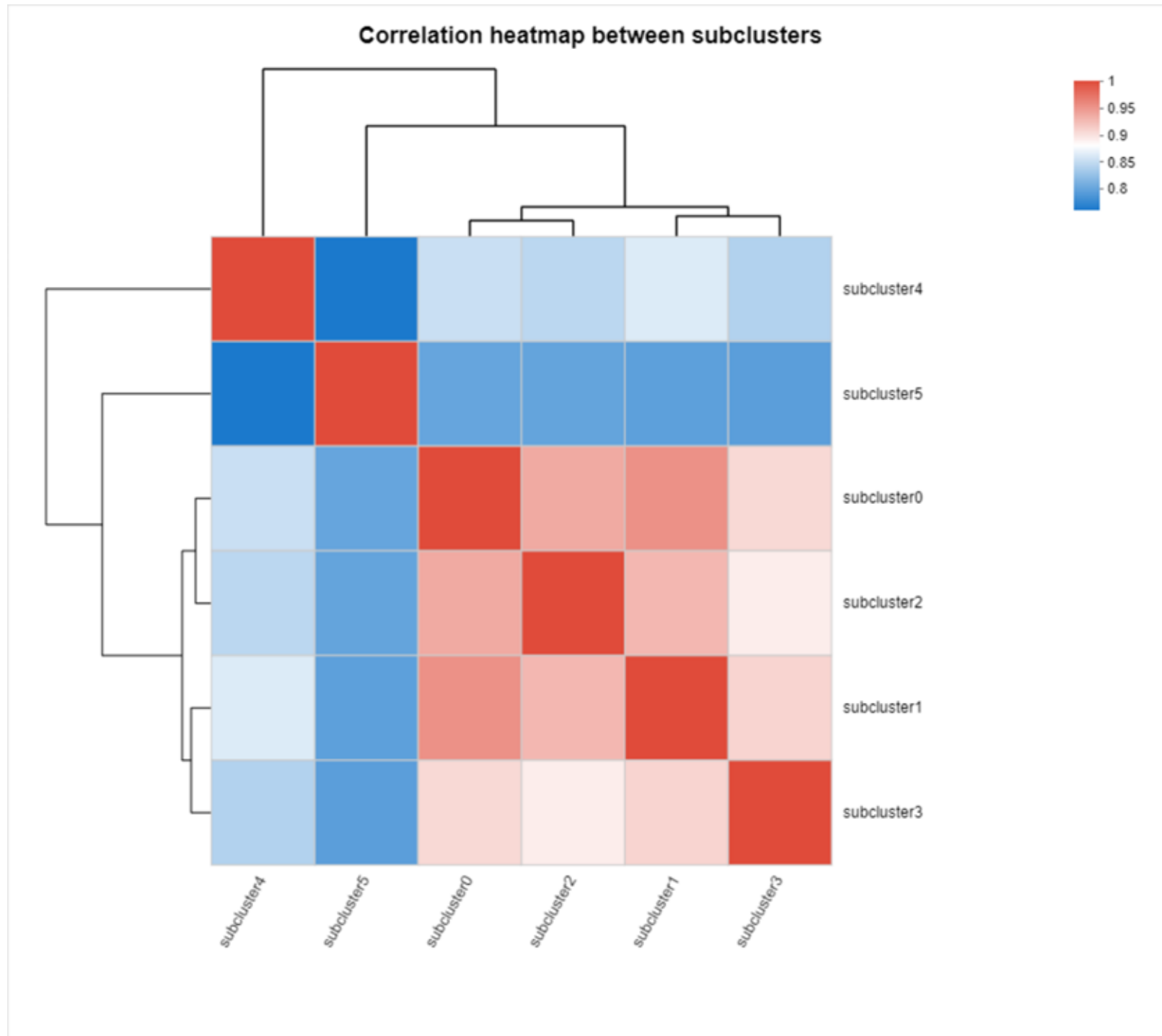
注：横纵坐标分别代表降维后的成分，图中的每个点代表一个细胞，距离相近的细胞认为是同一类型细胞，不同群细胞以不同的颜色区分。【可切换不同可视化形式进行展示】

各 subcluster 统计结果堆积图



注：每个样本各 subcluster 比例/数量堆积图。不同颜色代表不同的样本/组别/subcluster 等信息。【图表工具可以切换不同的展示形式】

subcluster 间相关性热图



注：图中右侧和下侧为本结果的 subcluster 名称，不同颜色的方块代表两个 subcluster 的相关性高低，数值越大，表示两个 subcluster 间的相关性越大，越相近，说明 subcluster 间的相似性越高。

2.5.2 subcluster 特征基因分析

通过上述聚类分析，已经将细胞区分成了不同的 subcluster。之后 Seurat 可以进行一个亚群与剩余其他亚群之间的比较分析，以获得不同细胞亚群之间的差异基因。Seurat 计算差异的方法是 FindAllMarkers，通常采用默认参数运行软件，同时根据实际情况，如测序数据量，基因组情况对参数做适当的调整。利用差异分析算法寻找每个 subcluster 的差异高表达基因，为进一步筛选 marker 基因提供依据。针对各 subcluster 中差异高表达的基因以多种形式进行展示。

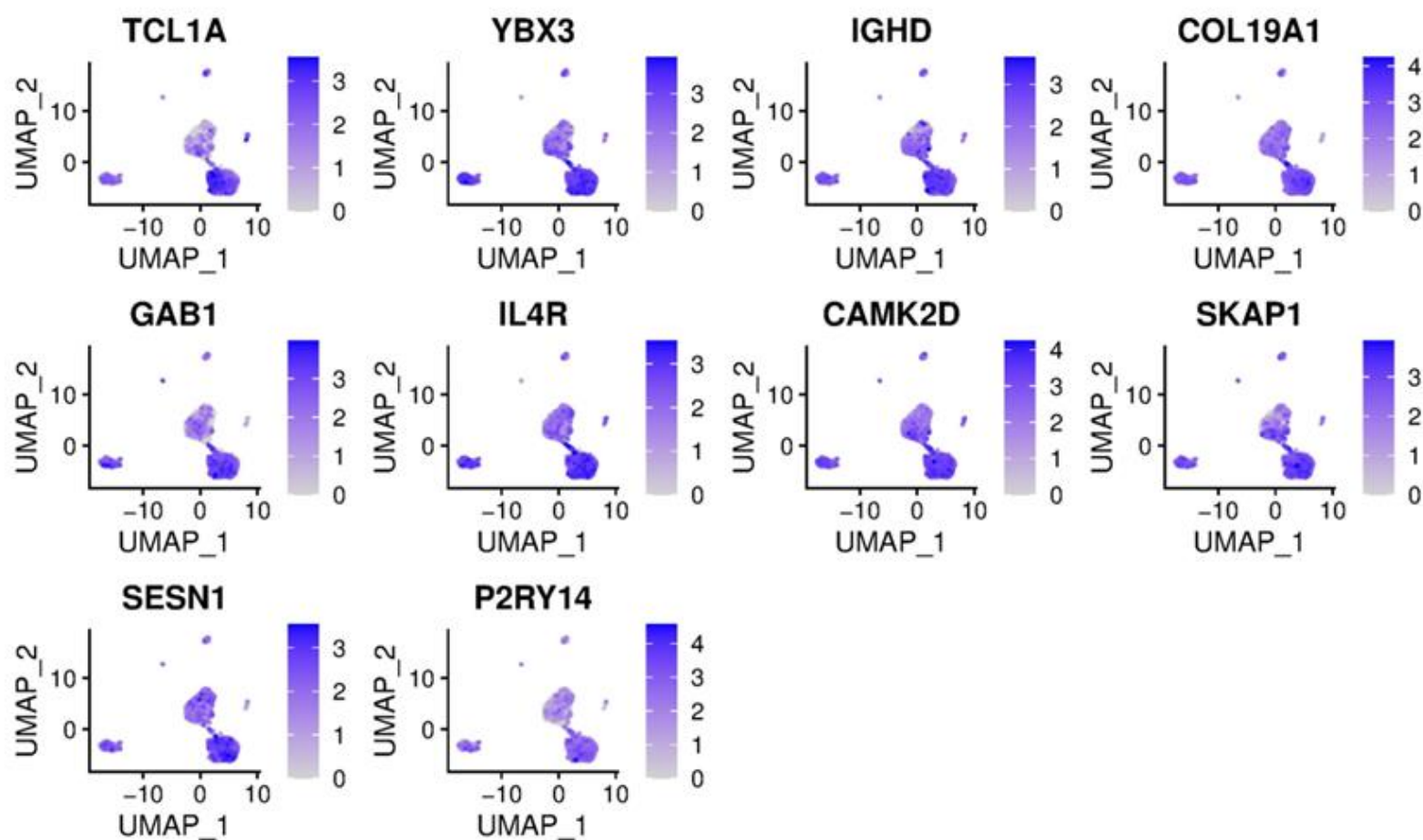
针对各个 subcluster 中差异表达基因的展示方法有很多，包括小提琴图、t-SNE 图、UMAP 图以及聚类热图。

subcluster 表达量差异统计表

Gene id	Gene name	Gene descript...	subcluster0	subcluster1	subcluster2	subcluster3	subcluster4	subcluster5	sum
ENSG000000000...	CD99	CD99 molecu...	yes down	yes up	None	None	None	None	2
ENSG000000000...	SNX11	sorting nexin ...	None	None	None	None	yes up	None	1
ENSG000000000...	CFLAR	CASP8 and FA...	yes down	yes up	yes down	None	None	None	3
ENSG000000000...	CD38	CD38 molecu...	None	yes down	None	None	yes up	None	2
ENSG000000000...	RBM6	RNA binding m...	None	None	None	None	yes up	None	1
ENSG000000000...	CDC27	cell division cy...	None	None	None	None	yes up	None	1
ENSG000000000...	UPF1	UPF1 RNA hel...	None	None	None	None	yes up	None	1
ENSG000000000...	SLC25A5	solute carrier...	None	None	None	None	None	yes down	1
ENSG000000000...	MCUB	mitochondria...	None	None	None	yes up	None	None	1
ENSG000000000...	ACSM3	acyl-CoA synt...	yes up	yes down	None	None	yes down	None	3

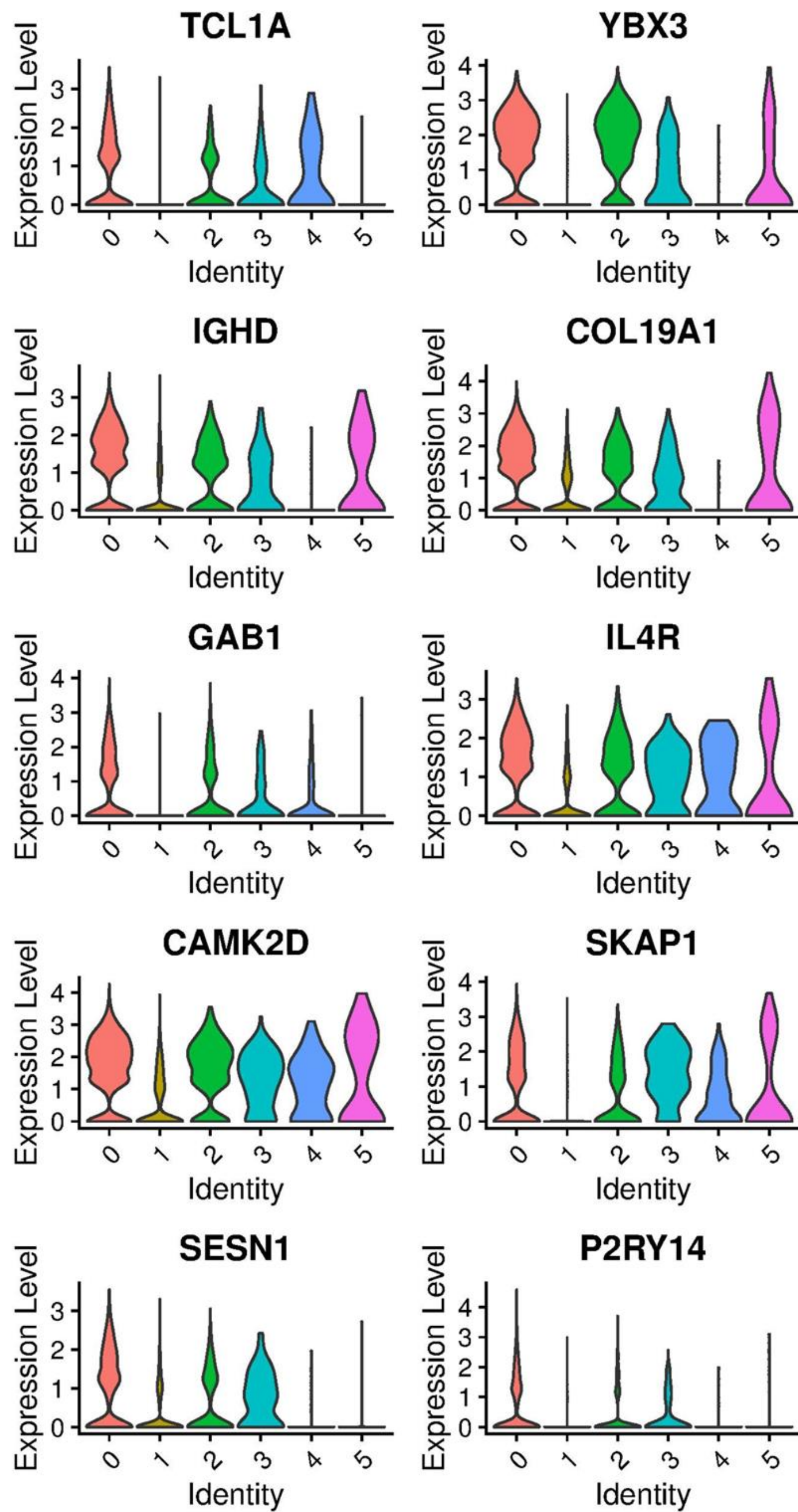
注：（1）Gene id：基因编号；（2）Gene name：基因名称；（3）Gene description：基因描述信息；（4）其余几列为该基因在 subcluster 中是否显著差异，“yes”表示显著性差异，“up”为上调，“down”为下调；（5）Sum：该基因在不同 subcluster 中显著差异表达的次数。

各 subcluster-Marker-top 基因分别展示 UAMP 图



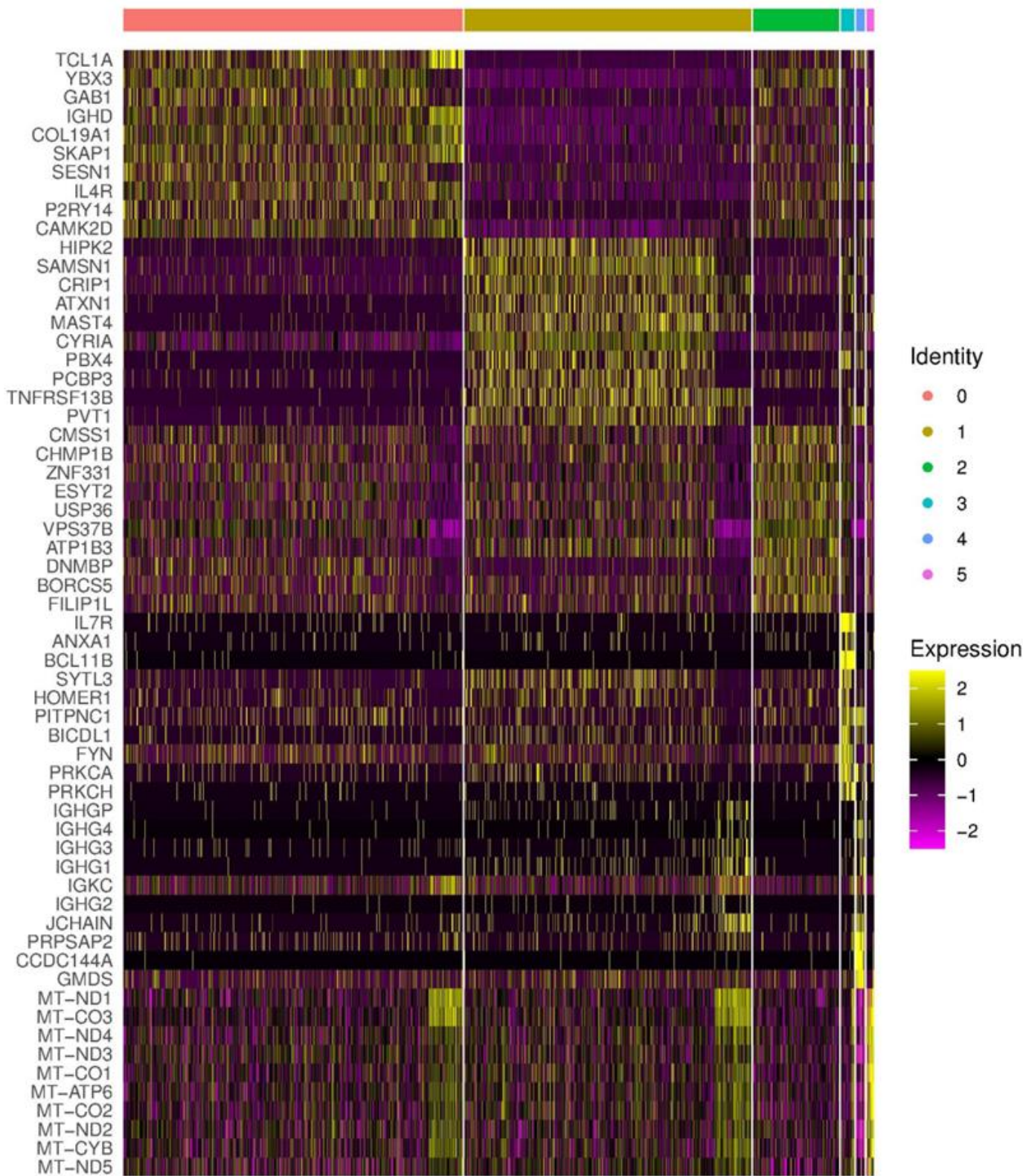
注：每个点代表每个细胞，颜色深浅表示基因表达量的高低。

各 subcluster-Marker-top10 基因展示-violin 图



注：横坐标为细胞群编号，纵坐标为标准化后的基因表达值，图中每个点表示每个细胞中差异基因的表达值。

各 subcluster-Marker-top 基因整体展示-热图



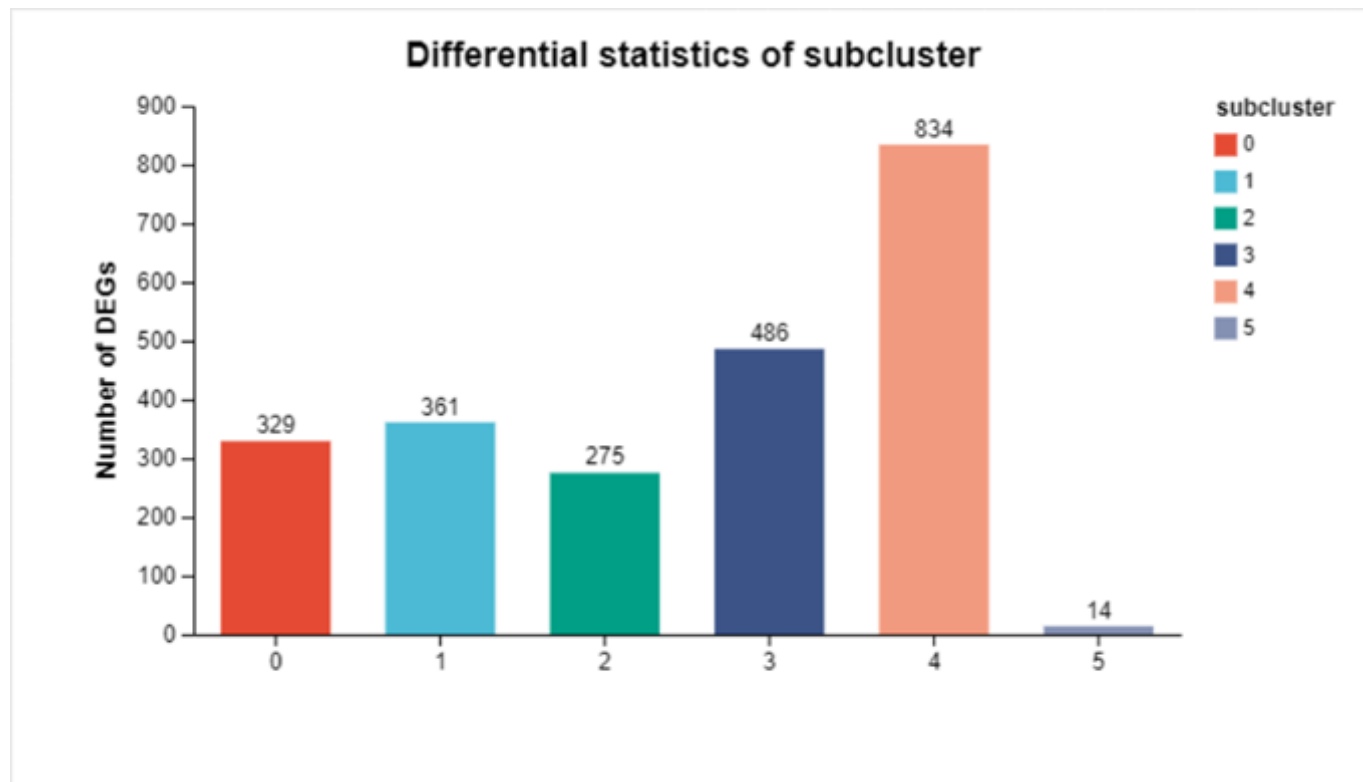
注：细胞群特异表达基因聚类热图，纵坐标表示差异基因，横坐标表示亚群，
一列就是一个细胞，图注为细胞亚群，图中黄色表示高表达，紫色表示低表达。

subcluster 表达量差异详情表

Subclusters	Gene id	Gene name	Gene descript...	avg_Log2FC	Pvalue	Padjust	pct.1(%)	pct.2(%)	Significant
0	ENSG00000010...	TCL1A	TCL1 family A...	1.4774278270...	0	0	57.7	15.4	yes
0	ENSG00000006...	YBX3	Y-box binding...	1.3082212326...	0	0	82.4	32.1	yes
0	ENSG00000010...	GAB1	GRB2 associat...	1.2928118679	3.3444708711...	9.8912726013...	55.9	18.5	yes
0	ENSG00000021...	IGHD	immunoglobu...	1.2525164869...	0	0	76.9	36.7	yes
0	ENSG00000008...	COL19A1	collagen type...	1.2076518443...	0	0	79.1	45.8	yes
0	ENSG00000014...	SKAP1	src kinase ass...	1.1651482793...	2.0895718215...	6.1799086622...	59.9	25.7	yes
0	ENSG00000008...	SESN1	sestrin 1 [Sour...	1.1047121184...	5.7124546045...	1.6894584492...	60.9	31.7	yes
0	ENSG00000007...	IL4R	interleukin 4 r...	1.1038735686...	6.4635636904...	1.9115989614...	73.3	40.1	yes
0	ENSG00000017...	P2RY14	purinergic rec...	1.0413028999...	7.8954398420...	2.3350763332...	41	14.8	yes
0	ENSG00000014...	CAMK2D	calcium/calm...	1.0328795845...	1.3230211738...	3.9128351215...	83.1	53.8	yes

注：Gene ID：基因编号；gene_name：基因 name；Gene description：基因描述信息；cluster为该基因作为此类群的 marker 基因；p_val为p value值；avg_logFC为此基因在对应 cluster 和除此 cluster 外的所有细胞的表达量差异倍数；pct.1 为该基因在此 cluster 中所有细胞中表达数目的占比；pct.2 为该基因在除此 cluster 外的所有细胞中表达数目的占比；p_val_adj 为经 FDR 矫正的 P 值。

subcluster 上调差异基因统计图



注：横坐标表示当某个 subcluster；纵坐标表示上调基因数目。

2.6 细胞类型差异分析

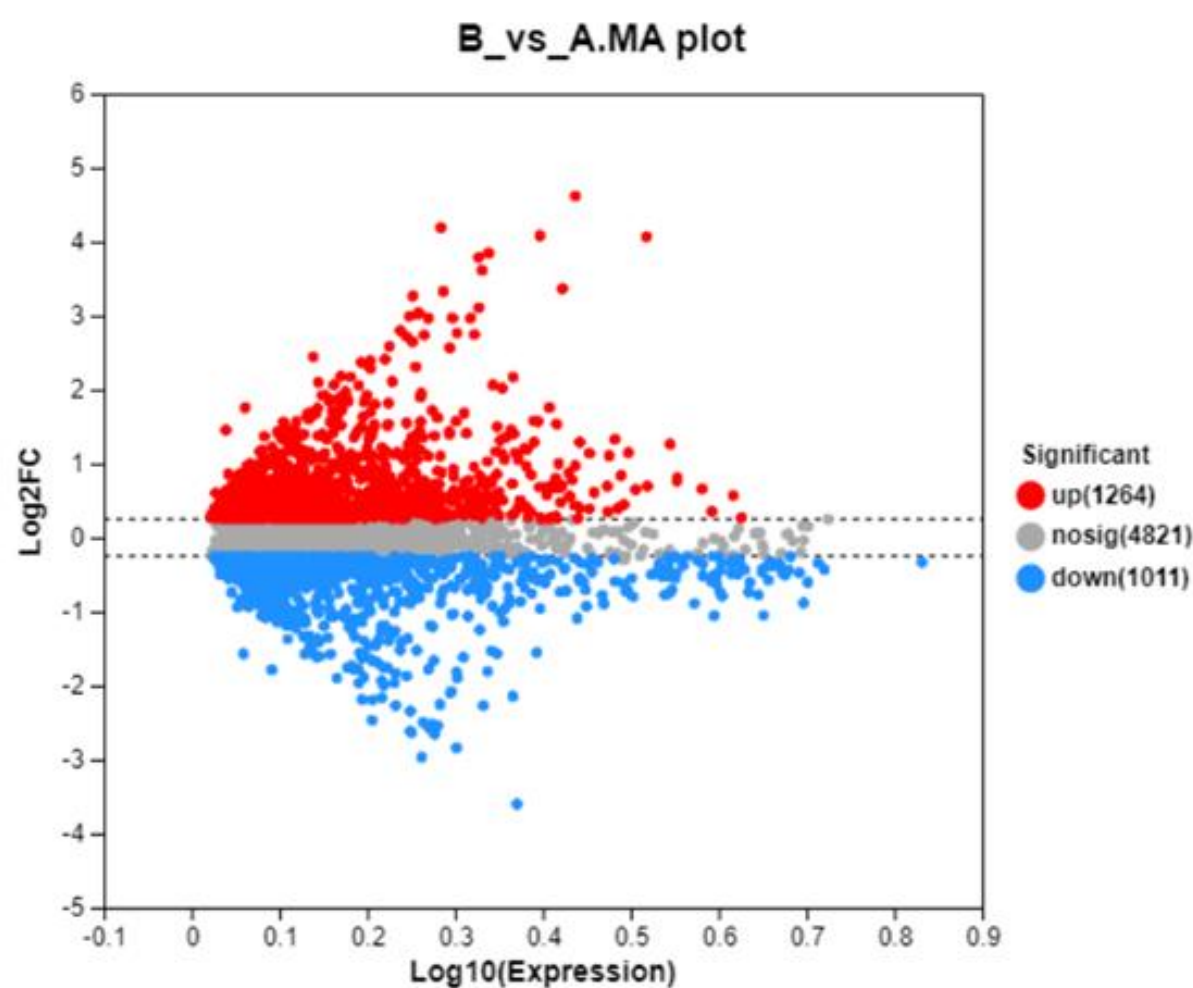
每个样本中的基因占比以及每个 cluster 占比情况都不一样。通过分析多样本比较结果，可以得到样本间占比高的基因、cluster 等，进一步解释生物学问题。每个样本细胞类型比例柱

状堆叠图和占比结果表等，可以反应每个样本中细胞类型的比例的异同，从而找到样本间差异或相同细胞类型。

2.6.1 细胞类型差异分析

对细胞类型/cluster 进行差异表达分析，使我们能够找到特定于每个 cluster 的标记基因,也有助于 cluster 的细胞类型鉴定。

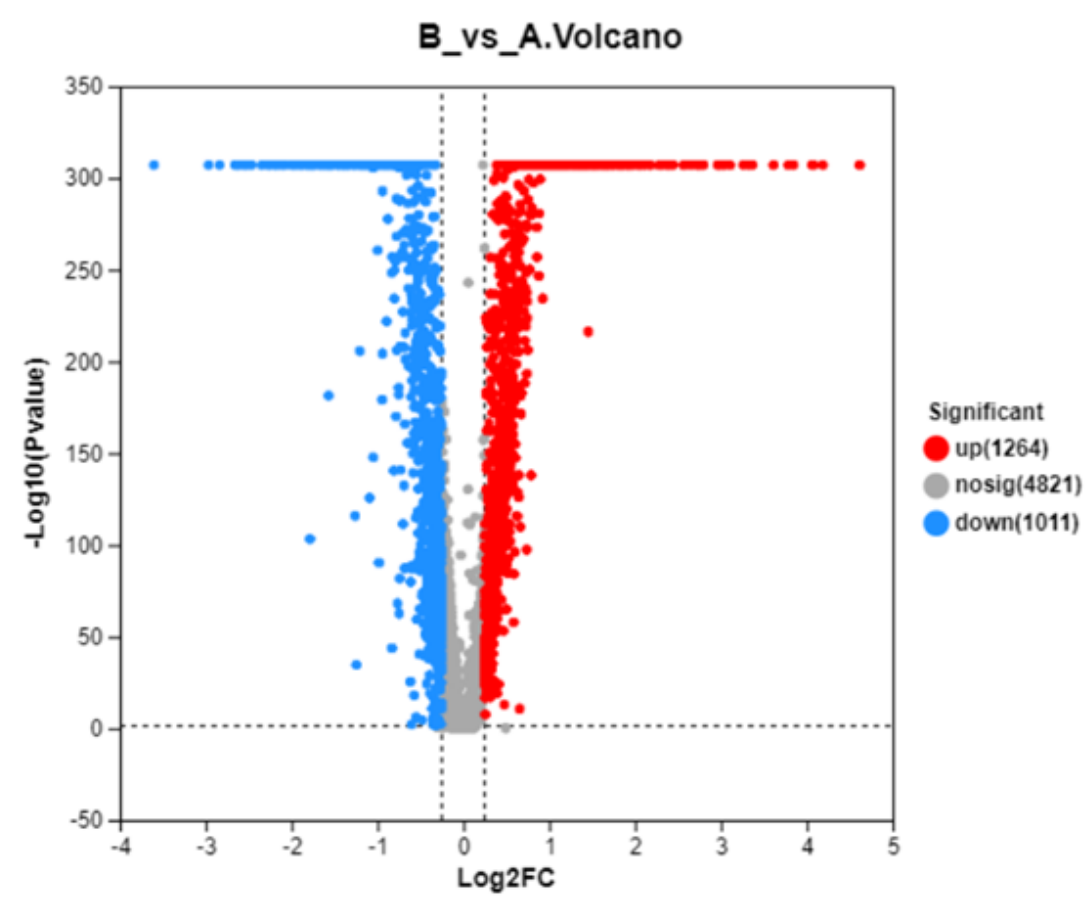
表达量差异 MA 图



注：横坐标为利用差异对比组/细胞集的表达量值进行计算，纵坐标为差异表达分析获得的差异对比组之间基因表达变化，横纵坐标的数值均进行了对数化处理。

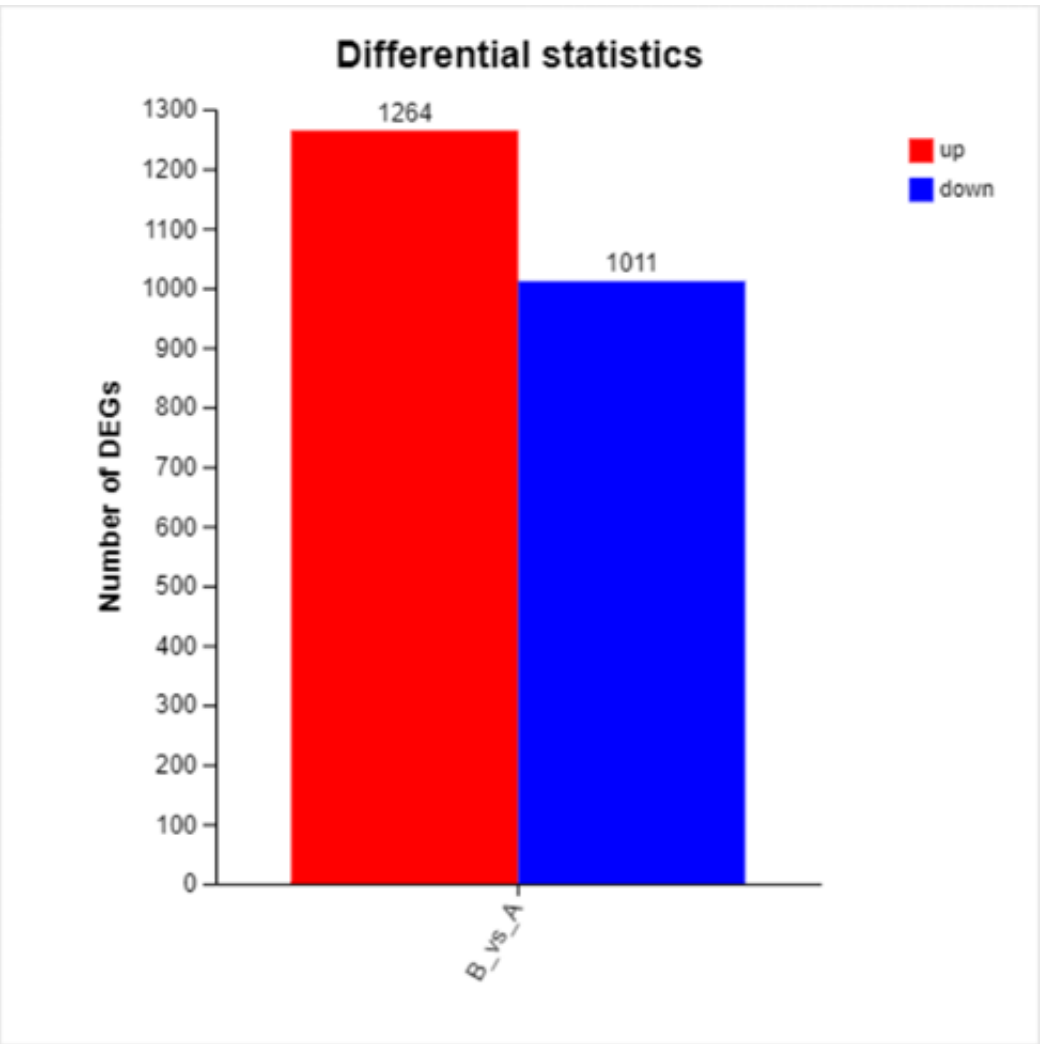
理。图中每个点代表一个特定的基因，默认红色点表示显著上调的基因，蓝色点表示显著下调的基因，灰色点为非显著差异基因/转录本。将所有基因映射上去之后，可以获知，在上边的点为表达差异上调的基因，下边的点为表达差异下调的基因，越靠上边和下边的点表达差异越显著。

表达量差异火山图



注：横坐标为基因在不同细胞间表达差异的倍数变化值，即处理组的表达量除以对照组的表达量得到的数值；纵坐标为基因表达量变化差异的统计学检验值，即 P_{adj} 。P 值越高则表达差异越显著，横纵坐标的数值均进行了对数化处理。图中每个点代表一个特定的基因，默认红色点表示显著上调的基因，蓝色点表示显著下调的基因，灰色点为非显著差异基因。将所有基因映射上去之后，可以获知，在左边的点为表达差异下调的基因，右边的点为表达差异上调的基因，越靠两边和上边的点表达差异越显著。

表达量差异统计图-柱状图



注：横坐标代表不同的差异比较组别/细胞集，纵坐标代表对应的上下调基因数目。默认红色代表上调，蓝色代表下调。

表达量差异详情表

Gene_id	Gene name	Gene descript...	avg_Log2FC(B...	Pvalue	Padjust	pct_1(%)	pct_2(%)	Significant	Regulate
ENSG0000020...	HLA-DRA	major histoco...	4.6169928359...	7.7681955519...	0	99.3	20.2	yes	up
ENSG0000021...	IGKC	immunoglobu...	4.1874425485...	7.7681955519...	0	67.5	12.7	yes	up
ENSG0000019...	HLA-DRB1	major histoco...	4.0799092763...	7.7681955519...	0	99	16.3	yes	up
ENSG0000001...	CD74	CD74 molecu...	4.0648827401...	7.7681955519...	0	99.9	61.4	yes	up
ENSG0000025...	LYN	LYN proto-on...	3.8449524133...	7.7681955519...	0	97.1	5.9	yes	up
ENSG0000015...	BANK1	B cell scaffold...	3.7881201707...	7.7681955519...	0	96.3	2.8	yes	up
ENSG0000019...	HLA-DQA1	major histoco...	3.6129565092...	7.7681955519...	0	97.2	6	yes	up
ENSG0000011...	CD83	CD83 molecu...	3.3655617954...	7.7681955519...	0	97.4	32.3	yes	up
ENSG0000013...	ARHGAP24	Rho GTPase a...	3.3272471587...	7.7681955519...	0	90.2	2.9	yes	up
ENSG0000021...	IGHM	immunoglobu...	3.2635064315...	7.7681955519...	0	74.2	4.4	yes	up

注：（1）Gene id：基因编号；（2）Gene name：基因名称；（3）Gene description：基因描述信息；（4）avg_Log2FC(s1/s2)：该基因在两个组别中差异倍数以 2 为底的对数值，s2 为对照组；（5）Pvalue：该基因在两个组别中差异显著性检验结果；（6）Padjust：多重检验校正后的 P 值；（7）pct_1（%）：该基因在此处理组中所选细胞中表达数目的占比；（8）pct_2（%）该基因在对照组的所选细胞中表达数目的占比；（9）Significant：该基因在两个组别中的差异是否显著，yes 为显著，no 为非显著；（10）Regulate：上下调情况，up 为上调，down 为下调；（11）其他列为基因在各大数据库的注释情况。【默认只展示 GO 和 KEGG2 个数据库的注释信息和所有样本的表达量信息，可通过筛选表格对其他数据库或者目标样本的表达量进行勾选展示】表格中蓝色字体均可以点击链接到基因详情页！

2.7 基因集分析

对差异基因/转录本集中的基因进行功能富集分析，获得该基因集中的基因/转录本主要具有哪些功能或主要参与哪些代谢通路，包括 GO 富集分析、KEGG 富集分析、Reactome 富集分析、DO 富集分析和 DisGeNET 富集分析。

注意：DO 和 Reactome 只支持基因水平分析，且 DO 和 DisGeNET 只能针对研究物种是人的项目进行分析。

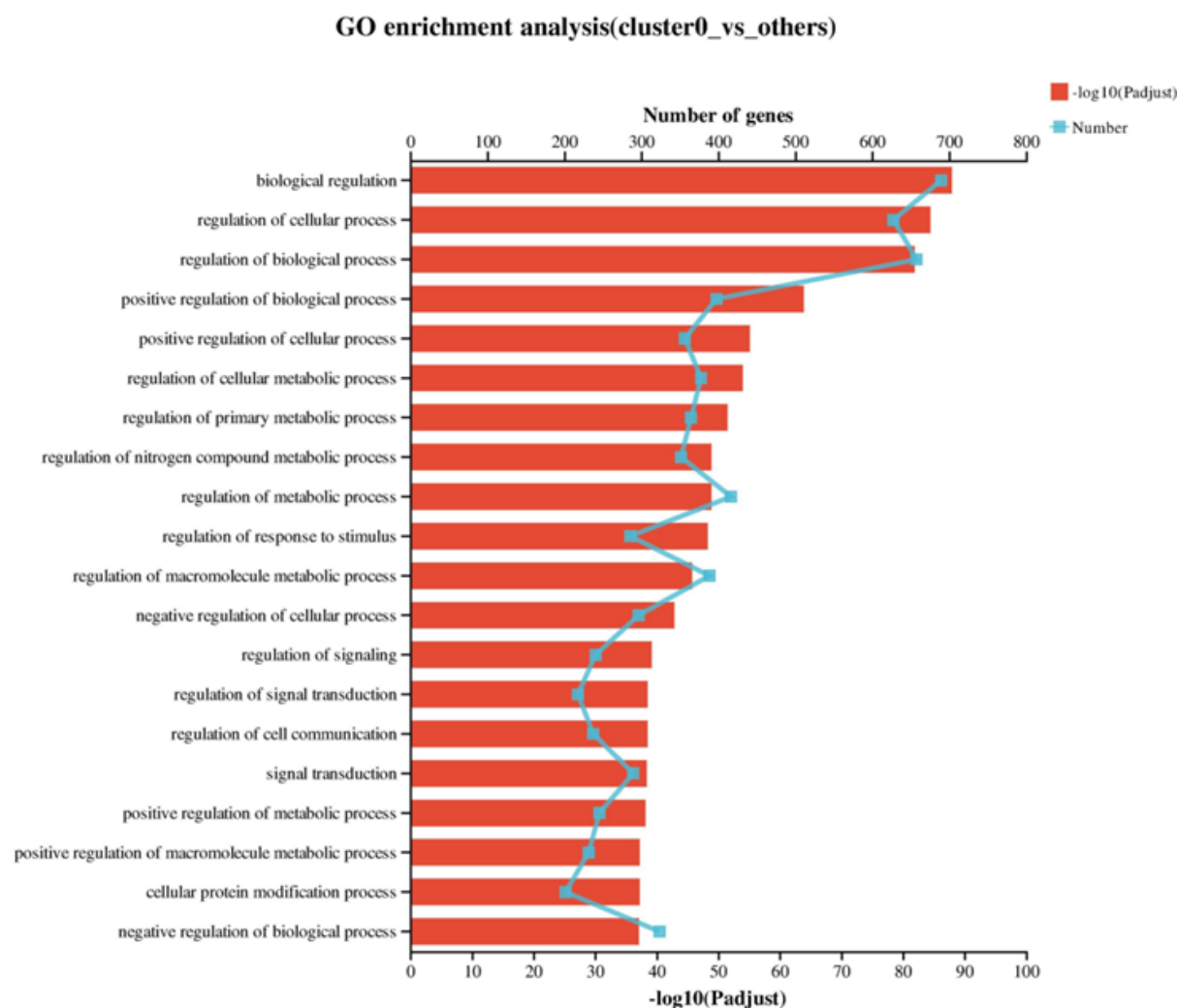
2.7.1 功能富集分析

对差异基因进行 GO 功能显著性富集分析，可以说明差异基因的功能富集情况，在基因功能水平阐明样本间的差异。

2.7.1.1 功能富集分析

功能富集分析是指对选择的目标基因集进行富集分析，使用超几何分布算法获得该基因集中的基因显著富集的功能。通常情况下，默认采用 BH 方法对 P 值进行校正，当经过校正的 P 值（Padjust）<0.05 时，认为此功能存在显著富集情况。

GO 富集分析柱形图（带折线）



注：纵坐标表示 GO term，下方横坐标表示比对上该 GO term 的基因数量，对应的是折线上的不同的点；上方横坐标表示富集的显著性水平，对应的是柱子的高度，其中，Padjust 越小， $-\log_{10}(\text{Padjust})$ 值越大，该 GO term 越显著富集。

【注：默认按 padjust <0.5 的前提下显示 top20 的富集结果，若提醒“没有数据”请确定下是否满足条件。】

GO 富集分析统计表

Number	GO id	Term type	Term description	Ratio_in_study	Ratio_in_pop	Rich factor	Pvalue	Padjust
690	GO:0065007	BP	biological regu...	690/ 1007	15078/ 40649	0.045762037405...	7.4141679713E-...	1.10196778558E...
628	GO:0050794	BP	regulation of ce...	628/ 1007	13000/ 40649	0.048307692307...	4.09932610126E...	3.04641419215E...
658	GO:0050789	BP	regulation of bi...	658/ 1007	14291/ 40649	0.046042964103...	2.17079718423E...	1.07548528497E...
609	GO:0043227	CC	membrane-bou...	609/ 1007	13547/ 40649	0.044954602495	6.05194031911E...	1.19767898915E...
398	GO:0048518	BP	positive regulat...	398/ 1007	6799/ 40649	0.058538020297...	3.03772144616E...	1.12874134636E...
490	GO:0005515	MF	protein binding	490/ 1007	9748/ 40649	0.050266721378...	2.07023588125E...	9.50031245908E...
358	GO:0005829	CC	cytosol	358/ 1007	5967/ 40649	0.059996648231...	1.69977064771E...	1.6819230559E-...
357	GO:0048522	BP	positive regulat...	357/ 1007	6111/ 40649	0.058419243986...	2.25713005443E...	6.7095447998E-...
378	GO:0031323	BP	regulation of ce...	378/ 1007	6789/ 40649	0.055678303137...	3.83186340222E...	9.49216429119E...
365	GO:0080090	BP	regulation of pr...	365/ 1007	6558/ 40649	0.055657212564...	1.38539596296E...	2.94159145679E...

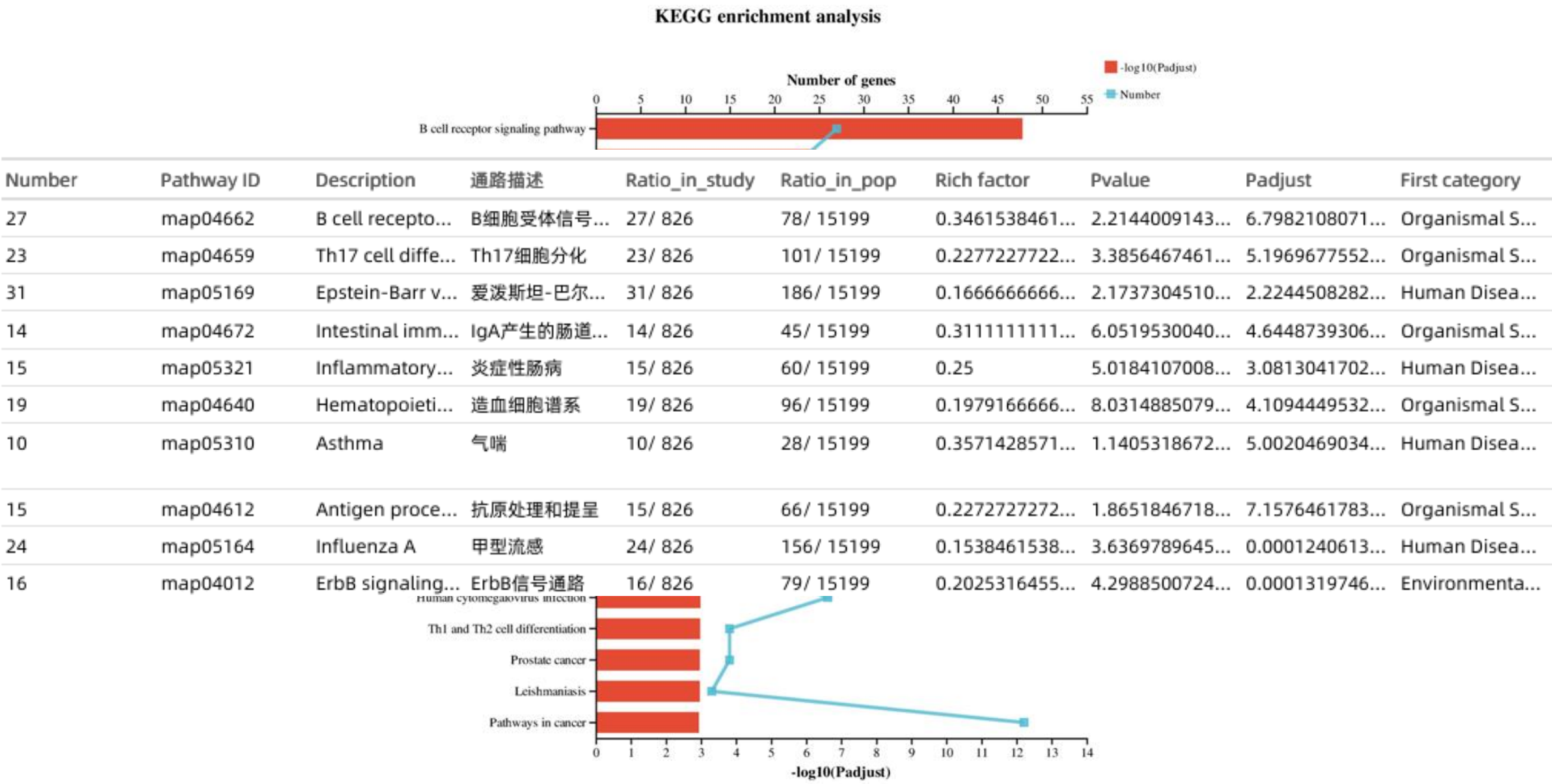
注：（1）Number：富集到该 GO term 的基因数目；（2）GO id：GO term 对应的编号；（3）Term type：GO 三大分类（即 BP、CC、MF）；（4）Term description：GO 功能描述；（5）Ratio_in_study：目标基因集含有 GO term 注释的总基因中落到该 GO term 的比例，分子为目标基因集中注释到该 GO term 的基因数目，分母为目

标基因集中具有 GO term 注释的总基因数目；（6）Ratio_in_pop: 本物种含有 GO term 注释的总基因中落到该 GO term 的比例，分子为本物种中注释到该 GO term 的基因数目，分母为本物种具有 GO term 注释的总基因数目；（7）Pvalue: 未经校正的 P 值，P 值代表富集出来的结果是否具有统计学上的显著意义，P 值越小，在统计学上就越有显著意义；（8）Padjust: 校正后的 P 值，默认采用 BH 方法对 P 值进行校正。

2.7.1.2 KEGG 富集分析

KEGG(Kyoto Encyclopedia of Genes and Genomes，京都基因和基因组百科全书，<http://www.genome.jp/kegg/>) 是基因组破译方面的公共数据库。富集分析方法通常是分析一组基因在某个功能节点上是否出现过，原理是由单个基因的注释分析发展为基因集合的注释分析。富集分析提出了研究的可靠性，能够识别出与□物现象最相关的生物学过程。

KEGG 富集分析带折线柱形图



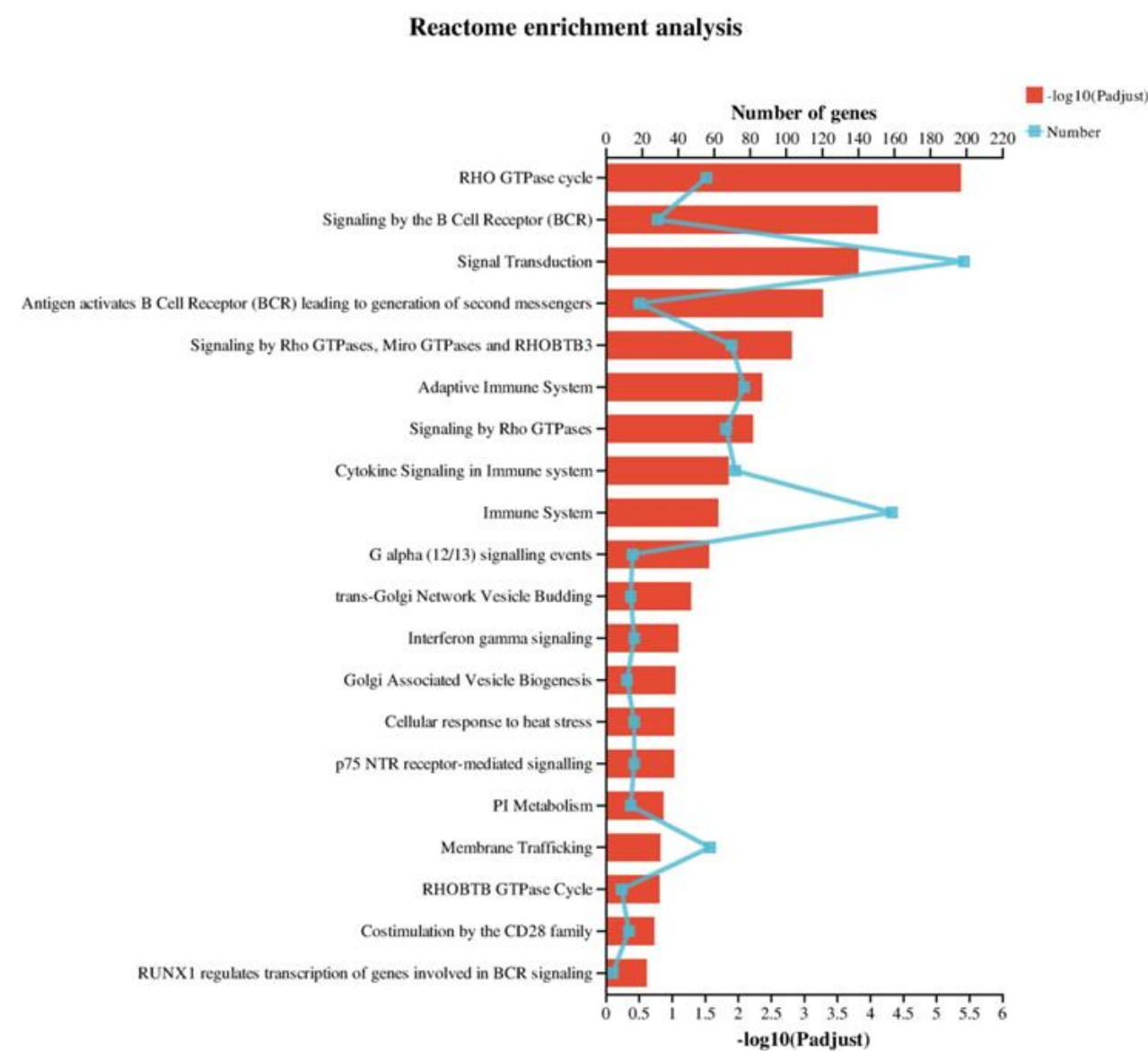
注：纵坐标表示 KEGG pathway，上方横坐标表示比对上该通路的基因数量，对应的是折线上的不同的点；下方横坐标表示富集的显著性水平，对应的是柱子的高度，其中，Padjust 越小，-log10（Padjust）值越大，该 KEGG pathway 越显著富集。【注：默认按 Padjust 从小到大排序显示 top20 的富集结果，可图表工具修改参数调整绘图形式。】

KEGG 富集分析统计表

注：（1）Number：富集到该通路的基因数目；（2）Pathway id：通路编号；（3）Pathway description：KEGG 通路具体描述；（4）Ratio_in_study：目标基因集含有 KEGG 注释的总基因中落到该 KEG G pathway 的比例，分子为目标基因集中注释到该 KEGG pathway 的基因数目，分母为目标基因集中具有 KEGG 注释的总基因数目；（5）Ratio_in_pop：本物种含有 KEGG 注释的总基因中落到该 KEGG pathway 的比例，分子为本物种中注释到该 KEGG pathway 的基因数目，分母为本物种具有 KEGG 注释的总基因数目；（6）Pvalue：未经校正的 P 值，P 值代表富集出来的结果是否具有统计学上的显著意义，P 值越小，在统计学上就越有显著意义，一般 P 值小于 0.05 认为该功能为显著富集项；（7）Padjust：校正后的 P 值，默认采用 BH 方法对 P 值进行校正。（8）First catergory：KEGG 代谢通路的 7 个分支：代谢（Metabolism），遗传信息处理（Genetic Information Processing），环境信息处理（Environmental Information Processing），细胞过程（Cellular Processes），生物体系统（Organismal Systems），人类疾病（Human Diseases），药物开发（Drug Development）；（9）Second catergory：KEGG 代谢通路的名称。

2.7.1.3 Reactome 富集分析

Reactome 富集分析带折线柱形图



注：纵坐标表示 Reactome pathway，上方横坐标表示比对上该通路的基因数量，对应的是折线上的不同的点；下方横坐标表示富集的显著性水平，对应的是柱子的高度，其中，Padjust 越小， $-\log_{10}(\text{P}_{\text{adjust}})$ 值越大，该 Reactome pathway 越显著富集。【注：默认按 Padjust 从小到大排序显示 top20 的富集结果，可图表工具修改参数调整绘图形式。】

Reactome 富集分析统计表

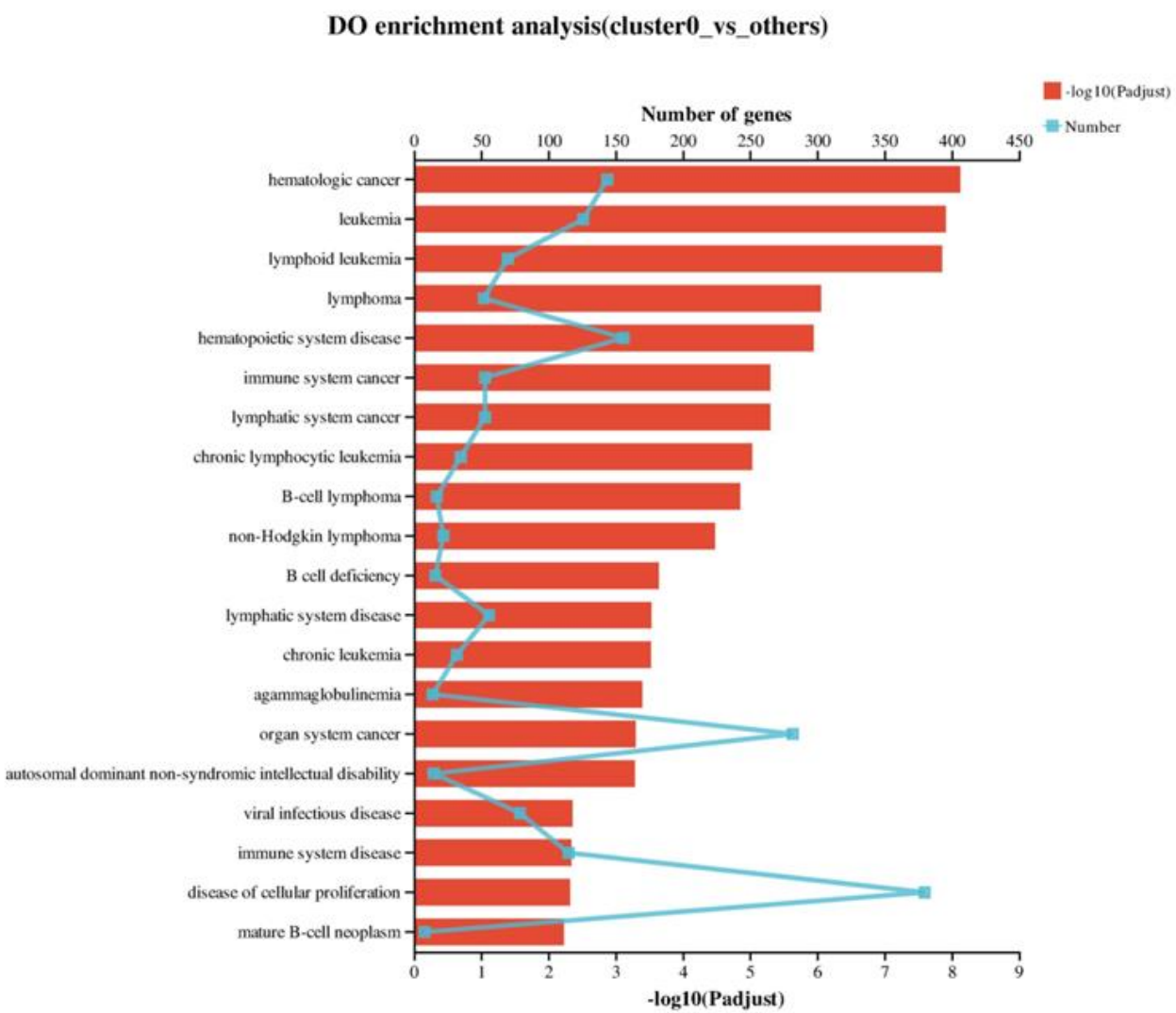
Number	Pathway ID	Pathway descr...	Ratio_in_study	Ratio_in_pop	Rich factor	Pvalue	Padjust	First category
56	R-HSA-9012999	RHO GTPase cy...	56/ 622	419/ 10758	0.133651551312...	2.67495765205E...	4.18630872546E...	Signal Transduc...
29	R-HSA-983705	Signaling by the...	29/ 622	163/ 10758	0.177914110429...	4.84683740574E...	7.58530053998E...	Immune System
199	R-HSA-162582	Signal Transduc...	199/ 622	2482/ 10758	0.080177276390...	9.42969372485E...	0.000147574706...	Signal Transduc...
19	R-HSA-983695	Antigen activate...	19/ 622	86/ 10758	0.220930232558...	3.24658954777E...	0.000508091264...	Immune System
70	R-HSA-9716542	Signaling by Rh...	70/ 622	675/ 10758	0.103703703703...	9.70346565358E...	0.001518592374...	Signal Transduc...
77	R-HSA-1280218	Adaptive Immu...	77/ 622	788/ 10758	0.097715736040...	2.72187103321E...	0.004259728166...	Immune System
67	R-HSA-194315	Signaling by Rh...	67/ 622	661/ 10758	0.101361573373...	3.7502996956E-...	0.005869219023...	Signal Transduc...
72	R-HSA-1280215	Cytokine Signa...	72/ 622	745/ 10758	0.096644295302...	8.73783583767E...	0.013674713086	Immune System
159	R-HSA-168256	Immune System	159/ 622	2027/ 10758	0.078441045880...	1.25744601349E...	0.019679030111...	Immune System
15	R-HSA-416482	G alpha (12/13)...	15/ 622	74/ 10758	0.202702702702...	1.71554928399E...	0.026848346294...	Signal Transduc...

注：（1）Number：富集到该通路的基因或转录本数目；（2）Pathway id：通路编号；（3）Pathway description：Reactome 通路具体描述；（4）Ratio_in_study：目标基因集含有 Reactome 注释的总基因中落到该 Reactome pathway 的比例，分子为目标基因集中注释到该 Reactome pathway 的基因数目，分母为目标基因集中具有 Reactome 注释的总基因数目；（5）Ratio_in_pop：本物种含有 Reactome 注释的总基因（包含本次新预测到的新基因）中落到该 Reactome pathway 的比例，分子为

本物种中注释到该 Reactome pathway 的基因数目，分母为本物种具有 Reactome 注释的总基因数目；（6）Pvalue：未经校正的 P 值，P 值代表富集出来的结果是否具有统计学上的显著意义，P 值越小，在统计学上就越有显著意义，一般 P 值小于 0.05 认为该功能为显著富集项；（7）Padjust：校正后的 P 值，默认采用 BH 方法对 P 值进行校正。（8）First caterygory：Reactome 代谢通路的分支，具体详见综述界面描述。

2.7.1.4 DO 富集分析

DO 富集分析带折线柱形图



注：纵坐标表示 DO term，上方横坐标表示比对上该 DO term 的基因数量，对应的是折线上的不同的点；下方横坐标表示富集的显著性水平，对应的是柱子的高度，其中，Padjust 越小，-log10（Padjust）值越大，该 DO term 越显著富集。

【注：默认按 Padjust 从小到大排序显示 top20 的富集结果，可图表工具修改参数调整绘图形式。】

DO 富集分析统计表

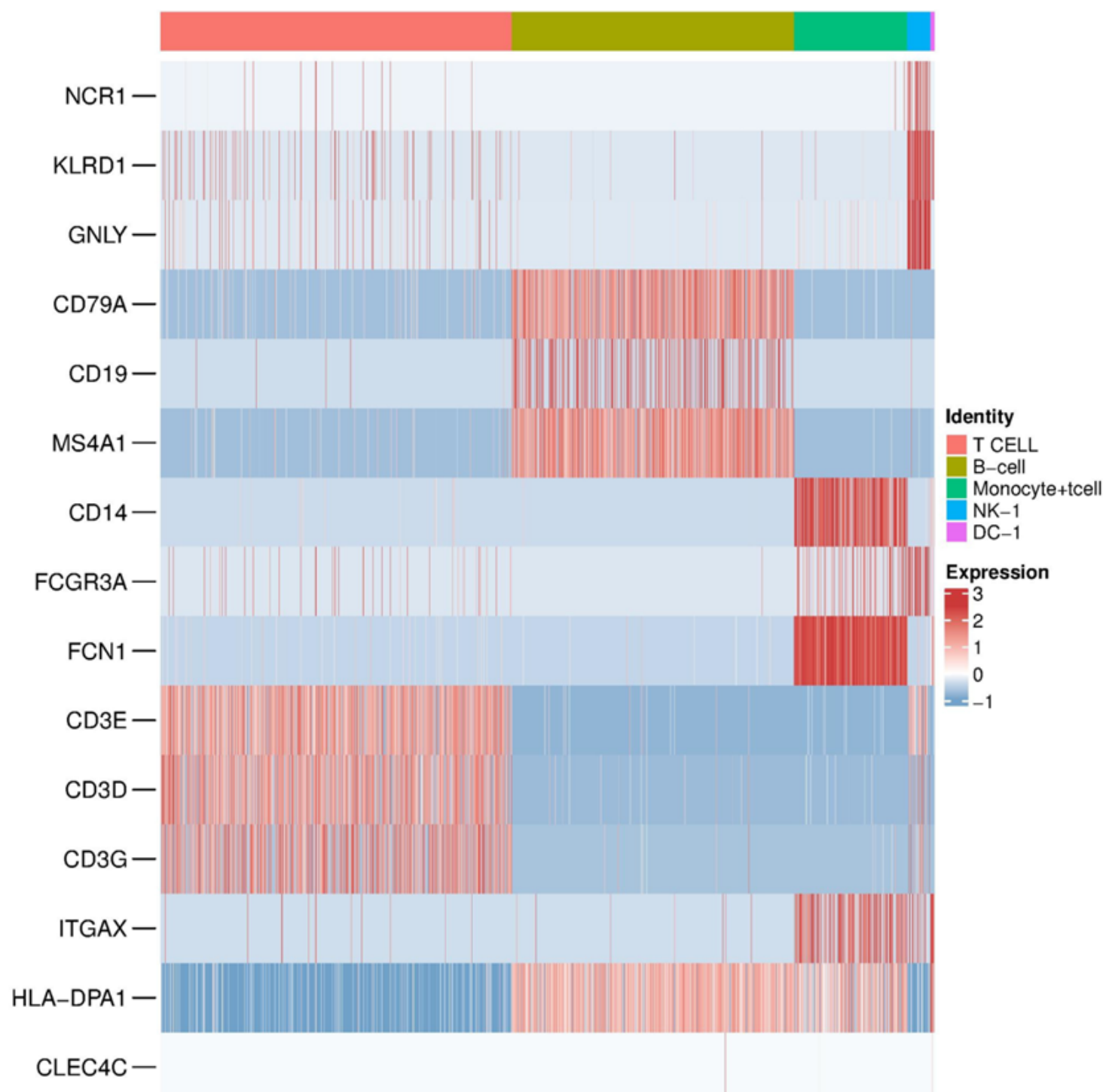
Number	Do id	Term type	DO name	Ratio_in_study	Ratio_in_pop	Rich factor	Pvalue	Padjust
144	DOID:2531	disease of cellu...	hematologic ca...	144/ 587	1315/ 9242	0.109505703422...	4.76158564267E...	7.45664311643E...
126	DOID:1240	disease of cellu...	leukemia	126/ 587	1112/ 9242	0.113309352517...	1.55867839588E...	1.22044518397E...
70	DOID:1037	disease of cellu...	lymphoid leuke...	70/ 587	481/ 9242	0.145530145530...	2.64550252636E...	1.38095231876E...
52	DOID:0060058	disease of cellu...	lymphoma	52/ 587	341/ 9242	0.152492668621...	2.226193051E-9	8.71554579466E...
156	DOID:74	disease of anat...	hematopoietic s...	156/ 587	1597/ 9242	0.097683155917...	3.57780186728E...	1.12056754483E...
53	DOID:0060083	disease of cellu...	immune system...	53/ 587	373/ 9242	0.142091152815...	1.90906869946E...	4.98266930559E...
53	DOID:0060073	disease of cellu...	lymphatic syste...	53/ 587	373/ 9242	0.142091152815...	1.90906869946E...	4.98266930559E...
35	DOID:1040	disease of cellu...	chronic lympho...	35/ 587	203/ 9242	0.172413793103...	4.74615695464E...	9.29060223871E...
17	DOID:707	disease of cellu...	B-cell lymphom...	17/ 587	59/ 9242	0.288135593220...	7.9918405114E-...	1.39058024898E...
22	DOID:0060060	disease of cellu...	non-Hodgkin ly...	22/ 587	100/ 9242	0.22	2.10222580087E...	3.29208560416E...

注：（1）Number：富集到该 DO term 的基因数目；（2）DO id：DO term 对应的编号；（3）Term type：DO 七大分类；（4）DO name：疾病名称；（5）Ratio_in_study：目标基因集含有 DO term 注释的总基因中落到该 DO term 的比例，分子为目标基因集中注释到该 DO term 的基因数目，分母为目标基因集中具有 DO term 注释的总基因数目；（6）Ratio_in_pop：本物种含有 DO term 注释的总基因中落到该 DO term 的比例，分子为本物种中注释到该 DO term 的基因数目，分母为本物种具有 DO term 注释的总基因数目；（7）Pvalue：未经校正的 P 值，P 值代表富集出来的结果是否具有统计学上的显著意义，P 值越小，在统计学上就越有显著意义；（8）Padjust：校正后的 P 值，默认采用 BH 方法对 P 值进行校正。

2.7.2 聚类分析

对选择的基因集在细胞群/样本/单个细胞的表达模式进行聚类分析，并提供热图可视化结果。

聚类热图



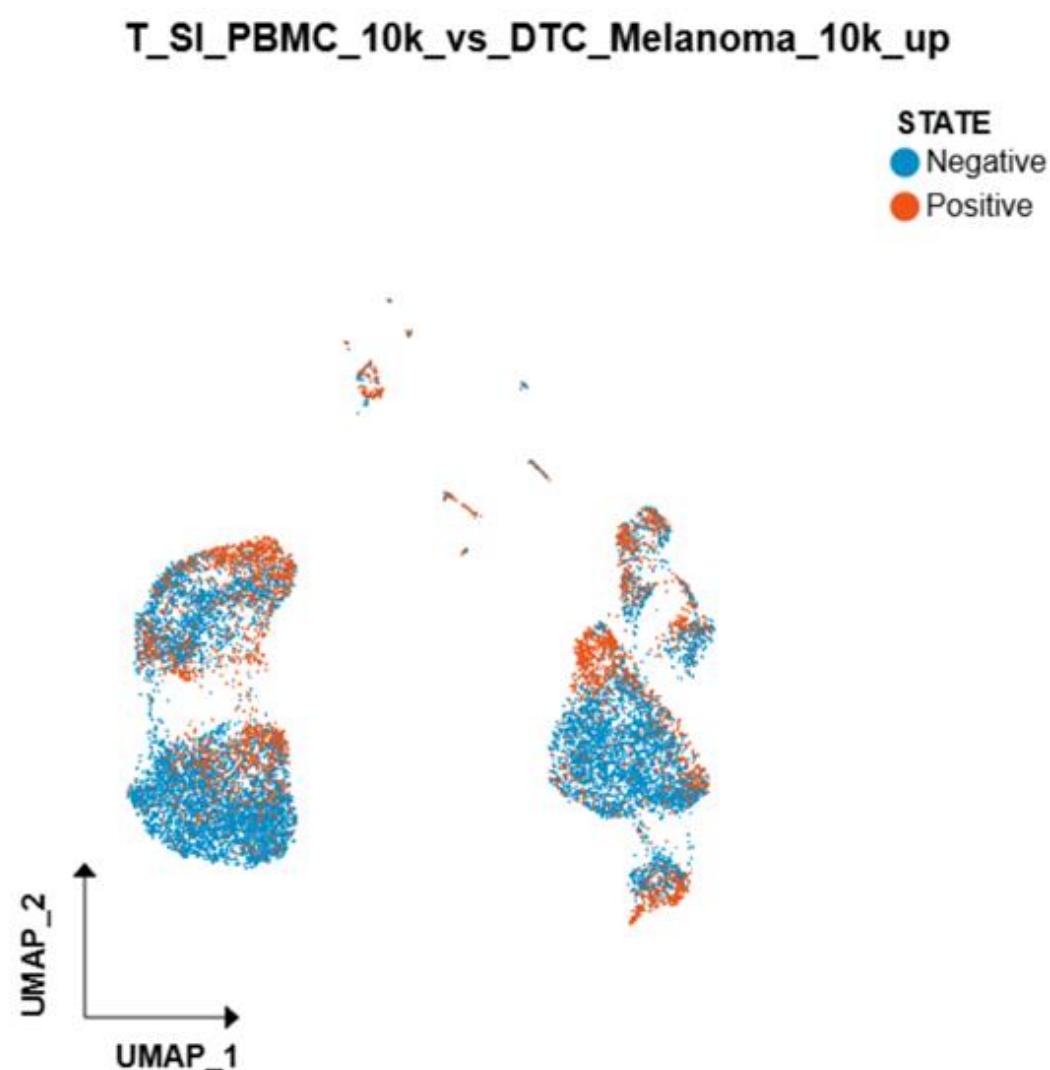
注：横坐标是表示细胞群，默认按照细胞数由多到少排序展示，纵坐标为指定绘图基因名称，颜色表示表达量。

2.7.3 基因集打分

在单细胞数据分析的过程中，Seurat中提供了 Calculate module scores for feature

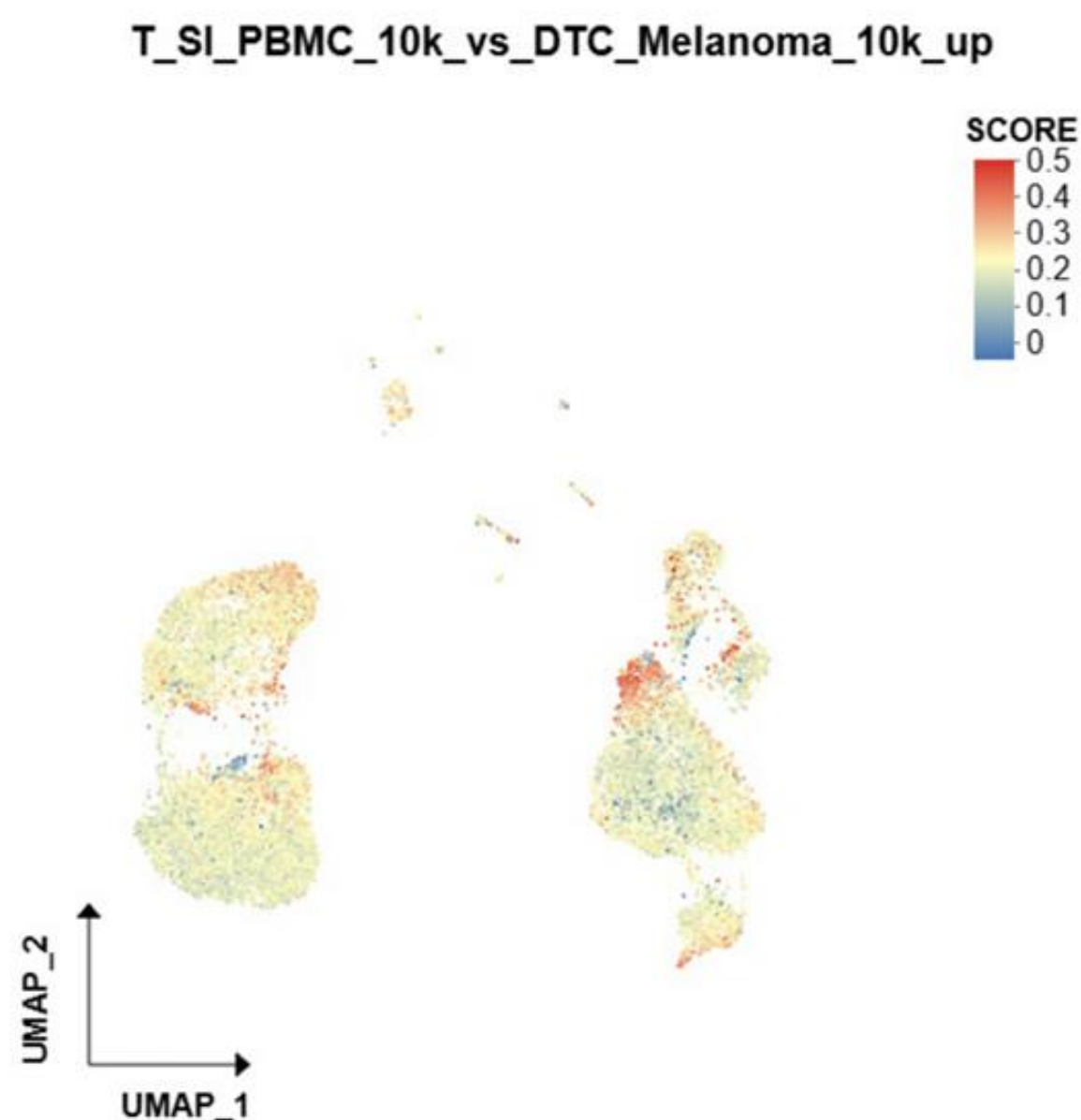
expression programs in single cells 的函数—AddModuleScore，其可以将提供的自定义基因集进行打分。除了进行常规的细胞集打分，本模块还可以基于打分结果，进行人工或者四分位法判定哪些细胞属于与该基因集高相关性的细胞，从而解决自己的生物学问题。

相关性阈值判定结果展示图



注：图中每个点代表一个细胞，Positive 表示高相关性细胞，默认黄色展示，Negative 表示低相关性细胞，默认红色展示。【可在上方切换可视化展示方式】

基因集打分结果 Score 值展示图

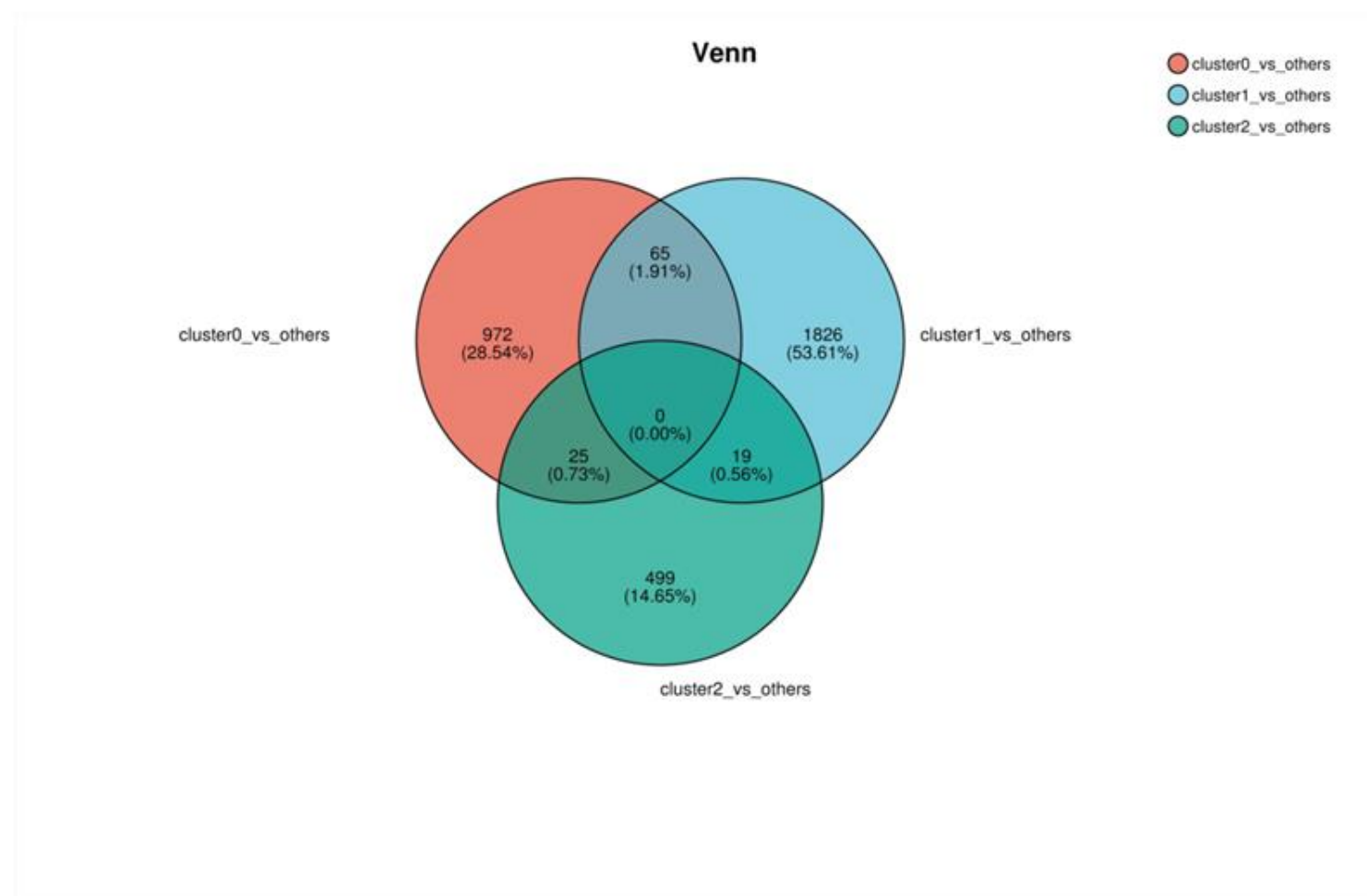


注：图中每个点代表一个细胞。颜色越深，说明 score 值越高。

2.7.4 Venn 分析

Venn 分析展示了各基因集中基因的个数，以及各基因集间基因的重叠关系。本图采用自主研发软件绘制而成，通过基因集 Venn 分析可直观展示基因在各基因集的数量分布状况。

基因集 venn 图



注：不同颜色的圆圈代表不同的基因集，数值代表不同基因集间共有和特有的基因数目。【当参数设置处基因集 ≤ 5 个的时候：圆圈内部所有数字之和代表该基因集基因个数的总和，圆圈的交叉区域代表各基因集中共有基因个数。当参数设置处基因集 > 6 个的时候：花瓣图展示两部分内容，所有基因集共有的基因数目（中心区域）和各基因集特有的基因数目（花瓣区域）。】【点击数字可创建基因集】

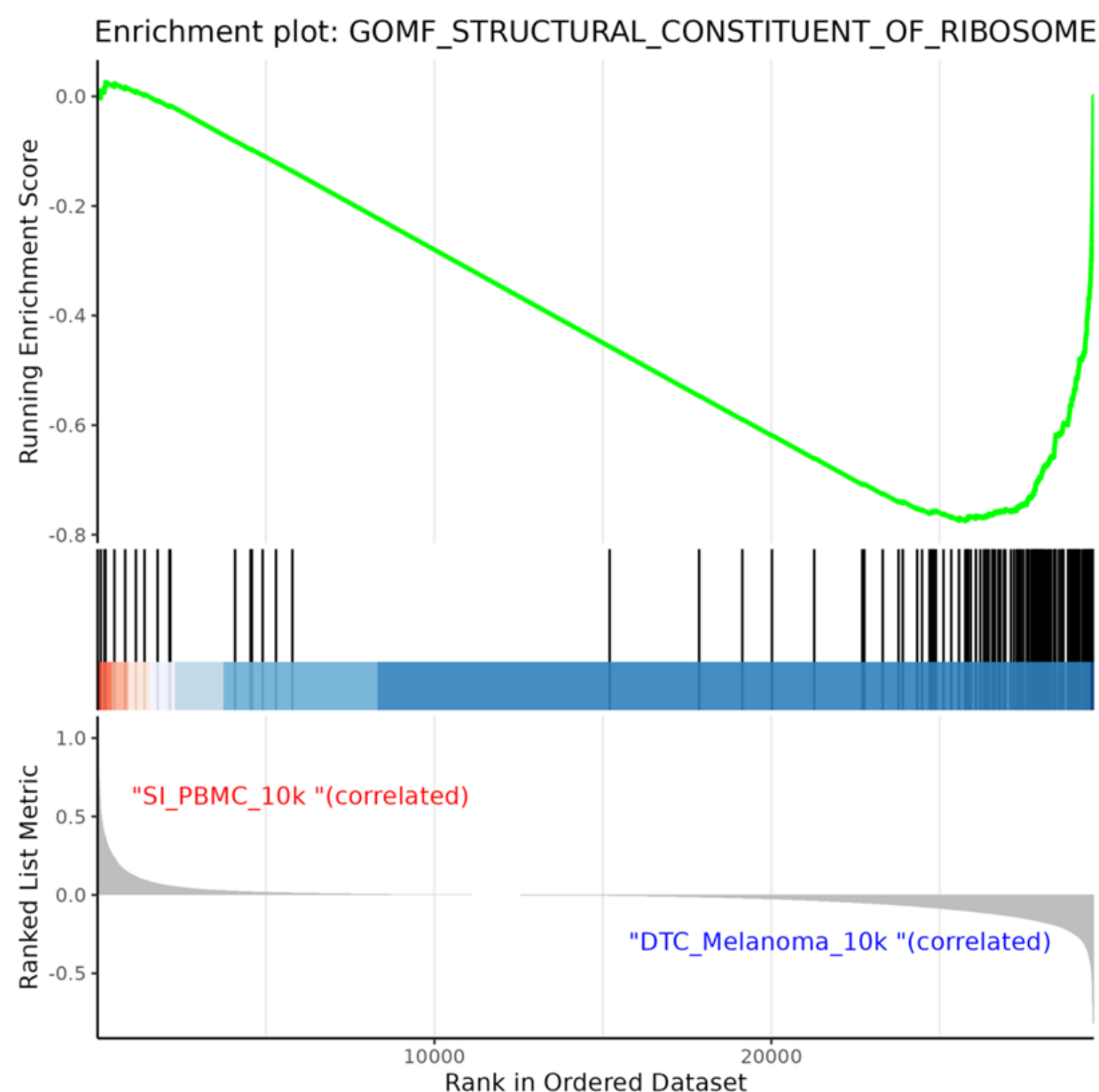
2.7.5 GSEA 富集分析

基因集富集分析（Gene Set Enrichment Analysis, GSEA）的基本思想是使用预定义的基因集，通常来自功能注释或先前实验的结果（收录于 MSigDB），将基因按照在两类样本中的差异表达程度排序，然后检验预先设定的基因集合是否聚集在排序表的顶端或者底端。

☆Tips: GSEA 分析不对基因进行预过滤，直接用所有基因的表达量进行分析，可检测出不显著但是一致的差异表达趋势的基因集，从而排除了传统富集分析中“固定阈值”过滤导致的，对微效基因的有效信息挖掘不足等问题。该分析除挖掘基因功能外，还可对特定细胞类型/cluster 进行功能富集分析，获得其功能富集情况，揭示生物学意义；通过其得到的 cluster 富集注释结果，判定细胞类型。如

巨噬细胞细分时，其中一种细胞亚群特异性注释到促炎功能且高表达促炎因子，则可以大致判定为促炎性 M1-like 巨噬细胞。

GSEA 富集图



注：GSEA 富集图分为 3 个部分。

上：基因富集得分（EnrichmentScore）折线图。横轴为该基因集下的每个基因，纵轴为对应的 ES 值，在折线图中有个峰值，该峰值是这个基因集的 EnrichmentScore，若富集得分为正值，则基因集富集在实验组，峰左侧的基因是该基因集下的核心基因；若富集得分为负值，则表示该基因集在对照组中富集，峰右侧的基因是该基因集下的核心基因。

中：每一条线代表基因集中的一个基因，及其在已排序的基因表达谱中的排序位置，红蓝相间的热图是表达丰度排列，红色越深的表示该位置的基因 logFC 越大，蓝色越深表示 logFC 越小，红色为与第一个表型（实验组）正相关，在该表型中表达高，蓝色与第二个表型（对照组）正相关。如果研究的功能注释基因集的成员显著聚集在表达数据集的顶部或底部，则说明功能基因数据集中的基因在数据集中高表达或低表达，若随机分配，则说明表达数据集与该通路无关。

下：排序后所有基因 rank 值得分布，每个基因对应的信噪比，以灰色面积图展示。ES 值折线呈现原理：排序后的表达基因集的第一个基因开始，如果排序

表达基因集中的基因出现在通路数据集中则加分，反之则减分（加减的值的的大小，需结合第三部分的 rank 值，列表中每个基因的权重不一样）。

GSEA 分析统计表

Gene set name	Description	Group	Size	ES	NES	Pvalue	Padjust	Rank at MAX	Leading_edg...
GOMF_STRUC...	The action of ...	DTC_Melanom...	158	-0.7750804841...	-2.0857257283...	2.0378577734...	1.6302862187...	3820	120
GOMF_ELECTR...	Any molecula...	DTC_Melanom...	115	-0.7901763439...	-2.0685755806...	9.5479662116...	3.8191864846...	3100	69
GOMF_OXIDO...	Primary active...	DTC_Melanom...	66	-0.8661533080...	-2.1342997746...	1.1760022194...	3.1360059184...	2762	53
GOMF_PRIMA...	Enables the tr...	DTC_Melanom...	160	-0.6965697819...	-1.8729629026...	8.9783318858...	1.7956663771...	3106	69
GOMF_NAD_P...	Catalysis of th...	DTC_Melanom...	45	-0.8503023364...	-1.9717269666...	6.1533182325...	9.8453091720...	2671	35
GOMF_OXIDO...	Catalysis of a...	DTC_Melanom...	57	-0.7948168586...	-1.9131066415...	1.2110969254...	1.6147959006...	2671	35
GOMF_OXIDO...	Catalysis of a...	DTC_Melanom...	87	-0.7181237468...	-1.8299772121...	9.6276390541...	1.1003016061...	4436	49
GOMF_PROTO...	Enables the tr...	DTC_Melanom...	136	-0.6550740025...	-1.7428670010...	3.9873701153...	3.5443289914...	3067	51
GOMF_STRUC...	The action of ...	SI_PBMC_10k	75	0.6777401111...	2.0465634049...	2.1576452874...	0.0001726116...	4036	40
GOMF_LIGASE...	Catalysis of th...	DTC_Melanom...	165	-0.6131632272...	-1.6506862714...	3.8316989612...	0.0002786690...	6159	88

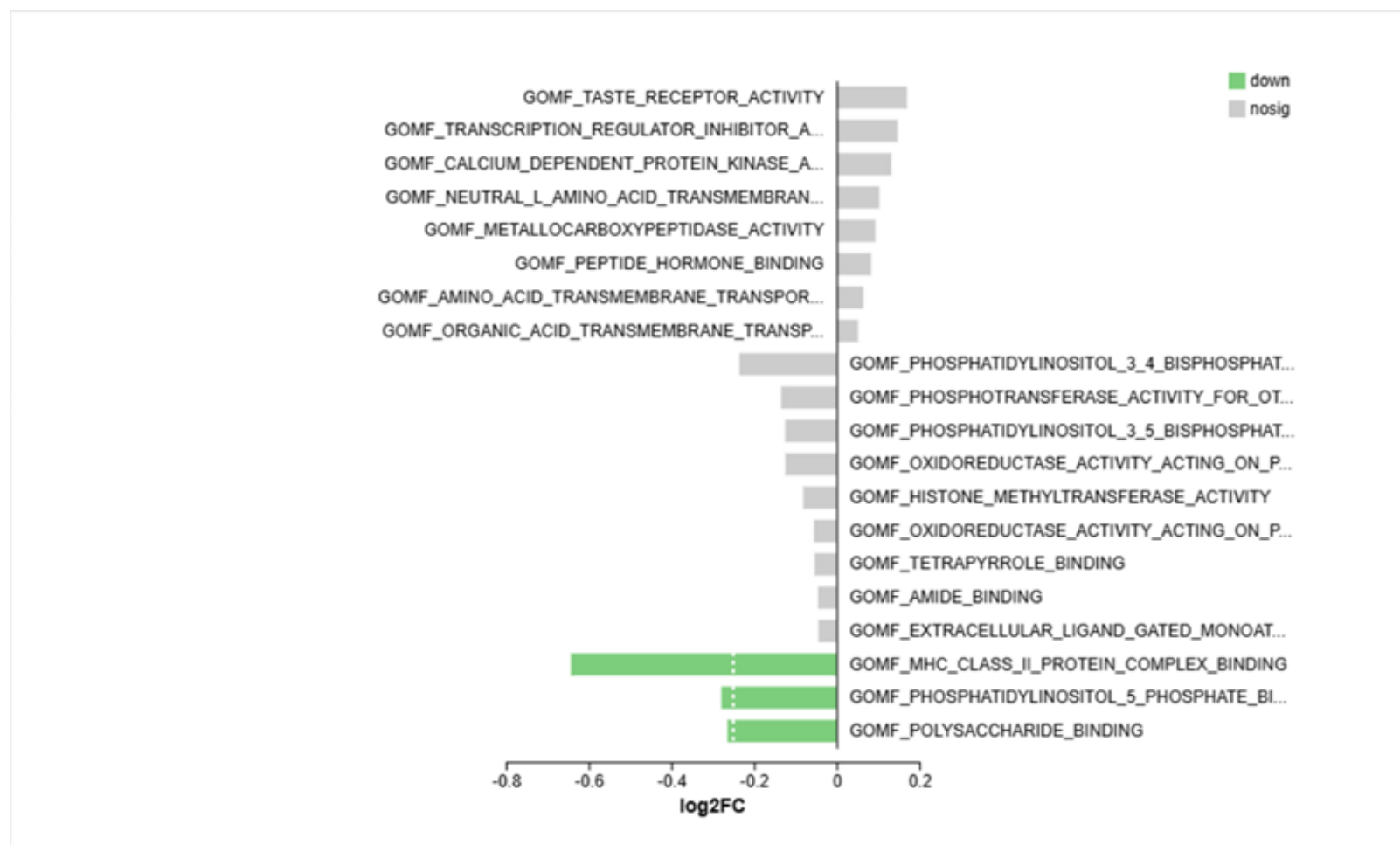
- 注：（1）Gene set name: 基因集名称；
- （2）Description: 基因集描述信息；
- （3）Group: 表示与富集基因集呈正相关的组别；
- （4）Size: 代表该基因集中包含在表达矩阵中的基因数目（经过条件筛选的值）；
- （5）ES 值: EnrichmentScore，基因集的富集分数，该基因集在表达数据中的基因排名列表的顶部或底部被过度代表的程度，反映基因集成员在排序列表的两端富集的程度；（6）NES 值: 标准化富集得分，归一化的 ES 值。通过标准化富集分数，GSEA 解决了基因集大小以及基因集与表达矩阵之间相关性的差异问题，可以使用 NES 来比较各个基因集的分析结果；（7）Pvalue: 针对获得的富集得分进行统计学检验，是对富集分数 ES 的统计学分析，用来表示富集结果的可信度；
- （8）Padjust: 错误发现率（FDR），归一化的富集得分代表假阳性结果的估计概率；
- （9）Rank at MAX: 该基因集 ES 值出现在排序列表中的位置；
- （10）Leading-edge: 对富集得分贡献最大的基因成员。

【说明：NES 绝对值越大，Padjust 越小，富集越显著，常见参数有 Pvalue<0.05，Padjust<0.25，同时，也可加入|NES|>1 作为额外的显著性判断标准。GSEA 官方没有固定判断标准，因此建议使用者根据自身情况或文献情况，对显著性阈值做出判断。】点击蓝色的基因集名称可切换展示下面的 GESA 富集图。

2.7.6 GSVA 富集分析

GSVA (Gene Set Variation Analysis), 是一种非参数, 无监督的算法。与 GSEA 不同, GSVA 不需要预先对样本进行分组, 可以计算每个样本中特定基因集的富集分数。换言之, GSVA 转化了基因表达数据, 从单个基因作为特征的表达矩阵, 转化为特定基因集作为特征的表达矩阵。GSVA 对基因富集结果进行了量化, 可以更方便地进行后续统计分析。如果用 limma 包做差异表达分析可以寻找样本间差异表达的基因, 同样地, 使用 limma 包对 GSVA 的结果 (依然是一个矩阵) 做同样的分析, 则可以寻找样本间有显著差异的基因集。这些“差异表达”的基因集, 相对于基因而言, 更加具有生物学意义, 更具有可解释性。

GSVA 差异富集图



注：此图为 GSVA 差异富集结果图，是两个基因集的比较分析。横坐标可做 log2FC 和 t 值切换，均表示基因集差异大小。纵坐标为通路或者其他注释基因集名称。该图默认 $P_{\text{adjust}} < 0.05$ 时， $|\log2FC| \geq 0.25$ 或者 $|t| \geq 1.96$ 即为差异显著的基因集。默认 Y 轴右侧蓝色基因集代表差异显著的上调基因集，Y 轴左侧绿色基因集代表差异显著的下调基因集。灰色部分为差异不显著的通路。

GSVA 聚类热图统计表

Gene set name	Description	Number
GOCC_XY_BODY	A structure found in a male mammalian spermatoc...	15
GOCC_VOLTAGE_GATED_CALCIUM_CHANNEL_COM...	A protein complex that forms a transmembrane ch...	54
GOCC_VESICLE_TETHERING_COMPLEX	Any protein complex that plays a role in vesicle teth...	57
GOCC_VESICLE_LUMEN	The volume enclosed by the membrane or protein...	306
GOCC_VESICLE_COAT	A membrane coat found on a coated vesicle. [GOC...	61
GOCC_VACUOLAR_MEMBRANE	The lipid bilayer surrounding the vacuole and sepa...	472
GOCC_VACUOLAR_LUMEN	The volume enclosed within the vacuolar membra...	166
GOCC_UBIQUITIN_LIGASE_COMPLEX	A protein complex that includes a ubiquitin-protein...	305
GOCC_U5_SNRNP	A ribonucleoprotein complex that contains small nu...	19
GOCC_U2_TYPE_SPLICEOSOMAL_COMPLEX	Any spliceosomal complex that forms during the sp...	90

注：（1）Gene set name：基因集名称；（2）Description：基因集描述信息；（3）Number：该基因集中包含的基因数目；【点击 Number 可查看相应基因集的基因详情】

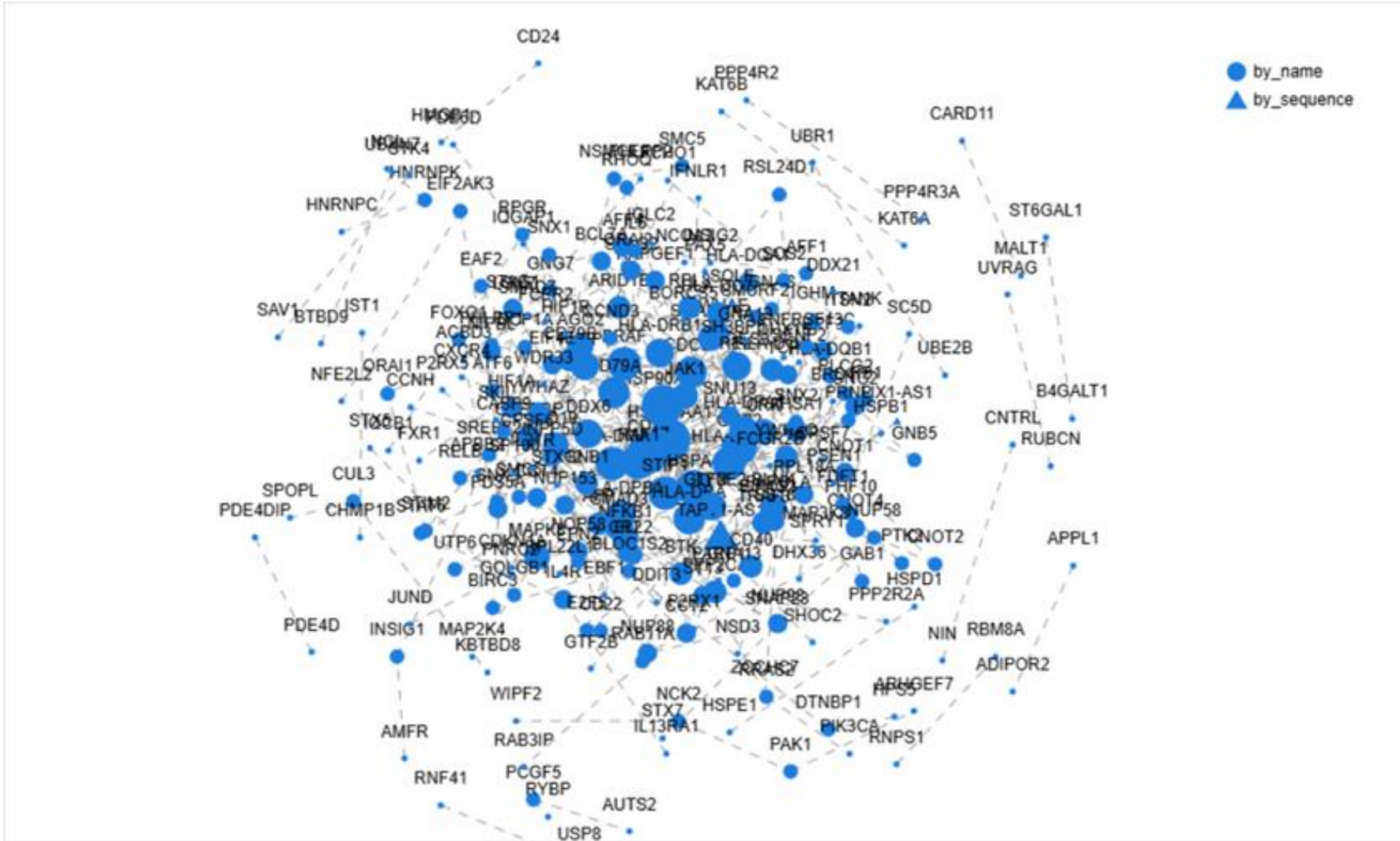
2.7.7 蛋白互作网络

蛋白互作网络分析用于研究基因产物—蛋白间是否存在相互作用，基于基因对应的蛋白信息进行分析。本项分析应用 STRING 数据库（<http://string-db.org/>）对感兴趣基因进行蛋白互作网络分析。针对数据库中存在蛋白互作关系的物种，直接从数据库中提取出感兴趣基因对应的互作关系来构建网络。

获得蛋白互作网络关系后，利用 Python 下的 networkX 对感兴趣基因进行网络可视化。统计基因互作关系网络的各种拓扑性质，如网络中各个节点的联通性（Degree），介数中心性（Betweenness），亲密系数（Closeness）以及聚类系数（Cluster Coefficient）等，可以获得互作网络中关键节点。

☆Tips：系统分析大量蛋白在生物系统中的相互作用关系，对了解生物系统中蛋白质的工作原理，了解疾病等特殊生理状态下生物信号和能量物质代谢的反应机制，以及了解蛋白之间的功能联系都有重要意义。在生物医药领域有助于从系统的角度研究疾病分子机制、发现新药靶点等等。

蛋白互作网络图



注：节点表示基因，边表示两基因间存在互作关系。节点的大小与该节点的连通性（即 **degree**）成正比，即与此节点相连的边越多，其度越大节点也就越大，表明基因在该网络中的重要性越强。

2.8 功能注释与查询

2.8.1 功能注释与查询

将基因在 GO、KEGG、NR、EggNOG、Swiss-Prot、Pfam、Reactome、DO 和 DisGeNet 等九大数据库的功能注释进行汇总，并提供多种检索方式，实现查询特定功能对应的基因信息，或根据基因信息检索其功能，以便迅速锁定所需信息。

功能注释信息表

Gene id	Gene name	Gene description
ENSG00000282221	-----	novel transcript, antisense to LAD1
ENSG00000249042	FAM151B-DT	FAM151B divergent transcript [Source:HGNC Symbo...
ENSG00000255112	CHMP1B	charged multivesicular body protein 1B [Source:HG...
ENSG00000235295	LINC01634	long intergenic non-protein coding RNA 1634 [Sour...
ENSG00000070814	TCOF1	treacle ribosome biogenesis factor 1 [Source:HGNC...
ENSG00000126653	NSRP1	nuclear speckle splicing regulatory protein 1 [Sourc...
ENSG00000261604	-----	novel transcript, antisense to HMGCS1
ENSG00000261602	-----	novel transcript
ENSG00000261600	-----	novel transcript
ENSG00000157837	SPPL3	signal peptide peptidase like 3 [Source:HGNC Symb...

注：（1）Gene id：基因编号；（2）Gene name：基因名称；（3）Gene description：基因描述信息；（4）其他列为基因在各大数据库的注释情况。【默认只展示 GO 和 KEGG2 个数据库的注释信息，可通过筛选表格对其他数据库进行勾选展示】表格中蓝色字体均可以点击链接到基因详情页！

2.9 个性化分析

2.9.1 拟时序分析-Monocle2

构建细胞谱系发育，也就是 pseudotime(拟时序分析)，也叫细胞轨迹(cell trajectory) 分析，主要是判断不同细胞表达量之间的关系，不同亚群之间表达量过渡的变化就是一条轨迹，类比于随着时间的发育过程的基因表达变化，但并不是真正的时间上的变化，因此我们叫它"拟时序"。拟时序分析可以通过构建细胞间的变化轨迹来重塑细胞随着时间的变化过程。而通过亚群基因的表达特征，可以揭示分化过程中的关键调控基因。该分析广泛应用于细胞发育过程中的变化轨迹或者某类细胞的分化来源。

本模块使用 Monocle2 软件包来进行分析。

2.9.1.1 拟时轨迹初步分析

构建细胞谱系发育，也就是 pseudotime(拟时序分析)，也叫细胞轨迹(cell trajectory) 分析，主要是判断不同细胞表达量之间的关系，不同亚群之间表达量过渡的变化就是一条轨迹，类比于随着时间的发育过程的基因表达变化，但并不是真正的时间上的变化，因此我们叫它"拟时序"。本模块使用 Monocle2 软件进行拟时序分析，其可以通过构建细胞间的变化轨迹来重塑细胞随着时间的变化过程。而通过亚群基因的表达特征，可以揭示分化过程中的关键调控基因。该分析广泛应用于细胞发育过程中的变化轨迹或者某类细胞的分化来源。

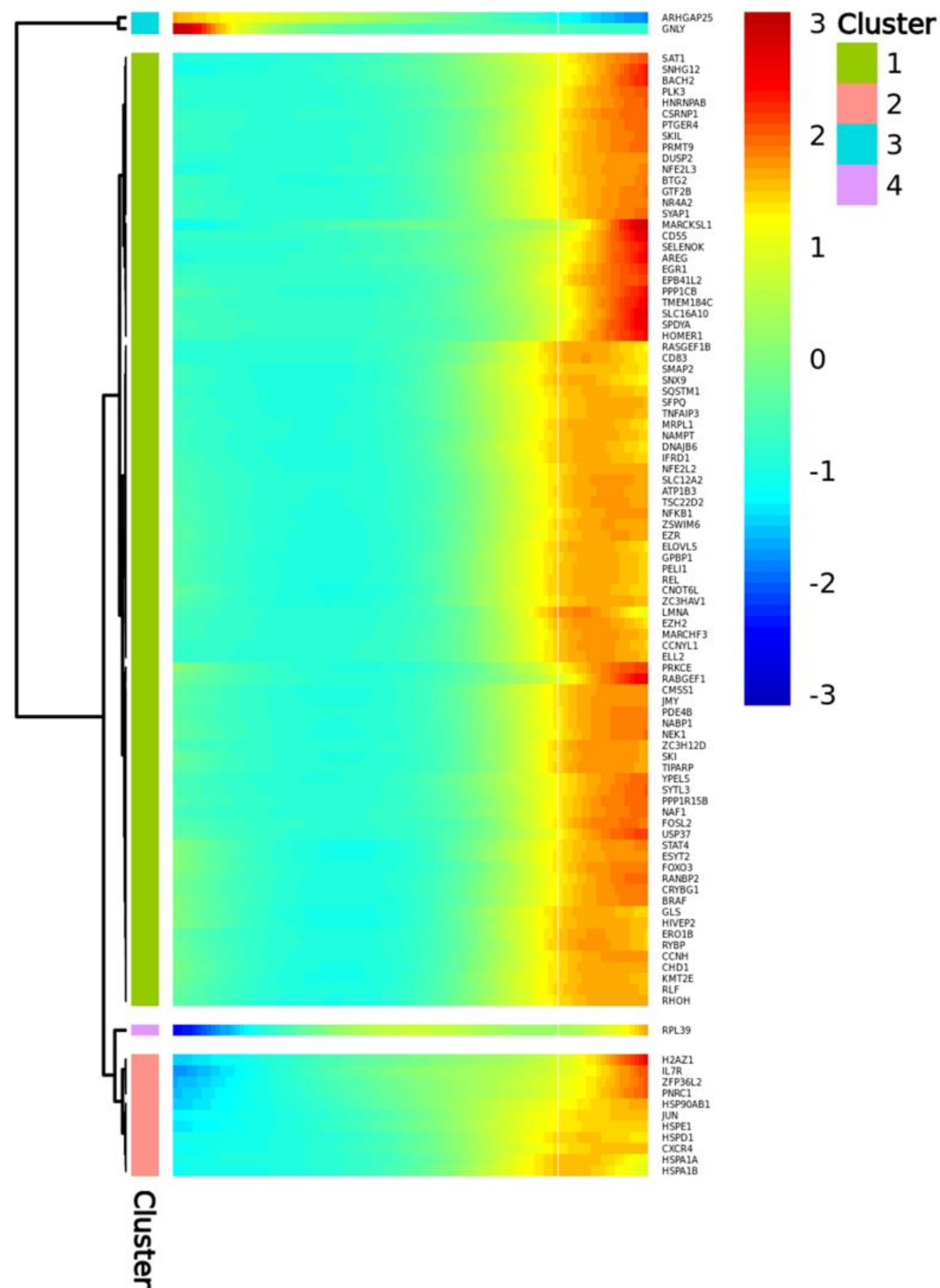
Tips: 注意计算机没法判断真实的起始点，要通过生物学意义去确定起始点。

拟时序分析统计表-按 Celltype 展示

Celltype	State	Cell_Num	Celltype_Num	State_Num
subcluster1	1	643	1604	2572
subcluster2	1	660	1498	2572
subcluster3	1	632	1357	2572
subcluster4	1	481	827	2572
subcluster5	1	130	703	2572
subcluster6	1	3	449	2572
subcluster7	1	23	115	2572
subcluster1	2	785	1604	2113
subcluster2	2	580	1498	2113
subcluster3	2	646	1357	2113

注：（1）Celltype：细胞类型；（2）State：拟时序分析得到的不同状态；（3）Cell_Num：该细胞类型（Celltype）在某状态（State）下的细胞个数；（4）Celltype_Num：该细胞类型总的细胞个数；（5）State_Num：处于某状态下的总细胞个数。

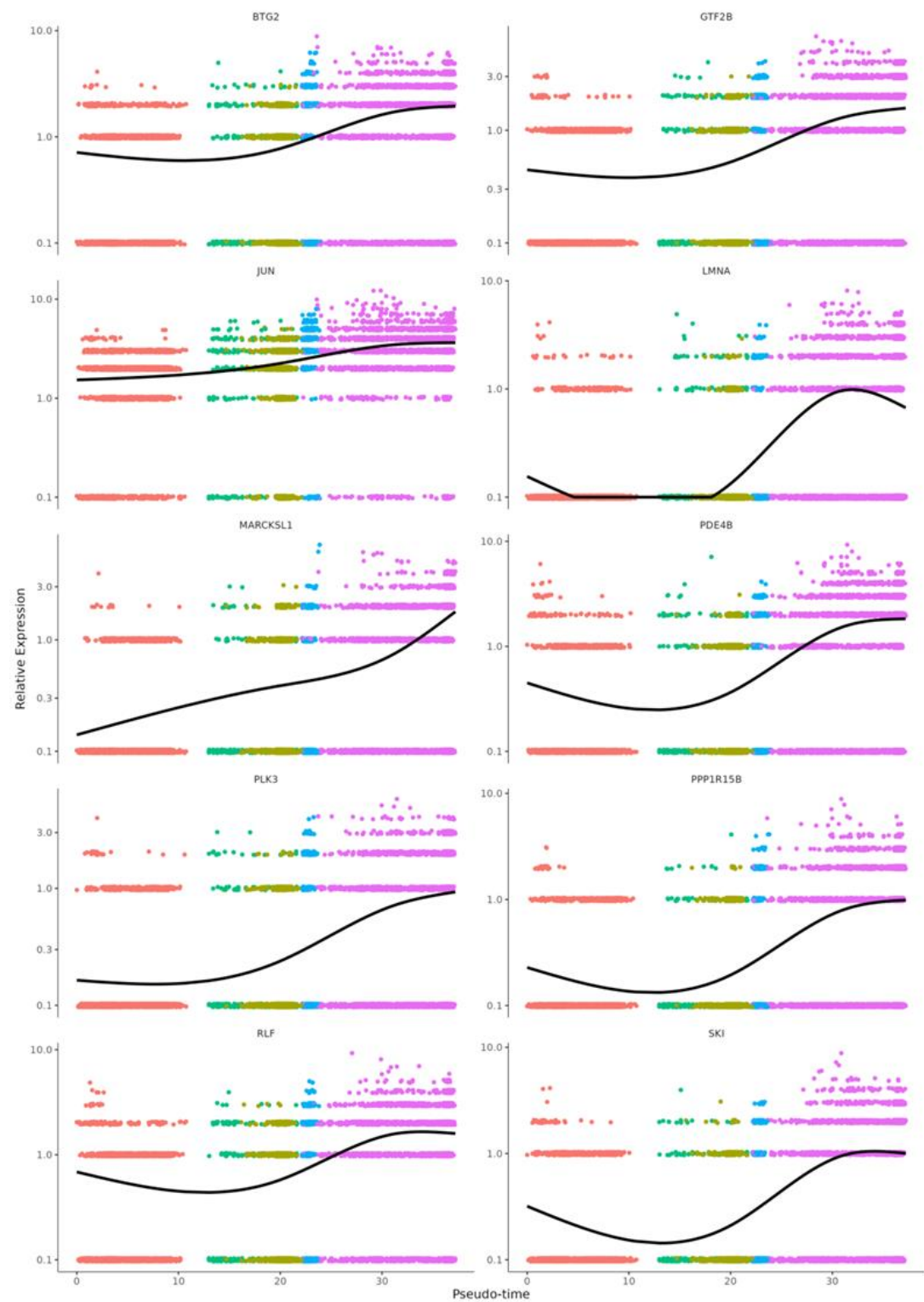
特定拟时相关差异基因热图



注：选取分化过程中差异最显著的100个基因，将这些基因聚类可视化，可以观察在拟时间过程中各个表达基因模块的变化（每个 cluster 代表一类表达趋势一致的基因集模块）。图中横坐标为拟时间顺序，

纵坐标表示不同的 Cluster 基因模块，每行代表一个基因，每列代表当前细胞状态下的平均表达值，颜色按照由红到蓝逐渐降低。【可切换 Top50 和 Top20 展示】

整体拟时相关差异基因表达趋势图



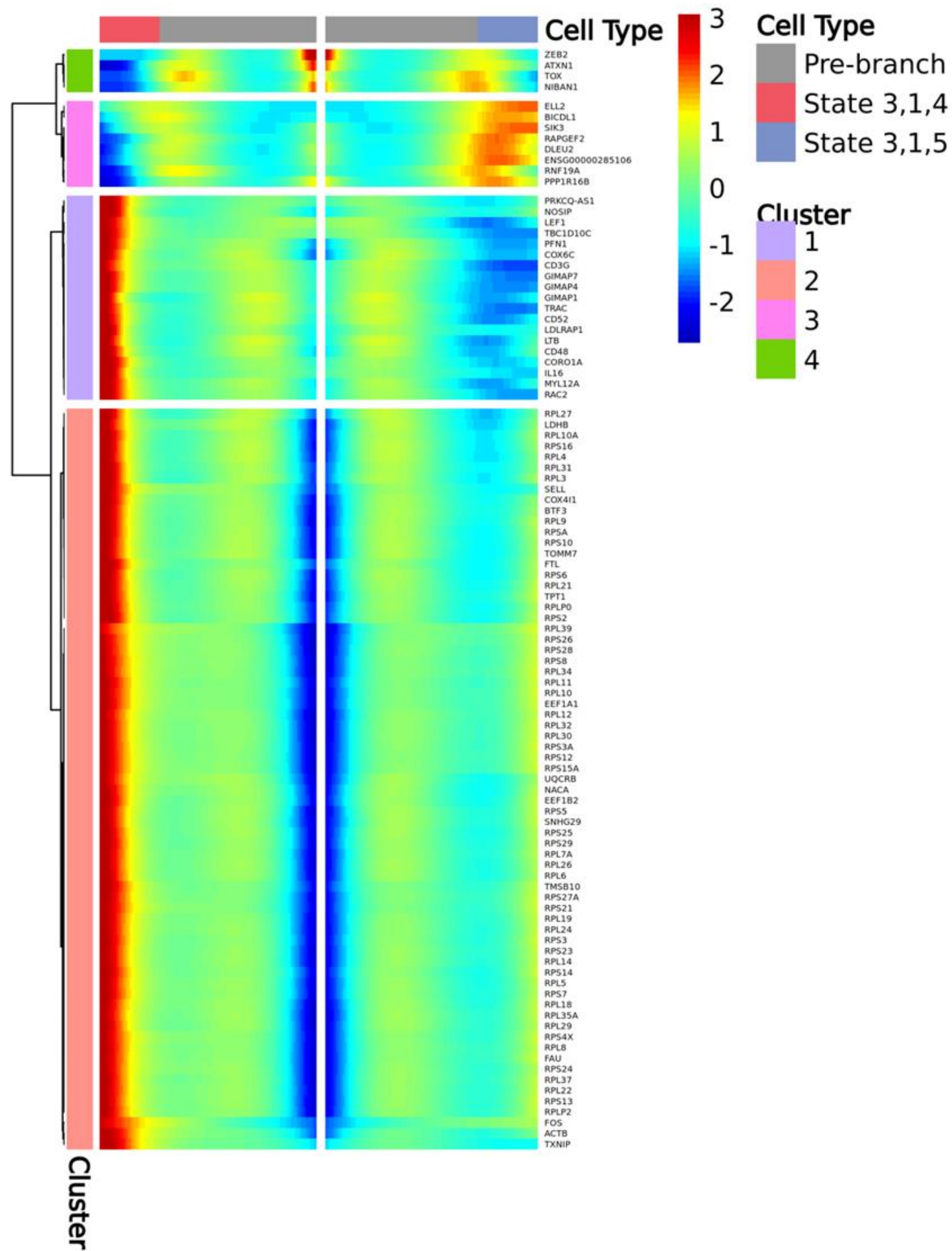
注：每个点的颜色表示不同的分化状态/细胞群/样本分组，横轴从左到右表示时间的早晚。从图中可以看出基因随时间发展的表达量变化。默认展示 Top10 基因随着拟定时间的表达变化。表达量越先出现上升趋势/下降的基因表示越早被激活/抑制。

拟时轨迹差异基因分析结果表

Gene_id	Gene name	Gene descript...	Pvalue	Padjust	Num_cells_Ex...	Use_for_Orde...	GO id	GO description	KO id
ENSG00000008...	ERO1B	endoplasmic ...	0	0	4275	1	GO:0006457;;...	protein foldin...	K10976
ENSG00000013...	NEK1	NIMA related...	0	0	4146	1	GO:0005737;;...	cytoplasm;;ce...	K08857
ENSG00000016...	SIK3	SIK family kin...	0	0	8145	1	GO:0005737;;...	cytoplasm;;p...	K19009
ENSG00000017...	RALGAPA1	Ral GTPase ac...	0	0	7079	1	GO:0051056;;...	regulation of ...	-----
ENSG00000010...	RAB11A	RAB11A, mem...	0	0	7324	1	GO:0055038;;...	recycling end...	K07904
ENSG00000018...	PDE4B	phosphodiest...	0	0	4651	1	GO:1901898;;...	negative regu...	K13293
ENSG00000017...	YES1	YES proto-onc...	0	0	2989	1	GO:0030154;;...	cell differenti...	K05705
ENSG00000007...	FOSL2	FOS like 2, AP...	0	0	2405	1	GO:0003700;;...	transcription f...	K09030
ENSG00000010...	RELB	RELB proto-on...	0	0	3534	1	GO:0032922;;...	circadian regu...	K09253
ENSG00000016...	TMEM184C	transmembra...	0	0	2425	1	GO:0016021	integral comp...	-----

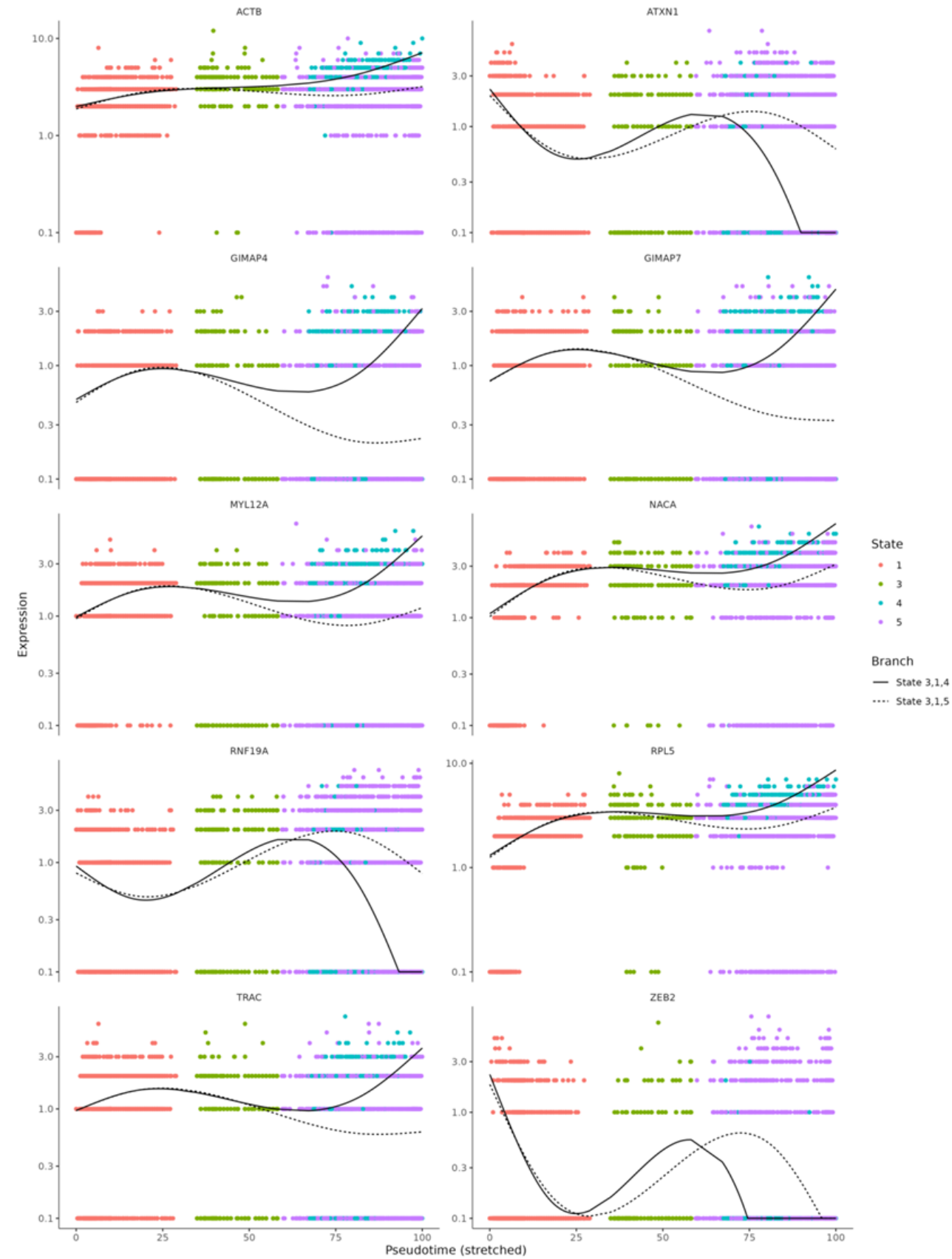
注：（1）Gene id：基因 id；（2）Gene name：基因名称；（3）Gene description：基因名称描述信息；（4）Pvalue：显著性水平；（5）Padjust：校正的 Pvalue 值；（6）Num_cells_Expressed：表达量细胞数目；（7）Use_for_Ordering：是否用于排序；（8）其他列为基因在各大数据库的注释情况。【默认只展示 GO 和 KEGG 2 个数据库的注释信息，可通过筛选表格对其他数据库进行勾选展示】表格中蓝色字体均可以点击链接到基因详情页！

特定分化分支点相关差异基因热图



注：沿着 pseudotime（在这个分支点产生的两条分化轨迹），变化显著的 100 个基因，将这些基因聚类可视化，可以观察在 拟时间过程中各个表达基因模块的变化（每个 cluster 代表一类表达趋势一致的基因集模块）。图中横坐标为拟时间顺序，每行代表一个基因，每列代表当前细胞状态下的平均表达值，颜色按照由红到蓝逐渐降低。

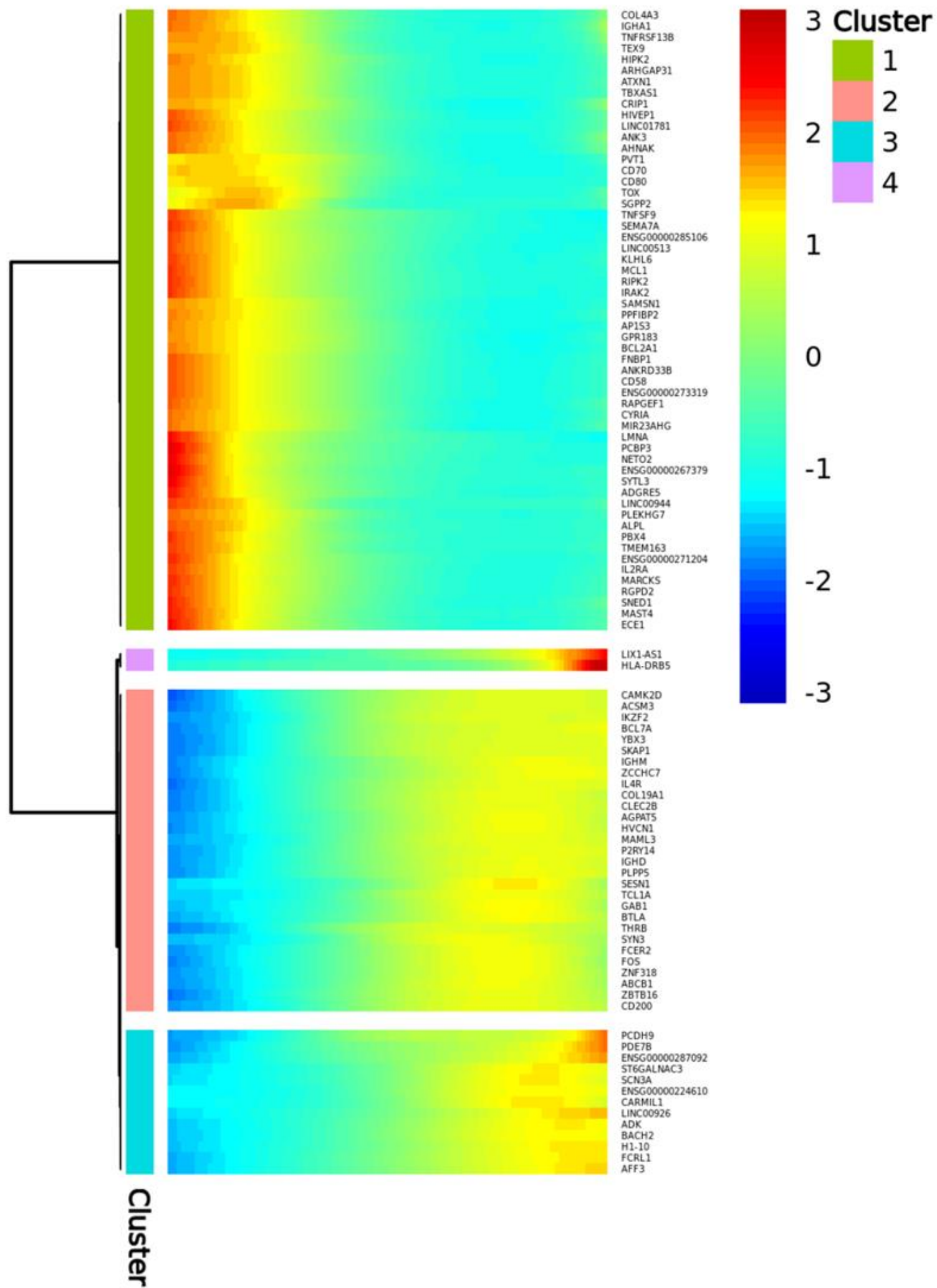
整体分化分支点差异基因表达趋势图



注：每个点的颜色表示不同的分化状态/细胞群/样本分组，横轴从左到右表示时间的早晚。从图中可以看出基因随时间发展的表达量变化。默认展示 Top10 基因随着拟定时间的表达变化。表达量越先出现上升趋势/下降的基因表示越早被激活/抑制。

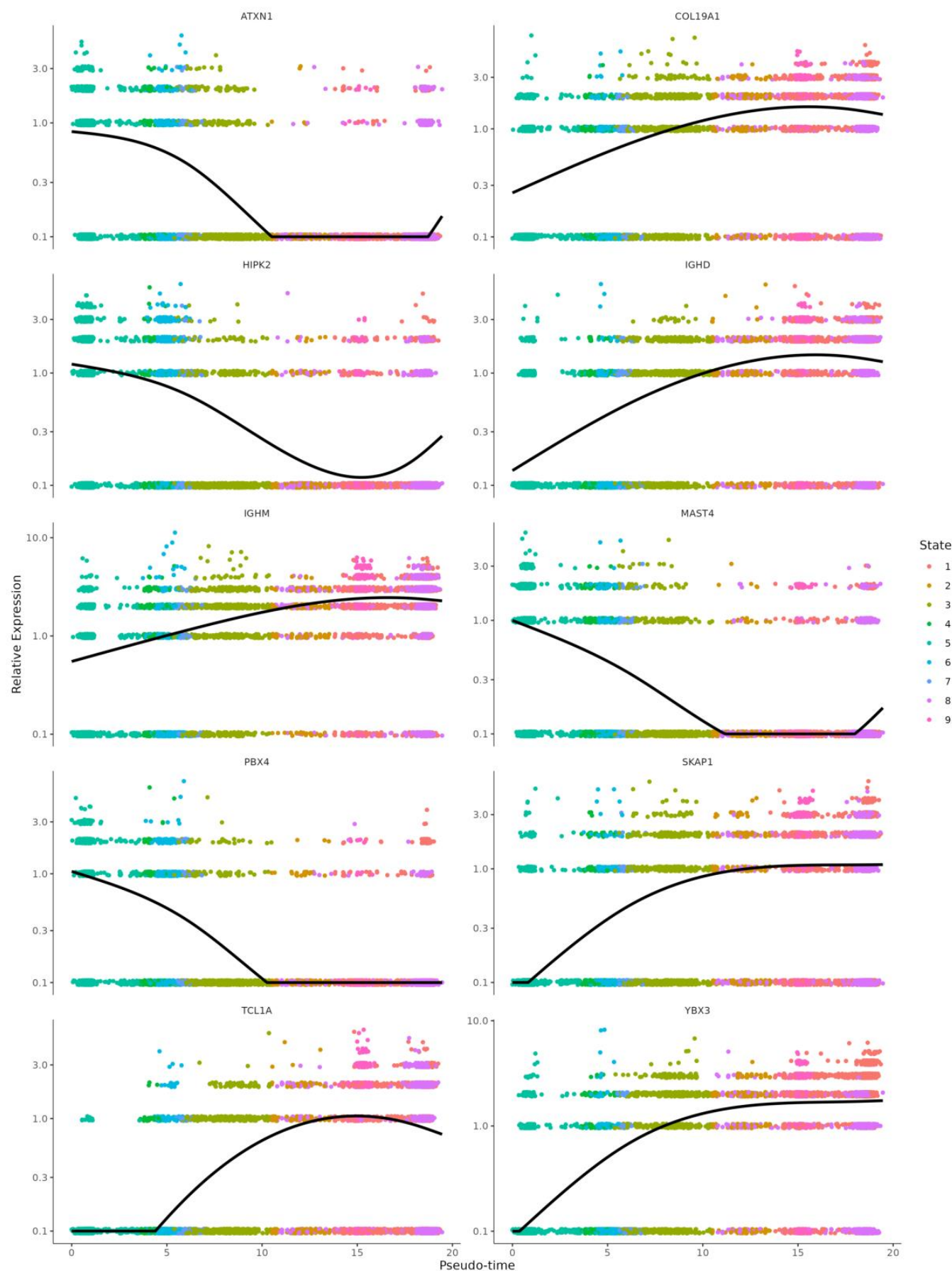
2.9.1.2 指定根节点拟时轨迹分析

特定拟时相关差异基因热图



注：选取分化过程中差异最显著的 100 个基因，将这些基因聚类可视化，可以观察在拟时间过程中各个表达基因模块的变化（每个 cluster 代表一类表达趋势一致的基因集模块）。图中横坐标为拟时间顺序，纵坐标表示不同的 Cluster 基因模块，每行代表一个基因，每列代表当前细胞状态下的平均表达值，颜色按照由红到蓝逐渐降低。【可切换 Top50 和 Top20 展示】

整体拟时相关差异基因表达趋势图



注：每个点的颜色表示不同的分化状态/细胞群/样本分组，横轴从左到右表示时间的早晚。从图中可以看出基因随时间发展的表达量变化。默认展示 Top10 基因随着拟定时间的表达变化。表达量越先出现上升趋势/下降的基因表示越早被激活/抑制。

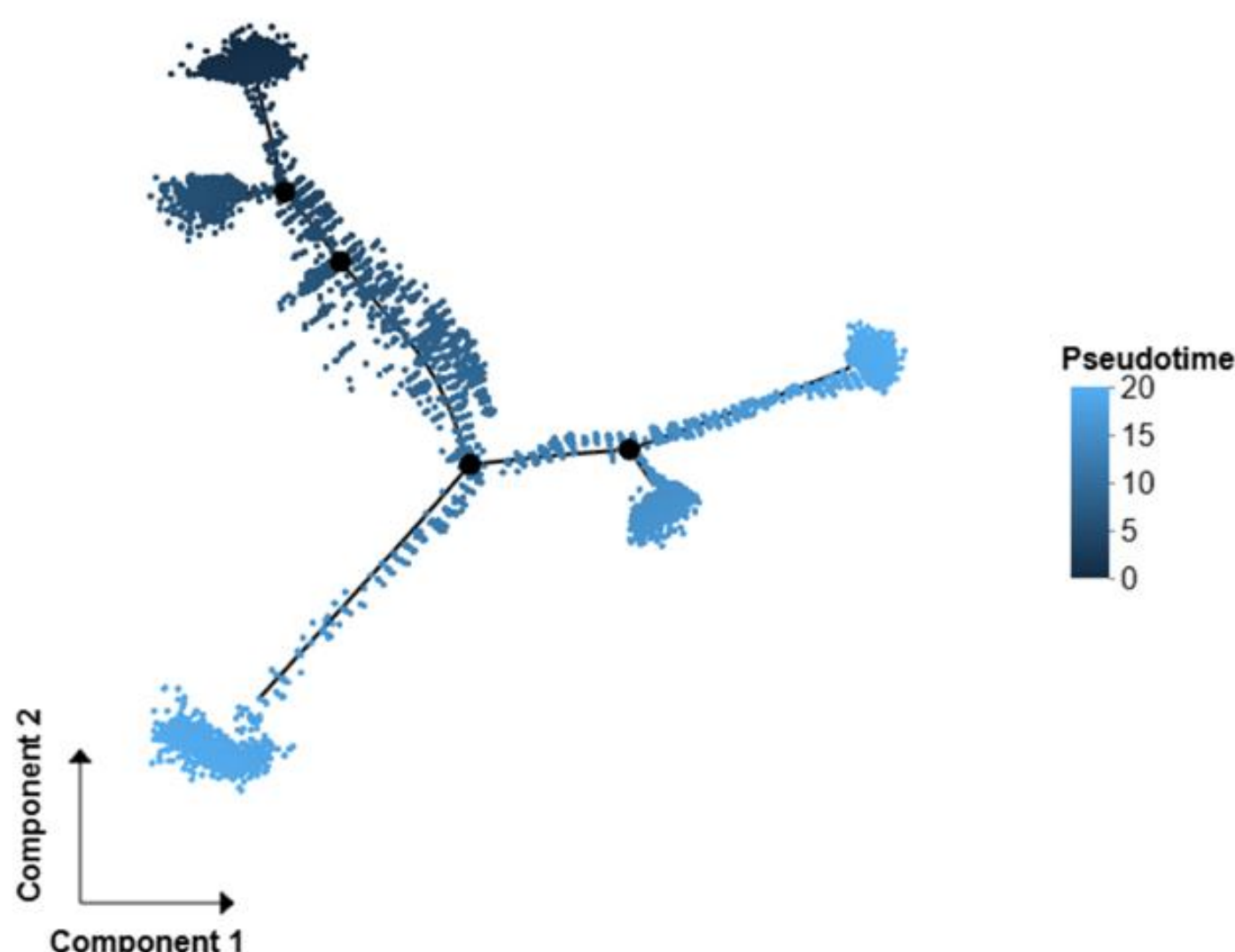
拟时轨迹差异基因分析结果表

Gene_id	Gene name	Gene descript...	Pvalue	Padjust	Num_cells_Ex...	Use_for_Orde...	GO id	GO description	KO id
ENSG0000014...	SKAP1	src kinase ass...	0	0	3184	1	GO:0034116;;...	positive regul...	K17699
ENSG0000013...	IL2RA	interleukin 2 r...	5.4663478361...	3.5926052723...	1379	1	GO:0042130;;...	negative regu...	K05068
ENSG0000011...	BCL7A	BAF chromati...	7.0102464970...	4.4112348968...	2104	1	GO:0045892	negative regu...	K25605
ENSG0000024...	TNFRSF13B	TNF receptor ...	2.1061959194...	4.1527162878...	1678	1	GO:0038023;;...	signaling rece...	K05150
ENSG0000012...	ATXN1	ataxin 1 [Sour...	0	0	1689	1	GO:0008022;;...	protein C-term...	K23616
ENSG0000016...	FCRL1	Fc receptor lik...	5.9739534269...	7.6817248957...	5311	1	GO:0009986;;...	cell surface;;...	K06727
ENSG0000006...	YBX3	Y-box binding...	0	0	4247	1	GO:1902219;;...	negative regu...	K06099
ENSG0000006...	MAST4	microtubule a...	0	0	1720	1	GO:0005737;;...	cytoplasm;;p...	K08789
ENSG0000019...	TOX	thymocyte se...	1.6964963244...	1.7919242427...	968	1	GO:0032825;;...	positive regul...	-----
ENSG0000014...	CAMK2D	calcium/calm...	1.4821969375...	3.1313880633...	5175	1	GO:0018107;;...	peptidyl-thre...	K04515

注：（1）Gene id：基因 id；（2）Gene name：基因名称；（3）Gene description：基因名称描述信息；（4）Pvalue：显著性水平；（5）Padjust：校正的 Pvalue 值；（6）Num_cells_Expressed：表达量细胞数目；（7）Use_for_Ordering：是否用于排序；（8）其他列为基因在各大数据库的注释情况。【默认只展示 GO 和 KEGG 2 个数据

库的注释信息，可通过筛选表格对其他数据库进行勾选展示】表格中蓝色字体均可以点击链接到基因详情页！

指定根节点拟时轨迹图



注：（1）按 Pseudotime 展示，Pseudotime 是一种用来衡量每个细胞在发育轨迹中的相对位置或虚拟时间的指标，它反映了每个细胞的发育进程和分化状态。图中的点代表单个细胞，它们根据 Pseudotime 值沿着一条路径排列。这条路径反映了细胞从一种状态向另一种状态转变的过程。细胞的颜色通常从深蓝色（表示 Pseudotime 较低，即较早的发育阶段）渐变到浅蓝色（表示 Pseudotime 较高，即较晚的发育阶段）。横纵坐标分别为两个主成分，图中每个圆点代表一个细胞，图中黑色圆圈内的数字代表轨迹分析中确定不同细胞状态的分支节点。【时间为软件计算拟合得到，不一定是真实的分化方向，需要结合实际项目背景进一步判断】

（2）按 State 展示：State 表示细胞的不同状态或阶段，例如在细胞分化或发育过程中的不同点。【此外也可切换按照样本、分组和 celltype 等维度进行展示】

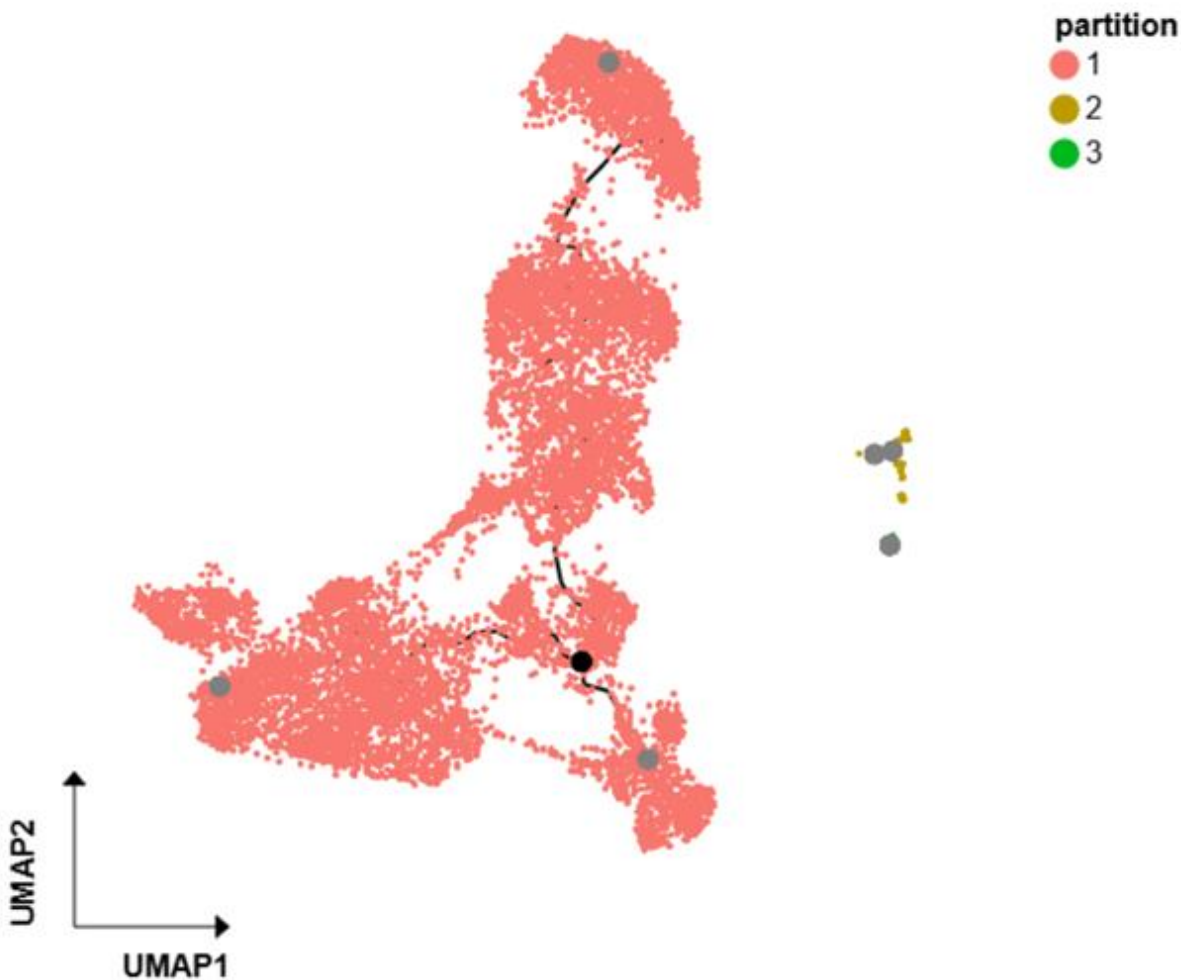
2.9.2 拟时序分析-Monocle3

Monocle3 是一个用于拟时序分析的 R 包，Monocle3 正式继承了 Monocle2 中迭代的灵魂算法，并舍弃了其中被诟病的轨迹图生成部分，引入了以 PAGA 为蓝本的基于图谱的轨迹推断流程，成为了目前单细胞拟时序分析工具中的“天花板”。Monocle 3 的核心算法具有高度可扩展性，可以处理数百万个细胞。软件主要可以

执行三类分析：1) 对细胞聚类、分群和计数；2) 构建单细胞轨迹；3) 差异表达分析。

2.9.2.1 拟时轨迹初步结果

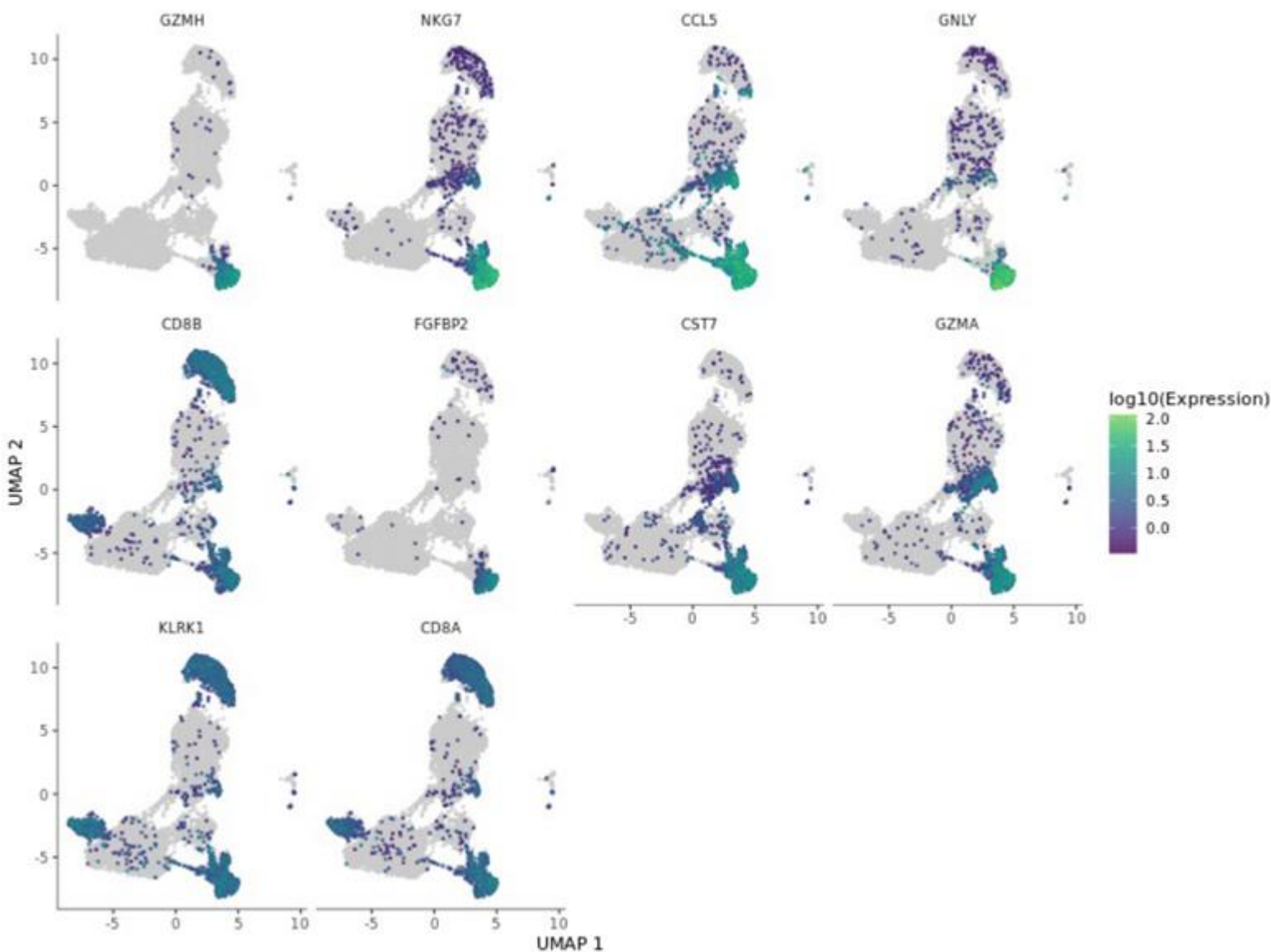
细胞分化轨迹图-整体展示-partition 水平



注：图中每个点代表一个细胞，具有相似状态的细胞被聚到一起，颜色标注不同的细胞群。黑色线条表示分化轨迹。黑色圆圈表示树干节点（branch nodes），表示从这个点开始，细胞可以有多个结局；灰色圆圈表示分化树叶节点（leaf nodes），即细胞的不同结局。图例表示不同的 partition。

2.9.2.2 指定根节点拟时轨迹结果

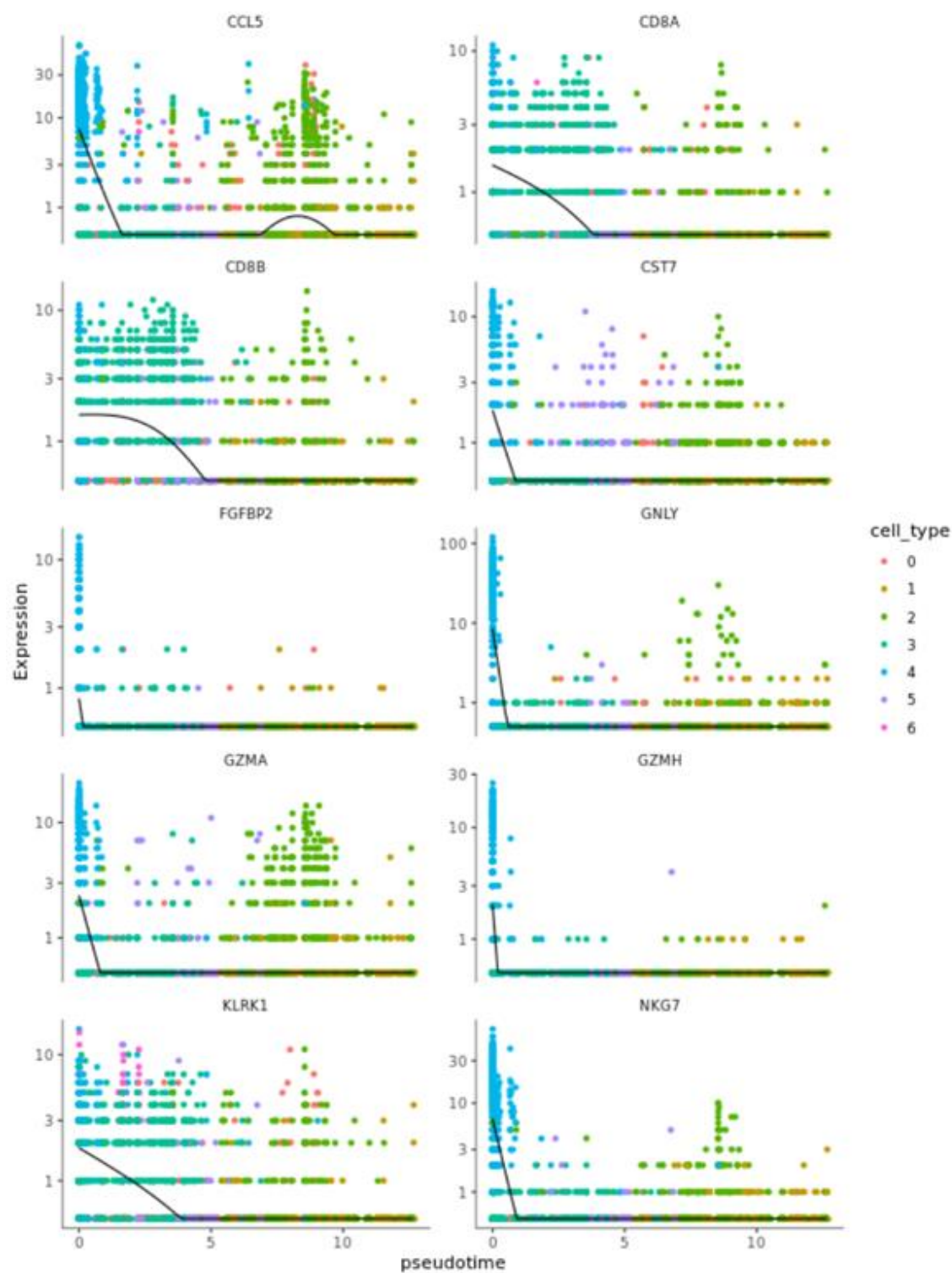
拟时间依赖基因 UMAP 图



注：该图反应同一基因在不同细胞群中的差异表达情况。背景为不同细胞群，

每个点代表基因在不同细胞中的表达情况，颜色的深浅表示基因表达量的高低，灰色部分表示该基因没有在这些细胞内进行表达。

拟时间依赖基因表达趋势图



注：每个点的颜色表示不同的细胞群，横轴从左到右表示时间的早晚。从图中可以看出基因随时间发展的表达量变化。默认展示 top10 基因随着拟定时间的表达变化。表达量越先出现上升趋势/下降的基因表示越早被激活/抑制。

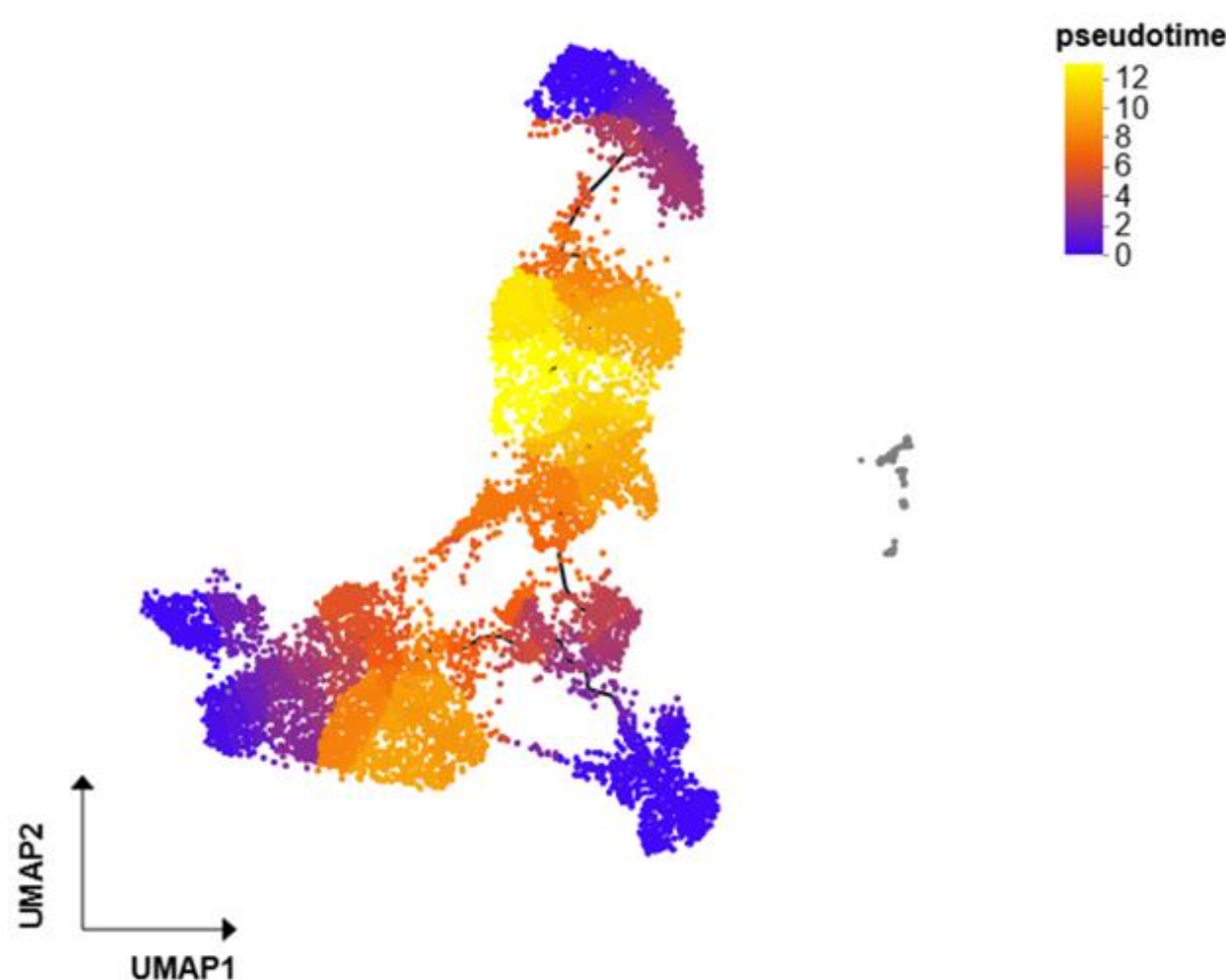
拟时间依赖基因结果表

Gene id	Gene name	Gene descript...	Pvalue	Padjust	Morans test st...	Morans I	Status	GO id	GO description
ENSG0000010...	NKG7	natural killer ...	0	0	262.74616864...	0.7882907948...	OK	GO:0005887;;...	integral comp...
ENSG0000013...	FGFBP2	fibroblast gro...	0	0	235.70127427...	0.7065295885...	OK	GO:0007267;;...	cell-cell signa...
ENSG0000019...	TOX	thymocyte se...	0	0	167.11439988...	0.5015043427...	OK	GO:0032825;;...	positive regul...
ENSG0000010...	GZMH	granzyme H [...	0	0	270.30018575...	0.8104124939...	OK	GO:0019835;;...	cytolysis;;seri...
ENSG0000011...	PLEK	pleckstrin [So...	0	0	194.70793670...	0.5839269584...	OK	GO:0005829;;...	cytosol;;cell p...
ENSG0000009...	EZR	ezrin [Source:...	0	0	171.48567315...	0.5146911147...	OK	GO:0045177;;...	apical part of...
ENSG0000017...	CD8B	CD8b molecu...	0	0	235.55189118...	0.7069705717...	OK	GO:0050776;;...	regulation of ...
ENSG0000014...	GZMA	granzyme A [...	0	0	221.26528748...	0.6639036784...	OK	GO:0019835;;...	cytolysis;;seri...
ENSG0000007...	CST7	cystatin F [Sou...	0	0	228.26021440...	0.6848849910...	OK	GO:0005794;;...	Golgi apparat...
ENSG0000017...	HOPX	HOP homeob...	0	0	161.64269338...	0.4848718356...	OK	GO:0005737;;...	cytoplasm;;ch...

注：（1）Gene id：基因 id；（2）Gene name：基因名称；（3）Gene description：基因名称描述信息；（4）Pvalue：显著性水平；（5）Padjust：校正的 Pvalue 值；（6）Morans_test_statistic：与拟时序相关的检验统计量；（7）Morans_I：Moran 的 I 统计量是一种多向多维空间自相关的度量。统计数据通过最近邻图对数据点之间的空间关系进行编码，使其特别适合于分析大型 scRNA-seq 数据集。Moran 指数的范围

为-1 到 1 之间，0 表示在空间中不相关，而 1 表示高度正相关，小于 0 的 Moran 指数一般都没有统计学意义；（8）Status：状态；（9）其他列为基因在各大数据库的注释情况。【默认只展示 GO 和 EKG2 个数据库的注释信息和所有样本的表达量信息，可通过筛选表格对其他数据库或者目标样本的表达量进行勾选展示】表格中蓝色字体均可以点击链接到基因详情页！

指定根节点拟时轨迹图-整体展示



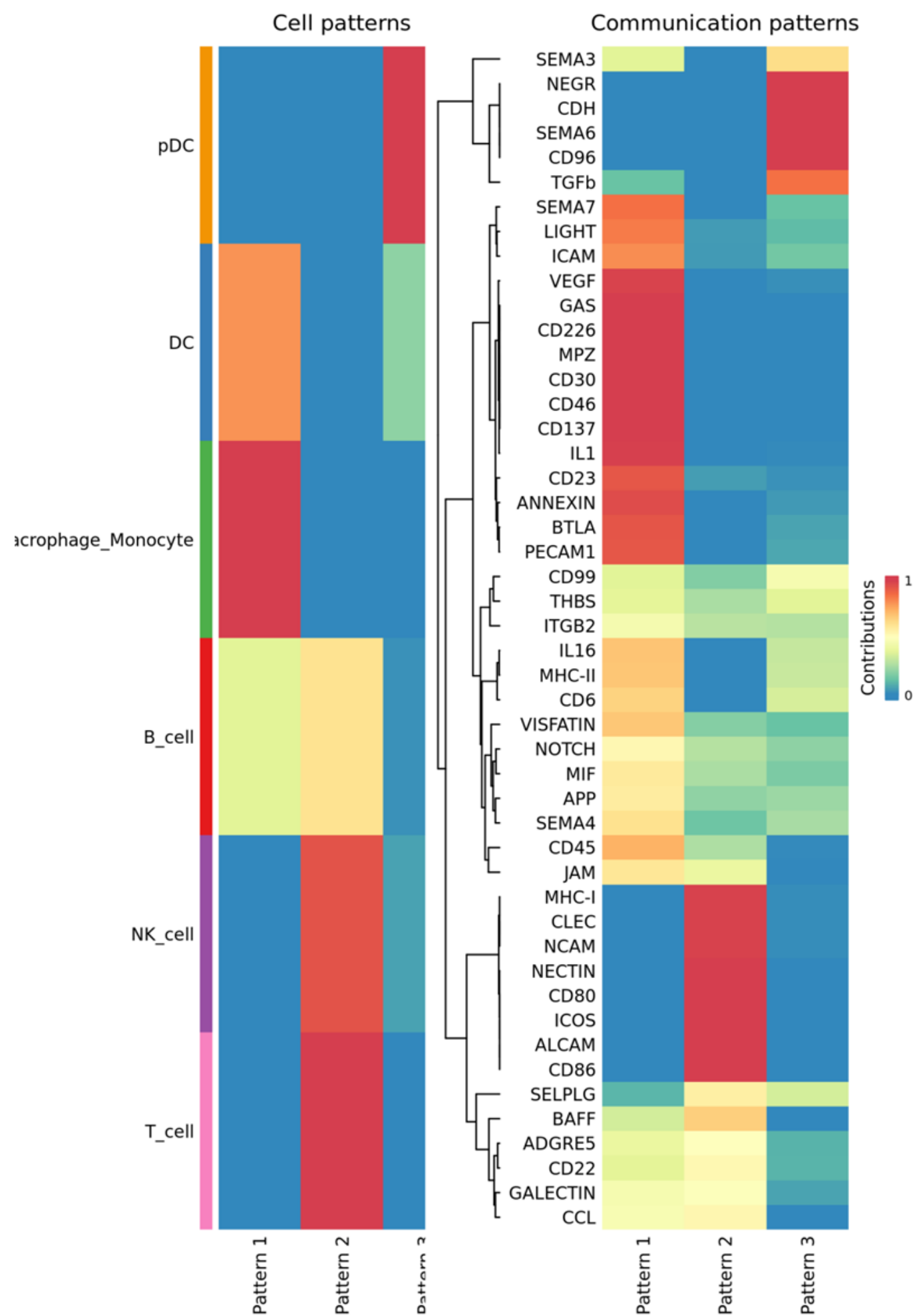
注：颜色由紫色到黄色，表示分化的时间由早及晚。数字表示分化的过程，数字越小，表示分化的时间越早。另外，如果自己指定的起始点，无法直接访问到某一部分细胞，这部分细胞就有无穷伪时间，用灰色表示。

2.9.3 细胞通讯分析-CellChat

细胞之间的相互通讯极为复杂，一般由激素、生长因子、趋化因子、细胞因子和神经递质等配体介导细胞之间的信息交流，主要分为自分泌，旁分泌，近分泌和内分泌等。当前基于单细胞转录组测序，进行细胞通讯分析越来越常见。CellChat（基于 R）细胞通讯分析可以给我们一些细胞类群之间相互调控/交流的信息，这种细胞之间的调控主要是通过受配体结合，传递信号来实现的。不同的分化、疾病过程，可能存在特异的细胞通讯关系，因此阐明这些通讯关系至关重要。

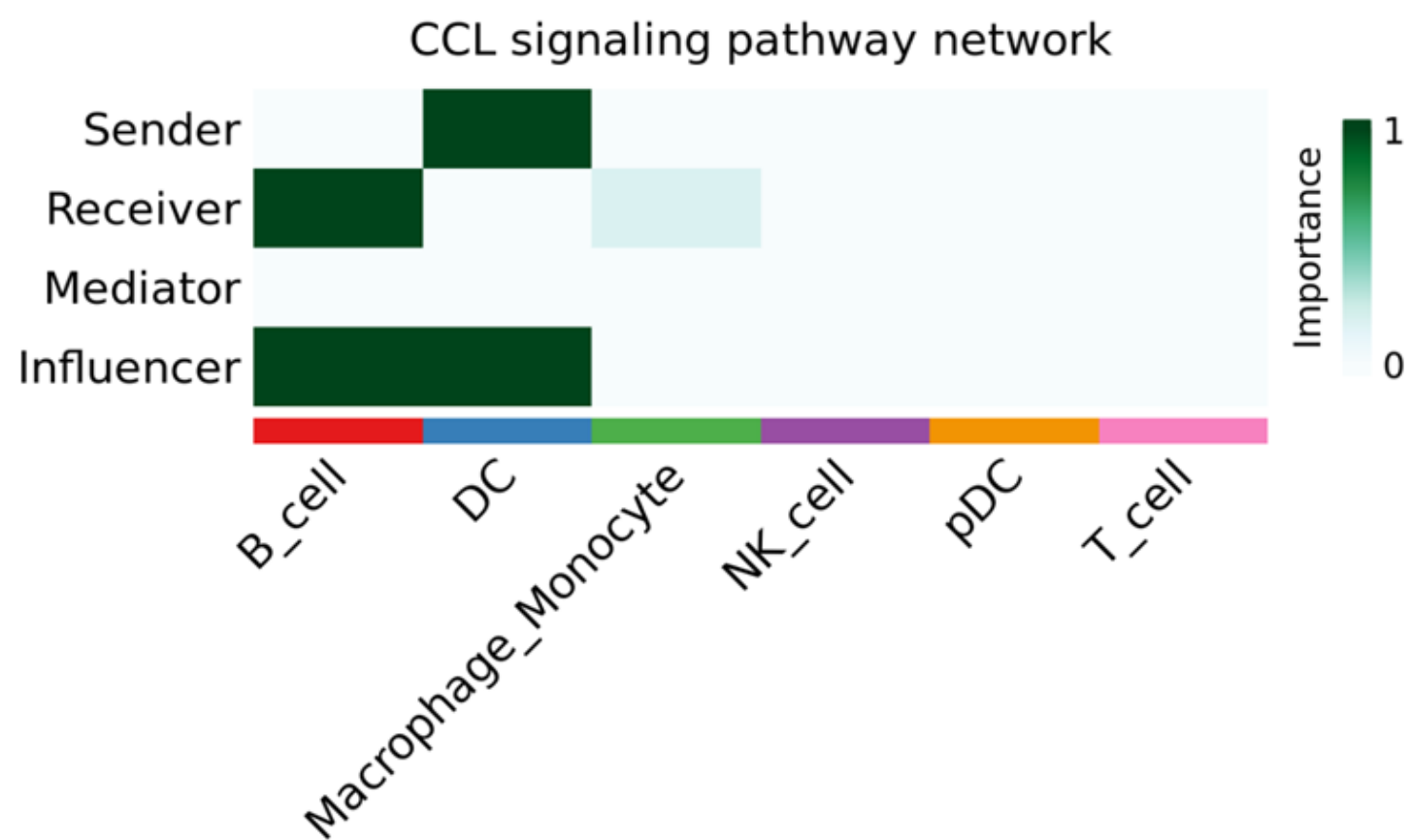
2.9.3.1 细胞通讯分析

通讯模式聚类图-信号输入热图



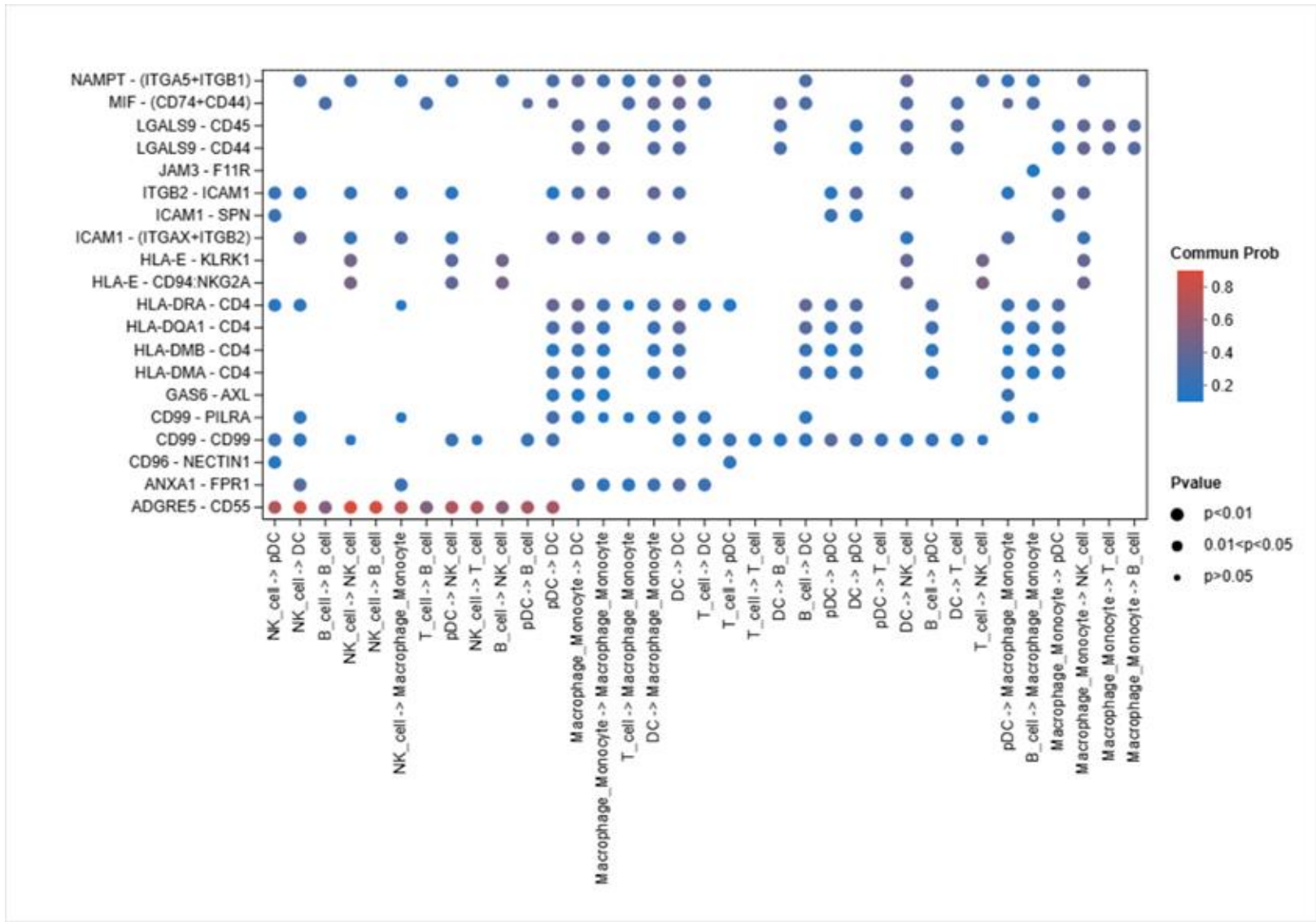
注：传入模式显示目标细胞（即作为信号接收器的细胞）如何相互协调以及它们如何与某些信号通路协调以响应传入信号。CellChat 采用模式识别方法来识别全局通信模式，横坐标为计算出的模式数量 pattern1、pattern2、pattern3...（其实类似于聚类的 cluster 数量），然后分别对细胞类型与信号通路进行聚类。按计算出的 pattern 值分为了 2 个大的 pattern，左图（Cell patterns）纵轴是细胞类型，右图（Communication patterns）纵轴是信号通路，也是被分为两个大的 pattern。颜色越红表示通路也就越活跃。

在 Pathway 中细胞类型的信号角色热图



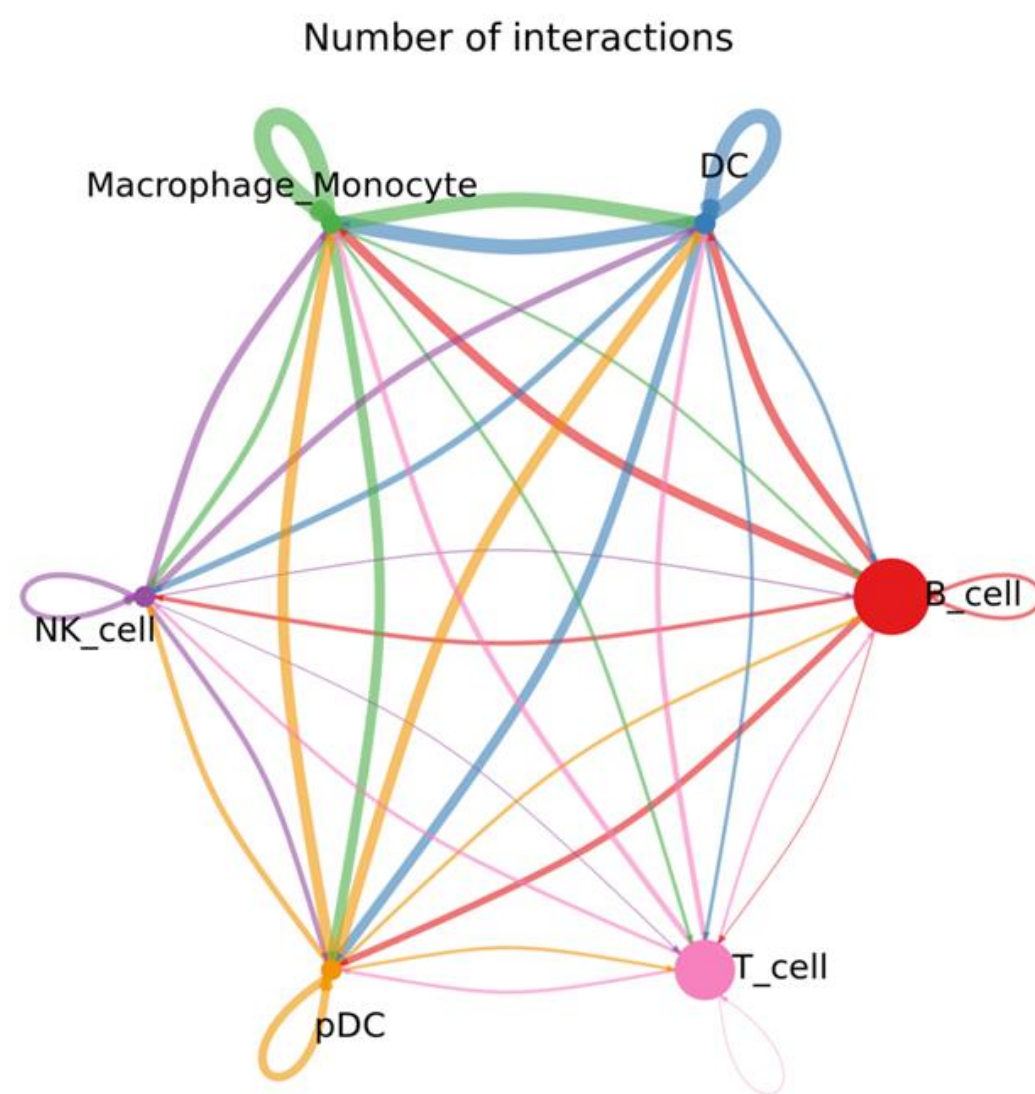
注：通过计算每个细胞群的网络中心性指标，CellChat 可以识别细胞间通信网络中的主要发送者（Sender）、接收者（Receiver）、调解者（Mediator）和影响者（Influencer），从而分析细胞在信号网络中扮演的角色。颜色深浅表示该细胞类型为某角色的概率大小。

多个配体-受体介导的细胞互作气泡图



注：横坐标是细胞类型互作，默认用“->”进行连接，前面为配体，后面为受体，eg: B_cell -> T_cells 即 B_cell 为配体， T_cells 为受体；纵坐标是配受体关系对，默认用“-”连接，前面为配体，后面为受体，eg: CCL3 - CCR1 即 CCL3 为配体， CCR1 为受体。点的颜色和大小分别表示配体-受体对介导细胞通讯的可能性和及该可能性对应的 Pvalue，空意味着在该数据集中，对应的配体-受体不介导该细胞通讯。点越大表示越显著，颜色越红表示可能性越大。

细胞间互作数量和强度网络图



注：细胞通讯网络（intercellular communication network）是一个由相互作用的细胞组之间的显著的配体-受体对组成的加权有向图，显示不同的细胞组之间检测到配体-受体相互作用的数量和强度。图中，不同颜色的实心圆表示不同细胞组，发出箭头的节点表示配体，箭头指向的节点表示受体。细胞间互作数量网络图，图中的节点表示不同细胞类型，节点圆圈越大表示该种细胞类型的细胞数目越多，线条越粗表示不同细胞类型之间检测到的互作数量越多，线条的颜色与配体细胞的颜色一致。细胞间互作强度网络图，节点表示不同细胞类型，线条越粗表示细胞类型之间的互作强度越高，线条的颜色与配体细胞的颜色一致。【可切换可视化展示方式：①细胞间互作数量网络图整体展示；②细胞间互作强度网络图整体展示；③细胞间互作数量网络图分亚群展示；④细胞间互作强度网络图分亚群展示】

温馨提示：当某个细胞类型作为配体没有匹配的受体时，分亚群展示的时候将进行过滤。

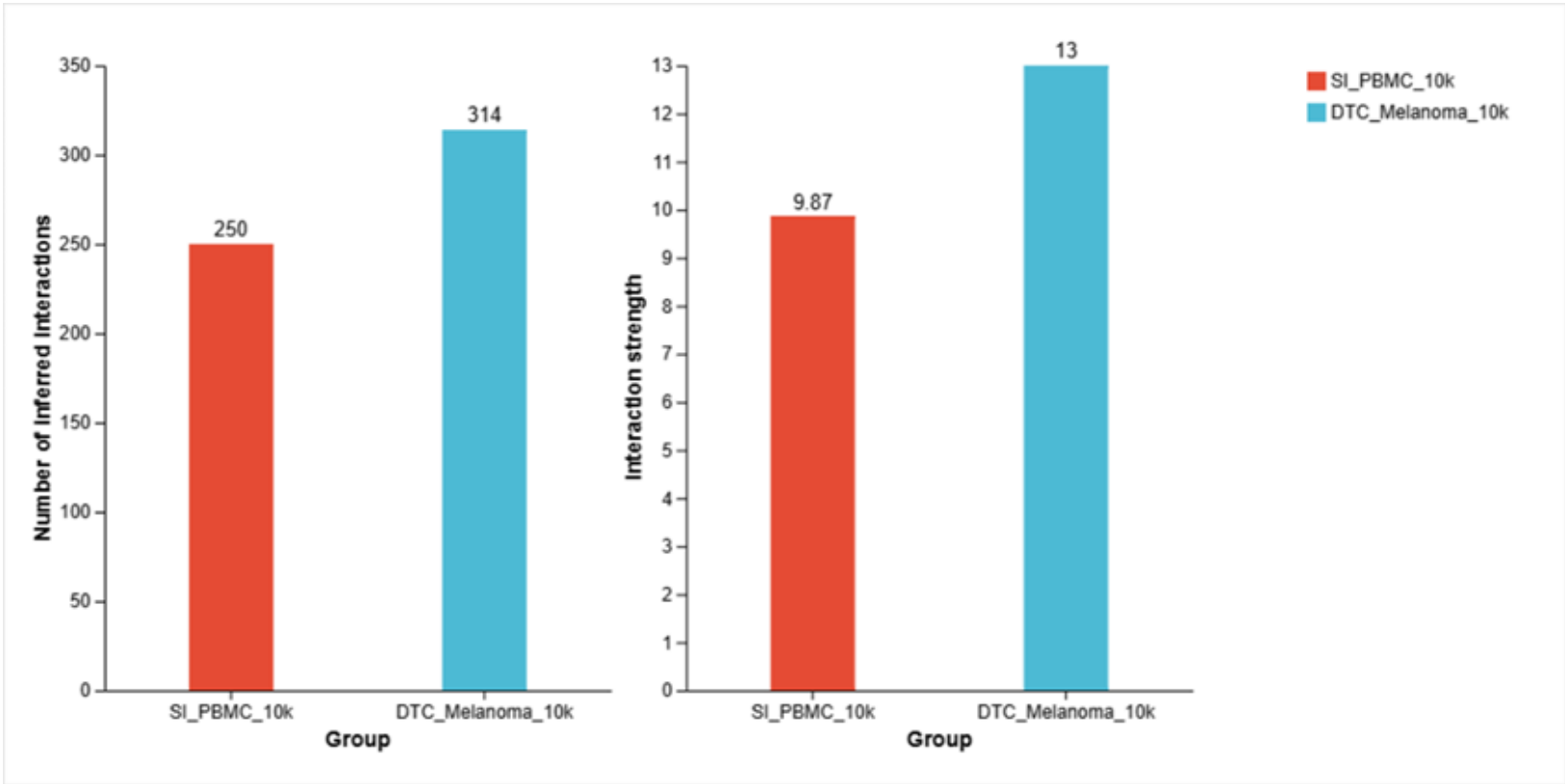
受配体关系对数目及权重统计表

Type I	Type II	Count	Weight
Macrophage_Monocyte	DC	35	1.4728446692401
NK_cell	DC	16	0.69196749776218
pDC	B_cell	8	0.76494827047455
DC	DC	35	1.7940146108158
T_cell	B_cell	6	0.3750545544454
NK_cell	B_cell	3	0.40604622634651
DC	B_cell	9	0.42567693703631
B_cell	B_cell	9	0.47485095545677
Macrophage_Monocyte	B_cell	7	0.23471884876876
B_cell	DC	21	0.94990605187487

注：（1）Type I：细胞类型 I，表示配体；（2）TypeII：细胞类型 II，表示受体；（3）Count：受配体关系对个数；（4）Weight：受配体关系对权重。

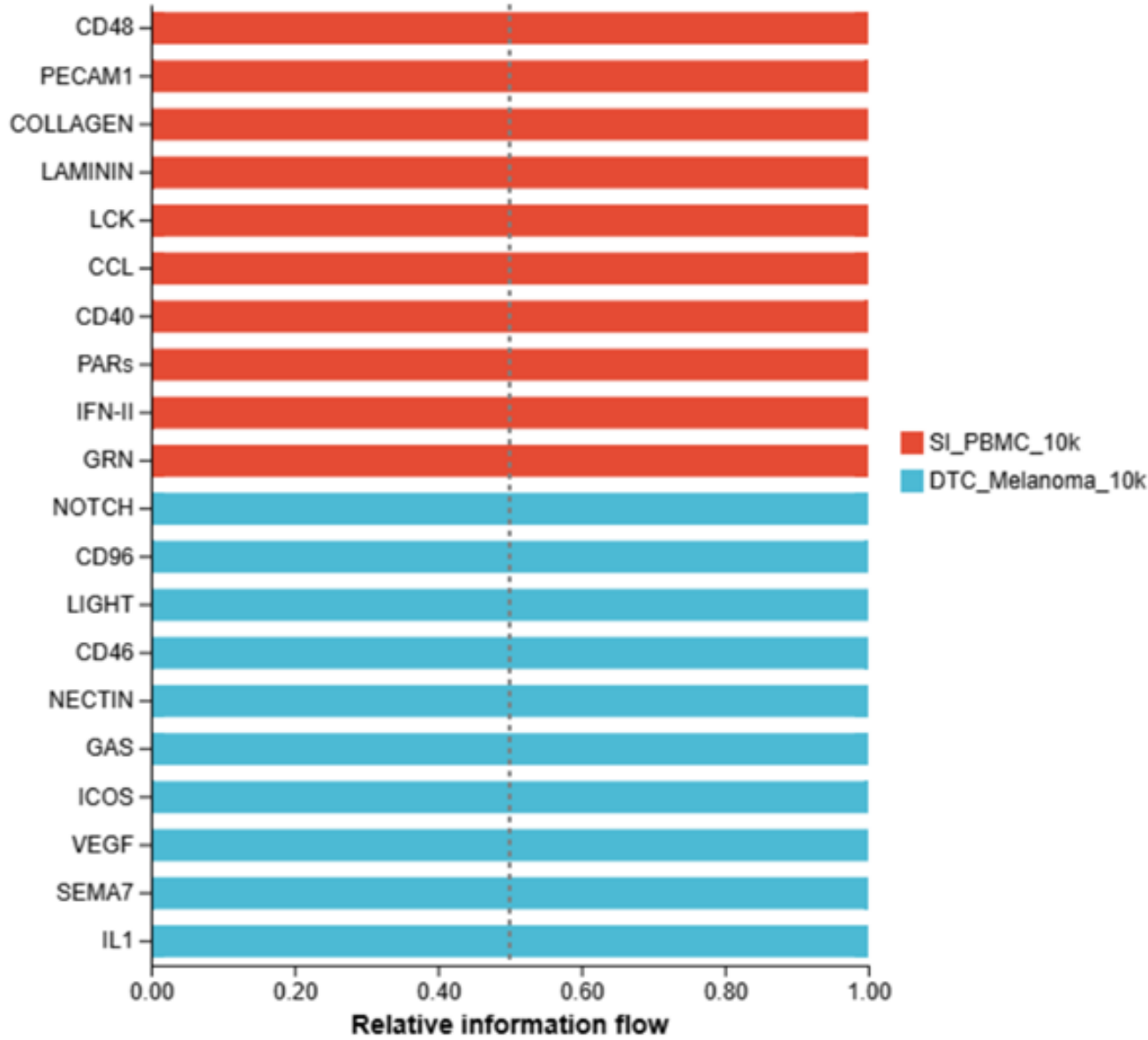
2.9.3.2 细胞通讯差异分析

相互作用数量及作用强度比较图-柱状图



注：细胞间互作数量和强度柱状对比图，比较交互数目和交互强度，左图展示不同组间通讯数量差异，右图展示不同组间的通讯强度差异。

比较各个信号通路的总体信息流图-堆积图



注：根据推断两个组别通讯网络之间整体信息流的差异，对重要的信号通路进行排名。红蓝色柱子分别表示在两组中富集的信号通路。**eg**：比如差异分组方案为 A_VS_B 时，当处理组（A）与对照组（B）的通路概率总和的比值 >1.05 且差异 $Pvalue<0.05$ （秩和检验）时，则该通路在处理组中的通讯强度显著增加（纵坐标为红色）；当比较组与对照组的通路概率总和的比值 <0.95 且差异 $Pvalue<0.05$ 时，则该通路在对照组中的通讯强度显著增加（纵坐标为蓝色）；纵坐标为黑色表示该通路在两组间没有差异。堆积图为比例图，柱形图为实际数值比对图。

细胞通讯差异详情表

source	target	ligand	receptor	prob	pval	interaction_na...	interaction_na...	pathway_nam...	annotation
B_cell	B_cell	ADGRE5	CD55	0.5477284775...	3	ADGRE5_CD55	ADGRE5 - CD55	ADGRE5	Cell-Cell Con...
NK_cell	B_cell	ADGRE5	CD55	0.8619549077...	3	ADGRE5_CD55	ADGRE5 - CD55	ADGRE5	Cell-Cell Con...
pDC	B_cell	ADGRE5	CD55	0.6870263346...	3	ADGRE5_CD55	ADGRE5 - CD55	ADGRE5	Cell-Cell Con...
T_cell	B_cell	ADGRE5	CD55	0.5355212277...	3	ADGRE5_CD55	ADGRE5 - CD55	ADGRE5	Cell-Cell Con...
NK_cell	Macrophage_...	ADGRE5	CD55	0.7621402050...	3	ADGRE5_CD55	ADGRE5 - CD55	ADGRE5	Cell-Cell Con...
B_cell	NK_cell	ADGRE5	CD55	0.5620996682...	3	ADGRE5_CD55	ADGRE5 - CD55	ADGRE5	Cell-Cell Con...
NK_cell	NK_cell	ADGRE5	CD55	0.8911741209...	3	ADGRE5_CD55	ADGRE5 - CD55	ADGRE5	Cell-Cell Con...
pDC	NK_cell	ADGRE5	CD55	0.7077277984...	3	ADGRE5_CD55	ADGRE5 - CD55	ADGRE5	Cell-Cell Con...
NK_cell	pDC	ADGRE5	CD55	0.7103470309...	3	ADGRE5_CD55	ADGRE5 - CD55	ADGRE5	Cell-Cell Con...
NK_cell	T_cell	ADGRE5	CD55	0.7131334451...	3	ADGRE5_CD55	ADGRE5 - CD55	ADGRE5	Cell-Cell Con...

注：（1）source：发送信号源亚群；（2）target：接收信号亚群；（3）ligand：配体，多个基因的为复合物 Heterodimers；（4）receptor：受体，多个基因的为复合物；（5）prob：根据表达值，推测互作的概率；（6）pval：pval分为3类（1,2,3）， $[pval>0.05]=1$ ， $[0.01 < pval \leq 0.05]=2$ ， $[pval\leq 0.01]=3$ ；（7）interaction_name：唯一标识，由 ligand-receptor 名构成；（8）interaction_name2：interaction_name 别名；（9）pathway_name：由受配体基因分类的 KEGG pathway，目前共收录 229 个 signaling pathway；（10）annotation：对受配体信号的分类注释，分3类相互作用，旁分泌/自分泌信号（Secreted Signaling）、细胞外基质受体（ECM-Receptor）、细胞接触（Cell-Cell Contact）；（11）evidence：数据来源，KEGG 或文献中；（12）source.target：发送者与接收者；（13）prob.original：原始互作概率， $prob=-1/\log(prob.original)$ ；（14）dataset：分组信息。

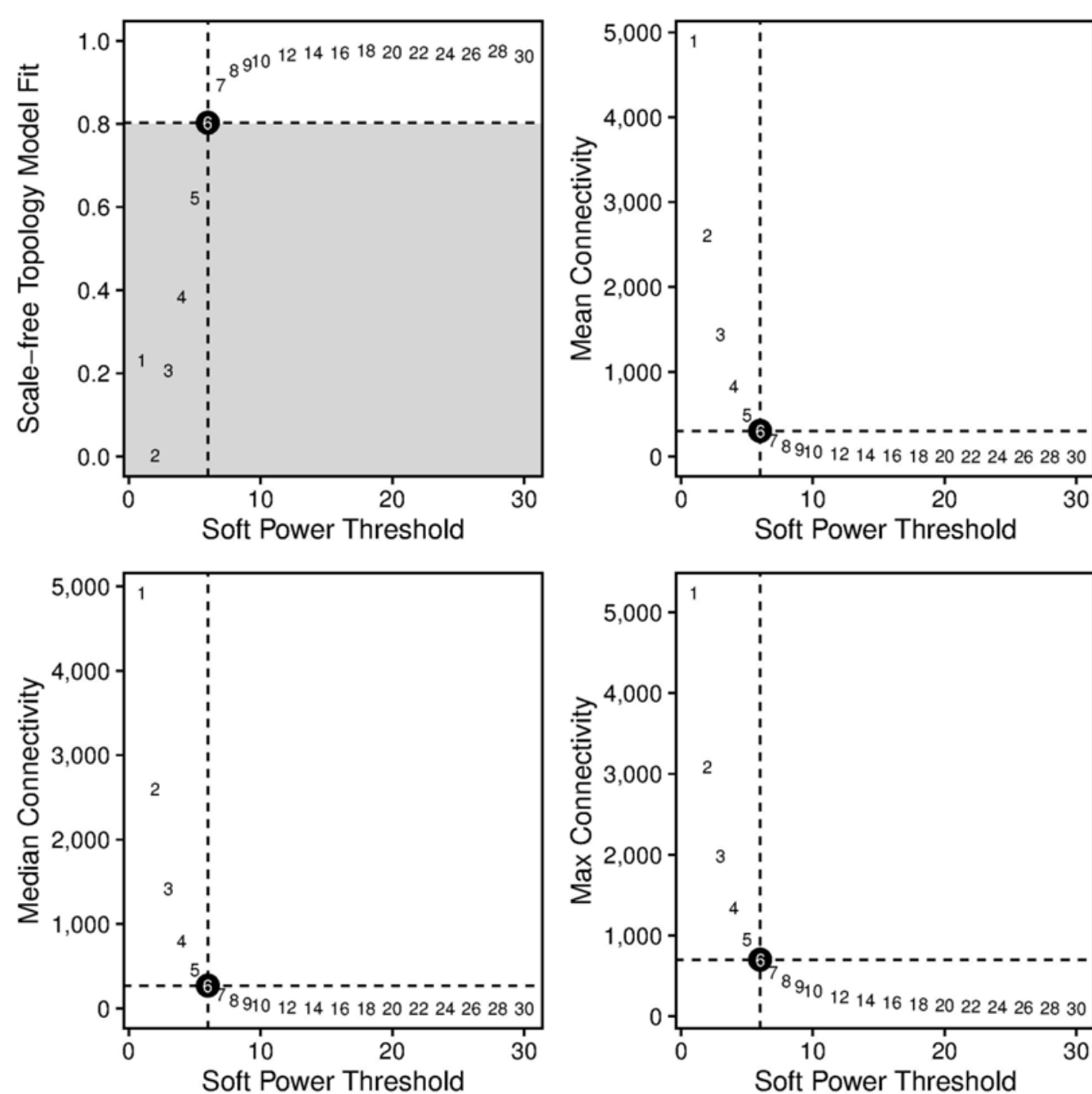
2.9.4 hdWGCNA 分析

WGCNA（weighted gene co-expression network analysis，加权基因共表达网络分析）是一种分析样本基因表达模式的分析方法，基于基因表达水平构建一个网络，根据基因间的共表达关系，鉴定表达模式相似的基因集合（modules），这些

基因可能存在共调控、功能相关或处于同一通路；并分析模块与特定性状或表型之间的关联关系，在疾病以及其他性状与基因关联分析等方面的研究中被广泛应用。但单细胞数据具有严重的稀疏性，数据纬度高计算量大，常规方法已无法分析单细胞数据，这里使用在单细胞 RNA-seq 或空间转录组学等高维转录组学数据中执行加权基因共表达网络分析(WGCNA) 的 R 包——hdWGCNA 进行此分析。hdWGCNA 采用了 MetaCell 概念(高度相似的细胞组)，通过 KNN 算法将输入数据集中邻近的 K 个细胞合并成一个 metacell，能够在降低数据稀疏性的同时也保留细胞间的异质性。

2.9.4.1 数据预处理

无尺度容适曲线和连通度曲线

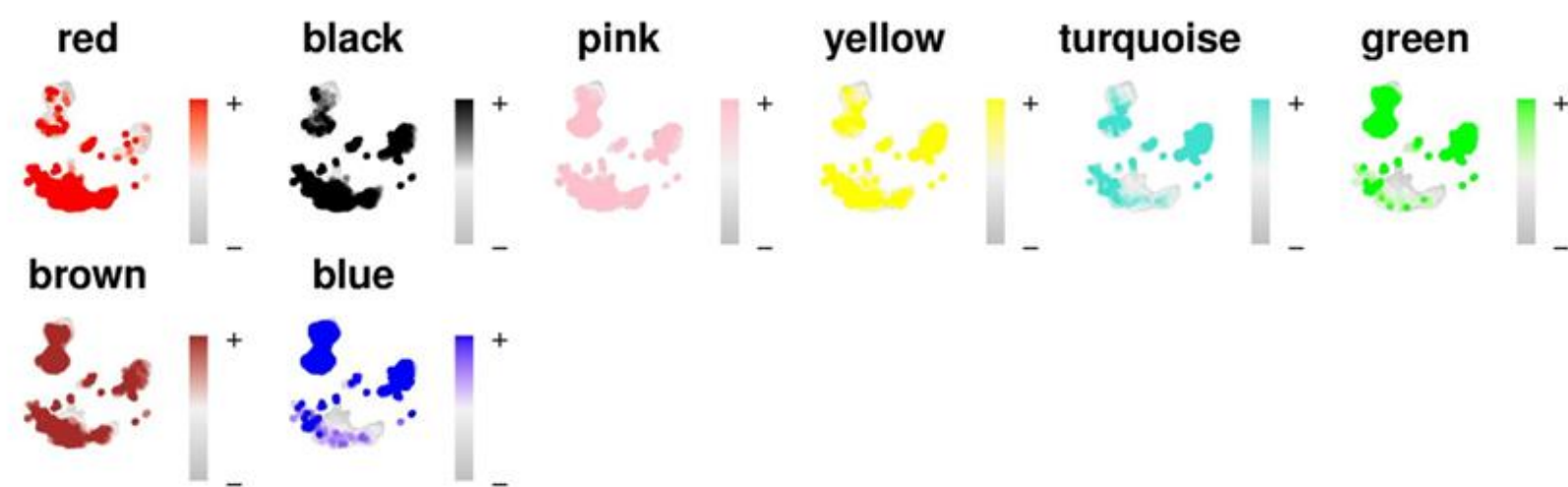


注：左上图：横坐标表示幂指数加权 β 值；纵坐标表示对应 β 值转化后的邻接矩阵与无尺度网络假设的拟合程度 R^2 。 R^2 越大，网络越符合无尺度分布。右上，左下，右下图：横坐标表示幂指数加权 β 值；纵

坐标表示对应 β 值转化后的邻接矩阵所代表网络中每个节点的平均/最小/最大连通度 (k/degree/connectivity)。

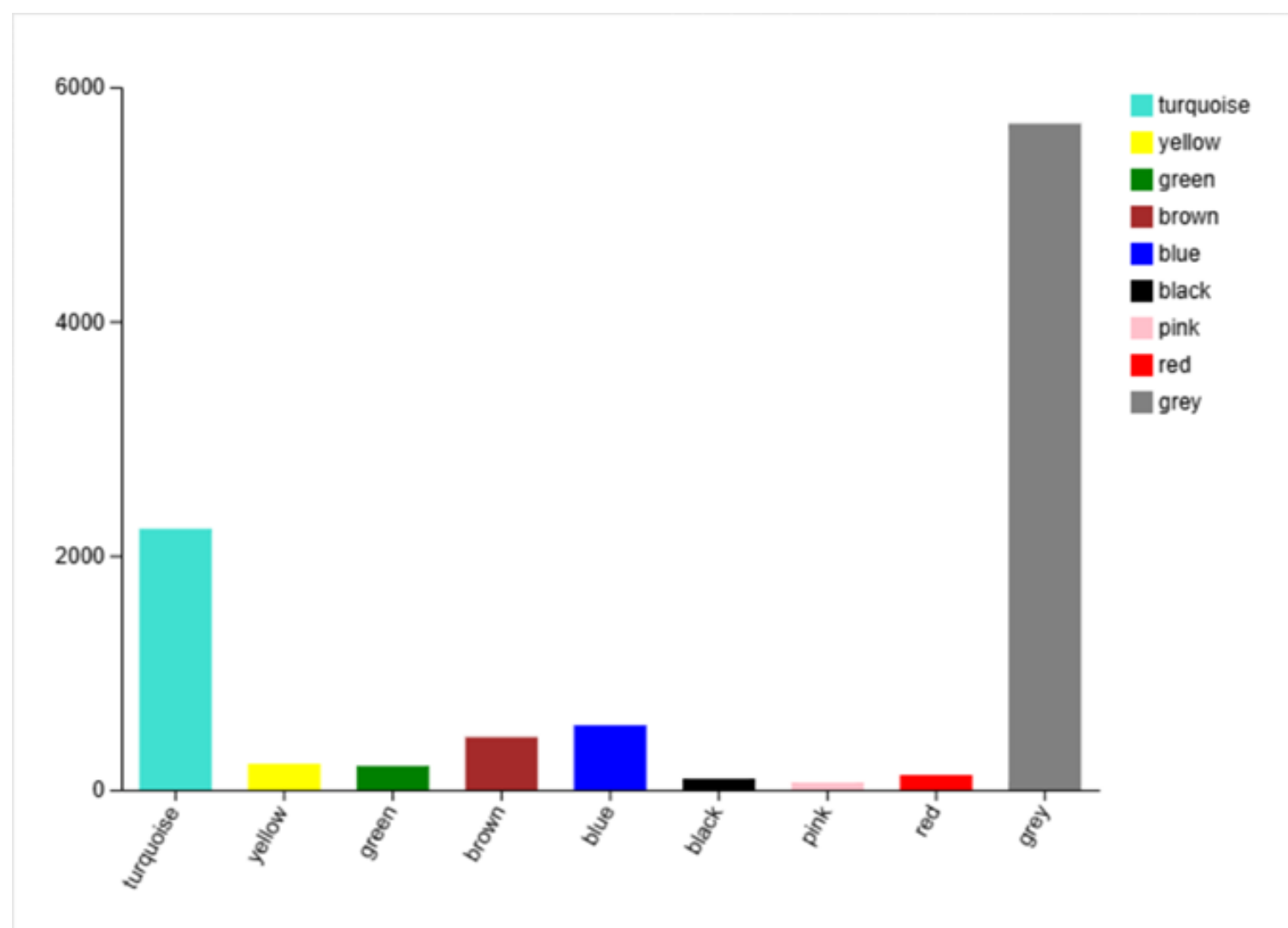
2.9.4.2 模块识别

模块成员可视化结果



注：提供 UMAP，tSNE 和 PCA 三种展示形式。展示不同颜色模块下的基因在不同细胞类群中的表达分布情况。

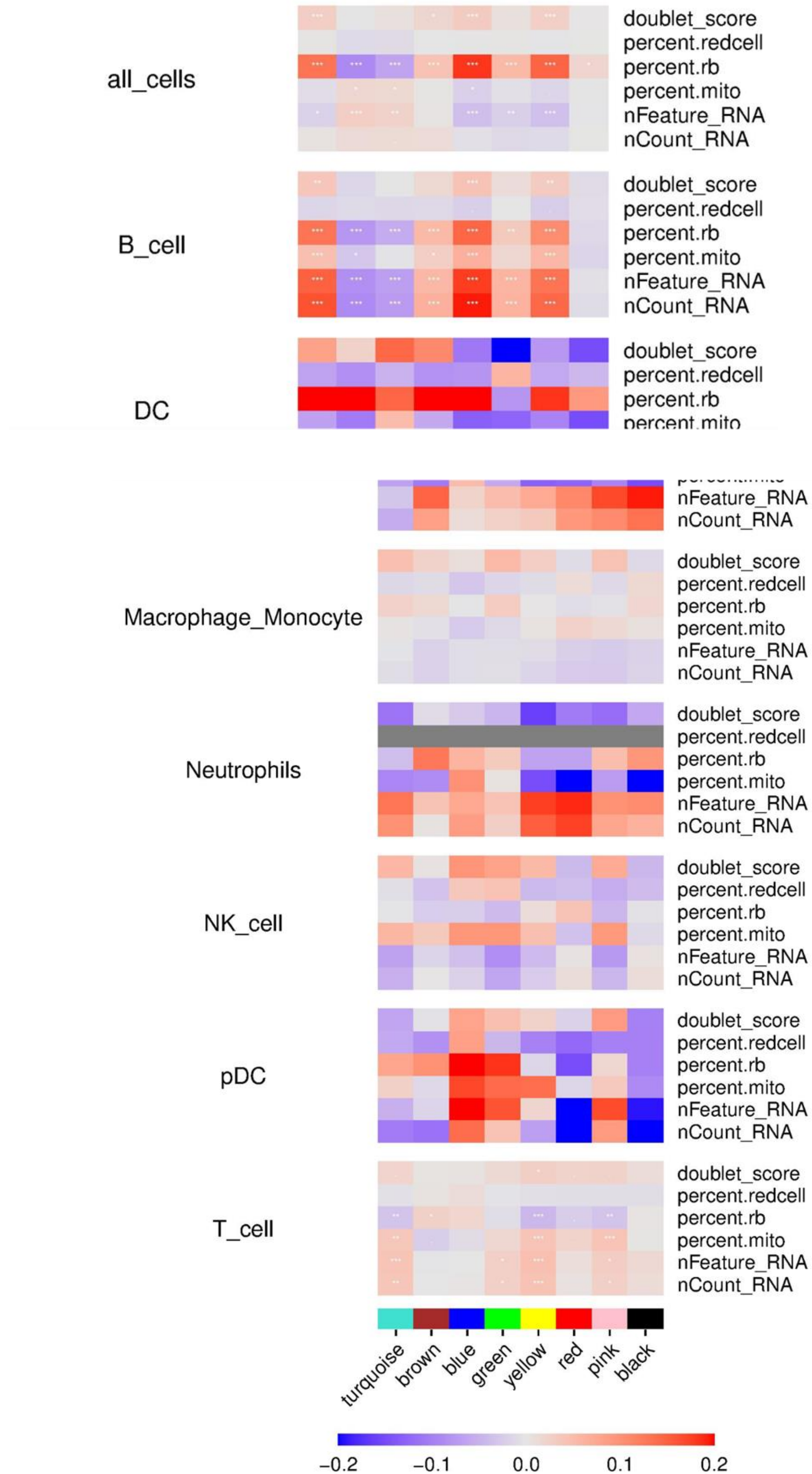
模块成员统计柱状图



注：直观展示各模块包含的基因数目统计图。横坐标代表模块（以模块颜色表示）；纵坐标代表属于该模块的成员数目。若颜色为灰色，则代表没有划分到具体模块的基因。

2.9.4.3 模块-表型相关性分析

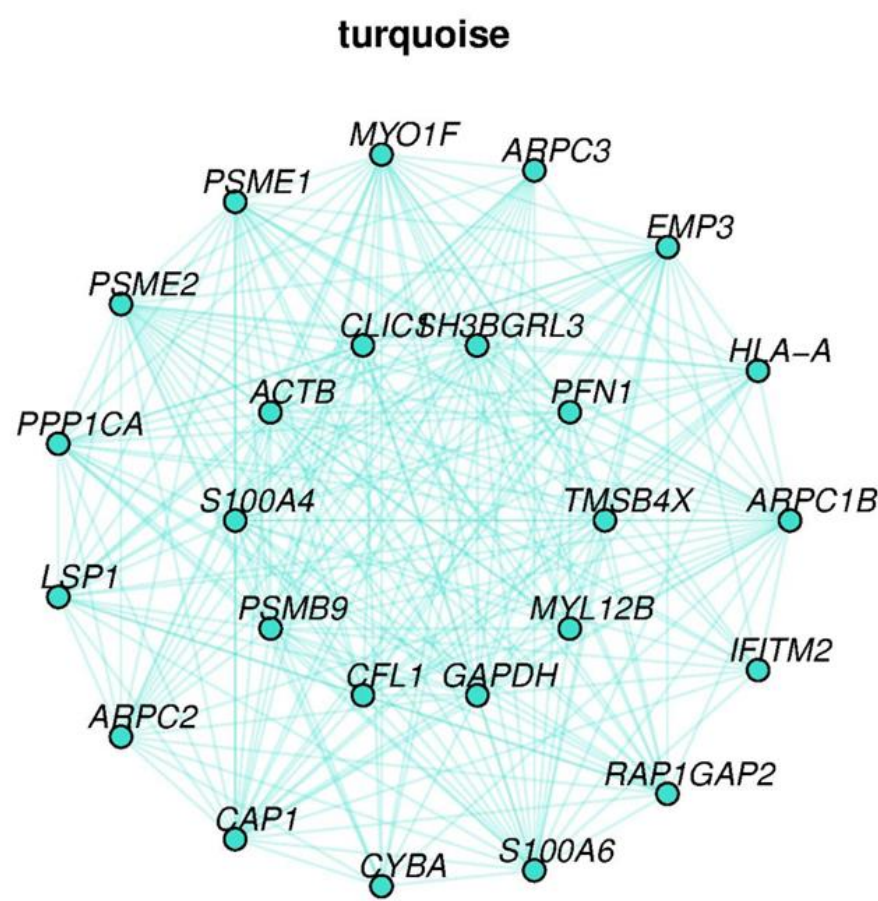
模块与表型相关性热图



注：横坐标表示模块，上方色块代表模块颜色，左侧纵坐标表示细胞类型（这里的细胞类型范围为一键化分析参数页面的细胞类型注释表下的所有细胞群，旨在查看对比同一模块下不同细胞群来源的基因与表型的相关性程度）。右侧表示不同的表型名称。当无表型数据时，默认将细胞6个质量值作为表型（具体含义请

注：根据 KME 值的大小，选取每个模块的 hub 基因和非 hub 基因（基因数量见上方参数选择），进行模块组合绘图。图中一种颜色代表一个模块。大的节点表示 hub gene, 小的节点表示非 hub 基因。点连线越多，表明该点越重要。默认每个模块调取前 6 个中心基因和 3 个其他基因进行绘图。

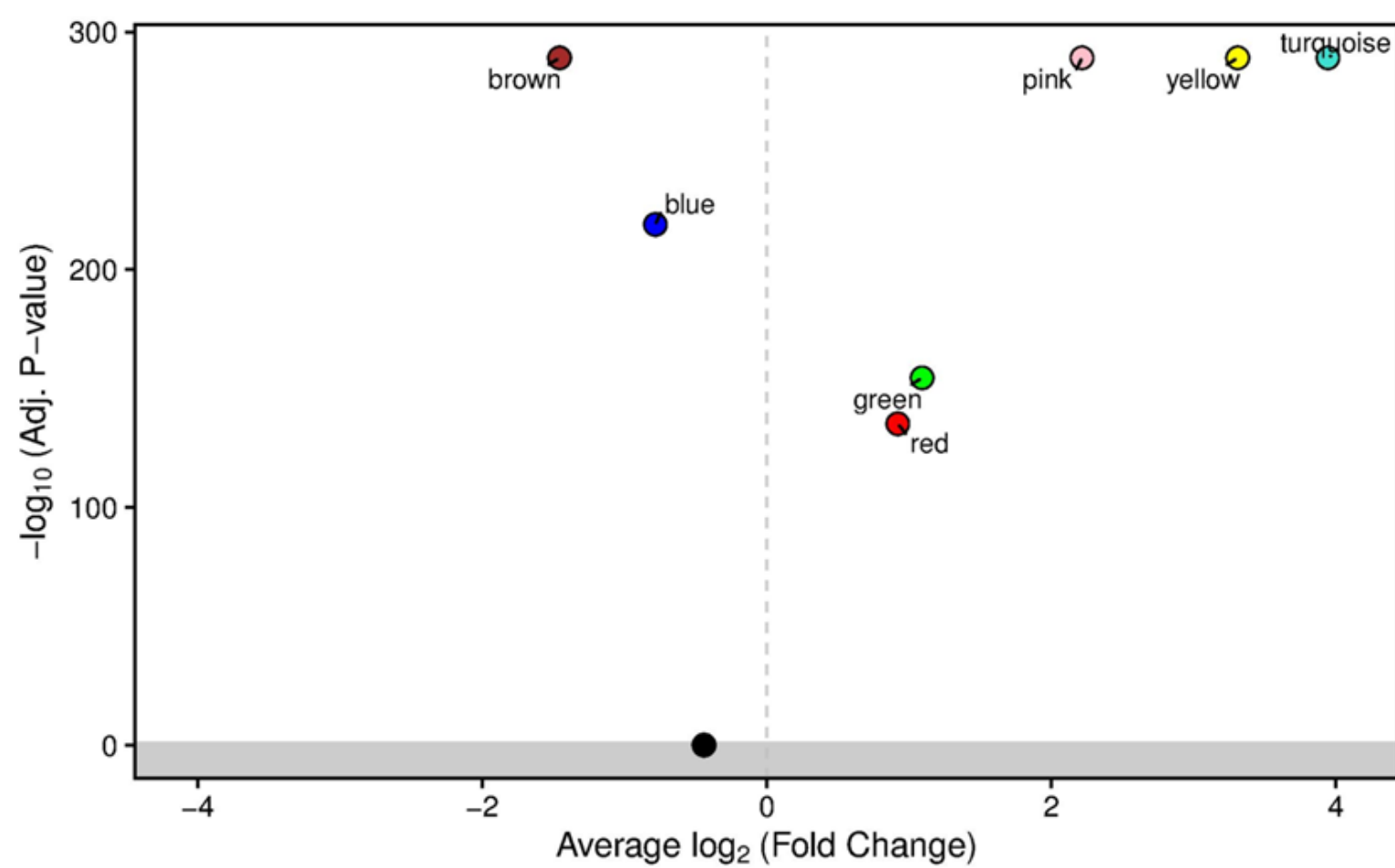
单个模块网络图-可视化网络图



注：每个节点 **nodes** 代表一个基因，每条边 **edges** 代表网络中两个基因之间的共表达关系，颜色对应模块的名称。根据 kME 值从小到大排序后选取模块中的 25 个基因进行绘图。网络图区分内外圈，内圈为 kME 前十的基因，外圈为剩余基因。

2.9.4.5 模块差异分析

差异火山图



注：横坐标为模块在不同组间/不同细胞间表达差异的倍数变化值；纵坐标为模块表达变化差异的统计学检验值，即 **P_{adjust}**。对数化处理后，P 值越高则表达差异越显著。图中每个点代表一个特定的模块，

将所有模块映射上去之后，可以获知，越靠两边和上边的模块表达差异越显著。

2.9.5 转录因子分析-SCENIC

SCENIC（single-cell regulatory network inference and clustering）是一个基于共表达和 motif 分析，计算单细胞转录组数据基因调控网络重建以及细胞状态鉴定的方法，可以利用其分析细胞或亚群的转录因子调控。

Motif logo 图

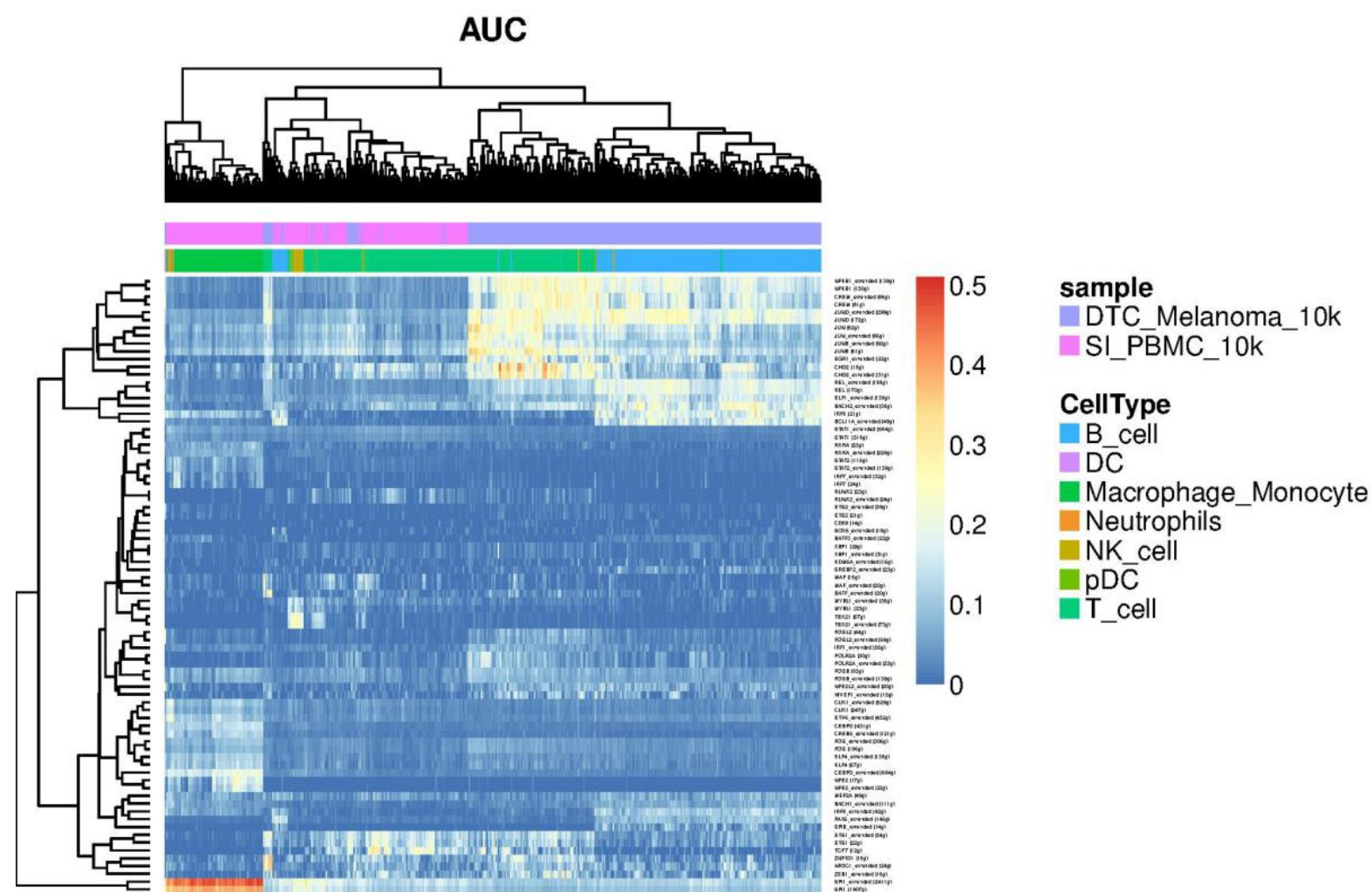
注：Motif logo 由每个位置的 ATGC 碱基组成。字母的相对大小表示它们在序列中出现的频率，与高度成正比，常以 bits 为单位。

Motif 富集结果表

logo	motifDb	geneSet	nEnrGenes	motif	NES	AUC	highlightedTFs	TFinDB	TF_highConf
	500bp	BACH1_top5p...	43	cisbp_M0283	3.44	0.063	BACH1	*	
	500bp	BACH1_top5p...	53	cisbp_M0287	3.74	0.067	BACH1	*	
	500bp	BACH1_top5p...	57	cisbp_M0346	4.47	0.076	BACH1	*	
	500bp	BACH1_top5p...	50	cisbp_M0361	3.63	0.066	BACH1	*	
	500bp	BACH1_top5p...	66	cisbp_M0362	3.63	0.066	BACH1	*	
	500bp	BACH1_top5p...	8	cisbp_M5290	3.07	0.059	BACH1	*	ARNTL (direct...
	500bp	BACH1_top5p...	42	cisbp_M5323	3.27	0.061	BACH1	*	CREB3 (directA...
	500bp	BACH1_top5p...	25	cisbp_M6039	3.41	0.063	BACH1	*	JDP2 (inferred...
	500bp	BACH1_top5p...	44	elemento_T...	3.25	0.061	BACH1	*	
	500bp	BACH1_top5p...	28	hocomoco_A...	3.16	0.06	BACH1	*	ATF7 (directA...

注：（1）logo：点击可以联动切换展示下方的 motif logo 图；（2）motifDB：数据库名称；（3）geneSet：基因集名称；（4）nEnrGenes：高度排名的基因数量；（5）motif：motif 的 ID；（6）NES：基因集中基序的标准化富集分数；（7）AUC：曲线下面积（用于计算 NES）；（8）highlightedTFs：高亮显示的转录因子；（9）TFinDB：指示突出显示的 TF 是包含在高置信度注释（两个星号）还是低置信度注释（一个星号）中；（10）TF_highConf：根据“ motifAnnot_highConfCat”注释的转录因子；（11）TF_lowConf：根据“ motifAnnot_lowConfCat”注释的转录因子；（12）rankAtMax：在最大富集时的排名，用于确定富集的基因数。

Regulon 活性热图



Subset of 500 random cells

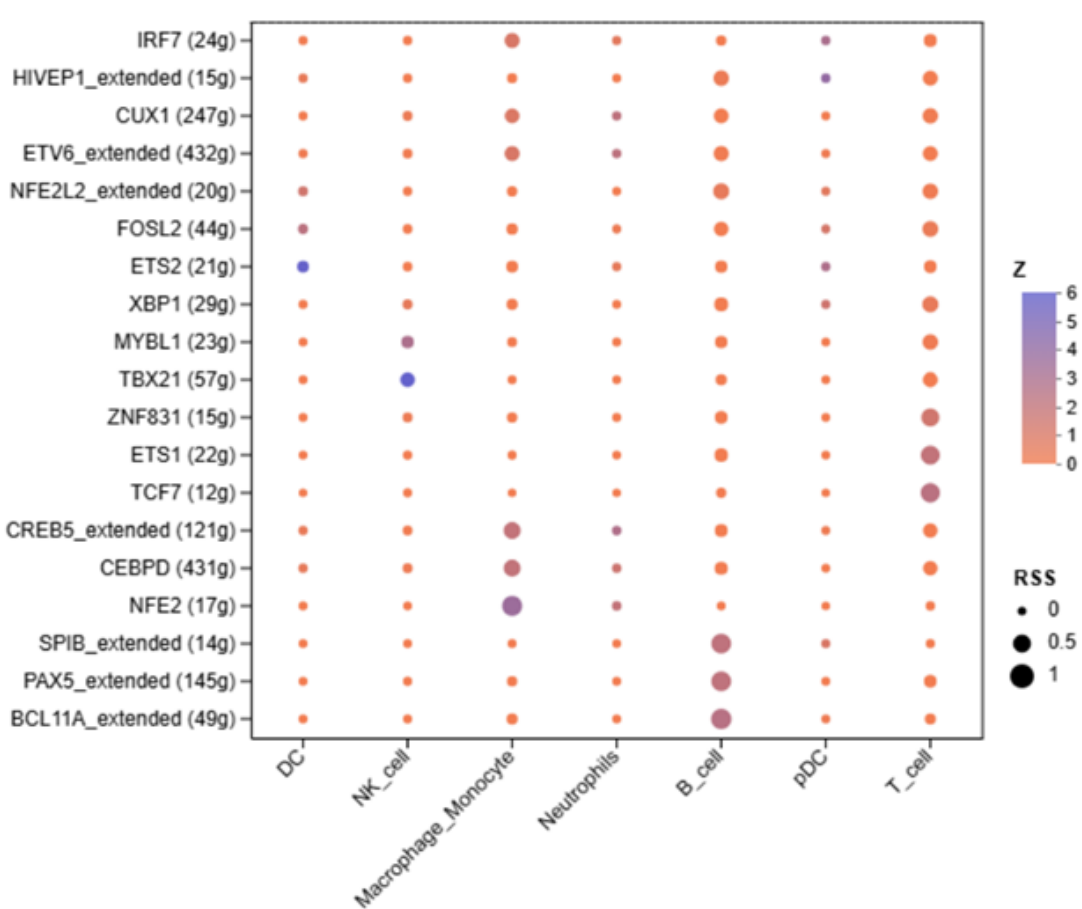
注：AUC 代表了与细胞内其他基因相比，特征基因中表达基因的比例及其相对表达值。AUCcell 使用 AUC 来计算输入基因集的关键子集是否在每个细胞的排名 Top 都得到了富集。软件会将每个细胞转为 AUC 值，对每个细胞的 AUC score 值进行可视化，随机提取 500 个细胞可视化。其中行为 Regulon，调控子，具体是指受同一个调控元件（转录因子）调控的一群基因的集合；列为细胞类型，有时候也指代 cluster /subcluster；颜色为 AUC score 值，颜色由蓝到红表示调控子的活性由低到高。Bar 分别代表样本以及 celltype。

Regulon 关联特异性指数（CSI）聚类图



注：通过 RAS 来计算不同 regulon 之间的相关性系数（PCC, Pearson Correlation Coefficient），基于 PCC，计算了 CSI（Connection Specificity Index）来衡量 regulon pairs 之间的相关性，之后基于 CSI 对 regulon 进行层次聚类。其中，行列均为 Regulon，颜色由浅到深表示 CSI 值由低到高。CSI 值较高的 Regulon 之间可能具有相似的细胞功能，共同调控下游基因。此图仅包含高可信的 Regulon，即不含 exten ded。

Regulon 特异性评分气泡图



注：行为 Regulon，列为 cluster/Celltype，颜色为 Z-score 值，圆圈大小为 RSS，

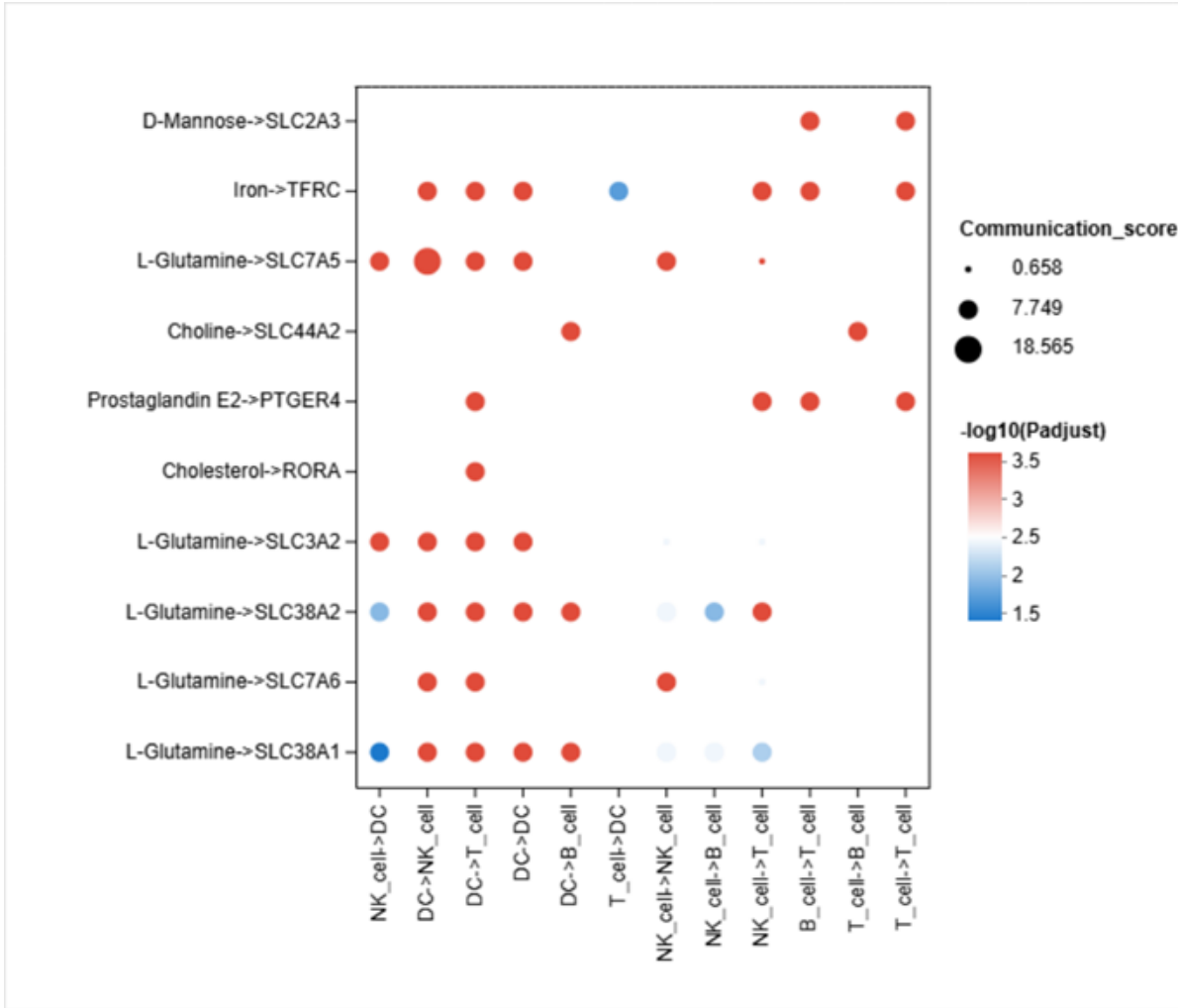
越大说明 Regulon 特异性越高。

2.9.6 MEBOCOST 代谢通讯分析

MEBOCOST 首先通过单细胞转录组数据和已知的代谢物数据库来预测代谢物在细胞中的存在，并结合不同代谢物的感受器在细胞中的表达量来预测不同细胞群之间的细胞通讯。数据来源上，作者选取了 scFEA(157),COMPASS(2597)和 HMDB(910)上被注释代谢物、相关酶和感受器的通路。

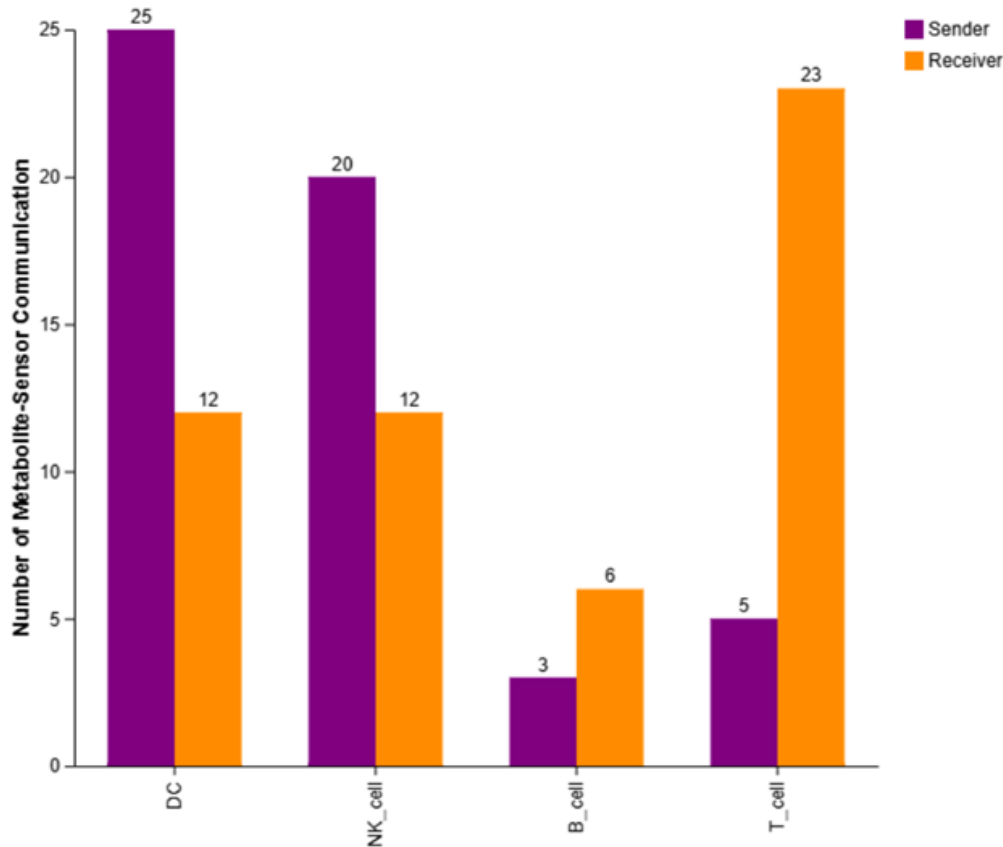
2.9.6.1 Mebocost 代谢整体分析

两两细胞间代谢通讯详情气泡图



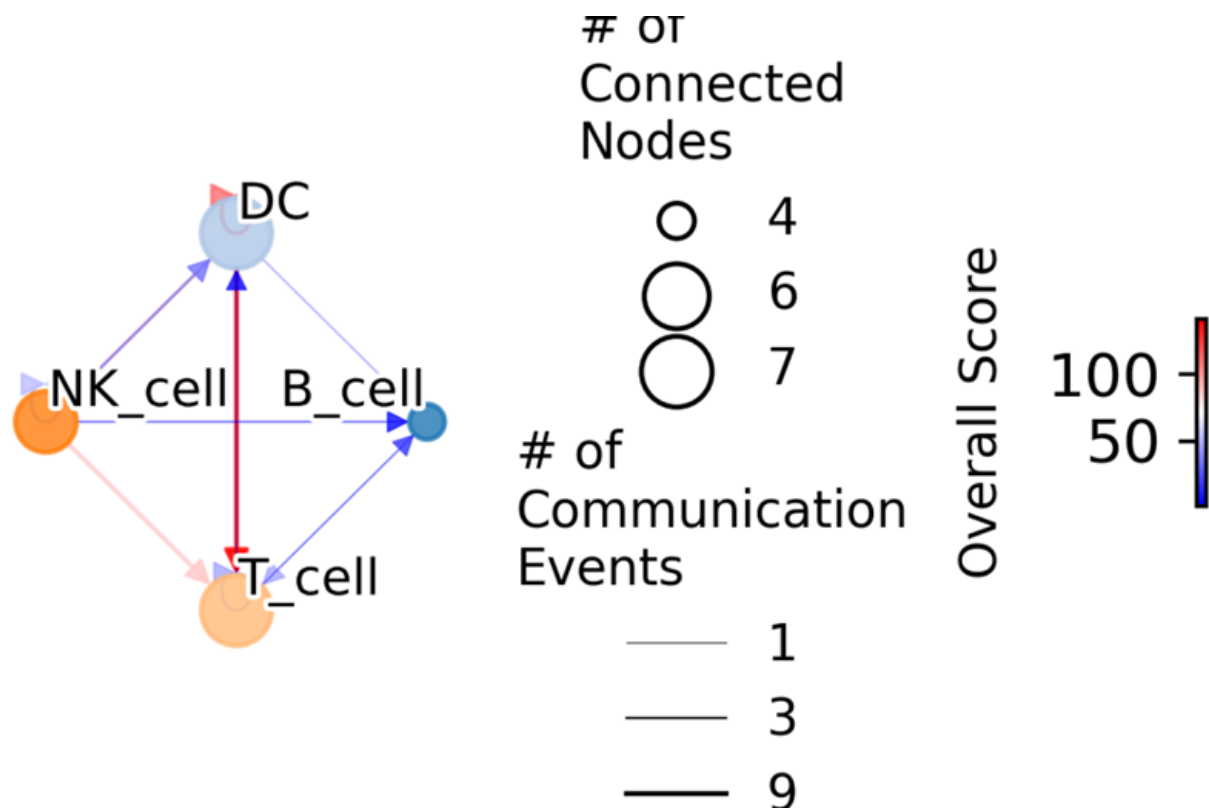
注：展示发送与接收两两细胞类型之间的通讯详情，纵坐标表示代谢物-传感器对（类似传统细胞通讯的配体受体对），横坐标为细胞类型对（发送者细胞类型-接收者细胞类型）。圆点大小表示 MEBOCOST 计算的通信分数，圆点的颜色是代谢物-传感器通信的 $-\log_{10}(\text{FDR})$ 值。【图表工具可以调整绘图参数】

代谢通讯柱形图



注：横坐标表示不同细胞类型，纵坐标表示各 Celltype 中代谢物发送者（紫色柱子）与传感器接收者（橙黄色柱子）的通讯数量。

细胞间代谢通讯可视化图-circle 图



注：每个点代表一种细胞类型。圆点的大小代表与其他细胞类型的通信数量，线条方向表示从发送细胞类型到接收细胞类型的通信，线条粗细表示发送和接收细胞类型之间代谢物传感器通信的数量，线条颜色显示了总体置信度，该置信度是由发送者和接收者之间所有代谢物传感器通信的 $-\log_{10}(\text{FDR})$ 之和计算出来的。

代谢通讯数目统计表

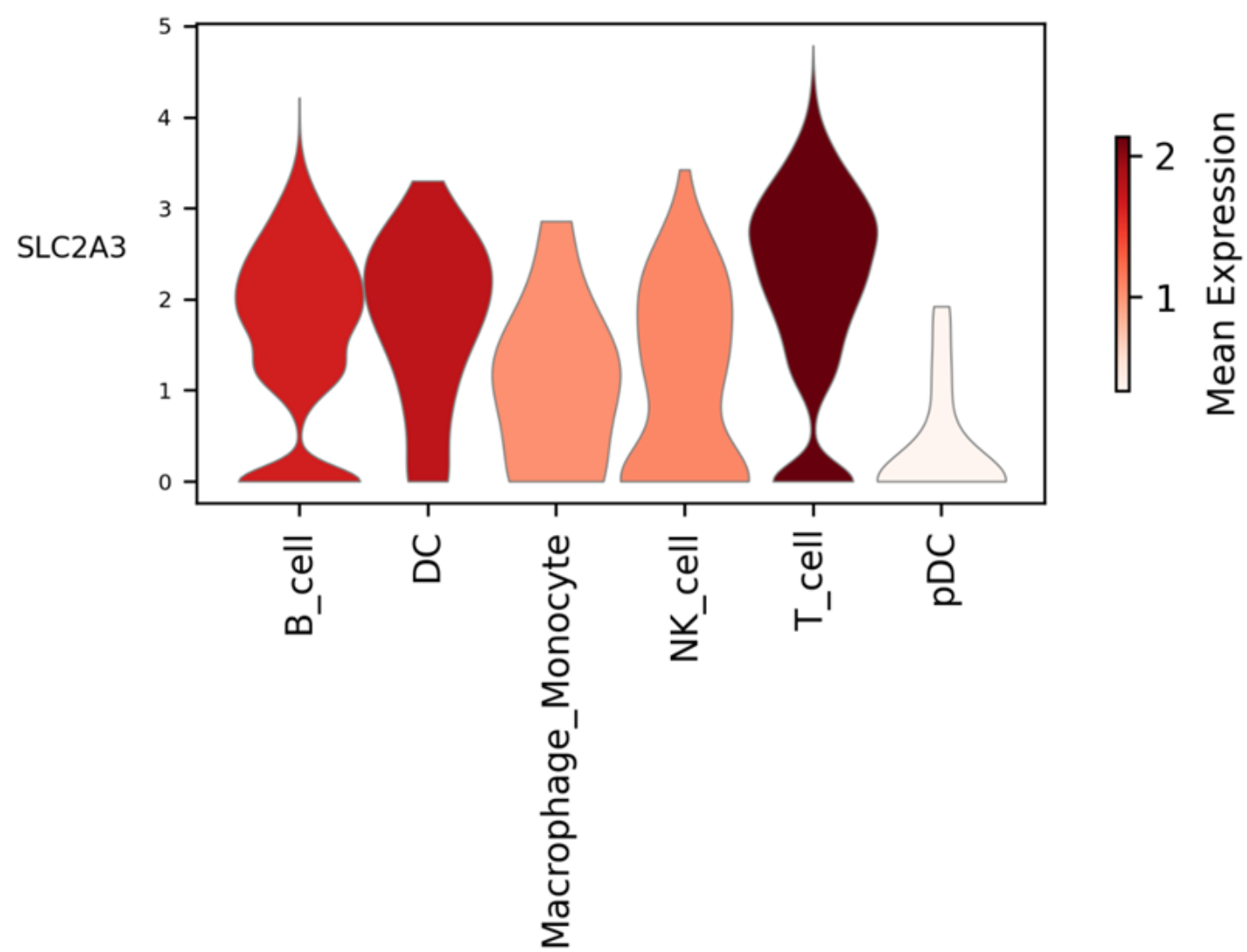
Celltype	Sender_Number	Receiver_Number
B_cell	3	3
DC	7	7
NK_cell	4	7
T_cell	4	11
Macrophage_Monocyte	0	0
Neutrophils	0	0
pDC	0	0

注：（1）Celltype：细胞类型；（2）Sender_Number：该 Celltype 作为发送者（Sender）发出的信号（代谢物）个数【点击蓝色数字可进入“代谢通讯详情表”页面】；（3）Receiver_Number：该 Celltype 作为接收器（Receiver）接收的信号（传感器基因）个数【点击蓝色数字可进入“基因”页面，勾选可创建基因集】。

温馨提示：统计表是软件之间按照 $\text{P}_{\text{adjust}} < 0.05, \text{Commu_Score} > 0$ 过滤后展示统计结果，如需看全部结果，可在详情表处查看。

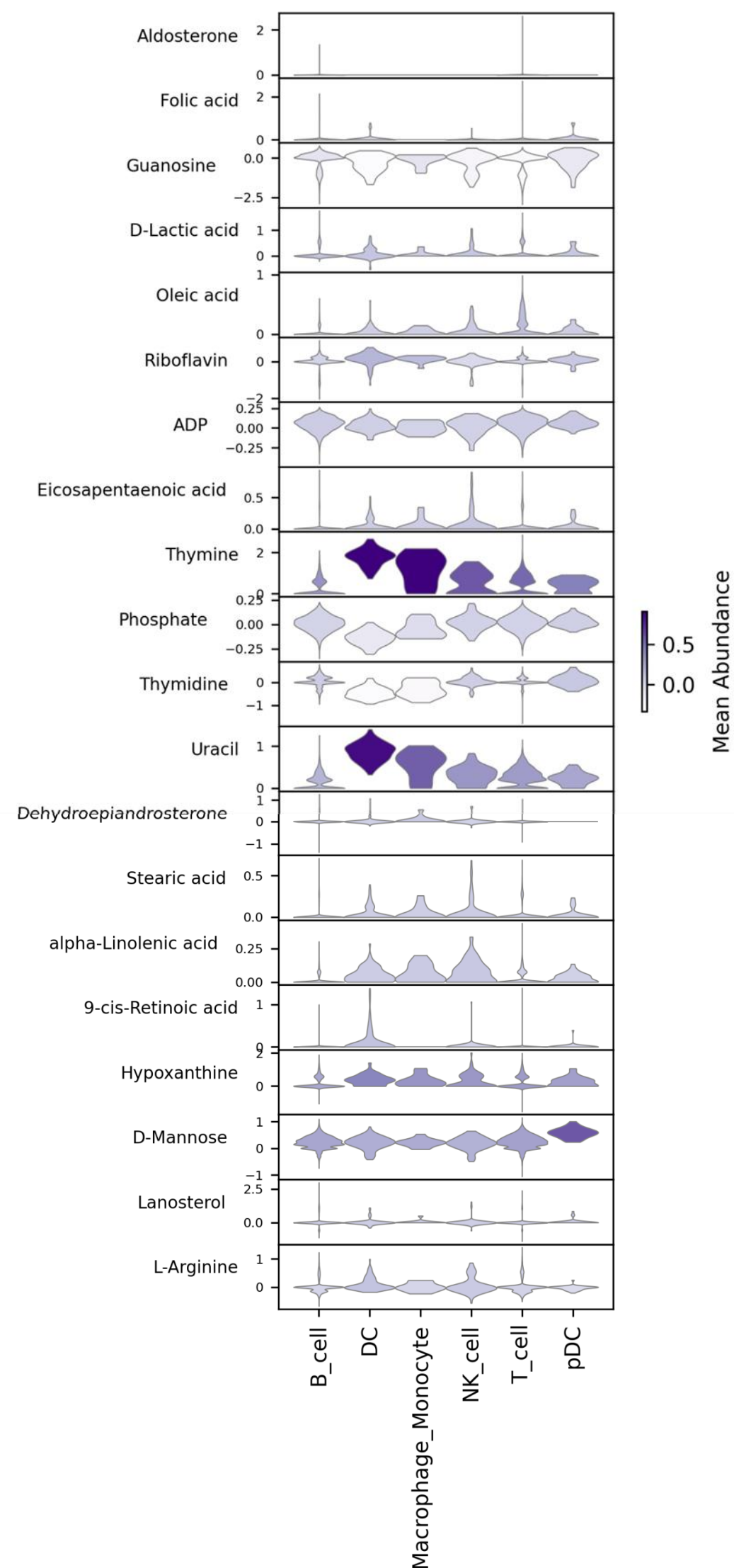
2.9.6.2 指定代谢物可视化

细胞类型中感受器基因表达量小提琴图



注：展示不同细胞类型中有较高表达的感受器基因表达量小提琴图。横坐标表示细胞类型，纵坐标为感受器基因，小提琴图每个点代表一个细胞。

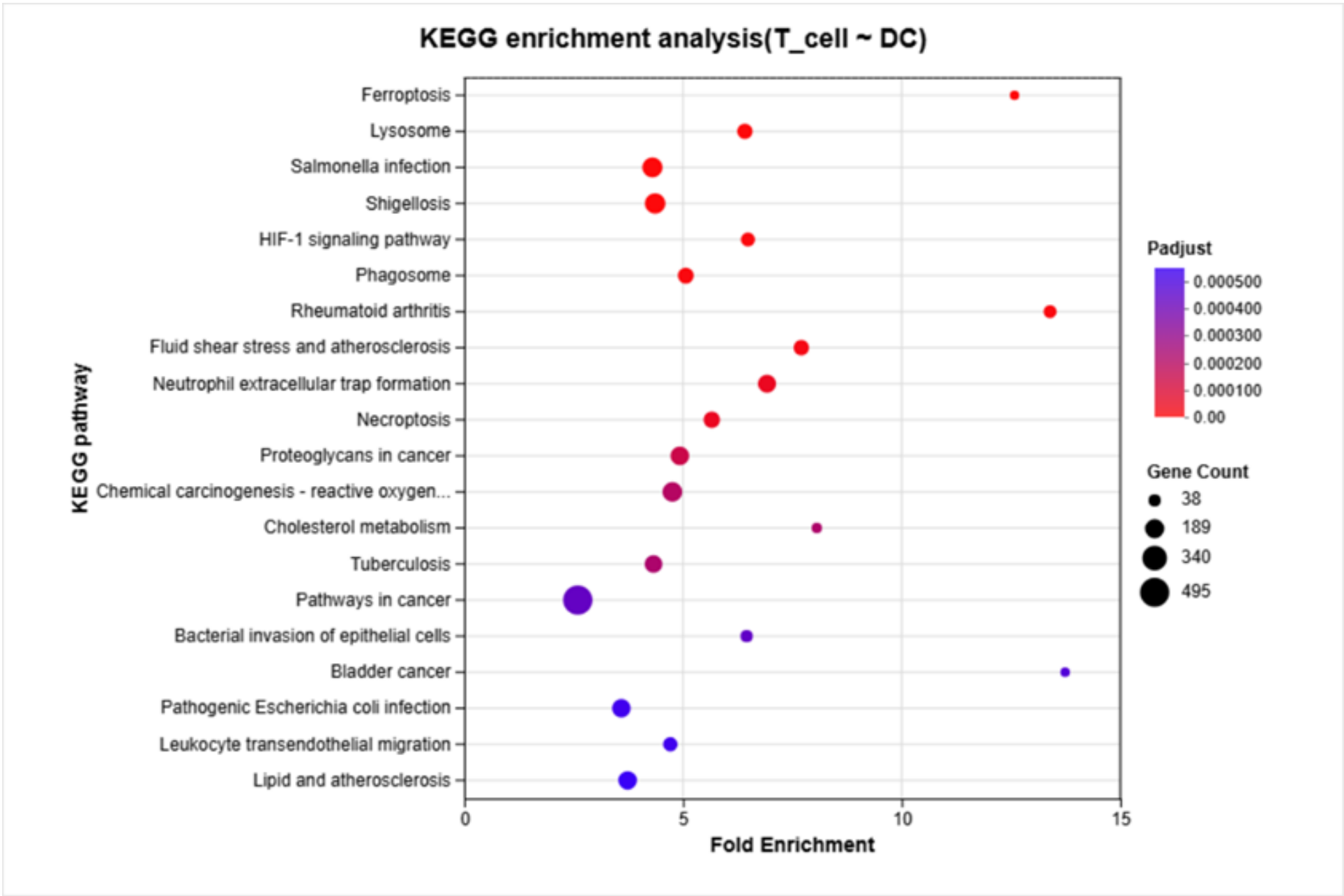
细胞类型中代谢物预测丰度小提琴图



注：展示不同细胞类型中有较高富集的代谢物预测丰度小提琴图。横坐标表示细胞类型，纵坐标为代谢物，小提琴图每个点代表一个细胞。

2.9.6.3 代谢通讯富集分析

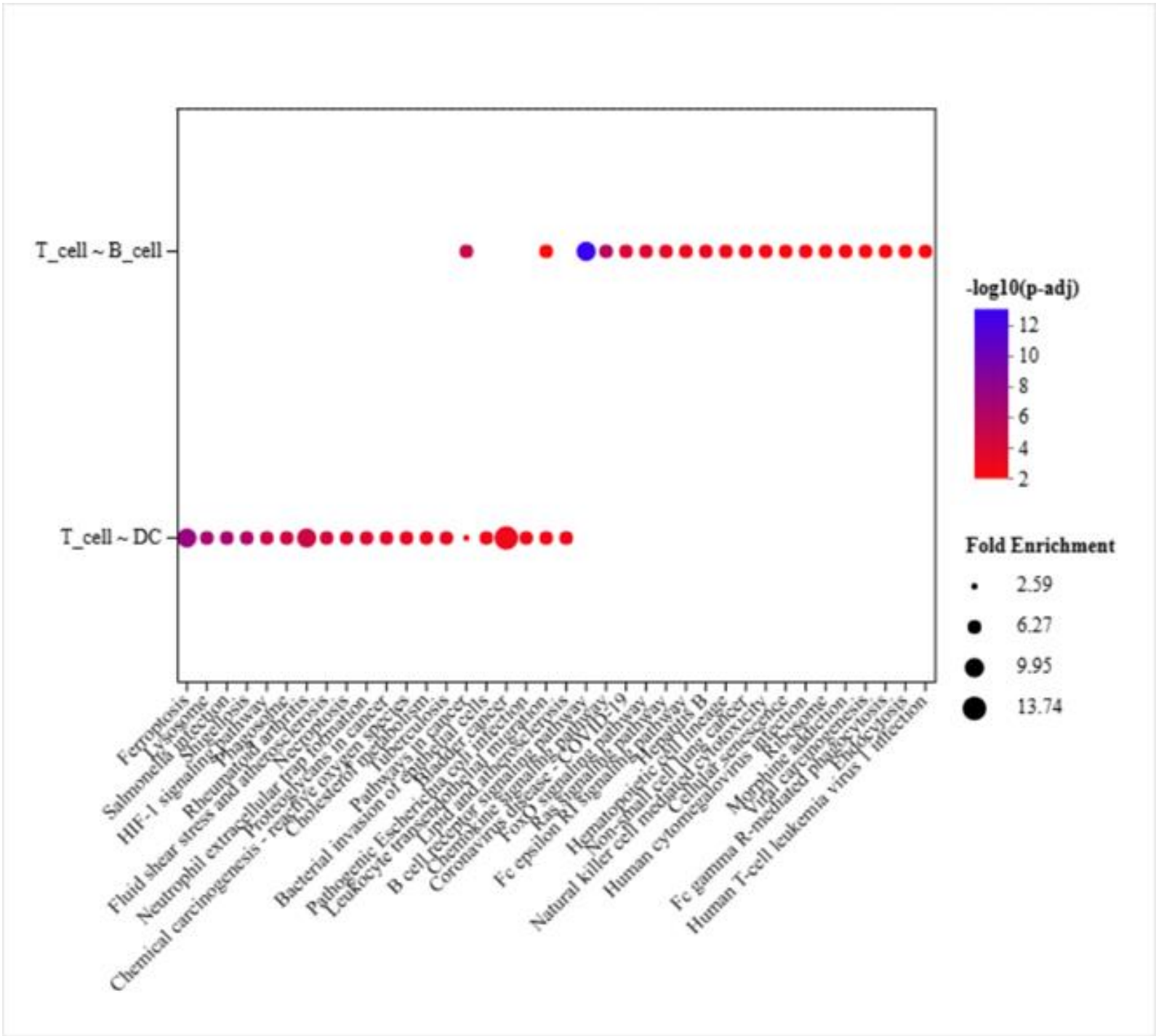
单个 pair 对富集分析结果图（气泡图）



注：纵轴表示 KEGG pathway 名称，横坐标表示 FoldEnrichment 值。

FoldEnrichment: 富集倍数，具体参见表格中注释，该值越大，表示富集的程度越大。点的大小表示此 KEGG pathway 中基因个数多少，点的颜色对应于不同的 Padjust 范围。【注：默认按 Padjust 从小到大排序显示 top20 的富集结果，可图表工具修改参数调整绘图形式。】

多个 pair 对间差异富集分析结果（气泡图）



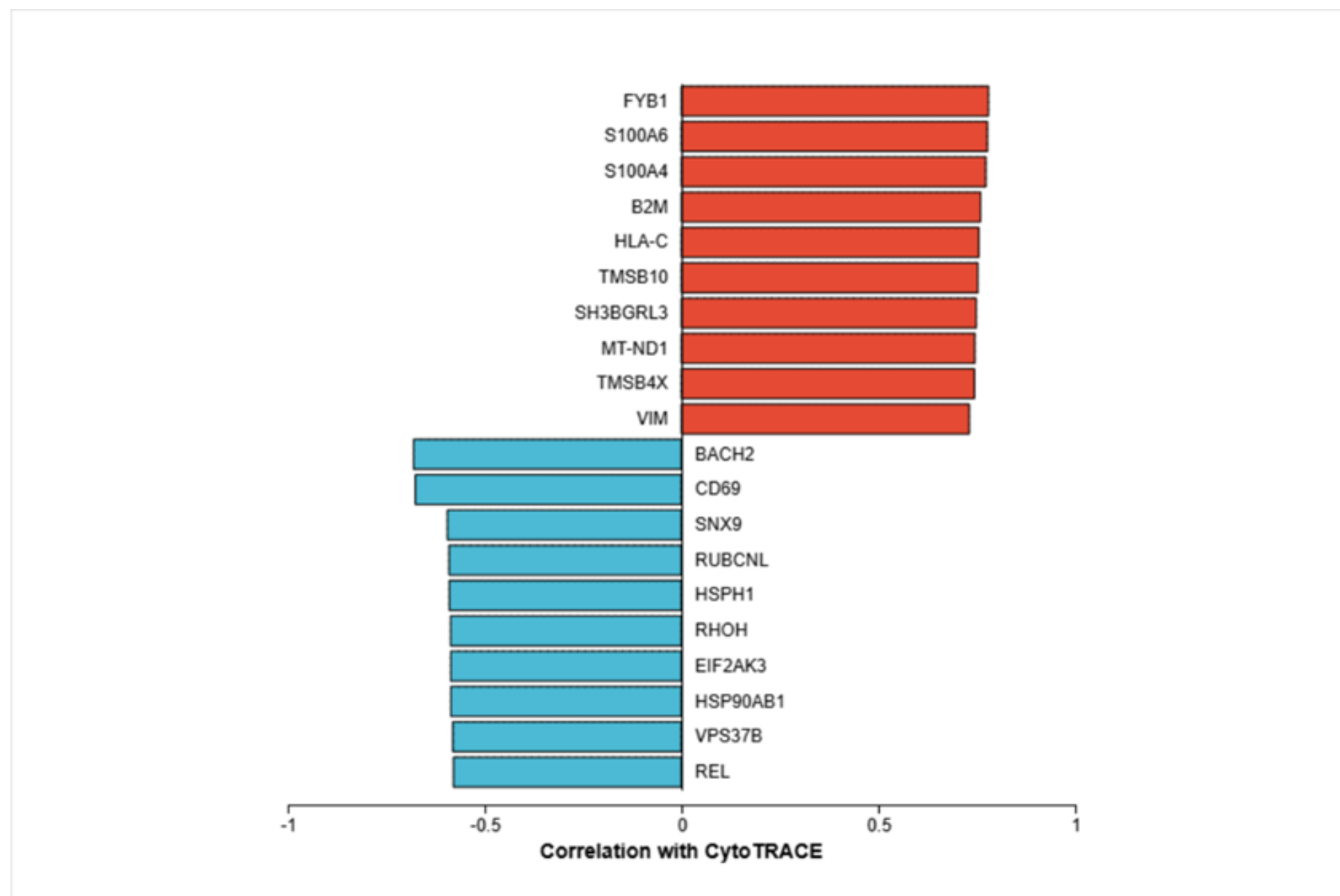
注：横坐标表示 KEGG pathway 名称，纵坐标表示 Pair 对。点的大小表示此

KEGG pathway 中基因个数多少，而点的颜色对应于不同的 Padjust 范围。
FoldEnrichment: 富集倍数，具体参见表格中注释，该值越大，表示富集的程度越大。【注：默认按 Padjust 从小到大排序显示 top20 的富集结果，可图表工具修改参数调整绘图形式。】

2.9.7 细胞分化潜能-CytoTRACE

CytoTRACE 是由斯坦福大学开发的，用于计算细胞分化能力的计算框架。可在不需先验知识的情况下，利用基因计数特征（gene counts signature, GCS）在单细胞数据上评估细胞分化水平。与大多数现有的谱系轨迹分析方法不同，CytoTRACE 能够独立于中间态（如 RNA velocity）或数据中连续发育过程（如 monocle）的存在与否，预测细胞分化的相对状态和方向。

CytoTRACE 打分值正负相关基因柱形图



注：将每个基因（每个基因在不同细胞中的表达量向量）与 CytoTRACE 结果值进行相关性计算，得到跟 CytoTRACE 结果（细胞分化状态）相关性较高的基因。双侧柱形图展示与分化程度较高（右侧）和分化

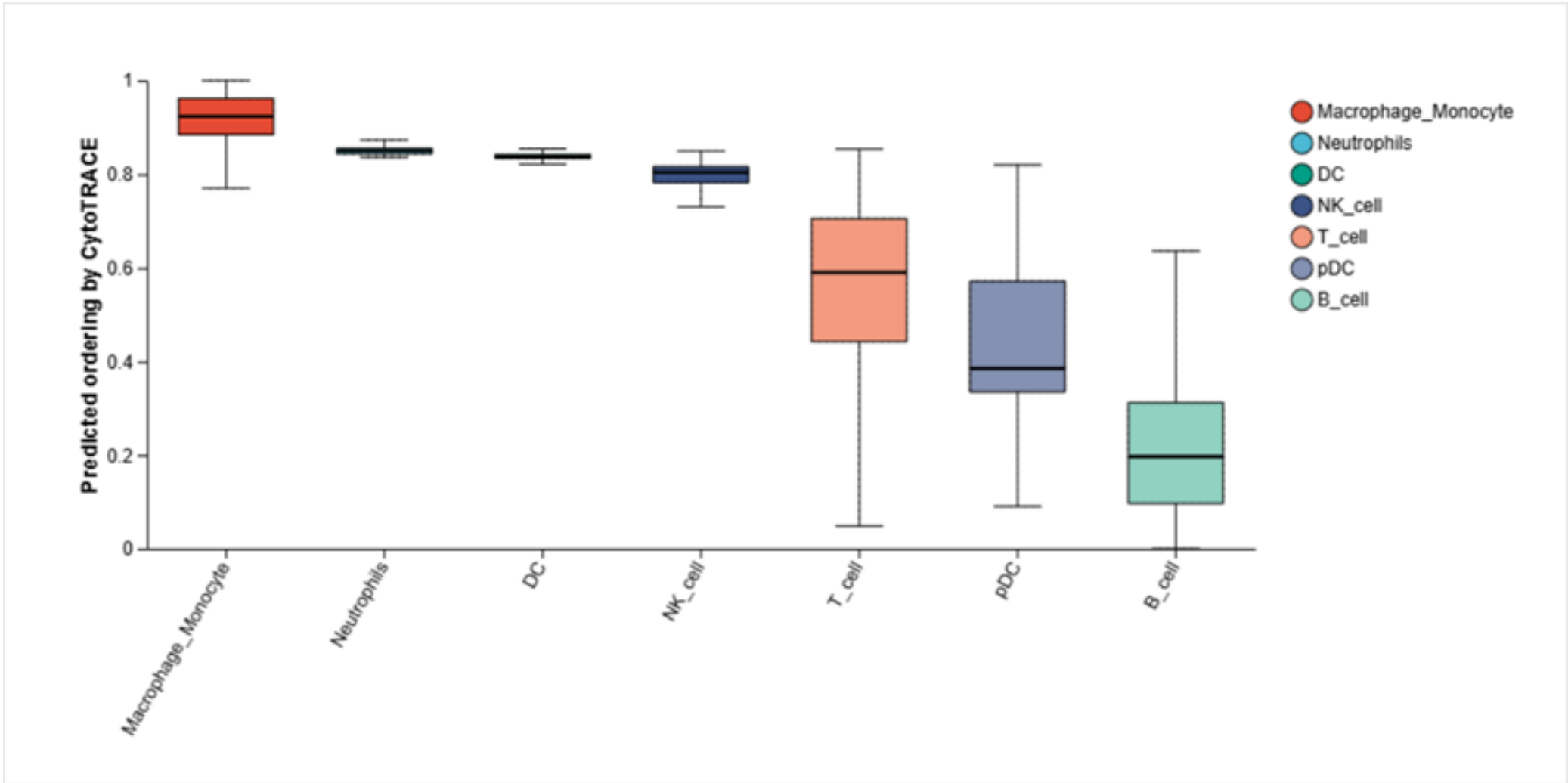
程度较低（左侧）的相关基因。【默认按照正负相关基因分别取值 TOP10 基因展示，可图表工具调整参数】

CytoTRACE 打分值正负相关基因结果表

Gene_id	Gene name	Gene descript...	Correlation	GO id	GO description	KO id	KO name	KEGG pathwa...	KEGG pathwa...
ENSG0000019...	S100A4	S100 calcium ...	0.7729982800...	GO:0070062;;...	extracellular ...	K23761	S100A4	-----	-----
ENSG0000007...	ACTB	actin beta [So...	0.6857836837...	GO:0005829;;...	cytosol;;kines...	K05692	ACTB_G1	map05164;; m...	Influenza A;;...
ENSG0000003...	TMSB10	thymosin beta...	0.7527096393...	GO:0005737;;...	cytoplasm;;a...	K13785	TMSB10	-----	-----
ENSG0000020...	HLA-C	major histoco...	0.7551532506...	GO:0002474;;...	antigen proce...	K06751	MHC1	map05166;; m...	Human T-cell...
ENSG0000020...	TMSB4X	thymosin beta...	0.7442394588...	GO:0033209;;...	tumor necros...	K05764	TMSB4	map04810	Regulation of...
ENSG0000008...	FTL	ferritin light ch...	0.6403911112...	GO:0005737;;...	cytoplasm;;ce...	K13625	FTL	map04978;; m...	Mineral absor...
ENSG0000011...	VAMP8	vesicle associ...	0.6284359017...	GO:0046718;;...	viral entry into...	K08512	VAMP8	map04611;; m...	Platelet activa...
ENSG0000008...	FYB1	FYN binding p...	0.7798657290...	GO:0050852;;...	T cell recepto...	K17698	FYB, ADAP	map04015;; m...	Rap1 signalin...
ENSG0000019...	MT-ND1	mitochondria...	0.7451372782...	GO:0032981;;...	mitochondria...	K03878	ND1	map04714;; m...	Thermogenes...
ENSG0000014...	SH3BGR L3	SH3 domain b...	0.7487043627...	GO:0005737;;...	cytoplasm;;n...	K23740	SH3BGR	-----	-----

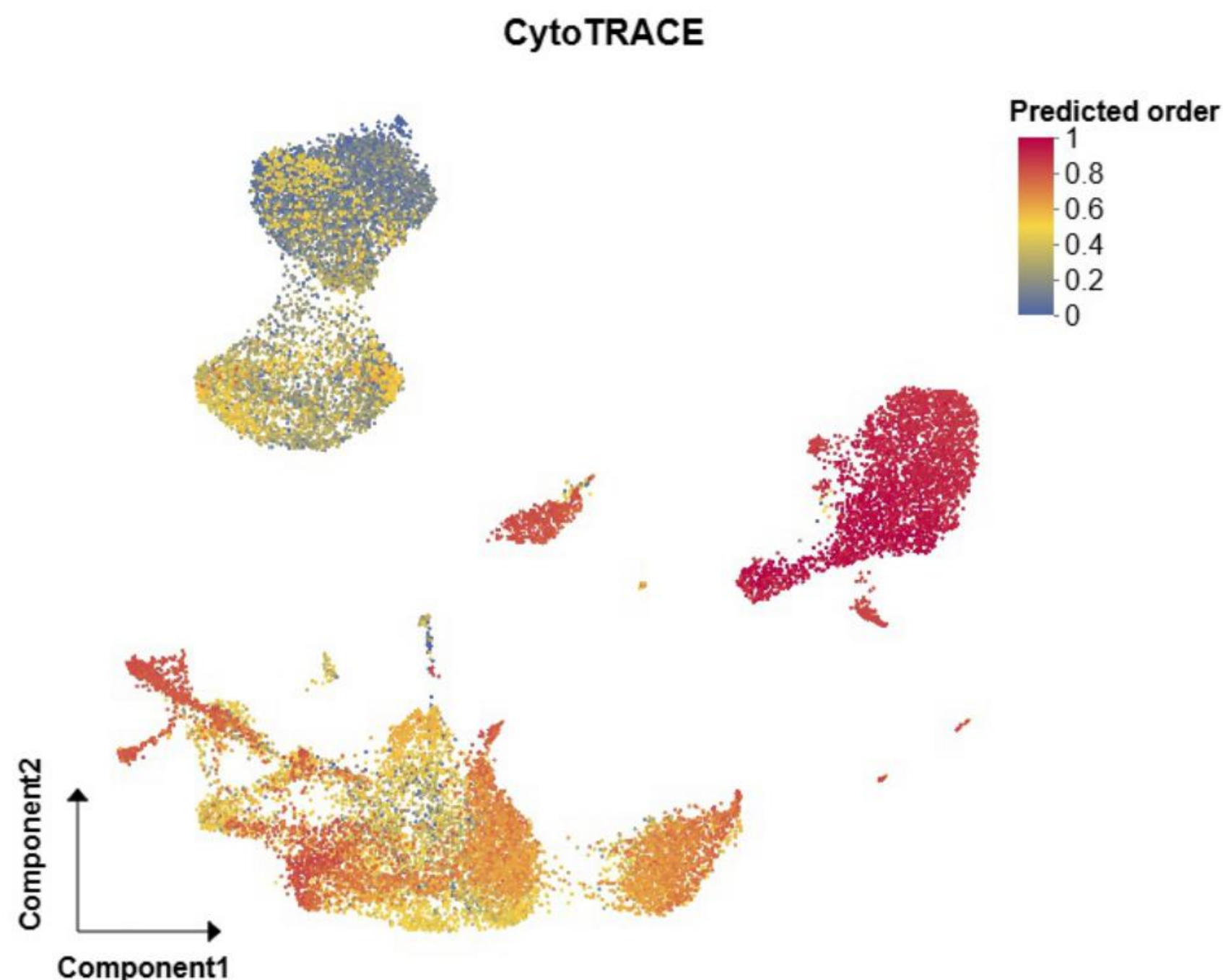
注：（1）Gene id：基因 id；（2）Gene name：基因名称；（3）Gene description：基因名称描述信息；（4）Correlation：将每个基因（每个基因在不同细胞中的表达量向量）与 CytoTRACE 结果值进行相关性计算得到相关性值；（5）其他列为基因在各大数据库的注释情况。【默认只展示 GO 和 KEGG 2 个数据库的注释信息，可通过筛选表格对其他数据库进行勾选展示】表格中蓝色字体均可以点击链接到基因详情页！

CytoTRACE 细胞类型分化潜力打分箱线图



注：根据每个单细胞的表型标签使用箱线图展示每个表型的 CytoTRACE 值的中位数和分布。横坐标为细胞类型，纵坐标为 CytoTRACE 分析值，打分值越高，表示细胞类型分化程度越低，干性越高。

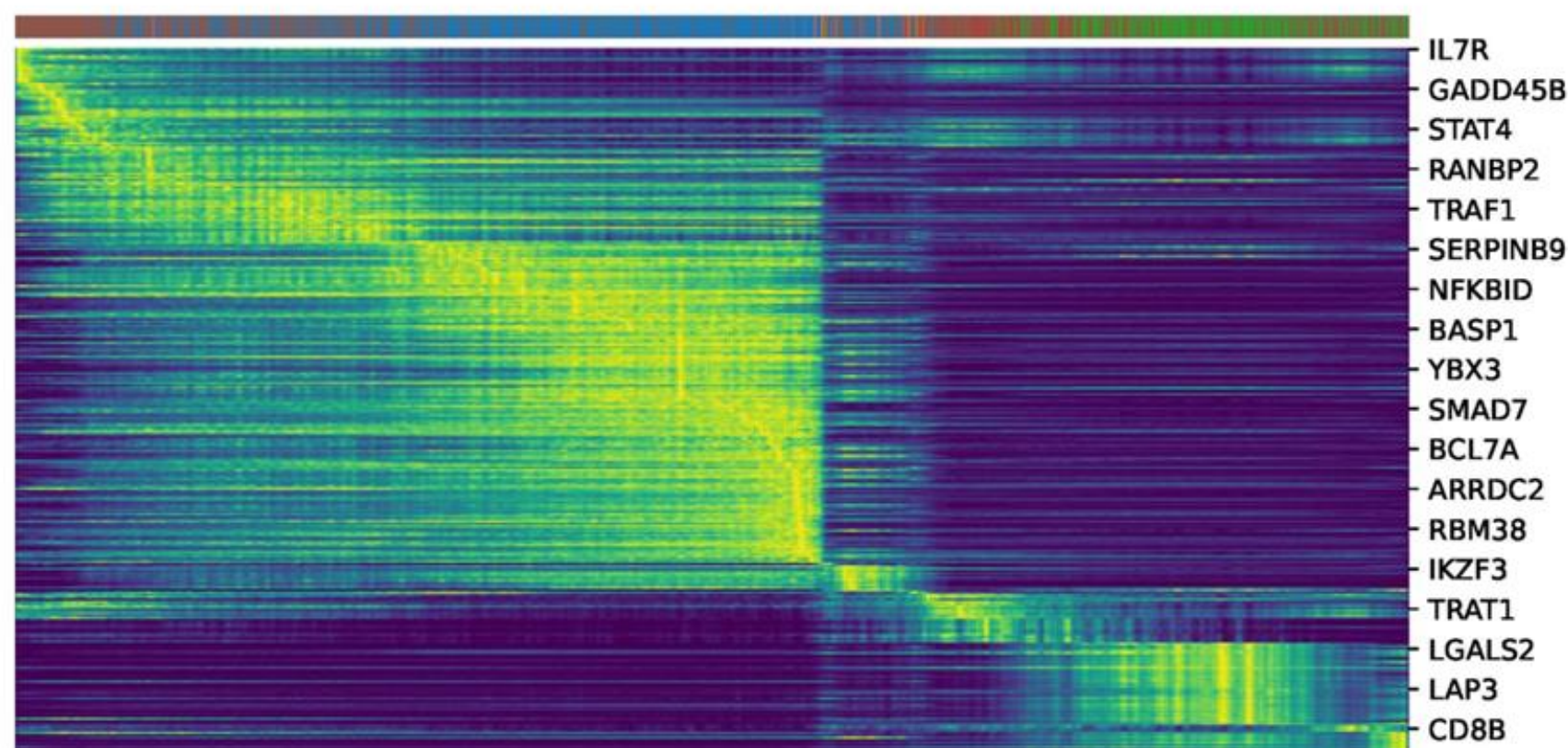
CytoTRACE 打分结果图



注：CytoTRACE 分析得到的结果值在 0 和 1 之间进行排序和缩放，代表细胞的相对分化状态的预测顺序（0：分化程度更高；1：分化程度较低），分化程度较低的细胞类型为起点细胞。

2.9.8 RNA 速率分析

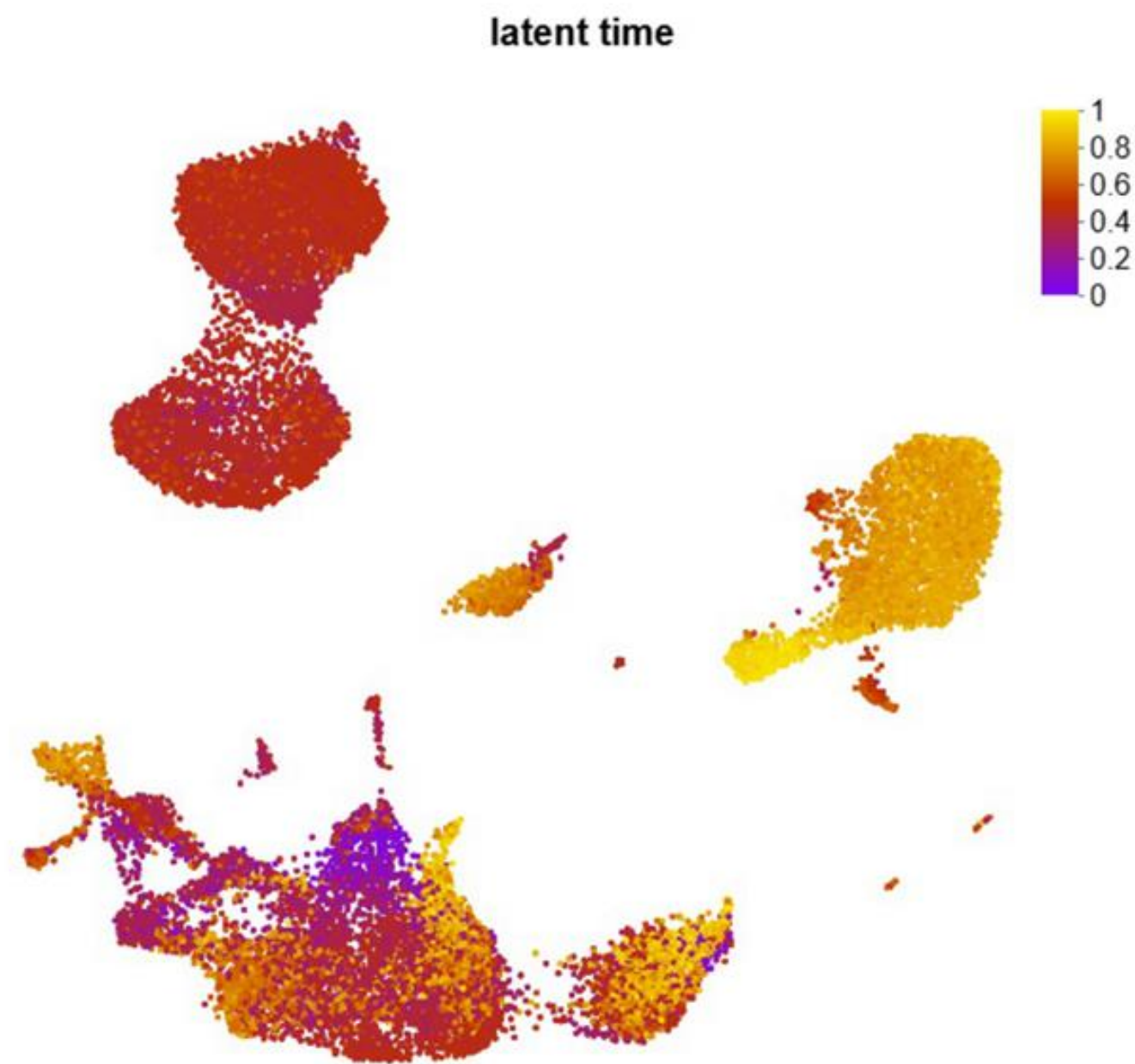
似然性排名 TOP300 基因热图



注：似然性排名 TOP300 基因热图是 scVELO 整体结果表（estimated_rate.csv）文件中按 fit_likelihoood 列排序后取前 300 个基因得到的表达谱热图（top_gene_heatmap）。其中行表示基因，列表示细胞，值为按基因归一化后的表达量，上方不同颜色色块代表不同细胞类型。备注：此图因为绘图基因较多，右侧展示的 name 是随机挑选展示的！温馨提示：若查看 300 个基因详情，请于下方

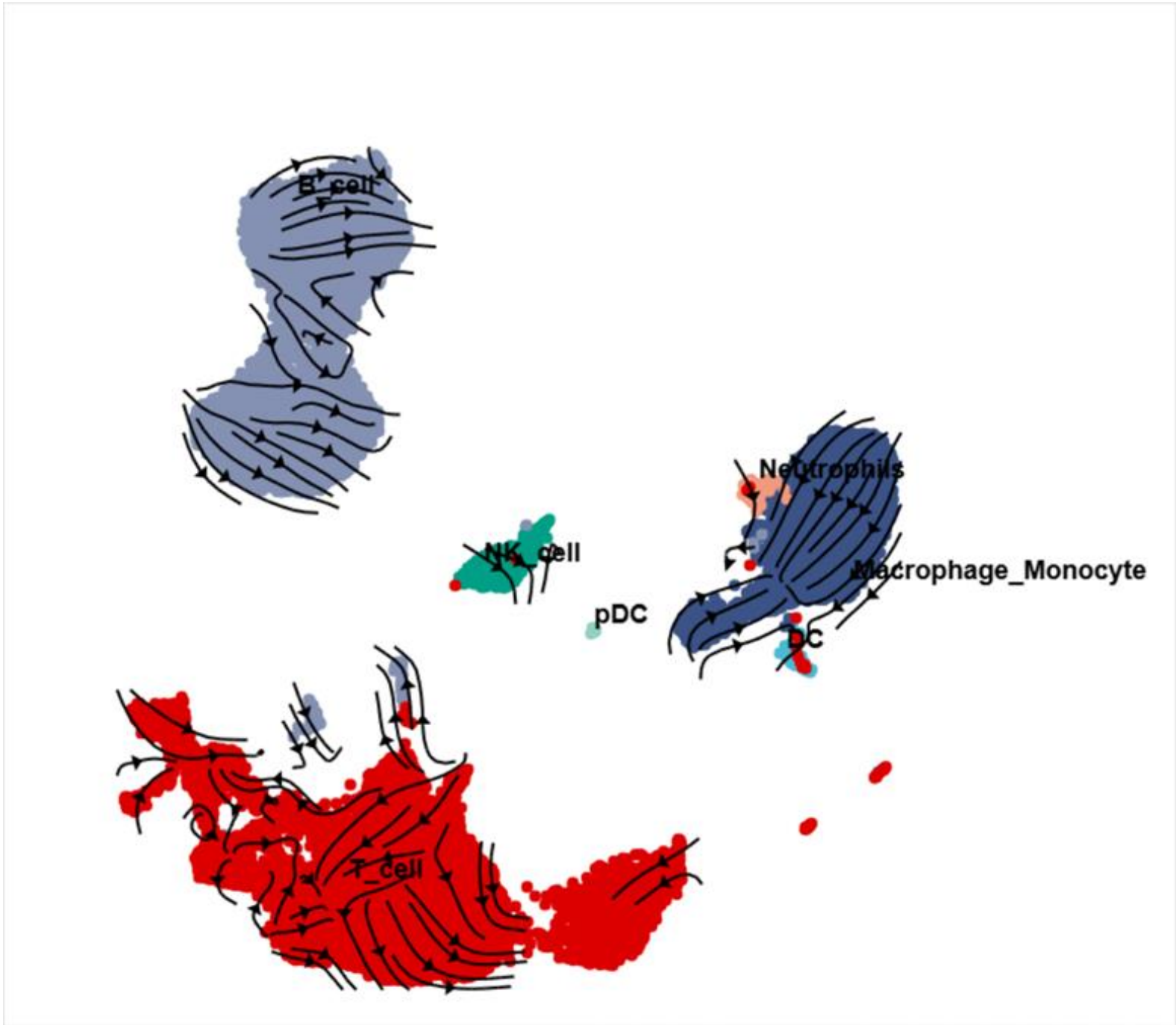
scVELO 整体结果表（estimated_rate.csv）处下载完整文件。

分化潜在时间（latent time）结果图



注：动力学模型恢复了潜在细胞过程的潜伏时间。这个潜伏时间代表了细胞的内部时钟，并且近似于细胞在分化时仅基于其转录动态所经历的实时时间。图中 0 到 1 的范围代表了细胞内部时钟的变化。0 代表细胞分化的早期，处于初始状态或干性状态，而 1 则代表细胞分化的末期或特化状态。

RNA 速率分化轨迹图



注：可以通过可视化二维降维顶部的矢量场，对所有基因和所有细胞的 RNA

速率进行广泛的可视化。箭头代表该处细胞速率预测的平均趋势，可以理解为细胞发育演化方向，不同颜色代表不同细胞类型。不需要人为指定分化的起点和终点。该图中每个点代表一个细胞，不同的颜色代表不同的细胞类型，而箭头的方向代表拟合的分化轨迹。

2.9.9 转录因子分析

转录因子（Transcription factor）是一类能与特异 DNA 序列结合的蛋白质，广泛存在于生物体中，能够通过特定的功能域（domain）识别并结合到基因上游调控区的顺式作用元件上，对基因的表达有激活或阻碍作用。转录因子通常由 DNA 结合功能域（DNA-binding domain, DBD）、转录活性功能域(Trans-activating domain, TAD)、其他转录因子结合功能域（Signal sensing domain, SSD）组成。根据功能域的不同，可将转录因子归为不同的转录因子家族，同一家族的转录因子一般含有相同的 DBD，不同的 TAD。

☆Tips：这部分可以进行后续的数据挖掘，老师在做完常规数据挖掘后，可以选择一些自己关心的差异表达基因进行转录因子预测，如果这些差异表达的基因中有转录因子，那么进一步还可以研究这些转录因子具体调控哪些基因的表达，即转录因子靶基因的预测。转录因子的调控导致基因表达的差异。与此同时，通常它们也是信号转导通路的靶标，也是被调控的对象。鉴于转录因子在生物体中发挥的重要作用，越来越多的研究人员开始着手转录因子研究，进而发现转录因子与多种人类重大疾病相关，如：p53、myc、ER、Id 与癌症相关；MEF2、HIF1 与心脏疾病相关；NFκB 与炎症相关；PPAR 与肥胖相关等。

2.9.9.1 转录因子预测

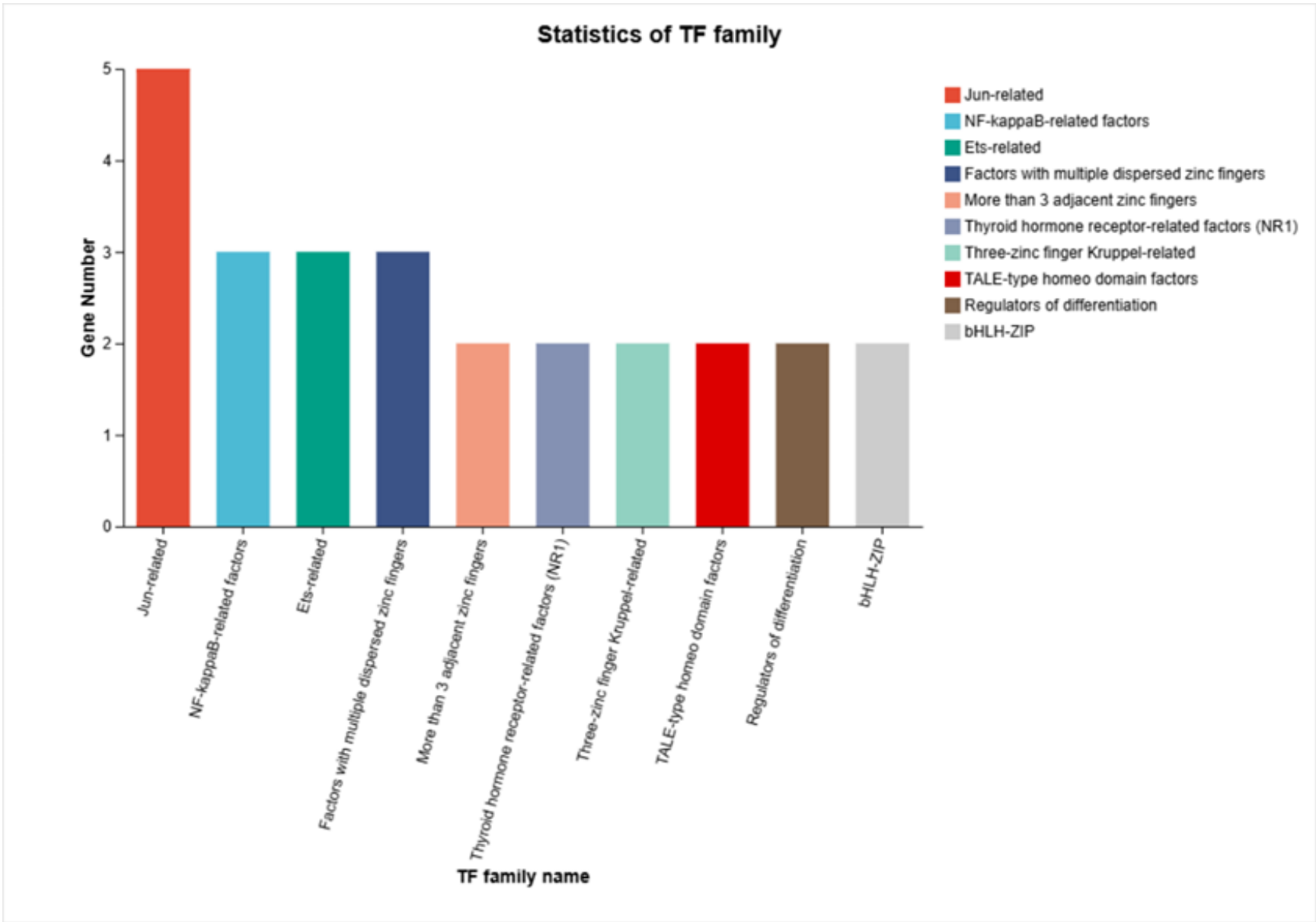
转录因子预测详情表

Gene id	Gene name	Gene descript...	Family	DNA domain	Domain descr...	Evalue	Score	TF id	Hit pident
ENSG00000003...	HOXC8	homeobox C8...	Homeobox					ENSP00000004...	
ENSG00000006...	GSC2	goosecoid ho...	Homeobox					ENSP00000008...	
ENSG00000008...	ZNF416	zinc finger pro...	zf-C2H2					ENSP00000019...	
ENSG00000010...	ZBED4	zinc finger BE...	zf-BED					ENSP00000021...	
ENSG00000010...	RHOXF1	Rhox homeob...	Homeobox					ENSP00000021...	
ENSG00000011...	MYF6	myogenic fac...	bHLH					ENSP00000022...	
ENSG00000011...	MYF5	myogenic fac...	bHLH					ENSP00000022...	
ENSG00000011...	ZBTB24	zinc finger an...	ZBTB					ENSP00000023...	
ENSG00000011...	HAND1	heart and neu...	bHLH					ENSP00000023...	
ENSG00000011...	CDX1	caudal type h...	Homeobox					ENSP00000023...	

注：（1）Gene id：基因编号；（2）Gene name：基因名称；（3）Gene description：基因描述信息；（4）TF name：转录因子名称；（5）Class：基于转录因子的 DNA 结合结构域对转录因子进行分类；（6）Family：基于功能或序列的相似性对转录因子进行分类；（7）Matrix id：转录因子结合位点的标识符。

2.9.9.2 转录因子家族统计

转录因子家族统计图（柱形图）



注：横坐标表示转录因子家族；纵坐标为属于该转录因子家族的基因数目。

2.9.9.3 转录因子靶基因预测

转录因子靶基因预测详情表

Matrix ID	TF gene id	TF gene name	Family	TF name	Target gene id	Target gene n...	Pvalue	Padjust	Start
MA1541.2	ENSG0000014...	NR6A1	GCMF-related...	NR6A1	ENSG0000018...	SYN3	2.05E-7	0.008	280
MA1541.2	ENSG0000014...	NR6A1	GCMF-related...	NR6A1	ENSG0000018...	SYN3	2.05E-7	0.008	294
MA1573.1	ENSG0000016...	THAP11	THAP-related...	THAP11	ENSG0000000...	ACSM3	9.82E-7	0.0591	461
MA0597.3	ENSG0000013...	THAP1	THAP-related...	THAP1	ENSG0000021...	IGHD	1.53E-5	0.0709	37
MA0597.3	ENSG0000013...	THAP1	THAP-related...	THAP1	ENSG0000011...	PHACTR1	1.53E-5	0.0709	120
MA0597.3	ENSG0000013...	THAP1	THAP-related...	THAP1	ENSG0000017...	P2RY14	1.53E-5	0.0709	122
MA0597.3	ENSG0000013...	THAP1	THAP-related...	THAP1	ENSG0000014...	CAMK2D	1.53E-5	0.0709	137
MA0597.3	ENSG0000013...	THAP1	THAP-related...	THAP1	ENSG0000021...	IGHM	1.53E-5	0.0709	162
MA0597.3	ENSG0000013...	THAP1	THAP-related...	THAP1	ENSG0000014...	CAMK2D	1.53E-5	0.0709	229
MA0597.3	ENSG0000013...	THAP1	THAP-related...	THAP1	ENSG0000014...	CAMK2D	1.53E-5	0.0709	326

【当数据库为 AnimalTFDB/PlantTFDB】时：（1）TF id：转录因子 ID；（2）TF gene id：编码转录因子对应基因 id；（3）TF gene name：编码转录因子对应的基因名称；（4）Family：转录因子家族（注：参数选择家族时出现）；（5）Motif id：数据库中转录因子对应的结合位点 id；（6）Target gene id：预测的转录因子潜在靶基因 id；（7）Target gene name：预测的转录因子潜在靶基因名称；【当数据库选择 JASPAR】时：（8）Matrix ID：转录因子结合位点的标识符；（9）TF gene id：转录因子基因 id；（10）TF gene name：转录因子基因名称；（11）Family：转录因子家族（注：参数选择家族时出现）；（12）TF name：转录因子名称；（13）Target gene id：靶标基因 id；（14）Target gene name：靶标基因名称；（15）Pvalue：显著性 P 值；（16）Padjust：校正后的 P 值；（17）Start：转录因子在靶基因上结合的起始位置；（18）Stop：转录因子在靶基因上结合的终止位置；（19）Matched sequence：转录因子结合靶基因的位点序列信息。

2.10 项目背景

2.10.1 基本信息

2.10.1.1 项目样本信息

展示该项目的样本基本信息，包括样品描述、样本名称以及组名。

项目样本信息表

Sample description	Sample initial name	Sample analysis name	Group name
样本信息表	DTC_Melanoma_10k	DTC_Melanoma_10k	DTC_Melanoma_10k
样本信息表	SI_PBMK_10k	SI_PBMK_10k	SI_PBMK_10k

注：展示本项目的样本基本信息，包括 Sample description（样品描述）、Sample initial/analysis name（样本初始/现用名称）以及 Group name（组名）。“样本描述”和“样本现用名称”可通过【批量编辑】进行再次编辑，修改后，页面展

示修改后的样本名。

2.10.1.2 分析软件信息

分析软件信息表

Soft/Database	Version	Analysis	Source
NR 数据库	Version 2023.07	基因注释	https://www.ncbi.nlm.nih.gov/publi...
GO 数据库	Version 2023.07	基因注释	http://www.geneontology.org/
NCBI 物种分类数据库	Version 2023.11	基因注释	ftp://ftp.ncbi.nlm.nih.gov/pub/taxon...
Swiss-prot 数据库	Version 2023.11	基因注释	ftp://ftp.uniprot.org/pub/databases...
Rfam 数据库	Version 14.10	核糖体比例评估	http://rfam.janelia.org/
eggNOG 数据库	Version 2020.06	基因注释	http://eggnogdb.embl.de/#/app/ho...
KEGG 数据库	Version 2023.09	基因注释	http://www.genome.jp/kegg/
Pfam 数据库	Version 36.0	基因注释	http://pfam.xfam.org/
Uniprot 数据库	Version 2023.11	基因注释	ftp://ftp.uniprot.org/pub/databases...
PIR idmapping 数据库	Version 2023.11	基因注释	ftp://ftp.pir.georgetown.edu/databa...
STRING数据库	v12.0	蛋白互作网络分析	https://string-db.org/
Seurat	4.1.1	降维与聚类	https://cloud.r-project.org/web/pac...
SingleR	1.8.1	singleR细胞注释	https://www.bioconductor.org/pack...
SCINA	1.2.0	maker注释(基于软件)	https://cran.r-project.org/web/pack...
Cell Ranger	v7.1.0	基础分析	https://support.10xgenomics.com/sin...
Monocle	2.22.0	拟时序分析	http://cole-trapnell-lab.github.io/m...
clustree	0.4.4	clusterTree分析/subclusterTree分析分...	https://mirrors.sjtug.sjtu.edu.cn/cran...
goatools	Version 1.4.4	基因集分析（GO富集）	https://pypi.org/project/goatools/
CellChat	Version 1.5.0	细胞通讯分析-CellChat	https://github.com/sqjin/CellChat
GSEA(clusterProfiler)	Version 4.6.2	GSEA分析	https://bioconductor.org/packages/...
GSVA	Version 1.52.3	GSVA分析	https://www.bioconductor.org/pack...
hdWGCNA	Version 0.2.15	hdWGCNA分析	https://smorabit.github.io/hdWGCNA...
Diamond	Version 2.1.9	基因功能注释	https://github.com/bbuchfink/diamo...
fastp	Version 0.23.4	测序数据质控	https://github.com/OpenGene/fastp
MSigDB 数据库	Version 2023.2	GSEA分析	https://docs.gsea-msigdb.org/#MSig...
SCENIC	Version 1.3.1	转录因子分析-SCENIC	https://github.com/aertslab/SCENIC
SciBet	version 1.0	SciBet细胞注释	https://github.com/YeehanXiao/scib...
scATOMIC	v2.0.3	恶性细胞scATOMIC分析	https://github.com/abelson-lab/scA...
infercnv	1.10.1	cnv分析	https://github.com/broadinstitute/in...
cytotrace	0.3.3	细胞分化潜能预测cytotrace分析	https://cytotrace.stanford.edu
PlantPhoneDB	1.0.0	细胞通讯-PlantPhoneDB	https://github.com/Jasonxu0109/Pla...
scVelo	0.2.4	RNA速率分析	https://github.com/theislab/scvelo
scanpy	1.9.1	降维与聚类	https://scanpy.readthedocs.io/en/st...
Seeksoultools	需与工作流页面联动	基础分析	https://www.seekgene.com/fxrj
JASPAR	Version 2024	转录因子预测	https://jaspar.elixir.no/
AnimalTFDB	Version 4.0	转录因子预测	http://bioinfo.life.hust.edu.cn/Anima...
PlantTFDB	Version 5.0	转录因子预测	http://planttfdb.gao-lab.org/

注：展示单细胞转录组生信分析中用到的软件名称/数据库、版本及其对应使用的分析项，并给出各软件的来源。【[点击链接可跳转到相关网页](#)】

三、备注

3.1 结果文件查看说明

结果文件格式	结果文件解释说明	结果文件查看方式
PNG	结果图像文件，位图，无损压缩。	unix/Linux/Mac用户使用display命令打开。
		windows用户可以使用图片浏览器打开，如photoshop等。
JPEG	结果图像文件，位图，有损压缩。	unix/Linux/Mac用户使用display命令打开。
		windows用户可以使用图片浏览器打开，如photoshop等。
SVG	结果图像文件，矢量图，可放大、缩小不失真，方便用户查看和编辑处理，可使用Adobe Illustrator进行编辑，用于文章发表等。	unix/Linux/Mac用户使用display命令打开。
		windows用户可以使用图片浏览器打开，如photoshop等。
xls, CSV	结果数据表格结果，文件以制表符（Tab）分隔	unix/Linux/Mac用户使用 less 或 more 命令查看；
		windows用户使用高级文本编辑器 Editplus/Notepad++等，也可以用Microsoft Excel打开。
PDF	结果图像文件，矢量图，可放大、缩小不失真，方便用户查看和编辑处理，可使用Adobe Illustrator进行编辑，用于文章发表等。	unix/Linux用户使用evince命令打开。
		windows/Mac用户可以使用Adobe Reader/福昕阅读器/网页浏览器等打开。

四、参考文献

[1].Le, D. T. et al. Mismatch repair deficiency predicts response of solid tumors to PD1 blockade. Science 357, 409 – 413 (2017). [2].Satija, R. et al. Spatial reconstruction of single-cell gene expression data. Nat. Biotechnol. 33, 495 – 502 (2015). [3].Satija R, Farrell JA, Gennert D, et al. Spatial reconstruction of single-cell gene expression data. Nature biotechnology 2015, 33(5):495-502.

[4].Stuart T, Butler A, Hoffman P, et al. Comprehensive Integration of Single-Cell Data[J]. Cell 2019, 177(7):1888-1902.e21. [5].Maaten L, Hinton G. Visualizing data using t-SNE. Journal of machine learning research 2008, 9:2579-2605.

[6].Blondel VD, Guillaume JL, Lambiotte R, et al. Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment 2008, 2008(10):P10008.

[7].Zhang Ze,Luo Danni,Zhong Xue et al. SCINA: A Semi-Supervised Subtyping Algorithm of Single Cells and Bulk Samples [J]. Genes (Basel), 2019, 10: undefined.