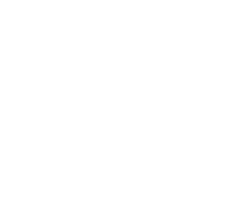


|  |
| --- |
| **非数据依赖型采集（DIA）蛋白质组学定量分析报告** |
|  |
| **May 11**  **上海中科新生命生物科技有限公司** |



**DIA技术**

|  |
| --- |
| 全扫描的蛋白质组学定性 定量技术  蛋白质组学技术，高度依赖于质谱分析。  现代生物质谱分析蛋白质组学的复杂样品，采取各种不同的扫描模式。 |
| *“新一代的DIA技术，结合卓越的生物信息学软件，能为您提供更优质、更全面的蛋白质组学产品”* |
| 由于质谱仪器的技术和物理限制，传统的蛋白质组学分析一般采用DDA（数据依赖性采集）进行扫描，结合动态排除等软件技术可以获得较高质量的数据。  要想获得更全面的扫描结果，最近几年技术进步，越来越多的情况下，质谱仪器可以按照DIA（非数据依赖性采集）的形式进行扫描，一次性获得整个质量范围内所有多肽分子的图谱，结合生物信息学软件的后期分析，能够获得更全面的数据结果。 |

FOR

Approved by

Project No.

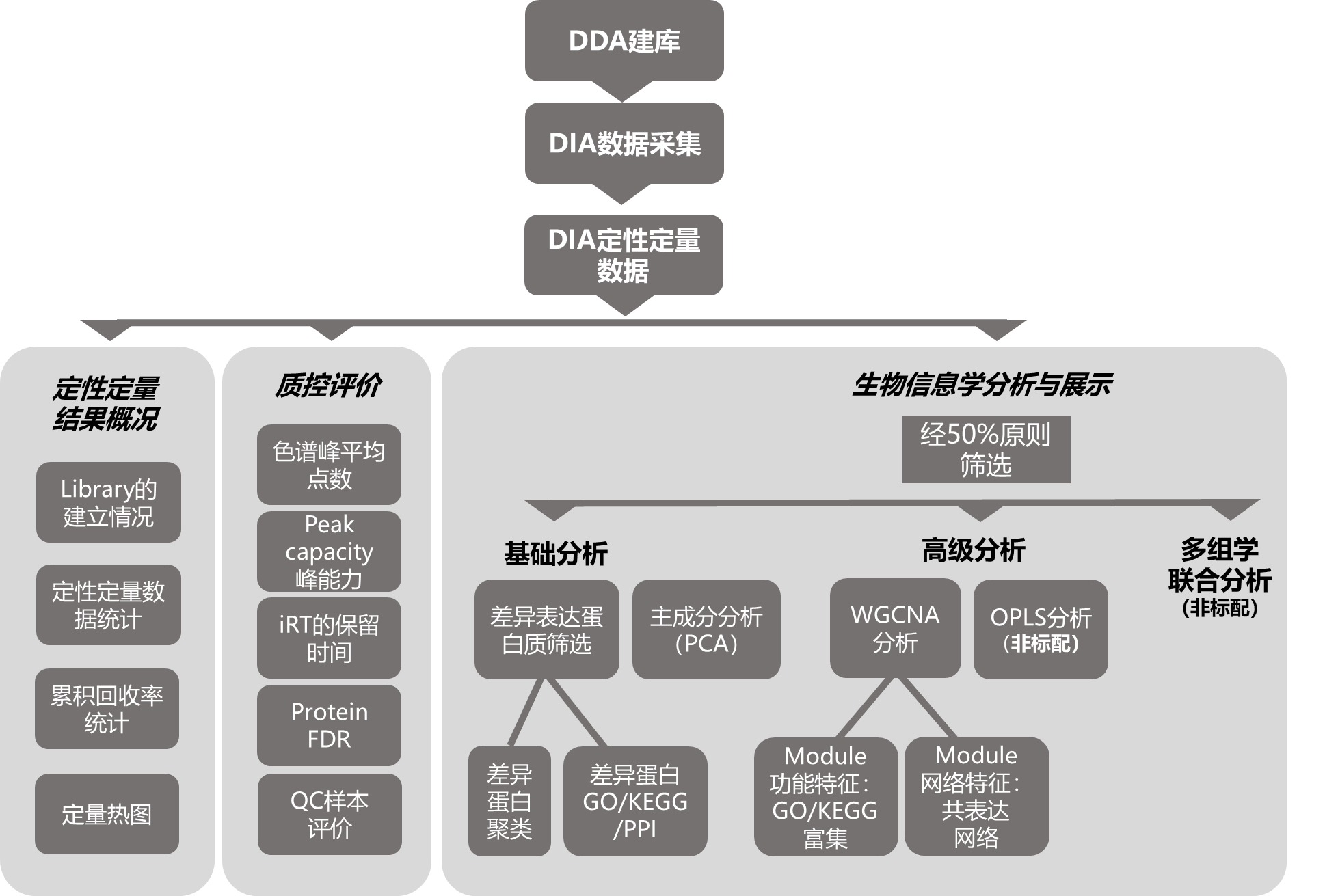
Date

**项目名称**

**Abstract**

上海中科新生命生物科技有限公司接受客户方的sample\_type样品，共sample\_num例，经过样品制备、蛋白质定量和电泳、蛋白质酶解、LC-MS/MS分析等步骤，完成了DIA分析工作。首先，其中使用等量合并后的样品作为pool，进行高低丰度分离后，将高低丰度样本进行酶解，低丰度样本进行 HPRP分级（HPRP\_num份），将上述样品分别进行了LC-MS/MS（QE-HFX\_DDA模式）分析，得到蛋白质组数为DDA\_protein的数据库，作为DIA工作的数据库。然后，对各样本分别进行了LC-MS/MS（QE-HFX\_DIA模式）分析，并采用上述数据库进行了定性定量分析。最后采用生物信息学分析工具对数据进行展示与深入挖掘。

**整体项目路线**，如下：



**注：**1. WGCNA与OPLS需要提前沟通，是否提供表型数据用于分析

2. 非标配内容，需要额外收取费用

**Index**

DIA技术 ii

全扫描的蛋白质组学定性定量技术 ii

Introduction 2

Glossary 3

**Chapter 1 定性定量结果概况** 4

**Chapter 2 DIA数据的质量控制** 7

**Chapter 3 生物信息学分析与展示** 12

**单维统计学分析与生信分析** 12

**Chapter 4 Methodology** 25

**DIA分析方法** 25

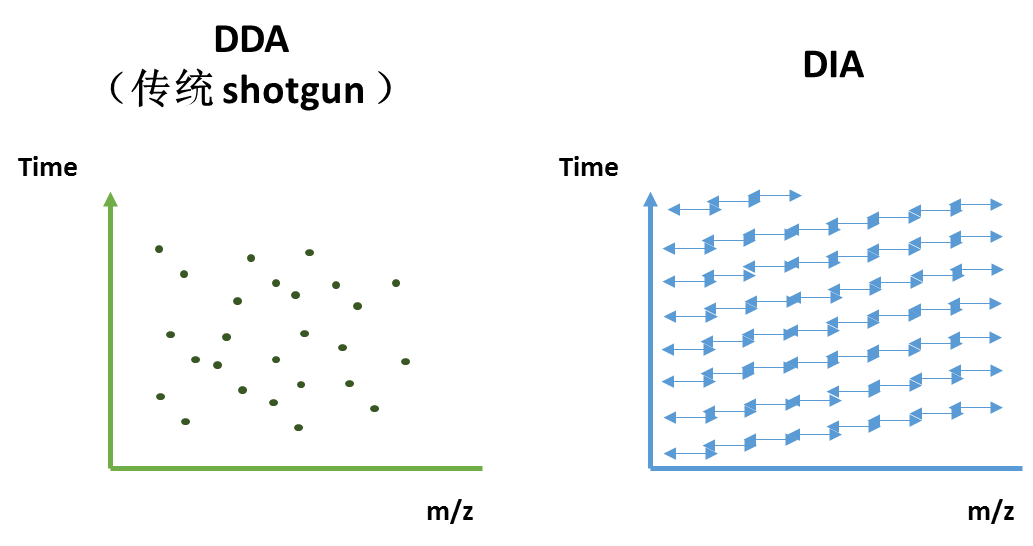
**生物信息学分析方法** 28

**Reference** 25

**Appendix A** 31

# **Introduction**

DIA（data-independent acquisition）技术是近年来发展起来的一种新的质谱技术。与传统的DDA（data-dependent acquisition）质谱技术相比，DIA采用了不同的数据扫描模式：将质谱整个全扫描范围分为若干个窗口，然后对每个窗口中的所有离子进行检测、碎裂，从而无遗漏、无差异地获得样本中所有离子的信息。与DDA技术相比，DIA技术的优势包括：（1）采集所有的离子信息，实现更高的数据覆盖度；（2）减少采集的随机性，实现极高的检测重现性、稳定性；（3）采用碎片离子定量，定量精密度、准确性、线性范围大大提高。基于上述技术优势，DIA技术尤其适用于大规模样本的高度覆盖、稳定和可追溯地分析。



DDA和DIA分析原理的比较

# **Glossary**

**DDA：** Data Dependent Acquisition，在本文中用于建立DIA分析流程中的library。

**DIA：**Data Independent Acquisition，本文中特指基于Thermo Scientific™ 静电场轨道阱 Orbitrap 的数据非依赖性采集。

**Spectronaut Pulsar X：** Proteomics software for the analysis of data independent acquisition (DIA) measurements. 来自Biognosys公司的著名DIA数据分析软件。

**iRT：** Retention time normalization kit for time-resolved mass spectrometry. 来自Biognosys公司的DIA校正试剂盒。

**Q Exactive HF/ Q Exactive HF-X：**The Thermo Scientific™ hybrid quadrupole-Orbitrap mass spectrometer features an ultra-high-field Orbitrap analyzer which doubles its speed and resolution – more samples run and amazing data quality.

**MaxQuant：**来自Mann M. 实验室的著名蛋白质组学软件，在本文中用于建立library时查库使用。

**Direct DIA：**Integrated database search engine called Pulsar enabling spectral-library free DIA workflow. 来自于Biognosys公司的算法，可以对DIA数据进行直接查库，用于增加library的容量，提高DIA最终定性定量结果的数量。

## **Chapter 1 定性定量结果概况**

#### Library的建立情况:

主要通过对所有样本的混合进行样品酶解和HPRP分级，经过data dependent acquisition (DDA)质谱数据采集，然后使用MaxQuant进行搜库鉴定，通过Spectronaut pulsar X软件构建图谱库，作为后续data independent acquisition (DIA)定量的数据库。本项目建立的DIA library（数据库），其中包含的肽段和蛋白质的统计信息如下：

表1-1  DDA library的数据统计

|  |  |  |
| --- | --- | --- |
|  | **Protein groups** | **Peptide** |
| Spectral library |  |  |

#### 定性定量数据统计

将项目中每个样品独立进行样品制备、蛋白质酶解后，分别上机进行DIA（data independent acquisition）分析。所得的DIA原始文件，导入Spectronaut Pulsar X进行分析，得到的部分样本的定性定量数量见表1-2：

表1-2 DIA蛋白质鉴定结果统计

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

附件：

1. Appendix-A\\附件1\\ protein\_Report：蛋白质鉴定定量列表

|  |  |  |
| --- | --- | --- |
| **表头** | **定义** | **描述** |
| PG.ProteinAccessions | 蛋白质登录号 | 蛋白质序列数据库中的蛋白质编号 |
| PG.ProteinDescriptions | 蛋白质信息描述 | 蛋白质名称 |
| X.raw.PG.Quantity | 蛋白质的定量值 | “X”为样本数据名称；“X”为QC时，该值为QC样本的蛋白质定量值 |

1. Appendix-A\\附件2 \\ peptide\_Report：肽段鉴定定量列表

|  |  |  |
| --- | --- | --- |
| **表头** | **定义** | **描述** |
| EG.PrecursorId | 肽段序列 | 肽段的氨基酸序列。以肽段“\_ADTLTPEEC[Carbamidomethyl (C)]QQFK\_.2”为例：“[Carbamidomethyl (C)]”为该半胱氨酸上存在碘乙酰胺化修饰；“\_.2”为二价肽 |
| X.raw.EG.TotalQuantity (Settings) | 肽段序列的定量值 | “X”为样本数据名称；“X”为QC时，该值为QC样本的肽段定量值 |
| PEP.IsProteinGroupSpecific | True or False | True or False. Tells you whether the peptide only belongs to one Protein Group. |
| PEP.IsProteotypic | True or False | Tells you whether a given peptide was proteotypic within the context of all provided FASTA files. |
| PG.Cscore | 得分 | A high score indicates high quality identifications |

#### 累积回收率统计

DIA数据的定性定量数量随着样本数量增加的累积结果见图1-1。可以看到随着样本数量的积累，总共定性定量到的数据量、各样本中共有的数据量，均逐渐达到平台期。

[Protein\_Group\_Profile.png]

图1-1  DIA数据的蛋白质鉴定数据累积回收率

横坐标：样本累积数量；纵坐标：定性定量到的数据量；绿色线：数据库中包含的数据量；红色纵列：所有样本累积总共定性定量到的数据量；蓝色纵列：所有样本中均出现的共有定性定量数据量；protein group profile：定性定量到的数据量占数据库中包含的数据量的比例。

附件：

1. Appendix-A\\附件3

#### 定量热图：

将各样本中鉴定到的蛋白质的定量信号进行热图展示，直观反映各样本的蛋白表达模式特点，呈现各样本中蛋白质表达与分布规律，如下为本项目的热图分析结果：

[Heatmap.png]

图1-2 DIA数据的蛋白质定量热图

横坐标：样本名称；纵坐标：蛋白质登录号；蛋白的表达量以不同颜色展示，黄色越深代表信号越强，蓝色越深代表信号越弱，白色代表没有定量信息

附件：

1. Appendix-A\\附件3

## **Chapter 2 DIA数据的质量控制**

#### 色谱峰的平均数据点

在DIA数据中，软件要实现对肽段色谱峰的准确积分定量，需要在LC-MS/MS数据中能够采集到足够密度的数据，以便实现有效的定量积分。一般认为，在DIA数据中数据采集的密度，应至少满足每个色谱峰有5个数据点。在本项目中，DIA数据的Average data points per peak为aver\_peak，满足定量分析的要求，具体数据如下：

[AVERAGE\_DATA\_POINT\_PER\_PEAK.png]

图2-1 Average data points per peak 色谱峰平均数据点

横坐标：样本上机顺序；纵坐标：数据点数量

附件：

1. Appendix-A\\附件3

#### Peak capacity 峰容量

代表色谱柱的分离分析能力，是LC-MS/MS分析工作中的重点。DIA分析，需要高效率的nano LC色谱分离能力，在本项目中，平均的peak capacity是peak\_capacity，（一般情况，正常的nano LC为200左右，越高越好）具体数据如下：

[Peak\_Capacity.png]

图2-2 DIA试验中色谱柱Peak capacity 统计数据

横坐标：样本上机顺序；纵坐标：peak数量；绿色线：所有肽段的数据；红色线：iRT内标的数据

附件：

1. Appendix-A\\附件3

#### iRT的保留时间

DIA实验的一个关键点是使用内标校正肽段（本项目使用iRT Kit），iRT的保留时间可用于对各样本间的数据进行对齐校正，进行各个样品LC-MS/MS原始数据的对比分析。因此，iRT肽段在各个样品分析时的色谱行为稳定性较为重要。如下图为本项目中的iRT Kit的各肽段在色谱上的洗脱时间数据，可见主要的iRT均被检测到，并且保留时间整体较稳定：

[iRT.png]

图2-3  iRT肽段（校正肽段）的洗脱时间

横坐标：样本上机顺序；纵坐标：保留时间

附件：

1. Appendix-A\\附件3

#### Protein FDR

DIA分析和其他蛋白质组学分析一样，在蛋白质的定性时，使用FDR方法筛选得分阈值。本项目中，使用Q Value 0.01作为项目的定性阈值标准，相当于FDR 0.01。如下为本项目的Protein FDR图谱，红色为decoy（反库）的分布情况，灰色为target（正库），可见在Q Value 0.01标准下的定性结果高度可靠：

[Protein\_FDR.png]

图 2-4 Protein FDR分布数据

Cscore：相当于蛋白质可信度打分，越大越好；横坐标：protein的Cscore值；纵坐标：某个Cscore得分下protein的数量；黑色虚线：1% Qvalue（相当于1% FDR）标准线，标准线处的Cscore越高越好

附件：

1. Appendix-A\\附件3

#### QC样本评价

为监测和评价系统的稳定性及实验数据的可靠性，在样本队列中每间隔一定数量的样本插入一个QC样本（一般为所有样本的混样），并对整个实验过程中插入的各QC样本的数据一致性进行评价。主要采用组内变异系数（CV）、主成分分析 (Principle Component Ananlysis, PCA)和Pearson相关性分析来评估QC。CV值越小，PCA中组内样本聚集性越高，相关性系数越接近1，说明实验体系稳定。该项目QC样品CV中值为CV\_median，%，相关系数大于0.9，说明体系稳定。

[CV.png]

图2-5 QC的CV值分布图

纵坐标：各QC样本的变异系数CV %

[DELE-QC.png]

图2-6 QC样本的PCA得分图

横坐标：第一主成分；纵坐标：第二主成分；每个点代表一个样本

[Scatterplot\_QC.png]

图2-7 QC样本相关性分析

横坐标和纵坐标分别标记强度值的对数值,一般相关系数大于0.9表明相关性较好。

附件：

1. Appendix-A\\附件3

## **Chapter 3 生物信息学分析与展示**

本实验结果中，为了保证后续生物信息与统计学分析的有效性、准确性，根据通用的原则，将在50%以上样本中均存在定量值的蛋白用于后续的统计学与生物信息学分析。本实验总共鉴定到total\_pro\_num个蛋白，其中在50%以上样本中均存在定量值的蛋白有half\_pro\_num个蛋白。

### **单维统计学分析与生信分析**

#### 差异表达蛋白质筛选

采用表达以倍数变化大于upRatio倍（上调大于upRatio倍或者下调小于downRatio倍）且p value小于0.05的标准筛选差异表达蛋白质。各比较组的差异表达蛋白质数目见下表。定量蛋白质结果以火山图（volcano plot）形式展示，参见图3-1。

表3-1 蛋白质定量结果统计

|  |  |  |  |
| --- | --- | --- | --- |
| **Comparisons** | **Up-** | **Down-** | **All-** |
|  |  |  |  |

Comparisons：差异比较组；Up-：上调差异表达蛋白质；Down-：下调差异表达蛋白质；All：所有差异表达蛋白质。

附件：

1. 单维统计学与生信分析[\\质谱鉴定和定量结果](file:///\\质谱鉴定和定量结果)

[Volcano\_Plot.png]

图3-1 groupvs组火山图

采用两组样本间的蛋白质表达差异倍数（Fold change）和T检验得到的Pvalue两个因素共同绘制火山图，用于显示两组样本数据的显著性差异。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），图中红色点为显著性差异表达的蛋白（倍数变化大于差异倍数且P value<0.05），黑色点为无差异变化的蛋白。

2) 单维统计学与生信分析\\ Volcano

#### Venn图：

[venn.png]

图3-2 Venn

1. ) 单维统计学与生信分析\\ Venn

#### 蛋白质丰度比分布图：

[Ratio\_Distribution]

图3-3 Ratio\_Distribution\_groupvs

1. ) 单维统计学与生信分析\\ Ratio\_Distribution

#### 差异表达蛋白质的聚类分析（Clustering）

本项目采用层次聚类算法（Hierarchical Cluster）对比较组的差异表达蛋白质分别进行聚类分析，并以热图（Heatmap）形式进行数据展示，如下图表明，通过倍数变化大于

聚类分析是一种常用的探索性数据分析方法，其目的是在相似性的基础上对数据进行分组、归类。聚类分组的结果中，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。在聚类分析过程中，聚类算法会对样本（Sample）和变量（Variable，在蛋白质组学研究中通常指蛋白质的定量信息）两个维度进行分类。对样本的聚类结果可以检验所筛选的目标蛋白质的合理性，即这些目标蛋白质表达量的变化可否代表生物学处理对样本造成的显著影响；目标蛋白质的聚类结果可以帮助我们从蛋白质集合中区分具有不同表达模式的蛋白质子集合，具有相近表达模式的蛋白质可能具有相似的功能或者参与相同的生物学途径，或者在通路中处于临近的调控位置。参考资料：<http://en.wikipedia.org/wiki/Cluster_analysis>。以一个比较组为例，所得的表达差异蛋白进行聚类分析的结果如下：

[cluster.png]

图3-4差异表达蛋白质聚类分析

层次聚类结果以树型热图表示，图中每行代表一个蛋白（即纵坐标为显著性差异表达的蛋白），每列代表一组样品（横坐标为样品信息），显著性差异表达的蛋白在不同样品中的表达量的对数值(Log2Expression）值以不同颜色在热图中展现，其中红色代表显著性上调蛋白，蓝色代表显著性下调蛋白，灰色部分代表无蛋白定量信息。

所有组别的聚类分析结果输出文件：

1. 单维统计学与生信分析\\比较组 \\ cluster

#### 差异表达蛋白质的Gene Ontology (GO)功能富集分析

在进行蛋白质组学研究时，我们的研究对象是细胞、组织或生物体中全部蛋白质的集合。对于后续功能研究和分析来说，了解每一个蛋白质的功能、细胞定位和参与的生物学过程固然重要，但是对于高通量组学来说，了解哪一些功能或生物学途径受到生物学处理的显著影响是首要任务。因此，需要从更为系统和概括的层次和角度，对所研究的蛋白及其功能进行概括和分析。我们采用Blast2Go（https://www.blast2go.com/）软件对本项目鉴定的所有蛋白质进行GO功能注释，然后通过Fisher精确检验方法对差异表达蛋白质进行GO功能富集分析，如下图所示，分析结果表明比较组groupvs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。

基因本体（Gene Ontology, http://www.geneontology.org/）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，并以以下三个方面描述生物体中基因和基因产物的属性：参与的生物过程（Biological Process, BP），分子功能（Molecular Function, MF）和细胞组分（Cellular Component, CC），关于GO功能注释的更多信息请参考: <http://en.wikipedia.org/wiki/Gene_Ontology>。GO数据库中的功能并不是完全并列的关系，彼此之间是（不完全严谨的）层级关系，并由此形成树状结构。每一个层级与其上一层级相比，对于蛋白质的属性描述都更为精准一些，层级越高（Level级别数字越大），对于蛋白质功能描述的更为具体详细。

GO功能注释的显著性富集分析是通过Fisher精确检验（Fisher’s Exact Test）来评价某个GO功能条目的蛋白质富集度的显著性水平。差异表达蛋白质的功能富集分析是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根据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）。不同于蛋白质功能注释以蛋白质为单位进行注释，差异表达蛋白质的功能富集分析以GO功能条目为单位，结果可以直接揭示所有差异表达蛋白质的整体功能富集特征，这些被显著富集的GO条目往往涉及到研究者最为关心的生物学功能。

[go\_enrichment.png]

图3-5 GO功能富集分析

图中横坐标表示富集到的GO功能分类，分为生物过程（Biological Process，BP），分子功能（Molecular Function，MF）和细胞组分（Cellular Component，CC）三大类；纵坐标表示每个功能分类下的差异蛋白质数目；条形图颜色表示富集的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，颜色由橘色至红色渐变，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；条形图上方标签显示富集因子（richFator<=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。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。

所有组别的GO富集结果输出文件：

1. 单维统计学与生信分析\\\比较组 [\\GO\\GO](\\\\GO\\\\GO)功能注释表
2. 单维统计学与生信分析\\\比较组 [\\GO](file:///\\GO)\\GO\_Enrichment.tif

#### 差异表达蛋白质的KEGG通路富集分析

在生物体中，蛋白质并不独立行使其功能，而是不同蛋白质相互协调完成一系列生化反应以行使其生物学功能。因此，通路分析是更系统、全面地了解细胞的生物学过程、性状或疾病的发生机理、药物作用机制等最直接和必要的途径。通过对显著性差异表达的蛋白质进行KEGG通路注释，旨在帮助我们了解这些蛋白可能参与的代谢或信号通路，从而显示蛋白质从细胞表面到细胞核的一系列变化过程，揭示参与该过程的一系列生物学事件和作用因子，提示某一过程的中断或变化可能导致的生物学后果等。如图所示，通过Fisher精确检验方法对比较组groupvs的差异表达蛋白质进行KEGG通路富集分析，KeggEnrich-top5等重要通路发生了显著变化。以KeggEnrich-top1 通路（排名第一位的通路图）为例，图展示了该通路图的详细信息。

KEGG (Kyoto Encyclopedia of Genes and Genomes, <http://www.kegg.jp/>)是常用于通路研究的数据库之一，它是由研究人员阅读海量文献以特定的图形语言描述众多的代谢途径以及各途径之间的相互关系。KEGG数据库收录了新陈代谢，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，KEGG数据库相关资料参见: <http://www.genome.jp/kegg/pathway.html>。KEGG通路富集分析方法与GO富集分析相似，即以KEGG通路为单位，以鉴定的总蛋白质为背景，通过Fisher精确检验（Fisher’s Exact Test），来分析计算各个通路蛋白质富集度的显著性水平，从而确定受到显著影响的代谢和信号转导途径。

[KEGG\_Enrichment.png]

图3-6 KEGG通路富集分析

图中纵坐标表示显著富集的KEGG通路；横坐标表示每条KEGG通路中包含的差异表达蛋白质数目；条形图颜色表示富集的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，颜色由橘色至红色渐变，越接近红色代表P值越小，对应KEGG通路富集度的显著性水平越高；条形图上方标签显示富集因子（richFator<=1）,富集因子表示参与某KEGG通路的差异表达蛋白质数目占所有鉴定到的蛋白质中参与该条通路的蛋白质数目的比例。一般情况下，KEGG通路富集结果中P值越小（P<0.05），统计学上KEGG通路富集越显著，而KEGG通路下包含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通路的影响程度大小，因此可以结合两方面因素，选择较为感兴趣的代谢或信号转导途径以及显著性影响这些途径的差异表达蛋白进行后续生物学实验验证或机制研究。

[Pathway.png]

图3-7 KEGG通路图举例

图中红色方框表示上调蛋白，深绿色方框表示下调蛋白。

所有组别的KEGG富集结果输出文件：

1. 单维统计学与生信分析\\比较组\\KEGG \\KEGG通路注释统计表
2. 单维统计学与生信分析\\比较组\\KEGG \\map文件夹
3. 单维统计学与生信分析\\比较组\\KEGG \\ KEGG\_Enrichment.tif

#### 差异表达蛋白质的相互作用网络分析（PPI）

在生物体中，蛋白质并不是独立存在的，其功能的行使必须借助于其他蛋白质的调节和介导。这种调节或介导作用的实现首先要求蛋白质之间有结合作用或相互作用。对蛋白质之间的相互作用及相互作用形成的网络进行研究，对于揭示蛋白质的功能具有重要意义。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。同时，将蛋白质相互作用网络分析和通路注释的结果相结合，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。以其中一个比较组的差异表达蛋白质构建的蛋白质互作网络图展示如下：

[ppi.png]

图3-8差异表达蛋白质相互作用网络

蛋白质互作网络中结点表示蛋白质，线表示蛋白质与蛋白质之间的相互作用。网络中黄色节点为差异表达蛋白质；蓝色节点为互作网络数据库中收录的可以直接与差异表达蛋白质相互作用的其他蛋白质。网络中，与某个蛋白质A直接相互作用的蛋白质数目称为该蛋白质A的连接度，通常来讲，蛋白质的连接度越大，该蛋白质发生变化时整个系统受到的扰动就越大，该蛋白质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。

所有组别的PPI分析结果输出文件：

1. 单维统计学与生信分析\\比较组 \\PPI分析结果文件夹\\PPI互作网络图

### **多维统计学分析与生信分析**

#### 主成分分析（PCA）

主成分分析（Principal Component Analysis, PCA）是一种非监督的数据分析方法。在主成分分析中，样本的蛋白表达轮廓越相似，则聚集程度越高，样本差异越大，则距离越原，因此能从总体上反映样本组间和组内的变异度。它将原本鉴定到的所有蛋白质重新线性组合，形成一组新的综合变量，同时根据所分析的问题从中选取几个综合变量，使它们尽可能多地反映原有变量的信息，从而实现降维的目的，以便从整体上反应各样本的蛋白组轮廓特点。

[pca.png]

图3-9 PCA得分图

图中t[1]代表主成分1、t[2]代表主成分2，每个点代表一个样本

附件：

1. 多维统计学与生信分析[\\PCA](file:///\\PCA)

#### WGCNA分析（weighted gene co-expression network analysis）

加权共表达网络分析用于分析多样本蛋白的表达模式，通过将表达模式相似的蛋白聚类分析，可鉴定出高度协同变化的蛋白模块（module），该模块中的蛋白可能协同参与某一生物过程并发挥重要作用；并根据蛋白模块的内连性和模块与特定性状或表型之间的关联，筛选候补生物标记物或治疗靶点，在疾病以及其他性状与蛋白关联分析等方面的研究中广泛应用。

WGCNA算法（weighted gene co-expression network analysis，基因加权共表达网络分析）是构建加权共表达网络的常用算法。首先假定蛋白网络服从无尺度分布，并定义蛋白共表达相关矩阵、蛋白网络形成的邻接函数，然后计算不同节点的相异系数，并据此构建分层聚类树(hierarchical clustering tree)，该聚类树的不同分支代表不同的蛋白模块(module)，模块内蛋白共表达程度高，而分属不同模块的蛋白共表达程度低。最后，探索蛋白模块与特定表型或疾病的关联关系，最终达到鉴定疾病治疗的靶点蛋白、蛋白网络的目的。

基于最优软阈值计算蛋白间表达量相关系数，构建蛋白层次聚类树划分共表达模块，并对表达模式相近的模块进行合并，获得最终划分的蛋白共表达模块：

[PlotModule.png]

图3-10 蛋白共表达模块划分

横坐标展示的是不同的模块，每个模块用不同的颜色表示，灰色表示无法进行归类的模块；每个树杈代表一个蛋白，树杈的距离体现了蛋白的相似程度，树杈越短相似性越高，表达模式相近的蛋白聚集在一个分支里。

模块特征值一定程度上代表模块内蛋白的表达模式，通过将模块特征值与表型性状或临床指标特征相关联，计算二者之间的相关系数，可以识别出与表型性状有显著关联的共表达模块（如果未有临床/样本性状数据，我们则是按照样本分组信息构建性状数据，从而将模块特征值与样本分组进行相关性分析）。

如下图所示，以分组作为性状，可以获得与该性状相关的共表达模块情况。进一步可从模块与表型性状相关性图中挑选感兴趣的性状所对应的模块，一般筛选标准：相关性数值越接近±1，相关性检验P值小于0.05，表明该模块是决定该性状的关键模块。

[Module-trait.relationships.png]

图3-11 模块与表型性状相关性热图

横坐标为性状因素； 左边纵坐标的色块为不同的 module 类型（以不同颜色命名不同 module）；中间的大色块代表各的蛋白 module，同时在色块上标注了相关性系数值及p-value（括号中的值）；右边 color bar 颜色代表相关系数大小，相关系数值介于-1 和+1 之间，红色代表正相关、绿色代表负相关，相关性越高则颜色越深，相关性越低则颜色越浅

观察模块特征特征值的聚类树杈图和聚类热图能有效的识别出表达模式更加相似的模块组。如下图所示：

[Eigengene.dendrogram.heatmap.png]

图3-12 模块（性状）关联聚类图

上半张聚类树杈图展示了不同模块（性状）之间的相异程度，距离越远则说明相异程度越高。下半张图展示了不同模块（性状）之间的相似程度，颜色越接近红色则说明相似程度越高

利用Topological Overlap Matrix（TOM）进行聚类热图分析，从而反映不同共表达模块内蛋白表达量之间的相关性

[ProteinNetwork\_heatmap\_plot.png]

图3-13 蛋白网络热图

聚类树的每个树杈表示一个蛋白，树杈的距离体现了蛋白相似程度，表达模式相近的蛋白聚集在一个分支里。聚类热图中每一行及每一列均对应一个蛋白，颜色越深，提示蛋白间的表达模式有高度的overlap，蛋白间的关系越密切。

根据上述结果， 挑选与表型性状高度关联的共表达模块，然后进一步对所挑选的模块中的蛋白在所有样本中的表达量进行聚类热图分析，并且研究每个模块在所有样本的特征值分布情况（一张图对应一个模块），如下图所示：

[module.heatmap\_barplot.png]

图3-14蛋白表达热图及模块特征值分布图

上部分热图展示了该模块中蛋白在各个样本中的表达模式热图，其中横坐标为不同样本，纵坐标为该模块中的各个蛋白； 红色代表上调，绿色代表下调。

下部分对应展示了该模块中各个样本的模块特征值分布， 绝对值越大说明模块蛋白在该样本中整体表达量上调/下调，值越接近横坐标说明蛋白整体表达量变化较小。

将所挑选出的模块中的蛋白进行 GO/KEGG 富集分析， 以查看该模块中共表达蛋白的功能/通路富集情况，反映该模块的蛋白功能/通路水平的特征，如下所示：

[WGCNA\_GO\_enrichment]

图3-15 GO富集分析

横坐标：GO功能类别，纵坐标：Rich factor，条形图数字标签标识与GO功能相关的蛋白数量；颜色深浅表示p值大小，即某个功能受到影响的显著程度

[WGCNA\_KEGG\_enrichment]

图3-16 KEGG富集分析

横坐标：Rich factor，纵坐标：KEGG通路，气泡的大小反映参与与KEGG通路的蛋白数量；气泡颜色的深浅表示p值大小（-log10），即某条通路受到影响的显著程度

在筛选出与性状（样本）高度相关的模块后，我们还可以观察Gene Significance（GS，蛋白与性状/样本的相关性）与Module Membership（MM，蛋白与模块的相关性）在每个模块中的散点分布情况，从而探究重点模块中蛋白-模块的相关性和蛋白-性状的相关性是否有较好的一致性，并且进一步从重点模块中筛选出相关性较高的关键蛋白，如下图所示：

[moduel\_GS\_vs\_MM]

图3-17散点分布图

横坐标为每个蛋白的表达量与模块特征值的相关系数，纵坐标为每个蛋白的表达量与性状（样本）数据的相关性，cor值为GS和MM值的相关系数，p值则是对相关系数的假设检验值

将所挑选出的模块中包含的所有蛋白构建蛋白共表达网络，以 cytoscape 软件进行可视化展示，反映模块中蛋白的相互关系，从另外一个角度筛选该模块共表达网络中的核心蛋白，进行后续深入研究，以下结果仅为示例：

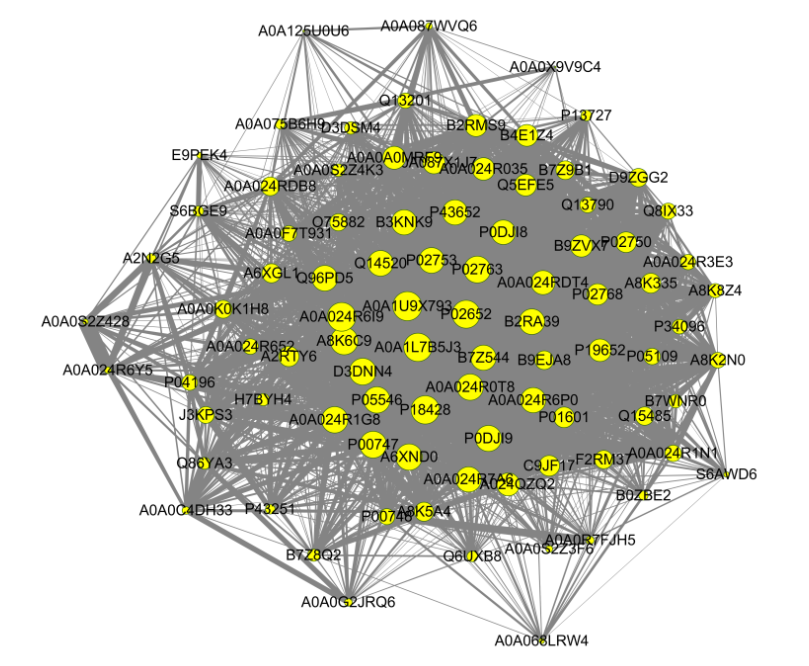


图3-18某模块内蛋白共表达网络

黄色节点代表蛋白， 节点大小与其连接度（Degree）正相关，即与该蛋白相关的蛋白越多， 则其连接度越大，节点尺寸就越大，其在网络中的地位越关键。 线条表示蛋白之间的互作关系，线条的粗细与相关系数的绝对值成正比，即相关程度越高。

WGCNA分析相关结果输出文件：

1. 多维统计学与生信分析\\WGCNA

## **Chapter 4** **Methodology**

### **DIA分析方法**

#### 实验基本流程

本项目实验示例流程见下图4-1，主要包括图谱构建和DIA分析两个部分：蛋白质提取、肽段酶解、色谱分级、液相色谱-串联质谱（LC-MS/MS）数据DDA采集、液相色谱-串联质谱（LC-MS/MS）数据DIA采集、library的构建、蛋白质DIA数据鉴定和定量分析、差异表达蛋白质筛选、差异表达蛋白质聚类分析、功能注释和通路分析等步骤。

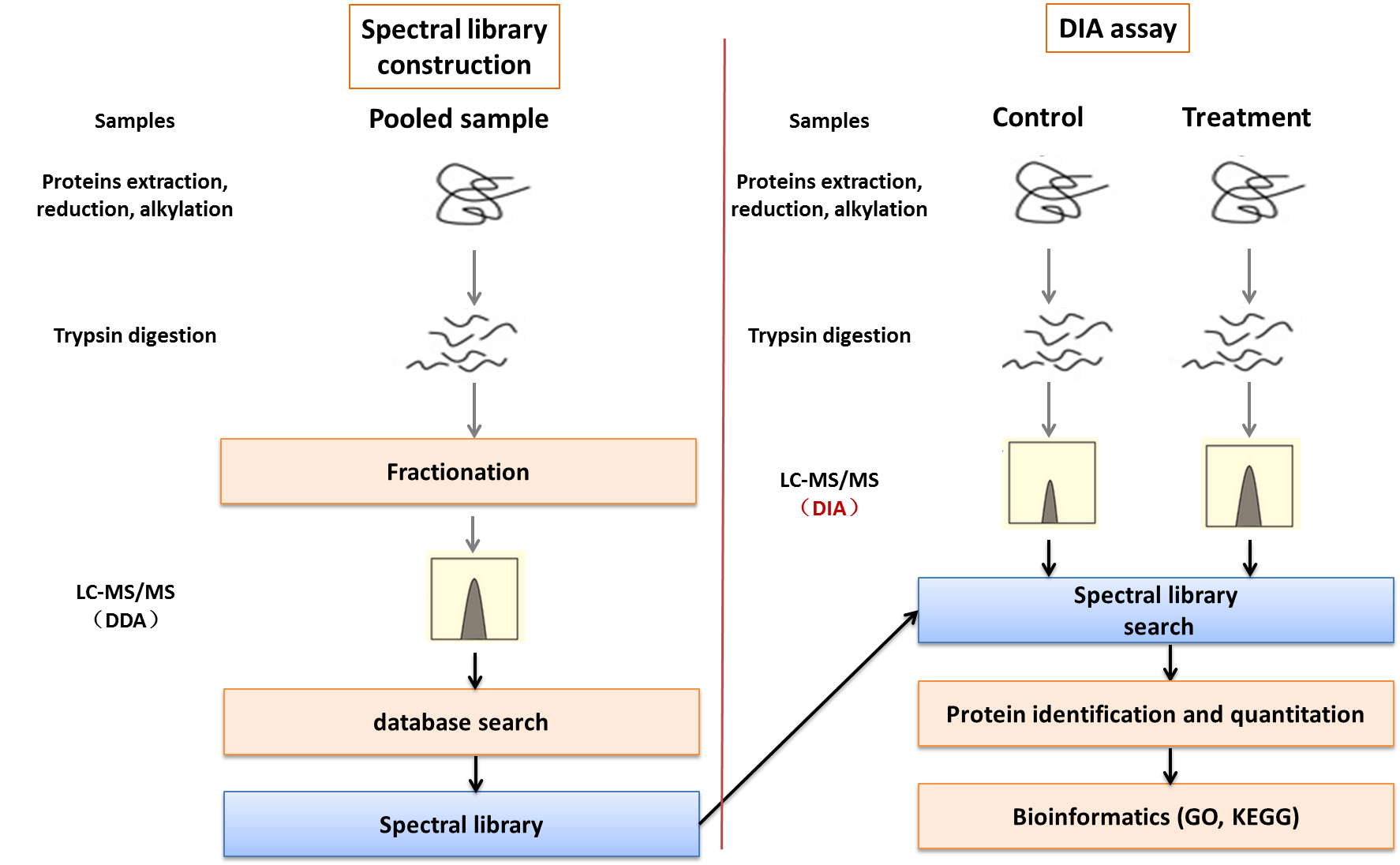


图 4-1 DIA实验和分析流程示例图

#### 实验方法（expr\_sample\_type）

Experiment\_method。

#### 质谱方法

DDA\_method。

DIA\_method。

#### LC-MS/MS数据分析

DDA数据采用Maxquant软件（Maxquant\_1.5.3.17）进行数据库检索，数据库采用human\_uniprot下载数据库，并将iRT肽段序列添加到数据库中（>Biognosys|iRT-Kit|Sequence\_fusionLGGNEQVTRYILAGVENSKGTFIIDPGGVIRGTFIIDPAAVIRGAGSSEPVTGLDAKTPVISGGPYEYRVEATFGVDESNAKTPVITGAPYEYRDGLDAASYYAPVRADVTPADFSEWSKLFLQFGAQGSPFLK）。检索参数设置如下：酶为trypsin，max miss cleavage site为2，固定修饰为Carbamidomethyl(C)，动态修饰设定为Oxidation(M)和Acetyl(Protein N-term)，数据库检索鉴定到的蛋白必须通过设定的过滤参数FDR<1%。将原始raw文件和搜库结果导入Spectronaut软件构建Spectral Library。

DIA数据采用Spectronaut软件（Spectronaut Pulsar X\_12.0.20491.4）进行数据处理，数据库与建库所用数据库相同。软件参数设置如下：retention time prediction type设置为 dynamic iRT, interference on MS2 level correction为enabled，cross run normalization为enabled，所有结果必须通过设定的过滤参数Q Value cutoff 为0.01（相当于FDR<1%）。

### **生物信息学分析方法**

#### 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。

#### 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。

#### GO注释和KEGG注释的富集分析

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KEGG通路注释的富集分析。

#### 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。

#### 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。

#### PCA分析

应用软件SIMCA-P 14.1（Umetrics，Umea，Sweden）进行模式识别，数据经Pareto-scaling预处理后，进行多维统计分析

#### WGCNA分析

使用R包WGCNA（R Version 3.4）撰写脚本构建加权共表达网络。

## **Reference**

1. Reiter L, Rinner O, Picotti P, Hüttenhain R, Beck M, Brusniak MY, Hengartner MO, Aebersold R. mProphet: automated data processing and statistical validation for large-scale SRM experiments. Nat Methods. 2011;8(5):430-5.

2. Rosenberger G, Koh CC, Guo T, Röst HL, Kouvonen P, Collins BC, Heusel M, et al. A repository of assays to quantify 10,000 human proteins by SWATH-MS. Scientific Data. 2014; 1: 140031.

3. Parker SJ, Rost H, Rosenberger G, Collins BC, Malmström L, et al. Identification of a Set of Conserved Eukaryotic Internal Retention Time Standards for Data-independent Acquisition Mass Spectrometry. Mol Cell Proteomics. 2015; 14(10): 2800-2813.

1. Li S, Cao Q, Xiao W, Guo Y, Yang Y, Duan X, Shui W. Optimization of Acquisition and Data-Processing Parameters for Improved Proteomic Quantification by Sequential Window Acquisition of All Theoretical Fragment Ion Mass Spectrometry. J Proteome Res. 2017; 16(2):738-747.
2. Gotz, S., et al., High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Res, 2008. 36(10): p. 3420-35.
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. Kanehisa, M., et al., KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res, 2012. 40(Database issue): p. D109-14.
5. Wisniewski, J.R., et al., Universal sample preparation method for proteome analysis. Nat Methods, 2009. 6(5): p. 359-62.
6. Ross, P.L., et al., Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. Mol Cell Proteomics, 2004. 3(12): p. 1154-69.
7. David M. Herrington., et al., Proteomic Architecture of Human Coronary and Aortic Atherosclerosis. Circulation. 2018;137:2741-2756

