基于多重假设检验与共表达网络的新型抗癌药物研究模型

摘要

靶向治疗是治疗肿瘤疾病的一种重要方法，它具有针对性强、疗效显著等特点。现有的靶向药物通常针对特定的基因突变靶点，容易出现耐药性。目前，一种由癌症诱发的血管新生作为靶点的靶向药物研究正成为该领域研究的热点。本文针对给定的药物影响血管新生基因表达的数据，建立了基于多重假设检验与共表达网络的新型抗癌药物研究模型。

针对问题 1，本文基于 t-2 检验与$Wilcoxon$秩和检验，对给定数据进行处理后，绘制火山图用以刻画不同基因之间的差异性关系，并给出相关参数估计，其中 PValue 参数为核心参数，使用其量化不同基因间的差异。

针对问题 2，本文使用网络建模方法来刻画基因表达的相互关系，使用给定的样本建立无向图模型，调整基因权重后设置阈值，用以建立相关性邻接矩阵，评估基因的协同作用。

针对问题 3，本文利用问题 2 中的相关结论，基于 FDR 校验修正的结果，经过两轮筛选，选择出 50 个可能的血管新生敏感基因。

经检验，模型在一般基因处理 RNA-seq 上具有一定的泛化能力与可行性，适用于一般生物医学信息处理领域中的参数估计、假设检验与基因信息提取问题。构建敏感基因筛选模型需要根据具体数据的特点和问题进行参数设置和调整。阈值的选择可以根据相关性的分布和网络的稀疏性进行调整。模型的推广与优化均保有较大空间。

关键词：t-2 检验、$Wilcoxon$秩和检验、共表达网络、邻接矩阵、FDR 校验

一、 问题重述

靶向治疗是治疗肿瘤疾病的一种重要方法，它具有针对性强、疗效显著等特点。现有的靶向药物通常针对特定的基因突变靶点，容易出现耐药性。目前，一种由癌症诱发的血管新生作为靶点的靶向药物研究正成为该领域研究的热点。

研究人员为研究某类药物队血管新生的作用，进行了如下对照实验：对某种动物使用药物 A 诱导其血管新生，加入药物 B 作用后发现其具有逆转 A 造成的血管新生作用（先加入药物 A，在其作用结束并清洗后，再加入药物 B），而药物 B 的结构类似物 C 对试验动物有明显的血管新生抑制作用。在对四组样品（正常对照组、加药物 A 组、加药物 B 组和加药物 C）适当处理（包括充分的培养时间和药液清洗）后，进行 RNA-seq 测序。本研究希望通过比对正常对照组（没有添加任何药物）、药物 A 添加组、药物 B 添加组和药物 C 添加组的基因表示，研究药物 A 诱导血管新生作用、药物 B 血管新生逆转作用和药物 C 对血管新生的抑制作用机理。

实验共获得了大量数据，包括基因 ID，2 个 Cont 对照组（Cont-1_count_fpkm 和 Cont-2_count_fpkm，对未添加任何药物样本测序，并计算基因表达量 FPKM）；1 个添加药物 A 组（A-1_count_fpkm，直接添加含药物 A 的培养液，经过足够长时间培养后对样本测序，并计算基因表达量 FPKM）；2 个添加药物 B 组（B-1_count_fpkm 和 B-2_count_fpkm，该实验是在添加含药物 A 培养液，经过适当时间，诱导血管新生后，洗去药液，再加入含药物 B 的培养液，经过足够长时间培养后对样本测序，并计算基因表达量 FPKM）；2 个添加药物 C 组（C-1_count_fpkm 和 C-2_count_fpkm，直接添加含药物 C 的培养液，经过足够长时间培养后对样本测序，并计算基因表达量 FPKM）。针对这些数据，需要解决的问题如下：

1.建立基因表达差异的显著性检验模型，并进行相关参数估计。因费用问题实际采集的样本较少，给出提高小样本显著性检验精度的方法；

2.在研究基因表达显著性差异时，一般假设基因表达是独立的。但事实上，生物学功能基因组的表达水平往往具有协同调节特点，请建立数学模型刻画基因表达的协同调节作用，并对模型的合理性进行评价；

3. 请建立模型，寻找与血管新生直接关联的基因。现有的方法是对表达显著性差异的基因利用 FDR校正以克服检验误差，但这样得到的基因数目通常还有数千个，请结合问题2模型，利用生物学功能基因组协同调节的特点减少敏感基因数目，并针对附

件中数据在论文中给出50个最敏感基因。

二、 问题分析

针对问题 1，对于基因表达差异的显著性检验可以使用 t 检验或者 DESeq2 等差异表达分析方法。这些方法可以帮助确定基因在不同处理组之间表达差异的显著性。同时，针对样本容量较小的特点，可以选择非参数方法进行显著性检验，如 Wilcoxon 秩和检验或 Mann-Whitney U 检验。非参数方法不依赖于总体分布的假设，通常在小样本情况下更具有鲁棒性。此外，可采用交叉验证、引导等方法以适应小样本数据特点。

针对问题 2，研究基因表达的协同调节作用时，可以使用网络建模方法来刻画基因表达的相互关系。如 WGCNA（全称为 weighted gene co-expression network analysis），即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，在研究表型性状与基因关联分析等方面的研究中被广泛应用。

针对问题 3，对显著性差异的基因进行进行 FDR（False Discovery Rate）校正后，可以控制多重假设检验的错误率，并筛选出显著基因。在筛选后，根据已知的基因调控关系或使用相关算法（如权重共享网络模型、因果推理方法等），建立基因调控网络模型，用以描述基因之间的协同调节作用，并识别出与血管新生直接关联的基因。通过计算节点的中心性指标（如节点度、介数中心性等）来评估基因的重要性，从而进一步具有较高中心性指标的基因，这些基因往往在调控网络中扮演重要角色，与血管新生直接关联的概率较大，可以作为敏感基因。

三、 模型假设

1.假设血管新生仅与给定数据中的基因有关；

2.假设本问题中基因控制的性状不受其他无关因素影响，如环境因素等；

3.假设无其他因素影响问题中网络图的建立。

四、 模型建立

4.1 基于 t-2 检验与$Wilcoxon$秩和检验的差异显著性检验模型

4.1.1 简介

（一）t-2 检验

t 检验用于检测小样本的平均值差异程度，通过 t 分布理论判断差异发生的概率，从而判断两个样本的平均数的差异是否显著。在使用独立样本 t 检验之前，需要满足如下三个前提条件：

（1） 每个样本的观察值必须是独立的；

（2） 两个样本对应的总体必须服从正态分布；

（3） 两个样本对应的总体必须有相同的方差。

为了判断样本是否满足条件（2）（3），需要先进行正态分布检验和方差齐性检验。若样本满足上述三个条件，接下来可进行独立样本 t 检验：

（1） 建立零假设H_0，确定显著性水平

独立样本 t 检验用于检验两个独立样本是否来自具有相同均值的总体，即检验两个正态分布总体的均值是否相等。若用μ_1,μ_2表示两个独立样本总体的平均值，则应做出如下假设：

$$H_0: \mu_1 - \mu_2 = 0 \tag{1}$$

同时，将显著性水平记为α。

（2） 计算检验统计量

由于两总体方差相等，即$\sigma_1^2 = \sigma_2^2$，两样本 t 检验的检验统计量可按照单样本 t 检验统计量公式进行计算。其检验统计量为：

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}} \tag{2}$$

自由度ν的值为：

$$\nu = n_1 + n_2 - 2 \tag{3}$$

其中，$\bar{x}$和$\bar{y}$分别代表两个样本的均值，s_1和s_2分别代表两个样本的标准差，n_1和n_2代表样本量。

若两样本总体服从正态分布，但方差不相等，可使用近似 t 检验方法，校正自由度。其检验统计量t'和自由度ν的计算公式如下：

$$t' = t$$
$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}} \tag{4}$$

（3） 求 p 值，与α比较，得出结论

得到检验统计量 t 的值后，可以确定样本均值之间是否存在显著差异。然而 t 值衡量的是样本均值与期望的差异相对于标准误差的大小，仅仅依靠 t 值无法得知这个差异是否显著，因为差异的大小与样本大小、数据的变异性等因素都有关。因此，需要根据 t 值得到对应的 p 值进一步判断。

通过 t 值获取对应 p 值涉及 t 分布累积积分函数的计算，计算过程相对复杂，通常需要使用软件或者计算工具进行。

P 值表示在零假设为真的情况下，统计量等于或更极端于实际观察值的概率。

如果$p \le \alpha$，则 p 值小于等于显著性水平，有足够的证据拒绝零假设，即结果是显著的；

如果$p > \alpha$，则 p 值大于显著性水平，没有足够证据拒绝零假设，即结果是不显著的。

（二）$Wilcoxon$秩和检验

$Wilcoxon$秩和检验为非参数方法，不依赖于总体分布的假设，通常在小样本情况下更具有鲁棒性。

基于样本数据秩和。先将两样本看成是单一样本(混合样本)，然后由小到大排列观察值统一编秩。如果原假设两个独立样本来自相同的总体为真，那么秩将大约均匀分布在两个样本中，即小的、中等的、大的秩值应该大约均匀被分在两个样本中。如果备选假设两个独立样本来自不相同的总体为真，那么其中一个样本将会有更多的小秩值，这样就得到一个较小秩和；另一个样本将会有更多的大秩值，因此就会得到一个较大的秩和。

设两个独立样本为：第一个x的样本容量为n_1，第二个y样本容量为n_2，在容量为$n = n_1 + n_2$的混合样本(第一个和第二个)中，x样本的秩和为W_x, y样本的秩和为W_y, 且有

$$W_x + W_y = 1 + 2 + \cdots + n = \frac{n(n+1)}{2} \tag{5}$$

我们定义

$$W_1 = Wx - n_1(n_1 + 1)/2 \tag{6}$$

$$W_2 = Wy - n_2(n_2 + 1)/2 \tag{7}$$

以x样本为例，若它们在混合样本中享有最小的n_1个秩，于是$Wx = n_1(n_1 + 1)/2$, 也是Wx可能取的最小值；同样Wy可能取的最小值为$n_2(n_2 + 1)/2$。

那么，Wx的最大取值等于混合样本的总秩和减去Wy的最小值，即$n(n+1)/2-n_2(n_2+1)/2$;同样,Wy的最大取值等于$n(n+1)/2-n_1(n_1+1)/2$。

所以,上式中的W_1和W_2均为取值在 0 与$n(n+1)/2-n_1(n_1+1)/2-n_2(n_2+1)/2=n1n2$的变量。当原假设为真时，所有的$x_i$和$y_i$相当于从同一总体中抽得的独立随机样本，$x_i$和$y_i$构成可分辨的排列情况，可看成一排$n$个样品随机地指$n_1$个为$x$,另$n_2$个为$y$,共有$C_n^{n_1}$种可能，而且它们是等可能的。基于这样的分析，在原假设为真的条件下不难求出W_1和W_2的概率分布，显然它们的分布还是相同的，这个分布称为样本大小为n_1和n_2的$Mann-Whitney-Wilcoxon$分布。

一个具有实际价值的方法是，对于每个样本中的观察数大于等于 8 的大样本来说，我们可以采用标准正态分布z来近似检验。由于W_1的中心点为$\frac{n_1n_2}{2}$,根据上式,Wx中心点μ为$\mu=\frac{n_1n_2}{2}+\frac{n_1(n_1+1)}{2}=\frac{n_1(n_1+n_2+1)}{2}$，$Wx$的方差为$\sigma^2$。

从数学上可推导出：

$$\sigma^2=\frac{n_1n_2(n_1+n_2+1)}{12} \tag{8}$$

如果样本中存在结，将影响到上式中方差，按结值调整方差的公式为

$$\sigma^2=\frac{\mathrm{n_1n_2(n_1+n_2+1)}}{12}-\frac{\mathrm{n_1n_2}\sum(\tau_j^3-\tau_j)}{12(\mathrm{n_1+n_2})(\mathrm{n_1+n_2-1})} \tag{9}$$

其中τ_j为第j个结值的个数。结值的存在将使原方差变小，这是一个显然正确的事实。标准化后W_x为

$$z=\frac{W_x-\mu\pm0.5}{\sigma}=\frac{W_x-\frac{n_1(n_1+n_2+1)}{2}\pm0.5}{\sqrt{\frac{n_1n_2(n_1+n_2+1)}{12}-\frac{n_1n_2\sum(\tau_j^3-\tau)}{12(n_1+n_2)(n_1+n_2-1)}}}\sim N(0,1) \tag{10}$$

其中分子加 0.5 或减 0.5 是为了对离散变量进行连续性修正，对于$W_x-\mu$大于 0 减 0.5 修正,对于$W_x-\mu$小于 0 加 0.5 修正。

4.1.2 相关检验结果

本模型中，选取α = 0.05，FC 阈值为 1.5，经 t-2 检验，所得火山图如图所示：

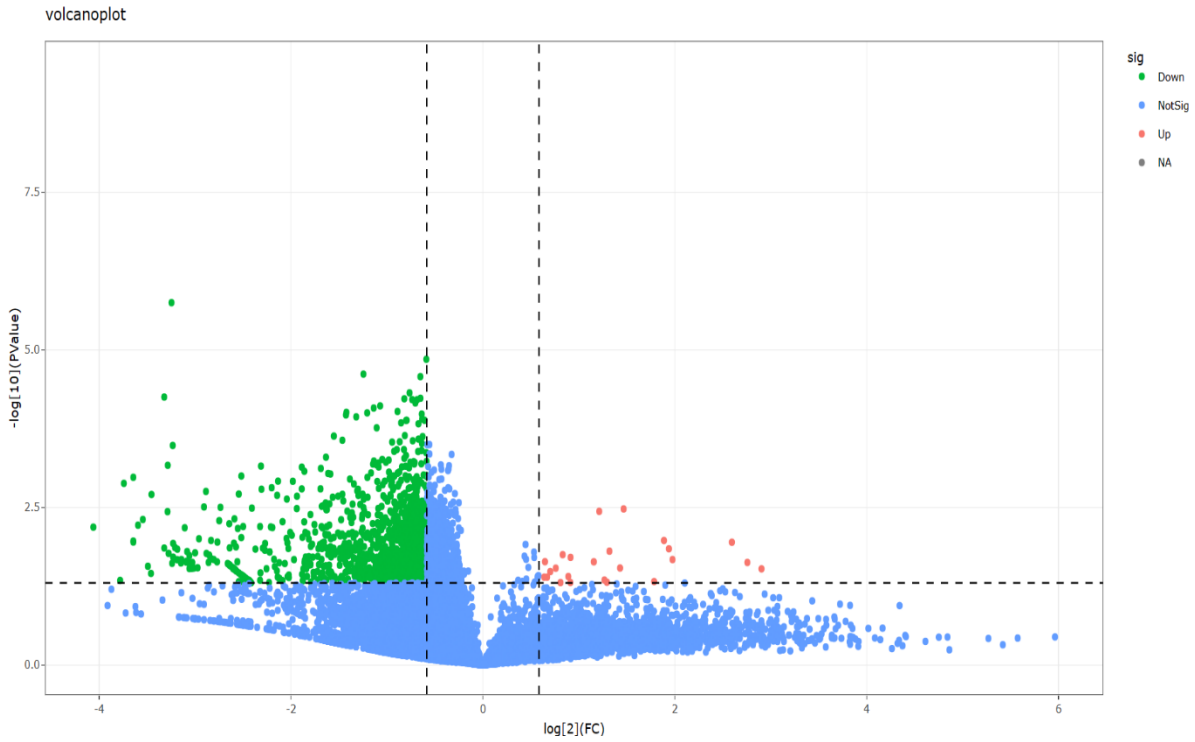


图 1. t-2 检验火山图

其中，绿色点（Down）为差异显著且下调的基因，红色点（Up）差异显著且上调的基因，蓝色点（NotSig）为无显著差异的基因，灰色点（NA）为缺失数据。相关参数估计结果较长，仅列举部分，完整参数参见附件。

表 1. 部分基因对应参数

id	FC	p	FDR	sig
ENSDARG00000000001	0.854545	0.752682	0.984748	NotSig
ENSDARG00000000002	1.290013	0.045494	0.636885	NotSig
ENSDARG00000000018	0.643277	0.000305	0.121408	Down
ENSDARG00000000019	1.004564	0.992798	0.999309	NotSig
ENSDARG00000000068	0.775194	0.491049	0.955127	NotSig
ENSDARG00000000069	0.784385	0.354551	0.955127	NotSig
ENSDARG00000000086	0.811727	0.671495	0.97371	NotSig
ENSDARG00000000103	0.794087	0.401536	0.955127	NotSig
ENSDARG00000000142	0.698921	0.511527	0.955127	NotSig
ENSDARG00000000151	0.694675	0.455305	0.955127	NotSig
ENSDARG00000000161	0.767147	0.443138	0.955127	NotSig
ENSDARG00000000175	0.798435	0.09889	0.808227	NotSig
ENSDARG00000000183	0.663502	0.049746	0.656781	Down

使用 Wilcoxon 秩和检验法对差异显著且下调的基因进行进一步检验，为体现差异，此处仅选取 Cont-1_count_fpkm 与 A-1_count_fpkm、B-1_count_fpkm、C-1_count_fpkm

进行检验，可得到如下火山图：

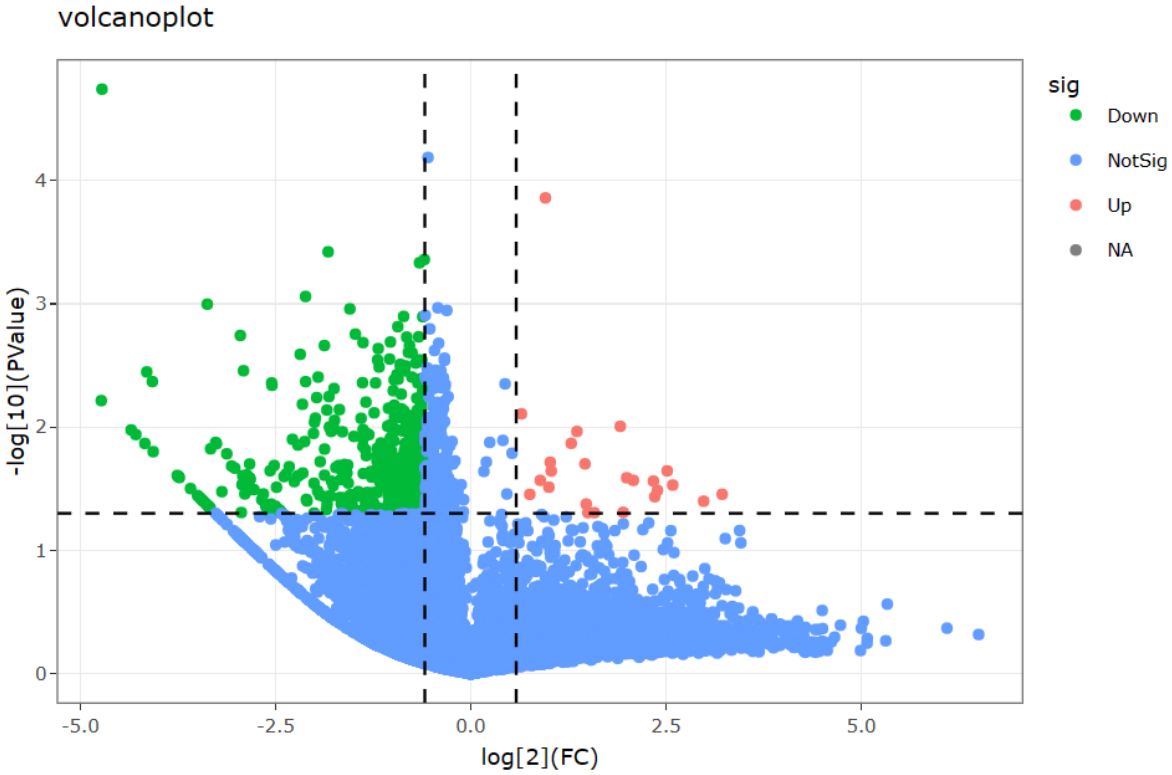


图 2. Wilcoxon 秩和检验火山图

4.2 加权基因共表达网络模型

4.2.1 简介

加权基因共表达网络分析 (WGCNA, Weighted 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。

该分析方法旨在寻找协同表达的基因模块(module)，并探索基因网络与关注的表型之间的关联关系，以及网络中的核心基因。其适用于复杂的数据模式，推荐 5 组(或者 15 个样品)以上的数据。一般可应用的研究方向有：不同器官或组织类型发育调控、同一组织不同发育调控、非生物胁迫不同时间点应答、病原菌侵染后不同时间点应答。

4.2.2 相关概念

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。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。如果没有合适的权重，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。

2.$Module$(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。

3.$Connectivity$ (连接度)：类似于图论中"度"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.$Module\ eigengene\ E$：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，使用传统 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。(降维除了 PCA，亦有 tSNE 等方法)

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.$Hub\ gene$：关键基因(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。

4.2.3 基本分析流程

（1）构建基因共表达网络：使用加权的表达相关性。

（2）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。

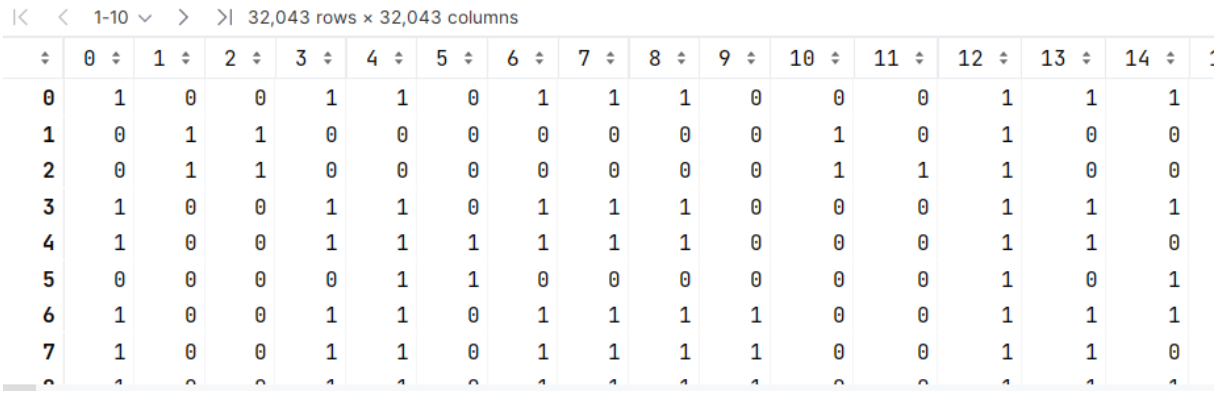
（3）研究模型之间的关系，从系统层面查看不同模型的互作网络。

（4）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。

（5）设置$Module\ membership$，导出邻接矩阵，绘制相关性图。

4.2.4 数据处理结果与分析

设置$Module\ membership\ =\ 0.5$，导出邻接矩阵如下图所示，完整的图模型请参见附件。



1-10 32,043 rows × 32,043 columns

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
0	1	0	0	1	1	0	1	1	1	0	0	0	1	1	1
1	0	1	1	0	0	0	0	0	0	0	1	0	1	0	0
2	0	1	1	0	0	0	0	0	0	0	1	1	1	0	0
3	1	0	0	1	1	0	1	1	1	0	0	0	1	1	1
4	1	0	0	1	1	1	1	1	1	0	0	0	1	1	0
5	0	0	0	0	1	1	0	0	0	0	0	0	1	0	1
6	1	0	0	1	1	0	1	1	1	1	0	0	1	1	1
7	1	0	0	1	1	0	1	1	1	1	0	0	1	1	0

图 3. 邻接矩阵建模示意图

4.3 基于共表达网络与度中心性的敏感基因筛选方法

4.3.1 简介

要寻找与血管新生直接关联的基因，可以结合问题 2 中提到的基因表达协同调节的特点来减少敏感基因的数目。以下是一种基于共表达网络的模型，结合问题 2 中的模型，用于筛选与血管新生直接关联的敏感基因：

1. 构建共表达网络：首先，根据附件中的基因表达数据，计算基因之间的相关系数（例如皮尔逊相关系数）。基于相关系数，构建基因的共表达网络，其中节点表示基因，边表示基因之间的相关性。

本问题中，选取 0.5 为相关系数 p 阈值，共表达网络使用邻接矩阵表示。

2. 识别血管新生关键基因：在共表达网络中，血管新生相关的基因往往与其他基因存在密切的关联。通过计算每个基因的网络度中心性，可评估基因在网络中的重要性。具有较高中心性指标的基因可能与血管新生直接关联。

度中心性是在网络分析中刻画节点中心性的最直接度量指标。在无向图中，度中心性测量网络中一个节点与所有其它节点直接相连的程度。对于一个拥有 g 个节点的无向图，节点 i 的度中心性是 i 与其它 g-1 个节点的直接联系总数（如果是有向图，则需要考虑的出度和入度的问题）：

$$C_D(N_i) = \sum_{j=1}^{g} x_{ij} (i \neq j) \tag{11}$$

一个节点的节点度越大，意味着该节点的度中心性越高，该节点在网络中就越重要。为消除网络规模变化对度中心性的影响，需要进行标准化，其最终计算公式如下：

$$DC_i = \frac{k_i}{N-1} \tag{12}$$

其中，k_i 表示现有的与节点 i 相连的边的数量，N - 1 表示节点 i 与其他节点都相连的边的数量。

3. 根据问题 2 中建立的基因表达调控模型，考虑基因之间的协同调节关系。根据模型参数（如回归系数、相关系数等）和基因表达数据，评估每个基因在调控网络中的重要性。具有较高回归系数和显著的表达差异的基因可能与血管新生直接关联。

4. 整合分析结果：结合共表达网络分析和基因表达调控模型的结果，选择在两个分析中均具有重要性和显著差异的基因作为与血管新生直接关联的敏感基因。

4.3.2 筛选结果

首先对数据进行 FDR 处理，采用$Benjamini - Hochberg$方法处理后，分布情况如图所示：

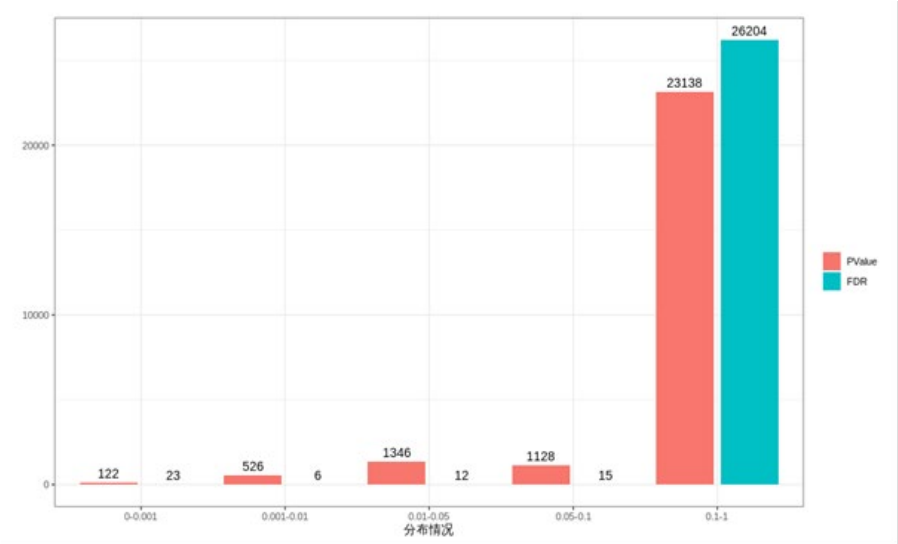


图 4. FDR 处理后分布情况

设置 PValue/FDR 阈值为 0.05，相关系数 p 阈值为 0.5，选取符合阈值要求的对应基因，计算其度中心性并进行排序，选取排序前 50 的基因，即为 50 个最敏感基因，id 列表如下：

表 2. 50 个最敏感基因 id

ENSDARG00000086490	ENSDARG00000004131
ENSDARG00000052242	ENSDARG00000068240
ENSDARG00000099772	ENSDARG00000042577
XLOC_032643	ENSDARG00000039850
ENSDARG00000043962	ENSDARG00000095130
ENSDARG00000023683	ENSDARG00000090499

ENSDARG00000098790	ENSDARG00000056504
ENSDARG00000092825	ENSDARG00000098604
ENSDARG00000095941	ENSDARG00000100120
XLOC_022960	ENSDARG00000043122
ENSDARG00000015592	ENSDARG00000022971
ENSDARG00000096949	ENSDARG00000015889
ENSDARG00000100381	ENSDARG00000089569
ENSDARG00000019304	ENSDARG00000071045
ENSDARG00000074604	ENSDARG00000053512
ENSDARG00000055781	ENSDARG00000087245
ENSDARG00000056004	ENSDARG00000017722
XLOC_005290	ENSDARG00000097443
ENSDARG00000091890	ENSDARG00000018611
ENSDARG00000086075	ENSDARG00000104037
ENSDARG00000098129	ENSDARG00000093237
ENSDARG00000101912	ENSDARG00000053876
ENSDARG00000069745	XLOC_032577
ENSDARG00000018361	ENSDARG00000076440
ENSDARG00000088766	ENSDARG00000043404

五、 模型评价与分析

总结：

本文的模型实现依赖于多种工具，包括 Python 数据分析工具、BioLadder 生物信息处理工具、spsspro 等。

经检验，模型在一般基因处理 RNA-seq 上具有一定的泛化能力与可行性，适用于一般生物医学信息处理领域中的参数估计、假设检验与基因信息提取问题。

需要注意，构建敏感基因筛选模型需要根据具体数据的特点和问题进行参数设置和调整。阈值的选择可以根据相关性的分布和网络的稀疏性进行调整。在评估基因重要性时，可以考虑其他中心性指标或使用更复杂的基因调控模型。此外，为了确保结果的可靠性，需要进行适当的统计显著性校正。

缺点：

（1）本模型的数据处理算法效率较低，实现耗时较长，有待进一步优化；

（2）共表达网络仅使用邻接矩阵、相关系数矩阵进行处理，不同基因的详细权重可通过机器学习、深度学习等方法进一步挖掘。

六、 模型推广

由于整合性状、基因表达和批量分析的结果极为复杂，可能出现多层数据，带来无意义的生物学见解。因此，高效、高质量帮助研究人员评估 RNA-seq 序列的工具十分重要。本文建立的模型可以推广适用于检验其他性状及基因表达的显著性差异，例如可以将结直肠癌投影到肿瘤微环境细胞，研究皮质投射神经元、神经元中间祖细胞对癌症的影响等。

参考文献

[1] Yang Ye, et al."Bioinformatics and Experimental Analyses Reveal NFIC as an Upstream Transcriptional Regulator for Ischemic Cardiomyopathy." genes 2022, 13, 1051.

[2] Ming-han Li, et al."Exploring the Mechanism of Active Components from Ginseng to Improve Diabetes Based on Network Pharmacology and Molecular Docking" Research Square 10.21203/rs.3.rs-1704245/v1

[3] Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis[J]. BMC bioinformatics, 2008, 9(1): 1-13.

[4] Zeng Z, Ma Y, Hu L, et al. OmicVerse: A single pipeline for exploring the entire transcriptome universe[J]. bioRxiv, 2023: 2023.06. 06.543913.

[5] van Iterson M, Boer J M, Menezes R X. Filtering, FDR and power[J]. BMC bioinformatics, 2010, 11(1): 1-11.

附录

附录 1：支撑材料清单

相关代码：
gene.ipynb

相关数据处理结果：
50 敏感基因.csv
参数估计.csv
协同调节.npy