Data dependent *versus* Data independent acquisition applied to samples with different protein dynamic range

Nicolas Smargiasso[1], Gabriel Mazzucchelli [1], Dominique Baiwir [2], Edouard Louis [3], Florence Quesada-Calvo[3], Marie-Alice Meuwis [3] and Edwin De Pauw [1]

[1] Laboratory of Mass Spectrometry, GIGA-R, chemistry department, CART, University of Liège, [2] Proteomic Facility, GIGA, University of Liège, [3] Hepato-Gastroenterology department, GIGA-R, Liège University Hospital, CHU - ULg

Introduction

Moderate to low invasive biomarker sources like serum are the gold standard for diagnosis or prognosis purposes
Biomarkers of disease state or progression lies most likely in the low concentration range

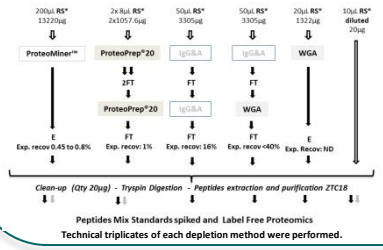
Plasma and serum are very complex due to : 1) Large protein concentration dynamic range and 2) High number of different proteins present depending on the physiological state → need to remove abundant proteins

GOALS OF THIS STUDY: - compare Data independent (Q Exactive) and dependent (Synapt G2 HDMS) analyses on raw, depleted serum and a biopsy sample.
- determine strenght and weekness of each analytical worflow with the same UPLC separation.

Qualitative and Quantitative Comparisons of: - results obtained with depletion kits or with raw serum
- both analytical setup for samples with different protein dynamic ranges

Methods

Serum sample: Pool of 3 healthy patients, 5 patients with colorectal cancer, 5 patients with Crohn's disease and 3 patients with ulcerative colitis. The pool was spiked with Invertase (INV1_YEAST, 1% w/w) and fetuin (FETUA_BOVIN, 0,5% w/w).
Colonic biopsy: one biopsy from colonic mucosa was collected and frosen, homogeonized using the Bioruptor® (Diagenode, Belgium) according to the manufacturer recommandations. Same purification steps and digestion protocol as for sera samples was applied.
UPLC: NanoAcquity (Waters): 2D with dilution configuration: - First dimension: X-Bridge precolumn : elution in 5 steps (10,8%, 14%, 16,7%, 20,4%, 65% solvent B) pH=10
- Second dimension: Symmetry precolumn and BEH C18 analytical col. 25 cm, 85 min gradient (97% to 60% solvent A) pH=3
Mass spectrometry: - Q-Exactive (Thermo): data dependent acquisitions: MS acquisition: res. 70000, AGC target 10^6, max accu time 200 ms; MS/MS acquisition: res. 17500, AGC target 10^5, max accu time 50 ms, Top12 selection.
- SYNAPT G2 (Waters): data independent (MS^E) method with IMS separation, source parameters: capillary 2,5 kV, sampling cone 35 V. IMS parameters: IMS cell pressure 2,5 mbar (N_2), Waves height 40 V, Waves velocity 650 m/s.
Protein identification and relative quantitation: MaxQuant 1.4.1.2 (Q Exactive data) and ProteinLynx Global Server 3.0 (Waters, SYNAPT G2 data). Protein FDR set to 1% for the identification, min. 2 peptides per protein.



Technical triplicates of each depletion method were performed.

Experimental scheme - serum depletion

Sample handling characteristics and recovery

Characteristics	ProteoPrep 20 plamsa	Proteominer	IgG&A	IgG&a+WGA	WGA	Raw Serum
Total starting protein quantity according to RCDC protein quantitation on serum pool (µg)	1056.7	13220	3305	3305	3305	20
Volum of spiked raw serum pool required and treated (µL)	16	200	50	50	50	0
Recovery of total proteins after depletion[1] (mean of 3 replicates +/-SD, expressed in percent of total quantity treated) (%)	1.91+/-0.17	2.35 +/-0.34	24.47+/-4.93	7.58+/-0.52	32.91+/-3.21	100+/-0
Expected protein recovery (% of starting protein quantity)[2]	1	0.4 to 0.7*	16	<40	ND	100
Manipulation time for kit application (hours)	3	4	3	5.5	2.5	0.1

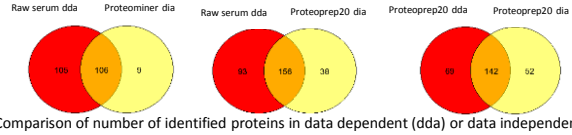
Notes :[1]recovery was calculated using SYNAPT G2 quantification, [2] according to manufacturer

Results

SYNAPT G2: analysis of 5 depleted samples and raw serum

Characteristics	ProteoPrep 20 plamsa	Proteominer	IgG&A	IgG&a+WGA	WGA	Raw Serum
Nbre of proteins identified in at least 1 replicate	258	199	201	182	229	225
Nbre of proteins identified in at least 3 replicatesout of 3	194	115	135	127	130	90
Protein conservation after the 3R3 filter application (% of number detected in at least 1 replicat)	75.2	57.8	67.2	69.8	56.8	40
Nbre of proteins identified within 3 replicates, range	222-246	157-212	162-184	138-184	156-197	156-188
Number of protein identified in 3R3 reported to mean protein quantity detected[2] (number prot/ng)	0.13	0.14	0.16	0.19	0.19	0.13
Number of proteins unique to the condition and detected in 1R3	49	27	5	0	7	39
Number of proteins unique to the condition and detected in 3R3	12	4	0	0	0	3
HSA residual quantity mean of 3 replicates (ion count table data) (% of theoretical starting quantity[4])	0.032	0.009	0.050	0.005	0.117	3.683
HSA quantity (% of the total protein quantity mean fOR each replicate per kit type +/- SD)[3]	3.11+/-1.42	0.09+/-0.03	0.68+/-0.56	0.24+/-0.05	3.39+/-0.5	9.24+/-2.66
Ranking according to HSA depletion efficiency and according to HSA residual final Qty quantified[3]	4	1	3	2	5	NR
Spiked protein invertase (% of initial yeast invertase Qty spiked in raw serum)	15	0	3.33	0	0	0
Spiked protein bovin fetuin (% of initial fetuin bovine Qty spiked in raw serum)	0	0	0	0	0	0
Ig family quantification - ion count table data (mean %tage of total quantity in 3R per type of kit)	2.24+/-1.75	4.82+/- 1.62	5.29+/-0.78	5.99+/-4.53	5.96+/-0.86	8.65+/-1.43

Comparisons of data dependent (Q Exactive) and data independent (SYNAPT G2) MS/MS

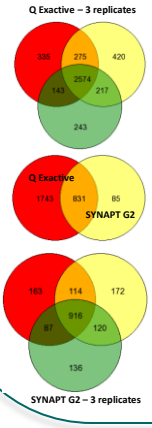


Comparison of number of identified proteins in data dependent (dda) or data independent (dia) methods - Only proteins found in 3 repilicates out of 3 are taken into account.

- Proteoprep 20 kit allows identifying the highest number of proteins but HSA depletion is not fully efficient.
- Using WGA kit, the depletion of IgG is lower because these proteins are glycosylated.
- Spiked proteins were only partially detected. A lower detection threshold allowed detecting them in Proteoprep20 treated samples.

- For a given sample, more proteins were identified using the data dependent MS/MS method.
- Analysis of raw serum using the data dependent MS/MS method gives better results than data independent MS/MS method even after treatment with Proteominer or Proteoprep20.

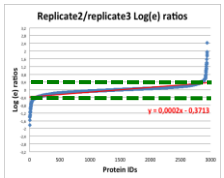
Protein groups identification



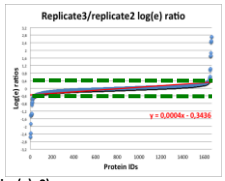
- More proteins were identified using the DDA method using the Qexactive compare to the DIA method using the Synapt G2 (same UPLC).
- No significant protein identification reproducibility difference was found between both analytical setups.
- 91% of the protein groups identified with the Synapt G2 setup were also identified with Q Exactive. This value drops to 55% when considering proteins for quantification.

Colonic Mucosa biopsy, 3 injection replicates
Protein quantification

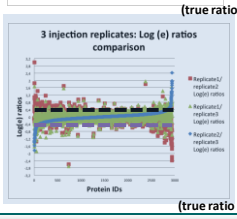
DDA – Q Exactive

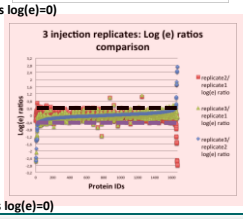


DIA – Synapt G2 HDMS



(true ratio is log(e)=0)





(true ratio is log(e)=0)

	DDA-Qexactive	DIA-Synapt G2 HDMS
Number of quantifiable proteins common in all replicates	2943	1667
Number of proteins showing a ratio between 1.2 and 0.833 (considering all replicates)	1301	1030
Number of ratios between 1.2 and 0.833 over the total number of ratios (%) (considering all replicates)	44%	62%
Number of proteins showing a ratio between 1.5 and 0.666 (considering all replicates)	2500	1547
Number of ratios between 1.5 and 0.666 over the total number of ratios (%) (considering all replicates)	85%	93%
Number of false positive ratios (>1.2 and <0.833) (considering all replicates)	55	18
False portive rate (>1.2 and <0.833) (considering all replicates)	1,87%	1,08%
Number of false positive ratios (>1.2 and <0.833) (considering 2 out of 3 replicates)	755	258
False portive rate (>1.2 and <0.833) (considering 2 out of 3 replicates)	25,65%	15,48%
Number of false positive ratios (>1.5 and <0.666) (considering all replicates)	4	0
False portive rate (>1.5 and <0.666) (considering all replicates)	0,14%	0,00%
Number of false positive ratios (>1.5 and <0.666) (considering 2 out of 3 replicates)	386	38
False portive rate (>1.5 and <0.666) (considering 2 out of 3 replicates)	13,12%	2,28%

Dash lines correspond to protein showing 50% over or under representation
- Q Exactive: 7.9% ± 1% of the measured ratios are outside the limit (n = 3).
- Synapt G2: 3.6% ± 0.9% of the measured ratios are outside the limit (n = 3).
- Relative quantitation using DIA method with the Synapt G2 HDMS is more precise compare to the DDA analysis using the Qexactive

Conclusions

- According to the number of proteins detected, Proteoprep 20 depletion samples analyzed on the Q Exactive (data dependent MS/MS) method instrument give the best results.
- Even after depletion, access to low concentration dynamic range proteins is still limited by the presence of few relatively abundant proteins.
- DDA method using the Q Exactive allows to identify more proteins especially in complex mixture with moderate protein dynamic range (≈77% more) compared to sera sample analysis (30% more).

Aknowledgements

This study was funded by the STORI project. The F.N.R.S., European Union (FEDER) and Walloon Region contributed the mass spectrometry facility funding.