Evaluating the benefit of dia-PASEF approaches and sample-specific database strategies for metaproteomics of very complex microbiomes

PROTEOAIX 2023

Thibaut Dumas1, Olivier Pible[1], Guylaine Miotello[1], Kristina Marx[2], Pierre-Olivier Schmit[3] and Jean Armengaud[1]

[1]CEA-Marcoule - Laboratory «Innovative technologies for Detection and Diagnostics», Bagnols-sur-Cèze, France
[2] Bruker Daltonics GmbH & Co.KG, Bremen, Germany
[3] Bruker France SAS, Wissembourg, France

Introduction

Understanding interaction of microorganisms with their host is crucial in microbiology. The identification and quantification of proteins from a complex mixture of a large variety of organisms, known as Metaproteomics, has emerged recently as a unique tool to get functional and taxonomical insights into microbiota and is currently one of the most challenging areas in proteomics. Species-specific signals can be low, highly diverse, and the search space is extremely large. For these reasons, the increased selectivity potential of LC-IMS-MS based approaches might prove beneficial for metaproteomics analyses. In this communication, we are using a specific sample representative of gut microbiome to benchmark dia-PASEF approaches against PASEF approaches and investigate the benefit of sample-specific database strategies.

Methods

A gut microbiome tryptic digest (200 or 800 ng) was injected on a 25cm X 75µm pulled emitter column (IonOptiks). Nano-HPLC separation was performed with a 36, 66 or 100 min gradient using a nanoElute (Bruker) connected to a timsTOF HT or a timsTOF Ultra mass spectrometer (Bruker). LCMSMS data were acquired in PASEF or dia-PASEF acquisition mode. PASEF data have been processed in real-time on PaSER (Bruker) or Mascot (MatrixScience), searching against an NCBI database. Dia-PASEF data have been searched against a reduced protein sequence database using TIMS DIA-NN on PaSER (Bruker) or Spectronaut 17 (Biognosys). The protein sequence reduced database was constructed after confidently proteotyping the most abundant organisms present in the standard sample. It comprised 893,451 protein entries from 57 taxa. DeNovo searches have been performed using Bruker ProteoScape Novor, against both the restricted library or the bacterial reference proteome of Uniprot. All searches performed with a 1% protein FDR threshold.

Results

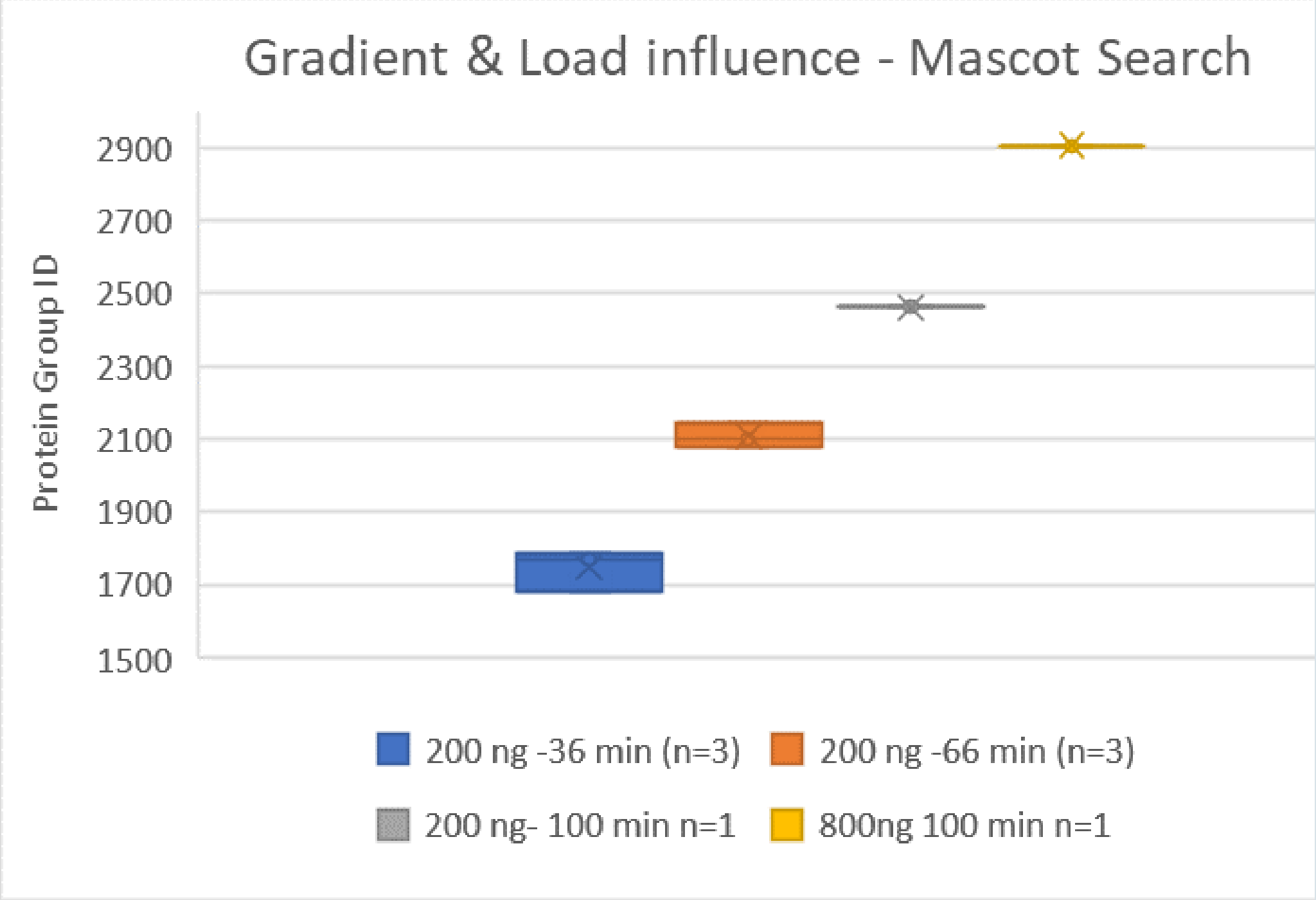

Fig. 1: Separation and Load influence

Various gradient lengths and loads have been tested, from 36 to 100min and from 200 to 800ng. PASEF (DDA) acquisition, search with Mascot (Ncbi nr). All consecutive acquisitions have beenperfromed with conditions similar to the CEA’s lab : 200ng load and 66 min gradient.

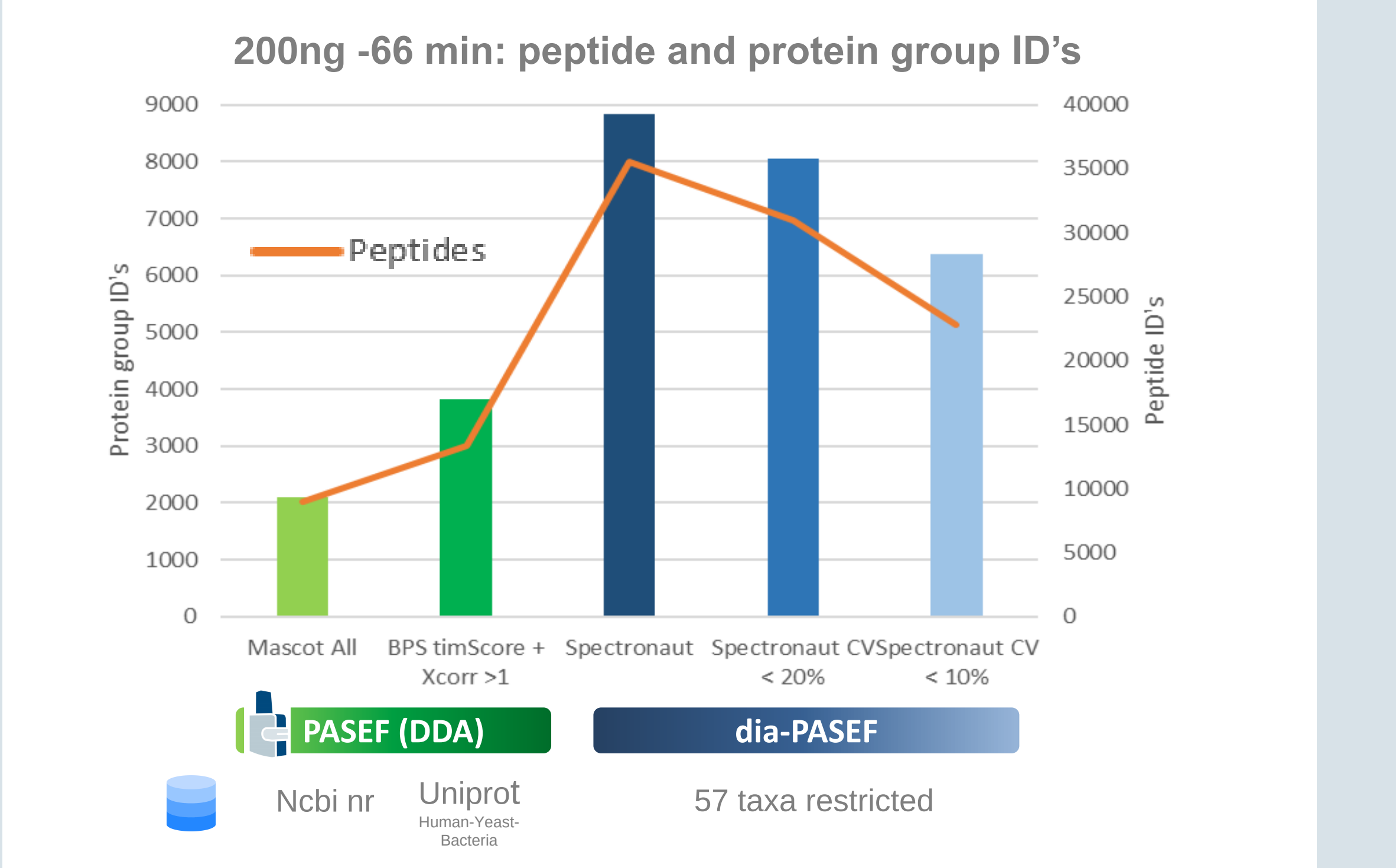

Fig. 2: Acquisition and search optimisation

Summary of the Protein group and peptide ID’s obtained with different acquisition methods and databse search strategies. PASEF data have been searched from Uniprot, dia-PASEF data have been searched from a library built using the 57-taxa restricted Db.

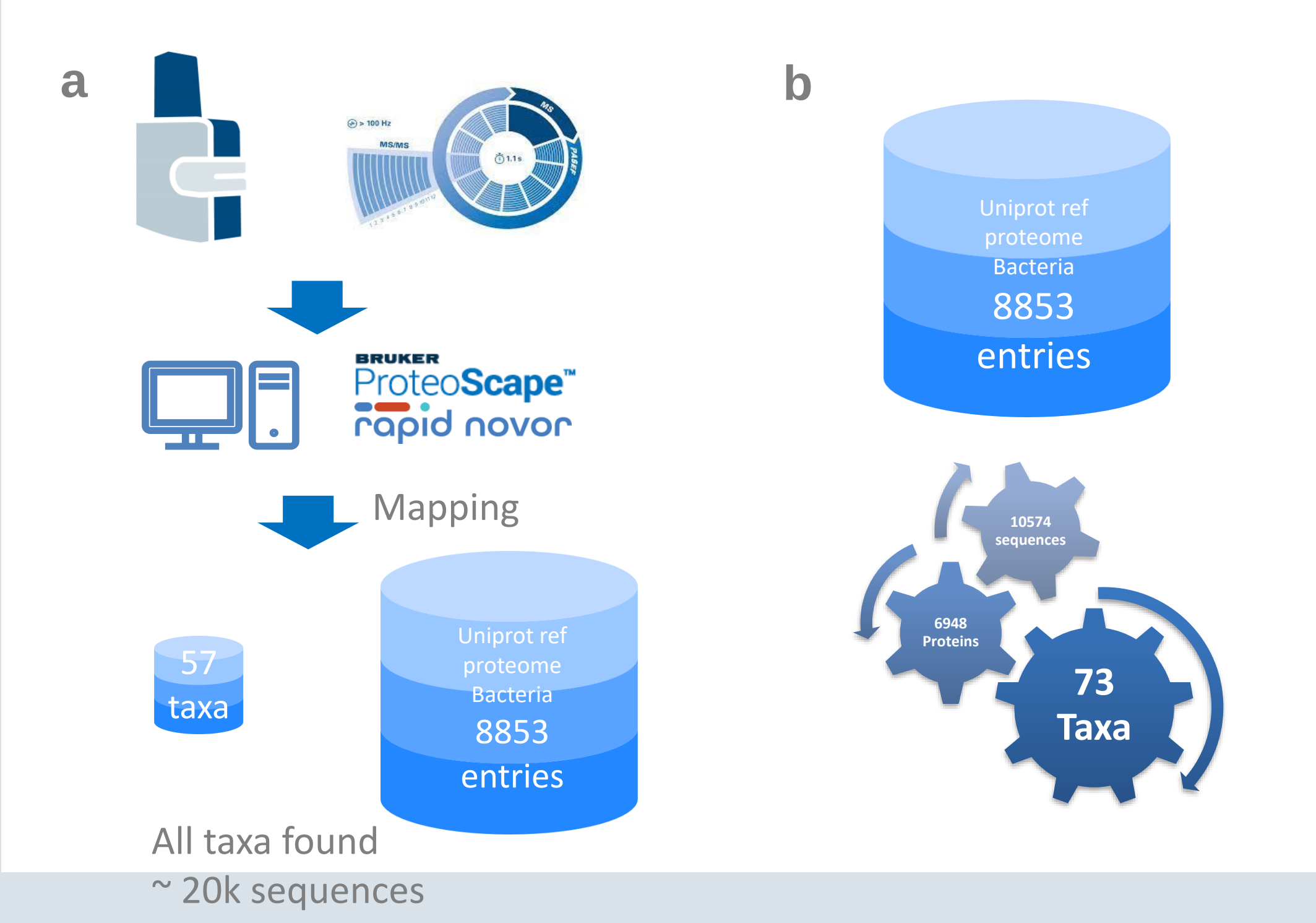

Fig. 3: Mapping DeNovo results on large databases.

PASEF data have been subjected to a DeNovo search and the sequence obtained have been mapped either on the 57 taxa database or on Uniprot’s Bacterial reference protome (8853 species / strain entries)

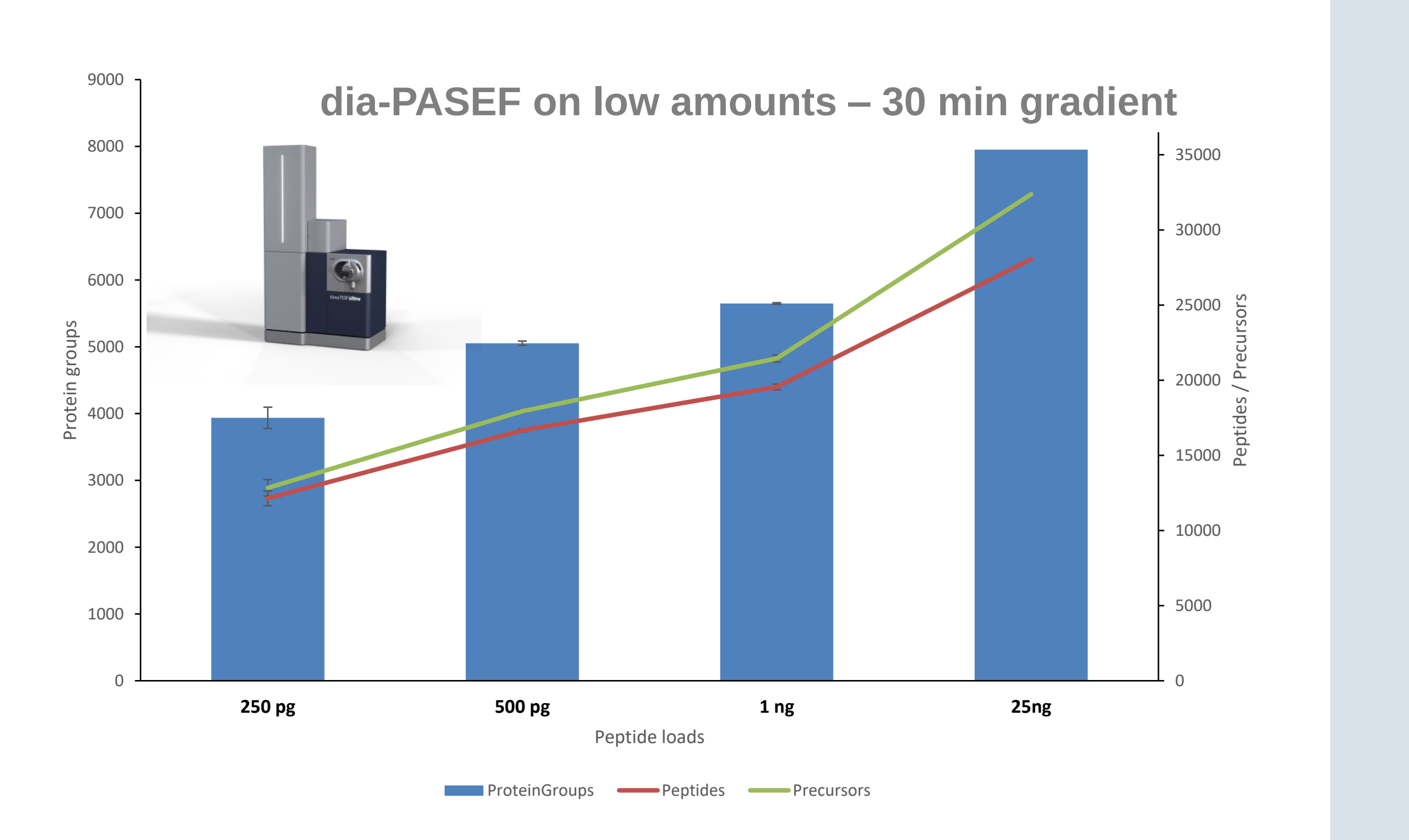

Fig. 4: high sensitivity searches.

250pg to 25ng of the sample were separated using a 25 min gradient and measured with the same dia-PASEF generic method as already used on a newly introduced timsTOF Ultra. Searches have been performed with Spectronaut as in Fig. 2, using the very same library.

The results initially obtained from a Mascot search from all 200 ng injections showed only a slight increase of the protein group ID number while jumping from a 36 min to a 66 min gradient and from a 66 min gradient to a 100min gradient (12 and 15%, respectively). There was more effect by quadrupling the injected amount (+18% protein group ID’s while jumping from 200 to 800ng injected with the 100 min gradient). In fine, 2110 protein groups could be identified from a 200ng injection and a 66min gradient (**Fig 1**). On average, 4 peptide sequences were identified for each protein Group ID’s. Using the ProLucid Algorithm on PaSER while allowing to take the peptides’s collisional cross section (CCS) value into account for the scoring process (TIMScore), a high stringency allowed to double the number of identified protein groups, while the number of sequences/proteins was down to 3 (**Fig. 2**, dark green). This suggest that the use of the TIMScore allowed to identify more of the low-intensity peptides, which also resulted into adding protein groups described by a lower number of peptides per protein.The use of dia-PASEF with the restricted database allowed to boost the number of protein group ID’s: 8,827 for 35,516 peptides for the 200ng/66 min injection (+35% protein group ID’s, +78% peptide sequences). Meanwhile, the average number of peptides per protein group was 4, giving high confidence in identification (**Fig2**, blue). If dia-PASEF allows for a real high analysis depth and optimal quantitative performances, performing a direct DIA search from a metaproteomic database is not realistic, both for computational power and FDR management reasons. There are a variety of strategies to generate restricted databases (proteotyping with double pass searches, Taxon spectrum mass ….), and we have deicided to evaluate the efficiency of a DeNovo based approach.h Using the PASEF data , we could map the generated sequence on all of the 57 taxa of the restricted database, and even map 73 from a bacterial reference proteome (**Fig .3**). All 57 taxa from the restricted database are included in these 73, suggesting that the approach is valid. Finally we have also tested an instrument optimized for high sensitivity to check if similar results could be obtained from lower sample amounts. Using a 30 min gradient (50% of the initial one), we could end up with a similar number of protein group and peptides while injecting only 12% of the initial amount, hence validating the process (**Fig. 4**).

Conclusions

- Using dia-PASEF with a restricted database allows to reach unprecedented depth for a direct metaproteome analysis
- A De-Novo based approach has proven efficient to map a large sequence database in order to extract a restricted one
- Similar results can be obtained from 8X less sample and half the gradient using the latest generation of timsTOF instruments
- Midia-PASEF, when operational, will be worth being tested to perform both DeNovo filtering and searches from the same dataset.

timsTOF HT / Ultra

For research use only. Not for use in diagnostic procedures.