博士論文

Studies on ribosome-associated lncRNAs using ribosome profiling data

(リボソームプロファイリングデータを用いたリボソーム関連

長鎖ノンコーディングRNAに関する研究)

曽　超

Chao Zeng

Studies on ribosome-associated lncRNAs using ribosome profiling data

(リボソームプロファイリングデータを用いた
リボソーム関連ノンコーディングRNAに関する研究)

曽　超 (Chao Zeng)

Supervisor: Prof. Kiyoshi Asai
Advisor: Prof. Michiaki Hamada

Graduate School of Frontier Sciences
The University of Tokyo

This dissertation is submitted for the degree of
Doctor of Philosophy

June 2018

Acknowledgements

First and foremost, I would like to express my respect and sincere gratitude to my supervisor Prof. Kiyoshi Asai for the continuous support of my Ph.D. study and related research, for his motivation, patience, and wide range of knowledge. It has been an honor to be his first Chinese student.

A very special appreciation goes out to Prof. Michiaki Hamada, who has guided me in all the time of research and writing of this thesis. Without his meticulous guidance and precious support, it would not be possible to conduct this research.

I also would like to express my special thanks of gratitude to the rest of my thesis committee: Prof. Hisanori Kiryu, Prof. Wataru Iwasaki, Prof. Shintaro Iwasaki, and Prof. Tetsuo Shibuya, for their insightful comments and suggestions, but also for the hard question which incented me to widen my research from various perspectives.

With a special mention to Dr. Tsukasa Fukunaga who is one of the co-authors of the journal manuscript. Thank you for your valuable discussions and helpful comments on the data analyses.

My sincere thanks also go to Prof. Osamu Gotoh, Dr. Junichi Iwakiri, Dr. Anish Man Singh Shrestha, Dr. Goro Terai, Prof. Martin Frith, Dr. Yutaka Saito, Prof. Nobuyoshi Akimitsu, Prof. Riu Yamashita, Dr. Masahiro Onoguchi and Dr. Anju Pratap who provided me valuable feedback. Their comments helped me a lot to conduct research in depth. And finally, I am grateful to all past and present members of Asai laboratory who gave me warm encouragements. My very special thanks and love go to my family for their support, understanding, love and tolerance.

Abstract

Although the number of discovered long non-coding RNAs (lncRNAs) has increased dramatically, their biological roles have not been established. Many recent studies have used ribosome profiling data to assess the protein-coding capacity of lncRNAs. However, very little work has been done to identify ribosome-associated lncRNAs, here defined as lncRNAs interacting with ribosomes related to protein synthesis as well as other unclear biological functions.

On average, 39.17% of expressed lncRNAs were observed to interact with ribosomes in human and 48.16% in mouse. We developed the ribosomal association index (RAI), which quantifies the evidence for ribosomal associability of lncRNAs over various tissues and cell types, to catalog 691 and 409 lncRNAs that are robustly associated with ribosomes in human and mouse, respectively. Moreover, we identified 78 and 42 lncRNAs with a high probability of coding peptides in human and mouse, respectively. Compared with ribosome-free lncRNAs, ribosome-associated lncRNAs were observed to be more likely to be located in the cytoplasm and more sensitive to nonsense-mediated decay. Furthermore, we tried to investigate the sequence features involved in the ribosomal association of lncRNA. We have extracted ninety-nine sequence features corresponding to different biological mechanisms (i.e., RNA splicing, putative ORF, k-mer frequency, RNA modification, RNA secondary structure, and repeat element). An $\mathcal{L}1$-regularized logistic regression model was applied to select these features. Finally, we obtained fifteen and nine important features for the ribosomal association of human and mouse lncRNAs, respectively.

To our knowledge, this is the first study to characterize ribosome-associated lncRNAs and ribosome-free lncRNAs from the perspective of sequence features. These sequence features that were identified in this study may shed light on the biological mechanism of the ribosomal association and provide important clues for functional analysis of lncRNAs. Our results suggest that RAI can be used as an integrative and evidence-based tool for distinguishing between ribosome-associated and free lncRNAs, providing a valuable resource for the study of lncRNA functions.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

1.1 Long noncoding RNAs

1.1.1 A brief history of lncRNAs

The knowledge of non-coding RNAs arises from a scientific understanding of non-coding sequences in the genome. As early as the 1970s scientists took note of the non-coding sequences and called them “junk DNAs” [1]. Transcripts from non-coding sequences, such as heterogeneous nuclear RNA (hnRNA [2]), have been discovered since the 1970s. In the 1980s, small nuclear RNA (snRNA [3]) and small nucleolar RNA (snoRNA, reviewed in [4]) were also discovered. This shows that the non-coding sequences in the genome are carrying information and they should have biological functions. Scientists have gained a more comprehensive understanding from the Human Genome Project (HGP) that began in the 1990s [5, 6]. This project led to the determination of the complete genome sequence of many species, thereby understanding the composition and structure of non-coding sequences in the genome. In the 21st century, with the advancement of the transcriptome research and the implementation of the ENCODE (Encyclopedia of DNA Elements) project, it was found that the vast majority of the genome is pervasively transcribed into tens of thousands of non-coding transcripts [7]. In the past decade, the MiTranscriptome project used RNA-seq data from various cancer tissues to obtain 91,013 expressed genes, of which over 68% (58,648) of genes were classified as long noncoding RNAs (lncRNAs) [8]. Meanwhile, the FANTOM (Functional ANnoTation Of the Mammalian genome) project used CAGE (cap analysis of gene expression) data from different primary cell types and tissues and detected 27,919 human lncRNAs [9]. So far, the number of lncRNAs has far exceeded the number of protein-coding mRNAs.

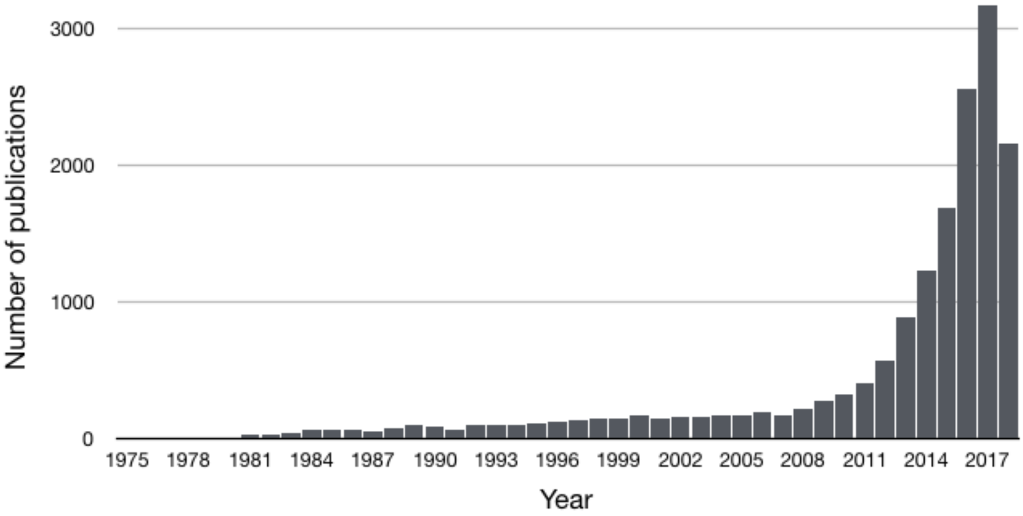


Fig. 1.1 **Number of Pubmed results for the keywords “lncRNA” or “long noncoding RNA” or “long non-coding RNA”.** Data accessed on 24th June 2018.

Is lncRNA a universal transcript or a functional element? It was initially a highly controversial problem. Due to the poor sequence conservation and low expression levels among model organisms, lncRNAs were considered as products of low-fidelity polymerases and without exact functionality [10]. However, this hypothesis has been ruled out by more and more depth sequencing analysis. For instance, the promoter region and the splicing site of lncRNAs have certain similarities to protein-coding genes [11]. Although the sequence conservation of lncRNAs is lower relative to that of mRNAs, lncRNAs to achieve their function may not rely on stringent sequence conservation but instead on RNA secondary structure [12]. The number of studies aimed at identifying lncRNAs and deciphering their roles and functions has dramatically increased (Fig. 1.1). This effort led to a more extensive annotation of their genomic organization and features as well as to a better understanding of their role in various biological processes, which span from the regulation of embryonic development to pathological conditions such as cancer [13–16].

1.1.2 Challenges in analysis of lncRNA

There are two challenging problems that scientists faced when analyzing the lncRNA population were how to annotate their genomic locations and structures accurately, and how to perform functional analysis and characterization of large numbers of lncRNAs.

The annotation of long non-coding RNAs depends on the genome-wide screening of transcripts that mainly annotated as non-coding sequences, such as the lack of an active open reading frame. Scientists are still faced with many challenges when annotating lncRNA genes in the genome. There are at least three reasons why are lncRNAs challenging to annotate. First, lncRNAs have low expression levels, implying that their transcripts will be weakly detected in any unbiased transcriptomic data, including expressed sequence tags (ESTs), RNA-seq and CAGE data. Second, lncRNAs tend to be weakly conserved during evolution [17], making it challenging to identify their homologs by sequence similarity. Third, our understanding of the relationship between lncRNA sequence and its function is insufficient. Sequence features or functional elements cannot be immediately used to identify novel lncRNAs. In contrast, we can use the ORF sequences to distinguish mRNAs from the transcriptome. There are several different annotation databases for the human genome. They are based on two main strategies of automatic and manual annotation. Automatic annotation (MiTranscriptome [8], Human Body Map 2.0 [18], FANTOM CAT [9], BIGTranscriptome [19]) usually applies a fast and economical method of transcription assembly, but it leads to incomplete and inaccurate annotations. Manual annotation (GENCODE [20], RefSeq [21]) generates high-quality lncRNAs, but it is slow and requires substantial long-term financial support.

Due to the majority of lncRNAs are still interpreted as having an unknown function, identification of lncRNA functions has become the most challenging problem. In general, there are two ways to study the functions of lncRNA: experimental and computational. Experimental approaches such as gene knockdown, knockout, overexpression, or editing are considered as golden-standard methods for the investigation of lncRNA functionality. Unfortunately, in consideration of both technical difficulties and limited resources in the form of time and money, these approaches are only suitable for the analysis of what is usually a limited pool of candidates. Alternatively, computational approaches can provide predictive biological functions for lncRNAs in genome-scale. For instance, with an amount of gene expression data across species, tissues and biological conditions in various public databases, it is possible to predict lncRNA functions based on the information from co-expressed transcripts [22].

1.2 Ribosome profiling

1.2.1 Basic principle of ribosome profiling

Ingolia and colleagues developed ribosome profiling technique to obtain a genome-wide snapshot of actively translating ribosomes on the transcriptome [23]. The core of this method is that a translating ribosome can protect approximate 30 nucleotides of mRNA from ribonuclease treatment [24]. These fragments (termed ribosome footprints) are subjected to deep sequencing to provide the position of the ribosome in a sub-codon revolution. As shown in Fig. 1.2, ribosome profiling is typically conducted on a split sample, with parallel libraries prepared for measuring RNA abundance by RNA-seq. The principle of ribosome profiling technique is as follows:

(1) Treatment of the cultured cells with cycloheximide (commonly used for eukaryotes) or chloramphenicol (for bacteria), which are inhibitors of translation elongation, making the translating ribosomes to freeze on the mRNAs. Then obtaining the cellular extract;

(2) Obtaining the cell extract and treating with ribonuclease to digest naked mRNA fragments. Sucrose density gradient centrifugation is used to separate ribosomes and protected mRNA fragments (ribosome footprints). Ribosome footprints are quantitatively transformed into cDNA libraries that could be deeply sequenced;

(3) Alignment of ribosome footprints to the transcriptome, which is derived from the known gene annotations corresponding reference genome sequence.

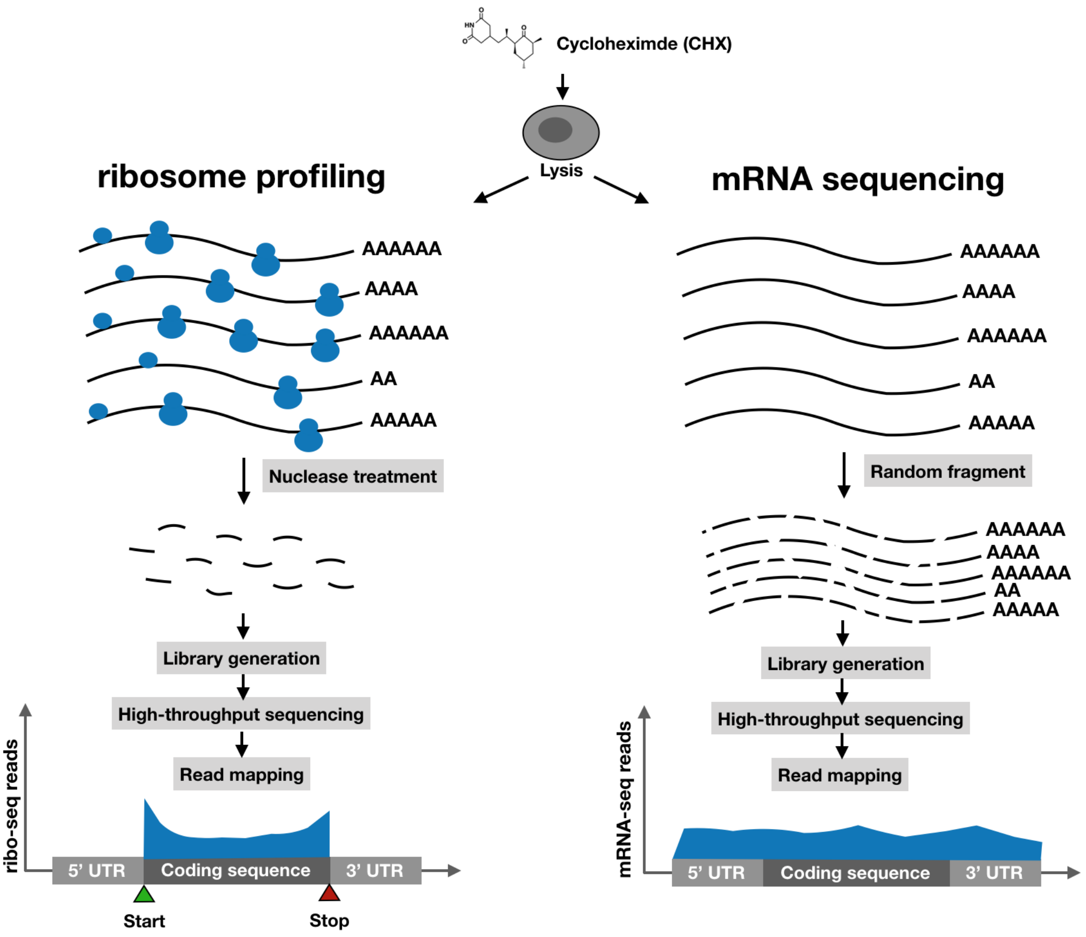


Fig. 1.2 **An overview of ribosome profiling technique.**

1.2.2 Applications and limitations of ribosome profiling

Ribosome profiling has rapidly become a widely used tool for studying diverse and complex biological problems. This technique can be applied to several research scenarios.

First, ribosome profiling can be applied to characterize translation efficiency or differential translation and measure protein abundance. Combining ribosome profiles and corresponding transcript abundances can be used for the characterization of protein translation efficiency (i.e., protein synthesis rate). Ingolia et al. found that the difference in translation efficiency between yeast proteins can be more than 100 times [23]. Analysis of protein translation changes during the differentiation of mouse embryonic stem cells into embryoid bodies revealed a large number of translational pause sites and unannotated translation products, as well as large uORF translation differences during differentiation [25]. In addition, due to the presence and complexity of translational regulation, it is difficult to accurately predict protein expression levels based on mRNA expression levels. Therefore, the use of translation efficiency may be more accurate in predicting protein abundance [26].

Second, analysis of whole genome-wide ribosome profiles facilitates us to define the proteome more precisely and sensitively. Using the ribosome profile, we can find some short-length translatable open reading frames, called upstream open reading frames (uORF), which refer to short translations that exist upstream of the coding region and may be involved in translational regulation and have an important biological role. By using ribosome profiling techniques, a large number of ribosomes have been found in many 5' UTRs of transcripts, implying that this region has high translation activity and may be translatable. A large number of typical AUG-initiated uORFs and non-AUG-initiated uORFs have been found in several species [23, 27, 25, 28]. Further analysis also revealed that the translational activity of uORF was higher during meiosis than vegetative, and the above two short ORFs had opposite effects on the downstream ORF translation efficiency [28]. The short open reading frame (sORF) is a short open reading frame (less than 80 to 100 amino acids) in unannotated transcripts. Studies on yeast meiosis have also revealed that the translation of some sORFs and they are highly regulatable [14]. However, the function of these sORFs remains to be further studied. Additionally, the use of ribosome profiling to analyze procedural shifts in the translation process can also lead to the discovery of dually decoded regions and stop codon read-throughs [29].

Finally, The use of ribosome profiles allows in-depth study of protein translation mechanisms. Studies by Li et al. [30] on bacteria found that the SD (Shine-Dalgarno) sequence inside the mRNA coding region was evolutionarily due to the translational pause. While the pairing binding of the anti-SD sequences on the rRNA (on the translating ribosomes) with SD-like sequences (on the transcripts) leads to translation pauses. In other words, the main factor in the translational pause and codon usage of bacteria is the presence of an anti-SD sequence in the 16S rRNA on the ribosome. This phenomenon can guide the expression of heterologous proteins in bacteria. Oh et al. [31] studied the function of the chaperone trigger factor (TF) in E. coli cells using ribosome profiling techniques and selective ribosome profiling techniques. By using the ribosome profiling and isotopic labeling analysis on the HEK293 cell line, mRNA translation was found to be more likely to bind ribosomes on the endoplasmic reticulum than ribosomes in the cytoplasm [32]. Therefore, using ribosome profiles can quickly and accurately discover which sequences of the genome are translated into proteins, where the translated transcripts are highly active, and the timing of some translation-related events.

Also, there are some remarkable limits for ribosome profiling. First, translation elongation inhibitors are particularly important when getting snapshots of the translating ribosomes. Particularly during the analysis of translation pausing, improper handling of inhibitors may cause many false positive or false negative results. Moreover, researchers have found that inhibitors can alter the local distributions of ribosomes on a mRNA [23, 33, 34]. Second, when analyzing ribosome profiling data, we consider that most of the footprints arise from

80S ribosomes and their sizes obey a particular distribution. Thus, we categorize the footprints that do not match this length distribution into the contamination. However, recent studies have shown that 80S mRNA footprints do not conform to the typical size pattern. The footprint size may also be various due to different ribosome conformations [35] and alternative mRNA properties [34]. Finally, how to deal with those ambiguous reads is also a problem we need to consider when analyzing ribosome profiling data. Due to the short length of the footprint and the presence of multiple isoforms in the genome, a footprint can often be mapped to multiple transcripts or multiple locations within a transcript. Thus, the probabilistic alignment may be more suitable for the analysis of ribosome profiling data.

Chapter 2

Identification and analysis of ribosome-associated lncRNAs

2.1 Introduction

Long non-coding RNAs (lncRNAs) are sequences longer than 200 nucleotides with no protein-coding capacity. Over 58,000 genes had been identified as human lncRNAs as of 2015 [8], and that number continues to grow [36, 9]. In contrast, only a small number of lncRNAs have been functionally annotated to date [37]. Because the majority of human lncRNAs are still interpreted as having an unknown function, identification of lncRNA functions has become a challenging problem [38].

Analysis of macromolecular lncRNA interactions has been used as an approach to conduct large-scale studies of lncRNA functions [39]. Ribosome profiling techniques adapt high-throughput sequencing methods to ribosome-protected fragment sequences, which provides a genome-wide dataset of ribosome–RNA interactions [23]. Ingolia et al. first developed ribosome profiling and applied it to studying long intergenic noncoding RNAs (lincRNAs) and reported that the majority of lincRNA fragments engaged by ribosomes represent a limited portion of different lincRNA sequences [25]. Other modified ribosome profiling techniques were applied to identify ribosome-associated lncRNAs and reduce false positives [40, 41].

Many previous studies have used ribosome profiling data to examine ribosome–lncRNA interactions, with a primary focus on detecting protein-coding signatures in lncRNAs. Hence, rigorous metrics and ignoring lncRNA characteristics can lead to underestimates of the association between lncRNAs and ribosomes. For instance, Guttman et. al defined the RRS, a ratio of counts of ribosome footprints from putative ORF to counts of ribosome footprints based on downstream sequences, to assess the sharp decrease in ribosome occupancy at the end of putative ORFs, ultimately demonstrating that lincRNAs do not produce proteins [42]. Wang et al. utilized the three-nucleotide periodicity and uniform distribution of ribosome occupancy to evaluate the translation potential of lincRNAs [43]. These two studies mainly focused on detecting lincRNAs with the ability to encode proteins while excluding any other forms or functions of ribosome-associated lncRNAs from consideration (e.g., storing ribosomes or translational regulation discussed in [44]). Ruiz-Orera et al. assessed ribosomal associations by measuring the breadth of ribosome coverage, which was defined as the number of nucleotides overlapped by Ribo-seq reads on a transcript or a transcript region [45]. This metric ignores the influences of the depth of ribosome coverage, the expression level of a transcript, and the length of a transcript with ribosomal association. Taken together, little attention has been given to ribosome–lncRNA interactions that may involve biological functions [46–48]. Efforts that focus on the identification of reliable ribosome-associated lncRNAs are insufficient.

Here, we define the term "ribosome-associated lncRNAs" as a class of lncRNAs that ribosomes associate with by sliding along regions on them or by binding to specific sites within them. In contrast, "ribosome-free lncRNAs" represent lncRNAs with little (or no) ribosomal association. Note that the term "ribosome-associated lncRNA" was frequently used in previous studies to refer to a rare fraction of lncRNAs with the predicted ability to encode peptides. By mapping ribosome profiling data to lncRNAs, we observed that an average of 39.17% (24.65–59.92%) and 48.16% (26.04–70.13%) of expressed lncRNAs interact with ribosomes in human and mouse, respectively. The protein-coding capacity remains relatively low for the total population of ribosome-associated lncRNAs compared with mRNAs. However, some evidence has emerged for the translation of ribosome-associated lncRNAs. As such, we newly present the ribosomal association index (RAI), an integrative and evidence-based tool that assigns a confidence score to a specific lncRNA representing its ribosomal associability. RAI can be applied to both tissue-specific and ubiquitous lncRNAs in combination with the tissue-specific expression metric *spec* (see "Methods" in this chapter). Focusing on ubiquitously expressed lncRNAs, we used RAI * (1 - *spec*) to measure ribosomal associability. (Note that RAI*spec can be used for analyzing tissue-specific lncRNAs.) Furthermore, we apply two threshold values (the 5th and 95th percentiles of RAI * (1 - *spec*) scores) to divide the lncRNAs into "noribo-lncRNAs" and "ribo-lncRNAs." Those lncRNAs that scored below the lower threshold are defined as "noribo-lncRNAs," representing a subset of reliable ribosome-free lncRNAs. Conversely, lncRNAs that scored above the upper threshold are referred to as "ribo-lncRNAs," representing a subset of high-confidence ribosome-associated lncRNAs. We show that transcript length may not be a major factor associated with ribosomal associability in lncRNAs. Moreover, we have obtained 78 human sequences (and 42 mouse sequences) that are putatively translated lncRNAs from ribo-lncRNAs, respectively. Finally, we investigated the relationship between the ribosome-associated lncRNAs and NMD and cell localization, and we conclude that RAI analyses are a valuable resource that will assist with determining the underlying lncRNA functions.

2.2 Methods

2.2.1 Data collection

We retrieved the original experimental data from NCBI GEO [49] as detailed in Tables 2.1 and 2.2. To calculate the transcript expression level and quantify potential lncRNA–ribosomal associations, we selected ribosome profiling experiments that contained both RNA-seq and ribosome footprint (Ribo-seq) measurements. For further analysis of lncRNA–ribosomal associations, we chose a single representative dataset for each tissue or cell type according to the following three empirical criteria: (i) The mapping rates of both RNA-seq and Ribo-seq are greater than 30%; (ii) The *dist* value is less than 0.15; (iii) For a tissue/cell type represented across multiple datasets, the dataset with the lowest *dist* value, indicating that the footprint length distribution for lncRNAs is closest to that of CDSs in this dataset, is selected. Here, *dist* is a metric of the length distribution similarity between two populations of ribosome footprints that mapped to lncRNAs and CDSs, respectively.

$$\mathrm{dist}(P,Q) = \frac{1}{2}\sum_{l\in L} |P(l) - Q(l)| \tag{2.1}$$

where P and Q denote length frequency distributions of ribosome footprints that mapped to CDSs and lncRNAs, respectively, and L is a finite length space. This value takes a real number between 0 and 1, and larger values indicate a greater difference between these two distributions (see Table S1 and Figs. S1 and S2). Finally, we selected ten different human datasets, which were derived from different tissues or cell types (i.e., brain, breast, fibroblasts, RPE-1, myeloma, ES, HEK293, HeLa, PC3, and U2OS). We selected eight mouse datasets, which were derived from different cell types (fibroblast, EB, and ES) and tissues (i.e., brain, hippocampi, skin, liver, and testis).

Table 2.1 Ribosome profiling datasets used in this study (human).

Source	Reference	Sample	RNA-seq	Ribo-seq	Description
Brain	Gonzalez2014 [50]	normal-A	GSM1495249	GSM1495244	Normal brain
		normal-B	GSM1495250	GSM1495245	
		normal-C	GSM1495251	GSM1495246	
		tumor-A	GSM1495252	GSM1495247	
		tumor-B	GSM1495253	GSM1495248	
Breast	Rubio2014 [51]	control-rep1	GSM1503444	GSM1503442	Breast cancer (cell type: Ductal breast carcinoma; cell line: MDA-MB-231)
		control-rep2	GSM1503438	GSM1503434	
Eye	Tanenbaum2015 [52]	G1-rep1	GSM1657726	GSM1657720	Retinal pigment epithelial cells (cell type: RPE-1)
		G1-rep2	GSM1657727	GSM1657721	
		G2-rep1	GSM1657728	GSM1657722	
		G2-rep2	GSM1657729	GSM1657723	
		M-rep1	GSM1657730	GSM1657724	
		M-rep2	GSM1657731	GSM1657725	
Fibroblasts	Shitrit2015 [53]	control	GSM1712278	GSM1712271	Primary fibroblasts
	Xu2016 [54]	wt-d-leucine	GSM1585204	GSM1585210	Fibroblast (supplemented with d-leucine or l-leucine)
		wt-l-leucine	GSM1585205	GSM1585211	
HEK	Eichhorn2014 [55]	mock	GSM1479597	GSM1479598	HEK293T (mock transfection)
	Iwasaki2016 [56]	dmso-rep1	GSM1720808	GSM1720803	HEK 293 T-REx cell (treatment: DMSO)
		dmso-rep2	GSM1720809	GSM1720804	
	Sidrauski2015 [57]	control-a	GSM1606099	GSM1606107	HEK293T (treatment: untreated)
		control-b	GSM1606100	GSM1606108	
	Subtelny2014 [58]	cyt	GSM1276541	GSM1276542	HEK293T (Cytoplasmically-enriched)
HeLa	Guo2010 [27]	mock12hr	GSM546927	GSM546926	HeLa (transfection: mock)
		mock32hr	GSM546921	GSM546920	
	Park2016 [59]	Mphase-rep1	GSM2100590	GSM2100598	Hela (RNA-seq oligo-dT)
		Mphase-rep2	GSM2100591	GSM2100599	
		Sphase-rep1	GSM2100587	GSM2100596	
		Sphase-rep2	GSM2100588	GSM2100597	
			GSM1898014	GSM1898018	
	Zur2016 [60]	G1phase-exp1	GSM1898015	GSM1898019	HeLa S3 cells
			GSM1898016	GSM1898020	
		G1phase-exp2	GSM1898017	GSM1898021	
			GSM1898006	GSM1898010	
		Mphase-exp1	GSM1898007	GSM1898011	
			GSM1898008	GSM1898012	
		Mphase-exp2	GSM1898009	GSM1898013	
KOPT-K1	Wolfe2014 [61]	dmso-rep1	GSM1370699	GSM1370695	KOPT-K1 T-ALL cell line
		dmso-rep2	GSM1370700	GSM1370696	
Lymphoblastoid	Cenik2015 [62]	GM12878-rep1	GSM1609427	GSM1609378	EBV-transformed lymphoblastoid cells
		GM12878-rep2	GSM1609428	GSM1609379	
		GM12891-rep1	GSM1609430	GSM1609382	
		GM12891-rep2	GSM1609431	GSM1609383	
		GM12892-rep1	GSM1609433	GSM1609384	
		GM12892-rep2	GSM1609434	GSM1609385	
		GM19238-rep1	GSM1609436	GSM1609417	
		GM19238-rep3	GSM1609438	GSM1609418	
		GM19239-rep2	GSM1609440	GSM1609413	
		GM19240-rep1	GSM1609442	GSM1609421	
		GM19240-rep2	GSM1609443	GSM1609422	
		GM19240-rep3	GSM1609444	GSM1609423	
Macrophages	Su2015 [63]	mock-rep1	GSM1632596	GSM1632594	human primary macrophages (TLR2 stimulated; micrococccal nuclease)
		mock-rep2	GSM1632600	GSM1632598	
Muscle	Wein2014 [64]	control	GSM1356677	GSM1356675	Skeletal muscle (normal control)
Myeloma	Wiita2013 [65]	control	GSM1184591	GSM1184592	MM1.S myeloma cell line
PC3	Hsieh2012 [66]	control-rep1	GSM869036	GSM869037	PC3 (prostate cancer cells; sample type: polyA RNA; treatment: vehicle)
		control-rep2	GSM869042	GSM869043	
U2OS	Eichhorn2014 [55]	mock	GSM1479587	GSM1479588	U2OS cell line (mock-transfected, tRNA and rRNA depleted RNA-seq)
	Guo2014 [67]	mock	GSM1248736	GSM1248735	U2OS cells (mock-transfected, poly(A)-selected RNA-seq)
	Jang2015 [[68]	CT00-rep1	GSM1371395	GSM1371443	U2OS (cell type: osteosarcoma)
		CT00-rep2	GSM1371407	GSM1371455	

Table 2.1 Ribosome profiling datasets used in this study (human, continued)

Source	Reference	Sample	RNA-seq	Ribo-seq	Description
		CT02-rep1	GSM1371396	GSM1371444	
		CT02-rep2	GSM1371408	GSM1371456	
		CT04-rep1	GSM1371397	GSM1371445	
		CT04-rep2	GSM1371409	GSM1371457	
		CT06-rep1	GSM1371398	GSM1371446	
		CT06-rep2	GSM1371410	GSM1371458	
		CT08-rep1	GSM1371399	GSM1371447	
		CT08-rep2	GSM1371411	GSM1371459	
		CT10-rep1	GSM1371400	GSM1371448	
		CT10-rep2	GSM1371412	GSM1371460	
		CT12-rep1	GSM1371401	GSM1371449	
		CT12-rep2	GSM1371413	GSM1371461	
		CT14-rep1	GSM1371402	GSM1371450	
		CT14-rep2	GSM1371414	GSM1371462	
		CT16-rep1	GSM1371403	GSM1371451	
		CT16-rep2	GSM1371415	GSM1371463	
		CT18-rep1	GSM1371404	GSM1371452	
		CT18-rep2	GSM1371416	GSM1371464	
		CT20-rep1	GSM1371405	GSM1371453	
		CT20-rep2	GSM1371417	GSM1371465	
		CT22-rep1	GSM1371406	GSM1371454	
		CT22-rep2	GSM1371418	GSM1371466	
hES	Werner2015 [69]	control-rep1	GSM1523640	GSM1523624	hES cell (cell line: H1)
		control-rep2	GSM1523648	GSM1523632	

Table 2.2 Ribosome profiling datasets used in this study (mouse).

Source	Reference	Sample	RNA-seq	Ribo-seq	Description
Brain	Gonzalez2014 [50]	normal-A	GSM1245211	GSM1245214	tissue: PDGF/Cre tumor; Stage: end-stage
		normal-B	GSM1245212	GSM1245215	
		normal-C	GSM1245213	GSM1245216	
		tumor-A	GSM1245217	GSM1245223	
		tumor-B	GSM1245218	GSM1245224	
		tumor-C	GSM1245219	GSM1245225	
Fibroblast	Thoreen2012 [70]	wild-vehicle	GSM904895	GSM904893	Embryonic fibroblast (genotype: 4EBP1/2 +/+ p53 -/-; genetic: 129/Svj
Hippocampi	Cho2015 [71]	10min-rep1	GSM1853990	GSM1853985	Hippocampal tissue (strain: C57BL/6N)
		10min-rep2	GSM1854000	GSM1853995	
		10min-rep3	GSM1854010	GSM1854005	
		30min-rep1	GSM1853991	GSM1853986	
		30min-rep2	GSM1854001	GSM1853996	
		30min-rep3	GSM1854011	GSM1854006	
		4hr-rep1	GSM1853992	GSM1853987	
		4hr-rep2	GSM1854002	GSM1853997	
		4hr-rep3	GSM1854012	GSM1854007	
		5min-rep1	GSM1853989	GSM1853984	
		5min-rep2	GSM1853999	GSM1853994	
		5min-rep3	GSM1854009	GSM1854004	
		control-rep1	GSM1853988	GSM1853983	
		control-rep2	GSM1853998	GSM1853993	
		control-rep3	GSM1854008	GSM1854003	
Liver	Alvarez2017 [72]	control-rep1	GSM2219150	GSM2219142	Fetal liver (cell type: Erythroid progenitors; strain: C57BL/6; age: E14.
		control-rep2	GSM2219151	GSM2219143	
	Eichhorn2014 [55]	wt	GSM1479601	GSM1479602	Primary liver tissue (strain: C57BL/6; age: 6 weeks; sex: male)
	Fradejas2017 [73]	secisbp2-wt-rep1	GSM2227376	GSM2227367	Liver (age: 5-8 weeks; genotype: wild type)
		secisbp2-wt-rep2	GSM2227377	GSM2227368	
		trsp-wt-rep1	GSM2227380	GSM2227371	
		trsp-wt-rep2	GSM2227382	GSM2227373	
	Frederic2015 [74]	ZT00-A	GSM1897722	GSM1897856	Liver (strain: C57BL/6; age: post natal day 12-16; genotype: wild type)
		ZT00-B	GSM1897734	GSM1897868	
		ZT00-C	GSM1897751	GSM1897880	
		ZT00-D	GSM1897777	GSM1897892	
		ZT02-A	GSM1897723	GSM1897857	
		ZT02-B	GSM1897735	GSM1897869	
		ZT02-C	GSM1897754	GSM1897881	
		ZT02-D	GSM1897779	GSM1897893	
		ZT04-A	GSM1897724	GSM1897858	
		ZT04-B	GSM1897736	GSM1897870	
		ZT04-C	GSM1897756	GSM1897882	
		ZT04-D	GSM1897781	GSM1897894	
		ZT06-A	GSM1897725	GSM1897859	
		ZT06-B	GSM1897737	GSM1897871	
		ZT06-C	GSM1897758	GSM1897883	
		ZT06-D	GSM1897783	GSM1897895	
		ZT08-A	GSM1897726	GSM1897860	
		ZT08-B	GSM1897738	GSM1897872	
		ZT08-C	GSM1897760	GSM1897884	
		ZT08-D	GSM1897785	GSM1897896	
		ZT10-A	GSM1897727	GSM1897861	
		ZT10-B	GSM1897739	GSM1897873	
		ZT10-C	GSM1897762	GSM1897885	
		ZT10-D	GSM1897788	GSM1897897	
		ZT12-A	GSM1897728	GSM1897862	
		ZT12-B	GSM1897740	GSM1897874	
		ZT12-C	GSM1897764	GSM1897886	
		ZT12-D	GSM1897790	GSM1897898	
		ZT14-A	GSM1897729	GSM1897863	
		ZT14-B	GSM1897742	GSM1897875	
		ZT14-C	GSM1897766	GSM1897887	

Table 2.2 Ribosome profiling datasets used in this study (mouse, continued)

Source	Reference	Sample	RNA-seq	Ribo-seq	Description
		ZT14-D	GSM1897791	GSM1897899	
		ZT16-A	GSM1897730	GSM1897864	
		ZT16-B	GSM1897743	GSM1897876	
		ZT16-C	GSM1897768	GSM1897888	
		ZT16-D	GSM1897793	GSM1897900	
		ZT18-A	GSM1897731	GSM1897865	
		ZT18-B	GSM1897745	GSM1897877	
		ZT18-C	GSM1897771	GSM1897889	
		ZT18-D	GSM1897796	GSM1897901	
		ZT20-A	GSM1897732	GSM1897866	
		ZT20-B	GSM1897747	GSM1897878	
		ZT20-C	GSM1897773	GSM1897890	
		ZT20-D	GSM1897798	GSM1897902	
		ZT22-A	GSM1897733	GSM1897867	
		ZT22-B	GSM1897749	GSM1897879	
		ZT22-C	GSM1897775	GSM1897891	
		ZT22-D	GSM1897800	GSM1897903	
	Howard2013 [75]	wt	GSM1122211	GSM1122205	Liver (strain: FVB/N; age: 3 weeks; treatment: 6 week diet 0 ppm sele
	Janich2015 [76]	ZT0-rep1	GSM1644100	GSM1644076	Liver (strain: C57BL/6JRj; age: 11-12 weeks; gender: male)
		ZT0-rep2	GSM1644101	GSM1644077	
		ZT10-rep1	GSM1644110	GSM1644086	
		ZT10-rep2	GSM1644111	GSM1644087	
		ZT12-rep1	GSM1644112	GSM1644088	
		ZT12-rep2	GSM1644113	GSM1644089	
		ZT14-rep1	GSM1644114	GSM1644090	
		ZT14-rep2	GSM1644115	GSM1644091	
		ZT16-rep1	GSM1644116	GSM1644092	
		ZT16-rep2	GSM1644117	GSM1644093	
		ZT18-rep1	GSM1644118	GSM1644094	
		ZT18-rep2	GSM1644119	GSM1644095	
		ZT2-rep1	GSM1644102	GSM1644078	
		ZT2-rep2	GSM1644103	GSM1644079	
		ZT20-rep1	GSM1644120	GSM1644096	
		ZT20-rep2	GSM1644121	GSM1644097	
		ZT22-rep1	GSM1644122	GSM1644098	
		ZT22-rep2	GSM1644123	GSM1644099	
		ZT4-rep1	GSM1644104	GSM1644080	
		ZT4-rep2	GSM1644105	GSM1644081	
		ZT6-rep1	GSM1644106	GSM1644082	
		ZT6-rep2	GSM1644107	GSM1644083	
		ZT8-rep1	GSM1644108	GSM1644084	
		ZT8-rep2	GSM1644109	GSM1644085	
Skin	Blanco2016 [77]	wt1	GSM1854037	GSM1854031	Skin (tumour stage: skin papilloma; strain: C57BL/6; age: 1 month)
		wt2	GSM1854038	GSM1854032	
		wt3	GSM1854039	GSM1854033	
	Sendoel2017 [78]	wt-invivo-rep0	GSM2199587	GSM2199581	Back skins (strain: R26-Sox2-IRES-eGFP fl/+; age: P4)
		wt-invivo-rep1	GSM2199588	GSM2199582	
Testis	Castaneda2014 [79]	wt-a	GSM1234250	GSM1234248	Testis (strain: 129SvJae; genotype: wild type)
		wt-b	GSM1234254	GSM1234252	
mEB	Ingolia2011 [25]	eb	GSM765286	GSM765291	Embryoid body (genotype: CReP+/-; GADD34+/+ (WT))
mES	Hurt2013 [80]	control	GSM1024299 GSM1024300	GSM1024311	Embryonic stem cells (v6.5 cell line)
	Ingolia2011 [25]	mes	GSM765288	GSM765300	ES cell (E14 cell line; genetic: 129/Ola)
	Reid2014 [81]	cyt	GSM1299862 GSM1299863	GSM1299858 GSM1299859	Embryonic fibroblasts (genotype: CReP+/-; GADD34+/+ (WT))
		er	GSM1299860 GSM1299861	GSM1299856 GSM1299857	

2.2.2 Transcriptome

The transcriptome (consisting of mRNAs and lncRNAs) was used as a reference for mapping RNA/Ribo-seq reads based on the following considerations. First, we restricted read mapping to annotated transcripts to avoid the identification of novel transcripts. Second, mapping reads to a genome is a complicated problem as the mapping rate is sensitive for short reads and those reads spanning splicing junctions. Thus, we downloaded genomic sequences and gene annotation files from GENCODE [36] and then utilized custom Python scripts to generate transcriptome sequences (see Table 2.4). By excluding lncRNAs that are derived from known protein coding genes, we finally obtained 27,545 and 14,609 lncRNAs for human and mouse, respectively, which primarily represent lincRNA and antisense RNA sequences (see Table 2.3).

Table 2.3 Long non-coding RNAs used in this study. See https://www.gencodegenes.org/gencode_biotypes.html for the details on transcript biotype.

Biotype	Human	Mouse
lincRNA	13245	6473
antisense	10980	3612
TEC	1072	2759
sense_intronic	984	294
retained_intron	517	294
processed_transcript	368	980
sense_overlapping	310	50
3prime_overlapping_ncRNA	34	3
pseudogene	20	24
bidirectional_promoter_lncRNA	11	118
non_coding	3	0
macro_lncRNA	1	2
Total	27545	14609

Table 2.4 Genomic sequences, gene annotations and contaminant sequences for human and mouse. Genomic sequences and gene annotation files were downloaded from GENCODE [36, 82]. The contaminant sequence collection consists of transfer RNAs (tRNAs), ribosomal RNAs (rRNAs), small nucleolar RNAs (snoRNAs), small nuclear RNAs (snRNAs) and microRNAs (miRNAs). Here, tRNAs were retrieved from GtRNAdb[83], rRNAs were obtained from NCBI[84], UCSC[85] and Ensembl[86].

Species	Files	Descriptions
Human	Genome (hg19)	http://hgdownload.cse.ucsc.edu/goldenPath/hg19/bigZips/hg19.2bit
	GTF (GENCODE_v25lift37)	ftp://ftp.sanger.ac.uk/pub/gencode/Gencode_human/release_25/GRCh37_mapping/gencode.v25lift37.annotation.gtf.gz
	tRNAs (GtRNAdb_hg19)	http://gtrnadb.ucsc.edu/genomes/eukaryota/Hsapi19/hg19-tRNAs.fa
	rRNAs (NCBI)	Search “Nucleotide” database by “rRNA[All Fields] AND “Homo sapiens”[porgn] AND biomol_rrna[PROP]”
	rRNAs (UCSC)	Search “Table Browser” by “genome:Human; assembly:Feb.2009 (GRCh37/hg19); group:All tables; table:rsmk; repClass = rRNA”
	rRNA (Ensembl)	Search “BioMart” by “Ensembl Genes 90; Human genes (GRCh38.p10); Transcript type: rRNA”
	snoRNAs/snRNAs/miRNAs	Search “BioMart” by “Ensembl Genes 90; Human genes (GRCh38.p10); Transcript type: snoRNA, snRNA, miRNA”
Mouse	Genome (mm10)	http://hgdownload.soe.ucsc.edu/goldenPath/mm10/bigZips/mm10.2bit
	GTF (GENCODE_v12)	ftp://ftp.sanger.ac.uk/pub/gencode/Gencode_mouse/release_M12/gencode.vM12.annotation.gtf.gz
	tRNAs (GtRNAdb_mm10)	http://gtrnadb.ucsc.edu/genomes/eukaryota/Mmusc10/mm10-tRNAs.fa
	rRNAs (NCBI)	Search “Nucleotide” database by “rRNA[All Fields] AND "Mus musculus"[porgn] AND biomol_rrna[PROP]”
	rRNAs (UCSC)	Search “Table Browser” by “genome:Mouse; assembly:Dec.2011 (GRCm38/mm10); group:All tables; table:rsmk; repClass = rRNA”
	rRNA (Ensembl)	Search “BioMart” by “Ensembl Genes 90; Mouse genes (GRCm38.p5); Transcript type: rRNA”
	snoRNAs/snRNAs/miRNAs	Search “BioMart” by “Ensembl Genes 90; Mouse genes (GRCm38.p5); Transcript type: snoRNA, snRNA, miRNA”

2.2.3 Alignment and quantification

RNA/Ribo-seq reads were mapped to the transcriptome using Bowtie2 [87] with the *–very-sensitive-local* option. Cutadapt [88] was used to trim adapter sequences from reads if the adapter sequence was described in the literature. Additionally, we performed a local read alignment to remove adapter sequences from one or both ends of the alignment. The Ribo-seq reads were produced by a strand-specific protocol, which means reads from $5'$ to $3'$ are mostly mapped to the transcript sense strand. This helps to determine whether reads were sequenced from the protein-coding transcript or the antisense transcript on the opposite strand. For each read, we allowed a maximum of 100 distinct alignments to take into account the high sequence similarity among transcript variants of the same gene locus or among transcripts with repetitive elements. Table 2.5 shows the details of the software parameters used in this procedure.

Table 2.5 Softwares and parameters used in this study.

Softwares	Parameters	Descriptions
Cutadapt v1.9.1 [88]	-a ADAPTER -m 15	Remove trimmed reads that are shorter than 15nt
Bowtie v2.3.2 [87]	–very-sensitive-local	Discard contaminant sequences from RNA-seq data
	–very-sensitive-local -k 100	Align RNA-seq data to the transcriptome (up to 100 alignments are allowed for a read)
	–very-sensitive-local –norc	Discard contaminant sequences from Ribo-seq data (only forward reference strand is considered)
	–very-sensitive-local -k 100 –norc –rdg 99999999,99999999 –rfg 99999999,99999999	Align Ribo-seq data to the transcriptome (insertion and deletion are not allowed)
RSEM v1.2.31 [89]	rsem-calculate-expression –alignments	Estimate transcript expression from SAM file

The transcript expression value RPKM (reads per kilobase per million mapped reads) was pre-computed from RNA-seq data using RSEM v1.2.31[89]. To quantify one Ribo-seq read that mapped to N $(1 \leq N \leq 100)$ different locations, we defined a metric $w(i)$ to represent the fraction of mapped reads assigned to the i-th location $(1 \leq i \leq N)$.

$$w(i) = \frac{\mathrm{RPKM}(i)}{\sum_{n=1}^{N} \mathrm{RPKM}(n)} \tag{2.2}$$

where RPKM(i) is the expression value for the transcript referring to the i-th location.

It is worth noting that reads need to be mapped to rRNA and tRNA databases before mapping to the transcriptome. This is because most of the RNAs in cells are derived from rRNAs and tRNAs. Using the reads originated from rRNAs and tRNAs will increase the workload of the downstream mapping process, and may also cause some unbiased predictions of transcripts including similar sequences with rRNA/tRNA. Surprisingly, after we performed the above processes, we still observed that the frequency distribution of reads with specific lengths on lncRNAs did not fit well to that on mRNAs (data not shown), and instead formed a local peak in the frequency distribution. We extracted reads from the local peak to examine if we can generate a consensus sequence. If that is the case, we further compared the consensus sequence with the human transcriptome by BLAST. As shown in Fig. 2.6, we found that the sources of contamination are snoRNA, snRNA, and miRNA. Thus, we added snoRNA, snRNA, and miRNA to the contamination list for filtering.

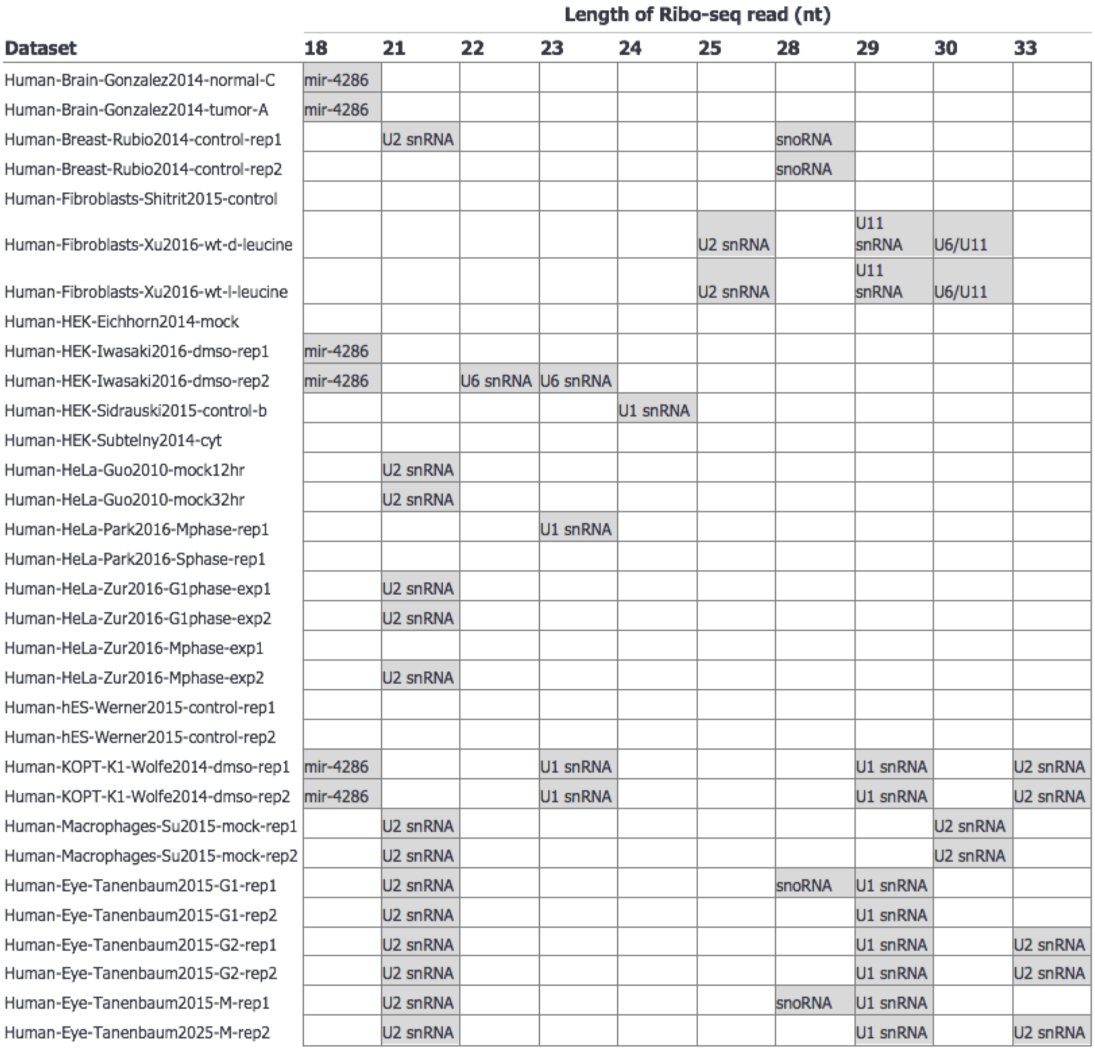
2.2.4 Expressed transcripts and tissue specificity

Although most previous studies are based on quantitative data over a single representative transcript for each gene, we used RSEM to estimate the abundance of total known transcript variants from RNA-seq data, defined by an expression threshold of 1 (i.e., ≥ 1 RPKM) for the purpose of identifying expressed transcripts [90, 91]. Where not otherwise specified, the following analyses were based on sets of expressed transcripts.

For a transcript, to measure the expression tissue specificity, we count the number (x) of tissues/cell types in which this transcript is expressed and transform it to a scale from 0 (ubiquitous) to 1 (specific) as follows:

$$\text{spec} = \frac{M-x}{M-1} \tag{2.3}$$

Table 2.6 Contaminant Ribo-seq reads derived from miRNAs, snRNAs and snoRNAs are enriched in lncRNAs.



	Length of Ribo-seq read (nt)									
Dataset	**18**	**21**	**22**	**23**	**24**	**25**	**28**	**29**	**30**	**33**
Human-Brain-Gonzalez2014-normal-C	mir-4286									
Human-Brain-Gonzalez2014-tumor-A	mir-4286									
Human-Breast-Rubio2014-control-rep1		U2 snRNA					snoRNA			
Human-Breast-Rubio2014-control-rep2							snoRNA			
Human-Fibroblasts-Shitrit2015-control										
Human-Fibroblasts-Xu2016-wt-d-leucine						U2 snRNA		U11 snRNA	U6/U11	
Human-Fibroblasts-Xu2016-wt-l-leucine						U2 snRNA		U11 snRNA	U6/U11	
Human-HEK-Eichhorn2014-mock										
Human-HEK-Iwasaki2016-dmso-rep1	mir-4286									
Human-HEK-Iwasaki2016-dmso-rep2	mir-4286		U6 snRNA	U6 snRNA						
Human-HEK-Sidrauski2015-control-b					U1 snRNA					
Human-HEK-Subtelny2014-cyt										
Human-HeLa-Guo2010-mock12hr		U2 snRNA								
Human-HeLa-Guo2010-mock32hr		U2 snRNA								
Human-HeLa-Park2016-Mphase-rep1				U1 snRNA						
Human-HeLa-Park2016-Sphase-rep1										
Human-HeLa-Zur2016-G1phase-exp1		U2 snRNA								
Human-HeLa-Zur2016-G1phase-exp2		U2 snRNA								
Human-HeLa-Zur2016-Mphase-exp1										
Human-HeLa-Zur2016-Mphase-exp2		U2 snRNA								
Human-hES-Werner2015-control-rep1										
Human-hES-Werner2015-control-rep2										
Human-KOPT-K1-Wolfe2014-dmso-rep1	mir-4286			U1 snRNA				U1 snRNA		U2 snRNA
Human-KOPT-K1-Wolfe2014-dmso-rep2	mir-4286			U1 snRNA				U1 snRNA		U2 snRNA
Human-Macrophages-Su2015-mock-rep1		U2 snRNA							U2 snRNA	
Human-Macrophages-Su2015-mock-rep2		U2 snRNA							U2 snRNA	
Human-Eye-Tanenbaum2015-G1-rep1		U2 snRNA					snoRNA	U1 snRNA		
Human-Eye-Tanenbaum2015-G1-rep2		U2 snRNA						U1 snRNA		
Human-Eye-Tanenbaum2015-G2-rep1		U2 snRNA						U1 snRNA		U2 snRNA
Human-Eye-Tanenbaum2015-G2-rep2		U2 snRNA						U1 snRNA		U2 snRNA
Human-Eye-Tanenbaum2015-M-rep1		U2 snRNA					snoRNA	U1 snRNA		
Human-Eye-Tanenbaum2025-M-rep2		U2 snRNA						U1 snRNA		U2 snRNA

where M is the total number of tissues and cell types used in this study. The $spec$ metric is consistent with the $counts$ metric mentioned in [92].

2.2.5 Ribosome density to distinguish ribosome-associated and ribosome-free lncRNAs in a single dataset

To measure the extent to which ribosomes are associated with a transcript or a region of a transcript, we used ribosome density, which is calculated as

$$\text{ribosome\_density}(i,j;T) = \frac{\text{ribo}(i,j)}{\text{RPKM}(T)\cdot|j-i+1|} \tag{2.4}$$

where $T = t_1...t_n$ is a transcript of length n, $\text{ribo}(i,j)$ is the number of Ribo-seq reads mapped on the substring $T(i,j) = t_i...t_j$ $(1 \leq i \leq j \leq n)$, and $\text{RPKM}(T)$ is the expression value of T. Thus, $\text{ribosome\_density}(1,n;T)$ represents the density of ribosome occupancy over the whole transcript T. In general, a ribosome will dissociate from mRNA once a stop codon is encountered, which makes the area downstream of the stop codon (the $3'$ UTR), a ribosome-free region and thus a suitable reference region for detecting ribosome-associated signals. To obtain a significant ribosome-associated lncRNA, we further derived an empirical distribution of ribosome density scores from $3'$ UTRs and then applied a 90th percentile cut-off value of ribosome density scores from $3'$ UTRs in order to distinguish between ribosome-associated and ribosome-free lncRNAs (see Fig. 2.2a). The rationale for choosing this seemingly less stringent cut-off value is that it (i) may enable the detection of ribosome rescue in $3'$ UTRs [93] and (ii) guarantees that the majority (i.e., >90%) of mRNAs that are associated with ribosomes and produce proteins are identified as expected [94](see Fig. 2.1).

2.2.6 Ribosomal association index (RAI) defines ribo-lncRNAs and noribo-lncRNAs across multiple datasets

For each lncRNA, we applied the newly proposed ribosomal association index (RAI) to quantify ribosome associability.

$$\mathrm{RAI} = \frac{\sum_{i=1}^{M} x(i) \cdot y(i)}{\sum_{i=1}^{M} x(i)} \tag{2.5}$$

where M is the number of independent experiments; $x(\cdot)$ is the indicator function of transcript expression, that is, $x(i) = 1$ if the lncRNA is expressed in the i-th experiment and 0 otherwise; and $y(\cdot)$ denotes the ribosomal association sign function, that is, $y(i) = 1$ if the ribosomal association was supported by the i-th experiment and -1 otherwise. Here, a continuous value of $y(i)$ will provide more information about the ribosomal association. However, it is difficult to directly compare the ribosomal association across different datasets by using ribosome density, which is normalized to transcript abundance in each dataset.

Furthermore, we used RAI * (1 - *spec*) to assign a more confident score of ribosome associability based on multiple pieces of experimental evidence. The RAI * (1 - *spec*) score can range between 1, for ribosome-associated lncRNAs, and -1, for the ribosome-free lncRNAs (see Fig. 2.4 and Table S3).

2.2.7 The putative ORF in lncRNAs

For lncRNAs, putative ORFs with lengths $\geq$ 30nt (including the stop codon) were considered to analyze their coding potential. A putative ORF is a continuous sequence of trinucleotides starting with an ATG trinucleotide and ending with TGA, TAA, or TAG.

2.2.8 Coding potential assessment

Fragment length organization similarity score

The fragment length organization similarity score (FLOSS) was computed as formulated and presented by [95].

$$\text{Original FLOSS} = \sum_{l=26}^{34} ||f(l) - f_{ref}(l)||. \tag{2.6}$$

where $f(l)$ is the fraction of reads at length l in the transcript histogram and $f_{ref}(l)$ is the corresponding fraction in the reference histogram (CDSs). Footprints derived from translating ribosomes are expected to have a specific length distribution. Thus, the idea behind the FLOSS analysis is to compare the histogram distributions of footprint lengths between a given transcript and the reference (i.e., CDSs), in which ribosomes are considered to translate proteins. To maintain the consistency of metrics of coding potential, we transformed the original FLOSS score to 1 - FLOSS. Thus, the transformed FLOSS (called FLOSS hereafter) value range is from 0 (non-translated) to 1 (high possibility of translating).

Ribosome release score

For a previously defined putative ORF of a lncRNA or a CDS, the ribosome release score (RRS) was calculated according to the description in [42]. For each ORF, we calculated the ratio of the number of footprints distributed in the ORF and the 3' UTR. At the same time, to exclude the influence of the different lengths of the two regions, we also calculated the ratio of RNA-seq reads in these two regions as normalization coefficients as below:

$$\text{RRS} = \frac{\left(\frac{Count_{ORF}}{Count_{3'UTR}}\right)_{footprint}}{\left(\frac{Count_{ORF}}{Count_{3'UTR}}\right)_{RNA-seq}} \tag{2.7}$$

A transcript undergoing translation tends to show ribosome coverage over the majority of an ORF, and thus the ribosome density over ORFs ends sharply at the translation termination site. Guttman et al. developed the ribosome release score (RRS) based on the drop signal at the translation termination site to detect the translation event [42]. Here, the RRS value was scaled to range from 0 to 1.

Framescore

As the ribosome moves three nucleotides in each step along each ORF during protein synthesis, the three-base periodicity can be represented by the frame distribution, which displays the frequency of Ribo-seq reads (from the $5'$ end of each read) in each frame. If the majority of lncRNAs also encode peptides, three-base periodicity would be expected in most of their putative ORFs. Note that the different experiments and different methods of processing reads may affect the shape of the frame distributions. Fortunately, the frame distribution of CDSs provides a good reference for the differentiation of ORFs between active and inactive translation. We proposed the Framescore to measure the dissimilarity in terms of frame distribution, which is the proportion of $5'$ ends of Ribo-seq reads mapped to all three frames. Here, Q is the frame distribution of Ribo-seq reads among all CDSs undergoing ribosomal translation, and P represents the frame distribution of reads in a (putative) ORF from a transcript. Framescore was used to calculate the Kullback–Leibler divergence from P to Q as

$$\text{Framescore}(P,Q) = \sum_{i=1}^{3} P(i) \log \frac{P(i)}{Q(i)}. \tag{2.8}$$

The difference between Framescore and ORFscore which is the other triplet phasing metric [96], is that ORFscore supposes footprints derived from translating ribosomes will be predominantly mapped to frame one and frame two. However, Framescore uses the mapping results onto CDSs as the reference to obtain a more stable performance.

Translation score

Taken together, we applied these three coding metrics (FLOSS, RRS, and Framescore) to assess the ability of each putative lncRNA ORF to encode a peptide. For each coding metric, a cut-off was generated such that 90% of mRNAs can be identified as having coding ability according to this threshold, which was then applied to lncRNAs. To integrate these three filtering results, we developed the translation score (TS) to evaluate the coding potential for a specific lncRNA across multiple datasets.

$$\mathrm{TS} = \sum_{i=1}^{N} w(\alpha(i)) \tag{2.9}$$

where N is the number of datasets in which the transcript is identified as ribosome-associated. In the i-th dataset, $\alpha(i)$ is a translation level function ranging from 0 to 3, indicating the maximum number of coding filters passed for a putative lncRNA ORF. While $w(\cdot)$ is a function that assigns the weight for each translation level (0, 1, 2, and 3 correspond to weights of -1, -0.5, 0.5, and 1, respectively). Finally, for a given lncRNA, TS is a weighted sum function, with a positive value indicating translation, and a negative value indicating no translation.

2.2.9 Mass spectrometry data

Peptide sequences derived from mass spectrometry data were downloaded from sORFs.org [97]. Peptide sequences aligned to protein coding transcripts (by tBlastn [98]) were removed, then the remaining peptides were aligned to lncRNAs.

2.2.10 Sequence conservation

PhyloP scores, which measure base-wise evolutionary conservation from multiple alignments, were downloaded from GENCODE [36]. Positive phyloP scores represent slower evolution than expected (in other words, conserved), and vice versa.

2.2.11 Nonsense-mediated decay (NMD) and cellular localization analysis

For the NMD analysis, we computed the fold change of RNA-seq expression levels from the control sample to those from the UPF1 knockdown sample. Here, UPF1 is one of the major NMD factors, and interfering with the expression of UPF1 is expected to cause increased expression levels of NMD-targeted transcripts. RNA-seq data from HeLa cells were downloaded from NCBI GEO (GSE86148) [99].

For the cellular localization analysis, cells were first separated into cellular fractions before the extraction of RNA. We calculated the fold change of RNA-seq data from the cytoplasmic fraction to that from the nuclear fraction of HeLa cells. RNA-seq data from the nucleus (ENCSR000CPQ) and the cytoplasm (ENCSR000CPP) were download from ENCODE [100].

We applied the same procedure to calculate the fold change for the NMD analysis and the cellular localization analysis. Reads mapped to tRNAs, rRNAs, snoRNAs, or miRNAs were first removed. For the remaining reads, their first 15 nucleotides with low sequencing qualities were trimmed by Cutadapt [88]. Trimmed reads were mapped to the transcriptome by Bowtie [101]. Transcript expression values were calculated by RSEM v1.2.31 [89]. Differential expression analysis was performed using EBSeq [102] to obtain the posterior fold change for each transcript.

2.3 Results

2.3.1 A large fraction of expressed lncRNAs are associated with ribosomes

To identify ribosome-associated lncRNAs in each dataset, we first calculated the ribosome density (i.e., the number of ribosomes per unit length of transcript) for each lncRNA and further derived the empirical distribution of ribosome density values from $3'$ UTRs. Then we adopted a cut-off value at the 90th percentile of the ribosome density values for $3'$ UTRs. The rationale for choosing this cut-off value is that it guarantees that the majority (i.e., >90%) of mRNAs that are associated with ribosomes and produce proteins are identified as expected [94] (see Fig. 2.2a and Fig. S3). Finally, a transcript with ribosome density greater than or equal to the cut-off value was defined as ribosome-associated and was otherwise defined as ribosome-free. For expressed mRNAs, an average of 97.36% (94.73–99.51%) and 98.30% (95.99–99.42%) of them were observed to interact with ribosomes in human and mouse, respectively. This is in agreement with the fact that mRNAs serve as protein-coding transcripts associated with ribosomes. Surprisingly, we found that an average of 39.17%(24.65–59.92%) of human-expressed lncRNAs and an average of 48.16% (26.04–70.13%) of mouse-expressed lncRNAs were also associated with ribosomes (see Fig. 2.1 and Table S2). In total, 7,153 and 3,577 lncRNAs were identified as associated with ribosomes in at least one human and mouse dataset, respectively. We also determined that ribosomal association was more difficult to detect among low-expression transcripts than among highly expressed ones, but this was not observed among all datasets (see Fig. 2.2b and Fig. S3). Despite the differences between the experiment samples, which may affect the expression level and the ribosomal association of lncRNAs, a substantial fraction of lncRNAs were observed to interact with ribosomes over all human and mouse ribosome profiling experiments.

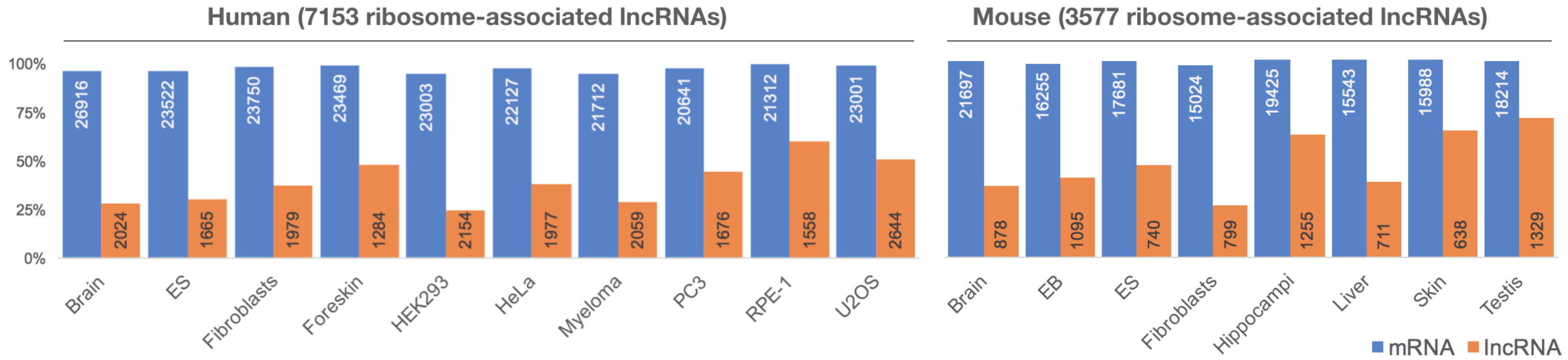


Fig. 2.1 **The percentages of expressed mRNAs (blue) and lncRNAs (orange) that are associated with ribosomes across multiple tissues or cell types.** For the human datasets, 94.73–99.51% of mRNAs and 24.65–59.92% of lncRNAs were associated with ribosomes. For the mouse datasets, 95.99–99.42% of mRNAs and 26.04–70.13% of lncRNAs were associated with ribosomes. In summary, 7,153 and 3,577 lncRNAs were identified to be associated with ribosomes in at least one dataset. The ribosome association was defined based on ribosome density (see “Methods" in this chapter). The number of expressed mRNAs or lncRNAs is shown in each bar.

2.3.2 Analysis of the coding potential for ribosome-associated lncRNAs based on Ribo-seq

To further examine whether the ribosome-associated lncRNAs encode peptides, we first defined the putative lncRNA ORFs (see “Methods" in this chapter), and then assessed the coding potential of their putative ORFs based on the following considerable characteristics for translating ORFs. (i) FLOSS (fragment length organization similarity score) was used to compare the length distributions of footprints from CDSs with the surveyed lncRNAs; (ii) RRS (ribosome release score) was used to measure the drop signal of footprints at the translation termination sites; (iii) Framescore, which was developed in this study, was used to measure the three-nucleotide periodicity. Note that such characteristics are measured by analyzing Ribo-seq reads across a given transcript (see “Methods" in this chapter for detailed description of FLOSS, RRS, and Framescore).

The above three different coding metrics were calculated after removing footprints corresponding to contaminants. To filter footprints from among potential nonribosomal RNA–protein complexes, we first compared Ribo-seq reads from lncRNAs to those from mRNAs and found that reads of a specific length were enriched among lncRNAs (see Table 2.6). By identifying the sequences that were most frequently observed in these enriched reads from the full transcripts, we found that Ribo-seq reads may also be obtained from snRNAs, snoRNAs, and miRNAs. This finding is consistent with previous observations [103]. To integrate these three coding metrics to more stringently assess the ability of each ribosome-associated lncRNA to encode a peptide, we first generated cut-offs from mRNAs based on these three metrics and then applied these cut-offs to filter lncRNAs. Figure 2.3b and S4 show the distribution of FLOSS, RRS, and Framescore values among mRNAs as well as ribosome-associated and ribosome-free lncRNAs. Based on these three coding metrics, mRNAs consistently have the strongest coding abilities. Conversely, both the ribosome-associated and ribosome-free lncRNAs showed weak coding potential. Note that there is

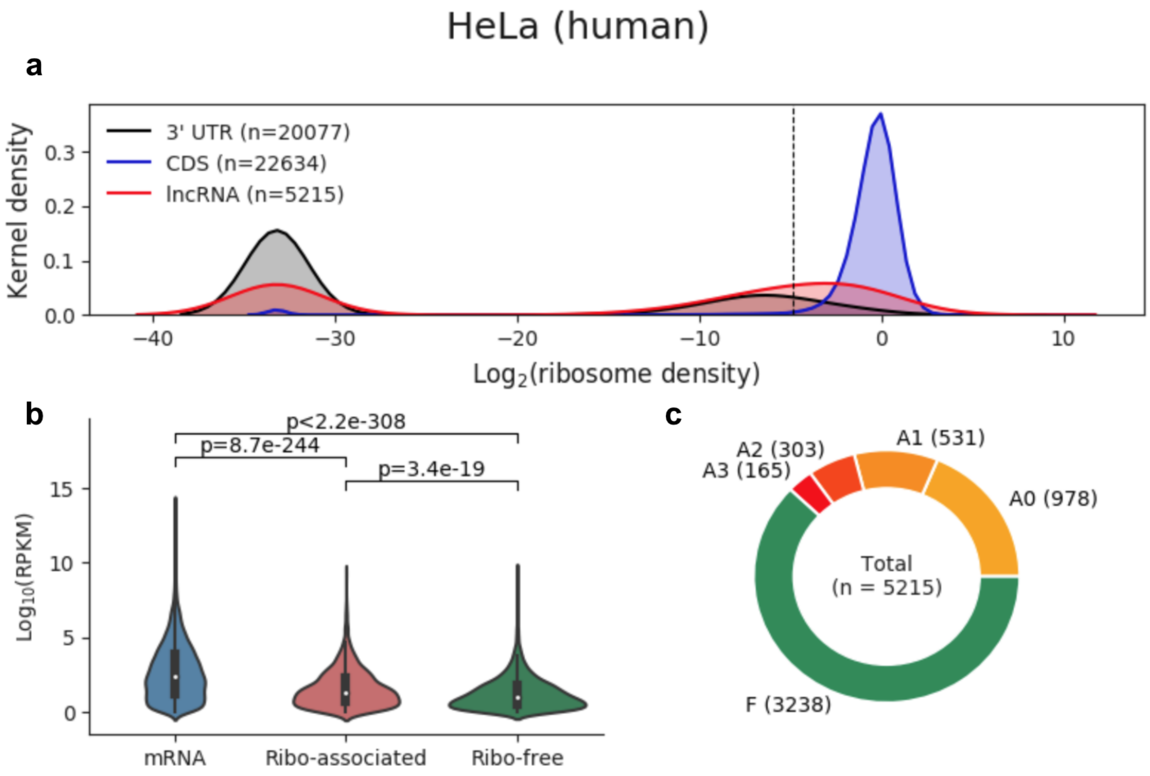


Fig. 2.2 **The discrimination of ribosome-associated and ribosome-free lncRNAs by ribosome density in the HeLa dataset.** (**a**) Kernel density distribution of ribosome density (log$_2$ scale) for 3′UTRs (gray), CDSs (blue), and lncRNAs (red). The vertical dashed line corresponds to the 90th percentile of the ribosome density scores for 3′ UTRs, which is used as the cut-off to distinguish between ribosome-associated lncRNAs and ribosome-free lncRNAs. Those lncRNAs to the right of this cut-off (including the cut-off itself) are identified as ribosome-associated lncRNAs; the rest are ribosome-free in this study. Note that transcripts or regions without any mapped Ribo-seq read correspond to a peak near -33 (owing to the addition of a pseudo value of 10e-10 prior to log transformation). (**b**) Violin plot of the expression levels (RPKM, log2 scale) of mRNAs as well as ribo-associated and free lncRNAs. The *p*-values correspond to two-sample *t*-tests. (**c**) Classification of lncRNAs by using FLOSS, RRS, and Framescore as filters to assess the coding potential for each ribosome-associated lncRNA. “F” means ribosome-free, “A0” means no coding filter has been passed, “A1”, “A2”, and “A3” denote that one, two, and three passed translation filter(s), respectively. (See Fig. S3 for the other datasets.)

still a tendency toward higher coding scores for ribosome-associated lncRNAs relative to ribosome-free lncRNAs across all datasets, suggesting that some of the ribosome-associated lncRNAs may even encode peptides. Figure 2.3a (see Fig. S4 for other datasets) indicates how many of the putative ORFs in ribosome-associated lncRNAs pass the cut-offs for those three coding scores (FLOSS, RRS, and Framescore). In HeLa cells, for example, we observed 275 putative ORFs that passed those three coding filters, implying that translation of these putative ORFs may occur.

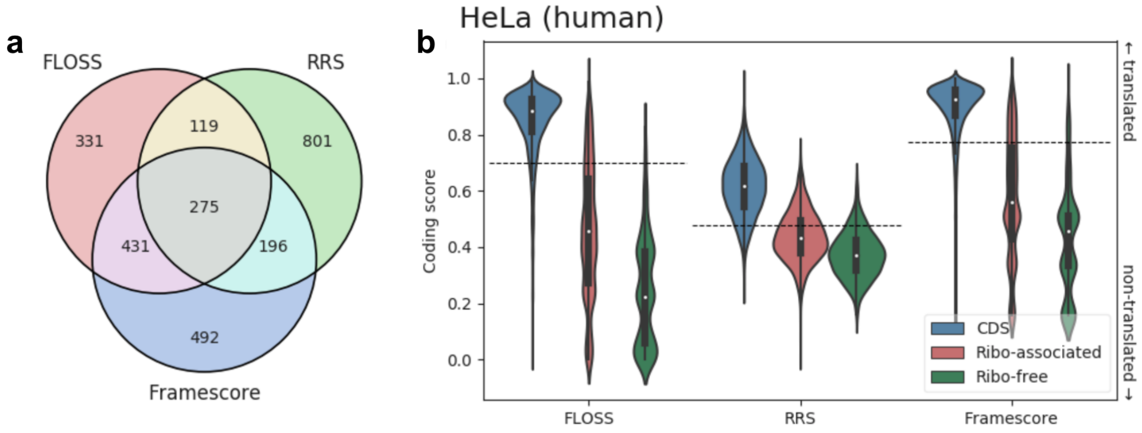
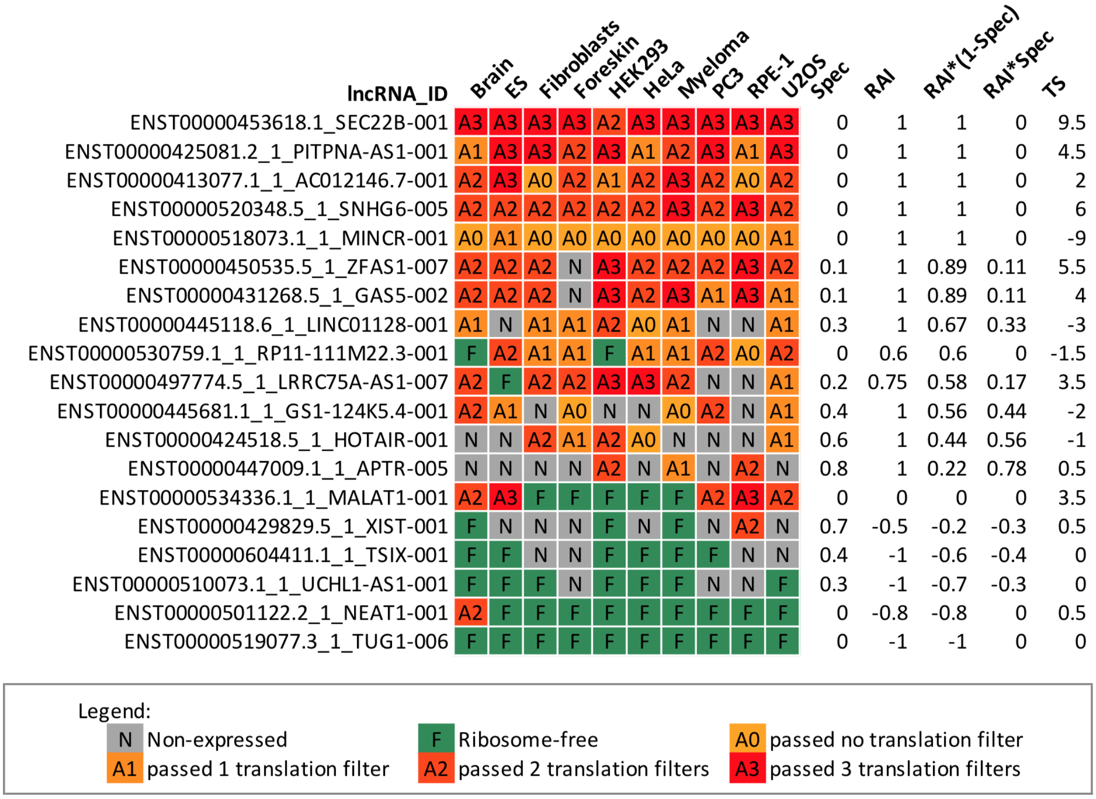


Fig. 2.3 **Analysis of coding potential by using FLOSS, RRS, and Framescore on the HeLa dataset.** (**a**) Venn diagram of putative ORFs in ribosome-associated lncRNAs evaluated by three coding filters (FLOSS, RRS, and Framescore). (**b**) Comparisons of the coding potential among CDSs (blue) and putative ORFs of ribosome-associated (red) or ribosome-free lncRNAs (green) for FLOSS, RRS, and Framescore, respectively. Based on these three coding metrics, we generated three cut-offs (the 10th percentiles represented as horizontal dashed lines) from CDSs to independently filter translation events for lncRNAs. For a coding filter of FLOSS, RRS, or Framescore, lncRNAs above the corresponding cut-off values (including the cut-off values) are identified as putatively translated lncRNAs according to this coding filter. (See Fig. S4 for the other datasets.)

For convenience, we label the translation of a lncRNA containing an ORF that passed 0–3 coding filters as “A0”–“A3”. When there are multiple ORFs in a lncRNA, we chose the one with the highest number of coding filters it passed. Finally, we obtained a preliminary classification of lncRNAs in each dataset. Figure 2.2c shows that 5,215 lncRNAs are expressed in HeLa cells, of which 3,238 are classified as ribosome-free, while the rest are classified as ribosome-associated. Furthermore, among the ribosome-associated lncRNAs, 978 were classified as “A0,” which means we have no evidence of translation events on these lncRNAs, while 165 were classified as “A3,” which indicates that at least one putative ORF has passed all three coding filters (FLOSS, RRS, and Framescore) and means that credible translation of them may be happening.

2.3.3 Identification of trans-lncRNAs, ribo-lncRNAs and noribo-lncRNAs across multiple datasets

As a measure of the reliability of ribosomal associations for a particular lncRNA, we developed RAI * (1 - *spec*) to assess the integrated confidence of specific ribosomal associations across multiple pieces of experimental evidence. Here, RAI is a metric that measures ribosomal association across datasets in which the target lncRNA was expressed; *spec* is a metric for transcript expression specificity. We used a binary value to represent the ribosome density for a lncRNA in each experiment, as the ribosome density is normalized to transcript abundance in each dataset, which complicates the use of ribosome density across different datasets directly. A lncRNA with an RAI * (1 - *spec*) value of 1 indicates that the transcript consistently interacts with ribosomes among multiple datasets, and an RAI * (1 - *spec*) value of -1 denotes that this transcript is highly dissociated from ribosomes. (See “Methods" in this chapter for the detailed definition of RAI * (1 - *spec*).) Table S3 lists the RAI * (1 - *spec*) values of all lncRNAs in the human and mouse datasets, respectively. As shown in Figs. 2.5a and 2.5d, we also used two threshold values—a low threshold at the 5th percentile and a high threshold at the 95th percentile—to determine high confident ribosome-free lncRNAs (termed “noribo-lncRNAs") and ribosome-associated lncRNAs (termed “ribo-lncRNAs"). A lncRNA was classified as a noribo-lncRNA when its RAI * (1 - *spec*) value fell below the lower threshold and as a ribo-lncRNA when its RAI * (1 - *spec*) value exceeded the upper threshold. It is worth noting that the terms “ribosome-associated lncRNAs" and



lncRNA_ID	Brain	ES	Fibroblasts	Foreskin	HEK293	HeLa	Myeloma	PC3	RPE-1	U2OS	Spec	RAI	RAI*(1-Spec)	RAI*Spec	TS
ENST00000453618.1_SEC22B-001	A3	A3	A3	A3	A2	A3	A3	A3	A3	A3	0	1	1	0	9.5
ENST00000425081.2_1_PITPNA-AS1-001	A1	A3	A3	A2	A3	A1	A2	A3	A1	A3	0	1	1	0	4.5
ENST00000413077.1_1_AC012146.7-001	A2	A3	A0	A2	A1	A2	A3	A2	A0	A2	0	1	1	0	2
ENST00000520348.5_1_SNHG6-005	A2	A2	A2	A2	A2	A2	A3	A2	A3	A2	0	1	1	0	6
ENST00000518073.1_1_MINCR-001	A0	A1	A0	A0	A0	A0	A0	A0	A0	A1	0	1	1	0	-9
ENST00000450535.5_1_ZFAS1-007	A2	A2	A2	N	A3	A2	A2	A2	A3	A2	0.1	1	0.89	0.11	5.5
ENST00000431268.5_1_GAS5-002	A2	A2	A2	N	A3	A2	A3	A1	A3	A1	0.1	1	0.89	0.11	4
ENST00000445118.6_1_LINC01128-001	A1	N	A1	A1	A2	A0	A1	N	N	A1	0.3	1	0.67	0.33	-3
ENST00000530759.1_1_RP11-111M22.3-001	F	A2	A1	A1	F	A1	A1	A2	A0	A2	0	0.6	0.6	0	-1.5
ENST00000497774.5_1_LRRC75A-AS1-007	A2	F	A2	A2	A3	A3	A2	N	N	A1	0.2	0.75	0.58	0.17	3.5
ENST00000445681.1_1_GS1-124K5.4-001	A2	A1	N	A0	N	N	A0	A2	N	A1	0.4	1	0.56	0.44	-2
ENST00000424518.5_1_HOTAIR-001	N	N	A2	A1	A2	A0	N	N	N	A1	0.6	1	0.44	0.56	-1
ENST00000447009.1_1_APTR-005	N	N	N	N	A2	N	A1	N	A2	N	0.8	1	0.22	0.78	0.5
ENST00000534336.1_1_MALAT1-001	A2	A3	F	F	F	F	F	A2	A3	A2	0	0	0	0	3.5
ENST00000429829.5_1_XIST-001	F	N	N	N	F	N	F	N	A2	N	0.7	-0.5	-0.2	-0.3	0.5
ENST00000604411.1_1_TSIX-001	F	F	N	N	F	F	F	F	N	N	0.4	-1	-0.6	-0.4	0
ENST00000510073.1_1_UCHL1-AS1-001	F	F	F	N	F	F	F	N	N	F	0.3	-1	-0.7	-0.3	0
ENST00000501122.2_1_NEAT1-001	A2	F	F	F	F	F	F	F	F	F	0	-0.8	-0.8	0	0.5
ENST00000519077.3_1_TUG1-006	F	F	F	F	F	F	F	F	F	F	0	-1	-1	0	0

Legend:

N Non-expressed	F Ribosome-free	A0 passed no translation filter
A1 passed 1 translation filter	A2 passed 2 translation filters	A3 passed 3 translation filters

Fig. 2.4 **The ribosomal association index (RAI) enables an integrative analysis of ribosome associability of lncRNAs across multiple independent datasets.** The table summarizes ribosomal association and translation for selected human lncRNAs. Rows represent lncRNAs, while colored columns denote datasets. For each lncRNA, “N (gray)” and “F (green)” cells correspond to unexpressed and ribosome-free lncRNAs, respectively. “A0”∼“A3” cells represent the lncRNA containing a putative ORF that passed 0–3 coding filters. The last four columns are statistics that describe the corresponding lncRNAs. Spec is the transcript expression specificity, ranging from 0 (ubiquitous) to 1 (specific). For a lncRNA, RAI is the ribosomal association index across datasets in which this lncRNA is expressed, ranging from -1 (ribosome-free) to 1 (ribosome-associated). RAI * (1 - *spec*) is a metric to measure the confidence of ribosomal association for a lncRNA that has a broad expression, ranging from -1 (lncRNA was observed as ribosome-free in most datasets) to 1. Conversely, RAI * spec can be used to select ribosome-associated or ribosome-free lncRNAs from the population of tissue-specific lncRNAs. TS can be used with RAI * (1 - *spec*) to filter the putatively translated lncRNAs. (See Table S3 for a complete list of human and mouse lncRNAs.)

“ribosome-free lncRNAs" mentioned above are particularly used to categorize lncRNAs in a single dataset, whereas the terms “ribo-lncRNAs" and “noribo-lncRNAs" are defined across multiple datasets.

Furthermore, for ribo-lncRNAs that were widely expressed and commonly associated with ribosomes across multiple tissues or cell types, we determined if there are lncRNAs that can be translated. We presented the translation score (TS), a weighted sum function of translation events (A0$\sim$A3), to evaluate the coding capacity for each ribo-lncRNA. The TS value for a lncRNA is expected to be positively related to the likelihood of this lncRNA contains an ORF encoding a peptide. We separated ribo-lncRNAs within the top 5% of TS values as putatively translated lncRNAs (termed “trans-lncRNAs"). (See Figs. 2.5b and 2.5e.) Overall, 746 noribo-lncRNAs, 613 ribo-lncRNAs, and 78 trans-lncRNAs in human (326 noribo-lncRNAs, 367 ribo-lncRNAs, and 42 trans-lncRNAs in mouse) were identified in this study (see Table S3 for the complete list of trans-lncRNAs, ribo-lncRNAs, and noribo-lncRNAs).

Footprint alignments, which were used to distinguish between ribosome-associated and ribosome-free lncRNAs, are more likely to occur on a longer transcript sequence. Thus, the first step is to evaluate the effect of transcript length on the RAI * (1 - *spec*) metric. We compared the transcript length among the trans-lncRNAs, ribo-lncRNAs, and noribo-lncRNAs (see Figs. 2.5c and 2.5f). Although we observed that the ribo-lncRNAs tended to be longer than noribo-lncRNAs in the human datasets ($p < 0.05$), we also found the opposite result in the mouse datasets ($p < 0.01$), which suggests that transcript length may not the dominant factor affecting the ribosomal association of lncRNAs. We also observed that, on average, the trans-lncRNAs were the longest in both human and mouse, suggesting transcript length is one of the important features that determines whether a transcript can encode a peptide.

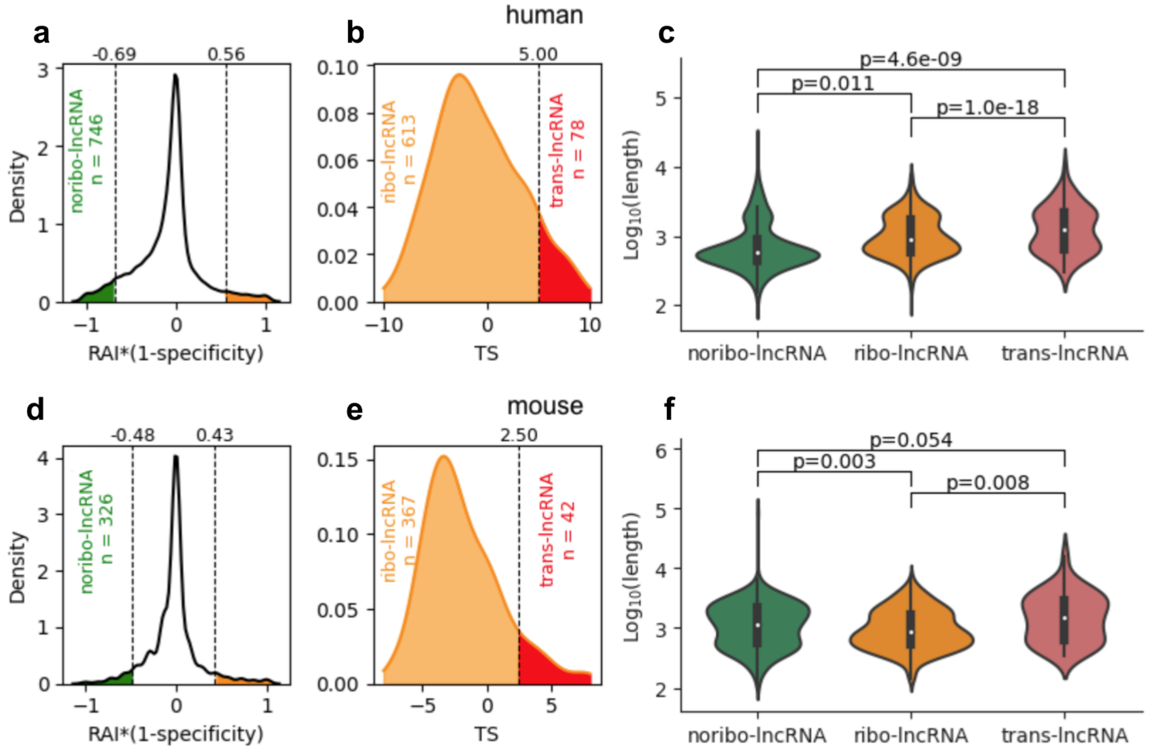


Fig. 2.5 **Classification of trans-lncRNAs, ribo-lncRNAs, and noribo-lncRNAs.** (**a**) The kernel density of the RAI * (1 - spec) scores for human lncRNAs. Two vertical dashed lines represent the 5th percentile (left, upper bound for the reliable ribosome-free lncRNAs, termed “noribo-lncRNAs (green)") and the 95th percentile (right, lower bound for the reliable ribosome-associated lncRNAs (orange) for further classification) of the RAI * (1 - spec) scores. (**b**) The kernel density of the TS scores for human ribosome-associated lncRNAs identified in (**a**). Top 5% of lncRNAs were classified as “trans-lncRNAs (red)" suggesting that stable translation events are likely to occur among them. The remaining lncRNAs were finally classified as “ribo-lncRNA (orange)" indicating that there is an interaction with ribosomes in this part of lncRNAs, but no strong translation activity was observed. (**c**(Comparisons among trans-lncRNAs, ribo-lncRNAs, and noribo-lncRNAs for their transcript lengths in human. (**d**)–(**f**) show the results for mouse; *p*-values in (**c**) and (**f**) were calculated using two-sample *t*-tests.

2.3.4 Exploring the biological characteristics of ribosome-associated lncRNAs

Next, we investigated the biological characteristics of ribosome-associated lncRNAs to determine their coding potential, sensitivity to nonsense-mediated decay, and cellular localization.

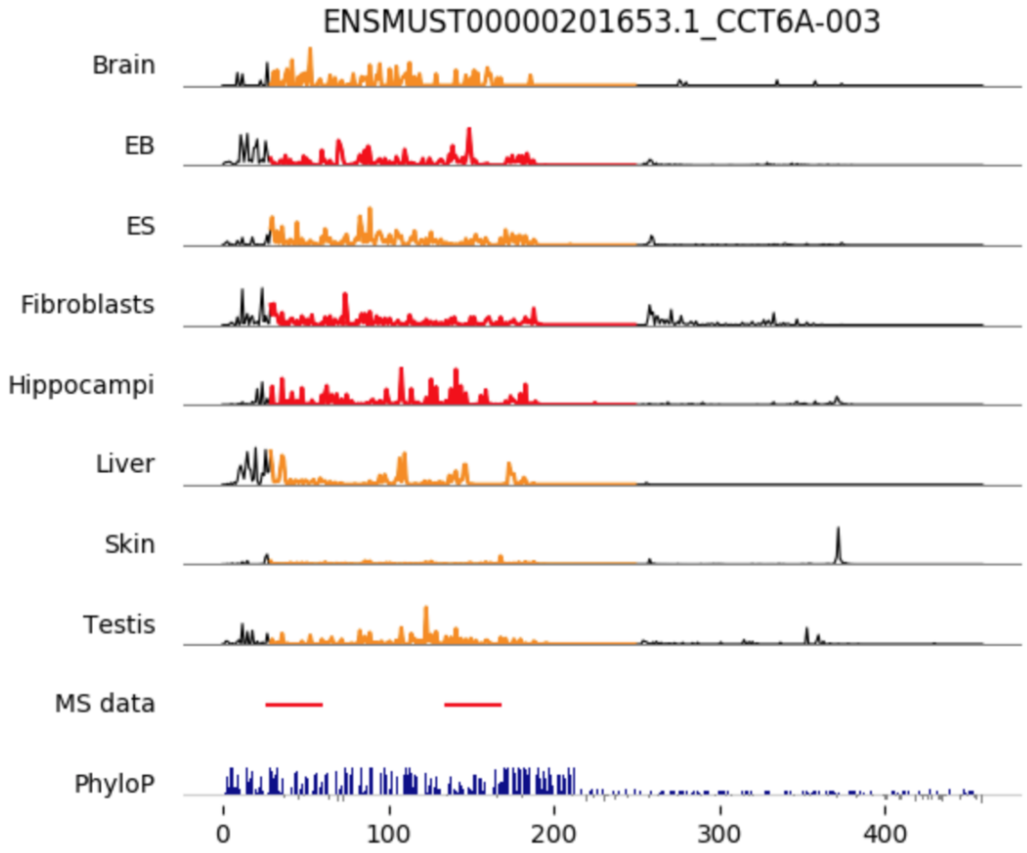


Fig. 2.6 **Overlapping of ribosome footprint coverage, mass spectrometry data, and sequence conservation (phyloP score) across mouse lncRNA CCT6A.** Top eight panels indicate the ribosome coverage (arbitrary unit) across the CCT6A-003 transcript, where the colored region represents a putatively translated ORF identified by applying three coding metrics (FLOSS, RRS, and Framescore). Orange and red regions indicate this putative ORF has passed two and three coding filters, respectively. The MS data panel shows the overlapping of peptides transformed from mass spectrometry data in this transcript. The phyloP panel shows the base-wise conservation scores with positive values (blue) meaning slower evolution than expected, and negative values (gray) suggesting faster evolution than expected.

Coding potential

To investigate whether the trans-lncRNAs detected in this study are consistent with mass spectrometry data, we aligned peptide sequences that were transformed from mass spectrometry data to lncRNAs. As expected, the lncRNAs with mappable peptides were significantly enriched among the trans-lncRNAs and ribo-lncRNAs for human and mouse (all $p < 0.001$, see Table 2.7). In particular, the trans-lncRNAs were associated with the highest odds ratios (8.28 and 23.03 for human and mouse, respectively), which indicates that trans-lncRNAs have the highest potential for coding peptides. Figure 2.6 shows the footprint coverage, peptide alignment, and sequence conversation (phyloP score) for trans-lncRNA ENSMUST00000201653.1_CCT6A-003. For the footprint coverage, a colored region indicates the putative ORF predicted in this lncRNA. The peptide sequences transformed from the mass spectrometry data are consistently mapped onto this putative ORF. Also, we observed positive phyloP scores for the putative ORF, which indicates that this putative ORF sequence is evolutionarily conserved. Both metrics supported the hypothesis that the trans-lncRNA can encode peptides (see Tables S4–S5 and S8–S9 for the details of other putative ORFs).

Table 2.7 Long non-coding RNAs supported by mass spectrometry data. (One-sided Fisher's exact test: ***p<0.001)

	Human		Mouse	
	#Total	#MS supported (odds ratio)	#Total	#MS supported (odds ratio)
trans-lncRNA	78	*** 5 (8.28)	42	*** 7 (23.03)
ribo-lncRNA	613	*** 18 (4.16)	367	*** 10 (3.82)
noribo-lncRNA	746	2 (0.32)	326	2 (0.73)
other	12209	85 (0.40)	5525	33 (0.23)
Total	13646	110	6260	52

Cellular localization

We sought to examine whether the ribosome-associated lncRNAs are enriched in the cytoplasm where the ribosomes are located. Here, we used expression fold change, which compared the abundance of lncRNAs from the nuclear or the cytoplasmic fraction, to quantify the subcellular localization in HeLa cells (see "Methods" in this chapter for details for generating the fold changes). Figure 2.7a indicates the kernel density of expression fold changes from the cytoplasmic fractions to the nuclear fractions for either ribosome-associated and ribosome-free lncRNAs. As expected, both the ribosome-associated lncRNAs and the ribosome-free lncRNAs were more likely to exist in the nucleus ($mean = 2.19$ and 1.12 for ribosome-free lncRNAs and ribosome-associated lncRNAs, respectively). However, if compared with the ribosome-free lncRNAs, the ribosome-associated lncRNAs have a significant tendency to be present in the cytoplasm ($p < 0.001$).

Sensitivity to nonsense-mediated decay

To test whether the ribo-lncRNAs are associated with nonsense-mediated decay (NMD), we investigated the differences in expression levels of various RNA populations in the presence (control) or absence (UPF1_KD) of NMD (see "Methods" in this chapter for details to generate the fold change values). In HeLa cells, Figure 2.7b is the kernel density of

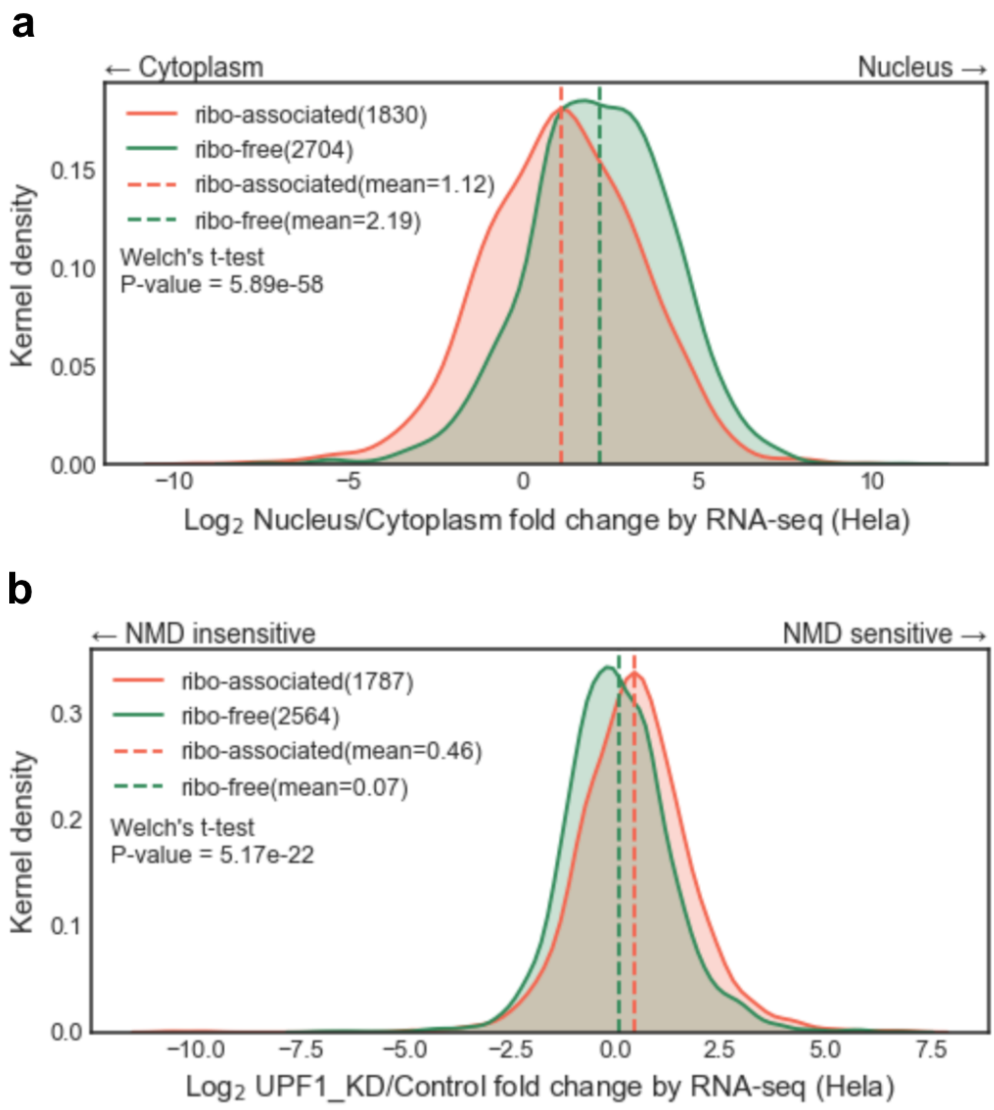


Fig. 2.7 **Comparisons between ribosome-associated lncRNAs and ribosome-free lncRNAs in HeLa cells.** (**a**) Cellular localization analysis. The fold changes of expression values were calculated between the nuclear and the cytoplasmic compartments to quantify the localization. (See Table S6 for the raw data used to generate this kernel density plot.) (**b**) Nonsense mediated decay (NMD) analysis. As UPF1 is an important NMD factor, we can use the fold changes of expression values between samples from a UPF1 knockdown and control to express NMD sensitivity. (See Table S7 for the raw data used to generate this kernel density plot.) The corresponding mean values are shown by vertical dashed lines; *p*-values were calculated using Welch's *t*-test.

expression fold changes from the control samples to UPF1 knockdown samples for either ribosome-associated or ribosome-free lncRNAs. In our observations, the expression level of ribosome-free lncRNAs was not affected by NMD (*mean* = 0.07). Interestingly, we found the expression level of ribosome-associated lncRNAs were significantly sensitive to NMD compared to ribosome-free lncRNAs (*mean* = 0.46, $p < 0.001$).

2.4 Discussion

We emphasize that the term ribosomal association in this study refers to the ribosome translating or binding of a transcript, as ribosomes not only translate proteins but may also carry out other unclear functions by interacting with transcripts. To our knowledge, this is the first comprehensive study of ribosome–lncRNA interactions across multiple ribosome-profiling experiments in mammals, and it has several differences from previous studies: (i) more lncRNAs, including lincRNAs (long intergenic RNAs), were examined; (ii) a main focus on human and mouse because of the well-annotated lncRNAs for these two species; (iii) the use of the ribosome density metric and the cut-off value derived from $3'$ UTRs to detect ribosomal associations of lncRNAs, which thus obtained robust detection rates of ribosome-associated lncRNAs over multiple independent datasets. We developed a novel tool, RAI * (1 - *spec*), to measure ribosomal association from multiple ribosome-profiling experiments. By using the RAI * (1 - *spec*) metric, we determined high-confidence ribosome-associated lncRNAs (ribo-lncRNAs) and ribosome-free lncRNAs (noribo-lncRNAs) and investigated the biological characteristics of ribosome-associated and ribosome-free lncRNAs involving coding potential, cellular localization, and NMD sensitivity.

Processed transcripts and retained introns were observed to prefer to associate with ribosomes, which suggests these two biotypes of lncRNAs are related to either protein-coding or ribosome-mediated regulation. For example, SEC22B has two transcript variants in the human genome, both of which were annotated as “processed transcript that does not contain an ORF" in GENCODE v25lift37(release 25 mapped to GRCh37). However, they had high RAI * (1 - *spec*) scores (both are 1, see Fig. 2.4 and Table S3), indicating their strong association with ribosomes. Moreover, we also observed a high translation score for SEC22B-001 (TS = 9.5), which indicates there is credible translation activity on this transcript. Strikingly, we found that SEC22B was removed from lncRNA category and annotated as a “protein coding" transcript in human genome h19 (GRCh38). Indeed, compared to ribosome-free lncRNAs, ribosome-associated lncRNAs have a higher protein-coding potential in the light of three variant coding metrics—FLOSS, RRS, and Framescore (see Figs. 2.3b and S4). This particular case of SEC22B suggests that some lncRNAs with high RAI * (1 - *spec*) and TS values could be protein/peptide coding transcripts. It seems plausible to use the RAI * (1 - *spec*) and TS in combination to examine the coding capacity of lncRNAs.

Table 2.8 LncRNAs derived from snoRNA host genes are enriched in trans-lncRNAs and ribo-lncRNAs. (One-sided Fisher’s exact test: **p<0.01, ***p<0.001)

	Human		Mouse	
	#Total	#snoRNA host (odds ratio)	#Total	#snoRNA host (odds ratio)
trans-lncRNA	78	*** 22 (20.46)	42	** 5 (7.26)
ribo-lncRNA	613	*** 70 (10.71)	367	*** 44 (11.21)
noribo-lncRNA	746	5 (0.42)	326	*** 57 (20.75)
other	12209	112 (0.14)	5525	1 (0.00)
Total	13646	209	6260	107

Most snoRNAs are located in introns of ribosomal protein genes and of genes encoding translation factors or nucleolar proteins. However, several noncoding genes are also reported as hosts for small nucleolar RNA (snoRNA) expression. Notably, as shown in Table 2.8, we observed snoRNA host gene-derived lncRNAs enriched in both trans-lncRNAs and ribo-lncRNAs, suggesting their interaction with ribosomes, which is consistent with previous studies [25, 104–106]. One possible reason for their association with ribosomes is that such lncRNAs are by-products of snoRNA production and are targeted to ribosomes, thus triggering the nonsense-mediated decay (NMD) pathway. Host gene-derived lncRNAs were reported to be sensitive to NMD [105], which provides indirect support of this hypothesis. In particular, GAS5 (growth arrest-specific 5) and ZFAS1 (ZNFX1 Antisense RNA 1), which were revealed by ribosomal association analysis in this study, have been reported to be associated with distinct biological functions. The GAS5 lncRNA sequence was determined to control transcriptional activity of apoptosis-related genes, while the NMD pathway appears to regulate the abundance of GAS5 transcripts [106]. The ZFAS1 lncRNA sequence was primarily identified to interact with the 40S ribosome subunit and reported to affect ribosomal protein modification [107]. The ZFAS1–ribosome interaction was also conserved in mouse (see Table S3), which suggests that the lncRNA may play a role in targeting the ribosome.

For lncRNAs, the dissociation of ribosomes illuminates lncRNA localization and functional studies. NEAT1 (nuclear enriched abundant transcript 1) is known to be a nuclear-enriched lncRNA. NEAT1 has been found to function as an important structural determinant of nuclear paraspeckles [108], which corresponds to the apparent ribosome-free NEAT1 (RAI *(1 - *spec*) = -0.8). TUG1 (taurine up-regulated gene 1) is a PRC2 (polycomb repressive complex 2)-associated lncRNA involved in cell-cycle regulation [109]. The longest transcript variant of TUG1 was highly ribosome-free (RAI * (1 - *spec*) = -1)). A TUG1 transcript variant of the human (ENST00000569149, RAI * (1 - *spec*) = 0.2)) and two transcript variants of the mouse (ENSMUST00000193809 and ENSMUST00000132077 with RAI * (1 - *spec*) = 1 and -1, respectively), on the contrary, displayed entirely different ribosomal association characteristics. This is also consistent with the finding that a unique peptide maps to TUG1 [110]. We, therefore, concluded that different transcript variants of lncRNAs act with different ribosome-associated properties, which may suggest a new functional class of lncRNAs regulated by alternative splicing coupled with ribosome targeting.

2.5 Conclusions

In this study, we applied ribosome profiling data to identify interactions between lncRNAs and ribosomes. To our knowledge, this is the first report showing that a large fraction of lncRNAs–ribosome interactions over multiple independent studies are consistent and reliable in human and mouse. We developed the ribosomal association index (RAI) and used it with transcript expression specificity (spec) to measure the degree of reliability of lncRNA-ribosome interactions across multiple datasets. Furthermore, we used three different coding metrics (FLOSS, RRS, and Framescore) to assess the coding potential for ribosome-associated lncRNAs. LncRNAs detected to associate with ribosomes were observed to be more likely to be located in the cytoplasm and be more sensitive to NMD compared to ribosome-free lncRNAs. We also noticed that many ribosome-associated lncRNAs are tissue- or splicing-specific, which suggests these lncRNAs may target ribosomes under specific conditions to perform certain special functions. An interesting goal for future research is determining the biological mechanism underlying the condition-specific ribosomal association for lncRNAs. Future research may also identify the genomic characteristics of ribosome-associated lncRNAs and develop a method for distinguishing ribosome-associated lncRNAs from other RNA species. The complete list of ribosome associations of known lncRNAs in human and mouse are available online, from Table S3, which will be a useful resource for functional lncRNA studies.

Chapter 3

Identifying sequence features that drive ribosomal association for lncRNA

3.1 Introduction

With the advancement of high-throughput sequencing technology, the lncRNA population has begun to emerge. In the past few decades, we have had a new understanding of this type of RNA that their number far exceeds the protein-coding gene in human and mouse [111]. However, it is still unclear what function most of the lncRNAs have [37]. Moreover, it is difficult to predict the lncRNA genes from other organisms without sequence characteristics of lncRNAs[111].

Here, we discuss ribosome-associated lncRNAs, which are interacting with the ribosomes although we did not have evidence for their protein translation. Such lncRNAs are considered to have the function of regulating translation [112, 113]. The ribosome-associated lncRNAs are also reported to serve as a source of new peptides [45]. Several individual studies have found encoded peptides from lncRNAs, which have been reviewed in [114]. However, due to the limited number of ribosome-associated lncRNAs, it is difficult to understand in depth what are the essential features (or regulatory elements) included in the lncRNAs that control their association with the ribosome. Characterization of ribosome-associated lncRNAs play a crucial role in understanding the involvement of lncRNA in specific biological functions or which possible regulatory mechanisms.

Ribosome profiling is a technique that collect and read RNA fragments, which are protected by the ribosome. It provides us a way to investigate the genome-wide association of lncRNAs with ribosomes. In the previous work [115], we have analyzed ribosome profiling data and identified 613 ribosome-associated lncRNAs (ribo-lncRNAs) and 746 ribosome-free lncRNAs (noribo-lncRNAs) from human (367 ribo-lncRNAs and 326 noribo-lncRNAs from mouse).

In this study, we investigated which sequence features could distinguish between these two lncRNAs. To our knowledge, this is a first study of characterizing ribosome-associated lncRNAs. Such sequence features identified in this study are possible to be considered as regulatory factors that play an essential role in the ribosomal association.

3.2 Methods

3.2.1 Datasets and potential features

Ribo-lncRNAs and noribo-lncRNAs were derived from our previous study [115]. We used Blast [116] to remove lncRNAs that share sequences of high similarity. If the sequence similarity between two lncRNAs exceeded 60% (of the shorter one), then it is considered as high similarity and hence the shorter one is discarded (Table 3.1 shows the statistics of dataset before and after removing lncRNAs of high similarity). All sequence features considered to affect ribosome association were listed in Table 3.2. For each feature column, we imputed missing data by using mean value.

Table 3.1 Statistics of dataset used in this study. The “reduced” column shows the number of lncRNAs after removing sequences of high similarity.

	Human		Mouse	
	Original	Reduced	Original	Reduced
ribo-lncRNA	613	**487**	367	**279**
noribo-lncRNA	746	**681**	326	**300**
Total	1359	**1168**	693	**579**

Primary/first/upstream ORF

We defined three different types of putative open reading frames (ORFs) on a lncRNA (Fig. 1). A primary ORF (pORF) is the longest ORF starting with ATG. A first ORF (fORF) starts with ATG and is closest to the $5'$ end of the lncRNA. An upstream ORF (uORF) starts with a near-cognate initiation site (i.e. CTG, GTG, or TTG [25]). Here, the uORF is considered only when an existing pORF located in the lncRNA; the beginning and end of uORF should be upstream of the pORF. These three types of ORFs above are all terminated with a TAG, TGA, or TAA. In addition, the upstream ORF overlapping with the primary ORF was not analyzed in this study.

Context/trimer/hexamer score

For the three types of ORFs mentioned above, we defined three scores based on frequency ratio between ribo-lncRNAs and noribo-lncRNAs. Context sequence score of ORF start (hereinafter abbreviated as “context score”) is the sum of frequency ratios of nucleotides at -6 to +3 positions relative to the ORF start. Trimer score and hexamer score are summed frequency ratios of trinucleotide or hexanucleotide, respectively, during ORFs. These three metrics can be calculated using the following formula (which is also applied to assess coding potential in CPAT[117]):

$$\text{Context/Trimer/Hexamer score} = \frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{F(x_i)}{F'(x_i)}\right) \tag{3.1}$$

where, for context score, $x_i \in [A, C, G, T]$ represents the nucleotide at the i-th position while i indicates the index of the relative position above (i = 1 .. 10). $F(\cdot)$ and $F'(\cdot)$ are the occurrence frequencies of position-specific nucleotide in categories of ribo-lncRNA and noribo-lncRNA, respectively. For trimer score and hexamer score, ORF sequence is converted into a sequence of length n in units of trinucleotide and hexanucleotide, respectively. Thus, x_i represents the unit (trimer or hexamer), $F(\cdot)$ and $F'(\cdot)$ are the occurrence frequencies of unit in ribo-lncRNAs and noribo-lncRNAs, respectively. Both $F(\cdot)$ and $F'(\cdot)$ need to be calculated in advance from a control dataset to generate a lookup table. Hence, we randomly selected 5,000 CDS sequences to calculate $F(\cdot)$ and shuffled those sequences to generated $F'(\cdot)$.

Stem probability

A higher stem probability means a stronger RNA secondary structure in this context. To investigate whether RNA secondary structure affects the ribosomal association, we used ParasoR [118], which is specifically designed for RNA secondary structure prediction of numerous and long RNAs, to predict the stem probability of each base in an lncRNA. We set the parameter $-constraint$ to $N-1$, where N is the length of the lncRNA, in order to consider all possible base pairs during the lncRNA. Except it was an extreme long ($>$ 9,500nt) RNA, we used the default parameter ($-constraint = 200$) to guarantee the prediction result in a limited time.

N^6-Methyladenosine modification, G-quadruplex, and repeat element

We used SRAMP [119] to predict N^6-Methyladenosine modification (m^{6}A) sites in an lncRNA. G-quadruplex (G4) segments were predicted by using QGRS [120]. G4 element with G-score ≥ 30 is considered as a stable G-quadruplex structure. Transposon elements (TEs) annotations were obtained from RepeatMasker [121]. We used the repeat library (build on 20140131) that mapped to human (hg19) and mouse (mm10), respectively. Repeat elements annotated as simple repeats, low-complexity, or non-coding RNA were removed.

Table 3.2 Sequence features were considered to influence the ribosomal association.

No.	Feature	Description
Basic		
1	fLen	Log$_{10}$(length+1) of the mature lncRNA
2	gc	G+C content of the mature lncRNA
RNA splicing		
3	nE	Number of exons
4	fELen	Log$_{10}$(length+1) of the first exon
5	minELen	Log$_{10}$(length+1) of the shortest exon
6	maxELen	Log$_{10}$(length+1) of the longest exon
7	avgELen	Log$_{10}$(averaged_length+1) of exons
8	fEgc	G+C content of the first exon
9	minEgc	G+C content of the shortest exon
10	maxEgc	G+C content of the longest exon
11	avgEgc	Averaged G+C content of exons
12	fILen	Log$_{10}$(length+1) of the first intron
13	minILen	Log$_{10}$(length+1) of the shortest intron
14	maxILen	Log$_{10}$(length+1) of the longest intron
15	avgILen	Log$_{10}$(averaged_length+1) of introns
16	fIgc	G+C content of the first intron
17	minIgc	G+C content of the shortest intron
18	maxIgc	G+C content of the longest intron
19	avgIgc	Averaged G+C content of introns

Table 3.2 Sequence features (continued)

No.	Feature	Description
Putative ORF (pORF: primary ORF; fORF: first ORF; uORF: upstream ORF)		
20-22	p/f/uOrfLen	Log$_{10}$(length + 1) of ORF
23-25	p/f/uOrfCov	Percentage of ORF length compared to that of lncRNA
26-28	p/f/uOrf5utrLen	Log$_{10}$(length + 1) of the upstream region of ORF (5$'$ UTR)
29-31	p/f/uOrf5utrCov	Percentage of the 5$'$ UTR length compared to that of lncRNA
32-34	p/f/uOrf3utrLen	Log$_{10}$(length + 1) of the downstream region of ORF (3$'$ UTR)
35-37	p/f/uOrf3utrCov	Percentage of the 3$'$ UTR length compared to that of lncRNA
K-mer frequency		
38-40	p/f/uOrfStartContext	Context sore of ORF start
41-43	p/f/uOrfSeqTrimer	Trimer score of ORF
44-46	p/f/uOrfSeqHexamer	Hexamer score of ORF
RNA secondary structure		
47-49	p/f/uOrfSp	Averaged RNA stem probability of ORF
50-52	p/f/uOrf5utrSp	Averaged RNA stem probability of 5$'$ UTR
53-55	p/f/uOrf5utrSpFC	Ratio of RNA stem probability of 5$'$UTR to that of ORF
56-58	p/f/uOrf3utrSp	Averaged RNA stem probability of 3$'$ UTR
59-61	p/f/uOrf3utrSpFC	Ratio of RNA stem probability of 3$'$UTR to that of ORF
62	g4NearTIS_log	Log$_{10}$(minimum distance) from G4 to transcription initiation
63	g4NearTTS_log	Log$_{10}$(minimum distance) from G4 to transcription termination
64-66	g4Near(p/f/u)ORFstart_log	Log$_{10}$(minimum distance) from G4 to ORF start
67-69	g4Near(p/f/u)ORFend_log	Log$_{10}$(minimum distance) from G4 to ORF end
70	g4NearTIS_%	Minimum distance from G4 to TIS divided by length of lncRNA
71	g4NearTTS_%	Minimum distance from G4 to TTS divided by length of lncRNA
72-74	g4Near(p/f/u)ORFstart_%	Minimum distance from G4 to ORF start divided by length of lncRNA
75-77	g4Near(p/f/u)ORFend_%	Minimum distance from G4 to ORF end divided by length of lncRNA
RNA modification		
78	m6aNearTIS_log	Log$_{10}$(minimum distance) from m^{6}A to transcription initiation
79	m6aNearTTS_log	Log$_{10}$(minimum distance) from m^{6}A to transcription termination
80-82	m6aNear(p/f/u)ORFstart_log	Log$_{10}$(minimum distance) from m^{6}A to ORF start

Table 3.2 Sequence features (continued)

No.	Feature	Description
83-85	m6aNear(p/f/u)ORFend_log	Log_{10}(minimum distance) from m^6A to ORF end
86	m6aNearTIS_%	Minimum distance from m^6A to TIS divided by length of lncRNA
87	m6aNearTTS_%	Minimum distance from m^6A to TTS divided by length of lncRNA
88-90	m6aNear(p/f/u)ORFstart_%	Minimum distance from m^6A to ORF start divided by length of lncRNA
91-93	m6aNear(p/f/u)ORFend_%	Minimum distance from m^6A to ORF end divided by length of lncRNA
Repeat element		
94	DNA	Containing DNA transposon or not
95	LINE	Containing LINE element or not
96	LTR	Containing LTR element or not
97	SINE	Containing SINE element or not
98	Retroposon	Containing Retroposon element or not
99	Satellite	Containing Satellite element or not

3.2.2 $\mathcal{L}1$-regularized logistic regression

Logistic regression (LR) model [122] can be used as a binary classifier which applies a logistic function to turn linear predictions to [0, 1]. Given a set of labeled training data X (feature vectors) and their labels y (i.e. 0 and 1 indicates noribo-lncRNA and ribo-lncRNA, respectively), LR model seeks to minimize the loss (or objective) function:

$$\min_{w,c} \|w\|_1 + C\sum_{i=1}^{n} \log(\exp(-y_i(X_i^T w + c)) + 1). \tag{3.2}$$

To avoid the over-fitting, in which a complicate (many parameters and parameters with a large variance) model can perform perfectly on training dataset but badly on testing dataset, a regularization term ($\|w\|_1$) was used to control the complexity (i.e. the number and the values of parameters) of model. Moreover, $\mathcal{L}1$-based regularization drives parameters to zero, which is a natural process of feature selection. After training the LR model, we get a small number of features with non-zero coefficients. Since the feature value has been scaled in the same range, the absolute value of the coefficient represents how much the change of this feature has an effect on the prediction of the model, and can be used to express the importance of this feature in classification. The choice of using the model is based on following reasons: First, the model uses a logistic function to transform the prediction results to a range of 0 to 1, which is suitable for a two-class problem involved in this study; Second, $\mathcal{L}1$-regularization drives the model to tend to adopt a sparse feature space during training, that is, the coefficients of many features will be zero, resulting in the model naturally selects features for us; Finally, a linear combination of all features is considered in the model. Thus, a positive/negative sign of the coefficient of the feature indicates that a positive/negative correlation with the result of prediction (i.e. ribo-lncRNA), and an absolute value of the coefficient can be used to describe the importance of the responding feature.

Feature selection by using the $\mathcal{L}1$-regularized logistic model becomes a univariate problem of how to select a hyperparameter C. Here, C represents the inverse of regularization strength. As C is increased, the number of features with non-zero coefficients is increased, and the model becomes more complicated. Thus, the criteria used in this study is that the most appropriate C should be to select fewer non-zero feature coefficients while still ensuring that the model has relatively high prediction accuracy. For this purpose, we divided all data into a training set and test set in a ratio of 80:20, and the training set was further applied for 5 fold cross validation. When we determine a value of C, the model optimizes all the feature coefficients on the training set. Then the performance of the optimized model was evaluated on the test set using accuracy metric:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{3.3}$$

where, TP is number of true positives, FP is number of false positives, TN is number of true negatives, and FP is number of false negatives. We used the Python scikit-learn library [123] to perform all the machine learning processes mentioned above.

Results

3.2.3 Defining ninety-nine features from lncRNA sequence

We considered factors that may cause lncRNA to associate with ribosome in terms of RNA splicing, putative ORF, k-mer frequency, RNA secondary structure, RNA modification, and repeat elements. A full list of extracted features is included in Table 3.2.

RNA splicing

To investigate the relationship between splicing and ribosomal association, we mainly examined length and G+C content of intron and exon. Because the first exon and intron was important for alternative splicing [124–126], their length and G+C content were also included in our feature set.

Putative ORF and k-mer frequency

We first defined three types of ORFs (primary, first, and upstream), then extracted sequence features based on them (see "Methods" in this chapter for more details). As shown in Fig. 3.1a, pORF is the longest ORF which is considered most frequently as a possible translated region; fORF is the ORF closest to the $5'$ end of the lncRNA which was selected because of the first-ATG rule [127]; uORF locates in the upstream of the primary ORF starting with near-cognate initiation site (i.e., CTG, GTG, or TTG). Other ORFs located inside or in the downstream of the primary ORF were excluded to ensure the simplicity of the problem.

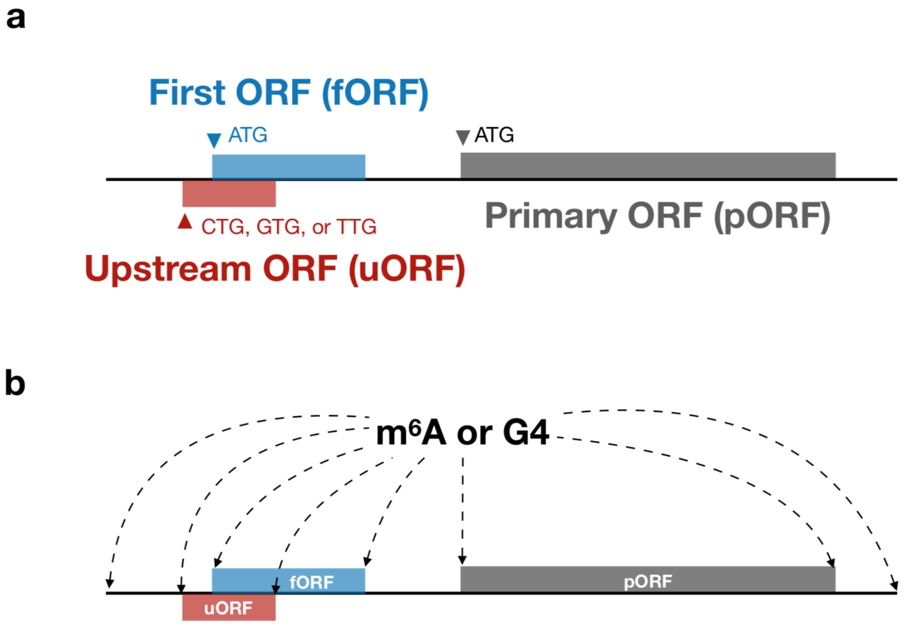


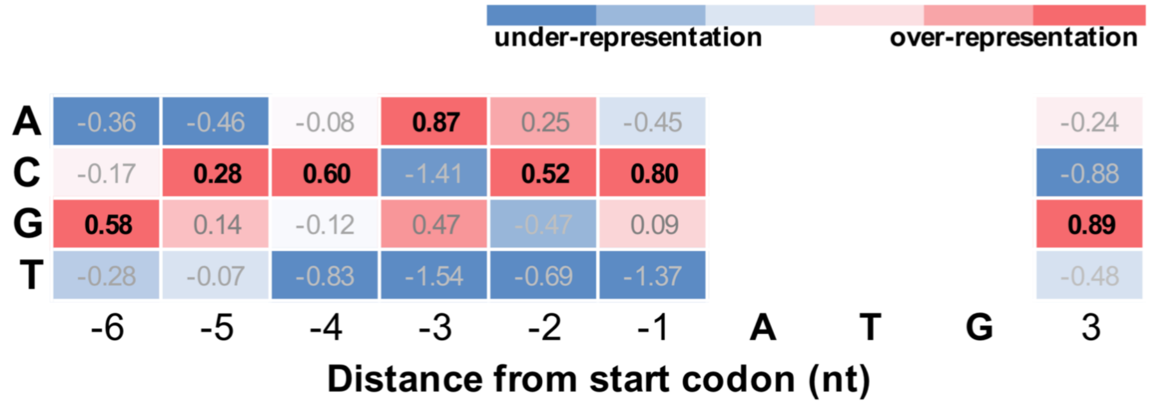
Fig. 3.1 **Example of feature extraction.** (**a**) Representation of primary ORF (pORF, gray), first ORF (fORF, blue), and upstream ORF (uORF, red) in a lncRNA. Horizontal line indicates a mature lncRNA, boxes represent putative open reading frames (ORFs) defined on this lncRNA. (**b**) Relationship (distance) between m^{6}A/G4 and transcript initiation site (TIS), transcript termination site (TTS), and starts or ends of u/f/pORF were used as features. Direct distance (bases in log scale) and relative distance (percentage of the length of lncRNA) were considered to express the relationship.

ORF length is a discriminating feature for coding and non-coding RNAs[117], hence we questioned whether this feature can also contribute to the detection of ribosome-associated lncRNA. As it was reported that 3′ UTR length may regulate the translation efficiency [128] and 5′ UTR may contain RNA modification [129] or regulatory motif (e.g., G-quadruplex [130]), they were also considered in this investigation. Moreover, we used trimer score and hexamer score to assess whether the codon usage and bi-codon frequency were similar to CDS. To calculate trimer (or hexamer) score, we first randomly selected 5,000 CDSs as active ORF reference and randomly shuffled their sequences as inactive ORF reference (Table S10). Each trimer (or hexamer) has a weight, which is the ratio of its occurrence frequency in the two reference groups. For a given putative ORF, we calculated the weight of all trimers (or hexamers), and then took the mean to represent its trimer (or hexamer) score (see "Methods" in this chapter). Thus, trimer (or hexamer) score measures the degree of trimer (or hexamer) usage bias in a specified putative ORF. A positive score indicates a possible active ORF, whereas a negative score indicates an inactive one.

A consensus sequence, termed Kozak sequence, surrounds the start codon in eukaryotic mRNAs and is reported to promote the translation initiation [131]. To take this into account, we developed context score to compare sequence motif surrounding the putative ORF start with that surrounding the start codons from mRNAs. The calculation of context score is similar to that of the trimer/hexamer score above. We calculated the weight of each base at -6 to +1 positions relative to the start codon. Indeed, we observed the Kozak sequence motif in this position-specific weight matrix (Fig. 3.2). Hence, the higher the context score, the more similar to the Kozak sequence.

RNA secondary structure

We considered the RNA stem probability as a metric of RNA secondary structure, and then defined RNA structure features with respect to 5′/3′ UTRs and ORF. Both experimental and



	-6	-5	-4	-3	-2	-1	A	T	G	3
A	-0.36	-0.46	-0.08	0.87	0.25	-0.45				-0.24
C	-0.17	0.28	0.60	-1.41	0.52	0.80				-0.88
G	0.58	0.14	-0.12	0.47	-0.47	0.09				0.89
T	-0.28	-0.07	-0.83	-1.54	-0.69	-1.37				-0.48

Fig. 3.2 **Context scoring matrix measures the similarity of Kozak sequence (human).** We calculated the context scoring matrix from 5,000 CDSs (see “Method”). This indicates a Kozak sequence motif (gcc[ag]ccATGg) surrounding the start codon.

computational studies have observed that ORF sequences were more structured comparing with other regions in the mRNAs [118, 132], and a change of RNA secondary structure can be often observed surrounding the start and the stop codon . Thus, we calculated the RNA stem probability which indicates the likelihood that each base is included in a RNA stem structure across the full RNA sequence. Then we could extract averaged stem probabilities for distinct regions corresponding to pre-defined putative ORFs. Furthermore, we proposed that a stem probability ratio of 5′ UTR to ORF is needed to quantify the RNA structure changes between these two regions. Similarly, we also defined the ratio between 3′ UTR and ORF.

G4 is a four-stranded helical structure which can form in RNA and may be involve in translational control. Although the study of G4 is still in its infancy, it is inferred from its stable RNA secondary structure that G4 may block the translational regulation of the relevant site when it is close to the 5′ cap structure, the start codon, and the stop codon [133]. Additionally, G4 may also provide a cap-independent initial entry for translation initiation factors, thereby facilitating RNA translation [130, 133]. To explore whether G4 affects the association of lncRNAs with the ribosome, we first predicted the possible G4 structure in lncRNAs using QGRS [120], and then considered the relative positions of these

G4s relative to transcription initiation site (TIS), transcription termination site (TTS), and the start and end of the putative ORF (Fig. 3.1b). In addition, for the definition of relative position, we used two kinds of measurement methods: direct distance and relative distance. Direct distance represents the number of nucleotides on the RNA between the G4 and the target site mentioned above. Relative distance is a measure of the direct distance normalized to the total length of the RNA, to prevent possible bias of different RNA lengths.

RNA modification and repeat element

We utilized SRAMP [119] to predict where an m^6A might occur in a lncRNA, and calculated the direct and relative distances of the m^6A to various locations (i.e. TIS, TTS, and start/stop codons) as features. This is because previous studies have found that the m^6A is often enriched in a $5'$ UTR or in a $3'$ UTR neighboring stop codon [134, 135]. The m^6A that located in the $5'$ UTR can promote cap-independent translation [129], while the m^6A located around the stop codon may promote translation initiation by a binding protein. Finally, we were interested in whether the lncRNA contains a particular repeat element as a binarized feature. For example, Alu element is reported to be related to the cellular localization of lncRNAs [136], and our previous work have shown that the ribosomal association of lncRNAs,indeed, is positively correlated with the nuclear localization of lncRNAs. SINEB2, which is one of SINE (short interspersed nuclear element) repeat sequence, is reported to be associated with the up-regulated translation [137]. Hence, we do not rule out that SINE or other repeat elements may have the potential to regulate the ribosomal association of lncRNA.

Figure 3.3 shows the distribution of all features in ribo-lncRNA and noribo-lncRNA in human (see Fig. 3.4 for mouse; the meaning of the features are described in Table 3.2). According to the KS importance (described below) of each feature, we ranked all the features from high to low in the figure. Interestingly, if only one feature was chosen to distinguish the two types of lncRNAs, the GC content of the first exon (fEgc) was the most discriminating feature. We observed that ribo-lncRNAs tend to have a higher GC content in their first exons both in human and mouse. Here, all feature values were transformed in a range of 0 to 1. Then, we used two-sample Kolmogorov—Smirnov (KS) statistic [138] to examine the ability of each feature to separate the two types of lncRNAs (KS importance). The two-sample KS statistic is a non-parametric test to compare two groups of samples. When a feature has a significant difference between the two groups of lncRNAs, a smaller P value will be obtained in the two-sample KS statistic. If we only consider the effect of an individual feature, we can rank the features according to the statistical significance level (-log P value) from high to low. This method can be used for feature selection. Since it only independently assesses the importance of a single feature, it is also referred to as a filter method. This method is fast and straightforward and works well in many scenarios, but it cannot consider the combination of various features in the classification. For this purpose, we will carry out a more systematic screening of these extracted features as below.

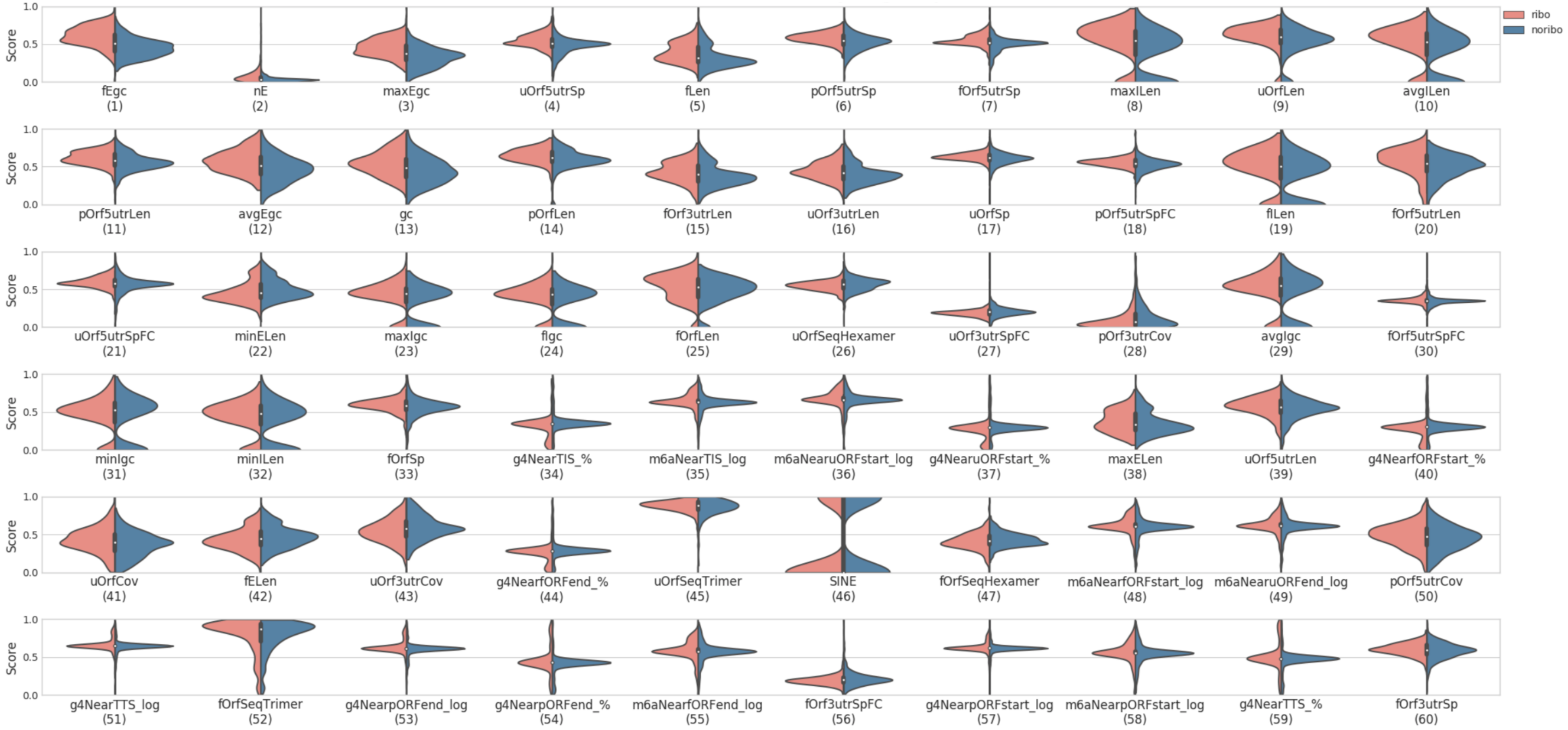


Fig. 3.3 **Distribution of all feature scores in human.** Each feature was ranked by -log(KS p-value), in which KS represents two samples Kolmogorov-Smirnov test between ribo-lncRNAs (red) and noribo- lncRNAs (blue)

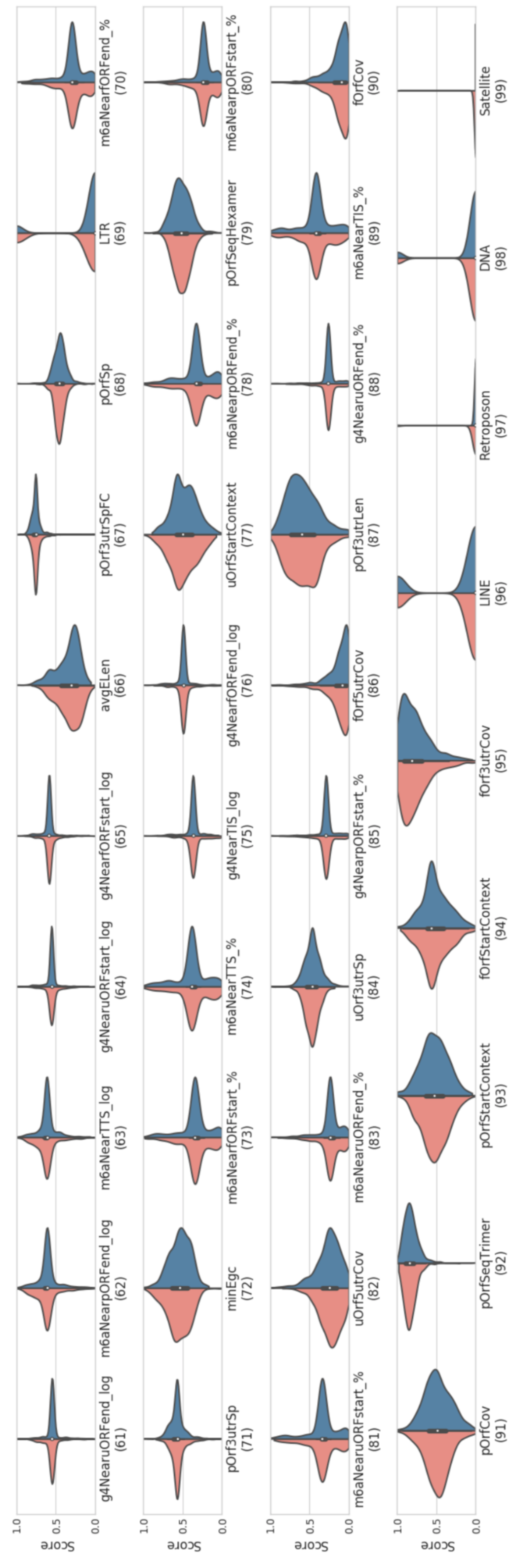


Fig. 3.3 Distribution of all feature scores in human (continued)

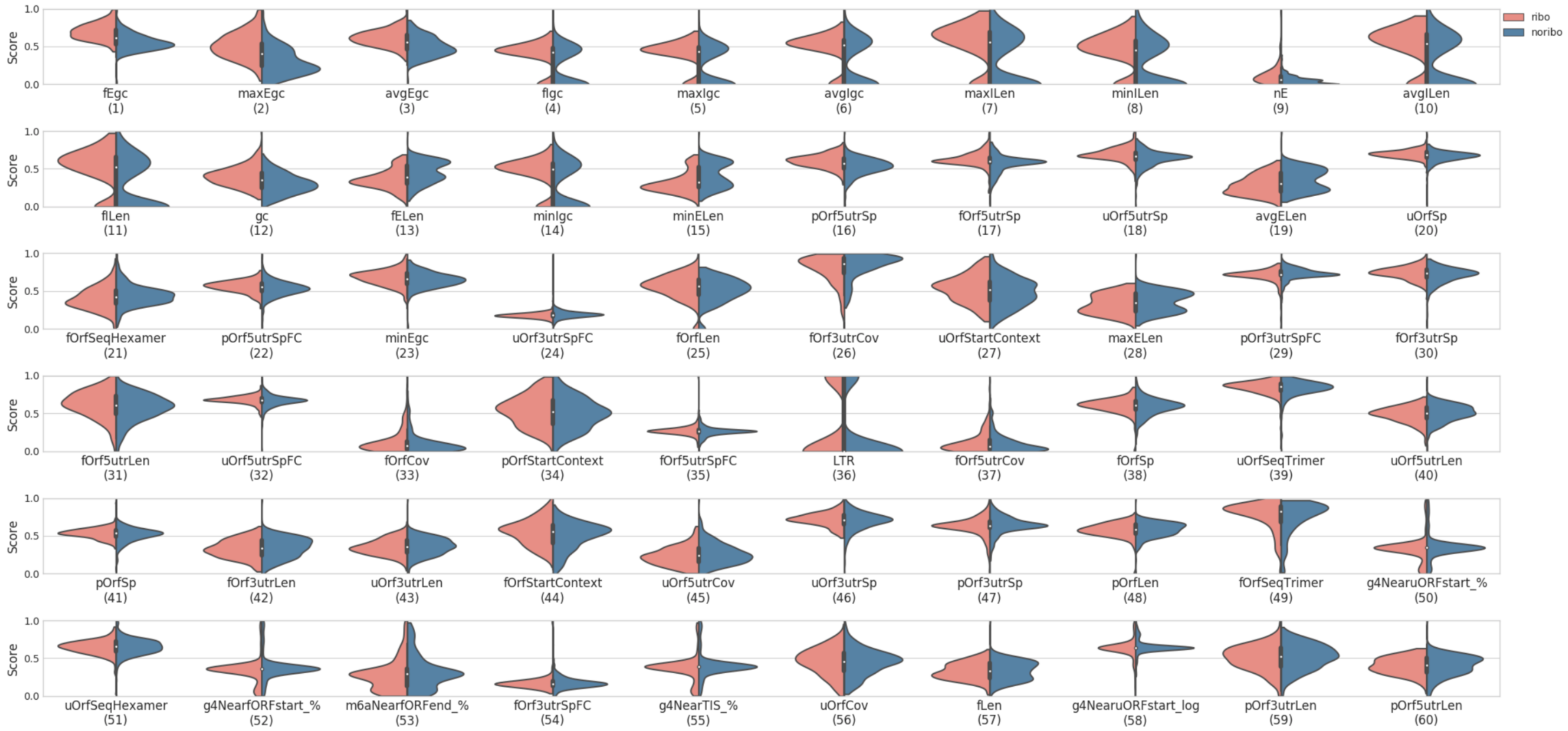


Fig. 3.4 **Distribution of all feature scores in human.** Each feature was ranked by -log(KS p-value), in which KS represents two samples Kolmogorov-Smirnov test between ribo-lncRNAs (red) and noribo- lncRNAs (blue)

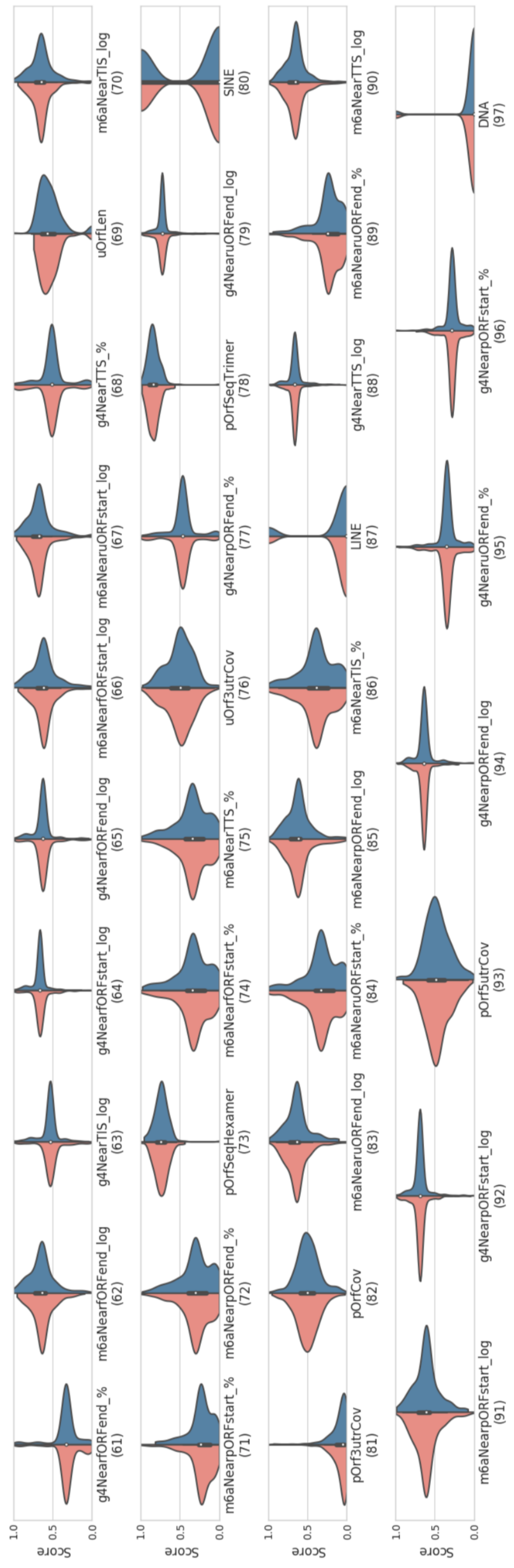


Fig. 3.4 Distribution of all feature scores in mouse (continued)

Removing high redundant features

One feature is considered to be redundant in the presence of another related feature with which is strongly correlated and can be removed without incurring much loss of information. To eliminate redundant features, we investigated the correlation coefficient between all features (Fig. 3.5a). The results show that the high redundant ($|r| > 0.8$) features are mainly clustered on exon/intron, G4, and m^{6}A in the form of length or distance. For example, in human, there is a high correlation between the lengths of a transcript and the longest exon in the transcript; the lengths of a pORF and the downstream $5'$ UTR, and the length of a $3'$ UTR of fORF and that of an uORF ($r > 0.8$, Table S10). The distance of m^{6}A relative to the transcript $5'$ end was highly correlated with its distance to the start of uORF ($r = 0.949$, Table S10). Similarly, there is a high correlation between the distance of G4 relative to the start of fORF and its distance to the start of uORF ($r = 0.928$, Table S10). We also observed similar results in mouse (Fig. 3.6a and Table S10).

After removing redundant features, we prepared low redundant features which were ready for a further feature selection. We removed one feature from each pair of redundant features to obtain the low redundant features (Table S10). Then, 59 and 55 sequence features were remained in the human and mouse, respectively. A list of low redundant features is given in Table 3.3. Figure 3.5b shows the correlation coefficient matrix between human low redundant features (see Fig. 3.6b for mouse). Although there are still some weak correlations between some features (e.g., the direct distance and the relative distance between m^{6}A and TIS), filtering of highly correlated features allows us to consider the importance of each feature more distinctly.

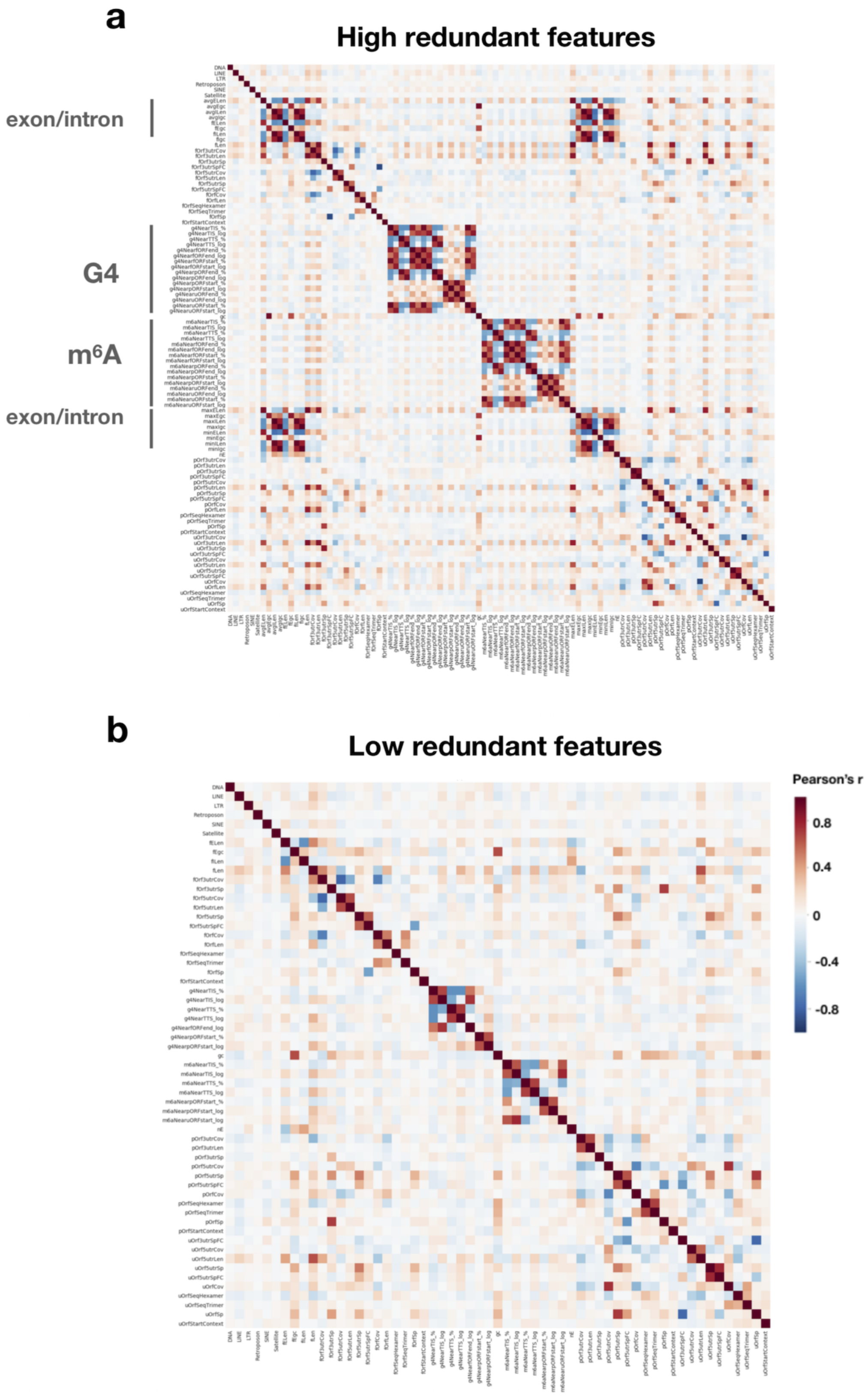


Fig. 3.5 **Correlations (r) of features indicate redundant features in human.** (**a**) Correlations of all extracted features show that features of several sub-regions are highly correlated (redundant). (**b**) After removing high redundant ($|r| > 0.8$) features, we obtained a low redundant feature set for further analysis in this study.

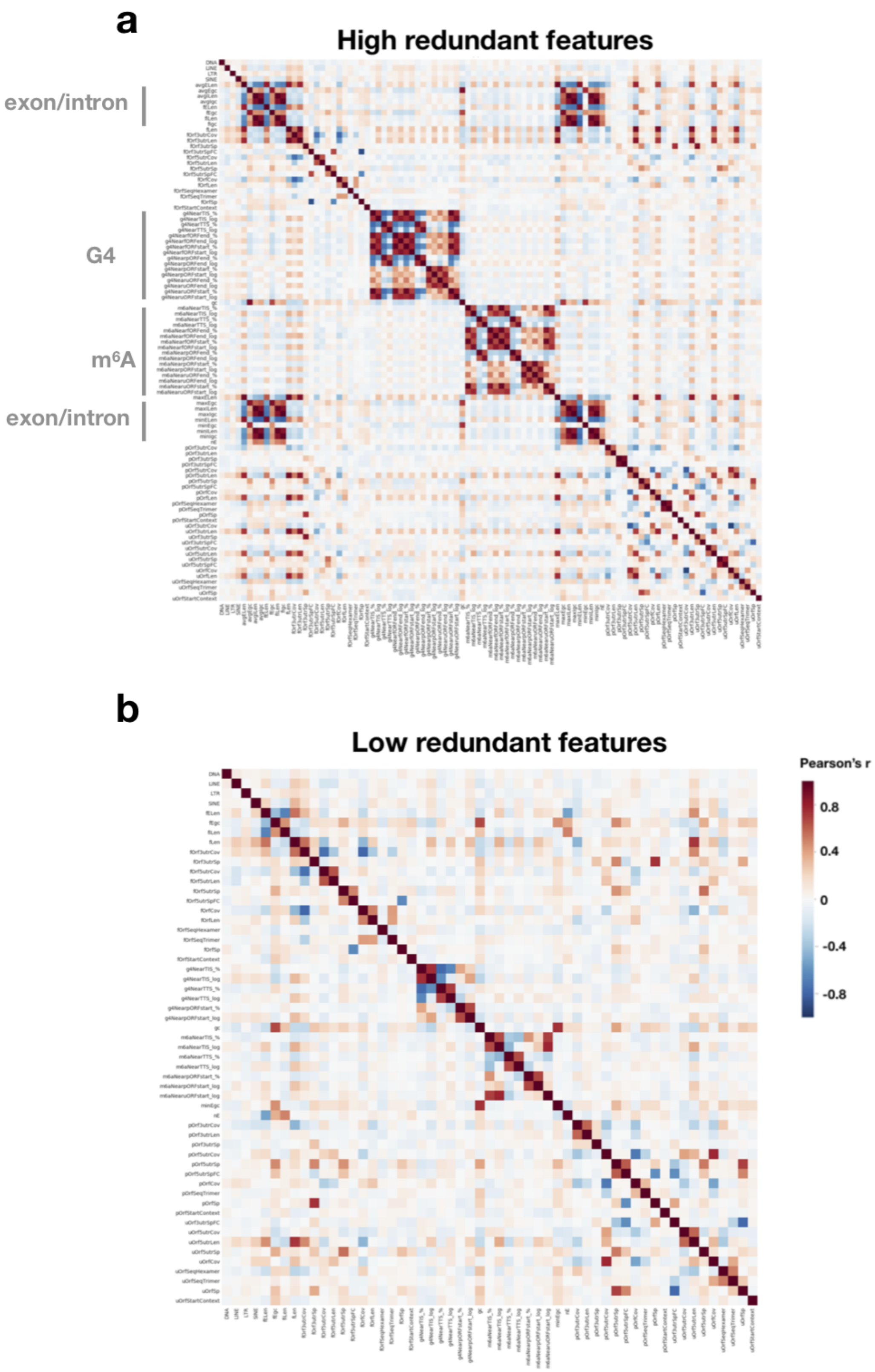
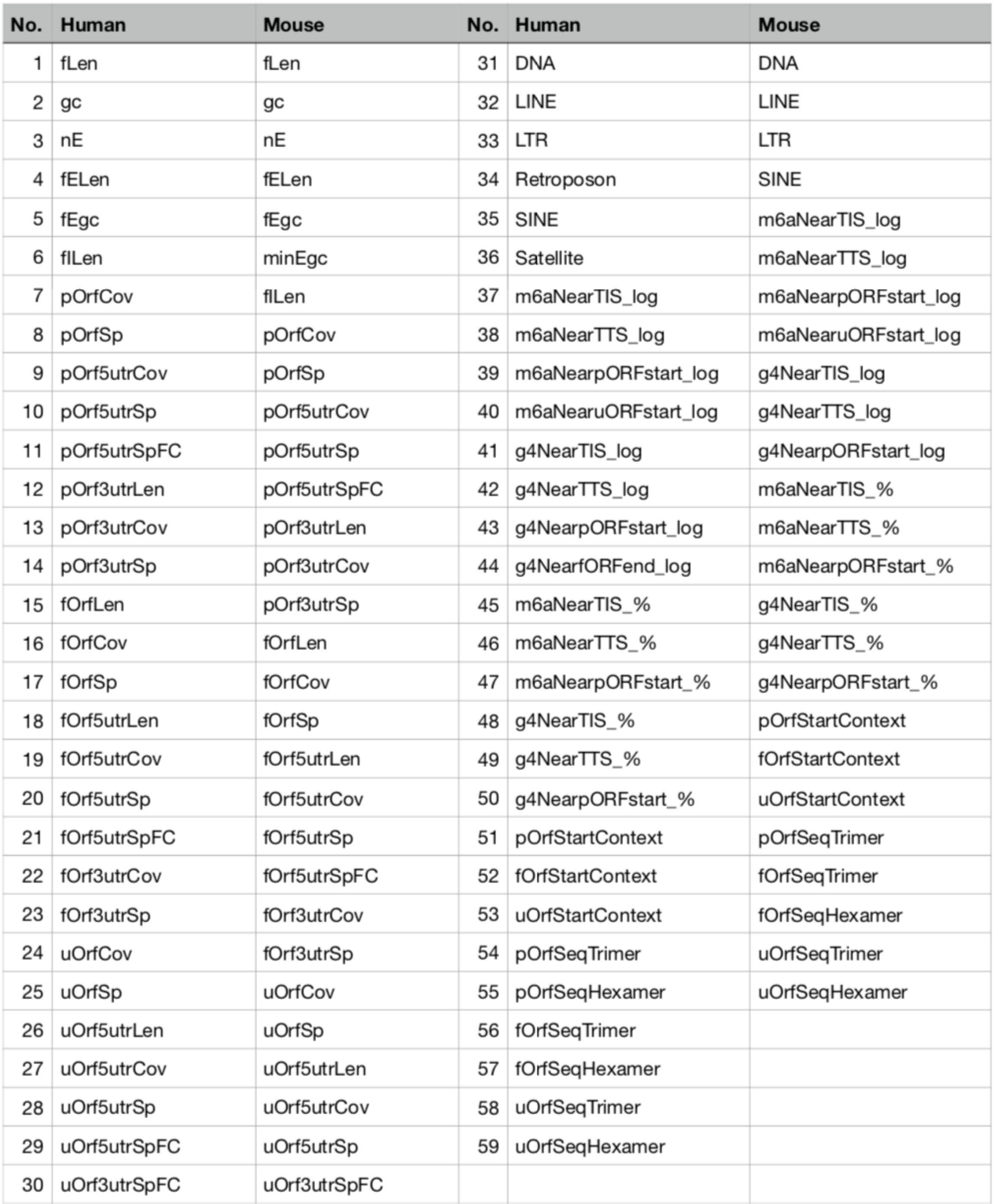


Fig. 3.6 **Correlations (r) of features indicate redundant features in mouse.** (**a**) Correlations of all extracted features show that features of several sub-regions are highly correlated (redundant). (**b**) After removing high redundant ($|r| > 0.8$) features, we obtained a low redundant feature set for further analysis in this study.

Table 3.3 Low-redundant features in human and mouse.



No.	Human	Mouse	No.	Human	Mouse
1	fLen	fLen	31	DNA	DNA
2	gc	gc	32	LINE	LINE
3	nE	nE	33	LTR	LTR
4	fELen	fELen	34	Retroposon	SINE
5	fEgc	fEgc	35	SINE	m6aNearTIS_log
6	fILen	minEgc	36	Satellite	m6aNearTTS_log
7	pOrfCov	fILen	37	m6aNearTIS_log	m6aNearpORFstart_log
8	pOrfSp	pOrfCov	38	m6aNearTTS_log	m6aNearuORFstart_log
9	pOrf5utrCov	pOrfSp	39	m6aNearpORFstart_log	g4NearTIS_log
10	pOrf5utrSp	pOrf5utrCov	40	m6aNearuORFstart_log	g4NearTTS_log
11	pOrf5utrSpFC	pOrf5utrSp	41	g4NearTIS_log	g4NearpORFstart_log
12	pOrf3utrLen	pOrf5utrSpFC	42	g4NearTTS_log	m6aNearTIS_%
13	pOrf3utrCov	pOrf3utrLen	43	g4NearpORFstart_log	m6aNearTTS_%
14	pOrf3utrSp	pOrf3utrCov	44	g4NearfORFend_log	m6aNearpORFstart_%
15	fOrfLen	pOrf3utrSp	45	m6aNearTIS_%	g4NearTIS_%
16	fOrfCov	fOrfLen	46	m6aNearTTS_%	g4NearTTS_%
17	fOrfSp	fOrfCov	47	m6aNearpORFstart_%	g4NearpORFstart_%
18	fOrf5utrLen	fOrfSp	48	g4NearTIS_%	pOrfStartContext
19	fOrf5utrCov	fOrf5utrLen	49	g4NearTTS_%	fOrfStartContext
20	fOrf5utrSp	fOrf5utrCov	50	g4NearpORFstart_%	uOrfStartContext
21	fOrf5utrSpFC	fOrf5utrSp	51	pOrfStartContext	pOrfSeqTrimer
22	fOrf3utrCov	fOrf5utrSpFC	52	fOrfStartContext	fOrfSeqTrimer
23	fOrf3utrSp	fOrf3utrCov	53	uOrfStartContext	fOrfSeqHexamer
24	uOrfCov	fOrf3utrSp	54	pOrfSeqTrimer	uOrfSeqTrimer
25	uOrfSp	uOrfCov	55	pOrfSeqHexamer	uOrfSeqHexamer
26	uOrf5utrLen	uOrfSp	56	fOrfSeqTrimer	
27	uOrf5utrCov	uOrf5utrLen	57	fOrfSeqHexamer	
28	uOrf5utrSp	uOrf5utrCov	58	uOrfSeqTrimer	
29	uOrf5utrSpFC	uOrf5utrSp	59	uOrfSeqHexamer	
30	uOrf3utrSpFC	uOrf3utrSpFC			

Feature selection by $\mathcal{L}1$-regularized logistic regression

Feature selection by using the $\mathcal{L}1$-regularized logistic model becomes a problem of how to select a hyperparameter C (see "Methods" in this chapter). As shown in Fig. 4, in a range of $[0.01, 1]$, we increased the value of C in steps of 0.001 and finally obtained the function between the C and the feature coefficients (colored solid lines), and the accuracy of prediction (blue dashed line). When the value of C is very small, the regularization strength is enormous and all of the feature coefficients are zeros, which means that no feature will be used as a predictor. At this time, the prediction accuracy implies that we predict all the results as positives (i.e., ribo-lncRNAs), which exactly reflects the proportion of positives in the test dataset. In human, for instance, the accuracy at this time is about 55%, which means that the number of positives and negatives in our test dataset is well-balanced. As the value of C increases, the more coefficients of the features turn to be non-zero, the prediction accuracy from the beginning of the rapid growth, to later stability or even a decrease. According to the criteria mentioned above, we choose $C = 0.257$ at the black vertical line in Fig.4, and the prediction accuracy at this time is 0.828. The features with non-zero coefficients corresponding to this are the critical features that we finally screen out. We can see that even if we continue to increase the value of C (to apply more features), this prediction accuracy has not improved considerably.

Taken together, we identified fifteen crucial sequence features of ribosomal association for human lncRNAs (nine for mouse lncRNAs). A list sorted by the importance of the crucial features is shown in the upper left corner of Fig. 3.7 (see Fig. 3.8 for mouse).

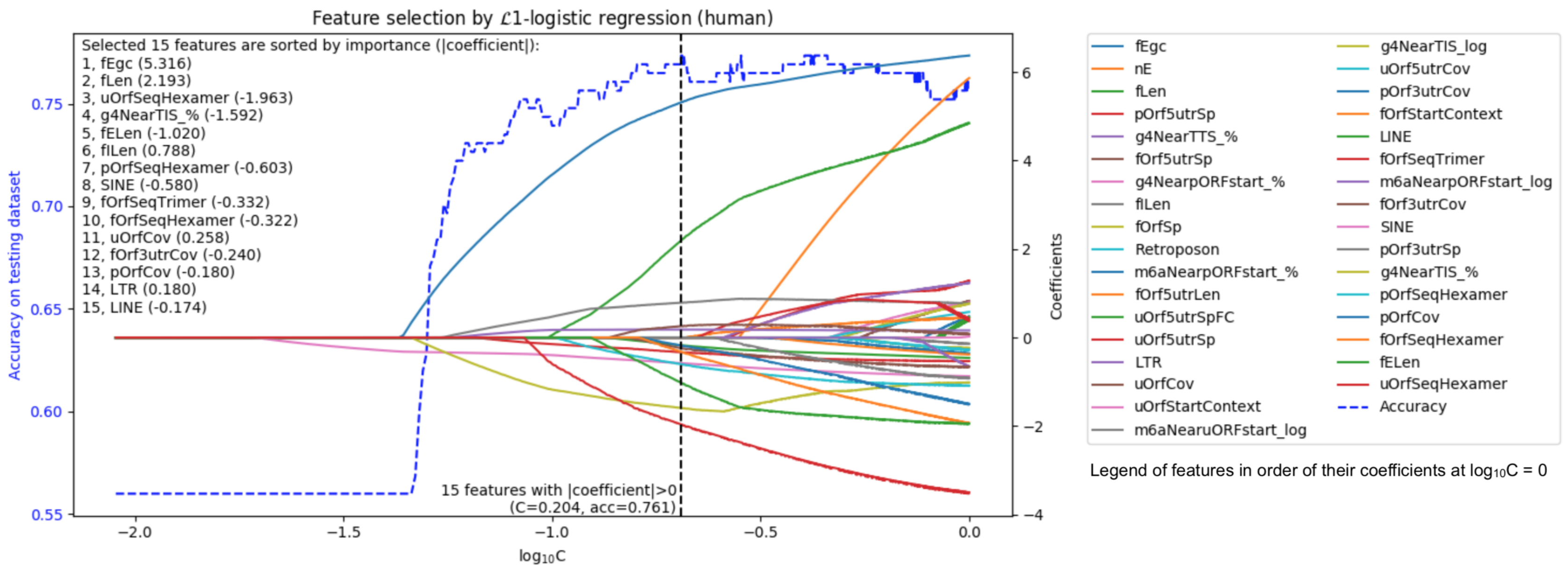


Fig. 3.7 **Feature selection by using $\mathcal{L}$1-logistic regression in human.** Total data was randomly separated into 80% for training the model and 20% for the calculation of accuracy (blue dashed line, left y-axis). On the x-axis, C indicates the inverse of regularization strength. As C is increased, the number of features with non-zero coefficients (right y-axis) is increased and the model becomes more complicated. The black dashed line shows the final model chosen in this study, and outputs 15 features with non-zero coefficients. These features were ranked by the absolute value of coefficient, which represents the importance for prediction, and shown in the upper left.

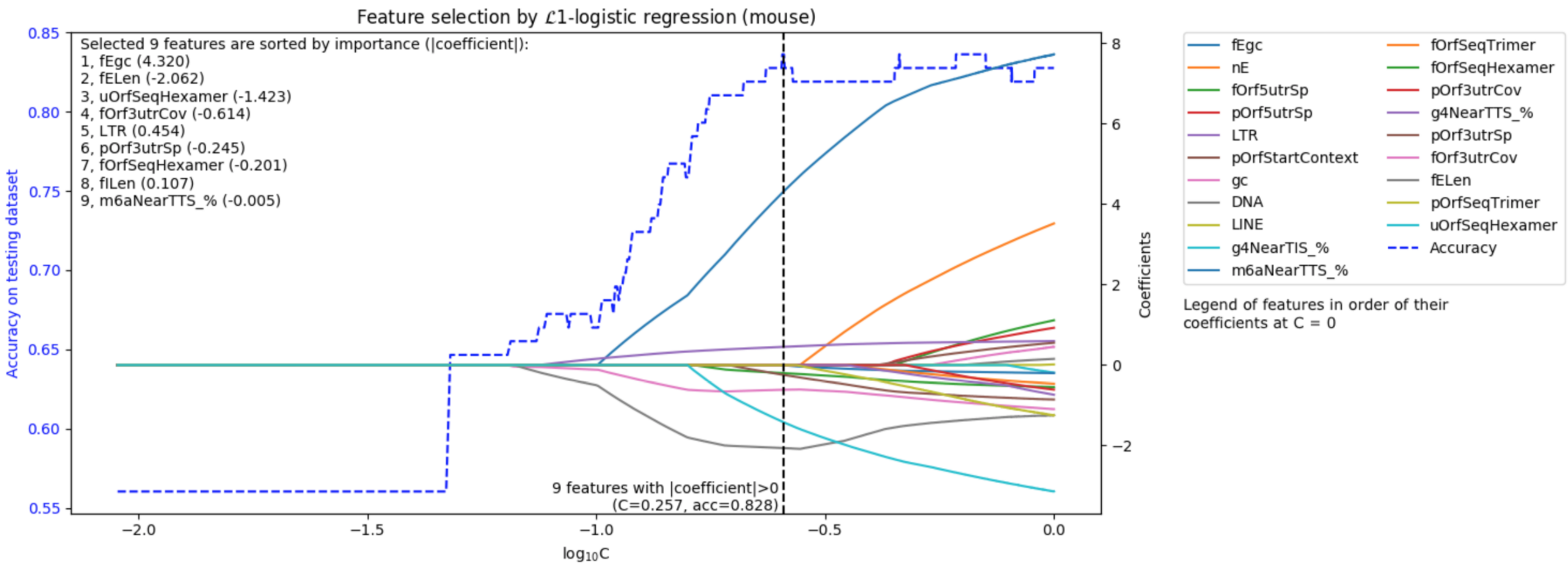


Fig. 3.8 **Feature selection by using $\mathcal{L}1$-logistic regression in mouse.** Total data was randomly separated into 80% for training the model and 20% for the calculation of accuracy (blue dashed line, left y-axis). On the x-axis, C indicates the inverse of regularization strength. As C is increased, the number of features with non-zero coefficients (right y-axis) is increased and the model becomes more complicated. The black dashed line shows the final model chosen in this study, and outputs 9 features with non-zero coefficients. These features were ranked by the absolute value of coefficient, which represents the importance for prediction, and shown in the upper left.

3.3 Discussion

By comparing the sequence features of the ribosomal association that we have identified in human and mouse lncRNAs, it is observed that seven features are conserved between the two species. It means that these common features may involve in the biological mechanisms of ribosomal association. Meanwhile, eight (human) and two (mouse) species-specific features are observed, which may involve species-specific regulatory mechanisms of the ribosomal association. In the following subsections, we discuss these features from the aspects of conserved and species-specific.

3.3.1 Conserved features

Conserved features include the fEgc, fELen, fILen, fOrfSeqHexamer, fOrf3utrCov, uOrfSeqHexamer, and LTR. Out of them, fEgc, fILen and LTR were positively correlated with the ribosomal association, while others vice versa. We observed that the G+C content and the length of the first exon had a high positive and negative correlation with the ribosomal association of lncRNA respectively. This finding matches with the results a study regarding the correlation between ribosome-associated mRNA and CDS [139]. High G+C content may indicate the occurrence of unexpected selection on ribosome-associated lncRNAs [140].

We could also observe that the longer the first intron, the more favorable lncRNAs are associated with the ribosome. The selection forces of intron-dependent nonsense-mediated RNA decay (NMD) on the first intron may be a reason for this situation [141]. This phenomenon is common among protein-coding genes, and a simple hypothesis is that longer introns are more likely to contain certain motifs [125], and these motifs may have essential factors that promote ribosomal association.

Surprisingly, the hexamer frequencies, which were used to assess the coding potential, of the first ORF and the first non-ATG ORF were inversely related to the ribosomal association. The reasons for this can be considered from two aspects: First, even if the ribosome has translation event on these two ORFs, the probability of detection of this event is low due to the length of the two ORFs is relatively shorter than that of the primary ORF. Moreover, the stronger the translation activity on these two ORFs will directly affect the ribosomal initiation of downstream pORFs, resulting in the failure of ribosome association on pORF to be detected. Second, we argue that the ribosomal association mentioned here not be the same as the ribosomal translation. The ribosome may use regulatory mechanisms other than the properties of the CDS sequence, to associate with particular RNAs (e.g., internal ribosomal entry site). Note that we did remove lncRNAs with translation potential when collecting ribosome-associated lncRNAs.

The results of human and mouse consistently demonstrated that lncRNAs containing a long terminal repeat (LTR), are more likely to associate with the ribosome. LTR is often used as a tool when viruses insert genetic material into a host genome. A well-known example of LTR is the human immunodeficiency virus (HIV), in which the LTR contains promoter, enhancer and other functional sequence elements [142]. Furthermore, our results indicate that LTR may be a functional element that promotes the ribosomal association or even translation.

3.3.2 Species-specific features

In human, the lncRNA length and the length of the non-ATG ORF are positively correlated with the ribosomal association. The remaining six features — the length and the hexamer frequency of the pORF, the trimer frequency of the fORF, the distance between G4 and TIS, and whether it contains LINE or SINE — have a negative correlation with the ribosomal association. In mouse, there are only two species-specific features — the RNA secondary structure of $3'$ UTR of pORF and the distance between m^6A and transcript $3'$ end — have a negative correlation with the ribosomal association.

Transcript length is one among the important features while distinguishing between protein-coding RNA and noncoding RNA [117]. As expected, this feature can also be used to distinguish ribo-lncRNA and noribo-lncRNA to some extent. The longer the transcript, the higher the probability that it may be associated with the ribosome (according to statistical point of view). Besides, the longer the sequence, the more likely it is to include functional motifs that promote ribosomal association. On the ORF, the features of the trimer/hexamer frequency and the length may be similar to those discussed above.

In contrast to LTR, SINE and LINE (long interspersed nuclear element) are more likely to appear in a ribosome-free lncRNA. This result is consistent with a report that Alu (a type of SINE) can drive the lncRNA in the nucleus [136]. We argue whether there is a set of complementary mechanisms controlling lncRNAs in the cytoplasm and nucleus by applying LTR and SINE/LINE. A systematic analysis of how these repeat elements affect the localization of lncRNAs can help us to understand the role of repeat elements in the evolution of genome, and the biological functions and mechanisms that lncRNAs may have involved.

G4 affects the ribosomal association when approaching transcript $5'$ end. This result is also discussed in many studies [130, 133]. Meanwhile, it further exhibits that the biological regulation of RNA in the secondary structure level. We observed that m^6A modification appears around transcript $3'$ end affecting the ribosomal association. Wang and colleagues mentioned that m^6A might form an RNA loop near the stop codon that brings the distance between the start and the stop codons closer to promote the translation efficiency [143]. However, the m^6A near TTS may hinder the formation of this mechanism. Finally, we compared mRNA with ribo-lncRNA and noribo-lncRNA (Fig. 3.9 and Fig. 3.10). It can be observed that in human, the length of the transcript can indeed be used to distinguish between lncRNA and mRNA. Additionally, we noticed that $5'/3'$ UTR of ribo-lncRNA seems to have a stronger RNA secondary structure compared with that of mRNA. In mouse, noribo-lncRNA has less number of exons compared with mRNA, which means the corresponding gene model is more straightforward.

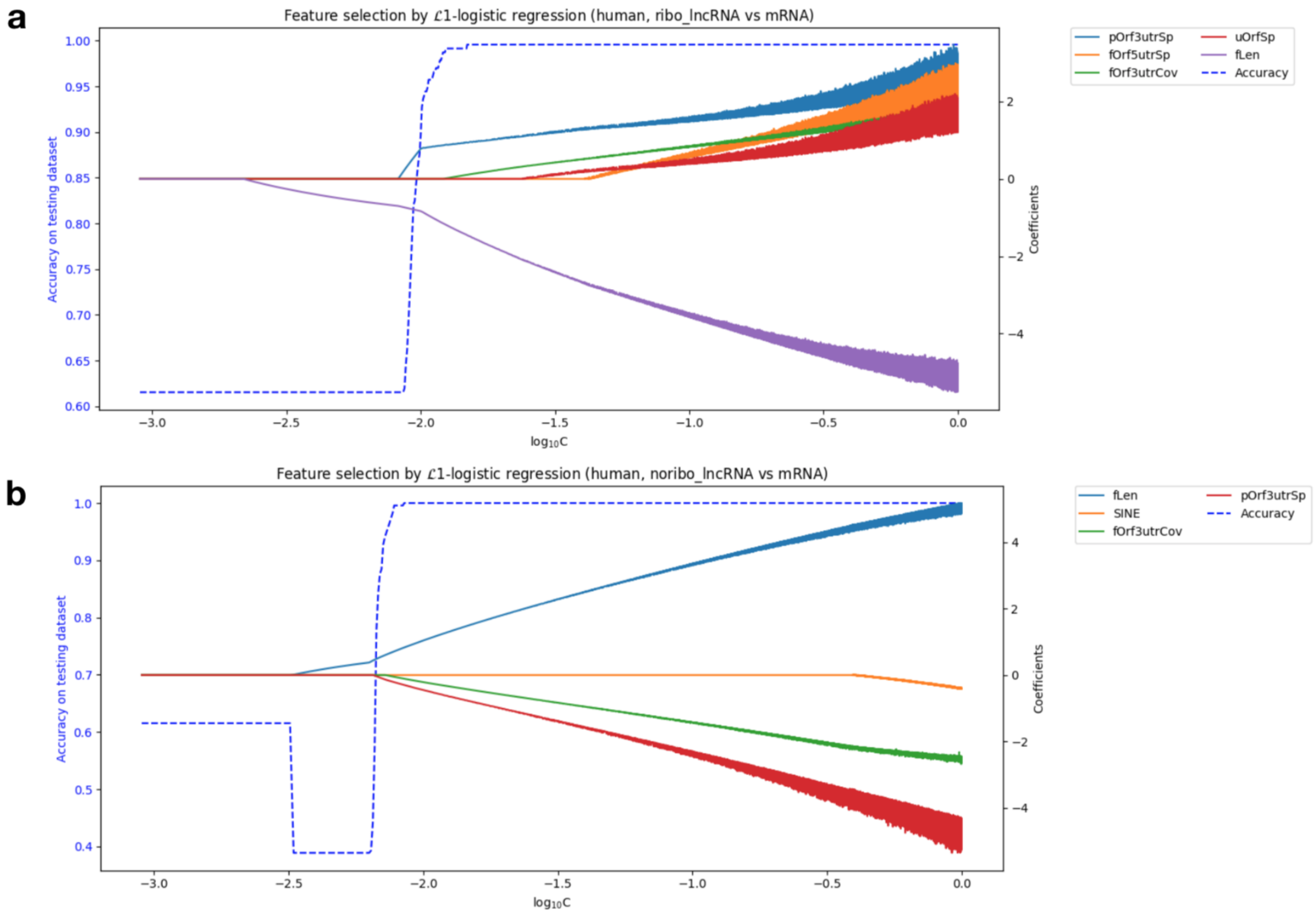


Fig. 3.9 Training $\mathcal{L}$1-logistic regression on the dataset of (**a**) ribo-lncRNAs and mRNAs; (**b**) noribo-lncRNAs and mRNAs in human.

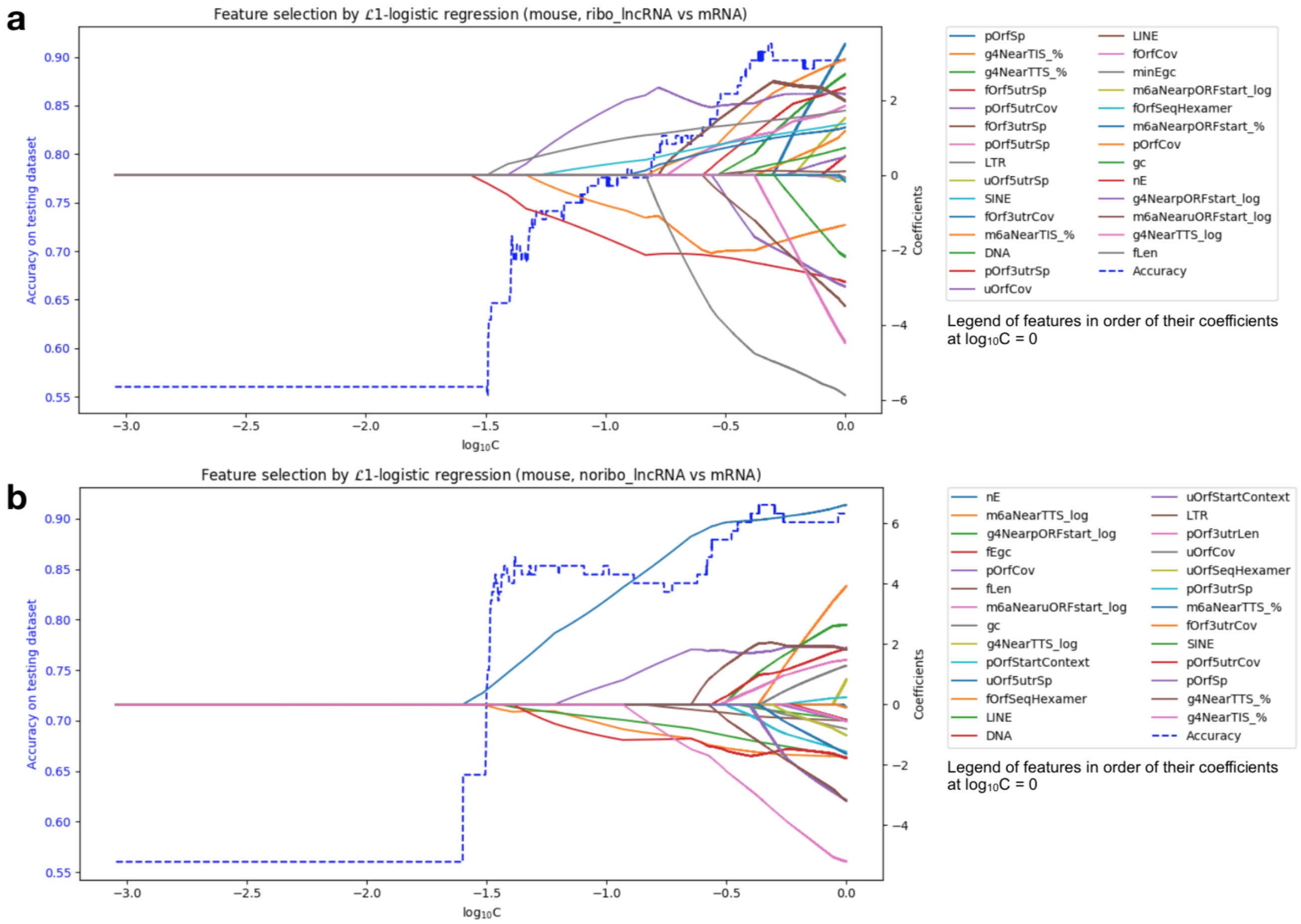


Fig. 3.10 Training $\mathcal{L}$1-logistic regression on the dataset of (**a**) ribo-lncRNAs and mRNAs; (**b**) noribo-lncRNAs and mRNAs in mouse.

3.4 Conclusions

This study analyzed the features of the ribosome-associated lncRNA at the level of sequence. Using the ribo-lncRNAs (ribosome-associated lncRNAs) and noribo-lncRNAs (ribosome-free lncRNAs) collected from human and mouse in our previous study [115], we analyzed which features are most important for distinguishing between the ribo-lncRNAs and the noribo-lncRNAs. Considering the reasons that a lncRNA may be involved in the ribosomal association, we mainly define sequence features based on distinct dimensions from several aspects such as RNA splicing, putative ORF, k-mer frequency, RNA secondary structure, RNA modification, and repeat element. Highly redundant features are removed by analyzing the correlation coefficient of each pair of features. Then, based on the $\mathcal{L}1$-regularized logistic regression model, we performed a feature selection while training feature parameters. Finally, we obtained fifteen and nine essential features for distinguishing between ribo-lncRNA and noribo-lncRNA from human and mouse, respectively, and discussed possible relationships between these features and the ribosomal association. To the best of our knowledge, this should be the first study of how to further divide ribo-lncRNA and noribo-lncRNA from the perspective of sequence features. This research describes how to extract sequence features to study lncRNAs and other biological phenotypes (e.g., subcellular localization), which provide research ideas for similar work. Moreover, the analysis of these sequence features has a critical reference value for us to understand further the ribosomal association, which is still an unknown mechanism, for lncRNA.

References

[1] S. Ohno, “So much" junk" dna in our genome. in" evolution of genetic systems",” in *Brookhaven Symposium in Biology*, vol. 23, pp. 366–370, 1972.

[2] J. R. Warner, R. Soeiro, H. 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*, vol. 19, no. 2, pp. 349–361, 1966.

[3] H. Busch, R. Reddy, L. Rothblum, and Y. Choi, “Snrnas, snrnps, and rna processing,” *Annual review of biochemistry*, vol. 51, no. 1, pp. 617–654, 1982.

[4] E. Maxwell and M. Fournier, “The small nucleolar rnas,” *Annual review of biochemistry*, vol. 64, no. 1, pp. 897–934, 1995.

[5] I. H. G. S. Consortium *et al.*, “Initial sequencing and analysis of the human genome,” *Nature*, vol. 409, no. 6822, p. 860, 2001.

[6] J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, *et al.*, “The sequence of the human genome,” *science*, vol. 291, no. 5507, pp. 1304–1351, 2001.

[7] P. Carninci, T. Kasukawa, S. Katayama, J. Gough, M. Frith, N. Maeda, R. Oyama, T. Ravasi, B. Lenhard, C. Wells, *et al.*, “The transcriptional landscape of the mammalian genome,” *Science*, vol. 309, no. 5740, pp. 1559–1563, 2005.

[8] M. K. Iyer, Y. S. Niknafs, R. Malik, U. Singhal, A. Sahu, Y. Hosono, T. R. Barrette, J. R. Prensner, J. R. Evans, S. Zhao, A. Poliakov, X. Cao, S. M. Dhanasekaran, Y.-M. Wu, D. R. Robinson, D. G. Beer, F. Y. Feng, H. K. Iyer, and A. M. Chinnaiyan, “The landscape of long noncoding RNAs in the human transcriptome,” *Nat. Genet.*, vol. 47, pp. 199–208, Mar. 2015.

[9] C.-C. Hon, J. A. Ramilowski, J. Harshbarger, N. Bertin, O. J. L. Rackham, J. Gough, E. Denisenko, S. Schmeier, T. M. Poulsen, J. Severin, M. Lizio, H. Kawaji, T. Kasukawa, M. Itoh, A. M. Burroughs, S. Noma, S. Djebali, T. Alam, Y. A. Medvedeva, A. C. Testa, L. Lipovich, C.-W. Yip, I. Abugessaisa, M. Mendez, A. Hasegawa, D. Tang, T. Lassmann, P. Heutink, M. Babina, C. A. Wells, S. Kojima, Y. Nakamura, H. Suzuki, C. O. Daub, M. J. L. de Hoon, E. Arner, Y. Hayashizaki, P. Carninci, and A. R. R. Forrest, “An atlas of human long non-coding RNAs with accurate 5’ ends,” *Nature*, vol. 543, pp. 199–204, Mar. 2017.

[10] K. Struhl, “Transcriptional noise and the fidelity of initiation by rna polymerase ii,” *Nature Structural and Molecular Biology*, vol. 14, no. 2, p. 103, 2007.

[11] J. Ponjavic, C. P. Ponting, and G. Lunter, “Functionality or transcriptional noise? evidence for selection within long noncoding rnas,” *Genome research*, vol. 17, no. 5, pp. 556–565, 2007.

[12] I. Ulitsky, A. Shkumatava, C. H. Jan, H. Sive, and D. P. Bartel, “Conserved function of lincrnas in vertebrate embryonic development despite rapid sequence evolution,” *Cell*, vol. 147, no. 7, pp. 1537–1550, 2011.

[13] S. U. Schmitz, P. Grote, and B. G. Herrmann, “Mechanisms of long noncoding rna function in development and disease,” *Cellular and molecular life sciences*, vol. 73, no. 13, pp. 2491–2509, 2016.

[14] T. R. Mercer, M. E. Dinger, and J. S. Mattick, “Long non-coding rnas: insights into functions,” *Nature Reviews Genetics*, vol. 10, no. 3, p. 155, 2009.

[15] K. C. Wang and H. Y. Chang, “Molecular mechanisms of long noncoding rnas,” *Molecular cell*, vol. 43, no. 6, pp. 904–914, 2011.

[16] O. Wapinski and H. Y. Chang, “Long noncoding rnas and human disease,” *Trends in cell biology*, vol. 21, no. 6, pp. 354–361, 2011.

[17] H. Hezroni, D. Koppstein, M. G. Schwartz, A. Avrutin, D. P. Bartel, and I. Ulitsky, “Principles of long noncoding rna evolution derived from direct comparison of transcriptomes in 17 species,” *Cell reports*, vol. 11, no. 7, pp. 1110–1122, 2015.

[18] 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 & development*, vol. 25, no. 18, pp. 1915–1927, 2011.

[19] B.-H. You, S.-H. Yoon, and J.-W. Nam, “High-confidence coding and noncoding transcriptome maps,” *Genome research*, vol. 27, no. 6, pp. 1050–1062, 2017.

[20] 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. Guigó, “The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression,” *Genome Res.*, vol. 22, pp. 1775–1789, Sept. 2012.

[21] N. A. O’Leary, M. W. Wright, J. R. Brister, S. Ciufo, D. Haddad, R. McVeigh, B. Rajput, B. Robbertse, B. Smith-White, D. Ako-Adjei, *et al.*, “Reference sequence (refseq) database at ncbi: current status, taxonomic expansion, and functional annotation,” *Nucleic acids research*, vol. 44, no. D1, pp. D733–D745, 2015.

[22] J. M. Stuart, E. Segal, D. Koller, and S. K. Kim, “A gene-coexpression network for global discovery of conserved genetic modules,” *science*, vol. 302, no. 5643, pp. 249–255, 2003.

[23] N. T. Ingolia, S. Ghaemmaghami, J. R. S. Newman, and J. S. Weissman, “Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling,” *Science*, vol. 324, pp. 218–223, Apr. 2009.

[24] S. L. Wolin and P. Walter, “Ribosome pausing and stacking during translation of a eukaryotic mrna.,” *The EMBO journal*, vol. 7, no. 11, pp. 3559–3569, 1988.

[25] 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*, vol. 147, no. 4, pp. 789–802, 2011.

[26] M. Siwiak and P. Zielenkiewicz, “A comprehensive, quantitative, and genome-wide model of translation,” *PLoS computational biology*, vol. 6, no. 7, p. e1000865, 2010.

[27] H. Guo, N. T. Ingolia, J. S. Weissman, and D. P. Bartel, “Mammalian micrornas predominantly act to decrease target mrna levels,” *Nature*, vol. 466, no. 7308, p. 835, 2010.

[28] G. A. Brar, M. Yassour, N. Friedman, A. Regev, N. T. Ingolia, and J. S. Weissman, “High-resolution view of the yeast meiotic program revealed by ribosome profiling,” *science*, vol. 335, no. 6068, pp. 552–557, 2012.

[29] A. M. Michel, K. R. Choudhury, A. E. Firth, N. T. Ingolia, J. F. Atkins, and P. V. Baranov, “Observation of dually decoded regions of the human genome using ribosome profiling data,” *Genome research*, vol. 22, no. 11, pp. 2219–2229, 2012.

[30] G.-W. Li, E. Oh, and J. S. Weissman, “The anti-shine–dalgarno sequence drives translational pausing and codon choice in bacteria,” *Nature*, vol. 484, no. 7395, p. 538, 2012.

[31] E. Oh, A. H. Becker, A. Sandikci, D. Huber, R. Chaba, F. Gloge, R. J. Nichols, A. Typas, C. A. Gross, G. Kramer, *et al.*, “Selective ribosome profiling reveals the cotranslational chaperone action of trigger factor in vivo,” *Cell*, vol. 147, no. 6, pp. 1295–1308, 2011.

[32] D. W. Reid and C. V. Nicchitta, “Primary role for endoplasmic reticulum-bound ribosomes in cellular translation identified by ribosome profiling,” *Journal of Biological Chemistry*, vol. 287, no. 8, pp. 5518–5527, 2012.

[33] S. Lee, B. Liu, S. Lee, S.-X. Huang, B. Shen, and S.-B. Qian, “Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution,” *Proceedings of the National Academy of Sciences*, vol. 109, no. 37, pp. E2424–E2432, 2012.

[34] N. R. Guydosh and R. Green, “Dom34 rescues ribosomes in 3’ untranslated regions,” *Cell*, vol. 156, no. 5, pp. 950–962, 2014.

[35] L. F. Lareau, D. H. Hite, G. J. Hogan, and P. O. Brown, “Distinct stages of the translation elongation cycle revealed by sequencing ribosome-protected mrna fragments,” *Elife*, vol. 3, 2014.

[36] J. Harrow, A. Frankish, J. M. Gonzalez, E. Tapanari, M. Diekhans, F. Kokocinski, B. L. Aken, D. Barrell, A. Zadissa, S. Searle, I. Barnes, A. Bignell, V. Boychenko, T. Hunt, M. Kay, G. Mukherjee, J. Rajan, G. Despacio-Reyes, G. Saunders, C. Steward, R. Harte, M. Lin, C. Howald, A. Tanzer, T. Derrien, J. Chrast, N. Walters, S. Balasubramanian, B. Pei, M. Tress, J. M. Rodriguez, I. Ezkurdia, J. van Baren, M. Brent, D. Haussler, M. Kellis, A. Valencia, A. Reymond, M. Gerstein, R. Guigó, and T. J. Hubbard, "GENCODE: the reference human genome annotation for the ENCODE project," *Genome Res.*, vol. 22, pp. 1760–1774, Sept. 2012.

[37] L. Ma, A. Li, D. Zou, X. Xu, L. Xia, J. Yu, V. B. Bajic, and Z. Zhang, "LncRNAWiki: harnessing community knowledge in collaborative curation of human long non-coding RNAs," *Nucleic Acids Research*, vol. 43, pp. D187–92, Jan. 2015.

[38] M. Baker, "Long noncoding RNAs: the search for function," *Nat. Methods*, vol. 8, pp. 379–383, Apr. 2011.

[39] J. Iwakiri, M. Hamada, and K. Asai, "Bioinformatics tools for lncRNA research," *Biochim. Biophys. Acta*, vol. 1859, pp. 23–30, Jan. 2016.

[40] P. Zhou, Y. Zhang, Q. Ma, F. Gu, D. S. Day, A. He, B. Zhou, J. Li, S. M. Stevens, D. Romo, and W. T. Pu, "Interrogating translational efficiency and lineage-specific transcriptomes using ribosome affinity purification," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 110, pp. 15395–15400, Sept. 2013.

[41] J. L. Aspden, Y. C. Eyre-Walker, R. J. Phillips, U. Amin, M. A. S. Mumtaz, M. Brocard, and J.-P. Couso, "Extensive translation of small open reading frames revealed by Poly-Ribo-Seq," *Elife*, vol. 3, p. e03528, Aug. 2014.

[42] M. Guttman, P. Russell, N. T. Ingolia, J. S. Weissman, and E. S. Lander, "Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins," *Cell*, vol. 154, pp. 240–251, July 2013.

[43] H. Wang, Y. Wang, S. Xie, Y. Liu, and Z. Xie, "Global and cell-type specific properties of lincRNAs with ribosome occupancy," *Nucleic Acids Res.*, vol. 45, pp. 2786–2796, Mar. 2017.

[44] A. Pircher, K. Bakowska-Zywicka, L. Schneider, M. Zywicki, and N. Polacek, "An mRNA-derived noncoding RNA targets and regulates the ribosome," *Mol. Cell*, vol. 54, pp. 147–155, Apr. 2014.

[45] J. Ruiz-Orera, X. Messeguer, J. A. Subirana, and M. M. Alba, "Long non-coding RNAs as a source of new peptides," *Elife*, vol. 3, p. e03523, Sept. 2014.

[46] B. Dallagiovanna, I. T. Pereira, A. C. Origa-Alves, P. Shigunov, H. Naya, and L. Spangenberg, "lncRNAs are associated with polysomes during adipose-derived stem cell differentiation," *Gene*, vol. 610, pp. 103–111, Apr. 2017.

[47] P. B. Essers, J. Nonnekens, Y. J. Goos, M. C. Betist, M. D. Viester, B. Mossink, N. Lansu, H. C. Korswagen, R. Jelier, A. B. Brenkman, and A. W. MacInnes, "A long noncoding RNA on the ribosome is required for lifespan extension," *Cell Rep.*, Jan. 2015.

[48] J. Carlevaro-Fita, A. Rahim, R. Guigó, L. A. Vardy, and R. Johnson, "Cytoplasmic long noncoding RNAs are frequently bound to and degraded at ribosomes in human cells," *RNA*, vol. 22, pp. 867–882, June 2016.

[49] R. Edgar, M. Domrachev, and A. E. Lash, "Gene expression omnibus: NCBI gene expression and hybridization array data repository," *Nucleic Acids Res.*, vol. 30, pp. 207–210, Jan. 2002.

[50] C. Gonzalez, J. S. Sims, N. Hornstein, A. Mela, F. Garcia, L. Lei, D. A. Gass, B. Amendolara, J. N. Bruce, P. Canoll, *et al.*, "Ribosome profiling reveals a cell-type-specific translational landscape in brain tumors," *Journal of Neuroscience*, vol. 34, no. 33, pp. 10924–10936, 2014.

[51] C. A. Rubio, B. Weisburd, M. Holderfield, C. Arias, E. Fang, J. L. DeRisi, and A. Fanidi, "Transcriptome-wide characterization of the eif4a signature highlights plasticity in translation regulation," *Genome biology*, vol. 15, no. 10, p. 476, 2014.

[52] M. E. Tanenbaum, N. Stern-Ginossar, J. S. Weissman, and R. D. Vale, "Regulation of mrna translation during mitosis," *Elife*, vol. 4, 2015.

[53] O. Tirosh, Y. Cohen, A. Shitrit, O. Shani, V. T. K. Le-Trilling, M. Trilling, G. Friedlander, M. Tanenbaum, and N. Stern-Ginossar, "The transcription and translation landscapes during human cytomegalovirus infection reveal novel host-pathogen interactions," *PLoS pathogens*, vol. 11, no. 11, p. e1005288, 2015.

[54] B. Xu, M. Gogol, K. Gaudenz, and J. L. Gerton, "Improved transcription and translation with l-leucine stimulation of mtorc1 in roberts syndrome," *BMC genomics*, vol. 17, no. 1, p. 25, 2016.

[55] S. W. Eichhorn, H. Guo, S. E. McGeary, R. A. Rodriguez-Mias, C. Shin, D. Baek, S.-h. Hsu, K. Ghoshal, J. Villén, and D. P. Bartel, "mrna destabilization is the dominant effect of mammalian micrornas by the time substantial repression ensues," *Molecular cell*, vol. 56, no. 1, pp. 104–115, 2014.

[56] S. Iwasaki, S. N. Floor, and N. T. Ingolia, "Rocaglates convert dead-box protein eif4a into a sequence-selective translational repressor," *Nature*, vol. 534, no. 7608, p. 558, 2016.

[57] C. Sidrauski, A. M. McGeachy, N. T. Ingolia, and P. Walter, "The small molecule isrib reverses the effects of eif2α phosphorylation on translation and stress granule assembly," *Elife*, vol. 4, 2015.

[58] A. O. Subtelny, S. W. Eichhorn, G. R. Chen, H. Sive, and D. P. Bartel, "Poly (a)-tail profiling reveals an embryonic switch in translational control," *Nature*, vol. 508, no. 7494, p. 66, 2014.

[59] J.-E. Park, H. Yi, Y. Kim, H. Chang, and V. N. Kim, "Regulation of poly (a) tail and translation during the somatic cell cycle," *Molecular cell*, vol. 62, no. 3, pp. 462–471, 2016.

[60] H. Zur, R. Aviner, and T. Tuller, “Complementary post transcriptional regulatory information is detected by punch-p and ribosome profiling,” *Scientific reports*, vol. 6, p. 21635, 2016.

[61] A. L. Wolfe, K. Singh, Y. Zhong, P. Drewe, V. K. Rajasekhar, V. R. Sanghvi, K. J. Mavrakis, M. Jiang, J. E. Roderick, J. Van der Meulen, *et al.*, “Rna g-quadruplexes cause eif4a-dependent oncogene translation in cancer,” *Nature*, vol. 513, no. 7516, p. 65, 2014.

[62] C. Cenik, E. S. Cenik, G. W. Byeon, F. Grubert, S. I. Candille, D. Spacek, B. Alsallakh, H. Tilgner, C. L. Araya, H. Tang, *et al.*, “Integrative analysis of rna, translation, and protein levels reveals distinct regulatory variation across humans,” *Genome research*, vol. 25, no. 11, pp. 1610–1621, 2015.

[63] X. Su, Y. Yu, Y. Zhong, E. G. Giannopoulou, X. Hu, H. Liu, J. R. Cross, G. Rätsch, C. M. Rice, and L. B. Ivashkiv, “Interferon-γ regulates cellular metabolism and mrna translation to potentiate macrophage activation,” *Nature immunology*, vol. 16, no. 8, p. 838, 2015.

[64] N. Wein, A. Vulin, M. S. Falzarano, C. A.-K. Szigyarto, B. Maiti, A. Findlay, K. N. Heller, M. Uhlén, B. Bakthavachalu, S. Messina, *et al.*, “Translation from a dmd exon 5 ires results in a functional dystrophin isoform that attenuates dystrophinopathy in humans and mice,” *Nature medicine*, vol. 20, no. 9, 2014.

[65] A. P. Wiita, E. Ziv, P. J. Wiita, A. Urisman, O. Julien, A. L. Burlingame, J. S. Weissman, and J. A. Wells, “Global cellular response to chemotherapy-induced apoptosis,” *Elife*, vol. 2, p. e01236, Oct. 2013.

[66] A. C. Hsieh, Y. Liu, M. P. Edlind, N. T. Ingolia, M. R. Janes, A. Sher, E. Y. Shi, C. R. Stumpf, C. Christensen, M. J. Bonham, *et al.*, “The translational landscape of mtor signalling steers cancer initiation and metastasis,” *Nature*, vol. 485, no. 7396, p. 55, 2012.

[67] J. U. Guo, V. Agarwal, H. Guo, and D. P. Bartel, “Expanded identification and characterization of mammalian circular rnas,” *Genome biology*, vol. 15, no. 7, p. 409, 2014.

[68] C. Jang, N. F. Lahens, J. B. Hogenesch, and A. Sehgal, “Ribosome profiling reveals an important role for translational control in circadian gene expression,” *Genome research*, vol. 25, no. 12, pp. 1836–1847, 2015.

[69] A. Werner, S. Iwasaki, C. A. McGourty, S. Medina-Ruiz, N. Teerikorpi, I. Fedrigo, N. T. Ingolia, and M. Rape, “Cell-fate determination by ubiquitin-dependent regulation of translation,” *Nature*, vol. 525, no. 7570, p. 523, 2015.

[70] C. C. Thoreen, L. Chantranupong, H. R. Keys, T. Wang, N. S. Gray, and D. M. Sabatini, “A unifying model for mtorc1-mediated regulation of mrna translation,” *Nature*, vol. 485, no. 7396, p. 109, 2012.

[71] J. Cho, N.-K. Yu, J.-H. Choi, S.-E. Sim, S. J. Kang, C. Kwak, S.-W. Lee, J.-i. Kim, D. I. Choi, V. N. Kim, *et al.*, “Multiple repressive mechanisms in the hippocampus during memory formation,” *Science*, vol. 350, no. 6256, pp. 82–87, 2015.

[72] J. R. Alvarez-Dominguez, X. Zhang, and W. Hu, “Widespread and dynamic translational control of red blood cell development,” *Blood*, vol. 129, no. 5, pp. 619–629, 2017.

[73] N. Fradejas-Villar, S. Seeher, C. B. Anderson, M. Doengi, B. A. Carlson, D. L. Hatfield, U. Schweizer, and M. T. Howard, “The rna-binding protein secisbp2 differentially modulates uga codon reassignment and rna decay,” *Nucleic acids research*, vol. 45, no. 7, pp. 4094–4107, 2016.

[74] F. Atger, C. Gobet, J. Marquis, E. Martin, J. Wang, B. Weger, G. Lefebvre, P. Descombes, F. Naef, and F. Gachon, “Circadian and feeding rhythms differentially affect rhythmic mrna transcription and translation in mouse liver,” *Proceedings of the National Academy of Sciences*, vol. 112, no. 47, pp. E6579–E6588, 2015.

[75] M. T. Howard, B. A. Carlson, C. B. Anderson, and D. L. Hatfield, “Translational redefinition of uga codons is regulated by selenium availability,” *Journal of Biological Chemistry*, vol. 288, no. 27, pp. 19401–19413, 2013.

[76] P. Janich, A. B. Arpat, V. Castelo-Szekely, M. Lopes, and D. Gatfield, “Ribosome profiling reveals the rhythmic liver translatome and circadian clock regulation by upstream open reading frames,” *Genome research*, vol. 25, no. 12, pp. 1848–1859, 2015.

[77] S. Blanco, R. Bandiera, M. Popis, S. Hussain, P. Lombard, J. Aleksic, A. Sajini, H. Tanna, R. Cortés-Garrido, N. Gkatza, *et al.*, “Stem cell function and stress response are controlled by protein synthesis,” *Nature*, vol. 534, no. 7607, p. 335, 2016.

[78] A. Sendoel, J. G. Dunn, E. H. Rodriguez, S. Naik, N. C. Gomez, B. Hurwitz, J. Levorse, B. D. Dill, D. Schramek, H. Molina, *et al.*, “Translation from unconventional 5’ start sites drives tumour initiation,” *Nature*, vol. 541, no. 7638, p. 494, 2017.

[79] J. Castañeda, P. Genzor, G. W. van der Heijden, A. Sarkeshik, J. R. Yates, N. T. Ingolia, and A. Bortvin, “Reduced pachytene pirnas and translation underlie spermiogenic arrest in maelstrom mutant mice,” *The EMBO journal*, p. e201386855, 2014.

[80] J. A. Hurt, A. D. Robertson, and C. B. Burge, “Global analyses of upf1 binding and function reveal expanded scope of nonsense-mediated mrna decay,” *Genome research*, vol. 23, no. 10, pp. 1636–1650, 2013.

[81] D. W. Reid, Q. Chen, A. S.-L. Tay, S. Shenolikar, and C. V. Nicchitta, “The unfolded protein response triggers selective mrna release from the endoplasmic reticulum,” *Cell*, vol. 158, no. 6, pp. 1362–1374, 2014.

[82] J. M. Mudge and J. Harrow, “Creating reference gene annotation for the mouse C57BL6/J genome assembly,” *Mamm. Genome*, vol. 26, pp. 366–378, Oct. 2015.

[83] P. P. Chan and T. M. Lowe, “GtRNAdb: a database of transfer RNA genes detected in genomic sequence,” *Nucleic Acids Res.*, vol. 37, pp. D93–7, Jan. 2009.

[84] NCBI Resource Coordinators, “Database resources of the national center for biotechnology information,” *Nucleic Acids Res.*, vol. 45, pp. D12–D17, Jan. 2017.

[85] D. Karolchik, "The UCSC Table Browser data retrieval tool," *Nucleic Acids Research*, vol. 32, no. 90001, pp. 493D–496, 2004.

[86] B. L. Aken, S. Ayling, D. Barrell, L. Clarke, V. Curwen, S. Fairley, J. Fernandez Banet, K. Billis, C. García Girón, T. Hourlier, K. Howe, A. Kähäri, F. Kokocinski, F. J. Martin, D. N. Murphy, R. Nag, M. Ruffier, M. Schuster, Y. A. Tang, J.-H. Vogel, S. White, A. Zadissa, P. Flicek, and S. M. J. Searle, "The ensembl gene annotation system," *Database*, vol. 2016, June 2016.

[87] B. Langmead and S. L. Salzberg, "Fast gapped-read alignment with bowtie 2," *Nat. Methods*, vol. 9, pp. 357–359, Mar. 2012.

[88] M. Martin, "Cutadapt removes adapter sequences from high-throughput sequencing reads," *EMBnet.journal*, vol. 17, pp. 10–12, May 2011.

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

[90] D. Hebenstreit, M. Fang, M. Gu, V. Charoensawan, A. van Oudenaarden, and S. A. Teichmann, "RNA sequencing reveals two major classes of gene expression levels in metazoan cells," *Mol. Syst. Biol.*, vol. 7, p. 497, June 2011.

[91] G. P. Wagner, K. Kin, and V. J. Lynch, "A model based criterion for gene expression calls using RNA-seq data," *Theory Biosci.*, vol. 132, pp. 159–164, Sept. 2013.

[92] N. Kryuchkova-Mostacci and M. Robinson-Rechavi, "A benchmark of gene expression tissue-specificity metrics," *Brief. Bioinform.*, vol. 18, pp. 205–214, Mar. 2017.

[93] N. R. Guydosh and R. Green, "Dom34 rescues ribosomes in 3' untranslated regions," *Cell*, vol. 156, pp. 950–962, Feb. 2014.

[94] N. Savage, "Proteomics: High-protein research," *Nature*, vol. 527, pp. S6–7, Nov. 2015.

[95] N. T. Ingolia, G. A. Brar, N. Stern-Ginossar, M. S. Harris, G. J. S. Talhouarne, S. E. Jackson, M. R. Wills, and J. S. Weissman, "Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes," *Cell Rep.*, vol. 8, pp. 1365–1379, Sept. 2014.

[96] A. A. Bazzini, T. G. Johnstone, R. Christiano, S. D. Mackowiak, B. Obermayer, E. S. Fleming, C. E. Vejnar, M. T. Lee, N. Rajewsky, T. C. Walther, and A. J. Giraldez, "Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation," *EMBO J.*, vol. 33, pp. 981–993, May 2014.

[97] V. Olexiouk, W. Van Criekinge, and G. Menschaert, "An update on sORFs.org: a repository of small ORFs identified by ribosome profiling," *Nucleic Acids Res.*, Nov. 2017.

[98] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," *Nucleic Acids Res.*, vol. 25, pp. 3389–3402, Sept. 1997.

[99] M. Colombo, E. D. Karousis, J. Bourquin, R. Bruggmann, and O. Mühlemann, "Transcriptome-wide identification of NMD-targeted human mRNAs reveals extensive redundancy between SMG6- and SMG7-mediated degradation pathways," *RNA*, vol. 23, pp. 189–201, Feb. 2017.

[100] 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. Röder, 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. Guigó, and T. R. Gingeras, "Landscape of transcription in human cells," *Nature*, vol. 489, pp. 101–108, Sept. 2012.

[101] B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg, "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome," *Genome Biol.*, vol. 10, p. R25, Mar. 2009.

[102] N. Leng, J. A. Dawson, J. A. Thomson, V. Ruotti, A. I. Rissman, B. M. G. Smits, J. D. Haag, M. N. Gould, R. M. Stewart, and C. Kendziorski, "EBSeq: an empirical bayes hierarchical model for inference in RNA-seq experiments," *Bioinformatics*, vol. 29, pp. 1035–1043, Apr. 2013.

[103] Z. Ji, R. Song, H. Huang, A. Regev, and K. Struhl, "Transcriptome-scale RNase-footprinting of RNA-protein complexes," *Nat. Biotechnol.*, vol. 34, pp. 410–413, Apr. 2016.

[104] J. A. Makarova and D. A. Kramerov, "Noncoding RNA of U87 host gene is associated with ribosomes and is relatively resistant to nonsense-mediated decay," *Gene*, vol. 363, pp. 51–60, Dec. 2005.

[105] S. Lykke-Andersen, Y. Chen, B. R. Ardal, B. Lilje, J. Waage, A. Sandelin, and T. H. Jensen, "Human nonsense-mediated RNA decay initiates widely by endonucleolysis and targets snoRNA host genes," *Genes Dev.*, vol. 28, pp. 2498–2517, Nov. 2014.

[106] H. Tani, M. Torimura, and N. Akimitsu, "The RNA degradation pathway regulates the function of GAS5 a non-coding RNA in mammalian cells," *PLoS One*, vol. 8, p. e55684, Jan. 2013.

[107] H. Hansji, E. Y. Leung, B. C. Baguley, G. J. Finlay, D. Cameron-Smith, V. C. Figueiredo, and M. E. Askarian-Amiri, "ZFAS1: a long noncoding RNA associated with ribosomes in breast cancer cells," *Biol. Direct*, vol. 11, p. 62, Nov. 2016.

[108] C. M. Clemson, J. N. Hutchinson, S. A. Sara, A. W. Ensminger, A. H. Fox, A. Chess, and J. B. Lawrence, "An architectural role for a nuclear noncoding RNA: NEAT1

RNA is essential for the structure of paraspeckles," *Mol. Cell*, vol. 33, pp. 717–726, Mar. 2009.

[109] 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," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 106, pp. 11667–11672, July 2009.

[110] D. K. Gascoigne, S. W. Cheetham, P. B. Cattenoz, M. B. Clark, P. P. Amaral, R. J. Taft, D. Wilhelm, M. E. Dinger, and J. S. Mattick, "Pinstripe: a suite of programs for integrating transcriptomic and proteomic datasets identifies novel proteins and improves differentiation of protein-coding and non-coding genes," *Bioinformatics*, vol. 28, pp. 3042–3050, Dec. 2012.

[111] B. Uszczynska-Ratajczak, J. Lagarde, A. Frankish, R. Guigó, and R. Johnson, "Towards a complete map of the human long non-coding RNA transcriptome," *resource*, vol. 8, no. 67, p. 276, 2018.

[112] A. Pircher, J. Gebetsberger, and N. Polacek, "Ribosome-associated ncRNAs: An emerging class of translation regulators," *RNA biology*, vol. 11, no. 11, pp. 1335–1339, 2014.

[113] J. Bazin, K. Baerenfaller, S. J. Gosai, B. D. Gregory, M. Crespi, and J. Bailey-Serres, "Global analysis of ribosome-associated noncoding RNAs unveils new modes of translational regulation," *Proceedings of the National Academy of Sciences*, vol. 114, no. 46, pp. E10018–E10027, 2017.

[114] F. Yeasmin, T. Yada, and N. Akimitsu, "Micropeptides encoded in transcripts previously identified as long noncoding RNAs: A new chapter in transcriptomics and proteomics," *Frontiers in Genetics*, vol. 9, p. 144, 2018.

[115] C. Zeng, T. Fukunaga, and M. Hamada, "Identification and analysis of ribosome-associated lncRNAs using ribosome profiling data," *BMC Genomics*, vol. 19, no. 414, 2018.

[116] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool.," *Journal of molecular biology*, vol. 215, no. 3, pp. 403–10, 1990.

[117] L. Wang, H. J. Park, S. Dasari, S. Wang, J. P. Kocher, and W. Li, "CPAT: Coding-potential assessment tool using an alignment-free logistic regression model," *Nucleic Acids Research*, vol. 41, no. 6, 2013.

[118] R. Kawaguchi and H. Kiryu, "Parallel computation of genome-scale RNA secondary structure to detect structural constraints on human genome," *BMC Bioinformatics*, vol. 17, no. 1, pp. 1–20, 2016.

[119] Y. Zhou, P. Zeng, Y. H. Li, Z. Zhang, and Q. Cui, "SRAMP: Prediction of mammalian N^6-methyladenosine (m^{6}A) sites based on sequence-derived features," *Nucleic Acids Research*, vol. 44, no. 10, 2016.

[120] C. Menendez, S. Frees, and P. S. Bagga, “QGRS-H Predictor: A web server for predicting homologous quadruplex forming G-rich sequence motifs in nucleotide sequences,” *Nucleic Acids Research*, vol. 40, no. W1, pp. 96–103, 2012.

[121] A. Smit, R. Hubley, and P. Green, “2013–2015. repeatmasker open-4.0,” 2013. accessed on 1 May 2018.

[122] D. R. Cox, “The regression analysis of binary sequences,” *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 215–242, 1958.

[123] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, *et al.*, “Scikit-learn: Machine learning in python,” *Journal of machine learning research*, vol. 12, no. Oct, pp. 2825–2830, 2011.

[124] J. Majewski, J. Majewski, J. Ott, and J. Ott, “Distribution and characterization of regulatory elements in the human genome,” *Genome Research*, vol. 12, no. 212, pp. 1827–1836, 2002.

[125] K. R. Bradnam and I. Korf, “Longer first introns are a general property of eukaryotic gene structure,” *PLoS ONE*, vol. 3, no. 8, 2008.

[126] N. I. Bieberstein, F. C. Oesterreich, K. Straube, and K. M. Neugebauer, “First exon length controls active chromatin signatures and transcription,” *Cell reports*, vol. 2, no. 1, pp. 62–68, 2012.

[127] M. Kozak, “The scanning model for translation: an update.,” *The Journal of cell biology*, vol. 108, no. 2, pp. 229–241, 1989.

[128] R. L. Tanguay and D. R. Gallie, “Translational efficiency is regulated by the length of the 3’ untranslated region,” *Molecular and cellular biology*, vol. 16, no. 1, pp. 146–156, 1996.

[129] K. D. Meyer, D. P. Patil, J. Zhou, A. Zinoviev, M. A. Skabkin, O. Elemento, T. V. Pestova, S. B. Qian, and S. R. Jaffrey, “5’ UTR m^{6}A Promotes Cap-Independent Translation,” *Cell*, vol. 163, no. 4, pp. 999–1010, 2015.

[130] A. Bugaut and S. Balasubramanian, “5’ UTR RNA G-quadruplexes: translation regulation and targeting,” *Nucleic acids research*, vol. 40, no. 11, pp. 4727–4741, 2012.

[131] M. Kozak, “An analysis of 5’-noncoding sequences from 699 vertebrate messenger RNAs,” *Nucleic acids research*, vol. 15, no. 20, pp. 8125–8148, 1987.

[132] M. Kertesz, Y. Wan, E. Mazor, J. L. Rinn, R. C. Nutter, H. Y. Chang, and E. Segal, “Genome-wide measurement of rna secondary structure in yeast,” *Nature*, vol. 467, no. 7311, p. 103, 2010.

[133] J. Song, J.-P. Perreault, I. Topisirovic, and S. Richard, “RNA G-quadruplexes and their potential regulatory roles in translation,” *Translation*, vol. 4, no. 2, p. e1244031, 2016.

[134] D. Dominissini, S. Moshitch-Moshkovitz, S. Schwartz, M. Salmon-Divon, L. Ungar, S. Osenberg, K. Cesarkas, J. Jacob-Hirsch, N. Amariglio, M. Kupiec, R. Sorek, and G. Rechavi, “Topology of the human and mouse m^{6}A RNA methylomes revealed by m^{6}A-seq,” *Nature*, vol. 485, no. 7397, pp. 201–206, 2012.

[135] K. D. Meyer, Y. Saletore, P. Zumbo, O. Elemento, C. E. Mason, and S. R. Jaffrey, “Comprehensive analysis of mRNA methylation reveals enrichment in 3’ UTRs and near stop codons,” *Cell*, vol. 149, no. 7, pp. 1635–1646, 2012.

[136] Y. Lubelsky and I. Ulitsky, “Sequences enriched in Alu repeats drive nuclear localization of long RNAs in human cells,” *Nature*, vol. 555, no. 7694, pp. 107–111, 2018.

[137] C. Carrieri, L. Cimatti, M. Biagioli, A. Beugnet, S. Zucchelli, S. Fedele, E. Pesce, I. Ferrer, L. Collavin, C. Santoro, *et al.*, “Long non-coding antisense RNA controls Uchl1 translation through an embedded SINEB2 repeat,” *Nature*, vol. 491, no. 7424, p. 454, 2012.

[138] W. J. Dixon, “Power under normality of several nonparametric tests,” *The Annals of Mathematical Statistics*, pp. 610–614, 1954.

[139] D. Zhao, J. P. Hamilton, M. Hardigan, D. Yin, T. He, B. Vaillancourt, M. Reynoso, G. Pauluzzi, S. Funkhouser, Y. Cui, J. Bailey-Serres, J. Jiang, C. R. Buell, and N. Jiang, “Analysis of ribosome-associated mRNAs in rice reveals the importance of transcript size and GC content in translation,” *G3: Genes | Genomes | Genetics*, vol. 7, no. 1, pp. 203–219, 2017.

[140] W. Haerty and C. P. Ponting, “Unexpected selection to retain high GC content and splicing enhancers within exons of multiexonic lncRNA loci,” *RNA*, vol. 21, no. 3, pp. 320–332, 2015.

[141] M. Lynch and A. Kewalramani, “Messenger RNA surveillance and the evolutionary proliferation of introns,” *Molecular biology and evolution*, vol. 20, no. 4, pp. 563–571, 2003.

[142] F. C. Krebs, T. H. Hogan, S. Quiterio, S. Gartner, and B. Wigdahl, “Lentiviral LTR-directed expression, sequence variation, and disease pathogenesis,” *HIV sequence compendium*, pp. 29–70, 2001.

[143] X. Wang, B. S. Zhao, I. A. Roundtree, Z. Lu, D. Han, H. Ma, X. Weng, K. Chen, H. Shi, and C. He, “N^6-methyladenosine modulates messenger rna translation efficiency,” *Cell*, vol. 161, no. 6, pp. 1388–1399, 2015.

Appendix A

Supplementary figures and tables

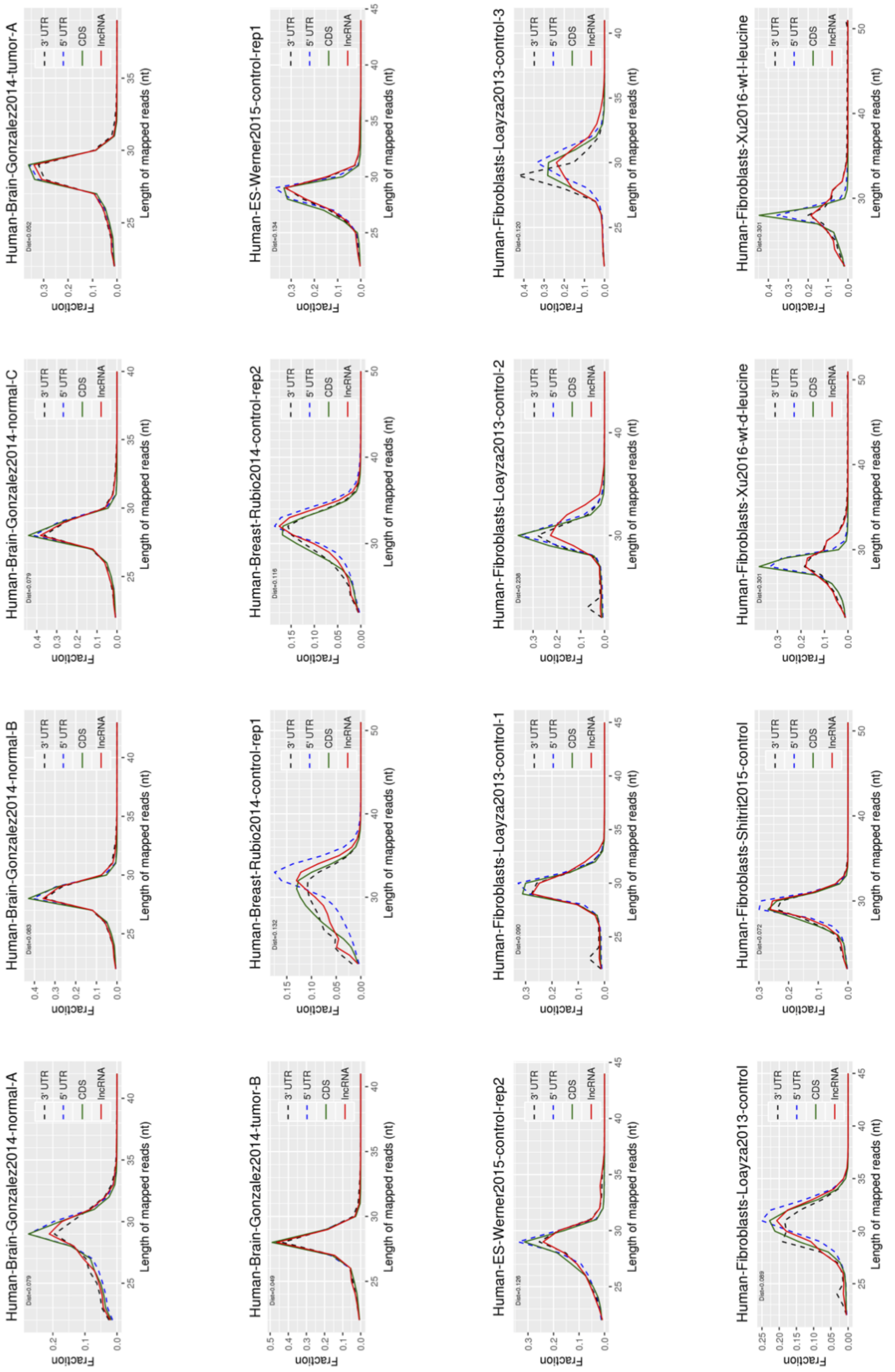


Fig. S1 Frequency distributions of Ribo-seq read lengths across CDSs, 5′/3′UTRs, and lncRNAs (human)

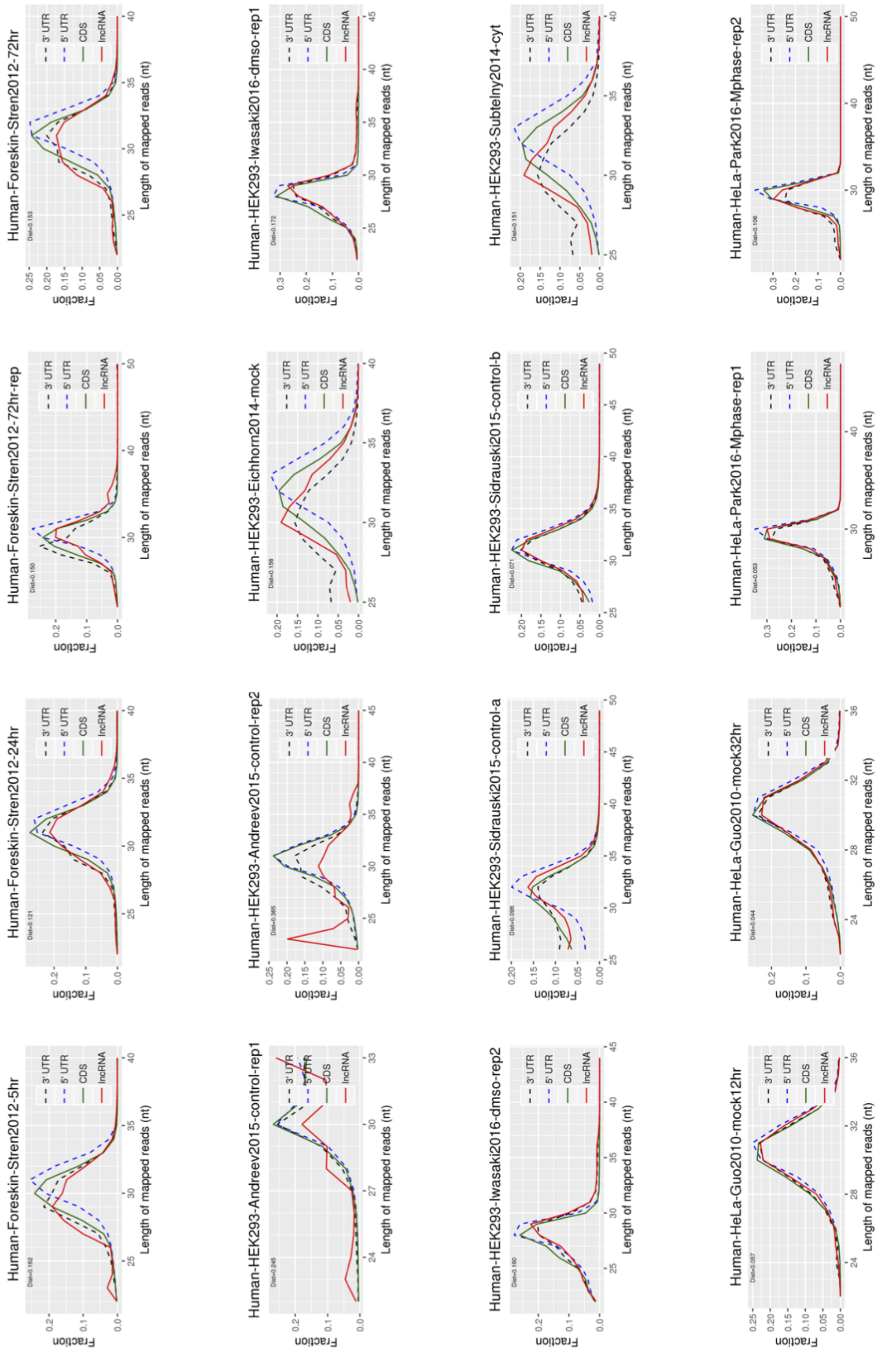


Fig. S1 (continued)

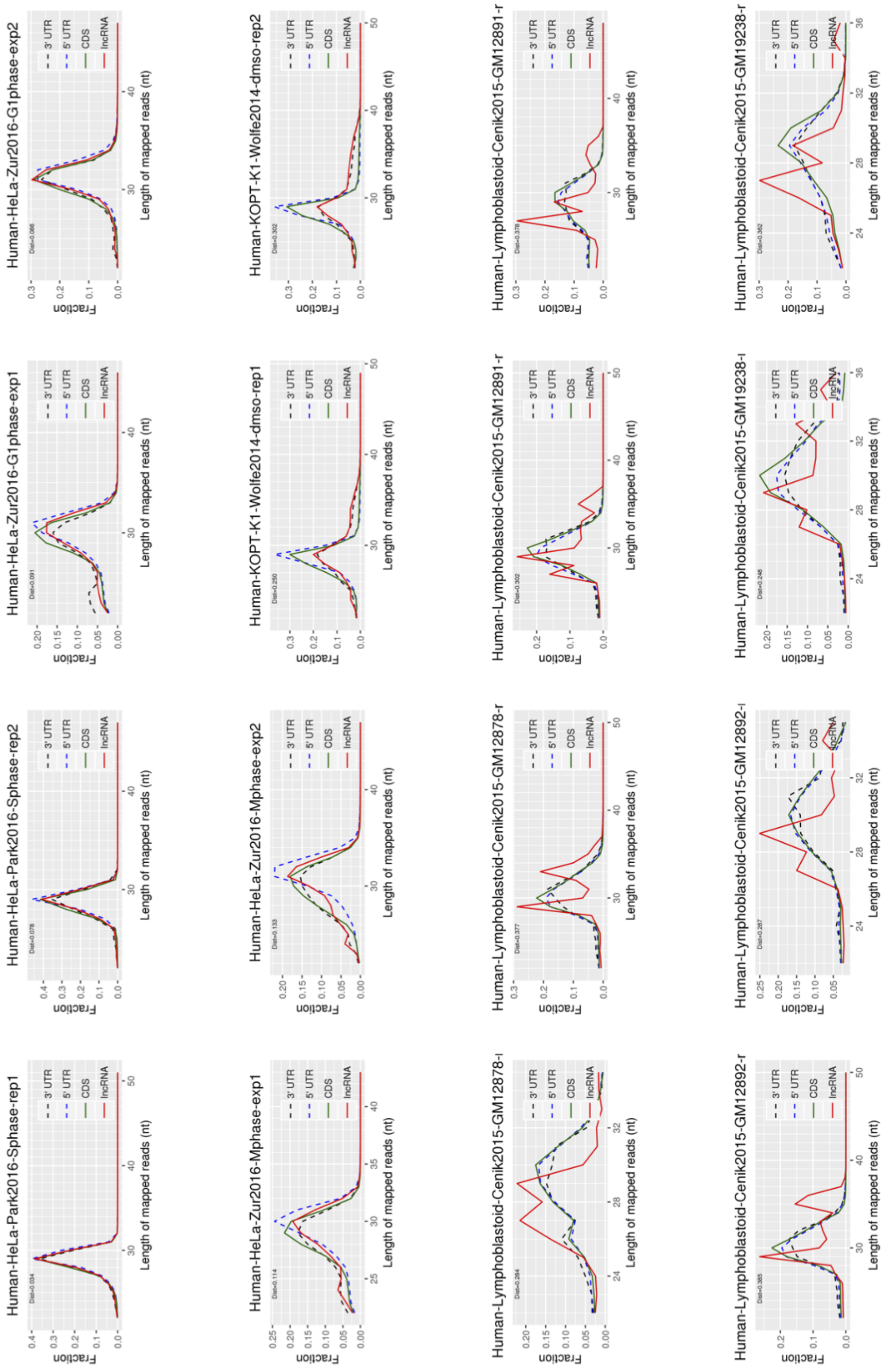


Fig. S1 (continued)

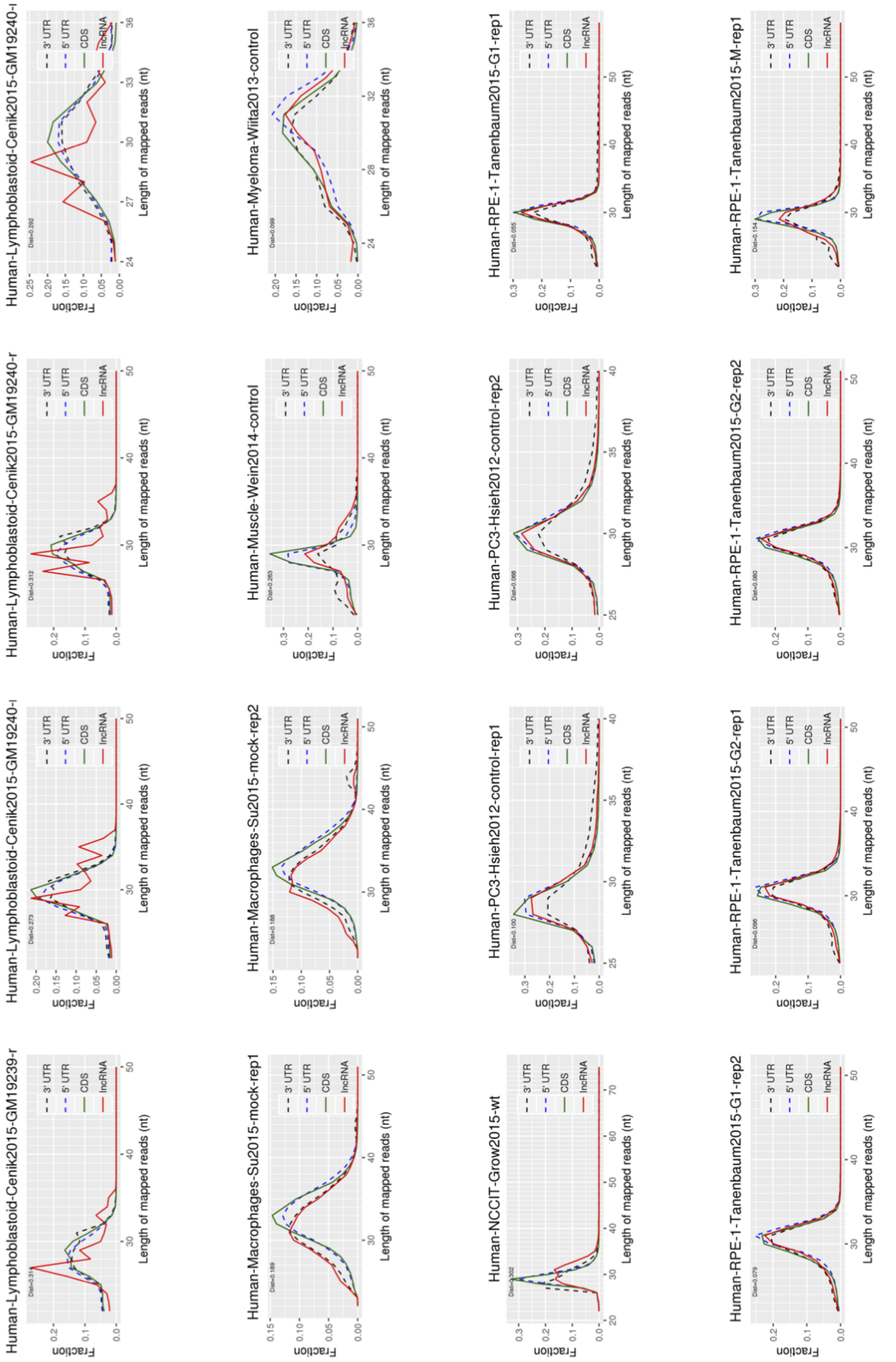


Fig. S1 (continued)

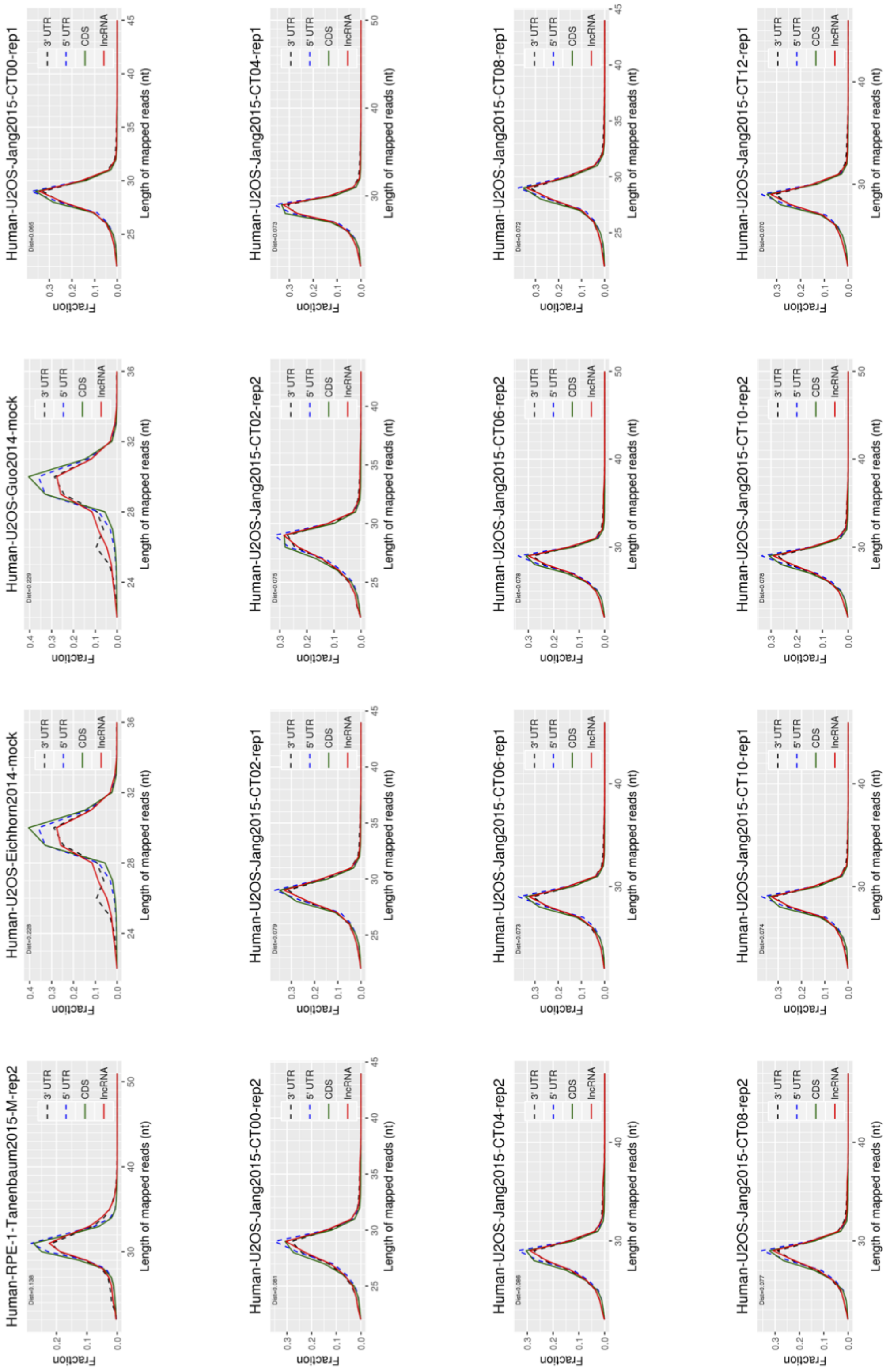


Fig. S1 (continued)

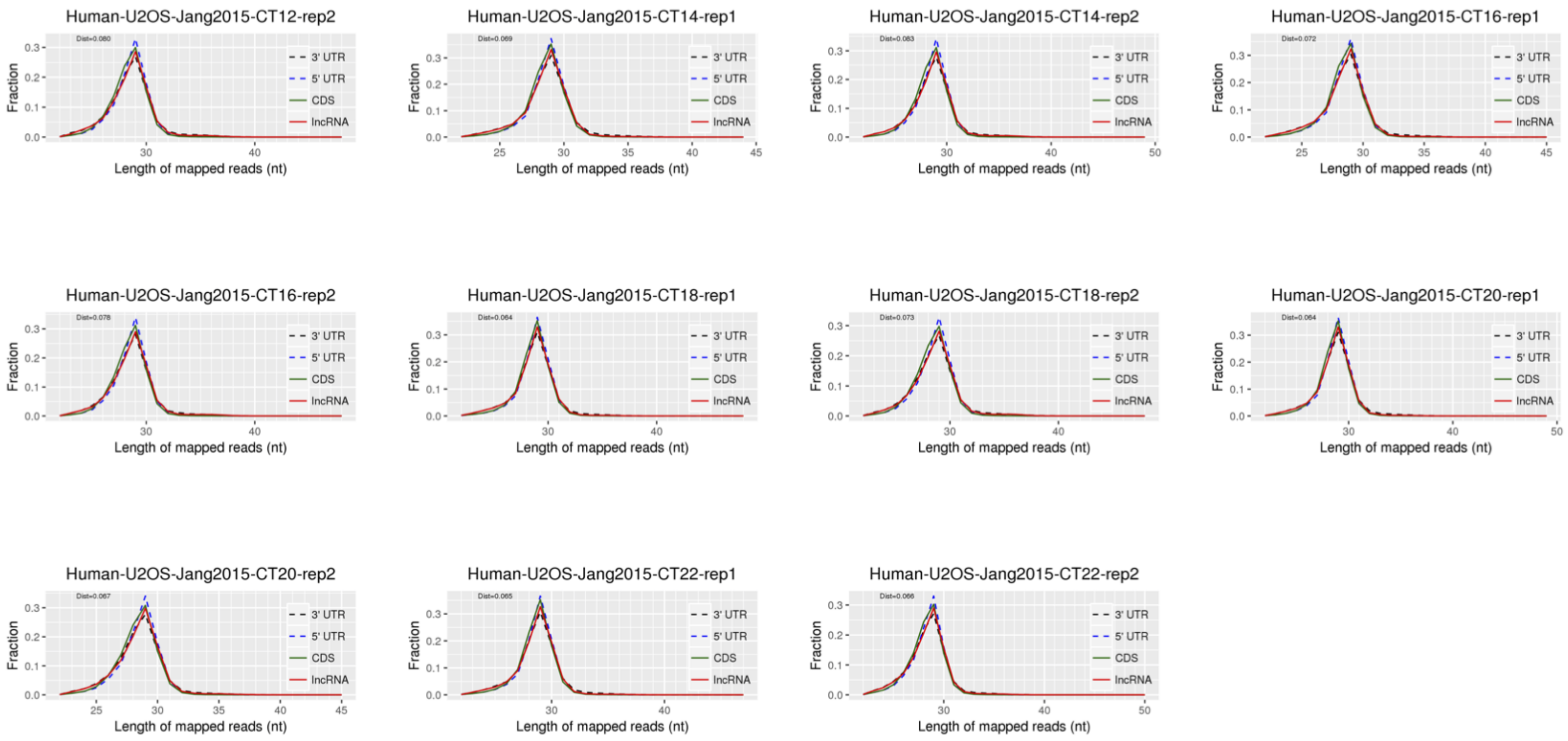


Fig. S1 (continued)

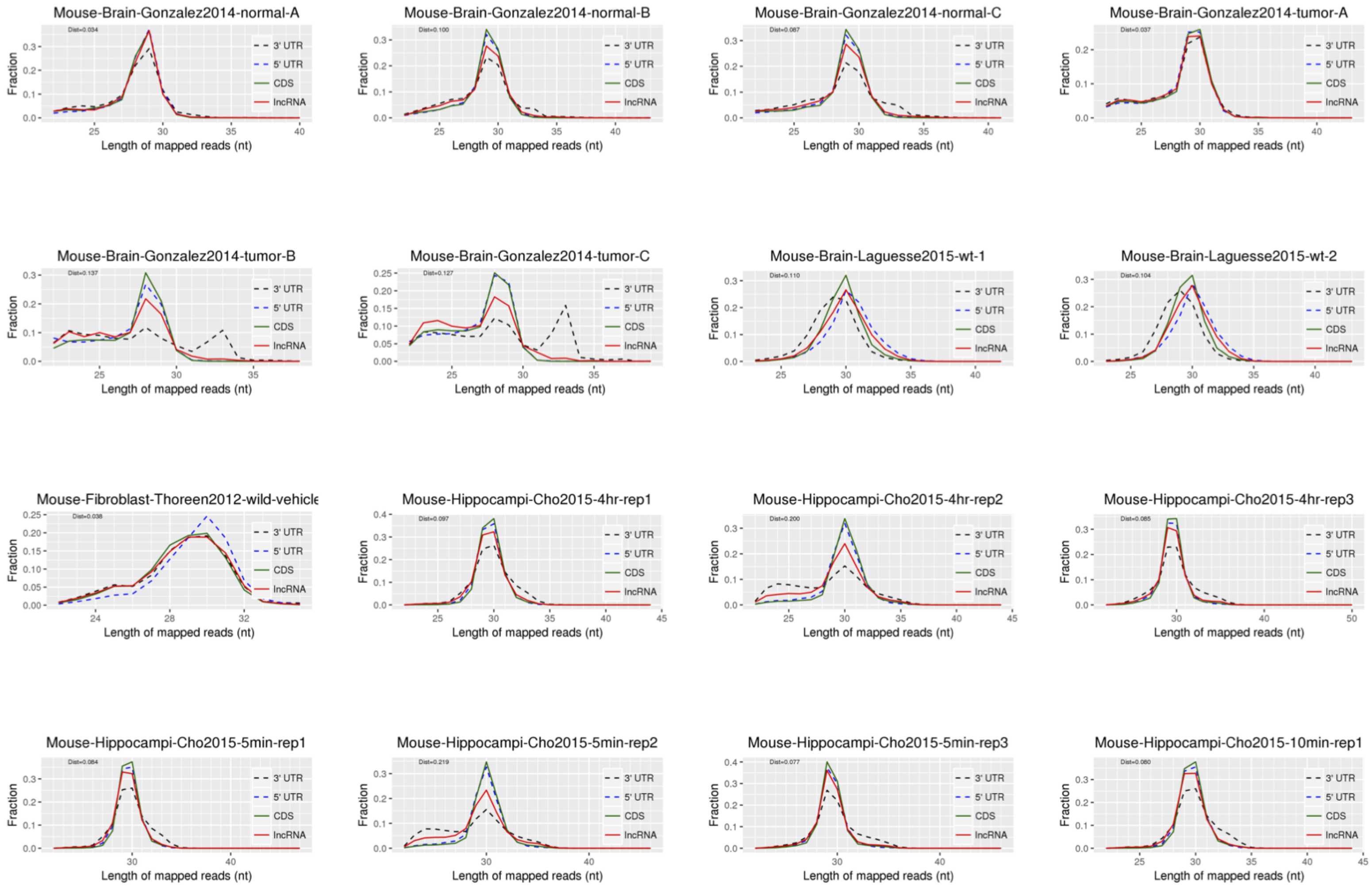
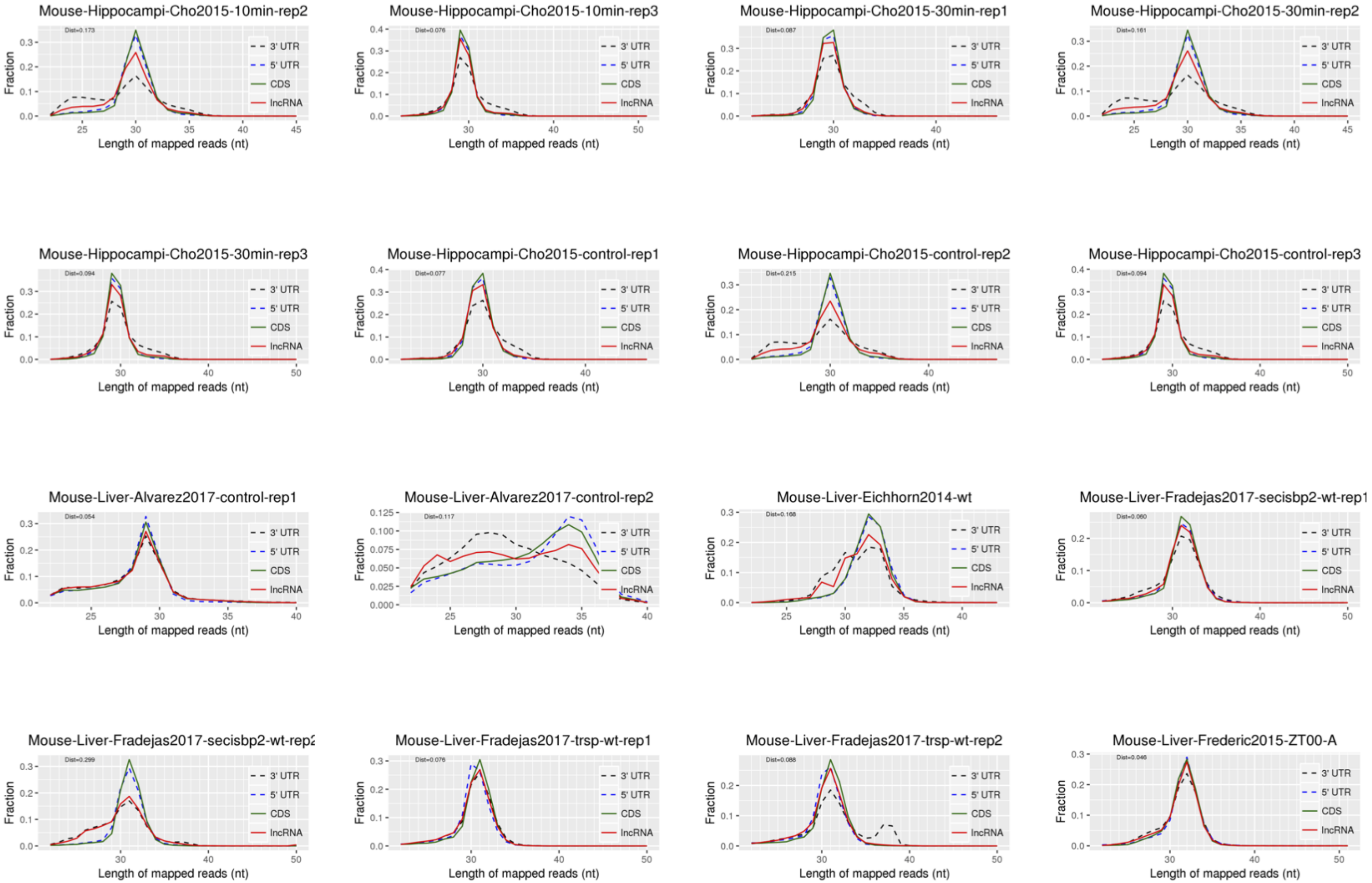


Fig. S2 Frequency distributions of Ribo-seq read lengths across CDSs, $5'/3'$UTRs, and lncRNAs (mouse)



Mouse-Hippocampi-Cho2015-10min-rep2

Mouse-Hippocampi-Cho2015-10min-rep3

Mouse-Hippocampi-Cho2015-30min-rep1

Mouse-Hippocampi-Cho2015-30min-rep2

Mouse-Hippocampi-Cho2015-30min-rep3

Mouse-Hippocampi-Cho2015-control-rep1

Mouse-Hippocampi-Cho2015-control-rep2

Mouse-Hippocampi-Cho2015-control-rep3

Mouse-Liver-Alvarez2017-control-rep1

Mouse-Liver-Alvarez2017-control-rep2

Mouse-Liver-Eichhorn2014-wt

Mouse-Liver-Fradejas2017-secisbp2-wt-rep1

Mouse-Liver-Fradejas2017-secisbp2-wt-rep2

Mouse-Liver-Fradejas2017-trsp-wt-rep1

Mouse-Liver-Fradejas2017-trsp-wt-rep2

Mouse-Liver-Frederic2015-ZT00-A

Fig. S2 (continued)

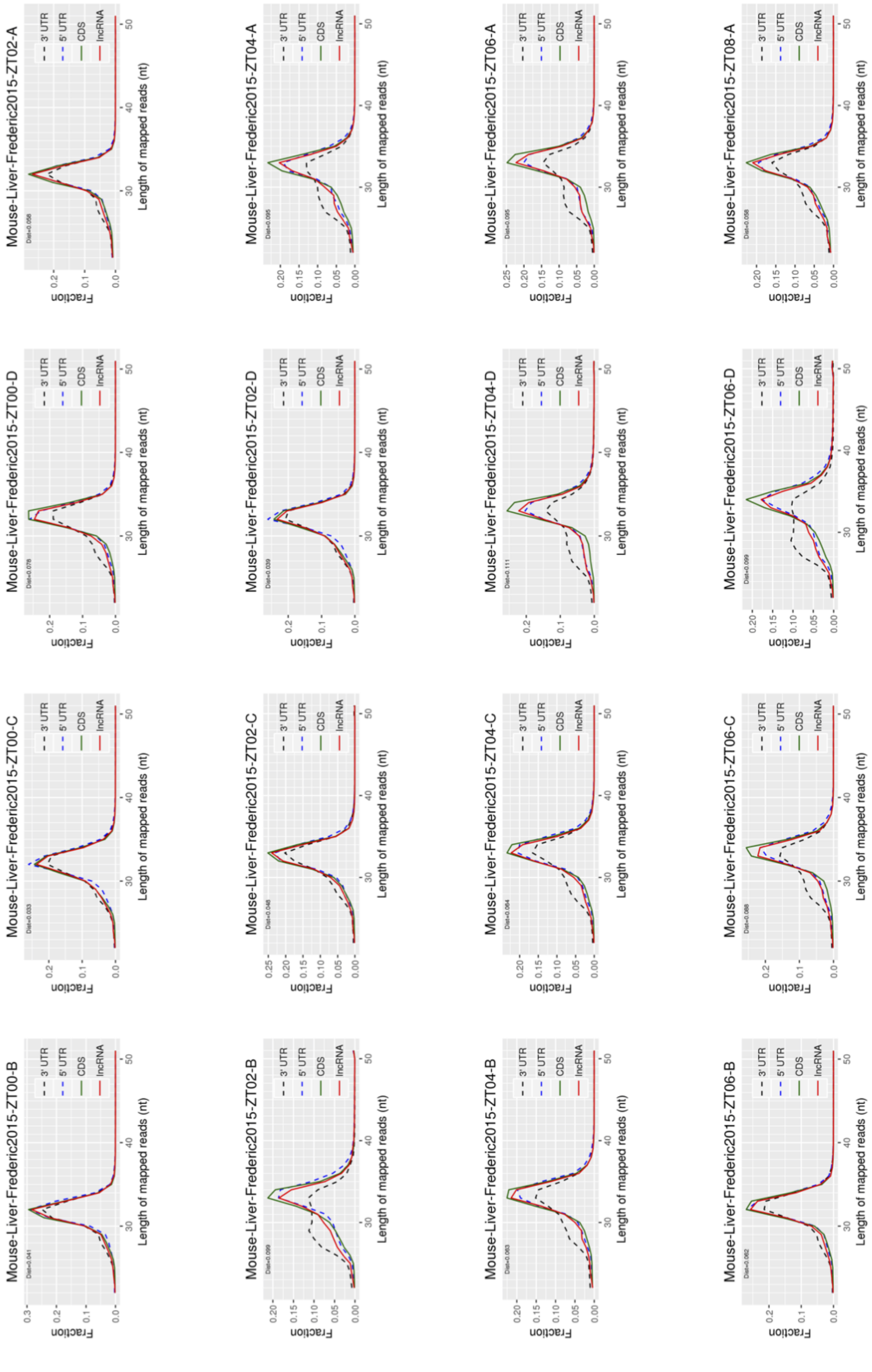


Fig. S2 (continued)

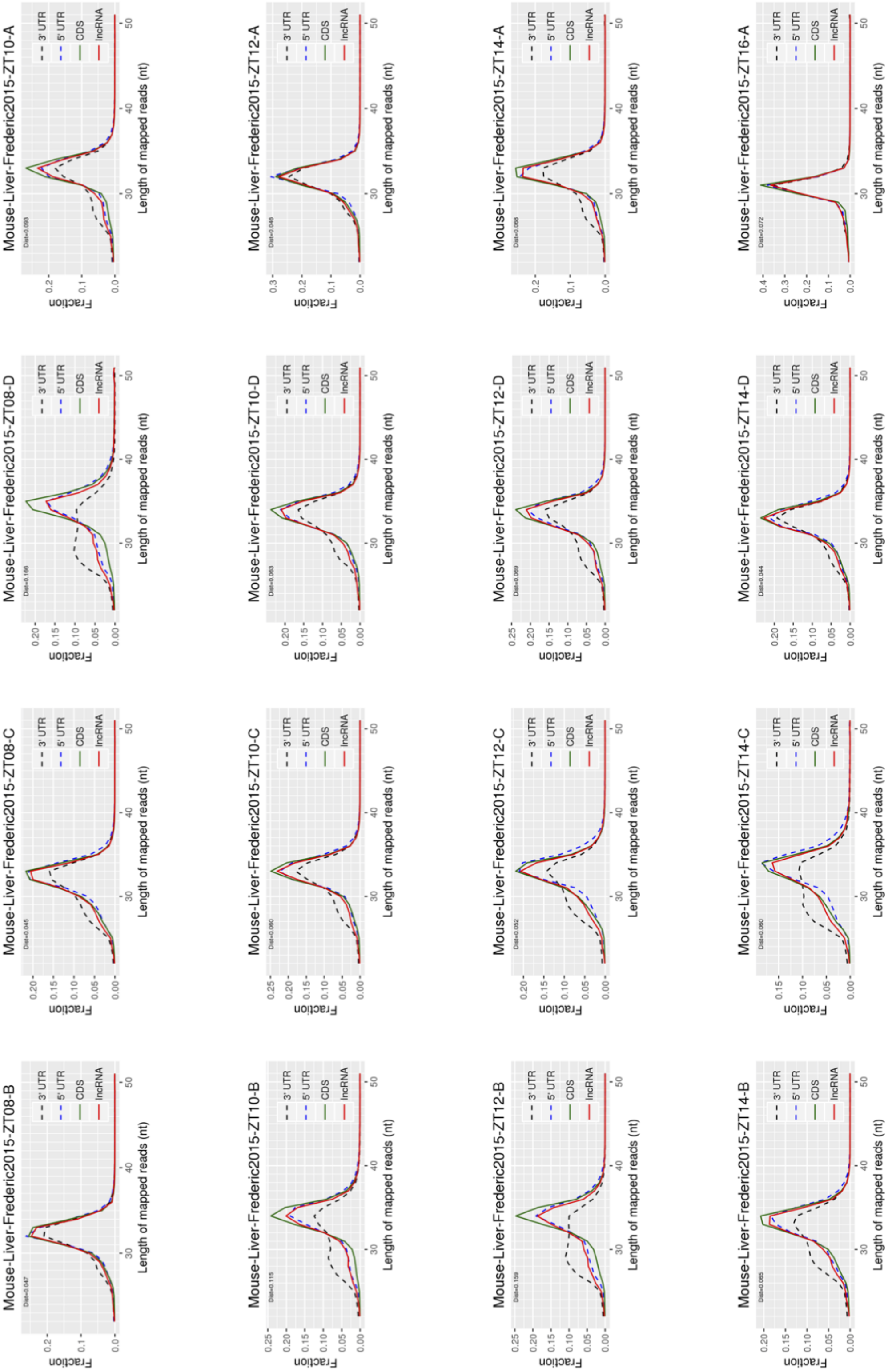


Fig. S2 (continued)

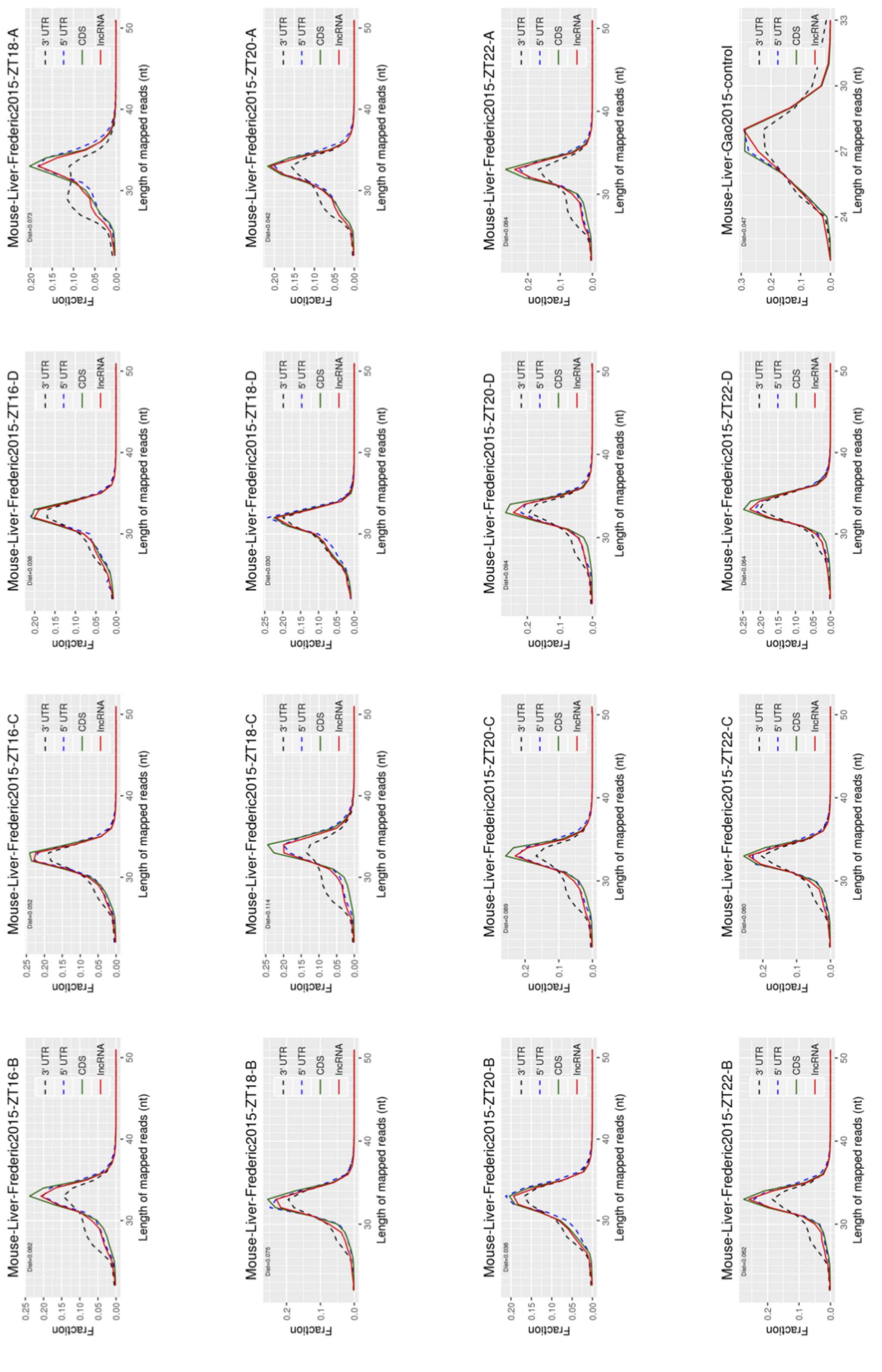


Fig. S2 (continued)

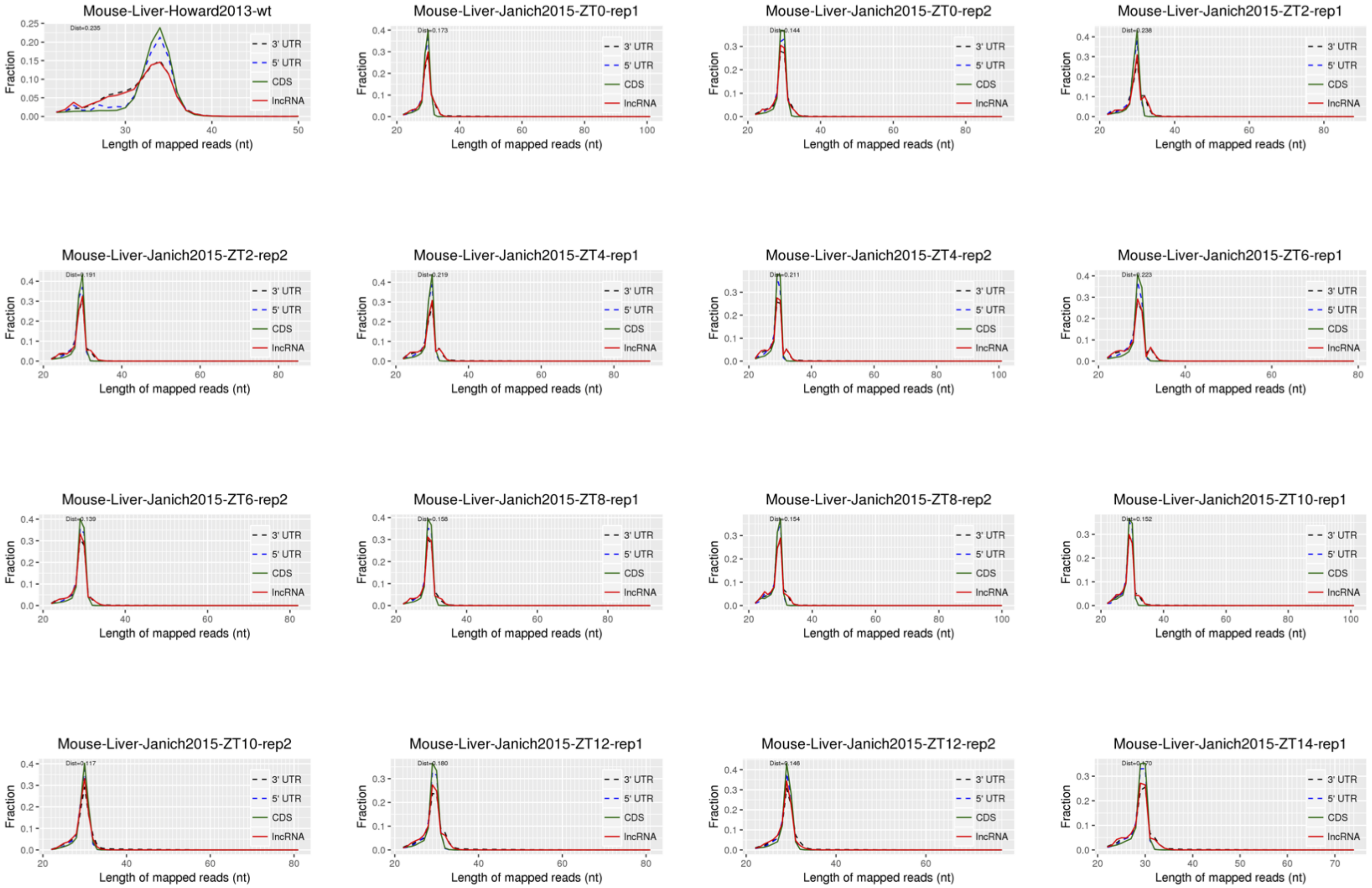


Fig. S2 (continued)

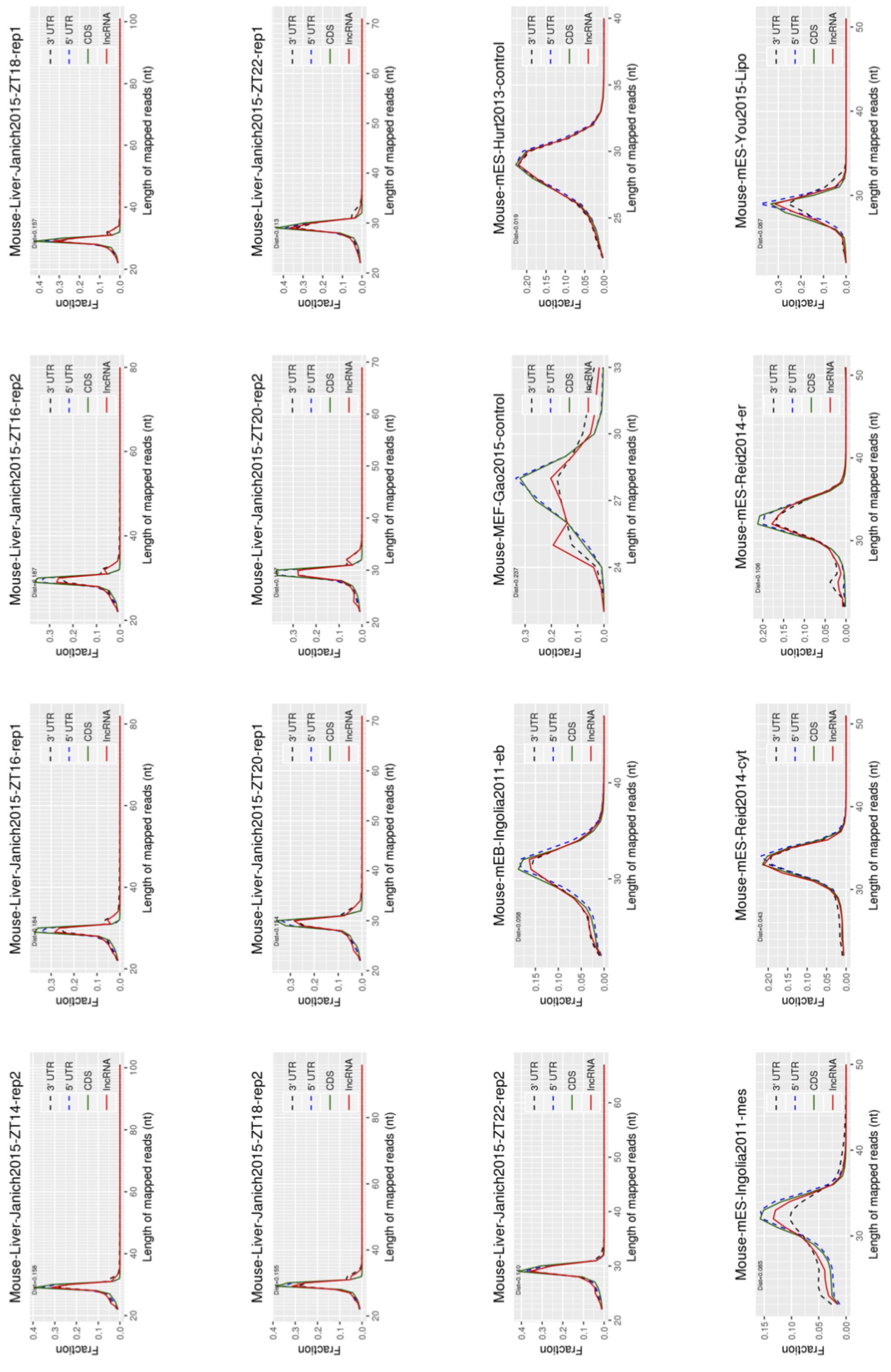


Fig. S2 (continued)

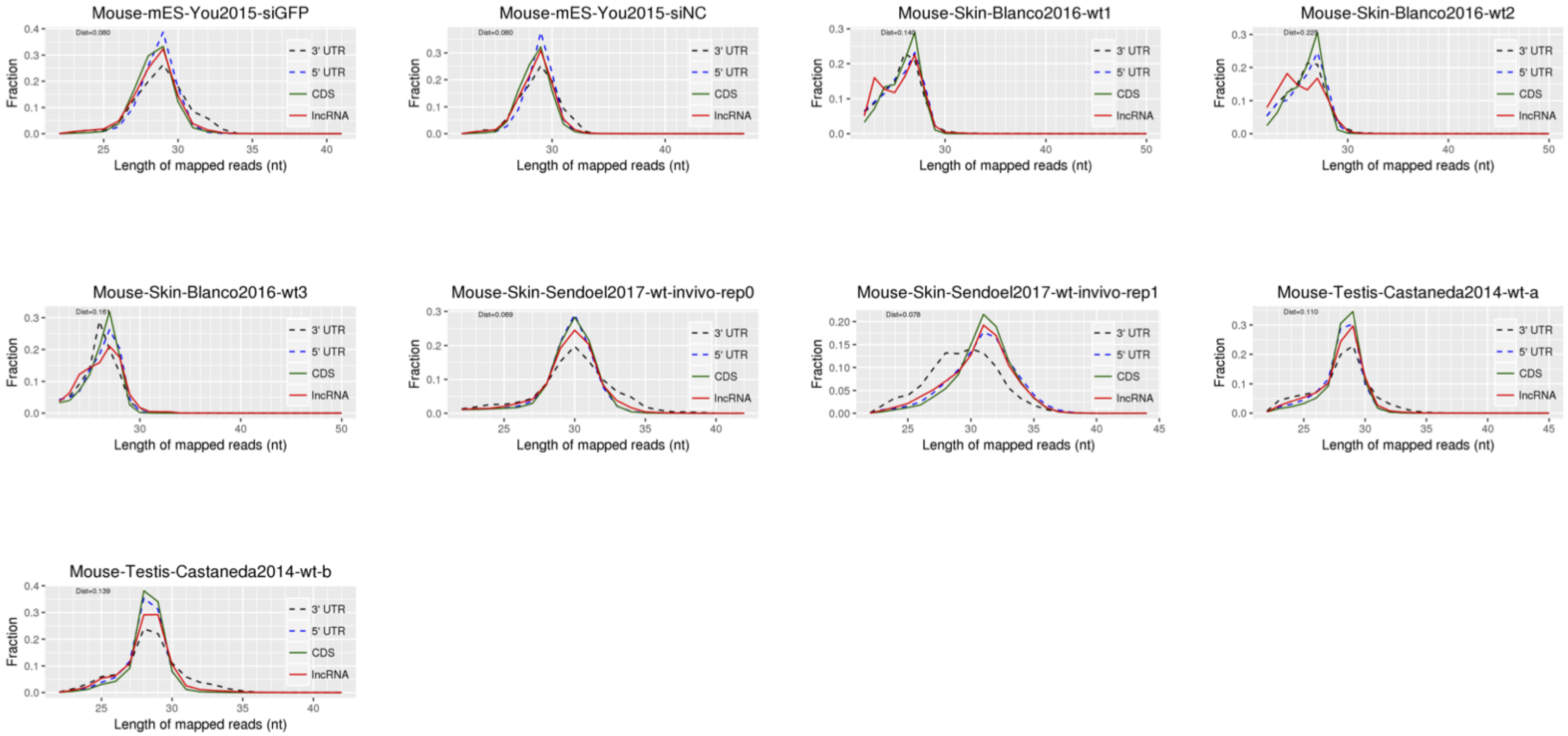


Fig. S2 (continued)

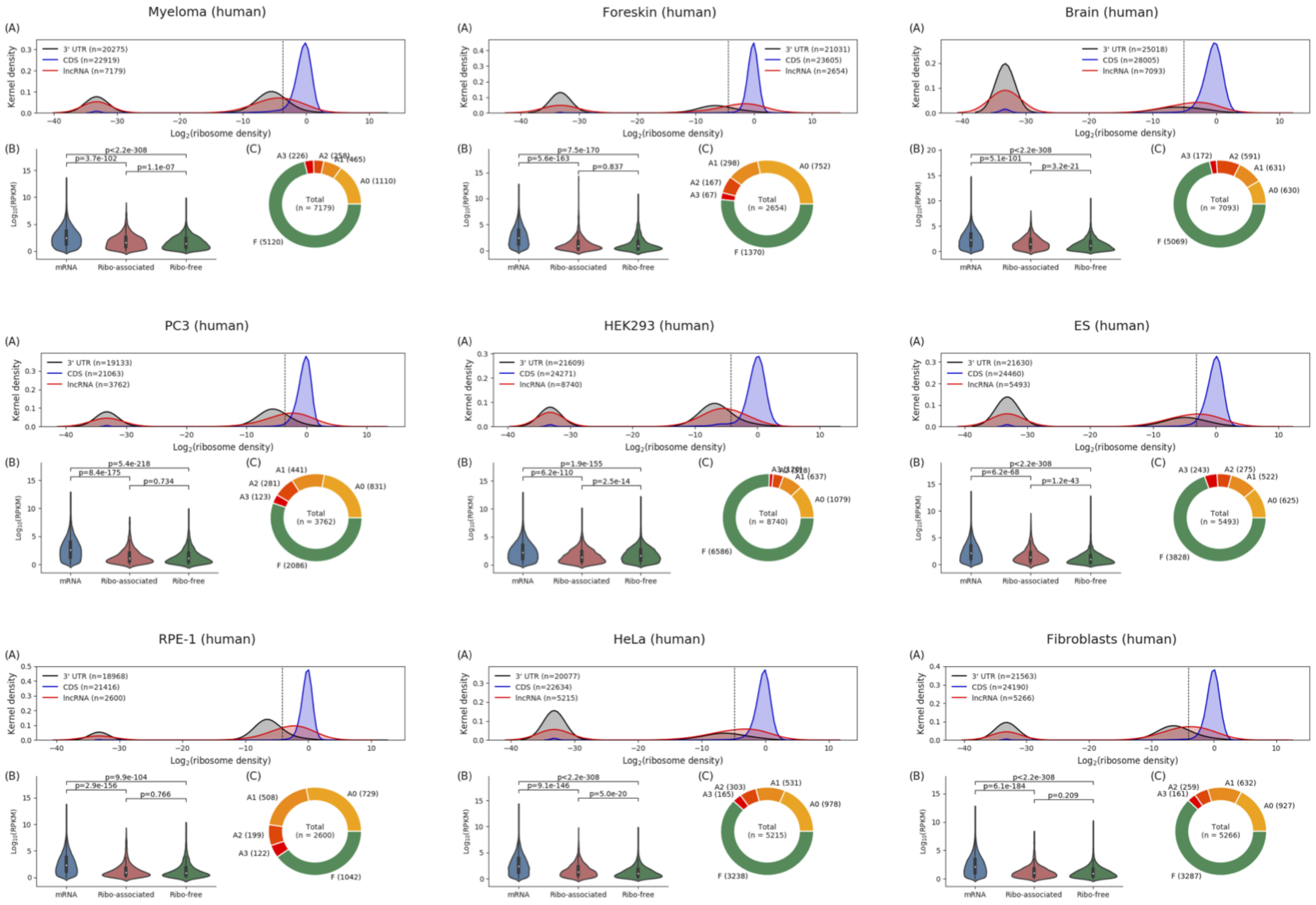


Fig. S3 The discrimination of ribosome-associated and ribosome-free lncRNAs by ribosome density in all selected datasets

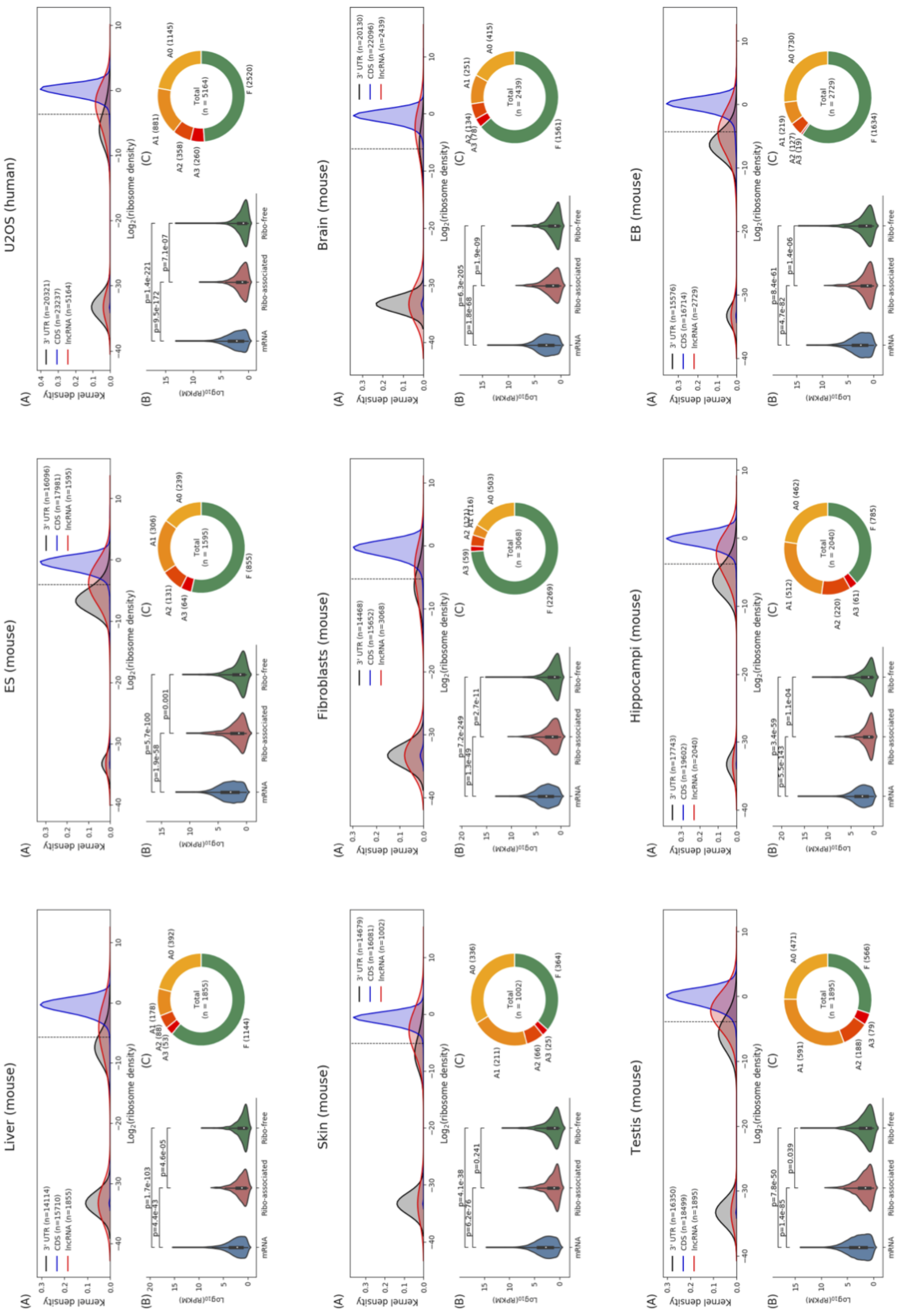


Fig. S3 (continued)

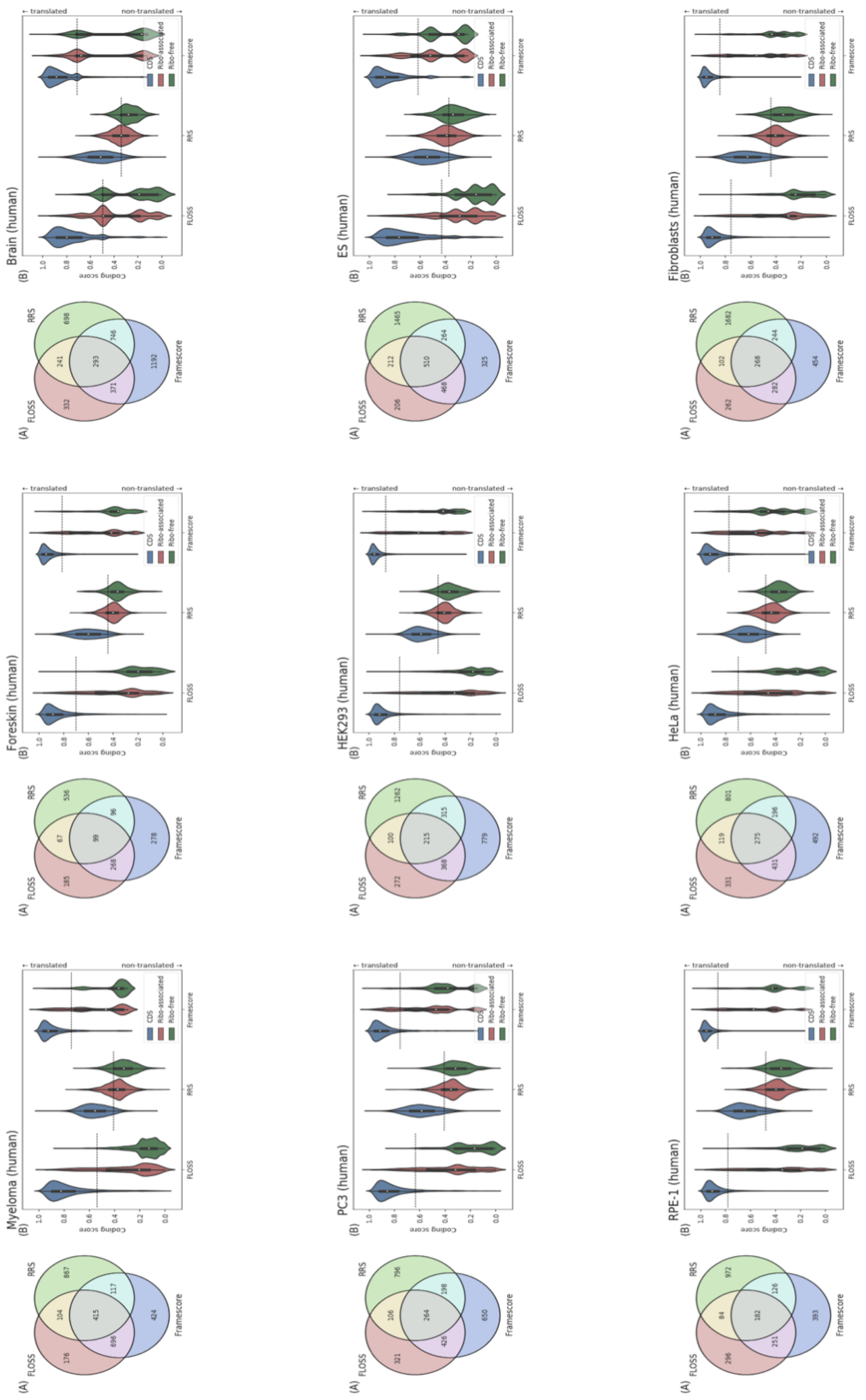


Fig. S4 Analysis of coding potential by using FLOSS, RRS, and Framescore in all selected datasets

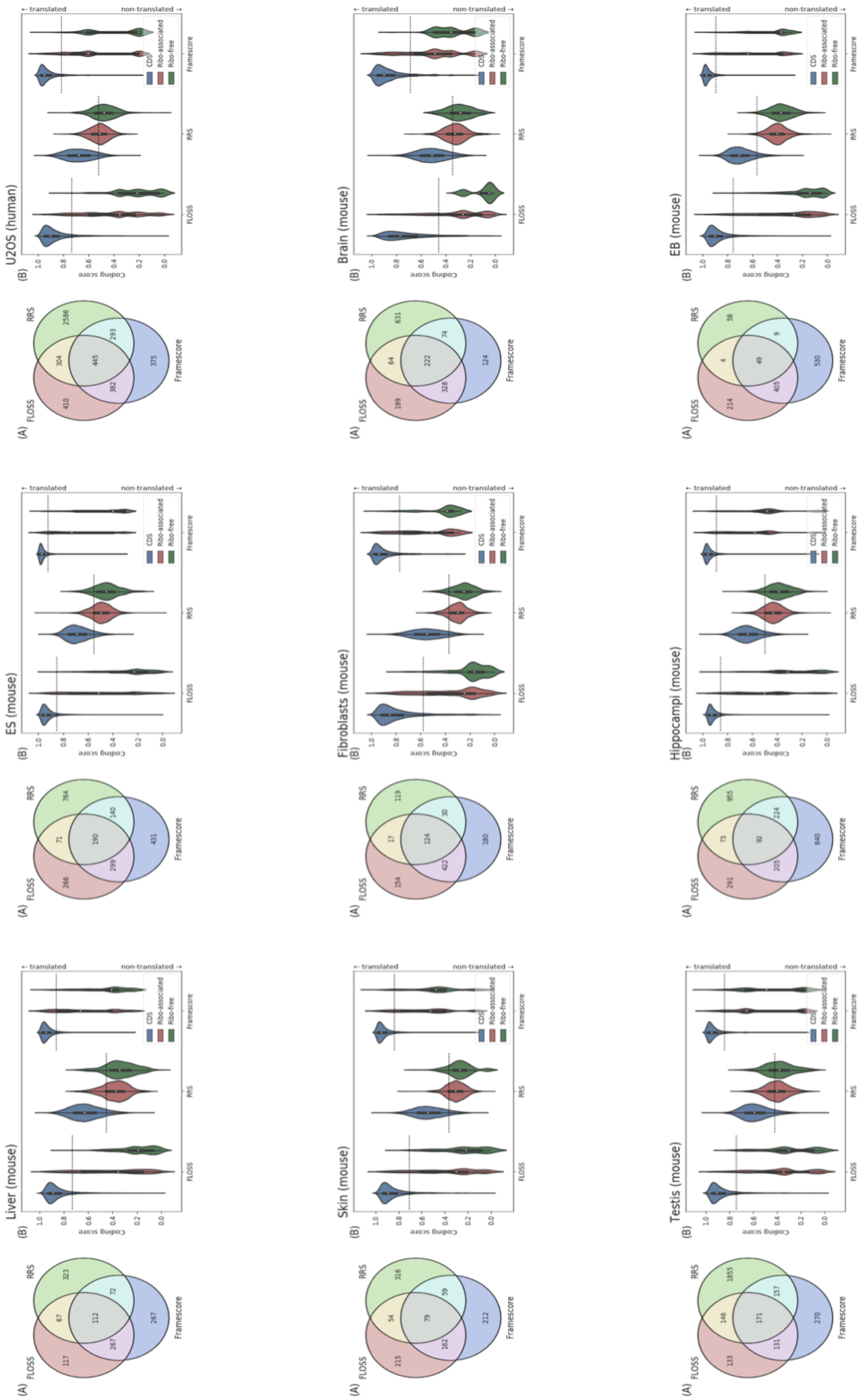
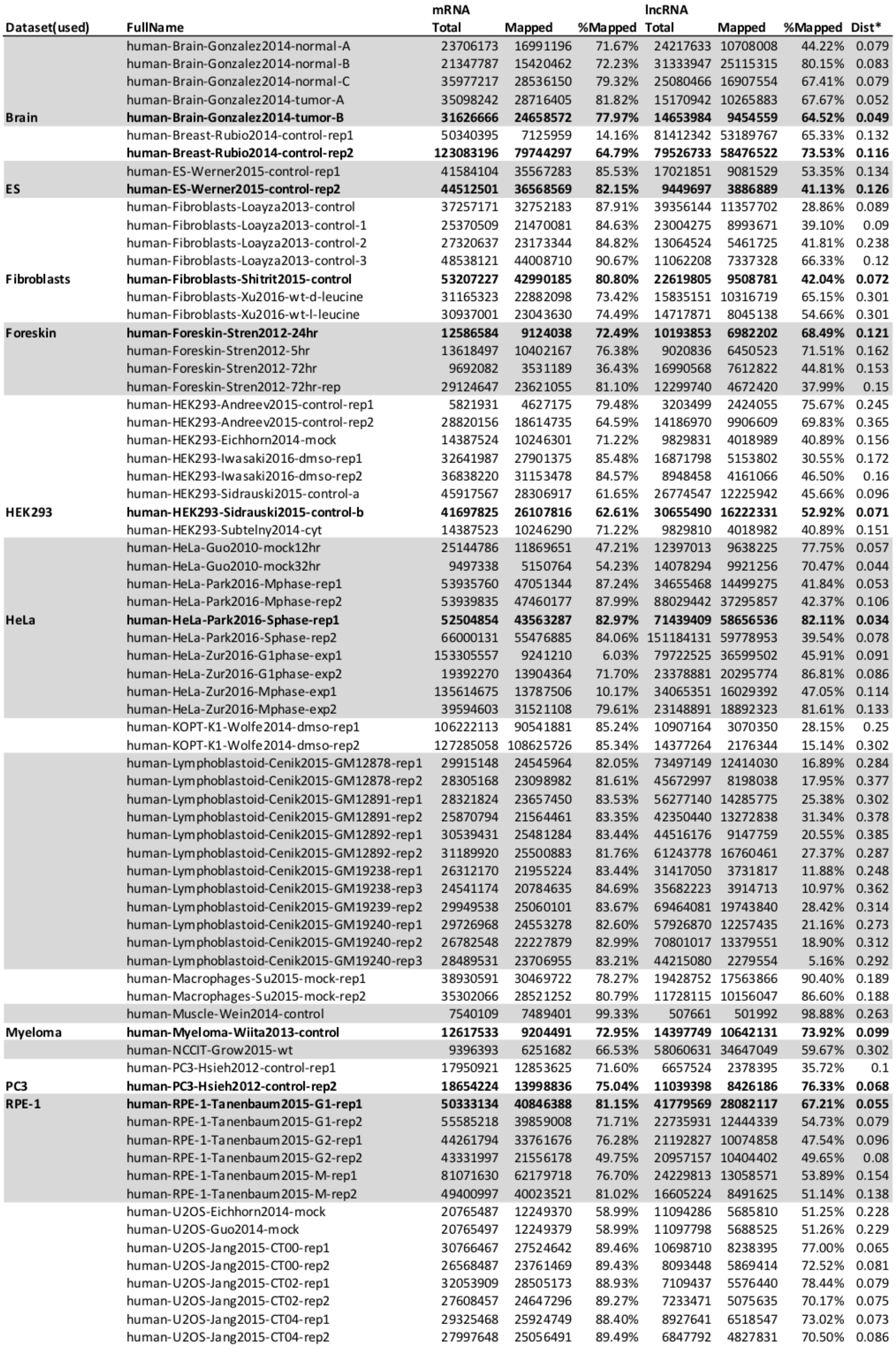


Fig. S4 (continued)

Table S1 Mapping statistics for RNA-seq and Ribo-seq reads



		mRNA			lncRNA			
Dataset(used)	FullName	Total	Mapped	%Mapped	Total	Mapped	%Mapped	Dist*
	human-Brain-Gonzalez2014-normal-A	23706173	16991196	71.67%	24217633	10708008	44.22%	0.079
	human-Brain-Gonzalez2014-normal-B	21347787	15420462	72.23%	31333947	25115315	80.15%	0.083
	human-Brain-Gonzalez2014-normal-C	35977217	28536150	79.32%	25080466	16907554	67.41%	0.079
	human-Brain-Gonzalez2014-tumor-A	35098242	28716405	81.82%	15170942	10265883	67.67%	0.052
Brain	**human-Brain-Gonzalez2014-tumor-B**	**31626666**	**24658572**	**77.97%**	**14653984**	**9454559**	**64.52%**	**0.049**
	human-Breast-Rubio2014-control-rep1	50340395	7125959	14.16%	81412342	53189767	65.33%	0.132
	human-Breast-Rubio2014-control-rep2	**123083196**	**79744297**	**64.79%**	**79526733**	**58476522**	**73.53%**	**0.116**
	human-ES-Werner2015-control-rep1	41584104	35567283	85.53%	17021851	9081529	53.35%	0.134
ES	**human-ES-Werner2015-control-rep2**	**44512501**	**36568569**	**82.15%**	**9449697**	**3886889**	**41.13%**	**0.126**
	human-Fibroblasts-Loayza2013-control	37257171	32752183	87.91%	39356144	11357702	28.86%	0.089
	human-Fibroblasts-Loayza2013-control-1	25370509	21470081	84.63%	23004275	8993671	39.10%	0.09
	human-Fibroblasts-Loayza2013-control-2	27320637	23173344	84.82%	13064524	5461725	41.81%	0.238
	human-Fibroblasts-Loayza2013-control-3	48538121	44008710	90.67%	11062208	7337328	66.33%	0.12
Fibroblasts	**human-Fibroblasts-Shitrit2015-control**	**53207227**	**42990185**	**80.80%**	**22619805**	**9508781**	**42.04%**	**0.072**
	human-Fibroblasts-Xu2016-wt-d-leucine	31165323	22882098	73.42%	15835151	10316719	65.15%	0.301
	human-Fibroblasts-Xu2016-wt-l-leucine	30937001	23043630	74.49%	14717871	8045138	54.66%	0.301
Foreskin	**human-Foreskin-Stren2012-24hr**	**12586584**	**9124038**	**72.49%**	**10193853**	**6982202**	**68.49%**	**0.121**
	human-Foreskin-Stren2012-5hr	13618497	10402167	76.38%	9020836	6450523	71.51%	0.162
	human-Foreskin-Stren2012-72hr	9692082	3531189	36.43%	16990568	7612822	44.81%	0.153
	human-Foreskin-Stren2012-72hr-rep	29124647	23621055	81.10%	12299740	4672420	37.99%	0.15
	human-HEK293-Andreev2015-control-rep1	5821931	4627175	79.48%	3203499	2424055	75.67%	0.245
	human-HEK293-Andreev2015-control-rep2	28820156	18614735	64.59%	14186970	9906609	69.83%	0.365
	human-HEK293-Eichhorn2014-mock	14387524	10246301	71.22%	9829831	4018989	40.89%	0.156
	human-HEK293-Iwasaki2016-dmso-rep1	32641987	27901375	85.48%	16871798	5153802	30.55%	0.172
	human-HEK293-Iwasaki2016-dmso-rep2	36838220	31153478	84.57%	8948458	4161066	46.50%	0.16
	human-HEK293-Sidrauski2015-control-a	45917567	28306917	61.65%	26774547	12225942	45.66%	0.096
HEK293	**human-HEK293-Sidrauski2015-control-b**	**41697825**	**26107816**	**62.61%**	**30655490**	**16222331**	**52.92%**	**0.071**
	human-HEK293-Subtelny2014-cyt	14387523	10246290	71.22%	9829810	4018982	40.89%	0.151
	human-HeLa-Guo2010-mock12hr	25144786	11869651	47.21%	12397013	9638225	77.75%	0.057
	human-HeLa-Guo2010-mock32hr	9497338	5150764	54.23%	14078294	9921256	70.47%	0.044
	human-HeLa-Park2016-Mphase-rep1	53935760	47051344	87.24%	34655468	14499275	41.84%	0.053
	human-HeLa-Park2016-Mphase-rep2	53939835	47460177	87.99%	88029442	37295857	42.37%	0.106
HeLa	**human-HeLa-Park2016-Sphase-rep1**	**52504854**	**43563287**	**82.97%**	**71439409**	**58656536**	**82.11%**	**0.034**
	human-HeLa-Park2016-Sphase-rep2	66000131	55476885	84.06%	151184131	59778953	39.54%	0.078
	human-HeLa-Zur2016-G1phase-exp1	153305557	9241210	6.03%	79722525	36599502	45.91%	0.091
	human-HeLa-Zur2016-G1phase-exp2	19392270	13904364	71.70%	23378881	20295774	86.81%	0.086
	human-HeLa-Zur2016-Mphase-exp1	135614675	13787506	10.17%	34065351	16029392	47.05%	0.114
	human-HeLa-Zur2016-Mphase-exp2	39594603	31521108	79.61%	23148891	18892323	81.61%	0.133
	human-KOPT-K1-Wolfe2014-dmso-rep1	106222113	90541881	85.24%	10907164	3070350	28.15%	0.25
	human-KOPT-K1-Wolfe2014-dmso-rep2	127285058	108625726	85.34%	14377264	2176344	15.14%	0.302
	human-Lymphoblastoid-Cenik2015-GM12878-rep1	29915148	24545964	82.05%	73497149	12414030	16.89%	0.284
	human-Lymphoblastoid-Cenik2015-GM12878-rep2	28305168	23098982	81.61%	45672997	8198038	17.95%	0.377
	human-Lymphoblastoid-Cenik2015-GM12891-rep1	28321824	23657450	83.53%	56277140	14285775	25.38%	0.302
	human-Lymphoblastoid-Cenik2015-GM12891-rep2	25870794	21564461	83.35%	42350440	13272838	31.34%	0.378
	human-Lymphoblastoid-Cenik2015-GM12892-rep1	30539431	25481284	83.44%	44516176	9147759	20.55%	0.385
	human-Lymphoblastoid-Cenik2015-GM12892-rep2	31189920	25500883	81.76%	61243778	16760461	27.37%	0.287
	human-Lymphoblastoid-Cenik2015-GM19238-rep1	26312170	21955224	83.44%	31417050	3731817	11.88%	0.248
	human-Lymphoblastoid-Cenik2015-GM19238-rep3	24541174	20784635	84.69%	35682223	3914713	10.97%	0.362
	human-Lymphoblastoid-Cenik2015-GM19239-rep2	29949538	25060101	83.67%	69464081	19743840	28.42%	0.314
	human-Lymphoblastoid-Cenik2015-GM19240-rep1	29726968	24553278	82.60%	57926870	12257435	21.16%	0.273
	human-Lymphoblastoid-Cenik2015-GM19240-rep2	26782548	22227879	82.99%	70801017	13379551	18.90%	0.312
	human-Lymphoblastoid-Cenik2015-GM19240-rep3	28489531	23706955	83.21%	44215080	2279554	5.16%	0.292
	human-Macrophages-Su2015-mock-rep1	38930591	30469722	78.27%	19428752	17563866	90.40%	0.189
	human-Macrophages-Su2015-mock-rep2	35302066	28521252	80.79%	11728115	10156047	86.60%	0.188
	human-Muscle-Wein2014-control	7540109	7489401	99.33%	507661	501992	98.88%	0.263
Myeloma	**human-Myeloma-Wiita2013-control**	**12617533**	**9204491**	**72.95%**	**14397749**	**10642131**	**73.92%**	**0.099**
	human-NCCIT-Grow2015-wt	9396393	6251682	66.53%	58060631	34647049	59.67%	0.302
	human-PC3-Hsieh2012-control-rep1	17950921	12853625	71.60%	6657524	2378395	35.72%	0.1
PC3	**human-PC3-Hsieh2012-control-rep2**	**18654224**	**13998836**	**75.04%**	**11039398**	**8426186**	**76.33%**	**0.068**
RPE-1	**human-RPE-1-Tanenbaum2015-G1-rep1**	**50333134**	**40846388**	**81.15%**	**41779569**	**28082117**	**67.21%**	**0.055**
	human-RPE-1-Tanenbaum2015-G1-rep2	55585218	39859008	71.71%	22735931	12444339	54.73%	0.079
	human-RPE-1-Tanenbaum2015-G2-rep1	44261794	33761676	76.28%	21192827	10074858	47.54%	0.096
	human-RPE-1-Tanenbaum2015-G2-rep2	43331997	21556178	49.75%	20957157	10404402	49.65%	0.08
	human-RPE-1-Tanenbaum2015-M-rep1	81071630	62179718	76.70%	24229813	13058571	53.89%	0.154
	human-RPE-1-Tanenbaum2015-M-rep2	49400997	40023521	81.02%	16605224	8491625	51.14%	0.138
	human-U2OS-Eichhorn2014-mock	20765487	12249370	58.99%	11094286	5685810	51.25%	0.228
	human-U2OS-Guo2014-mock	20765497	12249379	58.99%	11097798	5688525	51.26%	0.229
	human-U2OS-Jang2015-CT00-rep1	30766467	27524642	89.46%	10698710	8238395	77.00%	0.065
	human-U2OS-Jang2015-CT00-rep2	26568487	23761469	89.43%	8093448	5869414	72.52%	0.081
	human-U2OS-Jang2015-CT02-rep1	32053909	28505173	88.93%	7109437	5576440	78.44%	0.079
	human-U2OS-Jang2015-CT02-rep2	27608457	24647296	89.27%	7233471	5075635	70.17%	0.075
	human-U2OS-Jang2015-CT04-rep1	29325468	25924749	88.40%	8927641	6518547	73.02%	0.073
	human-U2OS-Jang2015-CT04-rep2	27997648	25056491	89.49%	6847792	4827831	70.50%	0.086

Table S1 (continued)

	human-U2OS-Jang2015-CT06-rep1	31538781	27599221	87.51%	9666763	7420551	76.76%	0.073
	human-U2OS-Jang2015-CT06-rep2	30991243	27708189	89.41%	12240724	9023725	73.72%	0.078
	human-U2OS-Jang2015-CT08-rep1	29291180	25973579	88.67%	10010105	8103457	80.95%	0.072
	human-U2OS-Jang2015-CT08-rep2	28186680	25103379	89.06%	8997894	7081787	78.70%	0.077
	human-U2OS-Jang2015-CT10-rep1	28867822	25496675	88.32%	8404007	6371930	75.82%	0.074
	human-U2OS-Jang2015-CT10-rep2	31234143	27526578	88.13%	9982689	7495258	75.08%	0.078
	human-U2OS-Jang2015-CT12-rep1	26613183	23555262	88.51%	11118970	7928171	71.30%	0.07
	human-U2OS-Jang2015-CT12-rep2	28002620	24927575	89.02%	7126120	4050069	56.83%	0.08
	human-U2OS-Jang2015-CT14-rep1	31000585	27548946	88.87%	8209464	6325949	77.06%	0.069
	human-U2OS-Jang2015-CT14-rep2	30829679	27415638	88.93%	10345199	7643952	73.89%	0.083
	human-U2OS-Jang2015-CT16-rep1	26693598	23855127	89.37%	5294481	4052150	76.54%	0.072
	human-U2OS-Jang2015-CT16-rep2	29921626	26733926	89.35%	8826312	6621706	75.02%	0.078
	human-U2OS-Jang2015-CT18-rep1	26105761	23154457	88.69%	6668649	4501141	67.50%	0.064
	human-U2OS-Jang2015-CT18-rep2	31325337	28107322	89.73%	9742329	6930868	71.14%	0.073
U2OS	**human-U2OS-Jang2015-CT20-rep1**	**29915338**	**26678709**	**89.18%**	**8502522**	**6290697**	**73.99%**	**0.064**
	human-U2OS-Jang2015-CT20-rep2	28135499	25308070	89.95%	10808655	8157323	75.47%	0.067
	human-U2OS-Jang2015-CT22-rep1	30156260	26776982	88.79%	12706599	9498408	74.75%	0.065
	human-U2OS-Jang2015-CT22-rep2	26765823	23998397	89.66%	8227433	5833278	70.90%	0.066
	mouse-Brain-Gonzalez2014-normal-A	33270791	27614324	83.00%	2522944	442908	17.56%	0.034
	mouse-Brain-Gonzalez2014-normal-B	33360725	27386092	82.09%	16698604	11583554	69.37%	0.1
	mouse-Brain-Gonzalez2014-normal-C	33881050	28359821	83.70%	11113005	6090392	54.80%	0.087
Brain	**mouse-Brain-Gonzalez2014-tumor-A**	**37120052**	**26551951**	**71.53%**	**14367807**	**6215221**	**43.26%**	**0.037**
	mouse-Brain-Gonzalez2014-tumor-B	28006296	22512124	80.38%	30945876	1047535	3.39%	0.137
	mouse-Brain-Gonzalez2014-tumor-C	31433374	25817920	82.14%	12901149	458937	3.56%	0.127
	mouse-Brain-Laguesse2015-wt-1	25468840	23146791	90.88%	19357901	11246784	58.10%	0.11
	mouse-Brain-Laguesse2015-wt-2	19128723	17476386	91.36%	21404954	12300424	57.47%	0.104
EB	**mouse-EB-Ingolia2011-eb**	**44474377**	**34359886**	**77.26%**	**39142735**	**25506644**	**65.16%**	**0.058**
ES	**mouse-ES-Hurt2013-control**	**44200486**	**40690743**	**92.06%**	**70747244**	**44057365**	**62.27%**	**0.019**
	mouse-ES-Ingolia2011-mes	59790833	49112252	82.14%	48811567	28303241	57.98%	0.085
	mouse-ES-Reid2014-cyt	45958696	37763078	82.17%	10184184	7401189	72.67%	0.043
	mouse-ES-Reid2014-er	56320987	24349018	43.23%	45081181	27375999	60.73%	0.106
	mouse-ES-You2015-Lipo	92713371	58873929	63.50%	54200488	41925126	77.35%	0.087
	mouse-ES-You2015-siGFP	94540100	60345402	63.83%	49928706	37901320	75.91%	0.08
	mouse-ES-You2015-siNC	100988376	64013168	63.39%	62157133	47750090	76.82%	0.08
Fibroblasts	**mouse-Fibroblast-Thoreen2012-wild-vehicle**	**6383082**	**3571804**	**55.96%**	**5996501**	**3997946**	**66.67%**	**0.038**
	mouse-Hippocampi-Cho2015-10min-rep1	101438895	31392345	30.95%	63906806	38476139	60.21%	0.08
	mouse-Hippocampi-Cho2015-10min-rep2	79682904	25551313	32.07%	71962828	26326298	36.58%	0.173
Hippocampi	**mouse-Hippocampi-Cho2015-10min-rep3**	**77253662**	**24824683**	**32.13%**	**59412790**	**27811123**	**46.81%**	**0.076**
	mouse-Hippocampi-Cho2015-30min-rep1	99165827	30980431	31.24%	68045914	42027923	61.76%	0.087
	mouse-Hippocampi-Cho2015-30min-rep2	83706253	25431361	30.38%	76954045	28258246	36.72%	0.161
	mouse-Hippocampi-Cho2015-30min-rep3	72231705	24568230	34.01%	58314315	29339899	50.31%	0.094
	mouse-Hippocampi-Cho2015-4hr-rep1	92607383	28704664	31.00%	58415171	36823640	63.04%	0.097
	mouse-Hippocampi-Cho2015-4hr-rep2	72710479	21696274	29.84%	60702086	19602931	32.29%	0.2
	mouse-Hippocampi-Cho2015-4hr-rep3	68806057	23027841	33.47%	60105894	28667564	47.70%	0.085
	mouse-Hippocampi-Cho2015-5min-rep1	86962503	25955207	29.85%	57225946	33246425	58.10%	0.084
	mouse-Hippocampi-Cho2015-5min-rep2	76613300	24340433	31.77%	62115557	23492391	37.82%	0.219
	mouse-Hippocampi-Cho2015-5min-rep3	82499197	26509424	32.13%	52286387	24643127	47.13%	0.077
	mouse-Hippocampi-Cho2015-control-rep1	89250671	27310697	30.60%	57407818	34276751	59.71%	0.077
	mouse-Hippocampi-Cho2015-control-rep2	84226124	28087405	33.35%	68682721	27642831	40.25%	0.215
	mouse-Hippocampi-Cho2015-control-rep3	78929229	26834524	34.00%	62221835	30705396	49.35%	0.094
	mouse-Liver-Alvarez2017-control-rep1	37544658	14983921	39.91%	21508992	3830928	17.81%	0.054
	mouse-Liver-Alvarez2017-control-rep2	21836373	13794543	63.17%	20241198	3107406	15.35%	0.117
	mouse-Liver-Eichhorn2014-wt	41427924	12417208	29.97%	40903891	19130139	46.77%	0.168
	mouse-Liver-Fradejas2017-secisbp2-wt-rep1	58481818	47567139	81.34%	18538329	13752643	74.18%	0.06
	mouse-Liver-Fradejas2017-secisbp2-wt-rep2	70260478	50268046	71.55%	23269466	17774081	76.38%	0.299
	mouse-Liver-Fradejas2017-trsp-wt-rep1	45357002	34404379	75.85%	13928674	10968035	78.74%	0.076
	mouse-Liver-Fradejas2017-trsp-wt-rep2	48240351	36338081	75.33%	11533817	7078147	61.37%	0.088
	mouse-Liver-Frederic2015-ZT00-A	61463219	46194716	75.16%	39881861	34195972	85.74%	0.046
	mouse-Liver-Frederic2015-ZT00-B	42213367	31209585	73.93%	43229442	37844873	87.54%	0.041
	mouse-Liver-Frederic2015-ZT00-C	58324789	44798969	76.81%	36190320	30479336	84.22%	0.033
	mouse-Liver-Frederic2015-ZT00-D	59970294	46049152	76.79%	45009390	39246747	87.20%	0.078
	mouse-Liver-Frederic2015-ZT02-A	59844576	45970942	76.82%	39250281	29287966	74.62%	0.058
	mouse-Liver-Frederic2015-ZT02-B	55928223	43494730	77.77%	40227347	27271002	67.79%	0.099
	mouse-Liver-Frederic2015-ZT02-C	59515906	46220179	77.66%	41028413	35148119	85.67%	0.048
	mouse-Liver-Frederic2015-ZT02-D	53866993	41218938	76.52%	36679695	31117317	84.84%	0.039
	mouse-Liver-Frederic2015-ZT04-A	59594923	45798470	76.85%	41313712	31477055	76.19%	0.095
	mouse-Liver-Frederic2015-ZT04-B	50506345	38364374	75.96%	35356679	24480446	69.24%	0.063
	mouse-Liver-Frederic2015-ZT04-C	57252425	43470398	75.93%	41725080	34959894	83.79%	0.064
	mouse-Liver-Frederic2015-ZT04-D	56999400	43761599	76.78%	30863503	23361131	75.69%	0.111
	mouse-Liver-Frederic2015-ZT06-A	59821414	43961477	73.49%	39179136	31735760	81.00%	0.095
	mouse-Liver-Frederic2015-ZT06-B	45754295	32620585	71.30%	43512282	37487683	86.15%	0.062
	mouse-Liver-Frederic2015-ZT06-C	60171428	45414939	75.48%	33242273	25698867	77.31%	0.088
	mouse-Liver-Frederic2015-ZT06-D	58974501	42779502	72.54%	32071831	23554355	73.44%	0.099
	mouse-Liver-Frederic2015-ZT08-A	58647867	44546814	75.96%	39619185	29267358	73.87%	0.058
	mouse-Liver-Frederic2015-ZT08-B	59903194	44778139	74.75%	36798863	31972355	86.88%	0.047

Table S1 (continued)

	mouse-Liver-Frederic2015-ZT08-C	53912120	40215367	74.59%	37556683	31380779	83.56%	0.045
	mouse-Liver-Frederic2015-ZT08-D	50594412	37597704	74.31%	32965845	24571731	74.54%	0.166
	mouse-Liver-Frederic2015-ZT10-A	60025634	47430382	79.02%	26921799	16154699	60.01%	0.093
	mouse-Liver-Frederic2015-ZT10-B	55529276	42145755	75.90%	33183723	25829744	77.84%	0.115
	mouse-Liver-Frederic2015-ZT10-C	54529232	41861010	76.77%	34384031	27353087	79.55%	0.06
	mouse-Liver-Frederic2015-ZT10-D	56513483	40245363	71.21%	41281247	34994695	84.77%	0.063
	mouse-Liver-Frederic2015-ZT12-A	55237730	41689510	75.47%	49840117	44213128	88.71%	0.046
	mouse-Liver-Frederic2015-ZT12-B	55688206	41154168	73.90%	32448529	25519673	78.65%	0.159
	mouse-Liver-Frederic2015-ZT12-C	58121704	44747826	76.99%	31482795	23921648	75.98%	0.052
	mouse-Liver-Frederic2015-ZT12-D	59053115	45838790	77.62%	40932029	35408928	86.51%	0.069
	mouse-Liver-Frederic2015-ZT14-A	56463229	44384461	78.61%	59055226	52009455	88.07%	0.068
	mouse-Liver-Frederic2015-ZT14-B	53422282	41171291	77.07%	30912332	22749841	73.59%	0.065
	mouse-Liver-Frederic2015-ZT14-C	62323830	47384097	76.03%	40918944	32538658	79.52%	0.06
	mouse-Liver-Frederic2015-ZT14-D	58084806	44899768	77.30%	39085568	33401367	85.46%	0.044
	mouse-Liver-Frederic2015-ZT16-A	41353484	30972309	74.90%	39034912	29044031	74.41%	0.072
	mouse-Liver-Frederic2015-ZT16-B	54506996	41807944	76.70%	34160229	27774246	81.31%	0.082
	mouse-Liver-Frederic2015-ZT16-C	57241604	43966661	76.81%	41161040	35406805	86.02%	0.052
	mouse-Liver-Frederic2015-ZT16-D	60331531	43953852	72.85%	32933404	24156727	73.35%	0.038
	mouse-Liver-Frederic2015-ZT18-A	41786138	32052623	76.71%	50169938	35723740	71.21%	0.073
	mouse-Liver-Frederic2015-ZT18-B	51185766	39490659	77.15%	42434381	37418200	88.18%	0.075
	mouse-Liver-Frederic2015-ZT18-C	58914821	45390640	77.04%	35702491	28521102	79.89%	0.114
Liver	**mouse-Liver-Frederic2015-ZT18-D**	**49164797**	**37639770**	**76.56%**	**35106576**	**21073274**	**60.03%**	**0.03**
	mouse-Liver-Frederic2015-ZT20-A	61424054	45079231	73.39%	36416542	30599080	84.03%	0.042
	mouse-Liver-Frederic2015-ZT20-B	60595356	45727009	75.46%	34056033	29622594	86.98%	0.038
	mouse-Liver-Frederic2015-ZT20-C	43300737	33681699	77.79%	45426759	36641154	80.66%	0.089
	mouse-Liver-Frederic2015-ZT20-D	43124830	32865998	76.21%	52675489	44377211	84.25%	0.094
	mouse-Liver-Frederic2015-ZT22-A	61355344	48356738	78.81%	41931302	34688743	82.73%	0.084
	mouse-Liver-Frederic2015-ZT22-B	54540169	40193986	73.70%	40405417	34065090	84.31%	0.062
	mouse-Liver-Frederic2015-ZT22-C	61292646	47593473	77.65%	47430108	41300385	87.08%	0.06
	mouse-Liver-Frederic2015-ZT22-D	54357767	39584984	72.82%	43931094	38882714	88.51%	0.064
	mouse-Liver-Gao2015-control	39838628	26610950	66.80%	3216308	2115663	65.78%	0.047
	mouse-Liver-Howard2013-wt	63513172	33665691	53.01%	53100660	39529814	74.44%	0.235
	mouse-Liver-Janich2015-ZT0-rep1	40066009	20109991	50.19%	26374605	22343259	84.72%	0.173
	mouse-Liver-Janich2015-ZT0-rep2	33034581	17120213	51.83%	28480769	23668038	83.10%	0.144
	mouse-Liver-Janich2015-ZT10-rep1	37166117	19685517	52.97%	39017872	31724041	81.31%	0.152
	mouse-Liver-Janich2015-ZT10-rep2	52434821	31210099	59.52%	55505348	49464297	89.12%	0.117
	mouse-Liver-Janich2015-ZT12-rep1	33460930	9804374	29.30%	26731927	18633958	69.71%	0.18
	mouse-Liver-Janich2015-ZT12-rep2	38172695	12292098	32.20%	28619553	20649016	72.15%	0.146
	mouse-Liver-Janich2015-ZT14-rep1	35825397	9840591	27.47%	36060159	24394028	67.65%	0.17
	mouse-Liver-Janich2015-ZT14-rep2	45538631	13482278	29.61%	26523033	18048726	68.05%	0.158
	mouse-Liver-Janich2015-ZT16-rep1	35396516	10252899	28.97%	35443540	25395529	71.65%	0.184
	mouse-Liver-Janich2015-ZT16-rep2	42037128	11826736	28.13%	43388802	27004249	62.24%	0.187
	mouse-Liver-Janich2015-ZT18-rep1	86299374	12793653	14.82%	38507184	25294086	65.69%	0.157
	mouse-Liver-Janich2015-ZT18-rep2	35876901	6646869	18.53%	31862312	22969489	72.09%	0.155
	mouse-Liver-Janich2015-ZT2-rep1	42598690	17739420	41.64%	43000318	28190435	65.56%	0.238
	mouse-Liver-Janich2015-ZT2-rep2	25950739	10319057	39.76%	22971346	18986684	82.65%	0.191
	mouse-Liver-Janich2015-ZT20-rep1	71828840	14588942	20.31%	34936315	22463937	64.30%	0.154
	mouse-Liver-Janich2015-ZT20-rep2	29445291	9492932	32.24%	26463626	20030154	75.69%	0.187
	mouse-Liver-Janich2015-ZT22-rep1	34173363	20544569	60.12%	32902738	27817250	84.54%	0.113
	mouse-Liver-Janich2015-ZT22-rep2	37395859	24264540	64.89%	26708844	21980535	82.30%	0.11
	mouse-Liver-Janich2015-ZT4-rep1	44391275	21490299	48.41%	26823442	22185260	82.71%	0.219
	mouse-Liver-Janich2015-ZT4-rep2	43488989	17884880	41.13%	51169822	27586940	53.91%	0.211
	mouse-Liver-Janich2015-ZT6-rep1	53967135	16548232	30.66%	28208501	22443144	79.56%	0.223
	mouse-Liver-Janich2015-ZT6-rep2	42780292	21238154	49.64%	28090962	24442046	87.01%	0.139
	mouse-Liver-Janich2015-ZT8-rep1	47958781	22402688	46.71%	27454799	23359208	85.08%	0.158
	mouse-Liver-Janich2015-ZT8-rep2	51511890	26596414	51.63%	34218062	27547193	80.50%	0.154
	mouse-MEF-Gao2015-control	31220314	18979484	60.79%	28602028	16707971	58.42%	0.237
	mouse-Skin-Blanco2016-wt1	31890403	17409393	54.59%	11278549	4210860	37.34%	0.14
	mouse-Skin-Blanco2016-wt2	32580339	20995627	64.44%	6979962	3168705	45.40%	0.225
	mouse-Skin-Blanco2016-wt3	29481496	15779595	53.52%	7412554	1867291	25.19%	0.161
Skin	**mouse-Skin-Sendoel2017-wt-invivo-rep0**	**39183466**	**35408652**	**90.37%**	**23305836**	**11309499**	**48.53%**	**0.069**
	mouse-Skin-Sendoel2017-wt-invivo-rep1	40175494	36264195	90.26%	15586905	4858355	31.17%	0.078
Testis	**mouse-Testis-Castaneda2014-wt-a**	**89242379**	**66147271**	**74.12%**	**15237617**	**7480401**	**49.09%**	**0.11**
	mouse-Testis-Castaneda2014-wt-b	40924228	28528077	69.71%	8669490	4604910	53.12%	0.139

Notes:
Dist is a metric of the length distribution similarity between reads mapped to lncRNAs and that mapped to CDSs. See "Methods" in the main text for the details on the Dist.

Table S2 Analysis of ribosomal associations of mRNAs and lncRNAs

		mRNA			lncRNA		
Dataset(used)	FullName	Expressed	ribo-associat	%ribo-associ	Expressed	ribo-associat	%ribo-associated
Brain	human-Brain-Gonzalez2014-tumor-B	28005	26916	96.11%	7093	2024	28.54%
ES	human-ES-Werner2015-control-rep2	24460	23522	96.17%	5493	1665	30.31%
Fibroblasts	human-Fibroblasts-Shitrit2015-control	24190	23750	98.18%	5266	1979	37.58%
Foreskin	human-Foreskin-Stren2012-24hr	23605	23469	99.42%	2654	1284	48.38%
HEK293	human-HEK293-Sidrauski2015-control-b	24271	23003	94.78%	8740	2154	24.65%
HeLa	human-HeLa-Park2016-Sphase-rep1	22634	22127	97.76%	5215	1977	37.91%
Myeloma	human-Myeloma-Wiita2013-control	22919	21712	94.73%	7179	2059	28.68%
PC3	human-PC3-Hsieh2012-control-rep2	21063	20641	98.00%	3762	1676	44.55%
RPE-1	human-RPE-1-Tanenbaum2015-G1-rep1	21416	21312	99.51%	2600	1558	59.92%
U2OS	human-U2OS-Jang2015-CT20-rep1	23237	23001	98.98%	5164	2644	51.20%
Brain	mouse-Brain-Gonzalez2014-tumor-A	22096	21697	98.19%	2439	878	36.00%
EB	mouse-EB-Ingolia2011-eb	16714	16255	97.25%	2729	1095	40.12%
ES	mouse-ES-Hurt2013-control	17981	17681	98.33%	1595	740	46.39%
Fibroblasts	mouse-Fibroblast-Thoreen2012-wild-vehicle	15652	15024	95.99%	3068	799	26.04%
Hippocampi	mouse-Hippocampi-Cho2015-10min-rep3	19602	19425	99.10%	2040	1255	61.52%
Liver	mouse-Liver-Frederic2015-ZT18-D	15710	15543	98.94%	1855	711	38.33%
Skin	mouse-Skin-Sendoel2017-wt-invivo-rep0	16081	15988	99.42%	1002	638	63.67%
Testis	mouse-Testis-Castaneda2014-wt-a	18499	18214	98.46%	1895	1329	70.13%

Due to the limited space, the following files are available in figshare.com

Table S3 Ribosome association for human and mouse lncRNAs. (XLSX 1.48 MB)
- https://ndownloader.figshare.com/files/11763305

Table S4 Alignment of mass spectrometry data to human trans-lncRNAs. (TSV 9.83 kb)
- https://ndownloader.figshare.com/files/11763065

Table S5 Alignment of mass spectrometry data to mouse trans-lncRNAs. (TSV 5.97 kb)
- https://ndownloader.figshare.com/files/11763083

Table S6 Fold change values for cellular localization analysis in HeLa cells. (XLSX 154 kb)
- https://ndownloader.figshare.com/files/11763149

Table S7 Fold change values for NMD analysis in HeLa cells. (XLSX 147 kb)
- https://ndownloader.figshare.com/files/11763173

Table S8 Putative ORFs in human lncRNAs. (TSV 7.8 MB)
- https://ndownloader.figshare.com/files/11763101

Table S9 Putative ORFs in mouse lncRNAs. (TSV 5.35 MB)
- https://ndownloader.figshare.com/files/11763116

Table S10 Raw data for human and mouse. (tar.gz 14.68 MB)
- https://figshare.com/s/ffbfcff93bce633908f9

Appendix B

List of abbreviations

- lncRNA, long noncoding RNA;
- ribo-lncRNA, ribosome-associated lncRNA;
- noribo-lncRNA, ribosome-free lncRNA;
- RAI, ribosomal association index;
- *spec*, transcript expression tissue-specificity;
- NMD, nonsense-mediated decay;
- lincRNA, long intergenic noncoding RNA;
- RRS, ribosome release score;
- TS, translation score;
- ORF, open reading frame;
- CDS, coding sequence;
- RNA-seq, RNA sequencing;

- Ribo-seq, ribosome profiling;
- RPKM, reads per kilobase per total million mapped reads;
- UTR, untranslated region;
- FLOSS, fragment length organization similarity score;
- snRNA, small nuclear RNA;
- snoRNA, small nucleolar RNA;
- miRNA, microRNA;
- TEC, to be experimentally confirmed
- TIS, transcript initiation site;
- TTS, transcript termination site;
- pORF, putative primary ORF in lncRNA;
- fORF, putative first ORF in lncRNA;
- uORF, putative upstream ORF in lncRNA;
- m^6A, N6-Methyladenosine modification;
- G4, G-quadruplex;
- LTR, Long terminal repeat;
- SINE, Short interspersed nuclear element;
- LINE, Long interspersed nuclear element;
- HIV, Human Immunodeficiency Virus;

- TE, Transposon element;

- TP, number of true positives;

- TN, number of true negatives;

- FP, number of false positives;

- FN, number of false negatives;

Appendix C

List of publications

Papers

1. **Chao Zeng**, Tsukasa Fukunaga and Michiaki Hamada, Identification and analysis of ribosome-associated lncRNAs using ribosome profiling data, BMC Genomics (2018) 19:414

2. **Chao Zeng**, Michiaki Hamada, Identifying sequence features that drive ribosomal association for lncRNA, The 29th International Conference on Genome Informatics (submitted on June 1, 2018)

Oral presentation

1. **Chao Zeng**,Tsukasa Fukunaga and Michiaki Hamada, Integrative analysis of multiple ribosome profiling datasets reveals widespread lncRNA-ribosome interaction in mammals, 第6回生命医薬情報学連合大会(IIBMP2017), 2017/09, 札幌

Poster presentation

1. **曽超**, 福永津嵩, 浅井潔, 浜田道昭, Identification and classification of translational pausing and ribosome-associated lncRNAs, NGS現場の会第５回研究会, 2017/05, 仙台

2. **Chao Zeng**, Tsukasa Fukunaga and Michiaki Hamada, Ribosome profiling reveals the plurality of ribosome-associated long non-coding RNAs in mammals, 第19回日本RNA学会年会, 2017/07, 富山

3. **曽超**, 福永津嵩, 浜田道昭, リボソームプロファイリングデータを用いたリボソーム関連する長鎖ノンコーディングRNAの同定と解析, 2018/07, 大阪

Awards

1. 生命医薬情報学連合大会2017大会研究奨励賞, 第6回生命医薬情報学連合大会(IIBMP2017), 2017/09, 札幌