Bruker Spatial Biology

Unlocking the Power of GeoMx® Digital Spatial Profiling: Best Practices for Protein Data Analysis

WHITEPAPER

Introduction

GeoMx® Digital Spatial Profiling provides a powerful tool to measure gene and protein expression with spatial context. In this paper, we outline core principles and best practices of analyzing a GeoMx DSP protein dataset.

There is no single correct approach to processing spatial biology data, and the methodologies often vary depending on experimental design, researcher preferences, or field-specific norms. However, certain best practices are widely recommended to ensure robust results. These include:

- Demonstrating QC of ROI/AOI and Probe Targets
- Showing QC plots for normalization factors
- Assessing concordance among the different factors
- Selecting an appropriate normalization method

The sample study used here represents a small protein dataset (~60-plex panel; two colorectal cancer slides; 90 regions of interest (ROI) each segmented into 2 areas of illumination (AOI): tumor and tumor microenvironment (TME)). However, these principles and tools can be applied to high-plex GeoMx protein assays such as the IO Proteome Atlas (IPA) with >570 targets. For additional guidance, refer to **Getting Started with Large Protein Datasets**. (On Page 9)

Note: we present here data that has passed the initial QC step in the GeoMx DSP Data Analysis Suite. The plots shown below can be generated using the Evaluate Normalization for Protein script and vignette, available from the GeoScript Hub.

AOI QC

The purpose of AOI-level QC is to identify AOIs with poor data that should be removed. We should look at both signal strength and background.

First, we will compute 2 metrics of AOI technical performance:

- Housekeeper geomean: this captures signal strength.
- IgG geomean: this captures background (negative controls), but in most experiments also reflects signal strength, as AOIs with more on-target signal also have more background.

We will evaluate the dataset for these metrics of signal strength and look for outliers with low signal. AOIs with extremely low signal strength (i.e. many standard deviations away from the mean signal) compared to the rest may be unreliable and should be considered carefully.

Here's a look at signal strength as measured by the housekeepers **(Figure 1)**. We use a log scale. (Note that log2(geomean) = mean(log2 expression).)

There are no clear outlier AOIs with low signal here, so we can safely use all AOIs. We do not see preferential clustering based on AOI Type or Tissue ID.

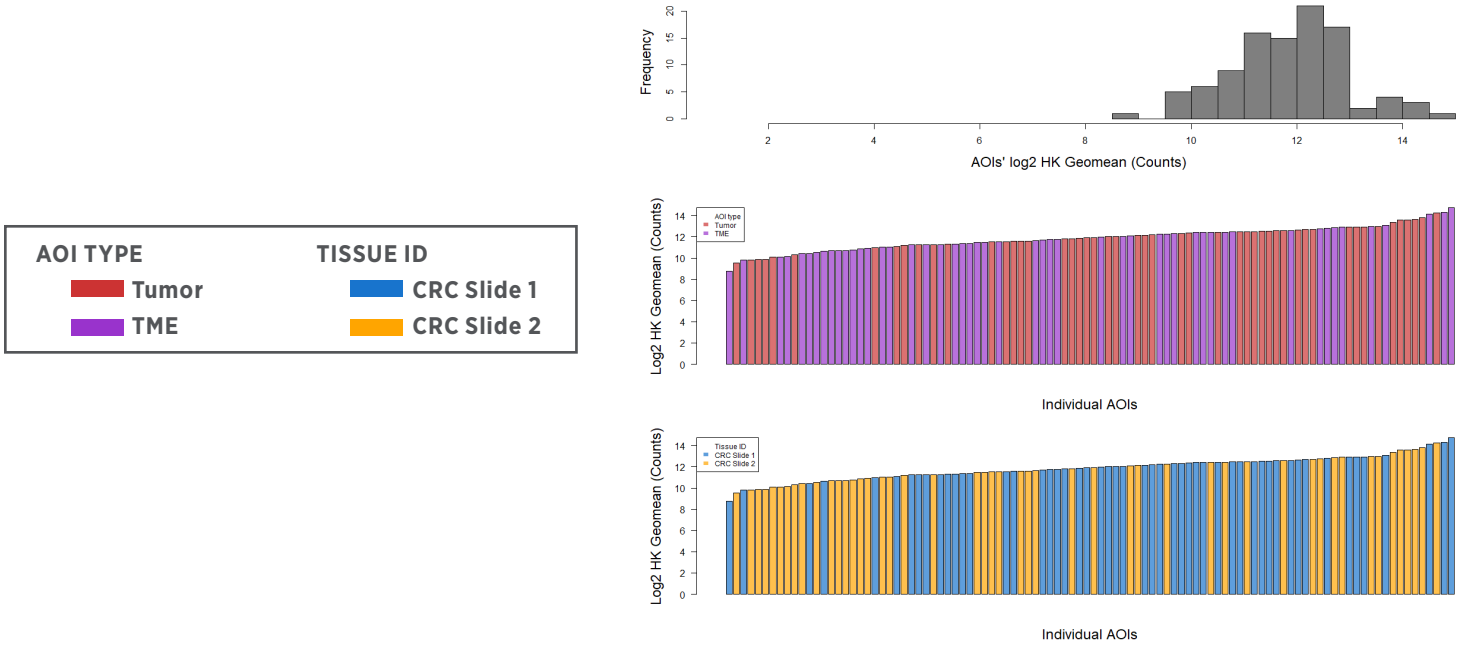


Figure 1

Now let's perform the identical analysis for background, as measured by IgGs.

We do see some preferential clustering by Tissue ID **(Figure 2**). Keep this in mind after normalization.

Now let's take a look for any discrepancies in signal and background **(Figure 3)**:

We see no points with dramatic departures from the line of best fit (i.e. points with unusual signal/background ratios). Based on this and the previous plots, we can use all AOIs in this study.

(Note: typically signal (HK) and background (IgG) are concordant with each other).

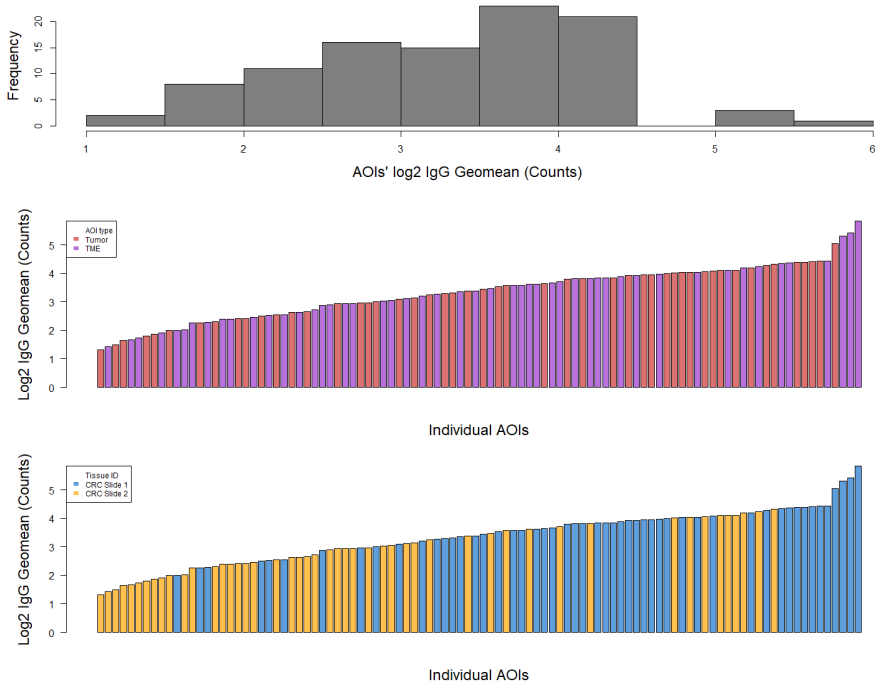


Figure 2

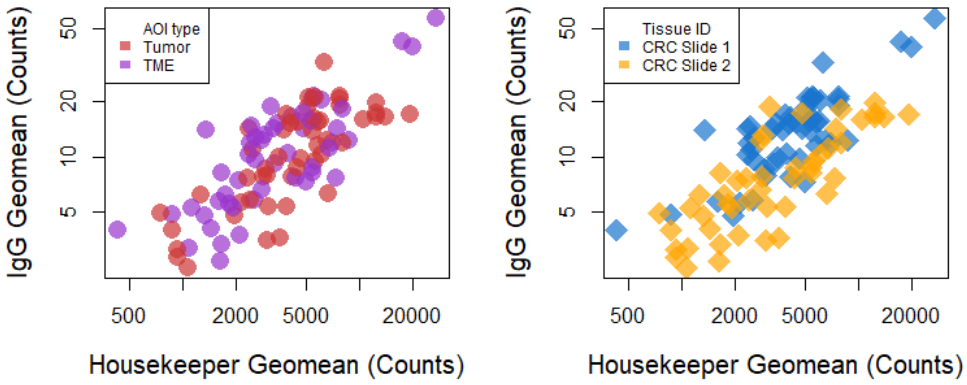


Figure 3

Probe QC

Probes who signals never rise above background should be interpreted carefully. The plot below shows a convenient way to identify poorly-performing probes. We first compute the "signal-to-background" ratio per target, which is simply each AOI's data divided by its IgG geomean **(Figure 4)**.

We should interpret the figure by target. For example: ER-alpha* has above-IgG counts (log signal-to-background > 0), but the fact that it is lower than Ms IgG1** and shows such a limited expression range could mean its signal is all in the background. This target should be interpreted carefully and/or potentially excluded from the study.

Note: within a given target, it is valuable to assess the distribution of AOIs to determine if certain AOIs are above signal while others are not; such differences can be the result of differing AOI type and/or Tissue ID and be of biological importance.

Normalization

This dataset presents four choices: normalization by housekeepers, IgG, area, or nuclei.

A hybrid approach of background-subtraction followed by housekeeper normalization is also possible, but it is not recommended unless the signal and background are highly divergent from each other with very poor correlation over a wide range of values.

Before choosing a normalization method, we must QC the potential normalization factors.

The theory is simple: if multiple probes all accurately measure signal strength, they should be highly correlated with each other.

Let's first investigate individual housekeeper counts in depth to learn more. Note the use of logscale in the plots below **(Figure 5-8)**.

There are multiple important findings from the plots **(Figure 5 and 6)**:

- S6 and GAPDH are consistent with each other.
- S6 and GAPDH do not display concerning trends, such as differing housekeeper ratios based on AOI type or Tissue ID.
- Histone H3 behaves divergently when comparing the two tissues (**Figure 6**). It should be removed as a housekeeper if we would like to make comparisons between the two tissues.

Note: The IO Proteome Atlas has five housekeepers, allowing users to drop up to two divergent HK and still have statistical power on the others.

Next, lets investigate background (negative IgGs) in depth. There are important findings from the plots **(Figure 7 and 8):**

- There is only weak correlation between the 3 IgGs.
- IgG counts are universally low (<30). This suggests their log-scale counts are statistically unstable, which explains their poor concordance with each other.

We conclude that negative IgGs should not be used as a normalization factor in this study.

Note that the IO Proteome Atlas includes five IgG targets, which allows for the exclusion of up to two inconsistent IgGs while maintaining statistical power with the remaining three. Although IgG normalization is not the ideal method in this example, it is often preferred for IO Proteome Atlas studies due to higher signal from NGS readout and because IgGs tend to not be confounded by biological factors.

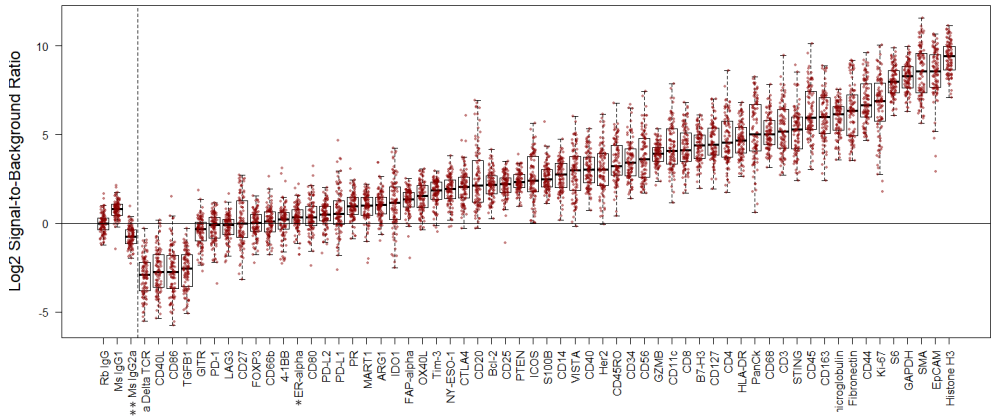
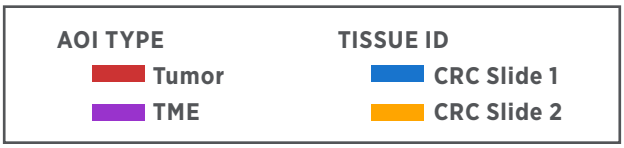


Figure 4



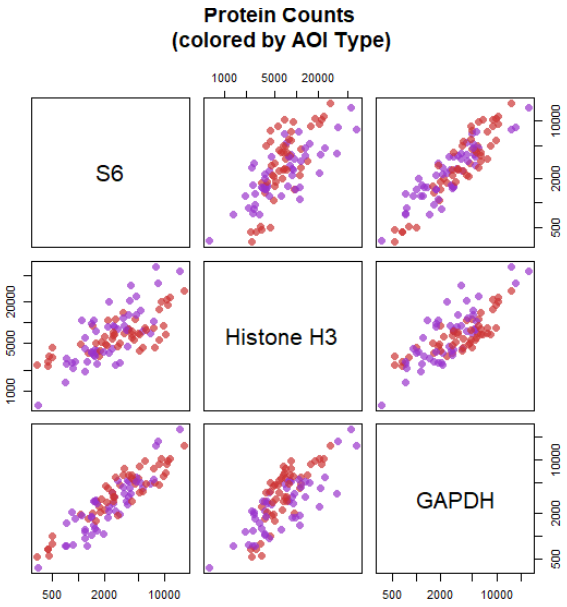


Figure 5

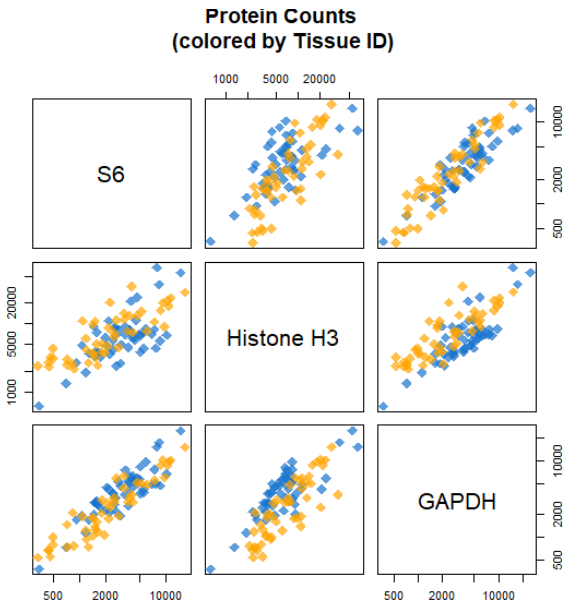


Figure 6

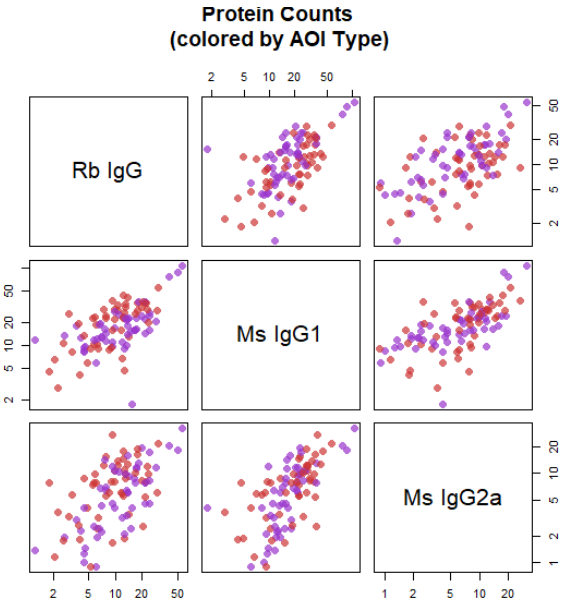


Figure 7

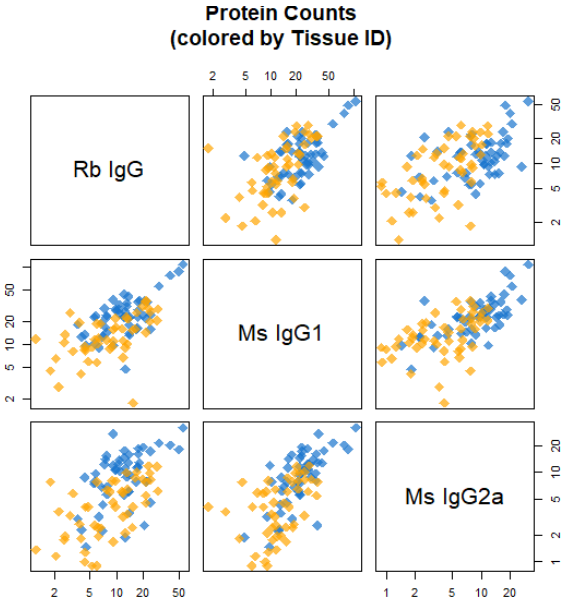


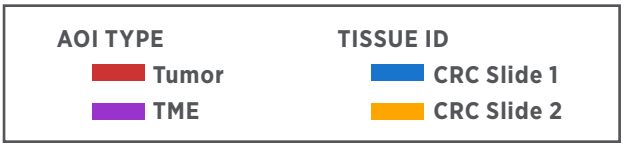
Figure 8

Finally, we look at the concordance of all possible normalization factors: the housekeeper geomean, the IgG geomean, Area, and Nuclei **(Figure 9 and 10)**. Note: as described above, we have removed Histone H3 from the calculation of housekeeper geomean; this step is important to best reflect the normalization factor's performance. We are also careful in how we interpret the IgG geomean.

We observe a mild correlation between signal/background. This is expected given the low IgG counts for this particular experiment.

Area and nuclei are strongly correlated but are only mildly concordant with signal strength as measured by housekeepers.

Given the strong correlation between the selected housekeepers (S6 and GAPDH), the low IgG counts/ concordance, and the weak correlation between area/nuclei vs. HKs, we would prefer HK normalization for this study.



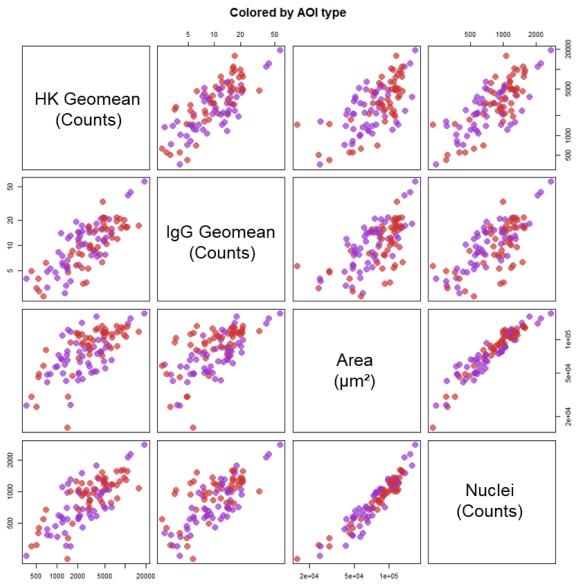


Figure 9

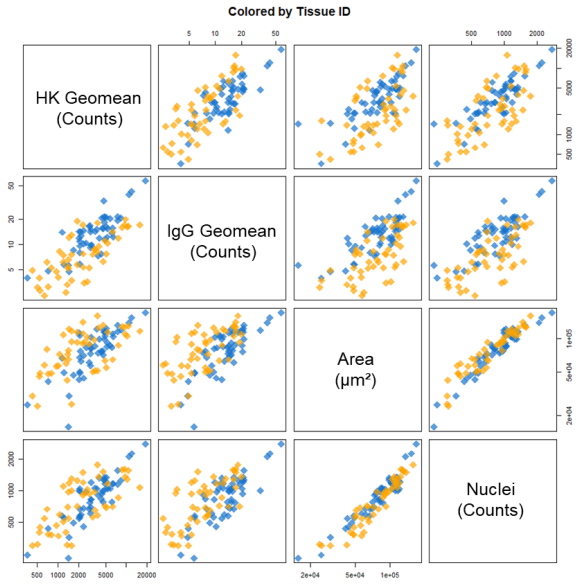


Figure 10

Getting Started with Large Protein Datasets

Analyzing large datasets, such as those generated with the IO Proteome Atlas, can be overwhelming at first. To help navigate your high-plex protein dataset, we recommend the following steps:

- Generate a box plot by target, sorted by ascending maximum value **(Figure 11)**. Targets with counts close to the IgG expression level are likely not detected in the assay. Note that computational steps may take longer with such large datasets.
- Use tools like Principal Component Analysis (PCA) or Uniform Manifold Approximation and Projection (UMAP) to reduce data complexity and visualize key patterns. PCA plots can be generated in the GeoMx DSP Data Analysis Suite without a custom script **(Figure 12)**. To customize PCA plots or generate UMAP or tSNE plots, use the Dimension Reduction custom script. Refer to resources such as the GeoMx DSP Data Analysis User Manual (MAN-10154) or GeomxTools package on Bioconductor.
- When working with samples known to have distinct biological populations, it's important to assess the distribution of target expression between these populations. A broad distribution is expected; however, if you observe tight clustering, this may indicate a technical issue with the assay. A quick way to verify population differences is by conducting a differential expression (DE) analysis **(Figure 13)**. This analysis offers an immediate snapshot of population profiles, such as immune versus non-immune cells, providing a useful preliminary check on the biological distinctions within your dataset. Refer to instructions to set up a DE analysis in the GeoMx DSP Data Analysis User Manual (MAN-10154).

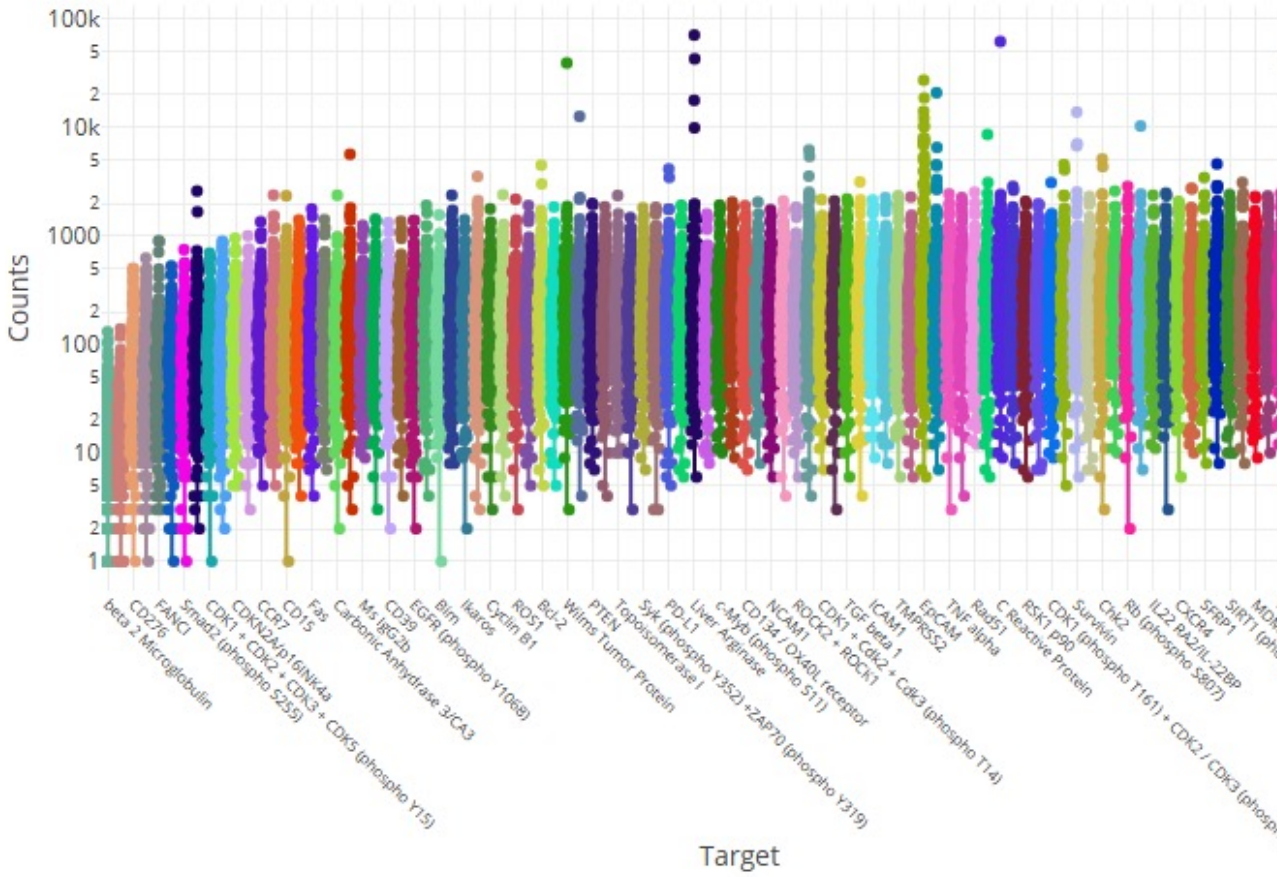


Figure 11

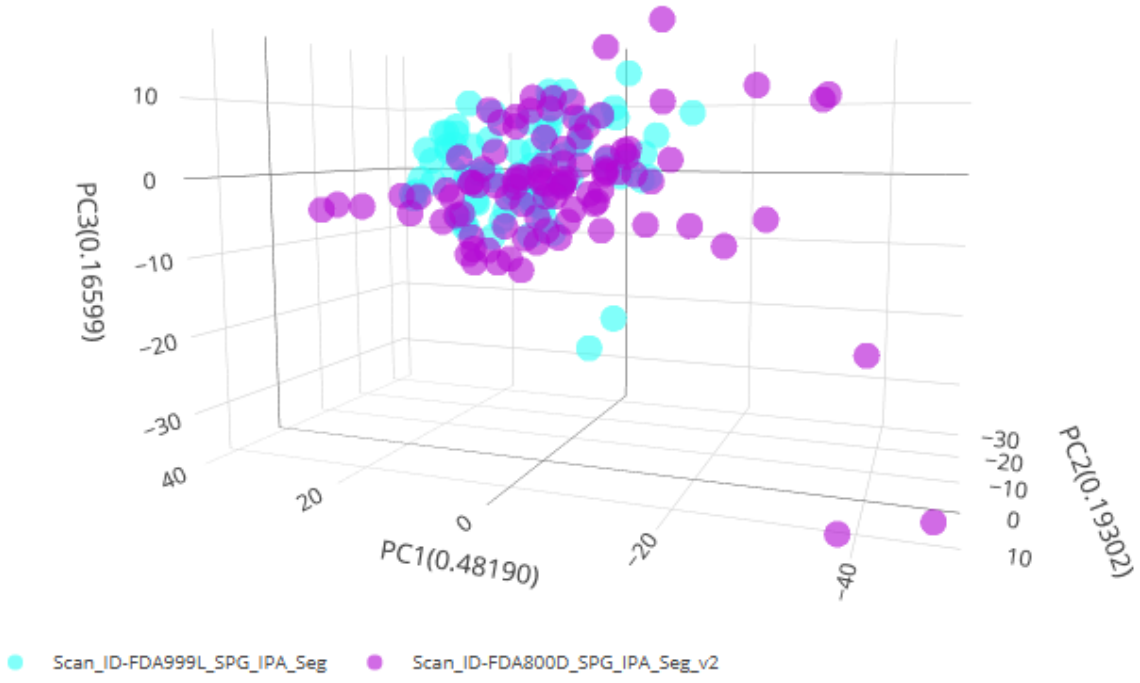


Figure 12

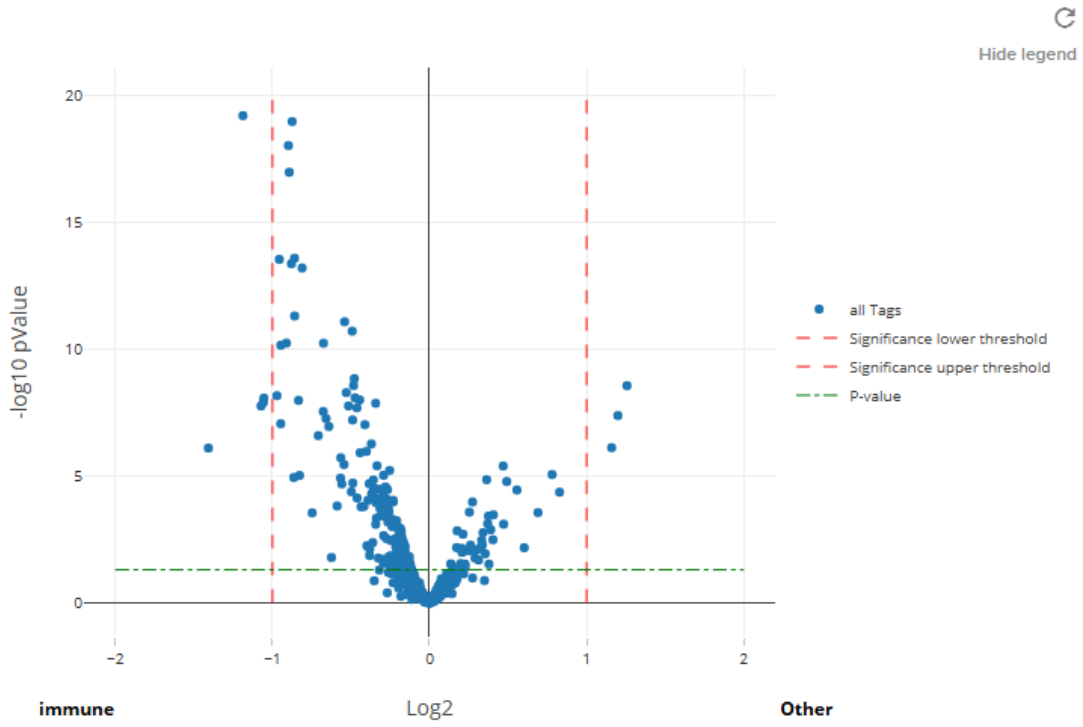


Figure 13

Considerations for Spatial Proteogenomic Datasets (RNA and Protein from one slide)

The entire field of multiomics is facing the challenge of analyzing and visualizing protein and RNA data together. To overcome the different scales of measurement of transcriptomes and proteomes, some researchers are employing a z-score or similar metric to bring the two analytes' datasets to the same scale. Others use overlaid volcano plots to bring the expression scales to the same frame of reference. While multiomic analysis tools are in a state of rapid development, many researchers simply generate box or violin plots, or heatmaps, for each analyte separately and look at them side-by-side. Targets that show a small change in one analyte and a concomitant large change in the other analyte are of particular interest and may lead to refinement or inspiration of research questions.

For additional resources on GeoMx DSP Data Analysis, please refer to documents and trainings in NanoString University.

For more information, please visit brukerspatialbiology.com

FEB 2025 MK6592