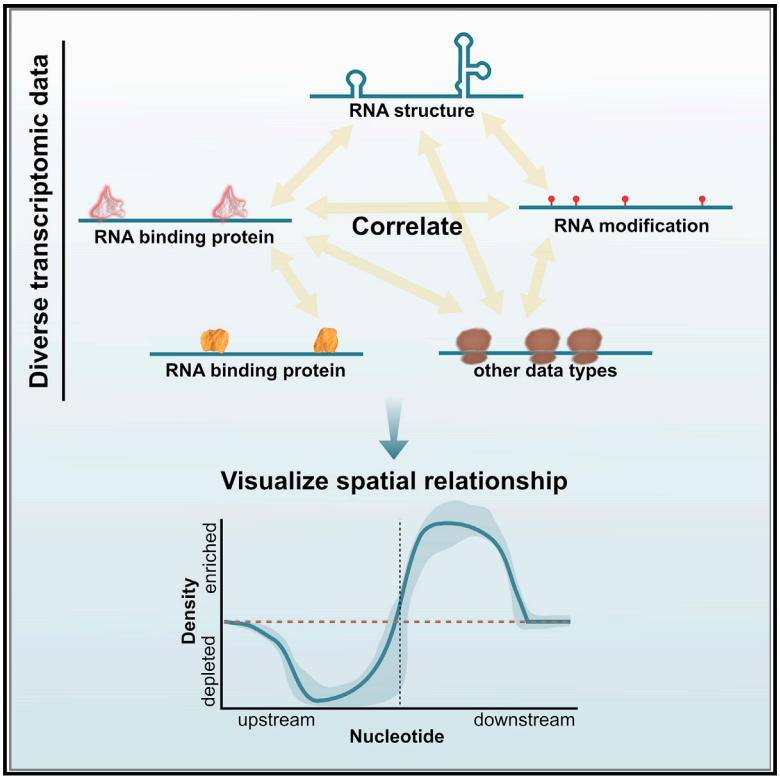
Cell Reports
Methods

Article

Spatial correlation statistics enable transcriptome-wide characterization of RNA structure binding

Graphical abstract



Highlights

- nearBynding is an R/Bioconductor algorithm to identify spatial relationships
- Flexible scaffold to cross-correlate diverse transcriptomic features
- Calculates features at or adjacent to annotation sites transcriptome-wide
- Accommodates interval or continuous data formats

Authors

Veronica F. Busa, Alexander V. Favorov, Elana J. Fertig, Anthony K.L. Leung

Correspondence

efertig@jhmi.edu (E.J.F.), anthony.leung@jhu.edu (A.K.L.L.)

In brief

The nearBynding algorithm calculates and visualizes spatial relationships across the transcriptome. Busa et al. demonstrate that nearBynding can recapitulate known protein-binding preferences for structured RNA and RNA modifications as well as known geometries between RNA-binding proteins. nearBynding's spatial correlations provide biological insights into protein binding of G-quadruplexes.

Busa et al., 2021, Cell Reports Methods *1*, 100088
October 25, 2021
https://doi.org/10.1016/j.crmeth.2021.100088

CellPress

Cell Reports Methods

CellPress
OPEN ACCESS

Article

Spatial correlation statistics enable transcriptome-wide characterization of RNA structure binding

Veronica F. Busa,[1,2] Alexander V. Favorov,[4,7] Elana J. Fertig,[1,4,5,6,*] and Anthony K.L. Leung[1,2,3,4,8,*]

[1]McKusick-Nathans Department of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA
[2]Department of Biochemistry and Molecular Biology, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD 21205, USA
[3]Department of Molecular Biology and Genetics, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA
[4]Department of Oncology, Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA
[5]Department of Biomedical Engineering, Johns Hopkins University Whiting School of Engineering, Baltimore, MD 21205, USA
[6]Department of Applied Mathematics and Statistics, Johns Hopkins University Whiting School of Engineering, Baltimore, MD 21205, USA
[7]Laboratory of Systems Biology and Computational Genetics, Vavilov Institute of General Genetics, Russian Academy of Sciences, Moscow, Russia
[8]Lead contact
*Correspondence: efertig@jhmi.edu (E.J.F.), anthony.leung@jhu.edu (A.K.L.L.)
https://doi.org/10.1016/j.crmeth.2021.100088

MOTIVATION Because RNA regulatory processes often occur at overlapping or adjacent transcriptional coordinates, assessing spatial relationships at a transcriptome-wide scale represents a powerful means of evaluating RNA structure, modification, and regulation. Application of available genome-wide correlation methods to discern transcriptomic spatial relationships is inefficient and/or imprecise. Machine-learning algorithms tailored to transcriptomic data exclusively rely on analysis of overlapping coordinates and cannot assess adjacent relationships. nearBynding is able to efficiently correlate diverse types of transcriptomic data such as protein binding, RNA structure, and RNA modification to calculate and visualize their spatial relationships.

SUMMARY

Molecular interactions at identical transcriptomic locations or at proximal but non-overlapping sites can mediate RNA modification and regulation, necessitating tools to uncover these spatial relationships. We present nearBynding, a flexible algorithm and software pipeline that models spatial correlation between transcriptome-wide tracks from diverse data types. nearBynding can process and correlate interval as well as continuous data and incorporate experimentally derived or *in silico* predicted transcriptomic tracks. nearBynding offers visualization functions for its statistics to identify colocalizations and adjacent features. We demonstrate the application of nearBynding to correlate RNA-binding protein (RBP) binding preferences with other RBPs, RNA structure, or RNA modification. By cross-correlating RBP binding and RNA structure data, we demonstrate that nearBynding recapitulates known RBP binding to structural motifs and provides biological insights into RBP binding preference of G-quadruplexes. nearBynding is available as an R/Bioconductor package and can run on a personal computer, making correlation of transcriptomic features broadly accessible.

INTRODUCTION

There has been an expansion in RNA profiling tools developed to measure transcriptional attributes associated with RNA structure, modification, and regulation. Sequencing-based tools have been developed to elucidate the RNAs bound by a particular protein (Sugimoto et al., 2012), to provide a snapshot of the precise locations of ribosomes on RNA (Ingolia et al., 2009), to identify loci with adenosine-to-inosine editing (Okada et al., 2019), and to provide RNA structure information (Lucks et al., 2011), to name only a small sample of diverse transcriptomic data types. Analysis approaches to integrate these diverse

Table 1. Comparison of functionality for correlation analysis programs

		deepTools *plotCorrelation*	bedtools *reldist*	StereoGene	nearBynding
Data type	optimized for single-nucleotide resolution				✓
	analyze interval data	✓	✓	✓	✓
	analyze non-binary data	✓		✓	✓
	analyze continuous data			✓	✓
Correlation capabilities	partial correlation			✓	✓
	cross-correlation			✓	✓
	correlate non-local features		✓	✓	✓
	correlation visualization	✓		✓	✓
Transcriptome tools	retains strand information				✓
	select regions of transcriptome				✓
	generate and integrate RNA structure predictions				✓

See also Figure S8 and STAR Methods.

data modalities and identify inter-related features can lead to novel hypotheses about biological regulation.

Efficient methods to correlate genome-wide features are available (Favorov et al., 2012; Stavrovskaya et al., 2017), but robust transcriptome-wide spatial correlation requires new tailored extensions. Transcriptomic data constitute only a fraction of the genome and their analysis often requires nucleotide-level spatial relationship resolution (Table 1), in contrast to the 100-bp to megabase resolution information that usually suffices for genomic data (Kravatsky et al., 2015; Stavrovskaya et al., 2017; Zhang et al., 2011). Therefore, applying genomic tools directly to calculate the spatial correlation of transcriptomic data analysis is computationally inefficient and imprecise. To overcome this limitation, the main approach currently used to assess colocalization of transcriptomic features adapts genome-based tools to compare features at identical transcriptomic locations or within windows (overlapping genomic coordinates) (Lee et al., 2020b; Luo et al., 2020; Sauer et al., 2019; Wolfe et al., 2020). However, biologically important relationships often occur at proximal but non-overlapping transcriptomic sites (adjacent genomic coordinates) (Beltran et al., 2019; Briese et al., 2019; Carlile et al., 2019; Van Nostrand et al., 2020a; Waldron et al., 2019), necessitating analyses that evaluate the relative positions of transcriptomic features. Moreover, these tools should also be flexible to allow for associations of binary features from processed RNA profiling data and continuous track data from assays that resolve quantitative features along the transcriptome.

One notable example in which transcriptome-wide correlation is particularly applicable is in the example of RNA-binding protein (RBP) preference for RNA structure and modification. Some RBPs recognize RNA structure more than sequence (Blaszczyk et al., 2004; Heller et al., 2017), but binding preferences to structured RNA have thoroughly been described for only a few proteins, and RNA structure surrounding protein-binding events is rarely characterized. Sequence motifs ascribed to RBPs are often insufficient for explaining a large proportion of binding occurrences (Bahrami-Samani et al., 2015; Edupuganti et al., 2017; Li et al., 2010, 2014; Taliaferro et al., 2016; Wilbert et al., 2012). Describing the unexplained binding of RBPs—especially for RBPs that bind structured RNAs—will increase our potential to elucidate the etiology of diseases driven by dysregulated protein-RNA interactions.

Several machine-learning algorithms have been developed to resolve structure-based RBP motifs using cross-linked immunoprecipitation (CLIP) data and RNA structure prediction (Maticzka et al., 2014; Pan and Shen, 2018; Yan et al., 2020). Algorithms such as GraphProt and iDeepS incorporate a post-processing step to easily visualize RBP sequence and structure preferences (Maticzka et al., 2014; Pan and Shen, 2018), but these algorithms only provide visualization of structure information for a short binding motif (7–12 nucleotides). The predictive power of many state-of-the-art algorithms may be limited by their reliance exclusively on sequence-based RNA structure prediction and their lack of accommodation for experimentally derived RNA structure information (reviewed by Chen et al., 2019; Sasse et al., 2018). The recent algorithm PrismNet has begun addressing these problems by allowing the incorporation of *in vivo* click selective 2′-hydroxyl acetylation and profiling experiment (icSHAPE) data (Sun et al., 2021), but all algorithms to resolve structure-based RBP motifs still rely exclusively on analysis of overlapping coordinates and do not offer insight about the RNA structure surrounding those motifs, despite evidence that such context can be important in RBP binding (Carlile et al., 2019; Jarmoskaite et al., 2019). Current methods interrogating RNA structure binding contexts at and adjacent to the binding site can only be performed in a low-throughput manner (Carlile et al., 2019; Jarmoskaite et al., 2019); therefore, RBPs known to have preferred secondary structures are sparse. Efficient methods to perform transcriptome-wide correlation are needed to overcome these limitations and resolve global RBP binding to RNA structure.

Here, we present a new algorithm, nearBynding, to calculate and visualize spatial correlations between transcriptomic data types. It is implemented in an R/Bioconductor package by the same name to promote accessibility and ease of use. The nearBynding algorithm is unique in three ways: first, it calculates correlations to indicate features colocalized at, or adjacent to,

CellPress
OPEN ACCESS

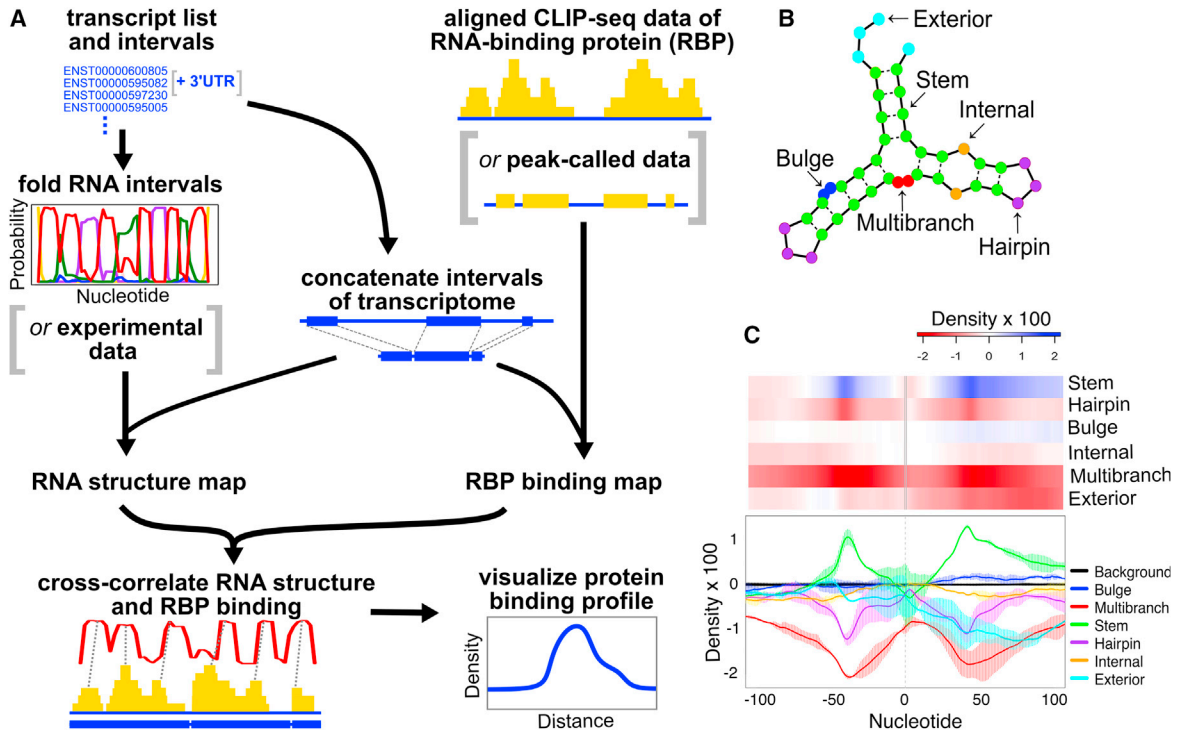


Figure 1. Overview of nearBynding

(A) The user inputs CLIP-seq data (aligned reads or called peaks), a list of transcripts, and an annotated genome. Optionally, an *in silico* predicted RNA structure track can be replaced by an experimentally derived one. RBP binding and RNA structure data are mapped to the concatenated transcriptome and cross-correlated. nearBynding outputs cross-correlation densities and their distributions to estimate RNA binding.

(B) Examples of six RNA structure contexts predicted by CapR (Fukunaga et al., 2014) for which nearBynding can be applied.

(C) Example heatmap and line plot visualizations of PUM2 binding from eCLIP data in two replicates of K562 cells estimated from the cross-correlation densities and visualized as part of the nearBynding software. The line plot shows the average signal and plus or minus three standard errors; the black line and error bars represent null signal (labeled as background) derived from recursively shuffled foreground signal. The heatmap only shows the average value at every position in cases where multiple samples are used to calculate density values.

annotation sites in a transcriptome-wide manner; second, it is a flexible scaffold to cross-correlate diverse transcriptomic data types; and third, it can analyze transcriptomic data in either interval or continuous data formats. The algorithm achieves transcriptome-wide correlation by expanding our current tool for genome-wide correlation, StereoGene (Stavrovskaya et al., 2017), to allow users to select and interrogate specific transcripts and specific regions of those transcripts (e.g., unspliced or mature; coding sequences [CDSs] or 3′ untranslated regions [UTRs]; see Table 1 and STAR Methods for capability comparisons). Users can then use nearBynding to input two tracks, calculate the pairwise correlation between those tracks, and visualize the correlation along the transcriptome.

We benchmark nearBynding using simulated data and replicates from enhanced CLIP (eCLIP) experiments. We demonstrate nearBynding's utility by comparing our results with known RBP binding preferences, such as RBP binding to RNA structure, RBP binding relative to a second RBP, and RBP binding to RNA modifications. We also demonstrate how nearBynding could be applied to discern binding preference differences between a wild-type (WT) and mutant RBP. Besides *in silico* predictions, we employ diverse experimentally derived data types (e.g., RNA G-quadruplex sequencing [rG4-seq], N^6-methyladenosine [m6A]-iCLIP sequencing [miCLIP-seq]) that are not utilized by currently available RBP motif-finding software in our correlation-based predictions of RBP binding preferences. We then use these discovered RBP binding preferences from our transcriptome-wide correlation analyses to hypothesize RBP characteristics that may predispose binding preferences for or against specific RNA structures. We show that aligned reads (continuous data format) provide qualitatively similar outputs and comparable reproducibility between technical replicates compared with peak-called data (interval data). The ability of nearBynding to correlate any interval or continuous feature annotated across selected regions of the transcriptome makes it a diverse, flexible tool to study RBP binding, RNA structure, RNA modification, and potentially other RNA features.

RESULTS

nearBynding probes transcriptome-wide RBP binding to RNA structures

We use the nearBynding algorithm to incorporate transcriptome-wide information of RBP binding sites and RNA structure to discern RNA structure for regions bound by an RBP as an example of its potential (Figure 1A). nearBynding requires only

a list of transcripts, an annotated genome, and aligned CLIP sequencing (CLIP-seq) data as inputs. RNA structure data are an optional input and can be predicted within the pipeline. nearBynding produces a concatenated transcriptome made of only the transcriptome regions being probed and maps the data tracks (e.g., CLIP-seq and RNA structure) to it, which drastically reduces the magnitude of the datasets to only the intervals of interest. With this extension enabling transcriptome-scale analysis, the algorithm for efficient spatial correlations in StereoGene (Stavrovskaya et al., 2017) is then applied to calculate the pairwise correlation between the two data tracks.

To enable analysis of RBP binding, the nearBynding software further includes visualization tools of the output statistics that are tailored to illustrate the relative positions of RBP binding and RNA structure. By default, nearBynding uses RNA structure probabilities predicted from sequence by CapR (Fukunaga et al., 2014) for the selected transcriptomic intervals (see STAR Methods). While CapR provides the default structural data input, the algorithm can also accept alternative inputs of custom RNA structure tracks or intervals, such as RNA modifications that affect RNA structure (e.g., m6A; see below). StereoGene generates cross-correlation densities for RNA folding contexts relative to RBP binding. Since cross-correlation shows the relative position of one track (e.g., RBP binding) to another track (e.g., RNA structure), we can use it as a tool to visually represent the location of the RNA structure relative to the RBP binding site (i.e., upstream, at, or downstream of RBP binding). To account for the case in which a transcriptional track is correlated to *in silico* predictions of structure, the visualizations are performed separately for RNA structures based on their categorization as double-stranded (stem) or one of five single-stranded types: hairpin, multibranch, internal, exterior, and bulge (Figure 1B).

Binding profiles illustrating RNA structures at and proximal to RBP binding can be visualized either as line plots with standard errors for cases with multiple replicates, or as heatmaps (Figure 1C). nearBynding, via StereoGene, calculates a null signal derived from the distribution of the correlation metrics for randomly shuffled windows (black line; Figure 1C, lower panel). nearBynding also calculates a plus or minus one standard error confidence interval for the foreground signal when more than one experimental replicate is input for analysis; statistical significance can be assessed by comparing the distribution of the foreground signal computed from replicates with the null distribution computed from randomly shuffled windows, and users can alter the number of standard errors shown by the error bars to increase or decrease confidence intervals (e.g., three standard errors were shown in all figures presented here). This combination of visualization and statistics can be used to predict RBP binding to and adjacent to RNA structures transcriptome-wide. In addition to allowing visual assessment, nearBynding includes functions to quantitatively compare RBP binding cross-correlation distributions between two different RBPs. Specifically, the software computes the Wasserstein, or earth-mover, distance (Schuhmacher et al., 2020) between pairs of RBP binding profiles. For example, a short Wasserstein distance suggests similarity between two RBP profiles and possible binding competition between RBPs.

RBP binding information from CLIP-seq data can be input to the nearBynding software as a BAM file of aligned reads or as a BED file of peak intervals or protein-RNA cross-linked sites. Our transcriptome-wide correlation algorithm is applicable to a variety of CLIP-seq data types. Some processing methods estimate binding peaks or cross-linking sites that correct the CLIP-seq data for a size-matched input (Drewe-Boss et al., 2018; Lovci et al., 2013) or modeled background signal (Uren et al., 2012; Zhang and Xing, 2017). When inputting raw CLIP data, the nearBynding software also allows for input of a size-matched input background track, the output of which is subtracted from the output signal of the foreground track prior to computing the correlation statistics. We hypothesized that including a background track would ensure that the observed signal is from the RBP of interest rather than from experimental artifacts, such as cell-specific transcript levels or size-matched input noise. To test this hypothesis, we applied nearBynding to calculate and visualize the RNA structure preferences of the poly(C)-binding protein HNRNPK (Matunis et al., 1992) and found that the binding profiles for HNRNPK in HepG2 and K562 cells were much more similar after background signal was removed (Wasserstein distance of 1.80 between non-corrected profiles versus 0.23 for background-corrected profiles) (Figure S1).

Benchmarking on simulated data demonstrates robust signal detection

Because of the uniqueness of this algorithm, we were unable to directly compare nearBynding with other algorithms available for spatial correlation of genomic tracks (see Table 1 and STAR Methods for comparisons with related spatial relationship algorithms). To evaluate its performance, we instead designed simulated tracks (we term these RBP binding and RNA structure but they could instead represent any two transcriptomic data types) to benchmark a full range of biological variables that may affect performance (Figure 2). Briefly, we tested three factors that may affect signal strength: peak concordance between tracks, foreground to background ratio, and peak width range. Each simulation contained a pairwise analysis of the cross-correlation between an RNA structure track (RNA) and a CLIP-seq track (CLIP), where a greater amplitude for cross-correlation density reflects better co-occurrence of the two tracks. The peak distances and heights of the RNA structure track were varied to simulate the range of predicted RNA structure probabilities and random distribution of these structures across the transcriptome. The RNA structure track consisted of 10,000 peaks 31 to 500 nucleotides apart (unless otherwise stated), 5 nucleotides in width, and 0.02 to 1 unit in height. The CLIP-seq track simulated signal from aligned CLIP-seq data and contained a mixture of background and foreground signal. The CLIP-seq track contained 30-unit-wide peaks (unless otherwise stated) to simulate the 30-nucleotide reads of CLIP-seq data deposited in the ENCODE portal (Davis et al., 2018). The CLIP-seq track was also shifted 12 units to the left of and equal in height to the RNA structure track peaks.

First, we tested the impact of the frequency of an RBP binding to its target RNA structure across the transcriptome, which may be affected by the accessibility of the RNA structure and the binding strength of the protein. To simulate this effect, we varied the frequency of the foreground signal peak concordance of the CLIP-seq track relative to the RNA structure track (Figure 2A). We hypothesized that tracks with a higher frequency of RBP

CellPress
OPEN ACCESS

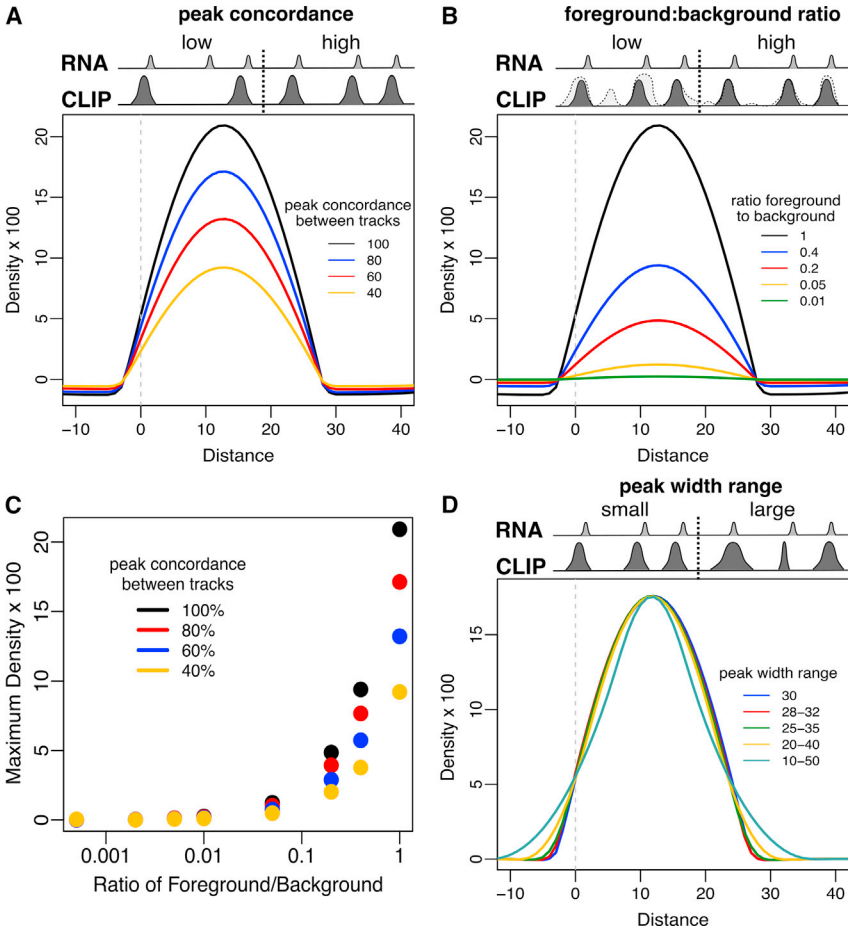


Figure 2. Cross-correlation distribution tracks of simulated RNA-binding data to benchmark the performance of nearBynding

The RNA structure track [RNA] is shifted 12 units to the right relative to the CLIP track to model proximal RNA structure. For (A), (B), and D) the middle gray peaks represent RNA structure data and the dark gray peaks represent signal from CLIP simulation data (i.e., foreground).

(A) Cross-correlation distribution tracks with differing peak concordance.

(B) Cross-correlation distribution tracks with differing foreground to background signal ratios. Foreground signal (dark gray) does not change, but different amounts of randomly distributed background signal is added to the foreground, as represented by the lightest gray regions of peaks.

(C) Maximum cross-correlation density values for pairs of tracks with varying peak concordance and foreground to background signal ratios.

(D) Cross-correlation distribution tracks with differing peak width range.

See also Figure S2.

binding to target RNA structure would provide stronger binding signals than RBPs with sparser target binding. Supporting this hypothesis, the result of the nearBynding algorithm for the simulated data showed that cross-correlation signal strength correlates positively with peak concordance.

Next, we simulated artifacts associated with collecting CLIP-seq reads, such as background signal from input, by varying the height of the background signal of the simulated CLIP-seq track relative to the foreground signal (Figure 2B). We hypothesized that simulations with a greater foreground (dark gray) to background signal (light gray) would have stronger RBP binding signals. As expected, cross-correlation signal strength correlated positively with the ratio of foreground to background. Both peak concordance and foreground to background ratio greatly affected signal strength, with nearBynding requiring a foreground to background signal greater than 0.05 to detect the binding signal (Figure 2C). Therefore, our algorithm performance may be superior when applied to data collected by protocols that minimize noise (e.g., via additional washing steps) rather than protocols that document all binding events at the expense of greater noise.

We further employed our simulated data to test the sensitivity of nearBynding to the uniformity of peak width. Specifically, we increased the range of the simulated CLIP-seq peak widths to accommodate the possibility that RBPs may have variable binding footprints (Figure 2D). Although the shape of the cross-correlation density track changed to reflect greater variation in peak widths, the amplitude and position of the signal maximum did not since the peaks were still centered at the same location. Therefore, we conclude that differences in peak width have no effect on signal amplitude.

nearBynding creates a concatenated transcriptome made of only the regions being interrogated as an input for StereoGene. However, signal may not be evenly distributed across the concatenated transcriptome. Therefore, we tested the dependence of nearBynding on the distribution of peaks along a concatenated transcriptome by shifting the locations of the simulated peaks such that they were all uniformly distributed or clustered near either end of the CLIP-seq track (Figure S2). Compared with peak concordance and foreground to background ratio, only a negligible loss in signal amplitude was observed for the most extremely skewed data (Figure S2). Overall, our results demonstrate that the order in which transcripts are concatenated, which could possibly affect the distribution of peaks, has negligible effect on binding signal relative to other variables tested.

Called peaks or aligned tracks for RBP binding produce similar binding profiles

Current practice for analyzing CLIP data is to call RBP-bound peaks using algorithms such as Piranha (Uren et al., 2012), CLIPper (Lovci et al., 2013), CLAM (Zhang and Xing, 2017), or omniCLIP (Drewe-Boss et al., 2018). The ENCODE portal (Davis et al., 2018) has eCLIP datasets for 103 RBPs in HepG2 cells and 120 RBPs in K562 cells, with each dataset containing two replicates and an input control. We selected 29 different RBPs in HepG2 and K562 cells that demonstrate strong, reproducible

binding signals at 3′ UTRs based on analysis from Van Nostrand et al. (2020b) (Figure S3A), which came to 40 unique cell-type-RBP combinations. We used these high-confidence datasets to test whether nearBynding can produce comparable peak binding profiles from peak callers and aligned reads. We collected eCLIP aligned reads and binding peaks called by two of the most commonly used peak callers, Piranha and CLIPper, for these RBPs. We ran Piranha on all replicates with parameters as described in the original publication (Uren et al., 2012). We also downloaded CLIPper-derived peaks of the eCLIP data from the ENCODE data portal (Lovci et al., 2013). These three different inputs—aligned eCLIP reads, Piranha peaks, and CLIPper peaks—were run through nearBynding, and RBP binding was assessed for 3′ UTR-annotated regions of the transcriptome. We used the Wasserstein distance (Schuhmacher et al., 2020) to determine the amplitude and distance required to transform one RBP binding profile into another. We calculated the sum of Wasserstein distances between the cross-correlation density tracks of all RNA structure contexts for all 40 unique cell-type-RBP combinations across the three different input types each. Visualizing their distances in 2D on a multidimensional scaling plot (Figure S3B) and comparing binding profiles across input types (Figure S3C) showed only minor differences in the binding profile for RBPs based on the input source. Iterative sampling of binding profiles further indicated that the binding profiles for the three input sources of the same protein are more closely clustered compared with randomly chosen binding profiles from other proteins in the same cell line ($p = 5.59 \times 10^{-7}$ in K562 cells and $p = 2.51 \times 10^{-10}$ in HepG2 cells, Kolmogorov-Smirnov test; Figure S3D). Therefore, the difference between profiles for different RBPs is greater than the difference within the same experiment queried via different inputs.

We next tested whether technical replicate reproducibility is comparable between peak-called and aligned read inputs by calculating the Wasserstein distance between replicates for the same 40 unique cell-type-RBP combinations using the three different input track types. The distance between aligned read binding profiles for technical replicates was smallest in 12 of 40 cases (30%), Piranha peak-called replicate profiles were closest in 23 of 40 cases, and CLIPper peak-called replicate profiles were closest in 5 of 40 cases (Figure S3E). Further, the distances between technical replicates for peak-called and aligned read tracks are quantitatively similar for the majority of cell-type-RBP combinations tested (Figure S3F). These data cumulatively suggest that aligned reads generate similar outputs for technical replicates via nearBynding compared with peak-called tracks.

Cross-correlation tracks reproducibly cluster RBP data across biological replicates

The context-dependence of RNA binding can be expected to cause variable signal concordance of binding predictions from CLIP-seq data for the same RBP. Replicates from the same cell type would manifest technical differences, whereas analyses of the same RBP across different cell types may show additional biological differences in RBP binding. We therefore expected that binding profiles of the same RBP between replicates within the same cell type would have greater concordance than profiles from different cell types. We assessed nearBynding's ability to reproducibly identify RBP binding contexts across replicates and cell types by clustering RBP binding profiles.

Within the ENCODE datasets, 73 RBPs are common across both cell lines. Genome-wide RNA structure profiling showed that 3′ UTRs, which are targets for many RBPs, are generally highly structured in cells (Beaudoin et al., 2014; Van Nostrand et al., 2020b). Therefore, in order to test our algorithm on a robust dataset, we restricted our analysis of RBP binding to 3′ UTRs. We collected isoform information of all 3′ UTRs expressed in HepG2 and K562 using RNA sequencing [RNA-seq] data from ENCODE (Davis et al., 2018). We generated cell-type-specific binding profiles by selecting eCLIP reads that aligned to isoforms expressed in the corresponding eCLIP cell type. The most abundant 3′ UTR isoforms for the expressed transcripts were then submitted to nearBynding to determine RBP binding preferences for these regions.

First, we wanted to test how well biological replicates of the same RBP in the same cell type clustered. We calculated the sum of Wasserstein distances between the cross-correlation density tracks of all RNA structure contexts for every sample within each cell type. Seventy-one of 206 replicates in HepG2 (34%, $p = 9.4 \times 10^{-109}$ for a one-tailed binomial test assuming random chance) and 115 of 240 replicates in K562 (48%, $p = 1.3 \times 10^{-203}$, one-tailed binomial test) most closely clustered in pairs with their corresponding biological replicate (Figure 3A). Replicates of RBPs with a large proportion of their binding events in 3′ UTRs tend to cluster more closely based on Wasserstein distance than those of other RBPs (Figure 3B), which is reasonable since the RBP binding profiles were generated from 3′ UTRs of the transcriptome (Van Nostrand et al., 2020a). Among these 3′ UTR-binding RBPs, 33 of 42 replicates in HepG2 (79%, $p = 2.2 \times 10^{-45}$, one-tailed binomial test) and 37 of 44 replicates in K562 (84%, $p = 1.2 \times 10^{-53}$, one-tailed binomial test) most closely clustered in pairs with their corresponding biological replicate, and 93% and 91% of replicates were clustered within the top two distances in K562 and HepG2, respectively (Figure 3C). These analyses demonstrate that biological replicates largely cluster together using Wasserstein distance.

Next, we interrogated the reproducibility of RBP binding profiles across cell lines. The cross-correlation densities of biological replicates for each RBP were averaged, and these averaged values were used to calculate the Wasserstein distances for all RNA structural contexts. For every RBP in K562 cells, we ranked how similar its binding profile was to RBPs in HepG2 cells. Fifteen of 73 RBPs (21%) clustered closest with their counterparts in the other cell line, and 21 of 73 RBP counterparts (29%) were within the top three closest distances in the other cell line (Figures 3D and 3E). Interestingly, there was no difference in clustering distances across cell lines between 3′ UTR-binding RBPs and all other RBPs (Figure S4A). The inverse comparison—the distance of HepG2 RBPs against all K562 RBPs—also had 29% of RBPs cluster within the top three distances of their counterparts (Figure S4B), suggesting poorer concordance across cell types than between biological replicates in the same cell line.

RNA structure cross-correlation signal is specific

We wanted to analyze negative control datasets to test nearBynding's specificity. Chromatin immunoprecipitation (ChIP)

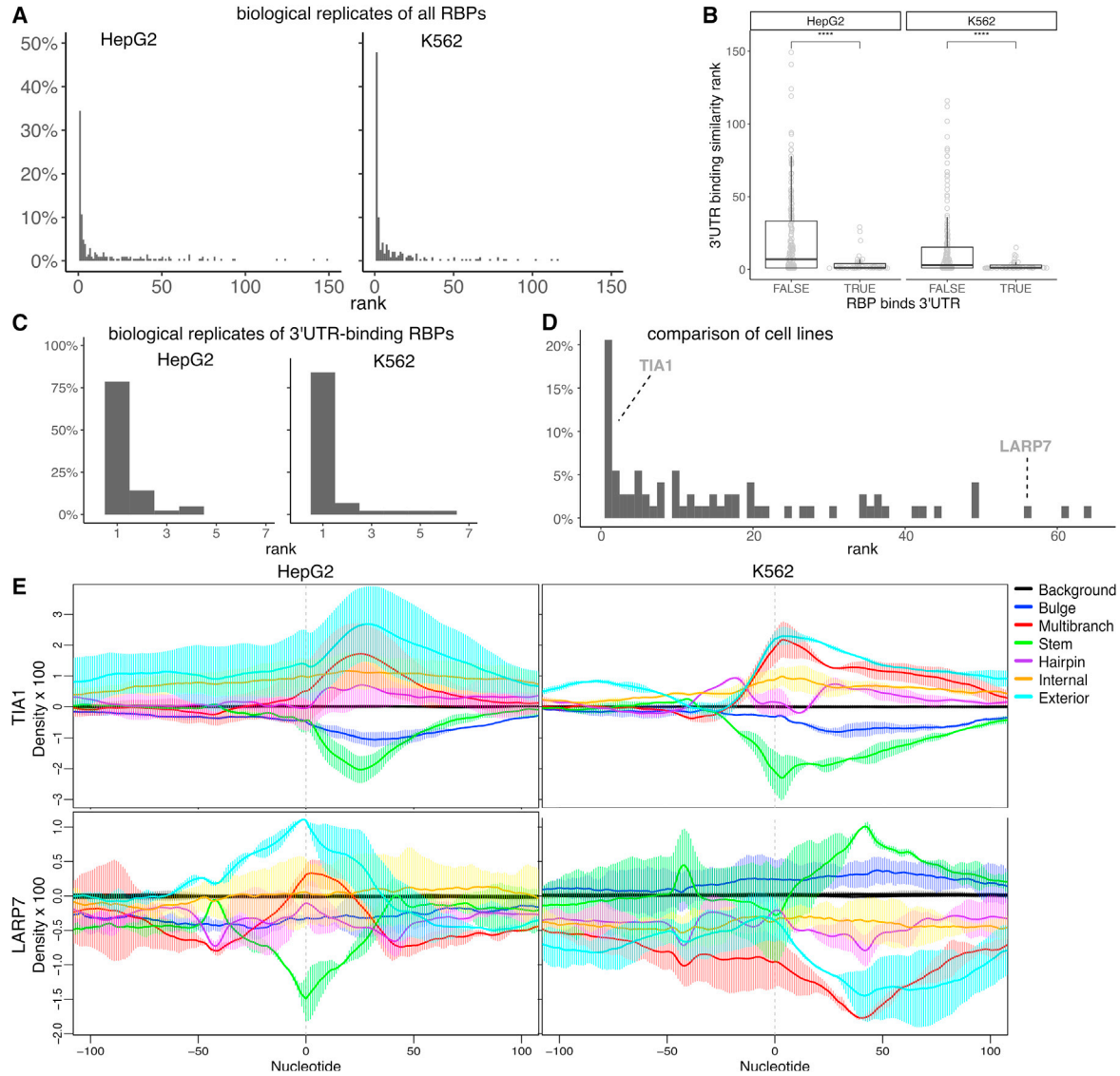


Figure 3. Binding profiles of all RBPs with eCLIP data in ENCODE clustered using Wasserstein distance

(A) Histogram of ranks for Wasserstein distances of paired biological replicates in HepG2 (left) and K562 (right) cells.

(B) Boxplots of biological replicate binding profile similarity rankings for RBPs that do (TRUE) or do not (FALSE) preferentially bind 3′ UTRs compared via t test ($p < 0.0001$ for both) in HepG2 (left) and K562 (right) cells.

(C) Histogram of similarity rankings for Wasserstein distances of paired biological replicates in HepG2 (left) and K562 (right) cells for RBPs that preferentially bind 3′ UTRs.

(D) Histogram of Wasserstein distance similarity rankings for the same RBP across HepG2 and K562 cell lines. The rankings of TIA1 and LARP7 across cell lines are indicated.

(E) Example binding profiles for TIA1 (top), an RBP that is similar across cell types, and LARP7 (bottom), an RBP that is dissimilar across cell types, in HepG2 (left) and K562 (right) cells. Error bars represent plus or minus three standard errors.

See also Figures S3 and S4.

sequencing data depict where proteins bind DNA and can be mapped to the transcriptome, but the protein-bound DNA is expected to have virtually no association with the RNA structure transcribed from the bound genomic regions, making it a reasonable negative control for RNA structure binding preferences. eCLIP and ChIP data in the same cell line were available for six 3′ UTR-binding proteins in the ENCODE portal (Davis et al., 2018). The eCLIP data were predicted to have significant preference for one or multiple RNA structure contexts, while the ChIP data were expected to have no significant binding preference signal. All tested ChIP datasets did not reach statistical significance based on a plus or minus three standard error confidence interval, whereas the eCLIP datasets for the same proteins demonstrated statistically significant RNA structure binding

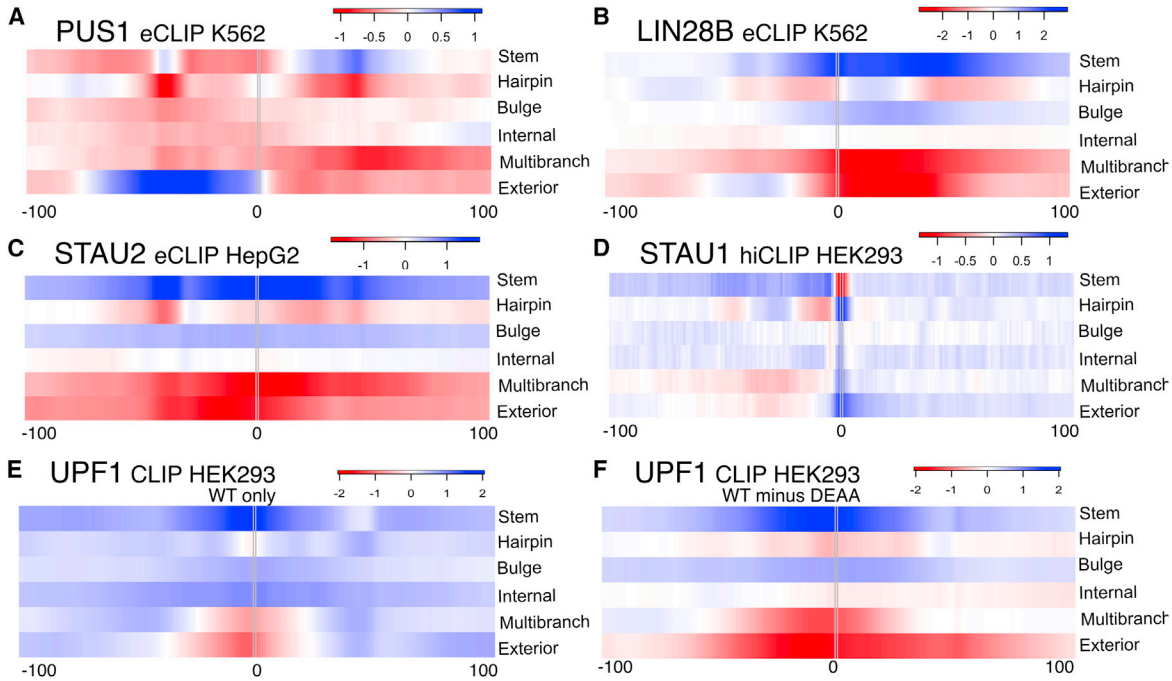


Figure 4. Application of nearBynding to analyze RBP binding profiles for proteins with known structure preference

(A–C) Binding profiles for (A) PUS1, (B) LIN28B, and (C) STAU2 from eCLIP data.

(D) Binding profile for STAU1 from hiCLIP cross-link site data.

(E) Binding profile for WT UPF1 binding from CLIP data.

(F) Binding profile of helicase-dependent UPF1 binding based on subtraction of DEAA UPF1 signal from WT.

See also Figure S5.

preferences (e.g., Figures S4C–S4E). These negative control results suggest that nearBynding's detection of RNA structure binding preferences is specific.

RBP binding profiles recapitulate known structural preferences

Next, we tested whether the binding profiles generated by nearBynding reflect known RBP structural binding preferences. We selected four diverse RBPs to represent RBPs with a defined sequence preference (PUM2), enzymatic activity (PUS1), multiple types of RNA-binding domains (LIN28B), and a well-characterized structure preference (STAU2).

We selected PUM2 for analysis because its modular structure of eight tandem repeats is known to recognize RNA in a sequence-specific manner (Wang et al., 2002). Although PUM2 preferentially binds 3′ UTRs in a sequence-specific manner, there is evidence that PUM2 also has a structural component to its binding preferences: *in vitro* analysis shows that PUM2 dissociates from double-stranded regions faster than single-stranded regions and that it stably binds regions flanked by stem structures (Jarmoskaite et al., 2019). The PUM2 binding profile (Figure 1C) showed that PUM2 has minimal structure preference at the point of binding (nucleotide = 0), but it does prefer stem context upstream and downstream of its point of contact.

The enzyme PUS1 adds a pseudouridine modification to target RNAs (Carlile et al., 2019). PUS1 has a weak trinucleotide binding sequence motif and modifies nucleotides at the 5′ end of stem loops flanked by single-stranded runs for the vast majority of its high-confidence targets (Carlile et al., 2019). Consistent with PUS1 binding and modifying the 5′ base of stems, its binding profile showed a preference for single-stranded regions at the end of the transcript (exterior context) upstream and double-stranded (stem) context downstream of PUS1 binding (Figure 4A).

LIN28B has two RNA-binding domains: a cold shock domain (CSD) and tandem zinc-binding motifs (zinc fingers [ZFs]). Although LIN28 has a preference for binding GGAGA motifs, target motifs are generally single stranded (Wilbert et al., 2012). NMR spectroscopy suggests that although LIN28 binds stem-rich regions, the CSD binds hairpins and the ZFs bind bulges containing the sequence motif associated with the stem (Nam et al., 2011). These same results were apparent in the binding profile, which showed enrichment for stem, bulge, and hairpin contexts at or proximal to the LIN28 binding site (Figure 4B).

STAU2 binds stretches of base-paired sequences of variable lengths via its three double-stranded RNA-binding domains (dsRBDs) (Ramos et al., 2000). Although the dsRBDs bind tightest to perfectly complementary stem structures, they are able to bind stems that contain bulges (Ramos et al., 2000). Consistent with expectations, the binding profile of STAU2 was strongly enriched for stem context, had slight enrichment for bulge context, and was generally depleted for single-stranded contexts such as hairpin, multibranch, and exterior (Figure 4C).

There is a range of nucleotide resolution derived from the various CLIP-seq techniques. For example, eCLIP provides

30-nucleotide reads surrounding the protein-RNA cross-linking site, whereas better resolution can be achieved with techniques such as individual-nucleotide resolution CLIP (iCLIP) and RNA hybrid iCLIP (hiCLIP) that are able to identify the protein-RNA cross-link site with single-nucleotide resolution. The resolution of nearBynding's profiles reflects the resolution of the input data. For example, by using hiCLIP cross-link sites of STAU1 (Sugimoto et al., 2015), which binds dsRNA similar to STAU2, nearBynding was able to demonstrate that STAU1 contacts single-stranded RNA (ssRNA)—preferably hairpin context, but there was enrichment for all ssRNA contexts at the binding point—but that this ssRNA was directly 3′ of double-stranded (stem) context (Figure 4D). Consistent with our nearBynding analyses, the authors of the hiCLIP data hypothesized that cross-linking sites were enriched at ssRNA because bases within the duplexes are inaccessible for protein-RNA cross-linking (Sugimoto et al., 2015). Further, the cross-linking site was often 3′ of a stem-hairpin-stem structure.

Overall, although there are only a few experimentally confirmed RNA structure binding preferences for us to use as true-positives, nearBynding-generated RBP binding profiles effectively recapitulate documented preferences for RNA structures.

Differentiating WT and mutant RBP structural preferences

Besides investigating WT protein binding relative to null signal, nearBynding can be applied to researching mutant RBPs by comparing WT and mutant protein binding. Although a comparison of WT versus input control depicts the full complement of RBP binding across the transcriptome, a comparison of WT versus a mutant allows visualization of the function-dependent binding of an RBP. For example, binding data are available for the processive RNA helicase UPF1, which is involved in many RNA decay pathways (Kim and Maquat, 2019), as well as for two helicase-dead UPF1 mutants, K498A and DEAA, which are deficient in ATP binding and hydrolysis, respectively (Lee et al., 2015). Both helicase-dead UPF1 mutants retain the ability to bind RNA, but they exhibit a complete loss in target discrimination (Lee et al., 2015). The WT-only UPF1 binding profile and the profiles corrected for helicase-dead mutant signals are all highly symmetrical (i.e., similar RNA structure binding preferences upstream and downstream of binding), but the mutant-corrected profiles indicate a broader span of structure signals (Figures 4E, 4F, and S5A). The mutant-corrected profiles suggest that UPF1 requires helicase activity to occupy stem contexts and select against the unstructured multibranch and exterior contexts.

Illustrate relative binding positions of RBPs

nearBynding can also use RBP binding as a track against another RBP binding track, allowing the user to assess binding preference of one RBP relative to another. To exemplify this functionality, we chose proteins known to occupy the 5′ and 3′ ends of introns and used nearBynding to observe the position and density of their preferred binding relative to each other across unspliced transcripts in K562 cells (Figure S5B). We studied four proteins important for pre-mRNA splicing via their roles in the spliceosome: PRPF8 and RBM22, both of which bind 5′ intronic termini, and BUD13 and U2AF2, which bind 3′ intronic termini (Briese et al., 2019). As expected, PRPF8 and RBM22 colocalized, and BUD13 and U2AF2 also colocalized (Figures S5C and S5D). PRPF8 and RBM22 have comparatively weak, broad signals roughly 100–300 nucleotides upstream of U2AF2 binding (Figure S5D), which corresponds to the $\sim 10^2$ bp intron length of the significant minimal intron peak common in mammalian genomes (JiaYan et al., 2013). Although these tests only reproduce known binding geometry, additional pairwise analyses of RBP binding using nearBynding could provide deeper insights into the arrangements of proteins relative to one another across the transcriptome.

Inform RBP binding preferences using experimentally derived RNA annotations

Analyzing RBP binding to G-quadruplexes

nearBynding is not restricted to *in silico* RNA structure prediction input, so we next interrogated RBP binding profiles with experimentally derived RNA structure data. Guanine-rich RNA sequences can interact via Hoogsteen base pairing and fold into non-canonical structural motifs called G-quadruplexes (G4s) (Brázda et al., 2014). Although many tools are available to predict putative G4s, they are prone to false-positives, since G4 folding is often dependent on the wider context of the RNA sequence and RBP regulation (Beaudoin et al., 2014; Puig Lombardi and Londoño-Vallejo, 2020). We therefore used rG4-seq data (Kwok et al., 2016) to map G4s that form in cells. Although the rG4-seq data were collected from HEK293 cells and ENCODE provided RBP binding data from HepG2 and K562 cells, we reasoned that these cell lines would have enough G4s in common that we could observe general G4-binding trends. Indeed, we observed strong RBP binding at G4s across the exome for multiple published G4-binding proteins such as NONO, FUS, GRSF1, DROSHA, and DDX42 (Figures 5A and S6A) (Pietras et al., 2018; Rouleau et al., 2017; Simko et al., 2020; Yagi et al., 2018; Zyner et al., 2019). Additionally, many of the RBPs that exhibited the strongest G4-binding signal—PRPF4, GTF2F1, FAM120A, CSTF2T, and DDX6—have recently been shown to bind at putative G4 sites in mRNA UTRs (Lee et al., 2020b). Notably, G4-binding profiles are consistent when analyzing RBP binding in exonic and unspliced transcriptomic regions for RBPs that have moderate to strong G4-binding preference (Figure S7). However, some published G4-binding proteins such as FMR1 did not exhibit a robust signal, perhaps due to cell-type-specific variations in binding (Figures 5A and S6A; Table S1). Our analysis also identified RBPs such as YBX3, PRPF8, ZNF800, PPIG, and NOLC1 that are depleted for G4s at their binding sites in HepG2 and K562. These proteins have not previously been documented for their preference against G4 binding, which warrants future investigation.

We hypothesized that our protein-level data can help identify domains that play a role in G4-binding preference. We pooled the exonic HepG2 and K562 data and used the Pfam database (El-Gebali et al., 2019) and protein sequence information to identify protein domains present within the RBPs. Across 13 common protein domains identified, most did not affect G4 binding (Table 2, Figure S6B). RGG repeats are the most common motif in G4-binding RBPs (e.g., FUS) (Brázda et al., 2018) and, based on our analysis, RBPs with RG-rich domains did demonstrate

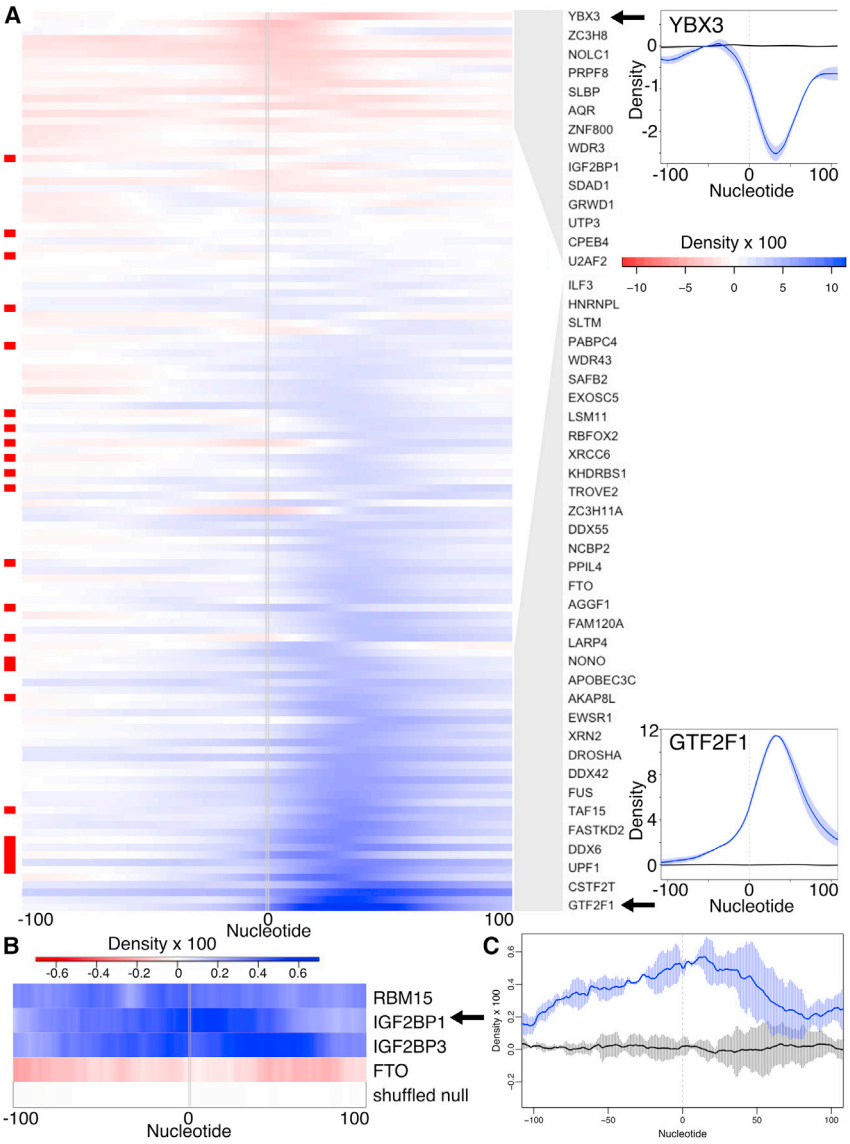


Figure 5. Application of nearBynding to analyze RBP binding profiles for experimentally derived RNA annotations

(A) G4-binding profiles for all 120 proteins with eCLIP exome data in K562 cells. RBPs with molecular evidence of G4 binding in the literature are indicated in red on the left. RBPs with the most positive and most negative correlation signals are highlighted by the gray blocks and listed on the right, with line graph examples to the far right.

(B) m6A binding profiles for RBM15, IGF2BP1, IGF2BP3, FTO, and shuffled null track based on miCLIP-seq data.

(C) Line plot of binding profile for IGF2BP1 and shuffled null track with error bars plus or minus three standard errors.

See also Figures S6 and S7; Table S1.

increased G4 binding. Proteins that contain SAP, dsRBD, or G-patch domains also had increased G4 binding, although there is no literature evidence of this preference. In contrast, RBPs that contain one or more armadillo domains had significantly decreased G4 binding, with six out of eight armadillo-containing proteins demonstrating G4 depletion in their binding preference.

Analyzing RBP preferences for RNA modification

m6A modification is an abundant RNA modification that affects RNA structure (Roost et al., 2015). Since m6A modifications affect RNA folding but are not considered in currently available *in silico* folding algorithms, we tested whether miCLIP-seq data could be used as an input for nearBynding to observe protein-binding contexts relative to m6A modification. Multiple RBPs are involved in the writing, reading, and erasing of m6A, such as RBM15, IGF2BP1/3, and FTO, respectively (Huang et al., 2018; Patil et al., 2016; Yu et al., 2018). We used miCLIP-seq data (Huang et al., 2019) and eCLIP data (Davis et al., 2018) from HepG2 to determine whether these m6A-interacting RBPs show binding preferences relative to m6A modification. As expected, RBM15, IGF2BP1, and IGF2BP3 all demonstrated a preference for binding m6A-modified RNA (Figure 5B). In contrast, FTO did not preferentially occupy m6A-modified regions of the transcriptome, perhaps reflecting its role as an m6A eraser. Despite their modest density amplitudes, likely due to a small signal to noise ratio in the miCLIP-seq data, these signals are significant given a plus or minus three standard error confidence interval above the shuffled null track (Figure 5C). These reproductions of previously observed results demonstrate the diversity of data types that can accurately inform RBP binding contexts using nearBynding.

DISCUSSION

Our analyses revealed that the concordance of RBP binding profiles across cell lines were lower than biological replicates of the same cell lines. This difference may be because (1) the eCLIP data are cleaner for one cell type; (2) the transcripts expressed and therefore available for binding differ; (3) there is differential expression of competitive binders, modifiers, or cofactors; or (4) the RNA folds differently between cell types. Prior analysis of these datasets has also uncovered cell-type-specific differences in preferred transcript region binding for some of the RBPs interrogated (Van Nostrand et al., 2020a). Literature precedents indicate many examples of RBPs that shift their binding preferences as a result of post-translational modifications or cofactor binding (Schmidt et al., 2017; Timchenko et al., 2006). For some RBPs, their cellular localization and therefore binding opportunities rely on cofactor binding (Heininger et al., 2016), whereas others' binding opportunities depend on the expression of competitive binders (Liu et al., 2015). Variations in protein expression and post-translational

Table 2. Influence of RBP domains on G4-binding signal

Domain	G4 signal difference	Effect size	p value
Armadillo	−2.41	0.913	0.00147
RGG-repeat	1.57	0.592	0.00612
SAP	0.836	0.312	0.0123
dsRBD	1.57	0.589	0.0219
G-patch	1.77	0.662	0.0385
Alpha-beta	0.476	0.178	0.198
Winged-helix	1.88	0.707	0.238
RRM	0.385	0.144	0.301
K-homology	−0.384	0.143	0.402
Helicase	0.428	0.16	0.436
WD40	−0.479	0.179	0.66
P loop	0.0385	0.0143	0.935
Zinc finger	0.000907	0.000338	0.999

Statistics from pooled HepG2 and K562 binding profiles: the difference in mean G4-binding signal between proteins with and without the indicated domain; Cohen's *d* effect size; and the p value of a t test comparing G4 signal of proteins with and without the indicated domain. See also Figure S6.

modification frequencies across cell lines may therefore drive differences in protein binding profiles. Molecular validation would be required to examine these intriguing cell-specific binding preferences for an RBP; in the absence of this proof, we suggest using signal similarities across replicates and cell types to bolster confidence in predictions of structure binding preference for a given RBP.

A statistically significant cross-correlation signal between RBP binding and RNA structures implies that an RBP binds that specific structure, but there could be an alternative explanation. It is possible that these RBP-associated sequences are prone to adopt a particular RNA structure only when it is not bound by the RBP. DROSHA, for example, binds G4-forming regions only when these regions are unfolded (Rouleau et al., 2017). Because many G4-forming sequences are actively unfolded *in vivo*, we cannot differentiate without further molecular experimentation whether an RBP binds to G4s or RBP-associated sequences are prone to forming G4s. We speculate that a phenomenon similar to DROSHA's binding drove the enrichment of dsRBD-containing RBPs among the higher G4 signals (Table 2), since G4-forming sequences are necessarily GC rich and likely form stable regions of dsRNA. Biochemical experimentation such as kinetics assays or crystal structures is necessary to definitively ascertain RBP binding.

nearBynding is able to process continuous datatypes, such as the probability of RNA structure and amplitudes of aligned CLIP-seq reads. Literature suggests modeling RBP binding as a list of bound regions across the transcriptome provides only a coarse approximation of RBP binding motifs (Maticzka et al., 2014), and therefore this nuanced read amplitude information may enable us to identify preferred RNA structure motifs based on RBP binding frequency. Although some differences may exist in RBP binding profiles when called peaks or aligned reads were used, the differences for distinct RBPs are far greater than for the same RBP interrogated using different inputs. These analyses demonstrate the similarity in results between interval and continuous datatypes, allowing for the possibility of omitting the step of peak calling for RBP binding analysis.

Most state-of-the-art algorithms that incorporate RNA structure into predictions of RBP binding motifs rely on RNA sequence alone to predict RNA secondary structure (Guo and Bartel, 2016). Similarly, nearBynding can call CapR to predict RNA structures. All these algorithms, however, assume that the mRNA being folded is naked and unmodified, with only the queried RBP binding it. nearBynding provides the flexibility for users to input an even broader range of experimentally derived RNA structure information, which could be used to study the binding of non-canonical RNA structures (e.g., G4s, triple helices) and RNA modifications (e.g., A-to-I editing, m6A, or N^4-acetylcytidine). In addition, users can improve the study of canonical RNA structure binding by incorporating structural information collected via chemical probing (e.g., SHAPE or dimethyl sulfate [DMS]).

Given that nearBynding allows for the comparisons of various interval or continuous features across user-selected regions of the transcriptome, this algorithm may be adapted to correlate many other transcriptome-related features. Users can study the binding of an RBP relative to RNA structure, one RBP relative to another, an RNA modification relative to RNA structure, or any pair of interval or continuous features so long as they can be annotated on a transcriptome. Further, mutant and WT data can be directly compared to understand how genetic changes affect elements such as RBP binding preferences or RNA modifications. Future work will take advantage of this flexibility to characterize RBP complexes relative to RNA structure via wider-ranging approaches than previously possible.

Limitations

nearBynding is only able to provide aggregate information about RBP binding preferences, in contrast to other available machine-learning tools that predict individual binding events across the transcriptome (Sun et al., 2021) or RBP binding to alternative RNA structures (Tomezsko et al., 2020). Another complexity of transcriptomic data is that a single gene may have multiple transcribed variants that are overlapping on a traditional genomic scale. Since nearBynding does not allow redundant mapping of data (e.g., one RBP binding event in an overlapping region of two variants cannot be duplicated in the generation of the transcriptome), it is unable to accommodate two or more variants from the same gene if the queried regions of the transcripts overlap.

The current software for nearBynding does not accommodate data from *in vitro* experiments of proteins bound to RNA oligos such as from SELEX or RNA Bind-N-Seq, since these methods use RNA sequences that do not correspond to transcripts that are mappable to the genome (Tuerk and Gold, 1990; Zykovich et al., 2009). To correlate *in vitro* data via nearBynding, the user would need to create a novel annotated genome containing sequences for every oligonucleotide probe in the queried experiment; however, the current pipeline could process a novel annotated genome without further customization. While nearBynding accepts two transcriptome-wide feature inputs for correlation,

the software currently only supports the consideration of replicates and background signal for one of these inputs, such as for an input control in a CLIP experiment. Future work will support the possibility of accommodating replicates and removing background signal for both input features, such as for analyzing RBP binding to RNA data derived from an RNA immunoprecipitation (RIP) experiment, which use antibodies targeting RNA structures or modifications (Huang et al., 2018; Park et al., 2019).

The resolution of the nearBynding output is limited by the lower-resolution input data track, so an output may imprecisely depict the correlation between one narrow data type (e.g., nucleotide sequence, which has functional peak widths of one nucleotide) and one broader data type (e.g., CLIP peaks). However, in cases of similar-width data types (e.g., single-nucleotide RBP binding information such as iCLIP or hiCLIP; Huppertz et al., 2014; Sugimoto et al., 2015; see Figure 4D), it is possible to, for example, assess RNA sequence preferences relative to RBP binding with single-nucleotide resolution. Current work has not probed RNA sequence correlations, but future work will integrate RNA sequence as a feature by separating each of the four nucleotides into individual RNA tracks, similar to how CapR separates six RNA structures into different tracks.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - nearBynding inputs
  - Map data to pseudochromosomes
  - CapR RNA structure prediction
  - Relative binding position calculation
  - nearBynding output analyses
  - Comparison of available spatial relationship algorithms
  - deepTools
  - BEDtools
  - StereoGene
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.crmeth.2021.100088.

ACKNOWLEDGMENTS

We thank Leung lab members, Fertig lab members, and Tim Nieuwenhuis for ongoing suggestions for the project; Kurt Weir for discussion and reading the manuscript; Arjun Bhutkar for the suggestion to use Wasserstein distances; Eric Van Nostrand for providing processed data from Van Nostrand et al. (2020b); and Andrey Miranov for technical advice on StereoGene. This work was supported by Johns Hopkins Bloomberg School of Public Health start-up fund and the Johns Hopkins President's Frontier Award program to A.K.L.L., the Predoctoral Fellowship in Informatics from the Pharmaceutical Research and Manufacturers of America Foundation to V.F.B., National Institutes of Health (T32GM07814 to V.F.B. and P50CA062924 and P30CA006973 to E.J.F.), and the Johns Hopkins University Discovery Award to A.K.L.L. and E.J.F. Funding for open access charge: Johns Hopkins President's Frontier Award program and Johns Hopkins University Discovery Award.

AUTHOR CONTRIBUTIONS

All authors were involved in conceptualization and methodology development of the project. V.F.B. wrote the software and performed all analyses. A.V.F. contributed coding suggestions to and reviewed the pipeline. V.F.B. wrote the original draft and prepared visuals. All authors reviewed and edited the manuscript. A.V.F., A.K.L.L., and E.J.F. provided supervision and guidance throughout the project. All authors read and approved the final manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: December 15, 2020
Revised: June 23, 2021
Accepted: August 30, 2021
Published: October 1, 2021

REFERENCES

Bahrami-Samani, E., Penalva, L.O.F., Smith, A.D., and Uren, P.J. (2015). Leveraging cross-link modification events in CLIP-seq for motif discovery. Nucleic Acids Res. *43*, 95–103.

Beaudoin, J.-D., Jodoin, R., and Perreault, J.-P. (2014). New scoring system to identify RNA G-quadruplex folding. Nucleic Acids Res. *42*, 1209–1223.

Beltran, M., Tavares, M., Justin, N., Khandelwal, G., Ambrose, J., Foster, B.M., Worlock, K.B., Tvardovskiy, A., Kunzelmann, S., Herrero, J., et al. (2019). G-tract RNA removes Polycomb repressive complex 2 from genes. Nat. Struct. Mol. Biol. *26*, 899–909.

Blaszczyk, J., Gan, J., Tropea, J.E., Court, D.L., Waugh, D.S., and Ji, X. (2004). Noncatalytic assembly of ribonuclease III with double-stranded RNA. Structure *12*, 457–466.

Brázda, V., Hároníková, L., Liao, J.C.C., and Fojta, M. (2014). DNA and RNA quadruplex-binding proteins. Int. J. Mol. Sci. *15*, 17493–17517.

Brázda, V., Červeň, J., Bartas, M., Mikysková, N., Coufal, J., and Pečinka, P. (2018). The amino acid composition of quadruplex binding proteins reveals a shared motif and predicts new potential quadruplex interactors. Molecules *23*. https://doi.org/10.3390/molecules23092341.

Briese, M., Haberman, N., Sibley, C.R., Faraway, R., Elser, A.S., Chakrabarti, A.M., Wang, Z., König, J., Perera, D., Wickramasinghe, V.O., et al. (2019). A systems view of spliceosomal assembly and branchpoints with iCLIP. Nat. Struct. Mol. Biol. *26*, 930–940.

Carlile, T.M., Martinez, N.M., Schaening, C., Su, A., Bell, T.A., Zinshteyn, B., and Gilbert, W.V. (2019). mRNA structure determines modification by pseudouridine synthase 1. Nat. Chem. Biol., 1–9.

Chen, X., Castro, S.A., Liu, Q., Hu, W., and Zhang, S. (2019). Practical considerations on performing and analyzing CLIP-seq experiments to identify transcriptomic-wide RNA-protein interactions. Methods *155*, 49–57.

Davis, C.A., Hitz, B.C., Sloan, C.A., Chan, E.T., Davidson, J.M., Gabdank, I., Hilton, J.A., Jain, K., Baymuradov, U.K., Narayanan, A.K., et al. (2018). The Encyclopedia of DNA elements (ENCODE): data portal update. Nucleic Acids Res. *46*, D794–D801.

Drewe-Boss, P., Wessels, H.-H., and Ohler, U. (2018). omniCLIP: probabilistic identification of protein-RNA interactions from CLIP-seq data. Genome Biol. *19*, 183.

Edupuganti, R.R., Geiger, S., Lindeboom, R.G.H., Shi, H., Hsu, P.J., Lu, Z., Wang, S.-Y., Baltissen, M.P.A., Jansen, P.W.T.C., Rossa, M., et al. (2017).

N6-methyladenosine (m6A) recruits and repels proteins to regulate mRNA homeostasis. Nat. Struct. Mol. Biol. *24*, 870–878.

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S.R., Luciani, A., Potter, S.C., Qureshi, M., Richardson, L.J., Salazar, G.A., Smart, A., et al. (2019). The Pfam protein families database in 2019. Nucleic Acids Res. *47*, D427–D432.

Favorov, A., Mularoni, L., Cope, L.M., Medvedeva, Y., Mironov, A.A., Makeev, V.J., and Wheelan, S.J. (2012). Exploring massive, genome scale datasets with the GenometriCorr package. PLoS Comput. Biol. *8*. https://doi.org/10.1371/journal.pcbi.1002529.

Fukunaga, T., Ozaki, H., Terai, G., Asai, K., Iwasaki, W., and Kiryu, H. (2014). CapR: revealing structural specificities of RNA-binding protein target recognition using CLIP-seq data. Genome Biol. *15*, R16.

Guo, J.U., and Bartel, D.P. (2016). RNA G-quadruplexes are globally unfolded in eukaryotic cells and depleted in bacteria. Science *353*. https://doi.org/10.1126/science.aaf5371.

Heininger, A.U., Hackert, P., Andreou, A.Z., Boon, K.-L., Memet, I., Prior, M., Clancy, A., Schmidt, B., Urlaub, H., Schleiff, E., et al. (2016). Protein cofactor competition regulates the action of a multifunctional RNA helicase in different pathways. RNA Biol. *13*, 320–330.

Heller, D., Krestel, R., Ohler, U., Vingron, M., and Marsico, A. (2017). ssHMM: extracting intuitive sequence-structure motifs from high-throughput RNA-binding protein data. Nucleic Acids Res. *45*, 11004–11018.

Howe, K.L., Contreras-Moreira, B., De Silva, N., Maslen, G., Akanni, W., Allen, J., Alvarez-Jarreta, J., Barba, M., Bolser, D.M., Cambell, L., et al. (2020). Ensembl Genomes 2020—enabling non-vertebrate genomic research. Nucleic Acids Res. *48*, D689–D695.

Huang, H., Weng, H., Sun, W., Qin, X., Shi, H., Wu, H., Zhao, B.S., Mesquita, A., Liu, C., Yuan, C.L., et al. (2018). Recognition of RNA N6-methyladenosine by IGF2BP proteins enhances mRNA stability and translation. Nat. Cell Biol. *20*, 285–295.

Huang, H., Weng, H., Zhou, K., Wu, T., Zhao, B.S., Sun, M., Chen, Z., Deng, X., Xiao, G., Auer, F., et al. (2019). Histone H3 trimethylation at lysine 36 guides m6A RNA modification co-transcriptionally. Nature *567*, 414–419.

Huppertz, I., Attig, J., D'Ambrogio, A., Easton, L.E., Sibley, C.R., Sugimoto, Y., Tajnik, M., König, J., and Ule, J. (2014). iCLIP: protein–RNA interactions at nucleotide resolution. Methods *65*, 274–287.

Ingolia, N.T., Ghaemmaghami, S., Newman, J.R.S., and Weissman, J.S. (2009). Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science *324*, 218–223.

Jarmoskaite, I., Denny, S.K., Vaidyanathan, P.P., Becker, W.R., Andreasson, J.O.L., Layton, C.J., Kappel, K., Shivashankar, V., Sreenivasan, R., Das, R., et al. (2019). A quantitative and predictive model for RNA binding by human pumilio proteins. Mol. Cell *74*, 966–981.e18.

JiaYan, W., JingFa, X., LingPing, W., Jun, Z., HongYan, Y., ShuangXiu, W., Zhang, Z., and Jun, Y. (2013). Systematic analysis of intron size and abundance parameters in diverse lineages. Sci. China Life Sci. *56*, 968–974.

Kim, Y.K., and Maquat, L.E. (2019). UPFront and center in RNA decay: UPF1 in nonsense-mediated mRNA decay and beyond. RNA *25*, 407–422.

Kravatsky, Y.V., Chechetkin, V.R., Tchurikov, N.A., and Kravatskaya, G.I. (2015). Genome-wide study of correlations between genomic features and their relationship with the regulation of gene expression. DNA Res. *22*, 109–119.

Kwok, C.K., Marsico, G., Sahakyan, A.B., Chambers, V.S., and Balasubramanian, S. (2016). rG4-seq reveals widespread formation of G-quadruplex structures in the human transcriptome. Nat. Methods *13*, 841–844.

Lee, C.M., Barber, G.P., Casper, J., Clawson, H., Diekhans, M., Gonzalez, J.N., Hinrichs, A.S., Lee, B.T., Nassar, L.R., Powell, C.C., et al. (2020a). UCSC Genome Browser enters 20th year. Nucleic Acids Res. *48*, D756–D761.

Lee, D.S.M., Ghanem, L.R., and Barash, Y. (2020b). Integrative analysis reveals RNA G-quadruplexes in UTRs are selectively constrained and enriched for functional associations. Nat. Commun. *11*, 1–12.

Lee, S.R., Pratt, G., Martinez, F., Yeo, G.W., and Lykke-Andersen, J. (2015). Target discrimination in nonsense-mediated mRNA decay requires Upf1 ATPase activity. Mol. Cell *59*, 413–425.

Li, X., Quon, G., Lipshitz, H.D., and Morris, Q. (2010). Predicting in vivo binding sites of RNA-binding proteins using mRNA secondary structure. RNA *16*, 1096–1107.

Li, X., Kazan, H., Lipshitz, H.D., and Morris, Q.D. (2014). Finding the target sites of RNA-binding proteins. Wiley Interdiscip. Rev. RNA *5*, 111–130.

Liu, L., Ouyang, M., Rao, J.N., Zou, T., Xiao, L., Chung, H.K., Wu, J., Donahue, J.M., Gorospe, M., and Wang, J.-Y. (2015). Competition between RNA-binding proteins CELF1 and HuR modulates MYC translation and intestinal epithelium renewal. Mol. Biol. Cell *26*, 1797–1810.

Lovci, M.T., Ghanem, D., Marr, H., Arnold, J., Gee, S., Parra, M., Liang, T.Y., Stark, T.J., Gehman, L.T., Hoon, S., et al. (2013). Rbfox proteins regulate alternative mRNA splicing through evolutionarily conserved RNA bridges. Nat. Struct. Mol. Biol. *20*, 1434–1442.

Lucks, J.B., Mortimer, S.A., Trapnell, C., Luo, S., Aviran, S., Schroth, G.P., Pachter, L., Doudna, J.A., and Arkin, A.P. (2011). Multiplexed RNA structure characterization with selective 2′-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-seq). Proc. Natl. Acad. Sci. U S A *108*, 11063–11068.

Luo, E.-C., Nathanson, J.L., Tan, F.E., Schwartz, J.L., Schmok, J.C., Shankar, A., Markmiller, S., Yee, B.A., Sathe, S., Pratt, G.A., et al. (2020). Large-scale tethered function assays identify factors that regulate mRNA stability and translation. Nat. Struct. Mol. Biol. *27*, 1–12.

Maticzka, D., Lange, S.J., Costa, F., and Backofen, R. (2014). GraphProt: modeling binding preferences of RNA-binding proteins. Genome Biol. *15*, R17.

Matunis, M.J., Michael, W.M., and Dreyfuss, G. (1992). Characterization and primary structure of the poly(C)-binding heterogeneous nuclear ribonucleoprotein complex K protein. Mol. Cell Biol. *12*, 164–171.

Nam, Y., Chen, C., Gregory, R.I., Chou, J.J., and Sliz, P. (2011). Molecular basis for interaction of let-7 microRNAs with Lin28. Cell *147*, 1080–1091.

Navarro, D. (2015). Learning Statistics with R: A Tutorial for Psychology Students and Other Beginners. (Version 0.5) (University of Adelaide).

Okada, S., Ueda, H., Noda, Y., and Suzuki, T. (2019). Transcriptome-wide identification of A-to-I RNA editing sites using ICE-seq. Methods *156*, 66–78.

Ozdilek, B.A., Thompson, V.F., Ahmed, N.S., White, C.I., Batey, R.T., and Schwartz, J.C. (2017). Intrinsically disordered RGG/RG domains mediate degenerate specificity in RNA binding. Nucleic Acids Res. *45*, 7984–7996.

Pan, X., and Shen, H.-B. (2018). Predicting RNA–protein binding sites and motifs through combining local and global deep convolutional neural networks. Bioinformatics *34*, 3427–3436.

Park, O.H., Ha, H., Lee, Y., Boo, S.H., Kwon, D.H., Song, H.K., and Kim, Y.K. (2019). Endoribonucleolytic cleavage of m6A-containing RNAs by RNase P/MRP complex. Mol. Cell *74*, 494–507.e8.

Patil, D.P., Chen, C.-K., Pickering, B.F., Chow, A., Jackson, C., Guttman, M., and Jaffrey, S.R. (2016). m6A RNA methylation promotes XIST-mediated transcriptional repression. Nature *537*, 369–373.

Pietras, Z., Wojcik, M.A., Borowski, L.S., Szewczyk, M., Kulinski, T.M., Cysewski, D., Stepien, P.P., Dziembowski, A., and Szczesny, R.J. (2018). Dedicated surveillance mechanism controls G-quadruplex forming non-coding RNAs in human mitochondria. Nat. Commun. *9*, 2558. https://doi.org/10.1038/s41467-018-05007-9.

Puig Lombardi, E., and Londoño-Vallejo, A. (2020). A guide to computational methods for G-quadruplex prediction. Nucleic Acids Res. *48*, 1–15.

Quinlan, A.R. (2014). BEDTools: the Swiss-army tool for genome feature analysis. Curr. Protoc. Bioinformatics *47*, 11.12.1–11.12.34.

Ramos, A., Grünert, S., Adams, J., Micklem, D.R., Proctor, M.R., Freund, S., Bycroft, M., St Johnston, D., and Varani, G. (2000). RNA recognition by a Staufen double-stranded RNA-binding domain. EMBO J. *19*, 997–1009.

Roost, C., Lynch, S.R., Batista, P.J., Qu, K., Chang, H.Y., and Kool, E.T. (2015). Structure and thermodynamics of N6-methyladenosine in RNA: a spring-loaded base modification. J. Am. Chem. Soc. *137*, 2107–2115.

Rouleau, S.G., Garant, J.-M., Bolduc, F., Bisaillon, M., and Perreault, J.-P. (2017). G-quadruplexes influence pri-microRNA processing. RNA Biol. *15*, 198–206.

Sasse, A., Laverty, K.U., Hughes, T.R., and Morris, Q.D. (2018). Motif models for RNA-binding proteins. Curr. Opin. Struct. Biol. *53*, 115–123.

Sauer, M., Juranek, S.A., Marks, J., Magis, A.D., Kazemier, H.G., Hilbig, D., Benhalevy, D., Wang, X., Hafner, M., and Paeschke, K. (2019). DHX36 prevents the accumulation of translationally inactive mRNAs with G4-structures in untranslated regions. Nat. Commun. *10*, 1–15.

Schmidt, T., Knick, P., Lilie, H., Friedrich, S., Golbik, R.P., and Behrens, S.-E. (2017). The properties of the RNA-binding protein NF90 are considerably modulated by complex formation with NF45. Biochem. J. *474*, 259–280.

Schuhmacher, D., Bähre, B., Bonneel, N., Gottschlich, C., Hartmann, V., Heinemann, F., Schmitzer, B., Schrieber, J., and Wilm, T. (2020). Transport: Computation of Optimal Transport Plans and Wasserstein Distances (CRAN).

Simko, E.A.J., Liu, H., Zhang, T., Velasquez, A., Teli, S., Haeusler, A.R., and Wang, J. (2020). G-quadruplexes offer a conserved structural motif for NONO recruitment to NEAT1 architectural lncRNA. Nucleic Acids Res. *48*, 7421–7438. https://doi.org/10.1093/nar/gkaa475.

Stavrovskaya, E.D., Niranjan, T., Fertig, E.J., Wheelan, S.J., Favorov, A.V., and Mironov, A.A. (2017). StereoGene: rapid estimation of genome-wide correlation of continuous or interval feature data. Bioinformatics *33*, 3158–3165.

Sugimoto, Y., König, J., Hussain, S., Zupan, B., Curk, T., Frye, M., and Ule, J. (2012). Analysis of CLIP and iCLIP methods for nucleotide-resolution studies of protein-RNA interactions. Genome Biol. *13*, R67.

Sugimoto, Y., Vigilante, A., Darbo, E., Zirra, A., Militti, C., D'Ambrogio, A., Luscombe, N.M., and Ule, J. (2015). hiCLIP reveals the in vivo atlas of mRNA secondary structures recognized by Staufen 1. Nature *519*, 491–494.

Sun, L., Xu, K., Huang, W., Yang, Y.T., Li, P., Tang, L., Xiong, T., and Zhang, Q.C. (2021). Predicting dynamic cellular protein–RNA interactions by deep learning using in vivo RNA structures. Cell Res., 1–22.

Sun, Z., Xue, S., Xu, H., Hu, X., Chen, S., Yang, Z., Yang, Y., Ouyang, J., and Cui, H. (2018). Effects of NSUN2 deficiency on the mRNA 5-methylcytosine modification and gene expression profile in HEK293 cells. Epigenomics *11*, 439–453.

Taliaferro, J.M., Lambert, N.J., Sudmant, P.H., Dominguez, D., Merkin, J.J., Alexis, M.S., Bazile, C., and Burge, C.B. (2016). RNA sequence context effects measured in vitro predict in vivo protein binding and regulation. Mol. Cell *64*, 294–306.

Timchenko, L.T., Salisbury, E., Wang, G.-L., Nguyen, H., Albrecht, J.H., Hershey, J.W.B., and Timchenko, N.A. (2006). Age-specific CUGBP1-eIF2 complex increases translation of CCAAT/Enhancer-binding protein β in old liver. J. Biol. Chem. *281*, 32806–32819.

Tomezsko, P.J., Corbin, V.D.A., Gupta, P., Swaminathan, H., Glasgow, M., Persad, S., Edwards, M.D., Mcintosh, L., Papenfuss, A.T., Emery, A., et al. (2020). Determination of RNA structural diversity and its role in HIV-1 RNA splicing. Nature *582*, 438–442.

Tuerk, C., and Gold, L. (1990). Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. Science *249*, 505–510.

Uren, P.J., Bahrami-Samani, E., Burns, S.C., Qiao, M., Karginov, F.V., Hodges, E., Hannon, G.J., Sanford, J.R., Penalva, L.O.F., and Smith, A.D. (2012). Site identification in high-throughput RNA–protein interaction data. Bioinformatics *28*, 3013–3020.

Van Nostrand, E.L., Freese, P., Pratt, G.A., Wang, X., Wei, X., Xiao, R., Blue, S.M., Chen, J.-Y., Cody, N.A.L., Dominguez, D., et al. (2020a). A large-scale binding and functional map of human RNA-binding proteins. Nature *583*, 711–719.

Van Nostrand, E.L., Pratt, G.A., Yee, B.A., Wheeler, E.C., Blue, S.M., Mueller, J., Park, S.S., Garcia, K.E., Gelboin-Burkhart, C., Nguyen, T.B., et al. (2020b). Principles of RNA processing from analysis of enhanced CLIP maps for 150 RNA binding proteins. Genome Biol. *21*, 90.

Waldron, J.A., Tack, D.C., Ritchey, L.E., Gillen, S.L., Wilczynska, A., Turro, E., Bevilacqua, P.C., Assmann, S.M., Bushell, M., and Le Quesne, J. (2019). mRNA structural elements immediately upstream of the start codon dictate dependence upon eIF4A helicase activity. Genome Biol. *20*, 300.

Wang, X., McLachlan, J., Zamore, P.D., and Hall, T.M.T. (2002). Modular recognition of RNA by a human pumilio-homology domain. Cell *110*, 501–512.

Wilbert, M.L., Huelga, S.C., Kapeli, K., Stark, T.J., Liang, T.Y., Chen, S.X., Yan, B.Y., Nathanson, J.L., Hutt, K.R., Lovci, M.T., et al. (2012). LIN28 binds messenger RNAs at GGAGA motifs and regulates splicing factor abundance. Mol. Cell *48*, 195–206.

Wolfe, M.B., Schagat, T.L., Paulsen, M.T., Magnuson, B., Ljungman, M., Park, D., Zhang, C., Campbell, Z.T., Goldstrohm, A.C., and Freddolino, P.L. (2020). Principles of mRNA control by human PUM proteins elucidated from multimodal experiments and integrative data analysis. RNA *26*, 1680–1703, 077362.120.

Yagi, R., Miyazaki, T., and Oyoshi, T. (2018). G-quadruplex binding ability of TLS/FUS depends on the β-spiral structure of the RGG domain. Nucleic Acids Res. *46*, 5894–5901.

Yan, Z., Hamilton, W.L., and Blanchette, M. (2020). Graph neural representational learning of RNA secondary structures for predicting RNA-protein interactions. Bioinformatics *36*, i276–i284.

Yu, J., Chen, M., Huang, H., Zhu, J., Song, H., Zhu, J., Park, J., and Ji, S.-J. (2018). Dynamic m6A modification regulates local translation of mRNA in axons. Nucleic Acids Res. *46*, 1412–1423.

Zhang, Z., and Xing, Y. (2017). CLIP-seq analysis of multi-mapped reads discovers novel functional RNA regulatory sites in the human transcriptome. Nucleic Acids Res. *45*, 9260–9271.

Zhang, Y., Liu, H., Lv, J., Xiao, X., Zhu, J., Liu, X., Su, J., Li, X., Wu, Q., Wang, F., et al. (2011). QDMR: a quantitative method for identification of differentially methylated regions by entropy. Nucleic Acids Res. *39*, e58.

Zykovich, A., Korf, I., and Segal, D.J. (2009). Bind-n-Seq: high-throughput analysis of in vitro protein–DNA interactions using massively parallel sequencing. Nucleic Acids Res. *37*, e151.

Zyner, K.G., Mulhearn, D.S., Adhikari, S., Martínez Cuesta, S., Di Antonio, M., Erard, N., Hannon, G.J., Tannahill, D., and Balasubramanian, S. (2019). Genetic interactions of G-quadruplexes in humans. eLife *8*, e46793.

CellPress
OPEN ACCESS

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
ENCODE eCLIP, ChIP, and RNA-seq	(Davis et al., 2018)	https://www.encodeproject.org/
iMaps STAU1 hiCLIP	(Sugimoto et al., 2015)	https://imaps.genialis.com/iclip/search/collection/hi-clip-reveals-m-rna-secondary-structures
UPF1 WT, K498A, and DEAA CLIP-seq	(Lee et al., 2015)	GEO: GSE69586
Ensembl release 100 FASTA and GTF files	(Howe et al., 2020)	ftp://ftp.ensembl.org/pub/release-100/
UCSC hg38 chain file	(Lee et al., 2020a)	http://hgdownload.soe.ucsc.edu/goldenPath/hg19/liftOver/hg19ToHg38.over.chain.gz
Pfam protein domains	(El-Gebali et al., 2019)	http://pfam.xfam.org/
RGG-repeat-containing proteins list	(Ozdilek et al., 2017)	Supplemental information
HepG2 miCLIP-seq	(Huang et al., 2019)	GEO: GSE121942
HEK293 rG4-seq	(Kwok et al., 2016)	GEO: GSE77282
HEK293 RNA-seq	(Sun et al., 2018)	GEO: GSE122425
Software and algorithms		
nearBynding	This paper	https://doi.org/10.5281/zenodo.5176831 https://doi.org/10.5281/zenodo.5176827
CapR	(Fukunaga et al., 2014)	https://github.com/fukunagatsu/CapR
StereoGene	(Stavrovskaya et al., 2017)	http://stereogene.bioinf.fbb.msu.ru/
BEDtools	(Quinlan, 2014)	https://bedtools.readthedocs.io/en/latest/
Piranha	(Uren et al., 2012)	http://smithlabresearch.org/software/piranha/
deepTools	(Ramírez et al., 2016)	https://deeptools.readthedocs.io/en/develop/index.html#

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Anthony Leung (anthony.leung@jhu.edu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- This paper analyzes existing, publicly available data. Accession numbers for the datasets are listed in the key resources table.
- All original code has been deposited at Zenodo and is publicly available as of the date of publication. DOIs are listed in the key resources table.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

nearBynding inputs

In order to provide predicted RNA structure context for RBP binding, nearBynding requires the following pieces of input data: (1) CLIP-seq alignment tracks for the RBP of interest, (2) an annotated genome and associated FASTA sequence, and (3) a list of transcripts of interest. It is recommended that all transcripts selected are expressed in the cell type used for the CLIP-seq experiment. Alternative RNA structure information can optionally be included, and it is recommended that the data is derived from the same cell type. All data must be converted to bedGraph format, and the functions *CleanBEDtoBG* and *CleanBAMtoBG* help users do this.

Map data to pseudochromosomes

Users must first choose which regions of the transcripts of interest to interrogate (e.g. UTRs, exons, whole transcript), based on annotations available in the genome GTF file. nearBynding creates (1) a chain file that will be used to map the selected regions of transcripts end-to-end, excluding the intergenic regions and undesired transcripts that comprise the majority of the genome, via the function *GenomeMappingtoChainFile* and (2) a file detailing the names and sizes of all the chain file's pseudochromosomes via the function *getChainChrSize*. The chain file can then be used to transfer genome references of the CLIP-seq data from the whole-genome alignment to the transcriptome alignment of interest via *liftOverToExomicBG*. If external RNA structure data is being studied, its genome references would need to be transferred to the transcriptome alignment as well using the same chain file. Chain files cannot accommodate overlapping intervals since they cause ambiguous regions in the transfer process, so it is recommended that users supply the highest-expressed isoform of every gene expressed in the cell type of the CLIP-seq data to create the concatenated pseudochromosomes.

CapR RNA structure prediction

nearBynding pulls the sequences of selected regions of transcripts of interest from the genome FASTA file based on genome annotations using BEDtools (Quinlan, 2014) wrapped in the function *ExtractTranscriptomeSequence*. Probabilistic RNA structure for the selected regions are derived from *in silico* folding predictions by CapR, which includes RNAfold software in its structure predictions (Fukunaga et al., 2014), via the function *runCapR*. Each nucleotide is scored as having a 0 to 1 probability of adopting one of six different contexts by CapR. The data for the six different folding conformations are then separated and the transcript fragments are concatenated into pseudo-chromosomes via *processCapRout*. In secondary structure representation, RNA bases are depicted as vertices of polygons with edges of RNA backbone or hydrogen bonds (Figure 1B). The six different RNA structure contexts are defined thus: stem context is if a base participates in hydrogen-bonding with another base; exterior context is if a base does not form a polygon such as that the end of a transcript; hairpin context is if a single-stranded base is involved in a polygon with only one hydrogen-bonding edge; bulge context is if a single-stranded base is involved in a polygon with two hydrogen-bonding edges and where all stem context vertices are contiguous in the polygon; internal context is if a single-stranded base is involved in a polygon with two hydrogen-bonding edges and where stem context vertices are not contiguous in the polygon; multibranch context is if a single-stranded base is involved in a polygon with at least three hydrogen-bonding edges (Fukunaga et al., 2014).

Relative binding position calculation

To visualize the RNA structure landscape surrounding protein binding, StereoGene (Stavrovskaya et al., 2017), wrapped within the functions *runStereogene* and *runStereogeneOnCapR*, is used to calculate the cross-correlation between RNA structure and protein binding. The *get_error* argument for *runStereogene* and *runStereogeneOnCapR* allows users to also get the standard error for the shuffled null track, and the *nShuffle* argument determines how many times the null track is shuffled; calculating the shuffled null standard error is optional because it requires substantially more computational time, especially in cases where many shuffling iterations are requested. nearBynding analyzes structure colocalization in single-nucleotide frames, which sacrifices some of the computational efficiency of StereoGene but maximizes RBP binding resolution. Cross-correlation densities are within the range -1 to +1, where -1 suggests perfect depletion of an RBP for a tested RNA structure context, 0 represents no binding preference, and +1 suggests perfect RBP binding for a tested RNA structure context. Since actual experimental correlation densities are far more modest, they are reported as density x 100 for visualization, which is conducted using the functions *visualizeStereogene* and *visualizeCapRStereogene*.

nearBynding output analyses

nearBynding allows users to calculate the similarity between output binding profiles via Wasserstein distance via *bindingContextDistance* and *bindingContextDistanceCapR*, where small values indicate greater similarity. Users can also assess the grouping of different categories of points via bootstrapping and the Kolmogorov–Smirnov test (Figure S3C) using *assessGrouping*.

Comparison of available spatial relationship algorithms

Multiple tools are available to evaluate the spatial relationships across genomic data types to draw conclusions about biological interactions (e.g. histone modifications and transcription start sites; CpG islands and gene promoters; splice sites and *Alu* elements). A common approach to assessing colocalization of features is to compare features at identical transcriptomic locations or within windows (overlapping coordinates, Figure S8A), such as with the *plotCorrelation* function in deepTools (Ramírez et al., 2016). Biologically important relationships can also occur at proximal but non-overlapping transcriptomic sites (adjacent coordinates), which can be identified and quantified by tools such as the *reldist* function in BEDtools (Favorov et al., 2012). StereoGene extended spatial relationship analysis to allow for correlation of continuous values, in addition to interval datatypes (Figure S8B) (Stavrovskaya et al., 2017). StereoGene is optimized for genomic annotation information and cannot selectively and efficiently analyze transcriptome data directly, so nearBynding expands its context of usability to transcriptomic data. Here, we provide an in-depth comparison of the spatial analysis capabilities of deepTools, BEDtools, and StereoGene and detail the ways in which nearBynding extends StereoGene to expand usability to transcriptomic analysis (summarized in Table 1).

We consider three main categories when comparing algorithms to correlate genomic relationships: the data types that can be analyzed, the types of correlation employed, and whether correlation results are visualizable. Different data types represent different information; for example, peak-called protein binding data is interval data, while aligned sequencing read data is continuous. Some data, such as G-quadruplex-forming intervals, is binary, whereas other data, such as methylation frequency, is non-binary. It is vital to choose a correlation algorithm that is able to accommodate the data type being studied. Some types of correlation, such as Spearman or Pearson correlation, can provide a single value for the overall correlation of two tracks; spatial correlation yields information about the overall relationship between two tracks as a distribution of correlations; cross-correlation shows the relative position of two tracks' features; and partial correlation can be used to remove a confounder that affects both the inputs being compared. No algorithm is able to compute all of these correlations, so users ought to first consider what correlation type can best answer the question being asked. Lastly, not all algorithms to correlate genomic relationships have the built-in capacity to graph results, which may be an important factor to weigh in the data analysis pipeline.

deepTools

deepTools *plotCorrelation* computes the Spearman or Pearson correlation between two or more files based on scores within genomic regions (default is 10 kb bins). The scores' correlation can only be computed between bins with overlapping intervals, so deepTools is blind to relationships between adjacent coordinates. However, the interval data analyzed can incorporate amplitude information such as total read coverage within the assigned bins. deepTools is unable to account for confounders affecting the tracks being correlated, but it does have the option to plot scatterplots or heatmaps to visualize the calculated correlation between files (Ramírez et al., 2016). *plotCorrelation* uses the output matrices from either *multiBamSummary* or *multiBigwigSummary*, so input data can only be in BAM or BigWig formats.

Since the CapR-predicted RNA structure data is generated as a BED file, we were unable to test how deepTools may be used to observe RNA structure and RBP binding correlation. However, we were able to test pairwise correlations between RBP eCLIP datasets. We input aligned BAM files from the intron-binding proteins PRPF8, RBM22, U2AF2, and BUD13 to see whether deepTools could observe the proteins' known colocalization patterns (*c.f.* Figures S5B–S5D). We input both replicate's BAM files for each protein as well as the size-matched inputs from RBM22 and U2AF2 as controls and visualized the correlations via *plotCorrelation* using the default arguments (Figure S8D). This method shows strong clustering of technical replicates but does not have conspicuous colocalization signal between U2AF2 and BUD13, which are known to both bind 3' intronic termini, and between PRPF8 and RBM22, which are known to both bind 5' intronic termini (Briese et al., 2019). Further, there is moderate to strong correlation between size-matched input files and RBP binding files (0.37-.91), suggesting that size-matched input information may confound the RBP binding data.

BEDtools

In addition to analyzing overlapping intervals, BEDtools *reldist* can calculate the correlation between non-overlapping features (i.e. features upstream and downstream of one another). Developed from GenometriCorr, BEDtools *reldist* allows users to identify the relative distance between two sets of intervals with consistent (i.e. non-random) spacing or proximity (Favorov et al., 2012). Unfortunately, BEDtools can only incorporate binarized interval data from BED, GFF, or VCF files in its correlation analysis (Figure S8C). And, like deepTools, it is unable to discern the effect of potential confounders of the correlation. There is no built-in visualization functionality.

To test whether BEDtools could be used to recapitulate known RNA structure binding preferences of RBPs (*c.f.* Figure 4), we input CapR-predicted RNA structure data and peak-called RBP binding data. Since BEDtools can only incorporate binarized interval data, we denoted a nucleotide as positive for a structure context (i.e. stem, hairpin, multibranch, bulge, internal, or exterior) if CapR predicted it to adopt that structure with at least 0.2 probability. We used CLIPper-called peaks for LIN28B and STAU2 eCLIP datasets (Figures S8E–S8G); PUS1 has only 35 reproducible CLIPper-called peaks, so it was omitted from the analysis. Graphing the *reldist* output shows strong correlation of RBP binding and stem structure within the closest colocalization bin for both proteins and weaker correlation for the other five structure contexts. The graphs for LIN28B in HepG2 and K562 look more similar to one another than to the STAU2 graph at relative distance < 0.05, but additional conclusions about similarities or differences in the proteins' binding preferences are difficult (Figures S8E–S8G). Since BEDtools' output is binned into intervals of 0.01 relative distance, there is poor resolution between RBP binding and RNA structure. BEDtools also cannot show regions of unfavorable binding (the correlation is always positive) or differentiate between upstream and downstream binding geometries.

The resolution limit of BEDtools is apparent when observing the binding of PRPF8, RBM22, U2AF2, and BUD13: PRPF8 and RBM22 appear to bind close to each other and U2AF2 and BUD13 bind close to each other as expected (Figures S8H and S8I); however, there is minimal difference in the correlation of U2AF2 with BUD13 versus U2AF2 with PRPF8 or RBM22 (Figure S8I), even though PRPF8 and RBM22 are known to generally bind further from U2AF2 than BUD13. Since BEDtools does not differentiate binding geometries, it is unable to show that PRPF8 and RBM22 have enriched binding $\sim$200 nt upstream of BUD13 and U2AF2 binding.

StereoGene

Whereas BEDtools' correlation of non-overlapping features is limited to binary profiles, StereoGene was extended to allow for spatial correlation of continuous values (Stavrovskaya et al., 2017). In addition to overall correlation of two tracks (i.e. the degree to which two

features are correlated across the tracks), as can be provided by deepTools and BEDtools, StereoGene also calculates positional cross-correlation to provide information about the relative position of two tracks. The cross-correlation calculation is ultimately the most valuable component of the output with regards to studying the relative position of two features, since it provides information about the geometry of the features' positions. The statistical significance of correlations with StereoGene is evaluated by a permutation-based test that compares the correlation between tracks of matched versus randomly shuffled windows. If a variable is expected to confound the correlation between the two tracks being tested, StereoGene can account for genome-wide confounders by partial correlation with another explanatory input track. StereoGene outputs an R script to visualize the cross-correlation and correlation distributions automatically.

nearBynding wraps StereoGene in biologically relevant methods and expands its context of usability from genome-scale to transcriptome-scale analysis. nearBynding alters the default StereoGene variables so that the size of the cross-correlation windows suits a transcriptomic rather than genomic scale (10 kb rather than 100 kb) and the correlation occurs in a single-nucleotide sliding frame for maximal resolution (rather than 100 nt bins). nearBynding provides additional functionality by allowing users to select regions of the transcriptome of interest; users may select specific transcripts and specific regions of those transcripts to interrogate. If the strand of the track information is available, this information is preserved for correlation analyses. The transition from genome-scale to transcriptome-scale analysis improves colocalization calculation specificity and efficiency so that correlation can be conducted at single-nucleotide resolution on a personal computer. nearBynding also determines RNA sequences and wraps CapR, an *in silico* RNA structure prediction software (Fukunaga et al., 2014), to provide RNA structure information as an optional track for correlation if only one data track (e.g. only protein binding position) is available. StereoGene only outputs the mean value at every position in the cross-correlation window of shuffled replicates; nearBynding can calculate and output the standard error of the shuffled null track to allow for a statistical assessment of foreground versus null signal. The StereoGene visualization function can only depict the correlation information for one replicate at a time; nearBynding replaces this with options to depict the mean cross-correlation signal with error bars for multiple replicates simultaneously and removes experimental background signal if an input track is also provided. Further, if the user chooses to study correlation relative to the CapR-derived RNA structure contexts, the cross-correlation signal for all six contexts can be depicted on the same graph.

QUANTIFICATION AND STATISTICAL ANALYSIS

nearBynding tracks were considered significant if +/- three standard error intervals of foreground signal computed from technical replicates did not overlap with +/- three standard error intervals of the shuffled null distribution (1,000 null iterations). Functions from the stats package were used to calculate binomial tests (*binom.test*), t-tests (*t.test*), and Kolmogorov-Smirnov tests (*ks.test*); *ks.test* is wrapped in nearBynding's *assessGrouping* function. The lsr package (Navarro, 2015) was used to calculate Cohen's *d* (*cohensD*). The transport R package function *wasserstein1d* (Schuhmacher et al., 2020), wrapped in nearBynding's *bindingContextDistance* and *bindingContextDistanceCapR* functions, was used to calculate Wasserstein distances.