基迪奥生物
GENE DENOVO

Omicshare 课堂第25期

初探WGCNA分析原理和应用

广州基迪奥生物科技有限公司
广州大学城外环东路280号健康产业产学研孵化基地
020-39340225 | service@genedenovo.com
http://www.genedenovo.com/

纲要

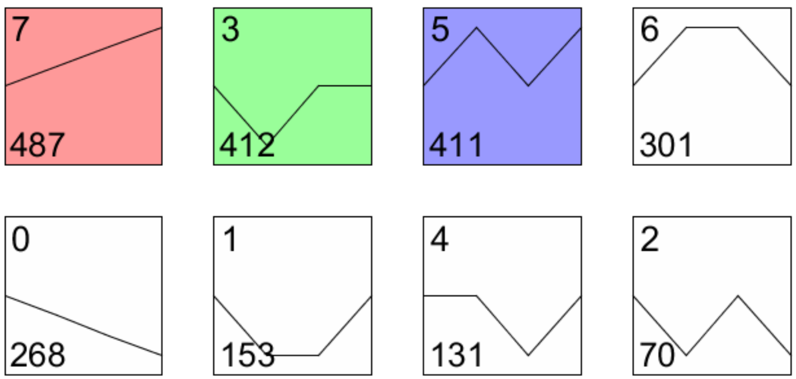
- 背景
- 与网络相关的一些基础概念
- WGCNA网络原理和构建过程
- WGCNA网络生物学意义的挖掘

RNA-seq数据挖掘的一般逻辑

（1）按照表达规律对基因归类

两组样本：上调、不变、下调

数组样本：趋势分析

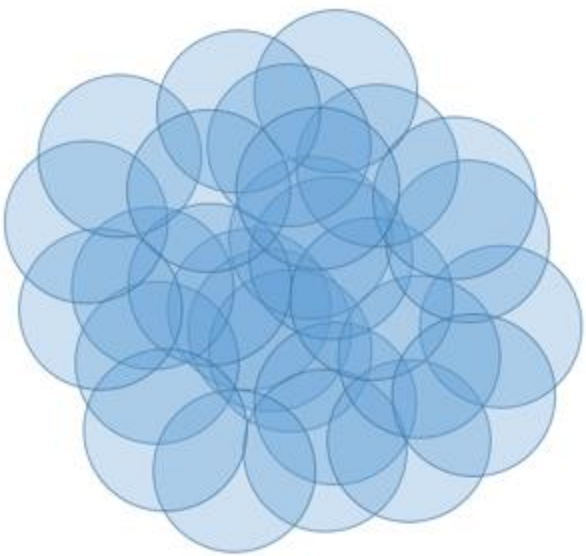


（2）对已分类的基因开展功能、调控分析

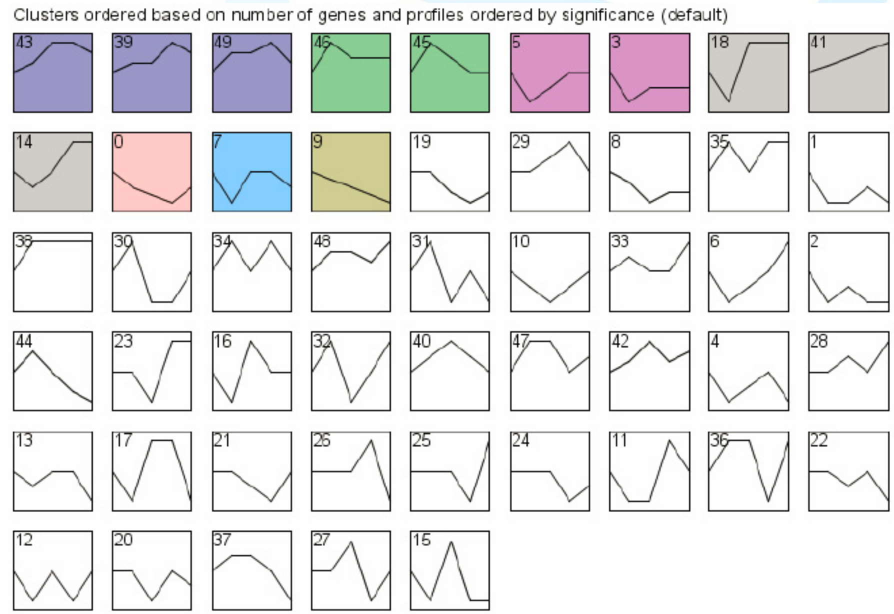
GO、KEGG分析

已有方法的局限

（1）在大样本中，差异分析或趋势分析无法对基因进行有效分类



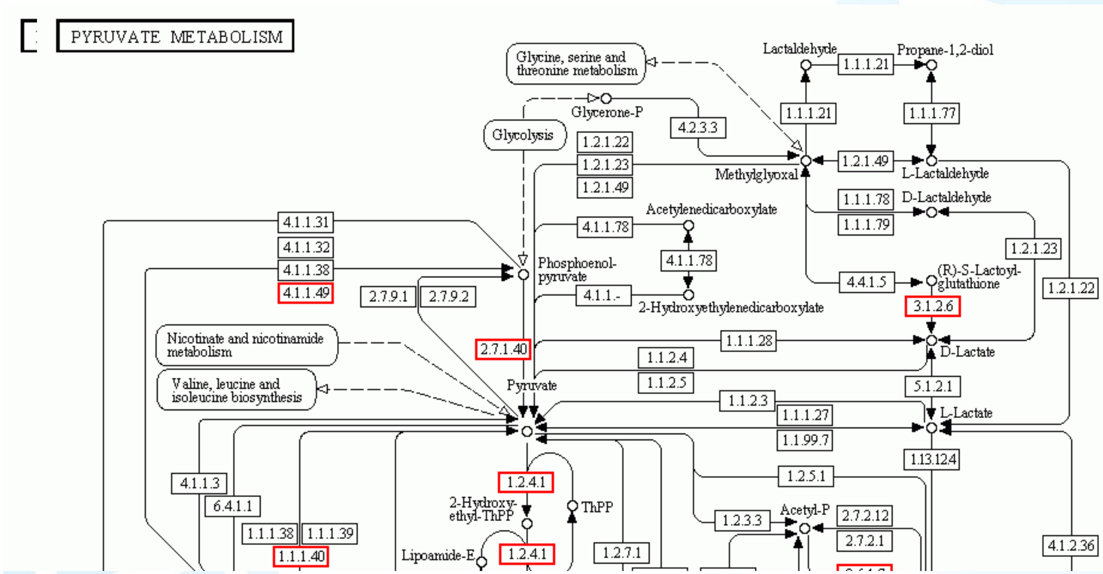
比较组太多的时候，维恩图无能为力



STEM类型的聚类软件，对样本构成复杂的情况下，也不能做高效、简洁的分类

已有方法的局限

（2）依赖数据的功能分析无法推测新的调控关系



Kegg的pathway都来源文献已报到的调控关系

如果你关注的调控关系在已有数据库未录入（或还没被报道），依赖这些分析是难以找到线索的。

基因表达和调控方式

1、组成型表达（constitutive gene expression）

细胞中持续进行的基因表达，且较少受环境因素的影响。这类基因通常被称为持家基因（housekeeping gene）。

2、基因间相互诱导和阻遏表达

即一个基因诱导（例如转录因子）或阻遏（例如 miRNA）另外一个基因的转录翻译。

3、协调表达

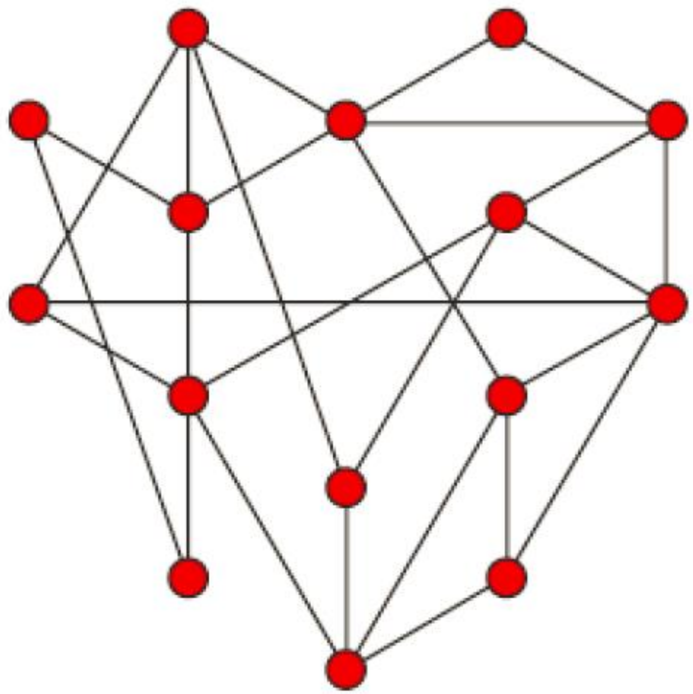
在一定机制控制下，功能相关的一组基因，协调一致，共同表达。

以上第二种和第三种情况，都会导致相关基因表达量间存在相关性。

纲要

- 背景
- 与网络相关的一些基础概念
- WGCNA网络原理和构建过程
- WGCNA网络生物学意义的挖掘

简单基因网络结构示意图



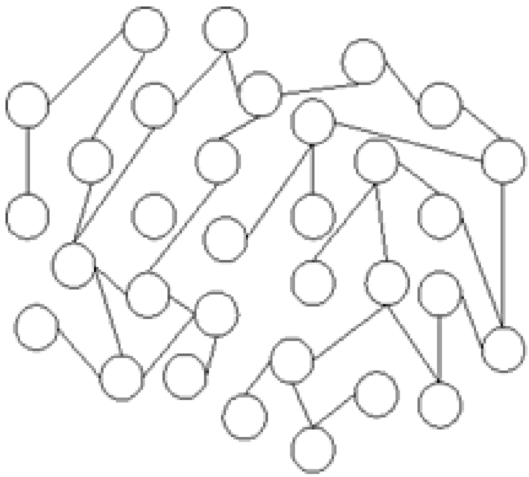
网络图的两大元素：

点：图中每一个圆圈代表一个节点，如基因。

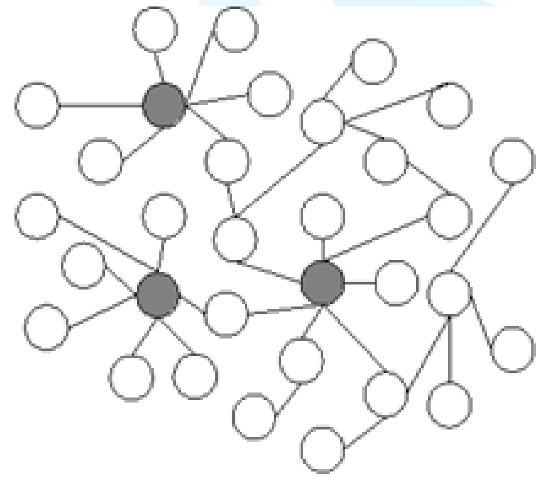
边：在基因调控网络中，基因相互间的调控关系构成了边。

无尺度网络（scale free network）

- 基因网络符合无尺度分布



随机网络（random network）



无尺度网络（Scale-free network）

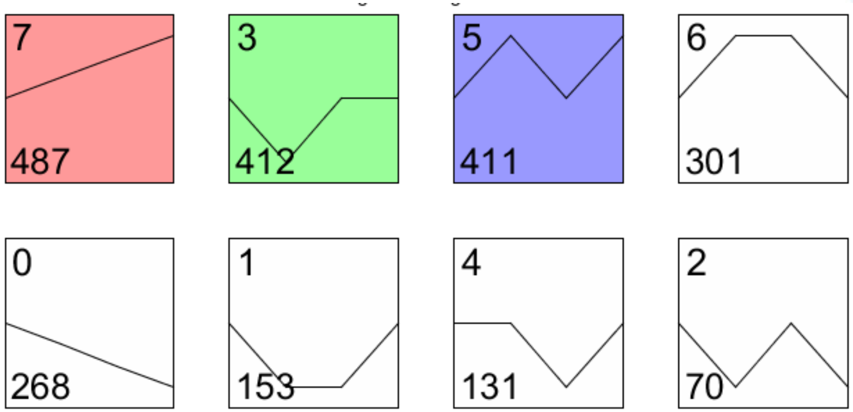
无尺度网络：大部分节点只和很少节点连接，而有极少的节点与非常多的节点连接。

数学上的描述：假设m为某节点的连接数。统计所有基因的m值，然后以m高低为指标对所有基因分类。n为节点连接数为m的基因的数量。m与n应该成反比。

模块以及模块特征值

- 模块(module)：表达模式相似的一组基因

趋势分析的本质也是分模块，但其功能没有WGCNA中的方法强大。



- 模块特征值（module eigengene）

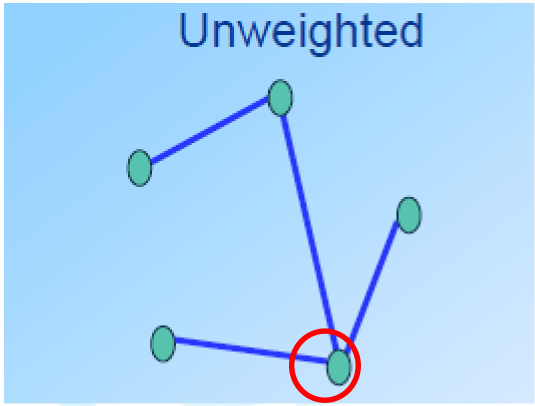
模块中的所有基因进行PCA分析，得到的主成分1（PC1）的值。PC1相当于模块中所有基因表达量的加权，可以代表这个模块的表达模式。

连通性

- 连通性（connectivity）：

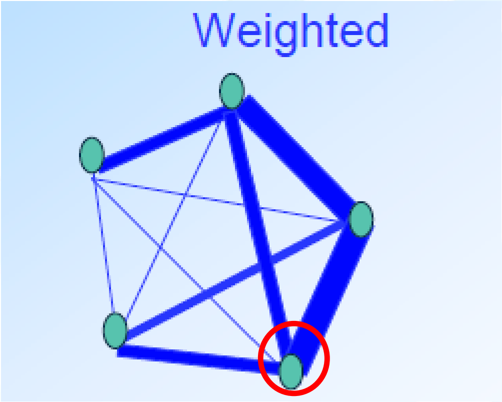
一个基因与其他基因的连接程度（通常只在模块内计算）。常称为connectivity或degree，或用数字k表示。有两种计算方法：

非权重网络



当两个基因的相关性大于某个值（例如0.6），才认为有相关性。最后某个基因的k值等于与其相关的基因的数量。红圈中的基因k=3。

权重网络



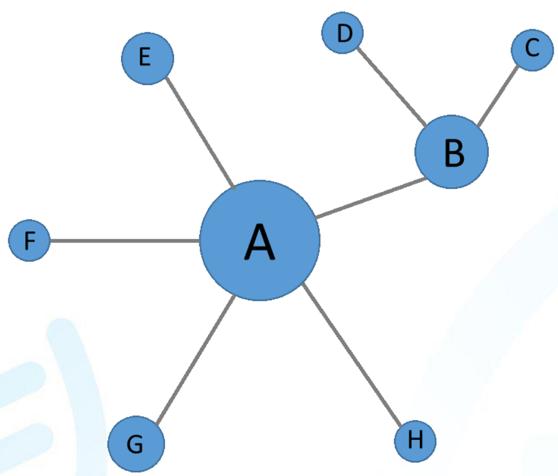
所有两两基因的相关性都被保留（无论强弱）。某个基因的k值等于其与各个基因的相关性之和。红圈中的基因k=四个相关性之和。

备注：WGCNA可以兼顾权重网络和非权重网络的特性（下文会介绍）

核心基因

- 核心基因（hub gene）

在一个模块中，连通性（k值）排名靠前的基因。K值排名靠前本身已经表明它们处于中心枢纽的位置。如下图中的A基因。



专注科研服务

核心基因的现实类比

- 现实的民航路线图，也是个“城市”间的调控网络。处于网络中心的城市包含北京、上海、广州、成都等中心城市，这有助于我们理解“核心基因”的概念

国内民航航线图

纲要

- 背景
- 与网络相关的一些基础概念
- WGCNA网络原理和构建过程
- WGCNA网络生物学意义的挖掘

WGCNA介绍

WGCNA，全称weighted gene co-expression network analysis，即权重基因共表达网络分析。

自2005年B Zhang, S Horvath等提出，至今已被引用694次。在疾病以及其他性状与基因关联分析等方面的研究中被广泛应用。

WGCNA方法的哲学理念

- 从“系统”的角度去解析关注的问题，而不是去罗列单一基因的清单。
  - 描述整个“发动机”的工作原理，而不是去摆放所有的螺丝钉。
- 研究的关注点是基因模块，而不是单一基因。
- 网络的概念让抽象的生物学问题更直观易懂。

WGCNA网络构建的两个核心步骤

Step1：计算基因间的相关性

Step2：将基因划分为模块

基因相关关系的无尺度化

- 处理过程

1）原始S矩阵：$S_{ij}=|cor(x_i, x_j)|$ # 计算基因间两两相关性

2）无尺度化（拉大贫富差距），确定最佳β值，得到A矩阵

A矩阵： $a_{ij}=power(S_{ij}, \beta)=|Smn|^{\beta}$

幂函数处理的作用：

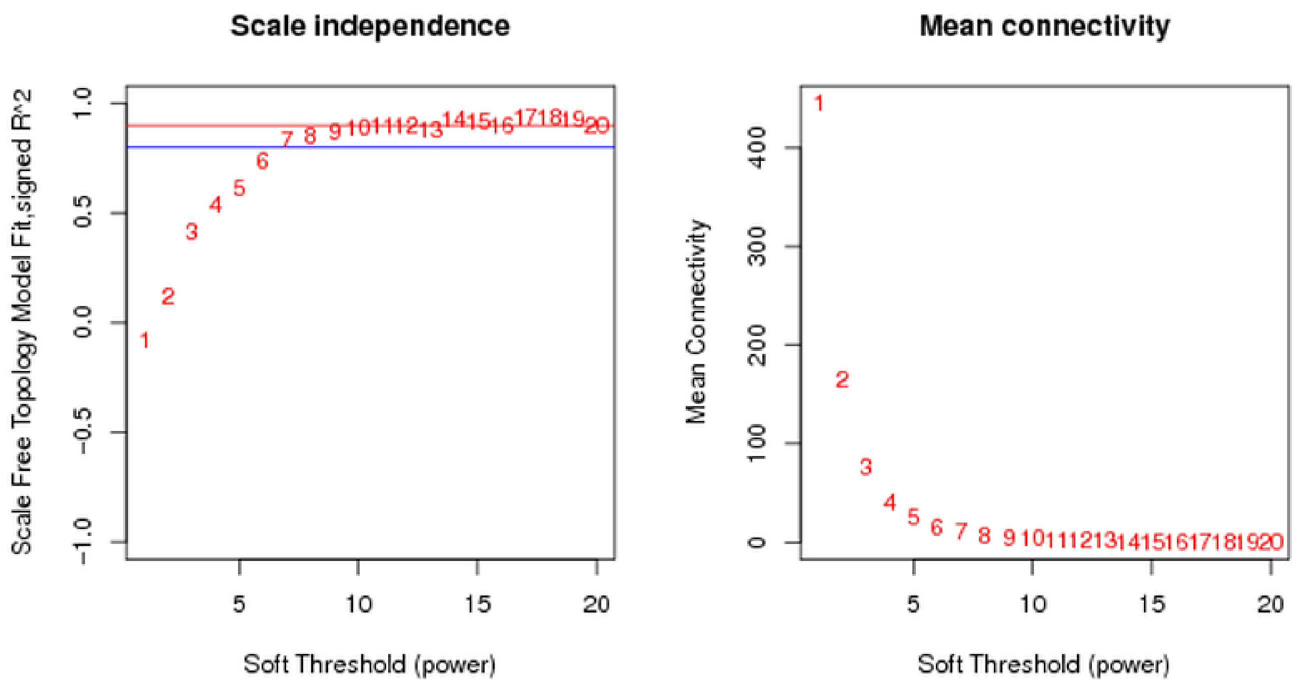
a）在幂函数处理后，少量强相关性的关系（例如：$r^2=0.999$）不受影响或影响较少，相关性弱的关系（例如：$r^2=0.1$）取n次幂后，相关性下降明显。

b) 最后导致网络关系无尺度化：强相关的关系少，弱相关的关系多；连通性高的基因少，连通性弱的基因多。

即，满足上文提到的无尺度网络的定义 "假设m为某节点的连接数。统计所有基因的m值，然后以m高低为指标对所有基因分类。n为节点连接数为m的基因的数量。m与n应该成反比。"

备注： 基因i的连通性： $k_i = \sum_j a_{ij}$ # 即其与其它基因的相关性之和。

关键参数β的确定



β值：WGCNA分析的第一个关键参数。

左图：不同β值下，m与n的相关性的变化（上一页提到，理论上是负值，应该做过转换）。一般认为取β值大于0.8或到达平台期时最小的β值用于构建网络。

右图：不同β值下，所有基因连通性的均值。

基因间表达调控的相关性

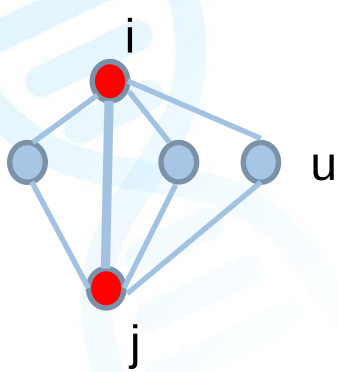
3）如何评估两个基因的表达模式的相关性

生物学逻辑：调控的相关性=直接相关 + 间接相关

数学实现过程：基于A矩阵,计算两两基因间的TOM值,

详析TOM值 ：TOM值= 直接相关 + 间接相关

$$TOM_{ij} = \frac{\sum_{u} a_{iu}a_{uj} + a_{ij}}{\min(k_i, k_j) + 1 - a_{ij}}$$



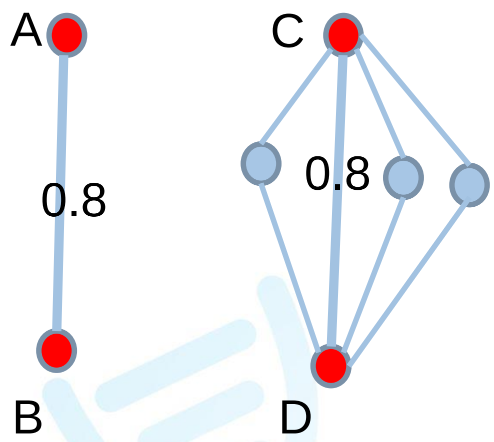
公式的亮点：

$$\sum_{u} a_{iu}a_{uj}$$

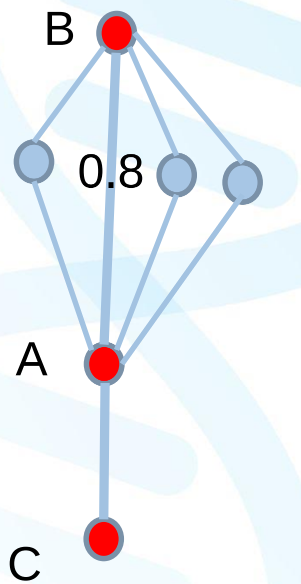
整合了基因i，j和第三个基因u的相关性，即间接相关

基因间表达调控的相关性

- TOM值在实际生物网络中的作用（TOM值在R语言的分析输出结果中是weight值）



A-B，C-D，他们间的直接相关性相同，但TOM_{CD}>TOM_{AB}

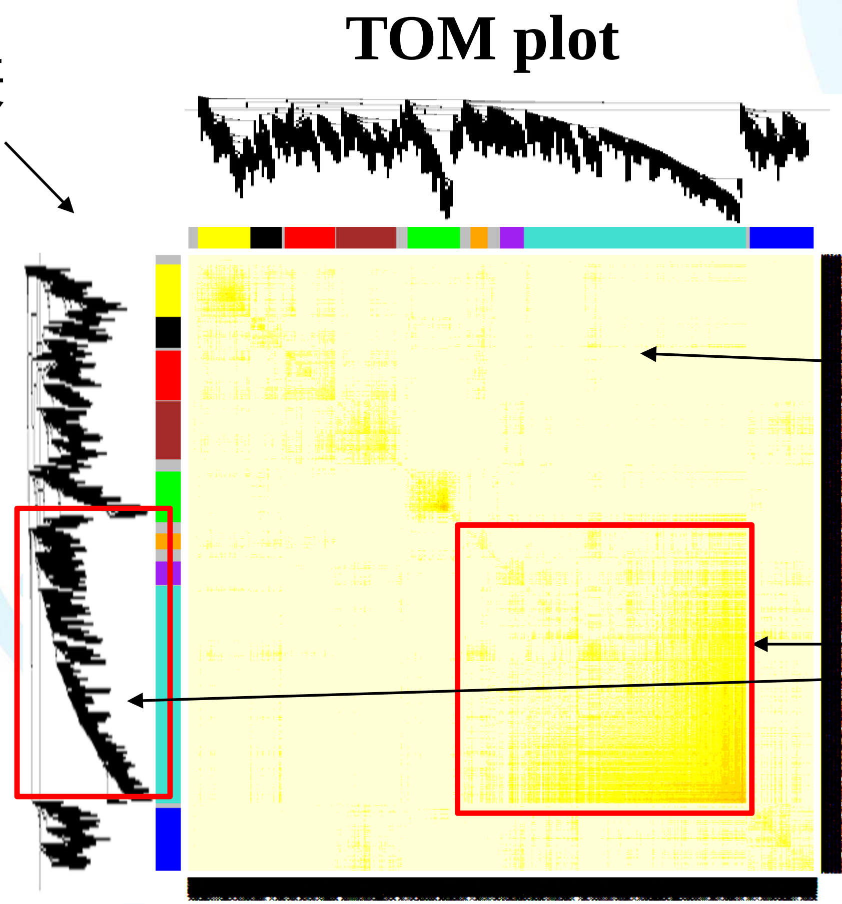


假设A是目标基因，B是上游的转录因子，C下游一个酶基因。B与A更可能形成更多间接相关（B是个核心基因），所以TOM_{AB} > TOM_{AC}

TOM矩阵与TOM聚类树

——这里相关性矩阵的模块对应树的某个分支

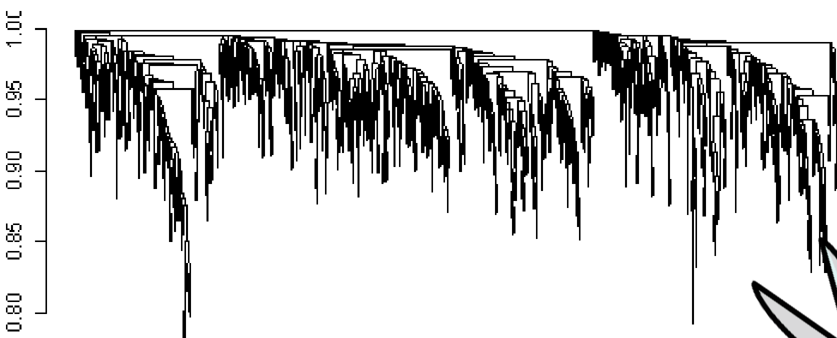
利用TOM（表达相关性）可以构建聚类树

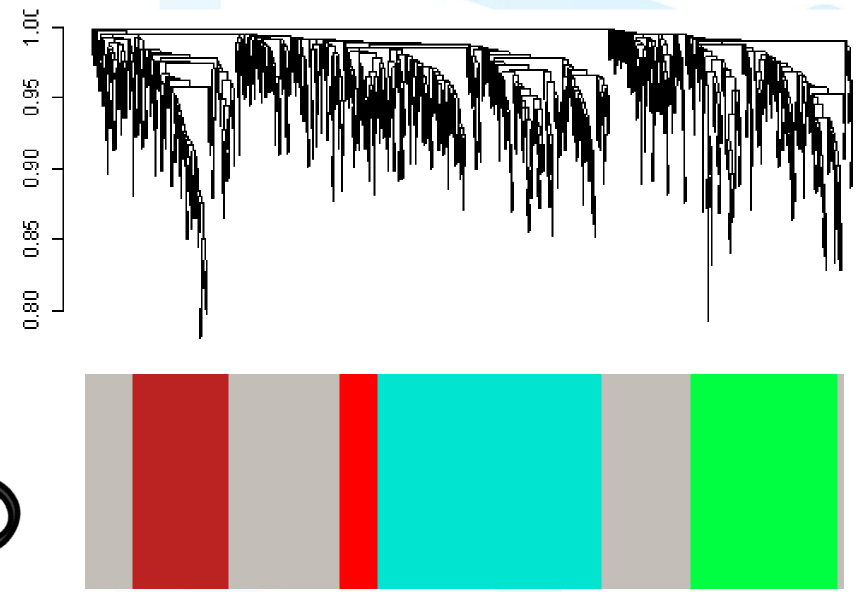


TOM 矩阵：Genes对应这个矩阵的行和列

模块:一簇表达量高度相关的基因集，对应树的一个分支

层级聚类树与区分模块

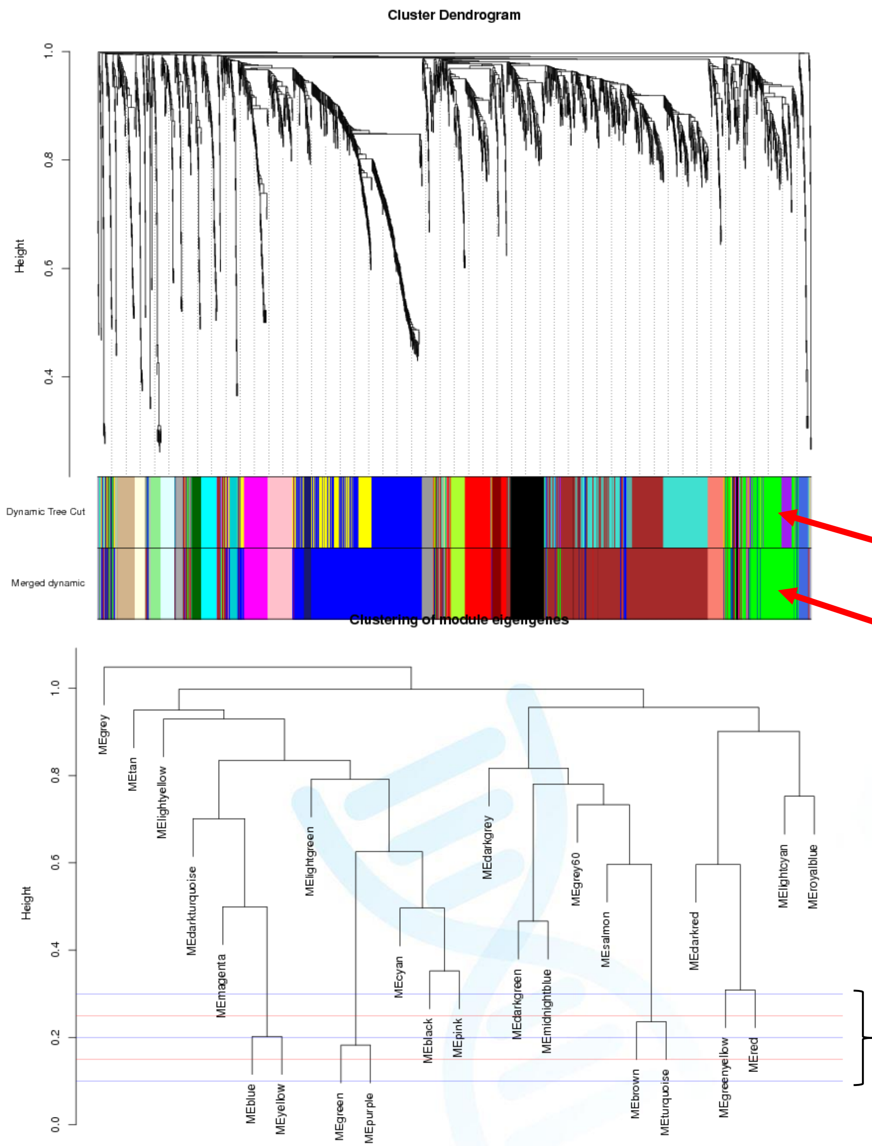




利用基因间的TOM值进行层级聚类，构建树（表达模式相似的基因属于1个分支）

- 对分支进剪切区分，产生不同的模块。
- 一个颜色代表一个模块
- 灰色的模块代表无法归入任何一个模块的基因

聚类和分模块的几个关键参数



WGCNA包中，有一个自动聚类的命令：blockwiseModules
其中包含的参数包括：

Power:即上文的β值；
minModuleSize：模块最少基因数
mergeCutHeight：在自动进行模块划分后，合并相似的模块

初步划分模块的结果

合并相似模块后的结果

mergecutHeight: 本质是利用模块的特征值构建树，然后将属于同一个树分支，且距离很近的模块合并（这里取值是0.25）

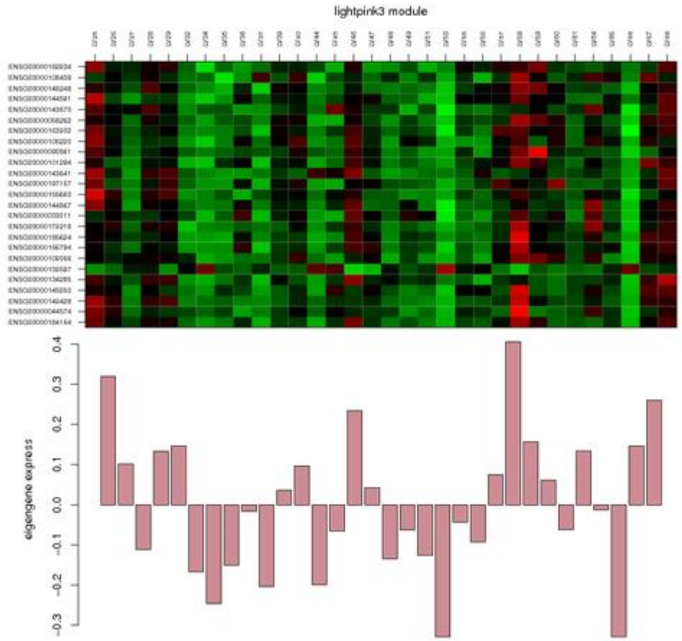
纲要

- 背景
- 与网络相关的一些基础概念
- WGCNA网络原理和构建过程
- WGCNA网络生物学意义的挖掘

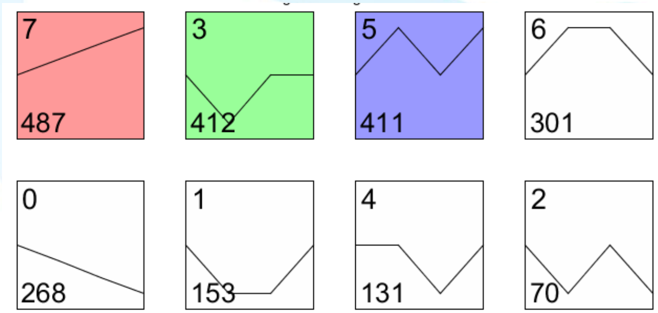
生物学意义挖掘

- 目标模块选取
  - 模块表达模式分析
  - 模块与其他指标的相关性分析
  - 富集分析（常见为功能富集分析）
  - 从目标基因直接入手
- 模块内的分析
  - 核心基因分析
  - 目标基因相关的局部调控网络
  - 关注特定类型的基因

目标模块选取——模块表达模式



VS



WGCNA的模块表达模式

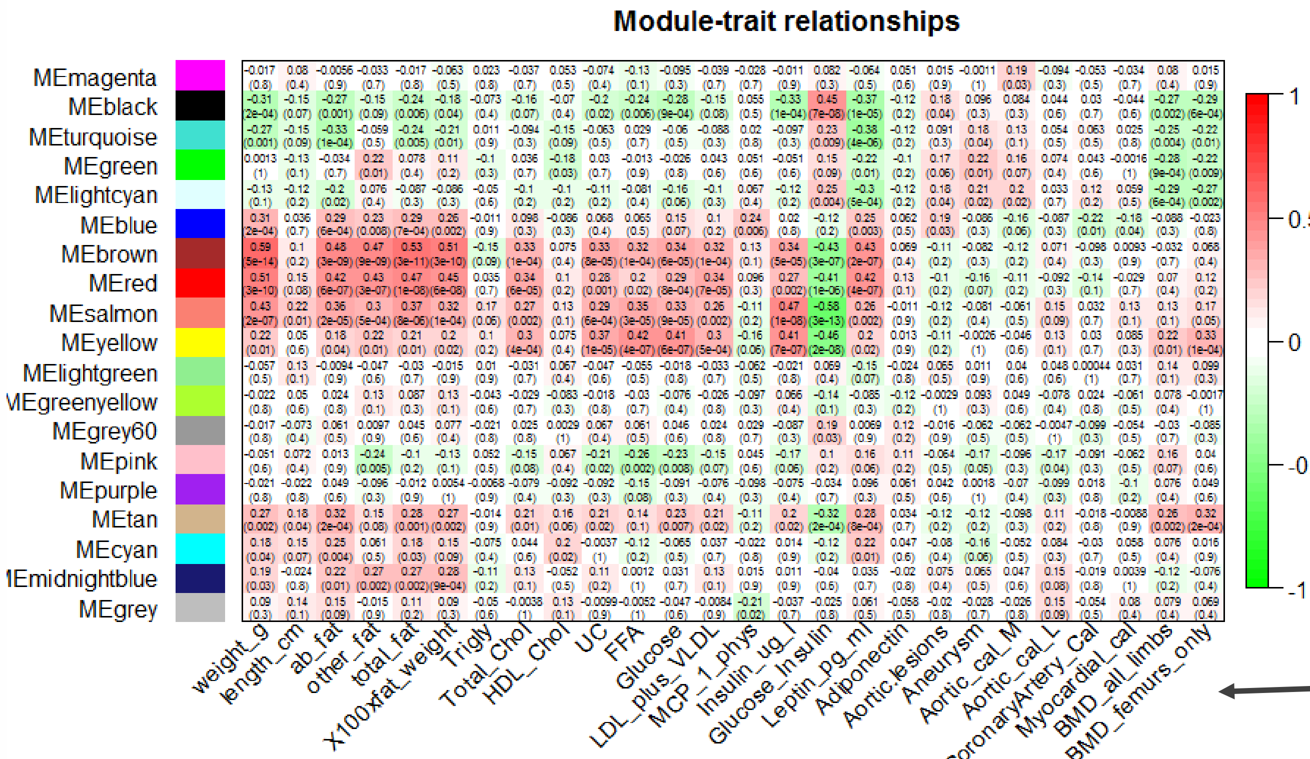
趋势分析中的表达模式

- 模块特征值（即PC1）在各个样本中的丰度，构成了这个模块的表达模式/规律
- 类似趋势分析中的表达趋势，但WGCNA分模块不受样本限制，所以通用性更强。

目标模块选取

——模块相关性分析

模块



性状

- 1个模块相当于代表一类基因，模块的特征值可以某种程度上代表一类基因的表达模式
- 样本在各个模块特征值可以和样本的性状，可以开展相关分析，找出与特定性状相关联的模块
- 模块也可以其他信息开展相关性分析，例如 样本SNP基因型，样本分组信息

目标模块选取——富集分析

如上文提到的 存在诱导/阻遏表达（TF和靶基因）或协调表达（例如被同一个TF调控一组基因）关系的一组基因，更容易出现在一个模块中。

而且在大样本的情况下，基因的表达分类更加有规律。

所以，对模块开展KEGG、GO功能富集分析，通常会找到很有规律的功能类型。

目标模块选取
——从模板基因入手

- 当然，如果有目标基因，也可以直接找其所在的模块，然后进行进行下一步分析——模块内的功能调控关系分析。

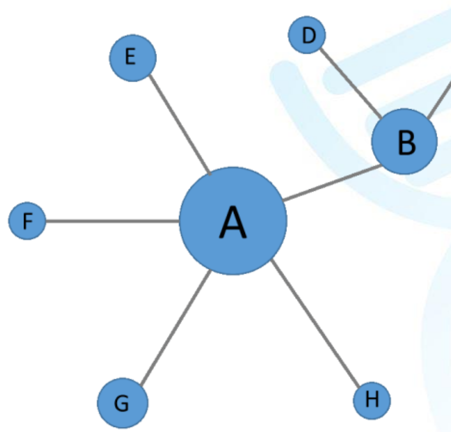
模块内的分析——基因的模块内连通性

- 模块内连通性（Module connectivity）

WGCNA模块内输出的Connectivity是模块内的连通性，是这个基因与其他基因相关性之和（记住：是a_{ij}，而不是TOM之和），所以这是软阈值（soft threshold）的计算方法（请回顾上文）。

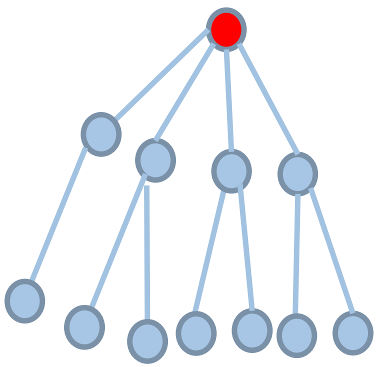
在某些文章里，也会使用硬阈值（hard threshold）的方法，即认为TOM值（就是模块调控关系表中的weight值）大于X（默认是0.15）的两个基因才认为是相关的，然后计算每个基因的连接数量（可以先过滤有足够强度的关系，然后导入cytoscape或OS-tools计算）

我们比较推荐使用前者。



fromNode	toNode	weight	direction	fromAltName	toAltName
MMT00000159	MMT00000793	0.074180225791694	undirected	Cdc2a	
MMT00000159	MMT00000840	0.0216775650220609	undirected	Cdc2a	
MMT00000159	MMT00001154	0.197142224192425	undirected	Cdc2a	
MMT00000159	MMT00001245	0.24126062408297	undirected	Cdc2a	
MMT00000159	MMT00001260	0.206752015676195	undirected	Cdc2a	
MMT00000159	MMT00001698	0.19874415334667	undirected	Cdc2a	
MMT00000159	MMT00002209	0.0223622257970916	undirected	Cdc2a	
MMT00000159	MMT00003188	0.125409921115193	undirected	Cdc2a	
MMT00000159	MMT00003410	0.241734122645486	undirected	Cdc2a	
MMT00000159	MMT00003994	0.024501365594856	undirected	Cdc2a	

模块内的分析——核心基因



调控网络中，偏上游的基因越容易获得更高的连通性。

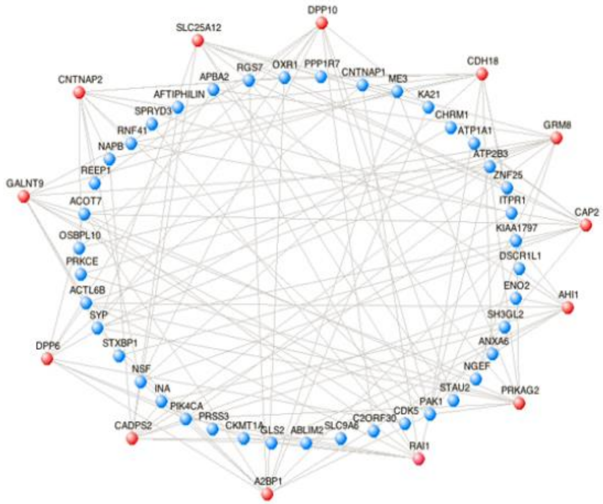
- 模块中连通性较高的基因（例如人为设定排名前30或前10%），被称为hub基因。
- 高连通性的Hub基因通常为调控因子（调控网络中处于偏上游的位置），而低连通性的基因通常为调控网络中偏下游的基因（例如，转运蛋白、催化酶等）
- 我们应该按照自己研究目的，充分利用连通性这个信息。

例如:

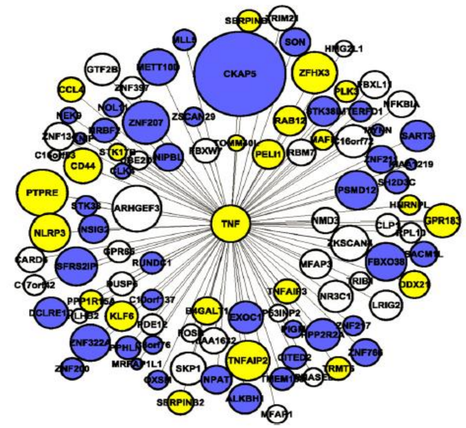
当只是对这个模块感兴趣，而没有特定目标，那么可以优先关注连通性高的基因。

但我已有特定目标，如转运蛋白，即使它连通性很低也是值得关注的。但我们可以优先挖掘与之强相关（TOM值很高）且自身连通性很高的基因（见下一页）。

模块内的分析
——目标基因相关的局部调控网络



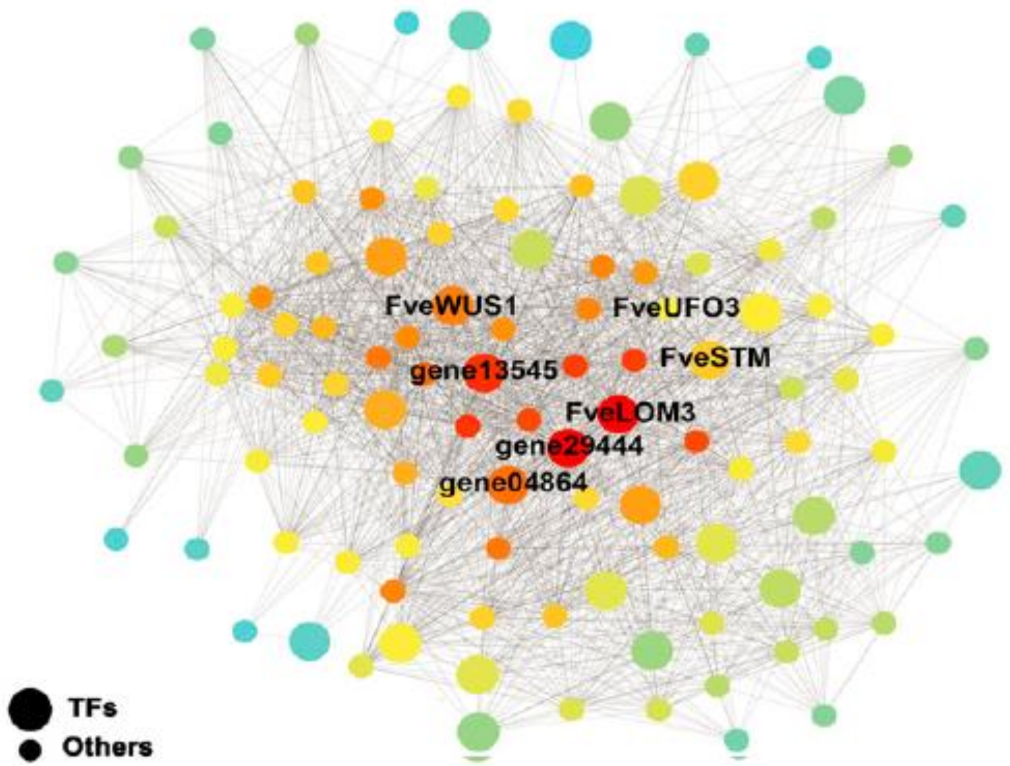
自闭症相关模块中，与13个自闭症候选基因（红色）相关性TOM值排名前10的基因（蓝色）
Voineagu I, Nature, 2011, 474(7351): 380-384.



与骨密度基因TNF的相关性TOM值大于0.15的一系列候选基因。
Farber C R. G3, 2013, 3(1): 119-129.

- 从目标基因入手，找与之TOM值排名靠前（例如前10）或TOM值大于某个阈值的基因列表。通过这一策略，可以准确筛选潜在与目标基因存在调控关系的候选基因，这些基因是下阶段功能验证的优先候选。

模块内的分析
——关注特定类型的基因



- 可以结合基因注释信息，关注模块内特定类型的基因，例如转录因子。

结果图形化

- 以上的大部分图形（除网络图），都是WGCNA R包可以输出。
- 网络图的必须自己定制：

（1）筛选目标基因（核心基因，已知目标基因）

（2） 筛选与目标基因相关的调控关系/基因（TOM值过滤）

（3） 整合多种信息，如目标调控关系、TOM值（调控强度）、连通性、基因种类（如转录因子），结合绘图软件（cytoscape 或 OS-tools网络图工具），就可以将以上得到的信息图形化。

网络图绘图工具推荐

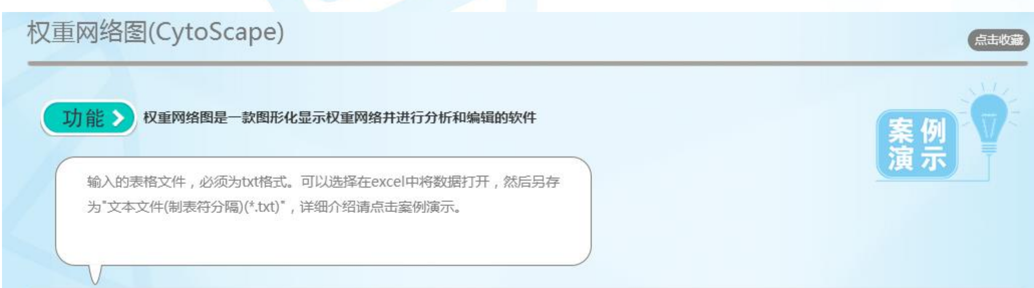
本地版Cytoscape教程：

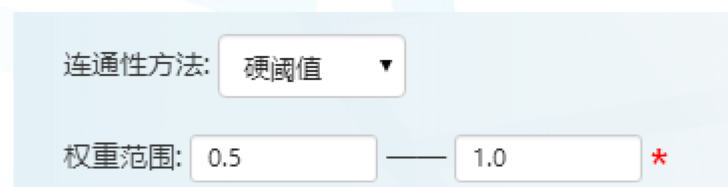
第19期在线交流cytoscape介绍与实操【视频】

http://www.omicshare.com/forum/thread-984-1-12.html

OS-tools 在线版cytoscape（做过优化，更适应cytoscape的原始输出结果和简单的数据过滤）

http://www.omicshare.com/tools/index.php/Home/Soft/cytoscape2





进行WGCNA分析的两个关键问题

- 1. WGCNA分析对样本有什么要求？

答：

我们推荐以下的样品数：

1）不含生物学重复的独立样本组：样本数≥8

2）包含生物学重复的样本组：样本数≥15

主要考虑的因素：

1）样本必须包含丰富的变化信息，才能区分为多个有意义的生物学模块（需要多个独立处理组）。

2）必须保证有多个样本，才能保证相关系数计算的准确性。

进行WGCNA分析的两个关键问题

- 2. 我如何实现WGCNA分析

答： 我们推荐两种途径：

1）如果你熟悉R语言，以及其他的配套分析方法（例如GO、KEGG富集分析）

可以自己动手练习和摸索

练习脚本地址（官方地址）：

https://labs.genetics.ucla.edu/horvath/CoexpressionNetwork/Rpackages/WGCNA/Tutorials/index.html

进行WGCNA分析的两个关键问题

- 2. 我如何实现WGCNA分析

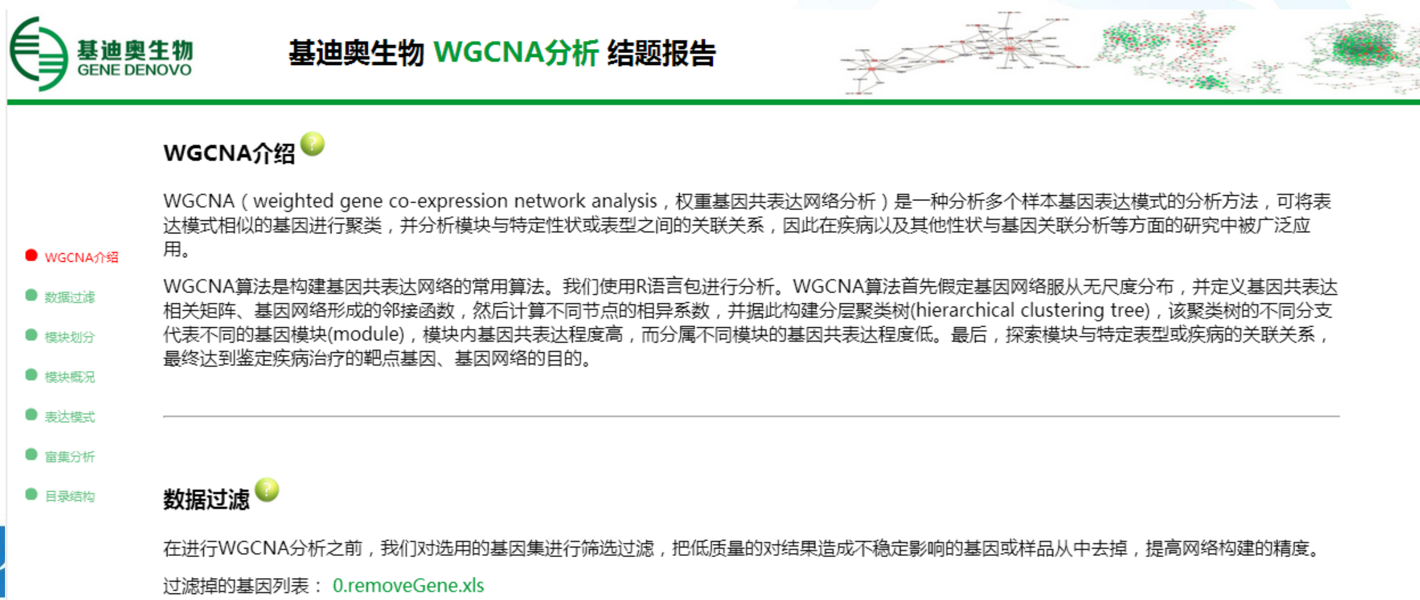
答： 我们推荐两种途径：

2）如果你不熟悉这一系列方法，或没有时间

可以交付给我们基迪奥公司，我们将提供

a）完善的WGCNA结题报告（丰富的分析结果和详细的说明文档）；

b）专业的售前，售后服务；


基迪奥生物 GENE DENOVO　基迪奥生物 WGCNA分析 结题报告

- WGCNA介绍
- 数据过滤
- 模块划分
- 模块概况
- 表达模式
- 富集分析
- 目录结构

WGCNA介绍

WGCNA（weighted gene co-expression network analysis，权重基因共表达网络分析）是一种分析多个样本基因表达模式的分析方法，可将表达模式相似的基因进行聚类，并分析模块与特定性状或表型之间的关联关系，因此在疾病以及其他性状与基因关联分析等方面的研究中被广泛应用。

WGCNA算法是构建基因共表达网络的常用算法。我们使用R语言包进行分析。WGCNA算法首先假定基因网络服从无尺度分布，并定义基因共表达相关矩阵、基因网络形成的邻接函数，然后计算不同节点的相异系数，并据此构建分层聚类树(hierarchical clustering tree)，该聚类树的不同分支代表不同的基因模块(module)，模块内基因共表达程度高，而分属不同模块的基因共表达程度低。最后，探索模块与特定表型或疾病的关联关系，最终达到鉴定疾病治疗的靶点基因、基因网络的目的。

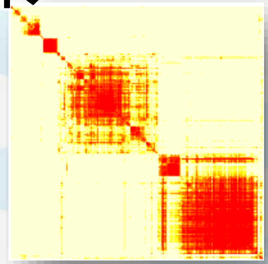
数据过滤

在进行WGCNA分析之前，我们对选用的基因集进行筛选过滤，把低质量的对结果造成不稳定影响的基因或样品从中去掉，提高网络构建的精度。

过滤掉的基因列表： 0.removeGene.xls

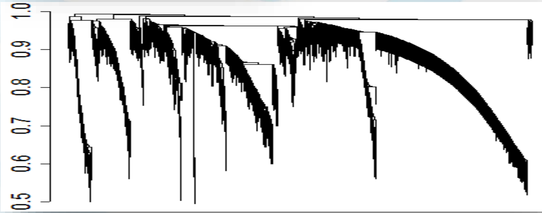
专注科

总结：WGCNA的步骤



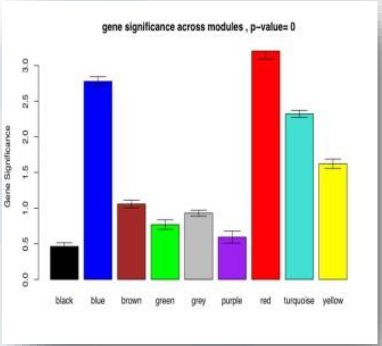
关系矩阵构建
基本原理: 利用基因间**表达量的相关系数**。

↓



模块识别
基本原理：利用拓扑树结构区分基因模块。

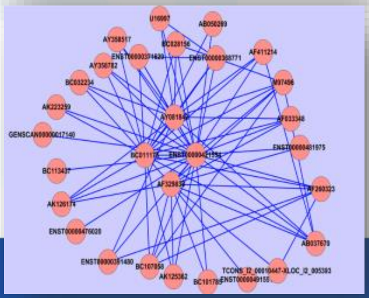
↓



核心模块挑选
基本原理：分析模块内基因的特性，进一步寻找有生物学意义的模块。

分析策略：模块特征值与表型间的相关性，模块内基因的KO、GO分析。

↓



核心基因的挑选，并构建网络
基本原理：利用基因连通性信息挑选核心基因，并围绕其构建网络

总结：WGCNA方法的特点

WGCNA 是一种系统的遗传分析方法，十分适用于分析复杂性状/疾病：

1）基于RNA表达量的调控关系分析，不需要基因间作用关系的先验信息。

2）强调模块 (通路)，而不是单一基因。

✓ 精简了组学水平的海量信息

✓ 在与其他信息（如表型）的相关性分析中，减少了多重检验校正的影响。

注：几十个模块的相关性分析 vs 几千个基因的相关性分析

3）使用基因连通性便于找到核心基因

✓核心基因在功能实验中往往有更高验证率。

4）调控网络相对稳定，受样本量影响小

✓相关调控网络基于相关系数，而不是p value（p value受样本量影响大）。因此，保证来源不同，样本数不同的数据间的可比较性

总结：WGCNA方法的特点

WGCNA 是一种系统的遗传分析方法，十分适用于分析复杂性状/疾病：

5）弱效应基因的挖掘对传统的DNA水平分析是个难点（例如，GWAS，连锁分析），但WGCNA分析系统挖掘的思路（模块相当于是众多微效基因的效应整合），对DNA水平的分析是很好的补充。

注：“GWAS分析 + WGCNA”的思路，在后续课程中可能会介绍。

还有一点：你可以获得免费的WGCNA分析

基迪奥生物
GENE DENOVO

All for Your Research