Integrative analysis of higher-order transcriptome organisation and RNA condensation principles

Tajda Klobučar*[1,2,5], Jona Novljan*[1], Ira A. Iosub*[2,3,4], Boštjan Kokot[6], Iztok Urbančič[6], D. Marc Jones[2], Anob M. Chakrabarti[2,7], Nicholas M. Luscombe[2,8], Jernej Ule[1,2,3,4], Miha Modic[1,2,3,4,#]

* These authors contributed equally

1 National Institute of Chemistry, Ljubljana, Slovenia

2 The Francis Crick Institute, London, UK

3 Dementia Research Institute at KCL, London, UK

4 Department of Basic and Clinical Neuroscience, Institute of Psychiatry, Psychology and Neuroscience, King's College London, UK

5 PhD Program 'Biosciences', Biotechnical Faculty, University of Ljubljana, Ljubljana, Slovenia

6 The Jožef Stefan Institute, Ljubljana, Slovenia

7 University College London, UCL Respiratory, London, United Kingdom

8 Okinawa Institute of Science and Technology, Okinawa, Japan

miha.modic@ki.si

Abstract

In the crowded cellular environment, the transcriptome is highly organised with a network of interactions between proteins and RNA molecules that are crucial for the formation of biomolecular condensates. While the role of proteins in this process has been extensively studied, the contribution of RNA has remained underexplored. In this study, we identify a distinct set of RNAs that are prone to condensation during early embryonic development and apply an integrative deep learning approach to deconvolve the sequence and structural determinants, and trans-acting regulators involved in RNA condensation. These RNAs form a highly interconnected subnetwork of RNA-RNA interactions with distinct local connectivity patterns, and encode proteins with intrinsically-disordered regions and a propensity for phase separation. We find that although the identity of condensation-prone RNAs differs across developmental stages, their organisational principle remains conserved. This suggests a multi-layered coordination of genes driving both RNA and protein phase-separating processes. This work highlights the unique organisation of transcripts encoding proteins with similar biophysical properties and provides insights into the principles governing RNA assembly within biomolecular condensates during early embryonic development.

Introduction

Macromolecular complexes, particularly those rich in RNA, physically partition the cellular milieu to create subcellular membraneless compartments also termed biomolecular condensates. These are mostly formed through phase separation, which depends on weak multivalent interactions between proteins and their cognate targets: RNAs and other proteins. The understanding of the molecular

composition of ribonucleoprotein (RNP) condensates, their physical properties, assembly principles, as well as biological significance has largely been focused towards the constituent proteins (Alberti and Hyman, 2021; Mittag and Pappu, 2022), for which simple changes to protein sequence or charge can drastically alter condensation properties (Patel et al., 2015; Wang et al., 2018). However, RNA molecules can also drive phase transitions, not acting solely as a scaffold for proteins but also by self-organising into RNA assemblies (Maharana et al., 2018; Trcek et al., 2020), thereby playing a critical role in regulating condensate assembly (Cabral et al., 2022; Han et al., 2024; Van Treeck et al., 2018). This is often achieved through a myriad of intermolecular interactions with other RNAs, which determine the content, biophysical properties and shape of RNP condensates (Eno et al., 2019; Guillén-Boixet et al., 2020; Khong et al., 2017; Tauber et al., 2020) in turn affecting processes like mRNA translation, stability, and stress responses(Fay and Anderson, 2018; Van Treeck et al., 2018).

While it is known that phase separation steers the formation of RNP condensates that are critical for development and disease, the composition and dynamics of most endogenous RNP condensates remains a significant challenge (Roden and Gladfelter, 2021; Spitale and Incarnato, 2023). Deciphering how specific RNA sequences, structures, and modifications contribute to condensate formation and function remains complex, especially given the dynamic and heterogeneous nature of RNA molecules, which adds further challenges to understanding their role in condensate biology. Furthermore, the ability of RNA-driven condensates to respond to cellular cues allows them to actively participate in regulating gene expression, stress responses, and signal transduction pathways(Knuckles et al., 2017; Wilbertz et al., 2019). Disruptions in these systems, such as misregulation of RNA condensation or aberrant phase separation, are linked to pathological conditions including impaired embryonic development, cancers, neurodegenerative diseases, and others (Boija et al., 2021; Modic et al., 2019; Ramaswami et al., 2013). Therefore, a deeper understanding of the principles behind RNA condensation and its role during both health development and disease is crucial for unravelling the broader implications of phase-separated cellular compartments (Alberti and Hyman, 2021).

To delineate the cartography of RNP condensates it is imperative to understand how RNA binding proteins (RBPs) interact with RNA and how this interplay influences RNA's higher-order organisation. Although molecular biology has been effective in providing tools to study protein-RNA interactions (Hafner et al., 2021) we are only beginning to unravel how RNA species and RNA binding proteins (RBPs) are organised in space. For this, various methods have identified a plethora of RNA-RNA interactions (RRIs) that can be mediated either by base pairing or in association with RBPs. Transcriptome-wide RNA proximity-ligation approaches and their improved sensitivity (reviewed in Ye et al., 2024) enabled novel insights into functional RRIs, such as those between enhancer RNAs/mRNAs (Cai et al., 2020), snoRNA/target RNAs (Wu et al., 2024), and viral/host transcripts (Cai et al., 2024). While these techniques identified specific pairwise interactions, the potential of network-based approaches for leveraging proximity ligation datasets to delineate higher-order transcriptome organisational domains and assembly principles at the RNA level remains underexplored.

Here we investigate the complex organisation of the higher-order transcriptome during early pluripotency transitions. We designated a novel class of transcripts as ‘smOOPs’ due to their semi-extractability and pronounced capacity for orthogonal organic phase separation (OOPS). Using a combinatorial deep learning approach, we deconvolved the core smOOPs features and revealed that

these condensation-prone transcripts code for similar highly disordered proteins with elevated phase-separation propensity. We additionally explored potential mechanisms of coordinated regulation to discover that smOOPs form highly interconnected RRI subnetworks that organise RNAs with similar features into densely interconnected substructures. Thereby this study provides a methodological framework to understand the complexity of higher-order transcriptome organisation and lays the foundation to uncover the principles of interconnected RNA assemblies, along with their potential role in coordinating post-transcriptional gene expression regulation.

Results

Atlas of condensation-prone RNAs across early development

To isolate key RNA components of biomolecular condensates, methods leveraging the biochemical properties of RNAs and their tendency to phase separate or associate with RBPs can be employed. We utilised a semi-extractability assay (Chujo et al., 2017) to capture RNP-entangled RNA molecules, and combined it with an orthogonal organic phase separation approach (OOPS; Queiroz et al., 2020), which specifically identifies highly RBP-bound RNAs upon UV-crosslinking. Together, these two methods provide a comprehensive strategy to identify condensation-prone RNAs. A previous study suggested that semi-extractable RNAs represent a diverse array of RNA species involved in various types of biomolecular condensates (Zeng et al., 2023). We reasoned that the RNAs enriched in both the semi-extractability assay and OOPS likely represent condensation-prone RNAs or key RNA components within RNP assemblies **(Figure 1A)**. We applied these methods to identify condensation-prone RNAs during three distinct timepoints of early embryonic development; naïve pluripotent stem cells (nPSC) in 2iLif conditions (Ying et al., 2008), primed epiblast stem cells (pPSC), in which lineage priming was prevented by Wnt inhibition (Kinoshita et al., 2021; Kurek et al., 2015), and the earliest Wnt-differentiated primitive streak progenitors (dPSC) (Gouti et al., 2014). Using a standard TRIzol RNA extraction protocol as a control, we used heating and shearing to release the semi-extractable RNAs into the aqueous phase, and independently isolated RNA from the TRIzol interphase upon UV-crosslinking to collect the OOPS RNAs. We generated total RNA-seq libraries that showed high correlations in gene-level counts within each developmental stage and assay type (**Figure S1A**). In a principal component analysis (PCA) the samples separated primarily by developmental stage (76% of the variance), with nPSCs being more distinct from pPSCs and dPSC. Assay type further contributed to separation (14% of the variance), with the OOPS samples being the most distinct (**Figure S1B**).

We next performed differential expression analysis to identify semi-extractable and OOPS-enriched genes compared to standard TRIzol RNA-seq controls at each developmental stage (**Figure S1C, Figure 1B**). By employing stringent effect-size and statistical cutoffs (at least two-fold enrichment and adjusted p-value<0.01), we identified 449, 1328 and 1390 high-confidence genes at each developmental stage with distinctly increased semi-extractability and elevated RBP occupancy in OOPS samples, henceforth collectively referred to as smOOPs (semi-extractable and OOPS-enriched RNAs) **(Figure 1B,C, Figure S1C, Table S1)**. Of the total 1828 unique smOOPs, 276 were common to all stages, with most occurring at later stages. Notably, there were fewer unique smOOPs identified in nPSCs (76) compared to the other stages (pPSCs, 304; dPSCs, 385, **Figure 1C**). A positive

correlation was observed between fold changes in the OOPS and semi-extractability assays, particularly upon the onset of cell fate commitment. However, the degree of enrichment (fold change) for a gene in one assay (OOPS or semi-extractability) did not always reflect the degree of enrichment in the other **(Figure 1B)**. Some smOOPs highly enriched in OOPS were not similarly enriched in semi-extractability **(Figure S1D)**, indicating that each assay preferentially captures distinct RNA characteristics, and that the smOOPs pool is rewired during developmental transitions **(Figure 1B,C, Figure S1D)**. In terms of gene biotype, smOOPs predominantly consist of protein-coding genes (74.6%-82.2%), a small fraction of lncRNAs (2.2%-4.1%) and TEC genes ("To be Experimentally Confirmed"; 14.5%-16.7%) **(Figure 1D).** Notably, the TEC proportion was more than double that of the remaining genes not classified as smOOPs (non-smOOPs) (6.3%). The overrepresentation of TECs in the smOOPs group suggests that these understudied transcripts might play previously unrecognised roles in RNA-protein interactions or in the formation of RNP complexes. In conclusion, our approach highlights smOOPs as a group of transcripts both bound by RBPs and with unique extractability properties. This provides an ontogenetic atlas of a class of condensation-prone RNAs that may have specialised roles in assembly of RNP complexes.

smOOPs: RNAs with high condensation propensity in RNP granules

RNA itself can serve as a scaffolding platform to concentrate bound RBPs driving the sorting and assembly of mRNPs networks (Chen et al., 2024; Ma et al., 2021; Schwartz et al., 2013; Tauber et al., 2020; Yamazaki et al., 2018). We validated the tendency of smOOPs for RNP condensation using global iCLIP, an orthogonal method for mapping the cumulative RBP occupancy across the transcriptome (Lee et al., 2021). This not only confirmed the elevated RBP binding compared to non-smOOPs, but also provided precise positional information on RBP interactions **(Figure 1E, Figure S1G)**. We hypothesised that smOOPs serve as hubs for concentrating RBPs and promoting the assembly of RNP condensates (Chen et al., 2024; Sanders et al., 2020; Yamazaki et al., 2018). To test this, we designed HCR-FISH probes (Choi et al., 2018) against the exons of 17 candidate protein-coding transcripts, including smOOPs and non-smOOPs **(Figure S1E,G)**, all with comparable expression levels. We postulated that condensation-prone smOOPs would appear as bright discrete foci similar to *NEAT1* in paraspeckles (Chujo et al., 2017; Clemson et al., 2009) or germ granule mRNAs known to form homotypic clusters (Trcek et al., 2015). Image analysis confirmed that smOOPs indeed formed larger foci with higher overall intensity compared to non-smOOPs **(Figure 1F,G)**. Notably, one of the imaged smOOPs, *Dync1h1*, has already been shown to form cytoplasmic foci of 3-7 copies at active translation sites (Pichon et al., 2016). Further analysis of RNA distribution found that only 36.9% of all smOOPs foci were nuclear, compared to 46.5% for control transcripts **(Figure S1F)**. These findings support the notion that smOOPs have unique biophysical properties, including higher RBP occupancy and a tendency to form homotypic clusters, characteristic of condensation-prone RNAs forming RNP granules.

Deep learning accurately predicts condensation-prone smOOPs

Cells exhibit a wide range of assemblies of RNA organisation. However, the molecular rules governing the RNA condensation and their local entrapment in RNP condensates (Chen and Mayr, 2022; Forrest and Gavis, 2003), has not been systematically explored. To address this challenge, we used an explainable deep learning approach to deconvolve in an unbiased manner the sequence and structural determinants, and trans-acting regulators predictive of RNA condensation propensity. For

our binary classification, smOOPs from nPSCs with processed transcript length under 20 kb were used as the positive class (n = 447), while genes that did not exhibit condensation-prone characteristics were defined as the control set. These genes were not differentially expressed in either OOPS and semi-extractability assay at any stage (|LFC| < 1.4 and padj > 0.01, see Methods) their intersection yielding a unified control set (n = 1232) that has similar expression levels to that of smOOPs **(Figure S2B)**. We focused on nPSCs due to the greater availability of public transcriptomic data compared to pPSCs, allowing us to utilise a more extensive set of features. To identify the features unique to the condensation prone transcripts we trained sequential deep learning classifiers based on multiple feature sets: RNA nucleotide sequence, global RBP occupancy (global iCLIP), m^6A modification sites (Modic et al., 2024), transcriptome-wide base pairing intra-molecular (PARIS-Intra) and inter-molecular interactions (PARIS-Inter) (Lu et al., 2016), *in silico* structure prediction with RNAfold (Lorenz et al., 2011) and RNA-binding sites of 46 RBPs determined via CLIP from the POSTAR3 database (Zhao et al., 2022). The global iCLIP, m^6A modification sites, PARIS-Intra/Inter data were first normalised to expression, after which all features were encoded as layers of information over each smOOPs and control transcript, preserving the positional importance and allowing for efficient feature extraction **(Figure 2A)**. To enable the model to learn the complex patterns and positional dependencies in the data, we implemented a deep learning (DL) architecture consisting of multiple convolutional neural network blocks (CNNs), a recurrent neural network (RNNs) and a multi-layer perceptron (MLPs) **(Figure 2A)**. The overarching goal of this deep learning approach was not merely prediction, but rather extraction and understanding the classifying information each dataset provides, both independently and in conjunction with others. Therefore, we devised a strategy of training an individual DL model on each subset in the powerset of the seven datasets, totalling to 127 unique models, each trained in 8 replicates (see Methods).

Given that smOOPs were significantly longer than control transcripts, particularly their coding sequence (CDS) **(Figure S2A)**, we initially trained a baseline model using sequence length as the sole feature, achieving a modest maximum accuracy of 74% and AUC score of 0.82 **(Figure 2B)**. The DL model trained on the sequence as a sole training feature already greatly improves the prediction accuracy of the smOOPs not seen by the model (AUROC = 0.91, accuracy = 81%), compared to the baseline length control model **(Figure 2B, S2C,D)**. As additional datasets were introduced, the predictive power gradually improved, ultimately reaching an AUROC of 0.94 and accuracy of 83% with all datasets included **(Figure 2B)**. To statistically evaluate the average improvement of the AUC score by individual information layers, we calculated the difference between the model performance trained on all subsets without the information layer in question and the model performance trained in addition to the layer. This approach highlighted the improvement in predictive power that each information layer contributed beyond any datasets already included. Our analysis revealed that the sequence, POSTAR3, global iCLIP, and PARIS-Intra layers significantly improved the performance of the models in which they were incorporated, giving us the confidence that these datasets contain a specific level of information about the features that distinguish smOOPs transcripts **(Figure 2C)**. Surprisingly, RNAfold predictions reduced performance, presumably by introducing noise due to the limited ability of energy minimization algorithms to predict secondary structures for long sequences and the absence of context-specific information (e.g. cellular factors like RNA-binding proteins (RBPs) or the influence of in vivo conditions that can affect folding). The AUC score improvement skews toward zero for most features **(Figure 2C)**, likely due to multicollinearity among the encoded features. To explore this overlap, we analysed feature pairs for the top four layers. For each pair, we excluded both features, then measured AUC after adding one and both, assessing their individual and combined

contributions. This revealed additive effects, particularly between RNA structure and total RBP occupancy. Notably, sequence alone encodes for most information represented by the other biological datasets, except for individual RBP binding profiles, which added further detail to the model, presumably by disentangling the sequence into distinct RBP binding partners enabling more efficient information extraction **(Figure 2D)**. Training models for the entire powerset of feature combinations allowed us to reveal nuanced interactions between datasets that would have been missed by analysing them either in conjunction or separately. This enables us to focus our model explainability efforts and elucidate the unique features models found to be predictive of the condensation-prone smOOPs.

Deconvolving the features of condensation-prone smOOPs

To identify the unique features distinguishing condensation-prone RNAs, we utilised integrated gradients to compute nucleotide-resolution attribution or importance scores for the models trained on each of the top four datasets that improved the average model performance (i.e. sequence, POSTAR3, global iCLIP, and PARIS-Intra) **(Figure 2C,D)**. By overlaying these importance scores over each transcript, we uncovered distinct patterns of feature importance for each dataset. To discern global feature patterns governing condensation, we divided each smOOPs transcript into 100 bins, transforming their absolute lengths into relative positions. For each feature, we averaged the importance scores within each bin, creating a 3D matrix of smOOPs transcripts, relative positions, and feature importance scores.

Since the DL models were able to identify several different features that contribute to classification, we flattened the matrix and reduced the dimensionality of this data using UMAP **(Figure 3A)**. The results showed two well-separated clusters of smOOPs, each of them highlighting different characteristics that were important for the model (cluster 1, n = 328, cluster 2, n = 119). Cluster 1 primarily contained mRNAs, while cluster 2 was dominated by TECs (54.6%) **(Figure 3A)**. Upon inspecting the average importance distribution for each cluster, distinct sequence compositions emerged. Cluster 1 exhibited a pronounced cytosine enrichment across the entire transcript, particularly within the 5'UTR and CDS, which overlapped with the PARIS-Intra importance **(Figure 3B)**. In contrast, cluster 2 displayed increased overall uridine importance, with adenine (A) being slightly more important at the 5' ends of the transcripts **(Figure 3C)**. Although global RBP occupancy (global iCLIP) is a key feature across all smOOPs, our analysis further highlighted the positional specificity of RBP binding with cluster 1 having the importance of RBP occupancy predominantly concentrated in the 3'UTR, as opposed to cluster 2, where it was distributed more uniformly across the entire transcript **(Figure 3B,C)**. Furthermore, the regions of POSTAR3 importance, summing individual CLIP binding profiles, portrayed more specific regions of importance, emphasising the precise nature of RBP interactions at these sites (Figure S3A,B). Cumulatively, our deep learning approach pinpoints the key biological information for RNA condensation propensity, with sequence, intramolecular structure and RBP binding information being most predictive.

Next, we looked at the actual transcripts to confirm the features deemed important by the model. Transcripts in cluster 1, predominantly mRNAs, are very C-rich and structured, particularly their CDS **(Figure 3B,D,H),** with the 'CCC' triplet being the most enriched, followed by 'GCC', 'CGC', 'CCG', and 'CCA' **(Figure S3C,E).** Global iCLIP and POSTAR3 data show enhanced RBP binding, especially in the 3'UTR region of the condensation-prone transcripts **(Figure 3F)**, with SRSF3, POU5F1, SRRM4,

AGO2, and FUS carrying key predictive information **(Figure 3I)** and are more frequently associated with smOOPs than control transcripts. Transcripts in cluster 2 were consistently highly enriched in uridine (U) across the entire length with a slight adenine (A) enrichment in the 5' end **(Figure 3C,E).** The most frequent triplets included 'UUU', followed by 'UUA', 'UCU', 'UGU', and 'UAU' **(Figure S3D)**. This cluster of transcripts also exhibited over a 10-fold greater median normalised RBP occupancy **(Figure 3G)** compared to control. While UV crosslinking preferentially targets uridines, we cannot determine the exact extent to which this bias influences the observed enrichment in RBP occupancy.

To validate the contribution of each feature across different transcript regions, we conducted masking experiments by sequentially omitting the 5'UTR, CDS, and 3'UTR signal **(Figure 3J)**. We observed that sequence information contributed to the predictive power the most in the CDS, whereas the global iCLIP and POSTAR3 were most informative in the 3'UTR. In contrast, the performance of models based on PARIS-Intra diminished when either the CDS or 3'UTR was masked.

These findings highlight two distinct clusters of condensation-prone RNAs in nPSCs: one comprising C-rich mRNAs and the other predominantly U-rich transcripts. In both clusters RBP binding plays a crucial role as a unique feature of RNA condensation, which may explain why PARIS inter-molecular interactions, formed by base-pairing, were not especially predictive - RBPs appear to be the key drivers of this process. This underscores the importance of understanding RBP involvement in RNP condensation, emphasising their role in shaping the behaviour of these unique transcripts.

smOOPs form distinct interconnected networks of RNA-RNA interactions

Given the well-established role of RBPs in regulating RNA assembly, we hypothesised that the condensation-prone smOOPs transcripts participate in broader RNA networks, facilitated by RBP-mediated contacts across the transcriptome. By identifying RNA-RNA interactions (RRI) involving smOOPs, we can explore potential mechanisms of their regulation or assembly at the RNA level. To globally map RBP-associated intra- and inter-molecular RNA-RNA contact sites, we performed RNA *in situ* conformation sequencing (RIC-seq (Cai et al., 2020)) **(Figure 4A)**. We focused on nPSCs and pPSCs, as the widespread changes across scales between these stages make them ideal for studying cell-type-specific RNA organisation (Yang et al., 2019). We detected 758,135 hybrid reads in nPSCs and 1,245,548 in pPSCs (excluding rRNA, tRNA, and mitochondrial reads), with 28% being inter-molecular in nPSCs and 44% in pPSCs **(Figure 4B)**. Gene-level inter-molecular hybrid frequency was highly reproducible across samples **(Figure S4A)**, with PCA showing developmental stage explained 88% of the variance **(Figure S4B)**. The global distribution of contacts remained highly consistent across stages, with most involving intronic regions (~66%, **Figure S4C**), aligning with previous observations (Cai et al., 2020).

We generated developmental stage-specific RNA-RNA interaction (RRI) networks using inter-molecular hybrid reads from RIC-seq, with genes as nodes and edges representing the frequency of interactions between them. The inferred networks showed typical characteristics of biological networks, such as protein-protein interaction (PPI) and RRI networks (Ju et al., 2018). Specifically, the networks displayed scale-free-like behaviour, where most nodes have few connections, but a small number of highly connected hubs dominate **(Figure S4D)**. This means that a few genes play central roles in the network, while most genes have fewer connections, creating a heavy-tailed distribution of connectivity **(Figure S4D)**. The RIC-seq networks also display small-world

properties, with a higher global clustering coefficient than random networks of the same size **(Figure S4E)** and a relatively short average path length of 3.5 **(Figure S4F)**. These features suggest structured RNA interaction groupings, potentially enabling efficient communication and modular organisation within RRI networks, rather than random connections.

Thus, we investigated whether the distinct properties of smOOPs are linked to their distinct higher-order organisation within these RRI networks. To explore this, we systematically analysed the behaviour of smOOPs in the derived RIC-seq networks. In the nPSCs network, 430 of 449 smOOPs were present, and in pPSCs, 1255 out of 1328 smOOPs. Across the RIC-seq networks, we observed that for genes of similar length, the degree of connectivity correlates strongly with expression, regardless of whether they are smOOPs or not **(Figure 4C)**. This is likely because RIC-seq is more sensitive to detecting interactions in abundant RNAs. Consequently, on a global scale, smOOPs do not show an unusually high number of connections with other transcripts.

However, a more detailed analysis of smOOPs RRIs and their properties is necessary to determine whether their interactions are specific and potentially arise from co-regulation or co-localisation. To explore potential interaction patterns among smOOPs, we examined the RIC-seq network focusing only on smOOPs-smOOPs interactions **(Figure 4D)**. The nPSCs and pPSCs smOOPs subnetworks appeared highly connected, as shown by several key metrics: edge density (the proportion of possible edges present), largest connected component (LCC, the size of the largest connected subnetwork), global clustering coefficient (GCC, the tendency of nodes to form tightly connected groups), and average degree (the average number of connections per node within the subnetwork, 4.1 in nPSCs and 21.3 in pPSCs) **(Figure 4D,E)**. To determine if the smOOPs connectivity is greater than expected by chance, we compared it to null network models generated using degree-matched sub-sampling (see Methods), which accounts for potential biases linked to degree distribution. Given the high degree of smOOPs **(Figure 4C)**, this ensured that the observed difference in connectivity reflects biological signals, not merely the tendency of high-degree nodes to form more connections. Compared to the degree-matched sampled networks, smOOPs are significantly more connected across all metrics in nPSCs and pPSCs **(Figure 4E)**. This indicates that smOOPs possess distinct local connectivity patterns that are not just a byproduct of their degree. Moreover, even though the identity of smOOPs varies across developmental stages **(Figure 1C)**, this organisational principle remains conserved across stages, persisting in both nPSCs and pPSCs **(Figure 4E)**. Together, these findings suggest that the higher connectivity of smOOPs is not an artefact of their degree but reflects a unique and specific network organisation that points to the possibility of coordinated regulation among smOOPs.

smOOPs mRNAs are translated into functionally related proteins

We reasoned that the distinct RNA signatures of smOOPs **(Figure 3)** and their interconnectivity **(Figure 4D,E)** could indicate co-regulation or co-assembly, which we further investigated in nPSCs. First, we compared RNA stability (Herzog et al., 2017) and translation efficiency (Ingolia et al., 2011) of the nPSCs smOOPs to those of control mRNAs both of which have similar expression levels **(Figure S2B)**, and observed that smOOPs mRNAs in cluster 1 exhibit comparable mRNA stability, but slightly increased translational efficiency **(Figure S4,G,H,I)**.

This prompted us to explore whether the distinct sequence features of cluster 1 smOOPs CDS could influence amino acid composition. For this we compared amino acid usage between smOOPs and control mRNAs, identifying a distinct bias towards increased proline and serine usage in smOOPs (1.9% and 1.0%, respectively, **Figure 4F**), with these residues showing the most pronounced differences even among other C-rich amino acids. Proline and serine are common building blocks of intrinsically disordered regions (IDRs), which are frequently found in proteins involved in essential cellular processes, including transcription, cell signalling, and subcellular organisation (Holehouse and Kragelund, 2024; Uversky et al., 2015). We examined the presence of IDRs in proteins encoded by smOOPs mRNAs compared to control mRNAs and found that cluster 1 smOOPs encode a higher proportion of IDR-containing proteins and have a higher overall disorder percentage **(Figure 4G,H)**. Since IDRs facilitate phase separation and biomolecular condensate formation, we applied the PICNIC tool (Proteins Involved in CoNdensates In Cells, (Hadarovich et al., 2023)) and confirmed that proteins encoded by smOOPs are more likely to be involved in condensates **(Figure 4I)**.

Taken together, our data suggests that condensation-prone smOOPs encode proteins with a propensity for phase separation and form a dense subnetwork of RRIs, pointing to a potential coordination of RNA and protein phase-separating processes in nPSCs.
Cells contain a wide array of RNA assemblies, the composition and biological roles of which remain largely uncharacterized

Discussion

Cells contain a wide array of RNA assemblies, the composition and biological roles of which remain largely unknown. Over three decades ago, an electron microscopic study identified a variety of branched reticular protein-rich structures which retained over 70% of nuclear RNA (He et al., 1990). Only recently it has become clear that the cytoplasm is similarly partitioned into extensive meshworks of RNA-binding proteins (RBPs) scaffolded on mRNAs (Chen et al., 2024; Masuda et al., 2024; Ma and Mayr, 2018). Despite this progress, the diversity of RNA organisation - from single mRNAs to mesoscopic condensates - remains largely unexplored (Cardona et al., 2023). Understanding the principles governing higher-order RNA assemblies in their native cellular context is crucial to uncovering how these structures shape cellular biology.

Here we provide the first systematic investigation of condensation-prone RNAs and their properties in early murine development. We identified smOOPs, a distinct class of long transcripts that consists of predominantly mRNAs that are both semi-extractable and RBP-associated. Notably, some of these transcripts exhibit subcellular localisation in larger foci, suggesting a potential role in spatial organisation and cellular compartmentalisation (**Figure 1**).

To decipher the properties driving the distinct biophysical behaviour of smOOPs, we implemented an integrative deep-learning approach that systematically extracted the key RNA signatures contributing to their condensation propensity (Figure 1, Figure 2). Notably, smOOPs are not a homogenous group; based on their features, they can be divided into two distinct clusters, with cluster 1 consisting almost exclusively of mRNAs. Cluster 1 smOOPs are notably cytosine-enriched, more structured, and have 3'UTRs extensively bound by RBPs. These shared RNA features co-occur with an increased coding

potential for IDRs in the proteins they encode, highlighting a principle where RNAs with common characteristics give rise to proteins with similar biophysical properties, such as phase-separation propensity. This suggests a coordinated role for these smOOPs in phase-separation-driven cellular processes. Our analysis also revealed that smOOPs form densely connected RRI networks, suggesting an additional layer of RNA-based coordination that could reflect co-regulation or co-assembly. Notably, this non-random association may indicate underlying biological interactions or shared functional roles within RNA networks. More specifically, this interconnectivity could represent a sorting mechanism to organise mRNA information and enable identity-dependent co-regulation of transcripts.

Despite their shared properties, it is likely that smOOPs comprise distinct subgroups, possibly manifesting their condensation potential at different cellular sites. Further work is needed to precisely map the cellular localisation of smOOPs, determine at which stage of the mRNA lifecycle their co-regulation or co-assembly occurs, and explore the functionalities of their higher-order RNA organisational domains.

More broadly, in terms of methodology, our study demonstrates the power of machine learning to not only make predictions but also reveal complex patterns across diverse datasets, enabling unbiased classification and characterisation of heterogeneous gene groups. This is especially valuable in the study of condensates, where distinguishing different assemblies and RNA functions is challenging using traditional approaches. By coupling network-based analyses with RNA-proximity ligation datasets, we provide a methodology to uncover new insights into the organisational principles of RNA networks involving specific genes, which lays the foundation for advancing our understanding of their role in post-transcriptional regulation.

Our findings suggest a sorting mechanism for organising mRNA information, revealing higher-order, identity-dependent coordination of transcripts. Our integrated approach paves the way for exploring the complexity and specificity of transcriptome organisation and the principles governing RNA assemblies in post-transcriptional gene regulation.

Methods

Cell culture

Low-passage, wildtype mouse pluripotent stem cells IDG3.2 PSCs (129S8/B6 background, (Hitz et al., 2007) and TGFP/+;Foxa2tagRFP/+ (G9) (Scheibner et al., 2021) were cultured feeder-free, on cell culture dishes (TPP) coated with 0.1% Gelatin (ES-006-B, Millipore). Naïve PSCs were maintained in N2B27 medium composed of 1:1 Neurobasal (21103049) and DMEM-F12 (11320074) medium containing N2 (17502001) and B27 (17504001) supplements, 1% Glutamax (35050061), 1% nonessential amino acids (11140050) and 0.1 mM 2-mercaptoethanol (31350-010) (all Thermo Fisher Scientific), 12 ng/mL LIF (104278, Qkine), with additional use of small molecule inhibitors: for a condition commonly named 2iLIF, 1 mM MEK inhibitor PD0325901 (1408, Axon Medchem) and 3 mM GSK3 inhibitor CHIR99021 (SML1046, Sigma). nPSCs were fed every day and split every 2-3 days using Accutase (A6964, Sigma). To transit from naïve to primed pluripotency state, the G9 PSCs were

accutased and seeded onto gelatine and FBS-coated (EmbryoMax® ES Cell Qualified FBS, ES-009-B, Merck) plates in N2B27 medium supplemented with 1,000 U/ml LIF (ESGRO ESG1107, Merck), 12 ng/mL bFGF (100-18B, Peprotech), 20 ng/mL Activin A (338-AC-050, R&D Systems), and 2 µM IWP2 (3533, Tocris). Once the nPSCs reached primed pluripotency state - pPSCs - they can be maintained indefinitely using the abovementioned FAI medium. For further expansion, pPSCs G9 cells were maintained in FAI medium on cell culture dishes (TPP) coated with Geltrex (A1413302, Gibco), fed every day and split every 2-3 days using Accutase (A6964, Sigma).
To generate differentiated primitive streak progenitors (dPSCs) pPSCs were grown in Wnt3a-differentiation medium (N2B27 medium supplemented with 20 ng/mL ActA, 12 ng/mL bFGF and 250 ng/mL Wnt3a) for 24 hours prior collecting the cells.

Semi-extractability and OOPS assays

pPSC G9 (p9) were prepared for experiment one day prior harvesting by plating them on 10 cm plates (TPP), and Wnt3a-differentiation medium was added to pPSCs for 24 hours to generate differentiated primitive streak progenitors (dPSCs). For this experiment only, 5% FBS (ES-009-B, Merck) was added to gelatin during coating for nPSCs (p20). Cells were prepared as follows: for semi-extractability and standard TRIzol RNA-seq controls one confluent 6-well was used per replicate. After washing cells 2x with PBS, 1 mL TRIzol-LS was added to the plate and scraped, then transferred to a 1.5 mL Eppendorf tube. Samples were snap-frozen and stored at -80°C. For OOPS, one confluent 10 cm plate was used per replicate, washed twice with ice-cold PBS before crosslinking them at 400 mJ/cm^2 (Crosslinker CL-3000 at 254 nm (AnalytikJena). After removing residual PBS, 1 mL TRIzol-LS was added to the plate before scraping off the cells content and transferring it to a 1.5 mL Eppendorf tube that was stored at -80°C.

The semi-extractability assay was performed according to the protocol from (Chujo et al., 2017): upon thawing and diluting TRIzol LS accordingly, the sample was heated for 10 minutes at 55°C and then sheared 40x through a 20G needle (the heating and shearing steps were omitted for control samples). Next, chloroform was added and mixed vigorously before the tube was centrifuged at 12,000x g for 15 minutes at 4°C. The aqueous phase was then mixed with 1.5x volume of 100% ethanol, and the isolation was continued with the RNeasy Plus Mini Kit (Qiagen) according to manufacturer's instructions.

For OOPS, we followed the protocol kindly provided by Dr. Eneko Villanueva (Villanueva et al., 2020). Briefly, the content of TRIzol LS tubes was first diluted with 1x volume nuclease-free water, then we added 1/5th volume of chloroform, mixed vigorously and centrifuged at 12,000x g for 15 minutes at 4°C. The aqueous and organic phase were carefully removed and the remaining interphase resuspended in fresh 1 mL TRIzol. Again, 200 µL chloroform was added before centrifugation and removal of aqueous and organic phases. Finally, the interphase was resuspended in 1 mL TRIzol once more before adding chloroform and separating the phases by centrifugation (12,000x g, 15 minutes, 4°C). This time, we removed as much of the aqueous and organic phases as possible and precipitated the remaining interphase (100 µL) using 100% Methanol (1 mL). After vortexing, the sample was centrifuged (14,000x g for 10 minutes at room temperature), supernatant was removed and the pellet digested using proteinase K (sro-3115828001, Roche) for two hours at 50°C. RNA was finally extracted by addition of 1 volume of phenol:chloroform (P3803, Sigma) and vortexing before centrifugation at 10,000x g for 10 minutes at room temperature. The aqueous phase was collected

and RNA was precipitated for 20 minutes at room temperature with isopropanol. After centrifugation (12,000xg, 10 min, 4°C) the RNA pellet was washed twice using 75% ethanol and finally resuspended in nuclease-free water (AM9930, Thermo Fisher). All RNA samples were treated with Turbo DNase (AM2238, Thermo Fisher) for 30 minutes at 37°C to remove DNA contaminants and bead-purified using Agencourt AMPure XP (A63881, Beckman Coulter). Sequencing libraries were prepared using CORALL Total RNA-Seq with RiboCop rRNA depletion (Lexogen) starting with 500 ng DNase-treated RNA as input. The resulting cDNA libraries were sequenced as single-end 100 bp reads on HiSeq at the Advanced Sequencing Facility at The Francis Crick Institute.

Global iCLIP

iCLIP was performed according to the iiCLIP protocol (Lee et al., 2021). First, 10 cm plates of nPSCs (G9, passage 17) and pPSCs (G9, passage 20) containing ~1.5 mg of total protein mass per sample were lysed. Three replicates were prepared for each developmental stage. Briefly, cells were lysed upon UV-crosslinking (150 mJ/cm^2 in Crosslinker CL-3000 at 254 nm (AnalytikJena)) and treated with RNase I before binding the RNA to an IR dye-labelled adaptor and separating the RNA-protein complexes on SDS-PAGE. After cutting the desired part of the membrane (from 40 kDa upwards), we digested the proteins and extracted the protein-bound RNA to prepare libraries for next-generation-sequencing. Libraries were sequenced as paired-end 150 bp reads on NovaSeq at Clinical Institute of Special Laboratory Diagnostics, University Children's Hospital at Ljubljana University Medical Centre.

RIC-seq

Experiments were performed using nPSCs (G9, passage 20) and pPSCs (G9, passage 13)) following the protocol by (Cao et al., 2021)) with some modifications. Two 10 cm confluent plates per sample were crosslinked using freshly prepared 0.5 mg/mL DSS (21655, Thermo Fisher Scientific) at room temperature for 30 minutes (rotating at 20 rpm). The reaction was quenched using 20 mM Tris (pH7.5) for 15 minutes at room temperature (rotating at 20 rpm). All centrifugation steps were carried out at 3500 rpm for 5 minutes at 4°C. After pelleting the cells, they were washed with PBS and then permeabilized (10 mM Tris-HCl (pH 7.5), 10 mM NaCl, 0.5 % Igepal, 0.3 % Triton X-100 and 0.1% Tween-20) for 15 minutes on ice. Cells were washed three times with 1X PNK buffer before fragmenting the RNA using 6U MNase (EN0181, ThermoFisher) at 37°C for 10 minutes. The following steps including FastAP treatment (EF0651, Thermo Fisher), pCp-biotin ligation (20160, Thermo Fisher), second FastAP and PNK treatment (M0201L, NEB) were performed as in the original protocol. Proximity ligation was performed using a different T4 RNA ligase (M0204L, NEB) and hence 1 mM ATP was added to the overnight reaction. Finally, cells were lysed using 200 µL proteinase K buffer (10 mM Tris-HCl pH 7.4, 100 mM NaCl, 1mM EDTA, 0,2% SDS) and 50 µL proteinase K (sro-3115828001, Roche) and an addition of 1.5 µLTurbo DNase (AM2238, Thermo Fisher), incubating at 37°C for 30 minutes, then 50°C for 60 minutes before adding Trizol-LS and snap freezing the samples. Once thawed and brought to room temperature, the samples were heated for 10 minutes at 55°C before adding the chloroform and precipitating RNA from the aqueous phase with isopropanol. The RNA was treated with Turbo DNAse once more and cleaned up using phenol:chloroform extraction as above. 21 µg of RNA was fragmented using 5x First Strand Synthesis Buffer (SuperScript IV, 18090050, Invitrogen) for 3.5 minutes at 94°C, immediately placed on ice and mixed with the MyOne Streptavidin C1 beads (65001, Invitrogen) to pull down biotinylated RNA (30

minutes at room temperature). Eluted RNA (10 µL) was extracted using phenol:chloroform method (see above). The biotinylation and pulldown efficiency were confirmed using dot blot assay before preparing sequencing libraries.

The biotin-enriched eluate was next subjected to 3'end dephosphorylation using PNK (M0201L, NEB) and FastAP (EF0654, Thermo Fisher) and purified using Agencourt AMPure XP beads (A63881, Beckman Coulter) before 3'end adapter ligation overnight at 20°C. Once again, RNA cleanup was done with Agencourt AMPure XP beads before adapter removal (using Deadenylase (M0331, NEB) and RecJf exonuclease (M0264S, NEB)). Reverse transcription was performed according to the Superscript IV RT kit (18090050, Invitrogen) manual with the use of a custom RT primer. Following cDNA cleanup (Agencourt AMPure XP beads) 5' cDNA adapter was ligated using T4 DNA ligase (EL001, Thermo Fisher, without ATP). Samples were loaded onto 6% TBE-Urea gel (EC6865BOX, Thermo Fisher) and cDNAs exceeding 200 nt were excised from the gel and extracted using Crush-Soak Gel buffer (as per iiCLIP protocol, (Lee et al., 2021)), followed by phenol:chloroform extraction. Precipitated cDNAs were stored at -20°C before performing PCR using P5/P7 standard Illumina primers and Phusion HF master mix (NE-M0531L, NEB). Ribosomal RNA contaminants were removed from the final library using Ribocutter which utilises Cas9-guided rRNA depletion (Wilkins and Ule, 2021); for this, 275 gRNAs were designed against mature 5S, 18S and 28S rRNAs (obtained as 50 pmol oPool from IDT). Final libraries (~10 nM) were treated with 4 µM sgRNAs for 30 minutes at 37°C and after beads purification, the libraries were reamplified with additional 6 cycles. They were sequenced as paired-end 150 bp reads on NovaSeq at Clinical Institute of Special Laboratory Diagnostics, University Children's Hospital at Ljubljana University Medical Centre.

HCR-FISH

Hybridization-chain reaction FISH was prepared by following the protocol from (Choi et al., 2018)) with slight modifications. The probes (8 pairs of probes per target) were designed using http://hcr.wustl.edu/input/quick/ against mature mRNAs, however not targeting exon-exon junctions (average expression for controls 23.6 +/- 16 TPM, for smOOPs 37.9 TPM +/- 31 TPM in semi-extractability total RNA-seq library). For imaging purposes WT IDG3.2 nPSCs were plated on Geltrex (A1413302, Gibco) coated 8-well glass-bottom ibidi plate (80827, ibidi) one day prior to fixation. After washing the cells with PBS, the cells were fixed by the fixation mixture (4% formaldehyde, 0.4% glyoxal, 0.1% methanol, 1x PBS). Amplification stage of a protocol to generate a tethered fluorescent amplification polymer lasted 10 hours. Cells were then washed as described in the protocol and finally mounted in 300 µl Fluoromount G (00-4958-02, Thermo Scientific). Images were acquired with a custom-built STED microscope (Abberior instruments) using a 60× water immersion objective and lasers running at 80 MHz repetition rate. We excited fluorescently labelled mRNA by one of the three lasers at either 488, 561 or 640 nm, with 120 ps pulse length, with maximal power of 116 µW, 111 µW and 200 µW in the sample plane and DAPI stained nuclei with 405 nm laser with maximal power 810 mW in the sample plane. The laser powers used were 10% for 405 laser and 30% for the other three lasers. We acquired the fluorescence intensity using an avalanche photodiode with 500-550 nm, 580-625 nm or 655-720 nm filters (Semrock) in front. The combinations of lasers and detectors were as follows: 405 nm laser and 500-550 nm, 488 nm laser and 500-550 nm, 561 nm laser and 680-620 nm filter, and 640 nm laser and 655-720 nm filter. The dwell time in the pixel was 10 µs, the pixel size was set to 50 nm and the pinhole size was set to 1.07 AU to achieve a good confocal resolution.

HCR-FISH image analysis

The image analysis was done in Fiji (Schindelin et al., 2012) using custom-made macros. Briefly, for each fluorophore background was subtracted (30.0 pixels rolling radius) and threshold was set accordingly to conform all different transcripts with the same fluorophore ((9,255) for channel for 561, (3,255) for 488 channel and (33,255) for 640 channel). Next, particle analysis was performed and all subsequent analysis was done using Total intensity was calculated by multiplying mean intensity and area for each foci. The average of total intensity for the control transcripts in each fluorophore was used to normalise other values and obtain “Total on average”.

Reference annotation

For all analyses we used the GRCm39 build of the mouse genome with the Gencode M27 annotation. We used a custom reference sequence built on this annotation for the alignment of hybrid reads, generated as previously described (Chakrabarti et al., 2023b). To unambiguously annotate the genes within hybrid reads, we used a flattened annotation produced by iCount-Mini (https://github.com/ulelab/icount-mini). Both are available for download with the Tosca pipeline (Chakrabarti et al., 2023b) (https://github.com/amchakra/tosca).

RNA-seq data analysis

The sequencing reads were processed using nf-core/rnaseq version 3.4 (Bushnell et al., 2017; Ewels et al., 2020) (https://nf-co.re/rnaseq/3.4). For differential expression analyses we used gene-summarised count tables generated by nf-core/rnaseq 3.4 (using Salmon) as input to DESeq2 version 1.44.0 (Love et al., 2014). The design incorporated both stage (nPSCs, pPSCs, dPSCs) and assay (control, semi-extractability, OOPS) factors. Prior to running DESeq2, we pre-filtered the count matrix to retain only genes with at least 10 normalised counts in at least 6 samples. For each desired contrast, we extracted results using the Wald test and applied ashr Log2FoldChange (LFC) shrinkage (Stephens, 2017). To identify genes enriched in OOPS or semi-extractability compared to control, we selected genes with padj < 0.01 and LFC > 1 for each contrast (semi-extractability vs control or OOPS vs control at each stage). Genes passing these thresholds in both semi-extractability and OOPS were defined as smOOPs. For data visualisation across assays, we converted rlog-normalised count data into Z-scores for each stage and plotted it using the ComplexHeatmap package version 2.20.0.

Global iCLIP data analysis

Sequencing reads were first demultiplexed using Ultraplex version 1.2.9 (https://github.com/ulelab/ultraplex) and then processed with the nf-core/clipseq version 1.0.0 (West et al., 2023) (https://nf-co.re/clipseq/1.0.0). We used BED files with crosslink positions and scores for all analyses. To normalise crosslinks by expression and correct for length for each transcript, we calculated crosslink density (CPM per kb) for exons and divided this by expression (semi-extractability TPM values obtained with Salmon).

RIC-seq data processing

We trimmed sequencing adapters using Cutadapt (Martin, 2011), then paired reads were merged with BBMerge (Bushnell et al., 2017). The merged FASTQ files were used as input for Tosca v1.0.0 (Chakrabarti et al., 2023b) to identify and analyse hybrid reads formed through RNA proximity ligation. To normalise RIC-seq gene-level counts to expression, we used TPM values calculated from RIC-seq nonhybrid reads summarised at the gene-level including all features (exons and introns) with featureCounts(Liao et al., 2013).

RIC-seq network inference and analysis

For each stage, we pooled the RIC-seq deduplicated hybrids files produced by Tosca, filtered out rRNA, tRNA and mitochondrial RNA containing hybrids, and retained the hybrid reads mapping to two different genes (representing inter-molecular interactions). We then summed the hybrid read counts corresponding to each gene pair to obtain gene-to-gene interaction frequencies. For each stage, we built a graph object, where genes were represented as nodes and edges were defined by the connections and their frequencies from the RIC-seq data, and analysed it using igraph (Csárdi et al., 2024; Csárdi and Nepusz, 2006) version 2.0.3 in R 4.4.0. To assess the connectivity of the smOOPs subgraph, we compared the observed subgraph with control networks generated by degree-matched sub-sampling. For the degree-matched sampling method, we sub-sampled random subnetworks from the full RIC-seq network maintaining the same number of nodes as the number of smOOPs identified in the RIC-seq data at each stage and a similar degree distribution (allowing for +/-1 degree variation). We performed 10,000 iterations to build a null distribution for comparison. Connectivity metrics were calculated for each random subset and compared to those of the observed smOOPs subnetworks for nPSCs and pPSC separately. While we also considered degree-preserving randomization methods, they were deemed less suitable for this analysis due to the network topology and the high degrees of the smOOPs, which could bias the results by preserving inherent connectivity patterns, especially among high-degree nodes (Horvát and Modes, 2021). We also considered random sub-sampling, but found it overestimated smOOPs interconnectivity due to their high degree relative to most other nodes. In contrast, degree-matched sub-sampling, which directly compares smOOPs to randomly selected groups of nodes with similar degrees, provided a more appropriate baseline for comparison, and was used as control.

PARIS data processing

We first trimmed sequencing adapters using Cutadapt and collapsed PCR duplicates with the readCollapse.pl script provided with the original publication (Lu et al., 2016) (https://github.com/qczhang/icSHAPE). Subsequent processing was performed using Tosca v1.0.0 (Chakrabarti et al., 2023b).

Data preparation for deep learning

To define genes that do not exhibit the condensation prones features in all stages, we selected those with padj > 0.01 and |LFC| < 1.4 in each stage and then selected their intersection as the unified control set. To create a reference transcriptome for the smOOPs and control genes, we selected the most highly expressed transcripts, based on semi-extractability, for each of the three stages. For

transcripts that varied across the stages, we selected the longest isoform among the most expressed transcripts at each stage.

To facilitate the training of the model on the selected datasets mapped across each transcript, we first encoded the data into a format suitable for deep learning. For both smOOPs and control transcripts, we extracted the genomic location of each exon from the reference genome. The RNA sequence for each exon was obtained based on their genomic coordinates using bedtools (Quinlan and Hall, 2010), and served as the raw sequence input for the model.

Several nucleotide-resolution transcriptomic datasets were extracted and aligned to the exon genomic locations: global iCLIP cross-links (this work); POSTAR3 database of CLIP binding peaks for 46 RBPs (Zhao et al., 2022) in mouse cell lines; Psoralen Analysis of RNA Interactions and Structures (PARIS) (Lu et al., 2016); RNAfold (Hofacker et al., 1994) predicting secondary RNA structures; and m6A methylation profiles (Modic et al., 2024). To precisely assign scores to exon positions, we generated a list of values corresponding to the length of each exon, initialised to zero. The experimental data were mapped onto the exon's genomic coordinates relative to its start position. Overlaps between experimental signals and exonic regions were located, and their scores were added to the list at the corresponding positions. The result was a list capturing experimental values at each relative position within the exon. For the POSTAR3 peaks, each RBP peak was mapped binary (only the presence of a peak or the lack of it was noted). Once scores were assigned, exons were concatenated back into transcripts, producing a list of feature scores across the entire transcript length for each feature. Additionally, RNAfold was used to predict secondary RNA structures in dot-bracket notation, classifying each position as either single- or double-stranded, which was subsequently encoded into binary form.

The datasets were randomly split into distinct training, validation, and testing subsets in a 70:15:15 ratio. Stratified sampling was employed to maintain consistent class proportions across subsets, and oversampling was applied to the minority group within each subset to balance class sizes. Each subset was stored independently and used directly for training, with dynamic data fetching and encoding implemented during the training process.

Data encoding

During training, batches of data were procedurally encoded by iterating over the training dataset and processing the mapped information to a DL compatible array. For each batch, the transcript sequence information was first converted from its native nucleotide string format into a numerical representation. To ensure transparency and interpretability, we employed one-hot encoding for the sequence data, with each nucleotide represented as a binary vector: Adenine (A) as [1, 0, 0, 0], Cytosine (C) as [0, 1, 0, 0], Guanine (G) as [0, 0, 1, 0], and Thymine (T) as [0, 0, 0, 1]. Non-sequence features, such as global iCLIP crosslink scores, were encoded as one-dimensional arrays of floats. All encoded data layers, including the nucleotide sequences and other transcriptomic features, were stacked vertically for each transcript. To then standardise input lengths for neural network processing, sequences were padded with -1 to match the length of the globally longest transcript. The output class was similarly encoded using a one-hot encoding scheme, where class membership was represented by a binary vector. Such encoded transcripts were then stacked to a new dimension representing the batch and as such processed by the training function.

Deep learning model architecture

The model architecture employs a previously optimised design based on (Modic et al., 2024), combining multiple convolutional, recurrent, and fully connected layers (MLP) to effectively classify sequential data. This hybrid approach leverages the strengths of each layer type to capture diverse patterns in the input data, progressively transforming the sequences into representations suited for binary classification. Initially, we wanted to enable the architecture of the model to have the ability to achieve optimal classification performance, therefore a hyperparameter optimization process was conducted - defining the models architecture. Due to the complexity and size of the hyperparameter search space, a manual approach was impractical. Instead, a systematic search via Bayesian optimization, using the Optuna framework, was implemented, exploring batch sizes (8, 16, 32, 64), number of convolutional blocks (2 to 8), final number of CNN units (16 to 1024), CNN unit increase percentages (0.0 to 0.5), kernel sizes (3, 5, 7, 9), kernel size increases (1 to 4), dilation increases (1 to 4), dropout probabilities (0.0 to 0.4), L1-L2 regularisation values (0 to 0.01), max pooling size (2, 3, 4), GRU units (16 to 1024), Dense units (16 to 1024), learning rates (1e-5 to 1e-3), and normalisation methods (None, BatchNormalization, LayerNormalization). The optimisation was performed over 100 trials to identify the hyperparameter set that maximised AUC on the validation set. The best-performing configuration was selected for final model training. The optimised model architecture starts with four convolutional blocks. Each block consists of a 1D convolutional layer with a progressively increasing number of filters of 108, 144, 192, and 256 respectively, and kernel size of 9. Dilation rates exponentially increase with each block, with the first block using a dilation of 1 and later blocks using 4, 16, and 64, respectively. Each convolutional layer is followed by layer normalisation, a ReLU activation function, and a Dropout layer with a rate of 0.34 to mitigate overfitting. Max-pooling (pool size of 4) is used after each convolutional block to reduce spatial dimensions and preserve key features. After the convolutional layers, a bidirectional GRU layer with 128 units per direction is employed to capture sequence dependencies, again followed by layer normalisation and dropout.
Following the GRU, two fully connected (Dense) layers are included, with 64 and 32 units, respectively, followed by layer normalisation and dropout. The final output layer consists of two units, with a softmax activation to predict binary classes. The Convolutional, GRU and Dense layers were configured with L1-L2 regularisation (λ = 1.2e-5) to further mitigate overfitting in the learning process. The model contains between 1,149,986 and 1,203,446 trainable parameters, depending on the input shape. It was trained using the Adam optimizer, with a learning rate optimised to 1e-4, and categorical cross-entropy as the loss function. The model was trained for unlimited epochs, employing early stopping with 20 epoch patience, based on validation AUC, to stop training and return best weights when overfitting.

The training of the powerset

To evaluate the predictive power of individual features and their combinations, we trained a separate model for each of the 127 subsets in the powerset of the 7 available datasets. For each subset, only the selected features were encoded as input to the model, ensuring that the model was trained specifically with that particular feature combination. Each model was trained in 8 replicates to account for variability and to ensure optimal performance for every subset. The trained models were saved after completion to enable further analysis of feature importance and interactions.

Assessing model performance

The final performance of each of the 1016 trained models was established by predicting the combined validation and testing dataset, both of which were not previously seen by the model, ensuring robust validation and mitigating the optimisation overfitting. Of the 8 replicates trained on each subset, the best performing model, based on the AUC score, was selected, ensuring the robustness of the predictive power assessment. To identify the datasets that significantly improved the predictive power of a model trained on any given subset of datasets, the difference in AUC score was calculated for all possible pairs of models where they were both trained on the same subset of datasets with the only difference being an additional inclusion of the unique dataset. This was further repeated for all possible pairs for each of the unique datasets and the differences were evaluated using a one-sample t-test was used to evaluate if the mean of differences is statistically different from zero, Using this technique, we could identify the datasets containing biological information that significantly improved the predictive capability of a model, irrespective of the number of other information layers included. To further assess the informational overlap between features, we conducted pairwise analyses of feature combinations. For each pair of features, we calculated AUC scores for each subset of each item in their powerset, first by adding one feature from the pair, and then by adding both features. This approach allowed us to quantify the individual and combined contributions of each feature to model performance. To evaluate the extent of overlap, we calculated the difference between the maximum AUC achieved when adding either feature individually and the AUC obtained when both features were included. These differences were calculated across all subsets of the remaining five features, resulting in a comprehensive set of values for each feature pair. Averaging these differences provided insight into the distinct contributions of each feature combination with a single sample t-test comparing the mean to zero.

Explaining the trained weights

To evaluate the contribution of individual positions and feature components to the model's predictions, we employed the integrated gradients (IG) method. The technique quantifies the importance of each nucleotide in the sequence by computing the gradient of the prediction output with respect to the input features, integrated along a path from a baseline of zeros to the actual input sequence. By summing these gradients across the entire path, IGs assign attribution/importance scores to each nucleotide, indicating its contribution to the prediction. We calculated the IGs for each of the models trained on individual datasets of sequence, global iCLIP, POSTAR3 peaks and PARIS-Intra. This approach enabled us to identify the regions within each transcript that had the greatest impact on the model's classification decision, providing insights into which specific sequence elements were driving the condensation-prone smOOPs RNA predictions compared to controls.

Gaining insight into the learned features

To gain insight into the global features the model has learned, not limited to a single example but integrated for the entire groups of transcripts, we firstly stacked the integrated gradients of each of the top 4 models to obtain the importance scores over all the predictive features for each transcript. These scores were grouped into 100 bins along the transcript length, and the average importance score within each bin was calculated to create a length-standardised distribution of importance across the sequence.We used Uniform Manifold Approximation and Projection (UMAP) to reduce the binned

transcripts to 2D space, using correlation as the distance metric. Agglomerative clustering was then applied to group the transcripts into 2 clusters.

To elucidate the features making each cluster condensation-prone we continued with the following analysis separately for each cluster. We calculated the importance of nucleotide triplets across the transcripts. For each transcript, a sliding window approach was used to capture triplet sequences, and their integrated gradient scores were averaged for each triplet at every position. This reduced the transcript length by two. The triplet importance scores were then divided into 100 equal-length segments, and the scores were averaged across all transcripts. To further analyse the predictive importance of each POSTAR3 track, we calculated the average importance scores across all bins for each POSTAR3 RBP binding site in each transcript.

Validation of the predictive importance by investigating original data

We calculated the percentage of bound nucleotides for each transcript region (5'UTR, CDS, and 3'UTR) in cluster 1 and over the entire transcript in cluster 2. For cluster 2, binding percentages were computed separately for each transcript part, and for cluster 2, they were averaged across the entire transcript. To assess nucleotide frequency differences between smOOPs and control transcripts in both clusters, we binned the sequences and averaged the presence of each nucleotide (A, C, G, T) across bins. For cluster 1, the sequence was separated into transcript regions, with the 5'UTR divided into 10 bins and the CDS and 3'UTR each divided into 50 bins. For cluster 2, the entire transcript was divided into 100 bins. The average nucleotide frequency was calculated separately for smOOPs and control transcripts within each bin. The differences in nucleotide frequency were determined by subtracting the control average from the smOOPs average for each nucleotide in each bin. To further investigate sequence composition, we calculated the frequency of each possible nucleotide triplet (3-mer) across the entire sequence for cluster 2 and within the transcript regions for cluster 1. The triplet occurrence frequency for control transcripts was subtracted from that of smOOPs transcripts for each cluster, yielding the difference in triplet usage.A similar approach was applied to the global iCLIP cross-linking signals. For cluster 1, the normalised signal was binned and averaged by transcript regions, with the 5'UTR divided into 10 bins and the CDS and 3'UTR into 50 bins each. For cluster 2, the signal was binned and averaged across the entire transcript into 100 bins. The median signal value for each bin was calculated for both smOOPs and control transcripts, and a 95% confidence interval for the median was estimated using bootstrapping with 1000 samples.

For cluster 1, we evaluated the impact of specific transcript regions on model predictions by masking either sequence, global iCLIP, POSTAR, and PARIS intramolecular interactions features. The entire region of the transcript was masked using zeros for 5'UTR, CDS, and 3'UTR individually. Models trained on each dataset were then used to predict outcomes based on these masked inputs. The ROC curves were used to identify the region with the highest impact on the prediction when masked.For cluster 1, the number of PARIS intramolecular hybrids was calculated for each transcript part (5'UTR, CDS, 3'UTR). These counts were normalised by transcript region length and expression levels.

Investigation of the features of proteins encoded by smOOPs

For cluster 1, we analysed the codon usage in protein-coding transcripts. We calculated the frequency of each codon per transcript and compared the average codon usage between smOOPs and control

transcripts. A p-value was computed to test if the mean codon usage differed significantly between the two groups. Additionally, we calculated the percentage of C nucleotides within codons for each amino acid. We retrieved the annotated locations of intrinsically disordered regions (IDRs) for the proteins encoded by the transcripts from the Uniprot database (UniProt Consortium 2023). We then calculated the percentage of proteins in each group (smOOPs and control) that contained any IDRs and determined the average percentage of each protein that is disordered.. Finally, we evaluated the likelihood of each protein to localise in biomolecular condensates using PICNIC (Hadarovich et al., 2023), a deep learning model that predicts condensate involvement based on sequence and structure-based features from Alphafold2 models.

Data availability

Newly produced and publicly available data were used for this work. Raw PARIS data is available from GEO at GSE74353. Raw Ribo-seq data is available from GEO at GSE30839, with processed data obtained from the supplementary files. Raw RNA metabolism data is available at GEO at GSE169549. POSTAR3 CLIP datasets were obtained from the POSTAR3 platform (http://postar.ncrnalab.org/).

Code availability

The code and notebooks to analyse the data and to produce the figures in this work are available at: https://github.com/ModicLab/smOOPs_project.

Figure legends

Figure 1: Atlas of condensation-prone RNAs across early development.
A) Experimental framework to identify RNAs that are both semi-extractable and highly RBP-bound RNAs (smOOPs) using three different TRIzol-based RNA extractions: aqueous phase of non-crosslinked sample as control, aqueous phase of heated and sheared TRIzol sample as semi-extractable RNAs and interphase of crosslinked sample to obtain OOPS-enriched RNAs. (nPSCs - naïve pluripotent embryonic stem cells (ESCs), pPSCs - primed pluripotent stem cells, dPSCs - 1day Wnt-differentiated pPSCs)
B) Scatter plots showing the overlap between semi-extractability and OOPS-enriched genes (padj < 0.01) as well as the correlation between their fold changes compared to control for each developmental stage
C) UpSet plot showing the intersections between smOOPs identified in nPSCs, pPSCs, and dPSCs.
D) Percentage of gene biotypes for smOOPs and non-smOOPs in each cell-state. (TEC: To be Experimentally Confirmed).
E) Increased global iCLIP normalised cross-linking signal adjusted for length and expression (CPM

density/semi-extractability TPM) for smOOPs compared to non-smOOPs.
F) Representative HCR-FISH photomicrographs. Right panel is a zoomed-in image of the region in the dotted white box.
G) HCR-FISH Quantifications. (Top) Boxplot showing the mean of foci size for each target transcript, calculated as the mean of all foci per each transcript. (Bottom) Boxplot showing the mean of total intensity (mean intensity * area) of all foci of each transcript, either smOOPs or non-smOOPs (Welch's t-test, *p<0.05, **p<0.01). n indicates the number of different mRNAs against which the HCR-FISH probes were designed. In total >70 nuclei for a single transcript were imaged and > 890 of foci counted for each transcript. Scale bar = 5 μm

Figure 2: Deep learning efficiently classifies condensation-prone smOOPs from control transcripts.
A) Schematic visualisation of the feature encoding and deep learning model training across 127 unique combinations of transcriptomic datasets. Each combination was used to train a model through convolutional and recurrent layers for nPSCs smOOPs and control transcript classification.
B) Model performance across dataset combinations. (Left) AUC scores of the best model trained on each unique dataset combination. Dots represent included datasets, with lines connecting models where one is a subset of the other, colour-coded by AUC improvement. (Right) Boxplots showing the median performance and interquartile range as more datasets are included.
C) AUC score improvement at the addition of a particular dataset to each combination of previously included datasets. Colour indicates the p-value from a two-sided Student's t-test.
D) Pairwise feature combination analysis quantifying the difference of maximum individual and combined contributions of features to model performance. An asterisk (*) marks statistically significant results (p-value < 0.05, two-sided Student's t-test).

Figure 3: Predictive feature analysis reveals both known and novel RNA condensation characteristics.
A) Scatter plot of UMAP components for nPSC smOOPs transcripts based on binned importance scores per feature. Each dot represents a transcript, colour-coded by cluster. On each side, the pie charts depict gene types found in each cluster. (TEC: To be Experimentally Confirmed)
B) Heatmap showing the average individual feature importance scores across all transcripts in cluster 1, divided into 10 bins for the 5'UTR and 50 bins each for the CDS and 3'UTR.
C) Heatmap showing the average individual feature importance scores across all transcripts in cluster 2, binned into 100 bins over the entire transcript.
D) The per-bin difference in average nucleotide content between cluster 1 smOOPs and control transcripts, calculated by binning the nucleotide content of each transcript part into 10 bins for the 5'UTR and 50 bins each for the CDS and 3'UTR.
E) The per-bin difference in average nucleotide content between cluster 2 smOOPs and control transcripts, calculated by binning nucleotide content into 100 bins across the transcripts.
F) Median global iCLIP signal, expression normalised and binned (10 bins for the 5'UTR and 50 bins each for the CDS and 3'UTR) for cluster 1 and control transcripts, with the shaded areas representing 95% confidence intervals at each bin, estimated via bootstrapping.
G) Median global iCLIP signal, expression normalised and binned (100 bins across the transcripts) for cluster 2 smOOPs and control transcripts, with the shaded areas representing 95% confidence intervals at each bin, estimated via bootstrapping.
H) Bar charts showing PARIS hybrid counts across individual transcript parts for cluster 1 smOOPs

compared to control transcripts, adjusted for length and expression.
I) Average importance scores of the top and bottom 10 CLIP datasets from POSTAR3 across all positions for transcripts in cluster 1, shown with 95% confidence intervals.
J) ROC curves showing the effect of masking specific transcript regions (5'UTR, CDS, 3'UTR) on the predictions of models trained solely on either sequence, global iCLIP, POSTAR, and PARIS-Intra features for cluster 1 smOOPs and control transcripts.

Figure 4: smOOPs are densely connected in RRI networks and encode proteins with phase separating ability.
A) Schematic overview of the approach for inferring RNA-RNA interaction (RRI) networks in nPSCs and pPSCs using the RIC-seq method (Cai et al., 2020), with key steps shown.
B) RIC-seq hybrid read counts in nPSCs and pPSCs (pooled from three replicates each), categorised by interaction type, excluding hybrid reads containing rRNA, tRNA, and mitochondrial RNA.
C) Analysis of smOOPs degree (number of distinct connections) in the RIC-seq network compared to non-smOOPs. (Left) Degree distribution for smOOPs versus non-smOOPs. (Right) Scatter plots showing the positive correlation between degree and expression across gene length quartiles (1-4). Blue and green Pearson coefficients (R) represent nPSCs and pPSCs smOOPs, respectively, while dark grey represents all genes.
D) Visualization of the nPSCs smOOPs RRI subnetwork from RIC-seq data, where nodes represent genes and edges represent interactions based on hybrid reads. Node size corresponds to degree, and edge width represents the number of hybrid reads between nodes. Unconnected smOOPs are displayed on the right, and connectivity metrics below.
E) smOOPs in nPSCs and pPSCs are more interconnected than expected by chance. Density plots show the distribution of connectivity metrics from 10,000 random or sub-sampled networks generated using three null models. The dashed line indicates observed values for the smOOPs RIC-seq subnetwork, with p-values from permutation tests comparing smOOPs subnetworks to degree-matched sub-sampled networks. (LCC: Largest connected component).) Differences in codon usage between smOOPs mRNAs in cluster 1 and control protein-coding transcripts, with significance assessed via Welch's t-test. The sizes represent the percentage of C nucleotides in codons for each amino acid.
G) The percentage of smOOPs mRNAs in cluster 1 and control transcripts coding for IDR-containing proteins.
H) The percentage of the disordered regions in proteins coded by smOOPs mRNAs in cluster 1 and control transcripts.
H) The predicted PICNIC (Proteins Involved in CoNdensates In Cells) (Hadarovich et al., 2023) score for the proteins coded for by mRNA cluster smOOPs and control transcripts.

Supplementary figure legends

Figure S1: Identification of condensation-prone RNAs.
A) Pairwise Pearson correlations between normalised gene-level counts for control, OOPS and semi-extractability assay samples in nPSCs, pPSCs and dPSCs shows high reproducibility among replicates. The dot size is proportional to the Pearson correlation coefficient.
B) Principal component analysis using normalised gene-level counts using the top 2000 most variable

genes separates samples on stage (PC1) and assay (PC2).
C) smOOPs identification: Venn diagrams showing the overlap between semi-extractability assay and OOPS-enriched genes (differentially expressed vs control; LFC > 1, padj < 0.01) in nPSCs, pPSCs and dPSCs.
D) Expression of example smOOPs for which fold changes in OOPS vs control are not equally high to the fold changes in semi-extractability assay vs control, and vice-versa. For each stage, the top five genes with highest fold changes in semi-extractability assay data and top five genes with highest fold changes in OOPS data are shown.
(E) Expression profiles across assays for the candidate mRNAs used for HCR-FISH.
(F) Fraction of nuclear signal for smOOPs compared to non-smOOPs based on the HCR-FISH data
(G) Comparative visualisation of an example smOOPs that was visualised by HCR-FISH, Peg3 . Normalised crosslinking signal and coverage across assays plotted using *clipplotr* (Chakrabarti et al., 2023a).

Figure S2: Training deep learning models to classify nPSCs smOOPs transcripts.
A) Distribution of full transcript lengths and individual transcript region lengths (Log_{10}) for nPSCs smOOPs and controls (Welch's t-test, *p<0.05, **p<0.01, ***p<0.001).
B) Distribution of expression levels (Log_{10} (TPM)) for nPSCs smOOPs and controls (Welch's t-test, *p<0.05, **p<0.01, ***p<0.001).
C) Training and validation loss, accuracy, and AUC curves over training epochs for models trained on sequence (left) and POSTAR3 peaks (right) features. The dashed lines indicate the early stopping points based on validation AUC performance.
D) Classification performance of models trained on each individual feature set. (Left) Confusion matrices displaying the distribution of true and predicted labels for nPSCs smOOPs and control transcripts. (Right) ROC curves that illustrate the true positive rate (TPR) versus the false positive rate (FPR), with the area under the curve (AUC) indicated.
E) Importance scores across transcript positions for Dnmt1 (left) and Tmem265 (right). Importance scores are derived from models trained on either global iCLIP, PARIS-Intra interactions, and the POSTAR3 peak dataset, shown in the respective tracks, with POSTAR3 RBP importance profiles shown for individual RBPs.

Figure S3: Identifying the positional importance of the most predictive features.
A) Heatmap showing the average individual feature importance scores across all transcripts in cluster 1, divided into 100 bins over the entire transcript, with expanded individual POSTAR3 RBP importance.
B) As in (A), but for cluster 2.
C) Heatmap showing the importance scores of the sequence model averaged over a sliding window for individual triplets, binned into 100 bins over the entire transcript and averaged per bin for cluster 1.
D) As in (C), but for cluster 2.
E) The difference in average triplet content between cluster 1 nPSCs smOOPs and control transcripts for each transcript region.

Figure S4: Characterisation of the RIC-seq data and networks in nPSCs and pPSCs.
A) Pearson correlation between samples using normalised inter-molecular gene-level hybrid read counts.
B) PCA of the RIC-seq replicates in nPSCs and pPSCs based on the inter-molecular hybrid reads

using top 2000 most variable genes.
C) Circos plots showing the region links based on inter-molecular hybrid reads, with percentages indicating the proportion of interactions for each region type.
D) Evaluation of the scale-free-like nature of the RIC-seq networks: degree distribution vs gene-counts on a log-log scale for the derived RRI networks. The red line indicates linear fit of the degree distribution on the log-log scale.
(E) Small world properties for the derived RRI networks: Mean global clustering coefficient compared to those of 100 random networks of the same size (equal number of nodes and edges).
(F) Small world properties for the derived RRI networks: Distribution of path lengths in the nPSCs and pPSCs RRI networks, with dashed vertical line indicating the average path length.
(G) Half-lives (Herzog et al., 2017) of nPSCs smOOPs mRNAs compared to those of control transcripts.
(H) Rate constants of decay of nPSCs smOOPs mRNAs compared to those of control transcripts.
(I) Violin plot showing translation efficiency in nPSCs for smOOPs and control transcripts obtained by Ribo-seq (Ingolia et al., 2011).

Acknowledgements

We would like to thank Flora Lee for establishing the RIC-seq library preparation method. We would like to thank Oscar Wilkins for designing guides for rRNA depletion using Ribocutter and helping with the Ultraplex demultiplexing tool, Eneko Villanueva for providing us with a detailed protocol for OOPS, Yuanchao Xue for providing a detailed RIC-seq protocol and Joel Ryan for continuous advice on HCR-FISH imaging together with Petra Čotar and Luka Čehovin Zajc for automated image acquisition and Thomas H Kapral with other POSTAR3 authors for access to the full CLIPdb dataset. We would also like to thank Tine Tesovnik and Jernej Kovač for their assistance and sequencing at the University Children's Hospital at Ljubljana University Medical Centre. We also wish to acknowledge the Advanced Sequencing Facility and Nemo HPC at The Francis Crick Institute and HPC VEGA at the IZUM, the Institute of Information Science.

Author contributions

T.K.: Conceptualization, Investigation, Project Administration, Formal Analysis, Writing - Original Draft Preparation, Writing - Review Editing; J.N.: Conceptualization, Methodology, Formal Analysis, Visualization, Data Curation, Writing - Original Draft Preparation, Writing - Review Editing; I.A.I.: Conceptualization, Methodology, Software, Formal Analysis, Visualization, Data Curation, Writing - Original Draft Preparation, Writing - Review Editing, Project Administration, Supervision; B.K.: Software, Formal Analysis; I.U.: Software; D.M.J.: Software, Formal Analysis; A.M.C.: Software, Writing - Review Editing, Supervision; N.M.L.: Resources, Funding Acquisition, Supervision; J.U.: Supervision, Resources, Funding Acquisition; M.M.: Conceptualization, Supervision, Project Administration, Resources, Funding Acquisition, Writing - Original Draft Preparation, Writing - Review Editing.

Grant Information

This work was supported by Wellcome (218672/Z/19/Z), Wellcome [215593], a Janko Jamnik PhD Fellowship awarded to TK, the Slovenian Research Agency (J4-50145, J7-2596, N1-0240, P1-0060), the European Research Council (ERC) under the EU's Horizon 2020 research and innovation programme (grant agreement no 835300), The Francis Crick Institute, which receives its core funding from Cancer Research UK (CC0102), the UK Medical Research Council (CC0102), and Wellcome (CC0102), and by the UK Dementia Research Institute [RE21605] which receives its funding from the UK Medical Research Council. HPC VEGA is financed through HPC RIVR consortium (www.hpc-rivr.si) and EuroHPC JU (eurohpc-ju.europa.eu)

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Alberti, S., Hyman, A.A., 2021. Biomolecular condensates at the nexus of cellular stress, protein aggregation disease and ageing. Nat. Rev. Mol. Cell Biol. 22, 196–213.

Boija, A., Klein, I.A., Young, R.A., 2021. Biomolecular condensates and cancer. Cancer Cell 39, 174–192.

Bushnell, B., Rood, J., Singer, E., 2017. BBMerge - Accurate paired shotgun read merging via overlap. PLoS One 12, e0185056.

Cabral, S.E., Otis, J.P., Mowry, K.L., 2022. Multivalent interactions with RNA drive recruitment and dynamics in biomolecular condensates in Xenopus oocytes. iScience 25, 104811.

Cai, Z., Cao, C., Ji, L., Ye, R., Wang, D., Xia, C., Wang, S., Du, Z., Hu, N., Yu, X., Chen, J., Wang, L., Yang, X., He, S., Xue, Y., 2020. RIC-seq for global in situ profiling of RNA–RNA spatial interactions. Nature 582, 432–437.

Cao, C., Cai, Z., Ye, R., Su, R., Hu, N., Zhao, H., Xue, Y., 2021. Global in situ profiling of RNA-RNA spatial interactions with RIC-seq. Nat. Protoc. 16, 2916–2946.

Cardona, A.H., Ecsedi, S., Khier, M., Yi, Z., Bahri, A., Ouertani, A., Valero, F., Labrosse, M., Rouquet, S., Robert, S., Loubat, A., Adekunle, D., Hubstenberger, A., 2023. Self-demixing of mRNA copies buffers mRNA:mRNA and mRNA:regulator stoichiometries. Cell 186, 4310–4324.e23.

Chakrabarti, A.M., Capitanchik, C., Ule, J., Luscombe, N.M., 2023a. clipplotr-a comparative visualization and analysis tool for CLIP data. RNA 29, 715–723.

Chakrabarti, A.M., Iosub, I.A., Lee, F.C.Y., Ule, J., Luscombe, N.M., 2023b. A computationally-enhanced hiCLIP atlas reveals Staufen1-RNA binding features and links 3′ UTR structure to RNA metabolism. Nucleic Acids Res. 51, 3573–3589.

Chen, X., Fansler, M.M., Janjoš, U., Ule, J., Mayr, C., 2024. The FXR1 network acts as a signaling scaffold for actomyosin remodeling. Cell. https://doi.org/10.1016/j.cell.2024.07.015

Chen, X., Mayr, C., 2022. A working model for condensate RNA-binding proteins as matchmakers for protein complex assembly. RNA 28, 76–87.

Choi, H.M.T., Schwarzkopf, M., Fornace, M.E., Acharya, A., Artavanis, G., Stegmaier, J., Cunha, A., Pierce, N.A., 2018. Third-generation in situ hybridization chain reaction: multiplexed, quantitative, sensitive, versatile, robust. Development 145. https://doi.org/10.1242/dev.165753

Chujo, T., Yamazaki, T., Kawaguchi, T., Kurosaka, S., Takumi, T., Nakagawa, S.,

Hirose, T., 2017. Unusual semi-extractability as a hallmark of nuclear body-associated architectural noncodingRNAs. EMBO J. 36, 1447–1462.
Clemson, C.M., Hutchinson, J.N., Sara, S.A., Ensminger, A.W., Fox, A.H., Chess, A., Lawrence, J.B., 2009. An architectural role for a nuclear noncoding RNA: NEAT1 RNA is essential for the structure of paraspeckles. Mol. Cell 33, 717–726.
Csárdi, G., Nepusz, T., 2006. The igraph software package for complex network research.
Csárdi, G., Nepusz, T., Müller, K., Horvát, S., Traag, V., Zanini, F., Noom, D., 2024. igraph for R: R interface of the igraph library for graph theory and network analysis. Zenodo. https://doi.org/10.5281/ZENODO.7682609
Eno, C., Hansen, C.L., Pelegri, F., 2019. Aggregation, segregation, and dispersal of homotypic germ plasm RNPs in the early zebrafish embryo. Dev. Dyn. 248, 306–318.
Ewels, P.A., Peltzer, A., Fillinger, S., Patel, H., Alneberg, J., Wilm, A., Garcia, M.U., Di Tommaso, P., Nahnsen, S., 2020. The nf-core framework for community-curated bioinformatics pipelines. Nat. Biotechnol. 38, 276–278.
Fay, M.M., Anderson, P.J., 2018. The role of RNA in biological phase separations. J. Mol. Biol. 430, 4685–4701.
Forrest, K.M., Gavis, E.R., 2003. Live imaging of endogenous RNA reveals a diffusion and entrapment mechanism for nanos mRNA localization in Drosophila. Curr. Biol. 13, 1159–1168.
Guillén-Boixet, J., Kopach, A., Holehouse, A.S., Wittmann, S., Jahnel, M., Schlüßler, R., Kim, K., Trussina, I.R.E.A., Wang, J., Mateju, D., Poser, I., Maharana, S., Ruer-Gruß, M., Richter, D., Zhang, X., Chang, Y.-T., Guck, J., Honigmann, A., Mahamid, J., Hyman, A.A., Pappu, R.V., Alberti, S., Franzmann, T.M., 2020. RNA-Induced Conformational Switching and Clustering of G3BP Drive Stress Granule Assembly by Condensation. Cell 181, 346–361.e17.
Hadarovich, A., Singh, H.R., Ghosh, S., Rostam, N., Hyman, A.A., Toth-Petroczy, A., 2023. PICNIC accurately predicts condensate-forming proteins regardless of their structural disorder across organisms. bioRxiv. https://doi.org/10.1101/2023.06.01.543229
Han, T.W., Portz, B., Young, R.A., Boija, A., Klein, I.A., 2024. RNA and condensates: Disease implications and therapeutic opportunities. Cell Chem. Biol. 31, 1593–1609.
He, D.C., Nickerson, J.A., Penman, S., 1990. Core filaments of the nuclear matrix. J. Cell Biol. 110, 569–580.
Herzog, V.A., Reichholf, B., Neumann, T., Rescheneder, P., Bhat, P., Burkard, T.R., Wlotzka, W., von Haeseler, A., Zuber, J., Ameres, S.L., 2017. Thiol-linked alkylation of RNA to assess expression dynamics. Nat. Methods 14, 1198–1204.
Hitz, C., Wurst, W., Kühn, R., 2007. Conditional brain-specific knockdown of MAPK using Cre/loxP regulated RNA interference. Nucleic Acids Res. 35, e90.
Hofacker, I.L., Fontana, W., Stadler, P.F., Bonhoeffer, L.S., Tacker, M., Schuster, P., 1994. Fast folding and comparison of RNA secondary structures. Monatsh. Chem. 125, 167–188.
Holehouse, A.S., Kragelund, B.B., 2024. The molecular basis for cellular function of intrinsically disordered protein regions. Nat. Rev. Mol. Cell Biol. 25, 187–211.
Horvát, S., Modes, C.D., 2021. Connectedness matters: construction and exact random sampling of connected networks. J. Phys. Complex. 2, 015008.
Ingolia, N.T., Lareau, L.F., Weissman, J.S., 2011. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. Cell 147, 789–802.
Khong, A., Matheny, T., Jain, S., Mitchell, S.F., Wheeler, J.R., Parker, R., 2017. The Stress Granule Transcriptome Reveals Principles of mRNA Accumulation in Stress

Granules. Mol. Cell 68, 808–820.e5.
Knuckles, P., Carl, S.H., Musheev, M., Niehrs, C., Wenger, A., Bühler, M., 2017. RNA fate determination through cotranscriptional adenosine methylation and microprocessor binding. Nat. Struct. Mol. Biol. 24, 561–569.
Lee, F.C.Y., Chakrabarti, A.M., Hänel, H., Monzón-Casanova, E., Hallegger, M., Militti, C., Capraro, F., Sadée, C., Toolan-Kerr, P., Wilkins, O., Turner, M., König, J., Sibley, C.R., Ule, J., 2021. An improved iCLIP protocol. bioRxiv. https://doi.org/10.1101/2021.08.27.457890
Liao, Y., Smyth, G.K., Shi, W., 2013. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 30, 923–930.
Lorenz, R., Bernhart, S.H., Höner Zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P.F., Hofacker, I.L., 2011. ViennaRNA Package 2.0. Algorithms Mol. Biol. 6, 26.
Love, M.I., Huber, W., Anders, S., 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15, 550.
Lu, Z., Zhang, Q.C., Lee, B., Flynn, R.A., Smith, M.A., Robinson, J.T., Davidovich, C., Gooding, A.R., Goodrich, K.J., Mattick, J.S., Mesirov, J.P., Cech, T.R., Chang, H.Y., 2016. RNA Duplex Map in Living Cells Reveals Higher-Order Transcriptome Structure. Cell 165, 1267–1279.
Maharana, S., Wang, J., Papadopoulos, D.K., Richter, D., Pozniakovsky, A., Poser, I., Bickle, M., Rizk, S., Guillén-Boixet, J., Franzmann, T.M., Jahnel, M., Marrone, L., Chang, Y.-T., Sterneckert, J., Tomancak, P., Hyman, A.A., Alberti, S., 2018. RNA buffers the phase separation behavior of prion-like RNA binding proteins. Science 360, 918–921.
Martin, M., 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 17, 10–12.
Masuda, A., Okamoto, T., Kawachi, T., Takeda, J.-I., Hamaguchi, T., Ohno, K., 2024. Blending and separating dynamics of RNA-binding proteins develop architectural splicing networks spreading throughout the nucleus. Mol. Cell 84, 2949–2965.e10.
Ma, W., Mayr, C., 2018. A Membraneless Organelle Associated with the Endoplasmic Reticulum Enables 3'UTR-Mediated Protein-Protein Interactions. Cell 175, 1492–1506.e19.
Ma, W., Zhen, G., Xie, W., Mayr, C., 2021. In vivo reconstitution finds multivalent RNA–RNA interactions as drivers of mesh-like condensates. Elife 10, e64252.
Mittag, T., Pappu, R.V., 2022. A conceptual framework for understanding phase separation and addressing open questions and challenges. Mol. Cell 82, 2201–2214.
Modic, M., Grosch, M., Rot, G., Schirge, S., Lepko, T., Yamazaki, T., Lee, F.C.Y., Rusha, E., Shaposhnikov, D., Palo, M., Merl-Pham, J., Cacchiarelli, D., Rogelj, B., Hauck, S.M., von Mering, C., Meissner, A., Lickert, H., Hirose, T., Ule, J., Drukker, M., 2019. Cross-Regulation between TDP-43 and Paraspeckles Promotes Pluripotency-Differentiation Transition. Mol. Cell 74, 951–965.e13.
Modic, M., Kuret, K., Steinhauser, S., Faraway, R., van Genderen, E., Ruiz de Los Mozos, I., Novljan, J., Vičič, Ž., Lee, F.C.Y., ten Berge, D., Luscombe, N.M., Ule, J., 2024. Poised PABP–RNA hubs implement signal-dependent mRNA decay in development. Nat. Struct. Mol. Biol. 1–9.
Patel, A., Lee, H.O., Jawerth, L., Maharana, S., Jahnel, M., Hein, M.Y., Stoynov, S., Mahamid, J., Saha, S., Franzmann, T.M., Pozniakovski, A., Poser, I., Maghelli, N., Royer, L.A., Weigert, M., Myers, E.W., Grill, S., Drechsel, D., Hyman, A.A., Alberti, S., 2015. A Liquid-to-Solid Phase Transition of the ALS Protein FUS Accelerated by Disease Mutation. Cell 162, 1066–1077.
Pichon, X., Bastide, A., Safieddine, A., Chouaib, R., Samacoits, A., Basyuk, E., Peter, M., Mueller, F., Bertrand, E., 2016. Visualization of single endogenous polysomes

reveals the dynamics of translation in live human cells. J. Cell Biol. 214, 769–781.
Quinlan, A.R., Hall, I.M., 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842.
Ramaswami, M., Taylor, J.P., Parker, R., 2013. Altered ribostasis: RNA-protein granules in degenerative disorders. Cell 154, 727–736.
Roden, C., Gladfelter, A.S., 2021. RNA contributions to the form and function of biomolecular condensates. Nat. Rev. Mol. Cell Biol. 22, 183–195.
Sanders, D.W., Kedersha, N., Lee, D.S.W., Strom, A.R., Drake, V., Riback, J.A., Bracha, D., Eeftens, J.M., Iwanicki, A., Wang, A., Wei, M.-T., Whitney, G., Lyons, S.M., Anderson, P., Jacobs, W.M., Ivanov, P., Brangwynne, C.P., 2020. Competing Protein-RNA Interaction Networks Control Multiphase Intracellular Organization. Cell 181, 306–324.e28.
Scheibner, K., Schirge, S., Burtscher, I., Büttner, M., Sterr, M., Yang, D., Böttcher, A., Ansarullah, Irmler, M., Beckers, J., Cernilogar, F.M., Schotta, G., Theis, F.J., Lickert, H., 2021. Epithelial cell plasticity drives endoderm formation during gastrulation. Nat. Cell Biol. 23, 692–703.
Schwartz, J.C., Wang, X., Podell, E.R., Cech, T.R., 2013. RNA seeds higher-order assembly of FUS protein. Cell Rep. 5, 918–925.
Spitale, R.C., Incarnato, D., 2023. Probing the dynamic RNA structurome and its functions. Nat. Rev. Genet. 24, 178–196.
Stephens, M., 2017. False discovery rates: a new deal. Biostatistics 18, 275–294.
Tauber, D., Tauber, G., Parker, R., 2020. Mechanisms and Regulation of RNA Condensation in RNP Granule Formation. Trends Biochem. Sci. 45, 764–778.
Trcek, T., Douglas, T.E., Grosch, M., Yin, Y., Eagle, W.V.I., Gavis, E.R., Shroff, H., Rothenberg, E., Lehmann, R., 2020. Sequence-Independent Self-Assembly of Germ Granule mRNAs into Homotypic Clusters. Mol. Cell 78, 941–950.e12.
Trcek, T., Grosch, M., York, A., Shroff, H., Lionnet, T., Lehmann, R., 2015. Drosophila germ granules are structured and contain homotypic mRNA clusters. Nat. Commun. 6, 1–12.
Uversky, V.N., Kuznetsova, I.M., Turoverov, K.K., Zaslavsky, B., 2015. Intrinsically disordered proteins as crucial constituents of cellular aqueous two phase systems and coacervates. FEBS Lett. 589, 15–22.
Van Treeck, B., Protter, D.S.W., Matheny, T., Khong, A., Link, C.D., Parker, R., 2018. RNA self-assembly contributes to stress granule formation and defining the stress granule transcriptome. Proc. Natl. Acad. Sci. U. S. A. 115, 2734–2739.
Villanueva, E., Smith, T., Queiroz, R.M.L., Monti, M., Pizzinga, M., Elzek, M., Dezi, V., Harvey, R.F., Ramakrishna, M., Willis, A.E., Lilley, K.S., 2020. Efficient recovery of the RNA-bound proteome and protein-bound transcriptome using phase separation (OOPS). Nat. Protoc. 15, 2568–2588.
Wang, J., Choi, J.-M., Holehouse, A.S., Lee, H.O., Zhang, X., Jahnel, M., Maharana, S., Lemaitre, R., Pozniakovsky, A., Drechsel, D., Poser, I., Pappu, R.V., Alberti, S., Hyman, A.A., 2018. A Molecular Grammar Governing the Driving Forces for Phase Separation of Prion-like RNA Binding Proteins. Cell 174, 688–699.e16.
West, C., Capitanchik, C., Cheshire, C., Luscombe, N.M., Chakrabarti, A., Ule, J., 2023. nf-core/clipseq - a robust Nextflow pipeline for comprehensive CLIP data analysis. Wellcome Open Res 8, 286.
Wilbertz, J.H., Voigt, F., Horvathova, I., Roth, G., Zhan, Y., Chao, J.A., 2019. Single-molecule imaging of mRNA localization and regulation during the integrated stress response. Mol. Cell 73, 946–958.e7.
Wilkins, O.G., Ule, J., 2021. Ribocutter: Cas9-mediated rRNA depletion from multiplexed Ribo-seq libraries. https://doi.org/10.1101/2021.07.14.451473
Yamazaki, T., Souquere, S., Chujo, T., Kobelke, S., Chong, Y.S., Fox, A.H., Bond, C.S., Nakagawa, S., Pierron, G., Hirose, T., 2018. Functional Domains of NEAT1

Architectural lncRNA Induce Paraspeckle Assembly through Phase Separation. Mol. Cell 70, 1038–1053.e7.
Yang, P., Humphrey, S.J., Cinghu, S., Pathania, R., Oldfield, A.J., Kumar, D., Perera, D., Yang, J.Y.H., James, D.E., Mann, M., Jothi, R., 2019. Multi-omic Profiling Reveals Dynamics of the Phased Progression of Pluripotency. Cell Systems. https://doi.org/10.1016/j.cels.2019.03.012
Ying, Q.-L., Wray, J., Nichols, J., Batlle-Morera, L., Doble, B., Woodgett, J., Cohen, P., Smith, A., 2008. The ground state of embryonic stem cell self-renewal. Nature 453, 519–523.
Zeng, C., Chujo, T., Hirose, T., Hamada, M., 2023. Landscape of semi-extractable RNAs across five human cell lines. Nucleic Acids Res. 51, 7820–7831.
Zhao, W., Zhang, S., Zhu, Y., Xi, X., Bao, P., Ma, Z., Kapral, T.H., Chen, S., Zagrovic, B., Yang, Y.T., Lu, Z.J., 2022. POSTAR3: an updated platform for exploring post-transcriptional regulation coordinated by RNA-binding proteins. Nucleic Acids Res. 50, D287–D294.

Figure 1

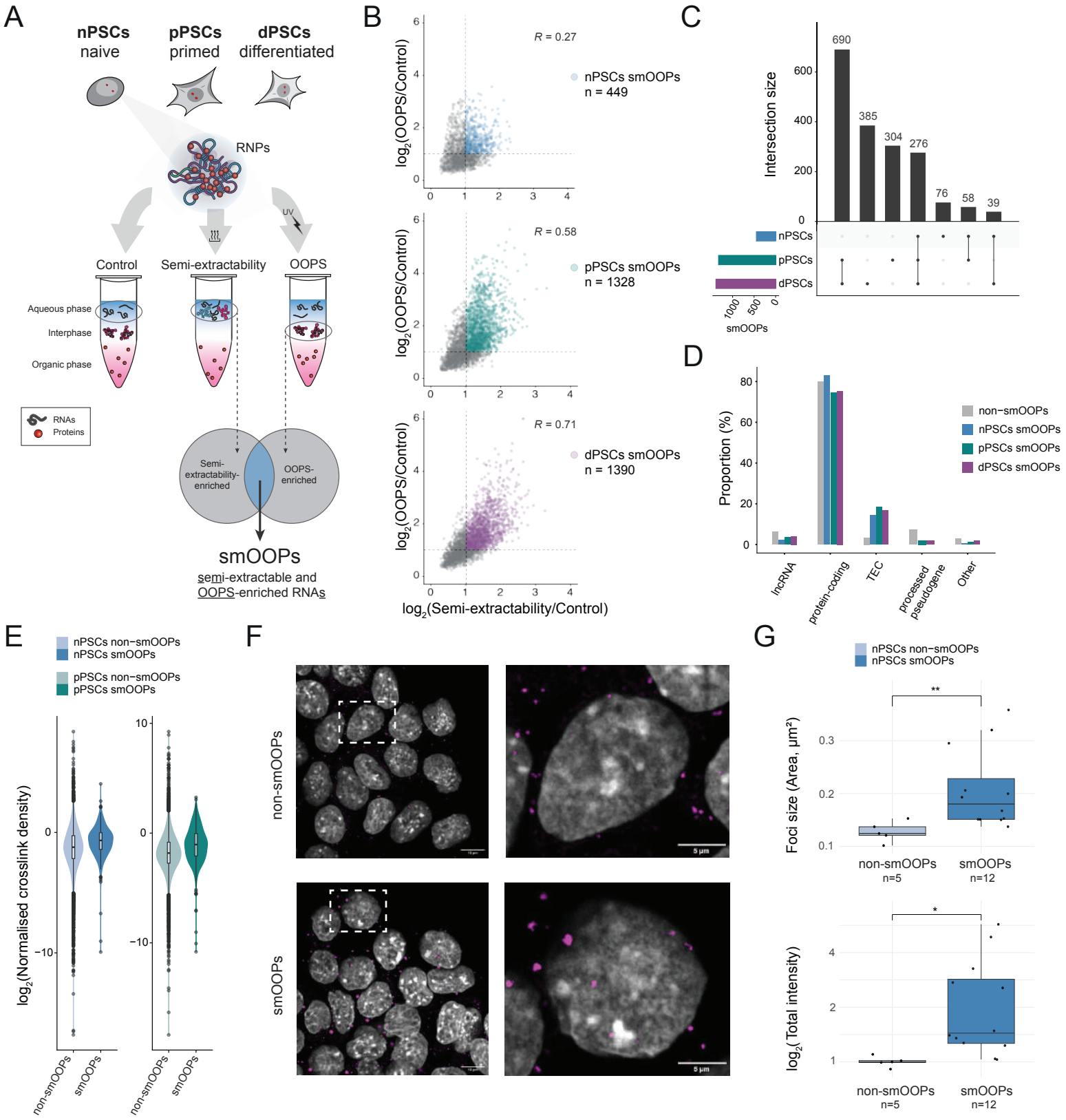

A
nPSCs
naive
pPSCs
primed
dPSCs
differentiated
RNPs
UV
Control
Semi-extractability
OOPS
Aqueous phase
Interphase
Organic phase
RNAs
Proteins
Semi-extractability-enriched
OOPS-enriched
smOOPs
semi-extractable and OOPS-enriched RNAs
B
R = 0.27
nPSCs smOOPs
n = 449
R = 0.58
pPSCs smOOPs
n = 1328
R = 0.71
dPSCs smOOPs
n = 1390
log2(OOPS/Control)
log2(Semi-extractability/Control)
C
Intersection size
690
385
304
276
76
58
39
nPSCs
pPSCs
dPSCs
smOOPs
D
Proportion (%)
non−smOOPs
nPSCs smOOPs
pPSCs smOOPs
dPSCs smOOPs
lncRNA
protein-coding
TEC
processed pseudogene
Other
E
nPSCs non−smOOPs
nPSCs smOOPs
pPSCs non−smOOPs
pPSCs smOOPs
log2(Normalised crosslink density)
non-smOOPs
smOOPs
F
non-smOOPs
smOOPs
5 µm
G
nPSCs non−smOOPs
nPSCs smOOPs
Foci size (Area, µm²)
**
non-smOOPs
n=5
smOOPs
n=12
log2(Total intensity)
*

Figure 2

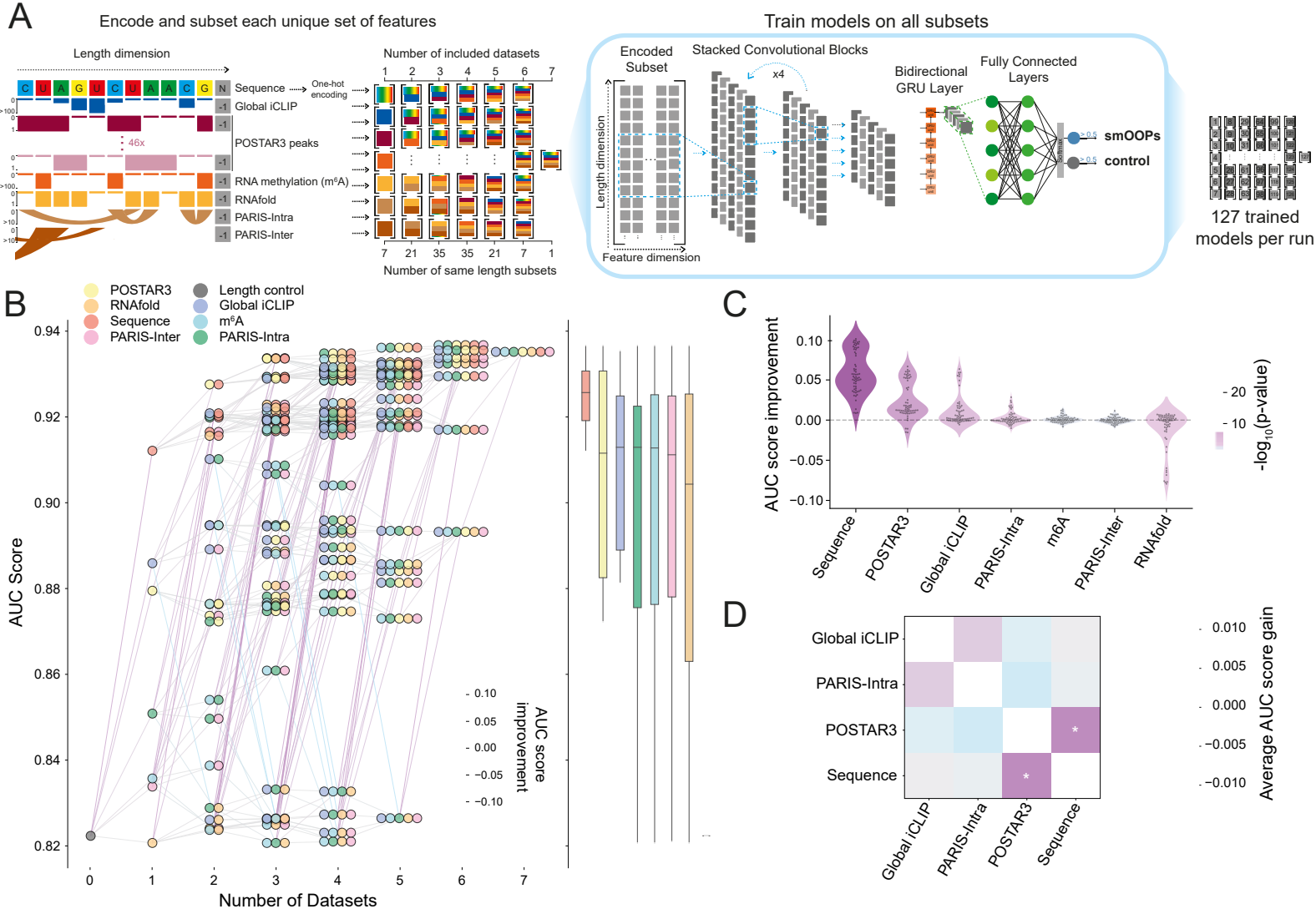

A
Encode and subset each unique set of features
Length dimension
Number of included datasets
Sequence
One-hot encoding
Global iCLIP
POSTAR3 peaks
46x
RNA methylation (m6A)
RNAfold
PARIS-Intra
PARIS-Inter
Number of same length subsets
Train models on all subsets
Encoded Subset
Stacked Convolutional Blocks
x4
Bidirectional GRU Layer
Fully Connected Layers
Length dimension
Feature dimension
smOOPs
control
127 trained models per run
B
POSTAR3
RNAfold
Sequence
PARIS-Inter
Length control
Global iCLIP
m6A
PARIS-Intra
AUC Score
Number of Datasets
AUC score improvement
C
AUC score improvement
-log10(p-value)
Sequence
POSTAR3
Global iCLIP
PARIS-Intra
m6A
PARIS-Inter
RNAfold
D
Global iCLIP
PARIS-Intra
POSTAR3
Sequence
Average AUC score gain

Figure 3

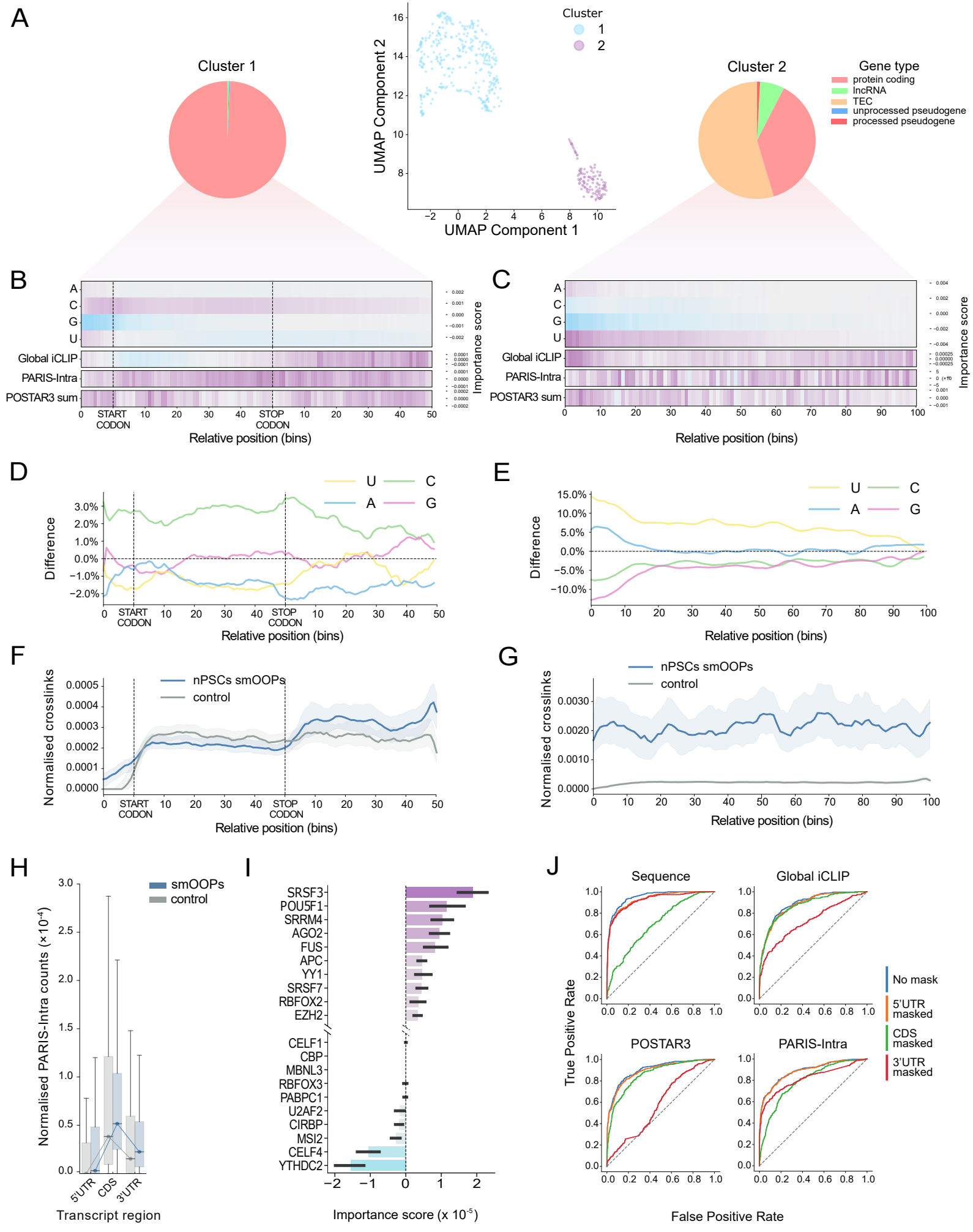

A
Cluster 1
Cluster 2
Cluster
1
2
UMAP Component 2
UMAP Component 1
Gene type
protein coding
lncRNA
TEC
unprocessed pseudogene
processed pseudogene
B
C
A
C
G
U
Global iCLIP
PARIS-Intra
POSTAR3 sum
START CODON
STOP CODON
Relative position (bins)
Importance score
D
E
U
C
A
G
Difference
F
G
nPSCs smOOPs
control
Normalised crosslinks
H
smOOPs
control
Normalised PARIS-Intra counts (×10⁻⁴)
5'UTR
CDS
3'UTR
Transcript region
I
SRSF3
POU5F1
SRRM4
AGO2
FUS
APC
YY1
SRSF7
RBFOX2
EZH2
CELF1
CBP
MBNL3
RBFOX3
PABPC1
U2AF2
CIRBP
MSI2
CELF4
YTHDC2
Importance score (x 10⁻⁵)
J
Sequence
Global iCLIP
POSTAR3
PARIS-Intra
True Positive Rate
False Positive Rate
No mask
5'UTR masked
CDS masked
3'UTR masked

Figure 4

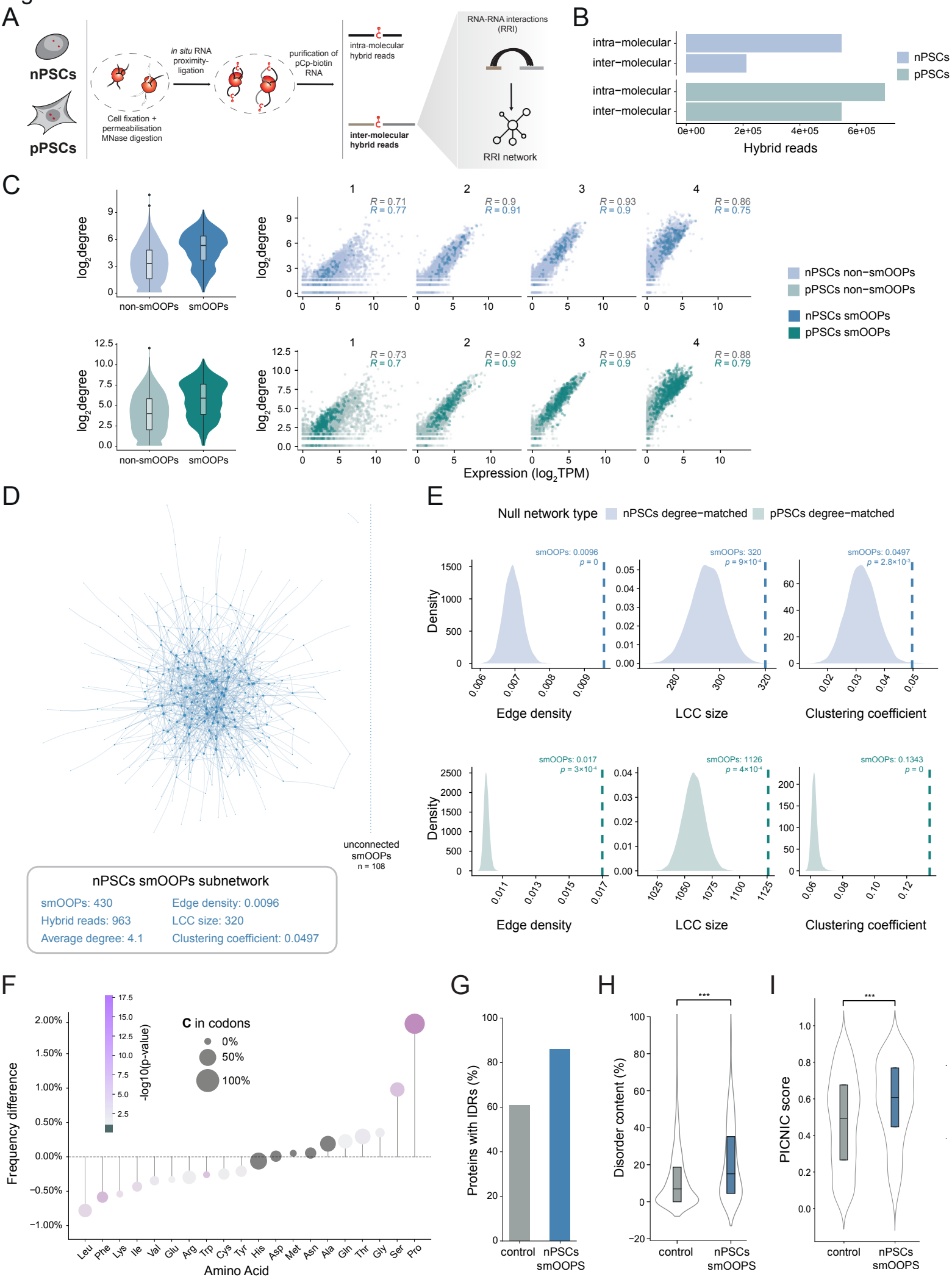

A
nPSCs
pPSCs
Cell fixation + permeabilisation MNase digestion
in situ RNA proximity-ligation
purification of pCp-biotin RNA
intra-molecular hybrid reads
inter-molecular hybrid reads
RNA-RNA interactions (RRI)
RRI network
B
intra−molecular
inter−molecular
intra−molecular
inter−molecular
nPSCs
pPSCs
0e+00
2e+05
4e+05
6e+05
Hybrid reads
C
log2degree
non-smOOPs
smOOPs
R = 0.71
R = 0.77
R = 0.9
R = 0.91
R = 0.93
R = 0.9
R = 0.86
R = 0.75
R = 0.73
R = 0.7
R = 0.92
R = 0.9
R = 0.95
R = 0.9
R = 0.88
R = 0.79
Expression (log2TPM)
nPSCs non−smOOPs
pPSCs non−smOOPs
nPSCs smOOPs
pPSCs smOOPs
D
unconnected smOOPs
n = 108
nPSCs smOOPs subnetwork
smOOPs: 430
Hybrid reads: 963
Average degree: 4.1
Edge density: 0.0096
LCC size: 320
Clustering coefficient: 0.0497
E
Null network type
nPSCs degree−matched
pPSCs degree−matched
Density
smOOPs: 0.0096
p = 0
smOOPs: 320
p = 9×10-4
smOOPs: 0.0497
p = 2.8×10-3
smOOPs: 0.017
p = 3×10-4
smOOPs: 1126
p = 4×10-4
smOOPs: 0.1343
p = 0
Edge density
LCC size
Clustering coefficient
F
Frequency difference
-log10(p-value)
C in codons
0%
50%
100%
Leu
Phe
Lys
Ile
Val
Glu
Arg
Trp
Cys
Tyr
His
Asp
Met
Asn
Ala
Gln
Thr
Gly
Ser
Pro
Amino Acid
G
Proteins with IDRs (%)
control
nPSCs smOOPS
H
Disorder content (%)
***
control
nPSCs smOOPS
I
PICNIC score
***
control
nPSCs smOOPS

Figure S1

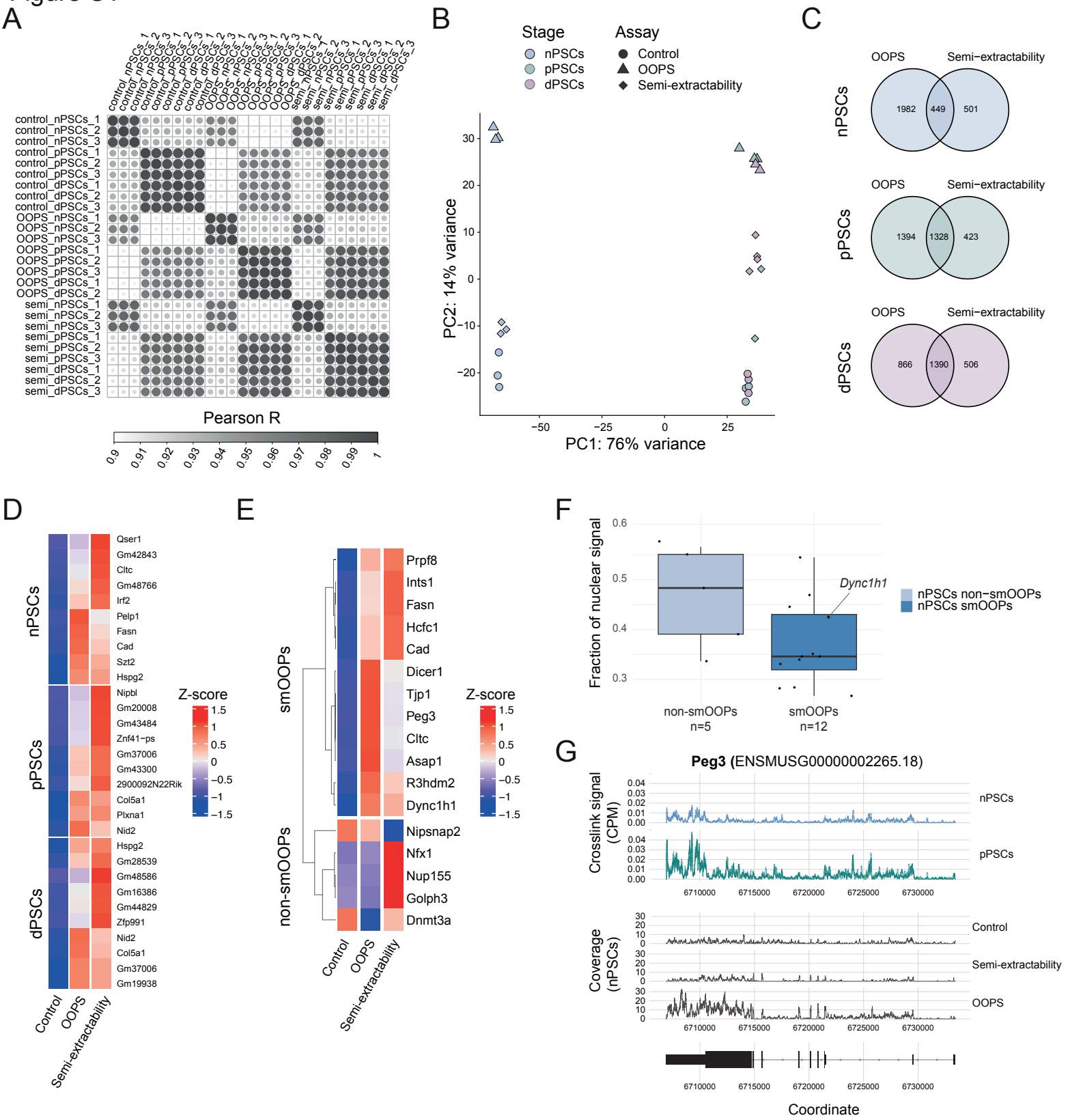

A
Pearson R
0.9
0.91
0.92
0.93
0.94
0.95
0.96
0.97
0.98
0.99
1
B
Stage
nPSCs
pPSCs
dPSCs
Assay
Control
OOPS
Semi-extractability
PC2: 14% variance
PC1: 76% variance
C
OOPS
Semi−extractability
nPSCs
1982
449
501
pPSCs
1394
1328
423
dPSCs
866
1390
506
D
nPSCs
pPSCs
dPSCs
Qser1
Gm42843
Cltc
Gm48766
Irf2
Pelp1
Fasn
Cad
Szt2
Hspg2
Nipbl
Gm20008
Gm43484
Znf41−ps
Gm37006
Gm43300
2900092N22Rik
Col5a1
Plxna1
Nid2
Hspg2
Gm28539
Gm48586
Gm16386
Gm44829
Zfp991
Nid2
Col5a1
Gm37006
Gm19938
Z-score
1.5
1
0.5
0
−0.5
−1
−1.5
Control
OOPS
Semi-extractability
E
smOOPs
non-smOOPs
Prpf8
Ints1
Fasn
Hcfc1
Cad
Dicer1
Tjp1
Peg3
Cltc
Asap1
R3hdm2
Dync1h1
Nipsnap2
Nfx1
Nup155
Golph3
Dnmt3a
Z-score
1.5
1
0.5
0
−0.5
−1
−1.5
Control
OOPS
Semi-extractability
F
Fraction of nuclear signal
Dync1h1
nPSCs non−smOOPs
nPSCs smOOPs
non-smOOPs
n=5
smOOPs
n=12
G
Peg3 (ENSMUSG00000002265.18)
Crosslink signal (CPM)
nPSCs
pPSCs
Coverage (nPSCs)
Control
Semi-extractability
OOPS
6710000
6715000
6720000
6725000
6730000
Coordinate

Figure S2

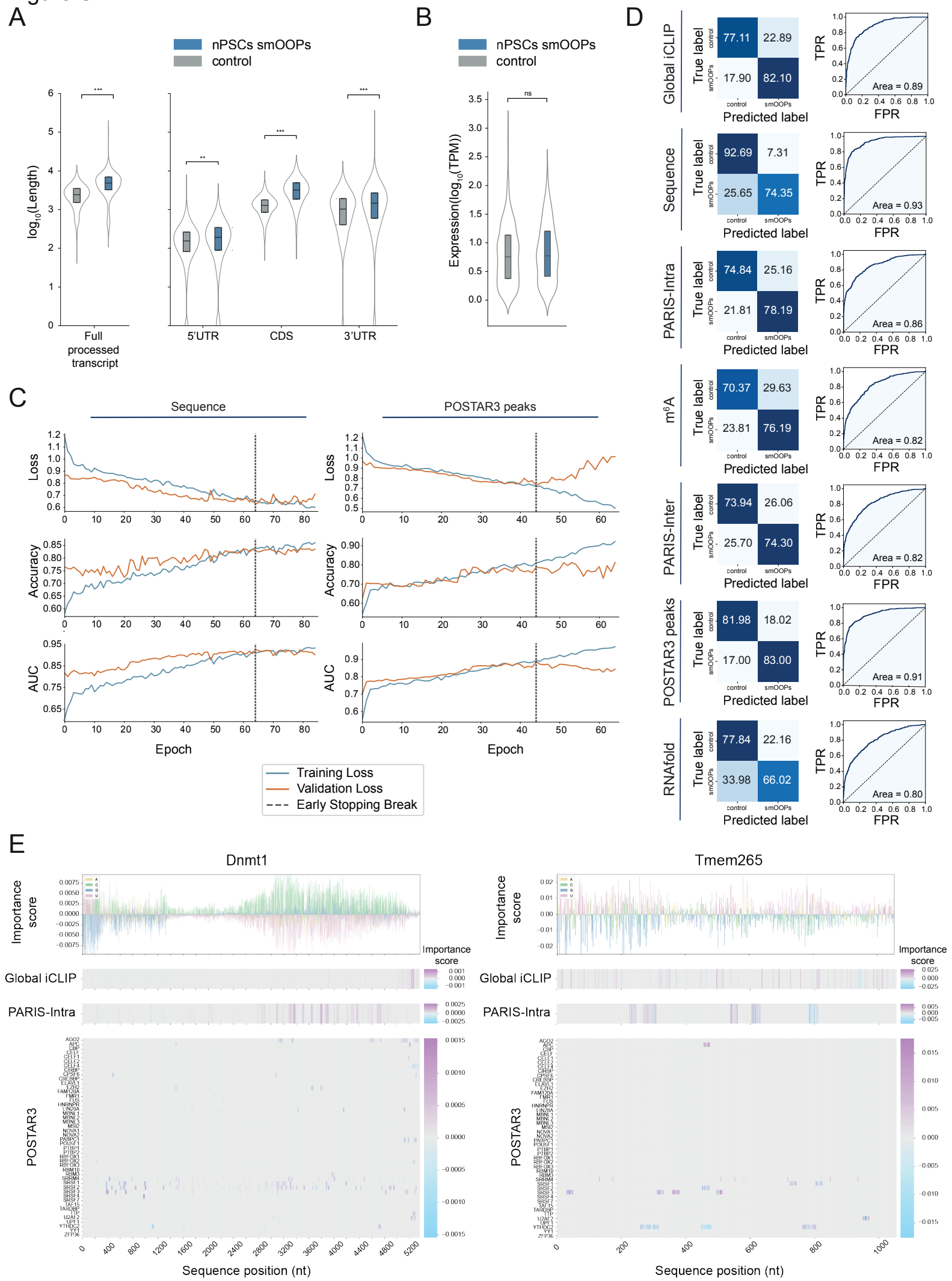

A
nPSCs smOOPs
control
log10(Length)
Full processed transcript
5'UTR
CDS
3'UTR
B
Expression(log10(TPM))
ns
C
Sequence
POSTAR3 peaks
Loss
Accuracy
AUC
Epoch
Training Loss
Validation Loss
Early Stopping Break
D
Global iCLIP
Sequence
PARIS-Intra
m6A
PARIS-Inter
POSTAR3 peaks
RNAfold
True label
Predicted label
control
smOOPs
TPR
FPR
Area = 0.89
Area = 0.93
Area = 0.86
Area = 0.82
Area = 0.82
Area = 0.91
Area = 0.80
E
Dnmt1
Tmem265
Importance score
Global iCLIP
PARIS-Intra
POSTAR3
Sequence position (nt)

Figure S3

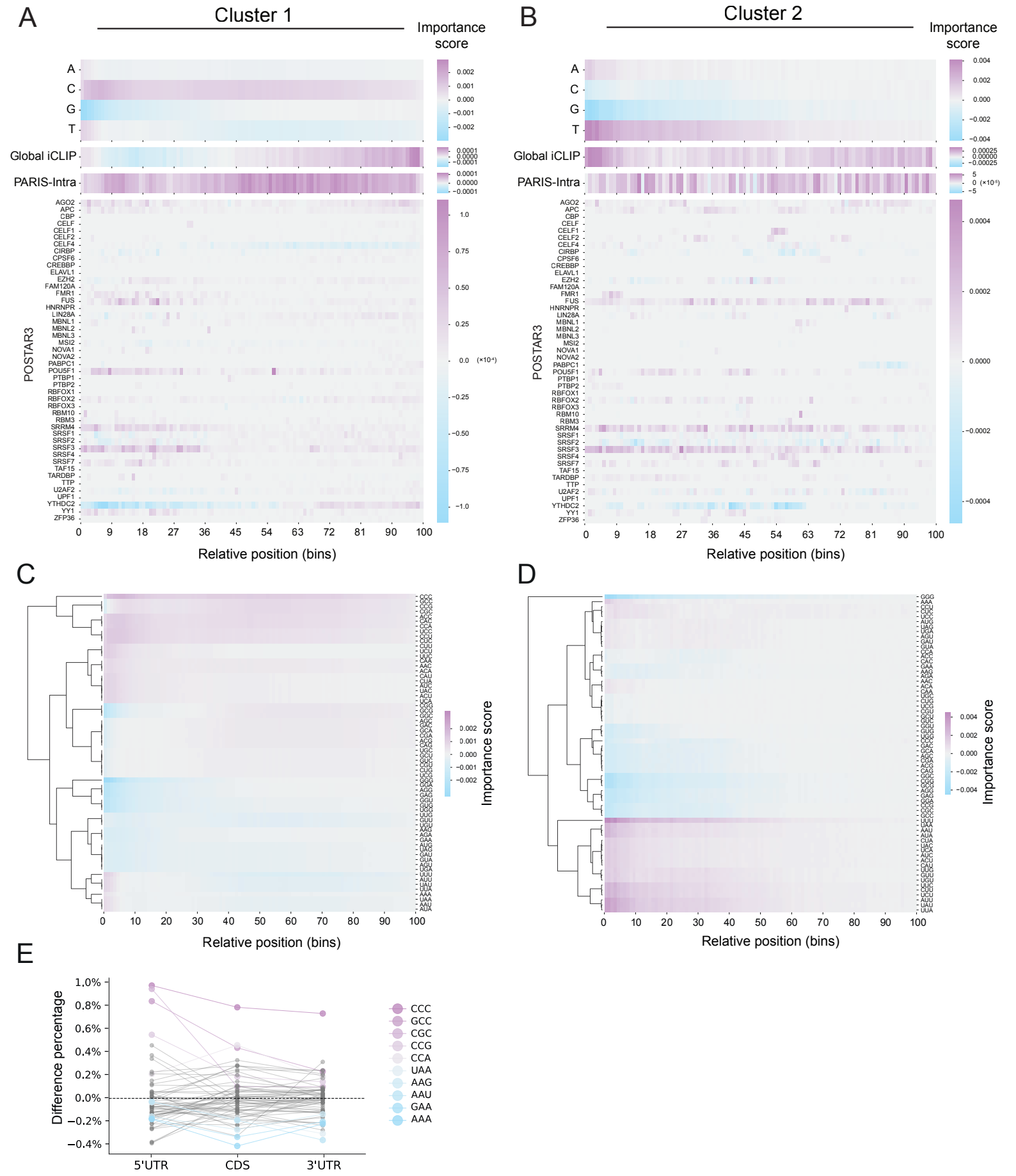
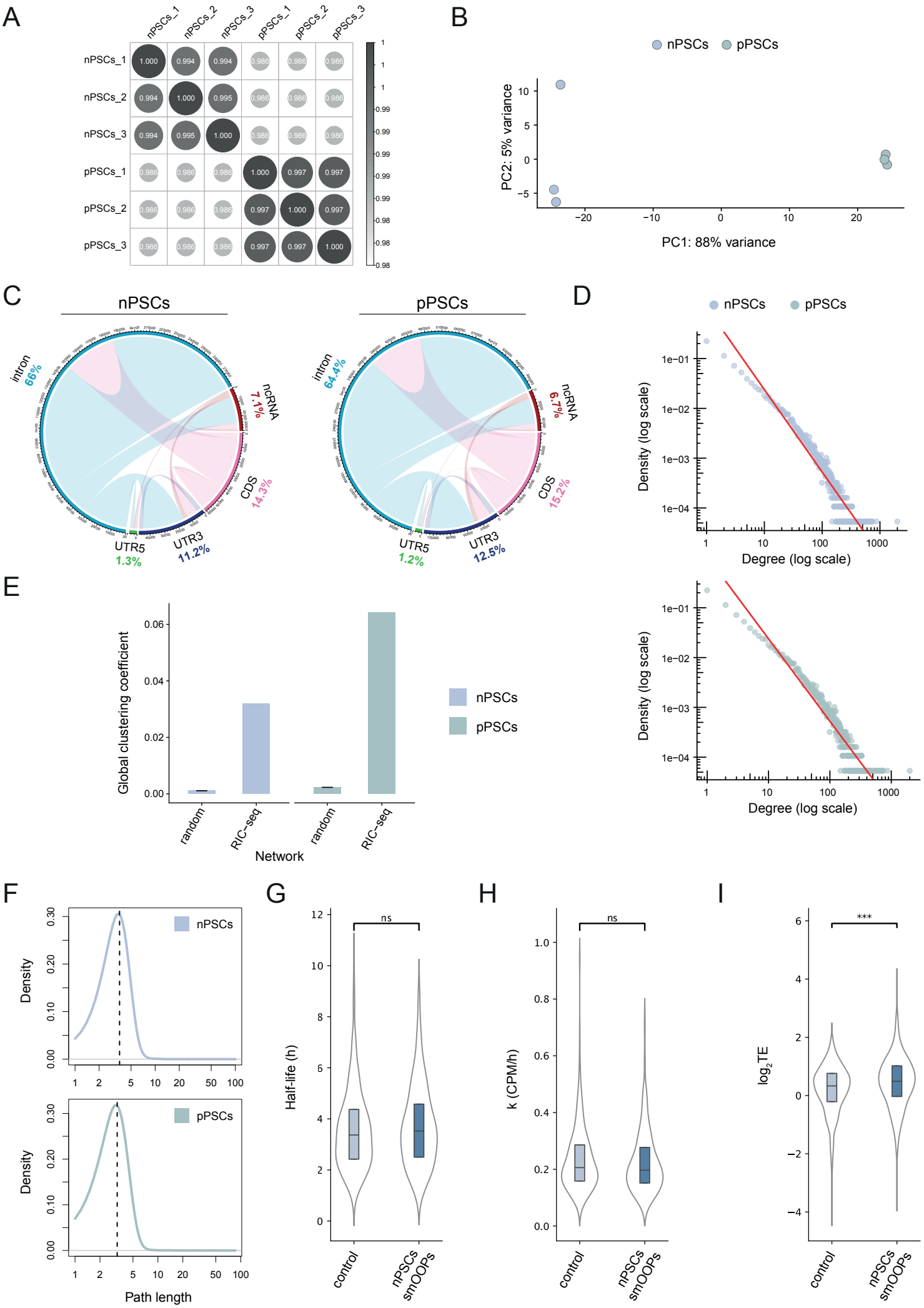

A
Cluster 1
Importance score
A
C
G
T
Global iCLIP
PARIS-Intra
POSTAR3
Relative position (bins)
B
Cluster 2
Importance score
C
Importance score
Relative position (bins)
D
Importance score
Relative position (bins)
E
Difference percentage
5'UTR
CDS
3'UTR
CCC
GCC
CGC
CCG
CCA
UAA
AAG
AAU
GAA
AAA

Figure S4


A
nPSCs_1
nPSCs_2
nPSCs_3
pPSCs_1
pPSCs_2
pPSCs_3
B
nPSCs
pPSCs
PC2: 5% variance
PC1: 88% variance
C
nPSCs
pPSCs
intron 66%
ncRNA 7.1%
CDS 14.3%
UTR3 11.2%
UTR5 1.3%
intron 64.4%
ncRNA 6.7%
CDS 15.2%
UTR3 12.5%
UTR5 1.2%
D
Density (log scale)
Degree (log scale)
E
Global clustering coefficient
random
RIC-seq
Network
F
Density
Path length
G
Half-life (h)
ns
control
nPSCs smOOPs
H
k (CPM/h)
I
log2TE
***