preprocessing-snakemake/

A custom snakemake pipeline was used to process RNA-SPRITE sequencing data. Briefly, data quality was assessed with fastqc (v0.12.1), barcodes were identified, and sequences without full barcodes were discarded (BarcodeIdentification_v1.2.0)[70]. MarkDuplicates (Picard) was used to remove PCR duplicates by filtering out similar sequences with the same barcode. Sequencing adapters and barcodes were removed before doing two pass RNAseq alignment. As described above, we performed paired-end sequencing with most of one read corresponding to the cluster barcode, therefore, alignments were based on approximately 150-200 nucleotides of sequencing. To avoid small, highly abundant RNAs spuriously aligning to mRNAs, we first aligned all sequences using Bowtie2 (v2.4.1) to a custom transcriptome with highly abundant non-coding RNAs (ncRNAs): snoRNAs, snRNAs, and rRNAs. We then took the unaligned reads from the first pass alignment and aligned them to a repeat-masked genome (GRCh38 or mRatBN7.2) using STAR (v2.7.11).

negative_binomial_regression/
1_rawcounts makeRDS negbinom position_bins.R

Further data processing and analysis was done using custom R scripts. Briefly, Ensembl transcript ID alignments were aggregated together by their associated Ensembl gene ID, and all analyses were performed at the level gene ID. The Ensembl transcript ID sequence with the best coverage of alignments was chosen as the one representative sequence for each Ensembl gene ID. The gene IDs in each dataset were filtered to be above a minimum number of observations (unique barcodes) cutoff. We did not allow for homotypic colocalization; each barcode could only map to a gene once. We made a gene by cluster matrix only using clusters mapping to 2-20 genes of interest. For cytoplasmic mRNA analyses, we removed all ncRNAs, mRNAs encoded by mitochondrial DNA, and all mRNAs in clusters that also had alignments to known nuclear RNAs (snRNAs, snoRNAs, scaRNAs, pre-rRNA, and RMRP).

Negative binomial regression

Motivated by its successful use in single-cell RNA-sequencing data analysis[75], we applied negative binomial regression (NBR, formally a generalized linear model, GLM) to model pairwise counts. Given the scale of the data, we developed a custom R package for fitting (https://github.com/davidaknowles/adaglm). Briefly, adaglm fits a NBR using Adam, an algorithm for stochastic gradient descent optimization. Since Adam supports minibatching, adaglm can run on very large datasets (tens of millions of data points). By default the fitting is done in two steps: 1) Fit a Poisson regression. 2) Starting with coefficients initialized from the Poisson regression, fit a NBR jointly optimizing the coefficients and dispersion parameter. This two step procedure avoids inefficient optimization dynamics that can result from attempting to learn the regression coefficients and the dispersion parameter simultaneously.

NBR models the observed colocalization count y_{ab}, between RNAs *a* and *b*, as negative binomially distributed with expectation μ_{ab} and variance $var(y_{ab}) = \mu_{ab} + \gamma\mu_{ab}^2$ where γ is the *dispersion* parameter which is shared across all pairs (note for $\gamma = 0$ the negative binomial is equivalent to a Poisson). In the baseline model, the expectation is determined by covariates X_{ab} according to,

$$\ln \mu_{ab} = X_{ab}\beta = \beta_0 + \beta_1 \ln(P_{ab}) + \sum_i \beta_i d_i^{ab} \quad (1)$$

where β are learned regression coefficients and P_{ab} is the preliminarily expected colocalization count between the RNA pair calculated as $P_{ab} = y_{a\bullet} y_{\bullet b} / y_{\bullet\bullet}$ with $y_{a\bullet}$ and $y_{\bullet b}$ being the total colocalizations for RNA a and b respectively, and $y_{\bullet\bullet}$ being the total colocalization count in the dataset. d_i^{ab} is a binary indicator denoting whether the genomic distance between a and b falls into distance bin *i.* The genomic distance bins refer to the distance between the genes encoding the two RNAs and whether they are encoded on the same DNA strand or the opposite strand. Eight genomic distance bins were used for NBR with the HEK and whole neuron datasets: 0 to 10 kb same strand, 10 to 50 kb same strand, 50 to 100 kb same strand, 100 to 500 kb same strand, 0 to 10 kb opposite strand, 10 to 50 kb opposite strand, 50 to 100 kb opposite strand, and 100 to 500 kb opposite strand. Only two genomic distance bins were used for the isolated neurite dataset because the other genomic bin terms were not significant in the NBR model: 0 to 10 kb same strand and 10 to 50 kb same strand.

Prior SPRITE studies looking at RNA-DNA colocalization demonstrated that co-transcriptional nucleic acid colocalizations can be picked up by SPRITE, and the increased colocalization for RNAs encoded within 500 kb (roughly exponentially decreasing with distance) are consistent with expectation for chromatin territories. Because the neurite dataset did not include material from cell nuclei, we believe that most of the increase in colocalizations between RNAs encoded within 50 kb on the same strand are due to the STAR alignment algorithm randomly choosing a transcript annotation when a sequencing read aligns equally well to multiple transcripts that are annotated for the same genomic locus. This is a common occurrence as many genomic loci, especially those that are highly transcriptionally active, can have many overlapping annotations with different Ensembl gene IDs. The genomic distance over which we observe this (0 to 50 kb) is consistent with the typical length of mammalian genes. As described above, we do not allow for homotypic colocalization (multiple alignments of a single gene ID to the same barcode) because we cannot discriminate whether the reads came from a single RNA molecule or separate molecules. However, that filter does not work if STAR calls one read as aligning to gene ID A and a second read from the same molecule as aligning to gene ID B. We used the genomic distance bins in the NBR to overcome this issue and account for the spurious increase in colocalizations between overlapping genes, as well as bona fide co-transcriptional colocalizations so that the genomic position of an RNA did not confound our analyses.

We defined colocalization “pairwise specificity” as the *Pearson residual* from the NBR,

$$r_{ab} = \frac{y_{ab} - \mu_{ab}}{\sqrt{var(y_{ab})}} \qquad (2)$$

where μ_{ab} is from the model fit in Equation (1). r_{ab} reflects the degree to which the number of colocalizations between a given pair of RNAs deviates from an expectation of random colocalization. We calculated “within group” mean pairwise specificity by taking each RNA (mRNA or ncRNA, as labeled) annotated to localize to a particular subcellular region from other studies (see above) and calculating the mean pairwise specificity for that RNA with all other RNAs in the group (mRNAs and ncRNAs). “Outside group” mean pairwise specificity was calculated by taking each RNA (mRNA or ncRNA, as labeled) annotated to localize to a particular subcellular region and calculating the mean pairwise specificity for that RNA with all other RNAs *not* annotated to be in the group (mRNAs and ncRNAs). To assess the statistical significance of individual colocalization counts y_{ab}, we calculated two-sided *p*-values taking the predictive distribution of Equation (1) as the null distribution. We convert these *p*-values into corresponding signed *z*-scores.

negative_binomial_regression/
2_Neg Binom result check local_global sequence homology.R

2_Neg Binom result check overlap sequence homology.R

With our current technique, we can't easily tell if two reads with the same barcode that align to different regions of the same transcript sequence are from a single RNA molecule or from two separate RNA copies from the same gene. RNA molecules are fragmented during the RNA-SPRITE protocol, and therefore a single RNA molecule can have multiple barcodes ligated to it in different locations along the transcript. To overcome this issue and a similar problem with homologs, we take a very conservative approach and ignore all colocalizations between transcripts that share sequence homology. As a result, we can't report colocalization of multiple copies of a single RNA species. To avoid reporting colocalizations between transcripts with sequence homology, we aggregate observations aligned to paralogs into a single gene ID. We identified paralogs through Ensembl annotation (at least 40% sequence identity). Importantly, to ensure that none of the RNAs we found to be significantly colocalizing had strong sequence homology, we also checked all RNA pairs with a NBR raw z-score above 3 for sequence homology using Biostrings (v2.70.3) pairwiseAlignment function ("local-global" and "overlap"). For gene pairs with strong sequence homology ("local-global" > 0 or "overlap" > 50), their observations were aggregated into a single gene ID as well, and NBR of the dataset was repeated until no significantly colocalizing pairs demonstrated strong sequence homology.

Explaining pairwise specificity

3_add terms to negbinom.R

To estimate the effect of pairwise binary features x_{ab}, e.g., known protein-protein interaction (PPI) between the proteins encoded by *a* and *b*, we fit the extended NBR model,

$$\ln \mu_{ab} = X_{ab}\beta = \beta_0 + \beta_1 \ln(P_{ab}) + \beta_2 x_{ab} + \beta_3(x_a + x_b) + \sum_i \beta_i d_i^{ab} \tag{3}$$

where x_a and x_b are binary indicators for *a* and *b* having an annotation in that category respectively (e.g., having any known PPI). The $x_a + x_b$ term controls for any non-specific effect of the annotation. The remaining terms are analogous to those in the null model. We obtain standard errors for each coefficient β and use these to calculate Wald statistic *p*-values for the significance of each term. To ease interpretation, we calculate $\alpha = \exp(\beta_2)$ which is a multiplicative factor representing how much more frequent colocalization is when $x_{ab} = 1$. For the plots with α, the p-values reported with refer to the Wald statistic for β_2 , and the 95% confidence intervals = exp(β_2 ± 2*standard error). We are primarily interested in the effect size reflected by α and whether it is statistically significantly different from 1, which implies the pairwise feature has a true effect on colocalization.

For testing whether shared binding of a specific RBP effects colocalization we use the extended NBR of Equation (3) where now $x_{ab} = x_a x_b = 1$ if and only if the RBP binds both *a* and *b.* The control term $x_a + x_b$ accounts for the possibility that binding of this RBP increases colocalization counts globally rather than specifically for partners to whom it also binds.

Dataset annotations

Protein interaction data was used from CORUM, Complex Portal, and STRINGdb databases[111–113]. For STRINGdb, we used a physical score threshold of 700 as an annotation of a physical interaction. Proteins were considered interacting if they or their human homologs were identified as interacting in at least two databases. mRNA-mRNA duplexes were identified using the RISE database[107]. Only experiments that identified direct (not protein-dependent) RNA-RNA interactions were used: PARIS, LIGRseq, and SPLASH[63,64,107,108]. Interactions annotated between rodent or human protein-coding genes were used, but “intronic” or “intergenic” were not used. There was not enough overlap among the datasets to do a 2 assay cutoff as we did for other annotations.

mRNA colocalization hub analysis
3_Louvain coms and UMAP.RmD

mRNA-mRNA colocalization counts in the whole cell neuron dataset were analyzed using custom R scripts. Barcodes aligning to any known nuclear RNAs were excluded to focus on cytoplasmic organization of mature mRNAs. To examine only mRNAs that had evidence of specific colocalization, only mRNAs with a maximum pairwise colocalization z-score of at least 3.2 were used (7702 out of 8322 mRNAs). Negative pairwise specificity scores were changed to zero to avoid modeling deviation from expectation in non-colocalizing pairs.

To reduce noise and extract key axes of variation, we performed singular value decomposition (SVD) of the pairwise specificity scores using the RSpectra package (v0.16-2) in R, which efficiently computes the top K singular values and vectors for large matrices. The top K=5 left singular vectors (U) were scaled by their corresponding singular values (D) to obtain the SVD-transformed data. To identify local structure, we constructed a nearest neighbor graph using the RANN package (v2.6.2), connecting each mRNA to its 100 nearest neighbors. The resulting adjacency matrix, where entries were set to 1 for nearest neighbors, was converted into a graph object weighted by Pearson residuals. We then applied Louvain community detection (cluster_louvain function in igraph v2.1.4), using a resolution parameter of 0.5 to identify clusters.

Separately, we applied Uniform Manifold Approximation and Projection (UMAP) to the SVD-transformed data for nonlinear dimensionality reduction using the umap R package (v0.2.10.0). The embedding was generated using a cosine distance metric and 50 nearest neighbors, allowing for visualization of mRNA colocalization hubs.

Free energy of folding estimates
nupack-calculations/

Gibbs free energy of folding (ΔG) was estimated using NUPACK 4.0[88] using custom scripts. Free energy of dimerization was calculated by subtracting the summed monomer folding free energies from the dimer free energy. For free energy of dimerization comparisons, we considered the top 1000 highest confidence colocalizing mRNA pairs (by z-score) in each dataset vs. control pairs where the second mRNA was replaced with a length-matched control that did not colocalize with the first mRNA more than expected by chance. We also calculated the free energy of dimerization for colocalizing mRNA pairs found to form a duplex by proximity ligation experiments[107] that also colocalized in HEK RNA-SPRITE data (850 pairs), and compared them to length-matched controls as above.