基于 limma 差异分析的特定功能基因集查找

摘要

核黄素（维生素 b2）对于维持正常代谢和身体健康非常重要，本文利用 **limma 差异分析**结合多种算法围绕 4088 个基因对核黄素产量的影响展开分析。

首先，依据核黄素产量数据将核黄素产量分为高产组和低产组，对这两个组的基因表达水平进行 **limma 差异分析**，在高产 vs 低产、$p<0.05$ 的情况下，得到 **67 个 1.5 倍差异表达基因**（见附录 B），其中 42 个为上调基因，25 个为下调基因；12 个 2 倍差异表达基因，全为下调基因。认为 1.5 倍差异表达基因集与产量关系显著。

接着，为验证 1.5 倍差异基因集与核黄素产量的关系是否显著，又进行了以下分析。首先经数据分析发现样本中存在部分异常样本点，因此，采取**孤立森林算法**，预设 15% 的样本为噪声，共去除了 11 个离群样本。随后运用**主成分分析法**，对 67 个差异基因提取主成分，最后得到 8 个主成分能解释 90% 的方差。最后划分数据集运用**多元线性回归**的方法，预测测试集产量数据，得到**均方误差为 0.0981**，R^2 **为 0.7812**，认为所选基因集对核黄素产量有重要影响。

然后，运用多种算法进行重复验证，包括 **Elastic 回归**、**Lasso 回归**，得到一个包含 36 个基因的基因集（见附录 C），与 1.5 倍差异基因有 **19 个重合**，反映 1.5 倍差异基因集解释核黄素合成效果良好

最后，进行生物学水平验证，通过 **KEGG 数据库**找到了有关枯草芽孢杆菌合成嘌呤、合成戊糖以及核黄素代谢的信号通路，下载相关基因集（见附录 A），发现存在与之重合的基因，例如 **ribA**、**GUAB**、**PURD** 等。认为所得基因集与核黄素的产量相关性显著。。

关键字： limma 差异分析　孤立森林　主成分分析法　多元线性回归　Elastic 回归

一、 问题重述

1.1 问题背景

核黄素作为维生素 B2，对于维持正常的代谢功能、促进生长和维持健康非常重要。在食品加工、化妆品和制药行业，核黄素作为添加剂被广泛使用。了解基因表达与核黄素产量的关系可以帮助改进食品和补充剂的质量，可以增强作物的营养价值，具有多方面的益处。

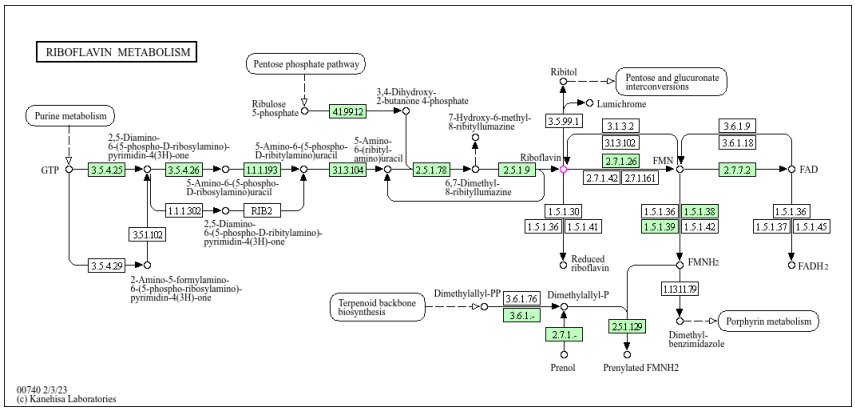


图 **1**　核黄素代谢信号通路

1.2 问题提出

现有关于影响核黄素（维生素 B2）产量的基因组数据集。数据集中包含 4088 个基因的基因表达水平和核黄素的产量。通过背景知识我们了解到仅有少量的基因对核黄素产量有影响，通过建立适当的模型识别出影响核黄素产量的基因，并通过适当的方法检验所选变量是否确实具有显著性。

二、 问题分析

针对这一问题，首先观察数据表，观察到样本命名方式代表的物种为枯草芽孢杆菌，学界对利用枯草芽孢杆菌进行核黄素合成的研究颇为丰富，因此，下面分析均围绕这一背景展开。

利用整体样本通过对各个基因与产量的相关性分析，来得出不同基因对产量的大致影响，在检验后得到在产量的不同梯度中表达有显著性差异的基因数据集作为筛选结果。但因为产量必然是多个基因共同作用的结果，所以我们要依据现有的初筛结果对样本进行一次筛选，剔除受其他基因影响大导致无法准确反映目标基因影响力的离群噪音

样本，为后续进一步细化分析奠定基础。最后量化出能反映筛选基因与产量关系的表达式，并通过划分测试集验证，作 R 方的二次显著性检验说明筛选结果的合理性。

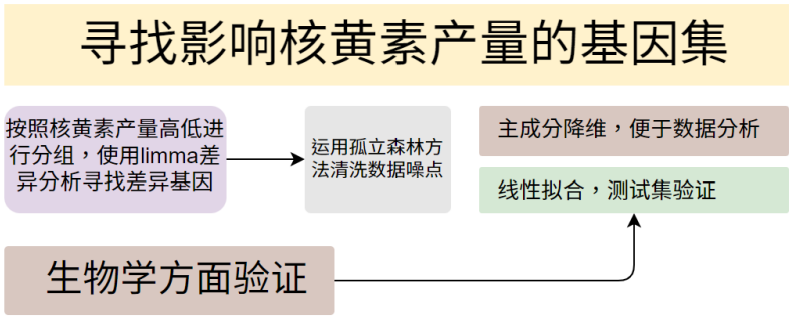


图 **2**　整体思路

三、 模型的假设

- 假设在基因组数据集中，只有少量基因对核黄素（维生素 B2）产量具有显著影响。即，核黄素的产量主要由少数几个基因调控，而大部分基因的表达对产量没有显著影响。
- 假设通过差异分析（如 Limma 差异分析）能够识别出在不同核黄素产量水平下表达有显著差异的基因，并且这些基因的差异表达是与核黄素产量相关的。
- 假设 limma 差异分析残差 ϵ_{ij} 来自同一分布，即ϵ_{ij} i.i.d.
- 假设通过孤立森林算法可以有效识别并剔除数据中的离群样本，离群样本是那些无法准确反映基因对核黄素产量影响的数据点。
- 假设主成分之间是无关的，即主成分是正交的。

四、 符号说明

符号	意义
y_{ij}	第 i 个基因在第 j 个样本中的表达值
μ_i	基因 i 的平均表达水平
x_{ij}	样本 j 的实验设计变量（如治疗组、对照组等）
ϵ_{ij}	残差，表示误差或噪音
ρ	Spearman 相关系数
d_i	第 i 对观测值的秩之差
n	样本的数量
$\bar{x_j}$	第 j 个指标的样本均值
s_j	第 j 个指标的样本标准差
R	相关系数矩阵
λ_j	相关系数矩阵的第 j 个特征值
u_{ij}	相关系数矩阵的第 j 个特征向量
y_i	第 i 主成分
b_j	主成分 y_j 的信息贡献率
a_p	前 p 个主成分的累计信息贡献率
z	综合得分

五、 模型的建立与求解

5.1 相关性分析

5.1.1 limma 差异分析

对于生物基因与性状数据，通过查阅文献，决定采用 Limma 差异分析来筛选与产量有高相关性的基因集合。

- **算法简介**

Limma（Linear Models for Microarray Data）是一种用于分析基因表达数据的统计方法，

特别适用于从微阵列数据和 RNA-Seq 数据中检测差异表达基因。Limma 的核心思想是通过线性模型来评估基因表达的差异，其融合了经验贝叶斯方法、多重检验校正来增强结果可信度。

Limma 使用线性模型来表示每个基因的表达数据。通常的模型形式为：

$$y_{ij} = \mu_i + x_{ij} + \epsilon_{ij} \tag{1}$$

其中：

y_{ij} 是第 i 个基因在第 j 个样本中的表达值。

μ_i 是基因 i 的平均表达水平。

x_{ij} 是样本 j 的实验设计变量（如治疗组、对照组等）。

ϵ_{ij} 是残差，表示误差或噪音。

- **模型建立**

首先，针对需要研究的目标：核黄素（维生素 B2）的产量，根据产量的高低分组将样本分为 high 组与 low 组作为实验设计变量，再利用公式 (1) 建立核黄素产量与基因表达的线性模型。

- **模型求解**

将分组数据与表达谱数据导入，利用 R 软件包 limma(version 3.40.6) 进行差异分析，以获得不同比较组与对照组间的差异基因，得到结果如下表，其中上调为 high 组高表达差异基因，下调为 low 组高表达差异基因

表 **1**　差异分析结果表

显著性阈值	1.3 倍差异		1.5 倍差异		2 倍差异	
	上调	下调	上调	下调	上调	下调
p<0.05	138	59	42	25	0	12
p<0.01	117	47	40	25	0	12
FDR<0.05	83	40	36	25	0	12
FDR<0.01	39	21	23	16	0	8

可以发现在 p 小于 0.05 的情况下差异倍数为 1.5 倍的基因数目有 67 个（结果见附录 B），其中 42 个基因在高产组显著上调，25 个基因在低产组显著下调。我们取这 67 个基因为显著基因集合，并绘制了 1.5 倍差异基因的火山图及 2 倍差异基因的表达热图。

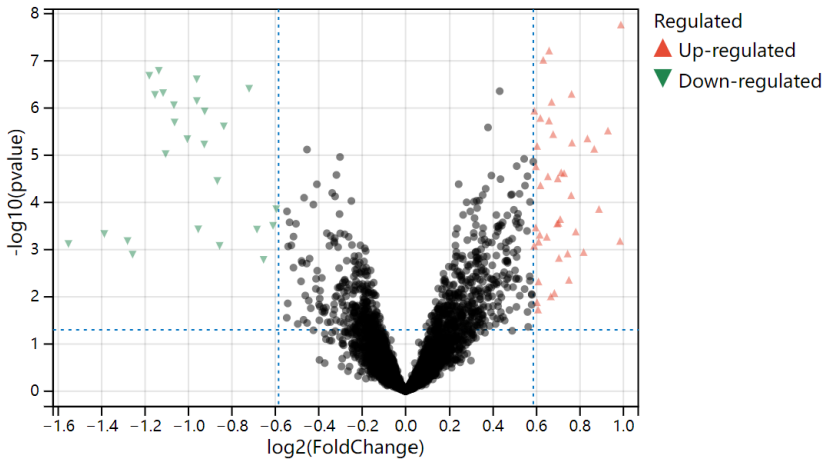


图 3　1.5 倍差异基因火山图

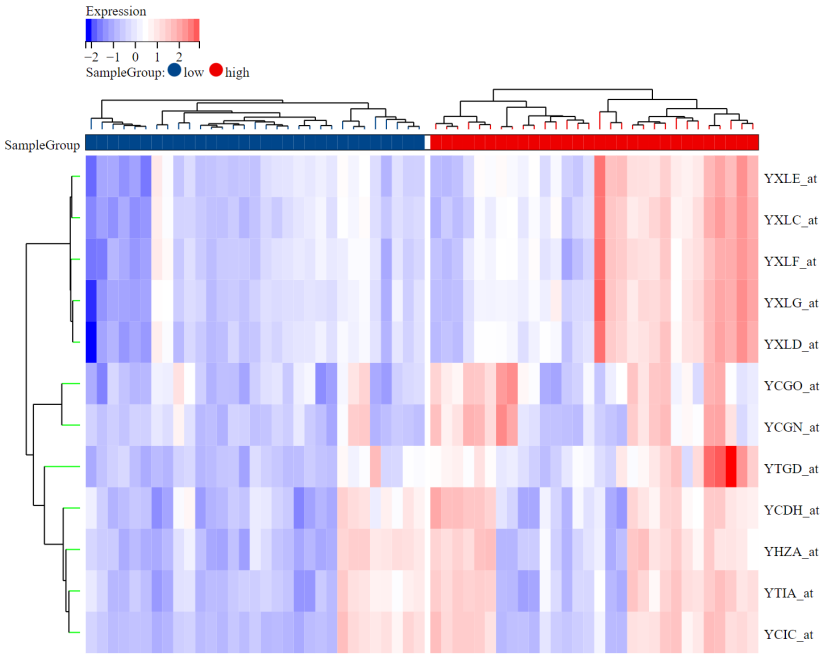


图 4　2 倍差异基因表达热图

5.1.2 Spearman 相关性系数评估

根据上文得出的差异基因集，初步断定，核黄素（维生素 B2）的产量与这些差异表达基因存在较强相关性。为探究这种相关性的具体特征及显著性，通过查阅文献与数据处理，决定采用 Spearman 相关系数进行线性相关分析。

- **Spearman 相关性系数简介**

Spearman 相关系数（Spearman's rank correlation coefficient），也称为 Spearman ρ（rho），是一种非参数统计方法，用于评估两个变量之间的单调关系。与 Pearson 相关系

数不同，Spearman 相关系数不假设数据的分布是正态的，因此在数据不满足线性关系或者含有异常值时尤为有用。

- Spearman 相关系数的计算方法

1 排序数据：将两个变量的数据分别排序。每个数据点用其在样本中的排名表示，称为秩（rank）。如果有相同值（平级），通常给它们赋予平均秩。

2 计算秩差：对于每对观测值，计算两个变量的秩之差 d_i。

3 计算 Spearman 相关系数：使用以下公式计算 Spearman 相关系数：

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \tag{2}$$

其中：- d_i 是第 i 对观测值的秩之差。

- n 是样本的数量。

- Spearman 相关系数的解释

值范围：Spearman 相关系数的值范围从-1 到 1。

- $\rho = 1$：表示两个变量之间存在完全的正单调关系。

- $\rho = -1$：表示两个变量之间存在完全的负单调关系。

- $\rho = 0$：表示没有单调关系。

- 单调关系：

Spearman 相关系数衡量的是单调关系，不要求变量之间的关系是线性的。例如，即使两个变量呈非线性但单调的关系，Spearman 相关系数仍然能够检测出它们之间的相关性。

- 使用场景

- 非正态分布的数据：当数据不满足正态分布假设时，Spearman 相关系数比 Pearson 相关系数更合适。

- 异常值存在的情况：由于 Spearman 相关系数基于秩，而不是原始数据，所以它对异常值不敏感。

- 探索单调关系：在研究中需要判断两个变量是否有单调关系，而不关心具体的线性关系时，Spearman 相关系数是一个合适的选择。

- **Spearman 相关性计算**

首先绘制出产量分布直方图，初步判断产量分布为偏态分布，适合采用 Spearman 相关性系数分析。

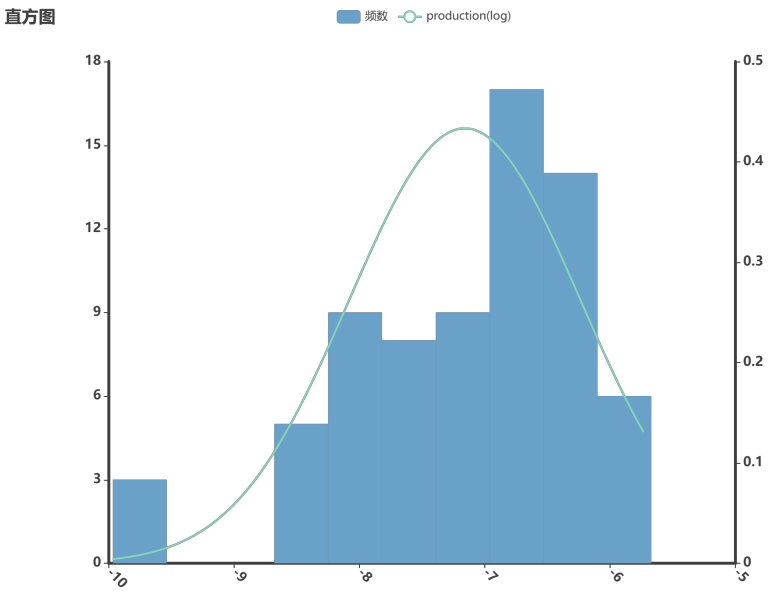


图 **5**　产量分布直方图

接着我们利用 Spearman 相关系数进行分析并绘制出 1.5 倍差异基因相关性点棒图与 2 倍差异基因相关性热图，可以看出其相关性非常明显，且 p<0.01。

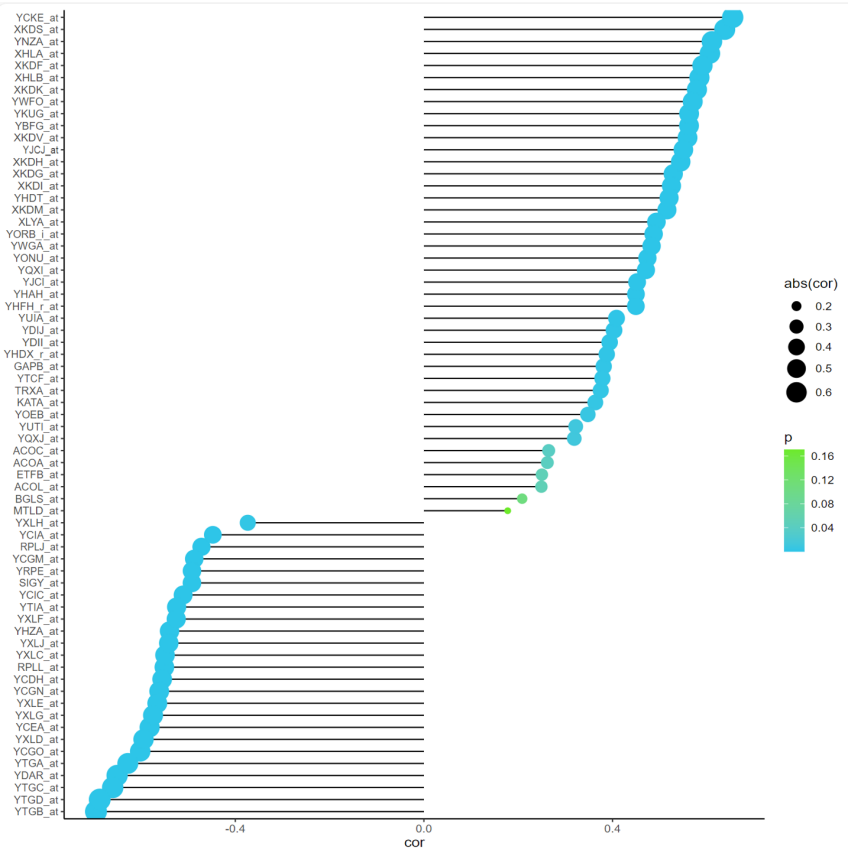


图 **6**　**1.5** 倍差异基因相关性点棒图

图 7　2 倍差异基因相关性热图

5.2 孤立森林清洗离群样本

5.2.1 样本数据可视化

挑选 2 倍差异基因绘制相关性散点图，可以看出基因表达量与产量大体上呈负相关，与相关性热图结果一致，并且散点图显示有部分数据噪点，符合问题分析，样本存在无法准确反映目标基因影响力的离群噪音样本，所以我们选用孤立森林算法清洗离群样本。

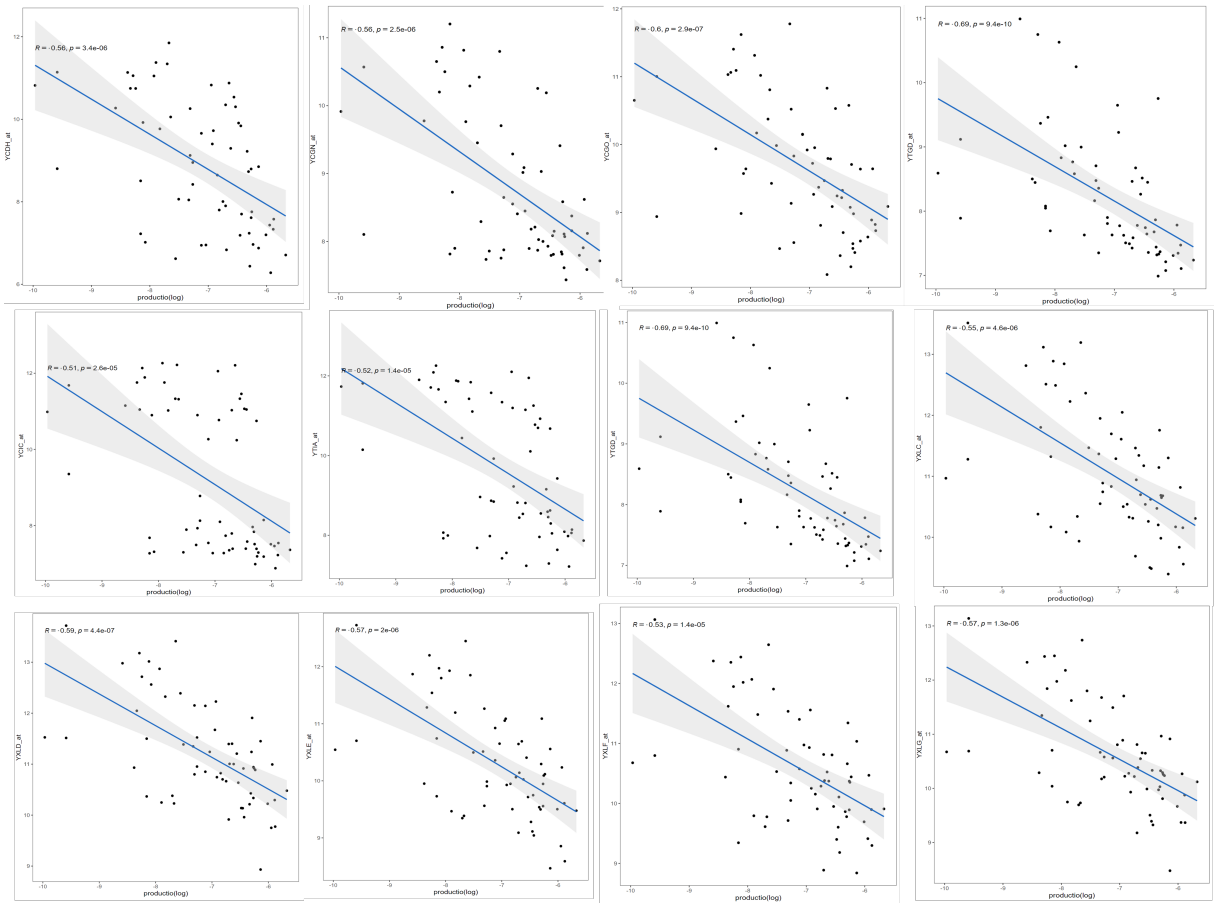


图 8　12 个 2 倍差异基因相关性散点图

5.2.2 孤立森林算法简介

孤立森林（Isolation Forest）是一种基于树的无监督机器学习算法，主要用于异常检测。与其他基于密度或距离的方法不同，孤立森林算法通过随机选择特征并随机选择分割值来隔离数据点。其涉及概念如下：

- 随机分割：在构建孤立森林时，每棵树是通过随机选择一个特征，并在该特征的取值范围内随机选择一个分割点来分割数据集。
- 孤立：数据点越容易被孤立（即需要的分割次数越少），它们越有可能是异常点。正常数据点通常需要更多的分割才能被完全孤立。
- 树结构：通过多次分割数据集，构建树结构，直到每个数据点被孤立或树的高度达到预设的最大值。

5.2.3 模型建立

结合孤立森林算法（具体代码见附录 D）对样本数据进行如下步骤：

1 从基因表达数据集中随机抽取多个子集 u。并预设 15% 的样本异常。
2 对于每个 u，构建孤立树，通过随机选择特征和分割点，不断分割数据，直到每个数据点被孤立。
3 对于每个数据点，计算在孤立树中被孤立所需的路径长度。
4 根据路径长度，计算异常分数。路径长度短的数据点更可能是异常点，路径长度长的数据点更可能是正常点。

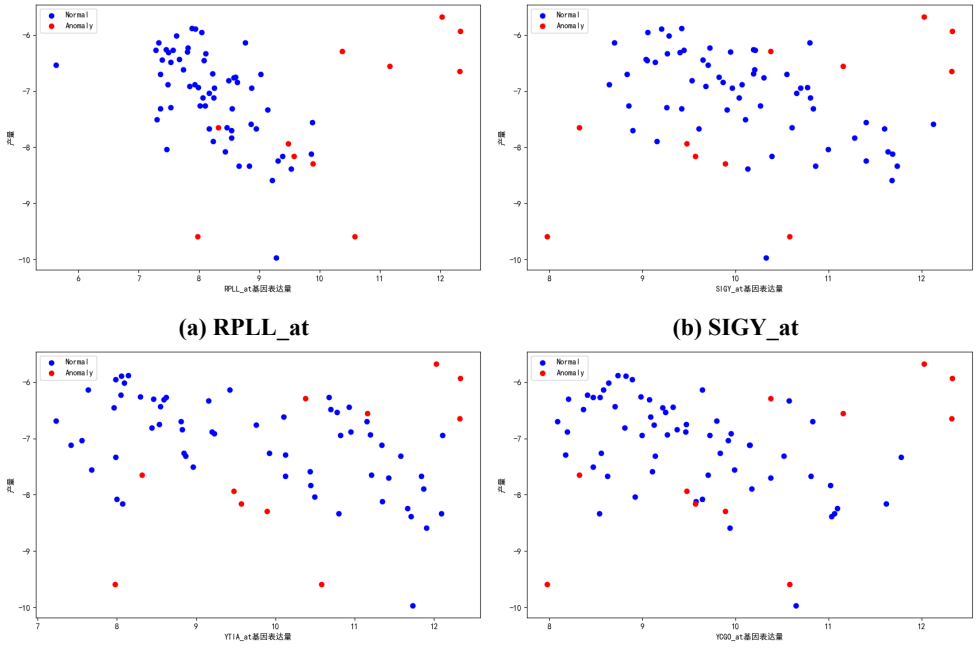
5.2.4 模型求解

经过孤立森林算法求解后，剔除离群样本，剔除的样本如下表所示：

表 **2**　离群样本表

离群样本名
b_Fbat107PT24.CEL
b_Fbat107PT48.CEL
b_Fbat107PT52.CEL
knh_491_BS0001_Fbat1070_PT30_1.Rep.CEL
knh_502_BS3416_E_Fbat1071_PT24_1.Rep.CEL
knh_503_BS3416_E_Fbat1071_PT30_1.Rep.CEL
knh_506_BS3416_E_Fbat1078_PT24_2.Rep.CEL
knh_661_BS5009_Fbat1374_PT48_1.Repl._No.4_rep_cDNA_new Yeast.CEL
knhb_090_Fbat284PT24.CEL
knhb_091_Fbat284PT48.CEL
knhb_247_Fbat525PT24.CEL

对异常点标注并绘制基因表达与产量关系的散点图，可以发现噪声数据确实被基本剔除。



(a) RPLL_at　　**(b) SIGY_at**

(c) YTIA_at　　**(d) YCGO_at**

图 **9**　单个基因表达量与产量关系散点图（部分）

5.3 主成分降维

5.3.1 主成分分析法简介

主成分分析法是利用降维的思想，在损失很少信息的前提下把多个指标转换为几个综合指标的多元统计方法。转换成的综合指标被称为主成分，其中每个主成分都是原始变量的线性组合，且各个主成分之间互不相关，这就使得主成分比原始变量具有某些更优越的性能。

主成分分析法模型的建立：

Step1：将原始数据标准化，以消除量纲影响，使不同维数据之间具有可比性。

假设进行主成分分析的指标变量有 m 个：x_1，x_2，…，x_m 共有 n 个评价对象。第 i 个评价对象的第 j 个指标的取值为 x_{ij}。将各指标值 x_{ij} 转换成标准化指标 $\widetilde{x_{ij}}$

$$x'_{ij} = \frac{(x_{ij} - \bar{x_j})}{S_j}\,(i = 1, 2, \ldots, n; j = 1, 2, \ldots, m) \tag{3}$$

其中,

$$\bar{x_j} = \frac{1}{n}\sum_i^n x_{iJ} \tag{4}$$

$$s_j = \frac{1}{n-1}\sum_{\bar{i}}^n (x_{ij} - \bar{x_J})^2 (i = 1, 2, \ldots, m) \tag{5}$$

即 $\bar{x_J}$，s_j 为第 j 个指标的样本均值和样本标准差。对应的 $\widetilde{x_{ij}} = \frac{x_i - \widetilde{x_i}}{s_i}, (i, j = 1, 2, \ldots, m)$ 为标准化指标变量。

Step2：建立相关系数矩阵 R

相关系数矩阵 $R = (r_{ij})_{m\times m}$，其中 $r_{ij} = \frac{\sum_{k=1}^n \widetilde{x_{k_i}} \cdot \widetilde{x_{k_j}}}{n-1} (i, j = 1, 2, \ldots, m)$

式中 $r_{ii} = 1$，。$r_{ij} = r_{ji}$ 是第 i 个指标与第 j 个指标的相关系数。

Step3：计算特征值与特征向量

计算相关系数矩阵 R 的特征值 $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m \geq 0$，及对应的特征向量 u_1，u_2，…，u_m，其中，$\mu_j = \left(\mu_{1j},\ \mu_{2j}, \ldots, \mu_{nj}\right)^T$ 由特征向量组成 m 个新的指标变量。

$$\begin{cases} y_1 = u_{11}\widetilde{X_1} + u_{21}\widetilde{X_2} + \cdots + u_{n1}\widetilde{X_n} \\ y_2 = u_{12}\widetilde{X_1} + u_{22}\widetilde{X_2} + \cdots + u_{n2}\widetilde{X_n} \\ \vdots \quad \vdots \quad \vdots \\ y_m = u_{1m}\widetilde{X_1} + u_{2m}\widetilde{X_2} + \cdots + u_{nm}\widetilde{X_n} \end{cases} \tag{6}$$

式中 y_1 是第一主成分,y_2 是第二主成分，…，y_m 是第 m 主成分。

Step4：计算综合得分

计算特征值 $_j$（$j=1,2,\ldots,m$）的信息贡献率和累计贡献率。

称 $b_j=\frac{\lambda_j}{\sum_{k=1}^{m}\lambda_m}(j=1,2,\ldots,m)$ 为主成分 y_j 的信息贡献率：

$$a_p=\frac{\sum_{k=1}^{P}\lambda_k}{\sum_{k=1}^{m}\lambda_k} \tag{7}$$

当 a_p 接近 1 时（a_p=0.85，0.90，0.95），则选择前 p 个指标变量为 p 个主成分，代替原来 m 个指标。

计算综合得分

$$z=\sum_{i}^{P}b \tag{8}$$

其中 b_j 为第 j 主成分的信息贡献率。

5.3.2 模型求解

- 相关性分析

  相关性分析是主成分分析的前置条件，这里运用 KMO 进行相关性分析，计算所得各基因的综合 KMO 值为 0.64，认为相关性适中，可以接受使用主成分分析。

- 模型求解

  在 67 个差异表达基因集中，最终提取到了 8 个主成分（具体代码见附录 E），保留足够的主成分来解释的方差。经过主成分降维后，数据仍然能够反映原基因集的特征，仍然可以说明原基因表达数据与产量之间的关系。但极大减少了数据回归的困难。

表 **3**　主成分贡献率表

	comp.1	comp.2	comp.3	comp.4	comp.5	comp.6	comp.7	comp.8
各主成分贡献表	0.359811	0.197192	0.121709	0.092978	0.0529	0.03619	0.025945	0.017712
累计主成分贡献表	0.359811	0.557003	0.678712	0.77169	0.824589	0.86078	0.886725	0.904437

5.4 利用主成分预测产量数据

接下来运用上文得出的 8 个主成分来进行线性拟合，并运用测试集验证模型效能，判断所得基因的显著性。（具体代码见附录 E）

首先，对数据清洗完的样本划分训练集与测试集，比例为 10 %

对各主成分得分与产量数据进行线性拟合，测试集评价结果如下表：

表 **4**　测试集结果

均方误差	决定系数 R^2
0.0981	0.7812

认为相关性良好，原基因集能够很好反映产量的变化，验证了这些基因与产量的相关性。

六、 模型评价

6.1 算法评价

6.1.1 limma 差异分析

- 优点

1 limma 支持多种实验设计，并且在各种高通量基因表达数据分析中都表现良好。
2 limma 通过贝叶斯方法对方差进行调整，在样本量较小的情况下。这种方法可以显著提高对差异表达的检测能力，
3 limma 提供了多重假设检验的调整方法，如 Benjamini-Hochberg 方法，帮助控制假阳性率。

- 缺点

1 计算复杂度高，对于非常大的数据集或非常高维的数据，计算和存储开销可能比较大。
2 对于低表达基因，因为这些基因的测量值可能噪音较大，limma 可能不够稳定。

6.1.2 孤立森林算法

- 优点

1 孤立森林在处理高维数据时也能保持较好的性能。这是因为它不依赖于距离度量或数据的分布，而是依赖于样本点的孤立性。
2 孤立森林是一种无监督学习算法，不需要标记样本数据即可进行异常检测。这使得它特别适合于未标记的数据集。

- 缺点

1 孤立森林的性能在不同的随机种子下可能会有所波动，特别是在训练数据量较少的情况下。算法的随机性可能导致结果的不稳定。

6.1.3 主成分分析

- 优点

1 PCA 可以有效地将高维数据降维到较低维度，同时保留数据的大部分信息和特征。这有助于减少计算复杂度和存储需求。

2 通过将数据投影到新的正交基上，PCA 能够消除特征之间的多重共线性和冗余信息，提取出主要的变化方向。

3 由于 PCA 基于线性代数方法，在现代计算机上计算高效，适用于大多数数据集。

- **缺点**

1 PCA 对异常值和噪声较为敏感。这些异常点可能会显著影响主成分的方向，从而影响降维效果。

6.2 结果评价

6.2.1 Elastic 回归分析对比

- **模型简介**

Elastic 回归（Elastic Net Regression）是一种回归分析方法，它结合了岭回归（Ridge Regression）和套索回归（Lasso Regression）的优点。Elastic Net 回归在处理特征选择和正则化方面具有很好的性能，尤其是在特征数量远大于样本数量时。它在减少模型复杂性和提高预测能力方面表现出色。

Elastic Net 回归的目标函数包含两部分：一个是最小化预测误差的部分（通常是均方误差），另一个是正则化项。公式如下：

$$\text{minimize} \quad \frac{1}{2n}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2 + \lambda_1\|\boldsymbol{\beta}\|_1 + \frac{\lambda_2}{2}\|\boldsymbol{\beta}\|_2^2 \tag{9}$$

其中：

- y_i 是第 i 个观测值的实际响应。
- $\mathbf{x}_i$ 是第 i 个观测值的特征向量。
- $\boldsymbol{\beta}$ 是回归系数向量。
- $\|\boldsymbol{\beta}\|_1$ 是 $\boldsymbol{\beta}$ 的 L_1 范数，即所有回归系数绝对值的和。
- $\|\boldsymbol{\beta}\|_2^2$ 是 $\boldsymbol{\beta}$ 的 L_2 范数的平方，即回归系数平方和。
- λ_1 和 λ_2 是正则化参数，分别控制 Lasso 和 Ridge 的影响程度。

- **结果误差分析**

经过 Elastic 回归分析（代码见附录 F）得到 36 个显著相关基因 (详见附录 C)，与 1.5 倍差异基因集有 19 个基因重叠

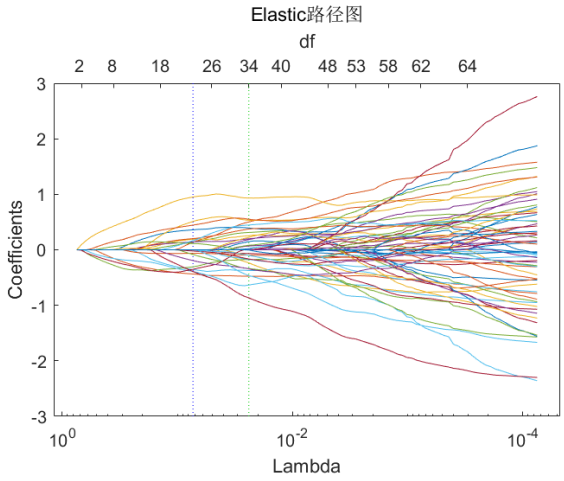


图 **10**　回归结果可视化

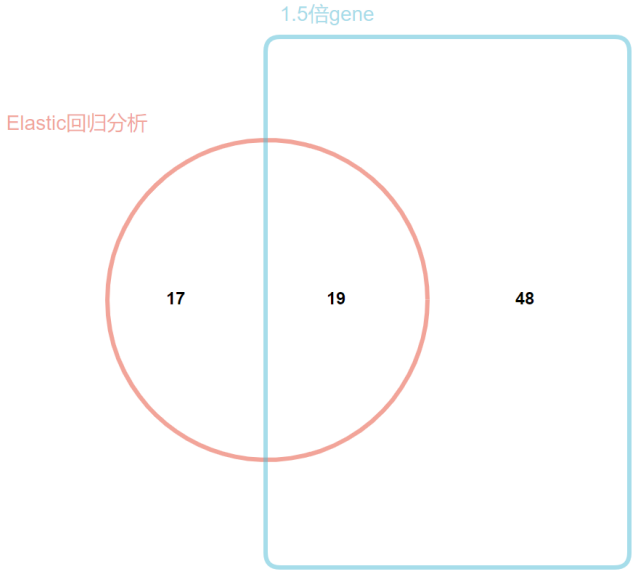


图 **11**　**1.5** 倍差异基因与 **Elastic** 韦恩图

重叠效果适中，认为 1.5 倍差异基因集反映产量的效果良好。

6.2.2 生物学角度验证结果

目前对于枯草芽孢杆菌核黄素合成代谢途径已有一个较为深入的了解，可将其为两个部分：

一是前体物和鸟嘌呤-5’-三磷酸（GTP）和 5-磷酸核酮糖（Ru-5-P）的合成，GTP 和 Ru-5-P 分别通过嘌呤途径（purine biosynthesis pathway）和磷酸戊糖途径（Pentose phosphate pathway）获得；

第二部分是前体物 Ru-5-P 和 GTP 以 2:1 的比例经由核黄素操纵子编码的一系列合成相关酶的催化生成核黄素，该过程一共包含 7 步反应。

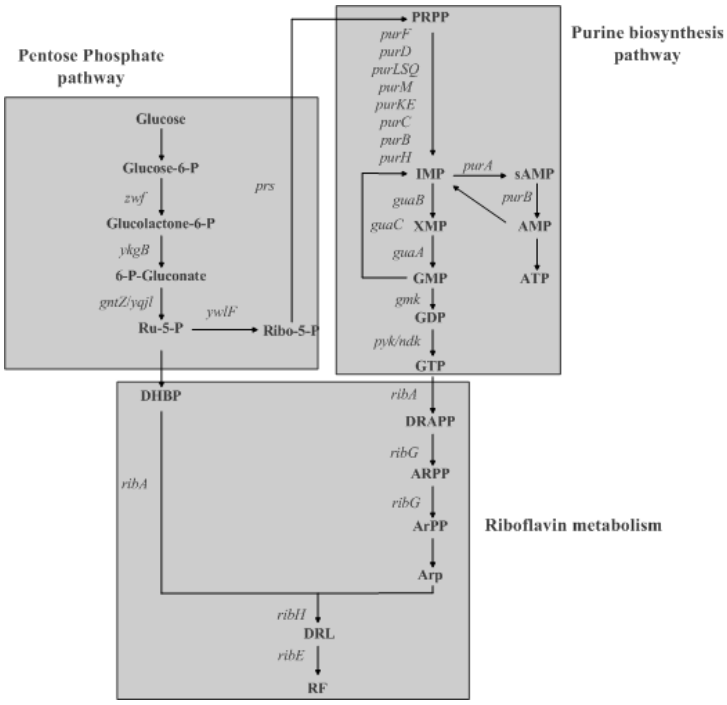


图 **12**　核黄素的合成的基本途径

由上图可以发现，与核黄素合成相关的型号通路一共有三个：

- 一是嘌呤合成的信号通路
- 二是戊糖合成的信号通路
- 三是核黄素代谢通路

因此，通过 KEGG 数据库进行检索，我们可以得到枯草芽孢杆菌在这些通路中的 marker 基因（结果见附录 A），这些基因被认为是调控核黄素生成与转运的重要基因。

于是将我们筛选所得基因集与核黄素生成与转运相关的 marker 基因集进行对比，发现了一些重复出现的基因。

- ribA

  ribA 基因编码核黄素合成途径中的一种多功能酶，称为 GTP 环化水解酶/3,4-二氢核黄素-5’-磷酸合成酶（GTP cyclohydrolase II/3,4-dihydroxy-2-butanone-4-phosphate synthase）。这两种酶的活性分别涉及核黄素合成的两个关键步骤

- GUAB

  GUAB（guaB）基因编码的 IMP 脱氢酶主要涉及鸟嘌呤核苷酸的合成，与核黄素的直接合成没有明确的关系。然而，鸟嘌呤核苷酸和其他嘌呤、嘧啶核苷酸的合成对细胞的整体代谢状态有重要影响。充足的核苷酸供应对于细胞的正常生长和代谢是必要的，而这些代谢活动也间接影响到核黄素的合成和利用。例如，在细胞能量代

谢的协调过程中，鸟嘌呤核苷酸的合成和核黄素代谢可能会相互影响。因此，GUAB 的活性对细胞的整体代谢健康有影响，这可能间接影响核黄素的合成和转运。

- PURD

  嘌呤代谢的产物（如 ATP、GTP）在细胞的能量代谢和遗传信息的传递中起重要作用。核黄素是重要的辅酶 FAD 和 FMN 的前体。这些辅酶在多种氧化还原反应中起作用，是能量代谢和细胞呼吸的核心。

参考文献

[1] Abbas CA, Sibirny AA2011.Genetic Control of Biosynthesis and Transport of Riboflavin and Flavin Nucleotides and Construction of Robust Biotechnological Producers. Microbiol Mol Biol Rev75:.https://doi.org/10.1128/mmbr.00030-10

[2] 刘峰, 曹东, 宗渊, 等.RNA-Seq 挖掘黄绿卷毛菇中核黄素合成途径差异表达基因 [J]. 分子植物育种,2024,22(01):77-84.DOI:10.13271/j.mpb.022.000077.

[3] 李健. 影响枯草芽孢杆菌核黄素过量合成的关键因素研究 [D]. 天津大学,2018.DOI:10.27356/d.cnki.gtjdu.2018.002533.

[4] Lienhart WD, Gudipati V, Macheroux P. The human flavoproteome. Arch Biochem Biophys. 2013 Jul 15;535(2):150-62. doi: 10.1016/j.abb.2013.02.015. Epub 2013 Mar 15. PMID: 23500531; PMCID: PMC3684772.

[5] 徐晶玉. 影响枯草芽孢杆菌核黄素合成的相关代谢途径研究 [D]. 天津大学,2022.DOI:10.27356/d.cnki.gtjdu.2022.001417.

[6] 黄灿. 核黄素高产枯草芽孢杆菌选育和发酵优化 [D]. 天津大学,2018.

[7] 王永成. 产核黄素枯草芽孢杆菌的代谢工程研究 [D]. 天津大学,2015.

[8] 刘露. 枯草芽孢杆菌嘌呤合成途径相关基因及 ribC 基因的遗传修饰 [D]. 天津大学,2014.

附录 A　KEGG 寻找所得基因集

YKGB	KDGA	PURF	PURE	HPRT	RELA	RELA	ADK
GNTZ	GDH	PURN	PURB	NDK	YJBM	YJBM	YITA
TKT	YVCT	PURT	PURH	YSNA	PURA	PURA	UREC
YWLF	GNTK	PURL	APT	XPT	YERA	YERA	UREB
RBSK	KDGK	PURQ	YUND	GUAA	ADEC	ADEC	UREA
DRM	FBAA	PURM	DEOD	NRDE	YFKN	YFKN	YURH
PRS	FBP	PURK	YLMD	NRDF	YMDB	YMDB	PGI
YCLB							

附录 B　1.5 倍差异基因

YCKE	YXLE	XHLA	YHDT	YCGN	YQXJ	RPLJ	YDIJ	YHDX
YNZA	YQXI	YTGC	XKDK	YHAH	YKUG	YORB	YHFH	GAPB
YWFO	YXLC	YWGA	YCGO	XKDG	XLYA	YRPE	YCDH	YHZA
YXLG	YTGA	YJCI	XKDF	XKDH	YCEA	RPLL	YUTI	YTCF
YXLD	YJCJ	YXLF	YCGM	YONU	XKDM	BGLS	YCIC	YCIA
YTGB	YTGD	YXLJ	XKDV	XKDI	TRXA	YTIA	KATA	ACOA
YDAR	XKDS	XHLB	YBFG	SIGY	YUIA	YDII	YXLH	YOEB
ACOL	ACOC	ETFB	MTLD					

附录 C　Elastic 所得结果

YLAJ	SPOIISA	PKSA	YOAB	PROJ	ccpB
YXLD	YHCL	YJCJ	YTGC	YONU	YTGB
YXLE	YKUG	XKDF	SIGY	YSXD	YXLJ
YHFH_r	YTGD	YXLG	YTGA	YBGB	YXLF
GGT	XKDX	YCKE	THIK	LYTS	YCEK
YNZA	YWFO	YHAH	YBAR	YCGT	XKDK

附录 D　随机森林去除噪声————源程序

```
import matplotlib.pyplot as plt
from sklearn.ensemble import IsolationForest

# 训练孤立森林模型
iso_forest = IsolationForest(contamination=0.15) # 假设15%的样本是异常的
y_pred = iso_forest.fit_predict(df4)

# 标记异常样本
outliers = y_pred == -1
X_df_clean = df4[~outliers]
y_df_clean = y1[~outliers]
```

附录 E　主成分分析代码

```
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error, r2_score

from sklearn.decomposition import PCA

# 初始化PCA并保留足够的主成分来解释90%的方差
pca = PCA(n_components=0.9)
X_pca = pca.fit_transform(X_df_clean)

# 查看主成分数量
print(f'Number of components: {X_pca.shape[1]}')

# 查看各主成分解释的方差比例
explained_variance_ratio = pca.explained_variance_ratio_
print("Explained variance ratio of each component:")
print(explained_variance_ratio)

# 查看累计解释的方差比例
cumulative_explained_variance = explained_variance_ratio.cumsum()
print("Cumulative explained variance ratio:")
print(cumulative_explained_variance)

X_train, X_test, y_train, y_test = train_test_split(X_pca, y_df_clean, test_size=0.1,
    random_state=42)

model = LinearRegression()
model.fit(X_train, y_train)
```

```
y_pred = model.predict(X_test)

# 计算均方误差和R方值
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)

print(f'Mean Squared Error: {mse}')
print(f'R^2 Score: {r2}')
```

附录 F　Elastic 分析代码

```
% 读取CSV文件为表格
dataTable = readtable('65.xlsx');

% 将表格转换为矩阵
X = table2array(dataTable(:, 2:end));
% 获取表格的列标题
columnTitles = dataTable.Properties.VariableNames;

% 将列标题转换为单元数组
columnTitlesCell = cellstr(columnTitles);
% 读取CSV文件为表格
dataTable_2 = readtable('y_data.csv');

% 将表格转换为矩阵
y = table2array(dataTable_2(:, 2:end));


% 使用lasso函数进行Elastic回归
[B, FitInfo] = lasso(X, y,'Alpha', 0.5, 'CV', 5);

% 找到最佳lambda对应的系数
idxLambda1SE = FitInfo.Index1SE;
coef = B(:, idxLambda1SE);

% 输出最佳lambda对应的系数
disp('最佳lambda对应的系数: ');
disp(coef);
% 找出非零系数的下标
nonZeroIdx = find(coef ~= 0);
nonZeroColumnTitles = columnTitlesCell(nonZeroIdx + 1);
%输出非零值的下标
disp('非零值的下标: ');
disp(nonZeroIdx);
```

```
% 绘制Lasso路径
lassoPlot(B, FitInfo, 'PlotType', 'Lambda', 'XScale', 'log');
xlabel('Lambda');
ylabel('Coefficients');
title('Elastic路径图');
```