第 3 次作业

崔晏菲 自硕 21 2021210976

1. 代码环境
   a) 编程语言：R 4.2.0
   b) 依赖包：Seurat = 4.3.0, SingleR = 2.0.0, celldex = 1.8.0, ggplot2 = 3.4.0, cowplot = 1.1.1, dplyr = 1.0.10, scCancer = 2.2.1, stCancer = 0.1.0, scPred = 1.9.2, magrittr = 2.0.3, hdf5r = 1.3.7
   c) 代码和图片结果详见文件 Homework3_code.ipynb，聚类后得到的 marker gene、stCancer 运行结果详见 output 文件夹。

2. scRNA-Seq 数据处理
   a) 读取数据
   使用 Seurat 包中的 Read10X 读取 P2 和 P3 的 scRNA-Seq 数据，创建 SeuratObject 类，并添加基本属性，代码如下：

1. P2 data

```
# read P2 data
P2_counts <- Read10X(data.dir = scDataPath_P2)
P2_object <- CreateSeuratObject(counts = P2_counts)
P2_object$status <- 'Tumor'
P2_object$batch <- 'P2'
P2_object$datatype <- 'HCC-sc'
```

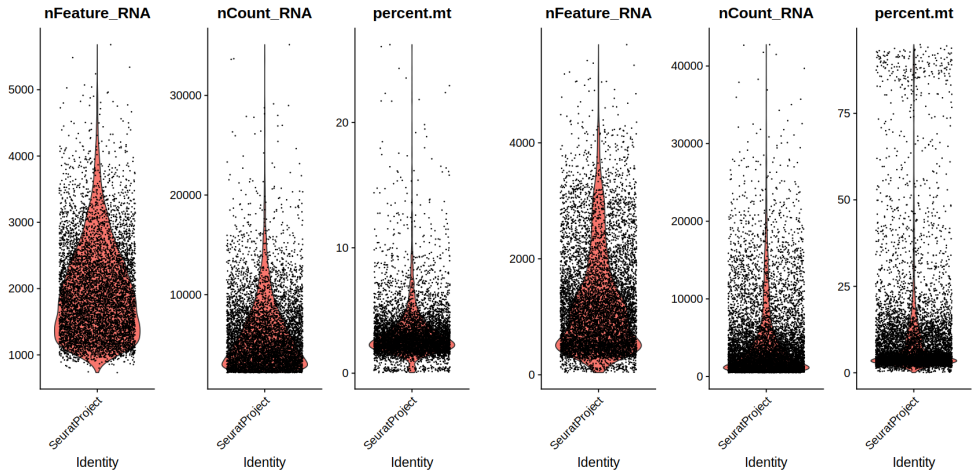
✓ 7.2s

2. P3 data

```
# read P3 data
P3_counts <- Read10X(data.dir = scDataPath_P3)
P3_object <- CreateSeuratObject(counts = P3_counts)
P3_object$status <- 'Tumor'
P3_object$batch <- 'P3'
P3_object$datatype <- 'HCC-sc'
```

✓ 4.6s

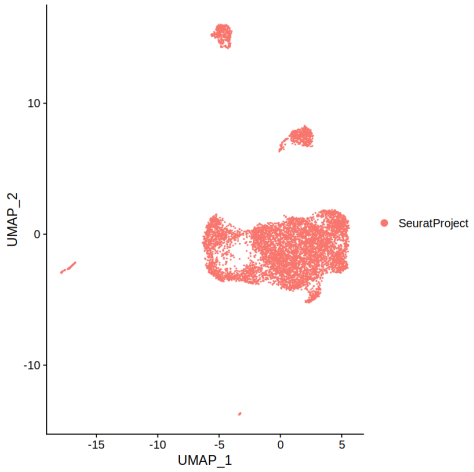
   b) 初始数据可视化
   计算两个数据集中线粒体基因的比例，并使用 VlnPlot()进行可视化，观察数据集 Feature_RNA, Count_RNA 以及线粒体基因比例的分布情况，可视化结果如下：



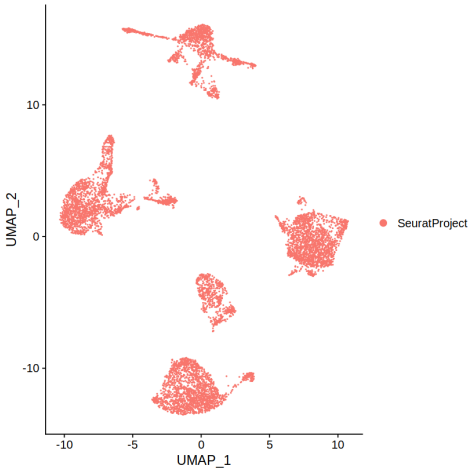
P2 样本分布图

P3 样本分布图

   c) Seurat 流程
   首先根据 nFeature_RNA, nCount_RNA 以及 percent.mt 的值对数据进行质控， 接着对数据进行归一化、再寻找高变基因、接着进行 PCA 降维找出最重要的特征、再使用 UMAP 降维进行可视化，结果如下：

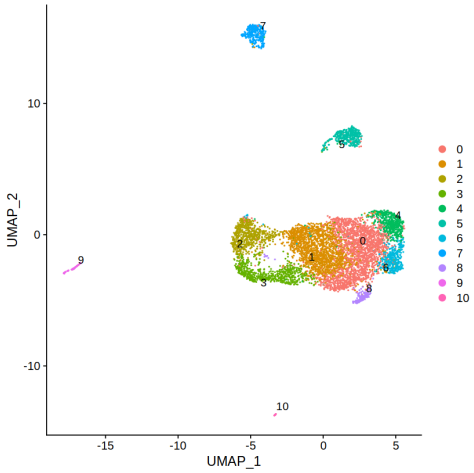


P2 UMAP

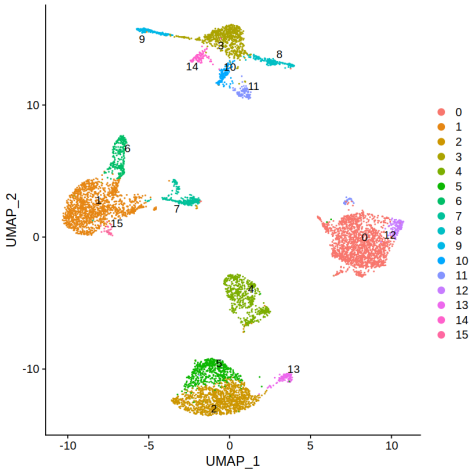


P3 UMAP

接着进行无监督聚类，聚类结果如下，可以看到 P2 数据集被分为了 11 类。P3 数据集被分为了 16 类。

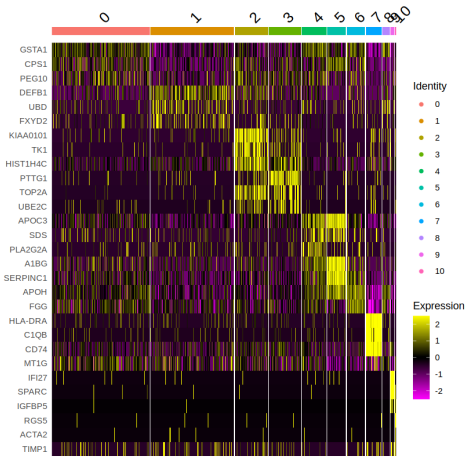


P2 clustering

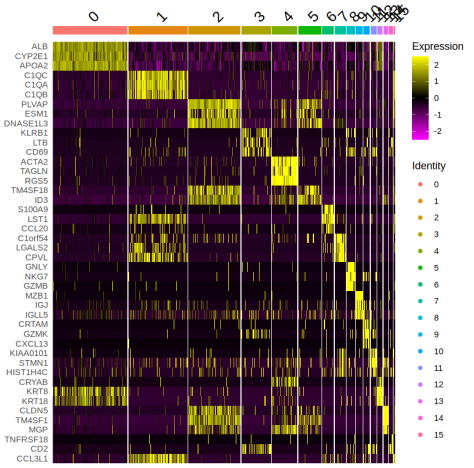


P3 clustering

接着再分析两个数据集中每个 cluster 的 marker gene，得到的结果详见文件夹 output/，本报告中只展示其热图，结果如下：



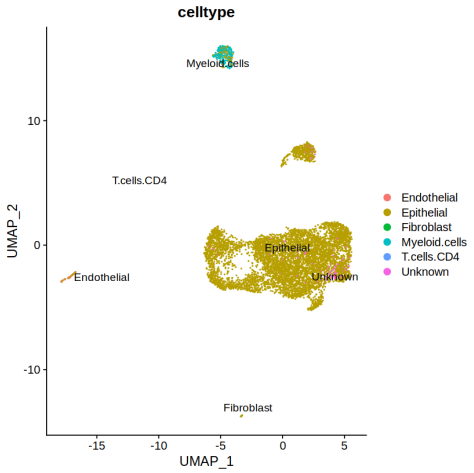
P2 heatmap



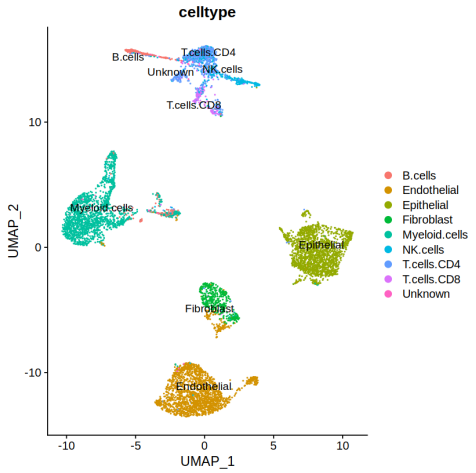
P3 heatmap

d) 细胞注释

使用 scCaner 对两个数据集进行自动标注，并画出标注结果，如下图所示。可以看到 P2 数据集的细胞被分成了 Endothelial, Epithelial, Fibroblast, Myeloid.cells, T.cells.CD4 以及 Unknown 六种，而 P3 数据集的细胞被分成了 B.cells, Endothelial, Epithelial, Fibroblast, Myeloid.cells, NK.cells, T.cells.CD4, T.cells.CD8 以及 Unknown 九种。说明我们在聚类的阈值上的选取和 scCancer 略有不同。



P2 scCancer



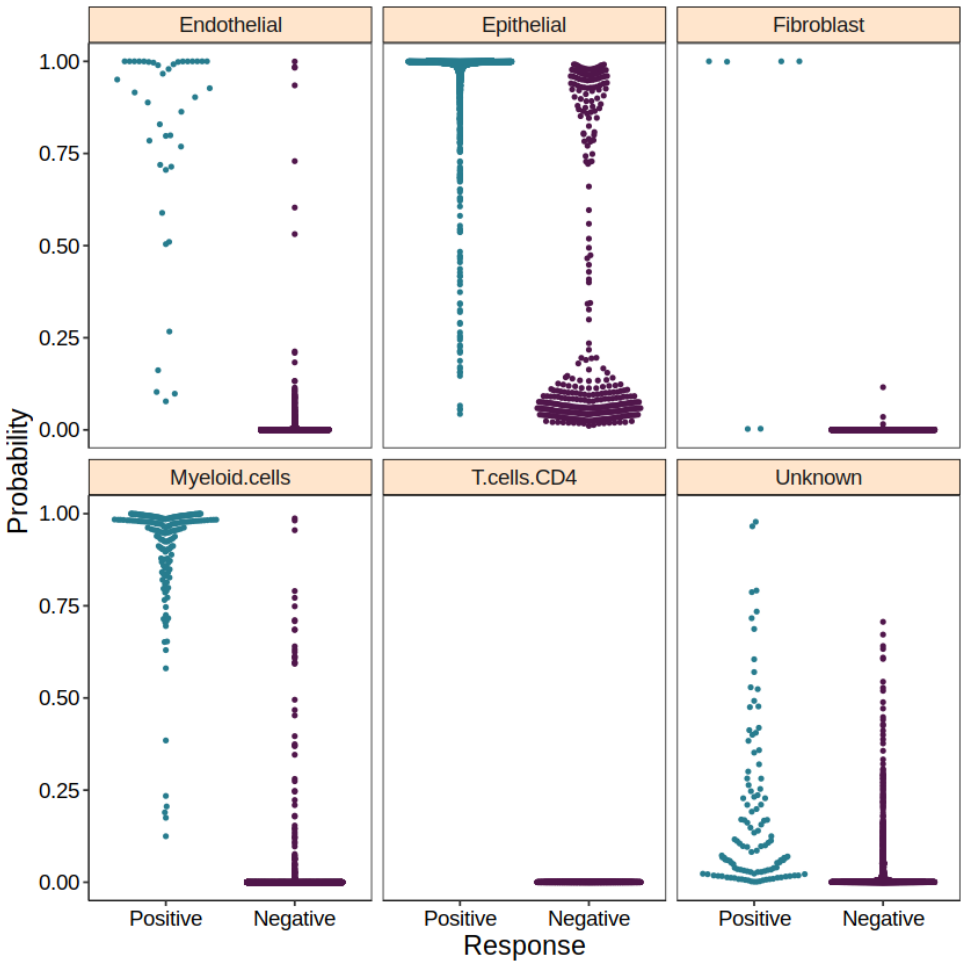
P3 scCancer

e) 训练分类器

使用 scPred 包中的 SVM 分类器用于训练。使用 P2 作为训练集，P3 作为测试集，scCancer 的标注结果作为 groundtruth。训练集正确率以及训练样本分布如下图所示：

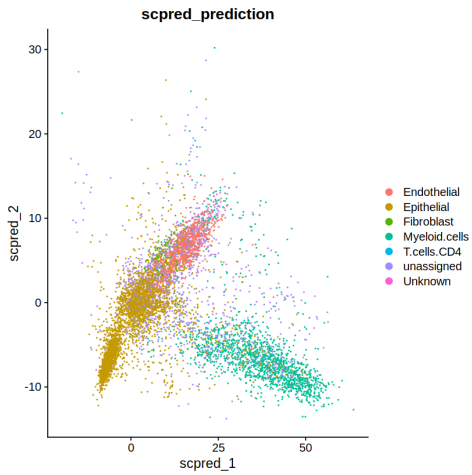
Cell type	n	Features	Method	Accuracy	Kappa
Endothelial	40	30	svmRadial	0.998	0.853
Epithelial	6345	30	svmRadial	0.978	0.767
Fibroblast	6	30	svmRadial	1.000	0.600
Myeloid.cells	215	30	svmRadial	0.996	0.943
T.cells.CD4	2	30	svmRadial	1.000	NaN
Unknown	113	30	svmRadial	0.983	0.152

训练集准确率



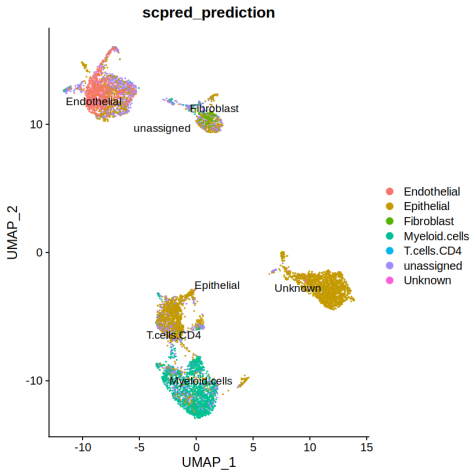
训练集样本分布

接着再使用训练好的模型对测试集进行分类，分类结果如下：

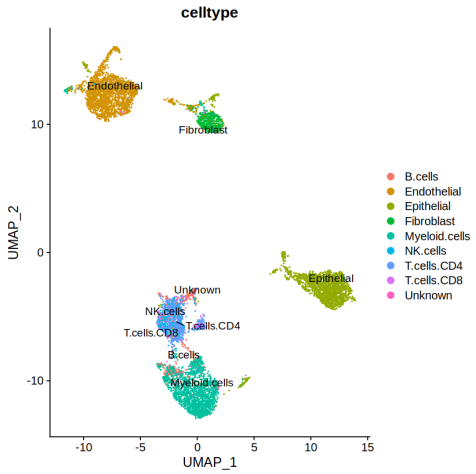


SVM 预测结果

将 SVM 预测结果和 scCancer 进行对比，发现准确率为 55.70%，具体对比结果如下：



SVM 预测结果



scCancer 分类结果

可以发现准确率低的主要原因是 SVM 预测结果中有很多 Unknown。

3. 空间转录组数据处理

a) 读取数据

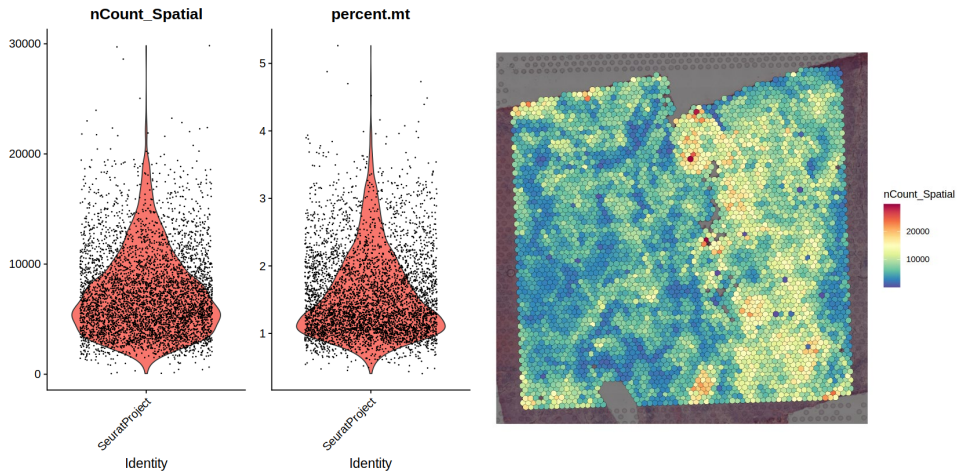
使用 Seurat 包的 Read10X 读取空间转录组数据 L2，创建 SeuratObject，并添加基本属性（包括 image），代码如下所示：

```
3. ST data

# read ST data
L2_counts <- Read10X_h5(filename = stDataPath_2L)
L2_object <- CreateSeuratObject(counts = L2_counts, assay = "Spatial")
img <- Read10X_Image(image.dir = stDataPath_2L_ST)
DefaultAssay(object = img) <- 'Spatial'
img <- img[colnames(x = L2_object)]
L2_object[['image']] <- img
✓ 3.3s
```

b) 初始数据可视化

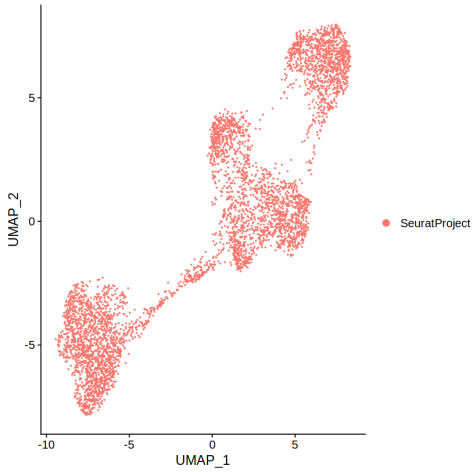
计算两个数据集中线粒体基因的比例，并使用 VlnPlot()进行可视化，观察数据集，nCount_Spatial 以及线粒体基因比例的分布情况，并观 nCount_Spatial 在空间上的分布，可视化结果如下：



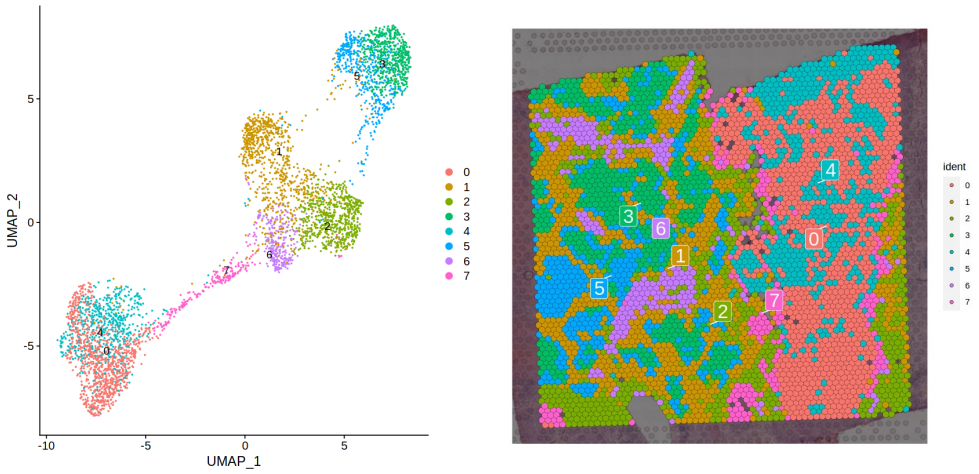
c) Seurat 流程

根据 nCount_Spatial 以及 percent.mt 的值对数据进行质控，接着对数据进行归一化、然后寻找高变基因、再进行 PCA 降维提取关键特征、再用

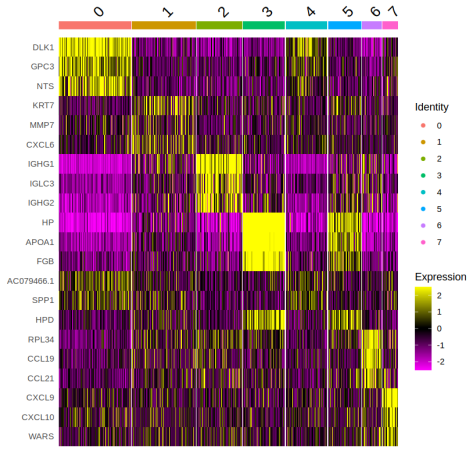
UMAP 降维进行可视化，得到的结果如下：



然后进行无监督聚类，得到的结果如下，可以看到数据被分成了 8 类：

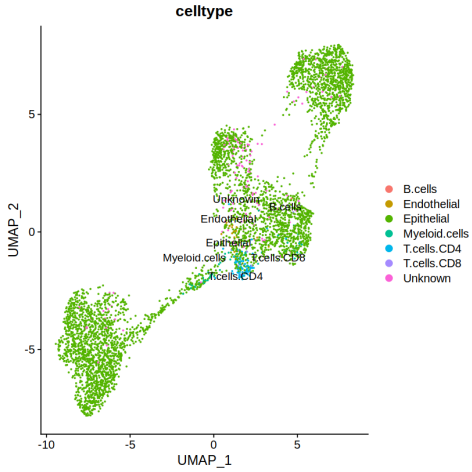


接着再分析每个类别的 marker gene，得到的热图如下：



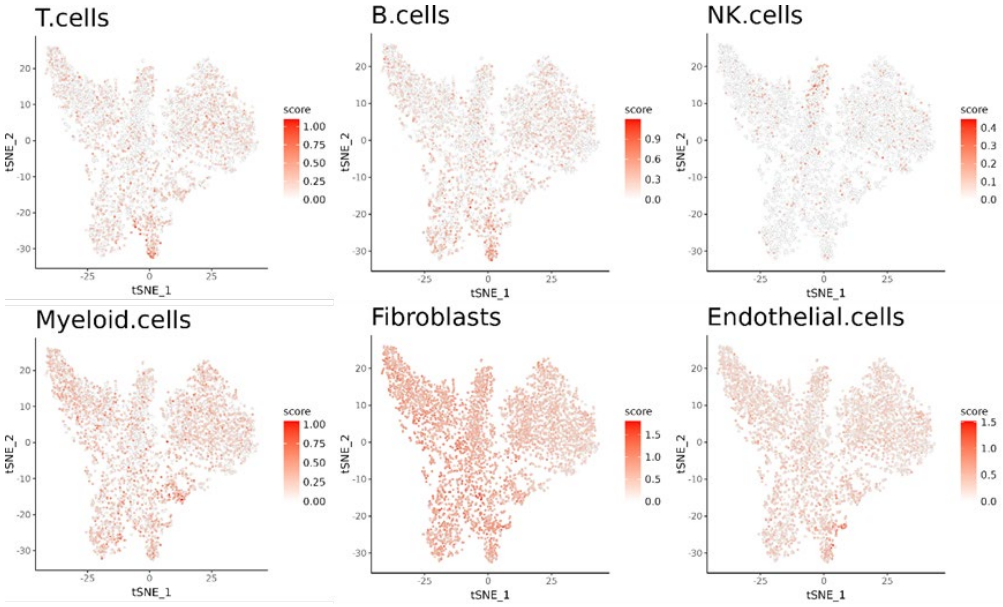
d) 细胞注释

使用 scCaner 包对 L2 数据集进行自动标注，得到结果如下：

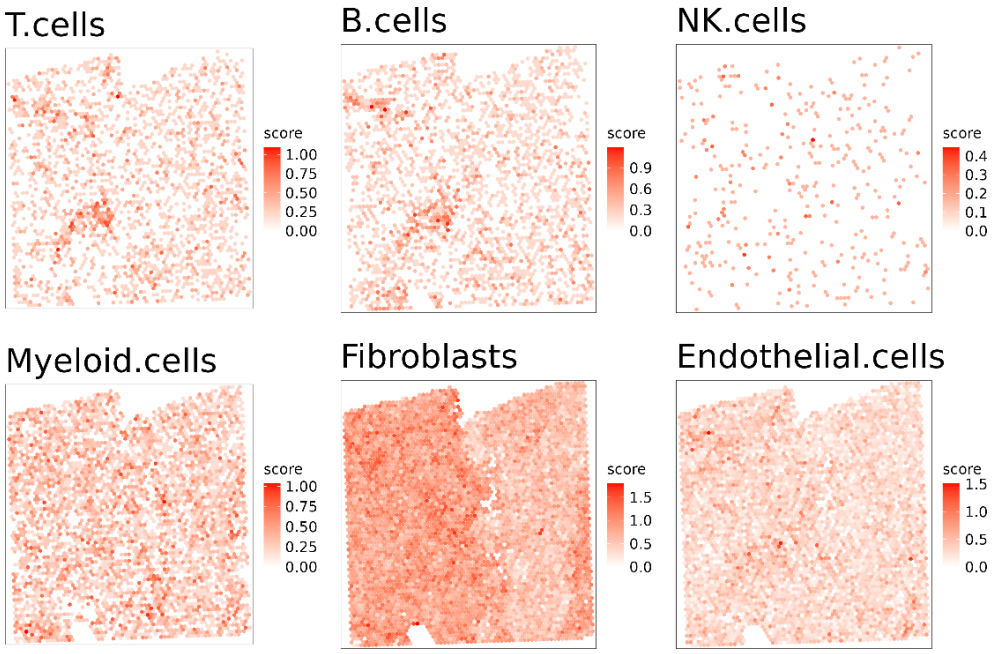


e) 定位主要细胞类型

使用 stCancer 包对 L2 数据进行处理，在物理空间和 t-SNE 空间上各自对主要的细胞类型进行定位，得到的结果如下：



每种细胞在 t-SNE 空间上的分布



每种细胞在物理空间上的分布