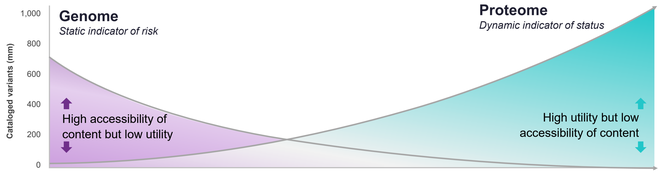
Plasma Proteomics at Scale Enabling Lung Cancer, Alzheimer's Disease and Proteogenomics Studies with the Proteograph™ Product Suite

Asim Siddiqui*, John E. Blume, Sangtae Kim, Margaret K. R. Donovan, Theodore L. Platt, Ryan W. Benz, Martin Goldberg, Serafim Batzoglou, Juan C. Cuevas, Philip Ma, and Omid C. Farokhzad

Proteograph Platform Delivers Unbiased, Deep and Rapid Proteomics at Scale

The ~20,000 genes in the human genome encode over one million protein variants, given alternative splice forms, allelic variation and protein modifications. Though large-scale genomics studies have expanded our understanding of biology, similarly-scaled unbiased, deep proteomics studies of biofluids have remained impractical due to complexity of workflows. Here we show how unbiased, deep proteomics enabled by the Proteograph Product Suite[1] enables large-scale proteomics and proteogenomic studies.



Utility of Information

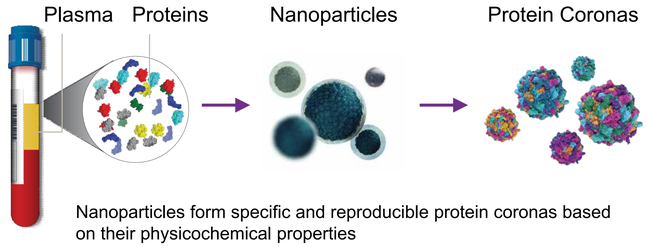
10M+ human exomes
1M+ genomes

~695M genetic variants catalogued

< 0.2% of genetic variants fully characterized

Core Technology

Proprietary Engineered Nanoparticles



Nanoparticles form specific and reproducible protein coronas based on their physicochemical properties

Proteograph Product Suite

Sample is ready to be analyzed on most LC-MS instruments

Proteograph Study of Non-small Cell Lung Cancer (NSCLC) and Alzheimer's Disease (AD)

NSCLC Study[1]

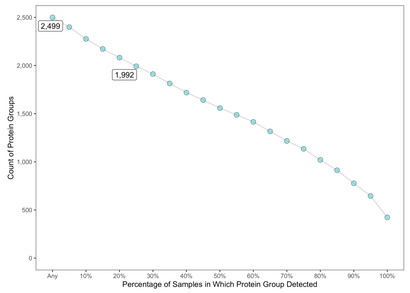
- 141 samples, NSCLC lung cancer (n=61) and control (n=80) plasma samples
- Total experiment time of ~2 weeks on 3 MS instruments

AD Study

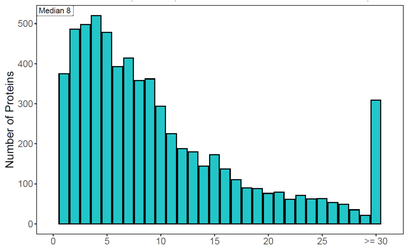
- 200 AD samples (n=50), mild cognitive impairment (MCI; n=50) and control (n=100) plasma samples
- Total experiment time of ~3 weeks on 2 MS instruments

Proteograph Enables Deep and Unbiased Plasma Proteomics

Protein Groups Detected Across 141 Subjects in Control vs. Early NSCLC Study



2499 protein groups are found across all subjects and **1,992** in **25%** of the subjects.

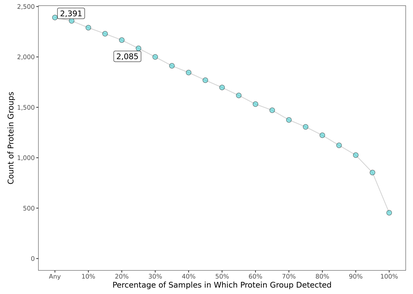


21,959 peptides were detected in total with a median **8** peptides per protein across NSCLC Study.

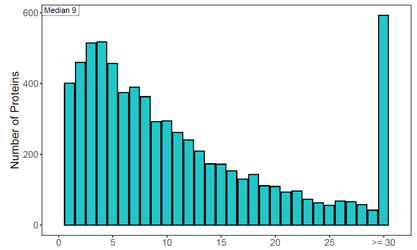
Figure 1.
Number of protein groups detected across a percentage of the subjects in these two Proteograph studies.

Figure 2.
Number of peptides for each protein group in each study.

Protein Groups Detected Across 200 Subjects in AD/MCI Study



2391 protein groups are found across all subjects and **2,085** in **25%** of all the subjects.



26,264 peptides were detected in total with a median **9** peptides per protein across AD Study.

Robust and Efficient Biomarker Discovery Workflow with Proteograph Technology

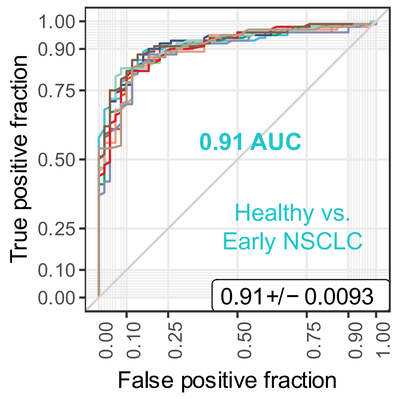
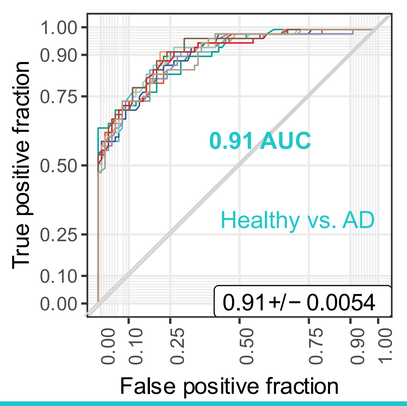


Figure 3.
ROC plot of the 10x10 cross validation of the results in each study. Further verification studies are required to validate the model.



Proteogenomic Studies at Scale Require Unbiased, Deep and Rapid Methods

To investigate the relative abundance of protein variants and isoforms across different biological states, there are three necessary things to consider; 1) abundance measurement of the variants present, which necessitates multiple measurements across the protein (i.e. multiple peptides), 2) Sufficient scale to measure across many samples and 3) Unbiased interrogation, so measurements are not restricted to the most frequent alleles as the rare alleles are known to play an important role in complex disease.

To test the application of the Proteograph solution to analyze protein variants, we exon sequenced 29 individuals from the NSCLC study, creating personalized mass spectrometry search libraries which were used to identify 464 amino acid variants in these individuals. Preliminary investigation of proteins containing these variants suggests putative allele specific presence in at least 178 genes. We observed an increased prevalence of low frequency alleles in the imbalanced variants (Figure 5) which is suggestive of functional implications of these differences. Further investigation is required to rule out technical artifacts of MS processing that may cause loss of these peptides.

Together, these results demonstrate how the Proteograph Product Suite can support unbiased, deep proteogenomic studies at scale.

AF Across 464 Protein Variant From 29 Subjects vs 1K Genome

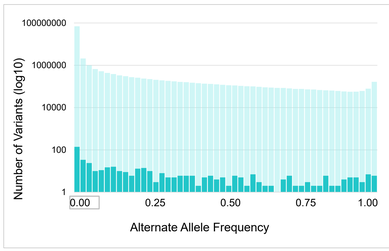


Figure 4.
Allele frequency analysis of protein variants. The allele frequency of the variants found in the 29 individuals against the background of allele frequencies in the 1000 genomes project[2] shows the distributions are similar, demonstrating the unbiased nature of the Proteograph solution.

Variants with ASE AF vs. Protein Variant AF

ASE:
allele specific expression

AF:
allele frequency

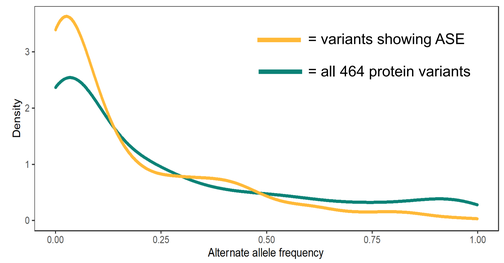


Figure 5.
Density plot of alleles with allele specific translation/expression as a function of AF. We define putative ASE where we observe one or more peptides mapping to only the reference or only the alternative allele, but not both, in all subjects with that genotype (i.e., monoallelic translation). Kernal smoothing used on density plot.

The Proteograph Product Suite has the throughput required for measurements of individual protein variants across the proteome in an unbiased manner, enabling unbiased proteogenomics at scale.

References:
1. Blume et al. Nat. Comm. (2020)
2. 1000 Genomes (2015)

Publications

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