iTRAQ 和磷酸化 iTRAQ 分析服务

Q&A

第一版

生效日期：2014 年 7 月 31 日

Email：info_apt@sibs.ac.cn

Tel：021-54665263

Addr：上海市漕宝路 500 号 5 号楼

目　　录

Q1． 如何解读 iTRAQ 项目的报告结果？

A1．iTRAQ 项目的报告结果除报告主体外共包括三个附件，分别为“附件 1 蛋白质定量表”，“附件 2 肽段定量表”，“附件 3 蛋白质显著性分析列表”，基于软件 **Proteome Discoverer** 分析得到的各表格相关说明如下：

- **附件 1 蛋白质定量表**

A 列为 Accession：蛋白质登录号。

B 列为 Description：蛋白质信息描述，包括蛋白质名称、物种名称等。

C 列为 Coverage：肽段覆盖率。

D 列为 Proteins：该**蛋白质组**中的蛋白质数量。

蛋白质组的含义：某些蛋白质序列的保守性或同源性非常高，unique peptides 已经无法将这些蛋白质进行进一步区分，这些蛋白质被归为一个蛋白质组，一个蛋白质组内的蛋白质也可能是数据库中同一个蛋白质的重复信息。最终鉴定结果列出的蛋白质信息是一个蛋白质组中得分最高、最可信的蛋白质。

E 列为 **Unique Peptides**：用于定量的唯一肽段数。

Unique peptides（唯一肽段）的含义：某一个蛋白质所特有的、能够将其与其他蛋白质特异性区分的肽段序列被称为这个蛋白质的 unique peptides。Unique peptides 数量越多，定量的准确度越高。

F-N（N=I+(比值个数*2-1）列为 Ratio 与 Ratio Variability[%]：归一化数据比值和比值变异性。

N+1 列为 MW[kDa]：理论分子量。

N+2 列为 calc. pI：理论等电点。

本表中的列出的蛋白质为实验鉴定到的全部蛋白质，其中包括少数无定量信息的蛋白质；

- **附件 2 肽段定量表**

A 列 Sequence：肽段氨基酸序列，小写字母为修饰氨基酸。

B 列为 Protein Group Accession：蛋白质登录号。

C 列为 Modifications：修饰氨基酸、位置及修饰方式。主要修饰有：N-Term/K(Itraq4/8plex)：N 端/赖氨酸的 iTRAQ 试剂修饰；C（Carbamidomethyl）：半胱氨酸碘乙酰化；M(Oxidation)：甲硫氨酸氧化。

D 列为 IonScore：Mascot 肽段得分。

E-N（N= E+(比值个数*2-1）列为 Ratio：比值及 Ratio count：用于比值计算的肽段数量。

N+1 列为 Charge：电荷。

N+2 列为 MH+[Da]：带一个电荷的肽段实际分子量。

N+3 列为 ΔM[ppm]：理论分子量和实验分子量的差异。

本表中的列出的肽段为实验鉴定到的全部肽段，因为蛋白质组学寻找的是蛋白质水平的差异，所以肽段的定量信息只做参考，一般不用于后续分析。

- **附件 3 蛋白质显著性分析列表**

A 列为 Accession: 蛋白质登录号。

B 列为 Description：蛋白质信息描述。

C 列为 MW[kDa]：分子量。

D 列为 calc.pI：等电点。

E 列为 Coverage：肽段覆盖率。

F 列为 Proteins：该蛋白质组中的蛋白质数量。

G 列为 Unique Peptides：唯一肽段数。

H-N（N 视实际需要而定）列为比值、平均值、P value，具体根据客户需求完成。

本表由“附件 1 蛋白质定量表”去除无定量信息的蛋白质后经数据分析（仅作数据的比值和统计显著性分析，不做差异数据筛选。差异数据需根据具体情况按照倍数和 p 值筛选，一般筛选建议筛选标准为：fold change > 1.2 and p value < 0.05）生成而来，所以该表中的蛋白质数量一般要少于附件 1 的蛋白质数量。对于数据分析方法的选择，一般视实验设计而定，如双样本（各三个技术或生物学重复）单因素分析选择 T-test 检验，多组样本单因素分析（每组包括两组以上技术或生物学重复）选择 one-way ANOVA 分析。

Q2． 如何解读 iTRAQ 磷酸化项目的蛋白鉴定列表与磷酸化肽段列表？

A2．iTRAQ 磷酸化项目的报告结果中包括两个附件，分别为“附件 1 蛋白质鉴定列表”与“附件 2 磷酸化肽段定量列表”。基于软件 **Proteome Discoverer** 分析得到的两个附件表格相关说明如下：

- **附件 1 蛋白质鉴定列表**

A 列为 Accession: 蛋白质登录号。

B 列为 Description：蛋白质信息描述，包括蛋白质名称、物种名称等。

C 列为 Coverage：肽段覆盖率。

D 列为 MW[kDa]：理论分子量。

E 列为 calc.pI：理论等电点。

质谱磷酸化定量分析的是发生磷酸化修饰的肽段的量的变化差异，从而揭示磷酸化位点修饰程度的改变。由于经过磷酸化富集后，大部分非磷酸化肽段都已去除，无法对蛋白质进行准确定量，所以只提

供磷酸化肽段的定量信息，而不提供蛋白质的定量信息。如需要对蛋白质的表达进行定量，需要另外开展蛋白质的 iTRAQ 定量分析。

另外，因为 TiO_2 对磷酸化肽段富集效率高达 95%而非 100%，故会有少量非磷酸化肽段被检测到，因而我们鉴定到的总蛋白质包括少量非磷酸化蛋白。由于软件原因，不能直接提供非磷酸化蛋白质和磷酸化蛋白质的两个独立列表。如果需要统计鉴定的磷酸化蛋白质总数目，可以将“附件 2 磷酸化肽段定量列表”中的“Protein Group Accession”列删除重复项后在 Excel 表中统计得到。

- **附件 2 磷酸化肽段定量列表**

A 列为 Sequence：肽段的氨基酸序列。

B 列为 Protein Group Accessions：蛋白质的登录号。

C 列为 Modifications：修饰氨基酸，位置及修饰方式。

D 列为 phosphoRS Isoform Probability：磷酸化肽段修饰可能性。

E 列为 pRS Site Probabilities：可能的磷酸化位点的打分， 0.75 及以上的可信度较高。

F 列为 phosphoRS Binomial Peptide Score：磷酸化肽段的得分，50 及以上的图谱质量较高。

G 列为 IonScore：肽段的 mascot 得分。

H 列为 Charge：电荷数。

I 列为 MH+ [Da]：带一个电荷的肽段的实际分子量。

J 列为 ΔM [ppm]：实测分子量与理论分子量的误差。

K-N（N 视样品组数而定）列为磷酸化肽段的定量比值及 p value 值。

同一个肽段可能有多个不同的磷酸化位点，因此可能检测到相同肽段的不同修饰形式，每种修饰形式都代表一种不同的肽段，在磷酸化肽段表中会被依次列出，被修饰的氨基酸用小写字母表示，如下图所示，展示了同一个肽段的三种不同修饰（磷酸化/氧化）状态。

	Sequence	Protein Group Accessions	Modifications	phosphoRS Isoform Probability	phosphoRS Site Probabilities	phosphoRS Binomial Peptide Score
1	aEAkEEsEEsDEDmGFGLFD	E9Q070	N-Term(iTRAQ8plex); K4(iTRAQ8plex); S7(Phospho); S10(Phospho); M14(Oxidation)	1	S(7): 100.0; S(10): 100.0	288.3973083
2	aEAkEEsEESDEDmGFGLFD	E9Q070	N-Term(iTRAQ8plex); K4(iTRAQ8plex); S7(Phospho); M14(Oxidation)	0.999986379	S(7): 100.0; S(10): 0.0	325.0328883
3	aEAkEEsEESDEDMGFGLFD	E9Q070	N-Term(iTRAQ8plex); K4(iTRAQ8plex); S7(Phospho)	0.999999723	S(7): 100.0; S(10): 0.0	343.5754254

磷酸化肽段表格的解读：首先可以通过 phosphoRS Binomial Peptide Score ≥50 判断图谱质量较好、可信度较高的磷酸化肽段，然后通过 phosphoRS Site Probability ≥75%判断磷酸化肽段中的可信磷酸化修饰位点。对于 phosphoRS Binomial Peptide Score ≥50 但是 phosphoRS Site Probability<75%的情况，说明该肽段的磷酸化存在形式可信，只是具体的修饰位点不能明确。同一个肽段中所有可能的修饰位点 phosphoRS Site Probability 相加如果是 100%，说明该肽段中仅有一个可能的修饰位点；如果相加为 200%，则说明该肽段中有两个可能的修饰位点；以此类推。

Q3． 蛋白质鉴定可信度的判断标准是什么？

A3．本公司开展的iTRAQ实验一般采用商业化软件Proteome Discoverer(Thermo Scientific)进行蛋白质定性和定量分析，蛋白质定性筛选标准为 peptide FDR≤0.01。FDR 是通过检索目标数据库(Target database) 和 Decoy 库(Decoy 库由 Proteome Discoverer 软件自动创建)后，根据得到的匹配图谱数量计算得来。设置该软件中的“High Confidence Filter Settings” 高可信度过滤参数即可得到符合 FDR≤0.01 的数据。FDR≤0.01 为公认的数据筛选标准，我们提供的报告中的数据均已经通过 FDR≤0.01 标准筛选，因此均是可信数据。

Q4． 关于蛋白质名称中的各部分含义?

A4．不同数据库有不同的命名方式及特征，现在较常用的 UniProt 数据库（www.uniprot.org）的命名格式如下：

tr|W6NF68|W6NF68_HAECO DNA RNA helicase domain containing protein OS=Haemonchus contortus GN=HCOI_01691200 PE=4 SV=1，代表的意义为 tr=TrEMBL（UniProtKB 中包含两类数据库：Swiss-Prot, which is manually annotated and reviewed；TrEMBL, which is automatically annotated and is not reviewed.），W6NF68 为蛋白质登录号，HAECO DNA RNA helicase domain containing protein 为蛋白质名称及其描述，OS=Organism Name(物种名称),GN=Gene Name(基因名称) ,PE=Protein Existence，SV=Sequence Version。

Q5． 为什么数据是以比值形式出现，而没有每个标签的原始信号强度值？

A5．iTRAQ 数据一般采用 Proteome Discoverer 软件进行查库和蛋白质定量分析，该软件在输出数据时只能输出标签比值的形式，没有原始信号强度值。我们给出的数据通常选择以所有标签通道的加和平均值为分母或以实验中设定的对照组为分母（在报告中显示为 REF）做比值之后的数据（数值范围在 1 左右）。该数据相当于经过均一化处理之后的各标签定量数据。

Q6． 在蛋白质定量数据表中，unique peptide 是不是越多越好，peptide=1 的蛋白质定量结果是否不太准确，最好将其排除？

A6．首先 unique peptide 在不同情况下有两种不同的含义：一种情况指用于定性的 unique peptide，在像 iTRAQ 这类定量的结果数据中我们并没有给出这个值（如果给出的话，则在表头中命名为 peptide）；第二种情况则是指用于定量的 unique peptide。也就是提供的附表 1 或者 3 中表头为“unique peptide”对应的值。定性的 unique peptide 的数量≥定量的 unique peptide 数量。

简单来说，肯定是检测到的蛋白质的（定性）unique peptide 越多，蛋白质可信度越高。早些年的发表文章一般会认为 unique peptide ≥2 的数据是可信的蛋白质定性数据。但是随着现在质谱仪精度和分辨率的提高，对应高精度质谱仪(如 Q-Exactive，AB5600 等等)产生的数据，像 MCP 或者 JPR 这类蛋白质组学领域最好的杂志，也明确表示认可 unique peptide=1 且对应图谱质量较高的数据，只是根据杂

志的不同要求需要提交不同格式的图谱数据。如 MCP 要求最高，需要提交质谱原始文件，查库结果文件，而且需要通过专门的软件提交（如 Pride, ProteomeXchange）；JPR 要求相对 MCP 低些，一般提供文章或附件列表中出现的所有鉴定差异蛋白质的 unique peptide=1 的肽段图谱即可。

另一方面，对于定量 unique peptide 的数量，文章一般都没有要求。可以理解的是，对于只有一段肽定量一个蛋白质的情况，蛋白质定量结果的准确性和可靠性会相对较低。通常情况下，通过设置生物学或者技术重复实验，根据蛋白质定量数据的平均值和统计学 p value 值可以获得准确度和可信度较高的蛋白质定量结果。纯粹的一次质谱行为则不能反映其定量准确性。

Q7． 附表 1（蛋白质定量表）中为什么有些蛋白质没有定量信息？在什么情况下才可能出现有定性但无定量的情况？

A7．可能的原因如下：

- 只检测到了用于定性的肽段，而没有检测到可用于定量的唯一肽段，所以无定量数据。这种蛋白质在附表一中显示的 unique peptide=0。
- 肽段没有被 iTRAQ 试剂标记上。iTRAQ 的标记效率超过 98%（iTRAQ 试剂技术报告，www.appliedbiosystems.com），保证了实验的可信度，但并未到达 100%，所以鉴定到几千个蛋白质中可能会有少数蛋白质肽段没有标记上。这种蛋白质一般表现为所有标记通道中只有部分通道没有定量信息。
- 肽段被标记上而且被检测到了，但是由于信号值太低，导致信号强度积分数据赋值为零。

Q8． 筛选差异蛋白质的基本原则是什么？

A8．同时满足统计学分析和差异倍数过滤标准，如 fold change> 1.2 以及显著性分析 p value <0.05。如果这样筛选的差异数据非常多，则可提高筛选标准，如 fold change>1.5 甚至>2 或者 p value <0.01 等。参考文献：

Moulder R, Lonnberg T, *et al*. Quantitative proteomics analysis of the nuclear fraction of human $CD4^+$ cells in the early phases of IL-4-induced Th2 differentiation. Mol Cell Proteomics. 2010; 9(9): 1937-1953. **（fold change >1.2 and p value < 0.05）**

Gan CS, Chong PK, *et al*. Technical, experimental, and biological variations in isobaric tags for relative and absolute quantitation (iTRAQ). J Proteome Res. 2007; 6(2): 821-827.

Q9． 蛋白质鉴定数量较少的可能原因是什么？

A9．数据库质量不好通常是鉴定数量较少的主要原因。如数据库不好（通常是数据库收录的蛋白质序列较少），可考虑换大一级的物种分类数据库，或者选择在进化树上同源性较近的、研究比较清楚的、基因组数据库相对完整的模式物种的数据库重新查库。

- 根据样品 SDS-PAGE 图，判断样本本身的蛋白质条带丰富程度，是否样本蛋白质条带较少。
- 根据 SDS-PAGE 图判断样品中是否存在高丰度蛋白质，高丰度蛋白质会影响样本整体蛋白质鉴定数量。
- 根据 TIC、Basepeak 图等，判断酶解样品、质谱仪器状态等是否正常，但这些都是有质控严格把关的，我们提供的数据都是符合质控标准的。

Q10． 差异蛋白质数量较少的原因是什么？

A10． 样本组别之间的差异蛋白质（fold change > 1.2）数量多少主要与样本本身的生物学差异大小有关。一般如正常组织和病理组织之间的样本差异比较显著，得到的差异蛋白质数量也较多（一般差异蛋白质数量占蛋白质鉴定总数的 10%左右）。然而，病理组和药物处理组之间的样本差异则不一定显著，如药物是否有效、药物处理浓度是否合适、取样时间是否恰当等因素都可能导致差异蛋白质数量较少。

如果仅按照 fold change>1.2 筛选得到蛋白质数量并不少，但是通过 p value <0.05 再做进一步筛选后得到的差异蛋白质数量显著减少，则主要是由组内样本的生物学个体变异较大造成的。该情况下可考虑剔除组内样本重复检测数据中波动较大的个别数据。

Q11． 通过 western blot 检测到的样本中的蛋白质或者蛋白质的磷酸化形式，为什么在 iTRAQ 或者磷酸化 iTRAQ 实验中没有检测到？

A11． western blot 检测是将目的蛋白质信号放大很多级之后进行检测的，灵敏度很高，(除非特异性结合之外)几乎不受复杂样品中背景蛋白质丰度的影响。对于纯蛋白质样品，质谱检测的灵敏度达 fmol 水平。但在复杂样品中，不同种类的蛋白质浓度动态范围较宽，而目前质谱检测的动态范围一般在 3-4 个数量级。样品中丰度较高的蛋白质优先、多次被检测到，而丰度较低的蛋白质则因为肽段信号过弱被淹没而不能被检测到。因此，如果样品中待检测的目标蛋白质丰度较低(或者蛋白质分子量较小)，即使 WB 能够检测到，质谱并不一定能检测到。

Q12． 什么是二级质谱图？什么是定量图谱？什么情况下需要用到这些图谱？

A12． 通常讲的二级质谱图包括两种，一种为母离子被碎裂后实际被检测的原始图谱，如图 1 所示，代表质荷比为 1140.14 的母离子的二级图谱，文章一般不要求提供此类图谱；

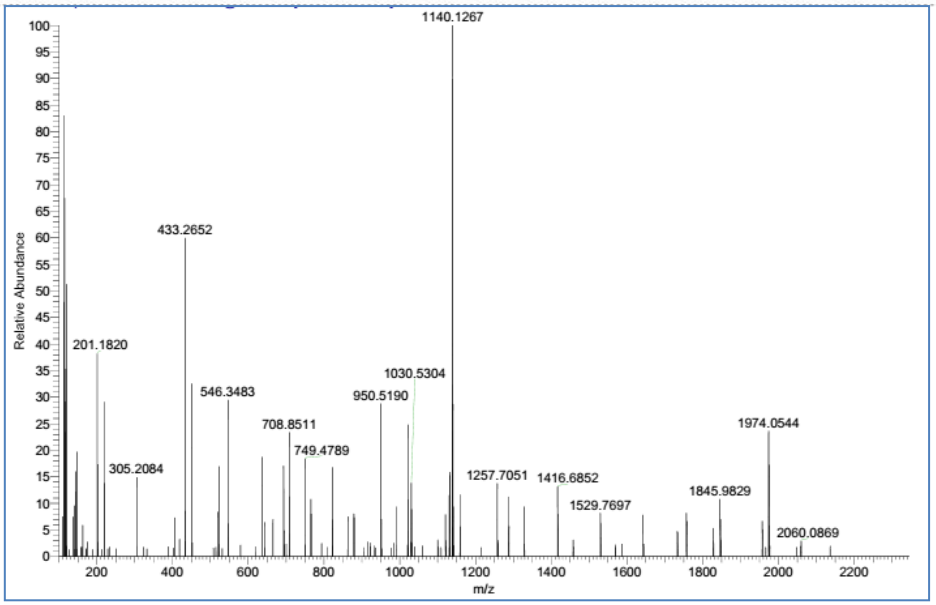

图 1. 二级质谱图

另外一种二级质谱图是 b,y 离子匹配图，是肽段按照理论方式碎裂与原始的二级质谱图匹配后生成的图谱，匹配上的峰用彩色表示，被标注为 b 离子或 y 离子，如图 2 所示。该图为定性图谱，可以证明特定肽段的存在，一般文章所要求的二级质谱图即为 b,y 离子匹配图。

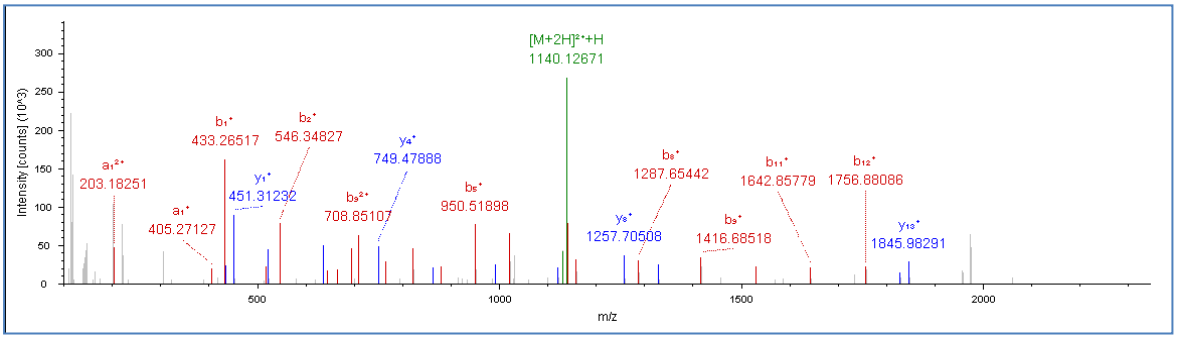

图 2. b, y 离子匹配图

二级定量图谱是标记标签 113,114,115……121（四标实验标记通道为 114,115,116,117）在对应的标记通道中的信号强度信息，可以认为是该肽段在不同标记样品中的相对表达量，如图 3 所示，它是图 1 中 113-121 质荷比范围的截图。例如有七组样品（A 组-G 组），标记通道分别为 113-119，即 A-113，B-114，……，G-119，该图谱中 113-119 离子报告峰的高低则代表特定肽段在 A 组-G 组共七组样品中的原始相对表达量。由于在进行数据定量分析时，需要进行数据归一化处理，以矫正各标记样本的上样量。所以最终输出的肽段定量表中的定量信息为归一化之后的数据，与图谱上反应的标签原始强度相对表达量不完全一致的，两者的趋势是基本符合的（偶尔也会出现趋势不一致的情况）。另外，蛋白质定量信息是综合所有唯一肽段定量值计算后得到的结果（蛋白质定量数据=所有唯一肽段定量值的中位数），而该图谱只是其中一段肽的定量信息，所以拥有两段及以上 unique peptide 的蛋白质定量信息与图谱中反应的定量信息的相关性不大。综合以上原因，该定量图谱对于投稿文章不是必需的，

也不需要提供。

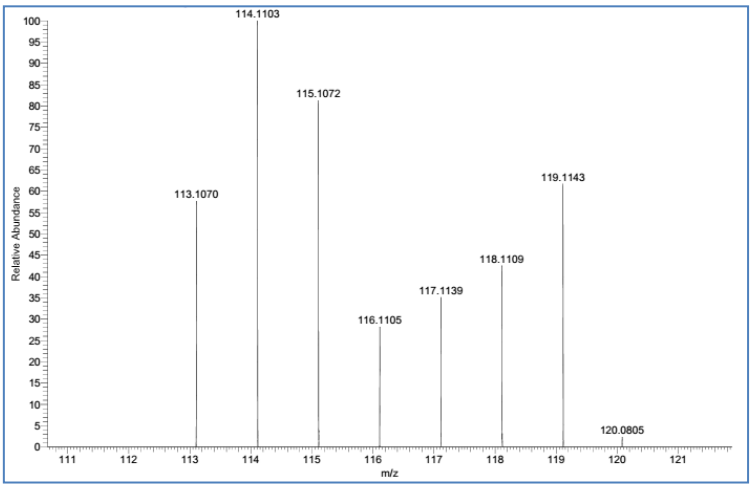

图 3. 肽段定量图

Q13． 在 UniProt 官网上发现鉴定到的蛋白质被标注 Obsolete 或 Replaced，为什么？

A13． UniPort 数据库是动态更新的，其中存在的一些重复、冗余、有误的等数据会被不断地纠正和删除。标注为 Obsolete 的蛋白质，一般是通过测序结果直接翻译而来，先存放在 TrEMBL 里，注释信息不全，甚至这个蛋白是否存在也不确定。经过人工校验和注解后，一部分不存在的蛋白质会被删除，状态显示为 Obsolete（如 F1MYU3 这个蛋白已经从数据库中删除了）；一部分经过注释后会被移到 SwisspProt，或者这个蛋白质已经在数据库中存在了，就和已存在的合并，显示为 Replaced。这些蛋白质可以通过对应网页中的“history…”追踪到数据库中该蛋白质信息的更新过程。

Q14． 检测结果中的蛋白质仅有登陆号码，没有蛋白质名称，没有蛋白质功能注释（通常描述为“uncharacterized protein”或“predicted protein”），也没有基因名称，为什么？这些蛋白质应如何进行后续分析？

A14． 对于研究较少、基因组或者蛋白质组数据库不完整的物种，这类蛋白质的出现频率很高。因为这类蛋白质序列通常只是从基因序列或者转录测序结果翻译而来，其功能尚未被研究过，所以在数据库注释为功能未知的蛋白质。客户可以通过 Blast 比对进行未知功能蛋白质的序列分析从而推测其功能，然后进行实验验证。

Q15． 差异表达蛋白质有无必要做进一步的验证？应如何挑？验证方法有哪些？

A15． 大规模的组学数据需要后期具体的验证实验，以支持组学数据和结论的可信度、说服力，也可以进一步提供文章发表的档次。如转录组数据需要进行 qPCR 的验证，蛋白质组学的数据则需要进行 WB 的验证。无法获得抗体的情况下，有时也可以采用 qPCR 进行间接验证。需要特别指出的是：qPCR 分析的是转录水平，不能代表蛋白质的表达水平；与蛋白质组学数据不符时，不能说明任何问题。一

般对差异蛋白质的验证数量并没有严格限制，通常可通过以下几种途径筛选用于验证的差异蛋白质：

- 来源于文献报道的，已知与该研究课题密切相关的蛋白质；
- 蛋白质组学定量差异倍数较大的数据；
- 结合功能分析或者通路分析（生物信息学的分析范畴），筛选得到的有意义的蛋白质。

Q16． 肽段 aDIQDEDGsLWGGsDEELDk 和 aDIQDEDGsLWGGsDEELDkTSDLDR 只是相差几个氨基酸，都是对应同一个蛋白，这种情况是不是酶解不充分？

A16． 蛋白质酶解过程中，可能因为理论酶切位点发生修饰等情况，导致蛋白酶漏切的情况出现。因此，为了保证存在漏切位点的肽段在数据库检索过程中也能匹配到，检索参数会设置最大漏切位点数为 2（报告参数中体现为 miss cleavage 2）。所以在肽段列表中会出现两种肽段序列相似，某一种肽段存在 K 或者 R 的漏切位点的情况。这种情况属于质谱检测的正常现象，在发表文章中写清楚查库参数就可以了。