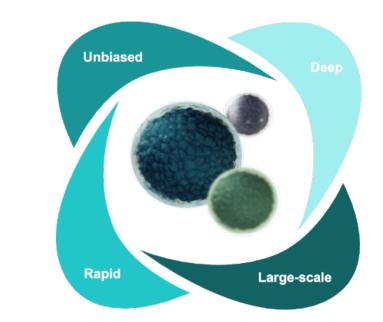
A Cloud-Scalable Software Suite for Large-Scale Proteogenomics Data Analysis and Visualization



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

Seer, Inc., Redwood City, CA 94065, USA

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

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

Introduction

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 Suite¹, 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 (LC-MS).

Protein

coronas

Nanoparticles



The PAS data processing, analysis, and visualization workflow





View Results

Export Results

 (\bigcirc)

Initiate

Analysis

Upload Data







A	seer		Settings Logout		Analysis			● ● ● ■ Summary Results < > □□□□□□□□□□□ □□ □ □ □ □ Q Search						
	Upload Progress		(Time remaining: 00:00:02)		Analysis Name *		Name		 Date Modified 	Size	Kind			
	file.raw		(Analysis Protocol: ODA System Provided DDA protocol - UP00005640			Peptide_NP.tsv	Oct 13, 2021 at 10:54 AM	655.9 MB	TSV Document			
	Queued files:	Uploaded files:			Description			Peptide_Panel.tsv	Oct 13, 2021 at 10:52 AM	146.7 MB	TSV Document			
								Protein_Group_NP.tsv	Oct 13, 2021 at 10:53 AM	1 GB	TSV Document			
					notes	<i>li</i>		Protein_Group_Panel.tsv	Oct 13, 2021 at 10:52 AM	322.8 MB	TSV Document			
					Exclude controls									
					Start Cancel									

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

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 LC-MS analyses.

LC/MS

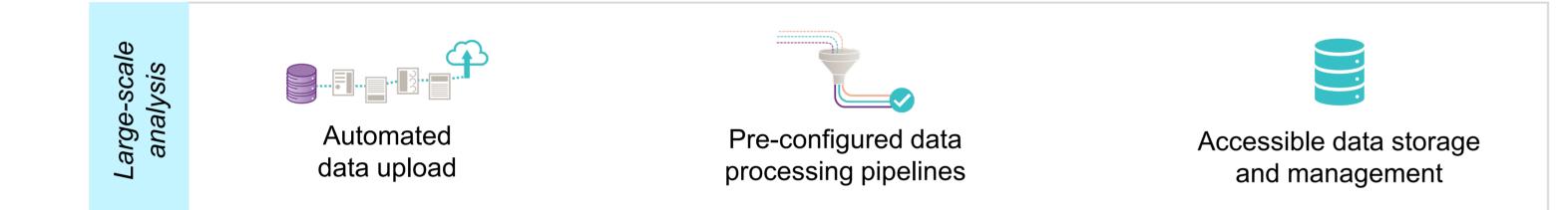
analysis

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

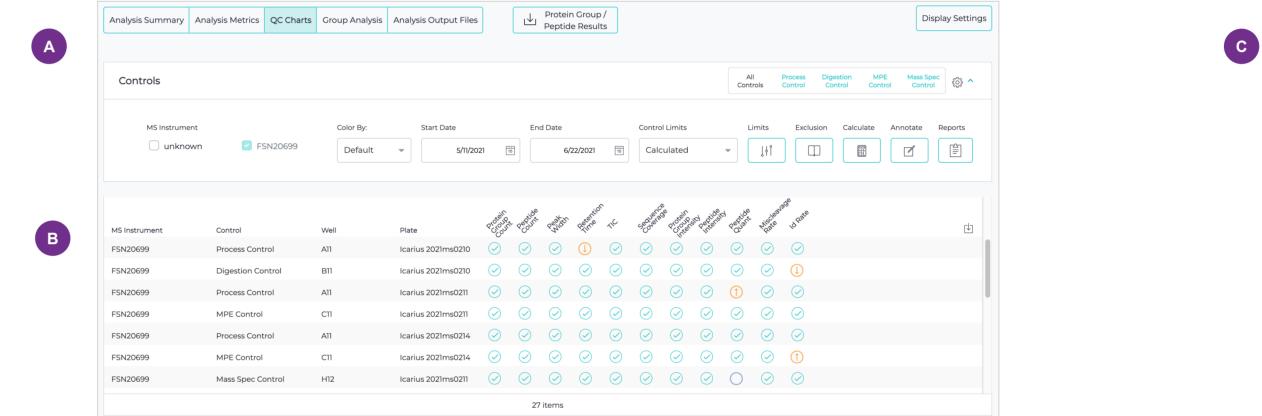
Tryptic

peptides

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.



QC tools enable assessment and assay performance monitoring





267 266 •	
g 264 g 259 g 259	•
the constraints of the constrain	1
	•
255 238	
May 16 2021 Jun 16 2021 May 16 2021 Jun 1	6 2021

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

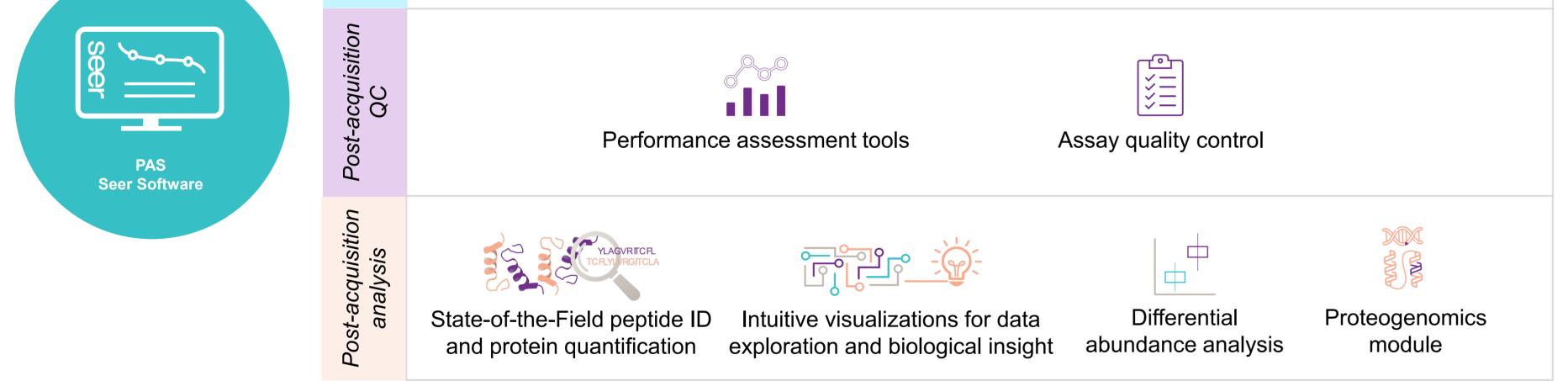
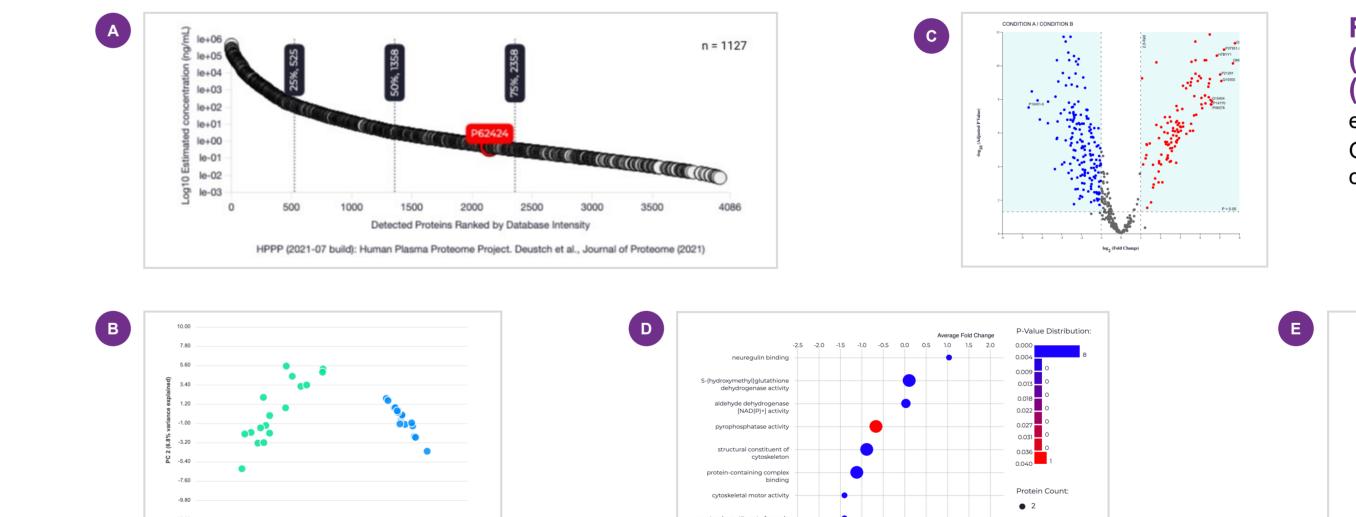
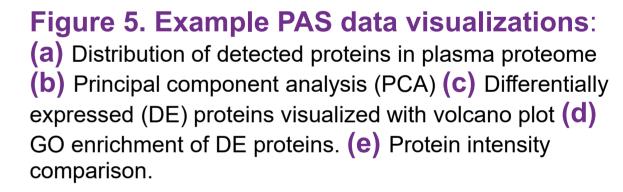
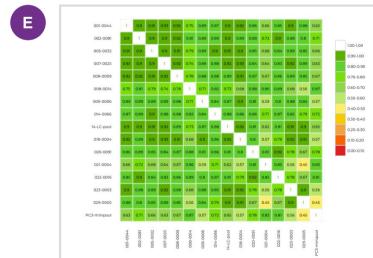


Figure 2. Data is seamlessly transferred from MS computer to PAS without manual intervention using the AutoUploader tool in PAS. PAS features multiple tools addressing large-scale proteogenomics analysis, post-acquisition QC and data visualization.



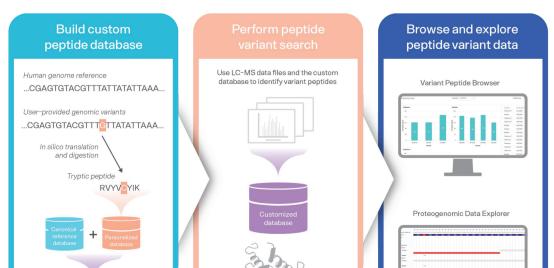
pre-mRNA binding





Identify and explore variant peptides with the PAS Proteogenomic workflow

Proteogenomics workflow links genomic variants with the proteome for variant peptide identification²



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.

Search for variant peptides: Using the customized protein sequence database, search your LC-MS DDA data for variant peptides utilizing MSFragger³ search algorithm in PAS.

Browse and explore your variant peptide results:

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

-17.00 -13.80 -10.60 -7.40 -4.20 -1.00 2.20 5.40 8.60 11.80 15.00

PC 1 (56.0% variance explained)

CONDITION A CONDITION B

Variant Pe	eptide Browser	Proteogenomic	cs Data Explorer														
Variant P	Peptide Browser	0															4
Analysis	Summary			Group By Condition +		Protein / Gene	Search Q					Filter Data	7				Summarize by Varian
	Number Of Measurements	Protein	Cene	Variant	Variant Type	Alele Frequency	T169529737C:M6V Nan	oparticle A	All	*							ample Intensity Group Intensity
+	7	A0A087X089	CCL14	<u>73598435\CKUE</u>	SNP	0.0721	«12»										NP5 NP4 NP1 NP3 NP2
+	4	A0A087X0I3	ADA2	117188416CH21R	SNP	0.39	CONDITION A - Variar	nt									
+	1	A0A087X1A8	ATXN3	6920710346C-03AVC92071010CCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTG	INSERTION	0.0695	6.8 - 6.6 -										
+	1	A0A096LP69	CD99	A2738220G-M	SNP	0.0539	6.4 - 6.2 -					-	-				
+	23	A0A096LPE2	SAA2-SAA4	C18231629T-C11Y	SNP	0.907	zi eo-					Ť	rth -				
+	4	A0A0A0MRF9	PLCG2	<u>A81889176T.H9L</u>	SNP	0.027	12 5.8- 01 5.4- 05 5.2-		\$				•				
+	1	A0A0A0MRF9	PLCG2	A81907742C/7V/C81908423C/P20R	SNP	0.00495	5.0 -		ц				1				
+	5	A0A0A0MR36	PCMTI	S149793609A3/71	SNP	0.533	4.8 - 4.6 -					-					
+	9	A0A0A0MR37	FS	A169515529C.M6T	SNP	0.0305	44- 42-										
+		A0A0A0MR37	FS	<u>Ti69529737C:M6V</u>	SNP	0.299	4.0 51	\$12	\$13	\$15	\$16	S18	\$20	ŝ	śś	\$9	3
+	38	AQAQAQMR37 AQAQAQMR37	F5 F5	T169542317CKSE	SNP	0.245	CONDITION A - Refer										
+	10	ADADADMRJ7 ADADAOMRJ7	FS	1169542347A-123511169542577C-K358	SNP	0.245	68 つ	rence									
1	5	A0A0A0MR37	FS	T169542347A-T235	SNP	0.0129	6.8 - 6.4 -	+			•_						
	22	A0A0A0MRJ7	FS	169542577CK08	SNP	0.246	62- ≧ 60-									-	

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

Variant Peptide Brow	wser Proteogenomics D	ata Explorer							
Proteogenomics I	Data Explorer								
Protein/Gene: A0A	0A0MRJ7	Gene: F5 Chromosome: chr1 Genon	e Version: hg38 Group By: condit	on 👻					
Coordinates	169586386	169578378	169570370	169562362		16955	1354	169546347	
Genomic Sequence					+				
Protein Domain		1					1 1 1		
Region								101 1	
Peptides									
CONDITION A	l l				1		1 1		

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



Review variant peptide results with the Variant Peptide Browser and interact with results with the Proteogenomic Data Explorer.



peptide intensity differences	CONDITION B
between the reference and variant	Variants
peptide across samples.	
μ eplice across samples.	





