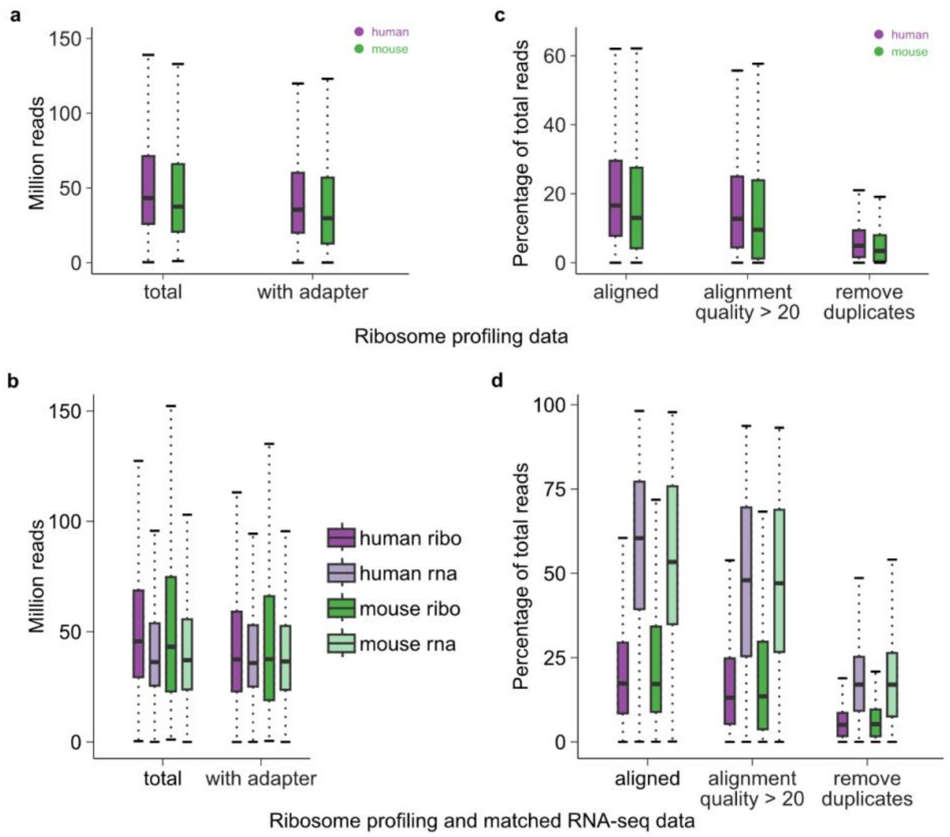
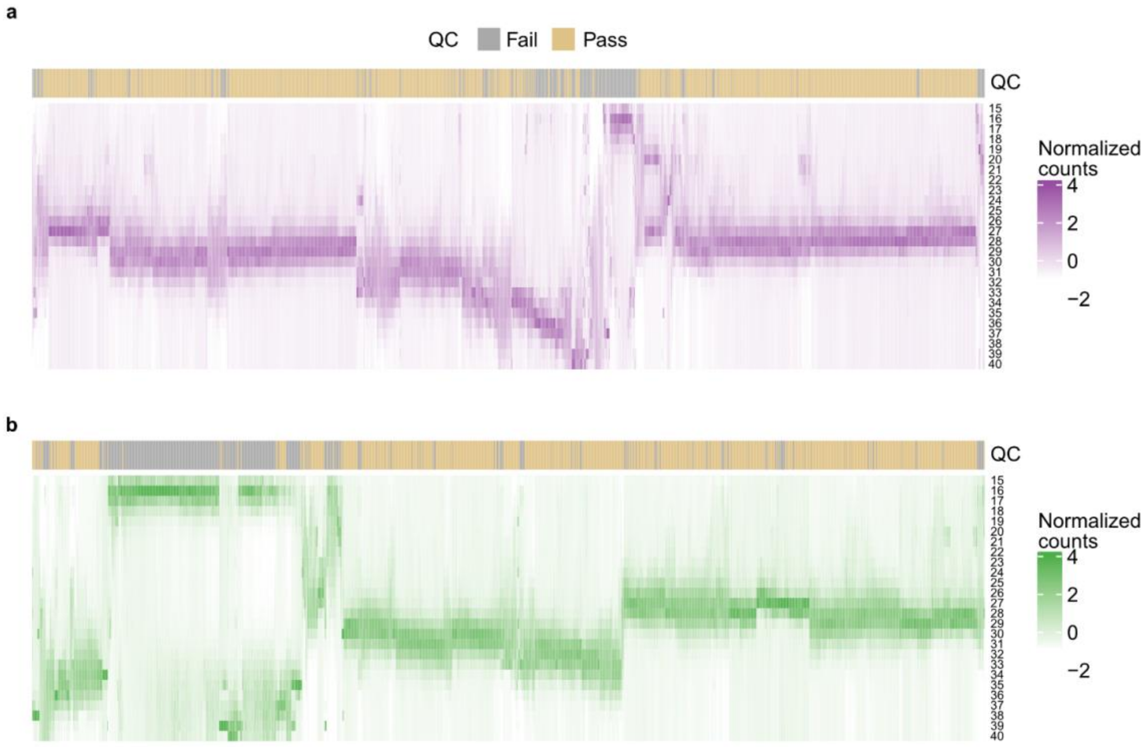
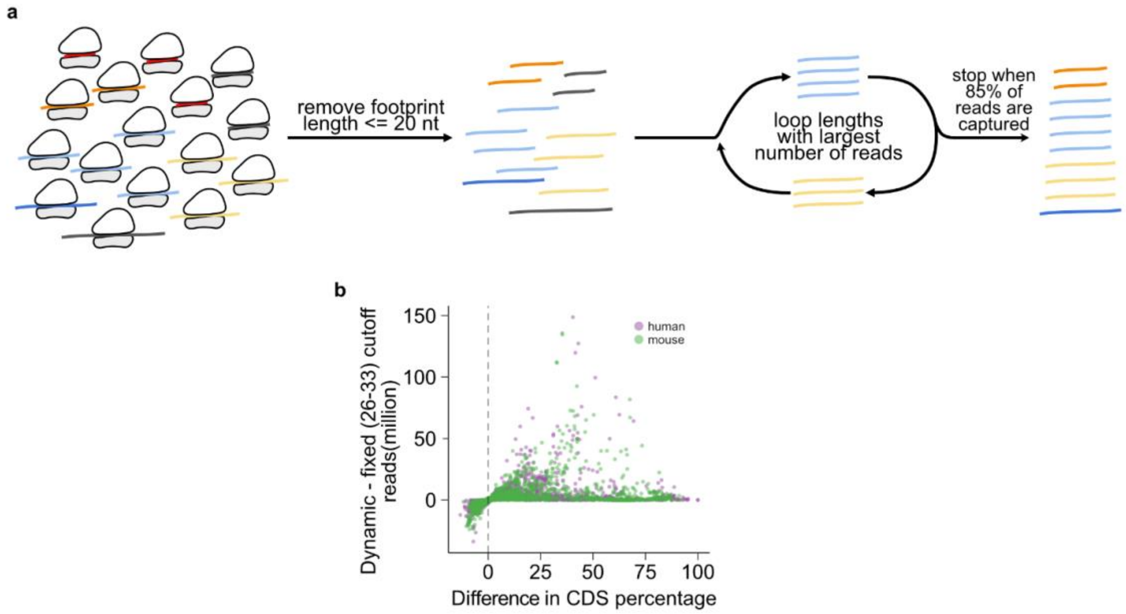
Extended Figures



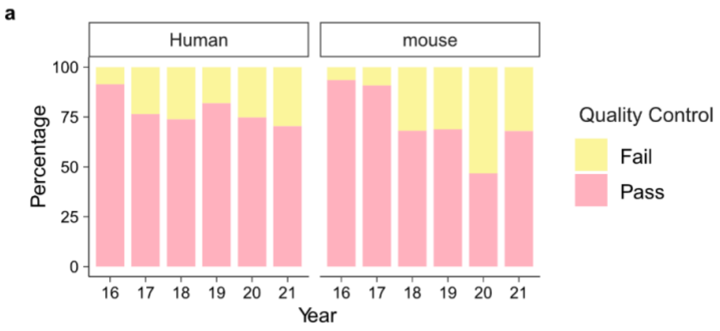
ExtendedDataFig. 1 | Sequencing quality of ribosome profiling data with matched RNA-seq data when available (supplementary text): a, Distribution of read counts for ribosome profiling data in RiboBase. In all figure panels, the horizontal line corresponds to the median. The box represents the interquartile range and the whiskers extend to 1.5 times of it. **b,** Distribution plot similar to panel A for ribosome profiling data with matched RNA-seq. **c,** Distribution of the proportion of read count aligned to transcripts, read counts with high-quality alignments, and the percentage of reads remaining after PCR deduplication, relative to the total number of reads from panel A. **d,** Similar plot as panel C for ribosome profiling with matched RNA-seq.



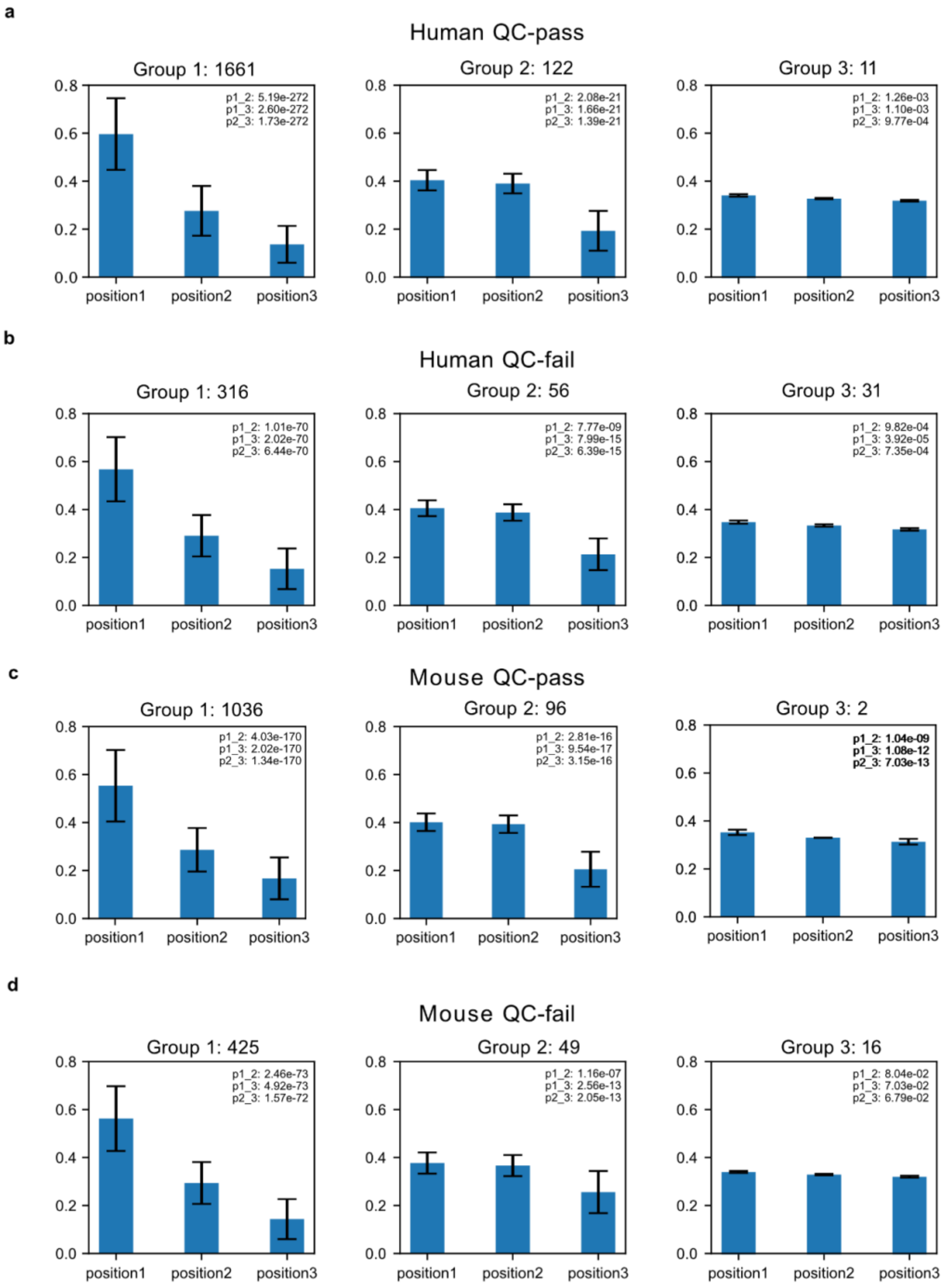
ExtendedDataFig. 2 | Length distribution of RPFs for human and mouse samples: a, The read length distribution of RPFs aligned to coding sequences for all human experiments. The color in the heatmap represents the z-score adjusted RPF counts (Methods). Each experiment where the percentage of RPFs mapping to CDS was greater than 70% and achieving sufficient coverage of the transcript (>= 0.1X) was annotated as QC-pass. **b,** Similar to panel A for mouse samples.



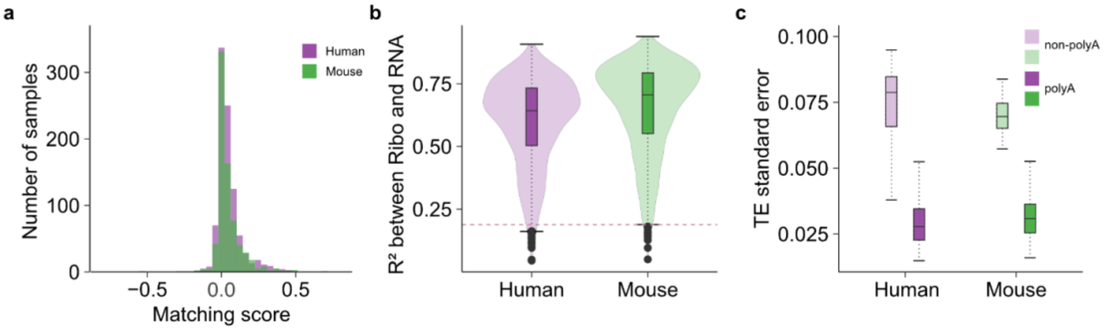
ExtendedDataFig. 3 | Schematic for method to select range of RPF lengths: a, RPFs shorter than 21 nucleotides were removed, then we identified the RPF length with the highest number of reads mapping to CDS to serve as the starting point. Subsequently, we compared one nucleotide longer or shorter than the first and chose the length with the most reads again. This looping process continued until at least 85% of the total CDS mapping RPFs were included. **b,** We compared the usable reads selected with two different boundary cutoffs (y-axis) and the proportion of these selected reads that map to the coding regions (x-axis) for each ribosome profiling experiment.



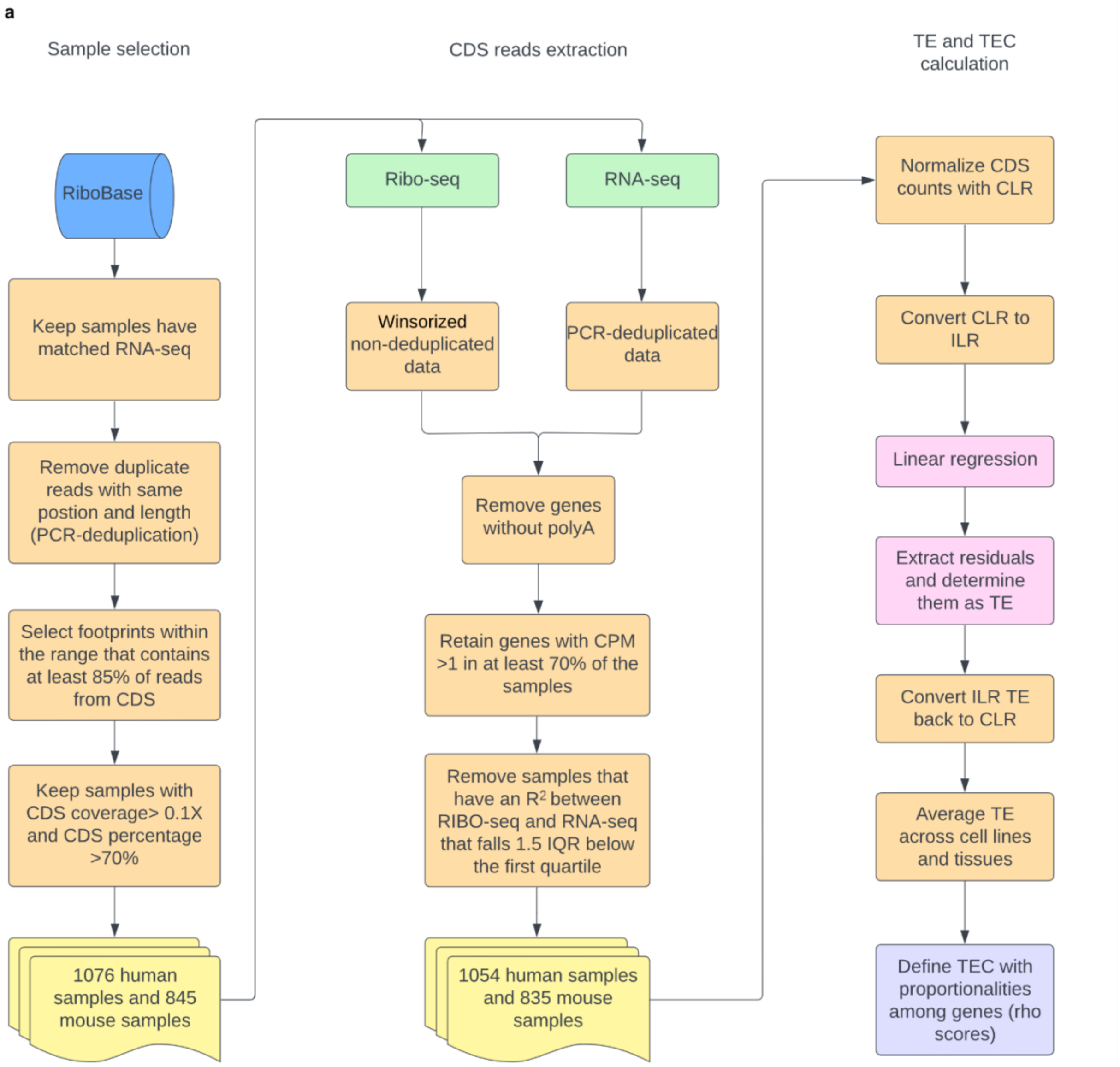
ExtendedDataFig. 4 | Data quality of ribosome profiling experiments from 2016 to 2021: a, The percentage of ribosome profiling experiments from GEO that pass or fail quality control (the percentage of RPFs mapping to CDS was greater than 70% and achieving at least 0.1X coverage of the transcript as QC pass).



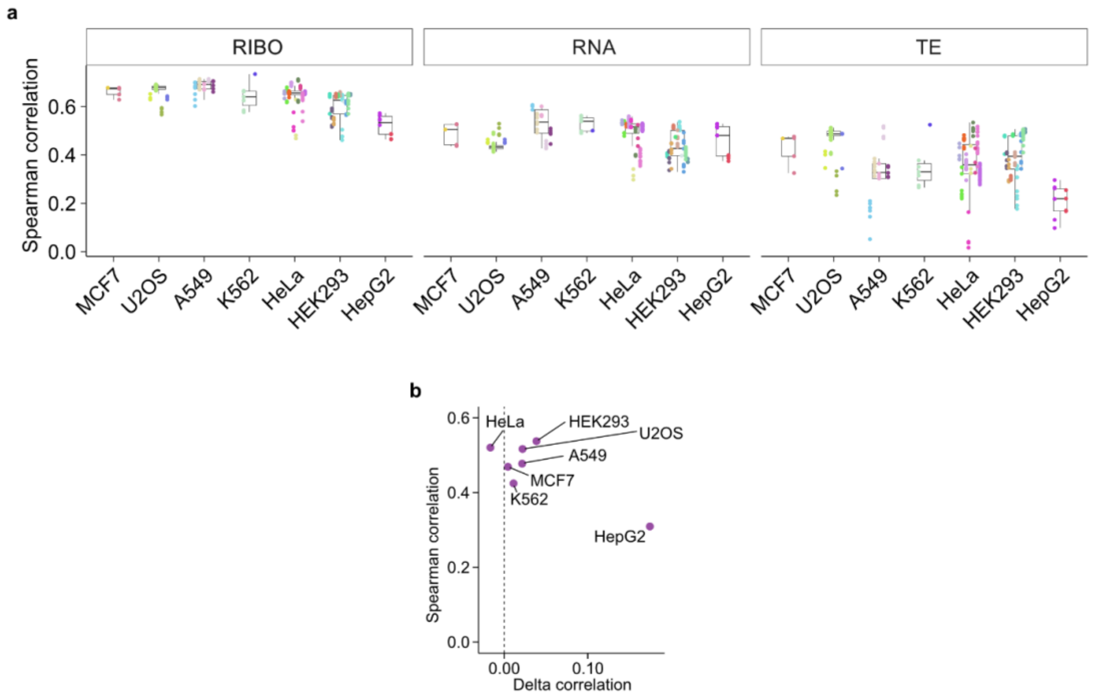
ExtendedDataFig. 5 | Three nucleotide periodicity of ribosome profiling data: a-d, In ribosome profiling experiments from RiboBase, samples were classified according to distinct periodicity patterns (Methods). For all figure panels, we added error bars to represent the standard deviation across samples. Statistical significance was assessed using the Wilcoxon test, and the p-values were subsequently adjusted for all 33 comparisons using the Benjamini-Hochberg method. We considered the Group 1 pattern as indicative of the expected three nucleotide periodicity patterns.



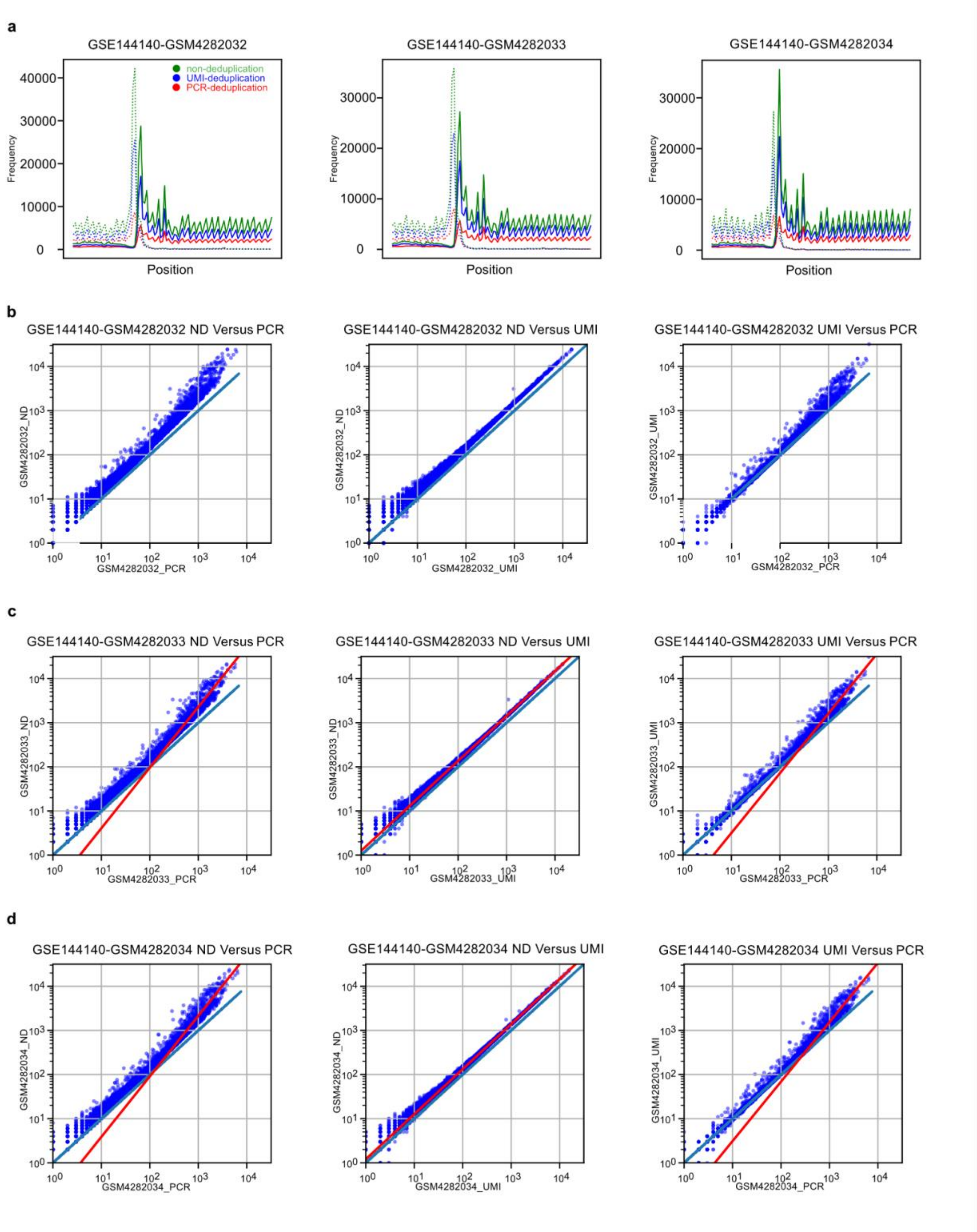
ExtendedDataFig. 6 | Validation of ribosome profiling and RNA-seq matching and gene selection for TE calculation: a, We calculated the coefficient of determination (R^2) between a specific ribosome profiling experiment and its corresponding RNA-seq from RiboBase. Additionally, we determined the average R^2 for all other pairings for the same ribosome profiling sample with other RNA-seq data from the same study. The matching score represents the difference in R^2 values between these two (x-axis; Methods). **b,** A dashed line at 0.188 serves as the threshold to identify samples with poor matching. In each figure panel containing boxplots, the horizontal line corresponds to the median. The box represents the IQR and the whiskers extend to 1.5 times of it. **c,** Distribution of standard error of TE values across tissue and cell lines (y-axis) for genes with polyA and without polyA tails.


a
Sample selection
CDS reads extraction
TE and TEC calculation
RiboBase
Keep samples have matched RNA-seq
Remove duplicate reads with same postion and length (PCR-deduplication)
Select footprints within the range that contains at least 85% of reads from CDS
Keep samples with CDS coverage> 0.1X and CDS percentage >70%
1076 human samples and 845 mouse samples
Ribo-seq
RNA-seq
Winsorized non-deduplicated data
PCR-deduplicated data
Remove genes without polyA
Retain genes with CPM >1 in at least 70% of the samples
Remove samples that have an R² between RIBO-seq and RNA-seq that falls 1.5 IQR below the first quartile
1054 human samples and 835 mouse samples
Normalize CDS counts with CLR
Convert CLR to ILR
Linear regression
Extract residuals and determine them as TE
Convert ILR TE back to CLR
Average TE across cell lines and tissues
Define TEC with proportionalities among genes (rho scores)

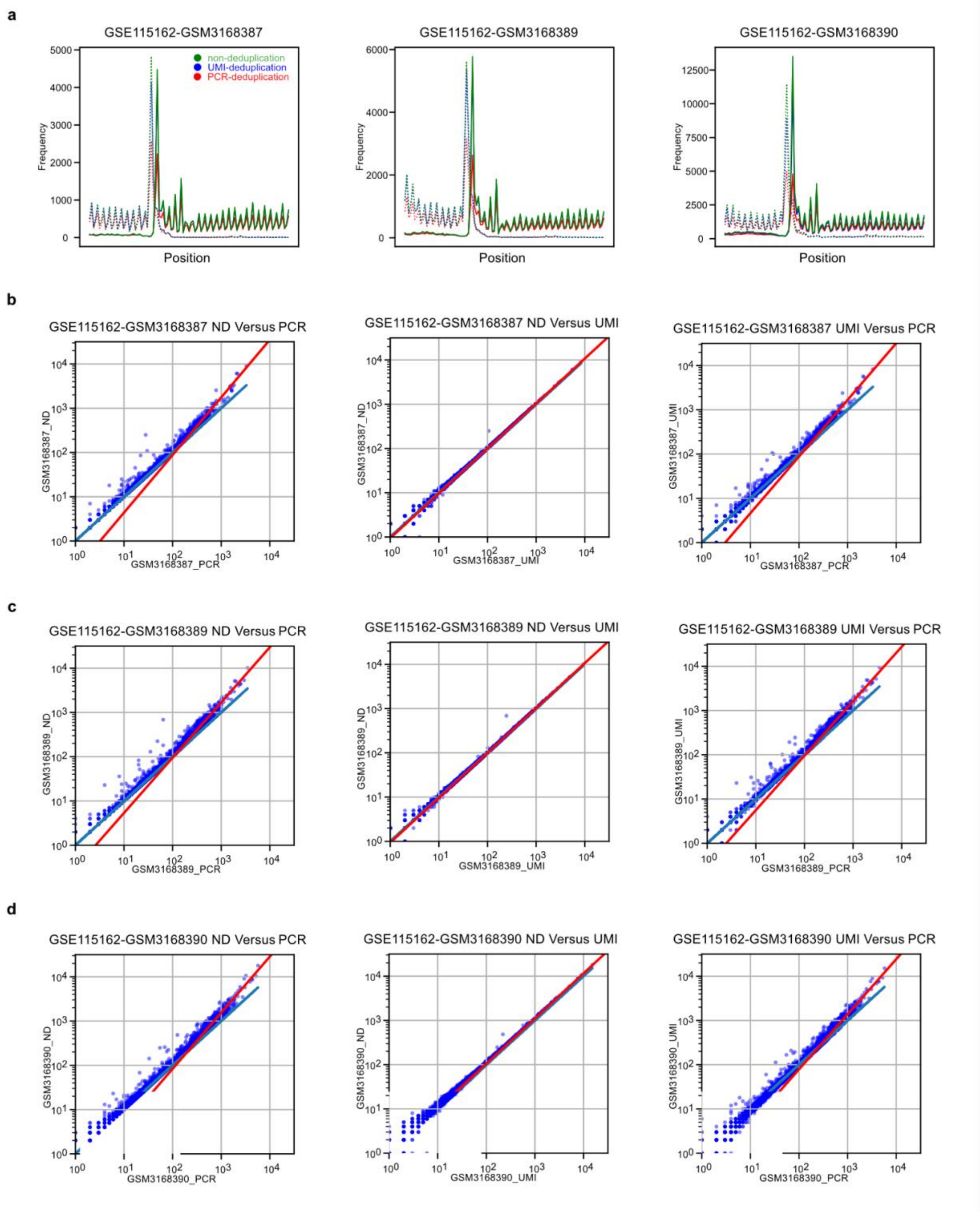
ExtendedDataFig. 7 | Detailed workflow of data processing for TE and TEC calculations: a, We selected ribosome profiling data with matched RNA-seq and removed duplicated reads with identical positions and lengths (PCR-deduplication). We set the RPF read length range for individual samples with our dynamic cutoff and filtered out ribosome profiling experiments that failed quality control. After selecting high-quality samples, we reprocessed all these ribosome profiling experiments using the winsorization method with non-deduplicated data. We removed genes without polyA tails and kept genes with sufficient counts per million RPFs. After obtaining RPF counts from the coding regions for both ribosome profiling and RNA-seq, we performed CLR normalization and compositional linear regression, defining the residuals as TE for each gene in each sample. We averaged this sample-level TE based on cell lines and tissues. TEC is further calculated with rho scores[50]. To build an RNA co-expression matrix, we transformed CDS counts from RNA-seq experiments using CLR, averaged them based on cell lines and tissue, and calculated pairwise proportionalities (rho scores).



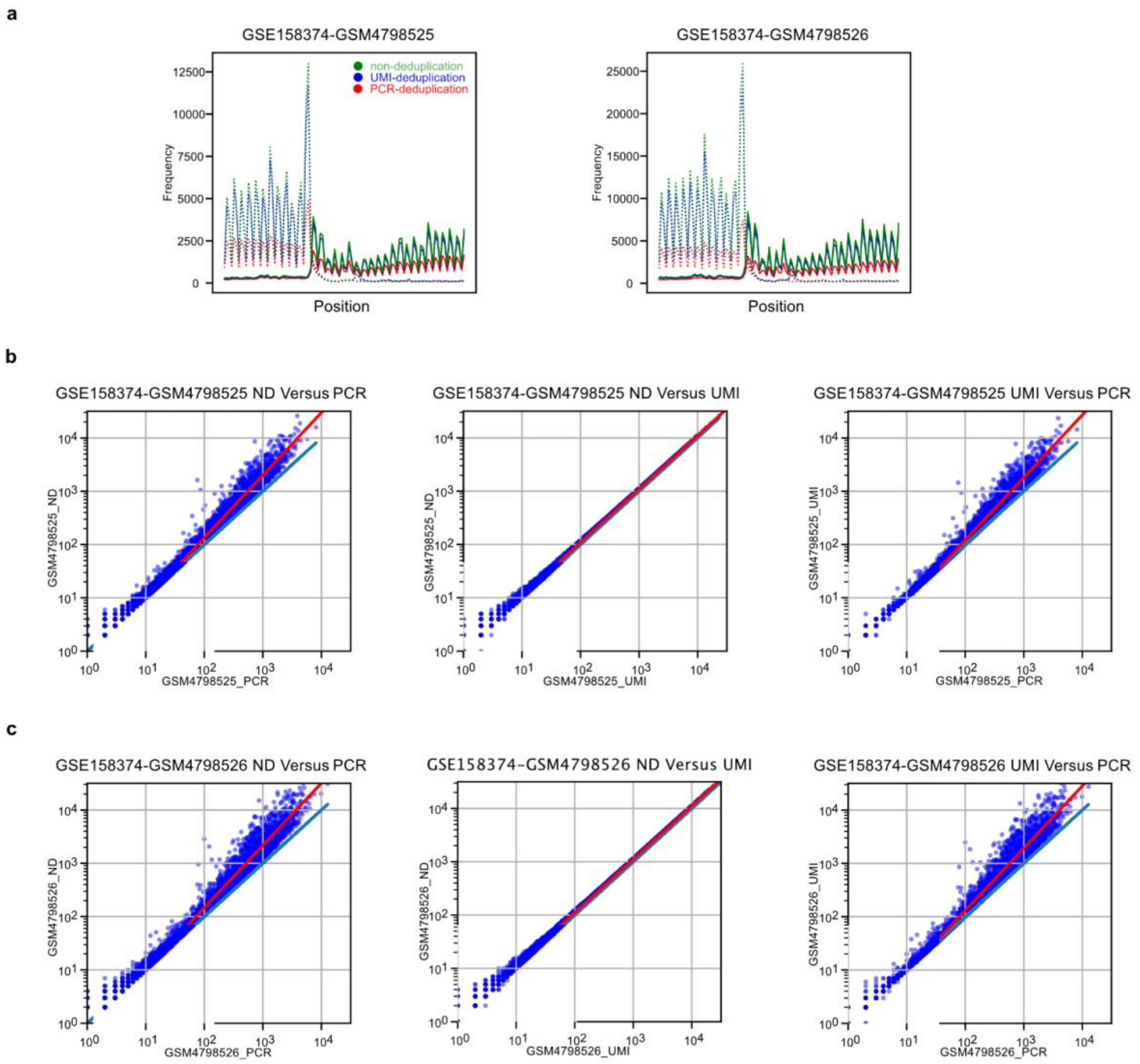
ExtendedDataFig. 8 | Spearman correlation between TE and protein abundance: a, The correlation between protein abundance and clr-transformed RPF counts from ribosome profiling (left), clr-transformed read counts from RNA-seq (middle), or TE calculated with winsorized RPFs counts using the linear regression model (right). Individual dots indicate specific experiments colored according to study. In the boxplot, the horizontal line corresponds to the median. The box represents the IQR and the whiskers extend to 1.5 times of this range. **b,** TE was calculated with winsorized RPF counts without deduplication or with deduplication based on position and fragment length. The Spearman correlation coefficient between TE calculated with winsorized RPF counts and protein abundance[100] (y-axis) was plotted against “delta correlation” (x-axis) defined by subtracting the correlation values obtained with PCR deduplication from those obtained with the method using winsorized RPF counts without deduplication.



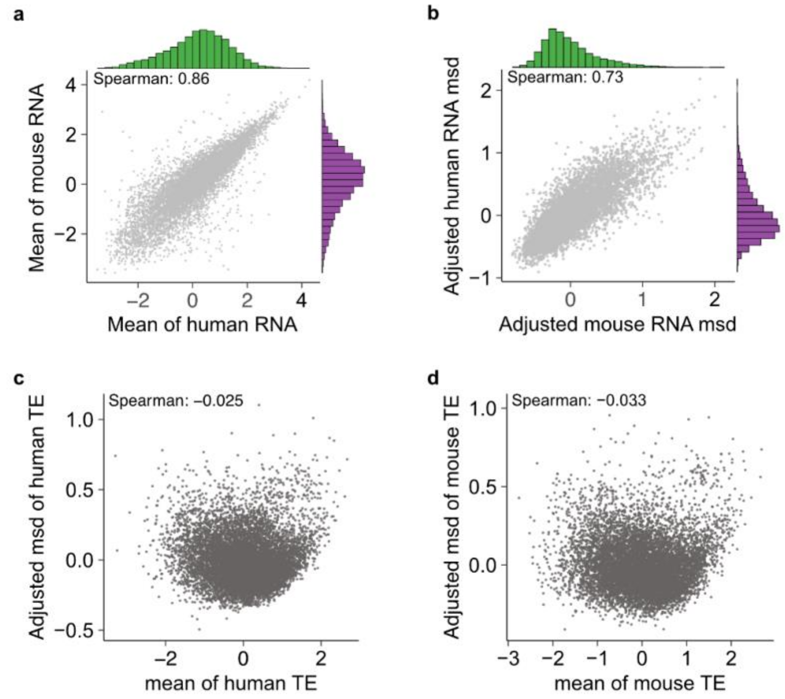
ExtendedDataFig. 9 | PCR vs. UMI deduplication comparison for GSE144140: a, Metagene plots centered on the start codon for samples GSM4282032 (RPFs range: 28-36 nt), GSM4282033 (RPFs range: 28-36 nt range), and GSM4282034 (RPFs range: 26-35 nt range) were plotted using three different deduplication methods: non-deduplication (ND), UMI-deduplication (UMI), and PCR-deduplication (PCR). **b,** Correlation of gene counts for GSM4282032 between the three deduplication methods. A blue diagonal line represents a 1:1 ratio in all figure panels. Same analysis as panel B for GSM4282033 **c,** and GSM4282034 **d**.



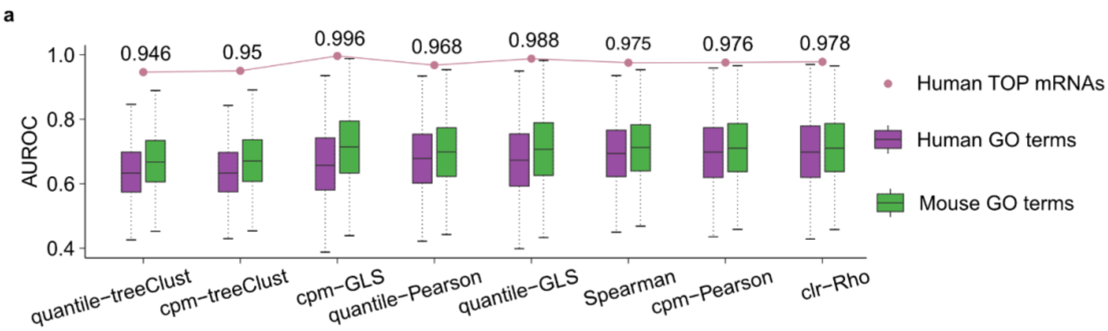
ExtendedDataFig. 10 | PCR vs. UMI deduplication comparison for GSE115162: Similar analysis as ExtendedDataFig. 7 for GSM3168387 (RPFs range: 24-34 nt), GSM3168389 (RPFs range: 23-33 nt), and GSM3168390 (RPFs range: 23-35 nt).



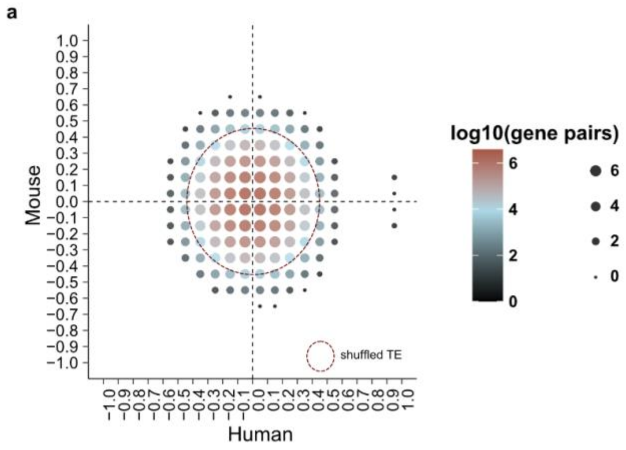
ExtendedDataFig. 11 | PCR vs. UMI deduplication comparison for GSE158374: Similar analysis as figure S7 and S8 for GSM4798525 and GSM4798526, both in the 28-32 nt RPFs range.



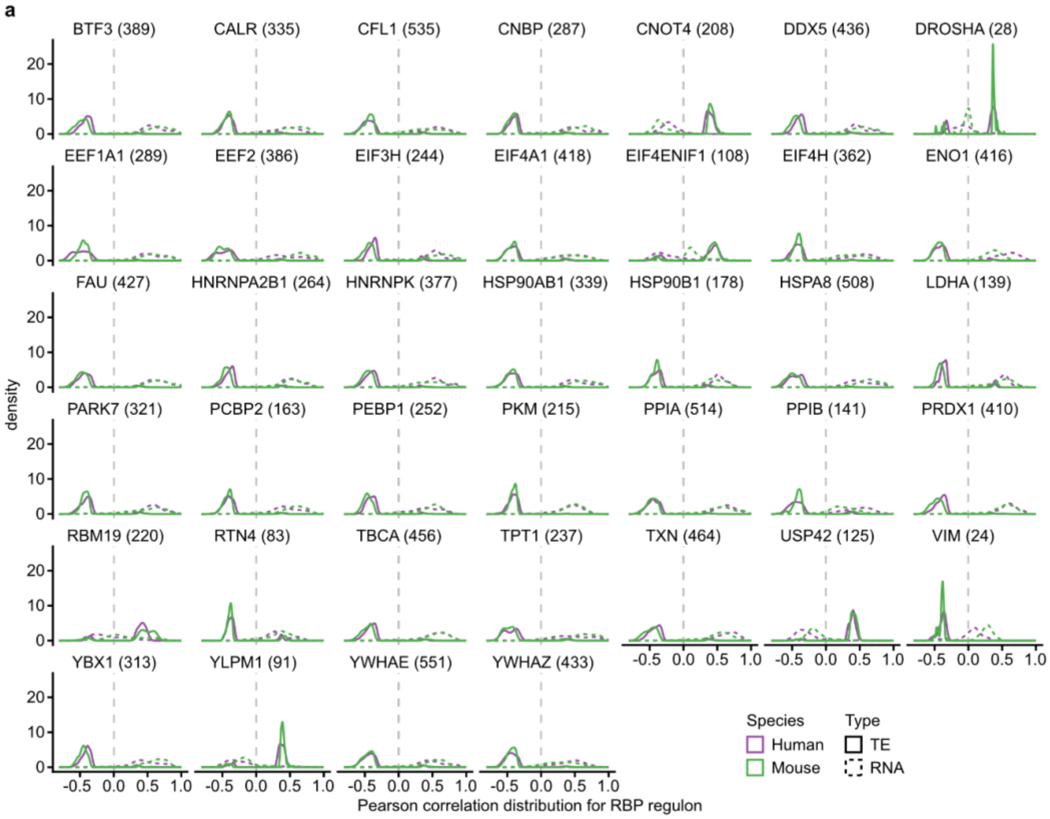
ExtendedDataFig. 12 | Conservation of gene expression between human and mouse: a, The relationship between the mean RNA expressions (clr-transformed counts) of 9,194 orthologous genes across two species is plotted. Dots represent genes in all figure panels. **b,** The variability of genes' RNA expression was quantified with metric standard deviation (msd; Methods) across different cell lines and tissues in either human or mouse. To account for the correlation between mean RNA expression and its variability, we adjusted the msd values with their mean values (Methods). **c,** The scatter plot shows the adjusted msd values (y-axis; Methods) and the average TE across different cell types (x-axis) for human genes. **d,** Similar analysis as in panel c for mouse genes.



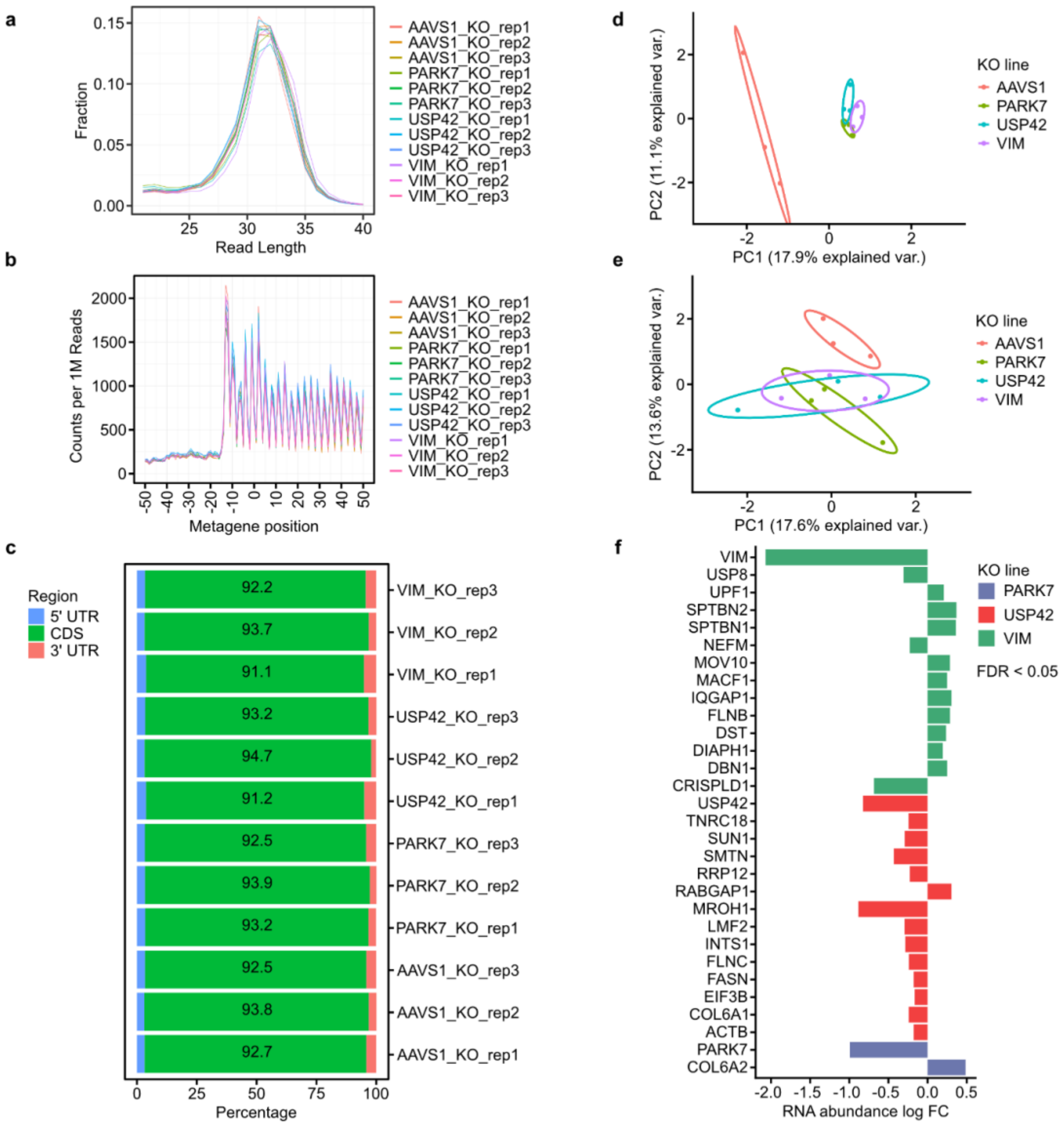
ExtendedDataFig. 13 | Evaluating the performance of eight methods to associate ribosome occupancy covariation with biological function: a, The AUROCs for biological functions were calculated using the similarity scores among genes at ribosome occupancy level determined by eight distinct methods (Methods). In the boxplot, 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 from the hinge. The dot in this figure represents the AUROC for human 5' TOP mRNAs.



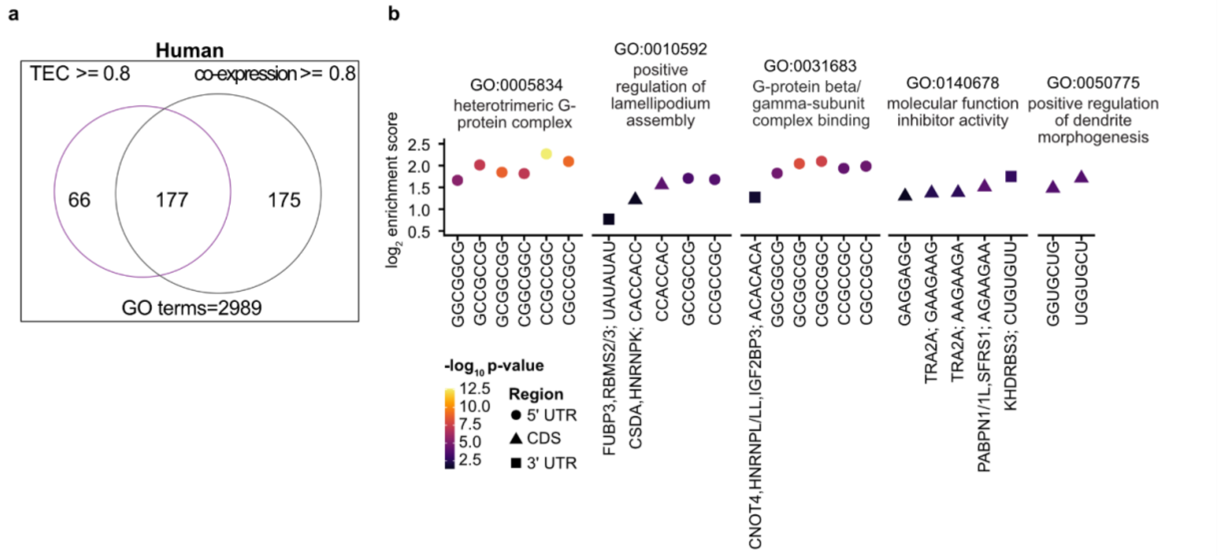
ExtendedDataFig. 14 | Lack of correlation in TEC across orthologous gene pairs between human and mouse using shuffled TE: a, TE values that were randomly reassigned from the original data for each gene (shuffled) and TEC was calculated. In the figure panel, we plotted the number of orthologous gene pairs within specified ranges. Each dot represents the aggregated $\log_{10}$-transformed counts of these gene pairs. The dashed line captures 95% of the data.



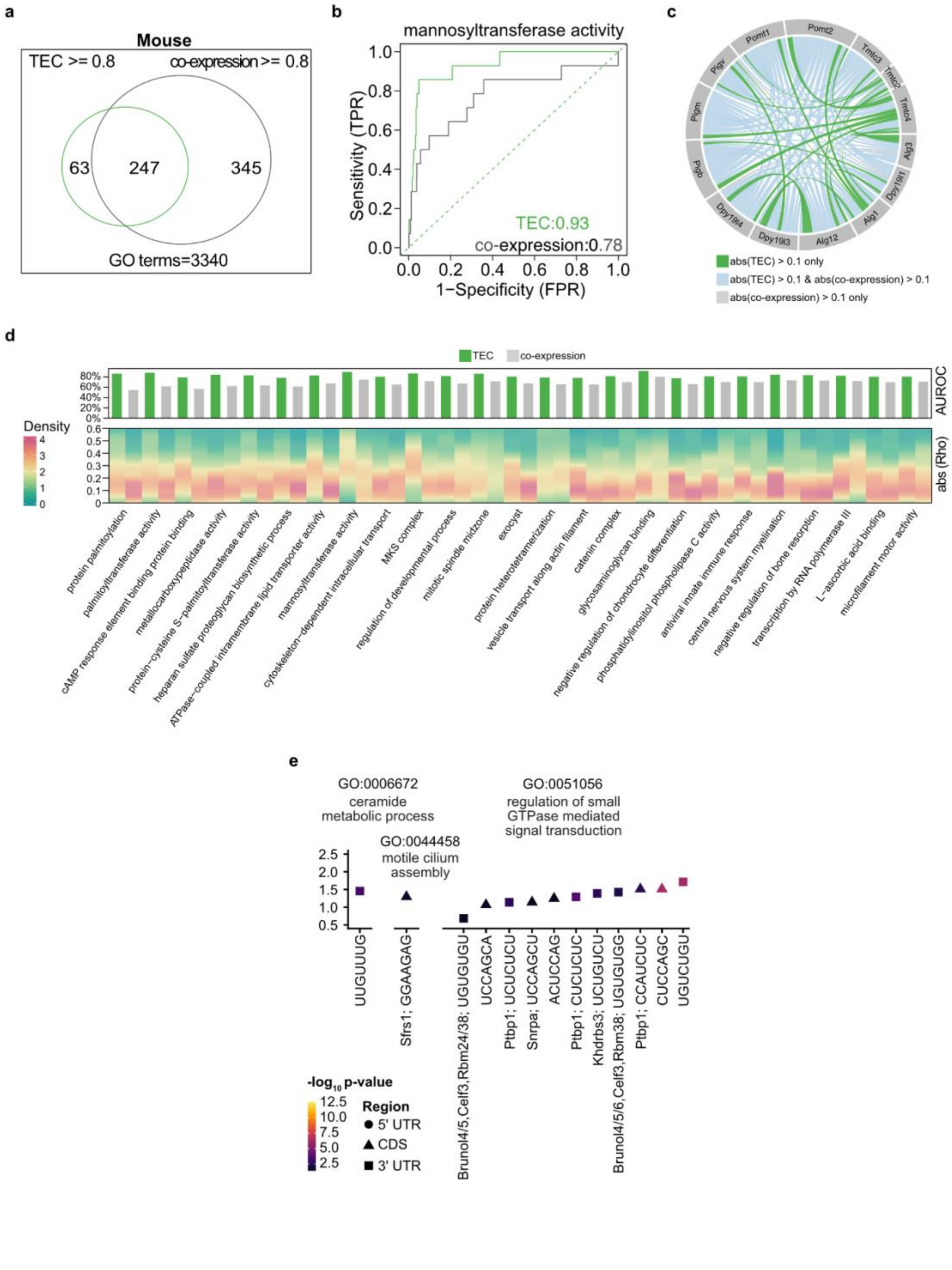
ExtendedDataFig. 15 | RBP regulon correlation distribution for regulons with high TEC: a, Distribution of Pearson correlation coefficients between RBP RNA expression and TE of the conserved regulon are shown for RBP regulons with mean abs(TE rho) > 90th percentile. Ribosomal protein genes are omitted except for FAU, as a representative example. Numbers in parentheses denote the number of genes in the conserved RBP regulon.



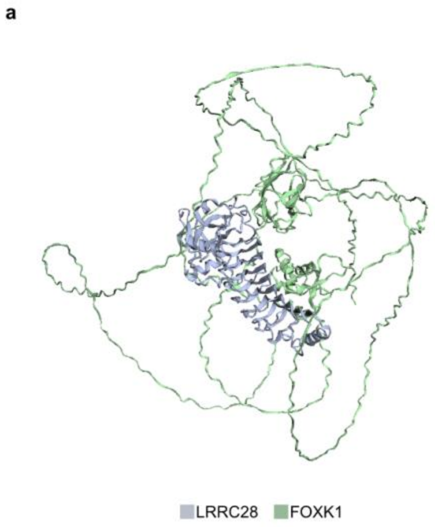
ExtendedDataFig. 16 | Ribosome profiling and RNA-seq of RBP KO cell lines: For b through e, ribosome footprints between 28 and 35 nt were used. **a,** Read length distributions of ribosome footprints. **b,** Metagene plot at the start site. **c,** Location of mapped ribosome footprints. **d,** PCA was performed on standardized counts per million (CPM) reads for transcripts whose sum of CPMs across cell lines and replicates is in the top 80th percentile. PCA of RNA-seq counts. **e,** Same as D for ribosome profiling read counts. **f,** Differential RNA expression of KO cell lines. A significance threshold of FDR < 0.05 was used.



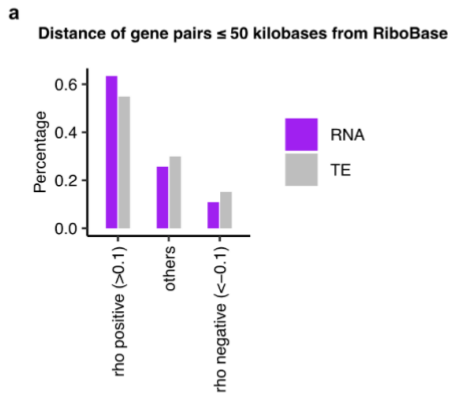
ExtendedDataFig. 17 | TEC and RNA co-expression among genes with shared functions in human: a, A comparison between the number of human GO terms that have AUROC of 0.8 or higher with either TEC or RNA co-expression. **b,** Motif enrichment in human GO terms. RNA binding proteins (RBPs) from oRNAment[134] or Transite[133] are indicated. P-values were corrected using the Holm method and those kmers with a p-value < 0.05 are shown.



ExtendedDataFig. 18 | TEC and RNA co-expression among genes with shared functions in mouse: a, Venn diagram for mouse GO terms that achieve an AUROC of 0.8 or higher with proportionality scores (rho) among genes at either TE or RNA expression level. **b,** The AUROC plot was calculated with genes associated with mannosyltransferase activity in mice. **c,** The connections represent absolute rho values above 0.1 in either TE pattern alone (green), in both RNA co-expression and TE pattern (blue), or RNA co-expression alone (gray). **d,** We summarized GO terms where genes exhibit greater similarity at the TE level than at the RNA expression level (AUROC with TEC > 0.8, and different AUROC between TEC and RNA co-expression > 0.1) in mice. We visualized the distribution of absolute rho score for gene pairs within each specific GO term (bottom; gene pairs with abs(rho) > 0.1) at the TE level. **e,** Motif enrichment in mouse GO terms. RNA binding proteins (RBPs) from oRNAment[134] or Transite[133] are indicated. P-values were corrected using the Holm method and those kmers with a p-value < 0.05 are shown.



ExtendedDataFig. 19 | 3D structure of the interaction between LRRC28 with FOXK1: a, AlphaFold2-multimer predicted binding between LRRC28 and FOXK1.



ExtendedDataFig. 20 | Rho scores enrichment of gene pairs with a distance of less than 50 kilobases on the same chromosome: a, Rho scores enrichment for 5,999 human gene pairs with a distance of less than 50 kilobases at either RNA expression or TE level.

Supplementary Text

1: Inaccurate and incomplete metadata examples from GEO

We observed recurrent issues regarding cell line identification in GEO. For instance, several studies categorize cell lines merely as “erythroid cells” without providing specifics. Similarly, descriptions of mouse embryonic stem cells (mESCs) often lack detail regarding subtypes, such as v6.5, which are either vague or missing. Inconsistencies in library strategies present additional challenges. While most researchers categorize ribosome profiling under OTHER in the library strategy, some entries label ribosome profiling as RNA-seq, ncRNA-seq, and miRNA-seq. Such nonstandard information lead to significant errors in large-scale data reanalysis, emphasizing the need for data curation.

2: Summary of sequencing quality for ribosome profiling and matched RNA-seq

The median number of reads for all human ribosome profiling samples was approximately 43.2 million, and after removing the adapter sequences the corresponding median was 35.5 million (ExtendedDataFig. 1a; table S2). For mouse samples, the median number of reads was around 37.5 million, compared to 29.7 million reads with adapters (ExtendedDataFig. 1a; table S3). On average, only 17% of reads could be aligned to the transcriptome, with 13% having a mapping quality higher than 20 in human samples (ExtendedDataFig. 1c). After removing duplicate reads with the same position and length (PCR-deduplication), 5% of the total ribosome profiling reads were retained (ExtendedDataFig. 1c). The mouse data showed a similar trend, with 13% alignment, 10% above a mapping quality of 20, and 3% retention after PCR-deduplication (ExtendedDataFig. 1c).

Furthermore, in our comparative analysis between ribosome profiling and the corresponding RNA-seq data, we observed that ribosome profiling experiments were generally sequenced at a higher depth compared to RNA-seq. The median reads for ribosome profiling were 45.6 million for human experiments and 43.1 million for mouse experiments, compared to 36.2 million and 37.1 million reads for the matched RNA-seq, respectively (ExtendedDataFig. 1b; table S4-5). However, ribosome profiling demonstrated a lower alignment percentage to transcriptome than RNA-seq, with only 13% in human and 14% in mouse experiments, as opposed to 48% and 47% in RNA-seq for human and mouse samples, respectively (ExtendedDataFig. 1d). This discrepancy is explained by the substantial presence of ribosomal RNA in ribosome profiling samples.

3: Comparison of different methods for removing duplicated reads

Removing duplicated reads with the same position and length is commonly used in sequencing data processing to minimize biases introduced by PCR amplification. A key concern is the inadvertent removal of ribosome footprints that are identical in sequence and length but originate from different templates, leading to misinterpretation of the data. To evaluate the impact of deduplication strategies on ribosome profiling data, we analyzed samples that incorporated unique molecular identifiers (UMIs). Our findings indicate a significant loss of reads originating from unique molecules when using PCR deduplication based on position and fragment length (ExtendedDataFig. 9-11). This discrepancy was exacerbated in samples with higher coverage.

Given the limited adoption of library preparation method that introduce UMIs in ribosome profiling experiments, we used a winsorizing-based method to process non-deduplicated ribosome profiling sequencing data, aiming to mitigate this bias by capping excessively high-depth regions (Methods).

We compared linear regression-based TE calculated by winsorized non-deduplicated and PCR-deduplicated data. The winsorized method showed a slightly higher mean correlation than the PCR-deduplicated method (ExtendedDataFig. 8b), indicating the PCR-based deduplication approach, which relies on identical position and length, could obscure the actual biological insights obtainable from ribosome profiling.

4: RBPs may coordinate TEC

We identified RBP regulons in which the component genes had high TEC, as this might indicate a direct influence of the RBP on TE. Salient examples of RBPs which were previously linked to translation regulation and had conserved regulons with high TEC were VIM and PARK7 (also known as DJ-1). Although VIM is a primary component of intermediate filaments, its RNA expression was negatively correlated with TE of genes encoding proteins in the electron transport chain and ribosomal proteins. VIM was previously found to repress translation of the mu opioid receptor[164]. Similarly, expression of *PARK7* was predominantly negatively correlated with TE, in line with a prior study that PARK7 represses translation[165,166] (68% and 79% of regulon genes having negative correlations in human and mouse, respectively). There was not a significant overlap between PARK7 targets determined by RIP-seq analysis in human neuroblastoma cells[165] and the human or mouse PARK7 regulons (hypergeometric test p-values 0.84 and 0.23, respectively). Nevertheless, thirty-three PARK7 RIP-seq targets were present in both human and mouse regulons, including glutathione peroxidase 4 (*GPX4*), Sm-like proteins (*LSM1/3/5*), and six genes encoding ubiquinone-oxidoreductase subunits, indicating PARK7 regulates a diverse set of biological processes extending beyond the oxidative stress response. Among genes with positive correlations with PARK7 expression, subunits of calcium channels such as CACNB1 and CACNA2D1 were notable, consistent with data that PARK7 increases nascent protein synthesis of CACNA2D1 despite not significantly binding it[166]. Altogether, these data suggest largely indirect influences of PARK7 on TE, and a smaller set of direct target genes.

We selected VIM, PARK7, and USP42 for further experiments, as their regulons exhibited distinct correlation distributions for RNA expression and gene TE (ExtendedDataFig. 15) and are not essential genes, facilitating knockout experiments. These RBPs had high HydRA[167] scores (>0.89, scale 0 to 1) and detectable RNA binding domains, supporting their role as *bona fide* RBPs. Surprisingly, knockout of these RBPs and subsequent matched ribosome profiling and RNA-seq (ExtendedDataFig. 16a-e, Methods) indicated no changes in TE for the genes in these RBPs' regulons, with one exception (*VIM* KO led to lower *VIM* TE). However, we found a small subset of genes with altered RNA abundance upon knockout of each RBP (ExtendedDataFig. 16f). For example, knockout of *VIM* led to increased RNA abundance of several genes involved in cytoskeletal function, including *SPTBN1*, *SPTBN2*, *MACF1*, *IQGAP1*, *FLNB*, *DST*, *DIAPH1*, and *DBN1*.

We note that the lack of genes with significantly altered TE upon KO of these RBPs may be due to several reasons: 1) these RBPs exert an indirect- rather than direct- influence on TE; 2) the associations between RBP expression and gene TE were identified across diverse cell lines, whereas the association was only tested in the HEK293T cell line; 3) use of heterogenous knockout populations (not single clones), and limited efficiency of knockout as measured by the observed RNA-seq fold changes (*PARK7*: 0.37, *USP42*: 0.44, *VIM*: 0.13) may limit sensitivity to observe effects on TE. Further work will be needed to validate the role of PARK7, USP42, and VIM on translational regulation.

5: TEC among genes is associated with shared biological functions in mouse

We identified 25 GO terms including protein palmitoylation, palmitoyltransferase activity, and metallocarboxypeptidase activity which exhibited AUROC scores that were at least 0.1 lower at the RNA expression level compared to the TE level (AUROC calculated with TEC > 0.8; ExtendedDataFig. 18). For example, mannosyltransferase activity demonstrated a significant difference between the two levels (ExtendedDataFig. 18c). This difference was further highlighted by the observation that 22 gene pairs in this biological function had absolute rho above 0.1 exclusively at the TE level, compared to only two at the RNA expression level for this term (ExtendedDataFig. 18d). In summary, we found genes from certain biological functions are more likely to be regulated at the translational level rather than the transcriptional level, in both humans and mice.

We predicted novel functions for genes associated with 31 mouse GO terms. These predictions are based on either significant covariation in TE greater than RNA expression (AUROC measured with TEC > 0.8; different AUROC measured with TEC and RNA co-expression > 0.1; table S16; Methods) or new functional predictions were only achievable with TEC (AUROC measured with TEC > 0.8, difference AUROC measured with TEC and RNA co-expression < 0.1, ranking of the predicted gene with RNA co-expression < top 50%; table S16; Methods). For instance, we identified Cenpf as highly correlated with the function of mitotic spindle midzone. This aligned with findings in human cell lines, where CENP-F has been observed assembling onto kinetochores at late G2 and detected at the spindle midzone during anaphase[151]. Another prediction linked Arhgap31 with the antiviral innate immune response. This prediction has been supported by previous research that has recognized the ARHGAP family as novel biomarkers associated with immune infiltration[158].