PCA之前可以挑选高变异的gene.
挑选方差，CV高的基因.
500, 1000, 3000个基因.

PCA的关键步骤及理解
① 获得数据，矩阵A进行归一化
② AA^T为A矩阵的协方差矩阵，记作B.
③ 对A正交旋转不影响A的协方差总和
④ B正交对角化 $B=QDQ^{-1}=QDQ^T$ $Q^TQ=QQ^T=I$

聚类
层次聚类：
hierarchical clustering method.
聚集法：按距离，每次设置cutoff，最后聚成一类.
分割法：由少到多

① cutoff设置？
② 计算距离？
③ 选择参考点

明考夫斯基距离

$$d(x,y)=\left[\sum_{i=1}^{p}|x_i-y_i|^p\right]^{1/p}$$

当$q=1$时，为绝对值距离（曼哈顿距离，街区距离）

$$d(x,y)=\sum_{i=1}^{p}|x_i-y_i|$$

当$q=2$时，为欧氏距离

当$q=\infty$时，为切比雪夫距离

$$d(x,y)=\max_{1\le i\le p}|x_i-y_i|$$

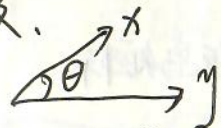
兰氏距离（Lance and williams distance）

所有数据为正，$d(x,y)=\sum_{i=1}^{p}\frac{|x_i-y_i|}{x_i+y_i}$

该距离与各变量的单位无关，且适用于高度偏斜或含异常值的数据.

定义 x, y 之间的距离满足：
① 非负性：$d(x,y)\ge 0$
② 对称性：$d(x,y)=d(y,x)$
③ 三角不等式：$d(x,y)\le d(x,z)+d(z,y)$

相似系数.

夹角余弦 

$\cos\theta=\frac{x^Ty}{\|x\|\|y\|}$

Pearson相关系数 $\rho=\rho(x,y)=\frac{Cov(x,y)}{\sqrt{V(x)\cdot V(y)}}$