Systematic analysis of DIA LC-MS protein rollup strategies and their impact on phenotype association and proteogenomic applications

Harendra Guturu[1]*, Guhan Venkataraman[1], Jian Wang[1], Yingxiang Huang[1], Amir Alavi[1], Ryan Benz[1], Khatereh Motamedchaboki[1], Matthijs B. De Geus[2], Sudeshna Das[2], Pia Kivisäkk[2], Steven E. Arnold[2], Asim Siddiqui[1], and Serafim Batzoglou[1]

[1]Seer, Inc., Redwood City, CA 94065, USA, [2]Massachusetts General Hospital, Boston, MA 02114, USA

The Proteograph™ Product Suite enables rapid sample processing for reproducible, deep plasma proteomic analysis

Introduction

Nanoparticle-based sample preparation coupled with liquid chromatography–mass spectrometry (LC–MS) enables untargeted measurement of the proteome from biofluids at unprecedented depth.[1] Deep plasma proteome coverage enables previously inaccessible applications such as phenotype diagnosis, biomarker discovery and proteogenomics from easy to obtain plasma samples. But, to improve the value of the measured data for these applications, we consider whether the LC-MS data from the Proteograph™ workflow is appropriately analyzed and summarized from the acquired peptide identifications.

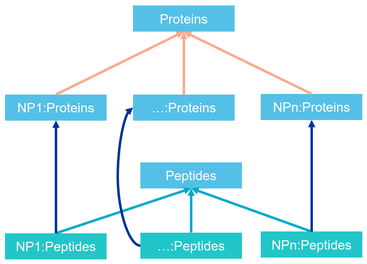
To address peptide to protein roll-up challenges using Data Independent Acquisition (DIA) methods, we have analyzed a case/control cohort of several hundred samples analyzed with DIA LC-MS using a 30-minute gradient on Bruker timsTOF Pro 2. We searched the DIA data using DIA-NN, which offers two different protein inference modes and three different quantification modes, including MaxLFQ. Combining the options provided by DIA-NN with Seer internally developed libraries and additional peptide to protein roll up strategies, we evaluated each strategy for summarizing the proteome by three metrics: 1) accuracy of the phenotype predictor, 2) number of significant marker associations with the phenotypes, and 3) number of protein quantitative trait loci (pQTLs) using the paired genotype data.

We found that the proteome measurements yield generally robust results across the applications irrespective of roll up strategies. Some of the rollup strategies yielded improvements in some cases over 10%. Interestingly, we found even unrolled up peptide level measurements gave robust results, sometimes outperforming the rolled-up values.

Study Design

Figure 1. Seer's High-resolution Measurements Provide Multiple Launch Points for Downstream Analysis.

The Proteograph™ Product Suite provides multiple quantitative measurements of the proteome at peptide level resolution. This data allows scientists to choose the level of detail they want to include in their downstream data analysis (e.g., incorporate the peptide quantities measured by each nanoparticle, or work with aggregated inferred protein quantities). This study aims to consider some aggregation methods and how the different levels of data impacts downstream analysis. For the dark blue line, the heuristically protein inference algorithm included in DIA-NN is used.



Methods

Label-Free DIA Analysis

Plasma of 186 case/control samples was processed using the Proteograph. The peptides from the Proteograph were analyzed using a 30-minute gradient DIA LC-MS protocol on a Bruker timsTOF Pro 2. The data was searched using DIA-NN[2] 1.8.1 in library free mode (**Library Free**), using the offline fractioned DDA library found in PAS 2.1 (**PAS 2.1 Library**) and a gas-phase fractioned DIA library built on a different partially overlapping cohort (**GPF Library**).

Each set of search results was kept unaggregated at the nanoparticle level (**NP**) or rolled up using one of 3 strategies (**Mean** – average the intensity measured by each nanoparticle, **Sum** – sum the intensity measured by each nanoparticle, **MaxRep** – keep the measurement from the nanoparticle that makes the most measurements for that peptide or protein group across the cohort).

Additionally, for each of the rollups, we considered up to four intensities (**intensity** – raw intensity computed by DIA-NN, **median_normalized_intensity** – intensity with median per injection removed and rescaled back to median intensity per nanoparticle, **diann_normalized_intensity** – DIA-NN variant of median normalized intensity and **maxlfq_normalized intensity** – DIA-NN's MaxLFQ intensity computed only for protein groups). In total this resulted in 84 configurations that were benchmarked using the classification pipeline shown in Figure 2.

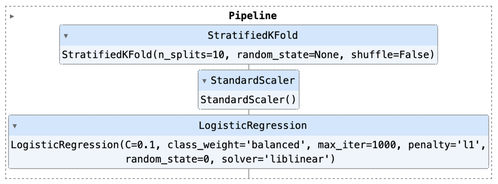


Figure 2. Machine Learning Case Classification Workflow.

To compare the performance of the different rollup strategies, a simple machine learning workflow was used to predict cases from controls. The classification workflow was kept simple to avoid more advanced models overcoming any signal disruptions due to aggregation. Features that were missing in more than 75% of samples were dropped and missing features were zero imputed.

Label-Free DIA Search Results

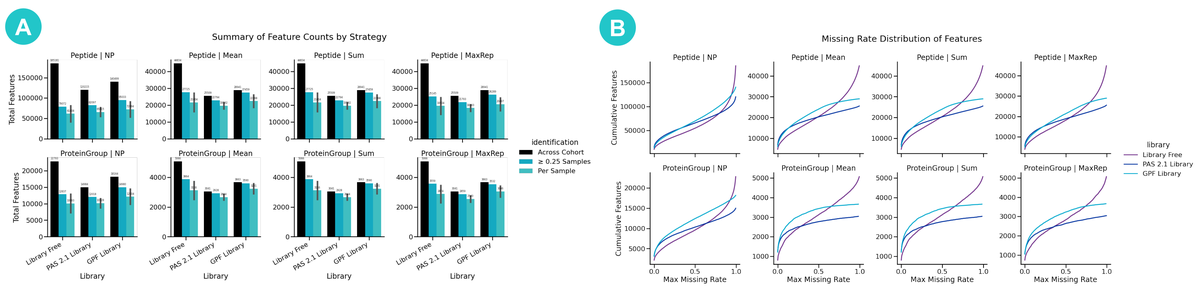


Figure 3. Identified Features and Data Sparsity Rate by Library and Rollup Strategy.

A) Feature counts show the **Library Free** search yields the most features across a large cohort but is challenged by data sparsity. The experimental libraries have better sparsity but sacrifice some extreme sensitivity. In general, we robustly identify over 20,000 peptides and approximately 3,000 protein groups per sample.

B) Depending on the study being constructed, the scientist may prefer very limited sparsity at the cost of sensitivity or may be willing to sacrifice data completeness for sensitivity. The accumulations curve show that in general the experimental libraries obtain lower data sparsity. The **GPF Library** performs especially well and can maintain better sensitivity than **Library Free** even for moderate levels of data sparsity.

Proteograph™ workflow with label-free DIA analysis provides high levels of reproducibility for large scale plasma proteomic studies

Classification Performance Across Rollup Strategies

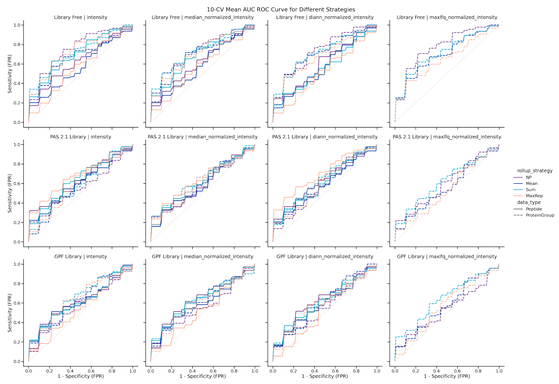


Figure 4. Classification Performance.

The mean receiver operator curves (ROC) of 10-fold cross validation show mean area under the curve (AUC) ranging from 0.52 to 0.78. The most visible phenomenon was the consistent strong performance of **Library Free**. It also is suggestive that protein group level rollup is strongly preferred over peptide level features.

The more advanced normalization of **diann_normalized_intensity** and **maxlfq_normalized_intensity** also seem to provide some gains in performance.

Table 1. Aggregated view of Mean and Standard Deviation of AUROC.

The tabular summary of the mean and standard deviation (green – best per row, red – worst per row) of the AUC reveal two interesting trends – 1) the **Sum** aggregation seems to provide robust performance (either the best or reasonably good with no worst case and 2) the **Library Free** mode with protein group features does very well. Curiously, both experimental libraries do poorly at the protein group level NP resolution, while the peptide level features give comparable performance to **Library Free**.

		rollup_strategy	NP		Mean		Sum		MaxRep	
library	data_type	intensity_type	mean_AUC	std_AUC	mean_AUC	std_AUC	mean_AUC	std_AUC	mean_AUC	std_AUC
Library Free	Peptide	intensity	0.63	0.16	0.57	0.13	0.70	0.09	0.57	0.13
Library Free	Peptide	median_normalized_intensity	0.62	0.16	0.59	0.13	0.69	0.09	0.58	0.14
Library Free	Peptide	diann_normalized_intensity	0.61	0.17	0.57	0.10	0.57	0.13	0.52	0.15
Library Free	**ProteinGroup**	intensity	0.76	0.13	0.68	0.10	0.72	0.12	0.72	0.13
Library Free	**ProteinGroup**	median_normalized_intensity	0.73	0.13	0.70	0.11	0.73	0.12	0.71	0.11
Library Free	**ProteinGroup**	diann_normalized_intensity	0.73	0.15	0.72	0.12	0.72	0.12	0.69	0.14
Library Free	**ProteinGroup**	maxlfq_normalized_intensity	0.78	0.12	0.70	0.13	0.72	0.11	0.65	0.13
PAS 2.1 Library	Peptide	intensity	0.63	0.12	0.61	0.13	0.64	0.10	0.67	0.14
PAS 2.1 Library	Peptide	median_normalized_intensity	0.59	0.14	0.59	0.14	0.62	0.10	0.64	0.14
PAS 2.1 Library	Peptide	diann_normalized_intensity	0.63	0.12	0.58	0.10	0.61	0.11	0.70	0.13
PAS 2.1 Library	**ProteinGroup**	intensity	0.55	0.13	0.57	0.12	0.60	0.15	0.60	0.12
PAS 2.1 Library	**ProteinGroup**	median_normalized_intensity	0.59	0.14	0.60	0.12	0.64	0.13	0.60	0.12
PAS 2.1 Library	**ProteinGroup**	diann_normalized_intensity	0.60	0.12	0.60	0.13	0.64	0.11	0.60	0.11
PAS 2.1 Library	**ProteinGroup**	maxlfq_normalized_intensity	0.60	0.14	0.55	0.17	0.62	0.17	0.53	0.11
GPF Library	Peptide	intensity	0.61	0.14	0.61	0.13	0.61	0.12	0.54	0.13
GPF Library	Peptide	median_normalized_intensity	0.63	0.12	0.60	0.12	0.63	0.14	0.53	0.12
GPF Library	Peptide	diann_normalized_intensity	0.63	0.12	0.59	0.09	0.61	0.10	0.53	0.11
GPF Library	**ProteinGroup**	intensity	0.57	0.11	0.62	0.14	0.66	0.16	0.64	0.09
GPF Library	**ProteinGroup**	median_normalized_intensity	0.56	0.14	0.60	0.14	0.63	0.16	0.60	0.11
GPF Library	**ProteinGroup**	diann_normalized_intensity	0.56	0.14	0.61	0.13	0.62	0.15	0.59	0.13
GPF Library	**ProteinGroup**	maxlfq_normalized_intensity	0.54	0.15	0.57	0.14	0.64	0.16	0.57	0.15

pQTLs Identified Across Rollup Strategies

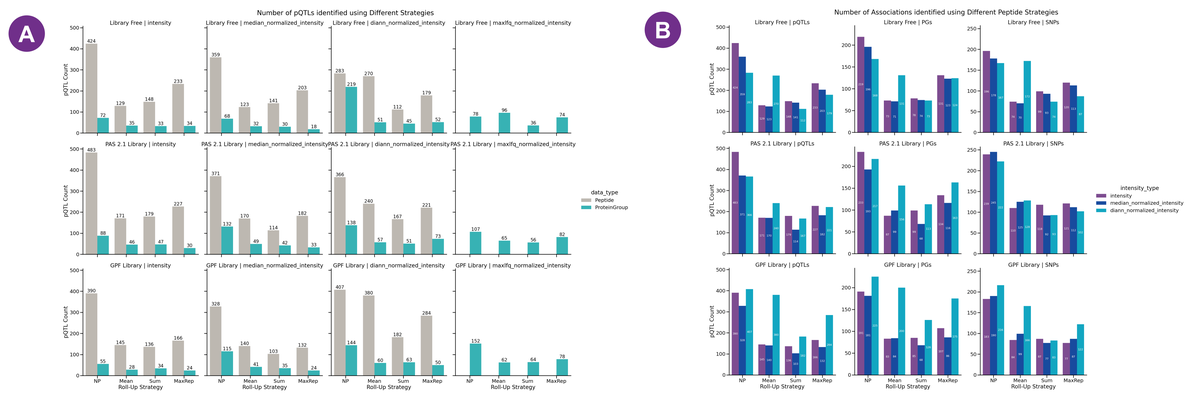


Figure 5. pQTL identified and breakdown of individual SNPs and Protein Groups.

A) Protein Quantitative Trail Loci (pQTL) identification rate across different rollup strategies reveals **Peptide**-level analysis increases sensitivity due to enhanced resolution. **B)** Further inquiry of only peptide level pQTLs indicates the **NP** level rollup identifying the most **pQTLs**, protein groups (**PGs**) and **SNPs** followed by **MaxRep**. In some cases, **diann_normalized_intensity** seems to be offering gains that need to be further investigated.

Future Work

- Interrogate the difference in performance between classification (generally better with **ProteinGroup** and **Sum** aggregation) and pQTLs (generally better with **Peptide** and **NP** aggregation).
- Incorporate additional metrics of success including regression of clinical measurements.
- Investigate the generalizability of the findings across multiple cohorts to enable recommendation of an optimal workflow.

Conclusion

- Incorporating measurements at **NP** resolution seems to work well in many cases, while the **Sum** across nanoparticles aggregation strategy seems to be a robust preserver of classification power.
- Library Free mode appears to work better with **NP** resolution for classification, while the experimental libraries seem to work better with **NP** aggregation for pQTL identification.
- The more advanced normalization strategies such as DIA-NN's RT dependent median normalization and MaxLFQ normalization seem to improve performance opening the door for NP aware variants of these methods to further increase the performance.

References

[1] Blume *et al. Nat. Comm.* (2020)
[2] Demichev *et al. Nat. Methods.* (2020)

Acknowledgements

This work was supported by Research Grant No. 5R44AG065051-02 from the National Institute of Health.

Publications

MASSACHUSETTS GENERAL HOSPITAL

Seer, Inc., Redwood City, CA 94065, USA | *hguturu@seer.bio

SBIR · STTR America's Seed Fund™ POWERED BY SBA

Research funded by SBIR grant

seer