Université de Montréal

Caractérisation systématique des motifs de régulation en *cis* à l'échelle transcriptomique et liens avec la localisation des ARN

par

Louis Philip Benoit Bouvrette

Département de biochimie et médecine moléculaire
Faculté de médecine

Thèse présentée à la Faculté des études supérieures et postdoctorales
en vue de l'obtention du grade de
Philosophiæ Doctor (Ph.D.)
en Bio-informatique

avril 2020

Université de Montréal

Faculté des études supérieures et postdoctorales

Cette thèse intitulée

Caractérisation systématique des motifs de régulation en ***cis*** à l'échelle transcriptomique et liens avec la localisation des ARN

présentée par

Louis Philip Benoit Bouvrette

a été évaluée par un jury composé des personnes suivantes :

Adrian Serohijos
(président-rapporteur)

Eric Lécuyer
(directeur de recherche)

François Major
(co-directeur)

Samer Hussein
(membre du jury)

Michelle Scott
(examinateur externe)

Franz Lang
(représentant du doyen de la FESP)

Thèse acceptée le :
1 Septembre 2020

Résumé

La localisation subcellulaire de l'ARN permet un déploiement prompt et spatialement restreint autant des activités protéiques que des ARN noncodant. Le trafic d'ARN est dirigé par des éléments de séquences (sous-séquences primaires, structures secondaires), aussi appelés motifs de régulation, présents en *cis* à même la molécule d'ARN. Ces motifs sont reconnus par des protéines de liaisons aux ARN qui médient l'acheminement des transcrits vers des sites précis dans la cellule. Des études récentes, chez l'embryon de *Drosophile*, indiquent que la majorité des ARN ont une localisation subcellulaire asymétrique, suggérant l'existence d'un « code de localisation » complexe. Cependant, ceci peut représenter un exemple exceptionnel et la question demeurait, jusqu'ici, si une prévalence comparable de localisation d'ARN est observable chez des cellules standards développées en culture. De plus, des informations facilement disponibles à propos des caractéristiques de distribution topologique d'instances de motifs à travers des transcriptomes complets étaient jusqu'à présent manquantes.

Afin d'avoir un aperçu de l'étendue et des propriétés impliquées dans la localisation des ARN, nous avons soumis des cellules de *Drosophile* (D17) et de l'humain (HepG2) à un fractionnement biochimique afin d'isoler les fractions nucléaire, cytosolique, membranaire et insoluble. Nous avons ensuite séquencé en profondeur l'ARN extrait et analysé par spectrométrie de masse les protéines extraites de ces fractions. Nous avons nommé cette méthode CeFra-Seq. Par des analyses bio-informatiques, j'ai ensuite cartographié l'enrichissement de divers biotypes d'ARN (p. ex. ARN messager, ARN long non codant, ARN circulaire) et protéines au sein des fractions subcellulaires. Ceci a révélé que la distribution d'un large éventail d'espèces d'ARN codants et non codants est asymétrique. Une analyse des gènes orthologues entre mouche et humain a aussi démontré de fortes similitudes, suggérant que le processus de localisation est évolutivement conservé. De plus, j'ai observé des attributs (p. ex. la taille des transcrits) distincts parmi les populations d'ARN messagers spécifiques

à une fraction. Finalement, j'ai observé des corrélations et anti-corrélations spécifiques entre certains groupes d'ARN messagers et leurs protéines.

Pour permettre l'étude de la topologie de motifs et de leurs conservations, j'ai créé oRNAment, une base de données d'instances présumée de sites de liaison de protéines chez des ARN codants et non codants. À partir de données de motifs de liaison protéique par RNAcompete et par RNA Bind-n-Seq, j'ai développé un algorithme permettant l'identification rapide d'instances potentielles de ces motifs dans un transcriptome complet. J'ai pu ainsi cataloguer les instances de 453 motifs provenant de 223 protéines liant l'ARN pour 525 718 transcrits chez cinq espèces. Les résultats obtenus ont été validés en les comparant à des données publiques de eCLIP.

J'ai, par la suite, utilisé oRNAment pour analyser en détail les aspects topologiques des instances présumées de ces motifs et leurs conservations évolutives relatives. Ceci a permis de démontrer que la plupart des motifs sont distribués de façon similaire entre espèces. De plus, j'ai discerné des points communs entre les sous-groupes de protéines liant des biotypes distincts ou des régions d'ARN spécifiques. La présence de tels patrons, similaires ou non, entre espèces est susceptible de refléter l'importance de leurs fonctions. D'ailleurs, l'analyse plus détaillée du positionnement d'un motif entre régions transcriptomiques comparables chez les vertébrés suggère une conservation synténique de ceux-ci, à divers degrés, pour tous les biotypes d'ARN. La topologie régionale de certaines instances de motifs répétées apparaît aussi comme évolutivement conservée et peut être importante afin de permettre une liaison adéquate de la protéine. Finalement, les résultats compilés avec oRNAment ont permis de postuler sur un nouveau rôle potentiel pour l'ARN long non codant *HELLPAR* comme éponge de protéines liant l'ARN.

La caractérisation systématique d'ARN localisés et de motifs de régulation en *cis* présentée dans cette thèse démontre comment l'intégration d'information à l'échelle transcriptomique permet d'évaluer la prévalence de l'asymétrie, les caractéristiques distinctes et la conservation évolutive de collections d'ARN.

Mots-clés : Localisation de l'ARN, Régulation post-transcriptionnelle, Transcriptomique, ARN messagers, ARN non codants, Protéine liant l'ARN, Motifs de régulation en *cis*, Fractionnement subcellulaire, Séquençage en profondeur de l'ARN, Conservation évolutive.

Abstract

The subcellular localization of RNA allows a rapid and spatially restricted deployment of protein and noncoding RNA activities. The trafficking of RNA is directed by sequence elements (primary subsequences, secondary structures), also called regulatory motifs, present in *cis* within the RNA molecule. These motifs are recognized by RNA-binding proteins that mediate the transport of transcripts to specific sites in the cell. Recent studies in the *Drosophila* embryo indicate that the majority of RNAs display an asymmetric subcellular localization, suggesting the existence of a complex "localization code". However, this may represent an exceptional example and the question remained, until now, whether a comparable prevalence of RNA localization is observable in standard cells grown in culture. In addition, readily available information about the topological distribution of pattern instances across full transcriptomes has been hitherto lacking.

In order to have a broad overview of the extent and properties involved in RNA localization, we subjected *Drosophila* (D17) and human (HepG2) cells to biochemical fractionation to isolate the nuclear, cytosolic, membrane and insoluble fractions. We then performed deep sequencing on the extracted RNA and analyzed through mass spectrometry the proteins extracted from these fractions. We named this method CeFra-Seq. Through bioinformatics analyses, I then profiled the enrichment of various RNA biotypes (e.g. messenger RNA, long noncoding RNA, circular RNA) and proteins within the subcellular fractions. This revealed the high prevalence of asymmetric distribution of both coding and noncoding RNA species. An analysis of orthologous genes between fly and human has also shown strong similarities, suggesting that the localization process is evolutionarily conserved. In addition, I have observed distinct attributes (e.g. transcript size) among fraction-specific messenger RNA populations. Finally, I observed specific correlations and anti-correlations between defined groups of messenger RNAs and the proteins they encode.

To study motifs topology and their conservation, I created oRNAment, a database of putative RNA-binding protein binding sites instances in coding and noncoding RNAs. Using data from protein binding motifs assessed by RNAcompete and by RNA Bind-n-Seq experiments, I have developed an algorithm allowing their rapid identification in a complete transcriptome. I was able to catalog the instances of 453 motifs from 223 RNA-binding proteins for 525,718 transcripts in five species. The results obtained were validated by comparing them with public data from eCLIP.

I then used oRNAment to further analyze the topological aspects of these motifs' instances and their relative evolutionary conservation. This showed that most motifs are distributed in a similar fashion between species. In addition, I have detected commonalities between the subgroups of proteins linking preferentially distinct biotypes or specific RNA regions. The presence or absence of such pattern between species is likely a reflection of the importance of their functions. Moreover, a more precise analysis of the position of a motif among comparable transcriptomic regions in vertebrates suggests a syntenic conservation, to varying degrees, in all RNA biotypes. The regional topology of certain motifs as repeated instances also appears to be evolutionarily conserved and may be important in order to allow adequate binding of the protein. Finally, the results compiled with oRNAment allowed to postulate on a potential new role for the long noncoding RNA *HELLPAR* as an RNA-binding protein sponge.

The systematic characterization of RNA localization and *cis* regulatory motifs presented in this thesis demonstrates how the integration of information at a transcriptomic scale enables the assessment of the prevalence of asymmetry, the distinct characteristics and the evolutionary conservation of RNA clusters.

Keywords : RNA localization, Post-transcriptional regulation, Transcriptomics, Messenger RNA, Noncoding RNA, RNA binding protein, *Cis*-regulatory motifs, Subcellular fractionation, RNA-sequencing, Evolutionary conservation.

Table des matières

Liste des tableaux

Table des figures

Liste des abréviations

ATtRACT	A daTabase of experimentally validated RNA binding proteins and AssoCiated moTifs
CDE	Constitutive Decay Element
cDNA	Complementary DNA
CDS	Coding Sequence
CeFra-Seq	Biochemical cell fractionation strategy coupled with RNA sequencing
CircRNA	Circular ribonucleic acid
CISBP-RNA	Catalog of Inferred Sequence Binding Proteins of RNA
CLIP	Cross-Linking and ImmunoPrecipitation
CLIP-seq	Cross-Linking and ImmunoPrecipitation RNA sequencing
CRM	*Cis*-Regulatory Motifs
CT	Cycle threshold
DNA	Deoxyribonucleic acid
DBMS	DataBase management system
eCLIP	Enhanced Cross-Linking and ImmunoPrecipitation
ENCODE	Encyclopedia of DNA Element
ER	Endoplasmic reticulum
eRNA	Enhancer Ribonucleic acid
EST	Expressed Sequence Tag
FDR	False discovery rate
FISH	Fluorescence *In Situ* Hybridization
FPKM	Fragments Per Kilobase per Million mapped reads
GO	Gene Ontology

HIST-CLIP	High-throughput Sequencing of RNA isolated by Cross-Linking ImmunoPrecipitation
HMM	Hidden Markov Model
hnRNA	Heterogeneous nuclear ribonucleic acid
HOMER	Hypergeometric Optimization of Motif EnRichment
iCLIP	Individual-nucleotide resolution Cross-Linking and ImmunoPrecipitation
IGV	Integrated Genome Browser
ISH	*In Situ* Hybridization
IUPAC	International Union of Pure and Applied Chemistry
KB	Kilobase
LC-MS/MS	Liquid chromatography tandem mass spectrometry
LESMoN	Local Enrichment of Sequence Motifs in biological Networks
lncRNA	Long noncoding Ribonucleic acid
MEME	Multiple EM for Motif Elicitation
miscRNA	miscellaneous RNA
miRNA	Micro ribonucleic acid
mRNA	Messenger ribonucleic acid
mRNP	Messenger ribonucleoprotein
MSS	Matrix Similarity Score
Mt rRNA	Mitochondrial ribosomal ribonucleic acid
Mt tRNA	Mitochondrial transfer ribonucleic acid
ncRNA	Noncoding RNA
nt	Nucleotides
oRNAment	o RNA motifs enrichment in transcriptomes
PAR-CLIP	PhotoActivatable Ribonucleoside-enhanced Cross-Linking and ImmunoPrecipitation
PA	PolyA-enrichment
PCA	Principal Component Analyses
pCRMI	Predicted RBP *cis*-regulatory motif instances
PDB	Protein Data Bank
pFPKM	Percent FPKM

piRNA	Piwi-interacting Ribonucleic acid
PWM	Position weight matrix
RBNS	RNA Bind-n-Seq
RBP	RNA-binding proteins
RBPDB	RNA-Binding Protein DataBase
RBD	RNA Binding Domains
RD	Ribosomal RNA-depletion
RIP	RNA ImmunoPrecipitation
RIP-seq	RNA ImmunoPrecipitation and Sequencing
RNA	Ribonucleic acid
RNAi	RNA interference
RNP	Ribonucleoprotein
rRNA	Ribosomal ribonucleic acid
scaRNA	Small Cajal body-specific ribonucleic acid
SGDB	System de gestion de base de données
snoRNA	Small nucleolar ribonucleic acid
snRNA	Small nuclear
spliRNA	Splice sites ribonucleic acid
SSHMM	Sequence-structure Hidden Markov Model
t-SNE	t-Distributed Stochastic Neighbour Embedding
tmRNA	Transfer-Messenger ribonucleic acid
TPM	Transcripts Per Millions
tRNA	Transfer ribonucleic acid
tiRNA	Transcription Initiation ribonucleic acid
UTR	UnTranslated Region

Remerciements

J'aimerais en premier lieu remercier Eric Lécuyer pour ta supervision, ton soutien et ta confiance tout au long de mes études doctorales. Tu m'as donné l'opportunité de travailler sur des projets accrocheurs et m'as laissé la liberté nécessaire à mon épanouissement scientifique. Ta porte fut toujours ouverte pour nos discussions animées qui m'ont permis d'avoir de précieuses réflexions et inspirations sur de nombreux aspects de la recherche. Ceci me guidera certainement pour les années à venir.

Je remercie aussi mon co-directeur, François Major, nos rencontres furent espacées, mais enrichissantes et toujours appréciées.

Je remercie les membres de mon comité de thèse. Pascal Chartrand pour vos conseils et suggestions. Mathieu Blanchette pour tous tes conseils et ton appui. Tu m'as accueilli comme un membre de ton propre laboratoire, nos discussions continueront de m'éclairer longtemps.

Je remercie aussi tous les membres de l'équipe ENCODE/ENCORE, principalement Eric Van Nostrand, Michael O. Duff et Xintao Wei, vos enseignements bio-informatiques à mes débuts m'auront été utile chaque jour.

Je remercie Élaine Meunier pour son assistance administrative exemplaire au cours de toutes ces années.

J'aimerais aussi remercier tous les membres du laboratoire, passés et présents, avec une pensée particulière pour Fabio Lefebvre, mon grand ami et éternel co-auteur, Julie Bergalet, la pro de l'apéro, Xiaofeng "Andy" Wang, *friendliest tenacious worker*, et Samantha Bovaird, *the other circos plot maker*.

Merci à mes parents pour leur soutien même s'ils ne comprennent toujours pas ce que je fais.

Finalement, puisque je dois remercier Valérie. Merci.

Avant-propos

Structure de la thèse.

Cette thèse de doctorat fût rédigée selon les directives décrites dans le *Guide des mémoires et des thèses* de l'Université de Montréal et est sous forme hybride telle qu'autorisée par le département de biochimie et médecine moléculaires.

Comme chapitre de cette thèse, j'ai inclus les textes et figures de quatre manuscrits qui ont été publiés ou sont en préparation et pour lesquels je suis le premier auteur. L'introduction est formé de deux chapitres. Le chapitre 1 propose une introduction générale, en français, alors que le chapitre 2 (premier article) présente une revue de littérature publiée en anglais. Le corps de cette thèse, soit les chapitres 3 (deuxième article), 4 (troisième article) et 5 (quatrième article), est sous forme d'articles rédigées en anglais. Une discussion et conclusion générale, en français, sont proposées aux chapitres 6 et 7. Chacun des chapitres 2, 3, 4 et 5 contient une préface qui les introduit contextuellement et qui détaille la contribution de chaque auteur. De plus, chacun des chapitres a sa section de références propre.

Manuscrits non inclus dans cette thèse.

En plus des manuscrits inclus dans cette thèse, j'ai contribué, durant ma candidature, aux manuscrits publiés suivant :

(1) Eric L. Van Nostrand, Peter Freese, Gabriel A Pratt, Xiaofeng Wang, Xintao Wei, Rui Xiao, Steven M. Blue, Daniel Dominguez, Neal A.L. Cody, Sara Olson, Balaji Sundararaman, Lijun Zhan, Cassandra Bazile, Louis Philip Benoit Bouvrette et al. *A Large-Scale Binding and Functional Map of Human RNA Binding Protein.* **Nature**, 583 (2020), 711-719. https://doi.org/10.1038/s41586-020-2077-3.

(2) The ENCODE Project Consortium, Jill E. Moore, Michael J. Purcaro, Henry E. Pratt, et al. *Expanded encyclopaedias of DNA elements in the human and mouse genomes.* **Nature**, 583 (2020), 699-710. https://doi.org/10.1038/s41586-020-2493-4.

(3) The ENCODE Project Consortium, Michael P. Snyder, Thomas R. Gingeras, Jill E. Moore, et al. *Perspective on ENCODE.* **Nature**, 583 (2020), 693-698. https://doi.org/10.1038/s41586-020-2449-8.

(4) Sami Hassine*, Florence Bonnet-Magnaval*, Louis Philip Benoit Bouvrette, Bellastrid Doran, Mehdi Ghram, Mathieu Bouthillette, Eric Lécuyer, Luc DesGroseillers. *Staufen1 localizes to the mitotic spindle and controls the localization of RNA populations to the spindle.* **Journal of Cell Science**, 133 (2020), jcs247155. https://doi.org/10.1242/jcs.247155.

(5) Julie Bergalet*, Dhara Patel*, Félix Legendre, Catherine Lapointe, Louis Philip Benoit Bouvrette, Ashley Chin, Mathieu Blanchette, Eunjeong Kwon, Eric Lécuyer. *Inter-dependent Centrosomal Co-localization of the cen and ik2 cis-Natural Antisense mRNAs in Drosophila.* **Cell Report**, 30 (2020), 3339–3352. https://doi.org/10.1016/j.celrep.2020.02.047.

(6) Jian Kong, Hong Han, Julie Bergalet, Louis Philip Benoit Bouvrette, Greco Hernández, Nam-Sung Moon, Hojatollah Vali, Eric Lécuyer, Paul Lasko. *A ribosomal protein S5 isoform is essential for oogenesis and interacts with distinct RNAs in Drosophila melanogaster*. **Scientific reports**, 9 (2019), 13779. https://doi.org/10.1038/s41598-019-50357-z.

(7) Fabio Alexis Lefebvre*, Neal A.L. Cody*, Louis Philip Benoit Bouvrette, Julie Bergalet, Xiaofeng Wang, Eric Lécuyer. *CeFra-seq : Systematic mapping of RNA subcellular distribution properties through cell fractionation coupled to deep-sequencing*. **Methods**, 126 (2017), 138-148. https://doi.org/10.1016/j.ymeth.2017.05.017.

(8) Fabio Alexis Lefebvre, Louis Philip Benoit Bouvrette, Julie Bergalet, Eric Lécuyer. *Data for the generation of RNA spatiotemporal distributions and interpretation of Chk1 and SLBP protein depletion phenotypes during Drosophila embryogenesis*. **Data in Brief**, 13 (2017), 28-31. https://doi.org/10.1016/j.dib.2017.05.008.

(9) Fabio Alexis Lefebvre, Louis Philip Benoit Bouvrette, Julie Bergalet, Eric Lécuyer. *Biochemical Fractionation of Time-Resolved Drosophila Embryos Reveals Similar Transcriptomic Alterations in Replication Checkpoint and Histone mRNA Processing Mutants*. **Journal of Molecular Biology**, 429 (2017), 3264–3279. http://dx.doi.org/10.1016/j.jmb.2017.01.022.

(10) Fabio Alexis Lefebvre, Louis Philip Benoit Bouvrette, Lilyanne Perras, Alexis Blanchet-Cohen, Delphine Garnier, Janusz Rak, Eric Lécuyer. *Comparative transcriptomic analysis of human and Drosophila extracellular vesicles*. **Scientific Reports**, 6 (2016), 27680. https://doi.org/10.1038/srep27680.

* Co-premier

Chapitre 1

Introduction générale

1.1. Historique de la découverte de l'ARN

L'ARN joue un rôle central non seulement dans une panoplie de processus et mécanismes cellulaires, mais aussi dans l'ontogénie et même l'évolution du vivant. Elle tiendrait plusieurs de ses caractéristiques depuis aussi tôt que l'époque du monde ARN, cette période hypothétique lors de laquelle les formes de vie primitives utilisaient l'ARN autant comme base de l'information génétique que de l'activité enzymatique [55, 149]. Bien que les cellules modernes aient changé de façon significative depuis cette période, l'ARN maintient un rôle central dans la biologie cellulaire.

Friedrich Miescher est considéré comme le premier scientifique, qui, en 1869, isola et caractérisa des composés faibles en sulfure et riches en phosphate dans des noyaux de cellules, et qu'il nomma nucléine [42, 161, 160]. Maintenant défini comme acides nucléiques, ils forment la base de l'ARN et de l'ADN [42, 161, 160]. Il aura fallu attendre au début des années 1940 pour que des analyses démontrent que l'ARN et l'ADN diffèrent en leur composition glucidique. L'ADN étant formé d'un désoxyribose et l'ARN d'un ribose et de bases azotées, la thymine chez l'ADN étant remplacée par l'uracile chez l'ARN [42]. Oswald Avery, en 1944, fut le premier à proposer que l'ADN forme le support de l'information génétique [5]. George Beadle et Edward Tatum furent parmi les premiers à développer l'idée que l'ARN fonctionne comme intermédiaire entre l'ADN et les protéines [14]. Dans les années 1950-1960, l'utilisation de techniques telles que la cristallographie par rayons X, la chromatographie et les techniques de centrifugation permettant d'isoler physiquement des composés biologiques et d'analyser leurs structures moléculaires fut naturellement regroupées dans une nouvelle discipline nommée biologie moléculaire [149]. En 1953, les travaux expérimentaux de Rosalind Franklin ont mené James Watson et Francis Crick à publier la structure de l'ADN et, en 1958, Francis Crick proposa le « dogme central de la biologie moléculaire » (Figure 1.1) [149, 194]. Celui-ci postule que l'information génétique est directionnelle et que l'ARN agit comme un intermédiaire entre l'ADN et les protéines [149]. En 1958, Schweet, Lamform et Allen produisirent les premières évidences expérimentales du rôle des ARN messagers qui fut ensuite établi, en 1961, par Brenner, Jacob et Meselson [23, 176]. En 1960, Francois Jacob et Jacques Monod furent les premiers à définir l'ARN messager (mRNA) suivant leurs analyses d'enzymes inductibles chez *E. coli* [23, 71, 88].

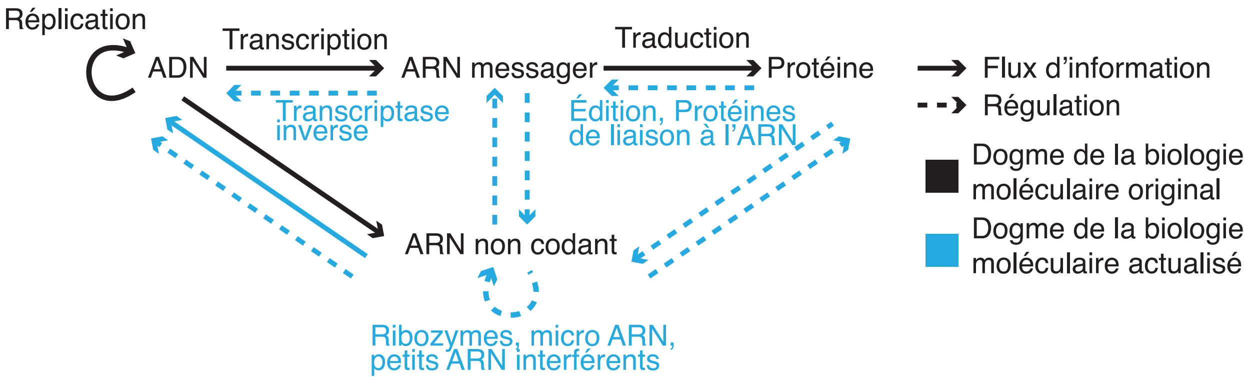

Réplication
ADN
Transcription
ARN messager
Traduction
Protéine
Transcriptase inverse
Édition, Protéines de liaison à l'ARN
ARN non codant
Ribozymes, micro ARN, petits ARN interférents
Flux d'information
Régulation
Dogme de la biologie moléculaire original
Dogme de la biologie moléculaire actualisé

FIGURE 1.1. Dogme central de la biologie moléculaire. Représentations schématiques du dogme central de la biologie moléculaire tel que postulé par Francis Crick (lignes noires) et actualisé (ligne cyan) suite aux décennies d'efforts de caractérisation en biologie des ARN.

Comme le décrivent les sections suivantes, les connaissances sur l'ARN et le dogme de la biologie moléculaire ne cessent d'être modifiés et nuancés afin d'y intégrer toutes les connaissances, sans cesse grandissantes, à propos des rôles et caractéristiques variés joués par différentes classes d'ARN et de molécules afférentes au sein de la cellule.

1.2. Transcriptome noncodant

Un des changements de paradigme le plus important en ce qui a trait aux connaissances liées aux ARN est assurément l'étendue des niveaux de transcription et la diversité fonctionnelle associée aux ARN non codants. En effet, les techniques de puce à ADN et de séquençage à haut débit ont démontré que la majorité des régions intergéniques et introniques sont différentiellement transcrites [38, 39]. Alors qu'environ 1,2 % des bases du génome humain encodent pour des ARN messagers, les observations les plus récentes indiquent qu'au moins

80 % du génome serait transcrit en ARN de biotypes variés, bien que leurs rôles fonctionnels restent controversés (Figure 1.2) [149, 38, 39, 33, 102, 70].

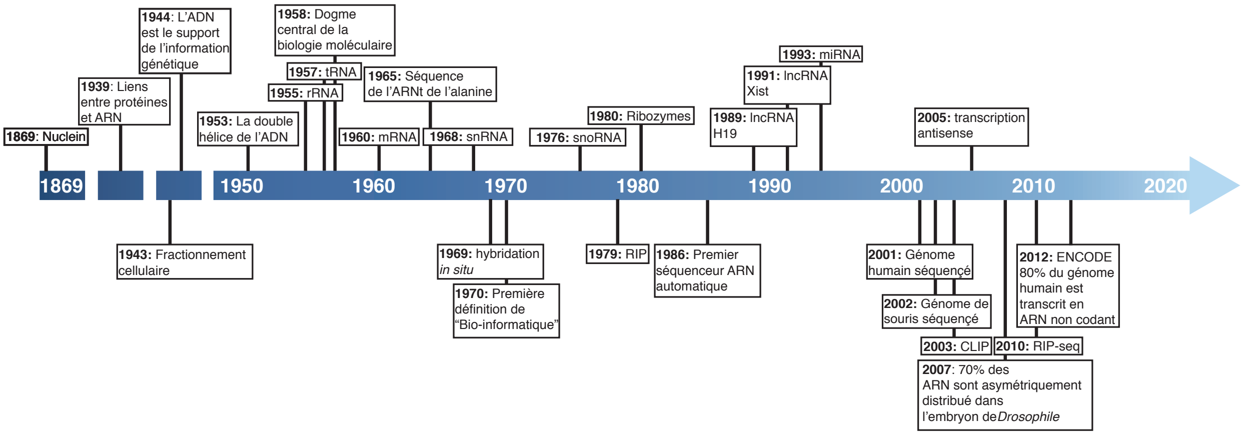

1944: L'ADN est le support de l'information génétique
1958: Dogme central de la biologie moléculaire
1993: miRNA
1939: Liens entre protéines et ARN
1957: tRNA
1955: rRNA
1965: Séquence de l'ARNt de l'alanine
1991: lncRNA Xist
1980: Ribozymes
1989: lncRNA H19
2005: transcription antisense
1869: Nuclein
1953: La double hélice de l'ADN
1960: mRNA
1968: snRNA
1976: snoRNA
1869
1950
1960
1970
1980
1990
2000
2010
2020
1943: Fractionnement cellulaire
1969: hybridation *in situ*
1970: Première définition de "Bio-informatique"
1979: RIP
1986: Premier séquenceur ARN automatique
2001: Génome humain séquencé
2002: Génome de souris séquencé
2003: CLIP
2010: RIP-seq
2012: ENCODE 80% du génome humain est transcrit en ARN non codant
2007: 70% des ARN sont asymétriquement distribué dans l'embryon de *Drosophile*

FIGURE 1.2. Chronologie des découvertes principales en biologie des ARN. Historique des événements et progrès majeurs des recherches en biologie des ARN mettant l'emphase sur les ARN non codants (haut) et les méthodes et technologies (bas).

Durant les années 1950, plusieurs recherches cruciales furent effectuées pour élucider les voies mécanistiques entre les gènes et les protéines. Parmi ceux-ci, les travaux de George Palade et Paul Zamecnik, publiés en 1955 et 1958 respectivement, ont mené aux découvertes des premiers ARN non codants [156, 81]. George Palade, d'un côté, s'intéressa aux ribosomes et aux liens avec les ARN ribosomaux [156]. Paul Zamecnik *et coll.* quant à eux, ont originalement décrit l' « ARN soluble », maintenant connu sous le nom ARN de transferts (tRNA) [81, 101]. Il s'agit d'une molécule « adaptatrice », d'une taille d'environ 76 nucléotides, servant d'intermédiaire à la traduction de l'information à partir de l'ARN messager [81, 101].

Par ailleurs, la découverte des ARN nucléaires hétérogènes (hnRNAs) et les observations que ceux-ci forment une population complexe enrichie principalement dans le noyau ont mené aux hypothèses que les ARN peuvent avoir des rôles supplémentaires [193]. Ces observations ont conduit Roy Britten *et coll.*, dès 1969, à raisonner que les ARN pourraient être fonctionnels et agir en tant que régulateurs et coordonnateurs, sans pour autant produire de protéines [43, 24, 25, 85].

À la suite des études décrivant les tRNAs et les rRNAs, des techniques de fractionnement cellulaire ont permis la découverte d'un grand nombre de petits ARN provenant du noyau des cellules [195, 52]. Parmi les ARN ainsi identifiés, les petits ARN nucléaires (snRNAs), d'une taille d'environ 150 nucléotides, tiennent un rôle dans l'épissage d'ARN messagers en formant de larges complexes ribonucléoprotéides nommés spliceosome [27, 192]. Les petits ARN nucléolaires (snoRNAs) d'une taille allant de 60 à 170 nucléotides, quant à eux, ont principalement un rôle dans la maturation des ARN ribosomaux. Des études ont même démontré qu'ils ciblent aussi les ARN messagers et sont différentiellement exprimés dans certains tissus, suggérant des rôles de régulations supplémentaires [142, 100, 29, 171, 6]. Des snoRNAs particuliers ont aussi été détectés dans des structures subnucléaires définies comme le corps de Cajal et ont donc été nommés petit ARN spécifique au corps de Cajal (scaRNAs) [89]. Par contre, leurs rôles restent encore mal définis [149].

En 1976, Heinz Sanger démontra que les viroïdes, particule virale élémentaire, composés uniquement d'un ARN circulaire sans capside, sont des molécules d'ARN liées aux extrémités de façon covalente pour former une boucle et ainsi il proposa le concept d'ARN circulaire (circRNA) [6]. Par contre, l'observation d'ARN circulaire chez les eucaryotes est venue beaucoup plus tard, soit en 2012 [173]. De nombreuses analyses subséquentes montrèrent qu'il existe au moins cinq classes d'ARN circulaire, soit les ARN circulaires génomiques (« genomic circular RNA »), les ARN circulaires intermédiaires au processus (« processing intermediate RNA »), les ARN non codants constitutifs circulaires (« circular housekeeping noncoding RNA »), les introns circularisés (« circular intron RNA »), et les exons épissés circularisés (« circular spliced-exon RNA ») [89, 173, 113, 73, 30, 31, 125]. Plusieurs études récentes tendent à démontrer qu'ils tiendraient un rôle d'éponge pour les micros ARN, bien que des mécanismes diversifiés leur sont pressentis [89, 169, 104].

Bien que plusieurs ARN non codants aient été détectés, un grand nombre de chercheurs doutaient encore de leurs rôles fonctionnels et les considéraient comme de simples résidus intermédiaires instables. En 1980, Tomas Cech et Sidney Altman découvrirent des ARN capables d'agir comme catalyseurs de réactions biochimiques en identifiant un intron capable d'effectuer son propre épissage, ils venaient de distinguer les ribozymes [113, 73, 30, 31].

Au début des années 1990, un nombre grandissant de chercheurs s'intéressèrent, principalement en effectuant des expériences de co-expressions transgéniques, à l'inhibition protéique médiée par les ARN et au phénomène de co-suppression. C'est ainsi qu'en 1993, Victor Ambros *et coll.* ont observé de petits ARN régulateurs d'une taille d'environ 22 nucléotides, ensuite définis comme des micros ARN (miRNAs) [125, 169]. Ils démontrèrent que ces miRNAs agissent comme inhibiteurs de la traduction et accélèrent la dégradation des ARN messagers en se liant typiquement par complémentarité de bases à leur région non transcrite en 3' [104]. Il fut ensuite établi que la majorité des miRNA ciblent des ARN messagers, que ceux-ci peuvent posséder plusieurs sites cibles de miRNAs et qu'un miRNA peut cibler plusieurs ARNm différents [98, 131, 59, 175]. Les analyses mécanistiques des miRNAs ont démontré qu'ils proviennent d'un clivage d'un ARN double brin endogène [9]. Par la suite, il fut reconnu que des protéines clés, telles que *Drosha*, *Dicer* et *Argonaute* sont impliquées dans la biogenèse des miRNAs [126, 18, 50]. De façon intéressante, une sous-classe des protéines *Argonautes*, nommées *PIWI*, sont requises pour la différenciation des cellules souches et germinales et sont associées à une classe distincte de petits ARN, d'une taille de 26 à 30 nucléotides, nommée ARN interagissant avec *PIWI* (piRNA) [135, 40, 115, 108, 187, 69, 116].

Vers la fin des années 1980, des travaux effectués par Vassilis Pachnis ont permis d'identifier, malgré le fait qu'il fût originalement interprété comme un ARN messager, le premier long ARN non codant (lncRNA), nommé *H19* [155, 8, 22]. Parallèlement, les travaux de Mary Lyon sur l'inactivation du chromosome X chez la souris ont permis à Carolyn Brown d'identifier le lncRNA *XIST* [139, 26]. Les lncRNAs furent finalement définis comme des ARN non codants étant plus longs que 200 nucléotides, taille principalement établie comme seuil de fractionnement biochimique pratique et permettant une différenciation simple des autres petits ARN non codants [94, 65]. Les lncRNAs sont par ailleurs distingués en plusieurs catégories selon qu'ils sont introniques, antisenses ou intergéniques. Par contre, bien qu'il puisse être utile de différencier leurs origines, leurs différences fonctionnelles demeurent incertaines [149, 103, 106]. De plus, les études récentes ont démontré que les lncRNAs ont des séquences faiblement conservées évolutivement, bien que leurs structures montrent des évidences de conservation et qu'ils ont une plus grande spécificité, autant cellulaire que d'expression, en moyenne, que les ARN messagers et les protéines qu'elles encodent [179, 99, 47, 28].

La combinaison de ces efforts de caractérisation a permis de définir des fonctions aux petits et longs ARN non codants et ainsi révolutionner les théories sur les gènes ne codant pas pour des protéines et leurs propriétés générales dans les cellules. Plus récemment, l'avènement de l'ère génomique et du séquençage ARN à haut débit ont révélé l'existence de plusieurs autres classes de petits ARN non codants qui peuvent être spécifiques à certaines branches du vivant, par exemple observés uniquement chez les animaux et non chez les plantes. Parmi ceux-ci, il y a les ARN d'initiation de transcription (tiRNAs) les ARN de site d'épissage (spliRNAs), les ARN transfert-messager (tmRNAs) et les « enhancer » ARN (eRNAs) [181, 182, 165, 203, 44]. Ceci démontre qu'il reste assurément beaucoup à apprendre sur la vaste diversité d'ARN non codants ayant des rôles fonctionnels importants au sein des cellules.

1.3. Principes, rôles biologiques et mécanismes de la localisation des ARN

En 1983, William Jeffrey *et coll.* identifièrent que la distribution subcellulaire des ARNm de la *β-actine* chez les embryons et les œufs d'ascidies, un animal marin, était asymétrique [96]. Cette découverte importante fut une étape majeure dans l'étude de la biologie d'ARN, car, jusque-là, les théories acceptées reposaient sur l'idée que la compartimentation des protéines était strictement due à un transport post-traductionnel [172]. Par conséquent, une nouvelle théorie dictant que les ARN qui codent pour des protéines peuvent être activement transportés vers une région subcellulaire précise avant d'être localement traduits émergea. Bien que la proportion d'ARN ainsi distribués de même que l'étendue de ce mécanisme à d'autres espèces furent originalement controversées, des travaux de caractérisation effectués dans les dernières décennies démontrèrent plusieurs exemples d'ARNm localisés dans une panoplie d'espèces. En effet, grâce à des techniques avancées en microscopie, en génomique et en bio-informatique, il fut établi que des milliers d'ARN sont régulés au niveau de leurs trafics subcellulaires chez les algues [177], les ascidies [162], les fongus [199], les insectes [121, 68], les invertébrés [79], les levures [80], les plantes [41], les procaryotes [105], les amphibiens, les poissons et les mammifères [143, 191, 19, 74, 146, 123, 201]. La localisation des ARN est maintenant acceptée comme une étape post-transcriptionnelle de la régulation génique et est un facteur important et prévalent dans la modulation des fonctions autant des ARN codants que non codants [11].

Selon les principes fondamentaux de la localisation des ARN, ceux-ci octroient de nombreux bénéfices à la cellule et impliquent plusieurs mécanismes (Figure 1.3) [16, 141, 21, 36, 91]. Puisqu'un seul ARNm peut être traduit simultanément par plusieurs ribosomes, une large quantité de protéines peut être créée par quelques ARN, réduisant considérablement l'énergie nécessaire pour obtenir la même quantité de protéines à une région subcellulaire précise [196, 190]. Par ailleurs, les ribosomes semblent être distribués dans la majorité des régions subcellulaires et des études ont démontré que la distribution des ARN montre généralement une corrélation avec les patrons de distribution des protéines qu'elles codent [123, 122, 15]. De plus, puisque les ARNm qui encodent pour des protéines possédant des rôles fonctionnels similaires partagent fréquemment des patrons de localisation similaire, il est présumé que leur transport et traduction collectivement synchronisés faciliteraient leurs assemblages en complexes protéiques [148, 10, 78, 152]. Cela étant, la localisation ciblée de la traduction des ARNm, en plus d'offrir un mécanisme précis pour contrôler la distribution et l'assemblage des complexes protéiques, permet, à l'inverse, d'écarter certaines protéines de régions où elles pourraient avoir un effet néfaste [16, 141, 21, 57, 2, 180]. D'autre part, elle permet d'en réguler la temporalité, où la traduction localisée peut être enclenchée par des stimuli externes. Par exemple, dans les neurones, la traduction d'ARNm préalablement acheminés à une région donnée peut être initiée en réponse à une stimulation exogène [143, 146]. Finalement, la localisation des ARN, en particulier des ARN non codants, a été impliquée dans des fonctions indépendantes de la traduction, par exemple en agissant comme échafaud dans l'assemblage de structures macromoléculaires [110, 97].

Des études ont démontré que la localisation des ARN semble pouvoir être effectuée selon trois mécanismes soit : 1) la diffusion générale couplée à un arrimage local ; 2) la protection contre la dégradation ; et 3) le transport direct à l'aide du réseau cytosquelettique (Figure 1.3) [16, 141, 21]. Notamment, il a été démontré que l'ARN localisé est généralement dans un état de répression de la traduction pendant son déplacement, un mécanisme nécessaire afin d'assurer un appariement étroit entre les patrons de distribution des ARNm et des protéines [95, 138]. De plus, un ARN peut exploiter plusieurs mécanismes afin de se rendre au bon endroit.

À l'origine de ces mécanismes de localisations, les ARN possèdent à même leurs séquences ou leurs structures des éléments, ou motifs, de régulation en *cis* qui sont liés par des facteurs, telles des protéines liant les ARN (« RNA binding proteins », RBP), en *trans*, ce qui impulse leur ciblage dirigé. Une description plus détaillée de ces motifs est proposée au chapitre 2.1.

Le premier mécanisme est la diffusion générale couplée à un arrimage local. Celui-ci est bien exemplifié par *nanos*, un ARNm localisé au pôle postérieur des oocytes de *Drosophile* [62]. Pendant les dernières étapes de l'oogenèse, *nanos* suit les forts courants cytoplasmiques

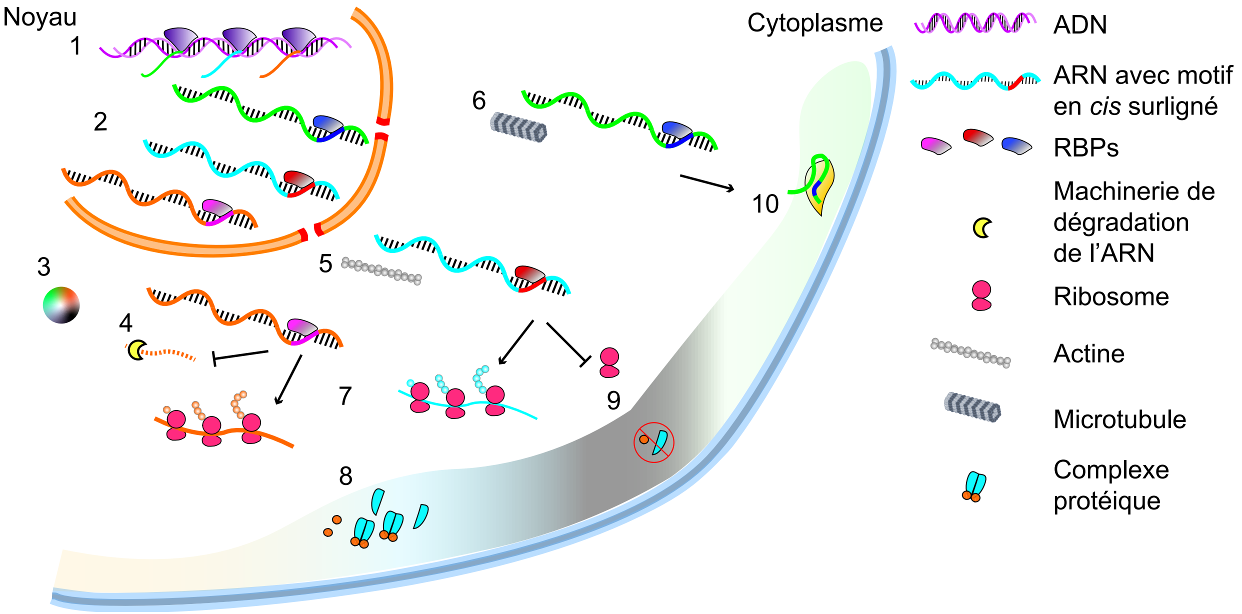


Figure 1.3. Mécanismes et bénéfices fonctionnels de la localisation des ARN. Schématisation de la localisation subcellulaire des ARN. Suivant la transcription (1), les ARN sont liés, dans le noyau, par des RBP, formant des ribonucléoprotéines (RNP) et sont exportés via des pores nucléaires (2). Dans le cytoplasme, les RNP transitent dans un état de répression de la traduction selon trois mécanismes possibles soit : (3) la diffusion générale couplée à un arrimage local ; (4) la protection contre la dégradation ; et (5, 6) le transport direct à l'aide du réseau cytosquelettique vers une région subcellulaire cible où ils sont, généralement, intégrés à des polysomes pour être traduits en protéines (7). En parallèle, la traduction ciblée confère des bénéfices importants dont : (8) faciliter la formation de complexes protéiques et (9) prévenir la localisation anormale de protéines à des régions où elles pourraient avoir un effet néfaste. (10) La localisation des ARN permet aussi des fonctions indépendantes de la traduction en agissant comme échafaud dans l'assemblage de structures macromoléculaires.

de l'oocyte et peut ainsi être rapidement mis en contact et ancré au cytosquelette à l'aide de molécule d'*actine* [62]. Évidemment, ce type de localisation n'est pas des plus efficients énergétiquement et seulement 4 % de *nanos* sont protégés d'une déadénylation et d'une dégradation par *Smaug* [17, 198]. La localisation de *nanos* au pôle postérieur implique donc un deuxième mécanisme, soit une protection contre la dégradation qui consiste en une dénaturation complète de l'ARN qui n'est pas à un endroit précis [16, 141, 62, 198]. Troisièmement, le mécanisme de localisation des ARN qui semble le plus prédominant est le transport direct à l'aide du réseau cytosquelettique [141, 21, 15]. Au moyen de celui-ci, lorsqu'un ARN atteint le cytoplasme sous forme de particule ribonucléique messager (mRNP), il peut subir un remodelage par l'ajout ou le retranchement d'éléments en *trans*, tels des RBP ou microARN, pour former une granule [141, 21, 15]. Ces granules peuvent s'associer à des protéines motrices et sont transportées à travers le réseau cytosquelettique vers leurs destinations [141, 21, 15]. Par exemple, la localisation de l'ARNm *ASH1* au bourgeon chez *S. cerevisiae* s'effectue grâce à la combinaison de quatre motifs de localisation qui agissent en synergie afin de lier l'ARN à plusieurs RBP, tels que *She2p*, *Loc1P*, *Puf6p*, et *Khd1p*, pour finalement se lier à *Myo4p*, une protéine motrice ayant un rôle dans le transport cellulaire et achemine cette mRNP de façon directe à destination par des filaments d'*actine* [93, 178, 150, 137, 72, 53, 87, 20, 58, 92, 136, 32, 112, 157]. Lorsqu'arrivé à destination, le *ASH1* est ancré par les protéines *She5p* et *Bud6p* et les protéines *Puf6p* et *Kdhlp* sont libérées, permettant ainsi à *ASH1* d'être traduit [13, 45, 158].

Ces découvertes, effectuées au cours des dernières décennies démontrent bien que la localisation des ARN couplés à une régulation de la traduction s'est révélée comme un mécanisme commun et fondamental pour une grande variété d'ARN, et ce, dans une panoplie de types cellulaires. Alors qu'originalement interprétée comme un simple moyen de restreindre la synthèse protéique chez des cellules polarisées, son importance est dorénavant établie pour réguler globalement l'expression génique de façon spatio-temporelle.

1.4. Méthodologies expérimentales et technologies appliquées à l'étude de la localisation des ARN

La biologie des ARN reflète la symbiose d'une pléthore de systèmes complexes et les chercheurs ont développé une grande variété de techniques permettant de documenter, entre autres, leurs mécanismes, prévalences et localisations de façon robuste. Parmi ces techniques, certaines à haut débit, de génomiques, de transcriptomiques, d'imageries et de bioinformatiques ont révélé qu'une panoplie d'ARN subissent une régulation stricte de leur trafic subcellulaire faisant passer le répertoire de transcrits localisés d'une centaine à plusieurs milliers (Figure 1.2) [143, 19, 146, 123, 122, 15, 56, 128, 90, 86, 60, 145].

1.4.1. Imageries

En 1969, Mary Lou Pardue *et coll.* publièrent une technique d'imagerie pour détecter la localisation subcellulaire d'acide nucléique, l'hybridation *in situ* (ISH) [159]. Cette méthode utilisait à l'origine des sondes radioactives, mais a été développée pour utiliser une combinaison de streptavidines et de biotines, ou la digoxigénine [123, 12]. Utilisée en conjonction avec des fluorophores, elle est nommée hybridation *in situ* en fluorescence (FISH). Plusieurs études utilisant ces techniques ont été effectuées dans une variété de types cellulaires et organismes, ce qui a permis de caractériser la localisation subcellulaire de plusieurs ARN, autant codants que non codants [123, 54, 140, 183, 109, 144, 129]. Par exemple, Eric Lécuyer s'est servi d'une optimisation à grande échelle d'une technique de FISH pour établir l'asymétrie de 3370 gènes de *Drosophile* durant les étapes précoces de l'embryogenèse et conclure qu'environ 70 % d'entre eux ont un patron de localisation asymétrique [123, 122]. Cette étude est la base d'une des hypothèses générales adressées dans cette thèse, soit que la localisation des ARN est prévalante dans des cellules standards humaines et de *Drosophiles*.

1.4.2. Fractionnement cellulaire biochimique

Une autre méthode importante pour caractériser les populations de transcrits est basée sur le fractionnement cellulaire, originalement développé par Albert Claude dans les années 1940 [164, 34, 35]. Cette technique repose sur le bris physique des cellules, autrefois effectué avec un mortier et un pilon, qui font ensuite l'objet de centrifugations différentielles, séparant ainsi les composés selon leurs densités, définies comme fraction. Les méthodes modernes utilisent plutôt une homogénéisation de Dounce pour lyser les cellules et des centrifugations plus précises, mais les principes restent les mêmes pour établir les enrichissements de molécules, dont les ARN, dans des compartiments subcellulaires spécifiques. Suivant cette approche, plusieurs ARN associés à un grand nombre d'organelles telles que les mitochondries, le réticulum endoplasmique, les membranes, les fuseaux mitotiques, ainsi que les fractions cytosoliques, nucléaires, insolubles et les vésicules extra cellulaires ont pu être identifiés [68, 75, 117, 151, 46, 197, 202, 66, 127]. Cette méthode peut aussi être utilisée lors d'une analyse de transcriptomique comparative afin d'établir les différentiations de localisation de populations d'ARN soit en suivant un processus cellulaire dynamique ou subséquemment à une manipulation génétique telle une déplétion d'un gène [191, 127]. L'utilisation de variation de ces techniques fut primordiale dans la création des jeux de données nécessaires aux analyses bio-informatiques effectuées pour cette thèse.

1.4.3. Techniques *in vitro* et *in vivo* de caractérisation de motifs

L'étude des processus et mécanismes des ARN est intimement liée à l'étude *in vivo* des réseaux de régulations et des protéines liant l'ARN (RBP). Plusieurs méthodes ont été développées pour identifier les interactions endogènes ARN-protéines [**170**, **124**]. En 1979, Michael Lerner *et coll.* furent les premiers à employer une méthode utilisant des anticorps contre la protéine d'épissage *Sm* pour identifier les petits ARN nucléaires formant ce petit complexe ribonucléoprotéide nucléaire [**130**]. Cette méthode fut nommée RIP pour « RNA immunoprecipitation » [**153**]. En 2010, les techniques de RIP furent combinées au séquençage à haut débit dans une méthode nommée RIP-Seq [**200**]. Par contre, ces techniques sont mal adaptées aux analyses des contacts directs entre ARN et protéines, car elles préservent les interactions protéiques [**147**]. Pour pallier cet obstacle, Jernej Ule *et coll.* élaborèrent, en 2003, le « Crosslinking and immunoprecipitation » (CLIP), méthode qui permet la conservation de contact RNA-protéine en s'assurant de purifier une seule RBP [**186**]. Pour leurs analyses, ils utilisèrent le séquençage de Sanger, décrit à la section 1.4.4, pour identifier 340 séquences d'interactions entre l'ARN et les facteurs d'épissage *Nova 1* et *Nova 2* [**186**]. Les optimisations subséquentes faisant appel aux micropuces ou aux séquençages à haut débit entrainèrent une panoplie de méthodes diverses tel le HITS-CLIP (ou CLIP-seq), le PAR-CLIP, le iCLIP, ou le eCLIP à être mis au point [**134**, **76**, **111**, **188**, **189**]. Ceux-ci ne cessent d'être optimisés afin d'améliorer la résolution, maintenant au nucléotide près, des régions d'interactions entre ARN et protéines dans un contexte cellulaire.

Ces techniques in vivo sont par contre difficiles d'application lorsque l'intérêt est de dériver les motifs à la base des interactions ARN-protéines. Ceci est principalement dû au cofacteur protéique et au niveau élevé d'arrière-plan non spécifique [**64**]. Des méthodes *in vitro* ont donc été développées dans ce but. En 2009, Debashish Ray *et coll.* proposèrent RNAcompete, une méthode de sélection à haut débit où des RBP purifiées sont incubées avec un ensemble d'ARN aléatoire, suivi d'un profilage par micropuce [**167**]. Ils ont ainsi pu établir les motifs consensus de près de 300 RBP [**168**]. Par contre, cette méthodologie est conceptuellement limitée, car elle est désignée pour identifier de façon prédominante des motifs dans des régions non structurées, soit des tiges-boucles [**167**, **168**, **166**]. Subséquemment, Nicole Lambert *et coll.* proposèrent une approche semblable utilisant des protéines purifiées, ou leurs domaines de liaison à l'ARN (« RNA binding domain », RBD), mais se basant plutôt sur le séquençage à haut débit, nommé RNA Bind-n-Seq (RBNS) [**118**, **119**, **51**]. Ils ont ainsi pu établir les motifs consensus plus complets de 70 RBP en y incluant la structure secondaire de l'ARN [**118**, **119**, **51**].

L'utilisation de ces approches démontre formellement comment les RBP peuvent s'associer aux ARN afin de moduler leurs fonctions, incluant leur localisation subcellulaire. L'inclusion des données publiques émanant de ces méthodes fut d'une grande valeur et est à l'origine des principales analyses et validations de cette thèse.

1.4.4. Séquençage à haut débit des ARN

Une des technologies récentes qui conduit à un changement révolutionnaire dans l'étude des ARN est sans doute le séquençage ARN à haut débit ou RNA-seq. Celle-ci est conceptuellement basée sur le séquençage ADN de Sanger en 1977 [**174**].

En 1965, Robert Holley *et coll.* furent les premiers à publier une séquence polynucléotidique naturelle, la séquence de 77 nucléotides de l'ARN de transfert de l'alanine, obtenue après sept ans d'efforts considérables leur permettant de récolter un gramme d'ARN [**84**, **163**]. Suivant ces efforts, Walter Fiers *et coll.*, en 1976, dévoilent la première séquence entière d'un génome, révélant ainsi les 3569 bases du bactériophage MS2 [**61**]. S'en suivirent le séquençage de plusieurs autres séquences géniques, dont celle du lysozyme du bactériophage T4, ou les 24 nucléotides de la séquence de l'opéron lac [**163**, **7**, **63**].

La première génération de séquenceurs automatisés fut mise en marché en 1986 et utilisait la méthode de Sanger couplée à des colorants fluorescents permettant une réaction unique plutôt que quatre réactions séparées [**7**, **63**]. L'avènement et la démocratisation de techniques telles que le PCR et de l'utilisation de molécules telles que la Taq polymérase ou la transcriptase inverse mena aux méthodes de séquençage utilisant l'ADN complémentaire (l'ADNc). En 1991, Marc Adams *et coll.* publièrent la première étude utilisant l'ADNc de façon systématique pour générer des séquences d'une taille moyenne de 397 nucléotides qu'il nomma marqueurs de séquences exprimées (« expressed sequence tags », ESTs) [**1**]. Ces ESTs sont toujours importants aux analyses d'identification des transcrits et permettent la reconnaissance de gènes [**1**].

Le séquençage d'ARN à haut débit a évolué de façon exponentielle à travers les années et est souvent catégorisé en plusieurs générations regroupant différentes méthodologies. Parmi les plateformes dites de deuxième génération il y en, entre autres, le DNA 454 GS FLX+ (ou le Roche 454 pyrosequencing), commercialisé par Roche, le « Sequencing by Oligonucleotide Ligation and Detection » ou « SOLiD » commercialisé par Life Technologies, et le HiSeq/NovaSeq commercialisé par Illumina. Ce dernier étant un joueur dominant dans le marché actuel, mais la compétition venant d'entreprises élaborant le séquençage de troisième génération, impliquant des séquences plus longues ou des analyses monocellulaires est forte.

Parallèlement aux développements des plateformes, la quantité considérable de données générées par le RNA-seq a orienté le développement d'infrastructures, d'outils et d'algorithmes efficaces permettant de chercher, de comparer et de visionner les séquences. Ceci a donné lieu à la création et au déploiement d'une panoplie d'algorithmes sophistiqués permettant le profilage transcriptomique, la détermination de l'expression différentielle, la visualisation et l'interprétation. [37, 114]. L'application la plus commune du RNA-seq est l'estimation de l'expression de transcrits à un génome de référence. Premièrement, il est nécessaire de cartographier (« mapping ») les « reads », c'est à dire les séquences de l'accumulation d'ADNc obtenues après un séquençage. Ceci est généralement effectué avec des logiciels comme Bowtie, TopHat, HISAT2, ou STAR [120, 184, 107, 48, 49]. Deuxièmement, une étape de quantification de transcrits peut être effectuée simplement en comptant le nombre de « reads » bruts alignés à une région annotée. Ceci peut être fait avec HTSeq-count ou featureCounts [3, 133]. Les comptes bruts sont rarements suffisant pour effectuer une comparaison robuste des niveaux d'expression de transcrits entre échantillons puisque ces valeurs peuvent différer selon la taille des séquences ou la profondeur du séquençage (le nombre total de « read » obtenu). Des mesures de normalisation telles que les « Fragment Per Kilobase of exon per Million mapped reads » (FPKM) ou les « Transcripts Per Million » (TPM) sont les valeurs de normalisation les plus courantes qui sont définies comme une normalisation du nombre de « reads » d'un transcrit par la taille de celui-ci et le nombre total de « reads » dans l'échantillon [3, 133, 185]. Des logiciels tels que Cufflinks, DeSEQ2 ou RSEM sont habituellement utilisés pour effectuer cette normalisation, mais ces outils font usage de formules beaucoup plus complexes dans le calcul des FPKMs, prenant en compte la distribution des « reads » à travers l'ensemble des échantillons et des conditions [3, 133, 185, 132]. Ces programmes calculent aussi l'expression différentielle d'un transcrit entre deux conditions et la valeur de probabilité (« p-value ») que celle-ci soit significative. L'étape ultime et cruciale se situe au niveau de l'interprétation des résultats. Pour faciliter la compréhension de jeux de données aussi larges, il peut être utile de caractériser les groupes de transcrits différentiellement selon leur ontologie génique (« Gene Ontology », *GO terms*). Les *GO terms*, initiative datant de 1998, définissent les termes représentant les propriétés des gènes et leurs voies métaboliques selon trois catégories : les composants cellulaires, les fonctions moléculaires et les processus biologiques [**67**, **4**]. Ceci permet d'obtenir rapidement une idée générale des fonctions communes à une longue liste de gènes. Le développement et l'optimisation d'une chaine de processus explorant des données issues de RNA-seq dans le contexte particulier du fractionnement cellulaire furent un travail critique et novateur effectué pour cette thèse.

La conjoncture des domaines scientifiques encadrant les techniques décrites dans cette section engendra une redéfinition de la bio-informatique. Le terme provient des publications de Paulien Hogeweg et Ben Hesper, en 1970, qui la définirent comme l'étude des processus d'information dans les systèmes biotiques, et reflète dorénavant un champ d'études multidisciplinaire ayant des ramifications considérablement plus vastes [**154**, **77**, **82**, **83**].

Références

[1] M. D. Adams, J. M. Kelley, J. D. Gocayne, M. Dubnick, M. H. Polymeropoulos, H. Xiao, C. R. Merril, A. Wu, B. Olde, R. F. Moreno, and et al., *Complementary DNA sequencing : expressed sequence tags and human genome project*, Science, 252 (1991), pp. 1651–6.

[2] K. Ainger, D. Avossa, F. Morgan, S. J. Hill, C. Barry, E. Barbarese, and J. H. Carson, *Transport and localization of exogenous myelin basic protein mRNA microinjected into oligodendrocytes*, Journal of Cell Biology, 123 (1993), pp. 431–41.

[3] S. Anders, P. T. Pyl, and W. Huber, *HTSeq–a python framework to work with high-throughput sequencing data*, Bioinformatics, 31 (2015), pp. 166–9.

[4] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock, *Gene ontology : tool for the unification of biology. The Gene Ontology Consortium*, Nature Genetics, 25 (2000), pp. 25–9.

[5] O. T. Avery, C. M. Macleod, and M. McCarty, *Studies on the chemical nature of the substance inducing transformation of pneumococcal types : Induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III*, Journal of Experimental Medicine, 79 (1944), pp. 137–58.

[6] J. P. Bachellerie, J. Cavaille, and A. Huttenhofer, *The expanding snoRNA world*, Biochimie, 84 (2002), pp. 775–90.

[7] M. Barba, H. Czosnek, and A. Hadidi, *Historical perspective, development and applications of next-generation sequencing in plant virology*, Viruses, 6 (2014), pp. 106–36.

[8] M. S. Bartolomei, S. Zemel, and S. M. Tilghman, *Parental imprinting of the mouse H19 gene*, Nature, 351 (1991), pp. 153–5.

[9] E. Basyuk, F. Suavet, A. Doglio, R. Bordonne, and E. Bertrand, *Human let-7 stem-loop precursors harbor features of RNase III cleavage products*, Nucleic Acids Researchearch, 31 (2003), pp. 6593–7.

[10] N. N. Batada, L. A. Shepp, and D. O. Siegmund, *Stochastic model of protein–protein interaction : Why signaling proteins need to be colocalized*, Proceedings of the National Academy of science, 101 (2004), pp. 6445–9.

[11] P. J. Batista and H. Y. Chang, *Cytotopic localization by long noncoding RNAs*, Current Opinion in Cell Biology, 25 (2013), pp. 195–9.

[12] J. G. Bauman, J. Wiegant, P. Borst, and P. van Duijn, *A new method for fluorescence microscopical localization of specific DNA sequences by in situ hybridization of fluorochromelabelled RNA*, Experimental Cell Research, 128 (1980), pp. 485–90.

[13] D. L. Beach, E. D. Salmon, and K. Bloom, *Localization and anchoring of mRNA in budding yeast*, Current Biology, 9 (1999), pp. 569–78.

[14] G. W. Beadle and E. L. Tatum, *Genetic control of biochemical reactions in neurospora*, Proceedings of the National Academy of Sciences of the United States of America, 27 (1941), pp. 499–506.

[15] L. P. Benoit Bouvrette, N. A. L. Cody, J. Bergalet, F. A. Lefebvre, C. Diot, X. Wang, M. Blanchette, and E. Lecuyer, *CeFra-seq reveals broad asymmetric mRNA and noncoding RNA distribution profiles in Drosophila and human cells*, RNA, 24 (2018), pp. 98–113.

[16] J. Bergalet and E. Lécuyer, *The functions and regulatory principles of mRNA intracellular trafficking*, Systems Biology of RNA Binding Proteins, 825 (2014), pp. 57–96.

[17] S. E. Bergsten and E. R. Gavis, *Role for mRNA localization in translational activation but not spatial restriction of nanos RNA*, Development, 126 (1999), pp. 659–69.

[18] E. Bernstein, A. A. Caudy, S. M. Hammond, and G. J. Hannon, *Role for a bidentate ribonuclease in the initiation step of RNA interference*, Nature, 409 (2001), pp. 363–6.

[19] M. D. Blower, E. Feric, K. Weis, and R. J. Heald, *Genome-wide analysis demonstrates conserved localization of messenger RNAs to mitotic microtubules*, Journal of cell biology, 179 (2007), pp. 1365–73.

[20] F. Bohl, C. Kruse, A. Frank, D. Ferring, and R. P. Jansen, *She2p, a novel RNA-binding protein tethers ASH1 mRNA to the Myo4p myosin motor via She3p*, The EMBO Journal, 19 (2000), pp. 5514–24.

[21] S. Bovaird, D. Patel, J. A. Padilla, and E. Lecuyer, *Biological functions, regulatory mechanisms, and disease relevance of RNA localization pathways*, FEBS Letters, 592 (2018), pp. 2948–72.

[22] C. I. Brannan, E. C. Dees, R. S. Ingram, and S. M. Tilghman, *The product of the H19 gene may function as an RNA*, Molecular and Cellular Biology, 10 (1990), pp. 28–36.

[23] S. Brenner, F. Jacob, and M. Meselson, *An unstable intermediate carrying information from genes to ribosomes for protein synthesis*, Nature, 190 (1961), pp. 576–81.

[24] R. J. Britten and E. H. Davidson, *Gene regulation for higher cells : a theory*, Science, 165 (1969), pp. 349–57.

[25] ———, *Repetitive and non-repetitive DNA sequences and a speculation on the origins of evolutionary novelty*, The Quarterly Review of Biology, 46 (1971), pp. 111–38.

[26] C. J. Brown, A. Ballabio, J. L. Rupert, R. G. Lafreniere, M. Grompe, R. Tonlorenzi, and H. F. Willard, *A gene from the region of the human X inactivation centre is expressed exclusively from the inactive X chromosome*, Nature, 349 (1991), pp. 38–44.

[27] S. E. Butcher and D. A. Brow, *Towards understanding the catalytic core structure of the spliceosome*, Biochemical Society Transactions, 33 (2005), pp. 447–9.

[28] M. N. Cabili, C. Trapnell, L. Goff, M. Koziol, B. Tazon-Vega, A. Regev, and J. L. Rinn, *Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses*, Genes and Development, 25 (2011), pp. 1915–27.

[29] J. Cavaille, H. Seitz, M. Paulsen, A. C. Ferguson-Smith, and J. P. Bachellerie, *Identification of tandemly-repeated C/D snoRNA genes at the imprinted human 14q32 domain reminiscent of those at the Prader-Willi/Angelman syndrome region*, Human Molecular Genetics, 11 (2002), pp. 1527–38.

[30] T. R. Cech, *Self-splicing of group I introns*, Annual Review of Biochemistry, 59 (1990), pp. 543–68.

[31] ———, *Structural biology. The ribosome is a ribozyme*, Science, 289 (2000), pp. 878–9.

[32] S. Chung and P. A. Takizawa, *Multiple Myo4 motors enhance ASH1 mRNA transport in Saccharomyces cerevisiae*, Journal of Cell Biology, 189 (2010), pp. 755–67.

[33] M. B. Clark, P. P. Amaral, F. J. Schlesinger, M. E. Dinger, R. J. Taft, J. L. Rinn, C. P. Ponting, P. F. Stadler, K. V. Morris, A. Morillon, J. S. Rozowsky, M. B. Gerstein,

C. Wahlestedt, Y. Hayashizaki, P. Carninci, T. R. Gingeras, and J. S. Mattick, *The reality of pervasive transcription*, PLoS Biology, 9 (2011), p. e1000625 ; discussion e1001102.

[34] A. Claude, *The constitution of protoplasm*, Science, 97 (1943), pp. 451–6.

[35] ———, *Fractionation of mammalian liver cells by differential centrifugation ; experimental procedures and results*, Journal of Experimental Medicine, 84 (1946), pp. 61–89.

[36] N. A. Cody, C. Iampietro, and E. Lecuyer, *The many functions of mRNA localization during normal development and disease : from pillar to post*, Wiley Interdisciplinary Reviews-Developmental Biology, 2 (2013), pp. 781–96.

[37] A. Conesa, P. Madrigal, S. Tarazona, D. Gomez-Cabrero, A. Cervera, A. McPherson, M. W. Szczesniak, D. J. Gaffney, L. L. Elo, X. Zhang, and A. Mortazavi, *A survey of best practices for RNA-seq data analysis*, Genome Biology, 17 (2016), p. 13.

[38] E. P. Consortium, *An integrated encyclopedia of DNA elements in the human genome*, Nature, 489 (2012), pp. 57–74.

[39] E. P. Consortium, E. Birney, J. A. Stamatoyannopoulos, A. Dutta, R. Guigo, T. R. Gingeras, E. H. Margulies, Z. Weng, M. Snyder, E. T. Dermitzakis, R. E. Thurman, M. S. Kuehn, C. M. Taylor, S. Neph, C. M. Koch, S. Asthana, A. Malhotra, I. Adzhubei, J. A. Greenbaum, R. M. Andrews, P. Flicek, P. J. Boyle, H. Cao, N. P. Carter, G. K. Clelland, S. Davis, N. Day, P. Dhami, S. C. Dillon, M. O. Dorschner, H. Fiegler, P. G. Giresi, J. Goldy, M. Hawrylycz, A. Haydock, R. Humbert, K. D. James, B. E. Johnson, E. M. Johnson, T. T. Frum, E. R. Rosenzweig, N. Karnani, K. Lee, G. C. Lefebvre, P. A. Navas, F. Neri, S. C. Parker, P. J. Sabo, R. Sandstrom, A. Shafer, D. Vetrie, M. Weaver, S. Wilcox, M. Yu, F. S. Collins, J. Dekker, J. D. Lieb, T. D. Tullius, G. E. Crawford, S. Sunyaev, W. S. Noble, I. Dunham, F. Denoeud, A. Reymond, P. Kapranov, J. Rozowsky, D. Zheng, R. Castelo, A. Frankish, J. Harrow, S. Ghosh, A. Sandelin, I. L. Hofacker, R. Baertsch, D. Keefe, S. Dike, J. Cheng, H. A. Hirsch, E. A. Sekinger, J. Lagarde, J. F. Abril, A. Shahab, C. Flamm, C. Fried, J. Hackermuller, J. Hertel, M. Lindemeyer, K. Missal, A. Tanzer, S. Washietl, J. Korbel, O. Emanuelsson, J. S. Pedersen, N. Holroyd, R. Taylor, D. Swarbreck, N. Matthews, M. C. Dickson, D. J. Thomas, M. T. Weirauch, et al., *Identification and analysis of functional elements in 1pilot project*, Nature, 447 (2007), pp. 799–816.

[40] D. N. Cox, A. Chao, J. Baker, L. Chang, D. Qiao, and H. Lin, *A novel class of evolutionarily conserved genes defined by piwi are essential for stem cell self-renewal*, Genes and Development, 12 (1998), pp. 3715–27.

[41] A. J. Crofts, H. Washida, T. W. Okita, M. Ogawa, T. Kumamaru, and H. Satoh, *Targeting of proteins to endoplasmic reticulum-derived compartments in plants. The importance of RNA localization*, Plant Physiologyogy, 136 (2004), pp. 3414–19.

[42] R. Dahm, *Discovering DNA : Friedrich Miescher and the early years of nucleic acid research*, Human Genetics, 122 (2008), pp. 565–81.

[43] E. H. Davidson, W. H. Klein, and R. J. Britten, *Sequence organization in animal DNA and a speculation on hnRNA as a coordinate regulatory transcript*, Developmental Biology, 55 (1977), pp. 69–84.

[44] F. De Santa, I. Barozzi, F. Mietton, S. Ghisletti, S. Polletti, B. K. Tusi, H. Muller, J. Ragoussis, C. L. Wei, and G. Natoli, *A large fraction of extragenic RNA pol II transcription sites overlap enhancers*, PLoS Biology, 8 (2010), p. e1000384.

[45] Y. Deng, R. H. Singer, and W. Gu, *Translation of ASH1 mRNA is repressed by Puf6p-Fun12p/eIF5B interaction and released by CK2 phosphorylation*, Genes and Development, 22 (2008), pp. 1037–50.

[46] P. P. Dennis and A. Omer, *Small non-coding RNAs in archaea*, Current Opinion in Microbiology, 8 (2005), pp. 685–94.

[47] S. Djebali, C. A. Davis, A. Merkel, A. Dobin, T. Lassmann, A. Mortazavi, A. Tanzer, J. Lagarde, W. Lin, F. Schlesinger, C. Xue, G. K. Marinov, J. Khatun, B. A. Williams, C. Zaleski, J. Rozowsky, M. Roder, F. Kokocinski, R. F. Abdelhamid, T. Alioto, I. Antoshechkin, M. T. Baer, N. S. Bar, P. Batut, K. Bell, I. Bell, S. Chakrabortty, X. Chen, J. Chrast, J. Curado, T. Derrien, J. Drenkow, E. Dumais, J. Dumais, R. Duttagupta, E. Falconnet, M. Fastuca, K. Fejes-Toth, P. Ferreira, S. Foissac, M. J. Fullwood, H. Gao, D. Gonzalez, A. Gordon, H. Gunawardena, C. Howald, S. Jha, R. Johnson, P. Kapranov, B. King, C. Kingswood, O. J. Luo, E. Park, K. Persaud, J. B. Preall, P. Ribeca, B. Risk, D. Robyr, M. Sammeth, L. Schaffer, L. H. See, A. Shahab, J. Skancke, A. M. Suzuki, H. Takahashi, H. Tilgner, D. Trout, N. Walters, H. Wang, J. Wrobel, Y. Yu, X. Ruan, Y. Hayashizaki, J. Harrow, M. Gerstein, T. Hubbard, A. Reymond, S. E. Antonarakis, G. Hannon, M. C. Giddings, Y. Ruan, B. Wold, P. Carninci, R. Guigo, and T. R. Gingeras, *Landscape of transcription in human cells*, Nature, 489 (2012), pp. 101–8.

[48] A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras, *STAR : ultrafast universal RNA-seq aligner*, Bioinformatics, 29 (2013), pp. 15–21.

[49] A. Dobin and T. R. Gingeras, *Mapping RNA-seq reads with STAR*, Current Protocols in Bioinformatics, 51 (2015), pp. 1–19.

[50] N. Doi, S. Zenno, R. Ueda, H. Ohki-Hamazaki, K. Ui-Tei, and K. Saigo, *Short-interfering-RNA-mediated gene silencing in mammalian cells requires Dicer and eIF2C translation initiation factors*, Current Biology, 13 (2003), pp. 41–6.

[51] D. Dominguez, P. Freese, M. S. Alexis, A. Su, M. Hochman, T. Palden, C. Bazile, N. J. Lambert, E. L. Van Nostrand, G. A. Pratt, G. W. Yeo, B. R. Graveley, and C. B. Burge,

Sequence, structure, and context preferences of human RNA binding proteins, Molecular Cell, 70 (2018), pp. 854–867 e9.

[52] G. Dreyfuss, L. Philipson, and I. W. Mattaj, *Ribonucleoprotein particles in cellular processes*, Journal of Cell Biology, 106 (1988), pp. 1419–25.

[53] T. G. Du, S. Jellbauer, M. Muller, M. Schmid, D. Niessing, and R. P. Jansen, *Nuclear transit of the RNA-binding protein She2 is required for translational control of localized ASH1 mRNA*, EMBO Reports, 9 (2008), pp. 781–7.

[54] J. Dubowy and P. M. Macdonald, *Localization of mRNAs to the oocyte is common in Drosophila ovaries*, Mechanisms of Development, 70 (1998), pp. 193–5.

[55] J. P. Dworkin, A. Lazcano, and S. L. Miller, *The roads to and from the RNA world*, Journal of Theoretical Biology, 222 (2003), pp. 127–34.

[56] J. Eberwine, B. Belt, J. E. Kacharmina, and K. Miyashiro, *Analysis of subcellularly localized mRNAs using in situ hybridization, mRNA amplification, and expression profiling*, Neurochemical research, 27 (2002), pp. 1065–77.

[57] A. Ephrussi and R. Lehmann, *Induction of germ cell formation by oskar*, Nature, 358 (1992), pp. 387–92.

[58] P. Estrada, J. Kim, J. Coleman, L. Walker, B. Dunn, P. Takizawa, P. Novick, and S. Ferro-Novick, *Myo4p and She3p are required for cortical ER inheritance in Saccharomyces cerevisiae*, Journal of Cell Biology, 163 (2003), pp. 1255–66.

[59] M. R. Fabian, N. Sonenberg, and W. Filipowicz, *Regulation of mRNA translation and stability by microRNAs*, Annual Review of Biochemistry, 79 (2010), pp. 351–79.

[60] S. Farris, G. Lewandowski, C. D. Cox, and O. Steward, *Selective localization of arc mRNA in dendrites involves activity- and translation-dependent mRNA degradation*, The Journal of Neuroscience, 34 (2014), pp. 4481–93.

[61] W. Fiers, R. Contreras, F. Duerinck, G. Haegeman, D. Iserentant, J. Merregaert, W. Min Jou, F. Molemans, A. Raeymaekers, A. Van den Berghe, G. Volckaert, and M. Ysebaert, *Complete nucleotide sequence of bacteriophage MS2 RNA : primary and secondary structure of the replicase gene*, Nature, 260 (1976), pp. 500–7.

[62] K. M. Forrest and E. R. Gavis, *Live imaging of endogenous RNA reveals a diffusion and entrapment mechanism for nanos mRNA localization in Drosophila*, Current biology, 13 (2003), pp. 1159–68.

[63] L. T. Franca, E. Carrilho, and T. B. Kist, *A review of DNA sequencing techniques*, Quarterly Reviews of Biophysics, 35 (2002), pp. 169–200.

[64] M. B. Friedersdorf and J. D. Keene, *Advancing the functional utility of PAR-CLIP by quantifying background binding to mRNAs and lncRNAs*, Genome Biology, 15 (2014), p. R2.

[65] M. Furuno, K. C. Pang, N. Ninomiya, S. Fukuda, M. C. Frith, C. Bult, C. Kai, J. Kawai, P. Carninci, Y. Hayashizaki, J. S. Mattick, and H. Suzuki, *Clusters of internally primed transcripts reveal novel long noncoding RNAs*, PLoS Genetics, 2 (2006), p. e37.

[66] J. A. Gagnon and K. L. Mowry, *Molecular motors : directing traffic during RNA localization*, Critical Reviews in Biochemistry and Molecular Biology, 46 (2011), pp. 229–39.

[67] C. Gene Ontology, *Gene Ontology Consortium : going forward*, Nucleic Acids Researchearch, 43 (2015), pp. D1049–56.

[68] S. Ghosh, V. Marchand, I. Gaspar, and A. Ephrussi, *Control of RNP motility and localization by a splicing-dependent structure in oskar mRNA*, Nature Structural and Molecular Biology, 19 (2012), pp. 441–9.

[69] A. Girard, R. Sachidanandam, G. J. Hannon, and M. A. Carmell, *A germline-specific class of small RNAs binds mammalian Piwi proteins*, Nature, 442 (2006), pp. 199–202.

[70] D. Graur, Y. Zheng, N. Price, R. B. Azevedo, R. A. Zufall, and E. Elhaik, *On the immortality of television sets : "function" in the human genome according to the evolution-free gospel of ENCODE*, Genome Biology and Evolution, 5 (2013), pp. 578–90.

[71] F. Gros, H. Hiatt, W. Gilbert, C. G. Kurland, R. W. Risebrough, and J. D. Watson, *Unstable ribonucleic acid revealed by pulse labelling of Escherichia coli*, Nature, 190 (1961), pp. 581–85.

[72] W. Gu, Y. Deng, D. Zenklusen, and R. H. Singer, *A new yeast PUF family protein, Puf6p, represses ASH1 mRNA translation and is required for its localization*, Genes and Development, 18 (2004), pp. 1452–65.

[73] C. Guerrier-Takada, K. Gardiner, T. Marsh, N. Pace, and S. Altman, *The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme*, Cell, 35 (1983), pp. 849–57.

[74] L. F. Gumy, G. S. Yeo, Y. C. Tung, K. H. Zivraj, D. Willis, G. Coppola, B. Y. Lam, J. L. Twiss, C. E. Holt, and J. W. Fawcett, *Transcriptome analysis of embryonic and adult sensory axons reveals changes in mRNA repertoire localization*, RNA, 17 (2011), pp. 85–98.

[75] O. Hachet and A. Ephrussi, *Splicing of oskar RNA in the nucleus is coupled to its cytoplasmic localization*, Nature, 428 (2004), pp. 959–63.

[76] M. Hafner, M. Landthaler, L. Burger, M. Khorshid, J. Hausser, P. Berninger, A. Rothballer, J. Ascano, M., A. C. Jungkamp, M. Munschauer, A. Ulrich, G. S. Wardle, S. Dewell, M. Zavolan, and T. Tuschl, *Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP*, Cell, 141 (2010), pp. 129–41.

[77] J. B. Hagen, *The origins of bioinformatics*, Nature Reviews Genetics, 1 (2000), pp. 231–6.

[78] J. W. Han, J. H. Park, M. Kim, and J. Lee, *mRNAs for microtubule proteins are specifically colocalized during the sequential formation of basal body, flagella, and cytoskeletal microtubules in the differentiation of Naegleria gruberi*, Journal of Cell Biology, 137 (1997), pp. 871–9.

[79] J. J. Henry, K. J. Perry, L. Fukui, and N. Alvi, *Differential localization of mRNAs during early development in the mollusc, Crepidula fornicata*, Integrative and Comparative Biology, 50 (2010), pp. 720–33.

[80] R. G. Heym and D. Niessing, *Principles of mRNA transport in yeast*, Cellular and Molecular Life Sciences, 69 (2012), pp. 1843–53.

[81] M. B. Hoagland, M. L. Stephenson, J. F. Scott, L. I. Hecht, and P. C. Zamecnik, *A soluble ribonucleic acid intermediate in protein synthesis*, Journal of Biological Chemistry, 231 (1958), pp. 241–57.

[82] P. Hogeweg, *The roots of bioinformatics in theoretical biology*, PLoS Computational Biology, 7 (2011), p. e1002021.

[83] P. Hogeweg and B. Hesper, *Bioinformatica : een werkconcept*, Kameleon, 1 (1970), pp. 28–29.

[84] R. W. Holley, J. Apgar, G. A. Everett, J. T. Madison, M. Marquisee, S. H. Merrill, J. R. Penswick, and A. Zamir, *Structure of a ribonucleic acid*, Science, 147 (1965), pp. 1462–5.

[85] D. S. Holmes, J. E. Mayfield, G. Sander, and J. Bonner, *Chromosomal RNA : its properties*, Science, 177 (1972), pp. 72–4.

[86] S. Hutten, T. Sharangdhar, M. Kiebler, S. Hutten, T. Sharangdhar, and M. Kiebler, *Unmasking the messenger*, RNA Biology, 11 (2014), pp. 992–7.

[87] K. Irie, T. Tadauchi, P. A. Takizawa, R. D. Vale, K. Matsumoto, and I. Herskowitz, *The Khd1 protein, which has three KH RNA-binding motifs, is required for proper localization of ASH1 mRNA in yeast*, The EMBO Journal, 21 (2002), pp. 1158–67.

[88] F. Jacob and J. Monod, *Genetic regulatory mechanisms in the synthesis of proteins*, Journal of Molecular Biology, 3 (1961), pp. 318–56.

[89] B. E. Jady, E. Bertrand, and T. Kiss, *Human telomerase RNA and box H/ACA scaRNAs share a common Cajal body-specific localization signal*, Journal of Cell Biology, 164 (2004), pp. 647–52.

[90] H. Jambor, V. Surendranath, A. T. Kalinka, P. Mejstrik, S. Saalfeld, and P. Tomancak, *Systematic imaging reveals features and changing localization of mRNAs in Drosophila development*, eLife, 4 (2015).

[91] R. P. Jansen, *mRNA localization : message on the move*, Nature Reviews Molecular Cell Biology, 2 (2001), pp. 247–56.

[92] R. P. Jansen, C. Dowzer, C. Michaelis, M. Galova, and K. Nasmyth, *Mother cell-specific HO expression in budding yeast depends on the unconventional myosin myo4p and other cytoplasmic proteins*, Cell, 84 (1996), pp. 687–97.

[93] R. P. Jansen and D. Niessing, *Assembly of mRNA-protein complexes for directional mRNA transport in eukaryotes–an overview*, Current Protein an Peptide Science, 13 (2012), pp. 284–93.

[94] J. Jarroux, A. Morillon, and M. Pinskaya, *History, discovery, and classification of lncRNAs*, Advances in Experimental Medicine and Biology, 1008 (2017), pp. 1–46.

[95] W. R. Jeffery, *The role of cytoplasmic determinants in embryonic development*, Developmental Biology, 5 (1988), pp. 3–56.

[96] W. R. Jeffery, C. R. Tomlinson, and R. D. Brodeur, *Localization of actin messenger RNA during early ascidian development*, Developmental Biology, 99 (1983), pp. 408–17.

[97] A. Jenny, O. Hachet, P. Zavorszky, A. Cyrklaff, M. D. Weston, D. S. Johnston, M. Erdelyi, and A. Ephrussi, *A translation-independent role of oskar RNA in early Drosophila oogenesis*, Development, 133 (2006), pp. 2827–33.

[98] B. John, A. J. Enright, A. Aravin, T. Tuschl, C. Sander, and D. S. Marks, *Human MicroRNA targets*, PLoS Biology, 2 (2004), p. e363.

[99] P. Johnsson, L. Lipovich, D. Grander, and K. V. Morris, *Evolutionary conservation of long non-coding RNAs ; sequence, structure, function*, Biochimica et Biophysica Acta, 1840 (2014), pp. 1063–71.

[100] H. Jorjani, S. Kehr, D. J. Jedlinski, R. Gumienny, J. Hertel, P. F. Stadler, M. Zavolan, and A. R. Gruber, *An updated human snoRNAome*, Nucleic Acids Research, 44 (2016), pp. 5068–82.

[101] F. Juhling, M. Morl, R. K. Hartmann, M. Sprinzl, P. F. Stadler, and J. Putz, *tRNAdb 2009 : compilation of tRNA sequences and tRNA genes*, Nucleic Acids Research, 37 (2009), pp. D159–62.

[102] P. Kapranov, S. E. Cawley, J. Drenkow, S. Bekiranov, R. L. Strausberg, S. P. Fodor, and T. R. Gingeras, *Large-scale transcriptional activity in chromosomes 21 and 22*, Science, 296 (2002), pp. 916–9.

[103] S. Katayama, Y. Tomaru, T. Kasukawa, K. Waki, M. Nakanishi, M. Nakamura, H. Nishida, C. C. Yap, M. Suzuki, J. Kawai, H. Suzuki, P. Carninci, Y. Hayashizaki, C. Wells, M. Frith, T. Ravasi, K. C. Pang, J. Hallinan, J. Mattick, D. A. Hume, L. Lipovich, S. Batalov, P. G. Engstrom, Y. Mizuno, M. A. Faghihi, A. Sandelin, A. M. Chalk, S. Mottagui-Tabar, Z. Liang, B. Lenhard, C. Wahlestedt, R. G. E. R. Group, G. Genome Science, and F. Consortium, *Antisense transcription in the mammalian transcriptome*, Science, 309 (2005), pp. 1564–6.

[104] K. D. Kaya, G. Karakulah, C. M. Yakicier, A. C. Acar, and O. Konu, *mESAdb : microRNA expression and sequence analysis database*, Nucleic Acids Research, 39 (2011), pp. D170–80.

[105] K. C. Keiler, *RNA localization in bacteria*, Current Opinion Microbiology, 14 (2011), pp. 155–9.

[106] A. M. Khalil, M. Guttman, M. Huarte, M. Garber, A. Raj, D. Rivea Morales, K. Thomas, A. Presser, B. E. Bernstein, A. van Oudenaarden, A. Regev, E. S. Lander, and J. L. Rinn, *Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression*, Proceedings of the National Academy of Sciences of the United States of America, 106 (2009), pp. 11667–72.

[107] D. Kim, J. M. Paggi, C. Park, C. Bennett, and S. L. Salzberg, *Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype*, Nature Biotechnology, 37 (2019), pp. 907–915.

[108] J. K. Kim, H. W. Gabel, R. S. Kamath, M. Tewari, A. Pasquinelli, J. F. Rual, S. Kennedy, M. Dybbs, N. Bertin, J. M. Kaplan, M. Vidal, and G. Ruvkun, *Functional genomic analysis of RNA interference in C. elegans*, Science, 308 (2005), pp. 1164–7.

[109] E. P. Kingsley, X. Y. Chan, Y. Duan, and J. D. Lambert, *Widespread RNA segregation in a spiralian embryo*, Evolution and Development, 9 (2007), pp. 527–39.

[110] M. Kloc, M. T. Dougherty, S. Bilinski, A. P. Chan, E. Brey, M. L. King, J. Patrick, C. W., and L. D. Etkin, *Three-dimensional ultrastructural analysis of RNA distribution within germinal granules of Xenopus*, Developmental Biology, 241 (2002), pp. 79–93.

[111] J. Konig, K. Zarnack, G. Rot, T. Curk, M. Kayikci, B. Zupan, D. J. Turner, N. M. Luscombe, and J. Ule, *iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution*, Nature Structural and Molecular Biology, 17 (2010), pp. 909–15.

[112] E. B. Krementsova, A. R. Hodges, C. S. Bookwalter, T. E. Sladewski, M. Travaglia, H. L. Sweeney, and K. M. Trybus, *Two single-headed myosin V motors bound to a tetrameric adapter protein form a processive complex*, Journal of Cell Biology, 195 (2011), pp. 631–41.

[113] K. Kruger, P. J. Grabowski, A. J. Zaug, J. Sands, D. E. Gottschling, and T. R. Cech, *Self-splicing RNA : autoexcision and autocyclization of the ribosomal RNA intervening sequence of tetrahymena*, Cell, 31 (1982), pp. 147–57.

[114] K. R. Kukurba and S. B. Montgomery, *Rna sequencing and analysis*, Cold Spring Harbor Protocols, 2015 (2015), pp. 951–69.

[115] S. Kuramochi-Miyagawa, T. Kimura, T. W. Ijiri, T. Isobe, N. Asada, Y. Fujita, M. Ikawa, N. Iwai, M. Okabe, W. Deng, H. Lin, Y. Matsuda, and T. Nakano, *Mili, a mammalian member of piwi family gene, is essential for spermatogenesis*, Development, 131 (2004), pp. 839–49.

[116] S. Kuramochi-Miyagawa, T. Watanabe, K. Gotoh, Y. Totoki, A. Toyoda, M. Ikawa, N. Asada, K. Kojima, Y. Yamaguchi, T. W. Ijiri, K. Hata, E. Li, Y. Matsuda, T. Kimura, M. Okabe, Y. Sakaki, H. Sasaki, and T. Nakano, *DNA methylation of retrotransposon genes is regulated by Piwi family members MILI and MIWI2 in murine fetal testes*, Genes and Development, 22 (2008), pp. 908–17.

[117] L. La Via, D. Bonini, I. Russo, C. Orlandi, S. Barlati, and A. Barbon, *Modulation of dendritic AMPA receptor mRNA trafficking by RNA splicing and editing*, Nucleic Acids Research, 41 (2013), pp. 617–31.

[118] N. Lambert, A. Robertson, M. Jangi, and S. McGeary, *RNA Bind-n-Seq : quantitative assessment of the sequence and structural binding specificity of RNA binding proteins*, Molecular cell, 54 (2014), pp. 887–900.

[119] N. J. Lambert, A. D. Robertson, and C. B. Burge, *RNA Bind-n-Seq : Measuring the binding affinity landscape of rna-binding proteins*, Methods in Enzymology, 558 (2015), pp. 465–93.

[120] B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg, *Ultrafast and memory-efficient alignment of short DNA sequences to the human genome*, Genome Biology, 10 (2009), p. R25.

[121] P. Lasko, *mRNA localization and translational control in Drosophila oogenesis*, Cold Spring Harbor Perspectives in Biology, 4 (2012).

[122] E. Lecuyer, S. Lariviere, M. C. Sincennes, A. Haman, R. Lahlil, M. Todorova, M. Tremblay, B. C. Wilkes, and T. Hoang, *Protein stability and transcription factor complex assembly determined by the SCL-LMO2 interaction*, Journal of biogical chemistry, 282 (2007), pp. 33649–58.

[123] E. Lecuyer, H. Yoshida, N. Parthasarathy, C. Alm, T. Babak, T. Cerovina, T. R. Hughes, P. Tomancak, and H. M. Krause, *Global analysis of mRNA localization reveals a prominent role in organizing cellular architecture and function*, Cell, 131 (2007), pp. 174–87.

[124] F. C. Y. Lee and J. Ule, *Advances in CLIP technologies for studies of protein-RNA interactions*, Molecular Cell, 69 (2018), pp. 354–69.

[125] R. C. Lee, R. L. Feinbaum, and V. Ambros, *The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14*, Cell, 75 (1993), pp. 843–54.

[126] Y. Lee, C. Ahn, J. Han, H. Choi, J. Kim, J. Yim, J. Lee, P. Provost, O. Radmark, S. Kim, and V. N. Kim, *The nuclear RNase III Drosha initiates microRNA processing*, Nature, 425 (2003), pp. 415–9.

[127] F. A. Lefebvre, L. P. Benoit Bouvrette, L. Perras, A. Blanchet-Cohen, D. Garnier, J. Rak, and E. Lecuyer, *Comparative transcriptomic analysis of human and Drosophila extracellular vesicles*, Scientific Reports, 6 (2016), p. 27680.

[128] F. A. Lefebvre, N. A. L. Cody, L. P. Benoit Bouvrette, J. Bergalet, X. Wang, and E. Lecuyer, *CeFra-seq : Systematic mapping of RNA subcellular distribution properties through cell fractionation coupled to deep-sequencing*, Methods, 126 (2017), pp. 138–48.

[129] E. S. Lein, M. J. Hawrylycz, N. Ao, M. Ayres, A. Bensinger, A. Bernard, A. F. Boe, M. S. Boguski, K. S. Brockway, E. J. Byrnes, L. Chen, L. Chen, T. M. Chen, M. C. Chin, J. Chong, B. E. Crook, A. Czaplinska, C. N. Dang, S. Datta, N. R. Dee, A. L. Desaki, T. Desta, E. Diep, T. A. Dolbeare, M. J. Donelan, H. W. Dong, J. G. Dougherty, B. J.

Duncan, A. J. Ebbert, G. Eichele, L. K. Estin, C. Faber, B. A. Facer, R. Fields, S. R. Fischer, T. P. Fliss, C. Frensley, S. N. Gates, K. J. Glattfelder, K. R. Halverson, M. R. Hart, J. G. Hohmann, M. P. Howell, D. P. Jeung, R. A. Johnson, P. T. Karr, R. Kawal, J. M. Kidney, R. H. Knapik, C. L. Kuan, J. H. Lake, A. R. Laramee, K. D. Larsen, C. Lau, T. A. Lemon, A. J. Liang, Y. Liu, L. T. Luong, J. Michaels, J. J. Morgan, R. J. Morgan, M. T. Mortrud, N. F. Mosqueda, L. L. Ng, R. Ng, G. J. Orta, C. C. Overly, T. H. Pak, S. E. Parry, S. D. Pathak, O. C. Pearson, R. B. Puchalski, Z. L. Riley, H. R. Rockett, S. A. Rowland, J. J. Royall, M. J. Ruiz, N. R. Sarno, K. Schaffnit, N. V. Shapovalova, T. Sivisay, C. R. Slaughterbeck, S. C. Smith, K. A. Smith, B. I. Smith, A. J. Sodt, N. N. Stewart, K. R. Stumpf, S. M. Sunkin, M. Sutram, A. Tam, C. D. Teemer, C. Thaller, C. L. Thompson, L. R. Varnam, A. Visel, R. M. Whitlock, P. E. Wohnoutka, C. K. Wolkey, V. Y. Wong, et al., *Genome-wide atlas of gene expression in the adult mouse brain*, Nature, 445 (2007), pp. 168–76.

[130] M. R. Lerner and J. A. Steitz, *Antibodies to small nuclear RNAs complexed with proteins are produced by patients with systemic lupus erythematosus*, Proceedings of the National Academy of Sciences of the United States of America, 76 (1979), pp. 5495–9.

[131] B. P. Lewis, C. B. Burge, and D. P. Bartel, *Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets*, Cell, 120 (2005), pp. 15–20.

[132] B. Li and C. N. Dewey, *RSEM : accurate transcript quantification from RNA-Seq data with or without a reference genome*, BMC Bioinformatics, 12 (2011), p. 323.

[133] Y. Liao, G. K. Smyth, and W. Shi, *featureCounts : an efficient general purpose program for assigning sequence reads to genomic features*, Bioinformatics, 30 (2014), pp. 923–30.

[134] D. D. Licatalosi, A. Mele, J. J. Fak, J. Ule, M. Kayikci, S. W. Chi, T. A. Clark, A. C. Schweitzer, J. E. Blume, X. Wang, J. C. Darnell, and R. B. Darnell, *HITS-CLIP yields genome-wide insights into brain alternative RNA processing*, Nature, 456 (2008), pp. 464–9.

[135] H. Lin and A. C. Spradling, *A novel group of pumilio mutations affects the asymmetric division of germline stem cells in the Drosophila ovary*, Development, 124 (1997), pp. 2463–76.

[136] R. M. Long, W. Gu, E. Lorimer, R. H. Singer, and P. Chartrand, *She2p is a novel RNA-binding protein that recruits the Myo4p-She3p complex to ASH1 mRNA*, The EMBO Journal, 19 (2000), pp. 6592–601.

[137] R. M. Long, W. Gu, X. Meng, G. Gonsalvez, R. H. Singer, and P. Chartrand, *An exclusively nuclear RNA-binding protein affects asymmetric localization of ASH1 mRNA and Ash1p in yeast*, Journal of Cell Biology, 153 (2001), pp. 307–18.

[138] R. M. Long, R. H. Singer, X. Meng, I. Gonzalez, K. Nasmyth, and R. P. Jansen, *Mating type switching in yeast controlled by asymmetric localization of ASH1 mRNA*, Science, 277 (1997), pp. 383–7.

[139] M. F. Lyon, *Gene action in the X-chromosome of the mouse (Mus musculus L.)*, Nature, 190 (1961), pp. 372–3.

[140] K. W. Makabe, T. Kawashima, S. Kawashima, T. Minokawa, A. Adachi, H. Kawamura, H. Ishikawa, R. Yasuda, H. Yamamoto, K. Kondoh, S. Arioka, Y. Sasakura, A. Kobayashi, K. Yagi, K. Shojima, Y. Kondoh, S. Kido, M. Tsujinami, N. Nishimura, M. Takahashi, T. Nakamura, M. Kanehisa, M. Ogasawara, T. Nishikata, and H. Nishida, *Large-scale cDNA analysis of the maternal genetic information in the egg of Halocynthia roretzi for a gene expression catalog of ascidian development*, Development, 128 (2001), pp. 2555–67.

[141] K. C. Martin and A. Ephrussi, *mRNA localization : gene expression in the spatial dimension*, Cell, 136 (2009), pp. 719–730.

[142] E. S. Maxwell and M. J. Fournier, *The small nucleolar RNAs*, Annual Review of Biochemistry, 64 (1995), pp. 897–934.

[143] C. Medioni, K. Mowry, and F. Besse, *Principles and roles of mRNA localization in animal development*, Development, 139 (2012), pp. 3263–76.

[144] T. R. Mercer, M. E. Dinger, S. M. Sunkin, M. F. Mehler, and J. S. Mattick, *Specific expression of long noncoding RNAs in the mouse brain*, Proceedings of the National Academy of Sciences of the United States of America, 105 (2008), pp. 716–21.

[145] M. Mikl, G. Vendra, and M. A. Kiebler, *Independent localization of MAP2, CaMKIIalpha and beta-actin RNAs in low copy numbers*, EMBO Reports, 12 (2011), pp. 1077–84.

[146] S. Mili, K. Moissoglu, and I. G. Macara, *Genome-wide screen reveals APC-associated RNAs enriched in cell protrusions*, Nature, 453 (2008), pp. 115–21.

[147] S. Mili and J. A. Steitz, *Evidence for reassociation of RNA-binding proteins after cell lysis : implications for the interpretation of immunoprecipitation analyses*, RNA, 10 (2004), pp. 1692–4.

[148] L. A. Mingle, N. N. Okuhama, J. Shi, R. H. Singer, J. Condeelis, and G. Liu, *Localization of all seven messenger RNAs for the actin-polymerization nucleator Arp2/3 complex in the protrusions of fibroblasts*, Journal of Cell Science, 118 (2005), pp. 2425–33.

[149] K. V. Morris and J. S. Mattick, *The rise of regulatory RNA*, Nature Reviews Genetics, 15 (2014), pp. 423–37.

[150] M. Muller, R. G. Heym, A. Mayer, K. Kramer, M. Schmid, P. Cramer, H. Urlaub, R. P. Jansen, and D. Niessing, *A cytoplasmic complex mediates specific mRNA recognition and localization in yeast*, PLoS Biology, 9 (2011), p. e1000611.

[151] I. A. Muslimov, E. Santi, P. Homel, S. Perini, D. Higgins, and H. Tiedge, *RNA transport in dendrites : a cis-acting targeting element is contained within neuronal BC1 RNA*, Journal of Neuroscience, 17 (1997), pp. 4722–33.

[152] E. Natan, J. N. Wells, S. A. Teichmann, and J. A. Marsh, *Regulation, evolution and consequences of cotranslational protein complex assembly*, Current Opinion in Structural Biology, 42 (2017), pp. 90–97.

[153] S. Niranjanakumari, E. Lasda, R. Brazas, and M. A. Garcia-Blanco, *Reversible cross-linking combined with immunoprecipitation to study RNA-protein interactions in vivo*, Methods, 26 (2002), pp. 182–90.

[154] C. A. Ouzounis and A. Valencia, *Early bioinformatics : the birth of a discipline–a personal view*, Bioinformatics, 19 (2003), pp. 2176–90.

[155] V. Pachnis, A. Belayew, and S. M. Tilghman, *Locus unlinked to alpha-fetoprotein under the control of the murine raf and Rif genes*, Proceedings of the National Academy of Sciences of the United States of America, 81 (1984), pp. 5523–7.

[156] G. E. Palade, *A small particulate component of the cytoplasm*, The Journal of Biophysical and Biochemical Cytology, 1 (1955), pp. 59–68.

[157] N. Paquin and P. Chartrand, *Local regulation of mRNA translation : new insights from the bud*, Trends in Cell Biology, 18 (2008), pp. 105–11.

[158] N. Paquin, M. Menade, G. Poirier, D. Donato, E. Drouet, and P. Chartrand, *Local activation of yeast ASH1 mRNA translation through phosphorylation of Khd1p by the casein kinase Yck1p*, Molecular Cell, 26 (2007), pp. 795–809.

[159] M. L. Pardue and J. G. Gall, *Molecular hybridization of radioactive DNA to the DNA of cytological preparations*, Proceedings of the National Academy of Sciences of the United States of America, 64 (1969), pp. 600–4.

[160] B. K. D. Pearce, R. E. Pudritz, D. A. Semenov, and T. K. Henning, *Origin of the RNA world : The fate of nucleobases in warm little ponds*, Proceedings of the National Academy of Sciences of the United States of America, 114 (2017), pp. 11327–32.

[161] C. Ponnamperuma, C. Sagan, and R. Mariner, *Synthesis of adenosine triphosphate under possible primitive earth conditions*, Nature, 199 (1963), pp. 222–6.

[162] F. Prodon, L. Yamada, M. Shirae-Kurabayashi, Y. Nakamura, and Y. Sasakura, *Postplasmic/PEM RNAs : a class of localized maternal mRNAs with multiple roles in cell polarity and development in ascidian embryos*, Developmental dynamic, 236 (2007), pp. 1698–715.

[163] U. L. RajBhandary and C. Kohrer, *Early days of tRNA research : discovery, function, purification and sequence analysis*, Journal of Biosciences, 31 (2006), pp. 439–51.

[164] N. Rasmussen, *Cell fractionation biochemistry and the origins of 'cell biology'*, Trends in Biochemical Sciences, 21 (1996), pp. 319–21.

[165] B. K. Ray and D. Apirion, *Characterization of 10S RNA : a new stable RNA molecule from Escherichia coli*, Molecular Genetics and Genomics, 174 (1979), pp. 25–32.

[166] D. Ray, K. C. H. Ha, K. Nie, H. Zheng, T. R. Hughes, and Q. D. Morris, *Rnacompete methodology and application to determine sequence preferences of unconventional RNA-binding proteins*, Methods, 118-119 (2017), pp. 3–15.

[167] D. Ray, H. Kazan, E. T. Chan, L. Pena Castillo, S. Chaudhry, S. Talukder, B. J. Blencowe, Q. Morris, and T. R. Hughes, *Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins*, Nature Biotechnology, 27 (2009), pp. 667–70.

[168] D. Ray, H. Kazan, K. B. Cook, M. T. Weirauch, H. S. Najafabadi, X. Li, S. Gueroussov, M. Albu, H. Zheng, A. Yang, H. Na, M. Irimia, L. H. Matzat, R. K. Dale, S. A. Smith, C. A. Yarosh, S. M. Kelly, B. Nabet, D. Mecenas, W. Li, R. S. Laishram, M. Qiao, H. D. Lipshitz, F. Piano, A. H. Corbett, R. P. Carstens, B. J. Frey, R. A. Anderson, K. W. Lynch, L. O. Penalva, E. P. Lei, A. G. Fraser, B. J. Blencowe, Q. D. Morris, and T. R. Hughes, *A compendium of RNA-binding motifs for decoding gene regulation*, Nature, 499 (2013), pp. 172–7.

[169] B. J. Reinhart, F. J. Slack, M. Basson, A. E. Pasquinelli, J. C. Bettinger, A. E. Rougvie, H. R. Horvitz, and G. Ruvkun, *The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans*, Nature, 403 (2000), pp. 901–6.

[170] K. J. Riley and J. A. Steitz, *The "observer effect" in genome-wide surveys of protein-RNA interactions*, Molecular Cell, 49 (2013), pp. 601–4.

[171] B. Rogelj, C. E. Hartmann, C. H. Yeo, S. P. Hunt, and K. P. Giese, *Contextual fear conditioning regulates the expression of brain-specific small nucleolar RNAs in hippocampus*, European Journal of Neuroscience, 18 (2003), pp. 3089–96.

[172] J. E. Rothman, *Mechanisms of intracellular protein transport*, Nature, 372 (1994), pp. 55–63.

[173] J. Salzman, C. Gawad, P. L. Wang, N. Lacayo, and P. O. Brown, *Circular RNAs are the predominant transcript isoform from hundreds of human genes in diverse cell types*, PLoS One, 7 (2012), p. e30733.

[174] F. Sanger, S. Nicklen, and A. R. Coulson, *DNA sequencing with chain-terminating inhibitors*, Proceedings of the National Academy of Sciences of the United States of America, 74 (1977), pp. 5463–7.

[175] M. Schnall-Levin, O. S. Rissland, W. K. Johnston, N. Perrimon, D. P. Bartel, and B. Berger, *Unusually effective microRNA targeting within repeat-rich coding regions of mammalian mRNAs*, Genome Research, 21 (2011), pp. 1395–403.

[176] R. Schweet, H. Lamfrom, and E. Allen, *The synthesis of hemoglobin in a cell-free system*, Proceedings of the National Academy of Sciences of the United States of America, 44 (1958), pp. 1029–35.

[177] K. A. Serikawa, D. M. Porterfield, and D. F. Mandoli, *Asymmetric subcellular mRNA distribution correlates with carbonic anhydrase activity in Acetabularia acetabulum*, Plant Physiology, 125 (2001), pp. 900–11.

[178] Z. Shen, N. Paquin, A. Forget, and P. Chartrand, *Nuclear shuttling of She2p couples ASH1 mRNA localization to its translational repression by recruiting Loc1p and Puf6p*, Molecular Biology of the Cell, 20 (2009), pp. 2265–75.

[179] M. A. Smith, T. Gesell, P. F. Stadler, and J. S. Mattick, *Widespread purifying selection on RNA structure in mammals*, Nucleic Acids Research, 41 (2013), pp. 8220–36.

[180] R. Smith, *Moving molecules : mRNA trafficking in mammalian oligodendrocytes and neurons*, Neuroscientist, 10 (2004), pp. 495–500.

[181] R. J. Taft, E. A. Glazov, N. Cloonan, C. Simons, S. Stephen, G. J. Faulkner, T. Lassmann, A. R. Forrest, S. M. Grimmond, K. Schroder, K. Irvine, T. Arakawa, M. Nakamura, A. Kubosaki, K. Hayashida, C. Kawazu, M. Murata, H. Nishiyori, S. Fukuda, J. Kawai, C. O. Daub, D. A. Hume, H. Suzuki, V. Orlando, P. Carninci, Y. Hayashizaki, and J. S. Mattick, *Tiny RNAs associated with transcription start sites in animals*, Nature Genetics, 41 (2009), pp. 572–8.

[182] R. J. Taft, C. Simons, S. Nahkuri, H. Oey, D. J. Korbie, T. R. Mercer, J. Holst, W. Ritchie, J. J. Wong, J. E. Rasko, D. S. Rokhsar, B. M. Degnan, and J. S. Mattick, *Nuclear-localized tiny RNAs are associated with transcription initiation and splice sites in metazoans*, Nature Structural and Molecular Biology, 17 (2010), pp. 1030–4.

[183] P. Tomancak, A. Beaton, R. Weiszmann, E. Kwan, S. Shu, S. E. Lewis, S. Richards, M. Ashburner, V. Hartenstein, S. E. Celniker, and G. M. Rubin, *Systematic determination of patterns of gene expression during Drosophila embryogenesis*, Genome Biology, 3 (2002), p. RESEARCH0088.

[184] C. Trapnell, L. Pachter, and S. L. Salzberg, *TopHat : discovering splice junctions with RNA-seq*, Bioinformatics, 25 (2009), pp. 1105–11.

[185] C. Trapnell, B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. van Baren, S. L. Salzberg, B. J. Wold, and L. Pachter, *Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation*, Nature Biotechnology, 28 (2010), pp. 511–5.

[186] J. Ule, K. B. Jensen, M. Ruggiu, A. Mele, A. Ule, and R. B. Darnell, *CLIP identifies Nova-regulated RNA networks in the brain*, Science, 302 (2003), pp. 1212–5.

[187] V. V. Vagin, A. Sigova, C. Li, H. Seitz, V. Gvozdev, and P. D. Zamore, *A distinct small RNA pathway silences selfish genetic elements in the germline*, Science, 313 (2006), pp. 320–4.

[188] E. L. Van Nostrand, G. A. Pratt, A. A. Shishkin, C. Gelboin-Burkhart, M. Y. Fang, B. Sundararaman, S. M. Blue, T. B. Nguyen, C. Surka, K. Elkins, R. Stanton, F. Rigo, M. Guttman, and G. W. Yeo, *Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP)*, Nature Methods, 13 (2016), pp. 508–14.

[189] E. L. Van Nostrand, A. A. Shishkin, G. A. Pratt, T. B. Nguyen, and G. W. Yeo, *Variation in single-nucleotide sensitivity of eCLIP derived from reverse transcription conditions*, Methods, 126 (2017), pp. 29–37.

[190] C. Wang, B. Han, R. Zhou, and X. Zhuang, *Real-time imaging of translation on single mRNA transcripts in live cells*, Cell, 165 (2016), pp. 990–1001.

[191] E. T. Wang, N. A. Cody, S. Jog, M. Biancolella, T. T. Wang, D. J. Treacy, S. Luo, G. P. Schroth, D. E. Housman, S. Reddy, E. Lecuyer, and C. B. Burge, *Transcriptome-wide regulation of pre-mRNA splicing and mRNA localization by muscleblind proteins*, Cell, 150 (2012), pp. 710–24.

[192] Z. Wang and C. B. Burge, *Splicing regulation : from a parts list of regulatory elements to an integrated splicing code*, RNA, 14 (2008), pp. 802–13.

[193] J. R. Warner, R. Soeiro, H. C. Birnboim, M. Girard, and J. E. Darnell, *Rapidly labeled HeLa cell nuclear RNA. I. identification by zone sedimentation of a heterogeneous fraction separate from ribosomal precursor RNA*, Journal of Molecular Biology, 19 (1966), pp. 349–61.

[194] J. D. Watson and F. H. Crick, *Molecular structure of nucleic acids ; a structure for deoxyribose nucleic acid*, Nature, 171 (1953), pp. 737–8.

[195] R. A. Weinberg and S. Penman, *Small molecular weight monodisperse nuclear RNA*, Journal of Molecular Biology, 38 (1968), pp. 289–304.

[196] X. Yan, T. A. Hoek, R. D. Vale, and M. E. Tanenbaum, *Dynamics of translation of single mRNA molecules in vivo*, Cell, 165 (2016), pp. 976–89.

[197] J. K. Yisraeli, *VICKZ proteins : a multi-talented family of regulatory RNA-binding proteins*, Biology of the Cell, 97 (2005), pp. 87–96.

[198] S. Zaessinger, I. Busseau, and M. Simonelig, *Oskar allows nanos mRNA translation in Drosophila embryos by preventing its deadenylation by Smaug/CCR4*, Development, 133 (2006), pp. 4573–83.

[199] K. Zarnack and M. Feldbrugge, *Microtubule-dependent mRNA transport in fungi*, Eukaryotic Cell, 9 (2010), pp. 982–90.

[200] J. Zhao, T. K. Ohsumi, J. T. Kung, Y. Ogawa, D. J. Grau, K. Sarma, J. J. Song, R. E. Kingston, M. Borowsky, and J. T. Lee, *Genome-wide identification of polycomb-associated RNAs by RIP-seq*, Molecular Cell, 40 (2010), pp. 939–53.

[201] W. Zhou, L. Wang, Y. Miao, and R. Xing, *Novel long noncoding RNA GACAT3 promotes colorectal cancer cell proliferation, invasion, and migration through miR-149*, OncoTargets and Therapy, 11 (2018), pp. 1543–52.

[202] K. H. Zivraj, M. Rehbein, J. Olschlager-Schutt, C. Schob, K. Falley, F. Buck, M. Schweizer, A. Schepis, E. Kremmer, D. Richter, H. J. Kreienkamp, and S. Kindler, *The RNA-binding protein MARTA2 regulates dendritic targeting of MAP2 mRNAs in rat neurons*, Journal of Neurochemistry, 124 (2013), pp. 670–84.

[203] C. Zwieb, J. Gorodkin, B. Knudsen, J. Burks, and J. Wower, *tmRDB (tmRNA database)*, Nucleic Acids Research, 31 (2003), pp. 446–7.

Premier article.

Bioinformatics approaches to gain insights about *cis*-regulatory motifs involved in mRNA localization

Louis Philip Benoit Bouvrette

Cet article de revue de littérature a été publié dans le livre *The Biology of mRNA : Structure and Function*, Advances in Experimental Medicine and Biology, vol 1203, Nature Springer (https://doi.org/10.1007/978-3-030-31434-7_7)

Préface et contributions

Ce chapitre d'introduction est présenté sous forme d'un article de revue et a aussi été publié comme chapitre dans le livre *The Biology of mRNA : Structure and Function*.

L'article discute, après un bref survol historique, de plusieurs aspects fondamentaux de la régulation post-transcriptionnelle des gènes, dont la localisation des ARN, ainsi que des outils informatiques disponibles pour en faire l'étude. Dans cet article, j'adresse premièrement le rôle de certains motifs de régulation en *cis* impliqués dans le trafic intracellulaire des ARN. Deuxièmement, je parle de la représentation de séquence de motifs en termes de contenu de l'information. Troisièmement, je cite les algorithmes et outils existants permettant l'analyse bio-informatique appliquée à l'étude des ARN. Finalement, je propose un exemple illustrant les points importants du chapitre, en reprenant, sous forme plus générale, des aspects étudiés plus en détail dans les articles subséquents de cette thèse.

J'ai effectué les analyses, créé l'ensemble des figures et écrit l'intégralité du manuscrit. Eric Lécuyer et Mathieu Blanchette ont supervisé les travaux et l'écriture du manuscrit et ont suggéré des modifications concernant, principalement et respectivement, les aspects biologique et informatique. Ils l'ont approuvé avant sa soumission à l'éditeur de *Nature Springer*, qui l'a fait évaluer par des paires avant de le publier.

Bioinformatics approaches to gain insights about *cis*-regulatory motifs involved in mRNA localization

par

Louis Philip Benoit Bouvrette[1, 2], Mathieu Blanchette[3] et Eric Lécuyer[1, 2, 4]

(1) Institut de Recherches Cliniques de Montréal (IRCM), Montréal, Québec, Canada
(2) Département de biochimie, Université de Montréal, Montréal, Québec, Canada
(3) School of computer science, McGill University, Montréal, Québec, Canada
(4) Division of experimental medicine, McGill University, Montréal, Québec, Canada

Cet article de revue de littérature a été publié dans le livre *The Biology of mRNA : Structure and Function*, Advances in Experimental Medicine and Biology, vol 1203, Nature Springer (https://doi.org/10.1007/978-3-030-31434-7_7).

Abstract. Messenger RNA (mRNA) is a fundamental intermediate in the expression of proteins. As an integral part of this important process, protein production can be localized by the targeting of mRNA to a specific subcellular compartment. The subcellular destination of mRNA is suggested to be governed by a region of its primary sequence or secondary structure, which consequently dictates the recruitment of *trans*-acting factors, such as RNA-binding proteins or regulatory RNAs, to form a messenger ribonucleoprotein particle. This molecular ensemble is requisite for precise and spatiotemporal control of gene expression. In the context of RNA localization, the description of the binding preferences of an RNA-binding protein defines a *motif* and one, or more, instances of a given motif is defined as a *localization element* (zip code). In this chapter, we first discuss the *cis*-regulatory motifs previously identified as mRNA localization elements. We then describe motif representation in terms of entropy and information content and offer an overview of motif databases and search algorithms. Finally, we provide an outline of the motif topology of asymmetrically localized mRNA molecules.

Keywords: RNA localization, RNA binding protein/RBP, *cis*-regulatory motifs

1. General Introduction

In 1950, it was first hypothesized that RNA was synthesized in the nucleus and then transferred into the cytoplasm, where it was aggregating with other molecules [90]. A better appreciation of the role of RNA was gained in 1961 when 3 publications revolutionized the way gene function was perceived by establishing messenger RNA (mRNA) as an information carrier in a transitional stage towards the synthesis of protein [26, 70, 86]. Following these breakthroughs, it was not immediately apparent whether mRNA could localize to specific subcellular sites. It was not until the mid-1980s that the first elements of the answer were identified when it was reported that the *actin* mRNA in ascidian oocytes and embryos was asymmetrically distributed [91]. The discovery of additional localized RNAs implicated in processes such as embryonic patterning and cell migration, led to the realization that regulated subcellular trafficking of mRNAs was biologically important [130, 38, 147, 92, 116, 81]. This work led to the model that mRNA transport is a multiple step process involving (1) the formation of a messenger ribonucleoprotein (mRNP) created by the association of a mRNA with RNA-binding proteins (RBPs), (2) the transport of this mRNP to a specific subcellular region, (3) the *in-situ* anchoring of this mRNP and (4) the local translation of the mRNA to produce the required protein [96, 182]. Since then, a broad diversity of mRNAs have been shown to be localized, through different mechanisms, in different cell types, organisms and developmental stages [116, 167, 40, 153, 78, 188, 95, 79, 112, 134, 178, 15]. With the advances of microscopy techniques, genomic approaches and, nowadays, bioinformatics modelling, it is now appreciated that a majority of mRNAs undergo regulated subcellular trafficking [116, 15, 52, 23, 138, 117, 89, 84, 54, 137, 12]. This growing body of evidence has underlined the importance of RNA localization as a key aspect of post-transcriptional gene regulation, while also emphasizing the potentially critical role played by *cis*-acting localization elements in this regulation process.

This chapter is aimed at the informatics-enthusiast biologists with an interest in RNA localization and who are keen to gain insights in the processing and analysis of RNA biology data. While the methods described herein to study *cis*-regulatory motifs, and their instances, may be applied to many aspects of post-transcriptional gene regulation, the examples given are focused on the specifics of RNA localization analysis. Additionally, we do not aim to provide a complete picture of the diverse resources available, but we cover useful examples to help guide the reader.

2. Fundamental aspects of RNA localization

Gene expression is modulated by a wide array of regulatory events that can be mediated by compartment-specific ribonucleoprotein (RNP) complexes. These complexes are involved in all aspects of the mRNA life cycle, from synthesis, processing, editing, nuclear export, cytoplasmic localization, translation and degradation [65]. These events are interdependent and can occur in different locales of a cell, from precise intra-nuclear regions, where nascent transcripts are synthesized, to the dispersion of mature transcripts in specific regions of the cytoplasm or extracellular milieu through secretion. An important facet of post-transcriptional gene regulation is the subcellular transit of mRNA, which may serve a variety of functions mechanistically. Firstly, when combined with localized translation, this process can serve to enrich protein products within a specific compartment of the cell in an efficient manner. Indeed, targeted translation has been proposed as a possible facilitator of the assembly of localized protein complexes [11, 110]. Consistent with this notion, transcripts that encode functionally related proteins can have similar localization patterns, which in turn, are often distinct amongst different functional classes [116, 89, 183]. Secondly, mRNA localization may also be important to avoid the aberrant targeting of proteins products, which could have deleterious effects if they were to accumulate in certain regions of the cell. Interestingly, while RNA localization has been known to have a special relevance in polarized cells, especially neurons, it has also been described to be highly prevalent in a myriad of cells types and appears to be deeply conserved evolutionarily [116, 167, 40, 153, 78, 188, 95, 79, 112, 134, 178, 15].

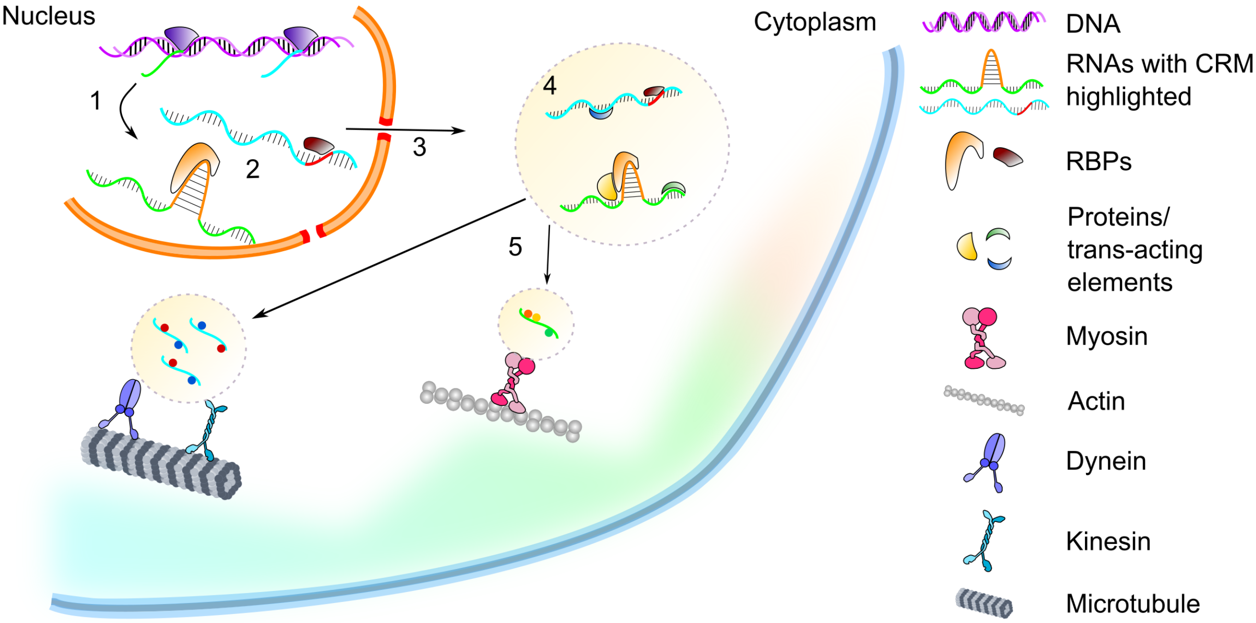


FIGURE 2.1. Distinct mRNA *cis*-regulatory motifs, acting as localization elements, guide the assembly with an RBP to form an mRNP that gets targeted to a specific subcellular region. Schema of RNA localization *cis*-regulatory motifs. Following transcription (1), mRNAs are bound in the nucleus by RBPs (2) that recognize CRMs formed by primary sequence (red) or secondary structure (orange) to form an mRNP. Following export into the cytoplasm via a nuclear pore (3), RBPs and *trans*-acting elements may be added, or removed, to remodel the mRNP and assemble it into RNA granules (4). These RNA granules associate with motor proteins and are transported by cytoskeletal elements towards their target subcellular location (5).

At the molecular level, mRNA localization is coordinated by *cis*-regulatory motifs (CRMs), where one or more instances of these motifs, present within the RNA molecule itself, is referred to as localization elements or zip codes that mediate interaction with *trans*-acting factors (Figure 2.1) [17, 87]. These CRMs are generally defined by their primary sequence and/or secondary/tertiary structure features [176]. CRMs are thought to be recognized by RNA-binding proteins (RBPs) that seed the formation of mRNP complexes necessary for transit. RBPs form a prominent and deeply conserved family of regulatory proteins, which are classified based on their RNA-binding domains (RBDs) [65]. While RBDs often confer binding to single-stranded RNA sequences, some RBP subfamilies mediate binding to structured regions of the target RNA. Different mechanisms may exist in order to target mRNA molecules and to keep them in a translationally repressed state during transport [43]. After nuclear export, an mRNP may acquire or discard a series of trans-regulatory factors (e.g., RBPs, miRNA) that will guide RNA fate by modulating its transport, translation and stability [66]. One of the major mechanisms characterized to achieve subcellular targeting implies the direct trafficking of a localization-competent RNP by association with specific molecular motor proteins that direct transport along cytoskeletal networks in the cytoplasm [17, 175]. Upon reaching its destination, the mRNP can be anchored and remodelled to enable translation to take place [59].

In this section, we survey some of the better documented CRMs implicated in the intracellular trafficking of RNA. For more comprehensive discussion of the functions and biological benefits of intracellular RNA trafficking or the molecular mechanisms involved, please refer to other recent reviews [130, 17, 87, 24].

2.1. *Cis*-regulatory motifs implicated in RNA localization

The characterization of CRMs involved in RNA localization is of great importance to gain insights into the mechanisms of this post-transcriptional regulatory process. CRMs are typically discrete intrinsic elements of information that can function independently from their host mRNA molecule, i.e. they can confer localization activity to a normally non-localized reporter RNA molecule (e.g., *gfp*, *lacz*). As such, CRMs can be identified via structure-function studies, by tracking the subcellular localization of fragments derived from an asymmetrically distributed mRNA, which is achieved by fusing such fragments to a reporter transcript. This chimeric transcript makes it possible to identify which region of an mRNA exhibits CRM activity and whether this component is sufficient for proper RNA targeting. For example, the vasopressin CRM was used to confer dendritic compartmentalization to *a-tubulin* mRNA, normally confined to the cell body [152]. This has allowed the delimitation of a number of CRMs from a wide array of localized mRNAs [87]. In Table 2.1, we compile a summarized list of few CRMs known to be involved in mRNA localization.

Table 2.1. Summary of *cis*-regulatory motifs involved in localized transcript.

mRNA (gene name)	*Cis*-regulatory motifs (localization element names)	Type of motifs	Position in mRNA (5'UTR, CDS, 3'UTR)	Recognized by RNA binding protein	Organisms	References
Arc	DTE	Primary sequence	3'UTR	Multiple	Mammals	[**50**, **51**]
ASH1	E1, E2A, E2B, E3 (43 nt stem-loop)	Secondary structure	3'UTR, CDS, 5'UTR	She2p	Yeast	[**19**, **33**, **34**, **69**, **87**, **88**, **108**, **121**, **122**, **144**, **174**]
Anxa2	G-quadruplex (18 nt region)	Secondary structure	3'UTR	SMN	Mouse	[**157**]
β-actin	54 nt region	Primary sequence	3'UTR	ZBP1, ZBP2	Chicken, Human	[**32**, **53**, **71**, **158**, **102**, **146**, **148**, **158**, **169**, **181**, **185**, **184**, **189**]
Bicoid	Domain III ; stem-loop IV-V ; BLE1	Primary sequence ; secondary structure	3'UTR	Staufen	*Drosophila*	[**56**, **109**, **124**, **125**, **56**, **127**, **171**, **180**]

mRNA (gene name)	*Cis*-regulatory motifs (localization element names)	Type of motifs	Position in mRNA (5'UTR, CDS, 3'UTR)	Recognized by RNA binding protein	Organisms	References
bitesize	BLR	Primary sequence	CDS		Drosophila	[165]
Bsg25D	2 sequence elements	Primary sequence	CDS, 3'UTR		*Drosophila*	[106]
CamKIIα	G-quadruplex (30 nt region)	Secondary structure	3'UTR	Staufen, hnRNP U, PSF, FMRP	Mammals	[21, 46, 82, 104, 133, 140]
Fatvg	FVLE (25 nt region)	Primary sequence	3'UTR		*Xenopus*	[30, 31]
GIRK2	YCAY element	Primary sequence	Introns and 3'UTR	Nova	Mammals	[155]
Gurken	GLE1 (35 nt), GLS (64 nt stem-loop)	Primary sequence ; secondary structure	5' end of ORF	Egalitarian	*Drosophila*	[28, 48, 109, 162, 177]
hairy	Structural element	Secondary structure		Egalitarian	*Drosophila*	[28, 48]
k10	TLS	Primary sequence ; secondary structure	3'UTR	Egalitarian	*Drosophila*	[28, 35, 48, 166]

mRNA (gene name)	*Cis*-regulatory motifs (localization element names)	Type of motifs	Position in mRNA (5'UTR, CDS, 3'UTR)	Recognized by RNA binding protein	Organisms	References
MAP2	DTE (640 nt region)	Secondary structure	3'UTR		Rat	[21, 22]
MBP	A2RE (11 nt region)	Primary sequence	3'UTR	hnRNP A2	Rat	[3, 2, 81, 142]
Nanos	4 redundant regions	Primary sequence	3'UTR	multiples	*Drosophila*	[59, 109, 18, 62, 61, 63]
Neurogranin	CPE (170 nt region)	Primary sequence	3'UTR		Mammals	[140]
orb	280 nt region	Primary sequence	3'UTR		*Drosophila*	[111]
Oskar	EJC, 150 nt region, SOLE	Primary sequence ; secondary structure	5' and 3'UTR	Y14 Staufen, hnRNP A/B	*Drosophila*	[25, 74, 75, 85, 98, 99, 100, 109, 159, 172, 187, 192]
sensorin	66 nt stem-loop region	Secondary strucure	5'UTR			[68, 135]
tau	91 nt region	Primary sequence	3'UTR		Rat	[6, 7, 14]
Vasopressin	ORF fragment	Primary sequence	3'UTR		Rat	[139, 152]

mRNA (gene name)	*Cis*-regulatory motifs (localization element names)	Type of motifs	Position in mRNA (5'UTR, CDS, 3'UTR)	Recognized by RNA binding protein	Organisms	References
Vg1	Vm1, E1-E4 (>300 nt region)	Primary sequence	3'UTR	40KoVe, hnRP U, Vg1RBP/Vera, Kinesin-1, Kinesin-2	*Xenopus*	[20, 42, 44, 45, 60, 107, 136, 141]
Xcat2	MCLE (250 nt region), GGLE (164 nt region)	Primary sequence	3'UTR		*Xenopus*	[103, 190, 191]
XNIF	300 nt region	Primary sequence	5' UTR		*Xenopus*	[37]
Xlsirt	71-81 nt repeats	Primary sequence	3' UTR		*Xenopus*	[36]
Xvelo	75 nt stem-loop	Primary sequence ; secondary structure	5' and 3' end		*Xenopus*	[37]

Interestingly, while many localization CRMs have been mapped to the 3' untranslated region (UTR) of mRNAs, some have also been characterized in the 5' UTR or coding regions [87, 176, 62, 141, 34, 165, 135, 106]. In addition to their variability in distribution across the mRNA molecule, the CRMs can also exhibit heterogeneity both in their sequence length and structure. The relative length of a CRM can vary greatly between transcripts, with some being only a few nucleotides long while others can run over kilobases of sequence [17, 87]. Moreover, as mentioned above, some CRMs are defined by simple primary sequence motifs or stem-loop elements, while others may be composed of more complex structural features, such as G-quadruplexes [87, 176]. For example, transcripts such as *β-actin*, *nanos*, *MBP*, or *vg1* have CRMs in the form of short primary sequence elements [81, 3, 59, 62, 106, 32, 53, 71, 143, 146, 158, 169, 189, 18, 61, 63, 83, 101, 118, 148, 1, 5, 20, 42, 44, 45, 60, 107, 142]. In most cases, the CRMs of these mRNAs are composed of multiple regions that may act sequentially or in concert to direct localization. In particular, *Drosophila nanos* mRNA bears four CRMs spanning a 280-nucleotide region of its 3' UTR, which govern localization in a combinatorial way and ultimately function in the patterning of the anterior-posterior body axis [59, 106, 62, 18, 61, 63, 118, 1]. By contrast, transcripts such as *Anxa2*, *ASH1*, *bicoid*, *CamKIIa*, and *Gurken*, have structural CRMs [79, 34, 1, 19, 33, 69, 121, 122, 157, 174, 127, 126, 56, 125, 171, 180, 109, 21, 27, 133, 140, 162, 177]. In particular, in *Drosophila* oogenesis, the localization of *bicoid* mRNA is driven by a 650-nucleotide segment of its 3' UTR, for which five domains of secondary structure have been shown to cooperate at the various steps of the transport process [127, 126, 56, 124, 125, 171, 180, 109]. Lastly, it is common to observe that multiple elements of different motifs cooperate in a combinatorial fashion and act at distinct steps of the localization process [60, 140]. On the other hand, recurring copies of a single motif can act synergistically to promote individual steps [45, 2]. While these examples convey the diversity of CRM topological organization within localized mRNA molecules, it has been difficult to glean consensus sequence or structure features within families of mRNAs that share similar localization properties. It is important to note that the variability in CRM features might be in part due to the experimental complexity inherent to their study, often requiring painstaking structure-function mapping via sequence trimming and mutagenesis. As such, in many cases, the characterization of minimal regions that define specific CRMs may have been imprecise.

Recent evidence supports the notion that RNAs have similar localization phenotypes in different cell types and species, suggesting that some CRMs might be evolutionarily conserved and operating via similar pathways [15, 27]. For example, strong correlations in the distribution profiles of ∼2500 mRNA orthologs between human and *Drosophila* were recently characterized, with shared general similarities with respect to their UTR and coding sequence lengths [15]. With the development of new experimental approaches to characterize subcellular transcriptomes, such as CeFra-seq or APEX-RIP, and the datasets generated, this establishes the basis for the implementation of bioinformatics approaches to map putative sequence motifs that may drive RNA localization [15, 93, 47].

3. Representation and information content of sequence motifs

RNA sequence motifs, regardless of their biological functions, can be viewed, from a more mathematical point of view, as blocks of regulatory information. This notion that information can be quantitatively measured is important as it allows for the modelling and discovery of additional instances of a given sequence motif. Here, we define a sequence motif as a specific pattern that is common to a set of DNA, RNA, or protein molecules, which are presumed to share particular biological properties or regulatory logic. In the case of RNA localization regulation, the sequence motifs can be the states and patterns that modulate the interaction of a transcript with specific RPBs that direct its targeting to a given subcellular destination. Below, we discuss the various ways by which RNA regulatory motifs can be represented and provide an overview of the different approaches used to map putative regulatory motifs.

There are numerous ways to describe sequence motifs within biological molecules in order to accurately annotate the binding preferences of a given RBP (Figure 2.2). For instance, one of the first biological motifs identified was the TATA box, which was identified by aligning gene promoter elements and transcription start sites and observing an over-representation of that short DNA substring. Therefore, the simplest representation of a motif is stating it as a short sequence. Similarly, if we were interested in the A2RE motif found in RNA targets of the HNRNP A2 protein, we could align multiple sequences containing the motifs and search for a cognate subsequence (Figure 2.2A). The consensus, or canonical, sequence, is obtained by selecting the most frequent nucleotide (or amino acid in the case of proteins) observed at each position (Figure 2.2B). While this is an adequate way of modelling a motif, it is

A

```
>Human_alphaCaMKII_A2RE
GCCAGGAGCCA
>Human_Putative_A2RE-like_sequence
GAGAAGGAGGG
>Human_Neurogranin_A2RE
CUGUUGAGGGC
>Human_ARC_A2RE
GCGGACGAGGA
```

B

```
GCGAAGANGGA
```

C

```
SHSDDSRRSSV
```

D

A	0.00	0.25	0.00	0.50	0.50	0.00	...	0.50
U	0.00	0.25	0.00	0.25	0.25	0.00	...	0.00
G	0.75	0.00	0.75	0.25	0.25	0.75	...	0.25
C	0.25	0.50	0.25	0.00	0.00	0.25	...	0.25

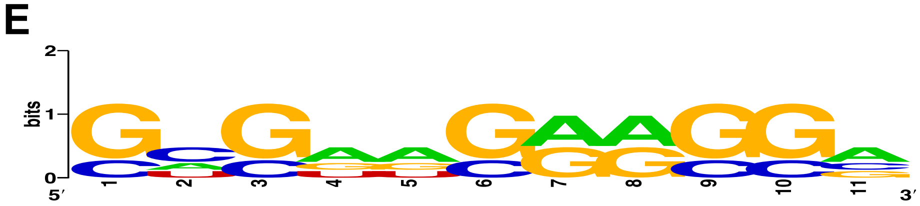


Figure 2.2. Various format can be used to describe the A2RE motif. A. Aligned fasta sequences of the A2RE localization element in 4 different human mRNA. **B.** Consensus motif of (A) showing the most represented nucleotide at each position. Ambiguous nucleotides, where all bases are equally represented are noted as "N". **C.** IUPAC representation of (A) **D.** Truncated position weight matrix (PWM) showing the percentage of each base observed at each position of (A). **E.** Sequence logo, assuming a uniform background nucleotide probability.

insufficient to fully capture its essence or identify other naturally occurring motifs, because RBPs tend to have flexible binding preferences. A motif is usually described as exact (precise), or degenerate (weak), according to the amount of deviations observed between its different instances. For example, the motif bound by HNRNP K is the fixed subsequence GCCGAC, which is considered an exact motif [49]. On the other hand, HNRNP A2 mediates trafficking of RNAs containing the A2RE motif, which display greater diversity and is therefore more degenerate (Figure 2.2A). One way to capture variations among instances is by way of a regular expression. For example, a *cis*-regulatory sequence motif might be formulated as [A][G][U][U or G][A][G], which can be abbreviated by the International Union of Pure and Applied Chemistry (IUPAC) nomenclature as AGU**K**AG, where K is the shorthand for either appearing nucleotide U or G (Figure 2.2C). Most scripting languages handle the search for regular expression (regex) well. Here the search for AGUKAG could simply be encoded as AGU[UG]AG.

Many alternative ways exist to describe a motif, of which the most popular is the position weight matrix (PWM), which is further described here (Figure 2.2D) [**173, 161, 119**]. This is a matrix with four rows (one for each base A, U, G, C,) and width k equal to the number of bases in the motif. A PWM assumes that each position has its own probability distribution over nucleotides, and that the choices of nucleotide at different positions are independent. This means that the columns of a PWM can be thought of as a set of independent multinomial distributions. This allows for the easy calculation of the probability of a subsequence given a PWM, done by simply multiplying each relevant probability. For example, the probability of the sequence $S = CUG$ would be calculated by multiplying the probability of having a "C" in position 1, a "U" in position 2, and a "G" in position 3. Taking the three first positions of the PWM of Figure 2.2D, this would be $0.25 \times 0.25 \times 0.75 = 0.0468$.

The level of specificity (or, inversely, flexibility) of a PWM is an important property that is captured in terms of the information theoretic notions of *information content* and *entropy*. Consider a given column of a PWM, with nucleotide probabilities $P_n(n = A, C, G, U)$. The *Shannon entropy* of a probability distribution is defined as $H(P) = -\sum_{n=1}^{4} P_n log_2 P_n$. This will yield a non-negative value, measured in bits. A bit represents the amount of information necessary to select between two equiprobable options [**128, 163**]. For DNA and RNA, which are each made of 4 bases, this value will be between 0 and $log_2 4 = 2$, whereas for protein motifs it can reach $log_2 20 \cong 4.32$. Since entropy is a measure of uncertainty, when P_n assigns a probability of 1 to a particular nucleotide, the entropy of P_n will be 0 bits, as there is no uncertainty. On the contrary, when all four bases are equiprobable, the entropy will be 2 bits. It requires 2 bits of information to determine which of the four bases occurs at that position. The first 1 bit of decision divides the set by half (e.g. purine vs. pyrimidine), leaving only 2 choices, A/G or C/T. A related notion is that of the *information content* H of a distribution

P (e.g., a column of a PWM) against a certain background distribution B (e.g., the genome-wide nucleotide frequencies), defined as $H(P) = \sum_{n=1}^{4} P_n log_2 \frac{P_n}{B_n}$. The information content of P against B, also known as the Kullback–Leibler divergence between the two distributions, is a measure of how different the two distributions are [73]. Note that when B is a uniform distribution ($B_n = 0.25$ for $n = A, C, G, U$), $H(P{,}B) = 2 - I(P)$.

An elegant way to visually represent a PWM while conveying its information content is called the graphical sequence logo (Figure 2.2E) [168, 164]. In a sequence logo, each position of the motif is represented as a stack of nucleotides, whose total height corresponds to the information content at that position. The height of each nucleotide is proportional to its probability at that position. Therefore, the sequence logo provides a rapid visual portrayal of the conservation and composition of each position in a motif [41].

Knowing the information content of a motif is useful when searching for additional instances as a motif with n bits of information will occur about once in every 2^n bases of random sequence. For example, the six-mer GCCCAC motif of HNRNP K has an information content of 12 bits (6 bases motifs with each 2 bits of information), it is expected that a putative motif instance for this RBP will be observed in an RNA sequence every $2^{12} = 4096$ bases (assume a uniform background), close to what has been described before [151]. By contrast the information content of the more degenerate HNRNP K motif [GC]CCCAC is $log_2(2) + 5 \times log_2(4) = 11$ and would be expected to occur twice as frequently as $2^{11} = 2048$. It is easy to see that GCCCAC or CCCCAC can occur two times more often than GCCCAC alone. However, this frequency of putative motif instances estimation is different than the frequency of actual RBP binding sites, as the former could include identifications of motif instances as false positive binding sites and therefore be much larger than the latter.

4. Algorithms and tools for finding motifs

4.1. Fundamentals of major motif discovery algorithms

One important question in bioinformatics applied to the study of RNA is : how to extract known and unknown regulatory motifs from an ensemble of given sequences ? This question comes in two flavours. *Motif scanning* aims to predict new instances of one or more known motifs in a given sequence. For example, one may use this approach to identify, in a given mRNA sequence, candidate binding sites for an RBP with a known PWM. *De novo motif* *discovery*, on the contrary, aims to determine, from a set of sequences thought to be co-regulated (e.g., identified through a CLIP-seq experiment on a given RBP) or colocalized, the motif(s) that best captures the binding preferences of the RBPs involved.

Motif scanning is simple and fast. When searching for matches given a PWM in a given sequence longer than k, the score of the k-mer starting at every possible positions in the sequence is evaluated as shown above, and high-scoring sites are reported [**96**, **13**, **132**, **160**]. The main issue is to decide on a score threshold above which sites should be reported. Various strategies have been proposed, aiming to maximize the sensitivity of the scan while maintaining an acceptable level of false positives [**96**, **67**, **150**, **120**]. One such approach is illustrated in the next section.

De novo motif discovery typically falls within one of three types : enumerative algorithms, probabilistic optimization, and deterministic optimization.

The first, and perhaps simplest, *de novo* motif discovery approach is designated as an enumerative, or dictionary, approach. In its basic form, it aims at discovering motifs represented as strict consensus sequences. For every possible consensus sequence w of length k (user-defined), these algorithms contrast the number of occurrences of w in a set of positive sequences (e.g. isolated RNA from a subcellular compartment), compared to a control set (unlocalized or random sequences). Enrichment within the positive set is then quantified statistically, to obtain an enrichment p-value. While effective, this approach is based on exact occurrence of specific strings of characters and is often too restrictive for a sensible application in biology where proteins generally bind RNA via degenerate motifs. As such, it is possible that none of the motifs would occur often enough to be observed in a statistically significant fashion. Fortunately, it is possible to generalize the method by being more flexible on the definition of the motifs to search. This alternative approach to the enumeration algorithm can be achieved by either using regular expression or allowing an explicit number of mismatches [**161**, **170**, **149**, **29**, **129**, **105**, **154**, **55**].

A second approach for finding motifs *de novo* is the probabilistic optimization strategy, which aims at inferring a PWM from a set of co-regulated sequences. It is perhaps best exemplified by the Gibbs Sampling algorithm, one of the earlier motif detection methods [**129**, **114**]. It works by first selecting a random position in each sequence and building a PWM from them. It further selects a sequence at random to scan and score all possible sites in this sequence using this predetermined PWM. It can then select a new motif site and update the motif instances and the weight matrix accordingly. Finally, the algorithm iterates over the last steps until a convergence is reached. This algorithm works well to find *de novo* motifs since a real motif is expected to be overrepresented and therefore should be encountered more often when searching at random, which will bias the original weight matrix. Updating the matrix will further lean it towards finding more motifs, until convergence. Since there is a random element involved, one caveat is that while it will always find a motif, there is no certainty that it will always converge towards the same motif.

A third strategy for finding *de novo* motifs, similar to the Gibbs Sampling, makes use of a deterministic optimization of the PWM for describing a motif and the binding probabilities for its associated sites and is referred to as the expectation maximization (EM) strategy [**129**, **9**, **10**, **115**]. EM class algorithms are often used for learning probabilistic models in problems that involve hidden states. In a motif-finding tool, this can be defined as the position(s) where the motif occurs in each sequence. Sequences can have 0, 1 or multiple occurrences of a given motif. This approach has the advantage of simultaneously identifying the position and characteristics of a motif. Briefly, this is achieved by initializing a weight matrix with a single k-mer and a subset of the background frequencies. Then, by scanning the possible space of motifs for each k-mer in the sequence set, it calculates the probability that this k-mer was generated by the motifs from the matrix, rather than by the background distribution. The matrix then gets updated based on these probabilities. A new and refined motif is therefore produced by alternating the calculation of the probability of each site based on the current matrix and calculating the new matrix based on these probabilities. By performing multiple iterations, this algorithm converges towards a maximum value for the motifs' matrix.

The algorithms described above are aimed at identifying *de novo* motifs. It is essential to consider that there is an understated yet important difference between searching for known and *de novo* motifs. While searching for known motifs in a set of sequences can be of great interest, the ultimate result will solely reveal which of these motifs are present, and at which positions in the sequences. Conversely, a *de novo* motif search is done by querying the sequences to identify which motifs are most enriched. This should be taken into consideration as it influences the interpretation of the results. For example, performing a *de novo* search on a set of sequences could result in the proper identification of the GAGAAGGAGGG in the human putative A2RE-like sequence (similar to figure 2.2A). On the other hand, if an unrelated known motif search were performed on these same sequences using a database of genome-wide annotations of transcription factors like JASPAR, hits like the myeloid zinc finger 1 (MZF1), whose canonical motif is GAGGGG, would be identified, perhaps erroneously, despite having a low *p*-value [97]. While biologically counterintuitive, this example shows the limits of motif searches. This demonstrates that motif search can be reduced to local multiple string alignments where context is easily lost at the algorithm level, but should be kept in consideration when performing such analyses.

While the two approaches aim to do different things, as one seeks to annotate sequences with known motifs and the other seeks to discover new motifs, they are often complementary. One decisive advantage of known motif searches is when the ensemble of sequences is limited as the accuracy of *de novo* searches can be reduced in such cases. For example, a *de novo* search is impossible on a single sequence. Otherwise, *de novo* searches are often thought to be less limiting. One common way to palliate this dilemma is to first perform a robust *de novo* motif search and then complete a detailed comparison of these hits to a database of known motifs. Tools to achieve this, like HOMER or the MEME suite, methodologies, and examples are detailed in the next sections.

To add to the complexity of robust identification of CRMs involved in localization, RNA often possesses additional *cis*-regulatory elements found scattered throughout its sequence, which may be needed for other aspects of post-transcriptional regulation, such as splicing and stability regulation. This can make it challenging to assign a specific localization function to a given signature motif. Furthermore, certain RBPs might bind only very short motifs that are quite prevalent in biological sequences (e.g. there might be cases where a CRM necessary and sufficient for localization is only 3 nucleotides long). One major challenge will be to distinguish these real but small motifs, from a background of specious motifs, for example stemming from common repeat elements bearing little information content. In other words, the challenge rapidly becomes to distinguish the true positive among the large number of false positives created by these short motifs that can be found throughout the sequence space.

4.2. Overview of existing computational tools to search for CRMs

Most bioinformatics tools available nowadays tend to be developed through open collaborations and are offered with open source licences, thus allowing the source code to be used, modified or shared under defined terms and conditions, often free of charge, especially for academic uses. They are mostly available only on Linux or Mac OS operating systems and available on platforms such as web-based version control repository hosting services (e.g., GitHub, Bitbucket). Furthermore, as there is often little use for an elaborate graphical user interface, they are predominantly offered as command-line tools (e.g., using Terminal, iterm). This provides the most flexibility and allows for a wide range of customizable options. The running time and memory requirements of these algorithms can be quite high ; therefore, it is often advisable to rely on high-performance computers (HPCs) allowing the use of parallel processing, which are generally accessible through major universities or private vendors (e.g., AWS). To be more accessible, many tools are offered as online databases and web servers, where analyses can be run without any local installation. However, web servers often come with strict limitations regarding the size of the inputs and local installation becomes necessary for larger-scale analyses.

In Table 2.2, we compile a non-exhaustive list of motif scanning and *de novo* motif discovery tools available to the community. These tools can be used, for example, to identify motifs that are likely to be candidates for potential regulatory roles in modulating different features of the RNA life cycle, including localization control. Dissecting the exact functions of a particular motif therefore requires the implementation of biological assays to assess the impact of the motif on RNA processing or activity (e.g., the use of reporter assays and site-specific mutagenesis to disrupt candidate motifs).

TABLE 2.2. A selection of motifs databases, web servers, and search algorithms.

Tools	Type	URL	Motifs types	References
ATtRAC	Database ; Webserver	https://attract.cnic.es	Primary sequence	[67]
CISBP-RNA	Database ; Webserver	http://cisbp-rna.ccbr.utoronto.ca	Primary sequence	[156]
Deepbind	Database ; stand-alone	http://tools.genes.toronto.edu/deepbind/	Primary sequence	[4]
Gibbs sampling	Algorithm		Primary sequence	[114]
GRAPHprot	Stand-alone	http://www.bioinf.uni-freiburg.de/Software/GraphProt/	Primary sequence ; secondary structure	[131]
Homer	Stand-alone	http://homer.ucsd.edu/homer/index.html	Primary sequence	[76]
LESMoN	Stand-alone	http://cs.mcgill.ca/~blanchem/LESMoN/	Primary sequence	[113]
MatrixREDUCE	Stand-alone	https://systemsbiology.columbia.edu/matrixreduce	Primary sequence	[57, 58, 179]
MEME	Stand-alone	http://meme-suite.org	Primary sequence	[9, 8]
MEMERIS	Stand-alone	http://www.bioinf.uni-freiburg.de/~hiller/MEMERIS/	Primary sequence	[80]

Tools	Type	URL	Motifs types	References
MotifMap-RNA	Database ; Webserver	http://motifmap-rna.ics.uci.edu	Primary sequence	[120]
oRNAment	Database	http://rnabiology.ircm.qc.ca/oRNAment/	Primary sequence	[16]
RBPDB	Database ; Webserver	http://rbpdb.ccbr.utoronto.ca	Primary sequence	[39]
RBPmap	Webserver	http://rbpmap.technion.ac.il	Primary sequence	[150]
RCK	Stand-alone	http://cb.csail.mit.edu/cb/rck/	Primary sequence	[145]
RNAcontext	Webserver ; Stand-alone	http://www.cs.toronto.edu/~hilal/rnacontext/	Primary sequence ; secondary structure	[94]
ssHMM	Stand-alone	https://github.molgen.mpg.de/heller/ssHMM	Primary sequence ; secondary structure	[77]

As there are an ever-increasing number of biologically validated motifs identified, databases are a valuable first place to search. The RNA-Binding Protein DataBase (RBPDB) is a large, manually curated, database grouping published observations of experimentally defined motifs [39]. This database has the advantage of allowing one to search by RNA-binding domain (RBD), by species or to use it as a web server to scan an RNA sequence for putative RBP binding sites. Along the same line, the Catalog of Inferred Sequence Binding Proteins of RNA (CISBP-RNA) is a database of RBP motifs and specificities derived from the impressive work compiling the results of systematic RNAcompete experiments. RNAcompete is a method through which the consensus binding motifs of ∼300 RBPs were characterized through an *in vitro* selection assay in which purified RBPs were incubated with a random RNA pool, followed by the profiling of the RNA molecules selectively bound by the RBP [156].

A separate database that extends RBPDB and CISBP-RNA, and which has rapidly established itself as a gold standard, is the 'A daTabase of experimentally validated RNA binding proteins and AssoCiated moTifs' (ATtRACT) resource [67]. This database currently compiles information on 370 RBPs and 1583 manually curated consensus RBP binding motifs, in addition to having integrated updates and information about protein-RNA complexes as described in the Protein Data Bank (PDB) database [64]. As with other databases, ATtRACT also provides the capacity to search for motifs in target sequences. Finally, MotifMap-RNA is another database and web server that expands on RBPDB/CISBP-RNA and allows for genome-wide motif searches [120]. While most databases described also offer web server capabilities to scan sequences and search for potential motifs, these tend to be limited. RBPmap is a web server that improves upon the scanning of sequences. Building on motifs compiled in all the previously mentioned databases, and with the possibility to input additional user-defined motifs, this algorithm can be quite efficient in predicting and mapping binding sites [150].

In order to gain more insights into CRMs, *de novo* motif search tools are a great complement to established motif databases. These algorithms use only the sequence, and do not consider structure, when calling a motif. A first suite of tools for *de novo* motif discovery is the Hypergeometric Optimization of Motif EnRichment (HOMER) [76]. HOMER is a powerful tool that identifies motifs by looking for subsequences with differential enrichment between two sets of sequences. While it is advised to use a background of meaningful sequences (e.g., localized vs. non-localized), the background set can be simply random sequences. Interestingly, HOMER will also make some attempts to compare the motifs observed to a database of known motifs and will identify similarities. When only one group of sequences is available, Multiple EM for Motif Elicitation (MEME) is perhaps best suited. It is a suite of tools that implement multiple motif-finding algorithms, each with their own specificities for sequence search and motif discovery, analysis, and comparison. It builds upon the EM algorithm described in section 4.1 [**9**, **8**]. Alternatively, MatrixREDUCE is a motif discovery algorithm that was originally designed to infer the binding specificity of transcription factors from microarray data, but can also be applied to the study of RNA sequence motifs [**57**, **58**, **179**]. Local Enrichment of Sequence Motifs in biological Networks (LESMoN) takes a different approach by being an enumerative motif discovery algorithm that integrates gene set enrichment and biological network analysis [**113**].

While primary sequence is a critical component of *cis* elements, RNA secondary and tertiary structures can also be key features that can influence the binding to *trans*-regulatory machineries. Indeed, depending on the type of RNA binding domain (RBD) they contain, RBPs can bind RNA based on primary sequence or structural motifs, although the most abundant classes of RBPs tend to bind specific primary sequence motifs [**65**, **156**]. As such, some regulatory motif prediction algorithms are taking structural prediction information into account. For example, the MEMERIS algorithm is built on the same principle as MEME but searches for RNA motifs enriched in any type of single-stranded regions (e.g., the loop of a hairpin). This has been shown to improve RNA binding site predictions [**80**]. Expanding on the idea that approaches making use of RNA sequence and structure can be used for better motifs predictions, the RNAcontext tool integrates predictions on whether a nucleotide is paired, in a hairpin loop, or unstructured region, to help define putative regulatory elements [**94**].

Machine-learning frameworks are proving to be quite efficient for identifying RBP binding preferences. In that category, GRAPHprot is able to detect motifs by taking into consideration both sequence and structure [**131**]. Alternatively, DeepBind, a state of the art in sequence models, only considers sequence and not structure, but has been shown to perform better than GRAPHprot [**4**]. RCK is an elegant machine learning algorithm that takes into account both sequence and structure and has established itself as an efficient and scalable tool for robust motif discovery [145]. Another tool named sequence-structure hidden Markov model (ssHMM) searches for motif based on a statistical model named hidden Markov model (HMM) and Gibbs sampling, which it performs while integrating the sequence and structure preference of an RBP [77].

Some algorithms have also been developed specifically to provide answers on localization. DeepLncRNA is a machine-learning algorithm that predicts the subcellular localization of lncRNA considering only its sequences [72]. Finally, RNATracker is a novel algorithm that takes advantage of deep neural network using both sequence and structural information to infer subcellular distribution of transcripts [186].

Individually, the results obtained from these databases, web servers, and stand-alone algorithms must be analyzed with great caution, as they are likely to produce a very large number of false positive predictions. This is unavoidable, given the low information content of certain motifs. Cross-validation of results from multiple tools, detailed literature consideration and experimental validation via mutational analysis or reporter assays is therefore of the utmost importance.

5. Examples of motif discovery applications

In order to exemplify the most important concepts addressed in this chapter, we performed different known and *de novo* motif searches on the complete human coding transcriptome (i.e. all portions of an mRNA) and between two sets of sequences that were observed to be localized to either the nucleus or the cytoplasm of human HepG2 cells [15, 117].

We first sought to assess the general distribution of motifs for 70 RBPs (listed in Figure 2.3A) for which PWMs were obtained by RNAcompete [156]. Sites were identified using the PWM scanning approach described in Section 4.1. For each PWM, we recorded sites whose score was greater than a certain PWM-specific threshold T, where T was established as the 99th percentile of the score distribution for that PWM. For example, for a PWM of length 4, we would calculate the score of all 256 possible 4-mers and kept only the two highest scores as a threshold. As RNAcompete motifs were designed for preferentially binding single strand RNA, we further reduce the list of putative motifs by selecting for those predicted to lie within single-stranded regions of each mRNA. For this we used RNAplfold, a gold

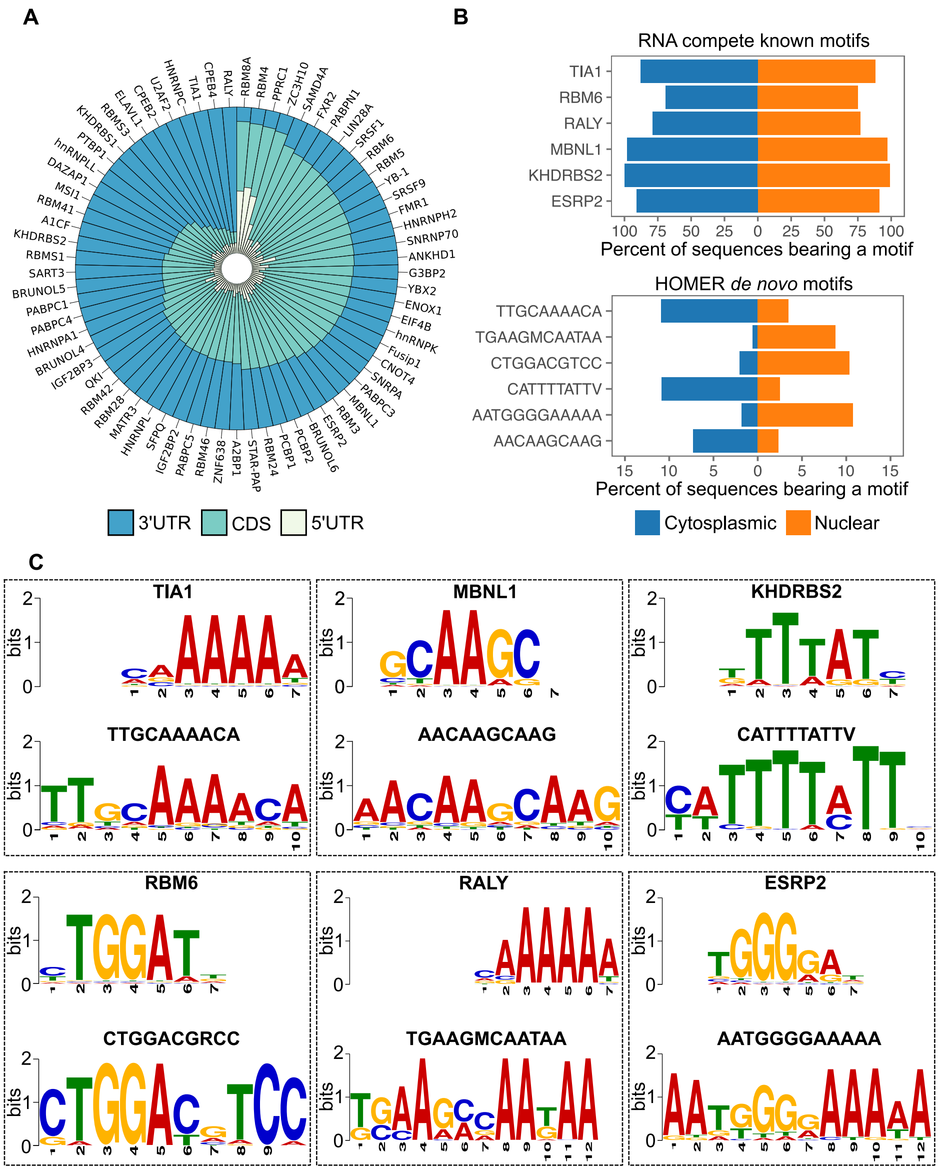


FIGURE 2.3. Global overview of known and de novo motifs and their putative role in RNA localization. A. Circos plot showing the relative regionalization towards the 5' UTR, coding sequence (CDS), and 3' UTR of 70 known motifs from RNAcompete. **B.** Histogram showing the percent of localized sequences, either enriched in the cytoplasm or the nucleus, harbouring a known motif from an RNAcompete experiment (upper panel) and from a *de novo* motifs search using HOMER (lower panel) in their 3' UTR. **C.** Sequence logos comparing the known motif from an RNAcompete experiment to the *de novo* motif identified by HOMER in the 3' UTR of its localized sequence, in the normal strand or its reverse complement.

standard RNA folding algorithm that calculates locally stable secondary structures and outputs base pairing probabilities for each nucleotide of an RNA of interest [123]. We retained only predicted sites located in regions with a higher than 90% probability of being unpaired for each nucleotide of the k-mer. This provided us with a comprehensive list of predicted binding sites for all 70 RBPs in all 179,236 annotated human mRNA transcripts. As shown in Figure 2.3A, each of these 70 sets of putative CRMs exhibited variable distribution profiles across the 5'UTR, coding region and 3' UTR of mRNAs. For example, target motifs of the CPEB4 protein are predominantly found in 3' UTRs, consistent with its previously established binding preferences [1].

Having a list of transcripts and their embedded motif instances, we next sought to determine whether any of these motifs could be correlated with localization. For this we took advantage of a recently published list of asymmetrically distributed mRNAs, determined using subcellular fraction and RNA sequencing, where we could cluster mRNAs based on their degree of enrichment within the nucleus or cytoplasm of HepG2 cells [15]. Starting with a naïve approach, we enumerated the percent of sequences bearing known RNAcompete motifs, within the nucleus and cytoplasm. As shown in Figure 2.3B (upper panel), the top 6 interrogated motifs tended to be roughly equally represented within nuclear and cytoplasmic mRNA populations. We therefore executed *de novo* searches using HOMER, on the same set of mRNA sequences. By doing so, it becomes apparent that specific subsequences are enriched in one group or the other (Figure 2.3B, lower panel). Strikingly, all the *de novo* motifs identified are longer than the ones previously defined using the RNAcompete *in vitro* pipeline. Interestingly, when we compared the known motifs with those found *de novo* using Tomtom, a motif comparison tool available in the MEME suite, we observed significant similarities between the two sets of results [73]. Indeed, these 6 motifs of length 7 derived from RNAcompete data can be embedded in the longer motifs identified by HOMER (Figure 2.3C). We can conclude from this that short motifs may not contain enough information to differentiate sets of RNA with distinctive biological features or behaviours. However, supplementing such analyses with *de novo* motif predictions strategies offers a promising avenue to identify biologically relevant CRM involved in localization.

6. Conclusion

As outlined in this review, mRNA localization has been shown to be a key layer of post-transcriptional gene regulation that impacts a wide array of biological processes. The targeting of a transcript to a precise subcellular location involves a complex coaction between a variety of CRMs, RBPs and additional factors to form an mRNP. Nevertheless, there is much to be discovered regarding the necessary and sufficient region of each mRNA dictating their subcellular distribution. Mathematical tools, such as information content and entropy, have been adapted to address the representation of biological motifs, like PWMs and sequence logos. This has laid the groundwork for the implementation of computational procedures, such as motif enumeration, that may help in deciphering and classifying individual CRMs. Already, a variety of programs exist that use these tools and procedures with the aim of filtering true motifs within a given subset of sequences. We demonstrated that it was possible to identify putative motifs involved in localization through the execution of these programs on sets of asymmetrically distributed transcripts. By combining the resulting motif inferences with classical molecular biology experiments, such as reporter assays, it is but a question of time before we have a more comprehensive knowledge of the regulatory code driving mRNA subcellular localization.

Références

[1] T. Afroz, L. Skrisovska, E. Belloc, J. Guillen-Boixet, R. Mendez, and F. H. Allain, *A fly trap mechanism provides sequence-specific RNA recognition by CPEB proteins*, Genes and Development, 28 (2014), pp. 1498–514.

[2] K. Ainger, D. Avossa, A. S. Diana, C. Barry, E. Barbarese, and J. H. Carson, *Transport and localization elements in myelin basic protein mRNA*, The Journal of Cell Biology, 138 (1997), pp. 1077–87.

[3] K. Ainger, D. Avossa, F. Morgan, S. J. Hill, C. Barry, E. Barbarese, and J. H. Carson, *Transport and localization of exogenous myelin basic protein mRNA microinjected into oligodendrocytes*, The Journal of Cell Biology, 123 (1993), pp. 431–41.

[4] B. Alipanahi, A. Delong, M. T. Weirauch, and B. J. Frey, *Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning*, Nature Biotechnology, 33 (2015), pp. 831–9.

[5] L. Allen, M. Kloc, and L. D. Etkin, *Identification and characterization of the Xlsirt cis-acting RNA localization element*, Differentiation, 71 (2003), pp. 311–21.

[6] G. E. Aranda-Abreu, L. Behar, S. Chung, H. Furmeaux, and I. Ginzburg, *Embryonic lethal abnormal vision-like RNA-binding proteins regulate neurite outgrowth and tau expression in PC12 cells*, Journal of Neuroscience, 19 (1999), pp. 6907–17.

[7] S. Aronov, G. Aranda, and I. Ginzburg, *Axonal tau mRNA localization coincides with tau protein in living neuronal cells and depends on axonal targeting signal*, Journal of Neuroscience, 21 (2001), pp. 6577–87.

[8] T. L. Bailey, M. Boden, F. A. Buske, M. Frith, C. E. Grant, L. Clementi, J. Ren, W. W. Li, and W. S. Noble, *MEME Suite : tools for motif discovery and searching*, Nucleic Acids Research, 37 (2009), pp. W202–8.

[9] T. L. Bailey and C. Elkan, *Fitting a mixture model by expectation maximization to discover motifs in bipolymers*, Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology, (1994), pp. 28–36.

[10] T. L. Bailey and C. Elkan, *Unsupervised learning of multiple motifs in biopolymers using expectation maximization*, Machine Learning, 21 (1995), pp. 51–80.

[11] N. N. Batada, L. A. Shepp, and D. O. Siegmund, *Stochastic model of protein–protein interaction : Why signaling proteins need to be colocalized*, Proceedings of the National Academy of Sciences of the United States of America, 101 (2004), pp. 6445–9.

[12] M. Batish, P. van den Bogaard, F. R. Kramer, and S. Tyagi, *Neuronal mRNAs travel singly into dendrites*, Proceedings of the National Academy of Sciences of the United States of America, 109 (2012), pp. 4645–50.

[13] M. Beckstette, R. Homann, R. Giegerich, and S. Kurtz, *Fast index based algorithms and software for matching position specific scoring matrices*, BMC Bioinformatics, 7 (2006), p. 389.

[14] L. Behar, R. Marx, E. Sadot, J. Barg, and I. Ginzburg, *cis-acting signals and trans-acting proteins are involved in tau mRNA targeting into neurites of differentiating neuronal cells*, International Journal of Developmental Neuroscience, 13 (1995), pp. 113–27.

[15] L. P. Benoit Bouvrette, N. A. L. Cody, J. Bergalet, F. A. Lefebvre, C. Diot, X. Wang, M. Blanchette, and E. Lecuyer, *CeFra-seq reveals broad asymmetric mRNA and noncoding RNA distribution profiles in Drosophila and human cells*, RNA, 24 (2018), pp. 98–113.

[16] L. P. Benoit Bouvrette, S. Bovaird, M. Blanchette, and E. Lécuyer, *oRNAment : a database of putative RNA binding protein target sites in the transcriptomes of model species*, Nucleic Acids Research, 48 (2019), pp. D166–73.

[17] J. Bergalet and E. Lécuyer, *The functions and regulatory principles of mRNA intracellular trafficking*, Systems Biology of RNA Binding Proteins, 825 (2014), pp. 57–96.

[18] S. E. Bergsten, T. Huang, S. Chatterjee, and E. R. Gavis, *Recognition and long-range interactions of a minimal nanos RNA localization signal element*, Development, 128 (2001), pp. 427–35.

[19] E. Bertrand, P. Chartrand, M. Schaefer, S. M. Shenoy, R. H. Singer, and R. M. Long, *Localization of ASH1 mRNA particles in living yeast*, Molecular Cell, 2 (1998), pp. 437–45.

[20] J. N. Betley, B. Heinrich, I. Vernos, C. Sardet, F. Prodon, and J. O. Deshler, *Kinesin II mediates Vg1 mRNA transport in Xenopus oocytes*, Current Biology, 14 (2004), pp. 219–24.

[21] A. Blichenberg, M. Rehbein, R. Müller, C. C. Garner, D. Richter, and S. Kindler, *Identification of a cis-acting dendritic targeting element in the mRNA encoding the asubunit of Ca2+/calmodulin-dependent protein kinase II*, The European Journal of Neuroscience, 13 (2001), pp. 1881–8.

[22] A. Blichenberg, B. Schwanke, M. Rehbein, C. C. Garner, D. Richter, and S. Kindler, *Identification of a cis-acting dendritic targeting element in MAP2 mRNAs*, Journal of Neuroscience, 19 (1999), pp. 8818–29.

[23] M. D. Blower, E. Feric, K. Weis, and R. J. Heald, *Genome-wide analysis demonstrates conserved localization of messenger RNAs to mitotic microtubules*, Journal of Cell Biology, 179 (2007), pp. 1365–73.

[24] S. Bovaird, D. Patel, J. A. Padilla, and E. Lecuyer, *Biological functions, regulatory mechanisms, and disease relevance of RNA localization pathways*, FEBS Letters, 592 (2018), pp. 2948–72.

[25] R. P. Brendza, L. R. Serbus, J. B. Duffy, and W. M. Saxton, *A function for kinesin I in the posterior transport of oskar mRNA and Staufen protein*, Science, 289 (2000), pp. 2120–2.

[26] S. Brenner, F. Jacob, and M. Meselson, *An unstable intermediate carrying information from genes to ribosomes for protein synthesis*, Nature, 190 (1961), pp. 576–81.

[27] S. L. Bullock and D. Ish-Horowicz, *Conserved signals and machinery for RNA transport in Drosophila oogenesis and embryogenesis*, Nature, 414 (2001), pp. 611–6.

[28] S. L. Bullock, I. Ringel, D. Ish-Horowicz, and P. J. Lukavsky, *A'-form RNA helices are required for cytoplasmic mRNA transport in Drosophila*, Nature Structural and Molecular Biology, 17 (2010), pp. 703–709.

[29] J. M. Carlson, A. Chakravarty, C. E. DeZiel, and R. H. Gross, *SCOPE : a web server for practical de novo motif discovery*, Nucleic Acids Research, 35 (2007), pp. W259–64.

[30] A. P. Chan, M. Kloc, S. Bilinski, and L. D. Etkin, *The vegetally localized mRNA fatvg is associated with the germ plasm in the early embryo and is later expressed in the fat body*, Mechanisms of Development, 100 (2001), pp. 137–40.

[31] A. P. Chan, M. Kloc, and L. D. Etkin, *fatvg encodes a new localized RNA that uses a 25-nucleotide element (FVLE1) to localize to the vegetal cortex of Xenopus oocytes*, Development, 126 (1999), pp. 4943–53.

[32] J. A. Chao, Y. Patskovsky, V. Patel, M. Levy, S. C. Almo, and R. Singer, *ZBP1 recognition of β-actin zipcode induces RNA looping*, Genes and Development, 24 (2010), pp. 148–58.

[33] P. Chartrand, X. H. Meng, S. Huttelmaier, D. Donato, and R. H. Singer, *Asymmetric sorting of ash1p in yeast results from inhibition of translation by localization elements in the mRNA*, Molecular Cell, 10 (2002), pp. 1319–30.

[34] P. Chartrand, X. H. Meng, R. H. Singer, and R. M. Long, *Structural elements required for the localization of ASH1 mRNA and of a green fluorescent protein reporter particle in vivo*, Current Biology, 9 (1999), pp. 333–6.

[35] H. K. Cheung, T. L. Serano, and R. S. Cohen, *Evidence for a highly selective RNA transport system and its role in establishing the dorsoventral axis of the Drosophila egg*, Development, 114 (1992), pp. 653–61.

[36] M. Claussen and T. Pieler, *Xvelo1 uses a novel 75-nucleotide signal sequence that drives vegetal localization along the late pathway in Xenopus oocytes*, Developmental Biology, 266 (2004), pp. 270–84.

[37] M. Claussen, K. Horvay, and T. Pieler, *Evidence for overlapping, but not identical, protein machineries operating in vegetal RNA localization along early and late pathways in Xenopus oocytes*, Development, 131 (2004), pp. 4263–73.

[38] N. A. Cody, C. Iampietro, and E. Lecuyer, *The many functions of mRNA localization during normal development and disease : From pillar to post*, Wiley Interdisciplinary Reviews-Developmental Biology, 2 (2013), pp. 781–96.

[39] K. B. Cook, H. Kazan, K. Zuberi, Q. Morris, and T. R. Hughes, *RBPDB : a database of RNA-binding specificities*, Nucleic Acids Research, 39 (2010), pp. D301–8.

[40] A. J. Crofts, H. Washida, T. W. Okita, M. Ogawa, T. Kumamaru, and H. Satoh, *Targeting of proteins to endoplasmic reticulum-derived compartments in plants. The importance of RNA localization*, Plant Physiology, 136 (2004), pp. 3414–9.

[41] G. E. Crooks, G. Hon, J. M. Chandonia, and S. E. Brenner, *WebLogo : a sequence logo generator*, Genome Research, 14 (2004), pp. 1188–90.

[42] K. Czaplinski, T. Köcher, M. Schelder, A. Segref, M. Wilm, and I. W. Mattaj, *Identification of 40LoVe, a Xenopus hnRNP D family protein involved in localizing a TGF-β-related mRNA during oogenesis*, Developmental Cell, 8 (2005), pp. 505–15.

[43] R. Dahm and M. Kiebler, *Cell biology : silenced RNA on the move*, Nature, 438 (2005), pp. 432–5.

[44] J. O. Deshler, M. I. Highett, T. Abramson, and B. J. Schnapp, *A highly conserved RNA-binding protein for cytoplasmic mRNA localization in vertebrates*, Current Biology, 8 (1998), pp. 489–96.

[45] J. O. Deshler, M. I. Highett, and B. J. Schnapp, *Localization of Xenopus Vg1 mRNA by Vera protein and the endoplasmic reticulum*, Science, 276 (1997), pp. 1128–31.

[46] J. B. Dictenberg, S. A. Swanger, L. N. Antar, R. H. Singer, and G. J. Bassell, *A direct role for FMRP in activity-dependent dendritic mRNA transport links filopodial-spine morphogenesis to fragile X syndrome*, Developmental Cell, 14 (2008), pp. 926–39.

[47] M. Diehn, R. Bhattacharya, D. Botstein, and P. O. Brown, *Genome-scale identification of membrane-associated human mRNAs*, PLoS Genetics, 2 (2006), p. e11.

[48] M. Dienstbier, F. Boehl, X. Li, and S. L. Bullock, *Egalitarian is a selective RNA-binding protein linking mRNA localization signals to the dynein motor*, Genes and Development, 23 (2009), pp. 1546–58.

[49] D. Dominguez, P. Freese, M. S. Alexis, A. Su, M. Hochman, T. Palden, C. Bazile, N. J. Lambert, E. L. Van Nostrand, G. A. Pratt, G. W. Yeo, B. R. Graveley, and C. B. Burge, *Sequence, structure, and context preferences of human RNA binding proteins*, Molecular Cell, 70 (2018), pp. 854–867 e9.

[50] J. L. Dynes and O. Steward, *Dynamics of bidirectional transport of arc mrna in neuronal dendrites*, The Journal of Comparative Neurology, 500 (2007), pp. 433–447.

[51] ——, *Arc mRNA docks precisely at the base of individual dendritic spines indicating the existence of a specialized microdomain for synapse-specific mRNA translation*, The Journal of Comparative Neurology, 520 (2012), pp. 3105–19.

[52] J. Eberwine, B. Belt, J. E. Kacharmina, and K. Miyashiro, *Analysis of subcellularly localized mRNAs using in situ hybridization, mRNA amplification, and expression profiling*, Neurochemical Research, 27 (2002), pp. 1065–77.

[53] K. L. Farina, S. Hüttelmaier, K. Musunuru, R. Darnell, and R. Singer, *Two ZBP1 KH domains facilitate β-actin mRNA localization, granule formation, and cytoskeletal attachment*, The Journal of Cell Biology, 160 (2003), pp. 77–87.

[54] S. Farris, G. Lewandowski, C. D. Cox, and O. Steward, *Selective localization of arc mRNA in dendrites involves activity- and translation-dependent mRNA degradation*, The Journal of Neuroscience, 34 (2014), pp. 4481–93.

[55] F. Fauteux, M. Blanchette, and M. V. Stromvik, *Seeder : discriminative seeding DNA motif discovery*, Bioinformatics, 24 (2008), pp. 2303–7.

[56] D. Ferrandon, I. Koch, E. Westhof, and C. Nusslein-Volhar, *RNA–RNA interaction is required for the formation of specific bicoid mRNA 3' UTR–STAUFEN ribonucleoprotein particles*, The EMBO Journal, 16 (1997), pp. 1751–997.

[57] B. C. Foat, S. S. Houshmandi, W. M. Olivas, and H. J. Bussemaker, *Profiling condition-specific, genome-wide regulation of mRNA stability in yeast*, Proceedings of the National Academy of Sciences of the United States of America, 102 (2005), pp. 17675–80.

[58] B. C. Foat, A. V. Morozov, and H. J. Bussemaker, *Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE*, Bioinformatics, 22 (2006), pp. e141–9.

[59] K. M. Forrest and E. R. Gavis, *Live imaging of endogenous RNA reveals a diffusion and entrapment mechanism for nanos mRNA localization in Drosophila*, Current Biology, 13 (2003), pp. 1159–68.

[60] D. Gautreau, C. A. Cote, and K. L. Mowry, *Two copies of a subelement from the Vg1 RNA localization sequence are sufficient to direct vegetal localization in Xenopus oocytes*, Development, 124 (1997), pp. 5013–20.

[61] E. R. Gavis, D. Curtis, and R. Lehmann, *Identification of cis-acting sequences that control nanos RNA localization*, Developmental Biology, 176 (1996), pp. 36–50.

[62] E. R. Gavis and R. Lehmann, *Localization of nanos RNA controls embryonic polarity*, Cell, 71 (1992), pp. 301–13.

[63] E. R. Gavis, L. Lunsford, S. E. Bergsten, and R. Lehmann, *A conserved 90 nucleotide element mediates translational repression of nanos RNA*, Development, 122 (1996), pp. 2791–800.

[64] C. Gene Ontology, *Gene Ontology Consortium : going forward*, Nucleic Acids Research, 43 (2015), pp. D1049–56.

[65] S. Gerstberger, M. Hafner, and T. Tuschl, *A census of human RNA-binding proteins*, Nature Reviews Genetics, 15 (2014), pp. 829–45.

[66] C. Giorgi and M. J. Moore, *The nuclear nurture and cytoplasmic nature of localized mRNPs*, Seminars in Cell and Developmental Biology, 18 (2007), pp. 186–93.

[67] G. Giudice, F. Sánchez-Cabo, C. Torroja, and E. Lara-Pezzi, *ATtRACT—a database of RNA-binding proteins and associated motifs*, Database, (2016).

[68] C. Gomes, T. T. Merianda, S. Lee, S. Yoo, and J. L. Twiss, *Molecular determinants of the axonal mRNA transcriptome*, Developmental Neurobiology, 74 (2014), pp. 218–232.

[69] I. Gonzalez, S. B. Buonomo, K. Nasmyth, and U. von Ahsen, *ASH1 mRNA localization in yeast involves multiple secondary structural elements and Ash1 protein translation*, Current Biology, 9 (1999), pp. 337–40.

[70] F. Gros, H. Hiatt, W. Gilbert, C. G. Kurland, R. W. Risebrough, and J. D. Watson, *Unstable ribonucleic acid revealed by pulse labelling of Escherichia coli*, Nature, 190 (1961), pp. 581–85.

[71] W. Gu, F. Pan, H. Zhang, G. J. Bassell, and R. H. J. Singer, *A predominantly nuclear protein affecting cytoplasmic localization of β-actin mRNA in fibroblasts and neurons*, Journal of Cell Biology, 156 (2002), pp. 41–51.

[72] B. L. Gudenas and L. Wang, *Prediction of LncRNA subcellular localization with deep learning from sequence features*, Scientific Report, 8 (2018), p. 16385.

[73] S. Gupta, J. A. Stamatoyannopoulos, T. L. Bailey, and W. S. Noble, *Quantifying similarity between motifs*, Genome Biology, 8 (2007), p. R24.

[74] O. Hachet and A. Ephrussi, *Drosophila Y14 shuttles to the posterior of the oocyte and is required for oskar mRNA transport*, Current Biology, 11 (2001), pp. 1666–1674.

[75] O. Hachet and A. Ephrussi, *Splicing of oskar RNA in the nucleus is coupled to its cytoplasmic localization*, Nature, 428 (2004), pp. 959–63.

[76] S. Heinz, C. Benner, N. Spann, E. Bertolino, Y. C. Lin, P. Lasslo, J. X. Cheng, C. Murre, H. Singh, and K. K. Glass, *Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell indentities*, Molecular Cell, 38 (2010), pp. 576–89.

[77] D. Heller, R. Krestel, U. Ohler, M. Vingron, and A. Marsico, *ssHMM : extracting intuitive sequence-structure motifs from high-throughput RNA-binding protein data*, Nucleic Acids Research, 45 (2017), pp. 11004–18.

[78] J. J. Henry, K. J. Perry, L. Fukui, and N. Alvi, *Differential localization of mRNAs during early development in the mollusc, Crepidula fornicata*, Integrative and Comparative Biology, 50 (2010), pp. 720–33.

[79] R. G. Heym and D. Niessing, *Principles of mRNA transport in yeast*, Cellular and Molecular Life Sciences, 69 (2012), pp. 1843–53.

[80] M. Hiller, R. Pudimat, A. Busch, and R. Backlofen, *Using RNA secondary structures to guide sequence motif finding towards single-stranded regions*, Nucleic Acids Research, 34 (2006).

[81] K. S. Hoek, G. J. Kidd, J. H. Carson, and R. Smith, *hnRNP A2 selectively binds the cytoplasmic transport sequence of myelin basic protein mRNA*, Biochemistry, 37 (1998), pp. 7021–9.

[82] Y. S. Huang, J. H. Carson, E. Barbarese, and D. Richter, *Facilitation of dendritic mRNA transport by CPEB*, Genes and Developments, 17 (2003), p. 638–53.

[83] S. Huttelmaier, D. Zenklusen, M. Lederer, J. Dictenberg, M. Lorenz, X. Meng, G. J. Bassell, J. Condeelis, and R. H. Singer, *Spatial regulation of β-actin translation by Src-dependent phosphorylation of ZBP1*, Nature, 438 (2005), pp. 512–5.

[84] S. Hutten, T. Sharangdhar, M. Kiebler, S. Hutten, T. Sharangdhar, and M. Kiebler, *Unmasking the messenger*, RNA Biology, 11 (2014), pp. 992–997.

[85] J.-R. Huynh, T. P. Munro, K. Smith-Litière, J.-A. Lepesant, and D. St Johnston, *The Drosophila hnRNPA/B homolog, Hrp48, is specifically required for a distinct step in osk mRNA localization*, Developmental Cell, 6 (2004), pp. 625–35.

[86] F. Jacob and J. Monod, *Genetic regulatory mechanisms in the synthesis of proteins*, Journal of Molecular Biology, 3 (1961), pp. 318–356.

[87] A. Jambhekar and J. L. DeRisi, *Cis-acting determinants of asymmetric, cytoplasmic RNA transport*, RNA, 12 (2007), pp. 625–42.

[88] A. Jambhekar, K. McDermott, S. Sorber, K. A. Shepard, R. D. Vale, P. Takizava, and J. L. DeRisi, *Unbiased selection of localization elements reveals cis-acting determinants of mRNA bud localization in Saccharomyces cerevisiae*, of the United StatesProceedings of the National Academy of Sciences of the United States of Americas, 102 (2005), pp. 18005–18010.

[89] H. Jambor, V. Surendranath, A. T. Kalinka, P. Mejstrik, S. Saalfeld, and P. Tomancak, *Systematic imaging reveals features and changing localization of mRNAs in Drosophila development*, eLife, 4 (2015).

[90] R. Jeener and D. Szafarz, *Relation between the rate of renewal and the intracellular localization of ribonucleic acid*, Archives of Biochemistry and Biophysics, 26 (1950), pp. 54–67.

[91] W. R. Jeffery, C. R. Tomlinson, and R. D. Brodeur, *Localization of actin messenger RNA during early ascidian development*, Developmental Biology, 99 (1983), pp. 408–17.

[92] O. Johnstone and P. Lasko, *Translational regulation and RNA localization in Drosophila oocytes and embryos*, Annual Review of Genetics, 35 (2001), pp. 365–406.

[93] P. Kaewsapsak, D. M. Shechner, W. Mallard, J. L. Rinn, and A. Y. Ting, *Live-cell mapping of organelle-associated RNAs via proximity biotinylation combined with protein-RNA crosslinking*, eLife, 6 (2017).

[94] H. Kazan, D. Ray, E. T. Chan, and T. R. Hughes, *RNAcontext : a new method for learning the sequence and structure binding preferences of RNA-binding proteins*, PLoS Computational Biology, 6 (2010).

[95] K. C. Keiler, *RNA localization in bacteria*, Current Opinion Microbiology, 14 (2011), pp. 155–9.

[96] A. E. Kel, E. Gossling, I. Reuter, E. Cheremushkin, O. V. Kel-Margoulis, and E. Wingender, *MATCH : A tool for searching transcription factor binding sites in DNA sequences*, Nucleic Acids Research, 31 (2003), pp. 3576–9.

[97] A. Khan, O. Fornes, A. Stigliani, M. Gheorghe, J. A. Castro-Mondragon, R. van der Lee, A. Bessy, J. Cheneby, S. R. Kulkarni, G. Tan, D. Baranasic, D. J. Arenillas, A. Sandelin, K. Vandepoele, B. Lenhard, B. Ballester, W. W. Wasserman, F. Parcy, and A. Mathelier, *JASPAR 2018 : update of the open-access database of transcription factor binding profiles and its web framework*, Nucleic Acids Research, 46 (2018), p. D1284.

[98] J. Kim, J. Lee, S. Lee, B. Lee, and J. Kim-Ha, *Phylogenetic comparison of oskar mRNA localization signals*, Biochemical and Biophysical Research Communications, 444 (2014), pp. 98–103.

[99] J. Kim-Ha, J. L. Smith, and P. M. Macdonald, *oskar mRNA is localized to the posterior pole of the Drosophila oocyte*, Cell, 66 (1991), pp. 23–35.

[100] J. Kim-Ha, P. J. Webster, J. L. Smith, and P. M. Macdonald, *Multiple RNA regulatory elements mediate distinct steps in localization of oskar mRNA*, Development, 119 (1993), pp. 169–78.

[101] E. H. Kislauskis and R. H. Singer, *Determinants of mRNA localization*, Current Opinion in Cell Biology, 4 (1992), pp. 975–8.

[102] E. H. Kislauskis, X. Zhu, and R. H. Singer, *Sequences responsible for intracellular localization of β-actin messenger RNA also affect cell phenotype*, The Journal of Cell Biology, 127 (1994), pp. 441–451.

[103] M. Kloc, S. Bilinski, A. Pui-Yee Chan, and L. D. Etkin, *The targeting of Xcat2 mRNA to the germinal granules depends on a cis-acting germinal granule localization element within the 3' UTR*, Developmental Biology, 217 (2000), pp. 221–9.

[104] R. B. Knowles, J. H. Sabry, M. E. Martone, T. Deerink, M. Ellisman, G. J. Bassell, and K. Kosik, *Translocation of RNA granules in living neurons*, Journal of Neuroscience, 16 (1996), pp. 7812–20.

[105] L. J. Korn, C. L. Queen, and M. N. Wegman, *Computer analysis of nucleic acid regulatory sequences*, Proceedings of the National Academy of Sciences of the United States of America, 74 (1977), pp. 4401–5.

[106] M. Kowanda, J. Bergalet, M. Wieczorek, G. Brouhard, E. Lécuyer, and P. Lasko, *Loss of function of the Drosophila Ninein-related centrosomal protein Bsg25D causes mitotic defects and impairs embryonic development*, Biology Open, 5 (2016), pp. 1040–51.

[107] T. L. Kress, Y. J. Yoon, and K. L. Mowry, *Nuclear RNP complex assembly initiates cytoplasmic RNA localization*, The Journal of Cell Biology, 165 (2004), pp. 203–11.

[108] C. Kruse, A. Jaedicke, J. Beaudouin, F. Böhl, D. Ferring, T. Güttler, J. Ellenberg, and R.-P. Jansen, *Ribonucleoprotein-dependent localization of the yeast class V myosin Myo4p*, The Journal of Cell Biology, 159 (2002), pp. 971–82.

[109] J. M. Kugler and P. Lasko, *Localization, anchoring and translational control of oskar, gurken, bicoid and nanos mRNA during Drosophila oogenesis*, Fly, 3 (2009), pp. 15–28.

[110] J. Kuriyan, D. Eisenberg, J. Kuriyan, and D. Eisenberg, *The origin of protein interactions and allostery in colocalization*, Nature, 450 (2007), pp. 983–90.

[111] V. Lantz and P. Schedl, *Multiple cis-acting targeting sequences are required for orb mRNA localization during Drosophila oogenesis*, Molecular and Cellular Biology, 14 (1994), pp. 2235–342.

[112] P. Lasko, *mRNA localization and translational control in Drosophila oogenesis*, Cold Spring Harbor Perspectives in Biology, 4 (2012).

[113] M. Lavallée-Adam, P. Cloutier, and B. Coulombe, *Functional 5' UTR motif discovery with LESMoN : Local enrichment of sequence motifs in biological networks*, Nucleic Acids Research, 45 (2017), pp. 10415–27.

[114] C. E. Lawrence, S. F. Altschul, M. S. Boguski, J. S. Liu, A. F. Neuwald, and J. C. Wootton, *Detecting subtle sequence signals : a Gibbs sampling strategy for multiple alignment*, Science, 262 (1993), pp. 208–14.

[115] C. E. Lawrence and A. A. Reilly, *An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences*, Proteins : Structure, Function, and Genetics, 7 (1990), pp. 41–51.

[116] E. Lecuyer, H. Yoshida, N. Parthasarathy, C. Alm, T. Babak, T. Cerovina, T. R. Hughes, P. Tomancak, and H. M. Krause, *Global analysis of mRNA localization reveals a prominent role in organizing cellular architecture and function*, Cell, 131 (2007), pp. 174–87.

[117] F. A. Lefebvre, N. A. L. Cody, L. P. Benoit Bouvrette, J. Bergalet, X. Wang, and E. Lecuyer, *CeFra-seq : Systematic mapping of RNA subcellular distribution properties through cell fractionation coupled to deep-sequencing*, Methods, 126 (2017), pp. 138–48.

[118] D. A. Lerit and E. R. Gavis, *Transport of germ plasm on astral microtubules directs germ cell development in Drosophila*, Current Biology, 21 (2011), pp. 439–48.

[119] A. Liefooghe, H. Touzet, and J. S. Varré, *Large scale matching for position weight matrices*, Annual Symposium on Combinatorial Pattern Matching, (2006).

[120] Y. Liu, S. Sun, T. Bredy, M. Wood, R. C. Spitale, and P. Baldi, *MotifMap-RNA : a genome-wide map of RBP binding sites*, Bioinformatics, 33 (2017), pp. 2029–31.

[121] R. M. Long, W. Gu, E. Lorimer, R. H. Singer, and P. Chartrand, *She2p is a novel RNA-binding protein that recruits the Myo4p-She3p complex to ASH1 mRNA*, The EMBO Journal, 19 (2000), pp. 6592–601.

[122] R. M. Long, R. H. Singer, X. Meng, I. Gonzalez, K. Nasmyth, and R.-P. Jansen, *Mating type switching in yeast controlled by asymmetric localization of ASH1 mRNA*, Science, 277 (1997), pp. 383–7.

[123] R. Lorenz, S. H. Bernhart, C. Honer Zu Siederdissen, H. Tafer, C. Flamm, P. F. Stadler, and I. L. Hofacker, *ViennaRNA package 2.0*, Algorithms for Molecular Biology, 6 (2011), p. 26.

[124] P. M. Macdonald and K. Kerr, *Redundant RNA recognition events in bicoid mRNA localization*, RNA, 3 (1997), pp. 1413–20.

[125] ——, *Mutational analysis of an RNA recognition element that mediates localization of bicoid mRNA*, Molecular and Cellular Biology, 18 (1998), p. 3788–95.

[126] P. M. Macdonald, K. Kerr, J. L. Smith, and A. Leask, *RNA regulatory element BLE1 directs the early steps of bicoid mRNA localization*, Development, 118 (1993), pp. 1233–43.

[127] P. M. Macdonald and G. Struhl, *Cis-acting sequences responsible for anterior localization of bicoid mRNA in drosophila embryos*, Nature, 336 (1988), pp. 595–8.

[128] J. Machta, *Entropy, information, and computation*, American Journal of Physics, 62 (1999), pp. 1074–7.

[129] K. D. MacIsaac and E. Fraenkel, *Practical strategies for discovering regulatory DNA sequence motifs*, PLoS Computational Biology, 2 (2006), p. e36.

[130] K. C. Martin and A. Ephrussi, *mRNA localization : gene expression in the spatial dimension*, Cell, 136 (2009), pp. 719–30.

[131] D. Maticzka, S. J. Lange, F. Costa, and R. Backlofen, *GraphProt : modeling binding preferences of RNA-binding proteins*, Genome Biology, 15 (2014).

[132] V. Matys, E. Fricke, R. Geffers, E. Gossling, M. Haubrock, R. Hehl, K. Hornischer, D. Karas, A. E. Kel, O. V. Kel-Margoulis, D. U. Kloos, S. Land, B. Lewicki-Potapov, H. Michael, R. Munch, I. Reuter, S. Rotert, H. Saxel, M. Scheer, S. Thiele, and E. Wingender, *TRANSFAC : transcriptional regulation, from patterns to profiles*, Nucleic Acids Research, 31 (2003), pp. 374–8.

[133] M. Mayford, D. Baranes, K. Podsypania, and E. Kandel, *The 3-untranslated region of CaMKIIaplha is a cis-acting signal for the localization and translation of mRNA in dendrites*, Proceedings of the National Academy of Science of the United States of America, 93 (1996), pp. 13250–5.

[134] C. Medioni, K. Mowry, and F. Besse, *Principles and roles of mRNA localization in animal development*, Development, 139 (2012), pp. 3263–76.

[135] E. J. Meer, D. Wang, S. Kim, I. Barr, F. Guo, K. C. Martin, E. J. Meer, D. Wang, S. Kim, I. Barr, F. Guo, and K. C. Martin, *Identification of a cis-acting element that localizes mRNA to synapses*, Proceedings of the National Academy of Sciences of the United States of Americas, 109 (2012), pp. 4639–44.

[136] T. J. Messitt, J. A. Gagnon, J. A. Kreiling, C. A. Pratt, Y. J. Yoon, and K. L. Mowry, *Multiple kinesin motors coordinate cytoplasmic RNA transport on a subpopulation of microtubules in Xenopus oocytes*, Developmental Cell, 15 (2008), pp. 426–36.

[137] M. Mikl, G. Vendra, and M. A. Kiebler, *Independent localization of MAP2, CaMKIIaand β-actin RNAs in low copy numbers*, EMBO Report, 12 (2011), pp. 1077–84.

[138] S. Mili, K. Moissoglu, and I. G. Macara, *Genome-wide screen reveals APC-associated RNAs enriched in cell protrusions*, Nature, 453 (2008), pp. 115–21.

[139] E. Mohr, J. F. Morris, and D. Richter, *Differential subcellular mRNA targeting : deletion of a single nucleotide prevents the transport to axons but not to dendrites of rat hypothalamic magnocellular neurons*, Proceedings of the National Academy of Sciences of the United States of America, 92 (1995), pp. 4377–81.

[140] Y. Mori, K. Imaizumi, T. Katayama, T. Yoneda, and M. Tohyama, *Two cis-acting elements in the 3' untranslated region of a-CaMKII regulate its dendritic targeting*, Nature Neuroscience, 3 (2000), pp. 1079–84.

[141] K. L. Mowry and D. A. Melton, *Vegetal messenger RNA localization directed by a 340-nt RNA sequence element in Xenopus oocytes*, Science, 255 (1992), pp. 991–4.

[142] T. P. Munro, R. J. Magee, G. J. Kidd, J. H. Carson, E. Barbarese, L. M. Smith, and R. Smith, *Mutational analysis of a heterogeneous nuclear ribonucleoprotein A2 response element for RNA trafficking*, Journal of Biological Chemistry, 274 (1999).

[143] Y. Oleynikov and R. H. Singer, *Real-time visualization of ZBP1 association with β-actin mRNA during transcription and localization*, Current Biology, 13 (2003), pp. 199–207.

[144] C. Olivier, G. Poirier, P. Gendron, A. Boisgontier, M. Major, and P. Chartrand, *Identification of a conserved RNA motif essential for She2p recognition and mRNA localization to the yeast bud*, Molecular and Cellular Biology, 25 (2005), pp. 4752–66.

[145] Y. Orenstein, Y. Wang, and B. Berger, *RCK : accurate and efficient inference of sequence- and structure-based protein–RNA binding models from RNAcompete data*, Bioinformatics, 32 (2016), pp. i351–9.

[146] F. Pan, S. Hüttelmaier, R. H. Singer, and W. Gu, *ZBP2 facilitates binding of ZBP1 to β-actin mRNA during transcription*, Molecular and Cellular Biology, 27 (2009), p. 8340–8351.

[147] N. Paquin and P. Chartrand, *Local regulation of mRNA translation : new insights from the bud*, Trends in Cell Biology, 18 (2008), pp. 105–11.

[148] V. L. Patel, S. Mitra, R. Harris, A. R. Buxbaum, T. Lionnet, M. Brenowitz, M. Girvin, M. Levy, S. C. Almo, and R. H. Singer, *Spatial arrangement of an RNA zipcode identifies mRNAs under post-transcriptional control*, Genes and Development, 26 (2012), pp. 43–53.

[149] G. Pavesi, P. Mereghetti, G. Mauri, and G. Pesole, *Weeder web : discovery of transcription factor binding sites in a set of sequences from co-regulated genes*, Nucleic Acids Research, 32 (2004), pp. W199–203.

[150] I. Paz, I. Kosti, M. Ares Jr, and M. Cline, *RBPmap : a web server for mapping binding sites of RNA-binding proteins*, Nucleic Acids Research, 42 (2014), pp. W361–7.

[151] A. Paziewska, L. S. Wyrwicz, J. M. Bujnicki, K. Bomsztyk, and J. Ostrowski, *Cooperative binding of the hnRNP K three KH domains to mRNA targets*, FEBS Letters, 577 (2004), pp. 134–40.

[152] N. Prakash, S. Fehr, E. Mohr, and D. Richter, *Dendritic localization of rat vasopressin mRNA : ultrastructural analysis and mapping of targeting elements*, The European Journal of Neuroscience, 9 (1997), pp. 523–32.

[153] F. Prodon, L. Yamada, M. Shirae-Kurabayashi, Y. Nakamura, and Y. Sasakura, *Postplasmic/PEM RNAs : a class of localized maternal mRNAs with multiple roles in cell polarity and development in ascidian embryos*, Developmental Dynamic, 236 (2007), pp. 1698–715.

[154] C. Queen, M. N. Wegman, and L. J. Korn, *Improvements to a program for DNA analysis : a procedure to find homologies among many sequences*, Nucleic Acids Research, 10 (1982), pp. 449–56.

[155] C. Racca, A. Gardiol, T. Eom, J. Ule, A. Triller, and R. B. Darnell, *The neuronal splicing factor Nova co-localizes with target RNAs in the dendrite*, Frontiers in Neural Circuits, 4 (2010).

[156] D. Ray, H. Kazan, K. B. Cook, M. T. Weirauch, H. S. Najafabadi, X. Li, S. Gueroussov, M. Albu, H. Zheng, A. Yang, H. Na, M. Irimia, L. H. Matzat, R. K. Dale, S. A. Smith, C. A. Yarosh, S. M. Kelly, B. Nabet, D. Mecenas, W. Li, R. S. Laishram, M. Qiao, H. D. Lipshitz, F. Piano, A. H. Corbett, R. P. Carstens, B. J. Frey, R. A. Anderson, K. W. Lynch, L. O. Penalva, E. P. Lei, A. G. Fraser, B. J. Blencowe, Q. D. Morris, and T. R. Hughes, *A compendium of RNA-binding motifs for decoding gene regulation*, Nature, 499 (2013), pp. 172–7.

[157] K. Rihan, E. Antoine, T. Maurin, B. Bardoni, R. Bordonné, J. Soret, and F. Rage, *A new cis-acting motif is required for the axonal SMN-dependent Anxa2 mRNA localization*, RNA, 23 (2017), p. 899.

[158] A. F. Ross, Y. Oleynikov, E. H. Kislauskis, K. L. Taneja, and R. Singer, *Characterization of a β-actin mRNA zipcode-binding protein*, Molecular and Cellular Biology, 17 (1997), pp. 2158–65.

[159] Y. H. Ryu, A. Kenny, Y. Gim, M. Snee, and P. M. J. Macdonald, *Multiple cis-acting signals, some weak by necessity, collectively direct robust transport of oskar mRNA to the oocyte*, Journal of Cell Science, 130 (2017).

[160] A. Sandelin, W. Alkema, P. Engstrom, W. W. Wasserman, and B. Lenhard, *JASPAR : an open-access database for eukaryotic transcription factor binding profiles*, Nucleic Acids Reseach, 32 (2004), pp. D91–4.

[161] G. K. Sandve and F. Drablos, *A survey of motif discovery methods in an integrated framework*, Biology Direct, 1 (2006), p. 11.

[162] C. Saunders and R. S. Cohen, *The role of oocyte transcription, the 5' UTR, and translation repression and derepression in Drosophila gurken mRNA and protein localization*, Molecular Cell, 3 (1999), pp. 43–54.

[163] T. D. Schneider, *A brief review of molecular information theory*, Nano Communication Networks, 1 (2010), pp. 173–80.

[164] T. D. Schneider and R. M. Stephens, *Sequence logos : a new way to display consensus sequences*, Nucleic Acids Research, 18 (1990), pp. 6097–100.

[165] J. Serano and G. M. Rubin, *The Drosophila synaptotagmin-like protein bitesize is required for growth and has mRNA localization sequences within its open reading frame*, Proceedings of the National Academy of Sciences of the United States of Americas, 100 (2003), pp. 13368–73.

[166] T. L. Serano and R. S. Cohen, *A small predicted stem-loop structure mediates oocyte localization of Drosophila K10 mRNA*, Development, 121 (1995), pp. 3809–3818.

[167] K. A. Serikawa, D. M. Porterfield, and D. F. Mandoli, *Asymmetric subcellular mRNA distribution correlates with carbonic anhydrase activity in Acetabularia acetabulum*, Plant Physiology, 125 (2001), pp. 900–11.

[168] M. C. Shaner, I. M. Blair, and T. D. Schneider, *Sequence logos : A powerful, yet simple, tool*, IEEE, 1 (1993), pp. 813–21.

[169] E. A. Shestakova, R. H. Singer, and J. Condells, *The physiological significance of β-actin mRNA localization in determining cell polarity and directional motility*, Proceedings of the National Academy of Science of the United States of America, 98 (2001), pp. 7045–50.

[170] S. Sinha and M. Tompa, *YMF : A program for discovery of novel transcription factor binding sites by statistical overrepresentation*, Nucleic Acids Research, 31 (2003), pp. 3586–8.

[171] M. J. Snee, E. A. Arn, S. L. Bullock, and P. M. Macdonald, *Recognition of the bcd mRNA localization signal in Drosophila embryos and ovaries*, Molecular and Cellular Biology, 25 (2005), pp. 1501–1510.

[172] D. St Johnston, D. Beuchle, and C. Nüsslein-Volhard, *Staufen, a gene required to localize maternal RNAs in the Drosophila egg*, Cell, 66 (1991), pp. 51–63.

[173] R. Staden, *Computer methods to locate signals in nucleic acid sequences*, Nucleic Acids Research, 12 (1984), pp. 505–19.

[174] P. A. Takizawa, A. Sil, J. R. Swedlow, I. Herskowitz, and R. D. Vale, *Actin-dependent localization of an RNA encoding a cell-fate determinant in yeast*, Nature, 389 (1997), pp. 90–93.

[175] H. Tekotte and I. Davis, *Intracellular mRNA localization : motors move messages*, Trends Genetics, 18 (2002), pp. 636–42.

[176] V. Van De Bor and I. Davis, *mRNA localisation gets more complex*, Current Opinion in Cell Biology, 16 (2004), pp. 300–307.

[177] V. Van De Bor, E. Hartswood, C. Jones, D. Finnegan, and I. Davis, *gurken and the I factor retrotransposon RNAs share common localization signals and machinery*, Developmental Cell, 9 (2005), pp. 51–62.

[178] E. T. Wang, N. A. Cody, S. Jog, M. Biancolella, T. T. Wang, D. J. Treacy, S. Luo, G. P. Schroth, D. E. Housman, S. Reddy, E. Lecuyer, and C. B. Burge, *Transcriptome-wide regulation of pre-mRNA splicing and mRNA localization by muscleblind proteins*, Cell, 150 (2012), pp. 710–24.

[179] L. D. Ward and H. J. Bussemaker, *Predicting functional transcription factor binding through alignment-free and affinity-based analysis of orthologous promoter sequences*, Bioinformatics, 24 (2008), pp. i165–71.

[180] T. T. Weil, K. M. Forrest, and E. R. Gavis, *Localization of bicoid mRNA in late oocytes is maintained by continual active transport*, Developmental Cell, 11 (2006), pp. 251–262.

[181] K. Welshhans and G. J. Bassell, *Netrin-1-induced local β-actin synthesis and growth cone guidance requires zipcode binding protein 1*, Journal of Neuroscience, 31 (2011), pp. 9800–13.

[182] J. E. Wilhelm and R. D. Vale, *RNA on the move : the mRNA localization pathway*, The Journal of Cell Biology, 123 (1993), pp. 269–74.

[183] R. Wilk, J. Hu, D. Blotsky, and H. M. Krause, *Diverse and pervasive subcellular distributions for both coding and long noncoding RNAs*, Genes and Development, 30 (2016), pp. 594–609.

[184] M. Yamagishi, Y. Ishihama, Y. Shirasaki, H. Kurama, and T. Funatsu, *Single-molecule imaging of β-actin mRNAs in the cytoplasm of a living cell*, Experimental Cell Research, 315 (2009), pp. 1142–7.

[185] M. Yamagishi, Y. Shirasaki, and T. Funatsu, *Size-dependent accumulation of mRNA at the leading edge of chicken embryo fibroblasts*, Biochemical and Biophysical Research Communications, 390 (2009), pp. 750–4.

[186] Z. Yan, E. Lecuyer, and M. Blanchette, *Prediction of mRNA subcellular localization using deep recurrent neural networks*, Proceedings of ISMB 2019, (2019).

[187] T. Yano, S. López de Quinto, Y. Matsui, A. Shevchenko, A. Shevchenko, and A. Ephrussi, *Hrp48, a Drosophila hnRNPA/B homolog, binds and regulates translation of oskar mRNA*, Developmental Cell, 6 (2004), pp. 637–48.

[188] K. Zarnack and M. Feldbrugge, *Microtubule-dependent mRNA transport in fungi*, Eukaryotics Cell, 9 (2010), pp. 982–90.

[189] H. L. Zhang, T. Eom, Y. Oleynikov, S. M. Shenoy, D. A. Liebelt, J. B. Dictenberg, R. H. Singer, and G. J. Bassell, *Neurotrophin-induced transport of a β-actin mRNP complex increases β-actin levels and stimulates growth cone motility*, Neuron, 31 (2001), pp. 261–75.

[190] Y. Zhou and M. L. King, *Localization of Xcat-2 RNA, a putative germ plasm component, to the mitochondrial cloud in Xenopus stage I oocytes*, Development, 122 (1996), pp. 2947–53.

[191] ——, *RNA transport to the vegetal cortex of Xenopus oocytes*, Developmental Biology, 179 (1996), pp. 173–83.

[192] V. L. Zimyanin, K. Belaya, J. Pecreaux, M. J. Gilchrist, A. Clark, I. Davis, and D. St Johnston, *In vivo imaging of oskar mRNA transport reveals the mechanism of posterior localization*, Cell, 134 (2008), pp. 843–53.

Deuxième article.

CeFra-seq reveals broad asymmetric mRNA and noncoding RNA distribution profiles in *Drosophila* and human cells

Louis Philip Benoit Bouvrette

Cet article à été publié dans le revue RNA (2018. 24 : 98-113, https://doi.org/10.1261/rna.063172.117)

Préface et contributions

Ce chapitre est présenté sous la forme d'un article de recherche et a aussi été publié dans le journal *RNA*.

L'article présente une méthode optimisée impliquant le fractionnement cellulaire biochimique et le séquençage d'ARN (CeFra-seq), en utilisant, en parallèle, des protocoles de déplétion en ARN ribosomal ou d'enrichissement poly-(A). Ceci m'a permis de profiler globalement la distribution asymétrique des ARN et leurs caractéristiques dans les modèles cellulaires épithéliaux humains HepG2 et de *Drosophile* Dm-D17. La procédure débute par l'isolement de quatre fractions subcellulaires (c.-à-d., noyau, cytosol, membrane et insoluble) de la même population cellulaire de départ, suivie d'une analyse à haut débit et détaillée de la composition en ARN et protéines. Les analyses transcriptomiques de CeFra-seq ont révélées un haut degré de distribution asymétrique des ARN ($> 80\,\%$ des transcrits détectés) dans les cellules humaines et de mouches, ainsi qu'une conservation évolutive des caractéristiques communes aux ARN localisés. Ces résultats sont cohérents avec une estimation antérieure de la prévalence de la localisation d'ARNm observée par imagerie systématique, par hybridation *in situ* en fluorescence (FISH), d'environ 3000 ARNm dans des embryons de *Drosophile* [41], et élargissent considérablement les travaux antérieurs en offrant une vue non seulement à l'échelle du transcriptome, mais incluant tous les biotypes d'ARN. En effet, cette étude révèle la forte prévalence de la localisation subcellulaire des transcrits non codants, tels que les longs ARN non codants (lncRNA) et les ARN circulaires (circRNA), qui présentent un ciblage important vers les fractions cytoplasmiques. Enfin, l'analyse comparative par spectrométrie de masse de protéines isolées à partir des mêmes échantillons fractionnés m'a permis d'évaluer les corrélations générales de distribution des ARNm et des protéines qu'elles encodent. Ces analyses ont révélé des signatures distinctes de distribution d'ARN/protéine corrélées et anti-corrélées et apportent un nouvel éclairage sur les fonctions potentielles de la localisation d'ARN dans le ciblage des modules protéiques.

J'ai effectué les analyses bio-informatiques, créé l'ensemble des figures et écrit l'intégralité du manuscrit. Neal Cody a fortement contribué au développement de la méthode CeFra-seq et a optimisé et effectué le fractionnement cellulaire. Julie Bergalet, Fabio Alexis Lefebvre, Cédric Diot et Xiaofeng Wang ont effectué diverses analyses biologiques, dont les extractions ARN et protéique sur les extraits subcellulaires, les immunobuvardages de Western, les

RT-qPCR, et la préparation des analyses pour la spectrométrie de masse. Eric Lécuyer et Mathieu Blanchette ont supervisé les travaux et l'écriture du manuscrit et ont suggéré des modifications concernant, principalement et respectivement, les aspects biologique et informatique. Tous les co-auteurs ont approuvé le manuscrit avant sa soumission à l'éditeur de *RNA*, qui l'a fait évaluer par des paires avant de le publier.

CeFra-seq reveals broad asymmetric mRNA and noncoding RNA distribution profiles in *Drosophila* and human cells

par

Louis Philip Benoit Bouvrette[1, 2], Neal A.L. Cody[1], Julie Bergalet[1], Fabio Alexis Lefebvre[1, 2], Cédric Diot[1, 2], Xiaofeng Wang[1], Mathieu Blanchette[3] et Eric Lécuyer[1, 2, 4]

(1) Institut de Recherches Clinique de Montréal (IRCM), Montréal, Québec, Canada

(2) Département de biochimie, Université de Montréal, Montréal, Québec, Canada

(3) School of computer science, McGill University, Montréal, Québec, Canada

(4) Division of experimental medicine, McGill University, Montréal, Québec, Canada

Cet article à été publié dans le revue RNA (2018. 24 : 98-113, https://doi.org/10.1261/rna.063172.117).

Abstract. Cells are highly asymmetrical, a feature that relies on the sorting of molecular constituents, including proteins, lipids and nucleic acids, to distinct subcellular locales. The localization of RNA molecules is an important layer of gene regulation required to modulate localized cellular activities, although its global prevalence remains unclear. We combine biochemical cell fractionation with RNA sequencing (CeFra-seq) analysis to assess the prevalence and conservation of RNA asymmetric distribution on a transcriptome-wide scale in *Drosophila* and human cells. This approach reveals that the majority (∼80%) of cellular RNA species are asymmetrically distributed, whether considering coding and noncoding transcript populations, in patterns that are broadly conserved evolutionarily. Notably, a large number of *Drosophila* and human long noncoding RNAs and circular RNAs display enriched levels within specific cytoplasmic compartments, suggesting that these RNAs fulfill extra-nuclear functions. Moreover, fraction-specific mRNA populations exhibit distinctive sequence characteristics. Comparative analysis of mRNA fractionation profiles with that of their encoded proteins reveals a general lack correlation in subcellular distribution, marked by strong cases of asymmetry. However, coincident distribution profiles are observed for mRNA/protein pairs related to a variety of functional protein modules, suggesting complex regulatory inputs of RNA localization to cellular organization.

Highlights

— CeFra-seq enables mapping of transcriptome localization features of human and *Drosophila* cells.

— The high prevalence and features of localized coding and noncoding RNAs are deeply conserved.

— Profiling of mRNA/protein distribution suggests diverse regulatory functions of RNA localization.

Keywords: RNA Localization, Subcellular fractionation, RNA-sequencing, Messenger RNA, Noncoding RNA, Cellular Organization

1. Introduction

In eukaryotic cells, biochemical reactions are often carried out within distinct subcellular compartments by localized molecular machineries. Indeed, most signal transduction systems rely on the colocalization of ligand-receptor pairs, as well as proteins that fulfill various molecular sensing, scaffolding and enzymatic functions [27]. Similarly, the diverse array of regulatory events that modulate gene expression are mediated by compartment-specific ribonucleoprotein (RNP) complexes involved in RNA synthesis, processing, nuclear export, cytoplasmic localization, translation and degradation [24]. By increasing the local concentrations of molecular constituents, colocalization is thought to enhance the probability of productive molecular interactions [40]. In the case of protein-protein interactions, subcellular localization strongly influences proteome organization and has been proposed to be a driving force in the evolution of functional binding interactions and allostery [40, 46]. Moreover, modeling studies suggest that coincident sites of synthesis may be crucial for ensuring the efficient assembly of protein complexes [2].

The intracellular trafficking of RNA molecules is an important and evolutionarily conserved mechanism for controlling cell polarity [50, 4]. This process has been most extensively studied in the context of messenger RNAs (mRNAs), for which localized translation at precise cytoplasmic destinations is implicated in a broad range of biological processes, including developmental patterning, cell fate determination, synaptic plasticity and cell migration [15]. Likewise, subcellular targeting strongly influences the function of various noncoding RNA species, such as long noncoding RNAs (lncRNAs) and small nucleolar RNAs (snoRNAs), and it has been proposed that such RNAs may act as key components of subcellular addressing systems [3]. Over the years, several transcriptome profiling surveys of purified organelles and subcellular compartments have revealed cofractionation of functionally coherent collections of mRNAs [38, 20, 22, 48, 45, 19, 6, 23, 53, 60, 77, 9, 72, 33, 74, 42]. Similarly, global RNA imaging-based screens in *Drosophila* oocytes and embryos have demonstrated that as much as 70% of coding transcripts are localized in patterns that broadly correlate with the distribution and function of their encoded proteins [41, 32, 73]. However, as *Drosophila* embryos may represent an exceptional case where mRNA localization is particularly prominent, due to their large size and syncytial nature, it remains unclear whether a comparably high prevalence of RNA localization is also manifest in standard cells grown in culture.

In this study, we combine subcellular fractionation with RNA sequencing in human and *Drosophila* cellular models, following poly(A)-enrichment or ribosomal RNA (rRNA)-depletion regimens, to assess the extent of RNA subcellular localization in eukaryotic cells. These results reveal the high prevalence of RNA asymmetric localization, with distinctive subcellular enrichments observed for a diverse array of cellular RNA species exhibiting discriminative sequence features. Comparative transcriptome and proteome profiling of cellular fractions further reveals functional coherence in the molecular components enriched within individual fractions, as well as diverse patterns of RNA-protein distribution suggestive of complex regulatory relationships.

2. Results

2.1. Subcellular fractionation and RNA sequencing (CeFra-Seq) of human and insect cells

To gain global insights into the subcellular localization properties of cellular RNAs in eukaryotic cells, and the degree of conservation of RNA distribution signatures, we applied a biochemical cell fractionation strategy coupled with RNA sequencing (CeFra-Seq) to human and *Drosophila* cellular models (Figure 3.1A) [**72**]. For this, we focused on two cell lines with epithelial-like features, human HepG2 hepatocellular carcinoma cells and *Drosophila* DM-D17-c3 (D17) cells, a cell line derived from imaginal discs [**16, 14**]. As outlined in Figure 3.1A, following harvesting, cells were swelled and lysed in hypotonic solution, then subjected to a low-speed centrifugation (1200×g) to isolate pelleted nuclei and a supernatant representing the general cytoplasmic extract. The pellet was further processed via centrifugation over a sucrose cushion to remove un-lysed cells and large cellular debris from the 'Nuclear' fraction. The general cytoplasmic extract was first subjected to high-speed ultracentrifugation at 100,000×g, after which the supernatant was retrieved as the 'Cytosolic' fraction. The recovered pellet was then incubated in buffer supplemented with Triton-X to solubilize endo-membranous components. Subsequent ultracentrifugation thus resulted in the isolation of a soluble 'Membrane' fraction and a pellet consisting of 'Insoluble' cellular material [**72, 28, 31**]. RNA and protein extracts were prepared from each sample and fractionation efficiency was evaluated via western blotting and RT-qPCR (Figure 3.1B, 3.S1A) analyses of fraction-specific markers, in comparison to total extracts from unfractionated cells. Western analysis revealed the expected distribution profiles of protein markers ; with enrichments observed for histone *H3* in the nuclear fraction, monomeric *a-Tubulin* in the cytosol, prominent membrane-targeting of proteins bearing the KDEL motif typically present on in endoplasmic reticulum proteins, and insoluble signatures for cytoskeletal and mitotic apparatus-associated proteins such as *Shot* and *Ninein* (Figure 3.1B). This was also generally the case at the RNA level, with the distinction that transcripts often exhibited a combination of nuclear and cytoplasmic localization signatures, reflecting the nuclear origin of most cellular RNAs. For instance, predominant nuclear targeting was observed for transcripts such *hsr-omega* and *SNORD17*, while others showed enrichments in the cytosolic

(*RN7SK*, *Rpl23a*), membrane (*MT-CO1*, *mt-NDF6*) and insoluble (*TJP-1*, *dlg-1*) fractions of the cytoplasm (Figure 3.S1A).

To evaluate global subcellular transcriptome distribution features, we next subjected RNA from biological replicate fractionation samples of HepG2 and D17 cells to strand-specific and paired-end RNA sequencing, following either poly(A)-enrichment (PA) or rRNA-depletion (RD) regimens. Sequencing reads were respectively aligned to the human and *Drosophila* reference genomes (GRCH 37.75 and BDGP 5.78). For D17 and HepG2 respectively, an average number of aligned reads of 19.9M and 30.5M was obtained for RD libraries and 20.6M and 22M for poly-A+ libraries (Table 3.S1 and Supplemental Files 1-4). Pearson correlation measurements and principal component analyses (PCA) revealed highly correlated transcriptomic signatures between biological replicate samples and distinctive gene expression profiles for each fraction type (Figure 3.1C and 3.S1B). The cumulative number of expressed transcripts, using a threshold of ≥ 1 average fragments per kilobase per million mapped reads (FPKM), for PA and RD libraries was respectively 8308 and 8505 for D17 cells, and 13,787 and 15,158 for HepG2 cells (Table 3.1). The majority of transcripts were detectable using both PA and RD regimens, although a subset of RNAs was only robustly detectable in either dataset (Table 3.1, blue numbers). Moreover, certain biotypes such as lncRNA, miRNAs (here primary miRNAs, pri-miRNAs), snoRNAs and snRNAs were more strongly represented in RD samples. Comparison of inter-fraction expression signatures revealed that most RNA species are detectable across all interrogated subcellular fractions. However, as will be detailed below, the majority display extensive asymmetry in relative fraction enrichment profiles, while many transcripts (2256/1565 and 762/533 for human and fly in RD/PA datasets respectively) were only reliably detected in one fraction compared to all other (Table 3.1, red numbers).

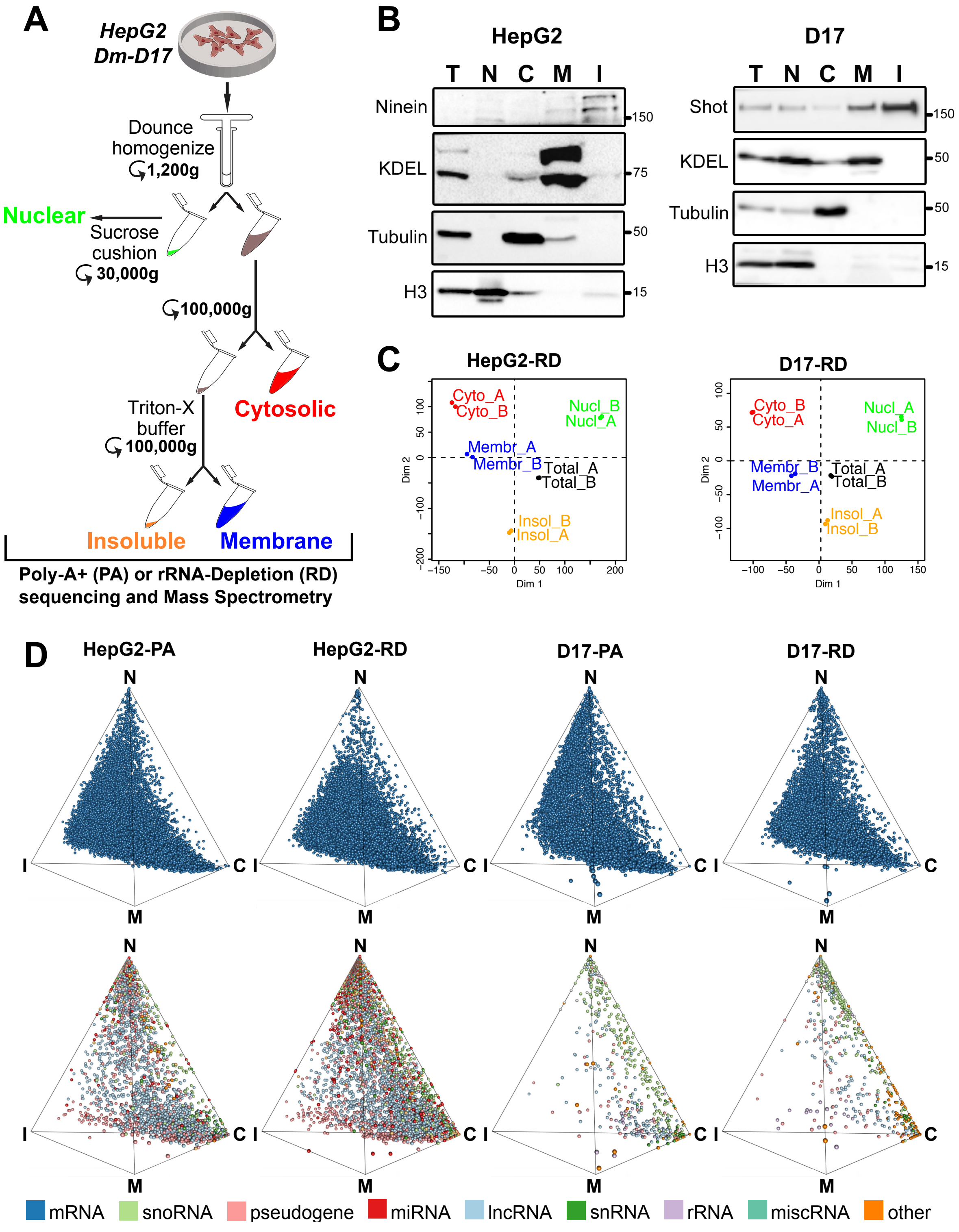


Figure 3.1. Cell fractionation combined with RNA sequencing (CeFra-seq) of human and *Drosophila* epithelial cell models. **(A)** Schematic diagram of the fractionation procedure based on Dounce homogenization, centrifugation and detergent extraction steps to obtain nuclear, cytosolic, membrane and insoluble fractions. **(B)** Western blots of proteins sample controls show fraction efficiency. The accumulation of the indicated protein markers was assessed in human HepG2 and *Drosophila* D17 cells. **(C)** Principal Component Analysis of RNA-seq replicates for HepG2 and D17 cells. **(D)** Simplex graph of the relative localization of mRNAs (*top row*) or noncoding RNAs (*bottom row*) across subcellular fractions, either assessed from poly(A)-enriched (PA) or rRNA-depleted (RD) sequencing datasets. T= Total, C=Cytosolic, M= Membrane, I= Insoluble, N= Nuclear.

TABLE 3.1. Total number of expressed transcripts (FPKM > 1) organized by biotypes.

HepG2

Biotypes	Cytosol		Membrane		Insoluble		Nuclear		Cummulative	
	RD	PA	RD	PA	RD	PA	RD	PA	RD	PA
mRNA	9,742 (1081/138)	9109 (448/146)	9864 (794/37)	9316 (246/13)	9740 (271/48)	9685 (216/64)	10,133 (395/195)	10,125 (387/273)	10,700 (303)	10,784 (387)
lncRNA	1264 (441/252)	1084 (261/215)	1044 (308/25)	916 (180/23)	573 (86/6)	580 (93/12)	1575 (685/557)	1072 (182/335)	2146 (763)	1687 (304)
pseudogene	550 (137/71)	566 (153/83)	484 (87/15)	488 (91/11)	398 (59/31)	386 (47/21)	786 (411/342)	485 (110/123)	1097 (465)	817 (176)
miRNA	135 (99/38)	49 (13/17)	124 (99/8)	31 (6/3)	49 (26/1)	27 (4/1)	384 (285/175)	119 (20/60)	451 (314)	150 (13)
misc RNA	42 (33/8)	11 (2/0)	31 (29/3)	4 (2/0)	8 (8/0)	0 (0/0)	135 (100/25)	41 (6/0)	154 (112)	48 (6)
snoRNA	234 (121/11)	116 (3/24)	199 (135/2)	66 (2/3)	116 (93/1)	25 (2/0)	309 (166/62)	151 (8/43)	332 (152)	186 (6)
snRNA	74 (48/16)	31 (5/22)	60 (51/1)	11 (2/2)	23 (20/0)	3 (0/0)	228 (182/124)	52 (6/32)	250 (179)	80 (9)
rRNA	4 (2/1)	6 (4/3)	3 (1/0)	5 (3/0)	1 (0/0)	4 (3/0)	9 (3/5)	13 (7/7)	10 (4)	16 (10)
Other	5 (33/11)	26 (12/5)	10 (33/0)	19 (11/2)	1 (9/0)	3 (3/0)	1 (103/75)	59 (8/22)	18 (2)	19 (2)
Total	12,050 (1962/546)	10,987 (884/515)	11,891 (1508/88)	10,852 (541/57)	10,909 (537/87)	10,713 (368/98)	12,160 (2230/1535)	12,076 (728/895)	15,158 (2294)	13,787 (913)

D17

Biotypes	Cytosol		Membrane		Insoluble		Nuclear		Cummulative	
	RD	PA	RD	PA	RD	PA	RD	PA	RD	PA
mRNA	7110 (559/120)	6857 (306/132)	6924 (270/3)	6921 (267/5)	6808 (125/7)	6842 (159/11)	7516 (275/416)	7351 (110/245)	7734 (239)	7738 (243)
lncRNA	105 (11/9)	114 (20/9)	89 (11/0)	110 (32/2)	67 (12/0)	74 (19/0)	131 (27/29)	108 (4/15)	142 (15)	142 (15)
pseudogene	43 (7/4)	48 (12/6)	37 (1/0)	42 (6/1)	35 (4/1)	33 (2/0)	57 (12/16)	46 (1/5)	66 (12)	59 (5)
miRNA	24 (17/1)	10 (3/2)	17 (16/0)	4 (3/0)	8 (8/0)	1 (1/0)	60 (25/16)	37 (2/10)	63 (26)	39 (2)
snoRNA	181 (33/1)	161 (13/15)	167 (60/0)	118 (11/2)	119 (103/0)	26 (10/0)	248 (59/27)	194 (5/30)	249 (27)	227 (5)
snRNA	29 (6/4)	23 (0/5)	22 (6/0)	17 (1/0)	18 (16/0)	3 (1/0)	25 (15/0)	10 (0/0)	30 (6)	24 (0)
rRNA	23 (3/0)	21 (1/2)	21 (3/0)	18 (0/0)	19 (4/0)	15 (0/0)	24 (5/1)	20 (1/2)	26 (4)	23 (1)
Other	183 (146/106)	41 (4/24)	31 (21/0)	19 (9/2)	20 (15/0)	6 (1/0)	65 (49/1)	21 (5/8)	195 (146)	56 (7)
Total	7698 (782/245)	7275 (359/195)	7308 (388/3)	7249 (329/12)	7094 (287/8)	7000 (193/11)	8126 (467/506)	7787 (128/315)	8505 (475)	8308 (278)

Numbers in parentheses represent the count of RNA transcripts uniquely identified within the RD or PA library for a given subcellular fraction. (Blue) Number of transcripts uniquely identified within the RD or PA library for a given fraction. (Red) Number of transcripts uniquely expressed in one fraction relative to all other fractions and to the total cell RNA sample.

To help visualize RNA distribution across subcellular fractions, we next built three-simplex graphs in which FPKM values are converted to Cartesian coordinates (see Material and Methods), such that each dot depicts the relative distribution of an individual RNA in relation to the interrogated fractions (Figure 3.1D). For this, we distinguished eight RNA subtypes : (i) mRNAs ; (ii) lncRNAs, including antisense, sense intronic, sense overlapping long intergenic noncoding RNA, and processed transcripts ; (iii) pseudogene-derived transcripts ; (iv) rRNAs ; (v) snoRNA ; (vi) snRNAs, (vi) miscellaneous RNA ; and (vii) pri-miRNAs. This representation conveys the tendency of coding (upper) and noncoding (lower) transcripts to be asymmetrically distributed towards specific fractions, both in HepG2 and D17 cells. We conclude that RNA expression signatures are highly reproducible across CeFra-seq replicate samples and clearly distinctive between subcellular fractions.

2.2. Subcellular fractions exhibit distinctive RNA biotype composition

To further characterize the RNA composition of interrogated fractions, we calculated transcript per million (TPM) values for each fraction, a measure that conveys the relative molar concentrations of transcripts within each sample [**71**]. Adding to our previous eight RNA subtypes, we grouped within the "other" category biotypes for which the highest TPM in any given fraction was below 1%. This analysis revealed clear distinctions in the relative RNA composition of subcellular fractions for both species (Figure 3.2A). For example, in RD samples of HepG2 cells, more than half of the TPM in the cytosolic and membrane fractions were derived from three abundant Pol III transcribed RNAs : *RN7SL*, the RNA component of the signal recognition particle involved in ER targeting of mRNAs encoding transmembrane proteins ; *RN7SK*, a lncRNA implicated in transcription elongation as a component of the pTEFB complex ; and Ribonuclease P RNA Component H1 (*RPPH1*), an endoribonuclease implicated in the maturation of nuclear and mitochondrial tRNAs. These findings are consistent with previous studies reporting strong expression of these RNAs in the cytoplasmic compartments of HEK293 cells and motor neurons [**66, 7**]. The *Drosophila* orthologs of these transcripts (*RNAseP :RNA*, *RN7SK* and *RN7SL*) were also abundant in the cytosol and membrane fractions of D17 cells, albeit to a lesser extent than HepG2. By contrast, a predominant mRNA signature was observed for the insoluble compartment in RD datasets of both cell types (Figure 3.2A), while snoRNAs and pri-miRNAs were primarily nuclear-enriched. As expected, mRNAs were the predominant RNA species represented in PA samples for all fractions tested (Figure 3.S2A).

In our analyses, we found that the total number of FPKMs was different between fractions, following the order cytosol > membrane > insoluble ≅ nuclear. For instance, total FPKMs ranged from 0.24×10^6 (nuclear) to 3.2×10^6 (cytosol) for HepG2 cells, despite having sequenced the fractions to similar depths (Table 3.S1). We reasoned that this variability in total FPKM values might reflect differences in the size distributions of the RNA populations isolated from each fraction. To address this question, we partitioned transcripts expressed in each fraction based on the size of their longest annotated isoform, following a log10 scale spanning 1.5-5 with increments of 0.1 (i.e., ranging from 31 to 100,000 nt in length), and calculated TPM values within each bin, both for RD and PA samples (Figure 3.2B, 3.S2B). Calculating the expected lengths of mRNAs for each fraction revealed an enrichment for shorter transcripts in the cytosolic fraction (1470 nt in PA / 1401 nt in RD samples), transcripts of intermediate length in the membrane fraction (1893 nt in PA / 1818 nt in RD) and a prevalence of longer mRNAs in the insoluble (2918 nt in PA / 2332 nt in RD) and nuclear fractions (2737 nt in PA / 2815 nt in RD). Similar expected length profiles were observed for *Drosophila* D17 samples (Figure 3.2B, 3.S2B), and these fraction-specific differences were apparent whether considering mRNAs or by combining all RNA biotypes together (designated as 'total'). Finally, while the nuclear fraction is enriched in short noncoding transcripts both in HepG2 and D17 cells, total FPKM counts were lower in this fraction due to the high abundance of intronic and intergenic reads, which ranged from 22-70% (Figure 3.S3).

To define the distribution profiles of individual transcripts across subcellular fractions, we next calculated a percent FPKM (pFPKM) value for each fraction, which depicts the proportion of FPKMs obtained within one fraction divided by the sum of FPKMs in all fractions combined ($pFPKM_i = FPKM_i / \sum_{k \in \{C,M,I,N\}} FPKM_k$ for $i \in \{C,M,I,N\}$). We found that pFPKM values closely reflect transcript expression signatures assessed by RT-qPCR validation ($r = 0.94$; Figure 3.S4), offering a reliable metric to assess the relative distribution profiles of individual transcripts across fractions. As depicted in Figure 3.2C and3.S2C, the median pFPKM values of all transcripts belonging to specific biotypes revealed distinctive distribution signatures across fractions. For instance, pFPKM values for mRNAs were similar

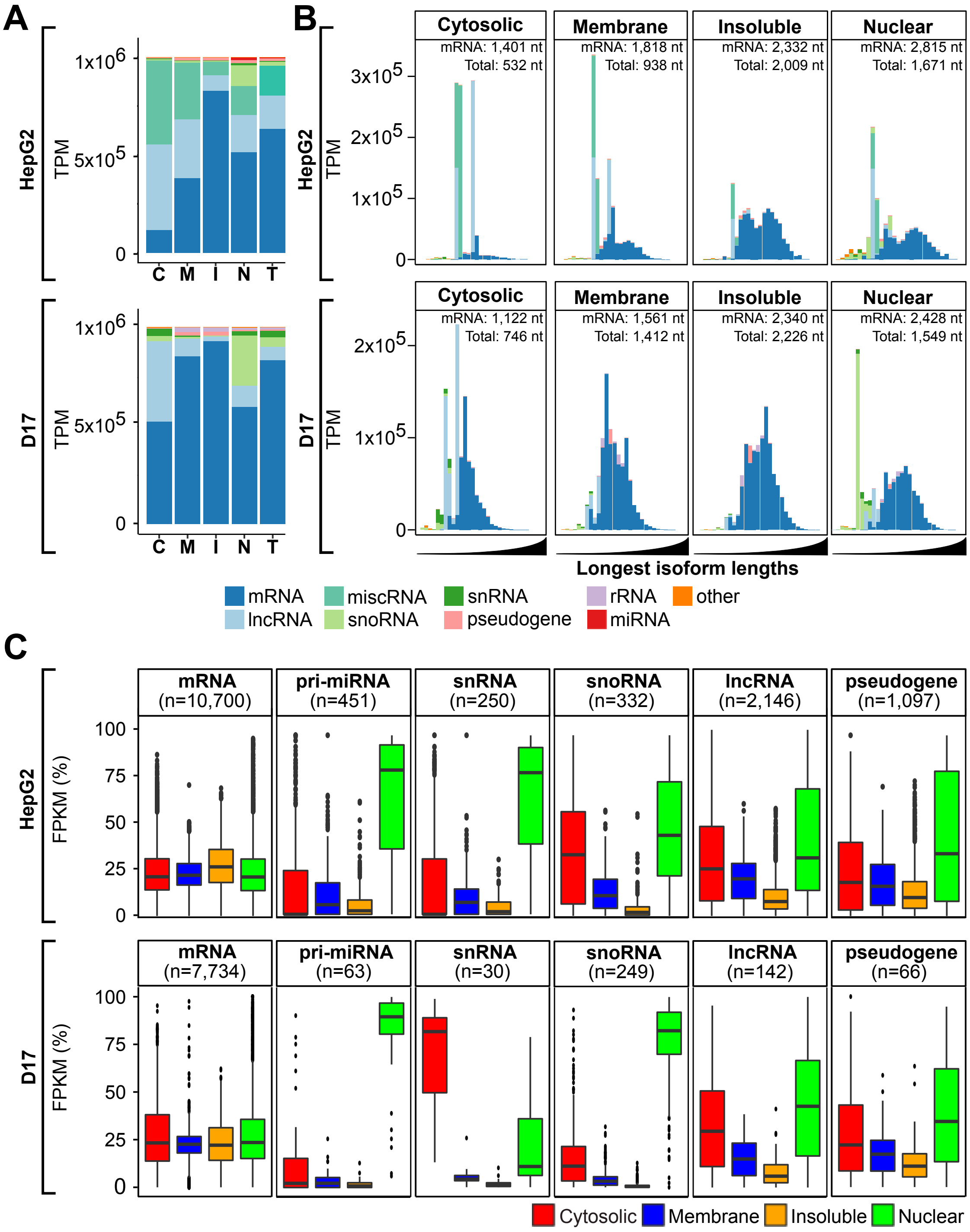


FIGURE 3.2. Distinctive transcript composition of subcellular fractions in rRNA-depletion dataset. **(A)** Histograms depicting the RNA biotype content, in TPM, detected via RD sequencing of cytosolic (C), membrane (M), insoluble (I) and nuclear (N) fractions or whole-cell RNA (T=total) from HepG2 (*upper panel*) and D17 cells (*lower panel*). **(B)** Histograms of the RNA biotype content of HepG2 (*upper panel*) and D17 (*lower panel*) cell fractions, binned according to the length of the longest annotated isoform of detected RNA species, following a log10 scale from 1.5-5 (i.e., ranging from 31-100,000 nt). The expected lengths for mRNA and total RNA populations are indicated for each fraction. For (A) and (B), biotypes accounting for less than 1% of the overall TPMs were grouped as "other". **(C)** Boxplots showing the fraction distribution profiles of different RNA biotypes in percent FPKM (pFPKM) for HepG2 (*upper panel*) and D17 (*lower panel*) cells. The number (n) of transcripts analyzed for each biotype is indicated.

across all fractions, each showing a median pFPKM of ∼25% in both RD and PA datasets. In contrast, transcript biotypes such as pri-miRNAs, snRNA, and snoRNAs showed higher prevalence in the nuclear compartment, while lncRNAs and pseudogene-derived transcripts were generally depleted from the insoluble fraction (Figure 3.2C). Thus, our data reveal that biochemically defined subcellular fractions exhibit distinct RNA composition, both in terms of RNA biotype profiles and expected transcript lengths, features that are strikingly similar between *Drosophila* and human cells.

2.3. Subcellular localization of mRNAs is prevalent and conserved

Previous FISH-based studies in *Drosophila* embryos revealed the high prevalence of RNA subcellular localization [**41**], a feature that has remained poorly defined in cultured cells. To address this question, we next took advantage of our CeFra-seq datasets, for which each fraction was successively isolated from the same starting cellular population, to quantify the global prevalence of RNA asymmetric distribution in HepG2 and D17 cells. For this, we categorized a given RNA as asymmetric when it exhibited a $\geq$ 2-fold difference in expression in at least one fraction when assessed by pair-wise comparisons of fraction FPKM values, either considering all fractions (including the nucleus) or only the cytoplasmic compartments. We further defined RNAs as fraction-specific when they showed $\geq$ 2-fold expression enrichment in one fraction compared to all others. Based on this metric, we found that $\sim$ 90% of mRNAs are asymmetrically distributed across all fractions, $\sim$ 60% when considering only cytoplasmic fractions, while $\sim$ 25% are fraction-specific (Figure 3.3A,B). A similar prevalence of asymmetry was observed in both cell lines and with both PA and RD datasets (Figure 3.3A,B, 3.S5A,B). Expectedly, transcripts classified as fraction-specific using this metric showed a clear regionalization towards the vertices of three-simplex graphs depicting HepG2 and D17 subcellular transcriptomes, in particular towards the nuclear, cytosolic and insoluble fractions (Figure 3.3C,D and3.S5C,D). Notably, few mRNAs were specific to the membrane fraction, as most abundant membrane transcripts were also abundant in the insoluble or cytosolic compartments.

To assess whether mRNA asymmetry signatures are conserved between species, we next compared the fractionation profiles of 2541 mRNAs with high confidence one-to-one orthologs between *Drosophila* and human from the Ensembl database. Strikingly, unsupervised

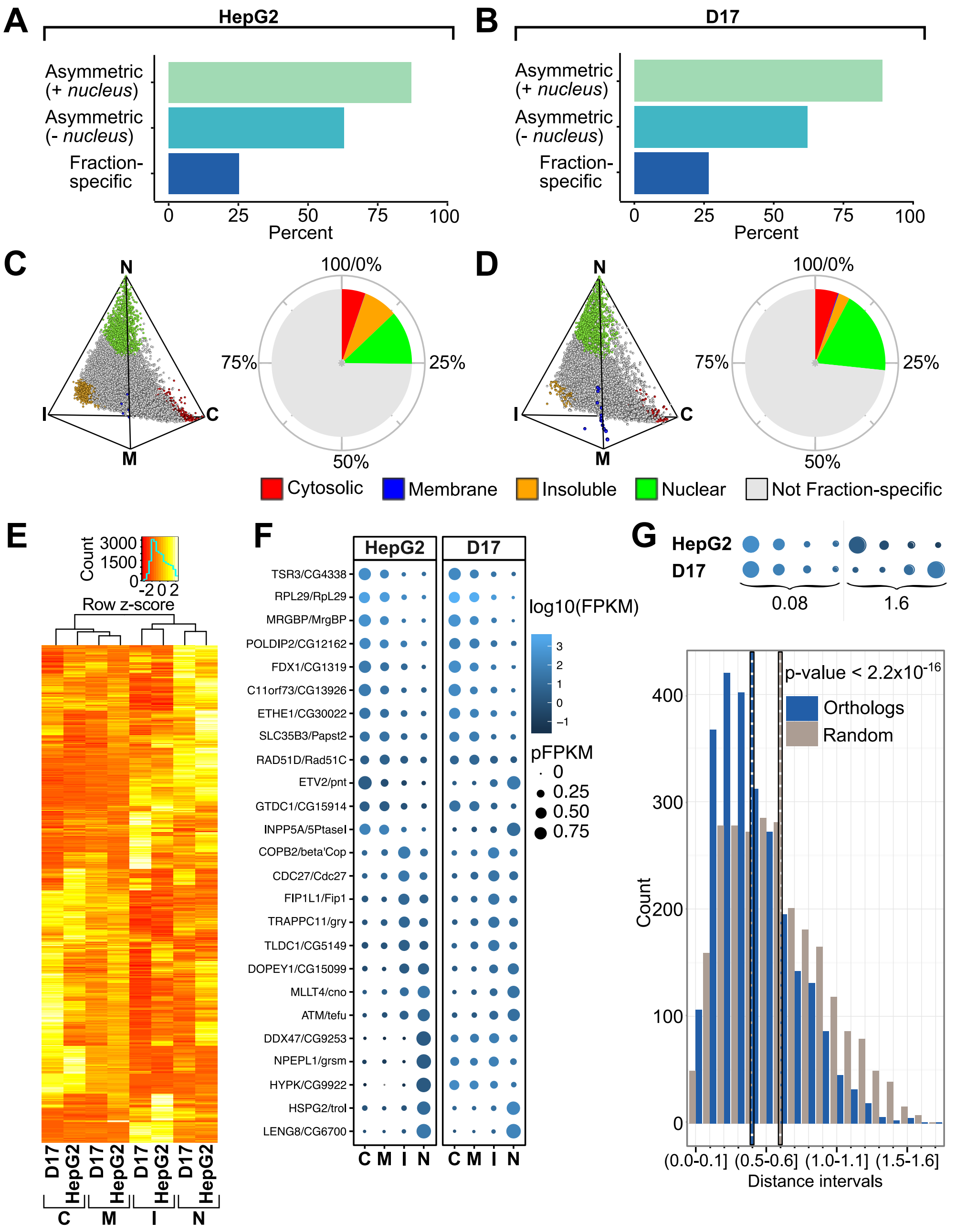


FIGURE 3.3. The subcellular distribution of mRNAs from PA datasets is highly asymmetric and evolutionarily conserved. (A-B) Histograms showing the percent of asymmetrically distributed and fraction-specific mRNAs in HepG2 (A) and D17 (B) cells. **(C-D)** Simplex graphs (*left panels*) and pie charts (*right panels*) depicting the relative distribution and proportion of fraction-specific mRNAs, coloured according to the fraction they are enriched in, relative to the total mRNA population in HepG2 (C) or D17 (D) cells. **(E)** Heatmap of the hierarchical clustering of percent FPKM of *Drosophila* and human orthologs. The hot metal color scale reflects the median-centered distributions of pFPKM. **(F)** Bubble plot showing examples of relative fraction enrichment profiles of orthologous fly and human mRNAs. Gradient blue coloration and bubble size represent log10(FPKM) and pFPKM values respectively. **(G)** A distance measurement score was devised by summing the differences in fraction-specific expression signatures for orthologous mRNAs (*upper schematic*). This metric was used to quantify the relative distance between all human and *Drosophila* orthologs (blue) and the values were binned according to distance range. Identical measurements were performed on the same population of mRNAs that were paired through random shuffling (grey). Vertical dotted lines indicate the median values of the distance distributions for othologous and shuffled pairs. C= Cytosolic, M= Membrane, I= Insoluble, N= Nuclear.

hierarchical clustering analysis revealed general similarities in the mRNA distribution profiles of corresponding HepG2 and D17 subcellular fractions, resulting in their co-clustering (Figure 3.3E). Analysis of pFPKM signatures revealed strong similarities for corresponding D17 and HepG2 subcellular fractions, with Pearson coefficients ranging from 0.33 (membrane) to 0.52 (cytosolic) ($p < 2.2 \times 10^{-16}$) (Figure 3.S6), and similar fractionation profiles of specific orthologs (Figure 3.3F). We further devised a distance measurement score (Figure 3.3G, examples in upper chart), defined as the sum of the differences in pFPKM values across fractions for each ortholog pair, with values ranging from 0 (perfect co-localisation) to 2 (perfectly asymmetric). This score conveys the similarity in relative localization of mRNA orthologs based on their respective pFPKM distribution profiles. By binning orthologous pairs by distance intervals (Figure 3.3G, lower histogram), we found that orthologous transcripts show a significantly shorter average distance (dashed blue line) across fractions, compared to randomly shuffled pairs (dashed grey line) generated from the same starting transcript list (Wilcoxon-Mann-Whitney, p-value $< 2.2 \times 10^{-16}$). Collectively, these results indicate that the asymmetric localization of human and *Drosophila* mRNAs is highly prevalent and broadly conserved evolutionarily.

2.4. Subcellular localization properties of lncRNAs and circular RNAs

We next evaluated the global subcellular distribution properties of lncRNAs. As with coding transcripts, lncRNAs displayed a high prevalence of distribution, with ∼90% detected as asymmetric across all fractions, ∼75% among cytoplasmic compartments, while ∼30% were fraction-specific (Figure 3.4A,B). Similar results were obtained whether considering RD or PA datasets (Figure 3.4A,B, 3.S7A,B) and the fraction-specific lncRNAs exhibited striking regionalization when visualized in three-simplex format (Figure 3.4C, 3.S7C). In both human and fly cells, the most highly asymmetric lncRNAs were primarily cytosolic or nuclear, whereas very few were specifically enriched within the insoluble or membrane fractions. While our standard asymmetry assessment considered all lncRNAs with a minimal expression threshold ≥ 1 FPKM, focusing our analysis on more highly expressed genes (≥ 10 FPKM) revealed a preferential enrichment within the cytosolic fraction (Figure 3.4D, 3.S7D). Notwithstanding their prevalence within the cytosolic and nuclear fractions, we also identified a variety of lncRNAs displaying predominant targeting to each subcellular compartment interrogated, both in HepG2 and D17 cells (Figure 3.4E). For example, the highly conserved RNA component of RNAse P (*RPPH1* and *RNAseP :RNA/CR32868*) is cytosolically-enriched in both human and fly cells. The imprinted maternally expressed *H19* lncRNA, which is aberrantly regulated in Beckwith Wiedermann Syndrome, shows a distinctive localization to the membrane fraction of HepG2 cells. Other fraction enriched transcripts include nuclear lncRNAs such as *MALAT1* and *CR42862*, while transcripts such as *RP11-342K6.1* and *CR31845* are enriched in the insoluble fraction.

With the recent emergence of circular RNAs (circRNA) as an intriguing class of cellular transcript generated through back splicing circularization [**11**], we next sought to interrogate our CeFra-seq data to glean insights into the subcellular compartmentalization features of these RNAs. For this, we first investigated the expression profiles of genes known to encode intron-derived circRNAs, such as *ANKRD52*. Indeed, the *ANKRD52* locus (Figure 3.5A), which transcribes an mRNA coding for a PP6 Phosphotase subunit, as well as a circRNA derived from its second intron [**76**], reveals a primarily insoluble mRNA signature and strong intron 2 read peak in the cytosol. This intronic peak was present in RD samples, but absent in PA samples, and presumably represents a circRNA signature. To gain a broader view of putative circRNA fractionation properties, we next assessed FPKM values for 103 intronic regions known to produce circRNAs, previously characterized by Zhang et al. [**76**]. As shown in circos plot (Figure 3.5B) and box plot (Figure 3.5C) representations, FPKMs aligning to circRNA-producing introns were enriched in the cytosolic and membrane fractions, with a striking depletion from the insoluble compartment. In contrast, analysis non-circRNA producing introns revealed weak expression signatures that were primarily restricted to the nucleus, likely attributable to the unspliced pri-mRNA, whether focusing on a random selection of 103 introns (Figure 3.5D) or all Ensembl-annotated introns (Figure 3.S8A,B). These results are consistent with recent findings that circRNAs are present in the cytoplasm and can be translated [**35**, **34**, **75**].

To evaluate circRNA distribution properties using an orthogonal approach, we adapted the CIRCexplorer protocol [**75**] to search for signature back-spliced junction reads corresponding to circRNAs. This enabled us to identify 1159 and 173 putative circRNAs in our human and fly RD datasets, respectively ; which were primarily detectable in the cytosolic or membrane fractions in both species (Figure 3.5E-H). As expected, no circRNA junction

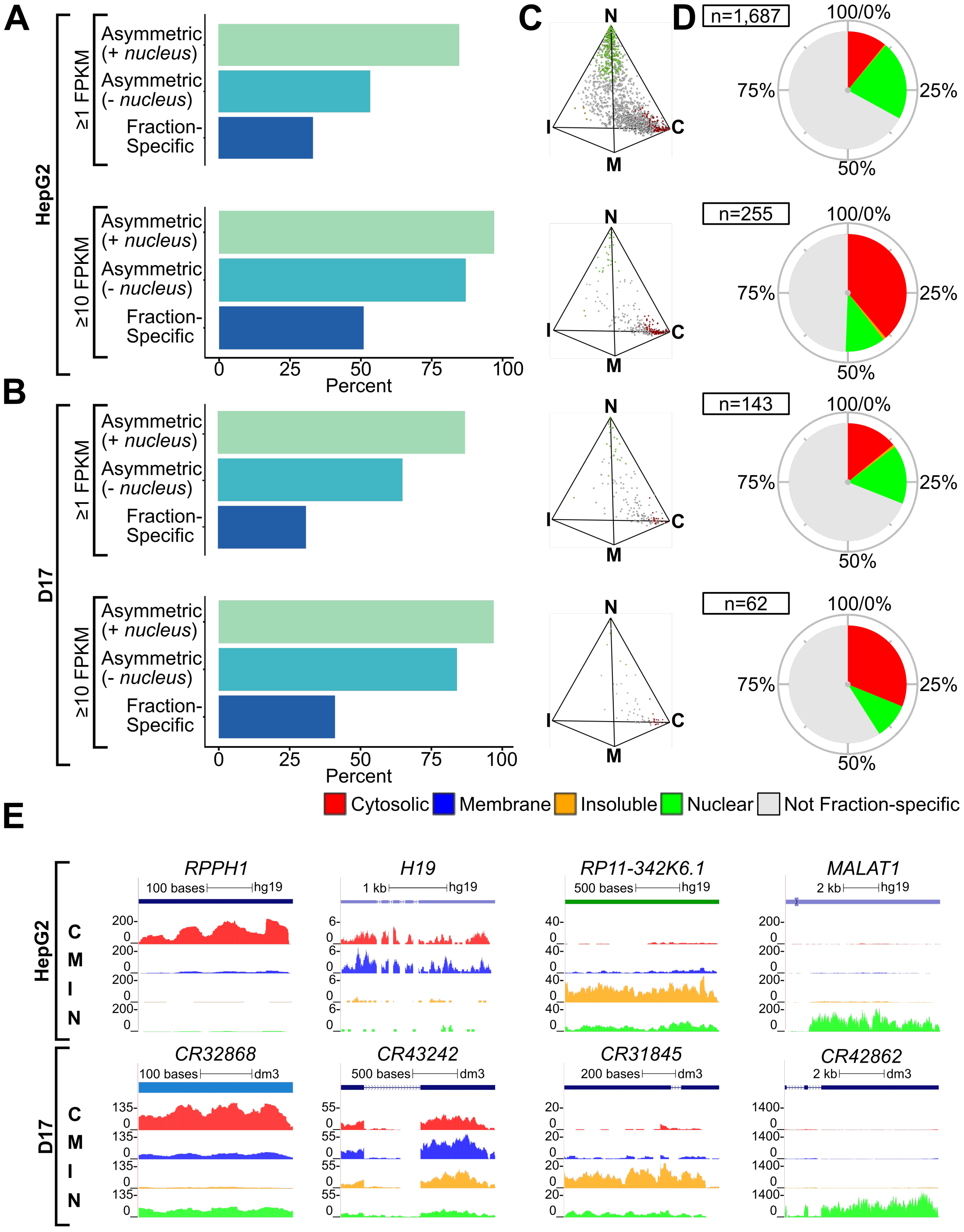


FIGURE 3.4. LncRNAs from PA datasets are asymmetrically distributed and exhibit preferential polarization towards the nucleus and cytosol. (A-B) Histogram showing the percent of asymmetrically distributed and fraction-specific lncRNAs expressed in HepG2 (A) and D17 (B) cells, either using a standard expression threshold (≥1FPKM) or focusing on highly expressed transcripts (≥10 FPKM). **(C-D)** Simplex graphs (C) and pie charts (D) depicting the relative distribution and proportion of fraction-specific lncRNAs, coloured according to the fraction they are enriched in, relative to the total lncRNA populations detected in HepG2 or D17 cells at the expression thresholds indicated in (A-B). **(E)** Genome browser views of candidate lncRNAs exhibiting fraction enrichment, either in HepG2 or D17 cells. C= Cytosolic, M= Membrane, I= Insoluble, N= Nuclear.

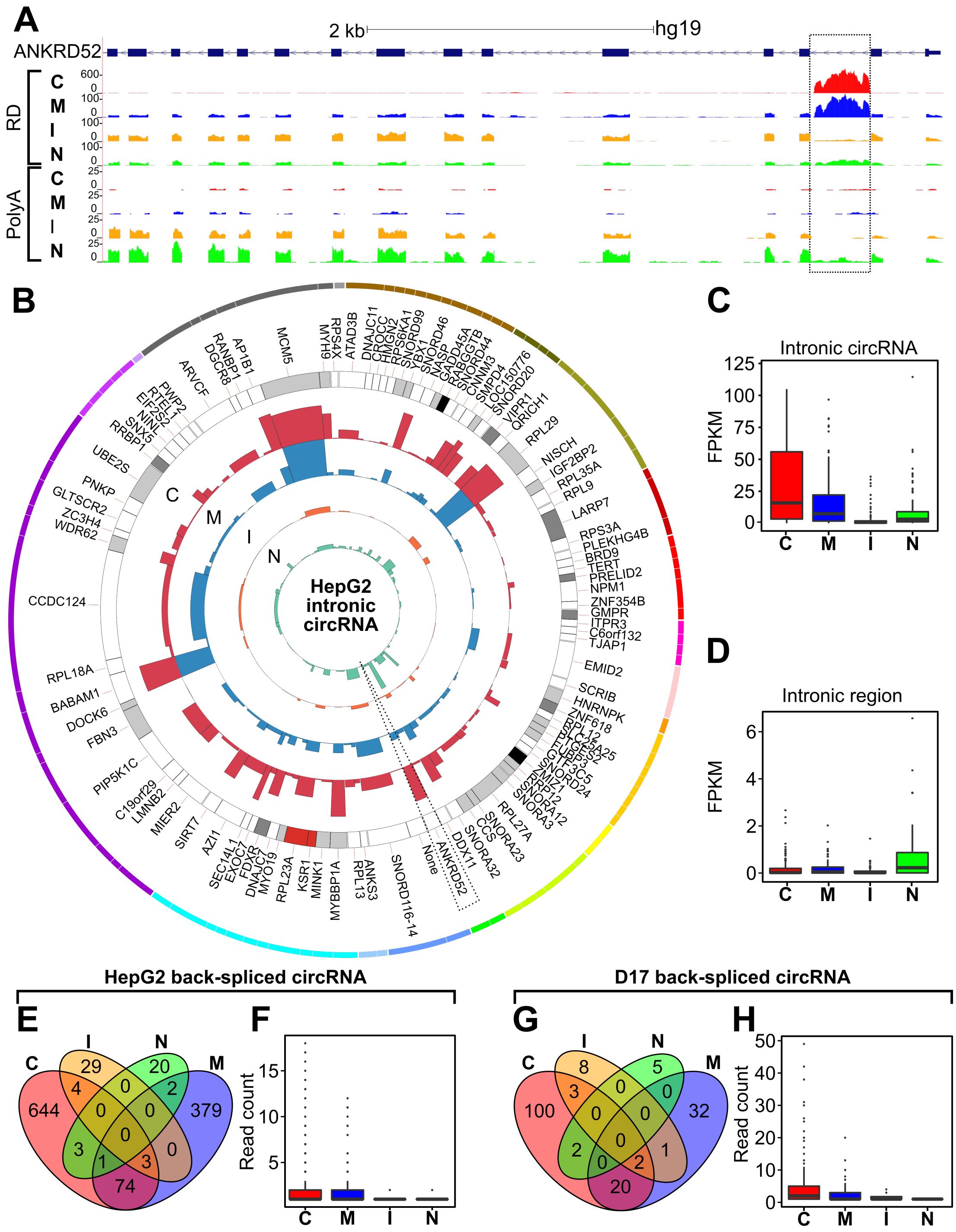


FIGURE 3.5. CircRNA exhibit distinct distribution compared to their host mRNAs and display cytosolic and membrane enrichments. (A) Genome browser view showing RNA-seq read coverage within the *ANKRD52* locus across HepG2 subcellular fractions detected via either RD or PA sequencing. **(B)** Circos plot showing the expression, in FPKM, of 103 intronic regions known to encode circRNAs in human cells. The width of the bar is relative to the length of the intron and the height to its expression within the indicated fractions of HepG2 cells. **(C-D)** Relative expression of the 103 circRNA producing introns (C) detailed in (B) and 103 randomly selected introns (D). **(E-H)** Identification of putative back-spliced circRNAs using the CIRCexplorer algorithm with HepG2 and D17 datasets. (E, G) Venn diagrams of the number of individual circular RNAs detected within the indicated fractions in HepG2 (E) and D17 (G) cells. (F, H) Boxplots showing the FPKM values of these putative circRNAs within the indicated fractions in HepG2 (F) and D17 (H) cells. C= Cytosolic, M= Membrane, I= Insoluble, N= Nuclear.

reads were identified in our PA datasets. We also sought to assess whether there is any relation between the distribution features of circRNAs and that of the cognate mRNAs from which they derive. For this, similarly to our analysis of orthologous mRNAs, we computed a distance measurement between the relative localization of an mRNA and its derived circRNA across fractions for each mRNA/circRNA pair (Figure 3.S8C). These analyses revealed that the distances measured for true mRNA/circRNA pairs was indistinguishable from randomly shuffled pairs, indicating that there is no broad concordance between the localization patterns of mRNA and circRNA transcripts originating from the same locus. Altogether, we conclude that noncoding transcripts, such as lncRNAs and circRNAs, exhibit a high prevalence of subcellular localization in eukaryotic cells.

2.5. Comparison of protein and mRNA distribution signatures across subcellular fractions

To characterize the proteomic signatures generated with the fractionation procedure, we next performed liquid chromatography tandem mass spectrometry (LC-MS/MS) analysis of proteins isolated from the same subcellular fractions as our RNA samples (Supplemental Files S5-S6). Using a stringent filtering procedure (see Material and Methods), we identified 1890 proteins expressed across HepG2 cell fractions. The proteomic signature of each fraction was distinctive (Figure 3.S9) and demarcated functionally coherent protein repertoires associated with specific Gene Ontology (GO) annotations (Tables 3.S2), which matched several of the GO term enrichments observed for fraction-specific mRNAs (Table 3.S3). To evaluate the fractionation similarities of mRNAs and their encoded proteins, we first calculated Spearman correlations of expression signatures, comparing FPKM and spectrum count values as well as asymmetry, comparing pFPKM and percent spectrum counts for mRNA/protein pairs in each fraction. These simple comparisons revealed modest but significant fraction-specific positive correlations in expression, with values ranging from 0.24-0.34 for HepG2 fractions and 0.3-0.47 for D17 fractions (p-values $< 2.2 \times 10^{-16}$) (Figure 3.6A, left panel). The fraction-specific correlation of asymmetry showed weaker values, ranging from -0.03 to 0.16 in human cells and -0.16 to 0.16 in fly cells (p-values $< 2.2 \times 10^{-16}$) (Figure 3.6A, right panel). These generally modest correlation scores were expected, as they are likely to be influenced by cases of mRNA-protein pairs exhibiting strong asymmetric distribution across fractions. Indeed, three-simplex graphs displaying the relative distribution of mRNA/protein pairs for specific protein complexes revealed several striking features (Figure 3.6B). Firstly, when compared amongst each other, functionally-related mRNAs tended to cluster to specific regions of the simplex, a feature that was even more pronounced at the protein level, underlining a general coherence in the distribution properties of mRNA and protein subgroups. In contrast, comparison of mRNA versus protein subsets revealed varying degrees of proximity (Figure 3.6B,C). For example, mRNAs for components of the actin-related protein 2/3 (*Arp2/3*) complex, which are known to undergo localized translation in the cellular cortex [54], were co-clustered with their encoded protein products (Figure 3.6B). Transcripts encoding ribosomal proteins localized to the cytosol, while their protein products show enrichment towards the insoluble/membrane fractions. Components of the PA700 regulatory complex of the 26S proteasome were also asymmetrically partitioned, with mRNAs displaying insoluble/membrane partitioning, while the protein components localize towards the cytosolic vertex (Figure 3.6B). These examples underline the variability in subcellular localization properties of specific classes of mRNA/protein pairs.

To further assess the relationship between the distribution features of mRNAs and that of their encoded proteins, we calculated a percent spectrum count, similar to the pFPKM metric used to define RNA signatures. We further used our distance measurement scores, summing the absolute values of the differences in pFPKM and percent spectrum counts across fractions, for all mRNA/protein pairs detected in our samples. To deconvolve the data in a functionally relevant manner, we evaluated the distance measurements displayed by mRNAs encoding subunits of experimentally-defined protein complexes tabulated within the CORUM database [62], focusing on complexes containing at least 3 subunits and for which we had localization data for at least 75% of the subunits (Figure 3.6C). Overall, more than a third of the clusters meeting our strict thresholds show relative proximity with a distance less than 1 and more than 10% could be defined as colocalizing with a distance less than 0.5. These results suggest that mRNA localization may serve to modulate the subcellular partitioning of several protein machineries through localized translation.

A

Spearman correlations expression	HepG2	D17
Cytosolic	0.24	0.30
Membrane	0.34	0.47
Insoluble	0.30	0.45
Nuclear	0.30	0.37

Spearman correlations asymmetry	HepG2	D17
Cytosolic	0.16	0.16
Membrane	0.05	0.13
Insoluble	-0.03	-0.16
Nuclear	0.04	0.12

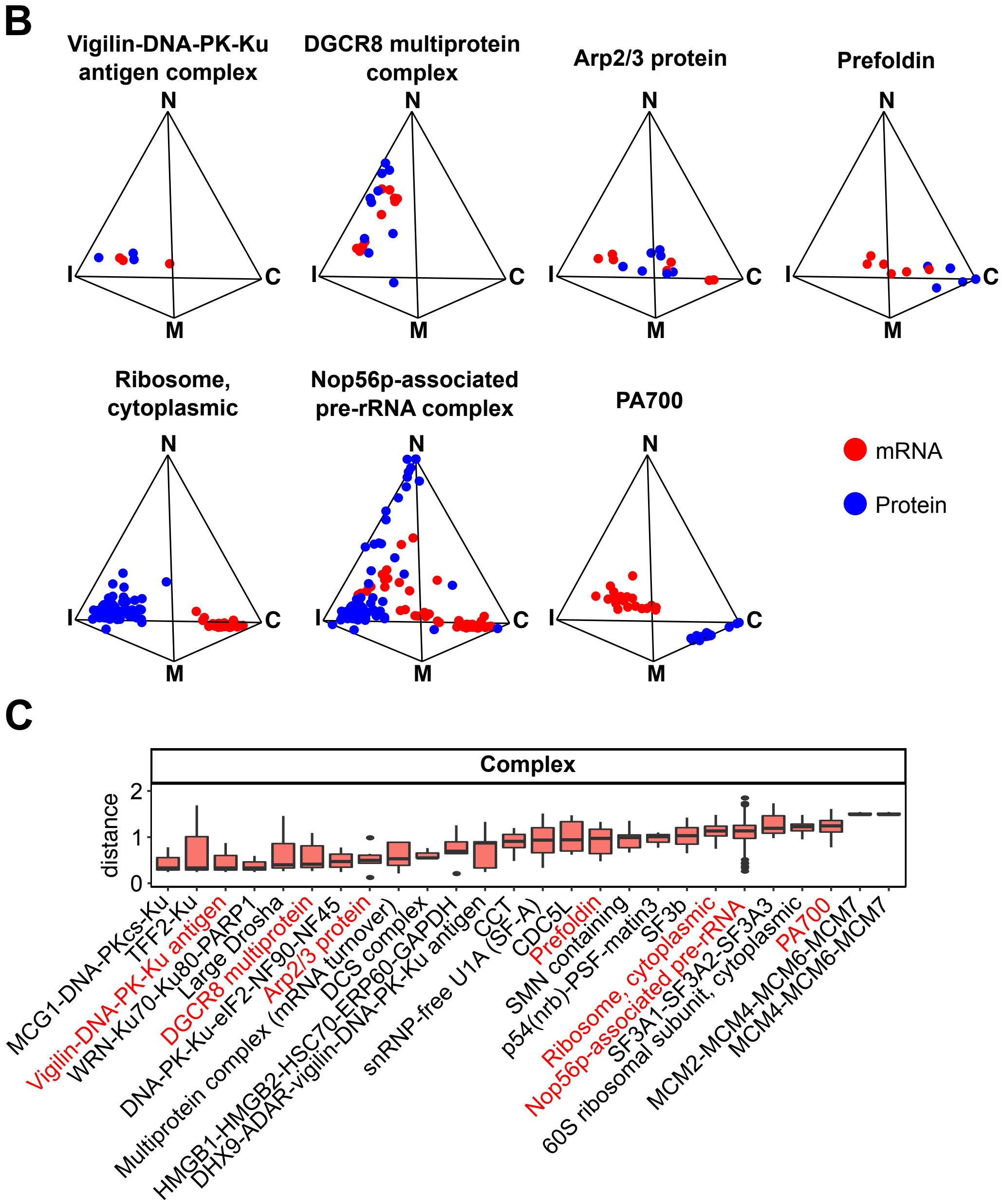


FIGURE 3.6. Proteins are asymmetrically detected in each fraction in patterns that demonstrate specific co-localization with mRNA for a specific subset of genes. (A) Summary table of the Spearman correlation values of mRNA/protein expression (*left*) and mRNA/protein asymmetry (*right*) within the indicated HepG2 and D17 fractions. **(B)** Simplex graphs depicting the relative distribution profiles of components of different protein complexes (blue) and their corresponding mRNAs (red). **(C)** Boxplot of the distance measurement scores of protein components of specific biochemically-defined protein complexes and their corresponding mRNAs. C= Cytosolic, M= Membrane, I= Insoluble, N= Nuclear.

2.6. Asymmetrically distributed mRNAs exhibit conserved features

Having identified subsets of fraction-specific mRNAs, we next sought to assess whether these transcripts exhibit distinctive features. For this, we investigated specific sequence attributes of fractionated mRNA populations, such as the average length of 5' untranslated regions (5' UTR), coding sequences (CDS) and 3' UTRs, or the exon (Figure 3.7A). While 5' UTR lengths were similar across cytoplasmic compartments in both HepG2 and D17 cells, significant differences were apparent when exploring other sequence features. For instance, cytosolic mRNAs exhibit significantly shorter CDS and 3' UTRs compared to other fractions, whereas the 3' UTRs of membrane (in both human and fly cells) and insoluble (in human cells) transcripts were longer on average, suggesting that these transcripts may be more susceptible to post-transcriptional regulatory events mediated by 3' UTR *trans*-regulatory factors. Finally, the CDS of mRNAs enriched in the insoluble and nuclear fractions are longer and contain a significantly larger number of exons compared to other fractions. This observation likely reflects the propensity for longer and more intricately spliced mRNAs to require a prolonged nuclear residence time for their synthesis and maturation. We next evaluated the fractionation properties of mRNA populations known to undergo specific modes of maturation control. For instance, when considering mRNAs encoding secretory proteins bearing signal peptides or transmembrane domains, we observed enrichment for these transcripts in the cytosolic and membrane fractions (Figure 3.S10A), with marked depletion in insoluble, consistent with the known transiting of these mRNAs from the cytosol to the ER via the signal recognition particle. In contrast, canonical histone mRNAs, which are nonpolyadenylated and undergo specialized 3' end processing steps involving a highly conserved stem-loop element, are enriched in the cytosolic fraction in both HepG2 and D17 cells (Figure 3.S10B). Thus, such transcript features help demarcate subcellular localized mRNAs pool and may influence localization control.

Our findings that fraction mRNAs exhibit distinctive features that are evolutionarily conserved in human and *Drosophila* cellular models prompted us to investigate whether RNA regulatory factors also show conserved localization properties. RNA-binding proteins (RBPs) are deeply conserved and essential modulators of RNA metabolism including RNA localization, which can be separated into distinct classes based on the nature of their RNA-binding

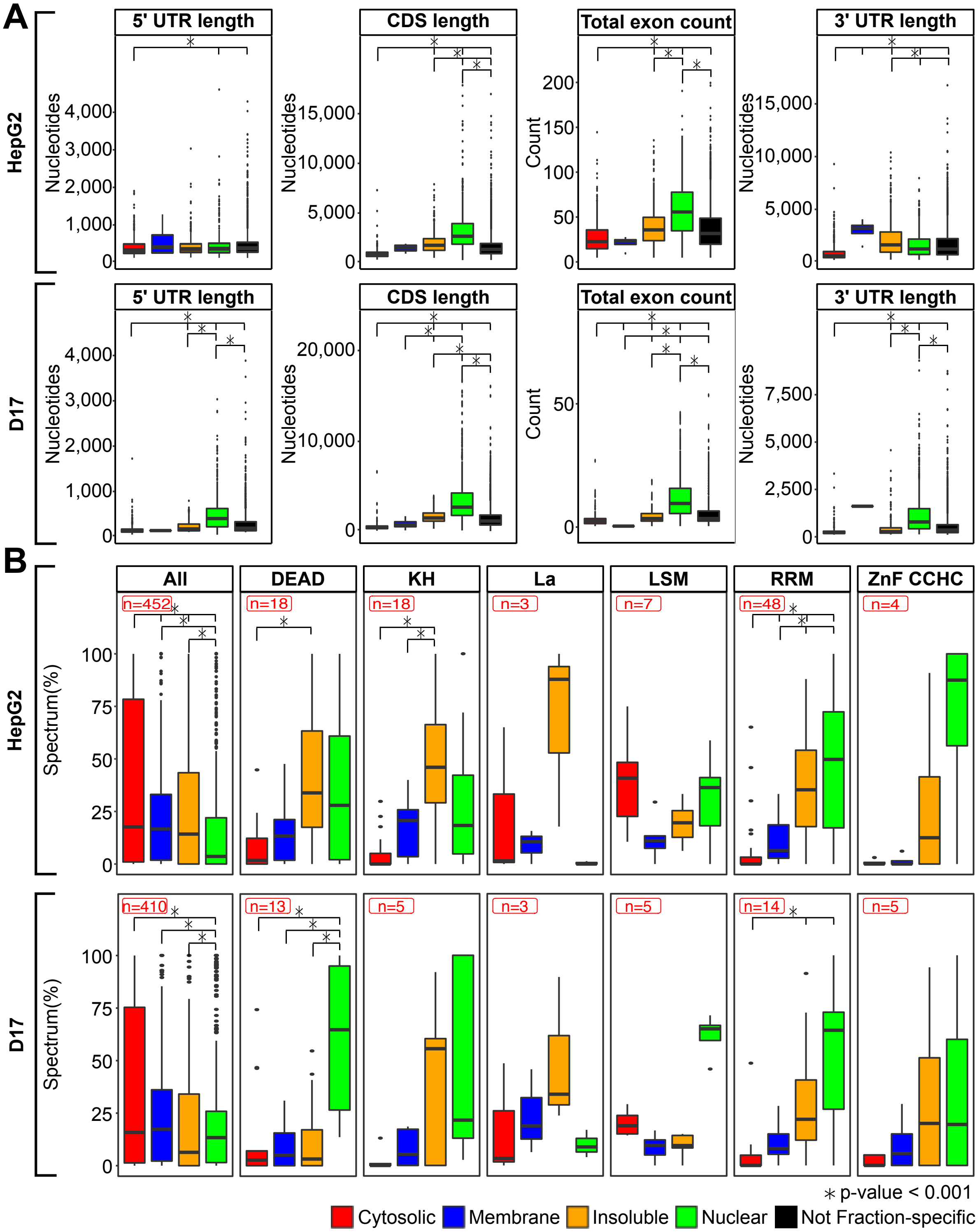


FIGURE 3.7. Subcellular fractions are enriched for mRNAs with distinctive sequence features and RNA binding protein (RBP) families in both human and *Drosophila* cells. (A) Boxplots of the longest isoform lengths of the indicated RNA regions (5' UTR, CDS or 3' UTR) and total exon numbers for fraction-specific and not fraction-specific mRNAs in HepG2 (*upper panel*) and D17 (*lower panel*) cell fractions. **(B)** Boxplots indicating the relative fraction distribution profiles of all RBP with identified orthologs in HepG2 and D17 or specific subfamilies of these RBP bearing distinctive RNA binding domains, in both HepG2 (*upper panel*) and D17 (*lower panel*).

domain (RBDs). To assess whether RBPs also exhibit conserved subcellular distribution features between human and fly, we next evaluated the fractionation profiles of specific classes of factors defined by the type of RBD they contain. Through orthology searches enabling one-to-many or many-to-many relationships, we were able to define a conserved set 410 and 452 orthologous RBPs respectively represented in our D17 and HepG2 LC-MS/MS datasets. As a whole, the pool of orthologous proteins displayed higher peptide abundance in the cytosol, with a lower expression in the membrane, insoluble or nuclear fractions (Figure 3.7B, left-most boxplot). This distribution was similar to that exhibited by the total pool of human or fly proteins detected in our LC-MS/MS datasets (Figure 3.S9). From the list of protein orthologs, we then sub-classified proteins with known RBDs, as defined within the Pfam and Interpro databases, and evaluated their distribution profiles in human or *Drosophila* cells (Figure 3.7B). Interestingly, RBPs belonging to different families exhibited similar subcellular distribution profiles in the interrogated cellular models. Moreover, the profiles exhibited by the different RBD families were distinctive. Indeed, several families of RBPs, including those containing KH, RRM, DEAD and CCHC-zinc finger domains, were enriched within the insoluble or nuclear fractions. In contrast, proteins bearing the La motifs, implicated in transcription and cell proliferation, tend to localize in the insoluble fraction ; whereas proteins bearing LSM domains were enriched in the cytosol and nucleus. As the distinctive asymmetric distribution patterns of RBP families appear to be generally conserved evolutionarily, this may explain why transcriptome distribution properties are also generally conserved.

3. Discussion

In the last few decades, the subcellular localization of RNA molecules has emerged as an important step in post-transcriptional gene regulation, impacting many biological processes that rely on polarized intracellular activities. However, the general prevalence of RNA asymmetry on a transcriptome-wide scale has remained unclear. To address this question, we utilized herein a cell fractionation and RNA-sequencing strategy, termed CeFra-seq, to probe the RNA content of subcellular compartments sequentially generated from starting cellular populations of human and *Drosophila* cells. This comparative profiling approach shows that isolated compartments exhibit distinctive profiles of RNA biotype composition and that these asymmetries are both highly prevalent and evolutionarily conserved.

Biochemical purification and high-throughput RNA expression analysis, either using microarray or RNA-seq as a read out, has been employed in several studies to identify specific populations of RNAs associated with structures such as the nucleus [42, 5, 68, 1], cytoplasm [5, 1, 10], cytosol [72, 68, 13, 69], the ER [38, 20, 45, 19, 77, 17], mitochondria [48], microtubules [6, 65], pseudopodia [53], and neuronal projections [22, 36, 55, 58, 26, 37, 63, 67]. However, these efforts generally focused on defining enriched transcript populations associated with single structures, without simultaneously probing RNA expression signatures across other compartments derived from the same starting cellular specimens, thus limiting the capacity to evaluate global RNA localization prevalence. By employing a comparative multi-compartment profiling strategy, we demonstrate that the majority of cellular RNA transcripts ($> 80\%$) are asymmetrically distributed, whether considering mRNAs or noncoding transcripts, patterns that appear conserved evolutionarily. While previous FISH-based imaging screens in syncytial embryos of *Drosophila* revealed a high degree of localization ($\sim 70\%$) among $\sim$6000 interrogated mRNA species [41, 73], the present study offers one of the most comprehensive surveys of transcriptome subcellular distribution to date, revealing that this phenomenon is a basic feature of cellular organization that can be generalized to standard cellular models.

We found that mRNAs targeted to different regions of the cytoplasm exhibit distinctive features in terms of the overall length of their coding regions and 3' UTRs ; with the cytosolic fraction preferentially composed of shorter mRNAs with lower exon complexity and shorter 3' UTR segments, while membrane and insoluble compartment mRNAs are longer and more complex. Since RNA localization and stability control elements often reside in mRNA 3' UTRs, this data suggests a model whereby targeting of mRNAs to membrane and insoluble (cytoskeletal) compartments may involve more elaborate regulatory information. In contrast, localization of RNAs to the cytosol, which can be achieved by simply exporting the mRNA from the nucleus, likely requires simpler targeting information. For example, mRNAs encoding histones and ribosomal proteins, which tend to be short and involve special regulatory mechanisms, exhibit preferential cytosolic targeting. Simple cytosolic targeting may prove beneficial for proteins that must reenter the nucleus to carry out their functions, as in the case of histones, which are central to chromatin formation, or for ribosomal proteins that are required in the nucleus for ribosome subunit assembly [39].

Our results also extend the assessment of RNA localization prevalence to noncoding components of the transcriptome, such as lncRNAs and circRNAs. The steady-state nuclear versus cytoplasmic distribution features of lncRNAs has been an issue of debate, with early reports suggesting an enrichment for these transcripts in the nucleus [56, 18], where they have been implicated in the epigenetic regulation of gene expression ; while more recent studies documented significant lncRNA signatures within cytoplasmic compartments and in association with ribosomes [10, 69, 30]. The CeFra-seq approach reveals higher representation of lncRNA reads within the cytosolic and nuclear fractions, and a general under-representation in membrane and insoluble compartments, signatures that were similar in human and fly cells. These findings are consistent with the notion that subcellular fates of lncRNAs are diverse [8, 12], possibly enabling these transcripts to carry out distinct regulatory functions in specific intracellular locales. In the case of circRNAs, early reports suggested that they tended to be nuclear-enriched [76], although recent studies have shown that these transcripts can associate with cytoplasmic ribosomes and undergo translation [57, 44]. Our results, gleaned by analyzing read coverage at intronic locations known to produce circRNAs and using a back-spliced junction read mapping algorithm, reveal a higher expression level of circRNAs within the cytosolic and membrane fractions, suggesting possible regulatory functions outside the nucleus. Altogether, these results highlight the potential usefulness of CeFra-seq methodology to segment noncoding RNAs into subgroups that may share common functional properties or interact physically to modulate cellular function.

Previous studies comparing global mRNA and protein expression signatures have generally revealed moderate levels of expression correlation, suggesting that post-transcriptional regulatory steps (e.g., translation rates, mRNA and protein decay) are a primary determinant of proteomic output of the transcriptome [47, 70, 64]. In this study, we sought to assess the potential relationship between mRNA and protein expression at the subcellular level by jointly profiling their expression signatures across our fractionated compartments. While the expression correlations within fractions for specific mRNA/protein pairs were generally good, comparison of asymmetry measurements (i.e., their distribution profiles across fractions) revealed a general absence of correlation, consistent with a previous study contrasting mRNA/protein profiles in cell bodies versus protrusions of migratory cells [49]. In light of the data presented here, it is clear that these overall distribution correlations are likely to be heavily influenced by examples of mRNA/protein pairs with strong asymmetric distribution. This led us to analyse the distribution signatures of mRNAs encoding components of well-defined protein complexes. A first striking characteristic to emerge is that functionally related mRNA subsets tend to cluster together in these graphs, as do the protein modules they specify, implying the existence of coherent sorting mechanisms. The second feature is that there is broad variability in co-distribution profiles of mRNA/protein sets depending on the protein modules under consideration, ranging from cases with more proximal targeting of mRNA/protein pairs (e.g., *Arp2/3*, *Vigilin*) and others that were very distant (e.g., ribosome, proteasome). This suggests that protein modules exist on a gradual continuum of codistribution with their encoding mRNAs, while also underlining the notion that regulated mRNA localization may serve different purposes mechanistically. Cases in which mRNA/protein pairs cofractionate are likely to represent instances in which localized translation contributes to the assembly of protein complexes, as has been shown for components of the *Arp2/3* complex [**54**]. In contrast, for mRNA/protein pairs exhibiting divergent fractionation behavior, i.e., with apparent steady-state accumulation in distinct subcellular locales, this may underlie cases where transcripts are subject a generalized storage mechanism or to localized repression, which may be altered under specific contexts, such as cellular state transitions, during the cell cycle or in response to environmental signaling cues [**59, 29**].

In summary, the CeFra-seq methodology presented herein offers an efficient approach to interrogate global subcellular transcriptome distribution features, as well as parallel analysis of recovered protein samples. In addition to offering insights into subcellular transcriptome targeting, the approach can allow detection of rare transcripts that display low overall cellular abundance, but may become detectable when profiling specific subcellular compartments. Indeed, we identified a significant number of transcripts that were fraction exclusive and otherwise would have escaped detection if solely focusing on whole-cell profiling at similar sequencing depth. In that sense, CeFra-seq may offer similar advantages to the capture-seq methodologies developed to deeply survey RNA species synthesized from precise genomic loci [**51**]. In light of the growing number of diseases in which RNA localization defects, CeFra-seq methodology will prove extremely useful for dissecting the specific molecular alterations associated with these disorders.

4. Material and Methods

4.1. Cell culture and antibodies

The ML-D17c3 cell line, stock 107, was obtained from the *Drosophila* Genomics Resource Center cell line repository (http://dgrc.cgb.indiana.edu/). Cells were grown on pre-treated tissue culture dishes with extracellular matrix as described [16] in M3 media (S-8398, Sigma-Aldrich) containing 1mg/ml of yeast extracts (Y-1000, Sigma-Aldrich) and 2.5mg/ml of bactopeptone (211677, Difco), 10% FBS (SH30070.02, Hyclone) and 10mg/ml of insulin (I0516-5ml Sigma). HepG2 cell line was kindly provided by B. Graveley (Institute for Systems Genomics, UCONN Health Center, Farmington, CT, USA) and maintained in Dulbecco's modified Eagle's medium (DMEM) (SH30022.01, Hyclone) supplemented with 10% FBS and 1% penicillin/streptavidin (15140-163, Invitrogen). Rabbit polyclonal anti-Histone H3 (ab1791) and mouse monoclonal anti-α-tubulin (clone DM1A) were obtained from Abcam and Sigma-aldrich respectively. Mouse monoclonal anti-KDel (ADI-SPA-827), mouse monoclonal anti-Shot (mABRod1) and mouse monoclonal anti-ninein (clone F-5) were purchased from Enzo Life Sciences, the Developmental Studies Hybridoma Bank and Santa Cruz, respectively.

4.2. Cell fractionation procedure

D17 (3.5×10^7) and HepG2 (2.5×10^7) cells were used for the fractionation procedure, as described in Lefebvre et al. [43]. Briefly, after PBS washes, 1/10$^{\text{th}}$ of the cells are kept aside as Total extract, and the remaining cells were resuspended and incubated in cold Hypotonic Buffer (20 mM Tris HCl (pH=7.5),10 mM KCl, 1.5 mM MgCl2, 5 mM EGTA, 1 mM EDTA, 1 mM DTT, 1 mM PMSF, 0.15 U/mL Aprotinin, 20µM Leupeptin, 40 U/mL RNase Out (Thermo Fisher Scientific) for 20 minutes. Swelled cells were transferred into a homogenizer chamber and dounced for 15 strokes for D17 and 5 strokes for HepG2. After centrifugation of the homogenate at 1200×g for 10 minutes at 4°C, the supernatant corresponding to the cytosolic fraction was conserved apart while the pellet was rinsed with 100µl of hypotonic buffer and mixed with 0.5ml of Sucrose Buffer 0.32M (0.32 M Sucrose, 3 mM CaCl2, 2 mM MgOAc, 0.1 mM EDTA, 10 mM Tris-HCl (pH=8),1 mM DTT, 0.5% v/v NP-40, protease inhibitors and RNase out) and 0.5ml of Sucrose Buffer 2.0M (2.0M Sucrose, 5 mM MgOAc, 0.1 mM EDTA, 10 mM Tris-HCl (pH=8), 1 mM DTT, protease inhibitors and RNase out).

The sucrose homogenate was loaded on top of 0.5ml of sucrose 2.0M in a polyallomer tube and centrifuged at 30,000×g for 30 minutes at 4°C in a Sorvall RPS55 rotor to collect a pellet corresponding to the nuclear fraction. The cytosolic, membrane and Insoluble fractions were prepared from the cytosolic homogenate by centrifugation at 100,000×g for 1 hour at 4°C in a Sorvall RP100 AT4 rotor. The supernatant was collected as the cytoplasmic fraction while the pellet was resuspended in 1ml of Hypotonic Lysis Buffer containing 1% Triton X-100, dounced for 40 times on ice and incubated in ice for 1 hour. After centrifugation at 100,000×g for 30 minutes at 4°C in the Sorvall RP100 AT4 rotor, the supernatant was collected as the membrane fraction and the pellet was rinsed and used as the insoluble fraction.

4.3. RNA and protein extractions

At each steps of the cell fractionation, the collected supernatants and pellets were immediately resuspended in 1ml of TRIzol-LS or TRIzol respectively. RNAs were isolated following TRIzol extraction procedure from the aqueous phase and resuspended in water, while proteins were extracted from the organic phase. For the proteins extraction, 0.3ml of ethanol was added to 0.6ml of organic phase and incubated 5 minutes at 25°C. After centrifugation at 3000×g for 5 minutes at 4°C, the supernatant was mixed with 0.750 ml of isopropanol, incubated for 10 minutes at 25°C and centrifuged at 13,000×g for 10 minutes at 4°C. The pellet was washed 3 times for 20 minutes at 25°C with 1 ml of 0.4M Guanidine hydrochloride in 95% isopropanol and once with 1 ml of ethanol 75%. The pellet was finally resuspended in 0.5M unbuffered Tris containing 5% SDS.

4.4. RT-qPCR and Western blot validations

RNA extracts from each fraction were subjected to reverse transcription using random hexamers and RT-MMLV (Invitrogen) followed by real-time (RT) quantitative PCR (qPCR) analyses using gene-specific primer pairs for : SNORD17 - Fw :5'-CTG CCA ACA CAC AAG CAG TT-3' ; Rv :5'-CTT GCA GCC TTG TGA AAT GA-3' RN7SK - Fw :5'-CCA TTT GTA GGA GAA CGT AGG-3' ; Rv :5'- CCT CAT TTG GAT GTG TCT GG-3' MT-CO1 - Fw :5'-CAA ACC ACA AAG ACA TTG GAA ; Rv :5'-GCA CCG ATT ATT AGG GGA AC-3' TJP1 - Fw :5'-GCT TAC CAC ACT GTG ATC CT ; Rv :5'-CAC AGT TTG CTC CAA CGA G-3' hsr-omega- Fw :5'-CCA CAA CAA AAT GAA CCA CAA ; Rv :5'-CAA TTT TGA ATT GGG GCA GT-3' Rpl23a- Fw :5'-GTG AAG CCC GTG ACC AAG ;

Rv :5'-AGG CGC CCT TGA TGA TCT-3' mt-NDF6 - Fw :5'-TCA TCC ATT AGC TTT AGG ATT AAC TTT-3' ; Rv : 5'-TTT CAT TAG AGG CTA AAG ATG TTA CG-3' dlg-1 - Fw :5'-CTG GAT AAG CAA TCG ACA TTG G-3' ; Rv : 5'- CAT TCT TCT CAT CGC GAC TC-3'

Quantitative PCR analyses were performed using PowerUp SYBR Green Master Mix kit (Applied Biosystem, Thermo Fisher Scientific) on Viia7 Real-Time PCR system (Applied Biosystem, Thermo Fisher Scientific). For the Western Blot, protein extracts from each fraction were loaded on a 11% SDS-PAGE gel and transferred onto nitrocellulose membranes. Following incubation (16 hours, 4°C) with primary antibodies corresponding to fraction specific markers, blots were washed and incubated with appropriate HRP-conjugated secondary antibodies for 1 hour. Signals were detected by enhanced luminescence (Clarity Western ECL Substrate, BIO-RAD) with the Gel Doc XR+ imaging system (BIO-RAD).

4.5. Library generation and RNA sequencing

Before library generation, the quality of the RNAs extracted from each sample was validated using the Agilent 2100 Bioanalyzer device and the RNA 6000 Pico Chip. RNA-seq libraries were prepared with Illumina TruSeq mRNA Stranded kit from rRNA depleted RNA samples (Ribo-Zero^{TM} Magnetic Gold Kit for Human and *Drosophila* kit, Epicentre) or from poly-A enriched RNA samples (NEBNext poly(A) mRNA, New England Biolabs). Deep sequencing was performed using the Illumina HiSeq2000 sequencer (paired-end 50 cycles).

4.6. *In silico* analysis of RNA sequencing and proteomics data

Read quality was assessed using FastQC v0.11.5. No trimming was deemed necessary. Read alignment was executed using TopHat v2.1.0 on the human GRCh37/hg19 and the *Drosophila* BDGP5.78/dm3 genomes respectively. Read count was obtained with featureCounts v1.5.0-p1. Normalized FPKM values and differential expression was computed with DESeq2 v1.10.1. Metrics about the alignment were obtained with picard CollectRnaSeqMetrics program. We only considered transcripts with fragments per kilobase per million mapped reads (FPKM) ≥ 1. Percent FPKM ($pFPKM_i = FPKM_i / \sum_{k \in \{C,M,I,N\}} FPKM_k$) where i is a given gene in a given fraction) was calculated as the relative distribution unit. Transcript per million ($TPM_x = (FPKM_x / \sum FPKM_y) \times 10^6$), where x is a given gene in a given fraction and y represent all the genes observed in this fraction) was calculated as the relative abundance unit. We grouped as "other" all biotypes where the highest TPM in any given fraction was below 1%.

Attributes such as biotypes and longest isoform length was obtained via the R biomaRt package [21]. Expected gene length was obtained by calculating the normalized weighted average of each gene. For each individual fraction, this is $\sum_i^j TPM_i \times L_i / \sum_i^j TPM_i$ Where TPM_i is the expression value of a given gene i, L is its length and j the total number of gene of a given biotype. All calculations and correlations where performed using R.

For proteomics data, spectrum count and probability was calculated with scaffold v4.4.8. We only conserve protein with a minimum number of 2 peptides and a peptide threshold and a protein probability of 95%. Gene ontology term statistical overrepresentation test was performed using Panther v11 [52]. Protein complex were obtained from the CORUM databases release 30-10-2016 [62].

4.7. Orthologs associations

We retrieved orthologous genes between fly and human via the R biomaRt package selecting those with high confidence [21]. We then filtered this list to keep only the genes with at least 1 FPKM in at least one fraction for both species.

4.8. Circular intronic RNA analysis

We downloaded the list of 103 circular intronic RNA identified and characterized by Zang et al. as a bed file from circbase and reported the counts of alignment in our bam files with bedtools multicov [76, 25, 61].

4.9. Regular 3-simplex (tetrahedron) representation of cellular compartments

Gene asymmetry within a 3-simplex space was obtained by computing a vector resulting from the relative distribution of a given gene from each fraction projected into each compartment represented by the 3-simplex. Assuming a 3-simplex centered at the origin with coordinates $V1 = (-\frac{1}{3}, -\frac{\sqrt{2}}{3}, -\sqrt{\frac{2}{3}}), V2 = (-\frac{1}{3}, -\frac{\sqrt{2}}{3}, \sqrt{\frac{2}{3}}), V3 = (-\frac{1}{3}, \frac{\sqrt{8}}{3}, 0), V4 = (1,0,0)$, in Cartesian space, this is defined as $X = pFPKM_{cyto} \times (-\frac{1}{3}) + pFPKM_{membr} \times (-\frac{1}{3}) + pFPKM_{insol} \times (-\frac{1}{3}) + pFPKM_{nucl} \times (1), Y = pFPKM_{cyto} \times (-\frac{\sqrt{2}}{3}) + pFPKM_{membr} \times (-\frac{\sqrt{2}}{3}) + pFPKM_{insol} \times (\frac{\sqrt{8}}{3}), Z = pFPKM_{cyto} \times (-\sqrt{\frac{2}{3}}) + pFPKM_{membr} \times (sqrt\frac{2}{3})$

4.10. Prediction back-spliced junctions for circular RNAs

Previously unmapped reads were re-aligned with tophat-fusion (–fusion-search –keep-fasta-order –bowtie1–no-coverage-search) and then we applied the circExplorer algorithm to identify putative circRNA [76].

4.11. Accession numbers

Raw sequencing data are available on the ENCODE portal (https://www.encodeproject.org/) under the experiment ID numbers : ENCSR931WGT (HepG2-cytosolic-PA) ; ENCSR541TIG (HepG2-membrane-PA) ; ENCSR019MXZ (Hepg2-insoluble-PA) ; ENCSR058OSL (HepG2-nuclear-PA) ; ENCSR862HPO (Hepg2-cytosolic-RD) ; ENCSR887ZSY (HepG2-membrane-RD) ; ENCSR813BDU (HepG2-insoluble-RD) ; ENCSR061SFU (HePG2-nuclear-RD) ; ENCSR283YJX (D17-cytosolic-PA) ; ENCSR053CWY (D17-membrane-PA) ; ENCSR622ROA (D17-insoluble-PA) ; ENCSR473SBP (D17-nuclear-PA) ; ENCSR432GTP (D17-cytosolic-RD) ; ENCSR302HSE (D17-membrane-RD) ; ENCSR772QDO (D17-insoluble-RD) ; ENCSR197ZHM (D17-nuclear-RD).

5. Authors' contributions

Conceptualization, E.L and N.A.L., L.P.B.B. ; Methodology, L.P.B.B., N.A.L. ; Investigation, L.P.B.B., N.A.L, J.B., F.A.L., C.D., X.W. ; Software, L.P.B.B. ; Formal Analysis, L.P.B.B. ; Visualization, L.P.B.B. ; Writing – Original Draft, L.P.B.B., E.L. ; Writing – Review & Editing, L.P.B.B., J.B, C.D, M.B., E.L. ; Funding Acquisition, E.L. ; Supervision, E.L. M.B.

6. Acknowledgements

We thank Christopher Burge (MIT), Eric T. Wang (University of Florida), Brenton Graveley, Michael Duff, Xintao Wei (UCONN Health Center), Sebastian Markmiller and Gene Yeo (UCSD), and Atefeh Mohajeri (McGill University) for insightful discussions about CeFra-seq over the years. We are grateful to Alexis Blanchet-Cohen and Odile Neyret (IRCM) for their help with transcriptomic analyses.

7. Funding

F.A.L. is supported by a Canadian Institute of Health Research (CIHR) scholarship, J.B. and C.D. by postdoctoral scholarships from the Fonds de Recherche Québec Santé (FRQS), while E.L. is a Junior 2 FRQS research scholar. This work was supported by grants to E.L. from CIHR, FRQS and the Canadian Cancer Society.

8. Conflict of interest

The authors declare no competing interests.

9. Supplemental Data

TABLE 3.S1. RNA-seq read statistics for Human HepG2 and *Drosophila* D17 subcellular fractions.

Library	Species	Fraction	Replicate	Number of reads	Duplicate (%)	Number of Aligned reads	Alignment (%)	Multiple alignment	Average FPKM
rRNA-depletion	*Drosophila*	Cytosolic	1	20,698,062	68.32	18,876,632	91.2	53.4	1.80×10^6
rRNA-depletion	*Drosophila*	Cytosolic	2	24,285,303	71.73	22,488,190	92.6	55.2	
rRNA-depletion	*Drosophila*	Membrane	1	21,843,697	43.58	19,069,547	87.3	17.1	0.61×10^6
rRNA-depletion	*Drosophila*	Membrane	2	21,134,180	44.12	18,513,541	87.6	18.4	
rRNA-depletion	*Drosophila*	Insoluble	1	22,607,476	30.85	19,781,541	87.5	23.3	0.34×10^6
rRNA-depletion	*Drosophila*	Insoluble	2	24,059,152	32.19	20,979,580	87.2	24.4	
rRNA-depletion	*Drosophila*	Nuclear	1	22,996,271	47.03	18,902,934	82.2	56.8	0.44×10^6
rRNA-depletion	*Drosophila*	Nuclear	2	24,902,681	43.5	20,694,127	83.1	52.8	
rRNA-depletion	Human	Cytosolic	1	32,552,634	64.76	27,344,212	84	24.6	3.20×10^6
rRNA-depletion	Human	Cytosolic	2	30,564,649	67.55	25,246,400	82.6	24.3	
rRNA-depletion	Human	Membrane	1	35,909,087	40.04	30,774,087	85.7	17.5	0.62×10^6
rRNA-depletion	Human	Membrane	2	35,061,741	39.84	30,819,270	87.9	16.5	
rRNA-depletion	Human	Insoluble	1	34,097,076	20.53	31,539,795	92.5	8.1	0.26×10^6
rRNA-depletion	Human	Insoluble	2	35,451,428	22.57	32,757,119	92.4	7.6	
rRNA-depletion	Human	Nuclear	1	35,747,174	10.16	32,065,215	89.7	7	0.24×10^6
rRNA-depletion	Human	Nuclear	2	37,390,540	11.56	33,913,219	90.7	6.8	
PolyA+	*Drosophila*	Cytosolic	1	22,668,517	53.28	20,628,350	91	7.4	2.10×10^6
PolyA+	*Drosophila*	Cytosolic	2	19,499,648	53.13	17,569,182	90.1	8	
PolyA+	*Drosophila*	Membrane	1	25,331,304	55.33	22,139,559	87.4	7.8	1.30×10^6

Library	Species	Fraction	Replicate	Number of	Duplicate	Number of	Alignment	Multiple	Average
				reads	(%)	Aligned reads	(%)	alignment	FPKM
PolyA+	*Drosophila*	Membrane	2	21,260,027	51.94	18,474,963	86.9	9.8	
PolyA+	*Drosophila*	Insoluble	1	20,515,878	42.19	17,664,170	86.1	34.7	0.42×10^6
PolyA+	*Drosophila*	Insoluble	2	20,678,208	41.61	17,783,258	86	32.9	
PolyA+	*Drosophila*	Nuclear	1	23,076,204	48.43	18,460,963	80	55.7	0.25×10^6
PolyA+	*Drosophila*	Nuclear	2	21,195,707	47.52	16,998,957	80.2	54.5	
PolyA+	Human	Cytosolic	1	31,719,232	53.46	30,101,551	94.9	6.8	0.99×10^6
PolyA+	Human	Cytosolic	2	24,755,754	49.65	22,948,583	92.7	6.4	
PolyA+	Human	Membrane	1	24,235,894	37.53	21,885,012	90.3	13.5	0.46×10^6
PolyA+	Human	Membrane	2	20,236,753	35.82	18,840,417	93.1	6.4	
PolyA+	Human	Insoluble	1	25,382,072	17.75	22,996,157	90.6	12.6	0.21×10^6
PolyA+	Human	Insoluble	2	23,337,888	17.33	21,354,167	91.5	12.2	
PolyA+	Human	Nuclear	1	21,059,227	11.22	19,795,673	94	6.5	0.19×10^6
PolyA+	Human	Nuclear	2	19,884,506	12.69	18,731,204	94.2	6.6	

Table 3.S2. Cell component gene ontology (GO) enrichments of HepG2 cell fraction-specific proteins.

Cytosolic	
Gene Category	**p-value**
cytosol	3.37E-81
cytoplasm	6.01E-57
extracellular vesicle	2.48E-50
membrane-bounded vesicle	3.94E-36
cytoplasmic part	1.05E-34
vesicle	8.38E-34
extracellular region	3.46E-22
nucleus	1.64E-11
nucleoplasm	1.65E-05
nuclear lumen	6.17E-05
nuclear part	7.35E-05
intracellular membrane-bounded organelle	4.80E-04
mitochondrion	6.90E-04
intracellular organelle	1.64E-03
intracellular organelle lumen	2.52E-03

Membrane	
Gene Category	**p-value**
endomembrane system	3.34E-47
bounding membrane of organelle	3.43E-39
endoplasmic reticulum	4.40E-37
vesicle	1.70E-23
extracellular exosome	2.18E-22
extracellular membrane-bounded organelle	2.24E-22
extracellular vesicle	3.06E-22
extracellular organelle	3.14E-22
cytoplasm	5.34E-18
membrane-bounded organelle	2.34E-17
extracellular region part	2.77E-16
Golgi apparatus	3.99E-14
integral component of membrane	6.22E-14
intrinsic component of membrane	2.55E-13
endosome	2.02E-08
cytoplasmic vesicle	5.05E-08
vacuole	6.52E-07
vacuole part	1.14E-05
lysosome	4.41E-05
lytic vacuole	4.71E-05
whole membrane	5.83E-05

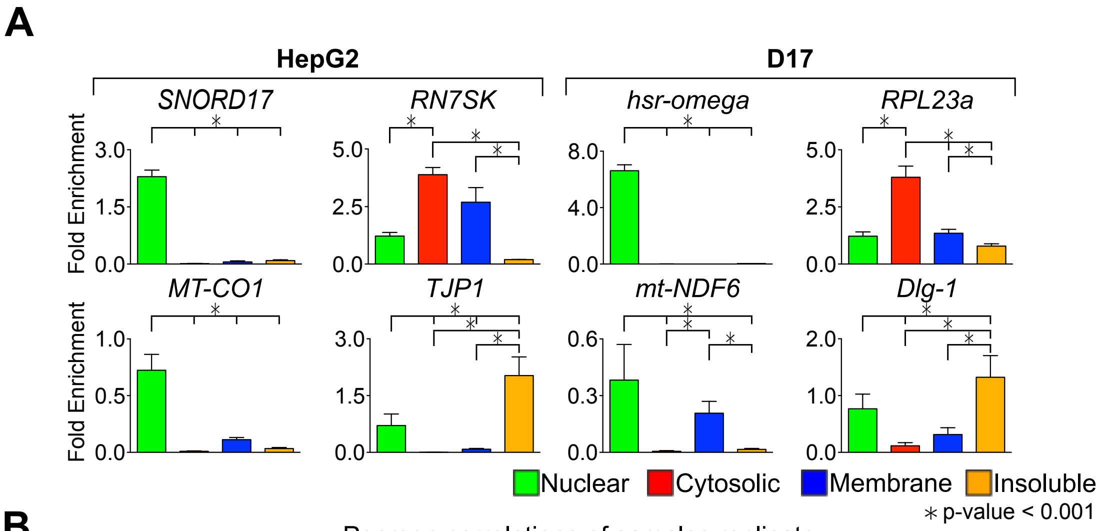
Insoluble	
Gene Category	**p-value**
macromolecular complex	1.89E-08
cytoplasmic part	4.84E-08
ribonucleoprotein complex	4.86E-08
intracellular ribonucleoprotein complex	4.86E-08
intracellular organelle part	6.79E-06
organelle part	2.38E-05
cytoplasm	2.59E-05
intracellular non-membrane-bounded organelle	1.76E-04
cytosol	1.87E-04
non-membrane-bounded organelle	1.92E-04
intracellular part	1.95E-04
organelle	2.77E-04
membrane-bounded organelle	4.48E-04
RNAi effector complex	4.94E-04
cytoplasmic ribonucleoprotein granule	5.25E-04
intracellular membrane-bounded organelle	6.33E-04
ribonucleoprotein granule	7.52E-04
polysome	1.98E-03
protein complex	2.71E-03

Nucleus	
Gene Category	**p-value**
membrane-enclosed lumen	3.48E-97
nuclear part	3.74E-92
nucleus	8.07E-66
membrane-bounded organelle	9.55E-64
nucleoplasm	5.96E-44
nucleolus	1.34E-42
chromosome	4.33E-36
organelle envelope	4.86E-34
envelope	1.51E-33
chromosomal part	9.82E-32
protein complex	1.11E-27
mitochondrial part	4.09E-26
preribosome	1.86E-25
chromatin	1.36E-20
chromosomal region	1.39E-19
nuclear chromosome	2.14E-19
ribonucleoprotein complex	2.88E-15
intracellular ribonucleoprotein complex	2.88E-15
chromosome, centromeric region	5.05E-15
nuclear pore	3.38E-14
SWI/SNF superfamily-type complex	4.31E-14
condensed chromosome	2.23E-08
kenetochore	2.64E-08
nuclear envelope	2.71E-07
histone methyltransferase complex	3.18E-07
nuclear pore outer ring	5.18E-07
chromosome, telomeric region	2.11E-06
PcG protein complex	4.45E-06
transcriptional repressor complex	5.11E-06
methyltransferase complex	5.11E-06
transferase complex	7.19E-06

Table 3.S3. Cell component gene ontology (GO) enrichments of HepG2 cell fraction-specific mRNAs.

Cytosolic		**Membrane**	
Gene Category	**p-value**	**Gene Category**	**p-value**
ribosome	3.50E-50	cytosolic ribosome	3.90E-31
cytosolic part	3.20E-46	ribosomal subunit	3.00E-22
ribonucleoprotein complex	5.80E-35	cytosolic part	7.20E-22
small ribosomal subunit	2.10E-28	ribosome	1.40E-18
large ribosomal subunit	5.70E-27	cytosol	3.50E-13
cytosol	6.10E-19	large ribosomal subunit	5.00E-13
mitodhondrion	5.30E-15	small ribosomal subunit	2.30E-09
mitochondrial part	3.90E-11	ribonucleoprotein complex	5.50E-09
mitochondrial membrane	4.90E-10	endoplasmic reticulum	3.40E-05
mitochondrial envelope	1.10E-09	endosome	8.10E-05
mitochondrial ribosome	3.80E-05	transcription factor complex	4.30E-04
non-membrane-bounded organelle	1.00E-04	nucleoplasm part	8.70E-04
respiratory chain	4.10E-04	extrinsic to membrane	1.10E-03
		Golgi apparatus part	7.10E-03
		organelle subcompartment	7.10E-03

Insoluble		**Nucleus**	
Gene Category	**p-value**	**Gene Category**	**p-value**
nucleoplasm	1.20E-43	extracellular matrix part	5.45E-06
nucleolus	1.00E-31	nuclear lumen	6.60E-06
chromosome	3.00E-24	membrane-enclosed lumen	1.40E-05
microtubule cytoskeleton	2.50E-22	basemen membrane	2.20E-05
nucleoplasm part	2.10E-20	nuclear speck	2.70E-05
nuclear body	3.00E-20	nucleoplasm	6.50E-05
spliceosome	1.10E-19	emdomembrane system	8.40E-05
spindle	3.70E-19	endoplasmic reticulum part	9.30E-05
chromosomal part	6.50E-18	collagen	1.00E-04
condensed chromosome	1.10-E-17	nucleoplasm part	2.90E-04
ribonucleoprotein complex	9.10E-16	anchoring collagen	2.30E-03
cytoskeleton	1.20E-15	cell-cell adherens junction	2.40E-03
kinetochore	7.10E-14	nuclear body	3.40E-03
nuclear speck	1.70E-12	anchoring juntion	4.50E-03
nuclear chromosome	2.60E-12	microtubule cytoskeleton	4.60E-03
nuclear pore	3.00E-10	nuclear envelope	4.80E-03
microtubule organizing center	6.20E-10	cell-cell junctions	7.30E-03
centrosome	9.30E-10		
nuclear envelope	1.90E-08		
proteasome complex	5.50E-08		



B

Pearson correlations of samples replicate

	HepG2		**D17**	
	PA	RD	PA	RD
Cytosolic	0.999	0.999	0.994	0.997
Membrane	0.998	0.999	0.995	0.996
Insoluble	0.998	0.999	0.995	0.985
Nuclear	0.996	0.994	0.997	0.996

Figure 3.S1. Fractionation validation and inter-replicate correlations. **(A)** RT-qPCR of RNAs sample controls show fractionation efficiency. The accumulation of the indicated RNA fraction-specific markers was assessed in HepG2 and D17 cells. **(B)** Summary table of inter-replicates Pearson correlations of transcript per million (TPM) values within each fractions for HepG2 and D17 cell RNA-seq samples. RD= rRNA-depletion, PA= poly-A+.

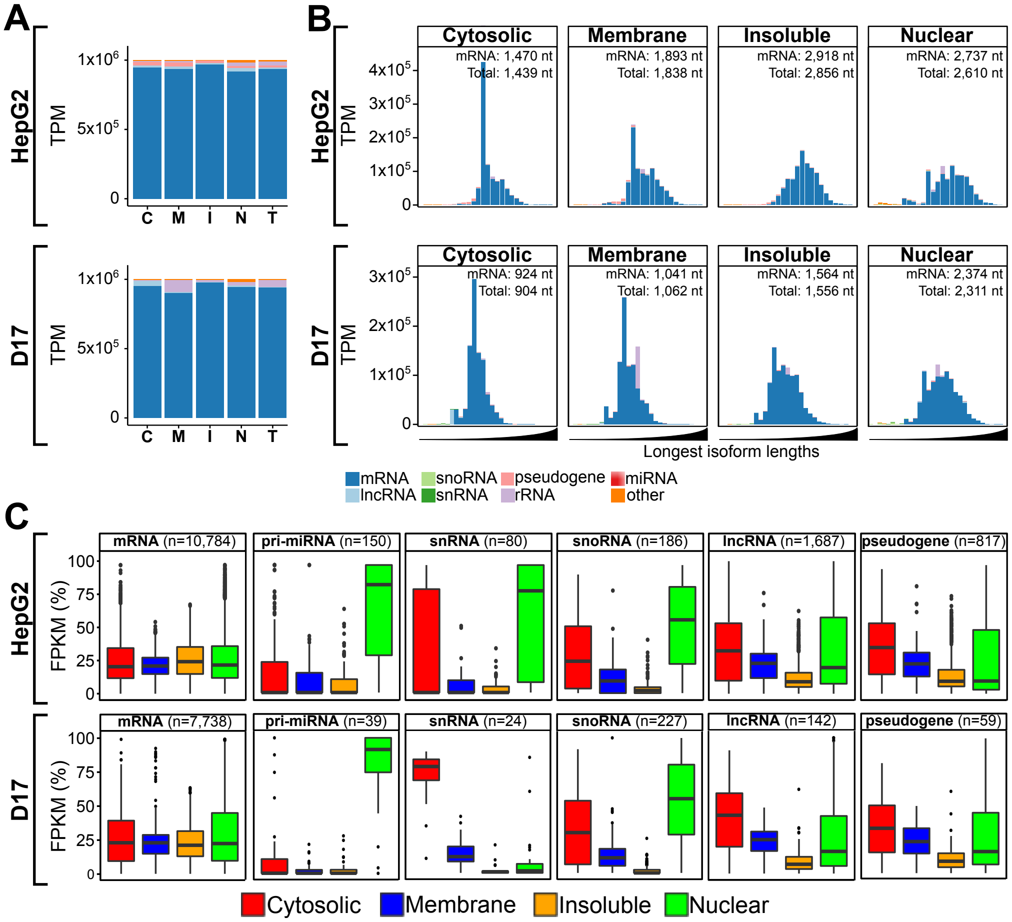


Figure 3.S2. Distinctive transcript composition of subcellular fractions in poly-A+ dataset. **(A)** Histograms depicting the RNA biotype content, in TPM, detected via PA sequencing of cytosolic (C), membrane (M), insoluble (I), and Nuclear (N) fractions or whole-cell RNA (T=total) from HepG2 (upper panel) and D17 cells (lower panel). **(B)** Histograms of the RNA biotype content of HepG2 and D17 cell fractions, binned according to the length of the longest annotated isoform of detected RNA species, following a log10 scale from 1.5-5 (i.e. ranging from 31-100,000 nt). The expected lengths for mRNA and total RNA populations are indicated for each fraction. For A and B, biotypes accounting for less than 1% of the overall TPMs were grouped as "other". **(C)** Boxplots showing the fraction distribution profiles of different RNA biotypes in percent FPKM (pFPKM) for HepG2 (upper plots) and D17 (lower plots) cells. The number (n) of transcripts analyzed for each biotype is indicated. TPM : Transcripts Per Millions ; FPKM : Fragment per kilobase per million ; lncRNA : long noncoding RNA ; mRNA : messenger RNA ; miRNA : microRNA ; miscRNA : miscellaneous other noncoding RNA ; snoRNA : small nucleolar RNA ; snRNA : small nuclear RNA ; rRNA : ribosomal RNA.

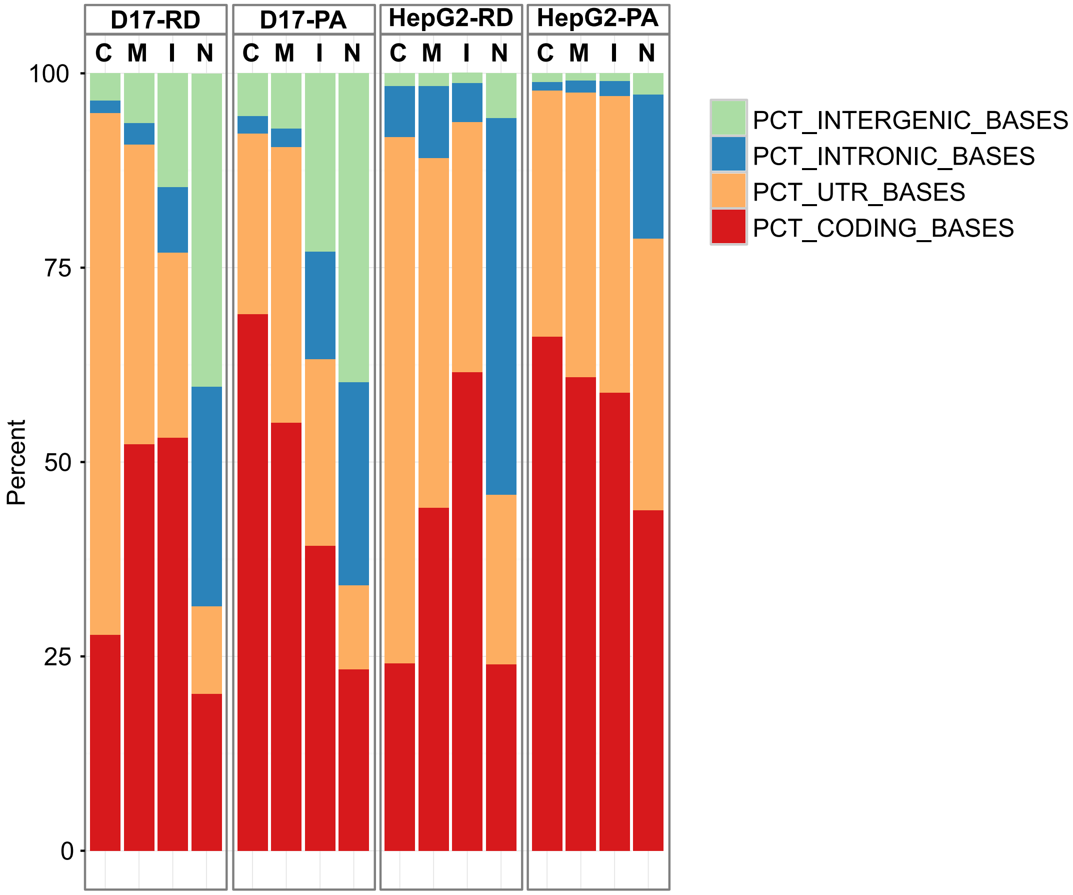


FIGURE 3.S3. RNA-seq read distribution profiles varies between fractions. Fraction of the total number fo aligned bases that mapped to either the protein coding region of genes, untranslated regions (UTR) of genes, gene introns, or intergenic regions of genomic DNA in HepG2 and D17 cells either assessed from poly(A)-enriched (PA) or rRNA-depleted (RD) sequencing datasets. C=Cytosolic, M=Membrane, I=Insoluble, N=Nuclear.

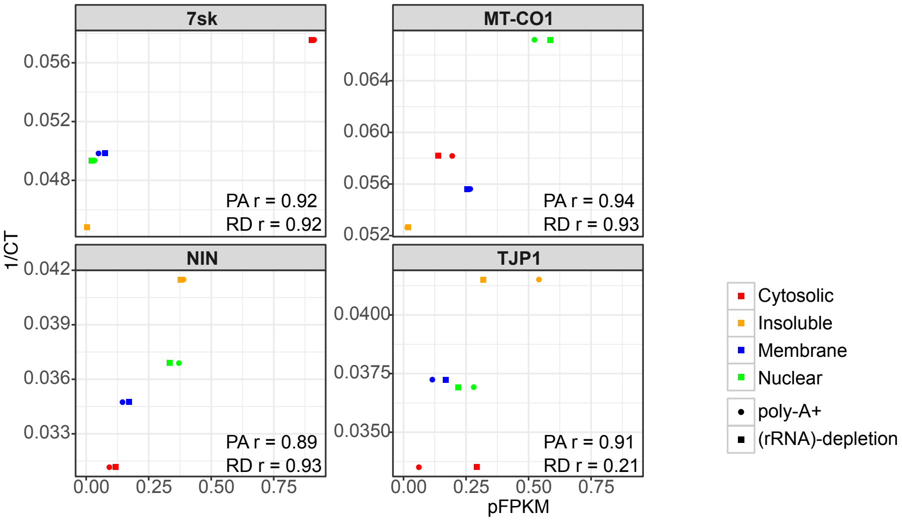


FIGURE 3.S4. pFPKM values are strongly correlated with the results of RT-qPCR. Scatter plot of RT-qPCR cycle threshold (CT) values and pFPKM from fraction-specific RNA markers demonstrate that CT values are strongly correlated with pFPKM values. Similar results were obtained with rRNA-depletion and poly(A) sequencing datasets. Pearson correlations are indicated on each graph.

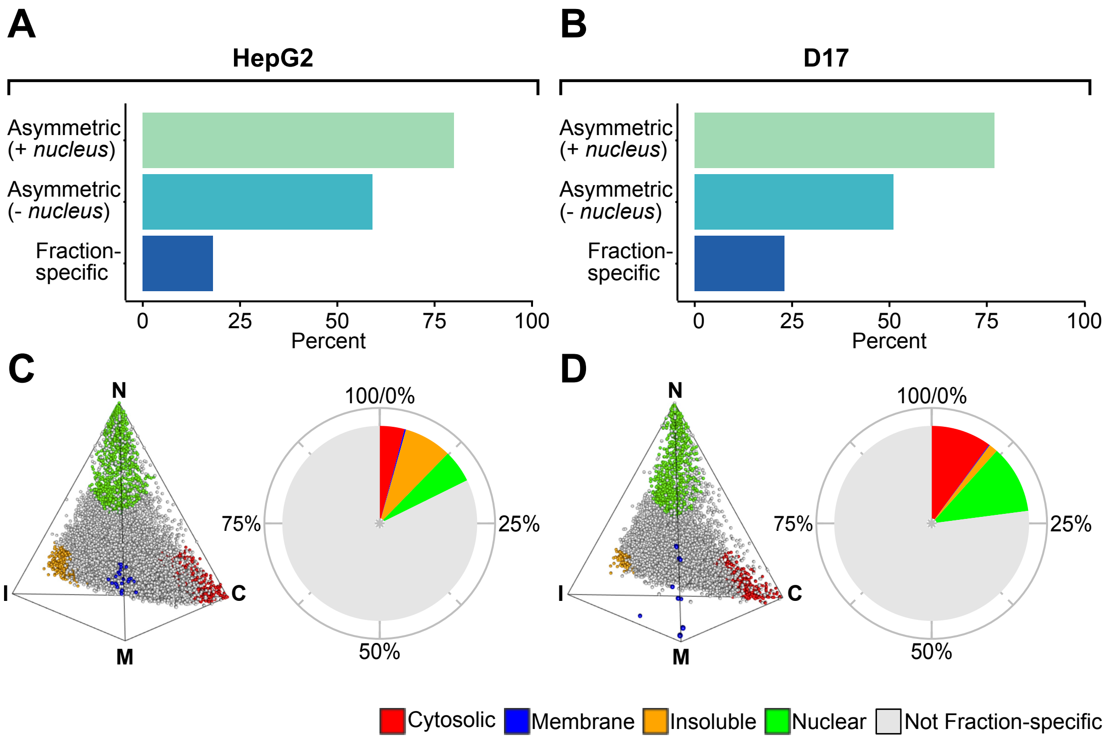


FIGURE 3.S5. Distribution profile of mRNAs in rRNA-depleted sequencing datasets. (A-B) Histogram showing the percent of asymmetrically distributed and fraction-specific mRNAs in HepG2 (A) and D17 (B) cells. **(C-D)** Simplex graphs (left panels) and pie charts (right panels) depicting the relative distribution and proportion of fraction-specific mRNAs, coloured according to the fraction they are enriched in, relative to the total mRNA population in HepG2 (C) or D17 (D) cells. C=Cytosolic, M=Membrane, I=Insoluble, N=Nuclear.

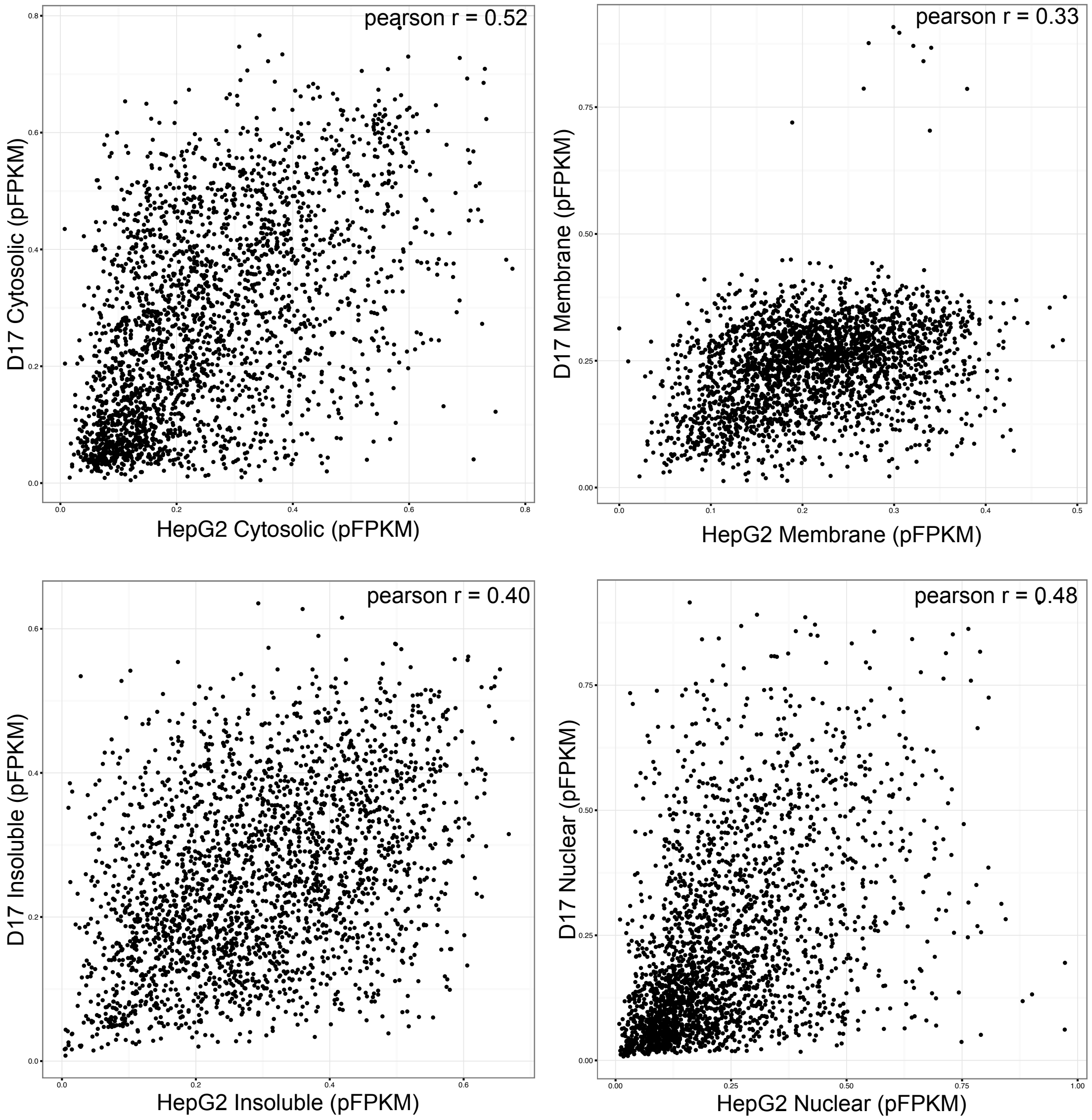


Figure 3.S6. Human and fly cell mRNA localization is strongly correlated between the same fraction. Scatter plot of mRNA relative localization in pFPKM between the same fraction in HepG2 and D17 cells. Pearson correlations are indicated on each graph. All p-values $< 2.2 \times 10^{-16}$.

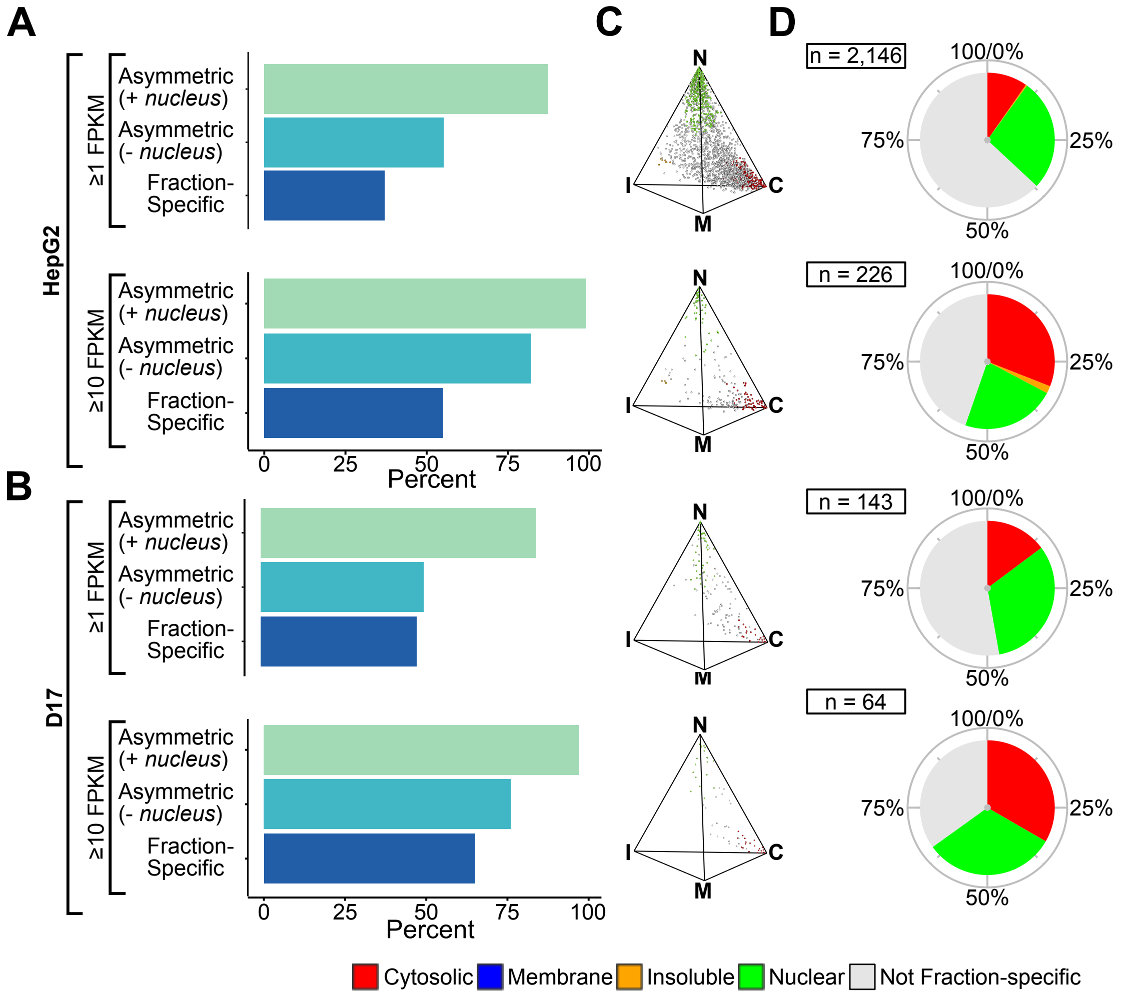


Figure 3.S7. LncRNAs from rRNA-depleted sequencing datasets are asymmetrically distributed and exhibit preferential polarization towards the nucleus and cytosol. (A-B) Histogram showing the percent of asymmetrically distributed and fraction-specific lncRNAs obtained fro mRD-sequencing, expressed in HepG2 (A) and D17 (B) cells, either using a standard expression threshold (≥ 1 FPKM) or focusing on highly expressed transcripts (≥ 10 FPKM). **(C-D)** Simplex graph (C) and pie chars (D) depicting the relative distribution and proportion of fraction-specific lncRNAs, coloured according to the fraction they are enriched in, relative to the total lncRNA population detected in HepG2 or D17 cells at the expression threshold indicated in (A-B).

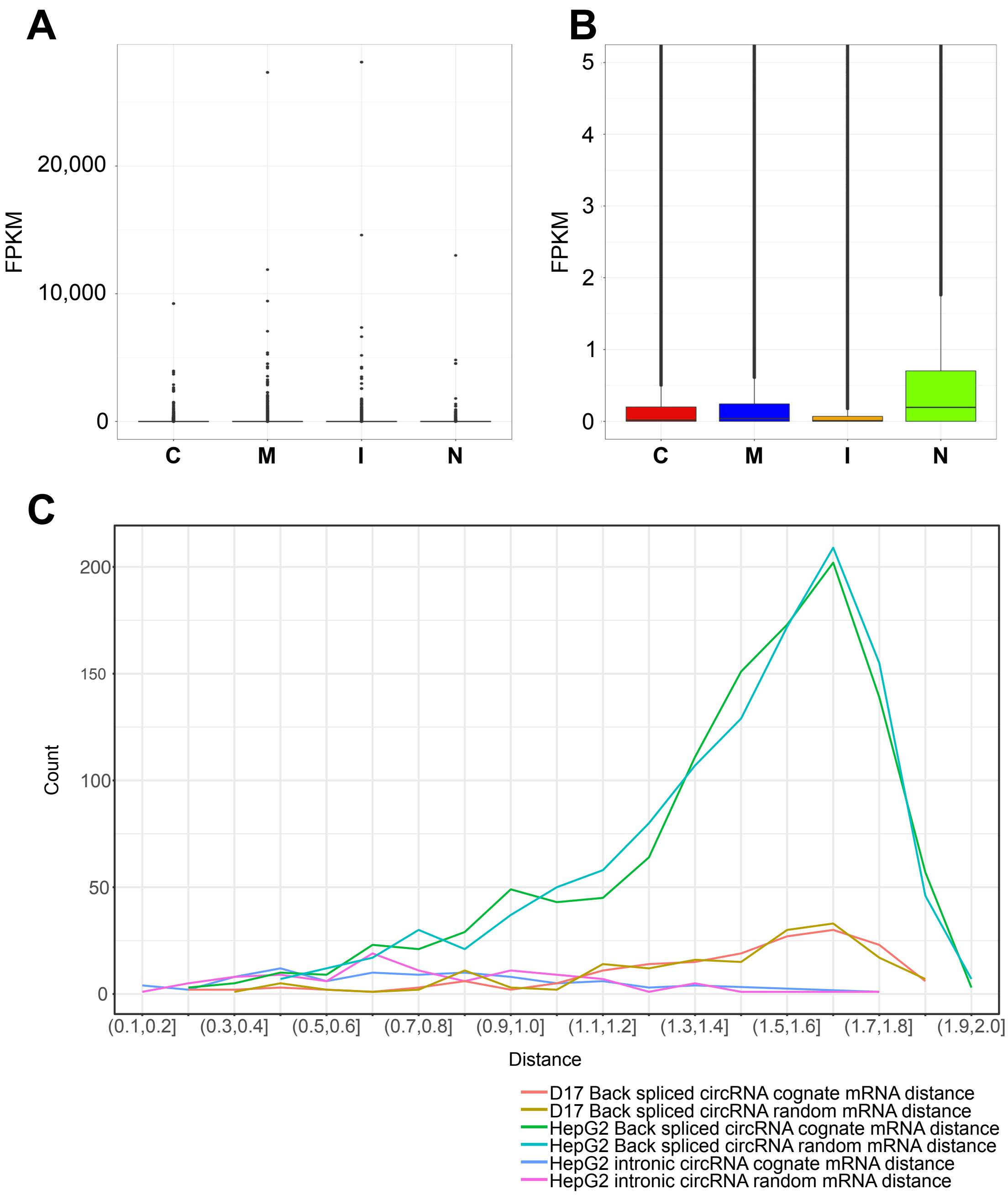


Figure 3.S8. CircRNA exhibit distinct distribution compared to other intronic regions or their host mRNAs. **(A)** Boxplot of the expression, in FPKM, of 1,040,283 introns as annotated by Ensembl in HepG2 cells. **(B)** Zoomed view of the boxplot described in (A). **(C)** The relative localization distance between a circRNA and a mRNA encompassing a putative circRNA for both back-spliced and intronic circRNA in both human and fly cells.

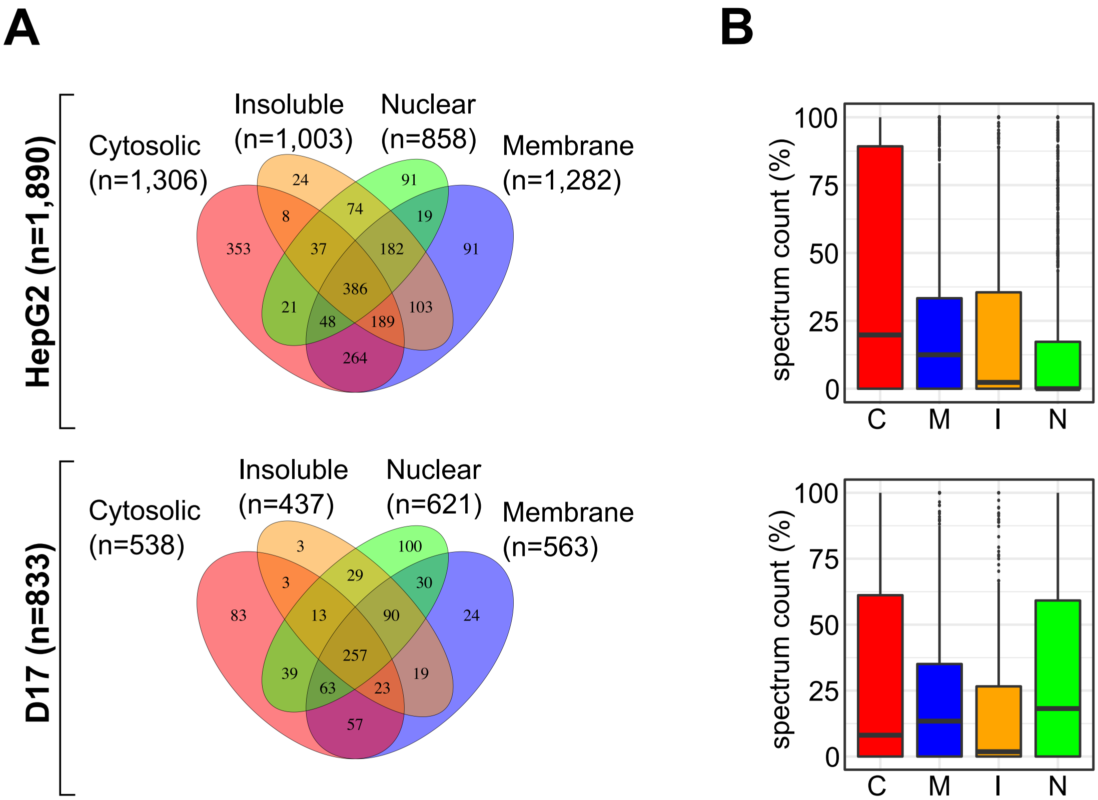


FIGURE 3.S9. Proteome distribution of each fraction. (A) Venn diagram depicting the relative distribution of proteins in HepG2 (upper panel) and D17 (lower panel) cell fractions following tandem mass spectrometry (LC-MS/MS), assessed by measuring their total spectrum count. **(B)** Boxplot showing the relative distribution, in percent spectrum count, of all proteins identified in our LC-MS/MS datasets for HepG2 (upper panel) and D17 (lower panel) cell fractions. C=Cytosolic, M=Membrane, I=Insoluble, N=Nuclear.

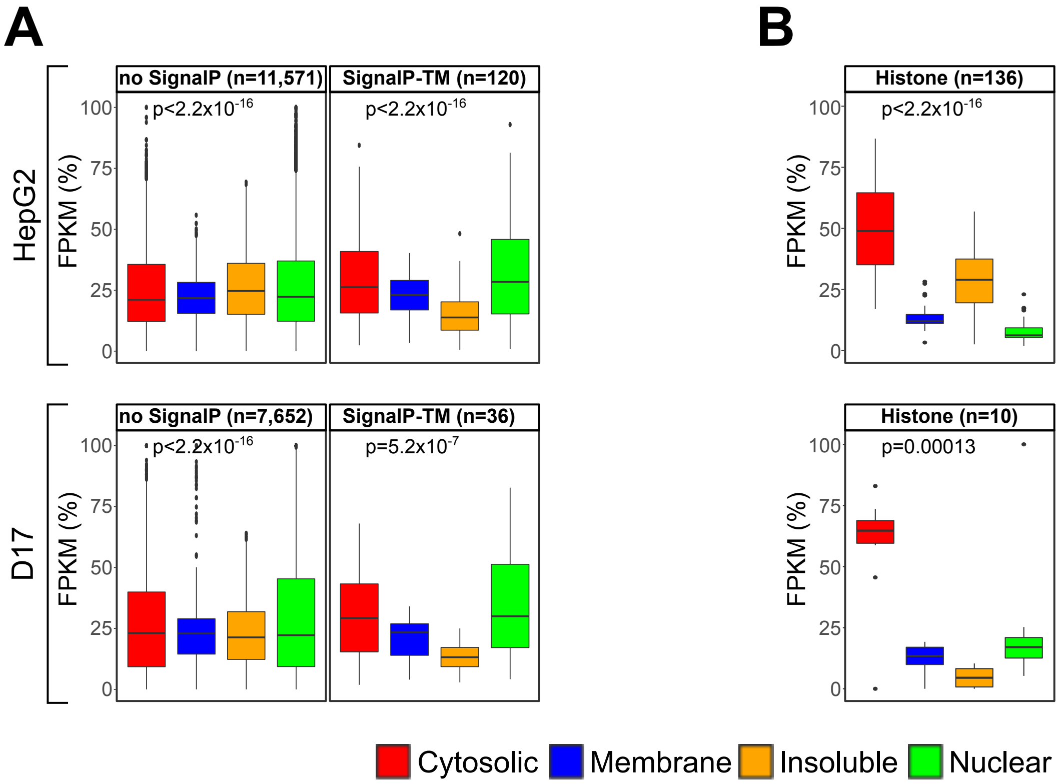


FIGURE 3.S10. mRNA bearing various motifs exhibits asymmetric distributions. **(A)** Boxplots showing the fraction distribution profiles of mRNAs bearing a signal peptide cleavage sites, in pFPKM, for HepG2 (upper plots) and D17 (lower plots) cells. **(B)** Boxplots showing the fractions distribution profiles of mRNA canonical histones, in pFPKM, for HepG2 (upper plots) and D17 (lower plots) cells. The number (n) of transcripts analyzed and Kruskal-Wallis p-value is indicated for each motifs.

Références

[1] K. Bahar Halpern, I. Caspi, D. Lemze, M. Levy, S. Landen, E. Elinav, I. Ulitsky, and S. Itzkovitz, *Nuclear retention of mRNA in mammalian tissues*, Cell Reports, 13 (2015), pp. 2653–62.

[2] N. N. Batada, L. A. Shepp, and D. O. Siegmund, *Stochastic model of protein-protein interaction : why signaling proteins need to be colocalized*, Proceedings of the National Academy of Sciences of the United States of America, 101 (2004), pp. 6445–9.

[3] P. J. Batista and H. Y. Chang, *Long noncoding RNAs : cellular address codes in development and disease*, Cell, 152 (2013), pp. 1298–307.

[4] J. Bergalet and E. Lecuyer, *The functions and regulatory principles of mRNA intracellular trafficking*, Systems Biology of RNA Binding Proteins, 825 (2014), pp. 57–96.

[5] D. M. Bhatt, A. Pandya-Jones, A. J. Tong, I. Barozzi, M. M. Lissner, G. Natoli, D. L. Black, and S. T. Smale, *Transcript dynamics of proinflammatory genes revealed by sequence analysis of subcellular RNA fractions*, Cell, 150 (2012), pp. 279–90.

[6] M. D. Blower, E. Feric, K. Weis, and R. Heald, *Genome-wide analysis demonstrates conserved localization of messenger RNAs to mitotic microtubules*, Journal of Cell Biology, 179 (2007), pp. 1365–73.

[7] M. Briese, L. Saal, S. Appenzeller, M. Moradi, A. Baluapuri, and M. Sendtner, *Whole transcriptome profiling reveals the RNA content of motor axons*, Nucleic Acids Research, 44 (2016), p. e33.

[8] M. N. Cabili, M. C. Dunagin, P. D. McClanahan, A. Biaesch, O. Padovan-Merhar, A. Regev, J. L. Rinn, and A. Raj, *Localization and abundance analysis of human lncRNAs at single-cell and single-molecule resolution*, Genome Biology, 16 (2015), p. 20.

[9] I. J. Cajigas, G. Tushev, T. J. Will, S. tom Dieck, N. Fuerst, and E. M. Schuman, *The local transcriptome in the synaptic neuropil revealed by deep sequencing and high-resolution imaging*, Neuron, 74 (2012), pp. 453–66.

[10] J. Carlevaro-Fita, A. Rahim, R. Guigo, L. A. Vardy, and R. Johnson, *Cytoplasmic long noncoding RNAs are frequently bound to and degraded at ribosomes in human cells*, RNA, 22 (2016), pp. 867–82.

[11] L. L. Chen, *The biogenesis and emerging roles of circular RNAs*, Nature Reviews Molecular Cell Biology, 17 (2016), pp. 205–11.

[12] ——, *Linking long noncoding RNA localization and function*, Trends in Biochemical Sciences, 41 (2016), pp. 761–72.

[13] Q. Chen, S. Jagannathan, D. W. Reid, T. Zheng, and C. V. Nicchitta, *Hierarchical regulation of mRNA partitioning between the cytoplasm and the endoplasmic reticulum of mammalian cells*, Molecular Biology of the Cell, 22 (2011), pp. 2646–58.

[14] L. Cherbas, A. Willingham, D. Zhang, L. Yang, Y. Zou, B. D. Eads, J. W. Carlson, J. M. Landolin, P. Kapranov, J. Dumais, A. Samsonova, J. H. Choi, J. Roberts, C. A. Davis, H. Tang, M. J. van Baren, S. Ghosh, A. Dobin, K. Bell, W. Lin, L. Langton, M. O. Duff, A. E. Tenney, C. Zaleski, M. R. Brent, R. A. Hoskins, T. C. Kaufman, J. Andrews, B. R. Graveley, N. Perrimon, S. E. Celniker, T. R. Gingeras, and P. Cherbas, *The transcriptional diversity of 25 Drosophila cell lines*, Genome Research, 21 (2011), pp. 301–14.

[15] N. A. Cody, C. Iampietro, and E. Lecuyer, *The many functions of mRNA localization during normal development and disease : From pillar to post*, Wiley Interdisciplinary Reviews-Developmental Biology, 2 (2013), pp. 781–96.

[16] J. D. Currie and S. L. Rogers, *Using the Drosophila melanogaster D17-c3 cell culture system to study cell motility*, Nature Protocols, 6 (2011), pp. 1632–41.

[17] M. de Jong, B. van Breukelen, F. R. Wittink, F. L. Menke, P. J. Weisbeek, and G. Van den Ackerveken, *Membrane-associated transcripts in Arabidopsis ; their isolation and characterization by DNA microarray analysis and bioinformatics*, The Plant Journal, 46 (2006), pp. 708–21.

[18] T. Derrien, R. Johnson, G. Bussotti, A. Tanzer, S. Djebali, H. Tilgner, G. Guernec, D. Martin, A. Merkel, D. G. Knowles, J. Lagarde, L. Veeravalli, X. Ruan, Y. Ruan, T. Lassmann, P. Carninci, J. B. Brown, L. Lipovich, J. M. Gonzalez, M. Thomas, C. A. Davis, R. Shiekhattar, T. R. Gingeras, T. J. Hubbard, C. Notredame, J. Harrow, and R. Guigo, *The GENCODE v7 catalog of human long noncoding RNAs : analysis of their gene structure, evolution, and expression*, Genome Research, 22 (2012), pp. 1775–89.

[19] M. Diehn, R. Bhattacharya, D. Botstein, and P. O. Brown, *Genome-scale identification of membrane-associated human mRNAs*, PLoS Genetics, 2 (2006), p. e11.

[20] M. Diehn, M. B. Eisen, D. Botstein, and P. O. Brown, *Large-scale identification of secreted and membrane-associated gene products using DNA microarrays*, Nature Genetics, 25 (2000), pp. 58–62.

[21] S. Durinck, P. T. Spellman, E. Birney, and W. Huber, *Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt*, Nature Protocols, 4 (2009), pp. 1184–91.

[22] J. Eberwine, K. Miyashiro, J. E. Kacharmina, and C. Job, *Local translation of classes of mRNAs that are targeted to neuronal dendrites*, Proceedings of the National Academy of Sciences of the United States of America, 98 (2001), pp. 7080–5.

[23] M. Garcia, X. Darzacq, T. Delaveau, L. Jourdren, R. H. Singer, and C. Jacq, *Mitochondria-associated yeast mRNAs and the biogenesis of molecular complexes*, Molecular Biology of the Cell, 18 (2007), pp. 362–8.

[24] S. Gerstberger, M. Hafner, and T. Tuschl, *A census of human RNA-binding proteins*, Nature Reviews Genetics, 15 (2014), pp. 829–45.

[25] P. Glazar, P. Papavasileiou, and N. Rajewsky, *circBase : a database for circular RNAs*, RNA, 20 (2014), pp. 1666–70.

[26] L. F. Gumy, G. S. Yeo, Y. C. Tung, K. H. Zivraj, D. Willis, G. Coppola, B. Y. Lam, J. L. Twiss, C. E. Holt, and J. W. Fawcett, *Transcriptome analysis of embryonic and adult sensory axons reveals changes in mRNA repertoire localization*, RNA, 17 (2011), pp. 85–98.

[27] L. H. Hartwell, J. J. Hopfield, S. Leibler, and A. W. Murray, *From molecular to modular cell biology*, Nature, 402 (1999), pp. C47–52.

[28] K. E. Howell, E. Devaney, and J. Gruenberg, *Subcellular fractionation of tissue culture cells*, Trends in Biochemical Sciences, 14 (1989), pp. 44–7.

[29] C. Iampietro, J. Bergalet, X. Wang, N. A. Cody, A. Chin, F. A. Lefebvre, M. Douziech, H. M. Krause, and E. Lecuyer, *Developmentally regulated elimination of damaged nuclei involves a Chk2-dependent mechanism of mRNA nuclear retention*, Developmental Cell, 29 (2014), pp. 468–81.

[30] N. T. Ingolia, L. F. Lareau, and J. S. Weissman, *Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes*, Cell, 147 (2011), pp. 789–802.

[31] S. Jagannathan, C. Nwosu, and C. V. Nicchitta, *Analyzing mRNA localization to the endoplasmic reticulum via cell fractionation*, Methods in Molecular Biology, 714 (2011), pp. 301–21.

[32] H. Jambor, V. Surendranath, A. T. Kalinka, P. Mejstrik, S. Saalfeld, and P. Tomancak, *Systematic imaging reveals features and changing localization of mRNAs in Drosophila development*, eLife, 4 (2015).

[33] C. H. Jan, C. C. Williams, and J. S. Weissman, *Principles of ER cotranslational translocation revealed by proximity-specific ribosome profiling*, Science, 346 (2014), p. 1257521.

[34] W. R. Jeck and N. E. Sharpless, *Detecting and characterizing circular RNAs*, Nature Biotechnology, 32 (2014), pp. 453–61.

[35] W. R. Jeck, J. A. Sorrentino, K. Wang, M. K. Slevin, C. E. Burd, J. Liu, W. F. Marzluff, and N. E. Sharpless, *Circular RNAs are abundant, conserved, and associated with ALU repeats*, RNA, 19 (2013), pp. 141–57.

[36] C. Job and J. Eberwine, *Localization and translation of mRNA in dendrites and axons*, Nature Reviews Neuroscience, 2 (2001), pp. 889–98.

[37] M. Khaladkar, P. T. Buckley, M. T. Lee, C. Francis, M. M. Eghbal, T. Chuong, S. Suresh, B. Kuhn, J. Eberwine, and J. Kim, *Subcellular RNA sequencing reveals broad presence of cytoplasmic intron-sequence retaining transcripts in mouse and rat neurons*, PLoS One, 8 (2013), p. e76194.

[38] C. C. Kopczynski, J. N. Noordermeer, T. L. Serano, W. Y. Chen, J. D. Pendleton, S. Lewis, C. S. Goodman, and G. M. Rubin, *A high throughput screen to identify secreted and transmembrane proteins involved in Drosophila embryogenesis*, Proceedings of the National Academy of Sciences of the United States of America, 95 (1998), pp. 9973–8.

[39] D. Kressler, E. Hurt, and J. Bassler, *A puzzle of life : Crafting ribosomal subunits*, Trends in Biochemical Sciences, (2017).

[40] J. Kuriyan and D. Eisenberg, *The origin of protein interactions and allostery in colocalization*, Nature, 450 (2007), pp. 983–90.

[41] E. Lecuyer, H. Yoshida, N. Parthasarathy, C. Alm, T. Babak, T. Cerovina, T. R. Hughes, P. Tomancak, and H. M. Krause, *Global analysis of mRNA localization reveals a prominent role in organizing cellular architecture and function*, Cell, 131 (2007), pp. 174–87.

[42] F. A. Lefebvre, L. P. Benoit Bouvrette, L. Perras, A. Blanchet-Cohen, D. Garnier, J. Rak, and E. Lecuyer, *Comparative transcriptomic analysis of human and Drosophila extracellular vesicles*, Scientific Reports, 6 (2016), p. 27680.

[43] F. A. Lefebvre, N. Cody, L. P. Benoit Bouvrette, J. Bergalet, X. Wang, and E. Lécuyer, *CeFra-seq : Systematic mapping of RNA subcellular distribution properties through cell fractionation coupled to deep-sequencing.*, Methods, 126 (2017), pp. 138–48.

[44] I. Legnini, G. Di Timoteo, F. Rossi, M. Morlando, F. Briganti, O. Sthandier, A. Fatica, T. Santini, A. Andronache, M. Wade, P. Laneve, N. Rajewsky, and I. Bozzoni, *Circ-ZNF609 is a circular rna that can be translated and functions in myogenesis*, Molecular Cell, 66 (2017), pp. 22–37 e9.

[45] R. S. Lerner, R. M. Seiser, T. Zheng, P. J. Lager, M. C. Reedy, J. D. Keene, and C. V. Nicchitta, *Partitioning and translation of mRNAs encoding soluble proteins on membrane-bound ribosomes*, RNA, 9 (2003), pp. 1123–37.

[46] E. D. Levy, J. Kowarzyk, and S. W. Michnick, *High-resolution mapping of protein concentration reveals principles of proteome architecture and adaptation*, Cell Reports, 7 (2014), pp. 1333–40.

[47] Y. Liu, A. Beyer, and R. Aebersold, *On the dependency of cellular protein levels on mRNA abundance*, Cell, 165 (2016), pp. 535–50.

[48] P. Marc, A. Margeot, F. Devaux, C. Blugeon, M. Corral-Debrinski, and C. Jacq, *Genome-wide analysis of mRNAs targeted to yeast mitochondria*, EMBO Reports, 3 (2002), pp. 159–64.

[49] F. K. Mardakheh, A. Paul, S. Kumper, A. Sadok, H. Paterson, A. McCarthy, Y. Yuan, and C. J. Marshall, *Global analysis of mRNA, translation, and protein localization : Local translation is a key regulator of cell protrusions*, Developmental Cell, 35 (2015), pp. 344–57.

[50] K. C. Martin and A. Ephrussi, *mRNA localization : gene expression in the spatial dimension*, Cell, 136 (2009), pp. 719–30.

[51] T. R. Mercer, D. J. Gerhardt, M. E. Dinger, J. Crawford, C. Trapnell, J. A. Jeddeloh, J. S. Mattick, and J. L. Rinn, *Targeted RNA sequencing reveals the deep complexity of the human transcriptome*, Nature Biotechnology, 30 (2011), pp. 99–104.

[52] H. Mi, X. Huang, A. Muruganujan, H. Tang, C. Mills, D. Kang, and P. D. Thomas, *PANTHER version 11 : expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements*, Nucleic Acids Research, 45 (2017), pp. D183–9.

[53] S. Mili, K. Moissoglu, and I. G. Macara, *Genome-wide screen reveals APC-associated RNAs enriched in cell protrusions*, Nature, 453 (2008), pp. 115–9.

[54] L. A. Mingle, N. N. Okuhama, J. Shi, R. H. Singer, J. Condeelis, and G. Liu, *Localization of all seven messenger RNAs for the actin-polymerization nucleator Arp2/3 complex in the protrusions of fibroblasts*, Journal of Cell Science, 118 (2005), pp. 2425–33.

[55] R. Moccia, D. Chen, V. Lyles, E. Kapuya, Y. E, S. Kalachikov, C. M. Spahn, J. Frank, E. R. Kandel, M. Barad, and K. C. Martin, *An unbiased cDNA library prepared from isolated Aplysia sensory neuron processes is enriched for cytoskeletal and translational mRNAs*, The Journal of Neuroscience, 23 (2003), pp. 9409–17.

[56] T. Mondal, M. Rasmussen, G. K. Pandey, A. Isaksson, and C. Kanduri, *Characterization of the RNA content of chromatin*, Genome Research, 20 (2010), pp. 899–907.

[57] N. R. Pamudurti, O. Bartok, M. Jens, R. Ashwal-Fluss, C. Stottmeister, L. Ruhe, M. Hanan, E. Wyler, D. Perez-Hernandez, E. Ramberger, S. Shenzis, M. Samson, G. Dittmar, M. Landthaler, M. Chekulaeva, N. Rajewsky, and S. Kadener, *Translation of CircRNAs*, Molecular Cell, 66 (2017), pp. 9–21 e7.

[58] M. M. Poon, S. H. Choi, C. A. Jamieson, D. H. Geschwind, and K. C. Martin, *Identification of process-localized mRNAs from cultured rodent hippocampal neurons*, The Journal of Neuroscience, 26 (2006), pp. 13390–9.

[59] K. V. Prasanth, S. G. Prasanth, Z. Xuan, S. Hearn, S. M. Freier, C. F. Bennett, M. Q. Zhang, and D. L. Spector, *Regulating gene expression through RNA nuclear retention*, Cell, 123 (2005), pp. 249–63.

[60] B. Pyhtila, T. Zheng, P. J. Lager, J. D. Keene, M. C. Reedy, and C. V. Nicchitta, *Signal sequence- and translation-independent mRNA localization to the endoplasmic reticulum*, RNA, 14 (2008), pp. 445–53.

[61] A. R. Quinlan and I. M. Hall, *BEDTools : a flexible suite of utilities for comparing genomic features*, Bioinformatics, 26 (2010), pp. 841–2.

[62] A. Ruepp, B. Waegele, M. Lechner, B. Brauner, I. Dunger-Kaltenbach, G. Fobo, G. Frishman, C. Montrone, and H. W. Mewes, *CORUM : the comprehensive resource of mammalian protein complexes–2009*, Nucleic Acids Research, 38 (2010), pp. D497–501.

[63] Y. Sasaki, C. Gross, L. Xing, Y. Goshima, and G. J. Bassell, *Identification of axon-enriched microRNAs localized to growth cones of cortical neurons*, Developmental Neurobiology, 74 (2014), pp. 397–406.

[64] B. Schwanhausser, D. Busse, N. Li, G. Dittmar, J. Schuchhardt, J. Wolf, W. Chen, and M. Selbach, *Global quantification of mammalian gene expression control*, Nature, 473 (2011), pp. 337–42.

[65] J. A. Sharp, J. J. Plant, T. K. Ohsumi, M. Borowsky, and M. D. Blower, *Functional analysis of the microtubule-interacting transcriptome*, Molecular Biology of the Cell, 22 (2011), pp. 4312–23.

[66] M. Sultan, V. Amstislavskiy, T. Risch, M. Schuette, S. Dokel, M. Ralser, D. Balzereit, H. Lehrach, and M. L. Yaspo, *Influence of RNA extraction methods and library selection schemes on RNA-seq data*, BMC Genomics, 15 (2014), p. 675.

[67] J. M. Taliaferro, M. Vidaki, R. Oliveira, S. Olson, L. Zhan, T. Saxena, E. T. Wang, B. R. Graveley, F. B. Gertler, M. S. Swanson, and C. B. Burge, *Distal alternative last exons localize mRNAs to neural projections*, Molecular Cell, 61 (2016), pp. 821–33.

[68] H. Tilgner, D. G. Knowles, R. Johnson, C. A. Davis, S. Chakrabortty, S. Djebali, J. Curado, M. Snyder, T. R. Gingeras, and R. Guigo, *Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs*, Genome Research, 22 (2012), pp. 1616–25.

[69] S. van Heesch, M. van Iterson, J. Jacobi, S. Boymans, P. B. Essers, E. de Bruijn, W. Hao, A. W. MacInnes, E. Cuppen, and M. Simonis, *Extensive localization of long noncoding RNAs to the cytosol and mono- and polyribosomal complexes*, Genome Biology, 15 (2014), p. R6.

[70] C. Vogel and E. M. Marcotte, *Insights into the regulation of protein abundance from proteomic and transcriptomic analyses*, Nature Reviews Genetics, 13 (2012), pp. 227–32.

[71] G. P. Wagner, K. Kin, and V. J. Lynch, *Measurement of mRNA abundance using RNA-seq data : RPKM measure is inconsistent among samples*, Theory in Biosciences, 131 (2012), pp. 281–5.

[72] E. T. Wang, N. A. Cody, S. Jog, M. Biancolella, T. T. Wang, D. J. Treacy, S. Luo, G. P. Schroth, D. E. Housman, S. Reddy, E. Lecuyer, and C. B. Burge, *Transcriptome-wide regulation of pre-mRNA splicing and mRNA localization by muscleblind proteins*, Cell, 150 (2012), pp. 710–24.

[73] R. Wilk, J. Hu, D. Blotsky, and H. M. Krause, *Diverse and pervasive subcellular distributions for both coding and long noncoding RNAs*, Genes and Development, 30 (2016), pp. 594–609.

[74] C. C. Williams, C. H. Jan, and J. S. Weissman, *Targeting and plasticity of mitochondrial proteins revealed by proximity-specific ribosome profiling*, Science, 346 (2014), pp. 748–51.

[75] X. O. Zhang, R. Dong, Y. Zhang, J. L. Zhang, Z. Luo, J. Zhang, L. L. Chen, and L. Yang, *Diverse alternative back-splicing and alternative splicing landscape of circular RNAs*, Genome Research, 26 (2016), pp. 1277–87.

[76] Y. Zhang, X. O. Zhang, T. Chen, J. F. Xiang, Q. F. Yin, Y. H. Xing, S. Zhu, L. Yang, and L. L. Chen, *Circular intronic long noncoding RNAs*, Molecular Cell, 51 (2013), pp. 792–806.

[77] K. H. Zivraj, Y. C. Tung, M. Piper, L. Gumy, J. W. Fawcett, G. S. Yeo, and C. E. Holt, *Subcellular profiling reveals distinct and developmentally regulated repertoire of growth cone mRNAs*, The Journal of Neuroscience, 30 (2010), pp. 15464–78.

Troisième article.

oRNAment : A database of putative RNA binding protein target sites in the transcriptomes of model species

Louis Philip Benoit Bouvrette

Cet article à été publié dans le revue *Nucleic Acid Research* (2020. 48 : D166-D173, https://doi.org/10.1093/nar/gkz986)

Préface et contributions

Ce chapitre est présenté sous la forme d'un article de recherche et a aussi été publié dans le journal *Nucleic Acid Research*.

L'article présente oRNAment (**o RNA m**otifs **e**nrichment in **t**ranscriptomes), une base de données qui répertorie pour la première fois les sites de liaison présumés de 223 protéines liant l'ARN (RBP), englobant 453 motifs, et ce, à l'échelle transcriptomique (excluant les introns) pour 5 espèces. Ces motifs ont été obtenus à partir de technologies de sélection *in vitro*, telles que RNAcompete [**18**] et RNA Bind-n-Seq (RBNS) [**3**]. Grâce à l'utilisation d'un algorithme établi [**7**], mais modernisé, notamment avec des ajustements à la méthodologie définissant une instance de motif et grâce à l'accès à une puissance de calcul supplémentaire fournie par Calcul Canada, j'ai pu identifier des instances de motifs à une échelle largement supérieure à celle qui a été obtenue par le biais d'autres ressources. De même, ma méthode prédit avec précision les sites de liaison RBP observés par eCLIP dans les cellules humaines. La base de données couvre actuellement 525 718 ARN codants et non codants à travers les transcriptomes humains et de quatre organismes modèles importants : *C. elegans*, *D. rerio*, *D. melanogaster* et *M. musculus*.

La ressource oRNAment a été développée pour afficher les résultats issus de la cartographie des motifs RBP, en utilisant un système de gestion de base de données (SGBD) de pointe et un cadre web moderne, permettant la récupération et le traitement efficace de grandes quantités de données. L'interface conviviale de oRNAment prend en entrée la sélection d'une espèce, d'un gène ou des noms/ID de transcrits, d'un nom d'une RBP et d'un seuil de score de similitude défini par l'utilisateur, ou d'un attribut spécifique [par exemple, la région non traduite en 3' (3' UTR) d'un ARNm]. Pour chaque type de recherche, plusieurs graphiques interactifs sont générés permettant la visualisation personnalisée de données résumant les résultats de la requête de l'utilisateur. Un tableau détaillé et des graphiques distincts définissant la position de toutes les instances de motif et leurs similitudes avec les motifs consensus d'une RBP dans chaque transcrit sont également produits. L'utilisateur peut parallèlement rechercher de manière interactive des instances de motif d'une RBP via un navigateur de génome intégré. Toutes les informations peuvent être facilement téléchargées sous forme de fichiers Excel, csv ou bed et la base de données contient une section proposant un tutoriel détaillée aidant l'utilisateur à naviguer à travers la ressource.

J'ai effectué les analyses bio-informatiques, développé l'algorithme de recherche de motifs, construit la base de données, créé l'interface web et instauré et configuré le serveur web (développement *full stack*). J'ai aussi créé l'ensemble des figures et écrit l'intégralité du manuscrit. Samantha Bovaird a effectué la validation systématique de la base de données et de l'interface web. Eric Lécuyer et Mathieu Blanchette ont supervisé les travaux et l'écriture du manuscrit et ont suggéré des modifications concernant, principalement et respectivement, les aspects biologique et informatique. Tous les co-auteurs ont approuvé le manuscrit avant sa soumission à l'éditeur de *Nucleic Acid Research*, qui l'a fait évaluer par des paires avant de le publier.

oRNAment : A database of putative RNA binding protein target sites in the transcriptomes of model species

par

Louis Philip Benoit Bouvrette[1, 2], Samantha Bovaird[3], Mathieu Blanchette[4] et Eric Lécuyer[1, 2, 3]

(1) Institut de Recherches Cliniques de Montréal (IRCM) Montréal, Québec, Canada.
(2) Département de biochimie, Université de Montréal, Montréal, Québec, Canada
(3) Division of experimental medicine, McGill University, Montréal, Québec, Canada
(4) School of computer science, McGill University, Montréal, Québec, Canada

Cet article à été publié dans le revue *Nucleic Acid Research* (2020. 48 : D166-D173, https://doi.org/10.1093/nar/gkz986).

ABSTRACT. Protein-RNA interactions are essential for controlling most aspects of RNA metabolism, including synthesis, processing, trafficking, stability and degradation. *In vitro* selection methods, such as RNAcompete and RNA Bind-n-Seq, have defined the consensus target motifs of hundreds of RNA-binding proteins (RBPs). However, readily available information about the distribution features of these motifs across full transcriptomes was hitherto lacking. Here, we introduce oRNAment (**o RNA m**otifs **en**richment in **t**ranscriptomes), a database that catalogues the putative motif instances of 223 RBPs, encompassing 453 motifs, in a transcriptome-wide fashion. The database covers 525,718 complete coding and noncoding RNA species across the transcriptomes of human and four prominent model organisms : *Caenorhabditis elegans*, *Danio rerio*, *Drosophila melanogaster*, and *Mus musculus*. The unique features of oRNAment include : (i) hosting of the most comprehensive mapping of RBP motif instances to date, with 421,133,612 putative binding sites described across five species ; (ii) options for the user to filter the data according to a specific threshold ; (iii) a user-friendly interface and efficient back-end allowing the rapid querying of the data through multiple angles (i.e. transcript, RBP, or sequence attributes) and (iv) generation of several interactive data visualization charts describing the results of user queries. oRNAment is freely available at http://rnabiology.ircm.qc.ca/oRNAment/.

Keywords: RNA-binding proteins, motifs, messenger RNA, noncoding RNA, transcriptome-wide, model species

1. Introduction

Throughout their life-cycle, RNA molecules undergo a variety of co- and post-transcriptional regulatory events that control their maturation, function and fate [4, 6, 10]. By modulating the assembly and function of ribonucleoprotein machineries, protein-RNA interactions play critical roles in virtually all facets of RNA metabolism. Indeed, RNA-binding proteins (RBPs) form an essential class of regulatory factors, which encompass among the most deeply evolutionarily conserved protein families [4, 13]. These proteins are primarily classified by the type of RNA-binding domain (RBD) they contain, which confers to them the capacity to interact with RNA molecules through binding sites defined by their sequence and/or structural properties [4, 13]. Recent studies, combining RNA-capture and mass spectrometry profiling, have characterized ∼1500 RBPs in human cells, hinting at the staggering complexity of post-transcriptional regulation [1, 2].

To characterize the binding specificities of candidate RBPs, binding site selection approaches, in particular RNAcompete and RNA Bind-n-Seq (RBNS) methodologies, have been systematically applied to a growing proportion of eukaryotic RBPs [3, 8, 17, 18]. Both of these methods involve *in vitro* binding assays combining a recombinantly purified RBP (or its RBD) and a randomized pool of RNA, followed by the biochemical purification of bound RNA molecules and their identification via microarray or RNA-sequencing [3, 8, 17, 18]. These approaches have enabled the identification of primary sequence consensus binding site motifs for a few hundred RBPs.

Several tools exist to scan user-provided RNA sequences for matches to these *in vitro* motifs, including servers such as CISBP-RNA, RBPmap, ATtRACT, and MotifMap-RNA [18, 5, 15, 11]. However, to date, no resources have been developed for identifying and cataloguing putative RBP motif instances in complete transcriptomes. Herein, we describe the oRNAment (**o RNA m**otifs **en**richment in **t**ranscriptomes) database, which catalogues the motif instances of 223 RBPs previously defined via the RNAcompete and RBNS platforms, across the coding and noncoding transcriptomes (excluding introns) of humans and four major model organisms. oRNAment is accessible at http://rnabiology.ircm.qc.ca/oRNAment/.

2. oRNAment analysis pipeline

2.1. Pre-processing of the oRNAment input data

oRNAment was created to characterize the distribution properties of potential RBP target sites across model organism transcriptomes from the most up-to-date RBP motif data available (Figure 4.1).

We acquired the data for 223 unique RBPs, totalling 453 consensus motifs in the form of position weight matrices (PWMs) obtained by either RNAcompete or RBNS (Figure 4.1i) [3, 8, 17, 18, 14, 9]. More precisely, we obtained 218 RNAcompete PWMs (172 RBPs), from the CISBP-RNA resource [18]. In parallel, we derived an additional 235 PWMs (78 RBPs) by executing the RBNS computational analysis pipeline for 7-mer enrichment on the RBNS data available from the ENCODE resource [3, 8, 14, 9]. Therefore, all motifs in the database are 7 nucleotides in length and are, as such, comparable. Overall, only 27 RBPs were profiled by both methods (Figure 4.1i, light grey lines). RBPs and their motifs were flagged for their species-specificities as defined by Ray *et al.* (Figure 4.1ii) [18]. Scans were performed for each PWM individually, regardless of similarities or discrepancies between

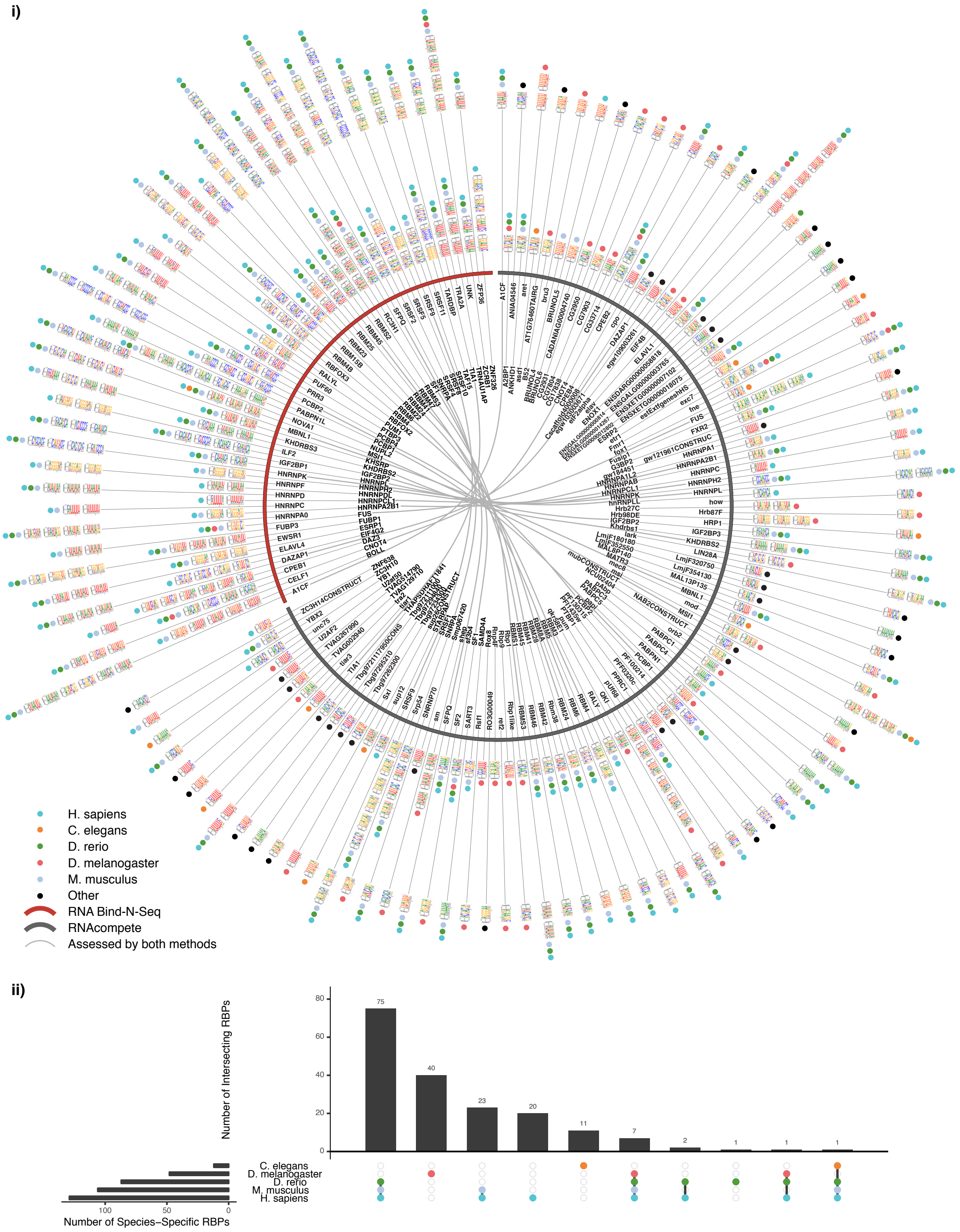


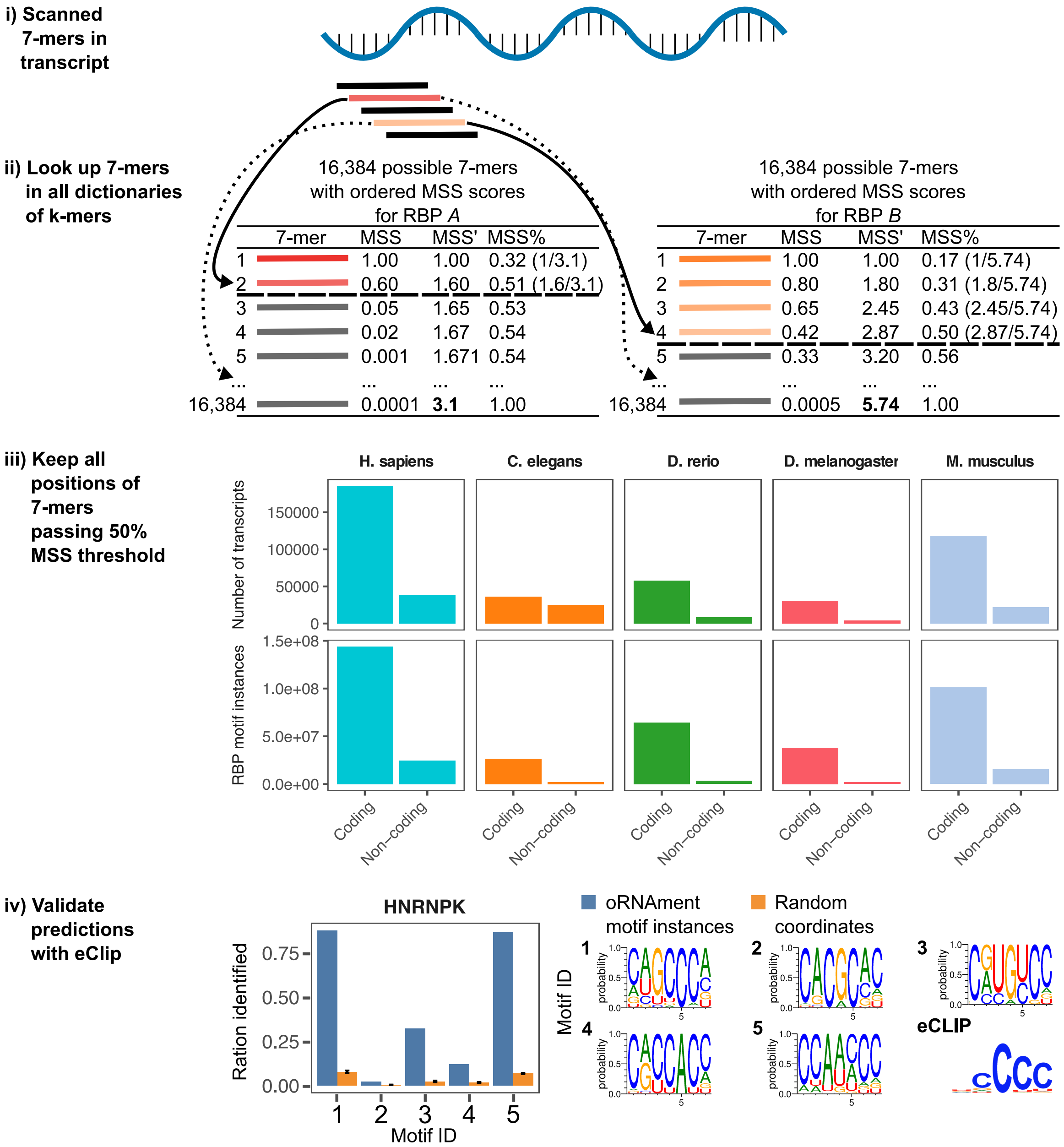
FIGURE 4.1. The oRNAment database contains 453 motifs attributable to 223 RBPs in 5 species. (i) Motifs obtained for each RBPs come from RNAcompete (red segment) and RBNS (dark grey segment) experiments. Coloured dots show the species-specificity of each motif according to Ray *et al.* [18]. There are 181 RBPs with binding specificities in the species included in the database and 42 from external species. Links shown between RBP (light grey lines) denote those that were assessed by both methods. **(ii)** Upset plot showing the distribution of interrogated species-specific RBPs across all five species.

RNAcompete and RBNS PWMs of the same RBP. Furthermore, motif scans were executed for motifs assigned to each RBP across all five species, regardless of the species representation of a given RBP. However, since RBP orthologs are expected to exhibit similar binding motif specificities if their RBD show >70% identity in amino acid sequence [18], we have flagged the species specificity of each factor so the user can take this information into consideration.

oRNAment is based upon a custom pipeline to perform efficient transcriptome-wide scans for instances of all 453 RBP motifs collected above (Figure 4.2). We based our pipeline on the previously published and widely used MATCH algorithm developed to scan for putative transcription factor binding sites across DNA sequences [7, 16]. This tool takes as input a motif, in the form of a PWM, and returns the position of the subsequences above a given score (Figure 4.2i-iii) [7, 16]. This is conceptually similar to scanning for RBP target motif instances, also taking a PWM as input, across RNA transcripts. Through the use of high-performance python 3.7 libraries (i.e., NumPy, Pandas) and data structures (pre-constructed hash tables of all heptanucleotides and score pairs), we developed a scanning algorithm that allows great efficiency, in terms of memory and speed, permitting timely execution across entire transcriptomes.

The search algorithm is based on the matrix similarity score (MSS), which measures the correspondence of a transcript region to a given RBP motif of the same length. This is defined as MSS = (current_score - minimum_score) / (maximum_score - minimum_score), where current_score is the product of each nucleotide probability at its respective position in the PWM, and the maximum_score and minimum_score are the product of each maximum or minimum probability value, respectively, in the PWM at each position. This provides a value between 0 and 1, where 1 is a perfect match to the top canonical binding motif of a given RBP (Figure 4.2).

In order to identify putative RBP motif instances, it is necessary to select an appropriate threshold for the MSS, which can vary depending on the user's objectives. This threshold allows the user to include motifs with varying degrees of similarity to the most probable *in vitro* defined consensus motif. For a given percentile P (e.g., $P = 50\%$) and a given PWM, the threshold T_P is chosen so that the probability that a 7-mer randomly generated based on the probabilities specified by the PWM obtains an MSS greater or equal to T_P is P. In other words, the fraction of sensitivity (or recall) of the search is P. In practice, T_P is obtained by


i) Scanned 7-mers in transcript
ii) Look up 7-mers in all dictionaries of k-mers
16,384 possible 7-mers with ordered MSS scores for RBP *A*
16,384 possible 7-mers with ordered MSS scores for RBP *B*
iii) Keep all positions of 7-mers passing 50% MSS threshold
H. sapiens
C. elegans
D. rerio
D. melanogaster
M. musculus
Number of transcripts
150000
100000
50000
0
RBP motif instances
1.5e+08
1.0e+08
5.0e+07
0.0e+00
Coding
Non-coding
iv) Validate predictions with eClip
HNRNPK
Ration identified
0.75
0.50
0.25
0.00
1
2
3
4
5
Motif ID
oRNAment motif instances
Random coordinates
Motif ID
1
2
3
4
5
probability
eCLIP

16,384 possible 7-mers with ordered MSS scores for RBP *A*

	7-mer	MSS	MSS'	MSS%
1		1.00	1.00	0.32 (1/3.1)
2		0.60	1.60	0.51 (1.6/3.1)
3		0.05	1.65	0.53
4		0.02	1.67	0.54
5		0.001	1.671	0.54
...		...	...	...
16,384		0.0001	**3.1**	1.00

16,384 possible 7-mers with ordered MSS scores for RBP *B*

	7-mer	MSS	MSS'	MSS%
1		1.00	1.00	0.17 (1/5.74)
2		0.80	1.80	0.31 (1.8/5.74)
3		0.65	2.45	0.43 (2.45/5.74)
4		0.42	2.87	0.50 (2.87/5.74)
5		0.33	3.20	0.56
...		...	...	...
16,384		0.0005	**5.74**	1.00

FIGURE 4.2. The oRNAment computational pipeline. **(i)** For a given transcript, the algorithm linearly scans for subsequences of length 7 and **(ii)** reports only those that have an MSS higher than the threshold, represented by the dashed line (table look-up, exemplified by the arrows, only shown for the second and fourth sequentially scanned 7-mers ; sum of MSS' used as denominator for MSS'% computation in bold). **(iii)** oRNAment reports all motif instances in all transcripts across five species. **(iv)** oRNAment reasonably predicts RBP binding sites observed by eCLIP in human K562 and HepG2 cells (blue bars in histogram), as shown for the five motifs bound by the HNRNPK RBP, in comparison to the same number of random sequences (orange bars).

calculating the MSS score of each of the 16,384 possible heptanucleotides, sorting them in decreasing order of MSS, and going down the sorted list until the sum of MSS of the selected heptanucleotides reaches $P\%$ of the total (Figure 4.2ii dash line). oRNAment contains motif matches for 10 different thresholds, for P ranging from 50% to 95% in increment of 5%, as well as for the special threshold MSS=1 (canonical motif) (Figure 4.2).

We observed that the analysis pipeline reasonably predicts RBP binding sites observed by eCLIP in human cells [14, 21]. We used as a validation set the group of 24 RBPs where eCLIP data was also available and compared the genomic coordinates of oRNAment motif instances, at a 50% threshold, to eCLIP peaks, at a ≥ 3 fold change and a p-value ≤ 0.001. For this, we first downloaded the bed narrowPeak files from ENCODE for both HepG2 and K562 cell lines and filtered them in order to only keep peaks in an annotated exon. This allowed a one-to-one comparison with the dataset scanned by oRNAment. We then collapsed peak regions from replicates when they showed any overlap. As the peak region rarely had the exact same coordinates, we kept as one region the coordinates englobing the shortest region between the two replicates (i.e. if replicate 1 had a peak between nucleotides 100–109 and replicate 2 a peak between 102–110, we kept as a peak a region between 102–109). We only kept peak regions of at least seven nucleotides. As eCLIP results tend to be cell dependent and we aimed to have a global dataset, we, on one hand, pooled all the data, replicated or not, from both cell lines and, on another hand, pooled only the data that was replicated within a cell line. We considered an oRNAment motif instance as matching an eCLIP peak when there was any type of overlap between the two coordinates. This revealed a good correspondence, as defined by the ratio of motif instances identified by oRNAment that are in an eCLIP peak. Furthermore, motif instances defined by oRNAment are always better enriched in eCLIP peaks compared to an equal number of random coordinates taken from the same transcriptomic space that was scanned by oRNAment. As an example, the five motifs recognized by HNRNPK are more highly enriched in HNRNPK eCLIP peaks compared to random coordinates (Figure 4.2iv), while additional examples are shown in Supplemental Figure 4.S1a, b and Supplemental Table S1 [21]. Furthermore, oRNAment displays reasonable false negative rates and precision (Supplemental Figure 4.S1c, d, e, f and Table S1).

The pipeline was executed on all coding (cDNA) and noncoding (ncRNA) transcripts obtained from the FASTA sequences of Ensembl genes release 97, for *Homo sapiens*

(GRCh38), *Caenorhabditis elegans* (WBcel235), *Danio rerio* (GRCz11), *Drosophila melanogaster* (BDGP6), and *Mus musculus* (GRCm38) [22].

2.2. Database implementation

oRNAment is built upon the column-oriented DBMS yandex ClickHouse version 19.5.3.1. The server-side back end of the web application makes use of Django version 2.1.9 and is written in Python 3.7.0. The client interface is implemented in Django's HTML template language with the inclusion, for a greater interactive experience, of several JavaScript libraries, including jquery version 3.3.1, datatables version 1.10.19, charts.js version 2.0, ViennaRNA/fornac.js version 1.1.8, and IGV.js version 2.2.13. The layout styling was created with Bootstrap 4 and Bootstrap-material-design version 4.1.1.

3. Primary features of oRNAment

3.1. Overall functionality

oRNAment contains the position of all motif matches for all PWMs defined by RNAcompete or RBNS, across the transcriptomes (excluding introns) of all five interrogated species. The user can narrow their search to only motifs for which RBPs are represented in a specific species or group of species.

For each type of search, the database outputs distinct figures summarizing the abundance and distribution of motifs across queried transcripts, subregion types (e.g., coding sequence, UTRs), or RNA biotypes (Figure 4.3i-v). It also outputs individual graphs showing the position of all motif instances and their MSS within each transcript for the selected species (Figure 4.3x). Moreover, a detailed table lists all motif instances along with their associated gene name, transcript ID, biotype, position along the transcript, MSS, genomic coordinates, and probability for the 7-mer region to be structurally unpaired, as assessed by RNAplfold predictions[12]. Further detailed information, including the predicted RNA secondary structure (Figure 4.3xi, as assessed by RNAfold, can be accessed for a specific transcript from the table [12]. For a multifaceted overview of multiple motifs, oRNAment also features an embedded Integrated Genome Browser (IGV) (Figure 4.4). All the above information can readily be downloaded as an Excel, CSV, or bed file. This can be achieved either by downloading a subset of the database from the detailed table stemming from a query or by downloading the entire database content.

The database contains a detailed tutorial page to help the user navigate the resource. This section documents the algorithm implemented and the RBP motif data used in oRNAment. Furthermore, it provides comprehensive instructions on how to use each functionality through step-by-step demonstrations using real examples.

3.2. Search by transcripts

This functionality allows the user to query the database for a specific gene, transcript, or group of genes or transcripts, in a specified species, and returns all their putative RBP binding sites. The results are visualized with interactive summarizing charts/histograms (Figure 4.3i–v) and a detailed table. First, a treemap, or histogram, shows the total number of putative instances associated with each RBP. Second, a polar plot, or histogram, illustrates the subregions where these motif instances are observed. Third, a box plot describes the distribution of motif instances within all transcripts searched within oRNAment. This is especially useful when searching for multiple transcripts to determine if they have a common RBP binding site.

3.3. Search by RBP

This functionality allows a user to query the database for a specific RBP, in a specified species, and returns all its putative binding sites in all coding and noncoding transcripts. The user can restrict or expand their query results by specifying the PWM's sensitivity threshold. The results of this query are visualized with interactive summarizing charts and a detailed table (Figure 4.3vi–ix). A doughnut plot, or histogram, shows the total number of putative motif instances identified for the queried RBP grouped by gene biotype allowing a user to, for example, predict protein-noncoding RNA interactions. Finally, a radar plot, or histogram, shows the subregions where these putative motif instances are observed.

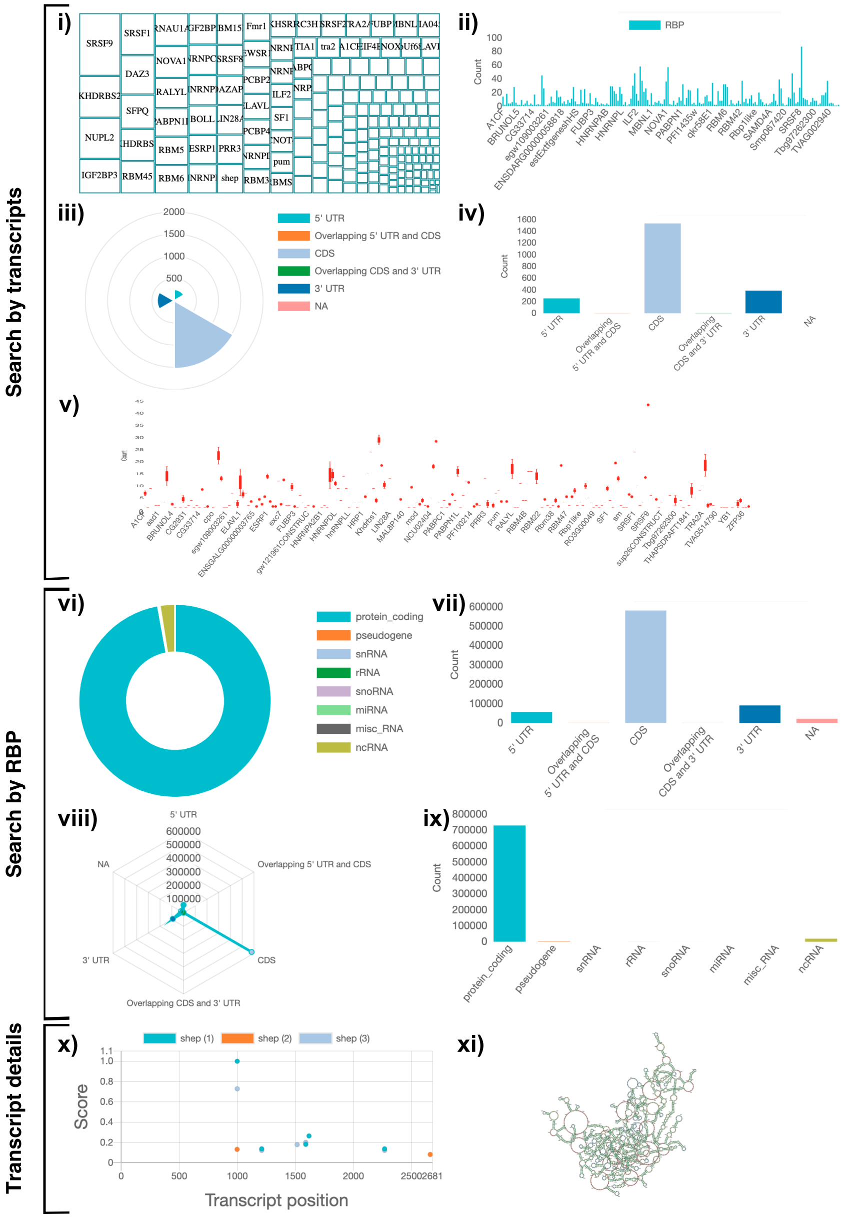


FIGURE 4.3. Examples of the figures generated by oRNAment when searching for motifs in specific RNAs or RBPs. Upon a user's query, either by transcript **(i–v)** or by RBP **(vi-ix)**, multiple figures summarizing the results are provided. **(i–v)** When searching by transcripts (here the *cen* mRNA in *Drosophila*), oRNAment provides : **(i)** a treemap of the most abundant RBP motif instances (likewise shown when searching by attributes) ; or **(ii)** a histogram of the same results ; **(iii)** a polar plot showing in which subregion of the transcript RBP motif instances are observed (here in *cen*) ; or **(iv)** a histogram of the same results ; **(v)** a box plot of the distribution of RBP motif instances in all transcripts queried (here, the boxplot shows the distribution of the number of motif instances among the two isoforms of *cen*). **(vi–ix)** When searching by RBP, oRNAment provides : **(vi)** a doughnut plot showing in which gene biotypes putative binding sites for the queried RBP are observed (here for *SRSF9*) ; or **(vii)** a histogram of the same results ; and **(viii)** a radar plot showing in what transcript subregion putative binding sites for the queried RBP are observed ; or **(ix)** a histogram of the same results. All search functionalities provide a table from which the user can access gene-level or transcript-level details. **(x–xi)** By selecting a gene/transcript and RBP pair, oRNAment will provide : **(x)** a scatter plot showing the position of each putative RBP binding site and corresponding MSS scores, here above the 50% MSS' threshold respectively for each motif of the *shep* RBP on the *cen* mRNA. The transcript positions, on the x-axis, end at the last motif instance + 10 nucleotides ; and **(xi)** a predicted 2D structure of the *cen* transcript as established by RNAfold with default parameters.

3.4. Search by attributes

This functionality allows the user to query the database for a specific combination of transcript attributes [e.g., 3' untranslated region (UTRs) of mRNA, rRNA] in a given species and returns all associated putative RBP binding sites. When an attribute is incompatible with other selections, it is shown as a blocked option (unclickable and greyed out text displaying "NA"). Contrastingly, when selecting the protein coding biotype, the region NA corresponds to the Ensembl annotation for unavailable information and it is selectable. The results are visualized with a treemap, or histogram, showing the total number of putative instances identified for each RBP and a detailed table.

3.5. Interactively visualize motif instances

oRNAment offers the possibility for a user to browse the genome of a given species and interactively visualize putative motif instances of up to three RBPs in an embedded Integrated Genomic Viewer (IGV) browser (Figure 4.4) [**19, 20**]. Unlike a detailed transcript query, which is designed to describe binding sites for specific and individual RBPs, this functionality allows the users to mine the data in a broader exploratory manner. The user can search for one or multiple loci, querying by genomic positions, and visualizing the RBP binding sites along each annotated exon.

4. Conclusion

oRNAment is a modern platform that offers access to a nucleotide-resolution mapping of putative RBP biding sites in the transcriptomes of human and four important model organisms, namely *C. elegans*, *D. rerio*, *D. melanogaster*, and *M. musculus*. The methodology and thresholds employed results in a computationally expensive analysis that produces a large quantity of data. oRNAment palliates this issue by having pre-computed all possible instances through high performance computing resources and by storing the data in a state-of-the-art column-oriented DBMS, which enables efficient retrieval and processing of large quantities of data up to 1000 times faster than traditional data management methods. Altogether, we propose a tool from which the searches and resulting figures are fully interactive and responsive on both desktops and tablets. oRNAment is the first database detailing the

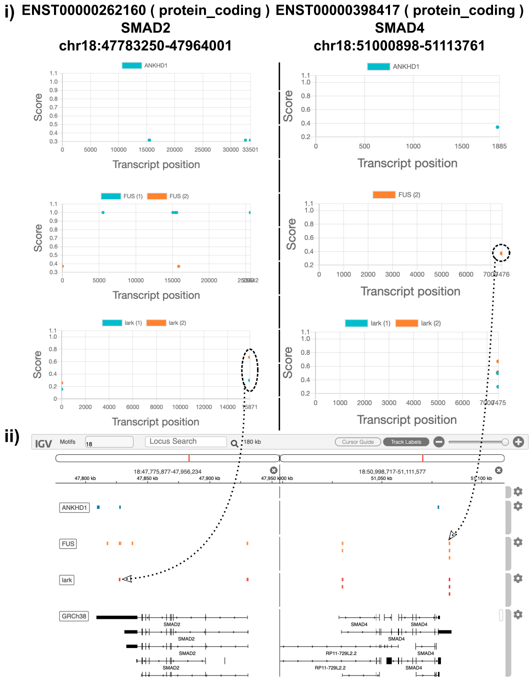


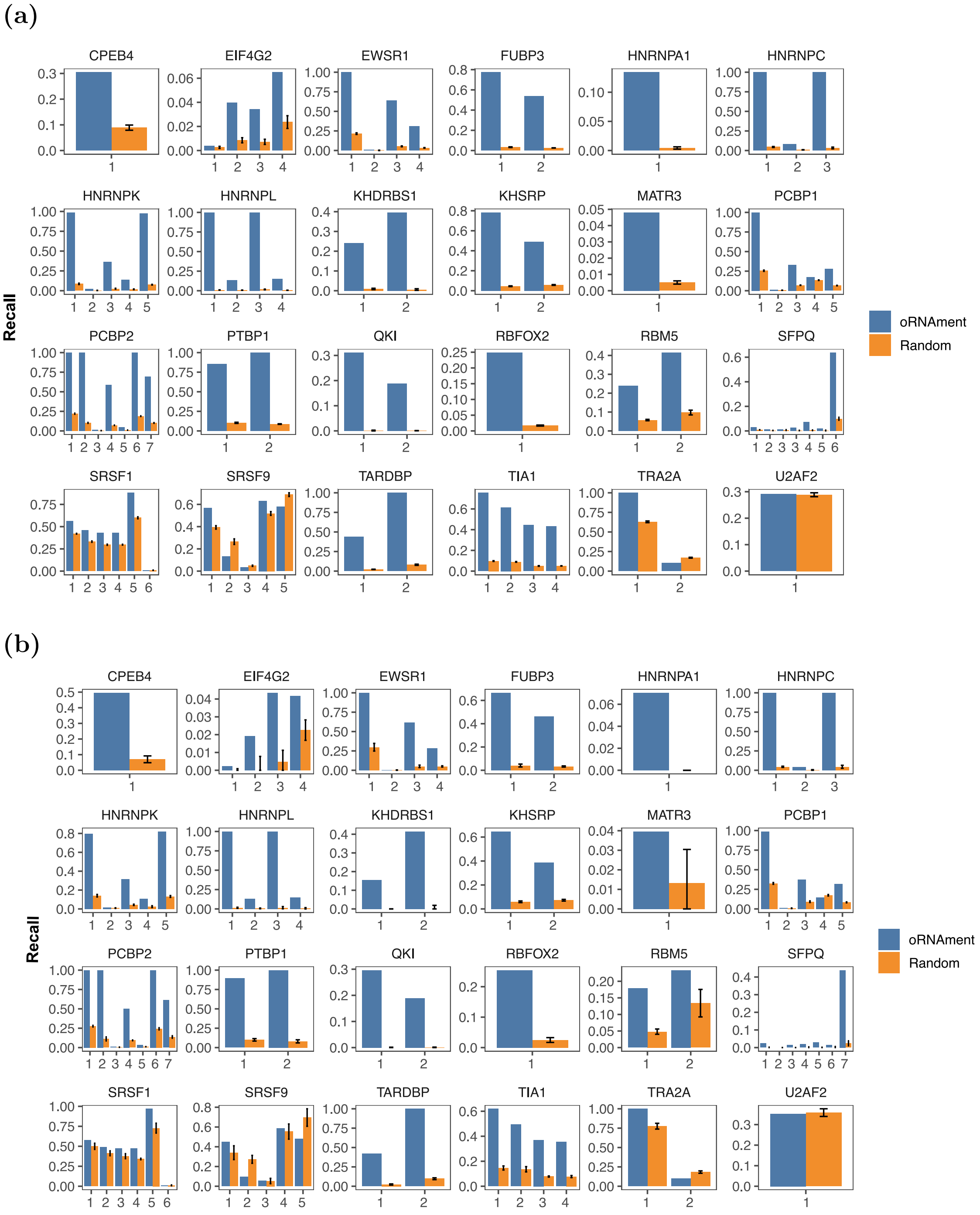
FIGURE 4.4. **Combined visualization of putative binding sites for three RBPs in two genes through a standard scatter plot and an embedded Integrative Genome Browser.** **(i)** Example of oRNAment transcript-level view scatter plot of three RBPs (*ANKHD1*, *FUS*, and *lark*) for two mRNAs (*SMAD2* and *SMAD4*) and **(ii)** Integrative Genome Browser view incorporating the same results when searching for their loci [IGV *Locus search* input in the form : 18 :47783250-47964001 18 :51000898-51113761 (i.e., with a space separating the coordinates)]. Two examples of corresponding motif instances (*lark* in *SMAD2* and *FUS* in *SMAD4*) between the two types of analysis are shown.

transcriptome-wide distribution features of putative RBP target motifs across multiple species. As such, it should prove very useful for users aiming to address hypotheses and to design experiments to study post-transcriptional gene regulation. Future versions will include the complete transcriptome of more species and the addition of other RBPs as their motifs are experimentally defined.

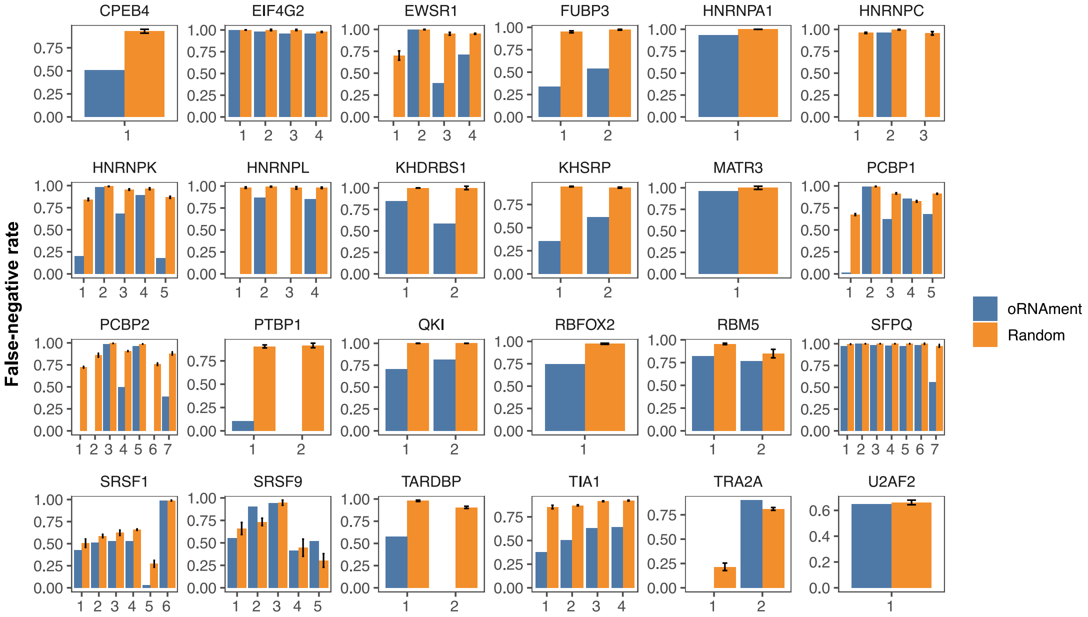
5. Acknowledgements

This research was enabled in part by support provided Calcul Quebec (www.calculquebec.ca), Compute Ontario (www.computeontario.ca) and Compute Canada (www.computecanada.ca). The work was supported by grants from the Canadian Institutes of Health Research (CIHR) to EL and from the Fonds de Recherche Québec - Nature et Technologies to MB and EL, as well as scholarships the Fonds de Recherche Québec - Santé to EL, and from CIHR to SB.

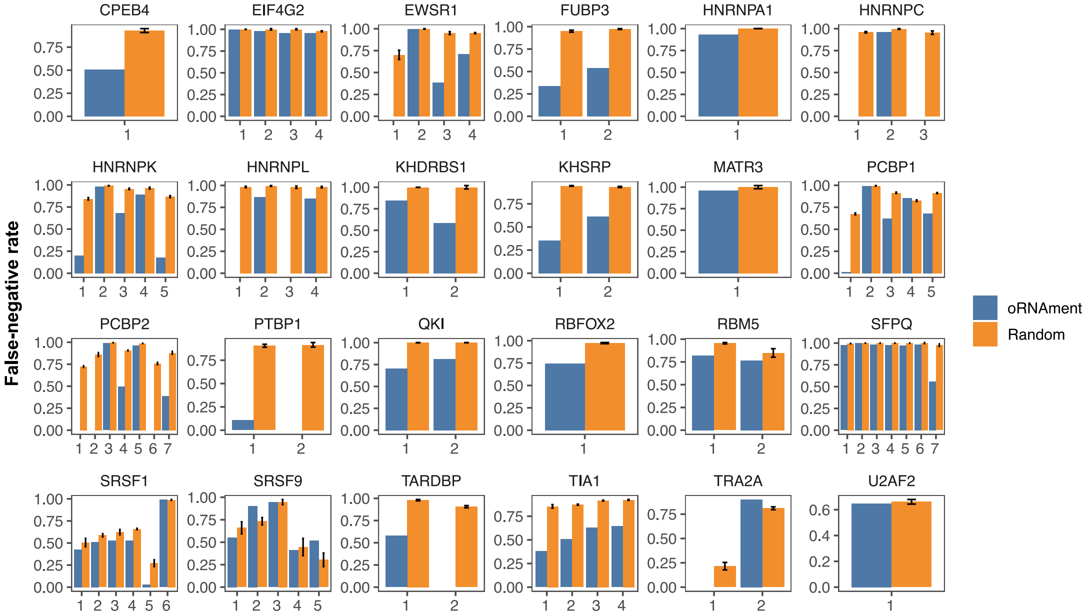
6. Supplemental Data



(c)


False-negative rate
CPEB4
EIF4G2
EWSR1
FUBP3
HNRNPA1
HNRNPC
HNRNPK
HNRNPL
KHDRBS1
KHSRP
MATR3
PCBP1
PCBP2
PTBP1
QKI
RBFOX2
RBM5
SFPQ
SRSF1
SRSF9
TARDBP
TIA1
TRA2A
U2AF2
0.00
0.25
0.50
0.75
1.00
0.0
0.2
0.4
0.6
1
2
3
4
5
6
7
oRNAment
Random

(d)


False-negative rate
CPEB4
EIF4G2
EWSR1
FUBP3
HNRNPA1
HNRNPC
HNRNPK
HNRNPL
KHDRBS1
KHSRP
MATR3
PCBP1
PCBP2
PTBP1
QKI
RBFOX2
RBM5
SFPQ
SRSF1
SRSF9
TARDBP
TIA1
TRA2A
U2AF2
0.00
0.25
0.50
0.75
1.00
0.0
0.2
0.4
0.6
1
2
3
4
5
6
7
oRNAment
Random

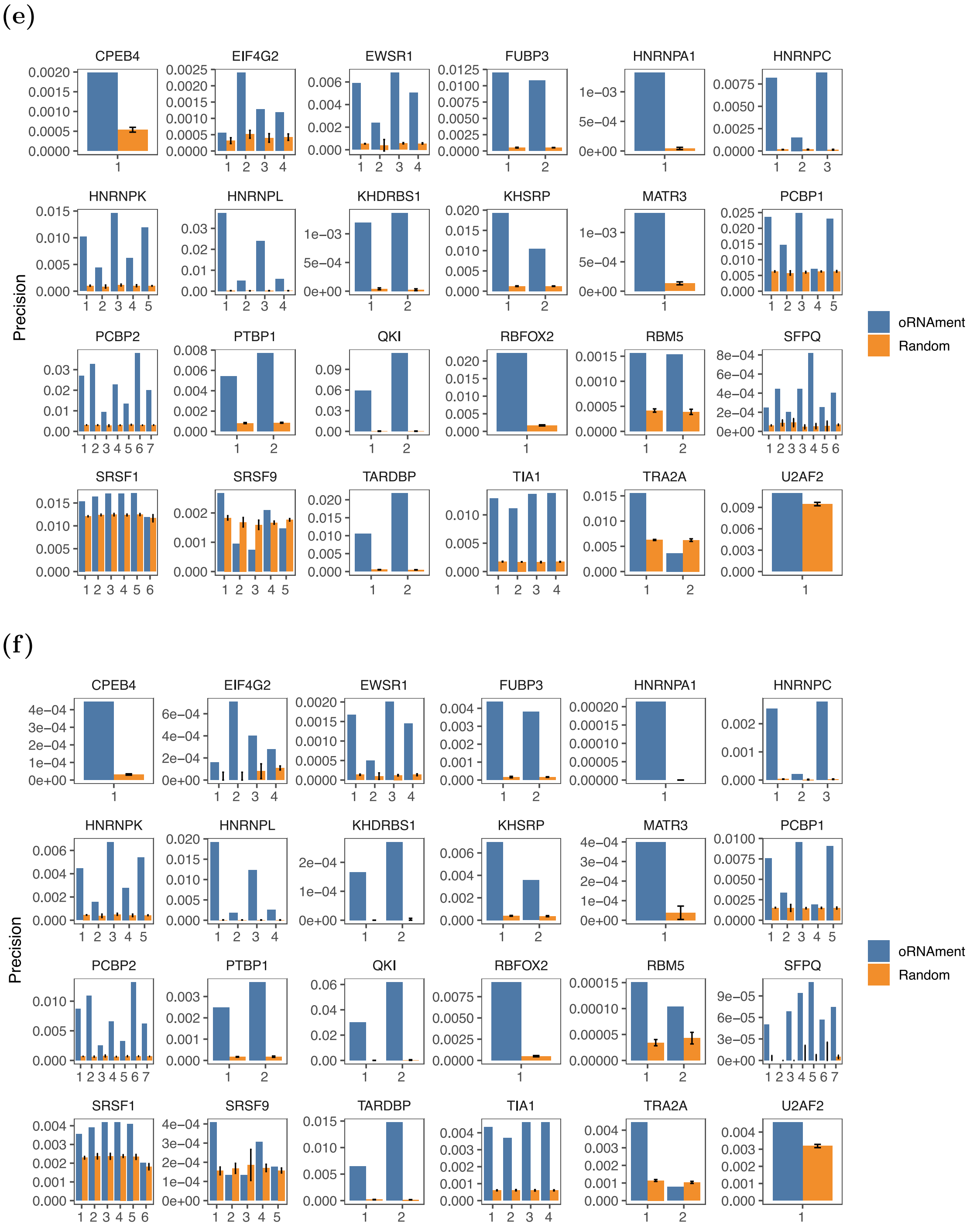


Figure 4.S1. oRNAment reasonably predicts RBP binding sites observed by eCLIP in HepG2 and K562 cell lines. a) Recall (fraction of eCLIP sites that contain a motif hit) for oRNAment predictions (blue) and an equal number of randomly selected regions (5 replicates) (orange) for 24 RBPs. **b)** Same as (a) but with replicated eCLIP hits only. **c)** False-negative rate (fraction of eCLIP binding sites that do not contain a predicted motif hit) for oRNAment predictions (blue) and randomly chosen regions (orange). **d)** Same as in (c) but with replicated eCLIP hits only. **e)** Precision (fraction of motif instances that intersect a eCLIP hit) for oRNAment predictions (blue) and randomly chosen regions. **f)** Same as in (e) but with replicated eCLIP hits only. For all plots, the height of random columns is the median of 5 replicates, with error bars of one standard deviation. Number on x axis represents the various PWMs of each RBP. oRNAment threshold 50% MSS, eCLIP peak calling threshold ≥ 3 fold change, p-value ≤ 0.001.

Références

[1] A. G. Baltz, M. Munschauer, B. Schwanhausser, A. Vasile, Y. Murakawa, M. Schueler, N. Youngs, D. Penfold-Brown, K. Drew, M. Milek, E. Wyler, R. Bonneau, M. Selbach, C. Dieterich, and M. Landthaler, *The mRNA-bound proteome and its global occupancy profile on protein-coding transcripts*, Molecular Cell, 46 (2012), pp. 674–90.

[2] A. Castello, B. Fischer, K. Eichelbaum, R. Horos, B. M. Beckmann, C. Strein, N. E. Davey, D. T. Humphreys, T. Preiss, L. M. Steinmetz, J. Krijgsveld, and M. W. Hentze, *Insights into RNA biology from an atlas of mammalian mRNA-binding proteins*, Cell, 149 (2012), pp. 1393–406.

[3] D. Dominguez, P. Freese, M. S. Alexis, A. Su, M. Hochman, T. Palden, C. Bazile, N. J. Lambert, E. L. Van Nostrand, G. A. Pratt, G. W. Yeo, B. R. Graveley, and C. B. Burge, *Sequence, structure, and context preferences of human RNA binding proteins*, Molecular Cell, 70 (2018), pp. 854–67 e9.

[4] S. Gerstberger, M. Hafner, and T. Tuschl, *A census of human RNA-binding proteins*, Nature Reviews Genetics, 15 (2014), pp. 829–845.

[5] G. Giudice, F. Sanchez-Cabo, C. Torroja, and E. Lara-Pezzi, *ATtRACT-a database of RNA-binding proteins and associated motifs*, Database, 2016 (2016).

[6] T. Glisovic, J. L. Bachorik, J. Yong, and G. Dreyfuss, *RNA-binding proteins and post-transcriptional gene regulation*, FEBS Letters, 582 (2008), pp. 1977–86.

[7] A. E. Kel, E. Gossling, I. Reuter, E. Cheremushkin, O. V. Kel-Margoulis, and E. Wingender, *MATCH : A tool for searching transcription factor binding sites in DNA sequences*, Nucleic Acids Research, 31 (2003), pp. 3576–9.

[8] N. Lambert, A. Robertson, M. Jangi, S. McGeary, P. A. Sharp, and C. B. Burge, *RNA Bind-n-Seq : quantitative assessment of the sequence and structural binding specificity of RNA binding proteins*, Molecular Cell, 54 (2014), pp. 887–900.

[9] N. J. Lambert, A. D. Robertson, and C. B. Burge, *RNA Bind-n-Seq : Measuring the binding affinity landscape of RNA-binding proteins*, Methods in Enzymology, 558 (2015), pp. 465–93.

[10] X. Li, H. Kazan, H. D. Lipshitz, and Q. D. Morris, *Finding the target sites of RNA-binding proteins*, Wiley Interdisciplinary Reviews–RNA, 5 (2014), pp. 111–30.

[11] Y. Liu, S. Sun, T. Bredy, M. Wood, R. C. Spitale, and P. Baldi, *MotifMap-RNA : a genome-wide map of RBP binding sites*, Bioinformatics, 33 (2017), pp. 2029–31.

[12] R. Lorenz, S. H. Bernhart, C. Honer Zu Siederdissen, H. Tafer, C. Flamm, P. F. Stadler, and I. L. Hofacker, *ViennaRNA package 2.0*, Algorithms for Molecular Biology, 6 (2011), p. 26.

[13] B. M. Lunde, C. Moore, and G. Varani, *RNA-binding proteins : modular design for efficient function*, Nature Reviews Molecular Cell Biology, 8 (2007), pp. 479–90.

[14] E. L. Nostrand, P. Freese, G. A. Pratt, X. Wang, X. Wei, S. M. Blue, D. Dominguez, N. A. L. Cody, S. Olson, B. Sundararaman, R. Xiao, L. Zhan, C. Bazile, L. P. Benoit Bouvrette, J. Chen, M. O. Duff, K. Garcia, C. Gelboin-Burkhart, M. Hochman, N. J. Lambert, H. Li, T. B. Nguyen, T. Palden, I. Rabano, S. Sathe, R. Stanton, A. L. Louie, S. Aigner, J. Bergalet, B. Zhou, A. Su, R. Wang, B. A. Yee, X.-D. Fu, E. Lecuyer, C. B. Burge, B. Graveley, and G. W. Yeo, *A large-scale binding and functional map of human RNA binding proteins*, bioRxiv, (2018), p. 179648.

[15] I. Paz, I. Kosti, J. Ares, M., M. Cline, and Y. Mandel-Gutfreund, *RBPmap : a web server for mapping binding sites of RNA-binding proteins*, Nucleic Acids Research, 42 (2014), pp. W361–7.

[16] K. Quandt, K. Frech, H. Karas, E. Wingender, and T. Werner, *MatInd and MatInspector : new fast and versatile tools for detection of consensus matches in nucleotide sequence data*, Nucleic Acids Research, 23 (1995), pp. 4878–84.

[17] D. Ray, H. Kazan, E. T. Chan, L. Pena Castillo, S. Chaudhry, S. Talukder, B. J. Blencowe, Q. Morris, and T. R. Hughes, *Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins*, Nature Biotechnology, 27 (2009), pp. 667–70.

[18] D. Ray, H. Kazan, K. B. Cook, M. T. Weirauch, H. S. Najafabadi, X. Li, S. Gueroussov, M. Albu, H. Zheng, A. Yang, H. Na, M. Irimia, L. H. Matzat, R. K. Dale, S. A. Smith, C. A. Yarosh, S. M. Kelly, B. Nabet, D. Mecenas, W. Li, R. S. Laishram, M. Qiao, H. D. Lipshitz, F. Piano, A. H. Corbett, R. P. Carstens, B. J. Frey, R. A. Anderson, K. W. Lynch, L. O. Penalva, E. P. Lei, A. G. Fraser, B. J. Blencowe, Q. D. Morris, and T. R. Hughes, *A compendium of RNA-binding motifs for decoding gene regulation*, Nature, 499 (2013), pp. 172–7.

[19] J. T. Robinson, H. Thorvaldsdottir, W. Winckler, M. Guttman, E. S. Lander, G. Getz, and J. P. Mesirov, *Integrative genomics viewer*, Nature Biotechnology, 29 (2011), pp. 24–6.

[20] H. Thorvaldsdottir, J. T. Robinson, and J. P. Mesirov, *Integrative Genomics Viewer (IGV) : high-performance genomics data visualization and exploration*, Briefings in Bioinformatics, 14 (2013), pp. 178–92.

[21] E. L. Van Nostrand, G. A. Pratt, A. A. Shishkin, C. Gelboin-Burkhart, M. Y. Fang, B. Sundararaman, S. M. Blue, T. B. Nguyen, C. Surka, K. Elkins, R. Stanton, F. Rigo, M. Guttman, and G. W. Yeo, *Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP)*, Nature Methods, 13 (2016), pp. 508–14.

[22] D. R. Zerbino, P. Achuthan, W. Akanni, M. R. Amode, D. Barrell, J. Bhai, K. Billis, C. Cummins, A. Gall, C. G. Giron, L. Gil, L. Gordon, L. Haggerty, E. Haskell, T. Hourlier, O. G. Izuogu, S. H. Janacek, T. Juettemann, J. K. To, M. R. Laird, I. Lavidas, Z. Liu, J. E. Loveland, T. Maurel, W. McLaren, B. Moore, J. Mudge, D. N. Murphy,

V. Newman, M. Nuhn, D. Ogeh, C. K. Ong, A. Parker, M. Patricio, H. S. Riat, H. Schuilenburg, D. Sheppard, H. Sparrow, K. Taylor, A. Thormann, A. Vullo, B. Walts, A. Zadissa, A. Frankish, S. E. Hunt, M. Kostadima, N. Langridge, F. J. Martin, M. Muffato, E. Perry, M. Ruffier, D. M. Staines, S. J. Trevanion, B. L. Aken, F. Cunningham, A. Yates, and P. Flicek, *Ensembl 2018*, Nucleic Acids Research, 46 (2018), pp. D754–61.

Quatrième article.

Broad multi-species survey of RNA-binding protein target motifs reveals conserved motif environments in coding and noncoding transcriptomes

Louis Philip Benoit Bouvrette

Article en préparation

Préface et contributions

Ce chapitre est présenté sous la forme d'un article de recherche que nous visons à soumettre au journal *RNA*.

L'article présente la suite de l'article précédant, décrivant la base de données oRNAment, et fournit la première étude comparative multiespèces pantranscriptomique cartographiant la position et la conservation des sites potentiels de liaisons RBP-ARN. Je me suis servie de cette ressource unique afin de caractériser précisément le contexte positionnel et environnemental de 223 protéines liant l'ARN chez cinq espèces, soit *Homo sapiens*, *Caenorhabditis elegans*, *Danio rerio*, *Drosophila melanogaster*, et *Mus musculus*. Ceci m'a permis de profiler globalement la distribution des sites de liaison des protéines pour chaque transcrit codant et non codant et d'évaluer leurs corrélations interespèces. Cette comparaison a démontré un large éventail de caractéristiques d'enrichissement régional de chaque protéine et a révélé des points communs entre des sous-groupes de protéines reliant préférentiellement des biotypes distincts ou des régions d'ARN spécifiques. De plus, je démontre que ces sous-groupes de protéines présentent de fortes corrélations et anti-corrélations entre espèces. L'analyse d'instances de motifs entre des régions transcriptomiques comparables chez les vertébrés suggère des conservations synténiques avec jusqu'à 59 % de sites dans les transcriptions codantes et 86 % dans les biotypes de transcrits non codants démontrant un site de liaison potentiel pour le même motif entre deux espèces. Mes résultats ont également permis d'analyser les propriétés topologiques des sites de liaison RBP. La topologie régionale de certains motifs en tant qu'instances répétées à intervalle égal semble également être conservée évolutivement comme pour la protéine *RC3H1* liant le transcrit *INSM1-201* et ses orthologues. De plus, nous décrivons un nouveau rôle potentiel pour le long ARN non codant *HELLPAR* comme une éponge de protéines liant l'ARN. Ce travail illustre la puissance de la cartographie à grande échelle de plusieurs protéines liant l'ARN avec des fonctions distinctes permettant de révéler des notions inédites de la biologie des ARN.

J'ai conceptualisé l'étude, énoncé les hypothèses, développé les scripts et effectué les analyses bio-informatiques. J'ai aussi créé l'ensemble des figures et écrit l'intégralité du manuscrit. Eric Lécuyer et Mathieu Blanchette ont supervisé les travaux et l'écriture du manuscrit et ont suggéré des modifications concernant, principalement et respectivement, les aspects biologique et informatique.

Broad multi-species survey of RNA-binding protein target motifs reveals conserved motif environments in coding and noncoding transcriptomes

par

Louis Philip Benoit Bouvrette[1, 2], Mathieu Blanchette[4] et Eric Lécuyer[1, 2, 3]

(1) Institut de Recherches Cliniques de Montréal (IRCM) Montréal, Québec, Canada.
(2) Département de biochimie, Université de Montréal, Montréal, Québec, Canada
(3) Division of experimental medicine, McGill University, Montréal, Québec, Canada
(4) School of computer science, McGill University, Montréal, Québec, Canada

Article en préparation .

ABSTRACT. Protein-RNA interactions carry fundamental roles in nearly every facet of RNA metabolism encompassing everything from synthesis, trafficking, stability to degradation. A critical step in understanding post-transcriptional gene regulation is to analyze the topological features of RNA-binding protein binding sites at a transcriptome-wide level and assess their broad evolutionary conservation between species. The consensus motif of a wide range of RBPs has been defined through *in vitro* selection methods which established significant datasets permitting for large-scale motif scanning. The previous development of an efficient computational scanning algorithm enabled us to comprehensively catalogue RNA motif instances in both coding and noncoding transcripts, creating a unique resource of RBP-RNA interactomes. This allowed us to characterize the situational context of 223 RBPs in five species (*Homo sapiens*, *Caenorhabditis elegans*, *Danio rerio*, *Drosophila melanogaster*, and *Mus musculus*), displaying a broad range of abundance features. The comparison of motif instances enrichment revealed commonalities between subgroups of proteins preferentially linking distinct biotypes or specific RNA regions with strong correlations and anti-correlations between species. Analysis of motif instances between comparable transcriptomic regions in vertebrates suggests syntenic conservation with up to 59% of sites in coding transcripts and between 30% and 86% in noncoding transcript biotypes bearing the same motif between two species. The regional topology of certain motifs, like *INSM1-201* and its orthologs, as repeated instances also appear to be evolutionarily conserved. Further, we describe a potential new role for the long noncoding RNA *HELLPAR* as an RNA-binding protein sponge. This work illustrates the power of large-scale scanning of multiple RBPs with distinct functions to reveal novel RNA biology.

Keywords: RNA-binding protein, motif, transcriptome-wide, evolutionary conservation, mRNA, noncoding RNA

1. Introduction

Protein-RNA interactions execute important roles in several biological functions. RNA-binding proteins (RBPs) can regulate a wide variety of cellular processes and post-transcriptional gene expression, such as RNA replication, repair, splicing, polyadenylation, capping, export, localization, stability and degradation [41, 32, 4]. It is currently hypothesized that the human genome encodes more than 1500 RBPs [13, 15, 26]. Most of these RBPs possess well defined RNA binding domains (RBDs) that bind RNA through a sequence- and/or structure-specific motif [13, 26, 35, 44, 40]. Although there is a high number of RBPs with established RBDs, the specific transcript subsequences bound by most RBPs are still incompletely mapped. It is suggested that gene expression is greatly influenced by post-transcriptional regulation and an increasing number of RBPs are being associated with diseases [4, 6, 54, 53, 5, 43, 28]. Therefore, a better characterization of transcriptome-wide and interspecies RNA motif instances is critically needed to improve our understanding of post-transcriptional gene regulation.

To this aim, several *in vitro* methods were developed to identify the transcript's motifs bound by RBPs. One such approach, named RNAcompete, involves a high-throughput *in vitro* selection strategy to assess the binding sequence preferences of select RBPs. For this, thousands of designed RNA oligonucleotides are used to assess the sequence binding motifs of different RBPs by microarray [40, 39]. Recently, an RNA-seq approach, named RNA Bind-n-Seq (RBNS), was developed [10, 22]. This method, where different concentrations of a recombinant RBP is incubated with a random pool of RNA oligonucleotides of a fixed concentration, allows for the estimation of bound RNA molecules compared to an input. While these methods allow for identification of motifs at a nucleotide resolution, they lack situational context. Indeed, despite these recent advances, the repertoires of transcript motif instances bound by different RBPs, along with their environment and their inter-species conservation remain mostly unknown.

The challenge of predicting *cis*-regulatory elements computationally has been widely studied and the combination of *in vitro* data and *in silico* algorithms has been proven useful in a plethora of conditions [20, 30, 33, 19, 9, 2, 29]. To this end, we recently introduced oRNAment, a computational tool and database that robustly and comprehensively catalogues RBP binding motif instances in the complete non-intronic coding and noncoding transcriptomes of human and 4 model species (*C. elegans*, *D. rerio*, *D. melanogaster*, *M. musculus*) [3]. As we were not invested on any specific RBP, but rather on portraying a systematic map of binding instances in multiple transcriptomes, we exploited the publicly available motifs for 223 RBPs described by RNAcompete and RBNS.

To gain insights into the RBP-mediated post-transcriptional regulatory logic, here we performed an in-depth characterization of RBP motif distribution properties across species. Our results reveal both general features of the RBP motif environment within specific RNA biotypes, as well as their conservation between species, while also providing specific mechanistic insights for individual RBPs and transcripts.

2. Transcriptome-wide profiling of putative RBP target sites across five species

We previously developed a computational pipeline that allows for the identification of putative RBP binding sites in a transcript (i.e., an observation of a transcript region that is sufficiently similar to an RBP binding motif to be considered a putative binding site). We refer to these regions as "predicted RBP *cis*-regulatory motif instances" or pCRMI. Our method was proven reasonably efficient, both in running time and memory usage, to scan the entire non-intronic transcriptome and correctly identify proper binding sites [**3**]. More specifically, we applied our scanning algorithm to 453 motifs for 223 RBPs derived from their *in vitro* characterization by either RNAcompete and/or RNA Bind-n-Seq (RBNS) experiments. This mapping was performed on the complete set of non-intronic coding and noncoding transcripts for both human and four model organisms : *Caenorhabditis elegans*, *Danio rerio*, *Drosophila melanogaster*, and *Mus musculus* (Figure 5.1A). As potential motif instances display varying degrees of similarity to the most probable *in vitro* defined consensus motif, we compiled each RBP motif independently according to individual thresholds (see methods). We hence built a database containing a comprehensive description of all pCRMI which covers 525,718 transcripts across five species (Figure 5.1A).

The dataset of 223 RBPs studied here covers a wide range of biological processes that most commonly have functions pertaining to gene expression, RNA splicing, and stability (Supplemental Figure 5.S1). While not all RBPs are expressed in all studied species, we nonetheless scanned each transcriptome for all RBPs in order to better capture potential evolutionary events, such as motif instance conservation. Consequentially, we have annotated each RBP with regards to their species specificity. That is, as RBP orthologs are expected to exhibit similar binding properties, we have grouped these RBPs together if their RNA-binding domains have amino acid sequences that are 70% or more similar as defined by Ray *et al.* [**40, 39**].

We first sought to establish a general and broad overview of the RBP repertoire. Standard motif instances cataloguing revealed a wide variety of binding modes to noncoding RNA and mRNAs. Importantly, our prediction method adjusts the threshold so as to yield the same FDR for every RBPs, therefore these observations are unlikely to be due to an FDR that would be more elevated in some pCRMI than others (see methods). Overall, we observed

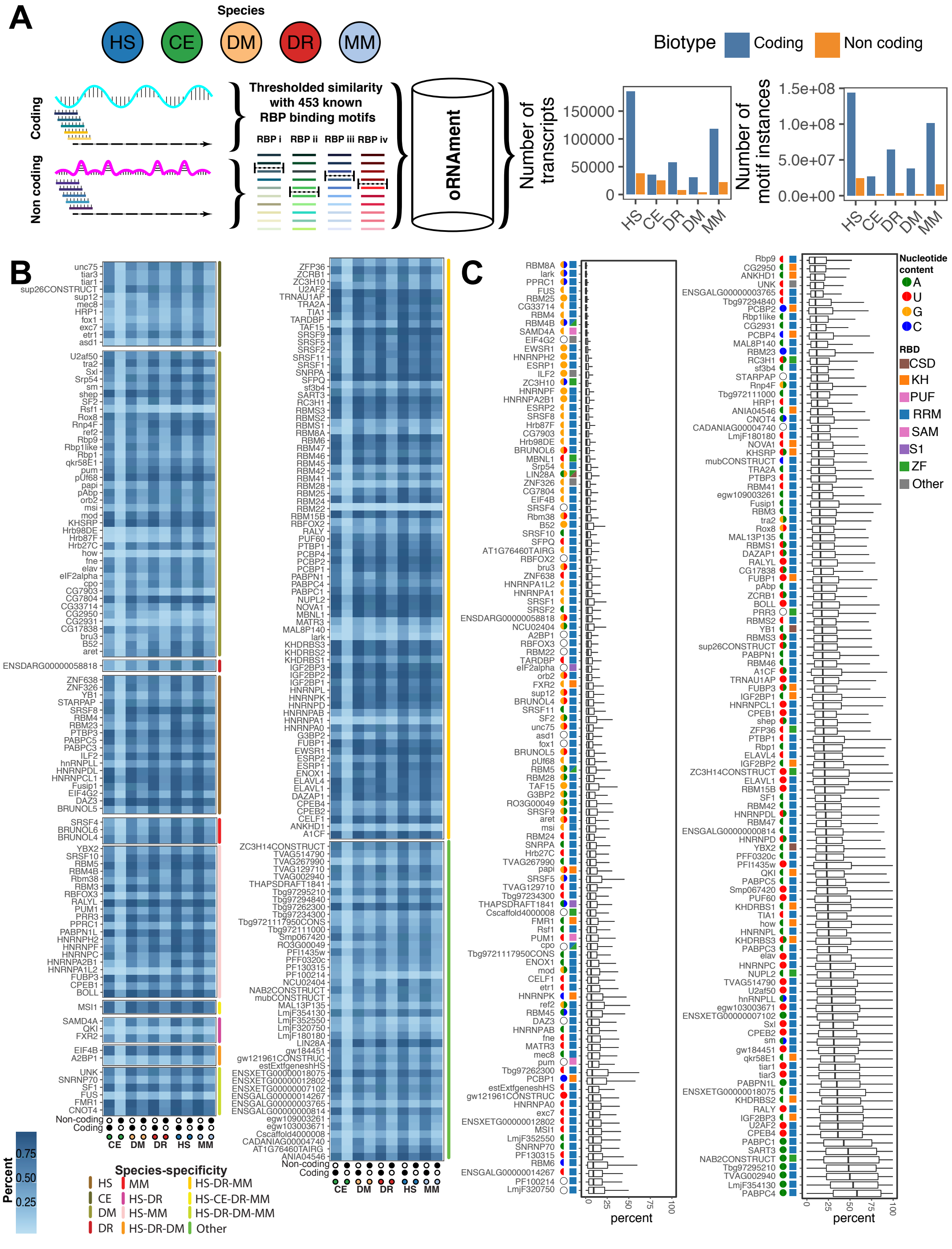


Figure 5.1. Overview of the oRNAment pipeline and data. **A)** Schematic of the motif search algorithm and database populating, executed on 5 species for both coding and noncoding transcripts. Transcripts 7-mers regions (dashed arrows) are compared to each known RBP motifs (left) where those above an individual RBP score threshold (coloured dash lines) populate a database describing all transcripts and their motifs instances (right). **B)** Heatmap showing the fraction of coding and noncoding transcripts exhibiting at least one motif instance for each RBP. RBP are grouped by their species specificity according to Ray *et al.* [40, 39]. **C)** Box plot showing the predicted percent of unpaired probability, determined by RNAplfold, of the 7-mer region bearing an RBP motif instance, ordered by their median. Coloured circles represent nucleotide content of each RBP motif (one circle of one colour if motif is >6 of one base, two half circles of different colours if motif is >3 of two distinct bases, half circle if motif is >3 of 1 base) for the canonical motif logo for each protein. Coloured squares represent RBD domain (2 tones half coloured when protein has 2 described RBD). HS : *Homo sapiens*, CE : *Caenorhabditis elegans*, DM : *Drosophila melanogaster*, DR : *Danio rerio*, MM : *Mus Musculus*.

that the pCRMI of *PCBP2* was the most common in human mRNA. *RALYL*, despite only being characterized here as *H. sapiens*- and *M. musculus*-specific with regards to amino-acid similarity, has the most pCRMI observed in *C. elegans* (where its closest homolog is named *ztf-4*), *D. rerio* and *M. musculus* mRNA, where both species have known orthologs (Figure 5.1B) [**45**]. *NUPL2* is the most observed RBP in *D. melanogaster* mRNA, where its ortholog is named *CG18789* [**47**]. In contrast, the pCRMI of *RBM22*, a protein with a known role in cell division and splicing, was the least observed RBP in the mRNA of all five species (Figure 5.1B). *RALYL* and *RBM22* were also the most and least observed pCRMI in the ncRNA of all five species. Manifestly, a non-negligible number of pCRMI are likely to be false positives but there are no *a priori* reasons to expect enrichments in particular regions of the transcriptomes which would offset interspecies comparisons. Furthermore, while the number of instances can greatly vary for each RBP pCRMI, we previously observed a good correspondence between human eCLIP data, and the ordered results of raw observations are comparable to previous such characterization efforts in human [**3, 50**].

We then established, for each RBP, the percent of transcripts exhibiting at least one instance of its motif (Figure 5.1B). Interestingly, we observed a broad range of RBP motif abundance features with some showing promiscuousness and others being much more selective. Expectedly, the most predominant pCRMI observed overall is also among the pCRMI observed in the greatest number of transcripts, regardless of length. Indeed, although the pCRMI of *SRSF9* was the most observed, identified at least once in 99% of coding transcript of all five species, the pCRMI of *PCBP2* and *RALYL* are observed at least once in 82% to 98% of all coding transcripts. Contrastingly, the pCRMI of *RBM22* is observed in 4% to 5% of transcripts. Notably, the median of coding transcripts containing at least one pCRMI for each RBP is roughly twice as much as for noncoding transcripts with a median between 64% (*C. elegans*) and 82% (*H. sapiens*) coding transcripts and a median between 10% (*C. elgans*) and 39% (*M. musculus*) of noncoding transcripts with at least one pCRMI. Interestingly, while the distribution of pCRMI throughout the transcriptome can be clustered with the three vertebrates together and the worm and fly together, these overall ratios are similar across all species studied (Supplemental table 5.S1).

We further characterized the context in which each RBP motif instance was observed by assessing the probability that its whole binding region is paired (structured) or unpaired. For this, we executed RNAfold on every transcript for each species. By ordering the predicted unpaired probability distribution of each pCRMI by their median, we observed them to be on a gradient (Figure 5.1C, Supplemental figure 5.S2). Interestingly, RBP motifs that are more G-rich tend to be mainly observed in structured regions. RBP motifs that are more A- and T-rich are more enriched in single strand regions. Notably, RBP motifs that are largely observed in paired regions tend to have much narrower distribution (22 of the 59 motifs with >3 Gs have a maximum unpaired probability less than 10%) while the A- or T-rich motifs are observed in a wider range of secondary structure.

We complemented the contextual characterization of each RBP by analyzing their structural binding preference with regards to their RNA-binding domain (RBD). For each RBP we gathered their known RBD from three sources : RBPDB, pfam, and the NCBI Conserved Domain Database [**7, 11, 27**]. Only four RBPs, namely *RBM4B*, *LIN28A*, *RBM5* and *CPO*, exhibited more than one type of RBD (Figure 5.1C). No correlation between an RBP motif and an RBD was found. Most RBPs have an RRM RBD and are evenly distributed over the variation of secondary structures with 95 RBPs below the median and 84 above (Figure 5.1C). Contrastingly, RBPs with a KH domain tends to have a pCRMI in structured regions with only 5 RBPs below the median while 18 are above (Figure 5.1C).

We conclude that, generally, most known RBP motifs are observed in a similar fashion across species, where highly, or lowly, abundant motifs are present in a corresponding proportion of transcripts. Additionally, most known RBP motifs are observed in a greater number of mRNAs than in all other noncoding biotypes combined. Furthermore, GC-rich motifs show a greater structured signature than AT rich motifs while RBPs with a KH domain bind regions that are less structured. Finally, this appears to hold true for all species studied regardless of the species specificity of the RBP.

3. Motifs instances within transcript biotypes and regions exhibit general apparent correlations

To gain global insights into the relative position of motif instances of RBPs within different biotypes and then within different regions (5'UTR, coding, 3'UTR) of mRNAs, we overlapped motif instances with the genome annotation of each species. As annotations can exhibit variations between sources, possibly leaving behind some transcripts even with a corresponding fasta sequence, we combined the annotations from Ensembl GTF, Ensembl GFF files and the ensembldb bioconductor R library (each with the version GRCh38.98, WBcel235.98, BDGP6.22.98, GRCz11.98, GRCm38.98 for human, worm, fly, zebrafish, and mouse respectively) [21, 56]. When a transcript or exon was documented differently in an annotation, the longest one was kept. In order to compare and establish motif instance enrichment correlations between species, we normalized the number of instances by kilobase of region (i.e., sum of all transcript exons annotated as a given region), for coding transcripts, or by kilobase of the sum of transcript length by biotype (i.e., all lncRNA, snoRNA, etc. lengths independently). For coding transcripts, we further established the ratio, for each RBP, as its percent distribution in each region. Overall, for human, we observed that RBP motif instances overlapping specific regions were generally consistent with previous findings. Specifically, when comparing with the distribution of eCLIP peak results within coding transcripts, we observed overall Pearson correlations of 0.48, 0.59, 0.63 for 5'UTR, CDS and 3'UTR respectively for the 27 RBPs comparable with HepG2 cells (all p-values < 0.01) and 0.66, 0.5, and 0.52 respectively for the 29 RBPs comparable with K562 cells (all p-values $<$ 0.006) (Supplemental figures 5.S3 and 5.S4) [50]. We conclude that our method reasonably predicts the distributions and enrichments of motif instances within a transcript region.

Based on the relative abundance of transcript biotype or region type containing a motif instance for each RBP for each species, we performed a t-SNE clustering analysis (Figure 5.2A). Interestingly, for coding transcripts, each region and species clustered strongly together suggesting a wide conservation of binding sites. By contrast, noncoding RNA biotypes exhibited more diverse patterns of clustering (Figure 5.2A). Nevertheless, many biotypes, such as snRNA, rRNA, Mt rRNA and Mt tRNA, showed clustering between species. These clusterings often exhibit distinct groupings with the vertebrates as one cluster and worm and fly as another, likely an attribute of their divergent evolution. Contrastingly, lncRNA did not display any grouping between species, consistent with their known diverse nature and expression even between cell types in the same species [36, 42].

To further evaluate the individual correspondence of each RBP motif instance distribution among biotypes between species, we calculated every possible pairwise Pearson correlation of their distribution (Figure 5.2B). Consistent with the t-SNE clustering, coding transcripts exhibited overall stronger correlations than noncoding transcripts. However, multiple RBPs

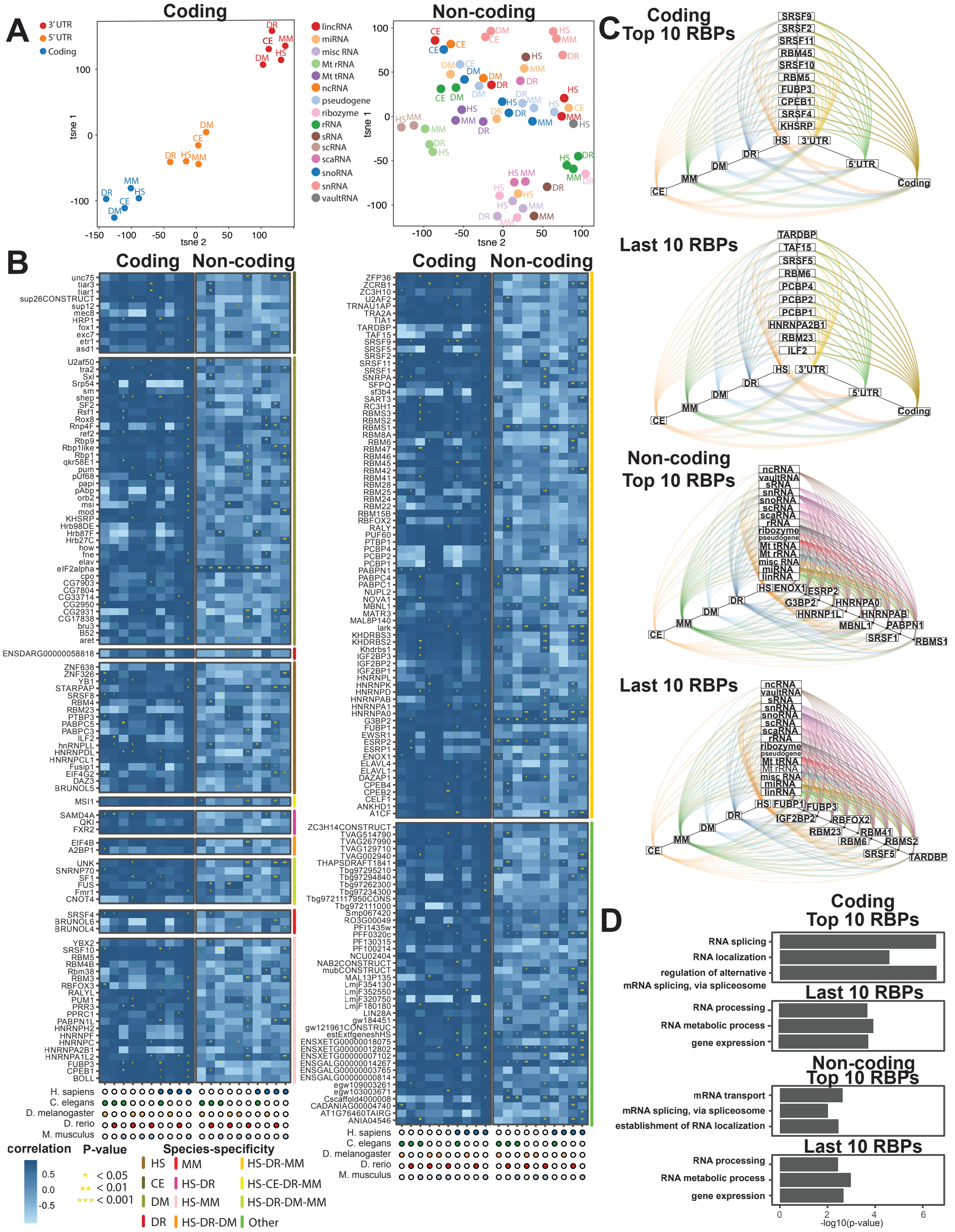


Figure 5.2. RBP binding site instances enrichment within mRNA regions or noncoding RNA biotypes exhibits overall prominent level of correlation. A) t-SNE plot showing the clustering of the relative instance positions, normalized in instances per KB, in each region (5'UTR, coding, 3'UTR) of all 223 RBPs in all mRNA (left) and the relative enrichment, normalized in instances per KB, of all RBPs in noncoding RNA (right) for 5 species. **B)** Heatmap showing the 10 pairwise Pearson correlations, for all 223 RBPs relative instance positions in each region, of the normalized, in instances per KB, (5'UTR, coding, 3'UTR) of all 223 RBPs in all mRNA (left) and the 10 pairwise correlations of the enrichment of all RBPs in noncoding RNA, normalized in instances per KB (right). Right side colour legend represents species specificity of each RBP according to Ray *et al.* [40, 39]. **C)** Hive plot showing the top 10 RBPs with the highest average correlation between the 5 species and the last 10 RBPs with the lowest average correlation between the 5 species for mRNA (top) and noncoding RNA (bottom). The line width is relative to the number of relative instance positions, normalized in instances per KB, in each region, for coding, or noncoding biotypes. **D)** Histogram showing 3 GO terms enrichment for the top 10 and last 10 RBPs for coding and noncoding RNA described in (C). HS : *Homo sapiens*, CE : *Caenorhabditis elegans*, DM : *Drosophila melanogaster*, DR : *Danio rerio*, MM : *Mus Musculus*.

showed high correlation between at least a subset of species. For example, human and mouse show the overall better correlation for most RBPs for both coding and noncoding transcripts. Interestingly, species specificities of the protein or its RBD does not appear to have an influence on the presence of motif instances for them in the interspecies correlations (Figure 5.2B). Our data does not permit to identify which RBPs have been gained or lost between each species during evolution. Nonetheless, this suggests that the transcript sequence among these closely related, and especially its RBP-binding region, has not been significantly mutated and is still similar to the canonical motif at our given threshold.

We next sought to assess the functionalities of RBPs that appear to have more, or less, evolutionarily conserved motif instances. For this we calculated the average correlation between all pairwise comparisons and established the 10 RBPs with the highest overall correlations and the 10 RBPs with the lowest correlations or that revealed anti-correlations (Figure 5.2C). We then performed statistical overrepresentation tests of GO terms on these RBPs. Strikingly, the RBP with the highest correlations, whether from coding or noncoding transcripts, are involved in important functions such as splicing and transport while RBPs with the lowest correlations have the same top-level GO terms (Figure 5.2D).

Overall, these results suggest binding site commonalities between each subset of RBPs binding specific biotypes and regions between species wherein extensive related patterns are likely associated with similar major functions. We hypothesize that the RBPs and genes involved in these mechanisms may require precise regulations, which, in turn, necessitate an exquisite set of *cis*-regulatory motif instances.

4. Conservation and divergence of RBP binding site instances

The extent of RBPs and transcripts catalogued provides to us a unique possibility to explore their conservation between species. We asked if an RBP pCRMI at a given genomic coordinate was also observed in the syntenic region of another species. Therefore, we translated every transcript coordinate with a pCRMI to its genomic coordinate with an in-house script and further converted it to the genomic coordinate of another species using the liftover tool from UCSC, which uses whole-genome pairwise alignments [18]. For this, we only considered the three vertebrates, because the transcriptomes of worm and fly are too divergent for robust genomic coordinate translation. We concentrated our effort on exact syntenic regions, excluding from our analysis everything off by even a single nucleotide. Moreover, to disentangle the data and remove ambiguities, when a position had pCRMI for multiple RBPs, often due to an RBP having many similar motifs, we only considered the one showing better similarity with the canonical motif.

We then aimed to determine the extent to which pCRMI predicted in one species were conserved at orthologous positions in other species. For this, we first evaluated the ratio of transcript positions between two species exhibiting a pCRMI for the same RBP. Expectedly, the comparison of human and mouse yielded the greatest number of common regions with the same motif instance for both coding and noncoding transcripts (Figure 5.3A). The count of comparable shared coordinates with a pCRMI for the same RBP was followed by the human and zebrafish pair, and, finally, the mouse and zebrafish pair. Moreover, there was 296 118 regions observed in all three species with a pCRMI for the same RBP (Figure 5.3A). Interestingly, when we established the ratio of these values compared to the total number of comparable genomic coordinates with a pCRMI but regardless of the type of equivalence (i.e., same RBP, different RBP, or no RBP), we observed that between 34 and 59% of the regions in coding transcript had a pCRMI for the same RBP (Figure 5.3B, solid columns). Notably, despite the contrasting number of shared coordinates between human and mouse and the other pairing of species, we observed similar conservation of pCRMI in all species compared. Furthermore, the region of the coding transcript (i.e., CDS or UTR) does not appear to have an impact on the conservation of putative RBP-binding sites. We observed a similar trend when analyzing noncoding RNA, where miRNA, scaRNA, and sRNA each exhibited ratios of inter-species conservation above 50% in all three pairwise species comparisons (Figure 5.3B). These ratios, both for coding and noncoding transcripts, are significantly higher than what we would expect at random (Figure 5.3B, unfilled columns). Overall, this suggests that the location of an RBP binding site within a transcript is important and well conserved.

We then evaluated the ratio of regions exhibiting a pCRMI for two different RBPs. The number of regions in pairwise species comparisons and their order were similar to what was observed for coding transcripts with human and mouse sharing the most common coordinates bearing different motif instances (Figure 5.3C). Most of these RBPs interchanges are likely explained by either false positives or motif resemblances. For example, the *SRSF9* to *SRSF1* exchange most observed in all pairwise species comparisons for coding transcripts can likely

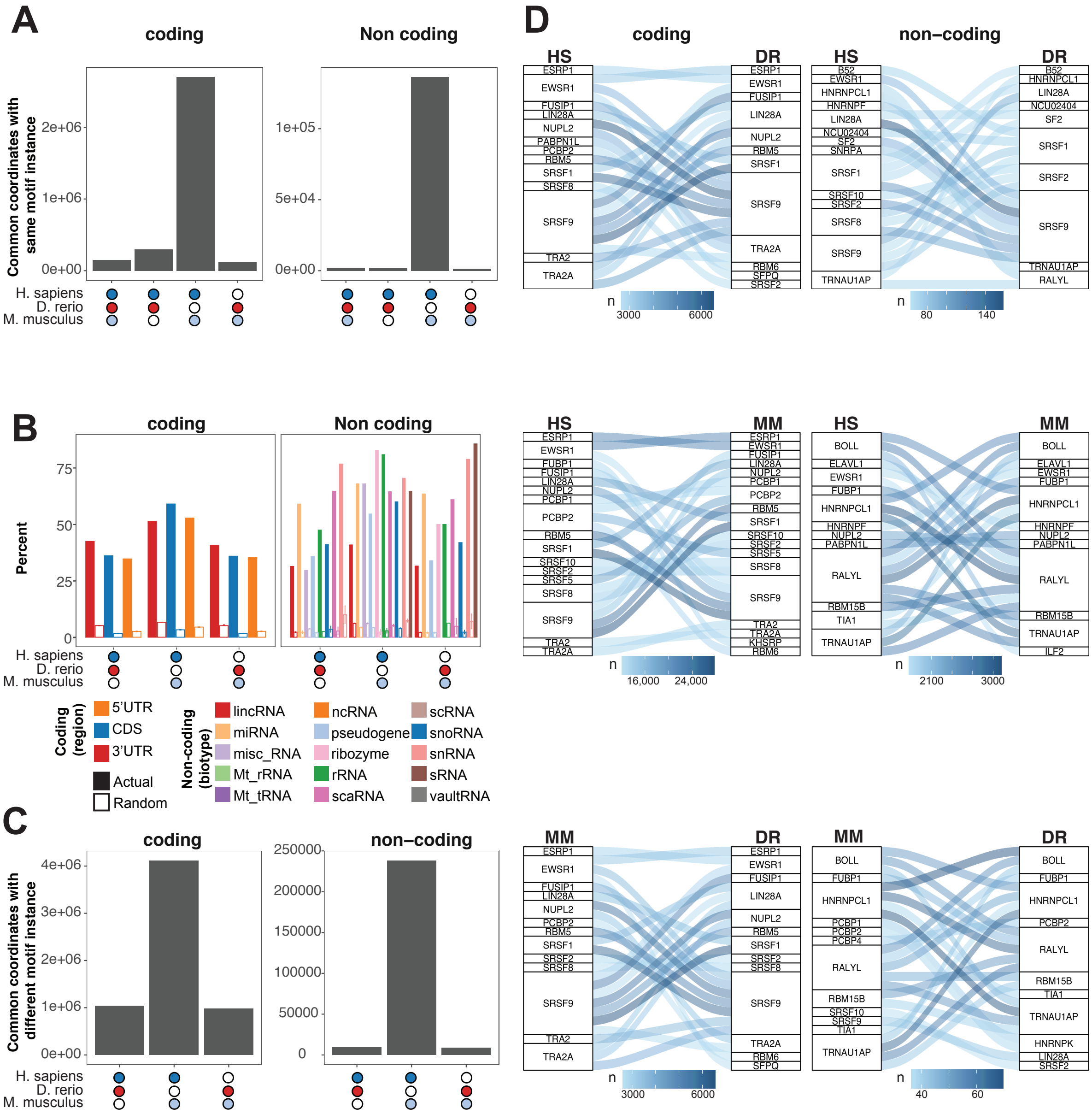


Figure 5.3. Interspecies comparable genomic coordinates often exhibit the conservation of the same RBP motif instance. A) Histogram showing the total number of comparable genomic coordinates with the same instance of a given RBP motif between all three species or pairwise by species for coding transcripts (left panel) and noncoding transcripts (right panel). **B)** Histogram showing the interspecies repartition, in percent, of their distinct comparable genomic coordinates with an instance of the same motif from all distinct comparable genomic region with a motif (i.e., same motif, different motif , or none) in at least one of the two species compared for coding transcript (left panel) and noncoding transcript (right panel). **C)** Histogram showing the total number of comparable genomic coordinates with instances of different RBP motif for each species pairwise comparison for coding transcripts (left panel) and noncoding transcripts (right panel). **D)** Alluvial plot showing the 25 most occurring changes of RBP motif between comparable genomic coordinate for each species pairwise comparison for coding transcripts (left panel) and noncoding transcripts (right panel). HS : *Homo sapiens*, CE : *Caenorhabditis elegans*, DM : *Drosophila melanogaster*, DR : *Danio rerio*, MM : *Mus Musculus*.

be explained by the AA to GG nucleotide substitution at the beginning of the canonical motif that ends in GGAG for both RBPs (Figure 5.3D). Indeed, a comparison with Tomtom results in a similarity probability with a p-value of 0.00965 [16]. However, the RBP *Fusip1*, labeled as *H. sapiens* only, is often interchanged with *SRSF9* in *D. rerio* and *M. musculus*, in coding transcripts, even though their motifs are appreciably different (Tomtom p-value = 0.25) (Figure 5.3D). Interestingly, both RBPs are members of the serine/arginine (SR)-rich family and have role as pre-mRNA splicing factors where they both, uncharacteristically of the SR family of protein, act as repressors [1, 12]. This suggests that one RBP can take over the role of another by binding to the same transcript location and preserve its function.

5. The regional topology of a motif is generally limited to a few instances with some lncRNA exceptions

To further characterize the distribution and interspecies relation of RBP motif pCRMI we calculated the ratio of transcripts that bear a binding site for each RBP within distinct regions or biotypes. Predictably, distinct mRNA and noncoding RNA biotypes do not carry a binding site for every given RBP (Figure 5.4A, 5.5A). Interestingly, for protein coding transcripts, once a biding site appears to be present in the 3'UTR or the CDS, it is most often present in 2 to 10 copies, with on average 21.7% and 27.7% of transcripts, respectively (Figure 5.4A). In comparison, 14.1% and 17.6% of transcripts only have 1 instance of a given motif, for 3'UTR and CDS respectively, and 7.5% and 4.5% of transcripts have more than 10 motifs. Inversely, the 5'UTR not only exhibits less transcripts with an RBP binding site, but when a transcript bears a binding site, it rarely contains more than one. Indeed, possibly due to their average shorter length, 10.7% of transcripts have a single instance in their 5'UTR while 9% have 2 to 10 copies and 1% have more than 10.

Likewise, we aimed to assess if some RBPs exhibit a more refined pCRMI topology. Therefore, we scanned our repertoire of pCRMI by windows of 100 nucleotides and tabulated the count of different groupings of pCRMI for the same RBP (i.e., considering all possible motifs for a given RBP). Importantly, to avoid finding multiple hits in the same sequence, especially in the case of motifs with a small repeat element, we only counted pairs of sites that are at a start position of more than seven nucleotides. Therefore, two motif instances cannot be overlapping. As a comparison, we repeated the exercise ten times with the same

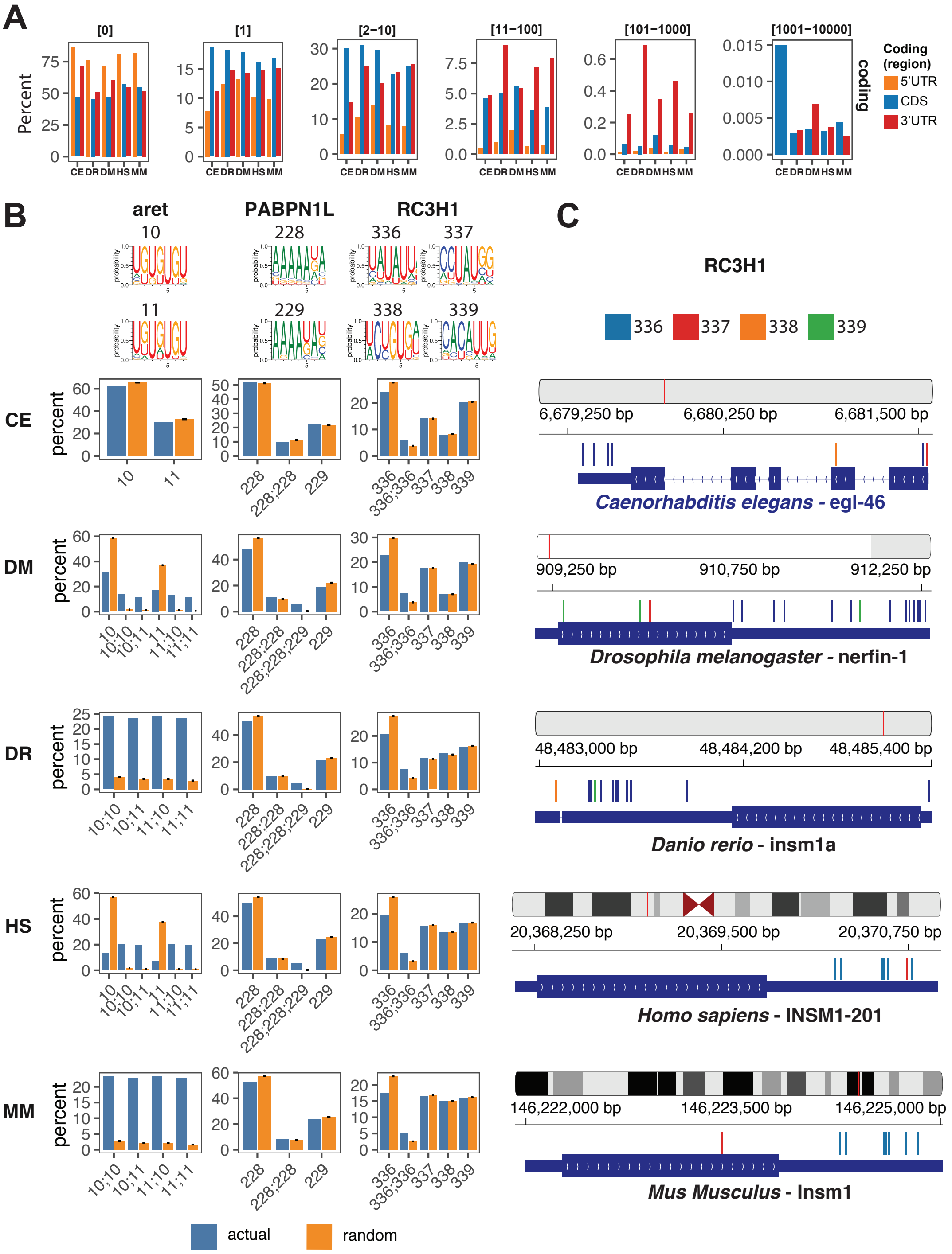


FIGURE 5.4. Coding transcripts often exhibit multiple instances of the same motif. **A)** Histogram showing the combined average percent of coding transcripts bearing RBP motif instances binned by the number of instances they contain (displayed top of each panel) separated by the region in which they are observed (coding or UTR). **B)** Histogram showing the percent of motif instances observed according to different topologies in a window of 100 nucleotides and compared with the average of 10 random permutations of motifs instances. Error bar indicates standard deviation between the 10 replicates. Repeated motifs within a window are denoted in a 5'-3' order with a semicolon separating the motif codes, a unique motif code indicating a single instance of the motif observed in the window. Motif code nomenclature defined as in Benoit Bouvrette *et al.* [3] with corresponding motif shown on top. **C)** Integrated genome browser screenshot showing the binding sites of each motif for the RBP RC3H1 in the 5 orthologs of the human transcript *INSM1-201*. HS : *Homo sapiens*, CE : *Caenorhabditis elegans*, DM : *Drosophila melanogaster*, DR : *Danio rerio*, MM : *Mus Musculus*.

set of genomic coordinates but with permuted motifs, hence keeping the same ratios and distances. Several RBPs displayed an enrichment of specific combinations (Figure 5.4B). For example, the pCRMI derived from the UG-rich motif of the RBP *aret* appears in pairs up to 5-fold more often than it would be expected stochastically (Figure 5.4B). This enrichment is observed in 4 out of 5 species studied, albeit to different ratios, and is only excluded from *C. elegans* (Figure 5.4B). Moreover, some pCRMI exhibit even more complex topology. The pCRMI of the A-rich motif of the RBP *PABPN1L* is arranged in triads in 5% of all their observations, while such a topology is expected to only be present less in than 0.3% of their instances.

Interestingly, the paired spatiality of the AU-rich motif of *RC3H1* is observed to be enriched by a 2-fold factor over what would be expected randomly (Figure 5.4B). Indeed, between 5.1% (*M. musculus*) and 7.5% (*D. rerio*) of *RC3H1* instances are observed in pairs in a 100-nucleotide window, compared with 2.5% and 4.2% stochastically (Figure 5.4B). The RBP *RC3H1* is known to bind a constitutive decay element (CDE) which folds in a short 17-nucleotide stem-loop and is present in the 3'UTR, leading to their degradation [**24**]. Notably, we observed that the pCRMI of *RC3H1* were present in pairs in the 3'UTR of the *H. sapiens* gene *INSM1-201*, a transcription factor involved in early embryonic neurogenesis, and all of its four orthologs, *Insm1*, *insm1a*, *nerfin-1*, and *egl-46*, as established by ensembl, in the studied species (Figure 5.4C) [**56, 55, 23**]. Furthermore, all five orthologs displayed one instance where the two pCRMI are always distanced exactly 13 nucleotides apart (Figure 5.4C). This suggests that in addition to the secondary structure necessary for the binding of *RC3H1*, primary sequence might play a role in RBP recognition and that this topology appears evolutionarily conserved between these orthologs.

Similar patterns of pCRMI content is observed for noncoding RNA where very few transcripts, with the notable exception of lncRNA and pseudogenes, bear more than 2 to 10 pCRMI, likely due to their average shorter length (Figure 5.5A, Supplemental figure 5.S5). Interestingly, in *H. sapiens*, one lncRNA, named *HELLPAR* (length of 205 KB), bears more than 1000 pCRMI for 42 RBPs (37 of which are species-specific) (Figure 5.5B, 5.5C). When comparing the pCRMI density observed for each RBP in *HELLPAR* with the ratio observed in the complete transcriptome, in motif instances per KB of transcript, we observed that these highly abundant RBPs are enriched on average 10-fold in *HELLPAR* (Figure 5.5B).

HELLPAR is a lncRNA associated with the Hellp syndrome, a rare haematological disease involving elevated liver enzymes and low platelets in pregnancy [49, 37]. The knockdown of *HELLPAR* has been associated with the upregulation of over 1000 genes [49]. Interestingly, the biological function GO term associated with the 42 RBPs with pCRMI in *HELLPAR* are 3'-UTR-mediated mRNA stabilization and 3'-UTR-mediated mRNA destabilization (FDR 2.76-07 and 2.15-05 respectively). While there is a possibility that most of these pCRMI are false positives, the abundance of pCRMI of RBPs with similar function suggests a potential role as an RBP sponge for this lncRNA.

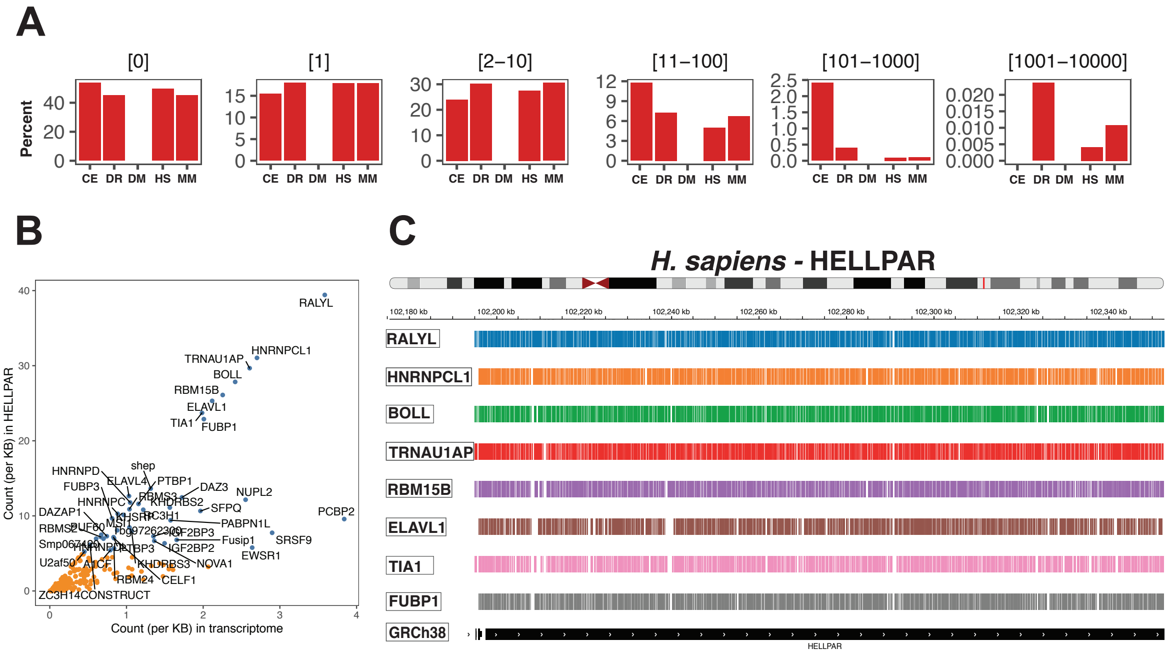


FIGURE 5.5. LncRNA *HELLPAR* is highly enriched for multiple motif instances. **A)** Histogram showing the combined average percent of lncRNA bearing RBP motif instances binned by the number of instances they contain (displayed top of each panel). **B)** Scatter plot showing the number of instances observed for each RBP, in count per KB, in the lncRNA *HELLPAR* compared with the number of instances, in count per KB, observed in the complete transcriptome. RBP with over 1000 motifs instances observed in *HELLPAR* are labelled. **C)** Integrated genome browser screenshot showing the binding sites of the 8 most abundant motifs human lncRNA *HELLPAR*. HS : *Homo sapiens*, CE : *Caenorhabditis elegans*, DM : *Drosophila melanogaster*, DR : *Danio rerio*, MM : *Mus Musculus*.

6. Discussion

In the last few decades, RNA-protein interaction has emerged as an important mediator in post-transcriptional gene regulation, impacting nearly all biological processes. However, the broad scope of potential RNA targets and specific correlational features of binding sites for a wide array of RBPs at a multi-species transcriptome-wide scale remained unknown. To address this question, we utilized herein a computational pipeline and database to probe the putative binding position of 223 RBPs in the full set of non-intronic coding and noncoding transcripts across five species. This global profiling approach, combined with precise comparative analyses, shows that RBPs exhibit broad variabilities in raw number of putative instances, transcript distribution and preferred binding structure where each arises on a gradual continuum and in patterns that appear, at diverse degrees, evolutionarily conserved whether considering mRNAs or noncoding transcripts.

Profiling of RBPs associated with specific RNAs has been used in several studies, providing unique insights into their mutual specialization. Highly reliable methods for identifying the positions of endogenous RNA–protein interactions such as RNP/RNA immunoprecipitation (RIP), cross-linking and immunoprecipitation (CLIP) and their many variations have recently been exploited to individually explore a wide range of proteins involved in a plethora of processes [**50**, **25**, **31**, **48**, **51**, **52**, **46**]. Even with progress and optimization now allowing for high-throughput studies, these are still painstaking and lengthy experiments often limited by cell lines and tissue specificities or the lack of good antibodies. Other *in vitro* and *in silico* studies have been designed to palliate these limits, however, this usually comes at the expense of detailed resolution. For example, RNAcompete and RBNS, when taken alone, are powerful methods to assess the precise binding motifs of an RBP, but does not conserve environmental context [**40**, **39**, **10**, **22**, **38**]. On the other hand, sourcing from the public databases of the results gathered from *in vitro* experiments, many computational frameworks of all kinds have been developed with the aim of providing a better portrayal of the RBP binding landscape [**29**, **17**, **33**, **34**, **14**]. However, the studies describing these computational pipelines always appear to stop short of utilizing them to systematically describe the transcriptome-wide features of every available RBP. Here, we harnessed the power of robust computational prediction based on motifs data coming from *in vitro* biological experiments and by further validating our results in human through comparisons with state-of-the-art *in vivo* large-scale experimental methods. While previous large-scale profiling of RBP binding sites has shown their individual landscape in one species, the present study offers one of the most comprehensive surveys of putative RBP binding sites and revealed that many RBP motif instances, and their clusters, are likely distributed across species within conserved motif environments. Furthermore, the relatively function-unbiased selection of 223 RBPs jointly profiled across all RNA biotypes demonstrated the power of our approach to expose potential new functions for RBPs or transcripts of interest.

We found that putative RBP motif instances were vastly and unevenly distributed across each transcript with some RBPs showing near-ubiquitous binding sites to some being scarcely observed. Intriguingly, the presence of these binding sites did not readily reflect the actual expression of its RBP in a given species, and while this may be a possible artefact of the likely non-negligible false positives inherent to our approach, we hypothesize that it may be an indicator of short evolutionary distances between the studied species. Moreover, we noticed a relatively well conserved structural binding preference for motifs with specific nucleotide enrichments, where G-rich motifs were observed more often in paired regions, while A-rich motifs were observed more in single-stranded regions, albeit to a less stringent degree. While this may simply be reflective of the fundamentals of RNA structure where GC-rich regions have more stable secondary structure, this may also be an indicator of the role played by structure in the binding recognition of these RBPs where the combination of sequence and structure, or even structure alone, can be the determining binding recognition factor. Interestingly, the RBD of the protein did not have an immediate correlation with neither the sequence nor the structure of the putative binding site. It will be interesting to integrate non-predictive structural data from novel *in vitro* binding assays, such as RNAcompete-S, as they become more available, to further refine the true combination of sequence/structure bound by each RBP [8].

Our results also provided significant insights relating to the topologies of RBP binding sites. Indeed, when comparing corresponding mRNA genome coordinates between two species we found that a large number, up to 59%, of these regions had the potential to be bound by the same RBP. This was even more striking for noncoding RNA biotypes with up to 86% of corresponding coordinates sharing a putative binding site for the same RBP. Interestingly, when the motif instances observed at a given region was different to the syntenic region of another species, the replacing RBP often has a similar role. This suggests that these regions are of high importance in the proper function of RNA-RBP interactions. We hypothesized that the presence, or absence, of such patterns between species is likely a reflection of the importance of their function. The importance of the positioning of the binding sites is even better reflected when studying the presence of repeated motif instances, often separated by equal intervals. We observed many occurrences where motif instances were positioned as pairs or triplets, topologies that were usually reflected in orthologs. While there are only a few orthologs shared between the five species studied, we were still capable of identifying this feature for the *RC3H1* RBP in the *INSM1* gene which has orthologs expressed in all species studied. Interestingly, the distance between two instances of *RC3H1* is similar to the size of the stem loop identified as the binding site. This suggests that a sequence/structure combination is important for its correct binding. This could be due to this RBP having multiple liaison sites or it may be combining with other *RC3H1* proteins to form a binding complex.

One important benefit of the unbiased approach, not concentrating on a specific RBP or RNA biotype, presented here is that it enables the identification of novel potential sites with hypothetical regulatory activity. As such our results also demonstrated that the lncRNA *HELLPAR*, one of the longest noncoding RNAs known with an annotated length over 205 KB, displays an enrichment, up to 10-fold over the expected ratio, for 42 RBPs. This intriguing observation makes this lncRNA a prime candidate as an RBP sponge that may influence an array of biological processes. As this lncRNA with perinuclear expression has been implicated in a severe disease, it would be interesting to study the mechanistic implications associated with the defect of this gene on an experimental level.

In conclusion, the computational methodology and derived analysis presented herein offers an efficient approach to comprehensively describe the position of RBP binding sites and their inherent distribution features. We provide examples on how cumulative analyses of RBP binding profiles across multiple species can yield novel insights about their interactions with both coding and noncoding transcripts, including the relative functional importance of evolutionary conserved patterns of clusters of RBPs and the role of RBP interactions with diverse RNA biotypes. As we further profile additional RBPs and species, we suggest that these emerging patterns and their relationships will allow for better functional and mechanistic prediction at high resolution. Moreover, we expect that the increasing identification of RBP features, at scale, will enable researchers to address hypotheses regarding post-transcriptional gene regulation, RNA processing mechanisms, and even conserved functional roles of groups of RBPs or transcripts.

7. Material and Methods

7.1. RBP motif acquisition and motif-finding algorithm

RBP motif instances datasets used were obtained from our oRNAment resource (http://rnabiology.ircm.qc.ca/oRNAment/). For a more detailed description of the algorithm for oRNAment see Benoit Bouvrette *et al.* [3]. Briefly, we retrieved the data for 223 unique RBPs in the form of position weight matrices (PWMs) from both RNAcompete and RBNS. Specifically, we downloaded 218, tallying 172 RBPs, RNAcompete PWMs, from CISBP-RNA and we derived an additional 235 PWMs, tallying 78 RBPs, by executing the RBNS computational analysis pipeline with parameters set for identifying 7-mer enrichment on the RBNS sequencing data available from the ENCODE resource. Our python algorithm, which establishes the correspondence of a transcript subsequence to a given RBP motif of the same length, takes as input a motif, in the form of a PWM, and returns the position of the subsequences above a given score. Therefore, all motifs in the database are 7 nucleotides in length and are, as such, comparable. Unless otherwise indicated, standardized motif instance calling for each RBP was used throughout this study. As such, we defined a threshold independently for each RBP with a rank score percentile of 50% over all possible scores.

7.2. Retrieving motif scores and transcriptomic position

All data was stored in a Yandex Clickhouse DBMS and all tallying of RBP and transcript repertoires of motif instances were performed using standard SQL statements.

7.3. Motif instances comparisons and correlations

Annotation data were retrieved from the Ensembl resource. Specifically, Ensembl GTF, Ensembl GFF files was downloaded from Ensembl and the ensembldb R library was downloaded from Bioconductor. Version GRCh38.98, WBcel235.98, BDGP6.22.98, GRCz11.98, GRCm38.98 for human, worm, fly, zebrafish, and mouse were respectively used for all annotations.

The combination of length annotations was compiled with R. All tSNE, pCRMI/eCLIP peaks correlations and pCRMI correlations between species with R.

7.4. GO terms statistical overrepresentation test

GO term enrichments for each set of clustered RBP were obtained with the gene ontology panther classification system by selecting for biological process with the default relevant species whole-genome reference lists as background.

7.5. Syntenic coordinates

Transcriptomic coordinated were translated to genomic coordinates with a python script by first creating a reference file for each species detailing chromosome, exon rank, the exon start base number and the exon end base number for each transcript isoform and converting them to a base 1 concatenated list of positions covering the whole exonic transcript. Secondly, each pCRMI transcript starting position was looked up into this table to establish its genomic coordinates. Validation was performed by manually loading example data to the UCSC genome browser and visually comparing each position.

Genomic coordinated between two species were established with the liftover tool from UCSC. The files hg38ToMm10.over.chain, hg38ToDanRer11.over.chain, and mm10ToDanRer11.over.chain were downloaded from UCSC. Coordinates were mapped for syntenic region by taking all genomic coordinates with a pCRMI from human and searching for the equivalent position in mouse or zebrafish. The mouse to zebrafish coordinated was established from the mouse pCRMI genomic coordinates. An R/SQL script was implemented and executed on the resulting files of syntenic coordinated to retrace species specific RBP pCRMI from the Clickhouse DBMS.

7.6. Motif PWM comparisons

Standard PWM files were manually converted to MEME Motif Format and motif were compared individually using pairwise analysis of query motif and target motif with the online version of the Tomtom motif comparison tool version 5.1.1.

8. Author's contributions

Conceptualization, L.P.B.B. ; Methodology, L.P.B.B. ; Investigation, L.P.B.B. ; Software, L.P.B.B. ; Formal Analysis, L.P.B.B. ; Visualization, L.P.B.B. ; Writing – Original Draft, L.P.B.B. ; Writing – Review and Editing, L.P.B.B., M.B., E.L. ; Funding Acquisition, E.L. ; Supervision, E.L. M.B.

9. Acknowledgements

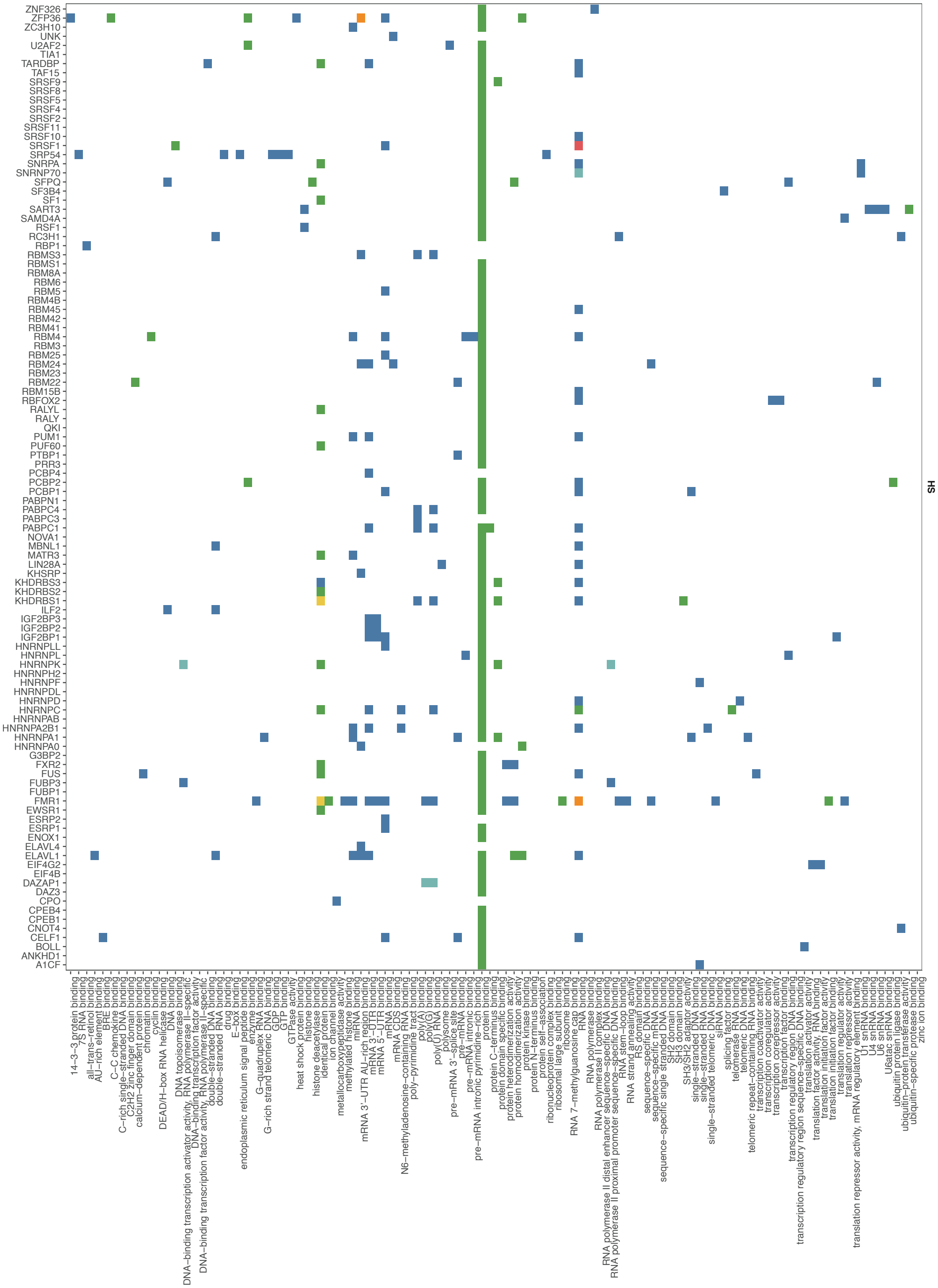
We wish to thank Eric Van Nostrand for kindly providing us eCLIP and peak distribution data. We also thank Samantha Bovaird and Eunjeong Kwon for their helpful comments and suggestions. This research was enabled in part by support provided Calcul Quebec (www.calculquebec.ca), Compute Ontario (www.computeontario.ca) and Compute Canada (www.computecanada.ca).

10. Supplemental Data

Table 5.S1. Summary of the overall distribution of transcripts bearing an RBP motif instance.

	H. sapiens	***C. elegans***	***D. rerio***	***D. melanogaster***	***M. musculus***
			Coding RNA		
Min	0.04539	0.04025	0.04752	0.05352	0.04104
1st Qu	0.68816	0.36300	0.57064	0.51896	0.66246
Median	0.82205	0.64811	0.74979	0.70529	0.81432
Mean	0.77635	0.60537	0.71182	0.68454	0.77199
3rd Qu	0.93493	0.84351	0.90164	0.90598	0.92886
Max	0.99114	0.98942	0.99083	0.99552	0.99086
			Noncoding RNA		
Min	0.006889	0.003817	0.009566	0.01394	0.006718
1st Qu	0.242290	0.048327	0.221075	0.23533	0.240121
Median	0.388485	0.108451	0.327052	0.37569	0.394261
Mean	0.421544	0.123469	0.342731	0.40500	0.420904
3rd Qu	0.595959	0.170522	0.482842	0.57630	0.607409
Max	0.873539	0.469080	0.812025	0.84951	0.878376

(a)


ZNF326
ZFP36
ZC3H10
UNK
U2AF2
TIA1
TARDBP
TAF15
SRSF9
SRSF8
SRSF5
SRSF4
SRSF2
SRSF11
SRSF10
SRSF1
SRP54
SNRPA
SNRNP70
SFPQ
SF3B4
SF1
SART3
SAMD4A
RSF1
RC3H1
RBP1
RBMS3
RBMS1
RBM8A
RBM6
RBM5
RBM4B
RBM45
RBM42
RBM41
RBM4
RBM3
RBM25
RBM24
RBM23
RBM22
RBM15B
RBFOX2
RALYL
RALY
QKI
PUM1
PUF60
PTBP1
PRR3
PCBP4
PCBP2
PCBP1
PABPN1
PABPC4
PABPC3
PABPC1
NOVA1
MBNL1
MATR3
LIN28A
KHSRP
KHDRBS3
KHDRBS2
KHDRBS1
ILF2
IGF2BP3
IGF2BP2
IGF2BP1
HNRNPLL
HNRNPL
HNRNPK
HNRNPH2
HNRNPF
HNRNPDL
HNRNPD
HNRNPC
HNRNPAB
HNRNPA2B1
HNRNPA1
HNRNPA0
G3BP2
FXR2
FUS
FUBP3
FUBP1
FMR1
EWSR1
ESRP2
ESRP1
ENOX1
ELAVL4
ELAVL1
EIF4G2
EIF4B
DAZAP1
DAZ3
CPO
CPEB4
CPEB1
CNOT4
CELF1
BOLL
ANKHD1
A1CF
14−3−3 protein binding
7S RNA binding
all−trans−retinol binding
AU−rich element binding
BRE binding
C−C chemokine binding
C−rich single−stranded DNA binding
C2H2 zinc finger domain binding
calcium−dependent protein binding
chromatin binding
cyclin binding
DEAD/H−box RNA helicase binding
DNA binding
DNA topoisomerase binding
DNA−binding transcription activator activity, RNA polymerase II−specific
DNA−binding transcription factor activity
DNA−binding transcription factor activity, RNA polymerase II−specific
double−stranded DNA binding
double−stranded RNA binding
drug binding
E−box binding
endoplasmic reticulum signal peptide binding
enzyme binding
G−quadruplex RNA binding
G−rich strand telomeric DNA binding
GTP binding
GTPase activity
heat shock protein binding
histone binding
histone deacetylase binding
identical protein binding
ion channel binding
metallocarboxypeptidase activity
methylated histone binding
miRNA binding
mRNA 3'−UTR AU−rich region binding
mRNA 3'−UTR binding
mRNA 5'−UTR binding
mRNA binding
mRNA CDS binding
N6−methyladenosine−containing RNA binding
poly−pyrimidine tract binding
poly(A) binding
poly(G) binding
poly(U) RNA binding
polysome binding
pre−mRNA 3'−splice site binding
pre−mRNA binding
pre−mRNA intronic binding
pre−mRNA intronic pyrimidine−rich binding
protein binding
protein C−terminus binding
protein domain specific binding
protein heterodimerization activity
protein homodimerization activity
protein kinase binding
protein N−terminus binding
protein self−association
ribonucleoprotein complex binding
ribosomal large subunit binding
ribosome binding
RNA 7−methylguanosine cap binding
RNA binding
RNA polymerase binding
RNA polymerase II complex binding
RNA polymerase II distal enhancer sequence−specific DNA binding
RNA polymerase II proximal promoter sequence−specific DNA binding
RNA stem−loop binding
RNA strand annealing activity
RS domain binding
sequence−specific DNA binding
sequence−specific mRNA binding
sequence−specific single stranded DNA binding
SH2 domain binding
SH3 domain binding
SH3/SH2 adaptor activity
single−stranded DNA binding
single−stranded RNA binding
single−stranded telomeric DNA binding
siRNA binding
splicing factor binding
telomerase RNA binding
telomeric DNA binding
telomeric repeat−containing RNA binding
transcription coactivator activity
transcription coregulator activity
transcription corepressor activity
transcription factor binding
transcription regulatory region DNA binding
transcription regulatory region sequence−specific DNA binding
translation activator activity
translation factor activity, RNA binding
translation initiation factor binding
translation regulator activity
translation repressor activity
translation repressor activity, mRNA regulatory element binding
U1 snRNA binding
U4 snRNA binding
U6 snRNA binding
U6atac snRNA binding
ubiquitin protein ligase binding
ubiquitin−protein transferase activity
ubiquitin−specific protease binding
zinc ion binding
HS

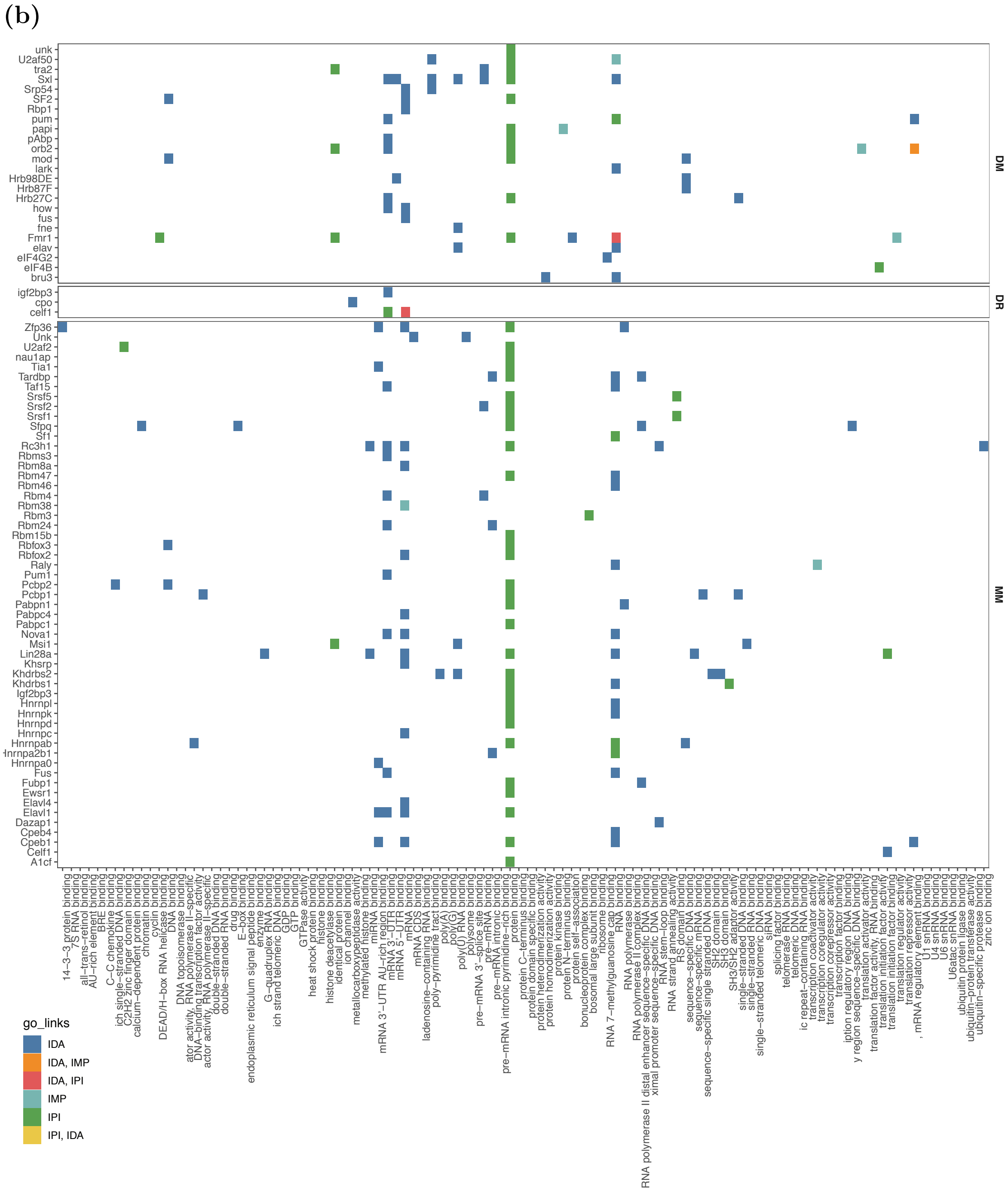


Figure 5.S1. RBPs studied are implicated in a wide range of biological processes. Binary heatmap of the biological process gene ontology annotation, with their evidence code, for each RBP considered in this study grouped by the species they were derived for RNAcompete and RBNS experiments. **(a)** *Homo sapiens* (HS) **(b)** *Drosophila melanogaster* (DM), *Danio rerio* (DR), *Mus Musculus* (MM). Only direct evidence code (go links) were considered. IDA : Inferred from Direct Assay ; IMP : Inferred from Mutant Phenotype ; IPI : Inferred from Physical Interaction.

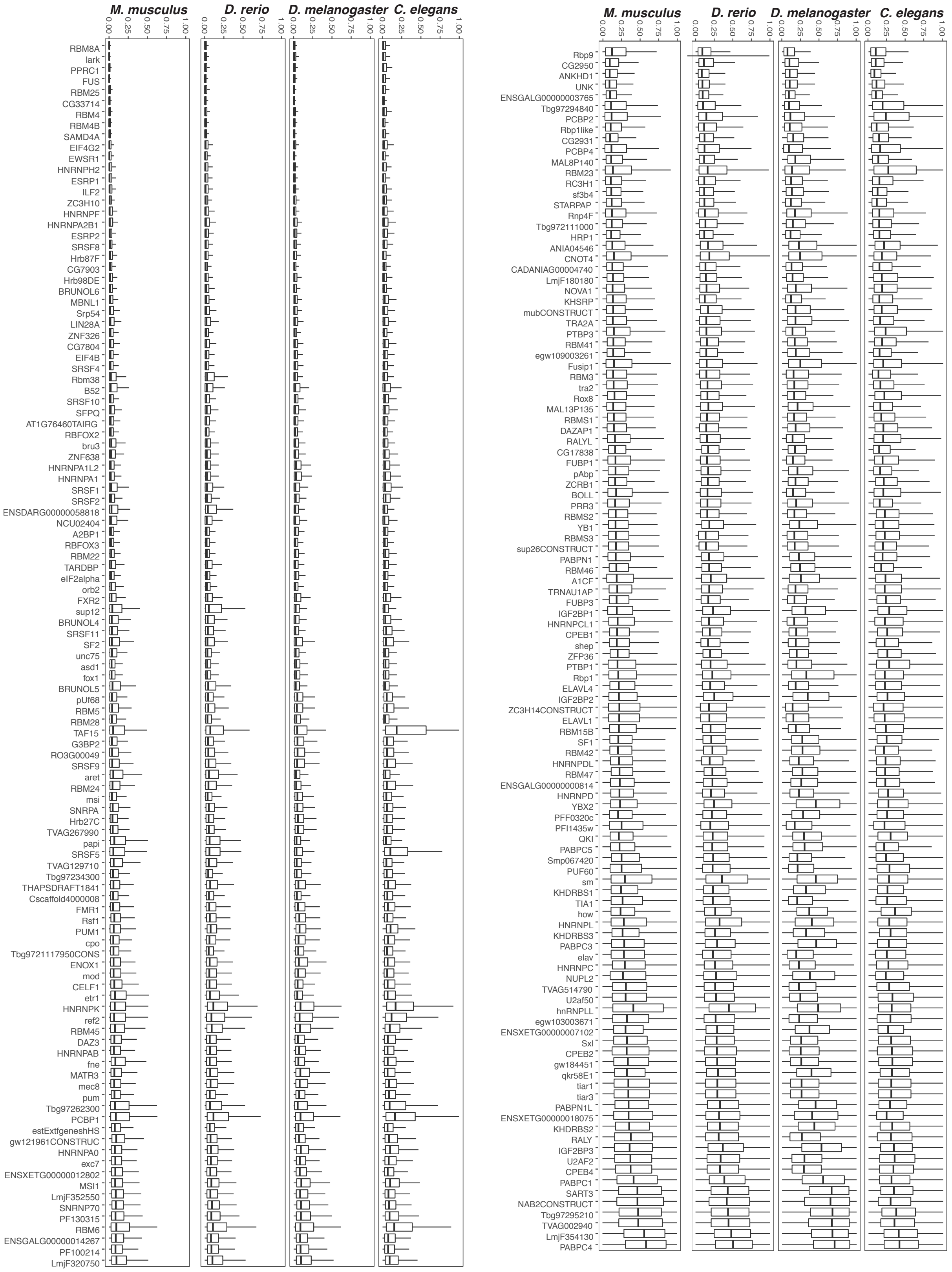
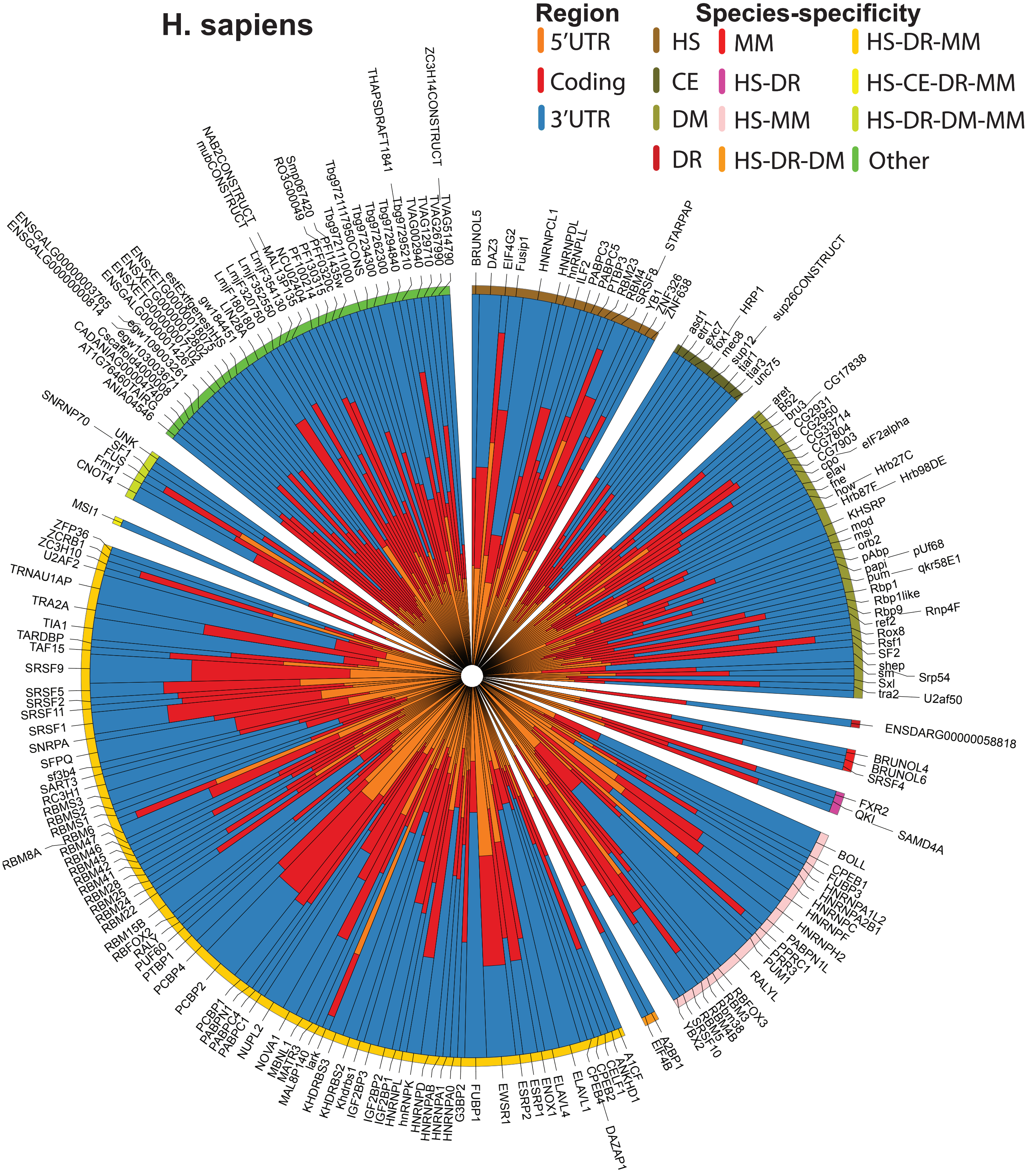
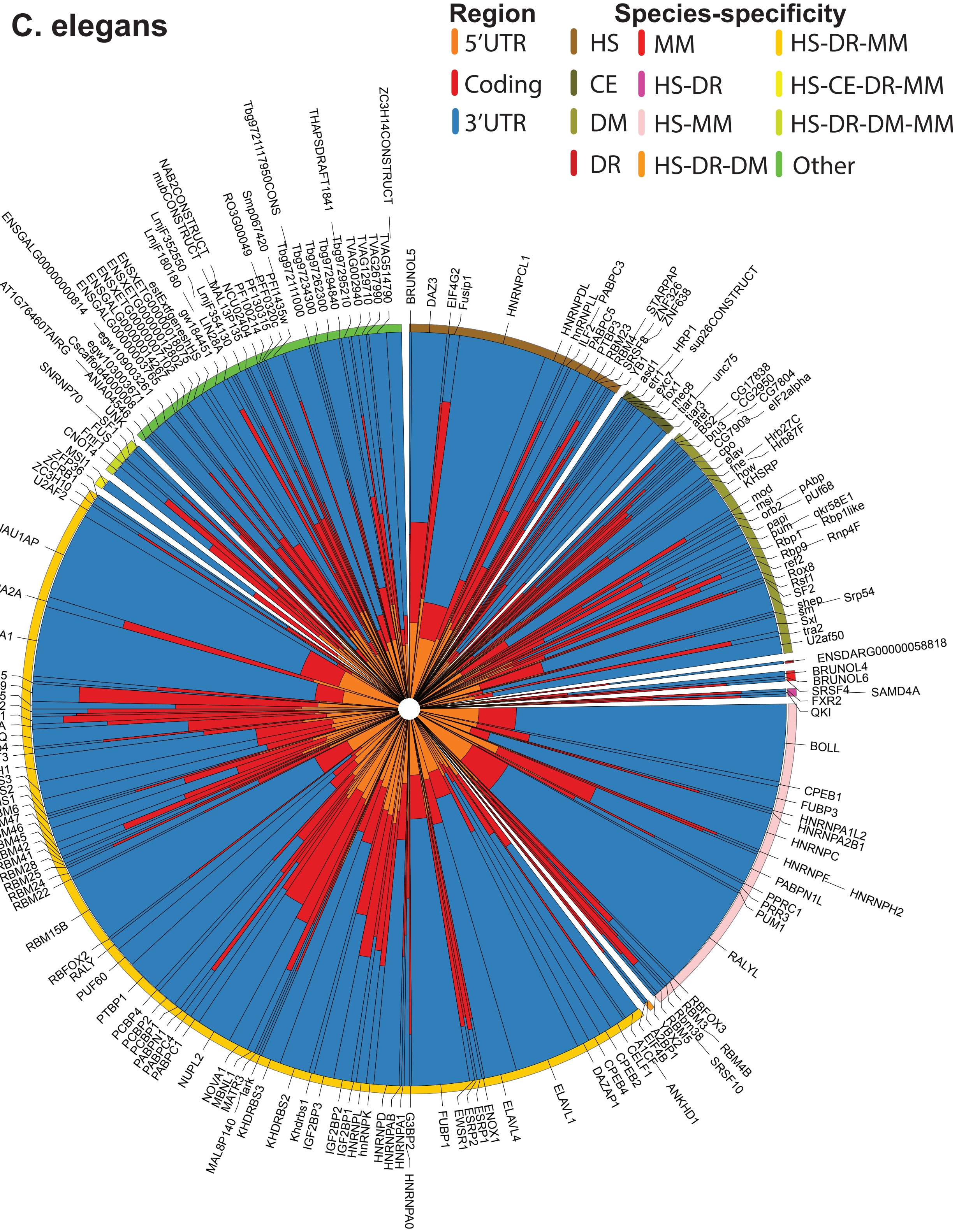


Figure 5.S2. Overview of the structural context of RBP motif instances. Box plot showing the predicted percent of unpaired probability, determined by RNAplfold, of the 7-mer region bearing an RBP motif instance, ordered by the median observed in *H. sapiens*.

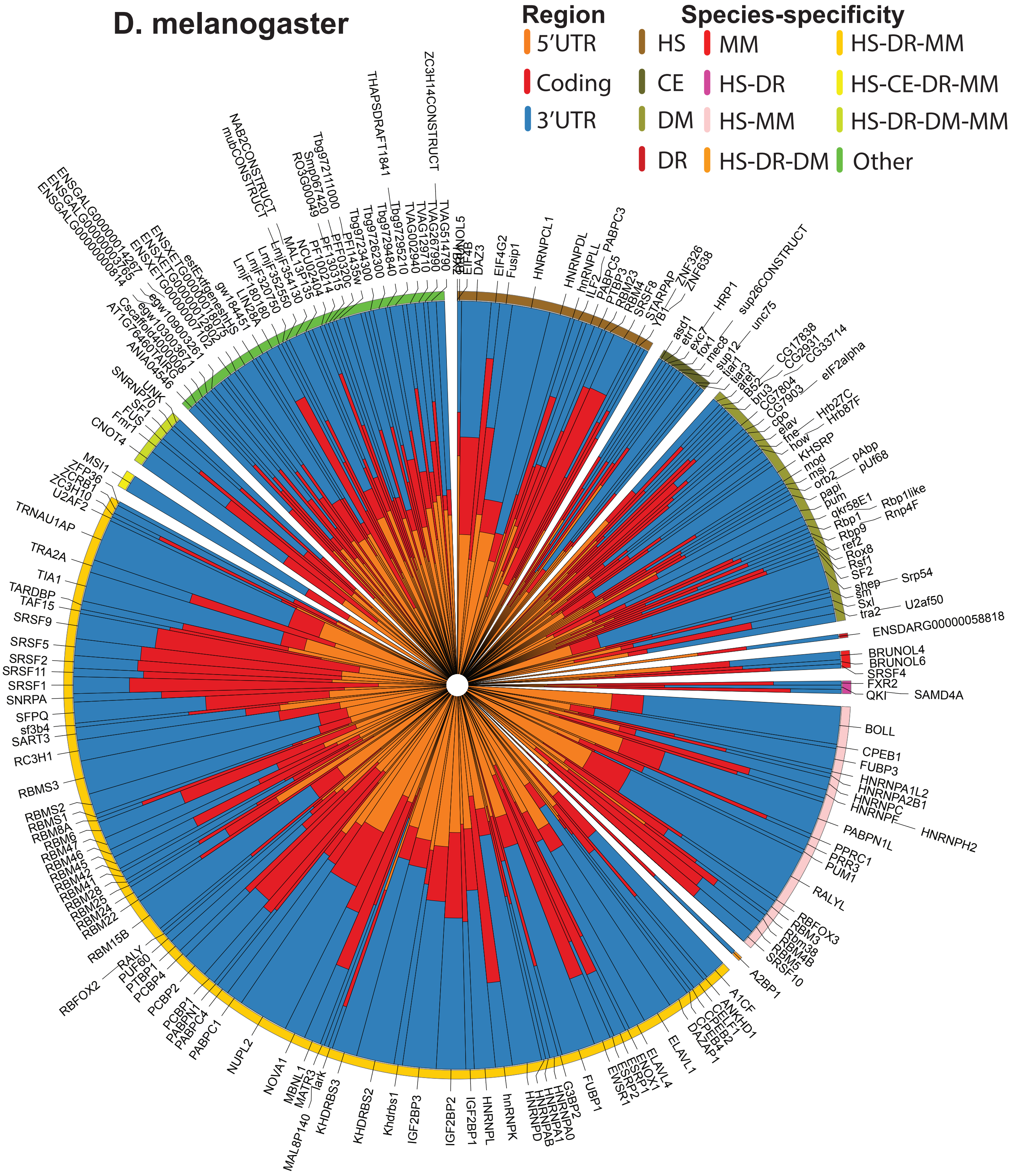
(a)


H. sapiens
Region
5′UTR
Coding
3′UTR
Species-specificity
HS
CE
DM
DR
MM
HS-DR
HS-MM
HS-DR-DM
HS-DR-MM
HS-CE-DR-MM
HS-DR-DM-MM
Other
ZC3H14CONSTRUCT
THAPSDRAFT1841
TVAG514790
TVAG267990
TVAG129710
TVAG002940
Tbg972.95210
Tbg972.94840
Tbg972.62300
Tbg972.34300
Tbg972.11950CONS
Tbg972.11.1000
PFF1435w
PFF0320c
PFL0213c
PF100204
NCU02404
MAL13P1.135
LmjF35.4130
LmjF35.2550
LmjF32.0750
LmjF18.0180
LIN28A
gw182451
estExt_fgenesh1_
Smp067420
RO3G00049
NAB2CONSTRUCT
mubCONSTRUCT
ENSXETG00000018075
ENSXETG00000017102
ENSXETG00000012803
ENSGALG00000014236
ENSGALG00000003765
ENSGALG00000000814
egw10903261
egw10303671
Cscaffold40000008
CADANIAG00004740
AT1G76460TAIR9
ANIA04546
SNRNP70
UNK
SF1
FUS
Fmr1
CNOT4
MSI1
ZFP36
ZCRB1
ZC3H10
U2AF2
TRNAU1AP
TRA2A
TIA1
TARDBP
TAF15
SRSF9
SRSF5
SRSF2
SRSF11
SRSF1
SNRPA
SFPQ
sf3b4
SART3
RC3H1
RBMS3
RBMS2
RBMS1
RBM6
RBM8A
RBM47
RBM46
RBM45
RBM42
RBM41
RBM28
RBM25
RBM24
RBM22
RBM15B
RBFOX2
RALY
PUF60
PTBP1
PCBP4
PCBP2
PCBP1
PABPN1
PABPC4
PABPC1
NUPL2
NOVA1
MBNL1
MATR3
MAL8P1.40
lark
KHDRBS3
KHDRBS2
Khdrbs1
IGF2BP3
IGF2BP2
IGF2BP1
HNRNPL
hnRNPK
HNRNPD
HNRNPAB
HNRNPA1
HNRNPA0
G3BP2
FUBP1
EWSR1
ESRP2
ESRP1
ENOX1
ELAVL4
ELAVL1
DAZAP1
CPEB4
CPEB2
CELF1
ANKHD1
A1CF
A2BP1
EIF4B
YBX2
SRSF7
SRSF10
RBM5
RBM4B
RBM38
RBM3
RBFOX3
RALYL
PUM1
PPRC1
PCBP3
PABPN1L
HNRNPH2
HNRNPF
HNRNPC
HNRNPA2B1
HNRNPA1L2
FUBP3
CPEB1
BOLL
SAMD4A
QKI
FXR2
SRSF4
BRUNOL6
BRUNOL4
ENSDARG00000058818
U2af50
tra2
Sxl
sm
Srp54
shep
SF2
Rsf1
Rox8
ref2
Rnp4F
Rbp9
Rbp1like
Rbp1
pum
qkr58E1
papi
pUf68
pAbp
orb2
msi
mod
KHSRP
Hrb98DE
Hrb87F
how
Hrb27C
fne
elav
eIF2alpha
cpo
CG7903
CG7804
CG33714
CG2950
CG2931
CG17838
bru3
B52
aret
unc75
tiar3
tiar1
sup26CONSTRUCT
sup12
mec8
HRP1
fox1
exc7
etr1
asd1
ZNF638
ZNF326
YB1
STARPAP
SRSF8
RBM4
RBM23
PTBP3
PABPC5
PABPC3
ILF2
HNRNPLL
HNRNPDL
HNRNPCL1
Fusip1
EIF4G2
DAZ3
BRUNOL5

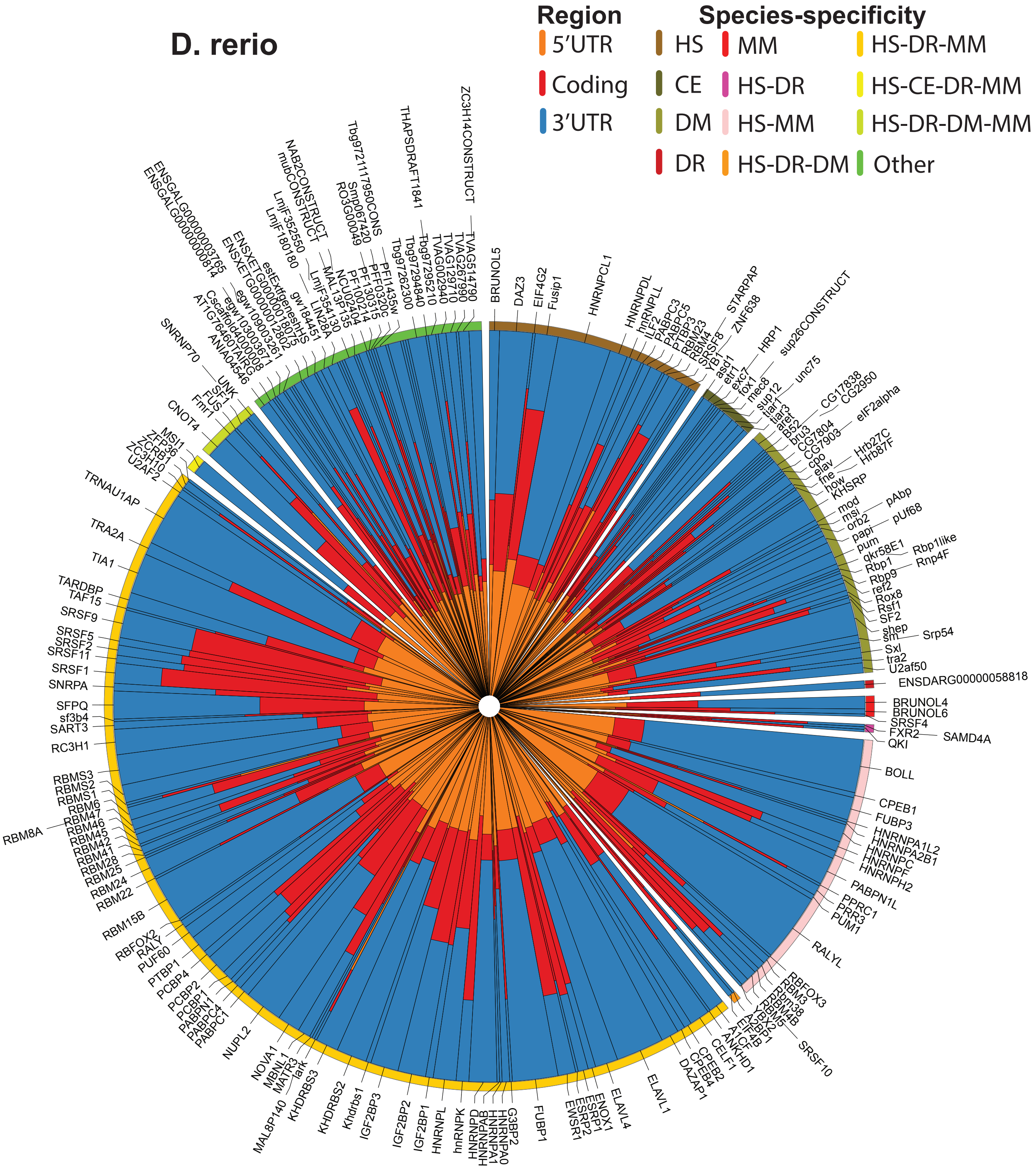
(b)


C. elegans
Region
5′UTR
Coding
3′UTR
Species-specificity
HS
CE
DM
DR
MM
HS-DR
HS-MM
HS-DR-DM
HS-DR-MM
HS-CE-DR-MM
HS-DR-DM-MM
Other
BRUNOL5
DAZ3
EIF4G2
Fusip1
HNRNPCL1
HNRNPDL
hnRNPLL
ILF2
PABPC3
PABPC5
PTBP3
RBM23
RBM4
SRSF8
STARPAP
ZNF326
ZNF638
YB1
asd1
HRP1
sup26CONSTRUCT
etr1
exc7
fox1
mec8
unc75
tiar1
tiar3
aret
B52
CG17838
CG2950
CG7804
eIF2alpha
bru3
CG7903
cpo
elav
Hrb27C
Hrb87F
fne
how
KHSRP
mod
msi
pAbp
pUf68
orb2
papi
pum
qkr58E1
Rbp1like
Rbp1
Rbp9
Rnp4F
ref2
Rox8
Rsf1
SF2
shep
Srp54
sm
Sxl
tra2
U2af50
ENSDARG00000058818
BRUNOL4
BRUNOL6
SRSF4
SAMD4A
FXR2
QKI
BOLL
CPEB1
FUBP3
HNRNPA1L2
HNRNPA2B1
HNRNPC
HNRNPF
HNRNPH2
PABPN1L
PPRC1
PRR3
PUM1
RALYL
RBFOX3
RBM3
RBM4B
Rbm38
SRSF10
RBM5
YBX2
A2BP1
EIF4B
A1CF
ANKHD1
CELF1
CPEB2
CPEB4
DAZAP1
ELAVL1
ELAVL4
ENOX1
ESRP1
ESRP2
EWSR1
FUBP1
G3BP2
HNRNPA0
HNRNPA1
HNRNPAB
HNRNPD
hnRNPK
HNRNPL
IGF2BP1
IGF2BP2
IGF2BP3
Khdrbs1
KHDRBS2
KHDRBS3
lark
MAL8P140
MATR3
MBNL1
NOVA1
NUPL2
PABPC1
PABPC4
PABPN1
PCBP1
PCBP2
PCBP4
PTBP1
PUF60
RALY
RBFOX2
RBM15B
RBM22
RBM24
RBM25
RBM28
RBM41
RBM42
RBM45
RBM46
RBM8A
RBM47
RBM6
RBMS1
RBMS2
RBMS3
RC3H1
SART3
sf3b4
SFPQ
SNRPA
SRSF11
SRSF1
SRSF2
SRSF5
SRSF9
TARDBP
TAF15
TIA1
TRA2A
TRNAU1AP
U2AF2
ZC3H10
ZCRB1
ZFP36
MSI1
CNOT4
Fmr1
FUS
SF1
SNRNP70
UNK
ANIA04546
AT1G76460TAIRG
Cscaffold4000008
egw103003671
egw109003261
ENSGALG00000000814
ENSGALG00000003765
ENSGALG00000014267
ENSXETG00000007102
ENSXETG00000012802
ENSXETG00000018075
estExtfgeneshHS
gw184451
LmjF180180
LmjF352550
LIN28A
LmjF354130
mubCONSTRUCT
NAB2CONSTRUCT
MAL13P135
NCU02404
PF100214
PF130315
RO3G00049
PFF0320c
Smp067420
PFI1435w
Tbg972111000
Tbg972117950CONS
Tbg972234300
Tbg972262300
Tbg972294840
THAPSDRAFT1841
Tbg972295210
TVAG002940
TVAG129710
TVAG267990
TVAG514790
ZC3H14CONSTRUCT

(c)


D. melanogaster
Region
5′UTR
Coding
3′UTR
Species-specificity
HS
CE
DM
DR
MM
HS-DR
HS-MM
HS-DR-DM
HS-DR-MM
HS-CE-DR-MM
HS-DR-DM-MM
Other
YBX2
BRUNOL5
EIF4B
DAZ3
EIF4G2
Fusip1
HNRNPCL1
HNRNPDL
hnRNPLL
ILF2
PABPC5
PABPC3
PTBP3
RBM23
RBM4
SRSF8
STARPAP
YB1
ZNF326
ZNF638
asd1
etr1
exc7
fox1
HRP1
mec8
sup26CONSTRUCT
sup12
unc75
tiar1
tiar3
aret
B52
CG17838
CG2931
CG33714
bru3
CG7804
CG7903
eIF2alpha
cpo
elav
fne
Hrb27C
Hrb87F
how
KHSRP
mod
msi
pAbp
pUf68
orb2
papi
pum
qkr58E1
Rbp1like
Rbp1
Rnp4F
Rbp9
ref2
Rox8
Rsf1
SF2
shep
Srp54
sm
Sxl
tra2
U2af50
ENSDARG00000058818
BRUNOL4
BRUNOL6
SRSF4
FXR2
QKI
SAMD4A
BOLL
CPEB1
FUBP3
HNRNPA1L2
HNRNPA2B1
HNRNPC
HNRNPF
HNRNPH2
PABPN1L
PPRC1
PRR3
PUM1
RALYL
RBFOX3
RBM3
Rbm38
RBM4B
RBM5
SRSF10
A2BP1
A1CF
ANKHD1
CELF1
CPEB2
CPEB4
DAZAP1
ELAVL1
ELAVL4
ENOX1
ESRP1
ESRP2
EWSR1
FUBP1
G3BP2
HNRNPA0
HNRNPA1
HNRNPAB
HNRNPD
hnRNPK
HNRNPL
IGF2BP1
IGF2BP2
IGF2BP3
Khdrbs1
KHDRBS2
KHDRBS3
lark
MATR3
MAL8P140
MBNL1
NOVA1
NUPL2
PABPC1
PABPC4
PABPN1
PCBP1
PCBP2
PCBP4
PTBP1
PUF60
RALY
RBFOX2
RBM15B
RBM22
RBM24
RBM25
RBM28
RBM41
RBM42
RBM45
RBM46
RBM47
RBM6
RBM8A
RBMS1
RBMS2
RBMS3
RC3H1
SART3
sf3b4
SFPQ
SNRPA
SRSF1
SRSF11
SRSF2
SRSF5
SRSF9
TAF15
TARDBP
TIA1
TRA2A
TRNAU1AP
U2AF2
ZC3H10
ZCRB1
ZFP36
MSI1
CNOT4
Fmr1
FUS
SF1
SNRNP70
UNK
ANIA04546
AT1G76460TAIRG
Cscaffold40000008
egw1030003671
egw1090003261
ENSXETG00000007102
ENSXETG00000012802
ENSXETG00000018075
estExtfgeneshHS
gw184451
ENSGALG00000000814
ENSGALG00000003765
ENSGALG00000014267
LIN28A
LmjF180180
LmjF320750
LmjF352550
LmjF354130
MAL13P135
NCU02404
PF100214
PF130315
PFF0320c
PFI1435w
NAB2CONSTRUCT
mubCONSTRUCT
RO3G00049
Smp067420
Tbg972111000
Tbg97234300
Tbg97262300
Tbg97294840
Tbg97295210
THAPSDRAFT1841
TVAG002940
TVAG129710
TVAG267990
TVAG514790
ZC3H14CONSTRUCT

(d)


D. rerio
Region
5′UTR
Coding
3′UTR
Species-specificity
HS
CE
DM
DR
MM
HS-DR
HS-MM
HS-DR-DM
HS-DR-MM
HS-CE-DR-MM
HS-DR-DM-MM
Other
BRUNOL5
DAZ3
EIF4G2
Fusip1
HNRNPCL1
HNRNPDL
hnRNPLL
ILF2
PABPC3
PABPC5
PTBP3
RBM23
RBM4
SRSF8
YB1
STARPAP
ZNF638
asd1
etr1
exc7
fox1
HRP1
sup26CONSTRUCT
mec8
sup12
unc75
tiar1
tiar3
aret
B52
bru3
CG17838
CG2950
CG7804
CG7903
eIF2alpha
cpo
elav
fne
how
Hrb27C
Hrb87F
KHSRP
mod
msi
orb2
papi
pum
pAbp
pUf68
qkr58E1
Rbp1
Rbp9
Rbp1like
Rnp4F
ref2
Rox8
Rsf1
SF2
shep
sm
Srp54
Sxl
tra2
U2af50
ENSDARG00000058818
BRUNOL4
BRUNOL6
SRSF4
FXR2
SAMD4A
QKI
BOLL
CPEB1
FUBP3
HNRNPA1L2
HNRNPA2B1
HNRNPC
HNRNPF
HNRNPH2
PABPN1L
PPRC1
PRR3
PUM1
RALYL
RBFOX3
RBM3
Rbm38
RBM4B
RBM5
SRSF10
YBX2
A2BP1
EIF4B
A1CF
ANKHD1
CELF1
CPEB2
CPEB4
DAZAP1
ELAVL1
ELAVL4
ENOX1
ESRP1
ESRP2
EWSR1
FUBP1
G3BP2
HNRNPA0
HNRNPA1
HNRNPAB
HNRNPD
hnRNPK
HNRNPL
IGF2BP1
IGF2BP2
IGF2BP3
Khdrbs1
KHDRBS2
KHDRBS3
MAL8P140
lark
MATR3
MBNL1
NOVA1
NUPL2
PABPC1
PABPC4
PABPN1
PCBP1
PCBP2
PCBP4
PTBP1
PUF60
RALY
RBFOX2
RBM15B
RBM22
RBM24
RBM25
RBM28
RBM41
RBM42
RBM45
RBM46
RBM47
RBM6
RBM8A
RBMS1
RBMS2
RBMS3
RC3H1
SART3
sf3b4
SFPQ
SNRPA
SRSF1
SRSF11
SRSF2
SRSF5
SRSF9
TAF15
TARDBP
TIA1
TRA2A
TRNAU1AP
U2AF2
ZC3H10
ZCRB1
ZFP36
MSI1
CNOT4
Fmr1
FUS
SF1
UNK
SNRNP70
ANIA04546
AT1G76460TAIRG
Cscaffold40000008
egw103003671
egw109003261
ENSXETG00000012802
ENSXETG00000018075
estExtfgeneshHS
gw184451
ENSGALG00000000814
ENSGALG00000003765
LmjF180180
LmjF352550
LmjF354130
LIN28A
mubCONSTRUCT
NAB2CONSTRUCT
MAL13P135
NCU02404
PF100214
PF130315
PFF0320c
PFI1435w
RO3G00049
Smp067420
Tbg972111795OCONS
Tbg97262300
Tbg97294840
Tbg97295210
THAPSDRAFT1841
TVAG002940
TVAG129710
TVAG267990
TVAG514790
ZC3H14CONSTRUCT

(e)

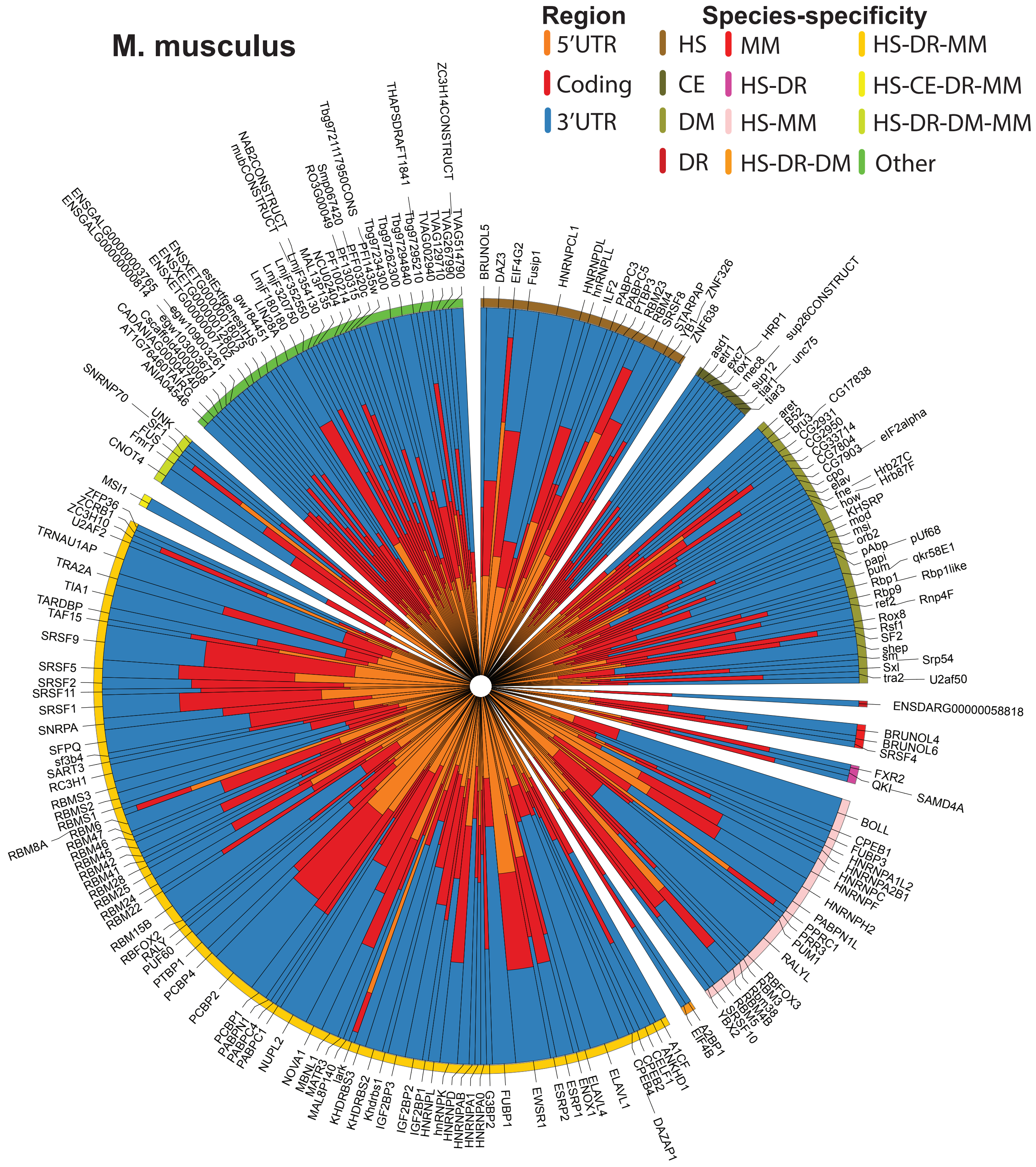
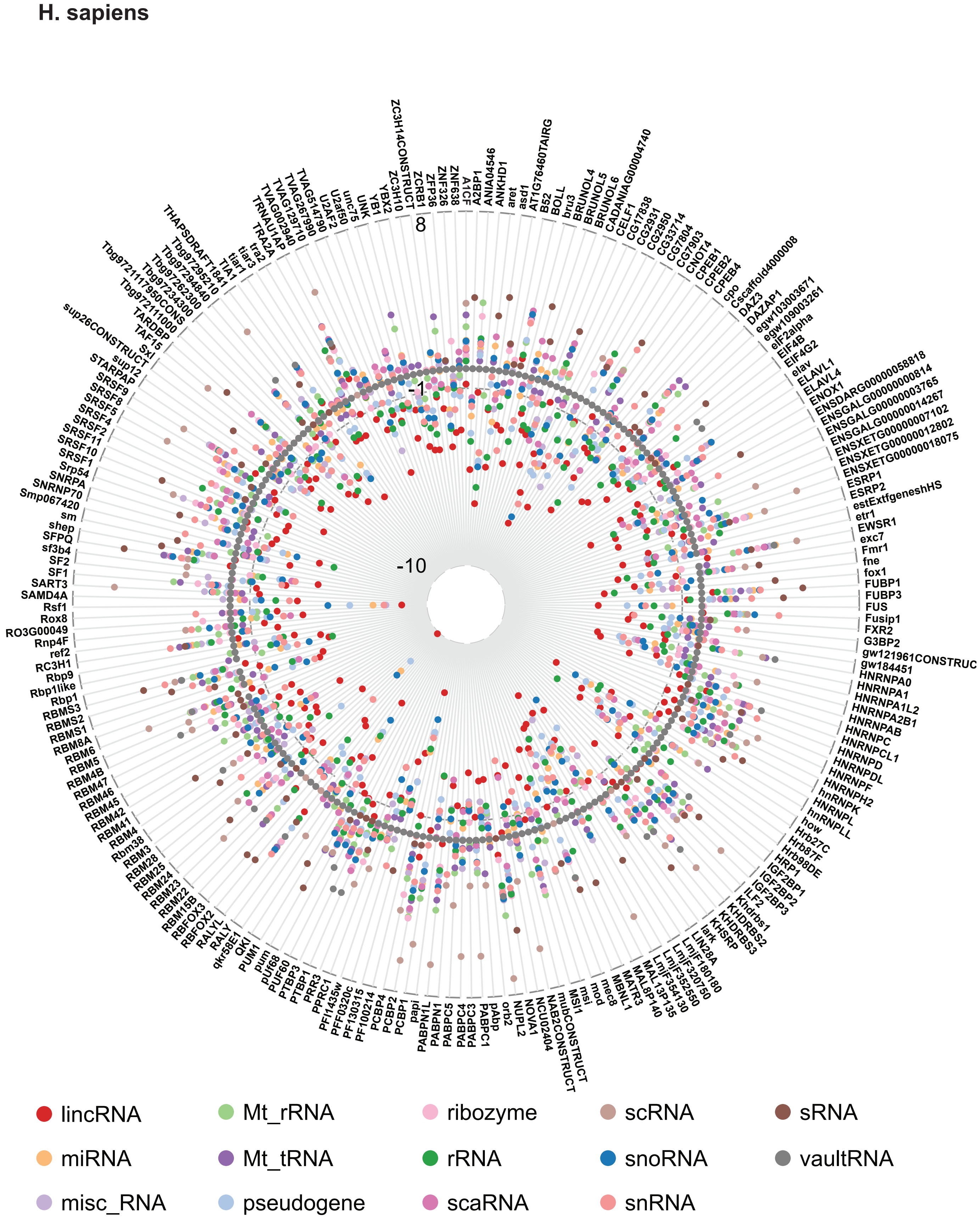


Figure 5.S3. Distribution of motif instances for each RBPs within each region of mRNAs. Circos plots of the relative distributions of RBP motif instances per 1 kilobase within each region (5'UTR, coding, and 3'UTR) for each species, **(a)** *H. sapiens*, **(b)** *C. elgans*, **(c)** *D. melanogaster*, **(d)** *D. rerio*, **(e)** *M. musculus*. The outer circle segments displays the species-specificities of each RBPs according to Ray *et al* [40, 39].

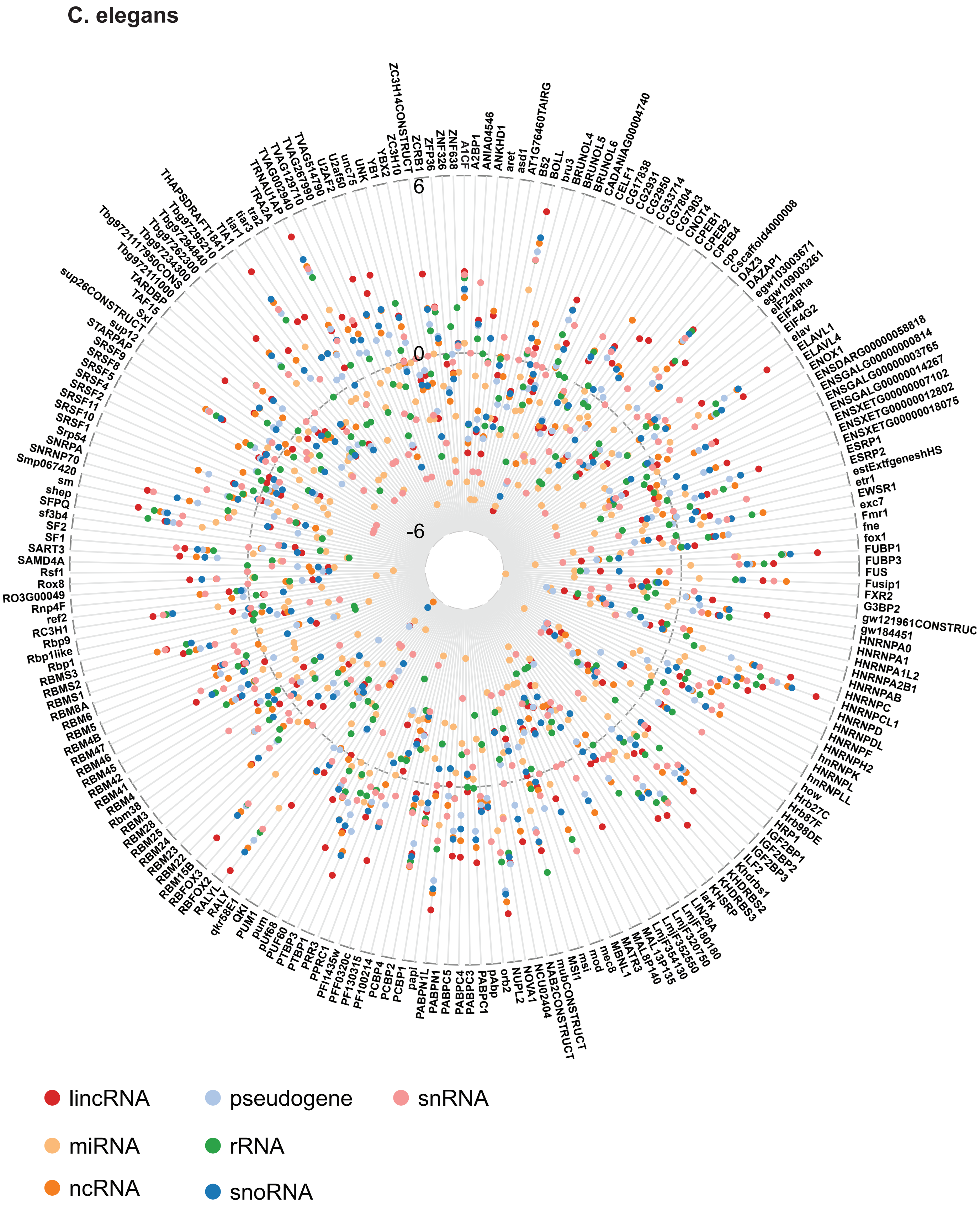
(a)



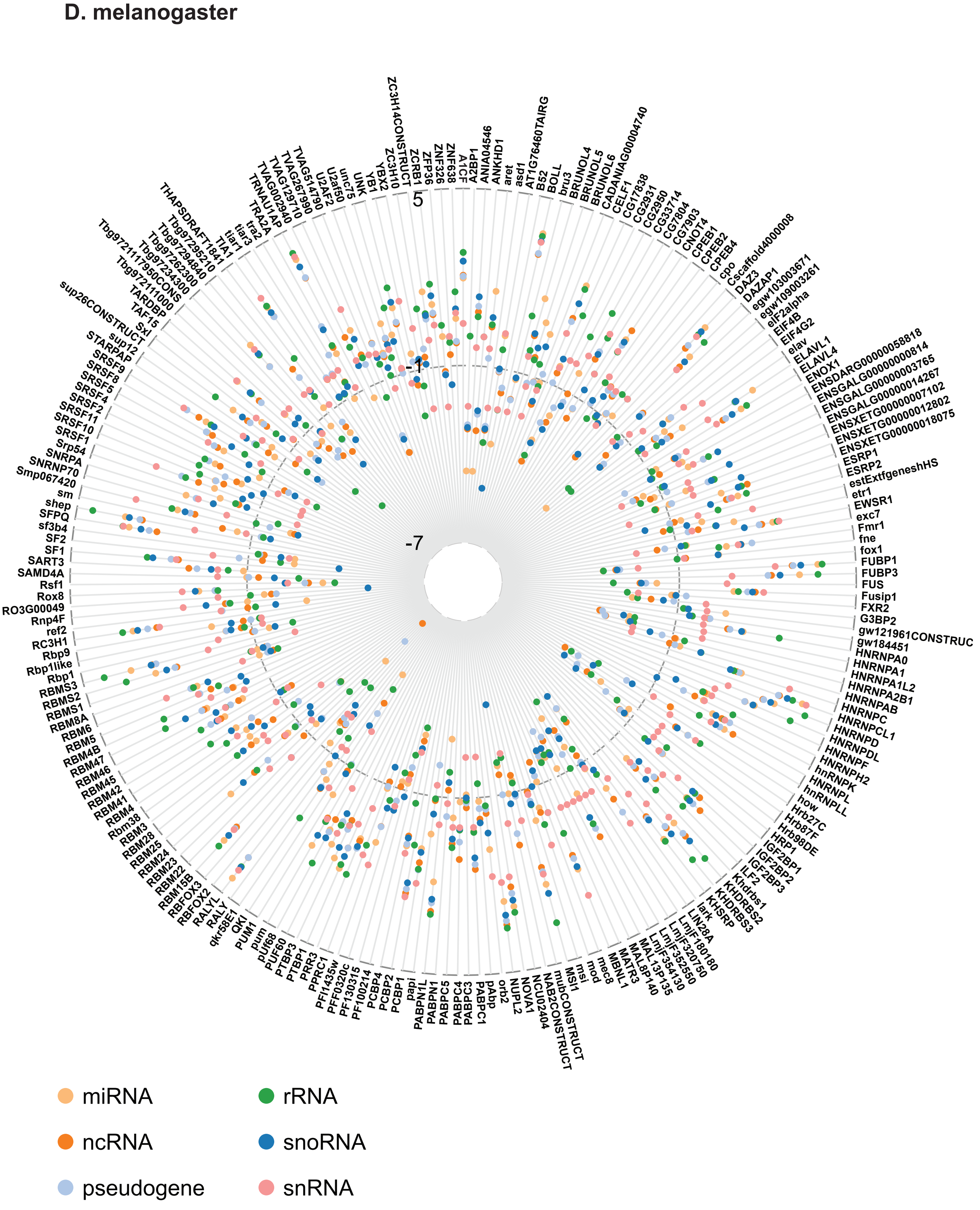
H. sapiens

8
-1
-10
A1CF
A2BP1
ANIA04546
ANKHD1
aret
asd1
AT1G76460TAIRG
B52
BOLL
bru3
BRUNOL4
BRUNOL5
BRUNOL6
CADANIAG00004740
CELF1
CG17838
CG2931
CG2950
CG33714
CG7804
CG7903
CNOT4
CPEB1
CPEB2
CPEB4
cpo
Cscaffold4000008
DAZ3
DAZAP1
egw103003671
egw109003261
elF2alpha
ElF4B
ElF4G2
elav
ELAVL1
ELAVL4
ENOX1
ENSDARG00000058818
ENSGALG00000000814
ENSGALG00000003765
ENSGALG00000014267
ENSXETG00000007102
ENSXETG00000012802
ENSXETG00000018075
ESRP1
ESRP2
estExtfgeneshHS
etr1
EWSR1
exc7
Fmr1
fne
fox1
FUBP1
FUBP3
FUS
Fusip1
FXR2
G3BP2
gw121961CONSTRUC
gw184451
HNRNPA0
HNRNPA1
HNRNPA1L2
HNRNPA2B1
HNRNPAB
HNRNPC
HNRNPCL1
HNRNPD
HNRNPDL
HNRNPF
HNRNPH2
hnRNPK
HNRNPL
hnRNPLL
how
Hrb27C
Hrb87F
Hrb98DE
HRP1
IGF2BP1
IGF2BP2
IGF2BP3
ILF2
Khdrbs1
KHDRBS2
KHDRBS3
KHSRP
lark
LIN28A
LmjF180180
LmjF320750
LmjF352550
LmjF354130
MAL13P135
MAL8P140
MATR3
MBNL1
mec8
mod
msi
MSI1
mubCONSTRUCT
NAB2CONSTRUCT
NCU02404
NOVA1
NUPL2
orb2
pAbp
PABPC1
PABPC3
PABPC4
PABPC5
PABPN1
PABPN1L
papi
PCBP1
PCBP2
PCBP4
PF100214
PF130315
PFF0320c
PFI1435w
PPRC1
PRR3
PTBP1
PTBP3
PUF60
pUf68
pum
PUM1
QKI
qkr58E1
RALY
RALYL
RBFOX2
RBFOX3
RBM15B
RBM22
RBM23
RBM24
RBM25
RBM28
RBM3
Rbm38
RBM4
RBM41
RBM42
RBM45
RBM46
RBM47
RBM4B
RBM5
RBM6
RBM8A
RBMS1
RBMS2
RBMS3
Rbp1
Rbp1like
Rbp9
RC3H1
ref2
Rnp4F
RO3G00049
Rox8
Rsf1
SAMD4A
SART3
SF1
SF2
sf3b4
SFPQ
shep
sm
Smp067420
SNRNP70
SNRPA
Srp54
SRSF1
SRSF10
SRSF11
SRSF2
SRSF4
SRSF5
SRSF8
SRSF9
STARPAP
sup12
sup26CONSTRUCT
Sxl
TAF15
TARDBP
Tbg972111000
Tbg9721117950CONS
Tbg97234300
Tbg97262300
Tbg97294840
Tbg97295210
THAPSDRAFT1841
TIA1
tiar1
tiar3
tra2
TRA2A
TRNAU1AP
TVAG002940
TVAG129710
TVAG267990
TVAG514790
U2AF2
U2af50
unc75
UNK
YB1
YBX2
ZC3H10
ZC3H14CONSTRUCT
ZCRB1
ZFP36
ZNF326
ZNF638
● lincRNA
● miRNA
● misc_RNA
● Mt_rRNA
● Mt_tRNA
● pseudogene
● ribozyme
● rRNA
● scaRNA
● scRNA
● snoRNA
● snRNA
● sRNA
● vaultRNA

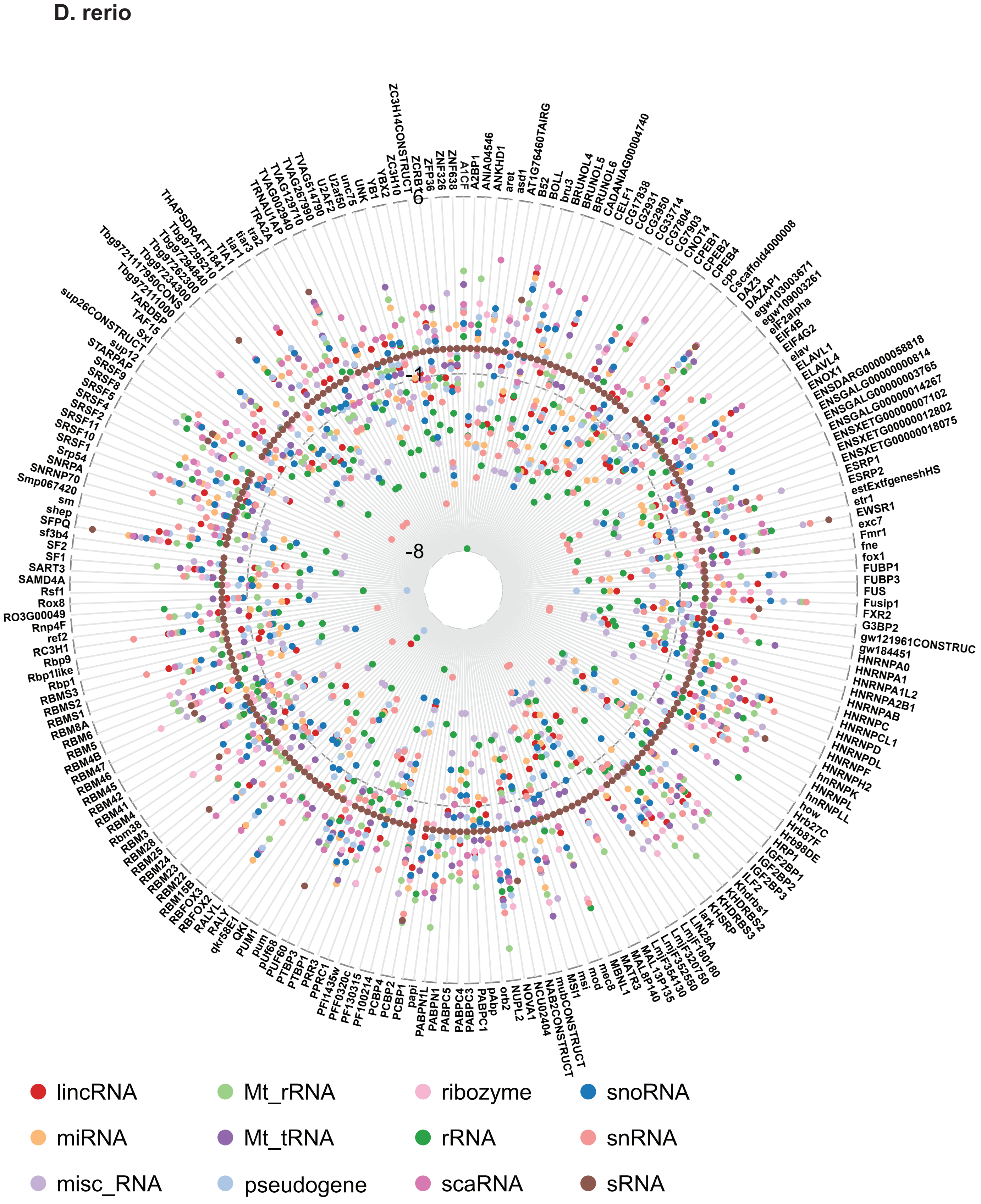
(b)


C. elegans
6
0
-6
A1CF
A2BP1
ANIA04546
ANKHD1
aret
asd1
AT1G76460TAIRG
B52
BOLL
bru3
BRUNOL4
BRUNOL5
BRUNOL6
CADANIAG00004740
CELF1
CG17838
CG2931
CG2950
CG33714
CG7804
CG7903
CNOT4
CPEB1
CPEB2
CPEB4
cpo
Cscaffold4000008
DAZ3
DAZAP1
egw103003671
egw109003261
eIF2alpha
EIF4B
EIF4G2
elav
ELAVL1
ELAVL4
ENOX1
ENSDARG00000058818
ENSGALG00000000814
ENSGALG00000003765
ENSGALG00000014267
ENSXETG00000007102
ENSXETG00000012802
ENSXETG00000018075
ESRP1
ESRP2
estExtfgeneshHS
etr1
EWSR1
exc7
Fmr1
fne
fox1
FUBP1
FUBP3
FUS
Fusip1
FXR2
G3BP2
gw121961CONSTRUC
gw184451
HNRNPA0
HNRNPA1
HNRNPA1L2
HNRNPA2B1
HNRNPAB
HNRNPC
HNRNPCL1
HNRNPD
HNRNPDL
HNRNPF
HNRNPH2
hnRNPK
HNRNPL
hnRNPLL
how
Hrb27C
Hrb87F
Hrb98DE
HRP1
IGF2BP1
IGF2BP2
IGF2BP3
ILF2
Khdrbs1
KHDRBS2
KHDRBS3
KHSRP
lark
LIN28A
LmjF180180
LmjF320750
LmjF352550
LmjF354130
MAL13P135
MAL8P140
MATR3
MBNL1
mec8
mod
msi
MSI1
mubCONSTRUCT
NAB2CONSTRUCT
NCU02404
NOVA1
NUPL2
orb2
pAbp
PABPC1
PABPC3
PABPC4
PABPC5
PABPN1
PABPN1L
papi
PCBP1
PCBP2
PCBP4
PF100214
PF130315
PFF0320c
PFI1435w
PPRC1
PRR3
PTBP1
PTBP3
PUF60
pUf68
pum
PUM1
QKI
qkr58E1
RALY
RALYL
RBFOX2
RBFOX3
RBM15B
RBM22
RBM23
RBM24
RBM25
RBM28
RBM3
Rbm38
RBM4
RBM41
RBM42
RBM45
RBM46
RBM47
RBM4B
RBM5
RBM6
RBM8A
RBMS1
RBMS2
RBMS3
Rbp1
Rbp1like
Rbp9
RC3H1
ref2
Rnp4F
RO3G00049
Rox8
Rsf1
SAMD4A
SART3
SF1
SF2
sf3b4
SFPQ
shep
sm
Smp067420
SNRNP70
SNRPA
Srp54
SRSF1
SRSF10
SRSF11
SRSF2
SRSF4
SRSF5
SRSF8
SRSF9
STARPAP
sup12
sup26CONSTRUCT
Sxl
TAF15
TARDBP
Tbg972111000
Tbg9721117950CONS
Tbg97234300
Tbg97262300
Tbg97294840
Tbg97295210
THAPSDRAFT1841
TIA1
tiar1
tiar3
tra2
TRA2A
TRNAU1AP
TVAG002940
TVAG129710
TVAG267990
TVAG514790
U2AF2
U2af50
unc75
UNK
YB1
YBX2
ZC3H10
ZC3H14CONSTRUCT
ZCRB1
ZFP36
ZNF326
ZNF638
● lincRNA
● pseudogene
● snRNA
● miRNA
● rRNA
● ncRNA
● snoRNA

(c)


D. melanogaster
5
-1
-7
ZC3H14CONSTRUCT
ZC3H10
ZCRB1
ZFP36
ZNF326
ZNF638
A1CF
A2BP1
ANIA04546
ANKHD1
aret
asd1
AT1G76460TAIRG
B52
BOLL
bru3
BRUNOL4
BRUNOL5
BRUNOL6
CADANIAG00004740
CELF1
CG17838
CG2931
CG2950
CG33714
CG7804
CG7903
CNOT4
CPEB1
CPEB2
CPEB4
cpo
Cscaffold4000008
DAZ3
DAZAP1
egw103003671
egw109003261
eIF2alpha
EIF4B
EIF4G2
elav
ELAVL1
ELAVL4
ENOX1
ENSDARG00000058818
ENSGALG00000000814
ENSGALG00000003765
ENSGALG00000014267
ENSXETG00000007102
ENSXETG00000012802
ENSXETG00000018075
ESRP1
ESRP2
estExtfgeneshHS
etr1
EWSR1
exc7
Fmr1
fne
fox1
FUBP1
FUBP3
FUS
Fusip1
FXR2
G3BP2
gw121961CONSTRUC
gw184451
HNRNPA0
HNRNPA1
HNRNPA1L2
HNRNPA2B1
HNRNPAB
HNRNPC
HNRNPCL1
HNRNPD
HNRNPDL
HNRNPF
HNRNPH2
hnRNPK
HNRNPL
hnRNPLL
how
Hrb27C
Hrb87F
Hrb98DE
HRP1
IGF2BP1
IGF2BP2
IGF2BP3
ILF2
Khdrbs1
KHDRBS2
KHDRBS3
KHSRP
lark
LIN28A
LmjF180180
LmjF320750
LmjF352550
LmjF354130
MAL13P135
MAL8P140
MATR3
MBNL1
mec8
mod
msi
MSI1
mubCONSTRUCT
NAB2CONSTRUCT
NCU02404
NOVA1
NUPL2
orb2
pAbp
PABPC1
PABPC3
PABPC4
PABPC5
PABPN1
PABPN1L
papi
PCBP1
PCBP2
PCBP4
PF100214
PF130315
PFF0320c
PFI1435w
PPRC1
PRR3
PTBP1
PTBP3
PUF60
puf68
pum
PUM1
QKI
qkr58E1
RALY
RALYL
RBFOX2
RBFOX3
RBM15B
RBM22
RBM23
RBM24
RBM25
RBM28
RBM3
Rbm38
RBM4
RBM41
RBM42
RBM45
RBM46
RBM47
RBM4B
RBM5
RBM6
RBM8A
RBMS1
RBMS2
RBMS3
Rbp1
Rbp1like
Rbp9
RC3H1
ref2
Rnp4F
RO3G00049
Rox8
Rsf1
SAMD4A
SART3
SF1
SF2
sf3b4
SFPQ
shep
sm
Smp067420
SNRNP70
SNRPA
Srp54
SRSF1
SRSF10
SRSF11
SRSF2
SRSF4
SRSF5
SRSF8
SRSF9
STARPAP
sup12
sup26CONSTRUCT
Sxl
TAF15
TARDBP
Tbg972111000
Tbg972117950CONS
Tbg97234300
Tbg97262300
Tbg97294840
Tbg97295210
THAPSDRAFT1841
TIA1
tiar1
tiar3
tra2
TRA2A
TRNAU1AP
TVAG002940
TVAG129710
TVAG267990
TVAG514790
U2AF2
U2af50
unc75
UNK
YB1
YBX2
● miRNA
● rRNA
● ncRNA
● snoRNA
● pseudogene
● snRNA

(d)


D. rerio
0
-1
-8
A1CF
A2BP1
ANIA04546
ANKHD1
aret
asd1
AT1G76460TAIRG
B52
BOLL
bru3
BRUNOL4
BRUNOL5
BRUNOL6
CADANIAG00004740
CELF1
CG17838
CG2931
CG2950
CG33714
CG7804
CG7903
CNOT4
CPEB1
CPEB2
CPEB4
cpo
Cscaffold4000008
DAZ3
DAZAP1
egw103003671
egw109003261
eIF2alpha
EIF4B
EIF4G2
elav
ELAVL1
ELAVL4
ENOX1
ENSDARG00000058818
ENSGALG00000000814
ENSGALG00000003765
ENSGALG00000014267
ENSXETG00000007102
ENSXETG00000012802
ENSXETG00000018075
ESRP1
ESRP2
estExtfgeneshHS
etr1
EWSR1
exc7
Fmr1
fne
fox1
FUBP1
FUBP3
FUS
Fusip1
FXR2
G3BP2
gw121961CONSTRUC
gw184451
HNRNPA0
HNRNPA1
HNRNPA1L2
HNRNPA2B1
HNRNPAB
HNRNPC
HNRNPCL1
HNRNPD
HNRNPDL
HNRNPF
HNRNPH2
hnRNPK
HNRNPL
hnRNPLL
how
Hrb27C
Hrb87F
Hrb98DE
HRP1
IGF2BP1
IGF2BP2
IGF2BP3
ILF2
Khdrbs1
KHDRBS2
KHDRBS3
KHSRP
lark
LIN28A
LmjF180180
LmjF320750
LmjF352550
LmjF354130
MAL13P135
MAL8P140
MATR3
MBNL1
mec8
mod
msi
MSI1
mubCONSTRUCT
NAB2CONSTRUCT
NCU02404
NOVA1
NUPL2
orb2
pAbp
PABPC1
PABPC3
PABPC4
PABPC5
PABPN1
PABPN1L
papi
PCBP1
PCBP2
PCBP4
PF100214
PF130315
PFF0320c
PFI1435w
PPRC1
PRR3
PTBP1
PTBP3
PUF60
pUf68
pum
PUM1
QKI
qkr58E1
RALY
RALYL
RBFOX2
RBFOX3
RBM15B
RBM22
RBM23
RBM24
RBM25
RBM28
RBM3
Rbm38
RBM4
RBM41
RBM42
RBM45
RBM46
RBM47
RBM4B
RBM5
RBM6
RBM8A
RBMS1
RBMS2
RBMS3
Rbp1
Rbp1like
Rbp9
RC3H1
ref2
Rnp4F
RO3G00049
Rox8
Rsf1
SAMD4A
SART3
SF1
SF2
sf3b4
SFPQ
shep
sm
Smp067420
SNRNP70
SNRPA
Srp54
SRSF1
SRSF10
SRSF11
SRSF2
SRSF4
SRSF5
SRSF8
SRSF9
STARPAP
sup12
sup26CONSTRUCT
Sxl
TAF15
TARDBP
Tbg972111000
Tbg9721117950CONS
Tbg97234300
Tbg97262300
Tbg97294840
Tbg97295210
THAPSDRAFT1841
TIA1
tiar1
tiar3
tra2
TRA2A
TRNAU1AP
TVAG002940
TVAG129710
TVAG267990
TVAG514790
U2AF2
U2af50
unc75
UNK
YB1
YBX2
ZC3H10
ZC3H14CONSTRUCT
ZCRB1
ZFP36
ZNF326
ZNF638
● lincRNA
● Mt_rRNA
● ribozyme
● snoRNA
● miRNA
● Mt_tRNA
● rRNA
● snRNA
● misc_RNA
● pseudogene
● scaRNA
● sRNA

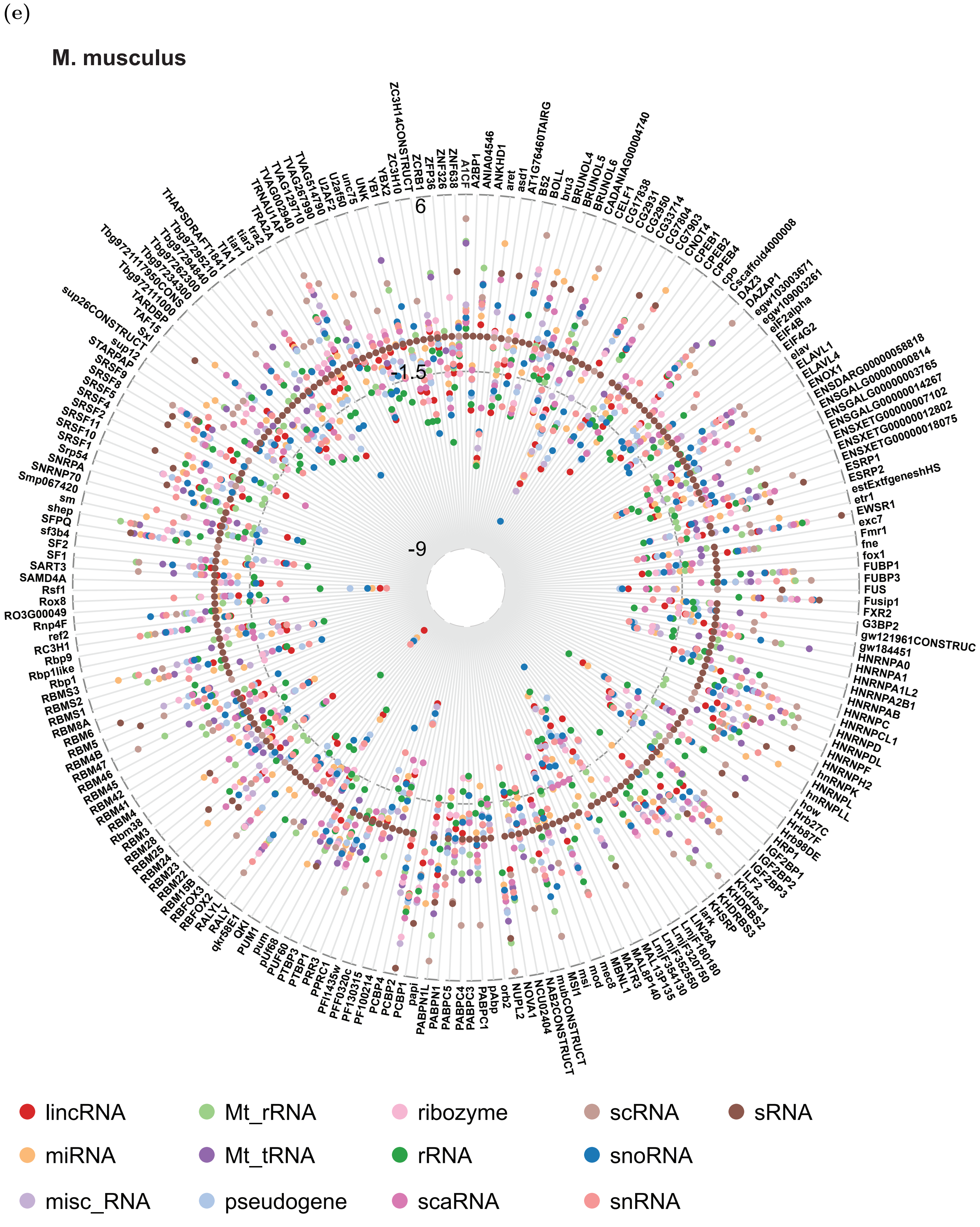


Figure 5.S4. Distribution of motif instances for each RBPs within each noncoding RNA. Radar plots of the relative distributions of RBP motif instances per 1 kilobase (log2 scale) of noncoding RNA per biotype for each species, **(a)** *H. sapiens*, **(b)** *C. elgans*, **(c)** *D. melanogaster*, **(d)** *D. rerio*, **(e)** *M. musculus*.

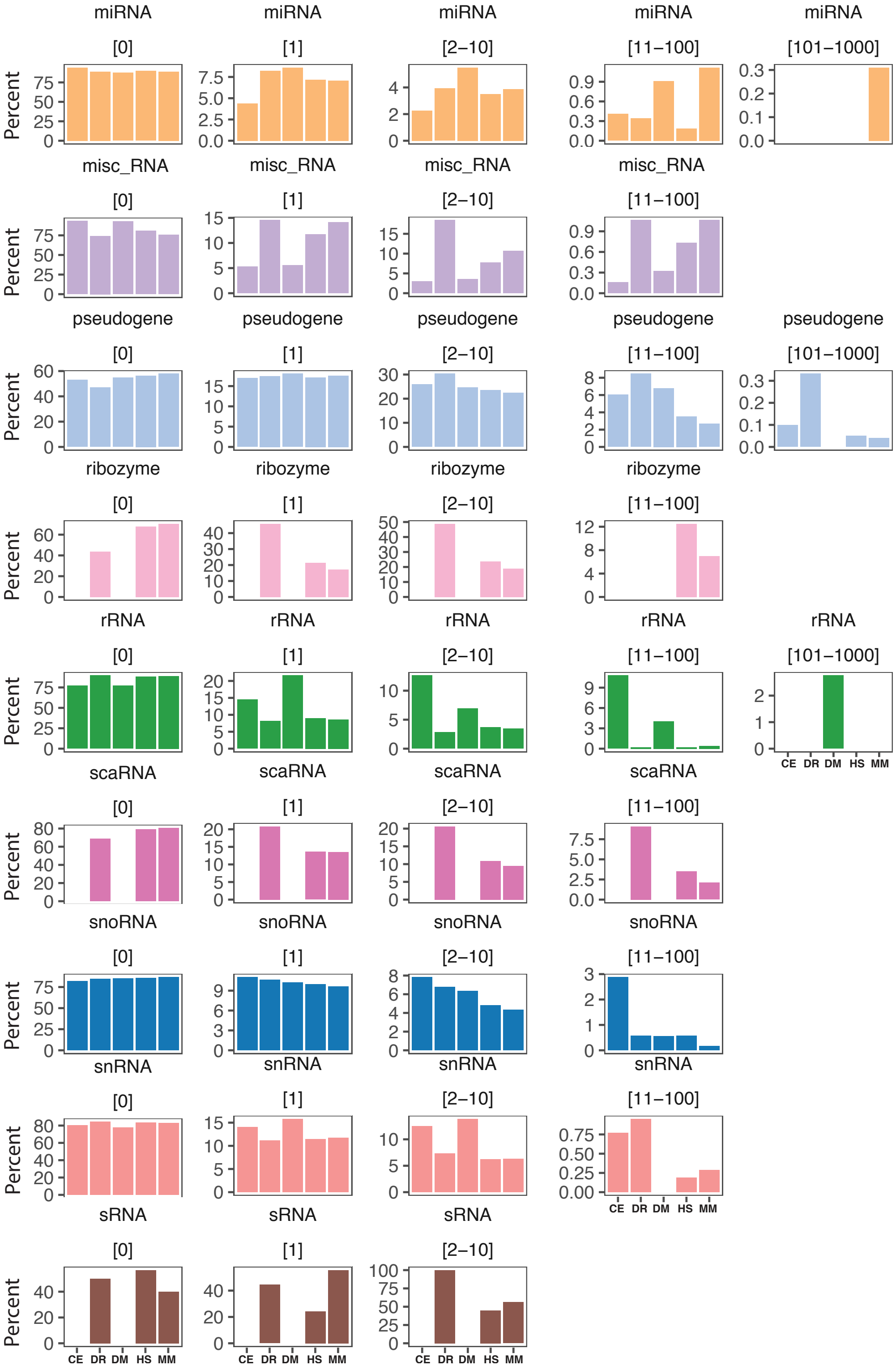


FIGURE 5.S5. Noncoding transcripts exhibit varying topologies of motifs instance. Histogram showing the combined average percent of different noncoding RNA bearing RBP motif instances binned by the number of instances they contain (displayed top of each panel). HS : *H. sapiens*, CE : *C. elegans*, DM : *D. melanogaster*, DR : *D. rerio*, MM : *M. Musculus*.

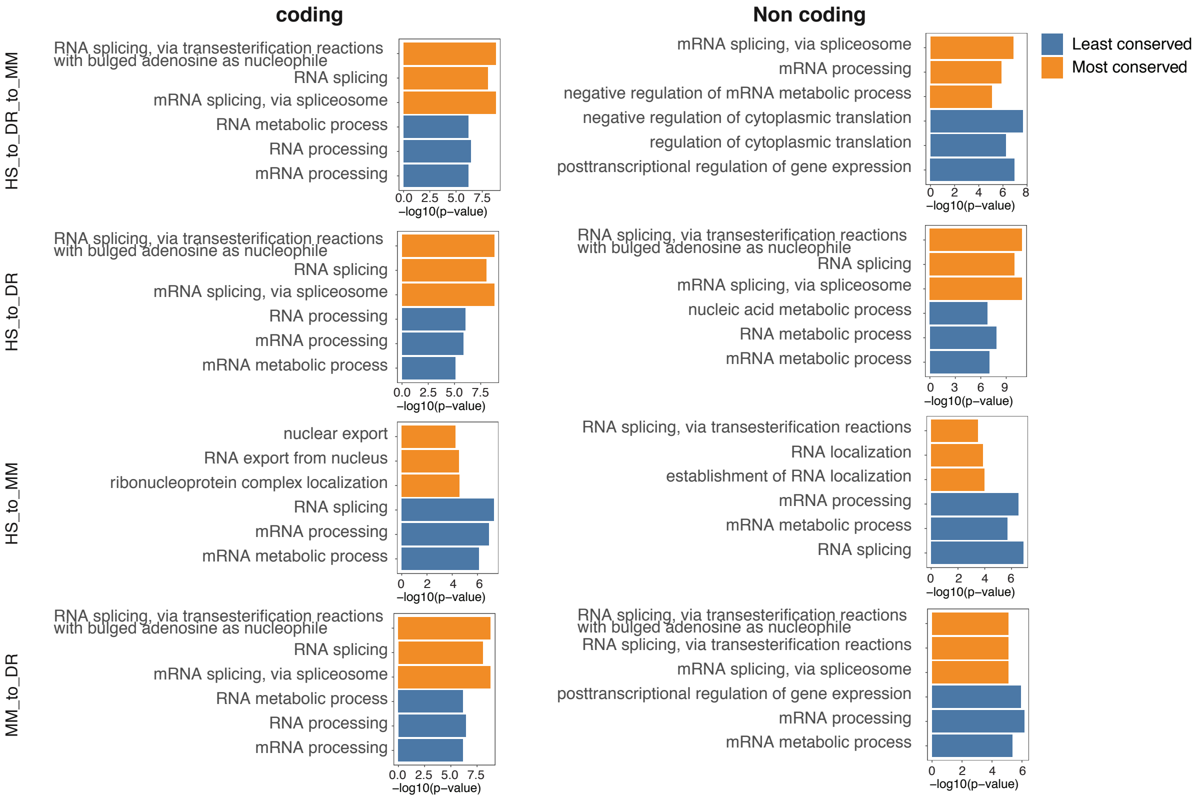


Figure 5.S6. RBPs conserved between syntenic regions of three vertebrates exhibits more specific biological processes gene ontologies. Histogram showing 3 biological process GO terms enrichment for the top 10 RBPs most conserved and 10 RBPs least conserved between comparable genomic coordinates of coding and noncoding RNA when comparing all three vertebrates or by pairwise comparison between each species. HS : *H. sapiens*, DR : *D. rerio*, MM : *M. Musculus*.

Références

[1] M. Akerman, O. I. Fregoso, S. Das, C. Ruse, M. A. Jensen, D. J. Pappin, M. Q. Zhang, and A. R. Krainer, *Differential connectivity of splicing activators and repressors to the human spliceosome*, Genome Biology, 16 (2015), p. 119.

[2] L. P. Benoit Bouvrette, M. Blanchette, and E. Lécuyer, *Bioinformatics approaches to gain insights into cis-regulatory motifs involved in mRNA localization*, Advances in Experimental Medicine and Biology, 1203 (2019), pp. 165–94.

[3] L. P. Benoit Bouvrette, S. Bovaird, M. Blanchette, and E. Lecuyer, *oRNAment : a database of putative RNA binding protein target sites in the transcriptomes of model species*, Nucleic Acids Research, 48 (2020), pp. D166–D173.

[4] J. Bergalet and E. Lecuyer, *The functions and regulatory principles of mRNA intracellular trafficking*, Systems Biology of RNA Binding Proteins, 825 (2014), pp. 57–96.

[5] G. Cestra, S. Rossi, M. Di Salvio, and M. Cozzolino, *Control of mRNA translation in ALS proteinopathy*, Frontiers in Molecular Neuroscience, 10 (2017), p. 85.

[6] N. A. Cody, C. Iampietro, and E. Lecuyer, *The many functions of mRNA localization during normal development and disease : From pillar to post*, Wiley Interdisciplinary Reviews-Developmental Biology, 2 (2013), pp. 781–96.

[7] K. B. Cook, H. Kazan, K. Zuberi, and Q. Morris, *RBPDB : a database of RNA-binding specificities*, Nucleic Acids Research, 39 (2010), pp. D301–8.

[8] K. B. Cook, S. Vembu, K. C. H. Ha, H. Zheng, K. U. Laverty, T. R. Hughes, D. Ray, and Q. D. Morris, *RNAcompete-S : Combined RNA sequence/structure preferences for RNA binding proteins derived from a single-step in vitro selection*, Methods, 126 (2017), pp. 18–28.

[9] M. K. Das and H. K. Dai, *A survey of DNA motif finding algorithms*, BMC Bioinformatics, 8 Suppl 7 (2007), p. S21.

[10] D. Dominguez, P. Freese, M. S. Alexis, A. Su, M. Hochman, T. Palden, C. Bazile, N. J. Lambert, E. L. Van Nostrand, G. A. Pratt, G. W. Yeo, B. R. Graveley, and C. B. Burge, *Sequence, structure, and context preferences of human RNA binding proteins*, Molecular Cell, 70 (2018), pp. 854–67 e9.

[11] S. El-Gebali, J. Mistry, A. Bateman, S. R. Eddy, A. Luciani, S. C. Potter, M. Qureshi, L. J. Richardson, G. A. Salazar, A. Smart, E. L. L. Sonnhammer, L. Hirsh, L. Paladin, D. Piovesan, S. C. E. Tosatto, and R. D. Finn, *The Pfam protein families database in 2019*, Nucleic Acids Research, 47 (2019), pp. D427–32.

[12] Y. Feng, M. Chen, and J. L. Manley, *Phosphorylation switches the general splicing repressor SRp38 to a sequence-specific activator*, Nature Structural and Molecular Biology, 15 (2008), pp. 1040–8.

[13] S. Gerstberger, M. Hafner, and T. Tuschl, *A census of human RNA-binding proteins*, Nature Reviews Genetics, 15 (2014), pp. 829–45.

[14] G. Giudice, F. Sánchez-Cabo, C. Torroja, and E. Lara-Pezzi, *ATtRACT—a database of RNA-binding proteins and associated motifs*, Database, (2016).

[15] T. Glisovic, J. L. Bachorik, J. Yong, and G. Dreyfuss, *RNA-binding proteins and post-transcriptional gene regulation*, FEBS Letters, 582 (2008), pp. 1977–86.

[16] S. Gupta, J. A. Stamatoyannopoulos, T. L. Bailey, and W. S. Noble, *Quantifying similarity between motifs*, Genome Biology, 8 (2007), p. R24.

[17] D. Heller, R. Krestel, U. Ohler, M. Vingron, and A. Marsico, *ssHMM : extracting intuitive sequence-structure motifs from high-throughput RNA-binding protein data*, Nucleic Acids Research, 45 (2017), pp. 11004–18.

[18] A. S. Hinrichs, D. Karolchik, R. Baertsch, G. P. Barber, G. Bejerano, H. Clawson, M. Diekhans, T. S. Furey, R. A. Harte, F. Hsu, J. Hillman-Jackson, R. M. Kuhn, J. S. Pedersen, A. Pohl, B. J. Raney, K. R. Rosenbloom, A. Siepel, K. E. Smith, C. W. Sugnet, A. Sultan-Qurraie, D. J. Thomas, H. Trumbower, R. J. Weber, M. Weirauch, A. S. Zweig, D. Haussler, and W. J. Kent, *The UCSC genome browser database : update 2006*, Nucleic Acids Research, 34 (2006), pp. D590–8.

[19] H. Kazan, D. Ray, E. T. Chan, and T. R. Hughes, *RNAcontext : a new method for learning the sequence and structure binding preferences of RNA-binding proteins*, PLoS Computational Biology, 6 (2010).

[20] A. E. Kel, E. Gossling, I. Reuter, E. Cheremushkin, O. V. Kel-Margoulis, and E. Wingender, *MATCH : A tool for searching transcription factor binding sites in DNA sequences*, Nucleic Acids Research, 31 (2003), pp. 3576–9.

[21] P. J. Kersey, J. E. Allen, A. Allot, M. Barba, S. Boddu, B. J. Bolt, D. Carvalho-Silva, M. Christensen, P. Davis, C. Grabmueller, N. Kumar, Z. Liu, T. Maurel, B. Moore, M. D. McDowall, U. Maheswari, G. Naamati, V. Newman, C. K. Ong, M. Paulini, H. Pedro, E. Perry, M. Russell, H. Sparrow, E. Tapanari, K. Taylor, A. Vullo, G. Williams, A. Zadissia, A. Olson, J. Stein, S. Wei, M. Tello-Ruiz, D. Ware, A. Luciani, S. Potter, R. D. Finn, M. Urban, K. E. Hammond-Kosack, D. M. Bolser, N. De Silva, K. L. Howe, N. Langridge, G. Maslen, D. M. Staines, and A. Yates, *Ensembl genomes 2018 : an integrated omics infrastructure for non-vertebrate species*, Nucleic Acids Research, 46 (2018), pp. D802–8.

[22] N. J. Lambert, A. D. Robertson, and C. B. Burge, *RNA Bind-n-Seq : Measuring the binding affinity landscape of RNA-binding proteins*, Methods Enzymology, 558 (2015), pp. 465–93.

[23] M. S. Lan and M. B. Breslin, *Structure, expression, and biological function of INSM1 transcription factor in neuroendocrine differentiation*, The FASEB Journal, 23 (2009), pp. 2024–33.

[24] K. Leppek, J. Schott, S. Reitter, F. Poetz, M. C. Hammond, and G. Stoecklin, *Roquin promotes constitutive mRNA decay via a conserved class of stem-loop recognition motifs*, Cell, 153 (2013), pp. 869–81.

[25] M. R. Lerner and J. A. Steitz, *Antibodies to small nuclear RNAs complexed with proteins are produced by patients with systemic lupus erythematosus*, Proceedings of the National Academy of Sciences of the United States of America, 76 (1979), pp. 5495–9.

[26] X. Li, H. Kazan, and H. D. Lipshitz, *Finding the target sites of RNA-binding proteins*, Wiley interdisciplinary reviews–RNA, 5 (2014), pp. 111–30.

[27] S. Lu, J. Wang, F. Chitsaz, M. K. Derbyshire, R. C. Geer, N. R. Gonzales, M. Gwadz, D. I. Hurwitz, G. H. Marchler, J. S. Song, N. Thanki, R. A. Yamashita, M. Yang, D. Zhang, C. Zheng, C. J. Lanczycki, and A. Marchler-Bauer, *CDD/SPARCLE : the conserved domain database in 2020*, Nucleic Acids Research, 48 (2020), pp. D265–8.

[28] K. E. Lukong, K. W. Chang, E. W. Khandjian, and S. Richard, *RNA-binding proteins in human genetic disease*, Trends in Genetics, 24 (2008), pp. 416–25.

[29] D. Maticzka, S. J. Lange, F. Costa, and R. Backlofen, *GraphProt : modeling binding preferences of RNA-binding proteins*, Genome Biology, 15 (2014).

[30] V. Matys, E. Fricke, R. Geffers, E. Gossling, M. Haubrock, R. Hehl, K. Hornischer, D. Karas, A. E. Kel, O. V. Kel-Margoulis, D. U. Kloos, S. Land, B. Lewicki-Potapov, H. Michael, R. Munch, I. Reuter, S. Rotert, H. Saxel, M. Scheer, S. Thiele, and E. Wingender, *TRANSFAC : transcriptional regulation, from patterns to profiles*, Nucleic Acids Research, 31 (2003), pp. 374–8.

[31] S. Mili and J. A. Steitz, *Evidence for reassociation of RNA-binding proteins after cell lysis : implications for the interpretation of immunoprecipitation analyses*, RNA, 10 (2004), pp. 1692–4.

[32] K. V. Morris and J. S. Mattick, *The rise of regulatory RNA*, Nature Reviews Genetics, 15 (2014), pp. 423–37.

[33] Y. Orenstein, Y. Wang, and B. Berger, *RCK : accurate and efficient inference of sequence- and structure-based protein–RNA binding models from RNAcompete data*, Bioinformatics, 32 (2016), pp. i351–9.

[34] I. Paz, I. Kosti, M. Ares Jr, and M. Cline, *RBPmap : a web server for mapping binding sites of RNA-binding proteins*, Nucleic Acids Research, 42 (2014), pp. W361–7.

[35] C. C. Query, R. C. Bentley, and J. D. Keene, *A common RNA recognition motif identified within a defined U1 RNA binding domain of the 70K U1 snRNP protein*, Cell, 57 (1989), pp. 89–101.

[36] J. D. Ransohoff, Y. Wei, and P. A. Khavari, *The functions and unique features of long intergenic non-coding RNA*, Nature Reviews Molecular Cell Biology, 19 (2018), pp. 143–157.

[37] N. Rappaport, M. Twik, I. Plaschkes, R. Nudel, T. Iny Stein, J. Levitt, M. Gershoni, C. P. Morrey, M. Safran, and D. Lancet, *MalaCards : an amalgamated human disease compendium with diverse clinical and genetic annotation and structured search*, Nucleic Acids Research, 45 (2017), pp. D877–87.

[38] D. Ray, K. C. H. Ha, K. Nie, H. Zheng, T. R. Hughes, and Q. D. Morris, *RNAcompete methodology and application to determine sequence preferences of unconventional RNA-binding proteins*, Methods, 118-9 (2017), pp. 3–15.

[39] D. Ray, H. Kazan, E. T. Chan, L. Pena Castillo, S. Chaudhry, S. Talukder, B. J. Blencowe, Q. Morris, and T. R. Hughes, *Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins*, Nature Biotechnology, 27 (2009), pp. 667–70.

[40] D. Ray, H. Kazan, K. B. Cook, M. T. Weirauch, H. S. Najafabadi, X. Li, S. Gueroussov, M. Albu, H. Zheng, A. Yang, H. Na, M. Irimia, L. H. Matzat, R. K. Dale, S. A. Smith, C. A. Yarosh, S. M. Kelly, B. Nabet, D. Mecenas, W. Li, R. S. Laishram, M. Qiao, H. D. Lipshitz, F. Piano, A. H. Corbett, R. P. Carstens, B. J. Frey, R. A. Anderson, K. W. Lynch, L. O. Penalva, E. P. Lei, A. G. Fraser, B. J. Blencowe, Q. D. Morris, and T. R. Hughes, *A compendium of RNA-binding motifs for decoding gene regulation*, Nature, 499 (2013), pp. 172–7.

[41] J. L. Rinn and J. Ule, *’oming in on RNA-protein interactions*, Genome Biology, 15 (2014), p. 401.

[42] M. Sauvageau, *Diverging RNPs : Toward understanding lncRNA-protein interactions and functions*, Advances in Experimental Medicine and Biology, 1203 (2019), pp. 285–312.

[43] H. Sidibe and C. Vande Velde, *RNA granules and their role in neurodegenerative diseases*, Advances in Experimental Medicine and Biology, 1203 (2019), pp. 195–245.

[44] H. Siomi, M. J. Matunis, W. M. Michael, and G. Dreyfuss, *The pre-mRNA binding K protein contains a novel evolutionarily conserved motif*, Nucleic Acids Research, 21 (1993), pp. 1193–8.

[45] G. Stelzer, N. Rosen, I. Plaschkes, S. Zimmerman, M. Twik, S. Fishilevich, T. I. Stein, R. Nudel, I. Lieder, Y. Mazor, S. Kaplan, D. Dahary, D. Warshawsky, Y. Guan-Golan, A. Kohn, N. Rappaport, M. Safran, and D. Lancet, *The GeneCards Suite : From gene data mining to disease genome sequence analyses*, Current Protocols Bioinformatics, 54 (2016), pp. 1.30.1–33.

[46] B. Sundararaman, L. Zhan, S. M. Blue, R. Stanton, K. Elkins, S. Olson, X. Wei, E. L. Van Nostrand, G. A. Pratt, S. C. Huelga, B. M. Smalec, X. Wang, E. L. Hong, J. M. Davidson, E. Lecuyer, B. R. Graveley, and G. W. Yeo, *Resources for the comprehensive discovery of functional RNA elements*, Molecular Cell, 61 (2016), pp. 903–13.

[47] J. Thurmond, J. L. Goodman, V. B. Strelets, H. Attrill, L. S. Gramates, S. J. Marygold, B. B. Matthews, G. Millburn, G. Antonazzo, V. Trovisco, T. C. Kaufman, B. R. Calvi, and C. FlyBase, *Flybase 2.0 : the next generation*, Nucleic Acids Research, 47 (2019), pp. D759–65.

[48] J. Ule, K. B. Jensen, M. Ruggiu, A. Mele, A. Ule, and R. B. Darnell, *CLIP identifies Nova-regulated RNA networks in the brain*, Science, 302 (2003), pp. 1212–5.

[49] M. van Dijk, H. K. Thulluru, J. Mulders, O. J. Michel, A. Poutsma, S. Windhorst, G. Kleiverda, D. Sie, A. M. Lachmeijer, and C. B. Oudejans, *HELLP babies link a novel lincRNA to the trophoblast cell cycle*, Journal of Clinical Investigation, 122 (2012), pp. 4003–11.

[50] E. L. Van Nostrand, P. Freese, G. A. Pratt, X. Wang, X. Wei, R. Xiao, S. M. Blue, J.-Y. Chen, N. A. Cody, D. Dominguez, S. Olson, B. Sundararaman, L. Zhan, C. Bazile, L. P. B. Bouvrette, J. Bergalet, M. O. Duff, K. E. Garcia, C. Gelboin-Burkhart, M. Hochman, N. J. Lambert, H. Li, T. B. Nguyen, T. Palden, I. Rabano, S. Sathe, R. Stanton, A. Su, R. Wang, B. A. Yee, B. Zhou, A. L. Louie, S. Aigner, X.-d. Fu, E. Lécuyer, C. B. Burge, B. R. Graveley, and G. W. Yeo, *A large-scale binding and functional map of human RNA binding proteins*, bioRxiv, (2018), p. 179648.

[51] E. L. Van Nostrand, G. A. Pratt, A. A. Shishkin, C. Gelboin-Burkhart, M. Y. Fang, B. Sundararaman, S. M. Blue, T. B. Nguyen, C. Surka, K. Elkins, R. Stanton, F. Rigo, M. Guttman, and G. W. Yeo, *Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP)*, Nature Methods, 13 (2016), pp. 508–14.

[52] E. L. Van Nostrand, A. A. Shishkin, G. A. Pratt, T. B. Nguyen, and G. W. Yeo, *Variation in single-nucleotide sensitivity of eCLIP derived from reverse transcription conditions*, Methods, 126 (2017), pp. 29–37.

[53] T. E. Vanderweyde and B. Wolozin, *RNA Binding Proteins in Health and Disease*, Springer International Publishing, Cham, 2017, pp. 299–312.

[54] E. T. Wang, N. A. Cody, S. Jog, M. Biancolella, T. T. Wang, D. J. Treacy, S. Luo, G. P. Schroth, D. E. Housman, S. Reddy, E. Lecuyer, and C. B. Burge, *Transcriptome-wide regulation of pre-mRNA splicing and mRNA localization by muscleblind proteins*, Cell, 150 (2012), pp. 710–24.

[55] H. Wang, C. Krishnan, and G. W. Charville, *INSM1 expression in peripheral neuroblastic tumors and other embryonal neoplasms*, Pediatric and Developmental Pathology, 22 (2019), pp. 440–8.

[56] D. R. Zerbino, P. Achuthan, W. Akanni, M. R. Amode, D. Barrell, J. Bhai, K. Billis, C. Cummins, A. Gall, C. G. Giron, L. Gil, L. Gordon, L. Haggerty, E. Haskell, T. Hourlier, O. G. Izuogu, S. H. Janacek, T. Juettemann, J. K. To, M. R. Laird, I. Lavidas, Z. Liu, J. E. Loveland, T. Maurel, W. McLaren, B. Moore, J. Mudge, D. N. Murphy, V. Newman, M. Nuhn, D. Ogeh, C. K. Ong, A. Parker, M. Patricio, H. S. Riat, H. Schuilenburg, D. Sheppard, H. Sparrow, K. Taylor, A. Thormann, A. Vullo, B. Walts, A. Zadissa, A. Frankish, S. E. Hunt, M. Kostadima, N. Langridge, F. J. Martin, M. Muffato, E. Perry, M. Ruffier, D. M. Staines, S. J. Trevanion, B. L. Aken, F. Cunningham, A. Yates, and P. Flicek, *Ensembl 2018*, Nucleic Acids Research, 46 (2018), pp. D754–61.

Chapitre 6

Discussion générale

Durant les dernières décennies, le trafic subcellulaire des ARN a été identifié et reconnu comme une étape importante de la régulation post-transcriptionnelle des gènes. Bien que la localisation des ARN fasse partie intégrante d'une panoplie de processus biologiques et ait un impact appréciable sur des activités cellulaires polarisées, la prévalence générale de l'asymétrie, à l'échelle transcriptomique, demeurait incertaine. De plus, en dépit des connaissances au sujet de leurs mécanismes médiés par des protéines, les caractéristiques des ARN localisés ainsi que la topologie de motifs de sites de liaison protéique à grande échelle et de leurs conservations entre espèces étaient inconnues. Pour adresser ces questions, j'ai utilisé des approches bio-informatiques pour tirer profit de données privées et publiques de fractionnement biochimique et de motifs de protéines liant l'ARN (RBP). Les études antécédentes de localisation des ARN chez la *Drosophile* ont établi qu'une majorité (>70 %) de quelque 3000 ARN messagers démontraient des patrons de distribution subcellulaires qui corrèlent en partie avec ceux des protéines qu'ils encodent. De plus, plusieurs autres études *in vivo* et *in vitro* ont cartographié les sites de liaison d'une RBP à l'échelle transcriptomique. Cette thèse adresse deux hypothèses principales, soit qu'il y a une prévalence de différents biotypes d'ARN localisés chez les cellules cultivées de mouche et d'humain et qu'il y a des topologies de sites de liaison protéique conservées entre espèces. Les objectifs généraux furent d'identifier la prévalence et les caractéristiques des ARN asymétriquement distribués en plus de définir des éléments de séquences pouvant être impliqués dans la régulation post-transcriptionnelle. En particulier, j'ai premièrement assemblé des collections d'ARN et de protéines asymétriquement distribuées pour évaluer l'ampleur de la localisation et, deuxièmement, j'ai cartographié 223 protéines liant l'ARN établissant 421 133 612 sites de liaison potentiels chez cinq espèces. Mes études ont ainsi permis, respectivement, de globalement conclure qu'une majorité (>80 %) des transcrits chez l'humain et la *Drosophile* sont distribués asymétriquement et que les protéines liant l'ARN montrent des variabilités étalées en gradients continus quant aux nombres de sites de liaison et à leurs sites de distribution. Mes efforts cataloguant la distribution des sites de RBP à l'échelle de transcriptome entier (excluant les introns), la première étude de la sorte, ont permis d'observer des tendances distinctives entres biotypes d'ARN. Notamment, l'ensemble de ces conclusions générales émergent autant lorsque l'on considère soit les ARN codants soit les ARN non codants et les patrons qui les caractérisent semblent évolutivement conservés. Tous ces projets ont été réalisés dans le but conjoint d'éclaircir les logiques de régulation des ARN.

6.1. L'approche CeFra-seq est une méthode à haut débit permettant d'identifier globalement la distribution subcellulaire des ARN

Nous avons développé l'approche CeFra-seq afin d'étendre la caractérisation des ARN asymétriquement distribués pour inclure tous les transcrits quantifiables par séquençage à haut débit. J'ai ainsi pu démontrer que la majorité des ARN ont une distribution asymétrique et que cette prévalence est observable autant chez les ARN codants que non codants (Chapitre 3). De plus, les ARN hautement enrichis dans une fraction subcellulaire ont des attributs distincts. Notamment, nous avons constaté que les ARN enrichis dans différentes régions du cytoplasme présentent des caractéristiques distinctives en terme de longueur totale de leurs régions codantes et 3'UTR. Par exemple, la fraction cytosolique est composée plutôt d'ARN plus courts démontrant une complexité d'exons plus faible et des 3'UTR plus courts, tandis que les ARN enrichis dans les fractions membranaires et insolubles sont plus longs et plus complexes. En outre, la présence de patrons de distribution localisées est corrélée entre compartiments subcellulaires chez les orthologues de l'humain et de la *Drosophile* ce qui suggère que la localisation des ARN est évolutivement conservée. J'ai aussi observé un étalement de sous-groupes ARNm-protéine montrant une colocalisation où les modules de complexes protéiques existent sur un continuum graduel de codistribution avec les ARNm qui les encodent. Cette étude révèle donc que la localisation de l'ARN est répandue et agit à travers des contraintes discriminantes, identifiant ainsi des machineries protéiques dont le ciblage intracellulaire est susceptibles d'être modulé par la traduction localisée.

Cette méthode à haut débit s'inclut parmi l'ensemble des protocoles permettant l'identification de transcrits localisés. Les autres méthodes préalablement établies se basent principalement sur l'imagerie. Par exemple, les méthodes d'hybridation in situ suivies de visualisations par microscopie offrent une résolution plus précise dans l'établissement de la distribution subcellulaire d'un transcrit [18, 11, 43]. Par contre, ces techniques deviennent rapidement laborieuses pour déterminer la distribution de plusieurs transcrits et seraient inenvisageables pour étudier l'ensemble des ARN de plusieurs espèces. Notre méthode de CeFra-seq présente donc un avantage clair pour estimer la distribution du transcriptome complet. Par contre, ceci se fait au détriment d'une résolution détaillée. En effet, notre méthode ne permet pour l'instant que le fractionnement cellulaire en quatre fractions, nucléaire, cytosolique, membranaire, et insoluble. Ces fractions sont le reflet de coefficients de sédimentation de diverses structures subcellulaires obtenues par centrifugations séquentielles. Elles ne correspondent donc pas à des localités juxtaposées. Les ARN y étant recensés ne sont donc pas nécessairement spatialement colocalisés, mais démontrent plutôt des patrons rudimentaires de localisation similaire. Conservant cette approche, il serait intéressant de développer le fractionnement pour inclure d'autres fractions. Par exemple, les mitochondries et autres organelles pourraient être séparées des fractions cytoplasmiques et le noyau pourrait être fractionné pour départager le nucléoplasme, l'euchromatine, les nucléoles, et l'hétérochromatine [36, 22].

Depuis le développement du CeFra-seq, d'autres techniques ne dépendant pas d'un fractionnement cellulaire ont été mises au point dans le but de détecter des ARN localisés et s'inscrivent dans les directions futures de la recherche transcriptomique. Parmi ces approches parallèles, celle dite de l'étiquetage d'ARN de proximité (« Proximity RNA labelling ») par APEX-seq est des plus prometteuse. Cette méthode permet, brièvement, d'utiliser l'ascorbate peroxydase APEX2 pour examiner l'organisation spatiale subcellulaire de transcriptome avec une résolution plus précise [12, 8]. Il serait intéressant d'intersecter les résultats issus des deux types d'analyses. Ceci permettrait de définir de façon plus robuste l'asymétrie du transcriptome et du protéome et fournirait de nouvelles perspectives à propos de l'organisation subcellulaire ARN-protéines. En outre, des techniques novatrices monomolécule ou monocellulaire sont à l'avant plan des technologies futures qui devraient être considérées dans les études de régulation post-transcriptionnelle. Certaines de ces méthodes peuvent être effectuées sur des échantillons fixés ou *in vivo* ce qui permet une visualisation en temps réel du mouvementent des transcrits [**10**, **42**, **1**, **2**].

D'autre part, les méthodes de séquençage et les chaines de processus des données utilisées dans l'approche actuelle du CeFra-seq ont aussi plusieurs limitations. Par exemple, nous avons séquencé nos échantillons avec le système HiSeq 2000 à une profondeur moyenne de 20 millions de « reads », ce qui est raisonnable pour effectuer une lecture simple du niveau d'expression d'un ARN moyen, mais qui peut s'avérer insuffisant pour détecter la présence de transcrits plus faiblement exprimés ou les isoformes distincts de chaque gène. L'utilisation d'un séquenceur plus moderne comme le système NovaSeq 6000, pouvant produire jusqu'à 2,5 milliards de « reads », serait une modification simple, mais appréciable dans la quantification de l'expression localisée de chaque transcrit [**27**, **34**].

En ce qui a trait aux analyses bio-informatiques subséquentes, nous avons effectué l'alignement sur un génome de référence avec TopHat et quantifié l'expression différentielle des transcrits avec DESeq2, les standards établis lorsque ces analyses ont été effectuées, il y a cinq ans [**37**, **38**, **21**]. Des outils d'alignement beaucoup plus précis et rapides, comme STAR, ont depuis été publiés [**5**, **6**]. D'autres algorithmes dits « alignment-free » ou « quasi-mapping » comme salmon, sailfish, ou kallisto se sont rapidement montrés supérieures pour quantifier précisément l'expression au niveau des transcrits, en normalisant pour certains biais expérimentaux [**23**, **24**, **3**]. Par exemple, ces outils pourraient permettre de quantifier la localisation différentielle d'isoformes d'ARN qui possèderaient des 3'UTR alternatifs ou la présence d'un intron retenu qui dicteraient leurs transports [**35**]. Ce type d'analyse s'est avéré infructueux avec les jeux de données que nous avons créés et les outils originalement disponibles tels que rMATS, car ils nécessitent une profondeur de séquençage d'au moins 60-80 millions de « reads » pour distinguer significativement les isoformes [**33**]. Globalement intégré, l'ajout de fractions combinées aux nouvelles technologies permettrait certainement une meilleure caractérisation des échantillons de fractionnement tels que présentés et serait susceptible de révéler des patrons de distribution sur la composition transcriptomique des corps cellulaires et sur la maturation des ARN épissés.

L'approche par CeFra-seq que nous avons développée apparaît comme une étape préalable déterminante, car elle permet de filtrer des groupes de transcrits présentant des tendances de localisation spécifiques communes en les départageant des autres transcrits non localisés.

De façon importante, les conclusions obtenues par CeFra-seq sont en accord avec les études antécédentes et l'idée que la plupart des ARN ont un degré de localisation asymétrique. Une analyse par CeFra-seq peut être centrale à l'élaboration de listes qui dicteraient les meilleurs candidats à analyser par différentes méthodes plus précises afin de consolider les conclusions établies.

6.2. L'élaboration d'une base de données de sites potentiels de liaison de protéines liant l'ARN a permis de caractériser leurs propriétés et organisations topologiques chez l'humain et chez des organismes modèles

Lorsque le projet « oRNAment » fut conceptualisé, aucune ressource centralisée répertoriant les motifs de protéines liant l'ARN et leurs sites de liaison n'existait. Pourtant, à travers les interactions avec la communauté scientifique s'intéressant au sujet, le besoin et la demande étaient évidents. Depuis les dernières décennies, une multitude d'informations furent compilées à propos des protéines de liaison à l'ARN et de leurs cibles. Des efforts cherchant à compiler ces informations en une ressource facilement explorable par la communauté devenaient nécessaires. D'ailleurs, d'autres groupes ont depuis publié des outils semblables, sans toutefois proposer des solutions précalculées et facilement accessibles à la communauté scientifique ni des résultats d'analyses à très grande échelle couvrant des transcriptomes entiers [**25**, **9**, **19**]. Durant le processus de recherche et développement de l'algorithme d'identification d'instance de motifs et de la mise en place de la base de données, l'importance de l'intégration d'une multitude d'angles, comme la structure, l'annotation de coordonnées, ou la topologie, se dessinait et plusieurs hypothèses et questions émergèrent. Alors qu'un des buts principaux habituels d'une création de bases de données est d'établir un moyen pour communiquer la recherche, elle peut aussi servir d'outil de recherche en procurant un point focal permettant d'établir des liens entre les données. C'est pourquoi, lorsque nous avons montré l'efficacité de notre approche sur des données humaines, nous y avons intégré d'autres espèces modèles. Ceci nous a permis de segmenter le projet pour offrir à la communauté scientifique, d'une part, une ressource robuste cataloguant les instances potentielles de sites de liaison RBP dans l'ARN codant et non codant dans diverses espèces en plus d'offrir, d'autre part, un outil nous amenant à caractériser le contexte situationnel et les propriétés de 223 RBP chez cinq espèces.

6.2.1. Fondements d'une base de données essentielle à l'exploration systématique et à la dissémination de résultats

J'ai développé l'approche oRNAment afin de caractériser la distribution des sites de liaison connus de protéines liant l'ARN à des transcriptomes complets, excluant les régions introniques (Chapitre 4). La stratégie employée pour oRNAment repose sur une chaine de processus en plusieurs étapes. Autant durant la création de l'algorithme de recherche de motifs que dans l'établissement de la basse de données et de l'interface web, j'ai investigué différentes approches dans le calcul de similarités entre un motif et une sous-séquence ARN ainsi que dans les moyens de stocker et d'accéder efficacement à la quantité volumineuse de données créées. Les recherches pionnières en ce sens, bien qu'elles s'adressent uniquement aux questions reliées aux facteurs de transcription, sont nombreuses et ont servi de source d'inspiration pour le projet [**13**, **15**, **16**, **32**]. Parmi la panoplie d'algorithmes publiés, une adaptation de MATCH a été déterminée comme la plus appropriée pour effectuer la recherche d'instances de motifs [**13**, **15**, **16**, **32**]. Celui-ci, qui prend en entrée des données sous forme de matrice poids-position (« position weight matrix » (PWM)), est un des fondements de JASPAR, une base de données de facteurs de transcription parmi les plus importantes [**15**, **16**, **32**]. La question demeurait si des motifs de différents RBP pouvaient être similairement cartographiés à l'échelle transcriptomique en un temps raisonnable.

Afin d'avoir des résultats fiables et représentatifs, il est évident que les données entrantes doivent être de haute qualité. Ce critère fut comblé grâce à l'obtention de données issues d'expériences biologiques de RNAcompete et plus récemment de RNA Bind-n-Seq (RBNS), qui font appel à des approches de sélection de sites *in vitro* pour identifier les motifs de liaison prédominants de quelques centaines de RBP [**29**, **30**, **7**, **17**]. Ces données moléculaires s'avèrent précieuses pour la communauté, et plusieurs bases de données furent développées pour les exploiter. Par exemple, les ressources CISBP-RNA, RBPmap, ATtRACT et MotifMap-RNA permettent à un utilisateur d'analyser une séquence d'intérêt afin d'identifier des instances de motifs [**25**, **9**, **19**, **31**]. Cependant, à ce jour, aucune ressource ne permettait de visualiser la distribution de ces motifs dans un transcriptome entier. De plus, ces petites bases de données sont régulièrement le fruit de développement indépendant faisant en sorte qu'ils n'adhèrent pas nécessairement à des standards communs. Par exemple, les données de RNAcompete sont publiques et facilement disponibles sous forme de PWM. D'ailleurs, les ressources similaires à oRNAment reposent uniquement sur celles-ci. Les données provenant de RBNS, bien que publiques, ne sont pas aussi facilement utilisables. En effet, les motifs résultant de l'exécution de leurs scripts n'étaient que disponibles sous forme d'image PNG. Il fut donc nécessaire d'adapter leur code pour qu'il puisse être exécuté sur les grappes de Calcul Canada. Ce point faible m'aura donné l'avantage de pouvoir formater les données issues de l'exécution du script associé au RBNS pour qu'elles soient comparables à celles de RNAcompete. La base de données oRNAment présente un exemple sur comment de telles données peuvent être gérées. La représentation uniforme sous plusieurs formats de toutes les informations et données accessibles sur le portail oRNAment a été pensée afin d'être utilisée par tous les membres de la communauté selon leurs besoins. Par exemple, tous les motifs sont disponibles pour téléchargement en format PWM et tous les résultats de recherches sont accessibles non seulement en format Excel, pour une visualisation rapide, mais aussi en CSV afin d'être facilement intégrés à des chaines de traitement informatique subséquentes et dans le format bed, qui permet une interopérabilité avec d'autres ressources telles que le « UCSC genome browser » ou les outils d'analyses de bedtools [**26**, **14**].

Un élément essentiel dans la production et la valorisation d'une base de données prédictive comme oRNAment tient à sa comparaison et validation des résultats avec des données issues d'expériences biologiques. Actuellement, oRNAment ne considère que la liaison aux exons, car les ressources de calcul nécessaires pour analyser les motifs à travers l'espace des séquences introniques étaient insuffisantes lors de sa création. Par conséquent, j'ai axé les analyses comparatives entre les sites de liaison prédits par oRNAment et les signatures exoniques observées par eCLIP [**40**, **41**]. Autrement dit, toutes les validations ont été effectuées dans les mêmes espaces transcriptomique. De façon importante, j'ai aussi procédé de la sorte pour les RBP liants préférablement les introns, comme *hnRNPC*. J'ai observé dans les données d'eCLIP que près de 11 % des sites de liaison cartographiés par ce facteur étaient exoniques. En outre, j'ai observé que ces résultats corrélaient bien entre eux et que les concordances générales entre oRNAment et eCLIP sont maintenues lorsqu'est considérée une RBP démontrant une prédominance de liaison exonique ou intronique. Par contre, il est important de noter que les données comparatives biologiques obtenues par eCLIP, bien que méthode étalon, ne sont pas absolues et ont leurs propres limitations majeures. Par exemple, elles nécessitent des anticorps validés de haute qualité et sont dépendantes du type cellulaire utilisé. De plus, les outils informatiques permettant l'identification de sites de liaison (« peak-caller ») peuvent montrer des biais pour les séquences riches en GC. D'ailleurs, les taux de reproductibilité sont estimés à 60 % [40, 41]. Un autre facteur important à considérer est que les motifs décrits par les analyses biologiques distinctes comme RNAcompete, RBNS et eCLIP ne concordent pas forcément et peuvent même être considérablement variables. En outre, il est aussi important de souligner que les motifs de liaison identifiés par RNAcompete et RBNS ne sont pas nécessairement ceux que la protéine manifeste dans tous les contextes cellulaires. En effet, les sites de liaison réels peuvent être influencés par le milieu cellulaire de différentes façons. Par exemple les liaisons ARN-protéine peuvent être modulées par des modifications post-traductionnelles sur la protéine, la présence de partenaires synergiques, la compétition pour les sites de liaison par d'autres protéines ou des caractéristiques des ARN telles que leur structure secondaire ou la présence de sites édités. Finalement, toutes les analyses comparatives ont été effectuées en conservant le motif prédit selon le seuil maximal pour oRNAment et le seuil publié pour le eCLIP. Ces seuils étant inévitablement subjectifs, l'utilisation de valeurs différentes se serait soldée par des estimations variables de la fiabilité des prédictions. Il est manifestement impossible, dans un temps raisonnable, d'optimiser ces valeurs pour chacun des RBP individuellement. Ces faiblesses devraient être considérées lors de l'interprétation de la validation de l'exactitude et de la précision des prédictions. Nous postulons que dans certains cas, les prédictions fournies par oRNAment peuvent être plus fiables que les données eCLIP où plusieurs instances de motifs potentiels pourraient émerger à la suite d'expériences effectuées dans d'autres modèles cellulaires. Une intégration comparative et cohérente de toutes les méthodologies biologiques et informatiques définissant des motifs de RBP constituerait une tâche monumentale, mais s'avèrerait certainement une source d'information profitable pour la communauté.

La compilation des procédés formant oRNAment forme une ressource intégrée centrale qui sert à la fois à stocker des données à propos des sites potentiels de liaison de protéines liant l'ARN et à conserver de nombreux types d'informations différents caractérisant une panoplie de propriétés connexes. Je présente ainsi la première ressource de recherche d'instances de motifs non seulement basée sur des seuils individuels pour chaque protéine, mais qui offre aussi la structure prédite, qui permet une recherche selon plusieurs facettes et qui est complémentée par une analyse biologique comparative. De plus, la ressource oRNAment est « transcrit-centrique » ce qui implique que deux transcrits, pouvant être de biotype différent, mais partageant les mêmes coordonnées génomiques, seront clairement décris de façon distincte. Puisqu'il est envisageable qu'une protéine lie cette sous-séquence dans les différents transcrits, ceci démontre bien l'utilité de la base de données pour un utilisateur désirant étudier, par exemple, l'effet éponge de certains types d'ARN.

Il serait intéressant d'ajouter quelques fonctionnalités à une version prochaine de oRNAment. Par exemple, RBNS fut conçu à partir de séquences de 150 nucléotides afin de conserver la structure locale à proximité du site de liaison. Lorsque j'ai exécuté le code pour RBNS, je me suis assuré de produire systématiquement ces données. Les données prédites structurales sont donc disponibles pour les 78 RBP provenant de RBNS. Il serait certainement profitable d'en faire usage. De plus, une approche optimisée de RNAcompete, nommée RNAcompete-S, qui intègre les informations structurales dans l'identification de sites de liaison a été proposée depuis la création de oRNAment [4]. La prochaine version pourrait donc inclure non seulement une recherche de similarité entre un PWM et une sous-séquence, mais aussi entre deux structures prédites. Pour l'instant seule RNAforester, de la suite ViennaRNA, permet d'aligner ainsi deux structures [20]. Malheureusement, cet algorithme ne semble pas pouvoir être facilement mis à l'échelle pour être exécuté sur une quantité aussi importante de données. Une telle adaptation est non triviale. Par contre, développer un tel algorithme résulterait en une meilleure résolution de la prédiction des sites de liaison, car elle permettrait d'éliminer plusieurs faux positifs et ajouterait une distinction possible entre les RBP liant préférentiellement la séquence ou la structure d'un transcrit. Il serait aussi digne d'intérêt d'exécuter oRNAment sur le transcriptome d'autres espèces et sur des génomes ancestraux prédits. Ceci permettrait d'établir l'histoire évolutive complète des gains et des pertes de motifs de RBP et apporterait un éclairage et un nouvel angle sur leur importance.

6.2.2. La cartographie à l'échelle transcriptomique d'un grand nombre de protéines liant l'ARN chez plusieurs espèces permet de faire émerger des notions déterminantes de la biologie des ARN

J'ai mis à profit les données cataloguées par oRNAment afin de caractériser les sites de liaison potentiels d'un large éventail de RBP et établir leur conservation évolutive entre cinq espèces (Chapitre 5). Grâce à cette approche de profilage globale, combinée à des analyses comparatives précises, j'ai démontré que les RBP présentent de grandes variabilités quant à leur nombre de sites de liaison potentiels dans différentes régions d'un ARN codants et chez divers types d'ARN non codants. J'ai aussi démontré que certaines RBP ont des sites de liaison dans des régions structurées ou non structurées de façon prédominante. L'analyse effectuée chez cinq espèces a permis de démontrer que ces patrons de distribution y sont observés de façon similaire. D'ailleurs, le nombre global de sites potentiels de liaison observés pour chaque type d'ARN non codant ou pour chaque région d'un ARN codant montrent des corrélations générales entre les espèces. De plus, des analyses approfondies de la conservation et de la divergence des motifs positionnés à des coordonnées génomiques comparables entre espèces montrent que près de 50 % des sites de liaison potentiels pour un RBP ont une correspondance pour le même RBP au site orthologue. Distinctement, certains sites de liaison potentiels ont, au contraire, une divergence de RBP au site orthologue, suggérant une compensation permettant une préservation de la fonction. En outre, plusieurs ARN codants présentent des instances multiples de la même RBP, instances qui peuvent être espacées de façon régulière et qui peuvent être conservées entre orthologues. Finalement, j'ai démontré le potentiel de l'approche oRNAment pour identifier de nouvelles fonctions pour des RBP ou des ARN d'intérêt en démontrant que l'ARN non codant *HELLPAR* est fortement enrichie en sites de liaison potentiels pour une panoplie de RBP, lui conférant une fonction présumée d'éponge. L'ARN long non codant *HELLPAR* est associé au syndrome de Hellp, une complication émergeant durant la grossesse et qui est caractérisée par une hémolyse, une élévation des enzymes hépatiques et une faible numération plaquettaire [**28**, **39**]. Puisque *HELLPAR* contient un enrichissement de sites de liaison pour des RBP impliqués dans la stabilisation des ARN, il serait intéressant d'étudier de façon expérimentale les implications mécanistiques associées à un défaut de ce gène. Cette étude révèle donc une approche efficace pour décrire de manière complète la position des sites de liaison potentiels de RBP ainsi que leurs caractéristiques de distribution inhérentes.

La combinaison de chaines de processus et de bases de données pouvant gérer de large quantité de données proposée par oRNAment s'inclut parmi l'ensemble des méthodes permettant l'identification de sites potentiels de liaison ARN-protéine et leur diffusion. Ces méthodes sont décrites en détail à la section 6.2.1. De façon surprenante, les études proposant des approches bio-informatiques pour identifier des sites de liaison semblent toujours omettre de les utiliser non seulement pour décrire systématiquement les sites de liaison potentiels, mais aussi pour faire l'étude des interactions ARN-protéine chez une variété d'espèces. L'approche oRNAment a donc l'avantage de s'inscrire plutôt parmi les études biologiques, comme l'eCLIP, profilant à grande échelle des sites de liaison des RBP et décrivant leur contexte individuel chez une espèce [**40**, **41**]. L'étude comparative proposée à partir de la base de données oRNAment est donc la première à offrir une analyse complète à propos des sites potentiels de liaison RBP entre espèces. Ceci m'a procuré l'avantage claire de pouvoir effectuer une analyse de la conservation et de la divergence de la présence de sites de liaison entre orthologues afin d'établir une certaine histoire évolutive. Par contre, cette étude se base uniquement sur des données prédictives et a inévitablement un taux non négligeable de faux positifs. J'ai cependant pu établir qu'il y a une correspondance adéquate avec des données biologiques. Conservant cette approche, il serait intéressant de développer l'étude comparative en se basant, en plus des données oRNAment, sur des données d'autres types de CLIP effectués chez l'humain et des espèces modèles pour inclure un plus grand ensemble de données provenant d'expérimentations biologiques [**44**].

Chapitre 7

Conclusion et perspectives

La recherche présentée dans cette thèse relate le développement, à l'échelle transcriptomique, d'approches novatrices dans la caractérisation systématique de, premièrement, la distribution asymétrique des ARN par CeFra-seq et, deuxièmement, la distribution de sites potentiels de liaison ARN-protéine par oRNAment. J'ai ainsi démontré comment des approches bio-informatiques originales permettent la synthèse de données biologiques à grande échelle et apporte de nouvelles perspectives sur la localisation des ARN. J'ai aussi démontré comment des données de sites de liaison de protéines peuvent être organisées et intégrées pour accroître notre compréhension à propos de notions déterminantes de la biologie des ARN. Les approches développées et détaillées dans cette thèse, soit une méthode d'analyse d'estimation de la prévalence générale de l'asymétrie des ARN et une base de données de sites potentiels de liaison de protéines liant l'ARN ont été mises à la disposition de la communauté scientifique et seront sans doute utiles pour des études futures.

Il reste encore beaucoup de travail à faire et mes recherches ne représentent qu'une infime partie des efforts nécessaires dans la compréhension de la régulation post-transcriptionnelle. Parmi les directions futures que pourrait prendre la recherche transcriptomique, il serait intéressant d'utiliser l'approche CeFra-seq dans des contextes de perturbation pour comprendre, par exemple, l'impact d'un stress ou d'une déplétion de facteurs RBP sur la distribution du transcriptome.

De plus, il serait certainement profitable d'améliorer oRNAment en modifiant l'algorithme afin qu'il soit assez efficace pour pouvoir inclure dans l'analyse toutes les séquences introniques de chaque espèce. De plus, il sera important d'intégrer les nouveaux motifs de RBP à mesure qu'ils seront identifiés ainsi que d'exécuter le script sur une plus grande variété d'espèces. Ensuite, il serait intéressant de développer les analyses de topologies de motifs afin d'y inclure plusieurs RBP et l'ensemble de leurs combinaisons possibles. En outre, il serait incontournable d'ajouter un angle régulatoire qui évaluerait la distribution de motifs de protéines avec des fonctions connues. Il serait possible de contraster les sites de liaison de protéines avec des données publiques variées telles que des sites de liaison de miRNA ou des analyses d'expression différentielle suite à une déplétion d'une RBP par RNAi. Finalement, la combinaison des données de CeFra-seq et de oRNAment pourrait permettre d'identifier des ARN localisés possédant des enrichissements de sites de liaison pour un ou des sous-groupes de RBP ce qui apporterait de nouvelles perspectives sur les logiques de régulation des ARN.

Références

[1] M. Batish, A. Raj, and S. Tyagi, *Single molecule imaging of RNA in situ*, Methods in Molecular Biologyl, 714 (2011), pp. 3–13.

[2] D. P. Bratu, B. J. Cha, M. M. Mhlanga, F. R. Kramer, and S. Tyagi, *Visualizing the distribution and transport of mRNAs in living cells*, Proceedings of the National Academy of Sciences of the United States of America, 100 (2003), pp. 13308–13.

[3] N. L. Bray, H. Pimentel, P. Melsted, and L. Pachter, *Near-optimal probabilistic RNA-seq quantification*, Nature Biotechnology, 34 (2016), pp. 525–7.

[4] K. B. Cook, S. Vembu, K. C. H. Ha, H. Zheng, and K. U. Laverty, *RNAcompete-S : Combined RNA sequence/structure preferences for RNA binding proteins derived from a single-step in vitro selection*, Methods, 15 (2017), pp. 18–28.

[5] A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras, *STAR : ultrafast universal RNA-seq aligner*, Bioinformatics, 29 (2013), pp. 15–21.

[6] A. Dobin and T. R. Gingeras, *Mapping RNA-seq reads with STAR*, Current Protocols in Bioinformatics, 51 (2015), pp. 11 14 1–19.

[7] D. Dominguez, P. Freese, M. S. Alexis, A. Su, M. Hochman, T. Palden, C. Bazile, N. J. Lambert, E. L. Van Nostrand, G. A. Pratt, G. W. Yeo, B. R. Graveley, and C. B. Burge, *Sequence, structure, and context preferences of human RNA binding proteins*, Molecular Cell, 70 (2018), pp. 854–67 e9.

[8] F. M. Fazal, S. Han, K. R. Parker, P. Kaewsapsak, J. Xu, A. N. Boettiger, H. Y. Chang, and A. Y. Ting, *Atlas of subcellular RNA localization revealed by APEX-Seq*, Cell, 178 (2019), pp. 473–90 e26.

[9] G. Giudice, F. Sánchez-Cabo, C. Torroja, and E. Lara-Pezzi, *ATtRACT—a database of RNA-binding proteins and associated motifs*, Database, (2016).

[10] R. Henriques, C. Griffiths, E. Hesper Rego, and M. M. Mhlanga, *PALM and STORM : unlocking live-cell super-resolution*, Biopolymers, 95 (2011), pp. 322–31.

[11] H. Jambor, V. Surendranath, A. T. Kalinka, P. Mejstrik, S. Saalfeld, and P. Tomancak, *Systematic imaging reveals features and changing localization of mRNAs in Drosophila development*, eLife, 4 (2015).

[12] P. Kaewsapsak, D. M. Shechner, W. Mallard, J. L. Rinn, and A. Y. Ting, *Live-cell mapping of organelle-associated RNAs via proximity biotinylation combined with protein-RNA crosslinking*, elife, 6 (2017).

[13] A. E. Kel, E. Gossling, I. Reuter, E. Cheremushkin, O. V. Kel-Margoulis, and E. Wingender, *MATCH : A tool for searching transcription factor binding sites in DNA sequences*, Nucleic Acids Research, 31 (2003), pp. 3576–9.

[14] W. J. Kent, C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler, and D. Haussler, *The human genome browser at UCSC*, Genome Research, 12 (2002), pp. 996–1006.

[15] A. Khan, O. Fornes, A. Stigliani, M. Gheorghe, J. A. Castro-Mondragon, R. van der Lee, A. Bessy, J. Cheneby, S. R. Kulkarni, G. Tan, D. Baranasic, D. J. Arenillas, A. Sandelin, K. Vandepoele, B. Lenhard, B. Ballester, W. W. Wasserman, F. Parcy, and A. Mathelier, *JASPAR 2018 : update of the open-access database of transcription factor binding profiles and its web framework*, Nucleic Acids Research, 46 (2018), p. D1284.

[16] A. Khan and A. Mathelier, *JASPAR RESTful API : accessing JASPAR data from any programming language*, Bioinformatics, (2017).

[17] N. J. Lambert, A. D. Robertson, and C. B. Burge, *RNA Bind-n-Seq : Measuring the binding affinity landscape of RNA-binding proteins*, Methods Enzymology, 558 (2015), pp. 465–93.

[18] E. Lecuyer, H. Yoshida, N. Parthasarathy, C. Alm, T. Babak, T. Cerovina, T. R. Hughes, P. Tomancak, and H. M. Krause, *Global analysis of mRNA localization reveals a prominent role in organizing cellular architecture and function*, Cell, 131 (2007), pp. 174–87.

[19] Y. Liu, S. Sun, T. Bredy, M. Wood, R. C. Spitale, and P. Baldi, *MotifMap-RNA : a genome-wide map of RBP binding sites*, Bioinformatics, 33 (2017), pp. 2029–31.

[20] R. Lorenz, S. H. Bernhart, C. Honer Zu Siederdissen, H. Tafer, C. Flamm, P. F. Stadler, and I. L. Hofacker, *ViennaRNA package 2.0*, Algorithms for Molecular Biology, 6 (2011), p. 26.

[21] M. I. Love, W. Huber, and S. Anders, *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2*, Genome Biology, 15 (2014), p. 550.

[22] L. M. Orre, M. Vesterlund, Y. Pan, T. Arslan, Y. Zhu, A. Fernandez Woodbridge, O. Frings, E. Fredlund, and J. Lehtio, *SubCellBarCode : Proteome-wide mapping of protein localization and relocalization*, Molecular Cell, 73 (2019), pp. 166–82 e7.

[23] R. Patro, G. Duggal, M. I. Love, R. A. Irizarry, and C. Kingsford, *Salmon provides fast and bias-aware quantification of transcript expression*, Nature Methods, 14 (2017), pp. 417–19.

[24] R. Patro, S. M. Mount, and C. Kingsford, *Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms*, Nature Biotechnology, 32 (2014), pp. 462–4.

[25] I. Paz, I. Kosti, M. Ares Jr, and M. Cline, *RBPmap : a web server for mapping binding sites of RNA-binding proteins*, Nucleic Acids Research, 42 (2014), pp. W361–7.

[26] A. R. Quinlan and I. M. Hall, *BEDTools : a flexible suite of utilities for comparing genomic features*, Bioinformatics, 26 (2010), pp. 841–2.

[27] A. Raine, U. Liljedahl, and J. Nordlund, *Data quality of whole genome bisulfite sequencing on Illumina platforms*, PLoS One, 13 (2018), p. e0195972.

[28] N. Rappaport, M. Twik, I. Plaschkes, R. Nudel, T. Iny Stein, J. Levitt, M. Gershoni, C. P. Morrey, M. Safran, and D. Lancet, *MalaCards : an amalgamated human disease compendium with diverse clinical and genetic annotation and structured search*, Nucleic Acids Research, 45 (2017), pp. D877–87.

[29] D. Ray, K. C. H. Ha, K. Nie, H. Zheng, T. R. Hughes, and Q. D. Morris, *RNAcompete methodology and application to determine sequence preferences of unconventional RNA-binding proteins*, Methods, 118-9 (2017), pp. 3–15.

[30] D. Ray, H. Kazan, E. T. Chan, L. Pena Castillo, S. Chaudhry, S. Talukder, B. J. Blencowe, Q. Morris, and T. R. Hughes, *Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins*, Nature Biotechnology, 27 (2009), pp. 667–70.

[31] D. Ray, H. Kazan, K. B. Cook, M. T. Weirauch, H. S. Najafabadi, X. Li, S. Gueroussov, M. Albu, H. Zheng, A. Yang, H. Na, M. Irimia, L. H. Matzat, R. K. Dale, S. A. Smith, C. A. Yarosh, S. M. Kelly, B. Nabet, D. Mecenas, W. Li, R. S. Laishram, M. Qiao, H. D. Lipshitz, F. Piano, A. H. Corbett, R. P. Carstens, B. J. Frey, R. A. Anderson, K. W. Lynch, L. O. Penalva, E. P. Lei, A. G. Fraser, B. J. Blencowe, Q. D. Morris, and T. R. Hughes, *A compendium of RNA-binding motifs for decoding gene regulation*, Nature, 499 (2013), pp. 172–7.

[32] A. Sandelin, W. Alkema, P. Engstrom, W. W. Wasserman, and B. Lenhard, *JASPAR : an open-access database for eukaryotic transcription factor binding profiles*, Nucleic Acids Research, 32 (2004), pp. D91–4.

[33] S. Shen, J. W. Park, Z. X. Lu, L. Lin, M. D. Henry, Y. N. Wu, Q. Zhou, and Y. Xing, *rMATS : robust and flexible detection of differential alternative splicing from replicate RNA-seq data*, Proceedings of the National Academy of Sciences of the United States of America, 111 (2014), pp. E5593–601.

[34] G. A. C. Singer, N. A. Fahner, J. G. Barnes, A. McCarthy, and M. Hajibabaei, *Comprehensive biodiversity analysis via ultra-deep patterned flow cell technology : a case study of eDNA metabarcoding seawater*, Scientific Reports, 9 (2019), p. 5991.

[35] J. M. Taliaferro, M. Vidaki, R. Oliveira, S. Olson, and L. Zhan, *Distal alternative last exons localize mRNAs to neural projections*, Molecular cell, 61 (2016), pp. 821–833.

[36] J. R. Tata and B. Baker, *Sub-nuclear fractionation. I. procedure and characterization of fractions*, Experimental Cell Research, 83 (1974), pp. 111–24.

[37] C. Trapnell, L. Pachter, and S. L. Salzberg, *TopHat : discovering splice junctions with RNA-seq*, Bioinformatics, 25 (2009), pp. 1105–11.

[38] C. Trapnell, A. Roberts, L. Goff, G. Pertea, D. Kim, D. R. Kelley, H. Pimentel, S. L. Salzberg, J. L. Rinn, and L. Pachter, *Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks*, Nature Protocols, 7 (2012), pp. 562–78.

[39] M. van Dijk, H. K. Thulluru, J. Mulders, O. J. Michel, A. Poutsma, S. Windhorst, G. Kleiverda, D. Sie, A. M. Lachmeijer, and C. B. Oudejans, *HELLP babies link a novel lincRNA to the trophoblast cell cycle*, Journal of Clinical Investigation, 122 (2012), pp. 4003–11.

[40] E. L. Van Nostrand, P. Freese, G. A. Pratt, X. Wang, X. Wei, R. Xiao, S. M. Blue, J.-Y. Chen, N. A. Cody, D. Dominguez, S. Olson, B. Sundararaman, L. Zhan, C. Bazile, L. P. B. Bouvrette, J. Bergalet, M. O. Duff, K. E. Garcia, C. Gelboin-Burkhart, M. Hochman, N. J. Lambert, H. Li, T. B. Nguyen, T. Palden, I. Rabano, S. Sathe, R. Stanton, A. Su, R. Wang, B. A. Yee, B. Zhou, A. L. Louie, S. Aigner, X.-d. Fu, E. Lécuyer, C. B. Burge, B. R. Graveley, and G. W. Yeo, *A large-scale binding and functional map of human RNA binding proteins*, bioRxiv, (2018), p. 179648.

[41] E. L. Van Nostrand, G. A. Pratt, A. A. Shishkin, C. Gelboin-Burkhart, M. Y. Fang, B. Sundararaman, S. M. Blue, T. B. Nguyen, C. Surka, K. Elkins, R. Stanton, F. Rigo, M. Guttman, and G. W. Yeo, *Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP)*, Nature Methods, 13 (2016), pp. 508–14.

[42] J. H. Wilbertz, F. Voigt, I. Horvathova, G. Roth, Y. Zhan, and J. A. Chao, *Single-molecule imaging of mRNA localization and regulation during the integrated stress response*, Molecular Cell, 73 (2019), pp. 946–58 e7.

[43] R. Wilk, J. Hu, D. Blotsky, and H. M. Krause, *Diverse and pervasive subcellular distributions for both coding and long noncoding RNAs*, Genes and Development, 30 (2016), pp. 594–609.

[44] Y. C. Yang, C. Di, B. Hu, M. Zhou, Y. Liu, N. Song, Y. Li, J. Umetsu, and Z. J. Lu, *Clipdb : a clip-seq database for protein-rna interactions*, BMC Genomics, 16 (2015), p. 51.