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

Yue Liu[a], Shilpa Rao[a], Ian Hoskins[a], Michael Geng[a], Qiuxia Zhao[a], Jonathan Chacko[a], Vighnesh Ghatpande[a], Kangsheng Qi[a], Logan Persyn[a], Jun Wang[b], Dinghai Zheng[b], Yochen Zhong[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

Characterizing 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. 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 candidate mechanisms driving shared patterns of translation regulation. TEC is conserved across human and mouse cells and uncovers gene functions that are not evident from RNA or protein co-expression. Moreover, our observations indicate that proteins that physically interact are highly enriched for positive covariation at both translational and transcriptional levels. Our findings establish TEC as a conserved organizing principle of mammalian transcriptomes.

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

INTRODUCTION

Over the past three decades, technological advances have progressively revealed the expression of RNAs with increasing spatial and cellular resolution[1–7]. These measurements have driven conceptual advances including the widespread use of RNA co-expression analysis, which quantifies the similarity of gene expression patterns across biological conditions[8–11]. Such analyses have proved informative for inferring gene functions[9,12,13], predicting protein-protein interactions[14,15], and identifying shared regulatory mechanisms via common transcription factor binding sites[16,17].

These findings suggest that RNA co-expression may serve as a proxy for the proteomic organization of cells. Quantification of protein abundance across numerous cell types and conditions has recently allowed this assumption to be explicitly tested. Surprisingly, coordinated patterns of protein abundance patterns are frequently not detected at the RNA level[18,19], and physically interacting proteins are much more likely to have coordinated protein abundance than RNA co-expression[18–20]. These discrepancies suggest that post-transcriptional regulation likely plays a significant role in proteome organization.

Translation regulation may bridge this gap, yet whether translational efficiency is coordinated among functionally related genes across biological contexts remains an open question. There are three lines of evidence that suggest its plausibility. First, mammalian mRNAs bind various proteins to form ribonucleoproteins that influence their lifecycle from export to translation[21]. 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, positing that functionally related mRNAs are regulated together post-transcriptionally[22,23]. Second, in *E. coli* and yeast, proteins within multiprotein complexes are synthesized in stoichiometric ratios[24,25]. However, in humans, evidence of such proportional synthesis is reported for only two complexes: ribosomes[25,26] and the oxidative phosphorylation machinery[27] in a limited number of cell lines. Third, the formation of many protein complexes is facilitated by the co-translational folding of nascent peptides[24,25,28,29,30].

Despite these precedents, a general framework for detecting translation-level coordination across cell types has been lacking. Here, we analyzed thousands of matched ribosome profiling and RNA-seq datasets from more than 140 human and mouse cell lines and tissues. To quantify the similarity in TE patterns across cell types, we introduce the concept of *Translation Efficiency Covariation* (TEC). Analogous to RNA co-expression, TEC is common among functionally related genes. We demonstrate that TEC nominates novel gene functions, is enriched among physically interacting proteins, and is conserved across species. Together, these findings suggest TEC is a functionally relevant and evolutionarily conserved organizing principle of mammalian gene expression.

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[31]. 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 (Supplementary Fig. 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 (Supplementary Fig. 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[32]. The choice of nuclease impacts the resulting read length distribution of RPFs[33] (Supplementary Fig. 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 (Supplementary Fig. 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 (Supplementary Fig. 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 (Supplementary Fig. 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; Supplementary Fig. 5; Methods). These findings underscore the necessity of meticulous quality control for the selection of experiments to enable large-scale data analyses.

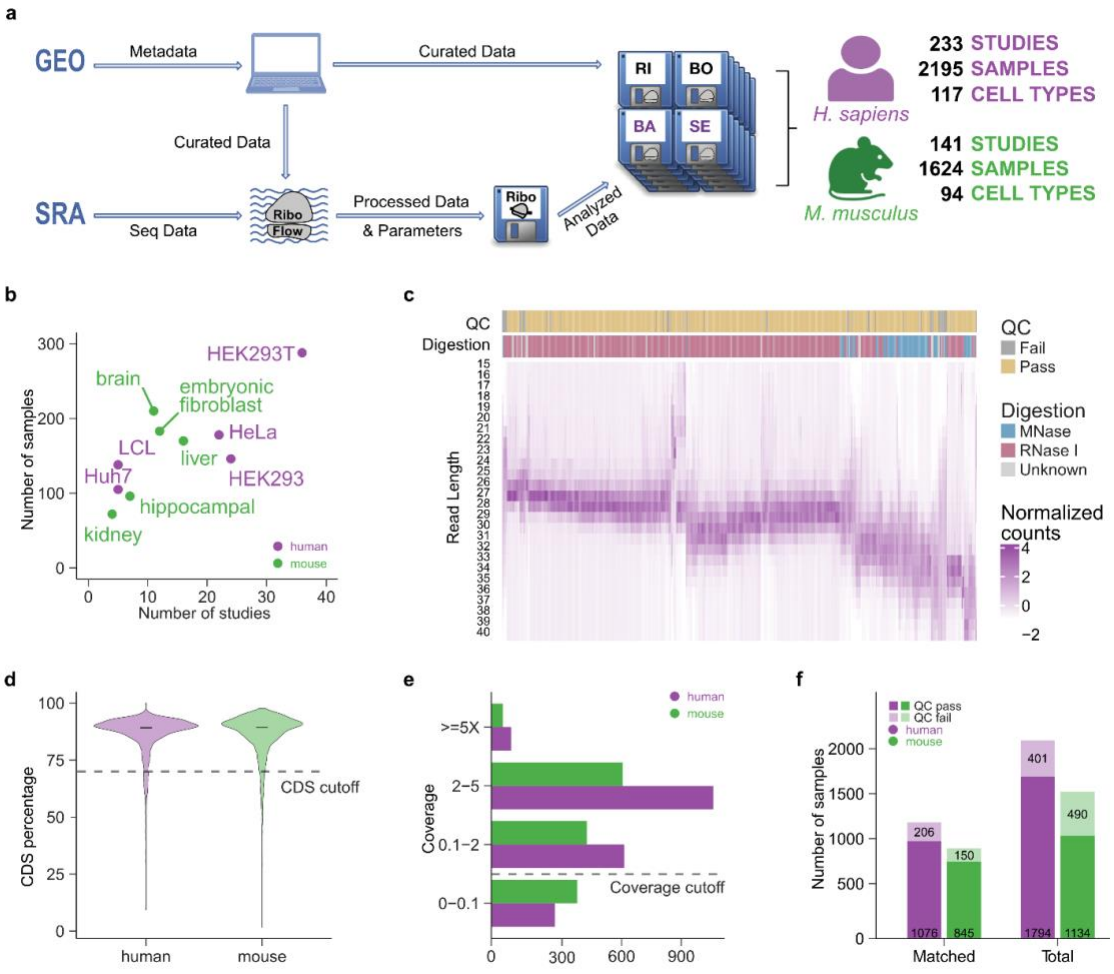


Fig. 1 | RiboBase: a comprehensive ribosome profiling database with thousands of experiments. a, Schematic of RiboBase. We manually curated metadata and processed the sequencing reads using a uniform pipeline (RiboFlow[31]). **b,** Top five most highly represented cell lines or tissues with respect to 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,**

128 For the 3,819 ribosome profiling experiments in RiboBase, we applied a function to select the
129 range of RPFs for further analysis (Methods). We calculated the proportion of the selected RPFs
130 that map to the coding regions (y-axis). The horizontal line represents the median of the
131 distribution. **e,** Experiments (x-axis) were grouped by the transcript coverage (y-axis). **f,** Among
132 the ribosome profiling experiments in RiboBase, 2,277 of them had corresponding RNA-seq data
133 (matched). The number of samples that pass quality controls was plotted.

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

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

146 Using the set of matched ribosome profiling and RNA-seq experiments, we next quantified TE,
147 which is typically defined as the ratio of ribosome footprints to RNA-seq reads, normalized as
148 counts per million[34]. However, this approach leads to biased estimates with significant
149 drawbacks[35]. To address this limitation, we calculate TE based on a regression model using a
150 compositional data analysis method[36–38], avoiding the mathematical shortcomings of using the
151 canonical log-ratio method (Fig. 2b; Supplementary Fig. 6a-c, 7; table S8-11; Methods;
152 supplementary text).

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

162 We expected that an accurate estimate of TE would correlate significantly with protein abundance.
163 We calculated for each transcript the cell type-specific TE by taking the average of TE values
164 across all experiments conducted with that particular cell line and found that TE derived using the
165 regression approach with winsorized read counts (Supplementary Fig. 8b; Supplementary Fig. 9-
166 11; supplementary text) was significantly correlated with protein abundance in seven cancer cell
167 lines (mean Spearman correlation coefficient of 0.465; Fig. 2e).

168 Furthermore, TE measurements from cell lines and tissues with the same biological origin (e.g.,
169 blood) tended to cluster together, supporting the existence of cell-type-specific in addition to
170 species-specific differences in TE (Fig. 2f). As expected, mean ribosome occupancy and RNA
171 expression across cell types showed a strong correlation (Spearman correlation: ~0.8), yet mean
172 TE was only weakly associated with RNA expression (Spearman correlation: ~0.2) (Fig. 2g).

173 Measurements of TE in two species across a large number of cell types enabled us to investigate
174 the conservation of TE, ribosome occupancy, and RNA expression. Transcriptomes, ribosome
175 occupancy, and proteomes exhibit a high degree of conservation across diverse organisms[39,40].
176 Consistently, we found average ribosome occupancy, RNA expression, and TE across different
177 cell lines and tissues were highly similar between orthologous genes in human and mouse (Fig. 2g;
178 table S12). Specifically, the Spearman correlation coefficient of mean TE across cell types and
179 tissues between human and mouse was 0.9 (Fig. 2h), which is comparable to the mean RNA
180 expression correlation between human and mouse (~0.86, Supplementary Fig. 12a). Using a 95%
181 prediction interval to identify outlier genes, we found that outlier genes with higher mean TE in
182 humans compared to mice were enriched in the gene ontology term 'RNA binding function' (Fig.
183 2i). In contrast, genes with elevated mean TE in mice were enriched for having functions related
184 to extracellular matrix and collagen-containing components (Fig. 2i). The enrichment of genes
185 with higher TE in mice, particularly those from the extracellular matrix and collagen-containing
186 components, may be due to the fact that many samples in mouse studies are derived from the early
187 developmental stage[41].

188 Despite the high correlation of mean TE across various cell lines and tissues between human and
189 mouse, TE distinctly exhibits cell-type specificity. While several studies compared the
190 conservation of TE between the same tissues of mammalians or model organisms[40,42,43], our dataset
191 uniquely enabled us to determine the conservation of variability of TE for transcripts across
192 different cell types. Intriguingly, we observed a moderately high similarity between the variability
193 of TE of orthologous genes in human and mouse (Spearman partial correlation coefficient = 0.63;
194 Fig. 2j; Supplementary Fig. 12b-d; Methods). Our results reveal that certain genes exhibit higher
195 variability of TE across cell types and this is a conserved property between human and mouse.

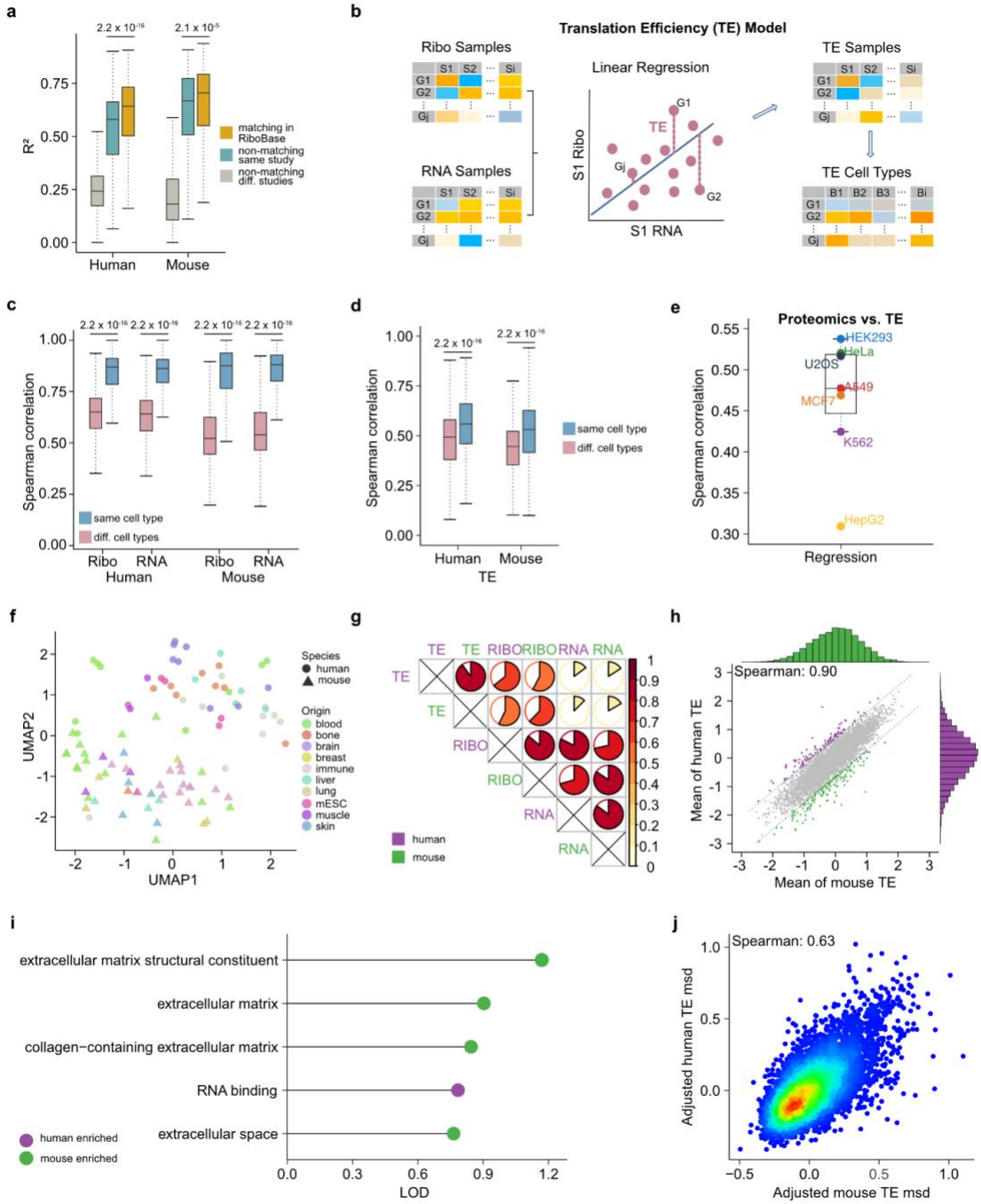


Fig. 2 | TE defined using a compositional linear regression model is conserved across cell types and species. a, The distribution of coefficient of determination (R^2, y-axis) between ribosome profiling data and RNA-seq in RiboBase was compared to random matching within the

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; Supplementary Fig. 7). **c,** Distribution of correlations of TE (linear regression model) across experiments. **d,** The distribution of Spearman correlations between experiments (y-axis) was calculated based on whether they originated from identical or different cell lines or tissues. **e,** Correlation between TE and protein abundance from seven human cell lines[44] was calculated using compositional regression method. **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; Supplementary Fig. 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)[37,38], 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 (Supplementary Fig. 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[39,45,46]. 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 (Supplementary Fig. 14, Pearson correlation coefficient 0.00022). Our findings imply that translation efficiency patterns are evolutionarily preserved, paralleling the conservation of RNA co-expression.

There are few known examples of transcript sets with shared translational control across cell types, including those with 5′ terminal oligopyrimidine (TOP) motifs[47] and transcripts regulated by

CSDE1[48]. We hypothesized that these sets of transcripts would exhibit significantly higher TEC compared to a control set, and this was indeed observed (Supplementary Fig. 15a; Wilcoxon p-value < 2.2e-16). These findings highlight that TEC effectively captures previously established patterns of shared translational regulation among transcripts.

RNA co-expression analyses led to the discovery of regulatory motifs and shared transcription factor binding sites[49]. We hypothesized that TEC among genes may similarly nominate RNA binding proteins (RBPs) as potential drivers of TEC[50]. We identified groups of transcripts whose TE is correlated with the RNA expression of RBPs[51] (1274 human and 1762 mouse RBPs; Methods; https://zenodo.org/uploads/11359114). For example, the RNA expression of ZC3H10 has largely positive correlations with TE of transcripts (71% of all significant correlations Fig. 3c). Conversely, the RNA expression of subunits of ubiquinone oxidoreductase (Ndufa7, Ndufv3) is negatively correlated with TE for many transcripts in mice (Fig. 3c).

We observed a correspondence between human and mouse in the proportion of transcripts whose TE is significantly correlated with an RBP's expression (Pearson correlation 0.44; Fig. 3d). However, knockout of three candidate RBPs did not lead to significant changes in TE (Supplementary Fig. 16), which may be due to several limitations including cell-type specificity, efficiency of knockout, and limited replicates (supplementary text). Taken together, our analyses nominate RBPs that may coordinate the TEC of evolutionarily conserved networks.

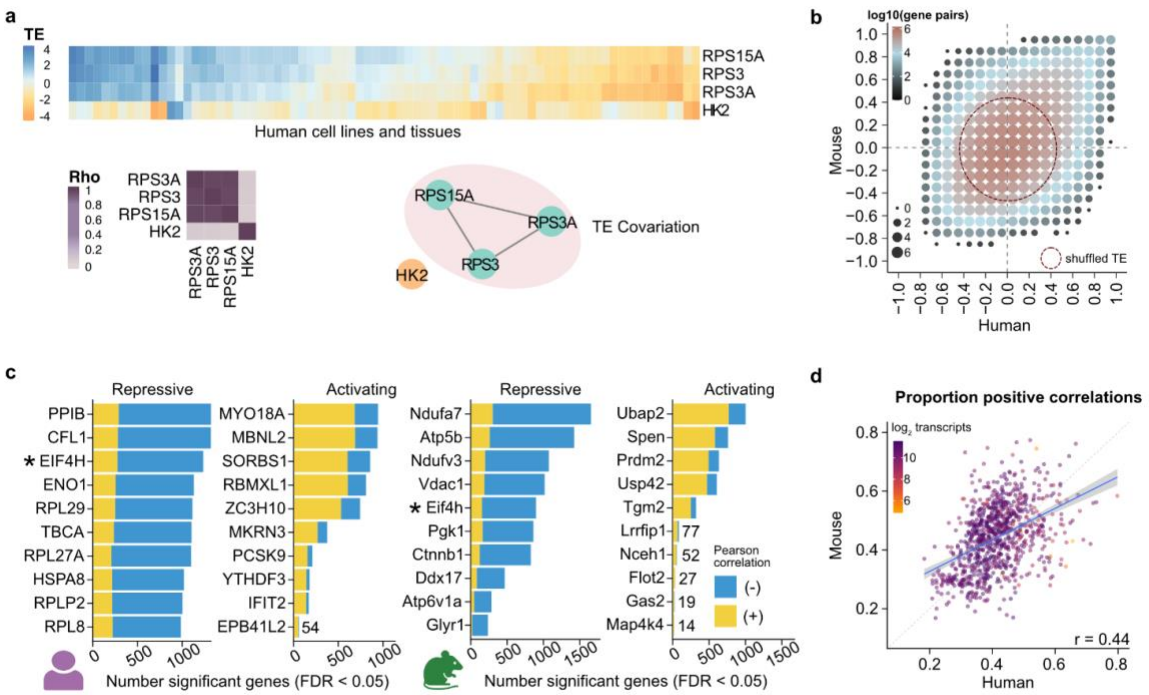


Fig. 3 | Translation efficiency covariation is conserved between human and mouse. a,

Example illustrating 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 size 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 (Supplementary Fig. 14). **c,** The top ten RBPs with the highest number of genes showing significant correlations between transcript TE and RBP RNA expression are displayed. An asterisk marks genes in the top ten in both species. **d,** For all RBPs (each dot), we plotted the proportion of positive correlations between TE of significant transcripts and the RNA expression of the RBP. Blue line is a linear fit with 95% confidence intervals in gray. Pearson correlation coefficient is shown.

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[11,52,53], 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; Supplementary Fig. 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[54]. 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[55,56]. Accordingly, we expected these genes to harbor shared sequence elements. 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; Supplementary Fig. 17b; Supplementary Fig. 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 (also see supplementary text for sequence features associated with TEC). In contrast, our analyses indicate miRNA is not a primary driver of TEC (Supplementary Fig. 19), in agreement with previous literature[57,58].

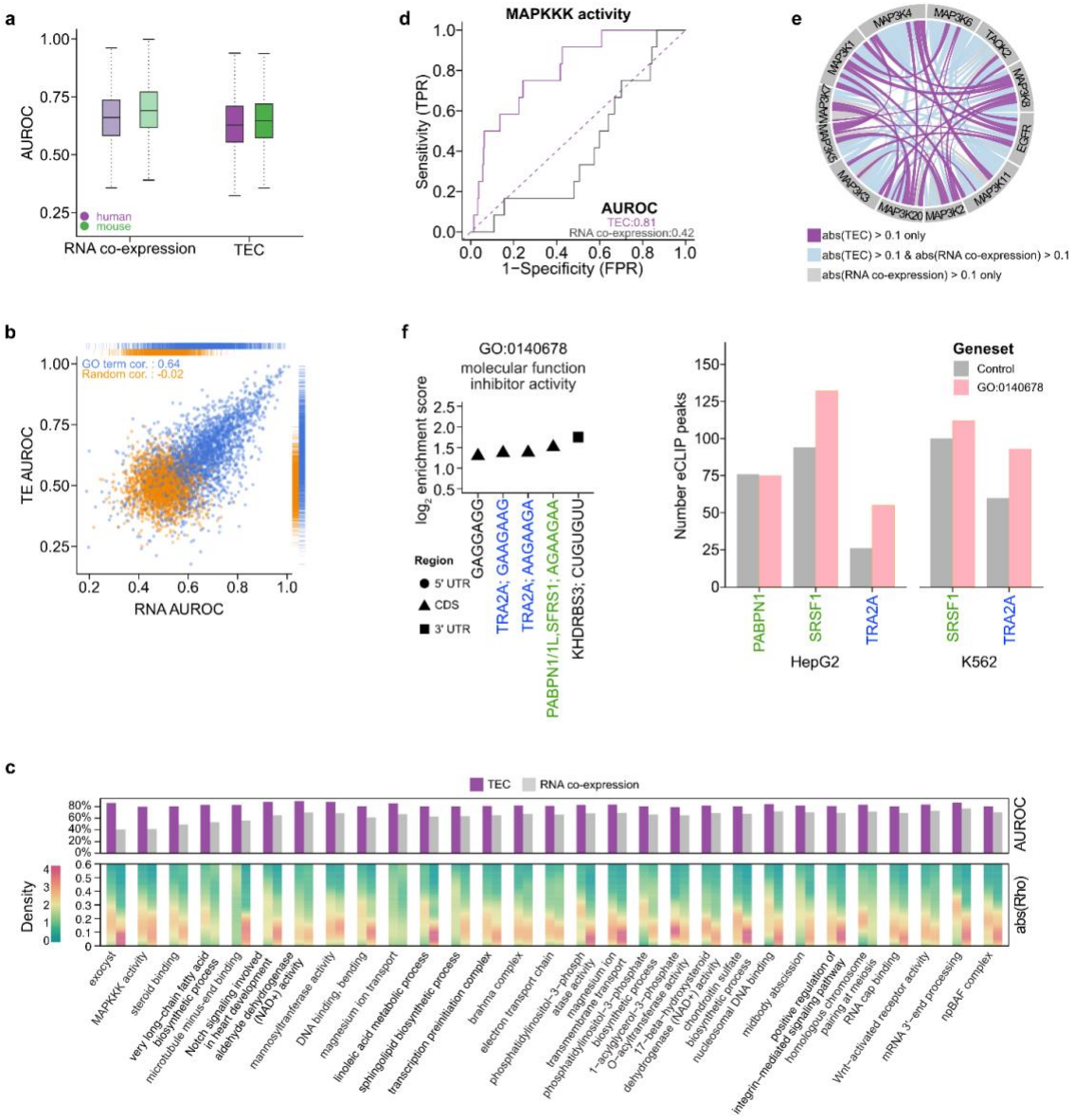


Fig. 4 | Genes associated with certain biological functions exhibit higher similarity patterns in TE than in RNA expression. a, We calculated the similarity of expression (quantified by AUROC; y-axis) among genes within 2,989 human and 3,340 mouse GO terms. In the box plot, the horizontal line corresponds 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

326 absolute rho scores for gene pairs (bottom; gene pairs with abs(rho) > 0.1). **d,** AUROC plot
327 calculated with genes associated with MAPKKK activity. **e,** In the circle plot, the connections
328 display absolute rho above 0.1 either at TE level alone (purple), at both RNA and TE levels (blue),
329 or RNA level alone (gray) for gene pairs involved in MAPKKK activity. **f,** Motif enrichment (left)
330 for the GO term 'molecular function inhibitor activity' (Supplementary Fig. 17b). RNA binding
331 proteins (RBPs) matching the motifs from oRNAment[59] or Transite[60] are indicated. Enhanced
332 cross-linking immunoprecipitation (eCLIP) data[61] indicates increased binding of TRA2A and
333 SRSF1 in the CDS of genes for this GO term compared to matched control genes with similar
334 sequence properties (Methods).

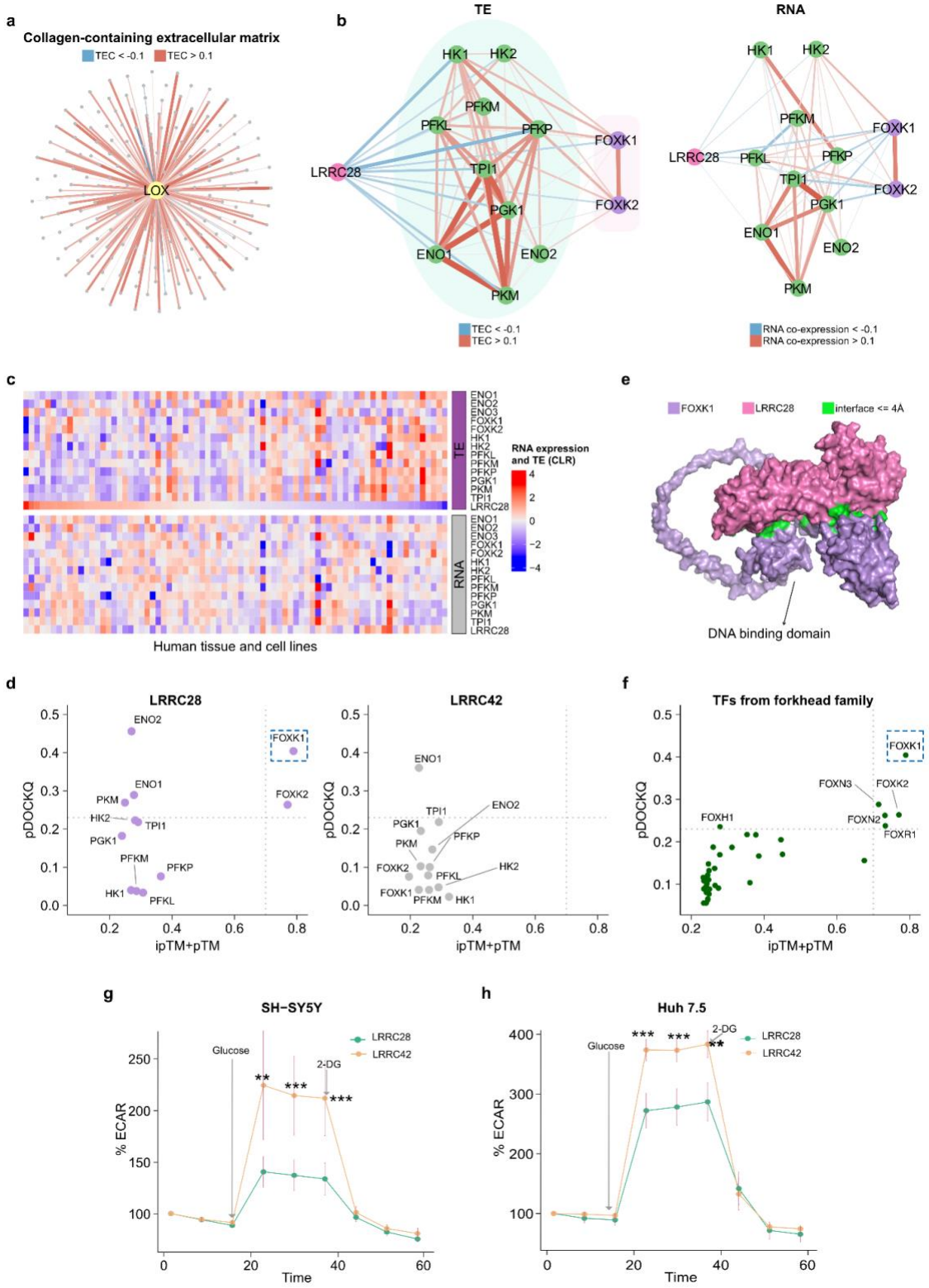
335 **TEC nominates gene functions**

336 We next investigated whether gene functions may be predicted by utilizing TEC, given the success
337 of RNA co-expression for this task[52,53]. The functional annotations of human genes are
338 continuously being updated, providing an opportunity to test this hypothesis using recently added
339 information to the knowledge base. Specifically, we used functional annotations from the GO
340 database from January 1, 2021, to determine functional groups that demonstrate strong TEC among
341 its members (AUROC > 0.8) and developed a framework to predict new functional associations
342 with these groups (Methods). By comparing our predictions to annotations from December 4, 2022,
343 we confirmed the predicted association of the *LOX* gene with the GO term 'collagen-containing
344 extracellular matrix'. LOX critically facilitates the formation, development, maturation, and
345 remodeling of the extracellular matrix by catalyzing the cross-linking of collagen fibers, thereby
346 enhancing the structural integrity and stability of tissues[62,63]. Our prediction successfully identified
347 this new addition, as *LOX* exhibits positive similarity in TE with the vast majority of genes in this
348 term (Fig. 5a).

349 Recognizing the capacity of TEC to elucidate biological functions, we utilized a recent version of
350 GO annotations (December 4, 2022) to systematically predict new associations for genes. To
351 underscore the unique insights gained from TEC, we focused on the 33 human and 31 mouse GO
352 terms that either exhibited significantly higher TEC than RNA co-expression (Table 1) or provided
353 new functional predictions that were only supported by TEC (the ranking of the newly predicted
354 gene with RNA co-expression fell beyond the top 50%, table S15-16; Methods). By focusing on
355 these GO terms, we aimed to identify similarity patterns based on TE, revealing functional
356 associations that would not be detected by RNA co-expression. We conducted a literature search
357 to determine if prior research supported these predictions, and found that 11 have already been
358 corroborated by previous publications, although they have not yet been reflected in the relevant
359 GO term annotations (Table 1; supplementary text). For example, cryo-electron microscopy
360 experiments demonstrated that human DNMT1 binds to hemimethylated DNA in conjunction with
361 ubiquitinated histone H3[64]. This binding facilitates the enzymatic activity of DNMT1 in
362 maintaining genomic DNA methylation. Our analysis revealed that DNMT1 was the highest
363 ranking prediction exhibiting strong TEC with genes associated with nucleosomal DNA binding
364 function. In mouse, we predicted Plekha7 to be a member of the regulation of developmental
365 processes. This prediction was recently validated by the observation of neural progenitor cell
366 delamination upon the disruption of Plekha7[65–69].

367 The high rate of validation of our predictions in the literature suggested that other predictions based
368 on TEC may reflect new and yet to be confirmed functions. In particular, we observed that the
369 human leucine-rich repeat-containing 28 (*LRRC28*) gene displays strong TEC with glycolytic
370 genes, but is not co-expressed at the RNA level (Fig. 5b-c, table S17). Specifically, *LRRC28*
371 displayed negatively correlated TE with key glycolytic genes including *HK1*, *HK2*, *PFKL*, *PFKM*,
372 *PFKP*, *TPI1*, *PGK1*, *ENO1*, *ENO2*, *PKM*, and two transcription factors *FOXK1* and *FOXK2* that
373 regulate glycolytic genes[70]. Given that the leucine-rich repeat domains typically facilitate protein-
374 protein interactions[71], LRRC28 may interact directly with one or more of the glycolytic proteins.
375 Using AlphaFold2-Multimer[72], we calculated the binding confidence scores between LRRC28 and
376 all glycolysis-associated proteins (Methods) and found that LRRC28 has a very high likelihood of
377 binding to FOXK1 (Fig. 5d-e).

378 FOXK1 is a member of the forkhead family of transcription factors that share a structurally similar
379 DNA-binding domain[73,74]. Interestingly, LRRC28 likely binds both the non-DNA-binding region
380 and DNA-binding domain of FOXK1 (distance < 4 angstroms; Fig. 5e; Supplementary Fig. 20a).
381 This observation led us to examine the specificity of the interaction between LRRC28 and FOXK1.
382 We calculated the binding probabilities of LRRC28 with 35 other forkhead family transcription
383 factors, finding that FOXK1 exhibits the strongest evidence of physical interaction with LRRC28
384 (Fig. 5f). This specificity is potentially due to a unique binding site between LRRC28 and
385 FOXK1's non-DNA-binding region (Fig. 5e). As an additional control, we selected LRRC42, a
386 protein with leucine-rich repeats that does not exhibit TEC with glycolytic genes. As expected,
387 LRRC42 showed a very low likelihood of interaction with any of the glycolytic genes, including
388 FOXK1 (Fig. 5d). These findings suggest that LRRC28 may serve as a regulator of glycolysis by
389 binding to FOXK1, thereby preventing FOXK1 from binding to the promoter regions of glycolytic
390 genes and leading to the downregulation of glycolysis. To experimentally test this prediction, we
391 stably overexpressed LRRC28 in four human cell lines and assessed the glycolytic capacity of the
392 cells by quantifying the extracellular acidification rate (ECAR). Stable overexpression of LRRC28
393 in Huh 7.5 resulted in slower growth therefore we resorted to transient overexpression strategy for
394 this particular cell line. Congruent to our expectation, ECAR was found to be significantly lower
395 in SH-SY5Y and Huh 7.5 cells overexpressing LRRC28 compared to cells overexpressing a
396 control gene; LRRC42 (Fig. 5g-h). However, HEK293T and MCF-7 showed non-significant
397 changes between the two conditions (Supplementary Fig. 20b-c). Interestingly, the varied response
398 in presence of LRRC28 likely reflects the metabolic state differences between cell lines and
399 suggests differential dependency on LRRC28. Taken together, TEC reveals shared biological
400 functions and predicts novel associations, providing insights not attainable with RNA co-
401 expression analysis alone.


a
Collagen-containing extracellular matrix
TEC < -0.1
TEC > 0.1
LOX
b
TE
HK1
HK2
PFKM
PFKL
PFKP
FOXK1
LRRC28
TPI1
FOXK2
PGK1
ENO1
ENO2
PKM
TEC < -0.1
TEC > 0.1
RNA
RNA co-expression < -0.1
RNA co-expression > 0.1
c
ENO1
ENO2
ENO3
FOXK1
FOXK2
HK1
HK2
PFKL
PFKM
PFKP
PGK1
PKM
TPI1
LRRC28
TE
RNA
RNA expression and TE (CLR)
Human tissue and cell lines
e
FOXK1
LRRC28
interface <= 4Å
DNA binding domain
d
LRRC28
LRRC42
pDOCKQ
ipTM+pTM
f
TFs from forkhead family
FOXN3
FOXK2
FOXH1
FOXN2
FOXR1
g
SH-SY5Y
LRRC28
LRRC42
Glucose
2-DG
% ECAR
Time
h
Huh 7.5

402

Fig. 5 | TEC enables the prediction of novel gene functions. a, We predicted that LOX belongs to the collagen-containing extracellular matrix using an older version of human GO terms (January 1, 2021) and confirmed this prediction with the newer version (December 4, 2022; Methods). The network displays the similarity in TE between LOX (yellow dot) and other genes (gray dots) from the collagen-containing extracellular matrix. Line weight in figure panels indicates the absolute value of rho from 0.1 to 1. **b,** The networks display the rho between LRRC28 and glycolytic genes at the TE level (on the left) and RNA level (on the right) in humans. Green dots represent genes 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[75] and pDOCKQ > 0.23[76,77] 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[78]. The dashed lines represent ipTM+pTM > 0.7 or pDOCKQ > 0.23. Kinetic ECAR response of **g,** SHSY-5Y cell line (n=6, stable overexpression) and **h,** Huh 7.5 cell line (n=6; transient overexpression) overexpressing LRRC28 or LRCC42 to 10 mM glucose and 100 mM 2-DG. Unpaired two-sided Student's *t*-test, ***$P < 0.005$, **$P < 0.05$. Panel g & h shows mean ± s.d.; *n* shows biological independent experiments.

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[14,79,80], and many protein complexes are assembled co-translationally[81], 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[24,25]. To test this hypothesis, we categorized gene pairs by whether they display positive or negative similarity in RNA expression or TE across cell types. 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 from STRING database[80]) 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). 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[82] which are enriched for positive rho values.

We then examined whether these patterns generalize to the higher-order organization of protein complexes. We observed protein complexes (as defined by hu.MAP[83]) 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[84,85]. Several known exocyst-binding partners are not required for its assembly and stability, indicating that the molecular details are still unclear[84]. 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[24,25,86].

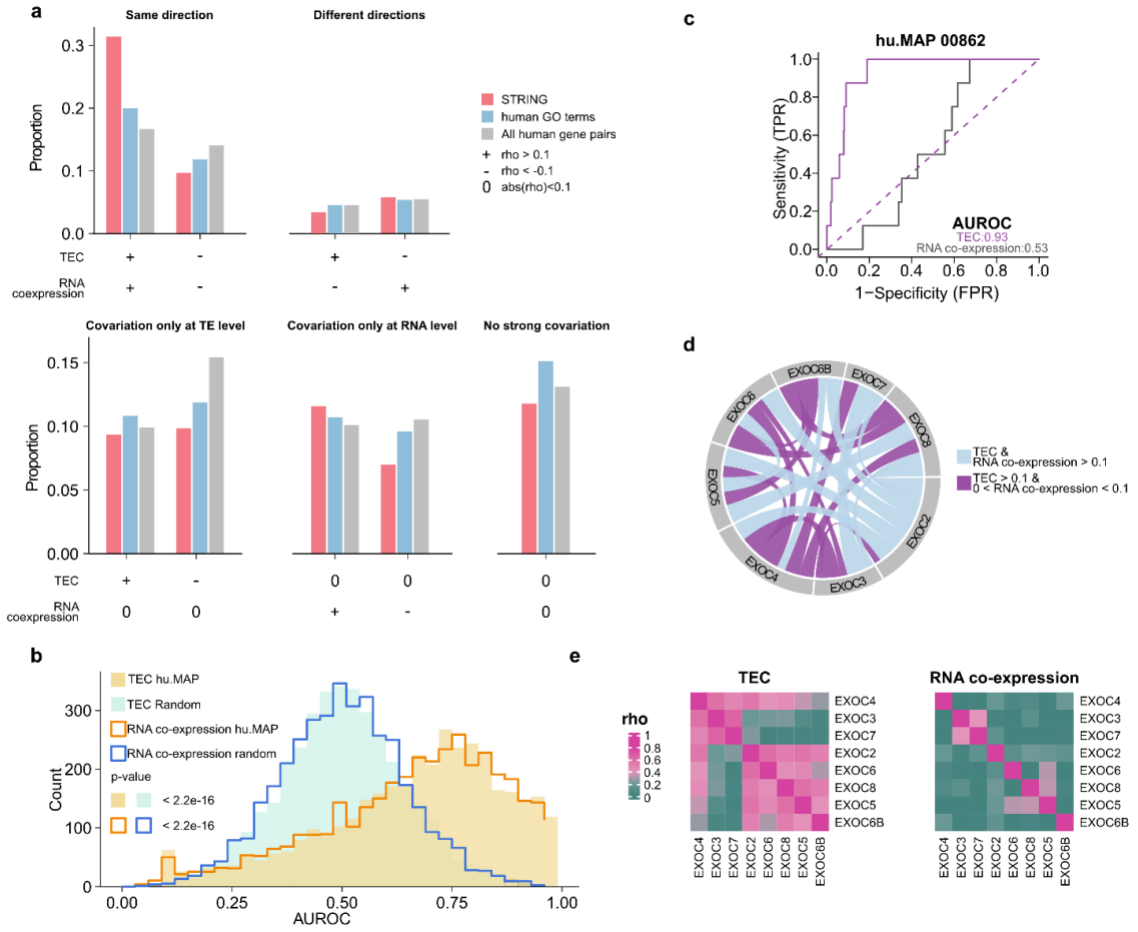


Fig. 6 | Physically interacting proteins display TEC. a, Number of pairs of genes among three sets (physical interaction-red; shared function-blue; all genes-gray) categorized based on the direction of similarity (Rho scores). **b,** The distribution AUROC calculated with either TEC or RNA co-expression 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 performed a large-scale analysis of ribosome profiling and matched RNA-seq data across diverse human and mouse cell types to quantify TE. A particular challenge was inadequate metadata associated with these experiments, which limits their reuse (supplementary text). To address this issue, we conducted a manual curation process and made several methodological advances including the selection of RPF read lengths, data normalization, and estimation of TE. Although the analyzed datasets are predominantly derived from peer-reviewed publications, over 20% failed basic quality control. Our findings point to a pressing need for community standards for data quality and structured metadata.

Consistent with its established use in prior literature, we used the term "translation efficiency" to refer to ribosome occupancy per unit of mRNA (ribosome density). Recent work has questioned its interpretation as the efficiency of protein synthesis at least in the context of base-modified reporter constructs[87]. However, our work and that of others indicate TE is significantly positively correlated with protein abundance and synthesis rate for endogenous transcripts[88]. Thus, for most endogenous mRNAs, TE (ribosome density) is positively associated with protein production.

In this study, we introduced the concept of ***Translation Efficiency Covariation*** (TEC) to quantify the similarity of TE patterns across cell types. Remarkably, we found that TEC relationships are evolutionarily conserved across orthologous gene pairs in humans and mice, suggesting that translational coordination is a conserved feature of mammalian transcriptome organization. Comparative analyses of TEC and RNA co-expression networks may help distinguish shared and distinct regulatory architectures, and reveal network-level conservation and rewiring.

Importantly, TEC is predictive of gene function. Although RNA expression and TE levels are typically weakly correlated across cell types, genes with shared functions frequently exhibit coordinated patterns of both. This coordination may enhance cellular energy conservation and responsiveness to environmental cues. Analysis of TEC revealed unique insights into protein function missed by transcriptomics and proteomics. For example, LRRC28 showed strong TEC covariation with glycolytic enzymes despite absent RNA co-expression. These patterns were also undetectable at the protein-level as LRRC28 is absent from most proteomic databases such as PAXdb and ProteomeHD[44,79].

We experimentally validated the role of LRRC28 in attenuating glycolysis in two human cell lines. Further, we identified a high confidence predicted interaction between LRRC28 and FOXK1, the key transcription factor controlling glycolytic enzyme expression, suggesting a potential mechanism. Given that LRRC28 is downregulated in multiple cancers[89–91], its regulation may represent a previously unrecognized axis of metabolic control with therapeutic relevance [70,92,93].

These findings prompted us to systematically analyze TEC among physically interacting proteins,

revealing a significant enrichment of positive TEC among such proteins. Coordination of TE between proteins may facilitate their co-translational assembly[81] into complexes and contribute to their stoichiometric production, hence reducing the energetic costs of orphan protein production. These results also highlight the potential utility of TEC for synthetic biology applications. Design of synthetic gene circuits often requires precise stoichiometric production of multi-subunit complexes or pathway enzymes. In the accompanying manuscript[94], we have developed machine learning approaches to predict TE from mRNA sequences. Such predictive modeling approaches can be combined with TEC to guide balanced translational output in engineered systems. This optimization may particularly be advantageous in applications for improving yield given that protein biosynthesis is the largest consumer of energy during cellular proliferation[25,28,86].

We acknowledge several limitations in our study. First, the limited number of samples for certain cell lines may affect the reliability of corresponding TE estimates. Second, our analyses focused on genes robustly expressed across most cell types, limiting our ability to assess TEC for genes with cell-type-specific or consistently low expression levels. Third, due to limitations in RNA-seq quality and the short length of ribosome-protected fragments, we restricted our analysis to one representative isoform per gene[95]. As a result, we may miss regulatory features that vary across isoforms such as alternatively spliced UTRs that influence miRNAs or RBP binding. Therefore, our analyses are unable to adequately capture isoform specific usage that may affect TEC. Fourth, given the compositional nature of all sequencing experiments, TE reflects the relative allocation of translational resources. Hence, a transcript can exhibit relatively higher TE even without changes in its polysome distribution, as observed in global translational capacity shifts during viral infections and other biological contexts[96,97]. Moreover, TEC is currently defined at the level of bulk cell populations. As single-cell ribosome profiling technologies mature[98,99], TEC could be extended to quantify translational coordination within heterogeneous cellular environments, enabling context-aware models of post-transcriptional regulation in development and disease.

In summary, TEC nominates new gene functions not captured by RNA or protein expression alone. The discovery of translation-level coordination among functionally related genes and physically interacting proteins across diverse cell types supports TEC as a conserved and functionally meaningful organizing principle of mammalian gene expression.

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 Yuna Shin and Victoria D. Chapman for their help with the experiments. We appreciate Dr. Milad Miladi for providing critical feedback. The original text in this paper was written by the authors. A LLM was used to suggest edits for clarity and grammar[100]. 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) and R35GM138340 (ESC). This work was also supported by the National Institutes of Health grant [HD110096], and the Welch Foundation grant [F-2027-20230405] (C.C.), [F-2133-20230405] (ESC). 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., I.H., and S.R. 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., and 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. S.R. conducted the experimental validation of LRRC28. S.R., I.H., V.G., and D.P. performed other 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. I.H. is an employee of Monoceros Biosystems.

METHODS

Acquisition and curation of ribosome profiling data

We used keyword search (“ribosome profiling”, “riboseq”, “ribo-seq”, “translation”, “ribo”,

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

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

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

611 **Ribosome profiling and RNA-seq data processing**

612 For each selected study in GEO, ribosome profiling and matching RNA-Seq reads (where available)
613 were downloaded, from SRA, using the SRA-Tools version 2.9.1[101], in FASTQ format using their
614 accession numbers. FASTQ files were processed using RiboFlow[31] where parameters were
615 determined using the metadata in RiboBase. The reference files for human and mouse
616 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. The detailed documentation and code used to select representative transcripts are also provided: https://github.com/ribosomeprofiling/references_for_riboflow/tree/master/transcriptome/human/v2 and https://github.com/ribosomeprofiling/references_for_riboflow/tree/master/transcriptome/mouse/v1. Briefly, the 3' adapters of the ribosome profiling reads were trimmed using Cutadapt version 1.18[102] 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[31]. All alignment steps used bowtie2 version 2.3.4.3[103]. 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[104].

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; Supplementary Fig. 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[33,105–107]. 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 (Supplementary Fig. 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

657 (PCR-deduplication) ribo files (Fig. 1d-e). We set two cutoffs for ribosome profiling quality
658 control. We required that on average each nucleotide of the transcript should be covered at least
659 0.1 times (0.1X). Coverage was calculated with the formula:

660 $$coverage = \frac{total\ nucleotides\ from\ reads\ mapped\ to\ transcripts}{total\ length\ of\ transcripts}$$

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

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

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

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

679 **PCR and Unique Molecular Identifiers (UMIs) deduplication comparison**

680 We selected eight ribosome profiling experiments that incorporated UMIs into the sequence library
681 preparation to assess the impact of different deduplication methods. Specifically, these samples
682 are GSM4282032, GSM4282033, and GSM4282034 from GSE144140[108]; GSM3168387,
683 GSM3168389, GSM3168390 from GSE115162[106]; and GSM4798525, GSM4798526 from
684 GSE158374[96]. We processed the data using Riboflow, applying three different deduplication
685 methods: non-deduplication, PCR deduplication, and UMI deduplication. The yaml files are
686 available at https://github.com/CenikLab/TE_model/tree/main/riboflow_scr. The RPF length
687 cutoffs for samples from GSE144140 and GSE115162 are listed in table S6. Since GSE158374 is
688 not currently included in RiboBase, we manually performed the dynamic module and selected 28
689 to 32 as the RPF cutoff for this study.

690 **Winsorization of CDS mapping read counts**

691 To address the issue of reduced usable reads resulting from PCR deduplication (supplementary
692 text), we employed a winsorization method, which was previously proposed for tackling this

693 problem[24,109]. For each gene's CDS region, we obtained the distribution of non-deduplicated
694 nucleotide counts and calculated the 99.5th percentile value. This calculation was based on reads
695 whose lengths fell within the RPF range determined by the RPF boundary selection function. RPF
696 counts that exceed the 99.5th percentile were capped to the value corresponding to the 99.5th
697 percentile. This method was designed to mitigate the impact of outlier values that might arise due
698 to disproportionate amplification during the PCR process[24].

699 **Gene filtering and normalization for ribosome profiling and RNA-seq**

700 RNA-seq experiments in RiboBase utilized several different strategies to enrich mRNAs. The two
701 most common approaches were the depletion of ribosomal RNAs and the enrichment of transcripts
702 by polyA-tail selection. This difference leads to dramatically different quantification of a subset
703 of genes that lack polyA-tails (e.g. histone genes, Supplementary Fig. 6c). Hence, we removed 166
704 human and 51 mouse genes identified as lacking polyA tails (table S8-9)[110,111].

705 We normalized both PCR-deduplicated RNA-seq data and winsorized non-deduplicated ribosome
706 profiling data with counts per million (CPM) after removing the genes without polyA-tails. Genes
707 with CPM less than one in over 30% of the total samples in both RNA-seq and ribosome profiling
708 for either human or mouse were removed in further analyses. 11,149 human and 11,434 mouse
709 genes were retained using this approach. We have summed the counts of all polyA genes that were
710 filtered out and grouped them under 'others' in the count table.

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

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

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

728 **Translation Efficiency (TE) calculation**

729 CLR normalized counts from PCR-deduplicated RNA-seq and winsorized non-deduplicated
730 ribosome profiling were used to calculate TE with compositional linear regression[36,112,113]. Missing

731 values can limit the power of TE calculations, and there is ongoing debate on how to address zeros
732 in compositional data [114]. Previous studies have shown that approximately 15–40% of genes in
733 bulk RNA-seq datasets from various tissues are not expressed [115]. To address the issue of missing
734 values, we followed the guidance from the propr library, replacing zeros with the lowest observed
735 read count (one). This approach ensures that these zeros are represented as a low proportion within
736 the sample. We also applied two other zero imputation methods: Geometric Bayesian-
737 Multiplicative and Square-root Bayesian-Multiplicative from zCompositions package
738 (version:1.5.0.4) for TE and TEC calculations.

739 In our linear regression approach, ribosome profiling data served as the dependent variable, while
740 the corresponding RNA-seq data provided the explanatory variable. The first step involved
741 transforming the gene count, which includes 'others', into CLR normalized compositional vectors.
742 Given the constraints of count data within a simplex, a further transformation from CLR to
743 Isometric Log Ratio (ILR) was necessary for linear regression[36]. This transformation is crucial as
744 it allows the compositional data to be decomposed into an array of uncorrelated variables while
745 preserving relative proportions. The ILR transformation projects the original data onto a set of
746 orthonormal basis vectors derived from the Aitchison simplex. Then the linear regression model
747 applied to these transformed variables can be represented as:

748 $$Y = b + B \times X$$

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

751 $$Y \sim N^{(D-1)}(Y, \Sigma\varepsilon)$$

752 Where Σε represents the residual variances. These residuals were then extracted from each sample
753 and reconverted to CLR coordinates which are used as the definition of TE for each gene in each
754 sample. Finally, we averaged TE for different cell lines and tissues (Fig. 2b, Supplementary Fig.
755 7), and reported the TE in table S10-11. The scripts to generate TE are available at
756 https://github.com/CenikLab/TE_model. We also provide uncertainty estimates calculated using
757 the metric standard deviation designed for compositional data from the Compositions (version:
758 2.0.8) library (table S21-22).

759 **Correlation between translation efficiency and protein abundance**

760 We assessed the correlation between TE and protein abundance from seven human cell lines (A549,
761 HEK293, HeLa, HepG2, K562, MCF7, and U2OS). The protein measurements were obtained from
762 PAXdb[44]. 9924 genes were shared between our TE and the protein abundance data. We calculated
763 the Spearman correlation coefficient for each cell line using the 'stats' package in R to evaluate
764 the relationship between TE and protein abundance. For a more detailed discussion of how to
765 interpret similarity between protein abundance and sequencing based measurements of RNA-
766 expression and translation see Csardi et al.[116].

767 **Conservation of translation efficiency between orthologous genes from human and mouse**

768 Orthologous genes between human and mouse were identified using the 'orthogene' package from
769 Bioconductor[117] using the parameters 'standardise_genes=TRUE,
770 method_all_genes="homologene", non121_strategy="keep_both_species"'. A single human gene
771 could correspond to multiple mouse orthologs or vice versa. To maintain all one-to-many matches
772 in our analysis, each correspondence is represented by multiple rows in our table (if a human gene
773 'A' is orthologous to mouse genes 'B' and 'C', we generate two separate rows: 'A-B' and 'A-C').
774 Human genes lacking corresponding mouse orthologs were excluded or vice versa. As a result, a
775 total of 9,194 gene pairs were identified as orthologous between human and mouse (table S12)

776 To capture the variability in TE and mRNA expression between orthologous genes in human and
777 mouse, we measured the standard deviation using the metric standard deviation (msd) function
778 from the 'compositions' package in R[118]. We observed a negative Spearman correlation coefficient
779 between msd of TE and mean TE, as well as msd of RNA expression and mean RNA expression,
780 in both species. To address the dependency between msd and mean values, we conducted a partial
781 correlation analysis. For example, we adjusted the human msd values using the mean TE from
782 both human and mouse with the 'pcor.test' function from the 'ppcor' package[119].

783 GO term analysis was performed using FuncAssociate 3.0, accessible at
784 http://llama.mshri.on.ca/funcassociate/[120]. For this analysis, we set either 9,194 mouse or 9,189
785 human orthologous genes as the background. We generated association files for these genes with
786 the December 4, 2022 version of human or mouse GO terms. In the human or mouse association
787 file, we only kept those GO terms containing at least 10 genes for further analysis.

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

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

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

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

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

797 Method 4 - CLR-normalized ribosome footprint counts were used to calculate the proportionality
798 (rho scores) between genes as implemented in the propr package with lr2rho function[38].

799 Method 5 - CPM-normalized ribosome footprint counts were used to calculate the similarity
800 between genes with a decision tree-based method as implemented in the treeClust package[79,121].
801 We applied the 'treeClust.dist' function with a dissimilarity specifier set to d.num=2.

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

804 Method 7 - CPM-normalized ribosome footprint counts were used to calculate gene similarity with
805 the generalized least squares (GLS) method[122].

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

808 We compared these eight ribosome occupancy similarity matrices to determine the most effective
809 method for constructing gene relationships with respect to biological functions. This assessment
810 employed the guilt by association principle to ascertain the functional coherence within a gene
811 matrix, determining if genes associated with a particular biological function (GO terms[123], TOP
812 mRNAs[124]) exhibit similar expression patterns and network interactions[125].

813 The complete ontology was sourced from the Gene Ontology website, with the files
814 goa_human.gpad.gz and mgi.gpad.gz, generated on December 4, 2022[123]. The annotation of Gene
815 Ontology terms was accomplished with the aid of the org.Hs.eg.db and org.Mm.eg.db R
816 packages[126,127]. We restricted the selection of GO terms to those associated with the 11,149 human
817 and 11,434 mouse genes that had passed gene filtering. We used GO terms associated with at least
818 10 but less than 1,000 genes for evaluation, yielding a total of 2,989 human and 3,340 mouse GO
819 terms.

820 We then employed the neighbor-voting algorithm to assess the covariations of ribosome
821 occupancy among genes from the same GO term with AUROC[125]. Specifically, we first converted
822 the similarity scores to absolute values. Then we extracted genes associated with a specific
823 function and implemented the leave-one-out cross-validation method. For this analysis, we
824 iteratively masked one gene at a time, treating it as if it did not belong to the function. In each
825 iteration, we calculated the total sum of similarity scores from all genes not belonging to the
826 function to all the remaining genes within the function. We normalized the sum of similarity scores
827 for each gene against the sum of similarity scores for that gene with all genes. After normalization,
828 we converted these normalized similarity scores into rankings. We retained the rankings only for
829 genes that belong to this specified functional property. Finally, we computed the AUROC for all
830 genes within this functional property based on these rankings. A detailed script for genes'
831 functional similarity pattern analysis can be found:
832 https://github.com/CenikLab/TE_model/blob/main/other_scr/benchmarking.R.

833 **RNA co-expression and translation efficiency covariation**

834 We introduce the concept of TEC, which employs a compositional data analysis approach[36,38] to
835 quantify the similarity patterns of TE across various cell and tissue sources, as described in Method
836 4 above. The proportionality scores were calculated with the following formula from the propr
837 package with lr2rho function[38]:

838 $$Rho(A_i, A_j) = \frac{1 - var(A_i - A_j)}{var(A_i) + var(A_j)}$$

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

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

844 **Evaluation of the ability of TEC to predict novel gene functions**

845 We compared the AUROC between an older version of GO terms (January 1, 2021) to the newer
846 version of GO terms (December 4, 2022) to identify genes that had been newly added to the GO
847 terms in this timeframe. GO terms were downloaded and filtered to include only those terms
848 containing between 10 and 1,000 genes with either human or mouse backgrounds (11,149 human
849 genes or 11,434 mouse genes). We selected 184 human and 238 mouse GO terms from the older
850 version that demonstrated high TEC similarity (AUROC > 0.8) among genes within the same term
851 for predicting novel gene functions. We first converted the rho scores for TEC between gene pairs
852 to absolute values. For genes not currently included in the GO terms, we calculated the sum of rho
853 for each gene relative to all genes within the term, based on either TE or mRNA expression levels.
854 We then normalized these rho sums for each gene against the total rho sum of that gene across all
855 11,149 human genes or 11,434 mouse genes. These normalized values were converted into ranking
856 percentages to reflect the likelihood of these genes being associated with the respective GO term.
857 Finally, we identified the top-ranking genes as potentially new additions and cross-validated them
858 with the newer version of the GO terms to confirm our predictions.

859 **Prediction of novel gene functions with TEC**

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

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

871 We computed the pair-wise interaction probabilities between LRRC28 or LRRC42 and glycolytic
872 proteins (HK1, HK2, PFKL, PFKM, PFKP, TPI1, PGK1, ENO1, ENO2, and PKM) with
873 AlphaFold2-Multimer 2.3.0[72,128]. In addition, we also calculated pairwise interaction probabilities
874 for LRRC28 with 35 proteins from the forkhead transcription TF family[78]. We extracted the
875 canonical amino acid sequence for each gene from UniPort[129] as the input file. We set 0.7 as the
876 cutoff of ipTM+pTM as a high-confidence protein structure and binding probability cutoff[75]. We
877 then evaluated the interfaces predicted by AlphaFold2-Multimer, using a pDOCKQ score greater
878 than 0.23 as our criterion for reliability[76,77].

879 **Benchmarking TEC and RNA co-expression for protein interactions**

880 Using a similar approach to our benchmarking with biological functions, we employed the
881 neighbor-voting algorithm to assess physical protein interactions based on rho scores among genes
882 at either the TE or mRNA expression level. We first kept the non-negative rho between genes and
883 set negative rho to zero. We then analyzed similarity patterns between genes from the same protein
884 complex, downloading from the hu.MAP 2.0 website[83]. In this process, we excluded genes from
885 hu.MAP terms that were not in the 11,149 human gene list, resulting in 8,024 overlapping genes
886 between our list and hu.MAP terms. Furthermore, we removed hu.MAP terms that included fewer
887 than three genes. This filtering process left us with 3,880 hu.MAP terms, among which 3,755
888 contained unique genes.

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

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

893 To reduce bias in motif enrichment analysis that may arise by ribosome footprint mapping to
894 paralogous genes, we removed predicted paralogs from each GO term using Paralog Explorer[130]
895 (DIOPT score > 1). Then, we enumerated heptamers in each transcript region using the Transite
896 kmer-TSMA method[60] with default parameters for each species (human, mouse), transcript region
897 (5' UTR, CDS, 3' UTR), and GO term (selected terms with TE AUROC > 0.7, TE-RNA AUROC
898 difference > 0.2, and number genes after paralog removal >= 12). We selected the three mouse
899 terms and top five terms in humans with the highest number of genes and greatest AUROC
900 difference.

901 After counting heptamers with Transite, we selected motifs that had >20 hits among genes in the
902 GO term to address assumptions of uniformity near p-values of 1 for some multiple-test correction
903 methods. Then, we used the Holm method to correct p-values for each species separately, and
904 selected motifs with an adjusted p-value < 0.05. Finally, heptamers were annotated with RBPs
905 included in the Transite[60] and oRNAment databases[59]. For annotation of RBPs in the oRNAment
906 database, we required that the heptamer have a matrix similarity score[59] of 0.8 or greater when
907 matching to each RBP position weight matrix. RBP motif hits from other species (*Drosophila*,
908 artificial constructs) were removed from RBP annotations, and the hits to the heptamer of
909 *Drosophila* tra2 were annotated as TRA2A for human genes with the term GO: 0140678.

910 eCLIP data for PABPN1, SRSF1, and TRA2A were downloaded from ENCODE[61] as BED files
911 (K562 and HepG2 cell lines, GRCh38 reference). The BED files for biological replicates were
912 concatenated and peaks that overlapped by at least one base pair were merged with 'bedtools merge
913 -s -c 4,6,7 -o collapse'[131]. The resulting merged peaks were intersected with transcripts in the GO
914 term of interest and an equal number of control transcripts (Gencode v34 GTF). The control
915 transcripts were selected by matching on length and GC content for each transcript region (5' UTR,
916 CDS, 3' UTR) using MatchIt[132] with default parameters. Because the gene *CARMIL2* in GO term
917 GO:0010592 does not have a 5' UTR, required for matching, we assigned it a dummy 5' UTR
918 with length and GC content equal to the median across all transcripts. The number of eCLIP peaks

919 in the CDS for each RBP were summed for genes in the GO term and control genes.

920 **Identification of RBP-transcript pairs with high correlation between RBP RNA expression**
921 **and transcript TE**

922 The Pearson correlation coefficient between transcript TE and the RNA expression of RBPs from
923 human and mouse[51] was tested using R stats::cor.test after taking the mean of these values by cell
924 types and tissues. P-values were corrected with the Benjamini-Hochberg procedure, and
925 correlations were deemed significant at a FDR < 0.05. For cloning the guides required for knockout
926 cell line generation, top two ranked guides were selected from the Brunello library[133] for each
927 RBPs (table S19). The guides were cloned in LentiCRISPRv2 (Addgene, 52961) as per protocol[134]
928 and confirmed by Sanger sequencing. Briefly, for lentiviral production, HEK293T cells were
929 seeded at a density of 1.2 x 10^6 cells per well in a 6-well plate in OPTI-MEM media supplemented
930 with 5% FBS and 100 mM Sodium Pyruvate, 24 h prior to transfection. Both the cloned gRNA
931 plasmids for each RBP (700 ng of each transfer plasmid) were co-transfected with the packaging
932 plasmids pMD2.G and psPAX2 (Addgene; 12259 and 12260) using Lipofectamine 3000
933 (Invitrogen) and the virus was collected as per the manufacturer's protocol. For generation of the
934 knockout clones, HEK293T cells were seeded at a density of 5 x 10^4 cells per well in a 6-well plate
935 in DMEM media supplemented with 10% FBS, 24 h prior to infection. Next day, the media was
936 replaced with 1.5 ml of 1:2 diluted lentivirus containing polybrene (8 µg/mL). After 16 h, the
937 lentivirus was replaced with fresh media and puromycin (2 µg/mL) was added to the cells 48 h
938 after transduction. The selection continued for 5 days followed by a period of recovery for 24 h
939 before harvesting the cells.

940 **Ribosome profiling and RNA sequencing of RBP knockout cell lines**

941 Three million cells for the PARK7, USP42, and VIM knockout cell lines, along with an AAVS1
942 (safe harbor control) knockout line, were plated in three 10 cm^2 dishes. 27 h later cells at ~60%
943 confluency were treated with 100 µg/mL cycloheximide (CHX) for 10 min at 37 ˚C, then collected
944 in ice-cold PBS with 100 µg/mL CHX. Cells were spun at 100 x g for 7 min at 4 ˚C, then flash
945 frozen in liquid nitrogen and stored at -80 ˚C. Cell pellets were lysed with 400 µL lysis buffer (20
946 mM Tris-HCl pH 7.4, 150 mM NaCl, 5 mM $MgCl_2$, 1 mM DTT, 1% Triton-X, 100 µg/mL CHX,
947 1x protease inhibitor EDTA free) for 10 min on ice. Lysates were clarified by centrifugation at
948 1,300 rpm for 10 min at 4 ˚C. 40 µL lysate was saved for total RNA extraction, and the rest of the
949 lysate was digested with 7 µL RNaseI for 1 h at 4 ˚C. Digestion was stopped by adding
950 ribonucleoside vanadyl complex to a final concentration of 20 mM. Digested lysates were then
951 loaded onto a 10 mL sucrose cushion (20 mM Tris-HCl pH 7.4, 150 mM NaCl, 5 mM $MgCl_2$, 1
952 mM DTT, 1 M sucrose) and centrifuged at 38,000 rpm for 2.5 h at 4 ˚C using a SW41-Ti rotor.
953 The pellet and the total RNA aliquot were both solubilized with 1 mL Trizol, and RNA was
954 purified with the Zymo Direct-zol RNA Miniprep Kit, including DNaseI digestion.

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

958 RPFs were size-selected on a 15% TBE urea gel by electrophoresing at 150 V for 1.5 h. RPFs

959 between 28-32 nt were sliced, using 28 nt and 35 nt markers as a guide. Slices were frozen at -20
960 ˚C for 1 h, crushed with pestles, and the RPFs were eluted in gel extraction buffer (300 mM sodium
961 acetate pH 5.5, 5 mM $MgCl_2$) by rotating overnight at room temperature. Eluates were passed
962 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
963 GlycoBlue, and 1 mL ethanol were added and the RPFs precipitated for two days at -20 ˚C. Pellets
964 were dried and resuspended in 16 µL water.

965 Libraries were generated from 8 µL RPF eluate using the Diagenode D-Plex Small RNA Kit with
966 minor modifications: in the 3' dephosphorylation step 0.5 µL T4 PNK was supplemented and
967 incubated for 25 min. The RTPM reverse transcription primer was used and 8 cycles were
968 performed for PCR. Libraries were quantified by Bioanalyzer High Sensitivity DNA Kit, pooled
969 equimolar according to the quantity of the peak for libraries with full-length inserts (~204 nt), and
970 cleaned up with 1.8X AMPure XP beads. Adapter dimers and empty libraries were removed by
971 size selection on a 12% TBE PAGE gel, followed by extraction with the crush and soak method,
972 and final libraries were resuspended in 20 µL water.

973 **Ribosome profiling and RNA sequencing analysis for RBP knockouts**

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

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

988 **Cell Culture**

989 HEK293T and MCF-7 cell lines were obtained from ATCC. Human hepatoma Huh-7.5 cells were
990 a gift from Charles Rice, Rockefeller University and human SH-SY5Y neuroblastoma cells were
991 obtained from Tanya Paull (University of Texas at Austin, Austin, TX, USA). All cell lines were
992 maintained in Dulbecco's modified Eagle's media (DMEM, GIBCO) supplemented with 10% Fetal
993 Bovine Serum (FBS, GIBCO, Life Technologies) and 1% Penicillin and Streptomycin (GIBCO,
994 Life Technologies) at 37°C in 5% CO_2 atmosphere. Cell lines were tested for mycoplasma
995 contamination every 6 months.

996 **Plasmid cloning**

997 pCMV SPORT6 vector with LRRC28 was obtained from transOMIC (TCH1003). LRRC42 ORF
998 was amplified using cDNA from HEK293T cells. The oligos used for PCR amplification of the
999 two genes had a FLAG sequence to be incorporated as a N-terminal tag along with BamHI and
1000 EcoRI restriction sites in forward and reverse primers, respectively.

1001 PCR amplification of the LRRC28 and LRRC42 ORF using the above mentioned templates were
1002 carried out using Q5 polymerase (NEB) with the following protocol; 98 °C for 30 sec, 30 cycles
1003 of 98 °C for 10 sec, 55 °C for 10 sec and 72 °C for 40 sec and a final extension at 72 °C for 2 min.
1004 The PCR amplified products were gel extracted followed by double digestion of the amplicon and
1005 the pLVX-M vector (Addgene, 125839) with EcoRI -HF (NEB) and BamHI-HF (NEB) at 37 °C
1006 for 1 h. Post digestion, the fragments were purified and cloned using T4 DNA ligase at a ratio of
1007 1:3 at 25 °C for 1 h followed by heat inactivation at 65 °C for 15 min and transformation by heat
1008 shock method. For LRRC42 cloning, ORF was amplified using cDNA from HEK293T cell line
1009 and two sequences (seq1 and seq 2) from the pCMV-SPORT6 adjacent to EcoRI and NotI site
1010 were amplified with LRRC42 overhangs using oligos in table S24. The vector; pCMV SPORT6
1011 vector with LRRC28 was digested with EcoRI and NotI followed by cloning of LRRC42 and two
1012 sequences by Gibson cloning. All clones were verified by Sanger sequencing (ACGT Inc.) and
1013 full plasmid sequencing (Plasmidsaurus).

1014 **Lentivirus and stable cell line generation:**

1015 For lentiviral generation, 1.2 X 10^6 HEK293T cells were seeded in a 6-well plate followed by
1016 transfection of the cloned plasmid along with psPAX2 (Addgene, 12260) and pMD2.G (Addgene,
1017 12259) plasmids using Lipofectamine 3000 (Invitrogen) as per the manufacturer's protocol. The
1018 generated lentivirus was used for infection using 8 µg/ ml of polybrene for HEK293T, MCF-7 and
1019 Huh 7.5 and 2 µg/ ml of polybrene for SH-SY5Y. After 48 h of infection, puromycin was added
1020 at concentrations of 2µg/ ml for HEK293T, Huh 7.5 and 1µg/ ml for MCF-7, SH-SY5Y. The
1021 selection was continued for 5 days following recovery in antibiotic-free media.

1022 **Measurement of extracellular acidification rates (ECAR):**

1023 20,000 cells of HEK293T, MCF-7 and 40,000 cells of SH-SY5Y stably expressing LRRC28 or
1024 LRRC48 and 20,000 cells of Huh 7.5 were seeded onto poly-L lysine coated XF Pro Cell Culture
1025 Microplates with Dulbecco's modified Eagle's media (DMEM, GIBCO) supplemented with 10%
1026 Fetal Bovine Serum (FBS, GIBCO, Life Technologies). Huh 7.5 was transiently transfected with
1027 either pCMV SPORT6 vector with LRRC28 or LRRC42 using lipofectamine 3000 as per
1028 manufacturer's protocol. The ECAR was measured using Seahorse XF Pro (Agilent). The
1029 cartridges were hydrated a day prior to the experiment as per the manufacturer's instructions. On
1030 the day of the experiment, prior to performing the assay, the cells were washed with Seahorse XF
1031 DMEM Medium pH 7.4 (Agilient, 103575), supplemented with 2 mM of L- Glutamine (Agilient;
1032 103579-100) followed by incubation in non-CO2 incubator for 1 h. To determine glycolytic
1033 capacity of cells expressing LRRC28 or LRRC42, glucose (Agilient, 103577-100) was injected at
1034 a final concentration of 10 mM (Injection 1) followed by injection of 100 mM 2-Deoxy-D- Glucose
1035 (Sigma, D6134-1G) (Injection 2). Three measurement cycles for each assay point i.e. baseline,
1036 injection 1 and 2 were carried out with each measurement cycle consisting of a mixing time of 3

1037 minutes and a data acquisition period of 3 minutes. The results were analyzed using Wave Pro
1038 10.2.1.

1039

1040 **REFERENCES CITED**

1041 1. Tang, F. *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* **6**,
1042 377–382 (2009).

1043 2. Nagalakshmi, U. *et al.* The transcriptional landscape of the yeast genome defined by RNA
1044 sequencing. *Science* **320**, 1344–1349 (2008).

1045 3. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and
1046 quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **5**, 621–628 (2008).

1047 4. Schena, M., Shalon, D., Davis, R. W. & Brown, P. O. Quantitative monitoring of gene
1048 expression patterns with a complementary DNA microarray. *Science* **270**, 467–470 (1995).

1049 5. Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S. & Zhuang, X. RNA imaging.
1050 Spatially resolved, highly multiplexed RNA profiling in single cells. *Science* **348**, aaa6090
1051 (2015).

1052 6. Combs, P. A. & Eisen, M. B. Sequencing mRNA from cryo-sliced Drosophila embryos to
1053 determine genome-wide spatial patterns of gene expression. *PLoS One* **8**, e71820 (2013).

1054 7. Achim, K. *et al.* High-throughput spatial mapping of single-cell RNA-seq data to tissue of
1055 origin. *Nat. Biotechnol.* **33**, 503–509 (2015).

1056 8. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network
1057 analysis. *BMC Bioinformatics* **9**, 559 (2008).

1058 9. Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. Cluster analysis and display of
1059 genome-wide expression patterns. *Proc. Natl. Acad. Sci. U. S. A.* **95**, 14863–14868 (1998).

1060 10. Skinnider, M. A., Squair, J. W. & Foster, L. J. Evaluating measures of association for
1061 single-cell transcriptomics. *Nat. Methods* **16**, 381–386 (2019).
1062 11. Stuart, J. M., Segal, E., Koller, D. & Kim, S. K. A gene-coexpression network for global
1063 discovery of conserved genetic modules. *Science* **302**, 249–255 (2003).
1064 12. Marcotte, E. M., Pellegrini, M., Thompson, M. J., Yeates, T. O. & Eisenberg, D. A
1065 combined algorithm for genome-wide prediction of protein function. *Nature* **402**, 83–86
1066 (1999).
1067 13. DeRisi, J. L., Iyer, V. R. & Brown, P. O. Exploring the metabolic and genetic control of
1068 gene expression on a genomic scale. *Science* **278**, 680–686 (1997).
1069 14. Jansen, R., Greenbaum, D. & Gerstein, M. Relating whole-genome expression data with
1070 protein-protein interactions. *Genome Res.* **12**, 37–46 (2002).
1071 15. Szklarczyk, D. *et al.* The STRING database in 2023: protein-protein association networks
1072 and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids*
1073 *Res.* **51**, D638–D646 (2023).
1074 16. Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J. & Church, G. M. Systematic
1075 determination of genetic network architecture. *Nat. Genet.* **22**, 281–285 (1999).
1076 17. Roth, F. P., Hughes, J. D., Estep, P. W. & Church, G. M. Finding DNA regulatory motifs
1077 within unaligned noncoding sequences clustered by whole-genome mRNA quantitation.
1078 *Nat. Biotechnol.* **16**, 939–945 (1998).
1079 18. Nusinow, D. P. *et al.* Quantitative Proteomics of the Cancer Cell Line Encyclopedia. *Cell*
1080 **180**, 387–402.e16 (2020).
1081 19. Gonçalves, E. *et al.* Pan-cancer proteomic map of 949 human cell lines. *Cancer Cell* **40**,

1082 835–849.e8 (2022).
1083 20. Ryan, C. J., Kennedy, S., Bajrami, I., Matallanas, D. & Lord, C. J. A Compendium of Co-
1084 regulated Protein Complexes in Breast Cancer Reveals Collateral Loss Events. *Cell Syst* **5**,
1085 399–409.e5 (2017).
1086 21. Singh, G., Pratt, G., Yeo, G. W. & Moore, M. J. The Clothes Make the mRNA: Past and
1087 Present Trends in mRNP Fashion. *Annu. Rev. Biochem.* **84**, 325–354 (2015).
1088 22. Keene, J. D. & Tenenbaum, S. A. Eukaryotic mRNPs may represent posttranscriptional
1089 operons. *Mol. Cell* **9**, 1161–1167 (2002).
1090 23. Keene, J. D. RNA regulons: coordination of post-transcriptional events. *Nat. Rev. Genet.* **8**,
1091 533–543 (2007).
1092 24. Li, G.-W., Burkhardt, D., Gross, C. & Weissman, J. S. Quantifying absolute protein
1093 synthesis rates reveals principles underlying allocation of cellular resources. *Cell* **157**, 624–
1094 635 (2014).
1095 25. Taggart, J. C. & Li, G.-W. Production of Protein-Complex Components Is Stoichiometric
1096 and Lacks General Feedback Regulation in Eukaryotes. *Cell Syst* **7**, 580–589.e4 (2018).
1097 26. Amirbeigiarab, S. *et al.* Invariable stoichiometry of ribosomal proteins in mouse brain
1098 tissues with aging. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 22567–22572 (2019).
1099 27. Soto, I. *et al.* Balanced mitochondrial and cytosolic translatomes underlie the biogenesis of
1100 human respiratory complexes. *Genome Biol.* **23**, 170 (2022).
1101 28. Natan, E. *et al.* Cotranslational protein assembly imposes evolutionary constraints on
1102 homomeric proteins. *Nat. Struct. Mol. Biol.* **25**, 279–288 (2018).
1103 29. Li, G.-W., Oh, E. & Weissman, J. S. The anti-Shine-Dalgarno sequence drives translational

1104 pausing and codon choice in bacteria. *Nature* **484**, 538–541 (2012).
1105 30. Bertolini, M. *et al.* Interactions between nascent proteins translated by adjacent ribosomes
1106 drive homomer assembly. *Science* **371**, 57–64 (2021).
1107 31. Ozadam, H., Geng, M. & Cenik, C. RiboFlow, RiboR and RiboPy: an ecosystem for
1108 analyzing ribosome profiling data at read length resolution. *Bioinformatics* **36**, 2929–2931
1109 (2020).
1110 32. Gerashchenko, M. V. & Gladyshev, V. N. Ribonuclease selection for ribosome profiling.
1111 *Nucleic Acids Res.* **45**, e6 (2017).
1112 33. Mohammad, F., Green, R. & Buskirk, A. R. A systematically-revised ribosome profiling
1113 method for bacteria reveals pauses at single-codon resolution. *Elife* **8**, (2019).
1114 34. Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. S. & Weissman, J. S. Genome-wide
1115 analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science*
1116 **324**, 218–223 (2009).
1117 35. Larsson, O., Sonenberg, N. & Nadon, R. Identification of differential translation in genome
1118 wide studies. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 21487–21492 (2010).
1119 36. van den Boogaart, K. G., Filzmoser, P., Hron, K., Templ, M. & Tolosana-Delgado, R.
1120 Classical and Robust Regression Analysis with Compositional Data. *Math. Geosci.* **53**,
1121 823–858 (2021).
1122 37. Quinn, T. P. *et al.* A field guide for the compositional analysis of any-omics data.
1123 *Gigascience* **8**, (2019).
1124 38. Quinn, T. P., Richardson, M. F., Lovell, D. & Crowley, T. M. propr: An R-package for
1125 Identifying Proportionally Abundant Features Using Compositional Data Analysis. *Sci. Rep.*

1126 **7**, 16252 (2017).

1127 39. Sudmant, P. H., Alexis, M. S. & Burge, C. B. Meta-analysis of RNA-seq expression data
1128 across species, tissues and studies. *Genome Biol.* **16**, 287 (2015).

1129 40. Wang, Z.-Y. *et al.* Transcriptome and translatome co-evolution in mammals. *Nature* **588**,
1130 642–647 (2020).

1131 41. Lu, P., Takai, K., Weaver, V. M. & Werb, Z. Extracellular matrix degradation and
1132 remodeling in development and disease. *Cold Spring Harb. Perspect. Biol.* **3**, (2011).

1133 42. Artieri, C. G. & Fraser, H. B. Evolution at two levels of gene expression in yeast. *Genome*
1134 *Res.* **24**, 411–421 (2014).

1135 43. McManus, C. J., May, G. E., Spealman, P. & Shteyman, A. Ribosome profiling reveals
1136 post-transcriptional buffering of divergent gene expression in yeast. *Genome Res.* **24**, 422–
1137 430 (2014).

1138 44. Wang, M., Herrmann, C. J., Simonovic, M., Szklarczyk, D. & von Mering, C. Version 4.0
1139 of PaxDb: Protein abundance data, integrated across model organisms, tissues, and cell-
1140 lines. *Proteomics* **15**, 3163–3168 (2015).

1141 45. Breschi, A., Gingeras, T. R. & Guigó, R. Comparative transcriptomics in human and mouse.
1142 *Nat. Rev. Genet.* **18**, 425–440 (2017).

1143 46. Crow, M., Suresh, H., Lee, J. & Gillis, J. Coexpression reveals conserved gene programs
1144 that co-vary with cell type across kingdoms. *Nucleic Acids Res.* **50**, 4302–4314 (2022).

1145 47. Thoreen, C. C. *et al.* A unifying model for mTORC1-mediated regulation of mRNA
1146 translation. *Nature* **485**, 109–113 (2012).

1147 48. Wurth, L. *et al.* UNR/CSDE1 Drives a Post-transcriptional Program to Promote Melanoma

1148 Invasion and Metastasis. *Cancer Cell* **36**, 337 (2019).

1149 49. Pierson, E. *et al.* Sharing and Specificity of Co-expression Networks across 35 Human
1150 Tissues. *PLoS Comput. Biol.* **11**, e1004220 (2015).

1151 50. Kershaw, C. J. *et al.* Translation factor and RNA binding protein mRNA interactomes
1152 support broader RNA regulons for posttranscriptional control. *J. Biol. Chem.* **299**, 105195
1153 (2023).

1154 51. Hentze, M. W., Castello, A., Schwarzl, T. & Preiss, T. A brave new world of RNA-binding
1155 proteins. *Nat. Rev. Mol. Cell Biol.* **19**, 327–341 (2018).

1156 52. Korbel, J. O., Jensen, L. J., von Mering, C. & Bork, P. Analysis of genomic context:
1157 prediction of functional associations from conserved bidirectionally transcribed gene pairs.
1158 *Nat. Biotechnol.* **22**, 911–917 (2004).

1159 53. Szklarczyk, R. *et al.* WeGET: predicting new genes for molecular systems by weighted co-
1160 expression. *Nucleic Acids Res.* **44**, D567–73 (2016).

1161 54. Zhang, M. *et al.* RNA-binding protein IMP3 is a novel regulator of MEK1/ERK signaling
1162 pathway in the progression of colorectal Cancer through the stabilization of MEKK1
1163 mRNA. *J. Exp. Clin. Cancer Res.* **40**, 200 (2021).

1164 55. Bodén, M. & Bailey, T. L. Associating transcription factor-binding site motifs with target
1165 GO terms and target genes. *Nucleic Acids Res.* **36**, 4108–4117 (2008).

1166 56. Machanick, P. & Bailey, T. L. MEME-ChIP: motif analysis of large DNA datasets.
1167 *Bioinformatics* **27**, 1696–1697 (2011).

1168 57. Eichhorn, S. W. *et al.* mRNA destabilization is the dominant effect of mammalian
1169 microRNAs by the time substantial repression ensues. *Mol. Cell* **56**, 104–115 (2014).

1170 58. Bartel, D. P. Metazoan MicroRNAs. *Cell* **173**, 20–51 (2018).
1171 59. Benoit Bouvrette, L. P., Bovaird, S., Blanchette, M. & Lécuyer, E. oRNAment: a database
1172 of putative RNA binding protein target sites in the transcriptomes of model species. *Nucleic*
1173 *Acids Res.* **48**, D166–D173 (2020).
1174 60. Krismer, K. *et al.* Transite: A Computational Motif-Based Analysis Platform That Identifies
1175 RNA-Binding Proteins Modulating Changes in Gene Expression. *Cell Rep.* **32**, 108064
1176 (2020).
1177 61. Van Nostrand, E. L. *et al.* A large-scale binding and functional map of human RNA-binding
1178 proteins. *Nature* **583**, 711–719 (2020).
1179 62. Mecham, R. *The Extracellular Matrix: An Overview*. (Springer Science & Business Media,
1180 2011).
1181 63. Kagan, H. M. & Li, W. Lysyl oxidase: properties, specificity, and biological roles inside
1182 and outside of the cell. *J. Cell. Biochem.* **88**, 660–672 (2003).
1183 64. Kikuchi, A. *et al.* Structural basis for activation of DNMT1. *Nat. Commun.* **13**, 7130 (2022).
1184 65. Wu, Y.-Y. *et al.* The hTERT-p50 homodimer inhibits PLEKHA7 expression to promote
1185 gastric cancer invasion and metastasis. *Oncogene* **42**, 1144–1156 (2023).
1186 66. Kurita, S., Yamada, T., Rikitsu, E., Ikeda, W. & Takai, Y. Binding between the junctional
1187 proteins afadin and PLEKHA7 and implication in the formation of adherens junction in
1188 epithelial cells. *J. Biol. Chem.* **288**, 29356–29368 (2013).
1189 67. Pulimeno, P., Paschoud, S. & Citi, S. A role for ZO-1 and PLEKHA7 in recruiting
1190 paracingulin to tight and adherens junctions of epithelial cells. *J. Biol. Chem.* **286**, 16743–
1191 16750 (2011).

1192 68. Jeung, H.-C. *et al.* PLEKHA7 signaling is necessary for the growth of mutant KRAS driven
1193 colorectal cancer. *Exp. Cell Res.* **409**, 112930 (2021).
1194 69. Tavano, S. *et al.* Insm1 Induces Neural Progenitor Delamination in Developing Neocortex
1195 via Downregulation of the Adherens Junction Belt-Specific Protein Plekha7. *Neuron* **97**,
1196 1299–1314.e8 (2018).
1197 70. Sukonina, V. *et al.* FOXK1 and FOXK2 regulate aerobic glycolysis. *Nature* **566**, 279–283
1198 (2019).
1199 71. Kobe, B. & Kajava, A. V. The leucine-rich repeat as a protein recognition motif. *Curr.*
1200 *Opin. Struct. Biol.* **11**, 725–732 (2001).
1201 72. Evans, R. *et al.* Protein complex prediction with AlphaFold-Multimer. *bioRxiv*
1202 2021.10.04.463034 (2021) doi:10.1101/2021.10.04.463034.
1203 73. Carlsson, P. & Mahlapuu, M. Forkhead transcription factors: key players in development
1204 and metabolism. *Dev. Biol.* **250**, 1–23 (2002).
1205 74. The Human Transcription Factors. http://humantfs.ccbr.utoronto.ca/cite.php.
1206 75. Hou, Y., Xie, T., He, L., Tao, L. & Huang, J. Topological links in predicted protein
1207 complex structures reveal limitations of AlphaFold. *Commun Biol* **6**, 1098 (2023).
1208 76. Burke, D. F. *et al.* Towards a structurally resolved human protein interaction network. *Nat.*
1209 *Struct. Mol. Biol.* **30**, 216–225 (2023).
1210 77. Bryant, P., Pozzati, G. & Elofsson, A. Improved prediction of protein-protein interactions
1211 using AlphaFold2. *Nat. Commun.* **13**, 1265 (2022).
1212 78. Lambert, S. A. *et al.* The Human Transcription Factors. *Cell* **172**, 650–665 (2018).
1213 79. Kustatscher, G. *et al.* Co-regulation map of the human proteome enables identification of

1214 protein functions. *Nat. Biotechnol.* **37**, 1361–1371 (2019).

1215 80. Szklarczyk, D. *et al.* STRING v10: protein-protein interaction networks, integrated over the
1216 tree of life. *Nucleic Acids Res.* **43**, D447–52 (2015).

1217 81. Shiber, A. *et al.* Cotranslational assembly of protein complexes in eukaryotes revealed by
1218 ribosome profiling. *Nature* **561**, 268–272 (2018).

1219 82. Ewing, R. M. *et al.* Large-scale mapping of human protein–protein interactions by mass
1220 spectrometry. *Mol. Syst. Biol.* **3**, 89 (2007).

1221 83. Drew, K., Wallingford, J. B. & Marcotte, E. M. hu.MAP 2.0: integration of over 15,000
1222 proteomic experiments builds a global compendium of human multiprotein assemblies. *Mol.*
1223 *Syst. Biol.* **17**, e10016 (2021).

1224 84. Heider, M. R. *et al.* Subunit connectivity, assembly determinants and architecture of the
1225 yeast exocyst complex. *Nat. Struct. Mol. Biol.* **23**, 59–66 (2016).

1226 85. Kee, Y. *et al.* Subunit structure of the mammalian exocyst complex. *Proc. Natl. Acad. Sci.*
1227 *U. S. A.* **94**, 14438–14443 (1997).

1228 86. Lalanne, J.-B. *et al.* Evolutionary Convergence of Pathway-Specific Enzyme Expression
1229 Stoichiometry. *Cell* **173**, 749–761.e38 (2018).

1230 87. Bicknell, A. A. *et al.* Attenuating ribosome load improves protein output from mRNA by
1231 limiting translation-dependent mRNA decay. *Cell Rep.* **43**, 114098 (2024).

1232 88. Liu, T.-Y. *et al.* Time-Resolved Proteomics Extends Ribosome Profiling-Based
1233 Measurements of Protein Synthesis Dynamics. *Cell Syst* **4**, 636–644.e9 (2017).

1234 89. Piepoli, A. *et al.* The expression of leucine-rich repeat gene family members in colorectal
1235 cancer. *Exp. Biol. Med.* **237**, 1123–1128 (2012).

1236 90. Liu, Y. *et al.* Identification of differential expression of genes in hepatocellular carcinoma
1237 by suppression subtractive hybridization combined cDNA microarray. *Oncol. Rep.* **18**, 943–
1238 951 (2007).
1239 91. Chen, H. *et al.* miR-218 contributes to drug resistance in multiple myeloma via targeting
1240 LRRC28. *J. Cell. Biochem.* **122**, 305–314 (2021).
1241 92. Vander Heiden, M. G., Cantley, L. C. & Thompson, C. B. Understanding the Warburg
1242 effect: the metabolic requirements of cell proliferation. *Science* **324**, 1029–1033 (2009).
1243 93. Liu, Y. *et al.* Histone H2AX promotes metastatic progression by preserving glycolysis via
1244 hexokinase-2. *Sci. Rep.* **12**, 3758 (2022).
1245 94. Zheng, D. *et al.* Predicting the translation efficiency of messenger RNA in mammalian
1246 cells. *bioRxiv* 2024.08.11.607362 (2024) doi:10.1101/2024.08.11.607362.
1247 95. Rodriguez, J. M. *et al.* APPRIS: annotation of principal and alternative splice isoforms.
1248 *Nucleic Acids Res.* **41**, D110–7 (2013).
1249 96. Rao, S. *et al.* Genes with 5' terminal oligopyrimidine tracts preferentially escape global
1250 suppression of translation by the SARS-CoV-2 Nsp1 protein. *RNA* **27**, 1025–1045 (2021).
1251 97. Mills, E. W. & Green, R. Ribosomopathies: There's strength in numbers. *Science* **358**,
1252 (2017).
1253 98. Ozadam, H. *et al.* Single-cell quantification of ribosome occupancy in early mouse
1254 development. *Nature* **618**, 1057–1064 (2023).
1255 99. VanInsberghe, M., van den Berg, J., Andersson-Rolf, A., Clevers, H. & van Oudenaarden,
1256 A. Single-cell Ribo-seq reveals cell cycle-dependent translational pausing. *Nature* **597**,
1257 561–565 (2021).

1258 100. ChatGPT. https://chat.openai.com.

1259 101. *Sra-Tools: SRA Tools*. (Github).

1260 102. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads.
1261 *EMBnet.journal* **17**, 10–12 (2011).

1262 103. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**,
1263 357–359 (2012).

1264 104. Creators Liu, O. *HeLa Ribosome Profiling Data*. doi:10.5281/zenodo.10594392.

1265 105. Gerashchenko, M. V. & Gladyshev, V. N. Translation inhibitors cause abnormalities in
1266 ribosome profiling experiments. *Nucleic Acids Res.* **42**, e134 (2014).

1267 106. Wu, C. C.-C., Zinshteyn, B., Wehner, K. A. & Green, R. High-Resolution Ribosome
1268 Profiling Defines Discrete Ribosome Elongation States and Translational Regulation during
1269 Cellular Stress. *Mol. Cell* **73**, 959–970.e5 (2019).

1270 107. Wolin, S. L. & Walter, P. Ribosome pausing and stacking during translation of a eukaryotic
1271 mRNA. *EMBO J.* **7**, 3559–3569 (1988).

1272 108. Sharma, J. *et al.* A small molecule that induces translational readthrough of CFTR nonsense
1273 mutations by eRF1 depletion. *Nat. Commun.* **12**, 4358 (2021).

1274 109. Tukey, J. W. The Future of Data Analysis. *Ann. Math. Stat.* **33**, 1–67 (1962).

1275 110. Zhang, X.-O., Yin, Q.-F., Chen, L.-L. & Yang, L. Gene expression profiling of non-
1276 polyadenylated RNA-seq across species. *Genom Data* **2**, 237–241 (2014).

1277 111. Yang, L., Duff, M. O., Graveley, B. R., Carmichael, G. G. & Chen, L.-L. Genomewide
1278 characterization of non-polyadenylated RNAs. *Genome Biol.* **12**, R16 (2011).

1279 112. van den Boogaart, K. G. & Tolosana-Delgado, R. *Analyzing Compositional Data with R*.

1280 (Springer Berlin Heidelberg).
1281 113. Cenik, C. *et al.* Integrative analysis of RNA, translation, and protein levels reveals distinct
1282 regulatory variation across humans. *Genome Res.* **25**, 1610–1621 (2015).
1283 114. Greenacre, M. Compositional data analysis. *Annu. Rev. Stat. Appl.* **8**, 271–299 (2021).
1284 115. Ramsköld, D., Wang, E. T., Burge, C. B. & Sandberg, R. An abundance of ubiquitously
1285 expressed genes revealed by tissue transcriptome sequence data. *PLoS Comput. Biol.* **5**,
1286 e1000598 (2009).
1287 116. Csárdi, G., Franks, A., Choi, D. S., Airoldi, E. M. & Drummond, D. A. Accounting for
1288 experimental noise reveals that mRNA levels, amplified by post-transcriptional processes,
1289 largely determine steady-state protein levels in yeast. *PLoS Genet.* **11**, e1005206 (2015).
1290 117. orthogene. *Bioconductor*
1291 https://bioconductor.org/packages/release/bioc/html/orthogene.html.
1292 118. van den Boogaart, K. G. & Tolosana-Delgado, R. 'compositions': A unified R package to
1293 analyze compositional data. *Comput. Geosci.* **34**, 320–338 (2008).
1294 119. Kim, S. ppcor: An R Package for a Fast Calculation to Semi-partial Correlation
1295 Coefficients. *Commun Stat Appl Methods* **22**, 665–674 (2015).
1296 120. Berriz, G. F., Beaver, J. E., Cenik, C., Tasan, M. & Roth, F. P. Next generation software for
1297 functional trend analysis. *Bioinformatics* **25**, 3043–3044 (2009).
1298 121. Buttrey, S. & Whitaker, L. TreeClust: An R package for tree-based clustering
1299 dissimilarities. *R J.* **7**, 227 (2015).
1300 122. Wainberg, M. *et al.* A genome-wide atlas of co-essential modules assigns function to
1301 uncharacterized genes. *Nat. Genet.* **53**, 638–649 (2021).

1302 123. Gene Ontology Consortium *et al.* The Gene Ontology knowledgebase in 2023. *Genetics*
1303 **224**, (2023).
1304 124. Philippe, L., van den Elzen, A. M. G., Watson, M. J. & Thoreen, C. C. Global analysis of
1305 LARP1 translation targets reveals tunable and dynamic features of 5′ TOP motifs.
1306 *Proceedings of the National Academy of Sciences* **117**, 5319–5328 (2020).
1307 125. Ballouz, S., Weber, M., Pavlidis, P. & Gillis, J. EGAD: ultra-fast functional analysis of
1308 gene networks. *Bioinformatics* **33**, 612–614 (2017).
1309 126. Carlson, M. org. Mm. eg. db: Genome wide annotation for Mouse. R package version 3.8.
1310 2. 2019.
1311 127. Carlson, M. org. Hs. eg. db: Genome wide annotation for Human. R package version 3.8. 2.
1312 2019.
1313 128. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**,
1314 583–589 (2021).
1315 129. UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids*
1316 *Res.* **51**, D523–D531 (2023).
1317 130. Hu, Y. *et al.* Paralog Explorer: A resource for mining information about paralogs in
1318 common research organisms. *Comput. Struct. Biotechnol. J.* **20**, 6570–6577 (2022).
1319 131. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic
1320 features. *Bioinformatics* **26**, 841–842 (2010).
1321 132. Stuart, E. A., King, G., Imai, K. & Ho, D. MatchIt: nonparametric preprocessing for
1322 parametric causal inference. *J. Stat. Softw.* (2011).
1323 133. Sanson, K. R. *et al.* Optimized libraries for CRISPR-Cas9 genetic screens with multiple

1324 modalities. *Nat. Commun.* **9**, 5416 (2018).
1325 134. Sanjana, N. E., Shalem, O. & Zhang, F. Improved vectors and genome-wide libraries for
1326 CRISPR screening. *Nat. Methods* **11**, 783–784 (2014).
1327 135. Rossi, G. P. *et al.* Endothelin-1 stimulates steroid secretion of human adrenocortical cells ex
1328 vivo via both ETA and ETB receptor subtypes. *J. Clin. Endocrinol. Metab.* **82**, 3445–3449
1329 (1997).
1330 136. Sánchez-Caballero, L. *et al.* TMEM70 functions in the assembly of complexes I and V.
1331 *Biochim. Biophys. Acta Bioenerg.* **1861**, 148202 (2020).
1332 137. Carroll, J., He, J., Ding, S., Fearnley, I. M. & Walker, J. E. TMEM70 and TMEM242 help
1333 to assemble the rotor ring of human ATP synthase and interact with assembly factors for
1334 complex I. *Proc. Natl. Acad. Sci. U. S. A.* **118**, (2021).
1335 138. Mii, Y. & Takada, S. Heparan Sulfate Proteoglycan Clustering in Wnt Signaling and
1336 Dispersal. *Front Cell Dev Biol* **8**, 631 (2020).
1337 139. Kamimura, K. *et al.* Perlecan regulates bidirectional Wnt signaling at the Drosophila
1338 neuromuscular junction. *J. Cell Biol.* **200**, 219–233 (2013).
1339 140. Camlin, N. J., McLaughlin, E. A. & Holt, J. E. Kif4 Is Essential for Mouse Oocyte Meiosis.
1340 *PLoS One* **12**, e0170650 (2017).
1341 141. Tang, F. *et al.* Involvement of Kif4a in Spindle Formation and Chromosome Segregation in
1342 Mouse Oocytes. *Aging Dis.* **9**, 623–633 (2018).
1343 142. Robertson, A. K., Geiman, T. M., Sankpal, U. T., Hager, G. L. & Robertson, K. D. Effects
1344 of chromatin structure on the enzymatic and DNA binding functions of DNA
1345 methyltransferases DNMT1 and Dnmt3a in vitro. *Biochem. Biophys. Res. Commun.* **322**,

1346 110–118 (2004).
1347 143. Schrader, A., Gross, T., Thalhammer, V. & Längst, G. Characterization of Dnmt1 Binding
1348 and DNA Methylation on Nucleosomes and Nucleosomal Arrays. *PLoS One* **10**, e0140076
1349 (2015).
1350 144. Ciossani, G. *et al.* The kinetochore proteins CENP-E and CENP-F directly and specifically
1351 interact with distinct BUB mitotic checkpoint Ser/Thr kinases. *J. Biol. Chem.* **293**, 10084–
1352 10101 (2018).
1353 145. Liao, H., Winkfein, R. J., Mack, G., Rattner, J. B. & Yen, T. J. CENP-F is a protein of the
1354 nuclear matrix that assembles onto kinetochores at late G2 and is rapidly degraded after
1355 mitosis. *J. Cell Biol.* **130**, 507–518 (1995).
1356 146. Chen, M. *et al.* FAT1 inhibits the proliferation and metastasis of cervical cancer cells by
1357 binding β-catenin. *Int. J. Clin. Exp. Pathol.* **12**, 3807–3818 (2019).
1358 147. Nishikawa, Y. *et al.* Human FAT1 cadherin controls cell migration and invasion of oral
1359 squamous cell carcinoma through the localization of β-catenin. *Oncol. Rep.* **26**, 587–592
1360 (2011).
1361 148. Morris, L. G. T. *et al.* Recurrent somatic mutation of FAT1 in multiple human cancers leads
1362 to aberrant Wnt activation. *Nat. Genet.* **45**, 253–261 (2013).
1363 149. Hou, R., Liu, L., Anees, S., Hiroyasu, S. & Sibinga, N. E. S. The Fat1 cadherin integrates
1364 vascular smooth muscle cell growth and migration signals. *J. Cell Biol.* **173**, 417–429
1365 (2006).
1366 150. Vallet, S. D., Berthollier, C., Salza, R., Muller, L. & Ricard-Blum, S. The Interactome of
1367 Cancer-Related Lysyl Oxidase and Lysyl Oxidase-Like Proteins. *Cancers* **13**, (2020).

1368 151. Vallet, S. D. *et al.* Insights into the structure and dynamics of lysyl oxidase propeptide, a
1369 flexible protein with numerous partners. *Sci. Rep.* **8**, 11768 (2018).
1370 152. Yang, C. *et al.* Transcriptomic Analysis Identified ARHGAP Family as a Novel Biomarker
1371 Associated With Tumor-Promoting Immune Infiltration and Nanomechanical
1372 Characteristics in Bladder Cancer. *Front Cell Dev Biol* **9**, 657219 (2021).
1373 153. Lamarche-Vane, N. & Hall, A. CdGAP, a novel proline-rich GTPase-activating protein for
1374 Cdc42 and Rac. *J. Biol. Chem.* **273**, 29172–29177 (1998).
1375 154. Yang, S. *et al.* Control of antiviral innate immune response by protein geranylgeranylation.
1376 *Sci Adv* **5**, eaav7999 (2019).
1377 155. Bouhaddou, M. *et al.* The Global Phosphorylation Landscape of SARS-CoV-2 Infection.
1378 *Cell* **182**, 685–712.e19 (2020).
1379 156. Swaine, T. & Dittmar, M. T. CDC42 Use in Viral Cell Entry Processes by RNA Viruses.
1380 *Viruses* **7**, 6526–6536 (2015).
1381 157. Redmond, S. A. *et al.* Somatodendritic Expression of JAM2 Inhibits Oligodendrocyte
1382 Myelination. *Neuron* **91**, 824–836 (2016).
1383 158. Skoufos, G. *et al.* TarBase-v9.0 extends experimentally supported miRNA-gene interactions
1384 to cell-types and virally encoded miRNAs. *Nucleic Acids Res.* **52**, D304–D310 (2024).
1385 159. Miettinen, T. P. & Björklund, M. Modified ribosome profiling reveals high abundance of
1386 ribosome protected mRNA fragments derived from 3' untranslated regions. *Nucleic Acids*
1387 *Res.* **43**, 1019–1034 (2015).
1388 160. Hsieh, A. C. *et al.* The translational landscape of mTOR signalling steers cancer initiation
1389 and metastasis. *Nature* **485**, 55–61 (2012).

1390 161. Song, K. Y., Choi, H. S., Law, P.-Y., Wei, L.-N. & Loh, H. H. Vimentin interacts with the
1391 5'-untranslated region of mouse mu opioid receptor (MOR) and is required for post-
1392 transcriptional regulation. *RNA Biol.* **10**, 256–266 (2013).
1393 162. van der Brug, M. P. *et al.* RNA binding activity of the recessive parkinsonism protein DJ-1
1394 supports involvement in multiple cellular pathways. *Proc. Natl. Acad. Sci. U. S. A.* **105**,
1395 10244–10249 (2008).
1396 163. Niere, F. *et al.* Aberrant DJ-1 expression underlies L-type calcium channel hypoactivity in
1397 dendrites in tuberous sclerosis complex and Alzheimer's disease. *Proc. Natl. Acad. Sci. U.*
1398 *S. A.* **120**, e2301534120 (2023).
1399 164. Jin, W. *et al.* HydRA: Deep-learning models for predicting RNA-binding capacity from
1400 protein interaction association context and protein sequence. *Mol. Cell* **83**, 2595–2611.e11
1401 (2023).
1402 165. Avsec, Ž. *et al.* Effective gene expression prediction from sequence by integrating long-
1403 range interactions. *Nat. Methods* **18**, 1196–1203 (2021).
1404 166. Agarwal, V. & Shendure, J. Predicting mRNA abundance directly from genomic sequence
1405 using deep convolutional neural networks. *Cell Rep.* **31**, 107663 (2020).
1406 167. Dong, X., Li, C., Chen, Y., Ding, G. & Li, Y. Human transcriptional interactome of
1407 chromatin contribute to gene co-expression. *BMC Genomics* **11**, 704 (2010).
1408 168. Zhang, K., Wang, C., Sun, L. & Zheng, J. Prediction of gene co-expression from chromatin
1409 contacts with graph attention network. *Bioinformatics* **38**, 4457–4465 (2022).
1410 169. Blum, M. *et al.* The InterPro protein families and domains database: 20 years on. *Nucleic*
1411 *Acids Res.* **49**, D344–D354 (2021).

1412 170. Cock, P. J. A. *et al.* Biopython: freely available Python tools for computational molecular
1413 biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).

1414

1415 **Table 1: Literature support for gene functions predicted using TEC.**

1416 In the table, we list the predictions that are supported by literature. To do the new gene function
1417 prediction, we selected GO terms with AUROC measured with TEC >=0.8, then focused on the
1418 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	[135]
GO:0022900	Human	electron transport chain	TMEM70	0.82	0.67	1611	[136,137]
GO:0042813	Human	Wnt-activated receptor activity	HSPG2	0.85	0.74	341	[138,139]
GO:0007129	Human	homologous chromosome pairing at meiosis	KIF4A	0.84	0.73	11	[140,141]
GO:0031492	Human	nucleosomal DNA binding	DNMT1	0.85	0.73	251	[64,142,143]
GO:0050793	Mouse	regulation of developmental	Plekha7	0.85	0.70	3929	[65–69]

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	[144,145]
GO:0016342	Mouse	catenin complex	Fat1	0.85	0.73	781	[146–149]
GO:0005539	Mouse	glycosaminoglycan binding	Lox	0.95	0.83	32	[150,151]
GO:0140374	Mouse	antiviral innate immune response	Arhgap31	0.84	0.73	1085	[152–156]
GO:0022010	Mouse	Central nervous system myelination	Jam2	0.88	0.76	584	[157]

1419