Pushing DIA proteomics analyses of neat plasma towards 1000 protein groups ID's/injection

We optimized our neat plasma proteomics acquisition workflow, working on the separation, MS acquisition and data processing parts, before applying it to a rare dermatological genetic disease study.

Abstract

Undepleted (neat) plasma samples have been analyzed with one "deep workflow" setup using a 30 min gradient with a nanoElute® and one "fast workflow" using a 60 SPD method on an Evosep One. The first allowed to identify an average of 1091 protein groups while the second allowed to identify up to 638 Protein groups per sample on average, using a dia-PASEF® approach on a timsTOF Pro. Some early real-life results obtained with these workflows will be displayed and compared.

Introduction

Its accessibility and biological role make plasma an attractive sample for diagnosis or prognosis operations, and therefore it's also a sample source of choice for biomarker research studies, especially when they need to be performed on a very large scale. However, proteomics analysis of neat plasma remains a challenge because of the huge dynamic range of the plasma proteins, and the current depletions technologies can be expensive. In this application note, we focus on workflow optimization of neat plasma analysis using the dia-PASEF® approach to maximize the number of proteins groups (PG) identified and quantified, while minimizing both gradient time and missing value levels. We describe the comparisons of different LC settings and IM range, and the optimization of data analysis using DIA-NN.

In particular, we report a reference "deep workflow" to quantify up to 1091 PG and 16417 unique peptides per sample (906 PG) when considering a minimum of 70% data completeness per group) using a 30 min gradient and a "fast workflow" leading to up to 638 PG and 6328 unique peptides per sample (542 PG when considering a minimum of 70% data completeness per group) using a 21 min gradient (60 SPD).

Ines Metatla[1], Kevin Roger*[1], Sara Ceccacci[1], Cerina Chhuon[1], Maria Leite-de-Moraes[2], Anne Jamet[2], Pierre-Olivier Schmit[3], Manuel Chapelle[3], Ida Chiara Guerrera[1]; [1]Necker Proteomics Platform, Faculty of Medecine, University Paris Cité, Paris, France; [2]Institut Necker Enfants Malades, Paris; [3]Bruker France SAS, Wissembourg, France.

Materials and Methods

Optimization tests have been performed using healthy patient plasma. For the small-sized clinical study we used a plasma sample cohort consisting of samples from 15 patients affected by a rare dermatological genetic disease (RDGD) and 18 age-matched controls patients (CP). Samples were digested with trypsin (Promega) using S-Trap™ micro spin column (Protifi, Hutington, USA). Peptide digest concentrations were measured using a nanodrop 2000 spectrophotometer system (Thermo Scientific, USA) and separated either by nano-HPLC (nanoElute, Bruker Daltonics GmbH & Co. KG, Bremen, Germany) using a 25 cm Aurora column (IonOpticks, Australia) and a 15 min or 30 min gradient, or using an Evosep One operated with the 100 or 60 Samples Per Day (100 SPD or 60 SPD) method and the 8 cm Performance column. Both LC's were connected to a timsTOF Pro™ instrument (Bruker Daltonics GmbH & Co. KG, Bremen, Germany) via its CaptiveSpray source. The timsTOF Pro was operated in PASEF or dia-PASEF acquisition modes. For the latter, we used a 19-windows method using 33da windows covering the described mobility ranges over a 400-1050 *m/z* range. Data processing was performed using MaxQuant V2.3.0 (Cox group, PASEF data) or DIA-NN v1.8.1 (Demichev group, dia-PASEF data).

Results and discussion

PASEF versus dia-PASEF: technical and biological reproducibility

Biological interpersonal variability can be huge on human samples, as it can depend on genetic, environmental or sampling factors, leading to high variations between individuals. Inferring a disease-related variation is challenging and requires minimizing technical variability while increasing analysis depth and, when possible, the number of individuals analyzed. Studying rare diseases, the number of patient samples available is naturally limited and drives even stronger requests on analysis depth and reproducibility. We evaluated technical and biological variation for both PASEF and dia-PASEF approaches on a subset of 10 healthy patient samples. Technical variability was assayed by injecting 10 times the same plasma sample while biological variability was evaluated by injecting 10 different plasma samples. All were separated using a 60 SPD method and measured in both PASEF and dia-PASEF modes, searches were performed with MaxQuant and DIA-NN respectively. Results are shown in Figure1. dia-PASEF allowed to identify roughly 50% more protein groups from the same samples (Figure 1A) while also reducing variability (Figure 1B): indeed, at any CV threshold chosen, dia-PASEF always allowed to quantify more proteins at the selected threshold. Consequently, the number of proteins we can expect to quantify is higher with dia-PASEF.
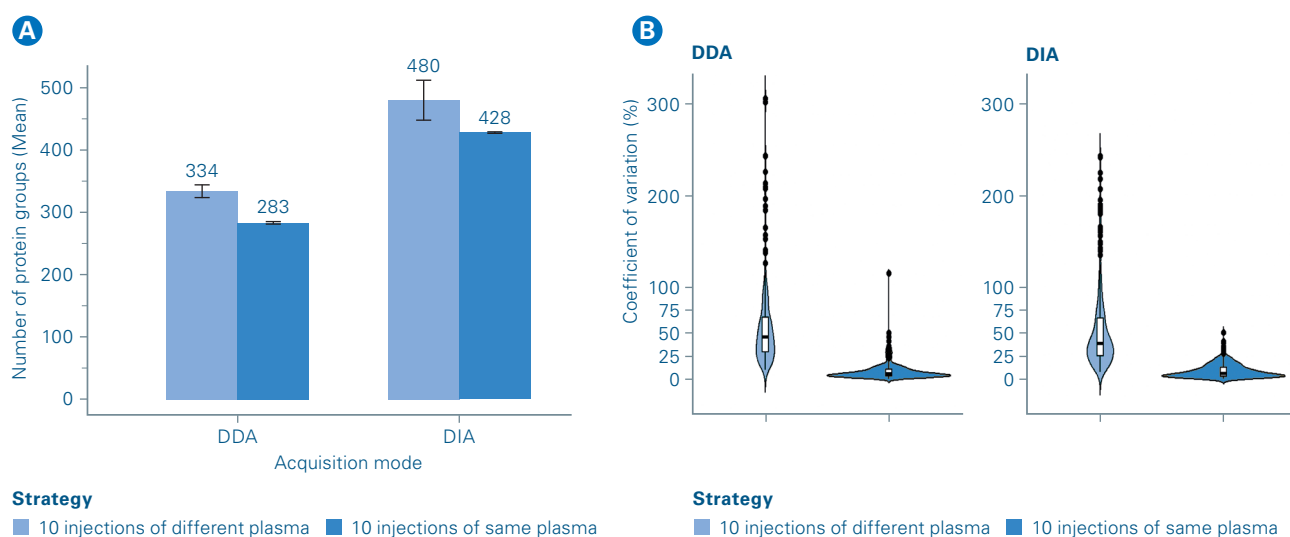

Figure 1

Technical vs. biological reproducibility

Results from 10 injections of 1 neat plasma sample (technical variability) or 10 injections of 10 different neat plasma samples (biological variability). Separation with a 60 SPD method on an Evosep One. Acquisition with optimized PASEF and dia-PASEF acquisition methods.
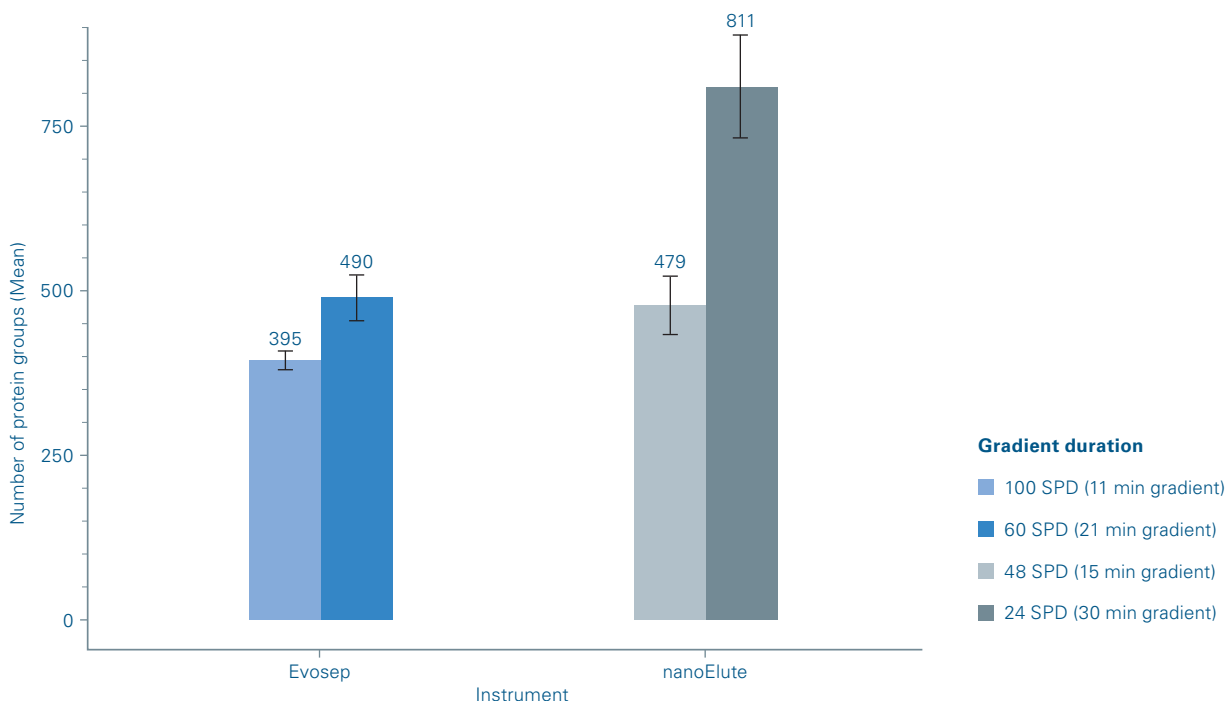
Figure 2

Gradient and column length influence

Outcome of a direct DIA search for 33 neat plasma samples. Evosep was used with an 8 cm Performance column while the nanoElute was equipped with the 25 cm IonOpticks Aurora column.
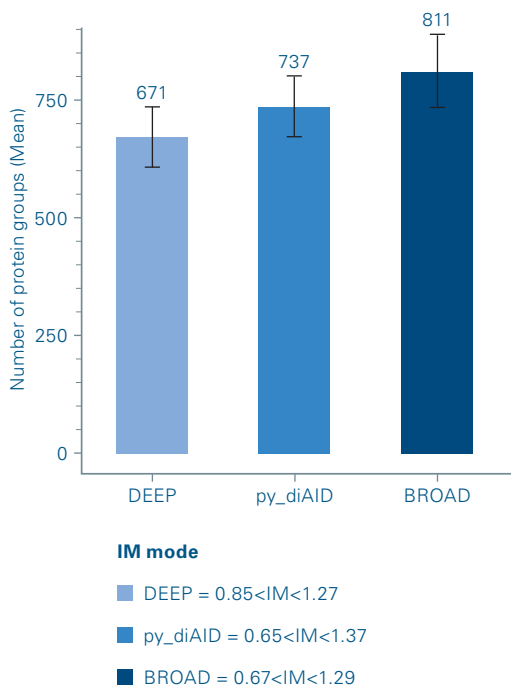
Impact of peptide separation: column length and gradient influence

We compared different gradient length on the Evosep One (100 SPD – 11 min gradient – and 60 SPD – 21 min gradient) using an 8 cm column as well as on the nanoElute (15 and 30 min gradient) using a 25 cm emitter embedded IonOpticks column (48 SPD and 24 SPD, respectively). This later column would not be the first choice for a 15 min gradient, as the associated washing and equilibration times it implies are reducing the interest of running a short gradient, but it allowed to check for the influence of the column length on the outcome (Figure 2). The results obtained from the 15 min gradient with the 25 cm column are close, with an average of 479 PG IDs, but still inferior to the ones obtained with the 60 SPD method (490 PG IDs) while the corresponding throughput is twice lower. Using the long column with short gradients makes therefore no sense for such short gradients. The 100 SPD method allowed to deliver 395 PG ID's on average, which is a 20% loss of PG ID's for a 66% increase of throughput.

Using the longer 30 min gradient with the 25 cm column, however, allowed to identify an average of 811 Protein Group ID's (65% PG IDs increase for a 63% loss of throughput).

Figure 3

dia-PASEF method optimization

Outcome of a direct DIA search for 33 neat plasma samples acquired with dia-PASEF methods covering the same mass range but different ion mobility ranges.

Impact of ion mobility acquisition range

We used three different dia-PASEF methods schemes covering similar *m/z* but different mobility ranges (Figure 3). A broader IM range allows to identify up to 20% more proteins. Unexpectedly, the outcome was the reverse of what we observe from cell line digests (data not shown), where best results are obtained from the method focusing on a smaller ion mobility range (Figure 3). Acquisitions with a method generated using PydiAID script did not allow to improve the PG identified. We therefore decided to use the "Broad" acquisition method to continue the evaluations.

Digging deeper: dia-PASEF method optimization

We further tried to improve the analysis depth we could achieve using dia-PASEF, by working on the search strategy.

For best results, we finally used dia-NN with the library free mode. In agreement with the high interpersonal variability of plasma protein abundance we report an increase of PG per group as more samples files are searched at the same time (Figure 4), in particular when Match Between Run (MBR) was allowed. This is a known effect, but it is accentuated in plasma samples compared to patients' tissue samples. Furthermore, we searched the 33 plasma samples in parallel with 5 runs of plasma extracellular vesicles purified from 5 healthy individuals independent form the cohort (search EV-boost). The number of PG obtained on average per samples using the 24 SPD method raised to 1091 PG (±181) compared to 811 PG (±78). Using the 60 SPD method the average number of PG was increased to 638 (±75). Data completeness, however, was lower, and when reduced to 50% data completeness we could retain 885 PG on average per sample (total matrix of 998 proteins), including 261 proteins obtained thanks to the EV-boost search.
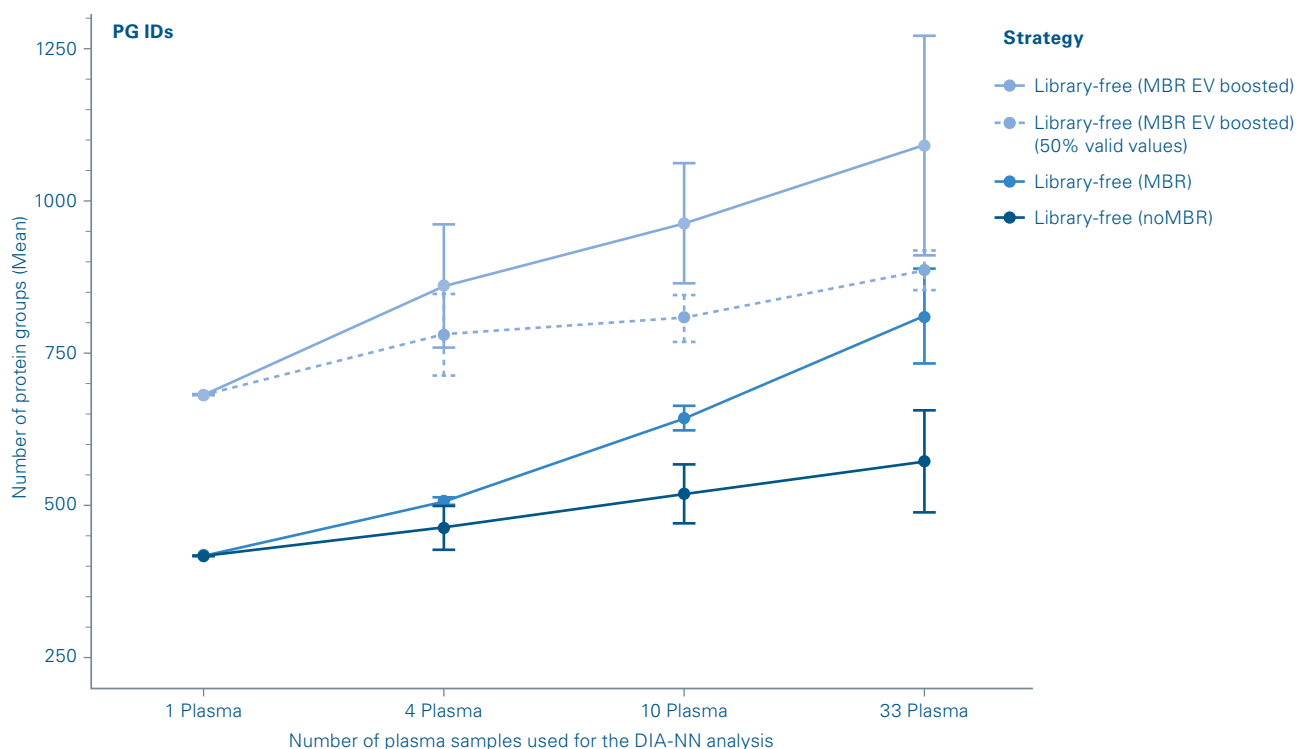

Figure 4

Search strategy optimization

Effect of the number of samples on the direct DIA search outcome. Comparison with or without MBR or searched with 5 runs of plasma extracellular vesicles purified from 5 healthy individuals independent form the cohort (search EV-boost).

Comparing deep and fast workflows: case study

The analyses of the 15 RDGD neat plasma samples vs. the 18 control samples (CP) allowed to detect 45 or 36 over-regulated protein groups and 73 or 37 down regulated PG's using the deep or fast workflow (applying a 1% FDR rate in both cases) (Figure 5A). PCA analysis on both datasets allowed to separate the controls form the RDGD, and even stratify the patients according to the severity of the disease (Figure 5B). Gene set enrichment analysis highlighted the increase of inflammation proteins in both datasets. However, the deep workflow allowed to also highlight the decrease of protein implicated in the skin extracellular matrix, highly disrupted in RDGB patients. Protein of the extracellular matrix (ECM) and collagen fibrin related proteins were hardly detectable in the fast workflow. As an example, dermatopontin (DPT) was only detected in all patients control and in mild forms of the RDGD (Figure 5C,D).
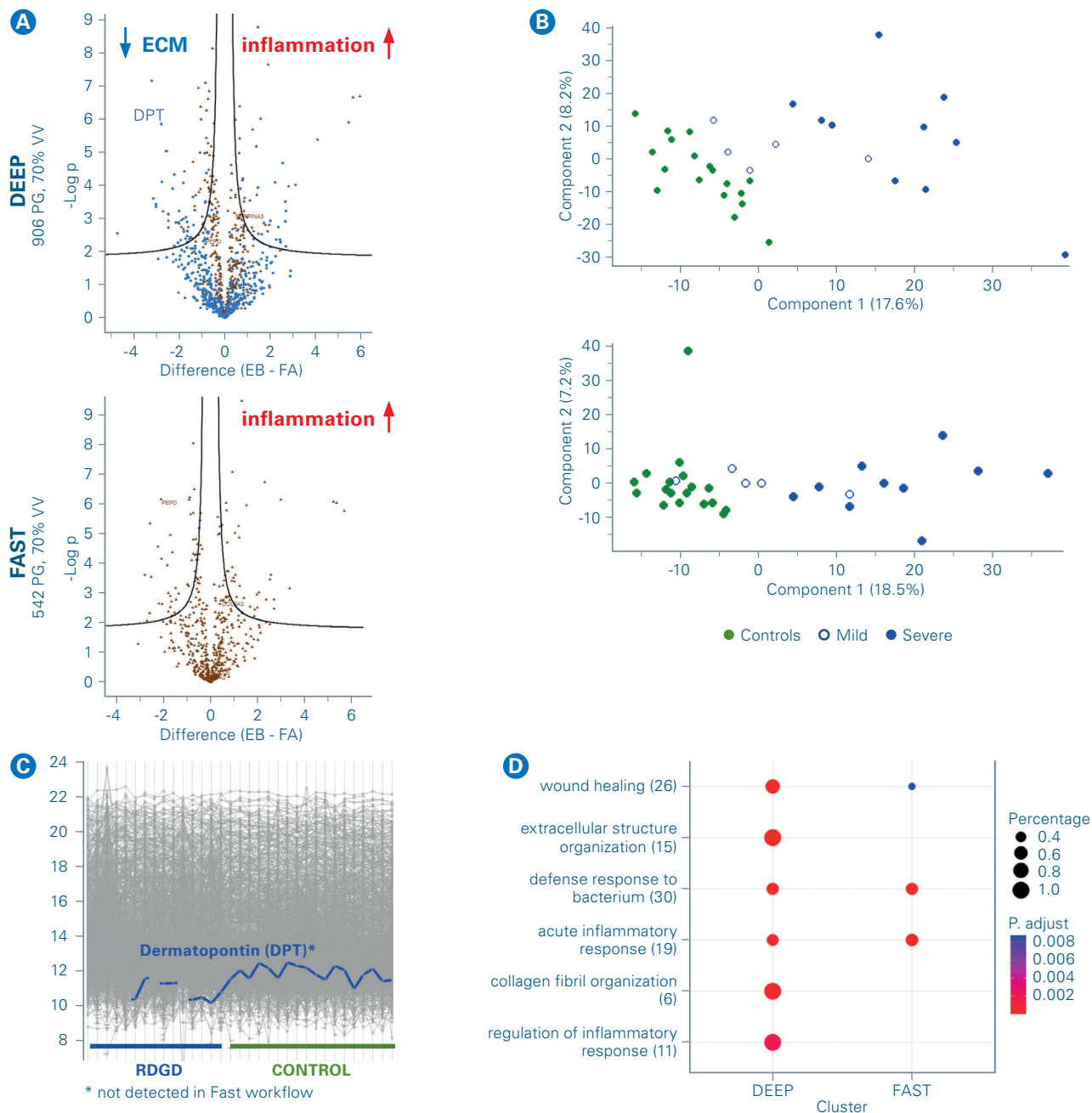

Figure 5

Snapshot of biological outcome

(A) LFQ volcano plots. Blues dots reveal the proteins uniquely detected with the DEEP approach. (B) PCA analysis outcome (control vs RDGD). (C) Profile plot of differential protein dermatopontin, detected exclusively with the deep workflow. Group related protein regulation (D) GO CC enrichment for the differential proteins detected using « deep » and « fast » workflow.

In conclusion both workflows can deliver the main biological messages, but deeper analysis gave us access to a new relevant class of proteins not detected (or rarely detected) in the faster pipeline.

Choice of depth over speed should be made as a function of the size of the cohort to analyze. We are now continuing further optimizing the methods to reduce time, while keeping the depth of the analysis (using 15 cm column, optimizing data analysis).

Conclusion

- The timsTOF Pro based dia-PASEF approach we are using allows for highly reproducible protein quantitation in neat plasma.
- We could quantify close to 550 protein groups from a 60 SPD neat plasma analysis, and more than 906 using a 30 min gradient.
- Those methods already allowed to deliver the physicians with relevant biochemical information.
- We are further investigating to improve these performances (improving data acquisition, chromatography and data analysis)