Bioinformatics for Immunoprofiling with Protein Microarrays

Introduction

Biomarker discovery has seen rapid growth over the last decade as researchers explore new techniques to identify disease as early as possible. New methods and tools have arisen, enabling high throughput discovery of highly specific and sensitive markers of disease. Immunoprofiling has grown from tools such as flow cytometry and protein microarrays where the milieu of immune cells, cytokines, antibodies, and other immunological markers can provide very specific, unique patient information.

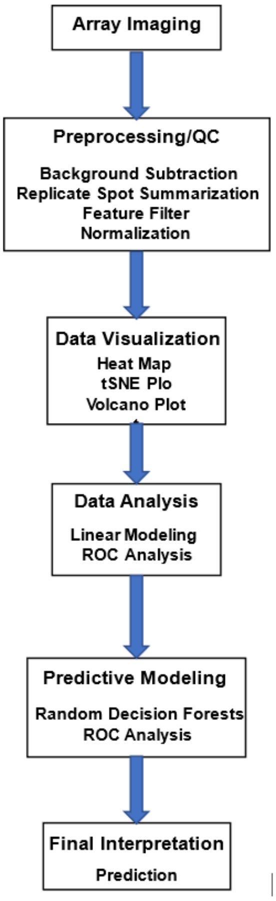

Figure 1. Protein Microarray Data Processing Workflow

All array data is computed in a similar manner. After the assay is run, the array is scanned using a microarray scanner. Protein microarrays often utilize indirect immunofluorescence to detect signal. Images are acquired, background is subtracted, and the data are normalized. Data are visualized and then examined for statistical significance across subjects for each protein. If desired, predictive modeling can be performed to look for sets of antibodies that may be prognostic. Antibodies are excellent predictive biomarkers.

While any biological molecule that can be objectively measured could be a potential biomarker, antibody-based biomarkers offer a unique opportunity to build a functional proteomics-based profile of the humoral immune repertoire, which is not possible through, for example, gene expression data or mass spectrometry-based proteomics. Profiling antibodies allows for both functional profiling (improved understanding of potential disease etiology and/or progression) and clinical applications (predictive, diagnostic biomarker candidates and identifying potential pathogenic or protective antibodies). Discovering novel antibody signatures can be accomplished with protein microarrays. Because protein arrays enable screening against hundreds or thousands of proteins, bioinformatics tools and expertise are necessary for acquiring, processing, analyzing, and interpreting antibody data. The workflow for this type of analysis is illustrated in *Figure 1*. Understanding the nuances of protein microarray data processing helps with data interpretation and the ensuing decisions regarding the next experiment.

Collecting the Data

Protein microarrays are capable of generating thousands of analyzable variables and therefore offer an ideal platform for immune profiling. Protein microarrays are constructed by printing expressed proteins onto glass slides as replicate spots distributed across an x by y array (*Figure 2*). Patient samples (most commonly serum) are then applied to the array. Once serum antibodies have bound their respective antigens (printed proteins), fluorescently labelled detection antibodies can be used to visualize the reaction. Historically, genomic arrays predate protein arrays and set the standards for array analysis. However, over time, it became evident that protein arrays presented challenges absent in gene arrays. For example, gene arrays utilize hybridization, direct binding of labeled DNA to two complimentary DNA strands, with less opportunity for noise. Protein arrays typically rely on indirect immunofluorescence detection, a technique with the potential for greater noise than in situ hybridization because there are multiple levels of antibody binding required for visualization. Poorly designed protein arrays consisting of unfolded/misfolded, denatured, or linear proteins can exacerbate noise. Importantly, antibodies primarily recognize discontinuous amino acid sequences and charge that form during proper protein folding (Barlow et al., 1986; Muro et al., 1994). Sengenics technology offers correctly folded proteins (*Figures 2, 3*). The use of correctly folded proteins in the array results in higher affinity binding of serum antibodies to biologically relevant epitopes patterned on the surface of the target antigen, producing better signal to noise than other technologies. In fact, this is the main reason other manufacturers struggle with protein arrays: noisy data due to non-specific binding (Tan et al., 1999). Data with better signal to noise requires less preprocessing prior to downstream analysis.

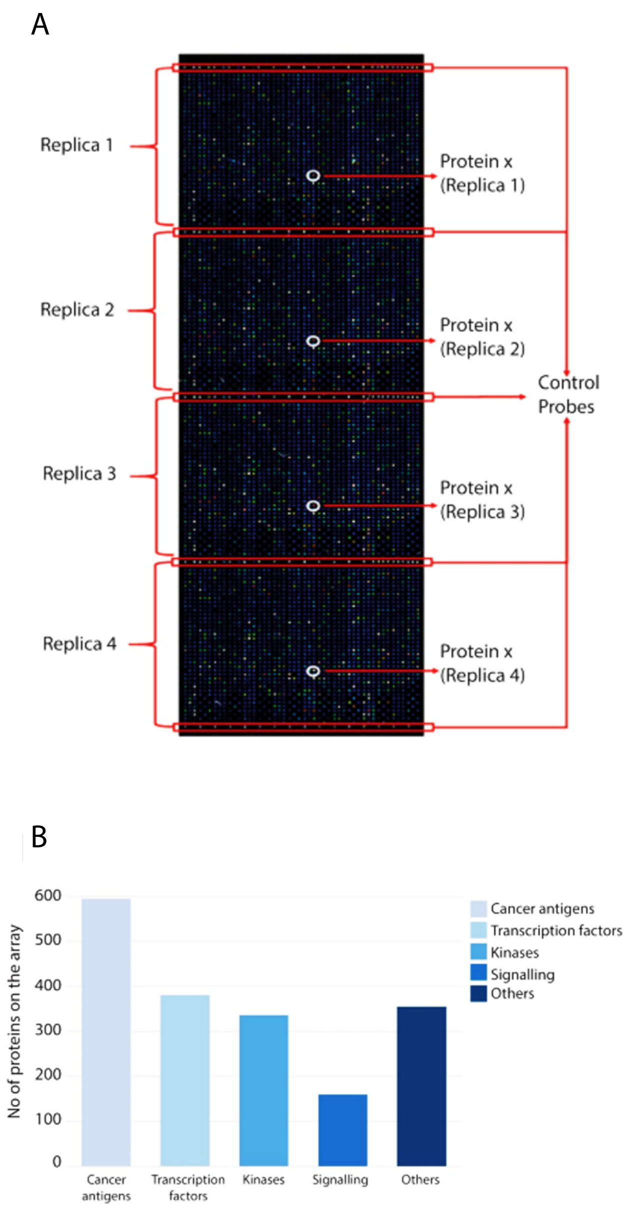

Figure 2. Example of the Sengenics i-Ome Microarray Design

Four replicates of each protein are printed across four quadrants on the slide. A) The slide is divided into four quadrants, each quadrant acting as a replicate array. B) The distribution of proteins on the Sengenics i-Ome microarray slide is representative of nearly 600 cancer antigens as well as antigens involved in normal cellular physiology (Poulsen et al., 2021). Arrays can be customized with the desired type and number of proteins.

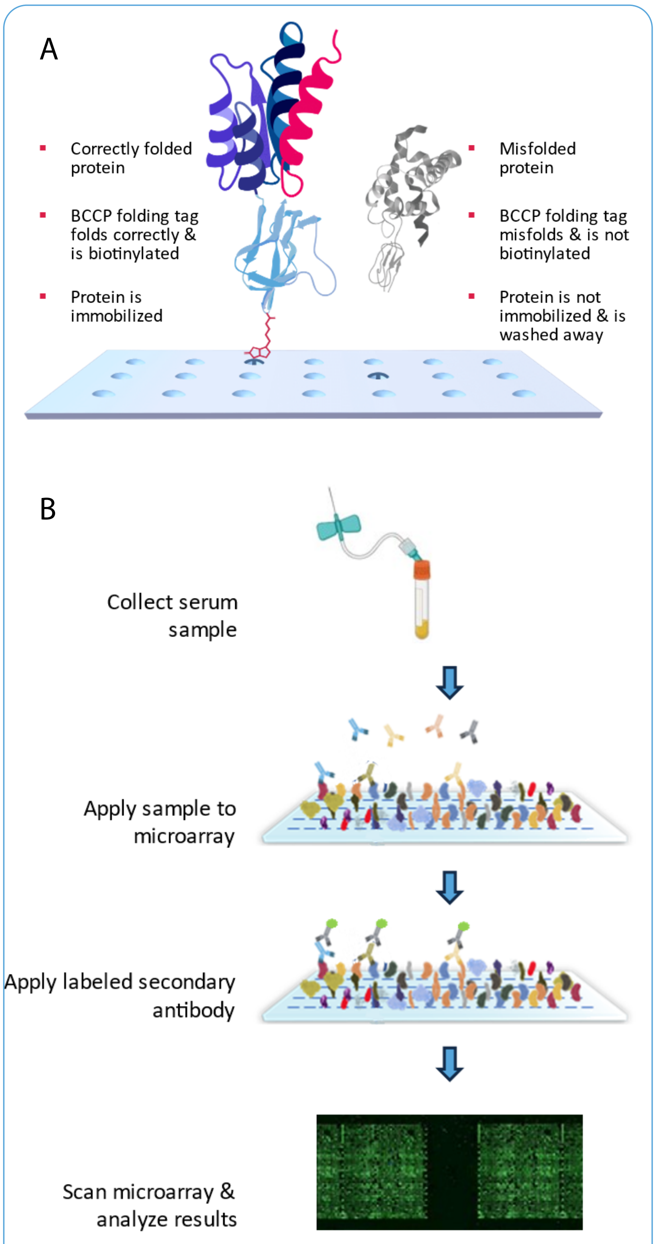
A
Correctly folded protein
BCCP folding tag folds correctly & is biotinylated
Protein is immobilized
Misfolded protein
BCCP folding tag misfolds & is not biotinylated
Protein is not immobilized & is washed away
B
Collect serum sample
Apply sample to microarray
Apply labeled secondary antibody
Scan microarray & analyze results

Figure 3. Sengenics Microarray Assay Workflow

Historically, protein microarrays were not properly designed to effectively capture antibodies that bind to conformational epitopes. Antibody-antigen binding is highly sensitive to antigen shape and not sequence. A. Sengenics technology overcomes this limitation by using full length, properly folded proteins on the microarray. KREX patented protein folding technology enables expression and immobilization of correctly folded, fully functional proteins on a microarray slide. Proteins are recombinantly expressed in frame with a biotin carboxyl carrier protein (BCCP). A misfolded or fragmented protein causes BCCP misfolding, and loss of the biotinylation site and binding to the streptavidin coated microarray slide cannot occur. Incorrectly folded proteins are washed away. This technology maintains conformational epitopes and ensures optimal antibody-epitope binding for the rigors of antibody screening. B. The Sengenics array assay workflow starts with sample application to the microarray followed by detection with a labeled secondary antibody, imaging and data analysis.

The data are visualized in a grid of thousands of spots (also called features) of varying fluorescence, captured via digital imaging. Computer software is used to extract these fluorescent intensities and to catalog them by their physical location on the array. Prior to analysis and interpretation, the data will be evaluated for quality.

Data Preprocessing and Quality Control

Data preprocessing involves examining the data for any overt abnormalities, removing background noise from each signal (by subtracting background from foreground signal), checking the level of precision between replicate spots from a given antigen (coefficient of variation), and performing between-array and between-experiment normalization. High quality microarray data visualization relies on careful handling and image acquisition processes where noise can be exacerbated by poor laboratory techniques.

Images are acquired by a laser microarray slide scanner (*Figure 4*). The signal intensity of the protein spot is called the foreground. The area surrounding the foreground, where no protein is present, is called the background. To improve data quality, it is common practice to calculate the median of the background and subtract this value from the median foreground intensity. This is a well-established filtering method called median filtering (Baxes, 1994) and the resulting net intensities become the variable or feature associated with each protein for each subject (Babu, 2004; Sumera et al., 2020). For example, if a subject's serum contains a higher concentration of p53 antibody, the p53 protein spot on the array will have a higher net intensity. Conversely, if the subject's serum contains very little or no p53 antibody, the net intensity will be lower.

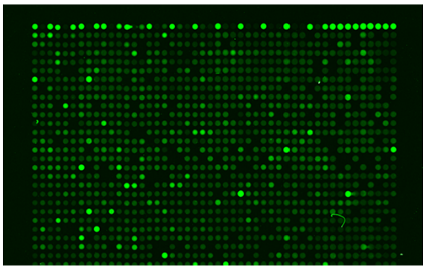

Figure 4. Example Image of Processed Microarray Slide

Example of feature staining from the Sengenics i-Ome technology. Pictured is a section illustrating a partial slide from the i-Ome array. Over 1600 different human proteins are printed on these slides.

The median and net intensities are used for quality control to ensure data accuracy and integrity. If, for example, the slide accidentally dries during processing, the median background intensities will be similar to the median foreground intensities. As a result, the net intensity will be very low or negative. Ideally, the distributions of the median intensities of the foreground and background plotted on the same graph should exhibit sharp peaks with no overlap. Where there is overlap, this suggests a processing error.

Control immunoglobulins are helpful in identifying printing errors or inconsistencies in reagents. Serial dilutions of control immunoglobulins are spotted on the array. The net intensities seen with their subsequent detection should increase linearly with increased concentration (Sumera et al., 2020). The coefficient of determination, or R^2 value, is used to determine linearity, with higher values indicating greater linearity and confirming reliable array printing. Lower R^2 values could indicate potential issues such as pipetting or array printing errors. Sengenics uses a very high R^2 value of 0.90 as a passing metric for quality control.

Lastly, the coefficient of variation (CV), a measure of variability across samples, can be used to determine the technical reproducibility of the overall data. The coefficient of variation is measured for the immunoglobulin dilution, a control probe (Cy3-BSA) and across each feature within the array, thus testing the quality of sample preparation and slide processing. The coefficient of variation for each of these measurements is often less than 5% with Sengenics arrays. Many labs set a standard of 30% coefficient of variation or less to pass the data. Features that fail to meet preset tolerances can be flagged in the dataset. Sengenics flags any antigens that have CV>20% and if more than 1% of all antigens show CVs exceeding 20%, the sample is subject to further inspection. Quality control can thus enable the researcher to identify problematic, inconsistent data from the array and make corrections where possible or re-run the sample.

The next step in data preprocessing is data transformation. Data transformation is a process of converting the data to a scale or range of numbers that are comparable and easier to handle and interpret in downstream analyses. Commonly, the data are log transformed, with the $\text{Log}_2 n$ being calculated for each net intensity, generating a data distribution that is closer to a normal distribution, enabling use of downstream parametric statistical methods.

The final step in pre-processing is normalization. The net intensity can vary from run to run and sample to sample for numerous reasons such as printing effects, lot differences in reagents, sample preparation differences, etc. (Mowoe et al., 2022; Smyth & Speed, 2003). Normalization standardizes the data across slides and experiments so that the data can be compared quantitatively without bias. The simplest form of normalization is to calculate fold change from a reference, a type of percent change. It is difficult to create a reference standard in protein arrays, so instead, each feature is compared with every other feature to produce a reference curve that simulates a standard. Features are moved towards the reference curve, either adding or subtracting intensity values. Specific antibody signal is retained, while the overall net intensity differences across arrays are reduced. In one method that is commonly applied to microarrays, the data are normalized (between samples) using a technique called Locally Estimated Scatterplot Smoothing or LoESS Normalization (Ballman et al., 2004; Liu et al., 2019; Smyth & Speed, 2003). LoESS uses math derived from both linear and nonlinear regression and is sometimes referred to as locally weighted polynomial regression (Figueira, 2020). This technique is a type of internal normalization whereby the data within the array or pairs of arrays are used for the normalization. A scatterplot of net intensity averages is plotted against the difference in net intensities for the same feature. A curve is fit to the scatterplot and becomes the factor for smoothing the data. The data are corrected for array-to-array differences while preserving actual biological differences. Sengenics processes arrays with LoESS normalization; however, it is important to note that LoESS cannot process negative intensity values. A common pitfall in proteomics datasets is missing values. Sengenics imputes missing or negative features by applying the mean minus two standard deviations to negative values. This correction results in positive numbers so LoESS can be applied and essentially re-introduces missing values to the dataset to improve normalization. LoESS normalization has been a very popular method for normalizing microarray data since 2003 (Liu et al., 2019; Smyth & Speed, 2003; Ting et al., 2009).

Background subtraction, measures of quality control and data normalization as outlined above are common practice with microarray data. While there are different approaches, all these elements are part of the pre-processing pipeline. There are multiple accepted techniques that can be applied to achieve the same goals. Nonetheless, all end users engage procedures to enhance the signal (background subtraction, normalization) and insure reliability (quality control) (*Figure 1*).

Examining the Data

Following data preprocessing and normalization, the next step is exploratory analysis via heatmaps and ordinations (*Figure 5*). Heat maps are plots in which the log2 of the net intensities, or transformed data, for specific antigens and specific samples are binned according to value, each bin representing a small range of contiguous values in ascending order from 0 to max (*Figure 6*). The data are pseudo-colored. High and low values are represented by different colors. A heat map is a visualization of the raw data, and for genomics data, has been very helpful in quickly identifying gene expression changes. For proteomics data, the heat map may be more subtle, and clear differences may sometimes be difficult to discern by visual inspection, especially across different arrays.

Heat maps colorize raw data for easy visualization. However, the heat map is linear in that all the values of all the conditions of all the samples are displayed in one flat picture as single units. An array with 1600 proteins and two sample sets of 120 subjects each will have a heatmap with 384,000 points! Relationships can be hard to locate. Most often researchers will only plot a subset of the data. Deciding what to plot, and where the most valuable data lie is assisted with ordination, a statistical procedure designed to re-plot multidimensional data in 2-dimensional space. The idea is to simplify the visualization of the data without losing any of the relationships within the data. A popular ordination technique is the t-distributed

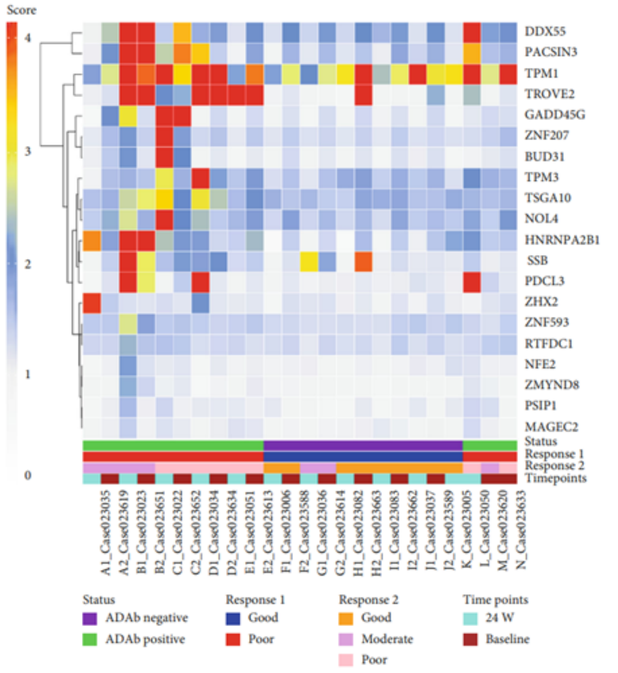

Figure 5. Example of a Protein Heatmap After Data Processing

Heat map illustrating antibody penetrance fold change among patients receiving adalimumab treatment. Patients can be differentiated by the presence of an adalimumab antibody in their sera, date of observation and response to therapy (x-axis). In this example, TROVE2 was observed in poor responding patients at baseline, before treatment started. The data indicate that antibodies such as TROVE2 and TPM1 may be valuable biomarkers for predicting patient outcomes to adalimumab therapy (Chen et al., 2021).

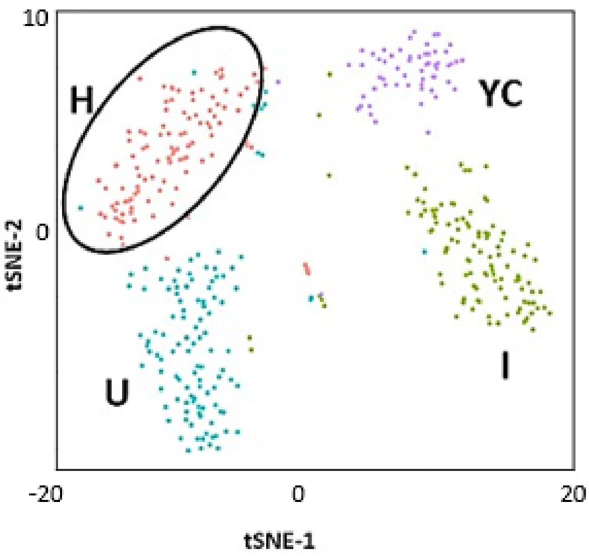

Figure 6. tSNE illustrates Unique Antibody Biomarkers Among Age Groups

tSNE can be used to graphically visualize different antibody biomarkers among subjects of different age groups. Antibodies that commonly occur among elderly patients are clustered into unique groups following tSNE application (unpublished data).

stochastic neighbor embedding or tSNE technique invented in 2008 by Laurens van der Maaten and Gregory Hinton (van der Maaten & Hinton, 2008). The value of this ordination technique is that it accurately clusters related multivariate data in two-dimensional space (*Figure 6*). In terms of proteomics, tSNE can help the researcher visualize related feature data in a simple scatter plot. For example, if you have data from healthy controls and rheumatoid arthritis patients, the tSNE will identify and segregate related feature data, if there are any. The value of tSNE plots lies in representing similarities amongst high dimensional data sets. The information obtained from tSNE plots can be used to inform further statistical analyses. This type of visualization is new compared to heatmaps and scatter plots and is gaining popularity in proteomics. Other ordination techniques are available. These are visualization and graphing tools, the data are not processed or changed.

To quickly identify important features in the population, the data can be visualized on a volcano plot (*Figure 7*) (Cui & Churchill, 2003). To express the data clearly, making it easy to visualize both fold change in feature and significance, the p-values are transformed into the negative Log_{10} of the p-value such that higher numbers on the Y-axis indicate greater significance. Features with high significance and high fold change are considered hits, typically a p-value <0.01 or <0.05 and a fold-change >1.5 or 2 depending on the experimental set up. The statistics applied to protein array data continues to evolve providing greater power and greater sensitivity for better pipelines in drug development and greater success in personalized medicine.

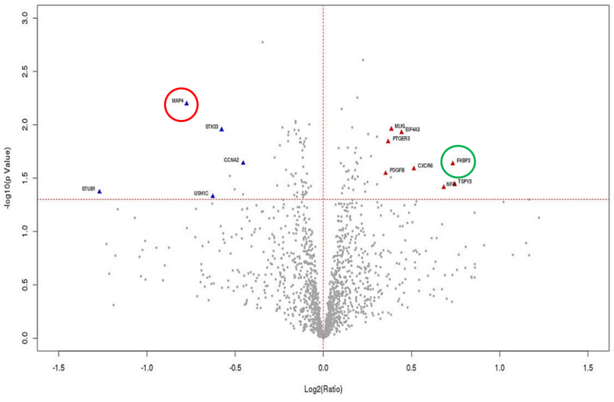

Figure 7. Volcano Plot Enables Quick Visualization of Significant Data

A Volcano Plot of data obtained from Parkinson's patients with or without *Helicobacter pylori*. The plot shows significant changes, both positive and negative, in the Penetrance Fold Change of 13 different antibodies. Microtubule associated protein 4 (MAP4) antibodies, for example, are significantly decreased in *Helicobacter pylori* infected Parkinson's patients (red circle) while FK506 Binding Protein 3 (FKBP3) antibodies, a protein folding chaperone, are significantly increased (green circle) when compared with non-*Helicobacter pylori* infected Parkinson's patients (Suwarnalata et al., 2016).

Analyzing the data

To focus on productive proteomics data, especially when analyzing antibody data, it is helpful to remove or ignore features with little information. A technique developed by Sengenics called Negative Control Filtering identifies a baseline low level net intensity for the array using control proteins that, by design, are not meant to react with sera. Features with intensity values that correlate closely with the negative controls across the samples possess no meaningful information and are not included in downstream analyses. Negative control filtering ensures that true signals are evaluated for statistical significance. In short, every Sengenics array contains positive (Cy3-BSA) and negative controls (negative control proteins). This is a standard assay design.

Proteomics array data have numerous statistical challenges. Protein microarray data have many unique characteristics that can increase error and reduce power such as missing observations (Ting et al., 2009), disparate intensity levels across samples, post translational modification of proteins, and handling and processing inconsistencies. Consequently, standard t-tests are not sufficient for determining statistical significance. In recent years, statistical methods have been developed to specifically address protein array analysis. Linear modeling is more flexible, easier to conduct and more robust than a standard t-test for proteomics data. Significance testing of the linear model is best accomplished with a moderated t-test (Ting et al., 2009). With proteomics data, the input for statistical analyses begins with the transformed intensity values of each feature, the Log_2 of the net intensity. Individual sample features are compared across samples using an expression ratio where the transformed net intensity values of each feature are divided by the mean transformed net intensities of the control samples, producing an Individual Fold Change for each feature (IFC). The IFC, however, does not consider population penetrance. Penetrance Fold-Change (PFC) incorporates the magnitude of the expression ratio as well as population occurrence into the analyses thereby identifying relevant features within the population. The PFC uses a cut-off threshold of the IFC data ≥2 to identify meaningful features and then a frequency or penetrance cutoff of greater than 10% to identify subjects with those features. PFC can be used to assess both frequency and magnitude of the occurrence of antibody/antigen binding, thus determining how often an antibody is present in the group of interest (Patel et al., 2022; Sumera et al., 2020). This is especially useful for smaller sample sizes when some subjects express a strong phenotype that could be missed by looking at only the IFC.

The moderated t-test applied to linear modeling is a powerful statistic for observing significant differences in the differential expression of protein microarray data between groups. It is similar to a regular t-test except in how the error is managed. Rather than using standard error per observation, the moderated t-test uses a pooled estimate for the entire array. This reduces the likelihood that proteins with small variance appear significant while also raising the power by increasing the allowable observations (degrees of freedom) (Ting et al., 2009). Further, a standard t-test assumes all observations are independent, but a moderated t-Test does not. ANOVA and linear regression are sometimes also conducted, depending on the assay design. ANOVA can be used to assess a significant difference across the all the planned comparisons. A post-hoc test can be used to look for specific differences between groups. Linear regression can also be used to model the relationship between variables and potentially predict outcomes.

A powerful emerging attribute of antibody profiling is the ability to use antibody panels versus individual markers to attain higher predictive value (Bizzaro, 2007; Kathrikolly et al., 2022). A panel of multiple markers of disease improves the ability to accurately diagnose and predict the outcome of the disease. It is rare to find a single antibody representative of a disease in all cases. Additionally, since some antibodies are produced to aberrantly expressed or aberrantly modified forms of proteins, pathway linkages identified between biomarkers can add mechanistic insight on underlying disease processes. Identifying a panel, however, requires adequate sample size to accommodate the large number of features and comparisons. Due to the high signal to noise ratio of Sengenics microarrays, variability is low and the recommended number of samples per group is comparatively low because the data are more robust. Sample number is dictated by experimental goals. Employing machine learning to identify patterns of autoantibody expression predictive of disease outcomes or therapeutic response typically requires more samples.

Making Predictions from the data

Protein microarrays are very powerful tools. There are many more proteins than genes. Proteins may be present while genes are downregulated or not expressed. Protein expression reflects present activities within the individual and proteins interact directly with one another so identification of one protein may link to others. Antibodies are a special breed of protein. Since the 1960's, it has been known that antibodies predate autoimmune diseases (Koffler et al., 1971; Kunkel & Tan, 1964; Tan, 1997; Tan & Kunkel, 1966; Tan et al., 1966). In the 1980's, it was determined that rheumatoid factor predated rheumatoid arthritis symptoms by an average of 4 years (Aho et al., 1991; Bizzaro, 2007; del Puente et al., 1988). Consequently, antibodies have great predictive potential, and have been examined in other diseases such as cancer (Patel et al., 2022) and neurological disorders (Wang et al., 2020) where antibody panels continue to demonstrate prognostic value. Running experiments to identify predictive antibodies from an array with 1000's of proteins necessitates predictive statistics and machine learning.

Machine learning with array data is used to make multiple comparisons of feature data against a response. In the case of antibodies, the machine "learns" the association of various antibodies with the response. The output will be a short list of antibodies with the greatest discriminative and/or predictive power from amongst the original array of proteins. For example, the presence of multiple cancer testis antigens is associated with poor prognosis among patients with non-small cell lung cancer (Patel et al., 2022). Usually, machine learning is conducted on a "training" cohort of individuals. There are numerous algorithms to select from, many based on decision trees. The random decision forests test is one of the most popular for array data. The test places the data into randomly organized parallel decision trees, each matrix deducing an answer or answers through a serial decision-making process. Once each tree returns a decision, the final answer is determined based off the most common answer deduced by each tree, i.e., a majority vote. The top several "voted" variables associated with the response are considered highly predictive for that response. This process is known as ensemble learning with the highest-ranking features determined by majority vote.

A significant panel, or subset, is often examined using the Receiver Operating Characteristic (ROC) curve, a data plot that calculates the sensitivity and specificity of the panel. A ROC curve plots the rates of true positives, subjects positive for a panel/classifier who respond, versus true negatives, subjects not positive for the panel/classifier who did not respond. The y-axis plots the observations that were truly positive. This is the sensitivity of the observations. The x-axis plots the observations that were truly false and is the specificity of the observations (*Figure 8*). A quickly rising curve indicates the classifier possesses high sensitivity and high specificity for the response. The Area Under the Curve (AUC) determines the overall strength of the model, higher AUC values indicate a stronger association between the variables, in this case an antibody panel and future outcome or prognosis. An AUC of 1.0 indicates the strongest association between the variables whereas an AUC of 0.5 indicates a random model with no predictive value. ROC analysis is conducted on both the training and validation cohorts. If the antibody panel (classifier) is suitably strong, then the AUCs, sensitivities and specificities of the validation cohort will be very similar to those of the training cohort. This helps confirm the linear modeling data identifying a predictive biomarker panel of antibodies.

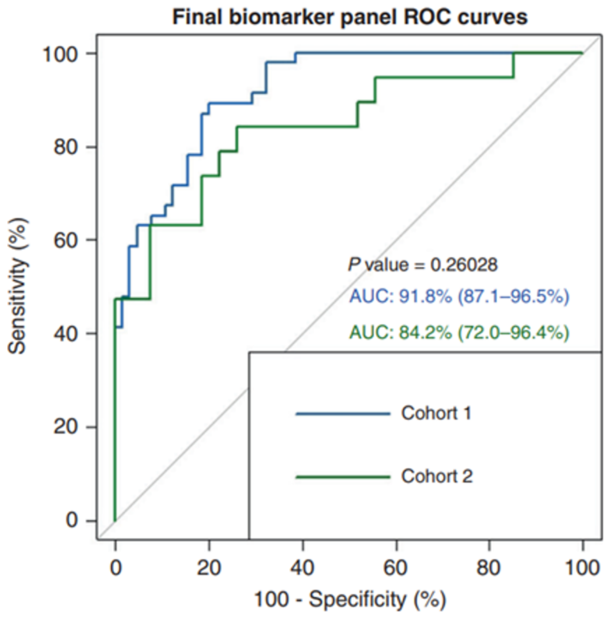

Figure 8. Receiver Operating Characteristic Analysis on Antibodies Predictive of Patient Prognosis

ROC curves for an antibody biomarker panel predictive of non-small cell lung cancer prognosis in both training and validation cohorts. This panel returned 13 novel antibodies with a high strength of reliability (AUC) in both cohorts. P value indicates no significant difference in the performance of this model between cohorts. AUC 95% confidence intervals are displayed within brackets (Patel et al., 2022).

Concluding Remarks

From the 1990's, as high throughput technologies developed, the amount of data to be analyzed became challenging to manage and interpret. Microarrays are capable of providing thousands of data points across subjects for gene expression, protein expression, and RNA expression, resulting in thousands of comparisons. While genomic data led the way with new visualization and statistical models, evaluation of protein array data did not always fit the genomic models. Especially for proteomics, the quality of the data must be high, or visualization, statistical analyses and interpretation suffer. Consequently, exemplary lab techniques with careful consideration of each step in the analysis process must be implemented. Sengenics technology has been developed with each step in mind. The microarrays focus on capturing antibody signatures. Antibodies are proven prognosticators that are directly related to the disease state. The proteins on Sengenics arrays are fully folded and functional, retaining shape for accurate and specific binding to sera antibodies thus resulting in high signal to noise on the array. As a result, pre- and downstream processing of the data is more reliable. These steps control the variables that have previously challenged protein micro array data interpretation. With quality processing and modern statistical analysis, this new generation of antibody profiling will accelerate biomarker discovery, and precision medicine.

References

1. Aho, K., Heliovaara, M., Maatela, J., Tuomi, T., & Palosuo, T. (1991). Rheumatoid factors antedating clinical rheumatoid arthritis. *J Rheumatol*, 18(9), 1282-1284. https://www.ncbi.nlm.nih.gov/pubmed/1757925
2. Aziz, F., & Blackburn, J. (2018). Autoantibody-Based Diagnostic Biomarkers:Technological Approaches to Discovery and Validation. In W. A. Khan (Ed.), *Autoantibodies and Cytokines* (pp. 159-188). IntechOpen. https://doi.org/10.5772/intechopen.75200
3. Babu, M. M. (2004). Introduction to Microarray Data Analysis. In R. Grant (Ed.), *Computational Genomics* - Theory and Application (pp. 225-249). Horizon Bioscience. https://www.mrc-lmb.cam.ac.uk/genomes/madanm/microarray/chapter-final.pdf
4. Ballman, K. V., Grill, D. E., Oberg, A. L., & Therneau, T. M. (2004). Faster cyclic loess: normalizing RNA arrays via linear models. *Bioinformatics*, 20(16), 2778-2786. https://doi.org/10.1093/bioinformatics/bth327
5. Barlow, D. J., Edwards, M. S., & Thornton, J. M. (1986). Continuous and discontinuous protein antigenic determinants. *Nature*, 322(6081), 747-748. https://doi.org/10.1038/322747a0
6. Baxes, G. (1994). *Digital Image Processing* (1st ed.). John Wiley and Sons, Inc.
7. Bizzaro, N. (2007). Autoantibodies as predictors of disease: the clinical and experimental evidence. *Autoimmun Rev*, 6(6), 325-333. https://doi.org/10.1016/j.autrev.2007.01.006
8. Chen, P. K., Lan, J. L., Chen, Y. M., Chen, H. H., Chang, S. H., Chung, C. M., Rutt, N. H., Tan, T. M., Mamat, R. N. R., Anuar, N. D., Blackburn, J. M., & Chen, D. Y. (2021). Anti-TROVE2 Antibody Determined by Immune-Related Array May Serve as a Predictive Marker for Adalimumab Immunogenicity and Effectiveness in RA. *J Immunol Res*, 2021, 6656121. https://doi.org/10.1155/2021/6656121
9. Cui, X., & Churchill, G. A. (2003). Statistical tests for differential expression in cDNA microarray experiments. *Genome Biol*, 4(4), 210. https://doi.org/10.1186/gb-2003-4-4-210
10. Damoiseaux, J., Andrade, L. E., Fritzler, M. J., & Shoenfeld, Y. (2015). Autoantibodies 2015: From diagnostic biomarkers toward prediction, prognosis and prevention. *Autoimmun Rev*, 14(6), 555-563. https://doi.org/10.1016/j.autrev.2015.01.017
11. del Puente, A., Knowler, W. C., Pettitt, D. J., & Bennett, P. H. (1988). The incidence of rheumatoid arthritis is predicted by rheumatoid factor titer in a longitudinal population study. *Arthritis Rheum*, 31(10), 1239-1244. https://doi.org/10.1002/art.1780311004
12. Duarte, J. S., J; Mulder, N; Blackburn, J. (2013). Protein Functional Microarrays: Design, Use and Bioinformatic Analysis in Cancer Biomarker Discovery and Quantitation. In X. Wang (Ed.), *Bioinformatics of Human Proteomics* (pp. 39-74). Springer Science+Business Media Dordrecht.
13. Figueira, J. P. (2020, 07/01/2020). *LOESS*. Tawrds Data Science. Retrieved 01/17/2023 from https://towardsdatascience.com/loess-373d43b03564
14. Kathrikolly, T., Nair, S. N., Mathew, A., Saxena, P. P. U., & Nair, S. (2022). Can serum autoantibodies be a potential early detection biomarker for breast cancer in women? A diagnostic test accuracy review and meta-analysis. *Syst Rev*, 11(1), 215. https://doi.org/10.1186/s13643-022-02088-y
15. Koffler, D., Carr, R., Agnello, V., Thoburn, R., & Kunkel, H. G. (1971). Antibodies to polynucleotides in human sera: antigenic specificity and relation to disease. *J Exp Med*, 134(1), 294-312. https://doi.org/10.1084/jem.134.1.294
16. Kunkel, H. G., & Tan, E. M. (1964). Autoantibodies and Disease. *Adv Immunol*, 27, 351-395. https://doi.org/10.1016/s0065-2776(08)60711-7
17. Liu, X., Li, N., Liu, S., Wang, J., Zhang, N., Zheng, X., Leung, K. S., & Cheng, L. (2019). Normalization Methods for the Analysis of Unbalanced Transcriptome Data: A Review. *Front Bioeng Biotechnol*, 7, 358. https://doi.org/10.3389/fbioe.2019.00358
18. Mowoe, M. O., Garnett, S., Lennard, K., Talbot, J., Townsend, P., Jonas, E., & Blackburn, J. M. (2022). Pro-MAP: a robust pipeline for the pre-processing of single channel protein microarray data. *BMC Bioinformatics*, 23(1), 534. https://doi.org/10.1186/s12859-022-05095-x
19. Muro, Y., Tsai, W. M., Houghten, R., & Tan, E. M. (1994). Synthetic compound peptide simulating antigenicity of conformation-dependent autoepitope. J Biol Chem, 269(28), 18529-18534. https://www.ncbi.nlm.nih.gov/pubmed/7518436

20. Patel, A. J., Tan, T. M., Richter, A. G., Naidu, B., Blackburn, J. M., & Middleton, G. W. (2022). A highly predictive autoantibody-based biomarker panel for prognosis in early-stage NSCLC with potential therapeutic implications. *Br J Cancer*, 126(2), 238-246. https://doi.org/10.1038/s41416-021-01572-x

21. Poulsen, T. B. G., Damgaard, D., Jorgensen, M. M., Senolt, L., Blackburn, J. M., Nielsen, C. H., & Stensballe, A. (2020). Identification of Novel Native Autoantigens in Rheumatoid Arthritis. *Biomedicines*, 8(6). https://doi.org/10.3390/biomedicines8060141

22. Poulsen, T. B. G., Damgaard, D., Jorgensen, M. M., Senolt, L., Blackburn, J. M., Nielsen, C. H., & Stensballe, A. (2021). Identification of potential autoantigens in anti-CCP-positive and anti-CCP-negative rheumatoid arthritis using citrulline-specific protein arrays. *Sci Rep*, 11(1), 17300. https://doi.org/10.1038/s41598-021-96675-z

23. Sexauer, D., Gray, E., & Zaenker, P. (2022). Tumour- associated autoantibodies as prognostic cancer biomarkers- a review. *Autoimmun Rev*, 21(4), 103041. https://doi.org/10.1016/j.autrev.2022.103041

24. Smyth, G. K., & Speed, T. (2003). Normalization of cDNA microarray data. *Methods*, 31(4), 265-273. https://doi.org/10.1016/s1046-2023(03)00155-5

25. Sumera, A., Anuar, N. D., Radhakrishnan, A. K., Ibrahim, H., Rutt, N. H., Ismail, N. H., Tan, T. M., & Baba, A. A. (2020). A Novel Method to Identify Autoantibodies against Putative Target Proteins in Serum from beta-Thalassemia Major: A Pilot Study. *Biomedicines*, 8(5). https://doi.org/10.3390/biomedicines8050097

26. Suwarnalata, G., Tan, A. H., Isa, H., Gudimella, R., Anwar, A., Loke, M. F., Mahadeva, S., Lim, S. Y., & Vadivelu, J. (2016). Augmentation of Autoantibodies by Helicobacter pylori in Parkinson's Disease Patients May Be Linked to Greater Severity. *PLoS One*, 11(4), e0153725. https://doi.org/10.1371/journal.pone.0153725

27. Tan, E. M. (1997). Autoantibodies and autoimmunity: a three-decade perspective. A tribute to Henry G. Kunkel. *Ann N Y Acad Sci*, 815, 1-14. https://doi.org/10.1111/j.1749-6632.1997.tb52040.x

28. Tan, E. M., & Kunkel, H. G. (1966). Characteristics of a soluble nuclear antigen precipitating with sera of patients with systemic lupus erythematosus. *J Immunol*, 96(3), 464-471. https://www.ncbi.nlm.nih.gov/pubmed/5932578

29. Tan, E. M., Schur, P. H., Carr, R. I., & Kunkel, H. G. (1966). Deoxybonucleic acid (DNA) and antibodies to DNA in the serum of patients with systemic lupus erythematosus. *J Clin Invest*, 45(11), 1732-1740. https://doi.org/10.1172/JCI105479

30. Tan, E. M., Smolen, J. S., McDougal, J. S., Butcher, B. T., Conn, D., Dawkins, R., Fritzler, M. J., Gordon, T., Hardin, J. A., Kalden, J. R., Lahita, R. G., Maini, R. N., Rothfield, N. F., Smeenk, R., Takasaki, Y., van Venrooij, W. J., Wiik, A., Wilson, M., & Koziol, J. A. (1999). A critical evaluation of enzyme immunoassays for detection of antinuclear autoantibodies of defined specificities. I. Precision, sensitivity, and specificity. *Arthritis Rheum*, 42(3), 455-464. https://doi.org/10.1002/1529-0131(199904)42:3<455::AID-ANR10>3.0.CO;2-3

31. Ting, L., Cowley, M. J., Hoon, S. L., Guilhaus, M., Raftery, M. J., & Cavicchioli, R. (2009). Normalization and Statistical Analysis of Quantitative Proteomics Data Generated by Metabolic Labeling. *Molecular & Cellular Proteomics*, 8(10), 2227-2242. https://doi.org/10.1074/mcp.m800462-mcp200

32. van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. The Journal of Machine Learning Research, 9(2579-2605), 85.

33. Vantaggiato, L., Cameli, P., Bergantini, L., d'Alessandro, M., Shaba, E., Carleo, A., Di Giuseppe, F., Angelucci, S., Sebastiani, G., Dotta, F., Bini, L., Bargagli, E., & Landi, C. (2022). Serum Proteomic Profile of Asthmatic Patients after Six Months of Benralizumab and Mepolizumab Treatment. *Biomedicines*, 10(4). https://doi.org/10.3390/biomedicines10040761

34. Wang, B. Z., Zailan, F. Z., Wong, B. Y. X., Ng, K. P., & Kandiah, N. (2020). Identification of novel candidate autoantibodies in Alzheimer's disease. *Eur J Neurol*, 27(11), 2292-2296. https://doi.org/10.1111/ene.14290

35. Zaenker, P., & Ziman, M. R. (2013). Serologic autoantibodies as diagnostic cancer biomarkers--a review. *Cancer Epidemiol Biomarkers Prev*, 22(12), 2161-2181. https://doi.org/10.1158/1055-9965.EPI-13-0621

36. Zhang, R., Siu, M. K. Y., Ngan, H. Y. S., & Chan, K. K. L. (2022). Molecular Biomarkers for the Early Detection of Ovarian Cancer. *Int J Mol Sci*, 23(19). https://doi.org/10.3390/ijms231912041