

# Step 系列：scRNA-seq 癌细胞鉴定

2024-02-21

LiChuang Huang



@ 立效研究院

# Contents

<b>1 摘要</b>	<b>1</b>
1.1 目的 . . . . .	1
1.2 解决的问题 (技术性的) . . . . .	1
<b>2 前情资料</b>	<b>1</b>
<b>3 适配性</b>	<b>1</b>
<b>4 方法</b>	<b>1</b>
<b>5 安装 (首次使用)</b>	<b>1</b>
5.1 安装依赖 . . . . .	1
<b>6 示例分析</b>	<b>2</b>
6.1 数据准备 . . . . .	2
6.1.1 快速获取示例数据 . . . . .	2
6.2 分析流程 . . . . .	2
6.2.1 单细胞数据的质控、聚类、Marker 鉴定、细胞注释等 . . . . .	2
6.2.2 癌细胞鉴定 . . . . .	2
6.2.2.1 As-job-kat 将前处理完毕的 <code>job_seurat</code> 数据对象转化 . . . . .	2
6.2.2.2 Step1 根据变异拷贝数鉴定癌细胞 . . . . .	3
6.2.2.3 Step2 可视化鉴定结果 . . . . .	3
6.2.2.4 (额外的) 保存 <code>job_kat</code> 并输出结果 . . . . .	4
6.2.2.5 Map 将结果映射回 Seurat . . . . .	5
6.2.3 拟时分析 . . . . .	6
6.2.3.1 do-monocle 对癌细胞进行拟时分析 . . . . .	7
6.2.3.2 Step1 构建拟时轨迹 . . . . .	7
6.2.3.3 Step2 选择拟时起点 . . . . .	8
6.2.3.4 Step3 拟时分析基础上的差异分析和基因表达模块 . . . . .	9
6.2.3.5 (进阶) 根据拟时分析结果重新划分细胞群体 . . . . .	9
6.3 完整示例代码 . . . . .	9
<b>Reference</b>	<b>9</b>

## List of Figures

1 Copykat prediction . . . . .	5
2 The scsa copykat . . . . .	6
3 SCSA Cell type annotation . . . . .	6
4 Cancer cell prin . . . . .	8
5 Cut tree . . . . .	9

**List of Tables**

1    CopyKAT results . . . . . 4

# 1 摘要

## 1.1 目的

解决癌组织单细胞数据集中，癌细胞鉴定的难题，并延续一般性的单细胞数据分析流程。

## 1.2 解决的问题（技术性的）

不同的 R 包或其他工具之间的数据转换和衔接。

# 2 前情资料

这份文档是以下的补充资料（以下的配置和使用是前提条件）：

- 《Step 系列：scRNA-seq 基本分析》

如果你还不知道 Step 系列的基本特性以及一些泛用的提取和存储方法，请先阅读：

- 《Step 系列：Prologue and Get-start》

# 3 适配性

同《Step 系列：scRNA-seq 基本分析》。

# 4 方法

以下是我在这个工作流中涉及的方法和程序：

Mainly used method:

- R package `copyKAT` used for aneuploid cell or cancer cell prediction<sup>1</sup>.
- R package `Monocle3` used for cell pseudotime analysis<sup>2,3</sup>.
- The R package `Seurat` used for scRNA-seq processing; `SCSA` (python) used for cell type annotation<sup>4-6</sup>.
- Other R packages (eg., `dplyr` and `ggplot2`) used for statistic analysis or data visualization.

# 5 安装（首次使用）

## 5.1 安装依赖

《Step 系列：scRNA-seq 基本分析》已经列举了大部分程序的安装方法，以下，仅展示 R 包 `copykat` 的安装。

R input

```
devtools::install_github("navinlabcode/copykat")
```

## 6 示例分析

在《Step 系列：scRNA-seq 基本分析》中，我以 GSE171306 做了 scRNA-seq 一般性的示例。但其实 GSE171306 是一批癌组织数据集，理应鉴定癌细胞，而 SC3 和多数其他自动注释工具，都无法鉴定出癌细胞。这样，就有必要专门鉴定癌细胞了。

以下针对癌细胞鉴定的问题展开示例分析，并依然使用 GSE171306 这批数据。

注：只要是《Step 系列：scRNA-seq 基本分析》中提及的，以下就不再赘述；只细述新的内容。

### 6.1 数据准备

#### 6.1.1 快速获取示例数据

运行以下代码获取数据 (和《Step 系列：scRNA-seq 基本分析》中的是一样的)：

```
R input

geo <- job_geo("GSE171306")
geo <- step1(geo)
geo <- step2(geo)
untar("./GSE171306/GSE171306_RAW.tar", exdir = "./GSE171306")
prepare_10x("./GSE171306/", "ccRCC1", single = F)
sr <- job_seurat("./GSE171306/GSM5222644_ccRCC1_barcode")
```

### 6.2 分析流程

#### 6.2.1 单细胞数据的质控、聚类、Marker 鉴定、细胞注释等

以下直接给出代码 (和《Step 系列：scRNA-seq 基本分析》相同)：

```
R input

sr <- step1(sr)
sr <- step2(sr, 0, 7500, 35)
sr <- step3(sr, 1:15, 1.2)
sr <- step4(sr, "")
sr <- step5(sr, 5)
sr <- step6(sr, "Kidney")
```

#### 6.2.2 癌细胞鉴定

##### 6.2.2.1 As-job-kat 将前处理完毕的 job\_seurat 数据对象转化

```
R input

kat <- asjob_kat(sr)
```

注意，这一步默认将 `sr@object` (也就是 `seurat` 数据对象) 中的 `assays` 数据槽中的第一个数据集用于鉴定癌细胞。转化完成后，你会在 `kat@object` 看到这个矩阵数据集。一般情况下，使用的都是第一个数据集。

你可以手动指定数据集，例如：

```
R input

# 不要运行
kat <- asjob_kat(sr, use = names(x@object@assays)[[1]])
```

#### 6.2.2.2 Step1 根据变异拷贝数鉴定癌细胞

`copykat` 有一个优点 (相对于 `inferCNV`)，不需要手动指定参考细胞，程序会自主在数据集中选择参考细胞。所以，将所有的细胞表达数据输入就可以了。

(这一步会比较耗时，1-5 小时)

```
R input

# 后一个参数指定线程数
kat <- step1(kat, 8)
# 你还可以通过添加 `path` 参数，指定其他文件夹用以存放中间数据
```

这里，我有必要做一些说明。`copykat` 内置一些参考数据集，用以参考和计算。对于人类，它内置的是 ‘hg20’ 参考集。目前遇到更多的可能是 ‘hg38’。`copykat` 内部并不会对基因符号 (symbol) 后缀的版本信息进行替换和匹配，因此，如果输入的基因即使是同一种，然而因为版本不同，也会无法匹配上 (例如，你输入的是 ‘AHR.5’，内置的参考集包含的是 ‘AHR.1’)。

因此，我在 `job_kat` 的 `step1` 中补充了这一部分工作，能够让这一步能够顺利进行下去。但我无法预料是否还会有其他特殊情况，所以以上事项需知悉。

如果你同样对以上事项保持警惕，可以用以下确认我做了哪些补充工作：

```
R input

selectMethod(step1, "job_kat")
```

我还需要备注的是，由于我只做到过人类的癌细胞鉴定，所以小鼠的数据集目前还不支持。

#### 6.2.2.3 Step2 可视化鉴定结果

`copykat` 自带可视化的函数；然而，由于其中的图例绘制的太糟糕，因此，这里我去除了 `copykat` 绘制的热图的图例，重新添加了一个新的图例。

这一步也会比较耗时 (热图很大)。

```
R input

kat <- step2(kat)
```

该热图可以直接提取查看；但最好不要这么做，因为加载这个热图太耗时：

- `kat@plots$step2$p.copykat`

推荐的做法是（见 6.2.2.4），运行下一部分（`clear`）之后，再将输出的 png 图片查看。

这里，我们也可以直接获取鉴定结果表格：

R input

```
kat@tables$step2$res_copykat
```

Table 1: CopyKAT results

cell.names	copykat.pred	copykat_cell
AAACCCAAGCTGTGCC-1	diploid	Normal cell
AAACCCACAGCCGGTT-1	diploid	Normal cell
AAACCCATCATGAGGG-1	diploid	Normal cell
AAACGAAAGTCACAGG-1	diploid	Normal cell
AAACGAACACACAGAG-1	diploid	Normal cell
AAACGAACACCTCTAC-1	diploid	Normal cell
AAACGAATCAACACGT-1	diploid	Normal cell
AAACGAATCACTACTT-1	diploid	Normal cell
AAACGAATCGGCATAT-1	aneuploid	Cancer cell
AAACGCTAGACAGTCG-1	diploid	Normal cell
AAACGCTAGTCCCAAT-1	aneuploid	Cancer cell
AAACGCTCATCAGTCA-1	diploid	Normal cell
AAACGCTCATGACTCA-1	diploid	Normal cell
AAACGCTGTAGGACCA-1	diploid	Normal cell
AAACGCTGTGAGCAGT-1	diploid	Normal cell
...	...	...

#### 6.2.2.4 （额外的）保存 job\_kat 并输出结果

R input

```
kat <- clear(kat)
```

这样，就能取得想要的结果了：

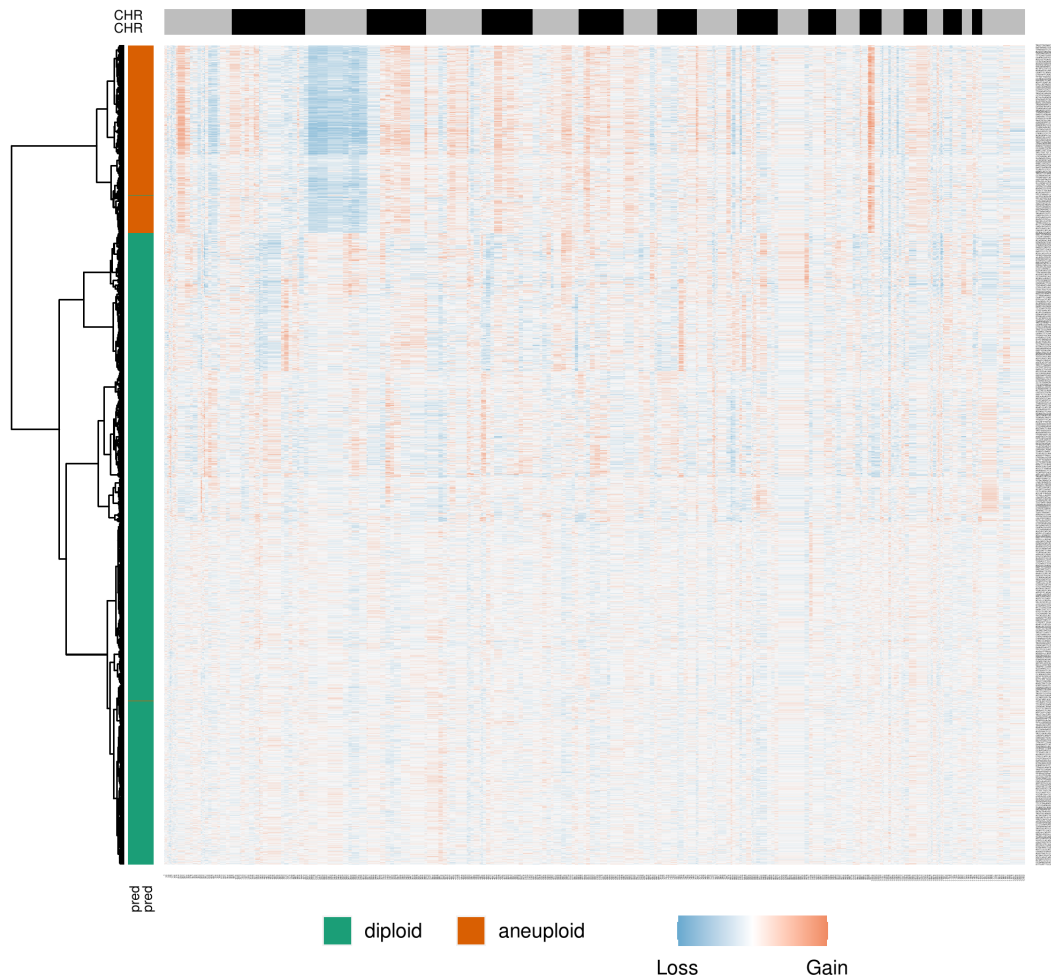


Figure 1: Copykat prediction

Fig. 1, 图中的 ‘aneuploidy’ 即为癌细胞。

#### 6.2.2.5 Map 将结果映射回 Seurat

R input

```
sr <- map(sr, kat)
p.sr_vis <- vis(sr, "scsa_copykat")
p.sr_vis
```

这样我们就能在 Fig. 2 中看到，被注释为 ‘Cancer cell’ 的细胞群体。



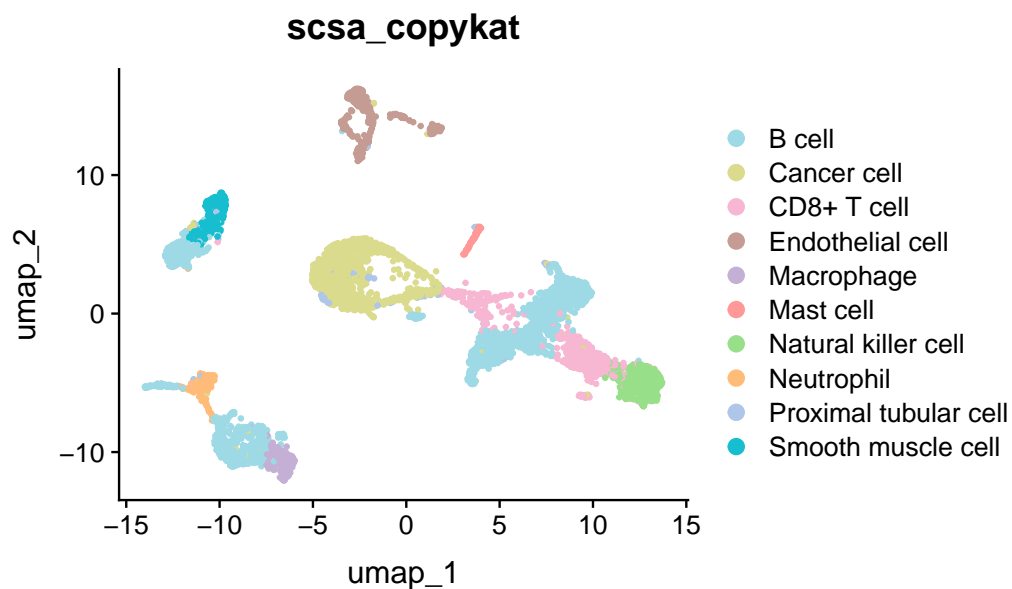


Figure 2: The scsa copykat

我们不妨对比一下 SCSA 的注释结果 (Fig. 3)，看看 ‘Cancer cell’ 可能的来源细胞。

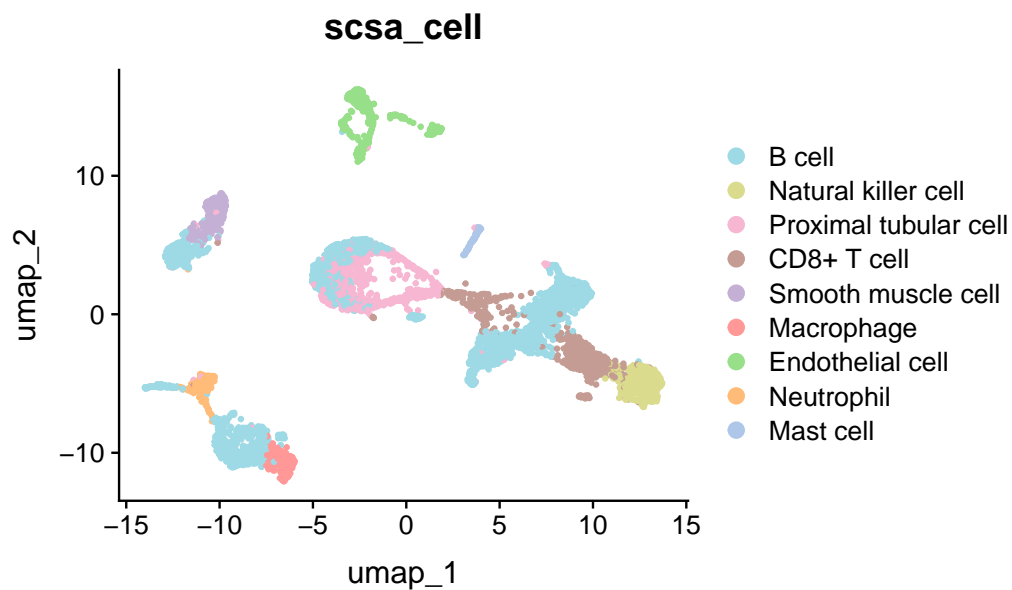


Figure 3: SCSA Cell type annotation

现在可以推断，‘Cancer cell’ 主要来源于 ‘Proximal tubular cell’。

### 6.2.3 拟时分析

其实到 6.2.2.5 为止，本文档的主要内容，即鉴定癌细胞，已经结束了。

然而如果我们进一步思考，如果将 copykat 依据变异拷贝数鉴定癌细胞的原理推进，结合拟时分析，也许

就能分析出，癌细胞或正常细胞是如何沿着‘拟时轨迹’，转变成癌细胞了。

这会是一种可以泛用于肿瘤组织单细胞数据的分析方法。

### 6.2.3.1 do-monocle 对癌细胞进行拟时分析

按照上述思路，这里需要将癌细胞单独取出，用以拟时分析。

我提供了一个便捷的方法，以快速达成这一目的：

R input

```
mn <- do_monocle(sr, kat)
```

在整个 Step 系列方法中，do\_\* 形式的目前还很少；这是一种根据传入的前两个参数的类来决定调用的函数的系列方法。而 asjob\_\* 形式的，只根据第一种参数来决定。

其实，do\_monocle 内部重新对传入的 sr 数据运行了 step3，即对分离的癌细胞重新聚类，区分出更多的群体（可能会是亚型）

### 6.2.3.2 Step1 构建拟时轨迹

R input

```
mn <- step1(mn)
```

Fig. 4 是用来确认选择拟时起点的。

R input

```
# 以 wrap 调整了长宽比例  
wrap(mn@plots$step1$p.prin, 6, 4)
```

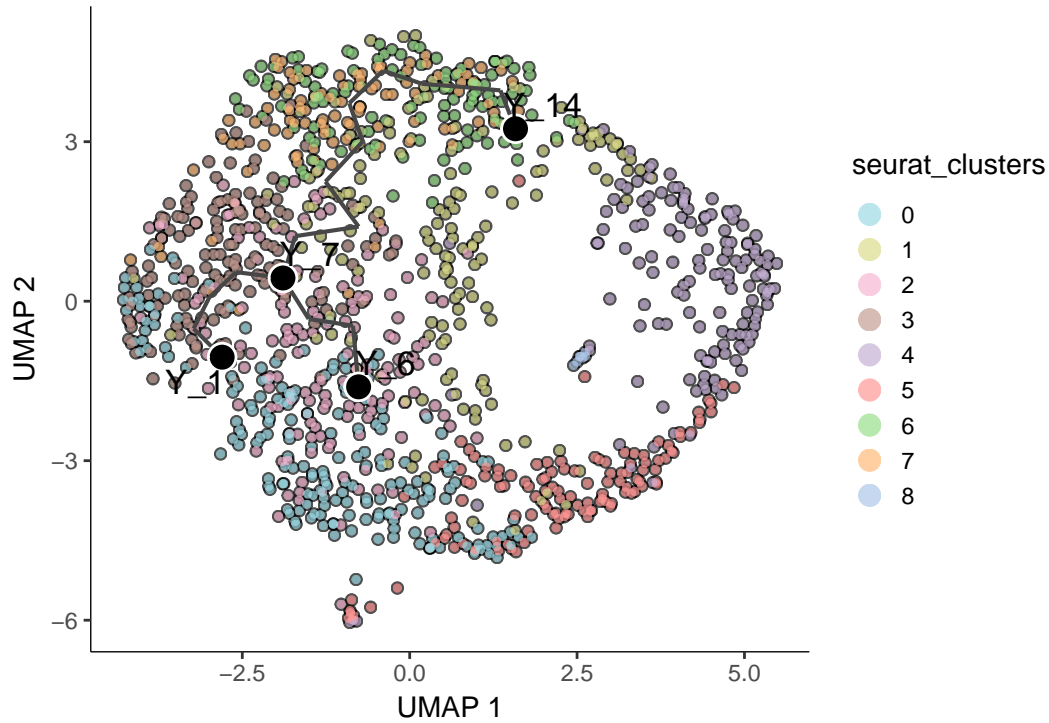


Figure 4: Cancer cell prin

### 6.2.3.3 Step2 选择拟时起点

选择合适的拟时起点依然会是一个难题。这里提供了一种可借鉴的，用以选择癌细胞拟时起点的思路。见 Fig. 1, 对于 ‘aneuploid’, 即癌细胞, ‘gain’, ‘loss’ 水平更接近 ‘diploid’ 的细胞, 或许更适合作为拟时起点。具体而言, 根据 Fig. 1 的侧边聚类树, 我们或许可以试着将肿瘤细胞切分为多个小群体, 然后根据它们相较于正常细胞的近似程度, 选择拟时起点。那么切分之后, 哪一群体更加接近正常细胞呢?

见 Fig. 1, 如果切分为两个群体, 下方高度 (Height) 更低的群体更近似正常细胞。这样, 我们就能大致决定: Height 更低的群体作为拟时起点。

我们可以把 copyKAT 的聚类结果映射到 UMAP 聚类上: `mn` 来自于 `do_monocle(sr, kat)`, 相较于《Step 系列: scRNA-seq 基本分析》中所述的, 有一点特殊之处, 即, 额外生成了图片, 也就是我们需要的映射图, 只不过, 它一共做了 1-30 种切分 (注意, 该切分是针对 Fig. 1 中的所有细胞进行的切分, 而不是单单癌细胞)。

R input

```
mn@plots$step1$p.cancer_position
```

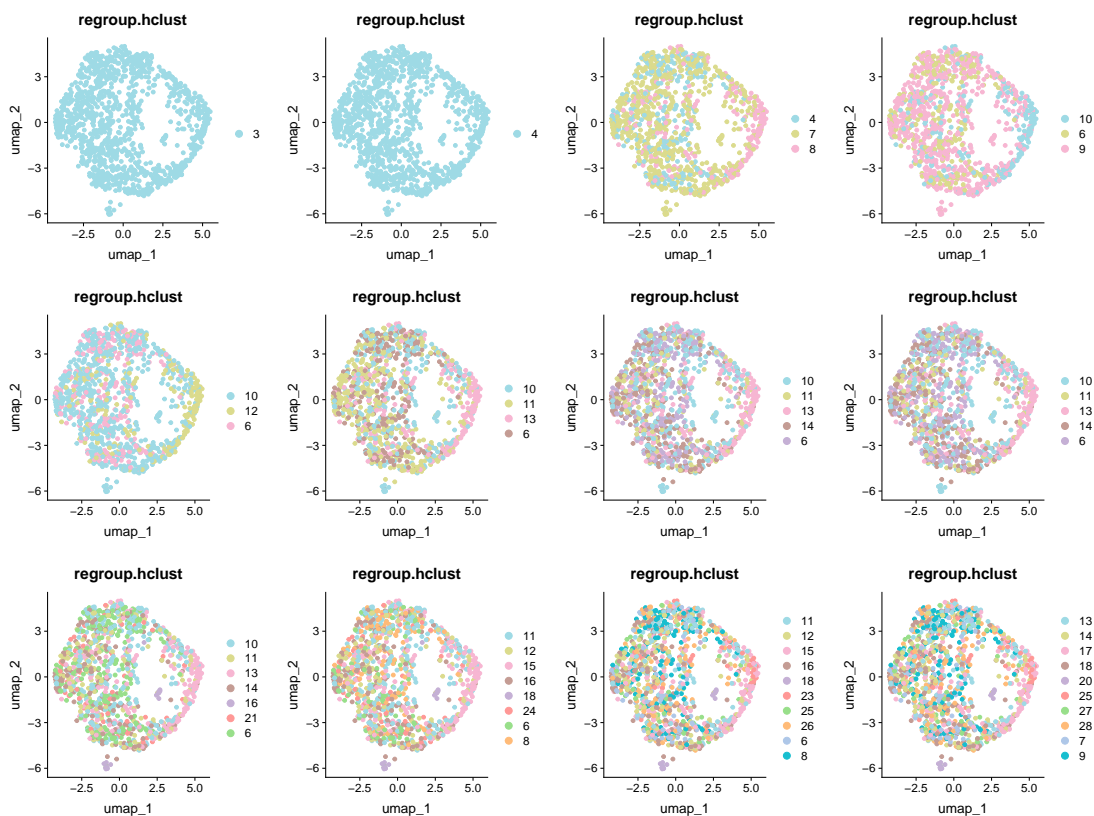


Figure 5: Cut tree

见 Fig. 5，图中的数值越小，代表 Height 越低。如果我们观察最后一副子图，会发现‘9’代表的群体，主要集中在 UMAP 的上半部分。因此，这里我们试着将更上半部分的细胞作为拟时起点。那么，也就是 Fig. 4 中的‘Y\_14’。

#### 6.2.3.4 Step3 拟时分析基础上的差异分析和基因表达模块

#### 6.2.3.5 (进阶) 根据拟时分析结果重新划分细胞群体

### 6.3 完整示例代码

## Reference

1. Gao, R. *et al.* Delineating copy number and clonal substructure in human tumors from single-cell transcriptomes. *Nature Biotechnology* **39**, 599–608 (2021).
2. Qiu, X. *et al.* Reversed graph embedding resolves complex single-cell trajectories. *Nature Methods* **14**, (2017).
3. Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature Biotechnology* **32**, (2014).
4. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, (2021).
5. Stuart, T. *et al.* Comprehensive integration of single-cell data. *Cell* **177**, (2019).

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