Translation efficiency covariation across cell types is a conserved organizing principle of mammalian transcriptomes

Yue Liu[a], Ian Hoskins[a], Michael Geng[a], Qiuxia Zhao[a], Jonathan Chacko[a], Kangsheng Qi[a], Logan Persyn[a], Jun Wang[b], Dinghai Zheng[b], Yochen Zhong[a], Shilpa Rao[a], Dayea Park[a], Elif Sarinay Cenik[a], Vikram Agarwal[b] , Hakan Ozadam[a,c], Can Cenik*

Affiliations:

[a] Department of Molecular Biosciences, University of Texas at Austin, Austin, TX 78712, USA

[b] mRNA Center of Excellence, Sanofi, Waltham, MA 02451, USA

[c] Present address: Sail Biomedicines, Cambridge, MA, 02141, USA

* Corresponding author**:** ccenik@austin.utexas.edu

ABSTRACT

Characterization of shared patterns of RNA expression between genes across conditions has led to the discovery of regulatory networks and novel biological functions. However, it is unclear if such coordination extends to translation, a critical step in gene expression. Here, we uniformly analyzed 3,819 ribosome profiling datasets from 117 human and 94 mouse tissues and cell lines. We introduce the concept of *Translation Efficiency Covariation* (TEC), identifying coordinated translation patterns across cell types. We nominate potential mechanisms driving shared patterns of translation regulation. TEC is conserved across human and mouse cells and helps uncover gene functions. Moreover, our observations indicate that proteins that physically interact are highly enriched for positive covariation at both translational and transcriptional levels. Our findings establish translational covariation as a conserved organizing principle of mammalian transcriptomes.

Keywords: Translation efficiency, Translational regulation, Ribosome profiling, Translation efficiency covariation, Gene functions, Regulatory networks

INTRODUCTION

In the last three decades, technological advances have progressively revealed the expression of RNAs with increasing spatial and cellular resolution[1–7]. These measurements have spurred conceptual advances, driven by computational approaches. Foremost among these is the concept of RNA co-expression, which quantifies the similarities in RNA expression changes among groups of genes across conditions[8–11].

RNA co-expression analysis across biological contexts reveals shared biological functions, informing us about underlying mechanisms and interactions[12–15]. By applying the principle of guilt-by-association, new functions for genes with previously unknown roles can be inferred by the similarity of RNA expression patterns with genes of known function[9,12,16]. Furthermore, RNA co-expression between transcripts is predictive of protein-protein interactions[17,18], and can indicate genes that are likely regulated by the same transcription factors, suggesting common regulatory mechanisms[19,20].

These findings suggest that RNA co-expression may serve as a proxy for the proteomic organization of cells. However, it is only recently that quantification of protein abundance across numerous cell types and conditions has become possible, allowing this assumption to be explicitly tested. Mass spectrometry-based measurements across hundreds of cell types have revealed that proteins similarly exhibit shared patterns of abundance, organized according to their functions and physical interactions[21–23]. Surprisingly, much of the proteome-level organization of co-abundance patterns are not detected at the RNA level[21,22]. Furthermore, physically interacting proteins are much more likely to have coordinated protein abundance than RNA co-expression[21,22,24]. RNA co-expression in both mouse and human cells often arises from the chromosomal proximity of genes even when they are functionally unrelated[25,26]. This likely unproductive co-expression pattern is absent at the protein level[27,28], suggesting that post-transcriptional regulation plays a significant role in proteome organization.

Translation regulation, a crucial post-transcriptional process, may bridge this gap, given its vital roles in development, maintaining cellular homeostasis, and responding to environmental changes[29–35]. There are three lines of evidence that suggest the possibility of coordinated translation of functionally and physically associated proteins across different biological contexts.

First, mammalian mRNAs bind various proteins to form ribonucleoproteins that influence their lifecycle from export to translation[36]. The set of proteins interacting with an mRNA varies with time and context, significantly altering the duration, efficiency, and localization of protein production. These observations led to the proposal of the post-transcriptional RNA regulon model over two decades ago, positing that functionally related mRNAs are regulated together post-transcriptionally[37,38]. Supporting this model, Puf3, a Pumilio family member in yeast, represses the translation of sequence-specific mRNAs encoding mitochondrial proteins[38]. Similarly, in human cells, CSDE1/UNR regulates the translation of mRNAs involved in epithelial-to-mesenchymal transition[39]. However, it remains to be determined if translation of functionally related mRNAs is coordinately regulated across different conditions.

Second, in both *E. coli* and yeast, proteins within multiprotein complexes are synthesized in

stoichiometric proportions needed for assembly[40,41]. This translational regulation likely tunes protein production to minimize the synthesis of excess protein components that would otherwise need to be degraded[42]. However, in human cells, evidence of such proportional synthesis is reported for only two complexes: ribosomes[41,43] and the oxidative phosphorylation machinery[44]. Furthermore, these observations have been made in a very limited number of cell lines, which limits the generalizability of this concept across diverse cell types and other functionally related protein groups.

Third, the formation of many protein complexes is facilitated by the co-translational folding of nascent peptides[40,41,45,46]. For instance, in bacteria, the anti-Shine-Dalgarno sequence induces translational pausing to modulate the co-translational folding of nascent peptides[46]. Co-translational assembly ensures that protein subunits are synthesized near each other, enabling near concurrent interactions, which are crucial for the biogenesis of some complex protein structures[47]. Recent evidence indicates that co-translational assembly may also be relatively common in human cells[48].

Co-translational assembly and stoichiometric synthesis rates of protein complexes suggest coordinated translation of several mRNAs within a given cell type.. However, due to the lack of robust, transcriptome-wide translational efficiency (TE) measurements across diverse biological conditions, it remains to be seen whether such coordination extends across different cell types or conditions. To address this, we analyzed thousands of matched ribosome profiling and RNA-seq datasets from >140 human and mouse cell lines and tissues. To quantify the similarity of translation efficiency patterns of transcripts across cell types and tissues, analogously to RNA co-expression, we introduce the concept of Translation Efficiency Covariation (TEC). based on a compositional data analysis approach[49,50] Our findings demonstrate that TEC can reveal gene functions not identified through RNA co-expression analysis alone and uncovered shared motifs for RNA binding proteins (RBPs) among genes exhibiting TEC. Physically interacting proteins are highly enriched for both TEC and RNA co-expression. Further supporting the functional significance of this concept, TEC among genes is highly conserved between humans and mice.

RESULTS

Integrated analysis of thousands of ribosome profiling and RNA-seq measurements enable quantitative assessment of data quality

We undertook a comprehensive, large-scale meta-analysis of ribosome profiling data to quantify TE across different cell lines and tissues. We collected 2,195 ribosome profiling datasets for humans and 1,624 experiments for mice, along with their metadata (Fig. 1a; Methods). Given that metadata is frequently reported in an unstructured manner and lacks a formal verification step, we conducted a manual curation process to rectify inaccuracies and collect missing information, such as experimental conditions and cell types used in experiments. One crucial aspect of our manual curation was pairing between ribosome profiling and corresponding RNA-seq when possible. Overall, 1,282 (58.4%) human and 995 (61.3%) mouse ribosome profiling samples were matched with corresponding RNA-seq data (table S1). The resulting curated metadata facilitated the uniform processing of ribosome profiling and corresponding RNA-seq data using an open-source

pipeline[51]. We call the resulting repository harboring these processed files RiboBase (table S1).

In RiboBase, the top cell types with the most experiments were HEK293T (13.1%) and HeLa (8.1%) for human; in mouse, the leading tissues were brain (9.6%), embryonic fibroblasts (8.3%), and liver (7.7%) (Fig. 1b; table S1). The median number of sequencing reads for ribosome profiling samples was ~43.2 million for humans and ~37.5 million for mice, respectively (ExtendedDataFig. 1a-b; table S2-3; supplementary text). A majority of reads contained adapter sequences included during library preparation (with medians of 82.2% and 79.2% of total reads having adapters for human and mouse, respectively). Due to the substantial presence of ribosomal RNA in ribosome profiling datasets, only around 15% of total reads aligned to the transcript reference (ExtendedDataFig. 1c-d; table S4-5; supplementary text).

The length of ribosome-protected mRNA footprints (RPFs) provides valuable information about data quality, the experimental protocol used, and translational activity[52]. The choice of nuclease impacts the resulting read length distribution of RPFs[53] (ExtendedDataFig. 2a-b). In agreement, we found that the peak position and range of RPF lengths were closely associated with the type of digestion enzymes used in human cancer samples (Fig. 1c). To account for the variability of RPF length distributions across the compendium of experiments, we developed a module that allowed for setting sample-specific RPF read length cutoffs (ExtendedDataFig. 3a; Methods). This dynamic approach proved more effective than using fixed minimum and maximum values for RPF lengths, resulting in a higher retrieval of usable reads (median increase of 10.8% for human and 17.1% for mouse) and an increased proportion of reads within the coding sequence (CDS) region (ExtendedDataFig. 3b).

After selecting a set of RPFs, we assessed the quality of ribosome profiling data within RiboBase using two additional criteria. Given that translating ribosomes should be highly enriched in annotated coding regions, we require that at least 70% of RPFs should be mapped to the CDS. We found that 160 human and 115 mouse samples failed to meet this criterion (Fig. 1d; table S6-7). Subsequently, we required a minimum number of RPFs that map to CDS to ensure sufficient coverage of translated genes (Methods). There were 318 human and 431 mouse samples with less than 0.1X transcript coverage (Fig. 1e; table S6-7). Altogether, 1,794 human samples and 1,134 mouse samples were retained for in-depth analysis. Of these, 1,076 human and 845 mouse samples were paired with matching RNA-seq data. Our results indicate a considerable fraction of publicly available ribosome profiling experiments had suboptimal quality (18.3% of the human and 30.1% of the mouse samples) (Fig. 1f). Interestingly, the data quality appeared to be independent of time (ExtendedDataFig. 4). Additionally, we found that samples that passed our quality thresholds were more likely to exhibit three-nucleotide periodicity compared to those that failed quality control (92.59% vs 78.30% for humans and 91.36% vs 86.73% for mice; ExtendedDataFig. 5; Methods). These findings underscore the necessity of meticulous quality control for the selection of experiments to enable large-scale data analyses.

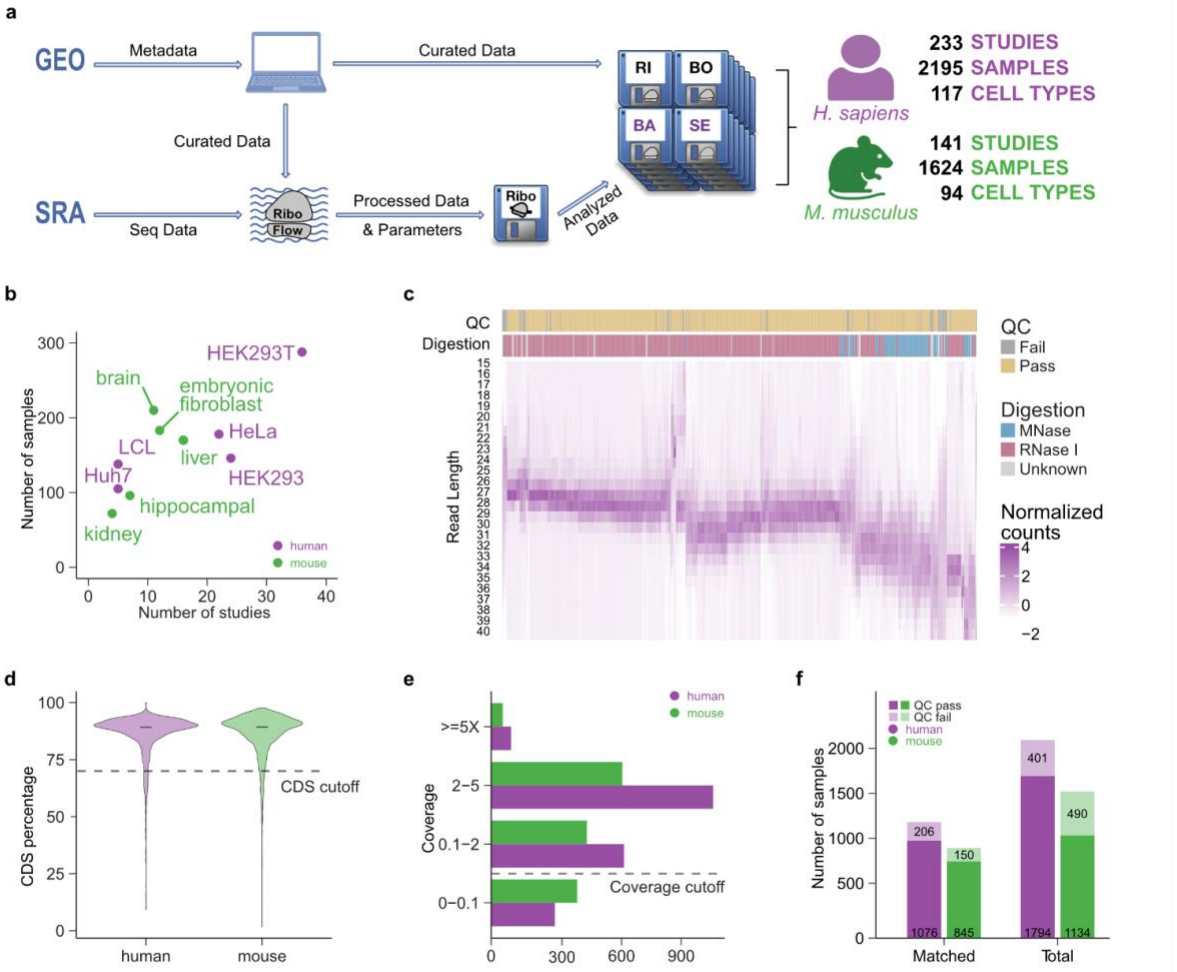


Fig. 1 | RiboBase: a comprehensive ribosome profiling database with thousands of experiments. a, Schematic of RiboBase. We manually curated metadata and processed the sequencing reads using a uniform pipeline (RiboFlow[51]). **b,** Top five most highly represented cell lines or tissues with respect to the number of experiments were plotted. **c,** We determined the ribonuclease used to generate ribosome profiling data for 680 experiments using human cancer cell lines. For each experiment, the read length distribution of RPFs mapping to coding regions was visualized as a heatmap. The color represents the z-score adjusted RPF counts (Methods). Each experiment where the percentage of RPFs mapping to CDS was greater than 70% and achieving sufficient coverage of the transcript (>= 0.1X) was annotated as QC-pass (Methods). **d,** For the 3,819 ribosome profiling experiments in RiboBase, we applied a function to select the range of RPFs for further analysis (Methods). We calculated the proportion of the selected RPFs that map to the coding regions (y-axis). The horizontal line represents the median of the distribution. **e,** Experiments (x-axis) were grouped by the transcript coverage (y-axis). **f,** Among

the ribosome profiling experiments in RiboBase, 2,277 of them had corresponding RNA-seq data (matched). The number of samples that pass quality controls were plotted.

Translation efficiency is conserved across species and is cell-type specific

Ribosome profiling measures ribosome occupancy, a variable influenced by both RNA expression and translation dynamics. Thus, estimating translation efficiency necessitates analysis of paired RNA-seq and ribosome profiling data. To assess accurate matching in RiboBase, we first compared the coefficient of determination (R^2) between matched ribosome profiling and RNA-seq data to that from other pairings within the same study. As would be expected from correct matching, we found that matched samples had significantly higher similarity on average (Fig. 2a; Welch two-sided t-test p-value = 2.2×10^{-16} for human and p-value = 2.1×10^{-5} for mouse). We then implemented a scoring system to quantitatively evaluate the correctness of our manual matching information (Methods). 99.2% of human samples and 98.5% of mouse samples had a sufficiently high matching score, demonstrating the effectiveness of our manual curation strategy (ExtendedDataFig. 6a; Methods).

Using the set of matched ribosome profiling and RNA-seq experiments, we next quantified TE, which is typically defined as the log ratio of ribosome footprints to RNA-seq reads, normalized as counts per million[54]. However, this approach leads to biased estimates with significant drawbacks[55]. To address this limitation, we calculate TE based on a regression model using a compositional data analysis method[49,50,56], avoiding the mathematical shortcomings of using a log-ratio (Fig. 2b; ExtendedDataFig. 6a-c, 7; table S8-11; Methods).

We next assessed whether measurement errors due to differences in experimental procedures dominate variability that would otherwise be attributed to biological variables of interest. Specifically, we compared similarities between experiments that used the same cell type or tissue in different studies (ExtendedDataFig. 8a). We found that ribosome profiling or RNA experiments from the same cell type or tissue exhibited higher similarity compared to those from different cell lines or tissues (Fig. 2c). Consistent with this observation, TE values displayed higher Spearman correlation coefficient within the same cell type or tissue (median correlation coefficient of 0.56 and 0.53 in human and mouse, respectively) compared to different cell lines and tissues (median correlation coefficient of 0.49 and 0.45 in human and mouse, respectively) (Fig. 2d).

We expected that a more accurate estimate of TE would show a stronger correlation with protein abundance. We calculated for each transcript the cell type-specific TE by taking the average of TE values across all experiments conducted with that particular cell line. Indeed, our results show that compared to the log-ratio definition, the TE derived using the regression approach with winsorized read counts (ExtendedDataFig. 8b; ExtendedDataFig. 9-11; supplementary text) is more strongly correlated with protein abundance in seven cancer cell lines (mean Spearman correlation coefficient of 0.465 vs 0.219; Fig. 2e).

Furthermore, TE measurements from cell lines and tissues with the same biological origin (e.g., blood) tended to cluster together, supporting the existence of cell-type-specific differences in TE (Fig. 2f). As expected, mean ribosome occupancy and RNA expression across cell types showed a strong correlation (Spearman correlation: ~0.8), yet mean TE was only weakly associated with

RNA expression (Spearman correlation: ~0.2) (Fig. 2g). Taken together, our analyses demonstrate that our compositional regression-based approach to calculating TE ensures more accurate and consistent measurements across different cell types and conditions.

Measurements of TE in two species across a large number of cell types enabled us to investigate the conservation of TE, ribosome occupancy, and RNA expression. Transcriptomes, ribosome occupancy, and proteomes exhibit a high degree of conservation across diverse organisms[57,58]. Consistently, we found average ribosome occupancy, RNA expression, and TE across different cell lines and tissues were highly similar between orthologous genes in human and mouse (Fig. 2g; table S12). Specifically, the Spearman correlation coefficient of mean TE across cell types and tissues between human and mouse was 0.9 (Fig. 2h), which is comparable to the mean RNA expression correlation between human and mouse (~0.86, ExtendedDataFig. 12a). Using a 95% prediction interval to identify outlier genes, we found that outlier genes with higher mean TE in humans compared to mice were enriched in the gene ontology term 'RNA binding function' (Fig. 2i). In contrast, genes with elevated mean TE in mice were enriched for having functions related to extracellular matrix and collagen-containing components (Fig. 2i). The enrichment of genes with higher TE in mice, particularly those from the extracellular matrix and collagen-containing components, may be due to the fact that many samples in mouse studies are derived from the early developmental stage[59].

Despite the high correlation of mean TE across various cell lines and tissues between human and mouse, TE distinctly exhibits cell-type specificity. While several studies compared the conservation of TE between the same tissues of mammalians or model organisms[58,60,61], our dataset uniquely enabled us to determine the conservation of variability of TE for transcripts across different cell types. Intriguingly, we observed a moderately high similarity between the variability of TE of orthologous genes in human and mouse (Spearman partial correlation coefficient = 0.63; Fig. 2j; ExtendedDataFig. 12b-d; Methods). Our results reveal that certain genes exhibit higher variability of TE across cell types and this is a conserved property between human and mouse.

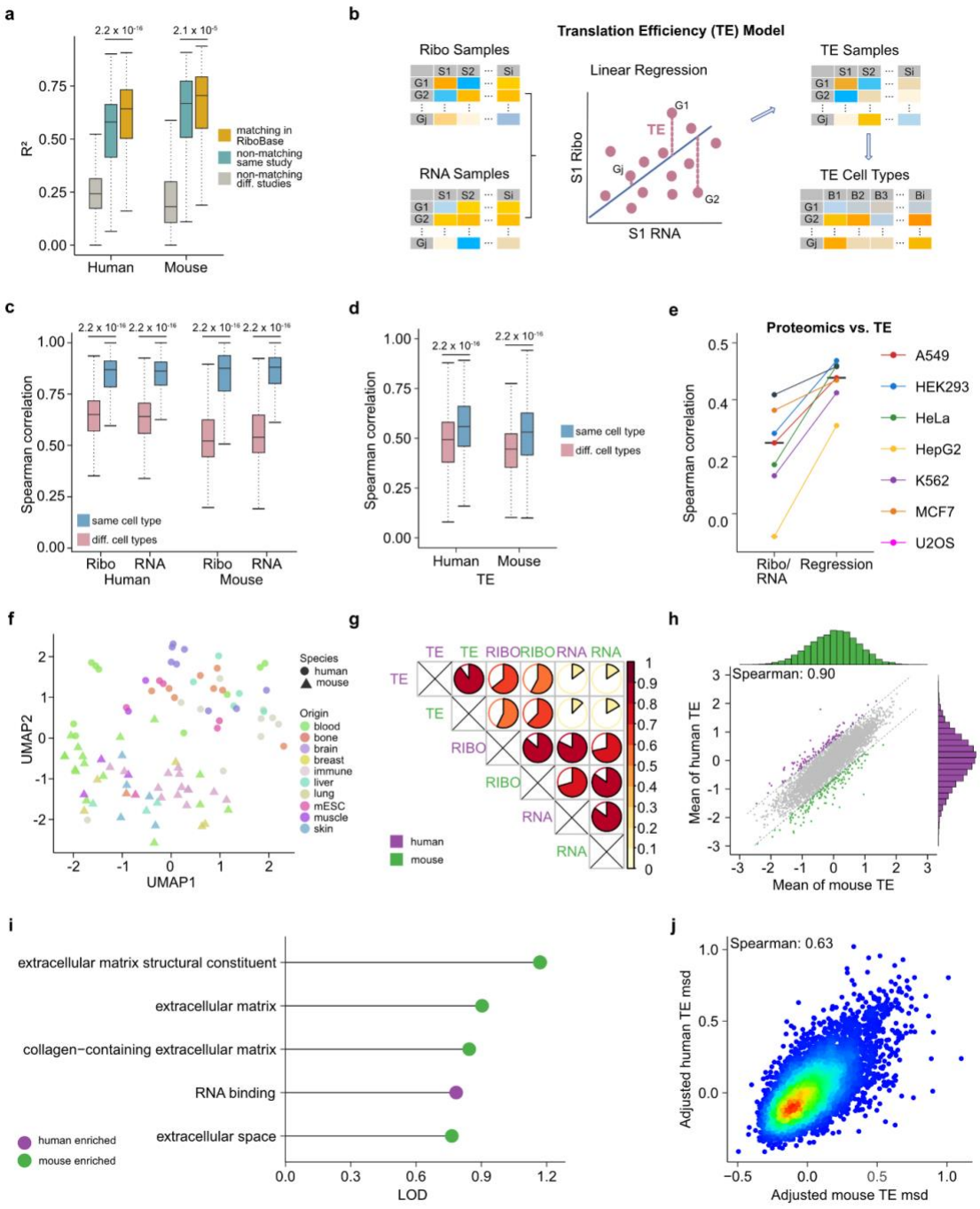


Fig. 2 | TE defined using a compositional linear regression model is conserved across cell types and species. a, The distribution of coefficient of determination (R^2, y-axis) between ribosome profiling data and RNA-seq in RiboBase was compared to random matching within the same study and across different studies. In each figure panel containing boxplots, the horizontal

line corresponds to the median. The box represents the interquartile range (IQR) and the whiskers extend to the largest value within 1.5 times the IQR. The significant p-value shown in this figure was calculated using the two-sided Wilcoxon test. **b,** Schematic of TE calculation using the linear regression model with compositional data (CLR transformed; Methods; ExtendedDataFig. 7). **c,** Distribution of correlations of TE (linear regression model) across experiments. **d,** Correlation between TE and protein abundance from seven human cell lines[100] was calculated using log-ratio of ribosome profiling and RNA expression or compositional regression method. The horizontal line corresponds to the median. **e,** The distribution of Spearman correlations between experiments (y-axis) was calculated based on whether they originated from identical or different cell lines or tissues. **f,** We used UMAP to cluster the TE values of all genes across different cell types, considering only those origins with at least five distinct cell types. **g,** The Spearman correlation of 9,194 orthologous genes between human and mouse across TE, ribosome profiling, and RNA-seq levels. The circles represent the value of the Spearman correlation between groups. **h,** TE values were averaged across cell types and tissues for either human and mouse. Each dot represents a gene, and a 95% prediction interval was plotted to identify outlier genes (highlighted in purple and green). **i,** We conducted GO term enrichment analysis for outlier genes from panel H. We ranked the GO terms (y-axis) by the logarithm of the odds (LOD; x-axis). **j,** The correlation of the standard deviation of TE (quantified with adjusted metric standard deviation (msd); Methods; ExtendedDataFig. 12c-d) for orthologous genes across different cell types between human and mouse.

Translation efficiency covariation (TEC) is conserved between human and mouse

Uniform quantification of TE enabled us to investigate the similarities in TE patterns across cell types. Given the usefulness of RNA co-expression in identifying shared regulation and biological functions, we aimed to establish an analogous method to detect patterns of translation efficiency similarity among genes. To achieve this, we employed the proportionality score (rho)[50,56], a statistical method that quantifies the consistency of how relative TE changes across different contexts (Methods). Recent work suggested that the proportionality score enhances cluster identification in high-dimensional single-cell RNA co-expression data[10]. Consistent with these findings, our analysis revealed its particular effectiveness in quantifying ribosome occupancy covariation (ExtendedDataFig. 13; Methods). We calculated rho scores for all pairs of human or mouse genes where a high absolute rho score indicates significant translation efficiency covariation (TEC) between pairs (Fig. 3a).

Previous studies have indicated that RNA co-expression between genes is conserved in mammals[57,62,63]. To assess the potential evolutionary significance of the newly introduced TEC concept, we evaluated its conservation across human and mouse transcripts. Indeed, TEC was highly similar for orthologous gene pairs in humans and mice (Fig. 3b, Pearson correlation coefficient 0.41), compared to a negligible correlation in TEC derived from shuffled TE values (ExtendedDataFig. 14, Pearson correlation coefficient 0.00022). Our findings imply that translation efficiency patterns are evolutionarily preserved, paralleling the conservation of RNA co-expression.

RNA co-expression analyses led to the discovery of regulatory motifs and shared transcription

factor binding sites[64]. We hypothesized that TEC among genes may nominate RNA binding proteins (RBPs) as potential drivers of TEC[65]. We identified groups of transcripts whose TE is correlated with the RNA expression of experimentally determined RBPs[66] (1274 human and 1762 mouse RBPs; Methods). The number of transcripts whose TE significantly correlates with the expression of each RBP differed widely, with ranges of 28-3052 (human) and 14-2393 (mouse) (https://zenodo.org/uploads/11359114; Pearson correlation FDR < 0.05). We refer to transcripts whose TE is significantly correlated with an RBP's expression as the RBP's regulon.

Interestingly, some RBP regulons were dominated by positive or negative correlations, suggesting activating or repressing functions for RBPs (Fig. 3c-d). For example, ZC3H10 has largely positive correlations (71% of RBP regulon) (Fig. 3c). Conversely, the RNA expression of subunits of ubiquinone oxidoreductase (Ndufa7, Ndufv3) is negatively correlated with TE for many genes in mice (Fig. 3c). Unexpectedly, we found that the RNA expression of ribosomal protein genes is negatively correlated with TE of many other transcripts (https://zenodo.org/uploads/11359114). This may indicate that transcriptome-wide TE is tempered during ribosome biogenesis, perhaps as a result of competition for ribosomes and other biosynthetic resources (tRNAs, amino acids) devoted to synthesizing new ribosomal proteins.

To identify evolutionarily conserved RBP regulons, we examined the intersection of significant RBP-gene correlations between human and mouse. At least some activating RBP functions may be evolutionarily conserved, as there was a correspondence between human and mouse in the proportion of regulon genes with positive correlations (Pearson correlation 0.44; Fig. 3d). To nominate RBPs that may modulate TE, we calculated the proportionality score of genes in each regulon and selected RBP regulons that had high absolute scores, reasoning that directional impacts on TE might be more likely if the RBP engages these transcripts. We found 85 RBPs where genes in the RBP's regulon had high TEC (mean absolute pairwise rho >90th percentile; ExtendedDataFig. 15; supplementary text). Some of these RBPs were previously known to regulate TE, including PARK7 and VIM (ExtendedDataFig. 16; supplementary text). Taken together, our analyses nominate RBPs that may coordinate the TEC of evolutionarily conserved RNA regulons.

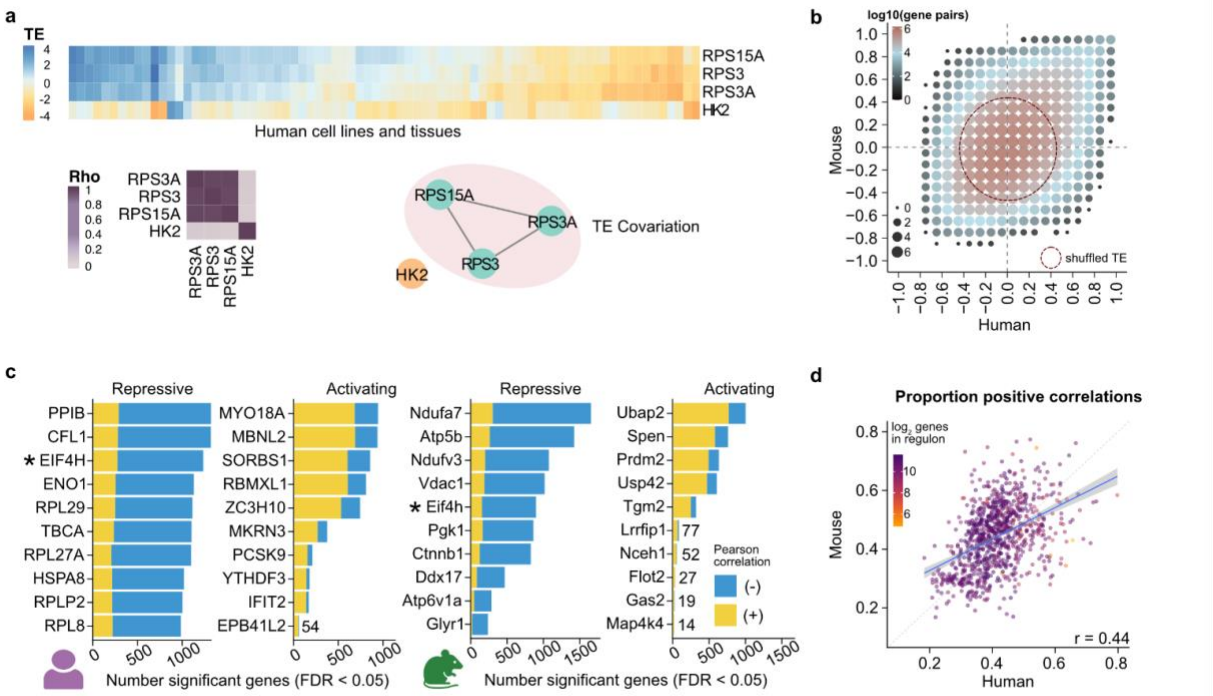


Fig. 3 | Translation efficiency covariation is conserved between human and mouse. a, Example illustrating translation efficiency covariation (TEC) between genes. The top section presents TE patterns across cell types in human. The bottom left part displays the similarity of the pattern between these genes quantified using proportionality scores. **b,** We calculated the TEC for gene pairs and compared their differences for the same orthologous gene pairs between human and mouse. In the figure panel, each dot represents the aggregated $\log_{10}$-transformed counts of gene pairs falling within specified ranges. We also calculated TEC using randomized TE for each gene (shuffled). The red dashed line in the figure captures the 95% gene pair TEC values obtained with shuffled TE (ExtendedDataFig. 14). **c,** Top ten candidates activating and repressive RBPs: human (left) and mouse (right). The number of genes with significant correlations between gene TE and RBP expression is shown. An asterisk marks genes in the top ten in both species. **d,** Each point is a RBP and plotted is the proportion of positive correlations between TE for genes in the regulon and the RNA expression of the RBP. Blue line is a linear fit with 95% confidence intervals in gray. Pearson correlation coefficient is shown.

Translation efficiency covariation (TEC) between transcripts across cell lines and tissues is associated with shared biological functions

Given that co-expression at the RNA level is predictive of shared biological functions[11,67,68], we next assessed whether TEC indicates common biological roles among genes. We calculated the area under the receiver operating characteristic curve (AUROC) to measure the ability of TEC in

distinguishing genes with the same biological functions (Methods). Genes that are annotated with a common GO term exhibited a similar degree of RNA co-expression and TEC, both of which were significantly higher than would be expected by chance (Median AUROC across GO terms calculated with TEC: 0.63 for human, 0.65 for mouse; RNA co-expression RNA: 0.66 for human, 0.69 for mouse; Fig. 4a; table S13-14; Methods). These findings demonstrate that TEC, similar to RNA co-expression, serves as an indicator of shared biological functions among genes.

Furthermore, we observed that biological functions whose members exhibit a high degree of RNA co-expression were also likely to have TEC. Specifically, the Spearman correlation between the AUROC scores calculated using TEC and RNA co-expression was ~0.64 for human GO terms in contrast to ~-0.02 when random genes were grouped (Fig. 4b). Despite the low correlation between average RNA expression and TE for human genes (Fig. 2g), our results highlight that members of specific biological functions whose RNA expression is coordinated across cell types tend to exhibit consistent translation efficiency patterns. This finding suggests coordinated regulation at both transcriptional and translational levels among functionally related genes.

While many gene functions were predicted accurately with both RNA co-expression and TEC, we noted specific exceptions. Notably, genes in 29 human GO terms demonstrated significantly stronger TEC than RNA co-expression (at least 0.1 higher AUROC; Fig. 4c; ExtendedDataFig. 17-18; supplementary text). An example of such a GO term is 'MAPKKK activity' (Fig. 4d-e). While there is limited evidence of direct translational regulation of the MAPKKK family, the RBP IMP3 may provide a potential mechanism for such regulation[69]. Additionally, there is post-translational regulation through the binding of activated RAS to genes from the MAPKKK family, leading to their activation[70]. These results indicate that some genes with specific biological functions exhibit greater similarity at the translational level.

We hypothesized that genes with shared functions and high TEC may be regulated through a common mechanism, analogous to shared transcription factor binding sites that mediate RNA co-expression[71,72]. Accordingly, we expected these genes to harbor sequence elements bound by RBPs. We identified enriched heptamers in the transcripts of five human and three mouse GO terms with significant TEC and at least 12 genes in the GO term (AUROC measured with TEC > 0.7, difference in AUROC between TEC and RNA co-expression > 0.2; Fig. 4f; ExtendedDataFig. 17b; ExtendedDataFig. 18e; Methods). For example, we found AG-rich motifs in coding regions of human genes with "molecular function inhibitor activity" (Fig. 4f). These motifs match the known binding sites of three RBPs (TRA2A, PABPN1, and SRSF1). In line with the enrichment of these motifs, analysis of eCLIP data revealed increased deposition of these RBPs in the coding sequences of genes in this GO term compared to matched control transcripts (Fig. 4f; Methods). Furthermore, we identified several additional enriched heptamers that currently have no RBP annotations, suggesting these motifs might be targets for RBPs that have not yet been characterized.

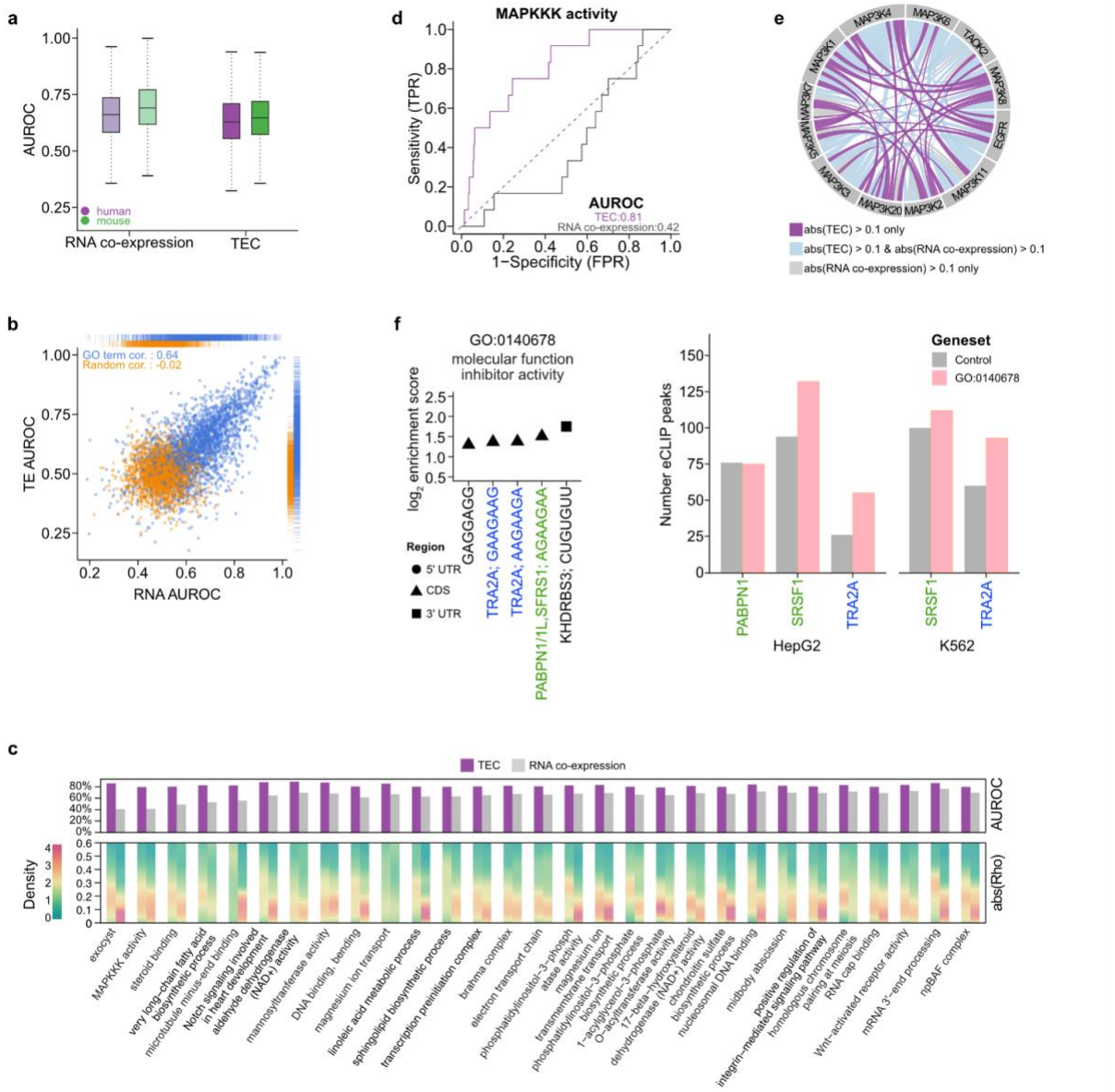


Fig. 4 | Genes associated with certain biological functions exhibit higher similarity patterns in TE than in RNA expression. a, We calculated the similarity of expression (quantified by AUROC; y-axis) among genes within 2,989 human and 3,340 mouse GO terms. In the box plot, the horizontal line corresponds to the median. The box represents the IQR and the whiskers extend to the largest value within 1.5 times the IQR. **b,** Each blue dot represents the AUROC calculated for a given GO term using TEC and RNA co-expression levels. Orange dots represent the same values for random grouping of genes (Methods). **c,** For GO terms where genes exhibit greater similarity at the TE level than at the RNA expression level (AUROC for TEC > 0.8, and difference of AUROC measured with TEC and RNA co-expression > 0.1), we visualized the distribution of

absolute rho scores for gene pairs (bottom; gene pairs with abs(rho) > 0.1). **d,** AUROC plot calculated with genes associated with MAPKKK activity. **e,** In the circle plot, the connections display absolute rho above 0.1 either at TE level alone (purple), at both RNA and TE levels (blue), or RNA level alone (gray) for gene pairs involved in MAPKKK activity. **f,** Motif enrichment (left) for the GO term 'molecular function inhibitor activity' (ExtendedDataFig. 17b). RNA binding proteins (RBPs) matching the motifs from oRNAment[134] or Transite[133] are indicated. Enhanced cross-linking immunoprecipitation (eCLIP) data[135] indicates increased binding of TRA2A and SRSF1 in the CDS of genes for this GO term compared to matched control genes with similar sequence properties (Methods).

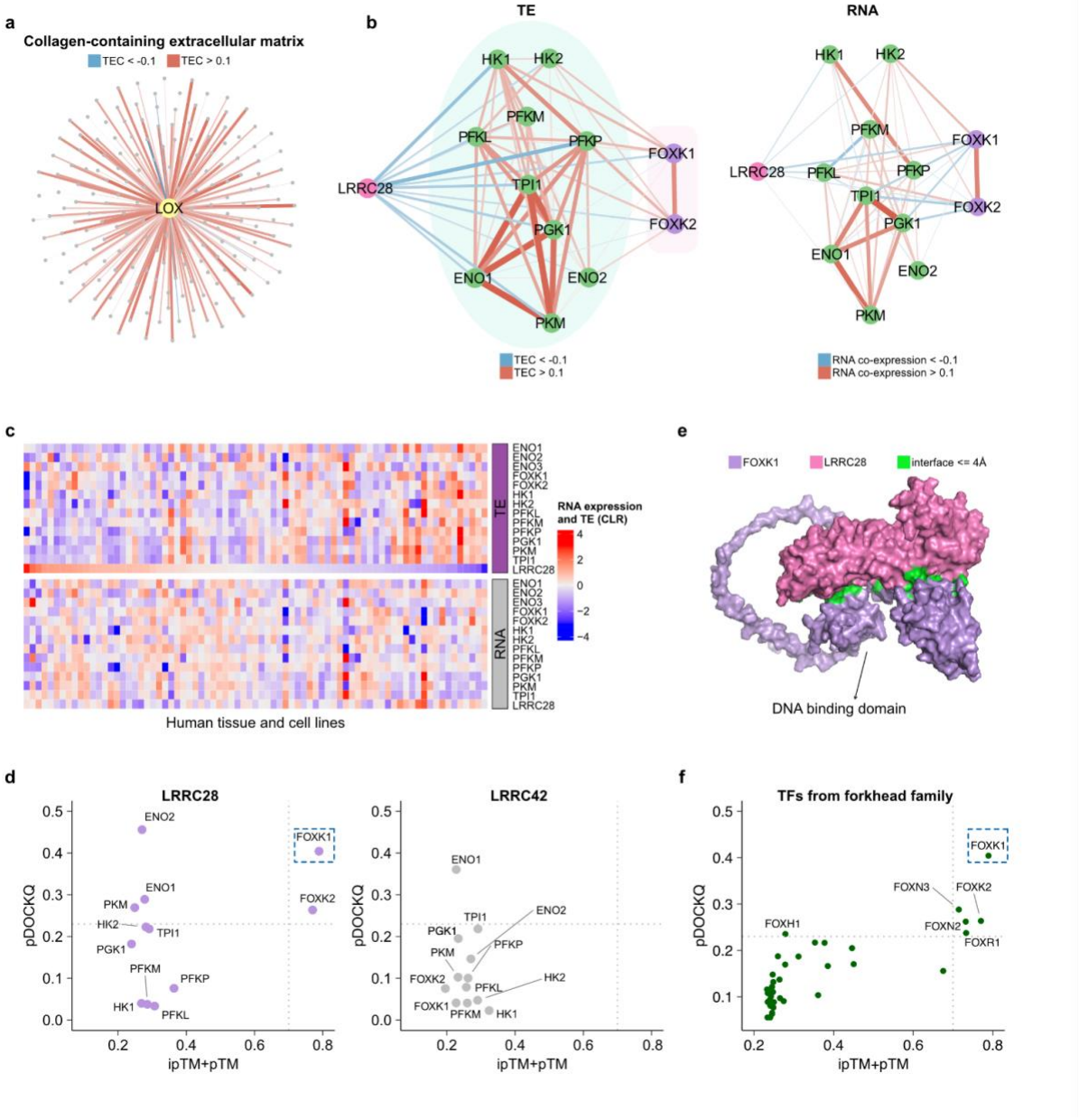
TEC reveals gene functions

We next investigated whether gene functions may be predicted by utilizing TEC, given the success of RNA co-expression for this task[67,68]. The functional annotations of human genes are continuously being updated, providing an opportunity to test this hypothesis using recently added information to the knowledge base. Specifically, we used functional annotations from the GO database from January 1, 2021, to determine functional groups that demonstrate strong TEC among its members (AUROC > 0.8) and developed a framework to predict new functional associations with these groups (Methods). By comparing our predictions to annotations from December 4, 2022, we confirmed the predicted association of the *LOX* gene with the GO term 'collagen-containing extracellular matrix'. LOX critically facilitates the formation, development, maturation, and remodeling of the extracellular matrix by catalyzing the cross-linking of collagen fibers, thereby enhancing the structural integrity and stability of tissues[73,74]. Our prediction successfully identified this new addition, as *LOX* exhibits positive similarity in TE with the vast majority of genes in this term (Fig. 5a).

Recognizing the capacity of TEC to elucidate biological functions, we utilized a recent version of GO annotations (December 4, 2022) to systematically predict new associations for genes. To underscore the unique insights gained from TEC, we focused on the 33 human and 31 mouse GO terms that either exhibited significantly higher TEC than RNA co-expression (Table 1) or provided new functional predictions that were only supported by TEC (the ranking of the newly predicted gene with RNA co-expression fell beyond the top 50%, table S15-16; Methods). By focusing on these GO terms, we aimed to identify similarity patterns based on TE, revealing functional associations that would not be detected by RNA co-expression. We conducted a literature search to determine if prior research supported these predictions, finding that 11 have already been corroborated by previous publications, although they have not yet been reflected in the relevant GO term annotations (Table 1; supplementary text). For example, cryo-electron microscopy experiments demonstrated that human DNMT1 binds to hemimethylated DNA in conjunction with ubiquitinated histone H3[75]. This binding facilitates the enzymatic activity of DNMT1 in maintaining genomic DNA methylation. Our analysis revealed that DNMT1 was the highest ranking prediction exhibiting strong TEC with genes associated with nucleosomal DNA binding function. In mouse, we predicted Plekha7 to be a member of the regulation of developmental processes. This prediction was recently validated by the observation of neural progenitor cell delamination upon the disruption of Plekha7[76–80].

The high rate of validation of our predictions in the literature suggested that other predictions based on TEC may reflect new and yet to be confirmed functions. In particular, we observed that the human leucine-rich repeat-containing 28 (LRRC28) gene displays strong TEC with glycolytic genes, but is not co-expressed at the RNA level (Fig. 5b-c, table S17). Specifically, *LRRC28* displayed negatively correlated TE with key glycolytic genes including *HK1*, *HK2*, *PFKL*, *PFKM*, *PFKP*, *TPI1*, *PGK1*, *ENO1*, *ENO2*, *PKM*, and two transcription factors *FOXK1* and *FOXK2* that regulate glycolytic genes[81]. Given that the leucine-rich repeat domains typically facilitate protein-protein interactions[82], LRRC28 may interact directly with one or more of the glycolytic proteins. Using AlphaFold2-Multimer[83], we calculated the binding confidence score between LRRC28 and all glycolysis-associated proteins (Methods) and found that LRRC28 has a very high likelihood of binding to FOXK1 (Fig. 5d-e).

FOXK1 is a member of the forkhead family of transcription factors that share a structurally similar DNA-binding domain[84,85]. Interestingly, LRRC28 likely binds both the non-DNA-binding region and DNA-binding domain of FOXK1 (distance < 4 angstroms; Fig. 5e; ExtendedDataFig. 19). This observation led us to examine the specificity of the interaction between LRRC28 and FOXK1. We calculated the binding probabilities of LRRC28 with 35 other forkhead family transcription factors, finding that FOXK1 exhibits the strongest evidence of physical interaction with LRRC28 (Fig. 5f). This specificity is potentially due to a unique binding site between LRRC28 and FOXK1's non-DNA-binding region (Fig. 5e). As an additional control, we selected LRRC42, a protein with leucine-rich repeats that does not exhibit TEC with glycolytic genes. As expected, LRRC42 showed a very low likelihood of interaction with any of the glycolytic genes, including FOXK1 (Fig. 5d). These findings suggest that LRRC28 may serve as a regulator of glycolysis by binding to FOXK1, thereby preventing FOXK1 from binding to the promoter regions of glycolytic genes and leading to the downregulation of glycolysis. Taken together, TEC reveals shared biological functions and predicts novel associations, providing insights not attainable with RNA co-expression analysis alone.



437

438 **Fig. 5 | TEC enables the prediction of novel gene functions. a,** We predicted that LOX belongs
439 to the collagen-containing extracellular matrix using an older version of human GO terms (January
440 1, 2021) and confirmed this prediction with the newer version (December 4, 2022; Methods). The
441 network displays the similarity in TE between LOX (yellow dot) and other genes (gray dots) from
442 the collagen-containing extracellular matrix. Line weight in figure panels indicates the absolute
443 value of rho from 0.1 to 1. **b,** The networks display the rho between LRRC28 and glycolytic genes
444 at the TE level (on the left) and RNA level (on the right) in humans. Green dots represent genes
445 that belong to the glycolysis pathway, purple nodes are transcription factors that regulate

glycolysis. **c,** TE and RNA expression of *LRRC28*, glycolytic genes, and transcription factors regulating glycolysis (*FOXK1*, *FOXK2*) across human cell types and tissues. **d,** We used AlphaFold2-Multimer to calculate the binding probabilities between the proteins LRRC28 or LRRC42 and glycolytic proteins (Methods). We evaluated the models with ipTM+pTM (x-axis) and precision of protein-protein interface binding predictions (pDOCKQ; y-axis). We set a threshold of ipTM+pTM > 0.7[129] and pDOCKQ > 0.23[130,131] as previously suggested to identify confident binding. **e,** 3D model of binding between LRRC28 and FOXK1. For visualization purposes, we removed residues 1-101 and 370-733 in FOXK1 (pLDDT scores below 50). **f,** Binding probabilities between LRRC28 and transcription factors belonging to the forkhead family[127]. The dashed lines represent ipTM+pTM > 0.7 or pDOCKQ > 0.23.

Genes with positive TEC are more likely to physically interact

The predicted binding between LRRC28 and FOXK1 suggests the utility of TEC to reveal physical interactions between proteins. Proteins that physically interact tend to be co-expressed at the RNA level[17,23,86], and many protein complexes are assembled co-translationally[87], leading us to hypothesize that the TE of interacting proteins may be coordinated across cell types. Specifically, we expect that there should be positive covariation between the TE of interacting proteins to ensure their coordinated production[40,41]. To test this hypothesis, we categorized gene pairs by whether they display positive or negative similarity in RNA expression or TE across cell types. We observed that nearly one-third of the known pairwise protein-protein interactions (STRING database[86], only considering the physical interaction subset) exhibited the same direction of similarity (positive rho scores) at both RNA expression and TE levels (Fig. 6a). Compared to all possible pairs (124,322,500), or those with the same biological function (6,492,564), physically interacting pairs of proteins (1,030,794) were substantially enriched for positive similarity of TE and RNA expression patterns (Fig. 6a; chi-square test $p < 2.2 \times 10^{-16}$ and 1.88-fold enrichment compared to all pairs; table S18). Additionally, we found that negative rho values were significantly depleted in protein-protein interactions compared to gene pairs derived from GO terms (Fig. 6a). Though we found enrichment of gene pairs only at RNA expression level, this may be due to neighboring genes being frequently coexpressed (ExtendedDataFig. 20[27,28]). This result aligns with the notion that genes with the same function can be regulated in opposite directions, as indicated by negative rho values, in contrast to physically interacting proteins[88,89].

We then examined whether these patterns generalize to the higher-order organization of protein complexes. We observed protein complexes (as defined by hu.MAP[90]) displayed positive TEC and RNA co-expression (Fig. 6b; Methods). Noticeably, while proteins within the same complex generally exhibited similar positive patterns in both TEC and RNA co-expression, certain interactions within protein complexes were particularly evident only at the TE level (Fig. 6c-e). For instance, members of the exocyst complex showed a strong positive TEC but not RNA co-expression (Fig. 6c-e). The exocyst complex consists of eight subunits in equal stoichiometry, forming two stable four-subunit modules[91,92]. Several known exocyst-binding partners are not required for its assembly and stability, indicating that the molecular details are still unclear[91]. Our finding suggests that translational regulation may play a role in maintaining the proper stoichiometry of the exocyst complex. In summary, physically interacting proteins are likely to have positive TEC in addition to positive RNA co-expression profiles. The positive correlation in

RNA abundance and TE among physically interacting proteins may reflect an evolutionary pressure to efficiently utilize energy resources[40,41,93].

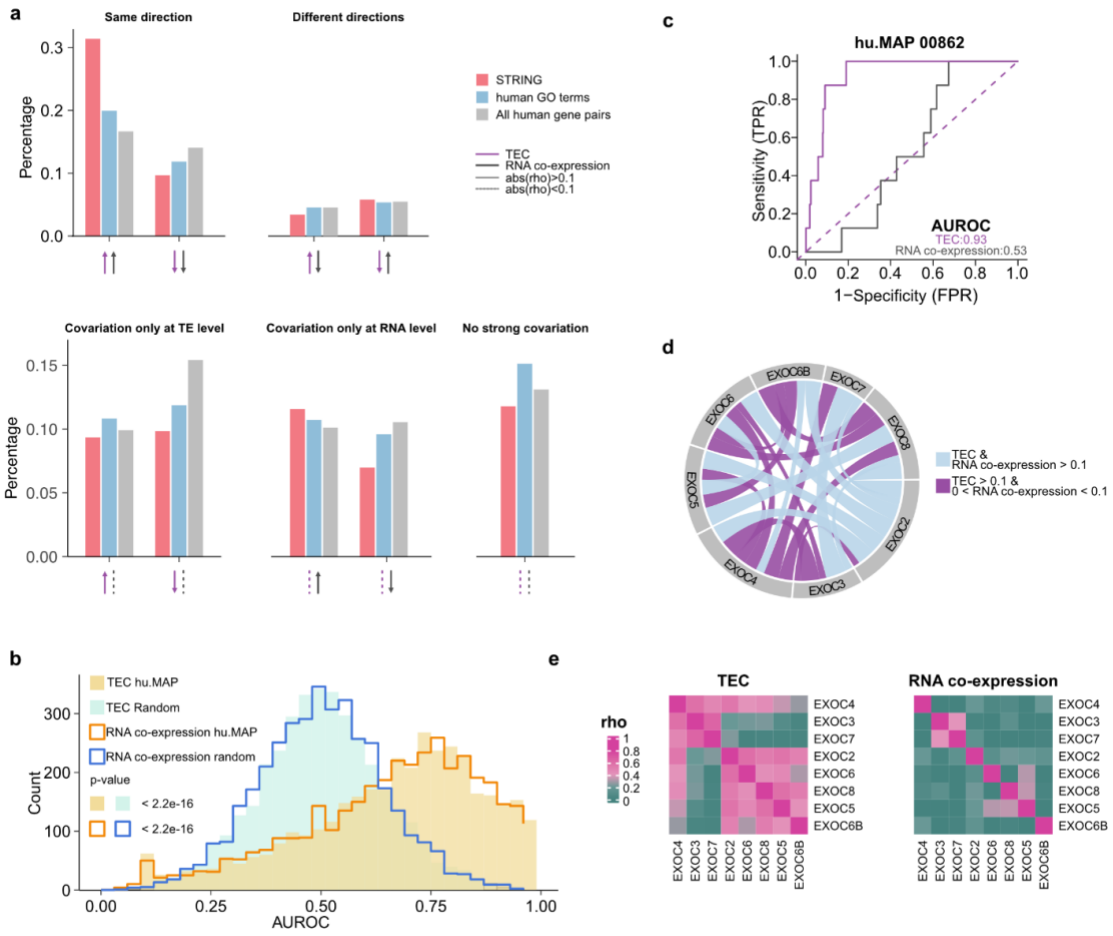


Fig. 6 | Physically interacting proteins display TEC. a, Solid lines indicate gene pairs with absolute rho greater than 0.1, while dashed lines represent those with absolute rho less than 0.1. Number of pairs of genes among three sets (physical interaction-red; shared function-blue; all genes-gray) categorized based on the direction of correlation. **b,** The distribution AUROC calculated with either TEC or RNA co-expression for 3,755 hu.MAP terms (Methods). The distribution was compared to AUROC for each term that is randomly assigned genes with size matched to the original hu.MAP term. P-values were calculated using a two-sided Wilcoxon test. **c,** AUROC plot for hu.MAP term 00862, which includes eight genes within the exocyst complex. **d,** Connections represent gene pairs with rho scores above 0.1. Purple lines indicate pairs connected at TE level alone, while blue lines depict those at both the RNA co-expression and TE levels. **e,** Heatmaps display the rho calculated among genes at the TE (left) and RNA expression

levels (right).

DISCUSSION

In this study, we analyzed thousands of matched ribosome profiling and RNA sequencing experiments across diverse human and mouse cell lines and tissues to quantify TE. A particular challenge in this effort was inadequate metadata associated with these experiments, which hampers their reuse. A particularly recurrent issue was inconsistencies in cell line identification (supplementary text). Additionally, metadata matching of RNA-seq and ribosome profiling data is necessary to quantify TE, yet this information is missing in current databases. To address these issues, we conducted a manual curation process. Given that the analyzed experiments were predominantly described in peer-reviewed publications, we anticipated the publicly accessible data would be of sufficient quality for large-scale analyses. However, more than 20% of the human and mouse experiments did not meet fundamental quality control criteria, such as ribosome footprints arising from coding regions and adequate transcript coverage, thereby deeming these studies unsuitable for further analyses. Our findings point to a pressing need for stricter data quality standards and more comprehensive, structured metadata in genomic databases.

We made several advances including the selection of RPF read lengths, data normalization, and estimation of TE (supplementary text). TE is typically defined as a log ratio of read counts from ribosome profiling and RNA expression measurements which often leads to spurious correlations between TE and RNA levels[55]. We instead employed a compositional data analysis framework for both ribosome profiling and RNA-seq[50,56,94], allowing for a more accurate estimation of TE as evidenced by improved correlation of these values with corresponding protein abundance. In this study, we used the term "translation efficiency" consistent with its established use in prior literature. Recent work has suggested that ribosome occupancy normalized for mRNA abundance may not directly indicate the efficiency of protein synthesis at least in the context of reporter constructs[95]. While there are mechanisms that lead to a decoupling between ribosome density and the rate of protein synthesis, our work and others indicate that TE as defined here is significantly correlated with protein abundance and synthesis rates for endogenous transcripts[96].

In this study, we introduce the concept of translation efficiency covariation (TEC) which quantifies the similarity of translation efficiency patterns across cell types. Among orthologous gene pairs, RNA co-expression relationships were shown to be conserved across evolution[11]. Our analyses demonstrated that covariation patterns of TE are also globally conserved between, highlighting the functional relevance of these patterns. Future research leveraging network level conservation metrics could provide further insights into TEC and RNA co-expression networks. Specifically, identifying conserved and divergent subnetwork properties between TE and RNA co-expression networks could elucidate specific regulatory interactions.

RNA co-expression among genes is known to be associated with shared functions[9–11]. Our analysis indicates that TEC is also informative regarding gene function (Table 1; supplementary text). Interestingly, while for a given transcript, average RNA expression and TE across cell types are only weakly correlated, genes with particular biological functions display highly coordinated patterns of both RNA expression and translation efficiency. This coordination may enhance

cellular energy conservation and responsiveness to environmental cues.

In addition, TEC revealed unique insights into protein function that elude RNA or protein-based analyses. A notable example is the covariation of TE between *LRRC28* and glycolytic genes, whose RNAs are not co-expressed. We discovered a high confidence predicted interaction between LRRC28 and FOXK1, the key transcription factor controlling glycolytic enzyme expression. Although LRRC28 is down-regulated in several cancers compared to normal tissues[97–99], the functional relevance, if any, remains unknown. These patterns were also not easily detectable at the protein-level as LRRC28 is absent from most proteomic databases such as PAXdb and ProteomeHD[23,100]. Taken together, these findings emphasize the unique insights provided by TEC that escape RNA and protein co-expression analyses.

TEC between LRRC28 and its potential physically interacting partner prompted us to systematically analyze the similarity of translation efficiency patterns across protein complexes. We found a significant enrichment of positive TEC between physically interacting protein pairs, establishing that physically interacting proteins often exhibit coordinated translation efficiencies across different cell types. This coordination may facilitate the co-translational assembly[87] of certain protein complexes and contribute to their stoichiometric production. Such RNA and translation level coordination between physically interacting proteins likely enhances the efficiency of complex formation and optimizes the energetic costs associated with these processes. This optimization is particularly advantageous given that protein biosynthesis is the largest consumer of energy during cellular proliferation[41,45,93].

It is important to acknowledge several limitations in our study that may impact the accuracy of TE calculations. First, the limited number of samples available for certain cell lines may lead to less accurate estimates of the translation for those cell types. Second, we only considered a representative transcript[102] for each gene based on criteria such as conservation, structure, and functional domains. Mapping RPFs to multiple isoforms of a single gene presents challenges due to the inherently short length of RPFs. This simplification may confound results for genes that have multiple isoforms with distinct expression patterns. In summary, our analyses reveal TEC is informative in uncovering gene functions, is conserved between humans and mice, and suggests simultaneous coordination of both RNA expression and translation among physically interacting proteins, establishing translation efficiency covariation as a fundamental organizing principle of mammalian transcriptomes.

ACKNOWLEDGMENTS

We thank all contributions to metadata curation: Hansel Chiang, Ashley Hoffman, Tori Tonn, Alia Segura, Charisma Tante, Eric Vasquez, and Liaoyi Xu. We also thank Dr. Vighnesh Ghatpande and Victoria D. Chapman for generating the KO cell lines and assisting with the preparation of the sequencing library. We appreciate Dr. Milad Miladi for providing critical feedback, and the original text in this paper was written by the authors. A LLM was used to suggest edits for clarity and grammar[140]. The authors acknowledge the Texas Advanced Computing Center (TACC) at The University of Texas at Austin for providing high-performance computing and storage resources that have contributed to the research results reported within this paper. URL:

http://www.tacc.utexas.edu.

Research reported in this publication was supported in part by the National Institute Of General Medical Sciences of the National Institutes of Health under Award Number R35GM150667 (CC). This work was also supported by the National Institutes of Health grant [HD110096], and the Welch Foundation grant [F-2027-20230405] (C.C.). C.C. was a CPRIT Scholar in Cancer Research supported by CPRIT Grant [RR180042].

AUTHOR CONTRIBUTIONS

Y.L., I.H., and C.C. co-wrote the original manuscript. Y.L. and I.H. generated the figures for the manuscript. H.O., M.G., and J.C. downloaded all the data from GEO and processed raw sequencing data. Y.L. and C.C. developed the translation efficiency calculation pipeline. J.C. and C.C. designed and implemented the winsorization method. Y.Z. performed the deduplication comparison. Y.L. and C.C. developed the translation efficiency covariation analysis and function prediction pipelines. Y.L., K.Q., H.O. performed the quality control analysis for all sequencing data. Y.L. carried out covariation analysis, gene function prediction, and AlphaFold2 analysis. I.H. conducted the RBP analysis. L.P., J.W., D.Z., and V.A. assessed the quality of TE measurements by developing machine learning approaches. H.O., J.W., D.Z., V.A., Q.Z., and E.S.C. provided suggestions for the manuscript. Y.L., Q.Z., and E.S.C. conducted literature search to evaluate gene function predictions. I.H., S.R., and D.P. performed all experiments. C.C. provided study oversight, conceptualized the study and acquired funding. All authors approved the final manuscript.

DATA AVAILABILITY

Metadata about RiboBase can be found in Supplementary table S1. Ribo files for the HeLa cell line are accessible at https://zenodo.org/records/10594392. Full TEC and RNA co-expression matrices are accessible via Zenodo repository at: https://zenodo.org/uploads/10373032. A RiboFlow configuration file and processed ribo files can be accessed at https://zenodo.org/uploads/11388478. Sequencing data and ribo files for the RBP knockout experiments are available on GEO GSE269734. Data will be publicly released upon successful review of this article.

CODE AVAILABILITY

The code used in the study is available at https://github.com/CenikLab/TE_model/tree/main. Code will be publicly released upon successful review of this article.

DECLARATION OF INTERESTS

D.Z., J.W. and V.A. are employees of Sanofi and may hold shares and/or stock options in the company. H.O. is an employee of Sail Biomedicines.

METHODS

Acquisition and curation of ribosome profiling data

We used keyword search ("ribosome profiling", "riboseq", "ribo-seq", "translation", "ribo", "ribosome protected footprint") to determine studies that may employ ribosome profiling in their experimental design, from the Gene Expression Omnibus (GEO) database, with a cutoff date of January 1, 2022. Search results were manually inspected and studies containing ribosome profiling data were kept. Organism, cell line, publication, and short read archive (SRA) identifiers were obtained by automatically parsing the GEO pages of the corresponding study and sample. There was no dedicated experiment-type field for ribosome profiling experiments in GEO. Therefore we determined the experiment type (ribosome profiling, RNA-Seq, or other) of each sample by manually inspecting the GEO metadata and the associated publication of the study. Typically, ribosome profiling samples were indicated in GEO using one of the following terms: "ribosome protected footprints", "ribo-seq", and "ribosome profiling" in various parts of the metadata such as title, extraction protocol, and library strategy. If there were RNA-Seq samples in the same study, they were matched with ribosome profiling experiments, where available, after inspecting the sample names, metadata, and the publication of the study.

Adapters are commonly observed on the 3' end of sequencing reads in ribosome profiling experiments, a consequence of the inherently short length of RPFs. If the 3' adapter sequence was listed in GEO, we extracted it as part of the manual data curation process. If this sequence was unavailable, we attempted to determine it from the corresponding publication of the study. If no explicit sequence was available, we computationally analyzed the sequencing reads and searched for commonly used adapters which are CTGTAGGCACCATCAAT, AAGATCGGAAGAGCACACGTCT, AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC, TGGAATTCTCGGGTGCCAAGG and AAAAAAAAAA. If any of these adapters were found in at least 50% of the reads, we used the detected sequence as the 3' adapter. If no match was found, we removed the first 25 nucleotides of the reads anchored 6 mers and tried to extend them. If any of these extensions reached 10 nucleotides and were still detected in at least 50% of the reads, we took the highest matching sequence as the 3' adapter. On the other hand, for sequencing reads from SRA having a length of less than 35 nucleotides, we assumed the 3' adapters had already been removed. Detailed code can be accessed from: https://github.com/RiboBase/snakescale/blob/main/scripts/guess_adapters.py.

RiboBase was pre-populated after mining GEO. Then data curators were assigned specific studies and used the web-based interface to access the database. Each study was curated independently by at least two people. In case of disagreements, an additional experienced scientist inspected the corresponding studies and publications to make the final decision. We supplemented any missing metadata from GEO by checking the corresponding publications to ensure completeness. The result of this data curation process with information such as cell line, organism, and matched RNA-seq can be found in table S1, which forms the metadata backbone of RiboBase.

Ribosome profiling and RNA-seq data processing

For each selected study in GEO, ribosome profiling and matching RNA-Seq reads (where available) were downloaded, from SRA, using the SRA-Tools version 2.9.1[103], in FASTQ format using their accession numbers. FASTQ files were processed using RiboFlow[51] where parameters were determined using the metadata in RiboBase. The reference files for human and mouse

transcriptomes, annotations, and non-coding RNA sequences are available at https://github.com/RiboBase/reference_homo-sapiens and https://github.com/RiboBase/reference_mus-musculus, respectively. Briefly, the 3' adapters of the ribosome profiling reads were trimmed using Cutadapt version 1.18[104] and reads having lengths between 15 and 40 nucleotides were kept. Then, reads were aligned against noncoding RNAs, and unaligned reads were kept. Next, reads were aligned against transcriptome reference, and alignments having mapping quality score above 20 were kept. Reads having the same length and mapping to the same transcriptome position were collapsed, which we refer to as "PCR deduplication". In the final step, we compiled the alignments into ribo files using RiboPy[51]. All alignment steps used bowtie2 version 2.3.4.3[105]. For each sample, we also performed the same run without the PCR deduplication step. We developed a pipeline, Snakescale, available at https://github.com/RiboBase/snakescale, to automate the entire process from downloading the data from SRA to generating the ribo files. Snakescale went over the selected list of studies and obtained their metadata from Ribobase, downloaded the sequencing data from SRA, generated Riboflow parameters file, and ran Riboflow to generate the ribo files. Examples of non-deduplicated ribo files for the HeLa cell line can be accessed at https://zenodo.org/records/10594392[106].

To visualize the length distribution of the RPFs, we applied the scale function (z-score) in R to normalize the count of RPFs mapped to CDS regions with PCR-deduplicated ribosome profiling data. Subsequently, we plotted the distribution of these normalized RPFs using the heatmap (Fig. 1c; ExtendedDataFig. 2).

Determination of cutoff for RPF lengths and quantification of ribosome occupancy

Ribosome profiling experiments employ a range of ribonucleases including RNase I, RNase A, RNase T1, and MNase (i.e., micrococcal nuclease) (S7). These different enzymes lead to variable RPF lengths[53,107–109]. To ensure that we retain high-quality RPFs for further analyses, we implemented a dynamic extraction module that automatically selected lower and higher boundaries of RPFs for each sample. Initially, we determined the first RPF length, ranging from 21 to 40 nucleotides, that contained the highest number of CDS mapping reads. Then, we examined the two positions adjacent to this selected position. The extension of the position was carried out on either side to include a higher number of CDS-aligned reads. This extension process was repeated until it encompassed at least 85% of the total CDS reads within the 21 to 40 nucleotides range (ExtendedDataFig. 3a). The final two positions identified were designated as the lower and upper boundaries. If these boundaries extended to either 21 or 40 nucleotides without including a sufficient number of reads, then 21 or 40 nucleotides, respectively, were set as the final boundaries. This approach was employed to establish the RPF cutoffs for each sample.

Transcript coverage and quality control for ribosome profiling data

We performed quality control using RPFs that were deduplicated based on the length and position (PCR-deduplication) ribo files (Fig. 1d-e). We set two cutoffs for ribosome profiling quality control. We required that on average each nucleotide of the transcript should be covered at least 0.1 times (0.1X). Coverage was calculated with the formula:

$$coverage = \frac{total\ nucleotides\ from\ reads\ mapped\ to\ transcripts}{total\ length\ of\ transcripts}$$

Additionally, samples with CDS mapping read percentage of 70% or higher were retained for subsequent analysis.

To assess the pattern of three nucleotide periodicity that is typically associated with ribosome profiling experiments, we first selected the length of RPFs with the highest number of counts from the PCR deduplicated ribo files. We then assigned all CDS mapping reads to one of three coding frames based on the position of their 5' end. We aggregated the results for all genes for each sample. To facilitate comparison, we reordered the counts for each position of the three nucleotide periodicity from highest to lowest and converted these counts into percentages for each sample.

We initially classified samples based on the differences between positions 1 and 2. We identified Group 1 by selecting samples where the difference did not exceed the 10th percentile of these differences between positions 1 and 2. For the remaining samples, we further classified them based on the differences between positions 2 and 3. Similarly, samples that did not exceed the 10th percentile of these differences between positions 2 and 3 among remaining samples were classified to Group 3, while the rest samples were Group 2. We further summarized the samples based on their QC status.

We classified samples from Group 1 as exhibiting three-nucleotide periodicity. The percentage of samples following three-nucleotide periodicity was calculated by dividing the number of Group 1 samples by the total number of samples across all three groups.

PCR and Unique Molecular Identifiers (UMIs) deduplication comparison

We selected eight ribosome profiling experiments that incorporated UMIs into the sequence library preparation to assess the impact of different deduplication methods. Specifically, these samples are GSM4282032, GSM4282033, and GSM4282034 from GSE144140[110]; GSM3168387, GSM3168389, GSM3168390 from GSE115162[108]; and GSM4798525, GSM4798526 from GSE158374[34]. We processed the data using Riboflow, applying three different deduplication methods: non-deduplication, PCR deduplication, and UMI deduplication. The yaml files are available at https://github.com/CenikLab/TE_model/tree/main/riboflow_scr. The RPF length cutoffs for samples from GSE144140 and GSE115162 are listed in table S6. Since GSE158374 is not currently included in RiboBase, we manually performed the dynamic module and selected 28 to 32 as the RPF cutoff for this study.

Winsorization of CDS mapping read counts

To address the issue of reduced usable reads resulting from PCR deduplication (supplementary text), we employed a winsorization method, which was previously proposed for tackling this problem[40,111]. For each gene's CDS region, we obtained the distribution of non-deduplicated nucleotide counts and calculated the 99.5th percentile value. This calculation was based on reads whose lengths fell within the RPF range determined by the RPF boundary selection function. RPF counts that exceed the 99.5th percentile were capped to the value corresponding to the 99.5th

percentile. This method was designed to mitigate the impact of outlier values that might arise due to disproportionate amplification during the PCR process[40].

Gene filtering and normalization for ribosome profiling and RNA-seq

RNA-seq experiments in RiboBase utilized several different strategies to enrich mRNAs. The two most common approaches were the depletion of ribosomal RNAs and the enrichment of transcripts by polyA-tail selection. This difference leads to dramatically different quantification of a subset of genes that lack polyA-tails (e.g. histone genes, ExtendedDataFig. 6c). Hence, we removed 166 human and 51 mouse genes identified as lacking polyA tails (table S8-9)[112,113].

We normalized both PCR-deduplicated RNA-seq data and winsorized non-deduplicated ribosome profiling data with counts per million (CPM) after removing the genes without polyA-tails. Genes with CPM greater than one in over 70% of the total samples in both RNA-seq and ribosome profiling for either human or mouse were included in further analyses. 11,149 human and 11,434 mouse genes were retained using this approach. We have summed the counts of all polyA genes that were filtered out and grouped them under 'others' in the count table.

Validation of manual curation and quality control by matching between RNA-seq and ribosome profiling from RiboBase

We assessed the manual matching of ribosome profiling (winsorization) and RNA-seq (PCR dedupication) data in RiboBase by establishing a matching score for the samples that successfully passed quality control (transcript coverage > 0.1X and CDS percentage > 70% with PCR-deduplicated ribosome profiling data). We calculated the coefficient of determination (R^2) using the Centered Log Ratio (CLR) transformed gene counts. This was done for each ribosome profiling sample against all corresponding RNA-seq samples within the same study. Subsequently, for each ribosome profiling sample, we calculated the difference between the R^2 of its matching pair from RiboBase and the mean R^2 of the non-matching pairs within the same study. The difference was defined as the matching score.

To remove poorly matched samples in both human and mouse datasets, we established a cutoff based on the R^2 from the matched ribosome profiling and RNA-seq data in RiboBase. Any sample with an R^2 lower than 0.188 in either human or mouse, which is Q1 - 1.5 * IQR of mouse R^2 distribution, was considered a poor match and consequently excluded from further analysis (ExtendedDataFig. 6b).Finally 1,054 human and 835 mouse ribosome profiling experiments with their matched RNA-seq were used for TE calculation.

Translation Efficiency (TE) calculation

CLR normalized counts from PCR-deduplicated RNA-seq and winsorized non-deduplicated ribosome profiling were used to calculate TE with compositional linear regression[49,94,114]. In our linear regression approach, ribosome profiling data served as the dependent variable, while the corresponding RNA-seq data provided the explanatory variable. The first step involved transforming the gene count, which includes 'others', into CLR normalized compositional vectors. Given the constraints of count data within a simplex, a further transformation from CLR to

Isometric Log Ratio (ILR) was necessary for linear regression[49]. This transformation is crucial as it allows the compositional data to be decomposed into an array of uncorrelated variables while preserving relative proportions. The ILR transformation projects the original data onto a set of orthonormal basis vectors derived from the Aitchison simplex. Then the linear regression model applied to these transformed variables can be represented as:

$$Y = b + B * X$$

Where Y is the ILR-transformed ribosome profiling data and X is the ILR-transformed RNA-seq data. The model assumes a normal distribution:

$$Y \sim N \wedge (D-1)\ (Y, \Sigma\varepsilon)$$

Where Σε represents the residual variances. These residuals were then extracted from each sample and reconverted to CLR coordinates which are used as the definition of TE for each gene in each sample. Finally, we averaged TE for different cell lines and tissues (Fig. 2b, ExtendedDataFig. 7), and reported the TE in table S10-11. The scripts to generate TE are available at https://github.com/CenikLab/TE_model.

Correlation between translation efficiency and protein abundance

We assessed the correlation between TE and protein abundance from seven human cell lines (A549, HEK293, HeLa, HepG2, K562, MCF7, and U2OS). The protein measurements were obtained from PAXdb[100]. 9924 genes were shared between our TE and the protein abundance data. We calculated the Spearman correlation coefficient for each cell line using the 'stats' package in R to evaluate the relationship between TE and protein abundance.

Conservation of translation efficiency between orthologous genes from human and mouse

Orthologous genes between human and mouse were identified using the 'orthogene' package from Bioconductor[115] using the parameters 'standardise_genes=TRUE, method_all_genes="homologene", non121_strategy="keep_both_species"'. A single human gene could correspond to multiple mouse orthologs or vice versa. To maintain all one-to-many matches in our analysis, each correspondence is represented by multiple rows in our table (if a human gene 'A' is orthologous to mouse genes 'B' and 'C', we generate two separate rows: 'A-B' and 'A-C'). Human genes lacking corresponding mouse orthologs were excluded or vice versa. As a result, a total of 9,194 gene pairs were identified as orthologous between human and mouse (table S12)

To capture the variability in TE and mRNA expression between orthologous genes in human and mouse, we measured the standard deviation using the metric standard deviation (msd) function from the 'compositions' package in R[116]. We observed a negative Spearman correlation coefficient between msd of TE and mean TE, as well as msd of RNA expression and mean RNA expression, in both species. To address the dependency between msd and mean values, we conducted a partial correlation analysis. For example, we adjusted the human msd values using the mean TE from both human and mouse with the 'pcor.test' function from the 'ppcor' package[117].

GO term analysis was performed using FuncAssociate 3.0, accessible at

http://llama.mshri.on.ca/funcassociate/[118]. For this analysis, we set either 9,194 mouse or 9,189 human orthologous genes as the background. We generated association files for these genes with the December 4, 2022 version of human or mouse GO terms. In the human or mouse association file, we only kept those GO terms containing at least 10 genes for further analysis.

Assessment of methods for calculating genes' similarity with ribosome occupancy data

We used eight commonly used methods to quantify the similarity of ribosome occupancy across cell types for all pairs of 11,149 human or 11,434 mouse genes in RiboBase.

Method 1 - CPM-normalized ribosome footprint counts were used to calculate the Pearson correlation coefficient as implemented in the stats R package.

Method 2 - Quantile-normalized (customized Python script) ribosome footprint counts were used to calculate the Pearson correlation coefficient.

Method 3 - Ranking of ribosome footprint counts was used to calculate the Spearman correlation coefficient as implemented in the stats R package.

Method 4 - CLR-normalized ribosome footprint counts were used to calculate the proportionality (rho scores) between genes as implemented in the propr package with lr2rho function[50].

Method 5 - CPM-normalized ribosome footprint counts were used to calculate the similarity between genes with a decision tree-based method as implemented in the treeClust package[23,119]. We applied the 'treeClust.dist' function with a dissimilarity specifier set to d.num=2.

Method 6 - Quantile-normalized ribosome footprint counts were used to calculate the similarity between genes with the decision tree-based method.

Method 7 - CPM-normalized ribosome footprint counts were used to calculate gene similarity with the generalized least squares (GLS) method[120].

Method 8 - Quantile-normalized ribosome footprint counts were used to calculate gene similarity with the GLS method.

We compared these eight ribosome occupancy similarity matrices to determine the most effective method for constructing gene relationships with respect to biological functions. This assessment employed the guilt by association principle to ascertain the functional coherence within a gene matrix, determining if genes associated with a particular biological function (GO terms[121], TOP mRNAs[122]) exhibit similar expression patterns and network interactions[123].

The complete ontology was sourced from the Gene Ontology website, with the files goa_human.gpad.gz and mgi.gpad.gz, generated on December 4, 2022[121]. The annotation of Gene Ontology terms was accomplished with the aid of the org.Hs.eg.db and org.Mm.eg.db R packages[124,125]. We restricted the selection of GO terms to those associated with the 11,149 human and 11,434 mouse genes that had passed gene filtering. We used GO terms associated with at least 10 but less than 1,000 genes for evaluation, yielding a total of 2,989 human and 3,340 mouse GO

terms.

We then employed the neighbor-voting algorithm to assess the covariations of ribosome occupancy among genes from the same GO term with AUROC[123]. Specifically, we first converted the similarity scores to absolute values. Then we extracted genes associated with a specific function and implemented the leave-one-out cross-validation method. For this analysis, we iteratively masked one gene at a time, treating it as if it did not belong to the function. In each iteration, we calculated the total sum of similarity scores from all genes not belonging to the function to all the remaining genes within the function. We normalized the sum of similarity scores for each gene against the sum of similarity scores for that gene with all genes. After normalization, we converted these normalized similarity scores into rankings. We retained the rankings only for genes that belong to this specified functional property. Finally, we computed the AUROC for all genes within this functional property based on these rankings. A detailed script for genes' functional similarity pattern analysis can be found: https://github.com/CenikLab/TE_model/blob/main/other_scr/benchmarking.R.

RNA co-expression and translation efficiency covariation

We introduce the concept of TEC, which employs a compositional data analysis approach[49,50] to quantify the similarity patterns of TE across various cell and tissue sources, as described in Method 4 above. The proportionality scores were calculated with the following formula from the propr package with lr2rho function[50]:

$$Rho(Ai, Aj) = 1 - var(Ai - Aj) / (var(Ai) + var(Aj))$$

Where Ai and Aj represent TE values for genes i and j from the TE matrix A.

In this study, the TEC was calculated with 77 human cell lines for 11,149 genes or 68 mouse cell lines for 11,434 genes. The proportionality coefficients (rho scores) generated from this method range from -1 to 1. Full TEC and RNA co-expression matrices are accessible via Zenodo repository at: https://zenodo.org/uploads/10373032.

Evaluation of the ability of TEC to predict novel gene functions

We compared the AUROC between an older version of GO terms (January 1, 2021) to the newer version of GO terms (December 4, 2022) to identify genes that had been newly added to from the GO terms in this timeframe. GO terms were downloaded and filtered to include only those terms containing between 10 and 1,000 genes with either human or mouse backgrounds (11,149 human genes or 11,434 mouse genes). We selected 184 human and 238 mouse GO terms from the older version that demonstrated high TEC similarity (AUROC > 0.8) among genes within the same term for predicting novel gene functions. We first converted the rho scores for TEC between gene pairs to absolute values. For genes not currently included in the GO terms, we calculated the sum of rho for each gene relative to all genes within the term, based on either TE or mRNA expression levels. We then normalized these rho sums for each gene against the total rho sum of that gene across all 11,149 human genes or 11,434 mouse genes. These normalized values were converted into ranking percentages to reflect the likelihood of these genes being associated with the respective GO term.

Finally, we identified the top-ranking genes as potentially new additions and cross-validated them with the newer version of the GO terms to confirm our predictions.

Prediction of novel gene functions with TEC

We analyzed 243 human and 310 mouse GO terms as of December 4, 2022, which demonstrated high similarity patterns between genes in TE level (AUROC > 0.8) to predict novel gene functions. Absolute TEC rho scores served as the input for biological function prediction (GO terms). The prediction method followed the same protocol as our previous evaluations of TEC's ability to predict novel gene functions. However, we added a filter step: a newly predicted gene was retained only if its average rho score with other genes within the same term exceeded the overall average rho score for all existing genes in that term. This prediction analysis was performed using a custom script that can be found at https://github.com/CenikLab/TE_model/blob/main/other_scr/prediction.R.

Computational evaluation of the interaction between LRRC28, glycolytic proteins, and proteins from forkhead TF family

We computed the pair-wise interaction probabilities between LRRC28 or LRRC42 and glycolytic proteins (HK1, HK2, PFKL, PFKM, PFKP, TPI1, PGK1, ENO1, ENO2, and PKM) with AlphaFold2-Multimer 2.3.0[83,126]. In addition, we also calculated pairwise interaction probabilities for LRRC28 with 35 proteins from the forkhead transcription TF family[127]. We extracted the canonical amino acid sequence for each gene from UniPort[128] as the input file. We set 0.7 as the cutoff of ipTM+pTM as a high-confidence protein structure and binding probability cutoff[129]. We then evaluated the interfaces predicted by AlphaFold2-Multimer, using a pDOCKQ score greater than 0.23 as our criterion for reliability[130,131].

Benchmarking TEC and RNA co-expression for protein interactions

Using a similar approach to our benchmarking with biological functions, we employed the neighbor-voting algorithm to assess physical protein interactions based on rho scores among genes at either the TE or mRNA expression level. We first kept the non-negative rho between genes and set negative rho to zero. We then analyzed similarity patterns between genes from the same protein complex, downloading from the hu.MAP 2.0 website[90]. In this process, we excluded genes from hu.MAP terms that were not in the 11,149 human gene list, resulting in 8,024 overlapping genes between our list and hu.MAP terms. Furthermore, we removed hu.MAP terms that included fewer than three genes. This filtering process left us with 3,880 hu.MAP terms, among which 3,755 contained unique genes.

Since proteins within the same complex may not physically interact, we used physical interaction pairs downloaded from the STRING website instead of gene pairs from hu.MAP terms to summarize the interactions in Fig. 6a.

Identification of enriched RNA motifs among genes with high degree of TEC

To reduce bias in motif enrichment analysis that may arise by ribosome footprint mapping to

paralogous genes, we removed predicted paralogs from each GO term using Paralog Explorer[132] (DIOPT score > 1). Then, we enumerated heptamers in each transcript region using the Transite kmer-TSMA method[133] with default parameters for each species (human, mouse), transcript region (5' UTR, CDS, 3' UTR), and GO term (selected terms with TE AUROC > 0.7, TE-RNA AUROC difference > 0.2, and number genes after paralog removal >= 12). For mice there were three terms and for humans there were eight. We selected the three mouse terms and top five terms in humans with the highest number of genes and greatest AUROC difference.

After counting heptamers with Transite, we selected motifs that had >20 hits among genes in the GO term to address assumptions of uniformity near p-values of 1 for some multiple-test correction methods. Then, we used the Holm method to correct p-values for each species separately, and selected motifs with an adjusted p-value < 0.05. Finally, heptamers were annotated with RBPs included in the Transite [133] and oRNAment databases[134]. For annotation of RBPs in the oRNAment database, we required that the heptamer have a matrix similarity score[134] of 0.8 or greater when matching to each RBP position weight matrix. RBP motif hits from other species (*Drosophila*, artificial constructs) were removed from RBP annotations, and the hits to the heptamer of *Drosophila* tra2 were annotated as TRA2A for human genes with the term GO: 0140678.

eCLIP data for PABPN1, SRSF1, and TRA2A were downloaded from ENCODE[135] as BED files (K562 and HepG2 cell lines, GRCh38 reference). The BED files for biological replicates were concatenated and peaks that overlapped by at least one base pair were merged with 'bedtools merge -s -c 4,6,7 -o collapse'[136]. The resulting merged peaks were intersected with transcripts in the GO term of interest and an equal number of control transcripts (Gencode v34 GTF). The control transcripts were selected by matching on length and GC content for each transcript region (5' UTR, CDS, 3' UTR) using MatchIt[137] with default parameters. Because the gene *CARMIL2* in GO term GO:0010592 does not have a 5' UTR, required for matching, we assigned it a dummy 5' UTR with length and GC content equal to the median across all transcripts. The number of eCLIP peaks in the CDS for each RBP were summed for genes in the GO term and control genes.

Identification of RBP-gene pairs with high correlation between RBP RNA expression and gene TE

The Pearson correlation coefficient between gene TE and the RNA expression of RBPs from human and mouse[66] was tested using R stats::cor.test after taking the mean of these values by cell types and tissues. P-values were corrected with the Benjamini-Hochberg procedure, and correlations were deemed significant at a FDR < 0.05. To select RBP candidates for experimental validation, the human and mouse regulons were intersected for each RBP, and those that had more than twenty genes in the intersection and that had a mean TEC > 0.35 between genes in the intersection were chosen.

Generation of RBP knockout cell lines

For cloning the guides required for knockout cell line generation, top two ranked guides were selected from the Brunello library[138] for each RBPs (table S18). The guides were cloned in LentiCRISPRv2 (Addgene, 52961) as per the protocol[139] and confirmed by Sanger sequencing. Briefly, for lentiviral production, HEK293T cells were seeded at a density of 1.2 x 10^6 cells per

well in a 6-well plate in OPTI-MEM media supplemented with 5% FBS and 100 mM Sodium Pyruvate, 24 h prior to transfection. Both the cloned gRNA plasmids for each RBPs (700 ng of each transfer plasmid) were co-transfected with the packaging plasmids pMD2.G and psPAX2 (Addgene; 12259 and 12260) using Lipofectamine 3000 (Invitrogen) and the virus was collected as per the manufacturer's protocol. For generation of the knockout clones, HEK293T cells were seeded at a density of 5 x 10^4 cells per well in a 6-well plate in DMEM media supplemented with 10% FBS, 24 h prior to infection. Next day, the media was replaced with 1.5 ml of 1:2 diluted lentivirus containing polybrene (8 µg/mL). After 16 h, the lentivirus was replaced with fresh media and, puromycin (2 µg/mL) was added to the cells 48 h after transduction. The selection continued for 5 days followed by a period of recovery for 24 h before harvesting the cells.

Ribosome profiling and RNA sequencing of RBP knockout cell lines

Three million cells for the PARK7, USP42, and VIM knockout cell lines, along with a AAVS1 (safe harbor control) knockout line, were plated in three 10 cm^2 dishes. 27 h later cells at ~60% confluency were treated with 100 µg/mL cycloheximide (CHX) for 10 min at 37 ˚C, then collected in ice cold PBS with 100 µg/mL CHX. Cells were spun at 100 x g for 7 min at 4 ˚C, then flash frozen in liquid nitrogen and stored at -80 ˚C. Cell pellets were lysed with 400 µL lysis buffer (20 mM Tris-HCl pH 7.4, 150 mM NaCl, 5 mM $MgCl_2$, 1 mM DTT, 1% Triton-X, 100 µg/mL CHX, 1x protease inhibitor EDTA free) for 10 min on ice. Lysates were clarified by centrifugation at 1,300 rpm for 10 min at 4 ˚C. 40 µL lysate was saved for total RNA extraction, and the rest of the lysate was digested with 7 µL RNaseI for 1 h at 4 ˚C. Digestion was stopped by adding ribonucleoside vanadyl complex to a final concentration of 20 mM. Digested lysates were then loaded onto 10 mL sucrose cushion (20 mM Tris-HCl pH 7.4, 150 mM NaCl, 5 mM $MgCl_2$, 1 mM DTT, 1 M sucrose) and centrifuged at 38,000 rpm for 2.5 h at 4 ˚C using a SW41-Ti rotor. The pellet and the total RNA aliquot were both solubilized with 1 mL Trizol, and RNA was purified with the Zymo Direct-zol RNA Miniprep Kit, including DNaseI digestion.

RNA-seq

Quality of total RNA was confirmed with Bioanalyzer RNA Pico. All RIN scores were >= 9.8. Libraries were prepared from 1 µg total RNA using the NEBNext Ultra II RNA Library Prep Kit for Illumina according to manufacturer's protocol and using 8 cycles for PCR.

Ribosome profiling

Ribosome protected fragments (RPFs) were size-selected on a 15% TBE urea gel by electrophoresing at 150 V for 1.5 h. RPFs between 28-32 nt were sliced, using 28 nt and 35 nt markers as a guide. Slices were frozen at -20 ˚C for 1 h, crushed with pestles, and the RPFs were eluted in gel extraction buffer (300 mM sodium acetate pH 5.5, 5 mM $MgCl_2$) by rotating overnight at room temperature. Eluates were passed through Costar Spin-X filter tubes at 12,000 x g for 1 min 30 s. Then 1 µL 1 M $MgCl_2$, 2.5 µL GlycoBlue, and 1 mL ethanol were added and the RPFs precipitated for two days at -20 ˚C. Pellets were dried and resuspended in 16 µL water.

Libraries were generated from 8 µL RPF eluate using the Diagenode D-Plex Small RNA Kit with minor modifications: in the 3' dephosphorylation step 0.5 µL T4 PNK was supplemented and

incubated for 25 min. The RTPM reverse transcription primer was used and 8 cycles were performed for PCR. Libraries were quantified by Bioanalyzer High Sensitivity DNA Kit, pooled equimolar according to the quantity of the peak for libraries with full-length inserts (~204 nt), and cleaned up with 1.8X AMPure XP beads. Adapter dimers and empty libraries were removed by size-selection on a 12% TBE PAGE gel, followed by extraction with the crush and soak method, and final libraries were resuspended in 20 µL water.

Ribosome profiling and RNA sequencing analysis for RBP knockouts

Analysis was conducted using RiboFlow v0.0.1 with deduplication of both Ribo-seq and RNA-seq data. A RiboFlow configuration file and processed ribo files can be accessed at https://zenodo.org/uploads/11388478.

We used edgeR to measure RBP KO effects on 1) RNA abundance and 2) gene TE. To do this, we respectively modeled 1) RNA-seq counts of a specific RBP KO line to that of the other two RBP KO lines; and 2) Ribo-seq counts, contrasted with RNA-seq counts, for a specific RBP KO line compared to the other two RBP KO lines. All counts were enumerated from mapped reads to the coding regions. We originally included a control KO line (AAVS1 locus) for comparison; however, by PCA, this KO line showed a distinct gene expression signature from that of the other KO lines, indicating it may not be suitable as a control (ExtendedDataFig. 16d-e). Using the AAVS1 KO line as a control, we observed highly similar hits for each RBP KO tested. We included filtering of counts using edgeR::filterByExpr with default parameters, the TMM method for calculation of size factors, and quasi-likelihood negative binomial models for fitting. Genes were considered differential at FDR < 0.05.

REFERENCES CITED

1. Tang, F. *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* **6**, 377–382 (2009).
2. Nagalakshmi, U. *et al.* The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320**, 1344–1349 (2008).
3. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **5**, 621–628 (2008).
4. Schena, M., Shalon, D., Davis, R. W. & Brown, P. O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**, 467–470 (1995).
5. Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S. & Zhuang, X. RNA imaging. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science* **348**, aaa6090 (2015).
6. Combs, P. A. & Eisen, M. B. Sequencing mRNA from cryo-sliced Drosophila embryos to determine genome-wide spatial patterns of gene expression. *PLoS One* **8**, e71820 (2013).
7. Achim, K. *et al.* High-throughput spatial mapping of single-cell RNA-seq data to tissue of origin. *Nat. Biotechnol.* **33**, 503–509 (2015).
8. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).
9. Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U. S. A.* **95**, 14863–14868 (1998).
10. Skinnider, M. A., Squair, J. W. & Foster, L. J. Evaluating measures of association for single-cell transcriptomics. *Nat. Methods* **16**, 381–386 (2019).
11. Stuart, J. M., Segal, E., Koller, D. & Kim, S. K. A gene-coexpression network for global discovery of conserved genetic modules. *Science* **302**, 249–255 (2003).
12. Marcotte, E. M., Pellegrini, M., Thompson, M. J., Yeates, T. O. & Eisenberg, D. A combined algorithm for genome-wide prediction of protein function. *Nature* **402**, 83–86 (1999).
13. Hughes, T. R. *et al.* Functional discovery via a compendium of expression profiles. *Cell* **102**, 109–126 (2000).
14. Kim, S. K. *et al.* A gene expression map for Caenorhabditis elegans. *Science* **293**, 2087–2092 (2001).
15. Hartl, C. L. *et al.* Coexpression network architecture reveals the brain-wide and multiregional basis of disease susceptibility. *Nat. Neurosci.* **24**, 1313–1323 (2021).
16. DeRisi, J. L., Iyer, V. R. & Brown, P. O. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278**, 680–686 (1997).
17. Jansen, R., Greenbaum, D. & Gerstein, M. Relating whole-genome expression data with protein-protein interactions. *Genome Res.* **12**, 37–46 (2002).
18. Szklarczyk, D. *et al.* The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res.* **51**, D638–D646 (2023).
19. Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J. & Church, G. M. Systematic determination of genetic network architecture. *Nat. Genet.* **22**, 281–285 (1999).
20. Roth, F. P., Hughes, J. D., Estep, P. W. & Church, G. M. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat. Biotechnol.* **16**, 939–945 (1998).

21. Nusinow, D. P. *et al.* Quantitative Proteomics of the Cancer Cell Line Encyclopedia. *Cell* **180**, 387–402.e16 (2020).
22. Gonçalves, E. *et al.* Pan-cancer proteomic map of 949 human cell lines. *Cancer Cell* **40**, 835–849.e8 (2022).
23. Kustatscher, G. *et al.* Co-regulation map of the human proteome enables identification of protein functions. *Nat. Biotechnol.* **37**, 1361–1371 (2019).
24. Ryan, C. J., Kennedy, S., Bajrami, I., Matallanas, D. & Lord, C. J. A Compendium of Co-regulated Protein Complexes in Breast Cancer Reveals Collateral Loss Events. *Cell Syst* **5**, 399–409.e5 (2017).
25. Furlong, E. E. M. & Levine, M. Developmental enhancers and chromosome topology. *Science* **361**, 1341–1345 (2018).
26. Hnisz, D., Shrinivas, K., Young, R. A., Chakraborty, A. K. & Sharp, P. A. A Phase Separation Model for Transcriptional Control. *Cell* **169**, 13–23 (2017).
27. Grabowski, P., Kustatscher, G. & Rappsilber, J. Epigenetic Variability Confounds Transcriptome but Not Proteome Profiling for Coexpression-based Gene Function Prediction *. *Mol. Cell. Proteomics* **17**, 2082–2090 (2018).
28. Kustatscher, G., Grabowski, P. & Rappsilber, J. Pervasive coexpression of spatially proximal genes is buffered at the protein level. *Mol. Syst. Biol.* **13**, 937 (2017).
29. Sonenberg, N., Hershey, J. W. B. & Mathews, M. B. *Translational Control of Gene Expression*. (CSHL Press, 2001).
30. Kuersten, S. & Goodwin, E. B. The power of the 3' UTR: translational control and development. *Nat. Rev. Genet.* **4**, 626–637 (2003).
31. Baker, S. A. & Rutter, J. Metabolites as signalling molecules. *Nat. Rev. Mol. Cell Biol.* **24**, 355–374 (2023).
32. Ozadam, H. *et al.* Single-cell quantification of ribosome occupancy in early mouse development. *Nature* **618**, 1057–1064 (2023).
33. King, R. W., Deshaies, R. J., Peters, J. M. & Kirschner, M. W. How proteolysis drives the cell cycle. *Science* **274**, 1652–1659 (1996).
34. Rao, S. *et al.* Genes with 5' terminal oligopyrimidine tracts preferentially escape global suppression of translation by the SARS-CoV-2 Nsp1 protein. *RNA* **27**, 1025–1045 (2021).
35. Slobodin, B. *et al.* Cap-independent translation and a precisely located RNA sequence enable SARS-CoV-2 to control host translation and escape anti-viral response. *Nucleic Acids Res.* **50**, 8080–8092 (2022).
36. Singh, G., Pratt, G., Yeo, G. W. & Moore, M. J. The Clothes Make the mRNA: Past and Present Trends in mRNP Fashion. *Annu. Rev. Biochem.* **84**, 325–354 (2015).
37. Keene, J. D. & Tenenbaum, S. A. Eukaryotic mRNPs may represent posttranscriptional operons. *Mol. Cell* **9**, 1161–1167 (2002).
38. Keene, J. D. RNA regulons: coordination of post-transcriptional events. *Nat. Rev. Genet.* **8**, 533–543 (2007).
39. Wurth, L. *et al.* UNR/CSDE1 Drives a Post-transcriptional Program to Promote Melanoma Invasion and Metastasis. *Cancer Cell* **36**, 337 (2019).
40. Li, G.-W., Burkhardt, D., Gross, C. & Weissman, J. S. Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources. *Cell* **157**, 624–635 (2014).
41. Taggart, J. C. & Li, G.-W. Production of Protein-Complex Components Is Stoichiometric and

Lacks General Feedback Regulation in Eukaryotes. *Cell Syst* **7**, 580–589.e4 (2018).
42. Ishikawa, K. Multilayered regulation of proteome stoichiometry. *Curr. Genet.* **67**, 883–890 (2021).
43. Amirbeigiarab, S. *et al.* Invariable stoichiometry of ribosomal proteins in mouse brain tissues with aging. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 22567–22572 (2019).
44. Soto, I. *et al.* Balanced mitochondrial and cytosolic translatomes underlie the biogenesis of human respiratory complexes. *Genome Biol.* **23**, 170 (2022).
45. Natan, E. *et al.* Cotranslational protein assembly imposes evolutionary constraints on homomeric proteins. *Nat. Struct. Mol. Biol.* **25**, 279–288 (2018).
46. Li, G.-W., Oh, E. & Weissman, J. S. The anti-Shine-Dalgarno sequence drives translational pausing and codon choice in bacteria. *Nature* **484**, 538–541 (2012).
47. Seidel, M. *et al.* Co-translational assembly orchestrates competing biogenesis pathways. *Nat. Commun.* **13**, 1224 (2022).
48. Bertolini, M. *et al.* Interactions between nascent proteins translated by adjacent ribosomes drive homomer assembly. *Science* **371**, 57–64 (2021).
49. van den Boogaart, K. G., Filzmoser, P., Hron, K., Templ, M. & Tolosana-Delgado, R. Classical and Robust Regression Analysis with Compositional Data. *Math. Geosci.* **53**, 823–858 (2021).
50. Quinn, T. P., Richardson, M. F., Lovell, D. & Crowley, T. M. propr: An R-package for Identifying Proportionally Abundant Features Using Compositional Data Analysis. *Sci. Rep.* **7**, 16252 (2017).
51. Ozadam, H., Geng, M. & Cenik, C. RiboFlow, RiboR and RiboPy: an ecosystem for analyzing ribosome profiling data at read length resolution. *Bioinformatics* **36**, 2929–2931 (2020).
52. Gerashchenko, M. V. & Gladyshev, V. N. Ribonuclease selection for ribosome profiling. *Nucleic Acids Res.* **45**, e6 (2017).
53. Mohammad, F., Green, R. & Buskirk, A. R. A systematically-revised ribosome profiling method for bacteria reveals pauses at single-codon resolution. *Elife* **8**, (2019).
54. Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. S. & Weissman, J. S. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* **324**, 218–223 (2009).
55. Larsson, O., Sonenberg, N. & Nadon, R. Identification of differential translation in genome wide studies. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 21487–21492 (2010).
56. Quinn, T. P. *et al.* A field guide for the compositional analysis of any-omics data. *Gigascience* **8**, (2019).
57. Sudmant, P. H., Alexis, M. S. & Burge, C. B. Meta-analysis of RNA-seq expression data across species, tissues and studies. *Genome Biol.* **16**, 287 (2015).
58. Wang, Z.-Y. *et al.* Transcriptome and translatome co-evolution in mammals. *Nature* **588**, 642–647 (2020).
59. Lu, P., Takai, K., Weaver, V. M. & Werb, Z. Extracellular matrix degradation and remodeling in development and disease. *Cold Spring Harb. Perspect. Biol.* **3**, (2011).
60. Artieri, C. G. & Fraser, H. B. Evolution at two levels of gene expression in yeast. *Genome Res.* **24**, 411–421 (2014).
61. McManus, C. J., May, G. E., Spealman, P. & Shteyman, A. Ribosome profiling reveals post-

transcriptional buffering of divergent gene expression in yeast. *Genome Res.* **24**, 422–430 (2014).
62. Breschi, A., Gingeras, T. R. & Guigó, R. Comparative transcriptomics in human and mouse. *Nat. Rev. Genet.* **18**, 425–440 (2017).
63. Crow, M., Suresh, H., Lee, J. & Gillis, J. Coexpression reveals conserved gene programs that co-vary with cell type across kingdoms. *Nucleic Acids Res.* **50**, 4302–4314 (2022).
64. Pierson, E. *et al.* Sharing and Specificity of Co-expression Networks across 35 Human Tissues. *PLoS Comput. Biol.* **11**, e1004220 (2015).
65. Kershaw, C. J. *et al.* Translation factor and RNA binding protein mRNA interactomes support broader RNA regulons for posttranscriptional control. *J. Biol. Chem.* **299**, 105195 (2023).
66. Hentze, M. W., Castello, A., Schwarzl, T. & Preiss, T. A brave new world of RNA-binding proteins. *Nat. Rev. Mol. Cell Biol.* **19**, 327–341 (2018).
67. Korbel, J. O., Jensen, L. J., von Mering, C. & Bork, P. Analysis of genomic context: prediction of functional associations from conserved bidirectionally transcribed gene pairs. *Nat. Biotechnol.* **22**, 911–917 (2004).
68. Szklarczyk, R. *et al.* WeGET: predicting new genes for molecular systems by weighted co-expression. *Nucleic Acids Res.* **44**, D567–73 (2016).
69. Zhang, M. *et al.* RNA-binding protein IMP3 is a novel regulator of MEK1/ERK signaling pathway in the progression of colorectal Cancer through the stabilization of MEKK1 mRNA. *J. Exp. Clin. Cancer Res.* **40**, 200 (2021).
70. Cargnello, M. & Roux, P. P. Activation and function of the MAPKs and their substrates, the MAPK-activated protein kinases. *Microbiol. Mol. Biol. Rev.* **75**, 50–83 (2011).
71. Bodén, M. & Bailey, T. L. Associating transcription factor-binding site motifs with target GO terms and target genes. *Nucleic Acids Res.* **36**, 4108–4117 (2008).
72. Machanick, P. & Bailey, T. L. MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics* **27**, 1696–1697 (2011).
73. Mecham, R. *The Extracellular Matrix: An Overview*. (Springer Science & Business Media, 2011).
74. Kagan, H. M. & Li, W. Lysyl oxidase: properties, specificity, and biological roles inside and outside of the cell. *J. Cell. Biochem.* **88**, 660–672 (2003).
75. Kikuchi, A. *et al.* Structural basis for activation of DNMT1. *Nat. Commun.* **13**, 7130 (2022).
76. Wu, Y.-Y. *et al.* The hTERT-p50 homodimer inhibits PLEKHA7 expression to promote gastric cancer invasion and metastasis. *Oncogene* **42**, 1144–1156 (2023).
77. Kurita, S., Yamada, T., Rikitsu, E., Ikeda, W. & Takai, Y. Binding between the junctional proteins afadin and PLEKHA7 and implication in the formation of adherens junction in epithelial cells. *J. Biol. Chem.* **288**, 29356–29368 (2013).
78. Pulimeno, P., Paschoud, S. & Citi, S. A role for ZO-1 and PLEKHA7 in recruiting paracingulin to tight and adherens junctions of epithelial cells. *J. Biol. Chem.* **286**, 16743–16750 (2011).
79. Jeung, H.-C. *et al.* PLEKHA7 signaling is necessary for the growth of mutant KRAS driven colorectal cancer. *Exp. Cell Res.* **409**, 112930 (2021).
80. Tavano, S. *et al.* Insm1 Induces Neural Progenitor Delamination in Developing Neocortex via Downregulation of the Adherens Junction Belt-Specific Protein Plekha7. *Neuron* **97**, 1299–1314.e8 (2018).

81. Sukonina, V. *et al.* FOXK1 and FOXK2 regulate aerobic glycolysis. *Nature* **566**, 279–283 (2019).
82. Kobe, B. & Kajava, A. V. The leucine-rich repeat as a protein recognition motif. *Curr. Opin. Struct. Biol.* **11**, 725–732 (2001).
83. Evans, R. *et al.* Protein complex prediction with AlphaFold-Multimer. *bioRxiv* 2021.10.04.463034 (2021) doi:10.1101/2021.10.04.463034.
84. Carlsson, P. & Mahlapuu, M. Forkhead transcription factors: key players in development and metabolism. *Dev. Biol.* **250**, 1–23 (2002).
85. The Human Transcription Factors. http://humantfs.ccbr.utoronto.ca/cite.php.
86. Szklarczyk, D. *et al.* STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* **43**, D447–52 (2015).
87. Shiber, A. *et al.* Cotranslational assembly of protein complexes in eukaryotes revealed by ribosome profiling. *Nature* **561**, 268–272 (2018).
88. Liesecke, F. *et al.* Ranking genome-wide correlation measurements improves microarray and RNA-seq based global and targeted co-expression networks. *Sci. Rep.* **8**, 10885 (2018).
89. Ewing, R. M. *et al.* Large-scale mapping of human protein–protein interactions by mass spectrometry. *Mol. Syst. Biol.* **3**, 89 (2007).
90. Drew, K., Wallingford, J. B. & Marcotte, E. M. hu.MAP 2.0: integration of over 15,000 proteomic experiments builds a global compendium of human multiprotein assemblies. *Mol. Syst. Biol.* **17**, e10016 (2021).
91. Heider, M. R. *et al.* Subunit connectivity, assembly determinants and architecture of the yeast exocyst complex. *Nat. Struct. Mol. Biol.* **23**, 59–66 (2016).
92. Kee, Y. *et al.* Subunit structure of the mammalian exocyst complex. *Proc. Natl. Acad. Sci. U. S. A.* **94**, 14438–14443 (1997).
93. Lalanne, J.-B. *et al.* Evolutionary Convergence of Pathway-Specific Enzyme Expression Stoichiometry. *Cell* **173**, 749–761.e38 (2018).
94. Cenik, C. *et al.* Integrative analysis of RNA, translation, and protein levels reveals distinct regulatory variation across humans. *Genome Res.* **25**, 1610–1621 (2015).
95. Bicknell, A. A. *et al.* Attenuating ribosome load improves protein output from mRNA by limiting translation-dependent mRNA decay. *Cell Rep.* **43**, 114098 (2024).
96. Liu, T.-Y. *et al.* Time-Resolved Proteomics Extends Ribosome Profiling-Based Measurements of Protein Synthesis Dynamics. *Cell Syst* **4**, 636–644.e9 (2017).
97. Piepoli, A. *et al.* The expression of leucine-rich repeat gene family members in colorectal cancer. *Exp. Biol. Med.* **237**, 1123–1128 (2012).
98. Liu, Y. *et al.* Identification of differential expression of genes in hepatocellular carcinoma by suppression subtractive hybridization combined cDNA microarray. *Oncol. Rep.* **18**, 943–951 (2007).
99. Chen, H. *et al.* miR-218 contributes to drug resistance in multiple myeloma via targeting LRRC28. *J. Cell. Biochem.* **122**, 305–314 (2021).
100. Wang, M., Herrmann, C. J., Simonovic, M., Szklarczyk, D. & von Mering, C. Version 4.0 of PaxDb: Protein abundance data, integrated across model organisms, tissues, and cell-lines. *Proteomics* **15**, 3163–3168 (2015).
101. Kulevich, S. E., Frey, B. L., Kreitinger, G. & Smith, L. M. Alkylating tryptic peptides to enhance electrospray ionization mass spectrometry analysis. *Anal. Chem.* **82**, 10135–10142

(2010).
102. Rodriguez, J. M. *et al.* APPRIS: annotation of principal and alternative splice isoforms. *Nucleic Acids Res.* **41**, D110–7 (2013).
103. *Sra-Tools: SRA Tools*. (Github).
104. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10–12 (2011).
105. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
106. Creators Liu, O. *HeLa Ribosome Profiling Data*. doi:10.5281/zenodo.10594392.
107. Gerashchenko, M. V. & Gladyshev, V. N. Translation inhibitors cause abnormalities in ribosome profiling experiments. *Nucleic Acids Res.* **42**, e134 (2014).
108. Wu, C. C.-C., Zinshteyn, B., Wehner, K. A. & Green, R. High-Resolution Ribosome Profiling Defines Discrete Ribosome Elongation States and Translational Regulation during Cellular Stress. *Mol. Cell* **73**, 959–970.e5 (2019).
109. Wolin, S. L. & Walter, P. Ribosome pausing and stacking during translation of a eukaryotic mRNA. *EMBO J.* **7**, 3559–3569 (1988).
110. Sharma, J. *et al.* A small molecule that induces translational readthrough of CFTR nonsense mutations by eRF1 depletion. *Nat. Commun.* **12**, 4358 (2021).
111. Tukey, J. W. The Future of Data Analysis. *Ann. Math. Stat.* **33**, 1–67 (1962).
112. Zhang, X.-O., Yin, Q.-F., Chen, L.-L. & Yang, L. Gene expression profiling of non-polyadenylated RNA-seq across species. *Genom Data* **2**, 237–241 (2014).
113. Yang, L., Duff, M. O., Graveley, B. R., Carmichael, G. G. & Chen, L.-L. Genomewide characterization of non-polyadenylated RNAs. *Genome Biol.* **12**, R16 (2011).
114. van den Boogaart, K. G. & Tolosana-Delgado, R. *Analyzing Compositional Data with R*. (Springer Berlin Heidelberg).
115. orthogene. *Bioconductor* https://bioconductor.org/packages/release/bioc/html/orthogene.html.
116. van den Boogaart, K. G. & Tolosana-Delgado, R. 'compositions': A unified R package to analyze compositional data. *Comput. Geosci.* **34**, 320–338 (2008).
117. Kim, S. ppcor: An R Package for a Fast Calculation to Semi-partial Correlation Coefficients. *Commun Stat Appl Methods* **22**, 665–674 (2015).
118. Berriz, G. F., Beaver, J. E., Cenik, C., Tasan, M. & Roth, F. P. Next generation software for functional trend analysis. *Bioinformatics* **25**, 3043–3044 (2009).
119. Buttrey, S. & Whitaker, L. TreeClust: An R package for tree-based clustering dissimilarities. *R J.* **7**, 227 (2015).
120. Wainberg, M. *et al.* A genome-wide atlas of co-essential modules assigns function to uncharacterized genes. *Nat. Genet.* **53**, 638–649 (2021).
121. Gene Ontology Consortium *et al.* The Gene Ontology knowledgebase in 2023. *Genetics* **224**, (2023).
122. Philippe, L., van den Elzen, A. M. G., Watson, M. J. & Thoreen, C. C. Global analysis of LARP1 translation targets reveals tunable and dynamic features of 5′ TOP motifs. *Proceedings of the National Academy of Sciences* **117**, 5319–5328 (2020).
123. Ballouz, S., Weber, M., Pavlidis, P. & Gillis, J. EGAD: ultra-fast functional analysis of gene networks. *Bioinformatics* **33**, 612–614 (2017).

124. Carlson, M. org. Mm. eg. db: Genome wide annotation for Mouse. R package version 3.8. 2. 2019.
125. Carlson, M. org. Hs. eg. db: Genome wide annotation for Human. R package version 3.8. 2. 2019.
126. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
127. Lambert, S. A. *et al.* The Human Transcription Factors. *Cell* **172**, 650–665 (2018).
128. UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res.* **51**, D523–D531 (2023).
129. Hou, Y., Xie, T., He, L., Tao, L. & Huang, J. Topological links in predicted protein complex structures reveal limitations of AlphaFold. *Commun Biol* **6**, 1098 (2023).
130. Burke, D. F. *et al.* Towards a structurally resolved human protein interaction network. *Nat. Struct. Mol. Biol.* **30**, 216–225 (2023).
131. Bryant, P., Pozzati, G. & Elofsson, A. Improved prediction of protein-protein interactions using AlphaFold2. *Nat. Commun.* **13**, 1265 (2022).
132. Hu, Y. *et al.* Paralog Explorer: A resource for mining information about paralogs in common research organisms. *Comput. Struct. Biotechnol. J.* **20**, 6570–6577 (2022).
133. Krismer, K. *et al.* Transite: A Computational Motif-Based Analysis Platform That Identifies RNA-Binding Proteins Modulating Changes in Gene Expression. *Cell Rep.* **32**, 108064 (2020).
134. Benoit Bouvrette, L. P., Bovaird, S., Blanchette, M. & Lécuyer, E. oRNAment: a database of putative RNA binding protein target sites in the transcriptomes of model species. *Nucleic Acids Res.* **48**, D166–D173 (2020).
135. Van Nostrand, E. L. *et al.* A large-scale binding and functional map of human RNA-binding proteins. *Nature* **583**, 711–719 (2020).
136. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
137. Stuart, E. A., King, G., Imai, K. & Ho, D. MatchIt: nonparametric preprocessing for parametric causal inference. *J. Stat. Softw.* (2011).
138. Sanson, K. R. *et al.* Optimized libraries for CRISPR-Cas9 genetic screens with multiple modalities. *Nat. Commun.* **9**, 5416 (2018).
139. Sanjana, N. E., Shalem, O. & Zhang, F. Improved vectors and genome-wide libraries for CRISPR screening. *Nat. Methods* **11**, 783–784 (2014).
140. ChatGPT. https://chat.openai.com.
141. Rossi, G. P. *et al.* Endothelin-1 stimulates steroid secretion of human adrenocortical cells ex vivo via both ETA and ETB receptor subtypes. *J. Clin. Endocrinol. Metab.* **82**, 3445–3449 (1997).
142. Sánchez-Caballero, L. *et al.* TMEM70 functions in the assembly of complexes I and V. *Biochim. Biophys. Acta Bioenerg.* **1861**, 148202 (2020).
143. Carroll, J., He, J., Ding, S., Fearnley, I. M. & Walker, J. E. TMEM70 and TMEM242 help to assemble the rotor ring of human ATP synthase and interact with assembly factors for complex I. *Proc. Natl. Acad. Sci. U. S. A.* **118**, (2021).
144. Mii, Y. & Takada, S. Heparan Sulfate Proteoglycan Clustering in Wnt Signaling and Dispersal. *Front Cell Dev Biol* **8**, 631 (2020).

145. Kamimura, K. *et al.* Perlecan regulates bidirectional Wnt signaling at the Drosophila neuromuscular junction. *J. Cell Biol.* **200**, 219–233 (2013).
146. Camlin, N. J., McLaughlin, E. A. & Holt, J. E. Kif4 Is Essential for Mouse Oocyte Meiosis. *PLoS One* **12**, e0170650 (2017).
147. Tang, F. *et al.* Involvement of Kif4a in Spindle Formation and Chromosome Segregation in Mouse Oocytes. *Aging Dis.* **9**, 623–633 (2018).
148. Robertson, A. K., Geiman, T. M., Sankpal, U. T., Hager, G. L. & Robertson, K. D. Effects of chromatin structure on the enzymatic and DNA binding functions of DNA methyltransferases DNMT1 and Dnmt3a in vitro. *Biochem. Biophys. Res. Commun.* **322**, 110–118 (2004).
149. Schrader, A., Gross, T., Thalhammer, V. & Längst, G. Characterization of Dnmt1 Binding and DNA Methylation on Nucleosomes and Nucleosomal Arrays. *PLoS One* **10**, e0140076 (2015).
150. Ciossani, G. *et al.* The kinetochore proteins CENP-E and CENP-F directly and specifically interact with distinct BUB mitotic checkpoint Ser/Thr kinases. *J. Biol. Chem.* **293**, 10084–10101 (2018).
151. Liao, H., Winkfein, R. J., Mack, G., Rattner, J. B. & Yen, T. J. CENP-F is a protein of the nuclear matrix that assembles onto kinetochores at late G2 and is rapidly degraded after mitosis. *J. Cell Biol.* **130**, 507–518 (1995).
152. Chen, M. *et al.* FAT1 inhibits the proliferation and metastasis of cervical cancer cells by binding β-catenin. *Int. J. Clin. Exp. Pathol.* **12**, 3807–3818 (2019).
153. Nishikawa, Y. *et al.* Human FAT1 cadherin controls cell migration and invasion of oral squamous cell carcinoma through the localization of β-catenin. *Oncol. Rep.* **26**, 587–592 (2011).
154. Morris, L. G. T. *et al.* Recurrent somatic mutation of FAT1 in multiple human cancers leads to aberrant Wnt activation. *Nat. Genet.* **45**, 253–261 (2013).
155. Hou, R., Liu, L., Anees, S., Hiroyasu, S. & Sibinga, N. E. S. The Fat1 cadherin integrates vascular smooth muscle cell growth and migration signals. *J. Cell Biol.* **173**, 417–429 (2006).
156. Vallet, S. D., Berthollier, C., Salza, R., Muller, L. & Ricard-Blum, S. The Interactome of Cancer-Related Lysyl Oxidase and Lysyl Oxidase-Like Proteins. *Cancers* **13**, (2020).
157. Vallet, S. D. *et al.* Insights into the structure and dynamics of lysyl oxidase propeptide, a flexible protein with numerous partners. *Sci. Rep.* **8**, 11768 (2018).
158. Yang, C. *et al.* Transcriptomic Analysis Identified ARHGAP Family as a Novel Biomarker Associated With Tumor-Promoting Immune Infiltration and Nanomechanical Characteristics in Bladder Cancer. *Front Cell Dev Biol* **9**, 657219 (2021).
159. Lamarche-Vane, N. & Hall, A. CdGAP, a novel proline-rich GTPase-activating protein for Cdc42 and Rac. *J. Biol. Chem.* **273**, 29172–29177 (1998).
160. Yang, S. *et al.* Control of antiviral innate immune response by protein geranylgeranylation. *Sci Adv* **5**, eaav7999 (2019).
161. Bouhaddou, M. *et al.* The Global Phosphorylation Landscape of SARS-CoV-2 Infection. *Cell* **182**, 685–712.e19 (2020).
162. Swaine, T. & Dittmar, M. T. CDC42 Use in Viral Cell Entry Processes by RNA Viruses. *Viruses* **7**, 6526–6536 (2015).
163. Redmond, S. A. *et al.* Somatodendritic Expression of JAM2 Inhibits Oligodendrocyte Myelination. *Neuron* **91**, 824–836 (2016).

164. Song, K. Y., Choi, H. S., Law, P.-Y., Wei, L.-N. & Loh, H. H. Vimentin interacts with the 5'-untranslated region of mouse mu opioid receptor (MOR) and is required for post-transcriptional regulation. *RNA Biol.* **10**, 256–266 (2013).
165. van der Brug, M. P. *et al.* RNA binding activity of the recessive parkinsonism protein DJ-1 supports involvement in multiple cellular pathways. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 10244–10249 (2008).
166. Niere, F. *et al.* Aberrant DJ-1 expression underlies L-type calcium channel hypoactivity in dendrites in tuberous sclerosis complex and Alzheimer's disease. *Proc. Natl. Acad. Sci. U. S. A.* **120**, e2301534120 (2023).
167. Jin, W. *et al.* HydRA: Deep-learning models for predicting RNA-binding capacity from protein interaction association context and protein sequence. *Mol. Cell* **83**, 2595–2611.e11 (2023).

Table 1: Literature support for gene functions predicted using TEC.

In the table, we list the predictions that are supported by literature. To do the new gene function prediction, we selected GO terms with AUROC measured with TEC >=0.8, then focused on the subset with differences between AUROC measured with TEC and RNA co-expression >= 0.1.

Term	Species	Description	New adding gene (top ranking in TE)	TEC AUROC	RNA co-expression AUROC	Ranking of new adding gene in RNA	Reference
GO:0005496	Human	steroid binding	EDNRA	0.81	0.50	3991	141
GO:0022900	Human	electron transport chain	TMEM70	0.82	0.67	1611	142,143
GO:0042813	Human	Wnt-activated receptor activity	HSPG2	0.85	0.74	341	144,145
GO:0007129	Human	homologous chromosome pairing at meiosis	KIF4A	0.84	0.73	11	146,147
GO:0031492	Human	nucleosomal DNA binding	DNMT1	0.85	0.73	251	75,148,149
GO:0050793	Mouse	regulation of developmental	Plekha7	0.85	0.70	3929	76–80

Term	Species	Description	New adding gene (top ranking in TE)	TEC AUROC	RNA co-expression AUROC	Ranking of new adding gene in RNA	Reference
		process					
GO:1990023	Mouse	mitotic spindle midzone	Cenpf	0.90	0.74	54	150,151
GO:0016342	Mouse	catenin complex	Fat1	0.85	0.73	781	152–155
GO:0005539	Mouse	glycosaminoglycan binding	Lox	0.95	0.83	32	156,157
GO:0140374	Mouse	antiviral innate immune response	Arhgap31	0.84	0.73	1085	158–162
GO:0022010	Mouse	Central nervous system myelination	Jam2	0.88	0.76	584	163

1389