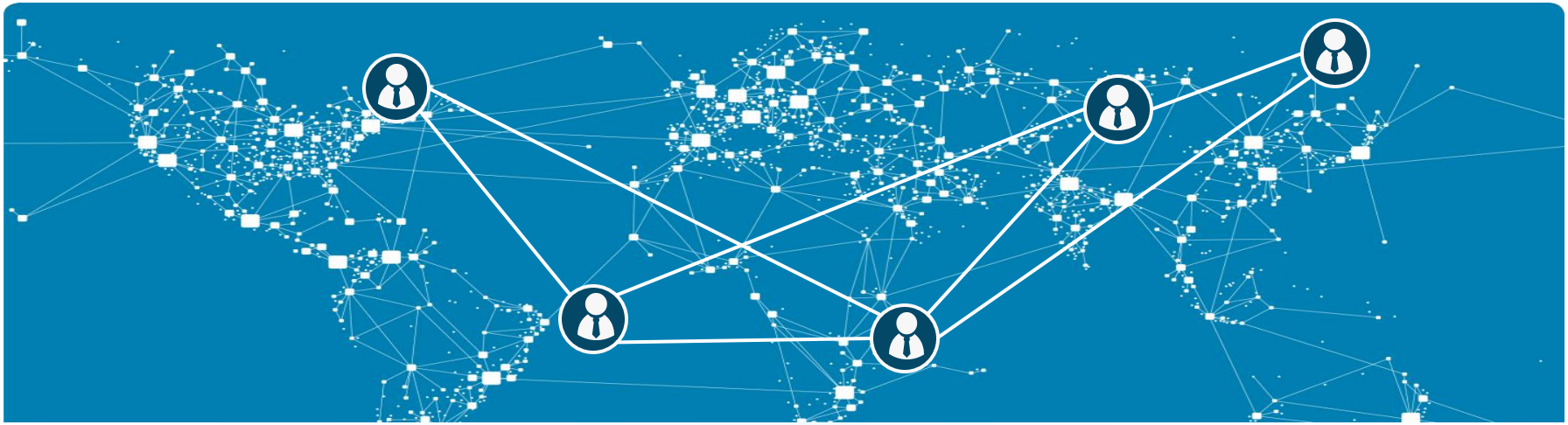
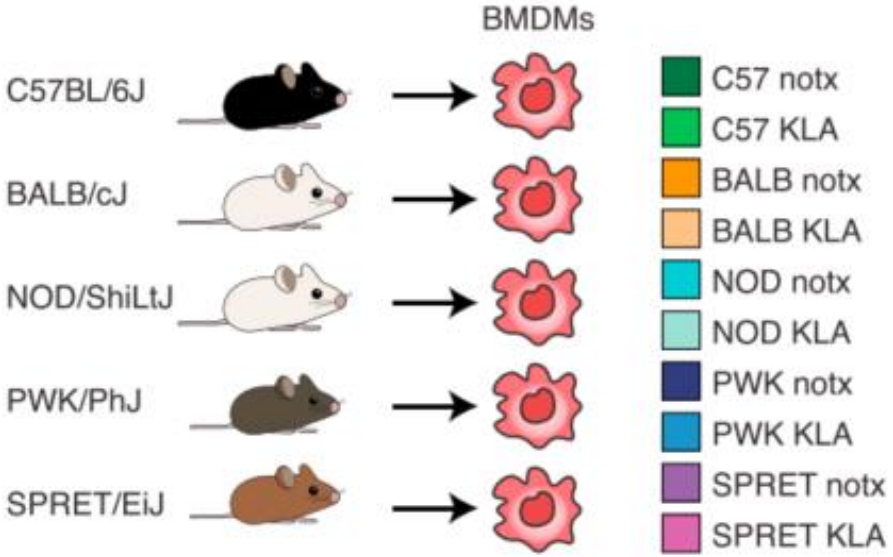
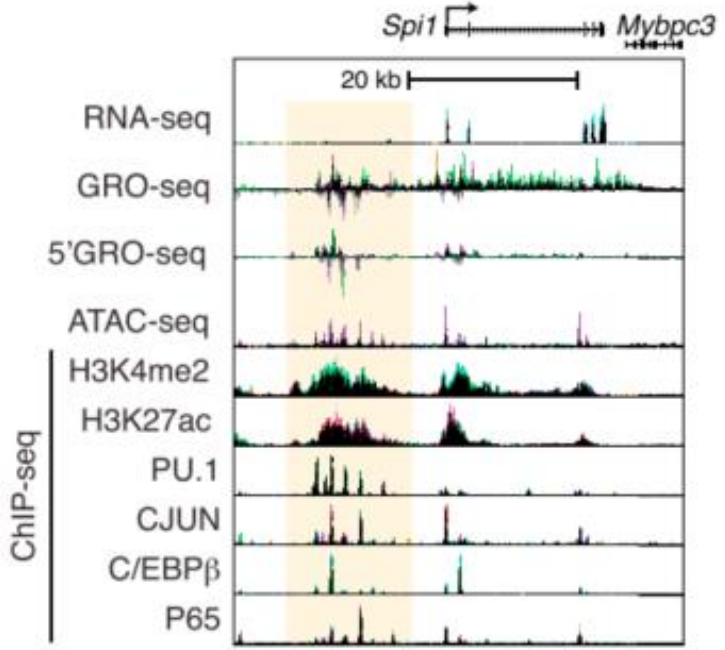


WGCNA共表达网络分析

WGCNA分析，简单全面的最新教程

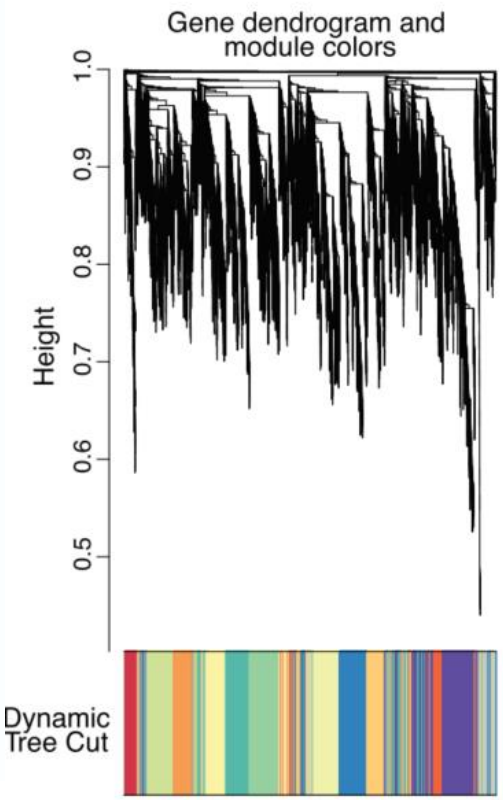
不同小鼠品系遗传突变对基因调控的影响

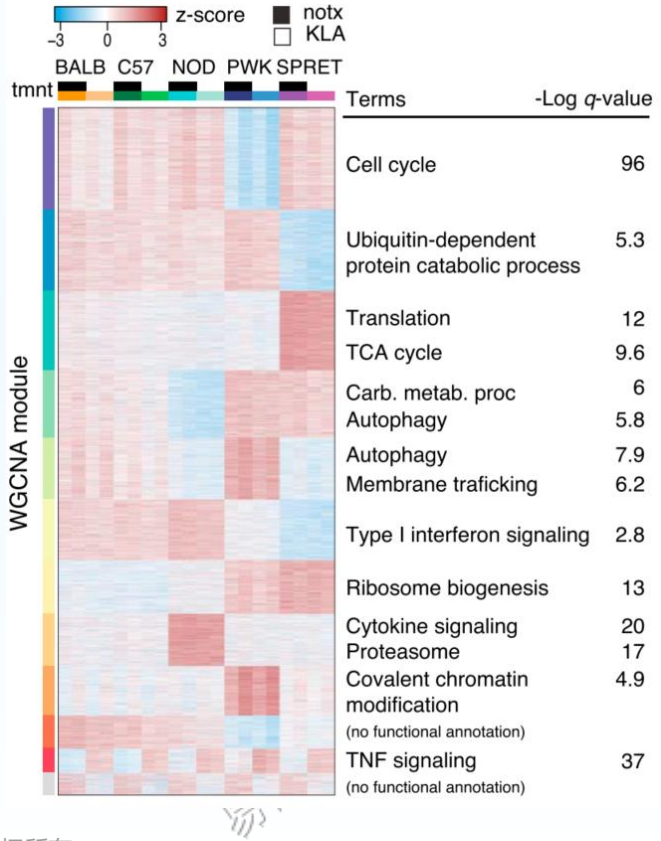




Analysis of Genetically Diverse Macrophages Reveals Local and Domain-wide mechanisms that Control Transcription Factor Binding and Function Cell

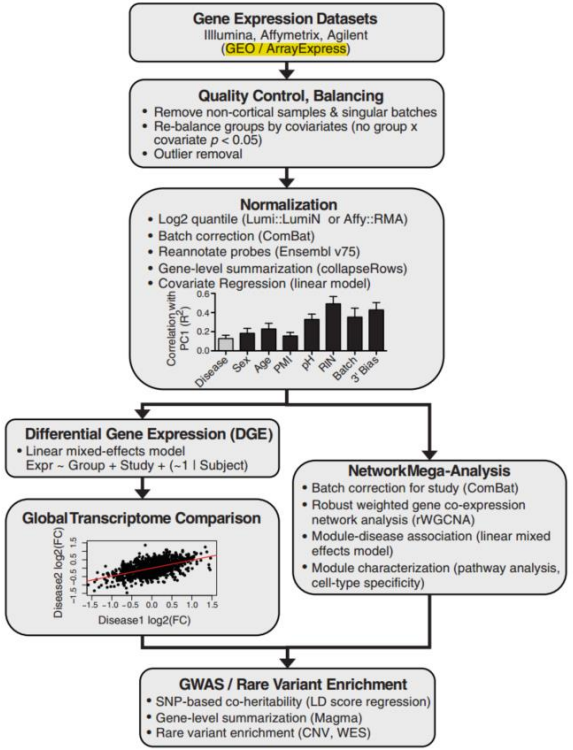
WGCNA鉴定模块并进行富集分析





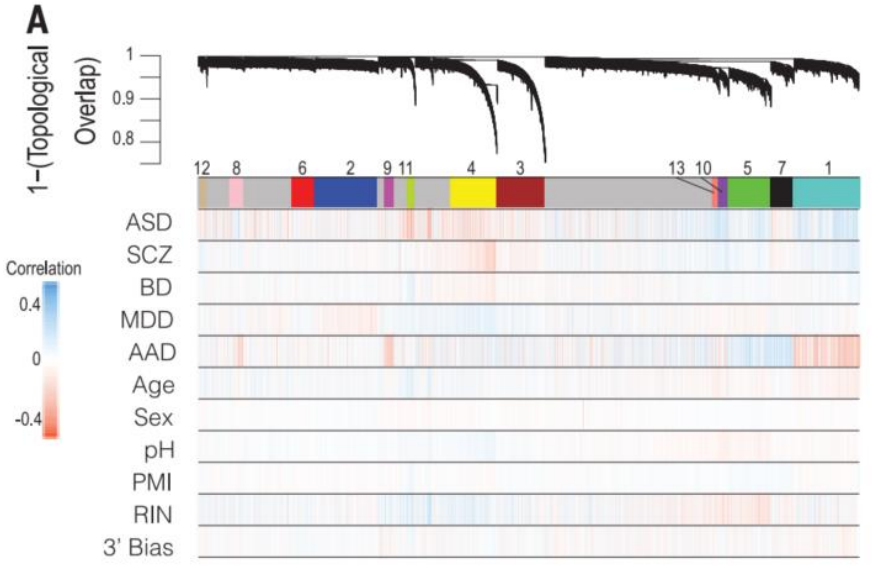
多个精神疾病间共有的分子神经病理学研究

- 从公共数据中下载5种神经类疾病的基因芯片数据，进行多重校正和WGCNA分析，鉴定出疾病相关的几个模块，并进行功能分析和与SNP关联，探索调控机制。



Shared molecular neuropathology across major psychiatric disorders parallels polygenic overlap Science

多个精神疾病间共有的分子神经病理学研究

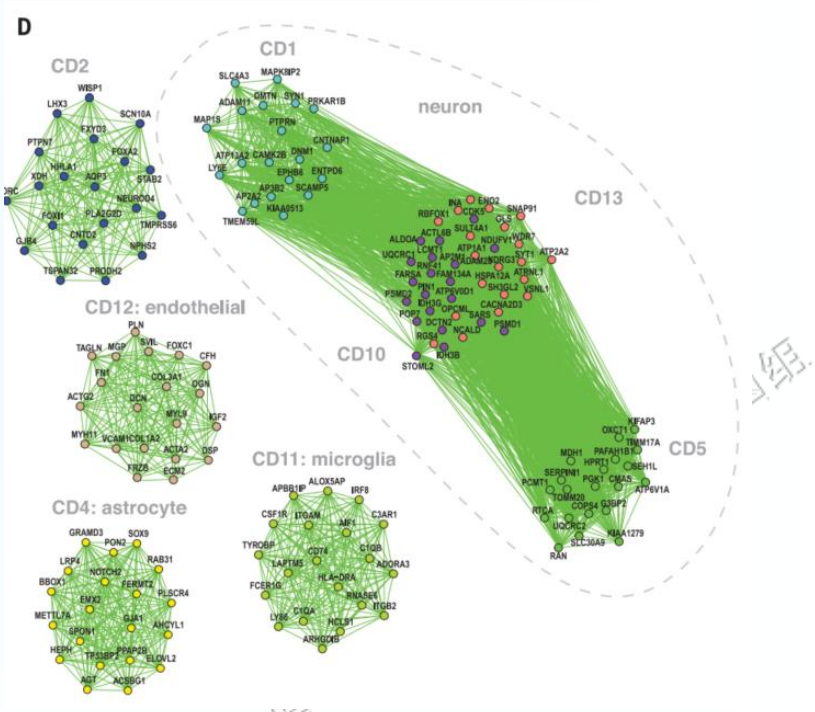


(**A**) Network dendrogram from co-expression topological overlap of genes across disorders. Color bars show correlation of gene expression with disease status, biological, and technical covariates.

Shared molecular neuropathology across major psychiatric disorders parallels polygenic overlap Science

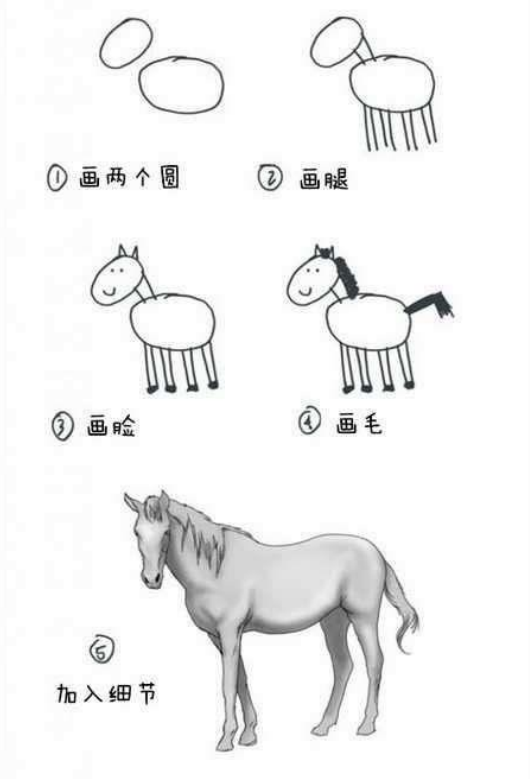
多个精神疾病间共有的分子神经病理学研究

- **D)** The top twenty hub genes are plotted for modules most disrupted in disease. See Data Table S2 for a complete list of genes' module membership (kME). Edges are weighted by the strength of correlation between genes.
- Modules are characterized by (**E)** Gene Ontology enrichment (top two pathways shown for each module)



Shared molecular neuropathology across major psychiatric disorders parallels polygenic overlap Science

方法在这了，能不能出成果就看自己打磨细节的时间了



WGCNA基本概念

- 加权基因共表达网络分析（WGCNA，Weighted gene correlation network analysis)是用来描述不同样品之间基因关联模式的系统生物学方法，可以用来鉴定高度**协同变化**的基因集，并根据基因集的内连性和基因集与表型之间的关联鉴定候补生物标记基因或治疗靶点。
- 相比于只关注差异表达的基因，WGCNA利用数千或近万个变化最大的基因或全部基因的信息识别感兴趣的基因集，并与表型进行显著性关联分析。一是充分利用了信息，二是把数千个基因与表型的关联转换为数个基因集与表型的关联，免去了多重假设检验校正的问题。

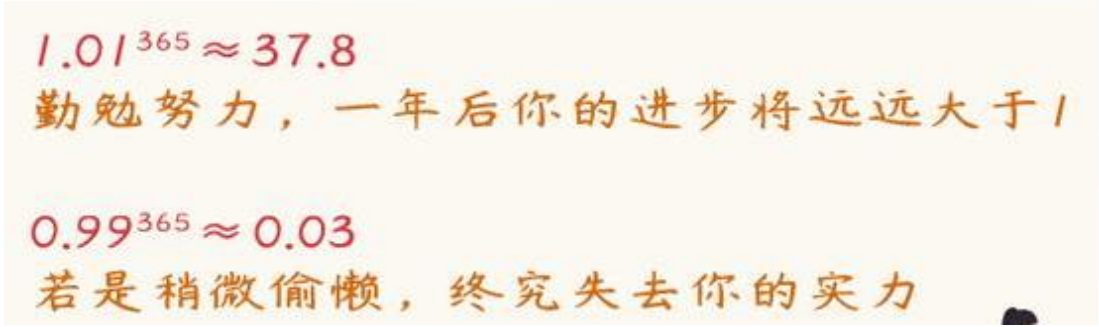
WGCNA术语解释 – 加权和方向

- 共表达网络：定义为加权基因网络。点代表基因，边代表基因表达相关性。加权是指对相关性值进行**幂次运算**（幂次的值也就是软阈值（power, pickSoftThreshold这个函数所做的就是确定合适的power））。

无向网络的边属性计算方式为 abs(cor(genex, geney)) ^ power；
有向网络的边属性计算方式为 (1+cor(genex, geney)/2) ^ power;
sign hybrid的边属性计算方式为cor(genex, geney)^power if cor>0 else 0。

WGCNA术语解释 – 加权和方向

- 这种处理方式强化了强相关，弱化了弱相关或负相关，使得相关性数值更符合无标度网络特征，更具有生物意义。如果没有合适的power，一般是由于部分样品与其它样品因为某种原因差别太大导致的，可根据具体问题移除部分样品或查看后面的经验值。



TOM矩阵

- TOM（Topological overlap matrix)：把邻接矩阵转换为拓扑重叠矩阵，以降低噪音和假相关，获得的新距离矩阵，这个信息可拿来构建网络或绘制TOM图。其定义依据是任何两个基因的相关性不只由它们自己的相关性决定，还依赖于与**这两个基因存在相关性的其它基因的互作**，把这些因素都考虑进来，才能更好地定义基因表达谱的相似性。

$$w_{i,j} = \frac{\sum_u (a_{iu} a_{uj}) + a_{ij}}{\min(\sum_u a_{iu}, \sum_u a_{ju}) + 1 - a_{ij}}$$

WGCNA术语解释 – 模块

- Module(模块)：高度内连的基因集。在无向网络中，模块内是高度相关的基因。在有向网络中，模块内是高度正相关的基因。
- 把基因聚类成模块后，可以对每个模块进行三个层次的分析：

1. 功能富集分析查看其功能特征是否与研究目的相符；
2. 模块与性状进行关联分析，找出与关注性状相关度最高的模块；
3. 模块与样本进行关联分析，找到样品特异高表达的模块。

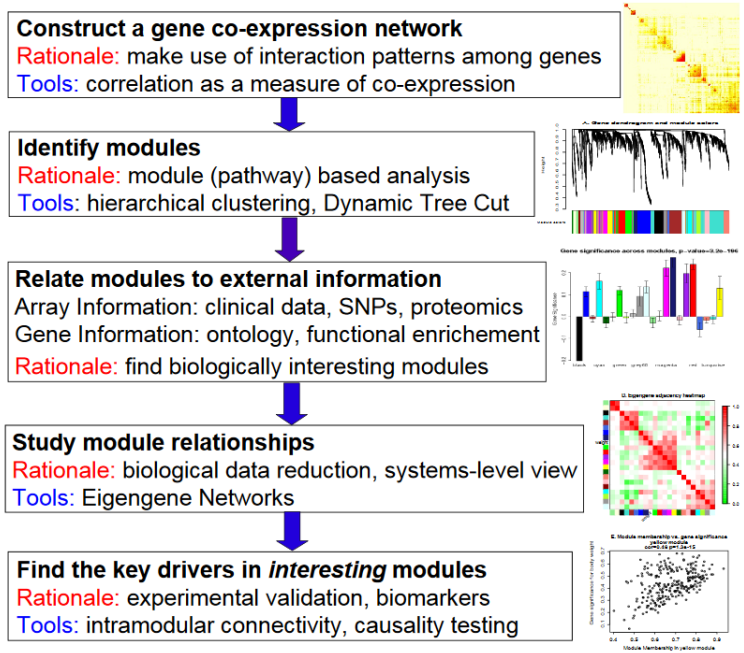
连接度，E，Hub gene

- Connectivity (连接度)：类似于网络中 “度” (degree)的概念。每个基因的连接度是与其相连的基因的边属性之和。
- Module eigengene E: 给定模型的第一主成分，代表整个模型的基因表达谱。这个是个很巧妙的梳理，我们之前讲过PCA分析的降维作用，之前主要是拿来做可视化，现在用到这个地方，很好的用一个向量代替了一个矩阵，方便后期计算
- Hub gene: 关键基因 (连接度最多或连接多个模块的基因)。

一文读懂PCA分析 （原理、算法、解释和可视化）

WGCNA分析的步骤

- 构建基因共表达网络：使用加权的表达相关性。
- 识别基因集：基于加权相关性，进行层级聚类分析，并根据设定标准切分聚类结果，获得不同的基因模块，用聚类树的分枝和不同颜色表示。
- 如果有表型信息，计算基因模块与表型的相关性，鉴定性状相关的模块。
- 研究模型之间的关系，从系统层面查看不同模型的互作网络。
- 从关键模型中选择感兴趣的驱动基因，或根据模型中已知基因的功能推测未知基因的功能。
- 导出TOM矩阵，绘制相关性图。



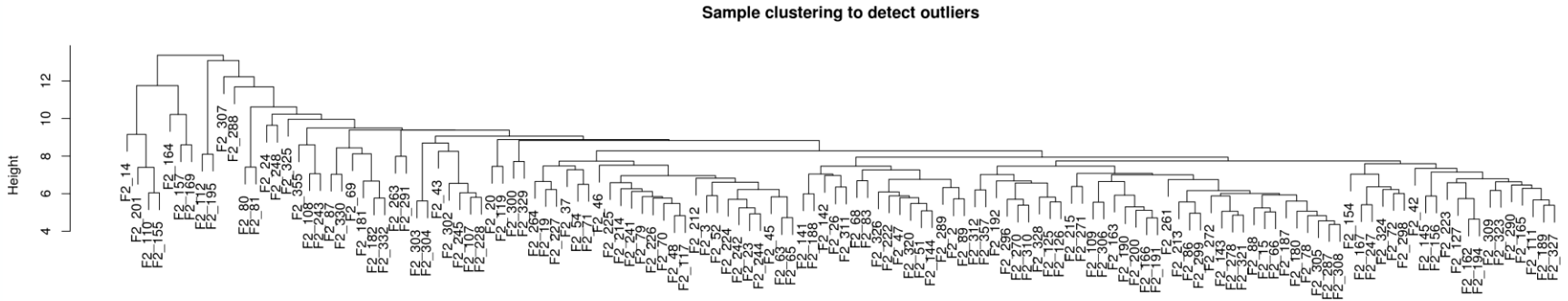
WGCNA输入数据和参数选择

- WGCNA本质是基于相关系数的网络分析方法，适用于多样品数据模式，一般要求样本数多于15个。样本数多于20时效果更好，样本越多，结果越稳定。
- 基因表达矩阵：常规表达矩阵即可，即基因在行，样品在列，进入分析前做一个转置。RPKM、FPKM或其它标准化方法影响不大，推荐使用Deseq2的varianceStabilizingTransformation或log2(x+1)对标准化后的数据做个转换。如果数据来自不同的批次，需要先移除批次效应（记得上次转录组培训课讲过如何操作)。如果数据存在系统偏移，需要做下quantile normalization。
- 性状矩阵：用于关联分析的性状必须是数值型特征（如下面示例中的Height, Weight，Diameter)。如果是区域或分类变量，需要转换为0-1矩阵的形式(1表示属于此组或有此属性，0表示不属于此组或无此属性，如样品分组信息WT，KO，OE)。

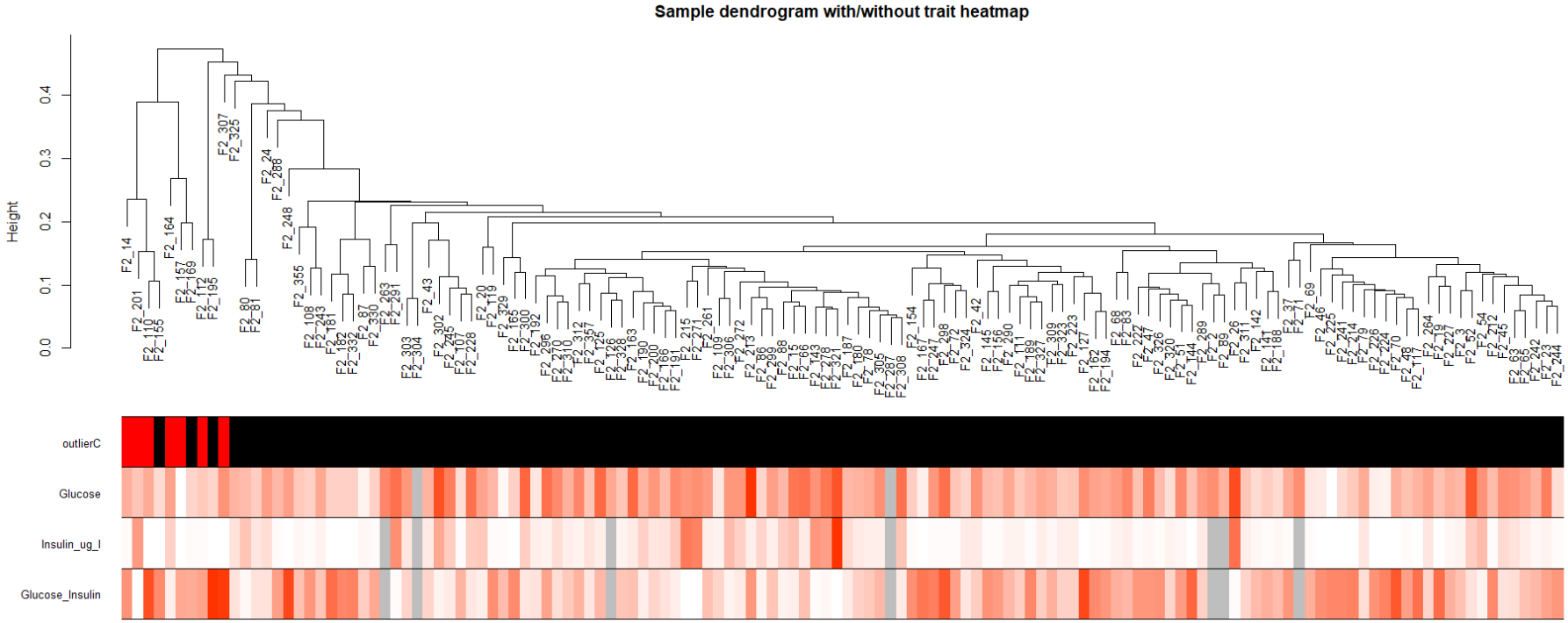
WGCNA输入数据和参数选择

- 推荐使用Signed network和Robust correlation (bicor)。(这个根据自己的需要，看看上面写的每个网络怎么计算的，更知道怎么选择)
- 无向网络在power小于15或有向网络power小于30内，没有一个power值可以使无标度网络图谱结构R^2达到0.8或平均连接度降到100以下，可能是由于部分样品与其他样品差别太大造成的。这可能由批次效应、样品异质性或实验条件对表达影响太大等造成，可以通过绘制样品聚类查看分组信息、关联批次信息、处理信息和有无异常样品 (可以使用之前讲过的热图简化，增加行或列属性)。如果这确实是由有意义的生物变化引起的，也可以使用后面程序中的经验power值。

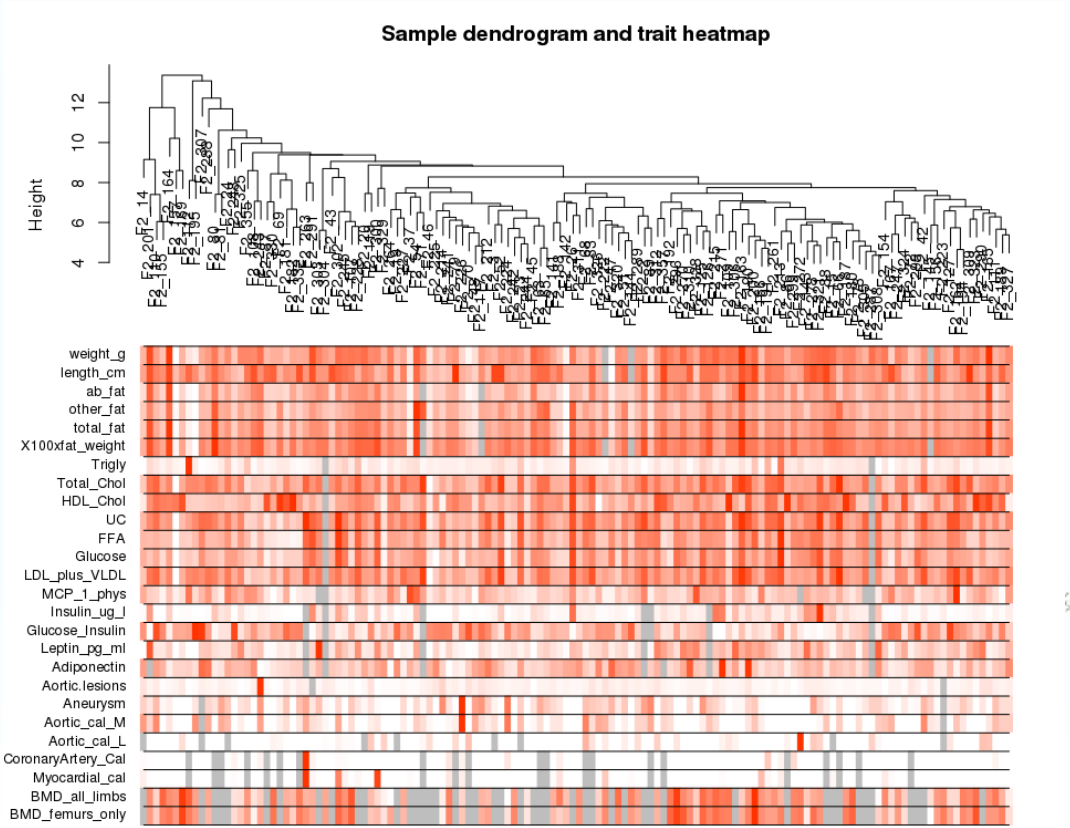
1. 样本聚类查看整体相关性



异常样品标记

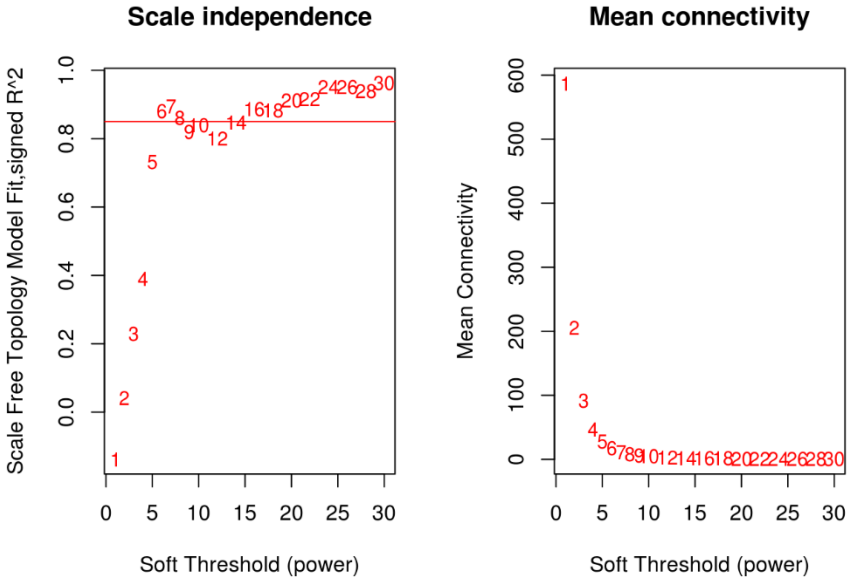


如果有表型信息，进行关联展示



2. 筛选合适的软阈值 (power)使得网络为无标度网络

- 横轴是 Soft threshold (power)，纵轴是无标度网络的评估参数，数值越高，网络越符合无标度特征 (non-scale)。
- 右图是连通性，Power越大，能连在一起的基因越少。

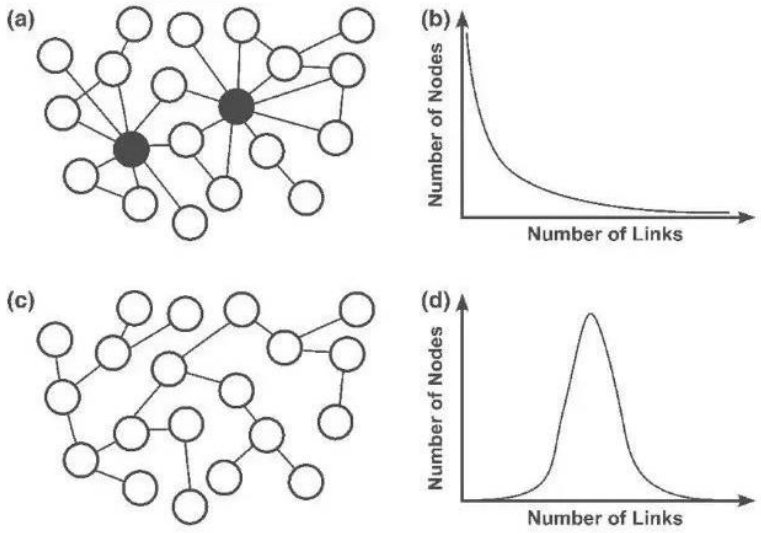


无标度网络

- 幂律分布广泛存在于物理学、生物学、社会学、经济学等众多领域中，也同样存在于复杂网络中。学者发现，对于许多现实世界中的复杂网络，如互联网、社会网络等，各节点拥有的连接数（**度 Degree**）服从幂律分布。也就是说，**大多数“普通”节点拥有很少的连接，而少数“热门”节点拥有极其多的连接。**这样的网络称作**无标度网络（Scale-free Network）**，网络中的“热门”节点称作**枢纽节点（Hub）**。

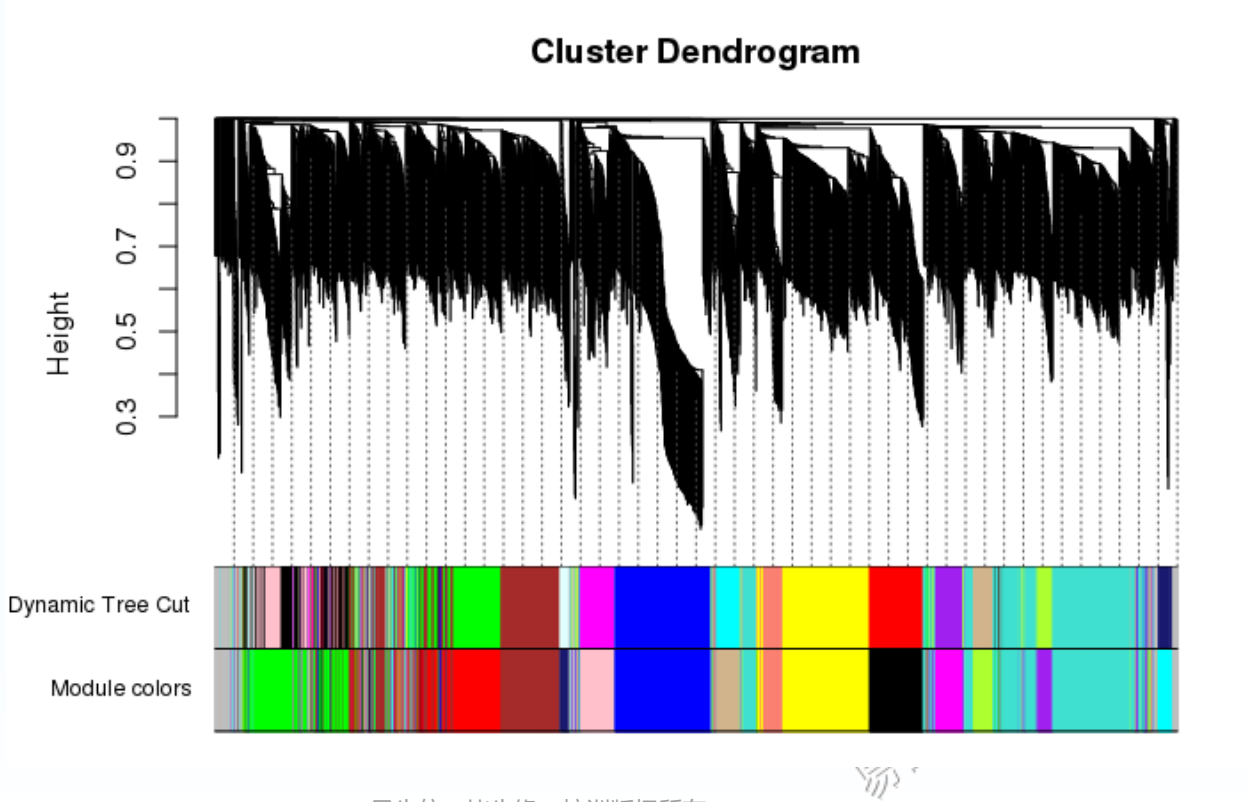
- 无标度网络是节点度分布（近似）为幂律分布的网络模型。如果用节点度概率分布 P(k) 表示网络中度为 k 的节点出现的频率，则有下面简洁的公式。

$$P(k) \sim k^{-\gamma}$$

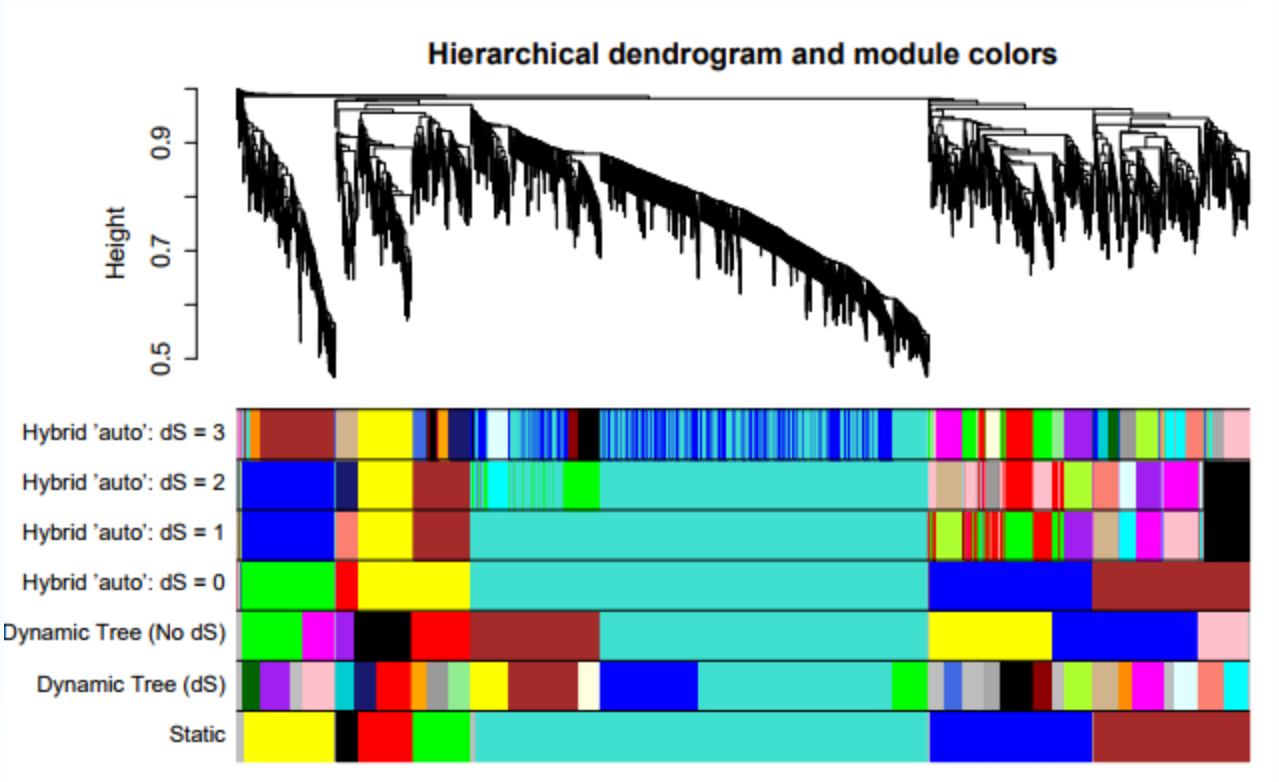


无标度网络

3. 鉴定共表达模块



dynamicTreeCut – deepsplit参数的意义



https://horvath.genetics.ucla.edu/html/CoexpressionNetwork/BranchCutting/Supplement.pdf

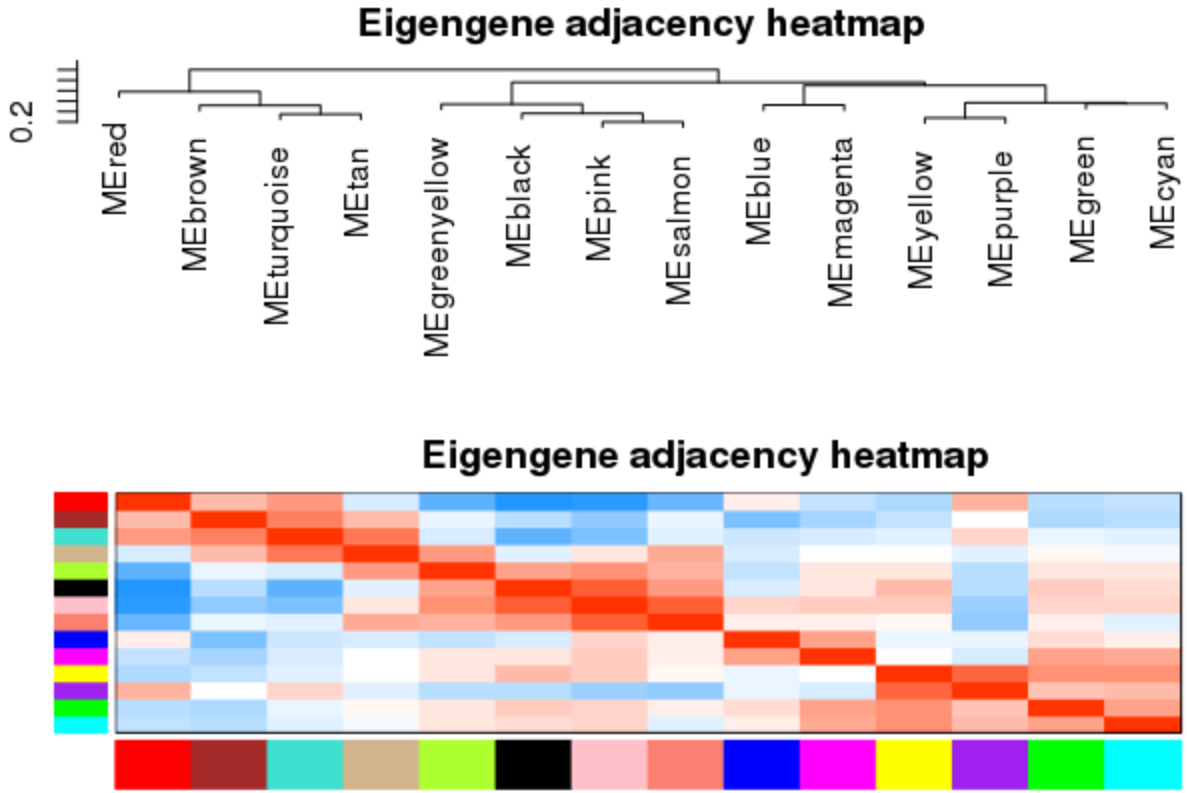
dynamicTreeCut – deepsplit参数的意义

- The value 0 is analogous to deepSplit = FALSE for the Dynamic Tree method and will produce relatively few, large and well-defined clusters. Values 1,2,3 will progressively produce a larger number of clusters that are allowed to exhibit larger core scatter and may be separated by smaller gaps, akin to setting deepSplit = TRUE in the Dynamic Tree variant. We recommend decreasing the minimum cluster size when using higher settings of deepSplit.

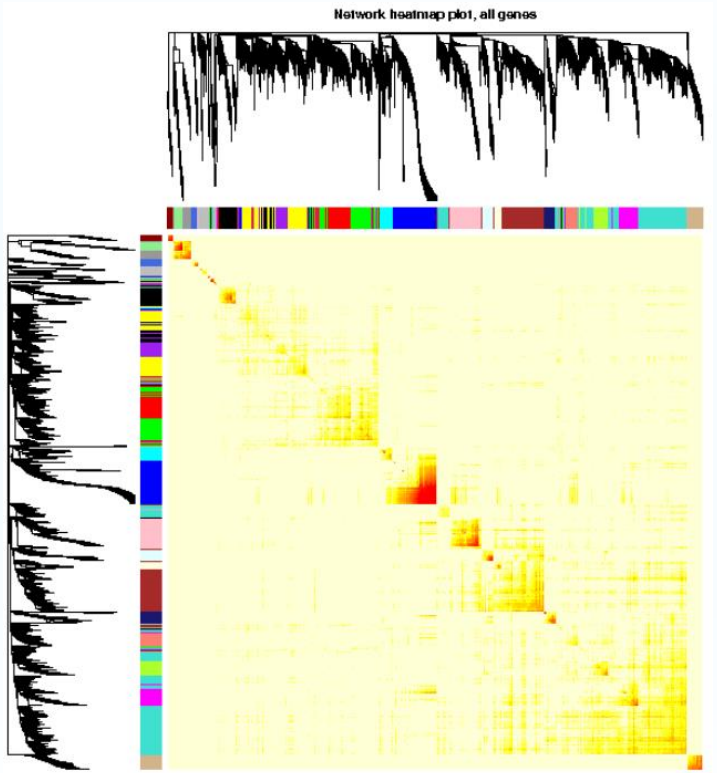
鉴定共表达模块的几个参数

- power: 上一步计算的软阈值
- maxBlockSize: 计算机能处理的最大模块的基因数量（默认5000)；4G内存电脑可处理8000-10000个，16G内存电脑可以处理2万个，32G内存电脑可以处理3万个。计算资源允许的情况下最好放在一个block里面。
- saveTOMs：存储TOM矩阵，这是最耗费时间的计算，存储起来，供后续使用
- mergeCutHeight: 合并模块的阈值，越大模块越少；越小模块越多，冗余度越大；一般在0.15-0.3之间
- loadTOMs: 避免重复计算

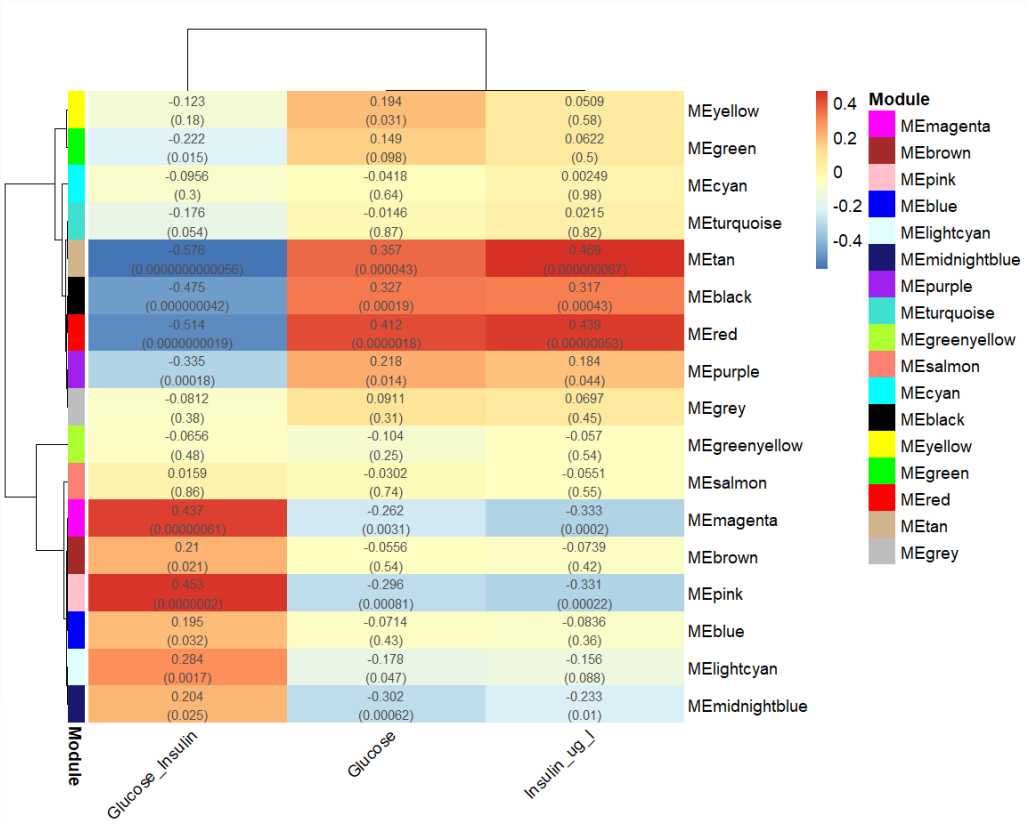
4. 模块之间的相似性



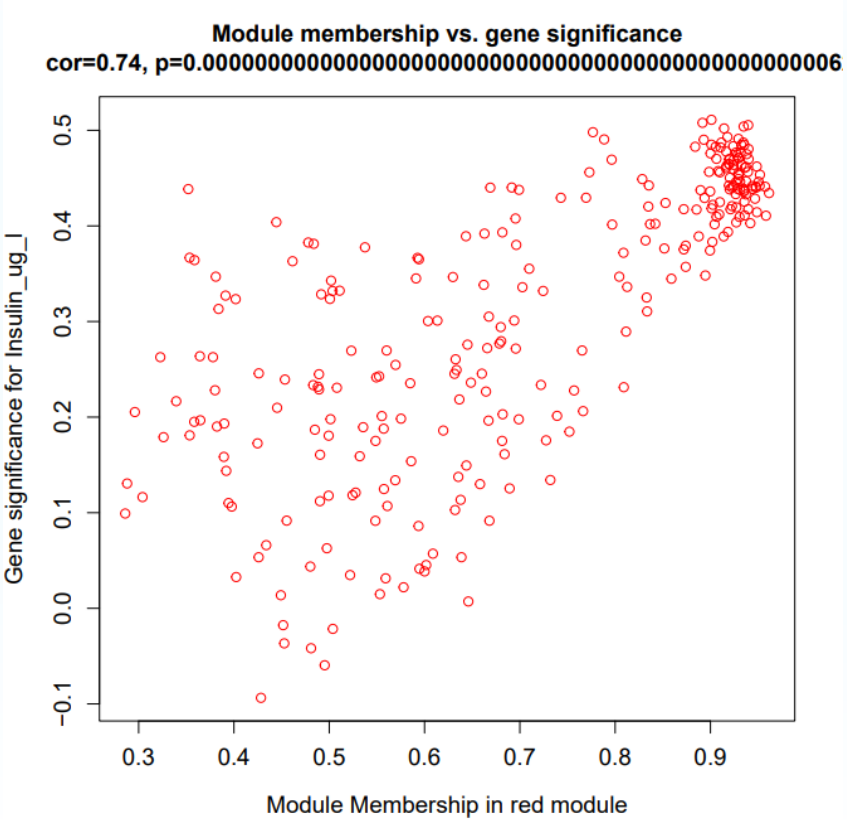
5. TOM矩阵聚类图



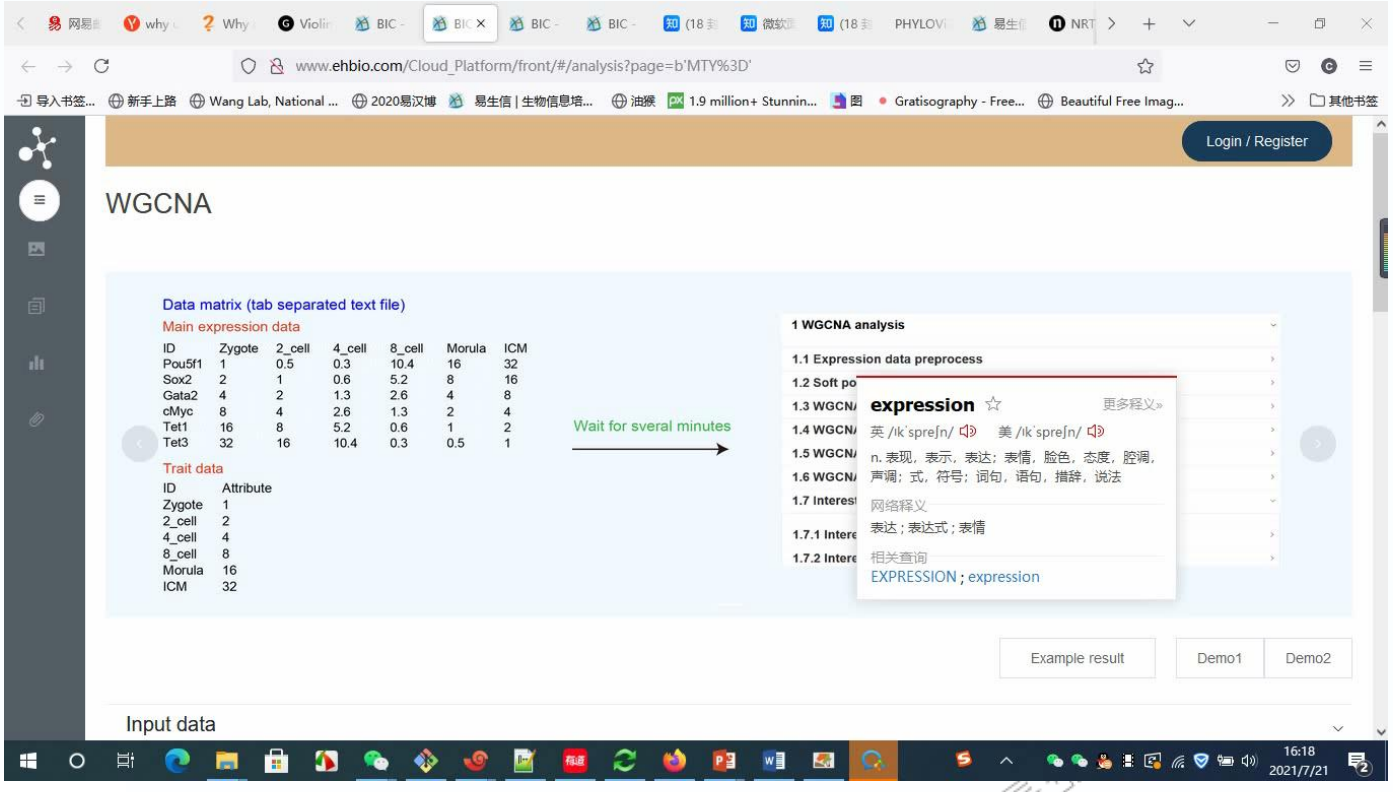
6. 关联表型数据



7. 筛选与模块和表型都相关的基因

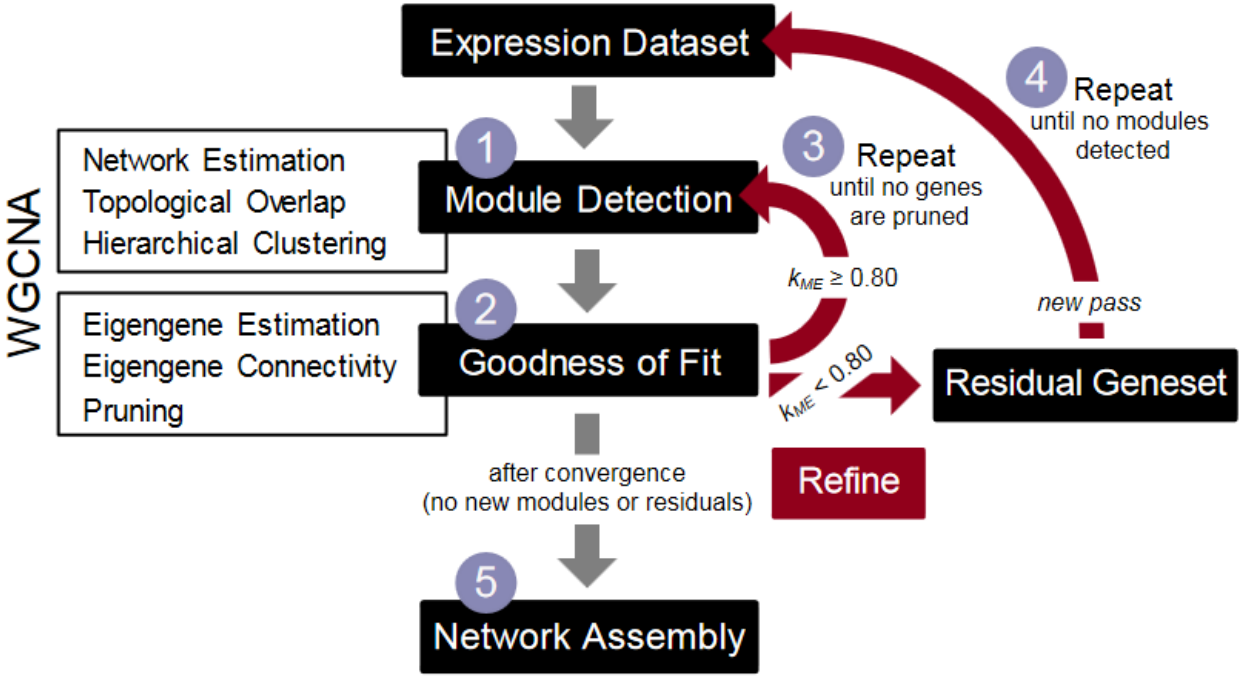


在线WGCNA分析



https://www.bic.ac.cn/BIC/#/analysis?page=b%27MTY%3D%27

迭代WGCNA



https://www.biorxiv.org/content/10.1101/234062v1.full.pdf

Sequencing costs a lot and gains more

长按关注 生信宝典 ， 简单入门，快速晋级