Interpreting the contributions of transcript features to protein abundance prediction models

[1,2]**Himangi Srivastava**, [1,2,*]**Edward Lau**

[1]Department of Medicine/Cardiology
[2]Consortium for Fibrosis Research and Translation
University of Colorado School of Medicine
Aurora, CO 80045

***Correspondence:**
Edward Lau, PhD, edward.lau@cuanschutz.edu

Abstract

Protein and mRNA levels correlate only moderately across samples. The availability of proteogenomics data sets with protein and transcript measurements from matching samples is providing new opportunities to assess the degree to which protein levels in a system can be predicted from mRNA information. Here we revisited the protein level prediction problem and analyzed large proteogenomics data from 8 cancer types within the Clinical Proteomic Tumor Analysis Consortium (CPTAC) data set. We trained models to predict the abundance of 13,637 proteins using matching transcriptome data from up to 958 tumor or normal adjacent tissue samples each, and compared predictive performances across algorithms and input features. As data depth increased, the incorporation of additional transcripts as input features increasingly improved the ability to predict sample protein abundance. We found widespread occurrences where the abundance of a protein is considerably less well explained by its own cognate transcript level than that of one or more trans locus transcripts, suggesting post-transcriptional regulations broadly influence protein and mRNA correlation. Transcripts that contribute to non-cognate protein abundance primarily involve those encoding known interaction partners and protein complex members of the protein of interest. Network analysis reveals a proteome-wide interdependency of protein abundance on the transcript levels of interacting proteins. Proteogenomic co-expression analysis may have utility for finding gene interactions and predicting expression changes in biological systems.

Introduction

It is well established that protein levels correlate imperfectly with the levels of their cognate transcripts [1–3]. More specifically, although a robust trend exists over the log scale between protein and mRNA measurements *across genes*, genewise correlation between proteins and their transcripts is much poorer *across observations* (i.e., samples, tissues, cell types, or subjects). In other words, abundant proteins have abundant transcripts but transcript variance within a group of samples does not necessarily predict or signify corresponding protein changes [4]. Multiple factors are known to contribute to this non-correlation, including technical reasons, as transcriptomics and proteomics data would have different sources of error. Proteins with lower baseline variance in mass spec data, presumably due to their higher abundance and detectability in mass spectrometry, are better predicted by their transcripts [5]. A substantial portion of the remaining unexplained protein variance is likely attributable to biological and biophysical regulations. It has been consistently observed across tissues that multi-protein complex membership invokes a buffer effect on protein levels [6,7], because a complex only fully folds and functions when all subunits are present, further induction of the transcript for a single subunit will not lead to additional complexes, and supernumerary proteins are quickly degraded [6,8]. Lastly, numerous post-transcriptional and post-translational mechanisms are known to modulate protein levels such as the gene- and context-dependent translation rates of mRNAs [9,10], the differential half-life and temporal distributions between mRNAs and proteins [2,11], and proteolytic degrading translated proteins in the cell.

The emergence of large-scale proteogenomics data from matching samples have created new opportunities to revisit protein-level predictions from transcriptomics data. The abundance of a protein may be the function of one or more transcripts. Most notably, available data sets from the Gene Tissue Expression (GTEx) project [7] and the Clinical Proteomic Tumor Analysis Consortium (CPTAC) [12], have spurred the use of machine learning approaches to evaluate how well one can predict protein level variance given a set of transcriptomics data, with the goal of developing strategies and algorithms that can boost the performance of protein level predictions. This culminated in a community based effort in the CPTAC Proteogenomics Dream Challenge Task 2 [12]. Perhaps surprisingly, many of the community submitted models performed no better than the baseline model, which is an elastic net taking into account all mRNA features available and has a median Pearson's correlation coefficient (r) of 0.47 for ovarian cancer, and some models performed no better than random, suggesting protein level prediction remains a challenging problem. Nevertheless, general lessons have emerged from the top performing models; for instance: (i) ensemble methods generally performed well [12–14]; (ii) combining more observations from multiple datasets that borrow information from each other led to improved predictions even if the data came from different cancer types [12]; and (iii) judicious feature pre-selection based on prior biological knowledge such as protein-protein interactions improved prediction performance [12,14]. Notwithstanding these general observations, the current literature reflects that much remains

to be learned about how best to predict protein levels from mRNA. This problem has several practical importances. Proteins carry out the majority of biological processes and hence are arguably the most relevant molecules to biological states. Despite rapid advances in proteomics however, bulk and single-cell RNA sequencing remain the most commonly used methods to interrogate gene expression status on a large scale and will likely remain so in the foreseeable future. Transcriptomics experiments often operate on the implicit assumption that identified differential regulation exert their biological effects via their cognate proteins, hence it is important to better understand the relationships between protein and mRNA levels to aid in data interpretation and determining potential protein level changes given a set of transcriptomics data. Alternatively, it would be useful to know the genewise difference in how well a gene's transcript can predict its protein counterpart as a means to filter or prioritize biologically relevant transcript signatures [15].

Here we revisit the predictability of protein levels from transcriptomics data. Since the time of the Dream Challenge, considerably more proteogenomics data have been made publicly available which increases the number of observations available for modeling training, as well as the number of proteins for which there is mass spectrometry information available. Individual CPTAC cancer studies have analyzed the protein and mRNA correlation in individual tumors and normal adjacent tissues and nominated specific pathways whose correlations are particularly poor; nevertheless, a detailed analysis remains lacking that interrogates the identity and regulatory modality of individual proteins. Accordingly, our goals here are to (1) evaluate how the availability of larger data sets from different tumors affects the performance of prediction algorithms and feature selection strategies; and (2) interpret the prediction results to identify principles regarding the architecture of protein regulation control.

Methods

Data retrieval and processing

Gene expression data were obtained from public data from the CPTAC project and included data from 8 cancer types: ovarian cancer [16], breast cancer [17], endometrial carcinoma [18], colorectal cancer [19], lung adenocarcinoma [20], clear cell renal carcinoma [21], glioblastoma [22], and lung squamous cell carcinoma [23]. The cumulative inclusions of each cancer type in the order above are sequentially referred to as CPTAC_2 to CPTAC_8 in the manuscript, such that CPTAC_2 refers to the union of ovarian and breast cancer; CPTAC_3 refers the union of ovarian, breast, and endometrial cancer; and so on. The mRNA and protein level expression data from the CPTAC cancer types was retrieved using the cptac package v.0.9.7 [24] in Python 3.9. Each column of the quantitative measurement of the transcriptomics data acted as an independent variable or feature variable whereas the normalized quantitative measurement of a particular protein of interest acted as the single

dependent or target variable in the protein model. Retrieved mRNA level gene expression data are standardized using the scikit-learn simple scaler. For each protein for which predictions are to be made, we retrieved five separate feature sets:

1. **Single:** Using only the single transcript coding for the protein of interest for model training, then running the pipeline to train a model for each protein separately.
2. **CORUM:** Using the transcripts of all proteins belonging to the same protein complex as the protein of interest, if any, in CORUM v.3.0 [25], where protein complexes are defined as two or more proteins that interact physically in a quaternary structure. These transcripts then act as independent variables (features) to predict the target variable (protein of interest). The pipeline is then run to train a model for each protein separately.
3. **STRING 800:** Using the transcripts of interacting partners of the protein of interest as input features. Interacting partners are retrieved from STRING v.11 [26], which documents functional associations including physical interactions, genetic interactions, co-expression, co-occurrence, and other associations. The STRING combined score represents the overall likelihood of interactions. Interacting pairs with a STRING combined score of 800 or above (high-confidence) are included. The pipeline is then run to train a model for each protein separately.
4. **STRING:** As above, except that interacting pairs with a STRING combined score of 200 or above (low- to high-confidence) are included. The pipeline is then run to train a model for each protein separately.
5. **Transcriptome:** A transcriptome-wide model where all qualifying transcripts in the data set are included as features, prior to the removal of low-variance features. The pipeline is then run to train a model for each protein separately.

Model training and evaluation

For each feature set, missing values for feature variables are imputed using median imputation, followed by the removal of features with variance of 0.2 or below. Models are not trained for proteins with fewer than 50 empirical observations. The data are then split 80:20 into training and test sets. No imputation or additional standardization was performed on the proteomics data. The input and target data from the training set are then used to train a model using either linear regression, elastic net with 5-fold cross validation, or random forest in scikit-learn v.1.0 [27], with the following specified parameters: random forest number of estimators: 500, criterion: squared error, max depth: elastic net cross validation L1 ratio: 0.1, 0.5, 0.9, 0.95, cv: 5, tolerance: 1e–3, max iterations: 2000. The trained models are saved as individual objects which include the predicted protein levels in the training set as well as the contribution of each feature to the overall prediction (coefficients in linear regression and elastic net; feature importance and trees in random forest), and are applied to predict protein levels in the test set data.

To evaluate model performance, the Pearson's correlation coefficients between predicted ($\hat{y}$) and actual (y) protein values are calculated individually for the train and test set data for each protein model using the numpy corrcoef function, or defaulted to 0 if the standard deviation of predicted values is 0. Goodness-of-fit (R^2) is calculated using scikit-learn.metrics.r2_score function. Normalized root mean square errors (NRMSE) are calculated as follows:

$$NRMSE = \sqrt{\frac{1}{n}\sum_{i=0}^{n-1}(y_i - \hat{y}_i)^2 * \frac{1}{(max(y)-min(y))}}$$

Network construction and analysis

To construct the protein regulation network, we first constructed an overall graph from a subset of transcript-protein relationships using the networkx package [28] in Python 3.9. Qualifying proteins are those whose prediction increased with the inclusion of more features, such that in the STRING or CORUM feature set, the correlation coefficient between the elastic net predicted and actual protein levels is greater than the single feature elastic nets by at least 0.25. Transcripts that contribute to the prediction of these proteins are therefore included if their elastic net coefficient *a* and random forest feature importance *b* are greater than a certain threshold, which was set to *a*=0.05, *b*=0.05 for the CORUM feature set and *a*=0.1, *b*=0.01 for the STRING feature set.

In the data, the list of proteins and their corresponding transcriptomics features act as the nodes and the interaction pattern is a function of the correlation value and feature importance between a protein and the transcriptomics. The direction of the edge therefore flows from transcriptomics to protein. All the text files for each protein obtained from the computational pipeline are compiled into a n overall data frame. This data frame is then converted into the form of a network data frame. In this data frame, all the protein becomes the target node, and the transcriptomics corresponding to the protein becomes the source node. Additional column of this data frame consists of the weight between a protein and its predicting transcripts, which was calculated using the sum of the [–1,1] clipped elastic net coefficient and the descending percentile rank of the random forest feature importance. An overall directed graph G(V, E) is then constructed from the edge lists with proteins as targets and their predicting transcripts as sources, such that it is a function of vertices V which depict protein and/or transcript nodes and edges E which flow in the direction from the transcript sources to the protein targets. Using the built-in functions in networkx, this overall directed graph is divided into weakly connected components, which resulted in individual subgraphs such that for the overall network G=(V, E) where V is the vertices and E is the edge then a subgraph of G = (V, E) is a graph S=(V′, E′) where vertex set V′ ⊆ V and

edge set $E' \subseteq E$ connects only nodes of V'. We then used Cytoscape v.3.9 [29] to visualize and perform additional topology analysis and functional annotations of each subgraph. HiDef persistent community detection [30] was performed with the aid of the CyCommunityDetection Cytoscape plugin [31] using the Leiden algorithm [32]. Functional enrichment of individual communities was performed using CyCommunityDetection with the g:Profiler web server [33]. Hub nodes were identified with the aid of the CytoNCA [34] Cytoscape plugin using the betweenness centrality algorithm with the top 10% of proteins with the highest centrality defined as hubs. Functional enrichment analysis on the hub proteins was performed with the aid of the stringApp [35] Cytoscape plugin using default settings and 5% false discovery rate cutoff.

Additional data analysis

Additional data analysis, statistics, and visualization were performed in R v.4.1.1 with the aid of the ReactomePA [36], clusterProfiler [37], and circlize [38] packages; and in Python 3.9 with the aid of the seaborn [39] package.

Data and code availability

The Python codes used to train the machine learning models and evaluate prediction performance have been uploaded to a containerized repository and can be accessed on CodeOcean.

Results

Feature selection improves the prediction of protein abundance from transcriptome data

We first evaluated whether increasing proteogenomics data depth would impact optimal feature selection strategies and algorithms in prediction protein levels. To do so, we retrieved up to data from up to 8 tumor types from the CPTAC data set and processed the data to collage a data table with matching protein and transcript data from each tumor sample (see Methods). We then trained models to predict the target variable (the normalized labeled mass spectrometry measured abundance of a particular protein) from various input features (transcriptomics data from the matching samples) using either a multiple linear regression, elastic net, and random forest regressor in scikit-learn. To compare the performance of prior knowledge based feature selection, we further compared supplying the models with input features based on (1) only RNA level data the corresponding transcript of the target variable protein (single feature) (2) self transcript plus the transcript of any protein within the same protein complex of the target variable protein in CORUM (CORUM feature); (3) self transcript plus the transcript of any proteins that are high-confidence protein interaction partners including physical interactions and other inferred association from STRINGdb (STRING 800 feature); (4) self transcript

plus the transcript of any proteins that are low to high- confidence protein interaction partners from STRINGdb; (5) using all qualifying transcripts (transcriptome features).

We then compared the genewise prediction performance of the transcriptome using the test set correlation coefficient between predicted and actual mass spectrometry protein values, R^2, and NRMSE as described (**Figure 1A; Supplementary Table S1**). As expected, all models performed better when given larger data sets, and elastic nets and random forest algorithms that were given multiple transcript features appeared to experience larger gains. In contrast, a multiple linear regression model performed little better with more included features and in fact failed to predict protein levels when a large number of features are given. For the elastic net and random forest models, performance gains began to saturate with additional data sets (**Figure 1A; Supplementary Figure S1**). The gain was the greatest when data set #5 (lung adenocarcinoma) was added. Overall, the data suggests that prediction improvements were due to other transcripts (number of features) rather than simply the gain in data set sizes (number of observations) or the use of algorithms (random forests) that can account for non-linearity.

This is corroborated when comparing feature sets across a single data set (CPTAC_8). Model performance continued to increase as more features were added in, and the number of non-predictors and negative predictors decreased while overall dispersion of predictability also decreased (**Supplementary Figure S2A–D**). Taking into consideration the number of features in the STRING 800 and STRING 200 sets (median feature sizes of 10 [IQR: 2–48] and 266 [IQR: 149–451], respectively) compared to the size of the whole transcriptome feature set, the results indicate that a substantial portion of performance gains over the single-transcript feature set is already achievable from from relatively few selected prior features. This is consistent with a majority of information about protein level information residing on the transcript levels of the genes encoding the protein as well as the protein association partners.

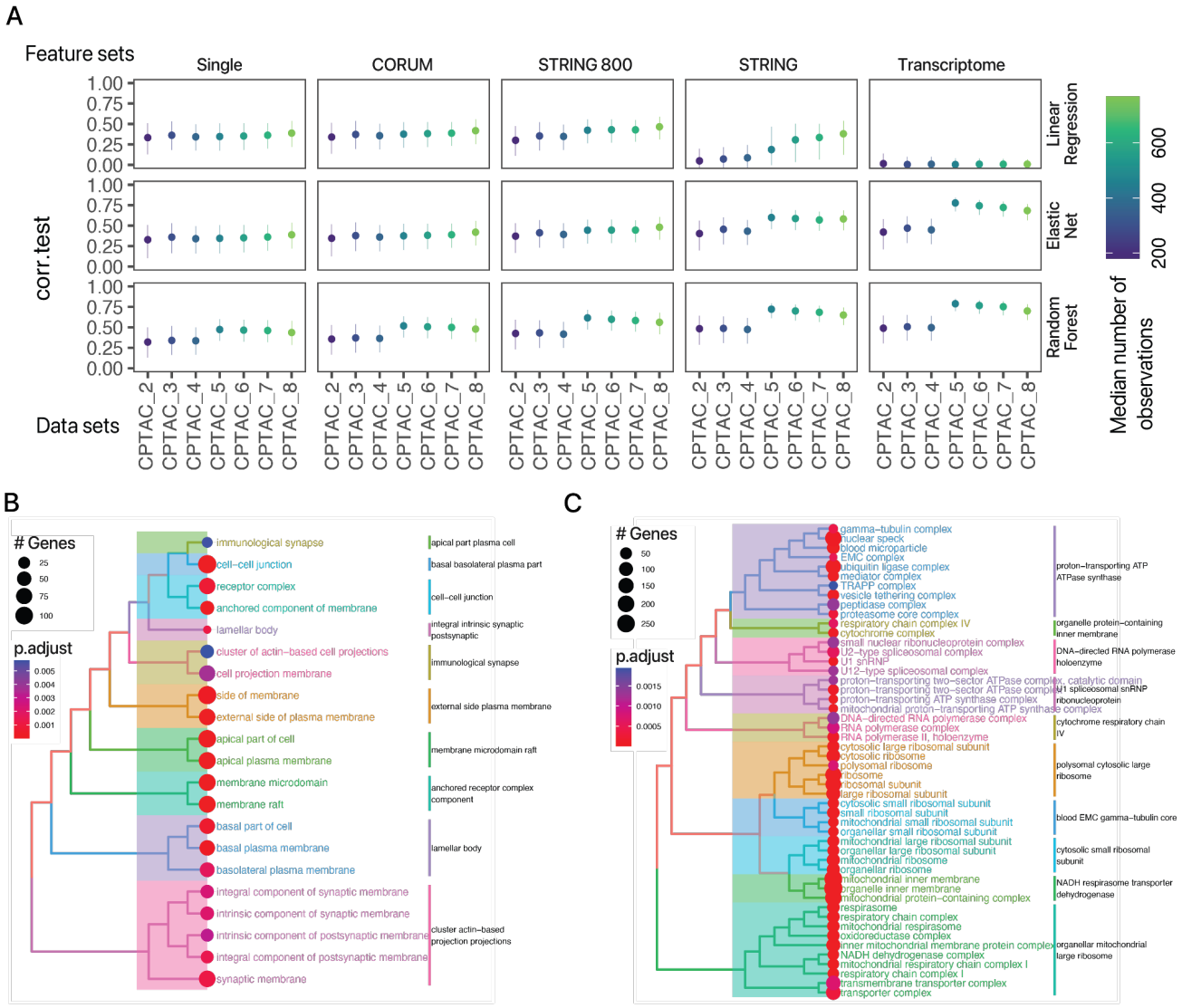


Figure 1. Genewise dispersion of protein predictability from transcriptome data. A. Point ranges (25–75th percentile) of test set correlation coefficients between the transcript-predicted and actual protein level for each protein are shown across five feature sets (column: single/self transcript, CORUM interactors, STRING high-confidence associated proteins; STRING low-confidence associated proteins, and all transcripts) and three algorithms (multiple linear regression, elastic net, and random forest). In each plot, x axis denotes the number of CPTAC data set used to train the models. **B.** Tree plots showing the clustering and relationships of gene ontology terms that are significantly enriched among proteins whose abundances are well predicted by their own transcripts ($r > 0.6$). **C.** Tree plots of terms enriched among proteins whose abundances are poorly predicted by their own transcripts ($r < 0.3$).

From the baseline self-transcript model in the CPTAC_8 data set containing 8 cancer types, we found that good protein predictors ($r \geq 0.6$) are enriched in genes participating in membrane proteins and cell junctions (**Figure 1C; Supplementary Table S2**) , whereas poor protein predictors ($r \leq 0.2$) are enriched in multiple large multi-protein complexes (**Figure 1D; Supplementary Table S3**). Hence the reanalysis of CPTAC_8 data set further confirms prior observations that protein complex membership presents a major source for transcript-protein non-correlation. We then compared the STRING feature elastic net model to the self transcript model and asked what explains the performance gains at a more granular level. Considering the 13 239 genes with protein-level prediction and with at least one STRING interactor, on average, each gene in the STRING model has a median of 273 [158–457] features, compared to 1 feature in the self transcript model. Incorporation of STRING features led to an average increase in test set correlation coefficients of 0.16 [IQR: 0.07–0.27] (**Supplementary Table S4**). There is a significant positive correlation between the number of features and prediction improvements (Pearson's r 0.22, P value < 2.2e–16) (**Figure 2A**).

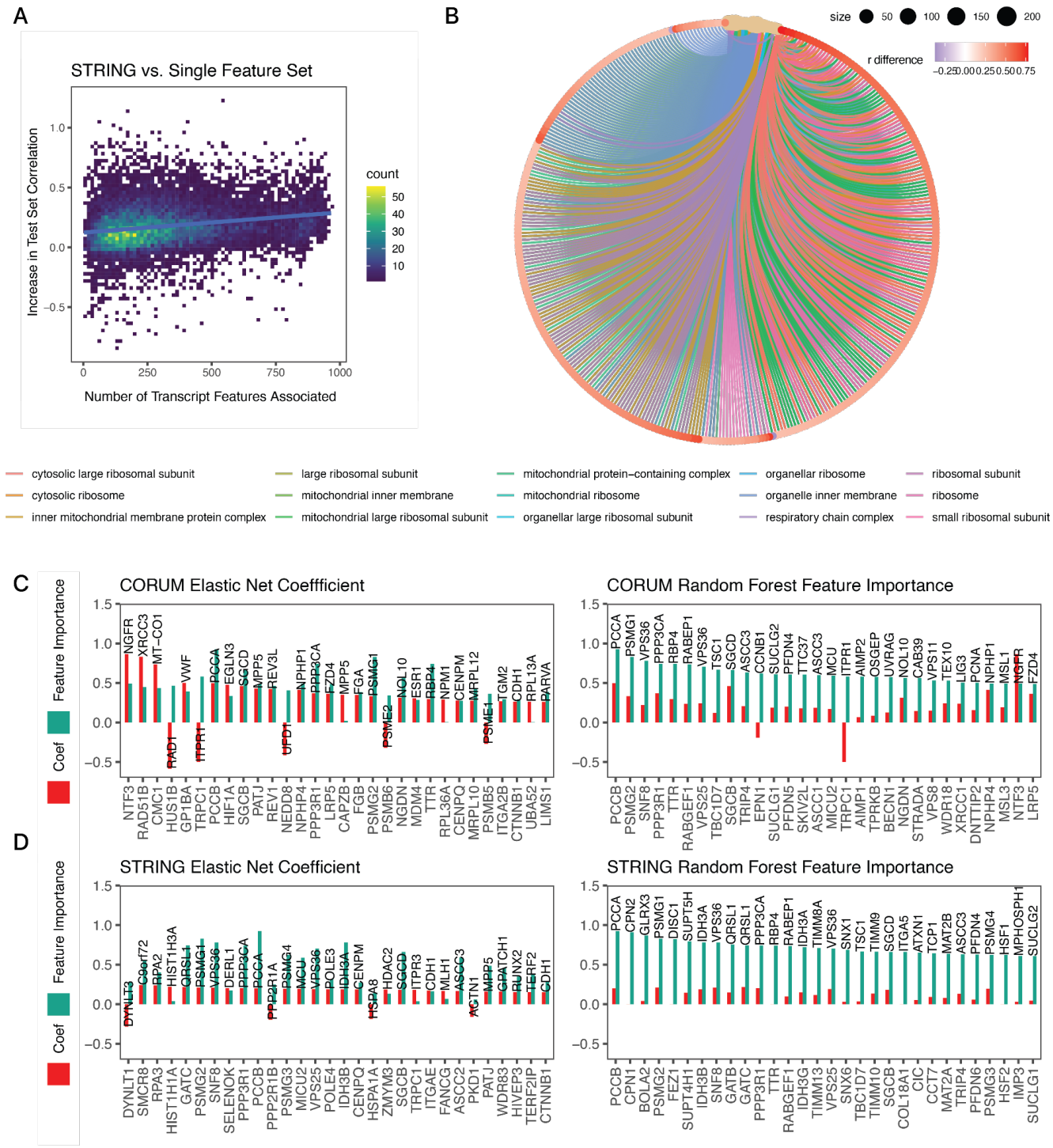


Figure 2. Improvements to protein prediction from the incorporation of additional transcript features. A. Scatterplot showing a significant linear relationship between the number of low-confidence interactors a protein has as annotated in STRING vs. its increase in test set correlation between predicted vs. actual protein levels in the Elastic Net models over the self-transcript feature set. **B.** Proteins with substantial increase in prediction performance are involved in

large multi-protein complexes including the ribosome, mitochondrial ribosome, and respiratory chain complexes. **C and D.** Bar charts showing 30 proteins (x axis) in the **C.** CORUM and **D.** STRING feature sets that are poorly predicted by their own transcripts but had a different transcript with high model coefficient (Coef; in the elastic nets; left) or feature importance (in the random forests, right) toward their protein levels. In each category, the top 30 proteins with the highest coefficient or feature importance contributed by another transcript are shown. The transcripts with the highest coefficients or feature importance for each protein are labeled on top of the bar.

We then queried the functional enrichment of proteins with strong improvements in prediction (difference in STRING vs. self-feature in predicted-actual correlation coefficient ≥ 0.25) and found that they were again highly enriched in large multi-protein complexes including the ribosome, mitochondrial ribosome, RNA polymerase, and spliceosome (**Figure 2B; Supplementary Table S5**). This suggests that the non-correlation between transcripts and proteins of multi-protein complexes levels can be partially 'rescued' by considering the information of interacting partner transcripts, which could participate in post-transcriptional regulations such as by being the stoichiometrically limiting transcript or crucial assembly parts, such that the transcript level of the lowest abundance protein can regulate the protein level of other subunits.

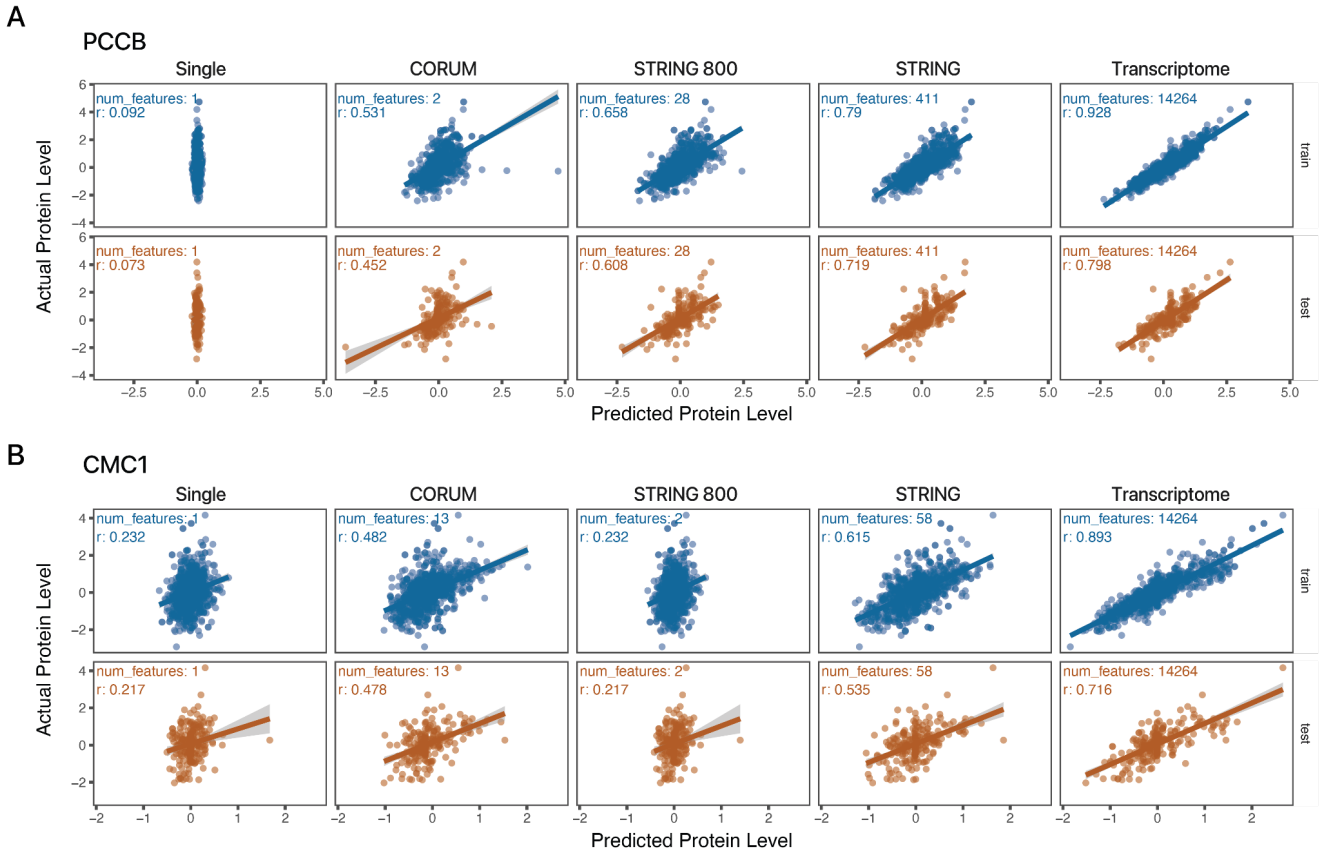


Figure 3. Proteins with improved predicted levels after inclusion of additional transcript features. A. PCCB and **B.** CMC1. For each protein, the transcript-trained prediction of protein level is plotted on the x axis and the actual protein level is plotted on the y axis. Blue: train set, brown: test set. Columns denote the transcript feature set used to train the model. The number of features used to train the model in each feature set is shown inside each plot. r: Correlation coefficient.

Proteins whose abundances are predicted by non-cognate transcripts are common

To investigate the underlying details behind the performance gains of protein predictions, we examined the genes whose protein prediction performance from transcripts increased substantially (test set correlation coefficient with additional features – correlation coefficient of self transcript >= 0.25) after the incorporation of additional transcript features. In total, we observed 434 such proteins in the CORUM feature sets, and 2692 genes from the STRING data set, suggesting non-cognate transcript contributions to protein level are common at the proteome level. Upon inspecting the model feature coefficients (elastic nets) and feature importance (random forests), we observed that a considerable portion of these proteins with improved prediction are associated with (1) poor contribution from the self-transcript, and (2) a substantial contribution from only a few non-cognate (i.e., trans locus) transcripts coding for other proteins (**Figure 2C**). In other words, although model performance continued to increase with increased feature set sizes, the contributions of trans locus transcript features to overall prediction performance is highly unevenly distributed, and are therefore attributable to a few high contribution genes rather than a simple scaling with feature size. Remarkably, despite the STRING feature set being substantially larger than the CORUM feature set, all the top 30 proteins with substantial contributions from trans locus transcripts in the STRING feature set involve transcripts encoding proteins that form part of the same CORUM complex as the protein of interest itself, suggesting that stable complex memberships play an outsized role in determining protein levels.

We next interrogated two proteins where a non-self transcript has an outsized effect on the protein level of the proteins in the CORUM and STRING feature sets, PCCB and CMC1. PCCB forms the propionyl-CoA carboxylase complex with two protein members, which breaks down certain amino acids in the cell (**Figure 3A**). CMC1 is an assembly factor that forms an early intermediate of the cytochrome c oxidase complex in the mitochondrion with 14 documented protein members in CORUM (**Figure 3B**). Both PCCB and CMC1 showed poor or moderate prediction by their own self-transcripts (r: 0.073 and 0.217, respectively) but experienced

high gains in test set correlation upon incorporating additional transcript features in the CORUM (r: 0.452 and 0.478, respectively) and STRING (r: 0.719 and 0.535) models

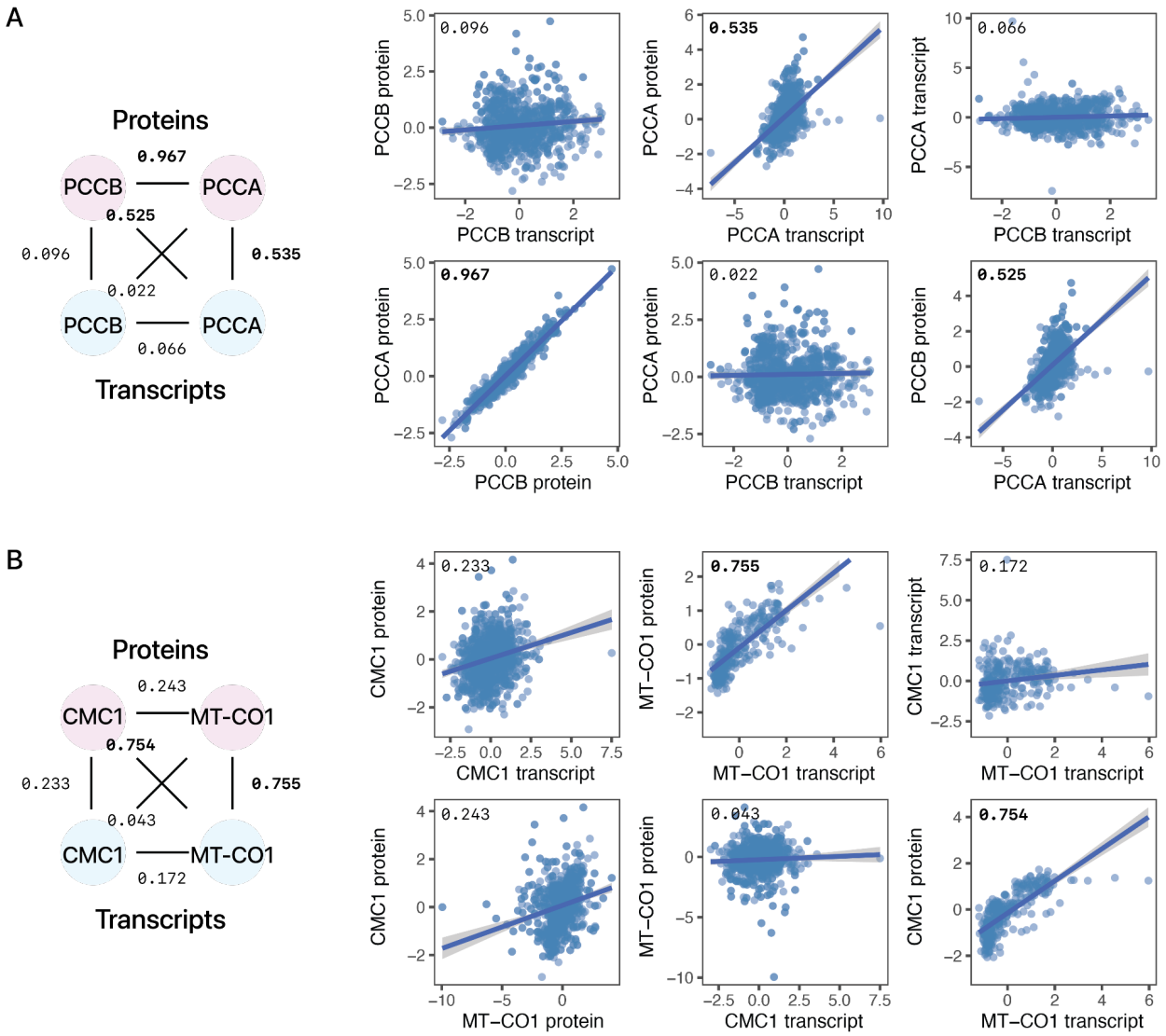


Figure 4: Cross-omics co-expression of functionally associated proteins. Two examples of proteins whose abundance is better explained by another transcript are shown. **A.** PCCB protein level is predicted by PCCA transcript but not its own transcript. **B.** CMC1 protein level is explained by MT-CO1 transcript level but not its own transcript. Substantial correlations across transcripts and proteins (≥ 0.4) are bolded.

We then considered the transcriptomics and proteomics data distributions of the features that had highest coefficients in the elastic net model from the CORUM feature set for PCCB and CMC1 . For PCCB, most protein-level predictions are recovered when the feature set includes PCCA, which together with PCCB forms the stable propionyl-CoA carboxylase enzyme that consists of six copies of PCCB and six copies of PCCA

each. There is a remarkably strong correlation with its propionyl-CoA carboxylase complex interacting partner PCCA at the protein level (r: 0.967), but this correlation is almost entirely absent at the transcript level (r: 0.066). This is partially explainable by the observation that PCCA transcript has a higher level of variability and is also strongly correlated with the PCCB protein level. The total protein level of PCCB is therefore primarily driven by PCCA rather than PCCB transcripts (**Figure 4A**).

In the case of CMC1, we found that MT-CO1 has the highest contribution to CMC1 protein level (**Figure 4B**). MT-CO1 is the mitochondrial genome encoded subunit of complex IV that is thought to act as the nascent scaffold around which the complex is assembled. CMC1 binds to MT-CO1 during the formation of the early module known as MITRAC and is subsequently released during assembly. [40,41]. CMC1 transcript and protein levels are moderately correlated with one another (r: 0.233). Remarkably, CMC1 and MT-CO1 show only moderate co-expression on either the transcript (r: 0.172) or protein (r: 0.243) levels, but MT-CO1 transcript is robustly correlated with the CMC1 protein level (r: 0.754), which suggests the possibility that proteogenomic co-expression may reveal additional functionally related protein pairs. Closer inspection suggests a non-linear relationship between MT-CO1 transcript and CMC1 protein where upon reaching a plateau, further increases in MT-CO1 transcript does not correlate to further increase in CMC1 protein levels, suggesting other transcripts may also contribute to CMC1 protein levels. Consistent with this, other complex IV subunits also have non-negligible coefficients (in the elastic nets) or feature importance (in the random forests) in the CMC1 model. When the feature set expanded to include all STRING proteins, MT-CO1 remained a high contributor in the random forest but not the elastic net model, suggesting the elastic nets may be more susceptible to collinear features than the random forests.

Two other examples of small stable complexes with interdependent protein and mRNA correlation are shown in **Supplementary Figure S3.** Proteasome Assembly Chaperone 2 (PSMG2) forms the heterodimer PAC-PAC4 complex (2 protein members in CORUM) along with Proteasome Assembly Chaperone 1 (PSMG1), which binds with 20S proteasome precursors and acts as a scaffold to promote proteasome ring assembly [42]. PSMG2 and PSMG1 are co-expressed strongly at the protein level (r: 0.743) but only moderate at the transcript level (0.273) (**Supplementary Figure S3A)**. PSMG2 protein has a higher correlation with PSMG1 transcript (r: 0.488) than its own (r: 0.136). From GTEx v8 data, PSMG2 is expressed more highly than PSMG1 across tissues (TPM ~40 vs. 15) [43], which suggests PSMG1 may act as a limiting factor in heterodimer formation. In another example, Smith-Magenis Syndrome Chromosomal Region Candidate Gene 8 (SMCR8) forms the heterotrimer C9orf72-SMCR8-WDR41 complex (3 protein members in CORUM) together with the C9orf72-SMCR8 Complex Subunit (C9orf72) and WD Repeat Domain 41 (WDR41) [44]. The C9orf72-SMCR8 complex modulates autophagy, likely by modulating the maturation of autophagosomes [45].

Mutations in the genes encoding the C9orf72-SMCR8 Complex are broadly implicated in diseases including amyotrophic lateral sclerosis and frontotemporal dementia. As in the other cases, we observed a co-expression of SMCR8 and C9orf72 at the protein level (0.632) but not the transcript level (–0.012), and SMCR8 protein levels correlate more strongly with the C9orf72 transcript (r: 0.436) than the SMCR8 transcript (r:0.231) (**Supplementary Figure S3B**).

Taken together, these observations exemplify widespread interdependency of protein levels on trans locus transcripts beyond very large metabolic and gene expression complexes, involving diverse biological processes in the cell. The CORUM feature set considered here alone contains over 2,700 human complexes derived from CORUM release 3.0 mappable to 3,689 proteins with distinct gene names, with a median of 3 proteins per complex (**Supplementary Figure S4**). Thus, small stable protein complexes are widespread and a large fraction of the proteome could be placed under complex post-transcriptional control that decouples protein levels from mRNA levels. Moreover, half of the proteins in the CORUM annotations belong to 2 or more distinct complexes (including intermediates and subcomplexes) which suggest further opportunities for more complex patterns of interdependent protein and mRNA levels. In some examined cases including PSMG2, the transcript coding for the protein of action is more abundant than that of its regulating interacting partner, which is compatible with a scenario where simple stoichiometric constraints control the protein level of the supernumerary subunit. It is however less clear whether this scenario applies to other examined cases including CMC1 and MT-CO1, where the MT-CO1 transcript is much more abundant than CMC1 owing to the multiple copies of mitochondrial genome in the tissue. Hence, how MT-CO1 regulates CMC1 protein level is not explainable directly from transcript stoichiometry alone and awaits further mechanistic studies.

Proteogenomic networks recapitulate the interdependencies of protein level regulations

To visualize the pattern of connections between transcript and protein levels on a proteome scale, we constructed graphical models that represent the interactions between each protein and the transcriptome features that contribute to its predicted level. Network analysis provides information about the nature and degree of interactions between individual protein and transcript levels and enables inferences into underpinning details. Topological analyses can further be applied to extract important nodes and significant patterns in the graphs to identify biological insights.

We first limited the graphs to proteins whose prediction improved following the inclusion of additional features. To find these proteins, we compared the results of computational models for the single feature dataset and the multiple feature dataset. From the CPTAC_8 data set, we obtained multiple correlation coefficients for each protein when the elastic net models are trained on the self transcript feature dataset and also trained on multiple

feature models (where the features are selected using the CORUM or STRING data). These correlation coefficients obtained for each protein from the single feature or multiple features using elastic nets were compared. If the coefficient of correlation for a particular target protein obtained using multiple feature data transformation and computational modeling was found to be more than the coefficient of correlation obtained for that particular target protein from the single feature data transformation and computational modeling by a certain threshold (0.25), that particular target protein is considered to meet the criteria to be included in the network model, and its saved scikit-learn models from both the elastic nets and random forests are exported as flat text files that contain the coefficients (for elastic nets) or feature importances (for random forests) of each feature for that protein in the CORUM or STRING models. A directed graph is then generated using a list of edges that connect feature variables (transcripts) to target variables (proteins), with the edge weights calculated using a function of the elastic net coefficients and random forest feature importance of the models for both the CORUM (**Supplementary Table S6**) and the STRING (**Supplementary Table S7**) feature sets. The resulting network is partitioned into connected components of transcript-protein relationships. **Figure 5** represents several example subgraphs from the CORUM feature set that contain only a few nodes, where a blue node is a gene that is represented as proteins or proteins and transcripts, whereas an orange node is a gene represented in the graph as transcript only. The edges are directed from transcript to protein, and the widths of the edges are scaled by the weight. Orange edges represent positive coefficients, and gray edges negative.

The subgraph that corresponds to the relationship between PCCA and PCCB as discussed above shows the PCCB node contributing (as transcript) to its own protein level weakly, and the PCCA node contributing strongly to the PCCB node (as protein) (**Figure 5A**). The relationship of CMC1 with other cytochrome c oxidase complex subunits, as described above, is represented in a wheel and spokes pattern where multiple transcripts (orange nodes) contribute either positively or negatively to the CMC1 protein level, with MT-CO1 having the strongest positive contributions (**Figure 5B**). Another member of the MITRAC module COA3 to which CMC1 binds likewise exerts a positive influence on the protein level of CMC1, whereas two subunits of the MT-CO2 assembly module (MT-CO2 and COX16) had a minor negative contribution to CMC1. Taken together, this is consistent with the possibility that interdependent protein and transcript relationships may reflect information related to protein complex assembly sequences.

Larger subgraphs are also observed where two or more protein target nodes are connected by shared transcript source nodes. For instance, in the actin related protein 2/3 complex, the transcripts of ARPC3 and ACTR3 regulates their respective protein levels reciprocally, and both are under the influence of the transcripts of ARPC2, and others (**Figure 5C**). The transcript level of ARPC3 appears to be negatively correlated with its

own protein level. In the CCR4-NOT transcription complex, the transcript level of CNOT1 drives the protein levels of both DDX6 and CNOT10 (**Figure 5D**). Due to the method by which the graphs are generated, protein hubs (with high in-degrees) are more prevalent. However, transcripts with high outflow activities are also seen. Two or more protein nodes may be connected when joined by a common source (transcript) node. A larger subgraph is shown where the transcript levels of RIN1, RAF1, and GNG2 both contribute to the predicted protein levels of multiple genes involved in the SNARE complex, the RIN1-STAM2-HRS complex, the AURKA-HRAS-RAF1 complex, the G protein complex, the GNAI1-GNB1-GNG3 complex, and other annotated complexes (**Figure 5E**).

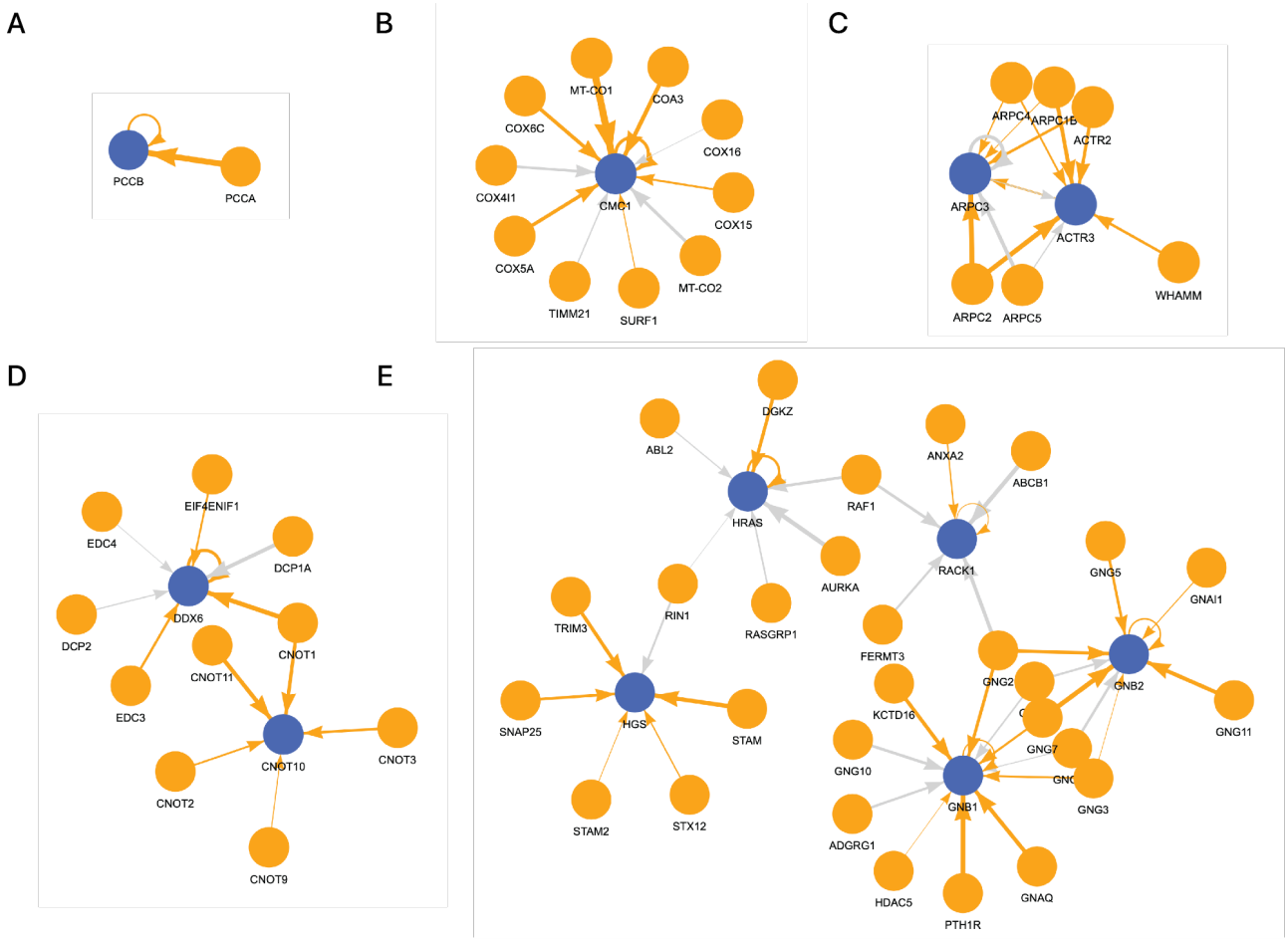


Figure 5: Example network representations of protein and transcript interrelationships. Examples of directed graphs constructed from genome-wide relationships of transcript-predicted proteins are shown. The subgraphs contain members of **A.** the propionyl-CoA carboxylase complex; **B.** the cytochrome c oxidase, mitochondrial complex; **C.** the Arp2/3 protein complex and the ARP3-WHAMM complex; **D.** the CCR4-NOT complex, the 4E-T-DDX6-CNOT1 complex, and the decapping complex; and **E**. the SNARE complex, the RIN1-STAM2-HRS complex, the AURKA-HRAS-RAF1 complex, the G protein complex, the GNAI1-GNB1-GNG3 complex, and others. In each panel, edges flow from transcript

to proteins. Orange nodes have outflow edges only (i.e., they are contributing transcripts in the prediction models). Blue nodes are nodes that are connected to other nodes via at least one inflow edge (i.e., they represent proteins, and optionally also transcripts if they also have outward edges). Orange edges represent positive coefficients of the transcripts to the target proteins in the elastic net models; gray edges represent negative coefficients.

We next considered a highly connected subgraph that represents the mitochondrial ribosome (**Figure 6A**). The machine learning results suggest that the abundances of a majority of complex subunit proteins are each contingent upon multiple transcripts. To help interpret the structure of the subgraph, we performed topological network analysis based on the hierarchical community detection framework (HiDeF) method [30] to identify persistent communities within the subgraph. Functional enrichment of subclusters reveal an expected hierarchical relationship between the 28S small and 39S large subunits in forming the 55S mitochondrial ribosome, suggesting the preferential connections within the proteogenomic graphs broadly recapitulate subcomplex assembly (**Figure 6B**). In parallel, we extracted the hub nodes from the subgraph using the betweenness centrality algorithm (**Figure 6C**). Notably, the analysis revealed mitochondrial serine beta-lactamase-like protein (LACTB) to be an upstream gene that is highly connected within the subgraph and exerts a regulatory effect on both small and large mitochondrial ribosome subunit components. Of note, LACTB was previously identified as a new subunit (MRP-L56) of the mitochondrial ribosomal large subunit isolated from sucrose gradients [46]. Subsequent work however has established the LACTB protein as a filament-forming protein that is localized to the intermembrane space instead away from the mitochondrial ribosomes [47] that possesses in vitro protease activity [48] and that acts as a tumor suppressor by maintaining post-mitotic differentiation states [48]. The physiological function and antiproliferative mechanism of LACTB are incompletely understood, although induced LACTB expression in cancer cells was found to affect mitochondrial phospholipid metabolism [48]. The analysis here suggests the possibility that LACTB may also help maintain preference for oxidative phosphorylation in differentiated cells by promoting mitochondrial ribosome biogenesis and protein synthesis.

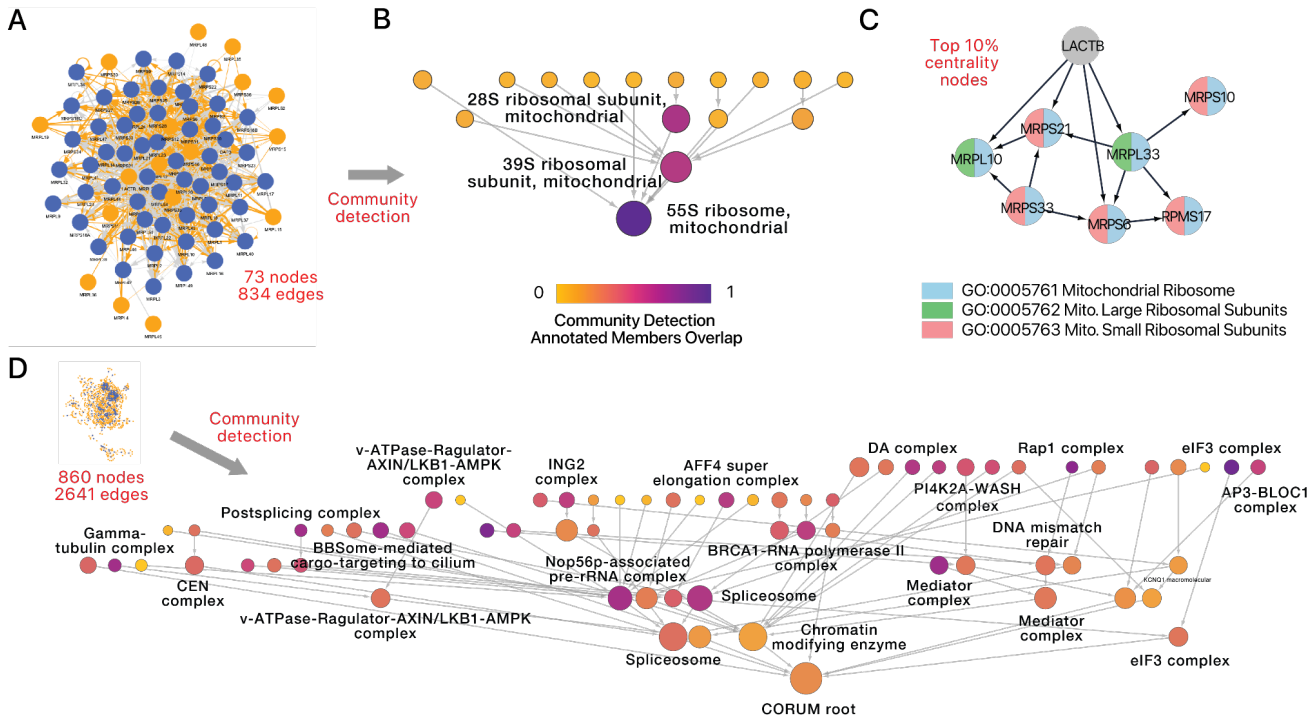


Figure 6: Topological network analysis of protein and mRNA correlations. A. A highly connected subgraph of mitochondrial ribosome subunits containing 73 nodes and 834 edges. **B.** Persistent community detection and network representation of preferential node connections, showing a hierarchical relationship between the 28S and 39S subcomplex with the assembled 55S mitochondrial ribosome. **C.** Network representation of hub nodes defined as 10% of nodes ranked by betweenness centrality, which predicts a potential role of LACTB as a critical hub that lies upstream of multiple large and small mitochondrial ribosomal protein subunits. Node colors represent the pie chart diagram of the corresponding GO biological process described in the table. **D.** The largest subgraph from the CORUM feature set contains 860 nodes and 2641 edges spanning multiple protein complexes. Community detection reveals a complex hierarchy between protein and mRNA levels across individual complexes.

Finally, larger subgraphs exist that connect multiple protein complexes and pathways, including the largest subgraph generated from the CORUM feature set that contains 860 nodes and 2,641 edges, and the largest subgraph in the STRING feature set that contains 9,444 nodes and 35,652 edges, both of which containing proteins belonging to multiple distinct cellular compartments and multi-protein complexes. In this case, although complex memberships of the involved proteins and transcripts are determined by the annotation sources used (i.e., CORUM and STRING), higher network structures such as the proximity and directionality between nodes and communities would be contingent upon protein and mRNA correlation from the proteogenomics datasets. Topological network analysis of the largest CORUM feature set subgraph shows that individual protein complex members within the subgraph are preferentially connected with each other

(**Figure 6D**). Moreover, a hierarchical relationship of the detected communities exists that connects individual multi-protein complexes, where for example, the community enriched in DNA mismatch repair is upstream of the mediator complex. Overall, the network topology hints at higher level interdependencies where the overall protein level of one multi-protein complex may be influenced by that of one or more other complexes and additional proteins in the proteome.

Discussion

Widespread contributions of trans locus transcripts to protein level

Building on the computational tasks of predicting across-sample protein levels using transcriptomics data, we examined cases where the transcript levels of interacting partners had an outsized contribution to the abundance of proteins of interest. Although the concept of proteins being modulated by their interacting partners is not novel per se, to our knowledge the prevalence of these regulations have not been examined at a proteome wide level. The results here suggest that protein expression control by other transcripts is widespread, involving not only large megadalton sized multi-protein complexes as previously observed, but also smaller stable complexes (e.g., propionyl-CoA carboxylase with two subunits) and potentially more transient interactions documented in the STRING feature set. The common presence of small stable complexes in the proteome therefore greatly expands the repertoire of proteins for which due considerations should be given when directly interpreting transcript level data as representing protein level information. The analysis here therefore reveals new details into the gene identity and modality of trans regulation of protein levels, and gives support to further development of prior transcript feature selection strategies to optimize protein prediction tasks.

The imperfect correlation between protein and mRNA points to orthogonal information that exists between transcript and protein regulations, which underpins the untapped potential for further multi-omics integration to derive new insights [49]. The use of protein correlations to find evidence for post-transcriptional regulations has previously been explored in other contexts, which looked for anti-correlation between E3 ubiquitin ligases with known proteins or ubiquitination sites of interest, which may control their protein level by virtue of post-translational degradation [17,23]. Extending this intuition towards multi-omics correlation, we used directed graphs generated from the model coefficients and feature importance of trans locus transcripts to represent the interaction patterns between the transcripts and proteins of interest. This in turn nominated a number of hub proteins whose abundance is contributed by multiple transcripts and hub transcripts that regulate multiple proteins, and enabled community detection analysis to find the relationships between

biological processes in protein and mRNA correlation networks. As more data continue to become available, we foresee that graphical models will be useful for finding more trans locus regulations of protein levels, such as those that represent known assembly sequential steps or post-transcriptional regulators. These graphical models can be used to generate testable hypotheses. For instance, in the case of the actin related protein 2/3 complex, one might expect that within a certain concentration range an overexpression of *ACTR3* and *ARPC4* will be more effective in modulating ARPC3 protein levels than augmenting the expression of *ARPC3* itself.

Limitations of the study

Significant differences in protein regulations likely occur in different cell and tissue types. The predictive models here are trained using publicly available CPTAC data from 8 cancer types, which contain transcriptome and proteome data from both tumor and normal adjacent tissue samples. In prior work, we found general concordance between protein and mRNA correlation in CPTAC samples vs. GTEx tissue proteomics data [15]. However, other studies that compared protein and mRNA correlation in tumors and normal adjacent tissues have found higher inter-sample correlation in tumor samples [20,21], which may be attributable to the increased translation rates in cancer. Hence, additional discordant cases between proteins and mRNA likely remain to be discovered that are omitted here. Sample difference may also explain the observation from GTEx proteomics that secreted proteins are associated with protein and mRNA discordance [7] as one might intuitively assume but which is not apparent in the analysis here. The cases used for training the model are not labeled by their cancer type or their tumor vs. normal adjacent tissue designation, which would likely have availed overall predictive performance.

We performed the analysis using expression data from TCGA/CPTAC RNA sequencing experiments through the cptac Python API, which retrieves the final data tables from the flagship CPTAC papers of each individual cancer type [24]. Although each TCGA/CPTAC cancer subtype project follows an overall consistent experimental design and data acquisition strategy, minute differences exist in the processing pipelines used to analyze the RNA sequencing data (e.g., STAR vs. Bowtie2) and gene expression measure (e.g., RPKM vs. FPKM) which could bias gene expression values across cancer types. Likewise, there are subtle differences in the protein expression data in each cancer type (e.g., search engine, isotope tags). As our goal here is not to compare trends across cancer types, we have taken inspiration from the CPTAC Dream Challenge submissions that improved overall predictions by borrowing information across cancer types. Future work may take advantage of harmonized processing pipelines which may further improve performance and reduce bias, such as using data from the TCGA PanCanAtlas which processed all TCGA cancer RNA sequencing data uniformly [50]. To our knowledge, uniformly processed protein data are not yet available at the time of writing.

Finally, other machine learning and deep learning algorithms can likely further boost protein level predictions. We have limited our scope here to comparisons of different feature sets and previously employed algorithms, and to result interpretation.

Conclusion

In summary, this study presents predictive models of protein levels using different transcript feature sets, and provides interpretations of the results by highlighting trans locus transcripts with substantial contributions to protein levels. With further refinement of methods and availability of large data sets, we foresee that this and similar approaches will provide valuable new insights into post-transcriptional mechanisms of protein regulations.

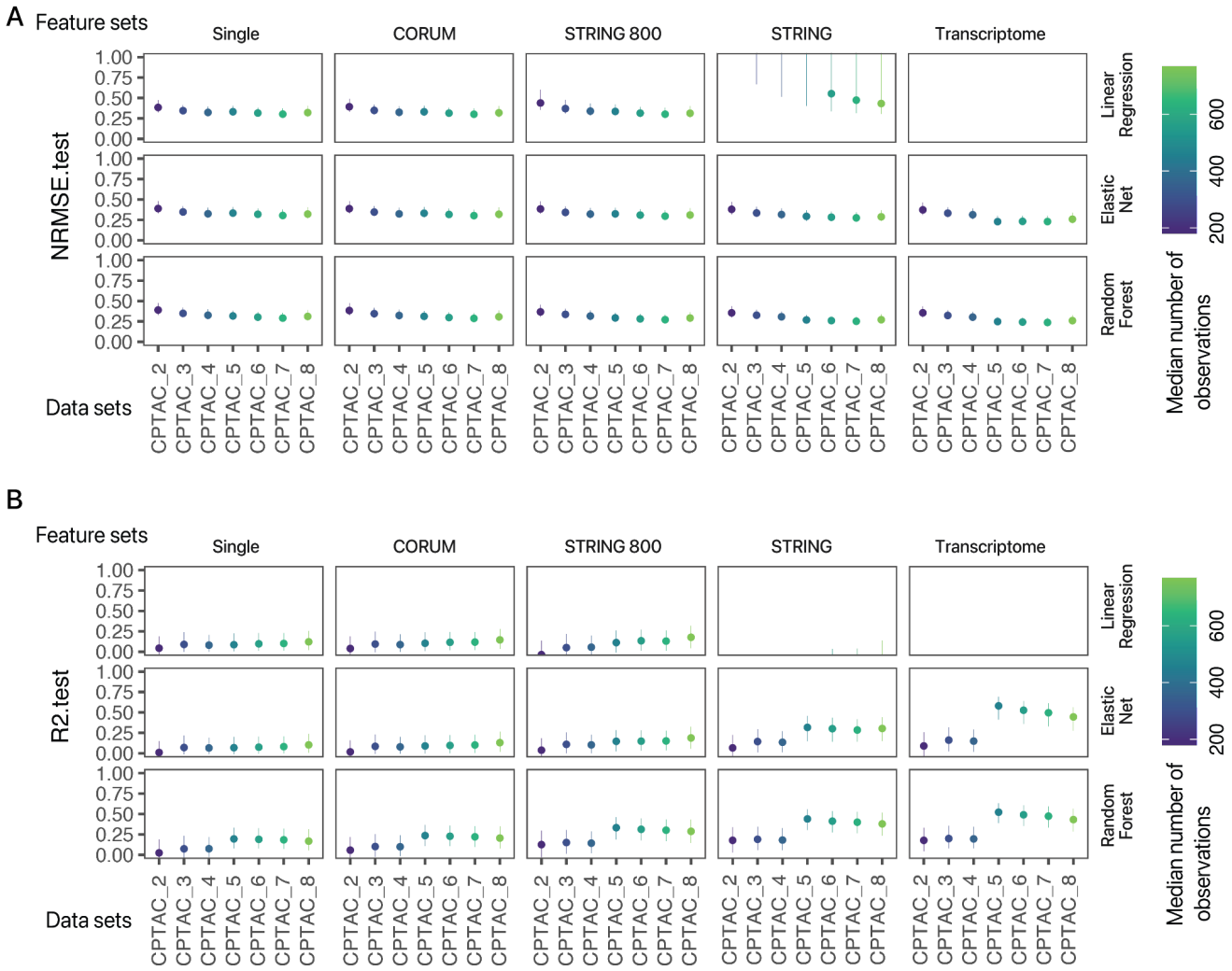
Author Contributions

HS and EL performed the analysis and drafted the manuscript.

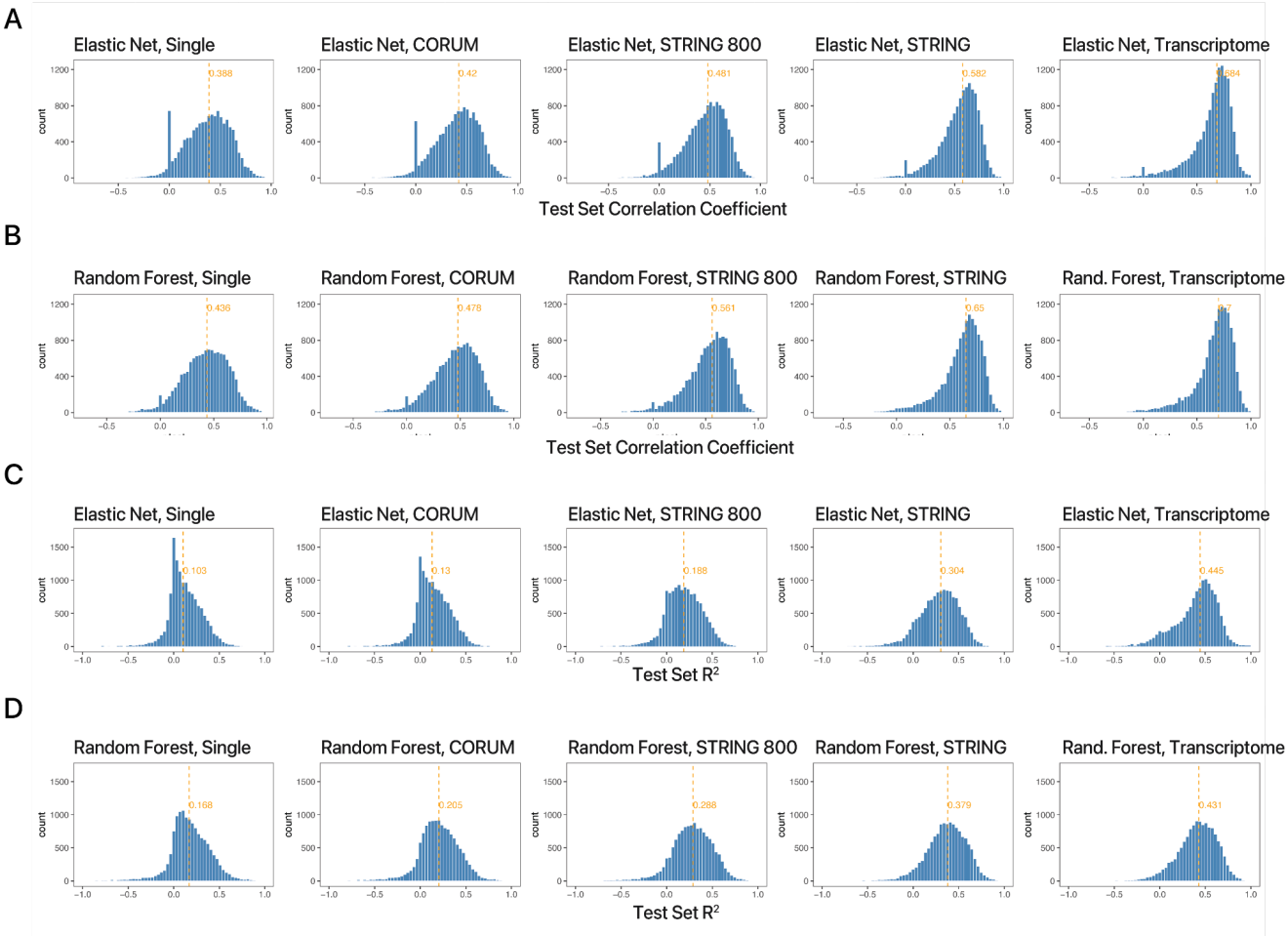
Acknowledgments

This work was supported in part by NIH Office of the Director award R03-OD032666 and NIH/NHLBI award R00-HL144829 to E.L.

Supplementary Figures

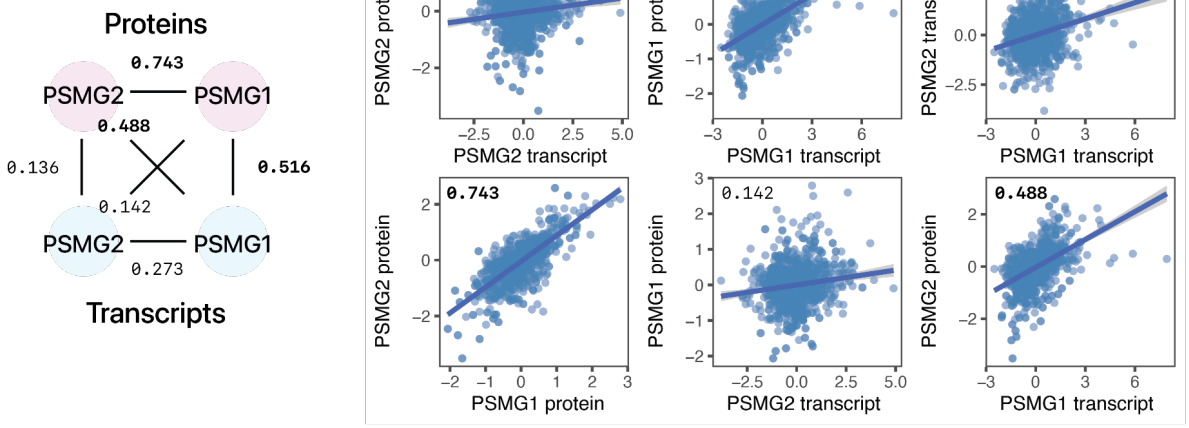


Supplementary Figure S1: Model performance measured by additional metrics. **A.** Point ranges (25–75th percentile) of test set normalized root mean square error (NRMSE) between the transcript-predicted and actual protein level for each protein are shown across five feature sets (column: single/self transcript, CORUM interactors, STRING high-confidence associated proteins; STRING low-confidence associated proteins, and all transcripts) and three algorithms (multiple linear regression, elastic net, and random forest). In each plot, x axis denotes the number of CPTAC data set used to train the models. **B.** As above, but for test set goodness-of-fit (R^2).

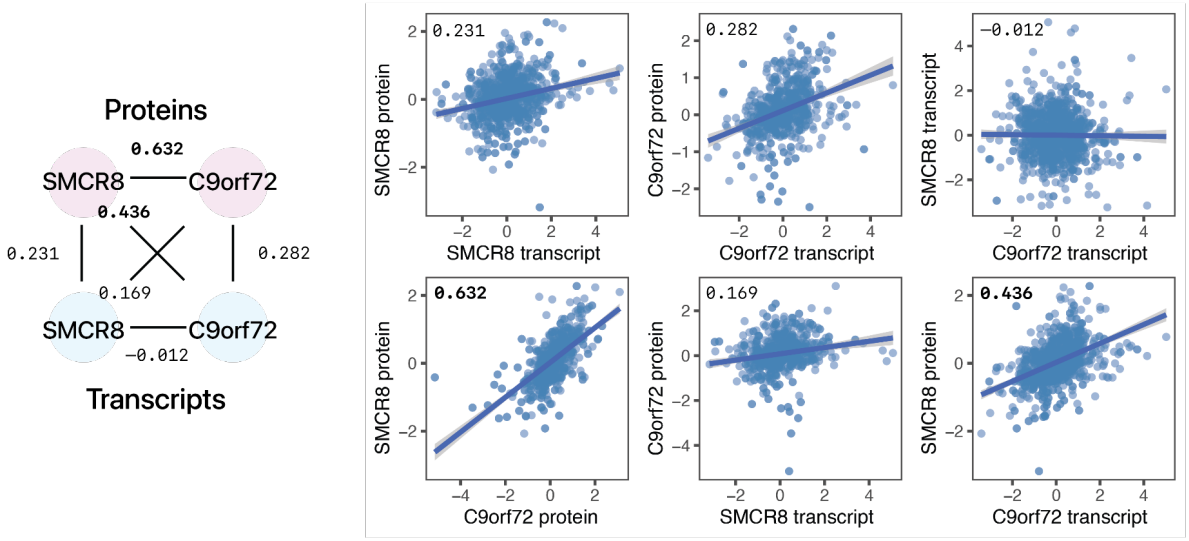


Supplementary Figure S2: Histograms of the test set correlation coefficients between the transcript-predicted and actual protein level in the CPTAC_8 data set across different feature selection strategies. **A.** The test set correlation coefficients (r) between predicted and actual protein levels in the elastic net models. **B.** The test set correlation coefficients (r) between predicted and actual protein levels in the random forest models. **C.** The test set goodness-of-fit (R^2) between predicted and actual protein levels in the elastic net models. **D.** The test set goodness-of-fit (R^2) between predicted and actual protein levels in the random forest models.

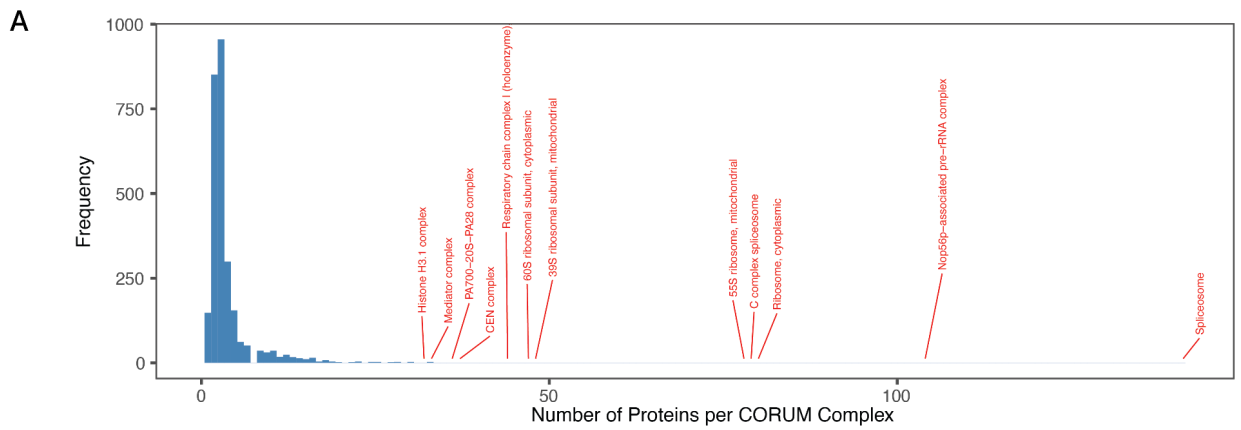
A



B



Supplementary Figure S3: Cross-omics co-expression of functionally associated proteins. Two examples of proteins whose abundance is better explained by another transcript are shown. **A.** PSMG2 protein level is better explained by PSMG1 transcript than its own transcript. **B.** SMCR8 protein level is better explained by C9orf72 transcript level than its own transcript. Substantial correlations across transcripts and proteins (≥ 0.4) are bolded.



Supplementary Figure S4: Histogram showing the distribution of complex sizes in the annotated feature set derived from CORUM v.3.0. The largest complexes are labeled. The majority of complexes are small with a median of 3 proteins per complex.

Supplementary Tables

Supplementary tables can be accessed on figshare at: https://doi.org/10.6084/m9.figshare.19330541.v1

Supplementary Table S1: Table listing performance metrics (test set correlation coefficients, R^2, NRMSE) of the tested models, feature sets, and data sets.

Supplementary Table S2: Table listing Gene Ontology cellular compartment term enrichment results for proteins that are well predicted by their transcript levels.

Supplementary Table S3: Table listing Gene Ontology cellular compartment term enrichment for proteins that are poorly predicted by their transcript levels.

Supplementary Table S4: Table listing individual genewise model performance metrics (test set correlation coefficients, R^2, NRMSE) in the CPTAC_8 data set using elastic net models on different feature sets.

Supplementary Table S5: Table listing Gene Ontology cellular compartment term enrichment for proteins with substantial improvements in prediction performance upon incorporating STRING features in elastic net models.

Supplementary Table S6: Network edge list for proteins with improved predictions in the CORUM feature set along with the coefficients and feature importances of contributing transcripts.

Supplementary Table S7: Network edge list for proteins with improved predictions in the STRING feature set along with the coefficients and feature importances of contributing transcripts.

References

(1) Gygi, S. P.; Rochon, Y.; Franza, B. R.; Aebersold, R. Correlation between Protein and MRNA Abundance in Yeast. *Mol. Cell. Biol.* **1999**, *19* (3), 1720–1730. https://doi.org/10.1128/MCB.19.3.1720.

(2) Liu, Y.; Beyer, A.; Aebersold, R. On the Dependency of Cellular Protein Levels on MRNA Abundance. *Cell* **2016**, *165* (3), 535–550. https://doi.org/10.1016/j.cell.2016.03.014.

(3) Vogel, C.; Marcotte, E. M. Insights into the Regulation of Protein Abundance from Proteomic and Transcriptomic Analyses. *Nat. Rev. Genet.* **2012**, *13* (4), 227–232. https://doi.org/10.1038/nrg3185.

(4) Franks, A.; Airoldi, E.; Slavov, N. Post-Transcriptional Regulation across Human Tissues. *PLoS Comput. Biol.* **2017**, *13* (5), e1005535. https://doi.org/10.1371/journal.pcbi.1005535.

(5) Upadhya, S. R.; Ryan, C. J. *Experimental Reproducibility Limits the Correlation between MRNA and Protein Abundances in Tumour Proteomic Profiles*; preprint; Systems Biology, 2021. https://doi.org/10.1101/2021.09.22.461108.

(6) An, H.; Harper, J. W. Ribosome Abundance Control Via the Ubiquitin–Proteasome System and Autophagy. *J. Mol. Biol.* **2020**, *432* (1), 170–184. https://doi.org/10.1016/j.jmb.2019.06.001.

(7) Jiang, L.; Wang, M.; Lin, S.; Jian, R.; Li, X.; Chan, J.; Dong, G.; Fang, H.; Robinson, A. E.; GTEx Consortium; Snyder, M. P. A Quantitative Proteome Map of the Human Body. *Cell* **2020**, *183* (1), 269-283.e19. https://doi.org/10.1016/j.cell.2020.08.036.

(8) Gonçalves, E.; Fragoulis, A.; Garcia-Alonso, L.; Cramer, T.; Saez-Rodriguez, J.; Beltrao, P. Widespread Post-Transcriptional Attenuation of Genomic Copy-Number Variation in Cancer. *Cell Syst.* **2017**, *5* (4), 386-398.e4. https://doi.org/10.1016/j.cels.2017.08.013.

(9) Eraslan, B.; Wang, D.; Gusic, M.; Prokisch, H.; Hallström, B. M.; Uhlén, M.; Asplund, A.; Pontén, F.; Wieland, T.; Hopf, T.; Hahne, H.; Kuster, B.; Gagneur, J. Quantification and Discovery of Sequence Determinants of Protein-per-MRNA Amount in 29 Human Tissues. *Mol. Syst. Biol.* **2019**, *15* (2), e8513. https://doi.org/10.15252/msb.20188513.

(10) Wang, D.; Eraslan, B.; Wieland, T.; Hallström, B.; Hopf, T.; Zolg, D. P.; Zecha, J.; Asplund, A.; Li, L.; Meng, C.; Frejno, M.; Schmidt, T.; Schnatbaum, K.; Wilhelm, M.; Ponten, F.; Uhlen, M.; Gagneur, J.; Hahne, H.; Kuster, B. A Deep Proteome and Transcriptome Abundance Atlas of 29 Healthy Human Tissues. *Mol. Syst. Biol.* **2019**, *15* (2).

https://doi.org/10.15252/msb.20188503.

(11) Waldbauer, J. R.; Rodrigue, S.; Coleman, M. L.; Chisholm, S. W. Transcriptome and Proteome Dynamics of a Light-Dark Synchronized Bacterial Cell Cycle. *PloS One* **2012**, *7* (8), e43432. https://doi.org/10.1371/journal.pone.0043432.

(12) Yang, M.; Petralia, F.; Li, Z.; Li, H.; Ma, W.; Song, X.; Kim, S.; Lee, H.; Yu, H.; Lee, B.; Bae, S.; Heo, E.; Kaczmarczyk, J.; Stępniak, P.; Warchoł, M.; Yu, T.; Calinawan, A. P.; Boutros, P. C.; Payne, S. H.; Reva, B.; NCI-CPTAC-DREAM Consortium; Boja, E.; Rodriguez, H.; Stolovitzky, G.; Guan, Y.; Kang, J.; Wang, P.; Fenyö, D.; Saez-Rodriguez, J. Community Assessment of the Predictability of Cancer Protein and Phosphoprotein Levels from Genomics and Transcriptomics. *Cell Syst.* **2020**, *11* (2), 186-195.e9. https://doi.org/10.1016/j.cels.2020.06.013.

(13) Eicher, T.; Patt, A.; Kautto, E.; Machiraju, R.; Mathé, E.; Zhang, Y. Challenges in Proteogenomics: A Comparison of Analysis Methods with the Case Study of the DREAM Proteogenomics Sub-Challenge. *BMC Bioinformatics* **2019**, *20* (Suppl 24), 669. https://doi.org/10.1186/s12859-019-3253-z.

(14) Li, H.; Siddiqui, O.; Zhang, H.; Guan, Y. Joint Learning Improves Protein Abundance Prediction in Cancers. *BMC Biol.* **2019**, *17* (1), 107. https://doi.org/10.1186/s12915-019-0730-9.

(15) Han, Y.; Li, L. Z.; Kastury, N. L.; Thomas, C. T.; Lam, M. P. Y.; Lau, E. Transcriptome Features of Striated Muscle Aging and Predictability of Protein Level Changes. *Mol. Omics* **2021**. https://doi.org/10.1039/d1mo00178g.

(16) Zhang, H.; Liu, T.; Zhang, Z.; Payne, S. H.; Zhang, B.; McDermott, J. E.; Zhou, J.-Y.; Petyuk, V. A.; Chen, L.; Ray, D.; Sun, S.; Yang, F.; Chen, L.; Wang, J.; Shah, P.; Cha, S. W.; Aiyetan, P.; Woo, S.; Tian, Y.; Gritsenko, M. A.; Clauss, T. R.; Choi, C.; Monroe, M. E.; Thomas, S.; Nie, S.; Wu, C.; Moore, R. J.; Yu, K.-H.; Tabb, D. L.; Fenyö, D.; Bafna, V.; Wang, Y.; Rodriguez, H.; Boja, E. S.; Hiltke, T.; Rivers, R. C.; Sokoll, L.; Zhu, H.; Shih, I.-M.; Cope, L.; Pandey, A.; Zhang, B.; Snyder, M. P.; Levine, D. A.; Smith, R. D.; Chan, D. W.; Rodland, K. D.; CPTAC Investigators. Integrated Proteogenomic Characterization of Human High-Grade Serous Ovarian Cancer. *Cell* **2016**, *166* (3), 755–765. https://doi.org/10.1016/j.cell.2016.05.069.

(17) Mertins, P.; Mani, D. R.; Ruggles, K. V.; Gillette, M. A.; Clauser, K. R.; Wang, P.; Wang, X.; Qiao, J. W.; Cao, S.; Petralia, F.; Kawaler, E.; Mundt, F.; Krug, K.; Tu, Z.; Lei, J. T.; Gatza, M. L.; Wilkerson, M.; Perou, C. M.; Yellapantula, V.; Huang, K.; Lin, C.; McLellan, M. D.; Yan, P.; Davies, S. R.; Townsend, R. R.; Skates, S. J.; Wang, J.; Zhang, B.; Kinsinger, C. R.; Mesri, M.; Rodriguez, H.; Ding, L.; Paulovich, A. G.; Fenyö, D.; Ellis, M. J.; Carr, S. A.; NCI CPTAC. Proteogenomics Connects Somatic Mutations to Signalling in Breast Cancer. *Nature* **2016**, *534*

(7605), 55–62. https://doi.org/10.1038/nature18003.

(18) Dou, Y.; Kawaler, E. A.; Cui Zhou, D.; Gritsenko, M. A.; Huang, C.; Blumenberg, L.; Karpova, A.; Petyuk, V. A.; Savage, S. R.; Satpathy, S.; Liu, W.; Wu, Y.; Tsai, C.-F.; Wen, B.; Li, Z.; Cao, S.; Moon, J.; Shi, Z.; Cornwell, M.; Wyczalkowski, M. A.; Chu, R. K.; Vasaikar, S.; Zhou, H.; Gao, Q.; Moore, R. J.; Li, K.; Sethuraman, S.; Monroe, M. E.; Zhao, R.; Heiman, D.; Krug, K.; Clauser, K.; Kothadia, R.; Maruvka, Y.; Pico, A. R.; Oliphant, A. E.; Hoskins, E. L.; Pugh, S. L.; Beecroft, S. J. I.; Adams, D. W.; Jarman, J. C.; Kong, A.; Chang, H.-Y.; Reva, B.; Liao, Y.; Rykunov, D.; Colaprico, A.; Chen, X. S.; Czekański, A.; Jędryka, M.; Matkowski, R.; Wiznerowicz, M.; Hiltke, T.; Boja, E.; Kinsinger, C. R.; Mesri, M.; Robles, A. I.; Rodriguez, H.; Mutch, D.; Fuh, K.; Ellis, M. J.; DeLair, D.; Thiagarajan, M.; Mani, D. R.; Getz, G.; Noble, M.; Nesvizhskii, A. I.; Wang, P.; Anderson, M. L.; Levine, D. A.; Smith, R. D.; Payne, S. H.; Ruggles, K. V.; Rodland, K. D.; Ding, L.; Zhang, B.; Liu, T.; Fenyö, D.; Agarwal, A.; Anurag, M.; Avtonomov, D.; Birger, C.; Birrer, M. J.; Boca, S. M.; Bocik, W. E.; Borate, U.; Borucki, M.; Burke, M. C.; Cai, S.; Calinawan, A.; Carr, S. A.; Carter, S.; Castro, P.; Cerda, S.; Chaikin, M.; Chan, D. W.; Chan, D.; Charamut, A.; Chen, F.; Chen, J.; Chen, L.; Chen, L. S.; Chesla, D.; Chheda, M. G.; Chinnaiyan, A. M.; Chowdhury, S.; Cieslik, M. P.; Clark, D. J.; Cottingham, S.; Culpepper, H.; Day, J.; De Young, S.; Demir, E.; Dhanasekaran, S. M.; Dhir, R.; Domagalski, M. J.; Dottino, P.; Druker, B.; Duffy, E.; Dyer, M.; Edwards, N. J.; Edwards, R.; Elburn, K.; Field, J. B.; Francis, A.; Gabriel, S.; Geffen, Y.; Geiszler, D.; Gillette, M. A.; Godwin, A. K.; Grady, P.; Hannick, L.; Hariharan, P.; Hilsenbeck, S.; Hindenach, B.; Hoadley, K. A.; Hong, R.; Hostetter, G.; Hsieh, J. J.; Hu, Y.; Ittmann, M. M.; Jaehnig, E.; Jewell, S. D.; Ji, J.; Jones, C. D.; Karabon, R.; Ketchum, K. A.; Khan, M.; Kim, B.-J.; Krek, A.; Krubit, T.; Kumar-Sinha, C.; Leprevost, F. D.; Lewis, M.; Li, Q. K.; Li, Y.; Liu, H.; Lubinski, J.; Ma, W.; Madan, R.; Malc, E.; Malovannaya, A.; Mareedu, S.; Markey, S. P.; Marrero-Oliveras, A.; Martignetti, J.; McDermott, J.; McGarvey, P. B.; McGee, J.; Mieczkowski, P.; Modugno, F.; Montgomery, R.; Newton, C. J.; Omenn, G. S.; Paulovich, A. G.; Perou, A. M.; Petralia, F.; Piehowski, P.; Polonskaya, L.; Qi, L.; Richey, S.; Robinson, K.; Roche, N.; Rohrer, D. C.; Schadt, E. E.; Schnaubelt, M.; Shi, Y.; Skelly, T.; Sokoll, L. J.; Song, X.; Stein, S. E.; Suh, J.; Tan, D.; Tansil, D.; Teo, G. C.; Thangudu, R. R.; Tognon, C.; Traer, E.; Tyner, J.; Um, K. S.; Valley, D. R.; Vatanian, N.; Vats, P.; Velvulou, U.; Vernon, M.; Wang, L.-B.; Wang, Y.; Webster, A.; Westbrook, T.; Wheeler, D.; Whiteaker, J. R.; Wilson, G. D.; Zakhartsev, Y.; Zelt, R.; Zhang, H.; Zhang, Y.; Zhang, Z.; Zhao, G. Proteogenomic Characterization of Endometrial Carcinoma. *Cell* **2020**, *180* (4), 729-748.e26. https://doi.org/10.1016/j.cell.2020.01.026.

(19) Vasaikar, S.; Huang, C.; Wang, X.; Petyuk, V. A.; Savage, S. R.; Wen, B.; Dou, Y.; Zhang, Y.;

Shi, Z.; Arshad, O. A.; Gritsenko, M. A.; Zimmerman, L. J.; McDermott, J. E.; Clauss, T. R.; Moore, R. J.; Zhao, R.; Monroe, M. E.; Wang, Y.-T.; Chambers, M. C.; Slebos, R. J. C.; Lau, K. S.; Mo, Q.; Ding, L.; Ellis, M.; Thiagarajan, M.; Kinsinger, C. R.; Rodriguez, H.; Smith, R. D.; Rodland, K. D.; Liebler, D. C.; Liu, T.; Zhang, B.; Clinical Proteomic Tumor Analysis Consortium. Proteogenomic Analysis of Human Colon Cancer Reveals New Therapeutic Opportunities. *Cell* **2019**, *177* (4), 1035-1049.e19. https://doi.org/10.1016/j.cell.2019.03.030.

(20) Gillette, M. A.; Satpathy, S.; Cao, S.; Dhanasekaran, S. M.; Vasaikar, S. V.; Krug, K.; Petralia, F.; Li, Y.; Liang, W.-W.; Reva, B.; Krek, A.; Ji, J.; Song, X.; Liu, W.; Hong, R.; Yao, L.; Blumenberg, L.; Savage, S. R.; Wendl, M. C.; Wen, B.; Li, K.; Tang, L. C.; MacMullan, M. A.; Avanessian, S. C.; Kane, M. H.; Newton, C. J.; Cornwell, M.; Kothadia, R. B.; Ma, W.; Yoo, S.; Mannan, R.; Vats, P.; Kumar-Sinha, C.; Kawaler, E. A.; Omelchenko, T.; Colaprico, A.; Geffen, Y.; Maruvka, Y. E.; da Veiga Leprevost, F.; Wiznerowicz, M.; Gümüş, Z. H.; Veluswamy, R. R.; Hostetter, G.; Heiman, D. I.; Wyczalkowski, M. A.; Hiltke, T.; Mesri, M.; Kinsinger, C. R.; Boja, E. S.; Omenn, G. S.; Chinnaiyan, A. M.; Rodriguez, H.; Li, Q. K.; Jewell, S. D.; Thiagarajan, M.; Getz, G.; Zhang, B.; Fenyö, D.; Ruggles, K. V.; Cieslik, M. P.; Robles, A. I.; Clauser, K. R.; Govindan, R.; Wang, P.; Nesvizhskii, A. I.; Ding, L.; Mani, D. R.; Carr, S. A.; Clinical Proteomic Tumor Analysis Consortium. Proteogenomic Characterization Reveals Therapeutic Vulnerabilities in Lung Adenocarcinoma. *Cell* **2020**, *182* (1), 200-225.e35. https://doi.org/10.1016/j.cell.2020.06.013.

(21) Clark, D. J.; Dhanasekaran, S. M.; Petralia, F.; Pan, J.; Song, X.; Hu, Y.; da Veiga Leprevost, F.; Reva, B.; Lih, T.-S. M.; Chang, H.-Y.; Ma, W.; Huang, C.; Ricketts, C. J.; Chen, L.; Krek, A.; Li, Y.; Rykunov, D.; Li, Q. K.; Chen, L. S.; Ozbek, U.; Vasaikar, S.; Wu, Y.; Yoo, S.; Chowdhury, S.; Wyczalkowski, M. A.; Ji, J.; Schnaubelt, M.; Kong, A.; Sethuraman, S.; Avtonomov, D. M.; Ao, M.; Colaprico, A.; Cao, S.; Cho, K.-C.; Kalayci, S.; Ma, S.; Liu, W.; Ruggles, K.; Calinawan, A.; Gümüş, Z. H.; Geiszler, D.; Kawaler, E.; Teo, G. C.; Wen, B.; Zhang, Y.; Keegan, S.; Li, K.; Chen, F.; Edwards, N.; Pierorazio, P. M.; Chen, X. S.; Pavlovich, C. P.; Hakimi, A. A.; Brominski, G.; Hsieh, J. J.; Antczak, A.; Omelchenko, T.; Lubinski, J.; Wiznerowicz, M.; Linehan, W. M.; Kinsinger, C. R.; Thiagarajan, M.; Boja, E. S.; Mesri, M.; Hiltke, T.; Robles, A. I.; Rodriguez, H.; Qian, J.; Fenyö, D.; Zhang, B.; Ding, L.; Schadt, E.; Chinnaiyan, A. M.; Zhang, Z.; Omenn, G. S.; Cieslik, M.; Chan, D. W.; Nesvizhskii, A. I.; Wang, P.; Zhang, H.; Clinical Proteomic Tumor Analysis Consortium. Integrated Proteogenomic Characterization of Clear Cell Renal Cell Carcinoma. *Cell* **2019**, *179* (4), 964-983.e31. https://doi.org/10.1016/j.cell.2019.10.007.

(22) Wang, L.-B.; Karpova, A.; Gritsenko, M. A.; Kyle, J. E.; Cao, S.; Li, Y.; Rykunov, D.; Colaprico,

A.; Rothstein, J. H.; Hong, R.; Stathias, V.; Cornwell, M.; Petralia, F.; Wu, Y.; Reva, B.; Krug, K.; Pugliese, P.; Kawaler, E.; Olsen, L. K.; Liang, W.-W.; Song, X.; Dou, Y.; Wendl, M. C.; Caravan, W.; Liu, W.; Cui Zhou, D.; Ji, J.; Tsai, C.-F.; Petyuk, V. A.; Moon, J.; Ma, W.; Chu, R. K.; Weitz, K. K.; Moore, R. J.; Monroe, M. E.; Zhao, R.; Yang, X.; Yoo, S.; Krek, A.; Demopoulos, A.; Zhu, H.; Wyczalkowski, M. A.; McMichael, J. F.; Henderson, B. L.; Lindgren, C. M.; Boekweg, H.; Lu, S.; Baral, J.; Yao, L.; Stratton, K. G.; Bramer, L. M.; Zink, E.; Couvillion, S. P.; Bloodsworth, K. J.; Satpathy, S.; Sieh, W.; Boca, S. M.; Schürer, S.; Chen, F.; Wiznerowicz, M.; Ketchum, K. A.; Boja, E. S.; Kinsinger, C. R.; Robles, A. I.; Hiltke, T.; Thiagarajan, M.; Nesvizhskii, A. I.; Zhang, B.; Mani, D. R.; Ceccarelli, M.; Chen, X. S.; Cottingham, S. L.; Li, Q. K.; Kim, A. H.; Fenyö, D.; Ruggles, K. V.; Rodriguez, H.; Mesri, M.; Payne, S. H.; Resnick, A. C.; Wang, P.; Smith, R. D.; Iavarone, A.; Chheda, M. G.; Barnholtz-Sloan, J. S.; Rodland, K. D.; Liu, T.; Ding, L.; Clinical Proteomic Tumor Analysis Consortium. Proteogenomic and Metabolomic Characterization of Human Glioblastoma. *Cancer Cell* **2021**, *39* (4), 509-528.e20. https://doi.org/10.1016/j.ccell.2021.01.006.

(23) Satpathy, S.; Krug, K.; Jean Beltran, P. M.; Savage, S. R.; Petralia, F.; Kumar-Sinha, C.; Dou, Y.; Reva, B.; Kane, M. H.; Avanessian, S. C.; Vasaikar, S. V.; Krek, A.; Lei, J. T.; Jaehnig, E. J.; Omelchenko, T.; Geffen, Y.; Bergstrom, E. J.; Stathias, V.; Christianson, K. E.; Heiman, D. I.; Cieslik, M. P.; Cao, S.; Song, X.; Ji, J.; Liu, W.; Li, K.; Wen, B.; Li, Y.; Gümüş, Z. H.; Selvan, M. E.; Soundararajan, R.; Visal, T. H.; Raso, M. G.; Parra, E. R.; Babur, Ö.; Vats, P.; Anand, S.; Schraink, T.; Cornwell, M.; Rodrigues, F. M.; Zhu, H.; Mo, C.-K.; Zhang, Y.; da Veiga Leprevost, F.; Huang, C.; Chinnaiyan, A. M.; Wyczalkowski, M. A.; Omenn, G. S.; Newton, C. J.; Schurer, S.; Ruggles, K. V.; Fenyö, D.; Jewell, S. D.; Thiagarajan, M.; Mesri, M.; Rodriguez, H.; Mani, S. A.; Udeshi, N. D.; Getz, G.; Suh, J.; Li, Q. K.; Hostetter, G.; Paik, P. K.; Dhanasekaran, S. M.; Govindan, R.; Ding, L.; Robles, A. I.; Clauser, K. R.; Nesvizhskii, A. I.; Wang, P.; Carr, S. A.; Zhang, B.; Mani, D. R.; Gillette, M. A.; Clinical Proteomic Tumor Analysis Consortium. A Proteogenomic Portrait of Lung Squamous Cell Carcinoma. *Cell* **2021**, *184* (16), 4348-4371.e40. https://doi.org/10.1016/j.cell.2021.07.016.

(24) Lindgren, C. M.; Adams, D. W.; Kimball, B.; Boekweg, H.; Tayler, S.; Pugh, S. L.; Payne, S. H. Simplified and Unified Access to Cancer Proteogenomic Data. *J. Proteome Res.* **2021**, *20* (4), 1902–1910. https://doi.org/10.1021/acs.jproteome.0c00919.

(25) Giurgiu, M.; Reinhard, J.; Brauner, B.; Dunger-Kaltenbach, I.; Fobo, G.; Frishman, G.; Montrone, C.; Ruepp, A. CORUM: The Comprehensive Resource of Mammalian Protein Complexes-2019. *Nucleic Acids Res.* **2019**, *47* (D1), D559–D563. https://doi.org/10.1093/nar/gky973.

(26) Szklarczyk, D.; Gable, A. L.; Nastou, K. C.; Lyon, D.; Kirsch, R.; Pyysalo, S.; Doncheva, N. T.; Legeay, M.; Fang, T.; Bork, P.; Jensen, L. J.; von Mering, C. The STRING Database in 2021: Customizable Protein-Protein Networks, and Functional Characterization of User-Uploaded Gene/Measurement Sets. *Nucleic Acids Res.* **2021**, *49* (D1), D605–D612. https://doi.org/10.1093/nar/gkaa1074.

(27) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, É. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12* (85), 2825–2830.

(28) Hagberg, A.; Swart, P.; S Chult, D. *Exploring Network Structure, Dynamics, and Function Using Networkx*; LA-UR-08-05495; LA-UR-08-5495; Los Alamos National Lab. (LANL), Los Alamos, NM (United States), 2008.

(29) Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N. S.; Wang, J. T.; Ramage, D.; Amin, N.; Schwikowski, B.; Ideker, T. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* **2003**, *13* (11), 2498–2504. https://doi.org/10.1101/gr.1239303.

(30) Zheng, F.; Zhang, S.; Churas, C.; Pratt, D.; Bahar, I.; Ideker, T. HiDeF: Identifying Persistent Structures in Multiscale ‘omics Data. *Genome Biol.* **2021**, *22* (1), 21. https://doi.org/10.1186/s13059-020-02228-4.

(31) Singhal, A.; Cao, S.; Churas, C.; Pratt, D.; Fortunato, S.; Zheng, F.; Ideker, T. Multiscale Community Detection in Cytoscape. *PLOS Comput. Biol.* **2020**, *16* (10), e1008239. https://doi.org/10.1371/journal.pcbi.1008239.

(32) Traag, V. A.; Waltman, L.; van Eck, N. J. From Louvain to Leiden: Guaranteeing Well-Connected Communities. *Sci. Rep.* **2019**, *9* (1), 5233. https://doi.org/10.1038/s41598-019-41695-z.

(33) Reimand, J.; Arak, T.; Adler, P.; Kolberg, L.; Reisberg, S.; Peterson, H.; Vilo, J. G:Profiler—a Web Server for Functional Interpretation of Gene Lists (2016 Update). *Nucleic Acids Res.* **2016**, *44* (W1), W83–W89. https://doi.org/10.1093/nar/gkw199.

(34) Tang, Y.; Li, M.; Wang, J.; Pan, Y.; Wu, F.-X. CytoNCA: A Cytoscape Plugin for Centrality Analysis and Evaluation of Protein Interaction Networks. *Biosystems* **2015**, *127*, 67–72. https://doi.org/10.1016/j.biosystems.2014.11.005.

(35) Doncheva, N. T.; Morris, J. H.; Gorodkin, J.; Jensen, L. J. Cytoscape StringApp: Network Analysis and Visualization of Proteomics Data. *J. Proteome Res.* **2019**, *18* (2), 623–632. https://doi.org/10.1021/acs.jproteome.8b00702.

(36) Yu, G.; He, Q.-Y. ReactomePA: An R/Bioconductor Package for Reactome Pathway Analysis and Visualization. *Mol. Biosyst.* **2016**, *12* (2), 477–479. https://doi.org/10.1039/C5MB00663E.

(37) Wu, T.; Hu, E.; Xu, S.; Chen, M.; Guo, P.; Dai, Z.; Feng, T.; Zhou, L.; Tang, W.; Zhan, L.; Fu, X.; Liu, S.; Bo, X.; Yu, G. ClusterProfiler 4.0: A Universal Enrichment Tool for Interpreting Omics Data. *The Innovation* **2021**, *2* (3), 100141. https://doi.org/10.1016/j.xinn.2021.100141.

(38) Gu, Z.; Gu, L.; Eils, R.; Schlesner, M.; Brors, B. Circlize Implements and Enhances Circular Visualization in R. *Bioinformatics* **2014**, *30* (19), 2811–2812. https://doi.org/10.1093/bioinformatics/btu393.

(39) Waskom, M. Seaborn: Statistical Data Visualization. *J. Open Source Softw.* **2021**, *6* (60), 3021. https://doi.org/10.21105/joss.03021.

(40) Mick, D. U.; Dennerlein, S.; Wiese, H.; Reinhold, R.; Pacheu-Grau, D.; Lorenzi, I.; Sasarman, F.; Weraarpachai, W.; Shoubridge, E. A.; Warscheid, B.; Rehling, P. MITRAC Links Mitochondrial Protein Translocation to Respiratory-Chain Assembly and Translational Regulation. *Cell* **2012**, *151* (7), 1528–1541. https://doi.org/10.1016/j.cell.2012.11.053.

(41) Signes, A.; Fernandez-Vizarra, E. Assembly of Mammalian Oxidative Phosphorylation Complexes I–V and Supercomplexes. *Essays Biochem.* **2018**, *62* (3), 255–270. https://doi.org/10.1042/EBC20170098.

(42) Hirano, Y.; Hendil, K. B.; Yashiroda, H.; Iemura, S.; Nagane, R.; Hioki, Y.; Natsume, T.; Tanaka, K.; Murata, S. A Heterodimeric Complex That Promotes the Assembly of Mammalian 20S Proteasomes. *Nature* **2005**, *437* (7063), 1381–1385. https://doi.org/10.1038/nature04106.

(43) The GTEx Consortium. The GTEx Consortium Atlas of Genetic Regulatory Effects across Human Tissues. *Science* **2020**, *369* (6509), 1318–1330. https://doi.org/10.1126/science.aaz1776.

(44) Tang, D.; Sheng, J.; Xu, L.; Zhan, X.; Liu, J.; Jiang, H.; Shu, X.; Liu, X.; Zhang, T.; Jiang, L.; Zhou, C.; Li, W.; Cheng, W.; Li, Z.; Wang, K.; Lu, K.; Yan, C.; Qi, S. Cryo-EM Structure of C9ORF72–SMCR8–WDR41 Reveals the Role as a GAP for Rab8a and Rab11a. *Proc. Natl. Acad. Sci.* **2020**, *117* (18), 9876–9883. https://doi.org/10.1073/pnas.2002110117.

(45) Sellier, C.; Campanari, M.-L.; Julie Corbier, C.; Gaucherot, A.; Kolb-Cheynel, I.; Oulad-Abdelghani, M.; Ruffenach, F.; Page, A.; Ciura, S.; Kabashi, E.; Charlet-Berguerand, N. Loss of C9ORF72 Impairs Autophagy and Synergizes with PolyQ Ataxin-2 to Induce Motor Neuron Dysfunction and Cell Death. *EMBO J.* **2016**, *35* (12), 1276–1297. https://doi.org/10.15252/embj.201593350.

(46) Koc, E. C.; Burkhart, W.; Blackburn, K.; Moyer, M. B.; Schlatzer, D. M.; Moseley, A.; Spremulli, L. L. The Large Subunit of the Mammalian Mitochondrial Ribosome: ANALYSIS OF THE

COMPLEMENT OF RIBOSOMAL PROTEINS PRESENT*. *J. Biol. Chem.* **2001**, *276* (47), 43958–43969. https://doi.org/10.1074/jbc.M106510200.

(47) Polianskyte, Z.; Peitsaro, N.; Dapkunas, A.; Liobikas, J.; Soliymani, R.; Lalowski, M.; Speer, O.; Seitsonen, J.; Butcher, S.; Cereghetti, G. M.; Linder, M. D.; Merckel, M.; Thompson, J.; Eriksson, O. LACTB Is a Filament-Forming Protein Localized in Mitochondria. *Proc. Natl. Acad. Sci.* **2009**, *106* (45), 18960–18965. https://doi.org/10.1073/pnas.0906734106.

(48) Keckesova, Z.; Donaher, J. L.; De Cock, J.; Freinkman, E.; Lingrell, S.; Bachovchin, D. A.; Bierie, B.; Tischler, V.; Noske, A.; Okondo, M. C.; Reinhardt, F.; Thiru, P.; Golub, T. R.; Vance, J. E.; Weinberg, R. A. LACTB Is a Tumour Suppressor That Modulates Lipid Metabolism and Cell State. *Nature* **2017**, *543* (7647), 681–686. https://doi.org/10.1038/nature21408.

(49) Payne, S. H. The Utility of Protein and MRNA Correlation. *Trends Biochem. Sci.* **2015**, *40* (1), 1–3. https://doi.org/10.1016/j.tibs.2014.10.010.

(50) Sanchez-Vega, F.; Mina, M.; Armenia, J.; Chatila, W. K.; Luna, A.; La, K. C.; Dimitriadoy, S.; Liu, D. L.; Kantheti, H. S.; Saghafinia, S.; Chakravarty, D.; Daian, F.; Gao, Q.; Bailey, M. H.; Liang, W.-W.; Foltz, S. M.; Shmulevich, I.; Ding, L.; Heins, Z.; Ochoa, A.; Gross, B.; Gao, J.; Zhang, H.; Kundra, R.; Kandoth, C.; Bahceci, I.; Dervishi, L.; Dogrusoz, U.; Zhou, W.; Shen, H.; Laird, P. W.; Way, G. P.; Greene, C. S.; Liang, H.; Xiao, Y.; Wang, C.; Iavarone, A.; Berger, A. H.; Bivona, T. G.; Lazar, A. J.; Hammer, G. D.; Giordano, T.; Kwong, L. N.; McArthur, G.; Huang, C.; Tward, A. D.; Frederick, M. J.; McCormick, F.; Meyerson, M.; Cancer Genome Atlas Research Network; Van Allen, E. M.; Cherniack, A. D.; Ciriello, G.; Sander, C.; Schultz, N. Oncogenic Signaling Pathways in The Cancer Genome Atlas. *Cell* **2018**, *173* (2), 321-337.e10. https://doi.org/10.1016/j.cell.2018.03.035.