bioRxiv preprint doi: https://doi.org/10.1101/2025.02.16.638500; this version posted February 16, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Decoding RNA-RNA Interactions: The Role of Low-Complexity Repeats and a Deep Learning Framework for Sequence-Based Prediction

Adriano Setti[1†], Giorgio Bini[2,3†], Valentino Maiorca[4], Flaminia Pellegrini[1,5], Gabriele Proietti[2], Dimitrios Miltiadis-Vrachnos[1], Alexandros Armaos[2], Julie Martone[6], Michele Monti[2,5], Giancarlo Ruocco[5], Emanuele Rodolà[4], Irene Bozzoni[1,2,5], Alessio Colantoni[1,5*], Gian Gaetano Tartaglia[2,7*]

† Equal Contribution

* To whom correspondence should be addressed: Tel: +39 010 28976204; Email: gian.tartaglia@iit.it
Correspondence may also be addressed to Alessio Colantoni. Tel: +39 06499122391; Email: alessio.colantoni@uniroma1.it

[1] Department of Biology and Biotechnologies “Charles Darwin”, Sapienza University of Rome, 00185 Rome, Italy
[2] Center for Human Technologies (CHT), Fondazione Istituto Italiano Di Tecnologia (IIT), 16152 Genova, Italy
[3] Department of Physics, University of Genova, Via Dodecaneso 33, 16146 Genova, Italy
[4] Department of Computer Science, Sapienza University of Rome, 00198 Rome, Italy
[5] Center for Life Nano- & Neuro-Science, Fondazione Istituto Italiano Di Tecnologia (IIT), 00161 Rome, Italy
[6] Institute of Molecular Biology and Pathology, CNR, 00185 Rome, Italy
[7] ICREA - Institució Catalana de Recerca I Estudis Avançats, Pg. Lluís Companys 23, 08010, Barcelona, Spain

bioRxiv preprint doi: https://doi.org/10.1101/2025.02.16.638500; this version posted February 16, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Abstract

RNA-RNA interactions (RRIs) are fundamental to gene regulation and RNA processing, yet their molecular determinants remain unclear. In this work, we analyzed several large-scale RRI datasets and identified low-complexity repeats (LCRs), including simple tandem repeats, as key drivers of RRIs. Our findings reveal that LCRs enable thermodynamically stable interactions with multiple partners, positioning them as key hubs in RNA-RNA interaction networks. RNA-sequencing of the interactors of the Lhx1os lncRNA allowed to validate the importance of LCRs in shaping interactions potentially involved in neuronal development.Recognizing the pivotal role of sequence determinants, we developed RIME, a deep learning model that predicts RRIs by leveraging embeddings from a nucleic acid language model. RIME outperforms traditional thermodynamics-based tools, successfully captures the role of LCRs and prioritizes high-confidence interactions, including those established by lncRNAs. RIME is freely available at https://tools.tartaglialab.com/rna_rna.

bioRxiv preprint doi: https://doi.org/10.1101/2025.02.16.638500; this version posted February 16, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Introduction

RNA molecules play key roles in cellular function, not only as carriers of genetic information but also as regulators of gene expression, structural components of molecular complexes, and guides for enzymatic processes. These functions often rely on RNA-RNA interactions (RRIs), which are central to diverse processes such as translation regulation, RNA splicing, stability control, and the formation of ribonucleoprotein complexes[1–5].

In recent years, technological advances have enabled systematic mapping of RRIs, unveiling their prevalence and complexity. High-throughput methods such as PARIS[6,7], SPLASH[8], and LIGR-Seq[9] combine Psoralen-based chemical crosslinking with sequencing to capture base-paired regions in RNAs. These techniques allow the detection of interactions at near-base resolution, revealing structural motifs and interaction networks within and between RNA molecules. Additionally, emerging methods such as KARR-Seq provide a complementary approach by capturing RNA interactions with the aid of click chemistry reactions[10]. Protein-based approaches, including RIC-seq[11] and MARIO[12], broaden the scope of RNA interactome studies by focusing on protein-facilitated RNA-RNA interactions. Together, these approaches have provided unprecedented insights into RNA interactomes, highlighting the pervasive role of RNA-RNA interactions in cellular regulation.

Despite these technological advances, the molecular determinants governing RRIs remain poorly understood. Sequence complementarity and RNA secondary structure are known to play fundamental roles[13,14], but recent findings suggest that specific sequence motifs and repeat elements contribute significantly to interaction specificity[15–17]. For instance, the A-repeat region of *Xist* long non-coding RNA (lncRNA) forms a series of inter- and intra-repeat bindings, suggesting that these interactions are vital for the multimerization necessary for *Xist* function[18,19]. Similarly, satellite repeats in a subset of the mRNAs targeted by miR-181 provide binding platforms for the miRNA; while miR-181 typically induces mRNA degradation, its interaction with these repeats is instead associated with translational repression[20]. Tandem telomeric repeats in the TERT gene allow to establish base pairing interactions between distal regions of the pre-mRNA, which are necessary for exons 7 and 8 skipping[21]. Interspersed repeat elements, such as SINEs, also play a significant role in RNA-RNA interactions. For example, STAU1 protein has been shown to recognize imperfect duplexes formed between pairs of Alu elements (primate-specific SINEs) embedded within a lncRNA and the 3'UTR of a target mRNA, leading to mRNA decay[22]. Collectively, these examples suggest a conserved mechanism by which repeat elements enhance RNA-RNA interactions, contributing to their specificity, stability, and functional versatility. However, a systematic study on the involvement of repeated elements in the establishment of RNA-RNA interactions is still lacking.

To complement experimental approaches, computational tools have emerged as powerful resources for predicting RRIs. Algorithms relying on thermodynamic models (e.g., RNAduplex and IntaRNA[23–25]) are effective for short RNA interactions but face limitations in capturing the full complexity of long RNA molecules and their in vivo dynamics[26]. Recent advances in neural networks[27], particularly deep learning[28], have created opportunities to enhance the prediction of both intra- and inter-RNA interactions. These developments harness large-scale datasets and diverse sequence encoding strategies to improve accuracy and broaden predictive capabilities. In particular, sequence representations generated by nucleic acid language models have demonstrated remarkable effectiveness in encoding intricate and informative sequence features.[29,30]. By combining high-throughput training data with advanced sequence representations, it becomes possible to develop predictive models that effectively capture the complexity of long RNA molecules and their interactions.

In this study, we investigated sequence determinants underlying the interactions between long RNA molecules, retrieved from several high-throughput datasets. Strikingly, we discovered that the interacting regions are enriched in simple and low-complexity repeats. These repetitive elements are associated with thermodynamically stable interactions and are often found in regions with a high

bioRxiv preprint doi: https://doi.org/10.1101/2025.02.16.638500; this version posted February 16, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

degree of connectivity, suggesting a potential role as interaction hubs. Notably, low-complexity repeats co-occur with interactions between transcripts involved in neuronal development. Recent research has highlighted the widespread upregulation of lncRNAs containing simple repeats during neuronal differentiation[31]. To explore this aspect further, we sequenced the RNA interactors of a lncRNA critical for motor neuron development. Our results showed that these interactors are enriched in low-complexity repeats that are likely to be involved in forming energetically favorable interactions with the lncRNA.

The discovery of these sequence features prompted us to model RNA-RNA interactions using a sequence-encoding strategy based on Language Models (LMs). To this end, we developed RIME, a deep learning-based predictor trained on Psoralen-based datasets. RIME outperforms thermodynamics-based tools, even when evaluated against interactions identified using protein-based methods. Notably, RIME achieves excellent performance on high-quality interaction datasets and demonstrates a clear ability to discern the importance of low-complexity repeats in RNA-RNA interactions. Our study provides the first systematic benchmarking of thermodynamics-based tools against RNA-RNA interactomes generated by high-throughput technologies.

bioRxiv preprint doi: https://doi.org/10.1101/2025.02.16.638500; this version posted February 16, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Results

Preparation of a transcriptome-wide RNA duplexes dataset.

We derived the main RRI dataset used in this study by collecting the results of five transcriptome-wide techniques that were developed to massively identify RNA-duplexes through RNA-sequencing: PARIS1, PARIS2, and SPLASH identify intramolecular and intermolecular helices crosslinked with Psoralen or its derivatives (Psoralen-based techniques), while RIC-Seq and MARIO detect RNA-RNA interactions mediated by proteins (protein-based techniques, **Supplementary Fig. 1a**). Overall, we collected a total of 7,813,978 RNA duplexes derived from both mouse (mESC, mouse Brain samples) and human systems (HEK293T, HeLa, Lymphoblastoid and ES cells treated or not with Retinoic Acid) (**Supplementary Table 1**). Since our aim was the characterization of interactions between long RNA molecules that are potentially involved in post-transcriptional gene expression regulation, we focused on duplexes derived from exonic RNA regions that are more likely to arise from processed transcripts (**Supplementary Fig. 1b** and **c**), and we discarded duplexes formed within the same transcripts, probably resulting from intramolecular RNA-RNA pairing (**Supplementary Fig. 1b** and **c**). Moreover, we removed interactions involving particularly abundant RNA species, which are most likely driven by their stoichiometric imbalance rather than by the specificity of RNA-RNA contact. These species include the ribosomal RNA, which directly contacts the mRNAs during translation. We also filtered out interactions carried by small RNAs, such as miRNAs and snoRNAs (**Supplementary Fig. 1b** and **c).** After these filters, we reduced our pool of RNA duplexes to 117,948 inter-molecular interactions (**Fig. 1a**, **Supplementary Table 1**), which account for 1.5% of the initial set (**Supplementary Fig. 1d**). Among these duplexes, the majority consist of mRNA pairs, with a significant proportion represented by 3'UTR-CDS and CDS-CDS interactions. Interactions between non-coding RNAs and mRNAs, particularly those involving 3'UTRs, are also well represented, with the exception of the SPLASH RRI sets, which are almost exclusively composed of interactions between protein-coding transcripts (**Supplementary Fig. 1e**).

RNA-RNA interactions detected by transcriptome-wide experiments are enriched in low-complexity repeats

After dataset preparation, we investigated the relationship between RRIs and repeated elements. We used RepeatMasker software[32] to identify known repeated elements in the 28,393 and 16,559 interacting RNAs detected in human and mouse samples, respectively. Human interacting RNAs contain 76,799 repeats belonging to 45 different families, while the murine ones harbor 43,825 repeats from 40 families (**Supplementary Fig. 1f**). To assess whether some repeat families are preferentially involved in RRIs, we performed a "nucleotide-level repeat enrichment analysis" (**Supplementary Fig. 1g**, see **Methods** section), testing for the enrichment or depletion of repeat nucleotides in the interacting RNA regions, while controlling for their overall abundance in transcripts. We found that, while most of the repeat classes displayed depletion or sporadic enrichment in RNA-RNA interacting regions, two classes stood out with widespread enrichments among datasets: Simple Repeats (SRs) and Low Complexity (LC) sequences. Indeed, the RRI identified in 12 and 8 (out of 14) experiments were significantly enriched ($\log_2[\text{Odds Ratio}] > 0$ and FDR < 0.05) in SRs and LCRs, respectively (**Fig. 1b**). Moreover, to confirm that the enrichment of these repeat families was not dependent on the software used to annotate repeated elements, we also employed the TanTan algorithm[33], observing that tandem repeats (not homopolymeric, TTnotHomo, see **Methods**) are enriched in RNA-RNA interactions from 9 experiments (**Fig. 1a**). We grouped the repeats belonging to the classes of SR, LC and TTnotHomo under the definition of Low-Complexity Repeats (LCRs)[34].

bioRxiv preprint doi: https://doi.org/10.1101/2025.02.16.638500; this version posted February 16, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

To further validate the robustness of our results, we repeated the analyses on a subset of RNA-RNA interactions with more stringent characteristics, such as longer interaction patches, that are more robust to mapping errors (**Supplementary Fig. 1h**, upper panel), or higher number of supporting reads (**Supplementary Fig. 1h**, lower panel), confirming the LCR enrichment in most of the sets. Moreover, we followed an orthogonal strategy by performing a "permutation-based repeat enrichment analysis" (see **Methods**), which estimates how much the observed overlap between interactions and repeats deviates from what would be expected by chance (**Supplementary Fig. 1i**). Overall, the results confirmed the strong enrichment of LCRs within the RNA-RNA interaction regions (**Supplementary Fig. 1j** and **k),** with even stronger enrichment observed when requiring the presence of repeats on both RNA interaction patches (**Supplementary Fig. 1j** and **l**).
By analyzing the sequence of the LCRs involved in RNA-RNA interactions, we found that the majority of them were composed of multiple repeats of G- and C-containing units (G-rich and C-rich SR, G-rich and GA-rich LC, G-rich and CU-rich TTnotHomo, **Fig. 1c**, **Supplementary Fig. 1m** and **Supplementary Table 1**).
Finally, to obtain some potential insight on regulatory processes in which the interactions could be implicated, we performed Gene Ontology (GO) term enrichment analysis on RNAs found to interact through LCRs. Interestingly, most of these RNAs encode for proteins involved in developmental processes and that act as transcription factors (**Supplementary Fig. 1n** and **Supplementary Table 1**), even in non-developmental systems like HeLa and HEK293T.

Low-complexity repeats are involved in multiple RNA-RNA contacts

We then assessed which mRNA regions are most involved in RNA-RNA interactions by performing a "region-level RRI enrichment analysis" (see **Methods**) on the transcriptome wide RNA-RNA interaction data. The results of this analysis, visualized via meta-transcript representation (**Fig. 2a**), revealed that, while interactions from almost all experiments are more prevalent in the 5'UTR, those identified in HeLa cells using PARIS1 are preferentially localized in the 3'UTR (**Fig. 2a**). Notably, LCRs over-representation in interactions showed a similar profile with respect to global RRIs enrichment, reinforcing the idea that the region-specific RRIs enrichment could be driven by low-complexity repeats (**Supplementary Fig. 2a** and **Fig. 2a**). Moreover, these results indicate that different systems could have different preferential regions for RNA-RNA interactions, albeit retaining the same properties. Indeed, the interaction sites at the 3'UTR in HeLa cells maintain the enrichment of low-complexity sequences (**Supplementary Fig. 2a**).
Additionally, both the RIC-seq and the PARIS2 experiments performed upon polyA enrichment (HEK293T-polyA and Mouse Brain) revealed widespread enrichments in the CDS (**Fig. 2a**), suggesting some preference of these datasets in capturing contacts involving translated regions.
Interestingly, we observed strong spikes in the meta-transcript profile of RNA-RNA contacts derived from PARIS1 experiments in HEK293T and mESC (**Fig. 2a** and **b**). We found that these spikes were due to single RNAs, TNPO2 for HEK293T and Zmpste24 for mESC, capable of generating large amounts of RNA-RNA interactions with multiple partners. Strikingly, we observed that 675 interactions of TNPO2 and 254 interactions of Zmpste24 were concentrated in specific regions of their transcripts, the first annotated as SR and the second as LC (**Fig. 2c**, left and right panel, respectively).
This finding prompted us to investigate whether LCRs have a natural tendency to form interactions with many different partners.
To this aim, we performed a network analysis comparing one random sampling of repeats-associated RNA-RNA interactions to an equal number of interactions not associated with repeats. RRI network analysis could be robustly performed only in datasets with a suitable amount of RRI, so we decided to focus on all the dataset with at least 1000 interactions. We utilized a spring

bioRxiv preprint doi: https://doi.org/10.1101/2025.02.16.638500; this version posted February 16, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

network model that tends to localize nodes with more edges centrally within the network. The analysis demonstrated that LCR-containing interactions exhibit a shorter distance from the network center compared to controls, confirming that they can engage more molecular partners (**Fig. 2d** and **e**). Notably, this is a common feature across the analyzed datasets. Indeed, we confirmed the same trend in all the datasets for at least one type of LCRs in at least 2 out of 5 independent samplings (**Supplementary Fig. 2b**).

We further tested the potential of LCR-containing interacting patches to be involved in multiple interactions with an orthogonal approach. First, we stratified the RNA patches from each experiment into five bins based on the interaction degree. We then assessed the proportion of LCR-containing patches within each bin. In line with the previous result, we found that, for 7 out of 8 experiments tested, LCRs were enriched in the bins with highest degree (**Fig. 2f**). Interestingly, we did not observe the same behavior for other abundant repeats such as LINE1 or Alu/B1 SINE (**Fig. 2f**).

Complementary low-complexity repeats mediate lncRNA-target interactions

Looking for further evidence to support our findings, we explored the possible role of repeated elements in mediating the RNA-RNA interactions identified through targeted approaches. First, we focused on TINCR and its RNA interactome, revealed by Kretz and colleagues through RIA-Seq methodology[16]. The lncRNA hosts a UC-rich SR (**Fig. 3a**, upper panel). Therefore, we decided to investigate the repeat content of its top 300 targets, as compared to that of non-interacting RNAs with matched properties (**Supplementary Fig. 3a**). This analysis revealed that SRs and LC sequences are over-represented in TINCR interactors and, among them, the most abundant one is a GA-rich LC sequence, found in 12% of the targets (**Fig. 3a** and **Supplementary Fig. 3b**). The complementarity between the UC-rich sequence hosted by TINCR and the GA-rich ones found among its targets suggests that such low-complexity repeat could be directly involved in multiple direct RNA-RNA interactions.

Given the link between RNA-RNA interactions involving LCRs and regulatory genes implicated in development, we decided to experimentally investigate the RNA interactome of the lncRNA Lhx1os, recently described as a potential regulator of motor neuron (MN) development[35]. To this aim, we set up an RNA pull-down experiment, which allowed the isolation of the lncRNA Lhx1os using 20 nt long antisense biotinylated oligonucleotide probes in mESC-derived MNs. The experiment was performed employing two sets of probes, ODD and EVEN, targeting all the Lhx1os isoforms (**Fig. 3b**). A third set of probes, targeting the LacZ mRNA, was used as a negative control (LacZ) (**Supplementary Table 2**). The proper enrichment of Lhx1os was validated through qRT-PCR (**Supplementary Fig. 3c**).

To identify the RNA interactors, we performed RNA sequencing of duplicated samples for each probe set, obtaining a set of 161 transcripts significantly enriched by ODD and EVEN, but not LacZ, purification (**Supplementary Fig. 3d** and **Supplementary Table 3**). We randomly selected 5 of them and validated their enrichment by qRT-PCR (**Supplementary Fig. 3e**).

Analyzing the Lhx1os sequence with RepeatMasker, we observed the presence of a SINE B1 element (also annotated as MurSatRep1) in antisense orientation in the third exon of the transcript. We investigated the possible enrichment of complementary repeats in the identified RNA targets with respect to a control set with matched properties (**Supplementary Fig. 3f**). First, we focused on the SINE B1, wondering whether, among the candidates, there was an enrichment of complementary SINEs. Indeed, while the presence of SINE B1 in antisense orientation was comparable between target RNAs and controls, the SINE B1 with sense orientation, which could in principle pair with the antisense SINE B1 hosted by Lhx1os, was over-represented in the interacting RNAs (**Supplementary Fig. 3g**). Then, we carried out a global repeat enrichment analysis, revealing a significant overrepresentation of MurSatRep1 and GA-rich LC sequences within the

bioRxiv preprint doi: https://doi.org/10.1101/2025.02.16.638500; this version posted February 16, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

targets (**Fig. 3c**). Since the GA-rich LC is the most frequently enriched repeat in the Lhx1os interactors (**Supplementary Fig. 3h**), we wondered if there was an association with the SINE B1 observed in Lhx1os. Interestingly, mapping the 7-mers complementary to the GA-rich LC on the Lhx1os sequence, we found that they are mainly located in the SINE B1 region, indicating a possible pairing of the two repeated elements (**Supplementary Fig. 3i**).

Low-complexity repeats are involved in thermodynamically stable interactions

After determining the widespread presence of repetitive elements within the regions involved in RRIs, we proceeded to characterize the thermodynamic properties of these interactions, starting from those determined using transcriptome-wide approaches. To do this, we used the IntaRNA 2 software[25] with the intent to predict the stability (ΔG) of each interaction. However, we noticed that, when using the default parameters, for roughly 10% of the interactions the software failed to identify any pairing in the evaluated regions (**Supplementary Fig. 4a**, left panel). This issue arose from the default seed constraint usage, which requires a perfectly complementary region of 5 to 8 nucleotides. By disabling the seed constraint, or by using IntaRNA 2 to emulate other prediction tools that do not require the presence of a seed (RNAup[36] and RNAhybrid[37]), we were able to obtain ΔG scores for each interaction (**Supplementary Fig. 4a**, right panel, and **b**). This evidence suggests that a significant portion of the experimentally identified interactions do not rely on a seed as their core. Notably, these seedless interactions were found to be less thermodynamically stable compared to others (**Supplementary Fig. 4c**).

Subsequently, we compared the ΔG distributions of experimentally identified interactions to a set of mock interactions derived from random RNA regions using multiple approaches (IntaRNA2 with and without seed, RNAup, and RNAhybrid). While the deviation from the mock interactions was modest, all methods consistently showed that the experimentally identified interactions are more thermodynamically favorable than the random controls (**Supplementary Fig. 4d**).

Next, we focused on repeats. We stratified the interactions from each experiment into five groups, from the most thermodynamically favorable to the least favorable (**Supplementary Fig. 4e**), evaluating the enrichment of repeats in the various strata. Interestingly, LCRs showed a strong enrichment among the most thermodynamically favorable interactions, which was not observed for other repeated elements such as Alu or LINE1, used as controls (**Fig. 4a**).

To validate the link between low ΔG values and LCRs, we examined the lncRNA-RNA interactions detected through targeted approaches. By performing an "RNA contact enrichment analysis" (see **Methods, Supplementary Fig. 4f**), we identified 5 TINCR regions which are more prone to interact with the target RNAs with respect to the control set for at least one of three algorithms (**Fig. 4b**, **Supplementary Fig. 4g** and **Supplementary Table 3**). Notably, 4 of the identified regions overlap with the TINCR motif boxes previously characterized as RNA-binding regions[16]. Of these identified regions, region 2 displayed the highest enrichment in contacts with the target RNAs, while region 4 exhibited the most favorable thermodynamics (**Fig. 4b**, **c** and **d**, and **Supplementary Fig. 4h**).

First, we focused on region 2 and noticed that it does not coincide with any annotated repeat (**Fig. 4b**). However, by looking at the regions of the targets predicted to contact TINCR region 2, we observed an enrichment of SRs compared to controls, indicating that the predicted interactions with this region are probably mediated by LCRs present in the target RNAs (**Fig. 4c**). In contrast, analyzing region 4, we found that it coincides with the TINCR UC-rich SR (**Fig. 4b**) and we observed an enrichment in GA-rich LCRs in the contact region of the predicted targets (**Fig. 4d**), supporting the hypothesis that the interactions involving region 4 of TINCR are mediated by contacts between low-complexity sequences.

Subsequently, we conducted an RNA contact enrichment analysis also on the Lhx1os pull-down dataset. As we hypothesized, thermodynamics-based predictions identified the SINE B1 as the

bioRxiv preprint doi: https://doi.org/10.1101/2025.02.16.638500; this version posted February 16, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

region most prone to interact with the target RNAs, compared to the controls (**Fig. 4e**, **Supplementary Fig. 4i** and **Supplementary Table 3**). Additionally, the analysis of the predicted target regions contacted by Lhx1os SINE B1 confirmed the enrichment of the previously identified GA-rich LCR (**Fig. 4f**).

Modelling RNA-RNA interactions with RIME, a deep learning approach based on language model embeddings

Our thermodynamic analysis of RNA-RNA interactions (**Supplementary Fig. 4d**) has revealed that, although the ΔG of experimentally detected duplexes is lower when compared to that of mock interactions generated with random RNA regions, it does not provide a neat separation between the two sets. Indeed, as already shown in other studies, thermodynamics-based computational tools often fail to predict biologically relevant interactions between long RNA molecules[26,38].

To handle the complexity of the interactions from a new perspective, beyond standard sequence representations like one-hot encoding[39–41] and thermodynamic features, we exploited the Nucleotide Transformer (NT) LM[30]. This model provides dense vectorial representations of nucleotide sequences, known as embeddings, which efficiently capture underlying biological patterns. Leveraging this advanced representation, we trained a deep neural network (RIME, RNA Interactions Model with Embeddings) to predict RNA-RNA interactions, aiming to evaluate whether LMs can effectively model the determinants of RRIs (**Supplementary Fig. 5a**). The model was trained to infer whether two input RNA sequences, modeled using a contact matrix representation (see **Methods**), interact (positive class) or not (negative class). To ensure the identification of direct RNA-RNA interactions, independent of protein mediation, the positive set consisted of contact matrix windows encompassing duplexes detected by Psoralen-based methodologies (PP, Paired regions in interacting transcript Pairs, **Fig. 1a**, **5a** and **b**, see **Methods**). The negative set was constructed from three distinct classes of non-interacting windows (**Fig. 5a** and **b**, see **Methods**). These included randomly selected Non-paired regions in interacting transcript Pairs (NP), randomly selected Non-paired regions in Non-interacting transcript pairs (NN), and Paired regions already seen in the positive set combined in Non-interacting transcript pairs (PN). This setup allowed us to approach RRI prediction from two angles (**Fig. 5b** and **c**). Discriminating RNA-RNA Patches (DRP) focuses on distinguishing interacting regions (PP) from non-interacting ones (NN, NP). Discriminating RNA-RNA Interactors (DRI) involves identifying interacting regions within transcript pairs that are truly interacting (PP) versus those in randomly combined transcript pairs that are not interacting with each other (PN).

The training set consisted of 70% of the total sample, while the remaining 30% was used for model evaluation (15% for validation and 15% for testing, **Fig. 5d**). To avoid potential biases we ensured similar RNA composition and length among the Positive and Negative set, as well as among the Training, Validation and Test sets (**Supplementary Fig. 5b**, **c**, **d** and **e**). Furthermore, to verify the ability of the model to generalize beyond Psoralen-based methodologies, we evaluated its performance on protein-based datasets (RIC-seq and MARIO) (**Supplementary Fig. 5f**).

RIME enhances RNA-RNA interaction prediction beyond thermodynamic models

The extensive RRI set we compiled allowed us to perform a thorough evaluation of RIME performances, and to benchmark it against the most well-known thermodynamics-based prediction tools: IntaRNA 2[25], ASSA[42], RNAplex[43], RNAhybrid[37], RIsearch2[44], pRIblast[45], RNAup[46], and RNAcofold[47].

We first assessed RIME's prediction scores by comparing them with those from these tools. Thermodynamics-based programs clustered into two main groups, reflecting differences in how

bioRxiv preprint doi: https://doi.org/10.1101/2025.02.16.638500; this version posted February 16, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

accessibility is considered into their predictions. Notably, RIME showed a weak to moderate correlation with thermodynamics-based tools, indicating that our method could capture novel interaction patterns (**Fig. 6a**, **Supplementary Fig. 6a**).
To evaluate performance across different tools and account for errors in both positive and negative predictions, we used the Area Under the ROC Curve (AUC) as a global evaluation metric. The tools were tested on 200 nucleotide-long RNA regions pairs (**Supplementary Fig. 6b**).
To evaluate the robustness of the tested methods, we assessed their performance on different subsets with progressively increasing quality, determined by the number of supporting reads (both for Psoralen-based and protein-based sets) and by the interacting regions length (only for the Psoralen-based methodologies, which rely on shorter chimeric reads).
In the DRP task, RIME consistently outperformed thermodynamics-based tools on human Psoralen-based RNA-RNA interactions, particularly when assessed against higher quality sets (**Supplementary Fig. 6c** and **d**, left panel). In particular, for the subset with at least 3 supporting reads or interacting regions length higher than 35 nucleotides (hereby named Psoralen HQ Human) (**Supplementary Fig. 6e)**, RIME demonstrated an AUC of 78%, while the top-performing thermodynamics-based tool (RIsearch2) reached 63% (**Fig. 6b**, left panel, and **Supplementary Table 4**). Noteworthy, we observed a similar trend for the RIC-Seq dataset (**Supplementary Fig. 6f**, left panel), with RIME achieving an AUC of 77% when evaluated against interactions supported by at least 4 reads (RIC-seq HQ, **Fig.6a**, right panel, and **Supplementary Fig. 6e**), whereas on the same set thermodynamics-based tools reached a peak AUC value of 66% with RIsearch2 (**Fig. 6c**, right panel, and **Supplementary Table 4**).
Although the overall performance metrics for the DRI task were comparatively lower, RIME still stood out as the top-performing tool, with its AUC improving alongside dataset quality (**Supplementary Fig. 6c**, **d** and **f,** right panel). On the Psoralen HQ Human dataset RIME demonstrated an AUC of 72%, while the leading thermodynamics-based tool (RIsearch2) achieved an AUC of 58% (**Fig. 6c**, left panel and **Supplementary Table 4**). Consistently, on the HQ RIC-Seq dataset, RIME's AUC (70%) was significantly better than RIsearch2's one (61%, **Fig. 6c**, right panel, and **Supplementary Table 4**).
We then shifted our focus to the human mouse RRI dataset and found that, although the performance gap between RIME and thermodynamics-based tools narrowed, RIME still emerged as the top-performing method with an AUC of 76% in DRP and 62% in DRI tasks, when evaluated against Psoralen HQ interactions (**Supplementary Fig. 6g** and **h**, and **Supplementary Table 4**). On the MARIO dataset, the performance of the tools was generally lower. In the DRP task, RIME ranked third with 0.65, while the leading tool, RNACofold, achieved 0.68. For the DRI task, all tools perform at a level comparable to random chance, suggesting a limited ability of the methodology to capture specific interactions.

RIME identifies key determinants of RNA interactions

The analysis of RIME's performance prompted us to further explore the relationship between three features: data quality, which reflects the types of evidence supporting the reliability of experimentally identified interactions; RIME's prediction score, indicating the model confidence in predicting positive or negative interactions; and RIME's performances, representing the model overall predictive capability (**Fig. 7a**, left panel). At first, we focused on the relationship between data quality and RIME's performances. As indicated by previous results, when selecting RRIs with gradually more stringent quality criteria, such as number of chimeric reads (**Supplementary Fig. 7a** and **6e**) or interacting region size (**Supplementary Fig. 7b**), we observed that RIME's performances increased monotonically. This result supports the idea that the model has understood the complex patterns behind true interactions and that the lower performance on the global dataset can be imputed to the higher proportion of false positives (**Figure 7a**, right panel). To further explore this aspect, we analyzed the connection between RIME's prediction confidence and data

bioRxiv preprint doi: https://doi.org/10.1101/2025.02.16.638500; this version posted February 16, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

quality. We found that the model's confidence increased with data quality, (**Supplementary Fig. 7c** and **d**), indicating that the prediction score can be effectively used to prioritize interactions likely to be true positive (**Figure 7a**, right panel). Additionally, we analyzed the relationship between RIME's performance and confidence for both positive and negative classifications, utilizing precision for positives and negative predictive value (NPV) for negatives (**Supplementary Fig. 7e**). We observed a positive correlation between the model's confidence and its performance, highlighting the prediction score's effectiveness in identifying reliable interactions (**Figure 7a**, right panel).

We previously demonstrated that LCRs are crucial sequence determinants involved in RNA-RNA interactions. Additionally, we found that thermodynamics-based tools are highly effective at detecting interactions involving LCRs, as these typically exhibit low ΔG values across nearly all the analyzed RRI sets (**Supplementary Fig. 7f**). To assess whether also our deep learning model captured the significance of these sequence features, we analyzed the confidence scores for all the positive sets. Notably, we found that, consistent with other thermodynamics-based tools, RIME identifies LCRs as key determinants in RNA-RNA interactions, assigning higher scores to RRIs involving them, particularly when both interaction patches contain these elements (**Fig. 7b**).

To gain deeper insight into the criteria driving RIME's predictions, we employed Grad-CAM[48], a gradient-based visualization technique that highlights the regions with the highest influence on the model's outputs. Notably, Grad-CAM localization maps consistently revealed high-intensity regions near actual interacting sites, with this proximity improving as the maximum intensity in the localization map increased and the confidence score of the model improved (**Supplementary Fig. 7g**). This indicates that, especially for high-confidence predictions, it is possible to precisely pinpoint the regions involved in the interaction by looking at the areas most important for the model's decisions (**Fig. 7c** and **Supplementary Fig. 7h**).

RIME accurately predicts high-confidence and functional interactions

To maximize data utilization and achieve the best possible model, we developed a final version of RIME, RIMEfull, by increasing the training data through the combination of the Psoralen-based training and validation datasets. To ensure the selection of the optimal model based on true interactions, we used the Psoralen HQ test set as the reference for model evaluation and selection. Notably, this expanded training dataset enhanced the model's performance when evaluated on the high-quality RIC-seq data (**Supplementary Fig. 8a** and **Supplementary Table 4**).

To further prove the reliability of RIMEfull's predictions on external datasets, we tested its ability to identify RNA-RNA interactions involving the NORAD lncRNA, which were recently profiled using the COMRADES technique[49]. By reanalyzing the COMRADES data, we identified 17 mRNAs that specifically interact with the lncRNA (see **Methods** section, **Supplementary Fig. 8b** and **c**). Mapping the interactions captured by the experiment onto the lncRNA sequence revealed regions more prone to contact the target RNAs (**Fig. 8a**, upper panel and **Supplementary Fig. 8d**). We used RIME to predict interacting regions between NORAD and each target RNA. Notably, we found that the top RIME predictions for each target clearly recapitulate the chimeric reads profile on the NORAD sequence, underscoring the ability of the model to recognize RNA-RNA interaction patches (**Fig. 8a**, lower panel). For instance, when examining EMP2, the NORAD interactor with the highest number of chimeric reads (**Supplementary Fig. 8e**), we observed a specific region near the 5' end of the transcript that engages in multiple interactions with NORAD RNA (**Supplementary Fig. 8f,** left panel). Notably, RIMEfull successfully pinpointed this primary target region contacted by NORAD (**Supplementary Fig. 8f**, right panel). Furthermore, analysis of all 17 RNA-RNA pairs showed that the regions predicted by RIMEfull as interacting were significantly enriched in chimeric reads compared to those predicted as non-interacting (Fisher's exact test p-value = 1.16×10^{-16}, **Supplementary Fig. 8g**).

bioRxiv preprint doi: https://doi.org/10.1101/2025.02.16.638500; this version posted February 16, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Finally, we evaluated the effectiveness of RIMEfull in predicting experimentally validated RNA-RNA interactions. As a first case study, we examined the interaction between TINCR and the PGLYRP3 mRNA, which depends on a region in its 3' end for binding[16]. Remarkably, RIME successfully identified this region as one of the most likely to interact with TINCR (**Fig. 8b**), underscoring the model's ability to precisely identify RNA regions that mediate contacts with other RNAs. Next, we tested whether RIME could identify RNA-RNA interaction regions with known functional roles. To do this, we analyzed the interactions through which the murine lncRNA SMaRT negatively regulates the translation of Mlx-γ[50] and Spire1[51] mRNAs. In both cases, RIME efficiently recognized the experimentally validated interacting regions, while also identifying additional regions that were found to be putatively involved in the interaction between the lncRNA and its targets (**Fig. 8c** and **Supplementary Fig. 8h**). Notably, in the case of Mlx, these additional regions may explain why pull-down experiments captured Mlx isoforms lacking the primary interaction region (**Fig. 8c**).

bioRxiv preprint doi: https://doi.org/10.1101/2025.02.16.638500; this version posted February 16, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Discussion

In 1979, Britten and Davidson hypothesized a crucial role for repetitive sequences in forming RNA-RNA duplexes that could potentially regulate gene expression[52]. This early idea laid the foundation for current research recognizing repetitive RNAs as essential players in gene regulation[53]. Here, we conducted the first systematic investigation into the role of repetitive elements in mediating interactions between long RNA molecules, detected through transcriptome-wide and targeted approaches. Our analysis identified low-complexity repeats, such as tandem repeats, as the most prominent repeat families involved in RNA-RNA interactions. The functional relevance of LCRs was further highlighted by their tendency to form interaction patches engaged in multiple contacts, as well as their significant enrichment in the RNA targets of long non-coding RNAs. These findings suggest that low-complexity repeats, by nature of their sequence composition, may act as "sticky" hubs within RNA interaction networks, promoting complementary interactions through multiple contact points that enable high-avidity binding. Further increasing the interest in their regulatory role, we found that LCRs-associated interactions frequently involve genes regulating developmental processes, consistent with a recent study highlighting connections between specific repeat subtypes and gene function[34]. Supporting this notion, we experimentally identified the RNA interactors of a lncRNA critical for motor neuron development and confirmed the presence of complementary LCRs that could be involved in contacts between the lncRNA and its targets. This reinforces the idea that LCRs play a role in coordinating the RNA interactome essential for cellular differentiation and tissue development.

On the pathological front, the association between LCRs and RNA-RNA interactions opens avenues for exploring disease mechanisms. Indeed, many repetitive elements are already linked to diseases, including Huntington's disease[54], autism[55] and ALS[56]. Our findings raise the question of whether these diseases could also be investigated as "RNA-RNA interaction pathologies". Notably, repeat-containing transcripts can phase-separate to form RNA *foci* within living cells, stabilized by intermolecular base-pairing interactions, which are frequently associated with disordered regions[57]. These structures may contribute to the long-term toxicity observed in neurodegenerative diseases[58–60], underscoring the critical role of repeat-mediated RNA-RNA interactions in the pathology of these conditions.

Our *in silico* analysis revealed that LCRs act as sequence determinants that enable the formation of thermodynamically stable RNA-RNA interactions. However, while thermodynamic stability, as calculated by predictive tools, offers valuable insights, it is not always sufficient to accurately infer whether two RNA molecules interact *in vivo*. To address this, we developed RIME, a deep learning model that utilizes context-specific nucleotide sequence embeddings to capture the sequence determinants driving the interaction between long RNA molecules. The model was rigorously tested by assessing its ability to identify regions involved in RNA-RNA interactions and to distinguish, among potential pairings of these regions, those that form actual interacting duplexes. To our knowledge, this comprehensive evaluation approach has not been applied before. Moreover, we curated RNA-RNA interaction datasets for model training and evaluation, which are made available to the community to foster further advancements in the field. RIME demonstrated superior performance compared to thermodynamics-based models. Notably, RIME excelled at predicting high-quality interactions, demonstrating its ability to extract meaningful patterns, such as the contribution of LCRs, from the noisy high-throughput experimental datasets used for its training. In fact, even if these sets are dominated by mRNA heteroduplexes, RIME successfully recapitulated the contacts formed between lncRNA regulators and their mRNA targets. Moreover, despite not

bioRxiv preprint doi: https://doi.org/10.1101/2025.02.16.638500; this version posted February 16, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

being explicitly trained to predict the exact positions of interacting regions, RIME showed a tendency to focus on areas in proximity to the actual contact sites to inform its decisions, which could be exploited to predict binding sites with higher resolution. Given its superior performance and ability to capture the complexities of RNA interactions, RIME holds great potential for accelerating research in RNA biology, especially if integrated with gene expression data[61], and may pave the way for novel therapeutic strategies targeting RNA-mediated regulatory networks.

bioRxiv preprint doi: https://doi.org/10.1101/2025.02.16.638500; this version posted February 16, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Methods

Construction of an RRI dataset from transcriptome-wide studies

The genomic coordinates of PARIS1 and MARIO interactions were retrieved from the RISE database[62]. The coordinates of PARIS2 interactions were retrieved from the supplementary tables provided in the original publication [7], and subsequently filtered to retain only canonical chromosomes. The RNA-RNA interactions coordinates from all these methodologies are referred to the hg38 genome assembly for experiments in *Homo sapiens* and to mm10 for those in *Mus musculus*. The interactions obtained *via* SPLASH experiments[63], described by coordinates on the RefSeq human transcriptome[64] , were obtained from the resource available at the link https://csb5.github.io/splash/. Finally, the RIC-seq interactions[11], provided by the authors upon request, were originally mapped to the hg19 genome assembly and subsequently converted to hg38 using the UCSC LiftOver tool[65]. Subsequently, the interactions of the RRI dataset annotated with genomic coordinates were mapped onto the reference human or mouse transcriptomes using the Ensembl 99 GTF annotation[65,66]. This mapping was performed using a custom Python script that runs the intersect module from BEDtools suite v2.29.1[67] with the parameters *-s -wao*. For genes with multiple isoforms, we employed the longest one, selecting the longest protein-coding variant in the case of protein-coding genes. Only RNA-RNA interactions where both patches were entirely mapped onto exons were retained, excluding interactions where both patches were mapped onto a single transcript, which likely correspond to intramolecular interactions. In cases where interaction patches mapped to ambiguous regions and were thus assigned to multiple genes, we selected a single representative gene using a specific prioritization strategy: genes with the most prevalent biotypes in the dataset and those with interactions in regions most involved in RNA-RNA duplex formation were given priority. Interactions involving small and/or abundant RNAs with the following biotypes were then filtered out: *Mt_rRNA*, *Mt_tRNA*, *misc_RNA*, rRNA, *rRNA_pseudogene*, *snRNA*, *snoRNA*, *scaRNA*, *miRNA*, and *scRNA*. Moreover, with the same intent, we filtered out interactions overlapping transcript regions annotated by RepeatMasker v4.1.1[32] as belonging to the following repeat families: *SINE_tRNA*, *SINE_tRNA-Deu*, *SINE_tRNA-RTE*, *SINE_5S-Deu-L2*, *scRNA_unknown*, *rRNA_unknown*, *snRNA_unknown*, *tRNA_unknown*, *srpRNA_unknown*, *SINE_7SL*. Finally, due to the high amount of interactions in the RIC-seq RRI set (134,550, **Supplementary Table 1**), we further refined this set by selecting the interactions supported by at least two chimeric reads (9,808).
The biomaRt R package v2.48.3[68] was used to annotate RRIs, classifying them by the biotype of the interacting transcripts and the specific transcript regions involved.

Detection of repeated elements

Interspersed repeats and low complexity sequences were found by scanning transcript sequences with RepeatMasker v4.1.1[32] with default settings, specifying the proper organism (*Homo sapiens* or *Mus musculus*) and using rmblastn v2.11.0+ as internal engine. The RepeatMasker output was then converted to a BED file using the RM2Bed.py util of the software. Tandem repeats were also detected using Tantan v49 software[33] with the *-f 4* parameter. The repeats identified by Tantan were divided in two groups: TanTanHomopolymeric (TTH), in which the repeating unit is composed by a single nucleotide and TanTan Not Homopolymeric (TTnH), in which the repeating unit is composed by multiple nucleotides.

bioRxiv preprint doi: https://doi.org/10.1101/2025.02.16.638500; this version posted February 16, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Repeat enrichment analyses

The "nucleotide-level repeat enrichment analysis" was performed as follows**:** for each repeat family, the BED file containing repeat coordinates was initially merged using the Bioframe Python library[69], and the same was done for the BED file of the interaction patches from each RRI detection experiment. Subsequently, the two BED files were intersected using Bioframe, and the number of nucleotides involved in interactions and belonging to the analyzed repeat family was calculated. The number of nucleotides specific to either the repeats or to the RRIs was computed by reciprocally applying the Bioframe *subtract* function. The number of nucleotides not involved in interactions and not belonging to repeats was determined by subtracting the sum of the nucleotides from the other three categories from the total length of the interacting transcripts. After that, a contingency table was populated with the four numbers (**Supplementary Fig. 1g**) and the statistical significance of the association between repeats and RNA patches was calculated using Fisher's exact test with Benjamini-Hochberg correction for multiple testing. The analysis was also repeated by using only interaction patches with a minimum length or supported by a minimum number of chimeric reads. Length thresholds were independently determined for each experiment by dividing the patch length distribution into quintiles and using the resulting breakpoints (**Supplementary Fig. 1h**, upper panel).

For each experiment, the repeat enrichment profile along the meta-transcript was computed by applying the same procedure described above to each meta-transcript bin. Meta-transcript bins were generated as described in the following "Region-level RRI enrichment analysis" section.

Using BEDtools coverage, nucleotides within each bin were categorized into four groups: those exclusively involved in RRIs, those exclusively belonging to a repeat family, those belonging to both categories, and those belonging to neither. The abundance of each category was quantified, and a "nucleotide-level repeat enrichment analysis" was conducted for each bin. Finally, p-values were adjusted using the Bonferroni multiple testing correction.

The "permutation-based repeat enrichment analysis" was performed as follows: initially, the localization of repeat regions within the functional regions of RNAs (5'UTR, CDS, 3'UTR, or the whole transcript if non-coding) was determined using transcript information retrieved from Ensembl using BioMart[70]. Subsequently, 100 sets of mock repeats were generated by shuffling the repeat positions 100 times, while preserving the localization within the same functional regions (**Supplementary Fig. 1i**). For each experiment, BEDtools intersect was used to assess the overlap between RRIs and both real and mock repeats. The number of RRIs overlapping with real repeats was transformed to a Z-score by using the mean and the standard deviation of the distribution obtained by counting the overlaps of RRIs with the 100 shuffled mock repeat sets.

For each repeat family, the configurations X-R or R-X were assigned to interactions where the repeat overlap with only one of the two interacting RNA patches, while the R-R configuration was assigned to interactions where repeats of the same family are present in both the patches (**Supplementary Fig. 1j**). The overlap between repeats and patches was determined using BEDtools intersect.

GO analysis

Gene Ontology (GO) term enrichment analyses of RNAs with SR or TTnH repeat-overlapping interaction patches were performed with the WebGestaltR R v0.4.6 package[71], using all the interacting RNAs of the specified experiments as background. Only experiments with at least 5000 interacting RNAs were taken into consideration.

bioRxiv preprint doi: https://doi.org/10.1101/2025.02.16.638500; this version posted February 16, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Region-level RRI enrichment analysis

The functional structure of interacting mRNAs, described by the lengths of the 5'UTR, CDS, and 3'UTR regions, was summarized into a meta-transcript representation. Region lengths, together with the transcript sequences, were retrieved from the Ensembl 99 database using the biomaRt R package v2.48.3 [68]. mRNA regions within the same class (5'UTR, CDS, or 3'UTR) were divided into an equal number of bins, depending on the ratio between the median length of that class and the median length of the other classes: 20 for the 5'UTR, 130 for the CDS and 150 for the 3'UTR. The bin ranges for each RNA were compiled into a BED file and, for each experiment, the amount of RRIs in meta-transcript bins was calculated using BEDtools intersect. Mock interaction sets were generated using BEDtools shuffle with the parameters *-chromFirst -bedpe.* A random distribution was obtained by counting the overlaps of meta-transcript bins for each of the 100 mock interaction sets, and its mean and standard deviation were used to calculate the Z-score for real RRIs in each bin. Graphical representations of the meta-transcript analyses were generated using the ComplexHeatmap R package v2.8.0 [72].

RRI network analysis

RRI network analysis was performed for each transcriptome-wide experiment using the NetworkX Python package [73]. Repeat family-specific networks were constructed by combining RRIs that overlap with repeats from a specific family with an equal number of randomly selected RRIs that do not overlap with repeats from that family. If the number of repeat-overlapping RRIs exceeded 500, a random subset of 500 was selected to build the network in which each node represents a transcript and the edges represent the interactions. The weights of the spring network were defined by the number of chimeric reads that supported each interaction. Nodes were positioned using the Fruchterman-Reingold force-directed algorithm provided by the *spring_layout* function. Then, for each node, the euclidean distance from the center (coordinates 0, 0) of the spring network was calculated by applying the *distance.euclidean* function from scipy.spatial Python module. For each edge, the distance from the network center was calculated as the average of the distances of the two nodes involved in the interaction.

The interaction degree was calculated as follows. Non-redundant interacting RNA patches were identified by collapsing all the overlapping RNA regions involved in RRI across each experimentally-derived set using BEDtools merge, considering the strand of the feature (**Supplementary Table 1**). Next, BEDtools intersect was used to map RRI patches to RNA regions, enabling the counting of unique RRI identifiers assigned to each RNA patch. BEDtools intersect was also applied to identify RNA patches overlapping with repetitive elements.

Thermodynamic stability prediction of the RRIs from transcriptome-wide studies

To calculate the ΔG for each RRI and obtain comparable scores within individual datasets, a fixed-size nucleotide region centered on the interaction was selected for each interaction patch. The size of this region was determined *ad hoc* for each dataset to ensure that it fully encompassed 95% of the interactions within the dataset. Then, the ΔG prediction was performed using IntaRNA 2 software v2.3.0 [25] with default parameters for the standard prediction; *--noSeed* for the prediction excluding the seed region; *--mode=M --noSeed --qAccW=0 --qAccL=0 --tAccW=0 --tAccL=0 --tIntLenMax=25 --qIntLenMax=25* and *--noSeed --tAcc=N --qAcc=N* parameters for RNAup[36] and RNAhybrid[37] emulation, respectively, as described in https://github.com/BackofenLab/IntaRNA. To ensure comparability across RRI sets of varying length, length-normalized ΔG scores from IntaRNA2 were used.

bioRxiv preprint doi: https://doi.org/10.1101/2025.02.16.638500; this version posted February 16, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

For comparison with mock interactions, the patches in each dataset were randomly shuffled within the transcripts using BEDtools shuffle with the following parameters: *-chromFirst -bedpe*, avoiding the formation of intramolecular interactions. Subsequently, the predictions were conducted using the same criteria described above.

Cell cultures conditions and treatments

All cell lines used in this study were grown at 37°C, 5% CO2 and tested for mycoplasma contamination. Murine HBG3 ES cells (embryonic stem cells derived from HB9::GFP transgenic mice) were cultured on gelatin-coated or MEF-coated dishes and maintained in mESC medium (Dulbecco's Modified Eagle's Medium for ES, 15% Fetal Bovine Serum for ES, 1X GlutaMAX, 1X Non Essential AmminoAcids, 1X nucleosides, 1X 2-mercaptoethanol and 1X Penicillin/Streptomycin) supplemented with LIF (103 unit/mL), FGFRi (1,5 μM) and Gsk-3i (1,5 μM) (LIF+2i condition). Medium was changed every day and cells were passaged every 2–3 days with 1X Trypsin-EDTA solution. Spinal motor neurons (MNs) were differentiated from mESCs HB9::GFP according to Wichterle et al.[74] and Errichelli et al.[75]. After papain dissociation, cells were plated on polyornithine/laminin-coated dishes with N2B27 medium (50% DMDM F12, 50% Neurobasal, 1X Glutamax, 1X Penicillin-Streptomycin, 1X B27, 1X N2, 200 ng/mL Ascorbic Acid, 20 ng/mL BDNF, 10 ng/mL CNTF, 10 ng/mL GDNF, 10 nM Rhok inhibitor) and differentiation was allowed to proceed for three additional days.

RNA extraction and quantification by qRT-PCR

Total RNA was isolated from cell cultures using miRNeasy Kit (Qiagen), and retrotranscribed with Superscript Vilo cDNA synthesis Kit (11754050, Invitrogen) according to the manufacturer's protocol, in a final reaction volume of 10 μL. qRT-PCR was performed using PowerUp SYBR Green Master Mix (A25742, Life Technologies). Relative RNA quantity was calculated as the fold change (FC, $2^{-\Delta\Delta Ct}$) with respect to the control sample set as 1, unless differently specified. Oligonucleotides used for qRT-PCR are provided in **Supplementary Table 2**. DNA amplification was monitored with an ABI 7500 Fast qPCR instrument. Data analysis was performed using the SDS Applied Biosystem 7500 Fast Real-Time PCR system software.

Native RNA pull-down

Native RNA pull-down was performed in mESC-derived neural mixed population obtained from EB cells at day 6 that were dissociated and replated, allowing differentiation to proceed for additional 3 days (DIV3) as described by Errichelli et al.[75]. The RNA pull-down procedure was performed as described by Martone et al.[50]. The sequences of the biotinylated oligonucleotides belonging to ODD and EVEN probe sets (consisting of six and five probes, respectively), and LacZ control are listed in **Supplementary Table 2**.

RNA pull-down sequencing and analysis

Library preparation for Input, ODD, EVEN, and LacZ duplicate samples was performed using Stranded Total RNA Prep with Ribo-Zero Plus (Illumina) according to manufacturer's documentation. The sequencing reaction was performed on an Novaseq 6000 sequencing system at the Genomic Facility in Istituto Italiano di Tecnologia (Genoa, Italy) and produced an average of 43,6 million 150-nucleotide long paired-end read pairs. Quality of reads was assessed using FASTQC software v0.11.9 (available at https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Reads were processed using Cutadapt v3.2[76] with *-u 1 -U 1 –trim-n –nextseq-trim=20 -m 35* parameters to trim low-quality bases and

bioRxiv preprint doi: https://doi.org/10.1101/2025.02.16.638500; this version posted February 16, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

sequencing artifacts. Reads aligning to the sequence of murine rRNAs (retrieved from NCBI Gene database[77]), 7SL or 7SK transcripts (retrieved from Ensembl database, Supplementary Table 2) were filtered out; this first alignment was conducted using Bowtie2 v2.4.2 [78], aligning the first read in each pair with the *--nofw* parameter and the second read with the *--norc* parameter. Only read pairs in which both mates did not map were retained for downstream analysis. STAR v2.7.7a software[79] was used to align reads to GRCm38 genome using the following parameters: *--outSAMstrandField intronMotif --outSAMattrIHstart 0 --outFilterType BySJout --outFilterMultimapNmax 20 --alignSJoverhangMin 8 --alignSJDBoverhangMin 1 --outFilterMismatchNmax 999 --outFilterMismatchNoverLmax 0.04 --outFilterIntronMotifs RemoveNoncanonical --peOverlapNbasesMin 50* l. PCR duplicates were removed from all samples using MarkDuplicates command from Picard suite[80]. Uniquely mapping fragments from de-duplicated BAM files were counted for each annotated gene (Ensembl release 99) using Htseq v0.13.5 software[81]. edgeR v3.34.1 R package[82] was used to compare pull-down-enriched RNAs to their relative Input samples. RNAs with p-value <0.05 and log2(FC) > 0.58 in both ODD versus Input and EVEN versus Input contrasts, and log2(FC) < 0 in the LacZ versus Input contrast were defined as enriched. Additionally, only RNAs with expression levels exceeding 1 FPKM (Fragments Per Kilobase of exon per Million mapped reads, calculated using the edgeR *rpkm* function) in the Input sample were kept. Isoform quantification was performed with Salmon v1.6.0 software[83] using *–libType ISR –validateMapping*s parameters and a full decoy transcriptome index created using the Ensembl 99 transcriptome and the GRCm38 genome. In case of genes with multiple isoforms, the most expressed isoform in the Input sample (according to Salmon TPM quantification) and with the same biotype of the gene was selected for downstream analyses.

The top 300 TINCR interactors were selected based on the 'log2_escore' column reported in the table of the 100 nucleotide-long regions enriched in the TINCR pull-down (retrieved from the GEO record GSE40121). Only interactors also present in the Ensembl release 99 database were included in the top set. A re-analysis of the TINCR pull-down experiment[16] was performed to calculate log2(FC) values between Pull-down samples and Input, enabling the identification of RNAs primarily depleted in the pull-down (log2[FC] < -1), from which controls were selected. The data were downloaded using fastq-dump from sratoolkit v2.10.9 (available at https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software). Subsequently, reads were aligned to the GRCh38 genome using STAR v2.7.7a[79] with the ENCODE standard parameters, along with the following additional settings: *--outSAMstrandField intronMotif --outSAMattrIHstart 0 --outFilterIntronMotifs RemoveNoncanonical*. The subsequent steps, including de-duplication and quantification of uniquely mapping fragments for each annotated gene (*Homo sapiens*, Ensembl release 99 annotation), were performed as previously described.

Generation of negative controls for lncRNAs' targets and repeat enrichment analysis

For each of the identified RNA interactors of the Lhx1os or TINCR lncRNAs, 5 and 3 negative controls, respectively, were selected for nucleotide-based RNA-RNA interaction prediction analysis and screened for the occurrence of interspersed repeats. The controls were chosen from transcripts expressed in the Input samples (FPKM > 1) but depleted in the pull-down extracts. For Lhx1os, we selected controls within the bottom 50% of log2(FC) values in the pull-down versus Input comparison. For TINCR, controls were selected from transcripts with log2(FC) values below -1 in the same comparison. The controls were selected to match the corresponding RNA targets in gene and transcript biotypes, as well as in 5' UTR, CDS, 3' UTR, and overall transcript length. The selection process followed these steps: (I) 5' UTR, CDS, 3' UTR, and overall lengths were obtained from Ensembl Release 99 using BioMart; (II) for each interactor, the percentage difference (%Δ) between the interactor and candidate controls was calculated for each parameter; (III) the %Δ values across all parameters were summed for each candidate; (IV) controls with the lowest total

bioRxiv preprint doi: https://doi.org/10.1101/2025.02.16.638500; this version posted February 16, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

%Δ and matching biotypes were selected. Each control was uniquely assigned to a single interactor to prevent redundancy. Finally, controls were tested using the Wilcoxon-Mann-Whitney test to confirm the absence of significant differences compared to interactors (p-value > 0.05).
For each repeat family and element, we used Fisher's exact test to compare the proportions of targets and controls containing them. P-values were then adjusted for multiple comparisons using the Benjamini-Hochberg procedure.

RNA contact enrichment analysis

Initially, TINCR and Lhx1os sequences were segmented into 50- or 10-nucleotide windows with steps of 10 or 1 nucleotide, respectively. The resulting bins were then compiled into a BED file. To predict the most favorable interactions between the analyzed RNA and each transcript in the target or control group, we used IntaRNA2 software with standard parameters, ASSA and RIblast (**Supplementary Table 3**). Interaction predictions for both the target and control groups were then mapped to the lncRNA bins using BEDtools intersect. For each bin, the statistical significance of the over-representation of contacts with pull-down RNAs compared to contacts with control RNAs was evaluated using Fisher's exact test. Additionally, the difference between the ΔG distributions obtained for pull-down RNAs and controls was assessed using the Mann-Whitney U-test. Benjamini-Hochberg correction was applied for both tests.
Meta-region analysis of predicted interaction patches in the pull-down and control groups was conducted as follows. A region extending from -100 nt to +100 nt around the middle point of interaction patches predicted on target and control RNAs was analyzed using a sliding window of 10 nucleotides, with steps of 1 nucleotide. Windows falling partially or totally outside transcripts were discarded. Then, RepeatMasker v4.1.1 was used to identify repeated elements in pull-down and control RNA groups, and BEDtools intersect was used to assign repeats to the overlapping windows. Also the interaction patches were assigned to windows in order to calculate the frequency of contacts in each window. Finally, a Fisher's exact test was performed to test the enrichments of repeat-overlapping predicted interaction patches in the pull-down group compared to control RNAs. P-values were corrected for multiple testing using the Benjamini-Hochberg method.
The identification of Lhx1os 7-mers complementary to GA-rich low-complexity sequences was performed as follows: since no specific annotations for GA-rich (-) or TC-rich (+) sequences are available in RepeatMasker, we retrieved GA-rich low-complexity sequences annotated by the software within the mESC-derived neural mixed population transcriptome (i.e. Input samples of Lhx1os pull-down experiment). These sequences were then partitioned into 7-mers, and the reverse-complementary sequences were mapped onto the Lhx1os transcript.

RIME Architecture

RIME is a deep learning model that leverages Nucleotide Transformer[30] (Multispecies 2.5B model) embeddings to predict RNA-RNA interactions. It first takes two RNA sequences as input, each represented as a series of 6-mer tokens. The NT model then transforms these tokenized sequences into 2,560-dimensional vectors, where each token captures the global sequence context. We introduced a Contact Matrix (CM) operation in our pipeline to map the two embeddings into a contact matrix, where each spatial point corresponds to a token pair combination from the two RNA sequences. Contact matrices provide a compact way to represent the spatial relationships between nucleotides and have been already used for secondary structure prediction[84]. The feature space of the contact matrix is formed by concatenating the respective embeddings of the token pairs. The output tensor can be seen as an image, and we designed specific modules commonly employed in computer vision. First, since the dimensionality of each 'pixel' is too high after concatenation (2x2,560), we jointly learn an independent transformation, denoted as ψ, to reduce the dimensionality to 800. Second, we designed a Feature Extractor transformation that effectively

bioRxiv preprint doi: https://doi.org/10.1101/2025.02.16.638500; this version posted February 16, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

mixes the information across both spatial and feature spaces using standard CNN-based layers. The advantage of convolutional layers is their length independence, which allows our pipeline to handle RNA sequences of varying lengths. Each input contact matrix has different dimensions, so we standardized them by applying global average pooling. The final layer for binary prediction is a linear layer with bias.
RIME was developed using the Pytorch v1.9.1 Python library[85].

Positive and negative classes and prediction tasks

The positive set used to train and evaluate the RIME model was generated from the RRI collection derived by Psoralen-based experiments. To enhance data consistency and avoid redundancy, we mapped overlapping interaction sites from different experiments into a unified interaction region, representing the minimal region encompassing all identified interaction sites and providing a cohesive view of the interaction dynamics. To provide sufficient context to the model, the positive set consisted of two-dimensional windows encompassing the interaction sites. The negative set was composed of non-interacting windows generated from the interacting transcripts according to the following criteria. The NP negative set was created by randomly selecting windows from the contact matrices of interacting transcript pairs, while ensuring that interacting sites were excluded from the selection. To generate the windows that compose the PN and NN classes, we first generated random transcript pairs, ensuring that they did not interact with each other, thus creating non-interacting contact matrices. Then, NN windows were generated by randomly selecting RNA regions from these matrices. Finally, PN windows were positioned in regions encompassing interaction patches of both transcripts. The described creation of non-interacting contact matrices was performed by a random pair-swapping approach, designed to prevent biases while preserving experimental context and biological relevance. We restricted swapping within each RRI dataset to ensure negative pairs reflected the specific experimental conditions and remained consistent with the original data. We also confined the non-interacting transcript pairs generation by preserving the biotype composition of interacting sites. In particular, we categorized each interacting site into four types - CDS, 3'UTR, 5'UTR (protein-coding transcripts), and ncRNA (non-coding transcripts) - yielding 16 possible combinations (e.g., CDS-CDS, CDS-3'UTR). We confined swapping within these combinations to ensure negative pairs matched the interaction region of their positive counterparts, maintaining a consistent "genetic grammar". We generated two non-interacting contact matrices for each interacting one where possible, though the exclusive presence of a single interaction type for certain biotype composition could prevent the generation of non-interacting contact matrices. In this process, we ensured that the generated transcript pairs were never seen interacting in the following datasets: RIC-seq, PARIS2 and MARIO, PARIS1, RAIDv2.0, CLASH, SPLASH, NPInterv3.0, RIAseq, RAINv1.0, RAP, LIGRseq from RISE database.
Building on this approach, we evaluated the problem of predicting RRIs by considering two distinct evaluation perspectives: DRP, whose aim is to classify PPs as positives and NNs and NPs as negatives, and DRI, whose aim is to classify PPs as positives and PNs as negatives. The DRP task assesses the model's ability to identify interaction-prone regions, while the DRI task evaluates its accuracy in identifying the transcript pairs that are actually interacting.
The Psoralen-based dataset was split into training (70%), validation (15%) and test (15%) sets. To ensure similar RNA composition between positives and negatives, we performed the randomed combination procedure generating the non-interacting contact matrices within each set. Furthermore, we ensured that the training, test, and validation sets maintained similar length distribution and RNA region composition, reflecting the properties of the original dataset (**Supplementary Fig. 5e**).

bioRxiv preprint doi: https://doi.org/10.1101/2025.02.16.638500; this version posted February 16, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Data preparation for RIME training

To accommodate the NT model's input sequence limit of 6,000 nucleotides, we processed each sample into sub-sequences of up to this length, resulting in sub-contact matrices with dimensions of up to 6,000×6,000 nucleotides. To note, Nucleotide Transformer operates on 6-mers, encoding each 6,000-nucleotide region as a sequence of 1,000 embedding tokens. To enhance the coverage of longer transcript pairs during training, additional embeddings were generated. Specifically, for contact matrices where both RNAs exceeded 7,000 nucleotides, two embeddings were created (22% of the training set), and for pairs exceeding 8,500 nucleotides, three embeddings were generated (2%).

Once embeddings were created, we sampled contact matrix windows corresponding to four classes—PP, PN, NP, and NN—which collectively formed the training set. Each window captured smaller segments of the original embeddings but, since NT embeddings are designed to encode each token within the context of the entire sequence, the broader sequence context was preserved. As a result, the feature space of each window reflected the influence of the full embedded transcript.

In each epoch, we used a balanced sampling strategy across three classes of negative samples: PNs made up 50% of the total negative samples, while NPs and NNs each contributed 25%. This resulted in a balanced sample distribution between DRI and DRP tasks.

To create a balanced training set and prevent the model from being biased toward positive or negative interactions, the total number of negative samples was matched to the number of PP samples. This was achieved by adjusting the sampling probabilities of NN, NP, and PN windows so that their combined total matched the number of PPs, while preserving the 50%-25%-25% distribution of PN, NN, and NP windows.

During training, the four classes—PN, NN, NP, and PP—were sampled from varying regions of the embeddings across epochs, increasing coverage of the 6,000-nucleotide embeddings, enhancing sequence context diversity, and mitigating overfitting.

Additionally, during training we sampled positive and negative windows with side lengths that varied according to a distribution centered at 200 nucleotides (**Supplementary Fig. 5c** and **d**), equalizing the length distributions across classes. This dynamic sampling ensured that the same interaction site (for PP) was sampled with varying granularity and sequence context, preventing memorization, enhancing generalization, and enabling the model to handle varying lengths.

To accommodate batch training with transcripts of different lengths, we normalized tensor lengths within each batch to match the shortest sequence by averaging groups of tokens. Specifically, for each longer tensor, we divided it into groups such that the number of groups equaled the shortest tensor length. We then computed the mean value for each group, resulting in a reduced tensor that matched the shortest tensor length. The number of tokens in each group was determined by the ratio of the tensor length to the shortest batch length (with any remainder grouped into the last group). This approach was made possible since the data within the batch represent a vector space comparable with the original embedding space, with original sequences treated as tensors with group sizes of one. This averaging acted as an additional form of augmentation, avoiding the inefficiencies of zero-padding, which would otherwise lead to the creation of sparse matrices. As an additional augmentation technique, we also reversed the order of transcript pairs during training, ensuring that the model did not develop an order-based bias, allowing it to learn from varied configurations of the same data.

Data preparation for RIME evaluation

The Psoralen-based validation set, Psoralen-based test set, as well as the RIC-seq and MARIO test sets, consisted of 200×200 nucleotide positive and negative windows.

For each interacting transcript pair, we generated a 6,000x6,000 contact matrix embedding, sampling one PP and one NP window. Similarly, for non-interacting pairs, we created embeddings

bioRxiv preprint doi: https://doi.org/10.1101/2025.02.16.638500; this version posted February 16, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

and sampled one NN and one PN window. For transcript pairs longer than 8,500 nucleotides (5% in RIC-seq, 1% in MARIO, 2% in the Psoralen-based test set), we generated additional embeddings without interaction sites (and without randomly combined sites for non-interacting pairs), sampling one NP window for interacting pairs and one NN window for non-interacting pairs.
We excluded all interacting transcript pairs observed in the Psoralen-based training set from both the RIC-seq and MARIO datasets to ensure an out-of-sample evaluation context. Additionally, we excluded pairs where the interaction site on either transcript exceeded 200 nucleotides.
All the results for the DRI and DRP tasks were evaluated by balancing the positive and negative sets according to the proportions of negatives considered for the specific task.
Testing was conducted on the full Psoralen-based, RIC-seq, and MARIO datasets, as well as on subsets with progressively higher quality, determined by selecting interactions based on the number of supporting reads and the minimum length of the interacting region. To calculate AUC values, for each threshold negative samples were randomly undersampled 100 times to match the number of positive samples, yielding 100 ROC-AUC values, which were then averaged to calculate the final performance metric. The same undersampling-based approach was adopted to calculate Accuracy, Precision, NPV, Recall, and F1 score (**Supplementary Table 4**). To calculate Precision and NPV against RIME's confidence (**Supplementary Fig. 7e**), the imbalance between the positive and the negative sets was addressed by undersampling before ranking and selecting the top and bottom predictions at each data proportion threshold. For each threshold the scores were averaged over 50 iterations.

RIME training and model selection

RIME was trained for 74 epochs with early stopping and a patience of X epochs, using a batch size of 32. The model that achieved the highest F1 score on the validation set was selected. The F1 score was calculated on a subset of the validation set that mirrored the composition of the training set, maintaining the 50%-25%-25% distribution of PN, NN, and NP windows and ensuring their total matched the number of the entire set of positive windows. To ensure an unbiased F1 score the negative set undersampling process was repeated 50 times, resulting in equal numbers of positive and negative pairs for each iteration. The F1 scores were then averaged across the iterations. In each epoch, we uniformly and randomly sampled half of all possible RRIs in the Psoralen-based training set, ensuring that, on average, the model was exposed to the entire set of interacting sites every two epochs.
To prioritize the reliability of positive interaction predictions, we optimized a weighted Cross Entropy loss function with a class weight ratio of 1.5:1.0, giving greater emphasis to the positive class. We used the AdamW optimizer with a learning rate of 5×10^{-5}, weight decay as 1×10^{-4}, and standard values for other hyperparameters: $\beta1 = 0.9$, $\beta2 = 0.999$, $\varepsilon = 1 \times 10^{-8}$.
We trained RIMEfull by expanding the dataset size through the incorporation of the validation set into the training phase. Model selection was performed on the high-quality Psoralen-based test set, chosen for its reliability and the strong performance of RIME on high-quality interactions. RIMEfull was trained for 86 epochs, with the F1 score used to select the best model, while all other hyperparameters were set identically to those used for RIME training. Since the entire Psoralen-based dataset was used for training and model selection, the performance of RIMEfull was evaluated on the external RIC-seq and MARIO datasets.

Comparative benchmark of RNA-RNA interaction prediction algorithms

The benchmark of RNA-RNA interaction prediction algorithms was performed evaluating RIME performances together with ASSA v1.0.1, IntaRNA 2 v3.4.0, pRIblast v0.03, RIsearch2 v2.1, RNAcofold v2.6.4, RNAhybrid, RNAplex 2.6.4 and RNAup. All the tools were evaluated on the

bioRxiv preprint doi: https://doi.org/10.1101/2025.02.16.638500; this version posted February 16, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

same set of 200x200 RNA regions already described in the "Data preparation for RIME evaluation" section. For each analyzed window, we measured the interaction propensity score of the algorithms as the minimum ΔG value returned by the software. ASSA was used with *--all_sites* parameter; RNAcofold with *--output-format=D*; for RIsearch2, we retained only RNA-RNA interactions filtering out predictions with negative strand; RNAhybrid algorithm was emulated by IntaRNA 2 software as previously specified. For all the other tools we used standard parameters.

Grad-CAM analysis

To enhance the interpretability of our classifier, we used Grad-CAM, an explainable Artificial Intelligence technique originally designed for computer vision tasks. Grad-CAM provides spatial explanations for model predictions by analyzing how gradients influence neurons in the last convolutional layer of the CNN, thereby highlighting the most relevant regions. Grad-CAM predictions are class-specific, and we focused on the positive class for our analyses. In our model, Grad-CAM was applied after the last convolutional layer, where it is expected to capture optimal high-level semantics. Since the convolutional blocks reduced the spatial dimensions of the RNA-RNA contact matrix, the resulting Grad-CAM matrix also had lower resolution. To align its dimensions with the original contact matrix, we resized the 2D Grad-CAM localization map using nearest-neighbor interpolation.

Preparation of RRI sets for RIMEfull evaluation

The COMRADES data relative to NORAD lncRNA were retrieved from the GEO database under accession number GSE188445, considering only untreated samples. Chimeric reads associated with the NORAD transcript (ENSG_ENST_NORAD_lincRNA) were retained, excluding those corresponding to secondary structures. For each potential RNA interactor we quantified the number of chimeric reads mapped in COMRADES experiments with proper cross-linking (GSM5683127_Nr1 [replicate 1], GSM5683128_Nr2 [replicate 2], and GSM5683129_Nr3 [replicate 3]) and in control samples (GSM5683135_Nr1C [replicate 1], GSM5683136_Nr2C [replicate 2], and GSM5683137_Nr3C [replicate 3]). To identify NORAD-specific interactors, we evaluated the enrichment of chimeric reads for each RNA target in the COMRADES samples compared to the controls. This analysis was conducted independently for each replicate, and NORAD interactors were defined as RNAs with a significant enrichment of chimeric reads in COMRADES in at least two replicates. The statistical significance of the enrichment was determined using Fisher's exact test, selecting targets with log2(Odds Ratio) > 1 and p-value < 0.05. To annotate the *bona fide* interacting regions, chimeric read arm sequences were re-mapped using the Python regex library to the NORAD sequence (NR_027451.1) and to that of the identified interactors, retrieved from Ensembl 99 annotations. Reads with multiple alignments or more than five nucleotide mismatches were excluded to ensure high-confidence mappings. RIMEfull predictions between NORAD and its identified targets was performed on this transcript sequence set.
The PGLYRP3 region that interact with TINCR lncRNA was retrieved from the work from Kretz and colleagues considering the TINCR box-depleted PGLYRP3 mutant (PGLYRP3Δ96□bp).
lnc-SMaRT interacting regions with Mlx-γ and Spire1 transcripts were retrieved from the works of Martone and colleagues[50] and Scalzitti and colleagues[51], respectively.

RIME web service

To enable access to RIME, we developed a user-friendly web service featuring the RIMEfull model for RNA-RNA interaction prediction. The platform, freely accessible at https://tools.tartaglialab.com/rna_rna, allows users to input two RNA molecules and compute

bioRxiv preprint doi: https://doi.org/10.1101/2025.02.16.638500; this version posted February 16, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

RIMEfull prediction scores across 200x200 nucleotide windows with a 100-nucleotide step. The results are visualized as a heatmap, which allows to identify RNA region pairs with a high likelihood of interaction.

Data and code availability

A web service leveraging the RIMEfull model for RNA-RNA interaction prediction is freely available at https://tools.tartaglialab.com/rna_rna.

Author contributions

A.S. and A.C. retrieved the RRI datasets analyzed in the present study. A.S. performed most of the bioinformatics data analysis, including repeat enrichment evaluation and thermodynamic characterization of RRIs. G.B. conceived, developed and tested RIME, formalizing the DRP and DRI model evaluation tasks, with help from V.M., M.M. and E.R. in model design and negative samples generation, from G.P. in GRAD-CAM analyses, and D.M.V. in benchmarking RIME against thermodynamics-based tools. F.P. performed and analyzed the Lhx1os pull-down experiment. A.C. and G.G.T. conceived the study and wrote the manuscript, with contributions from A.S. and G.B.. I.B., G.R., J.M. and M.M. provided valuable input in the interpretation of data and results. All the authors read and approved the manuscript.

Funding

The research leading to these results have been supported through ERC [ASTRA_855923 (to G.G.T., I.B. and G.R.), H2020 Projects IASIS 727658 and INFORE 825080 and IVBM4PAP 101098989 (to G.G.T. and G.R.)] and National Center for Gene Therapy and Drug-based on RNA Technology" (CN00000041), financed by NextGenerationEU PNRR MUR – M4C2 – Action 1.4- Call "Potenziamento strutture di ricerca e di campioni nazionali di R&S" (CUP J33C22001130001) (to G.G.T.).

Declaration of interests

The authors declare no conflict of interest.

Acknowledgments

The authors would like to thank the 'RNA initiative' at IIT and all the members of Tartaglia's and Irene Bozzoni's lab at Sapienza and IIT. Special recognition goes to Dr. Davide Mariani for expertly curating the sequencing library preparation.

bioRxiv preprint doi: https://doi.org/10.1101/2025.02.16.638500; this version posted February 16, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Legends

Fig. 1

RNA-RNA interaction landscapes reveal a predominance of low-complexity repeats.

a, Schematic representation of the RRIs derived from transcriptome-wide experiments that were employed in this work. The first row describes the RRI detection methodology; the second row describes the organism and the third one reports the cellular system; the fourth row reports the amount of intramolecular RNA-RNA interactions used in the present study. Columns are colored based on the type of the RRI detection methodology: Psoralen-based or protein-based. **b**, Heatmap depicting, for each transcriptome-wide experiment (rows), the log2(Odds Ratio) representing the nucleotide enrichment of repeat families (columns) in RRIs. Significant enrichment or depletion of repeats is indicated by red or blue colors, respectively; white (NS) denotes the absence of statistical significance, while gray (NA) indicates no nucleotide overlap. Statistical significance was assessed using Fisher's exact test with Benjamini-Hochberg correction for multiple testing. The bar plot above the heatmap shows, for each tested repeat family, the number of transcriptome-wide experiments where a statistically significant association between repeats and RRIs was found (red for enrichment, blue for depletion) or not (gray). The color bar below the heatmap indicates the proportion of repeat nucleotides involved in RRIs relative to the total RRI nucleotides, averaged across all experiments, using a color scale from yellow (low frequency) to violet (high frequency). **c**, Stacked bar plot reporting, for each transcriptome-wide experiment, the percentage of RRIs overlapping SR (left panel), LC (middle panel) or TTnH (right panel), categorized by their nucleotide composition. The colors in the legend describe the nucleotide composition categories used to classify repeated elements.

Fig. 2

LCRs exhibit high connectivity in RNA-RNA interaction networks.

a, Heatmap reporting, for each transcriptome-wide experiment (rows), the "region-level RRI enrichment analysis" Z-score calculated in each meta-transcript bin (columns). A representation of the coding and non-coding mRNA regions is depicted under the heatmap. The bar plot on the right reports the fraction of meta-transcript bins enriched in RRIs (Z-score > 2) that display LCR nucleotide enrichment in RRIs (FDR < 0.05, log2[OR]>0 for SR, LC or TTnH). **b**, Meta-transcript profile reporting, for PARIS1 HEK293T (left panel) and PARIS1 mESC (right panel) transcriptome-wide experiments, the amount of observed RRIs (red line) compared to the distribution of 100 random shuffles (gray area, with median value represented with a black line, region-level RRI enrichment analysis) across mRNA regions (x-axis). Dashed boxes highlight regions with the highest peak in CDS (PARIS1 HEK293T) or 3'UTR (PARIS1 mESC). **c**, Bar plot depicting the amount of RNA-RNA interactions identified in the binned sequence (x-axis) of TNPO2 (left panel) and Zmpste24 (right panel) transcripts, as well as the localization of repeated

bioRxiv preprint doi: https://doi.org/10.1101/2025.02.16.638500; this version posted February 16, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

elements. **b** and **e**, Left panel: Spring network of RRIs involving LCRs (red edges) or not (black edges), detected in PARIS1 HEK293T (**b**) or PARIS1 mESC (**e**) experiments. Right panel: density plot reporting the distribution of nodes distance from the center of the network for RRIs overlapping (red) or not (gray) with LCRs. Statistical significance was assessed using Mann Whitney U-Test. **f**, Heatmap illustrating the row-scaled (Z-score) frequency of overlap between various repeat families and the RRIs detected by transcriptome-wide experiments with over 1000 interactions (rows). Frequencies are measured across five equally sized groups of RNA-RNA interaction patches (strata), ranked by their number of contacts (degree level).

Fig. 3

Transcripts targeted by TINCR and Lhx1os lncRNAs are enriched in LCRs.

a, Left panel: schematic representation of the TINCR sequence, highlighting the position of the LC UC-rich element indicated. Lower panel: bar plot reporting the percentage of transcripts hosting LC GA-rich elements among the top TINCR interactors (targets) and negative controls. **b**, Schematic representation of Lhx1os RNA pull-down in mESC-derived motor-neurons (MN). Created with BioRender.com **c**, Left panel: schematic representation of the Lhx1os sequence, indicating the position of the UC-rich SINE B1 element. Lower panel: bar plot reporting the percentage of transcripts hosting LC GA-rich elements among the identified Lhx1os interactors (targets) and negative controls.

Fig. 4

LCR-bearing RNA-RNA interactions exhibit thermodynamic stability.

a, Heatmaps illustrating the row-scaled (Z-score) frequency of overlap between various repeat families and the RRIs detected by transcriptome-wide experiments (rows). Frequencies are measured across five equally sized groups of RNA-RNA interactions, ranked from left to right in order of decreasing thermodynamic stability. The ΔG prediction was performed using IntaRNA without seed constraint. **b**, Heatmap displaying the results of the "RNA contact enrichment analysis" for the TINCR RNA sequence. The sequence is represented from the 5' end (left) to the 3' end (right) in segments of 50 nucleotides, with a 10-nucleotide sliding step. The top panel line plot illustrates the frequency of RNAs from the target (red line) or the control (black line) group predicted by IntaRNA 2 software to be in contact with each segment. The first heatmap row displays log2(Odds Ratio) values, which measure the enrichment of contacts with target RNAs compared to control RNAs. Significant enrichment or depletion of contacts is indicated by red or blue colors, respectively; white denotes the absence of statistical significance. Statistical significance was assessed using Fisher's exact test. The second row of the heatmap reports the difference between the median ΔG values of contacts with the target and control groups. Significant negative or positive difference of ΔG medians is indicated by red or blue colors, respectively; white denotes the absence of statistical significance. Statistical significance was assessed using Mann-whitney U-test. Orange boxes in the bottom annotation of the heatmap correspond to segments containing the TINCR box motifs in sense (+) or antisense (-) orientation. Black boxes correspond to bins with repeated elements from different families. Red boxes, also highlighted with dashed rectangles, indicate the 5 TINCR regions that displayed preferential binding with target RNAs according to at least one out of three prediction tools (IntaRNA 2, RIblast, ASSA). **c** and **d**, Visual representation of the results of "RNA contact enrichment analysis" performed on TINCR region 2 (**c**) and 4 (**d**). The text in the top-left area reports the amount of RNAs from the target and control

bioRxiv preprint doi: https://doi.org/10.1101/2025.02.16.638500; this version posted February 16, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

groups that are predicted to contact the TINCR region. The box plot in the bottom-left area reports the ΔG distribution of predicted interactions between the TINCR region and the two transcript sets, calculated using IntaRNA 2. Statistical significance was assessed using Mann-whitney U-test. The heatmap in the right panel presents the results of the meta-region analysis of repeat families (rows), with columns corresponding to nucleotide positions within a 200-nucleotide zone centered on regions from target or control RNAs that are predicted to contact the TINCR region. Significant enrichment or depletion of repeats in target RNA contact regions with respect to controls is indicated by red or blue colors, respectively; white denotes the absence of statistical significance. Statistical significance was assessed using Fisher's exact test. **e**, Heatmap displaying the results of the "RNA contact enrichment analysis" for the Lhx1os RNA sequence. For a detailed explanation of the top rows, please refer to the **Fig. 4b** legend. The bottom row reports the frequency of UC-rich 7-mers found within the Lhx1os sequence. Dashed rectangles indicate the 3 Lhx1os regions that displayed preferential binding with target RNAs according to at least one out of three prediction tools (IntaRNA 2, RIblast, ASSA). **f**, Visual representation of the results of "RNA contact enrichment analysis" performed on Lhx1os region 2. For detailed explanations, refer to the legends of **Fig. 4c** and **d**.

Fig. 5

Preparation of datasets for training and evaluation of RIME.

a, Schematic representation of the negative data generation process. Non-interacting pairs were generated by randomly combining interacting transcript pairs. The interacting sites are shown as white rectangles along transcript sequences, represented with lines. **b**, Illustration of Interacting Transcript Pairs (left panel) and Non-interacting Transcript Pairs (right panel), represented with contact matrices. Training and evaluation data were generated by sampling windows of varying sizes within these matrices, classified into four categories (PP, NP, NN, or PN) based on their positions. Interacting sites, depicted as white rectangles along the transcript sequence axes, are included in PP and PN windows but excluded from NN and NP windows. **c**) Visual representation of the DRP and DRI model evaluation tasks. In both cases, PPs serve as the positive set, while non-interacting regions (NN and NP) are used as the negative set for DRP, and randomly permuted interacting regions (PN) for DRI. **d**, Table reporting the amount of interacting region pairs (RRIs) and transcript pairs used to build the training, validation, and test sets for the Psoralen-based dataset.

Fig. 6

RIME overcomes the limitations of thermodynamics-based tools.

a, Heatmap of Pearson correlation coefficients between RRI prediction tools, calculated using scores assigned to positive interactions in the Psoralen-based test set. For comparability with RIME scores, ΔG values from thermodynamics-based tools were converted by inverting their sign. Tools with similar scoring patterns were grouped using average linkage hierarchical clustering based on Euclidean distances. **b** and **c**, ROC curves describing the performances of the RRI prediction tools in the DRP (**b**) and DRI (**c**) tasks, shown for the Psoralen HQ Human (left) and the RIC-seq HQ (right) datasets. To address class imbalance, negatives were randomly undersampled 100 times to match the number of positives, and ROC curves were averaged across iterations.

Fig. 7

bioRxiv preprint doi: https://doi.org/10.1101/2025.02.16.638500; this version posted February 16, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

RIME captures critical features driving RNA-RNA interactions.

a, Left panel: triangular diagram illustrating the relationships between data quality, RIME's confidence, and RIME's performance. Right panel: heatmap showing Pearson correlation coefficients between the three features, calculated for the Protein-based and RIC-seq datasets. ρ_{QC} is the correlation between data quality, represented by the minimum number of reads supporting an interaction, and model confidence, reflected by the mean RIME prediction score. ρ_{QP} was calculated for DRP and DRI tasks, between data quality (minimum number of reads supporting an interaction) and model performance (ROC-AUC), averaged over 100 iterations with negatives undersampled at various thresholds to match the positive set and quality criteria. ρ_{PC} is calculated for DRP and DRI tasks, measuring the Pearson correlation between model performance (calculated using the Precision metric) and model confidence (RIME score), which ranks predicted positive samples across progressively stricter confidence thresholds (100% of predicted positives, top 76%, top 52%, top 28%, top 5%). **b**, Box plots illustrating the distribution of RIME scores across the Psoralen-based, RIC-seq, and MARIO test sets. The scores are categorized based on the overlap of LCRs with the interacting patches: pairs where neither patch overlaps with LCRs, pairs where LCRs overlap with one patch, and pairs where both patches overlap with LCRs. Mann–Whitney U test was used to test for differences between distributions. P-values were corrected with the Benjamini-Hochberg procedure. *, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$; ****, $p < 0.0001$. **c**, Grad-CAM visualization of the top 9 (from left to right, from top to bottom) interactions from the Psoralen-based dataset. Interactions were ranked by averaging separate rankings based on the RIME score and the maximum Grad-CAM intensity value. Grad-CAM intensities were normalized between 0 and 1. The red bounding box highlights the true interacting regions.

Fig. 8

RIMEfull reliably identifies lncRNA-mRNA interactions.

a. Plot summarizing the comparison between COMRADES-derived RRIs and RIME predictions for interactions between NORAD and its 17 mRNA partners. The NORAD sequence is represented from the 5′ end (left) to the 3′ end (right) in 50-nucleotide segments with a 10-nucleotide sliding step. The top bar plot reports, for each segment, the total number of COMRADES-derived chimeric reads connecting NORAD to its 17 target mRNAs across all experimental replicates. The central heatmap shows, for each replicate, the frequency of chimeric reads covering each segment, normalized on the maximum value observed across the segments. The bottom bar plot shows, for each segment, the frequency with which it was predicted as interacting by RIMEfull, considering the top 50 positive (prediction score > 0.5) windows for each target. **b** and **c**. Heatmaps showing the RIMEfull predictions for the TINCR–PGLYRP3 (**b**) and lnc-SMaRT–Mlx-γ (**c**) transcript pairs. Transcripts were segmented using a 200-nucleotide sliding window with a 100-nucleotide step. Each cell reports the RIMEfull prediction score for the interaction between the corresponding RNA windows. Previously validated interacting regions are highlighted with red boxes.

bioRxiv preprint doi: https://doi.org/10.1101/2025.02.16.638500; this version posted February 16, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References

1. Dai, X., Zhang, S. & Zaleta-Rivera, K. RNA: interactions drive functionalities. *Molecular Biology Reports* **47**, 1413–1434 (2019).
2. Romero-Barrios, N., Legascue, M. F., Benhamed, M., Ariel, F. & Crespi, M. Splicing regulation by long noncoding RNAs. *Nucleic Acids Research* **46**, 2169 (2018).
3. Werner, A., Kanhere, A., Wahlestedt, C. & Mattick, J. S. Natural antisense transcripts as versatile regulators of gene expression. *Nature Reviews Genetics* **25**, 730–744 (2024).
4. Singh, S., Shyamal, S. & Panda, A. C. Detecting RNA-RNA interactome. *Wiley interdisciplinary reviews. RNA* **13**, (2022).
5. Bose, M., Rankovic, B., Mahamid, J. & Ephrussi, A. An architectural role of specific RNA–RNA interactions in oskar granules. *Nature Cell Biology* **26**, 1934–1942 (2024).
6. Lu, Z. *et al.* RNA Duplex Map in Living Cells Reveals Higher-Order Transcriptome Structure. *Cell* **165**, 1267–1279 (2016).
7. Zhang, M. *et al.* Optimized photochemistry enables efficient analysis of dynamic RNA structuromes and interactomes in genetic and infectious diseases. *Nature Communications* **12**, 1–14 (2021).
8. Aw, J. G. A., Shen, Y., Nagarajan, N. & Wan, Y. Mapping RNA-RNA Interactions Globally Using Biotinylated Psoralen. *Journal of visualized experiments : JoVE* (2017) doi:10.3791/55255.
9. Sharma, E., Sterne-Weiler, T., O'Hanlon, D. & Blencowe, B. J. Global Mapping of Human RNA-RNA Interactions. *Molecular cell* **62**, (2016).
10. Wu, T. *et al.* KARR-seq reveals cellular higher-order RNA structures and RNA–RNA interactions. *Nature Biotechnology* 1–12 (2024).
11. Cai, Z. *et al.* RIC-seq for global in situ profiling of RNA–RNA spatial interactions. *Nature* **582**, 432–437 (2020).
12. Nguyen, T. C. *et al.* Mapping RNA–RNA interactome and RNA structure in vivo by MARIO. *Nature Communications* **7**, 1–12 (2016).
13. Skeparnias, I. & Zhang, J. Cooperativity and Interdependency between RNA Structure and RNA-RNA Interactions. *Non-coding RNA* **7**, (2021).
14. Tieng, F. Y. F. *et al.* A Hitchhiker's guide to RNA–RNA structure and interaction prediction tools.

bioRxiv preprint doi: https://doi.org/10.1101/2025.02.16.638500; this version posted February 16, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Briefings in Bioinformatics **25**, bbad421 (2023).

15. Adinolfi, M. *et al.* Discovering sequence and structure landscapes in RNA interaction motifs. *Nucleic acids research* **47**, (2019).
16. Kretz, M. *et al.* Control of somatic tissue differentiation by the long non-coding RNA TINCR. *Nature* **493**, (2013).
17. Liang, L. *et al.* Complementary Alu sequences mediate enhancer–promoter selectivity. *Nature* **619**, 868–875 (2023).
18. Duszczyk, M. M., Wutz, A., Rybin, V. & Sattler, M. The Xist RNA A-repeat comprises a novel AUCG tetraloop fold and a platform for multimerization. *RNA* **17**, 1973 (2011).
19. Pintacuda, G., Young, A. N. & Cerase, A. Function by Structure: Spotlights on Xist Long Non-coding RNA. *Front Mol Biosci* **4**, 90 (2017).
20. Verheyden, N. A. *et al.* A high-resolution map of functional miR-181 response elements in the thymus reveals the role of coding sequence targeting and an alternative seed match. *Nucleic Acids Res* **52**, 8515–8533 (2024).
21. Wong, M. S., Shay, J. W. & Wright, W. E. Regulation of human telomerase splicing by RNA:RNA pairing. *Nature Communications* **5**, 1–6 (2014).
22. Gong, C. & Maquat, L. E. lncRNAs transactivate STAU1-mediated mRNA decay by duplexing with 3′ UTRs via Alu elements. *Nature* **470**, 284–288 (2011).
23. Lorenz, R. *et al.* ViennaRNA Package 2.0. *Algorithms for Molecular Biology* **6**, 1–14 (2011).
24. Busch, A., Richter, A. S. & Backofen, R. IntaRNA: efficient prediction of bacterial sRNA targets incorporating target site accessibility and seed regions. *Bioinformatics (Oxford, England)* **24**, (2008).
25. Mann, M., Wright, P. R. & Backofen, R. IntaRNA 2.0: enhanced and customizable prediction of RNA-RNA interactions. *Nucleic Acids Res* **45**, W435–W439 (2017).
26. Umu, S. U. & Gardner, P. P. A comprehensive benchmark of RNA–RNA interaction prediction tools for all domains of life. *Bioinformatics* **33**, 988–996 (2016).
27. Delli Ponti, R., Marti, S., Armaos, A. & Tartaglia, G. G. A high-throughput approach to profile RNA structure. *Nucleic Acids Res.* **45**, e35 (2017).
28. Sato, K., Akiyama, M. & Sakakibara, Y. RNA secondary structure prediction using deep learning with

bioRxiv preprint doi: https://doi.org/10.1101/2025.02.16.638500; this version posted February 16, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

thermodynamic integration. *Nat. Commun.* **12**, 941 (2021).

29. Ji, Y., Zhou, Z., Liu, H. & Davuluri, R. V. DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics (Oxford, England)* **37**, (2021).
30. Dalla-Torre, H. *et al.* Nucleotide Transformer: building and evaluating robust foundation models for human genomics. *Nature Methods* 1–11 (2024).
31. Chung, T. H., Zhuravskaya, A. & Makeyev, E. V. Regulation potential of transcribed simple repeated sequences in developing neurons. *Hum. Genet.* **143**, 875–895 (2024).
32. Smit, A. F. A., Hubley, R. & Green, P. *RepeatMasker Open-4.0*.
33. Frith, M. C. A new repeat-masking method enables specific detection of homologous sequences. *Nucleic Acids Res* **39**, e23 (2011).
34. Lu, J. Y. *et al.* Genomic Repeats Categorize Genes with Distinct Functions for Orchestrated Regulation. *Cell Rep* **30**, 3296–3311.e5 (2020).
35. Pellegrini, F. *et al.* A KO mouse model for the lncRNA Lhx1os produces motor neuron alterations and locomotor impairment. *iScience* **26**, (2022).
36. Mückstein, U. *et al.* Thermodynamics of RNA-RNA binding. *Bioinformatics* **22**, 1177–1182 (2006).
37. Rehmsmeier, M., Steffen, P., Hochsmann, M. & Giegerich, R. Fast and effective prediction of microRNA/target duplexes. *RNA* **10**, 1507–1517 (2004).
38. Antonov, I. V., Mazurov, E., Borodovsky, M. & Medvedeva, Y. A. Prediction of lncRNAs and their interactions with nucleic acids: benchmarking bioinformatics tools. *Brief. Bioinform.* **20**, 551–564 (2019).
39. Kelley, D. R., Snoek, J. & Rinn, J. L. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Res* **26**, 990–999 (2016).
40. Zhou, J. & Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning-based sequence model. *Nature methods* **12**, (2015).
41. Avsec, Ž. *et al.* Effective gene expression prediction from sequence by integrating long-range interactions. *Nature Methods* **18**, 1196–1203 (2021).
42. Antonov, I., Marakhonov, A., Zamkova, M. & Medvedeva, Y. ASSA: Fast identification of statistically

bioRxiv preprint doi: https://doi.org/10.1101/2025.02.16.638500; this version posted February 16, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

significant interactions between long RNAs. *Journal of bioinformatics and computational biology* **16**, (2018).

43. Tafer, H. & Hofacker, I. L. RNAplex: a fast tool for RNA-RNA interaction search. *Bioinformatics (Oxford, England)* **24**, (2008).
44. Alkan, F. *et al.* RIsearch2: suffix array-based large-scale prediction of RNA-RNA interactions and siRNA off-targets. *Nucleic acids research* **45**, (2017).
45. Amatria-Barral, I., González-Domínguez, J. & Touriño, J. pRIblast: A highly efficient parallel application for comprehensive lncRNA–RNA interaction prediction. *Future Generation Computer Systems* **138**, 270–279 (2023).
46. Mückstein, U. *et al.* Translational control by RNA-RNA interaction: Improved computation of RNA-RNA binding thermodynamics. in *Communications in Computer and Information Science* 114–127 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2008).
47. Bernhart, S. H. *et al.* Partition function and base pairing probabilities of RNA heterodimers. *Algorithms Mol. Biol.* **1**, 3 (2006).
48. Selvaraju, R. R. *et al.* Grad-CAM: Visual explanations from deep networks via gradient-based localization. in *2017 IEEE International Conference on Computer Vision (ICCV)* 618–626 (IEEE, 2017).
49. Farberov, S. *et al.* Structural features within the NORAD long noncoding RNA underlie efficient repression of Pumilio activity. *Nat. Struct. Mol. Biol.* (2024) doi:10.1038/s41594-024-01393-5.
50. Martone, J. *et al.* SMaRT lncRNA controls translation of a G-quadruplex-containing mRNA antagonizing the DHX36 helicase. *EMBO Rep.* **21**, e49942 (2020).
51. Scalzitti, S. *et al.* Lnc-SMaRT translational regulation of Spire1, A new player in muscle differentiation. *J. Mol. Biol.* **434**, 167384 (2022).
52. Davidson, E. H. & Britten, R. J. Regulation of gene expression: possible role of repetitive sequences. *Science* **204**, 1052–1059 (1979).
53. Trigiante, G., Blanes Ruiz, N. & Cerase, A. Emerging roles of repetitive and repeat-containing RNA in nuclear and chromatin organization and gene expression. *Front. Cell Dev. Biol.* **9**, 735527 (2021).
54. Rué, L. *et al.* Targeting CAG repeat RNAs reduces Huntington's disease phenotype independently of

bioRxiv preprint doi: https://doi.org/10.1101/2025.02.16.638500; this version posted February 16, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

huntingtin levels. *J. Clin. Invest.* **126**, 4319–4330 (2016).

55. Trost, B. *et al.* Genome-wide detection of tandem DNA repeats that are expanded in autism. *Nature* **586**, 80–86 (2020).
56. DeJesus-Hernandez, M. *et al.* Expanded GGGGCC hexanucleotide repeat in noncoding region of C9ORF72 causes chromosome 9p-linked FTD and ALS. *Neuron* **72**, 245–256 (2011).
57. Vandelli, A., Cid Samper, F., Torrent Burgas, M., Sanchez de Groot, N. & Tartaglia, G. G. The interplay between disordered regions in RNAs and proteins modulates interactions within Stress Granules and Processing Bodies. *J. Mol. Biol.* **434**, 167159 (2022).
58. Wojciechowska, M. & Krzyzosiak, W. J. Cellular toxicity of expanded RNA repeats: focus on RNA foci. *Hum. Mol. Genet.* **20**, 3811–3821 (2011).
59. Jain, A. & Vale, R. D. RNA phase transitions in repeat expansion disorders. *Nature* **546**, 243–247 (2017).
60. Sanders, D. W. & Brangwynne, C. P. Neurodegenerative disease: RNA repeats put a freeze on cells. *Nature* **546**, 215–216 (2017).
61. Fiorentino, J., Armaos, A., Colantoni, A. & Tartaglia, G. G. Prediction of protein-RNA interactions from single-cell transcriptomic data. *Nucleic Acids Res.* **52**, e31 (2024).
62. Gong, J. *et al.* RISE: a database of RNA interactome from sequencing experiments. *Nucleic acids research* **46**, (2018).
63. Aw, J. G. A. *et al.* In Vivo Mapping of Eukaryotic RNA Interactomes Reveals Principles of Higher-Order Organization and Regulation. *Mol Cell* **62**, 603–617 (2016).
64. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* **44**, D733–45 (2016).
65. Speir, M. L. *et al.* The UCSC Genome Browser database: 2016 update. *Nucleic Acids Res* **44**, D717–25 (2016).
66. Yates, A. D. *et al.* Ensembl 2020. *Nucleic acids research* **48**, (2020).
67. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)* **26**, (2010).
68. Durinck, S., Spellman, P. T., Birney, E. & Huber, W. Mapping identifiers for the integration of genomic

bioRxiv preprint doi: https://doi.org/10.1101/2025.02.16.638500; this version posted February 16, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

datasets with the R/Bioconductor package biomaRt. *Nature protocols* **4**, (2009).

69. Abdennur, N. *et al.* Bioframe: operations on genomic intervals in Pandas dataframes. *Bioinformatics (Oxford, England)* **40**, (2024).
70. Smedley, D. *et al.* BioMart--biological queries made easy. *BMC genomics* **10**, (2009).
71. Liao, Y., Wang, J., Jaehnig, E. J., Shi, Z. & Zhang, B. WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. *Nucleic acids research* **47**, (2019).
72. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics (Oxford, England)* **32**, (2016).
73. Hagberg, A. A., Schult, D. A. & Swart, P. J. Exploring Network Structure, Dynamics, and Function using NetworkX. *scipy* (2008) doi:10.25080/TCWV9851.
74. Wichterle, H. & Peljto, M. Differentiation of mouse embryonic stem cells to spinal motor neurons. *Curr. Protoc. Stem Cell Biol.* **Chapter 1**, Unit 1H.1.1–1H.1.9 (2008).
75. Errichelli, L. *et al.* FUS affects circular RNA expression in murine embryonic stem cell-derived motor neurons. *Nat. Commun.* **8**, 14741 (2017).
76. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10–12 (2011).
77. Sayers, E. W. *et al.* Database resources of the national center for biotechnology information. *Nucleic Acids Res.* **50**, D20–D26 (2022).
78. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature methods* **9**, 357 (2012).
79. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
80. Broad Institute. Picard Toolkit. *Broad Institute, GitHub repository*. https://broadinstitute.github.io/picard/ (2019).
81. Anders, S., Pyl, P. T. & Huber, W. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).
82. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
83. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware

bioRxiv preprint doi: https://doi.org/10.1101/2025.02.16.638500; this version posted February 16, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

quantification of transcript expression. *Nature Methods* **14**, 417–419 (2017).

84. Fu, L. *et al.* UFold: fast and accurate RNA secondary structure prediction with deep learning. *Nucleic Acids Res.* **50**, e14 (2022).

85. Paszke, A. *et al.* PyTorch: An imperative style, high-performance deep learning library. *arXiv [cs.LG]* (2019).

Figure 1

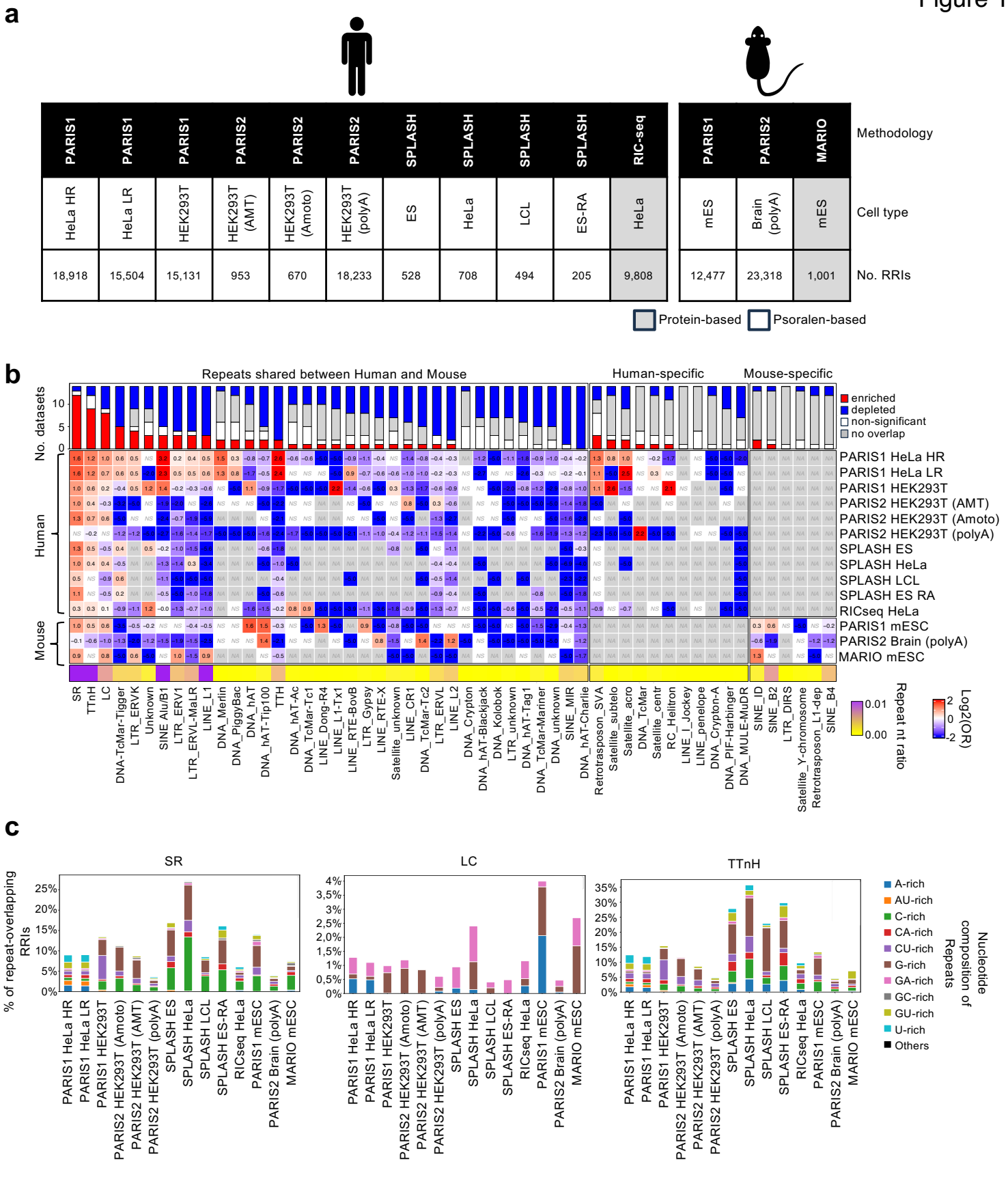
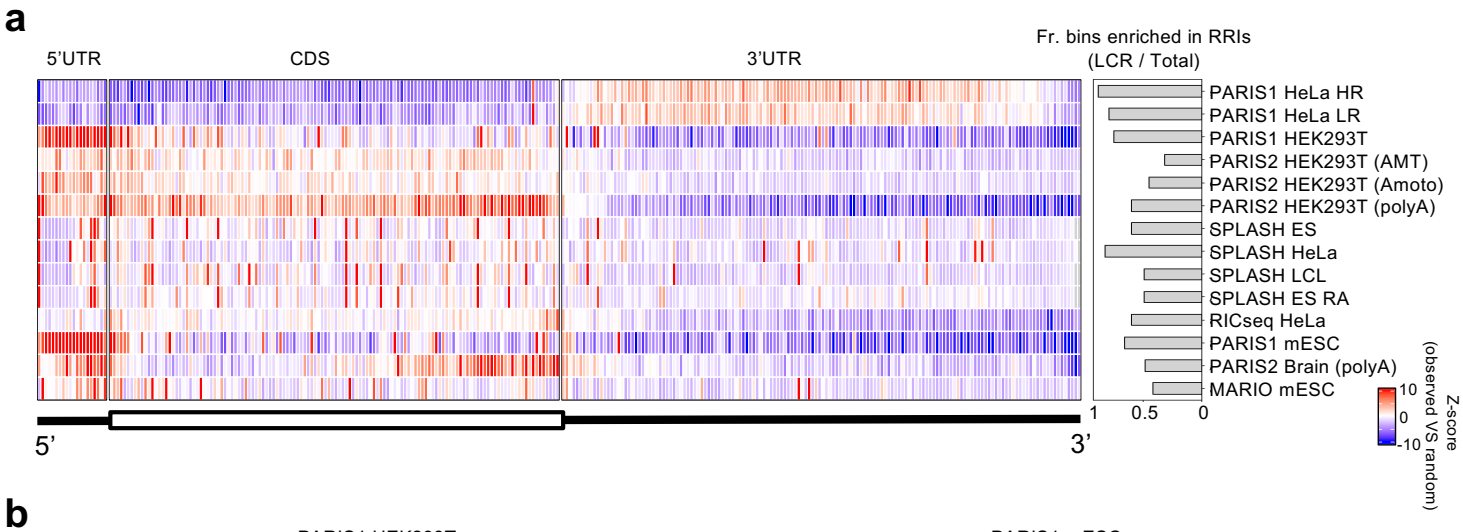
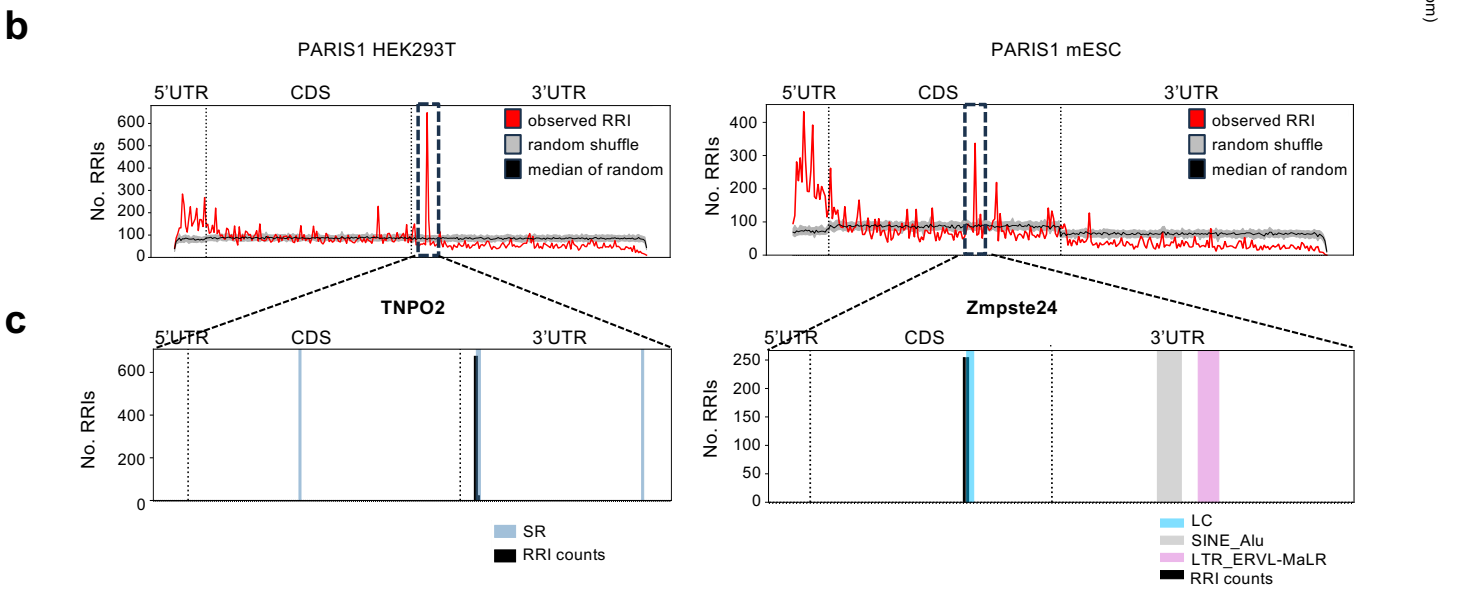
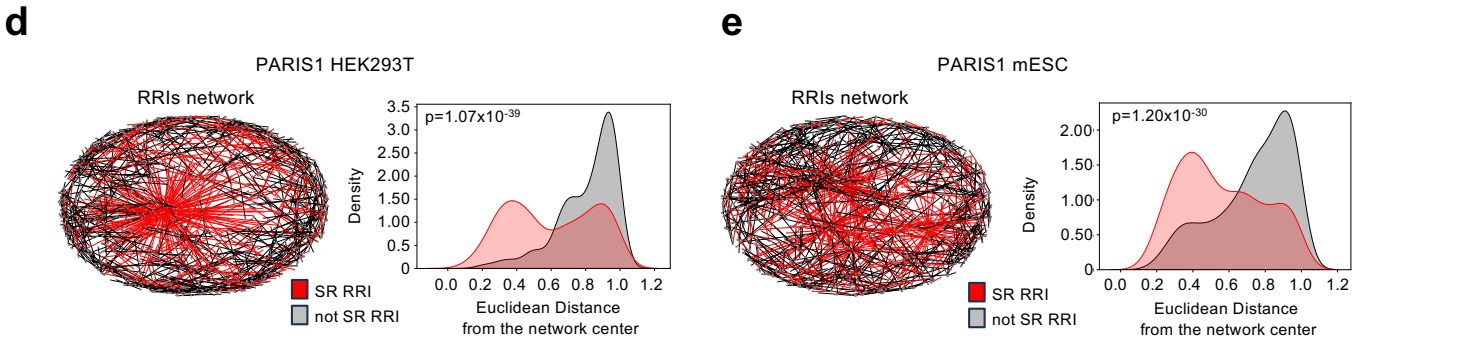

a
PARIS1 PARIS1 PARIS1 PARIS2 PARIS2 PARIS2 SPLASH SPLASH SPLASH SPLASH RIC-seq
PARIS1 PARIS2 MARIO
Methodology
HeLa HR | HeLa LR | HEK293T | HEK293T (AMT) | HEK293T (Amoto) | HEK293T (polyA) | ES | HeLa | LCL | ES-RA | HeLa
mES | Brain (polyA) | mES
Cell type
18,918 | 15,504 | 15,131 | 953 | 670 | 18,233 | 528 | 708 | 494 | 205 | 9,808
12,477 | 23,318 | 1,001
No. RRIs
Protein-based
Psoralen-based
b
Repeats shared between Human and Mouse
Human-specific
Mouse-specific
No. datasets
enriched
depleted
non-significant
no overlap
Human
Mouse
PARIS1 HeLa HR
PARIS1 HeLa LR
PARIS1 HEK293T
PARIS2 HEK293T (AMT)
PARIS2 HEK293T (Amoto)
PARIS2 HEK293T (polyA)
SPLASH ES
SPLASH HeLa
SPLASH LCL
SPLASH ES RA
RICseq HeLa
PARIS1 mESC
PARIS2 Brain (polyA)
MARIO mESC
SR TTnH LC DNA-TcMar-Tigger LTR_ERVK Unknown SINE Alu/B1 LTR_ERV1 LTR_ERVL-MaLR LINE_L1 DNA_Merlin DNA_PiggyBac DNA_hAT DNA_hAT-Tip100 TTH DNA_hAT-Ac DNA_TcMar-Tc1 LINE_Dong-R4 LINE_L1-Tx1 LINE_RTE-BovB LTR_Gypsy LINE_RTE-X Satellite_unknown LINE_CR1 DNA_TcMar-Tc2 LTR_ERVL LINE_L2 DNA_Crypton DNA_hAT-Blackjack DNA_Kolobok LTR_unknown DNA_hAT-Tag1 DNA_TcMar-Mariner DNA_unknown SINE_MIR DNA_hAT-Charlie Retrotrasposon_SVA Satellite_subtelo Satellite_acro DNA_TcMar Satellite_centr RC_Helitron LINE_I_Jockey LINE_penelope DNA_Crypton-A DNA_PIF-Harbinger DNA_MULE-MuDR SINE_ID SINE_B2 LTR_DIRS Satellite_Y-chromosome Retrotrasposon_L1-dep SINE_B4
Repeat ratio
Log2(OR)
c
SR
LC
TTnH
% of repeat-overlapping RRIs
PARIS1 HeLa HR
PARIS1 HeLa LR
PARIS1 HEK293T
PARIS2 HEK293T (Amoto)
PARIS2 HEK293T (AMT)
PARIS2 HEK293T (polyA)
SPLASH ES
SPLASH HeLa
SPLASH LCL
SPLASH ES-RA
RICseq HeLa
PARIS1 mESC
PARIS2 Brain (polyA)
MARIO mESC
Nucleotide composition of Repeats
A-rich
AU-rich
C-rich
CA-rich
CU-rich
G-rich
GA-rich
GC-rich
GU-rich
U-rich
Others

Figure 2


a
5'UTR
CDS
3'UTR
Fr. bins enriched in RRIs
(LCR / Total)
PARIS1 HeLa HR
PARIS1 HeLa LR
PARIS1 HEK293T
PARIS2 HEK293T (AMT)
PARIS2 HEK293T (Amoto)
PARIS2 HEK293T (polyA)
SPLASH ES
SPLASH HeLa
SPLASH LCL
SPLASH ES RA
RICseq HeLa
PARIS1 mESC
PARIS2 Brain (polyA)
MARIO mESC
1
0.5
0
5'
3'
Z-score
(observed VS random)
10
0
-10


b
PARIS1 HEK293T
PARIS1 mESC
5'UTR
CDS
3'UTR
No. RRIs
observed RRI
random shuffle
median of random
c
TNPO2
Zmpste24
SR
RRI counts
LC
SINE_Alu
LTR_ERVL-MaLR
RRI counts


d
PARIS1 HEK293T
RRIs network
SR RRI
not SR RRI
p=1.07x10-39
Density
Euclidean Distance
from the network center
e
PARIS1 mESC
RRIs network
SR RRI
not SR RRI
p=1.20x10-30
Density
Euclidean Distance
from the network center

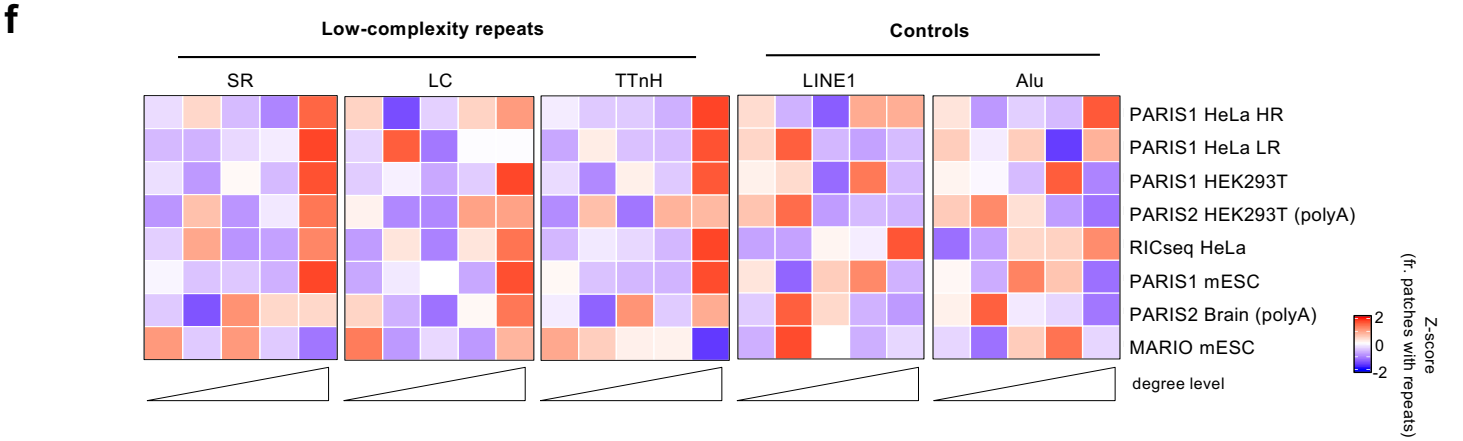

f
Low-complexity repeats
Controls
SR
LC
TTnH
LINE1
Alu
PARIS1 HeLa HR
PARIS1 HeLa LR
PARIS1 HEK293T
PARIS2 HEK293T (polyA)
RICseq HeLa
PARIS1 mESC
PARIS2 Brain (polyA)
MARIO mESC
degree level
Z-score
(fr. patches with repeats)
2
0
-2

Figure 3

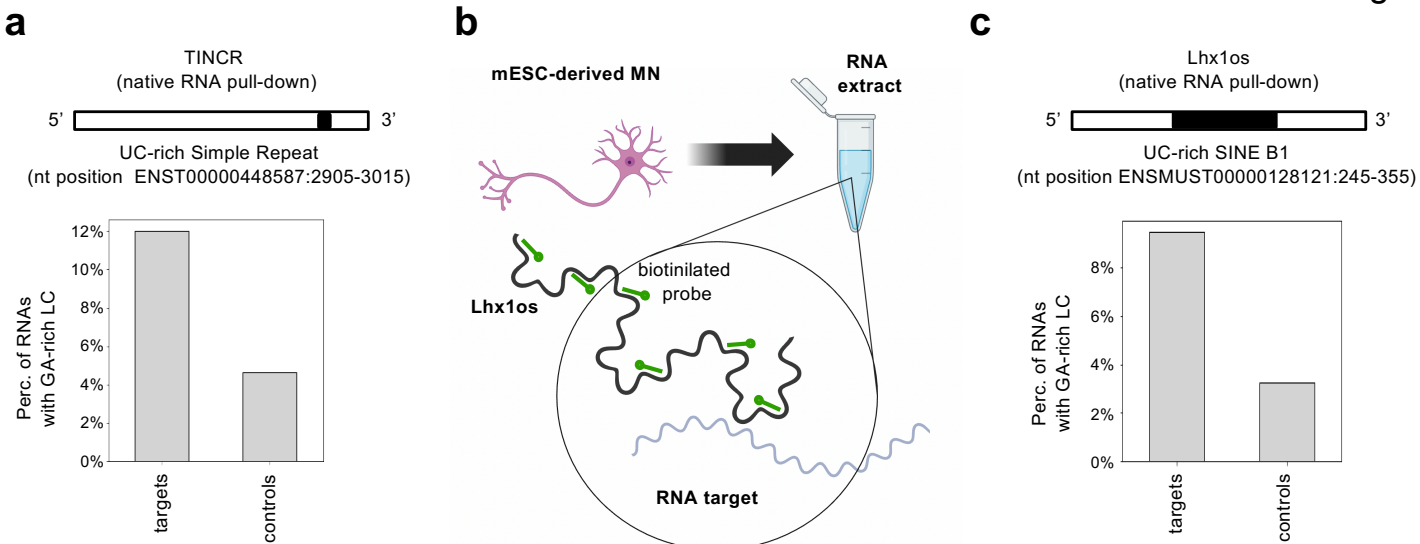
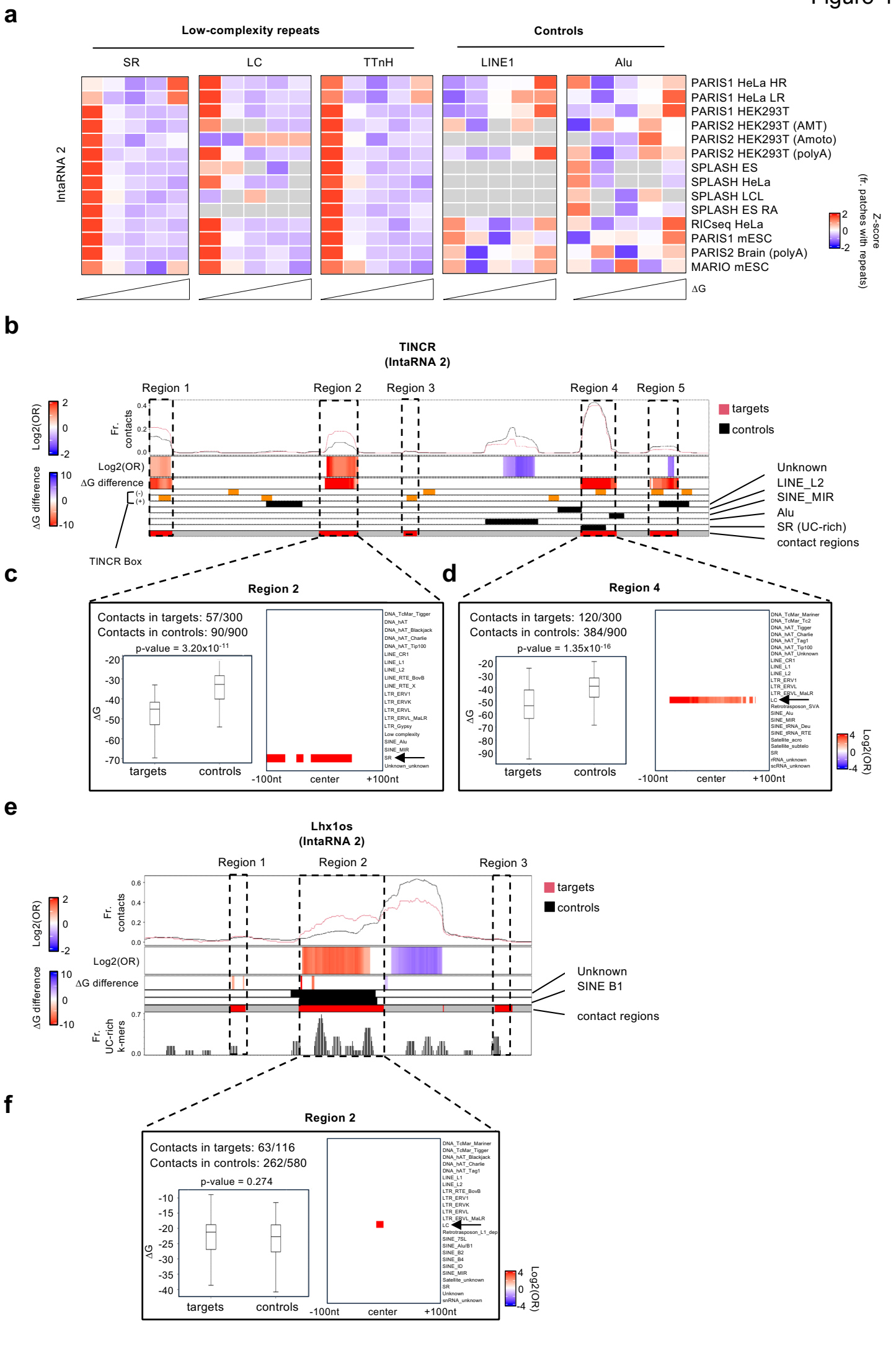
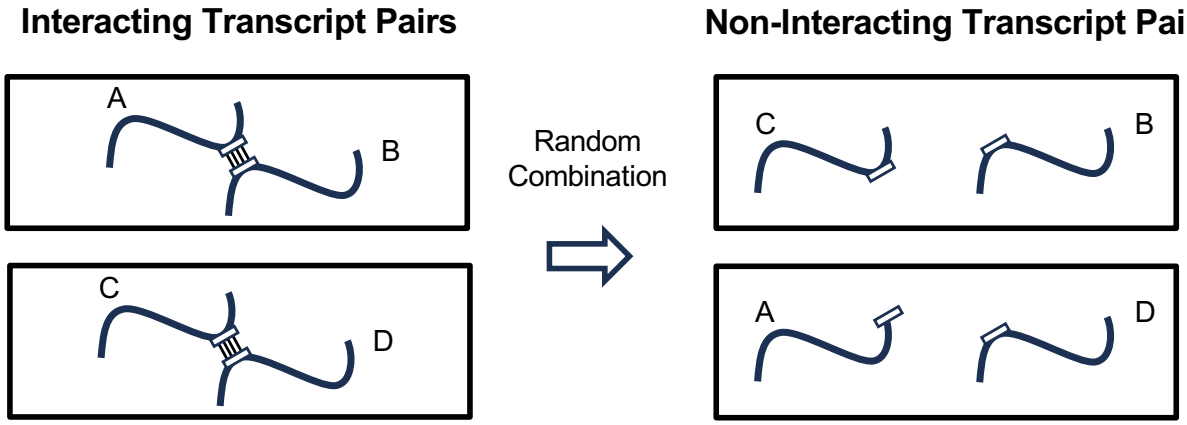

a
TINCR
(native RNA pull-down)
5'
3'
UC-rich Simple Repeat
(nt position ENST00000448587:2905-3015)
Perc. of RNAs with GA-rich LC
12%
10%
8%
6%
4%
2%
0%
targets
controls
b
mESC-derived MN
RNA extract
biotinilated probe
Lhx1os
RNA target
c
Lhx1os
(native RNA pull-down)
5'
3'
UC-rich SINE B1
(nt position ENSMUST00000128121:245-355)
Perc. of RNAs with GA-rich LC
8%
6%
4%
2%
0%
targets
controls

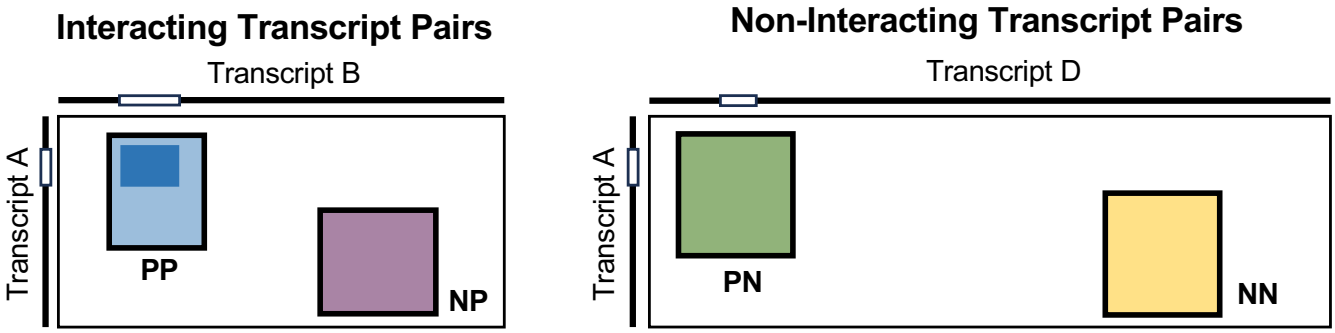
Figure 4


a
Low-complexity repeats
Controls
SR
LC
TTnH
LINE1
Alu
IntaRNA 2
PARIS1 HeLa HR
PARIS1 HeLa LR
PARIS1 HEK293T
PARIS2 HEK293T (AMT)
PARIS2 HEK293T (Amoto)
PARIS2 HEK293T (polyA)
SPLASH ES
SPLASH HeLa
SPLASH LCL
SPLASH ES RA
RICseq HeLa
PARIS1 mESC
PARIS2 Brain (polyA)
MARIO mESC
ΔG
Z-score
(fr. patches with repeats)
b
TINCR
(IntaRNA 2)
Region 1
Region 2
Region 3
Region 4
Region 5
Log2(OR)
ΔG difference
Fr. contacts
targets
controls
Unknown
LINE_L2
SINE_MIR
Alu
SR (UC-rich)
contact regions
TINCR Box
c
Region 2
Contacts in targets: 57/300
Contacts in controls: 90/900
p-value = 3.20x10-11
ΔG
targets
controls
-100nt
center
+100nt
d
Region 4
Contacts in targets: 120/300
Contacts in controls: 384/900
p-value = 1.35x10-16
e
Lhx1os
(IntaRNA 2)
Region 1
Region 2
Region 3
Fr. UC-rich k-mers
Unknown
SINE B1
contact regions
f
Region 2
Contacts in targets: 63/116
Contacts in controls: 262/580
p-value = 0.274

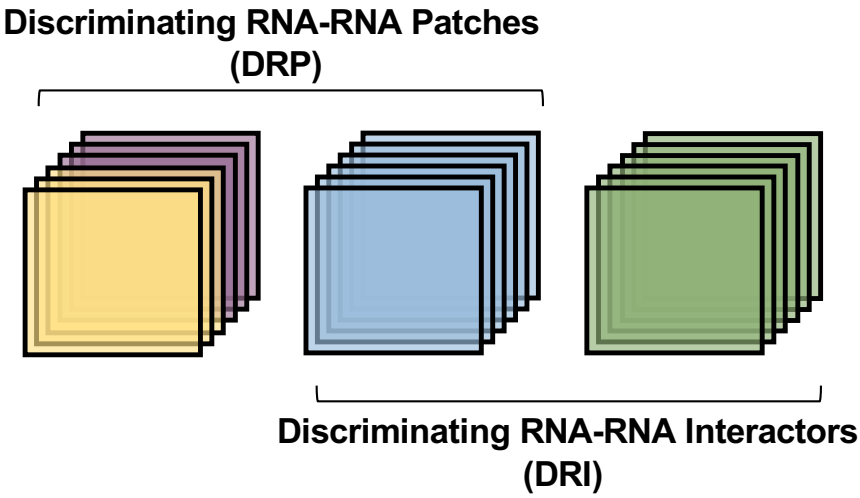
a



b



c

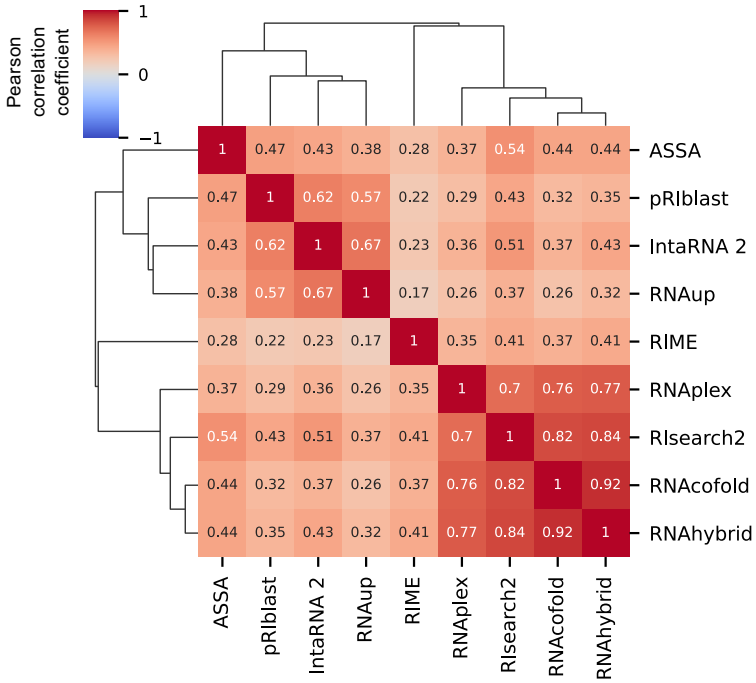


d

Psoralen-based	Human			Mouse		
	RRIs	Interacting Transcript Pairs	Non-Interacting Transcript Pairs	RRIs	Interacting Transcript Pairs	Non-Interacting Transcript Pairs
Training	49,369	45,264	47,934	25,139	24,243	23,914
Validation	10,417	9,532	10,062	5,327	5,200	5,044
Test	10,770	9,853	10,363	5,397	5,199	5,087

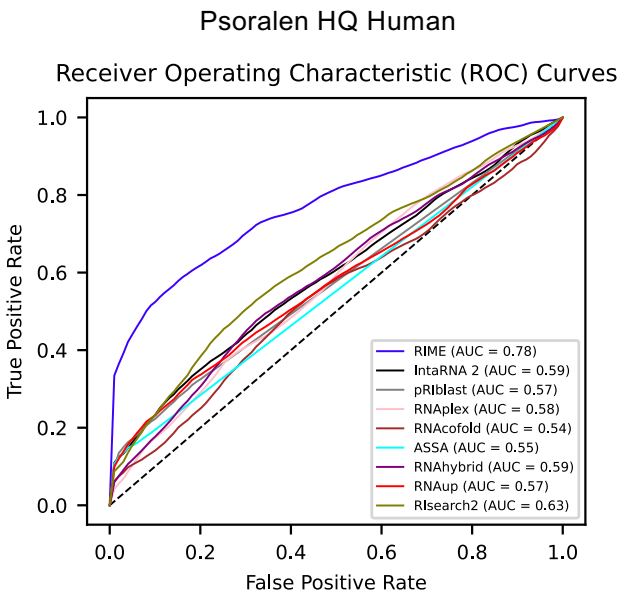
Figure 6

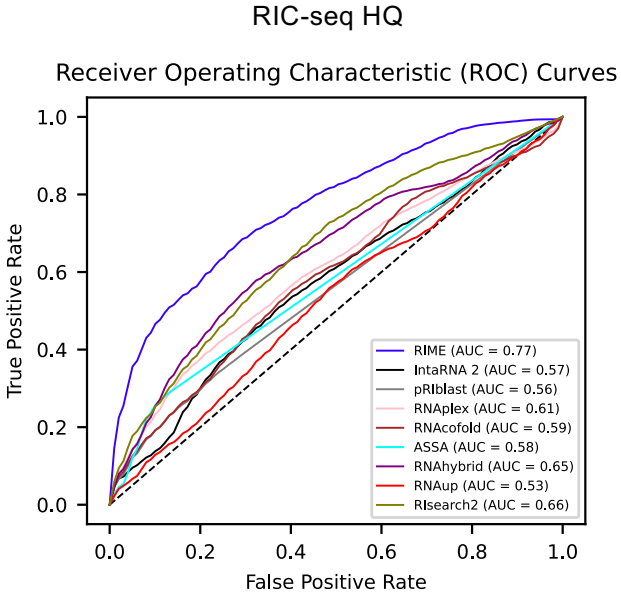
a



b

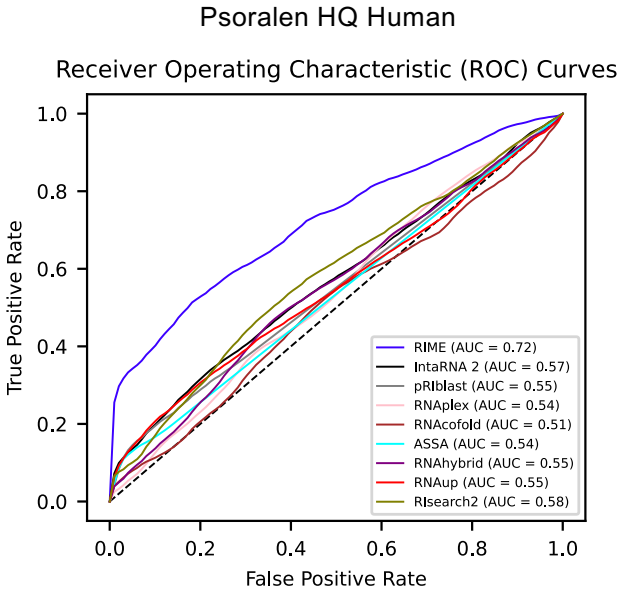
Discriminating RNA-RNA Patches (DRP)

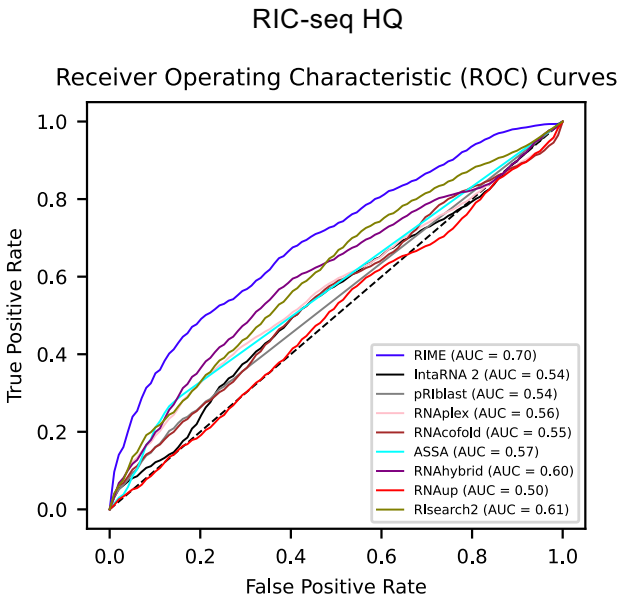




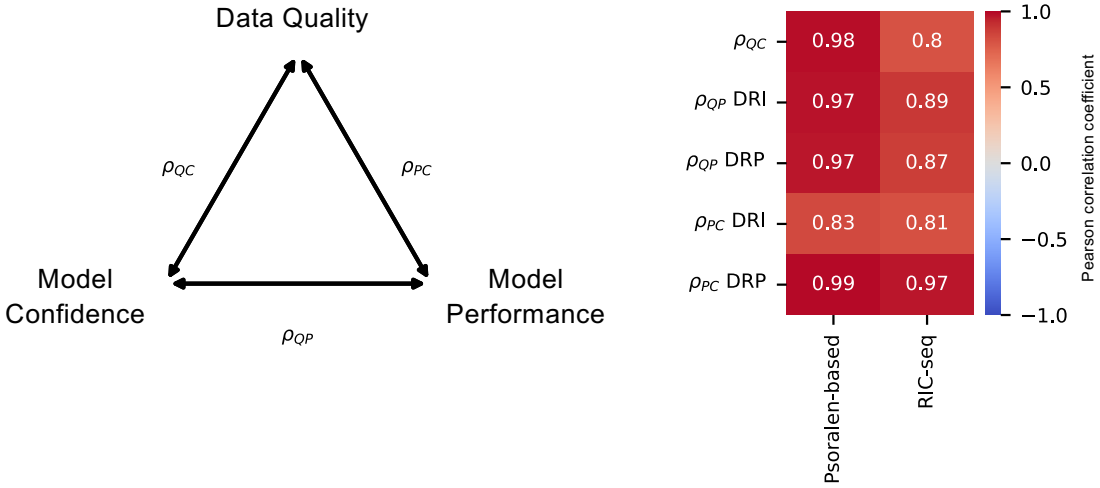
c

Discriminating RNA-RNA Interactors (DRI)

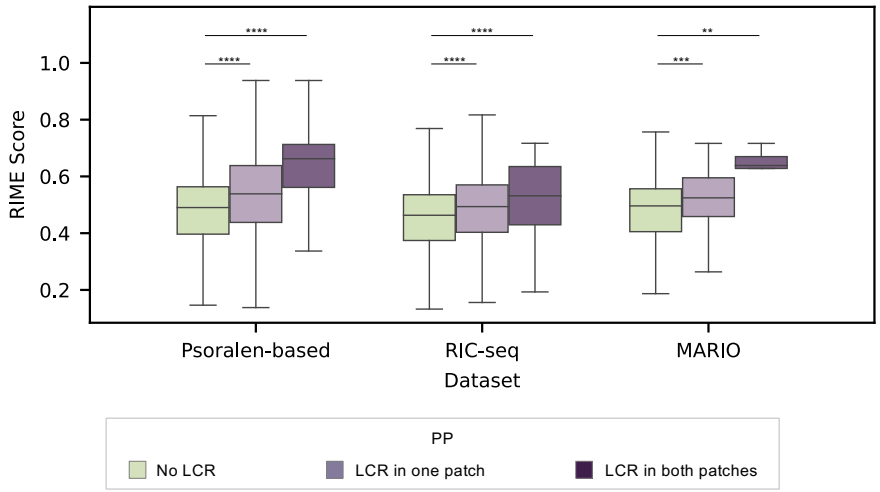




a


Data Quality
ρ_{QC}
ρ_{PC}
Model Confidence
Model Performance
ρ_{QP}
ρ_{QC} 0.98 0.8
ρ_{QP} DRI 0.97 0.89
ρ_{QP} DRP 0.97 0.87
ρ_{PC} DRI 0.83 0.81
ρ_{PC} DRP 0.99 0.97
Psoralen-based
RIC-seq
1.0
0.5
0.0
−0.5
−1.0
Pearson correlation coefficient

b


RIME Score
1.0
0.8
0.6
0.4
0.2
****
****
****
****
**
***
Psoralen-based
RIC-seq
MARIO
Dataset
PP
No LCR
LCR in one patch
LCR in both patches

c

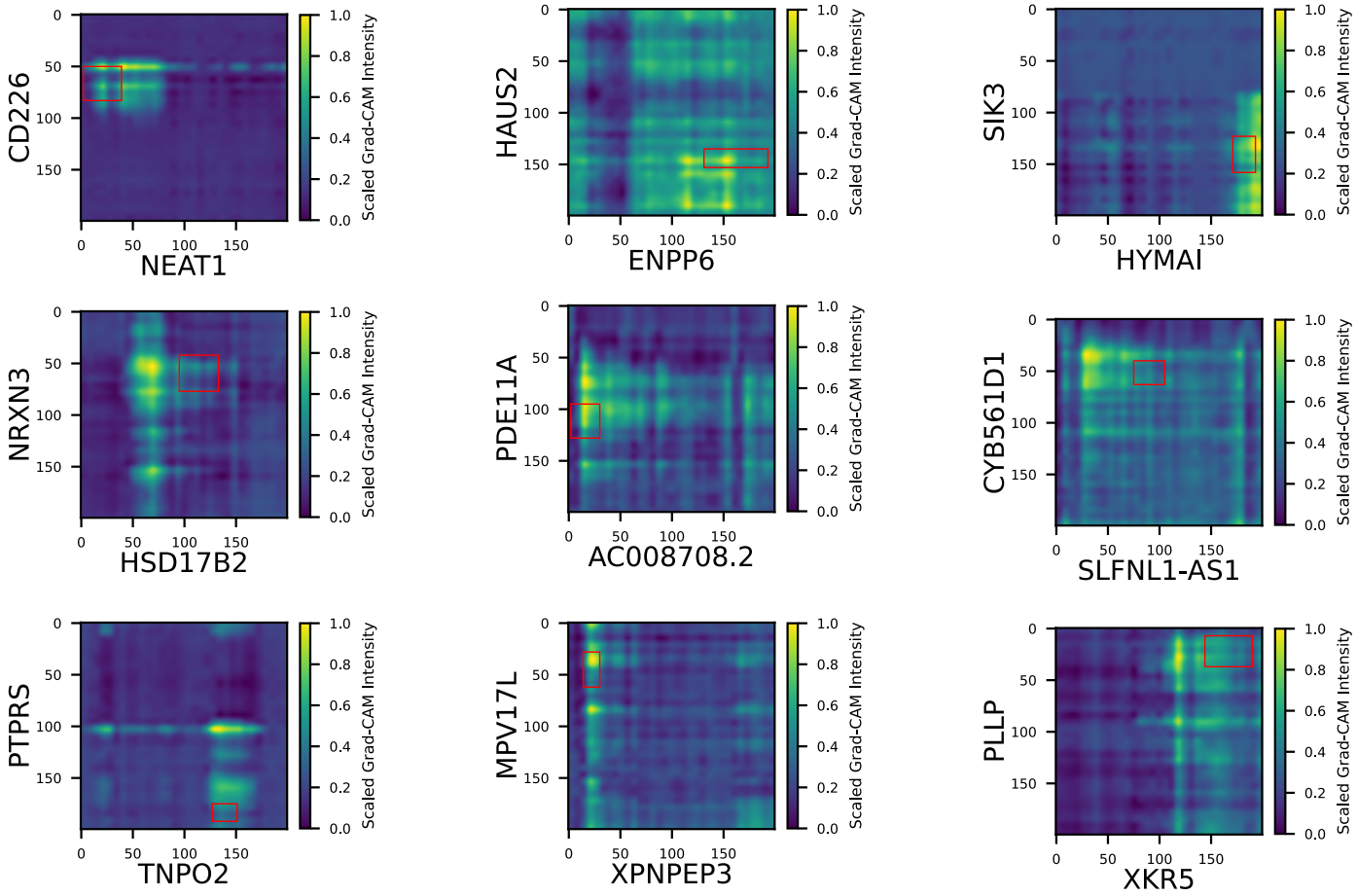
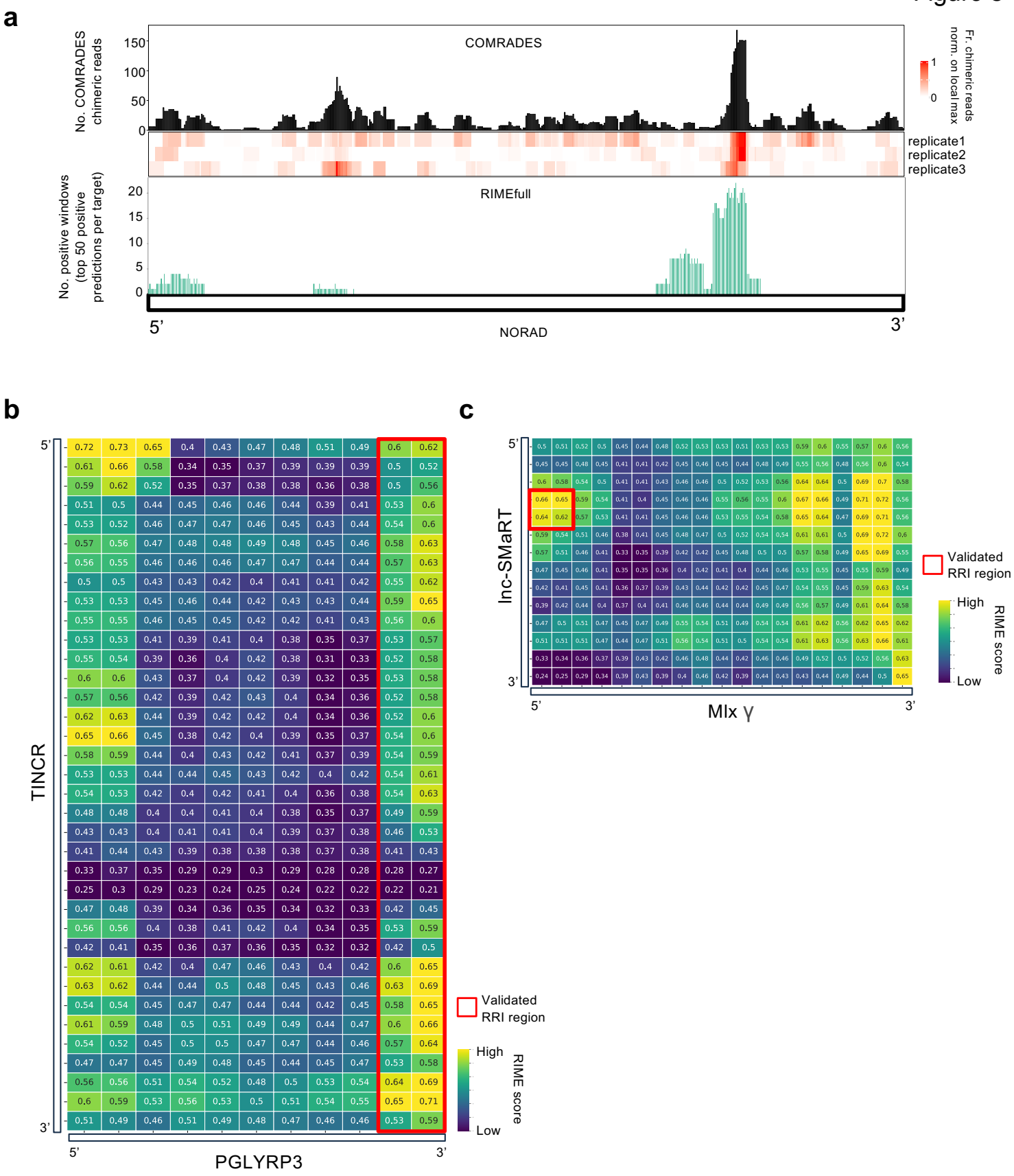

CD226
NEAT1
HAUS2
ENPP6
SIK3
HYMAI
NRXN3
HSD17B2
PDE11A
AC008708.2
CYB561D1
SLFNL1-AS1
PTPRS
TNPO2
MPV17L
XPNPEP3
PLLP
XKR5
Scaled Grad-CAM Intensity

Figure 8


a
No. COMRADES chimeric reads
COMRADES
Fr. chimeric reads norm. on local max
replicate1
replicate2
replicate3
No. positive windows (top 50 positive predictions per target)
RIMEfull
5'
NORAD
3'
b
TINCR
PGLYRP3
Validated RRI region
High
Low
RIME score
c
lnc-SMaRT
Mlx γ
Validated RRI region
High
Low
RIME score