A Cloud-scalable Software Suite for Large-cohort Proteogenomics Data Analysis and Visualization.

Mass Spectrometry

Aaron Gajadhar (agajadhar@seer.bio), Seer, **Margaret Donovan**, Seer, **Harsharn Auluck**, Seer, **Yan Berk**, Seer, **Yuandan Lou**, Seer, **Theo Platt**, Seer, **Serafim Batzoglou**, Seer

Comprehensive 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) is used to identify genetic variants and characterize gene function (e.g. transcriptome and epigenome), while mass spectrometry is used to assess the proteome through characterization of protein abundances, modifications, and interactions. A new plasma profiling platform, the ProteographTM Product Suite, leverages multiple nanoparticles with distinct physiochemical properties to enable deep plasma proteome analyses at scale. Here, we present a cloud-based, data analysis software platform called Proteograph Analysis Suite (PAS) for proteogenomic data analyses through the integration of proteomics data derived from the Proteograph with genomic variant information derived from NGS experiments.

PAS features include an experiment data management system, analysis protocols, an analysis setup wizard, and tools for reviewing and visualizing results. PAS can support both Data Independent Analysis (DIA) and Data Dependent Analysis (DDA) proteomics workflows and is compatible with variant call format (vcf) files from NGS workflows to enable personalized database searches. To assess quality of the resulting data PAS includes various metrics like peptide/protein group intensity, protein sequence coverage, relative protein abundance distribution, peptide and protein group counts. Visualizations including principal component analysis, hierarchical clustering, and heatmaps allow intuitive identification of experimental trends. To enable biological insights, differential expression analyses results are reported with interactive visualizations such as volcano plots, protein interaction maps, and protein-set enrichment. From data to insight, PAS provides an easy-to-use and efficient suite of functionality to enable proteogenomic data analysis.

Integration of proteomics and genomics data require a variety of tools, many of which require command-line interfaces and operating system-specific requirements that can act as a barrier for researchers to adapt new data analysis tools. Here, we demonstrate the utility of PAS by analyzing samples from the Proteograph NSCLC plasma dataset (1). PAS can analyze VCF files generated from NGS pipelines in combination with mass spec data to identify peptide variants using personalized libraries. Using the cloud-based architecture computational tasks are distributed for rapid analysis. The integrated proteogenomics viewers allow variant IDs to be interpreted in the context of genomic coordinates, protein sequence, functional domains and features. Together, these results show the utility of PAS for seamless and fast proteomic data analysis.

Reference: (1) Blume, J. E. et al. Nature Communications, 2020