**美吉生物结题报告3**

**iTRAQ定量蛋白质组学分析**

**（生信部分）**

客户姓名：王 琛

项目编号：MJ20161201030

时 间：2016年12月

**目 录**

[一、项目信息 3](#_Toc3594)

[二、信息分析流程 4](#_Toc19371)

[三、本项目信息学分析内容 5](#_Toc19290)

[四、生物信息分析 6](#_Toc2564)

[4.1全谱分析 6](#_Toc7825)

[4.1.1 GO功能分类注释 6](#_Toc4663)

[4.1.2 KEGG注释通路分析 8](#_Toc31379)

[4.1.3 COG功能分类注释 11](#_Toc12811)

[4.2差异蛋白分析 12](#_Toc19117)

[4.2.1统计分析 12](#_Toc20888)

[4.2.2差异蛋白GO分析 14](#_Toc24726)

[4.2.3差异蛋白KEGG分析 17](#_Toc17865)

[4.2.4差异蛋白表达模式聚类分析 19](#_Toc10071)

[4.2.5差异蛋白互作网络分析 21](#_Toc14382)

[4.2.6 Ipath整合分析 23](#_Toc27225)

[五、附录 25](#_Toc22067)

[5.1 结果文件列表 25](#_Toc14913)

[5.2常用数据库及专用名词介绍 27](#_Toc19253)

[5.3文件解压缩方法 27](#_Toc1335)

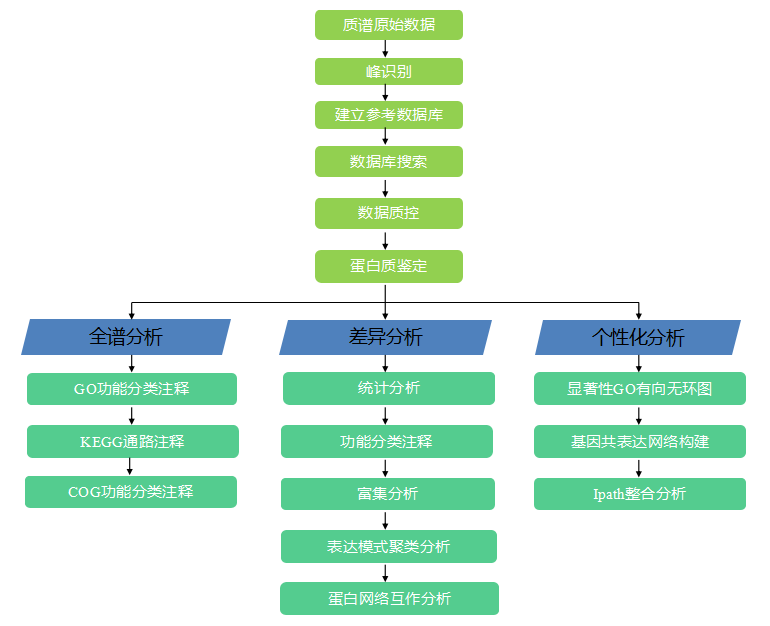
[5.4文件打开或浏览方法 27](#_Toc25606)

[5.5 联系方式 29](#_Toc26310)

# 一、项目信息

|  |  |  |  |
| --- | --- | --- | --- |
| **项目名称** | | | |
| 蛋白鉴定分析 | | | |
| **合同编号** | | | |
| MJ20161201030 | | | |
| **项目样本信息** | | | |
| **物种信息** | 小鼠 | | |
| **实验目的** | / | | |
| **客户信息** | | | |
| **单位名称** | / | | |
| **单位地址** | / | | |
| **实验室导师** | / | 电话 | / |
| 邮箱 | / |
| **项目联系人** | 王 琛 | 电话 | 18930177004 |
| 邮箱 | / |
| **美吉联系人信息** | | | |
| **销售员** | 谷 雨 | 电话 | 18521332283 |
| 邮箱 | / |
| **技术支持** | 沈娇娇 | 电话 | 021-51875086-8075 |
| 邮箱 | jiaojiao.shen@majorbio.com |

# 二、信息分析流程



**图2.1信息分析流程**

注：该图显示iTRAQ定量蛋白质组学的基本信息分析流程。首先对于质谱下机的原始文件，进行峰识别，得到峰列表。其次建立参考数据库，进行肽段及蛋白质的鉴定。对鉴定出的所有蛋白进行GO功能分类注释和KEGG通路注释；根据差异倍数和显著性P值对差异蛋白进行筛选，对差异蛋白进行分析，包括统计分析中的差异蛋白火山图和韦恩图，差异蛋白GO、KEGG注释和富集分析，表达模式聚类分析和差异蛋白互作网络分析；也可以对用户制定个性化分析，如显著性GO有向无环图，Ipath整合分析和PCA统计分析等。

# 三、本项目信息学分析内容

|  |  |  |  |
| --- | --- | --- | --- |
| **标准分析** | | **个性化分析** | |
| 数据质控 | √ | 显著性GO有向无环图 |  |
| 蛋白质鉴定 | √ | Ipath代谢通路整合分析 |  |
| GO功能分类注释 | √ | 多组样本PCA分析 |  |
| KEGG通路注释 | √ |  |  |
| COG功能分类注释 | √ |  |  |
| 蛋白差异分析（火山图+韦恩图） | √ |  |  |
| 差异蛋白GO分类统计 | √ |  |  |
| 差异蛋白KEGG通路可视化 | √ |  |  |
| 差异蛋白GO富集分析 | √ |  |  |
| 差异蛋白KEGG通路富集分析 | √ |  |  |
| 差异蛋白表达模式聚类分析 | √ |  |  |
| 蛋白互作网络分析(限物种) | √ |  |  |

注：见上表（打√部分）

# 四、生物信息分析

## 4.1全谱分析

将定量得到的全部蛋白进行功能注释，探究这些蛋白的生物学功能。

### 4.1.1 GO功能分类注释

GO (Gene Ontology, http://www.geneontology.org/) 是基因本体论联合会建立的将全世界所有与基因有关的研究结果进行分类汇总的综合数据库，其目的在于标准化不同数据库中关于基因和基因产物的生物学术语，对基因和蛋白功能进行统一的限定和描述。

利用 GO 数据库，可以将基因按照其参与的生物过程（Biological Process, BP）、细胞组分（Cellular Component, CC），分子功能（Molecular Function, MF）三个方面进行分类注释。在这三个大分支下面又分很多小层级（level），level级别数字越大，功能越细致。最顶层的三大分支视为level1，之后的分级依次为level2，level3和level4。因此GO注释有助于了解基因背后所代表的生物学意义。通过GO分类图，可以大致了解某个物种的全部基因产物的分类情况。

结果目录：**2.Annotation/2.1.GO**

各文件说明如下：

**GO.list：**针对每个蛋白，给出所有相应的GO功能的ID列表：

|  |  |
| --- | --- |
| **蛋白Accession号** | **对应的GO编号** |
| F0R446 | GO:0016491;GO:0055114 |
| R6FLS7 | GO:0016491;GO:0055114 |

**\*level2/3/4.xls：**记录GO二/三/四级分类的各个类型、术语和所有相应蛋白数目和蛋白名称，如下表：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **（GO注释分类的分支）**  **Term\_type** | **（GO分类的定义）**  **Term** | **（GO编号）**  **GO\_id** | **（该分类下的所有相关蛋白数目）**  **Number\_of\_protein** | **（该分类下的所有相关蛋白及对应的GO号）**  **Protein\_GO\_list** |
| molecular\_function | molecular transducer activity | GO:0060089 | 7 | R9I030(GO:0031992);B7AE02(GO:0031992);F0R189(GO:0031992);I9TW86(GO:0031992);E5WWM8(GO:0031992);R5CBD8(GO:0004871);U6R8Q0(GO:0031992) |
| molecular\_function | nucleic acid binding transcription factor activity | GO:0001071 | 1 | R6YH92(GO:0003700) |

**level2.go.txt.pdf：**GO二级分类统计条形图，如下图：

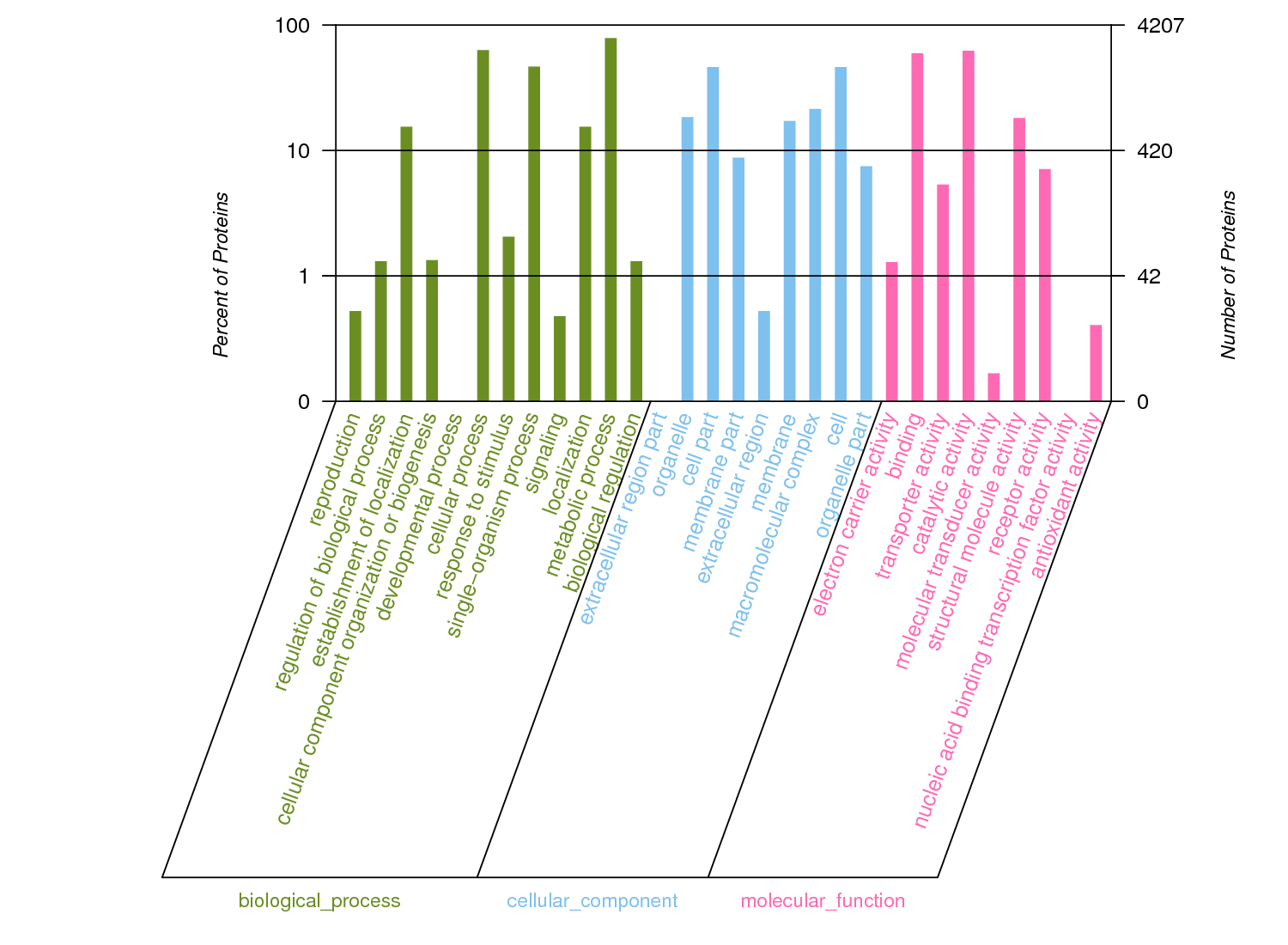


图4.1.1.1 GO二级分类统计图

(原始图片文件请查看2.Annotation/2.1.GO 文件夹)

注：图中每一个柱子表示一个GO的二级分类，柱条越高表示此二级分类的蛋白越多；横坐标表示GO的二级分类术语；纵坐标（左）表示包含在该二级分类中的蛋白占总数的百分比，纵坐标（右）表示该二级分类的蛋白数量；三个颜色表示三大类，其中绿色代表生物过程，蓝色代表细胞组分，红色代表分子功能。

**level234.pdf：**GO二级、三级、四级分类统计九饼图，如下图：

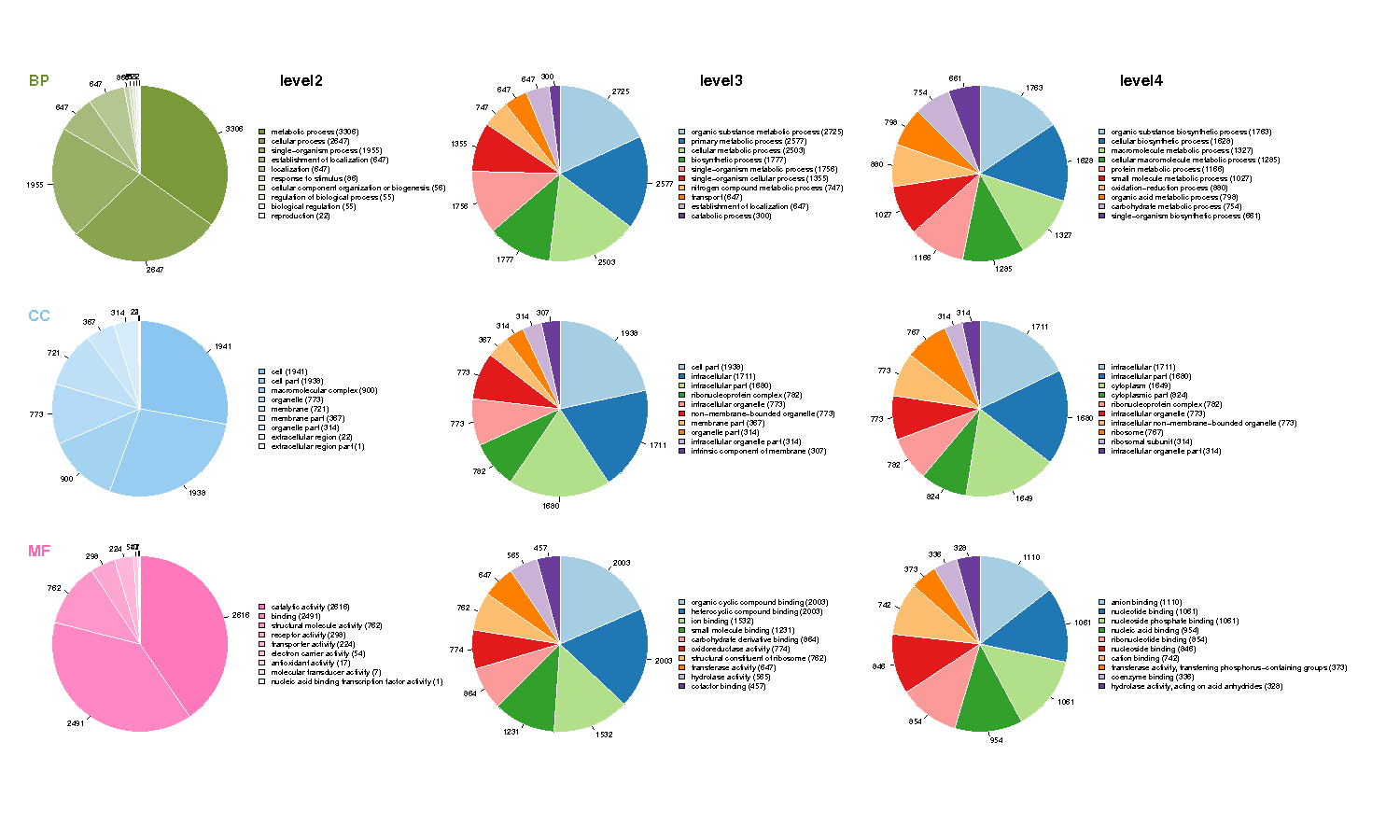


图4.1.1.2 GO二、三、四级分类统计图

(原始图片文件请查看2.Annotation/2.1.GO 文件夹)

注：图中每个饼图中的不同颜色代表了不同的GO Term，其面积表示该GO Term中蛋白所占的相对比例；此图包含9个饼图，从上至下共3排，分别表示GO的三大分支，即BP（生物过程，绿色饼）、CC（细胞组分，蓝色饼）、MF（分子功能，粉色饼）；从左至右共3列，分别表示GO三大分支下面的更详细的分类，即第2、3、4层（level2、3、4）。随着层数的增加，对BP、CC和MF的描述也就越详细，从level2 -> level3 -> level4该蛋白所具有的功能被越来越清晰的注释出来。

**参考文献：**

A. Conesa, S. Götz, J. M. Garcia-Gomez, J. Terol, M. Talon and M. Robles. "Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research", Bioinformatics, Vol. 21, September, 2005, pp. 3674-3676.

### 4.1.2 KEGG注释通路分析

KEGG（Kyoto Encyclopedia of Genes and Genomes，京都基因和基因组百科全书，http://www.genome.jp/kegg/）是基因组破译方面的公共数据库。该数据库是系统分析基因功能、联系基因组信息和功能信息的大型知识库，其中的基因组信息主要是从NCBI等数据库中获得的，包括完整和部分测序的基因组序列，存储于KEGG GENES数据库中；更高级的功能信息包括图形化的细胞过程如代谢、膜转运、信号传递、细胞周期等，还包括同系保守的子通路等信息，存储于KEGG PATHWAY数据库中；此外，关于化学物质、酶分子、酶化反应等相关的信息存储于KEGG LIGAND数据库中。

在生物体内，基因产物并不是孤立存在地作用的，不同基因产物之间通过有序的相互协调来行使其具体的生物学功能。因此，KEGG数据库中丰富的通路信息将有助于我们从系统水平去了解基因的生物学功能，例如代谢途径、遗传信息传递以及细胞过程等一些复杂的生物功能，这大大提高了该数据库在实际生产和应用中的价值。

结果目录：**2.Annotation/2.2.KEGG**

各文件说明如下：

**pathway.txt：**蛋白对应的KO编号，如下：

|  |  |
| --- | --- |
| **蛋白的Accession编号** | **对应的KO号** |
| F0R873 | K01610 |
| F0R5G0 | K03737 |

**pathway.top20.pdf：**包含蛋白数目最多的前20个通路（除以ko01开头的基础通路外），如下图：

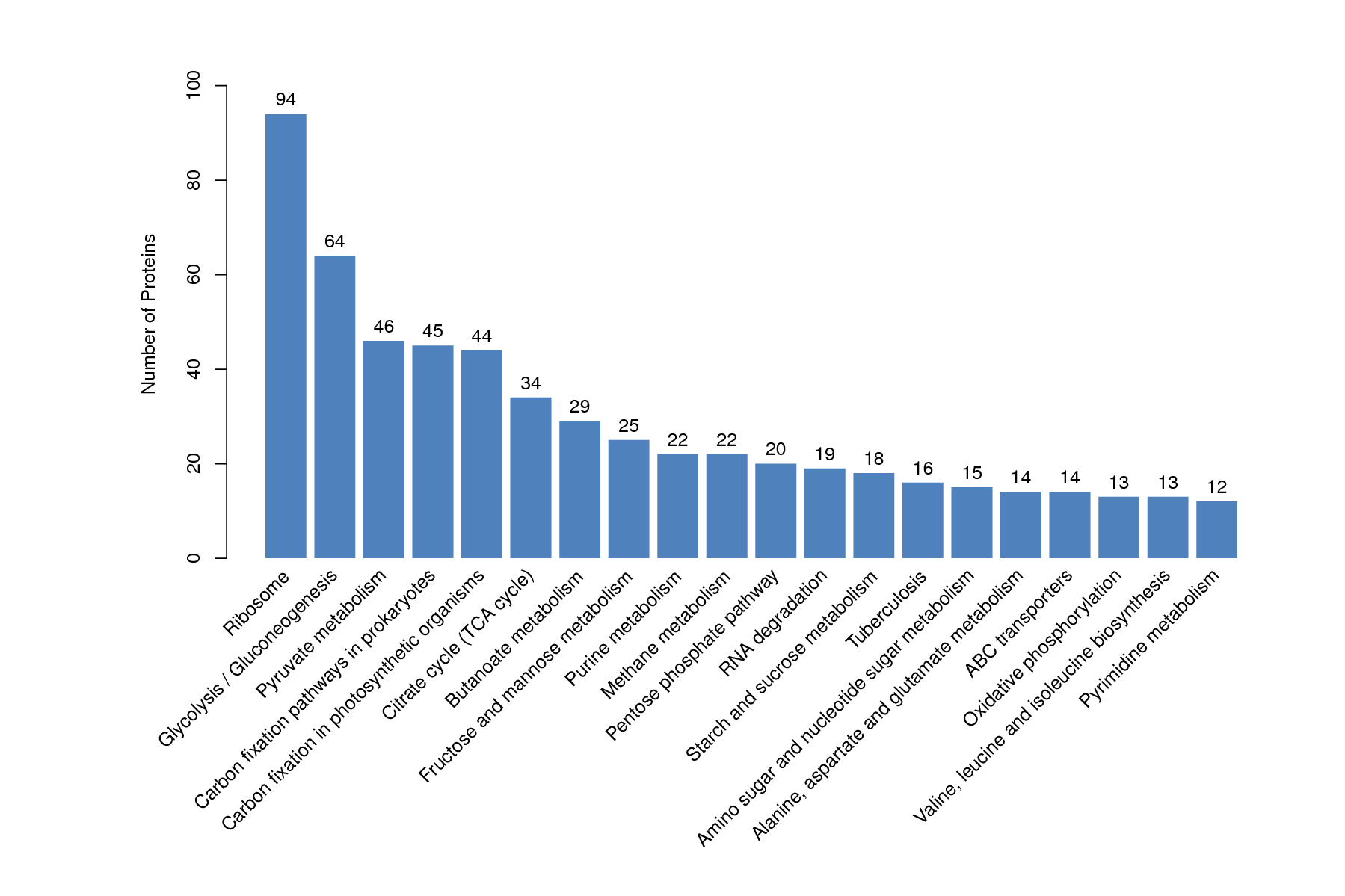


图4.1.2.1包含蛋白数目最多的前20个通路

(原始图片文件请查看2.Annotation/2.2.KEGG 文件夹)

注：从左至右按照包含蛋白数目从高到低依次排列，柱子越高表明该生物学通路在所测样本中越活跃，可以根据研究目的的不同选择排序靠前的通路进行后续深入分析。

**pathways文件夹下的pathway\_table.xls：**代谢通路分析结果说明文档，各列含义如下表：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **（通路的编号）**  **Pathway** | **（通路的定义）**  **Pathway\_definition** | **（蛋白数目）**  **Number\_of\_protein** | **（通路上所有蛋白和KO编号）**  **Protein\_ko\_list** | **（通路png图片名称）**  **Pathway\_image\_name** |
| path:ko00053 | Ascorbate and aldarate metabolism | 1 | F0R1J0(K03077); | ko00053.png |
| path:ko00061 | Fatty acid biosynthesis | 2 | D4LIK6(K02371);F0R5P5(K09458); | ko00061.png |

**pathways文件夹下的kegg\_table.xls：**代谢通路说明信息表，各列含义如下表：

|  |  |  |
| --- | --- | --- |
| **（蛋白名称）**  **Protein** | **（KO编号）**  **Ko\_id** | **（通路图片名称）**  **Paths** |
| F0R941 | K00688 | path:ko00500;path:ko01100;path:ko01110 |
| F0R1I9 | K01804 | path:ko00040;path:ko01100 |

**pathways文件夹下的.html文件：**pathway详细信息的网页介绍

**pathways文件夹下的.png文件：**KEGG通路图片展示，如下图：

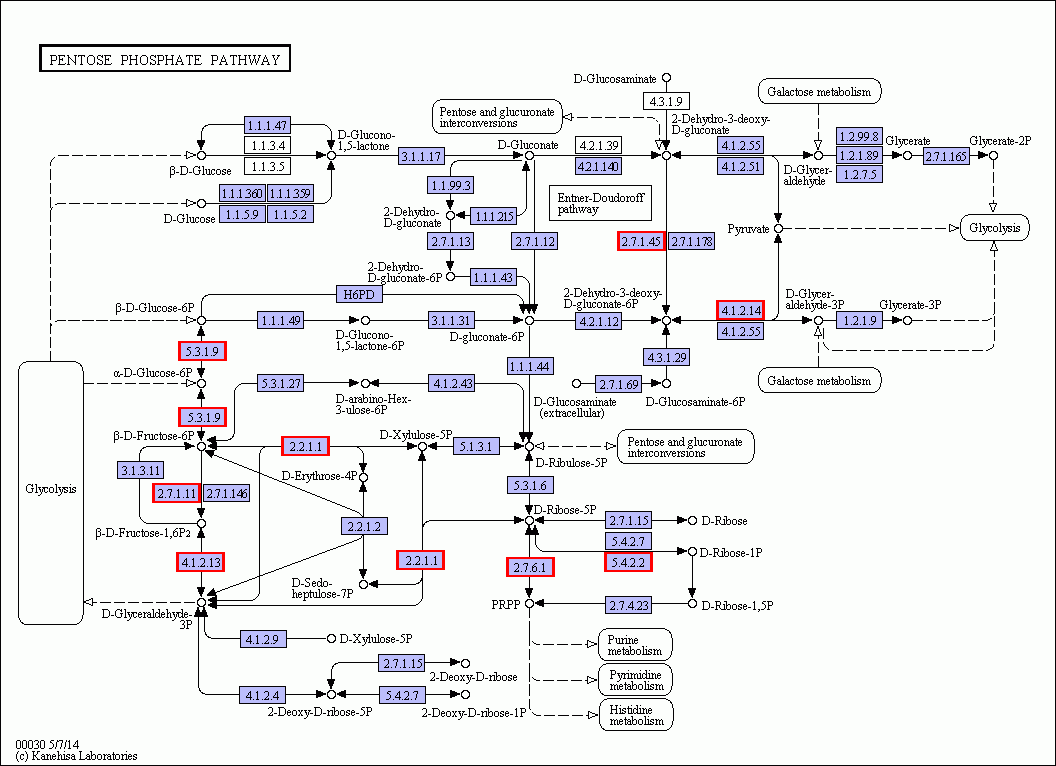


图4.1.2.2 KEGG通路数据库中的详细信息

(原始图片文件请查看2.Annotation/2.2.KEGG 文件夹)

注：图中长方形节点表示基因产物（如酶或一些RNA调节子）；所有蓝色背景的基因产物都属于KEGG ORTHOLOGY（KO）分类体系（序列高度相似，并在同一条通路上有相似功能的蛋白质被归为一组KO），而白色背景的基因产物则不在KO分类体系之列，红色表示本次研究中能够注释到这些基因产物上的蛋白（即认为具有与该节点基因产物相同或相似的功能）；圆形节点表示化合物（即底物或产物）；白色背景圆角长方形表示与本通路相关联的其他通路。箭头说明：酶反应方向或信息传递方向等；实线表示直接作用，虚线表示间接作用。详细说明请参见：[http://www.genome](http://www.genome.jp/kegg/document/help_pathway.html)[.jp](http://www.genome.jp/kegg/document/help_pathway.html)[/](http://www.genome.jp/kegg/document/help_pathway.html)[kegg](http://www.genome.jp/kegg/document/help_pathway.html)[/document/help\_pathway.html](http://www.genome.jp/kegg/document/help_pathway.html)。

**kegg\_classification.pdf：**对蛋白做KO注释后，可根据它们参与的KEGG代谢通路进行分类，如下图：

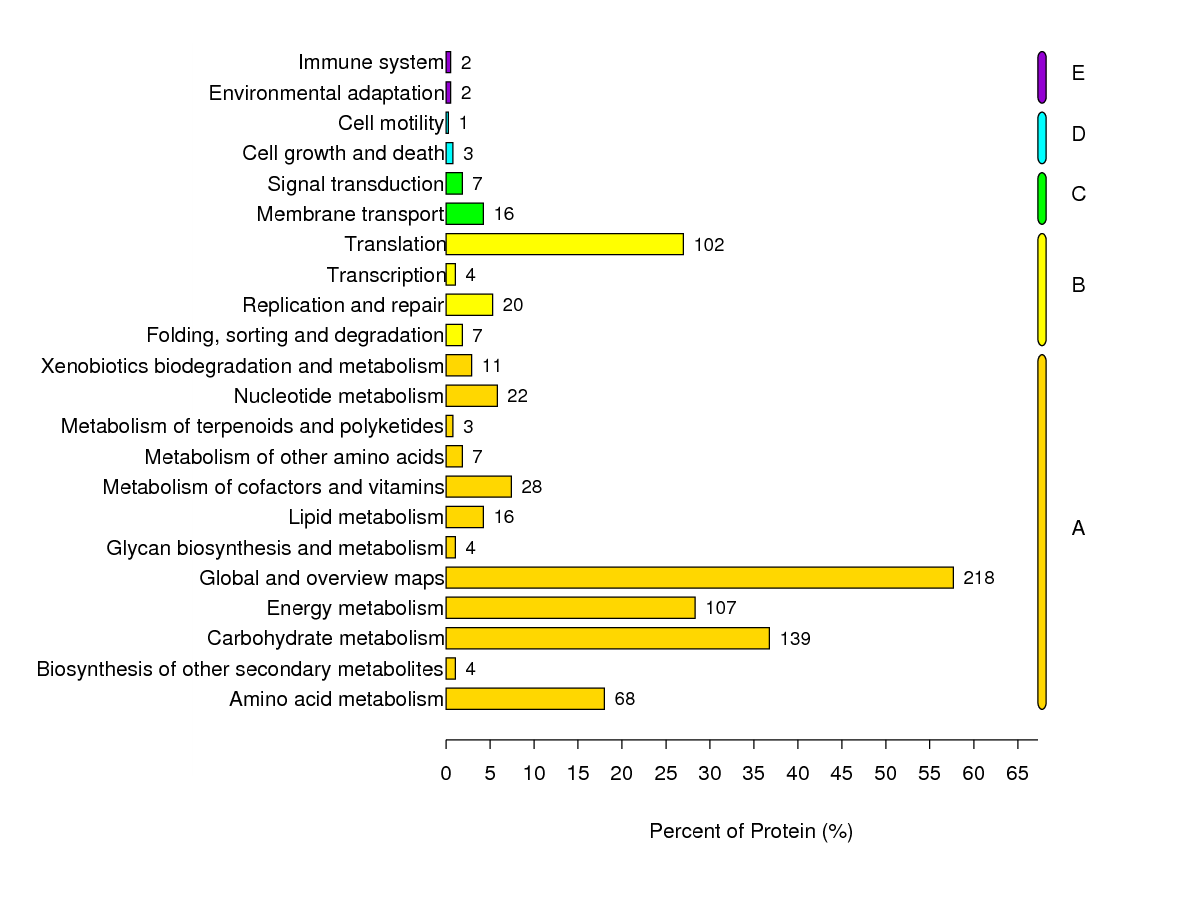


图4.1.2.3 KEGG注释统计

（原始图片文件请查看2.Annotation/2.2.KEGG 文件夹）

注：纵坐标为KEGG代谢通路的名称，横坐标为注释到该通路下的蛋白个数及其个数占被注释上的蛋白总数的比例。将蛋白根据参与的KEGG代谢通路分为5个分支：代谢（A，Metabolism），遗传信息处理（B，Genetic Information Processing）环境信息处理（C，Environmental Information Processing），细胞过程（D，Cellular Processes），有机系统（E，Organismal Systems）。

### 4.1.3 COG功能分类注释

COG，全称是Cluster of Orthologous Groups of proteins，由NCBI创建并维护的蛋白数据库，根据细菌、藻类和真核生物完整基因组的编码蛋白系统进化关系分类构建而成。通过比对可以将某个蛋白序列注释到某一个COG中，每一簇COG由直系同源序列构成，从而可以推测该序列的功能。

COG分为两类，一类是原核生物的，另一类是真核生物。原核生物的一般称为COG数据库；真核生物的一般称为KOG数据库。

结果目录：**.Annotation/2.3.COG**

各文件说明如下：

**COG.list：**蛋白对应的COG编号，如下：

|  |  |
| --- | --- |
| **蛋白Accession号** | **对应的COG号** |
| S0F728 | COG1014 |
| F0R873 | COG1866 |

**COG.annot.xls：**COG功能分类注释文档，如下：

|  |  |  |  |
| --- | --- | --- | --- |
| **（蛋白名称）**  **Accession** | **（COG编号）**  **COG** | **（COG所属功能分类）**  **Function** | **（COG功能具体描述）**  **Name** |
| S0F728 | COG1014 | [C] | Pyruvate:ferredoxin oxidoreductase and related 2-oxoacid:ferredoxin oxidoreductases, gamma subunit |
| F0R873 | COG1866 | [C] | Phosphoenolpyruvate carboxykinase (ATP) |

**COG.class.catalog.xls：**COG功能分类统计列表，如下：

|  |  |  |
| --- | --- | --- |
| **（COG 4种类型）**  **#Type** | **（COG功能分类，共25）**  **Functional\_categories** | **（该功能分类对应的COG数目）**  **COG** |
| INFORMATION STORAGE AND PROCESSING | [J] Translation, ribosomal structure and biogenesis | 272 |
| INFORMATION STORAGE AND PROCESSING | [K] Transcription | 26 |

**COG.class.catalog.pdf：**COG功能分类统计柱图，如下图：

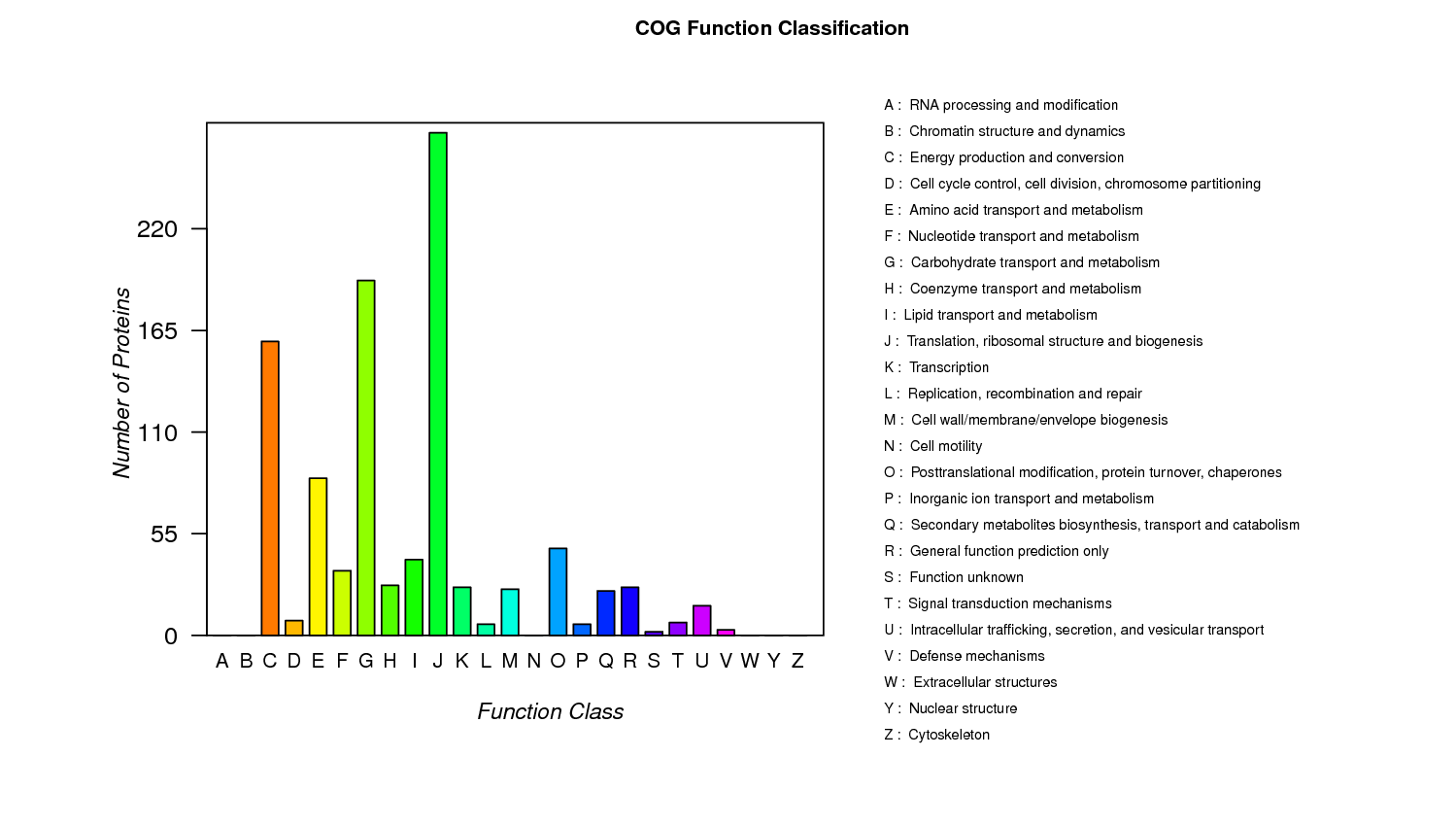


图4.1.3.1 COG功能分类统计柱图

（原始图片文件请查看2.Annotation/2.3.COG 文件夹）

## 4.2差异蛋白分析

### 4.2.1统计分析

使用R语言中t.test函数计算样本间差异显著性p值，本项目中，显著差异表达蛋白的筛选标准为：p＜0.05 &（FC＜0.83或 FC＞1.20）。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **名称** | **蛋白总数** | **差异蛋白数** | **上调蛋白数** | **下调蛋白数** |
| A\_vs\_B | 1557 | 163 | 125 | 38 |
| error | error | error | error | error |
| error | error | error | error | error |
| error | error | error | error | error |
| error | error | error | error | error |
| error | error | error | error | error |

#### 4.2.1.1火山图

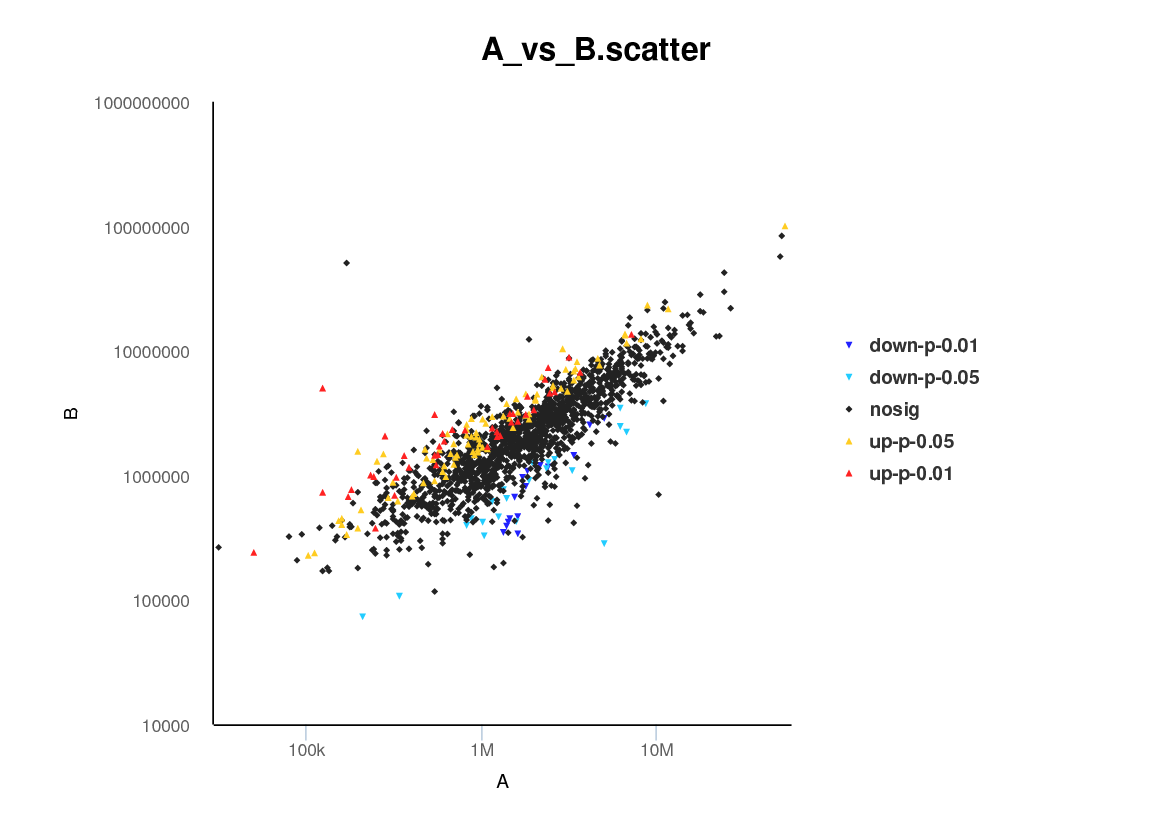
结果目录：**3.DiffExpAnalysis/3.1.Statistics/Volcano**

各文件说明如下：

**\*\_****vs\_\*.diff.exp.xls：**各分组样本所有蛋白差异表达详情，如下表：

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **（蛋白Accession编号）Accession** | **（样本1中该蛋白相对表达量均值）**  **Sample 1** | **（样本2中该蛋白相对表达量均值）**  **Sample 2** | **（该蛋白在两样本间差异倍数）FC(Sample1/ ample2)** | **（该蛋白在两样本间差异倍数的自然对数值）**  **Log2FC(Sample1/ Sample2)** | **（该蛋白在两样本间差异显著性检验结果P值）**  **Pvalue** | **（yes为显著，no为不显著）significant** | **(上下调情况，up为上调，down为下调)**  **regulate** |
| R6ID58 | 8986666.66666667 | 23300000 | 2.59272997032641 | 1.37447195908094 | 0.0310858677795758 | yes | up |
| B3JHG7 | 125100 | 5030000 | 40.2078337330136 | 5.32940470541443 | 2.32172917128381e-07 | yes | up |

**\*\_vs\_\*.scatter.pdf：**各分组样本差异蛋白可视化散点图，如下图：

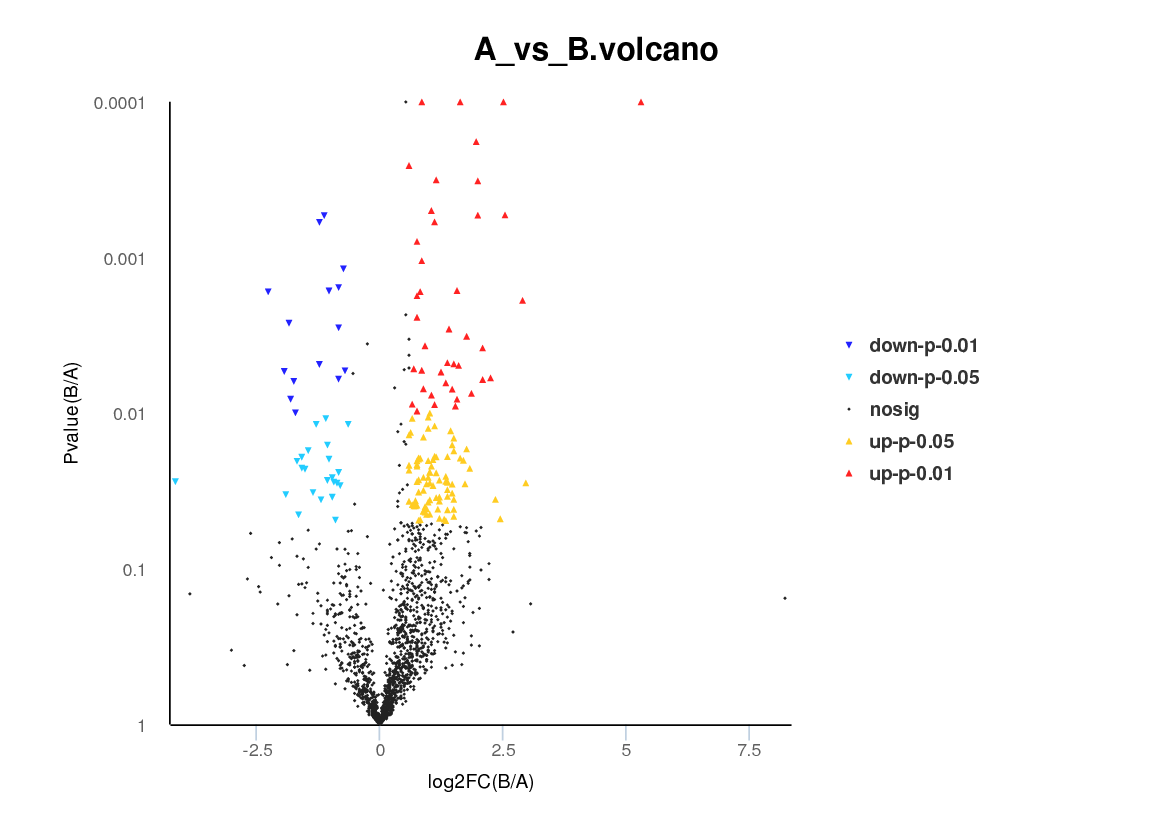


**图4.2.1.1差异蛋白可视化Scatter图**

**（原始图片文件请查看3.DiffExpAnalysis/3.1.Statistics/Volcano 文件夹）**

注：图中横纵坐标分别表示两个样本中蛋白的表达量，这里横纵坐标的数值都做了对数化处理，每个点代表一个特定的蛋白质，特定的一个点对应的横坐标值为该蛋白在样本1中的表达量，纵坐标值为该蛋白在样本2的表达量。图中红色点表示显著上调的蛋白，蓝色点表示显著下调的蛋白，黑色点为非显著差异蛋白；将所有蛋白质映射上去后，越接近 0 的点，说明表达量越低；那些偏离了对角线程度越大的点表明该蛋白在两个样本间表达差异越大。

**\*\_vs\_\*.volcano.pdf：**各分组样本差异蛋白可视化火山图，如下图：



**图4.2.1.2差异蛋白可视化Volcano图**

**（原始图片文件请查看3.DiffExpAnalysis/3.1.Statistics/Volcano 文件夹）**

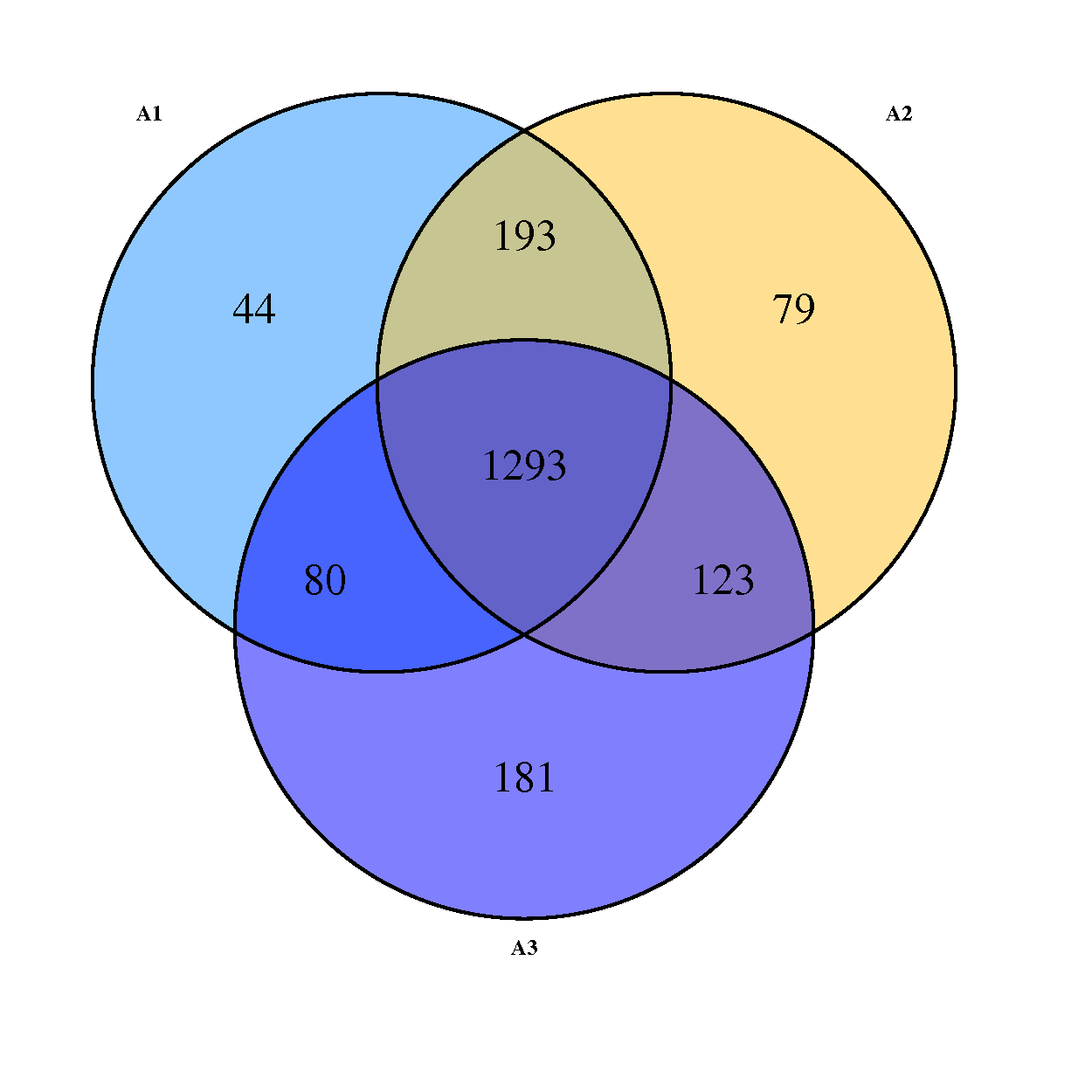
注：图中横坐标为蛋白在两个样本间差异的倍数变化值，即样本2的表达量除以样本1的表达量得到的数值，对此数值做了对数化处理；纵坐标为蛋白表达量变化差异的统计学t检验p值，p值越小则表达差异越显著。图中每个点代表一个特定的蛋白，黄色点表示在p<0.05条件下显著上调的蛋白，红色点表示在p<0.01条件下显著上调的蛋白，绿色点表示在p<0.05条件下显著下调的蛋白，蓝色点表示在p<0.01条件下显著下调的蛋白，黑色点为非显著差异蛋白；将所有蛋白映射上去之后，可以获知，在左边的点为表达差异下调的蛋白，右边的点为表达差异上调的蛋白，越靠左边和上边的点表达差异越显著。

#### 4.2.1.2韦恩图

结果目录为**3.DiffExpAnalysis/3.1.Statistics/Venn**，目录中各文件说明如下：

**\*.Venn.xls：**差异蛋白Venn图结果表格。

**\*.Venn.pdf：**差异蛋白Venn图，如下图：



**图4.2.1.2 差异蛋白韦恩图**

**（原始图片文件请查看 3.DiffExpAnalysis/3.1.Statistics/Venn文件夹）**

**参考文献：**

Chen H, Boutros PC. VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R.BMC Bioinformatics 2011;12:35.

### 4.2.2差异蛋白GO分析

#### 4.2.2.1差异蛋白GO分类

针对两两分组的差异表达蛋白进行GO注释的统计，以其中一个样本为对照，所得结果可绘制上下调蛋白GO注释柱形图。

结果目录：**3.DiffExpAnalysis/3.2.GO/Annotation**

各文件说明如下：

**\*gobars.xls：**上下调蛋白GO注释信息

**\*gobars.pdf：**上下调蛋白GO注释柱形图，如下图：

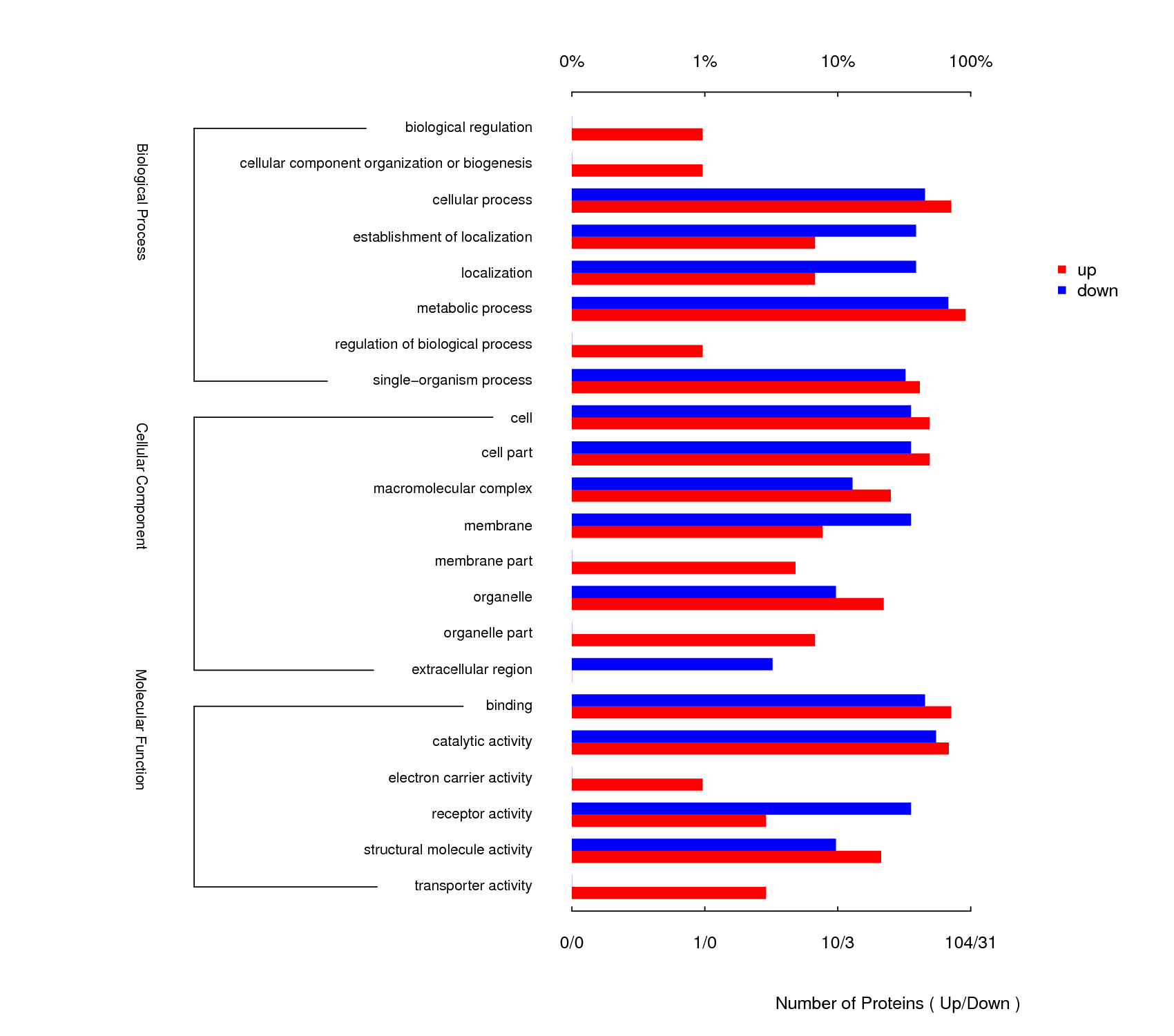


图4.2.2.1上下调蛋白GO注释柱形图

**（原始图片文件请查看3.DiffExpAnalysis/3.2.GO/Annotation文件夹）**

注：图中下方横坐标表示注释到某一GO term的蛋白个数，上方横坐标表示注释到某一GO term的蛋白数占所有GO注释蛋白总数的比例（蛋白和GO term是多对多的关系，即一个蛋白可包含多个GO term的注释，某一个GO term也会对应到多个蛋白，并不是一对一的关系）；纵坐标表示GO的每一详细分类，三个方块分别代表GO的三个二级分类，分别为生物过程（Biological Process）、细胞组分（Cellular Component）、分子功能（Molecular Function）；红色柱形代表上调蛋白，蓝色柱形代表下调蛋白。

#### 4.2.2.2差异蛋白GO富集分析

对差异蛋白进行GO功能显著性富集分析，可以说明差异蛋白的功能富集情况，在功能水平阐明样本间的差异。

本分析使用软件Goatools ([https://github.com/tanghaibao/GOatools](https://github.com/tanghaibao/goatools)) 进行富集分析，使用方法为Fisher精确检验。为控制计算的假阳性率使用Bonferroni检验方法 对p值进行了校正，通常情况下，当经过校正的p值（p\_bonferroni）≤0.05时，认为此GO功能存在显著富集情况。

假设对于一个GO功能A，如果注释到该功能的蛋白数量如下表所示：

|  |  |  |  |
| --- | --- | --- | --- |
|  | **差异蛋白数量** | **非差异蛋白数量** | **总和** |
| GO功能A | A | b | a + b |
| 其余GO功能 | C | d | c + d |
| 总和 | a + c | b + d | a + b + c + d (=n) |

Fisher精确检验理论指出得到这一组数据的概率可以由超几何分布计算，公式如下：



分析软件：Goatools ( <https://github.com/tanghaibao/Goatools> )

结果目录：**3.DiffExpAnalysis/3.2.GO/Enrichment**

各文件说明如下：

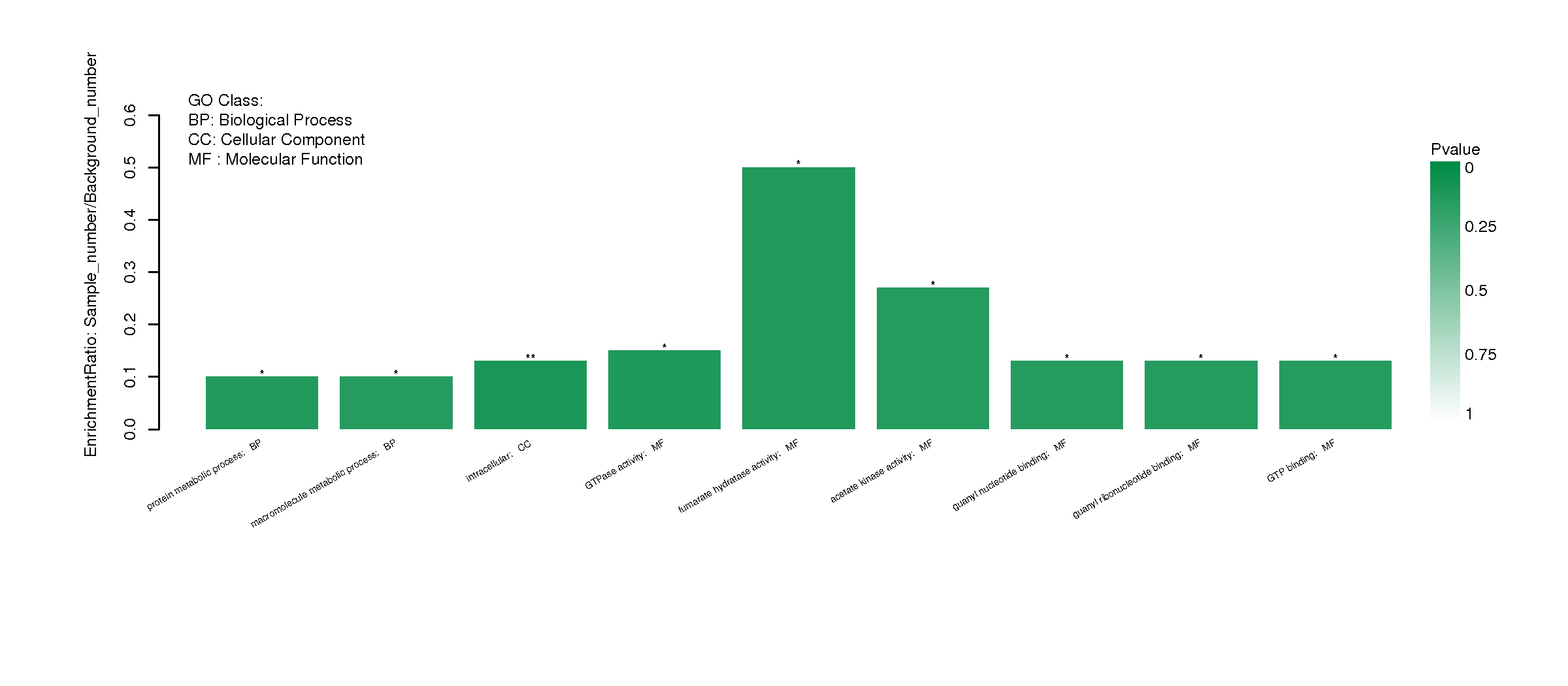
**\*.enrichment.detail.xls：**各分组差异蛋白GO功能显著性富集结果, 部分结果如下表：

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Id** | **Enrichment** | **Description** | **Ratio\_in\_study** | **Ratio\_in\_pop** | **P\_uncoorrected** | **P\_bonferroni** | **Type** | **Diff\_proteins** |
| GO:0004333 | e | fumarate hydratase activity | 2/163 | 4/2135 | 0.0314 | 1 | molecular\_function | F0R1G1;R5RUC2 |
| GO:0008776 | e | acetate kinase activity | 3/163 | 11/2135 | 0.0457 | 1 | molecular\_function | E5V7U8;F0R164;R6L429 |

表格中每列含义如下：

|  |  |
| --- | --- |
| **表头** | **描述** |
| id | GO号，GO数据库中针对每条GO功能给出的编号 |
| enrichment | 富集情况，“e”表示富集（enriched）， 代表此GO功能在目标蛋白（差异蛋白）中的富集程度显著高于在背景蛋白中的富集程度，即ratio\_in\_study大于ratio\_in\_pop的情况； “p”表示显著不富集（purified），代表此GO功能在目标蛋白（差异蛋白）中的富集程度显著低于在背景蛋白中的富集程度，即ratio\_in\_study小于ratio\_in\_pop的情况。 |
| description | GO详细注释，即该GO号所代表功能的详细解释 |
| ratio\_in\_study | 该GO信息在目标蛋白(差异蛋白)中占有的比例，分子为富集到该GO信息的差异蛋白数目，分母为本组差异蛋白的总数目。 |
| ratio\_in\_pop | 该GO信息在背景蛋白（检测到的所有蛋白）中占有的比例，分子为富集到该GO信息的背景蛋白的数目，分母为背景蛋白的总数目。 |
| p\_uncorrected | 未经校正的p值，p值代表富集出来的结果是否具有统计学上的显著意义，p值越小，在统计学上就越有显著意义，一般p值小于0.05认为该功能为显著富集项。 |
| p\_bonferroni | 使用bonferroni校正后的p值 |
| Type | 该GO信息的一级分类注释 |
| diff\_protein | 该GO信息的差异蛋白编号，即ratio\_in\_study中对应分子数目的蛋白编号 |

**\*.enrichment.detail.xls.go.pdf：**各分组差异蛋白GO功能富集分析柱状图，如下图：



**图4.2.2.2差异表达蛋白GO富集柱状图**

**（原始图片文件请查看3.DiffExpAnalysis/3.2.GO/Enrichment文件夹）**

注：图中每个柱子为一个GO term，横坐标文字表示GO的名称和分类，所属分类描述如左上角所示。柱子的高度即纵坐标表示富集率，计算公式如下：（EnrichmentRatio=SampleNumber/BackgroundNumber）。颜色表示富集的显著性即P值，颜色越深表示该GO term越显著富集，其中p<0.001的标记为\*\*\*，p<0.01的标记为\*\*，p<0.05的标记为\*，右侧颜色梯度表示P 值大小。

### 4.2.3差异蛋白KEGG分析

#### 4.2.3.1差异蛋白KEGG通路可视化

利用KEGG数据库，可将蛋白按照参与的pathway通路或行使的功能分类，并针对两两分组的差异蛋白，以其中一个样本为对照，可将差异蛋白显示在KEGG pathway通路图上，展示差异蛋白KEGG注释通路图。

结果目录：**3.DiffExpAnalysis/3.3.KEGG/Annotation/\*.diff.exp.xls.path/**

各文件说明如下：

**kegg\_table.xls：**代谢通路分析结果说明文档。

**pathway\_table.xls：**代谢通路说明信息表。

**\*.html：**差异蛋白pathway详细信息的网页介绍，将鼠标移至红色或绿色边框的蛋白可直接弹出该蛋白的上下调情况及蛋白编号等信息。

**\*.png：**差异蛋白KEGG通路图片展示，如下图：

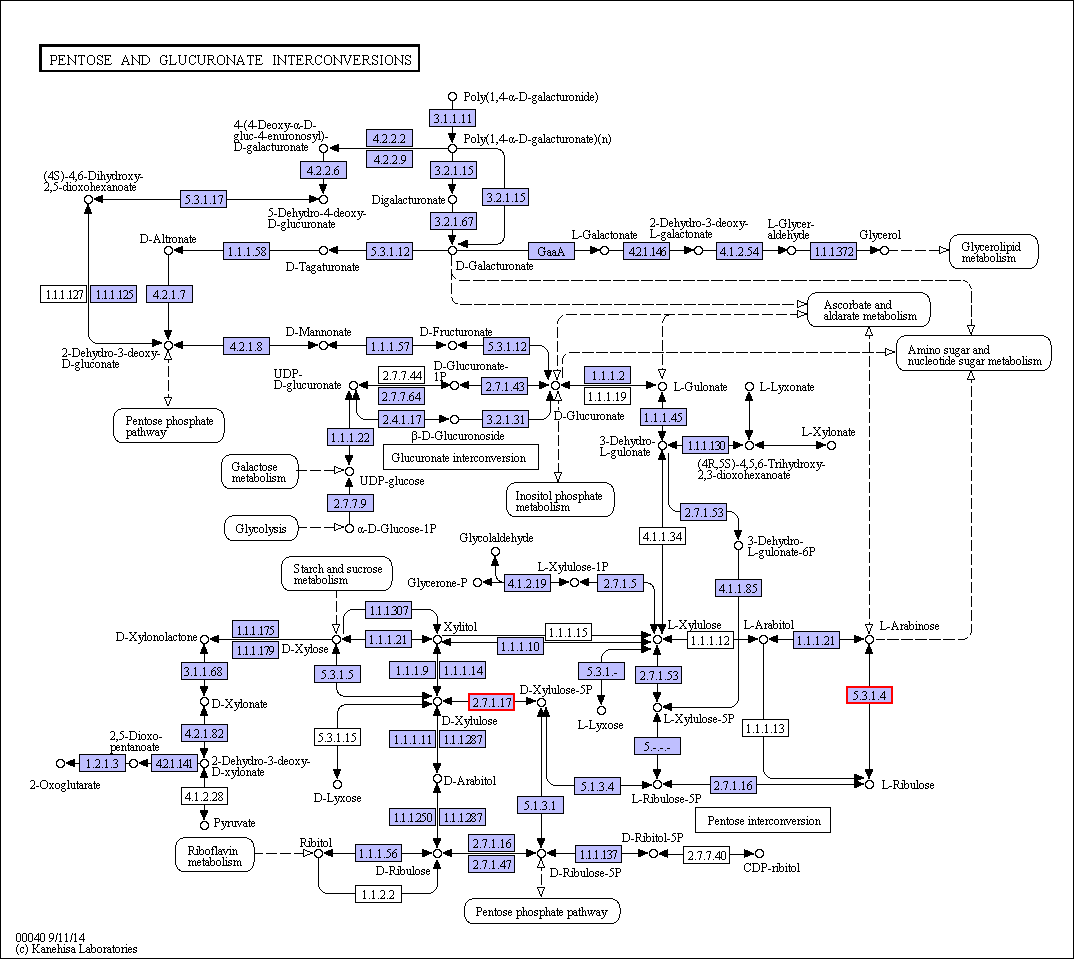


图4.2.3.1上下调蛋白KEGG注释通路图

**（原始图片文件请查看 3.DiffExpAnalysis/3.3.KEGG/Annotation/\*.diff.exp.xls.path文件夹）**

注：图中所有蓝色背景框的基因产物属于背景蛋白，白色背景框的基因产物表示不属于ko分类体系中。图中有红色/绿色边框的基因产物属于本次检测测到的差异蛋白，其中红色代表上调蛋白，绿色代表下调蛋白。

#### 4.2.3.2差异蛋白KEGG通路富集分析

富集分析方法通常是分析一组蛋白在某个功能节点上是否出现过，原理是由单个蛋白的注释分析发展为蛋白集合的注释分析。富集分析提高了研究的可靠性，能够识别出与生物现象最相关的生物学过程。

本次分析使用KOBAS( http://kobas.cbi.pku.edu.cn/home.do ) 进行KEGG PATHWAY富集分析, 计算原理同GO功能富集分析，使用Fisher精确检验进行计算。为控制计算假阳性率，采用BH(FDR) 方法进行多重检验，计算公式与上节相同，经过校正的p值（Corrected P-Value）以0.05为阈值，满足此条件的KEGG通路定义为在差异表达蛋白中显著富集的KEGG通路。

分析软件：KOBAS( http://kobas.cbi.pku.edu.cn/home.do )

相关参数：

-m f (statistic method, f is fisher exact test)，表明采用fisher精确检验法进行统计；

-n BH (false discovery rate (FDR) correction method:)，表明采用false discovery rate方法进行校验。

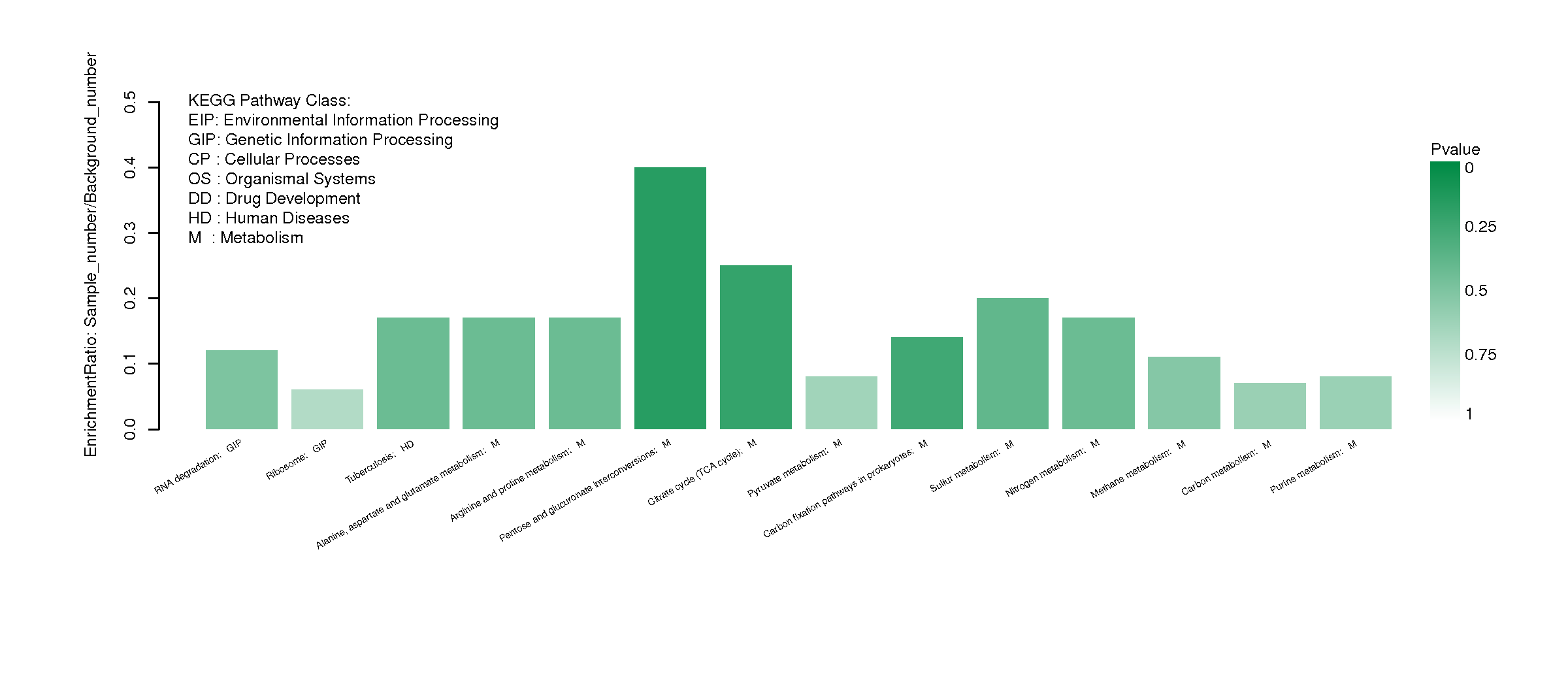
结果目录：**3.DiffExpAnalysis/3.3.KEGG/Enrichment**

文件说明如下：

**\*.pathway.xls：**各分组差异蛋白KEGG富集分析结果，部分表格见下表：

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **(KEGG pathway名称)**  **#Term** | **（数据库）Database** | **（KEGG通路ID）**  **Id** | **（该KEGG通路的差异蛋白数量）Sample number** | **（物种在该KEGG通路上的总蛋白数量）**  **Background number** | **（未经校正的p值，小于0.05任务该通路显著富集）**  **P-Value** | **（校正后的p值）Corrected**  **P-Value** | **（该KEGG通路的差异蛋白名）Proteins** | **（该KEGG通路的网页链接，包含该通路的详细信息描述，没有KEGG ID的通路显示“none”）**  **Hyperlink** |
| Citrate cycle (TCA cycle) | KEGG PATHWAY | ko00020 | 2 | 8 | 0.116480460206 | 0.62251021013 | F0R8U0|F0R1G1 | http://www.genome.jp/kegg-bin/show\_pathway?ko00020/K01676%09red/K00174%09red |
| Pentose and glucuronate interconversions | KEGG PATHWAY | ko00040 | 2 | 5 | 0.0606760629119 | 0.62251021013 | F0R2I5|F0R1I9 | http://www.genome.jp/kegg-bin/show\_pathway?ko00040/K00854%09red/K01804%09red |

**\*.pathway.pdf：**各分组差异蛋白KEGG pathway富集分析柱状图，如下图：



**图4.2.3.2 差异表达蛋白KEGG富集柱状图**

**（原始图片文件请查看 3.DiffExpAnalysis/3.3.KEGG/Enrichment 文件夹）**

注：图中每个柱子为一个通路，横坐标文字表示通路的名称和分类，所属分类描述如左上角所示。柱子的高度即纵坐标表示富集率，计算公式如下：（EnrichmentRatio=SampleNumber/BackgroundNumber）。颜色表示富集的显著性即P值，颜色越深表示该通路越显著富集，其中Pvalue<0.001的标记为\*\*\*，Pvalue<0.01的标记为\*\*,Pvalue<0.05的标记为\*，右侧颜色梯度表示P值大小。

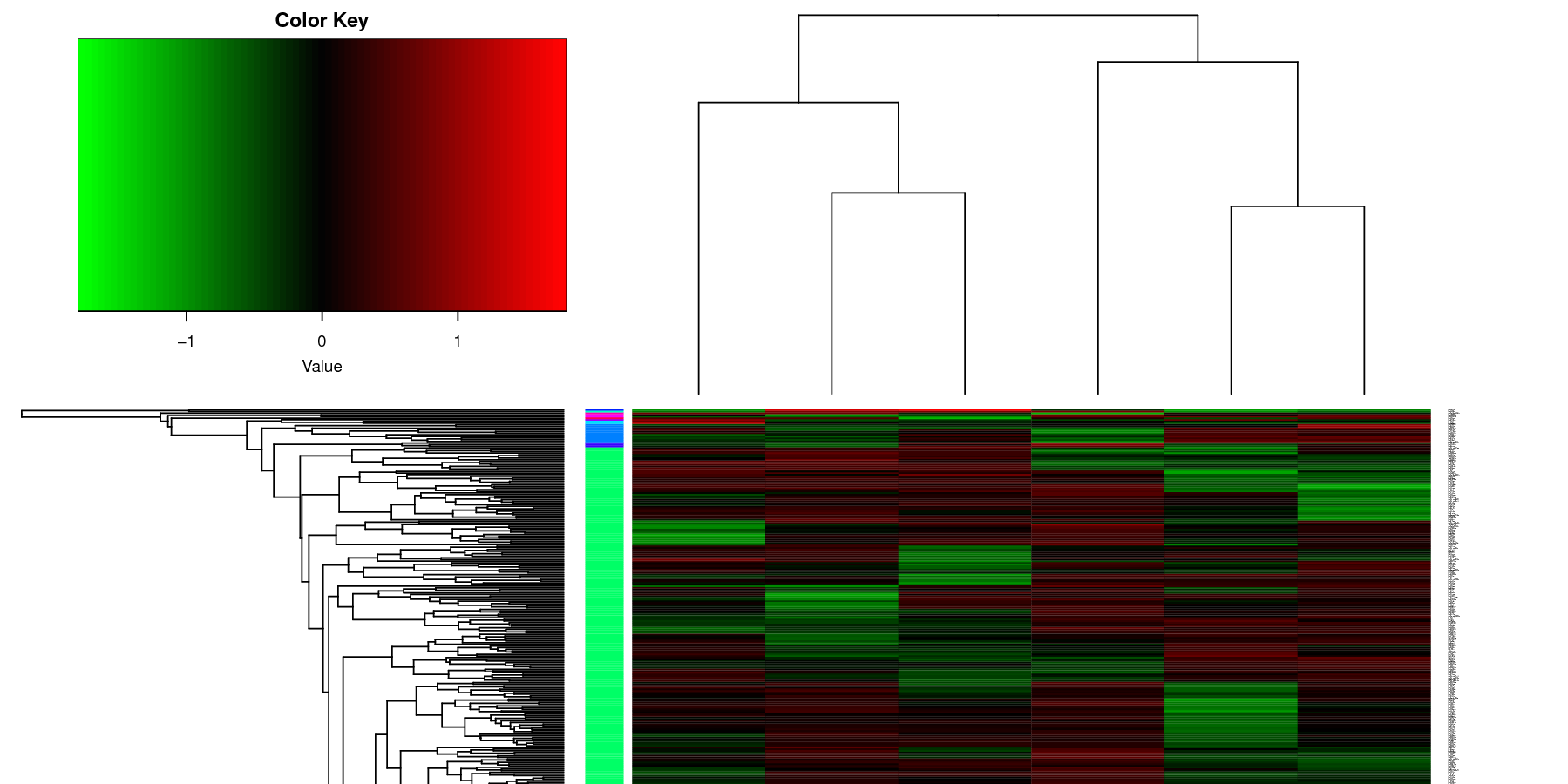
### 4.2.4差异蛋白表达模式聚类分析

用R语言中gplot包将有显著差异的蛋白进行表达模式聚类分析，采用距离计算算法：样本间为spearman，基因间为pearson，采用的聚类方法为hcluster（complete算法）。

结果目录：**3.DiffExpAnalysis/3.4.Cluster**

文件说明如下：

**Heatmap.pdf：**差异蛋白表达模式聚类图，如下图：



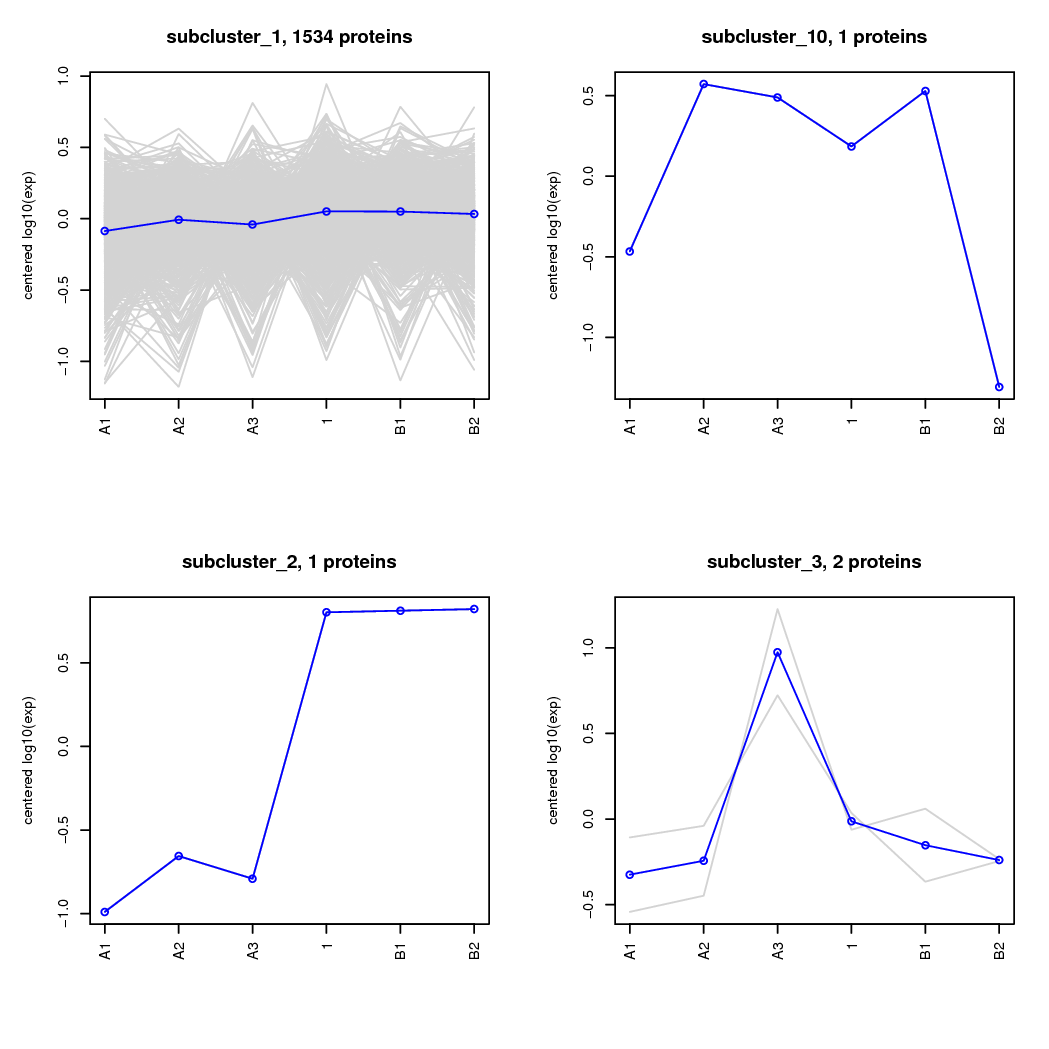
**Fig 4.2.5.1 Heatmap**

**（原始图片文件请查看3.DiffExpAnalysis/3.4.Cluster文件夹）**

注：图中每列表示一个样本，每行表示一个蛋白，图中的颜色表示蛋白在该组样本中相对表达量的大小，红色代表该蛋白在该样本中表达量较高，绿色代表表达量较低，具体表达量大小变化趋势请见左上方颜色条下的数字标注。左侧为蛋白聚类的树状图，右侧为蛋白的名称，两个蛋白分支离得越近，说明它们的表达量越接近；上方为样本聚类的树状图，下方为样本的名称，两个样本分支离的越近，说明这两个样本所有蛋白的表达模式越接近，即蛋白表达量变化趋势越接近。

**subclusters\_\*/：**差异蛋白模块（clusters）表达模式结果表。

**Heatmap\_trendlines\_for\_\*\_subclusters.pdf：**差异蛋白模块（clusters）表达趋势折线图，如下图：



**Fig 4.2.5.2 Cluster**

**（原始图片文件请查看3.DiffExpAnalysis/3.4.Cluster文件夹）**

注：图中为每个子cluster的表达趋势折线图，横坐标为各比较样本组，纵坐标为蛋白在该组样本中的表达量。图中每一条线表示一个蛋白，蓝色的线表示该子cluster中所有蛋白的表达量平均值；每张图展示一种类型的表达模式，即体现这组蛋白表达量变化的趋势。

**参考文献：**

Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, Simonovic M, Roth A, Santos A, Tsafou KP, Kuhn M, Bork P, Jensen LJ, von Mering C. STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res. 2015 Jan; 43:D447-52.

### 4.2.5差异蛋白互作网络分析

STRING数据库( http://string-db.org/ )是一个搜寻已知蛋白质之间和预测蛋白质之间相互作用的数据库。这种相互作用既包括蛋白质之间直接的物理作用，也包括蛋白质之间间接的功能相关性。它除了包含有实验数据、从PubMed摘要中文本挖掘的结果和综合其他数据库数据外，还利用生物信息学的方法预测的结果，包括：染色体临近、基因融合、系统进化和基于芯片数据的基因共表达。该系统利用一个打分机制对这些不同方法得来的结果给与一定的权重，最终给出一个综合的得分。  
 结果目录：**3.DiffExpAnalysis/3.5.Network**

各文件说明如下：

**\*\_vs\_\*.network.xls：**两组样本之间差异蛋白的蛋白互作关系表格，如下表：

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **node1** | **node2** | **node1\_Accession\_id** | **node2\_Accession\_id** | **neighborhood\_on\_chromosome** | **gene\_fusion** | **phylogenetic\_cooccurrence** | **homology** | **coexpression** | **experimentally\_determined\_interaction** | **database\_annotated** | **automated\_textmining** | **combined\_score** |
| F5 | Pros1 | O88783 | Q08761 | 0 | 0 | 0 | 0 | 0 | 0.381 | 0.961 | 0.818 | 0.995 |
| Apoc3 | Apoe | E9QP56 | Q3TXU4 | 0 | 0 | 0 | 0 | 0.111 | 0 | 0.900 | 0.953 | 0.995 |

表中各列含义如下：

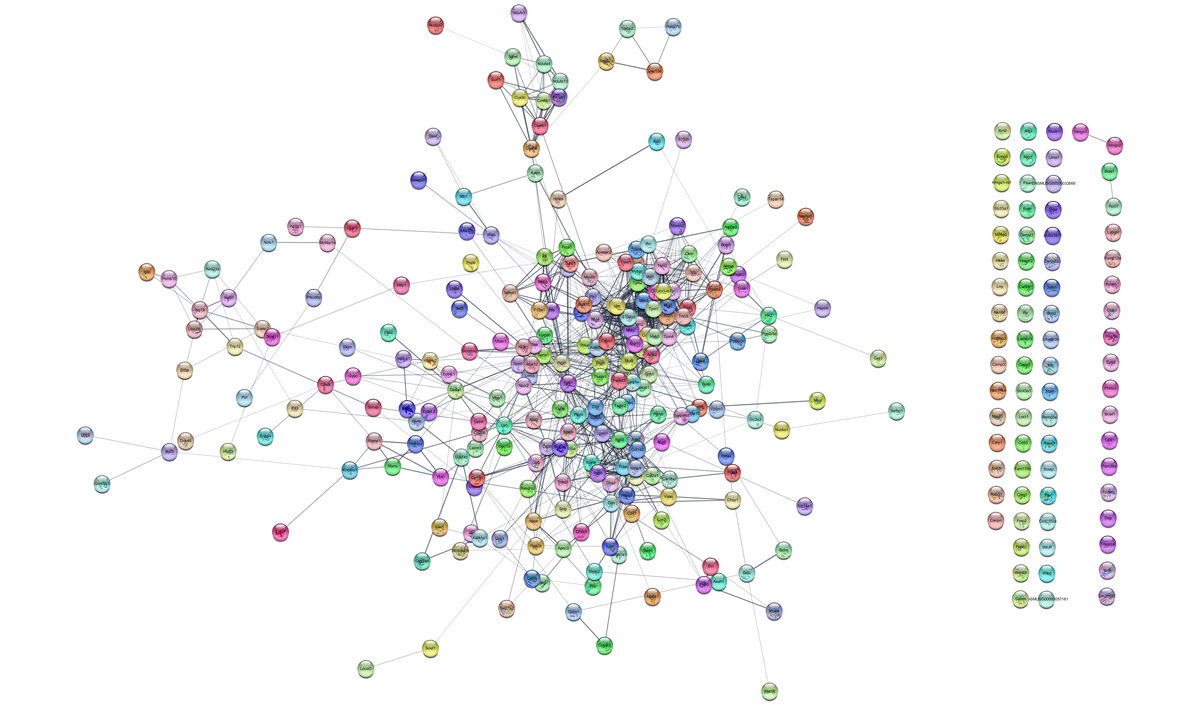
|  |  |  |
| --- | --- | --- |
| **列名** | **含义** | **举例** |
| node1 | 互作网络图中蛋白1的标签 | AGPS1 |
| node2 | 互作网络图中蛋白2的标签 | SS2 |
| node1\_Accession\_id | 蛋白1的uniprot登陆号 | P23Q14 |
| node2\_Accession\_id | 蛋白2的uniprot登陆号 | P452S1 |
| neighborhood\_on\_chromosome | 染色体上的临近关系得分值 | 0.6 |
| gene\_fusion | 基因融合得分值 | 0 |
| phylogenetic\_cooccurrence | 系统共发生得分值 | 0.505 |
| homolog | 同源性得分值 | 0 |
| coexpression | 共表达得分值 | 0.648 |
| experimentally\_determined\_interaction | 实验数据得分值 | 0 |
| database\_annotated | 数据库注释得分值 | 0.976 |
| automated\_textmining | 文本挖掘得分值 | 0.538 |
| combined\_score | 综合得分值 | 0.999 |

**\*\_vs\_\*.annotation.xls：**两组样本之间差异蛋白的string注释结果，如下表：

|  |  |  |  |
| --- | --- | --- | --- |
| **node**  **(蛋白名)** | **node\_Accession\_id**  **（蛋白uniprot登录号）** | **Annotation**  **（蛋白注释）** | **domain\_summary\_url**  **（结果与url链接）** |
| Apoh | Q01339 | apolipoprotein H; Binds to various kinds of negatively charged substances such as heparin, phospholipids, and dextran sulfate. May prevent activation of the intrinsic blood coagulation cascade by binding to phospholipids on the surface of damaged cells | http://smart.embl.de/smart/DD2.cgi?smart=345:SIGNAL(1|19)+CCP(23|79)+CCP(84|137)+CCP(142|200)+CCP(205|260)+CCP(264|325)+ |
| Alox12 | A2CF85 | arachidonate 12-lipoxygenase; Oxygenase and 14,15-leukotriene A4 synthase activity | http://smart.embl.de/smart/DD2.cgi?smart=663:LH2(2|111)+Pfam\_Lipoxygenase(146|655)+ |

**\*\_vs\_\*.****network.svg：**两组样本之间差异蛋白的蛋白互作图（svg格式）

**\*\_vs\_\*.network.png：**两组样本之间差异蛋白的蛋白互作图（png格式），如下图：



**Fig 4.2.6 Network**

**（原始图片文件请查看3.DiffExpAnalysis/3.5.Network文件夹）**

注：图中每个节点表示一个蛋白，大节点表示三维结构已知或者已预测的蛋白质，小节点表示三维结果未知的蛋白质；每条连线表示蛋白之间的相互作用关系，连线宽度表示蛋之间互作关系的综合得分大小（0.4-1），连线越宽，得分越高，连线越窄，得分越低；其中，0.15表示低得分，0.4表示中等得分，0.7表示高得分，0.9表示最高得分，本次研究选择默认得分最小值为中等得分（0.4）。

**参考文献：**

Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, Simonovic M, Roth A, Santos A, Tsafou KP, Kuhn M, Bork P, Jensen LJ, von Mering C. STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res. 2015 Jan; 43:D447-52.

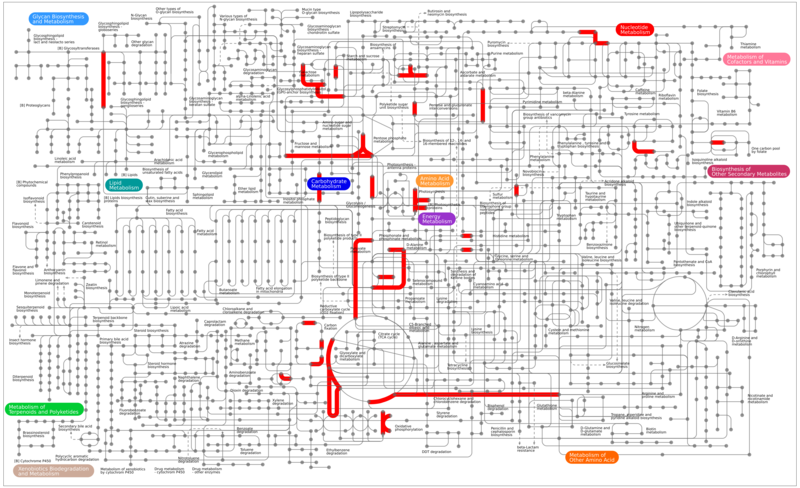
### 4.2.6 Ipath整合分析

利用 iPath2.0 （http://pathways.embl.de）对代谢途径进行可视化分析，可以查看整个生物系统的代谢通路信息。iPath2.0 能够概述次生代谢物的生物合成和重要的调控途径，轻松查找复杂的代谢通路。  
 结果目录：**3.DiffExpAnalysis/3.6.Ipath**

各文件说明如下：

**\*\_vs\_\*.Ipath.svg：**Ipath整合通路图（svg格式）

**\*\_vs\_\*.Ipath.png：**Ipath整合通路图（png格式），如下图：



**图4.2.6 Ipath整合分析图**

**（原始图片文件请查看3.DiffExpAnalysis/3.6.Ipath文件夹）**

注：每组差异蛋白可以绘制一张图，图中的节点代表不同的化合物，边界代表不同的酶促反应，其中红色的线代表在本次检测到的差异蛋白参与的反应，不同线型代表不同类型的代谢通路或功能。

# 五、附录

## 5.1 结果文件列表

**|-- 1.QualityControl/ 数据质控目录**

| |--dMass.pdf 肽段匹配误差分布图

| |--Peptide\_number\_distribution.pdf 肽段数量分布柱状图

| |--Peptide\_number\_distribution.xls 肽段数量分布表格

| |--Peptide\_length\_distribution.pdf 肽段长度分布柱状图

| |--Peptide\_length\_distribution.xls 肽段长度分布表格

| |--Protein\_molecular\_weight\_distribution.pdf 蛋白分子量分布柱状图

| |--Protein\_molecular\_weight\_distribution.xls 蛋白分子量分布表格

**|-- 2.Annotation/ 功能注释结果目录**

**| |--2.1.GO/ GO功能分类注释结果目录**

| | |--GO.list 记录每个蛋白所对应的GO编号

| | |--\*level2/3/4.xls 记录GO二/三/四级分类的各个类型、术语和所有相应蛋白数和蛋白名

| | |--level2.go.txt.pdf GO分类统计条形图

| | |--level234.pdf GO分类统计九饼图

**|** **|--2.2.KEGG/ KEGG注释通路分析结果目录**

| | |--pathway.txt 蛋白对应的KO编号

| | |--pathway.top20.pdf 包含蛋白数目最多的前20 个通路柱图

| | |--kegg\_classification.pdf 根据蛋白参与的代谢通路将其分为5类的注释统计图

**| | |--pathways/**

| | | |--pathway\_table.xls 代谢通路分析说明文档

| | | |--kegg\_table.xls 代谢通路说明信息表

| | | |--\*.html pathway详细信息的网页介绍

| | | |--\*.pngKEGG通路图片

**| |--2.3.COG/ COG功能注释结果目录**

| | |--COG.list 蛋白对应的COG编号

| | |--COG.annot.xls COG功能分类注释文档

| | |--COG.class.catalog.xls COG功能分类统计列表

| | |--COG.class.catalog.pdf COG功能分类统计柱图

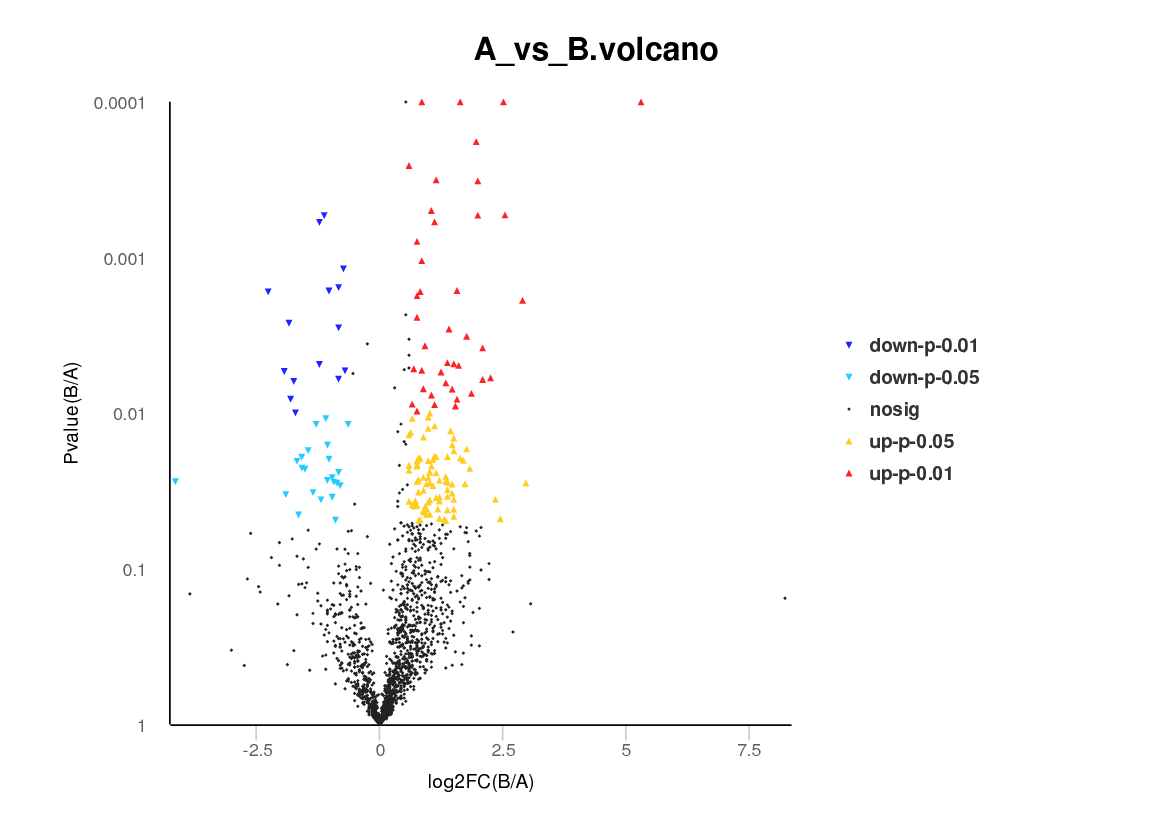
**|-- 3.DiffExpAnalysis/ 表达差异分析结果目录**

**| |--3.1.Statistics 差异表达蛋白统计情况**

**| | |--Volcano 火山图结果目录**

| | |--\*\_vs\_\*/\*.diff.exp.xls 各分组样本所有蛋白差异表达详情

| | |--\*\_vs\_\*/\*.volcano.pdf 各分组样本差异蛋白可视化火山图

**Venn 韦恩图结果目录**

| | |--\*Venn.xls 差异蛋白Venn图结果表格

| | |--\*Venn.pdf 差异蛋白Venn图

**| |--3.2.GO/ 差异蛋白GO分析结果目录**

**| | |--Annotation/ 差异蛋白GO分类统计结果目录**

| | |--\*-gobars.pdf 上下调蛋白的GO注释柱形图

| | |--\*-gobars.xls 上下调蛋白GO 注释信息

| | |--level234.pdf 差异蛋白GO分类统计九饼图

**| | |--Enrichment/ 差异蛋白GO富集分析结果目录**

| | | |--\*\_vs\_\*.DE.list.enrichment.detail.xls 样本间差异蛋白GO富集分析结果

| | | |--\*\_vs\_\*.DE.list.enrichment.detail.xls.go.pdf 样本间差异蛋白GO富集分析柱状图

**| |--3.3.KEGG/ 差异蛋白KEGG分析结果目录**

**| | |--Annotation/ 差异蛋白KEGG通路注释结果目录**

| | | |--\*\_vs\_\*.path/kegg\_table.xls 代谢通路分析结果说明文档

| | | |--\*\_vs\_\*.path/pathway\_table.xls 代谢通路说明信息表

| | | |--\*\_vs\_\*.path/\*.html pathway详细信息网页介绍

| | | |--\*\_vs\_\*.path/\*.png 差异蛋白的代谢通路图

**| | |--Enrichment/ 差异蛋白KEGG富集分析结果目录**

| | | |--\*\_vs\_\*.DE.list.kegg\_enrichment.xls.pathway.xls 样本间差异蛋白KEGG富集分析结果

| | | |--\*\_vs\_\*.DE.list.kegg\_enrichment.xls.pathway.pdf 样本间差异蛋白KEGG富集分析柱状图

**| |--3.4.Cluster/ 表达模式聚类分析结果目录**

| | |--Heatmap.pdf 差异蛋白表达模式聚类图

| | |--Heatmap\_trendlines\_\*\_subclusters.pdf 差异蛋白模块（clusters）表达趋势折线图

| | |--subclusters\_\*/ 差异蛋白表达模式聚类模块表达量结果

**| |--3.5.Network/ 差异蛋白互作网络分析目录**

| | |--\*\_vs\_\*.annotation.xls 样本间差异蛋白的string注释结果表格

| | |--\*\_vs\_\*.network.xls 样本间差异蛋白的互作关系表格

| | |--\*\_vs\_\*.network.png 样本间差异蛋白互作网络png图

**| |--3.6.Ipath/ 差异蛋白Ipath整合通路分析结果**

| | |--\*\_vs\_\*.mapped.xls 匹配到通路上的蛋白登录号

| | |--\*\_vs\_\*.Ipath.png Ipath整合通路png图

## 5.2常用数据库及专用名词介绍

1）GO(Gene Ontology, http://www.geneontology.org/) 是基因本体论联合会建立的数据库，其目的在于标准化不同数据库中的关于基因和基因产物的生物学术语，对基因和蛋白功能进行限定和描述。利用 GO 数据库，可以将基因按照它们参与的生物学过程、构成细胞的组分、实现的分子功能等进行分类。

GO功能包括3大方面：

* Cellular component(GO:0005575)：the parts of a cell or its extracellular environment，such as nucleus(GO:0005634),ribosome(GO:0005840);
* Molecular function(GO:0003674)：the elemental activities of a gene product at the molecular level, such as binding(GO:0005488),catalysis(GO:0043364);
* Biological process(GO:0008150) ：operations or sets of molecular events with a defined beginning and end, pertinent to the functioning of integrated living units: cells, tissues, organs, and organisms, such as cellular physiological process(GO:0009987),signal transduction(GO:0007165)。

2）KEGG 库(Kyoto encyclopedia of genes and genomes 数据库)：京都基因和基因组百科全书，是系统分析基因功能、联系基因组信息和功能信息的知识库。基因组信息存储在GENES 数据库里，包括完整和部分测序的基因组序列；更高级的功能信息存储在PATHWAY 数据库里，包括图解的细胞生化过程如代谢、膜转运、信号传递、细胞周期，还包括同系保守的子通路等信息；KEGG 的另一个数据库是LIGAND，包含化学物质、酶分子、酶反应等信息；KEGG还有个数据库是DISEASE，提供和人类疾病有关的的信息，整合了疾病的遗传和环境影响因素等信息，包括已知疾病相关基因，环境影响因子，诊断标记，对应治疗药物等。

3）STRING（Search Tool for the Retrieval of Interacting Genes）：是欧洲分子生物学实验室建立的数据库，这个数据库可用于预测蛋白质的相互作用，它利用了COG数据库的信息对蛋白进行直系同源聚类。通过blast比对String数据库，可以得到蛋白编码基因的COG分类信息，一个COG表示一组直系同源蛋白。

## 5.3文件解压缩方法

所有提供的文件均为Linux系统下的文件，压缩包使用linux下的命令压缩成\*.tar.gz格式，以下为不同系统用户解压缩的方法：

Unix/Linux/Mac用户: 使用 tar Linux/Mac缩的方法：点或者 gzip inux/Mac命令

Windows用户：推荐使用WinRAR软件解压缩

## 5.4文件打开或浏览方法

如果在本附录中无特殊说明，所有提供的文件均为Linux系统下文本文件，Unix/Linux用户可以使用more或less命令查看文本文件内容。对于Windows用户，一般文本文件（后缀名为.txt或者.list等）可以使用写字板或excel打开。推荐使用开源文本编辑器gedit for win32版本(http://projects.gnome.org/gedit/)或者商业文本编辑器UltraEdit。当文件比较大时，打开文件可能导致Windows系统死机，建议使用性能较好的计算机或者使用更适合处理大量数据的Unix/Linux系统打开。

数据中可能包含部分图像文件，一般图像文件后缀名为.png、.jpg、.gif等，对于图像文件，Windows用户可以使用图片浏览器打开，Linux/Unix用户使用display命令打开。

后缀名为.svg的文件为文本格式描述的图像文件，Windows用户需要安装Adobe Illustrator软件打开。Linux/Unix用户可以使用rsvg-view命令查看。公司默认提供矢量图，可利用"Adobe Illustrator"软件对该格式图片进行编辑。后缀名为.tsv的文件可以用excel软件查看具体信息。

Linux下的表格均为制表符(Tab)分割的文本，为了便于阅读，建议使用excel或openoffice等办公软件用表格形式打开，打开时请用“办公软分割”方式。

## 5.5 联系方式

|  |  |
| --- | --- |
|  | **上海总部** |
| 地址： | 上海市浦东新区康新公路3399号时代医创园3号楼 |
| 电话： | 400-624-3721 |
| E-mail： | protein@majorbio.com |
|  | **广州分公司** |
| 地址： | 广州市海珠区荔福路68号广州市微生物所六楼西 |
| 电话： | 020-61130189 |
| E-mail： | seqgz@majorbio.com |
|  | **北京子公司** |
| 地址： | 北京市海淀区安宁庄东路18号光华创新园新科研楼6楼 |
| 电话： | 010-51293026,010-51293126 |
| E-mail： | seqbj@majorbio.com |