BerryGenomics
贝瑞基因

ScRNA-seq之图谱绘制和Marker鉴定

王　鹏

科技服务事业部

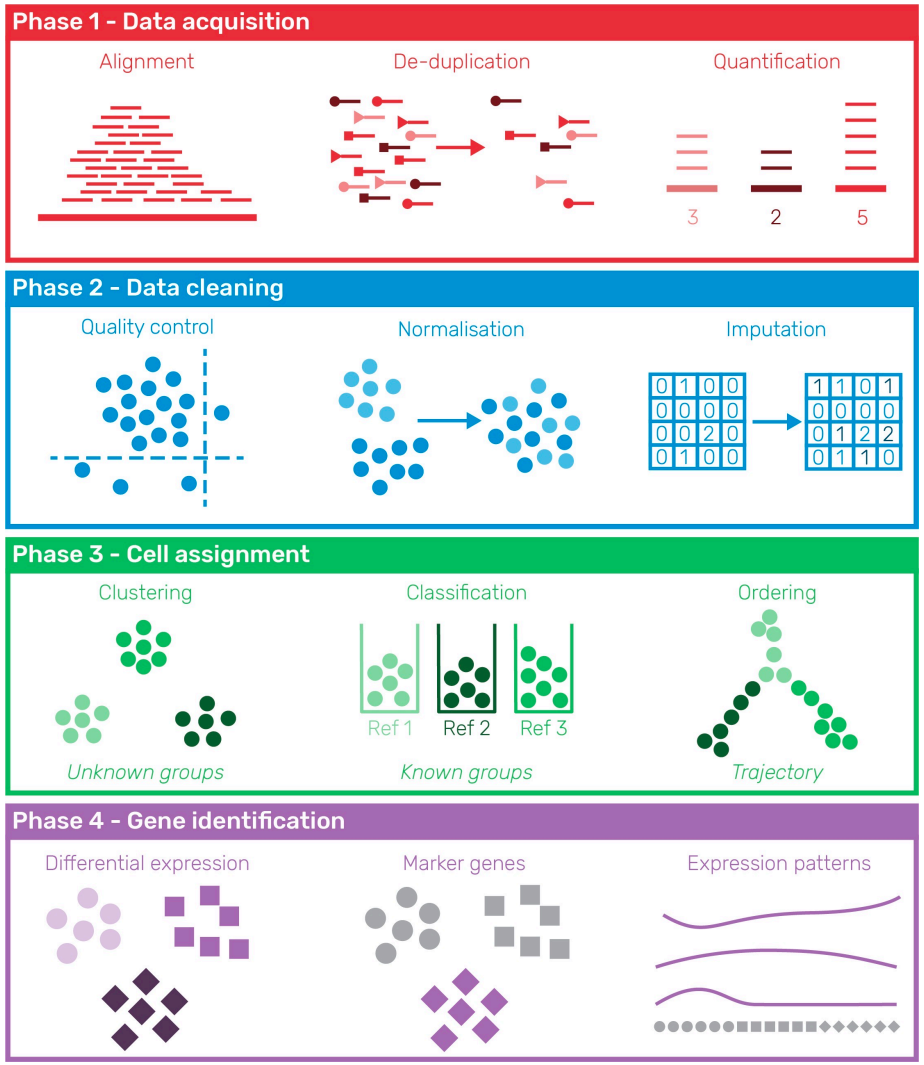
数据分析流程

数据获取

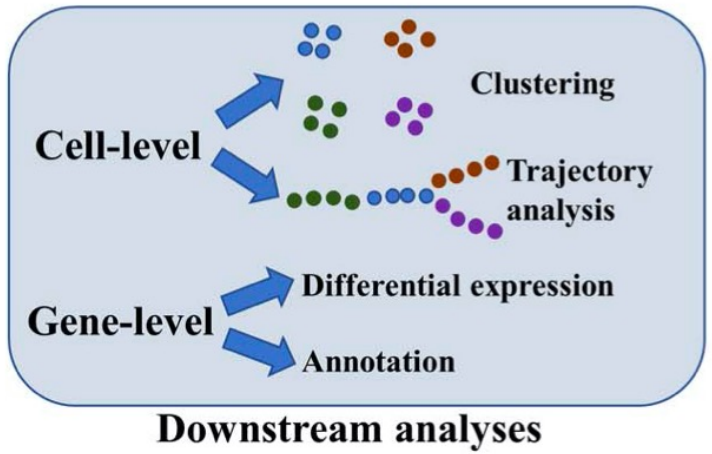
数据质控

细胞聚类

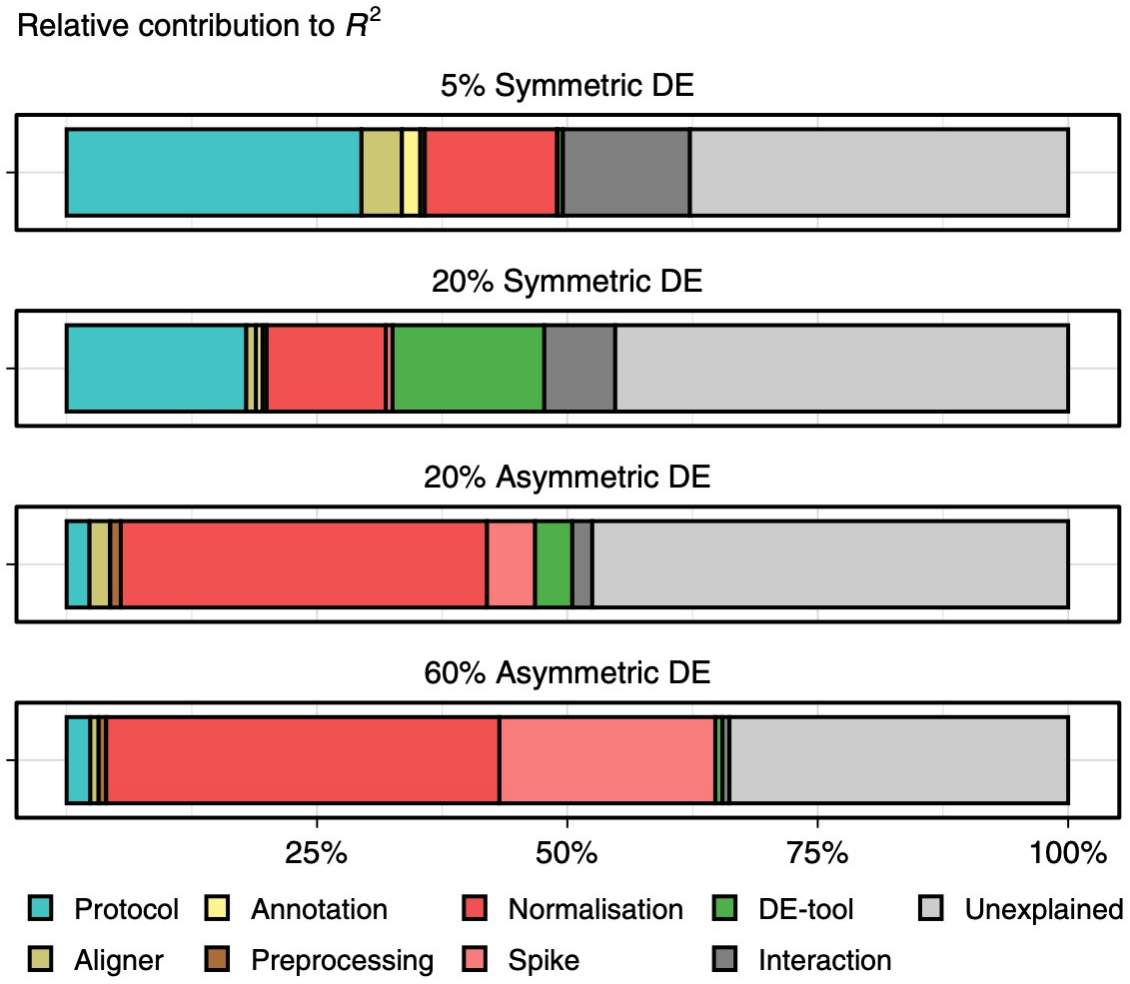
基因分析



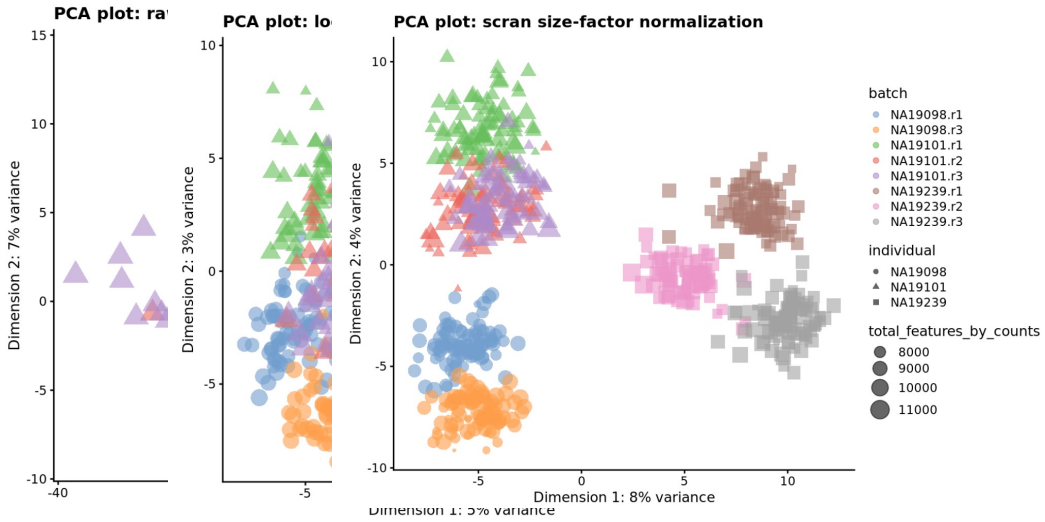
	Cell1	Cell2	...	CellN
Gene1	3	2	.	13
Gene2	2	3	.	1
Gene3	1	14	.	18
...	.	.	.	.
...	.	.	.	.
...	.	.	.	.
GeneM	25	0	.	0



聚类之前为什么要标准化



Normalisation is overall the most influential step. Because we tested a nearly exhaustive number of ~3000 possible scRNA-seq pipelines, starting with the choice of library preparation protocol and ending with DE-testing, we can estimate the contribution of each separate step to pipeline performance for our different DE-settings (Fig. 1b). We used a beta regression model to explain the variance in pipeline performance with the choices made at the seven pipeline steps (1) library preparation protocol, (2) spike-in usage, (3) alignment method, (4) annotation scheme, (5) preprocessing of counts, (6) normalisation and (7) DE-tool as explanatory variables. We used the



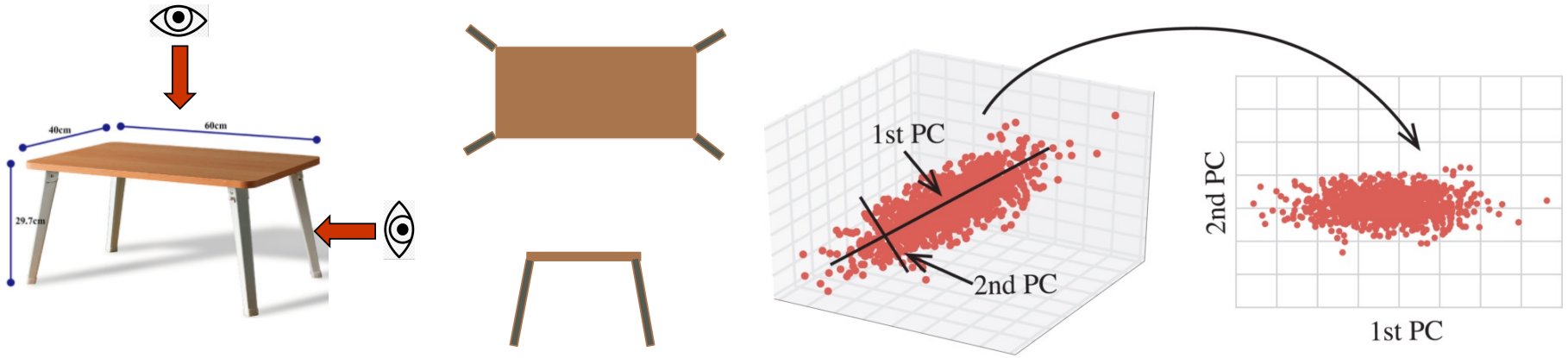
doi: 10.1038/s41467-019-12266-7

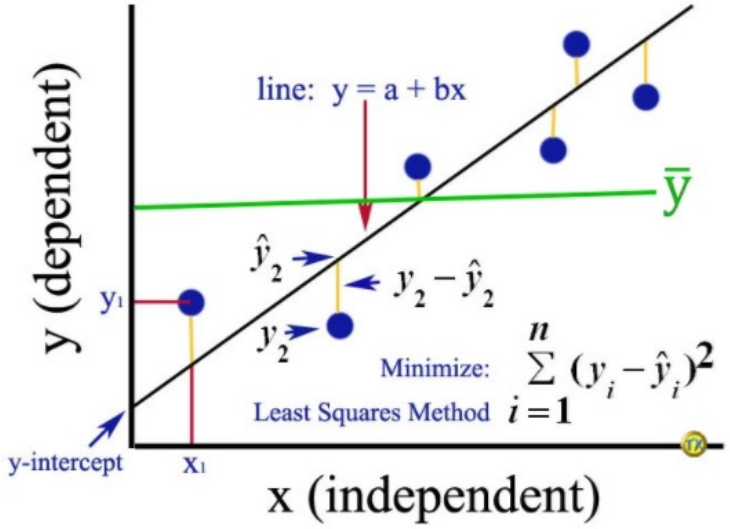
基于细胞的分析-图谱绘制（聚类或降维）

- PCA（principle components analysis）
- t-SNE（t-distributed stochastic neighbor embedding）
- UMAP（Uniform Manifold Approximation and Projection）

PCA

High dimension → Low dimension





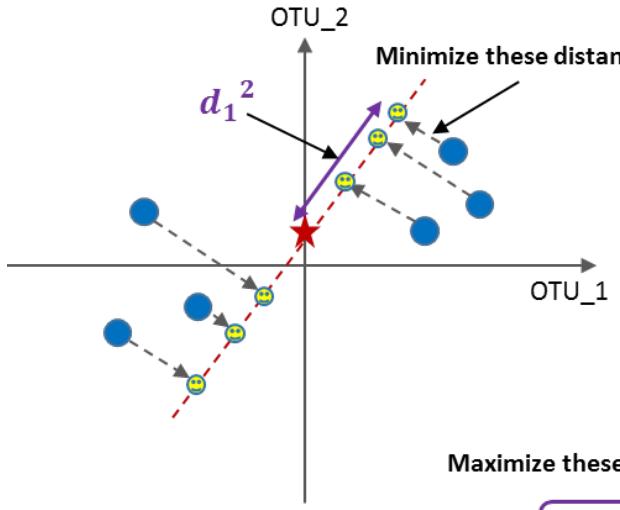
线性回归：

Step1：计算中心点（x和y的平均值）

Step2：线性拟合（最小二乘）

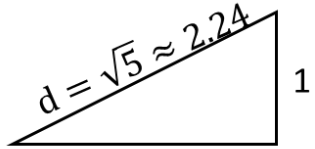
Step3：计算斜率b、截距a和决定系数R^2

PCA

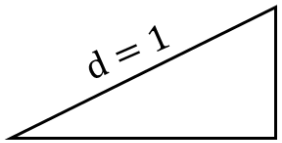


Maximize these distances

$$d_1{}^2 + d_2{}^2 + d_3{}^2 + d_4{}^2 + d_5{}^2 + d_6{}^2 = d_{total}{}^2$$

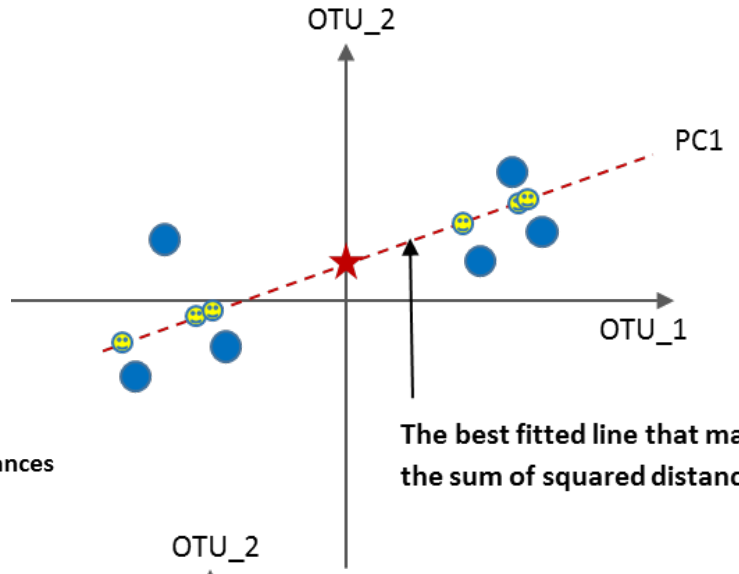


Scaled distance (d) to 1

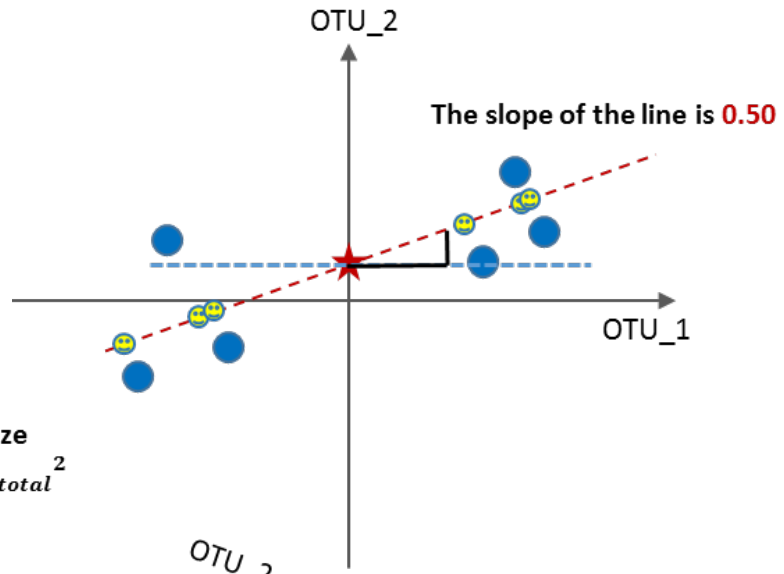


1/2.24=0.446

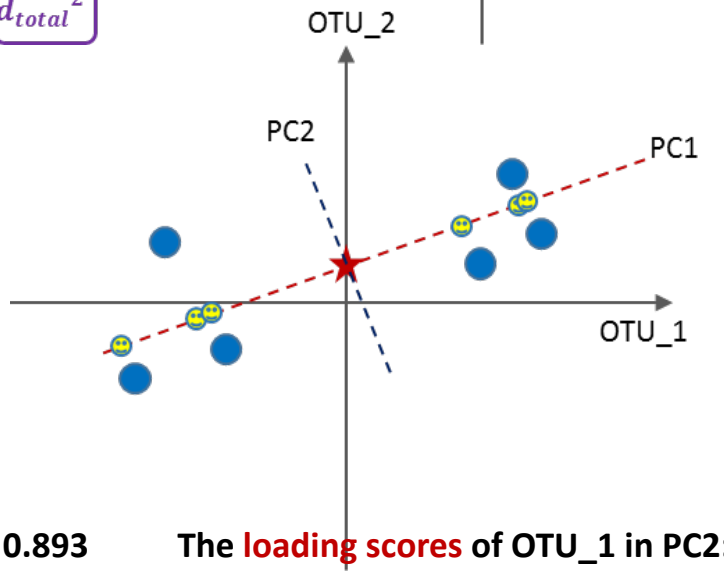
2/2.24=0.893

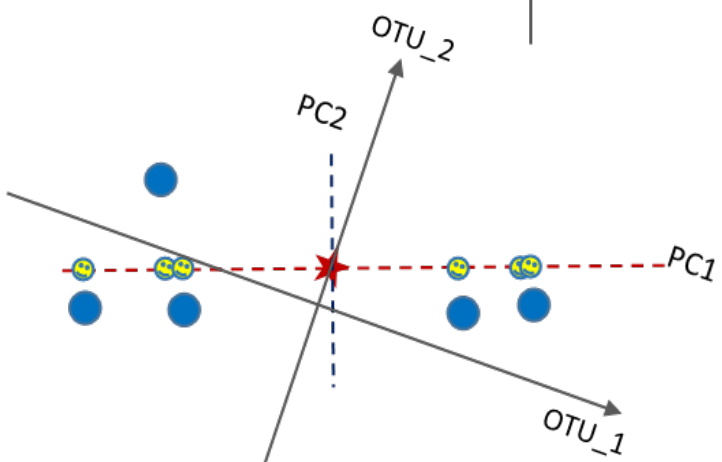


The best fitted line that maximize the sum of squared distances $d_{total}{}^2$



The slope of the line is 0.50





Eigenvalue: sum of squared distances $d_{total}{}^2$

Variation for PC1 = $d_{pc1_{total}}{}^2/(n-1)$

Variation for PC2 = $d_{pc2_{total}}{}^2/(n-1)$

The loading scores of OTU_1 in PC1: 0.893

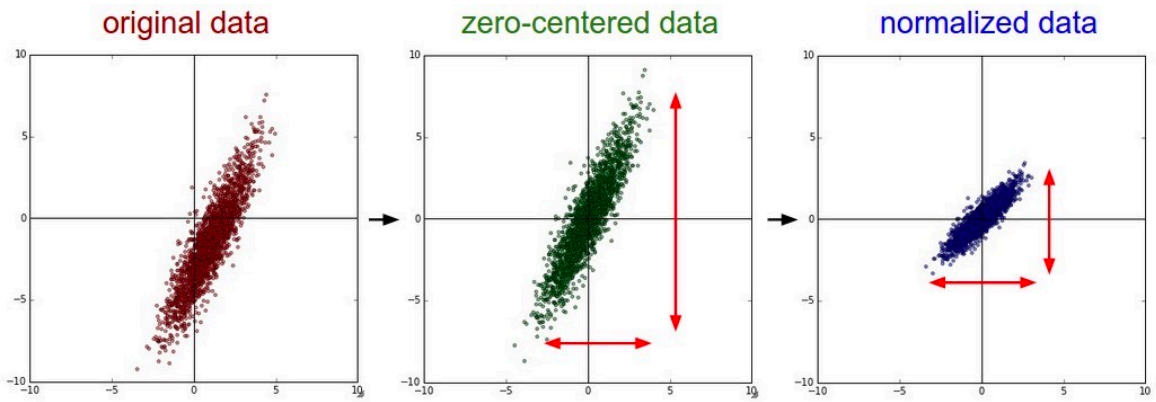
The loading scores of OTU_2 in PC1: 0.446

The loading scores of OTU_1 in PC2: xxx

The loading scores of OTU_2 in PC2: xxx

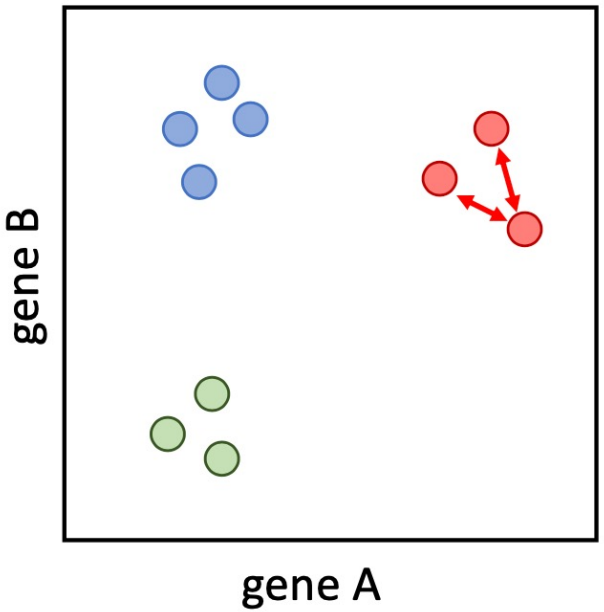
PCA-总结

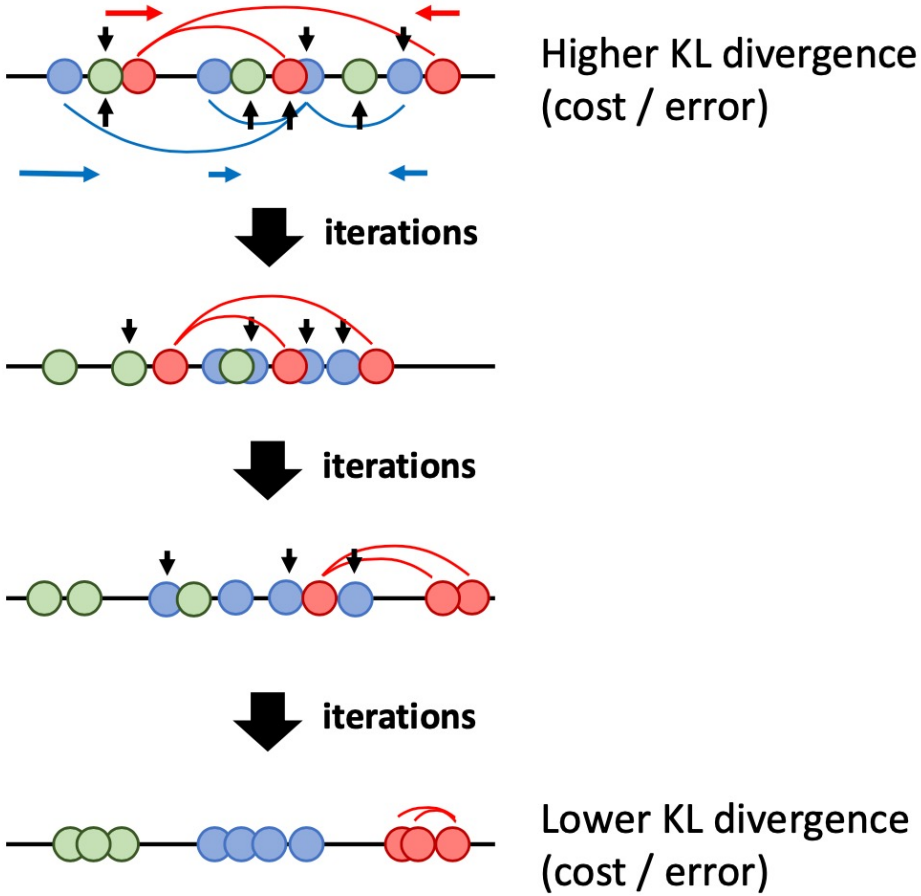
- 基于线性的降维方法
- PCA分析之前需要进行标准化（z-score）
- PCA的前几个主成分包含了数据的大部分变化信息
- 可以选取top n的PCs用于数据过滤（5-30 PCs）
- 对于0膨胀数据表现很差



使数据平均值=0，标准差=1，近似符合正态分布，方便下游分析

t-SNE





随机：开始的投射点是随机的；邻域嵌入：与相似的点不断靠近；t-分布：点与点之间的距离以t分布衡量

t-SNE总结

- 基于非线性的图形降维方法
- 通常取PCA的前20-30个PCs分析

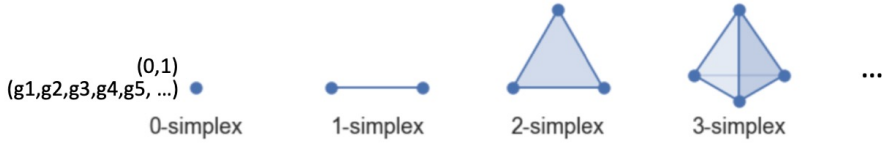
- 新样本点的加入会降低准确度
- 不能够保留全局的数据结构
- 只能嵌入到2-3维的数据中
- 耗时久，计算内存高

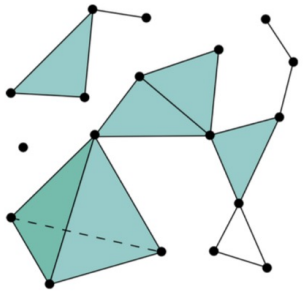
UMAP

It is based on **topological structures** in multidimensional space (simplices)

Points are connected with a line (edge) if the distance between them is below a threshold:

- Any distance metric can be used (euclidean)





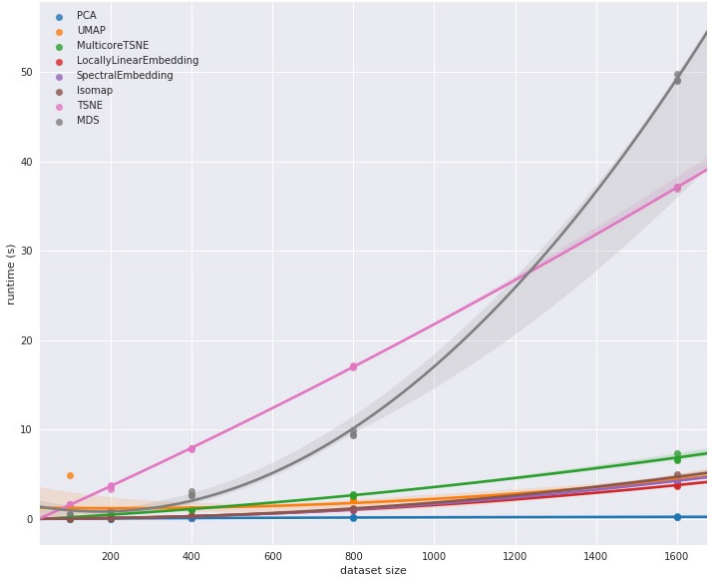
This way, by constructing the simplicial complexes beforehand allows UMAP to calculate the relative point distances in the lower dimension

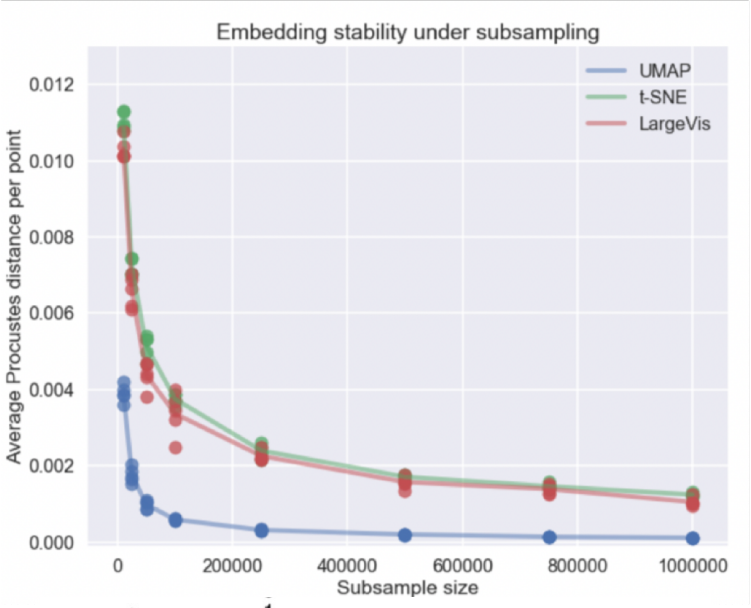
(instead of randomly assigning as in tSNE)

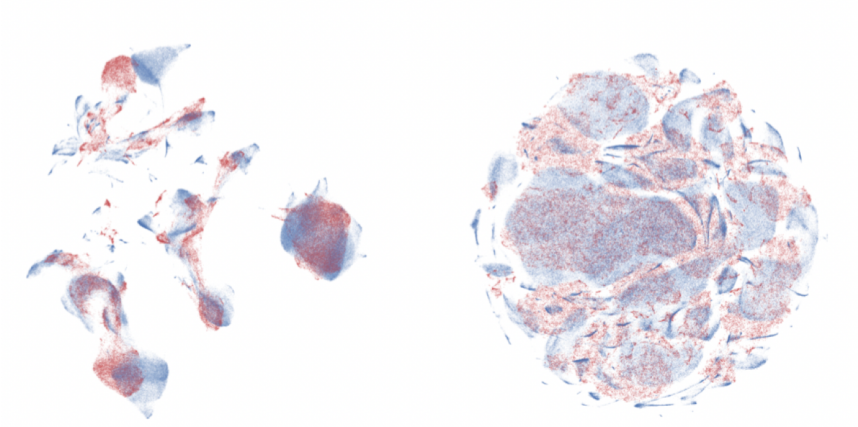
UMAP-总结

- It is a NON-LINEAR graph-based method of dimensionality reduction
- Very efficient - $O(n)$
- Can be run from the top PCs (e.g.: PC1 to PC10)
- Can use any distance metrics!
- Can integrate between different data types (text, numbers, classes)
- It is no longer completely stochastic as t-SNE
- Defines both LOCAL and GLOBAL distances
- Can be applied to new data points

各种降维方法性能比较

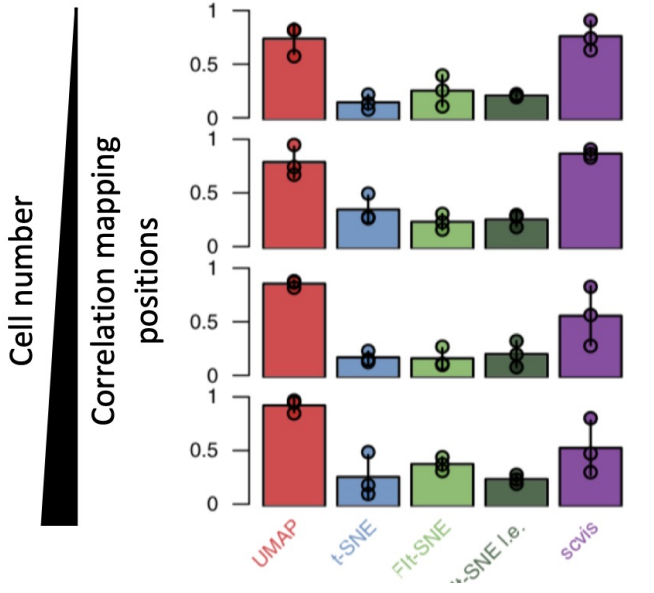






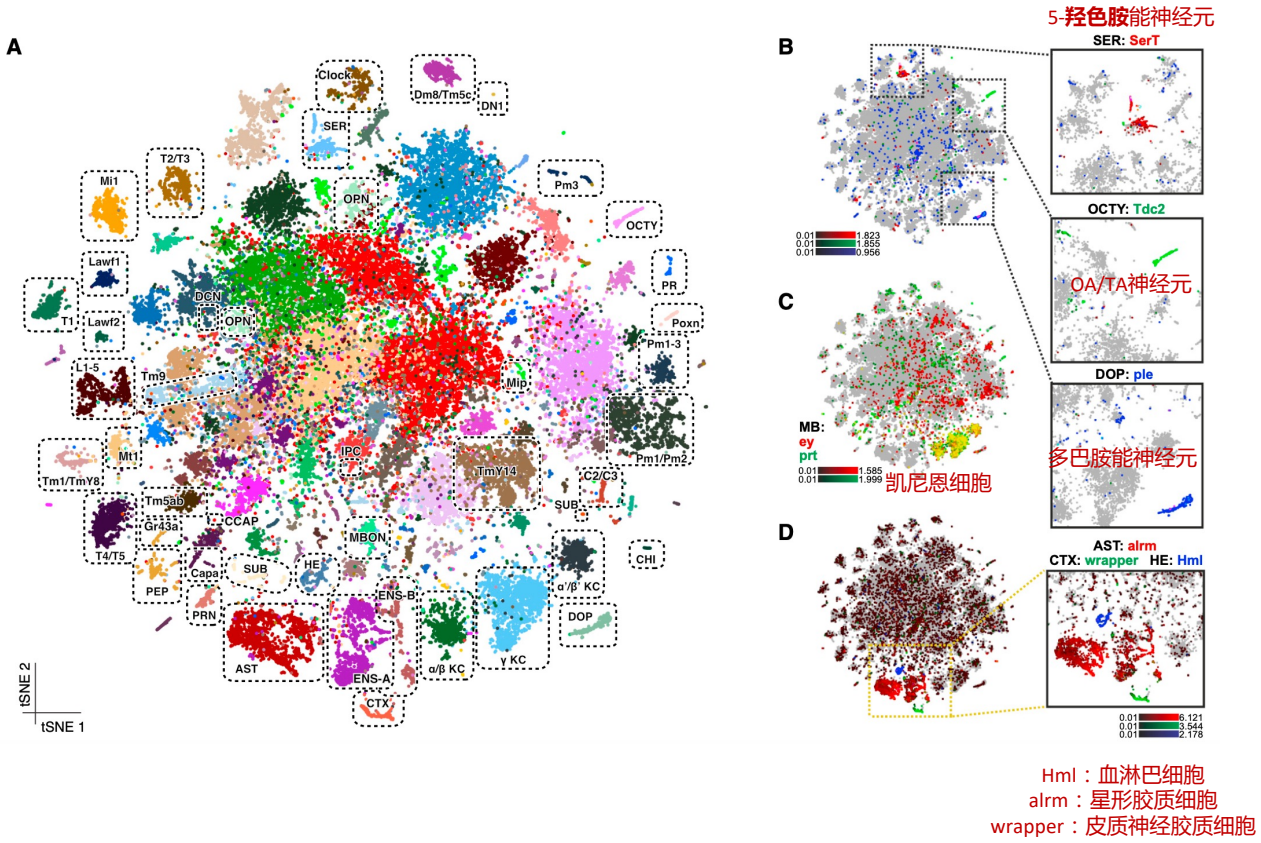
(a) UMAP

(b) t-SNE



案例1-果蝇大脑

- 果蝇衰老过程中的大脑细胞图谱



Hml：血淋巴细胞
alrm：星形胶质细胞
wrapper：皮质神经胶质细胞

20雌、20雄性果蝇，分布在8个时间点

1）过滤后得到57K高质量细胞（87个簇）

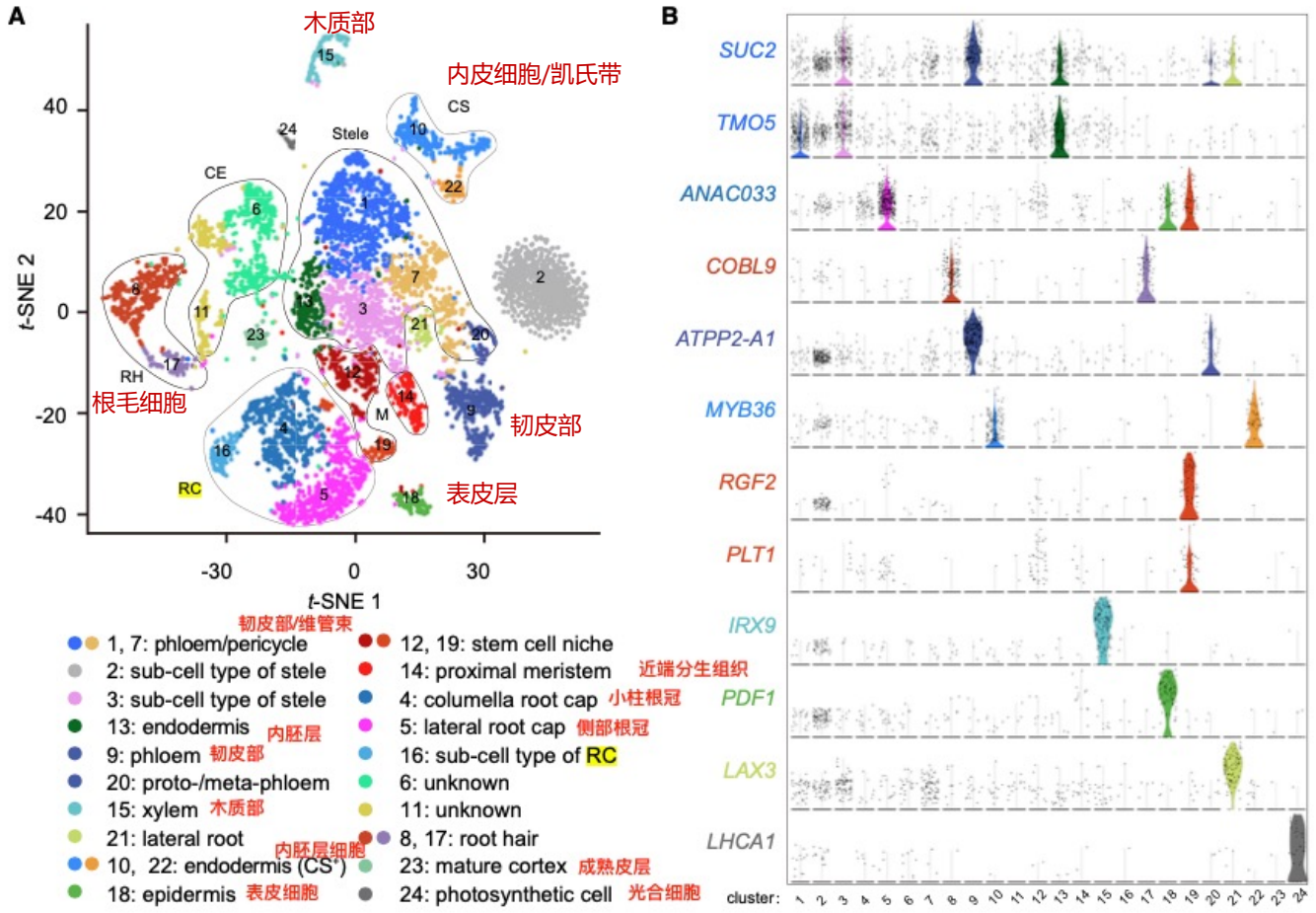
2）细胞身份：已发表marker基因（B）和果蝇大脑亚群转录组数据（嗅觉投射神经元等）

3）共鉴定到37个细胞身份类型，中间大部分细胞类型未能确定

doi: 10.1016/j.cell.2018.05.057

案例2-拟南芥根

- 拟南芥根部的细胞异质性

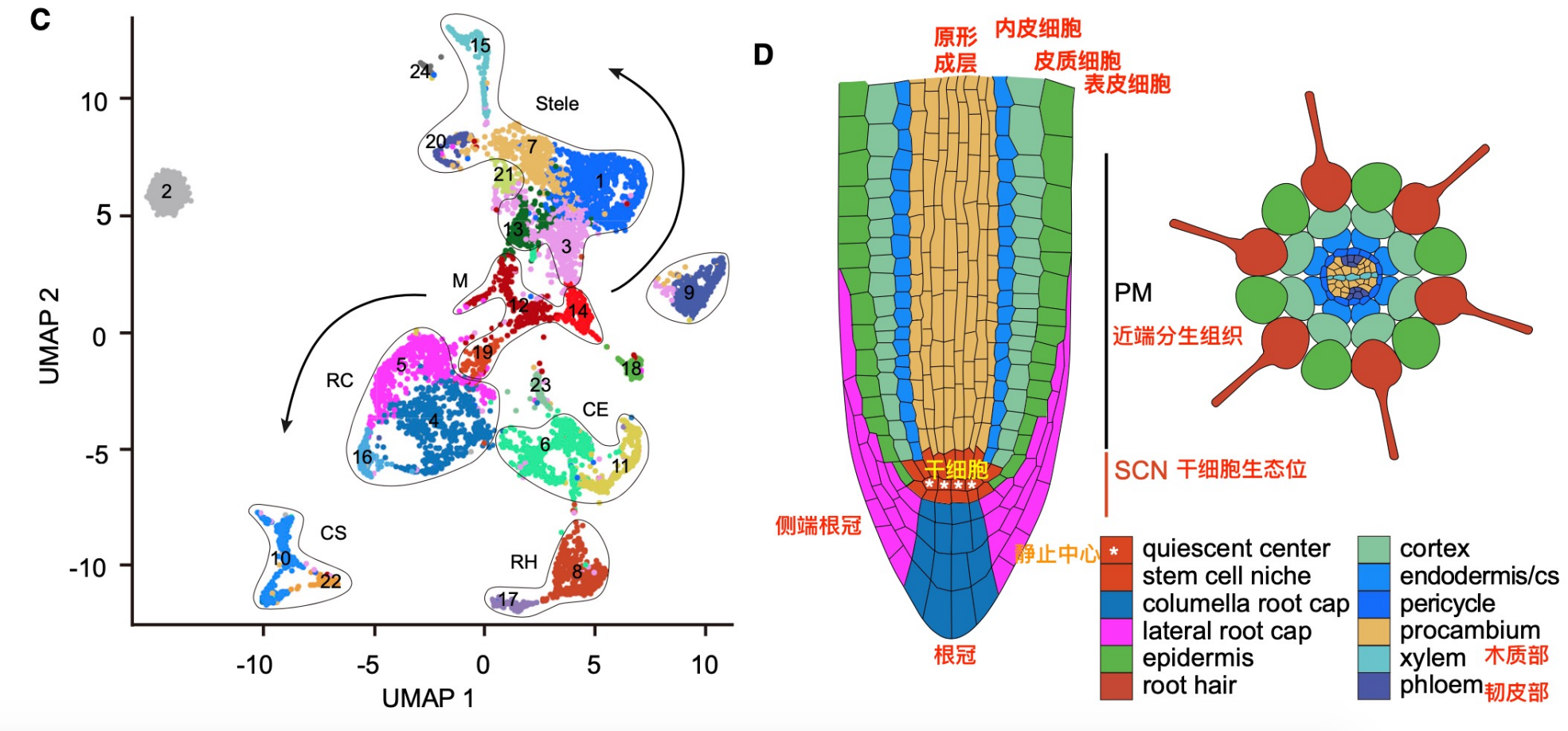


原生质体制备：种子消毒，MS培养基培养，10d生长收集根；Rnase-free的酶溶液（纤维素酶1.5%、离析酶1.5%、0.4M甘露醇、10mM KCl、10mM $CaCl_2$、0.1%BSA）中2h室温下消化，40um滤网过滤，8%甘露醇冲洗三次。台盼蓝染色计算活率（85%），调整浓度1500-2000cell/ul，10X建库。

1）上样15000个细胞，过滤后得到7695个23161个基因

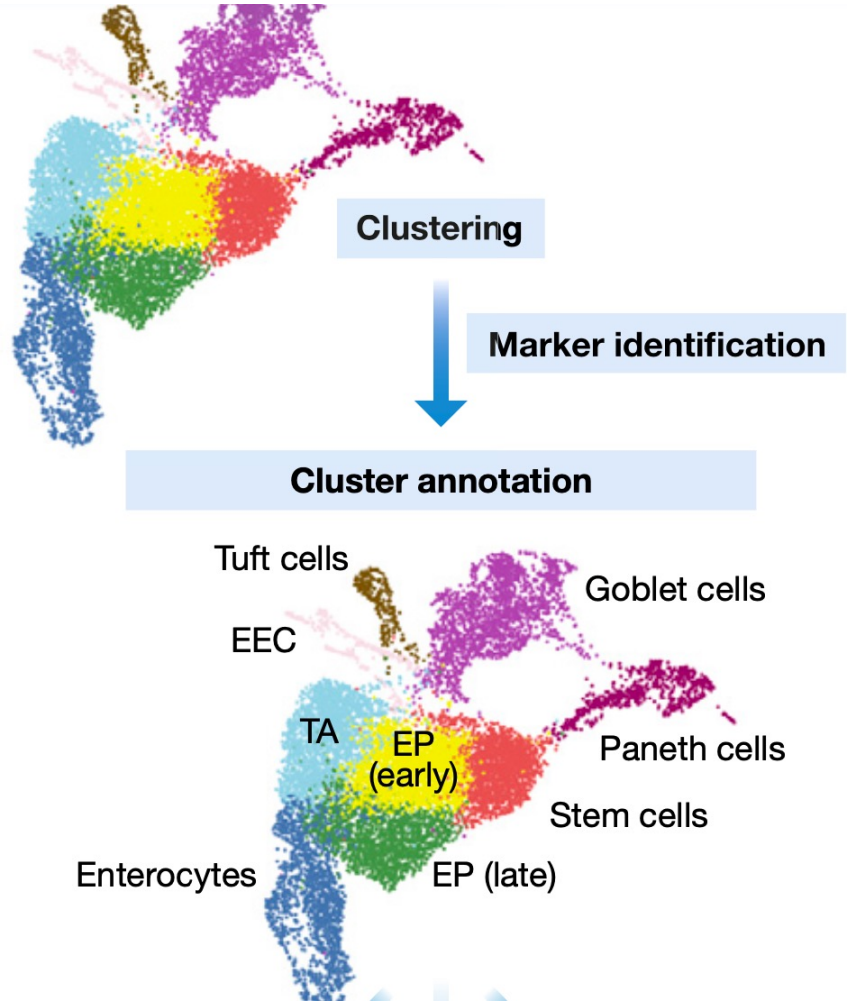
2）共得到24个cluster，通过已发表的103个marker基因去注释得到具体细胞类型

3）具有明显边界的细胞群；cluster内的异质性；新的细胞类群（cluster24，6和11无分类）



UMAP细胞图谱还可以看出细胞分化的方向

Marker鉴定

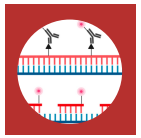


细胞聚类到细胞身份，还有很长一段距离…

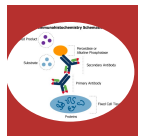
必须先有一个参考基因组及其注释信息

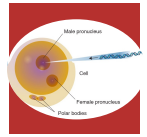
无Marker数据库	有Marker数据库
差异分析找Marker	差异分析找Marker
文献报道的Marker	与数据库比对
实验验证Marker	可不用实验验证

候选Marker基因的可视化？？？


1. RNA原位杂交


2. 免疫荧光


3. 免疫组化


4. 转基因技术

实战

- ✓ **Cellranger如何工作**
- ✓ **图谱如何绘制**
- ✓ **候选Marker如何寻找**

免疫荧光：https://www.youtube.com/watch?v=zO70nrWcAyk；

https://www.youtube.com/watch?v=8v77U1UoLnA

RNA原位杂交：https://www.youtube.com/watch?v=DEIL3KeXL9w&t=198s

BerryGenomics
贝瑞基因

Thank You!

官 方 网 站

官 方 微 信

www.berrygenomics.com