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*).

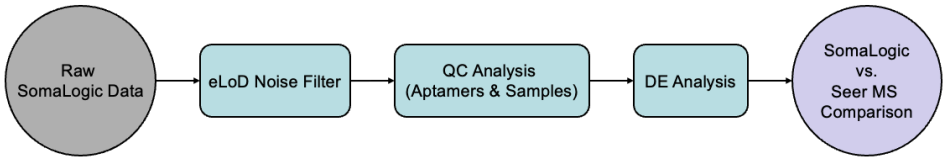


Figure 1: *Workflow diagram for this proteomic method comparison study between SomaLogic SomaScan and Seer Proteograph MS, applicable to both the murine and human datasets. The eLoD-based noise filter removes intensity results below background fluorescence threshold.*

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 intensity distribution per sample was assessed to check for any samples with abnormal intensity results for any aptamer.

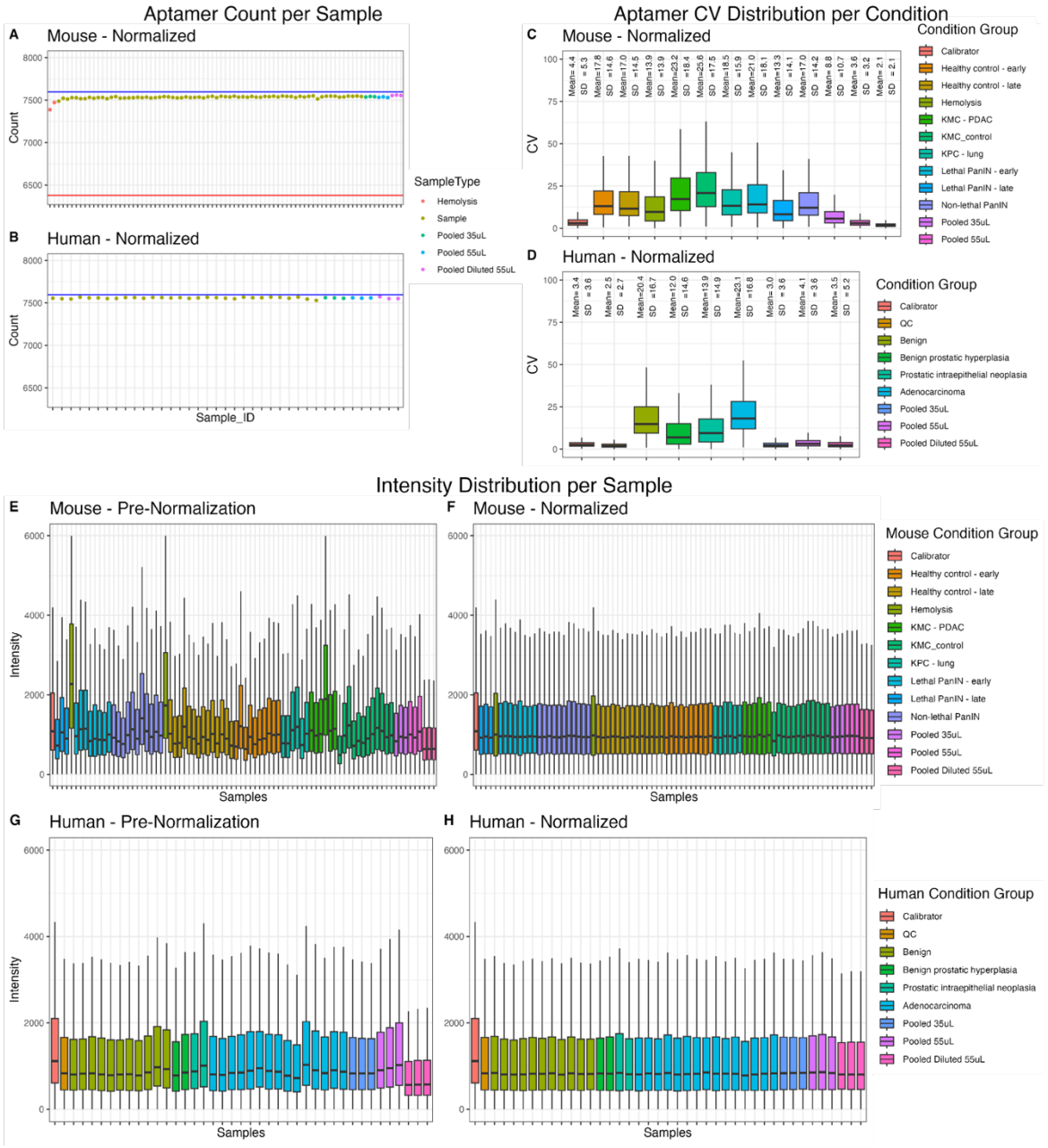


Figure 2: SomaLogic aptamer performance and data quality assessment.

A) Shows the number of significant aptamers detected in each of 81 murine samples
B) Same across 39 human samples. Blue line on each plot marks the total number of aptamers in the assay, 7596. Red line on the mouse plot shows the number of aptamers with expected good performance in murine studies according to SomaLogic (87% of aptamer panel). In both species datasets, a consistently high number of protein IDs are detected across all samples.

- The two pink/red suspected hemolysis samples – marked by the red arrow – have slightly lower aptamer counts, implying fewer aptamers with significant intensity signal in these samples.
- The dilution series pooled samples – low volume 35uL, typical 55uL, and low volume diluted to 55uL – are marked by the red brackets on both plots. No differences in aptamer count are apparent across either dilution series, potentially supporting the validity of these low plasma volume samples.

C) Intensity distribution boxplots divided by sample and color grouped by condition. Somalogic applies a species-specific final normalization step to their output data. As such, they supplied us with both pre-normalized and normalized data versions.

D) Distribution of protein fluorescent intensities across all murine samples for the pre-normalized data (left) and normalized data (right). As expected, the normalization does reduce variation in these distributions.
- Pink/red suspected hemolysis samples have higher shifted intensity distribution compared to other samples in both pre-normalized and normalized
- In the pre-normalized data, the diluted samples have a lower intensity distribution compared to the other dilution series samples. In the normalized data, the diluted intensity range is adjusted to be comparable to the others in the series. From the pattern observed in the pre-normalized dilution series, the fact that low intensities are observed for proteins in the diluted samples but not the low volume samples, suggests that submitting 35uL samples is possible, but dilutions should be avoided.

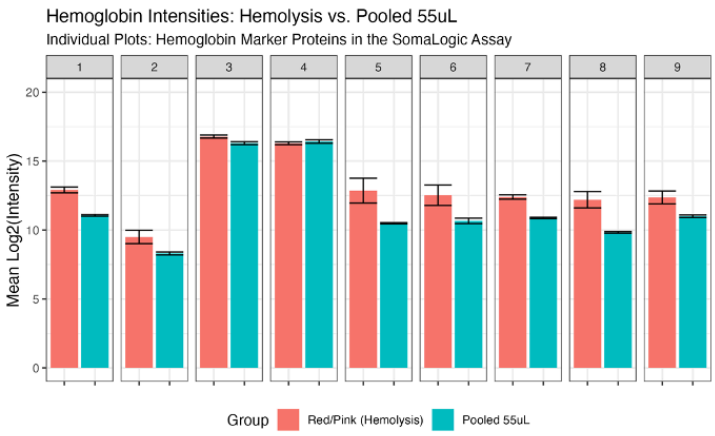
E) Same intensity distribution plots for the human dataset. The dilution series is again marked with red brackets, and the same pattern is observed for the dilution series between pre-normalized and normalized data.

F) Additionally, the human distributions appear to be even more consistent across samples compared to mouse. One likely reason is the SomaLogic assay has only recently started being used for mouse samples, and the assay has not yet been tuned to the same low-variance performance displayed in their analyses of human datasets.

G) In this plot for the mouse data, we calculated the aptamer intensity CVs, grouped by mouse condition to get an idea of the technical variability of the assay and biological variability of our samples.
- Technical assay variability observed in the pooled 55uL result is very low, implying promising technical reproducibility of the assay.
- Second, although the aptamer counts and intensity distributions were similar between the 35 and 55uL samples, there does appear to be more data variation in the low volume samples. Also of note is that the diluted samples have minimal variability, for which we don't yet have an explanation.
- Decent amount of biological variability in our main set of mouse samples – denoted in the red box. Although these samples were taken from genetically identical mice, this variability could be stemming from the murine sample collection process.

- Comparatively, the amount of biological variability observed here is relatively similar to that observed in the Seer Proteograph MS analysis of these samples.

H) Same aptamer CV by condition plot to assess the variability of our human samples. Even less aptamer variability is seen in the human sample groups – both in terms of technical and biological variability.



Table_Key	Aptamer_Target	Target_Full_Name
1	seq.17137.160_Beta-globin	Hemoglobin subunit beta
2	seq.18198.51_HBAT	Hemoglobin subunit theta-1
3	seq.19774.8_HBG2	Hemoglobin subunit gamma-2
4	seq.4915.64_Hemoglobin	Hemoglobin
5	seq.6919.3_HBAZ	Hemoglobin subunit zeta
6	seq.6992.67_HBD	Hemoglobin subunit delta
7	seq.7136.107_Hemoglobin epsilon chain	Hemoglobin subunit epsilon
8	seq.7965.25_HBAT	Hemoglobin subunit theta-1
9	seq.9025.5_AHSP	Alpha-hemoglobin-stabilizing protein

Figure 3: Hemolysis Results. Compared the mean log2 intensities of 9 hemoglobin marker proteins in SomaLogic's assay between the 2 pink/red murine samples and the 3 pooled 55uL samples. For at least 8/9 hemoglobin marker proteins, 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. Error bars represent +/- 1 SD between technical replicates.

Differential Abundance/Expression (DE) Analysis:

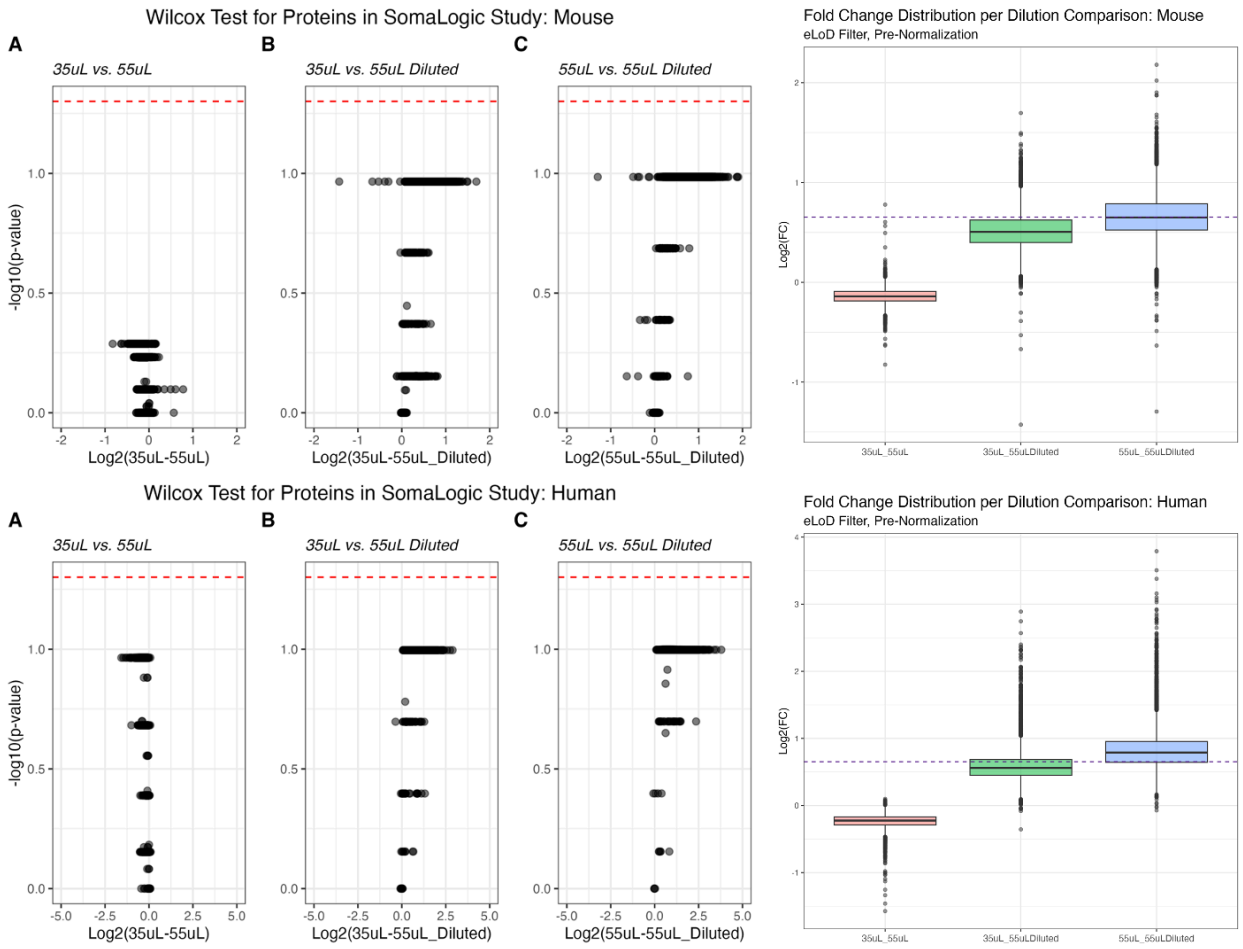


Figure 4: Comparing the dilution series samples to further assess viability of low volume samples.

A) Calculated log2(FC) and performed Wilcox tests with FDR correction to get adjusted pvalues for the differential abundance volcano plots shown here (7422 tested aptamers). No additional filters are applied other than the eLoD noise filter, and any protein IDs too sparse across samples for the test are removed.
   - No significant proteins of differential expression were identified, which is good because despite dilution differences, all samples here should be biologically similar.
   - Diluted samples had noticeably fewer IDs tested compared to the other series samples, around 100 less, implying sparser sets of protein IDs in the dilutions.

B) Matt had also noticed that these volcano plots appeared oddly shaped for dilution series comparisons he had seen previously, so he advised me to make this boxplot of FCs per dilution comparison to see if the median FC in each condition was approximately on target for what the expected dilution fold change should be.
   - Red: Compares low volume to typical volume protein FCs. The median FC being close to 0 on the y-axis means that the expected 1:1 relationship in FC with no dilution is in fact true here.

- Blue: Compares typical volume to the diluted sample protein FCs. The median FC close to the expected dilution FC marked by the purple line here indicates that the FC of most proteins do adjust as expected with the dilution from 35 to 55uL.
- Green: Further reinforces the takeaways from the other two comparisons, showing that low volume to diluted sample protein FCs follow the same pattern as the typical volume to diluted sample comparison, with the median FC close to the expected dilution target.

C) Same series of calculations and tests were performed for the human dataset as well – similarly no significant proteins identified (7491 tested aptamers)
- Same lower number of tested protein IDs in the diluted samples was again observed, implying similar data sparsity concerns with the dilution.

D) Same dilution FC boxplot for the human data. From brief observation, I would say that the human dilution appears to be even more on target for the expected FC, with less variation.

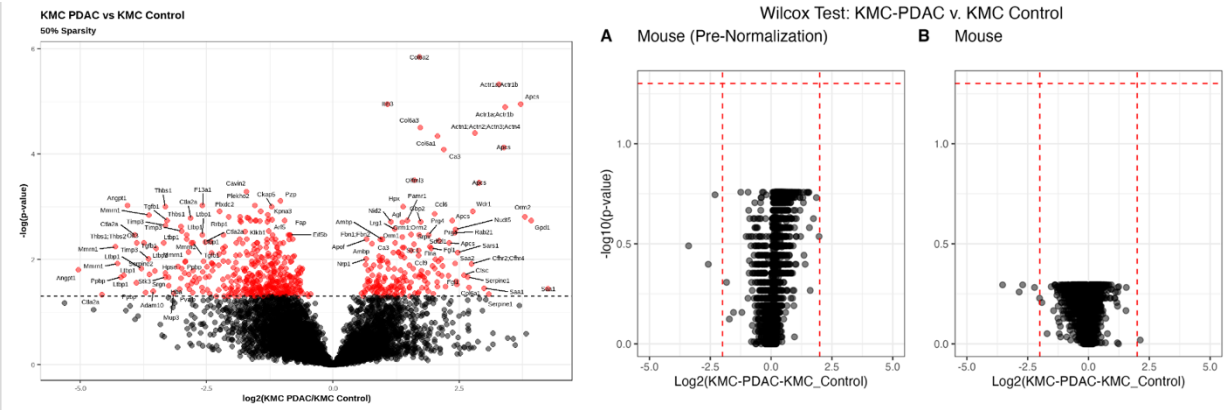


Figure 5: We knew that at least one of the mouse comparisons resulted in many significant protein results using the Seer Proteograph MS method. [Human differentials using initial Seer MS data did not ID any significant proteins.]

A) The Seer DE results from that known condition – KMC-PDAC vs. KMC control – generated from Welch's t-tests, are shown here with a 50% sparsity (sample coverage) filter applied [meaning the protein ID had to appear in at least 50% of samples for a condition to be included]

B) This is the SomaLogic result for that same mouse condition comparison, using the pre-normalized data on the left, normalized on the right.
- 6 v. 12 samples (Soma and Seer).
- No significant proteins appear to be identified between these two conditions by the SomaLogic assay.
- To confirm that it wasn't an issue with my code, I entered the same raw data files into SomaLogic's DataDelve online data analysis portal to see if they would get a similar result, and for the most part it did! Which was great news for me, but slightly more confusing news regarding this apparent difference between the SomaLogic and Seer murine results for this comparison

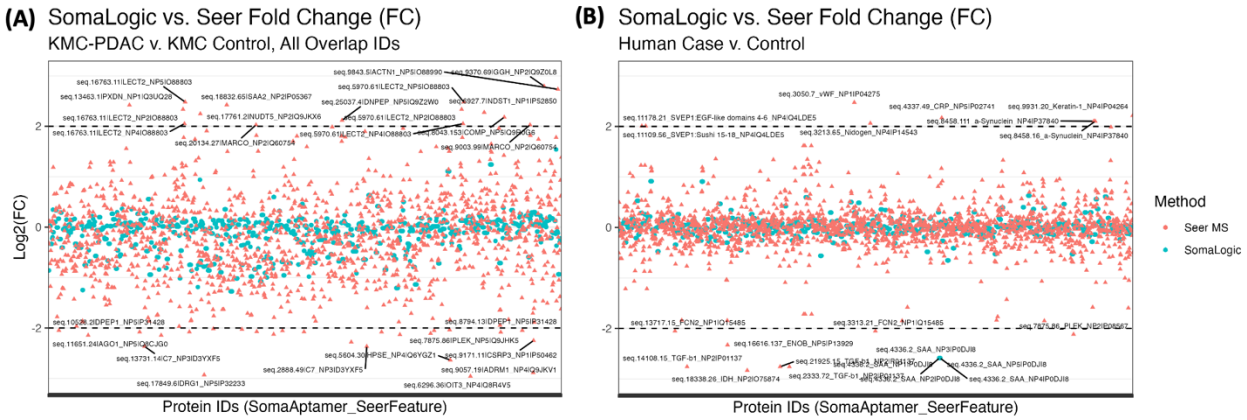


Figure 6: In the effort to begin understanding this difference in result between the two platforms, I started with looking for FC differences in overlapping proteins between the two methods.

A) I first found the number of overlapping proteins IDs between the SomaLogic and Seer data for this murine comparison – out of the 7500 (7545) SomaLogic aptamers and the 2500 (2571) total Seer proteins, 475 proteins were found in both datasets.
   - Plotted the log2(FC) for each method – Seer in red triangles, SomaLogic in blue circles – for each overlapping protein ID in the murine data. The plot on the left shows all 475 overlapping proteins

B) Middle plot shows only the 57 overlapping proteins that were identified as significant in the Seer DE results. There are actually more data points because of Seer's NP use creating up to 5 features for each protein.
   - Main takeaway here is that the murine Seer results appear to have higher magnitude fold changes per protein compared to the murine SomaLogic results.

C) Same FC plot but for the human DE comparison.
   - Here, 320 overlapping proteins between the 409 Seer proteins and 7577 SomaLogic aptamers compared in the Human Case (tumor) v. Control DE analysis. No significant proteins were identified by either platform.
   - The overlap is slightly better here (78% of the Seer proteins), and there might be less of a magnitude difference in FCs between these human Seer and SomaLogic results.
   - After seeing this magnitude difference in both datasets between methods, we wanted a more direct FC comparison between methods for each SomaLogicAptamer_SeerFeature pair.

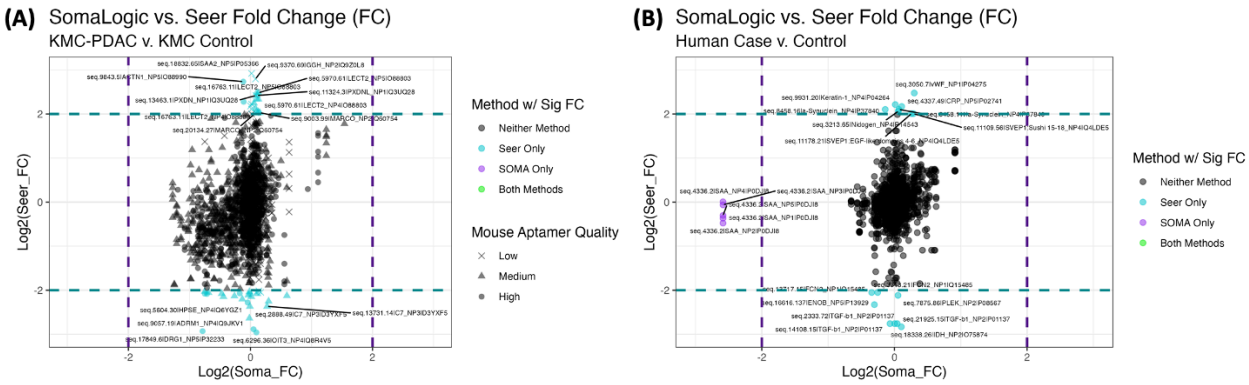


Figure 7: I plotted the SomaLogic FC on the x-axis – against Seer FC on the y-axis for each aptamer_feature match.

A) This plot is for the murine data. As noted by the teal data points, several protein IDs only have a detected FC in the Seer data, with almost none detected in the SomaLogic data.
   - Ideally, we'd like to see some manner of diagonal agreement here, which while maybe hinted at, is not very apparent.
   - Numerically this method difference is apparent given the low Pearson's correlation coefficient of 0.27.

B) Murine plot color coded by murine aptamer quality as determined by SomaLogic.
   - Aptamer quality does not appear to be affecting correlation in murine data.

C) This plot is for the human data. There are also teal data points here, indicating several protein IDs with FCs only detected in the Seer data, and one protein (SAA) with FC only detected in the SomaLogic data.
   - Method agreement is even less apparent here, with Pearson's r is still very low at 0.23.

Conclusion

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 Dr. Leslie Coonrod from University of Oregon BGMP for reviewing and providing feedback during the writing process.

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., Proteograph™ 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/).