Deep Plasma Proteomics at Scale Enabling Precision Analyses in a Lung Cancer (NSCLC) Study

Mahdi Zamanighomi*, Harendra Guturu, Jian Wang, Amir Alavi, Tristan Brown, Daniel Hornburg, Moaraj Hasan, Shadi Ferdosi, Khatereh Motamedchaboki, Margaret Donovan, Theodore Platt, Ryan Benz, Asim Siddiqui, and Serafim Batzoglou

Deep and unbiased plasma proteomics for disease cohort studies at scale

Introduction

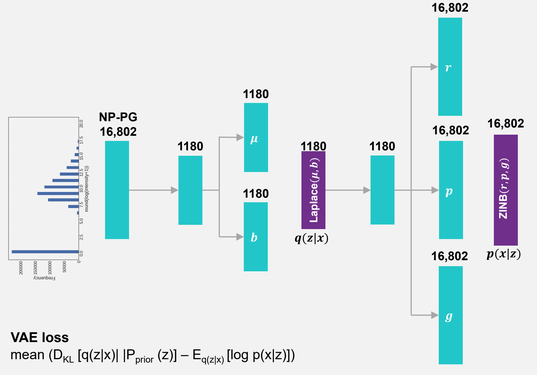
Our ~20,000 genes encode over one million protein variants, given alternative splice forms, allelic variation, and protein modification. Though large-scale genomics studies have expanded our understanding of biology, similarly, scaled deep and untargeted proteomics studies of biofluids have remained impractical due to complex workflows. To address this need, we have previously described Proteograph™, a novel platform that leverages the protein-corona interactions of nanoparticles (NP) for deep and untargeted proteomic sampling at scale.

Using Proteograph in a non-small cell lung cancer (NSCLC) cohort, we previously conducted a deep interrogation of plasma from using early-stage NSCLC subjects and non-cancer controls. We identified 2,499 plasma proteins, with 1,992 present in ≥ 25% of the samples. Leveraging this data, we created a biomarker classifier distinguishing NSCLC from controls with average area under the receiver operating characteristic curve of 0.91.[1] In this study, we now re-analyze the data with the recently released DIA-NN software and leveraged cloud architecture to successfully scale up and process large cohort group runs. This enabled enhanced proteome depth (41% increase) while improving the accuracy (0.95) of the classifier. Our results outline workflows for robust biomarker discovery and cohort subtyping.

The Proteograph™ platform interrogates the plasma proteome at previously impractical combinations of scale, depth and coverage, and enables the development of improved classification models and marker discovery.

Methods

- Cohort was analyzed with DIA-NN v1.8 in single group-run in library free mode against standard human uniport proteome using the `--relaxed-prof-inf` option[2].
- Differential expression of proteins was computed using DIA-NN estimated $\log_{10}(1 + \text{intensities})$ and Welch's t-test.
- Functional enrichment analysis was performed using g:Profiler (version e104_eg51_p15_3922dba) with g:SCS multiple testing correction method applying significance threshold of 0.01[3].
- Cohort classification was done with implementations of several machine learning classifiers applied to protein intensities and their embedding from a variational auto-encoder (VAE) that will allow for easier integration and visualization of proteogenomics data in the future.



VAE loss
mean (D_{KL} [q(z|x)| P_{prior} (z)] – $E_{q(z|x)}$ [log p(x|z)])

Figure 1. Architecture diagram of VAE neural network.

Each rectangle denotes a fully connected block. Laplace and Zero Inflated Negative Binomial (ZINB) distributions are considered to encourage sparsity and to model sparsity.

Deep proteomics enhances disease classification and biomarker characterization of early NSCLC cohort

Results

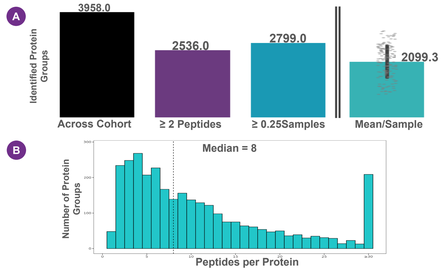


Figure 2. Distribution of identified protein group counts.

A) Using DIA-NN, we identified nearly 4,000 protein groups across the cohort (on average >2,000 per sample vs 439 in neat plasma) with the majority being supported by multiple peptides and found in multiple samples. **B)** The Proteograph™ successfully detected a median of 8 peptides per protein.

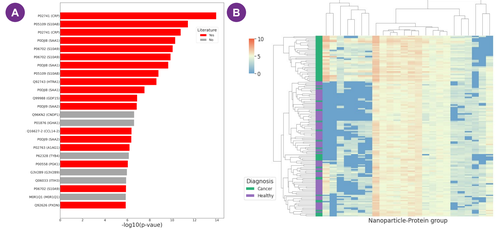


Figure 3. Differentially expressed protein groups across cohort.

A) The Welsh's t-test on protein intensities resulted in 17 unique protein groups that were significantly differentially expressed after Bonferroni corrected threshold of 0.01. Protein groups in red indicate an association to NSCLC was found in literature. **B)** Hierarchical clustering on the differentially expressed protein intensities depicts a distinct separation of cancer and healthy subjects.

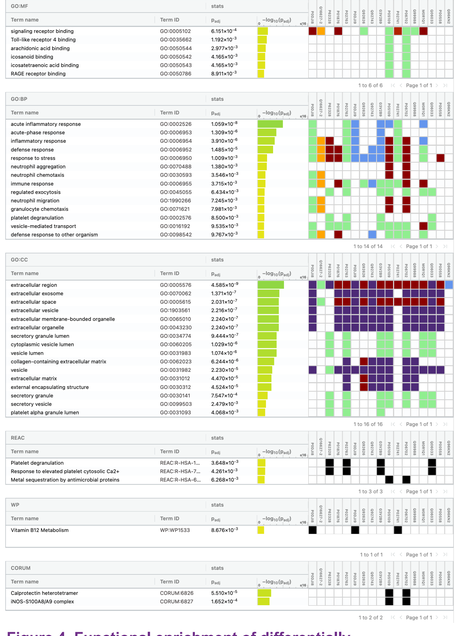


Figure 4. Functional enrichment of differentially expressed proteins.

Detailed list of enriched terms after correction for multi-correction testing using g:Profiler's g:SCS multiple hypothesis testing method that accounts for relationships between ontology terms.

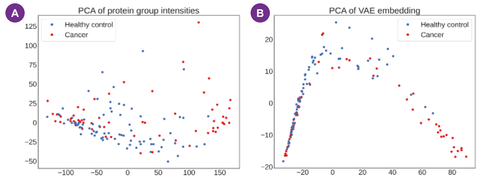


Figure 5. PCA projection of the cohort.

PCA of the **A)** protein group intensity matrix and **B)** VAE embedding visualizes the ability of the VAE to better separate the cohort based on disease status in the latent space.

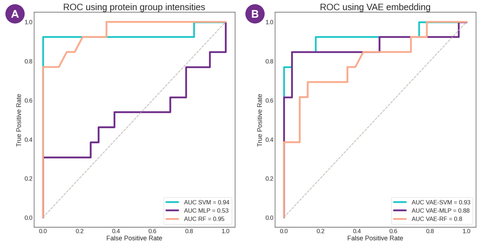


Figure 6. Classification performance using protein intensities and a VAE embedding.

The ROC curve using the **A)** protein group intensity matrix and **B)** VAE embedding reflects the classification power of different methods. All commonly used classifiers are robust and accurate in the VAE embedding.

Conclusion

- DIA-NN combined with the cloud-based large group run more sensitively identifies protein groups across the cohort (~4000 vs the previous ~2,500).
- Enhancement of classification performance and stability with VAE, Random Forest, and SVM to AUC-ROC 0.95 compared to the previous AUC-ROC 0.91.
- Differentially expressed proteins can lead to a better biomarker discovery and disease characterization.

References

[1] Blume et al. *Nat. Comm.* (2020)
[2] Demichev et al. *Nat Comm.* (2020)
[3] Raudvere, Kolberg et al. *Nucleic Acids Res.* (2019)