A Cloud-scalable Software Suite for Large-Scale Proteogenomic Data Analysis and Visualization

Margaret K.R. Donovan, Harsharn Auluck, Arjun Vadapalli, Yan Berk, Aaron S. Gajadhar, Khatereh Motamedchaboki, Yuandan Lou, Theo Platt, and Asim Siddiqui

The Proteograph Analysis Suite is an intuitive, scalable, data informatics solution

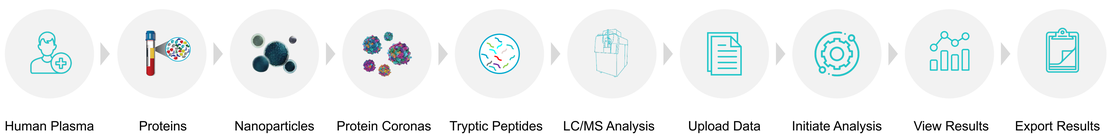
PAS enables automated results generation and intuitive, easy to interpret proteomics visualizations

Introduction

Assessment of the flow of genetic information through multi-omic data integration can reveal the molecular consequences of genetic variation underlying human disease. Next-generation sequencing (NGS) can be used to identify genetic variants, while mass spectrometry can be used to assess the proteome. The Proteograph™ Product Suite[1], leverages multiple nanoparticles with distinct physiochemical properties to enable large-scale, deep plasma proteome analyses. Integration of proteomics and genomics data requires many tools of which require complex workflows that can act as a barrier for researchers to adapt new analysis tools. We present a cloud-based, data analysis software platform called Proteograph Analysis Suite (PAS) for proteogenomic data analyses through the integration of Proteograph proteomics data with NGS variant information.

Here, we apply PAS by analyzing 141 Proteograph NSCLC plasma dataset[1] and performing a database search. This search was launched through the user interface requiring only 3 clicks, and in the background this search provisioned 142 servers and completed in approximately five and half hours. Together, these results show the utility of PAS for seamless and fast proteomic data analysis.

Seer core technology and the Proteograph Product Suite provides untargeted, deep, and rapid proteomics at scale



Proteograph Analysis Suite allows a seamless journey from raw data to biological insight

PAS includes an experiment data management system, analysis protocols, analysis setup wizard, and result visualizations. PAS can support both Data Independent Analysis (DIA) and Data Dependent Analysis (DDA) workflows and is compatible with variant call format (.vcf) files, enabling personalized database searches. To assess data quality, PAS includes metrics for identified peptides and protein groups like intensity, protein sequence coverage, abundance distributions, and counts. Visualizations, including principal component analysis, hierarchical clustering, and heatmaps, allowing identification of experimental trends. To enable biological insights, differential abundance analyses results are displayed as volcano plots, protein interaction maps, and protein-set enrichment. From data to insight, PAS provides an easy-to-use and efficient suite of tools to enable proteogenomic data analysis.

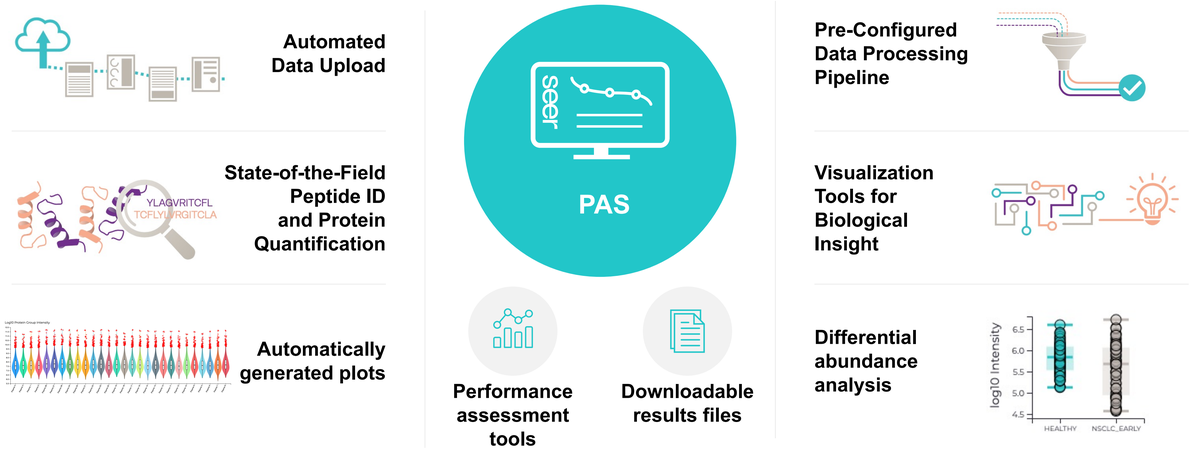


Figure 1.

Proteograph Analysis Software (PAS) is a scalable on the cloud solution to coordinate the data analysis for the entire Proteograph Product Suite including the Proteograph Assay Kit, SP100 automation instrument and LC-MS analyses. Data is seamlessly transferred from MS computer to PAS without manual intervention using the AutoUploader tool in PAS. PAS features multiple, integrated MS/MS database search engines, automatic results generation, QC tools to evaluate data quality, and differential expressions analysis wizard for seamless generation of proteomics results.

Result summaries and visualizations simplify evaluation of assay performance and data interpretation

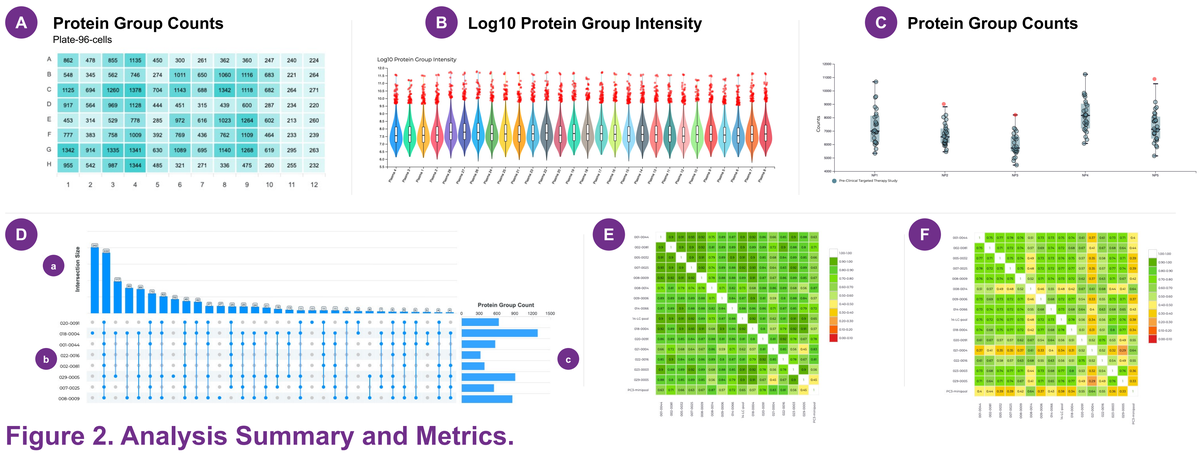


Figure 2. Analysis Summary and Metrics.

A) View results for protein groups (shown) and peptide counts, quant mass, miscleavage rate, oxidation ratio and ID rate in a simple and intuitive plate format. **B)** Distributions of protein group intensities and CVs across samples. **C)** Box plots showing the number of protein groups identified across NPs. Hovering over a dot reveals the peptide or protein count, file, and sample name. Hovering over a box shows the quantile for the NP. **D)** Graphs and a matrix show protein group overlaps; Intersection Size bar graph **b)** Protein Group Count bar graph **c)** Matrix **E, F)** A color-coded matrix displays sample comparability data using PCC (left) or the Jaccard index (right). Samples on the green end of the spectrum have high correlation, while samples on the red end of the spectrum have low correlation.

Differential abundance analysis identify proteome differences between conditions or groups, enabling new biological insights

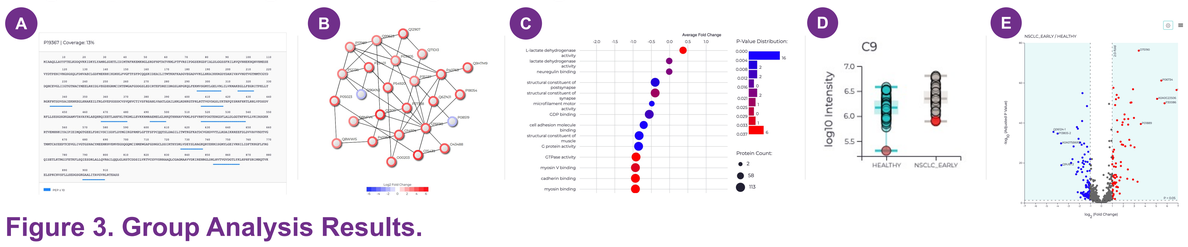


Figure 3. Group Analysis Results.

A) Sequence Coverage: Visualize where peptides map relative to the protein sequence. **B)** Protein-Protein-Interactions Comparison: Build a STRING-based PPI network to identify differences in protein interactors. **C)** GO Enrichment: Explore how proteins associated with a group differ functionally. **D)** Intensity Comparison: View how the intensity of a protein of interest differs between groups. **E)** Sample group analysis visualized with volcano plot.

Proteogenomics functionality allows integration of multi-omics datasets such as linking genomic variant results with the proteome

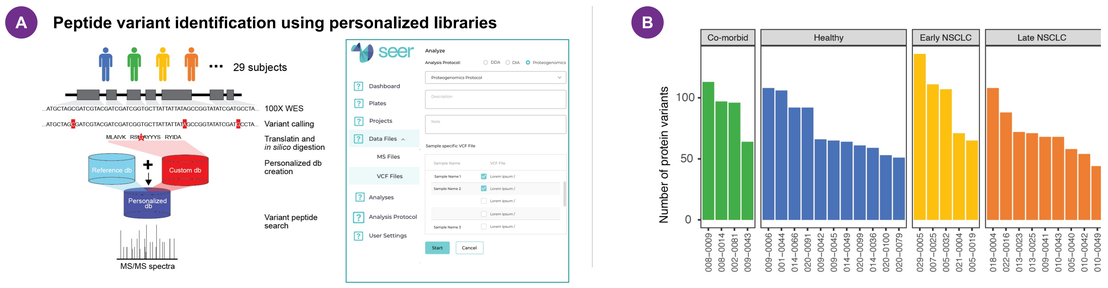


Figure 4.

A) PAS can analyze VCF files generated from NGS pipelines in combination with mass spec data to identify peptide variants[2] using personalized libraries. **B)** Bar charts showing the number of protein variants identified using this proteogenomic data analysis approach.

Conclusion

We present a comprehensive proteomic analysis software suite to enable user-friendly and reproducible multi-omics analyses of proteomic and genomic data.

References
[1] Blume et al. *Nat. Comm.* (2020)
[2] Donovan et al. *BioRxiv.* (2022)

Publications

Seer, Inc., Redwood City, CA 94065, USA | mdonovan@seer.bio