Why Unbiased Proteomics is Necessary

Contents

Complexity of the Human Proteome 2
Proteins in Human Biology and Disease 3
Technologies For Studying the Proteome 3
The Power of Nanoparticles is a Game Changer 5
The Proteograph Solution 6
Unbiased Methods Enable Biological Discovery in Plasma Proteomics 8
The Future is Unbiased 10
Key Takeaways 11

Methods to assess the human proteome for generating proteomic datasets for biomarker discovery or diagnostics must be reproducible, quantitative, and informative. This overview discusses the unmet need for unbiased proteomic approaches, potential shortcomings of targeted approaches, and the unique values of Seer's technology using recent data examples.

AUTHORS:

Margaret Donovan, Product Marketing Manager*
Khatereh Motamedchaboki, Associate Director, Product Marketing*
Gary Wilson, Research Scientist, Computational Proteomics
Daniel Hornburg, VP, Proteomics*
Asim Siddiqui, VP, Research & Tech Development*

*Seer, Inc. Protein Metrics, Inc.

seer

Complexity of the Human Proteome

Proteomics is the study of proteomes, the set of proteins produced in an organism, system, or biological context in any given time or state. DNA is the blueprint of living organisms, encoding instructions in RNA for proteins, which are the functional building blocks that support the molecular mechanisms of cells. With 20,000 genes encoded in the human genome, it is estimated that there may be over one million distinct protein variants, or proteoforms,[1] in each cell type, emerging from a broad range of modifications that modulate their structure and function.

Within the past decade, access to deep and unbiased genomic information at scale has resulted in decoding millions of genomes and exomes, identifying hundreds of millions of genetic variants. However, the functional context for most of this DNA information is not well characterized at the RNA or protein level.[2] Part of the reason for this knowledge gap is that biochemical complexity increases and becomes more dynamic from the genome to the proteome (Figure 1).

Proteomics can provide a wide range of biological insights, including how proteins change in abundance, how they are modified after they are translated from mRNA, where they are in cells and how they get there, which ones are interacting with others and much more. In other words, the genome tells us 'what could be', whereas the proteome tells us 'what is'.

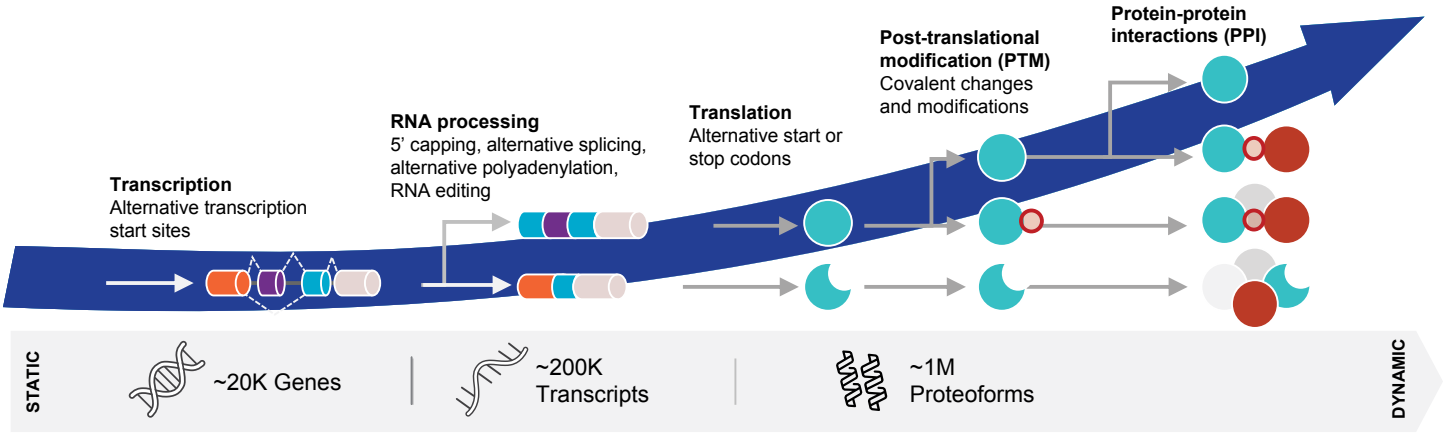
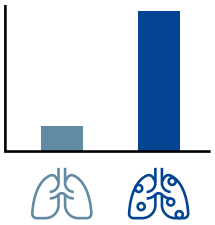


Figure 1. While the genome is a static indicator of risk, the proteome is a dynamic indicator of status.[3] Starting with the human genome on the left, biological complexity increases as it approaches the proteome, going from ~20,000 genes to over one million proteoforms. Many protein variants can be produced from a single gene during protein synthesis, and proteins are further modified creating a greater number of variants. Proteins also interact with each other, which has a significant impact on health and disease.

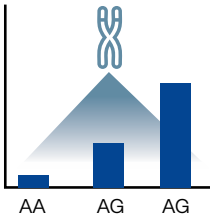
Proteins in Human Biology and Disease

Currently, most proteomic studies are focused on fundamental discovery and deciphering how biology works.[4] For human health, the promise of proteomics is to improve care by correlating specific proteoforms with a given disease state.[5] With less than one percent of functional protein mapping completed, proteomics has the potential to transform our understanding of human biology and disease at many levels, from individual cells to populations (Figure 2).



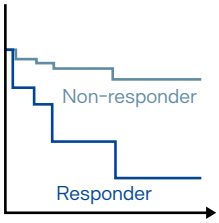
Early Detection of Disease

Proteins can be present in different amounts in diseased and healthy samples. For example, cancer researchers are using proteins to detect disease using a blood draw to identify cancer at the earliest stages when it is most treatable.



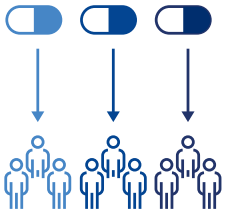
Molecular Implications of Genetic Variation

Genome-wide association studies have discovered thousands of risk loci associated with diseases. Unbiased proteomics can help reveal how genetic variation affects protein function, leading to a deeper understanding of human health and disease.



Drug Target Discovery

By identifying protein biomarkers driving diseases, researchers can refine which biomarkers to target and predict which patients aremost likely to be the best responders to a given therapy.



Personalized Medicine

With the addition of comprehensive, unbiased proteomics informing the dynamics of molecular mechanisms driving biology, a wider view of the complex landscape may offer more precise insights to guide treatments.[6]

Figure 2. Proteomics has the potential to transform our understanding of human biology and disease in several ways.

Technologies For Studying the Proteome

In addition to increased biological complexity, the proteome is also harder to access than the genome. In samples such as blood plasma, there exists a huge dynamic range of proteins, and the ability to accurately survey the least abundant but perhaps most interesting proteins remains a challenge. To generate proteomic datasets that can be used for biomarker discovery or for diagnostics, prognostics, and monitoring, methods to assess the human proteome must be both reproducible and quantitative. Two common methods for studying proteomics use targeted, affinity-based technologies and deep, mass spectrometry-based technologies.

Targeted, Affinity-based Technologies

Targeted, affinity-based approaches rely on analyte-specific reagents (ASR) such as antibodies or aptamers that bind to known protein domains and are used to screen for specific proteins for which the ligands were designed.

While affinity-based assays are built upon established techniques and can address problems associated with accessing low abundant proteins, several critical challenges exist (Figure 3).

Targeted assays are performed at the epitope level and distinguishing between proteoforms would require a panel design targeting all potential variants. Because targeted technologies cannot distinguish between all potential proteoforms, there is a growing belief that the presence of a protein-altering variant, degradation, or post-translational modifications may cause some epitopes to be inaccessible to ASRs and miss protein variants.[7] Additionally, ASR specificity can be impacted by assay variables including ionic strength, temperature, cross-reactivity, and protein-protein or protein-DNA interactions.

The most significant one is that researchers are limited to examining a catalogue of predefined proteins, introducing bias into the results by only assaying abundances of proteins within this set and excluding the possibility of discovering novel or unexpected biomarkers.

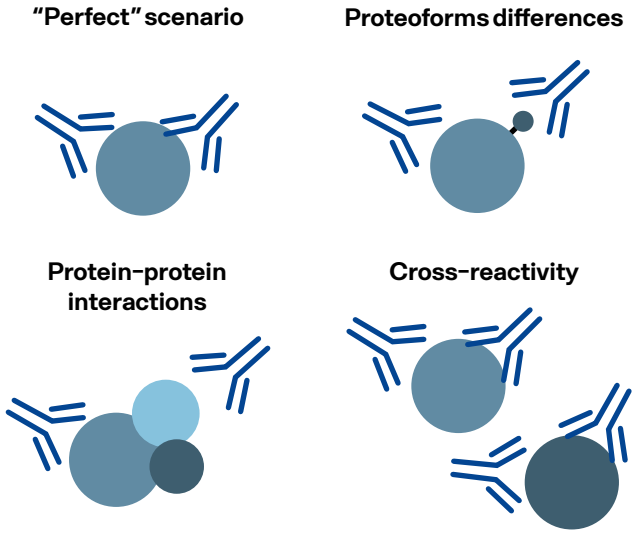


Figure 3. Limitations of affinity-based proteomics assays. From left to right; the "perfect" scenario with ASR epitopes available and no cross reaction, situations where the epitopes are unavailable and the ASR is not able to bind to proteoform differences or protein-protein interactions, and an example of cross reactivity when the ASR targets non-specific epitopes.

Deep, Mass Spectrometry-based Technologies

Mass spectrometry (MS)-based proteomics enable direct, unbiased analysis of the proteome through the detection and quantification of peptides obtained from digested proteins. This means the entire proteome can be examined, rather than looking only at specific proteins. With an unbiased, deep approach, you are not limited by what is thought to be known—and the more samples you study, the more you discover.

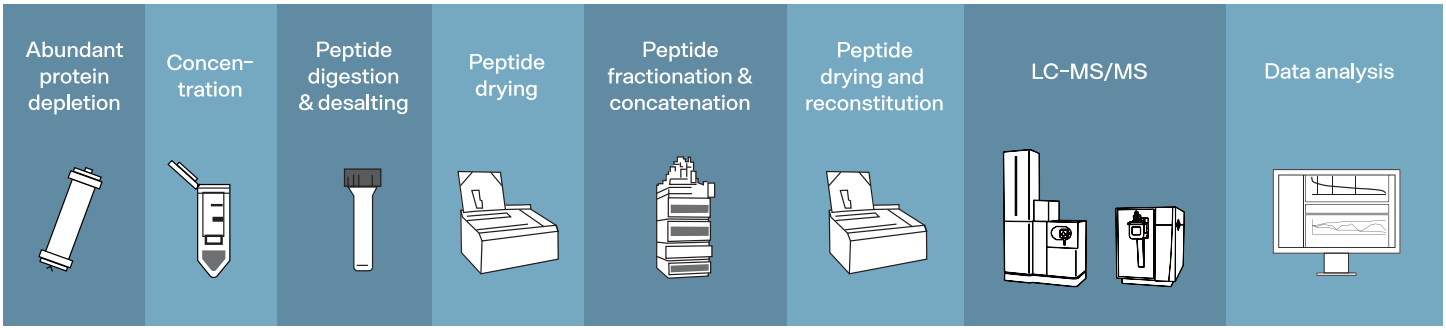


Figure 4. Conventional, deep MS-based plasma proteomic methods require complex depletion, concatenation, and fractionation steps that are difficult to scale and reproduce.[5]

Importantly, bottom-up MS-based proteomics approaches are peptide-centric, meaning researchers can differentiate between protein isoforms and detect post-translational modifications from a corresponding MS signal.

Mass spectrometry sample preparation, however, is not without its own challenges. Today's conventional, deep MS-based plasma proteomic methods do allow broad exploration across the dynamic range of the proteome but rely on fractionation and depletion steps to access lower abundance proteins and increase the sensitivity of detection (Figure 4).[8] These methods increase the number of times samples are handled, negatively impacting throughput and reproducibility, and create challenges for

normalizing data and scaling experiments. As a result, access to the deep proteome is limited, and these traditional workflows take months to complete and do not scale. Complex, labor-intensive workflows also prevent wide use in studies involving many samples and/or large patient cohorts.[9]

The Power of Nanoparticles is a Game Changer

Biology often happens at the log scale and exploring the proteins in blood plasma is no exception. The top 22 most abundant proteins in plasma account for over 99% of the proteome mass, whereas less abundant proteins are likely to contain biologically meaningful information.[10, 11] The technological challenge of reproducibly accessing proteomic information across the large dynamic range of the plasma proteome at scale has recently been improved through Seer's proprietary engineered nanoparticles (NPs).[12]

To overcome the limitation of deep plasma proteomics in large cohorts, Seer has developed a fast and scalable technology that employs protein-nano interactions. Introducing a proprietary engineered Seer nanoparticle into a biofluid such as blood plasma leads to the formation of a selective, specific, and reproducible protein corona at the nano-bio interface[13] driven by the relationship between protein-nanoparticle affinity, protein abundance and protein-protein interactions.

Once NPs are introduced into a biofluid sample, proteins assemble on the nanoparticle surface to form a protein corona via physical adsorption and/or electrostatic interactions. Before equilibrium, corona composition is governed by the relative concentration of proteins that diffuse to interacting moieties on the particle surface. At equilibrium, high-abundance, low-affinity proteins on the nanoparticle surface are displaced by low-abundance, high-affinity proteins, compressing the dynamic range.[14] The quantitative composition of protein coronas depends on the physicochemical properties of the nanoparticle, the presence and abundance of proteins with compatible surface epitopes, and the competition of proteins.

Nanoparticles have been explored for use in medicine and drug delivery for decades, providing a wealth of experience and knowledge for various designs and reproducible engineering, making them an attractive tool for proteomics.[15, 16] Importantly, Seer's proprietary engineered nanoparticles provide physicochemically distinct interaction domains, attracting a broad range of proteins. For instance, charge properties can be combined with hydrogen donors to provide high affinity to a large portion of proteins and render them more visible to any downstream detector.[17] As such, a panel of NPs has been designed to perform deep sampling of the entire plasma proteome, with application for different biosample types with similar dynamic range challenges. Synergistic panels of NPs can be engineered to reproducibly interact with thousands of proteoforms by their inherent ability to form complex protein coronas (Figure 5).[17, 18]

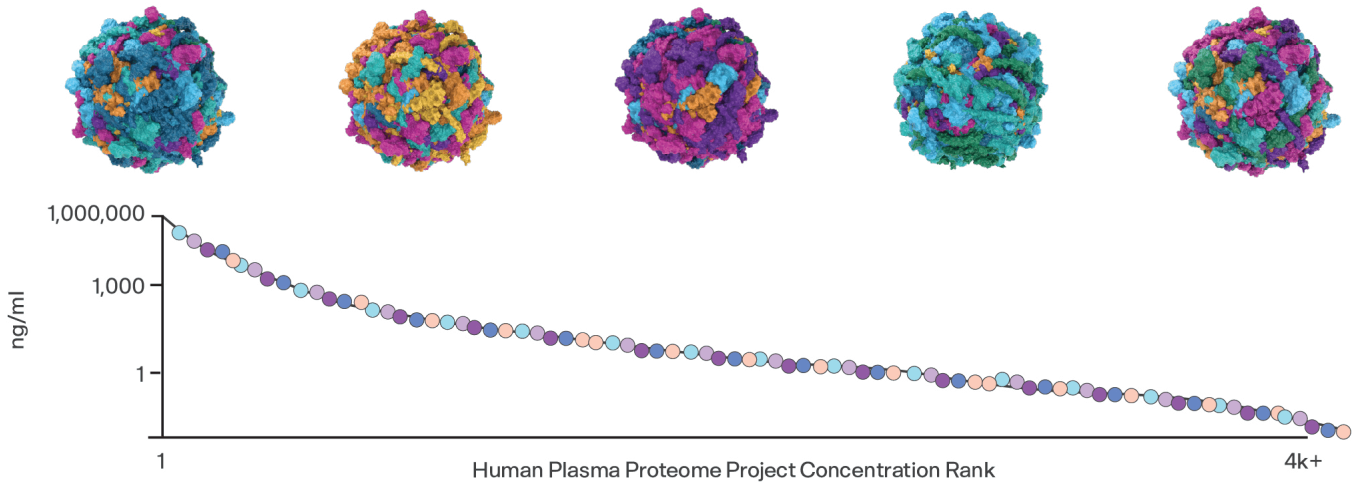


Figure 5. Seer's proprietary nanoparticles form stable protein coronas due to diverse physicochemical properties that interrogate the dynamic range by affinity.[19]

The Proteograph™ Solution

Seer offers the first and only automated platform for proteomic studies that enables sampling across the entire plasma proteome's wide dynamic range, allowing discovery of novel protein variants with the reproducibility and speed needed for high throughput studies.

The Proteograph Product Suite is designed for researchers performing discovery proteomics experiments with biofluids such as plasma, serum, urine, cerebrospinal fluid (CSF), or conditioned culture media. The Proteograph Product Suite includes the SP100 Automation Instrument, the Proteograph Assay Kit with quality controls, and the Proteograph™ Analysis Suite (PAS) software, and requires the use of an existing LC-MS instrument provided by the user (Figure 6).

The NP-protein corona-based proteomics workflow enables more precise quantification and deeper sampling of the proteome compared to conventional deep workflows. The NP design and engineering process can be guided by machine learning to diversify the proteomics content captured by a panel of NPs and to survey specific families of proteins.[18] Protein corona formation is a function of affinity and protein concentration in the sample, enabling precise and reproducible surveys of proteomic content (Figure 7).

The automated Proteograph workflow is optimized to run 16 plasma samples in 6 hours and 45 minutes and includes quality controls. The SP100 Automation Instrument includes optimized methods for automated peptide quantitation and reconstitution. Peptides can be analyzed on nearly any LC-MS instrument optimized for deep proteomics experiments and data files can be automatically uploaded into PAS for analysis and visualizations.

Deep coverage of the proteome across the dynamic range allows for identification of protein groups and the discovery of protein variants across a cohort. The Proteograph is not constrained in the protein variants it can detect, so it is able to pick up variation across very low to very high allele frequencies, as well as identifying different isoforms, PTMs, and other variants.

As more studies describing the vast complexity of the proteome are being published, it is becoming apparent that conventional proteomics technologies are not able to reach the depth, breath, and scale that are needed to explore the proteome efficiently.

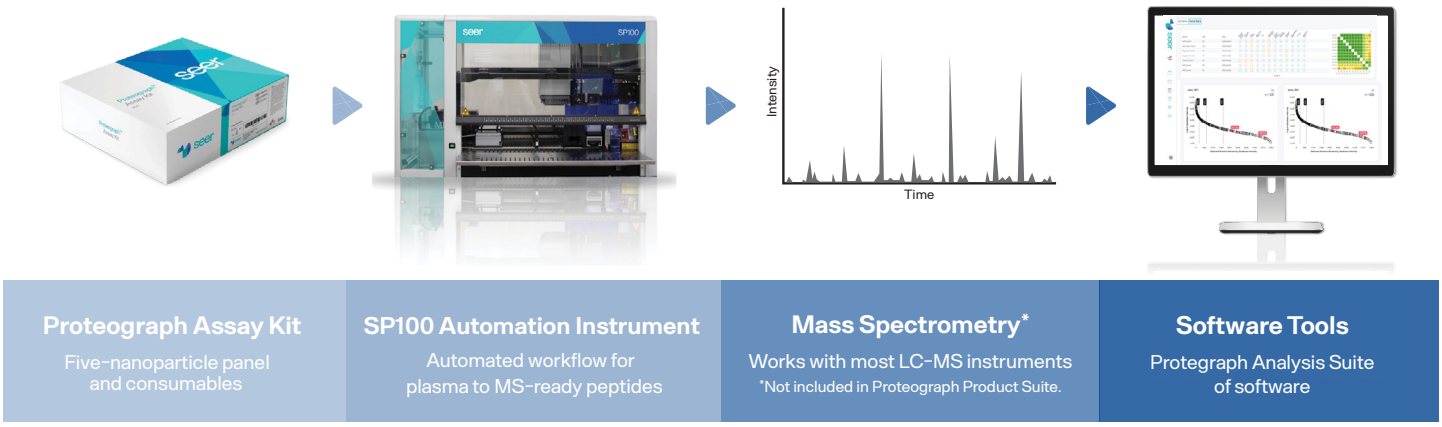


Figure 6. The Seer Proteograph Product Suite consists of the Proteograph Assay Kit containing all reagents and a five-nanoparticle panel, the SP100 Instrument that automates peptide preparation, LC-MS on the user's existing instrument, and data analysis using the Proteograph Analysis Suite software.

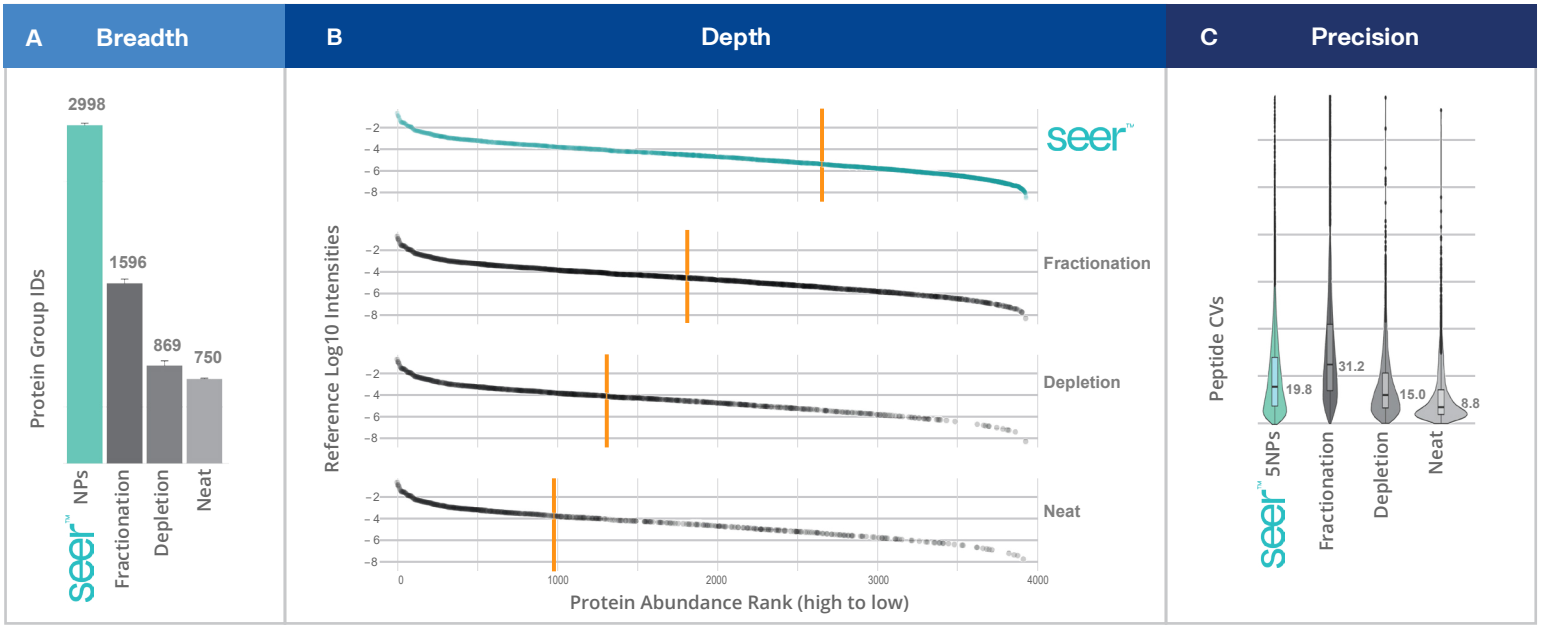


Figure 7. The Proteograph Product Suite yields more low abundant plasma proteins with higher precision. Fractionation of 19 high pH fractions of digested, depleted plasma concatenated to 9 injections. All samples were run with a 30-minute gradient and processed using DIA-NN software in library-free mode. **A:** A panel of five NPs identified ~3,000 protein groups compared to deep fractionation, depletion, and neat plasma, respectively. **B:** A panel of five NPs cover more protein at lower intensity than alternative workflows.[10] Vertical orange bars represent 75% percentile. **C**: A panel of five NPs has lower a coefficient of variation (%CV) than deep fractionation methods. Neat plasma has the lowest CV because of the presence of abundant proteins.[20]

Unbiased Methods Enable Biological Discovery in Plasma Proteomics

Unbiased approaches identify protein isoforms and provide a deeper, more nuanced assessment of the human proteome, supporting enhanced understanding of health and disease.

Examples:

1 Discovering Disease-associated Proteoforms in Non-Small Cell Lung Cancer

Unbiased techniques using a peptide-centric approach can identify proteoforms that otherwise would not be identified using protein affinity-based targeted technologies. Specifically, assessment of the plasma proteome at the peptide level reveals unique biological insight compared to protein-level evaluation. Four proteins with putative isoforms were identified. Importantly, none of these proteins showed a difference in abundance at the protein level. For BMP1, a long isoform associated with healthy patients and short isoform associated with NSCLC patients were identified. While not yet well-characterized, BMP1 is known to play a role in collagen processing, which has been shown to be linked to cancer. These findings may thus reveal BMP1 isoform-specific roles in cancer (Figure 8).[21]

2 Unbiased Biomarker Discovery in Alzheimer's Disease

To evaluate early detection of biomarkers from blood plasma, the Proteograph Product Suite was used to perform unbiased biomarker discovery research with a previously classified cohort of Alzheimer's Disease (AD) patients. Preliminary data from 100 Alzheimer's patient plasma samples and 100 healthy control plasma samples were analyzed using a Data Independent Acquisition (DIA) MS method.[22]

As shown in Figure 9, protein signals from AD and control samples enable high performing classification of health state, including important features that are both previously AD-linked biomarkers and potentially novel AD biomarkers.

3 PTM Identification: Glycosylation

Protein glycosylation is an important category of post-translational modifications that can have dramatic effects on protein structure and function. A number of studies suggest that identifying novel protein glycoforms may have critical implications in biomarker discovery and drug target identification.[23, 24] Improved methods for profiling the plasma glycoproteome can accelerate biomarker discovery, as protein glycosylation states can provide diagnostic evidence and a complementary insight when the total protein abundance is not informative.

For discovery of novel protein biomarkers, Seer's proprietary engineered nanoparticles are key for robust and reproducible measurement of the plasma proteome. Since nanoparticle coronas differentially interrogate complex samples at the proteoform level and plasma proteins are often glycosylated, researchers from Protein Metrics investigated whether this strategy could also provide insight to the plasma glycoproteome without the need for glycopeptide-specific enrichment. Glycoproteins and glycopeptide coverage with NPs and their replicate LC-MS analysis reproducibility is shown in Figure 10, demonstrating that the Proteograph Product Suite can be used to identify a range of plasma protein post-translational modifications, including differentially glycosylated proteoforms.

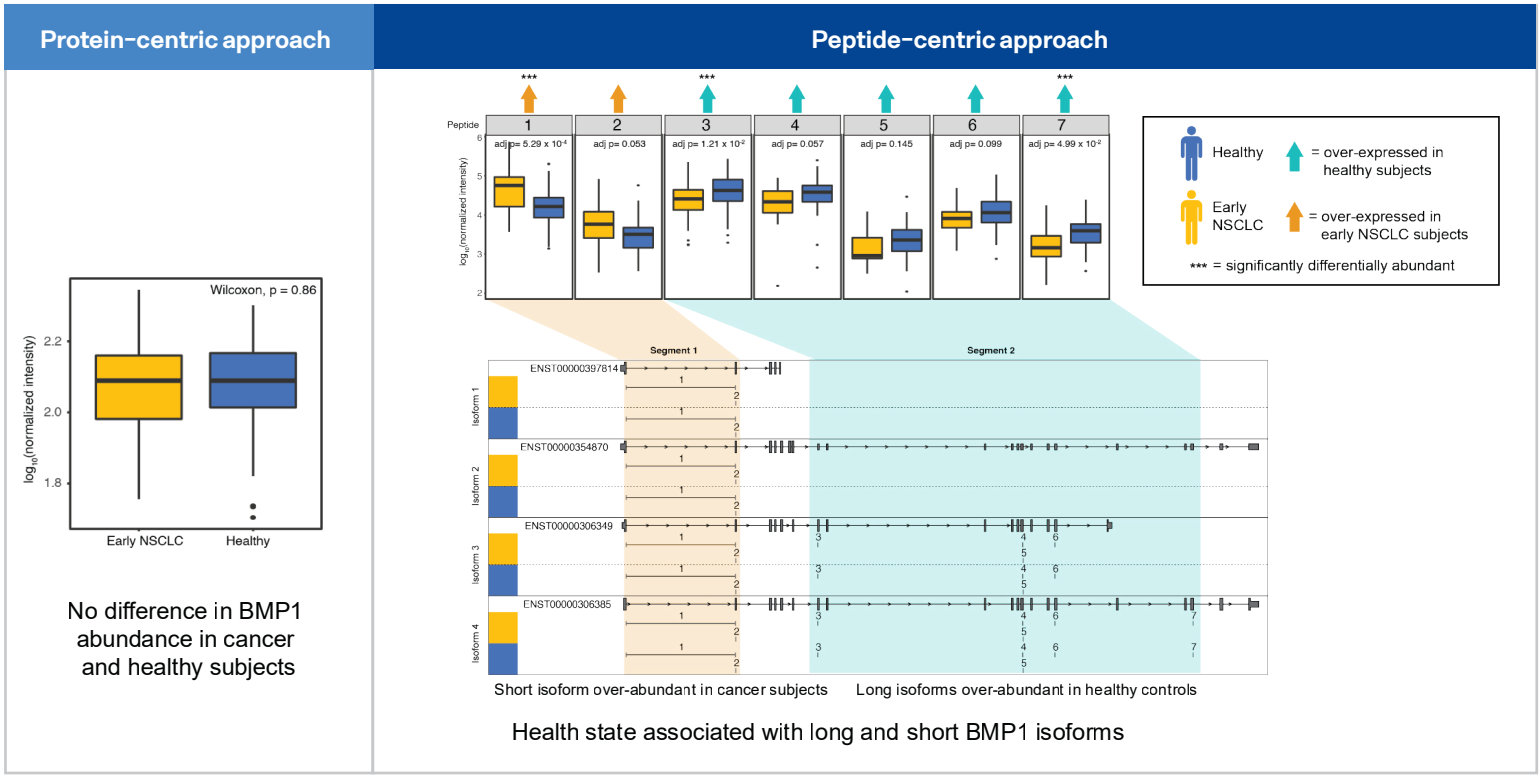


Figure 8. BMP1 has four protein coding isoforms, three of which are substantially longer than the other two isoforms covering additional exons. The peptides mapping to exons that cover all four protein isoforms have higher abundance in NSCLC patients relative to controls, whereas peptides mapping to exons that cover only the three longer isoforms have higher abundance in researcher controls.

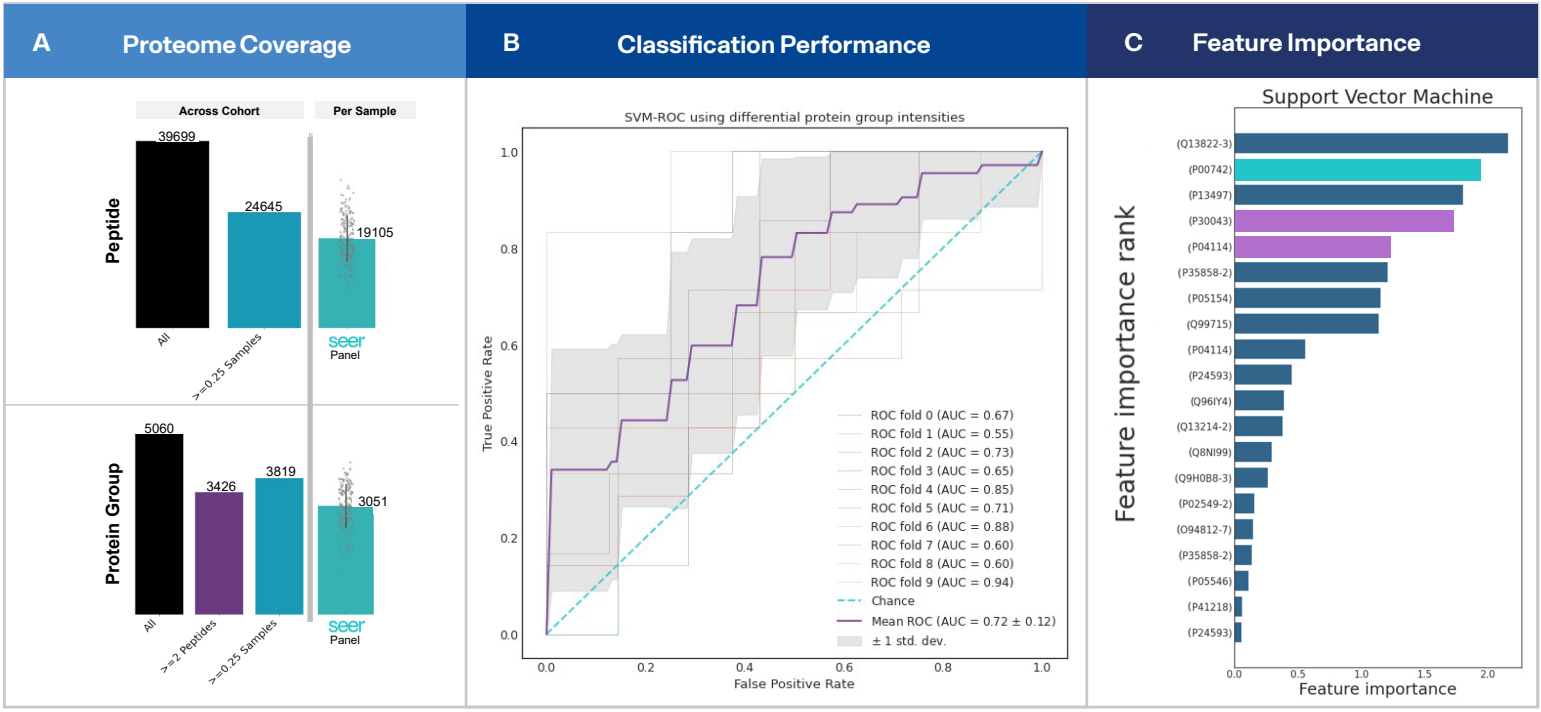


Figure 9. Proteome coverage across 100 AD patients and 100 healthy control patients using a Data Independent Acquisition (DIA) MS method. **A**. Using the Proteograph Product Suite, ~5,000 protein groups and ~40,000 peptides were identified across all 200 samples. **B**. Classification performance was measured as averaged area under curve (AUC) of 72% across 10 cross validation folds. **C**. The top 20 most important features, including features associated with known Alzheimer's Disease markers highlighted in purple25,26 and teal27, as well as other novel features in blue.

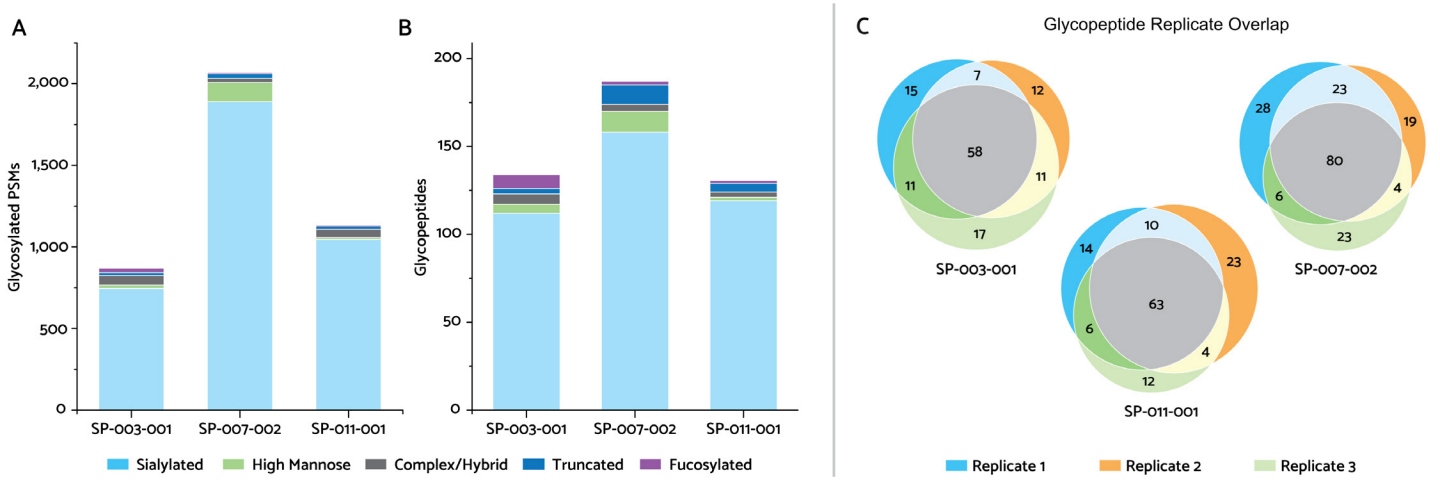


Figure 10. A and B. The number of glycosylated peptide spectrum matches and peptides in Byonic™ search results are broken down by glycan type. **C**. The overlap of glycopeptide identifications between individual replicates is displayed for each NP. Figure reproduced with permission from Protein Metrics, Inc.

The Future is Unbiased

Advances in unbiased, mass spectrometry-based workflows are pushing the boundaries of our understanding of human biology and disease. Unbiased and deep coverage across the broad dynamic range, rapid generation of data, and large-scale studies will bring us to the next generation of plasma proteomics, where we can look forward to:

Unbiased pQTLs:

Large cohorts such as biobanks have identified thousands of disease risk loci, but we still need to understand what molecular mechanisms these variants impact. Previous studies have linked genetic variation and protein abundance, but also have highlighted the inability of targeted technologies to distinguish between many proteoforms. Importantly, the role of protein-altering variants in causing some epitopes to be inaccessible to ASRs is not well understood and may confound results. With an unbiased study at the proteoform level, an improved picture of the impact of genetic variation on protein abundance emerges, giving a true picture of molecular underpinnings of human biology and disease.

Multi-omic, systems-level view of human biology:

Over the past 15 years there has been and continues to be an explosion of petabytes of genomic data. Using technologies like the Proteograph Product Suite, it is likely that unbiased, deep proteomics will rapidly catch up to deliver even more biological insight than genomics alone, with potential to be combined with other data types, including transcriptomics, metabolomics, and lipidomics. Multi-omics approaches will provide a systems-level view of human biology, shedding light on more complex interactions, such as the interplay between different molecular families.

Precision medicine:

To understand the complexity of biology, researchers need sensitive, scalable tools to see small changes and how combinations of changes are significantly associated with diseases in large cohorts. Therefore, looking for small or early trends in unbiased proteomic data may provide clues for early interventions to avoid acute or chronic disease.

Key Takeaways

- The genome is the plan for what the cell could do; the proteome provides dynamic information on what the cell does.
- Mass spectrometry (MS)-based proteomics enable direct, unbiased analysis of the human proteome through the detection and quantification of peptides obtained from digested proteins.
- Targeted methods have substantial challenges with specificity, bias, and inability to distinguish between proteoforms.
- The unique synergy of Seer's NP-based proteomics workflows coupled to unbiased LC-MS readouts to perform peptide-centric analyses, large-scale, rapid studies, and proteoform detection is not achievable by targeted affinity-based proteome screening technologies.
- Seer offers the first and only automated platform for proteomic studies that lets you sample across the entire proteome, enabling the discovery of novel plasma protein variants with the reproducibility and speed needed to do this at scale.

For more information about unbiased, deep, and rapid proteomic approaches, visit:

seer.bio/product/proteograph-product-suite

References

1. Abersold, R. et al. How many human proteoforms are there? *Nat Chem Biol* 14(3):206-214 (2018).
2. Karczewski, K. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 581,434-443 (2020).
3. Bludau, I. et al. Proteomic and interactomic insights into the molecular basis of cell functional diversity. *Nat Rev Molec Cell Biol* 21, 327–340 (2020).
4. Crutchfield, C.A. et al. Advances in mass spectrometry-based clinical biomarker discovery. *Clin. Proteomics* 13, 1 (2016).
5. Ignjatovic, V. et al. Mass spectrometry-based plasma proteomics: Considerations from sample collection to achieving translational data. *J Proteome Res*. 18(12):4085-4097 (2019).
6. Geyer, P.E. et al. Plasma proteome profiling to assess human health and disease. *Cell Syst*. 2, 185–195 (2016).
7. Pietzner, M. et al. Synergistic insights into human health from aptamer- and antibody-based proteomic profiling. *Nat. Commun.* 12 (2021).
8. Geyer, P.E. et al. Revisiting biomarker discovery by plasma proteomics. *Mol Sys Biol* 13(9):942 (2017).
9. Anderson, N.L. et al. The human plasma proteome: History, character and diagnostic prospects. *Mol Cell Proteomics* (11)845-67 (2002).
10. 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).
11. Nanjappa, V. et al. Plasma proteome database as a resource for proteomics research: 2014 update. *Nucleic Acids Res.* 42 D959-65 (2014).
12. Blume, J.E. et al. Rapid, deep and precise profiling of the plasma proteome with multi-nanoparticle protein corona. *Nat. Commun.* 11 3662 (2020).
13. Vroman, L. et al. Interaction of high molecular weight kininogen, factor XII, and fibrinogen in plasma at interfaces. *Blood* 55 (1): 156–159 (1980).
14. Video: Proteograph Product Suite: An automated workflow that scales with your studies.
15. Bertrand, N. et al. Mechanistic understanding of in vivo protein corona formation on polymeric nanoarticles and impact on pharmacokinetics. *Nat. Commun.* 8, 777 (2017).
16. Liu, Y. et al. Nano-bio interactions in cancer: From therapeutics delivery to early detection. *Acc Chem Res.* 54 291-301 (2021).
17. Hornburg, D. et al. Enhanced competitive protein exchange at the nano-bio interface enables ultra-deep coverage of the human plasma proteome *Adv. Mater.* (2022).
18. Ferdosi S., et al. Engineered nanoparticles enable deep proteomics studies at scale by leveraging tunable nano-bio interactions. *Proc. Natl. Acad. Sci USA* (2022).
19. Schwenk J.M. 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, 4299-4310 (2017).
20. Human Plasma Sample Preparation Workflow Using the Seer Proteograph™ Product Suite.
21. Donovan, M, et al. Peptide-centric analyses of human plasma enable increased resolution of biological Insights into non-small cell lung cancer relative to protein-centric analysis. *bioRxiV* (2022).
22. Siddiqui, A et al. Deep and unbiased plasma proteomics analysis of Alzheimer's and mild cognitive impairment subjects with a novel multi-nanoparticle approach. *YouTube* (2022).
23. Zhang, Q. et al. Integrative glycoproteomics reveals protein N-glycosylation aberrations and glycoproteomic network alterations in Alzheimer's disease. *Science Advances* 6(40) (2020).
24. Costa, A.F. et al. Targeting glycosylation: A new road for cancer drug discovery. *Trends Cancer* 6(9)757-766 (2020).
25. Mueller, C. et al. The heme degradation pathway is a promising serum biomarker source for the early detection of Alzheimer's disease. *J Alzheimers Dis.*19(3)1081-91 (2010).
26. Picard, C. et al. Apolipoprotein B is a novel marker for early tau pathology in Alzheimer's disease. *Alzheimers Dement.* 18(5) 875-887 (2022).
27. Begic, E. et al. Increased levels of coagulation factor XI in plasma are related to Alzheimer's disease diagnosis. *J Alzheimers Dis.* 77(1)375-386 (2020).