CNGBdb

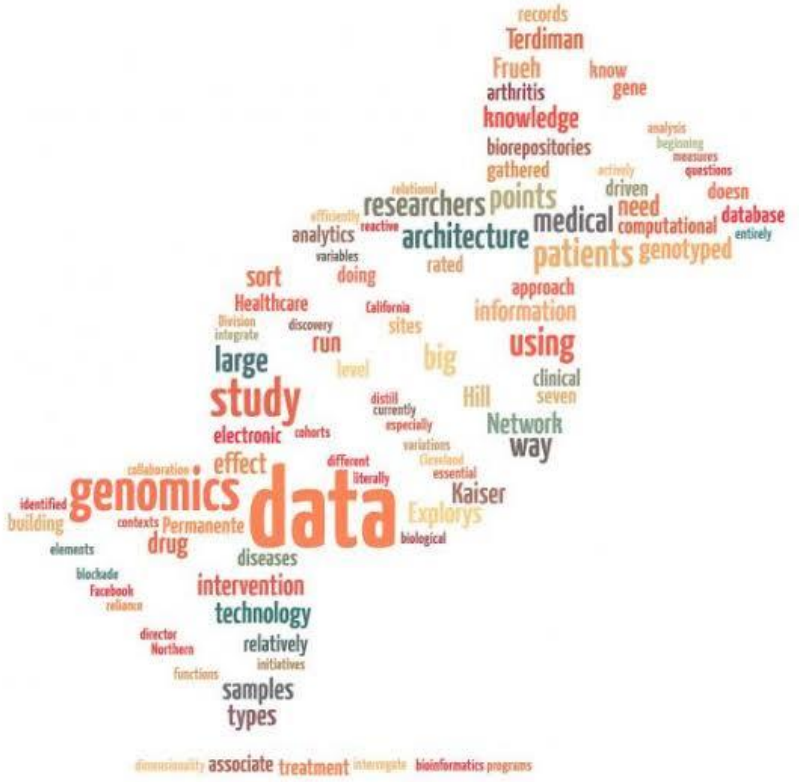
国家基因库生命大数据平台

百万单细胞基因表达数据算法大赛

CNGBdb
国家基因库生命大数据平台

赛题介绍

1. 数据背景
2. 单细胞聚类目标
3. 单细胞聚类常见方法
4. 难点与挑战
5. 评分方法



CNGBdb
国家基因库生命大数据平台

1.数据背景：什么是基因表达？

中心法则

源代码

```
#include "gzstream.h"
#include "lapack.h"
#include "lm.h"

using namespace std;

void LM::CopyFromParam(PARAM &cPar) { }

void LM::CopyToParam(PARAM &cPar) {
  cPar.time_opt = time_opt;
  cPar.ng_test = ng_test;
  return;
}

void LM::WriteFiles() {
  string file_str;
  file_str = path_out + "/" + file_out;
  file_str += ".assoc.txt";

  ofstream outfile(file_str.c_str(), ofstream::out);
  if (!outfile) {
    cout << "error writing file: " << file_str.c_str() << endl;
    return;
  }
```

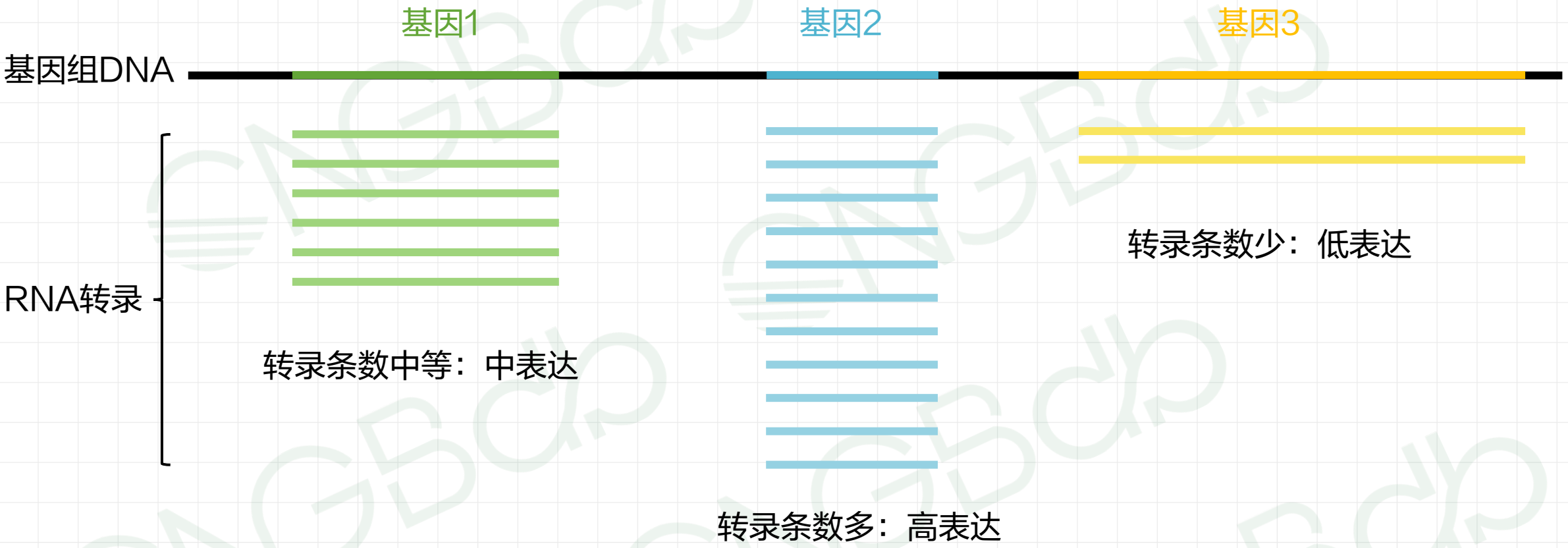
机器码

```
0000 1001 1100 0110 1010
1010 1111 0101 1000 0000
1100 0110 1010 1111 0101
0101 1000 0000 1001 1100
```

可执行程序

CNGBdb
国家基因库生命大数据平台

1.数据背景：什么是基因表达量？

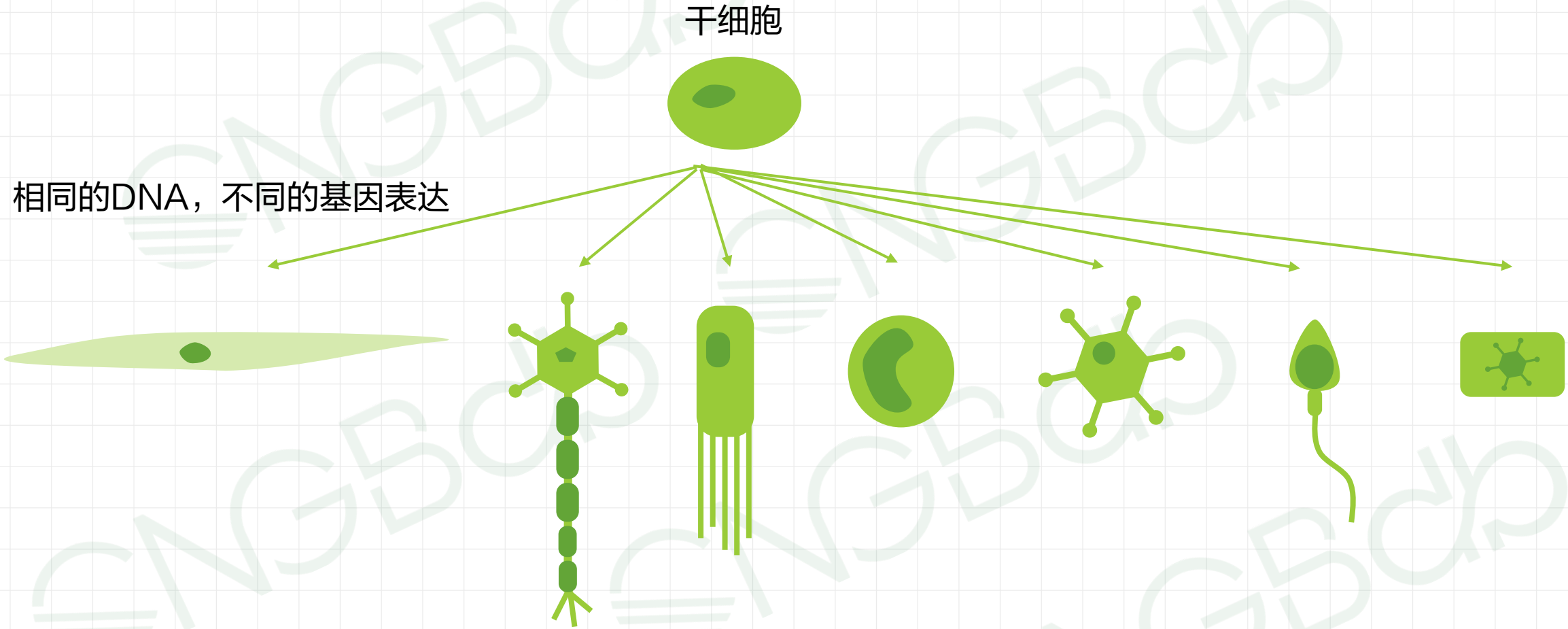


数据矩阵

样本ID	基因1	基因2	基因3
sample001	6	11	2
…	…	…	…

CNGBdb
国家基因库生命大数据平台

1.数据背景：基因表达的意义

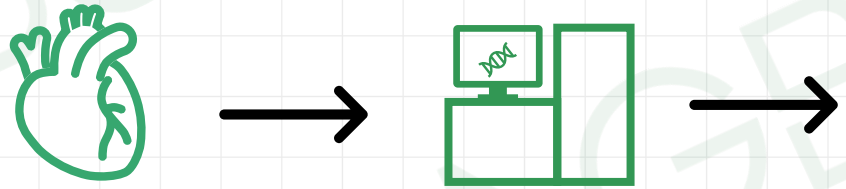


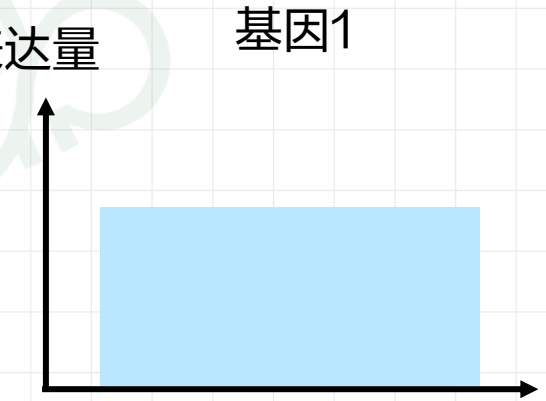
人类成人个体约有 $2x10^{13}$个细胞，主要细胞类型约有300种，每种主类型又可包含不同数量的亚类型，且仍有新的细胞类型尚待发现

CNGBdb
国家基因库生命大数据平台

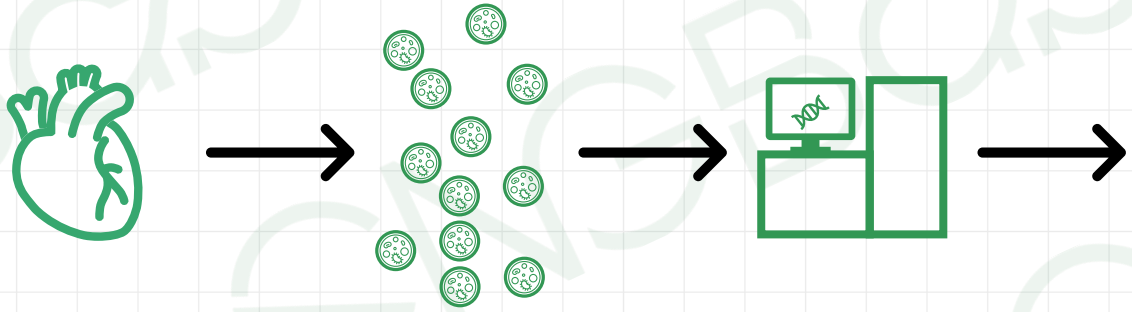
1.数据背景：单细胞基因表达

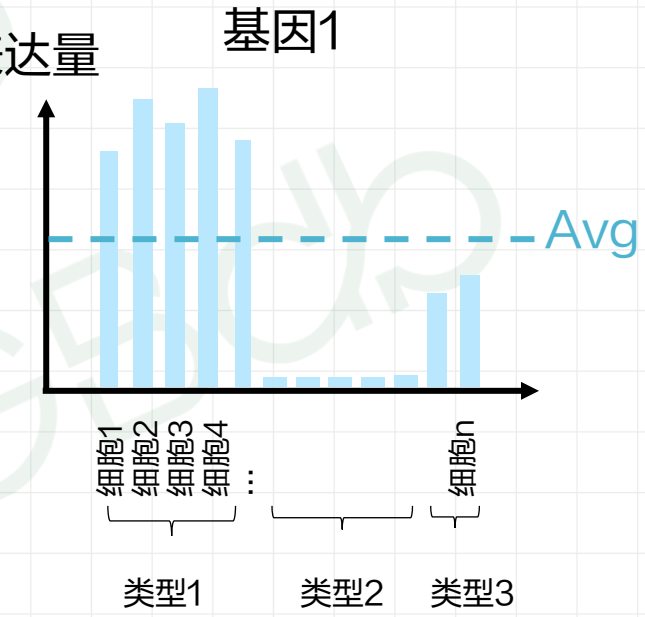
传统基因表达检测技术：
仅能获得大量细胞的平均表达值





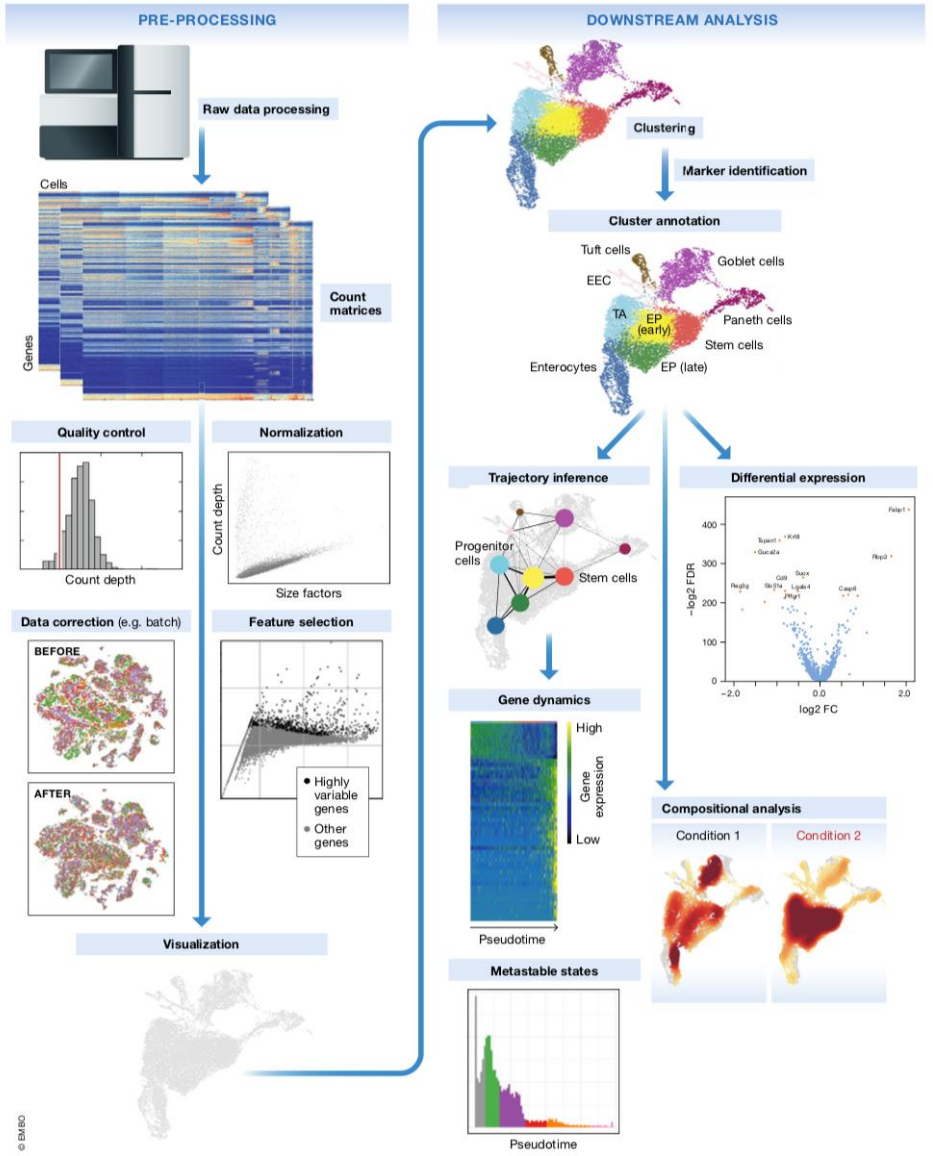
单细胞基因表达检测技术：
更高精度的基因表达



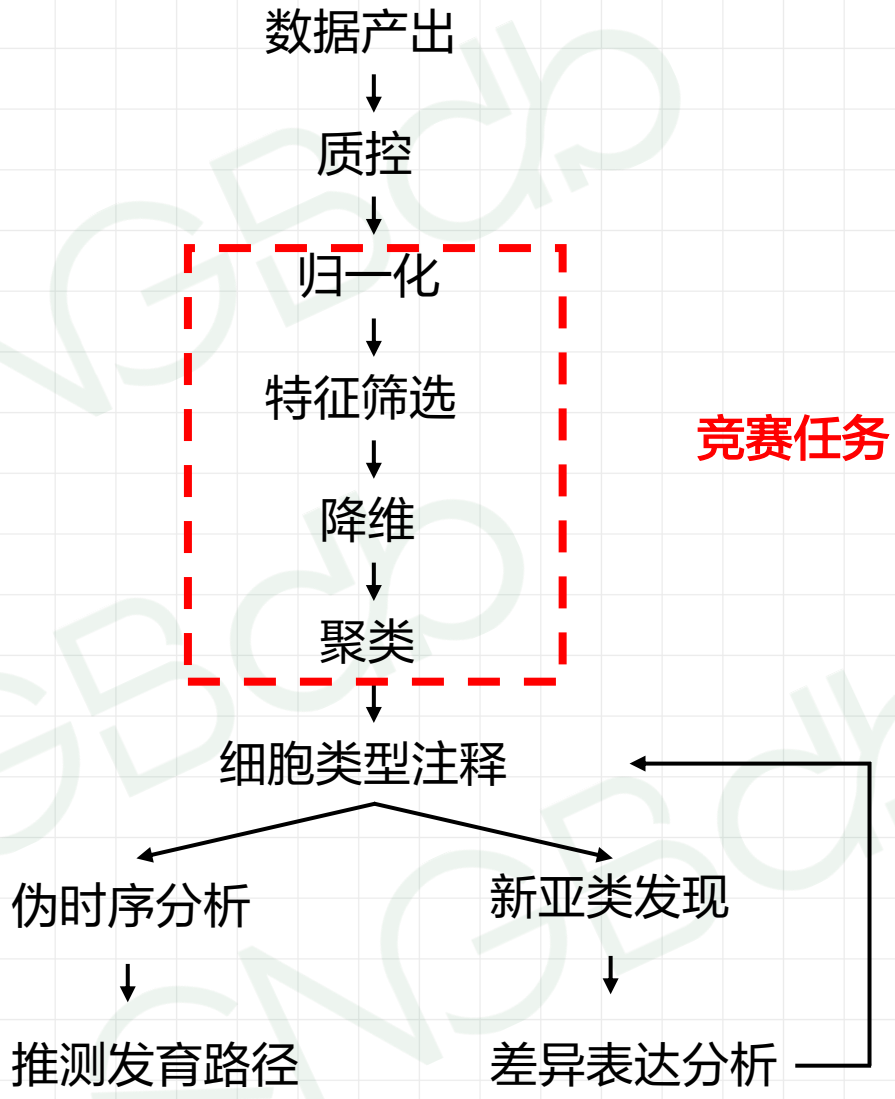


CNGBdb
国家基因库生命大数据平台

2.单细胞基因表达聚类：一般分析流程

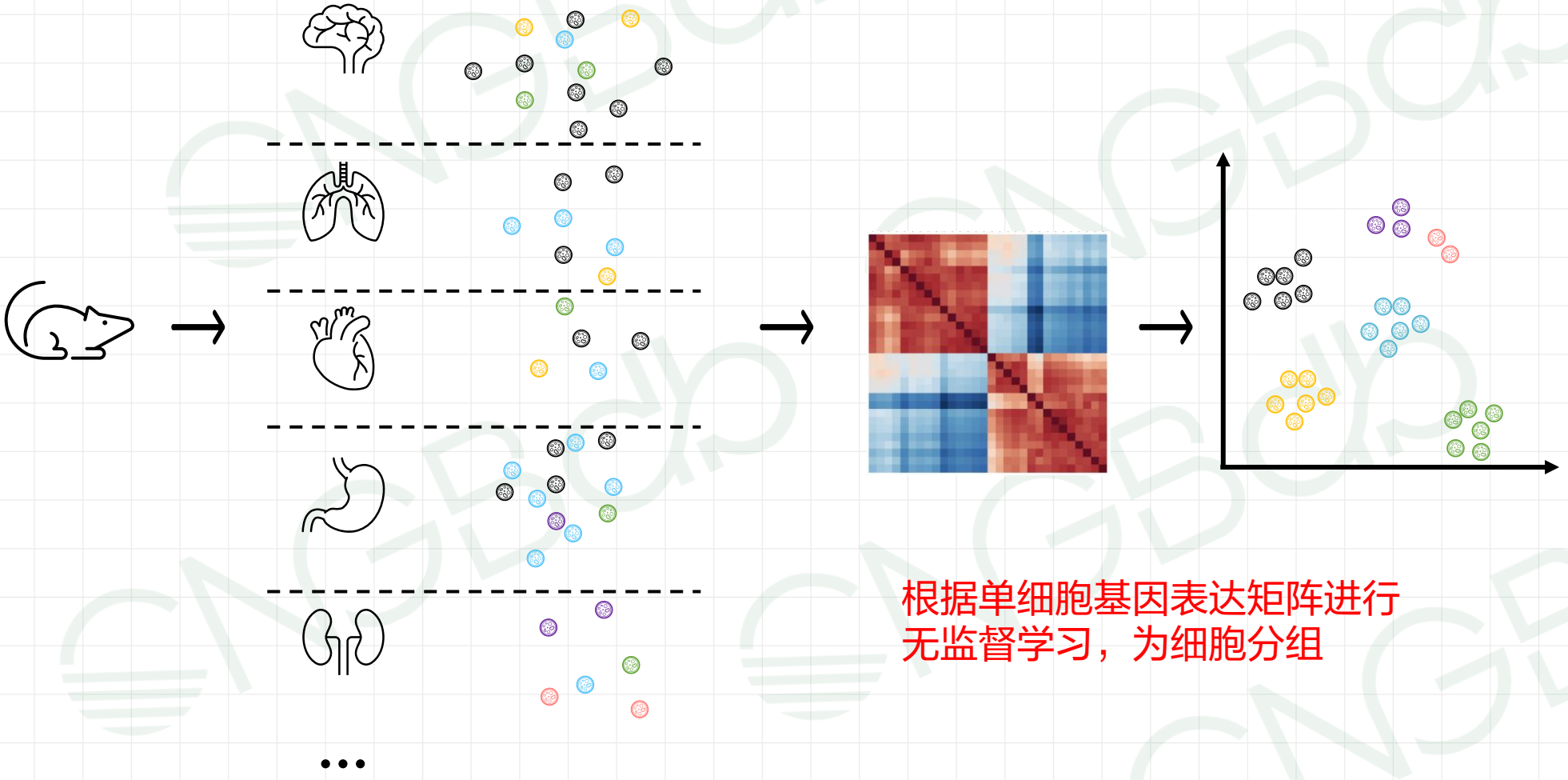


Luecken, M. D., & Theis, F. J. (2019). Current best practices in single - cell RNA - seq analysis: a tutorial. *Molecular systems*



竞赛任务

2.单细胞基因表达聚类目标



根据单细胞基因表达矩阵进行
无监督学习，为细胞分组

Monocyte

T cell

B cell

Smooth muscle cell

Endothelial cell

Epithelial cell

Dendritic cell

Stromal cell

...

CNGBdb
国家基因库生命大数据平台

2.单细胞基因表达聚类目标

比赛数据

不同细胞
初赛：约30万条
决赛：约70万条

样本ID	基因1	基因2	…	…	基因k	批次	物种	组织来源	取样时间	…
cell001						1	Mouse	Liver		
cell002						1	Mouse	Lung		
…						2	…			
…						…	…			
cellxxx						n	Mouse	Brain		

基因表达>20000列
~90% dropout

不同实验批次

物种
初赛：小鼠
决赛：人

初赛>50种

其它信息

预测聚类分组标签

样本ID	cluster
cell001	1
cell002	3
…	3
…	…
cellxx	i

与手工标注进行对比

样本ID	真实注释类型
cell001	A
cell002	A
…	B
…	…
cellxx	Z

3.单细胞聚类常见方法

- PCA+ hierarchical clustering
- Zero-imputed PCA + hierarchical clustering
- T-SNE + density-based clustering
- PCA + k-means
- K-medoids clustering
- SVM
- Nearest neighbor graph clustering
- Model based clustering

CNGBdb
国家基因库生命大数据平台

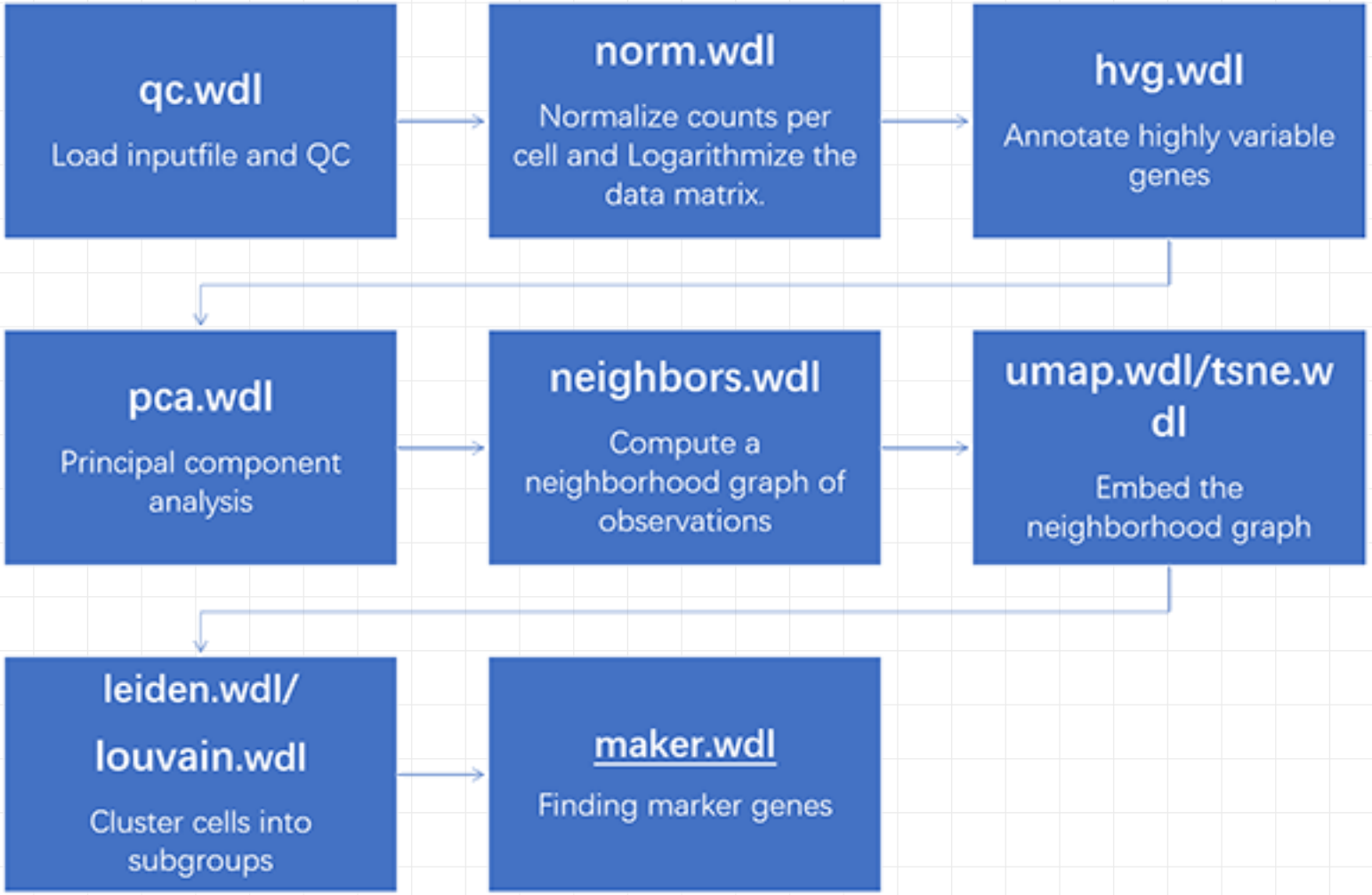
3.单细胞聚类常见方法

- 各类单细胞分析工具大集合
  - https://github.com/seandavi/awesome-single-cell

- 主流工具
  - Seurat
    - 一个基于R的单细胞数据质控和分析工具
      - Satija, R., Farrell, J. A., Gennert, D., Schier, A. F., & Regev, A. (2015). Spatial reconstruction of single-cell gene expression data. *Nature biotechnology*, *33*(5), 495-502.
  - ScanPy
    - 一个基于 Python 分析单细胞数据的主流软件包,内容包括预处理,降维,聚类等多个步骤
      - Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. Genome biology. 2018 Dec;19(1):15.

CNGBdb
国家基因库生命大数据平台

3.单细胞聚类常见方法例子：ScanPy



CNGBdb
国家基因库生命大数据平台

4.难点与挑战

- 标准化/归一化/特征筛选方法选择
- 批次效应
- 跨样本/组织/物种等的通用处理方法
- 表达量高缺失率，稀疏矩阵
  - 90%基因表达为0: 细胞无表达 vs 有表达但未检测出?
  - 常见处理手段
    - 特征筛选
    - 数据插补
- 细胞类型丰度差异大：如何鉴定<1%占比的细胞类型?
- 是否引入额外信息?
  - 已知的细胞类型特征基因：如http://bio-bigdata.hrbmu.edu.cn/CellMarker/
  - 基因互作关系
  - 参考细胞类型表达谱
  - …
- 大规模数据量（决赛挑战：高效处理百万级单细胞表达数据）

CNGBdb
国家基因库生命大数据平台

5.评分方法

- 初赛：Adjusted Rand Index
- 衡量聚类分组标签与人工标注类别的数据分布吻合程度
- 根据ARI排名选择进入决赛的团队

Given a set S of n elements, and two groupings or partitions (*e.g.* clusterings) of these elements, namely $X = \{X_1, X_2, \ldots, X_r\}$ and $Y = \{Y_1, Y_2, \ldots, Y_s\}$, the overlap between X and Y can be summarized in a contingency table $[n_{ij}]$ where each entry n_{ij} denotes the number of objects in common between X_i and Y_j : $n_{ij} = |X_i \cap Y_j|$.

$X \backslash Y$	Y_1	Y_2	$\cdots$	Y_s	sums
X_1	n_{11}	n_{12}	$\cdots$	n_{1s}	a_1
X_2	n_{21}	n_{22}	$\cdots$	n_{2s}	a_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
X_r	n_{r1}	n_{r2}	$\cdots$	n_{rs}	a_r
sums	b_1	b_2	$\cdots$	b_s	

The original Adjusted Rand Index using the Permutation Model is

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right] \Big/ \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right] \Big/ \binom{n}{2}}$$

where n_{ij}, a_i, b_j are values from the contingency table.

- 决赛：在指定的环境中完成大数据量聚类分析，主要考察性能指标

CNGBdb
国家基因库生命大数据平台

THANKS

共有、共为、共享
Owned by All, Completed by All and Shared by All