Comparative Proteomics: SomaLogic SomaScan vs. Seer Proteograph MS

Author: Sophia M. Soriano Date Submitted: 13 November 2023

Abstract

Clinical biomarker panels are key to the early detection of many cancer types. While many biomarker panels rely on genomic and transcriptomic markers, recent efforts in developing more accurate and effective panels have expanded to include proteomics. Two new methods of proteomic biomarker discovery – Seer Proteograph MS and SomaLogic SomaScan – are currently being assessed to determine if the aptamer-based SomaScan platform can be used as an orthogonal validation or complementary method to the mass spectrometry (MS)-based Seer Proteograph platform. In this study, two sample sets – one murine, one human – were analyzed using these two novel technologies, with the SomaLogic SomaScan results assessed for 1) overall dataset quality, 2) the validity of using low-volume or diluted samples, and 3) comparability to results produced by the Seer Proteograph MS platform. Concerning technical assay quality, low coefficients of variation (CVs) and high total protein IDs resulted across all samples, and low-volume samples were determined to have equivalent results to typical volume samples. The diluted samples, while comparable in CV and total protein IDs, resulted in decreased intensity measurements across most aptamers. Critically, very low correlation was observed in protein differential expression (DE) results between SomaLogic and Seer data for both human and mouse datasets.

Introduction

A major effort in current clinical cancer research focuses on the development of new disease detection methods capable of identifying cancer at the earliest possible stage. Some of the most prominent of these detection methods leverage clinical screening panels composed of molecules experimentally determined to serve as biological markers for a cancer type – either alone or in combination with other molecules in the panel (*1*). Traditionally composed of genomic or transcriptomic markers, recent biomarker panel development research has expanded to the discovery and validation of proteomic biomarker panels. As proteins are directly responsible for much of the molecular activity in a cell, clinical cancer researchers have advocated that earliest detection of cancerous cell activity will be most powerful with the inclusion of protein-based panels (*2,3*). Additionally, the clinical field has emphasized the need for tests that can be used to screen blood plasma (or serum) samples, given blood collection is a minimally invasive clinical standard (*4*). The discovery of novel protein biomarkers in plasma is performed using a variety of proteomic techniques ranging from non-targeted MS methods to targeted assays such as enzyme-linked immunosorbent assays (ELISAs) (*2*). MS-based methods can accommodate broad-range, non-targeted protein detection; however, deep sampling depth in plasma has been difficult at scale (*1*). Targeted technologies like antibody-based ELISAs typically have high sensitivity and technical accuracy – but they rely on prior knowledge of the protein of interest and possible biomarkers for antibody selection.

Recently, two new methods of proteomic biomarker discovery were developed to better facilitate blood-based proteomic discovery. The first novel method, originally published and now marketed by Seer Technology, is a specialized MS-based method employing their Proteograph technology. The automated Proteograph system leverages proprietary nanoparticle (NP) technology for non-specific protein capture based on the biophysical properties of plasma proteins (*5*). This replaces the time-intensive sample preparation steps previously necessary to widen the depth of coverage for MS plasma proteomics, and now feasibly enables automated proteomic detection for large sample cohorts. Because no specific targeting molecules are used, the usual advantage of MS-based methods – broad-range, non-targeted protein detection – is still largely applicable to this technique. Oregon

Health & Science University Cancer Early Detection Advanced Research Center (OHSU CEDAR) currently uses Seer Proteograph MS to enable proteomic studies in plasma and serum for ongoing cancer biomarker research and has recently been evaluating additional proteomic screening methods, particularly those capable of analysis with low sample volumes (precluding Seer MS) or those with sufficient panel depth for validation of Seer MS proteomic discoveries.

One potential candidate as an orthogonal complementary method is the SomaScan platform, created by SomaLogic. SomaScan is a targeted (non-MS) method that uses specially designed aptamer molecules for protein capture, in a process functionally similar to antibody behavior, but much less expensive per protein given the depth of the current SomaScan panel (*6,7*). This panel consists of 7596 aptamers, each targeting a specific protein or protein subunit, with anticipated expansion to 10000. These aptamers fluorescently label their target protein in a series of reactions during sample preparation, and protein quantification is performed through fluorescence intensity measurement on a DNA microarray (*1*).

While both novel methods mentioned above have been demonstrated to generate datasets with high technical quality, a current focus of clinical proteomics research is benchmarking these new platforms and determining how well results correspond. As such, OHSU CEDAR conducted a study to compare proteomic results from SomaLogic SomaScan to those from Seer Proteograph MS, with the end goal of assessing whether SomaScan is a suitable orthogonal validation method, or at minimum a complementary method, to Proteograph MS.

Methods

Two datasets were used in this study to compare proteomic analysis results of the two platforms of interest – Seer Proteograph MS and SomaLogic SomaScan (*8,9*). The first dataset consisted of 70 samples from a larger cohort collected for a murine pancreatic ductal adenocarcinoma (PDAC) proteomic study. Eight murine groups from the original study were represented in this subset: 10 Healthy Control – Early, 14 Healthy Control – Late, 6 KMC – PDAC (MYC mutation-driven PDAC), 12 KMC Control (non-cancer KM mice), 6 KPC – Lung (p53 mutation-driven lung cancer), 11 Non-Lethal Pancreatic Intraepithelial Neoplasia (PanIN) (non-lethal lesions), 8 Lethal PanIN – Early (lethal lesion precursors to PDAC), and 3 Lethal PanIN – Late (lethal lesion precursors to PDAC). The "Early" and "Late" designations refer to early and late stages of mice in the study. Two additional murine samples noted with a pink/red color, possibly indicating hemolysis, were also included to test SomaScan's ability to screen for hemolytic markers. Lastly, an additional nine samples of plasma pooled from several of the murine groups were included to evaluate SomaScan's performance with low volume or diluted samples: 3x55uL typical volume, 3x35uL low volume, 3x35uL plasma diluted with PBS buffer to 55uL total volume. All 70 murine group samples were run using Seer Proteograph combined with LC-MS on a Bruker timsTOF Pro MS to generate raw MS data, which was subsequently searched through DIA-NN in Seer's Proteograph Analysis Suite (PAS) (*10*). These samples, along with the other 11 test samples, were also submitted to SomaLogic, from which the resulting data was retrieved through SomaLogic's online portal. SomaLogic added four buffer samples and three calibration samples for quality control of their process for a total of 88 samples.

The second dataset consisted of 30 patient samples, subset from a human prostate cancer proteomic study evenly split between case (cancer) and control (non-cancer) groups. The same pooling strategy described above to test low volume and diluted samples was also used here to generate nine additional SomaLogic test samples. Raw data for the 30 case and control samples was generated with Seer Proteograph and LC-MS as described above, with these results then searched using MaxQuant (*11*). These samples, plus the nine test samples, were also sent to SomaLogic, and with their addition of five buffer samples, five calibration samples, and three QC samples (used to normalize human samples to their internal population standard), we received results on 52 total samples.

All raw SomaLogic data was processed in the following steps. First, application of an estimated limit of detection (eLoD)-based noise filter based on buffer intensities for each of the 7596 aptamers in the panel, as recommended by SomaLogic (eLoD/aptamer = Mean_buffer + 4.9MAD_buffer). Second, quality control (QC) analysis methods were applied to the filtered data, assessing the performance of aptamers across samples. Third, differential expression (DE) analyses were conducted comparing protein intensities between conditions of interest. Finally, the DE results were compared to their counterparts in the Proteograph MS data. SomaLogic provided two versions of each dataset – one with all internal normalization steps completed, and another without the final normalization step. This normalization step serves as a population variation control – the human data is normalized to SomaLogic's collected "normal" human population group, while the mouse data undergoes a median normalization. Each QC analysis performed compared the "Normalized" and "Pre-Normalization" data results. However, on the recommendation of SomaLogic, the "Normalized" data was used for all biological DE analyses. The "dilution sets" of nine samples each were the only groups where the "Pre-Normalization" data was recommended for DE analysis. Custom R scripts were used for data quality assessment and proteomic analysis (https://github.com/ssoriano22/SomaLogic_Benchmark). As Seer data quality has been previously confirmed by the OHSU CEDAR proteomics group, this study focused on evaluating SomaLogic data. SomaLogic's online portal for analysis – SomaLogic DataDelve – was used to confirm SomaLogic DE analyses performed with the custom analysis pipeline (*12*).

Results & Discussion

Quality Control (QC) Analysis:

The first step in assessing the technical quality of the SomaLogic data was to determine how many aptamers per sample ID fell below their eLoD cutoff. In the "Normalized" mouse dataset of 81 samples, all but three had at least 7500 aptamers with intensity signal above noise out of the maximum 7596 aptamers, indicating that the vast majority of aptamers detected signal across all samples (Figure 1A). Two of the lowest samples were the pink/red annotated suspected hemolysis samples, although still very high at 7387 and 7472 aptamers with signal above noise. No major difference in aptamers with signal above noise was observed between any of the nine "dilution set" samples and the other biological samples. The "Pre-Normalization" dataset version had slightly more samples filtered out than the "Normalized" version, with 10 samples under 7500 aptamers, the lowest sample having 7351 aptamers. Overall, there was very little difference between the two versions of the mouse dataset, with the number of aptamers above noise per sample generally very close to the maximum. This is well above SomaLogic's stated expectation of high-quality aptamer performance for 84% of their current panel (6379 aptamers) when applied to murine sample analysis. The human dataset of 39 samples showed even better performance of the aptamer panel across samples, with the "Normalized" dataset samples all having at least 7500 aptamers above noise, and the "Pre-Normalization" dataset only having one sample below at 7482 (Figure 1B).

Next, the distribution of aptamer coefficients of variation (CV) per condition group was surveyed to quantify the technical variability of the SomaLogic assay and the biological variability between samples. Both murine and human aptamer intensity CVs were calculated using log2-transformed intensity data, then grouped by condition in boxplots for each species dataset (Figure 1C-D) (*13*). The "Pre-Normalization" data for both datasets showed higher mean CVs across all groups compared to the "Normalized" data, confirming the normalization was applied as intended. However, no difference in overall trends between condition groups was observed, so the "Normalized" data version was the focus of analysis. The technical variability of the assay was very low in both datasets, as demonstrated by the pooled 55µL mean CVs of 3.6 (murine) and 4.1 (human), implying that the technical reproducibility of SomaLogic's assay is promising. Also of note is that the murine low volume 35µL mean CV (8.8) is higher than the mean CVs of the murine typical 55µL (3.6) and diluted 55µL (2.1). Given that the human "dilution set" mean CVs were all similar (3.0, 4.1, 3.5), it is possible that using SomaLogic's assay on lower volumes of murine samples results in higher technical variability. Mean aptamer CVs were higher across all conditions in the murine dataset compared to those in the human dataset, despite their controlled genetic and experimental background. A possible explanation is that murine plasma sample collection (cardiac puncture) is more traumatic and technically challenging than human blood collection, potentially introducing molecular differences between samples collected from different mice. Overall, the amount of biological variability observed in both the murine and human SomaLogic "Normalized" datasets is relatively similar to that observed in Seer Proteograph MS.

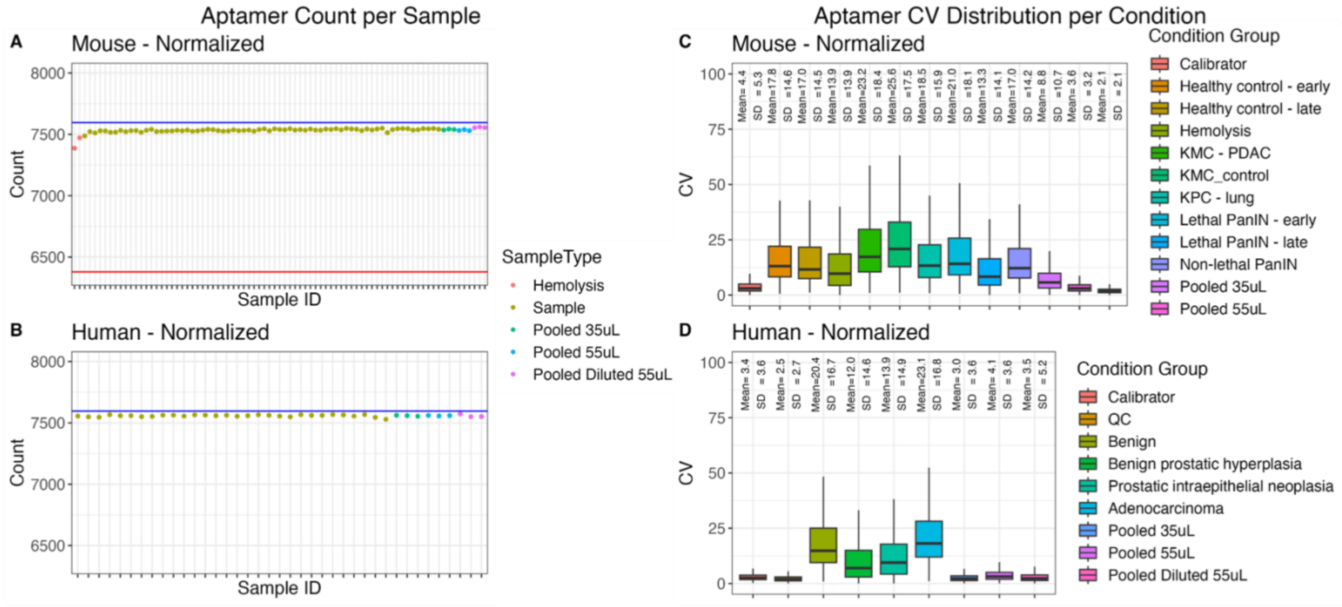

Figure 1: *SomaLogic data QC results evaluating aptamer panel performance across samples in both the murine and human datasets. A-B) Number of aptamers per sample (81 murine, 39 human "Normalized") after eLoD noise filter. Blue lines indicate maximum 7596 aptamers, red line marks 84% of aptamers in the panel with expected good performance in murine studies. C-D) Aptamer CV distributions per condition, using "Normalized" data. Outliers not shown.*

The intensity distribution per sample was evaluated to check for batch effects and sample outliers, and also to better visualize the effect of SomaLogic's final normalization step. Normalization of both datasets reduces overall variation in both the murine and human intensity distributions. Also, the murine dataset appeared to have more overall variation in intensity distributions across samples. This could be due to the traumatic murine sample collection described earlier, but it is also conceivable that the human-protein-targeted design of SomaLogic's aptamer panel could result in less specific protein-aptamer binding in murine samples. Diluted samples in both species "Pre-Normalization" datasets have lower intensity distributions than the typical volume and low volume samples, but normalization did adjust the diluted intensity distributions in both species datasets to be comparable to all other samples in the "dilution set". Additionally, it was noted that the two suspected hemolysis murine samples maintained higher intensity distributions than the other murine samples in both the "Pre-Normalization" and "Normalized" data, suggesting these samples contained more high-concentration proteins. This could be a hemolysis-related effect, as higher quantities of high-concentration blood proteins like hemoglobin and haptoglobin are expected in hemolysis-contaminated samples.

In the final QC analysis result, the mean log2-transformed intensities of nine hemoglobin-related proteins in the murine dataset were compared between the two pink/red suspected hemolysis samples and the three pooled typical volume samples, to determine whether screening for hemolysis contamination was possible with the SomaLogic panel. For at least eight of the nine hemoglobin marker proteins in the "Pre-Normalization" dataset, the pink/red samples have a higher mean log2 intensity than the 55uL samples, suggesting that SomaLogic's assay may be able to screen for hemolyzed samples. Normalization removed all of these observed discrepancies in protein intensities between the suspected hemolysis and pooled typical volume samples.

Differential Abundance/Expression (DE) Analysis:

The first differential analysis performed was comparing log2 protein intensities between the three conditions in the "dilution set" for each species – pooled 35µL v. 55µL (low volume validity), 35µL v. 55µL Diluted (low volume vs. dilution), and 55µL v. 55µL Diluted (dilution validity). The log2 fold change (FC) – difference between intensity values – was calculated per protein between conditions, and Wilcox tests were performed with Benjamini–Hochberg (BH) False Discovery Rate (FDR) correction (*14*). The resulting FCs and p-values were plotted as volcano plots – one for each dilution condition comparison per species dataset (Figure 2A). In the mouse dataset, 7422 proteins had sufficient data for comparison (present in at least 2/3 condition replicates), while the human dataset had 7491 proteins. No proteins with significant differential abundance were found in any of the mouse or human dilution condition comparisons at an FDR < 0.05, which is somewhat odd considering the planned 1.6x FC dilution. Lowering sample volume to 35µL or reducing concentration did not have a significant effect on protein quantification; however, the diluted replicates had 50-100 fewer proteins than the low volume and typical volume replicates in both mouse and human datasets, reaffirming that dilution did result in fewer aptamers with signal above noise. To further investigate quantitation, boxplots were used to confirm that proteins in the murine and human diluted samples showed the expected FC of 1.6x dilution (Figure 2B). The median protein FC for the low volume-typical volume comparison was close to log2 FC 0 (1:1 protein intensities), suggesting no abundance difference when using lower volumes in the assay, and the typical volume-diluted volume median protein FC was close to the 1.6x target.

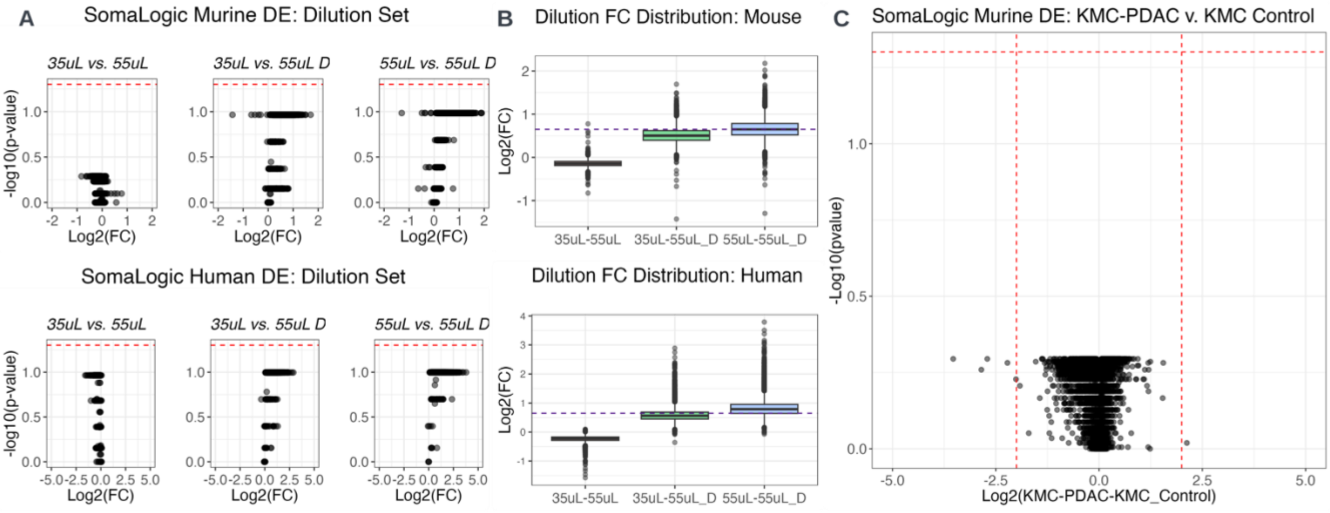

Figure 2: *Differential abundance analysis for pooled plasma "dilution sets" in both mouse and human datasets. A) Volcano plots evaluating proteins for significant FC (red line, $\alpha = 0.05$) between*

conditions. B) Boxplots confirming diluted protein median FC are on target (purple line = log2(1.6)), and low volume protein FCs are within an equivalent range (approx. 1:1) to typical volume protein FCs. C) SomaScan DE result for murine KMC-PDAC v. KMC Control (FC significance determined by Wilcox test $\alpha = 0.05$, horizontal red line. Vertical red lines mark FC=2, up or down regulated.

The next DE analysis focused on one of the murine condition comparisons of biological interest that showed differential expression in the Proteograph murine study – KMC-PDAC vs. KMC Control. Differential expression and statistical significance were evaluated as described previously. No proteins with significant differential expression were found in the SomaLogic data (Figure 2C), in contrast to the previous Proteograph MS analysis that identified 322 significant DE proteins. Both Seer and SomaLogic analysis used the same number of samples in each condition group (6 v. 12). As the Seer MS results available were generated using Welch's t-tests, the SomaLogic data was re-tested for significance using that test method, but there was no difference when compared to the original Wilcox test results. To ensure that the custom analysis pipeline was not the issue, the raw SomaLogic data (before eLoD filter) was entered into SomaLogic's DataDelve online portal for analysis. The resulting Wilcox test volcano plot from DataDelve confirmed the results of the custom analysis pipeline.

To explain this obvious discrepancy between Seer Proteograph MS and SomaLogic SomaScan DE results, the overlap of protein IDs detected across both platforms for these murine and human DE comparisons was assessed. UniProt IDs were used to define unique protein IDs in these method overlaps, to avoid losing any protein IDs that map to the same gene name. As SomaLogic only labels proteins by human IDs, the number of overlapping protein IDs in the murine comparison was defined by the intersection of protein IDs resulting from the nomenclature pivot from human to mouse and those resulting from the reverse mouse to human pivot (unique murine UniProt ID-human UniProt ID pairs found using custom function querying NCBI Entrez and KEGG databases). These paired mouse-human orthologue conversions for each dataset were used rather than converting to the same shared species in order to account for multiple protein IDs corresponding to the same gene during the orthologue pivot. Out of 2571 murine protein IDs identified with Seer MS and 6356 with SomaLogic in this KMC-PDAC v. KMC Control analysis, the overlap between platforms comprised 710 protein IDs (28% of Seer's total). Of these, 88 proteins were significantly differentially expressed in Seer MS results. In the human case v. control DE analysis – where no significant proteins were identified by either platform and both methods identified human IDs – from 409 protein IDs in Seer MS and 6377 in SomaLogic, 282 protein IDs overlapped between platforms (69% of Seer's total). From these numbers, the two platforms clearly have higher protein overlap in the human dataset compared to the murine. However, the murine protein overlap between platforms relies on a nomenclature pivot between species and is an estimate until SomaLogic allows for better specification of murine proteins targeted by their aptamer panel. Comparing FCs of these overlapping proteins, SomaLogic FCs for murine proteins were generally smaller magnitude compared to Seer FCs, while this difference is much less apparent in the comparison of human protein FCs (Figure 3). Finally, method correlation in FC was calculated (Pearson's correlation coefficient r) from the direct comparison of SomaLogic and Seer FCs for each overlapping protein (Figure 3). This correlation was found to be very low for both the murine overlap proteins (r = 0.15) and the human overlap proteins (r = 0.23). Given that the murine study used models with engineered genetic differences well studied in pancreatic cancer, it is surprising that SomaScan revealed no statistically significant, differential protein expression between cancer and control for any group.

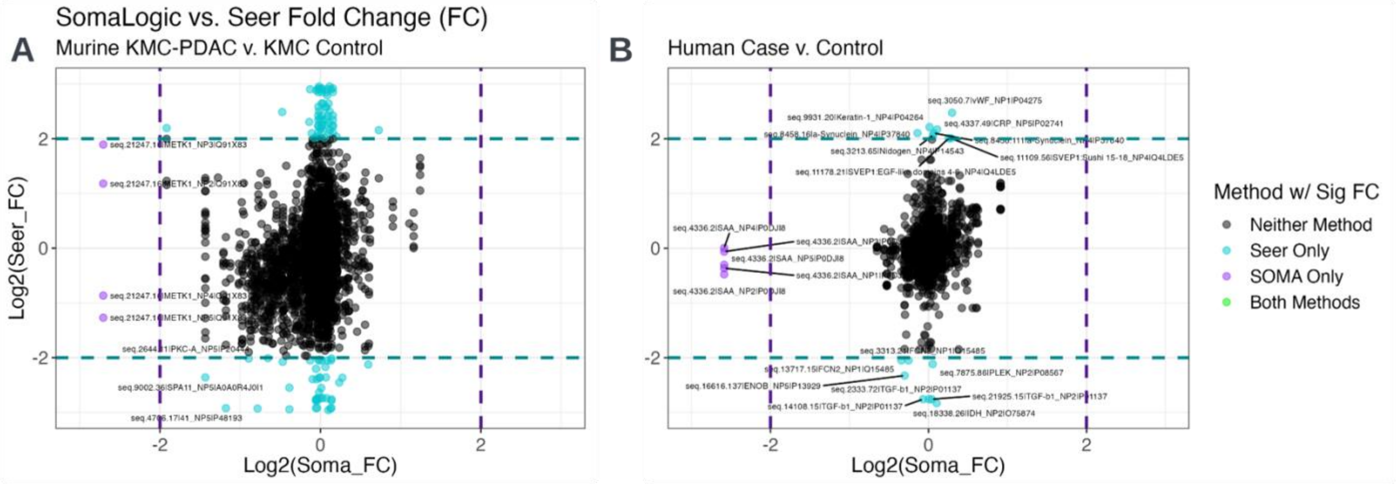

Figure 3: *Scatter plots comparing SomaLogic and Seer log2 FCs per overlap protein, identified by unique combinations of a SomaLogic aptamer and a Seer NP+UniProtID "feature" to assess method correlation. Horizontal teal lines indicate ± 2 Seer FC, vertical purple lines indicate ± 2 SomaLogic FC. A) Murine KMC-PDAC v. KMC Control overlap of 710 proteins, Pearson's correlation coefficient r = 0.15. B) Human case v. control overlap of 282 proteins, r = 0.23).*

Conclusion

As evidenced by the results above, the SomaLogic SomaScan platform is capable of providing proteomic quantification data with high technical quality – low assay variation demonstrated by low per-aptamer CVs and high panel sensitivity evidenced by the high number of aptamers with signal above noise across all samples. Additionally, the fact that differences in overall intensity distribution and elevated hemoglobin intensity were observed for the murine pink/red-annotated samples suggests that the SomaScan platform could screen for hemolysis contamination. Although dilution reduced aptamer panel sensitivity, low volume 35µL samples had comparable results to typical volume 55µL samples and likely could be used when limited sample quantities inhibit MS methods.

Despite these positive results from independent assessment of the SomaScan technology, very little correlation – in terms of FC and significance – was observed when comparing DE analysis results between the targeted SomaLogic platform and the non-targeted Seer Proteograph MS platform. Both technologies have been previously reported to independently show excellent proteomic coverage and high technical reproducibility, although MS-based methods have a more established record in the proteomic field than the SomaLogic aptamer-based method (*1,5,6,7*). However, discovery-phase MS-based methods like Seer Proteograph MS require orthogonal – non-MS – validation methods to confirm the results of potential clinical proteomic biomarkers. Most recently, several research groups present at the 2023 iHUPO (international Human Proteome Organization) conference shared discrepancies in DE results observed when comparing data produced by targeted (SomaLogic and Olink) and non-targeted (MS-based) proteomic analysis methods (*15*). The disagreement in DE FC and statistical significance for proteins detected by both platforms in this study provides further evidence to that discussed at the conference. Given the nature of this murine model experimental design that uses engineered genetic differences for expected extreme cancer phenotypic responses and the results described here, the current SomaLogic SomaScan is likely not a suitable validation or complementary method for Seer Proteograph MS.

Acknowledgements

I would like to acknowledge my mentors in the proteomics group at OHSU CEDAR – Dr. Matthew Chang, Dr. Mark Flory, and Jessie May Cartier – for recommending several of the key references for this paper, providing background information on the topics of MS-based proteomics research and clinical biomarker discovery, and sharing insights from iHUPO 2023. I would also like to thank the help team at SomaLogic for input on methods for data analysis, as well as the Seer and OHSU teams collaborating on the larger human prostate cancer and murine PDAC studies from which the data used here originated.

References

1. Y. Yan, S. Y. Yeon, C. Qian, S. You, W. Yang, On the Road to Accurate Protein Biomarkers in Prostate Cancer Diagnosis and Prognosis: Current Status and Future Advances. *IJMS*. **22**, 13537 (2021).

2. P. E. Geyer, L. M. Holdt, D. Teupser, M. Mann, Revisiting biomarker discovery by plasma proteomics. *Mol Syst Biol*. **13**, 942 (2017).

3. A. Sinha, V. Huang, J. Livingstone, J. Wang, N. S. Fox, N. Kurganovs, V. Ignatchenko, K. Fritsch, N. Donmez, L. E. Heisler, Y.-J. Shiah, C. Q. Yao, J. A. Alfaro, S. Volik, A. Lapuk, M. Fraser, K. Kron, A. Murison, M. Lupien, C. Sahinalp, C. C. Collins, B. Tetu, M. Masoomian, D. M. Berman, T. Van Der Kwast, R. G. Bristow, T. Kislinger, P. C. Boutros, The Proteogenomic Landscape of Curable Prostate Cancer. *Cancer Cell* **35**, 414-427.e6 (2019).

4. A. Khoo, L. Y. Liu, J. O. Nyalwidhe, O. J. Semmes, D. Vesprini, M. R. Downes, P. C. Boutros, S. K. Liu, T. Kislinger, Proteomic discovery of non-invasive biomarkers of localized prostate cancer using mass spectrometry. *Nat Rev Urol*. **18**, 707–724 (2021).

5. J. E. Blume, W. C. Manning, G. Troiano, D. Hornburg, M. Figa, L. Hesterberg, T. L. Platt, X. Zhao, R. A. Cuaresma, P. A. Everley, M. Ko, H. Liou, M. Mahoney, S. Ferdosi, E. M. Elgierari, C. Stolarczyk, B. Tangeysh, H. Xia, R. Benz, A. Siddiqui, S. A. Carr, P. Ma, R. Langer, V. Farias, O. C. Farokhzad, Rapid, deep and precise profiling of the plasma proteome with multi-nanoparticle protein corona. *Nat Commun*. **11**, 3662 (2020).

6. A. Joshi, M. Mayr, In Aptamers They Trust: Caveats of the SOMAscan Biomarker Discovery Platform From SomaLogic. *Circulation*. **138**, 2482–2485 (2018).

7. L. Gold, J. J. Walker, S. K. Wilcox, S. Williams, Advances in human proteomics at high scale with the SOMAscan proteomics platform. *New Biotechnology*. **29**, 543–549 (2012).

8. Seer Inc., Proteograph^{TM} Product Suite | Seer. *Seer* (2023), (available at https://seer.bio/products/proteograph-product-suite/).

9. Somalogic, The SOMASCAN platform - our science - platform. *SomaLogic* (2023), (available at https://somalogic.com/somascan-platform/).

10. V. Demichev, C. B. Messner, S. I. Vernardis, K. S. Lilley, M. Ralser, DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. *Nat Methods*. **17**, 41–44 (2020).

11. J. Cox, M. Mann, MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol*. **26**, 1367–1372 (2008).

12. SomaLogic, *DataDelve Statistics* (2023), (available at https://stats.somalogic.com/).

13. J. A. Canchola, Correct Use of Percent Coefficient of Variation (%CV) Formula for Log-Transformed Data. *MOJPB* **6** (2017).

14. Y. Benjamini, Y. Hochberg, Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)* **57**, 289–300 (1995).

15. Olink, Home - Olink. *Olink* (2023), (available at https://olink.com/).