APPLICATION NOTE

Accelerating throughput with the Proteograph™ XT Assay

Benjamin Lacar, Ting Huang, Harendra Guturu, Jian Wang, Paul Pease, Lucy Williamson, Gabriel Castro, Shadi Ferdosi, Taylor Page, Karthik Ganapathy, Khatereh Motamedchaboki, Margaret K. R. Donovan

Introduction

Plasma proteomics holds great potential for the evaluation of human health and disease states, including the early detection of life-threatening diseases like cancer. However, the large dynamic range and diversity of protein variants in plasma make it challenging to achieve simultaneous depth of coverage and throughput required for large-scale plasma proteome studies. Traditional methods, including digestion of neat or depleted plasma, or extensive peptide fractionations, require a trade-off between depth of coverage and sample analysis throughput. This limits the ability to perform large scale and deep proteomics studies. To address this challenge, Seer has developed a new two-nanoparticle (NP) suspension panel-based Proteograph™ XT workflow as part of the Proteograph™ Product Suite that facilitates deep and unbiased plasma proteomics at enhanced throughput.

This assay enables reproducible quantification of thousands of proteins in large cohort plasma studies without compromising on depth, and enhances the throughput of proteomics analysis, creating a unique opportunity to detect protein biomarkers in an unbiased and high-throughput manner.

Here, we demonstrate the performance of the new Proteograph XT workflow followed by a 1-hour per sample data independent acquisition (DIA) liquid chromatography mass spectrometry (LC-MS) workflow across a cohort of 1,790 human plasma samples, highlighting throughput, reproducibility, and depth of proteome coverage. Further, we compare the DIA LC-MS results from library-free, experiment-specific gas-phase fractionation (GPF), and Proteograph™ Analysis Suite (PAS 2.1) default library searches. Our findings suggest that while library-free searches identify more protein groups, an experiment-specific or the PAS default library results in improved data completeness, which enables higher powered studies.

Study design

The Proteograph™ XT Assay was performed on a collection of 1,790 human plasma samples. Samples were processed with three SP100 Automation Instruments using the Proteograph™ XT Assay Kit (40 samples), across 47 plates (Figure 1). The samples were processed with proprietary engineered NP suspensions. Trypsin-digested peptides were generated for downstream LC-MS analysis using a 33-minute DIA LC-MS method on one Orbitrap Exploris™ 480 MS for a total run time of 1 hour per sample, resulting in a total of ~11 weeks of whole sample preparation and data acquisition time. For each plate, additional peptides from all samples were pooled for each NP suspension to generate a sample-specific spectral library based on a gas phase

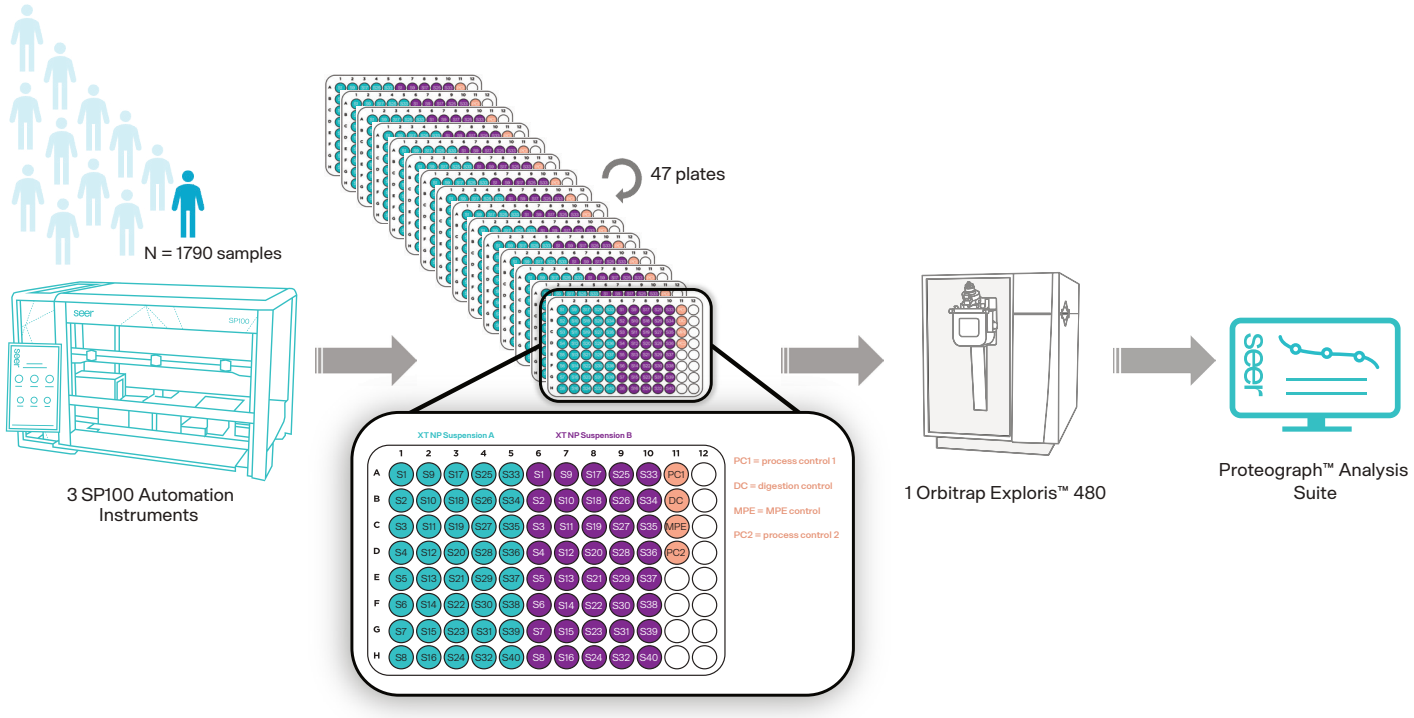

Figure 1. High-throughput Proteograph Product Suite with Proteograph XT workflow. Proteomics data from 1,790 samples were generated by running 47 plates across three SP100 Automation Instruments, followed by DIA LC-MS analysis of tryptic peptides on one Orbitrap Exploris 480 MS, and finally data analysis with the Proteograph Analysis Suite. Generation and processing of LC-MS data from all samples was achieved in ~11 weeks.

fractionation (GPF) approach, 1 DIA LC-MS data were searched, and database search performances were then compared to library-free and PAS library searches. The LC-MS raw data were analyzed using a library-free search and the following three search strategies: (i) PAS (v2.1) using DIA-NN 1.8.1, (ii) PAS 2.1 default spectral library, and (iii) GPF library search.

Methods

Sample preparation

Plasma from 1,790 individual samples and a plasma control sample (PC6), consisting of pooled citrate phosphate dextrose anticoagulant plasma from 15 healthy individuals, were processed with the Proteograph XT Assay Kit. Plasma tubes containing 240 µL of plasma were loaded onto the SP100 Automation Instrument for sample preparation to generate purified peptides for LC-MS analysis. Sample was incubated to form each of the two proprietary, physicochemically distinct nanoparticle suspensions for protein corona formation. Samples (40 samples/plate; 38-39 individual plasma samples and 1-2 PC6 samples) were automatically plated, including process controls, digestion control, and MPE peptide clean-up control. After a one-hour incubation, leveraging the paramagnetic property of NPs, NP-bound proteins were captured using magnetic isolation. A series of gentle washes removed non-specific and weakly bound proteins. This results in a highly specific and reproducible protein corona. Protein coronas are denatured, reduced, alkylated, and digested with Trypsin/Lys-C to generate tryptic peptides for LC-MS analysis. All steps were performed in a one-pot reaction directly on the NPs. The in-solution digestion mixture was then desalted and all detergent removed using a solid phase extraction and positive pressure (MPE^2) system on SP100 Automation Instrument. Clean peptides were eluted in a high-organic buffer into a deep-well collection plate and quantified. Equal volumes of the peptide elution were dried down in a SpeedVac (3 hours overnight), and the resulting dried peptides were stored at -80 °C or directly analyzed by LC-MS. Quantified peptides were reconstituted to a final concentration of 0.06 µg/µL in Proteograph XT Assay Kit Reconstitution Buffer.

LC-MS analysis

8 µL of the reconstituted peptides were loaded on an Acclaim PepMap 100 C18 (0.3 mm ID x 5 mm) trap column and then separated on an Ultimate 3000 HPLC System and a 50 cm µPAC HPLC column (Thermo Fisher Scientific) at a flow rate of 1 µL/minute using a gradient of 5–25% solvent B (0.1% FA, 100% ACN) in solvent A (0.1% FA, 100% water) over 22 minutes, resulting in a 33-minute total run time. For the MS analysis on the Thermo Fisher Scientific Orbitrap Exploris 480 MS, 480 ng of material per NP was analyzed in DIA mode using 10 m/z isolation windows from 380–1000 m/z. MS1 scans were acquired at 60k resolution and MS2 at 30k resolution.

Spectral library generation

Gas phase fractionation

We utilized a MS-only workflow that combines GPF and DIA LC-MS, saving significant experiment time while maintaining high data completeness and reproducibility.[1] This strategy generated a chromatogram spectral library with GPF deep scanning experiments, which consist of staggered m/z window analysis of the pooled peptides left over from Proteograph XT Assay plates by pooling up to 5 µL of tryptic peptides left for each sample in the plate into separate pools for each NP suspension. Six DIA LC-MS injections of 10 µL each containing a peptide concentration of 0.06 µg/µL from each NP pool were analyzed. The six injections covered mass over charge (m/z) ranges of 400–500 m/z, 500–600 m/z, 600-700 m/z, 700–800 m/z, 800–900 m/z, and 900–1000 m/z, with each injection having 50 staggered windows covering 4 m/z. MS1 was run in 60K resolution and MS2 was run in 30K resolution on another Orbitrap Exploris 480 MS with similar chromatographic (LC, trap, and column) setup. A library-free search of the DIA LC-MS data was conducted using DIA-NN 1.8.1 to create the empirically corrected GPF library.

PAS default spectral library

The PAS default spectral library was derived from data-dependent acquisition (DDA) LC-MS data collected on a Bruker timsTOF Pro from a mixture of 9 high-pH reverse-phase peptide fractions of PC2 plasma (pooled plasma from healthy and non-small-cell lung cancer subjects) and from PC2 and PC3 (pooled plasma from healthy subjects) using the Proteograph Assay with 5 NP panel and a neat digestion workflow. DDA LC-MS data analysis was conducted using MSFragger 3.6 to generate the PAS default spectral library.

Data analysis

All MS files were automatically transferred from the MS to a PAS account using the AutoUploader tool. The data were then processed using DIA-NN 1.8.1 and the PAS default spectral library. Custom bioinformatics pipelines were developed to run DIA-NN 1.8.1 with a GPF library search and a library-free search with default parameters, followed by further tertiary analyses to compare the results from the three search strategies, including comparison of peptide counts, protein counts, data completeness, and protein depth. Power analysis was computed with DIA-NN 1.8.1 library-free data using the residual variance (σ^2) estimated from a linear mixed model that was fitted on the median normalized protein intensities and adjusted for known co-variates (age and sex). All identifications are reported at 1% FDR.

Proteograph XT assay performance results

1,790 plasma samples were processed with the Proteograph Product Suite and the Proteograph XT Assay, and the resulting peptides were analyzed using DIA LC-MS method on an Orbitrap Exploris 480 MS. We investigated three library search strategies including library-free analysis, an experiment-specific GPF library created from pooled leftover peptides, and PAS default library to examine workflow performance, including protein and peptide identification rates, reproducibility, data missingness, depth of coverage, and power analysis.

Quality control

Five internal controls were run on each of the 47 plates, which are included on the Proteograph XT Assay plate, to assess the performance of each step of sample preparation and LC-MS analysis. These controls include Process Control 1, which serves as a control for the entire assay workflow for NP suspension A; Process Control 2, which serves as a control for the entire assay workflow for NP suspension B; Digestion Control, which serves as a control for the protein digestion process through peptide cleanup; MPE Control, which serves as a control for the peptide cleanup process; and MS Control, which serves as a control for LC-MS performance analysis. QC performances are monitored in PAS and shown to be stable for all five controls across the 47 plates (**Figure 2**).

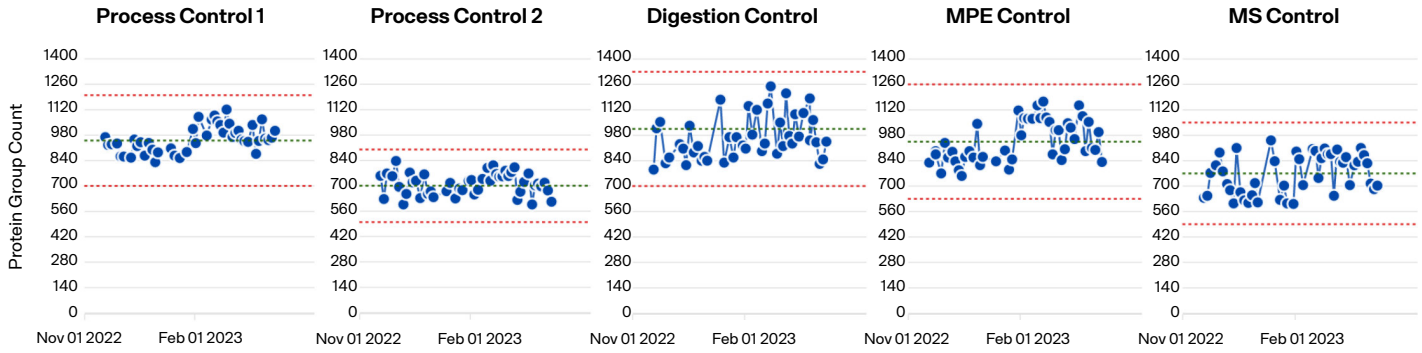

Figure 2. Quality controls assessment and Proteograph XT workflow performance monitoring. PAS provides automatic quality control evaluation of LC-MS data for five controls: process control 1, process control 2, digestion control, MPE control, and MS control. Available QC metrics include protein group counts (shown above), peptide counts, intensities, peak width, retention time, TIC, sequence coverage for FASTA-based searches, missed cleavage rate, peptide quant, and identification rates. Red dashed lines indicate the upper and lower bounds for each metric and the horizontal green dashed line highlights the mean for each metric.

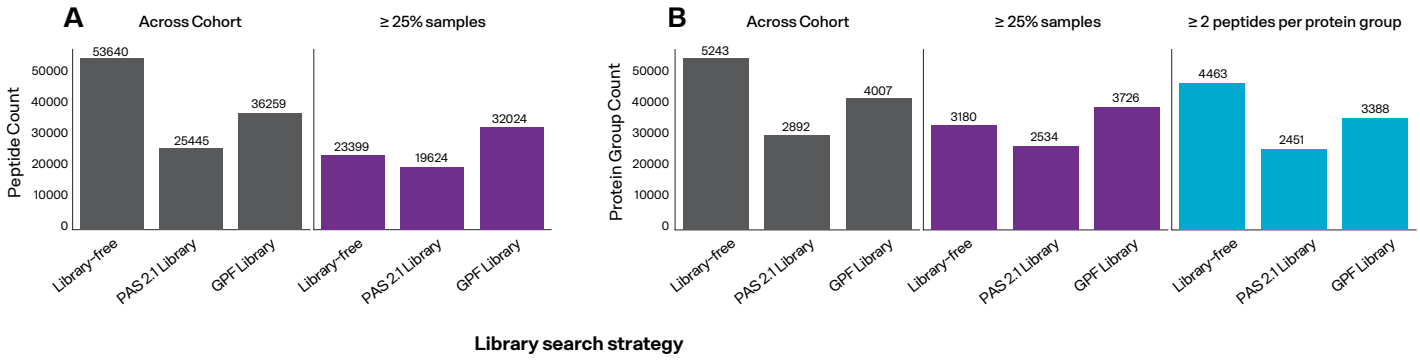

Figure 3. Protein groups and peptide identification performance. The number of identified peptides **(A)** and protein groups **(B)** by DIA-NN searching the DIA LC-MS data using library-free, PAS default library, and sample-specific GPF library. A total of ~5,200 protein groups and 54,000 peptides were identified across 1,790 plasma samples with library-free search with 1-hour per sample analysis throughput.

Protein group and peptide identification performance

In our analysis of 1,790 plasma samples, we identified 5,243 protein groups and 53,640 peptides across all samples (i.e., the total number of unique protein groups and peptides across the cohort) using the library-free search, 2,892 protein groups and 25,445 peptides using the PAS default library, and 4,007 protein groups and 36,259 peptides using the GPF library (Figure 3). When limiting protein groups to those with ≥2 unique peptides per protein, the results showed a relatively consistent level of reduction (~13-16%) in protein groups identification across the three search strategies, suggesting that the different search strategies result in comparable peptide coverage across the cohort. Conversely, limiting protein groups to those reported in at least 25% of samples shows a smaller decrease in protein group count for library-based searches (~7-12%) relative to a library-free search (~39%), indicating library-based searches result in increased data completeness across the cohort (Figure 4). Together these results indicate that library-free searches result in high protein group identification rates relative to library-based searches; however, sample-specific GPF library-based searches result in more complete protein group identification, which can impact the power of downstream statistical analyses.

Reproducibility

In our analysis of 1,790 plasma samples across 47 Proteograph XT Assay plates and across three SP100 Automation instruments, we assessed the technical coefficient of variation (CV) of median normalized intensities for peptides (Figure 5A) and protein groups (Figure 5B) found in at least 85% of plasma control samples. We observe that the majority of peptides and protein groups show a median technical CV of <40%. Moreover, we assessed the technical and biological variability of median normalized intensities for protein groups detected in at least 85% of samples (Figure 6). To evaluate technical variations, we used plasma control processed on each plate, while we employed 1,790 pre-classified plasma samples from healthy and diseased individuals with subtle protein alterations in plasma to assess biological variability. For the plasma control samples run across all 47 plates, we show a technical median CV of <37% (33%, 36%, and 31% for library-free, GPF library, and PAS default library, respectively). For biologically distinct samples run across all 47 plates, we show a median CV of >55% (59%, 58%, and 57% for library-free, GPF library, and PAS default library, respectively in this plasma cohort study). These results illustrate that biological variation surpasses technical variation even in studies involving subtle protein changes in blood plasma, thus indicating the potential to detect discernible biological differences among samples.

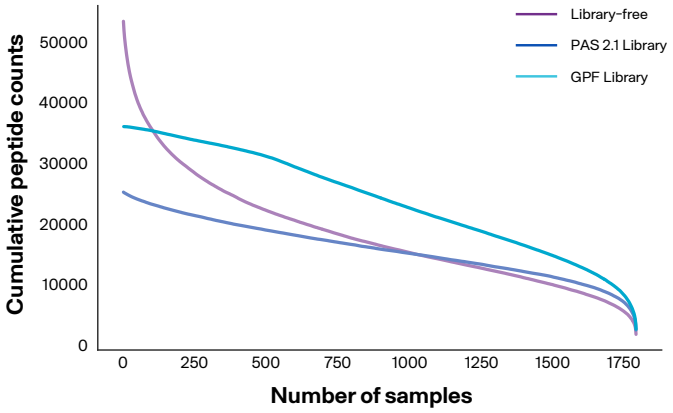

Figure 4. Proteograph XT workflow cumulative peptide coverage as a function of data completeness. Line plots showing cumulative number of peptides as a function of minimum data completeness across the cohort between searches.

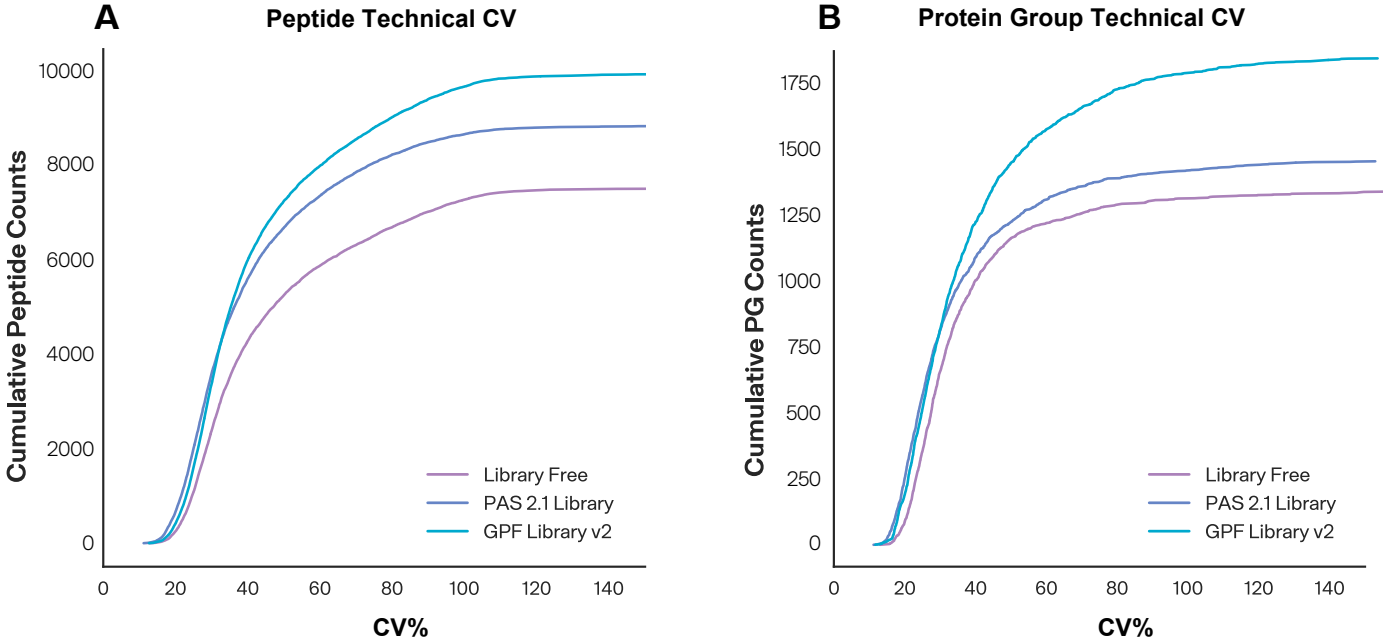

Figure 5. Cumulative distribution of peptides and protein group counts across CVs. Cumulative plot showing the number of peptides **(A)** and protein groups **(B)** identified across CVs (0–140%). The purple line indicates results from a library-free search, the dark blue line indicates results from the PAS default library search, and the light blue line indicates results from a GPF library search.

Power analysis

Given the observation that biological variation exceeded technical variation, supporting the opportunity to discover meaningful biological differences, we next aimed to estimate the minimum sample size required to detect differences across a range of fold changes with 0.7, 0.8, and 0.9 power (Figure 7). A 2-fold difference can be reliability detected with at least 46 samples or more, and a 1.5-fold difference can be reliability detected with at least 132 samples or more. These results indicate that the technical reproducibility of the Proteograph XT workflow powers the detection of small differences in protein abundance across a large-scale proteomics study.

Depth

The plasma proteome spans over 12 orders of magnitude, which is beyond the range detectable by modern LC-MS detection limits. To assess the extent to which the Proteograph XT Assay enables the assessment of proteins at both ends of the dynamic range (high and low abundant proteins), we mapped the identified proteins to the Human Plasma Proteome Project (HPPP) database3 (Figure 8). We observe that these proteins span nine orders of magnitude and the entire reported abundance range of HPPP, from highly abundant P02768 (7 x 105 ng/mL, ALB) to lowly abundant Q63HN8 and P46939 (0.004 ng/mL; RNF213 and UTRN).

Summary

- Seer's newly introduced Proteograph XT Assay empowers large-scale cohort studies while preserving optimal coverage, depth, and reproducibility.
- In a study on nearly 1,800 samples, over 5,000 plasma proteins were identified with library-free search across samples with a throughput of one hour per sample, LC-MS analysis time.
- The end-to-end workflow, encompassing sample processing and LC-MS analysis, was successfully executed on approximately 1,800 samples within a span of around 11 weeks.
- Proteograph XT Assay presents a new opportunity to expedite proteomics analytical analysis to gain enhanced biological insight.

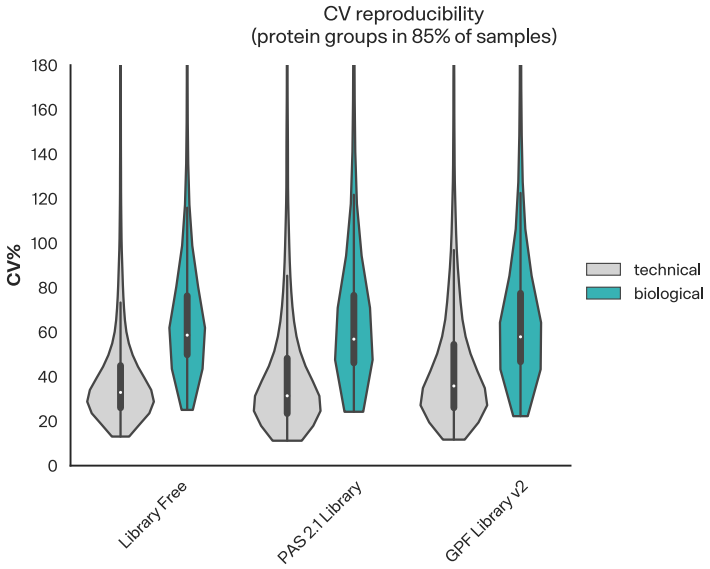

Figure 6. Reproducibility of the Proteograph XT Assay workflow in a large cohort. Violin plots display the coefficient of variation (CV) for proteins identified in a minimum of 85% of the samples. The grey violins illustrate the CVs attained between technical replicates, while the teal violins illustrate the CVs attained between biological replicates in a disease state characterized by subtle protein changes.

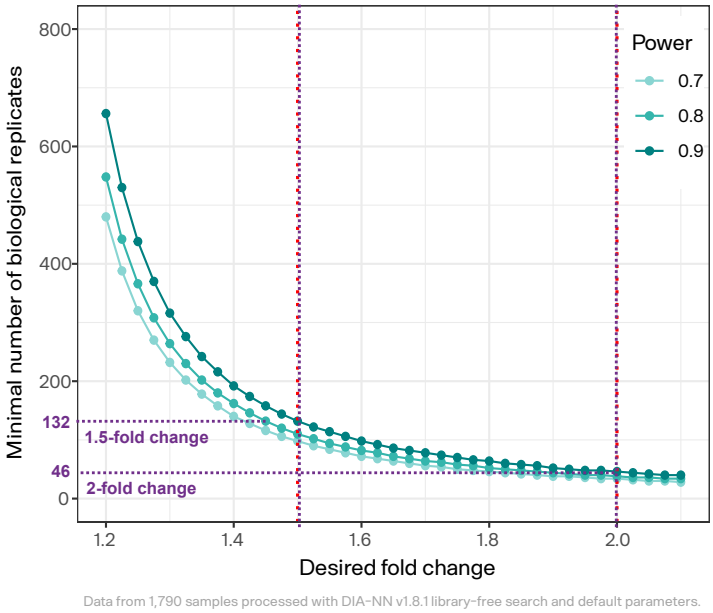

Figure 7. Power analysis for Proteograph XT workflow in large-scale discovery proteomics studies. The power analysis was conducted on 1,790 plasma samples at a false discovery rate (FDR) of 5%.

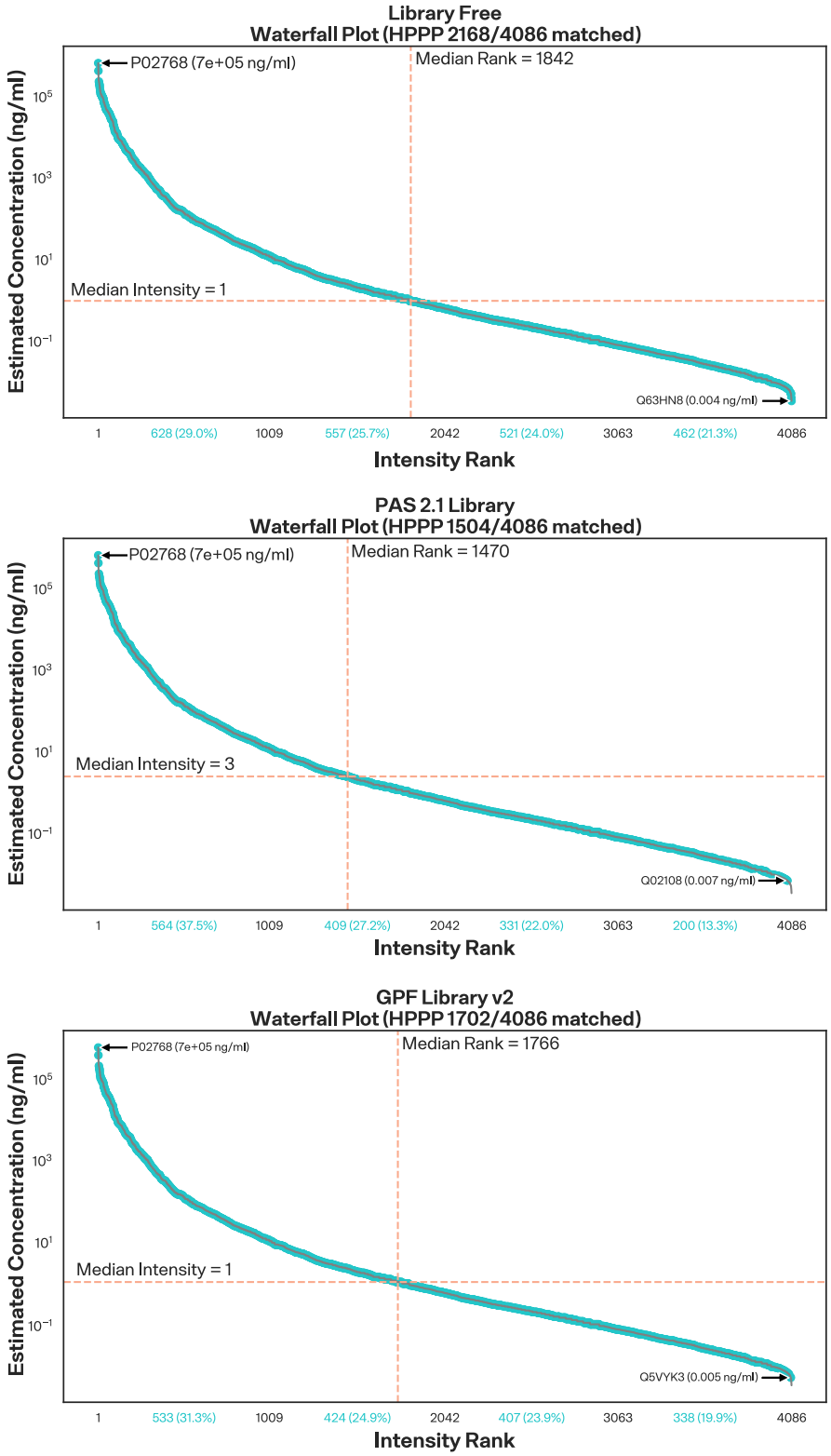

Figure 8. Depth of plasma proteome coverage based on HPPP intensity rank. The plot depicts the identified proteins and their overlap with HPPP plasma proteins, demonstrating an unparalleled depth of protein coverage and a significant boost in identification of low abundant human plasma proteins achieved using the Proteograph XT workflow.[3]

Reference

1. Pino, LK, et al. Acquiring and Analyzing Data Independent Acquisition Proteomics Experiments without Spectrum Libraries. *Mol Cell Proteomics* 19(7). 1088 –1103 (2020).
2. Keshishian, H, et al. Quantitative, multiplexed workflow for deep analysis of human blood plasma and biomarker discovery by mass spectrometry. *Nat Protoc* 12. 1683–1701 (2017).
3. Schwenk, J, et al. The Human Plasma Proteome Draft of 2017: Building on the Human Plasma Peptide Atlas from Mass Spectrometry and Complementary Assays. *J. Proteome Res* 16. 12, 4299–4310 (2017).

Find out more at
seer.bio/product/proteograph-product-suite