An integrated data processing and visualization suite leveraging cloud scalable architecture for large-cohort proteogenomics data analysis and interpretation

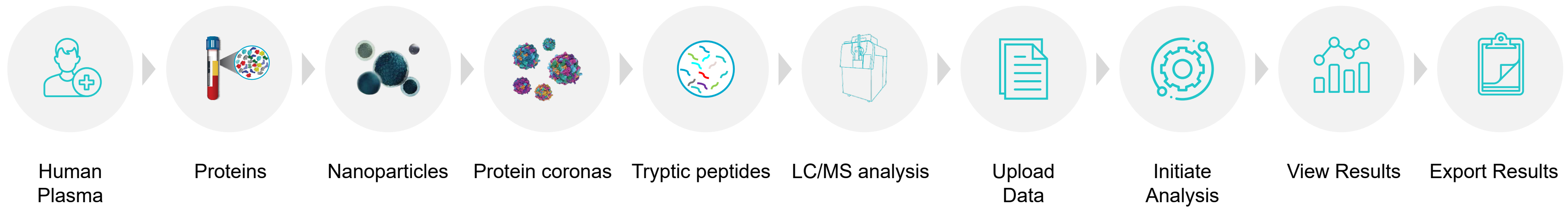
Aaron S Gajadhar*; Margaret K.R. Donovan; Harsharn Auluck; Yan Berk; Yuandan Lou; Theo Platt; and Serafim Batzoglou. Seer, Inc., Redwood City, CA

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

Introduction

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 Proteograph^{TM} Product Suite, leverages multiple nanoparticles with distinct physiochemical properties to enable deep plasma proteome analyses at scale. Here, we present an intuitive, scalable analysis 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.

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.

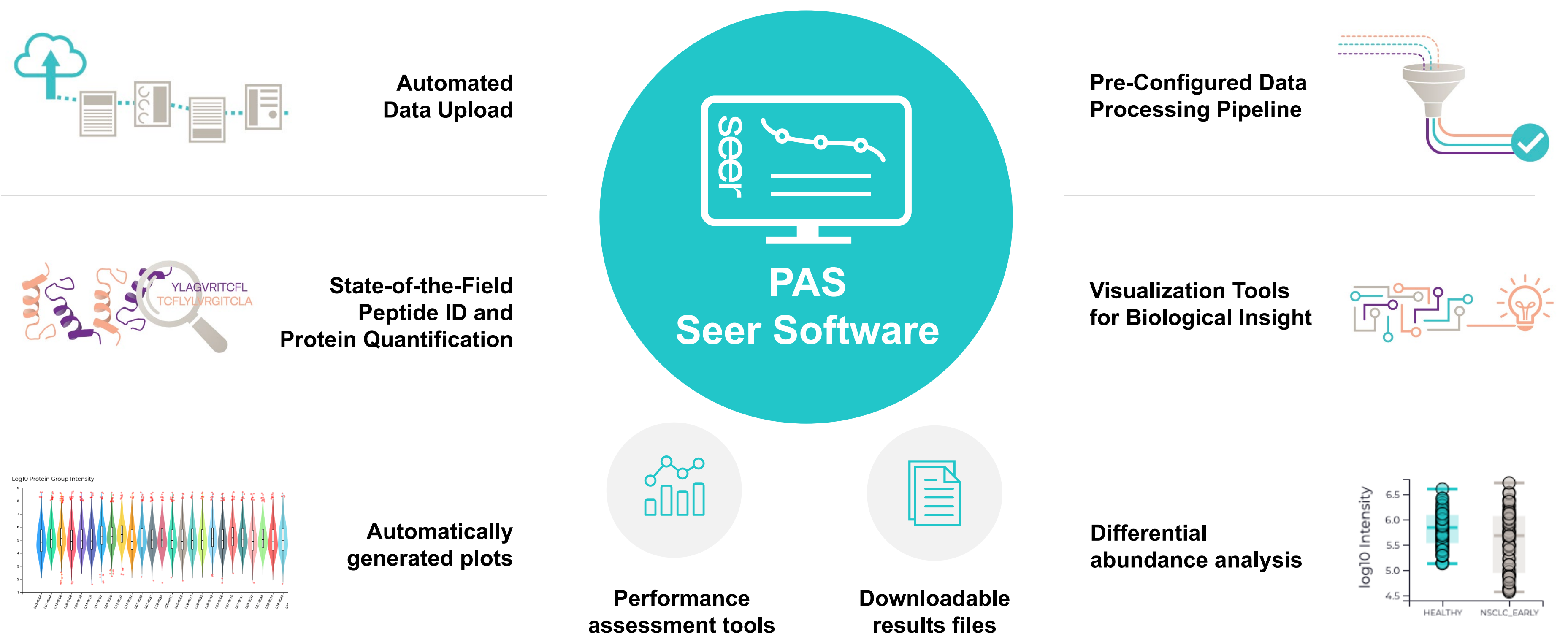

Automated Data Upload
State-of-the-Field Peptide ID and Protein Quantification
Automatically generated plots
PAS
Seer Software
Performance assessment tools
Downloadable results files
Pre-Configured Data Processing Pipeline
Visualization Tools for Biological Insight
Differential abundance analysis
log10 Intensity
HEALTHY
NSCLC_EARLY

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.

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

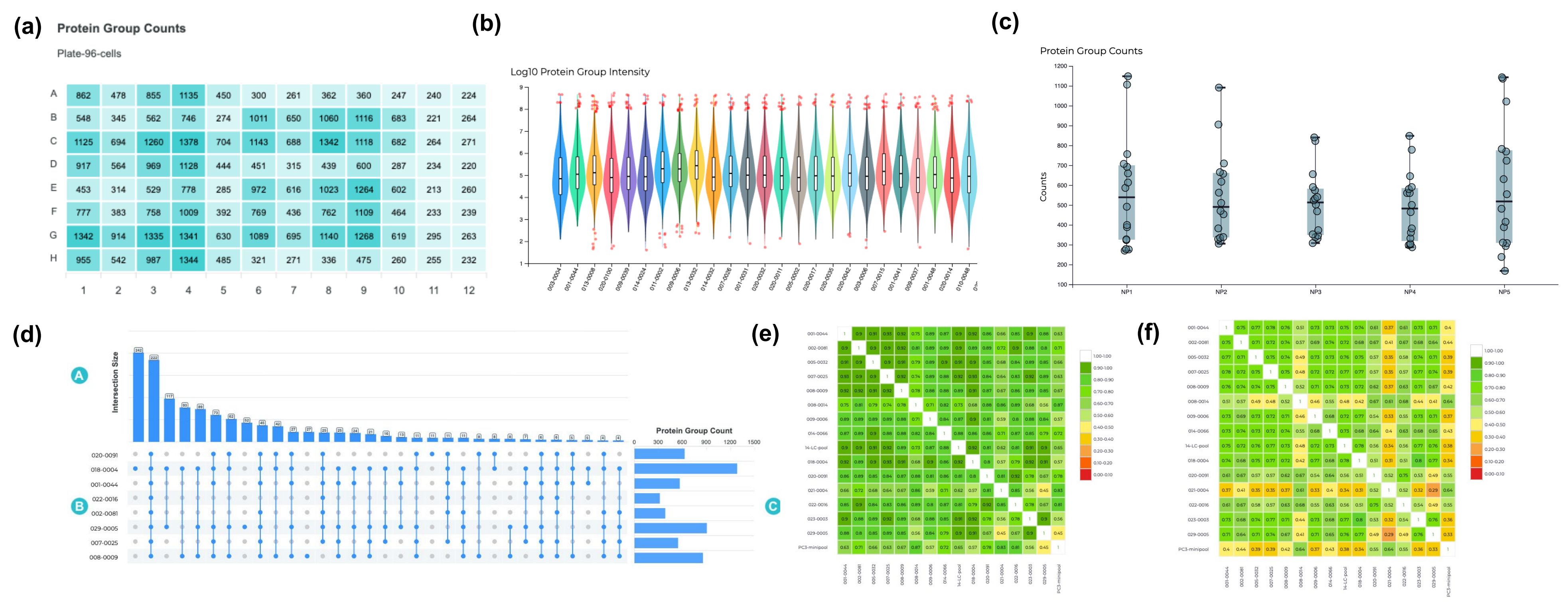


Figure 2. Analysis Summary and Metrics **(a)** View results for protein groups (shown) and peptide counts, quant mass, miscleavage rate, oxidation ratio and ID rate in a simple and intuitive plate format. **(b)** Distributions of protein group intensities and CVs across samples. **(c)** Box plots showing 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.

QC Charts automate statistical process control assessment allowing rapid run-to-run performance evaluation

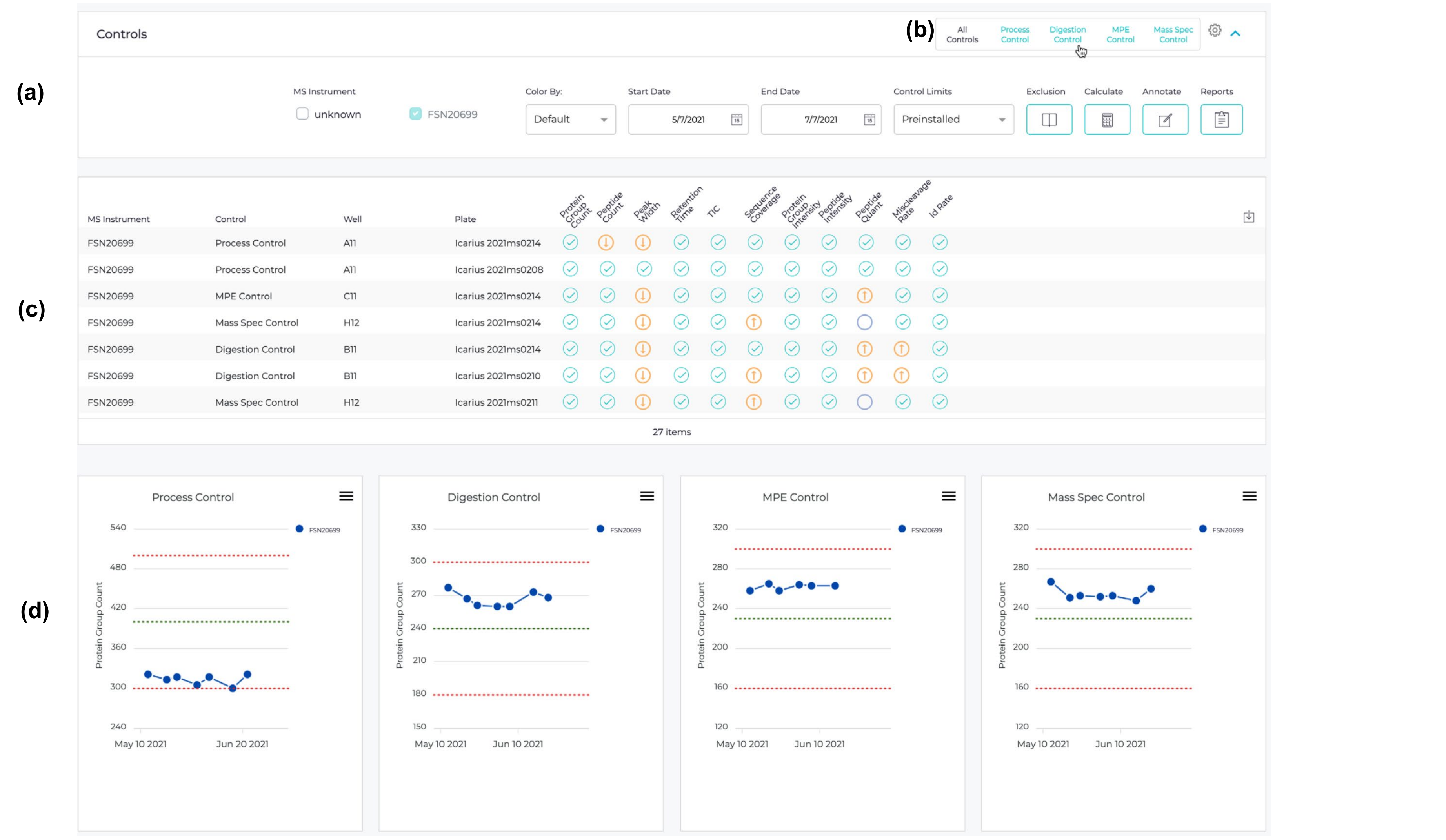


Figure 3. Control Results: **(a)** Filters for viewing charts for all controls or a selected control type. **(b)** Toolbar with additional filters and functions. **(c)** Summary of control data for the selected analysis time frame. **(d)** QC charts with metrics for each control.

Differential expression analysis tools simplify data interpretation allowing easy determination of biological insights

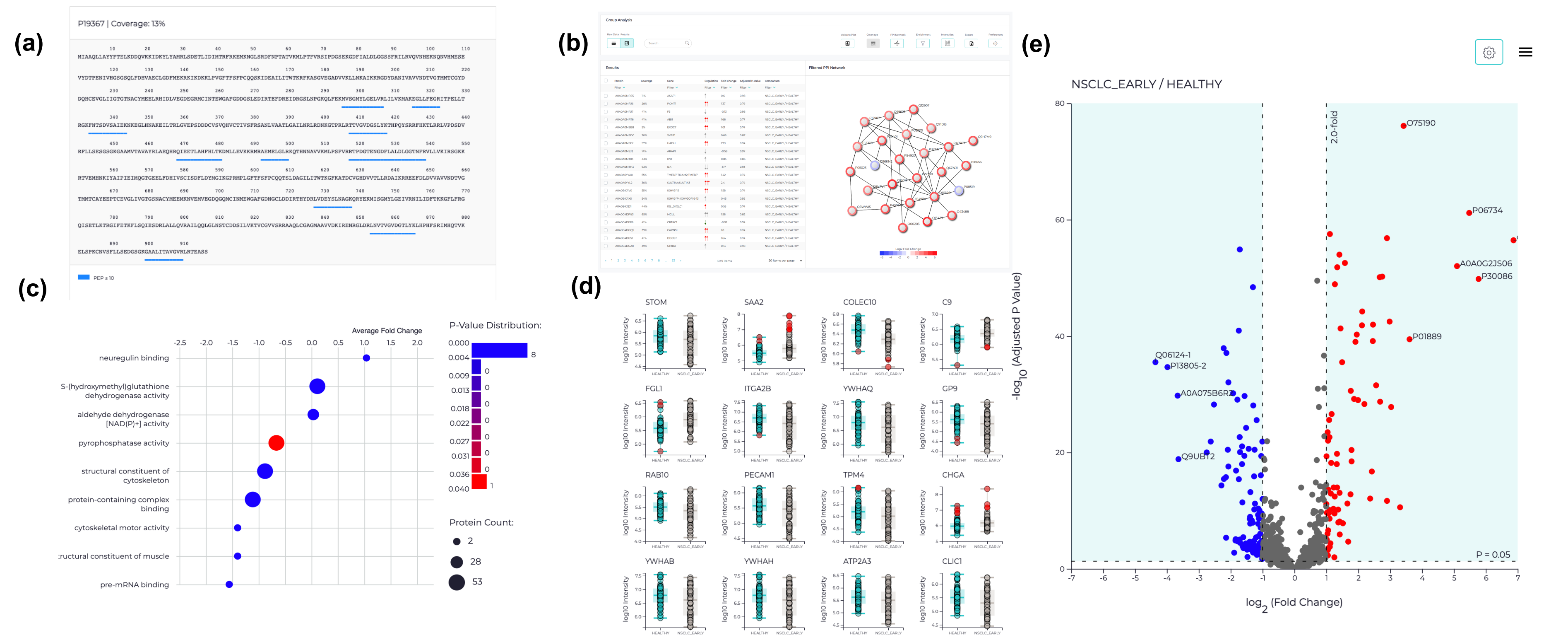
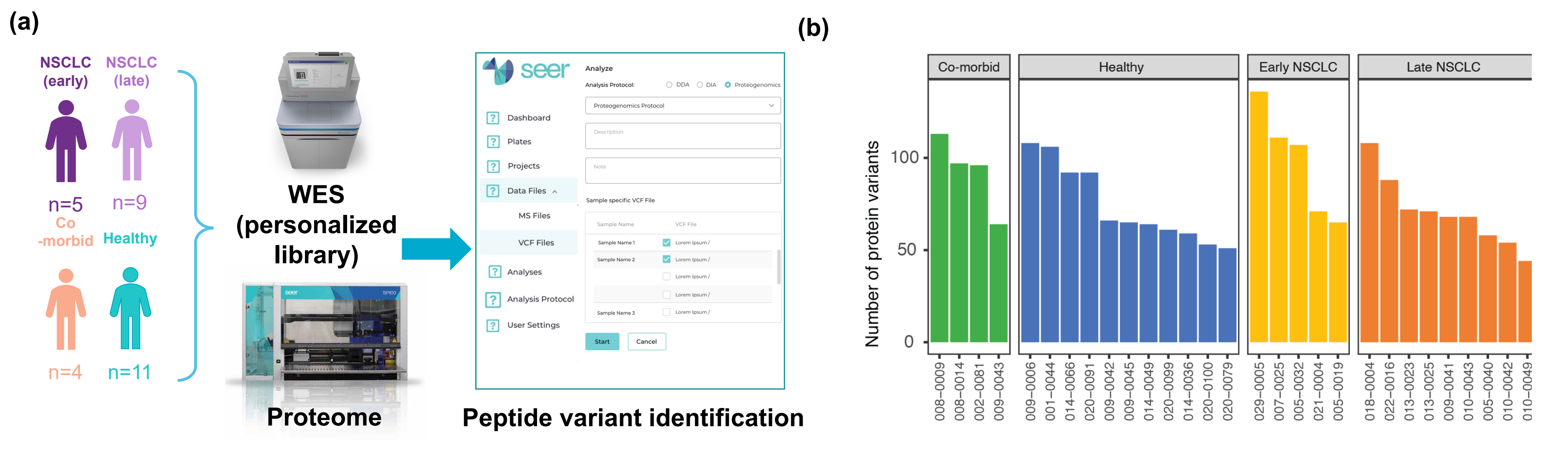


Figure 4. Group Analysis Results: **(a)** Sequence Coverage: Visualize where peptides map relative to the protein sequence. **(b)** Protein-Protein-Interactions Comparison: Build a STRING-based PPI network to identify differences in protein interactors. **(c)** GO Enrichment: Explore how proteins associated 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



Peptide variant identification using personalized libraries

Figure 5. (a) PAS can analyze VCF files generated from NGS pipelines in combination with mass spec data to identify peptide variants using personalized libraries. **(b)** An example of the variant peptides identified in 29 individuals grouped according to disease status.

Conclusions

We present a comprehensive proteogenomic 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. Searle, B.C et al. Nat Commun (2018)
3. Demichev, V et al. Nat Methods (2020)
4. Tyanova S et al. Nat Protocols (2016)

Publications