**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 discovered genetic cancer markers, recent efforts in developing more accurate and effective panels have expanded to the evaluation of proteins as potential biomarkers. Two newer, more efficient 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 MS-based Seer Proteograph platform. In this study, two sample sets – one murine PDAC, one human prostate cancer – 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. The quality of the produced SomaLogic data for both human and mouse datasets was determined to be very good with low coefficients of variation (CVs) and high numbers of protein IDs across all samples, and low-volume samples were determined to have equivalent quality and results to typical volume samples. The diluted samples, while comparable in CV and total protein IDs, unfortunately resulted in decreased intensity quantification measurements. Additionally, very low correlation was observed in protein differential expression (DE) results between the SomaLogic and Seer data results for both human and mouse datasets.

**Introduction**

Current clinical cancer research is largely focused on developing new disease detection methods that are capable of identifying cancer at the earliest possible stage. Some of the most prominent of these detection methods are 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*). While these biomarker panels were originally composed to screen genomic cancer markers, recent research related to biomarker panel development has expanded to the discovery and validation of proteomic biomarker panels as well. Since 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 may best be accomplished by proteomic biomarker panels (*2*). Additionally, the clinical field has emphasized the need for panels that can be used to screen blood plasma (or serum) samples, as opposed to tissue, as this is one of the least invasive types of sample collection (*3*). The discovery of novel protein biomarkers in plasma is performed using a variety of proteomic techniques ranging from non-targeted mass spectrometry (MS) methods to targeted enzyme-linked immunosorbent assays (ELISAs) (*2*). MS-based methods tend to be preferred at present for initial discovery studies, as they accommodate detection of a broad, non-targeted range of proteins, whereas current antibody-based technologies like ELISAs rely on prior knowledge of the proteome of interest to design antibodies for each targeted potential biomarker. However, most MS-based proteomic methods require extensive sample preparation procedures to broaden the proteomic depth of coverage of the MS, allowing quantification of candidate biomarker proteins typically present in plasma in the µM-nM range (*1*).

Recently, two new methods of proteomic biomarker discovery were developed in the effort to make this discovery process more efficient while maintaining a high level of accuracy. 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 performs a MS sample preparation method that uses proprietary nanoparticle (NP) technology for non-specific protein capture based on the biophysical properties of plasma proteins (*4*). This step replaces the time-intensive sample preparation steps previously necessary to widen the proteomic depth of coverage for MS, while accomplished this same end result. 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 Seer Proteograph MS technique. Oregon Health & Science University Cancer Early Detection Advanced Research Center (OHSU CEDAR) currently uses Seer Proteograph MS for plasma protein quantification and proteomic analysis in several ongoing studies for cancer proteomic biomarker research, and has recently been evaluating methods of orthogonal (non-MS) validation for the Seer Proteograph MS platform.

One potential candidate as an orthogonal validation method is the newly updated 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 (*5,6*). The current aptamer “panel” of this platform consists of 7596 aptamers, each targeting a specific protein isoform, and this panel is expected to expand to 10000 aptamers in the coming months. These aptamers fluorescently label their target proteins in a series of reactions during sample preparation, and protein quantification is performed through fluorescence intensity measurement on a DNA microarray (*1*).

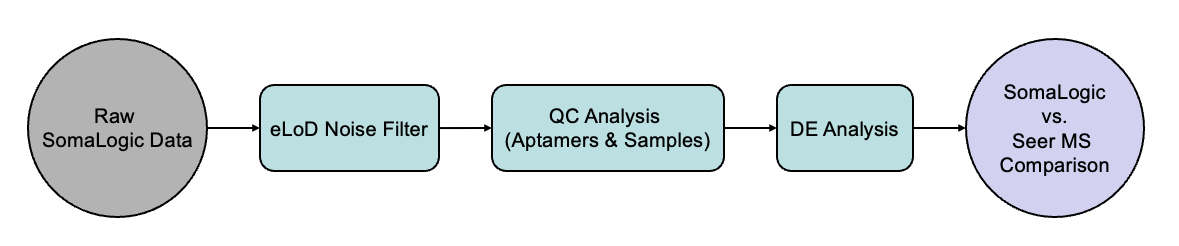
While both novel methods mentioned above have been demonstrated to generate high quality datasets, a current focus of clinical proteomics research is determining how well the results of an aptamer-based method of proteomic analysis correspond to the results of a MS-based method of proteomic analysis. As such, the OHSU CEDAR proteomics group conducted a study to compare the data quality and proteomic analysis results of the newer SomaLogic SOMAscan method to the currently used Seer Proteograph MS method, with the end goal of assessing whether the SomaLogic SomaScan platform is a suitable orthogonal validation method, or at least a complementary method, to Seer Proteograph MS.

**Methods**

Two datasets were used in the course of this study to compare proteomic analysis results of the two platforms of interest – Seer Proteograph MS and SomaLogic SomaScan (*7,8*). The first dataset consisted of 70 individual samples from a larger sample pool collected for a murine pancreatic ductal adenocarcinoma (PDAC) proteomic study. Eight conditions 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” designation refers to samples taken from mice at a chronologically earlier stage of the study, and the “Late” designation refers to samples taken from mice at a later stage. In addition to these biological conditions, two additional samples noted with a pink/red color possibly indicating hemolysis were also included in this murine dataset to fulfill a side objective of testing SomaScan’s ability to screen for such contaminated samples. Lastly, an additional nine samples of plasma pooled from the biological samples were included to evaluate the validity of low or diluted volume samples: 3x55uL typical volume, 3x35uL low volume, 3x35uL plasma diluted with PBS buffer to 55uL total volume. All 81 samples were run on the Seer Proteograph MS at OHSU CEDAR and these results subsequently searched using Seer PAS DIA-NN quantitative proteomics software (*9*). All samples were also submitted to SomaLogic for proteomic quantification via SomaScan, from which the resulting data was made available through SomaLogic’s online portal. SomaLogic added four buffer samples and three calibration samples for quality control of their process, and with the addition of those samples we received results on 88 total samples.

The second dataset consisted of 30 individual patient samples, subset from a larger human prostate cancer proteomic study, with this subset evenly split between Case (cancer) and Control (non-cancer) groups. Nine additional samples of pooled plasma were also prepared for this human sample set: 3x55uL typical volume, 3x35uL low volume, 3x35uL plasma diluted with PBS buffer to 55uL total volume. All 39 samples were run on the Seer Proteograph MS at OHSU CEDAR and the results searched using MaxQuant quantitative proteomics software (*10*). All samples were also sent to SomaLogic for proteomic quantification via SomaScan, 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 (Figure 1). First, application of an estimated limit of detection (eLoD)-based noise filter based on buffer intensity results for each of the 7596 aptamers in the panel, using a formula recommended by SomaLogic (eLoD per 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 Seer Proteograph MS data. SomaLogic provided two versions of each dataset – one with all of their internal normalization steps completed, and another without the final normalization step completed. This normalization step is intended 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 – these are available via GitHub (<https://github.com/ssoriano22/SomaLogic_Benchmark>). As Seer data quality has been previously confirmed by the OHSU CEDAR proteomics group, more focus in this study was placed on evaluating SomaLogic data quality. SomaLogic’s online portal for analysis – SomaLogic DataDelve – was used to confirm SomaLogic DE analyses performed with custom analysis pipeline (*11*).



***Figure 1:*** *Workflow diagram for this proteomic method comparison study between SomaLogic SomaScan and Seer Proteograph MS, applicable to both the murine and human datasets. The eLoD-based noise filter removes intensity results below background fluorescence threshold.*

**Results & Discussion**

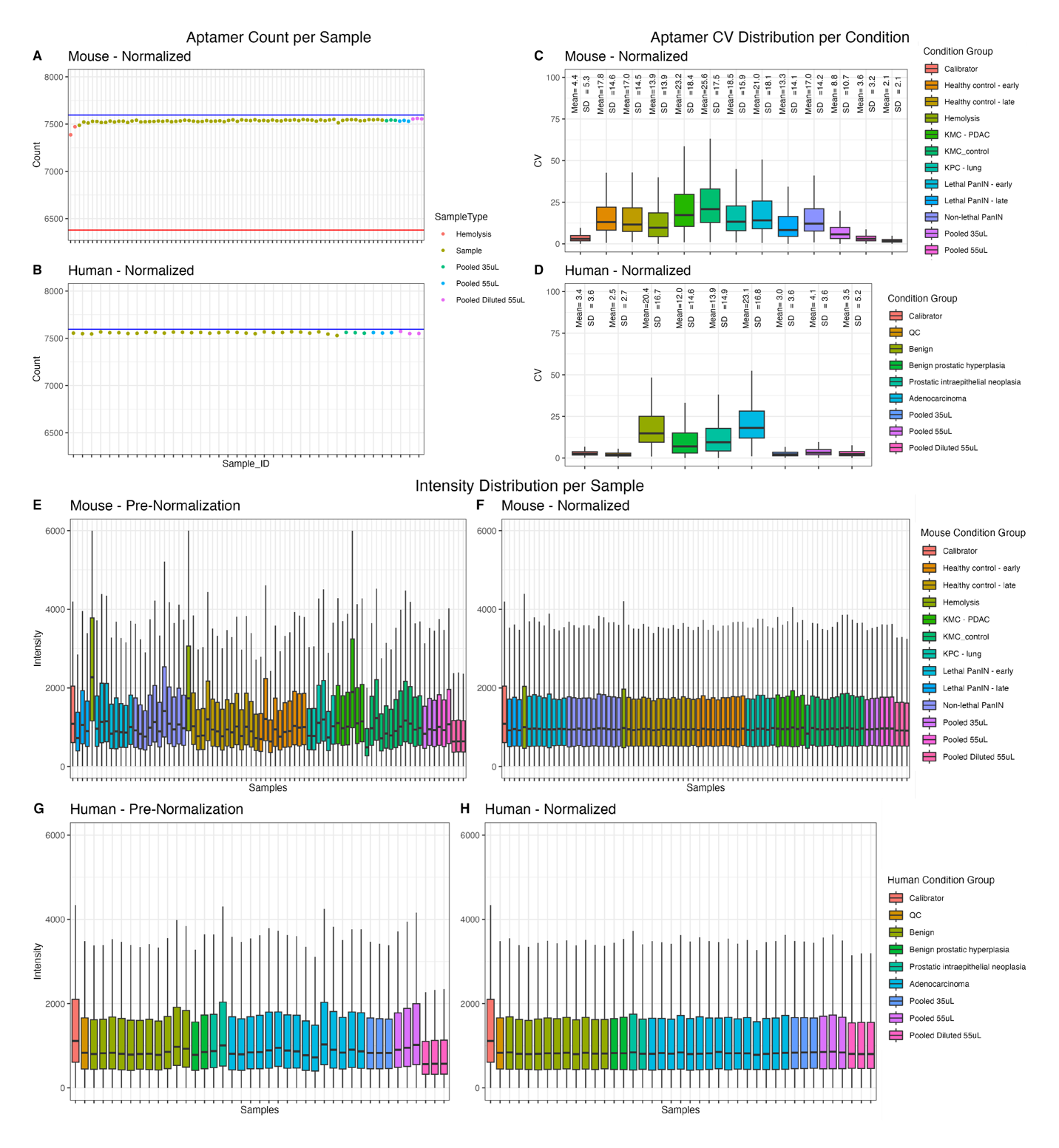
*Quality Control (QC) Analysis:*

The first step in assessing the 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 2A). Two of the lowest samples were the pink/red annotated suspected hemolysis samples, although their counts were still very high at 7387 and 7472 aptamers with signal above noise. No major aptamer count difference 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 a 7500 aptamer count and a lowest count of 7351. Overall, there was very little difference between the two versions of the mouse dataset, with aptamer counts 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 aptamer counts above 7500, and the “Pre-Normalization” dataset only had one sample with an aptamer count below at 7482 (Figure 2B).

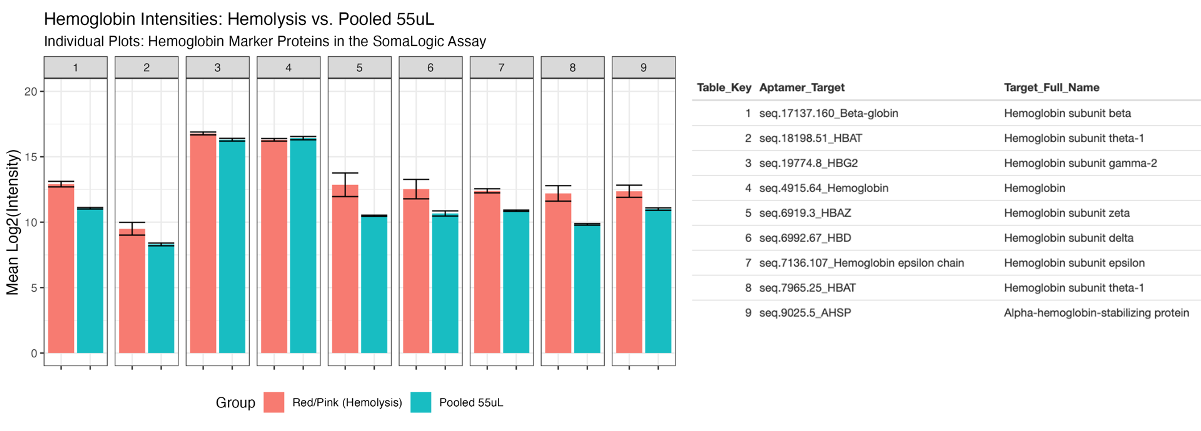
Next, the distribution of aptamer coefficients of variation (CV) per condition group was surveyed to quantify the technical variability of the SomaLogic assay and the biological variability between samples. Both murine and human aptamer intensity CVs were calculated using log2-transformed intensity data, then grouped by condition in boxplots for each species dataset (Figure 2C,D) (*12*). The “Pre-Normalization” data for both datasets showed higher mean CVs across all groups compared to the “Normalized” data, as expected. 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 observed to be very low in the analysis of 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 the fact 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. Concerning variability observed between the biological condition samples, mean aptamer CVs were higher across all conditions in the murine dataset compared to those in the human dataset, despite the genetic similarity of the mice these samples originated from. One theory to explain this involves the traumatic sample collection process for murine PDAC plasma samples introducing molecular differences between samples collected from different mice, whereas a clinical human blood draw does not typically result in systemic side effects. Overall, the amount of biological variability observed in both the murine and human SomaLogic “Normalized” datasets is relatively similar to that observed in other Seer Proteograph MS studies currently in progress.

The intensity distribution per sample was also evaluated, primarily to check for any samples and/or conditions with abnormal intensity results across aptamers, but also to better visualize the effect of SomaLogic’s final normalization step on both intensity datasets (Figure 2E-H). As anticipated, 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, likely due to the fact that SomaLogic’s panel had originally been designed to target human proteins, and has only recently expanded to use in murine projects. In addition to these expected results, 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. Furthermore, both “dilution sets” provided interesting results – the diluted samples in both species datasets have lower intensity distributions than the typical volume and low volume samples, suggesting that diluting a sample should be avoided, as it reduces the fluorescence signal used for quantification, but signal intensity is not affected in the low volume samples. It was noted that normalization did adjust the diluted intensity distributions in both species datasets to be comparable to all other samples in the “dilution set”.

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 (Figure 3). For at least eight of the nine hemoglobin marker proteins in the “Pre-Normalization” dataset, the pink/red samples do have a higher mean log2 intensity than the 55uL samples, suggesting that SomaLogic’s assay may be able to screen for hemolyzed samples in the future. It was also observed that normalization removed all of these observed discrepancies in protein intensities between the suspected hemolysis and pooled typical volume samples.



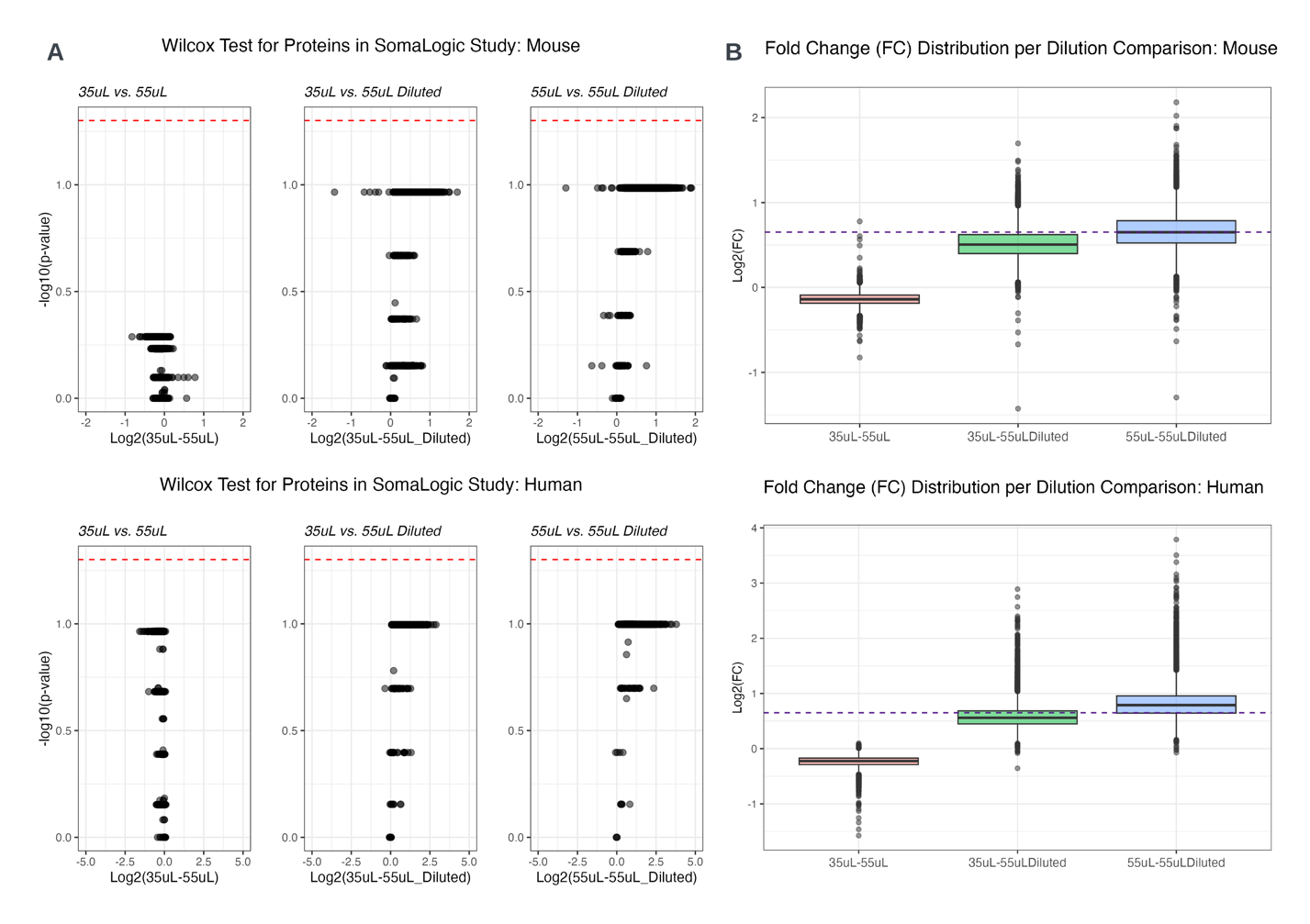
***Figure 2:*** *SomaLogic data QC results evaluating aptamer panel performance across samples in both the murine and human datasets. A,B) Aptamer counts per sample (81 murine, 39 human “Normalized”) after eLoD noise filter. Blue lines indicate maximum 7596 aptamers, red line marks the 84% of aptamers in the panel with expected good performance in murine studies. C,D) Aptamer CV distributions per condition. E-H) Intensity distributions divided by sample and colored by condition. Outliers are hidden on all boxplots.*



***Figure 3****: SomaLogic QC results comparing mean log2 protein intensities for nine hemoglobin-targeting aptamers in suspected hemolysis samples to those in pooled typical 55µL volume samples. Error bars represent +/- 1 SD between technical replicates.*

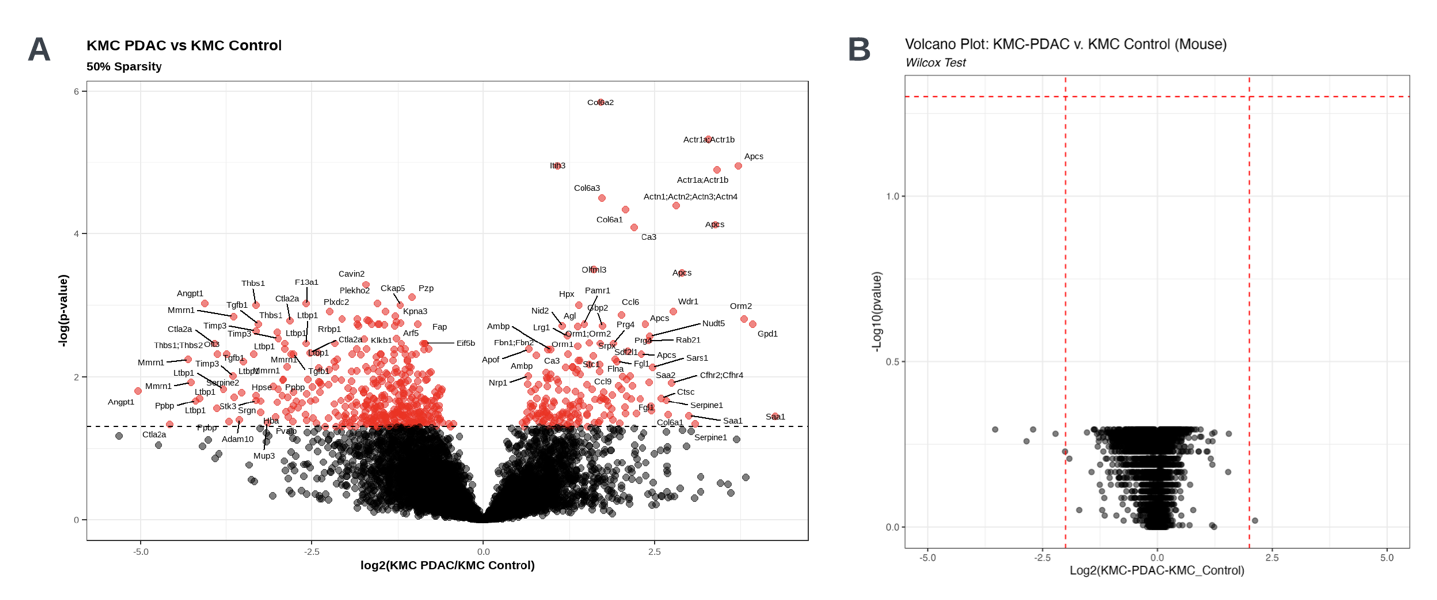
*Differential Abundance/Expression (DE) Analysis:*

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). For each protein in common between the compared conditions, log2 fold change (FC) – the difference between intensity values – was calculated, and Wilcox tests were performed with Benjamini–Hochberg (BH) False Discovery Rate (FDR) correction. The resulting FCs and p-values were plotted in volcano plots – one for each dilution condition comparison per species dataset (Figure 4A). In the mouse dataset, 7422 proteins had sufficient data for comparison (present in at least 2/3 condition replicates), while the human dataset had a slightly higher 7491 proteins. No proteins with significant differential abundance were found in any of the mouse or human dilution condition comparisons, confirming that lowering sample volume to 35µL or reducing concentration did not have a significant effect on protein quantifications. However, it was noted that the diluted replicates had 50-100 fewer proteins than the low volume and typical volume replicates in both the mouse and human datasets, reaffirming that the dilution did result in an unfortunate increase in data sparsity. To further investigate these diluted samples, boxplots were used to confirm that proteins in the murine and human diluted samples showed the expected FC for the planned 1.6x dilutions (Figure 4B). The median protein FC for the low volume-typical volume comparison was close to the expected 0 (1:1 protein intensities), and the typical volume-diluted volume median protein FC was close to the 1.6x target.



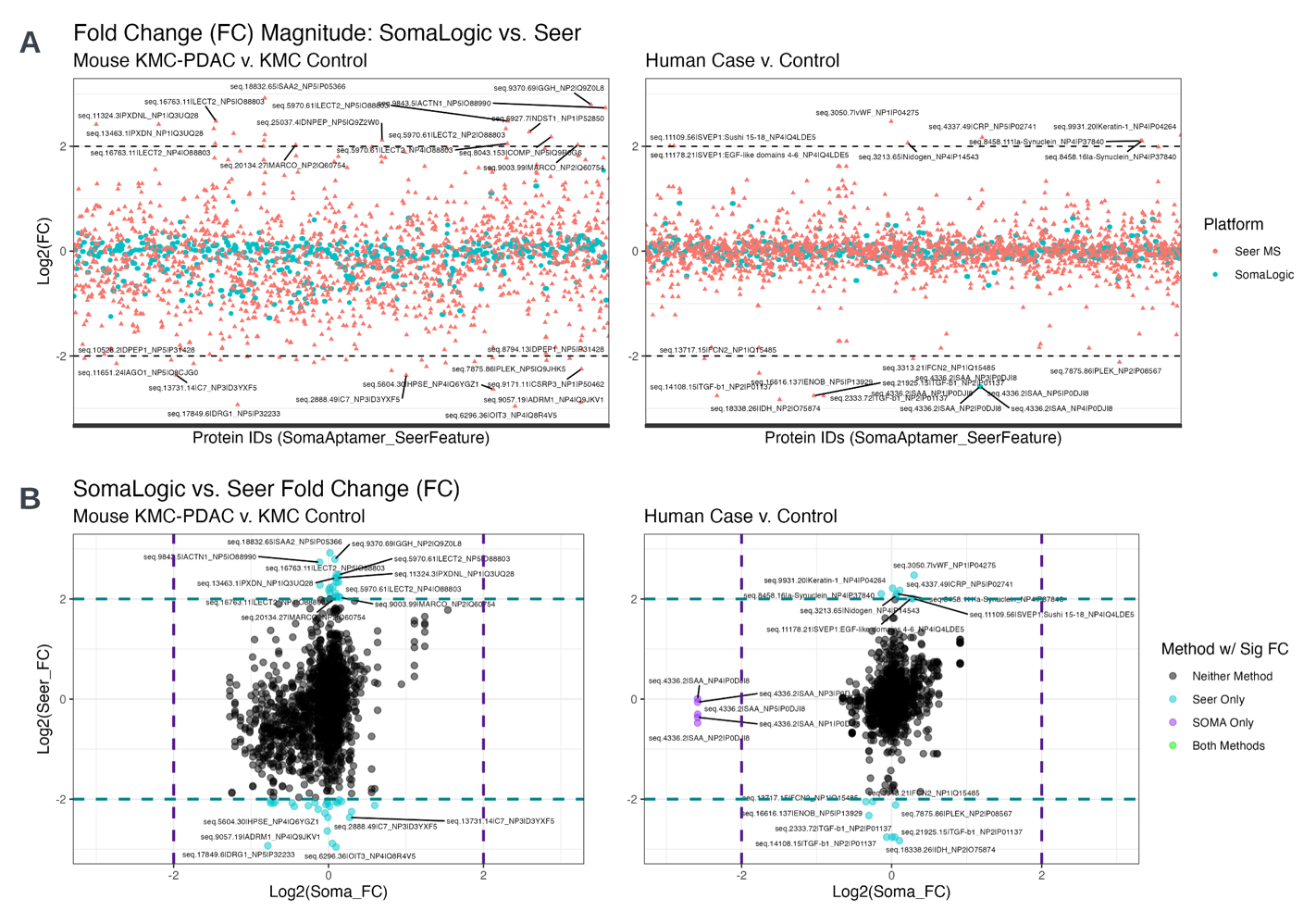
***Figure 4:*** *Differential abundance analysis for pooled plasma “dilution sets” in both mouse and human datasets. A) Volcano plots evaluating proteins for significant FC (red line, 𝛼 = 0.05) between conditions. B) Boxplots confirming diluted protein median FC are on target (purple line = log2(1.6)), and that the low volume protein FCs are within an equivalent range (approx. 1:1) to the typical volume protein FCs.*

The next DE analysis focused on one of the murine condition comparisons of biological interest that was known to have differentially expressed proteins in the larger Seer Proteograph murine study – KMC-PDAC vs. KMC Control. As described in the initial “dilution set” DE analyses, FC was calculated for each of the 7545 proteins with sufficient data, and significance determined by p-values resulting from BH-corrected Wilcox tests. No proteins with significant differential expression were found using the SomaLogic data, in contrast to the previous Seer Proteograph MS results for this comparison from the larger murine study in process at OHSU CEDAR (Figure 5). Both Seer and SomaLogic analysis used the same number of samples in each condition group for this differential result (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 no difference was apparent in the significance of the SomaLogic DE results 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 the DataDelve portal clearly validated the results of the custom analysis pipeline.



***Figure 5:*** *Murine DE analysis comparing KMC-PDAC v. KMC Control condition groups. A) Seer Proteograph MS results (50% sparsity filter – protein ID present in at least 50% of samples for a condition, FC significance determined by Welch’s t-test 𝛼 = 0.05). Seer NPs can introduce duplicate data points labeled with identical gene/protein names. B) SomaLogic SomaScan results (FC significance determined by Wilcox test 𝛼 = 0.05).*

In an effort to explain this obvious discrepancy between the Seer Proteograph MS and SomaLogic SomaScan DE results, the overlap of protein IDs detected with both platforms was assessed based on gene names present in both datasets for a species. Because gene names are being used as the basis for determining protein ID overlap in this study, it is possible that some protein IDs are lost in cases when multiple proteins map to a single gene, and the following overlap numbers are acknowledged to be approximations. The number of overlapping protein IDs for the murine KMC-PDAC v. KMC Control comparison was 475, out of 2571 Seer proteins and 7545 SomaLogic aptamers (18% of Seer’s protein total), with 57 of these overlap proteins identified as significantly differentially expressed in the Seer results. In the human Case v. Control DE analysis (for which there were no significant proteins identified by either platform), 320 proteins were found to overlap between platforms, from 409 Seer proteins and 7577 SomaLogic aptamers (78% of Seer’s protein total). From these numbers, the two platforms clearly have higher protein overlap in the human dataset compared to the murine. FCs of the overlap proteins were also graphically compared between methods for both datasets. The SomaLogic murine overlap proteins generally have smaller magnitude FCs compared to Seer FCs for the same proteins, while this difference is much less apparent in the comparison of human overlap protein FCs (Figure 6A). Finally, method correlation in FC was calculated (Pearson’s r) from the direct comparison of SomaLogic and Seer FCs for each overlapping protein (Figure 6B). This correlation was found to be very low for both the murine overlap proteins (r = 0.27) and the human overlap proteins (r = 0.23).



***Figure 6:*** *Survey of protein ID overlap between Seer and SomaLogic methods. A) Scatter plots of log2 FCs for 475 overlap murine proteins and 320 overlap human proteins showing FC magnitudes per platform. Cross-platform protein IDs used on the x-axis are unique combinations of a SomaLogic aptamer and a Seer NP+Protein “feature” for a given gene name in the overlap. B) Scatter plots directly comparing SomaLogic and Seer FCs per overlap protein to assess method correlation (murine Pearson’s r = 0.27, human r = 0.23).*

**Conclusion**

As evidenced by the results detailed above, the SomaLogic SomaScan platform does provide a high-quality proteomic quantification alternative to MS methods like Seer Proteograph MS. Additionally, the fact that differences in overall intensity distribution and hemoglobin intensity were observed for the murine pink/red suspected hemolysis samples suggests that the SomaScan platform could be used to screen for hemolysis contamination. However, despite these positive results documented when assessing the independent success of the SomaScan technology, very little correlation was observed when comparing DE analysis results between the targeted SomaLogic platform and the non-targeted Seer Proteograph MS platform. Both of these technologies have been previously established as high quality methods for proteomic quantification (*1,4,5,6*). The low level of agreement between methods demonstrated in these plasma proteomic analysis results, and noted at recent global proteomic conferences, shows a clear need for further study of this topic to determine a root cause. Until such a cause is found, the SomaLogic SomaScan platform is likely not an acceptable orthogonal validation method for proteomic quantification results produced by the currently used Seer Proteograph MS at OHSU CEDAR.

**Acknowledgements**

I would like to acknowledge my mentors in the proteomics group at OHSU CEDAR – Dr. Matthew Chang and Dr. Mark Flory – for recommending several of the key references for this paper, and for providing background information on the topic of MS-based proteomics research and clinical biomarker discovery related to cancer. 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. 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).

4. 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).

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

6. 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).

7. Seer Inc., ProteographTM Product Suite | Seer. *Seer* (2023), (available at <https://seer.bio/products/proteograph-product-suite/>).

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

9. 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).

10. 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).

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

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