Building Spectral Libraries for Large-Scale Quantitative Proteomic Studies in Human Plasma

Jian Wang[1], Harendra Guturu[1], Yingxiang Huang[1], Seth Just[1], Shadi Ferdosi[1], Xiaoyan Zhao[1], Andrew Nichols[1], **Lee. S Cantrell***[1], Alexey Stukalov[1], Iman Mohtashemi[1], Ting Huang[1], Lucy Williamson[1], Gabriel Castro[1], Eltaher Elgierari[1], Ryan W. Benz[1], Khatereh Motamedchaboki[1], Daniel Hornburg[1], Asim Siddiqui[1], Anna Halama[2], Frank Schmidt[2], Karten Suhre[2] and Serafim Batzoglou[1]

1. Seer, Inc., Redwood City, CA, 94065 USA; 2. Weill Cornell Medicine-Qatar, Doha, State of Qatar

Deep and unbiased plasma proteomics for disease cohort studies at scale

Access to quantitative information in the plasma proteome is important to study and monitor human health. However, due to the large dynamic range in the plasma proteome most previous studies were limited either in the depth of coverage or scale.

The Proteograph™ workflow allows the detection of thousands of proteins per sample across thousands of individuals.[1] To facilitate the extraction of quantitative information from these large-scale studies with high throughput Liquid Chromatography coupled to Data Independent Acquisition (DIA) Mass Spectrometry (LC-MS), we generated comprehensive spectral libraries from diverse plasma samples and demonstrate the utility of these peptide spectral libraries in extracting biological insights in large cohort plasma studies.

Methods

For multiple cohorts, project specific and generalized spectral libraries were generated with several conventional approaches

1. Pooled and fractionated or individual injection in either DIA or DDA LC-MS
2. Gas phase fractionation of pooled peptides, evaluated with DIA LC-MS
3. Predicted library generation from a FASTA sequence database

Spectral libraries were generated using conventional spectral library assembly tools. Search was performed in DIA-NN 1.8.1.[2]

For controlled experiments and cohort level experiments, figures of merit were assessed including:

- Identification rate and data completeness
- Precision
- Quantitative accuracy

For a cohort with known subject disease diagnosis, differential expression of disease markers was compared

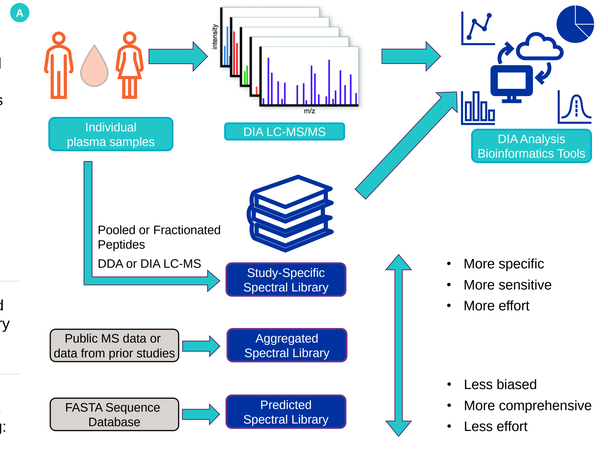

Figure 1. Workflow for spectral library generation. Plasma samples are evaluated by DIA LC-MS and can be searched against libraries generated by multiple strategies. These are defined as study specific, aggregate, and predicted or library-free peptide spectral libraries. Workflow for subset-library generation in library-free framework.

Evaluation of plasma proteomic with the Proteograph workflow using spectral library strategies

Results

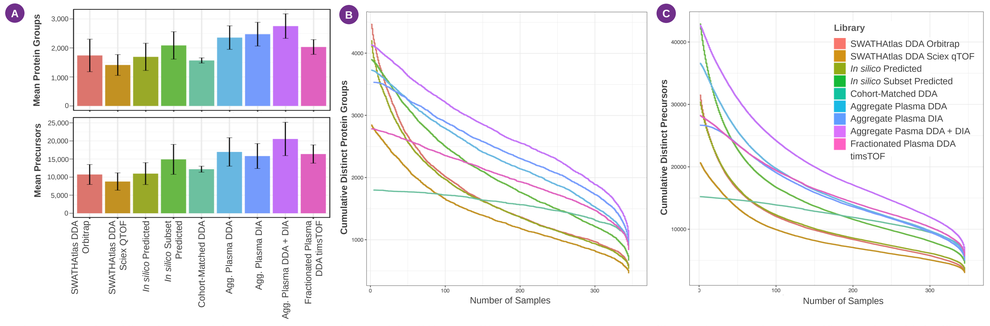

Figure 2. Identifications made on a large cohort with different library approaches. For library-free, project-specific, and multiple library aggregation strategies, precursor and protein identifications were assessed across a 345 human plasma cohort study using Seer's Proteograph workflow. **(A)** Mean distinct protein groups and precursors was assessed per sample for each library method, **(B)** Cumulative distinct protein group completeness was assessed for each library method, **(C)** Cumulative distinct precursors was assessed for each library method.

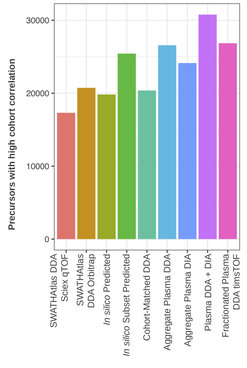

Figure 3. Comparison of library correlation above background. Following LC-MS analysis of human plasma samples with Proteograph workflow and normalization, quality of quantitation was assessed by correlation of precursors within a protein across a plasma sample cohort study. Relative to a null distribution, each library generated a distinct number of precursors with high correlation. Aggregate libraries demonstrate superior performance to shallow cohort libraries and library-free methods in this metric.

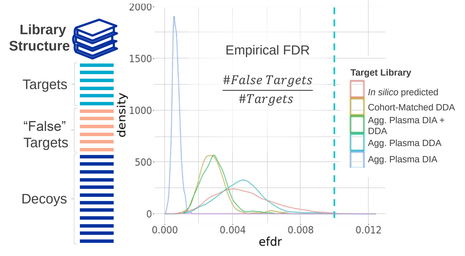

Figure 4. Empirical FDR demonstrates larger libraries superior performance to smaller libraries. Empirical FDR was assessed for aggregate, library-free, and cohort-specific DDA libraries. While each library approximately matched an assessed 1% FDR, q-value distribution has a lower mode in cohort-specific DDA libraries than aggregate, and a lowest mode in large aggregate libraries relative to DDA libraries and library-free methods.

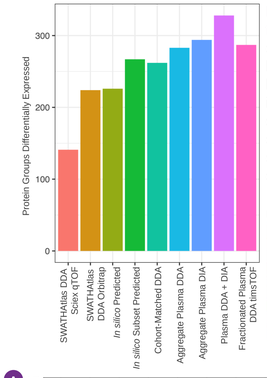

Figure 5. Differential expression measured as a function of search library. For each library, differential protein expression was assessed with limma. Groups were selected based on prior knowledge of disease diagnosis. Results here indicate that aggregate libraries may enhance the quantity of proteins measured as differentially expressed. Such analyses may be useful in discovery of single proteins or protein network biomarkers correlated with disease onset.

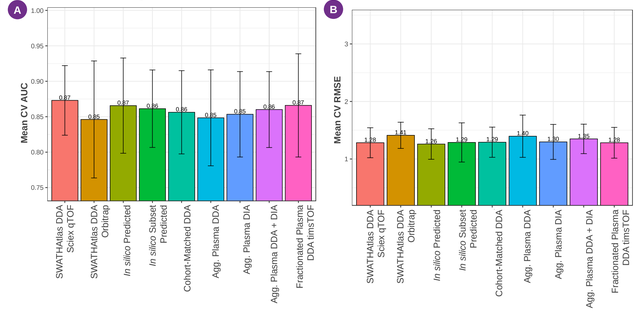

Figure 6. Supervised regression and classification analyses with a large cohort. The studied cohort had known subject classification of disease state and progression. Each analysis was done on the precursor level. **(A)** Mean CV Area Under Curve (AUC) – higher is better – was assessed for different library methods to classify subject disease state and **(B)** Mean CV Root Mean Squared Error (RMSE) – lower is better – was evaluated to assess the quality of regression related to a known disease biomarker.

Conclusion

- DIA-NN combined with a cloud-based platform allows efficient measurement of 5,677 proteins and 57,388 peptides between libraries.
- Lower abundance proteins may be measured with project-specific or aggregated libraries, while library-free search results in greater overall identifications at cost of data completeness
- Empirical FDR is sufficiently estimated in each evaluated library method. Deeper libraries result in greater differentially expressed protein groups. Regression and classification perform comparatively between library methods.

References

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