Deep Plasma Proteomics at Scale Enabling Precision Analyses in a Lung Cancer (NSCLC) Study

Mahdi Zamanighomi*, Harendra Guturu, Jian Wang, Amir Alavi, Tristan Brown, Daniel Hornburg, Moaraj Hasan, Shadi Ferdosi, Khatereh Motamedchaboki, Margaret Donovan, Theodore Platt, Ryan Benz, Asim Siddiqui, and Serafim Batzoglou

Deep and unbiased plasma proteomics for disease cohort studies at scale

Introduction

Our ~20,000 genes encode over one million protein variants, given alternative splice forms, allelic variation, and protein modification. Though large-scale genomics studies have expanded our understanding of biology, similarly, scaled deep and untargeted proteomics studies of biofluids have remained impractical due to complex workflows. To address this need, we have previously described Proteograph[™], a novel platform that leverages the protein-corona interactions of nanoparticles (NP) for deep and untargeted proteomic sampling at scale.

Using Proteograph in a non-small cell lung cancer (NSCLC) cohort, we previously conducted a deep interrogation of plasma from using early-stage NSCLC subjects and non-cancer controls. We identified 2,499 plasma proteins, with 1,992 present in \geq 25% of the samples. Leveraging this data, we created a biomarker classifier distinguishing NSCLC from controls with average area under the receiver operating characteristic curve of 0.91.1 In this study, we now re-analyze the data with the recently released DIA-NN software and leveraged cloud architecture to successfully scale up and process large cohort group runs. This enabled enhanced proteome depth (41% increase) while improving the accuracy (0.95) of the classifier. Our results outline workflows for robust biomarker discovery and cohort subtyping.

The Proteograph[™] platform interrogates the plasma proteome at previously impractical combinations of scale, depth and coverage, and enables the development of improved classification models and marker discovery.

Methods

Cohort was analyzed with DIA-NN v1.8 in single group-run in library free mode against standard human uniport proteome using the --relaxedprof-inf option².

Differential expression of proteins was computed using DIA-NN estimated log₁₀(1 + intensities) and Welch's t-test.

Functional enrichment analysis was performed using g:Profiler (version e104_eg51_p15_3922dba) with g:SCS multiple testing correction method applying significance threshold of 0.01³.

Cohort classification was done with implementations of several machine learning classifiers applied to protein intensities and their embedding from a variational auto-encoder (VAE) that will allow for easier integration and visualization of proteogenomics data in the future.



Figure 1. Architecture diagram of VAE neural network.



Each rectangle denotes a fully connected block. Laplace and Zero Inflated Negative Binomial (ZINB) distributions are considered to encourage sparsity and to model sparsity.







DIA-NN combined with the cloud-based large group run more sensitively identifies protein groups across the cohort (~4000 vs the previous ~ 2.500).



Deep proteomics enhances disease classification and biomarker characterization of early NSCLC cohort



Figure 2. Distribution of identified protein group counts.

A) Using DIA-NN, we identified nearly 4,000 protein groups across the cohort (on average >2,000 per sample vs 439 in neat plasma) with the majority being supported by multiple peptides and found in multiple samples. B) The Proteograph[™] successfully detected a median of 8 peptides per protein.



Figure 3. Differentially expressed protein groups across cohort.

A) The Welsh's t-test on protein intensities resulted in 17 unique protein groups that were significantly differentially expressed after Bonferroni corrected threshold of 0.01. Protein groups in red indicate an association to NSCLC was found in literature. B) Hierarchical clustering on the differentially expressed protein intensities depicts a distinct separation of cancer and healthy subjects.

 \bigcirc



Enhancement of classification performance and stability with VAE, Random Forest, and SVM to AUC-ROC 0.95 compared to the previous AUC-ROC 0.91.

Copyright Seer, Inc 2022 Seer, Inc., Redwood City, CA 94065, USA | *mzamanighomi@seer.bio

name	GO:ME		stats																	
contained intend print prin print print	Torm pama	Torm ID				POD	216627	P623	P018	P027	Q926	Q927	G3V2	P051	P027	P067	Q999	Q060 MOR1	P005	Q96K
aip water 00000000 00000000 00000000 00000000 000000000 000000000 000000000 0000000000 0000000000 0000000000 0000000000 0000000000 0000000000 0000000000 0000000000 00000000000 000000000000000000000000000000000000	ierm name	Termito	Padj	0 -109 ₁₀ (Padj)	≤16	8	-2	28	76	6	26	43	89	60	41	02	88	Q 33	58	N2
opinume response anisoning COUCUSESS2 1 1 5 4 4 4 9 1 Image: Counce of the counc	signaling receptor binding	GO:0005102	6.151×10 ⁻⁴	_		_			-	-		-	-						-	-
unamed 00000542 4.818-10 ³ 0 <th>Ioli-like receptor 4 binding</th> <th>GO:0035662</th> <th>1.192×10 °</th> <th>-</th> <th></th> <th>-</th> <th>-</th> <th></th> <th>-</th> <th>+</th> <th>+</th> <th>-</th> <th>-</th> <th></th> <th>-</th> <th></th> <th></th> <th>-</th> <th>+-</th> <th>-</th>	Ioli-like receptor 4 binding	GO:0035662	1.192×10 °	-		-	-		-	+	+	-	-		-			-	+-	-
Constrained COUSSIDAR 4 User (P) ² CoussiDAR 4 User (P) ² CoussiDAR Couss	icosanoid hinding	GO:0050544	2.977×10^{-3}	-		-			-	+	-	-	-					-	+	-
BAGE respects binding OC0050786 Byth (0) Interplant Inter	icosatetraenoic acid binding	GO:0050542	4.165×10^{-3}	-			-		+	+	-	-	-					-	+	-
Output tates masked maskdddddddddddddddddddddddddddddddddddd	RAGE receptor binding	GO:0050786	8.911×10 ⁻³	-		-			-	-	-	t	-						-	1
0.1)P tata P<											1 to	6 6	of 6	ŀ	< <	Pa	age '	1 of 1	>	>1
Term name Term ID Pail	GO:BP		stats				Q16	-					G			ч	Q	~ 0	P	Q
aute Humanatery response CO000256 1.059+10 * 0 <td>Term name</td> <td>Term ID</td> <td>Padj</td> <td>o -log₁₀(p_{adj})</td> <td>≤16</td> <td>PODJI8</td> <td>6627-2</td> <td>62328</td> <td>01876</td> <td>02763</td> <td>92626</td> <td>292743</td> <td>3V2B9</td> <td>05109</td> <td>P02741</td> <td>06702</td> <td>88666</td> <td>06033 00R1Q1</td> <td>00558</td> <td>96KN2</td>	Term name	Term ID	Padj	o -log ₁₀ (p _{adj})	≤16	PODJI8	6627-2	62328	01876	02763	92626	292743	3V2B9	05109	P02741	06702	88666	06033 00R1Q1	00558	96KN2
aduel-phase regionse aduel-phase regionse adoes not see regionse addees	acute inflammatory response	GO:0002526	1.059×10 ⁻⁸				—					Г							Т	Г
inflammane 000006661 3810-10 ⁻⁰ 0 <td< td=""><td>acute-phase response</td><td>GO:0006953</td><td>1.309×10⁻⁶</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></td<>	acute-phase response	GO:0006953	1.309×10 ⁻⁶																	
defense reporte 00000892 1.485.10 ⁴ 1 1	inflammatory response	GO:0006954	3.910×10 ⁻⁶																	
response to stress 000000689 1000-109 requised excytosis 000000689 1564-109 requised excytosis 000001762 1000-109 100	defense response	GO:0006952	1.485×10 ⁻⁵																	
meturophil demotybil aggregation GO-0070488 1.880×10 ⁻¹ Implication	response to stress	GO:0006950	1.009×10 ⁻³															_		L
mainto response GO 0006993 3.54810 ⁻¹ Image: Second	neutrophil aggregation	GO:0070488	1.380×10 ⁻³								_								+	-
minumer regione G0:006865 3.71s.10 ⁻¹ 0 0	neutrophil chemotaxis	GO:0030593	3.546×10 ⁻³								_								-	-
regulate degramation GOUDED B 434 ND ⁻ A 40 ND ⁻	immune response	GO:0006955	3.715×10 ⁻³								-									-
0.0.3902600 2.428/10 ⁻¹ gamular deformation G0.0071621 2.981/10 ⁻¹ 0.0.001612 0.0007678 8.500/10 ⁻¹ 0.0.001612 0.0007678 0.00071621 0.00071621 0.0.001612 0.00071621 0.00071621 0.00071621 0.00071621 0.0.0001612 0.00071621 0.00071	regulated exocytosis	GO:0045055	6.434×10 ⁻³	-								-							4	-
grammory channetables OU-UNICL1 ZMBYN P A A A A	neutrophil migration	GO:1990266	7.245×10 ⁻³	-					-	-	-	-	-					_	-	-
partner GGU 0002/r0 8.358/10 ² Partner	granulocyte chemotaxis	GO:0071621	7.981×10 ⁻³									-						_	H	-
Velocie Basicio Basicio <t< td=""><td>platelet degranulation</td><td>GO:0002576</td><td>8.500×10⁻³</td><td></td><td></td><td></td><td>_</td><td></td><td></td><td>_</td><td>-</td><td>-</td><td></td><td></td><td></td><td></td><td></td><td></td><td>-</td><td>-</td></t<>	platelet degranulation	GO:0002576	8.500×10 ⁻³				_			_	-	-							-	-
delenter regularse to outrier organism GO.0038542 9 / 2 / 2 / 2 / 2 / 2 / 2 / 2 / 2 / 2 /	vesicle-mediated transport	GO:0016192	9.535×10 ⁻³			_				-	-						_	_	4	-
Stats No. 10. 10. 10. 10. 10. 10. 10. 10. 10. 10	derense response to other organism	GO:0098542	9.767×10 °								l to 1	4 of	: 1/1			Da		1 of 1		
Term name Term ID Pell -log to Pell Pell <th< td=""><td>00:00</td><td></td><td>stats</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td>4 01</td><td>14</td><td></td><td></td><td>Fo</td><td>age</td><td></td><td></td><td></td></th<>	00:00		stats									4 01	14			Fo	age			
ranke point point <th< td=""><td>Term name</td><td>Term ID</td><td>Dedi</td><td>-log₁₀(p_{edi})</td><td></td><td>PODJ</td><td>Q16627-</td><td>P6232</td><td>P0187</td><td>P0276</td><td>Q9262</td><td>Q9274</td><td>G3V2B</td><td>P0510</td><td>P0274</td><td>P0670</td><td>8666Ď</td><td>Q0603 MOR10</td><td>P0055</td><td>Q96KN</td></th<>	Term name	Term ID	Dedi	-log ₁₀ (p _{edi})		PODJ	Q16627-	P6232	P0187	P0276	Q9262	Q9274	G3V2B	P0510	P0274	P0670	8666Ď	Q0603 MOR10	P0055	Q96KN
extracellular region GC00005/76 4.585×10 ⁻⁷ 0 <td></td> <td></td> <td>Padj</td> <td>0 10910(Padj)</td> <td>≤16</td> <td></td> <td>'n</td> <td>ö</td> <td>6</td> <td>ŭ</td> <td>5 6</td> <td>ü</td> <td>ő</td> <td>ğ</td> <td><u>1</u></td> <td>2</td> <td>õ</td> <td>2 8</td> <td></td> <td>12</td>			Padj	0 10910(Padj)	≤16		'n	ö	6	ŭ	5 6	ü	ő	ğ	<u>1</u>	2	õ	2 8		12
extra cellular existeme G0:000002 1.37 x10 ⁻¹ Image: Complex comp	extracellular region	GO:0005576	4.585×10 ⁻⁵						+		+	-	┢						-	
Autore (a) space 00-000010 2.03 k10 0		GO:0070082	2.021×10 ⁻⁷			_			+		╈	┢	┢							
$ \begin{array}{c c c c c c c c c c c c c c c c c c c $	extracellular space	GO:1003561	2.031×10^{-7}			_			+		+	-	┢							
Extracellular method metod regarine 000004320 2.240x10 ⁻⁷ 0 0	extracellular membrane-bounded organelle	GO:1905501	2.210×10						╋	╋	╋	┢	┢		-		\vdash	+		
Screetery granule lumen G0:0034774 9.444x10 ⁻⁷ 9.44x10 ⁻⁷ 9.44x		GO:0043230	2.240×10^{-7}			-			+	+	╈	┢	┢	\vdash		\vdash	\vdash	+		
corrors 000000000 1.029x10 ⁻⁶ 0 0	secretory granule lumen	GO:0034774	9 444×10 ⁻⁷																	1
vesicle lumen G0:0031983 1.074×10 ⁻⁶ Image: Containing extracellular matrix G0:0031982 1.074×10 ⁻⁶ Image: Containing extracellular matrix G0:0031982 2.230×10 ⁻⁶ Image: Containing extracellular matrix G0:003102 4.470×10 ⁻⁶ Image: Containing extracellular matrix G0:00000000000000000000000000000000000	cytoplasmic vesicle lumen	GO:0060205	1.029×10 ⁻⁶				-				-	-								1
collagen-containing extracellular matrix G0:0062023 6.244×10 ⁻⁶ Image: Collagen-containing extracellular matrix G0:0031982 2.230×10 ⁻⁵ Image: Collagen-containing extracellular matrix Image: Collagen-containing extracellular matrix G0:0031982 2.230×10 ⁻⁵ Image: Collagen-containing extracellular matrix Image: Collagen-containing extracellular matrix G0:0030312 4.470×10 ⁻⁴ Image: Collagen-containing extracellular matrix Image: Collagen-containing extracellular matrix G0:00303141 7.547×10 ⁻⁴ Image: Collagen-containing extracellular matrix Image: Collagen-containing extracellular matrix G0:00303141 7.547×10 ⁻⁴ Image: Collagen-containing extracellular matrix	vesicle lumen	GO:0031983	1.074×10 ⁻⁶								+									t
vesicle G0:0031982 2.230×10 ⁻⁵ Image: Comparison of the compa	collagen-containing extracellular matrix	GO:0062023	6.244×10 ⁻⁶						1			i -							1	t
extracellular matrix G0:0031012 4.470×10 ⁻⁵ Image: Comparison of the secretary granule G0:0030312 4.524×10 ⁻⁵ Image: Comparison of the secretary granule Image: Comparis	vesicle	GO:0031982	2.230×10 ⁻⁵																i se	E
external encapsulating structure G0:0030312 4.524×10 ⁻⁵ Image: Comparison of the structure G0:003011 7.547×10 ⁻⁴ secretory yesicle G0:003013 2.479×10 ⁻³ Image: Comparison of the structure Image	extracellular matrix	GO:0031012	4.470×10 ⁻⁵										\square							Γ
secretory granule G0:0030141 7.547×10 ⁻⁴	external encapsulating structure	GO:0030312	4.524×10 ⁻⁵																	Г
secretory vesicle G0:009903 2.479×10 ⁻³ Image: Comparison of the comparison of	secretory granule	GO:0030141	7.547×10 ⁻⁴																	Γ
platelet alpha granule lumen GO:0031093 4.068×10 ⁻³ Image: Comparison of the co	secretory vesicle	GO:0099503	2.479×10 ⁻³																	Γ
REAC stats and and <t< td=""><td>platelet alpha granule lumen</td><td>GO:0031093</td><td>4.068×10⁻³</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></t<>	platelet alpha granule lumen	GO:0031093	4.068×10 ⁻³																	
REAC stats 0<									_		1 to 1	6 of	16	ŀ	< <	Pa	age '	1 of 1	>	X
Term name Term ID Padj -logn(padj) ate District of the terotetramer CORUM CORUM:6826 5.510x10^-5 Action A	REAC		stats			P	Q166	P6	PO	PO	Qg	QS	G3	PC	PC	PO	Qg	MO	PO	Q9
Platelet degranulation REAC:R-HSA-1 3.648×10 ⁻³ Image: Constraint of the second of the se	Term name	Term ID	P _{adj}	o -log ₁₀ (p _{adj})	≤16	0DJI8	627-2	2328	11876	2763)2626	92743	V2B9	15109	02741	6702	8866)R1Q1	0558	6KN2
Response to elevated platelet cytosolic Ca2+ Metal sequestration by antimicrobial proteins REAC:R-HSA-6 4.261×10 ⁻³ Image: Carbon control of the contro	Platelet degranulation	REAC:R-HSA-1	3.648×10 ⁻³																	
Metal sequestration by antimicrobial proteins REAC:R-HSA-6 6.268×10 ⁻³ Image: Constraint of the sequestration by antimicrobial proteins REAC:R-HSA-6 6.268×10 ⁻³ Image: Constraint of the sequestration by antimicrobial proteins Image: Constraint of the sequestration by antimicrobial proteins Image: Constraint of the sequestration by antimicrobial proteins REAC:R-HSA-6 6.268×10 ⁻³ Image: Constraint of the sequestration by antimicrobial proteins REAC:R-HSA-6 6.268×10 ⁻³ Image: Constraint of the sequestration by antimicrobial proteins Image: Constraint of the sequestration by antipication	Response to elevated platelet cytosolic Ca2+	REAC:R-HSA-7	4.261×10 ⁻³																	
WP stats 0 <td>Metal sequestration by antimicrobial proteins</td> <td>REAC:R-HSA-6</td> <td>6.268×10⁻³</td> <td></td>	Metal sequestration by antimicrobial proteins	REAC:R-HSA-6	6.268×10 ⁻³																	
WP stats o <td></td> <td>1 to</td> <td>30</td> <td>of 3</td> <td>ŀ</td> <td>< <</td> <td>Pa</td> <td>age '</td> <td>1 of 1</td> <td>></td> <td>> </td>											1 to	30	of 3	ŀ	< <	Pa	age '	1 of 1	>	>
Term name Term ID padj -log10(padj) s16 Dig S0	WP		stats			PO	Q166	P62	PO	PON	Q9	Q9.	G3V	PO	PO	POG	Q99	MOR	POC	960
Vitamin B12 Metabolism WP:WP1533 8.676×10 ⁻³ Image: Construction of the state of the	Term name	Term ID	P _{adj}	o -log ₁₀ (p _{adj})	≤16	DJI8	27-2	2328	1876	2763	2626	2743	/289	5109	2741	6702	8866	6033 R1Q1	0558	SKN2
Ito 1 of 1 K < Page 1 of 1 > 1 CORUM stats 0	Vıtamin B12 Metabolism	WP:WP1533	8.676×10⁻³																	
CORUM stats 0											1 t	o 1 (of 1		< <	Pa	age '	1 of 1	>	>
Term name Term ID padj -log10(padj) ≤16 0	CORUM		stats				Q1						_				0	, ,	-	
Calprotectin heterotetramer CORUM:6826 5.510×10 ⁻⁵ Image: Core of the second seco	Term name	Term ID	Padj	o -log ₁₀ (p _{adj})	≤16	PODJI8	16627-2	P62328	P01876	P02763	Q92626	Q92743	G3V2B9	P05109	P02741	P06702	886665	Q06033 M0R1Q1	P00558	296KN2
iNOS-S100A8/A9 complex CORUM:6827 1.652×10 ⁻⁴	Calprotectin heterotetramer	CORUM:6826	5.510×10 ⁻⁵		10		1		T	T	T	T	Ē						f	f
	iNOS-S100A8/A9 complex	CORUM:6827	1.652×10 ⁻⁴																	

Figure 4. Functional enrichment of differentially expressed proteins.

Detailed list of enriched terms after correction for multi-correction testing using g:Profiler's g:SCS multiple hypothesis testing method that accounts for relationships between ontology terms.

Differentially expressed proteins can lead to a better biomarker discovery and disease characterization.

References







Figure 5. PCA projection of the cohort.

PCA of the A) protein group intensity matrix and B) VAE embedding visualizes the ability of the VAE to better separate the cohort based on disease status in the latent space.



Figure 6. Classification performance using protein intensities and a VAE embedding.

The ROC curve using the A) protein group intensity matrix and B) VAE embedding reflects the classification power of different methods. All commonly used classifiers are robust and accurate in the VAE embedding.

¹Blume et al. *Nat. Comm.* (2020) ² Demichev et al. *Nat Comm.* (2020) ³Raudvere, Kolberg et al. *Nucleic Acids Res.* (2019)

