

Proteograph™ 血浆蛋白质组高深度检测分析

(DEMO)

目录

一、简介	2
二、样本信息:	2
三、检测流程:	2
(一) 仪器、试剂与耗材:	2
(二) Proteograph™ 血浆蛋白质组富集及前处理:	3
(三) 质谱检测与蛋白质鉴定/定量:	3
(四) 数据统计与分析:	4
四、结果与分析:	4
(一) Proteograph™ 血浆蛋白质样品制备质控:	4
(二) 质谱检测与蛋白质鉴定:	4
(三) 蛋白质鉴定数据汇总分析:	4
(四) 样本总体质量评估分析:	6
(五) 鉴定蛋白质的动态分布范围分析:	6
(六) 组间差异表达蛋白质筛选分析:	8
(七) 差异表达蛋白质 KEGG 和 GO 分析:	9
五、总结:	9
六、参考文献:	10

Proteograph™ 血浆蛋白质组检测分析

一、简介

血液是人体最大的体液库，血液中的血浆组分 (Plasma) 中包含了来自人体各个组织和器官的信息分子(DNA、RNA、蛋白质和代谢物等)。因蛋白质分子是生命体的功能分子和执行者，血浆蛋白质组(Plasma Proteome) 的动态变化直接指针人体生理和病理的变化，使得从血浆中筛选与人体生理和病理相关联的蛋白质分子标志物 (Biomarker) 成为长久以来的研究热点[1, 2]。

然而，由于血浆蛋白质组本身的独特性：(1) 蛋白质种类超过 **1万**种，(2) 高丰度占比极高，**22种**高丰度蛋白占据血浆总蛋白质量的**99%**，(3) 蛋白质组的浓度动态分布范围跨越**10~12个数量级**，导致几十年来对血浆蛋白质组的解析几乎没有明显的进步。即使借助成熟的质谱技术，也仅能从血浆中鉴定到 500-800 种蛋白质，覆盖度不到血浆蛋白质组的 5%，“看”不到更多、更低丰度的蛋白，蛋白质 biomarker 的筛选、发现和转化应用都无从谈起[3]。

2020 年，美国 **Seer** 公司在《Nature Communications》发表原创技术文献，利用全新设计的纳米磁珠组合结合质谱技术实现对血浆蛋白质组深度和宽度的双重高深度覆盖，尤其是对血浆低丰度蛋白质的富集和检出，重启了血浆蛋白质分子标志物的筛选和发现的新征程[4]。基于其原创技术，Seer 公司推出了自动化的 Proteograph™ 系统，可快速实现大样本队列血浆蛋白质组的高深度分析。

本 DEMO 项目利用目前最新的 Proteograph™ 系统和技术，对 6 例样本进行高深度血浆蛋白质探索，考察 Proteograph™ 系统对血浆蛋白质组的解析**深度和宽度**。

二、样本信息：

Group H 和 Group D 两组样本，每组 3 例血浆样本：

样品组	样品编号
Group H	H1
	H2
	H4
Group D	D4
	D7
	D8

三、检测流程：

(一) 仪器、试剂与耗材：

序号	名称	生产商	应用
1	SP100 Automation Station	Seer	血浆蛋白质自动化富集及处理工作站
2	Proteograph™ Assay Kit	Seer	血浆蛋白质富集及肽段制备

3	Proteograph™ PQR Kit	Seer	肽段定量及重悬
4	酶标仪	BioTek	肽段定量
5	低温冻干浓缩仪	Labconco	肽段冻干

(二) Proteograph™ 血浆蛋白质组富集及前处理:

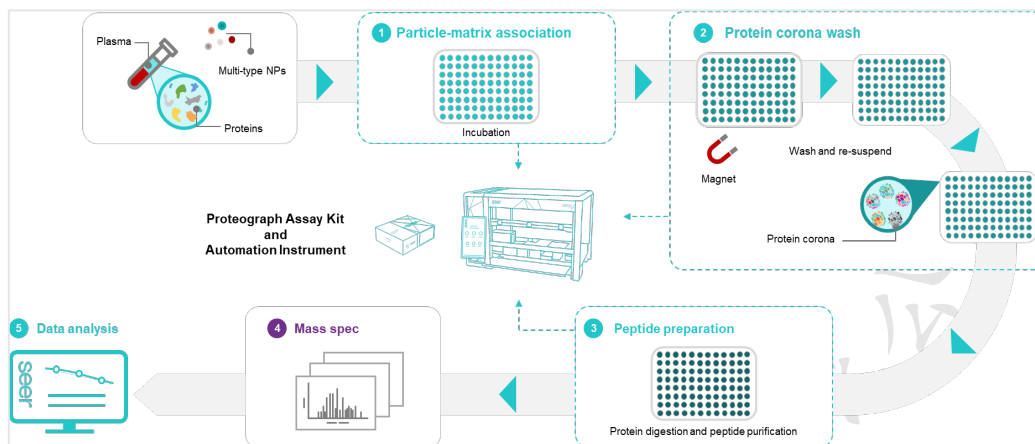


图 1. Proteograph™ 血浆蛋白质组高深度自动化分析流程图

1. 样品前处理使用 Proteograph™ Assay Kit (Seer, Inc.) 在 SP100 Automation Station 自动工作站 (Seer, Inc.) 中进行[4, 5]。取 1 个 Proteograph™ Assay Kit，按照说明书，将 16 例样本 (**Group H 和 Group D 的 6 个样品采用随机排列方式，与其他 12 个样本同时进行实验**)、5 种纳米磁珠 (Nanoparticle, NPs) 和其他相应的试剂/耗材放入 SP100 高深度血浆蛋白质组自动工作站。
2. 打开 SP100 自动工作站的电脑控制程序，开启 Proteograph Assay 自动化流程，无需人工值守，可在 **7 小时**内，完成 NPs 对血浆蛋白质富集、蛋白质酶解、肽段除盐和纯化步骤。最终获得 80 个肽段样品+3 个质控（收集在 96 孔的样品收集板中）。
3. 退出 Proteograph Assay 程序，打开 Peptide Quantitation 程序，将样品收集板（80 个肽段样品+3 个质控）、对应的肽段定量检测试剂（Proteograph™ PQR Kit (Seer, Inc.) 和耗材放入 SP100 工作站中后启动程序，约 30 分钟后，在 96 孔板中完成肽段标准品（标准曲线）和样本（80 个肽段样品+3 个质控）的反应制备。
4. 将步骤 3 的 96 孔板，放入荧光酶标仪（激发波：390nm，发射波：475nm），测定并计算每个孔中的肽段浓度和总量。

(三) 质谱检测与蛋白质鉴定/定量:

1. 因 Group H 和 Group D 的 6 个样品分别与 5 种 NPs 孵育，最终共计产生 $6 \times 5 = 30$ 个肽段样品，分别进行质谱检测。
2. NP 富集的肽段用 Evosep One (Evosep Biosystems) 液相系统串联 TimsTOF Pro (Bruker) 质谱系统进行检测。Evosep One (Evosep Biosystems) 分离时间设定为 44 分钟。质谱采用 DIA-PASEF 模式，参数设置为：扫描范围为 100 至 1700m/z；对于一个 100 毫秒的 TIMS 扫描，按照粒子迁移率（ion mobility）设定为 8 个窗口；离子迁移率（1/K0）范围设为 0.6-1.6。
3. 利用 DIA-NN (version 1.8.1) [6] 软件对 TimsTOF Pro 采集的数据进行检索，检索参数设定为：
 - (1) library-free 模式。(2) 检索蛋白质组数据库为人的蛋白质组(UniprotKB ID: UP000005640, 79052 蛋白质序列)。(3) 固定修饰设定为半胱氨酸（C）烷基化。(4) 可变修饰设定为甲硫

氨酸 (M) 氧化和 N-term M excision。

(四) 数据统计与分析:

1. 通过 DIA-NN (version 1.8.1) 对质谱原始文件进行检索, 完成蛋白质的鉴定和相对定量 (LFQ Intensity)。通过 label-free 相对定量方式, 实现对组间 (Group D vs. Group H) 差异表达蛋白质的筛选[7]。
2. 数据统计及差异表达蛋白筛选使用 R script (<https://www.r-project.org/>) 进行, 统计方法使用 T-test, 差异表达蛋白质 (Differentially Expressed Proteins, DEPs) 筛选卡值 (cutoff) 设定为: $p\text{-value} < 0.05$, $\log_2\text{FC} = |\pm 1|$ 。

四、结果与分析:

(一) Proteograph™ 血浆蛋白质样品制备质控:

1. Group H 和 Group D 共 6 例样本, 通过 Proteograph™ 系统约 7 个小时的自动化流程, 实现从血浆原液到质谱待检肽段的制备, 最终获得 $6 \times 5 = 30$ 个除盐的肽段样品。
2. 肽段定量数据 (Proteograph™ PQR Kit) 显示每例样本的肽段总量符合质控参考范围, 样品合格, 送交质谱检测。

(二) 质谱检测与蛋白质鉴定:

1. 对 30 个肽段样品, 分别用 Evosep One 串联 TimsTOF Pro 质谱进行检测, 共获得 30 个质谱文件。代表性的质谱文件见图 2:

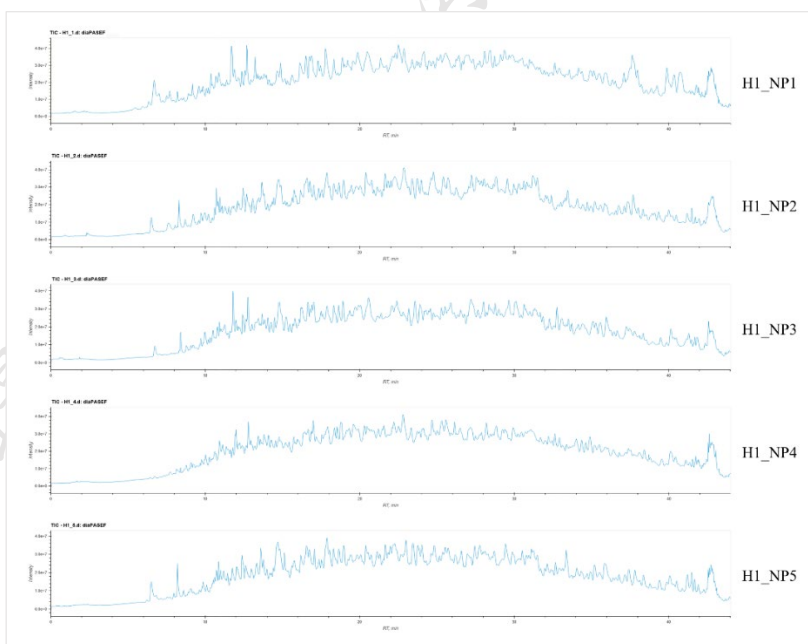


图 2. 代表性的样品质谱检测图。H1 样本 5 种 NP 富集的蛋白质质谱检测总离子图 (TIC)。

(三) 蛋白质鉴定数据汇总分析:

1. 使用 DIA-NN (version 1.8.1) [6] 软件对 30 个质谱文件进行检索, 完成蛋白质的鉴定和相对定量 (LFQ Intensity), 蛋白质鉴定信息汇总详见图 3。
2. 蛋白质鉴定表和肽段鉴定表详见《[pg_matrix.csv](#)》和《[pr_matrix.csv](#)》文件。表头参数参见:

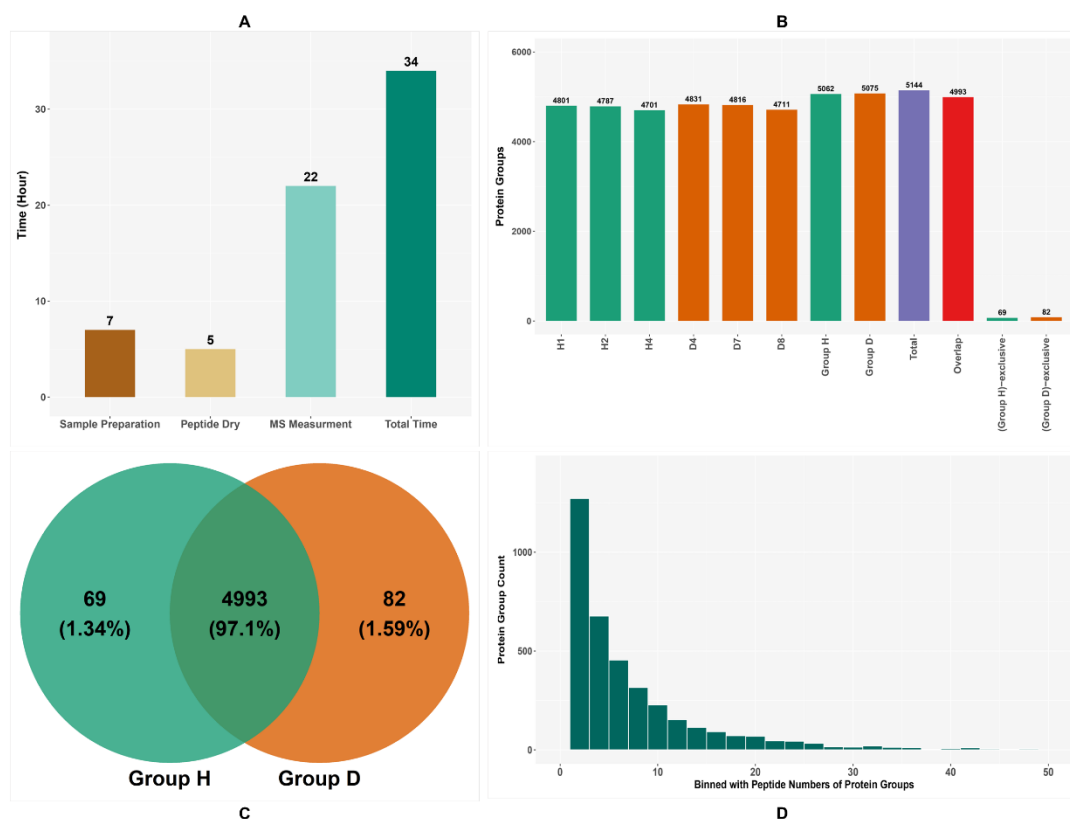


图 3. 蛋白质鉴定信息汇总。A. 从样品制备到质谱检测完成所需时间。B. 每个样本鉴定的蛋白质数目(H1~D8)、按组鉴定的数目(Group H, Group D)、两组全部鉴定的数目(Total)、两组共同鉴定的数目(Overlap) 和每组独有鉴定的数目((Group H)-exclusive, (Group D)-exclusive)。C. 两组(Group H, Group D)鉴定蛋白质的 Venn 图。D. 蛋白质鉴定肽段数目分布图。

结果分析

- 6 例样本，借助 Proteograph™ 的自动化系统，从样品制备到完成质谱检测，总实验时间仅为 **34 小时** (图 3A)! 并且实现了每例样本的蛋白质鉴定数目均超过 **4700 种** (图 3B) 的深度探查，总蛋白质鉴定数目为 **5144 种**!
- Group H 和 Group D 两组共同鉴定到的蛋白质 **4993 种**，重合度高达 **97.1%**(图 3C)! 显示了 Proteograph™ 系统的超高稳定性和重复性。**稳定和可重复才能保障数据质量。**
- 肽段数目分布图显示蛋白质鉴定的肽段数量丰富(图 3D)，数据质量高。

(四) 样本总体质量评估分析:

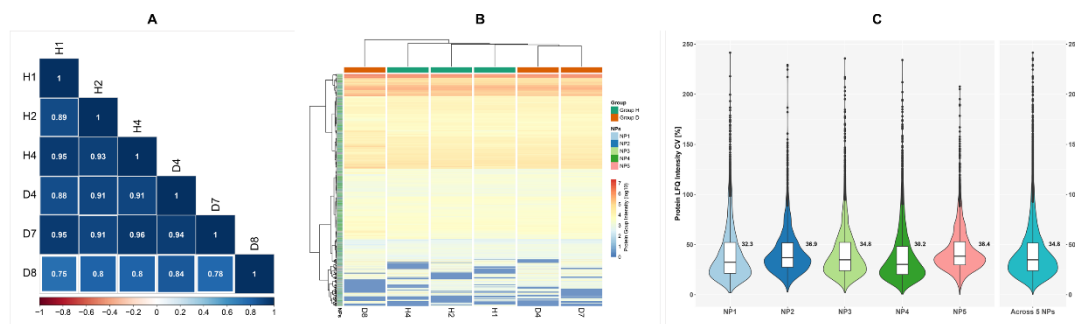


图 4. 基于蛋白质质谱定量强度 (LFQ Intensity) 的样本相关性分析及定量 CV (Coefficient of Variation) 分析。A. 样本 Pearson correlation 统计分析。B. 样本 Heatmap 统计分析。C. 5 种磁珠 (NP1~NP5) 及总体定量 CV 统计分析。

■ 结果分析

1. 基于蛋白质的质谱原始定量强度 (LFQ Intensity) 计算 6 例样本的 Pearson 相关系数, 图 4A 的结果显示 Group H 的 3 例样本的组内相关性 (≥ 0.89) 要优于 Group D 的 3 例样本 (≥ 0.78)。
2. 热图分析 (Heatmap, LFQ intensity (log10)) 结果与 Pearson 相关性分析相一致, Group H 的 3 例样本可聚类, Group D 的 D8 样本存在明显的异质性 (图 4B)。
3. 总体蛋白质的定量 CV 中值 (median, 图 4C) 为 34.8% (使用原始 LFQ Intensity 计算)。
4. 以上分析显示 :
 - (1) 个体差异导致的样本异质性可被 Proteograph™ 系统清晰揭示;
 - (2) 个体差异及小样本量, 导致整体定量 CV 值偏高。
 - (3) 提示对大队列样本 (large-scale cohort) 进行分析是筛选和发现血浆蛋白质分子标志物的必要条件。Proteograph™ 的自动化流程/系统正是为大队列样本分析而设计。

(五) 鉴定蛋白质的动态分布范围分析:

1. HPPP (Human Plasma Proteome Project) 是人类血浆蛋白质组研究数据库 (<https://peptideatlas.org/builds/human/plasma>) [8], 该数据库是用传统的血浆蛋白质组研究前处理流程 (高丰度蛋白去除 (depletion) + 肽段馏分分离 (Fractionation)) 结合质谱技术构建的。最新的 HPPP 数据库 (Build 2023-04) 显示通过对大量不同样本进行了 35,081 次质谱测试, 共鉴定并记录了 4608 种血浆蛋白质 (**大量不同样本检测到的蛋白质并集**) (图 5)。

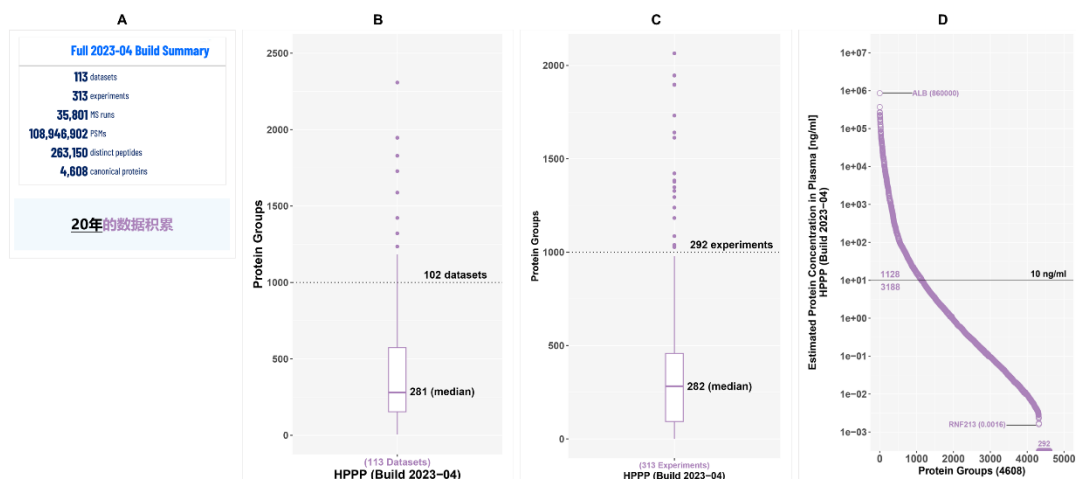


图 5. HPPP 数据库总览及传统技术流程对血浆蛋白质组的解析能力分析。A. HPPP 数据库 Build 2023-04 信息。B. HPPP 数据库 113 个数据集的蛋白质鉴定数目分布图，**90.3%** (102/113) 的数据集蛋白质鉴定数目 **<1000 种**。C. 313 次实验的蛋白质鉴定数目分布图，**93.3%** (292/313) 的实验蛋白质鉴定数目 **<1000 种**。D. 4608 种蛋白质在血浆种的浓度分布。传统技术流程对血浆蛋白质组的解析能力低下，无法在通量、深度和宽度上平衡/满足对大队列血浆样品的解析。

2. 本 DEMO 实验的 6 例样本共鉴定到 **5144** 种血浆蛋白，通过与 HPPP 数据库比对汇总结果如图 6 所示：

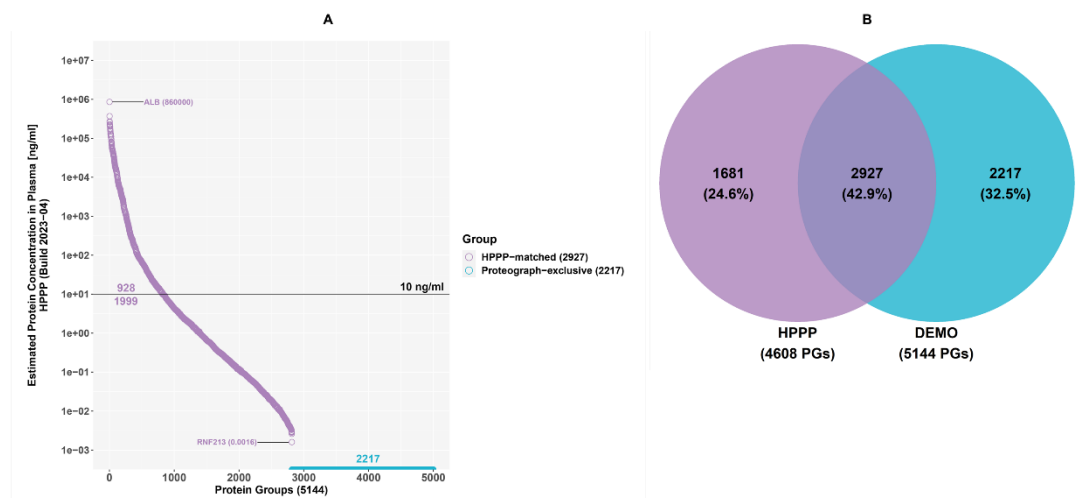


图 6. 鉴定蛋白质的动态分布范围 (dynamic range) 统计分析。A. 与 HPPP 数据库匹配蛋白质的动态分布范围 (跨越 9 个数量级) 分布图。B. 鉴定蛋白质 (5144 Protein Groups, PGs) 与 HPPP 数据库数据 (Build 2023-04) 记录的蛋白质 (4608 PGs) 比较 Venn 图。

结果分析

1. 有 **2927** 种蛋白质匹配到 HPPP 数据库中，蛋白质的浓度动态分布范围跨越 **9** 个数量级，浓度 ≤ 10 ng/ml 的蛋白为 **1999** 种。显示了 ProteographTM 系统对血浆蛋白质组的**高深度**覆盖 (图 6 A-B)。
2. 同时有多达 **2217** 种蛋白质 (Proteograph-exclusive)，是通过 ProteographTM 系统鉴定到的血浆“新”蛋白，显示了 ProteographTM 系统对血浆蛋白质组的**高宽度**覆盖 (图 6 A-B)。

(六) 组间差异表达蛋白质筛选分析:

1. 6 例样本，通过质谱检测，获得蛋白质的 LFQ Intensity 定量值。
2. 依据 LFQ Intensity，通过 label-free 定量方式[7]，筛选组间 (Group D vs. Group H) 的差异表达蛋白质 (Differentially Expressed Proteins, DEPs)。筛选与统计方式为 T-test, $p < 0.05$, $\log_2 FC$ (Fold Change) = $|\pm 1|$ 作为卡值 (cutoff)。筛选的 DEPs 信息汇总见图 7。

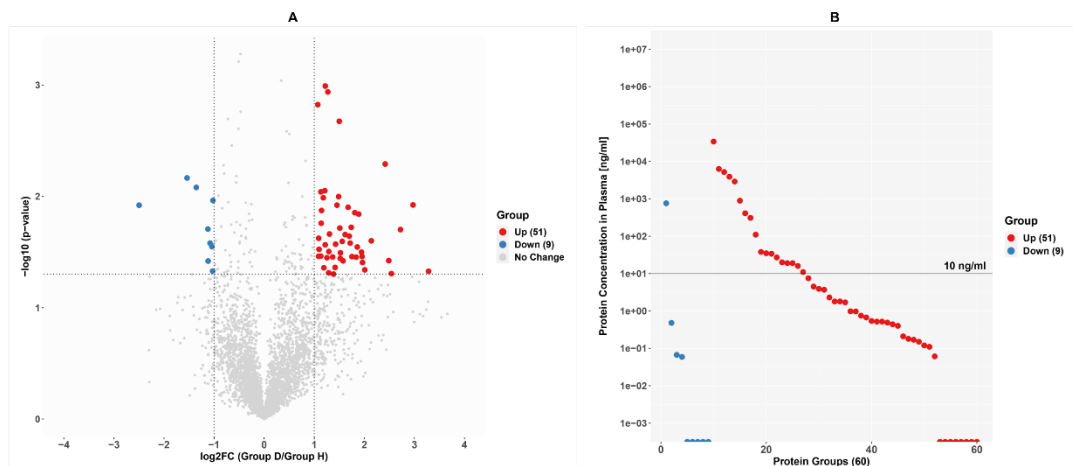


图 7. 差异表达蛋白质 (DEPs) 筛选及 DEPs 的血浆浓度分析。A. 基于蛋白质的质谱定量值 (LFQ Intensity)，通过 T-test 筛选组间 (Group D vs. Group H) 差异表达的蛋白质 (DEPs)。B. 差异表达蛋白质在血浆中的浓度分布分析。

结果分析

1. 通过统计分析，共筛选到上调蛋白 51 种 (Up)，下调蛋白 9 种 (Down) (图 7A)。
2. 60 种 DEPs 中：
 - (1) 有 28 种是浓度 $< 10 \text{ ng/ml}$ 的低丰度蛋白，显示了 ProteographTM 系统对血浆低丰度蛋白质变化的精细解析力 (逻辑上，血浆中的蛋白质分子标志物几乎都是中、低丰度的)
 - (2) 有 13 种 DEPs 是 ProteographTM 系统鉴定和发现的“新”血浆蛋白，显示了 ProteographTM 系统的创新探索能力 (图 7B)。

(七) 差异表达蛋白质 KEGG 和 GO 分析:

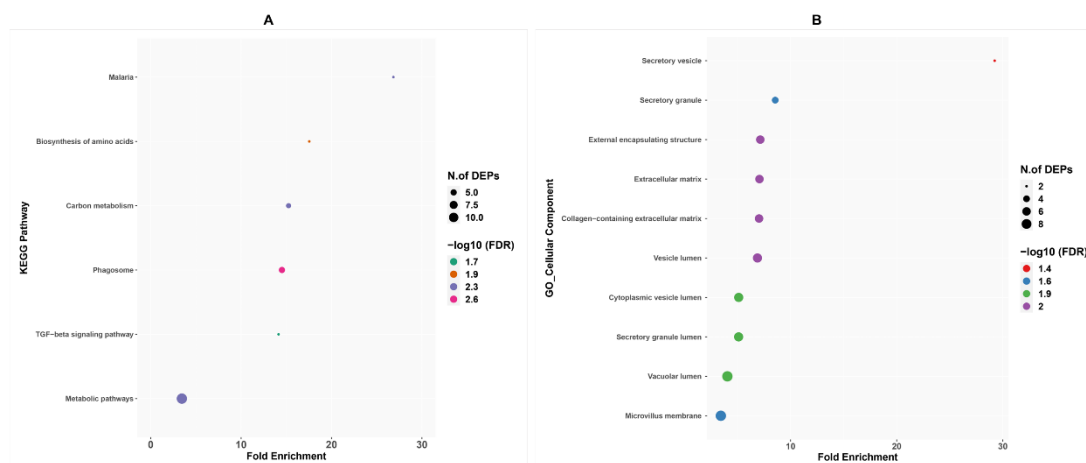


图 8. 差异表达蛋白质 (DEPs) KEGG 和 GO 富集分析。A. KEGG 富集分析。B. GO: Cellular Component (细胞组分)富集分析。

结果分析

- 对 60 种 DEPs 的 KEGG (<https://www.genome.jp/kegg/>) 富集分析显示这些差异表达的蛋白质被富集在多种信号通路中，其中富集数量最多的为代谢通路 (Metabolic pathway)，提示在 Group D 这些差异表达蛋白质与代谢异常相关 (图 8A)。
- 对 60 种 DEPs 的 GO(<https://geneontology.org/>) 细胞组分分析 (Cellular Component) 显示这些 DEPs 几乎都定位于囊泡和细胞外基质中 (图 8B)，提示这些蛋白质都是分泌蛋白，存在于血浆中是逻辑上合理的。

五、总结:

- 本 DEMO 实验，借助 Proteograph™ 的高深度血浆蛋白质组探索系统，对 6 例样本进行了探索性考察研究。
- 仅用 34 小时，完成了从样品制备到质谱检测的全部过程，共计鉴定到 5144 种血浆蛋白质，实现了对血浆蛋白质深度和宽度的双重高覆盖。
- 通过 label-free 定量筛选，发现了 60 种差异表达的血浆蛋白质，为该类疾病的血浆蛋白质分子标志物筛选提供了潜在的“靶标”。
- 因个体差异的原因，对大队列样本进行分析是筛选疾病相关血浆蛋白质分子标志物的必然要求。传统血浆蛋白质组分析技术流程无法兼顾样本通量和检测深度，导致血浆蛋白质组的探索研究远落后于基因组和转录组技术对血浆样本的分析。Proteograph™ 系统的高深度、高宽度血浆蛋白质探索能力结合自动化流程，同时实现了对血浆蛋白质组研究的通量、深度和宽度的探索需求，将重塑和强力推进血浆蛋白质组探索的新篇章。

六、参考文献:

1. Landegren U, Hammond M: **Cancer diagnostics based on plasma protein biomarkers: hard times but great expectations.** *Molecular oncology* 2021, **15**(6):1715-1726.
2. Anderson NL, Anderson NG: **The Human Plasma Proteome: History, Character, and Diagnostic Prospects *.** *Mol Cell Proteomics* 2002, **1**(11):845-867.
3. Geyer PE, Holdt LM, Teupser D, Mann M: **Revisiting biomarker discovery by plasma proteomics.** *Mol Syst Biol* 2017, **13**(9):942.
4. Blume JE, Manning WC, Troiano G, Hornburg D, Figa M, Hesterberg L, Platt TL, Zhao X, Cuaserna RA, Everley PA *et al*: **Rapid, deep and precise profiling of the plasma proteome with multi-nanoparticle protein corona.** *Nat Commun* 2020, **11**(1):3662.
5. Ferdosi S, Tangeysh B, Brown TR, Everley PA, Figa M, McLean M, Elgierari EM, Zhao X, Garcia VJ, Wang T *et al*: **Engineered nanoparticles enable deep proteomics studies at scale by leveraging tunable nano-bio interactions.** *Proceedings of the National Academy of Sciences* 2022, **119**(11):e2106053119.
6. Demichev V, Messner CB, Vernardis SI, Lilley KS, Ralser M: **DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput.** *Nat Methods* 2020, **17**(1):41-44.
7. Cox J, Hein MY, Lubner CA, Paron I, Nagaraj N, Mann M: **Accurate Proteome-wide Label-free Quantification by Delayed Normalization and Maximal Peptide Ratio Extraction, Termed MaxLFQ** *Mol Cell Proteomics* 2014, **13**(9):2513-2526.
8. Schwenk JM, Omenn GS, Sun Z, Campbell DS, Baker MS, Overall CM, Aebersold R, Moritz RL, Deutsch EW: **The Human Plasma Proteome Draft of 2017: Building on the Human Plasma PeptideAtlas from Mass Spectrometry and Complementary Assays.** *Journal of Proteome Research* 2017, **16**(12):4299-4310.