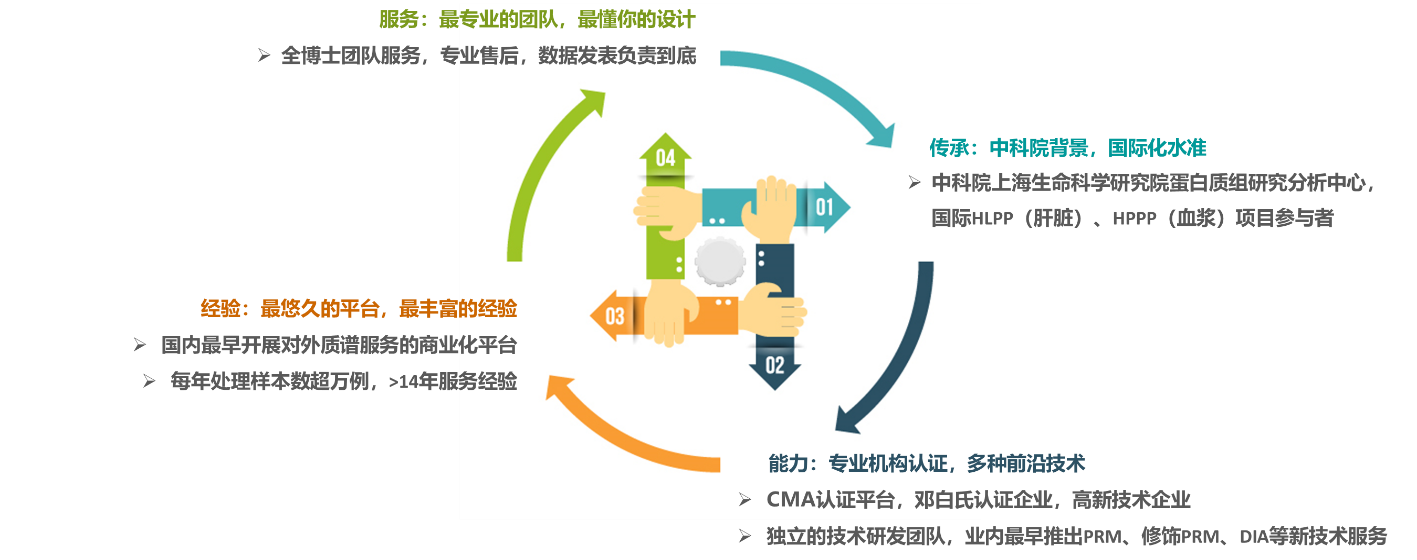
**蛋白质组项目报告**

****



**Label-free定量蛋白质组**

**生物信息学分析报告**

**项目名称：**

**委托单位：**

**项目编号：**

**报告时间：**

**目 录**

[1. 项目概述 Summary 4](#_Toc34828600)

[2. 分析流程 Flowchart 6](#_Toc34828601)

[3. 生物信息学分析 Bioinformatics analysis 8](#_Toc34828602)

[**3.1** **鉴定数量分析** 8](#_Toc34828603)

[**3.2** **表达差异分析** 10](#_Toc34828604)

[**3.3** **功能分析** 13](#_Toc34828605)

[4. 材料和方法 Materials and Methods 27](#_Toc34828606)

[**4.1** **质谱实验方法** 27](#_Toc34828607)

[**4.2** **生物信息学分析方法** 29](#_Toc34828608)

[5. 质量控制与说明文档 Quality control and introduction 31](#_Toc34828609)

[**5.1** **实验原理介绍** 31](#_Toc34828610)

[**5.2** **质量控制（QC）** 32](#_Toc34828611)

[**5.3** **鉴定蛋白质和肽段特性描述** 34](#_Toc34828612)

[**5.4** **质谱鉴定表格说明** 37](#_Toc34828613)

[**5.5** **数据库介绍** 41](#_Toc34828614)

[6. 参考文献References 43](#_Toc34828615)

[7. 生信数据分析云平台 Data analysis platform 45](#_Toc34828616)

[8. 拓展研究建议（供参考）Research suggestions 46](#_Toc34828617)

[9. 组学期刊投稿指南（供参考）Publication recommendation 47](#_Toc34828618)

[10. 部分合作发表文章Cooperative Projects Papers 48](#_Toc34828619)

# 项目概述 Summary

为了获得比较组中蛋白质表达差异变化，本项目采用Label-free定量蛋白质组学技术开展研究，具体项目结果概览如下表。

表 项目结果总览

|  |  |
| --- | --- |
| **样本类型** | |
| **Species** | **Database** |
|  |  |

|  |  |
| --- | --- |
| **样本信息** | |
|  |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **鉴定结果** | | | | | |
| **Spectrum** | | **Peptides** | | **Proteins** | |
| **Total spectrum** | **Matched spectrum** | **Peptides** | **Unique peptides** | **Identified** | **Quantified** |
|  |  |  |  |  |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **差异分析** | | | | | |
| **Filtering Criteria** | **Fold change>upRatio and P value<0.05** | | | **Proteins which identified in at least two of three replicates in one group while another group with all null value** | |
| **Comparisons** | **Significantly changing** | | | **Consistent presence/absence** | |
|
| **Downregulated** | **Upregulated** | **All** | **Downregulated** | **Upregulated** |
|  |  |  |  |  |  |

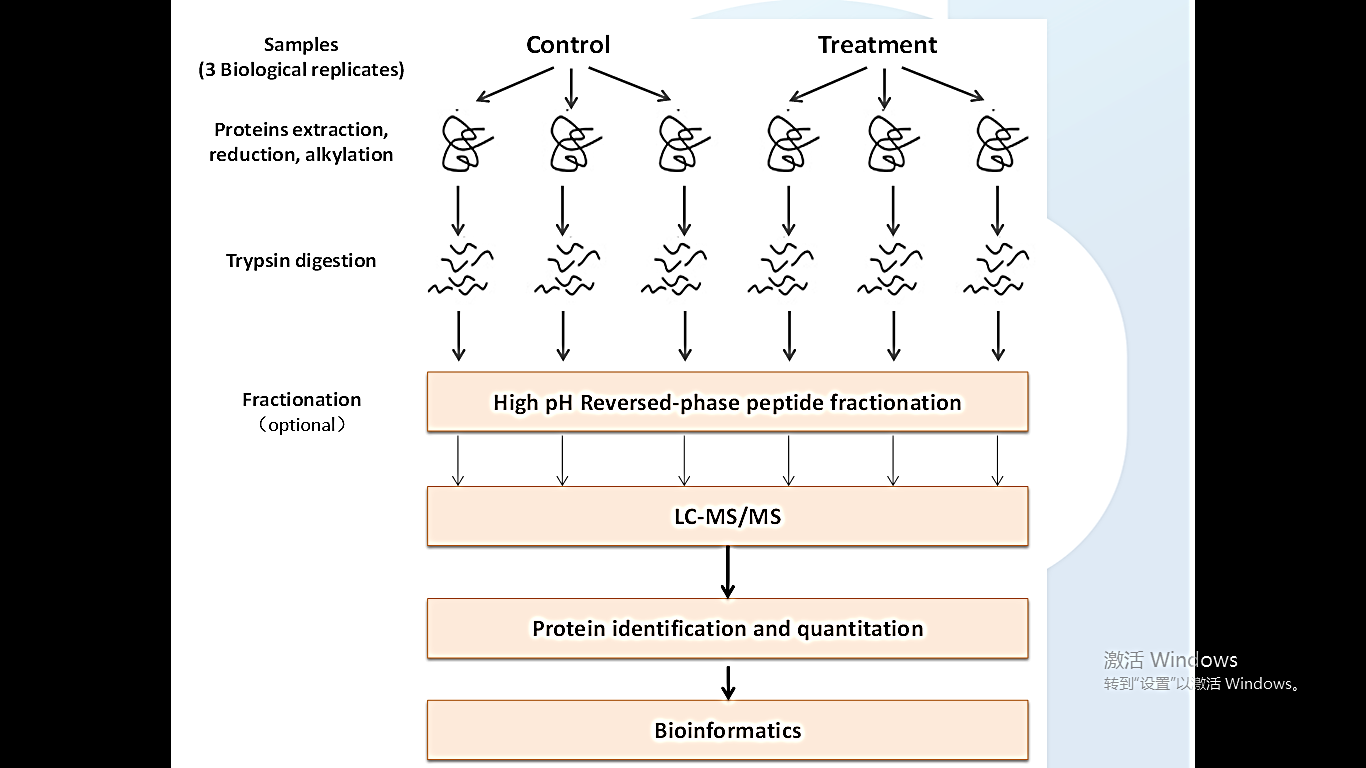
|  |  |
| --- | --- |
| **功能分析** | |
| **Comparisons** | **GO enrichment Top 5** |
|  |  |
|  |
|  |
|  |
|  |
| **Comparisons** | **KEGG pathway enrichment Top 5** |
|  |  |
|  |
|  |
|  |
|  |

备注：报告仅展示一组比较组功能分析信息。总项目鉴定与定量信息参见输出文件：[1项目鉴定总览](file:///D:\4%20产品相关资料\1%20产品-方案-报告模板\201910%20报告升级\2%20蛋白组label%20free\labef%20free蛋白组%20升级\1项目鉴定总览)。

# 分析流程 Flowchart

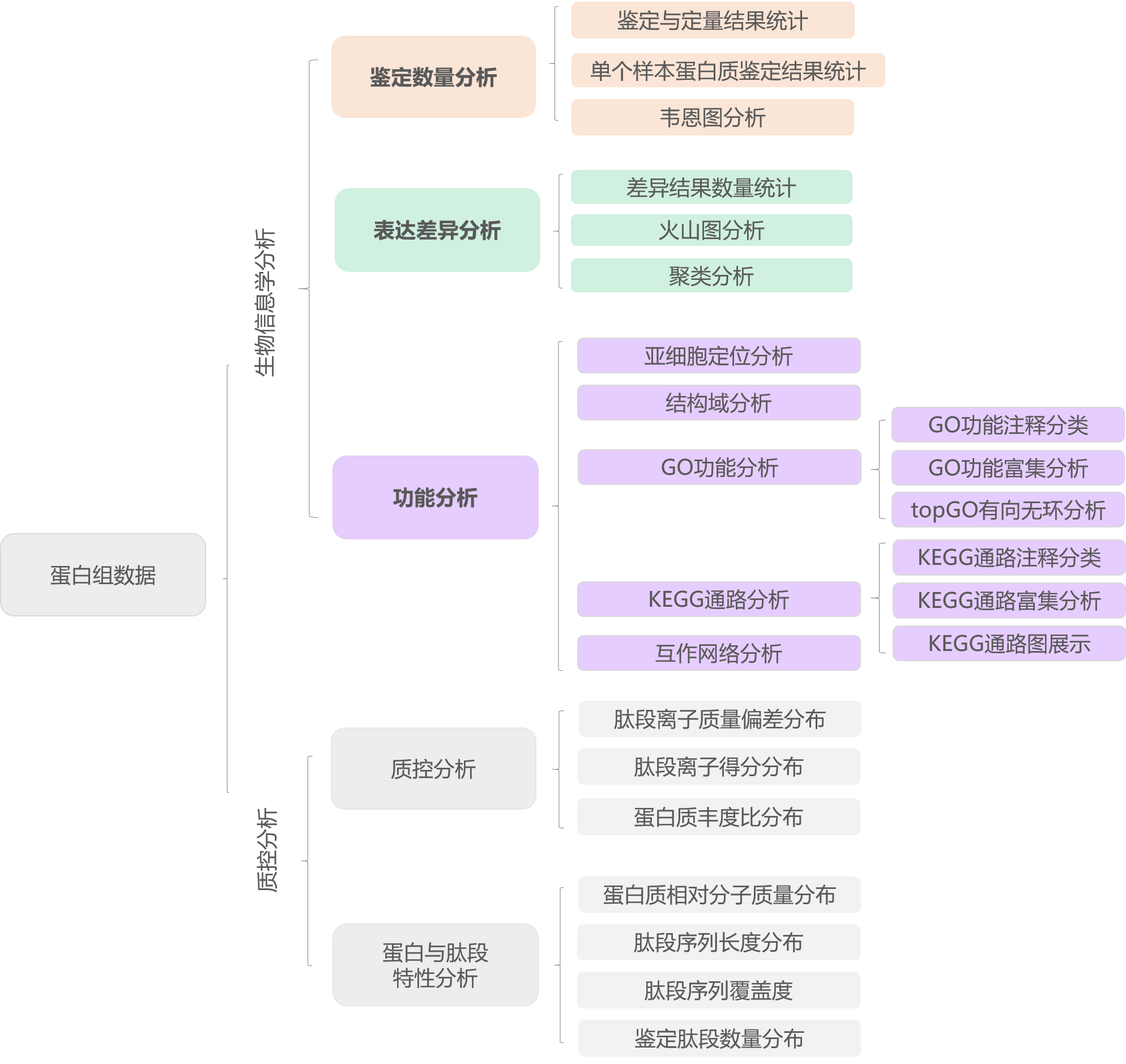
本项目分析流程主要包括质谱实验与数据分析两个阶段。

质谱实验分析流程主要包括蛋白质提取、肽段酶解、色谱分级（可选项）、液相色谱-串联质谱（LC-MS/MS）数据采集、数据库检索等步骤，见下图。



Label-free定量蛋白质组学实验流程图

随后对搜库后数据进行生物信息学分析，分析内容主要包括鉴定分析、表达差异分析、功能分析等，见下图。



生物信息学分析内容

# 生物信息学分析 Bioinformatics analysis

* + - 1. **鉴定数量分析**

### 3.1.1鉴定与定量结果统计

本项目共获得的谱图数、鉴定的肽段数、鉴定的蛋白数结果统计结果，如下表与下图。

鉴定与定量结果统计表

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Spectrum** | | **Peptides** | | **Proteins** | |
| **Total spectrum** | **Matched spectrum** | **Peptides** | **Unique peptides** | **Identified** | **Quantified** |
|  |  |  |  |  |  |

[Statistic]

鉴定与定量结果统计柱状图

说明： Total spectrum：二级谱图总数，Matched spectrum：数据库匹配谱图总数；Peptides：肽段总数目；Unique peptides：唯一肽段总数目； Identified Proteins：鉴定蛋白数。Quantified Proteins：可定量蛋白数，蛋白在至少一个组别中，有超过半数的生物学重复有定量信息。

输出文件：

[1) 3-1鉴定数量分析\3-1-1鉴定与定量结果统计](D:\\4 产品相关资料\\1 产品-方案-报告模板\\201910 报告升级\\2 蛋白组label free\\labef free蛋白组 升级\\3-1鉴定数量分析\\3-1-1 鉴定与定量结果统计)

**3.1.2单个样本蛋白质鉴定结果统计**

单个样本的蛋白质鉴定数量统计如下表。

单个样本蛋白质鉴定结果统计

|  |  |  |
| --- | --- | --- |
| **Database** | **Sample** | **Protein groups** |
|  |  |  |

说明： Database：选择使用的数据库物种名称；Sample：样本名称或者质谱实验名称； Protein groups：鉴定到的蛋白质总数。

输出文件：

1. [3-1鉴定数量分析/3-1-2单个样本蛋白质鉴定结果统计](D:\\4 产品相关资料\\1 产品-方案-报告模板\\201910 报告升级\\2 蛋白组label free\\labef free蛋白组 升级\\3-1鉴定数量分析\\3-1-2 修饰位点分布分析\\3-1-2-1磷酸化修饰位点的数量分布)

**3.1.3 韦恩图分析**

为了分析组内样本之间蛋白质鉴定的重复性，采用韦恩图分析组内样本的重叠情况。如下图，显示了组内所有生物学重复实验鉴定到的蛋白质集合的重叠情况。

[venn1]

该样本组内所有重复鉴定到的蛋白质Venn图

为了分析组间样本的蛋白质鉴定重叠情况，采用韦恩图分析不同组之间的鉴定重叠情况。如下图，分别显示了所有样本组间鉴定到的蛋白质和蛋白质集合的重叠情况。

[venn2]

所有样本组间鉴定到的蛋白质Venn图

输出文件：

1. [3-1鉴定数量分析/3-1-3 韦恩图分析](D:\\4 产品相关资料\\1 产品-方案-报告模板\\201910 报告升级\\2 蛋白组label free\\labef free蛋白组 升级\\3-1鉴定数量分析\\3-1-3 韦恩图分析)
   * + 1. **表达差异分析**

### 3.2.1 差异结果数量统计

为了分析不同组间具有表达差异的蛋白质，对实验数据进一步进行差异筛选。

在显著性差异蛋白质筛选中，以表达倍数(Fold Change, FC)> upRatio倍（上调大于upRatio倍或下调小于downRatio倍）且P value<0.05（T-test或其他）为标准，得到比较组间的上调、下调蛋白质数目，如下表中Significantly changing in abundance列。同时，将结果以柱状图形式呈现，其中上、下调> 10倍的蛋白数目以更深颜色标注，如下图。在“有无”差异比较中，以一组样品中两次及以上不为空值，另一组所有数据均为空值的筛选标准，得到蛋白质数目，如下表中Consistent presence/absence expression profile列，该部分蛋白除不参与火山图、聚类分析外，与上述显著性差异蛋白质一并进行其他生物信息学功能分析。

表 蛋白质定量差异结果统计

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Comparisons** | **Significantly changing**  **in abundance** | | | **Consistent presence/absence**  **expression profile** | |
| **Downregulated** | **Upregulated** | **All** | **Downregulated** | **Upregulated** |
|  |  |  |  |  |  |

[pro\_diff]

蛋白质定量差异结果柱状图

说明：Comparisons：差异比较组； Significantly changing in abundance: 符合筛选倍数和p value的差异表达蛋白质；Consistent presence/absence expression profile：一组样品中两次及以上不为空值，另一组所有数据均为空值的差异蛋白质。Upregulated：上调差异表达蛋白质；Downregulated：下调表达蛋白质；All：所有差异表达蛋白质；

输出文件：

1. [3-2表达差异分析/3-2-1差异结果数量统计](D:\\4 产品相关资料\\1 产品-方案-报告模板\\201910 报告升级\\2 蛋白组label free\\labef free蛋白组 升级\\3-2表达差异分析\\3-2-1差异结果数量统计)

### 3.2.2 火山图

为了展示比较组间蛋白质的显著性差异，将比较组中蛋白质以表达差异倍数（Fold change）和P value（T-test）两个因素为标准绘制火山图，其中显著下调的蛋白质以蓝色标注（FC< downRatio 且p<0.05），显著上调的蛋白质以红色标注（FC>upRatio 且p<0.05），无差异的蛋白质为灰色，如下图。

[volcano]

**groupvs**组的差异表达火山图

说明：横坐标为差异倍数（以2为底的对数变换），纵坐标为差异的显著性P-value（以10为底的对数变换），图中红色点为上调的显著性差异表达蛋白质，蓝色点为下调的著性差异表达蛋白质，灰点为无差异变化的蛋白质。

注：火山图形式展示适用于采用T检验算法获得的数据【无三次及以上重复实验设计的数据不适用火山图展示形式】。

输出文件：

1) [3-2表达差异分析/ 3-2-2火山图](file:///D:\4%20产品相关资料\1%20产品-方案-报告模板\201910%20报告升级\2%20蛋白组label%20free\labef%20free蛋白组%20升级\3-2表达差异分析\3-2-2火山图)

**3.2.3 聚类分析**

为了分析组间、组内样本的表达模式，检验本项目分组合理性，说明差异蛋白质表达量变化是否可代表生物学处理对样本造成的显著影响，采用层次聚类算法（Hierarchical Cluster）对比较组的差异表达蛋白质进行分组归类，并以热图（Heatmap）的形式展示。基于相似性基础，聚类分组结果中，一般组内的数据模式相似性较高，而组间的数据模式相似性较低，因此可以有效区分组别。

如下图，以倍数变化>upRatio倍且P value<0.05（T-test或其他）的筛选标准，得到的显著差异表达蛋白质可以有效的把比较组分开，说明差异表达蛋白质筛选能够代表生物学处理对样本影响。

[cluster]

**groupvs**组的差异表达蛋白质聚类分析

说明：层次聚类结果以树型热图表示，每列代表一组样品（横坐标为样品信息），每行代表一个蛋白质（即纵坐标为显著性差异表达的蛋白质），显著性差异蛋白质在不同样品中的表达量用Z-Score方法进行标准化后以不同颜色在热图中展现，其中红色代表显著性上调的蛋白质，蓝色代表显著性下调的蛋白质，灰色部分代表无蛋白质定量信息。

输出文件：

1. [3-2表达差异分析/ 3-2-3聚类分析](D:\\4 产品相关资料\\1 产品-方案-报告模板\\201910 报告升级\\2 蛋白组label free\\labef free蛋白组 升级\\3-2表达差异分析\\3-2-3聚类分析)

* + - 1. **功能分析**

### 3.3.1 亚细胞定位分析

亚细胞器（Organelle）是细胞质内具有一定形态和功能的微器官（如线粒体、内质网等），它是蛋白发挥不同功能的重要场所。不同亚细胞器往往行使不同细胞功能，故分析蛋白的亚细胞定位有助于我们进一步探究蛋白质在细胞中发挥的作用。

采用亚细胞结构预测软件CELLO[1]对所有差异表达的蛋白质进行亚细胞定位分析，分析结果以表格形式输出,参见输出文件。同时，以饼状图形式展示各亚细胞器中的蛋白质数目与分布比例，如下图。

[Subcellular\_Localization]

**groupvs**组的差异表达蛋白质亚细胞定位饼图

输出文件：

1) [3-3功能分析/3-3-1 亚细胞定位分析](file:///D:\4%20产品相关资料\1%20产品-方案-报告模板\201910%20报告升级\2%20蛋白组label%20free\labef%20free蛋白组%20升级\3-3功能分析\3-3-2亚细胞定位分析)

### 3.3.2 结构域分析

蛋白质结构域（domain)是在较大的蛋白质分子中，由于多肽链上相邻的超二级结构紧密联系，形成两个或多个在空间上可以明显区别的局部区域。一般每个结构域由几十至几百个氨基酸残基组成，各有独特的空间结构，并承担不同的生物学功能。一般来说，蛋白与蛋白（或其他小分子）的相互作用常以结构域为单位，结构域内氨基酸或修饰发生改变，可能引起蛋白关键功能的改变，故后续氨基酸突变功能实验可以以此为参考。因此，结构域预测对于研究蛋白关键功能区域及其发挥的潜在生物学作用具有重要意义。

采用结构域预测软件interproscan[2]对差异表达蛋白质进行结构域预测，分析结果以表格形式输出,参见输出文件。同时，以柱状图形式展示Domain中的蛋白质数目(前20 )，如下图。

[TopDomainStat]

**groupvs**组差异表达蛋白质结构域分析

为了揭示差异表达蛋白质的结构域富集特征，并通过评价某个结构域条目下的蛋白质富集度的显著性水平，找到研究者最关心的显著富集结构域及其对应差异蛋白，采用Fisher精确检验（Fisher’s Exact Test）对差异表达蛋白质进行结构域富集分析，如下图。

[Domain\_Enrichment]

**groupvs**组组结构域富集分析图

说明：图中横坐标为富集因子（Rich Fator≤1）。纵坐标表示每个结构域分类下的差异蛋白质统计结果；其中气泡颜色表示富集的结构域分类的显著性，即基于Fisher精确检验（Fisher’s Exact Test）计算P值，颜色梯度代表P值的大小（取-log10），颜色越接近红色代表P值越小，对应的结构域分类下富集度的显著性水平越高。

输出文件：

1) [3-3功能分析/3-3-2结构域分析](file:///D:\4%20产品相关资料\1%20产品-方案-报告模板\201910%20报告升级\2%20蛋白组label%20free\labef%20free蛋白组%20升级\3-3功能分析\3-3-3结构域分析)

### 3.3.3 GO功能分析

为了全面了解蛋白在生物体中的功能、定位及参与的生物学途径，通过基因本体（Gene Ontology, GO）对蛋白质进行注释。GO是一个标准化的功能分类体系，提供了一套动态更新的标准化词汇表用以描述生物体中基因和基因产物的属性。GO功能注释主要分为3类：生物过程（Biological Process, BP），分子功能（Molecular Function, MF）和细胞组分（Cellular Component, CC）[3]。

本项目采用Blast2Go（https://www.blast2go.com/）[4]软件对所有差异表达蛋白质进行GO功能注释，注释结果表格参见输出文件。同时，在GO二级功能注释层级上对差异蛋白数目进行统计，如下图。

更多信息请参考: <http://www.geneontology.org/><http://en.wikipedia.org/wiki/Gene_Ontology>

[GOLevel2]

**groupvs**组差异表达蛋白质的GO注释统计图（level 2）

说明：图中纵坐标表示GO 二级功能注释信息（GO Level2），包含生物过程（Biological Process），分子功能（Molecular Function）和细胞组分（Cellular Component），依次以蓝色，红色，橙色予以区分；横坐标（下）表示每个功能分类下的差异表达蛋白质数目，横坐标（上）表示每个功能分类下的差异表达蛋白质数目占总的差异表达蛋白质数目的百分比。一般情况下，某一功能类别对应的差异表达蛋白质数目越多，说明该功能越重要，需要重点关注或者进行后续深入机制的探讨。

为了揭示所有差异表达蛋白质的整体功能富集特征，并通过评价某个GO功能条目的蛋白质富集度的显著性水平，找到研究者最关心的显著富集GO条目，采用Fisher精确检验（Fisher’s Exact Test）对差异表达蛋白质进行GO功能富集分析。

将所有差异表达蛋白质与参考物种的全部蛋白质（或实验鉴定到的所有蛋白质）以GO功能的注释结果进行对照比较，通过Fisher精确检验 (Fisher’s Exact Test)得出两者差异的显著性，从而找到所有差异表达蛋白质富集的功能类别（P value <0.05）。用气泡图分别显示GO三大分类下的GO条目富集情况（柱状图参见输出文件），如下面三图所示，在该比较组中**BP-TOP5**等重要生物学过程；**MF-TOP5**等分子功能；**CC-TOP5**等定位蛋白质发生了显著性变化。

[BP\_Enrichment]

**groupvs**组生物过程分类下GO功能富集气泡图（Biological Process，BP）

[CC\_Enrichment]

**groupvs**组细胞组分分类下GO功能富集气泡图（Cellular Component，CC）

[MF\_Enrichment]

**groupvs**组分子功能分类下GO功能富集气泡图（Molecular Function，MF）

说明：图中横坐标为富集因子（Rich Fator≤1）,富集因子表示注释到某GO 功能类别的差异表达蛋白质数目占注释到该GO功能类别的所有鉴定到的蛋白质数目的比例。纵坐标表示每个GO功能分类下的差异蛋白质统计结果；其中气泡颜色表示富集的GO功能分类的显著性，即基于Fisher精确检验（Fisher’s Exact Test）计算P值，颜色梯度代表P值的大小（取-log10），颜色越接近红色代表P值越小，对应的GO功能类别富集度的显著性水平越高。**一般情况下，GO富集结果中P值越小（P<0.05），对应GO功能分类从统计学上讲富集越显著，而与GO功能分类相关的差异表达蛋白质数目在某种程度上反映实验设计中生物学处理对各个分类的影响程度大小，因此可以结合两方面因素，选择较为感兴趣的生物学功能以及显著性影响这些功能的差异表达蛋白质进行后续生物学实验验证或机制研究。**

为了展示差异蛋白质的富集GO 条目的层级关系，采用topGO有向无环图(DAG)展示自上而下的定义功能范围，其中分支代表包含关系，分支越往下代表定义功能范围越具体，如下图。

[BP\_DAG]

**groupvs**组GO富集有向无环图（Biological Process，BP）

[CC\_DAG]

**groupvs**组GO富集有向无环图（Cellular Component，CC）

[MF\_DAG]

**groupvs**组GO富集有向无环图(Molecular Function，MF)

说明：对GO三大分类的每一类均取富集程度最高的前10位作为有向无环图的主节点，以方框表示。并通过包含关系将相关联的GO条目一起展示，用圆框表示。颜色表明富集程度，越接近红色富集程度越高。

输出文件：

1) [3-3功能分析/3-3-3 GO功能分析](file:///D:\4%20产品相关资料\1%20产品-方案-报告模板\201910%20报告升级\2%20蛋白组label%20free\labef%20free蛋白组%20升级\3-3功能分析\3-3-4GO功能分析)

### 3.3.4 KEGG通路分析

为了更系统全面地解析生物学过程、疾病发生机理、药物作用机制等，往往需要从一系列蛋白质协调作用的角度阐述变化规律，如代谢通路变化。因此通过KEGG通路数据库（Kyoto Encyclopedia of Genes and Genomes）对蛋白质解析注释[5]。KEGG是由研究人员阅读海量文献后，将众多的代谢途径以特定的图形语言整理而成的数据库，其收录了新陈代谢，遗传信息加工，环境信息加工，细胞过程，生物体系统，人类疾病以及药物开发等多个方面的通路信息，常用于通路研究。

本项目将所有差异表达蛋白质进行KEGG通路注释，注释表格参见输出文件。同时，对差异表达蛋白质对应的蛋白质数目进行统计，其中差异表达蛋白数目最多的前几条通路为**kegg-map2query-top5** ，如下图。

更多信息请参考: <http://www.genome.jp/kegg/pathway.html>。

[TopMapStat]

**groupvs**组差异表达蛋白质的KEGG通路注释统计图（Top20）

说明：图中纵坐标是含差异表达蛋白质参与的通路名称，横坐标表示参与该通路的含差异表达蛋白质的数目。一般情况下，参与某一通路的差异表达蛋白质数目越多，说明该通路越重要，需要重点关注或者进行后续深入机制的探讨。

为了揭示所有差异表达蛋白质的整体代谢通路富集特征，并通过评价某个KEGG代谢通路的蛋白质富集度的显著性水平，找到研究者最关心的显著富集KEGG代谢通路，采用Fisher精确检验（Fisher’s Exact Test）对差异表达蛋白质进行KEGG通路富集分析。

将所有差异表达蛋白质与参考物种的全部蛋白质（或实验鉴定到的所有蛋白质）以KEGG的注释结果进行对照比较，通过Fisher精确检验 (Fisher’s Exact Test)得出两者差异的显著性，从而找到所有差异表达蛋白质富集的通路类别（P value <0.05）。如下图所示，通过Fisher精确检验方法对比较组的差异表达蛋白质进行KEGG通路富集分析，结果表明**KeggEnrich-top5** 等重要通路发生了显著变化。以**KeggEnrich-top1**通路进行可视化展示，如下图。

[KEGG\_Enrichment]

**groupvs**组KEGG通路富集气泡图（Molecular Function，MF）

说明：图中横坐标为富集因子（Rich Fator≤1）,富集因子表示注释到KEGG通路类别的差异表达蛋白质数目占注释到该类别的所有鉴定到的蛋白质数目的比例。纵坐标表示每个KEGG通路下的差异蛋白质统计结果；其中气泡颜色表示富集的KEGG通路的显著性，即基于Fisher精确检验（Fisher’s Exact Test）计算P值，颜色梯度代表P值的大小（-log10），颜色越接近红色代表P值越小，对应代谢通路富集度的显著性水平越高。**因此可以选择较为感兴趣的生物学功能以及显著性影响这些功能的差异表达蛋白质进行后续生物学实验验证或机制研究。**

[kegg\_pathway]

**groupvs**组KEGG通路图示例

说明：图中红底方框表示发生差异的蛋白质为上调, 绿底方框表示发生差异的蛋白质为下调，黄底方框表示有多个蛋白既有上调又有下调。小圆圈表示小分子代谢物，大圆框代表其他通路。备注：浅绿色底方框为物种专属蛋白，浅紫色底方框为不区分物种特异性蛋白，红色字体为KEGG默认与疾病相关蛋白。

输出文件：

1) [3-3功能分析/3-3-4 KEGG通路分析](file:///D:\4%20产品相关资料\1%20产品-方案-报告模板\201910%20报告升级\2%20蛋白组label%20free\labef%20free蛋白组%20升级\3-3功能分析\3-3-5KEGG通路分析)

### 3.3.5蛋白互作网络分析

蛋白质发挥功能的重要方式之一就是与其他蛋白发生相互作用，通过蛋白间介导的途径、或形成复合物进而发挥生物学调控作用。例如，高度聚集的蛋白质可能具有相同或相似的功能；连接度高的蛋白质可能是影响整个系统代谢或信号转导途径的关键点。因此研究蛋白-蛋白相互作用（Protein Protein Interaction,PPI）具有重要意义。此外，将蛋白质相互作用网络分析和通路注释的结果相结合，还可以获得更全面系统的分子层面的细胞活动模型，便于分子机制的深入研究和挖掘。

本项目基于STRING或IntAct数据库中的蛋白质相互作用关系，使用CytoScape软件，对比较组的差异表达蛋白质构建蛋白质互作网络图，如下图。

[ppi]

**groupvs**组差异表达蛋白质相互作用网络图

说明：图中圆圈结点表示差异表达蛋白质，线表示蛋白质与蛋白质之间的相互作用。其中圆圈颜色表示蛋白质表达差异（下调标注为蓝色、上调标注为红色），圆圈大小表明该蛋白质连接度（即与某蛋白直接相互作用的蛋白质数目）。**通常来讲，连接度越大，该蛋白质发生变化时整个系统受到的扰动就越大，更可能是维持系统平衡和稳定的关键，为后续重点研究的候选蛋白质。**

在PPI互作网络中，高度聚集的蛋白质往往可能具有相同或相似的功能，并通过协同作用发挥生物学功能。因此，基于拓扑结构识别原理，将互作网络图中聚集程度高的蛋白划分为不同簇（Cluster），具体划分簇展示如下图（每一类簇的展示图详见输出文件）。进一步，对每一类簇进行功能方向归类，功能归类表参见输出文件。

[Module\_ppi]

**groupvs**组相互作用网络图进行功能归类图

说明：通常来讲，同一网络模块内蛋白往往具有相似的生物学功能，可选区感兴趣功能模块内的蛋白作为后续研究重点。

输出文件：

1. [3-3功能分析/3-3-5蛋白互作网络分析](D:\\4 产品相关资料\\1 产品-方案-报告模板\\201910 报告升级\\2 蛋白组label free\\labef free蛋白组 升级\\3-3功能分析\\3-3-6蛋白互作网络分析)

# 材料和方法 Materials and Methods

1. **质谱实验方法**

**4.1.1蛋白质提取和肽段酶解**

样品（人、小鼠和大鼠血清样品需要先去除血清高丰度蛋白质）采用SDT（4%(w/v) SDS, 100mM Tris/HCl pH7.6, 0.1M DTT）裂解法提取蛋白质，然后采用BCA法进行蛋白质定量。每个样品取适量蛋白质采用Filter aided proteome preparation (FASP)方法进行胰蛋白酶酶解[6]，采用C18 Cartridge对肽段进行脱盐，肽段冻干后加入40μL 0.1%甲酸溶液复溶，肽段定量（OD280）。

**4.1.2 LC-MS/MS数据采集**

每份样品采用纳升流速的HPLC液相系统Easy nLC进行分离。缓冲液A液为0.1%甲酸水溶液，B液为0.1%甲酸乙腈水溶液（乙腈为84%）。色谱柱以95%的A液平衡，样品由自动进样器上样到上样柱（Thermo Scientific Acclaim PepMap100, 100μm\*2cm, nanoViper C18），经过分析柱（Thermo scientific EASY column, 10cm, ID75μm, 3μm, C18-A2）分离，流速为300 nL/min。

样品经色谱分离后用Q-Exactive质谱仪进行质谱分析。检测方式为正离子，母离子扫描范围300 – 1800 m/z，一级质谱分辨率为70,000 at 200 m/z，AGC(Automatic gain control) target为1e6，Maximum IT为50ms，动态排除时间（Dynamic exclusion）为60.0s。多肽和多肽碎片的质量电荷比按照下列方法采集：每次全扫描（full scan）后采集20个碎片图谱（MS2 scan），MS2 Activation Type为HCD，Isolation window为2 m/z，二级质谱分辨率17,500 at 200 m/z， Normalized Collision Energy为30eV，Underfill为0.1%。

**4.1.3蛋白质鉴定和定量分析**

质谱分析原始数据为RAW文件，采用MaxQuant软件（版本号1.5.3.17） [7]进行查库鉴定及定量分析。相关参数和说明如下：

Maxquant鉴定和定量参数表

|  |  |
| --- | --- |
| Item | Value |
| Enzyme | Trypsin |
| Max Missed Cleavages  （允许的最大漏切位点数目） | 2 |
| Main search  （一级离子质量容差) | 6 ppm |
| First search  （预检索一级离子质量容差) | 20 ppm |
| MS/MS Tolerance  （二级级离子质量容差) | 20 ppm |
| Fixed modifications(固定修饰) | Carbamidomethyl (C) |
| Variable modifications(可变修饰) | Oxidation (M) |
| Database  （查库所使用的蛋白质序列数据库） | 公共库命名如 uniprot\_mouse（物种）\_76417（蛋白质序列条数）\_20141212（下载日期）.fasta；自建库一般命名为项目号。具体数据库信息见项目报告。 |
| Database pattern  （用于计算FDR的数据库模式） | Reverse |
| Include contaminants | True  （包含常见污染蛋白质序列的数据库） |
| Protein FDR  （可信蛋白质的筛选标准） | ≤0.01 |
| Peptide FDR  （可信肽段的筛选标准） | ≤0.01 |
| Peptides used for protein quantification  （参与定量肽段） | Use razor and unique peptides(采用唯一肽段和razor肽段的定量强度值进行蛋白质定量) |
| Time window (match between runs)  （色谱对齐的时间窗口） | 2min |
| protein quantification  （蛋白质定量算法） | LFQ |
| min. ratio count  （用于蛋白质定量分析的最少比值数） | 1 |

说明： iBAQ（Intensity-based absolute quantification）和LFQ 属于Maxquant软件提供的两种不同的蛋白质定量算法。iBAQ一般用于样本的蛋白质绝对定量，主要算法是基于该蛋白质鉴定到的肽段的强度之和与理论肽段个数的比值[8]。LFQ一般用于多组间的两两定量比较，主要算法是经过肽段和蛋白质多层的pair-wise矫正[9]。

1. **生物信息学分析方法**

### 4.2.1蛋白质聚类分析

首先对目标蛋白质集合的定量信息进行归一化处理（归一化到（-1,1）区间）。然后，使用Complexheatmap R包（R Version 3.4）同时对样品和蛋白质的表达量两个维度进行分类（距离算法：欧几里得，连接方式：Average linkage），并生成层次聚类热图。

### 4.2.2亚细胞定位分析

采用CELLO（http://cello.life.nctu.edu.tw/）的方法进行亚细胞定位预测，该方法采用多重支持向量机（multi-class SVM）的机器学习的方法对公共数据库亚细胞定位信息已知的蛋白质序列数据建模，用于预测待检索蛋白亚细胞定位信息。

### 4.2.3蛋白结构域分析

蛋白结构域分析使用Pfam数据库，该数据库是一系列蛋白质家族的集合，其中每一个蛋白家族都以多序列比对和隐马尔科夫模型的形式来表示，该数据库包含有结构域信息。具体分析时，使用InterProScan软件包，以集成的方式从InterPro数据库运行扫描算法对序列进行功能表征，从而获得目标蛋白序列在Pfam数据库中的结构域注释信息。

### 4.2.4 GO功能注释

利用Blast2GO对目标蛋白质集合进行GO注释，过程大致可以归纳为序列比对（Blast）、GO条目提取（Mapping）、GO注释（Annotation）和InterProScan补充注释（Annotation Augmentation）等四个步骤。

### 4.2.5 KEGG通路注释

利用KAAS (KEGG Automatic Annotation Server) 软件，对目标蛋白质集合进行KEGG通路注释。

### 4.2.6富集分析

采用Fisher精确检验（Fisher’s Exact Test），比较各个GO分类（或KEGG通路、或Domain）在目标蛋白质集合和总体蛋白质集合中的分布情况，对目标蛋白质集合进行GO注释或（或KEGG通路、或Domain）注释的富集分析。

### 4.2.7 蛋白质相互作用网络分析

基于IntAct (<http://www.ebi.ac.uk/intact/main.xhtml>)或者STRING（http://string-db.org/）数据库中的信息查找目标蛋白质之间的直接和间接相互作用关系，并使用CytoScape软件（版本号：3.2.1）生成相互作用网络并对网络进行分析。

# 质量控制与说明文档 Quality control and introduction

1. **实验原理介绍**

非标记定量蛋白质组学（Label-free）技术近年来已成为重要的质谱定量方法[9-10]。Label-free技术的定量原理主要有两种：首先，spectrum counts类的非标记定量方法发展较早，也形成了多种定量算法，但核心原理都是根据MS2的鉴定结果作为定量的基础，各种方法的差别在于后期算法对高通量数据的修正；第二种非标记定量方法的原理是以MS1为基础，计算每个肽段信号在LCMS色谱上的积分。本项目采用的Maxquant算法即基于第二种原理，如下图。



图MaxQuant软件进行非标记定量的原理

1. **质量控制（QC）**

本实验采用具有高质量精度、高分辨率的Q Exactive质谱仪，在数据采集过程中可以保持良好的质量偏差，并最终获得高质量的MS1和MS2图谱。**如下图1所示，所有鉴定肽段的质量偏差主要分布在10ppm以内，说明鉴定结果准确可靠**。然后结合Andromeda这种严格的分析工具对MS图谱数据进行分析，获得每张MS2图谱的得分。**下图2表明，MS2的Andromeda得分较为理想，约Percentage以上的肽段得分在 60分以上，肽段得分中位数为： Median Score分。**每一套Label free数据在定性分析工作中均使用Peptide FDR≤0.01和Protein FDR≤0.01作为筛选标准，结合所得的优秀肽段得分分布情况，进一步说明MS实验数据质量较高。

1. **肽段离子质量偏差分布**

[mass\_error]

图 肽段离子质量偏差分布

说明：横坐标为肽段离子理论质荷比和质谱实验测得质荷比的质量偏差，单位ppm即百万分之一，是一个相对单位。纵坐标为Andromeda肽段得分。

输出文件：

1. [5附件与说明文档/5-2质量控制（QC）/MassError\_Distribution](D:\\4 产品相关资料\\1 产品-方案-报告模板\\201910 报告升级\\2 蛋白组label free\\labef free蛋白组 升级\\5附件\\5-2质量控制（QC）\\MassError_Distribution.png)
   * 1. **肽段离子得分分布**

[Andromeda\_Score\_Distribution]

图 肽段离子得分分布图

说明：横坐标为肽段得分；主纵坐标The Number of Peptides对应图中的柱状图，表示鉴定到的具有对应离子得分的肽段数量；次纵坐标对应图中的累积曲线，表示不高于对应离子得分的肽段累积百分比。

输出文件：

1. [5附件与说明文档/5-2质量控制（QC）/Score\_Distribution](D:\\4 产品相关资料\\1 产品-方案-报告模板\\201910 报告升级\\2 蛋白组label free\\labef free蛋白组 升级\\5附件\\5-2质量控制（QC）\\Score_Distribution.png)

* + 1. **蛋白质丰度比分布**

[Ratio\_Distribution]

**groupvs**组蛋白质丰度比分布图

说明：横坐标为差异倍数（以2为底的对数变换），纵坐标为鉴定到的蛋白质数量。

输出文件：

1. [5附件与说明文档/5-2质量控制（QC）/Ratio\_Distribution](D:\\4 产品相关资料\\1 产品-方案-报告模板\\201910 报告升级\\2 蛋白组label free\\labef free蛋白组 升级\\5附件\\5-2质量控制（QC）)
2. **鉴定蛋白质和肽段特性描述**
3. **蛋白质相对分子质量分布**

[MW\_Distribution]

图 鉴定蛋白质相对分子质量分布图

说明：横坐标为鉴定到的蛋白质的相对分子量；主纵坐标Number of proteins对应图中的柱状图，表示鉴定到的具有对应相对分子质量的蛋白质数量；次纵坐标对应图中的累积曲线，表示具有不高于对应相对分子质量的蛋白质的累积百分比。

输出文件：

1. [5附件与说明文档/5-3鉴定肽段和蛋白质特性描述/Molecular\_Weight\_Distribution](D:\\4 产品相关资料\\1 产品-方案-报告模板\\201910 报告升级\\2 蛋白组label free\\labef free蛋白组 升级\\5附件\\5-3鉴定磷酸化蛋白和磷酸化肽段特性描述\\Molecular_Weight_Distribution.png)
2. **肽段序列长度分布**

[PepCount\_Distribution]

图 肽段序列长度分布图

说明：横坐标为鉴定到的肽段序列的氨基酸个数；纵坐标为鉴定到的肽段质数量（百分比）。

输出文件：

1. [5附件与说明文档/5-3鉴定蛋白质化蛋白和蛋白质特性描述/Phosphorylated\_Peptide\_Length\_Distribution](D:\\4 产品相关资料\\1 产品-方案-报告模板\\201910 报告升级\\2 蛋白组label free\\labef free蛋白组 升级\\5附件\\5-3鉴定磷酸化蛋白和磷酸化肽段特性描述\\Phosphorylated_Peptide_Length_Distribution.png)
2. **肽段序列覆盖度**

[Protein\_Sequence\_Coverage\_Distribution]

图 蛋白质序列覆盖度分布图

说明：该图显示鉴定到的不同覆盖度的蛋白质比例分布情况。横坐标为鉴定到的蛋白质序列覆盖百分比；纵坐标为鉴定到的蛋白质数量。

输出文件：

1. [5附件与说明文档/5-3鉴定蛋白质化蛋白和蛋白质特性描述/ Protein\_Sequence\_Coverage\_Distribution](D:\\4 产品相关资料\\1 产品-方案-报告模板\\201910 报告升级\\2 蛋白组label free\\labef free蛋白组 升级\\5附件\\5-3鉴定磷酸化蛋白和磷酸化肽段特性描述\\Phosphorylated_Peptide_Length_Distribution.png)
2. **鉴定肽段数量分布**

**[Peptide\_Count\_Distribution]**

图 鉴定肽段数量分布图

说明：该图显示鉴定蛋白质所对应的鉴定肽段数量分布情况。横坐标为鉴定蛋白质的肽段数量；主纵坐标Number of proteins对应图中的柱状图，表示对应鉴定肽段数量的蛋白质数目；次纵坐标对应图中的累积曲线，表示鉴定蛋白质的肽段数量不高于对应数值的蛋白质累积百分比。

输出文件：

1. [5附件与说明文档/5-3鉴定蛋白质化蛋白和蛋白质特性描述/ Peptide\_Count\_Distribution](D:\\4 产品相关资料\\1 产品-方案-报告模板\\201910 报告升级\\2 蛋白组label free\\labef free蛋白组 升级\\5附件\\5-3鉴定磷酸化蛋白和磷酸化肽段特性描述\\Phosphorylated_Peptide_Length_Distribution.png)
2. **质谱鉴定表格说明**

**附件1\_蛋白质鉴定列表**

|  |  |  |
| --- | --- | --- |
| **表头** | **定义** | **描述** |
| Protein IDs | 蛋白质登录号 | 与数据库匹配到的该蛋白质组中的所有蛋白质登录号，按照各自匹配到的肽段数量降序排列。Identifier(s) of protein(s) contained in the protein group. They are sorted by number of identified peptides in descending order. |
| Protein names | 蛋白质名称 | 显示Fasta header列中注释的蛋白质名称，如果数据库中的注释信息不完整或没有蛋白质名称，则不显示该内容。 |
| Gene names | 基因名称 | 显示Fasta header列中注释的基因名称，如果数据库中的注释信息不完整或没有基因名称，则不显示该内容。 |
| Fasta headers | 蛋白质在数据库中命名及描述 | 蛋白质序列数据库中的详细信息描述（与Protein IDs对应）。如果一个蛋白质组包含多个蛋白质，则按照每个蛋白质鉴定到的肽段数量降序排列。搜库经常使用两类数据库：公共库（用的比较多的是UniProtKB，NCBI）与自建库（由转录组数据库翻译而来的蛋白质数据库）。UniProtKB数据库（www.uniprot.org）的命名格式为：>db|UniqueIdentifier|EntryName ProteinName OS=OrganismName [GN=GeneName] PE=Protein Existence SV=SequenceVersion。 |
| Proteins | 蛋白质组中的蛋白质数量 | Number of proteins contained within the group. This corresponds to the number of entries in the column 'Protein IDs' |
| Peptides | 肽段数 | 鉴定某蛋白质组的全部肽段数，相当于用于定性的肽段数目。The total number of peptide sequences associated with the protein group (i.e. for all the proteins in the group). |
| Razor + unique peptides | 用于定量的肽段数量 | 用于定量的肽段数量。The total number of razor + unique peptides associated with the protein group |
| Unique Peptides | 唯一肽段数 | 专属于某一蛋白质组的肽段的数量。The total number of unique peptides associated with the protein group (i.e. these peptides are not shared with another protein group). |
| Sequence coverage [%] | 肽段覆盖率 | 鉴定到的氨基酸数目占蛋白质组中得分最高的蛋白质(Protein IDs中排名第一位的蛋白质)总氨基酸数目的比例。Percentage of the sequence that is covered by the identified peptides of the best protein sequence contained in the group. |
| Mol. weight [kDa] | 蛋白质的理论分子量 | 蛋白质组中得分最高的蛋白质(Protein IDs中排名第一位的蛋白质)的理论分子量。Molecular weight of the leading protein sequence contained in the protein group. |
| PEP | 后验错误概率 | Posterior Error Probability of the identification. This value essentially operates as a p-value, where smaller is more significant. |
| iBAQ/LFQ  intensity X | 蛋白质表达量强度值 | iBAQ Intensity是基于iBAQ算法得到的样品X中的蛋白质表达量，近似等于该样品中的蛋白质绝对浓度。LFQ Intensity是基于LFQ算法得到的样品X的蛋白相对表达量，常应用于组间比较。一般Labelfree选择其中之一作为定量结果。 |

**附件2\_肽段鉴定列表**

|  |  |  |
| --- | --- | --- |
| **表头** | **定义** | **描述** |
| Sequence | 肽段序列 | 描述肽段氨基酸的组成。The amino acid sequence of the identified peptide. |
| Length | 肽段的氨基酸个数 | The length of the sequence stored in the column "Sequence". |
| Missed cleavages | 漏切位点数量 | 鉴定肽段序列中含有胰蛋白酶漏切位点（即K和R）的数量。Number of missed enzymatic cleavages. |
| Mass | 肽段分子量 | 肽段的单同位素分子量。Monoisotopic mass of the peptide. |
| Proteins | 蛋白登录号 | 肽段所属蛋白质组的所有蛋白质登录号。Identifiers of proteins this peptide is associated with. |
| Leading razor protein | 得分最高的蛋白质 | 蛋白质组中得分最高的蛋白质，一般为蛋白质组中排名第一位的蛋白质。Identifiers of the best scoring protein this peptide is associated with. |
| Gene names | 基因名称 | 显示Fasta header列中注释的基因名称，如果数据库中的注释信息不完整或没有基因名称，则不显示该内容。 |
| Protein names | 蛋白质名称 | 显示Fasta header列中注释的蛋白质名称，如果数据库中的注释信息不完整或没有蛋白质名称，则不显示该内容。 |
| Unique (Groups) | 是否是唯一肽段 | When marked with 'yes', this particular peptide is unique to a single protein group in the proteinGroups file |
| Charges | 电荷数 | 检测到的肽段电荷数。All charge states that have been observed. |
| PEP | 后验错误概率 | Posterior Error Probability of the identification. This value essentially operates as a p-value, where smaller is better. |
| Score | 肽段得分 | Andromeda肽段得分。Andromeda score for the best associated MS/MS spectrum. |
| Intensity | 肽段总强度值 | 所有样品中的肽段强度值总和。Summed up eXtracted Ion Current (XIC) of all isotopic clusters associated with the identified AA sequence. |
| Intensity X | 肽段强度值 | 样品X中的肽段强度值。 |
| Contaminant | 是否为污染肽段 | MaxQuant查库除分析目标数据库（正库）和反库外，还包含一个标准的污染库，该数据库中包含如keratins, BSA, and trypsin等常见污染蛋白质序列。如果检测到的肽段为常见的污染蛋白质序列，则会被标注“+”，不参与后续数据分析。When marked with ‘+’, this particular peptide was found to be part of a commonly occurring contaminant. These should be removed for further data analysis. |

**附件3\_蛋白质定量和差异分析列表**

|  |  |  |
| --- | --- | --- |
| **包含表格** | **表格说明** | |
| “定量和差异分析”sheet（如果有多个比较组则有多个表格） | 对符合同组三次重复数据中至少有两个非空值的数据进行比值计算和统计学分析，包含各比较组的LFQ或者iBAQ强度值比值和P-value。差异表达蛋白质采用红色标识上调差异表达蛋白质，绿色标识下调差异表达蛋白质。 | |
| “空值数据”sheet（如果有多个比较组则有多个表格） | 该表中的数据保留原则：在进行两两组间的数据比较分析时，其中一组样本的三次重复实验数据中出现两次及以上不为空值的数据，同时另一样本组中所有数据均为空值。 | |
| **表头** | **定义** | **描述** |
| Protein IDs | 蛋白质登录号 | 与数据库匹配到的该蛋白质组中的所有蛋白质登录号，按照各自匹配到的肽段数量降序排列。Identifier(s) of protein(s) contained in the protein group. They are sorted by number of identified peptides in descending order. |
| Protein names | 蛋白质名称 | 显示Fasta header列中注释的蛋白质名称，如果数据库中的注释信息不完整或没有蛋白质名称，则不显示该内容。 |
| Gene names | 基因名称 | 显示Fasta header列中注释的基因名称，如果数据库中的注释信息不完整或没有基因名称，则不显示该内容。 |
| Fasta headers | 蛋白质在数据库中命名及描述 | 蛋白质序列数据库中的详细信息描述（与Protein IDs对应）。如果一个蛋白质组包含多个蛋白质，则按照每个蛋白质鉴定到的肽段数量降序排列。搜库经常使用两类数据库：公共库（用的比较多的是UniProtKB，NCBI）与自建库（由转录组数据库翻译而来的蛋白质数据库）。UniProtKB数据库（www.uniprot.org）的命名格式为：>db|UniqueIdentifier|EntryName ProteinName OS=OrganismName [GN=GeneName] PE=Protein Existence SV=SequenceVersion。 |
| Peptides | 肽段数 | 与某蛋白质组相关的全部肽段数，相当于用于定性的肽段数目。The total number of peptide sequences associated with the protein group (i.e. for all the proteins in the group). |
| Razor + unique peptides | 用于定量的肽段数量 | 用于定量的肽段数量。The total number of razor + unique peptides associated with the protein group (i.e. these peptides are shared with another protein group). |
| Unique Peptides | 唯一肽段数 | 专属于某一蛋白质组的肽段数量。The total number of unique peptides associated with the protein group (i.e. these peptides are not shared with another protein group). |
| Sequence coverage [%] | 肽段覆盖率 | 鉴定到的氨基酸数目占蛋白质组中得分最高的蛋白质总氨基酸数目的比例。Percentage of the sequence that is covered by the identified peptides of the best protein sequence contained in the group. |
| Mol. weight [kDa] | 蛋白质的理论分子量 | 蛋白质组中得分最高的蛋白质的理论分子量。Molecular weight of the leading protein sequence contained in the protein group. 该分子量是软件根据数据库中的蛋白质序列计算得到的。如果用以参加计算的蛋白质序列不是完整的全长序列，比如由转录组翻译而来的蛋白质序列，由此计算得到的分子量会小于完整蛋白质的分子量。 |
| PEP | 后验错误概率 | Posterior Error Probability of the identification. This value essentially operates as a p-value, where smaller is more significant. |
| iBAQ /LFQ  intensity X | 蛋白质表达量强度值 | iBAQ Intensity是基于iBAQ算法得到的样品X中的蛋白质表达量，近似等于该样品中的蛋白质绝对浓度。LFQ Intensity是基于LFQ算法得到的样品X的蛋白相对表达量，常应用于组间比较。一般Labelfree可以选择其中之一作为定量结果。 |
| Average (XXX) | 组内样品蛋白质表达量的平均值 | 组内多个强度值的平均值 |
| *Ratio* (XXX/YYY) | 两组样品间的蛋白质表达丰度的比值 | 蛋白质在XXX组样品中强度值的平均值与在YYY组样品中强度值的平均值的比值 |
| P value | 两组样品间的蛋白质相对表达量的统计学检验 | 对于p value的两种常用计算方法适用情况如下：  1） t-test算法：适用于每组样品至少包括三次及以上生物学重复的实验设计。  2） significance B算法[7]：适用于每组样品至少包括三次及以上质谱技术重复的实验设计。 |

1. **数据库介绍**

### NR数据库

NR 数据库全称为无冗余蛋白数据库（non-redundant），由美国国家生物技术信息中心（NCBI）维护，它综合了GenBank CDS区的翻译序列、Refseq蛋白库、SwissProt蛋白数据库、PIR、PDF、PDB等多个蛋白据库。NR数据库可由用户可任意提交序列，信息量丰富全面，但蛋白注释信息不完善，大部分蛋白并没有得到验证，质量很难保证。

### UniProt 数据库

UniProt 数据库由欧洲生物信息学中心（EBI）维护，旨在帮助基因组和蛋白质组以及相关的分子生物学研究⼈员提供有关蛋白质氨基酸序列的最新信息，Uniprot数据库分为SwissProt数据库和TrEmbl数据库，SwissProt 中的蛋白均经过人工校验，数据可靠性高，注释完整，而TrEmbl 由基因组序列翻译而来，未经人工校验，注释信息不全。

### GO数据库

Gene Ontology (GO)数据库通过建立一套具有动态形式的控制字集，是描述基因及基因产物（蛋白质）在细胞中的生物学功能的词汇集合，是一种标准化的词汇集合，它以一种规范化的方式描述基因及基因产物（蛋白质）已知的客观规律,解释它们在细胞内所扮演的角色。这组词汇集合从最初整合果蝇数据库，酵母基因组数据库以及小鼠基因组数据库中对基因/基因产物的功能描述开始，到现在为止已经整合了包含数百种动物、植物以及微生物等多物种数据库，因此这种词汇集合具有物种特异性并随研究的进步不断积累和更新。

### KEGG通路数据库

KEGG通路数据库的全称是京都基因与基因组百科全书（Kyoto Encyclopedia of Genes and Genomes），由日本京都大学物信息学中研究人员人工阅读海量文献，根据相关知识手工绘制的通路图，手工绘制是指人工以特定的语言格式来确定通路中各组件的相互联系; 与其他数据库相比，KEGG 通路数据库的显著特点是具有强大的图形功能，它利用图形而非文字，介绍众多的代谢/信号通路之间的关系，这样可以使研究者能够对其所要研究的通路有一个直观全面的了解，因此KEGG通路数据库是国际最常用的生物信息数据库之一。KEGG 通路数据库包含以下几方面的分子间相互作用和反应网络：新陈代谢，遗传信息加工，环境信息加工，细胞过程，生物体系统，人类疾病以及药物开发七大方面。由于KEGG通路数据库最早主要是做代谢的通路，所以代谢通路是该数据库中里面最为完善的一类。

### STRING数据库

STRING 数据库（http://string-db.org/）是一个搜寻已知的和基于预测的蛋白质之间相互作用的系统。这种相互作用既包括蛋白质之间直接的物理相互作用，也包括蛋白质之间简介的功能的相关性。与 IntAct、MINT、UniProt 等人工收录有实验证据的互作信息的数据库不同，STRING 除了有实验数据的互作信息外，还包含从 PubMed 摘要中通过文本挖掘获取的互作信息，以及利用生物信息学的方法预测的互作信息。所应用的生物信息学方法有：染色体临近、基因融合、系统进化谱和基于芯片数据的基因共表达等方法。STRING 利用一个打分机制对这些不同方法得来的结果赋予一定的权重，最终得出一个综合的可信度得分。由于 STRING 中的互作信息来源广泛，因此对于互作研究程度不足的物种，可以获得更全面的互作信息和网络，供后续研究参考。但由于其中大部分信息是基于预测的，准确性不易评判，因此，以此形成的互作网络需谨慎参考。

对于蛋白质互作研究程度较好的物种，例如：人、小鼠、大鼠、拟南芥等，可获得的互作信息已较为完善。因此，我们为获取有实验证据的准确的互作信息和网络，使得研究蛋白之间的精确调控关系更具意义，一般以IntAct(<http://www.ebi.ac.uk/intact/main.xhtml>)为主。对于蛋白质互作研究程度不足的物种，为获取更多的互作信息，我们通常采用STRING数据库的数据。

# 参考文献References

1．Yu CS1, Lin CJ, Hwang JK. Predicting subcellular localization of proteins for Gram-negative bacteria by support vector machines based on n-peptide compositions. Protein Sci. 2004 May;13(5):1402-6.

2．Finn RD, Coggill P, et al., The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. 2016 Jan 4;44(D1):D279-85.

3．Ashburner, M., et al., Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet, 2000. 25(1): p. 25-9.

4．Gotz, S., et al., High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Res, 2008. 36(10): p. 3420-35.

5．Kanehisa M, Goto S, et al. KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res. 2012; 40(Database issue): D109-14.

6．Universal sample preparation method for proteome analysis. Wisniewski, J. R., A. Zougman, et al. Nat Methods.2009. 6(5): 359-362.

7．Cox, J. and M. Mann (2008). "MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification." Nat Biotechnol 26(12): 1367-1372.

8．Schwanhäusser B, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, Chen W,Selbach M: Global quantification of mammalian gene expression control. Nature 2011, 473:337–342.

9．Cox, J., M. Y. Hein, et al. (2014). "Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ." Mol Cell Proteomics 13(9): 2513-2526.

10．Cox, J., N. Neuhauser, et al. (2011). "Andromeda: a peptide search engine integrated into the MaxQuant environment." J Proteome Res 10(4): 1794-1805.

# 生信数据分析云平台 Data analysis platform

* **APT-BioCloud**

对于以上报告的分析内容，若您仍需对分析图片的字体、样式等做调整（或数据修改后需重新分析），您也可以登录中科新生命APT-BioCloud分析云平台，进行一站式自助数据分析。将数据导入至云平台后，您可自行进行统计学、火山图、聚类分析、富集分析等多种生物信息学分析，快速便捷输出个性化的生信分析图。注册与登录地址如下，如有疑问可联系中科新生命当地销售。

注册网址：[http//cloud.aptbiotech.com/#/register](file:///D:\4%20产品相关资料\1%20产品-方案-报告模板\201910%20报告升级\2%20蛋白组label%20free\labef%20free蛋白组%20升级\http\cloud.aptbiotech.com\#/register) 登录网址：[http//cloud.aptbiotech.com/#/login](file:///D:\4%20产品相关资料\1%20产品-方案-报告模板\201910%20报告升级\2%20蛋白组label%20free\labef%20free蛋白组%20升级\http\cloud.aptbiotech.com\#/login)



# 拓展研究建议（供参考）Research suggestions

* **机制功能的创新研究方案——翻译后修饰**

蛋白质的功能与作用，除了通过表达水平进行调控，还存在另外一个层面的关键调控方式——蛋白的翻译后修饰（PTM），如磷酸化修饰、泛素化修饰、乙酰化修饰等等。蛋白的翻译后修饰能够直接改变蛋白的功能与作用，在信号转导、代谢调控、物质转运、表观调控等各种生物过程中起到广泛而重要的作用。甚至在某些生物过程中，蛋白的翻译后修饰调控比蛋白表达调控更为重要。目前大多数研究领域对蛋白的翻译后修饰研究仍相对较少，翻译后修饰是更创新的研究方向和切入点。从文章发表的角度来说，目前翻译后修饰组也呈现出发文难度低、杂志档次相对高的特点。

中科新生命提供各种主流翻译后修饰检测服务：

|  |  |
| --- | --- |
| 修饰谱分析 | ■ 修饰表达谱 |
| 修饰定量分析 | Label free 修饰:  ■ 磷酸化 ■ N-糖基化 ■ 泛素化  ■ 乙酰化 ■ 甲基化 ■ 琥珀酰化  ■ 丙酰化 ■ 丙二酰化 ■ 戊二酰化  iTRAQ/TMT修饰：  ■ 磷酸化 ■ 磷酸化（深度） |
| “一站式”服务 | ■ 修饰组定量分析+PRM |
| 修饰鉴定分析 | ■ 纯化蛋白修饰位点鉴定 |

* **打通机制与表型的研究模式——代谢组分析**

生物活动是一个复杂的过程，由多个生物层面共同参与和决定。基因、转录、蛋白主要决定生物功能与调控机制，而代谢物则是主要的物质来源和表型基础。蛋白质组水平的研究仅能解释功能和机制，但缺乏对表型的直接描述。因此，将蛋白质组与代谢组进行联合分析，可以将功能、调控机制与代谢表型直接打通，从而更系统、全面的解析某个生物过程和机制，是目前提高组学文章档次的最有效手段之一。

中科新生命拥有业内最大的代谢物标准品数据库，已合作发表多篇高水平代谢组学文章，包括蛋白质组-代谢组联合分析文章。我们是业内仅有的同时拥有蛋白组、代谢组双自主尖端服务平台的服务商，提供蛋白组-代谢组、蛋白组-脂质组、修饰蛋白组-代谢组、修饰蛋白组-脂质组等多做蛋白-代谢联合分析服务。

* **系统性研究的高端方案——多组学联合研究**

以往研究主要集中在某一阶段或者某一层面，但是生物体是一个复杂的调控系统，生物性状多样性的主要遗传基础是基因表达多样性， 并且受多层级复杂而精细的遗传调控。每个调控过程不是独立的，而整个调控过程是互相补充的。因此要研究生命过程的调控机制，单一某个层次的研究明显受局限，需要考虑到多个层次组学的影响。可利用多组学的方式并对数据进行有机整合，全面性揭示生物学问题。目前中科新生命率先推出完整的多组学分析解决方案，包括转录+蛋白，蛋白（转录）+代谢，蛋白+修饰等等多种联合思路，可以满足解决不同的生物学问题

# 组学期刊投稿指南（供参考）Publication recommendation

* **第一档期刊（影响因子9 分以上）**

组学期刊：Nature Genetics，Genome Research， Microbiome，Molecular Systems Biology等

综合性期刊： Cell，Nature，Science，Nature Communications，PNAS，Cell Metabolism，Molecular Cell，Cell Research等

医学专业期刊：Nature Medicine，Cancer Cell，Cancer Research，European Heart Journal，Circulation，Journal of the American College of Cardiology，Gut，Gastroenterology，Hepatology，Diabetes Care，Molecular Psychiatry，Immunity 等

农林专业期刊：Nature Plants，Molecular Plant，Plant Cell等

**组学思路**：在这类高水平期刊中，组学驱动的研究主要包括：1.大队列研究，筛选临床生物标志物以及进行疾病分子分型等。2. 组学+后续功能机制研究。3. 结合时下的热点，如肠道微生物进行相关研究，阐明肠道微生物与疾病的相关性及相互作用机制。

* **第二档期刊（影响因子4~9分）**

组学期刊：Molecular & Cellular Proteomics，Metabolism，Metallomics，Gigascience，Proteomics & Bioinformatics等

综合期刊： Scientific Data，Scientific Reports，Cell Reports，Progress in Lipid Research，Frontiers in Microbiology，mSystems等

医学专业期刊：Advances in Cancer Research，International Journal of Cancer，Diabetologia，Thyroid，Diabetes，Oncogene，mBio等

农林领域专业期刊：New Phytologist，Plant Biotechnology Journal，Plant Physiology，Plant Journal，Plant Cell and Environment，Food Chemistry，Journal of Experimental Botany等

**组学思路**：发表在该分数段期刊的组学文章，1. 对于蛋白组学，修饰组学而言，要求组学+功能验证与机制研究（如表达量验证，基因敲除，点突变，位点特异性抗体验证，蛋白活性验证，后续功能实验，动物模型等）。2. 对于脂质组学是个例外，研究基础浅，新颖性高，纯脂质组学也能发表文章，无需后续验证。

* **第三档期刊（影响因子4分以下）**

组学期刊：BMC Genomics， Proteomics，Journal of Proteomics，Journal of Proteome Research，Metabolomics等

综合期刊：Plos One等

农林领域专业期刊: BMC Plant Biology，Journal of Agricultural and Food Chemistry，Frontiers in Plant Science等

**组学思路**：这类文章一般是最常见也是最简单的组学研究思路。1. 蛋白质组学方向：组学+差异蛋白WB或PRM验证。2.蛋白质翻译后修饰组学方向：则以修饰组学数据分析为主。3. 代谢组学方向：组学（+差异代谢物MRM验证）。

# 部分合作发表文章Cooperative Projects Papers

* **医学：**

[1] Lysozyme-Assisted Photothermal Eradication of Methicillin-Resistant Staphylococcus aureus Infection and Accelerated Tissue Repair with Natural Melanosome Nanostructures. **ACS Nano**

[2]One-Carbon Metabolism Supports SAdenosylmethionine and Histone Methylation to Drive Inflammatory Macrophages. **Molecular Cell**

[3] Early Prediction of Developing Type 2 Diabetes by Plasma Acylcarnitines: Population-Based Study. **Diabetes Care**

[4] Hepatocellular Carcinoma-Associated Protein TD26 Interacts and Enhances Sterol Regulatory Element-Binding Protein 1 Activity to Promote Tumor Cell Proliferation and Growth. **Hepatology**

[5] Polyunsaturated fatty acids metabolism, purine metabolism and inosine as potential independent diagnostic biomarkers for major depressive disorder in children and adolescents. **Mol Psychiatry**

[6] CLOCK Acetylates ASS1 to Drive Circadian Rhythm of Ureagenesis. **Mol Cell**

[7] EGFR modulates monounsaturated fatty acid synthesis through phosphorylation of SCD1 in lung cancer. **Molecular Cancer**

[8] Protein phosphatase 2A has an essential role in promoting thymocyte survival during selection. **Proc Natl Acad Sci U S A**

[9] O-GlcNAcylation destabilizes the active tetrameric PKM2 to promote the Warburg effect. **Proc Natl Acad Sci U S A**

[10] Proteomic analysis of solid pseudopapillary tumor of the pancreas reveals dysfunction of the endoplasmic reticulum protein processing pathway. **Molecular & Cellular Proteomics**

* **植物：**

[1] Natural alleles of a proteasome α2 subunit gene contribute to thermotolerance and adaptation of African rice. **Nat Genetics**

[2]Maize Oxalyl-CoA Decarboxylase1 Degrades Oxalate and Affects the Seed Metabolome and Nutritional Quality. **Plant Cell**

[3] OsSPL3, a SBP-Domain Protein, Regulates Crown Root Development in Rice. **Plant Cell**

[4] RRM Transcription Factors Interact with NLRs and Regulate Broad-Spectrum Blast Resistance in Rice. **Mol Cell**

[5] Global analysis of sumoylation function reveals novel insights into development and appressorium-mediated infection of the rice blast fungus.**New Phytologist**

[6] Quantitative Phosphoproteomic and Metabolomic Analyses Reveal GmMYB173 Optimizes Flavonoid Metabolism in Soybean under Salt Stress. **Molecular & Cellular Proteomics**

[7] Label-Free Quantitative Proteomics of Lysine Acetylome Identifies Substrates of Gcn5 in Magnaporthe oryzae Autophagy and Epigenetic Regulation.**mSystems**

[8] Phototrophy and starvation-based induction of autophagy upon removal of Gcn5-catalyzed acetylation of Atg7 in Magnaporthe oryzae. **Autophagy**

[9] Integrative Transcriptome and Proteome Analysis Identifies Major Metabolic Pathways Involved in Pepper Fruit Development. **J Proteome Res**

[10] Proteomics integrated with metabolomics: analysis of the internal causes of nutrient changes in alfalfa at different growth stages. **BMC Plant Biology**

* **动医/动科/食品等其他：**

[1] Fecal Microbiota Transplantation Beneficially Regulates Intestinal Mucosal Autophagy and Alleviates Gut Barrier Injury. **mSystems**

[2] Quantitative Phosphoproteomic Analysis among Muscles of Different Color Stability using Tandem Mass Tag Labeling.**Food Chem**

[3] Comparative Analysis of Whey N-Glycoproteins in Human Colostrum and Mature Milk Using Quantitative Glycoproteomics.**J Agric Food Chem**

[4] N-glycosylation proteomic characterization and cross-species comparison of milk fat globule membrane proteins from mammals. **Proteomics**

[5] Metabolomics Reveals that Crossbred Dairy Buffaloes Are More Thermotolerant than Holstein Cows under Chronic Heat Stress. **Journal of agricultural and food chemistry**

[6] The Synergistic Effect of Exogenous Glutamine and Rifampicin Against Mycobacterium Persisters. **Frontiers in Microbiology**

