A Cloud-Scalable Software Suite for Large-Scale Proteogenomics Data Analysis and Visualization

Harsharn Auluck*, *Taylor Page, Aaron S Gajadhar, Margaret K.R. Donovan, Khatereh Motamedchaboki, Yuandan Lou, Theo Platt, and Serafim Batzoglou*

Seer, Inc., Redwood City, CA

The Proteograph^{TM} Analysis Suite is an intuitive, scalable, data informatics solution

The Proteograph Product Suite provides unbiased, deep, and rapid proteomics at scale

Assessment of the flow of genetic information through multi-omics 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-based proteomic analysis can be used to assess the proteome. Integration of proteomics and genomics data requires many tools of which require complex workflows that can act as a barrier for researchers. The Proteograph Analysis Suite (PAS) software application, included in Proteograph™ Product Suite[1], is a dedicated, cloud-based software solution removes barriers for proteogenomics researchers by enabling processing, analyzing, and visualizing proteomics data sets generated by liquid chromatography-mass spectrometry (LCMS).

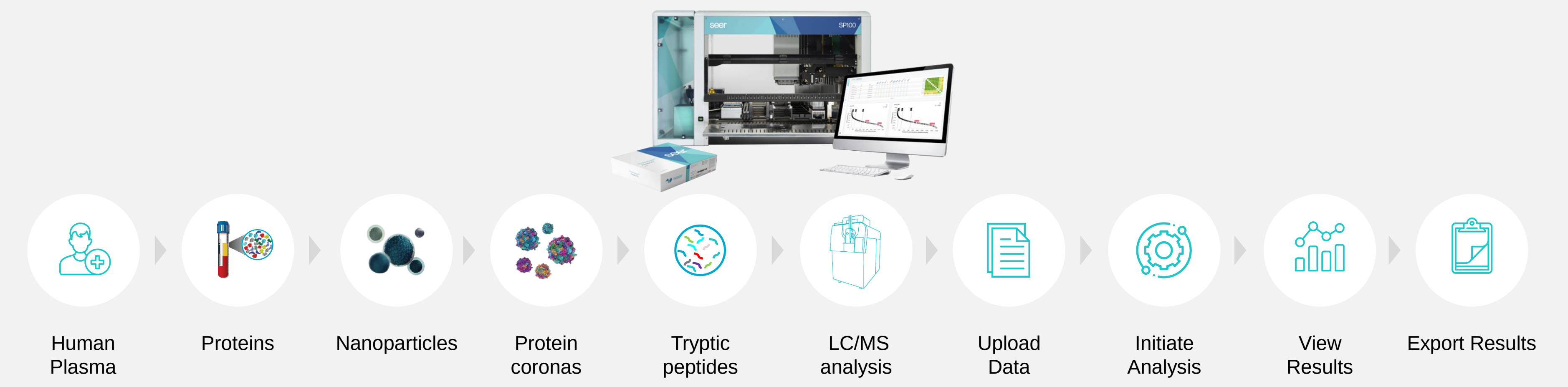


Figure 1. Proteograph Analysis Software (PAS) s a scalable on the cloud solution to integrate the data analysis for the entire Proteograph Product Suite including the Proteograph Assay Kit, SP100 automation instrument, and LCMS analyses.

Proteograph Analysis Suite enables 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 Acquisition (DIA) and Data Dependent Acquisition (DDA) Mass Spectrometry 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 peptide/protein intensities, 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 for large scale proteogenomics studies.

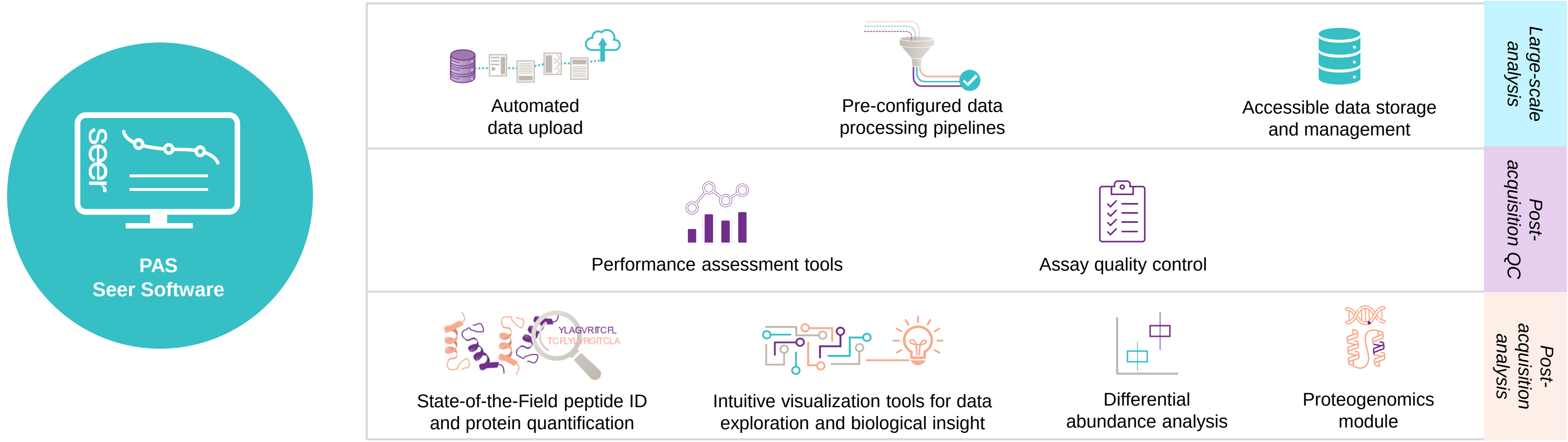

PAS
Seer Software
Automated data upload
Pre-configured data processing pipelines
Accessible data storage and management
Large-scale analysis
Performance assessment tools
Assay quality control
Post-acquisition QC
State-of-the-Field peptide ID and protein quantification
Intuitive visualization tools for data exploration and biological insight
Differential abundance analysis
Proteogenomics module
Post-acquisition analysis

Figure 2. Data is seamlessly transferred from MS computer to PAS without manual intervention using the AutoUploader tool in PAS. PAS features multiple tools addressing large-scale proteogenomics analysis, post-acquisition QC and data visualization.

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

The PAS data processing, analysis, and visualization workflow

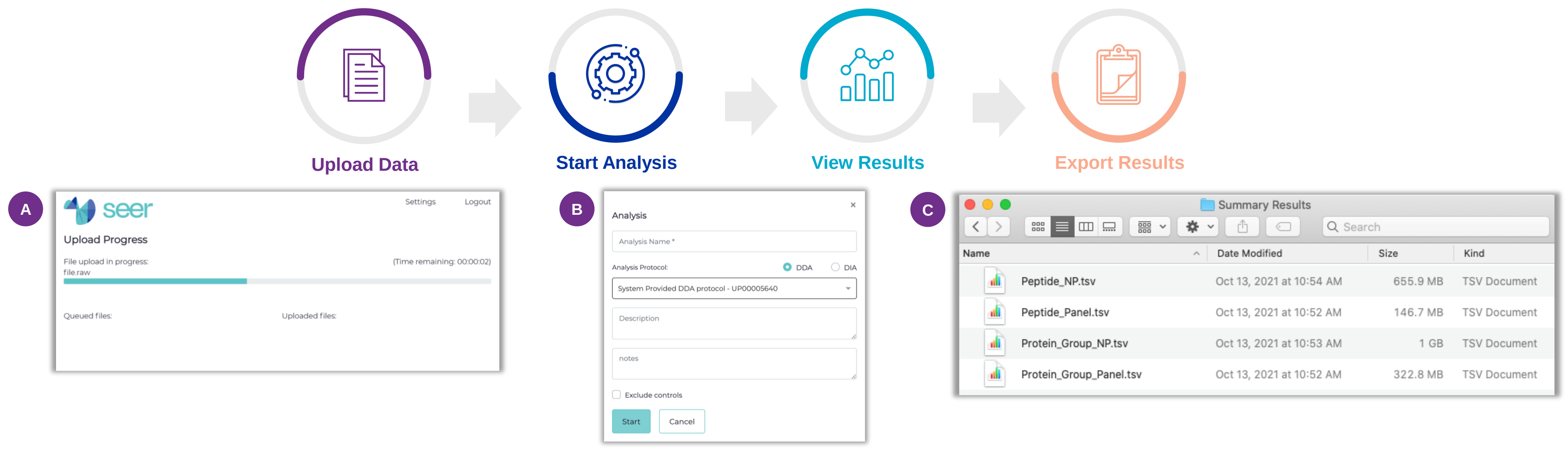


Figure 3. PAS workflow: (a) The AutoUploader tool automatically transfers raw LCMS data to your PAS account. **(b)** Raw DIA or DDA LCMS data can be analyzed using pre-configured data processing pipelines. **(c)** Processed data can be exported for further custom analysis.

QC tools enable assessment and assay performance monitoring

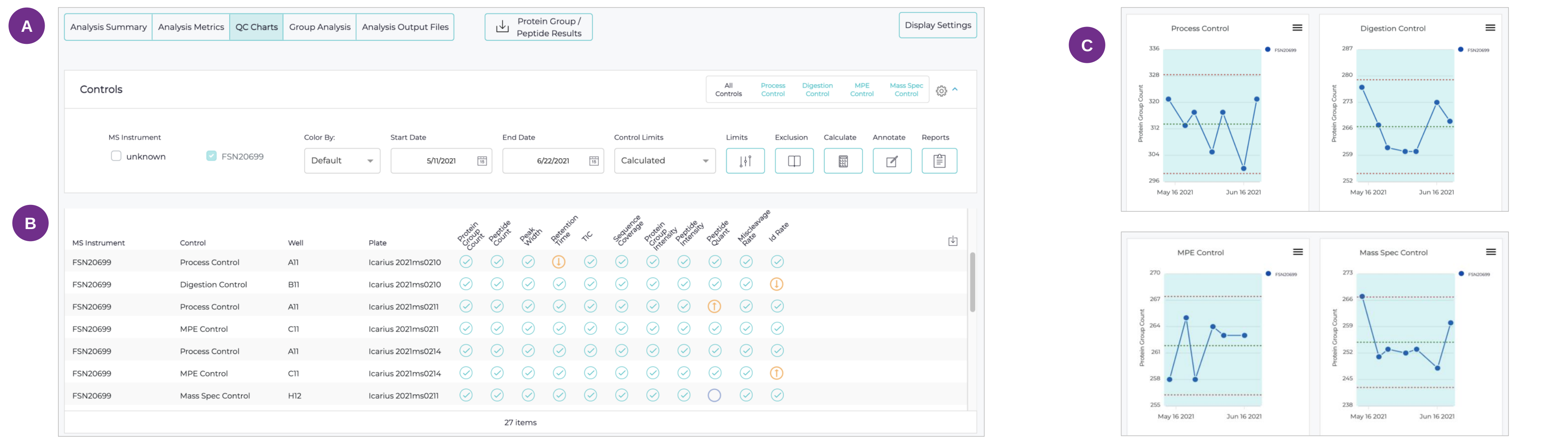


Figure 4. Control results dashboard: (a) Toolbar to select or filter control data, and to define expected limits. **(b)** Summary of control data for the selected analysis time frame. **(c)** QC charts with metrics for each control.

Data exploration and biological insights with PAS

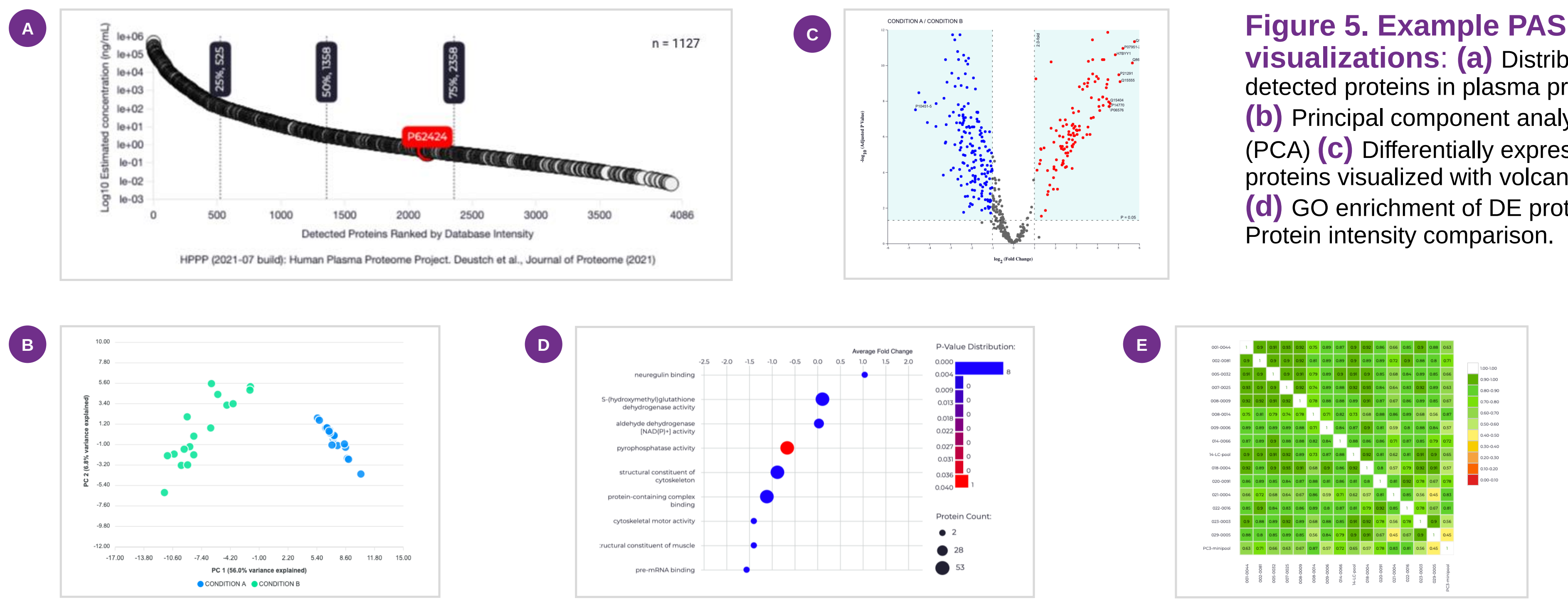
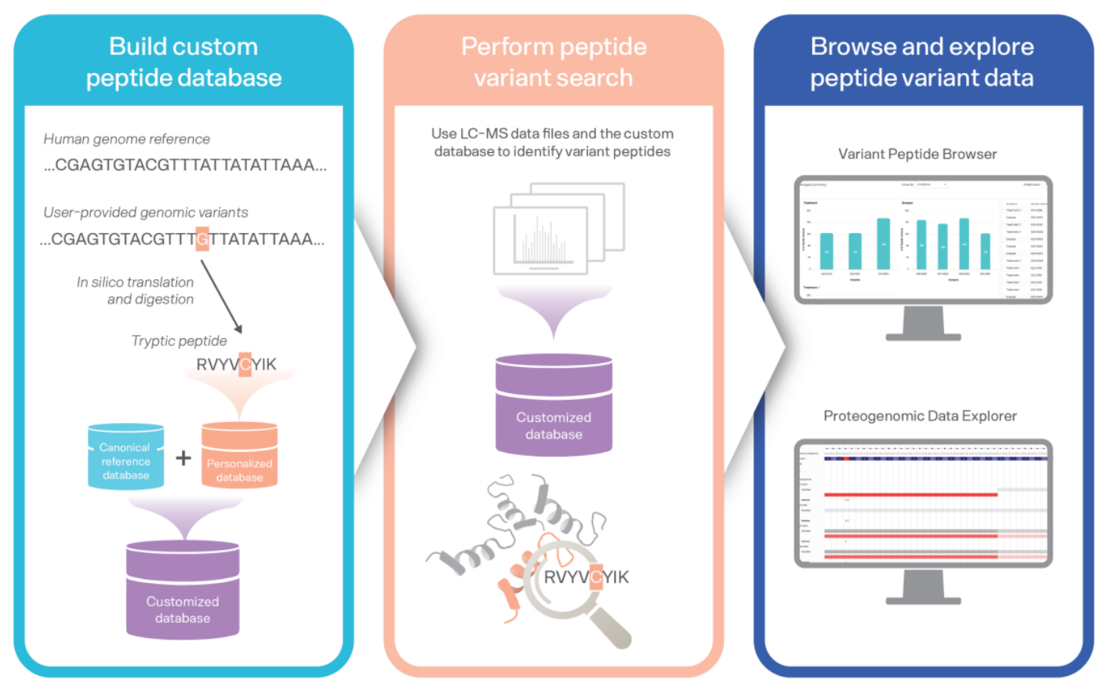


Figure 5. Example PAS data visualizations: **(a)** Distribution of detected proteins in plasma proteome **(b)** Principal component analysis (PCA) **(c)** Differentially expressed (DE) proteins visualized with volcano plot **(d)** GO enrichment of DE proteins. **(e)** Protein intensity comparison.

Identify and explore variant peptides with the PAS Proteogenomic workflow

Proteogenomics workflow links genomic variants with the proteome for variant peptide identification[2]


Build custom peptide database
Human genome reference
...CGAGTGTACGTTTATTATATTAAA...
User-provided genomic variants
...CGAGTGTACGTTTGTTATATTAAA...
In silico translation and digestion
Tryptic peptide
RVYVCYIK
Canonical reference database
Personalized database
Customized database
Perform peptide variant search
Use LC-MS data files and the custom database to identify variant peptides
Customized database
RVYVCYIK
Browse and explore peptide variant data
Variant Peptide Browser
Proteogenomic Data Explorer

A **Build a custom peptide database**: Use a custom or sample-specific vcf to predict protein altering variants not captured in the canonical reference database. Variant peptides are combined with the canonical reference database to generate a customized database.

B **Search for variant peptides**: Using the customized protein sequence database, search your LCMS DDA data for variant peptides utilizing MSFragger[3] search algorithm in PAS.

C **Browse and explore your variant peptide results:** Review variant peptide results with the Variant Peptide Browser and interact with results with the Proteogenomic Data Explorer.

Explore variant peptide results with the Variant Peptide Browser and Proteogenomic Data Explorer

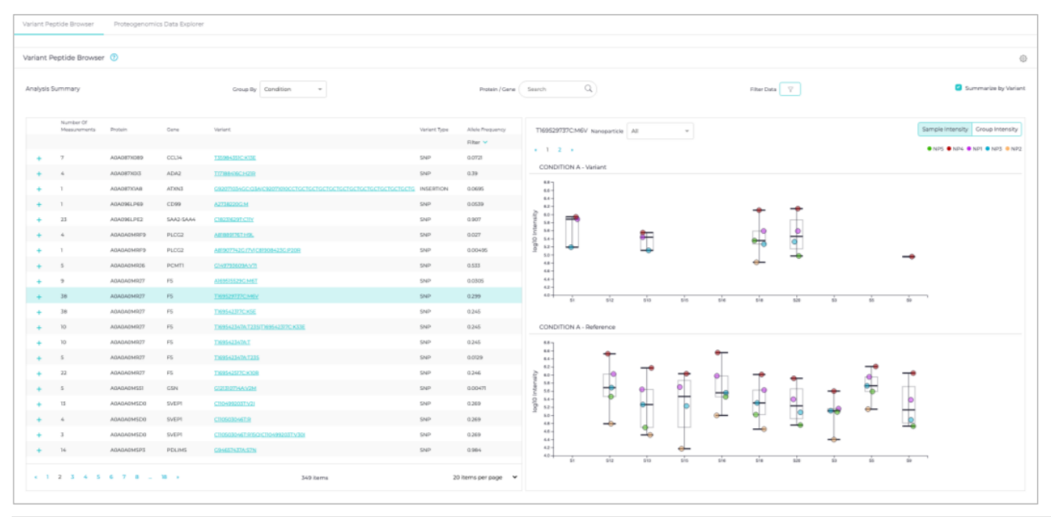


Figure 6. Variant Peptide Browser: Variant peptide search results are summarized in an interactive table and plots. The number of variant peptides per sample are summarized as a bar plot. Further, proteins of interest are searchable, and you can select a give protein harboring some variant peptide and explore the peptide intensity differences between the reference and variant peptide across samples.

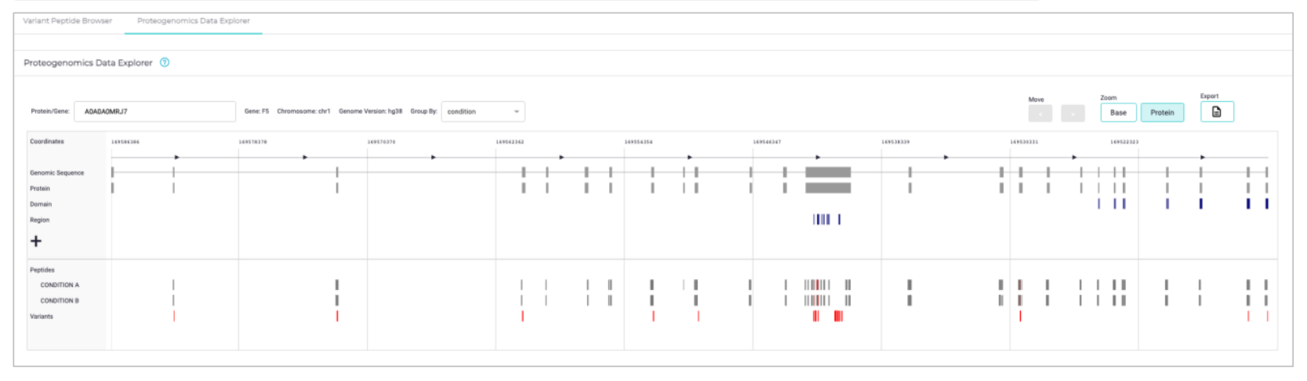


Figure 7. Proteogenomic Data Explorer: Browse how reference peptide and variant peptide data maps in genomic space at nucleic acid/amino acid and protein resolution. Visualize gene structure, protein domain information, and region information with respect to identified peptides.

Conclusion

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

References

1. Blume et al. Nat. Comm. (2020)
2. Donovan et al. BioRxiv (2022)
3. Nesvizhskii et al. Nature Methods (2017)

PAS Proteogenomics

Publications

[1]Seer, Inc., Redwood City, CA 94065, USA | *hauluck@seer.bio
[2]Massachusetts General Hospital, Boston, MA 02114 , USA