High-throughput plasma proteomics to identify diabetes associated protein biomarkers and pQTLs

Harendra Guturu[1]*, Guhan Venkataraman[1], Amir Alavi[1], Ryan Benz[1], Khatereh Motamedchaboki[1], Anna Halama[2], Frank Schmidt[2], Karsten Suhre[2], and Serafim Batzoglou[1]

[1]Seer, Inc., Redwood City, CA 94065, USA, [2]Weill Cornell Medicine-Qatar, Education City, 24144 Doha, Qatar

Deep and unbiased plasma proteomics for disease cohort studies at scale

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 assessment of the proteome enables previously inaccessible applications such as phenotype diagnosis, biomarker discovery, and proteogenomics analysis from easy to obtain sample types, like plasma. We used Seer Proteograph™ workflow to build a classifier to distinguish diabetics from controls and regressors to predict clinical markers. Additionally, we integrated the proteomics measurements with genotype data to identify protein quantitative trait loci (pQTLs) that are robust to potential epitope effects due to coding variations in the proteome.

Novel aspect

- Integrated unbiased and large-scale proteomic data generated from Proteograph™ Assay to detect biomarkers associated with diabetes.
- Developed method for using protein alternating variation in mass spectrometry data (MS-PAV) when performing pQTL analysis.

Methods

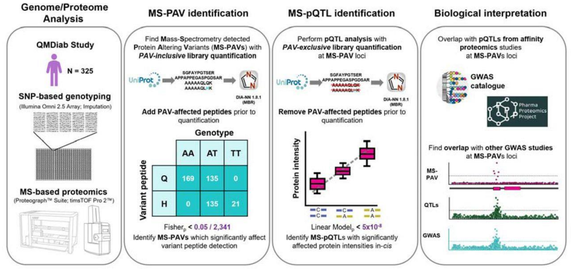
Cohort-specific library (QDIAB) generated by Liquid Chromatography coupled to Data Dependent Acquisition; DDA Mass Spectrometry DIA LC-MS) was built by running 90-minute gradient DDA of 7 pools of 16 samples each searched using MSFragger[2].

Cohort was analyzed with DIA-NN[3] v1.8.1 in separate group-runs in library free mode, library free mode with match between runs (MBR) enabled, and with a cohort-specific DDA .

For all classification and regression analysis the data `log(1 + intensities)` scaled and zero imputed. The regularization parameters were tuned using 10-fold CV and the model performance was estimated using 10-fold CV bootstraps.

MS-PAV pQTL analysis was generated using custom logic to identify variant peptides and the pQTL search was done using PLINK. Refer to Suhre *et al.*, bioRxiv 2023.04.20.537640 for more detailed methods[4].

Identification of MS-PAV pQTLs



We developed a novel bottom-up proteomics approach that accounts for protein altering variants in the detection of pQTLs. Using this approach, we identify novel protein altering variants in proteins of clinical relevance that may not be accessible to affinity proteomics. For more details on our MS-PAVs work to identify pQTLs free of biases in measurement due to epitope effects please refer at our pre-print.

Suhre *et al.* bioRxiv 2023.

Deep proteomics enhances disease classification and biomarker discovery

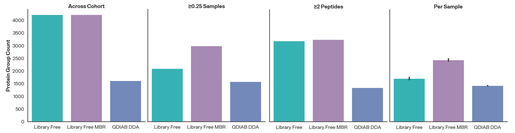


Figure 1. Library Performance Across Cohort.

Library free and library free with MBR identify more protein groups with better sequence coverage than the QDIAB DDA cohort-specific library.

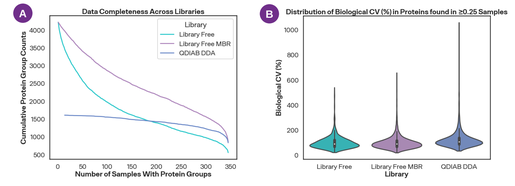


Figure 2. Data Completeness of Identified Protein Group.

A) The QDIAB DDA cohort-specific library shows lower proteome depth, but more complete (i.e., protein groups measured in more samples) data compared to library free search and **B)** potentially captures more biological variation. Library free search with MBR improves the data completeness for library free search.

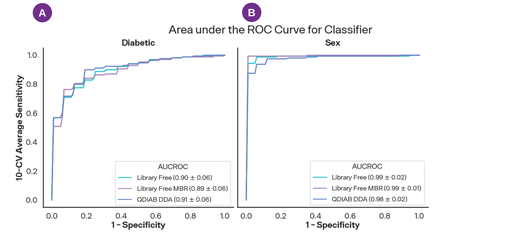
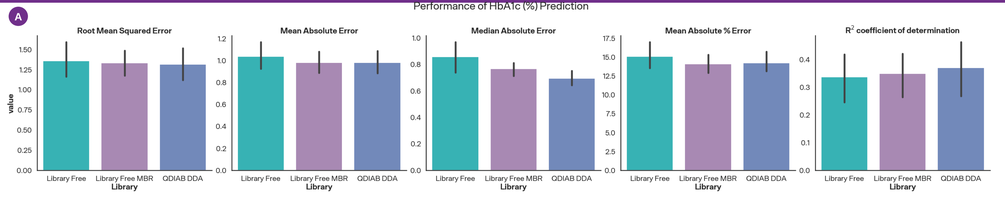
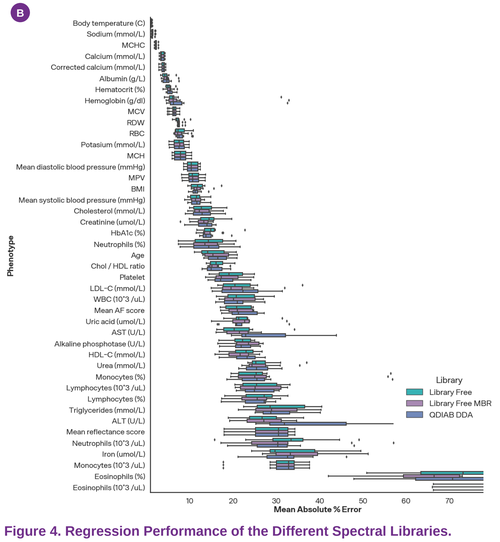


Figure 3. Classification Performance of the Different Spectral Libraries.

A) Classification performance across the three search modes for Diabetic status and **B)** Sex. Minimal to no difference is seen in classification accuracy across the three different library searches.





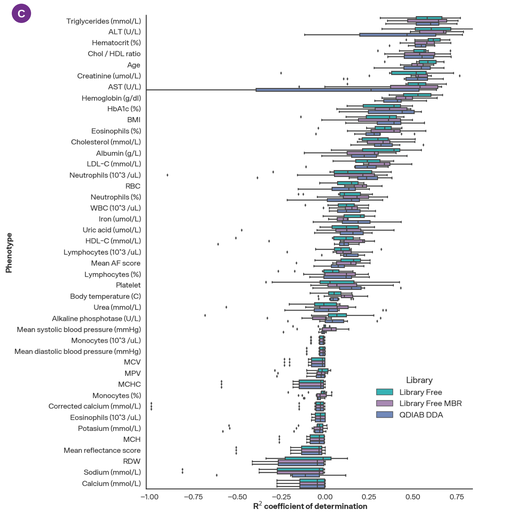


Figure 4. Regression Performance of the Different Spectral Libraries.

A) Five different metrics of model performance for predicting HbA1c values (a good clinical indicator of Diabetic status). **B)** Mean absolute percentage error and **C)** The R^2 value of 43 different regression models to predict clinical phenotypes. The two metrics indicate the need for multiple metrics since although some values such as body temperature are predicted with low percentage error that is primarily due to low variability of that measurement, while R^2 providers a better indicator of predictor fit and identifies measurements such as triglyceride levels are those that can be strongly predicted. Across the various regressors we see the cohort-specific library slightly outperforming the Library Free options.

Conclusion

- Library free with MBR increases data completeness for library free search and cohort-specific libraries potentially capture more biological variation.
- The different search modes don't seem to improve classification, possible due to these classification tasks being "too easy".
- The library free search with MBR appears to marginally outperform library free search without MBR suggesting the additions from MBR may outweigh the laxer FDR control. Additionally, the shallower cohort-specific library seems to outperform both suggesting cohort-specific libraries may better capture the relevant signals.

References

[1] Blume *et al. Nat. Comm.* (2020)
[2] Yu *et al. Molecular & Cellular Proteomics.* (2020)
[3] Demichev et al. *Nat Comm.* (2020)
[4] Suhre et al., bioRxiv 2023.04.20.537640

Publications