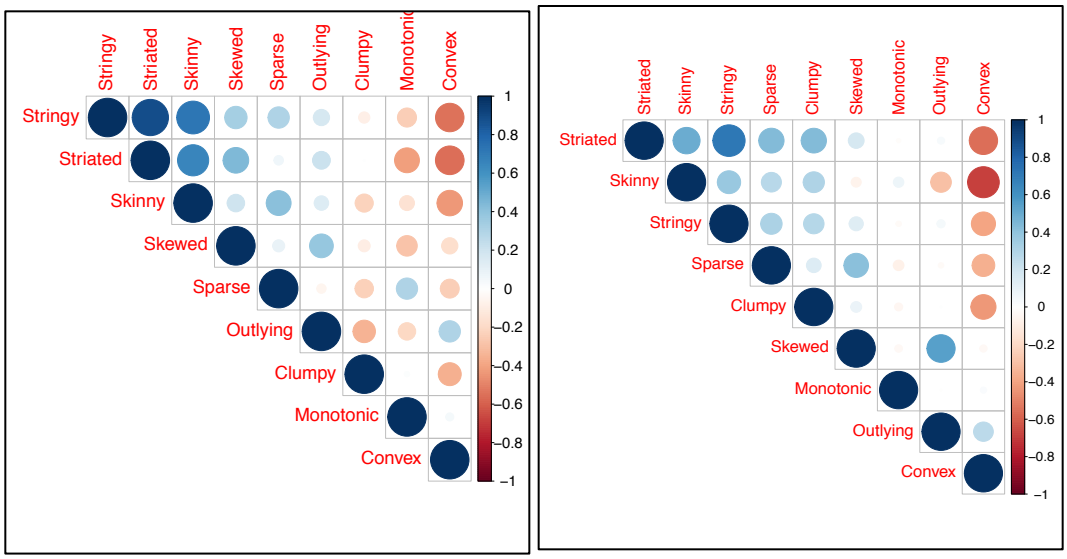
Supplementary Section 1:

Below we qualitatively assess three distinct points of comparison between Scagnostics and traditional methods of exploring proteogenomic data:

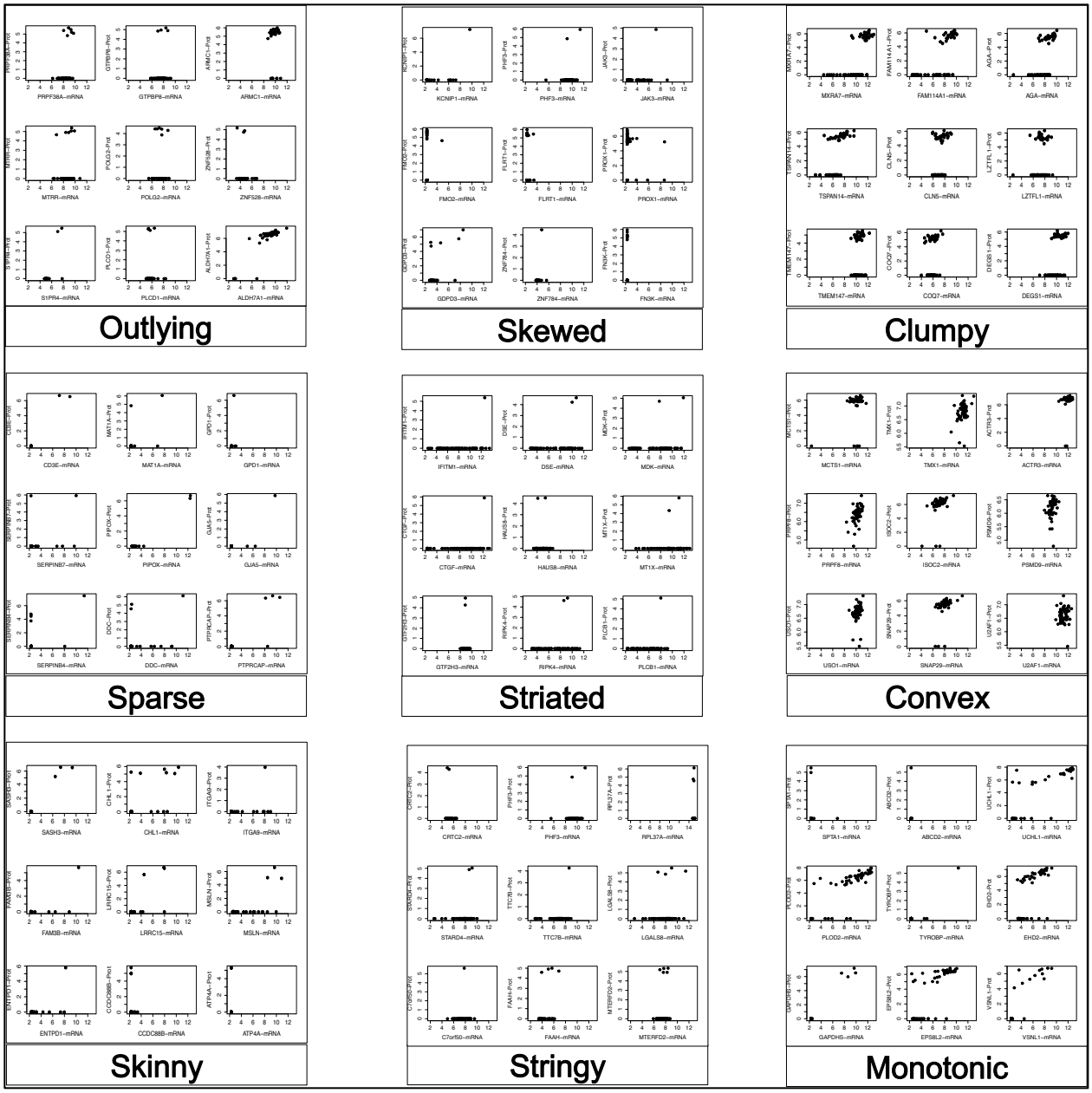
a) Integrative – Scagnostics based methodologies offer an integrative approach that simultaneously considers two data dimensions (transcriptomics and proteomics) to model proteogenomic relationships. Most other methods (bi-clustering, co-expression analysis, consensus clustering) process one dimension at a time and then further analysis may attempt to aggregate the results from each dimension.

b) Efficient relationship characterization – Regardless of the number of samples and number of gene-protein pairs, this methodology identifies a 9 dimensional feature set that presents an interpretable characterization of proteogenomic relationships. All other traditional methods, do not scale well when the number of samples or genes/proteins are increased. A scatterplot is an intuitive way to understand the regulatory relationship between a gene and the corresponding protein, and scagnostics is an efficient tool to characterize these scatterplots.

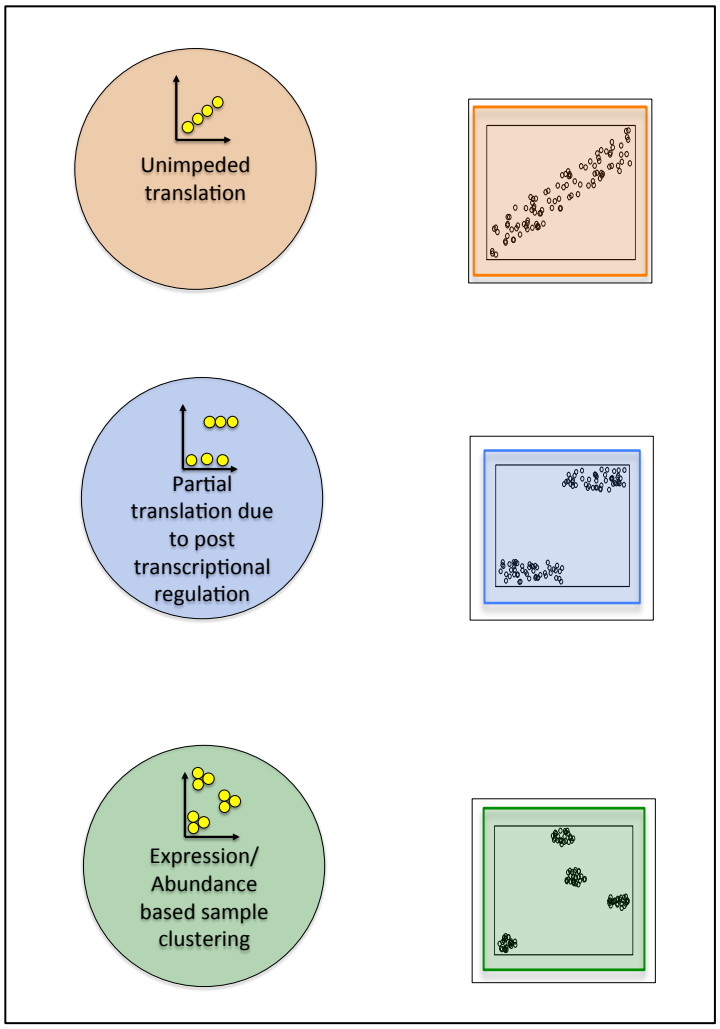
c) Biological relevance and dynamic range dependence – Transcriptomics and proteomics datasets measurements are vastly different and span different dynamic ranges. Most traditional methods are based on clustering genes/proteins according to expression/abundance, which require cumbersome preprocessing and to glean interpretable information describing transcription and translation is difficult. Scatterplots and scagnostics circumvent these issues by transforming the analytical space of modeling the gene-protein relationships by focusing on the trends that describe transcription and translation.



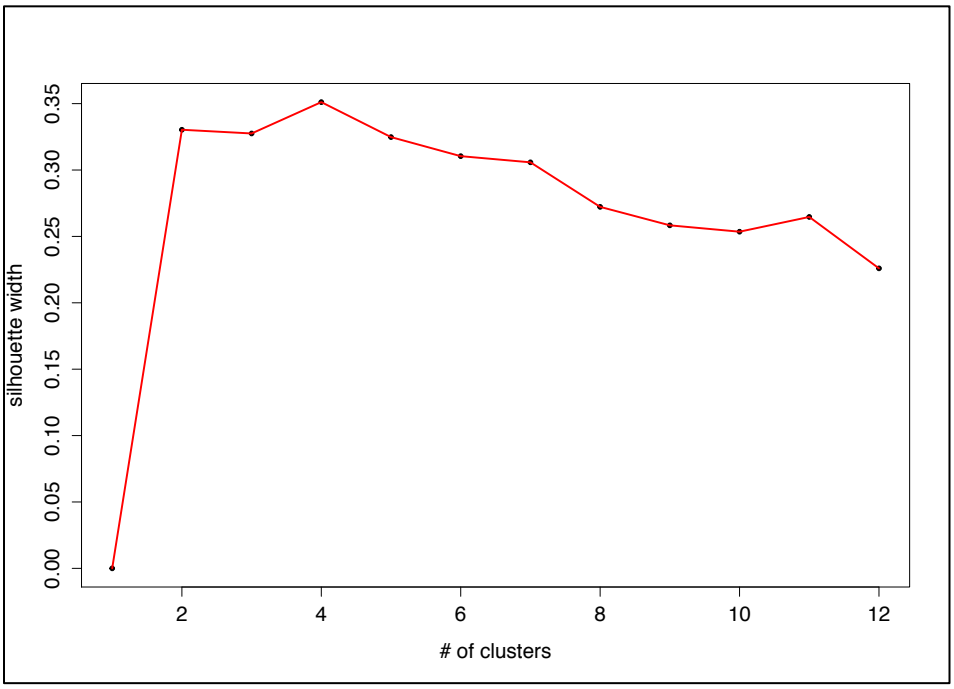
Supplementary figure 1. Correlation plot of the scagnostics features of (a) NCI-60 proteogenomic data (b) Colorectal Cancer proteogenomic data (Zhang 2014). We observe that Striated, Skinny and Stringy features showcase high positive correlation, which may indicate redundancy.



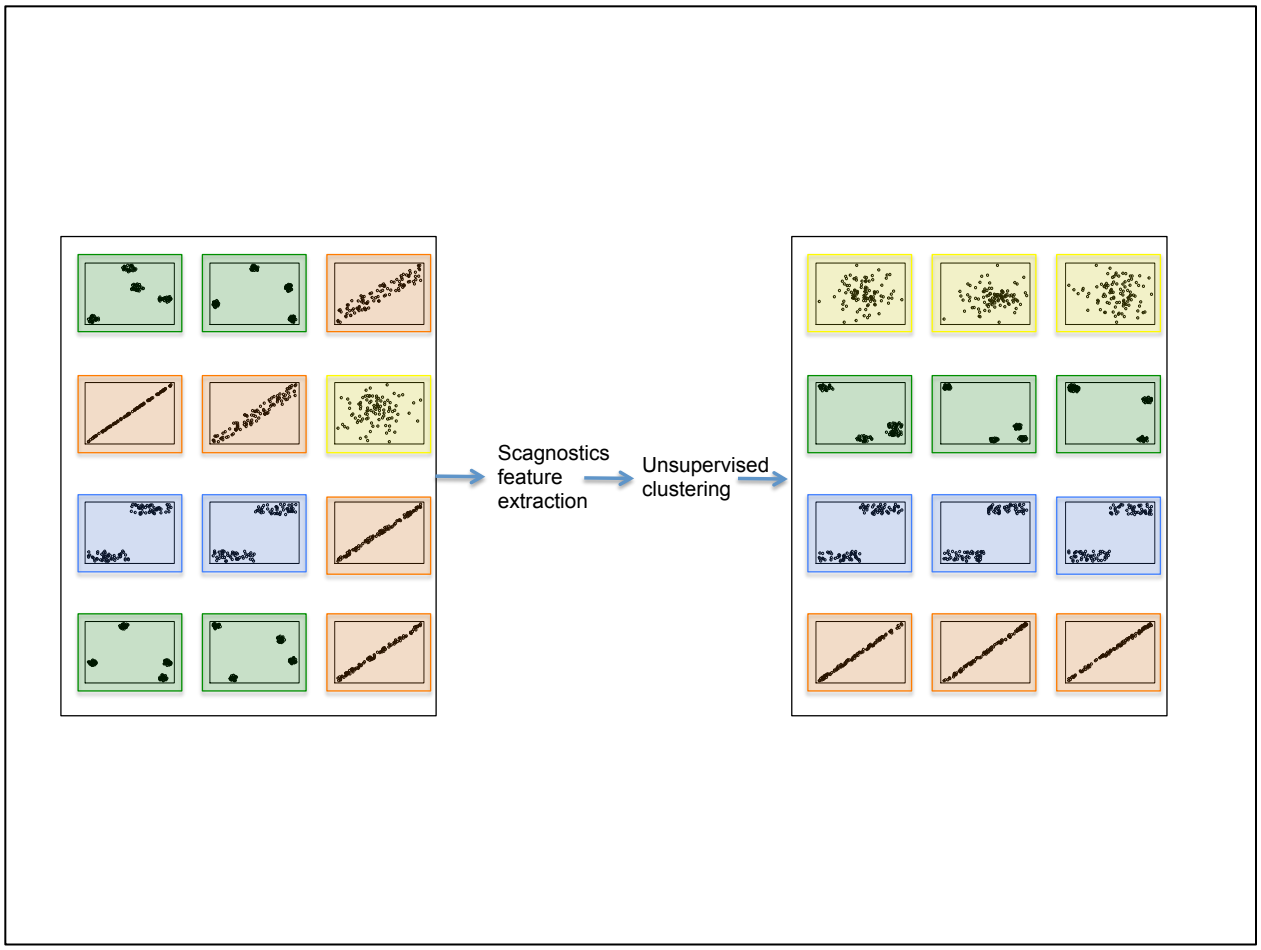
Supplementary figure 2. Random sampling of proteogenomic (gene-protein) pairs from NCI-60 that report high (top 1%) values for each scagnostics features



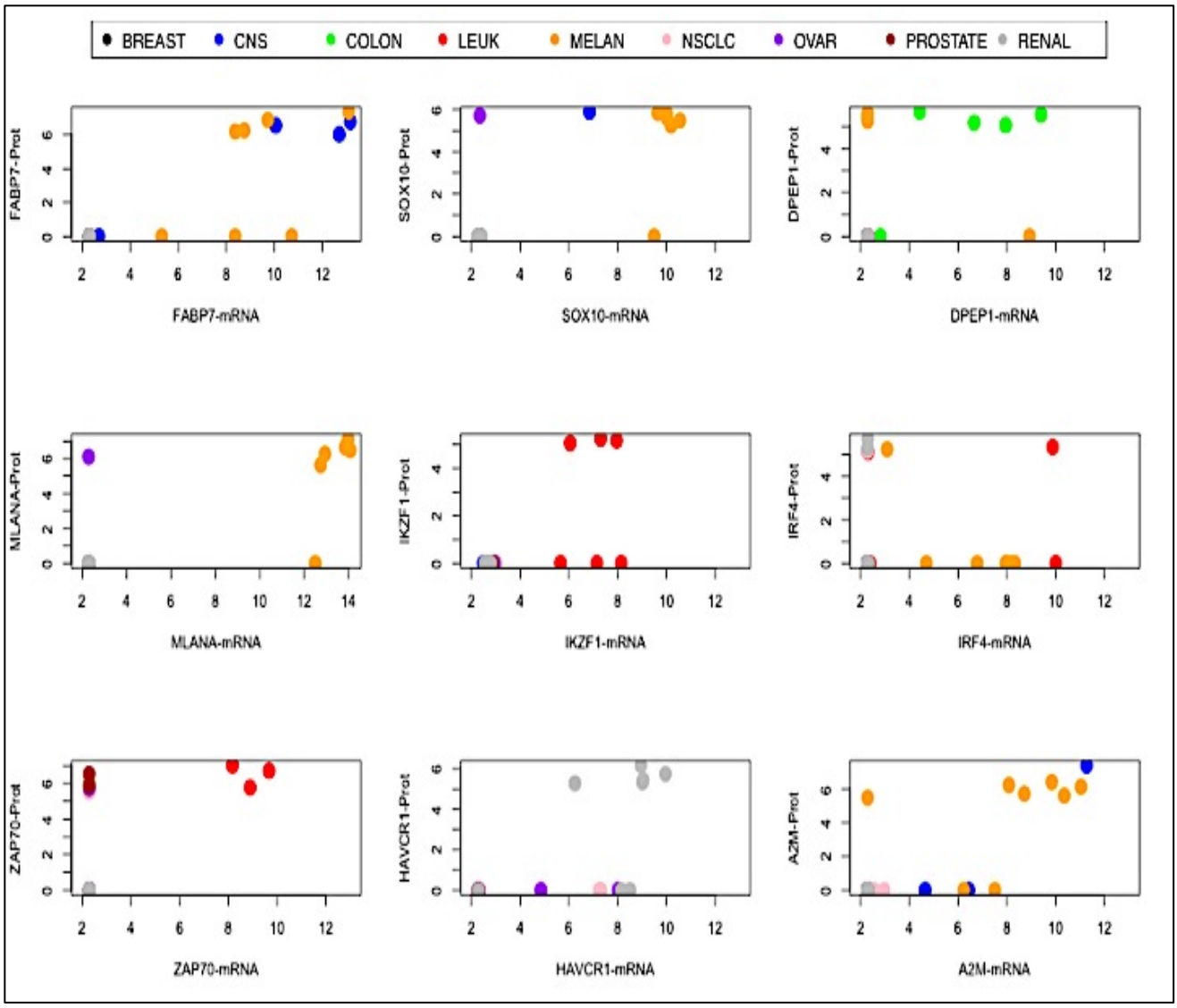
Supplementary Figure 3. Examples of artificial trends in synthetic data that mirror potential regulatory trends



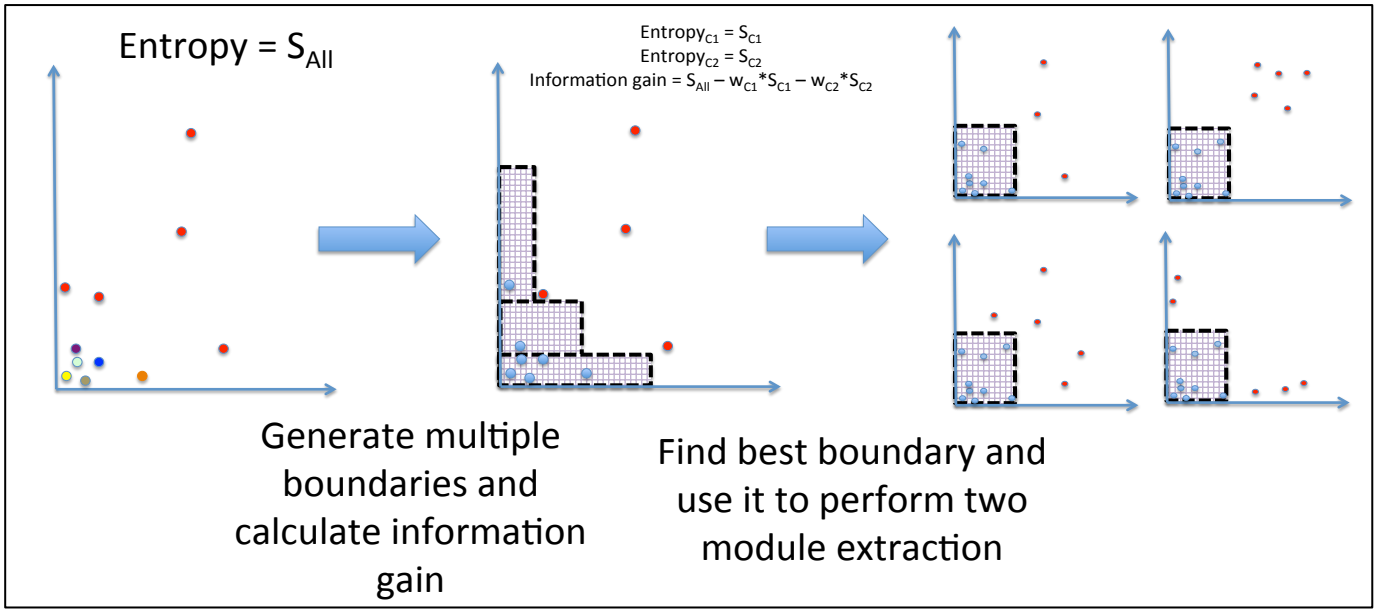
Supplementary Figure 4. Average silhouette width plot to find optimum number of clusters for unsupervised clustering (k-means) of proteogenomic relationships using scagnostics features



Supplementary figure 5. Summary figure representing (Left) Random sampling of the synthetic dataset and (Right) Random sampling of each cluster discovered using scagnostics features.



Supplementary Figure 6. Examples of cluster 3 members with high information gain. Each showcases, cancer tissue specificity in terms of expression and/or abundance.



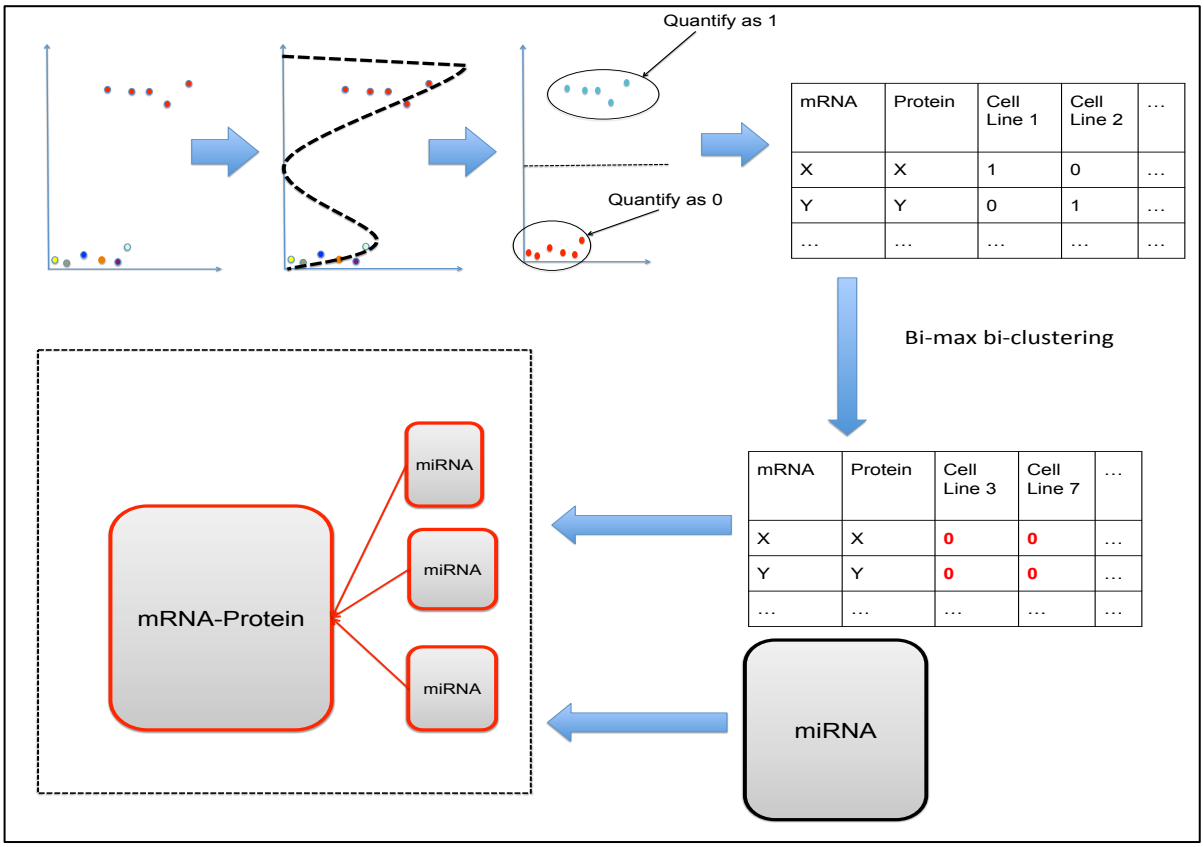
Supplementary figure 7. Workflow for two-module extraction clustering. Utilizes entropy and information gain to decide optimal boundary for separating dense low abundance data points from relatively high abundance data points.

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 Figure 9** 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[22] (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 in **Supplementary Equation 1**.

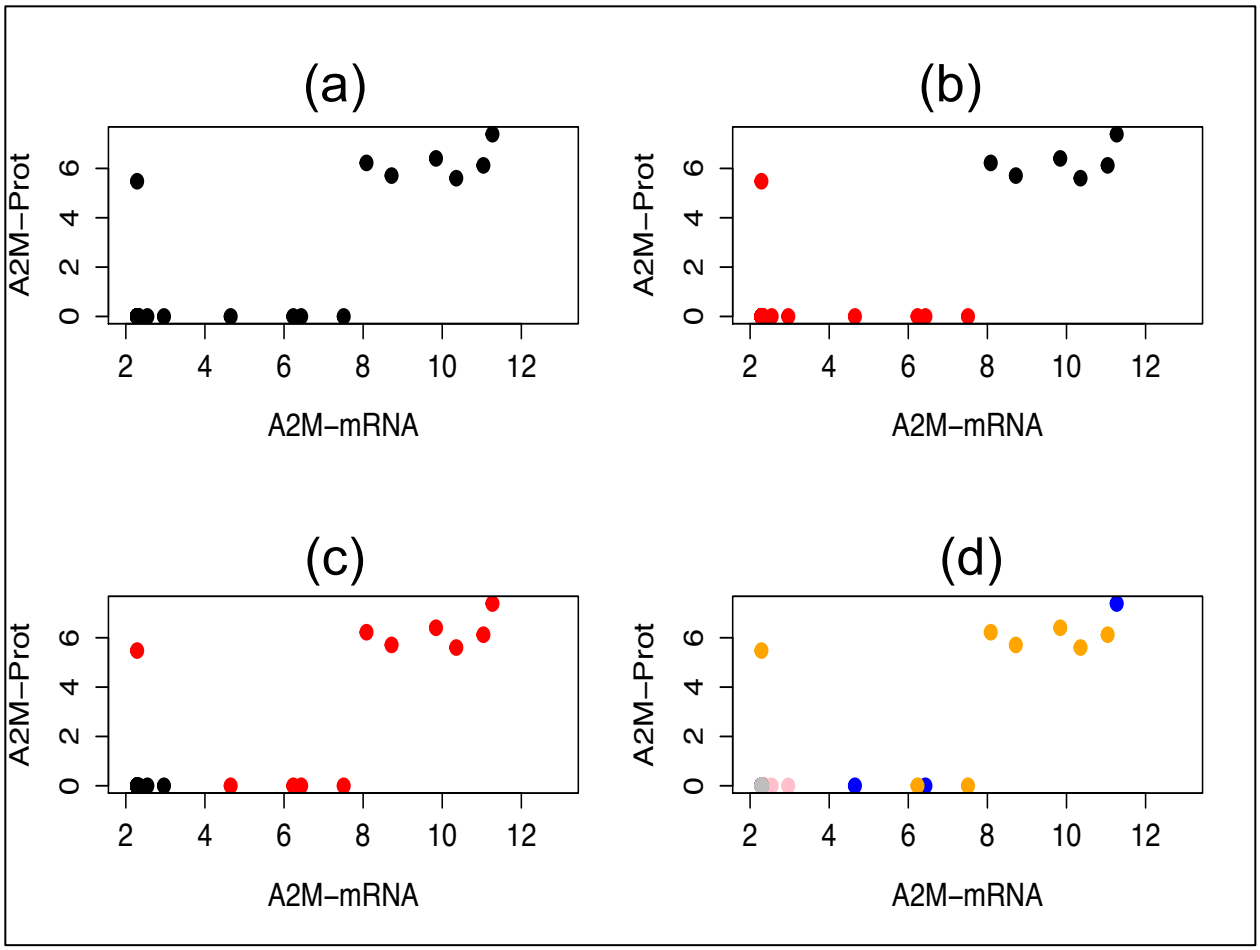
Next we isolated mRNA-protein pairs that show the highest information gain and mapped them to the cancer tissue they indicate specificity for. The two-module extraction was successful in highlighting mRNA-protein pairs that showed high specificity for specific cancer tissues and could potentially be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



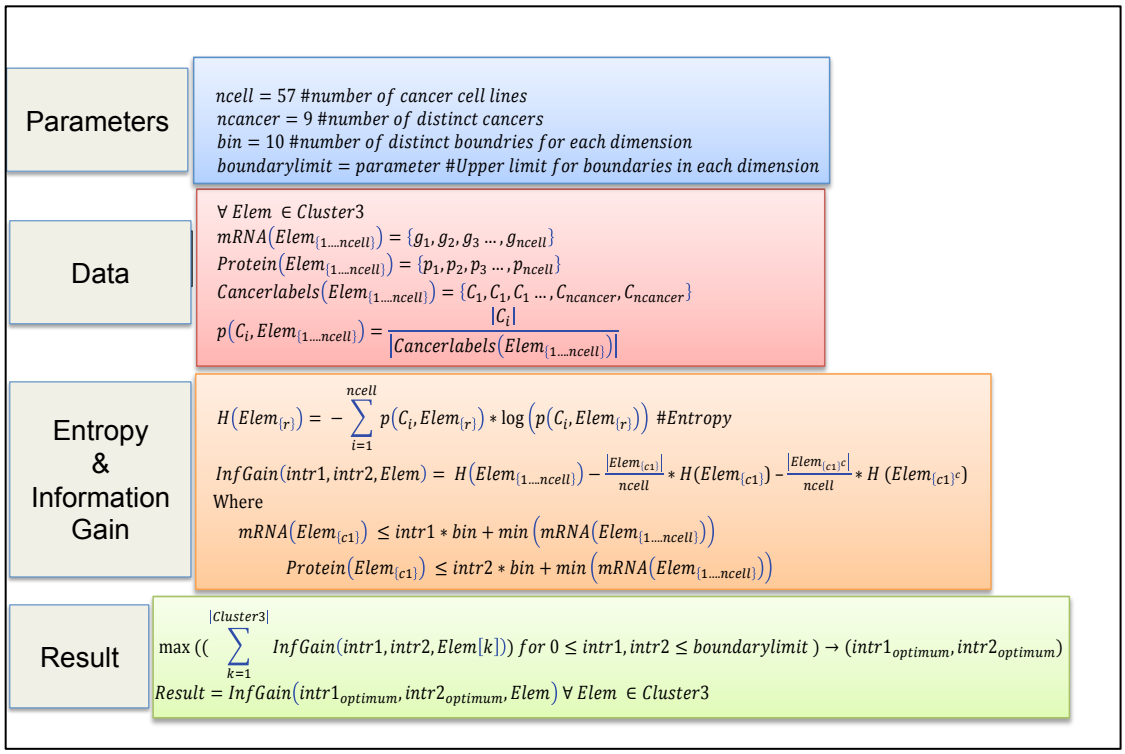
Supplementary figure 8. Modeling potential post-transcriptional regulation with mixture modeling, bi-clustering and miRNA integration.

Step1: Mixture modeling and bi-clustering - Since the defining characteristic of the fourth cluster were extremely high protein abundances for some cell lines and extremely low abundances for the rest, we employ mixture Gaussian modeling of protein abundances in all cell lines, for each transcript-protein relationship. The mixture modeling endeavors to fit two Gaussian distributions to the given data. The two Gaussian distributions report back two means (μ1, μ2) for each of the two Gaussians. The mid-point between μ1 and μ2 served as our linear separator to quantify the cell line data points as having "high" or "low" protein abundance. Thus, any data point of the corresponding protein, which reported expression larger than (μ1 + μ2)/2 was quantified as "high" protein expression (1) and the rest were quantified as "low" protein expression (0). Further, quantifying each data point as high (1) or low (0) converted the protein abundance matrix to a binary matrix, which was used for bi-clustering using the bi-max algorithm[24]. The mathematical formulation for this method can be found in **Supplementary Equation 2**.

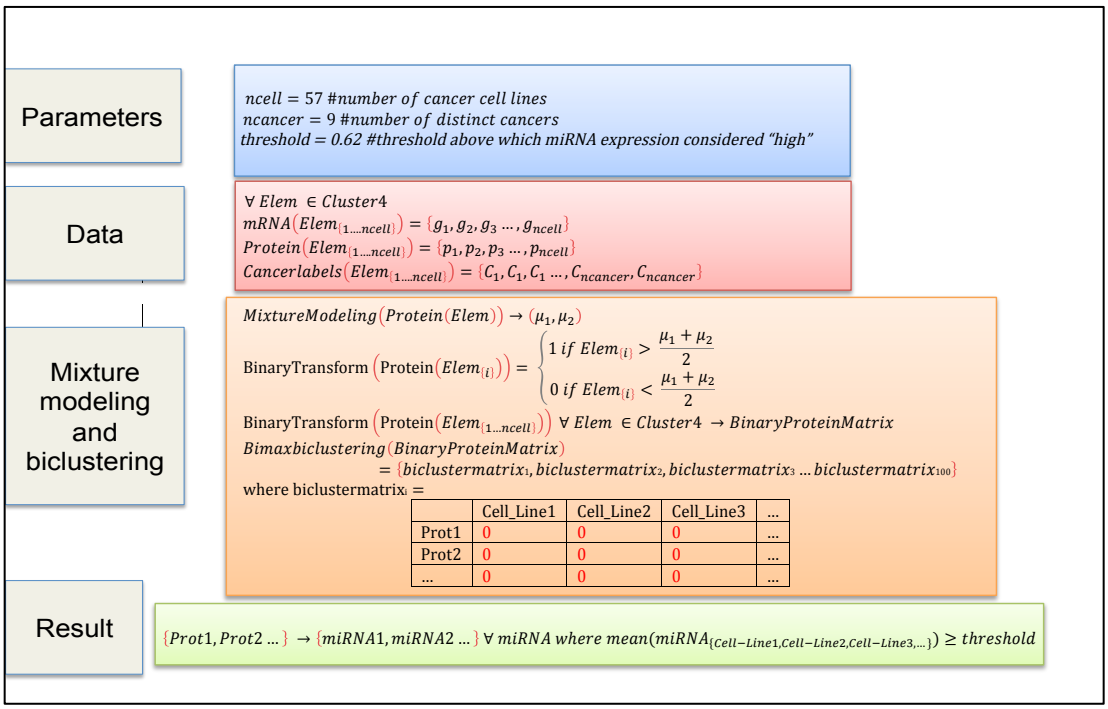
Step2: miRNA integration - We hypothesize that, subgroups of low protein abundances across subgroups of cell lines, isolated from the bi-clustering analysis could shed light on potential miRNAs that serve as post-transcriptional regulators. To recover these gene-miRNA mappings, we search for highly expressed miRNAs in bi-clusters that reported as low abundance in the previous step. We generate these mappings by calculating the average of each miRNA's expression in cell lines that form the low abundance bi-clusters, and extract the miRNAs that report high expression. All average miRNA expression associated with the top 100 bi-clusters reported that the top 25% of the data was contained between normalized expression levels of 0.62 and 8.4. Thus, we use this threshold to label average miRNA expression as "high" if it is reported to be above 0.62.



Supplementary figure 9. (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 Equation 1. Two-module extraction clustering of Cluster 3 members



Supplementary Equation 2. miRNAs to gene set mapping of Cluster 4 members using mixture modeling and bi-clustering of protein matrix