委托协议编号	

技术服务（测试化验加工）委托协议

委托任务名称 1200 例人血浆样本载脂蛋白检测

甲方 北京市心肺血管疾病研究所

单位通讯地址 北京市朝阳区安贞路 2 号

乙方 上海中科新生命生物科技有限公司

单位负责人 蒋春丽

联系人 蒋春丽

联系电话 18301047312

单位通讯地址 上海市闵行区园美路 58 号 1 号楼 15 楼

签订日期: ____年____月____日

签订地点：北京

有效期限：2025 年 5 月 26 日 至 2027 年 5 月 25 日

填 写 说 明

一、 本协议适用于我院科研人员在项目研究过程中支付给外单位的检验、测试、化验及加工等费用时需要签署的协议。

二、 合同封面的委托任务名称指本合同的测试加工等具体内容，应用简明规范的专业术语明确概括所要完成的服务内容。

三、 本合同的甲方和乙方名称，须按单位公章的详细名称填写，若涉及外文名称，首次出现时应写明全称及简称。

四、 本协议书未尽事项，可由当事人附页另行约定，并可作为本协议的组成部分。如协议研究内容涉及国家秘密或重大商业秘密的，双方应另行签署保密义务。

五、 使用本协议书时约定无须填写的条款，应在该条款处注明“无”等字样。

六、 协议书要求 A4 纸打印，一式 4 份，左侧装订，正文内容所用字型应不小于 5 号字，协议正本中所涉及与本协议约定事项有关的技术资料及其指定附件备齐后应合装成册，其规格大小应与协议书一致。

七、 乙方需提供测试化验加工的原始数据，甲方务必保留原始数据 10 年以上以备审计抽查。

八、 协议需法人或委托代理人签署意见后加盖医院公章方可生效。

依据《中华人民共和国民法典》及本协议书相关的科研项目、经费管理办法规定，为完成甲方承担的研究任务，经双方协商一致，各方在真实、充分地表达各自意愿的基础上，就本协议书中所描述的委托内容、经费支付、保密内容、知识产权等问题达成如下协议，签订本合同并由签约双方共同恪守。

第一条 委托工作的主要内容、加工方式和要求

1、测试加工内容

甲方委托乙方就 1200 例人血浆样本进行载脂蛋白检测

1.1 本项目实验分析流程包括以下步骤：

1.1.1 人血浆 DIA 相对定量检测：本项目分析流程主要采用 thermo top14 试剂盒的方法去除高丰度蛋白，再对血浆里中低丰度蛋白检测。DIA 质谱实验分析流程主要包括血浆样本通过 thermo top14 试剂盒去除高丰度蛋白，酶解后得到肽段进行液相色谱-串联质谱（LC-MS/MS）DIA 数据采集、单个样本 8min 检测时长、数据库检索等步骤对血浆样本分析。

1.1.2 人血浆 DIA 相对定量检测：采用 Thermo Scientific Vanquish Neo UHPLC+Orbitrap Astral 8min(最新一代高通量质谱，实现血液蛋白质组超深度鉴定)的全息扫描蛋白质组学科研服务。

1.1.3 组学检测：本项目分析流程主要包括 DDA 建库与 DIA 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、液相色谱-串联质谱（LC-MS/MS）DDA 数据采集、数据库检索等步骤；正式实验阶段主要包括 DIA 分析，质控分析，定性定量结果分析及生物信息学分析。下机数据采用 Spectronaut 进行定性定量分析，提供蛋白定性定量列表，肽段定性定量列表，GO，KEGG ，PPI 等生物信息学分析结果。

2、测试加工方式和要求

2.1 技术服务的方式：甲方提供样本，乙方完成全部检测工作。

2.2 技术服务的要求：客观检测，符合数据的质量要求。

第二条 考核指标及验收方式

双方确定以下列标准和方式对乙方的技术服务工作成果进行验收：

1．乙方完成技术服务工作的形式：按照合同要求客观检测。

2．技术服务工作成果的验收标准：达到合同中技术服务质量要求。

3．技术服务工作成果的验收方法：成果报告以纸质版和电子版两种形式发送给甲方。经甲方签字确认后，验收报告生效。

4．验收地点：北京市心肺血管疾病研究所

第三条　测试化验加工细目：

序号	测试化验加工的内容	测试结果的呈现方式	计量单位	单价 （万元/单位）	数量	金额（万元）
1	1200 例人血浆样本 DIA 蛋白质组学相对定量检测	电子版和纸质版的成果报告	例	0.0583	1200	69.96
	合计					69.96

第四条　经费支付方式：

1. 委托应支付费用共计 __69.96__万元，由甲方提供。

2. 支付方式一次：（一次或分期）支付乙方　（按以下第　③种方式）：

①一次总付：_____万元。乙方在甲方付款前，即需提供测试服务。

②分期支付：

第一次支付_____万元，甲方在合同签订后____日内支付。

第二次支付_____万元，甲方在乙方全部测试技术服务完成并通过验收后___日内支付。

③其它方式：

合同签订完成 30 日内，乙方向甲方提供 2 份履约保函，其中：合同总价 5%（叁万肆仟玖佰捌拾元整人民币）的履约保函，保函期限为一年，全部服务完成经甲方验收合格后退还 另外合同总价 5%（叁万肆仟玖佰捌拾元整人民币）的质量保函，保函期限为两年，待验收签字确认合格之日起免费售后服务执行 12 个月后（若售后服务无问题）退还。甲方收到乙方开户银行履约保函后甲方向财政办理合同支付手续。甲方支付费用 7 日前，乙方应将对应金额的法定发票提供甲方审核，待审核通过后甲方按照合同约定向乙方支付费用，如发票审核不合格，或者乙方未按规定提供保函的，甲方有权延期支付费用。

第五条　知识产权归属

1. 双方在申请本课题之前各自所获得的知识产权及相应权益均归各自所有，不因共同申请本课题而改变。

2. 本协议所产生的所有成果的知识产权全部归属于甲方，乙方不得利用测试结果单独申报任何形式的成果。

3. 在课题执行过程中各自向对方提供的相关信息,不构成向对方授予任何关于知识产权的许可行为。

4. 本合作协议不在各方之间建立任何商业上的代理、合作关系。

第六条　保密条款

1. 乙方保证不向甲方以外的人员提供或披露本合同的委托内容及未公开的信息和资料。包括但不限于本协议的委托内容及结果。
2. 双方保证采取一切合理和必要措施和方式对委托中知悉的对方商业秘密进行保密。

第七条　承诺

1. 如委托的任务涉及人类遗传资源采集、收集、买卖、出口、出境等，乙方承诺遵照《人类遗传资源管理暂行办法》相关规定执行。
2. 如委托任务涉及动物实验，乙方承诺自觉遵守《实验动物管理条例》，严格选用符合要求的合格动物进行实验，保障动物福利。
3. 如委托任务的研究对象涉及人类受试者，乙方承诺在签署协议前已经将委托任务的实施方案呈交单位伦理委员会讨论，并获得了伦理委员会批准。甲方在完成委托任务的过程中，自觉遵守国内外相关的医学伦理准则，保障保护受试者的安全和权益。
4. 在乙方从事委托事项中发生的不可归责于甲方的人身、财产损害，由乙方自行承担。
5. 乙方保证与甲方无直接经济利益关系，并保证委托关系及事项真实有效。

第八条　不可抗力

1. 本协议所指不可抗力是指不能预见、不能避免并不能克服的客观情况，包括但不限于地震、火灾、水灾、战争、政府行为等。

2. 乙方因不可抗力不能履行协议的，应当在不可抗力事件发生之日起七日内将不可抗力事由以书面方式通知甲方，并应当在合理期限内提供证明。

3. 因不可抗力不能履行本协议的，根据不可抗力的影响，部分或全部免除责任。乙方延迟履行后发生不可抗力的，不能免除责任。

第九条　违约责任

1. 如无正当理由，甲方未能按期拨付工作经费，且经乙方催促仍不能拨付或不能给出合理解释的，乙方有权暂停履行受托任务。如甲方违约行为给乙方造成损失的，甲方还应承担相应赔偿责任。

2. 如乙方在完成委托工作时出现弄虚作假情况、不履行本协议或履行义务不符合要求的，甲方有权追回全部已拨经费。如乙方违约行为造成甲方损失的，甲方有权要求赔偿并追究乙方相关责任人员的法律责任。

3. 非因甲方违约或非因不可抗力，乙方不能完成受托任务或乙方逾期不能提交全部产出成果的，甲方有权解除本委托。委托解除后，乙方应返还甲方已经拨付的项目经费。如乙方的违约行为给甲方造成损失的，乙方还应承担相应的赔偿责任。

4. 乙方在合作期间及合作结束后，未经甲方书面同意，不得在任何形式的宣传材料、广告、媒体发布或公开声明中，使用甲方的名称（包括全称和简称等）、商标、标志、域名、产品或服务进行宣传或暗示其与甲方存在任何形式的合作关系，包括但不限于技术合作、业务往来、信用担保等。如违反本条内容，甲方有权要求乙方停止此侵权行为，并要求乙方赔偿甲方由此遭受的损失（包括直接损失及间接损失）。

第十条　协议的变更、终止及解除

1. 本协议的变更应由双方协商一致后达成变更协议，并作为本协议的附件。

2. 本协议可由双方协商一致予以终止。

第十一条　争议解决：如在履行本协议的过程中发生争执，双方当事人应友好协商解决，如协商不成，任何一方可向甲方签署地（甲方所在地）有管辖权的人民法院提起诉讼。

第十二条　其他约定事项（如无其他事项，请填“无”）

无

第十三条　本协议一式五份，甲方四份，乙方一份，具有同等法律效力。

与本协议约定事项有关的技术资料附件清单：见附件

第十四条 签字盖章页

委托方（甲方）	单位名称	北京市心肺血管疾病研究所（盖章）		
	单位负责人	蔡军 (签字)		
	经办人	(签字)	经办人 联系电话	
乙方	单位名称	上海中科新生命生物科技有限公司（盖章）		
	乙方的单位负责人	陈薇 印 (签字)		
	经办人	(签字)	经办人 联系电话	18301047312
	开户名称	上海中科新生命生物科技有限公司		

	开户银行	农行漕河泾开发区支行
	银行账号	03390800040007818

附件一 投标分项报价表

序号	服务内容	单价（元）	数量	合价（元）
1	载脂蛋白检测	583.00	1200	699600.00
总价				699600.00

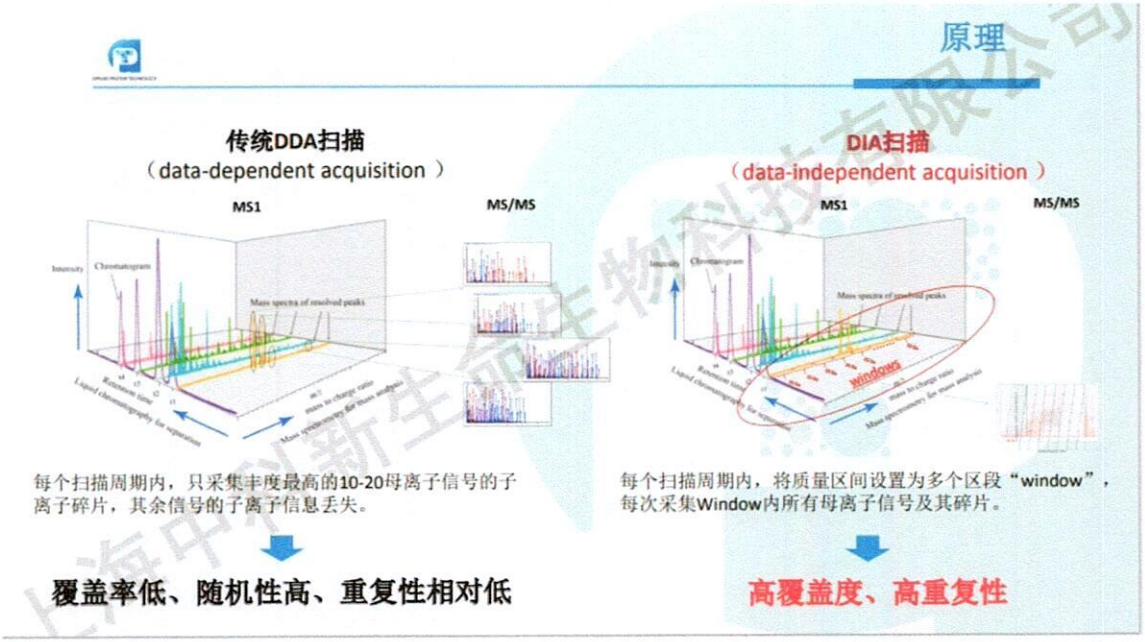
附件二　技术方案

1　DIA 蛋白组学技术原理

1.1　DIA蛋白组学技术原理

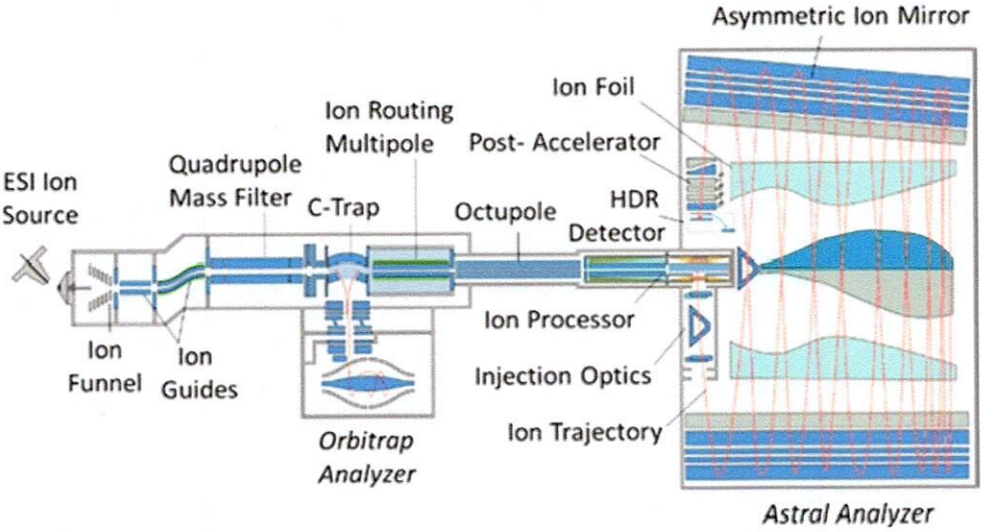
蛋白组学是通过液质联用技术对蛋白质进行质谱定量分析的。该技术通过大规模分析肽段所产生的质谱数据，比较不同样品中相应肽段的定量信息，从而对肽段对应的蛋白质进行相对定量。DIA（data-independentacquisition）技术是近年来发展起来的一种新的质谱技术。与传统的 DDA（data-dependent acquisition）质谱技术相比，DIA 采用了不同的数据扫描模式：将质谱整个全扫描范围分为若干个窗口，然后对每个窗口中的所有离子进行检测、碎裂，从而无遗漏、无差异地获得样本中所有离子的信息。

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。



1.2 新一代高通量质谱Orbitrap Astral仪原理

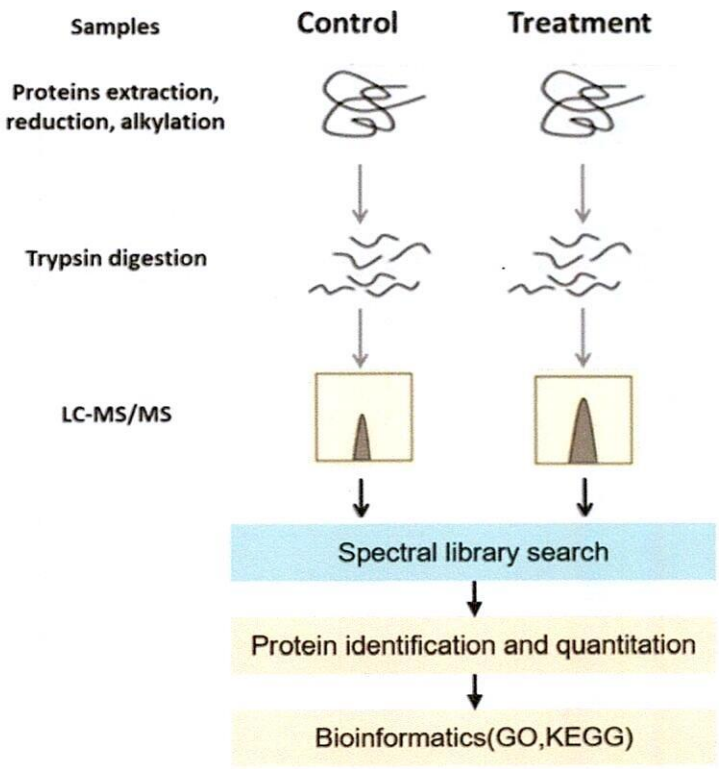
奥斯卡DIA 蛋白质组学是基于赛默飞Orbitrap™ Astral™高分辨质谱仪（简称Astral质谱仪）的蛋白质组学检测。Astral质谱仪集合了四级杆质量分析器、Orbitrap质量分析器和Asymmetric Track Lossless非对称轨道无损质量分析器，显著扩大了研究的范围和视角。前端（离子源至四极杆）最大限度提高仪器的灵敏度和耐用性。Orbitrap 质量分析器能够以高分辨率采集全景全扫描数据。Astral 质量分析器能够快速（高达 200 Hz）、灵敏地采集高动态范围 HRAM， 与Orbitrap 质量分析器的采集完全同步。因此， Astral质谱仪在多种数据采集策略下都具有出色表现。Astral质谱仪加持全扫描 DIA 技术，极大提升了质谱的鉴定能力。



2 DIA 全息扫描蛋白质组学整体方案

2.1 分析流程

本项目分析流程主要采用thermo top14试剂盒的方法去除高丰度蛋白，再对血浆里中低丰度蛋白检测。DIA质谱实验分析流程主要包括血浆样本通过thermo top14试剂盒去除高丰度蛋白，酶解后得到肽段进行液相色谱-串联质谱（LC-MS/MS）DIA数据采集、单个样本8min检测时长、数据库检索等步骤对血浆样本分析。检测流程示意图如下：



DIA定量蛋白质组学实验流程图

2.2 样品处理和预实验评估及分析方法

2.2.1 实验材料准备

2.2.1.1 样本采集的一般性原则

【一致性原则】：每例样本取样的部位、方式、预处理方法需要保持一致

【快速原则】：请务必提前设计准备好实验和材料，快速取样和分装。

【分装原则】：为避免样本反复冻融，建议样本采集后立即进行分装。

【联合分析原则】：尽可能保证样本同一批次，组数及生物学重复一致，并对不同组学样本进行分装。

【低温原则】：在采集过程中，请冰上操作，分离好的样本液氮速冻，取出保存于-80℃冰箱中。

2.2.1.2 样本的包装和运输指南

[1] 样品尽可能采用1.5ml或者2ml离心管（进口离心管）保存，运输时采用封口膜密封离心管（如管内为有机溶剂，务必采用螺旋口的冻存管并密封）。离心管上标记清楚样

品名称后，按顺序整齐排列在冻存盒中。将冻存盒中样品存放的顺序信息对应填写《APT科研项目送样表》（电子版）。

[2] 不方便存储在离心管中的体积较大的组织样品，推荐采用锡箔纸等材料仔细包装，标记清楚样品名称，按照组别整理整齐，放置在密封袋中。

[3] 推荐采用双层泡沫盒密封包装，盒中加入足量的干冰。

2.2.2 样品处理和预实验评估

客户自行收集样品，低温下运输至技术中心。血浆样本通过thermo top14试剂盒去除高丰度蛋白。在本实验中，从每个样品中取出等量的部分，合并为一个样品作为质量控制样品。

2.2.3 样本消化

向每个样品中加入二硫苏糖醇（DTT）以还原二硫键，于 37° C 反应 1.5 小时。接着加入碘乙酰胺（IAA）来封闭被还原的半胱氨酸残基，在室温下避光反应30分钟。随后向样品中加入胰蛋白酶（胰蛋白酶与蛋白质的质量比为 1:50），并在 37° C 孵育 15 - 18 小时（过夜）。将消化后的肽段在MCX脱盐柱（omicsolution，OS-MCX-1ML）上进行脱盐处理，通过真空离心浓缩，再用 20μL 含 0.1%（v/v）甲酸的水溶液复溶。通过 280nm 处的紫外光谱密度来估算肽段含量。对于数据非依赖性采集（DIA）实验，向样品中加入 iRT（校准保留时间）校准肽段。

2.2.4 DIA分析方法

每个样品中的肽段通过与 Vanquish Neo 系统液相色谱（赛默飞世尔科技）相连的OrbitrapAstral质谱仪（赛默飞世尔科技），以数据非依赖性采集（DIA）模式进行分析。检测模式：正离子，母离子扫描范围为 380-980m/z，一级质谱分辨率为 240000 at 200 m/z，Normalized AGC Target 为 500% Maximum IT 为 5ms。MS2 采用 DIA 数据采集模式，设置 299 个扫描窗口， Isolation Window 为 2 Th，HCD Collision Energy 为 25% ，Normalized AGC Target 为 500%，Maximum IT 为 3ms。

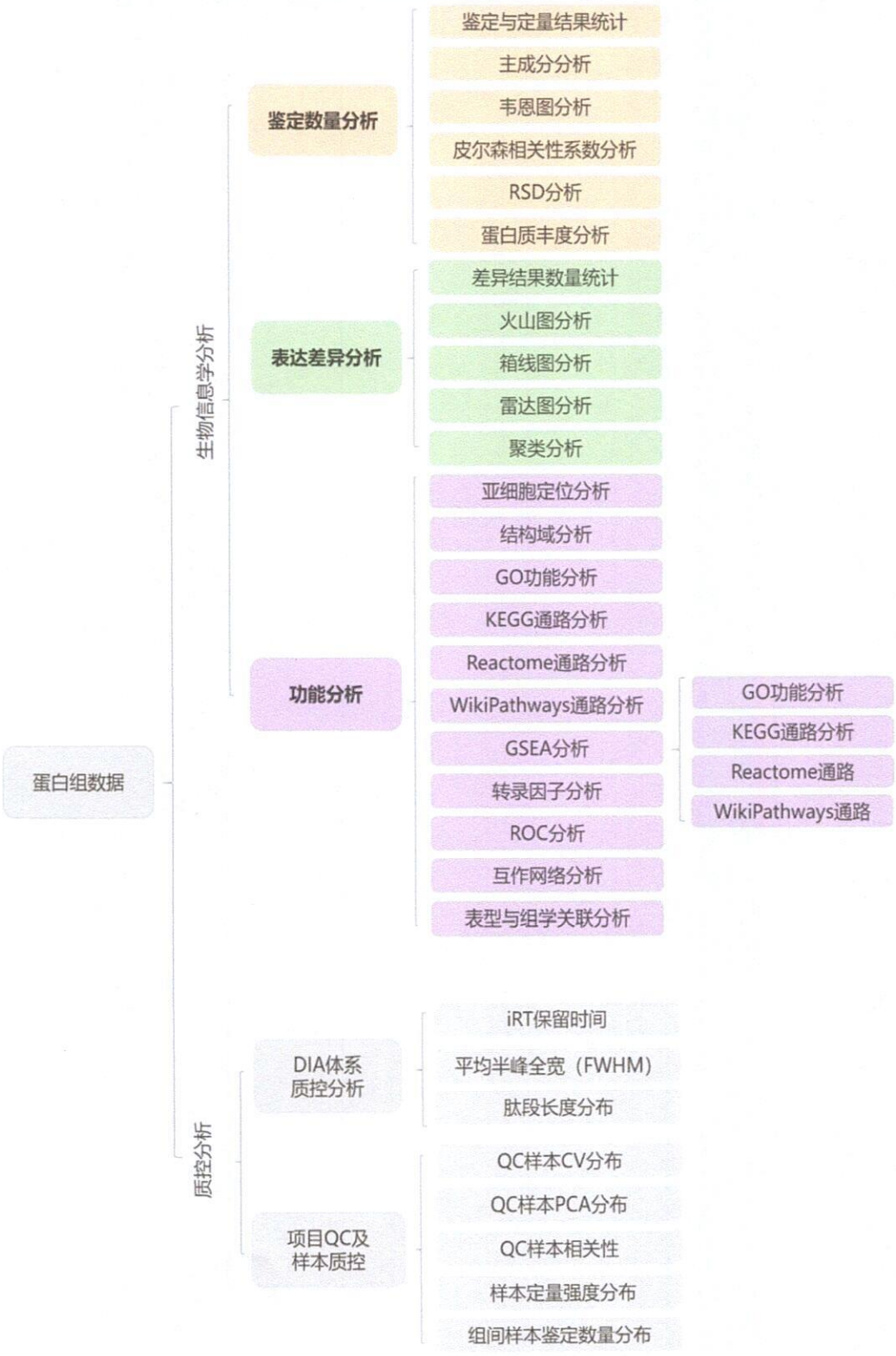
2.2.5 LC-MS/MS数据分析及交付

使用DIA-NN1.8.1软件对DIA数据进行分析。主要软件参数设置如下：酶为胰蛋白酶，最大漏切数为 1，固定修饰为半胱氨酸氨基甲基化（C），动态修饰为甲硫氨酸氧化

（M）和蛋白质 N 端乙酰化。所有报告的数据均基于蛋白质鉴定的置信度达到 99%，此置信度由错误发现率（FDR）≤1% 确定。

定性分析：血浆符合定性要求的鉴定数量不低于3000，所检出的蛋白包含载脂蛋白，包括APOA, APOB，APOC, APOD, APOE, APOF, APOH, APOM相关各亚型，原则上载脂蛋白检测数量不低于15种。

实验下机数据搜库后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，除此之外还提供与其他组学的联合分析，例如蛋白组与临床脂蛋白亚组分参数（甲方检测的临床指标，由甲方提供）、代谢组联合分析；蛋白组与转录组联合分析等。基本分析见下图。


蛋白组数据
生物信息学分析
鉴定数量分析
鉴定与定量结果统计
主成分分析
韦恩图分析
皮尔森相关性系数分析
RSD分析
蛋白质丰度分析
表达差异分析
差异结果数量统计
火山图分析
箱线图分析
雷达图分析
聚类分析
功能分析
亚细胞定位分析
结构域分析
GO功能分析
KEGG通路分析
Reactome通路分析
WikiPathways通路分析
GSEA分析
转录因子分析
ROC分析
互作网络分析
表型与组学关联分析
GO功能分析
KEGG通路分析
Reactome通路
WikiPathways通路
质控分析
DIA体系质控分析
iRT保留时间
平均半峰全宽（FWHM）
肽段长度分布
项目QC及样本质控
QC样本CV分布
QC样本PCA分布
QC样本相关性
样本定量强度分布
组间样本鉴定数量分布

3.生物信息学分析 Bioinformatics analysis

3.1 鉴定数量分析

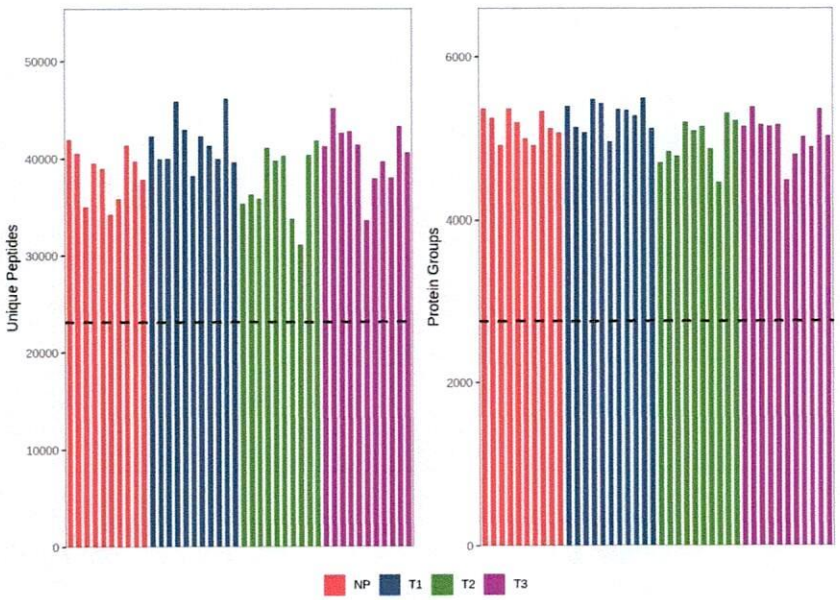
3.1.1 鉴定与定量结果统计

本项目每个样本鉴定的肽段数、鉴定的蛋白数结果统计，如下表与下图。(此报告模板中仅列举部分样本鉴定数目作为示例)

DIA 鉴定与定量结果统计表

Sample	Peptides	Proteins
.....	.....	.....
.....	.....	.....
.....	.....	.....

为整体观测不同组别样本鉴定到的蛋白及肽段数目，将每个样本的鉴定结果以柱状图展示如下



DIA 鉴定结果统计柱状图

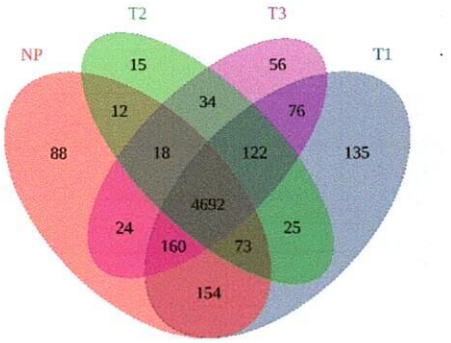
说明： Peptides：鉴定到的肽段总数目； Protein groups：鉴定到的蛋白质总数。不同颜色代表不同组别。图中虚线代表样本的蛋白/肽段鉴定数量最高的一半。

输出文件：

1) 3-1-1 鉴定与定量结果统计

3.1.2 组间样本鉴定重复性

为考察不同组别之间鉴定数量的重叠情况，以 Venn 图的形式将各组鉴定到的蛋白进行展示，结果如下图所示：



组间样本Venn图（超过五组出具花瓣图）

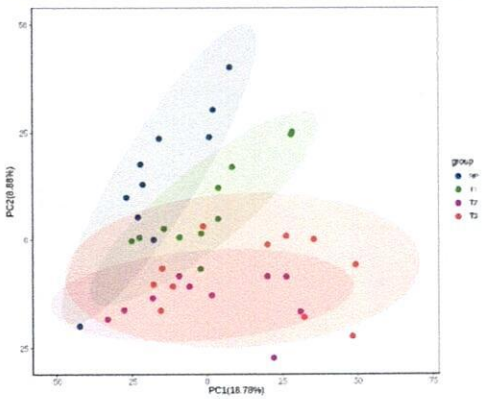
说明：每个颜色代表一个组别。

输出文件：

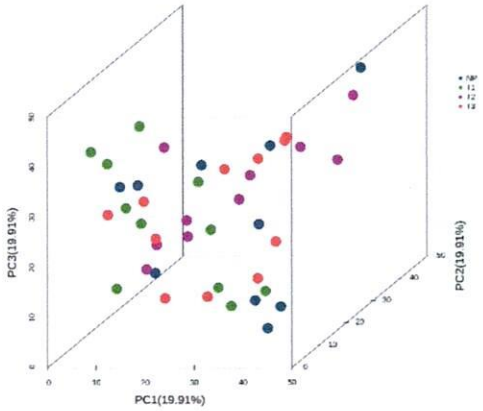
1) 3-1-2韦恩图分析

3.1.3 PCA 主成分分析

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，则距离越远，因此能从总体上反映样本组间和组内的变异度。本项目对所有样本进行2D和3D PCA分析，结果如下图展示：



所有样本2D PCA分布图



所有样本3D PCA分布图

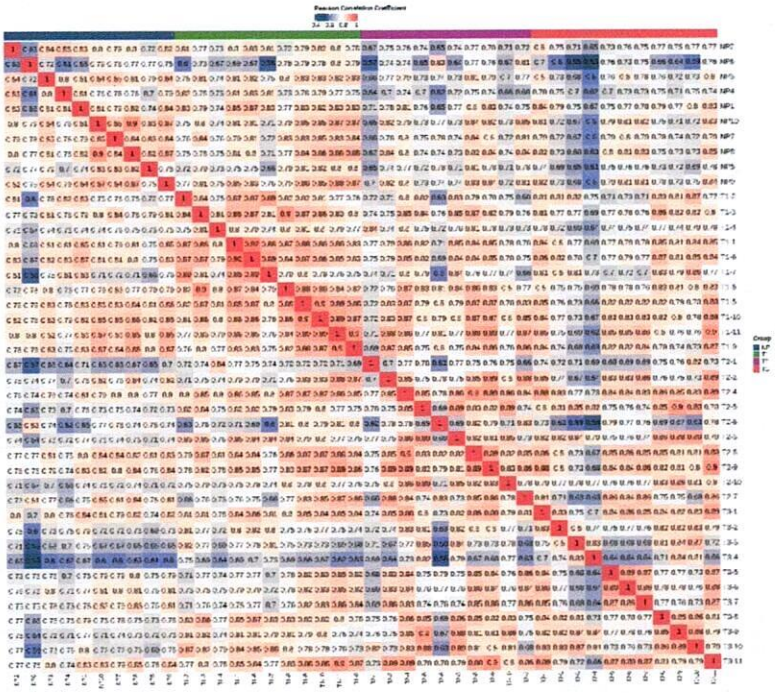
说明：图中PC1代表主成分1，PC2代表主成分2，PC3代表主成分3，每个点代表一个样本，不同颜色分别代表不同的组别。

输出文件：

1) 3-1-3 PCA分析

3.1.4 皮尔森相关性系数（Pearson's Correlation Coefficient，PCC）分析

所有样本两两之间计算皮尔森相关系数而绘制的热图。此系数是度量两组数据线性相关程度的值：当皮尔森系数越接近-1为负相关，越接近1为正相关，越接近0为不相关。结果如图所示。



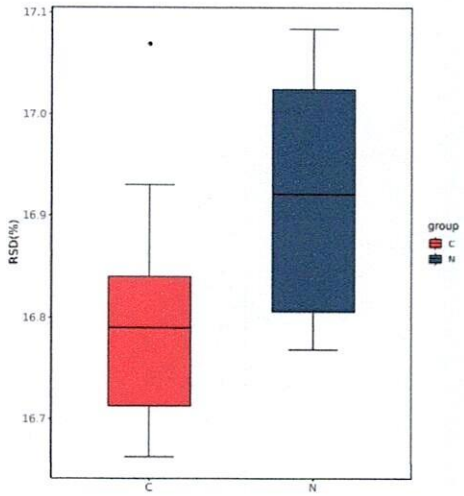
样本组间鉴定到的蛋白质PCC分析图

输出文件：

1）3-1-4 PCC分析

3.1.5 RSD 分析

样本间蛋白定量值的相对标准差（RSD）越小，表明蛋白质组学的定量重复性越好。



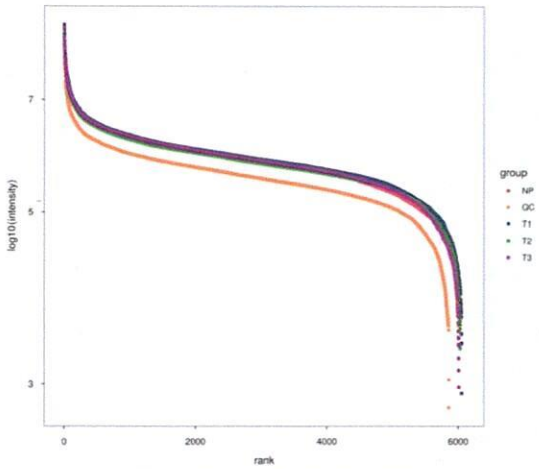
样本组间鉴定到的蛋白质RSD分析图

输出文件：

1) 3-1-5 BoxPlot分析

3.1.6 蛋白质丰度分析

对所有组的样本鉴定到的蛋白质丰度做散点图分析，如下所示。



蛋白质丰度分布散点图

说明：横坐标为蛋白表达量的排名，纵坐标为蛋白的强度值（log10转化）

输出文件：

1) 3-1-6 Scatter分析

3.2 表达差异分析

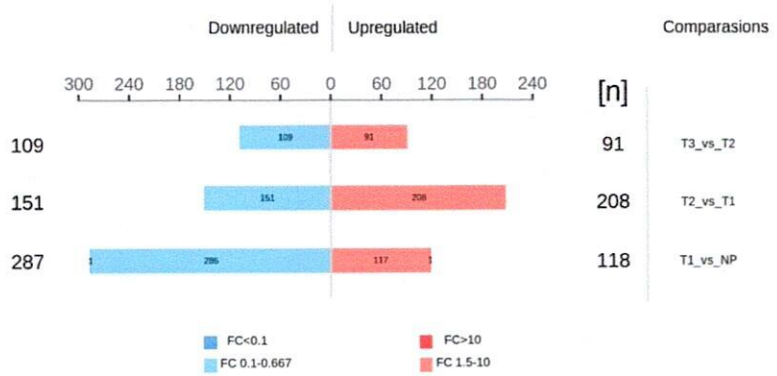
3.2.1 差异结果数量统计

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)> 1.5 倍（上调大于 1.5 倍或下调小于 0.67 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。

蛋白质定量差异结果统计表

Comparisons	Significantly changing in abundance			Consistent presence/absence expression profile	
	Upregulated	Downregulated	All	Upregulated	Downregulated
T3_vs_T2	109	91	200		
T2_vs_ T1	151	208	359		
T1_vs_NP	287	117	404		



蛋白质定量差异结果柱状图

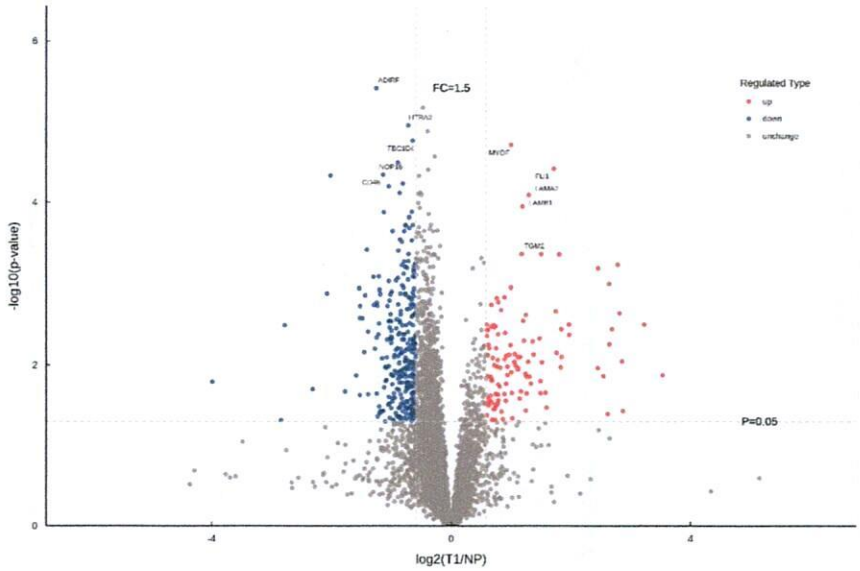
说明：Comparisons：差异比较组；Significantly changing in abundance：符合筛选倍数和 p value 的差异表达蛋白质；Consistent presence/absence expression profile：一组样品中半数及半数以上不为空值，另一组所有数据均为空值的差异蛋白质。Upregulated：上调差异表达蛋白质；Downregulated：下调表达蛋白质；All：所有差异表达蛋白质。

输出文件：

1) 3-2-1 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< 0.67 且p<0.05），显著上调的蛋白质以红色标注（FC>1.5 且p<0.05），无差异的蛋白质为灰色，并对上下调蛋白差异最显著的top5进行标注，结果如下图所示。



groupvs组火山图

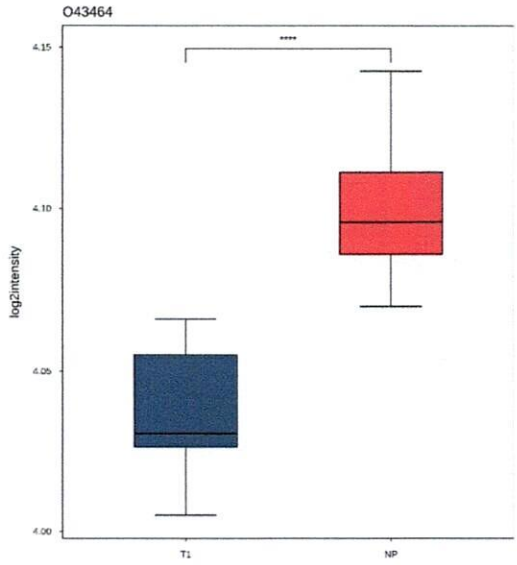
说明：横坐标为差异倍数（以2为底的对数变换），纵坐标为差异的显著性P-value（以10为底的对数变换）。图中红色点为上调的显著性差异表达蛋白质，蓝色点为下调的显著性差异表达蛋白质，灰点为无差异变化的蛋白质。标注ID的点为差异最显著的top5上、下调蛋白。

输出文件：

1) 3-2-2火山图

3.2.3 差异蛋白表达箱线图

为了更直观的展示差异蛋白在不同组之间的表达差异，利用箱线图的形式对两组间差异表达的蛋白进行展示。报告中仅展示了一个比较组一个差异蛋白的箱线图，其他比较组差异蛋白的在附件中展示箱线图。



groupvs组差异蛋白箱线图

说明：横坐标为组别，纵坐标为表达量（以2为底的对数变换），图中红色和蓝色分别代表该差异蛋白在不同样本中的表达量。 *表示差异显著性程度，***表示p<0.001,**表示0.001<p<0.01，*表示0.01<p<0.05。具体结果可查看附件。

输出文件：

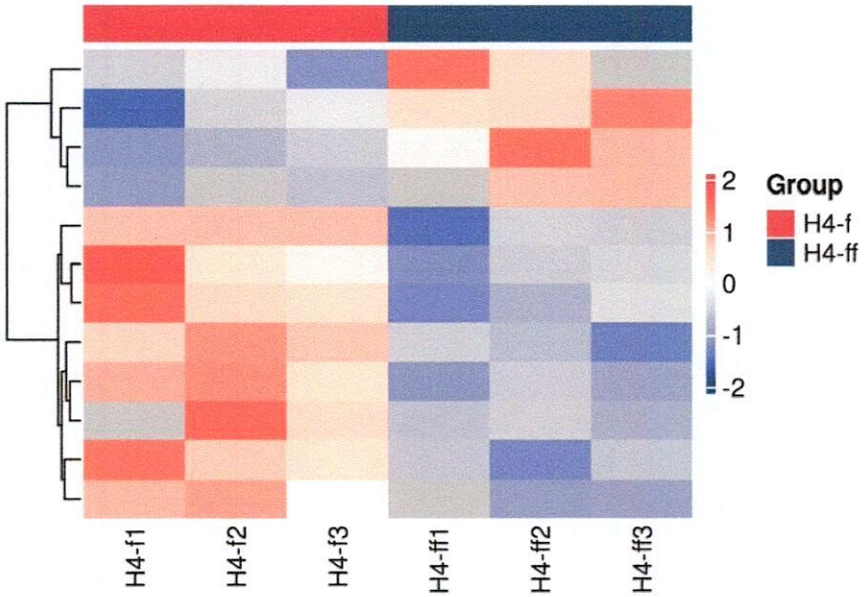
1) 3-2-3 差异蛋白箱线图

3.2.4 聚类分析

3.2.4.1 差异蛋白表达层次聚类分析

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>1.5 倍且 Pvalue<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。



groupvs组差异表达蛋白质聚类分析图

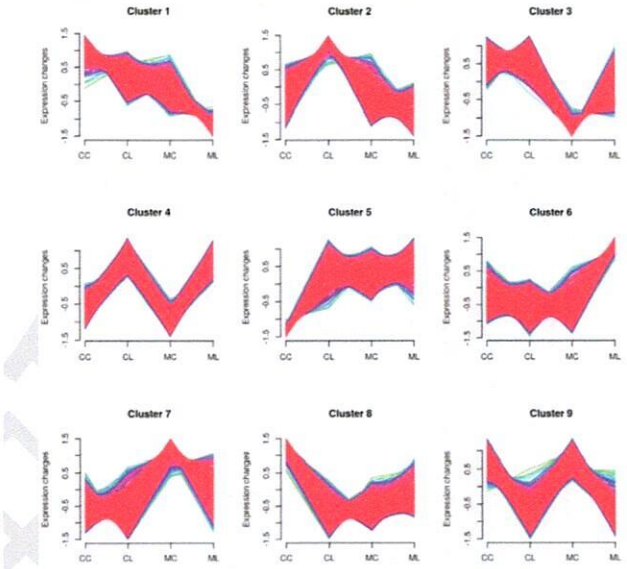
说明：层次聚类结果以树型热图表示，横轴表示样本，用不同颜色表示样本分组信息，纵轴代表差异表达的蛋白质（即纵坐标为显著性差异表达的蛋白质），显著性差异的蛋白质在不同样品中的表达量用Z-score方法进行标准化后以不用颜色在热图中展现，其中红色代表显著性上调的蛋白质，蓝色代表显著性下调的蛋白质，灰色部分代表无蛋白质定量信息。其他图示见附件。

输出文件：

1) 3-2-4差异蛋白聚类分析

3.2.4.2 多组蛋白表达模式聚类

为了分析多组样本的所有蛋白整体表达模式，说明蛋白质表达量变化趋势。采用Mfuzz软件的fuzzy c-means（FCM）算法进行分析，根据所有蛋白的表达趋势分为不同的表达模块。本项目的表达模式及趋势分类如下图展示。（本分析仅适用于3组及以上，2组及以下无此分析，如有具有时间梯度或者不同疾病进程阶段，需作图之前说明标注顺序）。



多组蛋白表达趋势聚类图

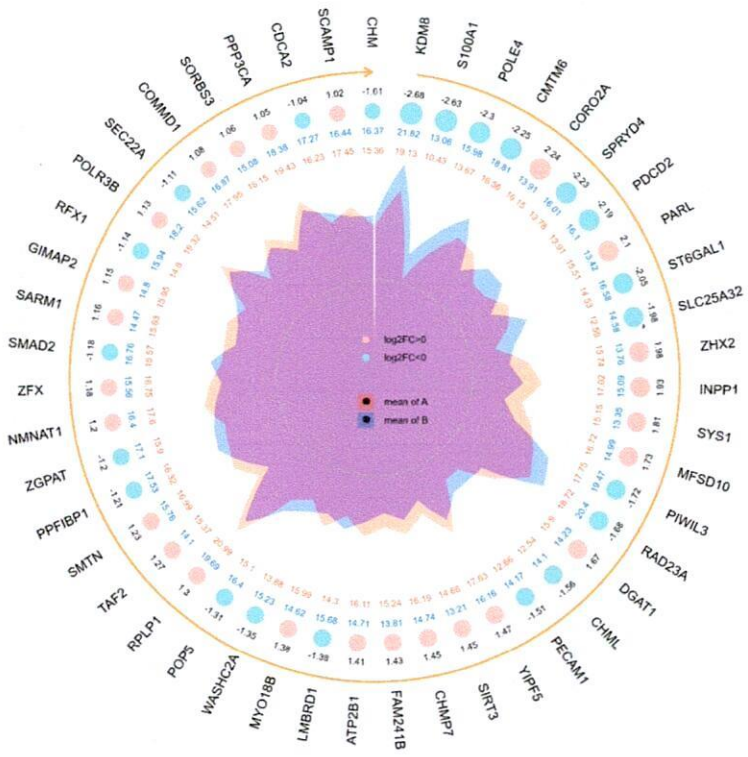
说明：横坐标代表不同组别，纵坐标表示均一化之后的表达量变化。每一个Cluster的线条指蛋白中表达趋势的一类蛋白。

输出文件：

1) 3-2-4模糊C-均值聚类分析

3.2.5 雷达图（Radar Chart）

用于展示多个差异蛋白在比较组中的相对表达水平。第一圈表示多个差异蛋白（人和小鼠展示的是差异蛋白所对应的基因）；第二圈的橙色箭头当样本有重复时，表示差异蛋白对应的P value或者CV值，由小到大排序，若样本无重复，则表示差异蛋白差异倍数Log2转换后的绝对值由大到小排序；第三圈表示Log2转换的比较组的差异倍数，粉红色表示上调，浅蓝色表示下调，点越大表示差异倍数越大；第四圈表示两组的平均定量值。详见：



groupvs组的差异表达雷达图

输出文件：

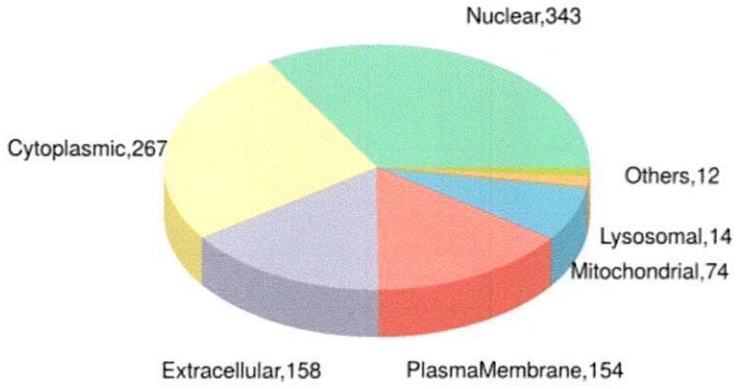
1) 3-2-5 差异蛋白雷达图

3.3 功能分析

3.3.1 亚细胞定位分析

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功能。

采用亚细胞结构预测软件CELLO（http://cello.life.nctu.edu.tw）[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细胞器中的蛋白质数目与分布比例，如下图。

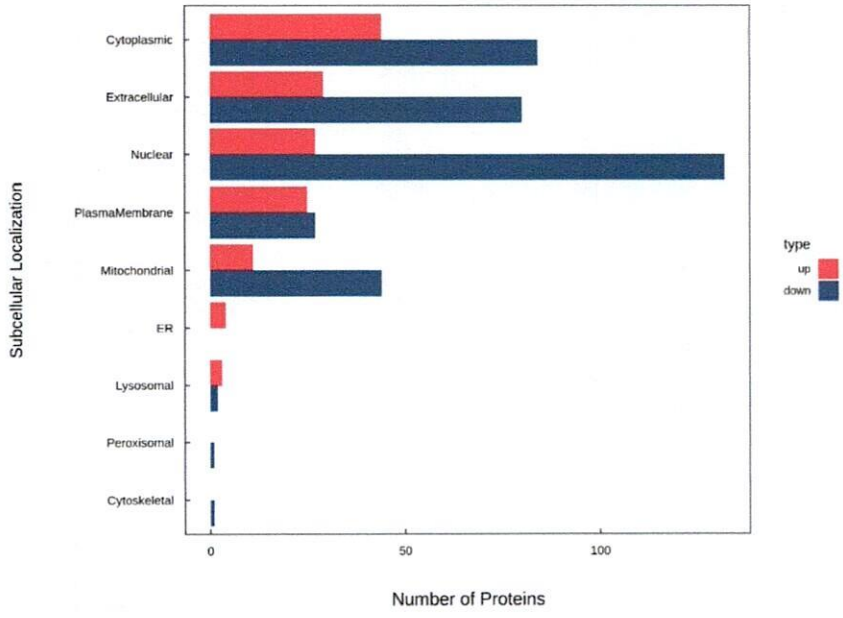


groupvs 组差异表达蛋白质亚细胞定位分布饼图

输出文件：

1）3-3-1亚细胞定位分析

对每个比较组的差异蛋白质进行亚细胞定位统计分析，分别统计上下调蛋白质数目，并以柱状图的形式展示；下图仅展示了一个比较组的结果，其他比较组结果见附件。



亚细胞定位结果上下调比较柱状图

说明：纵坐标代表亚细胞定位，横坐标代表该亚细胞注释到的差异蛋白质数量，红蓝色代表上下调的蛋白质。

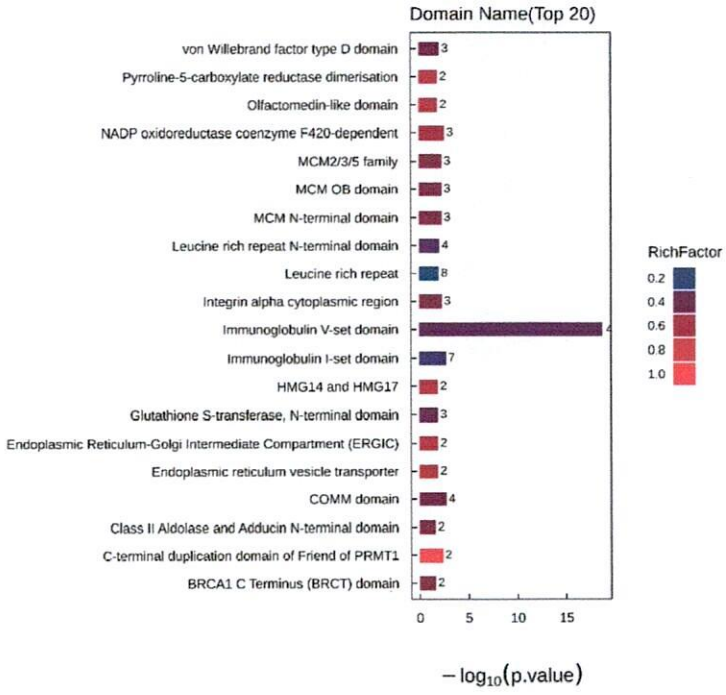
输出文件：

1）3-3-1亚细胞定位分析

3.3.2 结构域分析

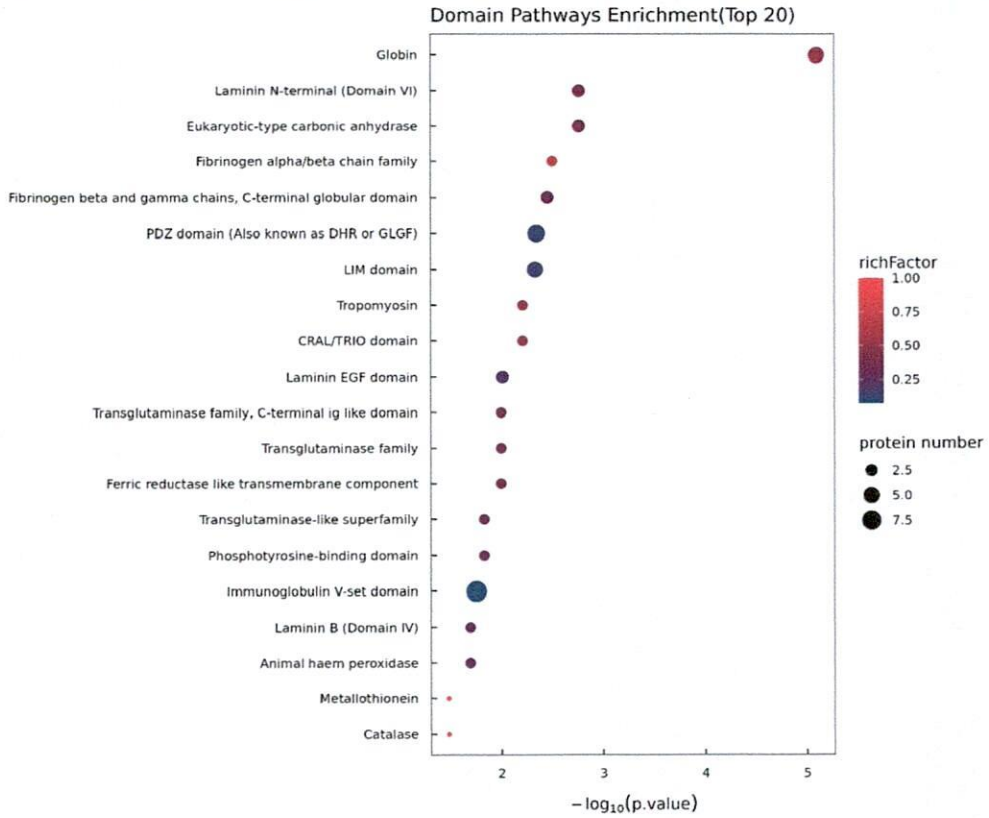
蛋白质结构域（domain)是在较大的蛋白质分子中，由于多肽链上相邻的超二级结构紧密联系，形成两个或多个在空间上可以明显区别的局部区域。一般每个结构域由几十至几百个氨基酸残基组成，各有独特的空间结构，并承担不同的生物学功能。一般来说，蛋白与蛋白（或其他小分子）的相互作用常以结构域为单位，结构域内氨基酸或修饰发生改变，可能引起蛋白关键功能的改变，故后续氨基酸突变功能实验可以以此为参考。因此，结构域预测对于研究蛋白关键功能区域及其发挥的潜在生物学作用具有重要意义。

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中的蛋白数目(前20)，如下图所示。



groupvs组差异表达蛋白质结构域分析柱状图

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。



groupvs组结构域富集分析气泡图

说明：图中横坐标为某结构域分类的富集显著性，即基于Fisher精确检验（Fisher’s Exact Test）计算P值（取-log10），横坐标的值越大表示对应的结构域分类下富集度的显著性水平越高，颜色梯度代表富集因子的大小（Rich Factor≤1），富集因子表示注释到某结构域的差异表达蛋白质数目占注释到该结构域的所有鉴定到的蛋白质数目的比例，颜色越接近红色代表Rich Factor值越大，气泡的大小表示每个结构域分类下差异蛋白质数目。

输出文件：

1）3-3-2结构域分析

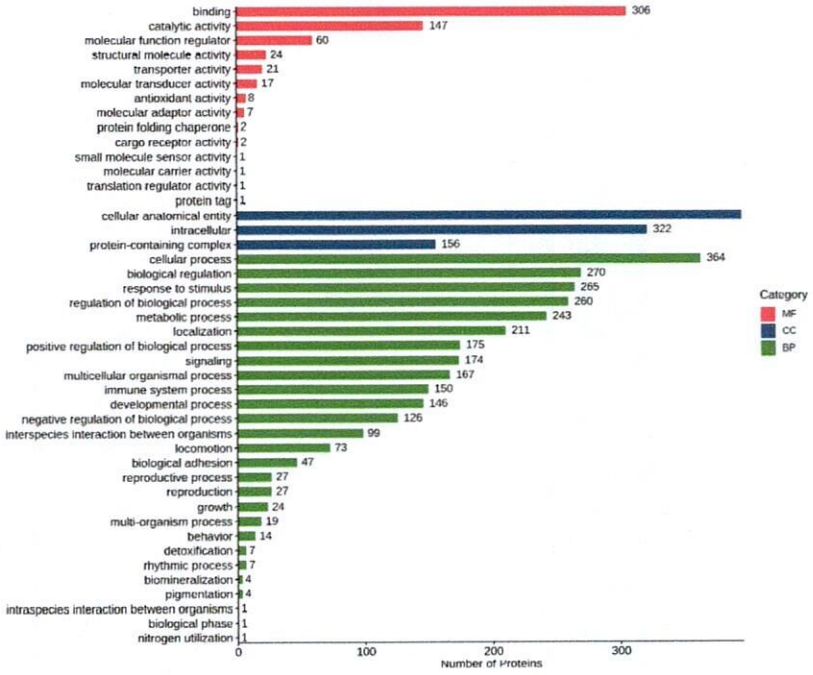
3.3.3 GO 功能分析

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]软件分别对所有显著差异表达蛋白质、显著上调差异蛋白质、显著下调差异蛋白质进行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显著差异蛋白数目进行统计，结果如下。

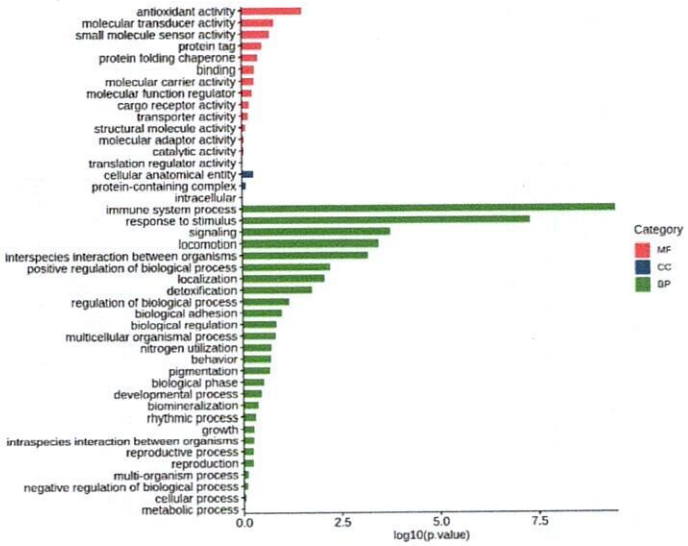
更多信息请参考：http://www.geneontology.org/http://en.wikipedia.org/wiki/Gene_Ontology

3.3.3.1 所有显著差异表达蛋白质 GO 功能分析



groupvs组所有显著差异表达蛋白质的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），包含分子功能（Molecular Function），细胞组分（Cellular Component）和生物过程（Biological Process），依次以红色，蓝色，绿色予以区分；横坐标表示每个功能分类下的显著差异表达蛋白质数目。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。



groupvs组所有显著差异表达蛋白质的GO注释统计图（level 2）

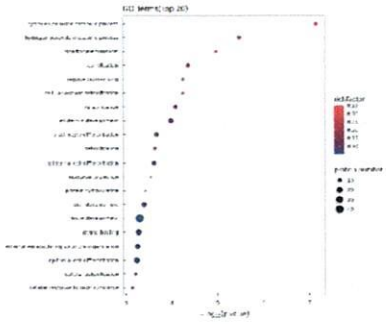
说明：图中纵坐标表示GO 二级功能注释信息（GO Level2），包含分子功能（Molecular Function），细胞组分（Cellular Component）和生物过程（Biological Process），依次以红色，蓝色，绿色予以区分；横坐标表示富集显著性，即基于Fisher精确检验（Fisher’s Exact Test）计算P值（取-log10），横坐标的值越大表示对应的GO功能下富集度的显著性水平越高。

输出文件：

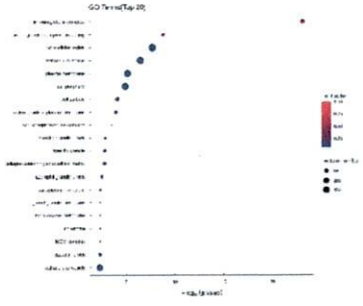
1）3-3-3 GO功能分析

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。

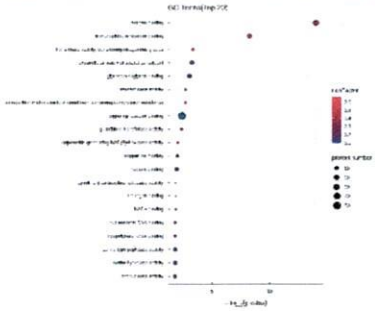
将所有差异表达蛋白质与参考物种的全部蛋白质（或实验鉴定到的所有蛋白质）以 GO 功能的注释结果进行对照比较，通过 Fisher 精确检验 (Fisher’s Exact Test)得出两者差异的显著性，从而找到所有差异表达蛋白质富集的功能类别 （P value <0.05）。用气泡图分别显示 GO 三大分类下的 GO 条目富集情况。



groupvs组所有差异蛋白质的GO富集气泡图(BP)



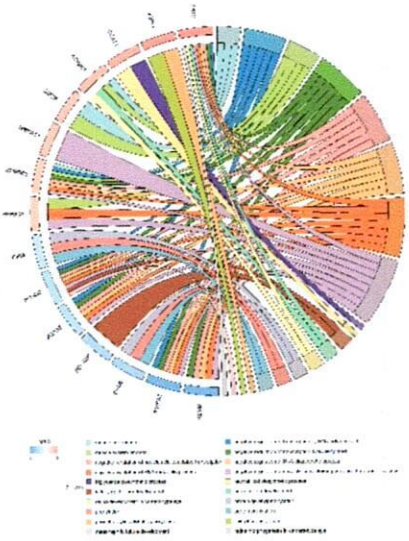
groupvs组所有差异蛋白质的GO富集气泡图(CC)



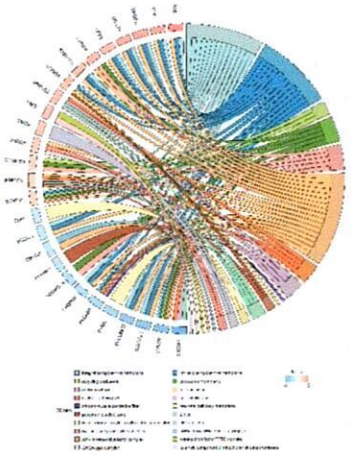
groupvs组所有差异蛋白质的GO富集气泡图(MF)

说明：图中横坐标为某GO功能的富集显著性，即基于Fisher精确检验（Fisher’s Exact Test）计算P值（取-log10），横坐标的值越大表示对应的GO功能下富集度的显著性水平越高，颜色梯度代表富集因子的大小（Rich Factor≤1），富集因子表示注释到某GO功能的差异表达蛋白质数目占注释到该GO功能的所有鉴定到的蛋白质数目的比例，颜色越接近红色代表Rich Factor值越大，气泡的大小表示每个GO功能分类下差异蛋白质数目。**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。**

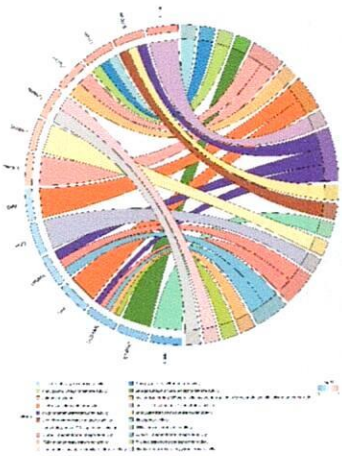
显著富集弦图，用于展示显著富集的GO功能与蛋白之间的关系，图的右侧表示富集到的GO功能，与右侧功能相连的是该功能中的差异蛋白，差异蛋白的顺序依据其Log2FC值从大到小排列。该图能直观的展示富集GO功能中每个蛋白的名字、差异程度。



groupvs组所有差异蛋白质的GO富集弦图（BP）



groupvs组所有差异蛋白质的GO富集弦图(CC)

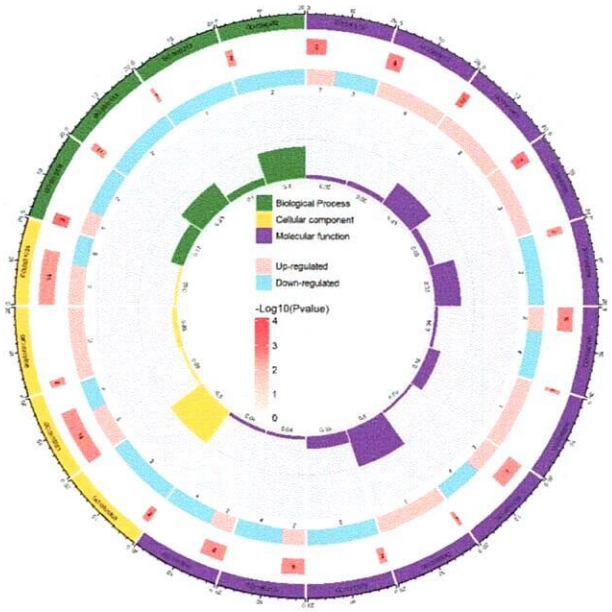


groupvs组所有差异蛋白质的GO富集弦图(MF)

输出文件：

1）3-3-3 GO功能分析

Circos每一圈含义（由外到内）：第一个圆圈：富集的GO功能一级分类，圆圈外是蛋白数量的坐标标尺。不同的颜色代表不同的类别；第二个圆圈：功能富集显著性P value经-Log10转换后的值。数值越大，颜色越红；第三圈：上、下调差异蛋白数量条形图，红色代表上调差异蛋白数量，蓝色代表下调差异蛋白数量；第四个圆圈：每个功能的富集因子的大小（Rich Factor≤1）。注：富集条目少于4不显示。



groupvs组所有差异蛋白质的GO富集Circos图

输出文件：

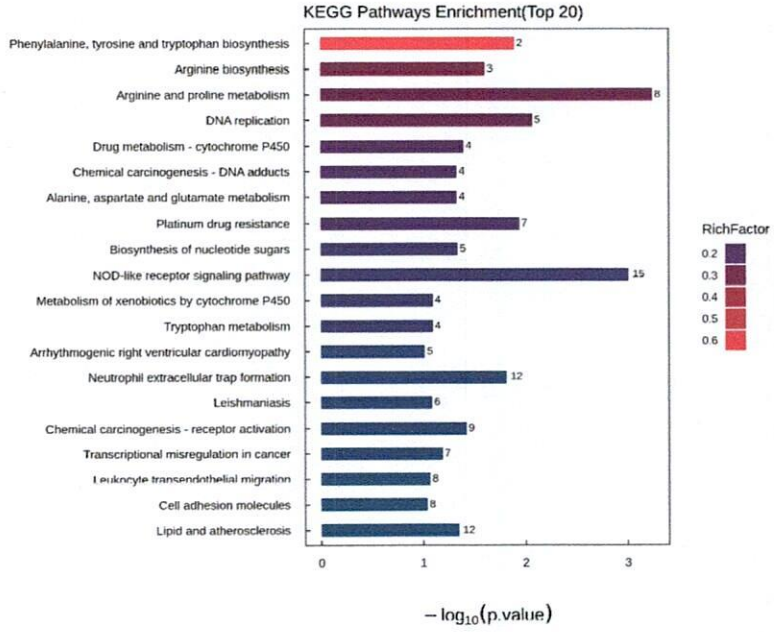
1）3-3-3 GO功能分析

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（Kyoto Encyclopedia of Genes and Genomes）数据库对蛋白质解析注释[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。

3.3.4.1 所有显著差异蛋白质 KEGG 通路注释分析

本项目将所有显著差异蛋白质进行KEGG通路注释，注释表格参见输出文件。结果如下图所示。更多信息请参考：http://www.genome.jp/kegg/pathway.html。



groupvs 组显著差异蛋白质的 KEGG 通路注释统计图（Top20）

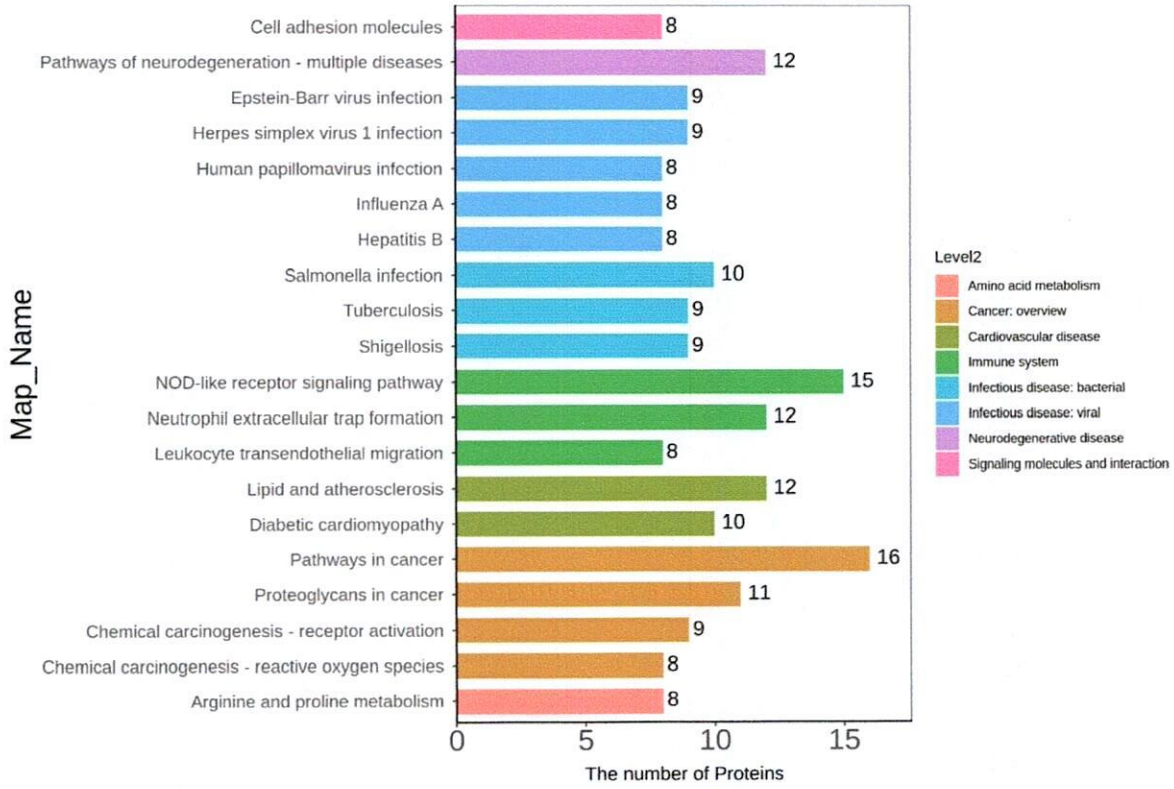
说明：图中纵坐标是含差异表达蛋白质参与的通路名称，横坐标表示富集显著性，即基于Fisher精确检验（Fisher's Exact Test）计算P值（取-log10），横坐标的值越大表示对应的通路下富集度的显著性水平越高。一般情况下，参与某一通路的差异表达蛋白质数目越多，说明该通路越重要，需要重点关注或者进行后续深入机制的探讨。

输出文件：

1) 3-3-4KEGG通路注释分析

3.3.4.2 所有显著差异蛋白质 KEGG 通路归属分析

在生物体内，不同蛋白相互协调行使其生物学行为，基于Pathway的分析有助于更进一步了解其生物学功能Pathway富集不同层级结果。 KEGG代谢通路共分为7个分支：细胞过程(Cellular Processes)、环境信息处理(Environmental Information Processing)、遗传信息处理(Genetic Information Processing)、人类疾病(Human Diseases)(仅限动物)、代谢(Metabolism)、有机系统(OrganismalSystems)、药物开发(DrugDevelopment)。本分析通过图形展示众多的代谢途径以及通路归属关系，以便更加直观观测到差异表达蛋白所参与的代谢途径。差异表达蛋白质参与的通路代谢注释如下展示。



groupvs 组显著差异蛋白质的 KEGG 通路注释及归属柱状图

说明：横坐标轴代表通路蛋白注释数目，纵坐标代表KEGG注释名称。不同颜色代表不同KEGG的代谢通路level2层级。。

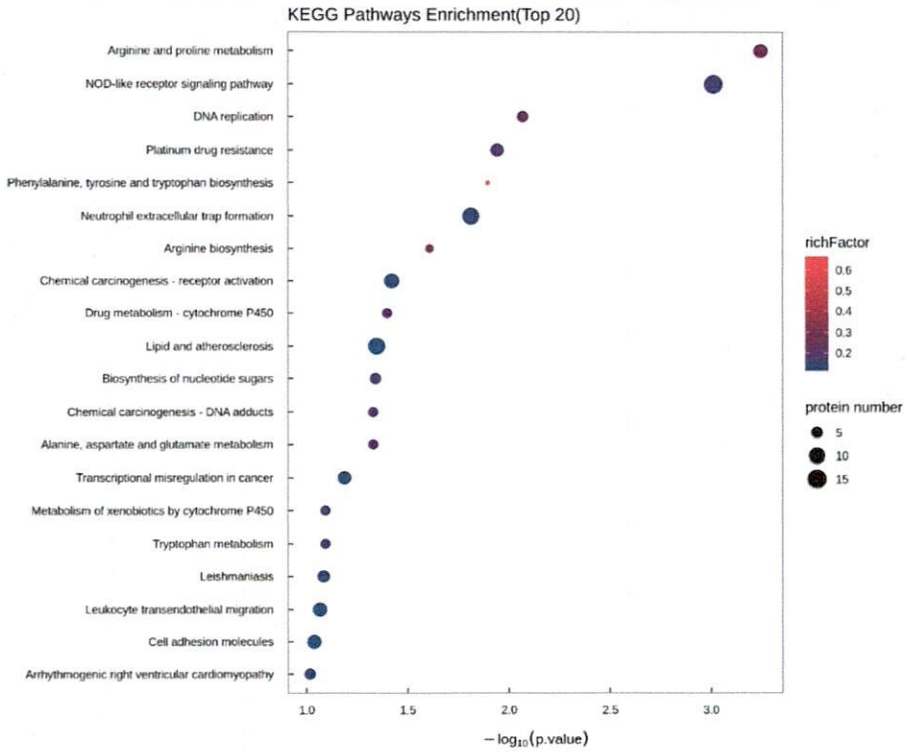
输出文件：

1) 3-3-4-1KEGG通路注释分析

3.3.4.3 所有显著差异蛋白质 KEGG 通路富集分析

为了揭示所有差异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。

将所有显著差异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 groupvs 比较组的显著差异蛋白质进行 KEGG 通路富集分析。



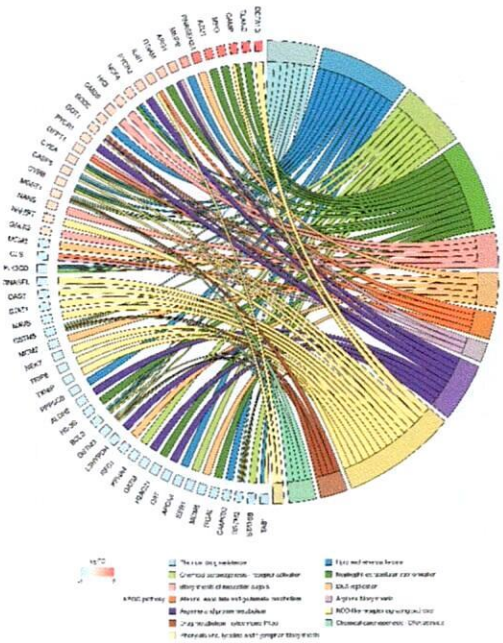
groupvs组所有差异蛋白质的KEGG通路富集气泡图

说明：图中横坐标为某KEGG通路的富集显著性，即基于Fisher精确检验（Fisher’s Exact Test）计算P值（取-log10），横坐标的值越大表示对应代谢通路富集度的显著性水平越高，颜色梯度代表富集因子的大小（Rich Fatcor≤1），富集因子表示注释到KEGG通路类别的显著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，颜色越接近红色代表Rich Factor值越大，气泡的大小表示每个KEGG通路下差异蛋白质数目。**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。**

输出文件：

1) 3-3-4KEGG通路富集分析

显著富集弦图，用于展示显著富集的 KEGG 通路与蛋白之间的关系，图的右侧表示富集到的 KEGG 通路，与右侧通路相连的是该通路中的差异蛋白，差异蛋白的顺序依据其 Log2FC 值从大到小排列。该图能直观的展示富集通路中每个蛋白的名字、差异程度。

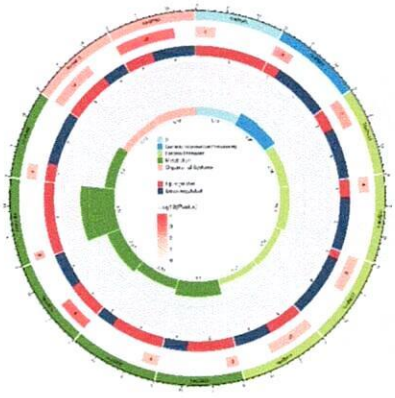


KEGG 通路的富集统计弦图

输出文件：

1) 3-3-4KEGG通路富集分析

对差异蛋白KEGG通路注释结果进行富集分析，以Circos图形式来进行结果展示。Circos每一圈含义（由外到内）：第一个圆圈：富集的KEGG通路，圆圈外是蛋白数量的坐标标尺；第二个圆圈：蛋白KEGG通路富集显著性P value经-Log10转换后的值。数值越大，颜色越红；第三圈：上、下调差异蛋白数量条形图，红色代表上调差异蛋白数量，蓝色代表下调差异蛋白数量；第四个圆圈：每个KEGG通路的富集因子的大小（Rich Factor≤1）。注：富集条目少于4不显示。



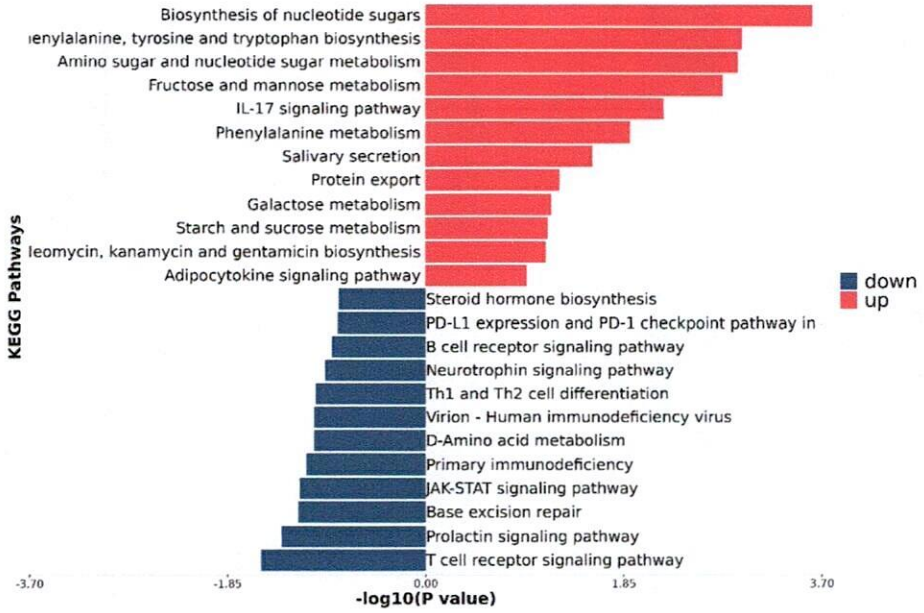
KEGG 通路的富集统计 Circos 图（top 20）

输出文件：

1） 3-3-4KEGG通路富集分析

3.3.4.4 显著上、下调差异蛋白质 KEGG 通路富集分析

为了更好考察差异蛋白的通路富集的显著性，分别对上、下调差异表达蛋白质进行KEGG通路富集分析，以蝴蝶图形式展示，结果展示如下：



groupvs组显著上、下调差异表达蛋白质的通路富集蝴蝶图

说明：横坐标为Fisher精确检验的p value值（取以10为底的对数），纵坐标表示通路名称。上调和下调蛋白参与的通路用红色（右）和蓝色（左）条表示。

每个差异蛋白注释到的通路图信息及在通路中的位置，在附件中展示。

输出文件：

1）3-3-4KEGG通路富集分析

每个差异蛋白注释到的通路图信息及在通路中的位置，在附件中展示。

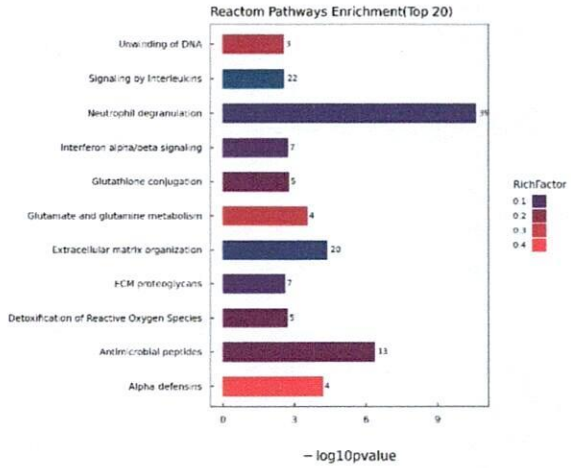
输出文件：

1）3-3-4KEGG通路富集分析

3.3.5 Reactome 通路富集分析

Reactome是一个免费提供的开源关系数据库，其中包含信号和代谢分子及其组织成生物途径和过程的关系。Reactome数据模型的核心单元是反应。参与反应的实体（核酸、蛋白质、复合物、疫苗、抗癌治疗剂和小分子）形成生物相互作用网络，并被分组为通路。Reactome中的生物学途径的例子包括经典的中间代谢、信号传导、转录调控、细胞凋亡和疾病。限人、大鼠、小鼠物种。

显著富集条形图，横轴表示-Log10转换的富集显著性P value，纵轴为对应Reactome通路描述信息。条形图长短表示富集显著性，越长代表富集显著性越强。

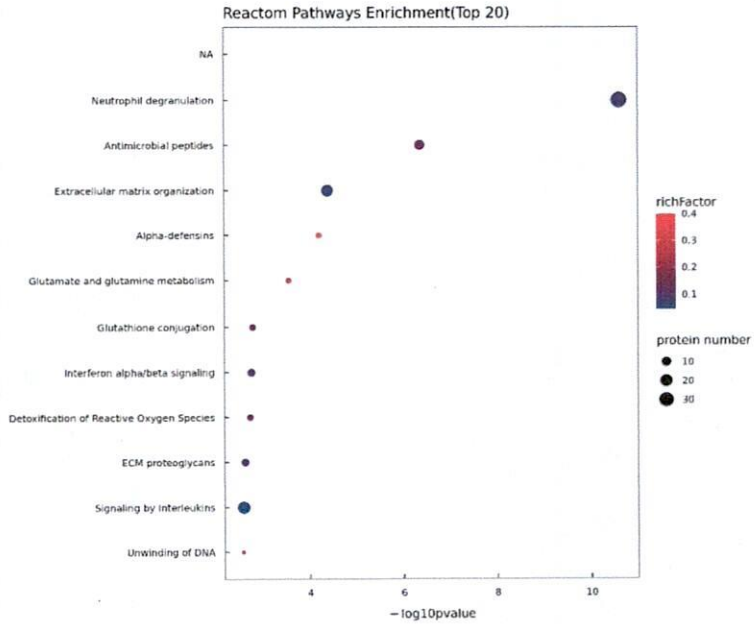


groupvs组所有差异蛋白质的富集条形图

输出文件：

1）3-3-5 Reactome通路富集分析

给出了最显著富集的前20个功能的结果，图中纵轴为Reactome通路描述信息，横坐标为某通路的富集显著性，即基于Fisher精确检验（Fisher’s Exact Test）计算P值（取-log10），横坐标的值越大表示对应代谢通路富集度的显著性水平越高，颜色梯度代表富集因子的大小（Rich Factor≤1），富集因子表示注释到该通路类别的显著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，颜色越接近红色代表Rich Factor值越大，气泡的大小表示该通路下差异蛋白质数目。

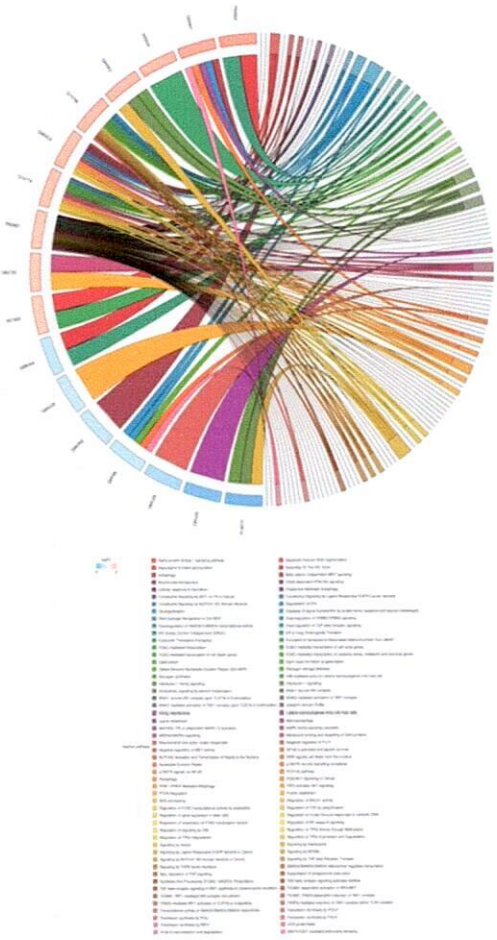


groupvs组所有差异蛋白质的富集气泡图

输出文件：

1）3-3-5 Reactome通路富集分析

显著富集弦图，用于展示显著富集的Reactome通路与蛋白之间的关系，图的右侧表示富集到的Reactome通路，与右侧Reactome通路相连的是该通路中的差异蛋白，差异蛋白的顺序依据其Log2FC值从大到小排列。该图能直观的展示富集通路中每个蛋白的名字、差异程度。

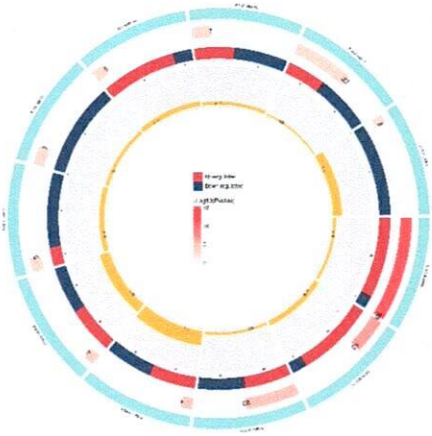


groupvs组所有差异蛋白质富集弦图

输出文件：

1）3-3-5 Reactome通路富集分析

Circos每一圈含义（由外到内）：第一个圆圈：富集的Reactome通路，圆圈外是蛋白数量的坐标标尺；第二个圆圈：Reactome通路富集显著性P value经-Log10转换后的值。数值越大，颜色越红；第三圈：上、下调差异蛋白数量条形图，红色代表上调差异蛋白数量，蓝色代表下调差异蛋白数量；第四个圆圈：每个功能的富集因子的大小（Rich Factor≤1）。



groupvs组所有差异蛋白质富集Circos图

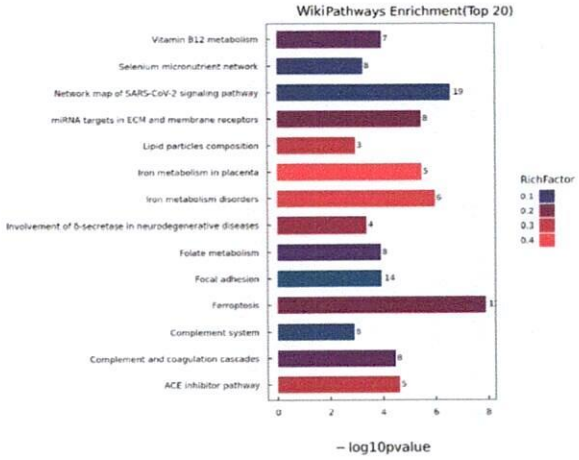
输出文件：

1）3-3-5 Reactome通路富集分析

3.3.6 WikiPathways通路富集分析

WikiPathways的建立是为了促进生物学界对通路信息的贡献和维护。WikiPathways是一个开放的致力于管理生物通路数据库。因此，WikiPathways为生物途径数据库提供了一种新模型，可增强和补充现有通路数据库信息，例如KEGG、Reactome和Pathway Commons等。限人、大鼠、小鼠物种。

显著富集条形图，横轴表示-Log10转换的富集显著性P value；纵轴为对应WikiPathways通路描述信息。条形图长短表示富集显著性，越长代表富集显著性越强。



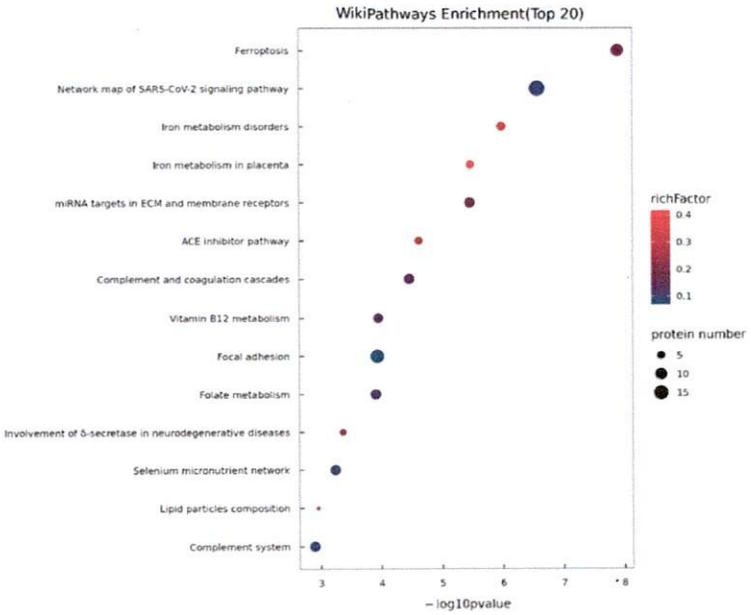
groupvs组所有差异蛋白质的富集条形图

输出文件：

1）3-3-6 WikiPathways通路富集分析

显著富集气泡图中给出了最显著富集的前20个功能的结果，图中纵轴为WikiPathways通路描述

信息，横坐标为某通路的富集显著性，即基于Fisher精确检验（Fisher's Exact Test)计算P值（取-log10），横坐标的值越大表示对应代谢通路富集度的显著性水平越高，颜色梯度代表富集因子的大小（Rich Factor≤1），富集因子表示注释到该通路类别的显著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，颜色越接近红色代表Rich Factor值越大，气泡的大小表示该通路下差异蛋白质数目。

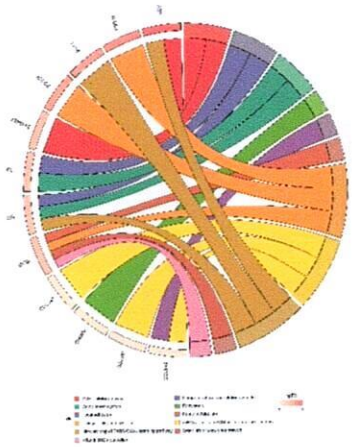


groupvs组所有差异蛋白质的富集气泡图

输出文件：

1）3-3-6 WikiPathways通路富集分析

显著富集弦图，用于展示显著富集的WikiPathways通路与蛋白之间的关系，图的右侧表示富集到的WikiPathways通路，与右侧WikiPathways通路相连的是该通路中的差异蛋白，差异蛋白的顺序依据其Log2FC值从大到小排列。该图能直观的展示富集通路中每个蛋白的名字、差异程度。

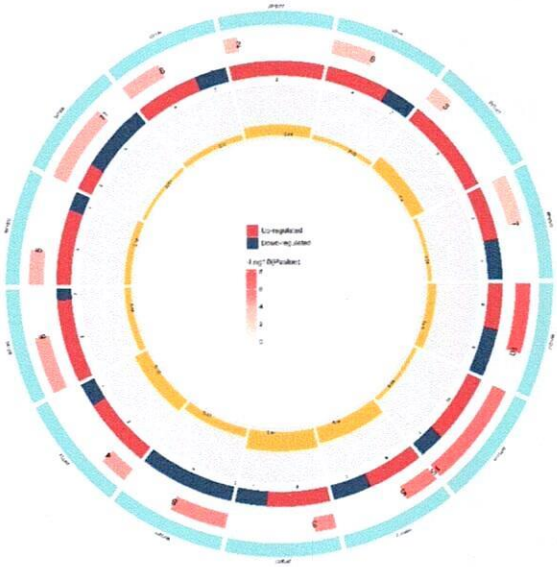


groupvs组所有差异蛋白质富集弦图

输出文件：

1）3-3-6 WikiPathways通路富集分析

Circos每一圈含义（由外到内）：第一个圆圈：富集的WikiPathways通路；第二个圆圈：WikiPathways通路富集显著性P value经-Log10转换后的值。数值越大，颜色越红；第三圈：上、下调差异蛋白数量条形图，红色代表上调差异蛋白数量，蓝色代表下调差异蛋白数量；第四个圆圈：每个功能的富集因子的大小（Rich Factor≤1）。注：富集条目少于4不显示。



groupvs组所有差异蛋白质富集Circos图

输出文件：

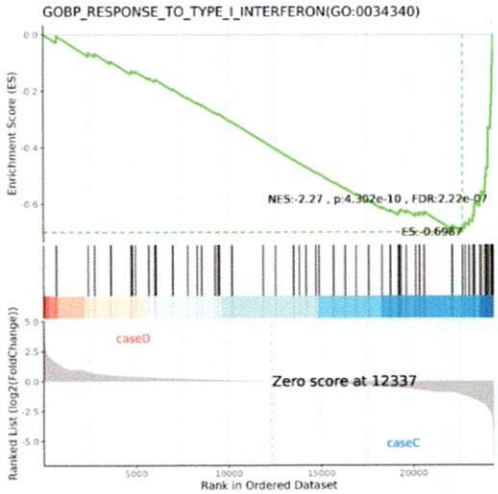
1）3-3-6 WikiPathway通路富集分析

3.3.7 GSEA 分析

传统的蛋白功能富集方法是基于超几何检验而针对差异表达蛋白进行富集的，但当单个蛋白表达变化不大时，基于传统富集分析得到结果可能会很少，甚至没有结果。GSEA分析（Gene Set Enrichment Analysis）能够有效弥补传统富集分析对信息挖掘不足等问题，更能全面地对某一功能单位（通路、GOterm或其他）的调节作用进行解释。它可以将那些在传统富集分析信息中容易遗漏掉的差异表达不显著却有着重要生物学意义的基因包含在内，也可以解决传统富集分析中因为得到的差异基因较少，而无法开展功能富集分析或者无法富集到感兴趣的通路的问题。其基本思想是使用预定义的蛋白集，将蛋白按照在两类样本中的差异表达程度排序，然后检验预先设定的蛋白集是否在这个排序表的顶端或者末端富集。此外，通过GSEA分析可以判断某条通路中基因的总体变化趋势，以及该通路到底是激活还是抑制状态。GSEA富集分析主要包括三个步骤：计算富集得分（Enrichment Score）；估计富集得分的显著性水平；矫正多重假设验证。我们分别对物种的GO、KEGG、Reactome、WikiPathways数据集进行GSEA分析，显著富集的蛋白集呈图展示见附件。限人、大鼠、小鼠、果蝇和酵母物种。

3.3.7.1 GSEA GO功能分析

通过GSEA的方式将可定量蛋白进行GO功能条目富集分析。



Groupvs组GSEA GO富集分析

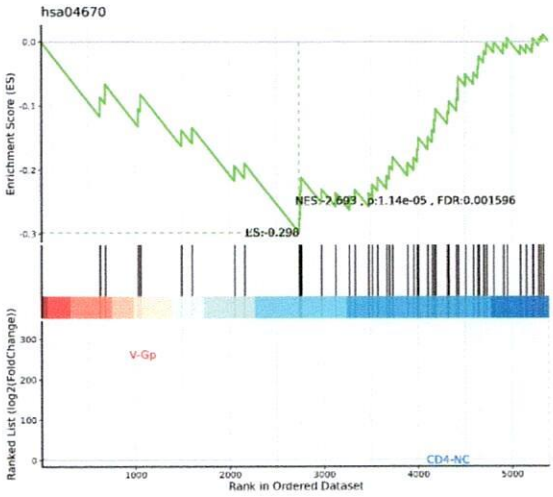
说明：横轴为比较组中的蛋白（蛋白集）根据其表达量变化值（Log2FC）由大到小的排序。GSEA富集图自上而下分为三部分：①上部分显示的是当分析沿着蛋白集按排序计算时，ES（Enrichment Score） 值在计算到每个蛋白位置时的展示(即分析过程中动态的ES值)，最高峰处的ES得分即为该通路的ES值。②中间部分俗称条形码图，用线条标记了该通路中涉及到的蛋白出现在蛋白集排序列表中的位置。红蓝相间的热图是表达丰度排列（红色越深的表示该位置的基因log2FC越大 ，蓝色越深表示log2FC越小）。③最下面部分为排序后比较组中蛋白Log2FC的排序，以灰色面积图展示。 图的右上侧的注释为GO通路富集pvalue和FDR值。

输出文件：

1) 3-3-7 GSEA GO分析

3.3.7.2 GSEA KEGG富集分析

通过GSEA的方式将可定量蛋白进行KEGG通路富集分析。



Groupvs组GSEA KEGG通路富集分析

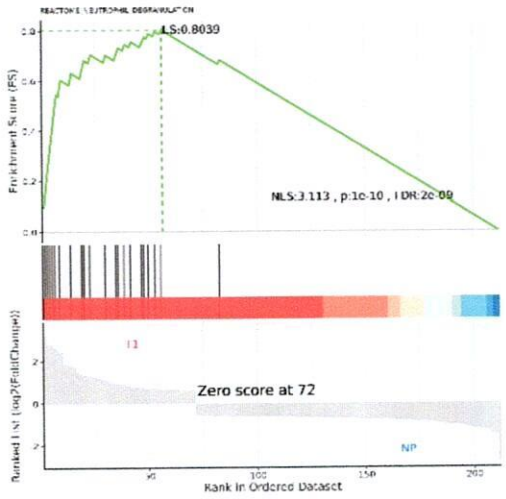
说明：横轴为比较组中的蛋白（蛋白集）根据其表达量变化值（Log2FC）由大到小的排序。GSEA富集图自上而下分为三部分：①上部分显示的是当分析沿着蛋白集按排序计算时，ES（Enrichment Score） 值在计算到每个蛋白位置时的展示(即分析过程中动态的ES值)，最高峰处的ES得分即为该通路的ES值。②中间部分俗称条形码图，用线条标记了该通路中涉及到的蛋白出现在蛋白集排序列表中的位置。红蓝相间的热图是表达丰度排列（红色越深的表示该位置的基因log2FC越大 ，蓝色越深表示log2FC越小）。③最下面部分为排序后比较组中蛋白Log2FC的排序，以灰色面积图展示。 图的右上侧的注释为KEGG通路富集pvalue和FDR值。

输出文件：

1) 3-3-7 GSEA KEGG分析

3.3.7.3 GSEA Reactome 富集分析

通过GSEA的方式将可定量蛋白进行Reactome通路富集分析。



Groupvs组GSEA Reactome富集分析

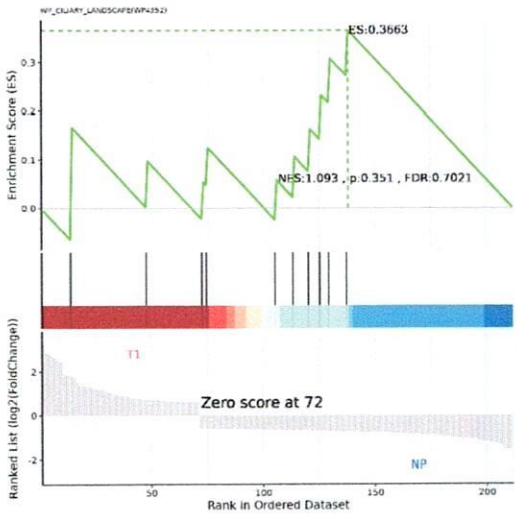
说明：横轴为比较组中的蛋白（蛋白集）根据其表达量变化值（Log2FC）由大到小的排序。GSEA富集图自上而下分为三部分：①上部分显示的是当分析沿着蛋白集按排序计算时，ES（Enrichment Score） 值在计算到每个蛋白位置时的展示(即分析过程中动态的ES值)，最高峰处的ES得分即为该通路的ES值。②中间部分俗称条形码图，用线条标记了该通路中涉及到的蛋白出现在蛋白集排序列表中的位置。红蓝相间的热图是表达丰度排列（红色越深的表示该位置的基因log2FC越大 ，蓝色越深表示log2FC越小）。③最下面部分为排序后比较组中蛋白Log2FC的排序，以灰色面积图展示。 图的右上侧的注释为Reactome通路富集pvalue和FDR值。

输出文件：

1) 3-3-7 GSEA Reactome分析

3.3.7.4 GSEA WikiPathways 富集分析

通过GSEA的方式将可定量蛋白进行WikiPathways通路富集分析。



Groupvs组GSEA WikiPathways富集分析

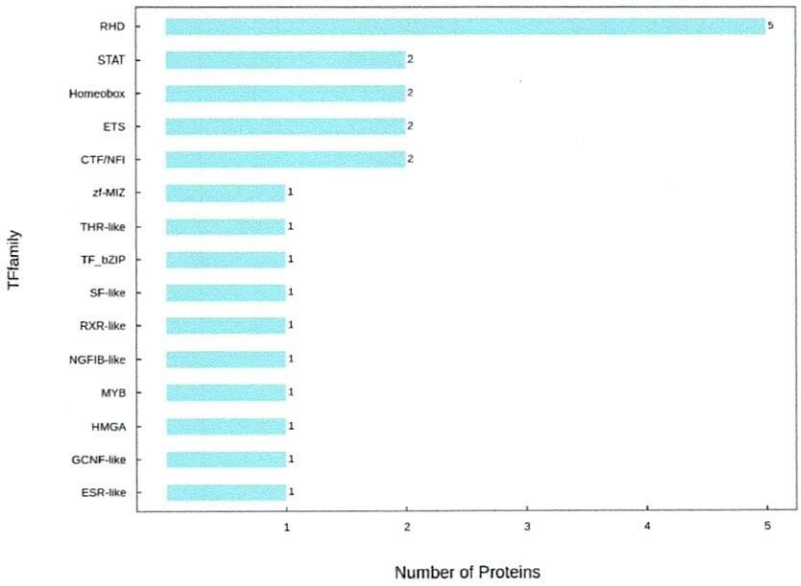
说明：横轴为比较组中的蛋白（蛋白集）根据其表达量变化值（Log2FC）由大到小的排序。GSEA富集图自上而下分为三部分：①上部分显示的是当分析沿着蛋白集按排序计算时，ES（Enrichment Score） 值在计算到每个蛋白位置时的展示(即分析过程中动态的ES值)，最高峰处的ES得分即为该通路的ES值。②中间部分俗称条形码图，用线条标记了该通路中涉及到的蛋白出现在蛋白集排序列表中的位置。红蓝相间的热图是表达丰度排列（红色越深的表示该位置的基因log2FC越大 ，蓝色越深表示log2FC越小）。③最下面部分为排序后比较组中蛋白Log2FC的排序，以灰色面积图展示。 图的右上侧的注释为WikiPathways通路富集pvalue和FDR值。

输出文件：

1) 3-3-7 GSEA WikiPathways分析

3.3.8 转录因子分析

转录因子(TranscriptionFactor，TF)是指能够以序列特异性方式结合DNA并且调节转录的蛋白质，由于转录因子有特殊的功能，会对这类蛋白进行注释并进行深入分析。PlantTFDB（Plant Transcription Factor Database）和AnimalTFDB（Animal Transcription Factor Database）数据库分别包含植物和动物的转录因子及转录因子家族信息，可预测所关注的蛋白是否为转录因子，以及所属的转录因子家族。



转录因子注释结果条形图

说明：纵坐标代表转录因子家族，横坐标代表注释到该转录因子家族的蛋白数目，浅蓝色为注释到该转录因子家族的差异蛋白数量，深蓝色为注释到该转录因子家族的鉴定所有蛋白数量。

输出文件：

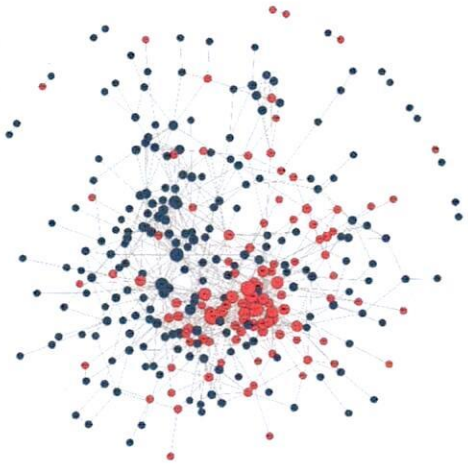
1) 3-3-8转录因子分析

3.3.9 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（ProteinProteinInteraction，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 数据库中的蛋白质相互作用关系，对 groupvs 比较组的差异表达蛋白质构建蛋白质互作网络图，如下图。



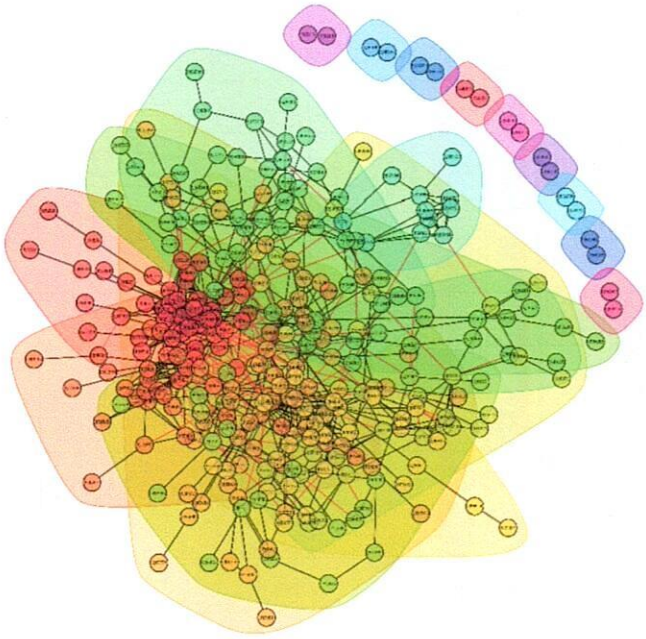
groupvs 组差异表达蛋白质相互作用网络

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（上调标注为红色、下调标注为蓝色、），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。**

输出文件：

1) 3-3-9蛋白互作网络分析

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。



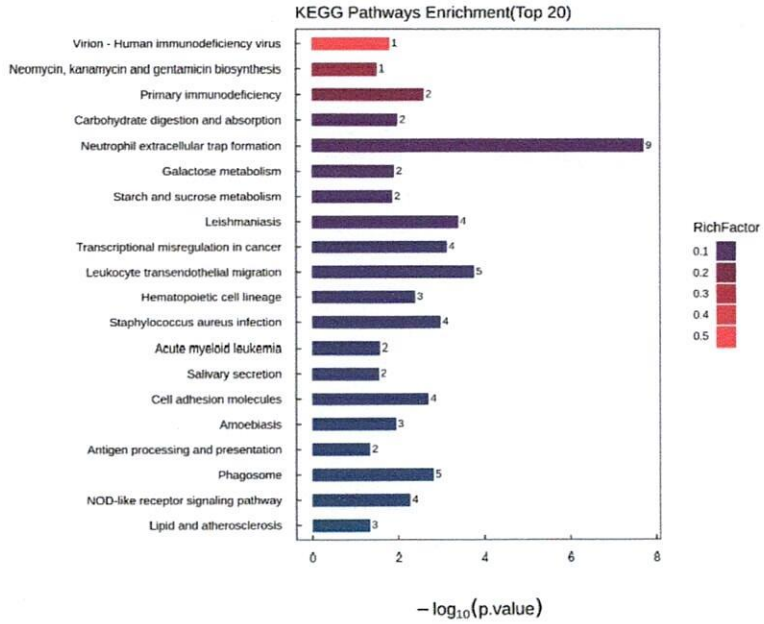
groupvs 组互作蛋白 module 分析图

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)3-3-9 蛋白互作网络分析

根据上图的高度聚集的蛋白质归类结果，从多到少分成多个 Cluster，选取前 5 个功能簇，每个 Cluster 中的相关高度聚集蛋白质进行 KEGG 富集分析，绘制 KEGG 富集条目柱形图和气泡图，文中仅展示一个 Cluster 的结果，其他结果见附件。进行 KEGG 通路注释统计，结果如下图所示。

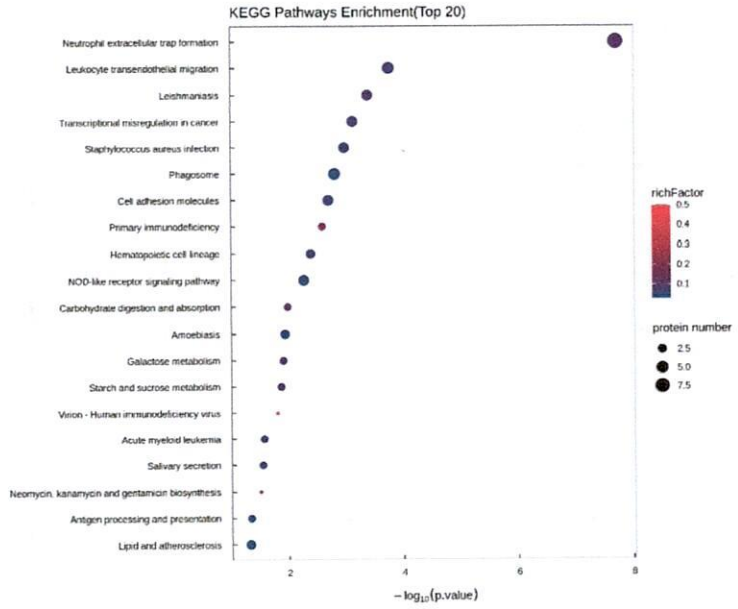


groupvs 组显著差异蛋白质的 KEGG 通路注释统计图（Top20）

说明：图中纵坐标是含差异表达蛋白质参与的通路名称，横坐标表示富集显著性，即基于 Fisher 精确检验（Fisher’s Exact Test）计算 P 值（取-log10），横坐标的值越大表示对应的通路下富集度的显著性水平越高。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。

输出文件：

1) 3-3-9 蛋白互作网络分析



groupvs 组所有差异蛋白质的 KEGG 通路富集气泡图（Top20）

输出文件：

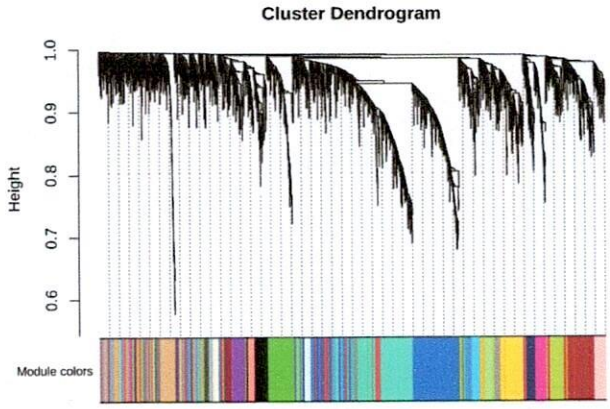
1）3-3-9 蛋白互作网络分析

3.3.10 表型与组学关联分析

WGCNA 分析(weighted gene co-expression network analysis)，主要原理是通过加权共表达网络分析的方式分析多样本蛋白的表达模式，鉴定出高度协同变化的蛋白模块（module），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。主要应用特色有两点：一是将表型数据如性别、年龄或者其他生理特征与组学进行关联分析，寻找表型与分子数据的关键的功能调控模块。二是通过基因间的表达相关性及权重挖掘关键功能分子，识别已知基因的新功能。

3.3.10.1 关键功能模块分析

首先，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：



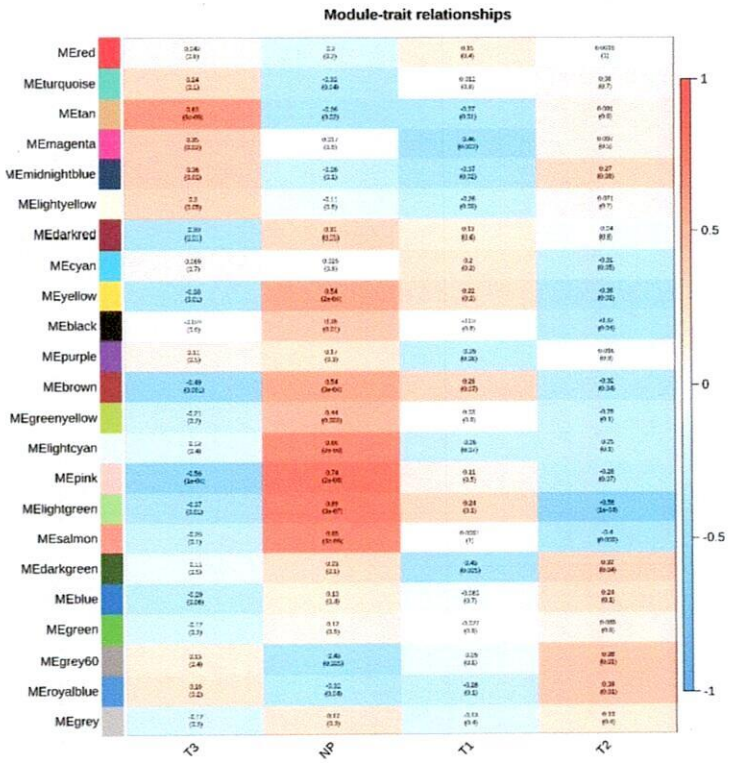
蛋白共表达模块划分图

说明：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。

输出文件：

1）3-3-10 Module构建

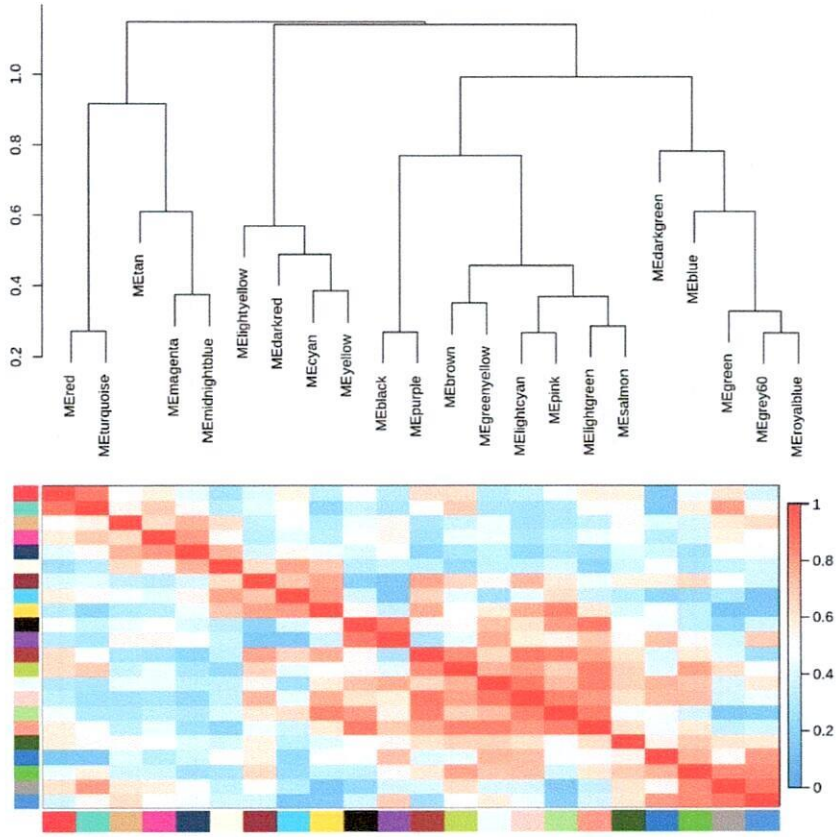
其次，以分组或临床表型作为性状，可以获得与该性状相关的共表达模块情况。进一步可从模块与表型性状相关性图中挑选感兴趣的性状所对应的模块，一般筛选标准：相关性数值越接近±1，相关性检验P值小于0.05，表明该模块是决定该性状的关键模块。



模块与表型性状相关性热图

说明：横坐标为性状或表型因素； 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蛋白 module，同时在色块上标注了 p-value （括号中的值）及相关性系数值；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、蓝色代表负相关，相关性越高则颜色越深，相关性越低则颜色越浅。

再者，观察模块特征值的聚类树权图，并通过聚类热图识别出表达模式更加相似的模块组。如下图所示：



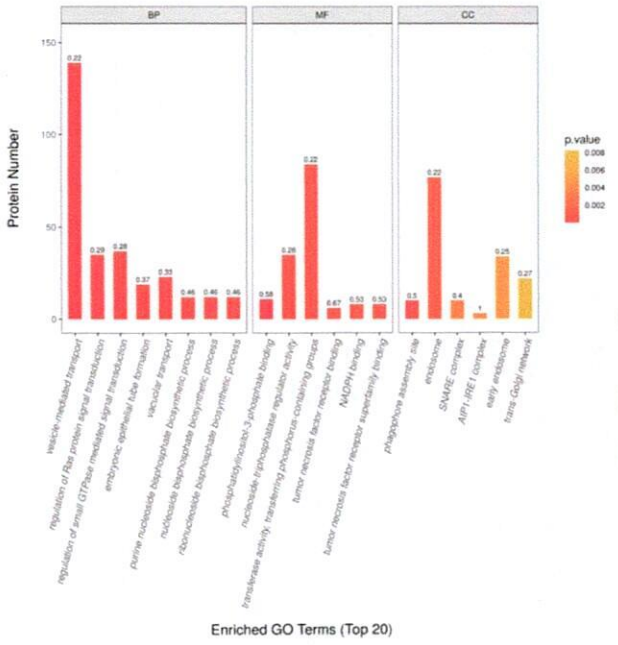
模块（性状）关联聚类图

说明：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

输出文件：

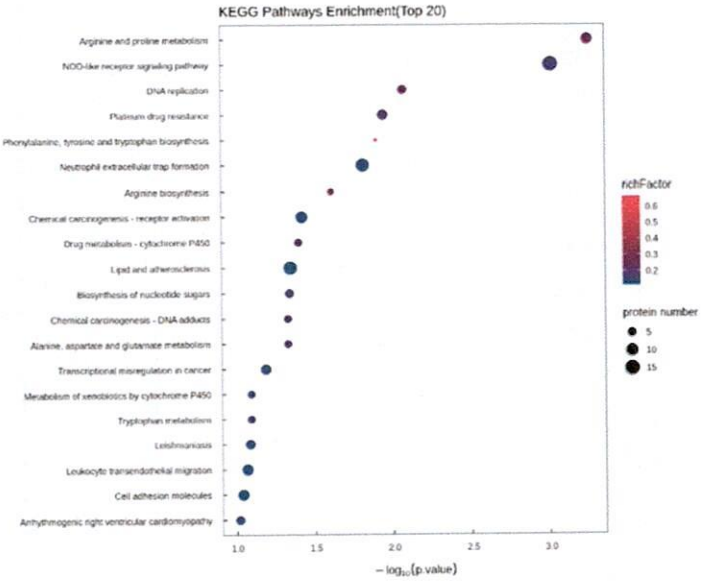
1）3-3-10 Module构建

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，如下图所示：



某模块蛋白的GO富集分析图

说明：横坐标：GO功能类别；纵坐标：与GO功能相关的蛋白质数量；条形图数字标签标识：富集因子（Rich factor）；颜色深浅表示p值大小，即某个功能受到影响的显著程度。



某模块蛋白的KEGG富集分析图

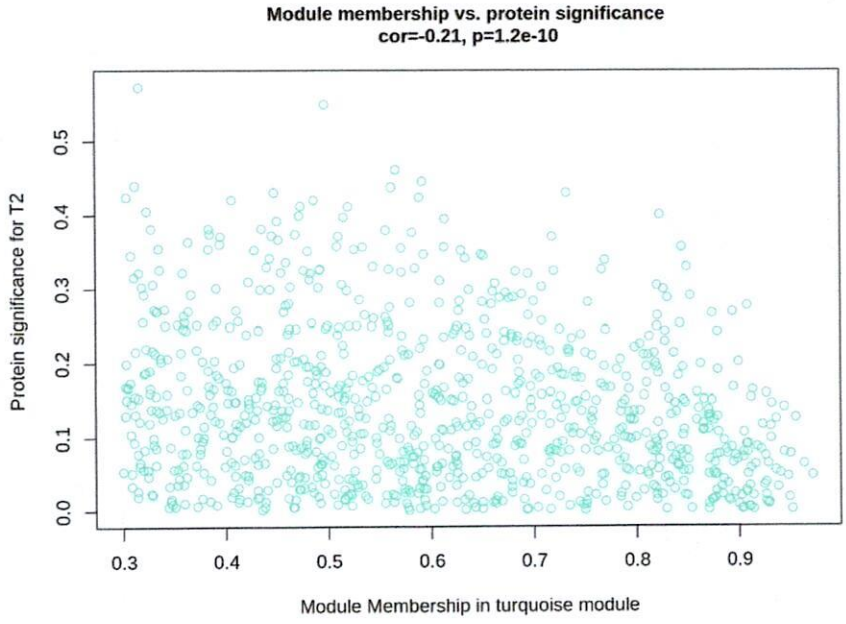
说明：图中横坐标为某KEGG通路的富集显著性，即基于Fisher精确检验（Fisher’s Exact Test）计算P值（取-log10），横坐标的值越大表示对应代谢通路富集度的显著性水平越高，颜色梯度代表富集因子的大小（Rich Factor≤1），富集因子表示注释到KEGG通路类别的显著差异表达蛋白质数目占注释到该类别的所有鉴定到的蛋白质数目的比例，颜色越接近红色代表Rich Factor值越大，气泡的大小表示每个KEGG通路下差异蛋白质数目。

输出文件：

1）3-3-10 Module构建

3.3.10.2 重要模块核心蛋白分析

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（蛋白与性状/样本的相关性）与Module Membership（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与模块的相关性和基因与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Hub protein（核心蛋白），如下图所示：



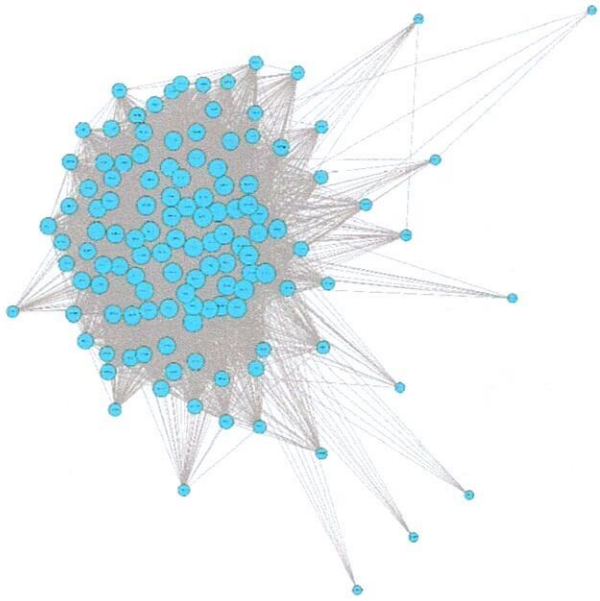
groupvs散点分布图

说明：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。

输出文件：

1）3-3-10 Module构建

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，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。



模块内蛋白共表达网络分析

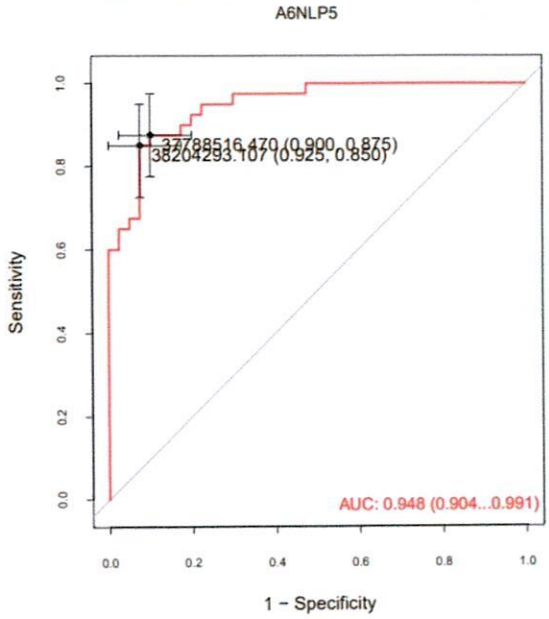
说明：圆圈颜色与模块颜色相同，如该图圆圈为蓝色，表示该互作网络为蓝色模块的蛋白构建的互作网络， 节点大小与其连接度（Degree）正相关，即与该蛋白相关的蛋白越多， 则其连接度越大，节点尺寸就越大，其在网络中的地位越关键。 线条表示蛋白之间的互作关系，线条的粗细与相关系数的绝对值成正比，即线条越粗，相关程度越高。

输出文件：

1）3-3-10 Module构建

3.3.11 ROC 分析

ROC 分析（receiver operating characteristic curve，受试者工作特征曲线）是把灵敏度和特异度结合起来综合评价诊断准确度或判别效果的一种方法，在医学领域中广泛用于临床诊断、人群筛选等研究。在蛋白组学中，对比较组间的差异蛋白进行ROC分析，可以用来指示biomarker区分两组间（实验组和对照组）的能力，展示AUC值TOP25的蛋白。



ROC分析

说明：横坐标为1-特异性，即假阳性率——阴性群体中，检测为阳性的概率，希望该值越低越好；纵坐标为敏感度，即真阳性率——阳性群体中，检测出阳性的概率，希望该值越高越好；曲线越往左上角说明预测准确率越高，曲线下面积越大，即AUC值越大说明预测准确率越高。图形中红色标识的文字为该曲线对应的AUC值和95%的置信区间；黑色标识的文字为原始的强度表最佳临界值，括号中位特异度和灵敏度。

输出文件：

1）3-3-11 ROC

3.3.12 高级生信分析以及临床转化方案

1. 高级生信分析方案

1.1分子分型方案

方案简介

利用**基于质谱的蛋白质组等多组学**方法，分析临床队列样本的多维度分子图谱，在分子分型基础上进一步寻找预后标志物及发现治疗靶点与药物。

方案特点

更前沿的顶刊研究模式

2019-2020年 Cell 及Nature 10篇、Cancer Cell 4篇等

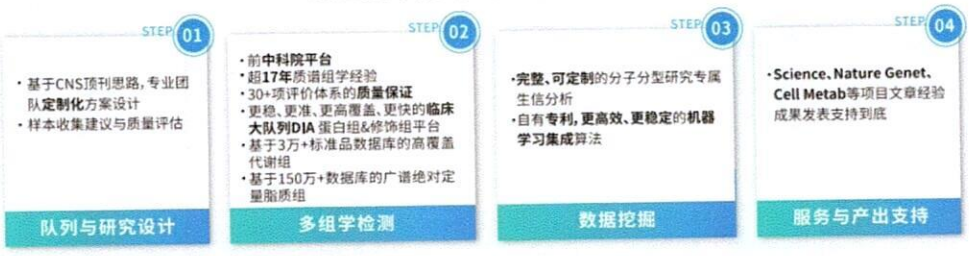
更可期的临床转化潜力

nature REVIEWS CLINICAL ONCOLOGY **Clinical potential of mass spectrometry-based proteogenomics** November 2018

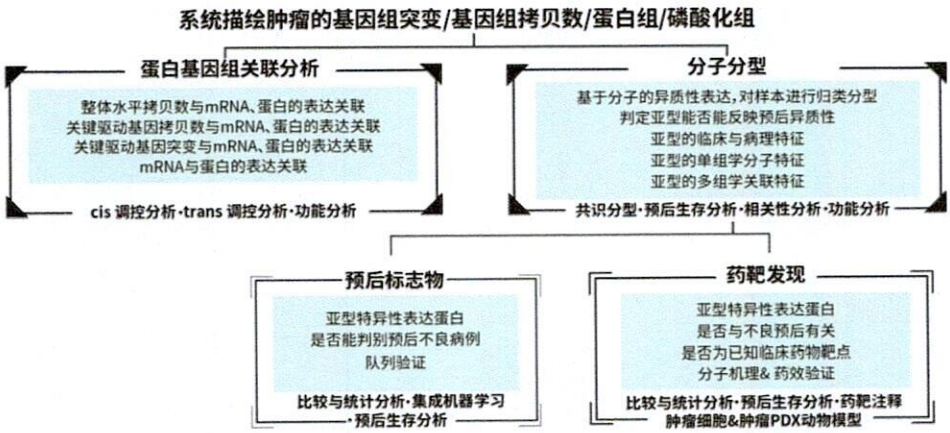
相比于单纯基因信息，蛋白质是功能分子与直接药靶

中科新生命提供基于蛋白组/蛋白基因组(Proteogenomics)的分子分型与诊疗研究的

完整解决方案与全方位支持



1）分析流程（肿瘤举例）



2）分析内容

1.1. 多维度质量控制分析

1.2. 基于无监督聚类的分子分型分析

1.3. 分子亚型的预后差异分析

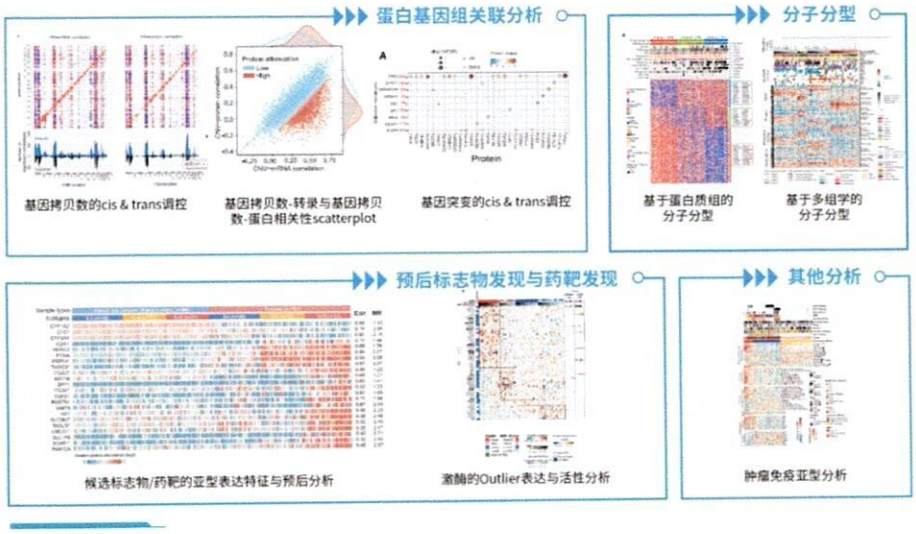
1.4. 分子亚型的病理、分子等临床关联与特征分析

1.5.分子亚型的预后标志物分析：机器学习标志物筛选，标志物临床关联分析、标志物分子关联分析

1.6. 分子亚型的药靶发现分析

1.7. 蛋白基因组分析：基因-转录-蛋白相关性分析，cis、trans调控分析

1.8分析图例展示



1.2WGCNA分析方案

1.2.1方法简介

Weighted Gene Co-Expression Network Analysis，即权重基因共表达网络分析。是一种从测序数据中挖掘模块(Module)信息的算法。Module被定义为一组具有类似表达谱的基因，如果某些基因在一个生理过程或不同组织中总是具有相类似的表达变化，那么我们有理由认为这些基因在功能上是相关的，可以把他们定义为一个模块。当基因module被定义出来后，我们可以利用这些模块搞事情了，例如跟表型和分组表型数据关联…这会大大降低问题的复杂程度。

加权共表达网络分析用于分析多样本代谢物/基因的表达模式，通过将表达模式相似的蛋白/基因聚类分析，可鉴定出高度协同变化的代谢和基因模块（module），该模块中的代谢物/基因可能协同参与某一生物过程并发挥重要作用；并根据代谢物/基因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。

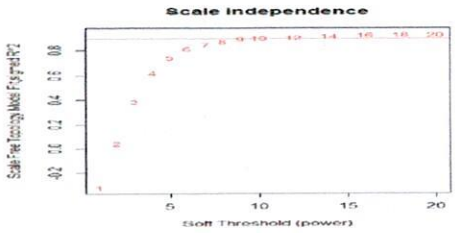
1.2.2分析方法

Step 1，计算基因间的相关性

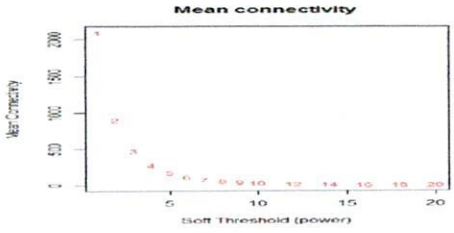
① 计算基因间两两相关性；

② 无尺度化/权重（拉大贫富差距）：对每个相关性系数取了β次幂。例，0.999取β次幂，影响较少，而0.01取β次幂，影响很大。这就实现了无尺度网络（网络中只有少数基因是核心基因），非常巧妙！

Step 2，把基因划分为模块



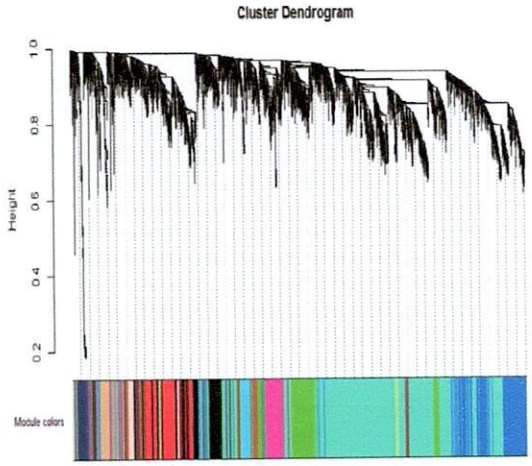
不同β值下，m（节点连接数）与 n（节点连接数为 m 的基因的数量）的相关性的变（一般取 R>0.8 或达到平台期的最小的β



不同β值下，所有基因连通性的均值（很明显，幂指

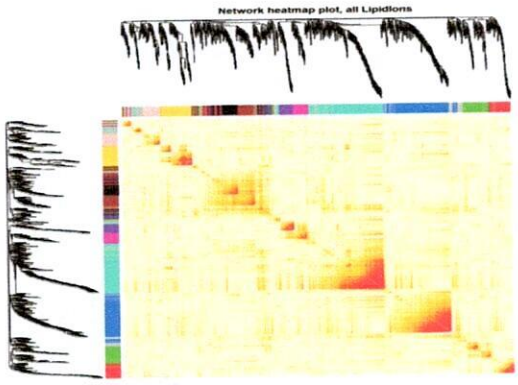
Step3：

利用TOM值构建聚类树—



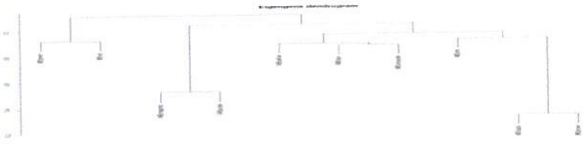
Step4:区分模块—

定一个指标，将表达模式相近的归为一类。那每个模块用一种颜色表示。灰色模块（grey），表示无法类的模块。（表达量变化不明显的基因；也与模块最少基因数有关）



Step5：

合并相似模块—将上述聚类树属于同一分支且距离很近的模块合并。



1.3集成机器学习分析

1.3.1分析方法简介

高通量组学数据的分析，通常会面临两个难题：存在大量的噪音信息；指标数量往往远大于样本数量。因此，从组学的海量数据中，识别出能够区分不同样本的特征信息，进而筛选出更具能力的潜在生物标志物，是利用组学工具进行标志物发现所面临的主要挑战之一。现有的单维统计学检验方法（如 T 检验、非参数检验等）：（1）通常单次检验只能分析一个生物分子与组别的关系，不能反映多个分子之间或其他变量因素对组别分类的影响，选出的潜在标志物之间很可能存在高度相关性，导致其分析结果的准确度和优化程度较低；（2）统计检验法没有分类预测的 功能，所以统计检验法的结果只适用本次实验样本，而不能用于预测新的样本分类。相比而言，机器学习法不仅适用于多维变量分析，可筛选出与组别相关性高且分子间相关性低的靶分子，而且还可构建具有预测新样本组别功能的模型。因此，单维统计学检验方法选择生物标记物受到方法本身的限制，对数据的信息挖掘能力有限，其结合机器学习技术能够更加高效地帮助我们实现上述目标。

在机器学习法中，特征选择算法被广泛应用于选择潜在的靶分子作为生物标志物来区分实验组和对照组的样本。目前主流的特征选择算法有过滤法（Filter）、包装法（Wrapper）、嵌入法（Embedded）等。然而，单一的特征选择方法仍存在一些弊端和问题。单一特征选择方法目的是选择特征变量最小的子集构造预测精度最高的分类器，但在算法设计中往往忽略了稳定性。并且，在实际数据中可能存在多组潜在的标志物集合，如果这些标志物具有高度相关性，可能在不同的设置下选择不同的标志物。另一方面，高维数据中样本量较少，在基因表达数据和蛋白质组学数据分析中，通常只有数百个样本，但有数千个物质特征。研究表明，高维数据中相对较少的样本是单一特征选择结果不稳定的主要来源之一。[7] 所以通过单一

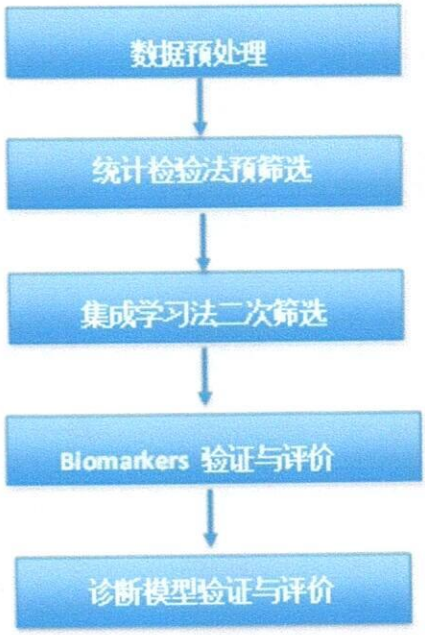
特征选择方法产生的最终的模型结果往往表现不够理想，导致从单个实验数据集里选泽出的生物标志物稳定性差，实际应用能力低，很难应用到其他同类型的实验样本上。集成的学习法选择生物标记物通过整合使用多个相同或者不同类型的机器学习算法，选出那些经常出现在具有高准确率的分类模型中的潜在靶分析，从而最大限度地提高生物标志物的稳定性 。

总的来说，集成学习具有比常规单维统计学、单个特征选择算法具有更优越的性能和表

现。本分析使用的是一种基于集成学习的特征选择技术，其整合了统计检验和目前主流使用的多种特征选择算法。该技术可以选择出一组稳健的诊断物质（如蛋白质、代谢物等），构建出效能更优的诊断标志物 panel 模型。

1.3.2分析流程本分析主要包括：数据预处理、预筛选、二次筛选、候选标志物评估与验证、诊

断 panel 模型构建与验证多个阶段，参考如下



1.3.3样本与数据信息

本次分析的样本组别、数据等信息如下：

表 1 分析信息表

No	组别名称	组内样本数量
1	根据客户要求	根据客户要求
2	***	***

1.3.3结果总结

许多不同的机器学习特征选择方法可以有效地用于生物标志物的识别鉴定。然而，每种方法都有优缺点，并且由于特征的多因素性质，使用不同的特征选择方法产生的生物标志物经常是不一致[4]。不同样本之间的个体差异变化，以及由于实验程序、机器和实验室之间的差异，也往往导致从一个数据集中选择出的特征，很难对另一个数据集进行有效分类。因此，我们利用了一个集成框架，整合多种统计和机器学习技术的多步骤数据挖掘的集成学习方法。该集成算法可识别和鉴定稳健且精确的生物标志物，适用范围广，可用于各种组学数据集构建生物标志物模型。同时，这种方法的筛选出的生物标志物分类表现好，稳定性强且数量少，应用于区分其他同类样本组别的能力佳。本实验通过对宏基因组/代谢组的数据进行上述分析，最终筛选出了关键的生物标志物。

1.4 大队列生信分析团队负责人介绍

4 DIA 蛋白组学质控

采集前仪器状态评价：293T细胞进行鉴能力测试，Astral DIA上机8分钟蛋白鉴定数量>8000。

4.1 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的质量。

4.1.1 QC 样本 CV 值分析

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中值小于20%，说明该项目检测过程中波动较小，体系较为稳定。

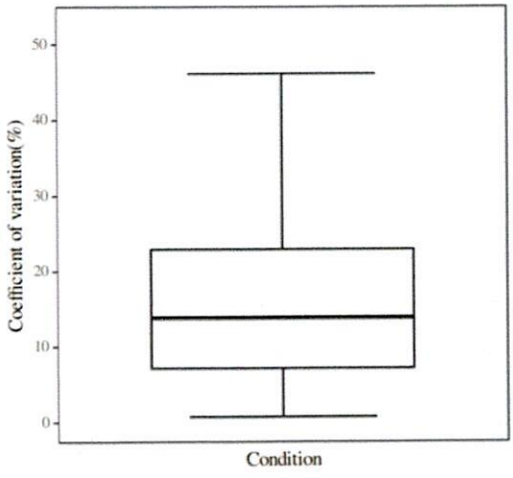


图1 QC的CV值分布箱线图

注：横坐标：QC样本；纵坐标：CV值在样本中的分布区间。一般QC的CV20%以内的蛋白比例越高越好

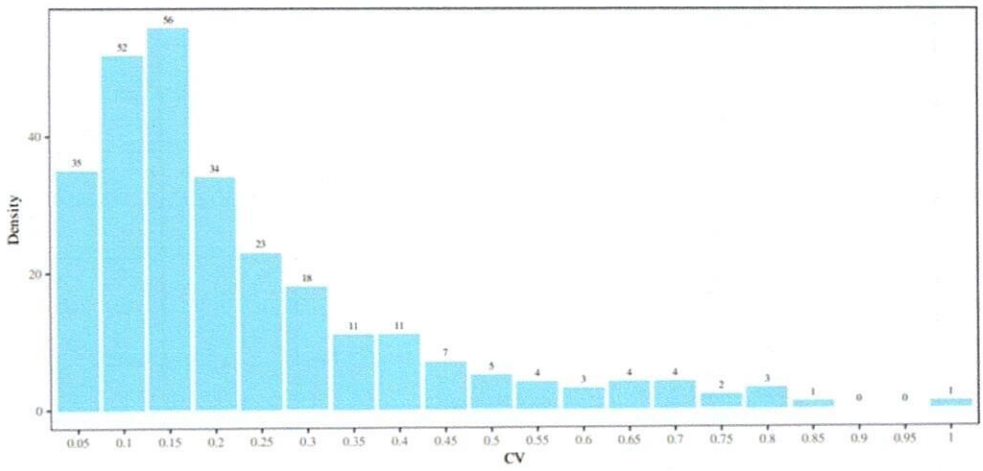


图2 QC的CV值分布图

注：横坐标：QC的CV值范围；纵坐标：各QC样本的变异系数CV区间的数量。一般QC的CV20%以内的蛋白比例越高越好

4.1.2 QC 样本 PCA 分析

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，则距离越远。QC样本的PCA聚集程度，反映了该项目采集过程中的稳定性，QC样本一般聚集在一起，说明样本采集过程中系统稳定。

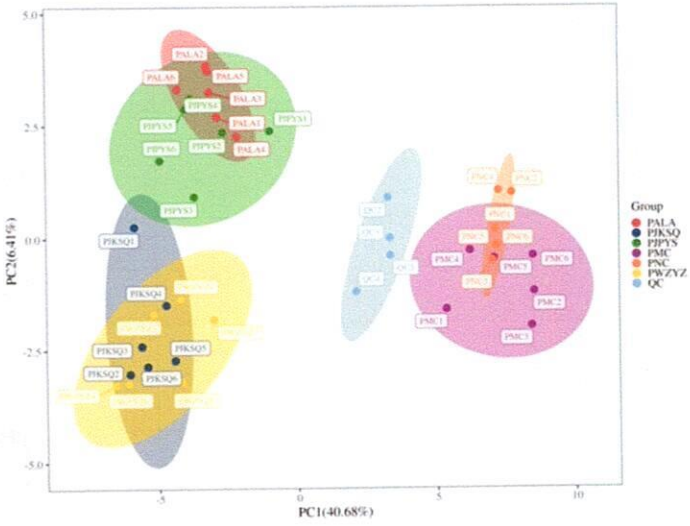


图3 QC样本的 2D-PCA图

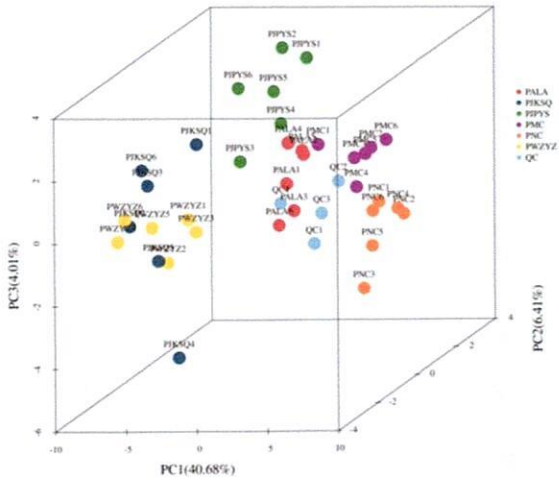


图4 QC样本的3D PCA 图

注：PC1表示第一主成分；PC2表示第二主成分；PC3表示第三主成分，每个点代表一个样本。不同颜色代表不同组别

4.1.3 QC 样本相关性分析

QC样本的强度相关性反映了项目检测过程中的稳定性，一般QC的相关性达到0.9以上，说明体系较为稳定。

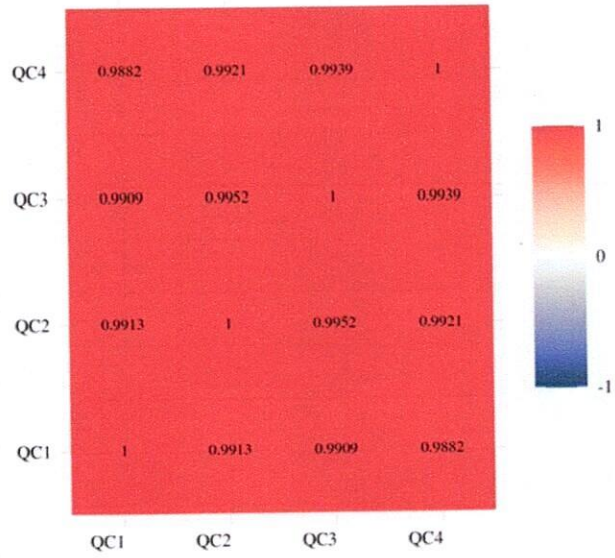


图5 QC样本相关性分析图

注：横坐标和纵坐标分别标记强度值的对数值

输出文件：3_1_QC_Sample_Evaluation\QC_Correlation_Heatmap

4.2 组间样本鉴定数量评价

4.2.1 组间样本鉴定数量评价

对不同组别之间的样本进行鉴定数量统计，反映组间样本鉴定数量是否有明显的差别，以箱线图形式进行展示。

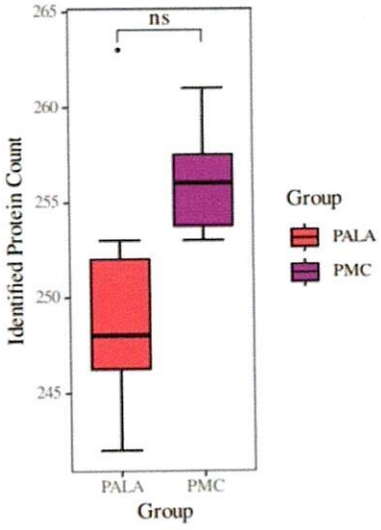


图6 样本蛋白鉴定数量箱线分布图

注: 横坐标代表组别，纵坐标表示蛋白鉴定数量。每个点表示一个样本，红色和蓝色分别代表不同的组别。P值反映了两组差异大小，如果组间没有显著性差异，用ns表示，一般*说明两组间有明显的差别

4.2.2 样本定量波动评价

从定量角度直观展示不同组样本以及样本内蛋白的定量值高低分布，从而观测是否有明显离群样本存在，本项目所有样本及QC样本的强度分布以箱线图形式展示如下。

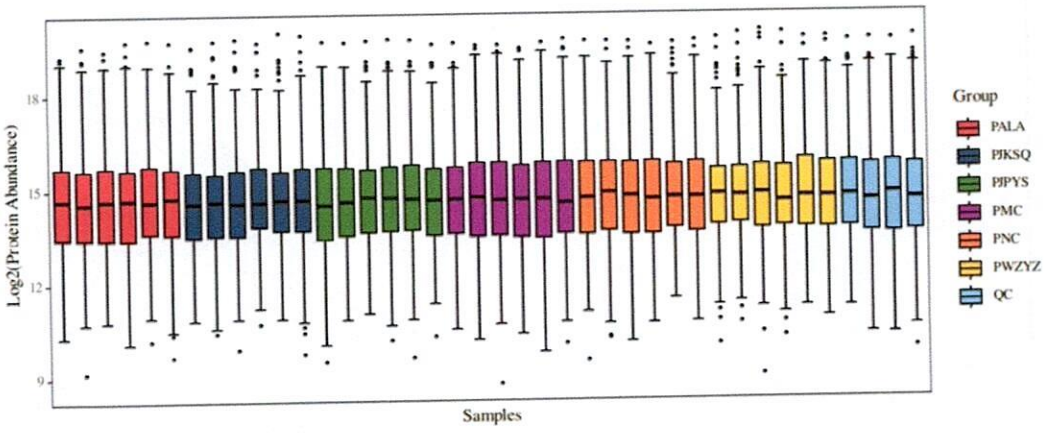


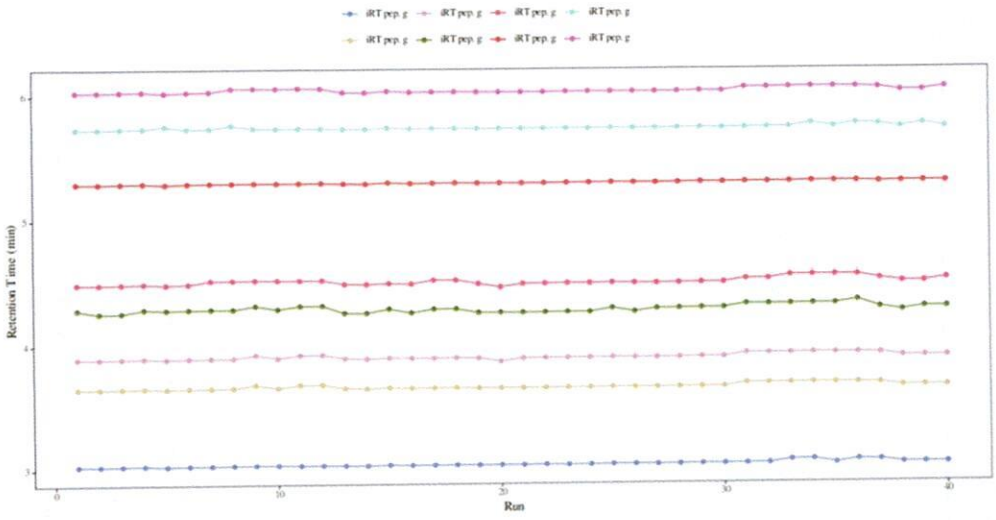
图7 样本定量强度抖动分布图

注：横坐标代表不同样本，纵坐标表示蛋白表达量（以2为底的对数）。不同颜色分别代表不同的组别

4.3 检测体系评价

4.3.1 iRT 保留时间

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[1]。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：



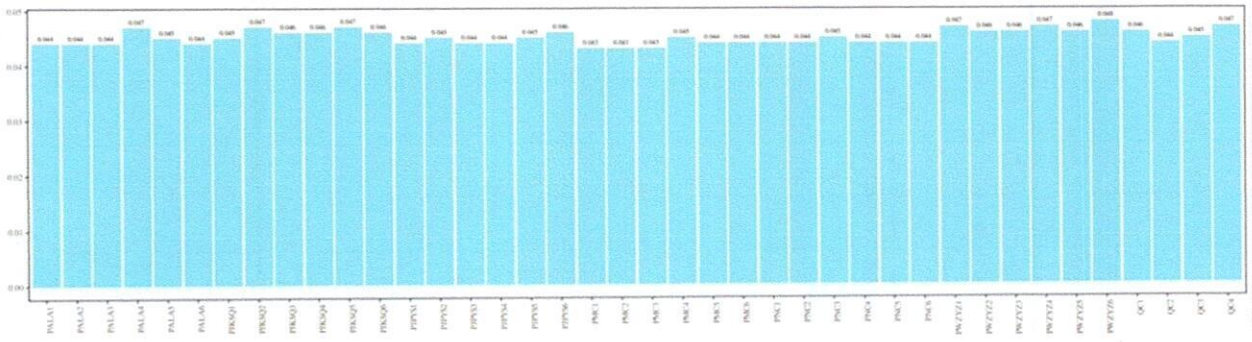
[iRT_Peptide_RT.png]

图8 iRT肽段（校正肽段）的洗脱时间图

注：横坐标为样本上机顺序；纵坐标为保留时间

4.3.2 平均半峰全宽

半峰全宽（FWHM）指色谱峰在半高度处的峰宽，反映了液相的峰分离情况。一般认为，所有样本的平均半峰全宽应无太大波动，说明体系稳定。在本项目中，平均半峰全宽波动满足要求，具体结果如下：



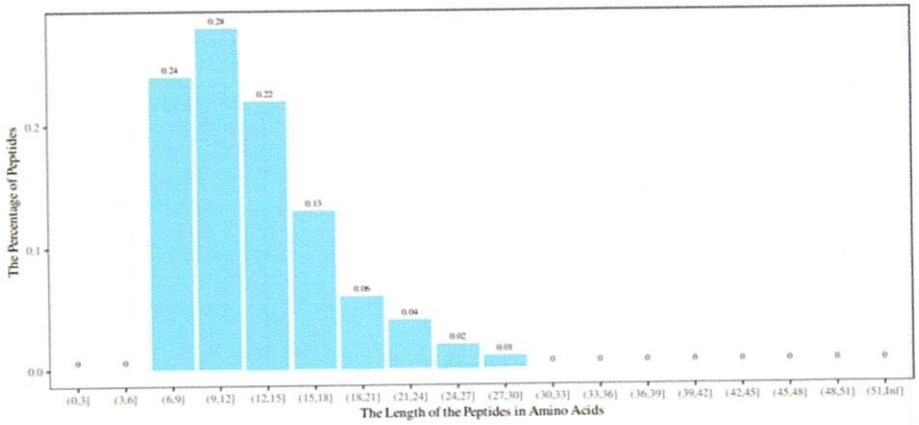
[FWHM.png]

图9 平均半峰全宽

注：横坐标为样本；纵坐标为平均半峰全宽

4.3.3 肽段长度分布

质谱仪的母离子扫描范围有限制，合适长度的肽段才能被质谱仪检测到。一般认为常规样本的酶切肽段长度主要分布在7-25个氨基酸。在本项目中，肽段长度分布满足要求，具体结果如下:



[Peptide_Length_Distribution.png]

图10 肽段序列长度分布图

注：横坐标为鉴定到的肽段序列的氨基酸长度区间；纵坐标为肽段长度区间内包含的肽段数占总肽段数的百分比

5.附件 Appendix

5. 1 数据库介绍

5.1.1 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。

5.1.2 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。

5.1.3 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。

5.1.4 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。

5.1.5 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。

5.1.6 Reactome 数据库

Reactome是一个开源、开放获取、手动管理和同行评议的通路数据库。目标是为通路知识的可视化、解释和分析提供直观的生物信息学工具，以支持基础和临床研究、基因组分析、建模、系统生物学和教育。包含信号和代谢分子及其组织成生物途径和过程的关系。Reactome数据模型的核心单元是反应。参与反应的实体（核酸、蛋白质、复合物、疫苗、抗癌治疗剂和小分子）形成生物相互作用网络，并被分组为通路。Reactome中的生物学途径的例子包括经典的中间代谢、信号传导、转录调控、细胞凋亡和疾病

5.1.7 WikiPathways 数据库

WikiPathways是一个开放式协同平台，用于数据可视化和分析所用的生物学模型的收集和传播，为了促进生物学界对通路信息的贡献和维护而建立的。WikiPathways是一个开放的致力于管理生物通路数据库，针对每一个物种，收录其特定的pathway信息。该数据库收录了超过20个物种的通路，其中人类的通路就包含了800多个通路，涵盖了7500种基因。此外，它还包含了超过1000个代谢产物的通路。

WikiPathways中的每一个通路都有自己专属的页面，呈现该通路最新的图解、描述、引用、下载选项、版本历史以及成分基因和蛋白质列表。因此，WikiPathways为生物途径数据库提供了一种新模型，可增强和补充现有通路数据库信息，例如KEGG、Reactome等。

5.1.8 ROC分析

R软件包(1.17.0.1版)用于ROC曲线分析。pROC是一个显示、平滑和比较ROC曲线的工具。曲线下的AUC可以通过统计检验进行比较，计算曲线的置信区间。

附件三　售后服务

1 服务内容

1.1提供专业的学术顾问咨询给出文章投稿建议，并提供文章杂志要求的分辨率和大小合格的图表；

1.2对审稿人的修回意见提供专业答复供参考；

1.3定制化信息分析：完成项目方案内的定制化信息分析内容

1.4 定制化绘图：在项目合作期内，为合作伙伴发表项目产出论文提供方案内的高端定制绘图服务

1.5应急解决方案：公司设立技术支持领导小组保证突发事件发生时,能够迅速召集技术人员,立即制定应急技术方案对一般性技术故障,如果检测意外失败、实验意外失败（未达到实验结果的质控标准），可以评估后安排立即复测。不另计费用。

1.6 其他售后服务：思路拓展、投稿建议、生信云平台等。

2 培训与维护

2.1根据需求提供技术培训及咨询服务；对DIA蛋白组学进行培训。

2.2技术咨询与指导：为招标方项目参与人员提供技术咨询与指导，包括质谱原理，实验操作，数据处理以及其他与本项目相关的技术指导

2.3 提供至少1次组学分析培训交流

3 售后服务方式

3.1电话服务，能7*24小时处理客户问题，全国范围内有相应的销售与技术，免费进行疑难问题解答。

3.2 远程连接技术支持人员，通过腾讯会议等在线会议方式，对数据分析结果进行指导.

3.3 往来信函、传真、电子邮件，解答用户在使用中碰到的各种技术问题.。

3.4 现场服务：在客户授权的情况下，针对客户已有数据结果进行分析，提供解决方案。

3.5定期汇报：项目进度以及数据分析处理进程。

4.服务响应时间

4.1我们将对用户提供全方位的售后服务,并提供最佳的服务响应时间。

4.2电话服务技术支持与服务时间为8:30-17:00,周一至周五(国家法定的休息日和节假

4.3数据验收合格后，至少提供2年的售后服务；

4.4 7*24小时处理客户问题

5.保密承诺

5.1保密信息范围

5.1.1 本条款所述"保密信息"指投标方在参与本项目过程中以任何形式获取的、或可被合理识别为具有保密性质的信息，包括但不限于：

（1）技术类信息：技术方案、设计图纸、源代码、实验数据、专利技术、工艺流程等

（2）商业类信息：商业计划、营销策略、成本构成、报价明细、供应商名录等

（3）运营类信息：客户数据库、内部研究报告、质量检测报告、生产计划等

（4）其他载体信息：包含上述内容的文档、电子数据、样品、模型、录音录像等

5.2 保密义务

5.2.1 保密期限自首次接触保密信息之日起持续至相关信息进入公知领域后五年，以较晚者为准

5.2.2 信息接收方应采取不低于自身商业秘密保护标准的措施，包括：

（1）建立分级保密制度，对核心机密信息实施物理隔离

（2）访问权限控制：仅限必要知悉的授权人员接触

（3）存储介质加密：对电子文档采用256位加密技术

（4）纸质文件管理：设置专用保密柜并建立借阅登记制度

5.3. 使用限制

5.3.1 未经信息披露方书面同意，投标方不得：

（1）向任何第三方（包括关联企业）披露保密信息

（2）将保密信息用于本项目以外的商业目的

（3）反向工程、复制或仿制基于保密信息开发的产品

（4）允许第三方查阅、摘抄、传播保密信息内容

5.4 例外情形

5.4.1 保密义务不适用于以下情况：

（1）信息接收方能证明在披露前已合法持有该信息

（2）根据法律法规要求或司法程序必须披露时（但须提前5个工作日书面通知）

（3）经国家保密行政管理部门认定已解密的信息

（4）通过公开渠道可获取的非涉密信息

5.5 违约责任

5.5.1 发生泄密事件时，违约方应：

（1）立即采取补救措施并书面报告详细情况

（2）承担不低于泄密信息评估价值三倍的违约金

（3）赔偿因此导致的所有直接和间接损失

（4）承担调查取证、诉讼仲裁等相关费用

5.6 特别约定

5.6.1 项目终止后30个工作日内，投标方应：

（1）返还所有载有保密信息的实体资料

（2）彻底销毁电子备份（需提供专业机构出具的销毁证明）

（3）保证其员工、顾问等关联人员继续履行保密义务

6.2 保密信息的知识产权始终归属信息披露方所有

5.7 法律效力

5.7.1 本条款独立于主合同存在，不因合同终止、解除或无效而失效

5.7.2 争议解决适用中华人民共和国法律，由信息披露方所在地法院管辖

5.8 承诺与保证

5.8.1 投标方确认已建立完善的保密管理体系并通过ISO/IEC 27001认证

5.8.2 保证每年对涉密人员进行不少于16学时的保密专项培训

5.8.3 承诺在合作期间及终止后五年内持续履行保密义务