**百迈客WGCNA分析结题报告**

**摘要**

利用hdWGCNA ,对 个样本进行WGCNA分析，绘制基因共表达调控网络。

**1、背景介绍**

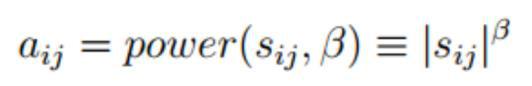
加权基因共表达网络分析（Weighted Gene Co-expression Network Analysis, WGCNA）是构建基因共表达网络的常用方法。基因共表达网络，我们定义每个节点为一个基因，在不同样本中存在表达共性的基因处于同一个基因网络，而基因之间的共表达关系一般由它们之间的表达相关系数衡量。WGCNA[1]算法首先假定基因网络服从无尺度分布,并定义基因共表达相关矩阵、基因网络形成的邻接函数,然后计算不同节点的相异系数,并据此构建分层聚类树(hierarchical clustering tree),该聚类树的不同分支代表不同的基因模块(module),模块内基因共表达程度高,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,最终达到鉴定基因网络的目的。

单细胞测序技术可以揭示特定肿瘤组织中的细胞特异性，对细胞进行分类，并且识别特定的标志物，但其检测的细胞数量和病例来源都是有限的。利用WGCNA分析单细胞转录组测序数据，可以提供一套有别于高変基因、差异分析的方法，不依赖于数据库直接用表达量的相关性值预测调控关系，筛选某些细胞亚群中有关联作用的基因集（称为模块），可以从成千上万的基因中挑选出高度相关的基因的模块，并将模块与表型进行关联，寻找marker gene或治疗靶点。

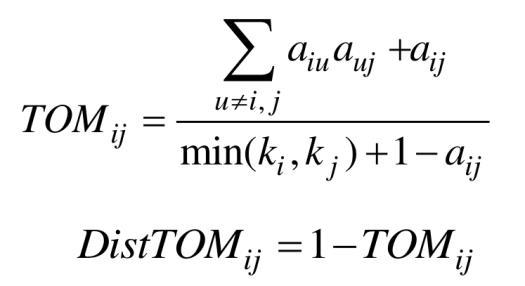
hdWGCNA[2,3]是一个专门针对单细胞转录组或空间转录组学等高维数据进行WGCNA的R包，又被称为scWGCNA，可以与Seurat对象直接兼容，能以细胞类型的特异表型构建共表达网络，识别相互连接基因的关键模块。

hdWGCNA的分析原理：

1）基因共表达相关矩阵的定义。基因共表达相关矩阵中的元素为基因之间的两两相关系数，即每对基因 i 和基因 j 的相关系数为 Sij，据此构成了基因共表达相关矩阵 S=[Sij]。其中 Sij 公式如下：

2）定义邻接函数。最直接的邻接函数通过指定基因之间的相关系数的阈值（如R=0.85）将基因对划分为相关和不相关的，但是这种方法会损失大量的信息。因此 WGCNA 用了幂指数邻接函数，用 aij 作为衡量他们关系的指标，其中 aij 公式如下：

3）确定邻接函数的参数β。根据无尺度网络的原则来确定合适的soft power阈值（β值），以减少相关性矩阵中存在的背景噪音，从而保留强连接并去除弱连接。

4）节点间的相异度衡量(Topological overlap dissimilarity measure,TOM)。确定了β值之后，矩阵 S 转换成邻接矩阵 A=[aij]。之后将某个基因和分析中其他所有基因之间的关系纳入考虑，邻接矩阵转换成拓扑矩阵Ω=ωij。 当基因 i 和 j 之间无连接，并且没有任何其它基因将这两个基因连接的情况下这个值为 0。节点的相异程度用来衡量，这个值是网络构建的基础。

5）聚类分析鉴定基因模块（module）。利用相异系数为基本元素构建分层聚类树，聚类树的不同分支代表不同的基因模块。

1. **分析内容**

在进行WGCNA分析之前，首先对样本的基因集和细胞进行过滤，提高共表达网络的精准度：首先选择至少在5%的细胞中表达的基因，然后通过KNN（k-Nearest Neighbors）算法进行聚类，将样本中的细胞聚类成相似的细胞组，计算这些细胞的平均表达水平，降低表达矩阵的稀疏性。

**2.1 构建邻接矩阵**

利用TestSoftPowers函数将相关性矩阵转换成邻接矩阵，通过迭代确定最适的soft power值，使该网络最接近无尺度网络，并尽可能保留连通性信息,本项目最适soft power值为\_。

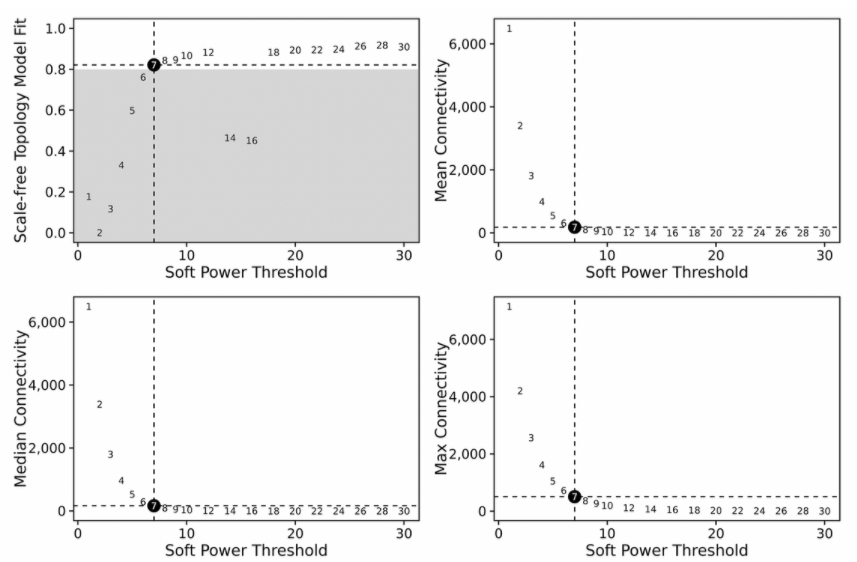


图1 参数soft power值曲线图

左图横坐标为soft power值（取值默认是1到30），纵坐标为无尺度拓扑模型拟合值，对应网络中log(k)与log(p(k))相关系数的平方，值越高表明该网络越接近无网络尺度的分布，一般选取拟合值大于等于0.8的最低soft power值；右图横坐标为soft power值，纵坐标代表对应的基因模块中所有基因邻接函数的均值。

**2.2 模块划分**

利用TOM根据基因间表达量的相关性构建聚类树，对基因进行聚类，构建拓扑重叠矩阵，并划分模块。

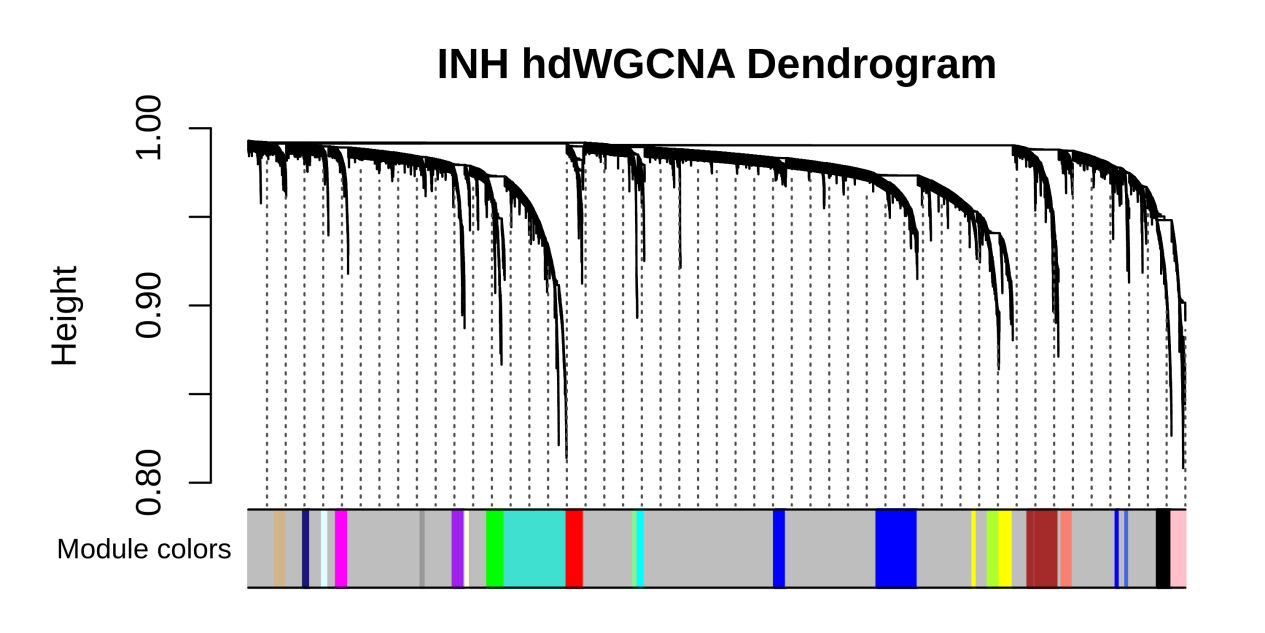


图2 基因共表达网络图

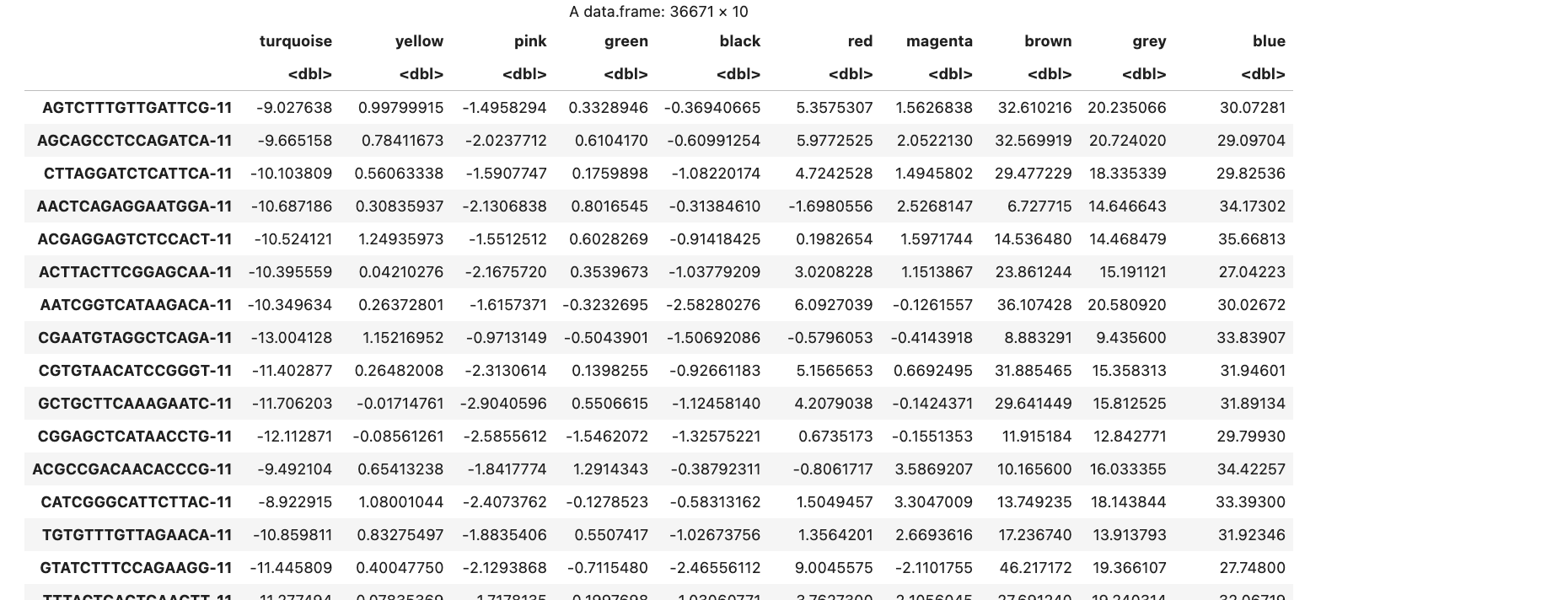
基因网络模块图共分三部分，第一部分为基于 TOM 绘制的基因系统聚类树，每条线代表一个基因，相似的基因被聚到一个分支，纵向距离代表两个基因间的距离；第二部分为对应第一部分结果，使用动态规划算法将基因划分模块的结果，不同模块用不同颜色表示，同一模块的基因通常据有类似的功能，灰色表示无法归入任何一个模块的基因（后续不对灰色模块进行分析）；第三部分为对第二部分结果进行优化合并，获得的最终模块划分结果。

**2.3 模块关键基因分析**

同一模块的基因通常具有类似的功能，利用降维对模块的特征基因进行分析，并基于特征基因计算每个基因的联通性，找到模块中高度连接的核心基因-hub gene，并进行可视化展示，为寻找marker gene或治疗靶点提供依据。

**2.3.1 模块特征基因筛选**

模块特征基因（Module Eigengenes，MEs）是一个用于分析共表达模块的基因表达谱特征的常用指标，通过对包含每个模块的基因表达矩阵子集进行主成分分析（Principal Component Analysis ，PCA），将每个模块PCA矩阵中的第一个主成成分做为MEs,并利用MEs进行相关性分析。MEs矩阵如下表所示，其中每一行为基因名称，每一列为模块：



注：由于表格较大，不在结果文件中展示。

基于MEs矩阵，绘制模块间相关性热图，筛选显著相关、可能功能相似的模块。

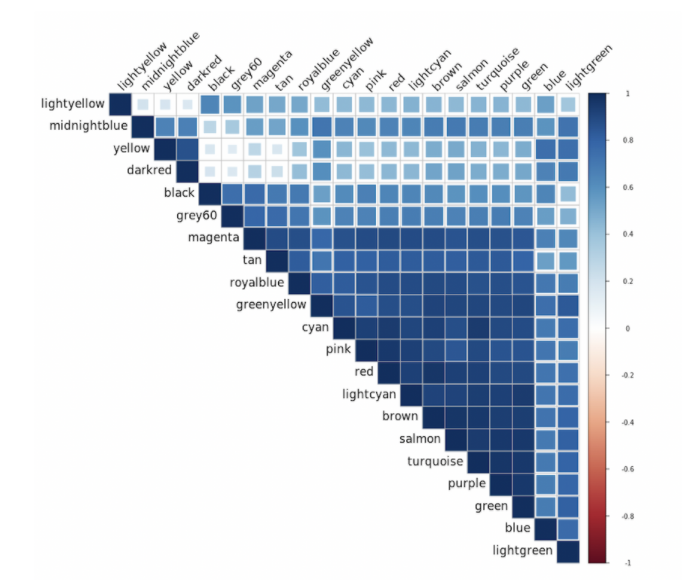


图3 模块-模块相关性热图（颜色与图3改成一致）

行和列均代表模块，负相关性越强呈现蓝色，正相关性越强呈现红色。

基于MEs矩阵，绘制模块与性状的相关性热图，筛选与性状显著相关的特定模块。

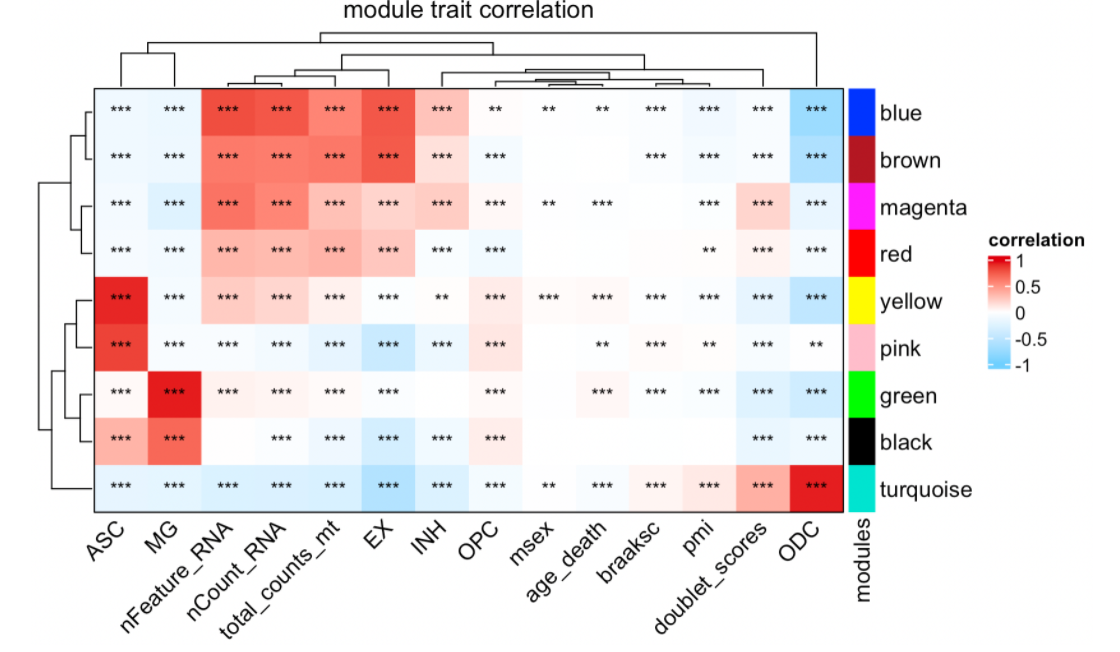


图4 模块-性状相关性热图

每一行代表模块，每一列代表性状（细胞类型、RNA count数等）。方块颜色表示相关性大小，负相关性越强呈现蓝色，正相关性越强呈现红色；方块内的\*为模块与性状相关性系数的Pvalue，\*\*\*表示Pvalue＜0.001，\*\*表示Pvalue＜0.01，\*表示Pvalue＜0.05；方块颜色，数值越大显著性越强。

**2.3.2 模块核心基因鉴定**

针对每个模块进一步挖掘核心基因，利用ModuleConnectivity函数，通过计算kME值来评估基因间有效连通性，默认选取kME值top100的基因作为该模块的核心基因，代表整个模块的表达趋势。

表2 核心基因在模块内kME值统计表

第一列为基因ID，第二列为所属模块，第三列为kME值。

按模块基因的kME值进行排序，将每个模块基因的连通性进行可视化。

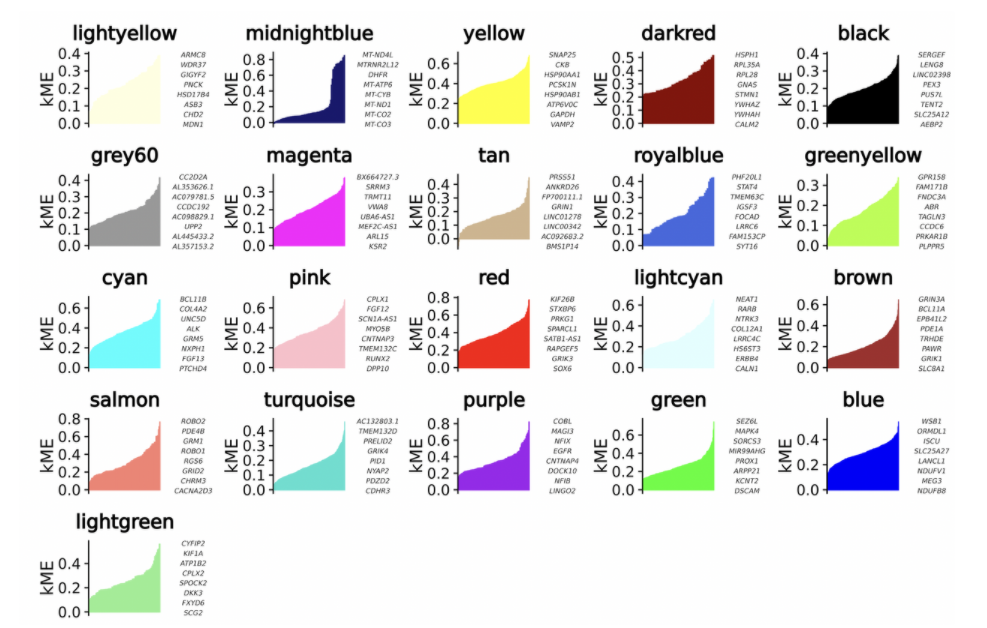


图5 模块基因kME值趋势分布图

子图标题为模块名称，横坐标为按kME值排序的模块基因，纵坐标为基因对应的kME值，每个子图右侧展示kME值为top8的核心基因。

计算每个细胞中每个模块基因的平均表达水平，绘制基因在细胞亚群中的UMAP分布图。

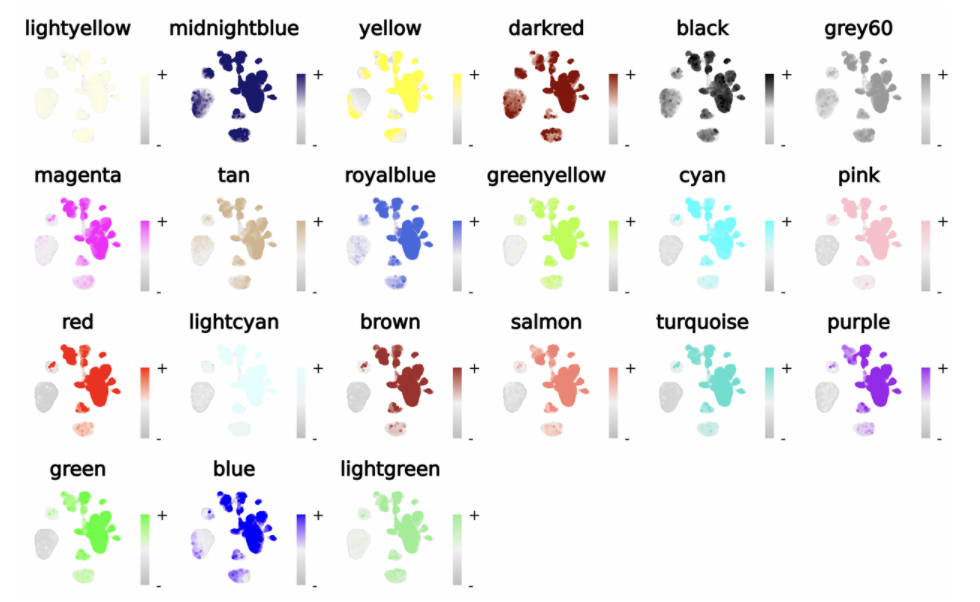


图6 模块基因表达水平UMAP分布图

子图标题为模块名称，图中每个点代表一个细胞，点颜色越深，hub基因在特定细胞中平均表达水平表越高。

利用UMAP对拓扑重叠矩阵TOM进行降维聚类，筛选每个模块top10的核心基因，绘制模块核心基因的UMAP分布图，UMAP的空间分布取决于核心基因的连通性。

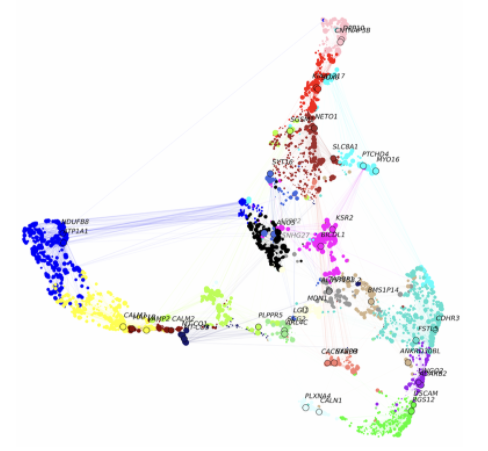


图7 所有模块核心基因连通性UMAP分布图

一个节点代表一个基因，不同节点颜色代表不同，节点大小代表基因kME值大小，图中展示每个模块top2核心基因名称；点之间的连线代表调控网络中两个基因的共表达关系，模块内基因连线的颜色与节点颜色一致，模块间基因的连线为灰色；连线颜色越深表示基因间的相关性越强，图中仅展示相关性前20%的连线。

挑选每个模块top3的核心基因，并随机选择6个其它基因，利用力导向算法（force-directed graph drawing algorithm），绘制模块间核心基因的共表达网络调控图，将模块间的核心基因调控关系进行可视化。

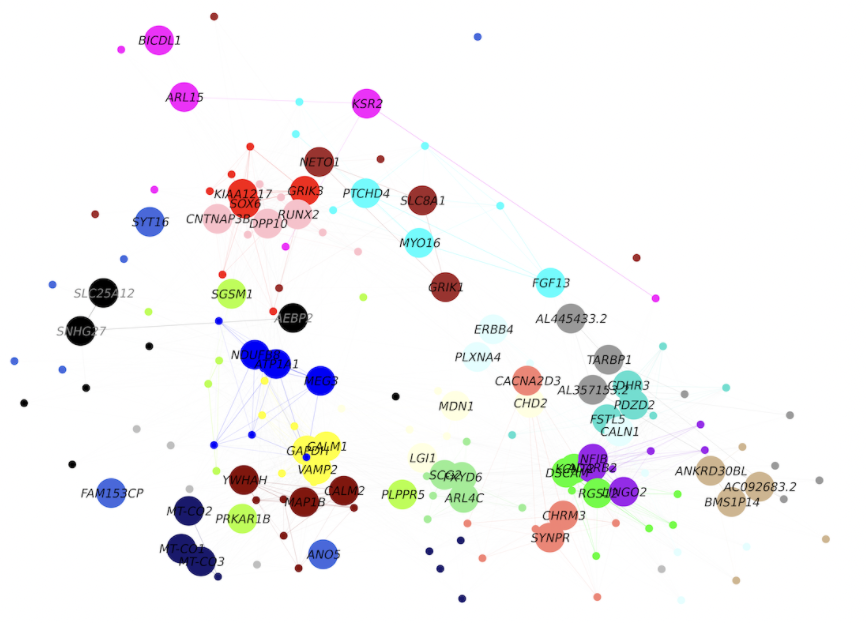


图8 所有模块核心基因共表达调控网络图

一个节点代表一个基因，节点的不同颜色代表不同模块模块；点之间的连线代表调控网络中两个基因的共表达关系，模块内基因连线的颜色与节点颜色一致，模块间基因的连线为灰色；连线颜色越深表示基因间的相关性越强。

利用每个模块内核心基因的kME平均值绘制小提琴图，展示核心基因kME值在不同性状中的分布特征。

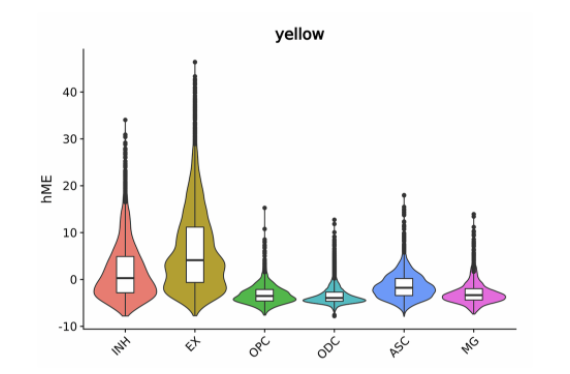


图9 模块核心基因在不同性状中的小提琴分布图

图的标题为模块名称，横坐标为性状，纵坐标为模块内每个细胞的核心基因kME平均值；符合目标基因ME值的细胞分布越多，外缘曲线越宽。

挑选每个模块top25的核心基因，绘制基因表达网络调控图，展示每个模块中核心基因的相互调控关系。

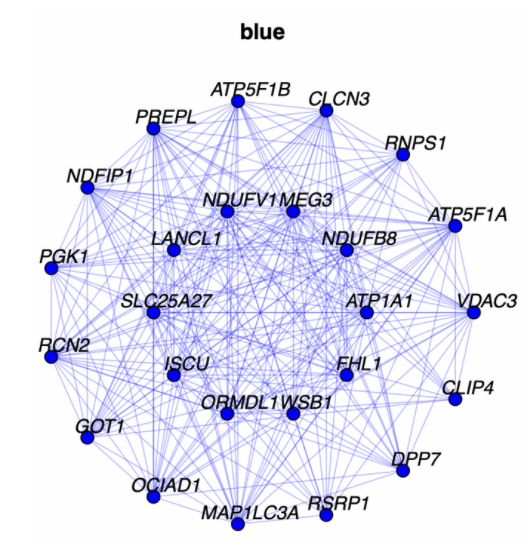


图10 模块top25核心基因共表达调控网络图

图的标题为模块名称，每个节点代表一个基因，点之间的连线代表调控网络中两个基因的共表达关系，颜色越深表示相关性越强；kME值top10的核心基因网络图中心，其余15个基因位于网络图外圈。

**2.4 模块核心基因功能富集分析**

**3、概念解释**

1）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冥次运算，冥次的值也就是软阈值（soft power）, TestSoftPowers这个函数所做的就是确定合适的soft power，通过强化强相关，弱化弱相关或负相关，使得相关性数值更符合无标度网络特征，更具有生物意义。

2）Module(模块)：高度关连的基因集。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：功能富集分析查看其功能特征是否与研究目的相符；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；模块与样本进行关联分析，找到样品特异高表达的模块。

3）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。

4）Module eigengene（模块特征基因）: 对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，可代表该模块内基因的整体表达模式。

5）Intramodular connectivity: 给定基因与给定模型内其他基因的关联度，判断基因所属关系。

6）Module membership: 给定基因表达谱与给定模型的eigengene的相关性。

7）Hub gene:核心基因，连接度最多或连接多个模块的基因。

8）Adjacency matrix (邻接矩阵)：基因和基因之间的加权相关性值构成的矩阵。

8）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。

**4、结果文件说明**

|-- netcolor2gene.xls 基因与模块对应结果表格

|-- Module\_Correlation.{pdf，png} 模块-样本/细胞类型相关性热图

|-- hubGeneHeatmap.{pdf，png} 模块间相关性热图

|-- networkHeatmap.{pdf，png} 基因共表达网络热图

|-- Heatmap 模块内基因热图文件夹

|--|-- \*\_heatmap.{pdf，png} 模块基因表达模式图

|-- kME 基于 kME 值确定模块成员文件夹

|--|-- kME\*\_kME.txt 模块 \*的成员及 kME 值

|-- All\_gene\_module\_kME.txt

|-- Cytoscape 可视化程序 Cytoscape 的分模块输入文件

|--|-- CytoscapeInput-edges-kME\*.txt 模块\*基因用于 Cytoscape 程序输入的“边”文件

|--|-- CytoscapeInput-nodes-kME\*.txt 模块\*基因用于 Cytoscape 程序输入的“点”文件

|-- Enrichment 模块内基因功能富集分析目录

|--|-- \* \*模块的富集目录

|--|--|-- \*.annotation.xls \*模块基因的注释结果整合文件

|--|--|-- \*.nr.anno.txt \*模块基因 nr 注释结果

|--|--|--go\_enrichment \*模块 Go 富集分析目录

|--|--|-- Graph \*模块topGo结果及 kegg 分类图

|--|--|-- pathway \*模块的 kegg 富集分析结果

hdWGCNA\_results/

├── step1\_co\_expression\_network 共表达调控网络分析

│   ├── TestSoftPowers.{pdf,png} 参数soft power值曲线图

│   ├── Dendrogram.{pdf,png} 基因共表达网络图

├── step2\_module\_plots 模块分析

│ ├── Correlogram.{pdf,png} 模块-模块相关性热图

│   ├── module\_trait\_correlation.{pdf,png} 模块-性状相关性热图

│   ├── hub\_df\_top100.xls 核心基因在模块内kME值统计表

│   ├── kME{pdf,png} 模块基因kME值趋势分布图

│   ├── kME\_umap.{pdf,png} 模块基因表达水平UMAP分布图

│   ├── UMAP\_hub.pdf.{pdf,png} 所有模块核心基因连通性UMAP分布图

│   ├── HubGeneNetworkPlot.{pdf,png} 所有模块核心基因共表达调控网络图

│   ├── vlnplot.{pdf,png} 模块核心基因在不同性状中的小提琴分布图

│   ├── ModuleNetworkPlot.{pdf,png} 模块top25核心基因共表达调控网络图

├──step3\_Enrichment 功能富集分析

│   ├── \* \*模块的富集目录

| │   ├── \*.annotation.xls \*模块基因的注释结果整合文件

| │   ├── \*.nr.anno.txt \*模块基因 nr 注释结果

| │   ├── go\_enrichment \*模块 Go 富集分析目录

| │   ├── Graph \*模块topGo结果及 kegg 分类图

| │   ├── pathway \*模块的 kegg 富集分析结果

**5、参考文献**

[1]Langfelder, Peter, and Steve Horvath. “WGCNA: an R package for weighted correlation network analysis.” BMC bioinformatics vol. 9 559. 29 Dec. 2008, doi:10.1186/1471-2105-9-559

[2]https://smorabit.github.io/hdWGCNA/index.html

[3]Morabito, Samuel et al. “Single-nucleus chromatin accessibility and transcriptomic characterization of Alzheimer's disease.” Nature genetics vol. 53,8 (2021): 1143-1155. doi:10.1038/s41588-021-00894-z