Deep, Rapid and Unbiased Plasma Proteomics with Differential Analysis of Proteoforms Enabling Proteogenomic Studies in a NSCLC Lung Cancer Study

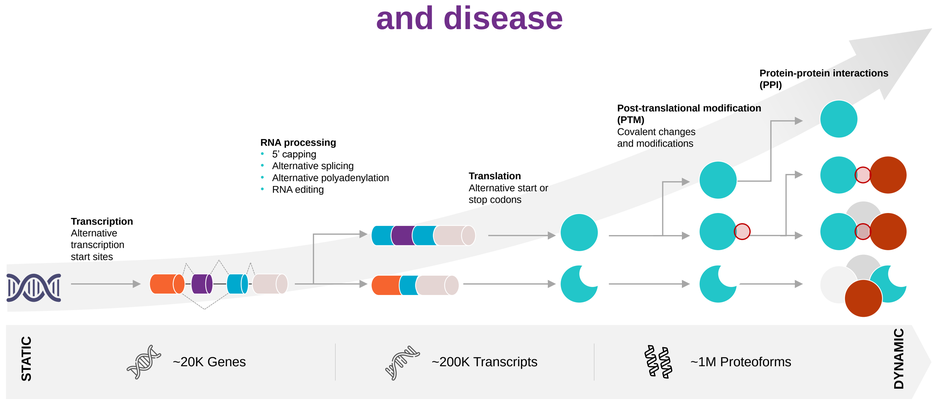
Henry Huang*, Margaret K. R. Donovan, John E. Blume, Jian Wang, Marwin Ko, Ryan W. Benz, Theodore L. Platt, Juan C. Cuevas, Serafim Batzoglou, Asim Siddiqui and Omid C. Farokhzad

Seer, Inc., Redwood City, CA 94065, USA

Proteomes are Dynamic and Far More Diverse than Genomes

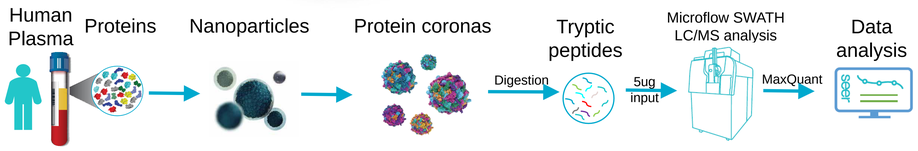
Comprehensive assessment of the human proteome remains elusive due to multiple forms of a protein, each of which can serve distinct functions, arising from alternative splicing, allelic variation, and protein post-translational modifications. Characterization of the variable protein forms, or proteoforms, will expand our understanding of the molecular mechanisms underlying disease, however identification of these variable forms requires unbiased protein coverage at sufficient scale. Scalable, deep, and unbiased proteomics studies have been impractical due to cumbersome and lengthy workflows required for complex samples, like blood plasma. Here, we demonstrate the power of Proteograph™ Product Suite in a proof-of-concept proteogenomic analysis of 80 healthy controls and 61 early-stage non-small-cell lung cancer (NSCLC) samples to dissect differences between protein isoforms arising from alternative gene splicing, as well as the identification of novel peptides arising from allelic variation.

Proteoforms are critical to understanding human health and disease



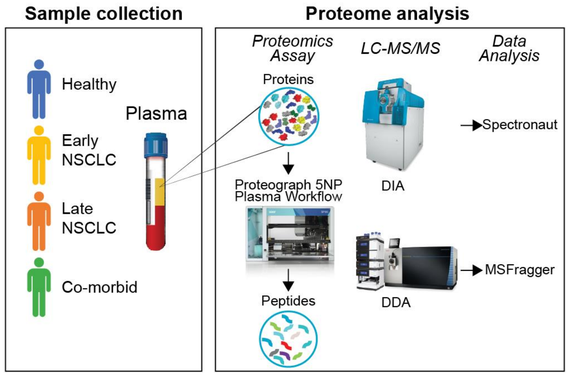
Proteograph solution

Proteograph Product Suite provides unbiased, deep, and rapid proteomics at scale



Proteoform inference methods

Using peptides with significant abundance differences ($p < 0.05$; Benjamini-Hochberg corrected), we extracted proteins comprised of peptides where at least one peptide had significantly higher plasma abundance, and another significantly lower plasma abundance in controls vs. cancer, resulting in a set of four proteins. To identify biologically relevant protein variants, we performed exome sequencing on 29 individuals from the NSCLC study and utilized it as a personalized mass spectrometry search libraries for each individual. Such personalized database allows for peptide search results from standard searches in addition to capturing novel and personalized peptides.



Proteoform Inference in a Non-small Cell Lung Cancer Plasma Proteome Study

Peptide-level analysis provide unique biological insights versus protein-level

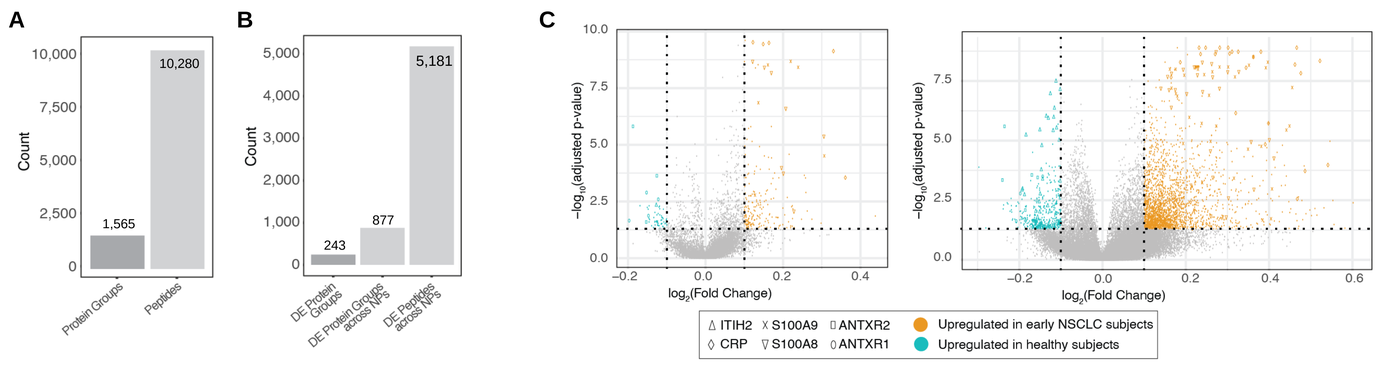


Figure 1. A) Peptides and proteins groups counts. **B)** Number of differentially abundant protein groups (collapsed NPs); protein groups (across NPs); and peptides (across NPs). **C)** Volcano plots showing differentially abundant protein groups (left) and peptides (right), across nanoparticles (NPs) between healthy and early NSCLC subjects.

Inference of BMP1 isoforms using a peptide-level discordant peptide search

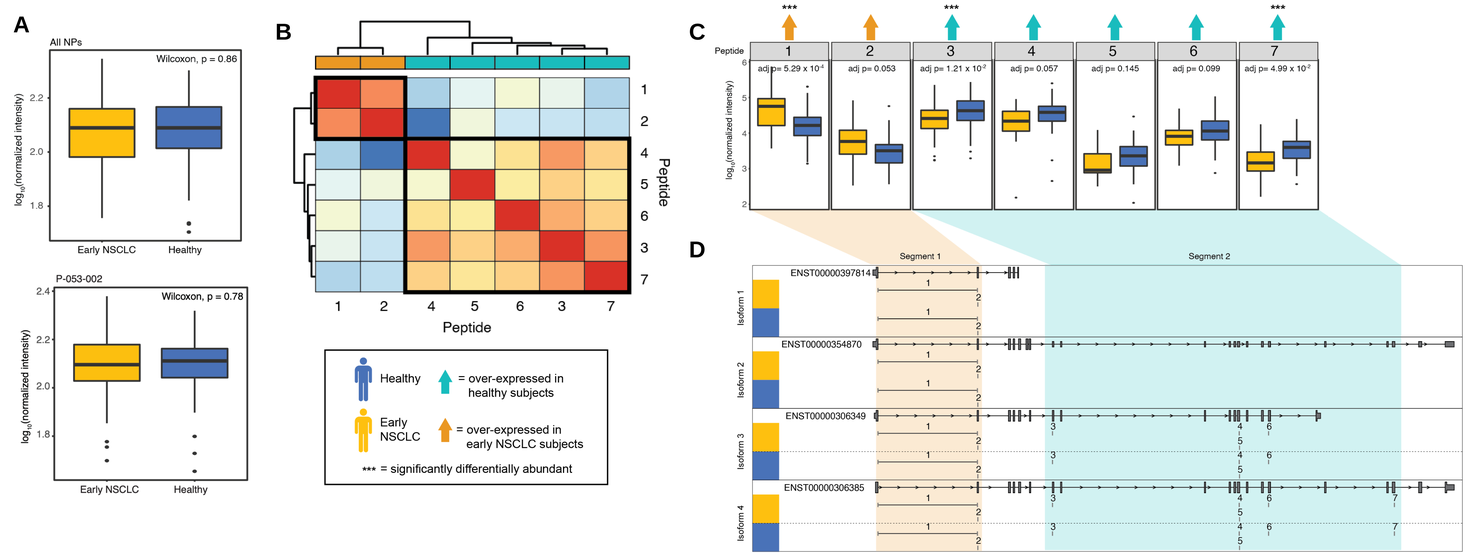


Figure 2. A) Log_{10} median normalized intensities of BMP1 in early NSCLC and in healthy subjects with collapsed abundances (top) and in a single NP (bottom). **B)** Heatmap showing the Pearson correlation of the BMP1 peptide abundances, where low correlation is indicated in shades of blue and high correlation is indicated in shades of red. Correlation values were clustered using hierarchical clustering. **C)** Log_{10} median normalized intensities of peptides mapped to BMP1 in early NSCLC and healthy subjects. **D)** Gene structure plots of known BMP1 protein coding transcripts (i.e., isoforms). BMP1's discordant abundance can be explained with known isoforms.

Proteogenomic inference of proteoforms arising from genetic variation

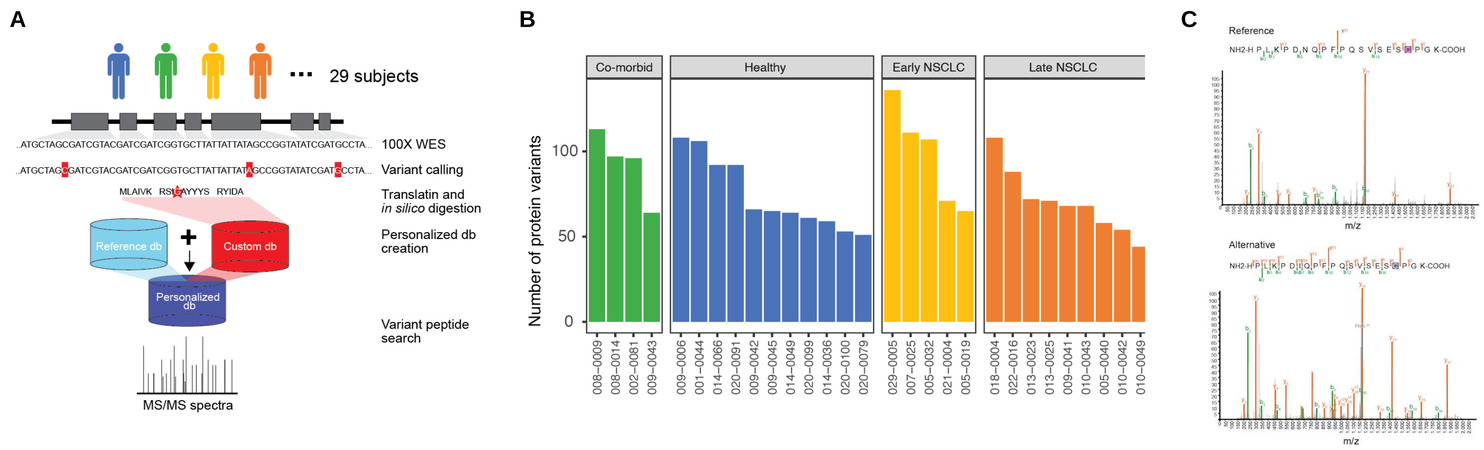


Figure 3. **A)** Experimental design for a proteogenomic search to identify protein variants. Exomes of 29 subjects were sequenced to 100X, followed by genomic variant calling. The genomic sequence, including identified variants, were translated to protein sequence and digested in silico. These custom sequences were then combined with a reference sequence database to generate a personalized database for searching variant peptides. **B)** Number of protein variants identified across 29 subjects. **C)** Tandem mass spectra of peptides arising from a heterozygous variant, where the alternative allele causes a single amino acid variant (SAAV; N → I).

Conclusions

Proteograph can generate unbiased and deep plasma proteome profiles that enable identification of protein variants and peptides present in plasma, at a scale sufficient to enable population-scale proteomic studies.

References

1. Blume et al. Nat. Comm. (2020)
2. Donovan et al. BioRxiv (2022)

Publications

Seer, Inc., Redwood City, CA - *yhuang@seer.bio