Deep, Rapid and Unbiased Plasma Proteomics with Peptide Correlation Analysis Enabling Proteoform Inference and Differential analysis in a NSCLC Lung Cancer Study

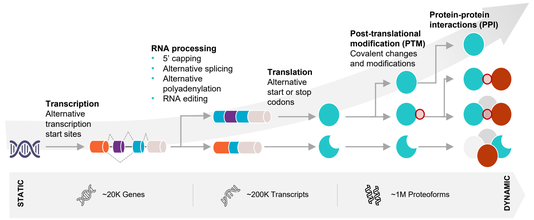
Henry Huang*, Jian Wang , Alexey Stukalov, Ting Wang, Margaret K. R. Donovan, Asim Siddiqui Serafim Batzoglou and Omid C. Farokhzad

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

Proteomes are Dynamic and Far More Diverse than Genomes

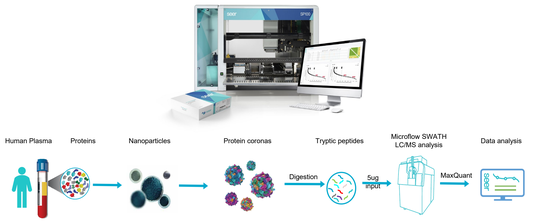
Comprehensive assessment of the human proteome remains elusive due to multiple forms of a protein, each of which can serve distinct functions, arising from alternative splicing, allelic variation, and protein post-translational modifications. Characterization of the variable protein forms, or proteoforms, 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 proteoform analysis of 80 healthy controls and 61 early-stage non-small-cell lung cancer (NSCLC) samples to infer proteoforms derived from alternative gene splicing or post-translational cleavage.

Proteoforms are critical to understanding human health and disease



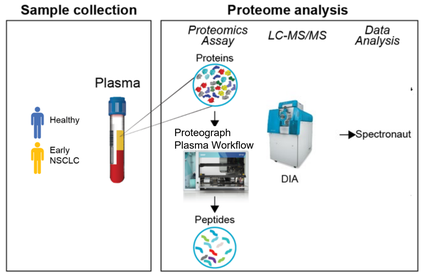
Proteograph solution

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



Proteoform inference methods

We implemented COPF[1] with minor adjustments and additional filtering steps. Pearson correlation was calculated for pairwise peptides using logged intensity of the peptides across samples. K-means clustering was then applied to the correlation matrix to group peptides into cluster/proteoforms. Proteoform score and p-value was calculated according to COPF. Of the proteins with potential proteoform according to COPF's proteoform score, we further filter for certain type of proteoforms, specifically post-translational cleavage. Using the Wilcoxon's rank test, we test for the significance that peptides from one cluster/proteoform are disproportionally located on one terminus of the protein.



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

Figure 1. K-means clustering and COPF's proteoform score indicate that there are 204 protein with significant clusters inferring potential proteoforms. Filtering for specifically post-translational cleaved proteoforms with Wilcoxon's rank test infer that there are 18 of such type of proteoform. Of these 18, 4 were able to map to known proteoforms.

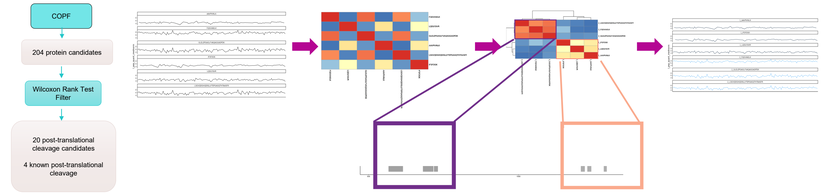


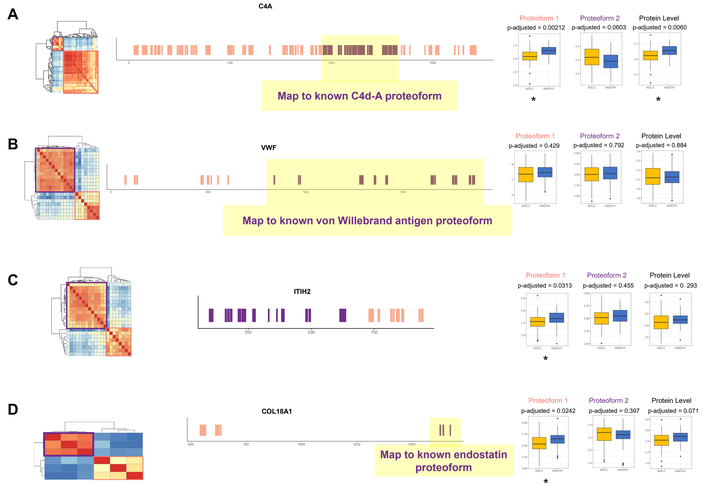
Figure 2. Four candidate clusters were able to map to know proteoforms, and all four candidates have known role in cancer, some even specifically in lung cancer. To compare proteoform intensities, peptides intensities from proteoforms were combined using MaxLFQ. Of the 18 post-translational cleavage candidates, 6 candidates have significant differentially expressed proteoform abundance between Healthy control and NSCLC samples, 3 of which have known proteoforms (P0C0L4, ITIH2, COL18A1). All peptides of the proteins were also combined to produce protein intensity for comparison.

A) C4A has two clusters. Cluster 2 is located in the C4d region of the protein. Elevation of C4d has been shown to be diagnostic of lung cancer[2]. Cluster 1 and 2 of C4A have opposite differential abundances, while protein abundance is upregulated in healthy control. We would not observe C4d down regulation at the protein level.

B) Cluster 1 of VWF is mapped to the von Willebrand antigen region. This proteoform has been shown to be elevated in NSCLC patient with poor prognosis[3]. However, we observe no significant differential abundance at any level.

C) Cluster 1 of ITIH2 is downregulated in NSCLC while there is no difference in cluster 2 and protein level. Down regulation of ITIH2 has been shown to be associated with progression of multiple malignancies including lung cancer[4]. We would not have observed this down regulation if we analyze differential expression at the protein level, but the phenomenon is present in cluster 1.

D) Cluster 2 of COL18A1 is mapped to the endostatin region. Endostatin has been shown to treat NSCLC in combination with radiation[5]. Cluster 1 of COL18A1 is downregulated in NSCLC while there is no difference in cluster 2 and protein level between healthy and NSCLC.



Conclusions

Using a modified COPF and Wilcoxon's rank test we found known post-translational cleaved proteoforms. We demonstrated that previously shown results in NSCLC is sometimes found at the proteoform level not at protein level. Only by looking at the proteoform level abundance would we find differential expression. In addition to known biology, There is potential new discovery in the differentially expressed clusters/proteoforms that do not map to known proteoforms.

References

1. Bludau et al. Nat. Comm. (2021)
2. Ajona et al. Plos One (2015)
3. Guo et al. J Clin Lab Anal. (2018)
4. Hamm et al. BMC Cancer (2008)
5. Zhang et al. Radiation Oncology (2020)

Publications

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