Unleashing the true power of DIA/SWATH data acquisition with short gradients - Rapid profiling of pre-clinical models and proteome dynamics

1. Introduction

Mass spectrometry-based proteomics and its ability to profile thousands of proteins from biological samples has evolved into the core technology to shed light on the interplay between proteins. However, achieving high-quality protein profiles with broad proteome coverage and high quantitative precision has been laborious and challenging. Firstly, protein concentrations in biological samples vary over a broader range than those of DNA or RNA. Secondly, informative protein abundance changes are often minute in response to varying biochemical conditions within cells, tissues or biofluids. Thus, achieving sufficient statistical power for the analysis of a comprehensive set of target proteins requires processing of a large number of samples. With finite resources, researchers face a difficult choice: should they aim at deeper proteomic coverage or higher sample throughput?
Ben Collins, former group leader affiliated with Ruedi Aebersold's laboratory at the ETH Zürich, and now Reader at Queen's University Belfast, strongly favors the latter. *"Of course, our goal is to obtain high-quality, quantitative measurements for as many proteins as possible. But we believe that increasing the number of samples is indispensable to obtain statistically significant, meaningful scientific content. This is especially true when we move beyond rather simplified cell-based studies where biological variation is low and – to some extent – controllable, and into the context of pre-clinical or clinical studies with potentially significant fluctuation in protein levels."*

Ben C. Collins, Queen's University Belfast

"With this faster and more robust protocol in place, we can now extend our scope and study metabolic, and potentially disease-related pathways in other tissues. The focus of our approach remains firmly on quantitative robustness. However, we are also excited to see an increasing depth of coverage in pilot studies interfacing the Evosep One with advanced mass spectrometers operated in DIA/SWATH mode."

The groundwork - Fast and standardized workflows

Higher throughput is only achievable with fast, robust and standardized procedures. Recent advances in the mass spectrometry technology and novel MS acquisition strategies have pushed the boundaries for what is possible, laying a solid foundation for clinical proteomics. The already widely used data-independent acquisition (DIA) methods, such as sequential window acquisition of all theoretical mass spectra (SWATH-MS), enable consistent quantification of thousands of proteins in hundreds of samples and exemplify some of the developments that hold great promise for boosting sample throughput. Albeit, to take full advantage of this MS performance and ensure a robust and efficient analytical setup, a fast and standardized frontend solution is highly recommended. Recent developments hold an even greater promise for future sample throughput increase. For example, the Evosep One (Bache *et al.* MCP 2018) delivers stable and consistent liquid chromatography (LC) performance over thousands of samples, even at significantly shortened separation gradients (Ludwig *et al.* Mol. Syst. Biol. 2018). Therefore, the sample throughput can be significantly increased by using DIA with the Evosep One, setting new standards for the number of samples that can be included in an experiment. When combined into a fast and standardized workflow, these state-of-the-art LC-MS methods allow measuring a complete elution profile constructed from high-confidence fragment ions in MS2 spectra for each detected peptide with high quantitative accuracy over a wide dynamic range. *"I am very optimistic that the near future will see a huge performance leap as the full potential of emerging technologies is fully explored and developed – which will be spearheaded by DIA strategies combined with short gradients,"* Ben says.

Exploring geno-phenotype relationships with high-throughput proteomics

To assess the impact of the LC gradient length on data quality, Evan Williams, a postdoc in Ruedi Aebersold's group, re-analyzed the liver tissue proteome of a genetically well-characterized mouse reference population cohort. In the previous LC-MS analyses of more than 400 samples, the team employed nanoflow LC with 60 min gradients and DIA/SWATH acquisition on a Sciex 6600 TripleTOF, allowing them to reliably quantify ~3,000 proteins and examine the effect of high- or low-fat diets on weight gain, diabetes and cardiovascular risk. When they repeated these experiments with the Evosep One, reducing the cycle time from 80 to 24 minutes and thereby increasing throughput from 18 to 60 samples per day, they still obtained 90% of information at the level of detectable proteins. In particular, they were able to retrieve known true positive quantitative trait loci (Figure 1).

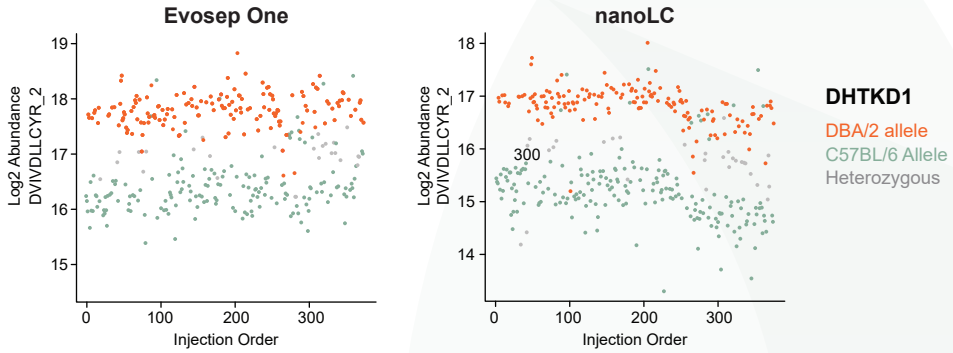

Figure 1: Reproduction of known biological experiments using the Evosep One

More importantly, the shorter cycle time plus less LC column changes and instrument cleaning allowed them to complete the measurements in one week instead of 2 months. Translating these studies from a mouse reference population with well-characterized genotypes to clinical profiling of human cohorts

with high inter-individual variability will, however, require another significant increase in sample numbers. This is because the higher population heterogeneity demands more samples measured, to correctly detect clinically relevant changes with sufficient statistical confidence. To achieve sufficient statistical power in proteomic analyses, clinical studies will need to scale up to hundreds or thousands of samples.

Questioning the depth vs. throughput dilemma

But is increasing sample throughput at the cost of quantifying fewer proteins also a reasonable strategy? In other words, if researchers accept a limited proteome depth analyzed in many more samples, can they expect to learn more with this strategy? Ting Huang, a Ph.D. student in Olga Vitek's lab at Northeastern University in Boston, took a systematic approach to this question. She analyzed data sets from several previously conducted studies, measured in tissue, cell culture, and plasma samples, using various mass spectrometry acquisition modes. In her simulations, she varied the number of detected proteins and the number of sample replicates to create data plots of the detected proteins vs. protein abundance. Her results were clear: Independent of acquisition mode, both accuracy and sensitivity decreased when she increased protein coverage by adding proteins of lower signal. Why? Because these proteins also generally displayed lower signal-to-noise ratios. On the other hand, Ting's simulations showed that analyzing larger cohorts improved both accuracy and sensitivity. *"This phenomenon is well known in machine learning and statistics, and is referred to as the 'curse of dimensionality"* Ting explains.

Ting Huang, Northeastern University, Boston

"As the number of quantified proteins increases, it becomes more difficult to distinguish the systematic patterns in protein abundance from random artifacts. A larger number of biological replicates allows us to understand better what's systematic and use it for biological investigations or disease studies."

Enabling novel applications: Profiling proteome organization dynamics

Higher throughput not only improves data quality and statistical power of biological results, but also permits extracting new layers of biological information, like analyzing protein complexes, or modular biology. *"We believe that the organization of molecules into functional modules is as important as the composition of the respective 'omes'. However, studying protein complexes in a global fashion is an immense technical challenge, especially as they dynamically react to the biochemical status of a cell."* Ben states. During their Ph.D. studies at the ETH, Moritz Heusel and Isabell Bludau jointly implemented a workflow (called complex-centric proteome profiling SEC-SWATH-MS) to detect hundreds of protein complexes from biological samples in a highly parallel manner. In this method, protein complexes are gently extracted from human cells and then lined up by decreasing size, using size exclusion chromatographic (SEC) fractionation. Each of the resulting 80 fractions is then subjected to quantitative, bottom-up DIA/SWATH mass spectrometry to derive peptide-level profiles, allowing to identify the proteins and protein complexes eluted from the columns. Data are analyzed by the program CcProfiler that implements a complex-centric data analysis strategy (Heusel *et al*. Mol. Syst. Biol. 2019). The first analysis of the HEK293 cell proteome detected 4,065 proteins, with 66% of these observed in at least one assembled state as judged by their apparent size. Some proteins were even found in multiple complexes, especially in proteasome components, ribosomal proteins and chaperones.

The study also identified novel subcomplexes and assembly intermediates of central players in the ubiquitin-proteasome-pathway and quantified the relative protein subunit distribution across them. The resulting protein SEC profiles have been published in a searchable format on the SECexplorer platform to enable other researchers to test novel or suspected models of complex formation in an interactive way. And this is only the beginning, as Ben explains: *"Eventually, we want to obtain deeper insight into modular proteome organization, including the dynamic changes observed in perturbed systems or disease states. This calls for analysis of many SEC fractions obtained from cells grown under different conditions, which is only possible if fast, standardized procedures using short gradients and MS acquisition times are applied. If we can get this working robustly, it could even become feasible to perform analysis of entire tissues."*
In a large nanoLC-based effort spearheaded Sabine Amon and Moritz Heusel in Zürich, the team at ETH Zürich further set out to compare complex association states between two cell cycle states (Heusel *et al*. Mol. Syst. Biol. 2020). This is now being accelerated using the Evosep One system by Charlotte Nicod, Claudia Martelli and Peng Xue, at the ETH. Early results indicate that analysis speed can be increased by one order of magnitude with the Evosep One, reducing analysis time from roughly 10 days to a little more than 1 day without significant loss of information at the protein complex level (Figure 2).

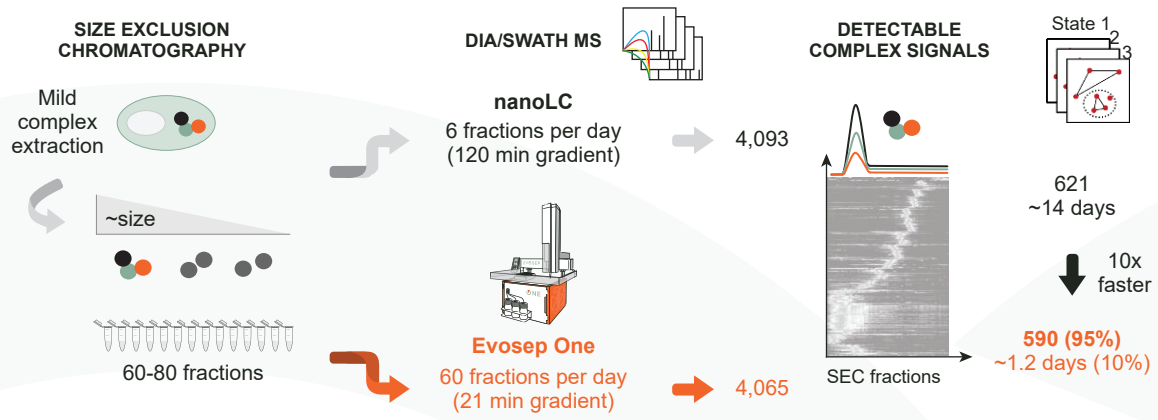

Figure 2: Complex-centric proteome profiling workflow accelerated ~10x using the Evosep One.

On the way to clinical proteomics?

So what will be the future of proteomics? As Ben summarizes: "*I believe we need a more holistic view on biology. How do we extract biologically relevant data, and what does speed buy you in terms of scientific information content? Already today, at least 50% of the MS-detectable proteome can be reproducibly measured within less than an hour per sample. Technical and scientific progress will facilitate additional types of experiments, expanding our knowledge on biological pathways and mechanisms, and also foster the dissemination of proteomics to routine and research applications.*"
Nicolai Bache, Head of Applications at Evosep, adds: *"For sure, we do not have all the answers yet but the proteomic technologies are advancing extremely fast. There is an ever-growing need and interest in clinical proteomics applications, and enabling both fast and standardized LC-MS workflows is our contribution and push in that direction."*