Network Analysis as a Supplementary Tool for Exploratory Data Analysis in Modeling Health Outcomes based on Proteomics Data

Monica Ghaly [*1] , Zachary Dwight [1], Susan Mockus [1]

Precision Biomarker Laboratories, Cedars-Sinai Medical Center, Beverly Hills, CA

Precision Biomarker Laboratories

Cedars Sinai

Introduction

The most difficult aspect of machine learning lies in the exploratory data analysis phase. There are many valid approaches such as using statistical tests and graphical methods. Within the field of proteomics, it is often the goal to find protein candidates that demonstrate a relationship to a disease state. Therefore, it is paramount to develop a protocol for exploratory data analysis that finds the maximum number of potential predictors of disease state. Additionally, there is interest in identifying communities of relevant proteins which may aid in the interpretability of the proteins' impact on the biological state of the subjects.

Methods

1) Exploratory Data Analysis

- An illustration of this procedure has been applied on the public data set, Plasma Proteomic Analysis Based on 4D-DIA Evaluates the Clinical Response to Imrecoxib in the Early Treatment of Osteoarthritis
- The outcome of interest is the response of the subject, 'Response'
- Filtering proteins for 90% completeness
- Creating a matrix of correlations between proteins after filtering for the level(s) of interest in the outcome
- Setting correlation less than the absolute value 0.7 to 0 and setting correlations along the diagonal equal to 0
- Clustering using edge betweenness through which modules are created of densely connected proteins that are sparsely connected to proteins outside their respective modules
- Performing principal component analysis and selecting the first component as a representative for the proteins of each module (Grp variables)
- Performing principal component analysis on single proteins (proteins that were not grouped with others in a module) and choosing 10 principal components to represent that set (PC variables)
- High degree (proteins with many correlations to other proteins) are also identified to be used in the final modeling phase (Figure 4)

2) StatisticalTests

- T-tests, Wilcoxon Rank Sum Tests
- Response ~ Grp+PC

Figures

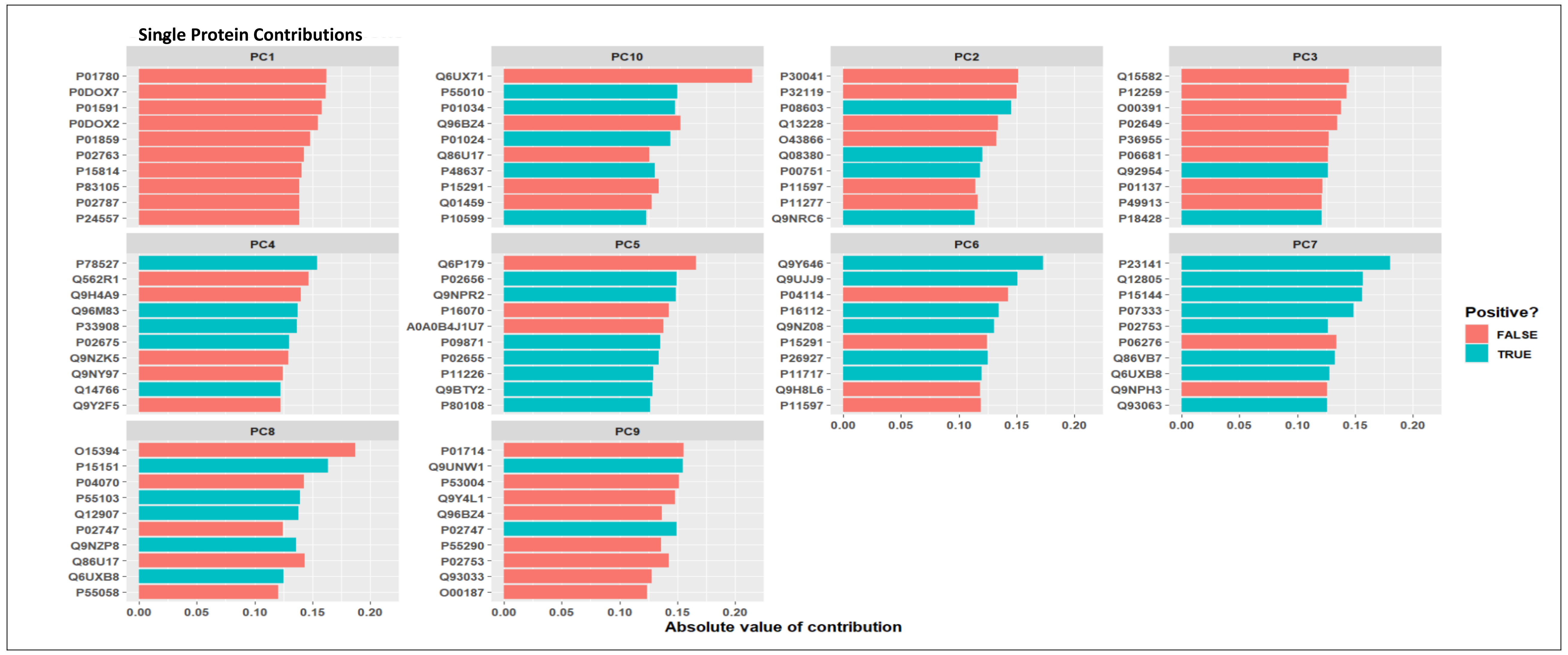

Figure 1 _ Single Protein Contributions
10 principal components representing proteins outside modules

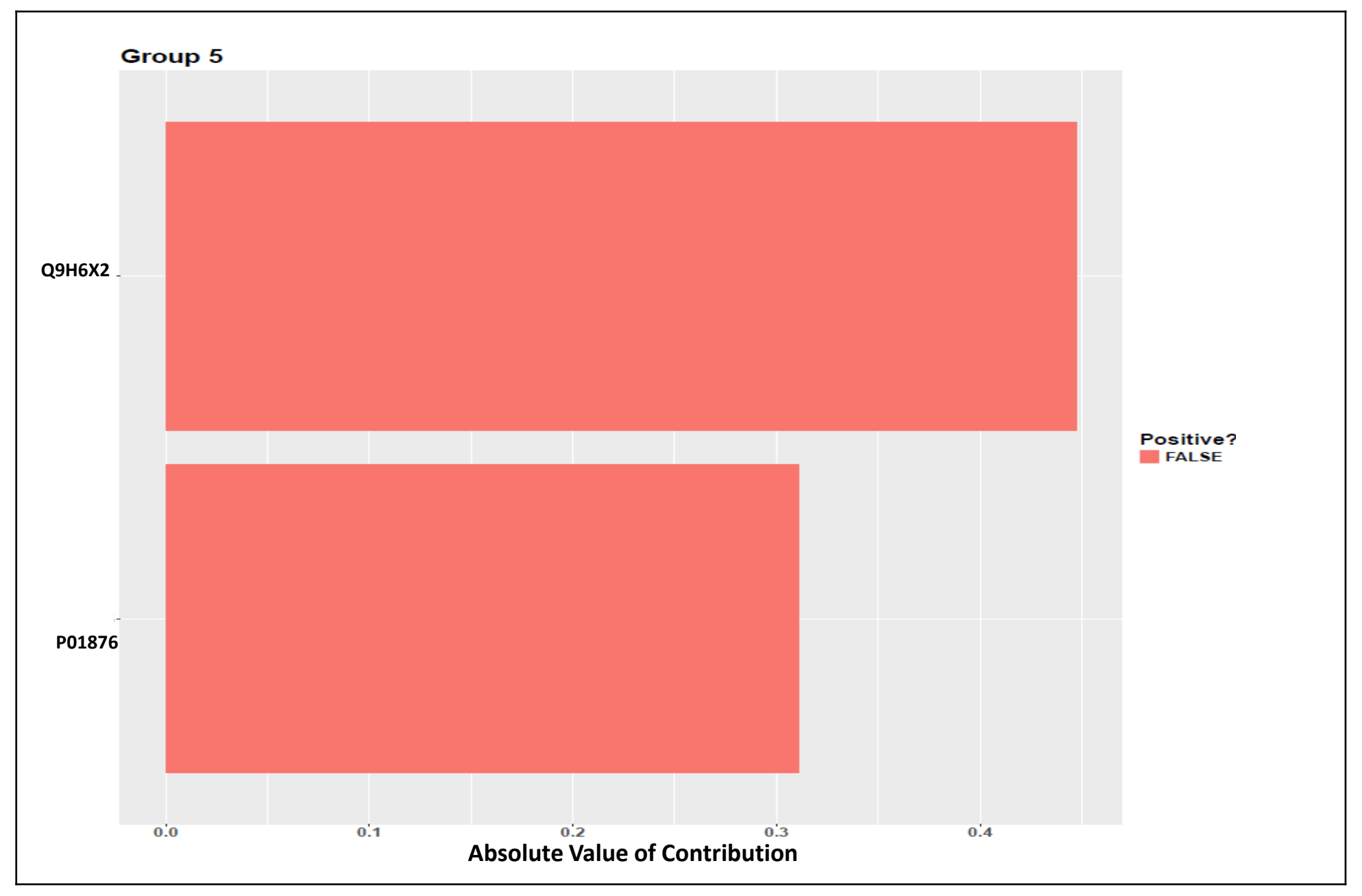

Figure 2_Group 5 contributing proteins
Q9H6X2 and P01876 contributions to the Grp 5 principal component

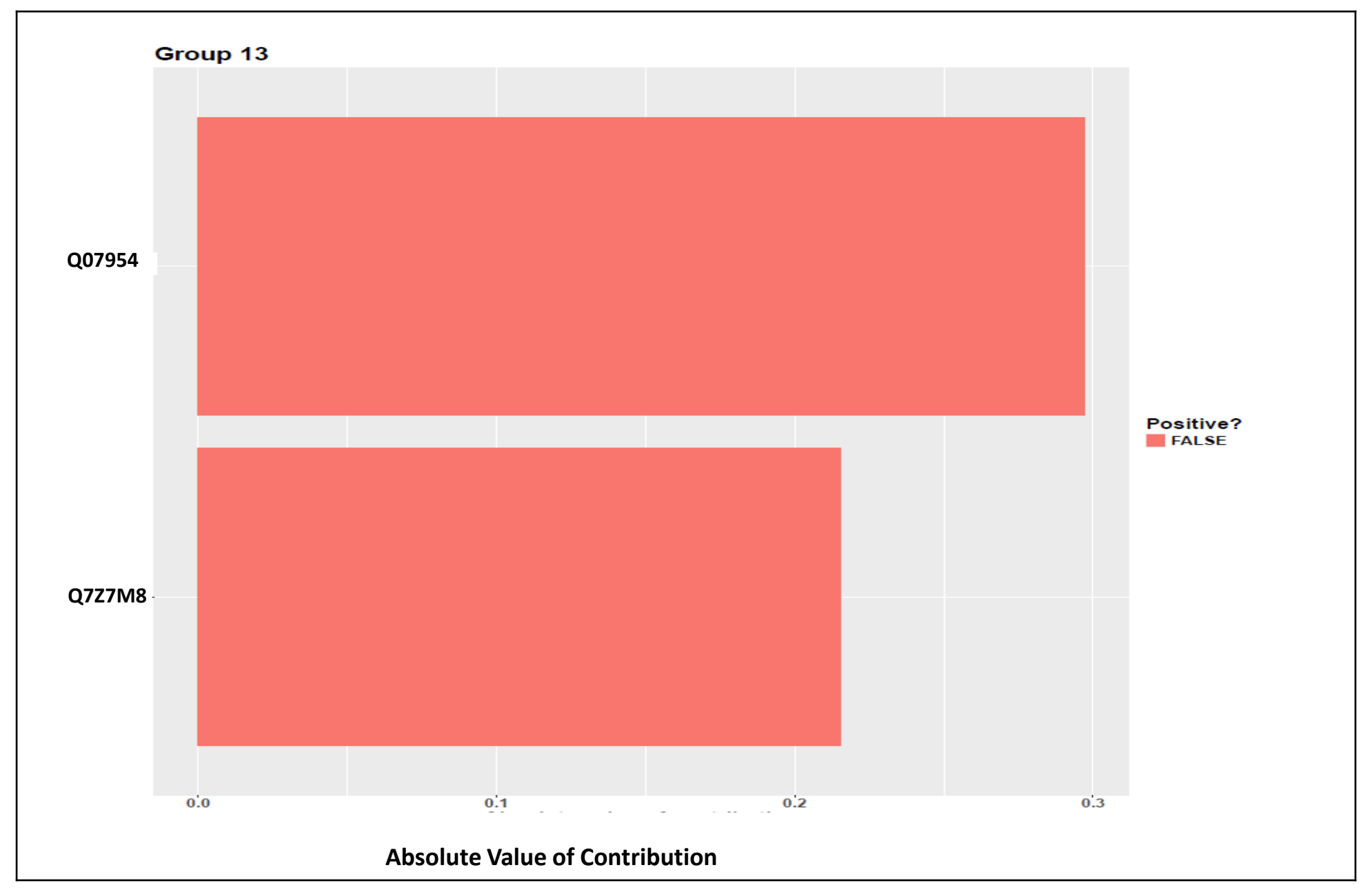

Figure 3 _ Group 13 contributing proteins
Q07954 and Q7Z7M8 contributions to the Grp 13 principal component

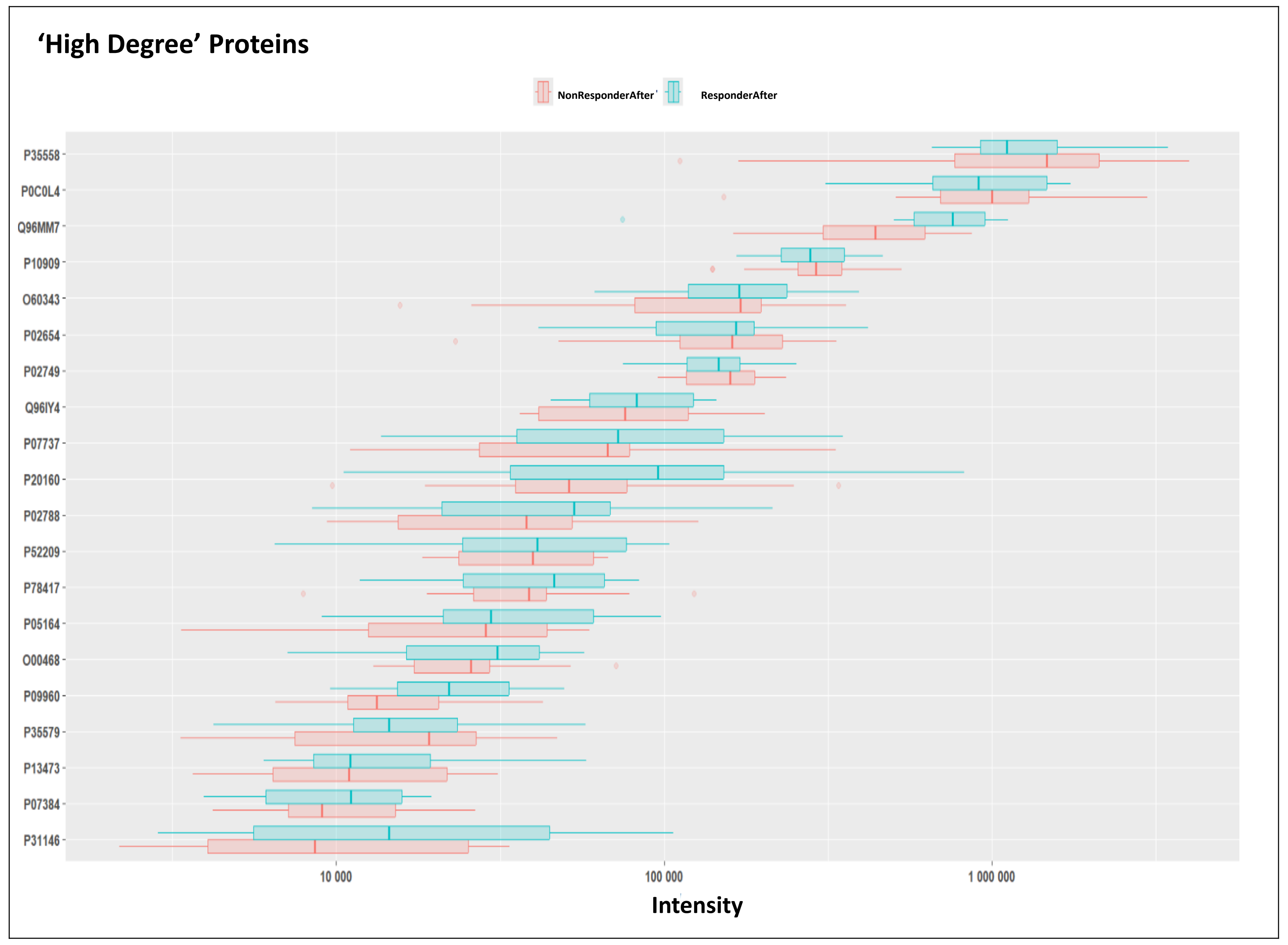

Figure 4_Intensities for 'High Degree' Proteins
Boxplot of log-scaled intensities for proteins with many high correlations with other proteins

Results

- Clustering produces 13 modules, each represented by 1 principal component
- 282 proteins remained outside modules and they are represented by 10 principal components (Figure 1)
- Wilcoxon Rank Sum tests show that PC 1, Grp 5, and Grp 13 have a statistically significant association with the Response (8.9010e-06, 1.3685e-02, and 4.3700e-02, respectively) (Figures 2, and 3)
- Within the framework of this technique, the components that are statistically significant would be broken into its individual protein components that would then be the predictors of the outcome.

Discussion

The technique of using Network Analysis as an antecedent to dimensional reduction allows for the retention of biological interpretation of the principal components. It is also a more efficient way of analyzing and modeling a large number of protein interactions with the outcome of interest. Due to the large number of variables involved in proteomics modeling, this procedure is best supplemented with other more visual techniques. Nonlinear dimensional reduction techniques may also serve to differentiate the components that have the most impact on the models' predictions.

Conclusions

The use of Network Analysis in the exploratory data analysis phase of a statistical investigation as a means of dimensional reduction allows for greater interpretability of protein group interactions in modeling the components as predictors of disease state in a machine learning model with feature importance. Along with proteins identified by statistical tests, and graphical methods, the proteins represented by the most important components widen the scope of the statistical and modeling analyses by exploring protein interactions in a compacts and intuitive manner.

Acknowledgements/Conflict of Interest

Conflict of Interest: **No Conflict of Interest**
Contact Information: **Monica.Ghaly@cshs.org**