PTR Explorer: An approach to identify and explore Post Transcriptional Regulatory mechanisms using proteogenomics

Arunima Srivastava[†]
Dept. of Computer Science and Engineering, The Ohio State University, 2015 Neil Ave Columbus, OH
Email: srivastava.92@osu.edu

Michael Sharpnack
Dept. of Biomedical Informatics, The Ohio State University, 1800 Cannon Drive Columbus, OH
Email: sharpnack.7@osu.edu

Kun Huang
Department of Medicine, Indiana University School of Medicine, 340 W 10th St #6200, Indianapolis, IN
Email: kunhuang@iu.edu

Parag Mallick[#]
Canary Center for Cancer Early Detection, Stanford University, 3155 Porter Dr., Palo Alto, CA
Email: paragm@stanford.edu

Raghu Machiraju[#]
Dept. of Computer Science and Engineering, The Ohio State University, 2015 Neil Avenue, Columbus, OH
Email: machiraju.1@osu.edu

[#]*Corresponding authors*

Integration of transcriptomic and proteomic data should reveal multi-layered regulatory processes governing cancer cell behaviors. Traditional correlation-based analyses have demonstrated limited ability to identify the post-transcriptional regulatory (PTR) processes that drive the non-linear relationship between transcript and protein abundances. In this work, we ideate an integrative approach to explore the variety of post-transcriptional mechanisms that dictate relationships between genes and corresponding proteins. The proposed workflow utilizes the intuitive technique of scatterplot diagnostics or scagnostics, to characterize and examine the diverse scatterplots built from transcript and protein abundances in a proteogenomic experiment. The workflow includes representing gene-protein relationships as scatterplots, clustering on geometric scagnostic features of these scatterplots, and finally identifying and grouping the potential gene-protein relationships according to their disposition to various PTR mechanisms. Our study verifies the efficacy of the implemented approach to excavate possible regulatory mechanisms by utilizing comprehensive tests on a synthetic dataset. We also propose a variety of 2D pattern-specific downstream analyses methodologies such as mixture modeling, and mapping miRNA post-transcriptional effects to

[†] Student supported by The Ohio State University's graduate presidential fellowship

explore each mechanism further. This work suggests that the proposed methodology has the potential for discovering and categorizing post-transcriptional regulatory mechanisms, manifesting in proteogenomic trends. These trends subsequently provide evidence for cancer specificity, miRNA targeting, and identification of regulation impacted by biological functionality and different types of degradation. (Supplementary Material - https://github.com/arunima2/PTRE_PSB_2020)

Keywords: Multiomics; Integrative; Multi-dimensional; Proteomics; Transcriptomics.

1. Introduction

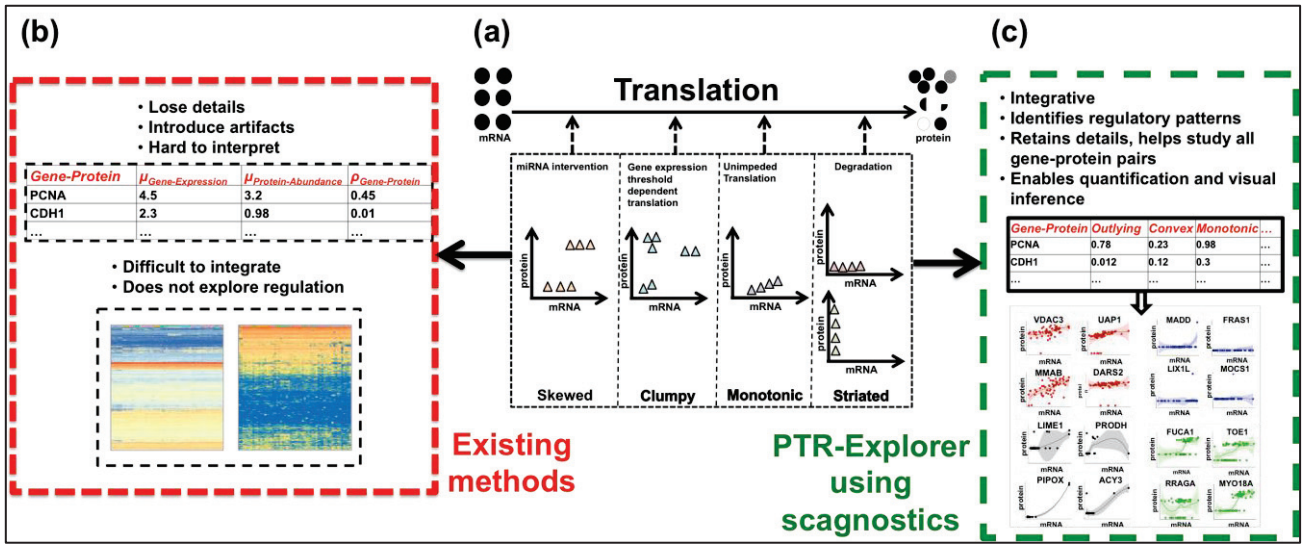

Figure 1. Exploring regulatory mechanisms with 2D scatterplots (a) Depiction of the various post-transcriptional regulatory mechanisms using gene expression and protein abundance 2D scatterplots, additionally annotated with the geometric feature that best represents these 2D patterns (b) Existing methods to quantify gene-protein relationships e.g. aggregation which loses information and individual clustering and visualization which fails to easily integrate information (c) Scagnostics based PTR Explorer, which intuitively assesses 2D patterns borne out of different regulatory mechanisms impacting gene protein relationships, using quantifiable geometric features

Recent advances in profiling techniques have made integrated analysis of the transcriptome and proteome (proteogenomics) a possibility and highlighted major gaps in our understanding of multi-omic regulatory processes[1]. Novel approaches for multi-omic analysis will be required to extract insight from these complex proteogenomic datasets. If successful, these novel approaches may ultimately illuminate fundamentals of multi-omic biological processes while concurrently enabling new classes of therapeutics that target those processes [2]. To date, multiple studies have proven that transcription and translation are impacted by a variety of factors[3] such as RNA and protein degradation, post-translational modification (PTM), the influence of non-coding RNAs, and epigenetic regulation[4–6]. Although these studies have been instrumental in revealing substantial disconnects between the transcriptome and proteome, they have not examined the diversity of mRNA-protein relationships or the regulatory mechanisms underlying those relationships. Additionally, sophisticated approaches have been utilized to explore intra transcript and protein relationships respectively (e.g. network approaches to build gene co-expression networks and protein-protein interaction networks), but such efforts lack intuitive integration of multi-omic data. Integrative multi-level analysis is an approach that can ideally quantify detailed information about regulatory mechanisms. Other studies have looked, in aggregate at the

exhaustive set of all gene (mRNA)-protein relationships within a sample[7,8]. While descriptive of general lack-of-correlations between all genes and proteins, they do not explore specific, nuanced trends between specific genes and their corresponding proteins, nor the regulatory origins of these relationships. There is a critical need for tools that examine the specific relationships between a given transcript and its associated protein across samples/conditions that are a result of specific post-transcriptional effects.

Here we present (P)ost-(T)ranscriptional (R)egulatory mechanism Explorer, a novel methodology to explore the presence and impact of post-transcriptional regulation by utilizing proteogenomic data. The approach is based on a technique that has not yet been utilized in the biological analyses space, but one that addresses the critical shortcomings of existing traditional multi-omic analyses by harnessing the elementary and intuitive visualizations of these multi-omic relationships. We posit that an examination of specific mRNA-protein relationships (rather than global trends) may reveal information about underlying regulatory mechanisms and disease processes. We further hypothesize that mRNA-protein relationships with common trends in their scatterplots may share common regulatory mechanisms. The simplest way to visualize mRNA-protein relationships is in a bivariate mRNA-protein scatterplot where, each point represents a gene's transcript and protein abundance in a given sample (**Figure 1(a)**). The usefulness of scatterplots is evident when we observe the shortcomings of utilizing summary statistics, correlative measures or isolated dataset clustering and visualization to infer any useful biological information across samples (**Figure 1(b)**). Not only do we lose the relevant nuances in the subsequent compression, but we may also introduce artifacts that affect downstream analysis.

Proteogenomic relationships have an underlying biological impetus offering a wealth of information about an experiment, apart from just the straightforward questions that can be answered by mRNA expression and protein abundance visualization and differentials (**Figure 1(b)**). While other methods have comprehensively explored the space of building inference networks from scatterplots[9,10] for a variety of different problems, they still lack implementation in the multi-omic space, and are missing an interpretable quantifiable layer from which one can easily perform visual inference. Certain examples of note are Sahoo et al[11] illustrating the usefulness of gleaning pairwise boolean implications using gene expression scatterplots and similarly Yates et al[9] proposing a visual analytical framework classifying scatterplots in a scatterplot matrix to reconstruct interaction networks.

Our methodology endeavors to solve a similar problem in the multi-omic space. As highlighted above, varying types of post-transcriptional regulation (protein and RNA degradation, functional responsibilities within the cell and miRNA intervention), all contribute to the uniqueness of mRNA-protein relationships and the study of their subsequent functionalities, making them a powerful tool in cancer studies. Our method aims at adding a layer of quantifiable features to existing visual analytics in order to extract useful relationships, their corresponding regulatory mechanisms and their functional attributes (**Figure 1(c)**). Our workflow takes advantage of scagnostics [12–16], to quantify the characteristics of the two-dimensional patterns present in scatterplots. Scagnostics features geometrically quantify a single scatterplot as gradations of nine features **(Figure 2)**[17]. They have been earlier used in classifying time series data including weather patterns assessed using temperature and pressure attributes[18].

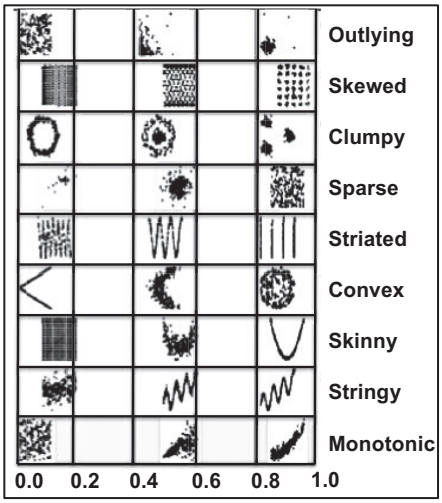

Figure 2. Scagnostic features quantify the diverse relationships that may be found in scatterplots [17]. The schematic showcases each scagnostic feature on the Y-axis and a relative measure of the scagnostics feature on the X-axis. The graph contains examples of scatterplots that would present high and low value of each scagnostics feature

We first examined whether our scagnostics based method could extract biologically relevant signals from proteogenomic data. To do this, we constructed a workflow that modeled mRNA-protein relationships based on their nine (9) dimensional scagnostic features. As a first test of the workflow, we generated a synthetic dataset that contained a mixture of random transcript-protein relationships (background) as well as multiple varieties of biologically relevant transcript-protein relationships (signal) that can possibly exist as a result of different post-transcriptional regulation mechanisms (e.g., the presence of an miRNA governing the transcript/protein relationship). We then tested whether scagnostic features of the 2D scatterplots generated from this data in tandem with clustering successfully differentiated signal from background. Further, we also performed comparisons with traditional methodologies.

Having demonstrated that scagnostics could successfully uncover biologically relevant mRNA-protein relationships and shed light on the corresponding PTRs, we apply our scagnostics-based methodology to proteogenomic data previously collected on the 57 cell types in the NCI-60[19]. We identified clusters of gene-protein relationships characteristic of a variety of different regulatory trends including protein degradation, post-translational modifications, and cancer specific signaling. These studies support the idea that a scagnostics based approach is able to uncover relevant proteogenomic relationships and their corresponding PTRs would not be uncovered by any other traditional method.

2. Methods and Materials

2.1. *Scagnostics analytical workflow*

Our data processing and analysis workflow is shown in **Figure 3**. In ***Step 1*** we (a) perform quality control, removal of missing data and de-noising as well as (b) map protein group identifiers/mRNA transcript identifiers to the relevant HUGO names/gene symbols. In ***Step 2***, we filter the datasets to retain only genes with corresponding protein profiles available. In ***Step 3***, we apply the R package *scagnostics*[14] to the scatterplot generated by examining the values of each transcript-protein pair across the conditions it has been measured in. Scagnostics accepts this scatterplot as input and outputs a nine dimensional feature vector describing how "Outlying", "Skewed" or "Clumpy" etc. the scatterplot is.

In ***Step 4*** we perform unsupervised clustering of mRNA-protein pairs. After examining diverse clustering approaches, we found that unsupervised *k*-means clustering worked well for grouping together mRNA-protein relationships. To estimate the number of underlying clusters, we used various techniques from the R package *NbClust*[20] to ensure that the groupings we found described the underlying dataset without biasing intervention. At this point, depending on the data and context, it may be prudent to refine the clusters achieved by unsupervised clustering to eliminate any gene-protein pairs that may show random patterns, which don't relate to regulatory impetus. Due to the nature of *k*-means, any such relationships will still be grouped within one of the resulting clusters. We propose ranking each gene-protein pair according to the likelihood that the pair presents a distinct pattern. This process is further detailed in **Supplementary section 1 (a and b).**

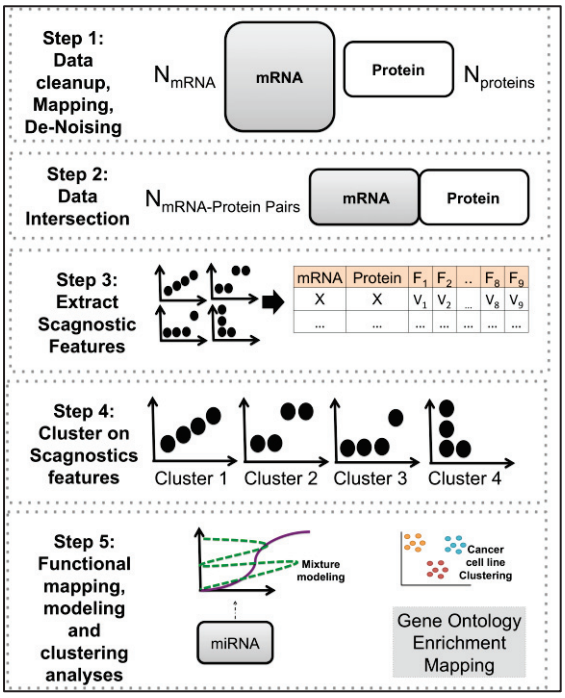

Figure 3. Five-step workflow to analyze the scagnostic feature modules in a biological context

In ***Step 5***, the resulting clusters were evaluated based on the 2D patterns indicating potential regulatory mechanisms. Depending on the type of pattern, these cluster-specific analyses included gene set enrichment analysis, miRNA target mapping, mixture modeling followed by bi-clustering, and further clustering of samples within each scatterplot.

2.2. *Validation of the workflow using synthetic data*

In order to verify that a scagnostics-based clustering approach was capable of grouping gene-protein relationships impacted by similar PTRs, we constructed a synthetic proteogenomic dataset that contained "biological" signals (**Figure 4(a)** describes the types of scatterplots that might be generated by data with these regulatory mechanisms; mRNA expression is presented on the *x*-axis and protein abundance on the *y*-axis). In **Figure 4(b)** we summarize our validation approach in which we apply the PTR explorer workflow to the synthetic data and assess generated clusters' quality using the measure of cluster purity[21].

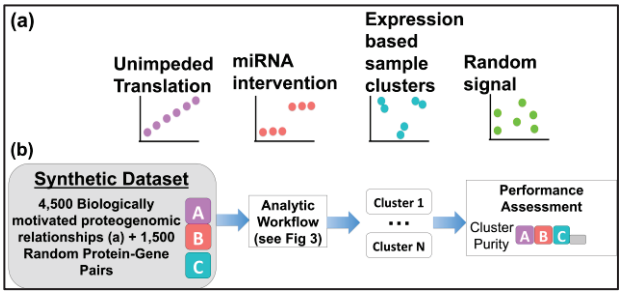

Figure 4. (a) Examples of scatterplots of patterns observed in plots of mRNA expression vs protein abundance in real-world data that were spiked into the synthetic dataset, along with unrelated relationships presenting random signal (b) Validation approach involved in employing the scagnostics based methodology on the synthetic dataset and assessing quality (cluster purity) of the resulting clusters

The synthetic dataset (**Supplementary Synthetic_data_and_processing.RData**) contained 6000 gene-protein pairs and was computationally generated from the distribution of NCI-60 transcriptomic and proteomic datasets. All the values in each of the NCI-60 datasets were aggregated and the resulting distributions were used to sample data points for each gene-protein relationship. Custom written functions, (stored within the supplementary RData file) coercing these data points to form certain types of relationships, were utilized. By generating the synthetic dataset from actual data, we ensure that the synthetic dataset retains the characteristics of real data. The signals included in this dataset are as follows, with data generation specifics in **Supplementary Section 4.**

a) Unimpeded translation: mRNA and protein abundances are highly correlated, implying efficient translation.
b) miRNA intervention: Selective active miRNA targeting resulting in bimodal protein abundance across the samples.
c) Translation rate dependent on gene expression: Sample clusters showcasing varying translation rates, or mRNA degradation, depending on different thresholds of gene expression.
d) Random: This trend consists of scatterplots with no distinguishable trend, essentially random values of mRNA expression and protein abundance paired together.

The synthetic dataset contained two matrices of 6,000 rows (genes and matched proteins), 100 columns (samples) and was spiked with three specific patterns that could arise due to PTRs ((a) through (c) above). There were 1,500 relationships for each specific pattern and 1,500 random relationships. Plotting the values from the synthetic mRNA and synthetic protein data, using matched rows in the two matrices, on each axis generated the 2D scatterplots of these four different types of signals.

Scagnostics features were extracted from the synthetic dataset above. This resulted in an eventual 9 dimensional feature vector for all the 6000 synthetic 2D scatterplots, each representing a synthetic gene-protein relationship. We further employed unsupervised *k*-means (with k=4) clustering on this feature set. The number of optimal clusters (k) was determined using the R package *NbClust*. The package evaluates 30 independent indices (including the measures of Dindex and the Hubert index[20] as referenced below in the Results section), which refer to widely known criteria, to deduce the optimum number of clusters in the underlying dataset.

To quantify the performance of our workflow we measured cluster purity (as defined in R package *IntNMF*[21]). This was feasible due to prior knowledge of the four types of relationships within the dataset. We additionally evaluated scagnostic features across increasing degrees of noise added to the synthetic data. To add or "jitter" the dataset with increasing levels of noise, a random value is added or subtracted from each data point in the synthetic dataset. This value lies in a range [$-nf*sd$, $nf*sd$] {nf = Noise Factor, sd = Standard Deviation of each dataset}. The *jitter* function in R, with increasing noise factors (0 to 2), was utilized to achieve this effect.

Lastly, to compare the performance of our workflow relative to traditional methods, we performed traditional *k*-means and hierarchical clustering (with *kmeans* and *hclust* in R) on the concatenated synthetic proteogenomic dataset to evaluate how much of the signal could be identified and clustered without using a scagnostic workflow, simply by clustering the mRNA expression and protein abundance of gene-protein pairs. Results of the same are highlighted in the

following Results section. We also employed a widely adopted integrative clustering method (*icluster*[22]), used to cluster multi-omic data. The details of the *icluster* implementation and results of this comparison are detailed in **Supplementary Section 3**.

2.3. *Application of workflow to real-world data*

After assessing the efficacy of scagnostics in the context of exploring gene-protein relationships and their regulatory mechanisms, we examined the performance of the PTR explorer on real world datasets. The aim of these experiments was (a) to test the robustness of a model built with scagnostics and (b) to extract meaningful biological results using the PTR explorer.

Firstly, to examine the robustness of a model built based on scagnostic features, we utilized the single cohort, colorectal cancer proteogenomic dataset from Zhang et al[23]. The details of preprocessing of the dataset are outlined in **Supplementary Section 2.** We aimed to verify that scagnostics features were effectively and uniquely able to characterize a biological model for a single cohort dataset, and the feature sets were largely impervious to varying sample sizes or differing samples. We evaluated the differences between scagnostics feature matrices when (a) varying sample sizes are used and (b) when different samples from the same cohort are used to generate the resulting features. The results of the same are detailed in the section below.

Secondly, to extract meaningful biological results using the scagnostics workflow, we utilized proteogenomic and miRNA data[24–27] from the NCI-60 panel of cancer cell lines (e.g., breast, ovarian, prostate)[26,28] used extensively to investigate cancer mechanisms and drug responses[19,27,29]. The details of access and preprocessing of this data are described in **Supplementary Section 2.** The final cluster specific analyses performed are noted in **Supplementary Table 1** [30–32].

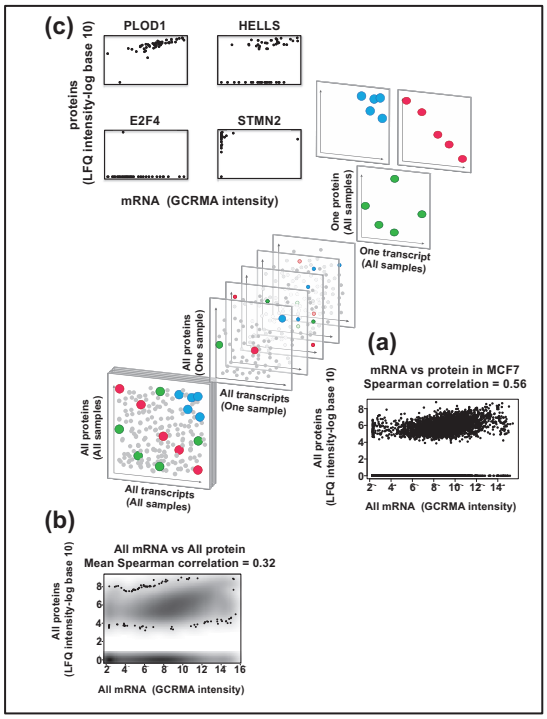

Figure 5. (a) Scatterplot of mRNA expression and protein abundance in a single NCI-60 cell line, breast cancer MCF7 cells. (b) Density plot of all mRNA expression vs. protein abundance in the NCI-60 proteogenomic dataset (c) Examples of the variety of trends we observe between mRNA expression and protein abundance analysis of data from all NCI-60 cell lines

3. Results

3.1. *Proteogenomic 2D scatterplots*

While mRNA expression and protein abundance have been studied by employing scatterplots previously, they have mostly been utilized to assess general trends across all gene-protein pairs in a single sample/cell-type or across all samples involved in an experiment. When a single scatter plot is made for all the transcripts and proteins in one sample **(Figure 5(a))** or compositing many samples (**Figure 5(b)**), the result is not conducive to examining specific post-transcriptional regulatory mechanisms affecting a specific gene's relationship with its corresponding protein. As has been previously

reported[23], the broad correlation between mRNA and protein, even within a single cell type is poor. Similarly, when examined across multiple cell types, such as the 57 cell lines from the NCI-60 cancer cell line panel[33], there is little correlation between mRNA and protein abundances (**Figure 5(b)**).

Here we instead examine a slice of the data cube, rather than a projection by looking at individual mRNA and protein measurements across multiple samples. Notably, this drastically increases the range of possible mRNA-protein scatterplots that are to be explored. Positing that these scatterplots may be indicative of potential mechanisms of PTR, we examined mRNA-protein scatterplots for individual genes across all samples from the NCI-60 panel. Several examples of these plots are shown in **Figure 5(c)**.

3.2. *Use of scagnostics to understand proteogenomic regulation*

Scagnostics are widely used to characterize 2D scatterplots effectively[12,13,16]. Extracting scagnostics features for 2D scatterplots of mRNA expression versus protein abundance, characterizes each mRNA - protein pairing and the PTR that may result in the 2D proteogenomic trend. This characterization takes into account shape (“stringy”,”convex”,”striated”), monotonicity (“monotonic”), density (“sparse”,”clumpy”), among other distinguishing attributes. We highlight how these features can potentially exemplify biological characteristics or phenomena in **Supplementary Table 2**. Scagnostics affords multiple advantages over traditional methods of analyzing proteogenomic data. First, it is an integrative technique that can easily assess two dimensions of data. Second, it reduces dimensionality to nine interpretable features regardless of the size of the dataset or dynamic ranges. Since calculation of scagnostics features requires identification of the convex hull, alpha hull and minimum spanning tree, depending on the algorithm and approximation technique for finding these geometric features, the time complexity for calculation scagnostics features may vary. However each scagnostics feature can be calculated independently for each scatterplot (gene-protein pair) in parallel, thereby ensuring swift processing of large-scale datasets.

3.3. *Scagnostics based methods are impervious to noise and uniquely characterize biological models*

We assessed changes in the scagnostic features generated from the synthetic dataset with increasing noise factors (*nf*), to judge sensitivity of scagnostics to noise. To each value in the synthetic dataset, a random value from the range [-*nf***sd* , *nf***sd*], where *sd* is standard deviation, was added or subtracted to create a new dataset. At *nf* values ranging from 0 to 1, we observe the median values for most features to be impervious to noise, whereas the numeric ranges (the minimum and maximum value of features observed) of a few features (e.g., clumpy, striated) were impacted by the addition of noise. The results are summarized in **Supplementary Figure 1**.

To gauge the robustness of scagnostics characterization, and verify how resilient the technique was to a reduction in number of samples, we employed the workflow on the colorectal cancer proteogenomic dataset from Zhang et al[23]. We assessed the changes in the scagnostics feature set when variable numbers of samples (number of samples reduced by ~25%) from this single cohort

dataset were used to generate the scatterplots and when non-identical but constant number of samples were used to do the same. Both experiments demonstrated invariance in the scagnostic feature vector regardless of the identity or number of samples utilized to generate the features (**Supplementary Figure 2 (a and b)**). The results confirm that (a) the scagnostic feature vectors uniquely characterize a biological model and (b) while discernible patterns exist, the feature set is impervious to a reduction in the number of samples in the dataset.

3.4. *Scagnostics workflow successfully isolates signals from synthetic data and out-performs traditional methods*

The scagnostics workflow (**Figure 3**) was used to extract scagnostic features from the mRNA-protein scatterplots created using the synthetic dataset (Section 2.2). Ten indices evaluated by the R package *NbClust* correctly predicted the number of clusters as four (the three spiked trends and the random relationships) in the scagnostics feature set (**Supplementary Figure 3**).

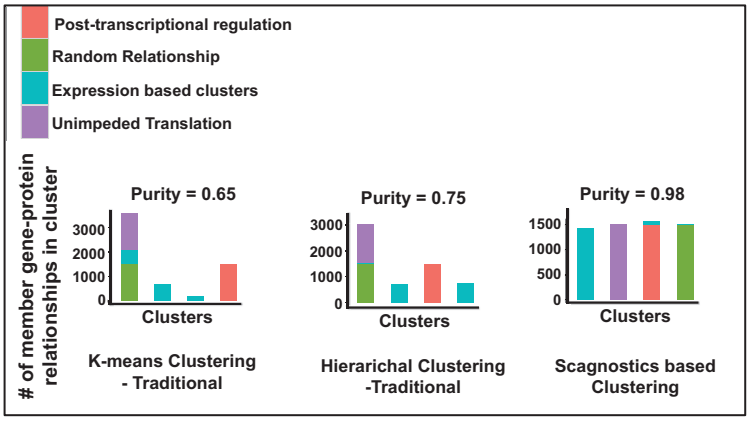

Figure 6. Comparison of cluster purity when traditional methods of hierarchal and *k*-means clustering and scagnostic feature-based clustering were employed on synthetic data

The cluster purities of the scagnostic clusters decreased as noise was increased in the synthetic dataset **(Supplementary Figure 4).** When clustering was compared across increasingly noisy datasets, we observed that the scagnostic feature-based clustering successfully isolated the known clusters till the trends were largely distinguishable.

We performed analyses to understand whether it was possible to discern similar insights about trends in the synthetic dataset using traditional methods of proteogenomic clustering. We performed traditional hierarchal clustering and *k*-means clustering (both with number of clusters equal to 4) on the synthetic dataset **(Figure 6).** For the original synthetic proteogenomic dataset sans noise (6,000 mRNA and protein abundances across 100 samples), we concatenated and clustered the data using the clustering methods above. This process of concatenating two datasets and performing clustering has been a traditional method of integrative analysis[34].

Neither hierarchical nor *k*-means clustering achieved the high cluster purity of scagnostics-based clustering (**Figure 6**). Only the very distinct expression-based trend in the post-transcriptional regulation signals was partially captured by traditional clustering methods. The other trends spiked into the synthetic dataset that were clearly visible in 2D scatterplots of mRNA-protein pairings were not captured by clustering on expression or abundance using traditional methods but were effectively identified when employing scagnostic feature clustering.

In addition to the traditional data clustering methods, the scagnostics-based clustering was more effective than employing an integrative clustering method for multi-omic data (*icluster*) on the synthetic data as well. The integrative method was unable to dissect all the relationships and presented a significantly lower purity than the scagnostics clustering (**Supplementary Section 3**).

3.5. *Application of scagnostics-based method to a real-world dataset*

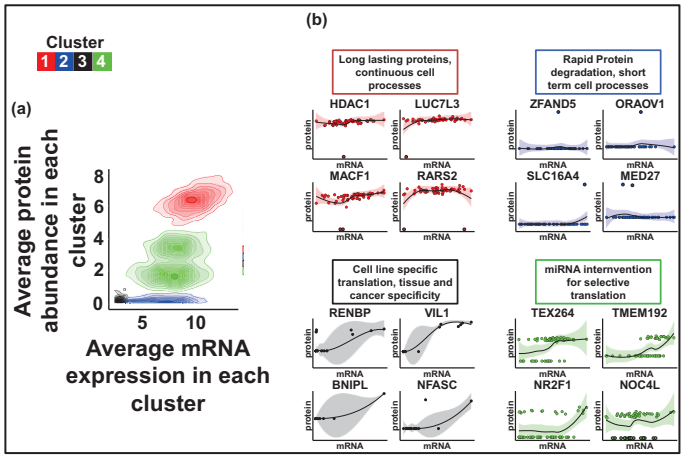

Figure 7. (a) Contour density plot of the average mRNA expression of each gene versus the average protein abundance of each protein across NCI-60 scagnostics feature driven clusters. (b) Sampling of proteogenomic relationships from each cluster.

PTR explorer followed by cluster pruning (Section 2.1) identified four groups from the NCI-60 proteogenomic pairs (**Figure 7**). Pruning reduced the size of the cluster by ~60%, and retained the most relevant proteogenomic pairs in each grouping. Cluster 1 showcases high abundance of both mRNA and protein, suggesting unimpeded translation and low levels of degradation. Cluster 2 presents moderate to high mRNA levels and close to no protein production. Suggesting the protein degradation mechanism at work as a PTR for the members of this cluster. Cluster 3 showcases sample specific protein and mRNA degradation, and upon further investigation, the cluster members predominantly showcase cancer and tissue specific behavior. And lastly, Cluster 4 mimics a pattern that results from sample selective miRNA intervention, or selective translation. Thus we observe, 4 different types of post-transcriptional regulation in effect in this pan-cancer dataset. Corroborating literature evidence and ontology/miRNA enrichment is detailed in **Supplementary Table 3**. The members of each cluster are listed in **Supplementary Folder 3**.

4. Conclusions

In this study we showed that trends and patterns unearthed by analyses of mRNA-protein expression profiles across cell types illuminate configurations and divisions in the data that are of biological relevance and significance. We developed a novel workflow, based on isolating patterns and trends in proteogenomic expression data using scatterplot diagnostics (scagnostics). Proteogenomic expression data from the NCI-60 cancer cell line panel were grouped using the scagnostic approach into four clusters, each with unique biological behaviors validated by literature examples, after cluster specific analyses. We plan to expand this workflow by building multilevel models to predict the behavior of mRNA-protein expression patterns and to build abstractions of network regulatory modules.

5. References

1. Schwanhausser B, Busse D, Li N, et al. Global quantification of mammalian gene expression control. *Nature*. 2011;473(7347):337-342. doi:10.1038/nature10098
2. Pavel AB, Sonkin D, Reddy A. Integrative modeling of multi-omics data to identify cancer drivers and infer patient-specific gene activity. *BMC Syst Biol*. 2016;10(1):16. doi:10.1186/s12918-016-0260-9
3. Payne SH. The utility of protein and mRNA correlation. *Trends Biochem Sci*. 2015;40(1):1-3. doi:10.1016/j.tibs.2014.10.010

4. Wilusz JE, Sunwoo H, Spector DL. Long noncoding RNAs: Functional surprises from the RNA world. *Genes Dev*. 2009;23(13):1494-1504. doi:10.1101/gad.1800909
5. Jansen RC, Nap J-P, Mlynárová L. Errors in genomics and proteomics. *Nat Biotechnol*. 2002;20(1):19. doi:10.1038/nbt0102-19b
6. Valencia-Sanchez MA, Liu J, Hannon GJ, Parker R. Control of translation and mRNA degradation by miRNAs and siRNAs. *Genes Dev*. 2006;20(5):515-524. doi:10.1101/gad.1399806
7. Maier T, Güell M, Serrano L. Correlation of mRNA and protein in complex biological samples. *FEBS Lett*. 2009;583(24):3966-3973. doi:10.1016/j.febslet.2009.10.036
8. Perl K, Ushakov K, Pozniak Y, et al. Reduced changes in protein compared to mRNA levels across non-proliferating tissues. *BMC Genomics*. 2017. doi:10.1186/s12864-017-3683-9
9. Yates A, Webb A, Sharpnack M, Chamberlin H, Huang K, Machiraju R. Visualizing Multidimensional Data with Glyph SPLOMs. *Comput Graph Forum*. 2014;33(3):301-310. doi:10.1111/cgf.12386
10. Krishnaswamy S, Spitzer MH, Mingueneau M, et al. Systems biology. Conditional density-based analysis of T cell signaling in single-cell data. *Science*. 2014;346(6213):1250689. doi:10.1126/science.1250689
11. Sahoo D, Dill DL, Gentles AJ, Tibshirani R, Plevritis SK. Boolean implication networks derived from large scale, whole genome microarray datasets. *Genome Biol*. 2008;9(10):R157. doi:10.1186/gb-2008-9-10-r157
12. Dang TN, Wilkinson L. Transforming Scagnostics to Reveal Hidden Features. *IEEE Trans Vis Comput Graph*. 2014;20(12):1624-1632. doi:10.1109/TVCG.2014.2346572
13. Wilkinson L, Anand A, Grossman R. Graph-theoretic scagnostics. *Proc - IEEE Symp Inf Vis INFO VIS*. 2005:157-164. doi:10.1109/INFVIS.2005.1532142
14. Wilkinson L, Anand A. scagnostics: Compute scagnostics - scatterplot diagnostics. 2012. http://cran.r-project.org/package=scagnostics.
15. Dang TN, Anand A, Wilkinson L. TimeSeer: Scagnostics for High-Dimensional Time Series. *IEEE Trans Vis Comput Graph*. 2013;19(3):470-483. doi:10.1109/TVCG.2012.128
16. Wilkinson L, Wills G. Scagnostics Distributions. *J Comput Graph Stat*. 2008;17(June 2014):473-491. doi:10.1198/106186008X320465
17. Tuan Nhon Dang, Wilkinson L. ScagExplorer: Exploring Scatterplots by Their Scagnostics. In: *2014 IEEE Pacific Visualization Symposium*. IEEE; 2014:73-80. doi:10.1109/PacificVis.2014.42
18. Dang TN, Anand A, Wilkinson L. TimeSeer: Scagnostics for high-dimensional time series. *IEEE Trans Vis Comput Graph*. 2013;19(3):470-483. doi:10.1109/TVCG.2012.128
19. Holbeck SL, Collins JM, Doroshow JH. Analysis of Food and Drug Administration-approved anticancer agents in the NCI60 panel of human tumor cell lines. *Mol Cancer Ther*. 2010;9(5):1451-1460. doi:10.1158/1535-7163.MCT-10-0106
20. Charrad M, Ghazzali N, Boiteau V, Niknafs A. NbClust : An R Package for Determining the Relevant Number of Clusters in a Data Set. *J Stat Softw*. 2014;61(6):1-36. doi:10.18637/jss.v061.i06

21. Chalise P, Raghavan R, Fridley B. IntNMF: Integrative Clustering of Multiple Genomic Dataset. 2016. http://cran.r-project.org/package=IntNMF.
22. Shen R, Olshen AB, Ladanyi M. (iCluster)Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics*. 2009;25(22):2906-2912. doi:10.1093/bioinformatics/btp543
23. Zhang B, Wang J, Wang X, et al. Proteogenomic characterization of human colon and rectal cancer. *Nature*. 2014. doi:10.1038/nature13438
24. Reinhold WC, Sunshine M, Liu H, et al. CellMiner: A Web-Based Suite of Genomic and Pharmacologic Tools to Explore Transcript and Drug Patterns in the NCI-60 Cell Line Set. *Cancer Res*. 2012;72(14):3499-3511. doi:10.1158/0008-5472.CAN-12-1370
25. Shankavaram UT, Varma S, Kane D, et al. CellMiner: a relational database and query tool for the NCI-60 cancer cell lines. *BMC Genomics*. 2009;10(1):277. doi:10.1186/1471-2164-10-277
26. Gholami AM, Hahne H, Wu Z, et al. Global Proteome Analysis of the NCI-60 Cell Line Panel. *Cell Rep*. 2013;4(3):609-620. doi:10.1016/j.celrep.2013.07.018
27. Weinstein JN. Spotlight on Molecular Profiling : Commentary Spotlight on molecular profiling : ‘“ Integromic ”’ analysis of the NCI-60 cancer cell lines. *Mol Cancer Ther*. 2006;5(November):2601-2605. doi:10.1158/1535-7163.MCT-06-0640
28. Reinhold WC, Mergny J-L, Liu H, et al. Exon Array Analyses across the NCI-60 Reveal Potential Regulation of TOP1 by Transcription Pausing at Guanosine Quartets in the First Intron. *Cancer Res*. 2010;70(6):2191-2203. doi:10.1158/0008-5472.CAN-09-3528
29. Body MR, Paull KD. Some practical considerations and applications of the national cancer institute in vitro anticancer drug discovery screen. *Drug Dev Res*. 1995;34(2):91-109. doi:10.1002/ddr.430340203
30. Filipowicz W, Bhattacharyya SN, Sonenberg N. Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? *Nat Rev Genet*. 2008;9(2):102-114. doi:10.1038/nrg2290
31. Huang DW, Lempicki RA, Sherman BT. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc*. 2009;4(1):44-57. doi:10.1038/nprot.2008.211
32. Huang DW, Sherman BT, Lempicki R a. Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res*. 2009;37(1):1-13. doi:10.1093/nar/gkn923
33. Shoemaker RH. The NCI60 human tumour cell line anticancer drug screen. *Nat Rev Cancer*. 2006;6(10):813-823. doi:10.1038/nrc1951
34. Ritchie MD, Holzinger ER, Li R, Pendergrass SA, Kim D. Methods of integrating data to uncover genotype–phenotype interactions. *Nat Rev Genet*. 2015;16(2):85-97. doi:10.1038/nrg3868