UNIVERSITY OF CALIFORNIA SAN DIEGO

Development of high-throughput technologies to map RNA structures and interactions

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy

in

Bioengineering

by

Tri C Nguyen

Committee in charge:

Professor Sheng Zhong, Chair
Professor Kun Zhang
Professor Gene Yeo
Professor Liangfang Zhang
Professor Bing Ren

2018

The Dissertation of Tri C. Nguyen is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____

Chair

University of California San Diego

2018

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I'd first like the thank Dr. Sheng Zhong for being a supportive mentor and excellent role model, scientifically and otherwise. I'd like to thank my other faculty mentors including Dr. Stephanie Ceman at University of Illinois at Urbana – Champaign and my thesis committee.

My sincerest thanks to all current and past members of the Zhong lab for their generous advice and friendship. I would also like to acknowledge all the co-authors for all of my publications, as the work presented here was truly a team effort. They all provided me with technical knowledge, experiment design feedback, and mental support.

Chapter 1, in full, is an adaptation of materials that appears in Nguyen, Tri C.; Zaleta-Rivera, Kathia; Huang, Xuerui; Dai, Xiaofeng; Zhong, Sheng. RNA, action through interactions. Trends in Genetics, 34(11):867-882, 2018. The dissertation author was the primary investigator and author of this material.

Chapter 2, in full, is an adaptation of materials that appears in Nguyen, Tri C., Cao, Xiaoyi, Yu, Pengfei, Xiao, Shu, Lu, Jia, Biase, Fernando H., Sridhar, Bharat, Huang, Norman, Zhang, Kang, Zhong, Sheng. "Mapping RNA-RNA interactome and RNA structure in vivo by MARIO". Nature Communications, 7:12023, 2016. The dissertation author was the primary investigator and author of this material.

Chapter 3, in full, is an adaptation of materials that appears in Sridhar, Bharat*, Rivas-Astroza, Marcelo*, Nguyen, Tri C.*, Chen, Weizhong, Yan, Zhangming, Cao, Xiaoyi, Hebert, Lucie, Zhong, Sheng. "Systematic mapping of RNA-chromatin interactions in vivo". Current

Biology, 27(4):602–609, 2017. (* co-first author). The dissertation author was one of the primary investigators and authors of this material.

Chapter 4, in full, is currently being prepared for submission for publication of the material. Nguyen, Tri C.; Zaleta-Rivera, Kathia; Hebert, Lucie; Huang, Norman; Zhong, Sheng. The dissertation author was one of the primary investigators and authors of this material.

VITA

2009	Bachelor of Engineering in Bioengineering, Nanyang Technological University
2015	Master of Science, University of California San Diego
2017	Candidate of Philosophy, University of California, San Diego
2018	Doctor of Philosophy, University of California San Diego

ABSTRACT OF THE DISSERTATION

Development of high-throughput technologies to map RNA structures and interactions

by

Tri C. Nguyen

Doctor of Philosophy in Bioengineering

University of California San Diego, 2018

Professor Sheng Zhong, Chair

At any time hundreds of thousands of macromolecular interactions occur in a cell, mediating functions that maintain normal cellular activities. A fundamental problem in molecular biology is to catalog these interactions and to decipher their functional consequences. High throughput sequencing has made it possible to characterize some of these interactions rapidly, at high-resolution, and *in vivo* (e.g., protein-DNA binding via ChIP-Seq and protein-RNA binding via CLIP-Seq). But many interactions are not susceptible to these methods (e.g., RNA- RNA complexes, ncRNA-DNA binding, protein-protein interactions).

This thesis aims to address this gap by coupling high-throughput sequencing with proximity-ligation-based methods. In proximity ligation, spatially proximate nucleic acids ligate to one another, forming a chimeric oligonucleotide. Observation of a chimera composed of X and Y suggests that X and Y must have been near one another in the original sample. As a result, questions about spatial arrangement become questions about sequence composition, making it possible to take advantage of high-throughput sequencing. Using this general approach, we developed high-throughput technologies to study RNA interactions with different types of molecular partners: RNA, chromatin, and lipid.

In Chapter 1, I review and discuss many different high-throughput techniques to map RNA structure and RNA-RNA as well as RNA-chromatin interactions. I also provide the biological insight that can be gained from the type of data generated by the new technologies.

In Chapter 2, I describe the development of MARIO, a technology to map RNA-RNA interactions. This method produces a global map of RNA-RNA interactome and RNA structures in vivo. The information will provide roadmaps to systems level understanding of cellular regulations through RNA-RNA interactions.

In Chapter 3, I describe the development of MARGI, a technology to map RNA-chromatin interactions. Mapping RNA-chromatin interactions identify RNAs that bind to the chromatin and shed light on the functional roles of chromatin associated RNAs in gene regulations.

In Chapter 4, I describe the discovery of a new class of RNA, named cell surface RNA. These are RNAs that are expressed outside of the cell membrane. I also develop methods to isolate and identify these surface RNAs and protein binding partners of the RNAs.

CHAPTER 1 – Introduction: Mapping genome-wide RNA structures and interactions

Abstract

As transcription of the human genome is quite pervasive it is possible that many novel functions of the noncoding genome have yet to be identified. Often the noncoding genome's functions are carried out by their RNA transcripts which may rely on their structures and/or extensive interactions with other molecules. Recent technology developments are transforming the fields of RNA biology from studying one-RNA-at-a-time to transcriptome-wide mapping of structures and interactions. Here, we highlight the recent advances in transcriptome-wide RNA interaction analysis. These technologies revealed surprising versatility of RNA to participate in diverse molecular systems. For example, tens of thousands of RNA-RNA interactions have been revealed in cultured cells as well as in mouse brain, including interactions between transposon-produced transcripts and mRNAs. Additionally, most transcription start sites in the human genome are associated with noncoding RNA transcribed from other genomic loci. These recent discoveries expanded our understanding of RNAs' roles in chromatin organization, gene regulation, and intracellular signaling.

RNA interactions regulate diverse molecular functions

RNA is produced from most human genomic sequences, although only a relatively small portion of these transcripts are translated and/or have known associated functions. The vast amounts of transcripts with unknown functions may not be translated and present an opportunity

to investigate the functions of the noncoding genome. Previous studies of RNA-RNA interactions have uncovered essential functions for these RNAs. These include the discoveries of RNA interference by studying miRNA-mRNA interactions and siRNA-mRNA interactions (Fire et al., 1998), the essential steps of RNA splicing through snRNA binding to intronic splice sites, and site-specific rRNA pseudouridination through snoRNA-rRNA interactions. Studies of RNA-chromatin interactions facilitated the discoveries of chromosome silencing mechanisms, RNAi mediated epigenetic inheritance, and transcriptional activations mediated by miRNA-promoter, and lncRNA-promoter interactions. We therefore anticipate novel functions to be revealed by identifying novel classes of RNA-participating interactions.

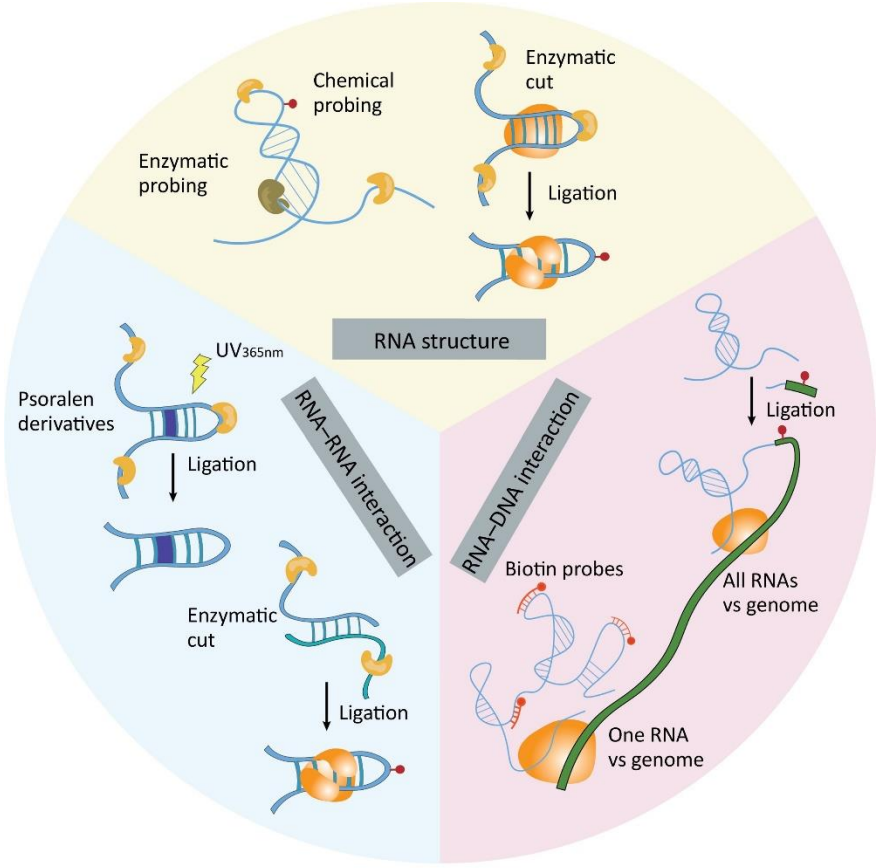


Figure 1.1*: Overview of Sequencing-Based Technologies for Mapping RNA Structures*, RNA–RNA Interactions, and RNA–DNA Interactions

Recent technology developments are transforming the analysis of RNA structure and interactions. Instead of studying one RNA or one interaction at-a-time, recent technologies enabled transcriptome-wide analysis of RNA structures, RNA-RNA interactions, and RNA-DNA interactions (Figure 1.1). These developments were achieved by combining biochemical reactions with next-generation sequencing. The general strategy of probing RNA structure is to utilize chemical and enzymes that cleave or modify at either single- or double-stranded regions and leverage sequencing to reveal these chemical- or enzyme-processed regions. The general strategy of investigating RNA-RNA and RNA-DNA interactions was to convert interacting sequence (RNA-RNA or RNA-DNA) pairs into chimeric DNA, and leverage DNA sequencing as a high-throughput readout of the underlying interactions. Many of these technologies are applicable to analyze intact cells and primary tissues without requiring genetic perturbation or ectopic expression. In this article, we review sequencing-based approaches for mapping RNA structures, RNA-RNA and RNA-DNA interactions, summarize the major findings, and point out the new hypotheses derived from these findings.

RNA Structure

Sequencing-based methods for mapping RNA structures

The flexibility of RNA provides a physical basis for forming a diverse array of secondary and tertiary structures. The structures of RNA and their interactions with other molecules are modulated by physiochemical environment (Downey et al., 2007; Egli et al., 2005; Higgs, 2000; Kolkenbeck and Zundel, 1975; Ribitsch et al., 1985), RNA sequences, and posttranscriptional modifications (Allen and Noller, 1989; H. et al., 2004). A general strategy employed in systematic mapping of RNA structures is to leverage enzymes or chemicals that specifically react with certain local structures. These reactions include RNA cleavage or

modification. The cleaved or modified sites could then be systematically revealed by sequencing. We have classified the sequencing-based RNA structure analysis methods: 1) by reagents, into enzyme-based and chemical-based approaches (columns, Figure 1.2, Panel A), and 2) by application scenarios, into *in vitro* and *in vivo* approaches (rows, Figure 1.2, Panel A). Briefly, enzyme-based *in vitro* RNA structure analysis methods include PIP-seq (Silverman and Gregory, 2015; Silverman et al., 2014), PARS (Kertesz et al., 2010), PARTE (Wan et al., 2012), FragSeq (Underwood et al., 2010). Chemical-based *in vitro* methods include DMS-seq (Rouskin et al., 2014), icSHAPE (Spitale et al., 2015), Structure-seq (Ding et al., 2014), and Mod-seq (Talkish et al., 2014). Chemical-based *in vivo* methods include CIRS-seq (Incarnato et al., 2014) and SHAPE-MaP (Siegfried et al., 2014). Finally, MARIO is an enzyme-based analysis method that in theory captures *in vivo* structures (Nguyen et al., 2016). In addition to revealing the single-stranded regions, MARIO also identifies all the spatially proximal regions of an RNA molecule, thus providing unique information about the secondary and tertiary structures. We selected representative methods to describe their major experimental steps (Figure 1.2, Panels B-E).

The enzyme-based approaches leverage different ribonucleases (RNases) based on their selectivity in cutting either single-stranded or double-stranded regions. The resulting mixture of RNA fragments when analyzed by sequencing, allows for assessment of nucleotide accessibility and base-pairing regions, and thereby inference of secondary structures. The commonly used RNases with predefined structural preferences include RNase V1, which is dsRNA-specific, but the specificity is not absolute (Lowman and Draper, 1986). Whereas RNase S1 (process all four nucleotides), RNase P1 (process all four nucleotides), RNase A (ssC/U-specific) and RNase T1 (ssG-specific) are ssRNA-specific, but these enzymes may miss small bulges, loops, or

mismatches (Underwood et al., 2010). Thus, integration of the sequencing data obtained from treatments of different RNases may generate more complete mapping of single- and double-stranded regions. A limitation of enzyme-based methods is that the applications are often limited to *in vitro* structural analysis. This is in part due to the large sizes of RNases (>10 kDa) and hence the difficulty of crossing cell membranes and susceptibility to steric hindrance in the presence of bound proteins or other RNA-associated macromolecules.

Chemical-based methods utilize small molecules (<500 Da) to probe RNA structure. These membrane permeable molecules are utilized for *in vivo* analyses of RNA structures, which often achieve single nucleotide resolution. Frequently used chemicals include nucleobase-specific chemicals, carbodiimide modifying reagents, and ribose-specific probes. Nucleobase-specific chemicals including dimethyl sulfate (DMS) can modify the functional groups on the Watson-Crick (WC) face of the base. DMS alkylates the unprotected N1 position of adenine (N_1A), unprotected N3 position of cytosine (N_3C), and unprotected N7 position of guanine (N_7G) (Kwok, 2016). Carbodiimide modifying reagents react with guanosine and uridine. These chemicals detect the presence of base-paired regions, allowing for mapping of the secondary structures and protein binding sites. Ribose-specific probes acylate the flexible C2'-hydroxyl group of the ribose (C2'-OH). Using such a probe, Selective 2'-Hydroxyl acylation Analyzed by Primer Extension (SHAPE) resolves local structural environment at nucleotide resolution (Watts et al., 2009). Flexible bases exhibit a higher tendency to adapt to specific local structural environments, which facilitates acetylation, resulting in higher SHAPE activity (McGinnis et al., 2012). An advantage of SHAPE reagents over nucleobase-specific probes lies in their capability of targeting the ribose of all four nucleotides. Ideally, combining the sequencing data generated from treatments of

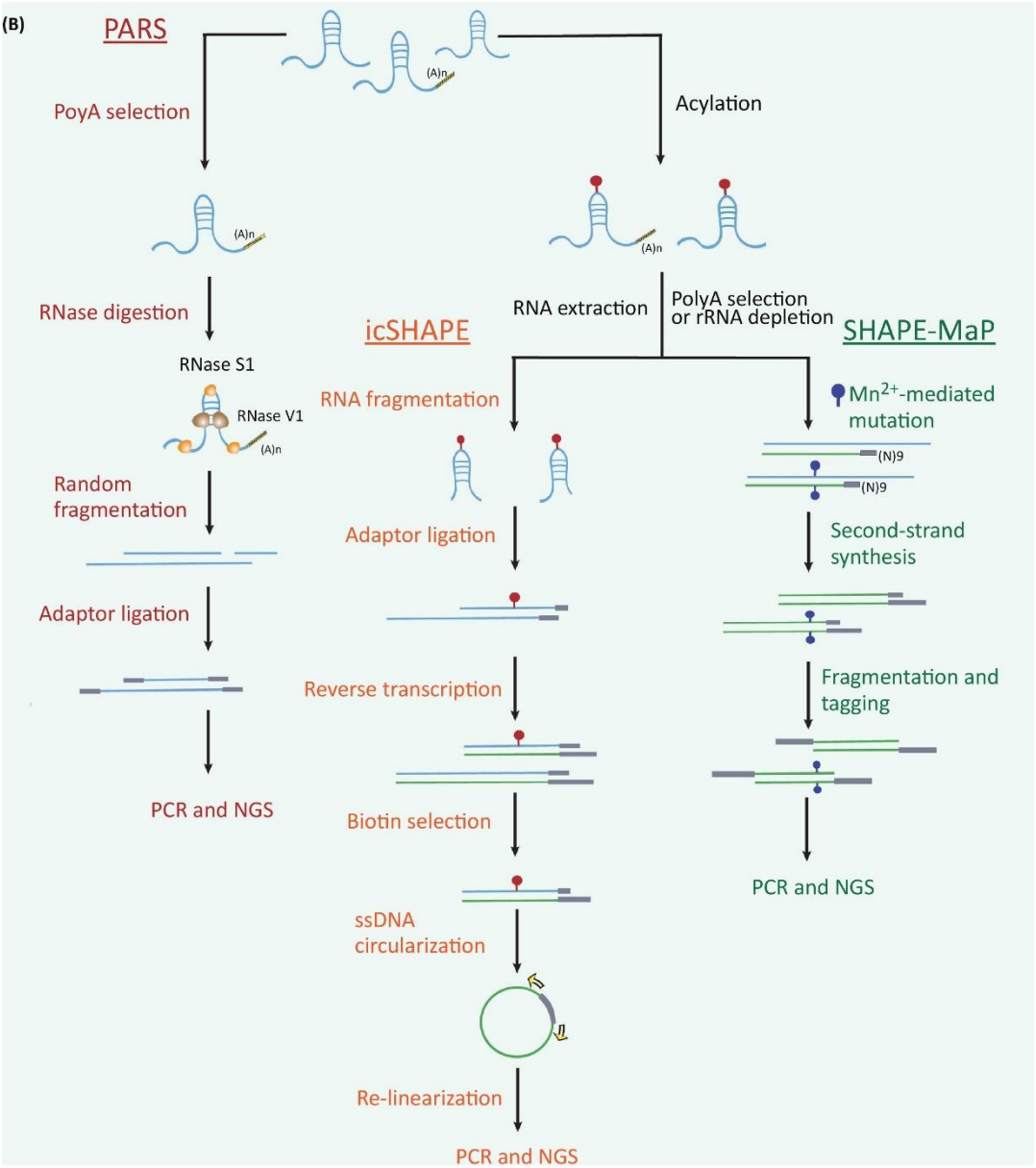
multiple chemicals and enzymes may release the complimentary advantages of these methods and potentially reveal more comprehensive structural information.

Figure 1.2: ***Sequencing-Based Technologies for Mapping RNA Structures***. (A) Summary of enzyme- and chemical-based RNA structure technologies (columns) and their application domains (rows). Selected technologies (underlined) are expanded in detail in B. (B) Major steps of selected technologies. In PARS, polyA-tailed RNA is selected and divided into two pools. One pool is treated with RNase S1 that cleaves single-stranded sequence, and the other pool is treated with RNase V1 that cuts at double-stranded regions. The produced RNA segments are subjected to random fragmentation and converted into a sequencing library. In icSHAPE, cells are treated with NAI–azide, allowing for attaching a biotin moiety through copper-free CLICK reactions. SHAPE-reacted RNA segments are enriched by streptavidin–biotin interaction and are subsequently converted into a sequencing library. In SHAPE-MaP, RNA is treated with 1M7 and is reverse transcribed in a reaction mixture that induces mutation at SHAPE-reacted sites. Abbreviations: CIRS-seq, chemical inference of RNA structures sequencing; DMS-seq, dimethyl sulfate sequencing; FragSeq, fragmentation sequencing; icSHAPE, in vivo click selective 2-hydroxyl acylation and profiling experiment; MARIO, mapping RNA interactome and structure in vivo; Mod-seq, high-throughput sequencing for chemical probing of RNA structure; NAI-azide, 2-methylnicotinic acid imidazoline-azide; NAI, 2-methylnicotinic acid imidazoline; NGS, next-generation sequencing; PARS, parallel analysis of RNA structure; PARTE, parallel analysis of RNA structures with temperature elevation; PIP-seq, protein interaction profile sequencing; RNase, ribonuclease; SHAPE, selective 2'-hydroxyl acylation analyzed by primer extension; SHAPE-MaP, selective 2'-hydroxyl acylation analyzed by primer extension and mutational profiling; ssDNA, single-stranded DNA.

(A)

	Enzyme based				Chemical based			
In vitro	PARS	PIP-seq	PARTE	FragSeq	SHAPE-MaP		CIRS-seq	
In vivo	MARIO				icSHAPE	Structure-seq	DMS-seq	Mod-seq

(B)



Toward understanding sequence and environmental determinants of RNA structures

Genome-wide mapping of RNA structures has been completed for the HIV-1 RNA genome (Watts et al., 2009), yeast (Kertesz et al., 2010), *Escherichia coli* (Del Campo et al., 2015), *Arabidopsis* (Ding et al., 2014), *Drosophila* (Li et al., 2012), *Caenorhabditis elegans* (Li et al., 2012), and selected cell types in mouse (Incarnato et al., 2014; Spitale et al., 2015) and human (Rouskin et al., 2014; Silverman et al., 2014; Wan et al., 2014). These genome-wide analyses offered insights to sequence and environmental determinants of RNA structures.

Sequence motifs and grammar have been searched for. A triplet repeat pattern emerged from both *in vitro* (Del Campo et al., 2015) and *in vivo* (Ding et al., 2014) experiments. This repeat pattern in chemical/enzyme reactivity is indicative of existence of a sequence grammar for RNA structure (Bevilacqua et al., 2016; Del Campo et al., 2015; Ding et al., 2014; Incarnato et al., 2014; Kertesz et al., 2010; Shabalina et al., 2006; Wan et al., 2014). Furthermore, single-nucleotide polymorphisms (SNP) were found to correlate with variations in RNA structures (Wan et al., 2014). Wan et al. discovered thousands of riboSNitches (SNP-mediated RNA structure switch) in healthy human parent–offspring trios, thus connected personal genomic variation with RNA structural differences (Wan et al., 2014).

Diverse environmental factors can modulate RNA structures (Downey et al., 2007; Egli et al., 2005; Higgs, 2000; Kolkenbeck and Zundel, 1975; Ribitsch et al., 1985). *In vitro*, RNA generally appears more structured than *in vivo* (Ding et al., 2014; Rouskin et al., 2014; Spitale et al., 2015), which is partially attributable to different Mg^{2+} concentrations (Rouskin et al., 2014) and accessibility to RNA-binding proteins. By quantifying the reactivity differences between *in vivo* and *in vitro* conditions (Spitale et al., 2015) and between *in cellulo* and *ex vivo* conditions

(Smola et al., 2015), two teams were able to reveal protein-bound RNA regions. By adding cross-linking and proximity ligation steps, the MARIO team identified a case of protein-assisted RNA folding (Nguyen et al., 2016). It remains a challenge to integrate RNA sequence and cellular context for deriving the most compatible structure from diverse types of structure probing assays.

RNA-RNA interactions

Sequencing-based methods for mapping RNA-RNA interactions

Methods for analysis of intermolecular RNA-RNA interactions were restricted to targeting a specific RNA that participates in RNA-RNA interactions, until Tollervey et al. discovered chimeric RNAs can be extracted from RNA sequencing data (Kudla et al., 2011). Although these chimeric RNAs were present in low frequencies, they could represent pairs of interacting RNAs (Bohnsack et al., 2012). Two subsequent methods, CLASH (Helwak et al., 2018) and hiCLIP (Sugimoto et al., 2015) enriched for the interacting RNAs by purifying a specific protein that is required for such interactions. The major difference between these two methods lies in utility of ectopic expression of a tagged protein (CLASH) versus antibody-based isolation of the protein of interest from unperturbed cells (hiCLIP). CLASH and hiCLIP broke the barrier of having to target a specific RNA in identifying RNA-RNA interactions. These technologies enabled identification of RNA interactions mediated by a specific protein.

Genome-wide RNA interactome analysis was enabled by a cohort of four methods, including psoralen analysis of RNA interactions and structures (PARIS) (Lu et al., 2016), sequencing of psoralen-crosslinked, ligated, and selected hybrids (SPLASH) (Aw et al., 2016), ligation of

interacting RNA followed by high-throughput sequencing (LIGR-seq) (Sharma et al., 2016), and mapping RNA interactome *in vivo* (MARIO) (Nguyen et al., 2016) (Figure 1.3). The central idea of these technologies is to leverage proximity ligation to produce chimeric sequences. All methods used *in vivo* crosslinking of RNA, either by UV-mediated RNA-protein crosslinking (MARIO), or RNA duplex crosslinking enabled by psoralen derivatives (PARIS, SPLASH, LIGR-seq), followed by RNA fragmentation to produce single-stranded RNA ends, which were subjected to intramolecular ligation and reverse-crosslinking to convert into a sequencing library. The different choice of crosslinking reagents led to revelation of several types of RNA interactions. PARIS, SPLASH and LIGR-Seq used psoralen or its derivatives including 4'-aminomethyltrioxsalen (AMT) and biotinylated psoralen, which intercalate in RNA helices and undergo interstranded cross-link upon 365 nm UV irradiation. MARIO crosslinked RNAs with proteins and ligated the RNAs bound by the same protein molecules. PARIS, SPLASH, LIGR-seq were designed for identifying hybridized RNA pairs, whereas MARIO was designed for identifying all RNA pairs brought together by any protein without requiring RNA-RNA hybridization.

RNA interactome as a scale-free network

Lack of specificity was considered a theme in miRNA interaction with its target mRNAs. This phenomenon was also referred to as promiscuity in miRNA targeting. The promiscuity was supported by many complementary sequences in the transcriptome, as well as changes in transcript abundances when the endogenous concentration of a miRNA was perturbed (Bartel, 2004; Chi et al., 2009; Du and Zamore, 2007). However, when applied to unperturbed cells, none of the four genome-wide assays (PARIS, SPLASHs, LIGR-seq, MARIO) reported many targets for most of the miRNAs. Instead, in embryonic stem cells and in mouse brain, most of the miRNAs exhibited

only 1 to 3 mRNA targets (Nguyen et al., 2016). Only a handful of miRNAs exhibited more than 10 mRNA targets. In addition, most of lincRNAs also appeared to each target only one or a few mRNAs. More generally, the MARIO authors found that RNA interactome follows the power-law and is a scale-free network (Nguyen et al., 2016). Nearly all other molecular networks being studied were reported to be scale-free (Barabasi and Oltvai, 2004; Oltvai and Barabasi, 2002), whereas the promiscuity of miRNA-involved interactions would argue against the scale-free property in an RNA interactome. However, the genome-wide data derived from recent technologies suggested that in endogenous cellular conditions, RNA interactome does not appear to be an exception to the power-law, a physics rule of biological networks (Barabasi and Oltvai, 2004).

Sno-miR: A new gene repertoire of regulatory RNAs

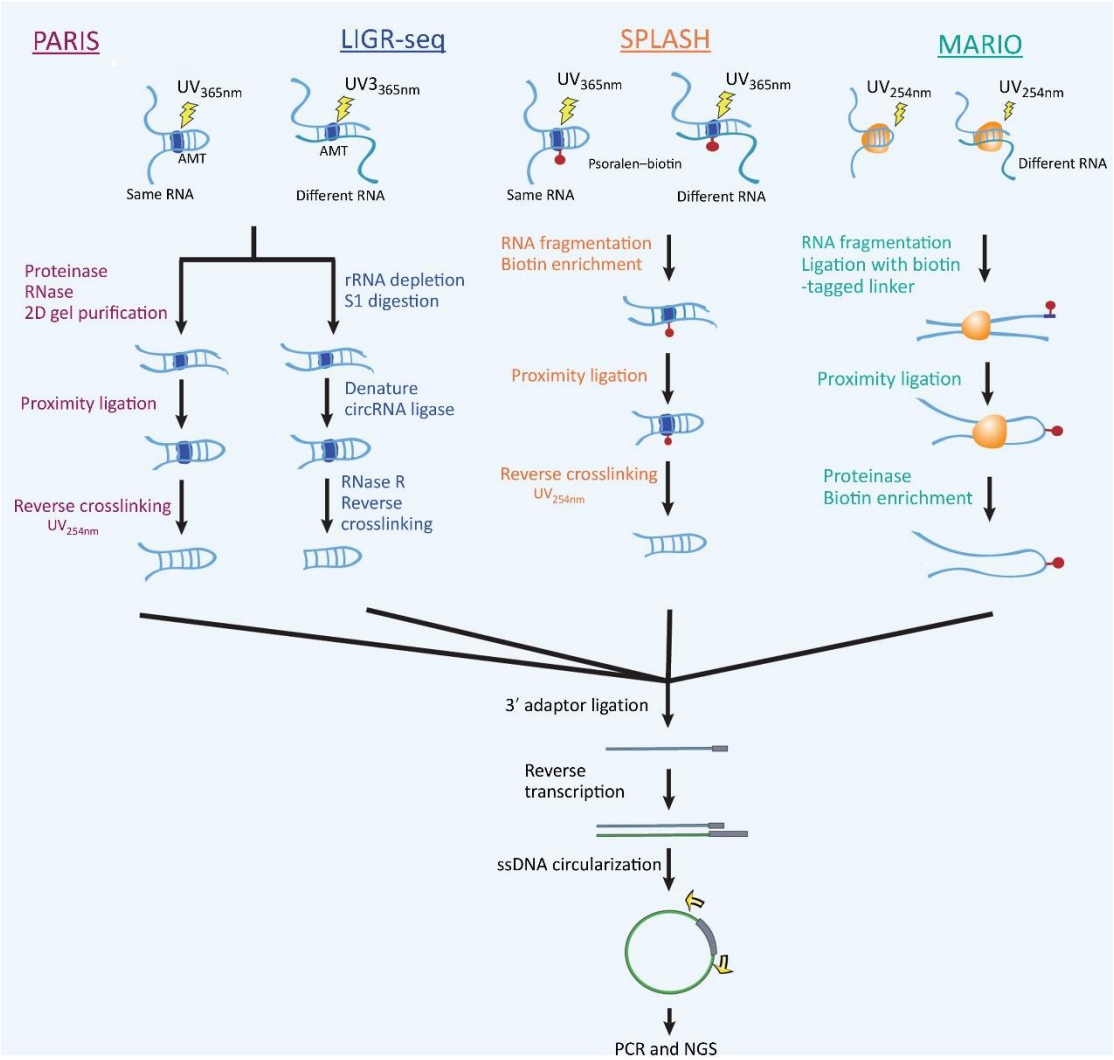
Abundant interactions between snoRNAs and mRNAs were reported from all four assays (PARIS, SPLASH, LIGR-seq, MARIO). The identified snoRNA interaction sites on mRNAs were enriched with pseudouridylation sites (Nguyen et al., 2016), consistent with the contribution of snoRNAs to the pseudouridylation process. However, many identified interactions involved truncated forms of snoRNAs rather than the entire snoRNAs (Nguyen et al., 2016). These truncation forms were present in the cells, as revealed by small RNA sequencing, and were bound by AGO2 as revealed by CLIP-seq (Nguyen et al., 2016). Taken together, more than 170 snoRNA genes appeared to produce miRNA-like RNAs, which interact with mRNAs. The snoRNA-originated miRNA-like RNAs (sno-miR) could be a new repertoire of regulatory RNAs. Indeed, one of the human snoRNAs is processed by DICER and mediate mRNA silencing through AGO1 and AGO2 (Ender et al., 2008).

Figure 1.3: ***Sequencing-Based Technologies for Mapping RNA–RNA Interactions***. (A) Summary of antibody-based methods that analyze interactions mediated by a specific protein (left column) and genome-wide methods without targeting any specific proteins (right column). Selected technologies (underlined) are expanded in B. (B) Major steps of selected technologies. In PARIS, double-stranded RNA regions are crosslinked by AMT and UV. RNA is purified and subjected to proximity ligation. The resulting RNA is ligated with a 3' adaptor and converted into a sequencing library. SPLASH procedure is similar to PARIS, except that instead of AMT, biotinylated psoralen is used as the crosslinking reagent, which allows for enrichment of double-stranded regions. LIGR-seq used a similar experimental strategy, with different choices of RNA purification, treatment, and ligation steps. In MARIO, RNA–protein complexes are crosslinked by UV. RNA is randomly fragmented and ligated with a biotinylated linker sequence and then subjected to proximity ligation. The resulting RNA–linker–RNA chimeric sequences are purified by streptavidin–biotin interaction and converted into a sequencing library. Abbreviations: AMT, 4'-aminomethyl trioxsalen; circRNA, circular RNA; CLASH, crosslinking, ligation, and sequencing of hybrids; hiCLIP, RNA hybrid and individual-nucleotide resolution UV crosslinking and immunoprecipitation; LIGR-seq, ligation of interacting RNA followed by high-throughput sequencing; MARIO, mapping RNA interactome and structure in vivo; NGS, next-generation sequencing; PARIS, psoralen analysis of RNA interactions and structures; RNase, ribonuclease; SPLASH, sequencing of psoralen-crosslinked, ligated, and selected hybrids; ssDNA, single-stranded DNA.

(A)

Mediated by a specific protein		Genome-wide			
CLASH	hiCLIP	PARIS	SPLASH	LIGR-seq	MARIO

(B)



Connections between mRNA-mRNA interaction and translational regulation

Many mRNA-mRNA interactions were identified by all four assays (PARIS, SPLASH, LIGR-seq, and MARIO) (Bartel, 2004; Chi et al., 2009; Du and Zamore, 2007). In humans and mice, approximately 1,000 mRNA pairs interact by base pairing (Aw et al., 2016), and more than 5,000 mRNA pairs were brought together by the same protein (Nguyen et al., 2016). Base complementation was significant even in MARIO identified mRNA-mRNA interactions, where the experimental procedure did not select for base paired RNA pairs (Nguyen et al., 2016). Interactions at sites near start codons negatively correlated with translation efficiency, whereas intra-molecular interactions of the two ends of mRNA molecules positively correlated translation efficiency, suggesting a link between RNA interaction and translational control (Aw et al., 2016). Furthermore, interacting mRNA pairs tended to encode for proteins that co-localize to the same subcellular compartments and sometimes exhibited similar translation efficiencies or RNA decay rates, suggesting mRNA-mRNA interaction as a means of co-regulation of gene expression (Aw et al., 2016).

Pseudogenes and transposons produce RNAs that interact with mRNA

Large numbers of transcripts produced from pseudogenes and transposons were reported to interact with mRNAs (Nguyen et al., 2016). Pseudogene RNAs interacted with both exonic and intronic regions of mRNAs. Both pseudogene-exon and pseudogene-intron interactions exhibited significant base pairing (Nguyen et al., 2016). Significant base pairing was also observed in interacting LINE_RNA-mRNA pairs and LTR_RNA-mRNA pairs (Nguyen et al., 2016). The interaction sites on pseudogene RNAs and mRNAs exhibited increased interspecies conservation

levels than other parts of the pseudogenes and mRNAs, suggesting the pseudogene_RNA-mRNA interactions were evolutionarily selected for (Nguyen et al., 2016). These novel interactions indicate a subset of pseudogenes and transposons may function by providing mRNA-interacting transcripts.

RNA–DNA Interactions

Sequencing-based methods for mapping RNA-DNA interactions

Earlier technology developments focused mapping genome-wide locations of a specific RNA (one RNA versus the genome, Figure 1.4A), including ChIRP-seq (Chu et al., 2011), CHART-seq (Simon et al., 2011) and RAP-seq (Engreitz et al., 2013). These technologies utilize biotinylated complementary oligonucleotides to pull down a specific target RNA together with its binding partners. The identities of its DNA- or protein-binding partners are subsequently revealed by sequencing or mass spectrometry. A more recent cohort of technologies enabled mapping (possibly) all chromatin-interacting RNAs together with each RNA's genomic interacting regions (all RNAs versus the genome, Figure 1.4A), including MARGI (Sridhar et al., 2017), ChAR-seq (Bell et al., 2018) and GRID-seq (Li et al., 2017). These methods leverage proximity ligation to convert RNA and its proximal DNA sequence into a chimeric sequence that can be read out by sequencing. A major advantage of these ligation-based methods is their capability of discovering *de novo* chromatin-associated RNAs.

In all of these techniques, cells are first subjected to cross-linking reagents to preserve protein-nucleic acid interactions. ChIRP-seq, CHART-seq and RAP-seq focus on capturing

chromatin interactions of individual RNAs. All use synthetic biotinylated antisense DNA oligonucleotides designed specifically to capture and purify lncRNA–chromatin complexes from the cells. Due to the inherent stickiness of RNA and propelled by the need to maximize both specificity and recovery of the RNA of interest, chemical crosslinking is used to allow stringent manipulations of pulldown experiments. Crosslinking coupled with sonication as well as denaturing washing conditions are to ensure that non-physiological bindings formed *in vitro* upon cell lysis are removed. ChIRP-Seq, CHART-Seq and RAP-Seq are very similar in overall approach with differences in specific crosslinking reagents, strength of chromatin shearing, strength of washing buffer, density and length of antisense probes. Without much prior knowledge about the local structures of RNA such as folding, interacting proteins, it is difficult to design only a few probes that ensure consistent performance of every pulldown experiments. Taking this consideration into account, ChIRP and RAP do not rely on any knowledge of the RNA of target. Instead, tiling probes that are spaced across the entire RNA are used. This design maximized the chances of capturing the entire length of fragmented RNAs, which is usually sheared in advance into smaller species. Chromatin shearing by sonication, however brief, is almost always required to efficiently lyse chemically crosslinked cells. ChIRP uses substantial sonication to fragment RNA into hundreds of nucleotides length. On the other hand, RAP only employs brief sonication to solubilize the chromatin while keeping the target RNA as long as possible. ChIRP uses 20-mer probes that are cheaper while RAP uses much longer 120-mer probes which can be cost-inhibitive to synthesize or cumbersome to prepare by *in vitro* transcription (Chu et al., 2015). A variation of ChIRP, domain ChIRP (dChIRP) (Quinn et al., 2014) designs probe sets by iteratively finding the minimal set of probes targeting the chromatin-interacting region of a RNA, which can result in higher signal-to-noise ratio.

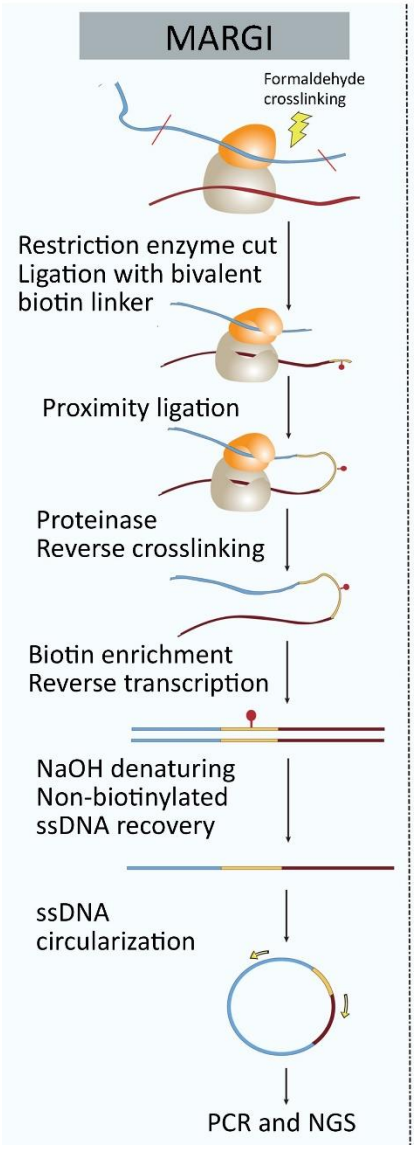
In all-RNA-versus-genome approaches (MARGI (Sridhar et al., 2017), ChAR-Seq (Bell et al., 2018), GRID-seq (Li et al., 2017)), a bivalent, and biotinylated linker comprising both single-stranded RNA at one end and double-stranded DNA at another end is used to link RNA to DNA by proximity ligation. The RNA end of the linker is first ligated to RNA molecule. Next, the DNA end of the linker is proximity-ligated to the DNA molecule. The biotinylated linker enables enrichment of desirable chimeric RNA-DNA segments. The procedure is followed by amplification and deep sequencing to detect the RNA-DNA interactions. In MARGI, the linker ligation and proximity ligation were performed on RNA-DNA protein complexes that are tethered to the solid surface of streptavidin beads (Sridhar et al., 2017). In contrast, ChAR-Seq and GRID-seq performed those steps in intact nuclei (Bell et al., 2018; Li et al., 2017).

Figure 1.4: ***Sequencing-Based Technologies for Mapping RNA–DNA Interactions***. (A) Summary of technologies for RNA–DNA interactions based on a specific RNA (left column) or any RNA (right column). Selected technologies (underlined) are expanded in detail in B–D. (B–D) Major steps of selected technologies. (B) In MARGI, protein–RNA–DNA complexes are crosslinked by formaldehyde. DNA is fragmented. RNA is ligated with the RNA end of a biotinylated half-RNA–half-DNA linker, and the DNA end of this linker is subsequently ligated to DNA through proximity ligation. The resulting chimeric RNA–DNA sequences are selected by streptavidin–biotin interactions and converted into a sequencing library. (C and D) The ChAR-seq and GRID-seq procedures are similar to MARGI. The major difference is that many steps are conducted in intact nuclei, including restriction enzyme digestion, RNA-linker ligation, and proximity ligation. Abbreviations: ChAR-seq, chromatin-associated RNA sequencing; CHART-seq, capture hybridization analysis of RNA targets sequencing; ChIRP-seq, chromatin isolation by RNA purification sequencing; GRID-seq, mapping global RNA interactions with DNA by deep sequencing; MARGI, mapping RNA–genome interactions; NGS, next-generation sequencing; RAP-seq, RNA antisense purification sequencing; ssDNA, single-stranded DNA.

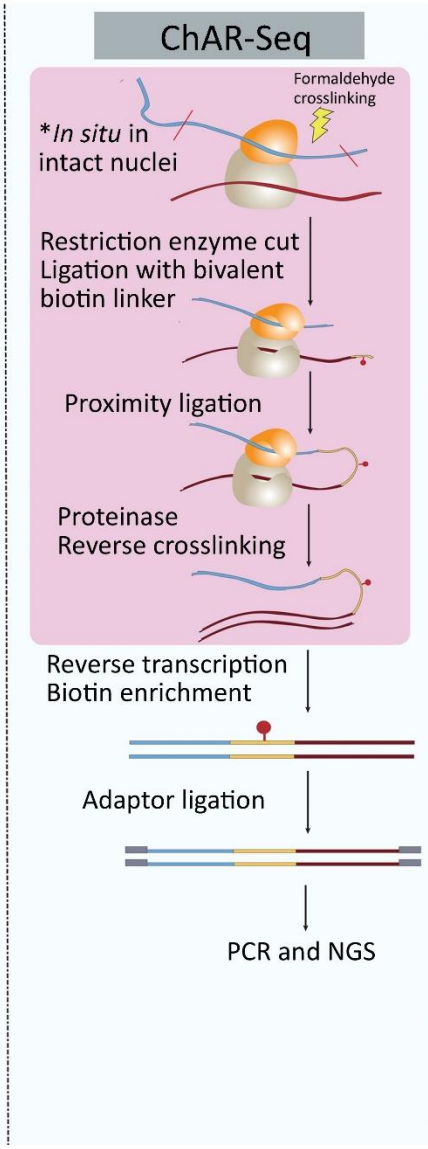
(A)

One RNA versus the genome			All RNAs versus the genome		
ChIRP-seq	CHART-seq	RAP-seq	MARGI	ChAR-seq	GRID-seq

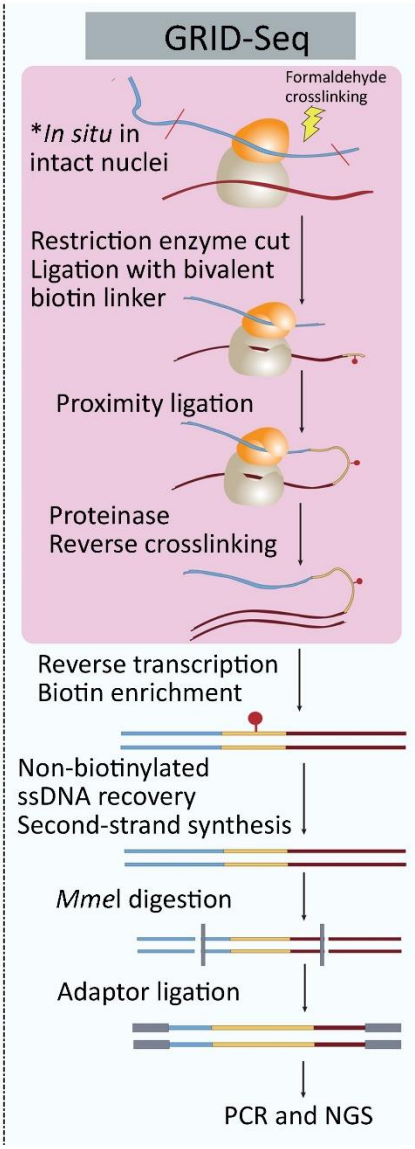
(B)



(C)



(D)



Diverse modes of RNA-chromatin interactions

Several modes of RNA-chromatin interactions have been revealed. Some lincRNAs bind to chromatin in a localized fashion reminiscent of transcription factor binding while others spread wider in binding locations (Chu et al., 2011; Engreitz et al., 2013; Simon et al., 2011, 2013). lincRNAs can interact in cis near their sites of transcription or in trans to regions on different chromosomes (Chu et al., 2011; Colak et al., 2014; Liu et al., 2013; Miao et al., 2018). Some lncRNAs only interact with a few genomic loci while others interact promiscuously with multiple genomic regions. Some lncRNAs act as repressors, whereas others function as activators of gene expression. Some lncRNAs, Xist for example, spread their binding regions by using the three-dimensional organization of the chromatin to spread to spatially proximal genomic loci. It remains to explore whether there are any rules or principles that explain how different RNAs would interact with the genome.

RNA-DNA interaction on transcription start sites: a genome-wide phenomenon

An overriding theme emerged from the genome perspective. That is nearly all promoters (Li et al., 2017) or more specifically nearly all transcription start sites (TSS) (Sridhar et al., 2017) are associated with trans-interacting RNAs. Even the TSSs of silent genes are attached with RNAs. However, the amount of TSS-associated RNAs exhibited weak correlation with the expression level of TSS-specified genes (Sridhar et al., 2017), suggesting RNA attachment on TSS may promote transcription. Consistent with this idea, a case study of TSS-associated antisense RNA suggested such an interaction promotes the transcription of the TSS-specified gene (Colak et al., 2014).

Enhancer-promoter interaction offers a possible explanation to the large amounts of TSS-RNA interactions. In this model, transcripts produced from enhancers can associate with promoters either as a result of or as a contributing cause to enhancer-promoter interactions. Conversely, many enhancers were found associated with the transcripts of their supposedly regulating genes (Li et al., 2017). However, enhancer-promoter interactions cannot completely explain the symmetric pattern of RNA attachment, centering at each TSS (Sridhar et al., 2017). It remains to be tested if RNA attachment on TSSs is a molecular mechanism for specifying the start sites of transcription. After all, unlike prokaryotes that have characteristic binding sites including TATA and CAT boxes located at defined distances which could help the transcription machinery to pinpoint the locations to initiate transcription, the vertebrates do not seem to have a comparable mechanism that would allow for precise TSS recognition.

RNA decoration on chromatin as a new layer of epigenome

RNA attachment was found to positively correlate with histone modifications H3K27ac and H3K4me3, both were associated with more open chromatin regions and active transcription (Sridhar et al., 2017). The genomic regions enriched with trans- interacting RNAs (RNA attachment hotspots) were clearly correlated with H3K9me3 depleted regions (Sridhar et al., 2017). These general observations, however, do not seem to apply to every type of RNAs. An exception lies in snoRNAs, of which both MARGI (Sridhar et al., 2017) and CHAR-seq (Bell et al., 2018) reported extensive interactions with chromatin, but primarily with heterochromatin (Bell et al., 2018). More intriguingly, ChAR-Seq analysis revealed that TSS-associated RNAs are enriched at TAD boundaries, corroborating with the potential role of active transcription in shaping the topological organization of the genome (Bell et al., 2018). Although the above mentioned

associations await further validations, these RNA attachments may act inter-dependently or coordinately to the hitherto better characterized DNA- and histone- modifications, thus constitute a novel layer of chromatin modifications that contribute to gene regulation.

Acknowledgement

Chapter 1, in full, is an adaptation of materials that appears in Nguyen, Tri C.; Zaleta-Rivera, Kathia; Huang, Xuerui; Dai, Xiaofeng; Zhong, Sheng. RNA, action through interactions. Trends in Genetics, 34(11):867-882, 2018. The dissertation author was the primary investigator and author of this material.

CHAPTER 2 – High-throughput mapping of RNA-RNA interactions with MARIO

Introduction

Many biological processes are regulated by RNA-RNA interactions (Kretz et al., 2013). One example of such important regulatory class of RNAs is lnRNAs. lnRNAs are a class of RNAs transcripts longer than 200 nucleotides, which are devoid of protein-coding potential. Across the human genome there are four-times more lncRNAs than traditional coding RNAs. Many lncRNAs are expressed in very specific anatomical or developmental patterns, suggesting that their regulation is of biological importance. Increasingly, recent research is showing not only their powerful role in tumorigenesis, but also in normal cell differentiation. When researchers knocked down particular lncRNAs associated with p53, it affected the expression of hundreds of genes that p53 normally repressed. The idea that lncRNAs play a direct role in tumor-suppressor or oncogenic pathways may pave the way for identifying new novel cancer therapeutics. However, it remains formidable to analyze the entire RNA interactome. Interactions between RNA molecules exert key regulatory roles and are often mediated by RNA binding proteins (Ray et al., 2013) such as ARGONAUTE proteins (AGO) (Chi et al., 2009), PUM2, QKI (Hafner et al.), and snoRNP proteins (Granneman et al., 2009). Despite recent advances such as PAR-CLIP (Hafner et al.), HITS-CLIP (Chi et al., 2009; Licatalosi et al., 2008; Zhang and Darnell, 2011), and CLASH (Granneman et al., 2009; Helwak et al., 2018), it remains a formidable challenge to map all protein-assisted RNA-RNA interactions. We therefore develop the MARIO method to detect protein-assisted RNA-RNA interactions in vivo. In this procedure, RNA is crosslinked with its bound

proteins then ligated to a biotinylated RNA linker such that RNAs co-bound by the same protein form a chimeric RNA of the form RNA1-Linker-RNA2. These linker-containing chimeric RNAs are isolated using streptavidin coated magnetic beads and subjected to pair-end sequencing. Thus, each non-redundant pair-end read reflects a molecular interaction.

We developed a technology with the following main steps (Figure 2.1). We crosslinked cells, causing the interacting RNAs mediated by a RNA binding protein to be covalently linked to this protein. We then fragmented RNAs with RNAase I and and biotinylate the cysteine residues on proteins. The proteins including protein-RNA complexes were immobilized on streptavidin beads. The 5' end of the RNA was then ligated with a biotin-tagged RNA linker (24nt) to facilitate subsequent selective purification of chimeric RNAs. Next, proximity-based ligation was carried out on beads under dilute conditions that favor ligations between crosslinked RNA fragments. Protein-RNA complex was then eluted from streptavidin beads and RNA is recovered by digesting the bound protein. Eluted RNA was subjected to rigorous DNase treatment to eliminate DNA contamination. Purified RNAs were then hybridized with a DNA probe that was complementary to the 24nt RNA linker and treated with T7 exonuclease to remove the non-ligated biotinylated RNA linkers. As a result, mainly the successfully ligated chimeric RNAs contained a biotin-tagged linker at the junction. This chimeric RNA library was fragmented again to an average of 150 nucleotides, and the ligation junctions were pulled-down with streptavidin-coated magnetic beads. The end product was a library of ~150nt chimeric RNAs. This library was enriched with chimeras of in the form of R1-linker-R2, where R1 and R2 were fragments of interacting RNAs. This library was converted into cDNAs and sequenced with paired-end next-generation sequencing.

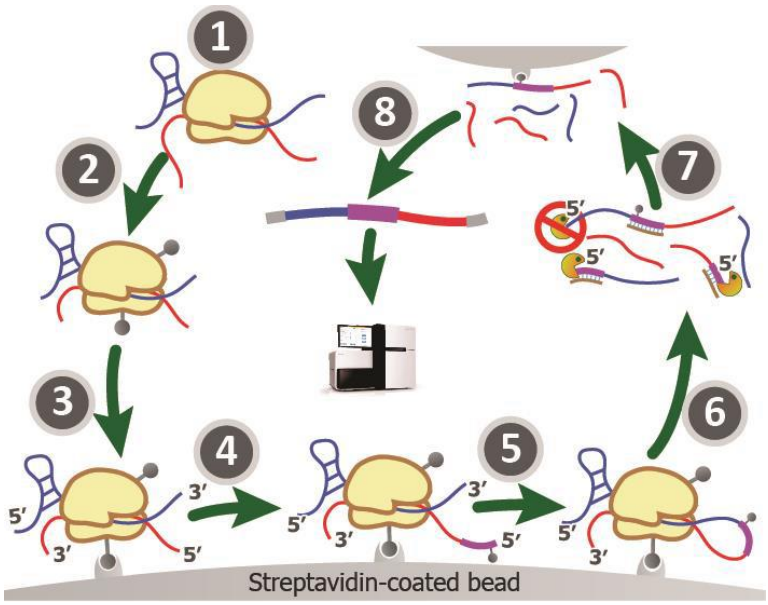


Figure 2.1*: A sequencing-based technology to map RNA-RNA interactions*. The major experimental steps: 1. crosslinking RNAs to proteins, 2. RNA fragmentation and protein biotinylation, 3. immobilization, 4. ligation of a biotinylated RNA linker, 5. proximity ligation under an extremely dilute condition, 6. RNA purification and reverse transcription, 7. biotin pull-down. 8. construction of sequencing library.

Methods

Experimental method

MARIO was designed to: (1) capture interacting RNAs in vivo in an unbiased manner without genetically or transiently introducing exogenous molecules; (2) allow stringent removal of non-physiologic associations that form after cell lysis (Mili and Steitz, 2004); (3) select the proximity-ligated chimeric RNAs; (4) allow straightforward and unambiguous bioinformatic identification of interacting RNAs. We achieved these by: (1) in vivo crosslinking and immobilization of all RNA-binding proteins in streptavidin beads and stringently washing away non-specific binding by harsh conditions; (2) introducing a biotin-tagged RNA linker to facilitate selective enrichment of chimeric transcript; (3) using the linker sequence to unambiguously split the interacting RNAs from a sequencing read.

Step 1: crosslinking RNAs to proteins

UV irradiation (254 nm, 400mJ/cm2) was used to form covalent bonds between photoreactive nucleotide bases and amino acids in 108 cells. UV irradiation generates highly reactive, short-lived states of the nucleotide bases within the RNA, inducing covalent bond formation only with amino acids at their contact points without additional elements that might cause conformational perturbation (Pashev et al., 1991). UV irradiation at 254 nm does not promote protein-protein crosslinking due to the different wave lengths absorbed by amino acids. In the control experiment (ES-indirect) where we wanted to crosslink proteins as well, we applied an in vivo dual crosslinking method with previously validated parameters, namely 45 minute treatment of ethylene glycol bis[succinimidylsuccinate] (EGS) followed by 10 minute of formaldehyde treatment (Kurdistani and Grunstein, 2003; Nowak et al., 2005; Zhang et al., 2012).

Step 2: cell lysis, RNA fragmentation, and protein biotinylation

The crosslinked cells were lysed with a mild lysis buffer (0.1% SDS, 1% NP-40, 0.5% Na-DOC). Chromatin was subsequently removed by centrifugation. The RNAs were digested by RNase I into ~1000-2000 nt (ES-1) or ~1000 nt (ES-2) fragments. The dual crosslinked cells (ES-indirect) were fragmented with sonication. Both RNase I and sonication-based fragmentation leave 5'-OH and 3'-P ends, incompatible with RNA ligation, which suppress undesirable RNA ligations. We then biotinylated the cysteine residues of proteins with EZlink Iodoacetyl-PEG2-Biotin (Pierce).

Step 3: immobilization on beads

The protein-RNA complexes were immobilized at low surface density on streptavidin-coated beads (800 µL MyOne Streptavidin T1 beads, providing a larger surface area (~200 cm2) than previously described [25]). The advantages of immobilization on a solid surface include (1)

reducing random intermolecular ligations between non-crosslinked oligonucleotides (Kalhor et al., 2011), (2) simplifying buffer exchange, (3) allowing for stringent washing to remove non-physiologic interactions. To saturate the excessive streptavidin after immobilization, the beads were incubated with free biotin (40 μL of 25 mM IPB for 600 μL of beads, proportionally scaled up from previously described (Kalhor et al., 2011)). To remove the RNAs associated with proteins either noncovalently or via nonspecific protein-protein interactions, the beads were washed with highly denaturing washing buffer containing 0.5% lithium dodecyl sulfate and 500 mM lithium chloride (as described in (Castello et al., 2013; Kwon et al., 2013)).

Step 4: ligation of a biotin-tagged RNA linker

Next, a biotin-tagged RNA linker (5'-rCrUrArG/iBiodT/rArGrCrCrCrArUrGrCrArArUrG rCrGrArGrGrA) was attached to the RNA's 5' end. The biotin-tagged linker serves as a selection marker to enrich for the ligated the RNAs; it also delineates a clear boundary to unambiguously split any sequencing read that covered a ligation junction. The 5'-end of the RNA linker was temporarily "blocked" from ligation to avoid linker circularization or concatenation. This was achieved by synthesizing the linker with a 5'-OH group, which is incompatible with ligation but can be "re-activated" by phosphorylation. However, RNase I leaves 5'-OH end, which is incompatible for linker ligation, thus we first phosphorylated the 5' end with T4 Polynucleotide Kinase (PNK), 3' phosphatase minus (NEB). We did not use the wild-type T4 PNK due to its additional 3' phosphatase activities, which modifies the 3'-ends of RNAs from 3'-P into 3'-OH, thus susceptible to self-ligation.

Step 5: Proximity ligation

We next dephosphorylated the RNA 3'-end to convert the cyclic phosphate group that was formed by RNase I digestion into a hydroxyl group to prepare for proximity ligation. Although a

phosphatase, such as Shrimp Alkaline Phosphatase (SAP) or Calf Intestinal Alkaline Phosphatase is normally employed for this purpose, we instead used the 3' phosphatase activity of T4 PNK since it has been shown to be more efficient in removing the inhibitory cyclic phosphate from 3'-ends (Huppertz et al., 2014). Proximity ligation was then performed under extremely dilute conditions to minimize inter-complex chimeric ligations which do not represent in vivo interactions. PEG couldn't be used to enhance intramolecular ligation at this stage. Because RNA secondary structure can deteriorate the ligation efficiency, we both heat-shocked the RNA and added 15% (v/v) dimethylsulfoxide (DMSO), which was shown to stimulate ligation to highly structured RNAs.

Step 6. Selection and extraction of desired RNA-RNA interactions and reverse transcription

Protein-RNA complexes were then eluted from streptavidin beads and RNAs were recovered by Proteinase K digestion of the bound proteins. The purified RNAs at this stage were a mixture of RNAs without linkers (RNA1 or RNA2), RNAs ligated with linkers but were not proximity-ligated with other RNAs (5'-linker-RNA2), and the desirable chimeric constructs in the form of 5'-RNA1-linker-RNA2. RNA1 can be depleted by selection of the biotin tagged linker. We therefore proceeded with depleting the non-informative 5'-linker-RNA2 as well in the next reaction with T7 exonuclease.

6.1. Removing biotin from terminal linkers (5'-linker-RNA2). This was based on the RNase H activity of T7 exonuclease, which not only removes 5' mononucleotides from duplex DNA but also exert exonucleolytic activity on the RNA strand from a RNA-DNA hybrid (11). A complementary DNA oligonucleotide (5'-T*C*G*C*ATTGCATGGGCTACTAGCAT, where * denotes the phosphorothioate bond to block its digestion by T7 exonuclease (12)) was annealed to

the RNA linker, creating a double stranded DNA-RNA hybrid between the RNA linker and the complementary DNA strand. The complementary DNA strand was designed so that after annealed, the 5'-end of the RNA linker was recessed while the 3'-end of the DNA strand was protruding. The annealed products were then treated with T7 exonuclease.

6.2. Removal of rRNAs. Since ~90% of total RNA is ribosomal RNA (rRNA). The presence of rRNA decreases the amount of informative sequence data obtained. Therefore, it is critical to deplete the amount of rRNA present in the sample. We tested 2 strategies for rRNA depletion: (1) Duplex-specific nuclease (DSN) for ES samples 1 and 3; (2) Antibody based removal of rRNAs using GeneRead rRNA Depletion Kit (Qiagen) for ES sample 2 and the MEF sample. We recognized that removing rRNA at this stage is not very efficient because the RNAs have been fragmented. However, we did not have another choice because most rRNA is protected within the ribosome, precluding its removal preceding protein digestion and RNA purification.

6.3. RNA shearing. RNA was fragmented into size range of 150 – 400 bp optimal for sequencing by Illumina HiSeq.

6.4. Ligation with reverse transcription adapter. Next, the RNAs were ligated with a 3' reverse transcription (RT) adapter (/5rApp/AGATCGGAAGAGCGGTTCAG/3ddC/) that served as primer for RT reaction.

6.5. Reverse transcription. For each experiment or replicate, a different RT primer containing individual experimental barcode sequence was used. Each RT primer has the form of 5'-

/5Phos/NNXXXXNNNNAGATCGGAAGAGCGTCGTGgatcCTGAACCGCTCTTCCGATCT.

According to this scheme, the first read of every sequencing read pairs contains a barcode that

takes the configuration of NNNNXXXXNN (reverse complement of that from the RT primer), where the Ns are a random 6nt barcode for removing PCR duplicates (13-16). Only when two pair-end reads have identical mapped locations AND random barcodes would they be counted as only one. The XXXX is a fixed 4nt sample barcode for multiplexed sequencing (AGGT for ES-1, CGCC for ES-2, CATT for ES-indirect, CGCC for MEF). Any two 4nt sample barcodes differs by three nucleotides to avoid potential confusions from mutations or sequencing errors.

Step 7. Biotin pull-down of chimeric RNA-DNA hybrids

Streptavidin-biotin affinity purification was used to enrich for chimeric RNA-DNA hybrids. This pull-down was carried out after the second RNA fragmentation and reverse transcription in order to allow a substantial fraction of the sequencing read pairs to cover the RNA-linker or linker-RNA junctions, in one end of the read pair.

Step 8. Construction of sequencing library

Considering the UV-induced crosslink site sometimes stalls reverse transcription, resulting in truncated cDNAs that lack the 5' adapter (Sugimoto et al., 2012), we adopted a circularization strategy that allowed for constructing sequencing libraries even from truncated cDNAs (Huppertz et al., 2014) (Figure 2.2). The RT primer contained the adapter regions to prime PCR amplification by Illumina PE PCR Forward Primer 1.0 and PE PCR Reverse Primer 2.0, flanking a BamHI restriction site and a sequencing barcode. Linearization by restriction digestion at a BamHI restriction site within the RT primer would generate suitable templates for PCR amplification. However, because the cDNAs at this stage were single-stranded, which are not substrates of BamHI, we annealed an oligo complementary to the RT primer to introduce a short double-

stranded region suitable for BamHI restriction. Besides, this strategy also prevents BamHI activities on other endogenous BamHI restriction sites.

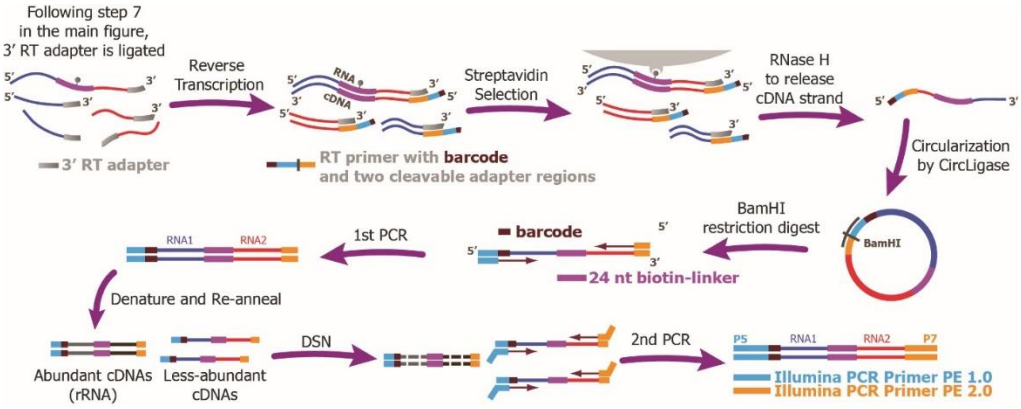


Figure 2.2*: Library construction*. The total of five random nucleotides in yellow (2 and 3 on each side of the sample barcode) are collectively form the "random barcode", which is used to remove PCR artifacts.

Computational method

The computational pipeline (MARIO tools) takes pair-end sequencing reads as input. The oligonucleotide sequences of the RNA linker and the sample barcodes used for multiplexed sequencing should also be provided to the pipeline (Figure 2.3). The main outputs include: (1) a parsed cDNA library, including the list of chimeric cDNAs in the form of RNA1–Linker–RNA2; (2) the genomic locations of RNA1 and RNA2 of every chimeric cDNA; and (3) interacting RNA pairs inferred from statistical enrichment of chimeric cDNAs. The major analysis steps of MARIO tools are as follows: (1) removing PCR duplicates; (2) assigning multiplexed sequencing reads into corresponding experimental samples; (3) recovering the cDNAs in the sequencing library; (4) parsing the chimeric cDNAs; (5) mapping to the genome; (6) identifying interacting RNA pairs; and (7) identifying RNA interaction sites. Detailed documentation of MARIO tools is available at http://mariotools.ucsd.edu

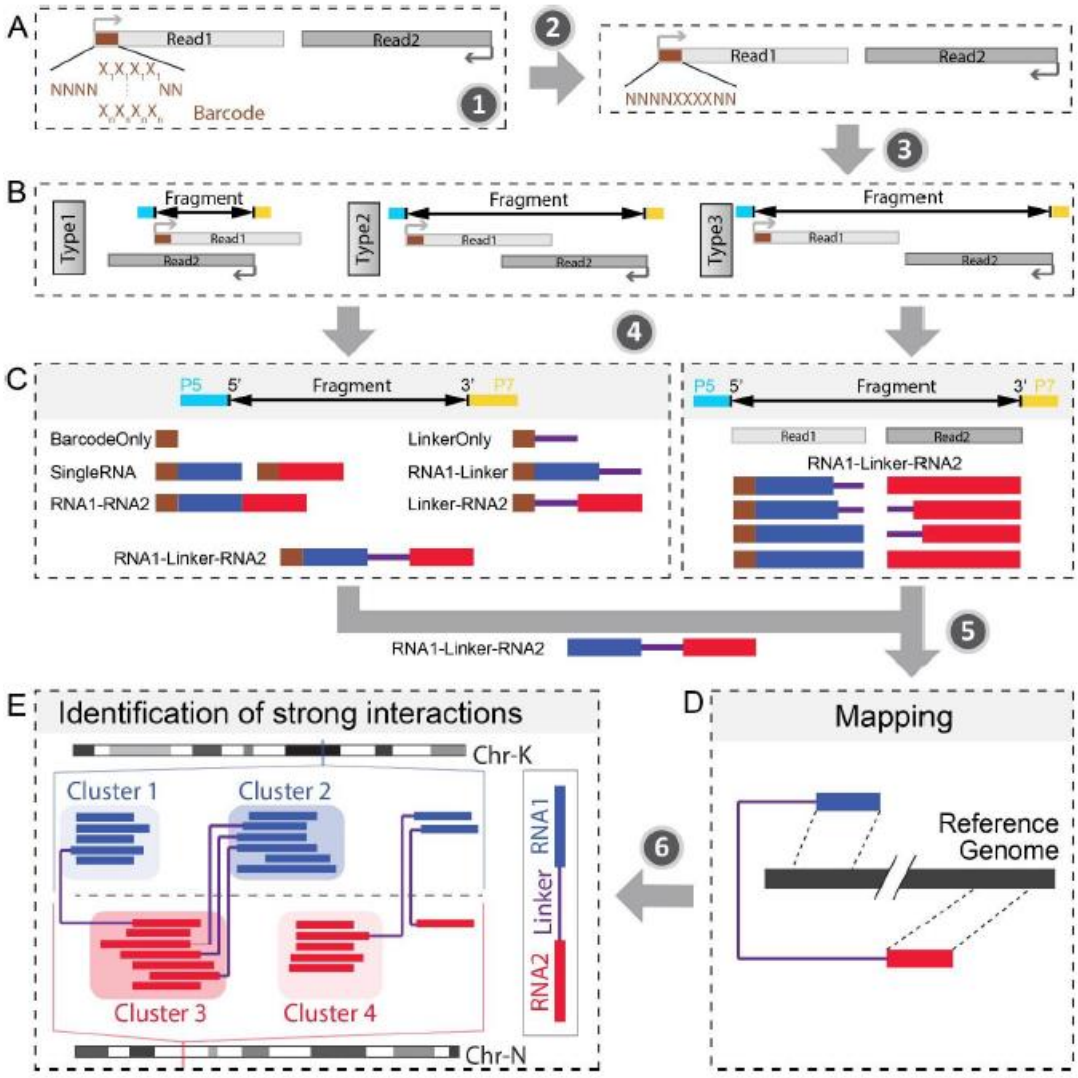


Figure 2.3*: The computational pipeline for analysis of MARIO data.* (A) PCR duplicates were removed from the pair-end sequencing reads (Step 1). Multiplexed samples were separated based on the 4nt experimental barcodes ('XXXX', Step 2). 'N': a nucleotide of the random barcode. 'X': a nucleotide of the experimental barcode. (B) Each pair of forward (Read1) and reverse (Read2) reads were used to recover a cDNA in the input sequencing library, if possible. (C) The recovered cDNA were categorized based on the configuration of the RNA fragments and the linker sequence (Step 4). The RNA1-Linker-RNA2 type of cDNAs were provided as the output. (D) The RNA1 and the RNA2 parts were separately mapped to the genome. The output was the cDNAs where both RNA1 and RNA2 were uniquely mapped to the genome. (E) RNA-RNA interactions were identified based on association tests.

Results

RNA Hi-C protocol selectively enriches RNAs in the form of 5'-RNA1-linker-RNA2

We characterized several experimental factors that can contribute to the success of procedure: We first calibrated the concentration of RNase I in the first fragmentation to trim RNAs in the lysate down to ~500-1000 nt on average. After cell lysis, the lysate was treated with increasing concentration of RNase I. RNAs were then purified from RNaseI-treated ES cell lysate by adding equal volume of 2x Proteinase K buffer (100 mM Tris-HCl pH 7.5, 100 mM NaCl, 2% SDS, 20 mM EDTA) and 1:5 volume of 20 mg/ml Proteinase K (NEB) and incubating at 55°C for 2 hours before phenol:chloroform treatment and ethanol precipitation. The concentrations tested were: no RNase I (Sample 1), 2.5U RNase I/ml lysate (Sample 2), 3.3U RNase I/ml lysate (Sample 3), 5.0U RNase I/ml lysate (Sample 4), and 12.5U RNase/ml lysate (Sample 5). The size distribution of of RNAs was monitored with the Agilent Bioanalyzer using RNA 6000 Pico Bioanalyzer Assay (**Figure 2.4A**). The concentration of 5.0U RNase I/ml lysate that produced 500-1000nt RNA fragments (Sample 4) was chosen for RNA Hi-C Step 2 in Figure 1.

We tested the efficiency of linker ligation on beads (**Figure 2.4B**). RNase-I-treated lysate was immobilized on streptavidin beads and then ligated with the biotin-labelled RNA linker (1). After ligation and proteinase K digestion to remove the proteins, RNAs were purified and quantified (1.3ug) (2). The purified RNAs were then subjected to streptavidin-biotin pull-down to select for those ligated to the biotin-labelled linker (3). After washing and eluting RNAs that were bound to streptavidin beads and ethanol precipitated, 0.22ug RNAs were collected. In parallel, the biotin-labeled RNA linkers were subjected to the same streptavidin-biotin pulldown, elution and ethanol precipitation (4). Assuming that the efficiencies of biotin pulldown, RNA elution and ethanol

precipitation in Steps 3 and 4 were the same, which was about 19.6% (1.96ug / 10.0ug), we estimated the ligation efficiency as (0.22ug/19.6%)/1.3ug = 86%.

We monitored RNA size distributions at different steps of the RNA Hi-C procedure (**Figure 2.5A**). Size distributions of RNAs in the lysates of MEF (Lane 1) and ES-indirect (Lane 2) before being tethered onto streptavidin beads, in the supernatant after immobilization (Lanes 3 and 4) and immobilized on beads after proximity ligation (ES-indirect: Lane 5, MEF: Lane 6). RNAs were denatured in 2X RNA loading dye (NEB) at 70^{o}C for 5 minutes, run on 1.5% Native Agarose gel and stained with SYBR Gold (Invitrogen).

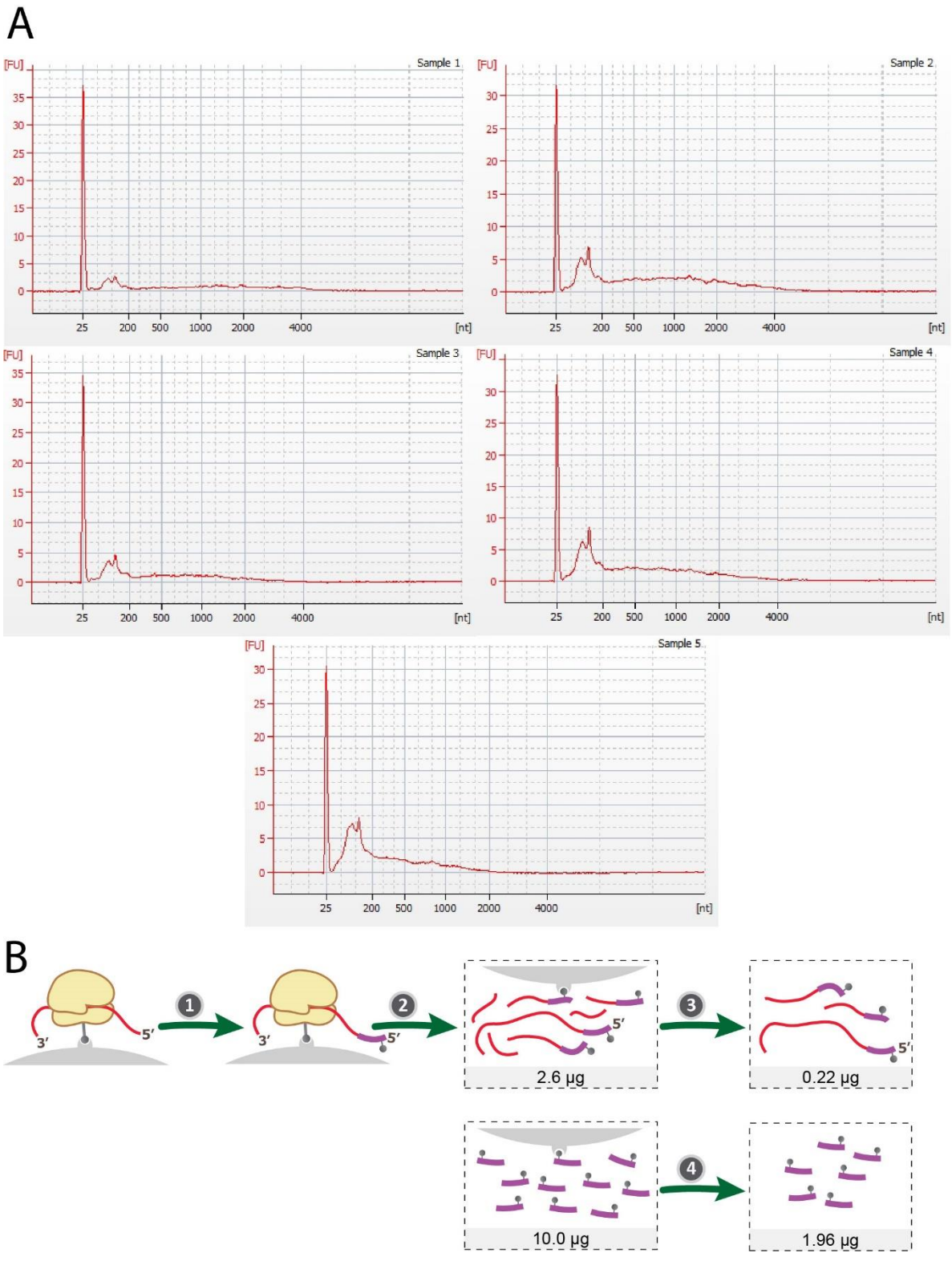


Figure 2.4*: **Calibration of RNase I concentration*** (A) and testing the efficiency of linker ligation on beads (B)

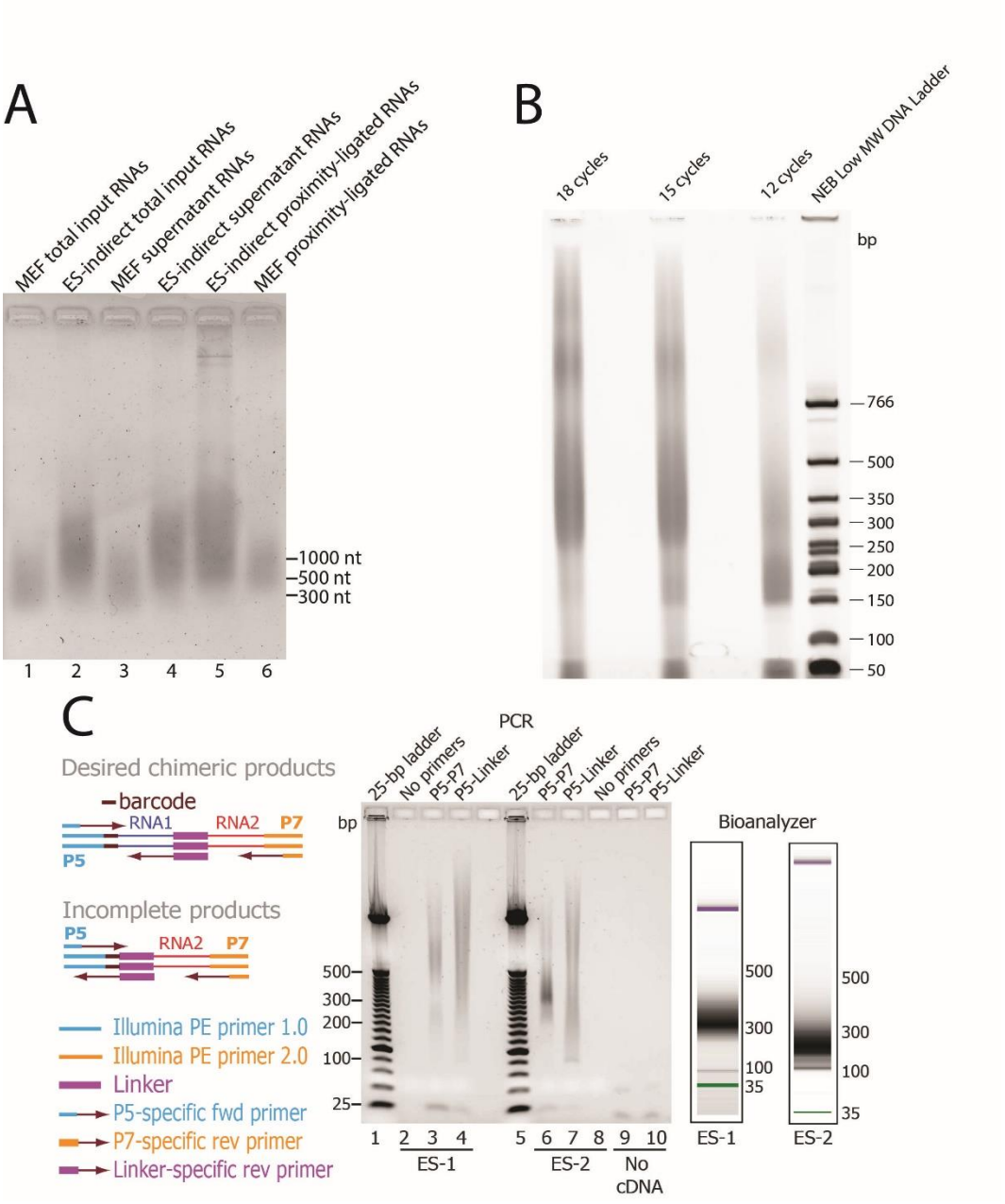


Figure 2.5*: **RNA size distributions at different steps of the RNA Hi-C procedure*** (A); optimization of the number of PCR cycles for construction of sequencing library (B); and PCR validation of RNA1-Linker-RNA2 chimeras (C)

We also optimized the number of PCR cycles for construction of sequencing library (**Figure 2.5B**). In Step 8 of the RNA Hi-C procedure, single-stranded cDNAs of the ES-1 sample were pre-amplified with 12 cycles of PCR using a truncated form of Illumina PCR sequencing

primers (DP5, DP3). The PCR products were purified with 1.8x SPRISelect beads, which produced 86 ng of double-stranded DNAs before the depletion of rRNA cDNAs by duplex-specific nuclease. One µl aliquots from a total of 22 µl rRNA-depleted double-stranded cDNAs were amplified with different PCR cycle numbers: 12, 15, 18, using NEBNext High-Fidelity 2X PCR Master Mix (NEB) and Illumina PE Primer 1.0 and 2.0. The PCR products were run on 6% TBE PAGE gel and stained with SYBR Gold (Invitrogen). Based on the gel result, 18 µl of rRNA depleted double-stranded DNAs were then amplified with 11 cycles of PCR to generate the sequencing library. Before high throughput sequencing, we experimentally verified by PCR that the majority of fragments in the final sequencing library containing the desirable chimeras in the form of RNA1-linker-RNA2 (**Figure 2.5C**). RNA1-Linker-RNA2 chimeras are expected to be above 91 bp from the P5 sequencing primer to the linker (purple) and above 200 bp from P5 to P7 sequencing primers. The failure to include RNA1 would create 91 bp products from P5 to the linker. The failure to include RNA2 would create similar sized products from P5 to the linker and from P5 to P7. The PCR primers are marked on top of each lane. The size distribution of the sequencing libraries was also assessed by Bioanalyzer (**Figure 2.5C**).

We carried out six independent RNA Hi-C samples on two different cell lines: mouse embryonic stem (ES) cells E14 and mouse embryonic fibroblasts (MEFs) with slightly procedural modifications. A summary of the six samples is given in **Table 2.1**. Using a threshold of at least 5 nt long for RNA1 and RNA2, we obtained, on average, approximately15.1 million non-redundant pair-end reads represented the desired chimeric form RNA1-linker-RNA2 out of 47.3 million pair-end reads.

Table 2.1: Description of the RNA Hi-C samples

Sample name	ES-1	ES-2	ES-indirect	MEF-1	MEF-2	MEF-3
Cell type	ES cells	ES cells	ES cells	MEF	MEF	MEF
Crosslinking	254nm UV	254nm UV	Dual crosslinking	254nm UV	254nm UV	254nm UV
RNA-protein interactions	Direct	Direct	Indirect	Direct	Direct	Direct
Protein solubilization	Detergents	Detergents	Sonication	Detergents	Detergents	Detergents
Targeted functional group for biotinylation	Sulfhydryls (-SH)	Sulfhydryls (-SH)	Sulfhydryls (-SH)	Sulfhydryls (-SH)	Primary amines (-NH2)	Primary amines (-NH2)
First fragmentation	1000-2000 nt	~1000 nt	~1000 nt	~300 nt	~1000 nt	~1000 nt
rRNA removal	Duplex-specific nuclease	Antibody based	Duplex-specific nuclease	Antibody based	Antibody based	Antibody based
Cellular compartment	Enriched for cytoplasm	Enriched for cytoplasm	Entire cell	Enriched for cytoplasm	Entire cell	Entire cell
Sample barcode	ACCT	GGCG	AATG	GGCG	AATG	GGCG
Total # of read pairs	45,702,794	49,316,127	74,009,386	83,083,324	68,477,373	62,157,745
# of reads RNA1-Linker-RNA2	9,884,229	3,325,392	10,647,838	5,994,815	1,076,391	906,420

RNA Hi-C identifies statistically significant mRNA-snoRNA, lincRNA-mRNA, pseudogeneRNA-mRNA, miRNA-mRNA interactions

Because ES-1 and ES-2 are similar judged by correlations of FPKMs and the interacting RNA pairs identified from ES-1 and those from ES-2 exhibited strong overlaps (p-value<10^{-35}, permutation test), we merged the two datasets to infer the RNA interactome in ES cells. This produces 4.54 million non-duplicated pair-end reads that were unambiguously split into two RNA fragments with both fragments uniquely mapping to the genome (mm9). We identified 46,780 inter-RNA interactions (FDR < 0.05, Fisher's exact test). mRNA-snoRNA interactions were the most abundant type, although thousands of mRNA-mRNA and hundreds of lincRNA-mRNA, pseudogeneRNA-mRNA, miRNA-mRNA interactions were also detected. Specifically, we identified 172 snoRNAs that interacted with mRNAs and also were supported by AGO HITS-CLIP and small RNA sequencing data, suggesting that most of the expressed snoRNA genes were enzymatically processed into miRNA-like small RNAs and interacted with mRNAs in RISC complex. An example of such interaction is the one between the 3' UTR of Trim25 RNA and small nucleolar RNA (snoRNA) Snora1, which was supported by 24 and 22 pair-end reads in ES-1 and ES-2 samples.

Validation of Malat1 and Slc2a3 interaction using single molecule FISH (smFISH)

Among interactions involving lncRNAs, the largest lncRNA hub is Malat1, which interacted with 4 mRNAs. We independently confirm the interaction between Malat1 lncRNA and Slc2a3 mRNA by visualizing their co-localization *in situ* by two-color single molecule RNA fluorescence in situ hybridization (smRNA-FISH). Co-localization of Malat1 and Scl2a3 RNAs were detected

in 10 cells, with 1 to 3 co-localized RNA pairs per cell (p-value = 4×10^{-40}, Fisher's exact test) (Figure 2.6).

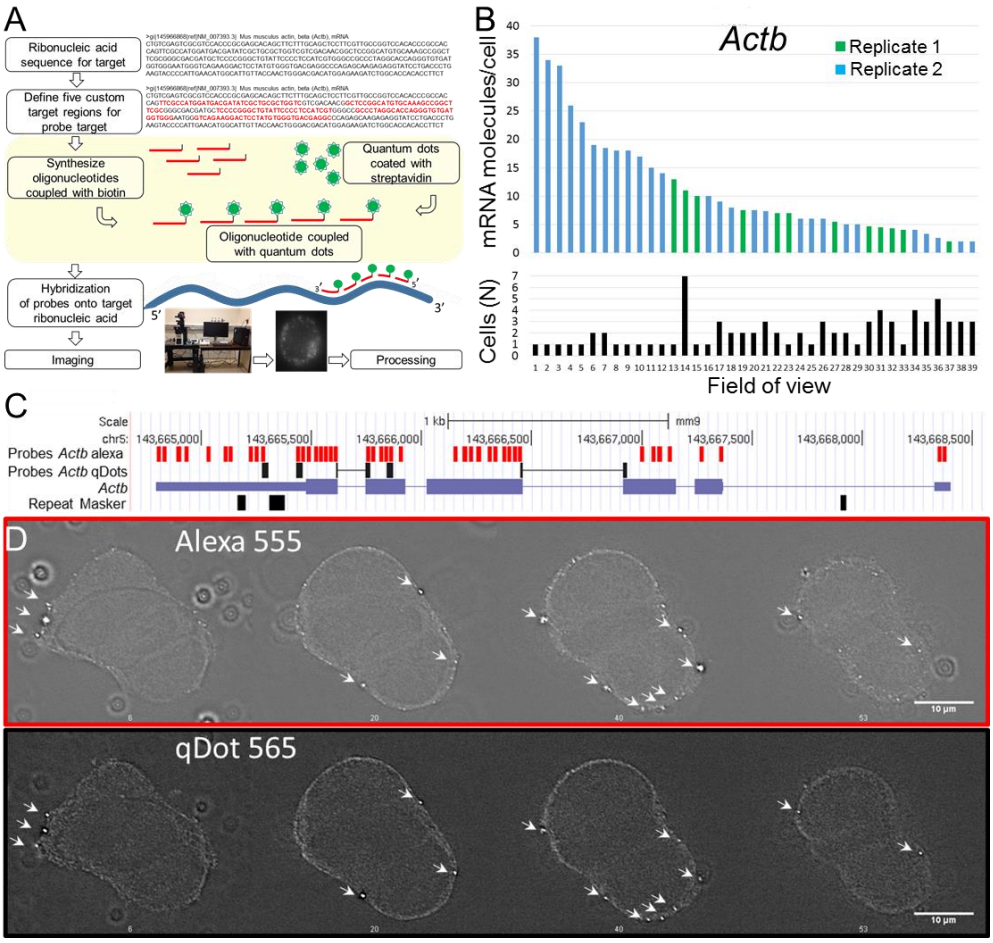


Figure 2.6*: Detection of RNA molecules with smRNA-FISH*. (A) Scheme of single molecule RNA-FISH with probes labeled with quantum dots. (B) Distribution of *Actb* mRNA molecules in 82 single ES cells, in 39 fields of view, from two independent experiments (Replicates 1 and 2). (C) Genomic positions for smRNA-FISH probes labeled with organic dyes (red) and probes labeled with qDots (black) for the same gene (*Actb*). (D) Co-localization of signals detected (arrows) from probes labeled with organic (Alexa 555) and inorganic (qDot 565) dyes.

Identification of Malat1 interactome by RIA-Seq

Because Malat1 is the largest lncRNA hub identified by RNA Hi-C procedure, we set out to identify associated RNAs that are targeted by Malat1 using an independent technique, RNA

interactome analysis, followed by deep sequencing (RIA-Seq) developed by Kretz et al. (Kretz et al., 2013). In RIA-Seq, biotinylated anti-sense DNA probes were designed in even- and odd-numbered pools. These two pools were used separately for pull-down of endogenous lncRNA of interest and associated RNAs (Kretz et al., 2013).

We first ensured that the RIA-Seq worked in our hands. We used the biotinylated anti-sense DNA FISH probes against Malat1 and β-actin that were designed for smFISH experiment to pull-down Malat1 and β-actin and associated RNAs, respectively. Five probes, each 25-30 nt long, were designed for Malat 1 and β-actin. Before sequencing the RNAs associated with Malat1, and β-actin, we verified that the enriched RNAs were specific by qPCR (**Figure 2.7**). We designed two pairs of qPCR primers surrounding each probes, one pair upstream and one pair downstream from the probes. Although the expression level Malat1 and β-actin is comparable (1.3 fold difference as determined by qPCR), Malat1 but not β-actin is specifically enriched in Malat1 RNA RNA interactome analysis and vice versa (**Figure 2.7**). **Figure 2.7** showing the enrichment of the region surrounding β-actin probe 1 to 5 as well as Malat1 probe 1 to 5 for Malat1 RIA (**Figure 2.7A**) and β-actin RIA (**Figure 2.7B**).

After ensuring that RIA worked in our hand, the next step is to extend the number of probes to cover the entire length of Malat1. Because Malat1 is 6982 nt long in mouse, we designed 94 probes to cover the entire length of Malat1, each probe is 20 nt long, spacing 40-120 nt from each other. We divided these probes into odd and even sets, 47 probes in each set and repeated the experiment using each set separately as internal control. We will prepare two separate pull-down of Malat1 and associated RNAs from each odd and seven probe set. The pull-downed RNAs will be converted into sequencing libraries and analyzed to identify overlapping RNAs from the two probe sets. Those overlapping RNAs will be more likely to be true targets of Malat1. We will then

verify whether Malat1 indeed interact with the RNA partners that we have identified from our RNA Hi-C procedure.

Figure 2.7*: **qPCR validation of Malat1 RIA*** (A) and β-actin RIA (B)

The constructed library was sequenced. We did both Malat1 RIA-seq and Actb RIA-seq (control) to test the interactions involving Malat1. Malat1 RNA itself exhibited a 5.81 fold increase in Malat1 RAP-seq over Actb RAP-seq, confirming the validity of the RAP purification. Malat1 interacting RNAs reported by MARIO showed 14.6 (0610007P14Rik), 4.53 (Slc2a3), 3.38 (Eif4a2), and 2.39 (Tfrc) fold increase in Malat1 RAP-seq over Actb RAP-seq (p-value < 0.0003, Chi-squared test). This suggests a strong overlap of Malat1 targets in MARIO and Malat1 RAP-seq. Next, we asked whether Tfrc RAP could reversely identify Malat1 by Tfrc RAP-seq (Supplementary Note 4). The Tfrc RNA itself showed 2.87 fold of increase in Tfrc RAP-seq compared to Actb RAP-seq. Malat1 exhibited 3.84 fold increase (p–value < 2.2×10^{-16}, derived

from testing the null hypothesis fold change = 1), suggesting that antisense purification of Tfrc could reversely pull-down Malat1. In addition, three out of four other Tfrc interacting RNAs identified by MARIO exhibited 1.4 – 13.6 fold increases (p-value < 0.00002, Chi-squared test). Taken together, 7 additional MARIO identified interactions were validated by RAP-seq.

Assessment of false positives caused by inter-molecular ligations of RNAs that come from different complexes in vitro

As a control experiment, we prepared a sample wherein the lysate obtained from UV-irradiated mouse ES cells are mixed with an equal quantity (as measured by RNA content) of Drosophila S2 lysates prior to MARIO analysis. The frequency of drosophila-mouse RNA chimeras indicates the frequency at which RNA-RNA interactions recovered by MARIO are formed in vitro following cell lysis. These drosophila-mouse chimeras are formed due to inter-molecular ligation between RNAs that come from different complexes. The mixture was subjected to the rest of the experimental procedure and resulted in a sequenced library (Fly-Mm). The proportion of RNA pairs mapped to two species is in the range of 2.5% and 6.8%, depending on whether the drosophila genome and the mouse genome were assembled into a pan-genome before mapping). We chose the more conservative estimate (derived from mapping to the pan-genome) that 6.8% of the ligation products were probably generated from random ligations.

Overview of ES cell RNA interactome

We analyzed the ES cell RNA interactome based on the merged data of ES-1 and ES-2, which included 4.54 million unambiguous and non-redundant chimeric RNAs. As an example, an interaction between the 3' UTR of Trim25 and SNORA1 was supported by 24 and 22 pair-end reads in ES-1 and ES-2 libraries, respectively, but not supported by any reads in ES-indirect or

MEF libraries (Figure 2.8). The predicted \interacting site" on Trim25 by the overlapping RNA Hi-C reads was associated with the Argonaute protein in ES cells, so was the SNORA1 RNA (AGO CLIP-seq, Figure 2.8). Since the ES-indirect sample used dual crosslinking method, which also included indirect RNA-RNA interactions for RNAs bound to different interacting proteins, we decided to use this sample as control to identify only direct RNA-RNA interactions where only RNAs bound to same protein were ligated. We identified 46,780 inter-RNA interactions under the criteria of FDR < 0:05 (p-values from hypergeometric test and FDR from Benjamini-Hochberg procedure) and each interaction being supported by fold change between test and control no less than 3. The most abundant types of interactions were between an mRNA and a snoRNA (27,375 unique interactions), two mRNAs (7,076 interactions), an mRNA and a tRNA (2,894 interactions) and two snoRNAs (2,781 interactions). Among other types of detected interactions were mRNA-snRNA, lincRNA-mRNA, and miRNA-mRNA interactions (Figure 2.9A).

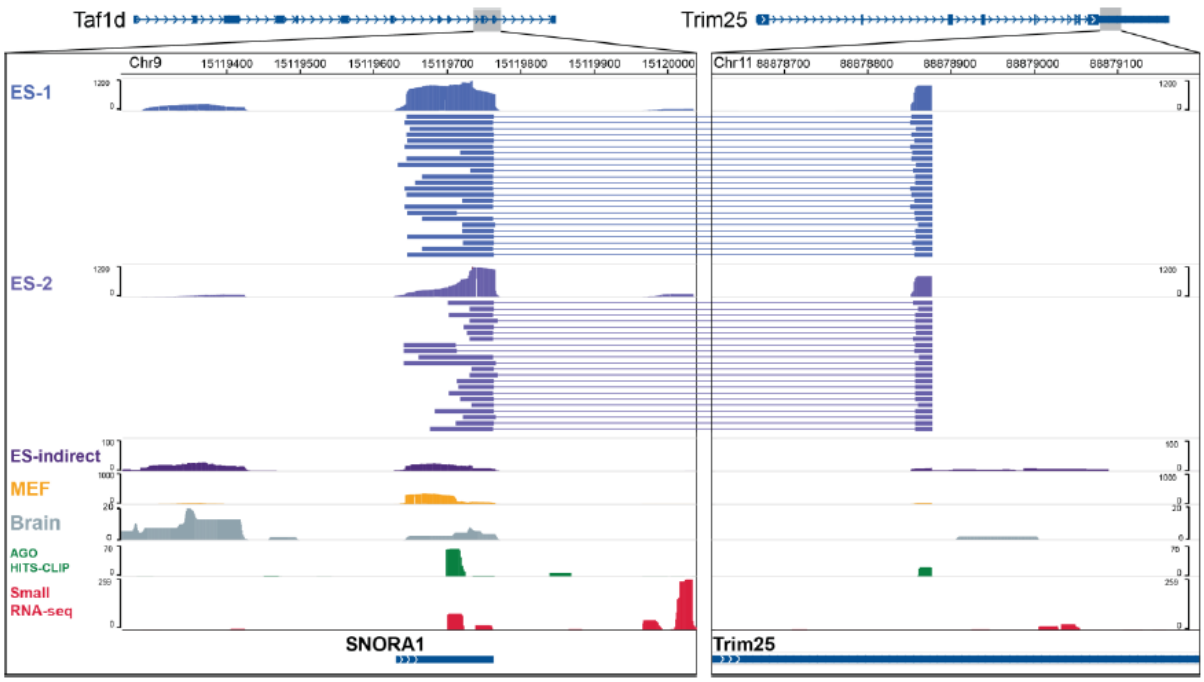


Figure 2.8*: MARIO data mapped to the genome*. Ligation of Trim25 and Snora1 RNAs was supported by multiple pair-end reads in ES-1 and ES-2 libraries. Ago CLIP-seq: AGO HITS-CLIP of mouse ES cells (GEO: GSM622570). Small RNA-seq: sequencing of small RNAs with a 3' hydroxyl group resulting from enzymatic cleavage (GEO: GSM945907).

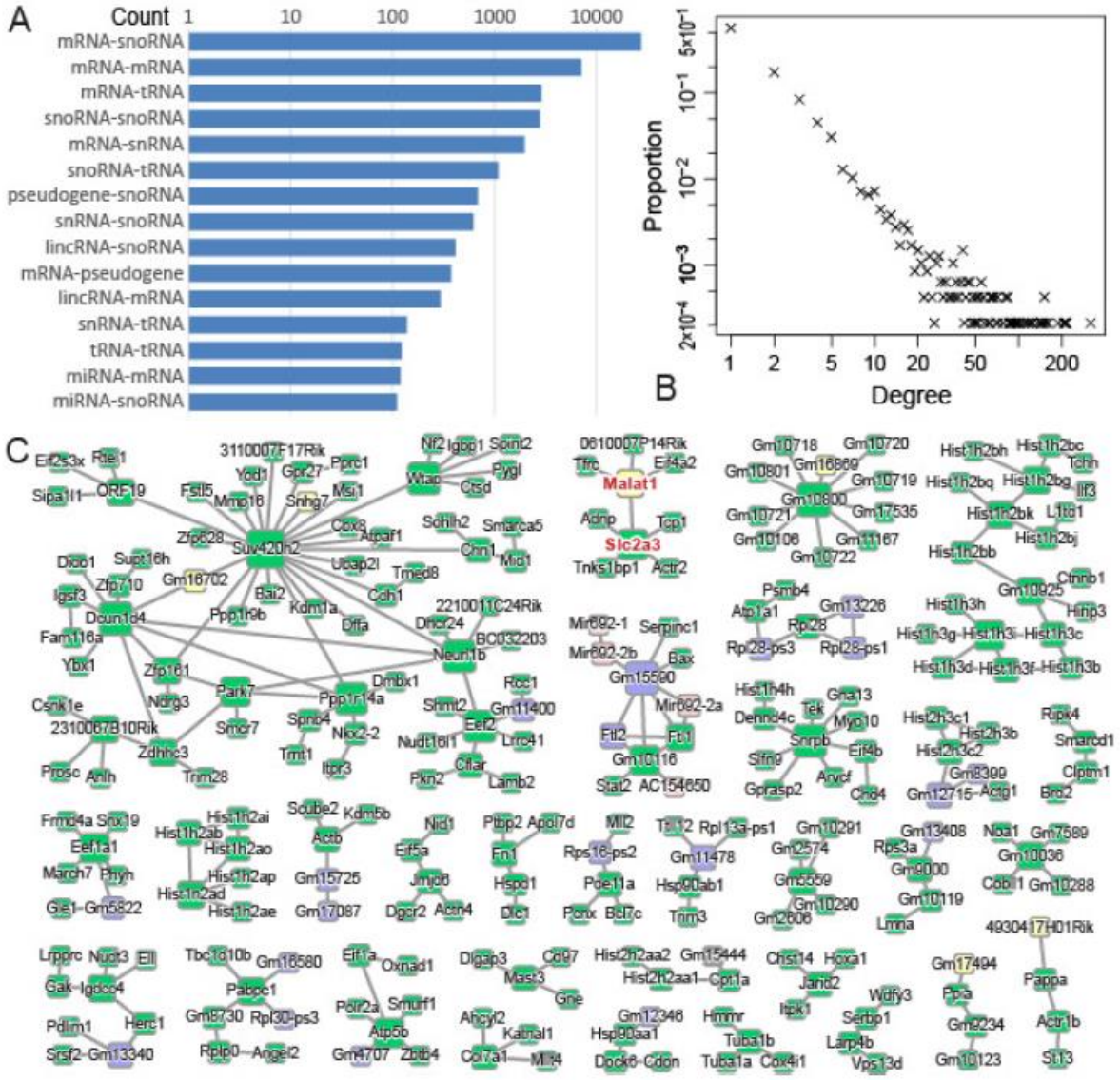


Figure 2.9*: The RNA interactome in ES cells*. (A) Distribution of detected RNA-RNA interactions among different types of RNAs. (B) Degree of network nodes against its corresponding proportion in the network (4738 nodes in total). Log-log scale plot was used to show that the network satisfies power law distribution (C) The part of the RNA interaction network, excluding snoRNA, snRNA, and tRNA and the modules of less than 4 nodes. Green: mRNAs; purple: pseudogene; yellow: lincRNA; red: miRNA; grey: antisense RNA.

To see if the detected interactions were biologically meaningful, we constructed an RNA-RNA interaction network based on the identified inter-RNA interactions. Each RNA was linked with another RNA if they showed interactions from our merged ES data. The ES cell RNA-RNA

interactome was a scale-free network, with a degree distribution conformed to power law (Figure 2.9B). To see whether the scale-free property was driven by a small number of highly connected snoRNAs, snRNAs, and tRNAs, we removed them from the network. The interactions composed only of mRNAs, lincRNAs, miRNAs, pseudogeneRNAs, and antisenseRNAs remained scale-free. A number of mRNAs, pseudogene RNAs, and lincRNAs emerged as hubs (Figure 2.9C). The largest mRNA hub was Suv420h2, which interacted with 21 mRNAs and 2 lincRNAs. The largest lincRNA hub was Malat1, which interacted 4 mRNAs, including an mRNA hub of Slc2a3. The scale-free property of RNA-RNA interactome network confirmed that the identified inter-RNA interactions were not random.

Increased interspecies conservation of RNA interaction sites

If these RNA–RNA interactions are sequence specific, the RNA interaction sites should be under selective pressure. We found that the interspecies conservation levels are strongly increased at the interaction sites and the peak of conservation precisely pinpointed the junction of the two RNA fragments (Figure 2.10). When interacting with lincRNAs, pseudogene RNAs, transposon RNAs or other mRNAs, the interaction sites on mRNAs were more conserved than the rest of the transcripts (Figure 2.11). The interactions sites on lincRNAs and pseudogene RNAs exhibited increased conservation in lincRNA–mRNA, pseudogeneRNA–mRNA and pseudogeneRNA–transposonRNA interactions (Figure 2.11). The increased conservation at interaction sites was not due to exon-intron boundaries (Figure 2.12). Taken together, base complementarity is widespread in the interactions of long RNAs. The complementary regions are evolutionarily conserved.

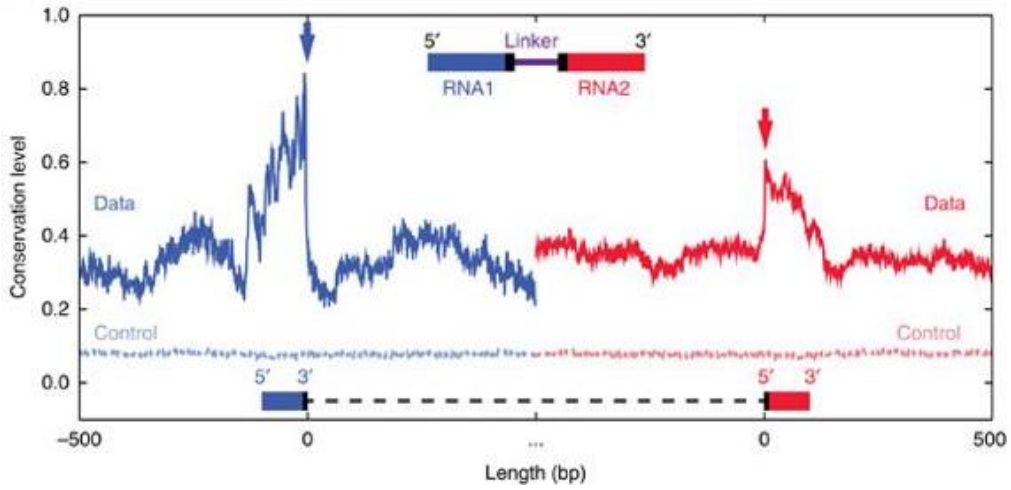


Figure 2.10*: Conservation levels*, measured by average PhyloP scores peaked at the junction (black bar, position 0 on the x axis) of the ligated RNA fragments. Control: conservation levels of randomly selected genomic regions.

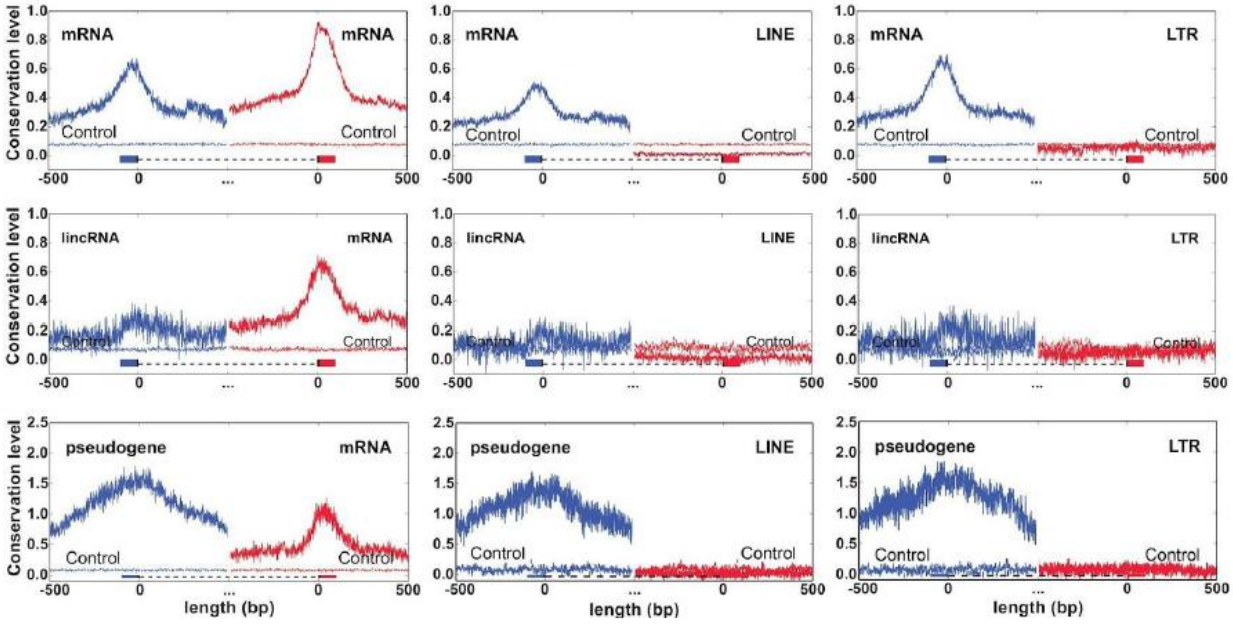


Figure 2.11*: Conservation levels of interacting RNAs*. Interactions were categorized by RNA types. For each type of interactions, the conservation level was approximated by the average PhyloP scores of the genomic regions (1000bp) centered at the RNA ligation junctions (position 0 on the x axis). The conservation levels of random genomic regions of the same lengths were plotted as controls. Blue and red bars: the RNA1 and RNA2 fragments of a RNA1-Linker-RNA2 chimeric RNA. Dashed line: the linker.

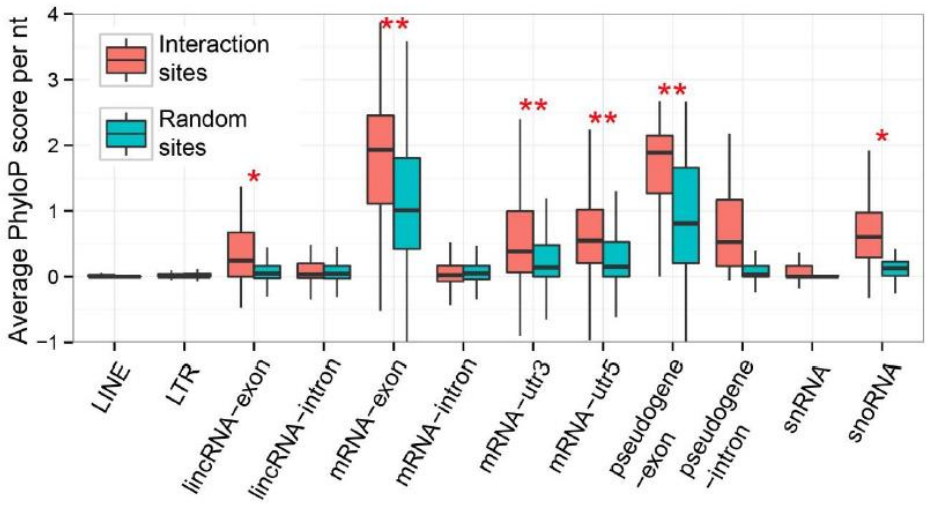


Figure 2.12*: Comparison of the conservation levels.* Conservation levels were quantified by the average PhyloP score per nucleotide of the interaction sites (y axis). To adjust for the difference of conservation of exons, introns, and UTRs, the interaction sites (red) in annotated exons, introns, and UTRs (dubbed genomic features) were compared to 200,000 randomly sampled genomic sequences from the same genomic feature (blue). The sizes of the randomly sampled genomic sequences shared the same mean and variation as the sizes of interaction sites. P-values were calculated from one-sided two-sample t-test. **: p-value $<10^{-12}$; *: p-value $< 10^{-6}$.

MARIO reveals unique information on RNA structure

Although we originally designed MARIO for mapping inter-molecule interactions, we also found that MARIO revealed RNA secondary and tertiary structures. All the analyses above were based on intermolecular reads. By looking at intramolecular reads, we learned two characteristics of RNA structure. First, the footprint of single-stranded regions of an RNA were identified by the density of RNase I digestion sites (RNase I digestion was applied before ligation, see Step 2 in Figure 2.1). Second, the spatially proximal sites of each RNA were captured by proximity ligation (Step 5 in Figure 2.1). Over 60,000 linker-containing read pairs were mapped to individual genes and thus were determined as intramolecule cutting and ligation. Each cut-and-ligated sequence can be unambiguously assigned to one of two structural classes by comparing the orientations of RNA1 and RNA2 in the sequencing read with their orientations in the genome (Figure 2.13a). These reads

provided spatial proximity information for 2,374 RNAs, including those from 1,696 known genes and 678 novel genes. For example, 277 cut-and-ligated sequences were produced from Snora73 transcripts (Figure 2.13b). The density of RNase I digestion sites (Figure 2.13c) was strongly predictive of the single-stranded regions of the RNA (heatmap; Figure 2.13e). Six pairs of proximal sites were detected (circles; Figure 2.13d). Each pair was supported by three or more cut-and-ligated sequences with overlapping ligation positions (black spots; Figure 2.13b). Five out of the six proximal site pairs were physically close in the generally accepted secondary structure (arrows of the same colour; Figure 2.13e). On Snora14, a pair of inferred proximal sites appeared distant, according to sequenced inferred secondary structure. However, ribonucleoprotein DYSKERIN bent Snora14 transcript in vivo, making the two pseudouridylation loops close to each other, as predicted by the cut-and-ligated sequence (green arrows; Figure 2.13f). Structural information can even be derived on novel transcripts and some parts of mRNAs. To date, resolving the spatially proximal bases of any individual RNA in three-dimensional space remains a grand challenge. MARIO provides intra-molecule spatial proximity information for the thousands of RNAs. In addition, the single-strand footprints of every RNA are mapped at the same time. Thus, MARIO largely expanded our capacity to examine RNA structures.

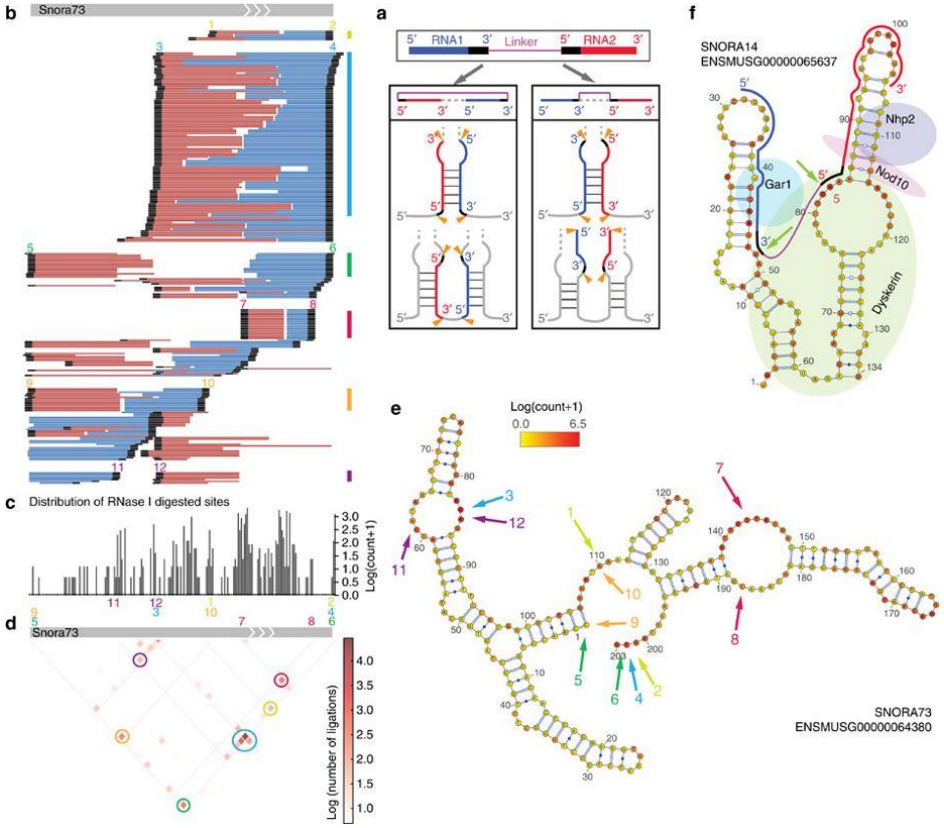


Figure 2.13*: Schematic depiction of resolving the proximal sites of an RNA*. Orange arrow: (a) RNase I cutting site. (**b**) The 'cut and ligated' products mapped to *Snora73*. Black regions: ligation junctions. Vertical colour bar: a cluster of read pairs supporting a pair of proximity sites. (**c**) Density of RNase I cuts. (**d**) Heatmap of the ligation frequencies between any two positions of the RNA. Each coloured circle corresponds to a vertical colour bar in **a** and represents a pair of proximal sites. (**e**) Footprint of single-stranded regions (red in colour scale) and inferred proximal sites (arrows of the same colour) on the accepted secondary structure. (**f**) A pair of inferred proximal sites, which were not supported by sequenced-based secondary structure, are physically close *in vivo*, supposedly due to protein-assisted RNA folding.

Discussion

The MARIO method offers several advantages for mapping RNA–RNA interactions. First, MARIO directly analyses the endogenous cellular features without introducing any exogenous nucleotides or protein-coding genes before cross-linking. This eliminates the uncertainty of reporting spurious interactions produced by changing the RNA or protein expression levels.

Moreover, it makes MARIO well-suited for assaying tissue samples. Second, the introduction of a selectable linker enables an unbiased selection of interacting RNAs, making it possible to globally map an RNA–RNA interactome. This method circumvents the requirement for a protein-specific antibody or the need to express a tagged protein. It also removes the limit of working with one RNA-binding protein at a time. Third, this method only captures the RNA molecules co-bound with a single protein molecule, avoiding capture of RNA molecules that are independently bound to different copies of a protein, which would potentially lead to reporting spurious interactions. Fourth, false positives that result from RNAs ligating randomly to other nearby RNAs are minimized by performing the RNA ligation step on streptavidin beads in extremely dilute conditions. Fifth, the RNA linker provides a clear boundary delineating the position of ligation site in the sequencing reads, thus avoiding ambiguities in mapping the ligated chimeric RNA. Sixth, potential PCR amplification biases are removed by attaching a random four- or six-nucleotide barcode to each chimeric RNA before PCR amplification and subsequently counting completely overlapping sequencing reads with identical barcodes only once. MARIO should facilitate future investigations of RNA functions and regulatory roles.

Acknowledgements

Chapter 2, in full, is an adaptation of materials that appears in Nguyen, Tri C., Cao, Xiaoyi, Yu, Pengfei, Xiao, Shu, Lu, Jia, Biase, Fernando H., Sridhar, Bharat, Huang, Norman, Zhang, Kang, Zhong, Sheng. "Mapping RNA-RNA interactome and RNA structure in vivo by MARIO". Nature Communications, 7:12023, 2016. The dissertation author was the primary investigator and author of this material.

CHAPTER 3 – High-throughput mapping of RNA-chromatin interactions with MARGI

Introduction

Besides engaging in extensive RNA-RNA interactions, lncRNA is discovered recently to exert their functions by interacting with the chromatin, modulating epigenetic events (Chu et al., 2011; Rinn and Chang, 2012; Tsai et al., 2010). It has been hypothesized that long non-coding RNAs have a role to play in creating unique epigenetic profiles (Chu et al., 2011; Engreitz et al., 2013; Simon et al., 2011). Both cis- and trans-acting models of epigenetic modulations mediated by RNAs were found. In cis, non-coding RNAs, like Xist have been shown to have a role in guiding chromatin modifying enzymes and other proteins to specific genomic loci that they have been transcribed from (Colak et al., 2014; Engreitz et al., 2013). In trans, non-coding RNA can direct chromatin modifying enzymes to genomic loci other than those from which they are transcribed, through non-Watson-Crick base paring or by acting as a scaffold molecule that assembles regulatory enzyme and is itself targeted to specific genomic loci by other adapter proteins (Miao et al., 2018). As trans-regulators, it has also been proposed that non-coding RNA can have a role in allosteric regulation of some enzymes or as a decoy molecule that can sequester enzymes away from the chromatin (Tsai et al., 2010). Specifically, long non-coding RNAs have been shown to modulate chromatin structure to facilitate processes such as dosage compensation, transcriptional gene silencing and imprinting of genes (Franke and Baker, 1999).

Despite the growing body of evidences implicating non-coding RNA in cellular regulations, there have been very few reported techniques to identify chromatin interacting RNAs. A few

techniques were developed to identify the genomic targets of known chromatin interacting RNAs. A few examples are RNA antisense purification – seq (RAP-seq) (Engreitz et al., 2013), Chromatin Isolation by RNA purification coupled with sequencing (ChIRP-seq) (Chu et al., 2011), and Capture hybridization analysis of RNA targets (CHART) (Simon et al., 2011). These techniques use slightly different strategies to pull down specific RNA and then sequence the chromatin associated with this RNA. Using RAP-seq researchers were able to show the genomic binding loci of Xist, the lncRNA required for silencing X chromosome in mammalian females (Engreitz et al., 2013). Using ChIRP-seq researchers were able to show genomic binding locations of rox2 RNA and human telomerase RNA terc as well as showing the order of RNA-chromatin interactions in the case of HOTAIR (Ilik et al., 2013). Using CHART, researchers have discovered the genomic binding locations of rox2 RNA (Simon et al., 2011). However, these techniques suffer from the obvious drawback that they can be applied to only one RNA molecule at a time. Also, in these assays, researchers need to know the identities of chromatin-associated RNAs. This is a major obstacle as only very few ncRNAs with possible chromatin interacting functions are known. Before this research, there was no reported methods that can identify chromatin interacting RNA and their genomic targets in a global and pairwise manner. This chapter fills this void by developing a technique that can identify the targets of non-coding RNA at a genomic scale.

Methods

An overview of the approach is demonstrated in Figure 3.1

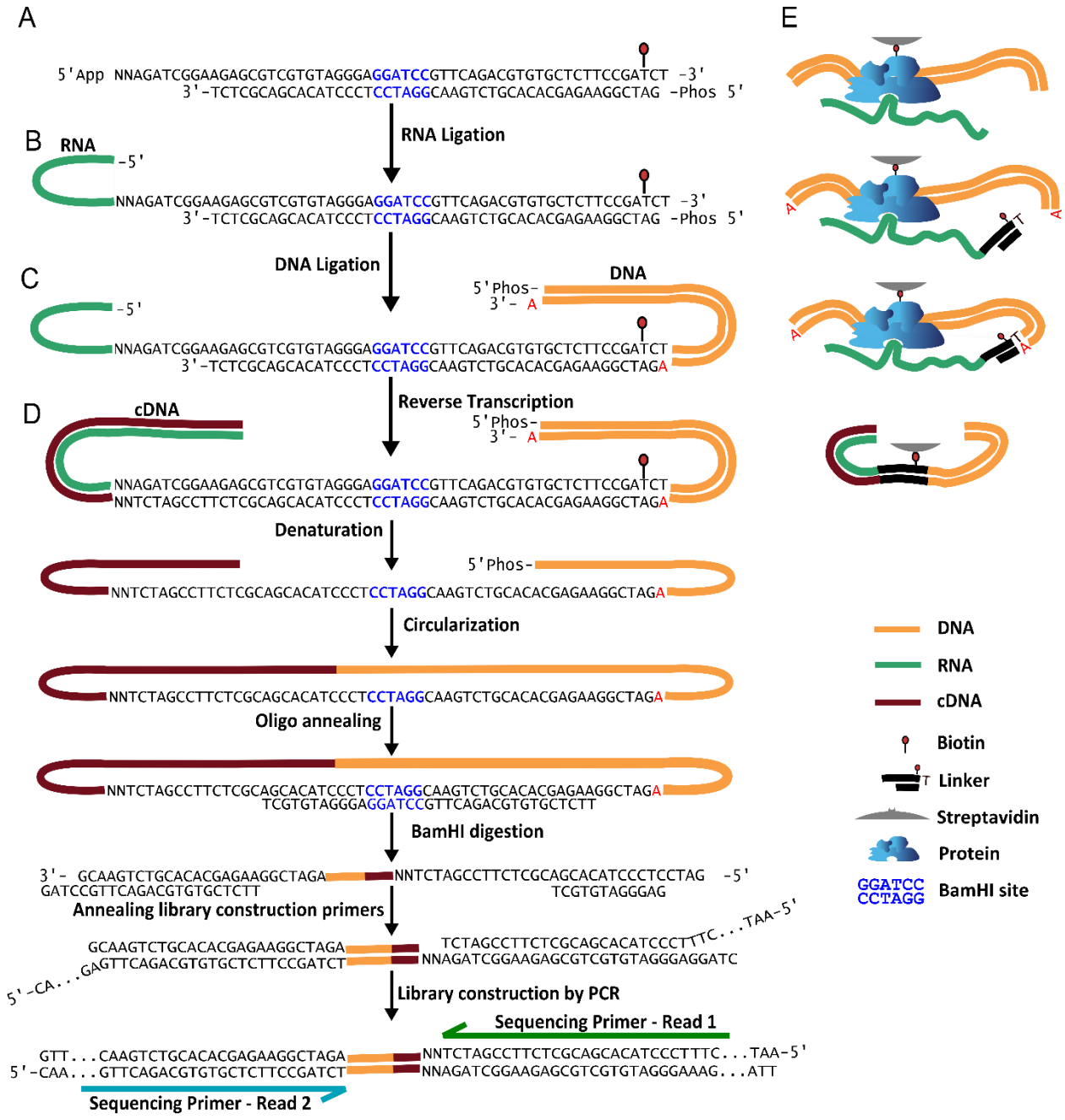


Figure 3.1: ***Overview of MARGI procedure***

The molecular interactions that brings RNAs into contact with DNAs are diverse. We want to maximize the detection of indirect RNA-DNA interactions formed by intermediate proteins. To

achieve this goal, we used a dual crosslinking approach: formaldehyde to crosslink nucleic acids to proteins followed by DSG to crosslink proteins with proteins. RNAs are inherently sticky and are prone to form non-specific interactions with other molecules. We, thus, strictly relied on crosslinked interactions and got rid of any non-crosslinked interactions to ensure that the interactions that we detect are formed in vivo. To this end, we applied stringent and completely denaturing (of proteins) washing and binding buffers to remove non-crosslinked interactions. Similarly to ChIA-PET, ChIP-loop and e4C procedures, we focused only on 'true' complexes of regulatory RNA-DNA interactions joined by protein bridges. We therefore strictly depended on solubilization of crosslinked chromatin fragments and only use the soluble part of the chromatin for downstream analysis. However, there is potentially one caveat as the percentage of such crosslinked complexes is relatively low. Consequently, more PCR amplifications and deep sequencing was necessary to detect RNA-DNA interaction signals.

Step 1: Cell lysis and chromatin sonication

Cell membrane wwas lysed in NP40 Lysis Buffer (10 mM Tris, pH 7.5, 10 mM NaCl, 0.2 % NP40, 1X Protease inhibitor cocktail) to lyse the cell membranes. This step is to destroy the cytoplasmic membranes and isolate the cell nuclei, and wash away most of cytoplasmic materials. Nuclear material was isolated and sonicated. After this, the chromatin was broken up into large DNA fragments.

Step 2: DNA restriction digestion and RNA fragmentation

Following this, DNA from crosslinked chromatin was digested to a smaller size. A restriction enzyme, HaeIII (New England Biolabs) was chosen to perform this task. HaeIII has a 4 base recognition sequence (GGCC) and produces blunt ends. The enzyme digests DNA between

the G and C in the recognition sequence. We also trimmed the size of RNAs by RNase I. The purpose of this step is to reduce the size of DNAs and RNAs into shorter fragments.

Step 3: Purification of nucleic acid-protein complexes and removal of irrelevant, free proteins from the lysate by SILANE beads

Next, 3.5x lysate volume of buffer RLT (Qiagen) was added. The lysate was cleared by centrifuging to pellet down the insoluble materials. The supernatant which contains soluble chromatin (we will call this soluble lysate) was collected and used for subsequent steps.

Step 4: Protein biotinylation and immobilization of proteins onto Streptavidin beads

An equal volume of isopropanol and 3 ml of SILANE beads were added to the soluble lysate. **T**his step is to isolate nucleic acid-protein crosslinks and minimize the amount of irrelevant proteins that are not bound to nucleic acids. The protein-nucleic acid complexes were next protein-biotinylated with NHS-PEG4-Biotin. NHS-PEG4-Biotin biotinylates the amine group on the lysine side chain and the N-termini.

We next remove excess unreacted NHS-PEG4-Biotin also by SILANE beads.

The biotinylated lysate was next incubated with GE streptavidin beads in completely denaturing binding buffer to tether biotinylated protein-nucleic-acid complexes onto the surface of streptavidin beads.

Step 5: Ligation of biotin-tagged linker and proximity ligation

After immobilizing protein-nucleic acid complexes onto streptavidin beads, RNA ligation was performed on (streptavidin) beads between the 3'-end of the tethered RNAs and the 5'-end of biotin-tagged linker. The biotin-tagged linker has two end: the single-stranded DNA 5'-end that will be ligated to the 3'-end of RNAs; and the double stranded DNA 3'-end that will be later proximity-ligated to DNAs. After RNA ligation overnight, streptavidin beads were washed

substantially to remove unligated biotin-linker. The end result is that now most RNAs had been attached with biotin-tagged linker at the 3'-end. DNA proximity ligation was performed under diluted conditions in total reaction volume of 20 mL.

Step 6: Reverse crosslinking and nucleic acid purification

After DNA proximity ligation, proteinase K digestion, reverse crosslinking and nucleic acid purification by SILANE beads was carried out.

Step 7: Removal of terminal biotin from non-proximity ligated linkers and selection of RNA-DNA chimeras

Next, unligated terminal biotin was removed in the sample by making use of the exonuclease activities of T4 DNA polymerase and exonuclease I. The exonuclease activities remove the terminal biotin from un-proximity-ligated biotin-linker. The second streptavidin-biotin pulldown was performed on the sample to select for chimeric RNA-DNA molecules that were not affected by exonuclease activities. Reverse transcription was performed on beads to generate cDNA complement to the RNA sides. The double-stranded DNA-RNA hybrids were denatured by 0.1 M NaOH into single-stranded strand. The strand containing RNA-biotin-DNA still binds to the streptavidin beads while the strand with DNA-cDNA was released. We collected and purified the DNA-cDNA strands.

Step 8: Circularization, re-linearization and PCR amplification

The single-stranded DNAs were then circularized with CircLigase II. A short oligo that complement to the linker region was hybridize to the linker. The linker region was designed to contain a recognition site for BamHI. The circular ssDNA was relinearized by digestion with BamHI. The relinearized products were amplified by PCR.

Results

pxMARGI data demonstrated that the protocol selectively enriches chimeric fragments in the form of linker-RNA-DNA-linker

When we designed the procedure, two models of RNA-DNA interactions were considered. In the chromatin-cage model, chromatin-associated RNA (caRNAs) are trapped in a dense mesh of highly-crosslinked chromatin (Cai et al., 2003). In a direct-interaction model, specific protein or protein complex may mediate caRNA's interactions with specific genomic locations (Engreitz et al., 2013; Licatalosi et al., 2008; Rinn and Chang, 2012).

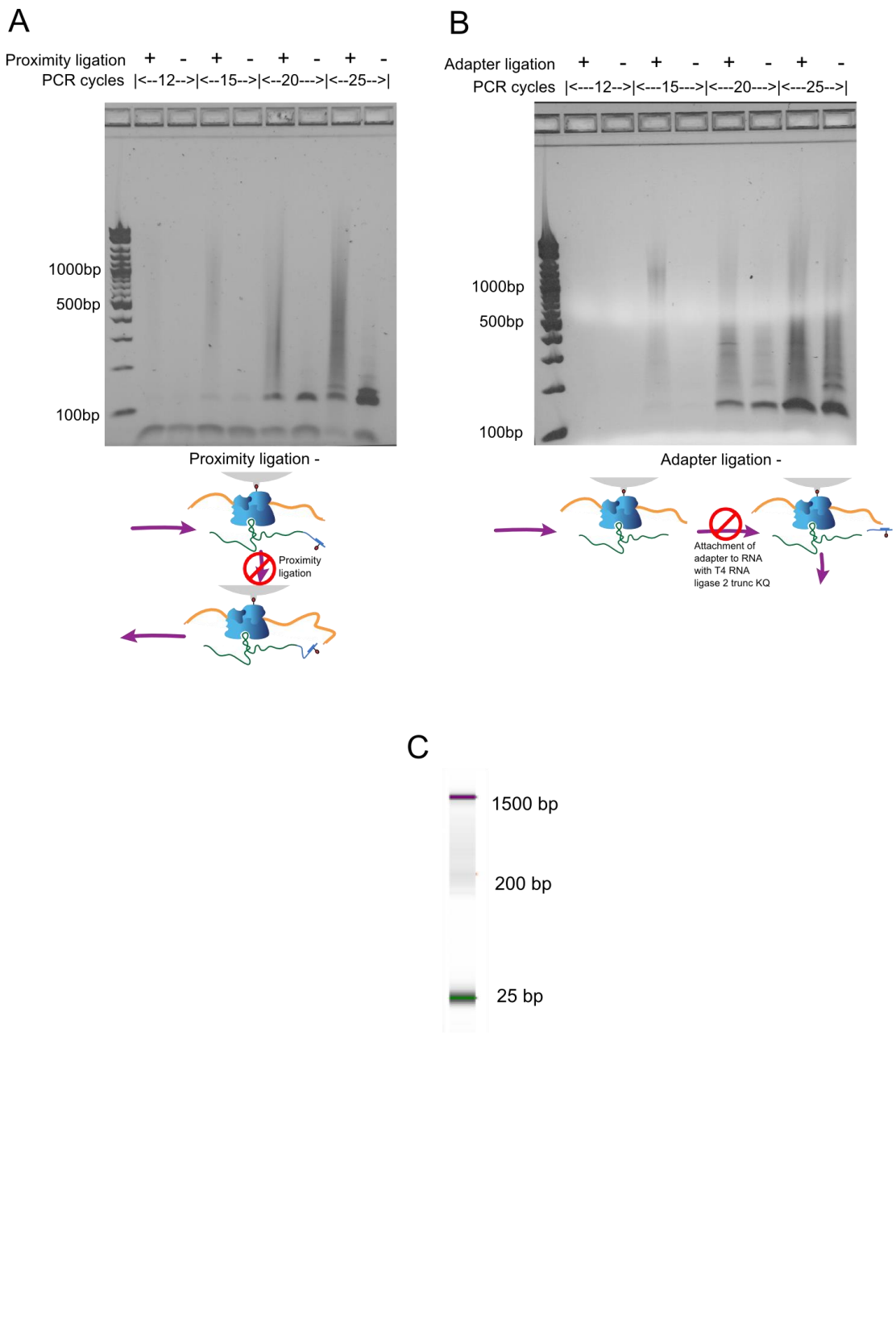
Two versions of the protocol were developed. The first version, namely proximity MARGI (pxMARGI) was developed to detect chromatin cage model of RNA-chromatin interactions. The second version, called direct-interaction MARGI (diMARGI), was tailored for the direct interaction model. pxMARGI procedure was developed by Bharat Sridhar, a graduate student from our lab, while I developed diMARGI procedure.

pxMARGI results generated by Bharat were used to validate the feasibility of our approach. Mouse E14 cells were used to generate pxMARGI libraries. Data were reproduced in three biological replicates. Two controls were performed. In one case ligation of adapter to RNA was not performed and in the other control, proximity ligation was not performed. In the absence of adapter ligation to RNA, the adapter will be washed away prior to proximity ligation. Thus, no library should be generated from this. In the absence of proximity ligation, adapter is ligated to RNA but exo I and T4 DNA polymerase are used to remove the biotin from constructs where adapter only ligates to RNA. T4 DNA polymerase has 5' $\rightarrow$ 3' DNA polymerase activity and 3'

→ 5’ exonuclease activity. The polymerase activity is far more processive than the exonuclease activity. Thus the exonuclease can access the biotinylated nucleotide only if it is close to the 3’ end. This would be true only in the case where the adapter did not ligate to DNA. Since the biotin is clipped from these molecules we will not be able to pull these molecules down using Streptavidin. Thus, a control where proximity ligation was not performed would also not be also to generate libraries as we would not be able to pull adapter containing material.

As expected, only the actual sample was able to generate material that could be sequenced. Both the controls failed to generate sufficient material to sequence. As shown in Figure 3.2, pulled down material was used to generate sequencing libraries by performing PCR. In the actual sample, enough material to load onto the sequencer was generated by cycle 15. With increasing cycle numbers amount of library produced also increased. In the case of the control even at high cycle numbers very little library was produced. Any material obtained after library generation had a band like pattern that most likely was an artifact of just the adapter getting amplified.

Figure 3.2*: **Library was generated from E14 cells and sent for sequencing***. Two controls were also performed. Material was amplified by PCR to generate the library. A. Library generated from RDI-seq sample and a proximity ligation control was run on a gel after PCR of different cycle numbers. RDI-seq sample produced increasing amount of material with increasing cycle numbers. Control didn't produce library that could be sequenced even at high cycle numbers. B. Library generated from an independent biological replicate of RDI-seq and a control where adapter was not ligated to RNA was run on a gel after PCR of different cycles. RDI-seq again produced increasing amounts of library with increasing cycle number. Control did not produce at lower cycle numbers. At high cycle samples some material was obtained but they had a band like pattern, which is most likely an artifact of the adapter amplification. C. Gel like image of sample sent for sequencing generated using tape station for E14 biological replicate 1.


A
Proximity ligation + - + - + - + -
PCR cycles |<--12-->|<--15-->|<--20--->|<--25-->|
1000bp
500bp
100bp
Proximity ligation -
Proximity ligation
B
Adapter ligation + - + - + - + -
PCR cycles |<---12-->|<---15--->|<---20--->|<---25-->|
1000bp
500bp
100bp
Adapter ligation -
Attachment of adapter to RNA with T4 RNA ligase 2 trunc KQ
C
1500 bp
200 bp
25 bp

First, the distribution of bases at various positions of the reads was analyzed. As per the design of the adapter, the DNA end should begin with restriction enzyme signature GG. The rest of the DNA should have a regular distribution of the bases. The entire RNA read should have a regular distribution. This is exactly what was observed (Figure 3.3).

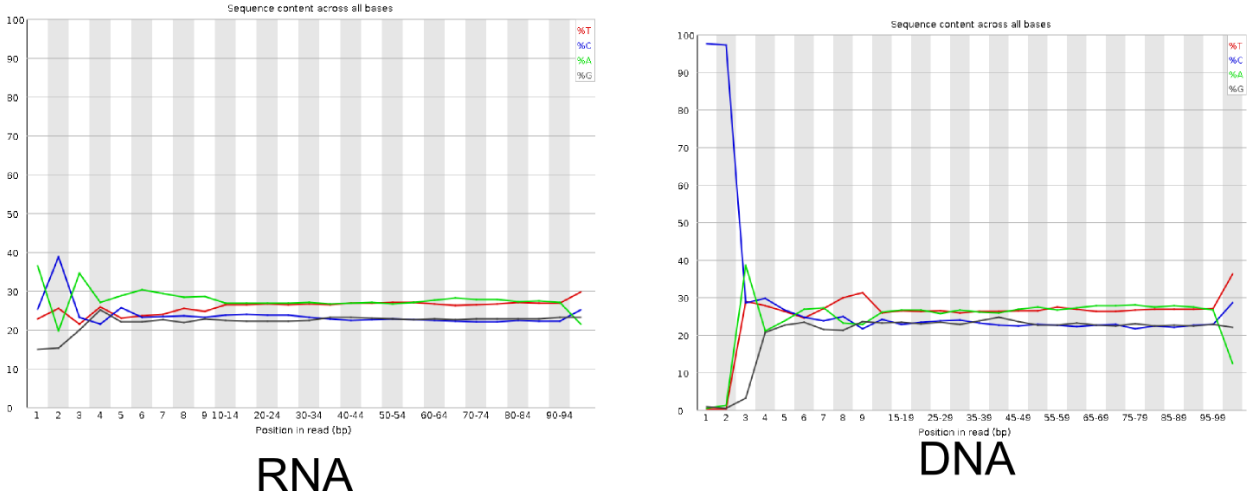


Figure 3.3*: Percentage of different bases at different read positions* for read 1(RNA) and read 2 (DNA). Almost 100% of the first two bases of DNA are GG (the restriction enzyme digestion signature)

diMARGI generated millions of reads that can be parsed to identify RNAs interacting with the chromatin in multiple cell lines

We generated one diMARGI library from HEK and one library from human ES cells and one from mouse ES cells. Paired-end reads from chimeric fragments were classified based on the possible configurations of which the two paired reads were mapped. They can be properly mapped, that is, when aligned to the same chromosome the left-most and right-most mates are aligned to

the positive and reverse strands, and with a distance between their outer-most coordinates not farther apart than the length used for fragment selection (**Figure 3.4**).

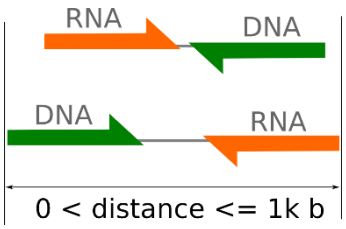


Figure 3.4*: Graphical depiction of the two scenarios were properly paired reads*. A pair of reads was deemed as properly paired only if (1) the left and right mates were aligned to the positive and reverse strands, respectively, and (2) their outer-most coordinates were not farther apart than the size of the fragment selection (1000 bases in this example).

However, as there is no restriction on how RNA and DNA mates are ligated, they can also end up in other non-traditional alignment configurations. For instance, both mates may align to different chromosomes, or –when aligned to the same chromosome– both mates align to the same strand or have distances between mates several times the size of the length of the fragment selection. It may also be the case that contamination of non-chimeric fragments (only RNA or DNA) show up in the sequencing results. Non-chimeric fragments, however, can only be aligned to the reference genome as properly mapped reads. We used this restricted alignment scheme of non-chimeric fragments to distinguish them from truly chimeric fragments. Specifically, we categorized the mapped read-pairs as:

- Proper: These are read-pairs properly aligned but where only the DNA mates start with CC at the 5' end and that have a distance between their outer-most coordinates lower than 1000 bases.
- Proximal: These are read-pairs not properly aligned, where the distance between their outer-most coordinates is lower than 1000 bases.

- Distal: These are read-pairs where both mates are mapped to the same chromosome and the distance between their outer-most coordinates is larger 1000 bases.
- Inter-chromosomal: These are reads-pairs where mates are aligned to different chromosomes.

The proximal read pairs constituted 94% and 95% in HEK and ES cells, respectively, distal pairs constituted 1% in both cell types, inter-chromosomal pairs 5% and 4%.

Using mm10 as reference genome for mouse, the results (**Table 3.1**) show that all replicates had a high alignment rate.

Table 3.1: Number of mapped reads and the different types of mapped reads

Sample	Protocol	Number of mapped pairs	Number of proximal pairs	Number of distal pairs	Number of inter-chromosomal pairs
E14 Biological replicate 1	pxMARGI	62,402,150	39,399,542	608,058	2,249,270
E14 Biological replicate 2	pxMARGI	31,201,075	15,453,443	642,280	5,979,803
E14 biological replicate 3	pxMARGI	23,087,666	21,235,241	74,989	416,502
E14	diMARGI	4,183,169	3,331,533	44,555	787,081

Besides the mouse ES cell line, we applied the MARGI seq protocol in two human cell lines. We hypothesized that the interactions between some of the RNAs and chromatin in embryonic stem cell and differentiated cells will be different. We therefore chose human H9 ES cells and HEK 293T as the human two cell types we wished to analyze. In this chapter I will provide you with the analysis of the data that we generated in these three cell lines. I will also provide a discussion of these results and some of the similarities and differences observed in the interactions between DNA and RNA in a differentiated cell type versus embryonic stem cells. We performed two replicates of pxMARGI for both the cell lines and performed one replicate each for the diMARGI protocol. The samples were mapped and classified as before into three classes – proximal, distal and inter chromosomal. The reads of each class have been summarized in Table 3.2 below:

Table 3.2: Read classification for human cells

Sample ID	Cell type	Technology	Mapped read pairs			
			Total	Proximal	Distal	Inter-chromosomal
1	HEK293	pxMARGI	45,187,015	35,307,650 (78%)	722,819 (2%)	9,156,546 (20%)
2	HEK293	pxMARGI	61,390,133	51,238,202 (83%)	971,270 (2%)	9,180,661 (15%)
3	HEK293	diMARGI	9,606,682	9,006,803 (94%)	78,134 (1%)	521,745 (5%)
4	H9	pxMARGI	29,774,645	24,015,455 (81%)	444,294 (1%)	5,314,896 (18%)
5	H9	pxMARGI	35,899,884	29,330,169 (82%)	1,414,592 (4%)	5,155,123 (14%)
6	H9	diMARGI	4,682,327	4,470,815 (95%)	42,479 (1%)	169,033 (4%)

As can be seen in the case of pxMARGI about 15% of interactions were observed to be long range (distal and proximal) for both cell types. In the case of diMARGI this number reduced to about 5%. Thus, using both methods we were able to detect a non-trivial amount of long-range interactions.

We then checked if the pxMARGI data still reflected the restriction enzyme signature in the pxMARGI protocol for the human cells. We mapped the base composition per base for the human cell types. As expected in pxMARGI almost all the reads had GG in the first two bases

whereas in diMARGI the distribution was random (Figure 3.5)

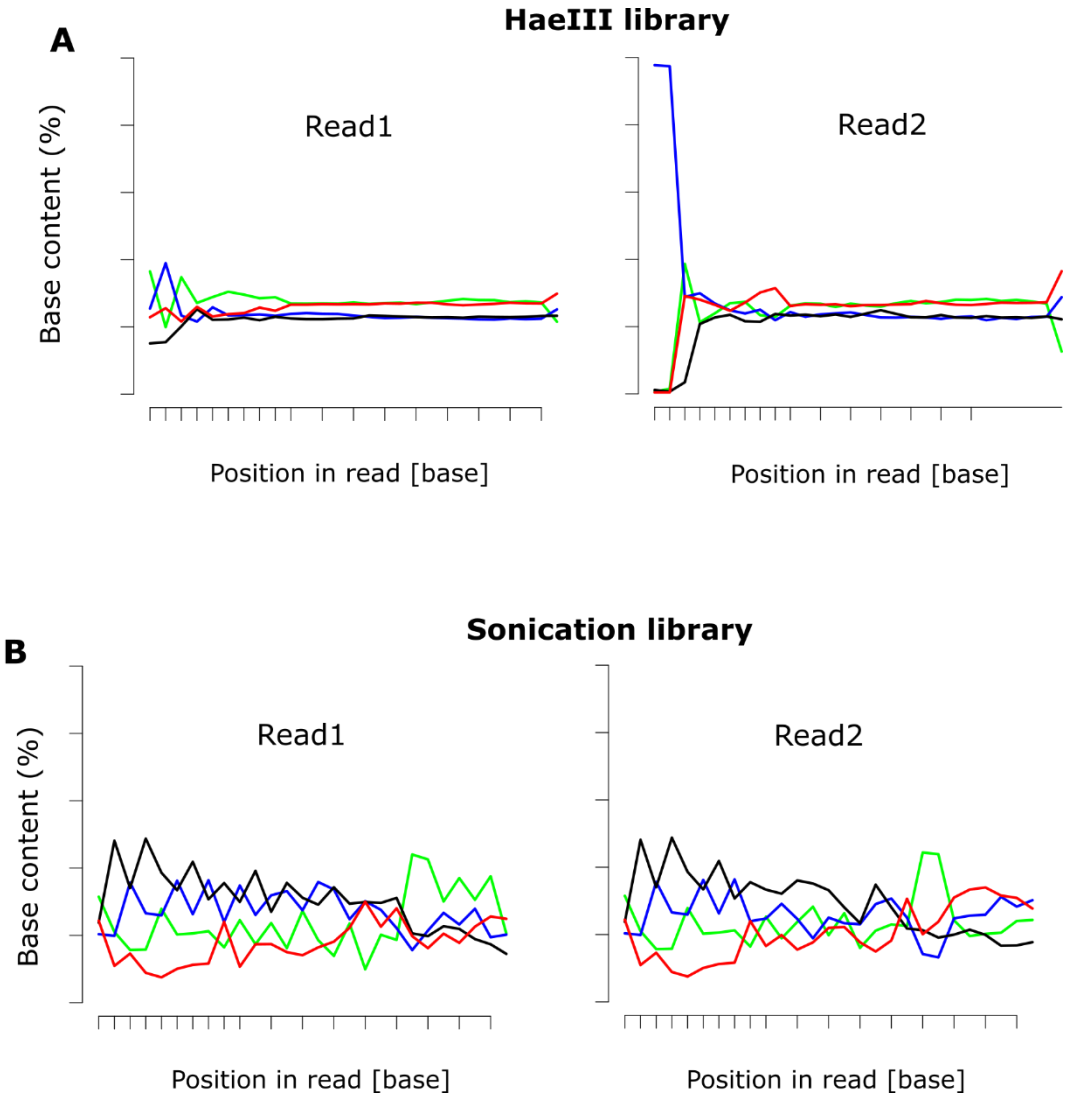


Figure 3.5*: Distribution of the four bases at each position*; C(Blue), G(black), T(red) and A(green) in libraries generated in 293 T cells using restriction enzyme (A) and sonication (B). In the HaeIII library the first two bases consist entirely of CC.

Once this was observed we proceeded with the analysis of the caRNA. For the purpose of this analysis we prioritized the long-range interactions. Short range interactions could have multiple sources. They could be nascent RNA currently undergoing transcription. There could be some artifacts of the sequencing procedure whereby just the DNA or just the RNA was sequenced. In order to avoid this, we prioritized the long-range interactions.

Identification of caRNAs

The RPKM (Reads per kilobase per million reads) was calculated for the read 1 to identify which RNAs could be classified as caRNAs. If the RPKM of a RNA was > 0 it was considered to be a caRNA. caRNAs identified by pxMARGI are called as pxRNA while those identified from diMARGI protocol were called as diRNA. The total number of non-coding pxRNAs and diRNAs (with a FDR < 0.0001) identified have been provided in Table 3.3.

Table 3.3: Number of non coding pxRNAs and diRNAs

Cell type	Number of pxRNAs	Number of diRNAs
HEK 293 T	2864	747
H9 Human ES	1993	467
E14 Mouse ES	1541	523

Since the nature of interactions being detected by pxMARGI and diMARGI are different we expected the RNAs to be detected to also be different. We performed a clustering analysis to check if this was indeed the case. As can be seen in Figure 3.6 the proximity and direct samples clustered separately.

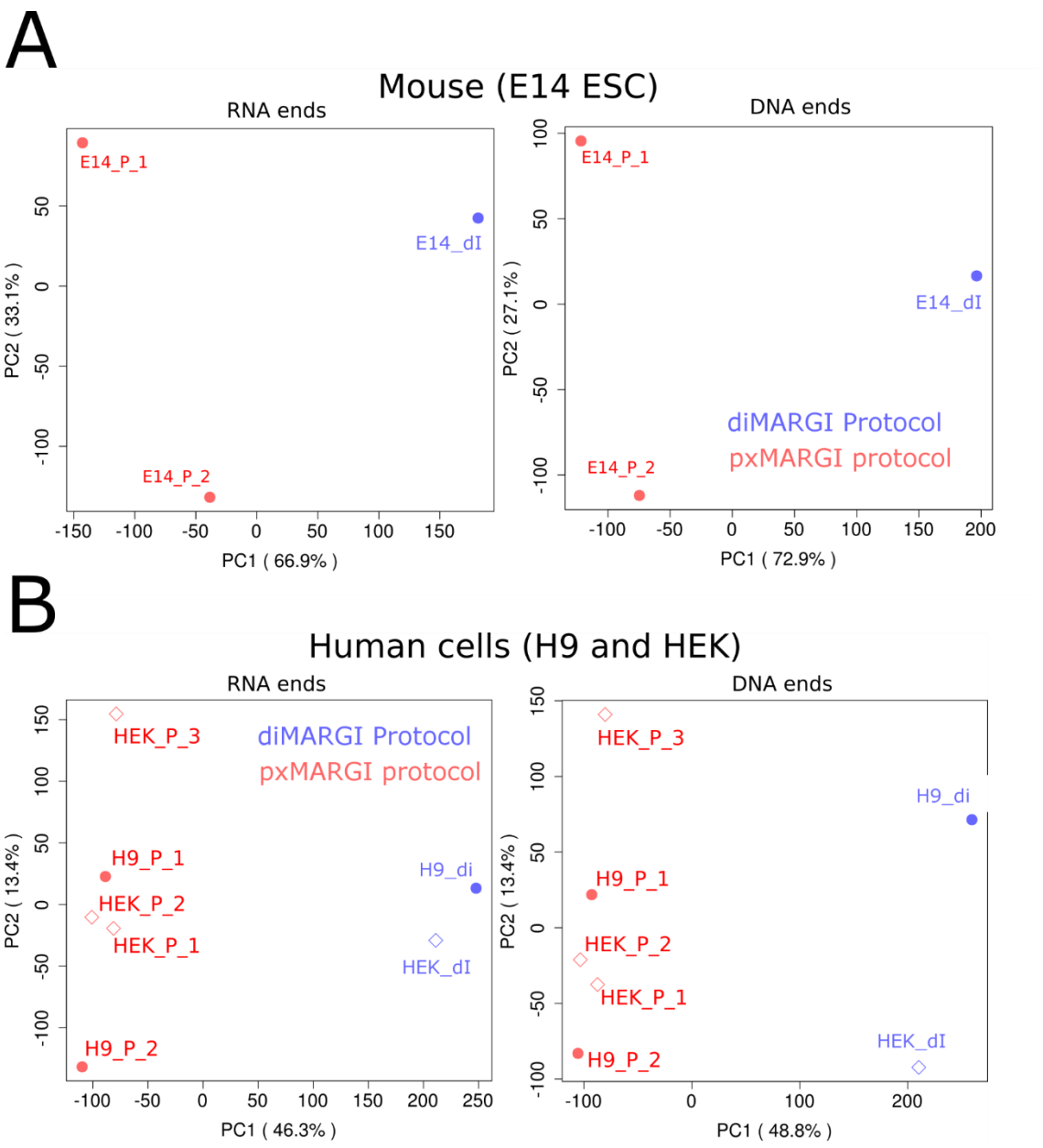


Figure 3.6*: Differences in enrichment of RNA and DNA reads among mmouse E14* (A) and human H9 and HEK cell lines (B) as a result of using proximity and direct MARGI protocols. For each library, we computed genes' enrichment of either RNA or DNA reads (RPKM), discarding genes with zero values among all libraries. Plotting their first two principal components, it can be clearly observed that libraries clustered by protocol.

We then classified the non-coding pxRNAs and diRNAs by biotype. The largest components of non-coding pxRNAs were pseudogene, antisense, and lincRNA, which was the

same for all the cell types (Figure 3.7 A). The largest component of diRNAs in both human cell types was snoRNA (Figure 3.7 B), consistent with their known activities for modification of nascent target transcripts. The other categories represented in the HEK cells include snRNA and antisense RNA. Among the H9 cells the other categories represented include miRNA and snRNA. Among the mouse ES cells the largest represented category was the snRNA. The other categories are snoRNA and miRNA. All these biotypes have known functions in the nucleus.

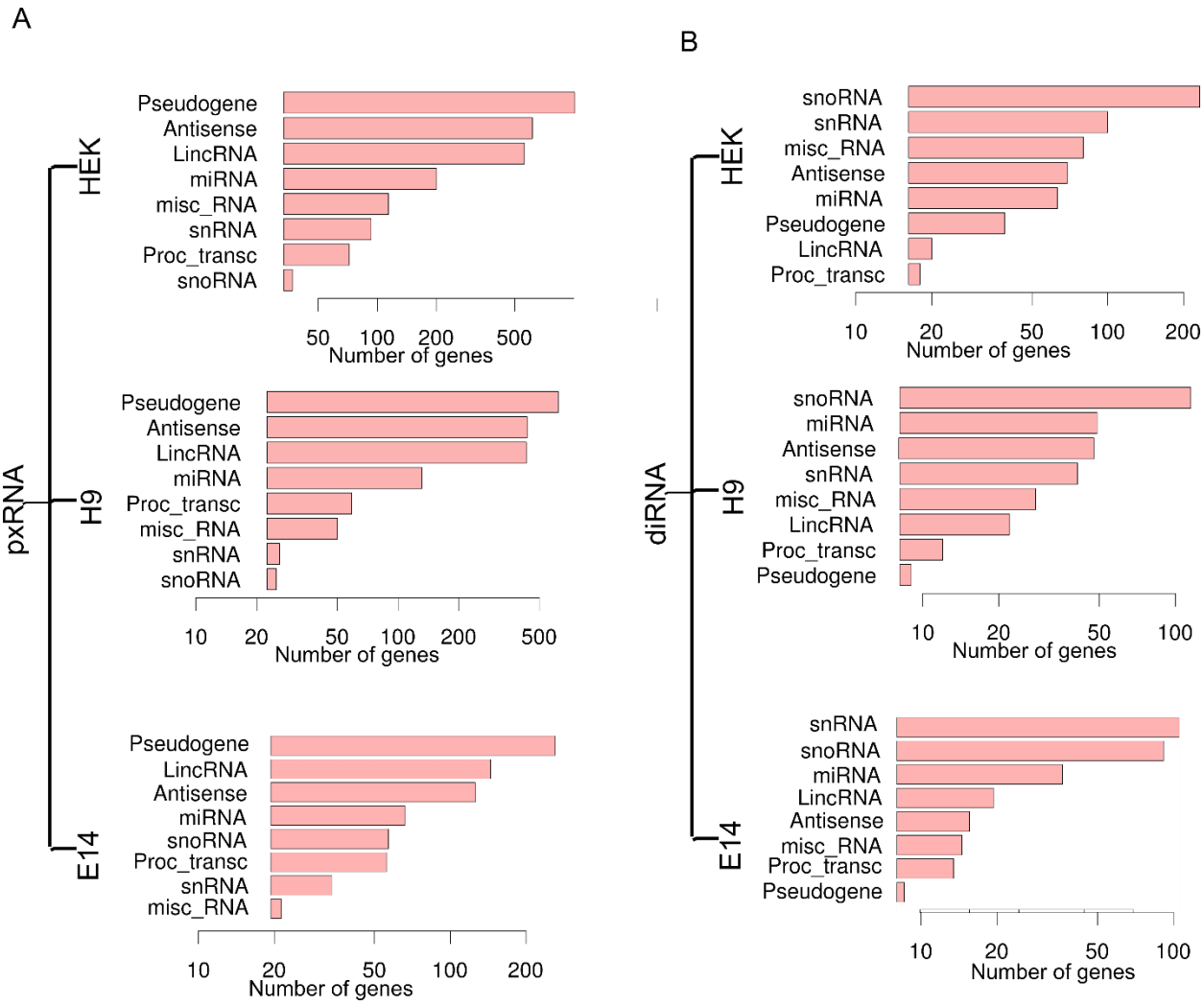


Figure 3.7*: Classification of non-coding pxRNAs(A) and diRNAs (B)* in HEK, H9 and E14 cells. Pseudogenes, antisense and lincRNA were the top categories among the pxRNAs while snoRNA, snRNA and miRNA were among the top categories in the diRNAs.

We then looked for specific examples of RNA with known nuclear locations and functions. We concentrated on human cell types for this analysis. The first RNA we looked at was XIST. We detected XIST in both H9 and HEK cells among pxRNAs (Figure 3.8A). XIST expression in ES cells is unstable and naïve ES cells have both X chromosomes active. Human ES cells are primed

pluripotent cells. X inactivation begins at this stage but XIST transcription is unstable at this stage. Thus the presence of XIST in the pxRNA data is consistent with the reported findings about XIST. For the diMARGI data, we noticed that the ES cells had low signal for XIST but among HEK cells, XIST was among the highest ranked RNAs (Figure 3.8B). This indicates that in differentiated female cells, XIST is indeed among the most important RNAs. But in ES cells these interactions haven't yet fully developed yet.

We then looked at other knows RNAs. We found other RNAs including MALAT1, NEAT1 and SNHG1. These were identified as caRNAs in pxMARGI in both cell types. MALAT1, SNHG1, NEAT1 became even more significant in diMARGI, and MALAT1 and SNHG1 rose into the few most significant diRNAs in both cell types. The other RNA we looked at were the 7SK snRNA in the HEK cells. This was also enriched and I will discuss it further in the genomic targets of the caRNA section.

All the RNAs identified by the pxMARGI and diMARGI techniques have been listed in Appendix A.

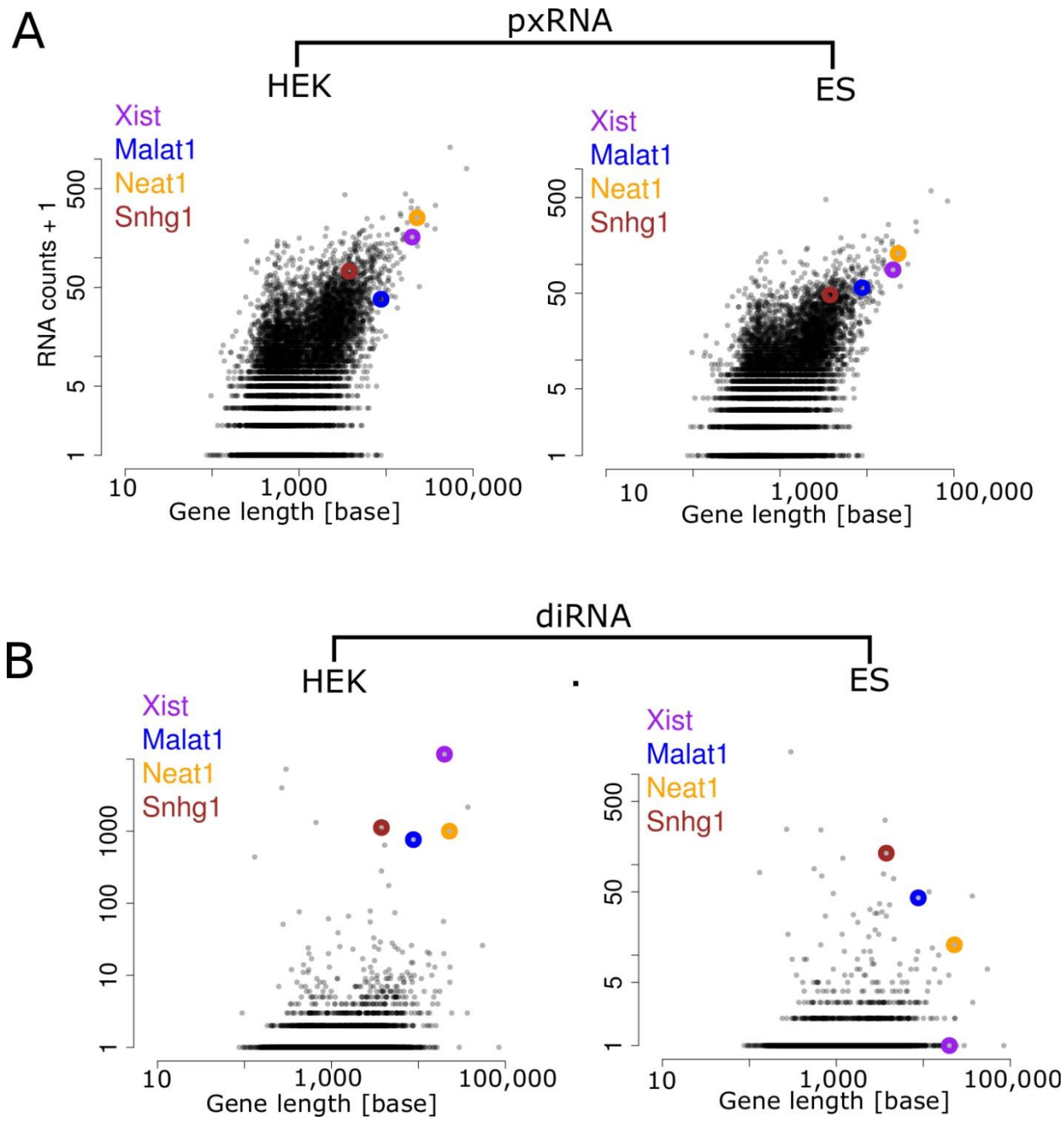


Figure 3.8*: RNA counts of all the pxRNAs (A) and diRNAs (B) identified in our data*. Known caRNAs XIST, MALAT1, NEAT1 and SNHG1 are highlighted.

Genomic targets of caRNA

We then began analyzing the DNA targets of the caRNAs. We focused on human cells for analysis on the DNA ends.

We called peaks from the DNA ends of the pxMARGI and diMARGI data using MACS v1.4.2 which we call pxPeaks and diPeaks respectively. As expected pxpeaks were larger in number than diPeaks. Also pxPeaks were on average larger than the diPeaks. (Figure 3.9A). HEK cells yielded 120,872 pxPeaks, whilH9 ES cells yielded 57,154 pxPeaks. In comparison, the amount of diPeaks from HEK (7,212) remains larger than that from H9 (5,247), but the difference was not as great as that in pxPeaks. This is reminiscent of the idea that pxRNA may be trapped in closed chromatin, because stem cell differentiation is usually coupled with chromatin condensation.

We then assessed if the genomic targets were associated with genomic features. We used the Upset tool to identify if DNA peaks overlapped with genomic features including promoters, 5' UTR, 3' UTR, exons, introns, downstream sequence (3 kb), and intergenic sequence, while accounting for overlaps to multiple genomic features. In HEK cells, approximately 37% of pxPeaks overlapped with intergenic regions and 17% overlapped with promoters (Figure 15B). Adjusting for the sizes of these genomic features, pxPeaks were enriched in promoters (Odds ratio = 1.7, p-value $< 2\times10^{-16}$, Chi-squared test). H9 cells exhibited very similar proportions, and an enrichment in promoters (p-value $< 2\times10^{-16}$) .

The overlaps of diPeaks to promoters increased to 61% (4,391) in HEK and 63% (3,306) in ES cells (Figure 15B), and the odds ratios for these overlaps increased to 16.8 (HEK, p-value < 2×10^{-16}, Chi-squared test) and 17.7 (ES, p-value < 2×10^{-16}, Chi-squared test) (Figure 15B).

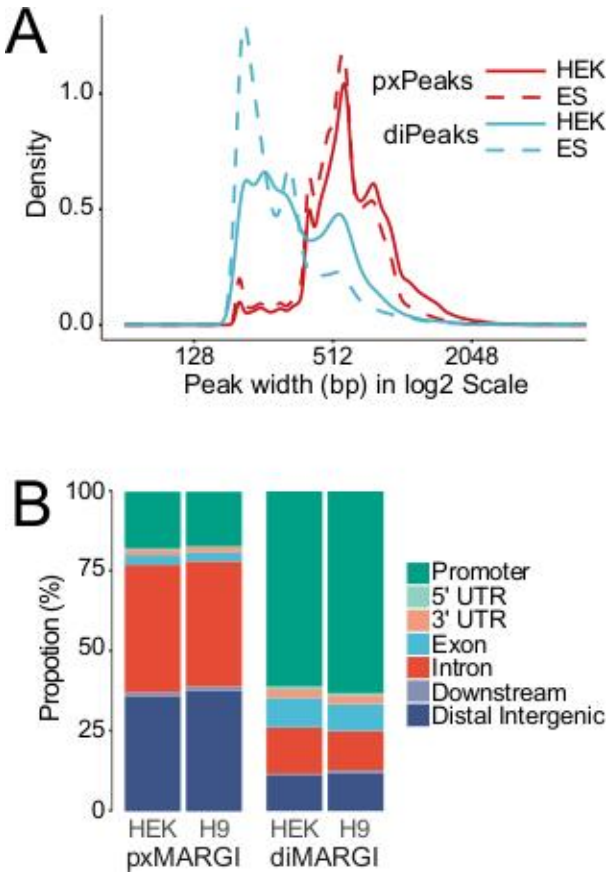


Figure 3.9*: Size distribution of pxPeaks(Red) and diPeaks(blues)* in HEK (solid lines) and H9(dashed lines). The size of pxPeaks was on average larger than the size of diPeaks. Proportion of pxPeaks and diPeaks overlapping with known genomic features as computed by the Upset tool. A large portion of the diPEAKS resided in promoters. The enrichment of pxPeaks in promoters was also significant.

Targets of specific caRNAs

We then analyzed the targets of specific caRNAs. Among the diRNAs originating from the X chromosome in HEK 293T cells XIST was the biggest hit. A large portion of the targets of XIST,

94 % were on the X chromosome as expected with its function in X inactivation. These targets were spread across the X chromosome (Figure 3.10 A).

The largest targets of MALAT1 and NEAT1 are each other respectively. Analyzing the DNA ends of these RNAs was consistent with this (Figure 3.10B). Similarly, a relationship was observed between SNHG1 and RNU2-2P (Figure 3.10 C).

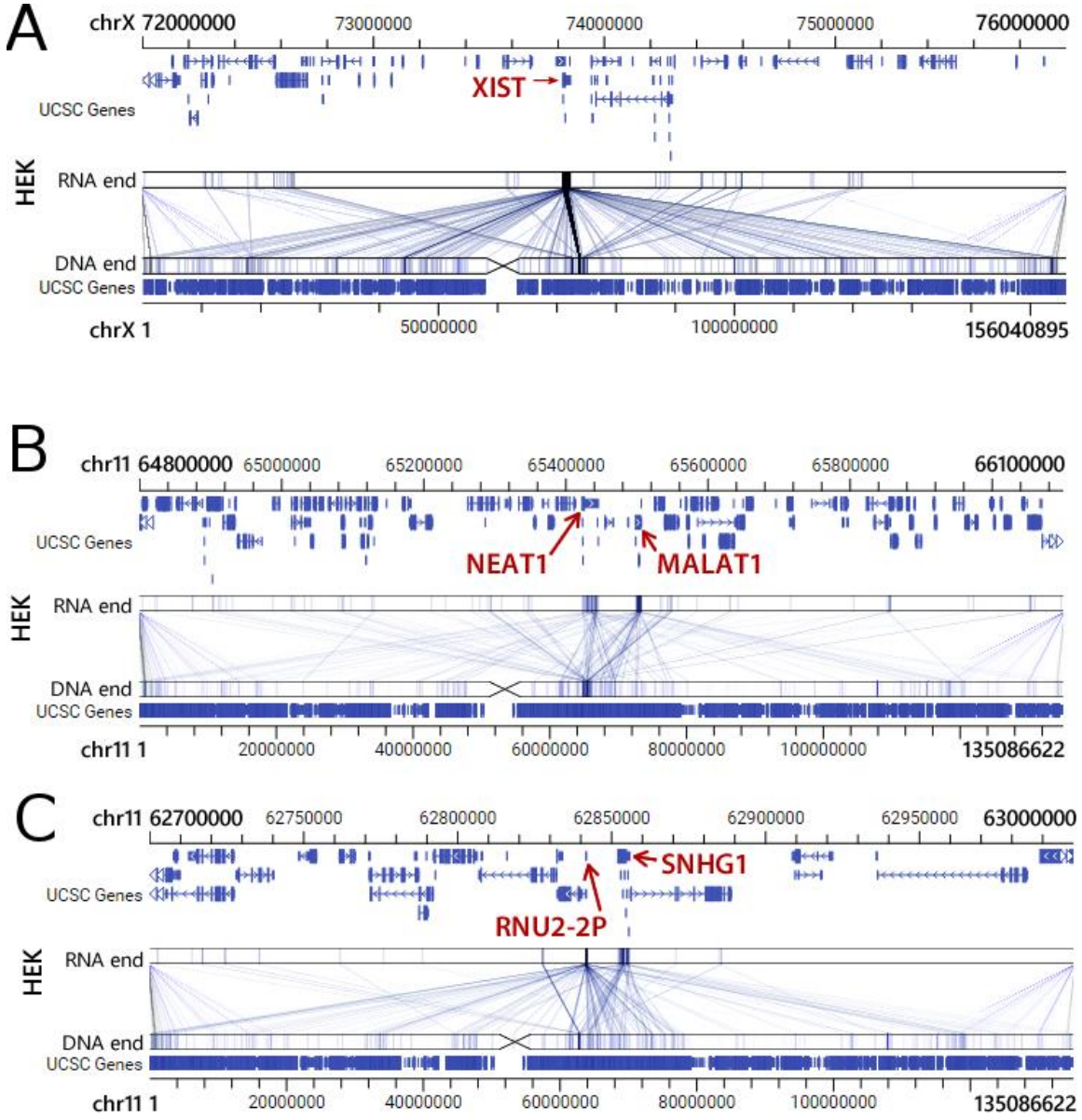


Figure 3.10*: The mapped MARGI reads are plotted with Genome Interaction Visualizer* (GIVe, http://give.genemo.org/?hg38), where the reference genome is plotted horizontally, twice (top and bottom bars). The top and the bottom bars can be zoomed in or out independent of each other. The mapped RNA ends are shown on the top bar (genome), and the DNA ends are shown on the bottom bar (genome). Each MARGI read pair is represented as a line linking the locations of RNA end (top) and the DNA end (bottom). The HEK diMARGI data at are shown with the *XIST* locus versus the entire Chromosome X (A), *MALAT1* locus (top) versus the entire Chromosome 11 (bottom) (B), and the *SNHG1* locus versus the entire Chromosome 11 (C)

We also analyzed the targets of 7SK snRNA. We tried to look at the overlaps of the DNA ends of diMARGI with ChIRP-seq data available for 7SK snRNA in 293T cells (Figure 3.11A). About 12 % of the diMARGI data overlapped with ChIRP-seq data (Figure 3.11B). 91 of 8472 ChIRP seq peaks were represented. This represented a p value of less than 1e -9

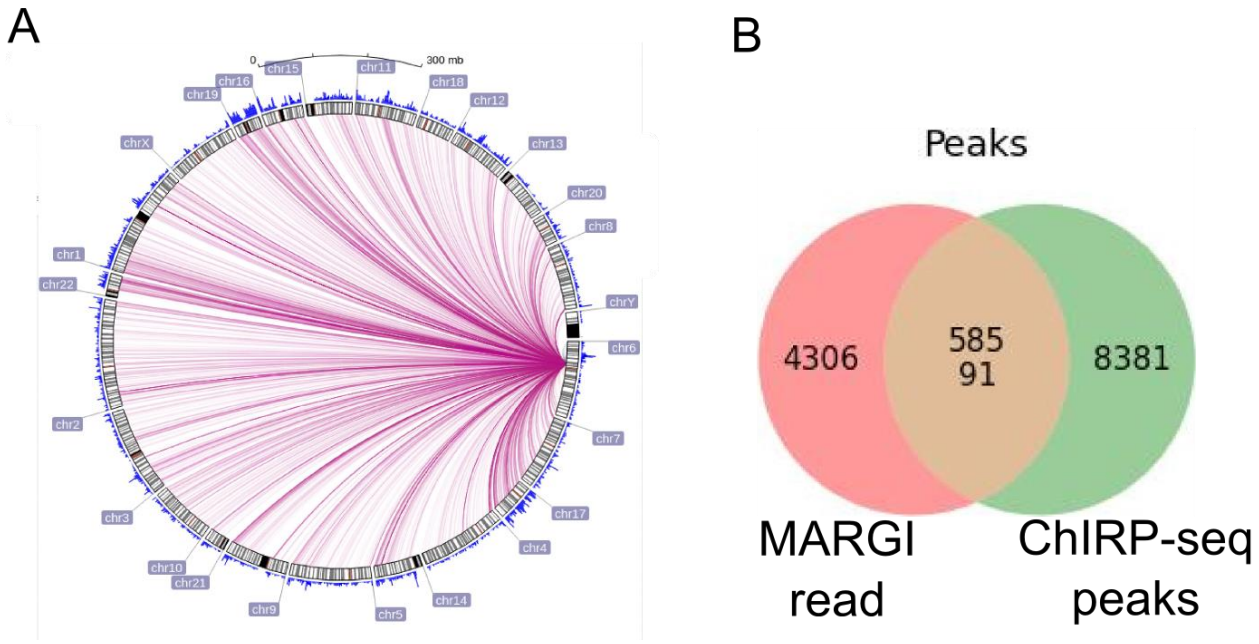


Figure 3.11*: A Circos representation of targets of 7sK*(encoded in chromosome 6) but has targets throughout the genome. B Venn diagram showing overlap of MARGI reads with ChIRP seq peaks.

Co-relation of the genomic targets with histone modifications

The genome wide observations led us to directly assess the degree of association between promoters and pxRNA. We plotted the density of pxRNA across the 20,000 bp flanking regions of every TSS (Figure 3.12A). Out of 34,475 human genes (GRCh38) with non-redundant TSSs, 23,838 (69.1%) exhibited increased pxRNA intensities at their TSSs in HEK cells. Even more TSSs (25,392, 73.7%) exhibited increased pxRNA intensity in H9 cells (Figure 3.12A).

Similar to pxRNA, diRNA intensities increased in promoters, but became more concentrated; forming sharp peaks centered at TSSs (Figure 3.12B). A total of 18,135 TSSs in HEK and 6,551 TSSs in ES exhibited clear increases of diRNA attachments

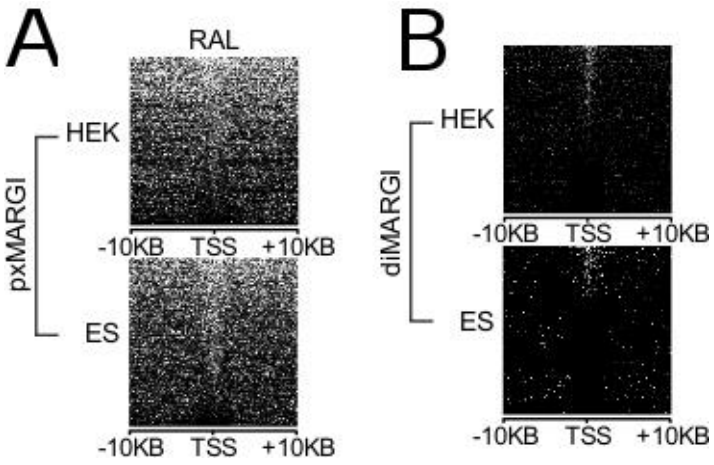


Figure 3.12*: RNA density in a 20KB region surrounding TSS of every gene*. An increase in RNA attachment at the TSS was observed

We asked whether caRNA intensity is correlated with ChIP-seq defined histone modification levels. To this end, we calculated RAL (RNA attachment level) for each genomic segment as the average read count of the DNA-end of distal and inter-chromosomal read pairs. Proximal read pairs were excluded from RAL calculation. Across all TSSs, RAL exhibited positive correlations with H3K4me3, H3K27ac, and negative correlation with H3K9me3 (Figure 3.13 A). These correlations were preserved in the datasets generated by pxMARGI and diMARGI.

We proceeded to analyze the entire genome by scanning the genome with 1,000 bp windows. diRNA RAL exhibited positive correlations with H3K4me3 and H3K27ac, and a negative correlation with H3K9me3 (Figure 3.13B). In comparison, pxRNA RAL did not exhibit clear genome-wide correlations to H3K4me3 and H3K27ac (Figure 3.13B), possibly attributable

to lack of H3K4me3 and H3K27ac in condensed chromatin. Interestingly, pxRNA RAL retained genome-wide anti-correlation with H3K9me3 (pxMARGI, Figure 3.13B). Moreover, H3K9me3 was depleted in nearly all diPeaks and all pxPeaks (Figure 3.13C), suggesting a competition between RNA attachment and H3K9me3 event in closed chromatin.

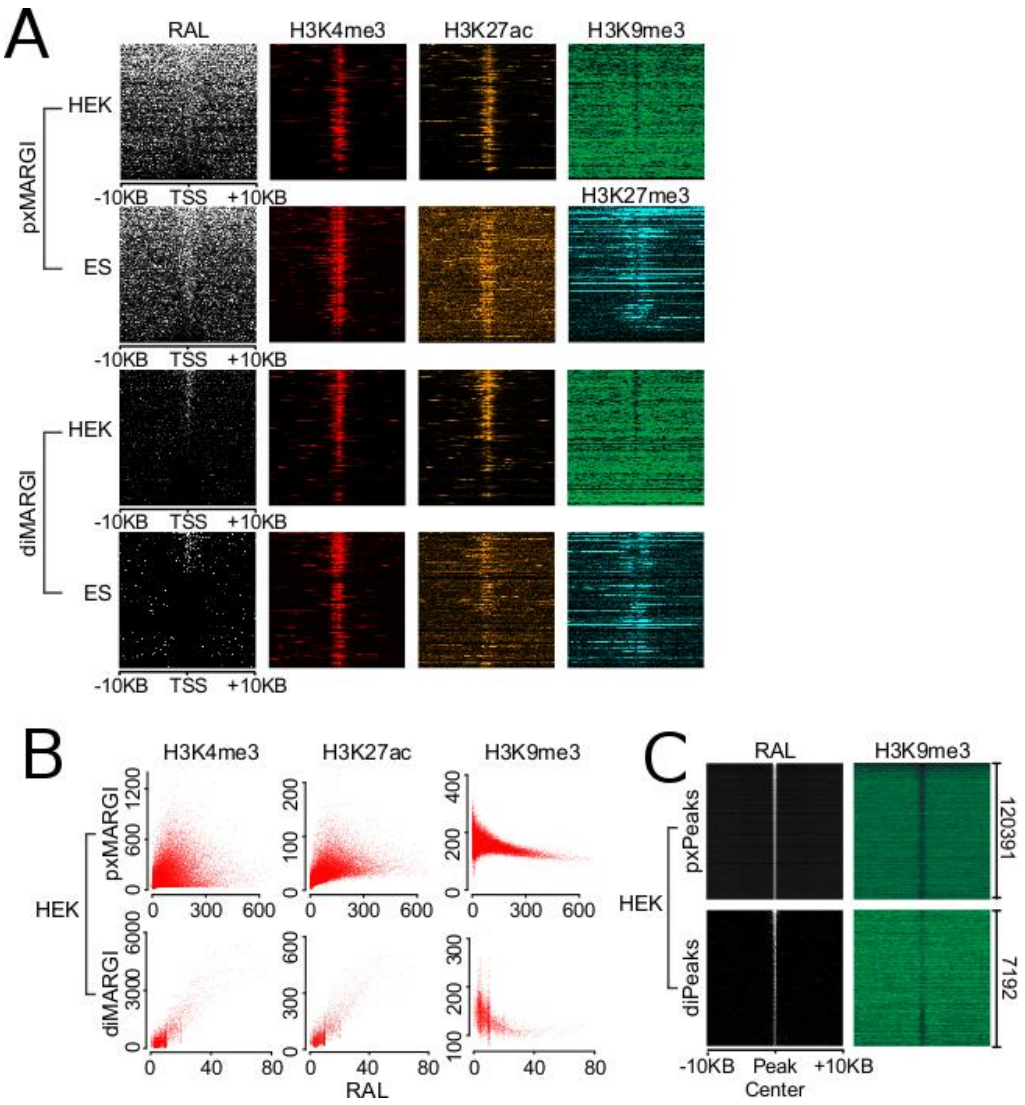


Figure 3.13*: H3K4me3, H3K27ac, H3K9me3* (HEK only, no available data in H9), and H3K27me3 (ES only), are shown in parallel to the RAL in 20 kB region flanking the TSS. (B) Scatter plots of 1,000 bp genomic windows with histone modification levels (y axis) versus RALs (x axis). (C) RAL were plotted for all identified pxPeaks and diPeaks and their 20,000 flanking regions. Shown in parallel are H3K9me3 levels on the same regions (green).

Discussion

In this chapter, we present MARGI (mapping RNA-genome interactions), a technology to massively reveal native RNA-chromatin interactions from unperturbed cells. The gist of this technology is to ligate chromatin-associated RNAs (caRNAs) with their target genomic sequences by proximity ligation, forming RNA-DNA chimeric sequences, which are converted to a sequencing library for paired-end sequencing. Using MARGI, we produced RNA-genome interaction maps for human embryonic stem cells (ESCs) and human embryonic kidney (HEK) cells. MARGI revealed hundreds of caRNAs, including previously known XIST, SNHG1, NEAT1, and MALAT1, as well as each caRNA's genomic interaction loci. Using a cross-species experiment, we estimated that approximately 2.2% of MARGI-identified interactions were false positives. In ESCs and HEK cells, the RNA ends of more than 5% of MARGI read pairs were mapped to distal or inter-chromosomal locations as compared to the locations of their corresponding DNA ends. The majority of transcription start sites are associated with distal or inter-chromosomal caRNAs. Chromatin-immunoprecipitation-sequencing (ChIP-seq)-reported H3K27ac and H3K4me3 levels are positively correlated, while H3K9me3 is negatively correlated, with MARGI-reported RNA attachment levels. The MARGI technology should facilitate revealing novel RNA functions and their genomic target regions.

Acknowledgements

Chapter 3, in full, is an adaptation of materials that appears in Sridhar, Bharat*, Rivas-Astroza, Marcelo*, Nguyen, Tri C.*, Chen, Weizhong, Yan, Zhangming, Cao, Xiaoyi, Hebert, Lucie, Zhong, Sheng. "Systematic mapping of RNA-chromatin interactions in vivo". Current Biology, 27(4):602–609, 2017. (* co-first author). The dissertation author was one of the primary investigators and authors of this material.

CHAPTER 4 – Visualization and affinity pulldown of cell surface RNA together with their protein binding partners

Introduction

A vast majority of non-coding RNAs remain unidentified and poorly understood. Indeed, though the cells expend a vast amount of energy to produce these RNAs we do not know what role a vast majority of these RNAs play in cellular function. One possibility that is as yet unexplored is that ncRNAs might localize to the extracellular surface of cell membrane and aid in cell identity or cell-to-cell communications. Even though cell-to-cell communications via RNAs encapsulated in vesicles are well understood, the concept that intercellular communication is mediated via RNA through the cell surface hasn't been described. To demonstrate the existence of cell surface membrane RNAs, our lab has successfully isolated RNA from the cell surface membrane of EL4 mice cells by using membrane-coated nanoparticles and RNA capture via adaptor ligation methods by Zaleta et al. (unpublished data). Cell surface RNA sequencing analysis identified more than 3500 possible RNA candidates associated with the cell membrane, and quantum dot-coupled RNA probes visually confirmed the presence of at least 2 candidates at the cell surface. Although these results are very promising and provided evidence of the existence of cell surface RNA, the major challenge remains to ensure specificity of these cell surface molecules, which is avoiding contamination of intracellular RNAs and cell free RNAs, and to retrieve as much as specific cell surface RNA with good integrity. Here we aim to combine visualization and isolation in a single experiment by using a fluorescent dye coupled with biotin. Specific RNA localization will be validated by microscopic visualization and RNA isolation will be performed by streptavidin pull-

down, all this by using a unique molecule. Several cell types will be tested for the presence of cell surface RNAs: 786-0 human kidney cancer cell line, HEK293T cell line and EL4 mouse lymphoma cell line.

Methods

Visualization of cell surface RNA based on Click chemistry

"Click" chemistry, more commonly called tagging, is a class of biocompatible reactions. The classic click reaction is the Copper-catalyzed reaction of an azide with an alkyne to form a 5-membered heteroatom ring (named Cu(I)-Catalyzed Azide-Alkyne Cycloaddition (CuAAC)). Here we propose to use copper-based Click chemistry to specifically label and isolate cell surface RNAs in a single click reaction. We will specifically conjugate RNA with an impermeable molecule that enable both the visualization and isolation of cell surface RNAs. We will use the Cy5-Biotin-Azide (Click Chemistry Tools) which contains a negatively charged sulfonate.

5' Ethynil-Uridine (EU) is an analogue of Uridine that contains an alkyne group, thus incorporates specifically into nascent RNA in an unbiased manner. We propose to use EU as the RNA-specific probe and use Click chemistry to conjugate alkyne-EU-labeled RNAs to an azide coupled with both a fluorophore and a biotin molecules. The specificity of cell surface RNAs will be ensured by the use of molecules unable to penetrate the cell membrane. This specificity will be confirmed by direct microscopic observation, and the RNA will be pulled down using streptavidin-biotin affinity and then sequenced. One caveat to this approach is that the copper-catalyzed click reaction also releases oxidative byproducts that will damage cell membranes. Therefore, click

reaction condition will be optimized to biocompatible and ensure that the cells are still alive and intact.

Two major difficulties will have to be overcome: (1) Make sure of the impermeability of the probes; (2) Make to label and pull down sufficient amount of cell surface RNA in an unbiased manner for further sequencing.

EU is highly permeable to the cell membrane and can be incorporated into RNA of all types (intracellular and cell surface RNAs). We propose the following steps to optimize the desired click protocol:

Determine the EU treatment conditions (concentrations and treatment duration):

It is known that EU is highly permeable to the cell membrane and can be incorporated into RNA in several minutes in the literature. We do not expect to prevent EU from entering the cells but we want optimize the EU labelling to prevent to harm the integrity of the cell membrane.

Determine the most impermeable labeling conditions:

This is the most critical part of the protocol, because it will ensure the specificity of RNA pull-down. We first optimize the click reaction (fixation protocols and buffers) to maintain membrane integrity

It is known that standard fixation procedures can affect protein detection and preservation of cellular structure (Schnell et al, Nature Method, 2012). We will then chose the most impermeable label (fluorophore + biotin-containing label). Some dyes are known to be membrane-impermeable dyes (Nikic´ et al, Nature Protocol, 2015), such as sulfonated-Cy5, Atto532 and Alexa Fluor 647. We are testing sulfonated-Cy5 dye combined with biotin.

Determine the quantity of cells required to pull down a sufficient amount of labelled-RNAs:

The final aim of the procedure is to identify specific cell surface RNA. The first step (microscopic observation) will ensure specificity of the labelling, and the second strep (pull-down) will ensure collecting these molecules. This step needs to be performed while keeping in mind that we want to gather in an unbiased manner as much as cell membrane specific RNA molecules to generate an identification card of our cell line model. Thus we should optimize the number of cells needed to reach this goal without impairing the specificity. The last step will be to generate a RNA library and send to sequencing.

An overview of the work-flow is demonstrated in Figure 4.1

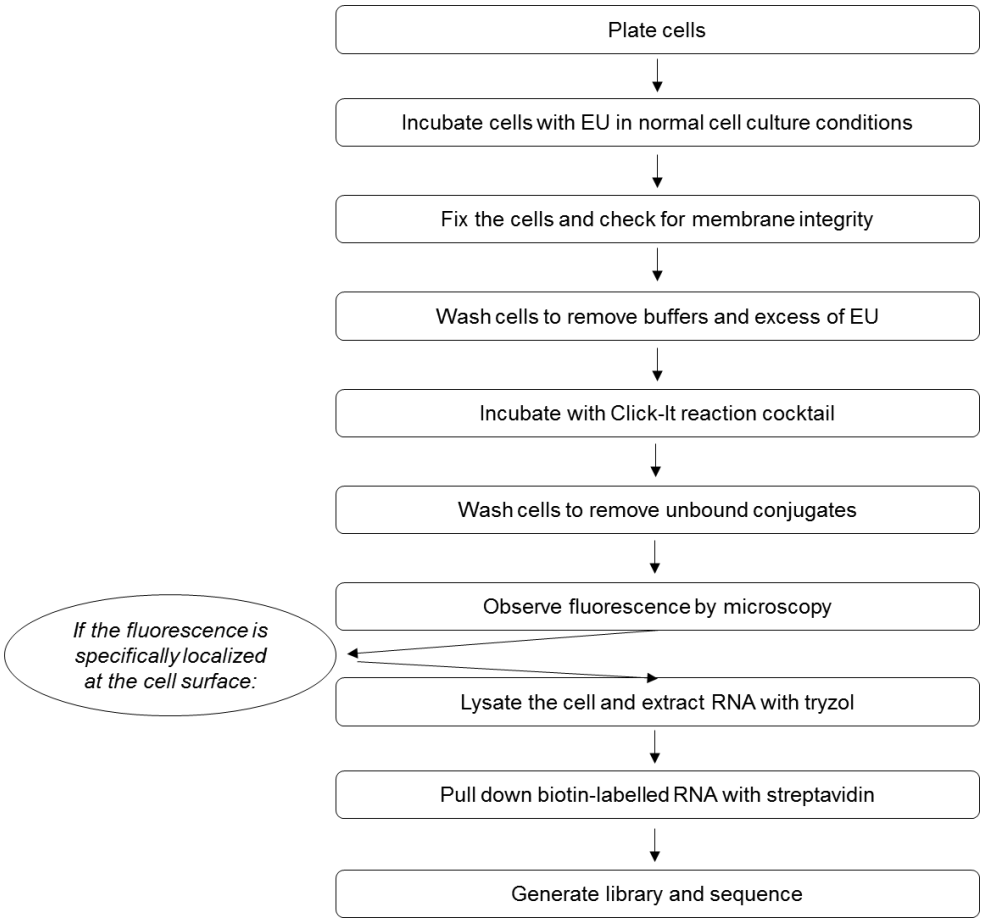


Figure 4.1*: **Procedure to visualize and affinity pulldown of surface RNA***

Affinity pulldown and high-throughput sequencing to identify cell surface RNAs

After click reaction, total RNA was isolated using Trizol reagent and 500ug of total RNA was fragmented using NEBNext Magnesium RNA fragmentation kit for 3.5min at 94C. Average RNA size should be 200nt. Fragmented RNA was then incubated with 300ul of T1 streptavidin beads slurry in 10ml of urea binding buffer (6M urea, 10mM tris pH7.5, 1mM EDTA, 0.1%NP-40). Samples were put on rotation overnight at room temperature. Non-biotinylated RNA carry-over was removed from beads using two rounds of stringent washes at 55C. 6 washes of 10min each with 10ml of 8M urea buffer (8M-UB) (8M urea, 10mM tris pH7.5, 1mM EDTA, 0.1%NP-40, 150mM NaCl, 0.5% SDS) followed by 6 washes of 10min each with 10ml of high salt buffer (HS-B) (4M NaCl, 10mM tris pH7.5, 1mM EDTA, 0.2%Tween20). Salt and urea were then removed by washing the beads twice with PNK buffer (20mM tris pH7.5, 10mM MgCl2, 0.2%Tween20). After ensuring for RNA binding specifity, cDNA library was generated on beads using an adapter version of NEBNext® Small RNA Library Prep Set for Illumina®. Beads were first incubated with T4 Polynucleotic kinase (T4 PNK) for 1h to phosphorylate the 5' end for subsequent ligation. Beads were washed twice with PNK buffer. Beads were then incubated with 1ul 3' SR Adaptor for Illumina, 6 ul H2O at 70C for 2min on a thermomixer for denaturation, then 3' ligation reaction buffer 2X and 3ul 3' Ligation Enzyme mix were added to the mixture and incubated at 16C overnight for ligation. 3 washes of 10min each with 1ml 8M-UB and 3 washes of 10min each with 1ml HS-B. 5' adaptors were then denatured separately in a thermocycler for 2 min at 70C and subsequently added to the beads previously resuspended in 60ul of the following reaction mix: 3' Ligation buffer (2X), RNAse inhibitor, 5' Ligation reaction buffer (10X), 5' Ligation Enzyme mix) and incubated at 16C overnight for ligation on a thermomixer. 3 washes of 10min each with 1ml 8M-UB and 3 washes of 10min each with 1ml HS-B. Beads were

then incubated with 2 ul of SR RT Primer for Illumina and First Strand Synthesis Buffer (5X) for 5min at 70C for denaturation and Protoscript II Reverse Transcriptase and Murine RNAse inhinitors were added to the mix and incubated on a thermomixer with the following program: 60 min 50C, 15 min 55C, 15 min 70C, ∞4C. The following PCR mix was subsequently added to the beads: LongAmp 2X Master Mix, 3.75 ul SR Primer for Illumina, 3.75 ul Illumina Index Primer, 27, 5 ul H2O. Number of cycles were evaluated with 20ul of sample + mix and sample library were subsequently amplified using the following PCR program: 94°C 30 sec, 12-15 cycles of (94°C 15 sec, 62°C 30 sec, 70°C 15 sec), 70°C 5 min.

Libraries were run on a gel and sized-selected to remove adaptors and dimers. Size-selected libraries were sequenced using Illumina Miniseq.

Affinity pulldown and protein mass-spectrometry to identify protein binding partners of cell surface RNA

Several batches of ~500M cells were generated as follows:

Cells were expanded in DMEM+10%HS. The day of the experiment, cells were counted, strained and concentrated into to flasks of 40ml. Half of the cells were treated with EU for 3h prior to the click reaction "(+) EU", the other half with PBS only "(-) EU". Cells were then washed thrice in warm culture medium and CLICK reaction was performed at RT on rotation in the dark in 20 ml Tyrode's Buffer supplemented with: 100uM CuSO4, 2.5 mM Na ascorbate, 150 uM TEMPOL, 500 uM BTTAA, 0.25mM Cy5-Biotin-Azide. Cells were then washed twice in 20 ml of warm cell culture medium and cells were immediately incubated with 3% formaldehyde in Tyrode's buffer for 30 min on rotation at RT. 0.1 M glycine was added to the cells for 5 min to quench any remaining free formaldehyde molecules and centrifuged at 4C at 2000g for 5 min. Cells were then washed thrice with cold PBS, flash frozen and store at -80C.

For each 3 mass spectrometry samples, three batches of cells were pooled together. Each pellet was weighted and resuspended in 10X worth is volume of Lysis Buffer (50 mM Tris-Cl pH 7.0, 10 mM EDTA, 1% SDS in H2O) supplemented with protease inhibitor cocktail (Roche # 04693132001). Lysates were homogenized with the Dounce 20 times and sonicated for 40min (Covaris E220 instrument, Chromatin Shearing protocol).

A small portion of lysate was used to isolate RNA to adjust for the quantities before proceeding to the pull-down. After adjustment, lysates were incubated O/N at RT with 1ml of streptavidin T1 beads in Urea-Binding Buffer (8M Urea, 10 mM Tris-Cl pH 7.5, 1 mM EDTA, 0.1 % NP-40, 150 mM NaCl, 0.5% SDS). Beads were washed stringently with Urea-Wash Buffer (8M Urea, 50 mM Tris-Cl pH 7.5, 5 mM EDTA, 0.1 % NP-40, 500 mM LiCl, 2% SDS) 6 times, 10 min each, on rotation at 55C. Another round of 6 washes with High-Salt-Buffer (4M NaCl, 10 mM Tris-Cl pH 7.5, 1 mM EDTA, 0.2 % Tween 100). Beads were then washed thrice with PNK buffer (20 mM Tris-Cl pH 7.5, 10mM MgCl2, 0.2 % Tween 100) to remove any remaining urea or salt.

Results

Detection and visualization cell surface RNA in EL4 and N2A

We tested several cell lines and were able to confirm the presence of cell surface RNAs through visualization in EL4 and N2A cell lines

Detection of cell Surface RNA in EL4

Two experimental replicates were performed. Cells were expanded in DMEM+10%HS. 2/3 of cells (500M) were treated with EU for 3h prior to the click reaction "(+) EU", 1/3 incubated with PBS "(-) EU". (+) EU cells were then either incubated with all CLICK reagents "(+) CuS04", or were incubated with all reagents but copper "(-) CuS04". (+) EU, (-) CuS04 negative control is the click reaction control and assessed for RNA pull-down specificity, and (-) EU (+) CuS04 control assesses the copper toxicity in live cells, mostly for the microscopic observation of RNA signal before the RNA pull-down. The images shown in Figures 4.2 and 4.3 illustrate a portion of live cells imaged prior to the pull-down to check on the RNA-labeling specificity.

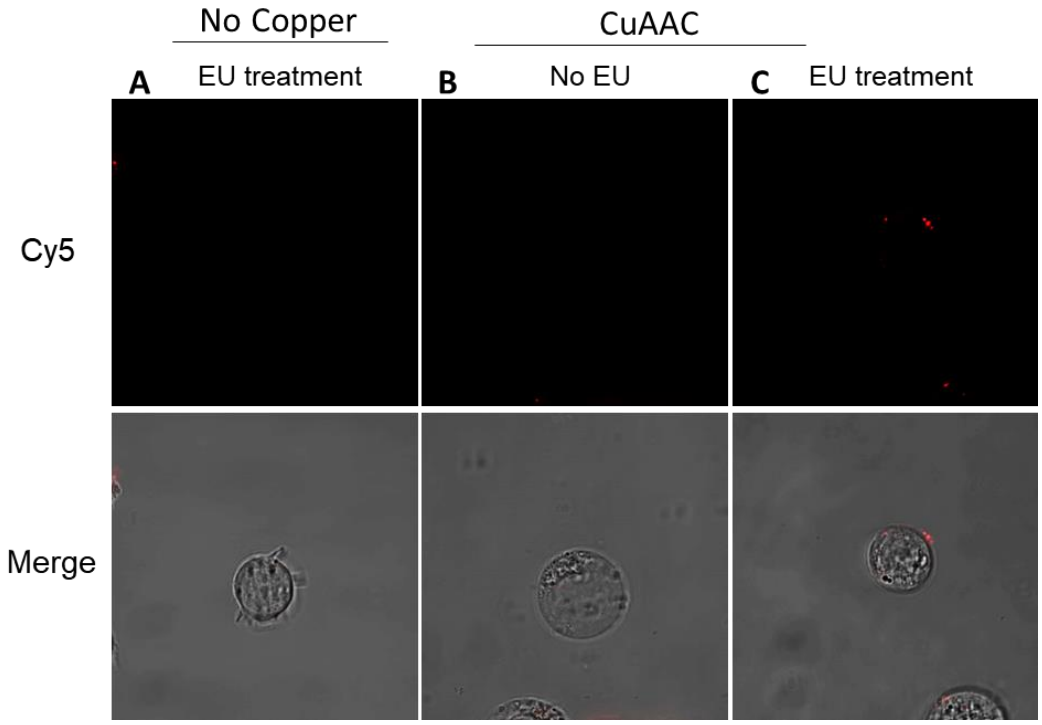


Figure 4.2*: Live EL4 cells after Click Reaction*. Experiment of 12-12-2017. (Labeling of cell surface RNA with the cy5 dye coupled with an azide through CuAAC). A. EL4 cells treated with Ethynyl-Uridine (EU) for 3h and then supplemented with all reagents for CuAAC reaction but the CuSO4 for 10 min. B. EL4 cells treated with PBS for 2h instead of EU and supplemented with all reagents for CuAAC reaction for 10 min. C. EL4 cells treated with Ethynyl-Uridine (EU) for 2h and then supplemented with all reagents but the CuSO4 for 10 min. The upper panel shows Cy5 channel in red (633 nm laser line). The lower panel shows differential interference contrast (DIC) merged with Cy5 channel. Scale bar, 10µm.

Table 4.1: Statistics of EL4 Click Experiment

Experimental condition	Cell surface staining	No staining	Full stain / aspecific	Total Cells Counted
EU + CuAAC	29%	47%	23.5%	17
EU + no Cu	0	50%	50%	30
No EU + CuAAC	8%	22%	32%	37

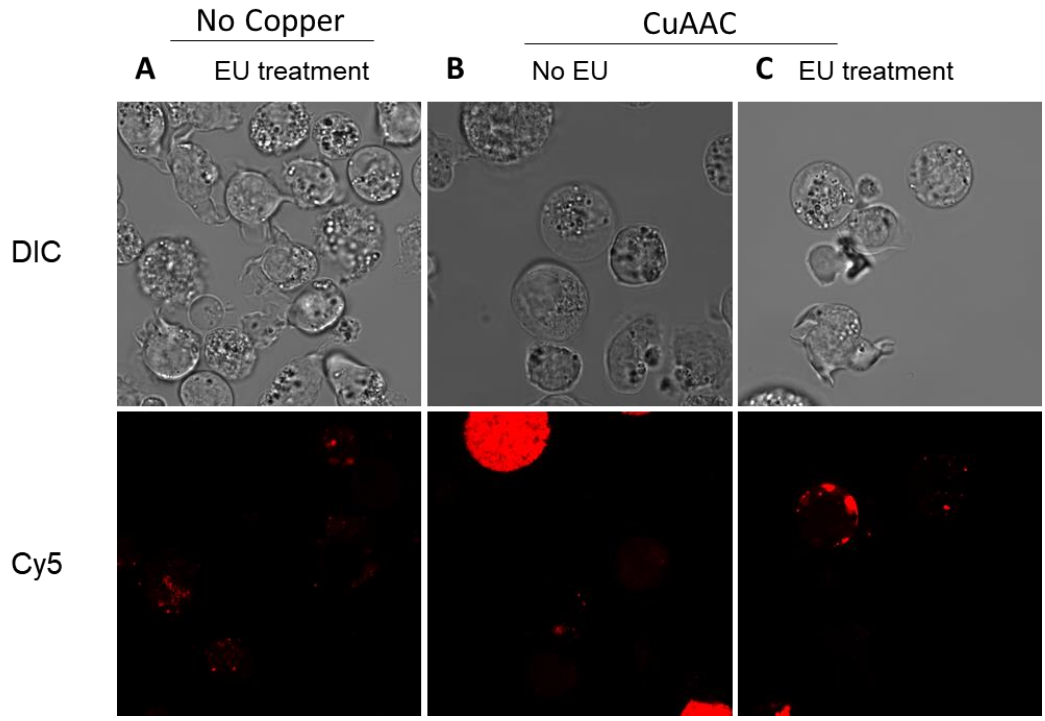


Figure 4.3: ***Live EL4 cells after Click Reaction***. (Labeling of cell surface RNA with the cy5 dye coupled with an azide through CuAAC). A. EL4 cells treated with Ethynyl-Uridine (EU) for 3h and then supplemented with all reagents for CuAAC reaction but the CuSO4 for 10 min. B. EL4 cells treated with PBS for 2h instead of EU and supplemented with all reagents for CuAAC reaction for 10 min. C. EL4 cells treated with Ethynyl-Uridine (EU) for 2h and then supplemented with all reagents including the CuSO4 for 10 min. The upper panel shows the differential interference contrast (DIC). The lower panel shows Cy5 channel in red (633 nm laser line). Scale bar, 10μm.

The two experiments showed cell-surface signal in the positive sample, though some aspecific signal was also observed in the NoEU + CuAAC. This was consistently observed in other experiments and is interpreted as a consequence of copper toxicity damaging the cell membrane and leading to aspecific binding of the dye – despite treatment with bathocuproin, a copper scavenger.

Detection of cell surface RNA in N2A cells

Cells were expanded 10 T125cm2 flasks in DMEM+10%FBS. All the cells were treated with EU for 3h prior to the click reaction. Half of cells were then either incubated with all CLICK reagents "(+) CuS04", or were incubated with all reagents but copper "(-) CuS04". (+) EU, (-) CuSO4 negative control is the click reaction control and assessed for RNA pull-down specificity. Of note, the click experiment was not validated by microscopic observation prior to pull-down.

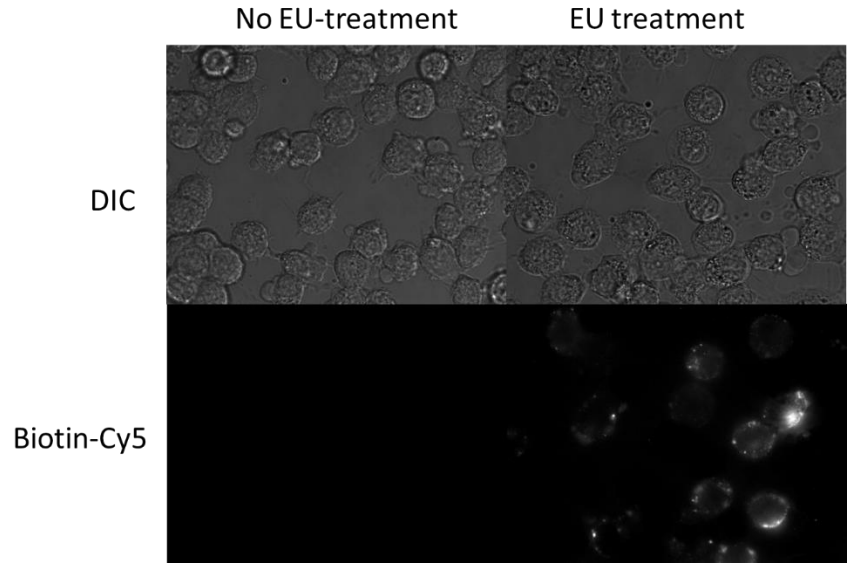


Figure 4.4: ***Live N2A cells after Click Reaction***

Table 4.2: Statistics of N2A Click Experiment

Experimental condition	**Cell surface staining**	**No staining**	**Full stain / aspecific**	**Total Cells Counted**
EU + CuAAC	10%	50%	40%	60
EU + no Cu	0	78%	22%	114
No EU + CuAAC	5%	61.5%	34%	80

Sequencing and analysis of cell surface RNA isolated from EL4 and N2A cells

The following libraries were generated from cell surface RNAs isolated from EL4 and N2A cells

Table 4.3: Description of RNA-Seq libraries generated for cell surface RNAs isolated from EL4 and N2A cells

Sample Name	Cell type	Species	Number of cells	Library lenght (nt)
EL4-wEU-wCu-11-06-17-20180112	EL4	Mouse	200M	210
EL4-wEU-wCu-12-12-17-20180112	EL4	Mouse	200M	318
N2A-wEU-wCu-11-17-17-20180112	N2A	Mouse	300M	280
N2A-wEU-woCu-11-17-17-20180112	N2A	Mouse	300M	290
EL4_Total_EU-Labelled-RNA	EL4	Mouse	50M	323
N2A_Total_EU-labelled-RNA	N2A	Mouse	50M	373

In order to see any RNA enriched at the surface of the cells, "cell-surface RNA libraries" were compared to "total EU-RNA libraries" using differential analysis. The analysis was performed separately for each cell line in order to keep the cell specificity.

EL4 cell line: Two cell-surface libraries were generated for EL4, whereas no negative libraries were amplified. The scRNA species were identified by calculating the mean of FPKM for each gene prior to the differential analysis:

FPKM_csRNA = MeanPos= (FPKM_EL4-11-06, FPKM_EL4-12-12)

N2A cell line: Since a negative control library was generated for N2A cell line, the difference FPKM between negative and sample library was calculated prior to the differential analysis:

FPKM_csRNA = FPKM_N2apos – FPKM_N2Aneg

For two cell lines, all FPKM values were logged using log10 (FPKM+1) formula.

The differential analysis was performed by comparing either the positive cell-surface libraries (previously analyzed with the negative libraries if any) or the total-RNA libraries using the formula below

Differential analysis formula:

Log Ratio (LR) = Log (Pos-Neg_FPKM+1)]-[Log (FPKM_tot1 + 1)]

Table 4.4 summarizes the number of genes identified in the libraries.

Table 4.4: Number of genes identified per libraries with FPKM>1

LIBRARY	NUMBER OF GENES WITH FPKM > 1
EL4_1	1963
EL4_2	836
N2A	1183
N2A_Negative control	2515
EL4_Total_EU-Labelled-RNA	4353
N2A_Total_EU-labelled-RNA	3840

Table 4.5 and 4.6 shows the top enriched surface RNAs in EL4 and N2A, respectively. The higher the ratio, the higher the enrichment of RNA species identified in the cell-surface libraries compared

to total RNA libraries. The tables 8, 9 and 10 show the top RNA candidates for which log ratio LR> 2. For the three cell lines. 28 top candidates were identified for EL4 cell line, 11 for N2A cell line. 4 RNAs were identified in both EL4 and N2A top differential genes and marked in red in Table 4.5 and 4.6. Most of top targets are snoRNAs for the two mouse cell lines with a few miRNA.

Table 4.5: EL4 csRNA species for which Log Ratio LR > 2

EL4		
gene name	**Biotype**	**LR(EL4)**
Snord2	snoRNA	5.22
Snord45b	snoRNA	4.54
Snord1b	snoRNA	4.36
Snord43	snoRNA	4.27
Snord12	snoRNA	4.23
Snord65	snoRNA	4.08
Snord68	snoRNA	3.94
Gm24148	snoRNA	3.88
Snord1c	snoRNA	3.77
Snord53	snoRNA	3.74
Snord66	snoRNA	3.57
Snord52	snoRNA	3.54

Table 4.6: N2A csRNA species for which Log Ratio LR > 2

N2A		
gene name	**Biotype**	**LR(N2A)**
Snord2	snoRNA	4.75
Rn45s	rRNA	4.56
Snord49a	snoRNA	4.12
Snord49b	snoRNA	4.08
Snord83b	snoRNA	3.88
Snord57	snoRNA	3.78
DQ267102	snoRNA	3.55
Mir881	miRNA	3.38
Snord1c	snoRNA	3.23
Snord12	snoRNA	2.87
Snora81	snoRNA	2.32

The RNA types in each cell lines are found in similar proportions in the two cell lines. These proportions reflect the major type of RNA enriched in cell surface click pull-down is from protein-coding genes (Figure 4.4)

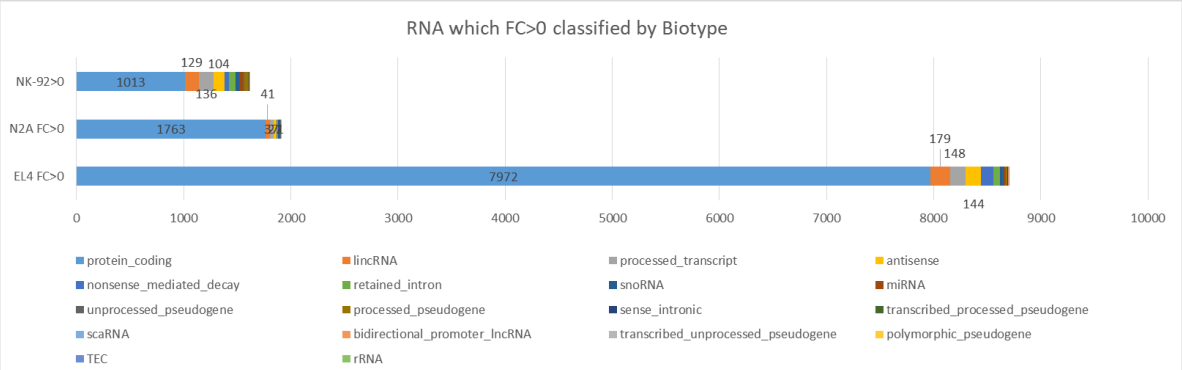


Figure 4.5*: Classification of candidate RNAs* which FC>0 by biotype for the three cell lines

The RNA types in each cell lines are found in similar proportions in the three cell lines. These proportions reflect the major type of RNA enriched in cell surface click pull-down is from protein-coding genes (Table 4.7)

Table 4.7: Classification of candidate RNAs which Log Ratio LR>0 by biotype for the three cell lines

RNA Biotype	EL4 LR>0	N2A LR>0
protein_coding	7972	1763
lincRNA	179	41
processed_transcript	148	37
antisense	144	27
nonsense_mediated_decay	111	3
retained_intron	63	8
snoRNA	39	23
miRNA	11	1
unprocessed_pseudogene	9	1
processed_pseudogene	8	1
sense_intronic	7	0
transcribed_processed_pseudogene	4	1
scaRNA	2	0
bidirectional_promoter_lncRNA	2	0
transcribed_unprocessed_pseudogene	1	0
polymorphic_pseudogene	1	0
TEC	1	0
rRNA	0	1
transcribed_unitary_pseudogene	0	0
Total	**8702**	**1907**

Analysis of protein pulldown assay and protein mass-spectrometry to identify protein-binding partners of cell surface RNAs

Two independent experiments of protein co-purified with cell surface RNAs were prepared for mass spectrometry. Table 4.8 summarizes their features

Table 4.8: Mass spectrometry protein sample description

Experiment #	Experiment Date	Elution 1 Type	Elution 2 Type	Elution 2 Date	Sample
1	April 3-6th	D-Biotin Elution (twice)	RNAse (twice)	April 6th	100 ul PBS
2	May 7-9th	RNAse (twice)	D-Biotin Elution (twice)	May 14th	100 ul PBS (pool of the 2 elutions)

The quantification file was used to assess for protein enrichment in the +EU+Click+RNAse over controls (no EU, no RNAse, both EU and no RNAse). The ratio between the samples was done and proteins which ratio are greater than 2 were used to find potential candidates. A pseudo count of 1 was used before doing the ratio to account for non-identified proteins in some of the conditions. Table 4.9 and 4.10 describe the samples submitted for two independent replicates.

Table 4.9: Description of the samples from experimental replicate #1

Sample Name	Full Name	Description of the sample
Inp -	Input -	Adjusted cell lysate of -EU+Click cells
inp +	Input +	Adjusted cell lysate of +EU+Click cells
Elu -	Eluent +EU	Eluent from +EU+Click beads
Elu +	Eluent -EU	Eluent from NoEU+Click beads
Exos	EL4 Exosomes	Proteins extracted from exosomes isolated from EL4 supernatant
Memb	EL4 Membrane proteins	Proteins extracted from EL4 membranes isolated by serial centrifugation

Table 4.10: Description of the samples from experimental replicate #2

Sample Name	Full Name	Description of the sample
Inp -	Input -	Adjusted cell lysate of -EU+Click cells 05-09
inp +	Input +	Adjusted cell lysate of +EU+Click cells 05-09
A	A	Eluent from +EU+Click, noRNAse treatment (pooled with D-biotin Eluent) 05-09
B	B	Eluent from +EU+Click, with RNAse treatment (pooled with D-biotin Eluent) 05-09
C	C	Eluent from NoEU+Click, noRNAse treatment (pooled with D-biotin Eluent)05-09
D	D	Eluent from NoEU+Click, with RNAse treatment (pooled with D-biotin Eluent) 05-09

Number of proteins detected in each replicate is described in Table 4.11 and 4.12

Table 4.11: Number of proteins detected in replicated #1

Sample	Nb protein detected
Inp -	1571
inp +	1594
Elu -	541
Elu +	403
Exos	525
Memb	1549

Table 4.12: Number of proteins detected in replicated #2

Sample	Nb protein detected
Inp -	874
inp +	778
A	176
B	100
C	138
D	209

In order to identify putative surface-RNA-binding protein candidates and discriminate candidates between non-specific proteins eluted from the beads, stringent criteria were applied.

We first compared the ratio of Elu+/Elu- and B/D (pool with RNAse) and looked for proteins commonly found in the two experiments for which FC>2.

7 proteins were commonly found as seen in Table 4.13

Table 4.13: Proteins commonly found in the duplicate experiments with B/D ratio

Accession		Name	FC (Elu+/Elu -)	FC (B/D)
Q9DBP5	KCY_MOUSE	UMP-CMP kinase	130001.00	1760001.00
Q8QZT1	THIL_MOUSE	Acetyl-CoA acetyltransferase mitochondrial	66801.00	3740001.00
P80316	TCPE_MOUSE	T-complex protein 1 subunit epsilon OS=Mus musculus GN=Cct5 PE=1 SV=1	50201.00	9.32
P35486	ODPA_MOUSE	Pyruvate dehydrogenase E1 component subunit alpha somatic form mitochondrial OS=Mus musculus GN=Pdha1 PE=1 SV=1	10001.00	43001.00
O08547	SC22B_MOUSE	Vesicle-trafficking protein SEC22b	3.22	317001.00
Q61233	PLSL_MOUSE	Plastin-2	2.45	1080001.00
P70333	HNRH2_MOUSE	Heterogeneous nuclear ribonucleoprotein H2	2.39	20201.00

To increase stringency, we looked at proteins commonly found in Elu+/Elu- and the ratio (B/A)/(D/C). This was to subtract proteins eluted without RNAse treatment, thus theoretically binding to the beads with no RNA interaction. As seen in Table 4.14, 5 proteins out of 7 are commonly found after adding the RNAse control stringency.

Table 4.14: Proteins commonly found in the duplicate experiments with (B/A)/(D/C) ratio

Accession	Symbol	Name	(Elu+) / (Elu -)	(B/A)/(D/C)
Q9DBP5	KCY_MOUSE	UMP-CMP kinase	130001.00	880001.00
Q8QZT1	THIL_MOUSE	Acetyl-CoA acetyltransferase mitochondrial	66801.00	1870001.00
O08547	SC22B_MOUSE	Vesicle-trafficking protein SEC22b	3.22	158501.00
Q61233	PLSL_MOUSE	Plastin-2	2.45	1079999.02
P70333	HNRH2_MOUSE	Heterogeneous nuclear ribonucleoprotein H2	2.39	10101.00

We next tested the significance of the overlap between the two biological duplicates. Hypergeometric test showed that the our overlap data is significant with p-value of 0.007776

We performed a hypergeometric test with the following parameters:

Sample #1 Elu+/Elu- proteins FC>2

Sample #2 B/D proteins FC>2

Universe=Union of the four inputs of the two experiments.

To better understand the possible roles of proteins binding to csRNAs, we went back to the two experiments independently, and submitted the most stringent list of 38 proteins from (**B/A) / (D/C)** ratios to DAVID gene ontology public software. The picture below summarizes the 5 top annotation clusters, listing the pathway enrichments and the associated p-values (Table 4.15).

Figure 4.15: DAVID pathway enrichment analysis for protein list of (B/A) / (D/C)

Annotation Cluster 1	Enrichment Score: 2.03	Count	P_Value	Benjamini
GOTERM_MF_DIRECT	nucleotide binding	11	1.00E-03	1.30E-01
UP_KEYWORDS	Nucleotide-binding	9	2.00E-03	4.90E-02
INTERPRO	P-loop containing nucleoside triphosphate hydrolase	7	2.10E-03	2.30E-01
UP_KEYWORDS	ATP-binding	7	9.40E-03	1.10E-01
GOTERM_MF_DIRECT	ATP binding	7	4.00E-02	6.80E-01
UP_KEYWORDS	Coiled coil	9	5.00E-02	2.80E-01
UP_SEQ_FEATURE	nucleotide phosphate-binding region:ATP	5	8.10E-02	8.90E-01
Annotation Cluster 2	**Enrichment Score: 1.77**	**Count**	**P_Value**	**Benjamini**
GOTERM_CC_DIRECT	cell-cell adherens junction	4	1.10E-02	2.90E-01
GOTERM_MF_DIRECT	cadherin binding involved in cell-cell adhesion	4	1.20E-02	4.20E-01
GOTERM_BP_DIRECT	cell-cell adhesion	3	3.70E-02	7.50E-01
Annotation Cluster 3	**Enrichment Score: 1.71**	**Count**	**P_Value**	**Benjamini**
UP_KEYWORDS	ER-Golgi transport	4	2.30E-04	7.90E-03
GOTERM_BP_DIRECT	vesicle-mediated transport	5	3.50E-04	7.50E-02
GOTERM_CC_DIRECT	membrane coat	3	8.90E-04	3.60E-02
GOTERM_BP_DIRECT	retrograde vesicle-mediated transport, Golgi to ER	3	9.80E-04	7.00E-02
GOTERM_BP_DIRECT	ER to Golgi vesicle-mediated transport	3	6.90E-03	3.20E-01
UP_KEYWORDS	Protein transport	5	7.60E-03	1.20E-01
GOTERM_BP_DIRECT	protein transport	5	1.40E-02	4.70E-01
UP_KEYWORDS	Golgi apparatus	5	1.90E-02	1.50E-01
GOTERM_BP_DIRECT	intracellular protein transport	3	5.30E-02	7.80E-01
GOTERM_CC_DIRECT	cytoplasmic vesicle	4	6.90E-02	5.90E-01
GOTERM_CC_DIRECT	Golgi apparatus	5	9.50E-02	6.20E-01
GOTERM_CC_DIRECT	Golgi membrane	3	1.20E-01	6.80E-01
UP_KEYWORDS	Cytoplasmic vesicle	3	1.40E-01	5.60E-01
UP_KEYWORDS	Transport	5	2.60E-01	7.40E-01
GOTERM_BP_DIRECT	transport	5	3.30E-01	1.00E+00
UP_KEYWORDS	Membrane	10	9.00E-01	1.00E+00
Annotation Cluster 4	**Enrichment Score: 1.59**	**Count**	**P_Value**	**Benjamini**
KEGG_PATHWAY	Carbon metabolism	4	2.70E-03	1.30E-01
UP_KEYWORDS	Transit peptide	5	4.90E-03	9.60E-02
UP_SEQ_FEATURE	transit peptide:Mitochondrion	5	9.90E-03	7.90E-01
UP_KEYWORDS	Mitochondrion	6	1.40E-02	1.30E-01
KEGG_PATHWAY	Biosynthesis of antibiotics	4	1.50E-02	3.20E-01
GOTERM_CC_DIRECT	mitochondrial matrix	3	3.10E-02	4.80E-01
GOTERM_CC_DIRECT	mitochondrion	7	3.70E-02	4.40E-01
GOTERM_BP_DIRECT	metabolic process	4	3.70E-02	7.00E-01
UP_KEYWORDS	Oxidoreductase	4	5.60E-02	2.90E-01
GOTERM_MF_DIRECT	oxidoreductase activity	4	8.40E-02	8.20E-01
GOTERM_BP_DIRECT	oxidation-reduction process	4	9.30E-02	8.80E-01
KEGG_PATHWAY	Metabolic pathways	6	1.90E-01	9.40E-01
Annotation Cluster 5	**Enrichment Score: 1.39**	**Count**	**P_Value**	**Benjamini**
GOTERM_MF_DIRECT	poly(A) RNA binding	8	2.30E-03	1.50E-01
GOTERM_MF_DIRECT	RNA binding	5	4.30E-02	6.40E-01
UP_KEYWORDS	RNA-binding	4	4.80E-02	2.80E-01
INTERPRO	RNA recognition motif domain	3	5.10E-02	9.60E-01
INTERPRO	Nucleotide-binding, alpha-beta plait	3	6.70E-02	8.80E-01
SMART	RRM	3	7.00E-02	7.80E-01
GOTERM_MF_DIRECT	nucleic acid binding	5	1.60E-01	9.30E-01

Interestingly, the first and fifth clusters relate to nucleotide and RNA binding, which makes sense considering our crosslinking/RNA-pulldown protocol. The second cluster appears to be more cell-type specific and related to protein functions, which also make sense considering the target protein localization. The third cluster relates to protein transport to the membrane. It is expected to have many proteins indirectly binding to csRNA, either through other RNA or with huge protein complexes. This pathway could be of interest if we would like to investigate how cell-surface RNA get to the surface of the cells.

Proteins corresponding to each interesting category are listed below. Highlighted proteins: ArF5, Sec22b and Rars appear in two independent categories, indicating a possible RNA interaction AND membrane localization/transport to the membrane. Of note, Sec22b is one of the candidate commonly found enriched in the tow independent experiments.

Table 4.16: Proteins involved in "Nucleotide binding" by DAVID - 38 protein list submission (B/A)/(D/C)

UNIPROT_ACCESSION	"Nucleic acid binding" GENE NAME
P84084	ADP-ribosylation factor 5(Arf5)
Q80Y44	DEAD (Asp-Glu-Ala-Asp) box polypeptide 10(Ddx10)
O89086	RNA binding motif protein 3(Rbm3)
Q9D0I9	arginyl-tRNA synthetase(Rars)
P80316	chaperonin containing Tcp1, subunit 5 (epsilon)(Cct5)
Q9DBP5	cytidine monophosphate (UMP-CMP) kinase 1(Cmpk1)
Q9Z0N1	eukaryotic translation initiation factor 2, subunit 3, structural gene X-linked(Eif2s3x)
P63017	heat shock protein 8(Hspa8)
P70333	heterogeneous nuclear ribonucleoprotein H2(Hnrnph2)
Q9QZH3	peptidylprolyl isomerase E (cyclophilin E)(Ppie)
P62196	protease (prosome, macropain) 26S subunit, ATPase 5(Psmc5)
Q8CG48	structural maintenance of chromosomes 2(Smc2)

Table 4.17: Proteins involved in "Transport/ Vesicle-mediated transport"by DAVID - 38 protein list submission (B/A)/(D/C)

UNIPROT_ACCESSION	GENE NAME
P84084	ADP-ribosylation factor 5(Arf5)
O08547	SEC22 homolog B, vesicle trafficking protein(Sec22b)
Q9JME5	adaptor-related protein complex 3, beta 2 subunit(Ap3b2)
Q8CIE6	coatomer protein complex subunit alpha(Copa)
O55029	coatomer protein complex, subunit beta 2 (beta prime)(Copb2)
P07724	albumin(Alb)

Table 4.18: Proteins involved in "cell-cell adherent junction" by DAVID - 38 protein list submission (B/A)/(D/C)

UNIPROT_ACCESSION	" cell-cell adherent junction" GENE NAME
Q9D0I9	arginyl-tRNA synthetase(Rars)
Q9Z0N1	eukaryotic translation initiation factor 2, subunit 3, structural gene X-linked(Eif2s3x)
P63017	heat shock protein 8(Hspa8)
Q9QXS1	plectin(Plec)

Table 4.19: Proteins involved in "Transmembrane" by DAVID - 38 protein list submission (B/A)/(D/C)

UNIPROT_ACCESSION	GENE NAME
V9GX34	CUB and Sushi multiple domains 2(Csmd2)
O08547	SEC22 homolog B, vesicle trafficking protein(Sec22b)
Q6P5G3	mbt domain containing 1(Mbtd1)
Q6ZWV3	ribosomal protein L10(Rpl10)
P49660	somatostatin receptor 4(Sstr4)

Acknowledgements

Chapter 4, in full, is currently being prepared for submission for publication of the material. Nguyen, Tri C.; Zaleta-Rivera, Kathia; Hebert, Lucie; Huang, Norman; Zhong, Sheng. The dissertation author was one of the primary investigators and authors of this material. The dissertation author was one of the primary investigators and authors of this material.

REFERENCES

Allen, P.N., and Noller, H.F. (1989). Mutations in ribosomal proteins S4 and S12 influence the higher order structure of 16 S ribosomal RNA. J. Mol. Biol. *208*, 457–468.

Aw, J.G.A., Shen, Y., Wilm, A., Sun, M., Lim, X.N., Boon, K.-L., Tapsin, S., Chan, Y.-S., Tan, C.-P., Sim, A.Y.L., et al. (2016). In Vivo Mapping of Eukaryotic RNA Interactomes Reveals Principles of Higher-Order Organization and Regulation. Mol. Cell *62*, 603–617.

Barabasi, A.-L., and Oltvai, Z.N. (2004). Network biology: understanding the cell's functional organization. Nat. Rev. Genet. *5*, 101–113.

Bartel, D.P. (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. Cell *116*, 281–297.

Bell, J.C., Jukam, D., Teran, N.A., Risca, V.I., Smith, O.K., Johnson, W.L., Skotheim, J.M., Greenleaf, W.J., and Straight, A.F. (2018). Chromatin-associated RNA sequencing (ChAR-seq) maps genome-wide RNA-to-DNA contacts. Elife *7*.

Bevilacqua, P.C., Ritchey, L.E., Su, Z., and Assmann, S.M. (2016). Genome-Wide Analysis of RNA Secondary Structure. Annu. Rev. Genet. *50*, 235–266.

Bohnsack, M.T., Tollervey, D., and Granneman, S. (2012). Identification of RNA helicase target sites by UV cross-linking and analysis of cDNA. Methods Enzymol. *511*, 275–288.

Cai, S., Han, H.J., and Kohwi-Shigematsu, T. (2003). Tissue-specific nuclear architecture and gene expession regulated by SATB1. Nat. Genet. *34*, 42–51.

Del Campo, C., Bartholomaeus, A., Fedyunin, I., and Ignatova, Z. (2015). Secondary Structure across the Bacterial Transcriptome Reveals Versatile Roles in mRNA Regulation and Function. PLOS Genet. *11*.

Castello, A., Horos, R., Strein, C., Fischer, B., Eichelbaum, K., Steinmetz, L.M., Krijgsveld, J., and Hentze, M.W. (2013). System-wide identification of RNA-binding proteins by interactome capture. Nat. Protoc. *8*, 491–500.

Chi, S.W., Zang, J.B., Mele, A., and Darnell, R.B. (2009). Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. Nature *460*, 479–486.

Chu, C., Qu, K., Zhong, F.L., Artandi, S.E., and Chang, H.Y. (2011). Genomic Maps of Long Noncoding RNA Occupancy Reveal Principles of RNA-Chromatin Interactions. Mol. Cell *44*, 667–678.

Chu, C., Spitale, R.C., and Chang, H.Y. (2015). Technologies to probe functions and mechanisms of long noncoding RNAs. Nat. Struct. Mol. Biol. *22*, 29–35.

Colak, D., Zaninovic, N., Cohen, M.S., Rosenwaks, Z., Yang, W.-Y., Gerhardt, J., Disney, M.D., and Jaffrey, S.R. (2014). Promoter-Bound Trinucleotide Repeat mRNA Drives Epigenetic Silencing in Fragile X Syndrome. Science (80-.). *343*, 1002–1005.

Ding, Y., Tang, Y., Kwok, C.K., Zhang, Y., Bevilacqua, P.C., and Assmann, S.M. (2014). In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features. Nature *505*, 696+.

Downey, C.D., Crisman, R.L., Randolph, T.W., and Pardi, A. (2007). Influence of Hydrostatic Pressure and Cosolutes on RNA Tertiary Structure. J. Am. Chem. Soc. *129*, 9290–9291.

Du, T., and Zamore, P.D. (2007). Beginning to understand microRNA function. Cell Res. *17*, 661.

Egli, M., Minasov, G., Tereshko, V., Pallan, P.S., Teplova, M., Inamati, G.B., Lesnik, E.A., Owens, S.R., Ross, B.S., Prakash, T.P., et al. (2005). Probing the influence of stereoelectronic effects on the biophysical properties of oligonucleotides: comprehensive analysis of the RNA affinity, nuclease resistance, and crystal structure of ten 2'-O-ribonucleic acid modifications. Biochemistry *44*, 9045–9057.

Ender, C., Krek, A., Friedlander, M.R., Beitzinger, M., Weinmann, L., Chen, W., Pfeffer, S., Rajewsky, N., and Meister, G. (2008). A human snoRNA with microRNA-like functions. Mol. Cell *32*, 519–528.

Engreitz, J.M., Pandya-Jones, A., McDonel, P., Shishkin, A., Sirokman, K., Surka, C., Kadri, S., Xing, J., Goren, A., Lander, E.S., et al. (2013). The Xist lncRNA Exploits Three-Dimensional Genome Architecture to Spread Across the X Chromosome. Science (80-.). *341*, 767+.

Fire, A., Xu, S., Montgomery, M.K., Kostas, S.A., Driver, S.E., and Mello, C.C. (1998). Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. Nature *391*, 806–811.

Franke, A., and Baker, B.S. (1999). The rox1 and rox2 RNAs are essential components of the compensasome, which mediates dosage compensation in Drosophila. Mol. Cell *4*, 117–122.

Granneman, S., Kudla, G., Petfalski, E., and Tollervey, D. (2009). Identification of protein binding sites on U3 snoRNA and pre-rRNA by UV cross-linking and high-throughput analysis of cDNAs. Proc. Natl. Acad. Sci. *106*, 9613–9618.

H., B., R., C., N., L.J., and D., V. (2004). Influence of the sugar configuration on the structure of RNA by conformational analysis of the ribose-phosphate unit. Biopolymers *14*, 695–713.

Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., Rothballer, A., Ascano Jr., M., Jungkamp, A.-C., Munschauer, M., et al. Transcriptome-wide Identification of RNA-Binding Protein and MicroRNA Target Sites by PAR-CLIP. Cell *141*, 129–141.

Helwak, A., Kudla, G., Dudnakova, T., and Tollervey, D. (2018). Mapping the Human miRNA Interactome by CLASH Reveals Frequent Noncanonical Binding. Cell *153*, 654–665.

Higgs, P.G. (2000). RNA secondary structure: physical and computational aspects. Q. Rev. Biophys. *33*, 199–253.

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

65, 274–287.

Ilik, I.A., Quinn, J.J., Georgiev, P., Tavares-Cadete, F., Maticzka, D., Toscano, S., Wan, Y., Spitale, R.C., Luscombe, N., Backofen, R., et al. (2013). Tandem Stem-Loops in roX RNAs Act Together to Mediate X Chromosome Dosage Compensation in Drosophila. Mol. Cell *51*, 156–173.

Incarnato, D., Neri, F., Anselmi, F., and Oliviero, S. (2014). Genome-wide profiling of mouse RNA secondary structures reveals key features of the mammalian transcriptome. GENOME Biol. *15*.

Kalhor, R., Tjong, H., Jayathilaka, N., Alber, F., and Chen, L. (2011). Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. Nat. Biotechnol. *30*, 90–98.

Kertesz, M., Wan, Y., Mazor, E., Rinn, J.L., Nutter, R.C., Chang, H.Y., and Segal, E. (2010). Genome-wide measurement of RNA secondary structure in yeast. Nature *467*, 103–107.

Kolkenbeck, K., and Zundel, G. (1975). The significance of the 2' OH group and the influence of cations on the secondary structure of the RNA backbone. Biophys. Struct. Mech. *1*, 203–219.

Kretz, M., Siprashvili, Z., Chu, C., Webster, D.E., Zehnder, A., Qu, K., Lee, C.S., Flockhart, R.J., Groff, A.F., Chow, J., et al. (2013). Control of somatic tissue differentiation by the long non-coding RNA TINCR. Nature *493*, 231-U245.

Kudla, G., Granneman, S., Hahn, D., Beggs, J.D., and Tollervey, D. (2011). Cross-linking, ligation, and sequencing of hybrids reveals RNA-RNA interactions in yeast. Proc. Natl. Acad. Sci. U. S. A. *108*, 10010–10015.

Kurdistani, S.K., and Grunstein, M. (2003). In vivo protein-protein and protein-DNA crosslinking for genomewide binding microarray. Methods *31*, 90–95.

Kwok, C.K. (2016). Dawn of the *in vivo* RNA structurome and interactome. Biochem. Soc. Trans. *44*, 1395–1410.

Kwon, S.C., Yi, H., Eichelbaum, K., Föhr, S., Fischer, B., You, K.T., Castello, A., Krijgsveld, J., Hentze, M.W., and Kim, V.N. (2013). The RNA-binding protein repertoire of embryonic stem cells. Nat Struct Mol Biol *20*, 1122–1130.

Li, F., Zheng, Q., Ryvkin, P., Dragomir, I., Desai, Y., Aiyer, S., Valladares, O., Yang, J., Bambina, S., Sabin, L.R., et al. (2012). Global Analysis of RNA Secondary Structure in Two Metazoans. Cell Rep. *1*, 69–82.

Li, X., Zhou, B., Chen, L., Gou, L.-T., Li, H., and Fu, X.-D. (2017). GRID-seq reveals the global RNA–chromatin interactome. Nat. Biotechnol. *35*, 940.

Licatalosi, D.D., Mele, A., Fak, J.J., Ule, J., Kayikci, M., Chi, S.W., Clark, T.A., Schweitzer, A.C., Blume, J.E., Wang, X., et al. (2008). HITS-CLIP yields genome-wide insights into brain alternative RNA processing. Nature *456*, 464–469.

Liu, W., Ma, Q., Wong, K., Li, W., Ohgi, K., Zhang, J., Aggarwal, A.K., and Rosenfeld, M.G. (2013). Brd4 and JMJD6-Associated Anti-Pause Enhancers in Regulation of Transcriptional Pause Release. Cell *155*, 1581–1595.

Lowman, H.B., and Draper, D.E. (1986). On the recognition of helical RNA by cobra venom V1 nuclease. J. Biol. Chem. *261*, 5396–5403.

Lu, Z., Zhang, Q.C., Lee, B., Flynn, R.A., Smith, M.A., Robinson, J.T., Davidovich, C., Gooding, A.R., Goodrich, K.J., Mattick, J.S., et al. (2016). RNA Duplex Map in Living Cells Reveals Higher-Order Transcriptome Structure. Cell *165*, 1267–1279.

McGinnis, J.L., Dunkle, J.A., Cate, J.H.D., and Weeks, K.M. (2012). The Mechanisms of RNA SHAPE Chemistry. J. Am. Chem. Soc. *134*, 6617–6624.

Miao, Y., Ajami, N.E., Huang, T.-S., Lin, F.-M., Lou, C.-H., Wang, Y.-T., Li, S., Kang, J., Munkacsi, H., Maurya, M.R., et al. (2018). Enhancer-associated long non-coding RNA LEENE regulates endothelial nitric oxide synthase and endothelial function. Nat. Commun. *9*, 292.

Mili, S., and Steitz, J.A. (2004). Evidence for reassociation of RNA-binding proteins after cell lysis: Implications for the interpretation of immunoprecipitation analyses. RNA *10*, 1692–1694.

Nguyen, T.C., Cao, X., Yu, P., Xiao, S., Lu, J., Biase, F.H., Sridhar, B., Huang, N., Zhang, K., and Zhong, S. (2016). Mapping RNA–RNA interactome and RNA structure in vivo by MARIO. Nat. Commun. *7*, 12023.

Nowak, D.E., Tian, B., and Brasier, A.R. (2005). Two-step cross-linking method for identification of NF-κB gene network by chromatin immunoprecipitation. Biotechniques *39*, 715–724.

Oltvai, Z.N., and Barabasi, A.-L. (2002). Systems biology. Life's complexity pyramid. Science *298*, 763–764.

Pashev, I.G., Dimitrov, S.I., and Angelov, D. (1991). Crosslinking proteins to nucleic acids by ultraviolet laser irradiation. Trends Biochem. Sci. *16*, 323–326.

Quinn, J.J., Ilik, I.A., Qu, K., Georgiev, P., Chu, C., Alchtar, A., and Chang, H.Y. (2014). Revealing long noncoding RNA architecture and functions using domain-specific chromatin isolation by RNA purification. Nat. Biotechnol. *32*, 933–940.

Ray, D., Kazan, H., Cook, K.B., Weirauch, M.T., Najafabadi, H.S., Li, X., Gueroussov, S., Albu, M., Zheng, H., Yang, A., et al. (2013). A compendium of RNA-binding motifs for decoding gene regulation. Nature *499*, 172–177.

Ribitsch, G., De Clercq, R., Folkhard, W., Zipper, P., Schurz, J., and Clauwaert, J. (1985). Small-angle X-ray and light scattering studies on the influence of Mg2+ ions on the structure of the RNA from bacteriophage MS2. Zeitschrift Fur Naturforschung. Sect. C, Biosci. *40*, 234–241.

Rinn, J.L., and Chang, H.Y. (2012). Genome Regulation by Long Noncoding RNAs. In ANNUAL REVIEW OF BIOCHEMISTRY, VOL 81, Kornberg, RD, ed. pp. 145–166.

Rouskin, S., Zubradt, M., Washietl, S., Kellis, M., and Weissman, J.S. (2014). Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. Nature *505*, 701+.

Shabalina, S.A., Ogurtsov, A.Y., and Spiridonov, N.A. (2006). A periodic pattern of mRNA secondary structure created by the genetic code. Nucleic Acids Res. *34*, 2428–2437.

Sharma, E., Sterne-Weiler, T., O'Hanlon, D., and Blencowe, B.J. (2016). Global Mapping of Human RNA-RNA Interactions. Mol. Cell *62*, 618–626.

Siegfried, N.A., Busan, S., Rice, G.M., Nelson, J.A.E., and Weeks, K.M. (2014). RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP). Nat. Methods *11*, 959–965.

Silverman, I.M., and Gregory, B.D. (2015). Transcriptome-wide ribonuclease-mediated protein footprinting to identify RNA-protein interaction sites. Methods *72*, 76–85.

Silverman, I.M., Li, F., Alexander, A., Goff, L., Trapnell, C., Rinn, J.L., and Gregory, B.D. (2014). RNase-mediated protein footprint sequencing reveals protein-binding sites throughout the human transcriptome. Genome Biol. *15*.

Simon, M.D., Wang, C.I., Kharchenko, P. V, West, J.A., Chapman, B.A., Alekseyenko, A.A., Borowsky, M.L., Kuroda, M.I., and Kingston, R.E. (2011). The genomic binding sites of a noncoding RNA. Proc. Natl. Acad. Sci. U. S. A. *108*, 20497–20502.

Simon, M.D., Pinter, S.F., Fang, R., Sarma, K., Rutenberg-Schoenberg, M., Bowman, S.K., Kesner, B.A., Maier, V.K., Kingston, R.E., and Lee, J.T. (2013). High-resolution Xist binding maps reveal two-step spreading during X-chromosome inactivation. Nature *504*, 465+.

Smola, M.J., Calabrese, J.M., and Weeks, K.M. (2015). Detection of RNA-Protein Interactions in Living Cells with SHAPE. Biochemistry *54*, 6867–6875.

Spitale, R.C., Flynn, R.A., Zhang, Q.C., Crisalli, P., Lee, B., Jung, J.-W., Kuchelmeister, H.Y., Batista, P.J., Torre, E.A., Kool, E.T., et al. (2015). Structural imprints in vivo decode RNA regulatory mechanisms. Nature *519*, 486+.

Sridhar, B., Rivas-Astroza, M., Nguyen, T.C., Chen, W., Yan, Z., Cao, X., Hebert, L., and Zhong, S. (2017). Systematic Mapping of RNA-Chromatin Interactions In Vivo. Curr. Biol. *27*, 602–609.

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

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

Talkish, J., May, G., Lin, Y., Woolford Jr., J.L., and McManus, C.J. (2014). Mod-seq: high-throughput sequencing for chemical probing of RNA structure. RNA *20*, 713–720.

Tsai, M.-C., Manor, O., Wan, Y., Mosammaparast, N., Wang, J.K., Lan, F., Shi, Y., Segal, E., and Chang, H.Y. (2010). Long Noncoding RNA as Modular Scaffold of Histone Modification Complexes. Science (80-.). *329*, 689–693.

Underwood, J.G., Uzilov, A. V, Katzman, S., Onodera, C.S., Mainzer, J.E., Mathews, D.H., Lowe, T.M., Salama, S.R., and Haussler, D. (2010). FragSeq: transcriptome-wide RNA structure probing using high-throughput sequencing. Nat. Methods *7*, 995-U81.

Wan, Y., Qu, K., Ouyang, Z., Kertesz, M., Li, J., Tibshirani, R., Makino, D.L., Nutter, R.C., Segal, E., and Chang, H.Y. (2012). Genome-wide Measurement of RNA Folding Energies. Mol. Cell *48*, 169–181.

Wan, Y., Qu, K., Zhang, Q.C., Flynn, R.A., Manor, O., Ouyang, Z., Zhang, J., Spitale, R.C., Snyder, M.P., Segal, E., et al. (2014). Landscape and variation of RNA secondary structure across the human transcriptome. Nature *505*, 706+.

Watts, J.M., Dang, K.K., Gorelick, R.J., Leonard, C.W., Bess Jr., J.W., Swanstrom, R., Burch, C.L., and Weeks, K.M. (2009). Architecture and secondary structure of an entire HIV-1 RNA genome. Nature *460*, 711-U87.

Zhang, C., and Darnell, R.B. (2011). Mapping in vivo protein-RNA interactions at single-nucleotide resolution from HITS-CLIP data. Nat Biotechnol *29*, 607–614.

Zhang, J., Poh, H.M., Peh, S.Q., Sia, Y.Y., Li, G., Mulawadi, F.H., Goh, Y., Fullwood, M.J., Sung, W.K., Ruan, X., et al. (2012). ChIA-PET analysis of transcriptional chromatin interactions. Methods *58*, 289–299.