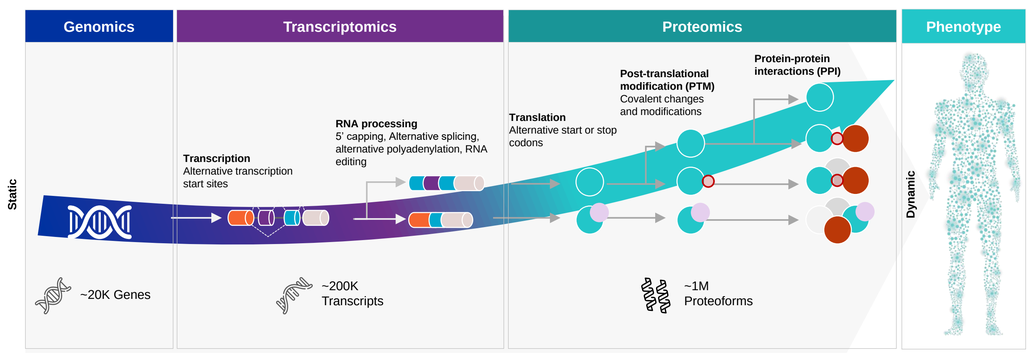
Proteoform inference using a proteogenomic approach in non-small cell lung cancer and healthy control plasma proteomes reveals disease-associated protein isoforms

Margaret K. R. Donovan*, *Yingxiang Huang, Jian Wang, Alexey Stukalov, Ting Wang, Asim Siddiqui, and Serafim Batzoglou*

Proteoforms are critical to understanding human health and disease

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.

Proteomes are dynamic and far more diverse than genomes

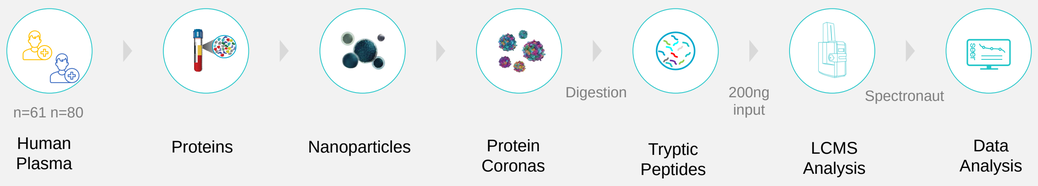


The Proteograph Product Suite enables deep, unbiased, rapid, and scalable access to the plasma proteome

Proteograph Product Suite

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

From sample to peptides to analysis on most LCMS instruments with a variety of proteomics methods.



Data Generation

- Plasma samples from 141 subjects comprising 61 NSCLC patients and 80 healthy controls were collected.
- Plasma samples were analyzed using Proteograph plasma protein profiling platform[1].
- Using 5 injections per sample, proteins were quantified using a 30-minute SWATH DIA method on SCIEX Triple TOF 6600+.
- The vwSWATH data were processed using Spectronaut.
- Proteoforms were inferred using a modified form of the COPF[2] method.

Proteoform inference using COPF

We implemented COrrelation-basaed functional ProteoForm[1] (COPF) assessment with minor adjustments and additional filtering steps to infer protein proteoforms. 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

Inference of proteoforms associated with NSCLC

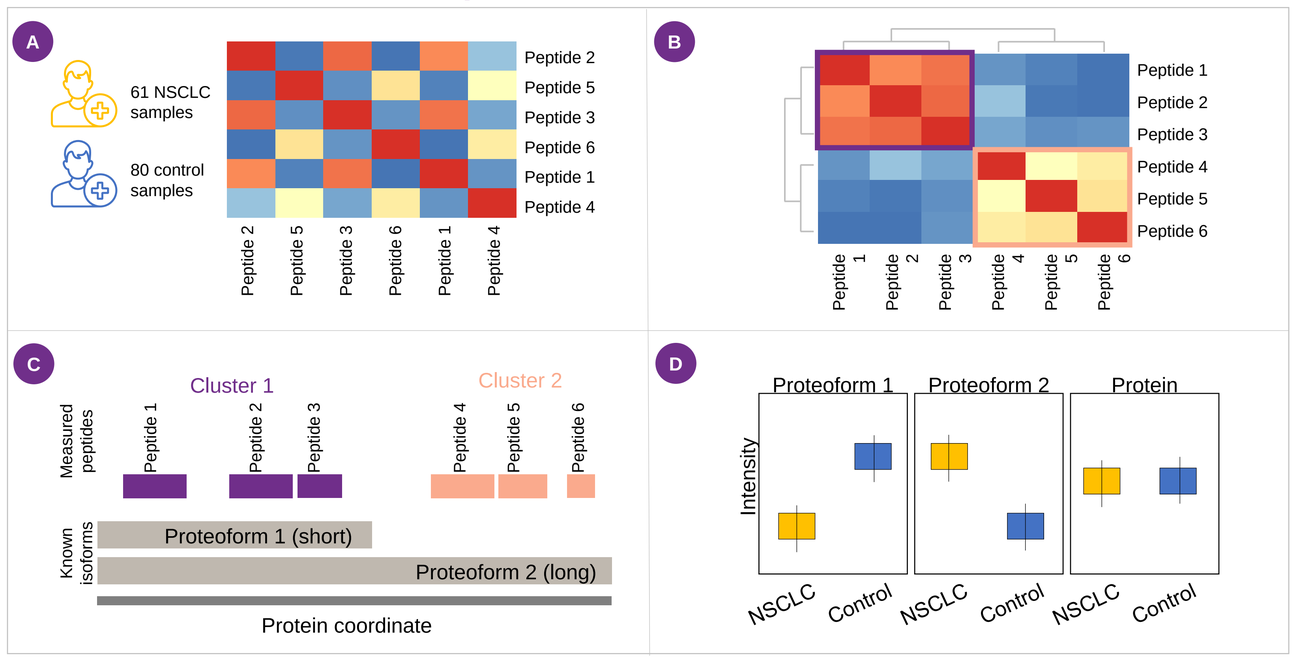


Figure 1. Proteoform inference using COPF and association with cancer

A) For each protein, the Pearson correlation is computed for all pairwise comparisons of logged peptides intensities across samples (61 NSCLC samples and 80 control samples).
B) K-means clustering is then applied to the correlation matrix to group peptides into clusters.
C) After filtering proteins to those that exceeded COPF proteoform score, candidate proteoforms arising from post-translational cleavage were identified by mapping the peptides to the protein coordinates and computing if the peptides were spatially arranged according to their identified k-means clusters by using the Wilcoxon's rank test to test for the significance that peptides from one cluster/proteoform are disproportionally located on one terminus of the protein. Of these candidate proteoforms, we further extracted the proteins whose peptide clusters corresponded to known proteoforms.
D) For each of the candidate isoforms, we calculated if there was a difference in proteoform intensity between NSCLC and healthy control samples.

Four proteoforms with potential clinical significance in cancer

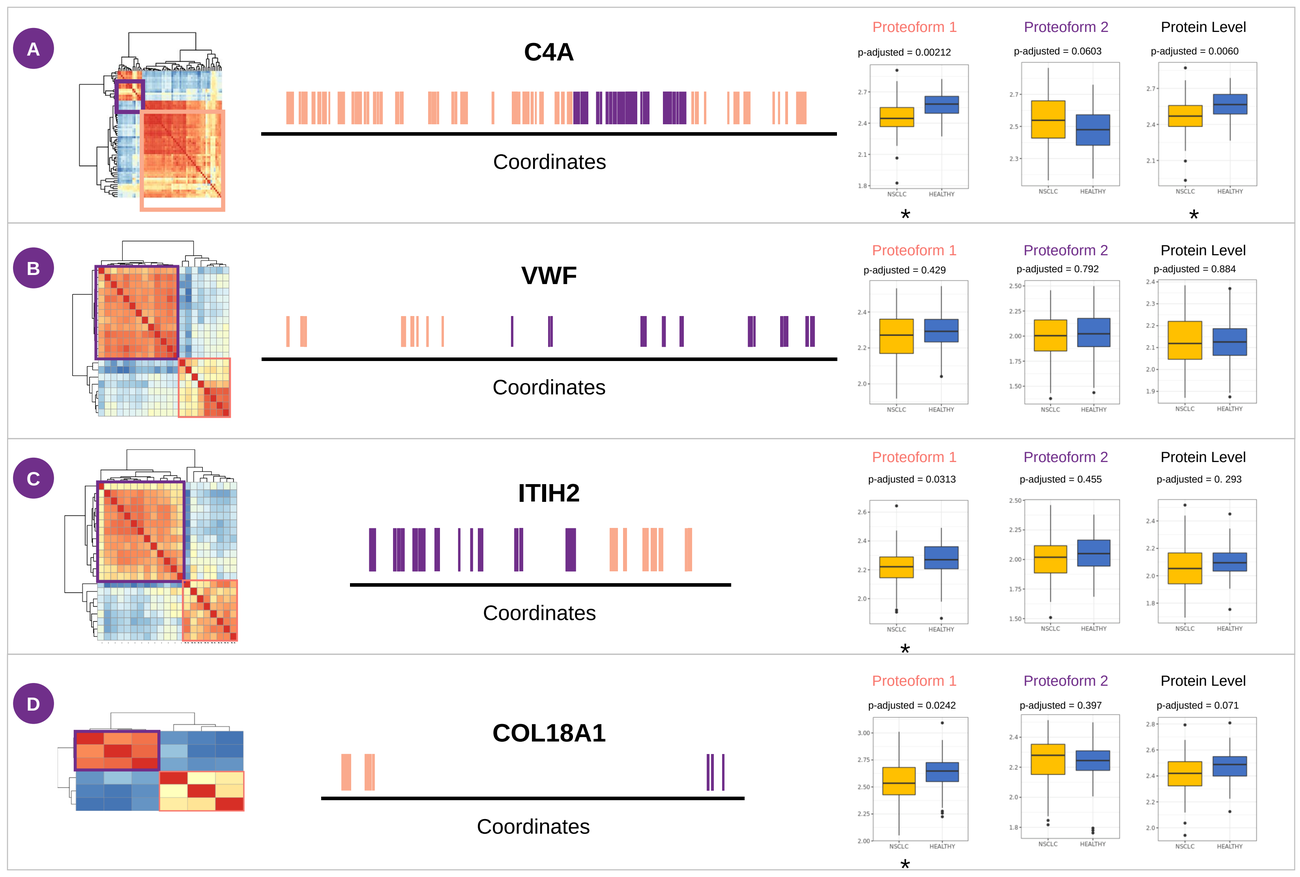


Figure 2. Four post-translationally cleaved proteoform candidates

- 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.
- 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.
- 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.
- 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.

Conclusion

- Using a modified COPF and Wilcoxon's rank test, we identified cancer-associated post-translationally cleaved proteoforms.
- We demonstrate that protein-level analysis can mask biologically important results that can only be identified at the peptide-level.
- Further analysis of protein candidates that do not map to known protein isoforms may reveal novel, clinically important 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)

PAS Proteogenomics

Publications

[1]Seer, Inc., Redwood City, CA 94065, USA | *mdonovan@seer.bio
[2]Massachusetts General Hospital, Boston, MA 02114 , USA