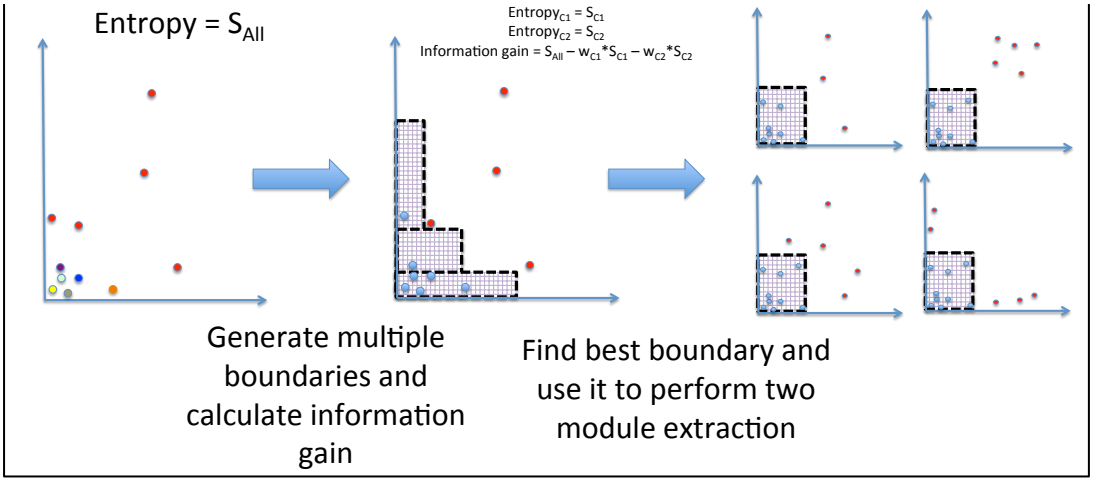


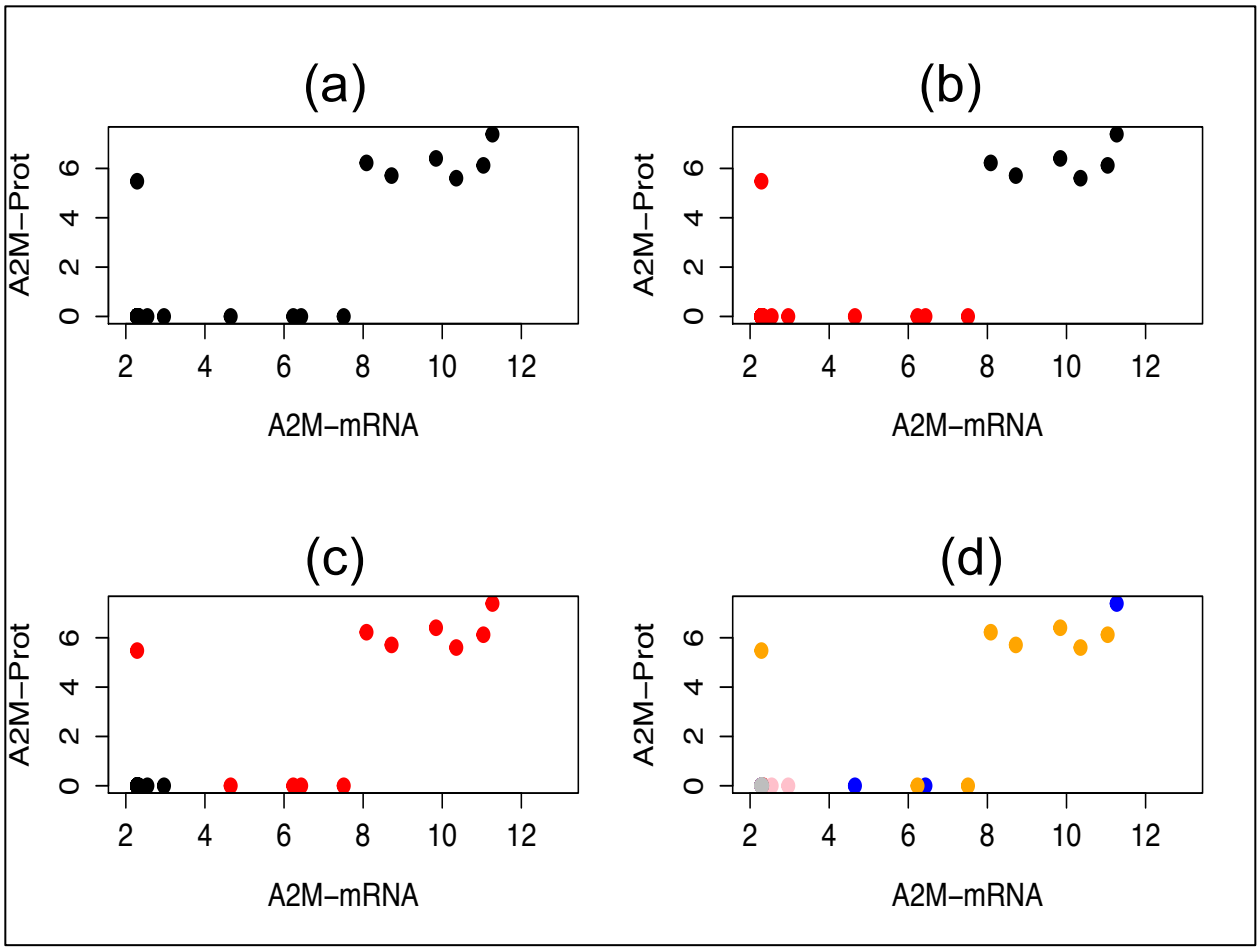
Supplementary Appendix 1 (Two Module Clustering). Workflow for two-module extraction clustering. Utilizes entropy and information gain to decide optimal boundary for separating dense low abundance data points from relatively high abundance data points.

Since each cancer in the NCI-60, on an average has only ~six cell lines in the dataset (ranging from two to nine), generic, unsupervised clustering methods would not be sensitive to isolating scatterplots that visually showcase cancer specific expression. Also, unsupervised clustering methods do not provide a feature that allows us to rank and separate the truly cancer tissue specific mRNA-protein pair from a non-specific one. We address both these issues in our custom clustering formulation

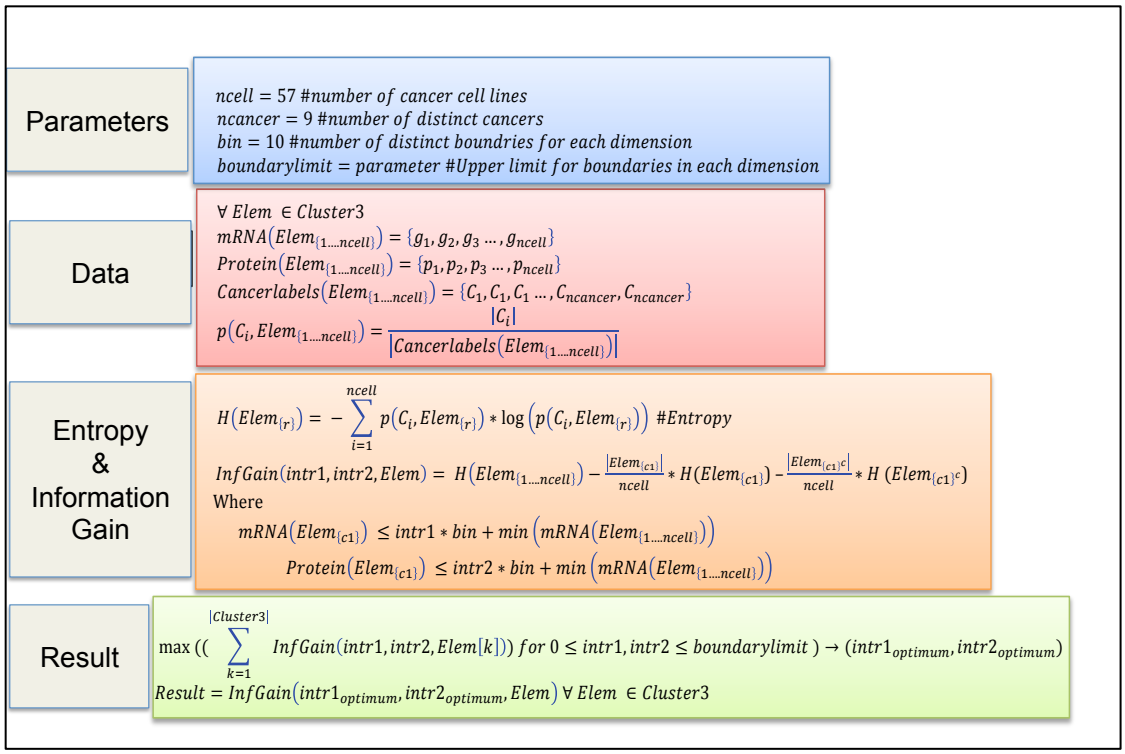
Two-module extraction clustering: By visualizing the third scagnostics feature cluster we see that the scatterplot trends can be characterized by two clusters in each mRNA-protein scatterplot. One of the clusters, is relatively dense, and is concentrated at low expression and abundance. The other is extremely sparse and is spread across the dynamic range of mRNA-protein measurements Especially since our goal with the clustering is to find cancer tissue specificity, the clustering is required to be sensitive to even the slightest deviation from the low expression and abundance. **Supplementary Appendix 1 Figure 1** indicates the type of patterns we will miss if we employ traditional unsupervised clustering to isolate the two clusters in the scatterplots of this module. We utilize the entropy and information gain produced by dividing the data points in each mRNA-protein relationship to quantify cancer tissue specificity of an mRNA-protein pair. The process is outlined above. Across all the mRNA-protein relationships from *scagnostics cluster 3*, we annotate the "low expression and abundance" points crowded around low expression and abundance for a particular relationship. All the points that are not annotated as low expression and low abundance are labeled as members of the "sparse" cluster. We parameterize the boundaries within which a point can be labeled low expression and abundance by distance from the relative origin for each of the 2D mRNA-protein scatterplots. Further, we judge the effectiveness of each boundary and subsequent clustering by calculating the entropy and information gain (using the cancer labels of each cell line) of this particular clustering. We optimize this boundary parameter across all members of the scagnostic cluster 3 and using the resulting optimum boundary parameters, calculate the final information gain for each mRNA-protein relationship.

The information gain feature for these relationships can now be used to rank them in terms of whether the sparse data points, are enriched in specific type of cancer tissue(s). This approach addresses the issues that generic unsupervised clustering methodologies present. Namely, we have an approach that defines a workflow for finding the optimum arbitrarily shaped clusters in a set of data with two obvious clusterings (one dense and one sparse) and allows us to quantify cancer tissue specificity of mRNA-protein pairs. The mathematical formulation of this method is provided in **Supplementary Appendix 1 Equation 1**.

Next we isolated mRNA-protein pairs that show the highest information gain and mapped them to the cancer tissue they indicate specificity for. The two-module extraction was successful in highlighting mRNA-protein pairs that showed high specificity for specific cancer tissues and could potentially be investigated as biomarkers for the corresponding cancers.



Supplementary Appendix 1 Figure 1. (a) A2M mRNA-protein scatterplot without color coding (b) A2M mRNA-protein scatterplot with color coding according to clusters defined by 2-means analysis (c) A2M mRNA-protein scatterplot with color coding according to clusters defined by two module extraction clustering (d) A2M mRNA-protein scatterplot with color coding according cancer labeling in data



Supplementary Appendix 1 Equation 1. Two-module extraction clustering of Cluster 3 members.