



# Clustering multilayer omics data using MuCut

主讲：张媛媛

2018.03.29

Hidalgo, Sebastian J. Teran, and Shuangge Ma. "Clustering multilayer omics data using MuNCut." *BMC Genomics* 19.2 (2018): 198.

2016-2017最新影响因子	3.729
2016-2017自引率	5.70%

中科院SCI期刊分区 (2017最新版本)	大类学科	小类学科	Top期刊	综述期刊
	生物	BIOTECHNOLOGY & APPLIED MICROBIOLOGY 生物工程与应用微生物 GENETICS & HEREDITY 遗传学	否	否

# 主要内容（摘要）

- 研究背景：多组学数据的产生使得人们可以整合多组学数据进行聚类研究，如发现组学单元（omics units）的未知功能、降维等。现有的聚类方法不适用于多组学数据（原因：不同组学数据间的相互作用）、
- 方法：基于Ncut技术（谱聚类图分割算法），提出MuNCut聚类方法，同时考虑层内和层间的连接。
- 结果：a、仿真数据上与其他方法比较的结果  
b、TCGA数据库的BRCA和CESC数据的结果分析。

# Introduction

- 本文的目标：设计一类聚类方法使之适用于多组学数据。
- 已存在的多组学聚类方法：
  - 1、目标多是聚类patients，如iCluster和SNF（相似网络融合）。
  - 2、基于节点中心性或社团识别的多组学数据分析。

# 谱聚类图分割算法

- 谱聚类算法是将样本点看成为顶点，将顶点之间用带权的边连接起来，带权的边可以看成是顶点之间的相似度。聚类从而可以看成如何分割这些带权的边，继而将聚类问题转化为怎么进行图分割的问题。

## 常用的分割方法有：

### 最小分割法 (Minimum Cut)

假如将一个图G 划分为A, B 两个子图, 那么目标函数可以理解为

$$\text{Cut}(A, B) = \sum_{u \in A, v \in B} w(u, v)$$

### 规范化分割 (Normalized Cut)

$$\text{Normalized Cut}(A, B) = \frac{\text{Cut}(A, B)}{\text{sum}(A, V)} + \frac{\text{Cut}(A, B)}{\text{sum}(B, V)}$$

其中Cut(A,B)表示 A, B 两个子图的相似程度, sumA 表示 A 图中所有点的权值之和

### 最小最大分割准则 (Min-max Cut)

最小最大分割准则要求最小化Cut (A,B) 的同时, 最大化sum(A,A)与sum (B,B) 。

$$\text{Mcut} = \frac{\text{Cut}(A, B)}{\text{sum}(A, A)} + \frac{\text{Cut}(A, B)}{\text{sum}(B, B)}$$

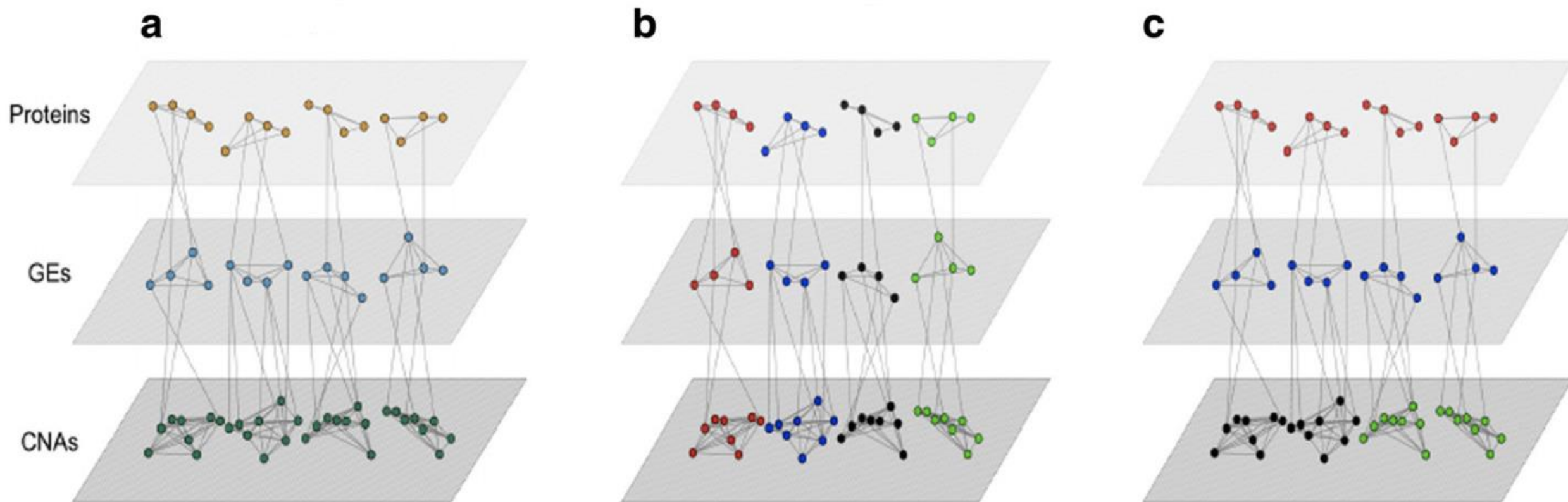
### 比例分割法 (Ratio Cut)

$$\text{Rcut} = \frac{\text{cut}(A, B)}{\min(|A|, |B|)}$$

其中|A|, |B|分别表示子图 A, B中顶点的个数, 目标 Rcut 函数只简单考虑到如何使 A, B两个子图间的相似性最小, 这样可以减少分割的次数。

# Methods—Ncut (Normalized cut)

- 本文研究三层网络：CNV，gene，protein



# Methods—Ncut (Normalized cut)

Denote  $Z = (Z_1, \dots, Z_q)$ ,  $Y = (Y_1, \dots, Y_p)$ , and  $X = (X_1, \dots, X_r)$  as the length  $q$ ,  $p$ , and  $r$  vectors of proteins, GEs, and CNVs, respectively.

- 以CNV为例,  $\bar{W}_C = (w_{jl,c})_{r \times r}$  表示加权邻接矩阵, 其中  $w_{jl}$  表示两个CNV的相似性 (高斯核)。

$$w_{jl} = \exp \left( \frac{(X_j - X_l)^2}{2\sigma^2} \right)$$



# Methods—Ncut (Normalized cut)

- 假设将所有的CNV划分为K个不相交的类，记为

$A_{1,C}, A_{2,C}, \dots, A_{K,C}$ ; 使用 $A_{k,C}^c$ 第k个类的补集。

则，

$$\text{NCut}_C = \sum_{k=1}^K \frac{\text{cut}(A_{k,C}, A_{k,C}^c; \mathbf{W}_C)}{\text{cutvol}(A_{k,C}; \mathbf{W}_C)},$$

where

$$\text{cut}(A_{k,C}, A_{k,C}^c; \mathbf{W}_C) = \sum_{j \in A_{k,C}, l \in A_{k,C}^c} w_{jl,c},$$

$$\text{NCut}_{\text{single}} = \text{NCut}_C + \text{NCut}_G + \text{NCut}_P.$$

and

$$\text{cutvol}(A_{k,C}; \mathbf{W}_C) = \sum_{j,l \in A_{k,C}} w_{jl,c}.$$

# Methods—MuNcut

- 首先采用回归方法描述组学间的调控:

$$Y = X\beta_1 + \epsilon_1, \quad Z = Y\beta_2 + \epsilon_2, \quad (5)$$

where  $\beta_1$  and  $\beta_2$  are the  $r \times p$  and  $p \times q$  matrices of unknown regression coefficients, and  $\epsilon_1$  and  $\epsilon_2$  are

- 使用罚函数的方法估计回归系数

$$\hat{\beta}_1 = \underset{R}{\operatorname{argmin}} \left\{ \|Y - X\beta_1\|_2^2 + \lambda \left( (1-\alpha)\|\beta_1\|_2^2 + \alpha\|\beta_1\|_1 \right) \right\}.$$

# Methods—MuNcut

- 进行图分割的**关键之一是定义相似矩阵**，因此，定义组间的相似矩阵：

其中  $W_{\hat{Z}:\hat{Y}}$  定义为回归系数矩阵  $\hat{\beta}_2$ .

最终网络有  $r+p+q$  个异质节点

define  $\hat{Y} = X\hat{\beta}_1$  and  $\hat{Z} = Y\hat{\beta}_2$ .

For  $(X, \hat{Y}, \hat{Z})$ , the length  $r+p+q$  “mega” vector of omics measurements, define the  $(r+p+q) \times (r+p+q)$  weight matrix

$$\tilde{W} = \begin{pmatrix} 0 & W_{\hat{Z}:\hat{Y}} & 0 \\ W_{\hat{Z}:\hat{Y}}^T & 0 & W_{\hat{Y}:X} \\ 0 & W_{\hat{Y}:X}^T & 0 \end{pmatrix}, \quad (7)$$

# Methods—MuNcut

## ■ 定义分割测度

$$\text{MuNcut}(A) = \text{NCut}_{\text{multi}} + \gamma \times \text{NCut}_{\text{single}},$$

## ■ 使用模拟退火算法对上述目标函数进行优化求最小值。

温度函数:  $\tilde{T}(t) = L \log(t + 1)$

**Step 1** Randomly initialize  $A^{(0)} = \{A_1^{(0)}, \dots, A_K^{(0)}\}$ . In our numerical study, different initial values lead to similar results.

**Step 2** Set  $t = t + 1$ . For  $k = 1, \dots, K$ , compute  $p_k$  as the number of  $(j, l)$  pairs such that  $j, l \in A_k^{(t-1)}$ . Draw  $k(-)$  and  $k(+)$  from  $\{1, \dots, K\}$  with probabilities proportional and inversely proportional to  $p_k$ .

**Step 3** Draw  $i$  randomly from  $A_{k(-)}^{(t)}$ . Set  $A_{k(+)}^{(t)} = A_{k(+)}^{(t-1)} \cup \{i\}$  and  $A_{k(-)}^{(t)} = A_{k(-)}^{(t-1)} \setminus \{i\}$ . For  $j \neq k(+), k(-)$ ,  $A_j^{(t)} := A_j^{(t-1)}$ .

**Step 4** If  $\text{MuNcut}(t) \leq \text{MuNcut}(t - 1)$ , then keep  $A^{(t)}$  as it is. If not, keep  $A^{(t)}$  as it is with probability  $\exp\left(-\frac{\text{MuNcut}(t) - \text{MuNcut}(t-1)}{\tilde{T}(t)}\right)$ , and otherwise  $A^{(t)} = A^{(t-1)}$ .

**Step 5** Repeat Steps 2-4 until  $t = B$ .

- 模型中的参数均使用交叉验证（cross-validation）进行设置。
- R包NCutYX可用

follows: `clust ← muncut(Z, Y, X, K = 2, B = 3000, L = 1000, gamma = 0.5, dist = "gaussian", sigma = 1)` In

# Results ——真实数据的结果分析

- 真实数据来源：TCGA中的BRCA和CESC

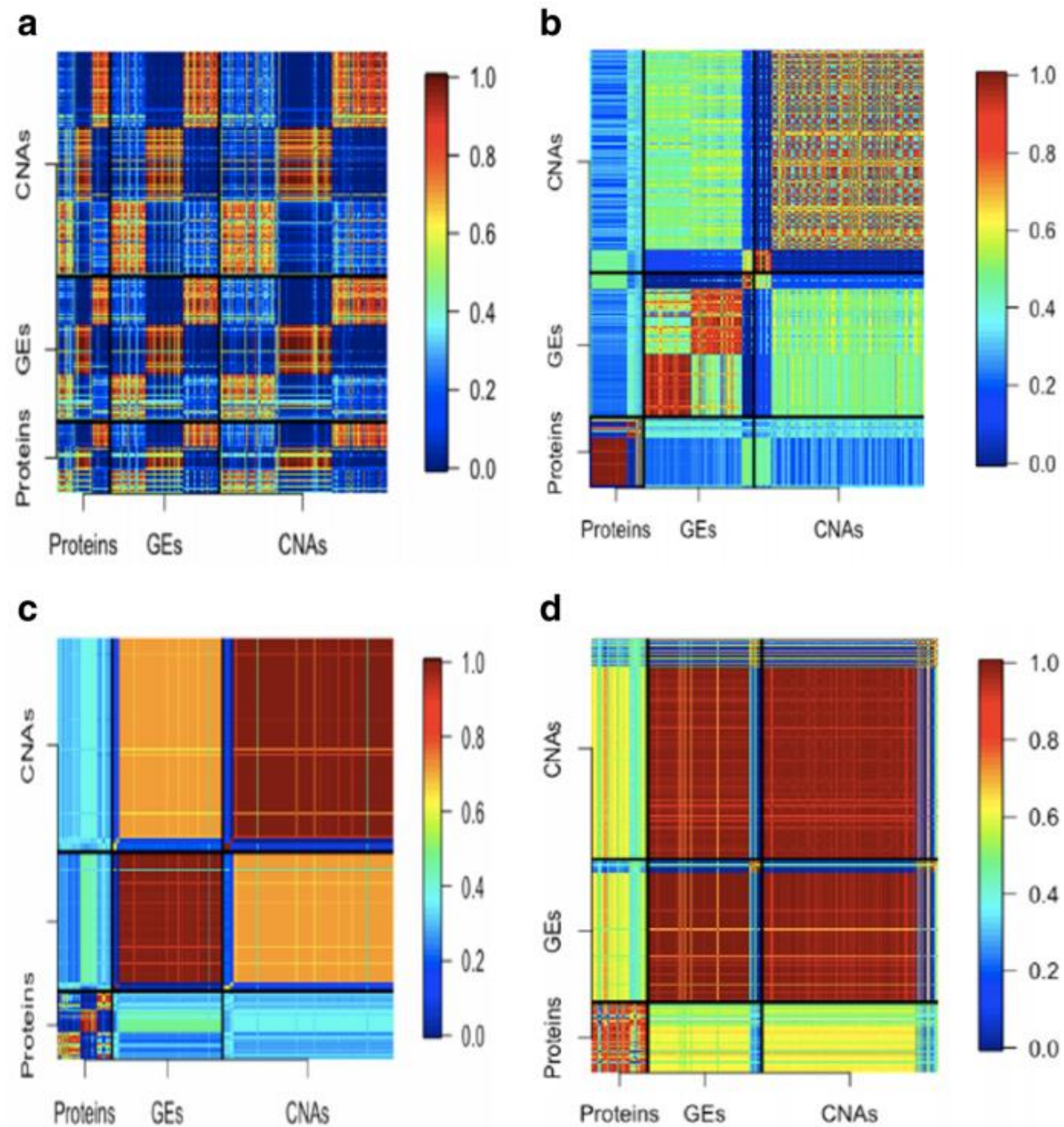
- 评价测度：

- 1、稳定性（stability）：不重复的选择n/2个样本，计算其聚类的邻接矩阵，重复N次，定义稳定性如下：

$$M_{stability} = N^{-1} \sum_{1 \leq k \leq N} \hat{A}^{(k)}.$$

元素（i,j）表示两个变量聚成一类的概率

理想的方法应该有如下特点：某些元素有很大的值，其他的值较小。



**Fig. 2** Analysis of BRCA data: stability of heatmaps. **a** MuNCut; **b** KM\*; **c** SC\*; **d** HC\*. The  $(i, j)$ th entry is the probability that the  $i$ th and  $j$  elements belong to the same cluster. Higher/lower probabilities are presented using warmer/colder colors



# Results——真实数据的结果分析

## ■ 评价测度

一致性（concordance）：比较不同方法下邻接矩阵的相似性，值越大表示两种方法越相似。

Define the concordance of method  $B$  with respect to  $A$  as

$$M(B|A) = \sum_{j,l}^m (\hat{A} \odot \hat{B})_{jl} / \sum_{j,l}^m (\hat{A})_{jl}, \quad (10)$$

该测度不具备对称性



**Table 1** Data analysis: concordance between the analysis results using different methods. In each cell,  $M(B|A)$ , where B and A are the clustering methods in the column and row, respectively

BRCA	MuNCCut	KM*	SC*	HC*
MuNCCut	100%	59.4%	72.7%	80.1%
KM*	44.7%	100%	74%	82.3%
SC*	34.5%	46.7%	100%	90.1%
HC*	36.3%	49.6%	85.9%	100%
CESC	MuNCCut	KM*	SC*	HC*
MuNCCut	100%	48.3%	44.6%	52.5%
KM*	37.8%	100%	51.3%	64.7%
SC*	38.7%	56.9%	100%	61.5%
HC*	35.6%	56.2%	48.1%	100%

## 仿真数据的结果分析

- 与该方法进行比较的方法：K-means (KM) , spectral clustering (SC) 和Hierarchical clustering (HC) ; 两个社团检测方法：Louvain (LC) 和fast Greedy Clustering (FGC)
- 定义测度描述聚类的精度：

$$M_{accuracy} = 1 - \sum_{j,l}^m (A_T \odot \hat{A})_{jl} / m^2,$$

$m=q+p+r$ ,  $A$ 表示邻接矩阵, 矩阵中的元素为1时, 表明两个节点属于同一类, 否则为0;  $A_T$ 表示真实的结果。该测度的值越小, 则聚类越精确。

# 仿真Scenario I

- $p=q=r$ , 共划分为4类, 类中节点数分别为 $p/5$ ,  $2p/5$ ,  $p/5$ ,  $p/5$ 。
- CNV数据 $X$ 由均值为0, 方差为1的**多变量正态分布产生**, 协方差 $\Sigma$ 是块对角结构, 前三个块对应前三个类, 而且远离对角线的元素设置为 $\rho$  ( $\rho=0.2$ 或 $0.4$ , 分别表示弱和中等相关); 第四个块对应第四个类。
- 基于上述的回归模型, 回归系数 $\beta_1$ 和 $\beta_2$ 也表示为块对角结构, 因此**GE是仅依赖于CNV, Protein仅依赖于GEs**。
- 在前三个类中, **随机选择20%的元素置为非零**, 满足Unif ( $h/2$ ,  $h$ )。  $h$ 取两个不同的值表示弱和中等调控; 第四个类, 对应的回归系数中为全零元素。

**Table 2** Simulation results for Scenario I

Parameters				<i>M</i> <sub>accuracy</sub>								
<i>n</i>	<i>q</i>	<i>h</i>	$\rho$	MuNCu	KM	SC	HC	KM*	SC*	HC*	LC	FGC
200	400	0.15	0.20	0.023	0.411	0.47	0.565	0.13	0.126	0.159	0.155	0.157
200	400	0.15	0.40	0.016	0.364	0.468	0.585	0.134	0.115	0.17	0.160	0.159
200	400	0.25	0.20	0.054	0.368	0.474	0.587	0.131	0.123	0.157	0.155	0.152
200	400	0.25	0.40	0.068	0.363	0.477	0.586	0.133	0.117	0.193	0.151	0.149
400	400	0.15	0.20	0.022	0.373	0.460	0.588	0.129	0.124	0.165	0.160	0.159
400	400	0.15	0.40	0.014	0.364	0.468	0.585	0.129	0.123	0.174	0.160	0.160
400	400	0.25	0.20	0.048	0.367	0.462	0.586	0.127	0.115	0.175	0.153	0.151
400	400	0.25	0.40	0.063	0.361	0.464	0.584	0.12	0.11	0.176	0.148	0.147
200	800	0.15	0.20	0.095	0.322	0.44	0.576	0.122	0.124	0.152	0.150	0.149
200	800	0.15	0.40	0.103	0.319	0.432	0.575	0.127	0.129	0.173	0.146	0.145
200	800	0.25	0.20	0.111	0.33	0.366	0.582	0.126	0.123	0.153	0.141	0.162
200	800	0.25	0.40	0.128	0.315	0.433	0.577	0.129	0.134	0.17	0.134	0.138
400	800	0.15	0.20	0.092	0.318	0.423	0.577	0.134	0.114	0.168	0.148	0.148
400	800	0.15	0.40	0.102	0.324	0.428	0.579	0.111	0.107	0.149	0.143	0.143
400	800	0.25	0.20	0.109	0.319	0.431	0.579	0.119	0.115	0.162	0.138	0.154
400	800	0.25	0.40	0.123	0.324	0.427	0.58	0.135	0.139	0.174	0.188	0.133
200	1200	0.15	0.20	0.11	0.312	0.384	0.578	0.139	0.139	0.157	0.145	0.156
200	1200	0.15	0.40	0.104	0.304	0.395	0.577	0.135	0.14	0.17	0.138	0.144
200	1200	0.25	0.20	0.124	0.308	0.4	0.576	0.132	0.131	0.153	0.212	0.162
200	1200	0.25	0.40	0.131	0.309	0.395	0.582	0.133	0.136	0.168	0.207	0.212
400	1200	0.15	0.20	0.112	0.316	0.388	0.58	0.122	0.124	0.154	0.141	0.154
400	1200	0.15	0.40	0.122	0.314	0.396	0.58	0.123	0.123	0.161	0.160	0.126
400	1200	0.25	0.20	0.127	0.315	0.403	0.573	0.13	0.132	0.162	0.186	0.197
400	1200	0.25	0.40	0.127	0.309	0.40	0.579	0.135	0.137	0.173	0.157	0.231

*n* is the sample size; *q* is the number omics measurements in each layer;

*h* measures the strength of regulation across layers;  $\rho$  is the correlation coefficient among CNVs

# 仿真Scenario II

- 上述仿真中，不在同一类中的节点是相互独立的
- 增强类内的相关性，在 $\Sigma$ 的设置中，将前三个类中的远离对角线的元素设为 $2\rho$ ，其他的设为 $\rho$ ；第四个类中节点以及不同类中的节点仍然相关。

**Table 3** Simulation results for Scenario II

Parameters				<i>M</i> <sub>accuracy</sub>								
<i>n</i>	<i>q</i>	<i>h</i>	$\rho$	MuNCu	KM	SC	HC	KM*	SC*	HC*	LC	FGC
200	400	0.15	0.20	0.026	0.365	0.462	0.582	0.13	0.122	0.188	0.139	0.155
200	400	0.15	0.40	0.025	0.411	0.476	0.564	0.133	0.119	0.202	0.158	0.157
200	400	0.25	0.20	0.095	0.409	0.475	0.566	0.131	0.122	0.19	0.163	0.163
200	400	0.25	0.40	0.118	0.409	0.473	0.563	0.124	0.12	0.202	0.157	0.155
400	400	0.15	0.20	0.024	0.412	0.469	0.564	0.13	0.125	0.204	0.155	0.152
400	400	0.15	0.40	0.024	0.413	0.475	0.567	0.129	0.123	0.197	0.155	0.153
400	400	0.25	0.20	0.096	0.413	0.469	0.564	0.128	0.113	0.20	0.162	0.159
400	400	0.25	0.40	0.111	0.411	0.479	0.565	0.125	0.134	0.203	0.153	0.151
200	800	0.15	0.20	0.113	0.399	0.436	0.561	0.129	0.118	0.179	0.152	0.174
200	800	0.15	0.40	0.132	0.397	0.443	0.560	0.138	0.138	0.194	0.144	0.143
200	800	0.25	0.20	0.132	0.405	0.432	0.562	0.127	0.12	0.18	0.181	0.151
200	800	0.25	0.40	0.142	0.397	0.442	0.56	0.138	0.137	0.197	0.208	0.164
400	800	0.15	0.20	0.106	0.402	0.443	0.560	0.129	0.129	0.184	0.148	0.149
400	800	0.15	0.40	0.134	0.394	0.452	0.559	0.14	0.137	0.198	0.141	0.140
400	800	0.25	0.20	0.13	0.391	0.431	0.546	0.125	0.122	0.189	0.180	0.149
400	800	0.25	0.40	0.141	0.396	0.429	0.561	0.143	0.142	0.196	0.165	0.213
200	1200	0.15	0.20	0.127	0.383	0.412	0.554	0.137	0.131	0.161	0.145	0.176
200	1200	0.15	0.40	0.143	0.404	0.441	0.558	0.14	0.138	0.186	0.218	0.149
200	1200	0.25	0.20	0.137	0.393	0.417	0.558	0.142	0.14	0.178	0.224	0.219
200	1200	0.25	0.40	0.148	0.393	0.434	0.56	0.141	0.14	0.188	0.163	0.238
400	1200	0.15	0.20	0.126	0.398	0.426	0.559	0.14	0.142	0.183	0.194	0.147
400	1200	0.15	0.40	0.14	0.396	0.427	0.559	0.142	0.141	0.184	0.192	0.221
400	1200	0.25	0.20	0.126	0.401	0.428	0.560	0.139	0.141	0.181	0.165	0.220
400	1200	0.25	0.40	0.142	0.397	0.420	0.559	0.143	0.147	0.187	0.165	0.242

*n* is the sample size; *q* is the number omics measurements in each layer;  
*h* measures the strength of regulation across layers;  $\rho$  is the correlation coefficient among CNVs

# 仿真Scenario II

## ■ 数据产生模型

$$Y = X\beta_1 + U_1\gamma_1 + \epsilon_1, \quad Z = Y\beta_2 + U_2\gamma_2 + \epsilon_2. \quad (12)$$

其中 $U_1$ 和 $U_2$ 是描述影响 $Y$ 和 $Z$ 的其他未知的调控机制。



**Table 4** Simulation results for Scenario III

Parameters				$M_{\text{accuracy}}$								
$n$	$q$	$h$	$\rho$	MuNCut	KM	SC	HC	KM*	SC*	HC*	LC	FGC
200	400	0.15	0.20	0.064	0.359	0.459	0.583	0.124	0.125	0.186	0.172	0.200
200	400	0.15	0.40	0.108	0.354	0.464	0.582	0.124	0.126	0.194	0.171	0.171
200	400	0.25	0.20	0.126	0.360	0.462	0.584	0.127	0.128	0.186	0.192	0.224
200	400	0.25	0.40	0.141	0.355	0.468	0.578	0.131	0.129	0.198	0.147	0.144
400	400	0.15	0.20	0.06	0.356	0.457	0.583	0.121	0.123	0.185	0.171	0.169
400	400	0.15	0.40	0.097	0.354	0.46	0.587	0.12	0.124	0.193	0.164	0.162
400	400	0.25	0.20	0.121	0.358	0.456	0.585	0.122	0.123	0.185	0.174	0.152
400	400	0.25	0.40	0.138	0.357	0.462	0.586	0.124	0.124	0.191	0.134	0.136
200	800	0.15	0.20	0.122	0.314	0.434	0.578	0.13	0.132	0.189	0.175	0.172
200	800	0.15	0.40	0.134	0.315	0.431	0.579	0.139	0.134	0.19	0.212	0.172
200	800	0.25	0.20	0.142	0.32	0.402	0.567	0.128	0.128	0.19	0.202	0.195
200	800	0.25	0.40	0.144	0.318	0.414	0.58	0.146	0.144	0.196	0.166	0.206
400	800	0.15	0.20	0.121	0.321	0.421	0.578	0.129	0.129	0.174	0.191	0.153
400	800	0.15	0.40	0.144	0.321	0.427	0.577	0.146	0.144	0.193	0.165	0.216
400	800	0.25	0.20	0.141	0.315	0.424	0.578	0.127	0.128	0.173	0.152	0.197
400	800	0.25	0.40	0.143	0.312	0.439	0.579	0.131	0.134	0.188	0.168	0.228
200	1200	0.15	0.20	0.138	0.307	0.391	0.578	0.139	0.139	0.168	0.207	0.212
200	1200	0.15	0.40	0.146	0.314	0.389	0.575	0.148	0.147	0.19	0.160	0.235
200	1200	0.25	0.20	0.136	0.30	0.374	0.575	0.136	0.133	0.169	0.187	0.225
200	1200	0.25	0.40	0.144	0.308	0.405	0.572	0.146	0.145	0.189	0.169	0.232
400	1200	0.15	0.20	0.136	0.316	0.406	0.573	0.138	0.139	0.163	0.159	0.223
400	1200	0.15	0.40	0.144	0.30	0.389	0.571	0.146	0.145	0.189	0.165	0.239
400	1200	0.25	0.20	0.141	0.316	0.376	0.577	0.135	0.139	0.183	0.171	0.228
400	1200	0.25	0.40	0.141	0.308	0.391	0.575	0.139	0.14	0.186	0.146	0.219


$n$  is the sample size;  $q$  is the number omics measurements in each layer;

$h$  measures the strength of regulation across layers;  $\rho$  is the correlation coefficient among CNVs



## 思考：

- 研究多组学聚类时，每一层中类的数目并不一定相同；
- 文章使用的聚类方法需要预先知道类的个数，当节点数目很多时不适用。
- 针对多组学数据的网络融合方法，对融合网络进行聚类？？
- 首先对单组学数据进行聚类，然后评估不同组学间的类-类或者子类-子类的关系，从而找到这种channel？？为理解生物功能等有意义。



谢谢