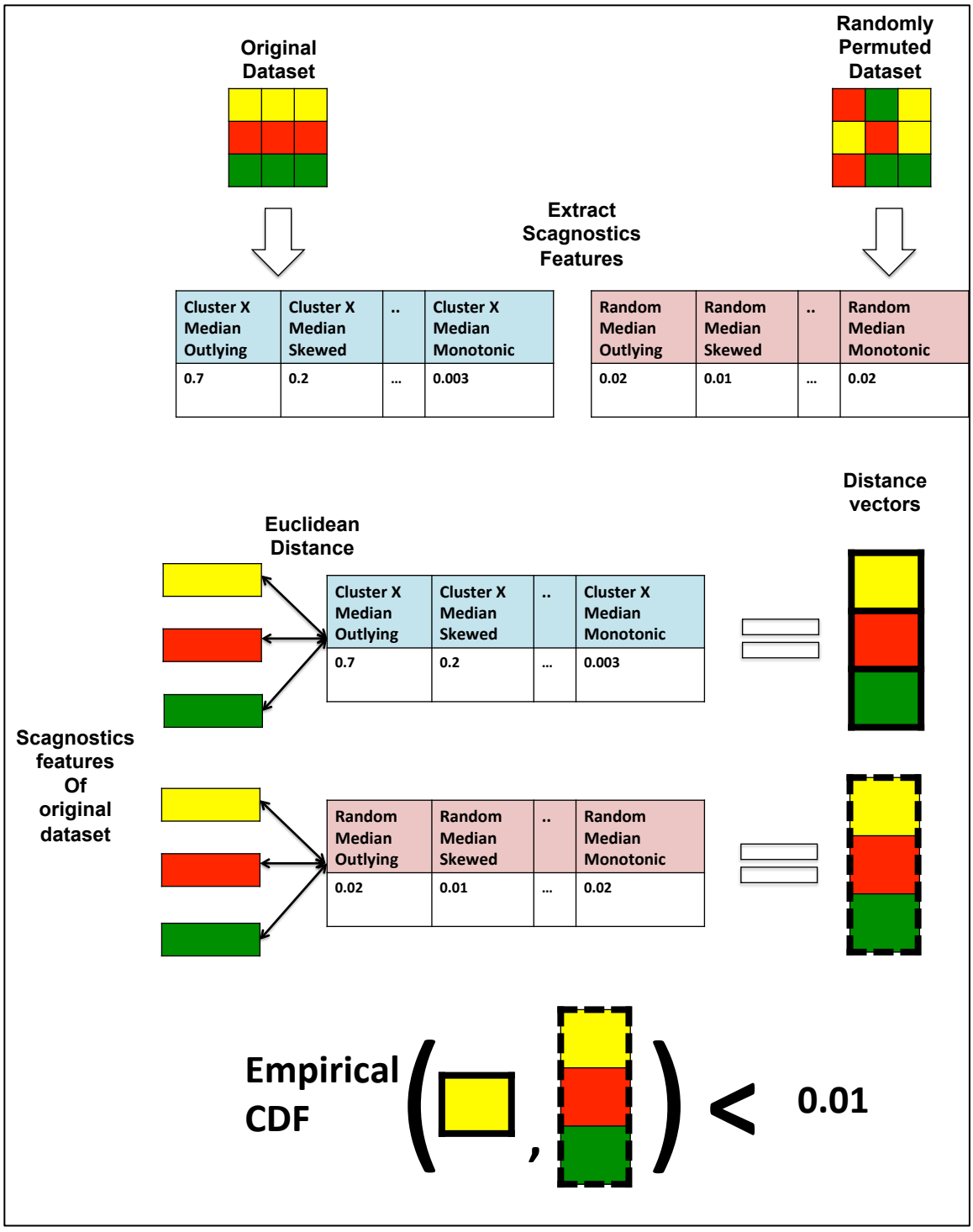
## Cluster filtering/pruning

To prune clusters achieved as a result of scagnostics based clustering on a proteogenomic dataset, we wish to discard the mRNA-protein pairs that do not present a scagnostics vector that is unique to the cluster they have been assigned. Due to the nature of *k*-means clustering, it assigns a cluster to every input mRNA-protein pair based on the closest cluster median to the feature vector. We want to assess the probability or likelihood of the scagnostics feature vector arising from random scatterplots (generated from the distribution of the data), rather than due to a specific regulatory mechanism. To achieve this we randomly permuted our input dataset and subsequently generated another scagnostics feature matrix from these now random and unrelated scatterplots. We then calculated the Euclidean distances between the members of each cluster to the median of this random scagnostics feature matrix (*Dist_to_random*). We also calculated Euclidean distances for each cluster member to its respective cluster median (*Dist_to_cluster*). Now for each of these distances, we assessed the likelihood of the same distance separating each scagnostics vector from the median random scagnostics vector. This likelihood is evaluated by utilizing the empirical cdf of *Dist_to_random*. If this probability was >0.01, the mRNA-protein pair was discarded.

**Supplementary Section 1 (a) (Pruning Clusters).** Method of pruning clusters based on the likelihood of the cluster member's scagnostics feature vector resulting from random data.




**Supplementary Section 1 (b) (Pruning Clusters).** Schematic of the workflow to prune scagnostics based clusters.

## Zhang et al. TCGA Colorectal Cancer Dataset preprocessing and intersection

The TCGA and CPTAC dataset for colorectal cancer has been analyzed in various studies. We utilized the preprocessed and intersected dataset used in the study "Integrative Omics Analysis Reveals Post-Transcriptionally Enhanced Protective Host Response in Colorectal Cancers with Microsatellite Instability". The dataset is located at http://bioinfo.vanderbilt.edu/zhanglab/msi/index.html and has been preprocessed, filtered and intersected as detailed in the above publication. The dataset was filtered to 3764 unique gene-protein pairs across 87 colorectal cancer samples.

## The NCI-60 collection preprocessing and data intersection

NCI-60 is a well-characterized panel of cancer cell lines (e.g. breast, ovarian, prostate, etc.) used extensively to investigate cancer mechanisms and drug response, leading to many new discoveries about cancer. The normalized proteogenomic data are downloaded from the Cellminer database and the NCI-60 proteome resource. The protein dataset identified 8,114 unique protein IDs while the microarray data reported 54,675 investigative probesets. Proteogenomic datasets were mapped to HUGO gene symbols as provided by Gholami et Al (2013) supplementary material and Cellminer annotation. The gene expression profile of the probeset, or protein abundance profile of the protein ID (IPI – International Protein Index) reporting the highest average was retained in the case of one gene symbol mapping to multiple profiles in each matrix. Both datasets were filtered down to 6449 genes across 57 cell lines. In order to investigate post transcriptional regulation, the NCI-60 miRNA data was also utilized. It presented normalized expression values for 306 unique human miRNAs and was also downloaded (normalized and annotated) from Cellminer.

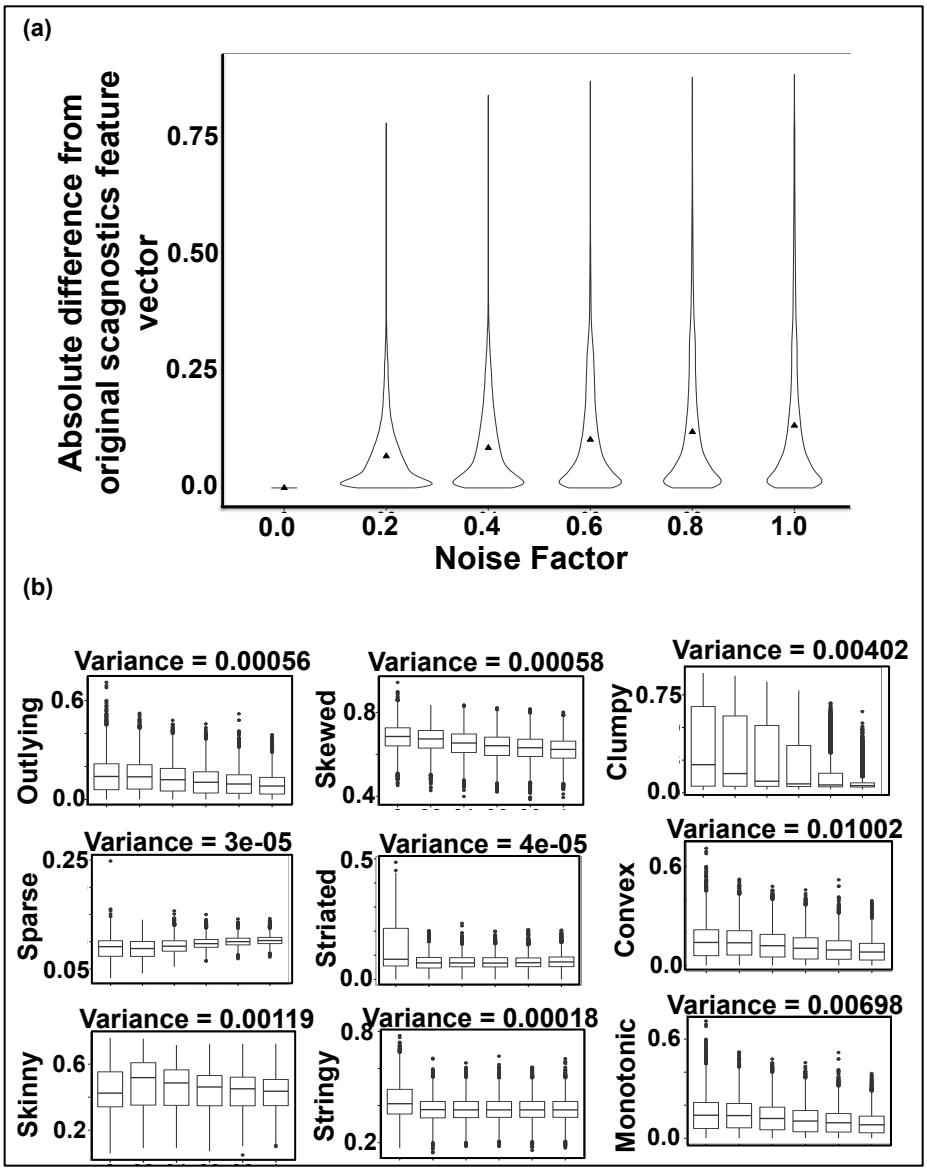
**Supplementary Section 2 (Data preprocessing).** Data preprocessing and intersection details for real world datasets utilized in this study.

| Cluster # | Member scatterplot characteristics | Potential regulatory mechanism | Cluster specific analytical technique |
|---|---|---|---|
| 1 | - Linear mode, high correlation, high values of mRNA expression and protein abundance<br>- ALL data points follow similar trends | All transcription and translation efficient and unimpeded | Gene set functional enrichment using DAVID |
| 2 | - Average/median to high mRNA expression, protein abundance mostly zero, linear model with invariant protein<br>- ALL data points follow similar trends | All transcription efficient, no translation evident OR high rate of protein degradation | Gene set functional enrichment using DAVID |
| 3 | - Cell lines clustered around minimum mRNA and minimum protein expression and abundance, except for a few cell lines deviating from with high mRNA expression or protein abundance | - Only a few cell lines' translation and transcription efficient and unimpeded<br>- Potential cancer tissue specificity due to the nature of the NCI-60 dataset | A custom two-module extraction clustering procedure, to correctly isolate these deviant data points (cell lines) that may be presenting cancer tissue specificity.<br>Details of this method are discussed and the requirement outlined in **Supplementary – Appendix 1 File (Two module clustering)** |
| 4 | Approximately half of the cell lines present high mRNA expression and high protein abundance, the rest record zero protein abundance | - All transcription but cell line specific translation efficient and unimpeded<br>- Possible miRNA ( post transcriptional regulators [CITE] ) intervention, causing translation to be "switched off" in certain cell lines by silencing mRNA | Find bi-clusters (groups of gene-protein pairs across groups of cell lines) where translation is seemingly impeded, and which also showcase high miRNA values.<br>Steps outlined in **Supplementary – Appendix 2 File (Bi-clustering)** |

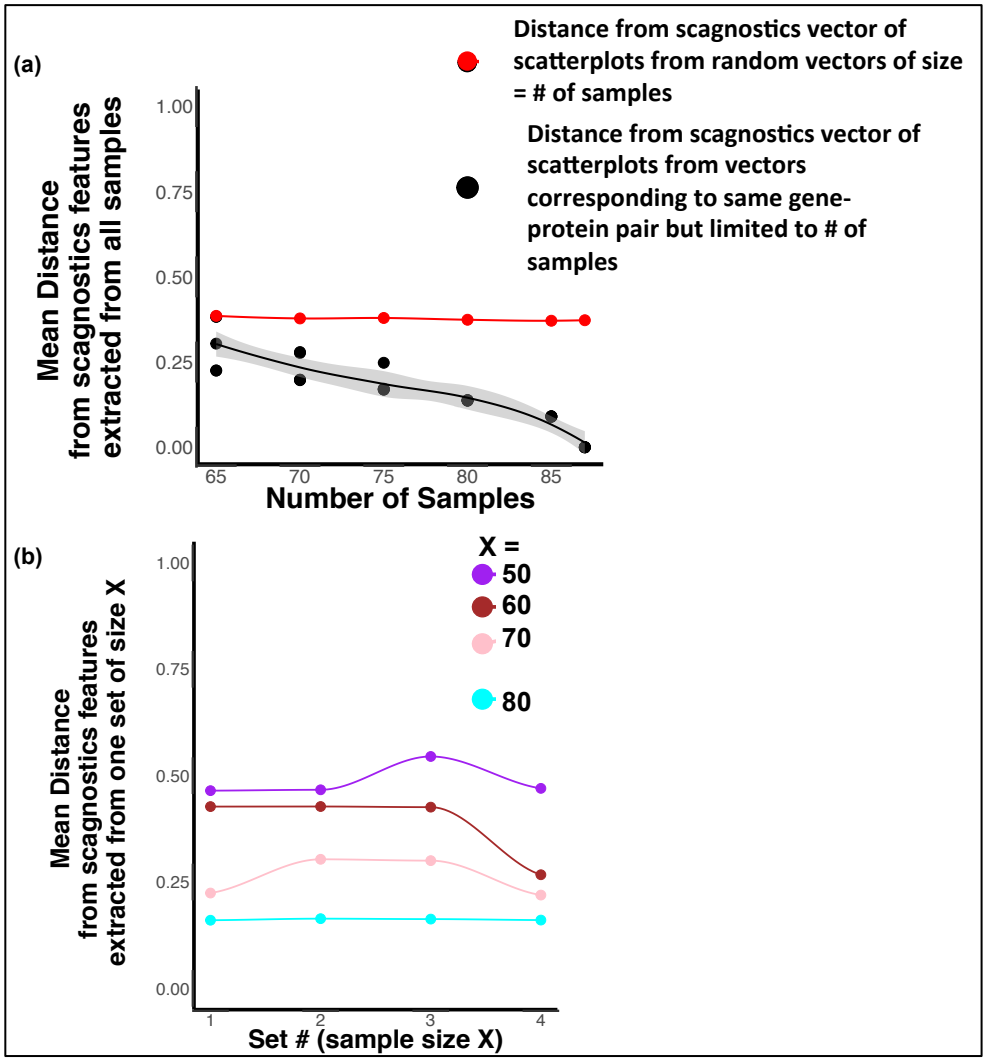
**Supplementary Table 1 (Cluster specific analyses).** Details of cluster specific analysis performed on NCI-60 clusters resulting from scagnostics based analysis.

| Scagnostics Feature | Biological Relevance |
|---|---|
| **Outlying** | The “outlying” feature can be utilized to identify mRNA-protein pairings that show discordant behavior in certain samples, which would further annotate sample specific behavior. |
| **Skewed and Clumpy** | Both “skewed” and “clumpy” features capture the tendency of the data to form natural clusters in the scatterplot. This could highlight potential proteogenomic biomarkers that are capable of subtyping or separating samples especially when highly skewed or clumpy. |
| **Dense** | The “dense” feature provides us the ability to isolate mRNA-protein relationships where mRNA expression and/or protein abundance is largely invariant across samples or experiments. This could be a result of poor or redundant transcription factor activity or dysfunction due to miRNA activation. |
| **Striated** | The “striated” feature could indicate translational thresholds for mRNA and protein abundances. If striations (nearly straight lines in the vertical or horizontal directions) are exhibited in a 2D scatterplot of a proteogenomic pair, it indicates expression invariance across samples of either the mRNA or the protein. This can indicate threshold of mRNA expression or protein abundance at which changing levels of one variable has no effect on the other. This trend may be typical of translational and transcriptional impediments. |
| **Convex, Skinny and Stringy** | The “convex”, ”skinny” and “stringy” features categorize non-linear mRNA-protein relationships and can potentially assist modeling the relationships. Helping model degradation and transcription/translation rates. |
| **Monotonic** | “Monotonic” is a feature that describes conformity to a traditional linear mRNA-protein relationship. It is important to note here that the “monotonic” feature describes the linear relationship regardless of positive or negative correlation, and is a function of spearman’s correlation. |

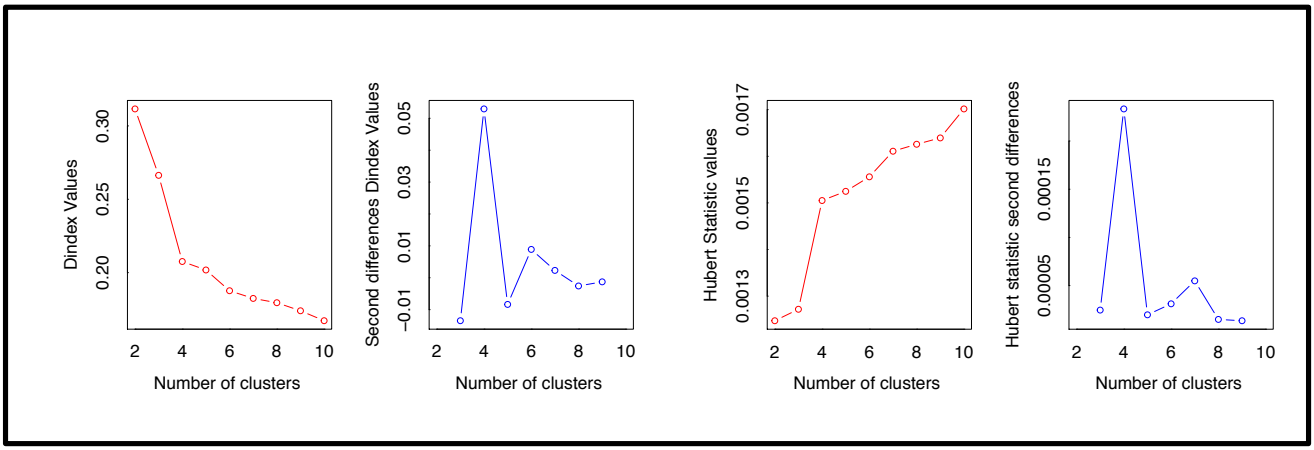
**Supplementary Table 2 (Scagnostics features relevance).** Biological relevance of each scagnostics feature when categorizing mRNA-protein relationships.



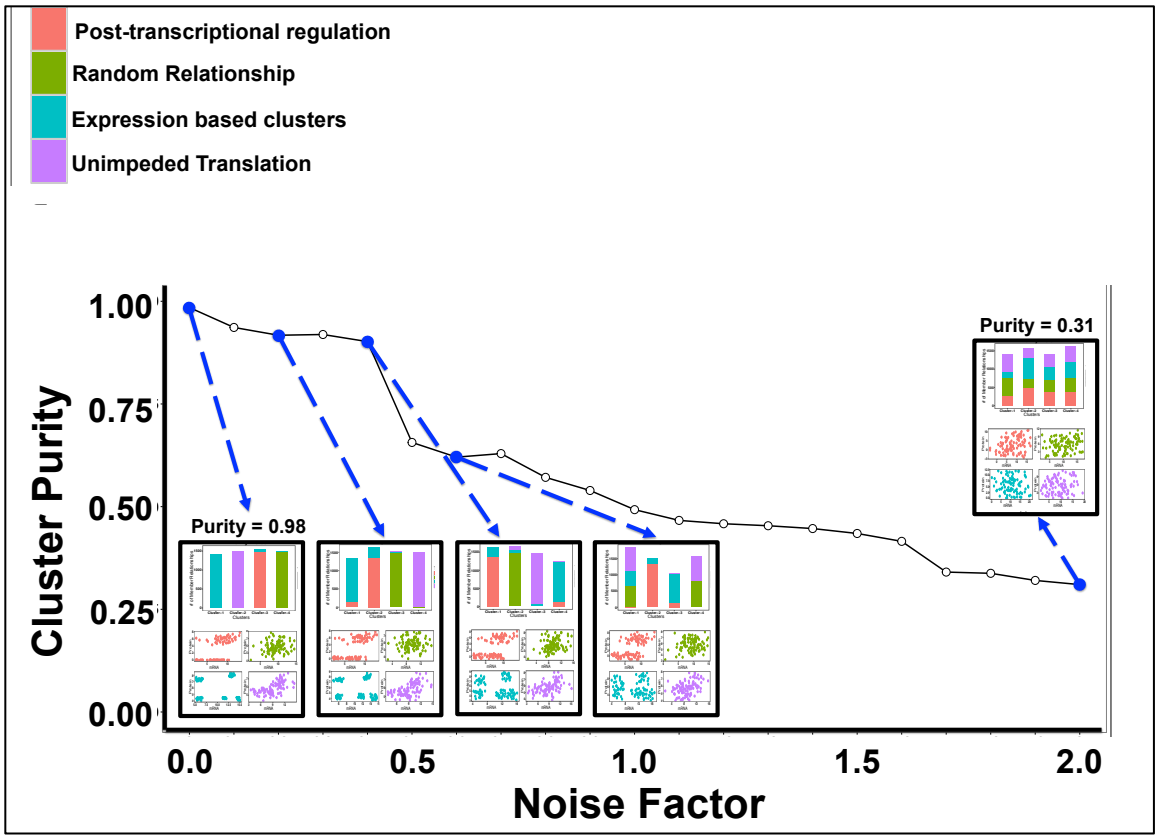

**Supplementary Figure 1 (Robustness against noise)** (a) Violin plot visualizing the change in the entire scagnostics feature matrix by showcasing absolute differences between the scagnostics feature matrix extracted from increasingly noisy synthetic datasets and scagnostics feature matrix extracted from original (noiseless) synthetic data. (b) Visualize change in each scagnostics feature value for all the scatterplots in the incrementally noisy synthetic dataset. The variance amongst the medians of feature values is reported.



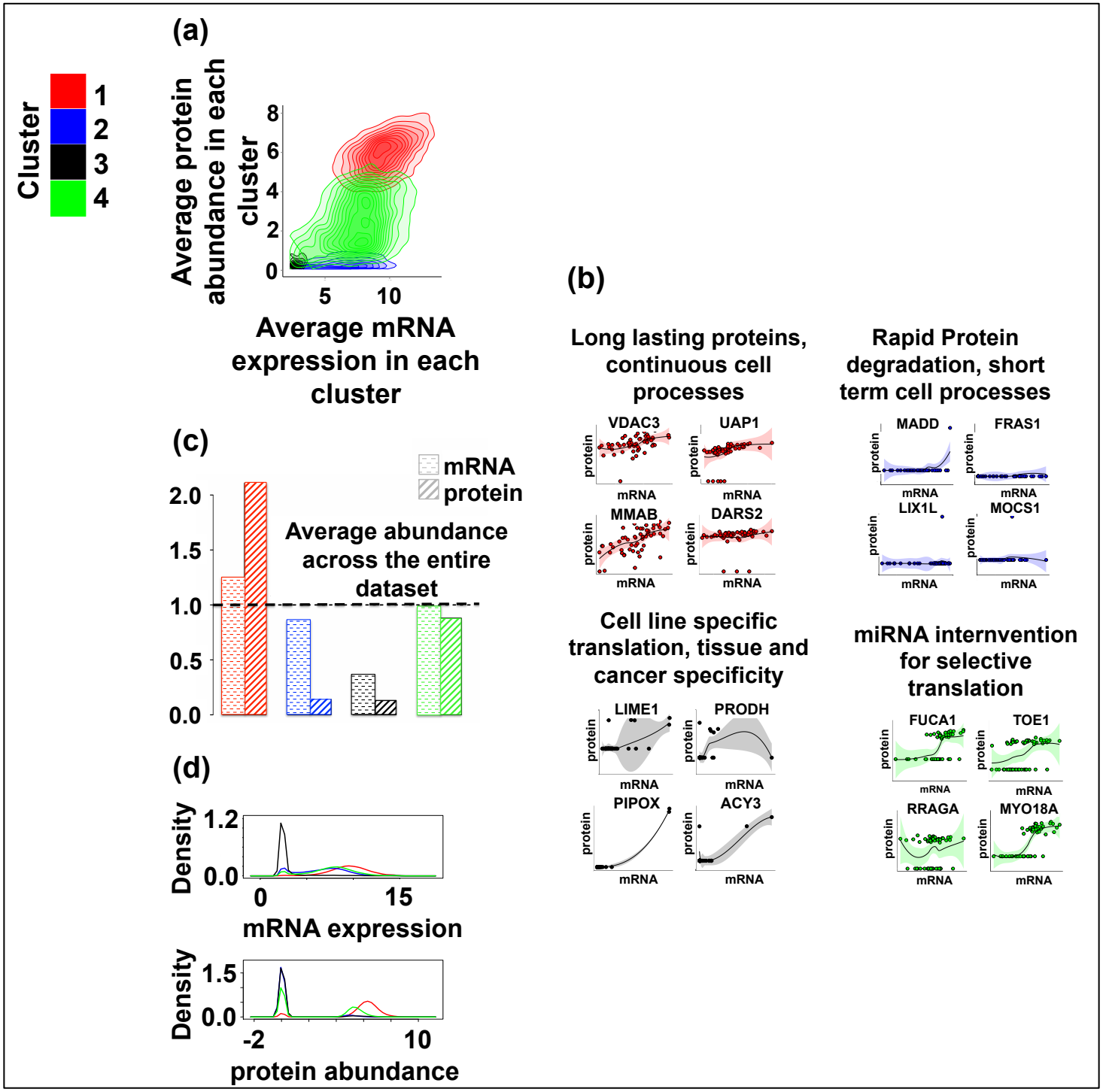

**Supplementary Figure 2 (Sampling robustness and consistency)** (a) Euclidean distance from the scagnostics matrix of the original dataset (n=87) of scagnostics matrices generated from data with varying number of samples. In red the data from which scagnostics feature matrices are generated is randomly permuted to create random relationships (serving as a negative control) whereas in black the data is the same as the original dataset. This showcases that the variance in the scagnostics feature vector is very small even when ~25% of the samples are removed. (b) Showcases the difference between scagnostics feature matrices generated from different sample sets containing the same number of samples, but constituting of varying samples from the same cohort. The multiple lines denote multiple different sample sizes. We observe that whereas the difference between scagnostics feature matrices of same sample sizes but varying samples is constant, implying that the scagnostics feature matrix is consistent, the difference reduces as the number of samples increase.



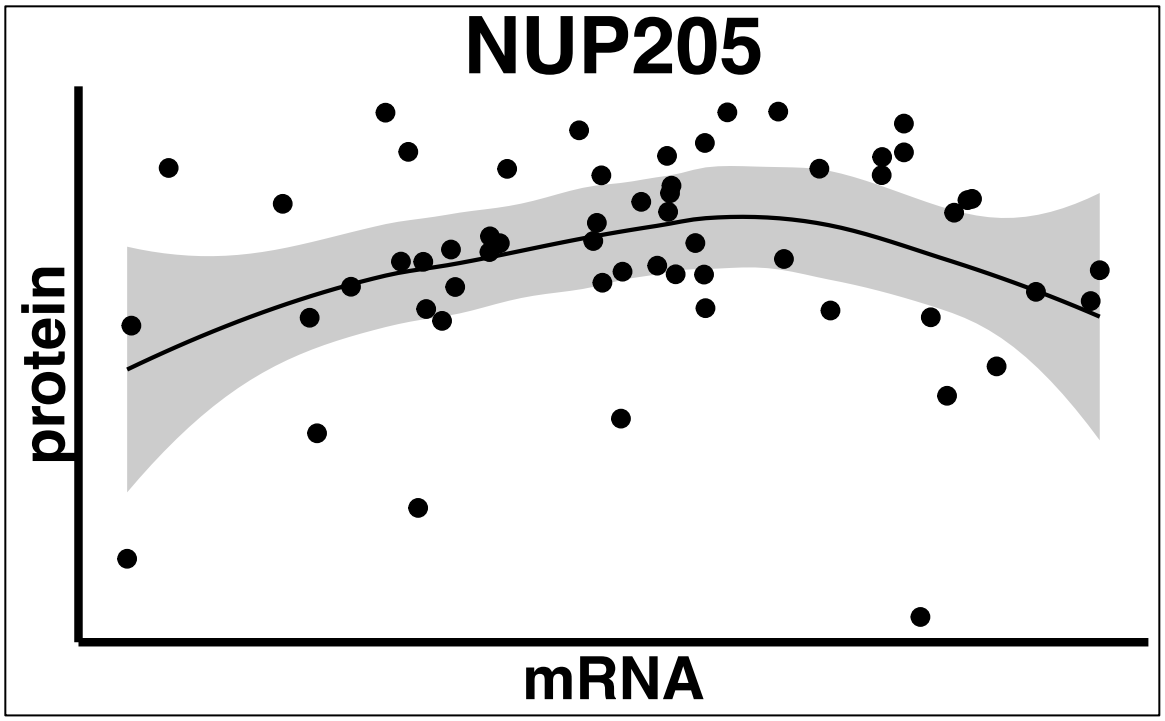

**Supplementary Figure 3 (Optimal Clusters)** The Dindex and Hubert index both indicate that the scagnostics feature matrix of the Synthetic proteogenomic dataset presents 4 underlying clusters. The knee in the red plots and the peak in the blue plots both indicate the optimal number of underlying clusters.



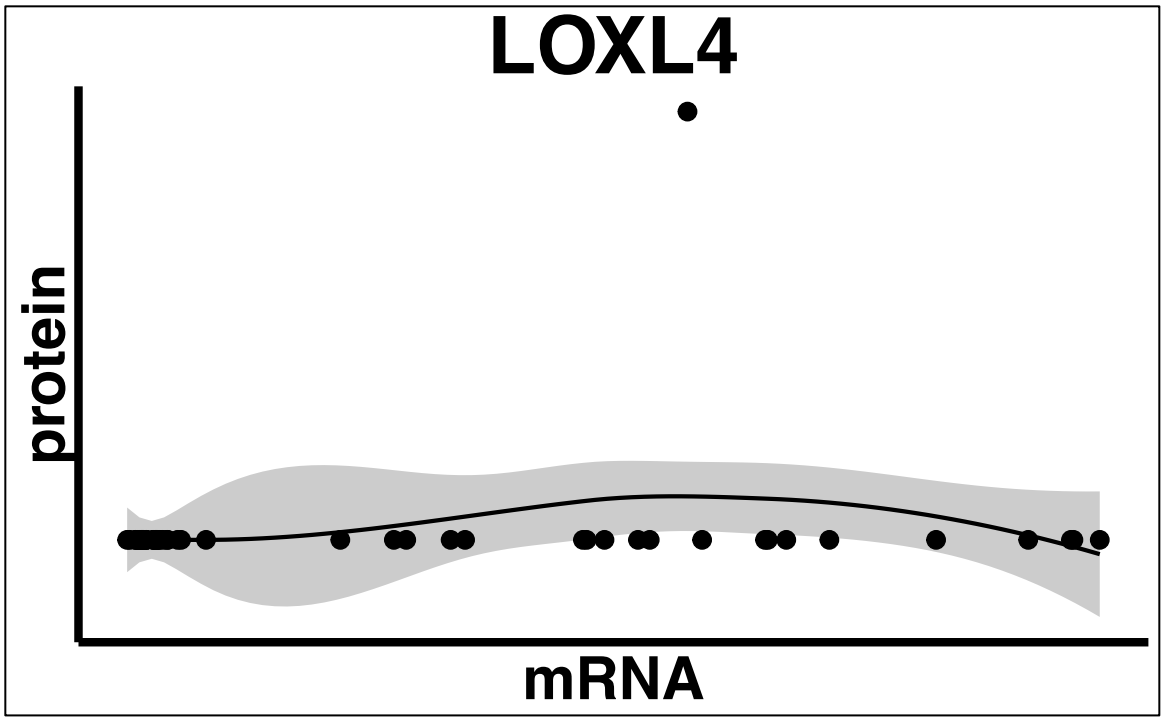

**Supplementary Figure 4 (Clustering across noise)** Progressively higher noise is added to the synthetic dataset (from left to right) and the resulting scagnostics feature based clusters are evaluated and visualized at Noise factor = 0, 0.2,0.4,0.6 and 2. The bar graphs indicate cluster purity of scagnostics based clustering at each noise level. The scatterplots provide a sample of how each spiked in trend is transformed as a result of noise.




**Supplementary Figure 5 (NCI-60 before pruning).** (a) Contour density plot of the average mRNA expression of each gene versus the average protein abundance of each protein across NCI-60 scagnostics feature driven clusters. (b) Sampling of proteogenomic relationships from each cluster. (c) Relative average mRNA and protein abundance in each cluster, as compared to average across the entire dataset denoted by the dotted line. (d) Densities of mRNA and protein abundance in each cluster.

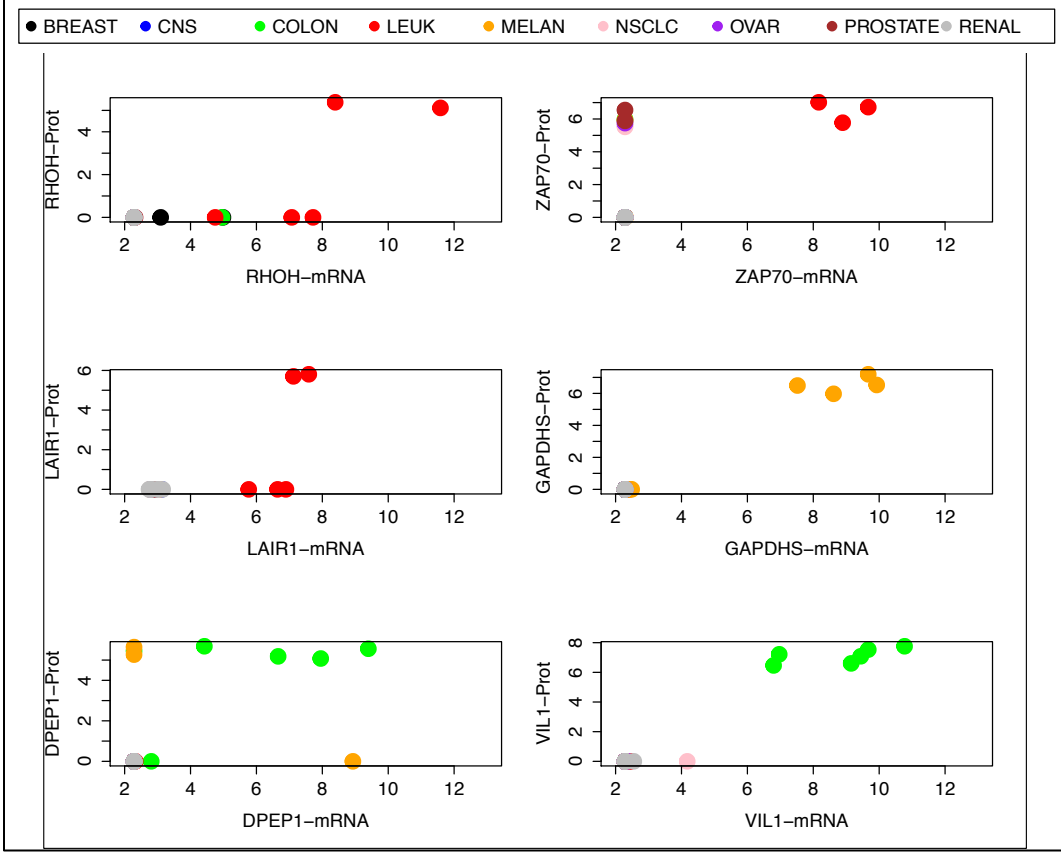



**Supplementary Figure 6 (a) (NCI-60 Cluster Examples).** Example of a gene-protein pair from Cluster 1. Scatterplot of the NUP205 gene-protein pair. Verified as an inherently long lasting protein.

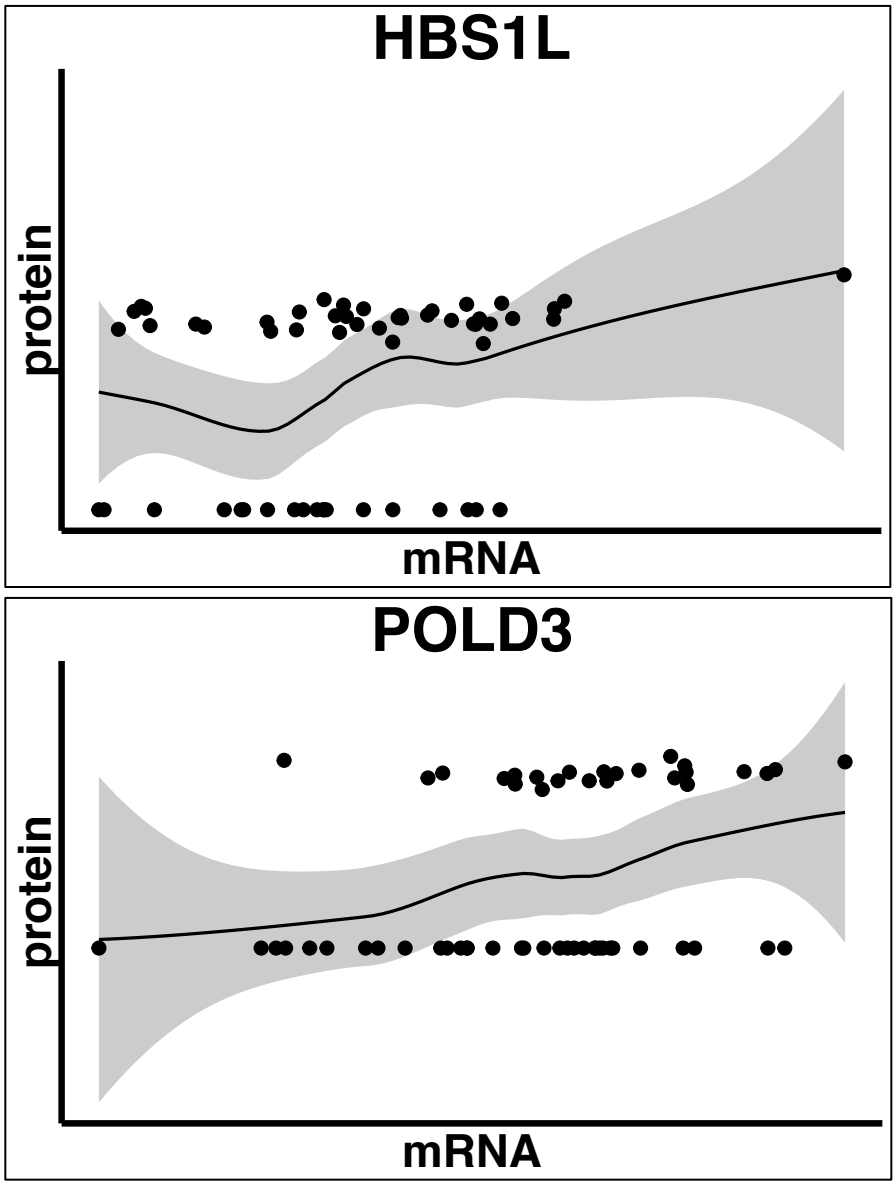



**Supplementary Figure 6 (b) (NCI-60 Cluster Examples).** Example of a gene-protein pair from Cluster 2. Scatterplot of the LOXL4 gene-protein pair. Verified as presenting vastly different half-lives for mRNA vs protein.

| mRNA-protein | Cancer Specificity | Past experimental evidence |
|---|---|---|
| RHOH | Leukemia | RhoH GTPase is involved in the development of B-cell chronic lymphocytic leukemia |
| ZAP70 | Leukemia | ZAP-70 expression is applied clinically and yields important prognostic information for patients with chronic lymphocytic leukemia |
| LAIR1 | Leukemia | The ITIM-containing receptor LAIR1 is essential for acute myeloid leukaemia development |
| GAPDHS | Melanoma | High expression of glycolytic (GAPDHS and others) and pigment proteins is associated with worse clinical outcome in stage III melanoma. |
| DPEP1 | Colon Cancer | DPEP1, expressed in the early stages of colon carcinogenesis, affects cancer cell invasiveness |
| VIL1 | Colon Cancer | Has a role in crosslinking actin fibers at the core of microvilli in intestinal epithelium |




**Supplementary Figure 6 (c) (NCI-60 Cluster Examples).** Example of a gene-protein pairs from Cluster 3.All of them present verifiable cancer specificity.




**Supplementary Figure 6 (d) (NCI-60 Cluster Examples).** Example of a gene-protein pair from Cluster 4. Scatterplot of the HBS1L and POLD3 gene-protein pair. The NCI-60 miRNA dataset, showcases significant correlation between HBS1L protein abundance and miRNA hsa-mir-32 expression, as well as POLD3 protein abundance and miRNA hsa-mir-765 expression. These pairings have been validated, as they are included in miRTarBase, the experimentally validated microRNA-target interactions database.