The Discovery and Characterization of the lncRNA Firre

The discovery and characterization of the lncRNA Firre

A dissertation presented

by

Fatma Ezgi Hacisuleyman

to

The Department of Molecular and Cellular Biology

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Biochemistry

Harvard University

Cambridge, Massachusetts

April 16, 2015

Dissertation Advisor: John L. Rinn Fatma Ezgi Hacisuleyman

The discovery and characterization of the lncRNA Firre

Abstract

RNAs, including long noncoding RNAs (lncRNA), are known to be abundant and important structural components of the nuclear infrastructure. Yet, the identities, functional roles, and localization dynamics of lncRNAs that influence nuclear architecture remain poorly understood. Another unexplored territory is the molecular nature of the nuclear lncRNAs, which hampers a mechanistic understanding of how these RNAs establish proper epigenetic states and drive and modulate nuclear compartmentalization.

Here, we identify a lncRNA that we discovered and termed Functional Intergenic RNA Repeat Element (Firre). Firre is a strictly nuclear lncRNA that interacts with the nuclear matrix protein hnRNPU, through a 156 bp repeat motif in its mature transcript sequence. This conserved and unique repeat motif, Repeating RNA Domain (RRD), is not only necessary to localize Firre around its site of transcription in the nucleus but also sufficient to act as a nuclear localization signal for any RNA in a species-specific manner.

Furthermore, Firre spreads across a ~5 Megabase (Mb) domain around its transcription site on the X chromosome and localizes across at least five distinct trans-chromosomal loci in the genome. The trans-chromosomal targets reside in spatial proximity to the *Firre* locus, the genetic deletion of which results in the loss of co-localization of these trans-chromosomal interacting loci. Interestingly, the knockdown of hnRNPU also impedes these trans sites to be brought into the vicinity of the *Firre* locus. Thus, our data suggest a new form of lncRNA-mediated regulation in the nucleus, in which lncRNAs, such as Firre,

via their unique repetitive domains, can interface with and modulate nuclear architecture across chromosomes.

Abbreviations

RNA Ribonucleic Acid
DNA Deoxyribonucleic Acid
ncRNA Noncoding RNA
lnc Long Noncoding RNA
mRNA Messenger RNA
rRNA Ribosomal RNA
mirNA Micro RNA
siRNA Small interfering RNA
snRNA Small nuclear RNA
snoRNA Small nucleolar RNA
tRNA Transfer RNA
piRNA Piwi interacting RNA
H3 Histone 3
K Lysine
me Methylation
Xist X inactive specific transcript
Firre Functional Intergenic RNA Repeat Element
RRD Repeating RNA Domain
TADs Topologically Associated Domains
ESC Embryonic Stem Cell
TE Transposable Element
TR Tandem Repeat
LR Local Repeat
LOF Loss of function
GOF Gain of function
UCSC University of California at Santa Cruz
ChIP Chromatin Immunoprecipitation
GO Gene Ontology
ORO Oil red O
JSD Jensen-Shannon distance
GSEA Gene Set Enrichment Analysis
CSF Codon Substitution Frequency
ORF Open Reading Frame
EST Expressed Sequence Tags
RNP Ribonucleoprotein
Mb Mega base
kb Kilo base
bp Base pair
FISH Fluorescence *in situ* Hybridization
CTCF CCCTC-binding factor
mLF Mouse Lung Fibroblast
hLF Human Lung Fibroblast
hFF Human Foreskin Fibroblast
RAP RNA Affinity Purification

Acknowledgments

First and foremost I would like to take this opportunity to express my deepest gratitude to my Ph.D advisor, John L. Rinn, for his continuous support, enthusiasm, and faith. He has given me the courage and ardor to pursue what I believe in and fueled my passion for science. He has not only been a great mentor but also provided me with a family that I can rely on during the challenges I have faced during this journey.

The completion of this dissertation would also have not been possible without the generous help, advice, and support of my thesis committee members. I would like to deeply thank Professor Catherine Dulac, with whom I started graduate school as my neutral advisor, for giving me precious advice throughout my Ph.D at every fork on the road and always taking the time to talk to me through difficult decisions. Professor Victoria D'Souza has been an excellent mentor and introduced me to area of expertise. Professor Alexander Meissner, who has been one of the prominent figures in guiding me for my graduate school decision, has also been a very valuable source of advice for my thesis project. I am very fortunate to have been guided by such great minds.

Not only my dissertation but any of the successes I have achieved would not have been realized without my biggest source of strength, which is my family. My parents and my sister have made me believe that fear is what limits your vision. With their unconditional love and support from miles away, they have instilled in me a core set of values that guided me throughout every step. They taught me to be strong and courageous, work hard, and learn to belong wherever I am by accepting myself. Their values and presence gave me the courage to be authentic, honest, and imperfect and to understand my flaws and work on them.

There are also those who have become like a family to me throughout my Ph.D. Out of the close circle of friends who have never given up their incredible friendship, I would like to emphasize a few people. David Hendrickson, Cole Trapnell, and Mike Morse, with whom I had the opportunity to work with in the Rinn Lab, have been the brothers I have never had. They have been a great resource of love, support, scientific wisdom, and critical thinking. I was also extremely lucky to meet and become close friends with three inspiring women: Abigail Groff, Selen Yanmaz, and Iris Odstricil. Abigail Groff very soon after joining the Rinn Lab has become not only an inspiring person to work with but also a dear friend and a role model. I find in her the strength any woman needs to be independent, hardworking, successful, assertive, and at the same time compassionate. Selen Yanmaz, who I met at a political protest in Cambridge, made me see a world quite different than what I had been used to in my science bubble. Her endless love for every living thing, perspectives on life, and continuous support transformed me in ways I have never imagined and made me a significantly better person. Finally, I would like to end by expressing my deepest thanks and love for Iris Odstricil, who is my classmate and who has become a sister to me. She was always there when I needed help the most. She brings out the best in me, allows me to grow, and gives me hope and total freedom to be myself. Her affection, compassion, and closeness have given me the power to endure scientific and personal hardships. I am deeply grateful for having someone like her to give value to my survival.

Table of Contents

List of Figures

Chapter 1: Introduction

1.1 The noncoding genome and noncoding RNAs

The advent of next generation sequencing approaches allowed the identification of numerous RNA molecules that arise from noncoding regions of the genome, previously known as the "junk" DNA[1-6]. RNA has always been in the center of information flow from genomic content to functional output; however, the roles that have been ascribed to RNA have surpassed just being a "messenger." In addition to messenger RNAs (mRNAs) that code for proteins, ribosomal RNAs (rRNAs), transfer RNAs (tRNAs), and small RNA (sRNAs) that perform functions as RNAs constitute the majority of RNA population in the cell. Amongst these noncoding RNA (ncRNA) forms, rRNA is the most abundant species. It is common to all life forms and have been used to map evolutionary divergence across organisms[7,8]. Furthermore, rRNA constitutes the structural and the catalytic core of the ribosome, the structure of which has been characterized to great detail along with the multitude of proteins that rRNA contacts[9-13]. Along with rRNAs, tRNAs are are also necessary components of protein translation. With their specialized structure, tRNAs serve as an intermediary between the DNA code and the amino acids and provide a dynamic interface between the ribosome and catalytic proteins[14-16]. In addition to the core regulatory ncRNAs, small ncRNAs were discovered by accident but turned out to have various functional roles in: gene regulation, genome stability, and chromatin organization (interfering)[17,18], transposon defense (Piwi-interacting)[19], nucleotide modification (nucleolar)[20], and splicing (nuclear)[21]. All these examples illustrate the breadth of biological functions that RNA can execute.

In addition to the incredible diversity of RNA species, what gave scientists a strong drive to study RNA was the discovery of pervasive transcription of the genome, especially of

the noncoding segments. At a given time, 68% of the RefSeq genes are active, resulting in many transcripts, ~1% of which is generated form the coding regions[2, 22]. Tiling arrays and RNA sequencing efforts revealed that this is a pertinent phenomenon across a diverse range of eukaryotes[23-25]. Using cap analysis of gene expression (CAGE) and 3' sequencing, ~180,000 cDNAs were identified along with ~20,000 protein-coding genes in mouse[26,27], which is a number shared with humans, flies, and worms[28-30]. The similarity in the numbers of genes but the stark contrast in complexity led to a rising interest in the RNA processing pathways and the noncoding genome. Interestingly, the majority of the transcripts resulted from RNAs that are alternatively spliced and are generated from alternative promoters or from noncoding regions[26,31,32].

1.2 Pervasively transcribed long noncoding RNAs

The idea of pervasive transcription supported with genome and RNA sequencing technologies initiated an unprecedented survey of the genome, which revealed longer transcripts that did not fit into the same category as previously described RNAs[33,34]. Deep sequencing of cDNAs, termed RNA sequencing, coupled with intensive computational efforts allowed for the reconstruction and identification of these new transcripts at single nucleotide resolution[35-40]. These RNAs are called long noncoding RNAs (lncRNAs) based on being larger than 200 base pairs (bp) due to experimental constraints. There are five broad categories of lncRNAs according to their derivation from the genome: sense, antisense, bidirectional, intronic, and intergenic[41]. Similar to mRNAs, lncRNAs have a promoter and contain the characteristic promoter and gene body chromatin marks associated with active transcription: histone 3 lysine 4 trimethylation (H3K4me3) and histone 3 lysine 36 trimethylation (H3K36me3), respectively (42,43).

There has been a cumulative and extensive effort to characterize these transcripts that resemble mRNAs in multiple ways but differ in others. Using protein homology queries (BLASTX) and codon substitution frequency analyses, lncRNAs have been assessed to lack coding potential although ribosome profiling indicates that small peptides can be encoded within lncRNAs[44-46]. Although lncRNAs do not code for proteins like mRNAs, they are transcribed by RNA polymerase II (RNAP II), can be poly-adenylated and spliced, and can localize in various cellular compartments in similar ways to mRNAs[42,47-50]. In fact, new studies show that there might be purifying selection to conserve the efficient splicing sites of multiexonic lncRNAs[51]. On the other hand, in contrast to mRNAs, primary sequence conservation and expression levels of lncRNAs are modest-to-low, rendering lncRNA studies really challenging[52,53]. However, although there might be low primary sequence conservation, functional, structural, and modular conservation were found to be important for lncRNA function[51,53-56], and rapid turnover of lncRNAs can suggest an evolutionary trend for species-specific gene expression[57].

Shortly after the characterization of lncRNAs, numerous studies have highlighted that these new transcripts might be dynamic, versatile, and critical regulators of the genome. Three important features, which particularly challenged the central dogma of RNA being a sole messenger in biological functions, emerged: lncRNAs 1) show tissue-specific expression[40,58,59], 2) are developmentally regulated[40,60-63], and 3) are associated with disease loci and can be used as biomarkers[40,64-68]. All these studies underscored a clear understanding of lncRNAs exhibiting architecture and coordination, leading to an elegantly choreographed regulation of DNA and protein by RNA and in turn biological functional output.

1.3 The emerging roles and mechanisms of long noncoding RNAs

One of the first lncRNAs to be discovered, Xist[69,70], preceded the bloom of lncRNA research and presented a lot of insights about the potential roles and mechanisms of lncRNAs. The lncRNA Xist, via its repeat domains, binds to multiple proteins as a "scaffold" to establish proper epigenetic silencing of genes on the X chromosome[71], thereby causing a structural condensation of the whole chromosome in females. Several lncRNAs in this locus have been shown to provide additional layers of regulation and recruit epigenetic regulatory complexes[72,73], some of which are brought to the future inactive X chromosome[74,75]. Moreover, Xist RNA localization is governed by tertiary chromosomal confirmations, supporting a model, where genomic proximity governs the association of lncRNAs and chromatin[76,77]. The findings in the Xist field pointed to general roles for lncRNAs in binding and guiding multiple proteins, regulating gene expression, and facilitating higher-order genomic interactions.

With the mechanistic study of Xist and additional lncRNAs, a few themes have emerged in the universe of lncRNA mechanisms: decoy, scaffold, guide, and signal[41]. In the decoy example, the lncRNA titrates its protein target away from the protein's target loci, resulting in gene repression; examples include lncRNAs Gas5 and PANDA[78,79]. The scaffold and guide mechanisms are similar in the way that the lncRNA binds to one or more protein targets; the lncRNA can act as a scaffold and concentrate proteins in certain sub-cellular domains or it can further actively recruit proteins to certain loci on the genome as a guide. For example, the HOTAIR lncRNA was found to bind PRC2 and LSD1-CoREST complexes at the same time via specific domains within the RNA sequence[80,81], and lncRNA-p21 binds and recruits hnRNP K to certain promoters upon DNA damage[82]. Lastly, the lncRNAs have been associated with being a signal to activate or repress gene expression or change the chromatin

conformation thus affect three-dimensional interaction networks. This has been exemplified in the context of enhancer RNAs (eRNAs), which upon expression induce activation of the nearby protein coding gene[83-85], or in the context of HOTTIP, which induces chromosome looping and up-regulates transcription of its targets[86].

The effort to investigate the mechanisms of lncRNAs brought about the development of a wealth of new biochemical and genomic tools, which further illuminated the roles lncRNAs play in cells. Especially, the predominance and roles of lncRNAs in the nucleus and their relatively lower expression levels required novel approaches that differed significantly than those for mRNAs[87-89]. In addition to the computational tools necessary to analyze these novel transcripts, new experimental methods to understand how RNA can interface with DNA were developed: ChIRP (chromatin isolation by RNA purification)[90], CHART (capture hybridization analysis of RNA targets)[91], and RAP (RNA affinity purification)[77,92]. The combination of RNA biochemistry and next generation sequencing approaches revealed that lncRNAs can 1) interfere with transcription or activate and transport transcription factors to initiate transcription, 2) recruit chromatin modifiers to specific sites, 3) regulate splicing, 4) serve as structural/organization components to form protein complexes, 5) alter protein localization, 6) function in telomere biology[93-98].

1.4 Organizational principles and roles of RNAs in the nucleus

The observed roles and mechanisms of lncRNAs on DNA and their localization properties raised intriguing questions as to whether they might be important for organization in the nucleus, which is a phenomenon still not well understood. The nucleus is an incredibly complex environment; while packing long stretches of DNA, it also has to accommodate transcription, DNA repair, replication, and all the other regulatory events and ensure that they

are carried out in a timely and organized manner. In agreement with the incredible coordination of events, the nucleus is inherently very structured. It has been shown that DNA is packaged into a higher-order chromatin structure[99], and chromosomes occupy distinct territories[100,101]. Depending on the expression or repression of genes, the parts of the chromosomes that house these genes can loop out to move towards the interior of the nucleus or towards the nuclear membrane, respectively[102-104]. Furthermore, the nucleus is highly compartmentalized; however, these compartments lack a membranous outer layer unlike their cytoplasmic counterparts, which renders nuclear organization very dynamic. In fact, the genome is compartmentalized on a larger scale into topologically associated domains (TADs), which are conserved across cell types and even across species[105,106]. However, how these domains are determined and how the organizational dynamics of the nucleus change to bring about a variety of gene regulatory programs in various cell types remain unknown.

Understanding the dynamics of organization in the nucleus is crucial because it is now well known that as cells differentiate and become more specialized, the structure of the chromatin and its associated marks, locations of genes, interactions within and across chromosomes, and sub-domains within TADs change significantly[107-114]. A fascinating organization of the genome is observed in the zygote during pre-implantation development in mice: repetitive centromeric regions move to the interior of the nucleus and are subsequently remodeled, clustering around precursors of nucleoli[115]. This movement and restructuring of the chromosomes allow for cross-talk across chromosomes and a concurrent gene activation and are required for normal embryonic development. Furthermore, as the zygote divides and differentiates, these repetitive motifs move around and adopt new arrangements[116]. Similarly, heterochromatic rearrangement and changes in gene positions have been found to be necessary

for embryonic stem cell (ESC) differentiation. One of the core pluripotency genes, Nanog, localizes towards the interior of the nucleus in ESCs and forms new long-range interactions with genes on other chromosomes during differentiation, which is hindered when cross-chromosomal contacts are inhibited[117,118]. Concordantly, another core factor, Oct4, loops out of its chromosome to regulate its targets in ESCs[117]. Since lncRNAs are generally expressed in a cell-type and/or context specific manner, they might help elucidate some of the cell-type specific organizational principles in the nucleus.

In the dynamic structure of the nucleus, it was in fact found that RNA molecules play important roles (Figure 1). Firstly, a variety of RNA species has been found to be key constituents of the nuclear matrix[119-122], and necessary for the maintenance of chromatin morphology. Secondly, there are regions, termed "transcription factories," in the nucleus that are dynamically brought together by means of active transcription[102,123]. Thirdly, several noncoding RNAs[122,124] have been demonstrated to be involved in the formation of nuclear sub-compartments such as the nucleolus and paraspeckles[125] as well as the facilitation of higher-order chromosomal architecture[76,77,126]. However, the diversity of lncRNA mechanisms, their influence on nuclear architecture, and consequent cellular roles remain enigmatic.

1.5 Repetitive motifs and their roles in the context of lncRNA function

One interesting and common aspect of lncRNA and nuclear architecture biology left unexplored is the potential role of repetitive elements as eluded to above. For example, Xist was found to regulate X chromosome inactivation through the use of its repeat motifs [71,74,75]. Although lncRNAs harbor more repetitive elements than mRNAs and the rest of the genome[127], it is not known what the roles of these repeats might be. Similarly, the sub-domains of the nucleus, such as the centromere hubs as described above, and polycomb bodies, which consist

of multiple loci from multiple chromosomes that are silenced by the Polycomb complex, are formed around repetitive elements[128-131].

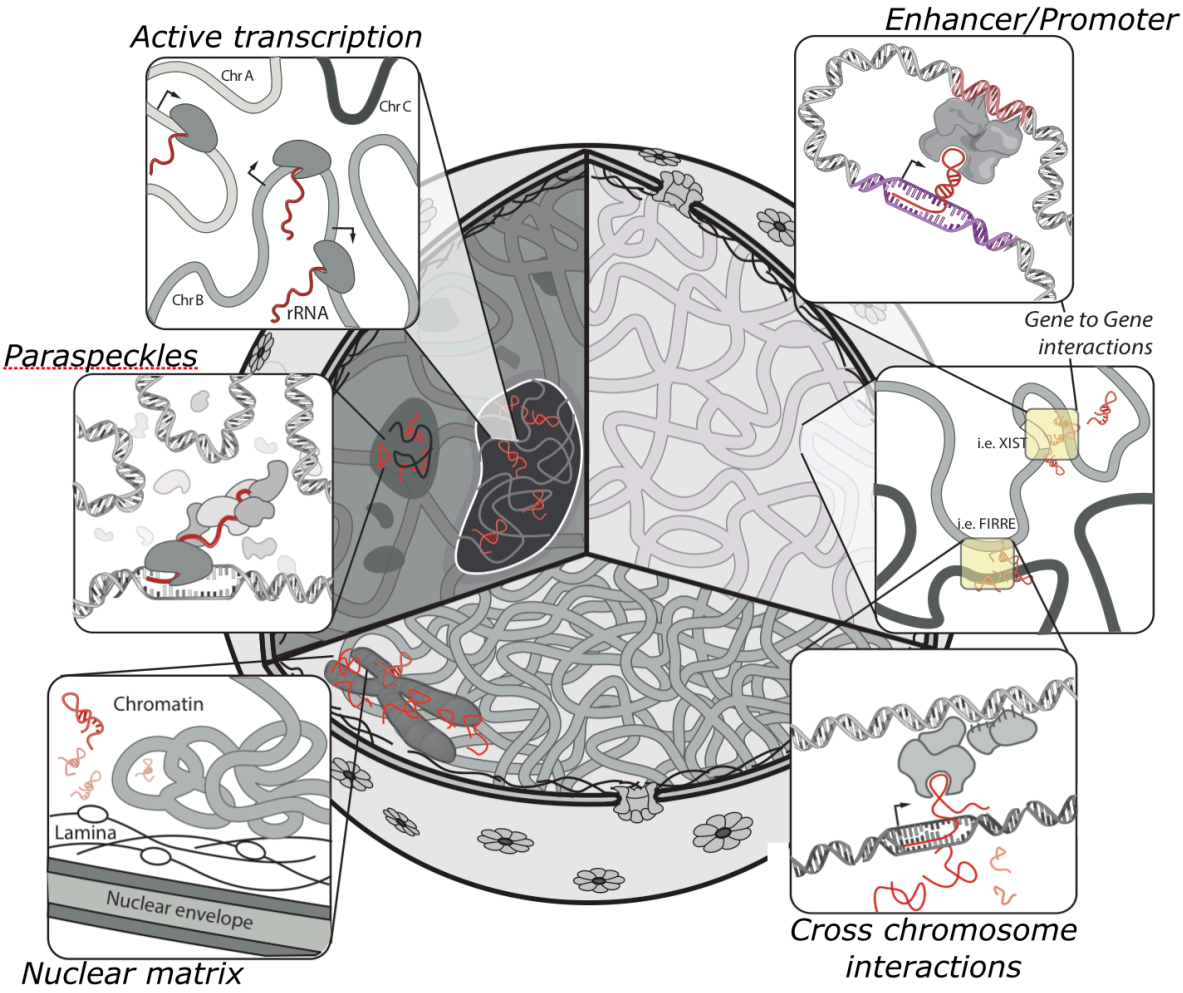


Figure 1: The proposed organizational roles of RNA in the nucleus.

What renders the findings above more interesting is the prevalence of repetitive sequences in the genome[132]. The types of repetitive elements are transposable elements (TEs), tandem repeats (TRs), and local repeats (LRs), the most widely studied of which are the TEs with numerous subclasses. TEs are repetitive DNA sequences that can either be immobile or mobile, which allows them propagate to new places in the genome[133]. Recent research has

revealed that TEs can play important roles in post-transcriptional regulation in primates[134,135], serve as a source of miRNA derivation[136,137] and impact the rewiring of regulatory networks in pluripotent cells[138-141].

Similar to TEs, TRs also constitute a large portion of the genome[142]; however, tandem repeats are classified according to their sizes: micro, mini, macro and megasatellite repeats. TRs are crucial for fundamental protein functions since 14% of all proteins contain them[143]. In addition, TRs can be used as genetic markers since they are highly variable across individuals but they have been intensely studied in the context of telomeres and centromeres, where TRs maintain the structural integrity of these important regions of the chromosomes[144]. Further investigation of noncentromeric TRs, such as DXZ4 and D4Z4, has illuminated how these repeat units can play roles in X chromosome dynamics and epigenetic regulation, respectively, and recruit transcription factors[145-152] and maintain genome stability via triggering gene silencing[153]. Interestingly, a significant portion of these repetitive regions, including TEs and DXZ4 and D4Z4, is transcribed into ncRNA, specifically lncRNAs[127,154,155].

The discovery of the presence of extensive silencing marks on TEs and TRs despite all the crucial roles they play as outlined above has brought new perspectives for the mechanisms of repetitive elements in the genome. One interesting area of biology has focused on the spatio-temporal silencing of TEs by the Polycomb complex to regulate cell fate specification[156,157]. This critical finding shed light on how the Polycomb complex might be functioning in the context of Polycomb bodies and how the Polycomb-regulated loci might associate in particular compartments in the nucleus[128-131,145]. Another area has been X chromosome inactivation: TEs aid in the formation of the heterochromatic core of the inactive X, and multiple TRs in the sequence of the Xist lncRNA mediate the localization and the recruitment of the factors

necessary for X chromosome inactivation[71,158-162]. Overall, these two approaches rendered repetitive elements more relevant for deciphering the complexity of the organization of the chromatin and nucleus.

Although the abundance of LRs besides TEs and TRs was well recognized[163], computational and experimental rigor has hindered further progress. The roles of LRs that exist in the introns and exons of lncRNAs can potentially be important both at the DNA level (for the intronic ones), by regulating the binding of protein factors, and at the RNA level (for exonic ones), by impacting the fate of the transcripts.

Throughout the body of this work, we set out to explore some of the unknown questions that are mentioned above. The first pressing question was to investigate whether lncRNAs can play a role in the establishment of a cell specific gene regulatory program, which is addressed in Chapter 2. The second question, which constitutes Chapter 3, aimed to understand how the lncRNA we discovered performed its role from a mechanistic perspective, which is largely unexplored in the lncRNA field. Thirdly, to further dissect the properties of the lncRNA under investigation and extrapolate the principles of nuclear organization, we asked how lncRNAs with unique repetitive motifs can help regulate nuclear infrastructure in Chapter 4.

1.6 References

1 Amaral, P. P., M. E. Dinger, T. R. Mercer, and J. S. Mattick. 2008. 'The eukaryotic genome as an RNA machine', *Science*, 319: 1787-9.

2 Ponting, C. P., P. L. Oliver, and W. Reik. 2009. 'Evolution and functions of long noncoding RNAs', *Cell*, 136: 629-41.

3 Weinberg, R. A., and S. Penman. 1968. 'Small molecular weight monodisperse nuclear RNA', *J Mol Biol*, 38: 289-304.

4 Paul, J., and J. D. Duerksen. 1975. 'Chromatin-associated RNA content of heterochromatin and euchromatin', *Mol Cell Biochem*, 9: 9-16.

5 Salditt-Georgieff, M., M.M Harpold, M.C. Wilson, J.E. Jr. Darnell. 1981. 'Large heterogeneous nuclear ribonucleic acid has three times as many 5′ caps as polyadenylic acid segments, and most caps do not enter polyribosomes', *Mol. Cell. Biol.* 1:179–87.

6 Salditt-Georgieff, M., J.E. Jr Darnell. 1982. 'Further evidence that the majority of primary nuclear RNA transcripts in mammalian cells do not contribute to mRNA', *Mol. Cell. Biol.* 2:701–7.

7 Smit, S., Widmann, J., Knight, R. 2007. 'Evolutionary rates vary among rRNA structural elements', *Nucleic Acids Res* 35: 3339–54.

8 Cole, J. R., B. Chai, T. L. Marsh, R. J. Farris, Q. Wang, S. A. Kulam, S. Chandra, D. M. McGarrell, T. M. Schmidt, G. M. Garrity, J. M. Tiedje, and Project Ribosomal Database. 2003. 'The Ribosomal Database Project (RDP-II): previewing a new autoaligner that allows regular updates and the new prokaryotic taxonomy', *Nucleic Acids Res*, 31: 442-3.

9 Schluenzen, F., A. Tocilj, R. Zarivach, J. Harms, M. Gluehmann, D. Janell, A. Bashan, H. Bartels, I. Agmon, F. Franceschi, and A. Yonath. 2000. 'Structure of functionally activated small ribosomal subunit at 3.3 angstroms resolution', *Cell*, 102: 615-23.

10 Korostelev, A., S. Trakhanov, M. Laurberg, and H. F. Noller. 2006. 'Crystal structure of a 70S ribosome-tRNA complex reveals functional interactions and rearrangements', *Cell*, 126: 1065-77.

11 Cech, T. 2000. 'Structural biology. The ribosome is a ribozyme', *Science* 289: 878–9.

12 Selmer, M., C. M. Dunham, F. V. th Murphy, A. Weixlbaumer, S. Petry, A. C. Kelley, J. R. Weir, and V. Ramakrishnan. 2006. 'Structure of the 70S ribosome complexed with mRNA and tRNA', *Science*, 313: 1935-42.

13 Ban, N., P. Nissen, J. Hansen, P. B. Moore, and T. A. Steitz. 2000. 'The complete atomic structure of the large ribosomal subunit at 2.4 A resolution', *Science*, 289: 905-20.

14 Itoh, Y., S. Sekine, S. Suetsugu, and S. Yokoyama. 2013. 'Tertiary structure of bacterial selenocysteine tRNA', *Nucleic Acids Res*, 41: 6729-38.

15 Agirrezabala, X., and J. Frank. 2009. 'Elongation in translation as a dynamic interaction among the ribosome, tRNA, and elongation factors EF-G and EF-Tu', *Q Rev Biophys*, 42: 159-200.

16 Clark, B. F. 2006. 'The crystal structure of tRNA', *J Biosci*, 31: 453-7.

17 Ahmad, K., and S. Henikoff. 2002. 'Epigenetic consequences of nucleosome dynamics', *Cell*, 111: 281-4.

18 van Wolfswinkel, J. C., and R. F. Ketting. 2010. 'The role of small non-coding RNAs in genome stability and chromatin organization', *J Cell Sci*, 123: 1825-39.

19 Ghildiyal, M., and P. D. Zamore. 2009. 'Small silencing RNAs: an expanding universe', *Nat Rev Genet*, 10: 94-108.

20 Kiss, T. 2001. 'Small nucleolar RNA-guided post-transcriptional modification of cellular RNAs', *EMBO J*, 20: 3617-22.

21 Thore, S., C. Mayer, C. Sauter, S. Weeks, and D. Suck. 2003. 'Crystal structures of the Pyrococcus abyssi Sm core and its complex with RNA. Common features of RNA binding in archaea and eukarya', *J Biol Chem*, 278: 1239-47.

22 Core, L. J., J. J. Waterfall, and J. T. Lis. 2008. 'Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters', *Science*, 322: 1845-8.

23 Kapranov, P., A. T. Willingham, and T. R. Gingeras. 2007. 'Genome-wide transcription and the implications for genomic organization', *Nat Rev Genet*, 8: 413-23.

24 Nagalakshmi, U., Z. Wang, K. Waern, C. Shou, D. Raha, M. Gerstein, and M. Snyder. 2008. 'The transcriptional landscape of the yeast genome defined by RNA sequencing', *Science*, 320: 1344-9.

25 Wilhelm, B. T., S. Marguerat, S. Watt, F. Schubert, V. Wood, I. Goodhead, C. J. Penkett, J. Rogers, and J. Bahler. 2008. 'Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution', *Nature*, 453: 1239-43.

26 Carninci, P., T. Kasukawa, S. Katayama, J. Gough, M. C. Frith, N. Maeda, R. Oyama, T. Ravasi, B. Lenhard, C. Wells, R. Kodzius, K. Shimokawa, V. B. Bajic, S. E. Brenner, S. Batalov, A. R. Forrest, M. Zavolan, M. J. Davis, L. G. Wilming, V. Aidinis, J. E. Allen, A. Ambesi-Impiombato, R. Apweiler, R. N. Aturaliya, T. L. Bailey, M. Bansal, L.

Baxter, K. W. Beisel, T. Bersano, H. Bono, A. M. Chalk, K. P. Chiu, V. Choudhary, A. Christoffels, D. R. Clutterbuck, M. L. Crowe, E. Dalla, B. P. Dalrymple, B. de Bono, G. Della Gatta, D. di Bernardo, T. Down, P. Engstrom, M. Fagiolini, G. Faulkner, C. F. Fletcher, T. Fukushima, M. Furuno, S. Futaki, M. Gariboldi, P. Georgii-Hemming, T. R. Gingeras, T. Gojobori, R. E. Green, S. Gustincich, M. Harbers, Y. Hayashi, T. K. Hensch, N. Hirokawa, D. Hill, L. Huminiecki, M. Iacono, K. Ikeo, A. Iwama, T. Ishikawa, M. Jakt, A. Kanapin, M. Katoh, Y. Kawasawa, J. Kelso, H. Kitamura, H. Kitano, G. Kollias, S. P. Krishnan, A. Kruger, S. K. Kummerfeld, I. V. Kurochkin, L. F. Lareau, D. Lazarevic, L. Lipovich, J. Liu, S. Liuni, S. McWilliam, M. Madan Babu, M. Madera, L. Marchionni, H. Matsuda, S. Matsuzawa, H. Miki, F. Mignone, S. Miyake, K. Morris, S. Mottagui- Tabar, N. Mulder, N. Nakano, H. Nakauchi, P. Ng, R. Nilsson, S. Nishiguchi, S. Nishikawa, F. Nori, O. Ohara, Y. Okazaki, V. Orlando, K. C. Pang, W. J. Pavan, G. Pavesi, G. Pesole, N. Petrovsky, S. Piazza, J. Reed, J. F. Reid, B. Z. Ring, M. Ringwald, B. Rost, Y. Ruan, S. L. Salzberg, A. Sandelin, C. Schneider, C. Schonbach, K. Sekiguchi, C. A. Semple, S. Seno, L. Sessa, Y. Sheng, Y. Shibata, H. Shimada, K. Shimada, D. Silva, B. Sinclair, S. Sperling, E. Stupka, K. Sugiura, R. Sultana, Y. Takenaka, K. Taki, K. Tammoja, S. L. Tan, S. Tang, M. S. Taylor, J. Tegner, S. A. Teichmann, H. R. Ueda, E. van Nimwegen, R. Verardo, C. L. Wei, K. Yagi, H. Yamanishi, E. Zabarovsky, S. Zhu, A. Zimmer, W. Hide, C. Bult, S. M. Grimmond, R. D. Teasdale, E. T. Liu, V. Brusic, J. Quackenbush, C. Wahlestedt, J. S. Mattick, D. A. Hume, C. Kai, D. Sasaki, Y. Tomaru, S. Fukuda, M. Kanamori-Katayama, M. Suzuki, J. Aoki, T. Arakawa, J. Iida, K. Imamura, M. Itoh, T. Kato, H. Kawaji, N. Kawagashira, T. Kawashima, M. Kojima, S. Kondo, H. Konno, K. Nakano, N. Ninomiya, T. Nishio, M. Okada, C. Plessy, K. Shibata, T. Shiraki, S. Suzuki, M. Tagami, K. Waki, A. Watahiki, Y. Okamura-Oho, H. Suzuki, J. Kawai, Y. Hayashizaki, Fantom Consortium, Riken Genome Exploration Research Group, and Group Genome Science. 2005. 'The transcriptional landscape of the mammalian genome', *Science*, 309: 1559-63.

27 Carninci, P., A. Sandelin, B. Lenhard, S. Katayama, K. Shimokawa, J. Ponjavic, C. A. Semple, M. S. Taylor, P. G. Engstrom, M. C. Frith, A. R. Forrest, W. B. Alkema, S. L. Tan, C. Plessy, R. Kodzius, T. Ravasi, T. Kasukawa, S. Fukuda, M. Kanamori-Katayama, Y. Kitazume, H. Kawaji, C. Kai, M. Nakamura, H. Konno, K. Nakano, S. Mottagui-Tabar, P. Arner, A. Chesi, S. Gustincich, F. Persichetti, H. Suzuki, S. M. Grimmond, C. A. Wells, V. Orlando, C. Wahlestedt, E. T. Liu, M. Harbers, J. Kawai, V. B. Bajic, D. A. Hume, and Y. Hayashizaki. 2006. 'Genome-wide analysis of mammalian promoter architecture and evolution', *Nat Genet*, 38: 626-35.

28 Clamp, M., B. Fry, M. Kamal, X. Xie, J. Cuff, M. F. Lin, M. Kellis, K. Lindblad-Toh, and E. S. Lander. 2007. 'Distinguishing protein-coding and noncoding genes in the human genome', *Proc Natl Acad Sci U S A*, 104: 19428-33.

29 WS227 Release Letter. WormBase. 10 August 2011.

30 NCBI (National Center for Bitechnology Information) Genome Database.

31 Kampa, D., J. Cheng, P. Kapranov, M. Yamanaka, S. Brubaker, S. Cawley, J. Drenkow, A. Piccolboni, S. Bekiranov, G. Helt, H. Tammana, and T. R. Gingeras. 2004. 'Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22', *Genome Res*, 14: 331-42.

32 Ravasi, T., H. Suzuki, K. C. Pang, S. Katayama, M. Furuno, R. Okunishi, S. Fukuda, K. Ru, M. C. Frith, M. M. Gongora, S. M. Grimmond, D. A. Hume, Y. Hayashizaki, and J. S. Mattick. 2006. 'Experimental validation of the regulated expression of large numbers of non-coding RNAs from the mouse genome', *Genome Res*, 16: 11-9.

33 Mattick, J. S., and I. V. Makunin. 2006. 'Non-coding RNA', *Hum Mol Genet*, 15 Spec No 1: R17-29.

34 Dinger, M. E., K. C. Pang, T. R. Mercer, M. L. Crowe, S. M. Grimmond, and J. S. Mattick. 2009. 'NRED: a database of long noncoding RNA expression', *Nucleic Acids Res*, 37: D122-6.

35 Mortazavi, A., B. A. Williams, K. McCue, L. Schaeffer, and B. Wold. 2008. 'Mapping and quantifying mammalian transcriptomes by RNA-Seq', *Nat Methods*, 5: 621-8.

36 Marioni, J. C., C. E. Mason, S. M. Mane, M. Stephens, and Y. Gilad. 2008. 'RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays', *Genome Res*, 18: 1509-17.

37 Trapnell, C., L. Pachter, and S. L. Salzberg. 2009. 'TopHat: discovering splice junctions with RNA-Seq', *Bioinformatics*, 25: 1105-11.

38 Guttman, M., M. Garber, J. Z. Levin, J. Donaghey, J. Robinson, X. Adiconis, L. Fan, M. J. Koziol, A. Gnirke, C. Nusbaum, J. L. Rinn, E. S. Lander, and A. Regev. 2010. 'Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs', *Nat Biotechnol*, 28: 503-10.

39 Garber, M., M. G. Grabherr, M. Guttman, and C. Trapnell. 2011. 'Computational methods for transcriptome annotation and quantification using RNA-seq', *Nat Methods*, 8: 469-77.

40 Cabili, M. N., C. Trapnell, L. Goff, M. Koziol, B. Tazon-Vega, A. Regev, and J. L. Rinn. 2011. 'Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses', *Genes Dev*, 25: 1915-27.

41 Rinn, J. L., and H. Y. Chang. 2012. 'Genome regulation by long noncoding RNAs', *Annu Rev Biochem*, 81: 145-66.

42 Guttman, M., I. Amit, M. Garber, C. French, M. F. Lin, D. Feldser, M. Huarte, O. Zuk, B. W. Carey, J. P. Cassady, M. N. Cabili, R. Jaenisch, T. S. Mikkelsen, T. Jacks, N. Hacohen, B. E. Bernstein, M. Kellis, A. Regev, J. L. Rinn, and E. S. Lander. 2009.

'Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals', *Nature*, 458: 223-7.

43 Khalil, A. M., M. Guttman, M. Huarte, M. Garber, A. Raj, D. Rivea Morales, K. Thomas, A. Presser, B. E. Bernstein, A. van Oudenaarden, A. Regev, E. S. Lander, and J. L. Rinn. 2009. 'Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression', *Proc Natl Acad Sci U S A*, 106: 11667- 72.

44 Lin, M. F., J. W. Carlson, M. A. Crosby, B. B. Matthews, C. Yu, S. Park, K. H. Wan, A. J. Schroeder, L. S. Gramates, S. E. St Pierre, M. Roark, K. L. Wiley, Jr., R. J. Kulathinal, P. Zhang, K. V. Myrick, J. V. Antone, S. E. Celniker, W. M. Gelbart, and M. Kellis. 2007. 'Revisiting the protein-coding gene catalog of Drosophila melanogaster using 12 fly genomes', *Genome Res*, 17: 1823-36.

45 Lin, M. F., A. N. Deoras, M. D. Rasmussen, and M. Kellis. 2008. 'Performance and scalability of discriminative metrics for comparative gene identification in 12 Drosophila genomes', *PLoS Comput Biol*, 4: e1000067.

46 Ingolia, N. T., L. F. Lareau, and J. S. Weissman. 2011. 'Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes', *Cell*, 147: 789-802.

47 Nie, L., H. J. Wu, J. M. Hsu, S. S. Chang, A. M. Labaff, C. W. Li, Y. Wang, J. L. Hsu, and M. C. Hung. 2012. 'Long non-coding RNAs: versatile master regulators of gene expression and crucial players in cancer', *Am J Transl Res*, 4: 127-50.

48 Derrien, T., R. Johnson, G. Bussotti, A. Tanzer, S. Djebali, H. Tilgner, G. Guernec, D. Martin, A. Merkel, D. G. Knowles, J. Lagarde, L. Veeravalli, X. Ruan, Y. Ruan, T. Lassmann, P. Carninci, J. B. Brown, L. Lipovich, J. M. Gonzalez, M. Thomas, C. A. Davis, R. Shiekhattar, T. R. Gingeras, T. J. Hubbard, C. Notredame, J. Harrow, and R. Guigo. 2012. 'The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression', *Genome Res*, 22: 1775-89.

49 van Heesch, S., M. van Iterson, J. Jacobi, S. Boymans, P. B. Essers, E. de Bruijn, W. Hao, A. W. MacInnes, E. Cuppen, and M. Simonis. 2014. 'Extensive localization of long noncoding RNAs to the cytosol and mono- and polyribosomal complexes', *Genome Biol*, 15: R6.

50 Cabili, M. N., M. C. Dunagin, P. D. McClanahan, A. Biaesch, O. Padovan-Merhar, A. Regev, J. L. Rinn, and A. Raj. 2015. 'Localization and abundance analysis of human lncRNAs at single-cell and single-molecule resolution', *Genome Biol*, 16: 20.

51 Haerty, W., and C. P. Ponting. 2015. 'Unexpected selection to retain high GC content and splicing enhancers within exons of multiexonic lncRNA loci', *RNA*.

52 Marques, A. C., and C. P. Ponting. 2009. 'Catalogues of mammalian long noncoding RNAs: modest conservation and incompleteness', *Genome Biol*, 10: R124.

53 Ulitsky, I., and D. P. Bartel. 2013. 'lincRNAs: genomics, evolution, and mechanisms', *Cell*, 154: 26-46.

54 Ulitsky, I., A. Shkumatava, C. H. Jan, H. Sive, and D. P. Bartel. 2011. 'Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution', *Cell*, 147: 1537-50.

55 Johnsson, P., Lipovich, L., Grandér, D., & Morris, K. V. 2014. 'Evolutionary conservation of long noncoding RNAs; sequence, structure, function', *Biochimica et Biophysica Acta*, *1840*(3): 1063–1071.

56 Zappulla, D. C., and T. R. Cech. 2004. 'Yeast telomerase RNA: a flexible scaffold for protein subunits', *Proc Natl Acad Sci U S A*, 101: 10024-9.

57 Kutter, C., S. Watt, K. Stefflova, M. D. Wilson, A. Goncalves, C. P. Ponting, D. T. Odom, and A. C. Marques. 2012. 'Rapid turnover of long noncoding RNAs and the evolution of gene expression', *PLoS Genet*, 8: e1002841.

58 Kaushik, K., V. E. Leonard, S. Kv, M. K. Lalwani, S. Jalali, A. Patowary, A. Joshi, V. Scaria, and S. Sivasubbu. 2013. 'Dynamic expression of long non-coding RNAs (lncRNAs) in adult zebrafish', *PLoS One*, 8: e83616.

59 Bao, J., J. Wu, A. S. Schuster, G. W. Hennig, and W. Yan. 2013. 'Expression profiling reveals developmentally regulated lncRNA repertoire in the mouse male germline', *Biol Reprod*, 89: 107.

60 Xia, F., F. Dong, Y. Yang, A. Huang, S. Chen, D. Sun, S. Xiong, and J. Zhang. 2014. 'Dynamic transcription of long non-coding RNA genes during CD4+ T cell development and activation', *PLoS One*, 9: e101588.

61 Sun, L., L. A. Goff, C. Trapnell, R. Alexander, K. A. Lo, E. Hacisuleyman, M. Sauvageau, B. Tazon-Vega, D. R. Kelley, D. G. Hendrickson, B. Yuan, M. Kellis, H. F. Lodish, and J. L. Rinn. 2013. 'Long noncoding RNAs regulate adipogenesis', *Proc Natl Acad Sci U S A*, 110: 3387-92.

62 Sauvageau, M., L. A. Goff, S. Lodato, B. Bonev, A. F. Groff, C. Gerhardinger, D. B. Sanchez-Gomez, E. Hacisuleyman, E. Li, M. Spence, S. C. Liapis, W. Mallard, M. Morse, M. R. Swerdel, M. F. D'Ecclessis, J. C. Moore, V. Lai, G. Gong, G. D. Yancopoulos, D. Frendewey, M. Kellis, R. P. Hart, D. M. Valenzuela, P. Arlotta, and J. L. Rinn. 2013. 'Multiple knockout mouse models reveal lincRNAs are required for life and brain development', *Elife*, 2: e01749.

63 Hacisuleyman, E., M. N. Cabili, and J. L. Rinn. 2012. 'A Keystone for ncRNA', *Genome Biol*, 13: 315.

64 Chen, J., R. Wang, K. Zhang, and L. B. Chen. 2014. 'Long non-coding RNAs in non- small cell lung cancer as biomarkers and therapeutic targets', *J Cell Mol Med*, 18: 2425-36.

65 Hauptman, N., and D. Glavac. 2013. 'MicroRNAs and long non-coding RNAs: prospects in diagnostics and therapy of cancer', *Radiol Oncol*, 47: 311-8.

66 Shen, Z., Q. Li, H. Deng, D. Lu, H. Song, and J. Guo. 2014. 'Long non-coding RNA profiling in laryngeal squamous cell carcinoma and its clinical significance: potential biomarkers for LSCC', *PLoS One*, 9: e108237.

67 Mirza, A. H., S. Kaur, C. A. Brorsson, and F. Pociot. 2014. 'Effects of GWAS- associated genetic variants on lncRNAs within IBD and T1D candidate loci', *PLoS One*, 9: e105723.

68 Qureshi, I. A., and M. F. Mehler. 2012. 'Emerging roles of non-coding RNAs in brain evolution, development, plasticity and disease', *Nat Rev Neurosci*, 13: 528-41.

69 Brown, C. J., A. Ballabio, J. L. Rupert, R. G. Lafreniere, M. Grompe, R. Tonlorenzi, and H. F. Willard. 1991. 'A gene from the region of the human X inactivation centre is expressed exclusively from the inactive X chromosome', *Nature*, 349: 38-44.

70 Brockdorff, N., A. Ashworth, G. F. Kay, V. M. McCabe, D. P. Norris, P. J. Cooper, S. Swift, and S. Rastan. 1992. 'The product of the mouse Xist gene is a 15 kb inactive X-specific transcript containing no conserved ORF and located in the nucleus', *Cell*, 71: 515-26.

71 Zhao, J., B. K. Sun, J. A. Erwin, J. J. Song, and J. T. Lee. 2008. 'Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome', *Science*, 322: 750-6.

72 Davidovich, C., L. Zheng, K. J. Goodrich, and T. R. Cech. 2013. 'Promiscuous RNA binding by Polycomb repressive complex 2', *Nat Struct Mol Biol*, 20: 1250-7.

73 Kaneko, S., J. Son, S. S. Shen, D. Reinberg, and R. Bonasio. 2013. 'PRC2 binds active promoters and contacts nascent RNAs in embryonic stem cells', *Nat Struct Mol Biol*, 20: 1258-64.

74 Jeon, Y., K. Sarma, and J. T. Lee. 2012. 'New and Xisting regulatory mechanisms of X chromosome inactivation', *Curr Opin Genet Dev*, 22: 62-71.

75 Plath, K., S. Mlynarczyk-Evans, D. A. Nusinow, and B. Panning. 2002. 'Xist RNA and the mechanism of X chromosome inactivation', *Annu Rev Genet*, 36: 233-78.

76 Simon, M. D., S. F. Pinter, R. Fang, K. Sarma, M. Rutenberg-Schoenberg, S. K. Bowman, B. A. Kesner, V. K. Maier, R. E. Kingston, and J. T. Lee. 2013. 'High-resolution Xist binding maps reveal two-step spreading during X-chromosome inactivation', *Nature*, 504: 465-9.

77 Engreitz, J. M., A. Pandya-Jones, P. McDonel, A. Shishkin, K. Sirokman, C. Surka, S. Kadri, J. Xing, A. Goren, E. S. Lander, K. Plath, and M. Guttman. 2013. 'The Xist lncRNA exploits three-dimensional genome architecture to spread across the X chromosome', *Science*, 341: 1237973.

78 Kino, T., D. E. Hurt, T. Ichijo, N. Nader, and G. P. Chrousos. 2010. 'Noncoding RNA gas5 is a growth arrest- and starvation-associated repressor of the glucocorticoid receptor', *Sci Signal*, 3: ra8.

79 Hung, T., Y. Wang, M. F. Lin, A. K. Koegel, Y. Kotake, G. D. Grant, H. M. Horlings, N. Shah, C. Umbricht, P. Wang, Y. Wang, B. Kong, A. Langerod, A. L. Borresen-Dale, S. K. Kim, M. van de Vijver, S. Sukumar, M. L. Whitfield, M. Kellis, Y. Xiong, D. J. Wong, and H. Y. Chang. 2011. 'Extensive and coordinated transcription of noncoding RNAs within cell-cycle promoters', *Nat Genet*, 43: 621-9.

80 Rinn, J. L., M. Kertesz, J. K. Wang, S. L. Squazzo, X. Xu, S. A. Brugmann, L. H. Goodnough, J. A. Helms, P. J. Farnham, E. Segal, and H. Y. Chang. 2007. 'Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs', *Cell*, 129: 1311-23.

81 Tsai, M. C., O. Manor, Y. Wan, N. Mosammaparast, J. K. Wang, F. Lan, Y. Shi, E. Segal, and H. Y. Chang. 2010. 'Long noncoding RNA as modular scaffold of histone modification complexes', *Science*, 329: 689-93.

82 Huarte, M., M. Guttman, D. Feldser, M. Garber, M. J. Koziol, D. Kenzelmann-Broz, A. M. Khalil, O. Zuk, I. Amit, M. Rabani, L. D. Attardi, A. Regev, E. S. Lander, T. Jacks, and J. L. Rinn. 2010. 'A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response', *Cell*, 142: 409-19.

83 Lam, M. T., W. Li, M. G. Rosenfeld, and C. K. Glass. 2014. 'Enhancer RNAs and regulated transcriptional programs', *Trends Biochem Sci*, 39: 170-82.

84 Hsieh, C. L., T. Fei, Y. Chen, T. Li, Y. Gao, X. Wang, T. Sun, C. J. Sweeney, G. S. Lee, S. Chen, S. P. Balk, X. S. Liu, M. Brown, and P. W. Kantoff. 2014. 'Enhancer RNAs participate in androgen receptor-driven looping that selectively enhances gene activation', *Proc Natl Acad Sci U S A*, 111: 7319-24.

85 Lai, F., and R. Shiekhattar. 2014. 'Enhancer RNAs: the new molecules of transcription', *Curr Opin Genet Dev*, 25: 38-42.

86 Wang, K. C., Y. W. Yang, B. Liu, A. Sanyal, R. Corces-Zimmerman, Y. Chen, B. R. Lajoie, A. Protacio, R. A. Flynn, R. A. Gupta, J. Wysocka, M. Lei, J. Dekker, J. A. Helms, and H. Y. Chang. 2011. 'A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression', *Nature*, 472: 120-4.

87 Washietl, S., I. L. Hofacker, and P. F. Stadler. 2005. 'Fast and reliable prediction of noncoding RNAs', *Proc Natl Acad Sci U S A*, 102: 2454-9.

88 De Santa, F., I. Barozzi, F. Mietton, S. Ghisletti, S. Polletti, B. K. Tusi, H. Muller, J. Ragoussis, C. L. Wei, and G. Natoli. 2010. 'A large fraction of extragenic RNA pol II transcription sites overlap enhancers', *PLoS Biol*, 8: e1000384.

89 Davis, C. A., and M. Ares, Jr. 2006. 'Accumulation of unstable promoter-associated transcripts upon loss of the nuclear exosome subunit Rrp6p in Saccharomyces cerevisiae', *Proc Natl Acad Sci U S A*, 103: 3262-7.

90 Chu, C., J. Quinn, and H. Y. Chang. 2012. 'Chromatin isolation by RNA purification (ChIRP)', *J Vis Exp*.

91 Simon, M. D., C. I. Wang, P. V. Kharchenko, J. A. West, B. A. Chapman, A. A. Alekseyenko, M. L. Borowsky, M. I. Kuroda, and R. E. Kingston. 2011. 'The genomic binding sites of a noncoding RNA', *Proc Natl Acad Sci U S A*, 108: 20497-502.

92 Engreitz, J., E. S. Lander, and M. Guttman. 2015. 'RNA antisense purification (RAP) for mapping RNA interactions with chromatin', *Methods Mol Biol*, 1262: 183-97.

93 Martianov, I., A. Ramadass, A. Serra Barros, N. Chow, and A. Akoulitchev. 2007. 'Repression of the human dihydrofolate reductase gene by a non-coding interfering transcript', *Nature*, 445: 666-70.

94 Willingham, A. T., A. P. Orth, S. Batalov, E. C. Peters, B. G. Wen, P. Aza-Blanc, J. B. Hogenesch, and P. G. Schultz. 2005. 'A strategy for probing the function of noncoding RNAs finds a repressor of NFAT', *Science*, 309: 1570-3.

95 Hellwig, S., and B. L. Bass. 2008. 'A starvation-induced noncoding RNA modulates expression of Dicer-regulated genes', *Proc Natl Acad Sci U S A*, 105: 12897-902.

96 Ishizuka, A., Y. Hasegawa, K. Ishida, K. Yanaka, and S. Nakagawa. 2014. 'Formation of nuclear bodies by the lncRNA Gomafu-associating proteins Celf3 and SF1', *Genes Cells*, 19: 704-21.

97 Wang, X., S. Arai, X. Song, D. Reichart, K. Du, G. Pascual, P. Tempst, M. G. Rosenfeld, C. K. Glass, and R. Kurokawa. 2008. 'Induced ncRNAs allosterically modify RNA-binding proteins in cis to inhibit transcription', *Nature*, 454: 126-30.

98 Azzalin, C. M., P. Reichenbach, L. Khoriauli, E. Giulotto, and J. Lingner. 2007. 'Telomeric repeat containing RNA and RNA surveillance factors at mammalian chromosome ends', *Science*, 318: 798-801.

99 Cremer, T., C. Cremer, H. Baumann, E. K. Luedtke, K. Sperling, V. Teuber, and C. Zorn. 1982. 'Rabl's model of the interphase chromosome arrangement tested in Chinese hamster cells by premature chromosome condensation and laser-UV-microbeam experiments', *Hum Genet*, 60: 46-56.

100 Borden, J., and L. Manuelidis. 1988. 'Movement of the X chromosome in epilepsy', *Science*, 242: 1687-91.

101 Cremer, T., P. Lichter, J. Borden, D. C. Ward, and L. Manuelidis. 1988. 'Detection of chromosome aberrations in metaphase and interphase tumor cells by in situ hybridization using chromosome-specific library probes', *Hum Genet*, 80: 235-46.

102 Simonis, M., P. Klous, E. Splinter, Y. Moshkin, R. Willemsen, E. de Wit, B. van Steensel, and W. de Laat. 2006. 'Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C)', *Nat Genet*, 38: 1348-54.

103 Croft, J. A., J. M. Bridger, S. Boyle, P. Perry, P. Teague, and W. A. Bickmore. 1999. 'Differences in the localization and morphology of chromosomes in the human nucleus', *J Cell Biol*, 145: 1119-31.

104 Schneider, R., and R. Grosschedl. 2007. 'Dynamics and interplay of nuclear architecture, genome organization, and gene expression', *Genes Dev*, 21: 3027-43.

105 Dixon, J. R., S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, M. Hu, J. S. Liu, and B. Ren. 2012. 'Topological domains in mammalian genomes identified by analysis of chromatin interactions', *Nature*, 485: 376-80.

106 Pope, B. D., T. Ryba, V. Dileep, F. Yue, W. Wu, O. Denas, D. L. Vera, Y. Wang, R. S. Hansen, T. K. Canfield, R. E. Thurman, Y. Cheng, G. Gulsoy, J. H. Dennis, M. P. Snyder, J. A. Stamatoyannopoulos, J. Taylor, R. C. Hardison, T. Kahveci, B. Ren, and D. M. Gilbert. 2014. 'Topologically associating domains are stable units of replication-timing regulation', *Nature*, 515: 402-5.

107 Brown, K. E., S. Amoils, J. M. Horn, V. J. Buckle, D. R. Higgs, M. Merkenschlager, and A. G. Fisher. 2001. 'Expression of alpha- and beta-globin genes occurs within different nuclear domains in haemopoietic cells', *Nat Cell Biol*, 3: 602-6.

108 Meshorer, E., and T. Misteli. 2006. 'Chromatin in pluripotent embryonic stem cells and differentiation', *Nat Rev Mol Cell Biol*, 7: 540-6.

109 Meshorer, E., D. Yellajoshula, E. George, P. J. Scambler, D. T. Brown, and T. Misteli. 2006. 'Hyperdynamic plasticity of chromatin proteins in pluripotent embryonic stem cells', *Dev Cell*, 10: 105-16.

110 Kosak, S. T., J. A. Skok, K. L. Medina, R. Riblet, M. M. Le Beau, A. G. Fisher, and H. Singh. 2002. 'Subnuclear compartmentalization of immunoglobulin loci during lymphocyte development', *Science*, 296: 158-62.

111 Spilianakis, C. G., M. D. Lalioti, T. Town, G. R. Lee, and R. A. Flavell. 2005. 'Interchromosomal associations between alternatively expressed loci', *Nature*, 435: 637-45.

112 Spivakov, M., and A. G. Fisher. 2007. 'Epigenetic signatures of stem-cell identity', *Nat Rev Genet*, 8: 263-71.

113 Williams, R. R., V. Azuara, P. Perry, S. Sauer, M. Dvorkina, H. Jorgensen, J. Roix, P. McQueen, T. Misteli, M. Merkenschlager, and A. G. Fisher. 2006. 'Neural induction promotes large-scale chromatin reorganisation of the Mash1 locus', *J Cell Sci*, 119: 132-40.

114 Dixon, J. R., I. Jung, S. Selvaraj, Y. Shen, J. E. Antosiewicz-Bourget, A. Y. Lee, Z. Ye, A. Kim, N. Rajagopal, W. Xie, Y. Diao, J. Liang, H. Zhao, V. V. Lobanenkov, J. R. Ecker, J. A. Thomson, and B. Ren. 2015. 'Chromatin architecture reorganization during stem cell differentiation', *Nature*, 518: 331-6.

115 Probst, A. V., F. Santos, W. Reik, G. Almouzni, and W. Dean. 2007. 'Structural differences in centromeric heterochromatin are spatially reconciled on fertilisation in the mouse zygote', *Chromosoma*, 116: 403-15.

116 Martin, C., N. Beaujean, V. Brochard, C. Audouard, D. Zink, and P. Debey. 2006. 'Genome restructuring in mouse embryos during reprogramming and early development', *Dev Biol*, 292: 317-32.

117 Wiblin, A. E., W. Cui, A. J. Clark, and W. A. Bickmore. 2005. 'Distinctive nuclear organisation of centromeres and regions involved in pluripotency in human embryonic stem cells', *J Cell Sci*, 118: 3861-8.

118 Apostolou, E., F. Ferrari, R. M. Walsh, O. Bar-Nur, M. Stadtfeld, S. Cheloufi, H. T. Stuart, J. M. Polo, T. K. Ohsumi, M. L. Borowsky, P. V. Kharchenko, P. J. Park, and K. Hochedlinger. 2013. 'Genome-wide chromatin interactions of the Nanog locus in pluripotency, differentiation, and reprogramming', *Cell Stem Cell*, 12: 699-712.

119 Bouvier, D., J. Hubert, A. P. Seve, and M. Bouteille. 1985. 'Nuclear RNA-associated proteins and their relationship to the nuclear matrix and related structures in HeLa cells', *Can J Biochem Cell Biol*, 63: 631-43.

120 Nickerson, J. A., G. Krochmalnic, K. M. Wan, and S. Penman. 1989. 'Chromatin architecture and nuclear RNA', *Proc Natl Acad Sci U S A*, 86: 177-81.

121 Pederson, T., and J. S. Bhorjee. 1979. 'Evidence for a role of RNA in eukaryotic chromosome structure. Metabolically stable, small nuclear RNA species are covalently linked to chromosomal DNA in HeLa cells', *J Mol Biol*, 128: 451-80.

122 Umlauf, D., P. Fraser, and T. Nagano. 2008. 'The role of long non-coding RNAs in chromatin structure and gene regulation: variations on a theme', *Biol Chem*, 389: 323-31.

123 Osborne, C. S., L. Chakalova, K. E. Brown, D. Carter, A. Horton, E. Debrand, B. Goyenechea, J. A. Mitchell, S. Lopes, W. Reik, and P. Fraser. 2004. 'Active genes dynamically colocalize to shared sites of ongoing transcription', *Nat Genet*, 36: 1065-71.

124 Wilusz, J. E., H. Sunwoo, and D. L. Spector. 2009. 'Long noncoding RNAs: functional surprises from the RNA world', *Genes Dev*, 23: 1494-504.

125 Mao, Y. S., H. Sunwoo, B. Zhang, and D. L. Spector. 2011. 'Direct visualization of the co-transcriptional assembly of a nuclear body by noncoding RNAs', *Nat Cell Biol*, 13: 95-101.

126 Delpretti, S., T. Montavon, M. Leleu, E. Joye, A. Tzika, M. Milinkovitch, and D. Duboule. 2013. 'Multiple enhancers regulate Hoxd genes and the Hotdog LncRNA during cecum budding', *Cell Rep*, 5: 137-50.

127 Kelley, D., and J. Rinn. 2012. 'Transposable elements reveal a stem cell-specific class of long noncoding RNAs', *Genome Biol*, 13: R107.

128 Cleard, F., Y. Moshkin, F. Karch, and R. K. Maeda. 2006. 'Probing long-distance regulatory interactions in the Drosophila melanogaster bithorax complex using Dam identification', *Nat Genet*, 38: 931-5.

129 Comet, I., B. Schuettengruber, T. Sexton, and G. Cavalli. 2011. 'A chromatin insulator driving three-dimensional Polycomb response element (PRE) contacts and Polycomb association with the chromatin fiber', *Proc Natl Acad Sci U S A*, 108: 2294-9.

130 Tiwari, V. K., K. M. McGarvey, J. D. Licchesi, J. E. Ohm, J. G. Herman, D. Schubeler, and S. B. Baylin. 2008. 'PcG proteins, DNA methylation, and gene repression by chromatin looping', *PLoS Biol*, 6: 2911-27.

131 Lanzuolo, C., V. Roure, J. Dekker, F. Bantignies, and V. Orlando. 2007. 'Polycomb response elements mediate the formation of chromosome higher-order structures in the bithorax complex', *Nat Cell Biol*, 9: 1167-74.

132 de Koning, A. P., W. Gu, T. A. Castoe, M. A. Batzer, and D. D. Pollock. 2011. 'Repetitive elements may comprise over two-thirds of the human genome', *PLoS Genet*, 7: e1002384.

133 Doolittle, W. F., and C. Sapienza. 1980. 'Selfish genes, the phenotype paradigm and genome evolution', *Nature*, 284: 601-3.

134 Athanasiadis, A., A. Rich, and S. Maas. 2004. 'Widespread A-to-I RNA editing of Alu-containing mRNAs in the human transcriptome', *PLoS Biol*, 2: e391.

135 Bazak, L., E. Y. Levanon, and E. Eisenberg. 2014. 'Genome-wide analysis of Alu editability', *Nucleic Acids Res*, 42: 6876-84.

136 Piriyapongsa, J., and I. K. Jordan. 2007. 'A family of human microRNA genes from miniature inverted-repeat transposable elements', *PLoS One*, 2: e203.

137 Devor, E. J., A. S. Peek, W. Lanier, and P. B. Samollow. 2009. 'Marsupial-specific microRNAs evolved from marsupial-specific transposable elements', *Gene*, 448: 187-91.

138 Feschotte, C. 2008. 'Transposable elements and the evolution of regulatory networks', *Nat Rev Genet*, 9: 397-405.

139 Schmidt, D., P. C. Schwalie, M. D. Wilson, B. Ballester, A. Goncalves, C. Kutter, G. D. Brown, A. Marshall, P. Flicek, and D. T. Odom. 2012. 'Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages', *Cell*, 148: 335-48.

140 Wang, T., J. Zeng, C. B. Lowe, R. G. Sellers, S. R. Salama, M. Yang, S. M. Burgess, R. K. Brachmann, and D. Haussler. 2007. 'Species-specific endogenous retroviruses shape the transcriptional network of the human tumor suppressor protein p53', *Proc Natl Acad Sci U S A*, 104: 18613-8.

141 Kunarso, G., N. Y. Chia, J. Jeyakani, C. Hwang, X. Lu, Y. S. Chan, H. H. Ng, and G. Bourque. 2010. 'Transposable elements have rewired the core regulatory network of human embryonic stem cells', *Nat Genet*, 42: 631-4.

142 Warburton, P. E., D. Hasson, F. Guillem, C. Lescale, X. Jin, and G. Abrusan. 2008. 'Analysis of the largest tandemly repeated DNA families in the human genome', *BMC Genomics*, 9: 533.

143 Pellegrini, M., E. M. Marcotte, and T. O. Yeates. 1999. 'A fast algorithm for genome- wide analysis of proteins with repeated sequences', *Proteins*, 35: 440-6.

144 Blackburn, E. H. 1984. 'The molecular structure of centromeres and telomeres', *Annu Rev Biochem*, 53: 163-94.

145 Cabianca, D. S., V. Casa, B. Bodega, A. Xynos, E. Ginelli, Y. Tanaka, and D. Gabellini. 2012. 'A long ncRNA links copy number variation to a polycomb/trithorax epigenetic switch in FSHD muscular dystrophy', *Cell*, 149: 819-31.

146 Chadwick, B. P. 2009. 'Macrosatellite epigenetics: the two faces of DXZ4 and D4Z4', *Chromosoma*, 118: 675-81.

147 Filippova, G. N. 2008. 'Genetics and epigenetics of the multifunctional protein CTCF', *Curr Top Dev Biol*, 80: 337-60.

148 Gabellini, D., M. R. Green, and R. Tupler. 2002. 'Inappropriate gene activation in FSHD: a repressor complex binds a chromosomal repeat deleted in dystrophic muscle', *Cell*, 110: 339-48.

149 Bodega, B., G. D. Ramirez, F. Grasser, S. Cheli, S. Brunelli, M. Mora, R. Meneveri, A. Marozzi, S. Mueller, E. Battaglioli, and E. Ginelli. 2009. 'Remodeling of the chromatin structure of the facioscapulohumeral muscular dystrophy (FSHD) locus and upregulation of FSHD-related gene 1 (FRG1) expression during human myogenic differentiation', *BMC Biol*, 7: 41.

150 Zeng, W., J. C. de Greef, Y. Y. Chen, R. Chien, X. Kong, H. C. Gregson, S. T. Winokur, A. Pyle, K. D. Robertson, J. A. Schmiesing, V. E. Kimonis, J. Balog, R. R. Frants, A. R. Ball, Jr., L. F. Lock, P. J. Donovan, S. M. van der Maarel, and K. Yokomori. 2009. 'Specific loss of histone H3 lysine 9 trimethylation and HP1gamma/cohesin binding at D4Z4 repeats is associated with facioscapulohumeral dystrophy (FSHD)', *PLoS Genet*, 5: e1000559.

151 van Overveld, P. G., R. J. Lemmers, L. A. Sandkuijl, L. Enthoven, S. T. Winokur, F. Bakels, G. W. Padberg, G. J. van Ommen, R. R. Frants, and S. M. van der Maarel. 2003. 'Hypomethylation of D4Z4 in 4q-linked and non-4q-linked facioscapulohumeral muscular dystrophy', *Nat Genet*, 35: 315-7.

152 Ottaviani, A., S. Rival-Gervier, A. Boussouar, A. M. Foerster, D. Rondier, S. Sacconi, C. Desnuelle, E. Gilson, and F. Magdinier. 2009. 'The D4Z4 macrosatellite repeat acts as a CTCF and A-type lamins-dependent insulator in facio-scapulo-humeral dystrophy', *PLoS Genet*, 5: e1000394.

153 Yang, Q., Q. A. Ye, and Y. Liu. 2015. 'Mechanism of siRNA production from repetitive DNA', *Genes Dev*, 29: 526-37.

154 Snider, L., A. Asawachaicharn, A. E. Tyler, L. N. Geng, L. M. Petek, L. Maves, D. G. Miller, R. J. Lemmers, S. T. Winokur, R. Tawil, S. M. van der Maarel, G. N. Filippova, and S. J. Tapscott. 2009. 'RNA transcripts, miRNA-sized fragments and proteins

produced from D4Z4 units: new candidates for the pathophysiology of facioscapulohumeral dystrophy', *Hum Mol Genet*, 18: 2414-30.

155 Chadwick, B. P. 2008. 'DXZ4 chromatin adopts an opposing conformation to that of the surrounding chromosome and acquires a novel inactive X-specific role involving CTCF and antisense transcripts', *Genome Res*, 18: 1259-69.

156 Macfarlan, T. S., W. D. Gifford, S. Driscoll, K. Lettieri, H. M. Rowe, D. Bonanomi, A. Firth, O. Singer, D. Trono, and S. L. Pfaff. 2012. 'Embryonic stem cell potency fluctuates with endogenous retrovirus activity', *Nature*, 487: 57-63.

157 Day, D. S., L. J. Luquette, P. J. Park, and P. V. Kharchenko. 2010. 'Estimating enrichment of repetitive elements from high-throughput sequence data', *Genome Biol*, 11: R69.

158 Duszczyk, M. M., A. Wutz, V. Rybin, and M. Sattler. 2011. 'The Xist RNA A-repeat comprises a novel AUCG tetraloop fold and a platform for multimerization', *RNA*, 17: 1973-82.

159 Hall, L. L., and J. B. Lawrence. 2010. 'XIST RNA and architecture of the inactive X chromosome: implications for the repeat genome', *Cold Spring Harb Symp Quant Biol*, 75: 345-56.

160 Jeon, Y., and J. T. Lee. 2011. 'YY1 tethers Xist RNA to the inactive X nucleation center', *Cell*, 146: 119-33.

161 Sarma, K., P. Levasseur, A. Aristarkhov, and J. T. Lee. 2010. 'Locked nucleic acids (LNAs) reveal sequence requirements and kinetics of Xist RNA localization to the X chromosome', *Proc Natl Acad Sci U S A*, 107: 22196-201.

162 Wutz, A., T. P. Rasmussen, and R. Jaenisch. 2002. 'Chromosomal silencing and localization are mediated by different domains of Xist RNA', *Nat Genet*, 30: 167-74.

163 Costas, J., C. P. Vieira, F. Casares, and J. Vieira. 2003. 'Genomic characterization of a repetitive motif strongly associated with developmental genes in Drosophila', *BMC Genomics*, 4: 52.

Chapter 2: The discovery of the Firre lncRNA

2.1 Introduction

Multicellular organisms are composed of many different types of cells that are produced from one single cell, the zygote. The genetic programs that follow multiple rounds of cell divisions and finally establish various cell identities are still not well understood. The discovery of the core pluripotency factors, Oct4, Sox2, Klf4, Nanog, cMyc, opened up new avenues to study cellular transitions[1-3]; however, cell fate determination exceeds the network of a cocktail of protein factors[4-7]. Additional epigenetic considerations as well as the role of RNAs have been found to be critical for differentiation, lineage commitment, and cellular memory.

Adipogenesis, or the formation of fat cells, is one of the most studied and well defined differentiation systems and is governed by a known transcriptional cascade mainly driven by peroxisome proliferator-activated receptor γ (PPARγ) and CCAAT/enhancer-binding protein α (CEBPα)[8-11]. Together these two transcription factors drive the expression of many genes that are required for terminal differentiation into mature adipocytes[10,11]. In addition to the transcription factors that bind DNA directly, many co-factors have been found to play critical roles in activating gene expression. Co-factors can serve as molecular scaffolds to mediate the interaction between accessory proteins and the transcription machinery and modify the chromatin to assist in transcriptional activation[12]. Certain co-factors, such as CBP and p300, can change the chromatin conformation by enzymatically modifying the histone proteins and allow the chromatin to be more or less accessible to the transcription machinery[12-14]. In line with this finding, dramatic changes have been observed in the epigenetic landscape during adipogenesis, suggesting an important role for the epigenome in regulating this differentiation process[15].

The complex and precise patterns of expression of ncRNAs have been described in differentiation and development, including adipogenesis. Various miRNAs have been found to repress and up-regulate specific gene sets during adipogenesis[16]. Recently, many groups have identified lncRNAs as pivotal molecules for regulation of important developmental processes, including X chromosome inactivation[17], p53-mediated apoptosis[18,19], reprogramming of induced pluripotent stem cells[20], and cancer metastasis[21]. Furthermore, the specific spatiotemporal expression of lncRNAs across various stages of differentiation[22-24] might be ascribed to them being 'fine-tuners' of gene expression programs, thus establishment of cell fates. By being involved in positive and negative feedback loops, lncRNAs can help program molecular differences that control cell identity and lineage commitment[25-27]. Therefore, we hypothesized that lncRNAs participate in the regulatory network governing adipogenesis.

Here, we used deep RNA sequencing to identify mRNA and lncRNAs that are regulated during adipogenesis. In order to test, whether the lncRNAs functionally contribute to the differentiation process, we performed RNAi-mediated loss of function (LOF) experiments for 26 candidate lncRNAs. The scoring for each LOF assay was done using a novel method developed by two post-docs in the Rinn laboratory, which incorporated Jensen-Shannon distance metric to quantify the lncRNA-dependent gene expression changes across differentiation time points. With our screen, we identified lncRNAs with subtle and critical impacts (four lncRNAs) on adipogenesis and were the first to suggest that lncRNAs can comprise an as yet unexplored and important layer in adipogenic regulation.

2.2 Results

2.2.1 Global identification of lncRNAs regulated during adipogenesis

To identify the global transcriptome changes during adipogenesis, we sequenced polyA-selected RNAs from cultured brown and white pre-adiocytes, *in vitro* differentiated brown and white adipocytes, and primary brown and white adipocytes directly isolated from mice. Of the 214 million 36 bp reads, 77% was mapped to the mouse genome using TopHat[28], with gene annotations provided to maximize spliced alignment accuracy. The differential expression of all transcripts, corresponding to University of California at Santa Cruz (UCSC) and RIKEN clones, between preadipocytes and mature adipocytes was quantified using RNA-seq analysis program Cuffdiff[29]. We were able to identify 4,506 coding genes (2,390 up-regulated) and 481 lncRNAs (340 up-regulated) that were significantly regulated in one or both types of adipocyte. For up-regulation, we took genes with >2 fold change during differentiation. To further focus our target pool, we took the genes that were common to brown and white fat, which brought the numbers down to 1,734 coding genes and 175 lncRNAs that were up- or down-regulated at least 2 fold during differentiation (FDR <5%) (Figure 2.2.1.1).

As a validation for our technique and analysis, we examined the expression of known adipigenesis regulators: fatty acid binding protein 4 (Fabp4), adiponectin (AdipoQ), and glucose transporter type 4 (Glut4) (Figure 2.2.1.2), as well as several additional markers, such as preadipocyte factor 1 (Pref1), cell death-inducing DFFA-like effector a (Cidea), and uncoupling protein 1 (Ucp1). All of these master regulators were up-regulated as previously described. A few of the lncRNAs that show similar expression patterns with the master protein regulators are also shown in Figure 2.2.1.2. For simplicity, we refer to the lncRNAs involved in adipogenesis as *R*egulated in *A*di*p*ogenesis, lnc-RAPn. In addition, to further test the accuracy

of our data, we employed a Global Gene Ontology analysis, using the protein coding gene changes across the time courses of differentiation. Consistent with *bona fide* adipocytes and many previous studies, we observed significant enrichments in lipid metabolism and adipocyte terms and depletions in cell cycle and fibroblast terms (Figure 2.2.1.3), demonstrating that our RNA-seq data truly reflect the adipogenesis process.

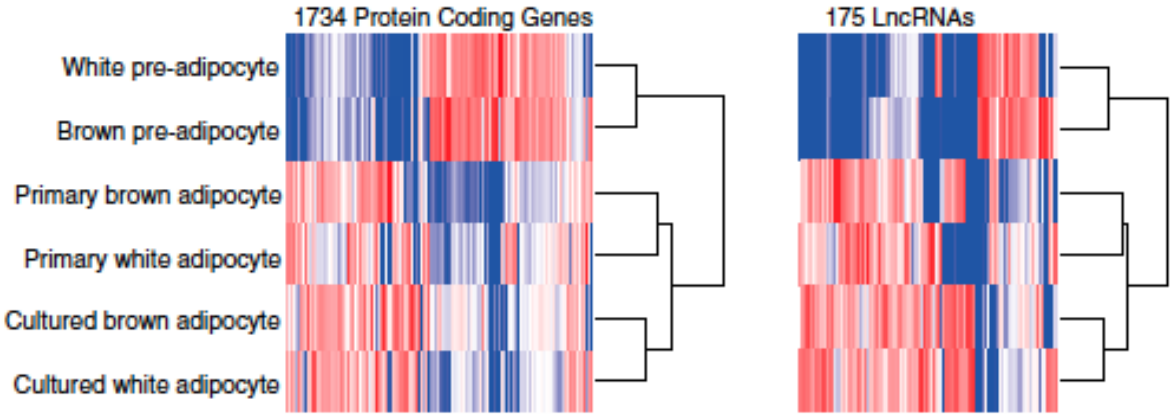


Figure 2.2.1.1: Independent hierarchical clustering of protein coding genes and lncRNAs during adipogenesis. Up-regulated shown in red and down-regulated in blue.

2.2.2 The expression of lncRNAs is tightly regulated during adipogenesis

Our analysis indicated that lncRNAs, similar to their coding counterparts, the mRNAs, distinguished precursors from mature adipocytes (Figure 2.2.1.1). The precursor cells cluster together and are different than the cultured and primary adipocytes, both of which share similar gene expression patterns, further suggesting that our *in vitro* culture system accurately reflects primary adipocytes.

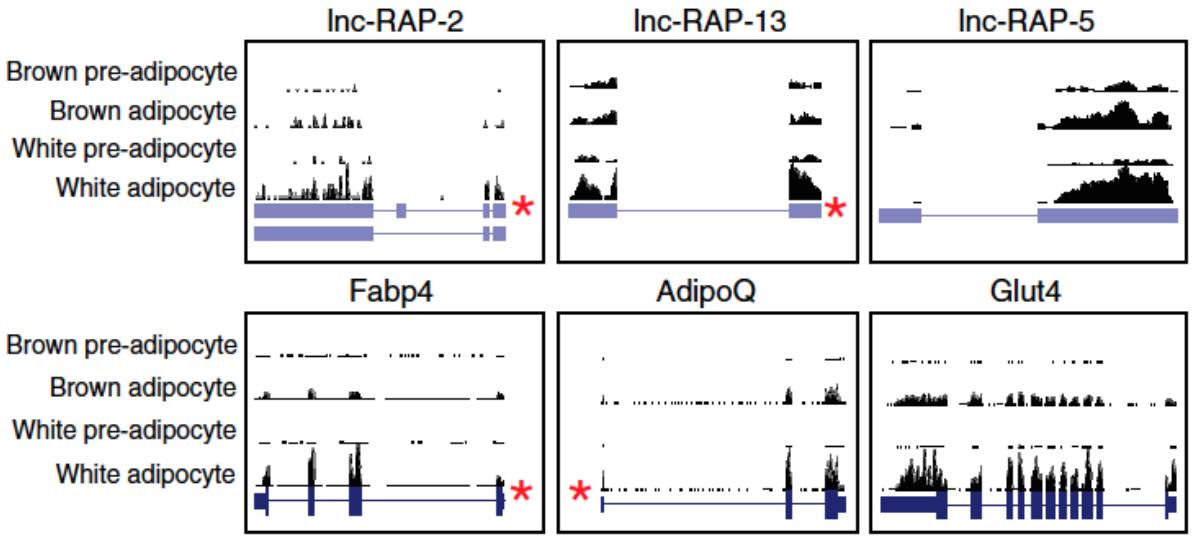


Figure 2.2.1.2: RNA-seq alignment and coverage of three lncRNAs and the master protein regulators of adiopogenesis in brown and white preadipocytes and mature adipocytes. Red asterisk indicating PPARγ ChIP binding site, blue bar indicating exons.

The reflection of the differentiation process by the expression patterns of the lncRNAs along with previous findings of key transcription factors (Oct4, Sox2, and p53) directly regulating ncRNAs impelled us to explore whether lncRNAs in our data are controlled by the same master factors, PPARγ and CEBPα[18,20,30-32]. Therefore, we explored the genome-wide binding sites of PPARγ[15] and CEBPα[33] by chromatin immunoprecipitation sequencing (ChIP-seq) using previously published data sets and observed that PPARγ binds within 2 kb upstream of the transcription start site of 23 (13%) out of 175 up-regulated lncRNAs, and CEBPα 34 (19%) of them. This observation is similar to those for mRNAs: PPARγ at 215 (14%) mRNA promoters and CEBPα at 352 (20%). Overall, our analysis suggests that lncRNAs and mRNAs are similarly bound and coordinated during adipogenesis.

Downregulated genes

GO Term	Ontology	#Hits in group	Group size	#Hits expected	p-value
lipid metabolic process	Biological process	34	1115	5	6.50025E-18
organic acid metabolic process	Biological process	27	673	4	1.06246E-16
lipid catabolic process	Biological process	16	139	1	2.23181E-16
regulation of lipid metabolic process	Biological process	17	190	1	9.81468E-16
monocarboxylic acid metabolic process	Biological process	22	432	2	1.1214E-15
cellular lipid metabolic process	Biological process	29	935	5	1.26534E-15
oxoacid metabolic process	Biological process	25	667	3	3.78043E-15
carboxylic acid metabolic process	Biological process	25	667	3	4.3205E-15
cellular ketone metabolic process	Biological process	25	678	4	4.93096E-15
fatty acid metabolic process	Biological process	19	341	2	2.59375E-14

BKL Expression Term	Location	#Hits in group	Group size	#Hits expected	p-value
adipocytes	Cell types	27	380	3	4.25124E-20
white adipocytes	Cell types	10	57	1	3.98909E-11
mesenchymally derived cells	Cell types	42	2867	19	8.41607E-09
brown adipocytes	Cell types	8	68	1	1.59265E-07
fibroblasts	Cell types	16	1084	8	0.00740825
fat	Organs, tissues, fluids	25	312	3	5.78837E-20
white fat	Organs, tissues, fluids	15	133	1	8.97881E-14
brown fat	Organs, tissues, fluids	9	110	1	3.87239E-07

Upregulated genes

GO Term	Ontology	#Hits in group	Group size	#Hits expected	p-value
cytokinesis	Biological process	8	89	1	1.19413E-06
M phase of mitotic cell cycle	Biological process	10	196	1	1.51029E-06
mitotic cell cycle	Biological process	14	461	3	1.67919E-06
DNA replication	Biological process	10	209	1	2.43058E-06
nuclear division	Biological process	9	182	1	6.69739E-06
mitosis	Biological process	9	181	1	7.02857E-06
organelle fission	Biological process	9	188	1	7.47131E-06
chromosome segregation	Biological process	7	90	1	7.50834E-06
cell division	Biological process	8	149	1	1.35029E-05
cell cycle	Biological process	19	1153	6	2.44111E-05

BKL Expression Term	Location	#Hits in g	Group size	#Hits expec	p-value
fibroblasts	Cell types	17	1084	7	0.00710516

Figure 2.2.1.3: Enriched Gene Ontology terms identified by the significantly regulated protein coding genes using RNA-seq.

To identify lncRNAs functionally contributing to adipocyte differentiation, we first ranked candidate lncRNAs according to their up-regulation in brown and white fat, and the presence of the PPARγ and CEBPα binding sites at the lncRNA promoter. These criteria resulted in 32 top targets, which we have independently analyzed by quantitative real time PCR (qRT-PCR), and used Fabp4 as control to monitor differentiation (Figure 2.2.2.1). Out of the 32 targets, 26 had a similar pattern to that of Fabp4.

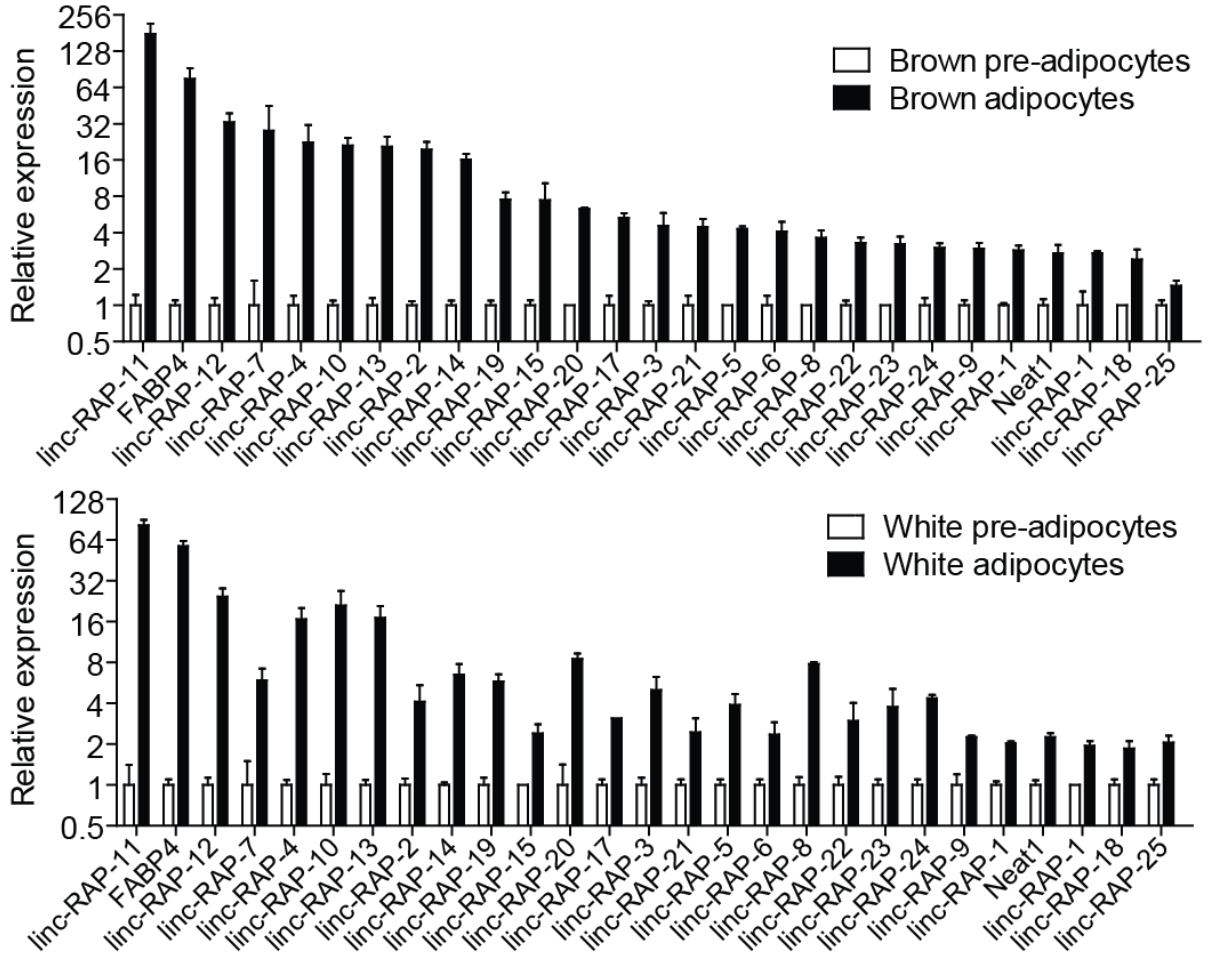


Figure 2.2.2.1: qRT-PCR validation of selected lncRNAs in primary brown preadipocytes and brown adipocytes. Preadipocytes are isolated (day 0), cultured, and differentiated (day 6) (n=3, $P<0.05$; means ± SEM).

2.2.3 LOF screening reveals functional lncRNAs during adipogenesis

Upon confirming the expression patterns of 26 lncRNAs, we wanted to test by RNAi-mediated LOF whether they directly functionally contribute to adipogenesis. We next screened 20 lncRNA genes identified by the above criteria of significant up-regulation in both brown and white fat cultures, PPARγ and CEBPα promoter binding, and independent validation of adipose-specific expression. To these criteria, we added one more, which was the down-regulation following TNF-α treatment to further establish a direct functional link (Figure 2.2.3.1). With this assay, we tested for lncRNAs that are down-regulated in a de-differentiation condition in a similar way to the adipocyte protein markers, which is a well established stimulation to induce insulin resistance[34,35].

After narrowing down our targets, we separately transfected three small interfering RNAs (siRNAs) targeting each lncRNA into subcutaneous preadipocyte cultures one day before differentiation. Two siRNAs targeting PPARγ were used as positive controls and non-targeting siRNAs as negative controls. Transfected adipogenic precursors were induced to differentiate. After four days of differentiation, lipid accumulation was evaluated via Oil Red O staining (ORO) staining. In addition to the ORO staining, lncRNA and mRNA levels were monitored using qRT-PCR. The knockdown of PPARγ resulted in a marked decrease in lipid accumulation, as expected. Ten of the targeted lncRNAs were not effectively depleted or did not result in functional outcome by ORO staining relative to the non-targeting controls. However, 10 of our targets exhibited moderate to strong reductions in lipid accumulation (Figure 2.2.3.2).

To further test the direct functional outcome of the lncRNA knockdown, we investigated the effects of the lncRNA LOF on key adipogenesis regulators: PPARγ, CEBPα,

Fabp4, and AdipoQ. All of the top 10 targets resulted in a significant decrease in at least 3 out of 4 adipogenic markers (Figure 2.2.3.3). Collectively, we were able to identify 10 lncRNAs out of 20 in our LOF screen that show important key regulatory roles in the proper differentiation of adipocyte precursors.

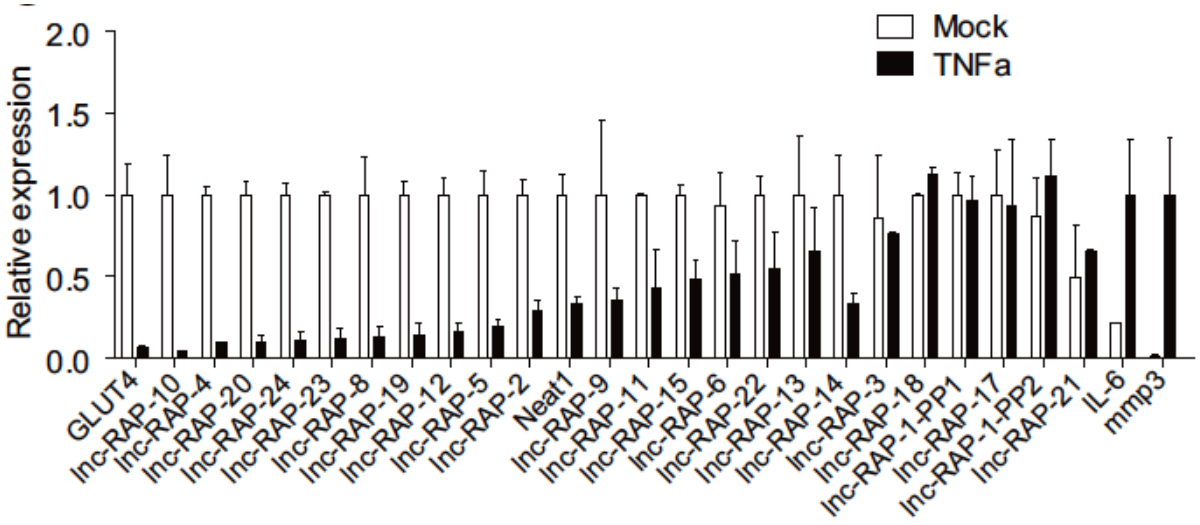


Figure 2.2.3.1: Expression of lncRNA candidates upon overnight TNFα treatment of mature adipocyte cultures. (n=3)

2.2.4 Information theoric metric scores cellular phenotypes

We next set out to analyze the phenotypes upon LOF of each lncRNA more systematically. We aimed to develop a scoring method which should accurately reflect how the expression profile from a knockdown differs from the precursor and adipocyte states and correspond to the ORO staining. Existing scoring functions, such as the Pearson correlation or the Euclidean distance have been used for similar purposes. However, neither of these methods accurately corresponded to the ORO staining results (Figure 2.2.4.1). This is likely due to some inherent limitations of these approaches. For example: (i) Pearson correlation is

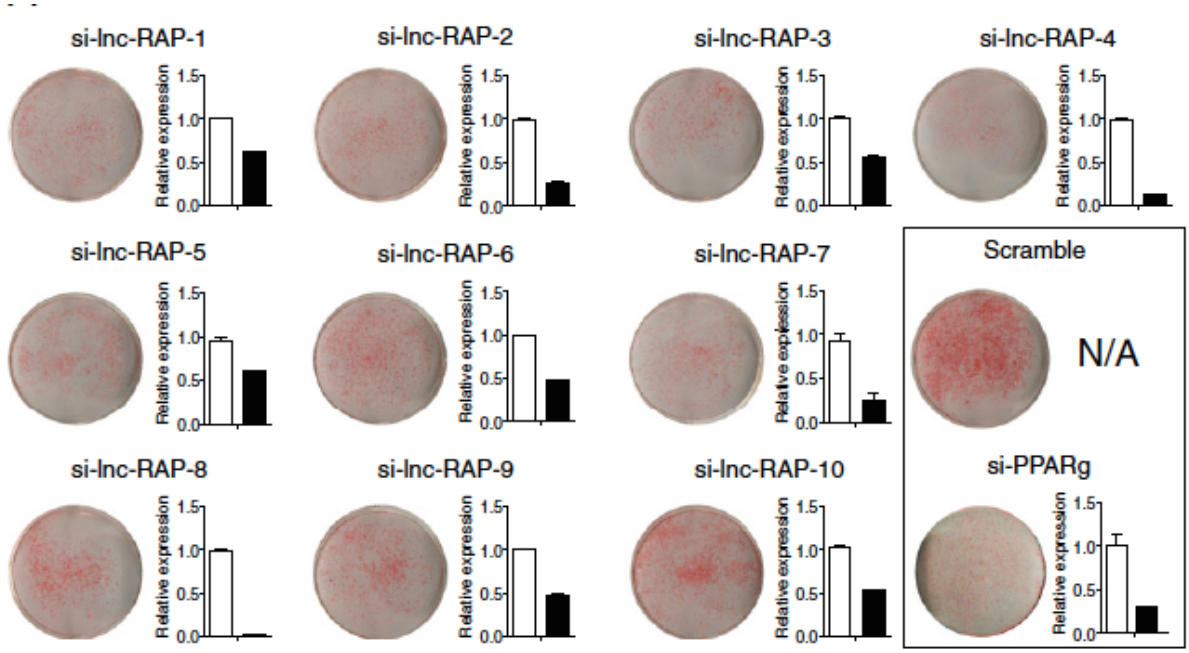


Figure 2.2.3.2: Oil Red O staining and qRT-PCR analysis following the knockdown of each lncRNA candidate. Knockdown is performed 1 day before differentiation of preadipocytes (n=3; $P< 0.05$; means ± SEM)

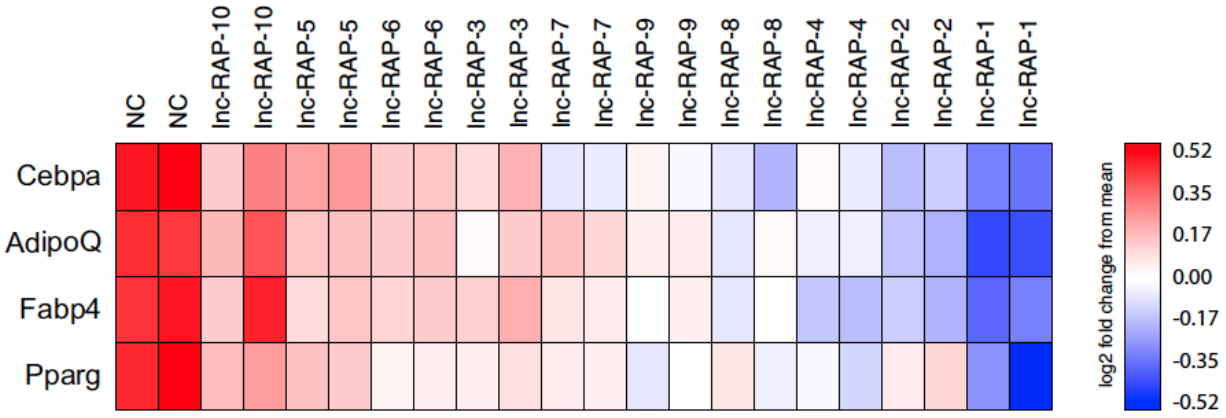


Figure 2.2.3.3: qRT-PCR analysis of key adipogenic regulators upon knockdown of each lncRNA from the top 10 candidates.

capable of identifying similar expression patterns across a range of different intensities; however, correlation is not a true distance metric and is sensitive to outliers, which can interfere with meaningful hierarchical clustering of samples; and (ii) Euclidean distance is greatly influenced by the absolute level of expression; differences in a few abundant genes may cause two profiles that are otherwise qualitatively very similar to appear distant.

To overcome these limitations, we turned to a metric of similarity between two frequency profiles based on Shannon Entropy, termed Jensen-Shannon distance (JSD). This metric has been used previously for quantifying differential splicing in high-throughput data and for machine learning applications[27]. To employ JSD, we first identified the gene signature that best distinguishes the precursor state (D0) from mature cultured white adipocytes (D4). Using the Affymetrix mouse 430a2 platform, we analyzed total RNA from three biological replicates from D0 and D4. We found 2,200 genes that were significantly differentially expressed between precursors and mature adipocytes (SAM, FDR<4.5%). We took a cautious approach to remove the genes that change due to siRNA treatment (non-targeting control). The resulting gene expression pattern comprised 1,727 genes that clearly distinguished precursors from mature adipocytes.

To monitor lncRNA-dependent perturbations to the adipogenic signature, subcutaneous preadipocytes were transfected with siRNAs targeting each lncRNA one day before differentiation. Total RNA was extracted after 4 days of differentiation. We hypothesized that if a lncRNA were required for proper adipose differentiation, then we would observe minimal differences between the expression profiles of the 1,727 genes in the lncRNA depletion samples and the profile of the same genes in the undifferentiated (D0) preadipocyte control samples. Our goal is to use this gene signature to score the similarity of each lncRNA depletion

to the precursor state.

We, therefore, quantified the shift in the transcriptome towards the mean D0 expression pattern across 1,727 significant genes for each of the 10 lncRNA depletion profile and scrambled controls. The mean-centered expression profiles are ordered by JSD to the D0 profile. Concordant with the quality of this new metric that we developed, the scramble control most closely resembles that of D4 differentiated adipocytes (Figure 2.2.4.2). We also observed interesting patterns, in which one lncRNA, RAP-10, had a differentiation profile similar to the scramble control and the D4 adipocytes, suggesting no adipogenic function for this lncRNA (Figure 2.2.4.2).

For the rest of our lncRNA candidates, we in deed observed a partial or near complete reversion of the mature adipocyte (D4) to precursor (D0) expression signature (Figure 2.2.4.2). This indicates that we have found lncRNAs that are required for the proper regulation of the transcriptional network in adipogenesis. JSD scores accurately represent the observed levels of lipid accumulation for each lncRNA knockdown as measured by ORO staining, and thus serve as a great tool for quantifying phenotypic differences in other cell systems and biological conditions.

2.2.5 lncRNA LOF specifically perturbs adipogenic pathways

To further examine the gene pathways that are perturbed upon the knockdown of each lncRNA, we conducted a Gene Set Enrichment Analysis (GSEA). Ran-ordered lists were generated for all genes comparing each lncRNA knockdown to the scramble control. These lists were used as input to a preranked GSEA. For each lncRNA knockdown vs. scramble control, normalized enrichment scores and significance values were determined across the C2, a curated gene set collection from MsigDB. To specifically investigate the perturbations to the adipose-

associated pathways, we separated all the other gene sets from the adipogenesis one (Figure 2.2.5.1) and found that the mean of the distribution of *P* values for the adipose-associated gene sets was lower than for the nonadipose gene sets (Figure 2.2.5.2). This enrichment for reduction in adipose-associated genes at either tail of the rank-ordered lists correlates with the JSD analysis and ORO staining. Collectively, our results indicate that adipogenesis differentiation pathway is tightly controlled by our top 9 lncRNA candidates.

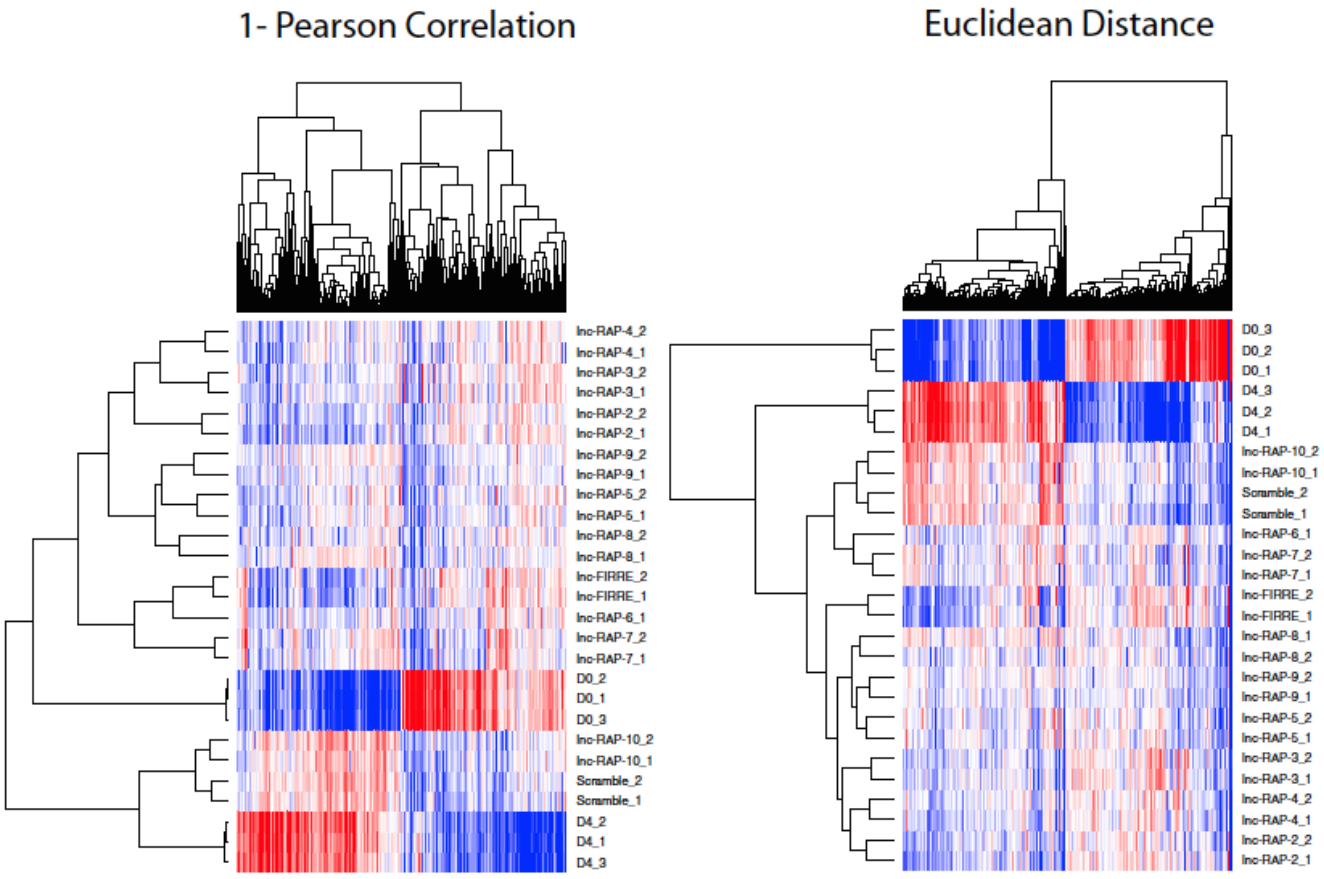


Figure 2.2.4.1: Comparison of common hierarchical clustering metrics, Pearson Correlation and Euclidean Distance. Used in determining relationships between genes and conditions in our gene expression studies.

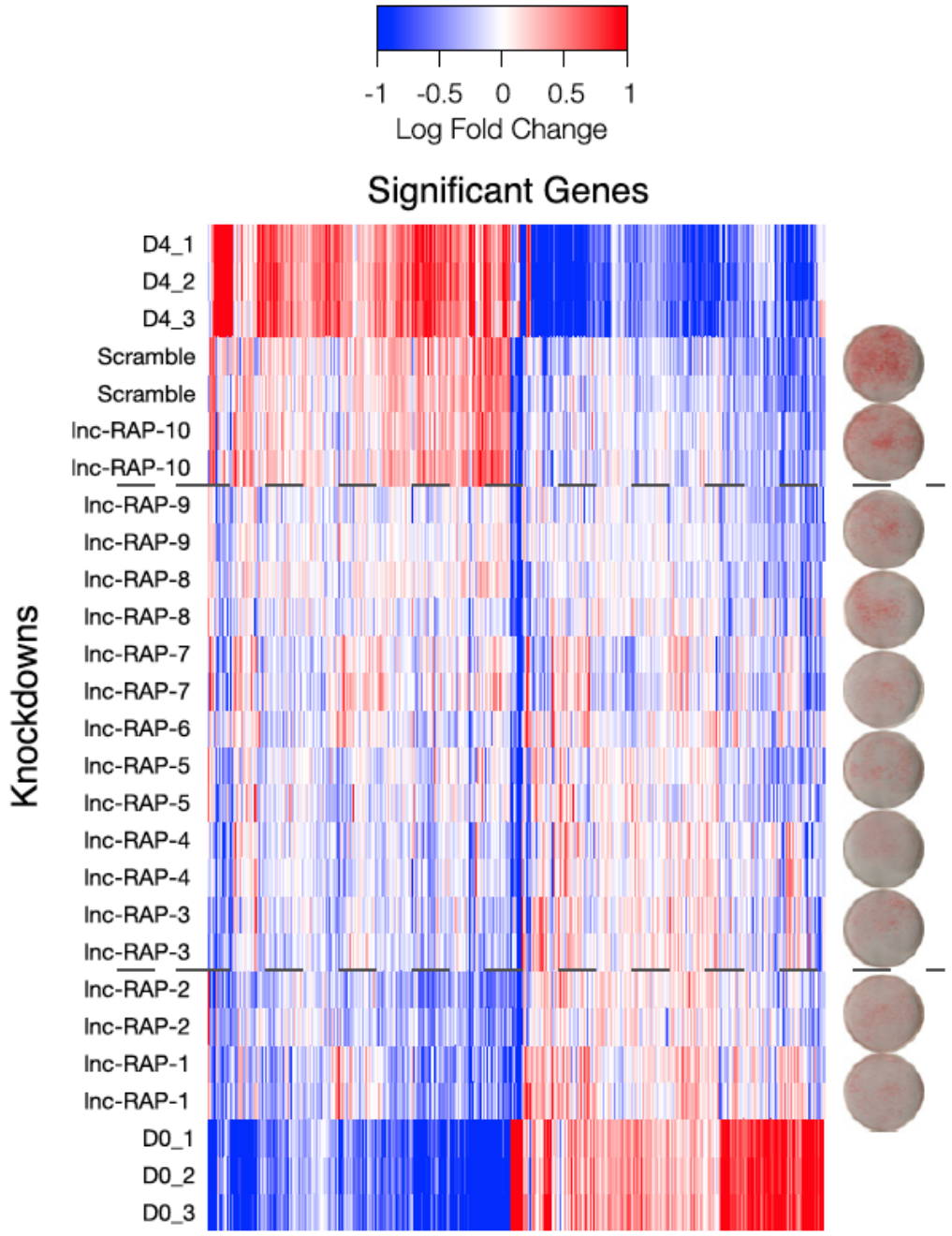


Figure 2.2.4.2: Jensen-Shannon distance ranking of expression profile of 1,727 genes. These genes are determinant between precursor (D0) and differentiated adipocytes (D4), upon knockdown of adipogenesis-regulated lncRNAs.

2.2.6 Two functional lncRNAs may encode small peptides

Although the 10 genes we discuss above were previously annotated as noncoding RNAs in the UCSC genome annotation, we scrutinized them for evidence of coding capacity. We first searched the amino acid databases SWISSPROT, PDB, and the RefSeq protein collections (currently comprised of over 12 million amino acid sequences) for homology to all possible translations of the annotated RNAs. This analysis revealed no significant coding potential for any of the lncRNAs we knocked down. Next, we calculated the codon substitution frequency (CSF) score for all open reading frames longer than 30 amino acids in these lncRNAs. Briefly, CSF utilizes multiple DNA alignments from 29 vertebrates genomes[37] to find regions in an RNA that have a substitution signature consistent with pressure to conserve putative codons more than the region as a whole. For example, an ORF that displays higher mutation frequency at the wobble base than at other bases in each in-phase triplet indicates coding potential, thus would have a high CSF score.

Our analysis identified two lncRNAs, lnc-RAP5 and lnc-RAP2, which have moderate CSF scores indicating coding potential of a small ORF (56 and 36 amino acids, respectively). Notably, the small ORF in lnc-RAP5 is strongly conserved across metazoans from human to African clawed frog, yet does not resemble any of the over 200,000 known amino acid sequences across the 4 kingdoms of life. However, the remaining 8 lnc-RAP genes display no detectable codon-level conservation signature, reinforcing their annotation as noncoding RNAs.

2.2.7 Orthology mapping of functional lncRNAs

We next investigated if lncRNA-RAPs have orthologous transcripts in the human genome. Using TransMap, a database of syntenic, homologous transcripts available through

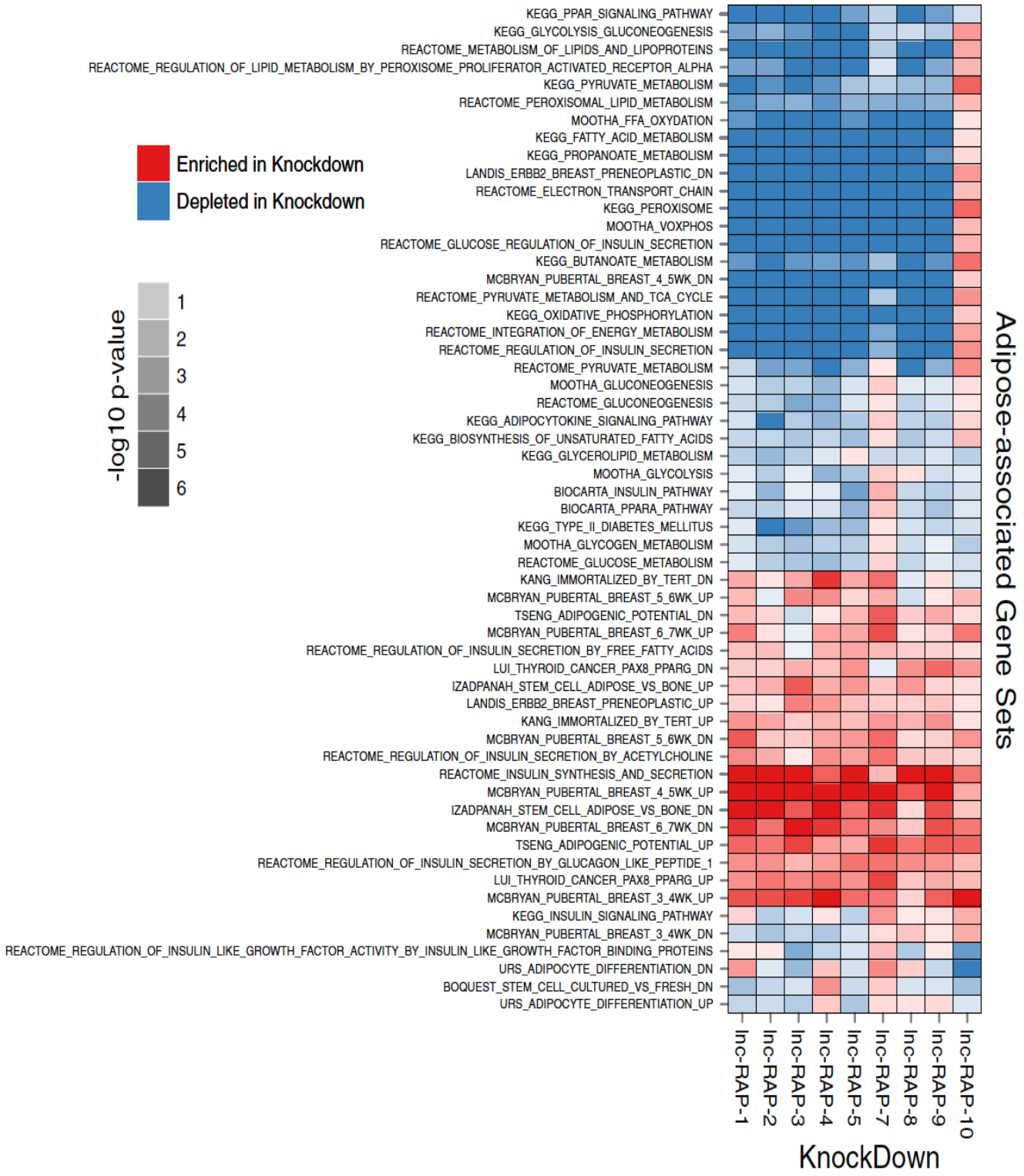


Figure 2.2.5.1: Adipose-associated gene sets from MsigDB C2 analyzed in lncRNA knockdown conditions.

the UCSC genome browser, we identified clear human orthologs for several lncRNAs used in our knockdown screen. Of the 10 mouse genes we knocked down, four have strong homology to human expressed sequence tags (EST). However, because EST coverage tends to be incomplete for lncRNAs in both the mouse and human genomes (presumably due to their low abundance), the true proportion of murine lncRNAs with human orthologs may be higher. Two of the mouse genes with human orthologs, lnc-RAP5 and lnc-RAP2, may encode small peptides, as discussed above. The other two, lnc-RAP1 and lnc-RAP8, display no evidence of coding function and are annotated as noncoding RNAs in the human genome.

We hypothesized that if the human orthologs of these 4 murine lnc-RAPs played a role in adipogenesis, they would similarly be induced upon differentiation of human adipocyte precursors to mature adipocytes. To this end, we examined the expression of these 4 genes in human primary preadipocytes and differentiated adipocytes.

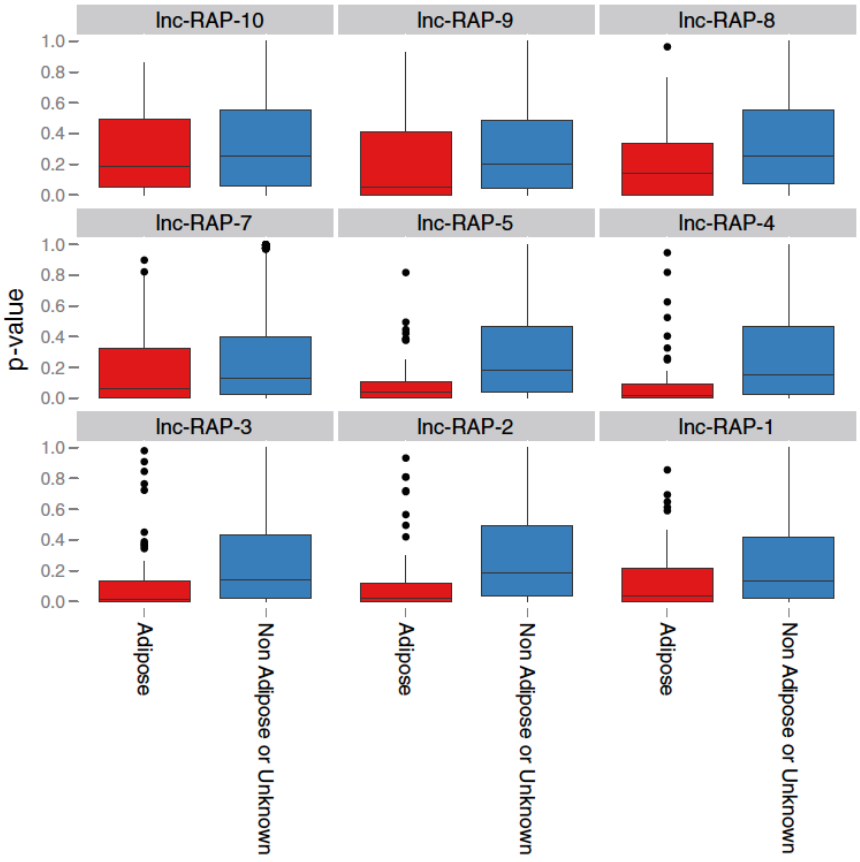


Figure 2.2.5.2: All MsigDB curated gene sets divided into "adipose" and "nonadipose or unknown." Gene sets are analyzed in lncRNA knockdown conditions.

As shown in Figure 2.2.7.1, all 4 genes were significantly up-regulated during human adipogenesis suggestive of a conserved function. It should be noted that lncRNA-RAP1, which demonstrates a strong adipogenic phenotype in mouse, exhibits striking structural similarity to the human noncoding RNA LOC286467 (Figure 2.2.7.1).

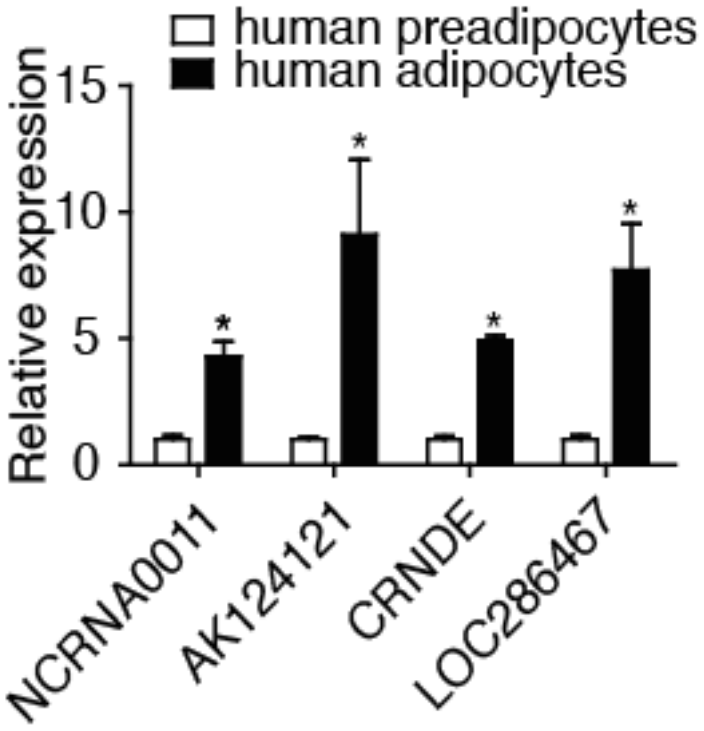


Figure 2.2.7.1: The expression profile of human orthologs of murine adipogenic lncRNA-RAPs. lncRNA RAPs 5, 2, 8, 1 are shown during in vitro differentiation of human adipocytes.

2.2.8 An orthologous RNA sequence domain

Interestingly, lncRNA-RAP1, which upon LOF almost completely blocks adipocyte differentiation, contains a primary RNA sequence domain of 163 bp that is repeated throughout a large portion of both the human and mouse orthologs (Figure 2.2.8.1). The mouse genome contains 12 instances of this domain, which we refer to as Repeated RNA Domain, RRD, while the human genome contains 7. Strikingly, most instances of RRD occur as nearly complete

exons in both the mouse and human version of the gene yet are not found anywhere else in either genome. We also note that alternative splice isoforms of lnc-RAP1 exist in both organisms. Repeated primary sequence domains have been found in other lncRNAs, notably Xist. Xist contains a repeated structural domain called RepA that is responsible for initial recruitment of PRC2 to the X chromosome during inactivation. RRD is reminiscent of RepA in both size, number of copies, and the tandem layout of its copies within its noncoding RNA.

We further searched for orthologs of lnc-RAP1 in the cow, dog, and elephant genomes, and while these species lack sufficient EST evidence to perform a TransMap-based analysis, their genomes do contain regions homologous to the lnc-RAP1 locus, and the cow genome contains three instances of a genomic repeat with substantial homology to RRD. Next, we performed multiple alignment analyses of the 7 human and 12 mouse instances of RRD (Figure 2.2.8.2). This analysis revealed that intra-species RRDs were more similar than inter-species RRDs, suggesting that these repeats may be derived from a gene-conversion-like event. The pairwise similarity between any two mouse instances of RRD is very high (~90% identity), and most human RRD copies are also highly similar to one another. This may simply reflect a lack of complete annotation in this locus but it is also suggestive of a mechanism that has expanded (or disrupted) RRDs in lnc-RAP1 differently since mouse and human share a common ancestor. To continue the analogy of RepA to Xist, which forms repeated secondary structures required for recruitment of polycomb proteins, RRD may be required for lnc-RAP1 function, and that function may depend on the number and layout of RRDs in the mature transcript.

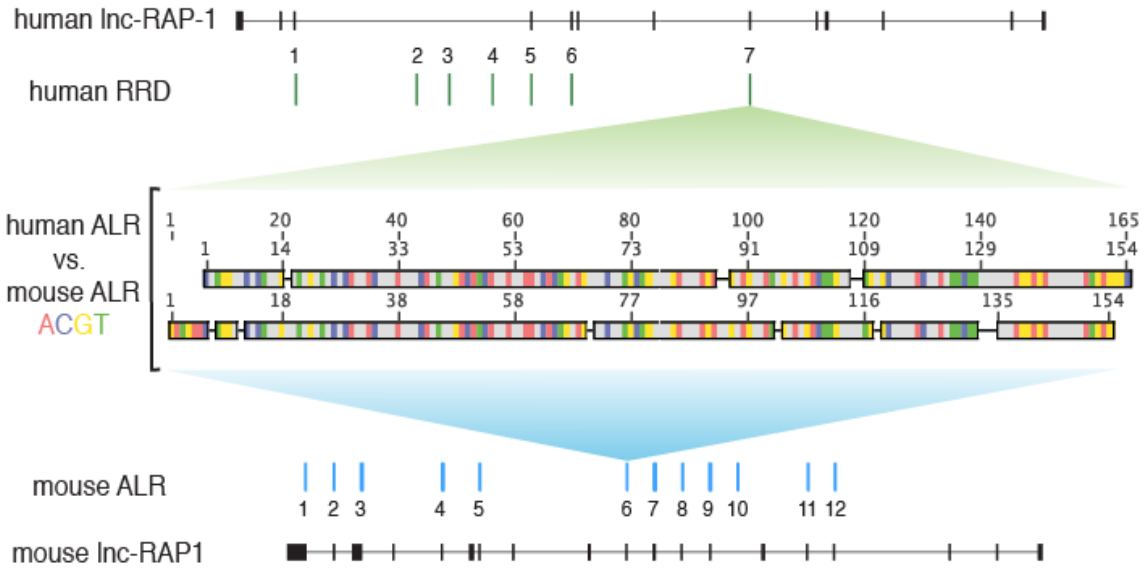


Figure 2.2.8.1: The unique repeat unit RRD that exists in both mouse and human lnc-RAP1 loci.

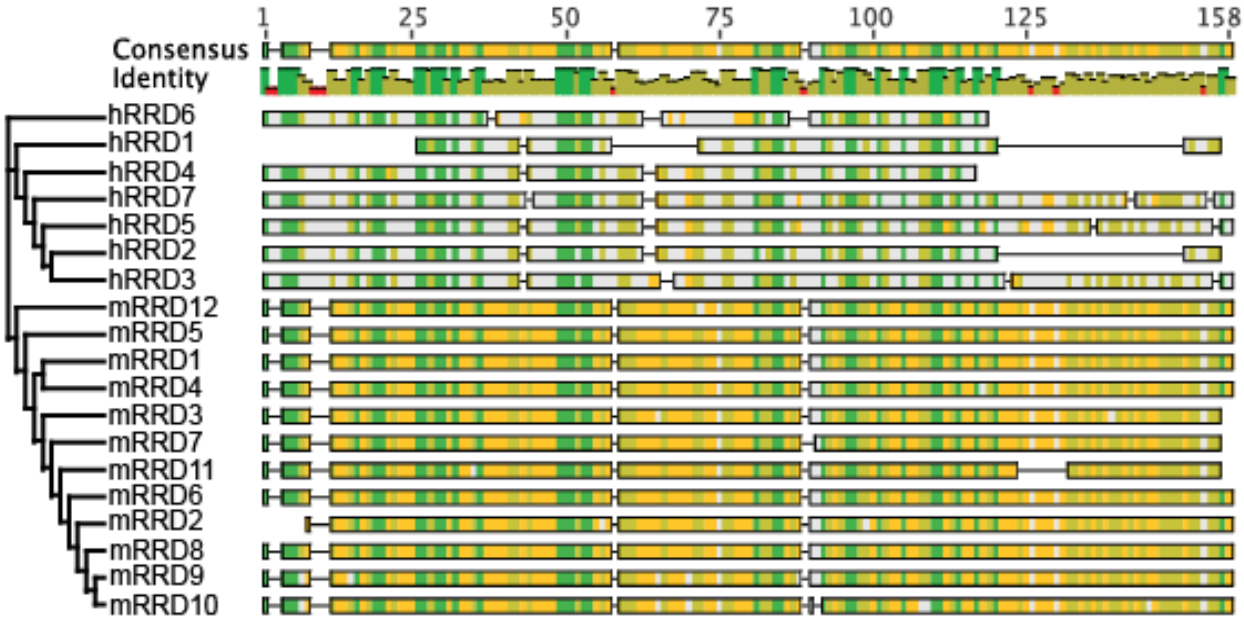


Figure 2.2.8.2: Multiple alignment of murine and human RRD instances.

This work was published: Sun, L.*, L. A. Goff*, C. Trapnell*, R. Alexander, K. A. Lo, **E. Hacisuleyman**, M. Sauvageau, B. Tazon-Vega, D. R. Kelley, D. G. Hendrickson, B. Yuan, M. Kellis, H. F. Lodish, and J. L. Rinn. 2013. 'Long noncoding RNAs regulate adipogenesis', *Proc Natl Acad Sci U S A*, 110: 3387-92. *Authors contributed equally to this work.

2.3 Discussion

Thousands of lncRNAs in both mouse and human have been discovered[30,32,38,39] but the functions of a vast majority of them remain elusive. Only a handful lncRNAs have been tested through loss of function experiments to determine whether they are required to establish the biological context in which they appear. The number of lncRNAs that are truly important remains unknown.

Our study has identified hundreds of lncRNAs that are regulated on multiple levels during adipogenic differentiation, a substantial fraction of which is regulated by the same core transcription factors. Of the lncRNAs that we tested by multiple criteria, most are specific to the adipose tissue and show important developmental phentotypes in adipogenesis. Our findings suggest the lncRNAs are regulated in a similar manner to mRNAs in adipogenesis and other developmental processes. Interestingly, depletion of several lncRNAs resulted in dramatic phenotypes by globally inhibiting adipogenic gene expression programs, hinting at the key roles these lncRNAs might play in inducing adipocyte-specific genes. However, it should be noted that RNAi-mediated studies are prone to off-target effects; thus, it needs to be further investigated to confirm that the phenotypes result from the loss of the particular lncRNA. To that end, these studies should be supported by gain of function experiments.

Large scale loss-of-function and gain-of-function (GOF) screens of lncRNAs will offer a wealth of useful information about the importance of these genes for development. We have presented an application of the Jensen-Shannon distance metric on gene expression profiles to quantify the phenotypic contributions of lncRNAs to adipogenesis. JSD circumvents shortcomings inherent to other metrics, and we expect it to be widely applicable to future LOF and GOF screens. Such a metric is particularly important for screens where *in vitro* functional

assays (e.g. the Oil Red Staining used here) are unavailable or excessively laborious, expensive, or unreliable. This metric on profiles can help triage screened lncRNAs for more detailed mechanistic follow-up.

Further studies will be required to decipher the molecular mechanism, by which the lncRNAs discussed in this study act to regulate adipogenesis. It is likely that some lncRNAs can serve as a modular scaffold and tether protein factors to form a chromatin modifying complex to regulate the epigenetic architectures during adipogenesis, as has been proposed[31,39-43]. For example, the Xist lncRNA required for X chromosome inactivation contains a RNA sequence domain, termed repeat A (RepA), that is repeated in tandem numerous times in the 5' exon but is not present elsewhere in the genome. The RepA adopts a specific secondary structure that is required for its physical association with and recruitment of chromatin modifying complexes to the inactive the X chromosome and is conserved in both human and mouse. Similarly, we observed an intriguing sequence domain, RRD, in lnc-RAP1 that is reminiscent of RepA and is also conserved between human and mouse. Future experiments will focus on the structural and functional features of this sequence domain, and how it affects the role of lnc-RAP1.

The ensemble of noncoding RNAs presented here should provide insights regarding the mechanisms by which lncRNAs regulate adipogenesis and possibly other developmental processes. By performing detailed biochemical studies on the whole collection, it should be possible to assess whether these genes all serve as scaffolds for protein complexes or by more diverse mechanisms. Collectively, our approach can be universally applied to any cell-based differentiation system to quickly screen candidate lncRNAs and score the gene-expression phenotype to unravel the regulatory circuits influenced by lncRNAs. Further investigation of

their roles involved in obesity may lead to identification of novel therapeutic targets and strategies against obesity and related metabolic disorders.

The lnc-RAP1 identified in this study was termed Functional Intergenic RNA Repeat Element, Firre, which will be further explored in the following chapter.

2.4 Materials and Methods

2.4.1 Cell isolation and tissue culture

Primary adipocytes are isolated according to published methods with few modifications[44,45]. Interscapular brown adipose tissues, epididymal fat pads and subcutaneous fat pads are harvested from five 8-week old male mice. Fat tissues are minced, digested in collagenase, and fractioned by centrifugation. Adipocytes are collected from the top layer. Brown preadipocytes are isolated from interscapular brown fat and white preadipocytes are isolated from subcutaneous fat from young mice (2 week old). After collagenase digestion and fractionation, preadipocytes, enriched in bottom stromal vascular fraction (VSC), are resuspended and cultured to confluence in DMEM supplemented with 10% New-born Bovine Serum. The cells are then exposed to differentiation medium: 10% FBS DMEM, Insulin 850 nM (Sigma), Dexamethasone 0.5 μM (Sigma), IBMX 250 μM (Sigma), Rosiglitazone 1μM (Cayman Chemical) (brown adipocyte cultures, T3 1nM and Indomethacin 125 nM were also added into the medium). After 2 days, cells are incubated in culture medium containing insulin 160 nM (supplemented with T3 for brown adipocyte cultures) for another 2 days, and then are switched to 10% FBS DMEM. Human preadipocytes were purchased from (Lonza PT-5020) and cultured according to manufacturer's instruction. Cells were exposed to PGM-2 (PT-9502 & PT-8202) media plus Rosiglitazone 1μM to induce differentiation for 2 weeks.

2.4.2 RNA extraction

RNAs of brown fat and white fat are extracted using Qiagen Kit according to manufacturer's Instructions. RNAs of other tissues are purchased from Ambion (AM7800).

2.4.3 TNFa treatment

Primary white preadipocytes are cultured and differentiated as described above. 6 days after induction of differentiation, 1 nmol/l human TNF-α (PeproTech) is added to the growth medium, and cell cultures are incubated with 1nM TNF-a (PeproTech) for 24 h. *2.4.4 Library preparation and sequencing* Total RNAs are extracted using Qiagen kit and 10 μg of total RNAs for each sample is used to prepare mRNA-seq library according to manufacturer's instruction (Illumina). cDNA libraries were prepared and sequenced by Illumina GAII according to the manufacturer's instructions.

2.4.5 Differential expression analysis of RNA-Seq

Reads from each sample were mapped against the mouse genome (mm9 build) using TopHat (version 1.1.0), using options "--no-novel-juncs -a 5 -F 0.0". A splice junction index derived from the combined UCSC and RefSeq mm9 annotations together with previously discovered lncRNA transcript models[46] built using RNA-Seq from several cell lines. This set of annotated transcripts was quantified in each sample using Cuffdiff[29] (version 0.9.3), which estimates transcript and gene expression in each condition using a generative statistical model of RNA-seq. Cuffdiff calculates the abundances in each condition for all transcripts that maximize the likelihood of observing the reads in the experiment under this model. Cuffdiff attaches statistical significance to observed changes to gene expression, and we restricted analysis to genes significantly differentially expressed by least two-fold between pre-adipocytes and adipocytes. We also required that differentially expressed genes used in downstream analysis were supported by at least 10 reads in either condition.

2.4.6 Promoter analysis

ChIP-Seq peak calls made for Pparg[15] and Cebpa[33] were compared against the transcript catalog used above by defining a 2 Kb window upstream of each annotated TSS and intersecting these regions using the windowBed program from BEDTools (version 2.0.12). The following arguments were provided to windowBed: "-sw -l 2000 -r 0". The significance of enrichment among up-regulated genes with peaks for these factors was calculated by Monte Carlo sampling: 1000 sets of randomly selected genes were selected from all genes in the catalog to estimate the empirical distribution of enrichment among gene sets as large as the set being tested (e.g. lncRNAs up-regulated during adipogenesis. This distribution was used to derive an upper bound on the statistical significance of the observed enrichment in the set being tested.

2.4.7 Quantitative real time PCR (qRT-PCR)

For qRT-PCR, 200 ng total RNA is reverse-transcribed using random primers and SuperScript II Reverse Transcriptase (Invitrogen), and cDNA is amplified with gene specific primers and SYBR Green PCR master mix using ABI 7900HT (Applied Biosystems). Primer sequences are listed in Appendix 1. 18S and b-actin were used as internal controls for mouse samples and human samples, respectively. Data are analyzed by the relative quantification ($\Delta\Delta C_t$) method and expressed as means ± SEM. Student's t test (unpaired, two-tailed) was used to compare two groups. $P < 0.05$ was considered as statistically significant. The expression of ncRNAs across different mouse tissues is plotted as heat map using Cluster 3.0 and Treeview[47].

2.4.8 Oil-Red O staining

To prepare Oil-Red O solution, 0.5g Oil Red O (Sigma, catalog # 0-0625) was

dissolved in 100 ml isopropanol, mixed with H_2O at a ratio 6:4, and filtered with Whatman #1 filter paper. Primary preadipocytes were differentiated in 6 well plates. Cells were washed with PBS twice and fixed with formalin for 15 minutes in room temperature. After formalin fix, Cells were washed with PBS and stained with freshly prepared Oil-Red O solution for 1 hour at room temperature. Cells were washed with H_2O to remove the residual Oil-Red O. *2.4.9 Knockdown of adipogenically regulated lncRNAs* When cultured primary preadipocytes reach 70-80% confluence, siRNAs (200 nM) were transfected by DharmaFect 2 (final 6 μl/ml) according to the manufacturer's instruction (Dharmacon). 24 hrs after transfection, cells were recovered in full culture media and grown to confluence for differentiation. siRNAs were purchased from GenePharma. siRNA sequences are provided in Appendix 2.

2.4.10 Expression microarray profiling

Three replicate RNA samples each from D0 undifferentiated preadipocytes, D4 PPARγ-induced adipocytes, and D4 scramble siRNA controls were labeled and hybridized to Affymetrix Mouse 430a2 arrays. Additionally, two replicate RNA samples from each of the 10 lncRNA siRNA studies were labeled and hybridized as well using standard Affymetrix protocols. Data were collected and analyzed using the 'affy' package in R/Bioconductor. Briefly, probes were background corrected using the 'mas' algorithm. Probe values were quantile normalized, and probe sets were summarized using the average difference of perfect matches only. All significance tests were performed using SAM[48] with Benjamini-Hochberg MTC. Significant gene lists were selected with a delta that constrained the FDR < 4.5%. Significant gene expression data were biclustered using the Jensen-Shannon Distance, which is derived from the Shannon entropy

$$H(p) = -\sum_{i=1}^{n} p_i \ln(p_i)$$

Where p is a discrete probability distribution. The Jensen-Shannon divergence is

$$JSD(p,q) = \sqrt{H\left(\frac{p+q}{2}\right) - \frac{1}{2}\left(H(p) + H(q)\right)}$$

where p and q are two discrete probability distributions. In our array analysis, these distributions each represent either the relative abundance of all genes in a condition (for condition-level clustering) or the density of a gene's expression across all conditions (for gene-level clustering). Pairwise condition distances thus reflect an information-theoretic summary of how similar the program of gene expression in two cell states (e.g. preadipocyte and adipocyte). Pairwise gene distances reflect the information-theoretic similarity of expression across all conditions (potentially implying coordinated regulation).

Since only two replicates were available for each knockdown vs. scramble control analysis, rank-ordered list of all genes were generated and used as input for the non-parametric preranked GSEA analysis[49]. Rank-ordered gene lists were compared to all 'curated gene sets' (C2) in MSigDB with the following parameters (collapse=True, norm=meandiv, scoring_scheme=weighted, mode=Max_probe, set_min=12, set_max=500, nperm=1000, chip=Mouse430A_2.chip).

2.4.11 Repetitive sequence analysis

The sequence for adipogenic lncRNAs was scanned for repetitive elements using the

ab initio repeat detection algorithm RepeatScout[50]. This sequence was then aligned back to the genome (mm9 or hg19) using BLAT[51] with the following parameters: "-stepSize=5 -repMatch=2253 -minScore=50 –minIdentity=0". Genomic DNA for the hits was extracted and multiply aligned with the Fast Statistical Aligner (FSA), a probabilistic multiple alignment tool specifically engineered to accommodate multiple alignment of sequences with potentially non-uniform evolutionary constraint[52]. FSA uses pair hidden Markov models to estimate gap and substitution parameters for the alignment multiple alignment scoring function, improving alignment robustness.

2.5 References

1 Takahashi, K., and S. Yamanaka. 2006. 'Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors', *Cell*, 126: 663-76.

2 Smith, A. G., J. K. Heath, D. D. Donaldson, G. G. Wong, J. Moreau, M. Stahl, and D. Rogers. 1988. 'Inhibition of pluripotential embryonic stem cell differentiation by purified polypeptides', *Nature*, 336: 688-90.

3 Takahashi, K., K. Tanabe, M. Ohnuki, M. Narita, T. Ichisaka, K. Tomoda, and S. Yamanaka. 2007. 'Induction of pluripotent stem cells from adult human fibroblasts by defined factors', *Cell*, 131: 861-72.

4 Kondo, T., M. Asai, K. Tsukita, Y. Kutoku, Y. Ohsawa, Y. Sunada, K. Imamura, N. Egawa, N. Yahata, K. Okita, K. Takahashi, I. Asaka, T. Aoi, A. Watanabe, K. Watanabe, C. Kadoya, R. Nakano, D. Watanabe, K. Maruyama, O. Hori, S. Hibino, T. Choshi, T. Nakahata, H. Hioki, T. Kaneko, M. Naitoh, K. Yoshikawa, S. Yamawaki, S. Suzuki, R. Hata, S. Ueno, T. Seki, K. Kobayashi, T. Toda, K. Murakami, K. Irie, W. L. Klein, H. Mori, T. Asada, R. Takahashi, N. Iwata, S. Yamanaka, and H. Inoue. 2013. 'Modeling Alzheimer's disease with iPSCs reveals stress phenotypes associated with intracellular Abeta and differential drug responsiveness', *Cell Stem Cell*, 12: 487-96.

5 Orlando, V. 2003. 'Polycomb, epigenomes, and control of cell identity', *Cell*, 112: 599-606.

6 Rijnkels, M., E. Kabotyanski, M. B. Montazer-Torbati, C. Hue Beauvais, Y. Vassetzky, J. M. Rosen, and E. Devinoy. 2010. 'The epigenetic landscape of mammary gland development and functional differentiation', *J Mammary Gland Biol Neoplasia*, 15: 85-100.

7 Mercer, T. R., I. A. Qureshi, S. Gokhan, M. E. Dinger, G. Li, J. S. Mattick, and M. F. Mehler. 2010. 'Long noncoding RNAs in neuronal-glial fate specification and oligodendrocyte lineage maturation', *BMC Neurosci*, 11: 14.

8 Rosen, E. D., and O. A. MacDougald. 2006. 'Adipocyte differentiation from the inside out', *Nat Rev Mol Cell Biol*, 7: 885-96.

9 Rosen, E. D., C. J. Walkey, P. Puigserver, and B. M. Spiegelman. 2000. 'Transcriptional regulation of adipogenesis', *Genes Dev*, 14: 1293-307.

10 Tontonoz, P., R. A. Graves, A. I. Budavari, H. Erdjument-Bromage, M. Lui, E. Hu, P. Tempst, and B. M. Spiegelman. 1994. 'Adipocyte-specific transcription factor ARF6 is a heterodimeric complex of two nuclear hormone receptors, PPAR gamma and RXR alpha', *Nucleic Acids Res*, 22: 5628-34.

11 Lefterova, M. I., Y. Zhang, D. J. Steger, M. Schupp, J. Schug, A. Cristancho, D. Feng, D. Zhuo, C. J. Stoeckert, Jr., X. S. Liu, and M. A. Lazar. 2008. 'PPARgamma and C/EBP factors orchestrate adipocyte biology via adjacent binding on a genome-wide scale', *Genes Dev*, 22: 2941-52.

12 Spiegelman, B. M., and R. Heinrich. 2004. 'Biological control through regulated transcriptional coactivators', *Cell*, 119: 157-67.

13 Musri, M. M., R. Gomis, and M. Parrizas. 2007. 'Chromatin and chromatin-modifying proteins in adipogenesis', *Biochem Cell Biol*, 85: 397-410.

14 Musri, M. M., R. Gomis, and M. Parrizas. 2010. 'A chromatin perspective of adipogenesis', *Organogenesis*, 6: 15-23.

15 Mikkelsen, T. S., Z. Xu, X. Zhang, L. Wang, J. M. Gimble, E. S. Lander, and E. D. Rosen. 2010. 'Comparative epigenomic analysis of murine and human adipogenesis', *Cell*, 143: 156-69.

16 Alexander, R., H. Lodish, and L. Sun. 2011. 'MicroRNAs in adipogenesis and as therapeutic targets for obesity', *Expert Opin Ther Targets*, 15: 623-36.

17 Lee, J. T. 2009. 'Lessons from X-chromosome inactivation: long ncRNA as guides and tethers to the epigenome', *Genes Dev*, 23: 1831-42.

18 Huarte, M., M. Guttman, D. Feldser, M. Garber, M. J. Koziol, D. Kenzelmann-Broz, A. M. Khalil, O. Zuk, I. Amit, M. Rabani, L. D. Attardi, A. Regev, E. S. Lander, T. Jacks, and J. L. Rinn. 2010. 'A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response', *Cell*, 142: 409-19.

19 Huarte, M., and J. L. Rinn. 2010. 'Large non-coding RNAs: missing links in cancer?', *Hum Mol Genet*, 19: R152-61.

20 Loewer, S., M. N. Cabili, M. Guttman, Y. H. Loh, K. Thomas, I. H. Park, M. Garber, M. Curran, T. Onder, S. Agarwal, P. D. Manos, S. Datta, E. S. Lander, T. M. Schlaeger, G. Q. Daley, and J. L. Rinn. 2010. 'Large intergenic non-coding RNA-RoR modulates reprogramming of human induced pluripotent stem cells', *Nat Genet*, 42: 1113-7.

21 Gupta, R. A., N. Shah, K. C. Wang, J. Kim, H. M. Horlings, D. J. Wong, M. C. Tsai, T. Hung, P. Argani, J. L. Rinn, Y. Wang, P. Brzoska, B. Kong, R. Li, R. B. West, M. J. van de Vijver, S. Sukumar, and H. Y. Chang. 2010. 'Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis', *Nature*, 464: 1071-6.

22 Rinn, J. L., and H. Y. Chang. 2012. 'Genome regulation by long noncoding RNAs', *Annu Rev Biochem*, 81: 145-66.

23 Derrien, T., R. Johnson, G. Bussotti, A. Tanzer, S. Djebali, H. Tilgner, G. Guernec, D. Martin, A. Merkel, D. G. Knowles, J. Lagarde, L. Veeravalli, X. Ruan, Y. Ruan, T. Lassmann, P. Carninci, J. B. Brown, L. Lipovich, J. M. Gonzalez, M. Thomas, C. A. Davis, R. Shiekhattar, T. R. Gingeras, T. J. Hubbard, C. Notredame, J. Harrow, and R. Guigo. 2012. 'The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression', *Genome Res*, 22: 1775-89.

24 Batista, P. J., and H. Y. Chang. 2013. 'Long noncoding RNAs: cellular address codes in development and disease', *Cell*, 152: 1298-307.

25 Cesana, M., D. Cacchiarelli, I. Legnini, T. Santini, O. Sthandier, M. Chinappi, A. Tramontano, and I. Bozzoni. 2011. 'A long noncoding RNA controls muscle differentiation by functioning as a competing endogenous RNA', *Cell*, 147: 358-69.

26 Ebert, M. S., and P. A. Sharp. 2012. 'Roles for microRNAs in conferring robustness to biological processes', *Cell*, 149: 515-24.

27 Wang, Y., Z. Xu, J. Jiang, C. Xu, J. Kang, L. Xiao, M. Wu, J. Xiong, X. Guo, and H. Liu. 2013. 'Endogenous miRNA sponge lincRNA-RoR regulates Oct4, Nanog, and Sox2 in human embryonic stem cell self-renewal', *Dev Cell*, 25: 69-80.

28 Trapnell, C., L. Pachter, and S. L. Salzberg. 2009. 'TopHat: discovering splice junctions with RNA-Seq', *Bioinformatics*, 25: 1105-11.

29 Trapnell, C., B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. van Baren, S. L. Salzberg, B. J. Wold, and L. Pachter. 2010. 'Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation', *Nat Biotechnol*, 28: 511-5.

30 Guttman, M., I. Amit, M. Garber, C. French, M. F. Lin, D. Feldser, M. Huarte, O. Zuk, B. W. Carey, J. P. Cassady, M. N. Cabili, R. Jaenisch, T. S. Mikkelsen, T. Jacks, N. Hacohen, B. E. Bernstein, M. Kellis, A. Regev, J. L. Rinn, and E. S. Lander. 2009. ' Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals', *Nature*, 458: 223-7.

31 Guttman, M., J. Donaghey, B. W. Carey, M. Garber, J. K. Grenier, G. Munson, G. Young, A. B. Lucas, R. Ach, L. Bruhn, X. Yang, I. Amit, A. Meissner, A. Regev, J. L. Rinn, D. E. Root, and E. S. Lander. 2011. 'lincRNAs act in the circuitry controlling pluripotency and differentiation', *Nature*, 477: 295-300.

32 Mercer, T. R., M. E. Dinger, and J. S. Mattick. 2009. 'Long non-coding RNAs: insights into functions', *Nat Rev Genet*, 10: 155-9.

33 MacIsaac, K. D., K. A. Lo, W. Gordon, S. Motola, T. Mazor, and E. Fraenkel. 2010. 'A quantitative model of transcriptional regulation reveals the influence of binding location on expression', *PLoS Comput Biol*, 6: e1000773.

34 Ruan, H., N. Hacohen, T. R. Golub, L. Van Parijs, and H. F. Lodish. 2002. 'Tumor necrosis factor-alpha suppresses adipocyte-specific genes and activates expression of preadipocyte genes in 3T3-L1 adipocytes: nuclear factor-kappaB activation by TNF-alpha is obligatory', *Diabetes*, 51: 1319-36.

35 Ruan, H., P. D. Miles, C. M. Ladd, K. Ross, T. R. Golub, J. M. Olefsky, and H. F. Lodish. 2002. 'Profiling gene transcription in vivo reveals adipose tissue as an immediate target of tumor necrosis factor-alpha: implications for insulin resistance', *Diabetes*, 51: 3176-88.

36 Ritchie, W., S. Granjeaud, D. Puthier, and D. Gautheret. 2008. 'Entropy measures quantify global splicing disorders in cancer', *PLoS Comput Biol*, 4: e1000011.

37 Lin, M. F., J. W. Carlson, M. A. Crosby, B. B. Matthews, C. Yu, S. Park, K. H. Wan, A. J. Schroeder, L. S. Gramates, S. E. St Pierre, M. Roark, K. L. Wiley, Jr., R. J. Kulathinal, P. Zhang, K. V. Myrick, J. V. Antone, S. E. Celniker, W. M. Gelbart, and M. Kellis. 2007. 'Revisiting the protein-coding gene catalog of Drosophila melanogaster using 12 fly genomes', *Genome Res*, 17: 1823-36.

38 Ponting, C. P., P. L. Oliver, and W. Reik. 2009. 'Evolution and functions of long noncoding RNAs', *Cell*, 136: 629-41.

39 Khalil, A. M., M. Guttman, M. Huarte, M. Garber, A. Raj, D. Rivea Morales, K. Thomas, A. Presser, B. E. Bernstein, A. van Oudenaarden, A. Regev, E. S. Lander, and J. L. Rinn. 2009. 'Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression', *Proc Natl Acad Sci U S A*, 106: 11667-72.

40 Wang, K. C., Y. W. Yang, B. Liu, A. Sanyal, R. Corces-Zimmerman, Y. Chen, B. R. Lajoie, A. Protacio, R. A. Flynn, R. A. Gupta, J. Wysocka, M. Lei, J. Dekker, J. A. Helms, and H. Y. Chang. 2011. 'A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression', *Nature*, 472: 120-4.

41 Tsai, M. C., O. Manor, Y. Wan, N. Mosammaparast, J. K. Wang, F. Lan, Y. Shi, E. Segal, and H. Y. Chang. 2010. 'Long noncoding RNA as modular scaffold of histone modification complexes', *Science*, 329: 689-93.

42 Koziol, M. J., and J. L. Rinn. 2010. 'RNA traffic control of chromatin complexes', *Curr Opin Genet Dev*, 20: 142-8.

43 Zappulla, D. C., and T. R. Cech. 2004. 'Yeast telomerase RNA: a flexible scaffold for protein subunits', *Proc Natl Acad Sci U S A*, 101: 10024-9.

44 Cannon, B., and J. Nedergaard. 2001. 'Cultures of adipose precursor cells from brown adipose tissue and of clonal brown-adipocyte-like cell lines', *Methods Mol Biol*, 155: 213-24.

45 Seale, P., S. Kajimura, W. Yang, S. Chin, L. M. Rohas, M. Uldry, G. Tavernier, D. Langin, and B. M. Spiegelman. 2007. 'Transcriptional control of brown fat determination by PRDM16', *Cell Metab*, 6: 38-54.

46 Guttman, M., M. Garber, J. Z. Levin, J. Donaghey, J. Robinson, X. Adiconis, L. Fan, M. J. Koziol, A. Gnirke, C. Nusbaum, J. L. Rinn, E. S. Lander, and A. Regev. 2010. 'Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs', *Nat Biotechnol*, 28: 503-10.

47 de Hoon, M. J., S. Imoto, J. Nolan, and S. Miyano. 2004. 'Open source clustering software', *Bioinformatics*, 20: 1453-4.

48 Tusher, V. G., R. Tibshirani, and G. Chu. 2001. 'Significance analysis of microarrays applied to the ionizing radiation response', *Proc Natl Acad Sci U S A*, 98: 5116-21.

49 Subramanian, A., P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov. 2005. 'Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles', *Proc Natl Acad Sci U S A*, 102: 15545-50.

50 Price, A. L., N. C. Jones, and P. A. Pevzner. 2005. 'De novo identification of repeat families in large genomes', *Bioinformatics*, 21 Suppl 1: i351-8.

51 Kent, W. J. 2002. 'BLAT--the BLAST-like alignment tool', *Genome Res*, 12: 656-64.

52 Bradley, R. K., A. Roberts, M. Smoot, S. Juvekar, J. Do, C. Dewey, I. Holmes, and L. Pachter. 2009. 'Fast statistical alignment', *PLoS Comput Biol*, 5: e1000392.

Chapter 3: The molecular and mechanistic characterization of Firre

3.1 Introduction

It has become clear that the mammalian genomes encode many lncRNAs with diverse functions in development and disease[1-6]. Recent work has begun to identify the myriad roles for lncRNAs, including but not limited to forming ribonucleoprotein (RNP) complexes with epigenetic regulatory machinery, transcriptional and post-transcriptional regulation of gene expression, and the formation sub-compartments in the nucleus to mediate higher-order chromosomal architecture, all of which result in the modulation of the genome to determine cellular states[7-14]. Examples of these phenomena can be observed in X chromosome dosage compensation in mammals. Several lncRNAs have been shown to recruit epigenetic regulatory complexes (e.g. the polycomb complex)[15,16], some of which are brought to the future inactive X chromosome[17,18]. Specifically, the lncRNA Xist binds to multiple proteins as a "scaffold" to mediate the silencing of genes on the X chromosome and affect the higher-order chromosomal architecture needed to establish proper epigenetic silencing[14,19,20]. More recently, a nuclear lncRNA termed *XACT* was shown to be associated with the active X chromosome, suggesting that additional lncRNAs may be involved in dosage compensation[13].

A majority of lncRNAs has been found to localize in the nucleus but an understanding of the roles of these molecules in the complex dynamics of the nucleus is missing. Beyond localizing protein complexes to their target loci, lncRNAs have been implicated in the dynamics of nuclear organization and the formation of sub-compartments such as the paraspeckles, nucleolus, and nuclear speckles[7,21-23]. Furthermore, RNA has also been shown to be an important structural component of the nuclear matrix[24-27] that is required for proper higher order chromosomal architecture. Although RNA has been associated with establishing

higher order nuclear architecture[8,28,29], the specific molecules and their mechanisms remain unknown. Recently, a lncRNA, *CISTR-ACT,* was discovered to facilitate *cis* and *trans* interactions in the nucleus, supporting a role for lncRNAs in organizing genomic architecture[8].

Given their properties and roles as described above, it has been hypothesized that lncRNAs might provide another layer of regulation for the establishment of cellular identities[30]. This is of interest because the question of how a wide variety of cell types are generated still remains unanswered. Xist, among other examples, exemplifies that lncRNAs can play important roles in driving cell fates.

In the discovery paper, we reported a loss of function screen for numerous lncRNAs that regulate adipogenesis in murine cell model systems[31]. We found that one of the most critically required lncRNAs from the screen is lnc-RAP1. One of the difficulties in studying the function and mechanism of lncRNAs has been the lack of a phenotypic outcome in the overexpression or knockdown studies. This might be due to the low basal expression level and/or rapid degradation. Another challenge in investigating lncRNAs is that there are no widespread functional mechanisms. lnc-RAP1, having a very strong phenotype and other intriguing properties, is a promising candidate to study and explore a possibly widespread mechanism of lncRNA function to regulate gene expression.

To that end, here, we characterize this intergenic lncRNA, lnc-RAP1, and term it Functional Intergenic Repeating RNA Element (Firre) that localizes across a 5 Mb domain around its site of transcription. This domain of Firre localization is also in spatial proximity to at least 5 other trans-chromosomal loci within the nucleus. This cross-chromosomal co-localization requires Firre as genetic deletion of *Firre* results in a loss of spatial proximity between its trans-chromosomal binding sites. We further identified a unique 156 bp repeating

RNA domain in the Firre sequence that is required to both interact with the nuclear matrix factor hnRNPU and localize Firre transcripts in a punctate manner in the nucleus. Strikingly, the knockdown of hnRNPU similar to deletion of *Firre* locus results in a loss of spatial proximity between the *Firre* locus and its trans-chromosomal binding sites. Collectively, these findings suggest a model where lncRNAs, such as Firre, can function as nuclear organization factors that interact with and influence higher order nuclear architecture across chromosomes.

3.2 Results

3.2.1 Firre is a novel and intriguing X chromosome-localized lncRNA

We previously identified Firre (previously referred to lnc-RAP1) as being required for proper adipogenesis in a loss of function screen in murine adipocyte precursors. A detailed subsequent analysis of Firre revealed many interesting and distinguishing features: (i) a diverse expression pattern of Firre *in vivo*, with enrichment in neural crest tissues as shown by *in situ* hybridization (Figure 3.2.1.1-9); (ii) a conserved intergenic human ortholog located on the X-chromosome (Figure 3.2.1.10), found using Transmap[32] to map the mouse lncRNA to its corresponding syntenic human locus; (iii) a unique 156bp Repeating RNA Domain (RRD) that occurs 16 and 8 times in *Mus musculus* (mouse) and *Homo sapien* (human) transcripts, respectively, with 96% sequence identity within species and 68% across species, detected using Fast Statistical Alignment[33] (Figure 3.2.1.11); (iv) numerous alternatively-spliced isoforms with differential inclusion or exclusion of RRD sequences (3.2.1.12). (v) Firre transcripts remain stable even after 6 hours of actinomycin D (ActD) treatment, as shown by RNA fluorescence in situ hybridization (FISH) (3.2.1.13).

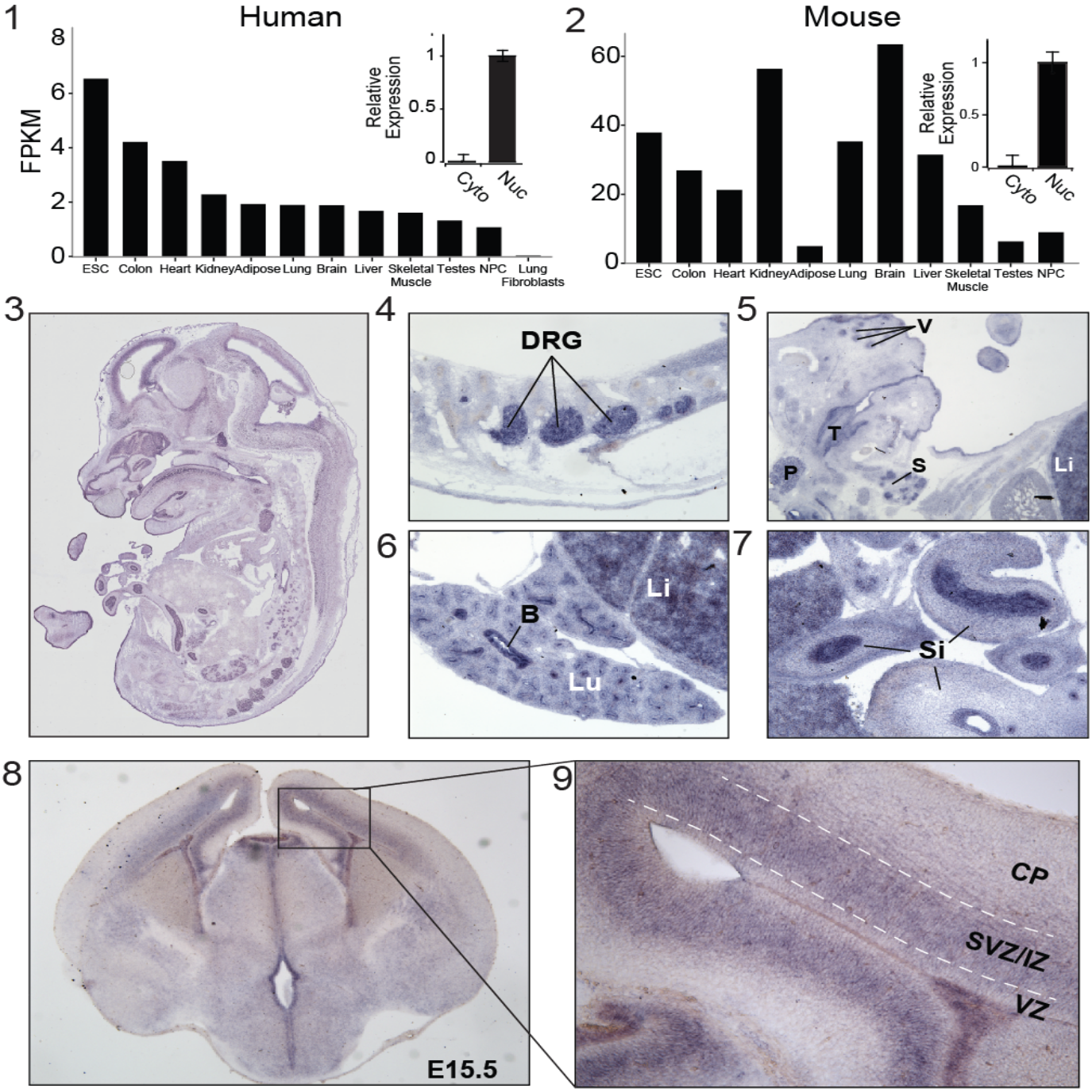


Figure 3.2.1.1-9: Firre expression in mouse and human tissues. RNA-seq and cellular fractionation (1,2) and by *in situ* hybridizations in the whole mouse embryo (E14.5) (3), dorsal root ganglia (DRG) (4), developing vibrissae (V) vibrissae (V), tongue (T), pituitary gland (P), salivary gland (S) (5), fetal liver (Li), lung (Lu), bronchi (B) (6), and small intestine (Si) (7), in the proliferative ventricular (VZ) and subventricular zones (SVZ) at E15.5 (8,9).

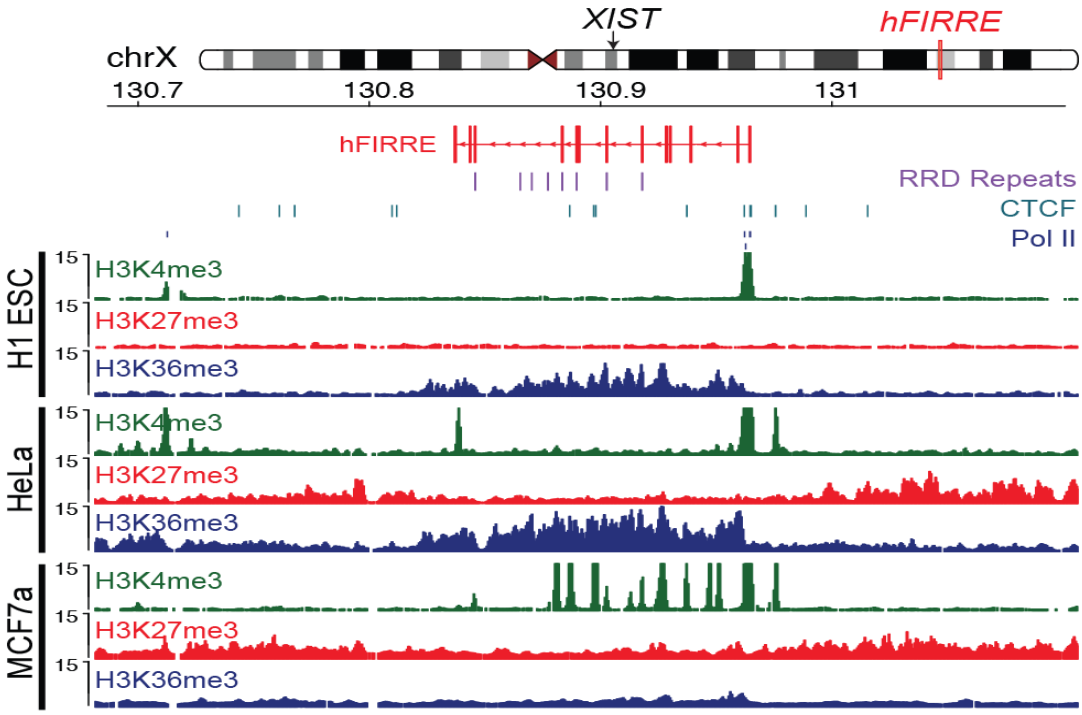


Figure 3.2.1.10: Human Firre locus in human embryonic stem cells (ESC), HeLa, and MCF7a. CTCF (light blue), H3K4me3 (green), H3K27me3 (red), and H3K36me3 (dark blue) ChIP peaks.

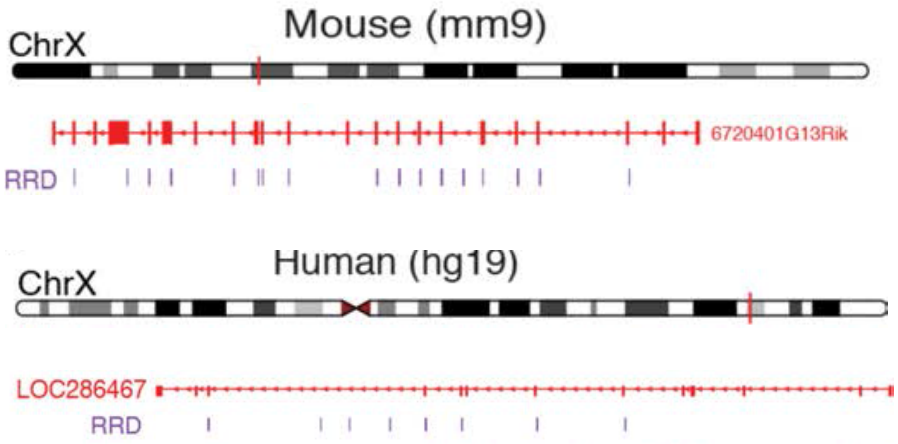


Figure 3.2.1.11: The exon/intron structure of mouse and human Firre RNA and the RRD repeats in both species.

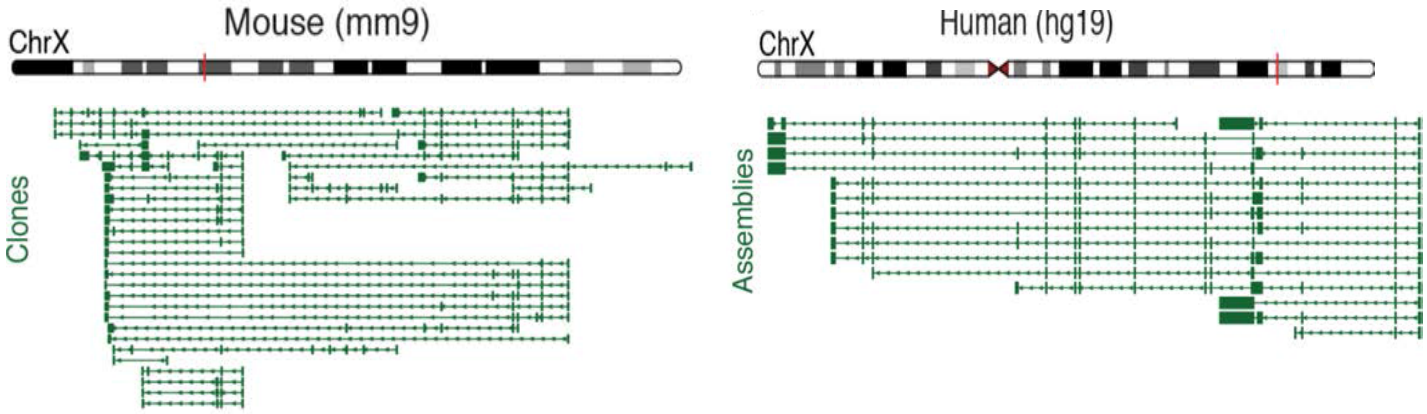


Figure 3.2.1.12: Clones of 50 different mouse Firre isoforms and clones and assemblies of human Firre isoforms.

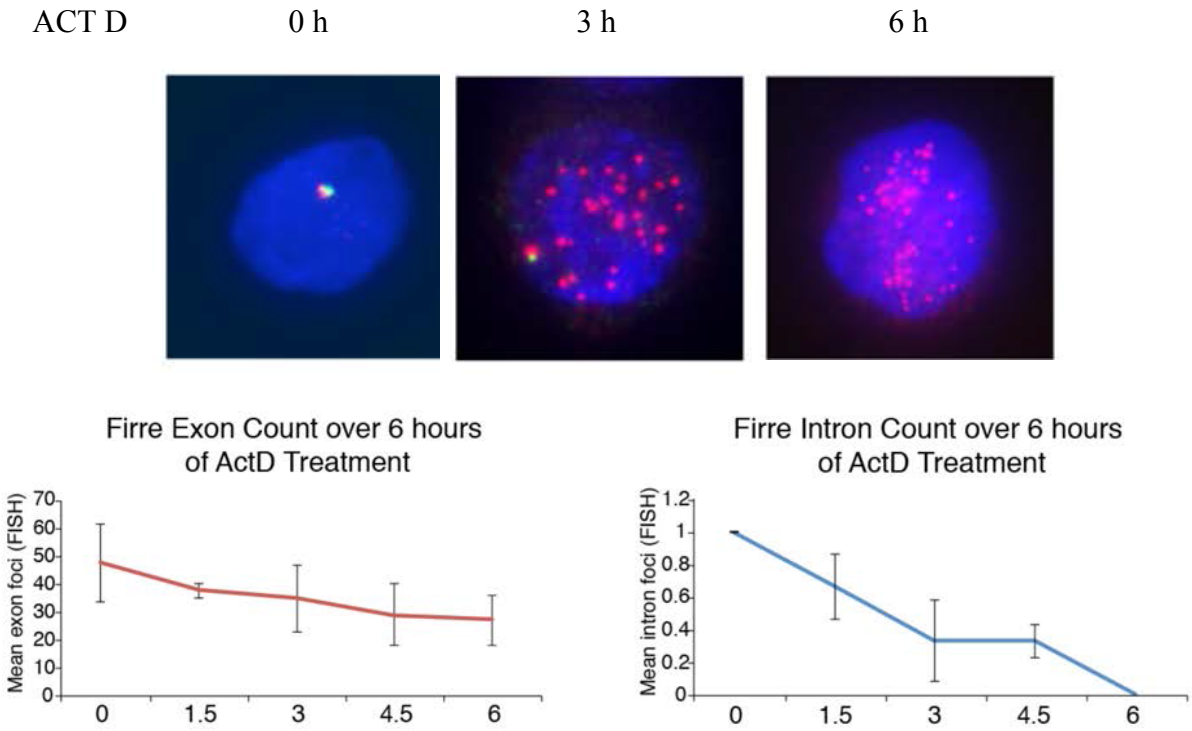


Figure 3.2.1.13: Actinomycin D treatment of mouse ESCs followed by RNA FISH. FISH images and their quantification using StarSearch after 0, 3, and 6 hours of ActD. Green corresponds to intron labeling with Alexa 594 and red corresponds to exon labeling with Cy3 in the FISH images.

3.2.2 Firre is a nuclear retained and chromatin associated RNA

To further determine the subcellular localization of Firre, we used single molecule RNA fluorescence *in situ* hybridization (FISH) targeting Firre (as described in [34]). We adopted a dual labeling strategy to independently target the introns and exons of Firre, thus marking the site of transcription on the X chromosome (intronic probes) and the location of the mature transcripts (exonic probes) separately (Figure 3.2.2.1, Appendix 3). RNA-FISH analysis revealed an exclusively nuclear and focal distribution for Firre in all cells tested. Notably, Firre exhibits strong expression foci near its site of transcription in both male and female mouse and human ESCs (mESCs, hESCs) (Figure 3.2.2.1). We also note that

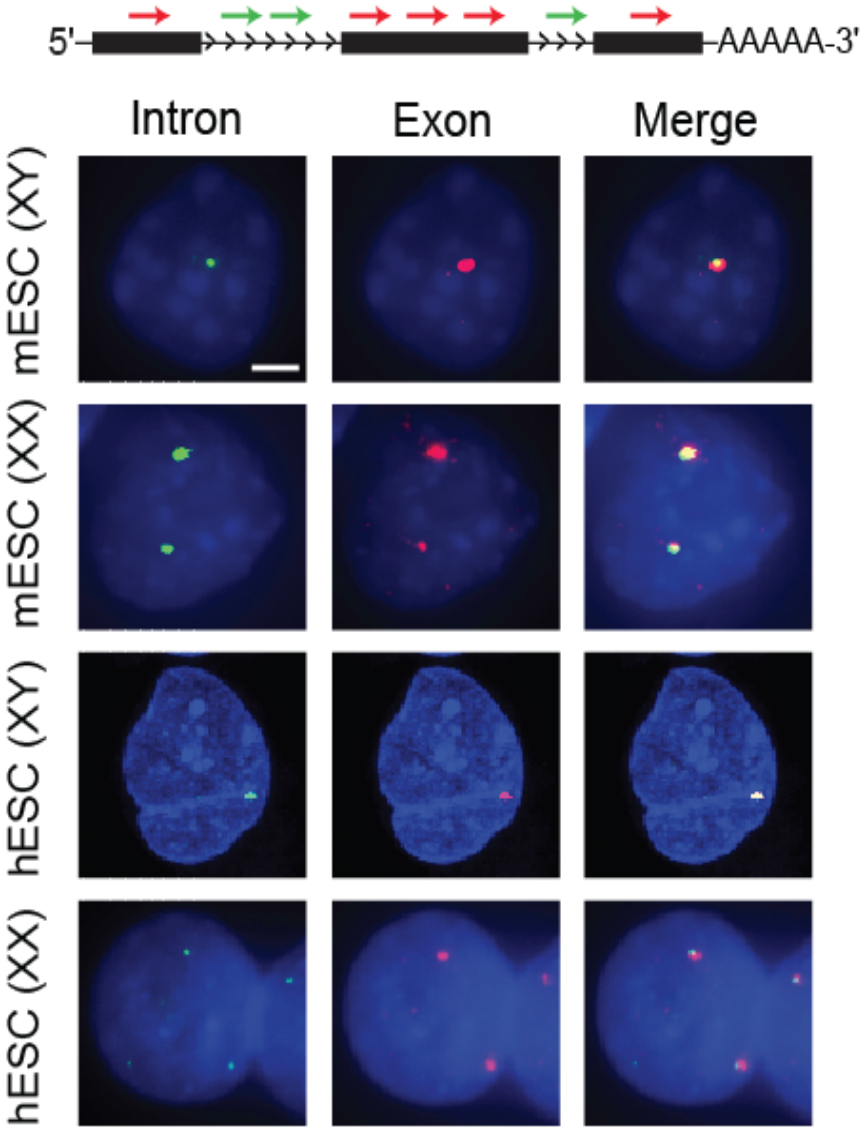


Figure 3.2.2.1: Single molecule RNA FISH in mES and hES cells. Intron probes in "green" (A594) and exon probes in "red" (Cy3) targeting Firre. Scale bar: 20 μm, nuclei by DAPI.

the Firre RNA is localized around its site of transcription but extend slightly beyond this site in all six human and mouse cell lines tested. Sub-cellular localization and expression of Firre in cell lines with and without inactive X-chromosomes was similar to that observed in ESCs (Figure 3.2.2.2). Thus, Firre is nuclear-localized and forms expression foci on both X chromosomes prior to, and after X chromosome inactivation.

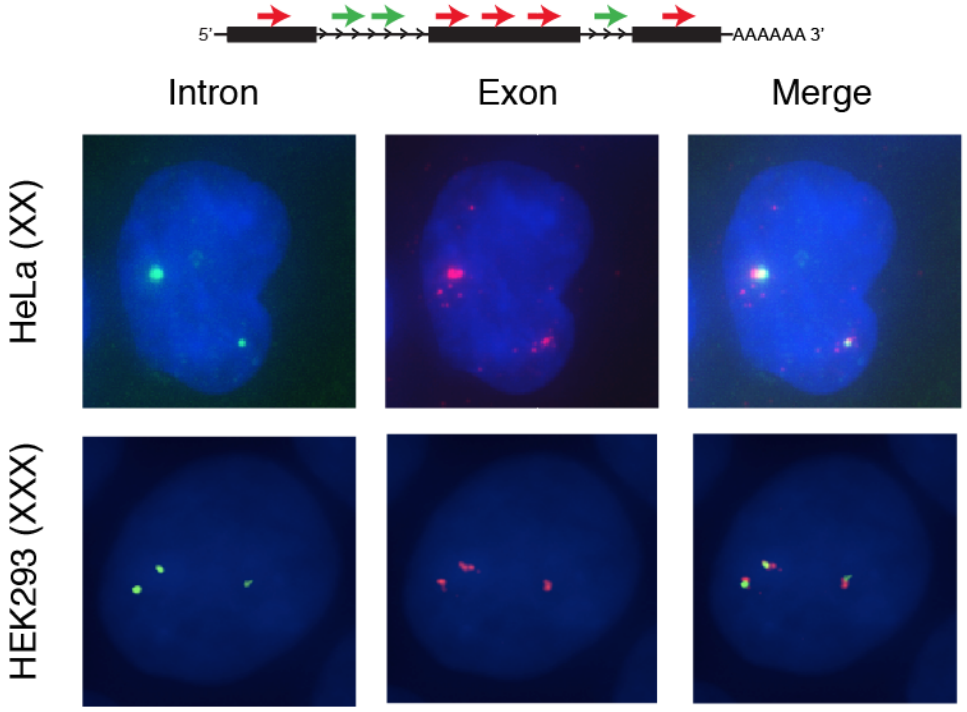


Figure 3.2.2.2: Single molecule RNA FISH in HEK293 and HeLa cells. Intron probes in "green" (A594) and exon probes in "red" (Cy3) targeting Firre. Scale bar: 10 μm, nuclei by DAPI.

We tested if over-expression of Firre was sufficient to form the observed endogenous local expression foci. Briefly, we ectopically expressed Firre via retrovirus-mediated integration in human and mouse lung fibroblasts, which do not express Firre (Figure 3.2.2.3). We observed many sites of focal accumulation of Firre. We obtained similar results upon ectopic expression of Firre in human cells that endogenously express this lncRNA (HEK293)

(Figure 3.3.2.4). We repeated the experiments above using isoforms of Firre with (isoform 4; one repeat) or without the RRD (isoform 6). Strikingly, in the absence of RRD, the nuclear localization of Firre is disrupted, and Firre RNA is detected in the cytoplasm, given that the expression levels of both constructs are comparable (Figure 3.2.2.5). Thus, RRD is required for the focal nuclear localization of Firre.

3.2.3 The Firre locus escapes X chromosome inactivation

Our observation from RNA FISH in female mouse and human ESCs and HEK293s led us to hypothesize that *Firre* might escape X chromosome inactivation (XCI). To test this, we analyzed the local chromatin environment within the *Firre* locus using existing chromatin immunoprecipitation (ChIP) data for numerous histone modifications and transcription factors. Several of these data are consistent with *Firre* escaping XCI: First, we observed an appreciable depletion of LaminB1 across the mouse *Firre* locus and across the human *FIRRE* locus in various cell lines (Figure 3.2.3.1). LaminB1 is a matrix protein involved in nuclear stability and chromatin organization and is known to mark heterochromatin[35,36]. The domain of LaminB1 depletion extends precisely across the body of the *Firre* gene but not into the upstream or downstream regions. Second, the *Firre* locus is specifically and significantly ($p<1.0x10^{-8}$) depleted of trimethylated histone 3 lysine 27 (H3K27me3) in differentiated mESCs and in human cells prior to and after X-chromosome inactivation (Figure 3.2.3.1). Third, the *Firre* locus is enriched for trimethylated histone 3 lysine 4 (H3K4me3) with and without Firre transcription Figure 3.2.3.1). Finally, we observed a striking pattern of CCCTC-binding factor (CTCF) (Figure 3.2.3.1), which can function as an insulator between chromatin domains and facilitates inter-chromosomal interactions, localization adjacent to almost every exon of Firre[37].

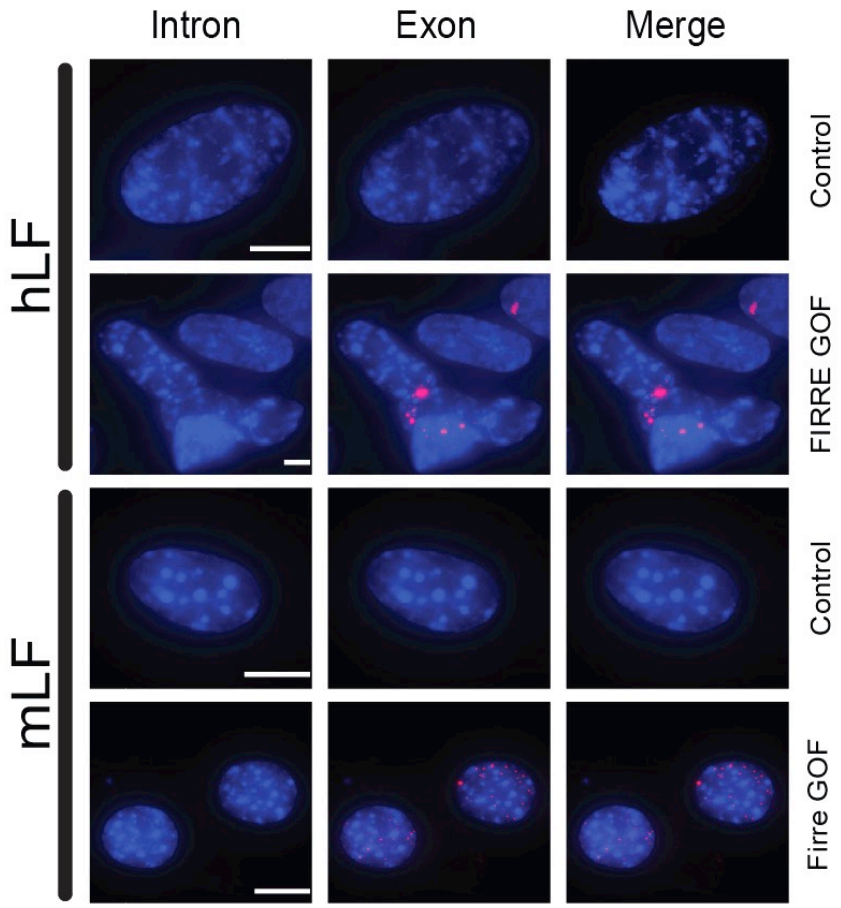


Figure 3.2.2.3: Viral overexpression of Firre in human and mouse lung fibroblasts (hLF, mLF). hLFs and mLFs do not endogenously express Firre. Scale bar: 15 μm, introns in "green" (A594), exons in "red" (Cy3), and nuclei by DAPI.

To further test the hypothesis that *Firre* escapes X-chromosome inactivation, we investigated whether Xist RNA itself localizes on the *Firre* locus upon X-chromosome inactivation. Specifically, we examined the localization of Xist on DNA using data generated by RNA Affinity Purification (RAP) in mouse lung fibroblasts (mLF). In contrast to the enrichment of Xist across most of the X-chromosome[19], we observed a strong and focal depletion in Xist binding at the *Firre* locus; similar to what was observed at genes known to escape XCI (Figure 3.2.3.1, bottom track). Interestingly, we note that the Xist-depleted boundaries are consistent with the previously identified boundaries for the lamin-depleted

regions. Collectively, these data indicate that the *Firre* locus escapes X chromosome inactivation and has a notable enrichment for CTCF and H3K4me3 and depletion for H3K27me3 and LaminB1.

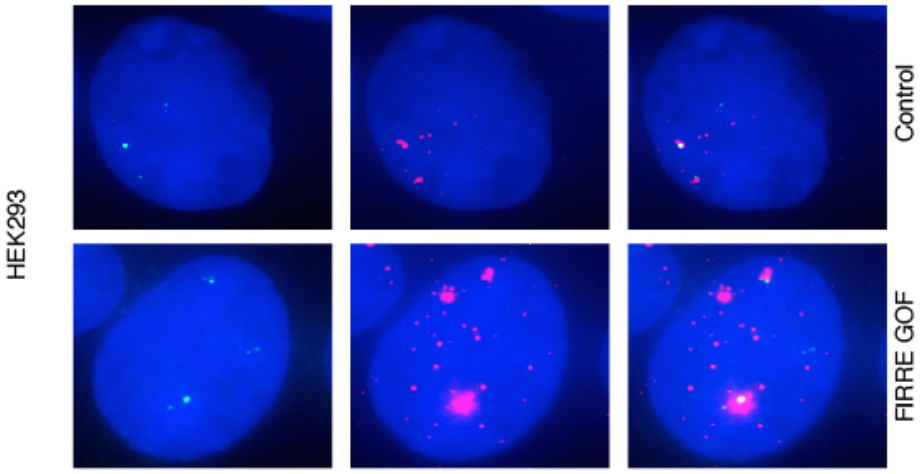


Figure 3.2.2.4: Viral overexpression of Firre in HEK293 cells shown by RNA FISH. Scale bar: 10 μm, introns in "green" (A594), exons in "red" (Cy3), and nuclei by DAPI.

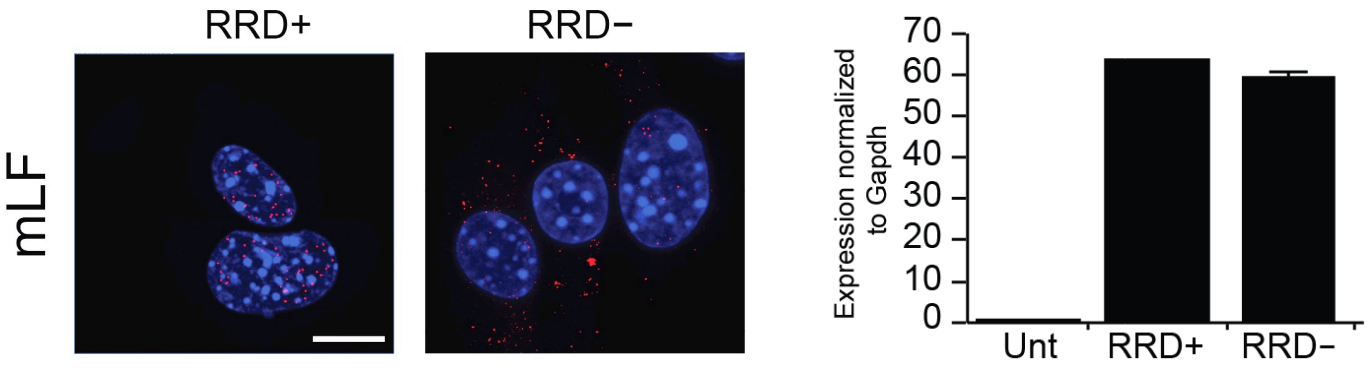


Figure 3.2.2.5: Viral overexpression of Firre isoforms with or without RRD in mLFs. Scale bar: 15 μm, exons in "red" (Cy3), and nuclei by DAPI. Errors are 1 s.d., n=3 for each condition.

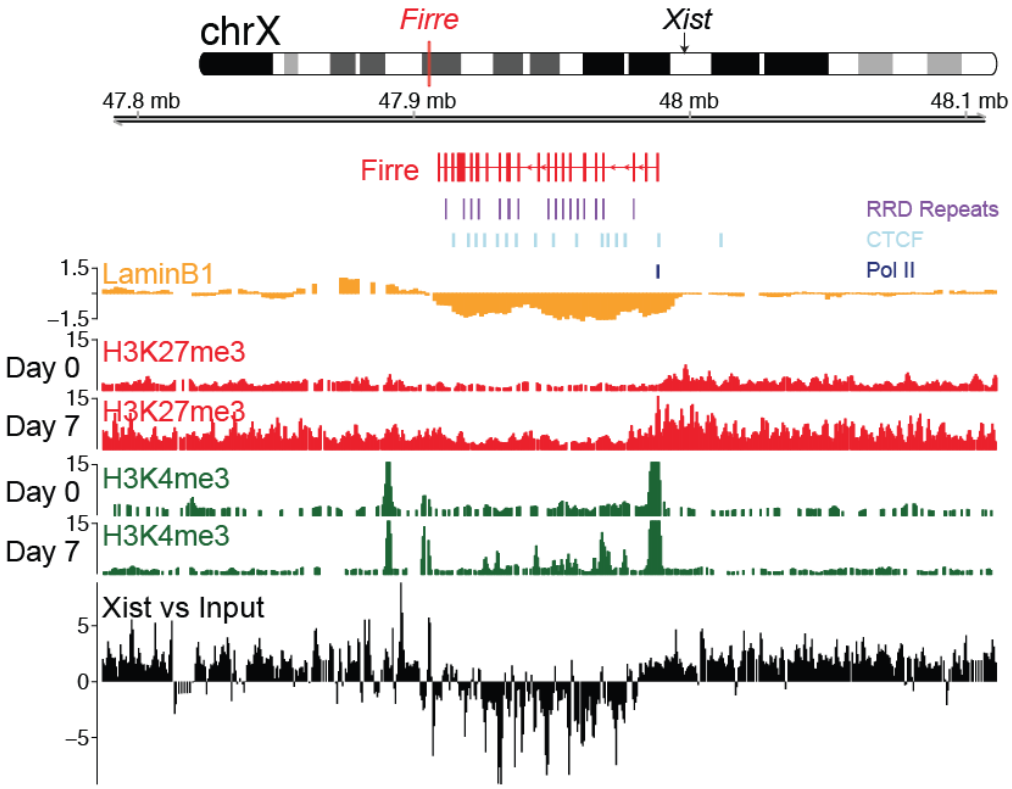


Figure 3.2.3.1: The mouse Firre locus. Repeat motif RRD (purple), CTCF (blue), LaminB1 (orange), H3K27me3 (red) and H3K4me3 (green), and Xist RAP (black). LaminB1 and Xist plotted as log fold-change on the y-axis relative to input; and chromatin modifications as raw counts.

3.2.4 Firre localizes to chromatin in cis and trans

The focal nuclear localization of Firre near its site of transcription led us to identify the direct interactions between Firre and chromatin in the nucleus. To resolve the DNA binding sites of Firre genome-wide, we performed RAP[19]. RAP along with ChiRP[38], CHART[39], ChOP[40] provide genome-wide DNA-binding locations for RNAs by cross-linking chromatin and RNA, followed by the targeting and pull-down of a specific RNA using antisense oligos. We performed RAP in male mESCs using two sets of 120 bp antisense probes targeting Firre

and two sets of sense probes as negative controls, followed by sequencing to identify genomic regions directly bound by Firre. We observed a ~5 Mega base (Mb) domain of Firre localization around the *Firre* locus (Figure 3.2.4.1). Strikingly, we also observed five significantly enriched peaks (Cuffdiff2; 1% FDR) of Firre located on chromosomes 2, 9, 15, and 17 that overlap known genes including *Slc25a12*, *Ypel4*, *Eef1a1*, *Atf4*, and *Ppp1r10* (Figure 3.2.4.2). Notably, 4 out of 5 of these genes have previously described regulatory roles during adipogenesis[41-44], consistent with our previous study showing the role of Firre in adipogenesis[31]. Expanding this search to regions not overlapping with mRNAs, we observed a total of 34 additional significant (Cuffdiff2; 1% FDR) localization sites for Firre.

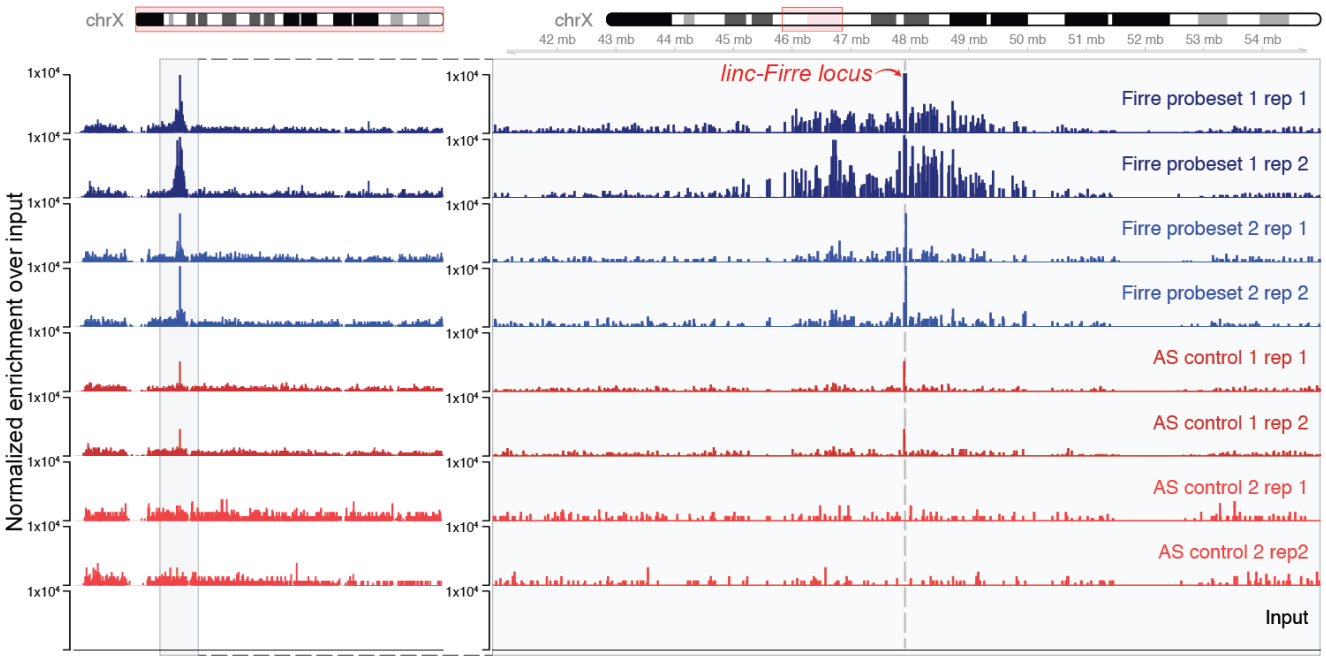


Figure 3.2.4.1: RNA Affinity Purification (RAP) by Firre along the X chromosome in male mESCs. Peaks shown as fold enrichment relative to input after normalization for library depth. Two Firre-targeting probes (blue tracks) and two sets of sense probes (negative controls) (red tracks) normalized to the input (black).

Collectively these data suggest that Firre is localized on multiple chromosomes, yet has only one predominant nuclear localization site in male and two in female cells around its site of transcription. These observations suggest two possible models. One possibility is that Firre could be shuttled from its site of transcription to these sites on other chromosomes. Alternatively, the focal localization of Firre to its own genomic locus could serve as a regional organizing factor to bring the trans-interacting sites into the three-dimensional proximity of the *Firre* locus on the X chromosome.

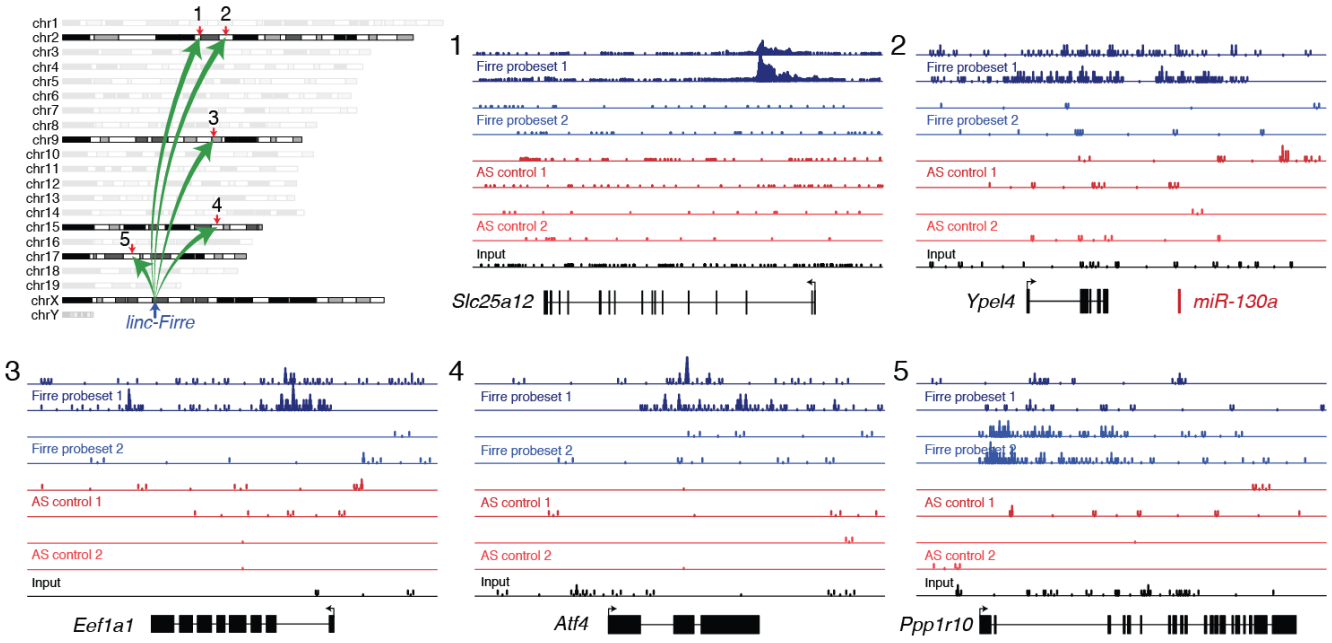


Figure 3.2.4.2: RAP by Firre shown for 5 distinct inter-chromosomal genomic loci. Loci are *Slc25a12, Ypel4, Eef1a1, Atf4, and Ppp1r10*. Counts for the trans-chromosomal contacts shown after normalization for sequencing depth.

3.2.5 Firre trans-chromosomal sites are in spatial proximity

In order to determine the nature of the trans-chromosomal interactions for Firre, we performed single molecule RNA co-FISH in mESCs on the transcription sites of Firre and three

of the *trans*-interacting genes (*Slc25a12, Ypel4* and *Ppp1r10*) (Figure 3.2.5.1). As negative controls, we performed similar RNA co-FISH for Firre and several genes with high expression in mESC that were not detected by RAP as *trans* targets, (e.g. *Pouf51 (Oct4), Nanog* and *Sox2*) (Figure 3.2.5.2). Remarkably, we observed co-localizations between Firre *and* all three trans-sites tested: Slc25a12 (73.9% of cells), Ypel4 (79.4% of cells), and Ppp1r10 (78.1% of cells) (Figure 3.2.5.1), and between these trans sites (Figure 3.2.5.3). Conversely, we did not observe any co-localization of Firre and unbound targets Oct4, Nanog, or Sox2 (Figure 3.2.5.2). Thus, these results are consistent with the latter model, where the *Firre* locus resides in three-dimensional proximity to these trans-chromosomal binding sites.

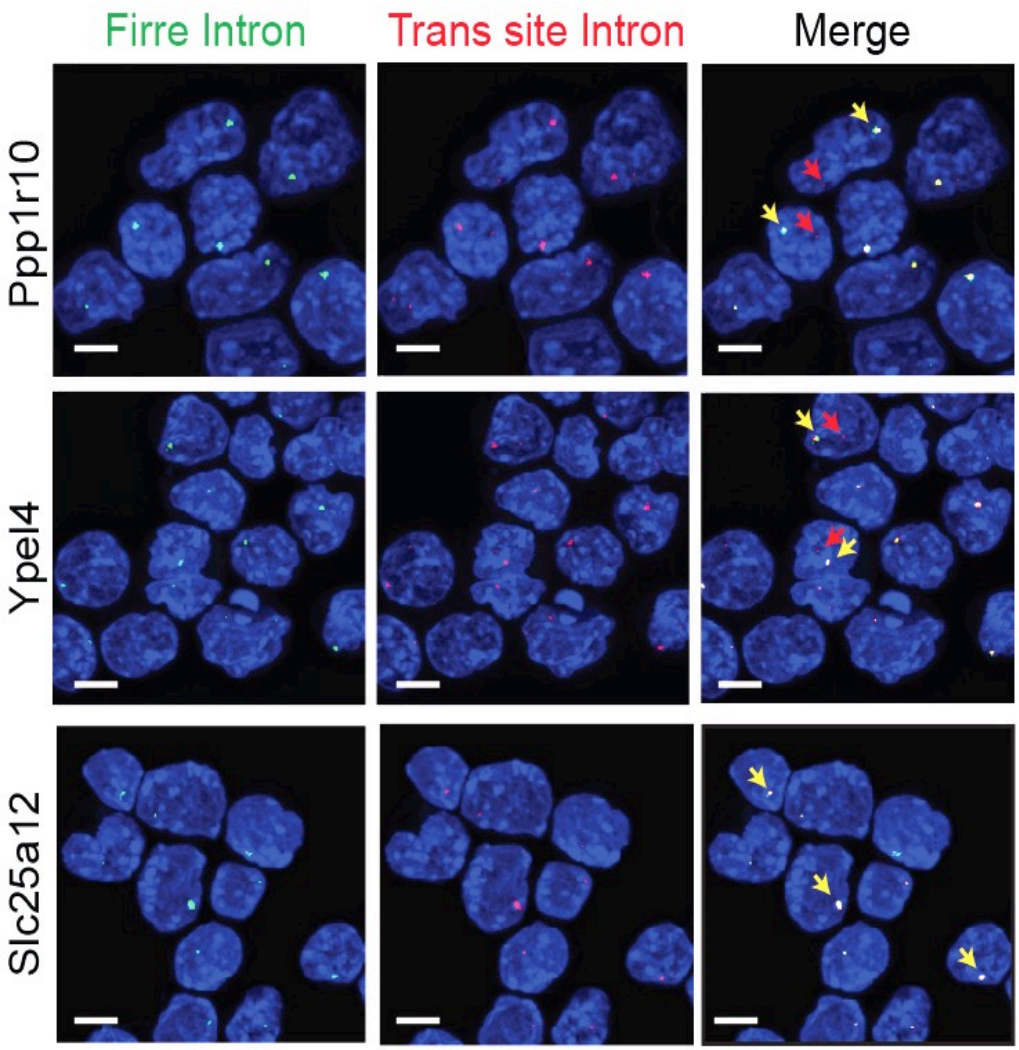


Figure 3.2.5.1: Co-localization of Firre with its trans targets: Ppp1r10, Ypel4, and Slc25a12. Scale bar: 40 μm, intron of Firre in "green" (A594) and introns of its trans targets in "red" (Cy3), and the nuclei by DAPI.

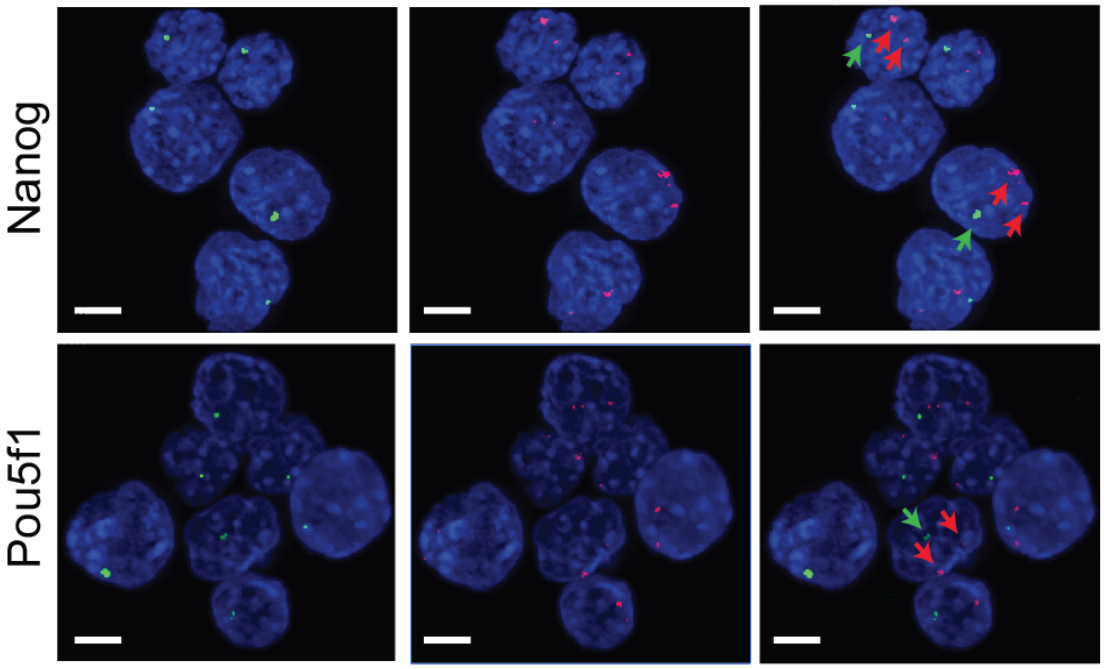


Figure 3.2.5.2: Co-RNA FISH of Firre with Nanog and Oct4. Scale bar: 40 μm, intron of Firre in "green" (A594) and introns of its trans targets in "red" (Cy3), and the nuclei by DAPI.

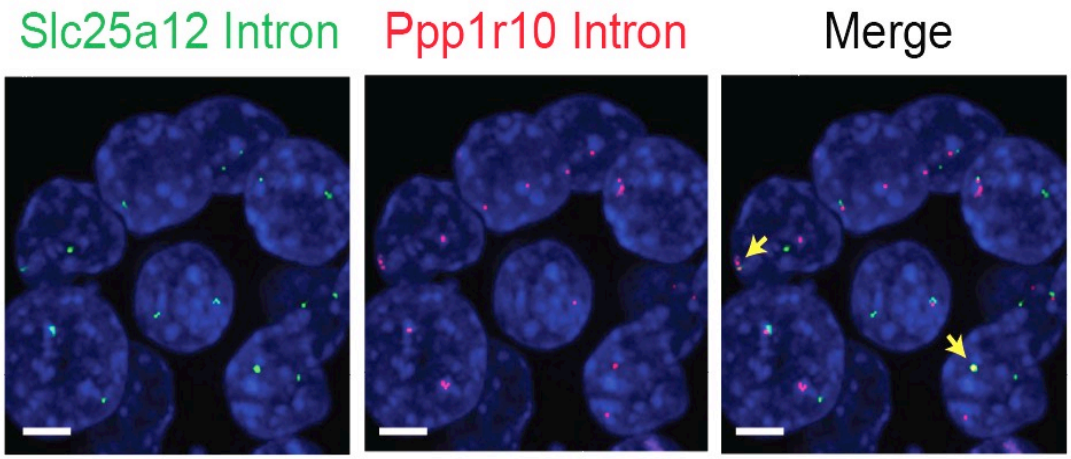


Figure 3.2.5.3: Co-localization of the trans-interacting loci Ppp1r10 and Ypel4. Scale bar: 40 μm, introns of Slc25a12 in "green" (A594) and introns of Ppp1r10 in "Red" (Cy3), and nuclei by DAPI.

3. 2.6 Firre regulates key pluripotency pathways

To determine the functional role of Firre, we generated a Firre knockout male mESC line by deleting the entire *Firre* locus on the X-chromosome (*ΔFirre*). Briefly, we introduced loxP sites in the 5' and 3' of the *Firre* locus by a two-step targeting strategy. Then, we infected the cells that harbor this locus with a Cre plasmid and clonally selected the cells with the proper deletion. Comparison of wild-type and *ΔFirre* growth rates revealed a marked retardation in growth rate and colony formation (Figure 3.2.6.1). We also note an intermediate growth defect when the cells were grown on a mouse embryonic fibroblast feeder layer (Figure 3.2.6.1). The *ΔFirre* cells on feeders were able to form bigger and more colonies in the same amount of time when compared to the *ΔFirre* cells grown without feeders (Figure 3.2.6.1).

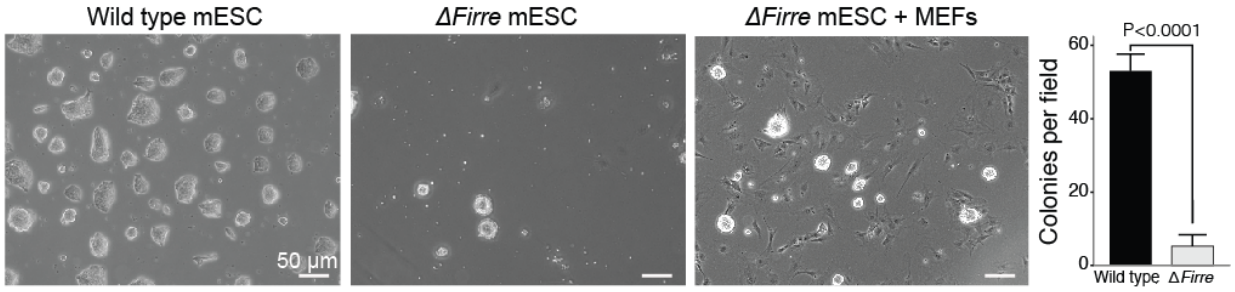


Figure 3.2.6.1: Deletion of *Firre* locus in male mESC. (Δ*Firre)* with (+MEF) and without feeder cells and quantification of the number of colonies per field for Δ*Firre* and wild type mESCs (Student's T-test).

To identify the gene-pathways and molecular signature that are altered upon deletion of Firre, we conducted massively parallel RNA-sequencing (RNA-Seq) comparisons between wild-type and *ΔFirre*. Briefly, RNA was isolated from three wild-type (WT) and two *ΔFirre* replicate cultures and subjected to paired-end illumina sequencing to a mean depth of ~$9x10^6$ fragments aligned per replicate. We identified 1077 genes with significant differential

expression (Cuffdiff2; 5% FDR) between the WT and *ΔFirre* mESCs (Figure 3.2.6.2). Preranked GSEA analysis demonstrated that *ΔFirre* cells were significantly enriched ($p<0.01$) for genes involved in extracellular matrix organization, and cell surface receptor-ligand

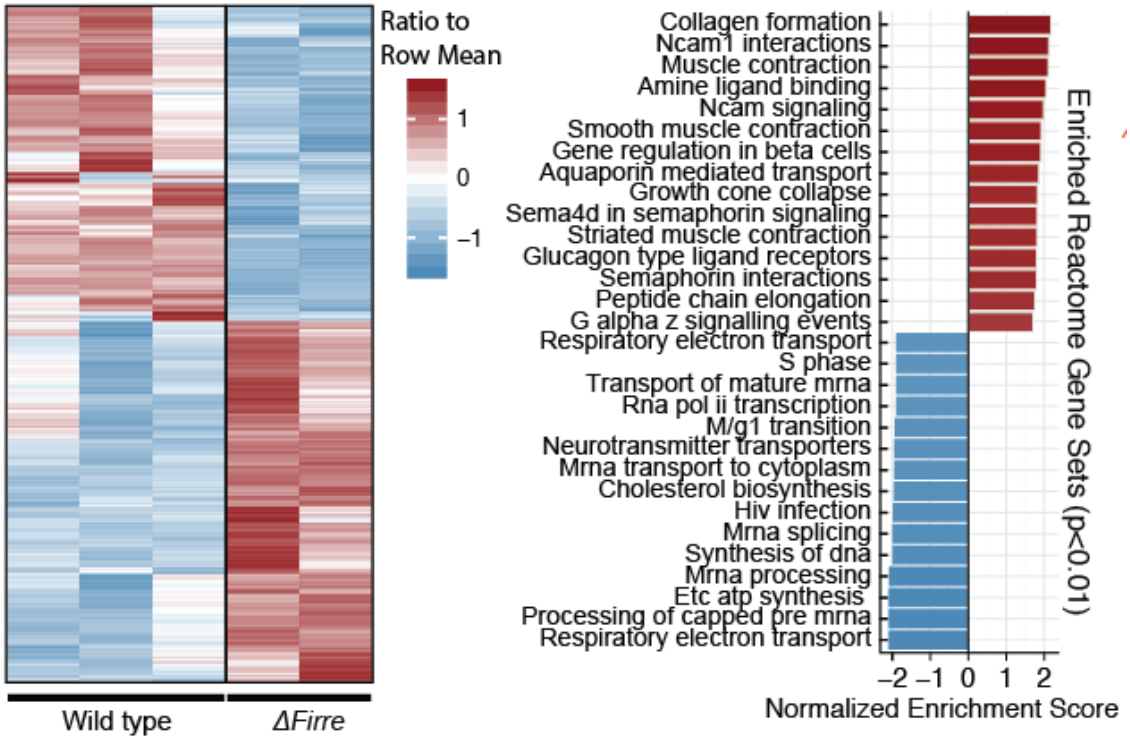


Figure 3.2.6.2: Heatmap of 892 significantlly differentially expressed genes between wild type and Δ*Firre* male mESCs. (Cuffdiff2; 1%FDR). Right: Top 15 enriched and depleted significant ($p<0.01$; Mann-Whitney U-test) Reactome gene sets from a pre-ranked GSEA analysis on Cuffdiff2 test statistics.

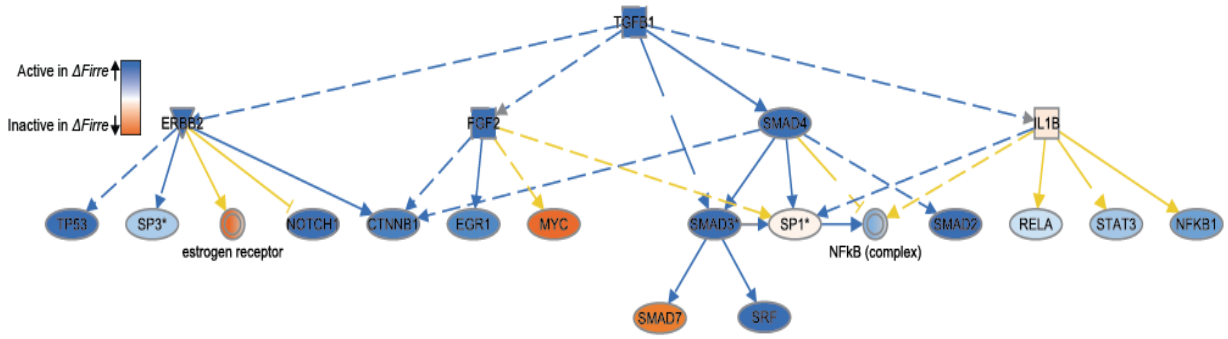


Figure 3.2.6.3: Ingenuity Pathway Analysis mechanistic network diagram. Showing significant ($p<6.31x10^{-44}$) increase in predicted downstream Tgfβ signaling activity in the Δ*Firre* male mESC relative to wild type.

interactions (Figure 3.2.6.2) Conversely, *ΔFirre* mESCs were depleted for genes involved in mRNA processing, nuclear export, and electron transport chain-mediated energy metabolism, and glucose metabolism relative to WT (Figure 3.2.6.2). Notably, we observed an increase in Tgfβ signaling in the *ΔFirre* mESCs (Figure 3.2.6.3). Interestingly Tgfβ signaling is known to be a potent inhibitor of adipogenesis[45], consistent with our previous observation that knockdown of Firre strongly inhibits adipogenesis in mouse preadipocytes[31] and growth defects observed in mESC cultures (Figure 3.2.6.1).

We next tested whether or not the Firre trans localization sites were affected by the absence of Firre (Figure 3.2.6.4). We did not observe a global enrichment for the five trans-site genes in the list of significantly differentially expressed genes ($p<1.0$; Hypergeometric test). We did observe one key exception, Ppp1r10, one of the 3 validated trans-sites, that was significantly decreased (Cuffdiff2; 1% FDR) in the *ΔFirre* cells relative to WT. However, we cannot preclude the possibility of perturbations to mRNA stability, translation, or processing at any of the remaining trans-sites.

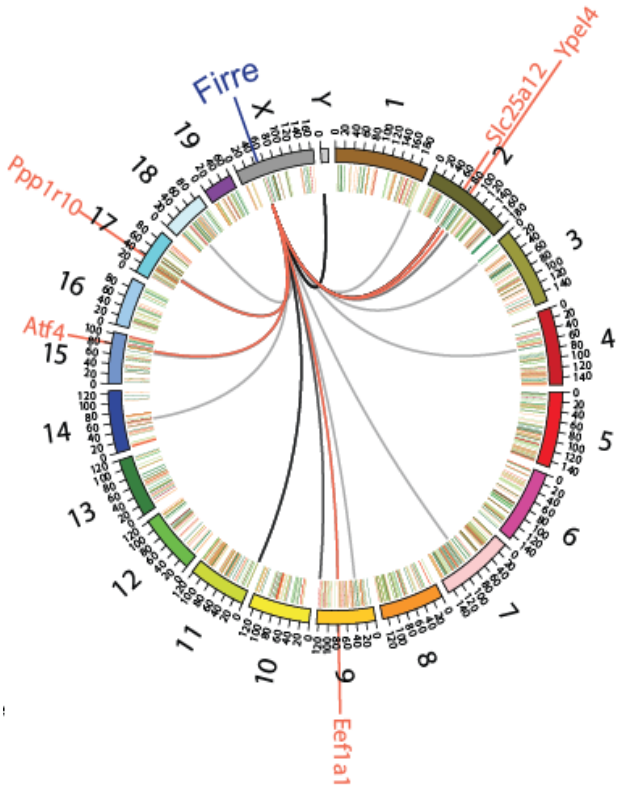


Figure 3.2.6.4: Circos diagram of significant Firre RAP peaks (links) interacting with the *Firre* genomic locus (blue) in male mESCs. Peaks intersecting genic regions are highlighted in red and specifically labeled. Log_2 fold changes for significant (Cuffdiff2; 1% FDR) differentially expressed genes (Δ*Firre*/WT) are inscribed at corresponding genomic locations within the circle.

3.2.7 Firre binds hnRNPU in an RRD-dependent manner

We next turned to identify proteins that interact with Firre that might mediate its 5Mb X chromosome localization. As a first approach to identify the candidate protein partners of Firre, we performed RNA pull-down assays in mouse ESC and adipocyte lysates by biotinylating the RNA, either by body-labeling (*in vitro* transcription) or 3' end-labeling (pCp-biotin), followed by mass-spectrometry (Appendix 4). Unrelated ncRNAs (sense and antisense telomerase RNA TERC) were used as negative controls. To identify proteins that preferentially co-precipitated with Firre in an RRD-dependent manner, we used five different RRD-positive isoforms and one RRD-negative isoform and took the difference between the peptide counts of the RRD-positive and RRD-negative isoforms (Appendix 4). We repeated the differential analysis for each of the five isoforms and took the top 10% of the differential peptide count scores for each isoform identified in both mESC and adipocyte lysates. Based on the highest unique peptide counts, we

identified 8 candidate proteins that physically associate with Firre in an RRD-dependent manner (Appendix 5).

The highest ranked candidate from this analysis was heterogeneous ribonucleoprotein U (hnRNPU) (Appendix 5). hnRNPU was of particular interest because it is required for the proper localization of Xist, interaction with the scaffold attachment regions on DNA, and the formation of highly structured chromatin territories[46-48]. To confirm the interaction between hnRNPU and Firre, we repeated the RNA pull-down experiments described above, and assayed for hnRNPU via Western blotting. In both mESC and adipocyte lysates, hnRNPU co-precipitated with Firre, but not with the negative controls (Figure 3.2.7.1, Appendix 4). We further tested additional hnRNP family members and found either no association or association independent of RRD (Figure 3.2.7.1, Appendix 4). Together these data suggest that Firre associates with hnRNPU.

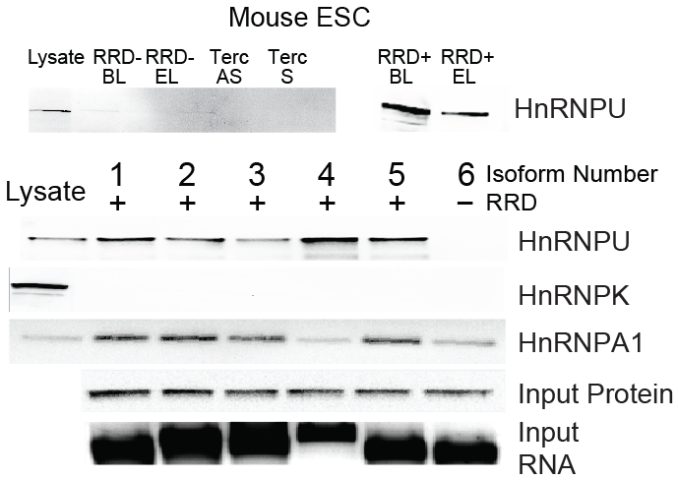


Figure 3.2.7.1: Western blots for RNA pull-downs. Five with RRD positive isoforms of Firre and one without. Biotin end labeling (EL) versus body labeling (BL) of RNA for pull-downs. TERC antisense (AS) and sense (S) as negative controls. hnRNPK and A1 shown for RRD specificity. 20% of lysate was loaded. Input protein lysate and input RNA shown as loading controls.

To test if the Firre RRD element is required and sufficient to interact with hnRNPU, we performed RNA pull-downs using single and double copies of the mouse and human RRD sequences in mESC lysates. We observed binding of hnRNPU to both mouse and human synthetic RRD constructs (Figure 3.2.7.2, panel 1). Finally, western blot analysis confirmed that hnRNPU binds to the Firre isoform harboring RRD but not to the ΔRRD isoform (Figure 3.2.7.2 panel 2). To determine if the hnRNPU–Firre interaction is biologically relevant at endogenous levels, we captured endogenous Firre using complimentary DNA oligos. Briefly, desthiobiotin-18 linker 23-25 bp DNA oligos targeting either RRD (capture efficiency 22% of the total input Firre, Figure 3.2.7.3) or non-targeting scramble controls were incubated with the mESC and HEK293 lysates and the co-precipitated RNP complexes were isolated with streptavidin beads (Figure 3.2.7.3). With this method, we confirmed that hnRNPU co-purifies specifically with the lncRNA (Figure 3.2.7.2, panels 3&4, Figure 3.2.7.3). To further test the specificity of the oligos targeting Firre, these experiments were repeated in mESC lysates depleted of Firre (greater than 85% knockdown) (Figure 3.2.7.3), resulting in 90% decreased recovery of hnRNPU despite similar RNA capture efficiency (Figure 3.2.7.3).

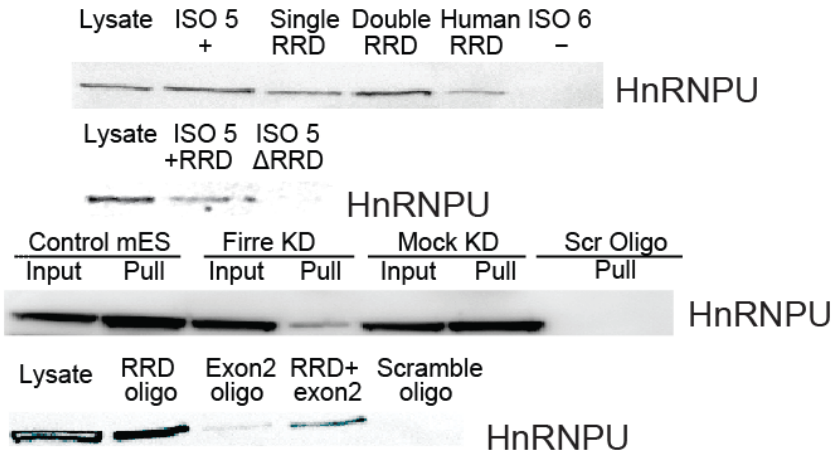


Figure 3.2.7.2: Western blots for RNA pull-downs performed with synthetic RRD constructs. Single, double, and human RRDs used to pull-down hnRNPU in mESCs (panel 1); RRD deleted isoform in mESCs (panel 2). Western blots for endogenous RNA pull-downs: desthiobiotin-DNA oligos complementary to Firre compared to non-targeting scramble oligos in mESCs and HEK293s (panels 3&4).

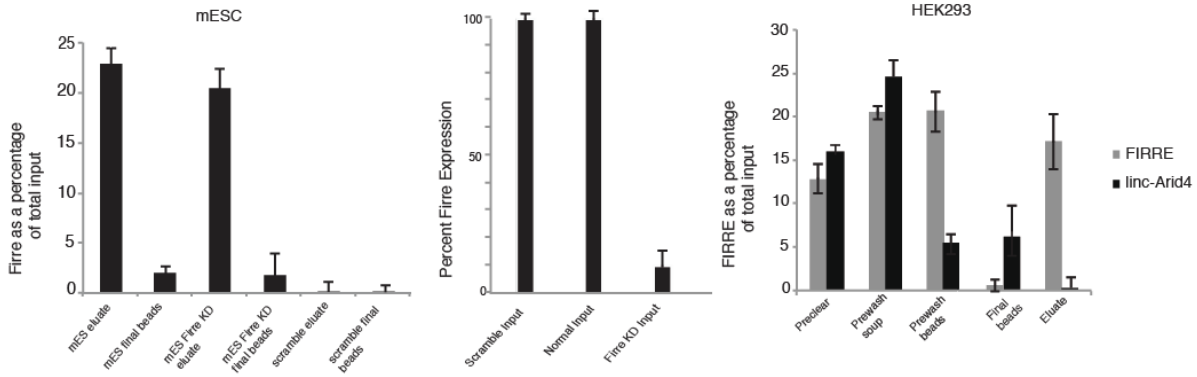


Figure 3.2.7.3: Endogenous Firre capture using desthiobiotin-modified DNA oligos in mESCs and HEK293s. Efficiencies measured by qRT-PCR, comparing the eluate and post-eluate fractions by normalizing to 10% total input. Specificity of oligos tested by scramble pull-downs, pull-downs in knockdown Firre mESCs, and unrelated lncRNA (Arid4) pull-downs.

To further validate the hnRNPU–Firre interaction, we performed the reciprocal experiment: RNA immunoprecipitation (RIP) targeting hnRNPU. Consistent, with the RNA pull-downs, we found strong enrichment of Firre relative to IgG controls after normalization to total input (Figure 3.2.7.4). Finally, analysis of publicly available data from UV-crosslinked RIP targeting hnRNPU verified that Firre directly and specifically binds to hnRNPU (Figure 3.2.7.4)[49,50]. Collectively, these results suggest that Firre interacts with hnRNPU, and that the RRD is both required and sufficient for this interaction.

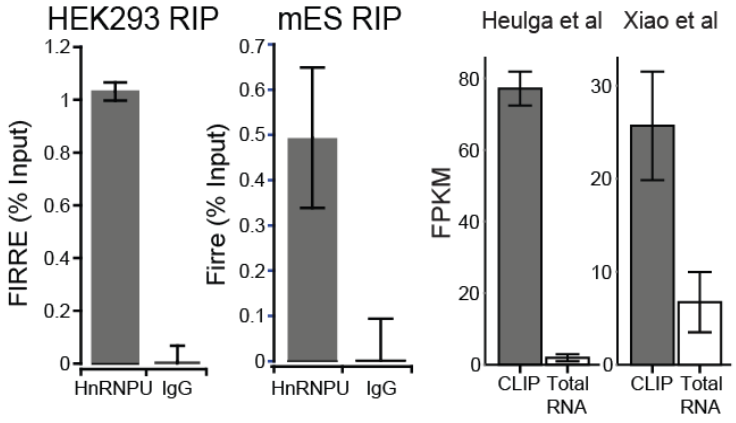


Figure 3.2.7.4: RIP with hnRNPU in HEK293s and mESCs. Shown as a percentage of input and publicly available hnRNPU CLIP data analyzed as gene level FPKM values.

3.2.8 hnRNPU is required for focal localization of Firre

Based on the requirement for RRD in establishing the proper localization of Firre and its interaction with hnRNPU, we investigated whether hnRNPU, in turn, regulates the spatial expression of Firre. Briefly, we transfected siRNAs targeting hnRNPU in mouse (mESC) and human (HEK293 and HeLa) cell lines and observed a >90% decrease in hnRNPU expression (Figure 3.2.8.1). We confirmed the previously described role of hnRNPU

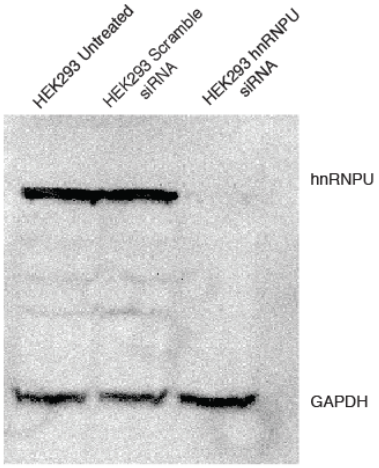


Figure 3.2.8.1: Knockdown of hnNRPU by siRNAs in HEK293.

in Xist localization in HEK293 cells (Appendix 6). Following the knockdown of hnRNPU, we repeated RNA-FISH targeting Firre as above. In all cell lines tested, we observed a strong delocalization of Firre and in several instances even translocation into the cytoplasm (Figure 3.2.8.2). Thus, both RRD and hnRNPU are required for the proper focal and nuclear localization of Firre and its retention in the nucleus.

3.2.9 hnRNPU is required for proximal trans-localization of Firre

Having found that hnRNPU regulates the specific localization of Firre, we next tested whether hnRNPU is required to maintain the proximal localization of the Firre locus and its trans-chromosomal localization sites. To this end, we repeated the RNA co-FISH between Firre and either Slc25a12 or Ypel4 upon siRNA-mediated depletion of hnRNPU in mESCs. In both cases, we observed a considerable decrease in co-localization of each *trans*-site with Firre in the absence of hnRNPU (Figure 3.2.9.1).

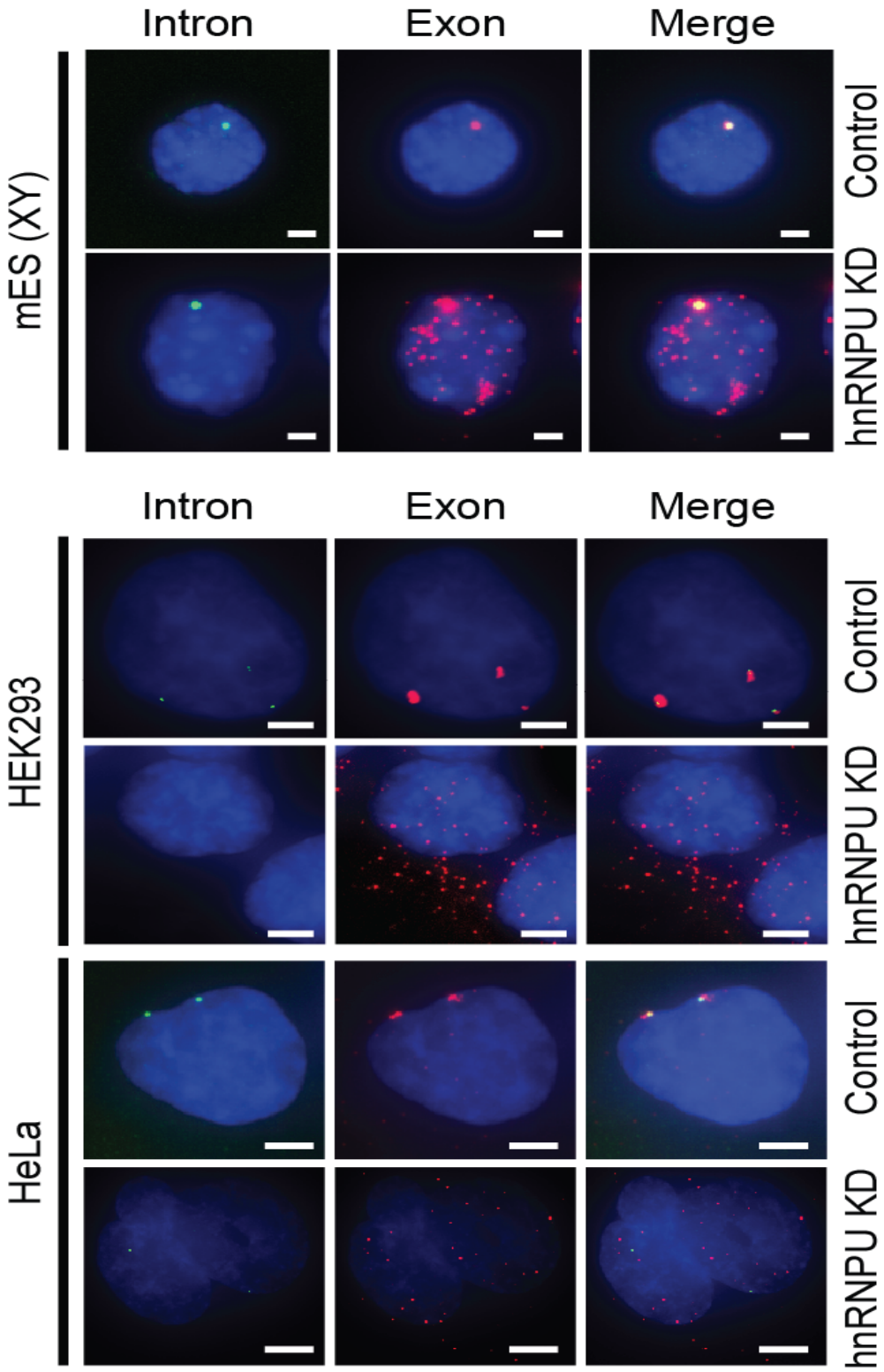


Figure 3.2.8.2: RNA-FISH targeting Firre in mESCs, HEK293s, and HeLas in the absence of hnRNPU. Introns ("green"); exons ("red"); nuclei (blue). Scale bars, mES: 20 μm; HEK293 and HeLa: 5 μm.

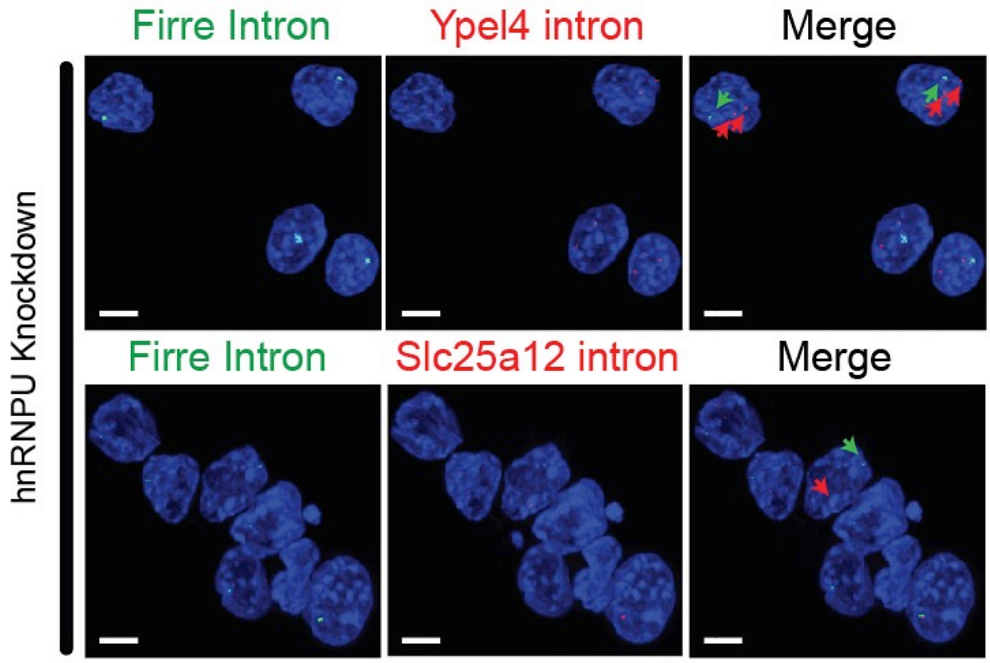


Figure 3.2.9.1: RNA-FISH co-localization of Firre in the absence of hnRNPU in male mESCs. Introns in "green" (A594) with Ypel4 or Slc25a12 introns in "red" (Cy3). Scale bar 40 μm, nuclei by DAPI.

3.2.10 Firre is required for trans-chromosomal co-localization

To test the functional contribution of Firre to trans-chromosomal co-localization, we repeated the co-FISH experiments between the trans sites in male *ΔFirre* mESCs. Strikingly, the Ppp1r10 and Ypel4 gene loci no longer co-localize in the absence of Firre (Figure 3.2.10.1) (15% co-localization in *ΔFirre* relative to 72% in wild-type), thus suggesting a requirement for the *Firre* gene locus facilitating the formation of cross-chromosomal interactions in mESCs. Furthermore, we do not observe co-localization of trans-sites in mLFs that do not express Firre (Figure 3.2.10.2). Collectively, the above results suggest a potential role for Firre, along with hnRNPU, in either maintaining or establishing higher-order nuclear architecture.

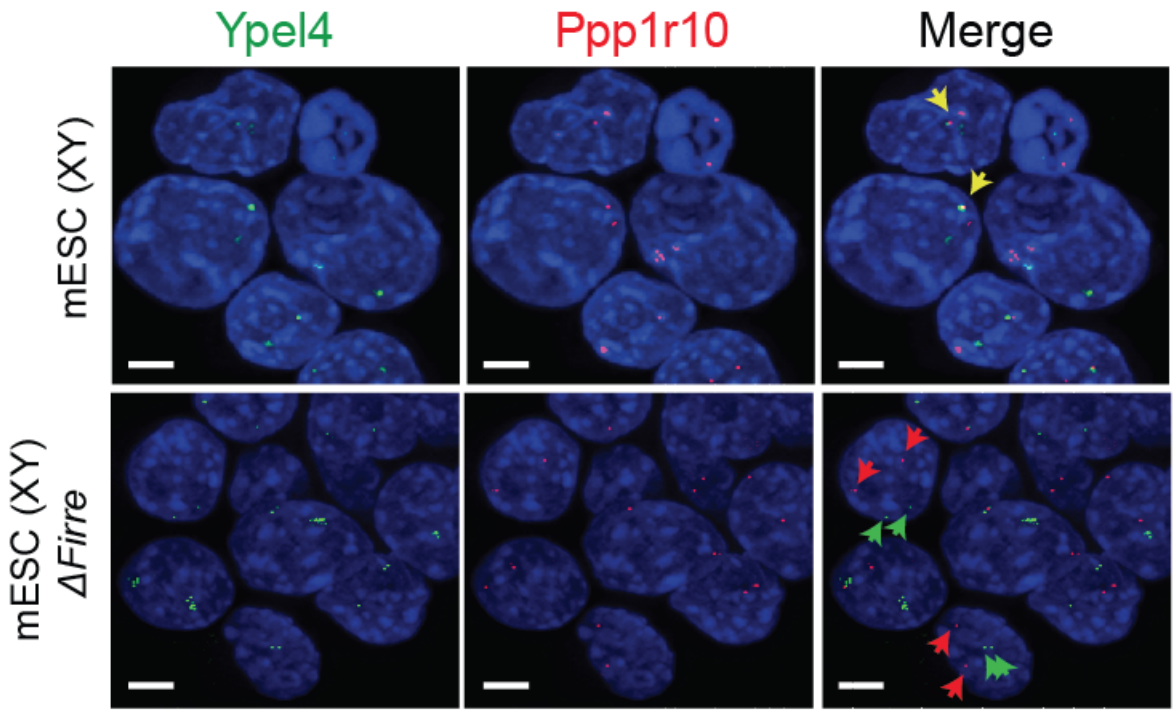


Figure 3.2.10.1: RNA-FISH co-localization of the trans-interacting loci Ypel4 and Ppp1r10. Ypel4 introns in "green" (A594) and Ppp1r10 introns in "red" (Cy3) in the absence of Firre expression in *ΔFirre* male mESCs compared to the wild-type mESCs Figure 3.2.9.1. Scale bar 40 μm, nuclei by DAPI.

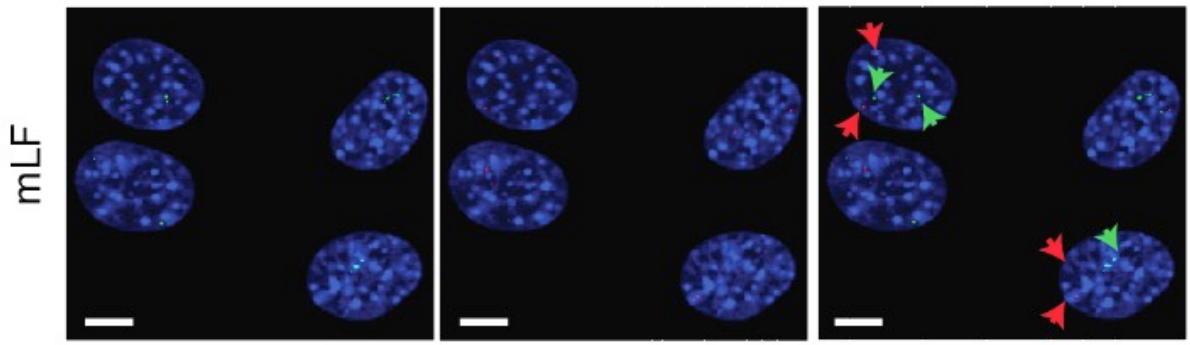


Figure 3.2.10.2: RNA-FISH co-localization of trans-sites in mLFs. Ypel4 introns in "green" (A594) and Ppp1r10 introns in "red" (Cy3). Scale bar, 40 μm., nuclei by DAPI.

This work was published: **Hacisuleyman, E**., L. A. Goff, C. Trapnell, A. Williams, J. Henao-Mejia, L. Sun, P. McClanahan, D. G. Hendrickson, M. Sauvageau, D. R. Kelley, M. Morse, J. Engreitz, E. S. Lander, M. Guttman, H. F. Lodish, R. Flavell, A. Raj, and J. L. Rinn. 2014. 'Topological organization of multichromosomal regions by the long intergenic noncoding RNA Firre', *Nat Struct Mol Biol*, 21: 198-206.

3.3 Discussion

Here, we have identified and characterized a novel ncRNA RNA gene, Firre, that has important roles in both cell physiology and nuclear architecture. This lncRNA escapes X chromosome inactivation and is required for proper cellular differentiation of adipocytes. Firre localizes across a 5Mb domain on both X chromosomes and makes *trans*-chromosomal contacts with several energy metabolism genes. Proper focal localization of Firre to the X chromosome requires a repetitive RNA domain that facilitates its interaction with nuclear matrix proteins, such as hnRNPU. Loss of either hnRNPU expression or the repetitive RNA domain results in mislocalization of Firre and misregulation of key genes. Taken together, these results suggest the requirement of a lncRNA to govern nuclear matrix interactions that in turn lead to proper transcriptional regulation.

The numerous properties of Firre shed new insights into: how RNA sequences can result in both *cis* and *trans* localization on chromatin, how these interactions between specific RNA sequences and nuclear matrix factors influence nuclear organization, and how these RNA-protein interaction properties summate to modulate properties of higher order nuclear architecture and the subsequent consequences to cell physiology.

Together these observations propose an intriguing model, in which Firre, and potentially many other lncRNAs[13,24,51,52], function as nuclear organization factors. One possibility is that lncRNAs can spread from their site of transcription to form interactions with nuclear matrix complexes and subsequently organize *trans*-chromosomal loci into local proximity. Specifically, Firre may serve to interface with and modulate the topological organization of multiple chromosomes (Figure 3.3.1). Consistent with this model, genetic deletion of *Firre,* results in a loss of nuclear proximity of several trans-chromosomal loci to the

Firre locus. Moreover, the proper localization of Firre requires both a specific 156 bp sequence and a physical interaction with hnRNPU to maintain the multi-chromosomal nuclear interactions. Thus, we propose that lncRNAs, through the interaction with nuclear matrix proteins, such as hnRNPU, might impart specificity in organizing a proper "zipcode" of chromosomal territories within the nucleus. For example, either the chromosomal binding of Firre or sequence specific interactions may serve as a cis localization signal in order to initiate the formation of or maintain sub-compartments within the nucleus.

Consistent with our model, two recent studies[19,20] demonstrate that the Xist RNA uses a "local proximity search" to guide its localization across large expanses of the X chromosome during X chromosome inactivation. Here, we broaden this phenomenon and show that these interactions are not merely restricted to a single chromosome but extend across multiple chromosomes in regional proximity. Several other observations in this study highlight potential gene regulatory roles of Firre, mediated by trans-chromosomal interactions. Intriguingly, we observe an array of CTCF binding sites across the *Firre* locus. CTCF has previously been shown to play a critical role in X chromosome pairing and counting[53] and interact with hnRNPU[54]. Similarly, this array of CTCF binding sites across the *Firre* locus might further facilitate inter-chromosomal interactions with the 5 Mb X chromosome localization domain. Finally, our study demonstrates that the formation of these cross-chromosomal interactions is altered upon genetic depletion of the *Firre* locus.

This model has several implications for potential new roles for lncRNA-mediated gene-regulation. For example, lncRNAs could bring genes involved in a similar biological process into close proximity allowing for co-regulation in space and time, serving as nuclear organization factors. This appears to be the case for Firre, where several genes involved in

energy metabolism and adipogenesis are organized together in spatial-proximity and are typically co-expressed, consistent with the previously described role of Firre in adipogenesis[31]. Underscoring the physiological relevance of such a model, either genetic deletion or transcriptional depletion of Firre results in the perturbation of cell physiology in both mESCs and adipocytes, respectively[31]. Future studies will require genetic studies in mouse models to further illuminate the role of Firre in mammalian development and disease.

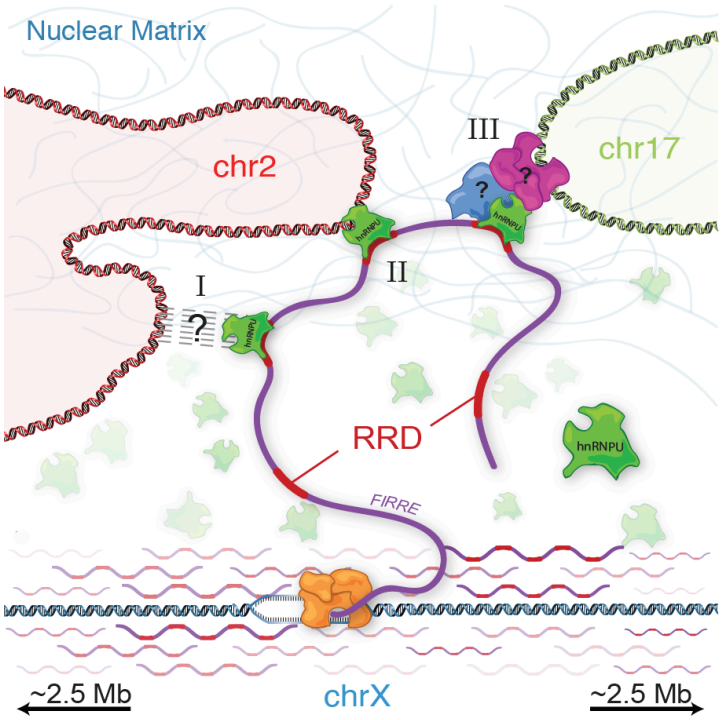


Figure 3.3.1: A model for Firre as a 'regional organization factor.' Firre transcripts accumulate at the site of their transcription. hnRNPU binds to the RRD of Firre and facilitates interactions with trans-chromosomal regions through one of several possible mechanisms: I) Tertiary interactions with nuclear matrix components, II) Direct binding of hnRNPU to matrix attachment regions *in trans*, or III) As yet undetermined interactions with other protein complexes to facilitate indirect binding to DNA.

3.4 Materials and Methods

All sequencing and related data is deposited in the Gene Expression Omnibus (GSE45157).

3.4.1 Repetitive sequence analysis (FSA)

The sequence for Firre was scanned for repetitive elements using the *ab initio* repeat detection algorithm RepeatScout[55]. This sequence was then aligned back to the genome (mm9 or hg19) using BLAT with the following parameters: "-stepSize=5 -repMatch=2253 -minScore=50 –minIdentity=0". Genomic DNA for the hits was extracted and multiply aligned with the Fast Statistical Aligner, a probabilistic multiple alignment tool specifically engineered to accommodate multiple alignment of sequences with potentially non-uniform evolutionary constraint. Fast Statistical Alignment uses pair hidden Markov models to estimate gap and substitution parameters for the multiple alignment scoring function, improving alignment robustness.

3.4.2 Cloning Firre

Total RNA (1 μg) was reverse transcribed following the instructions in the Superscript III kit (Life Technologies, #18080-051). The thermocycling conditions were: 25°C for 10 minutes, 55°C for 1 hour, 70°C for 15 minutes and 4°C final. 2 μL of the cDNA was mixed with 21 μL of water, 2 μL of 10 μM primers and 25 uL 2x Phusion Mastermix (New England Biolabs, # M0531S). The PCR conditions were: 1) 98°C for 30 seconds, 2) 98°C for 10 seconds, 3) 66°C for 30 seconds, 4) 72°C for 3 minutes, 5) 72°C for 5 minutes, 6) and 4°C final, with 45 cycles repeating steps 2-4. The extension time varied with the length of the lincRNA. The products were checked on 1% agarose gel. Nested PCR was performed when necessary using purified PCR products instead of cDNA. Longer isoforms were gel purified and then subjected to the following cleaning steps. The PCR products were purified using SPRI

beads (Beckman Coulter, #A63880), following the instructions in the manual. SPRI beads were added to the PCR product and incubated at room temperature for 2 minutes. The mix was put on a magnet for 4 minutes and the supernatant was removed. The beads were washed with 100 µL of 70% EtOH for 30 seconds twice and placed at 37°C for 5 minutes until the beads appeared dry. The PCR product immobilized on the beads was eluted with 30 µL of water on the magnet for 5 minutes.

The purified PCR product was quantified and used in BP reactions. The amount of DNA to be added was calculated as described in the Gateway cloning manual (Invitrogen). The BP reaction was set up according to the BP Clonase II instructions (Life Technologies, #11789020).

For transformations, 1 vial of Omnimax 2T1R (Life Technologies, #8540-03) cells were used for four BP reactions. The steps outlined in the Omnimax 2T1R manual were followed. The transformation plates were incubated at 37°C overnight, and the colonies were sequenced through Genewiz. When the inserts were verified, the plasmids were prepared using the Qiagen mini-prep kit (Qiagen, #27104).

3.4.3 In situ hybridization

The Firre probe was generated by PCR from adult brain cDNA and subcloned in pCRII-TOPO (Life Technologies, #K4610-20). The antisense riboprobe was generated by *in vitro* transcription using SP6 polymerase (Roche Applied Science, #10810274001) as previously described[56]. For non-radioactive *in situ* hybridizations, staged embryos were dissected in 1X PBS (Invitrogen) and fixed in 4% paraformaldehyde overnight at 4°C. For E14.5 cross-sections, embryos were washed overnight at 4°C in 30% sucrose/PBS followed by 1:1 ratio of 30% sucrose/OCT Clear Frozen Section Compound (VWR, #95057-838) for 1 hour. Embryos

were then placed in fresh OCT, frozen and stored at -80°C until sectioning. Frozen serial sections, 20 µm thick, were prepared using a HM550 cryostat (Thermo Scientific) and mounted onto Superfrost Plus slides (VWR #48311-703). Sections were permeabilized with 10 µg/ml proteinase K (Roche Diagnostics) for 10 minutes, washed with 1X PBS, treated 10 minutes in RIPA buffer and cross-linked again for 5 minutes in cold 4% paraformaldehyde. Sections were then pre-hybridized for 1 hour at room temperature at 70°C (50% formamide, 5X SSC, 5X Denhardts, 500 µg/ml Salmon sperm DNA, 250 µg/ml Yeast RNA) and then incubated overnight at 70°C in the same solution containing 2 µg/ml of DIG-labeled riboprobe. Sections were washed, blocked 1 hour with 10% sheep serum and incubated overnight at 4°C with 0.375 U/ml alkaline phosphatase-labeled anti-DIG antibody (Roche Diagnostics). Signal was detected by exposing sections to NBT-BCIP (Sigma #B1911), 0.1% Tween-20. Reaction was stopped with washes in 1X PBS supplemented with 0.1% Tween-20, and sections were mounted in Fluoromount-G (Southern Biotech #0100-01). The E14.5 whole embryo cross-section image (Fig. 1C) was taken from the Eurexpress Database (Assay ID #euxassay_013928, www.eurexpress.org). Non-radioactive *in situ* hybridizations of E15.5 embryo brains were performed on 40 µm vibratome sections (Leica) mounted on Superfrost Plus slides (VWR) using reported methods[57]. Sense probes were used as negative controls in all experiments. The PCR primers used to generate the Firre probe are: forward 5'-GAGAACCCATTGGAGGTTGA-3' and reverse 5'-CCCGTTCTTGTGCATCCT-3'. The Firre riboprobe sequence used for the hybridizations was:

5'-
CCCGUUCUUGUGCAUCCUCUCUGAAGACCCGGGAACCACAAGUAACAGCAUAGA
CAAUGACAAGCCUGCACUUCUUCUCUUCCUCAGGCAGCAGUGUUCCAGCUCCAG
UGAUUGCUCACGGUACCUGGUGGUCUUGGGAUCGCGGGGGCAUGUCUCAGCAUC
CAGUUCUGAGGCAUGUACACUCCUCAUGCAUCUUCUCUUUGGAUAAAGUCAGCA

UCCAGGCAAUGACGAAAUUCUGCAUUUCUGCCUUUCUCCAUCCACCCAGCGCUG
UUCAACCUCCAAUGGGUUCUC-3'.

3.4.4 Cellular fractionation

The cells grown in 15 cm dishes were washed with 5 mL of 1X PBS and trypsinized with 3 mL of TrypLE (Invitrogen) at 37°C for 3-5 minutes. The trypsin was quenched with 5 volumes of ice-cold growth media (DMEM (Invitrogen), 10% FBS (Invitrogen), 1% Pen-Strep (Invitrogen), and 1% L-Glutamine (Invitrogen)) and the cells were pelleted at 200xg for 3 minutes and resuspended in 1 ml of ice-cold 1X PBS. The resuspension was centrifuged at 200xg for 10 minutes at 4°C. The supernatant was carefully removed without disturbing the pellet; the remaining packed pellet volume was estimated for the next steps. The pellet was resuspended in 5 packed pellet volumes of ice-cold cytoplasmic extraction buffer (20 mM Tris, pH 7.6 (Ambion), 0.1 mM EDTA (Ambion), 2 mM MgCl2 (Ambion), 1X protease inhibitors (VWR), 0.5 U/μL RNaseOUT (Invitrogen)). The cells were incubated first at room temperature for 2 minutes then on ice for 10 minutes. The cells were lysed by adding CHAPS to a final concentration of 0.6%. The sample was then homogenized by passing it through a 1 ml syringe and centrifuged at 500x for 5 minutes at 4°C. The 70-80% of the supernatant was taken and saved at -80°C; this was the cytoplasmic fraction. The remaining supernatant was carefully removed and the pellet was washed with cytoplasmic extraction buffer supplemented with 0.6% (w/v) CHAPS. The sample was centrifuged at 500xg for 5 minutes at 4°C, and the entire supernatant was discarded. The wash step was repeated one more time. The pellet was then resuspended in 2 packed pellet volumes of nuclei suspension buffer (10 mM Tris, pH 7.5, 150 mM NaCl, 0.15% (v/v) NP–40, 1X protease inhibitors, 0.5 U/μL RNaseOUT). The nuclear suspension was layered on 5 packed pellet volumes of sucrose cushion (10 mM Tris, pH 7.5, 150 mM NaCl, 24% (w/v) Sucrose, 1X protease inhibitors, 0.5 U/μL RNaseOUT) and pelleted

at 14,000 rpm for 10 minutes at 4°C. The supernatant was discarded and the pellet was washed with 10 packed pellet volumes of ice-cold 1X PBS supplemented with 1 mM EDTA. The sample was then centrifuged at 500xg for 5 minutes at 4°C. The pellet constituted the nuclear fraction.

3.4.5 Fluorescence In Situ Hybridization (FISH)

The FISH protocol was followed as described previously[34]. Briefly, oligonucleotide probes targeting and tiling the intron of Firre were conjugated to Alexa594 fluorophores, and the probes targeting and tiling the exon were conjugated to tetramethylrhodamine (TMR) and HPLC purified. Before the hybridization, the adherent cells were fixed (10 minutes with 4% formaldehyde) and permeabilized with 70% EtOH in two-chamber coverglasses. mESCs and hESCs were fixed in solution after they were collected from the plate: the cells were incubated at room temperature in 2% formaldehyde solution for 10 minutes, followed by centrifuging at 1000xg for 3 minutes. The cells were washed with 1X PBS twice with centrifuging at 1000xg for 3 minutes in between. The cells were permeabilized with 70% EtOH. The ESCs were then plated on gelatinized coverglasses. Prior to the hybridization, the cells were rehydrated with the wash buffer containing 10% formamide (Ambion, #AM9342) and 2X SSC (Ambion, #AM9765) for 5 minutes. Then the probes (0.5 ng/μL final) were hybridized in 10% dextran sulfate (Sigma, #D8906), 10% formamide, and 2x SSC at 37°C overnight. After hybridization, the cells were washed in wash buffer at 37°C for 30 minutes twice (with the addition of DAPI in the second wash), followed by washing with 2X SSC twice. The imaging was done immediately after using 2X SSC as the mounting medium.

The same protocol was followed for the co-FISH experiments. The probes targeting and tiling the introns of the trans sites (Slc25a12, Ypel4, Ppp1r10) were conjugated to Quasar570.

Co-FISH assays were conducted as indicated in wild type male mESC, *ΔFirre* male mESC, or mLF. Quasar670 was used as an additional fluorophore when working with three colors and trans sites: Quasar570 and Quasar670 were used together when staining for trans sites.

3.4.6 Actinomycin-D treatment

Actinomycin-D (Act-D) (Sigma, # A9415-2MG) was resuspended in DMSO with a final concentration of 2 mg/mL. Act-D was thoroughly mixed with the 2I media (2 μg/ml) and added on the male mESCs at 0, 1.5, 3, and 6 hours.

3.4.7 RNA Antisense Purification (RAP) Analysis

RAP was performed as described[19]. Briefly, the RNA of interest is tiled with 120 bp antisense nucleotides that have been biotinylated. Two distinct pools of antisense probes targeting Firre and one pool containing sense probes (negative control) were generated. The hybridization was done in duplicate crosslinked and precleared lysates with 20 ng (350 fmol) of oligos. The oligos were then captured by streptavidin beads and, the elutions for RNA and DNA were performed. Consistent with standard ChIP-Seq assays duplicate pull-downs were performed and sequenced to control for technical variability.

For X-chromosome enrichment analysis, the X-chromosome was divided into 10Kb bins and a linear regression of counts per bin was performed against each replicate and the input control. The slope of the linear regression was used as a normalization factor (alpha) between the two libraries. Enrichment levels relative to input were calculated by dividing the experimental counts for each bin by the input counts times alpha.

To identify significant regions bound by the Firre RNA *in trans*, we used the Scripture peak-calling algorithm to call significant peaks across each of the replicate sequencing bam files, including the input control and anti-sense control. All peaks were merged using the

Bedtools mergeBed[54] program to obtain the universe of significant peaks across all samples. A .gtf file of significant peaks, along with the replicate .bam files for each of the samples was used as input for Cuffdiff2 for quantification and differential testing. Cuffdiff2 was run using default parameters with the addition of the '--no-length-correction' argument to disable length correction. Significant peaks were called using the Cuffdiff2 test-statistic with $p<0.1$.

3.4.8 Retroviral Overexpression of Firre

The overexpression vector for overexpressing Firre was made by modifying the pLenti6.3/TO/V5-DEST (Snap Gene) destination vector. We modified by removing the WPRE, the SV40 promoter and the blastacidin resistance gene, keeping the gateway tails the same, to prevent any interference with the lincRNA structure and function. All the transductions were done as follows: the cells were split into 12-well dishes and resuspended in media with 4 μg/mL polybrene. Immediately after, 100 μL of virus (of the same titer; if not the same titer, the volume was adjusted accordingly) was added to each well. The untransduced control was used to measure the overexpression levels by qRT-PCR. The sequences of the isoforms 1.1 (+RRD), 1.4 (+RRD), 1.6 (-RRD) used for mouse transductions are in Supplementary Table 1 (bed file) and the sequence of the human isoform used for HEK293 and HeLa transductions is:

GCTTGATGAGGGCATGGATCACTAAGGTCTGTTCCCAATACAAGAAGACTCTTTGA
CATCATAATAAAATACTGCAGATACGATGCTGAGTGAAAAAGAGTAGAAATGGGA
AGACTTGGTTGTGCAGAAACTGAGTTCTTAAAGAGAGGAGATACTTTATGAGGGCT
GGAGTGCACTGGAGCAATCCTGGCTGGCCAATACTGAGTTCTTGAAAACAGGAGA
TGCTTGATGAGGATGGGATCATCTAGTTGCAGGAAAACAAGGCTCAGGGTGCCTA
CTGATTCTACATTATGCTTGGTCCGGAGAGCTGCCTTGGACTCTCCAATGCAGCTGC
CTCTGTTACCTTGACTCGTCTCAGATTTCATTGACCCAAGATGGGTCCTGGCACTAG
GAATGTAAGACTGTCTCCATTCTTTTGGCTTGTTCCAGAGAAGCCCATGCAAGGTT
CTTACTGACCATATGTTTCTTTCTTTTCTTTTGTTTTTTGAGACGGAGTTTCGCTGTT
GTTGCCCAGGCTGGAGTGCAATGGCGCGATCTCCGCTCACTGCAACATCCGCCTCC
CGGACTCAAGCGATTCTCCTGCCTCAGCCTCCCGAGTAGCTGGGATTACAGGCATG
CGCCACCACGCCCGGCTAATTTTGTATTTCTAGTAGAGTTAGAGTTTCTCCATATTG
GTCAGGCTGGTTTCGAACTCCTGGCCTCAGGTGATCTGCCCTCCTTGGCCTCCCAAA

GCGCTGGGATTATAGGCGTGAGCCACCGCCCCCGGCCTTGACCATATGTTTTATTT
CTAGCTTTGATGTCTGGGCATCGATTTCCCTCGGTTAAACTATTTGCTCAATGTTAA
GGCCACTCTGTAGAAATTTGTCTGTGTAACTGAGGTGCTATGCAGGCCTGCCTGTG
TGACTGTCATGCAGGCCTGTCTGTGTGATTGTCAGGGAGAATTGGTCTGCCACAAT
CCTTTTCTAAGCATAGCCAATAGAGGTAGTTAGGCATAATTTGTATATTACAGAAA
TTGCCTTACAGAGAGTAACACATTTCTATACTCTCCTTCCATAACAGACACTTAAA
AAAACAAAAAAAGTAATGTATGCTTGCTGTGGACCTCATTTAAGATTGCAACAGA
AGCACTTTTCAATACTATTAAACAGCTTTTTACTTTCC.

3.4.9 Cell Culture

HEK293 (ATCC: CRL-1573), HeLa (ATCC: CCL-2), mLF (ATCC: CCL-206) and hLF (ATCC: IMR90) cells were grown in growth media (see above) at 37°C at 5% CO_2. Male (Novus: NBP1-41162) and female (RIKEN: AES0010) mouse ES cells were grown in previously gelatinized (0.2%) dishes with 2I media containing 125 mL DMEM/F12 (Invitrogen), 83.5 µL BSA fraction V (50 µg/ml relative to DMEM) (Invitrogen, 15260-037, 75mg/ml), 125 mL Neurobasal medium (Invitrogen, 21103-049), 625 µL of the Ndiff Neuro2 (200X, relative to Neurobasal medium) (Millipore, SCM012), 2.5 mL B27 minus vitamin A (50X, relative to Neurobasal medium) (Invitrogen, 12587-010), 2 µL beta-mercaptoethanol, 1 µM PD0325901 (Stemgent, 04-0006), 3 µM CHIR99021 (Stemgent 04-0004), 25 µL LIF ESGRO (from Chemicon, ESG1106), 1% pen-strep (Invitrogen, 15140-163), 1% Non-Essential Amino Acids (Invitrogen, 11140-076), 1% L-glutamine (Invitrogen, 25030-164). The plating density of mES cells was chosen to be 30,000-50,000/cm^2. Adipocytes were grown as described previously [28].

3.4.10 Pull-down with in vitro biotinylated RNA

The cloned in transcript in pdest14 plasmid vector was linearized with Nhe1. Phenol-chloroform extracted and ethanol precipitated template was then used in *in vitro* transcription, which included: 20 ug/ml DNA template, 40 mM Tris (pH 7.9), 2.5 mM Spermidine, 26 mM $MgCl_2$, 0.01% Triton X-100, 8 mM GTP, 5 mM ATP, 5 mM CTP, 1.3 mM UTP, 0.7 mM Bio-

16-UTP (Epicentre), 5 mM DTT, 20 mM $MgCl_2$, RNAseOut 80 U/ml, 20 U of T7 RNA polymerase (Life Technologies, #18033-019). The mix was incubated at 37°C until a white precipitate formed. After the reaction reached completion, 60 mM-final EDTA was added to dissolve the precipitate. Due to the biotin, RNA will partition into the organic layer if it is phenol-extracted. Therefore, the *in vitro* transcriptions were first treated with DNase (Worthington, #LS006353) (37°C for 10 minutes followed by EDTA addition and 75°C for 10 minutes), then cleaned with the Bio-Spin 30 columns (BIO-RAD, # 732-6231).

For end labeling, the *in vitro* transcribed RNA (without labeled UTPs) was treated with the 3' end biotinylation kit of Thermo Scientific Pierce (#20160).

The lysate was prepared by lysing 15 cm dishes into 1 ml of lysis buffer (150 mM KCl, 25 mM TRIS- HCl pH 7.4, 5mM EDTA, 5 mM $MgCl_2$, 1% NP-40, 1X protease inhibitor, 0.5 mM DTT, 100 U/ml RNAseOut) for 30 minutes at 4°C. The lysate was centrifuged at 13,000 rpm for 30 minutes and filtered with 0.45 um filter. The concentration of the lysate was measured by BCA protein assay (Thermo Scientific, # 23225).

For the pull-down, 1.5 mg of the lysate was initially pre-cleared with the Magnetic MyOne Streptavidin T1 beads (Life Technologies, #65601) for 30 minutes at 4°C. The beads were prepared as described in the manual. The precleared lysate was 2X diluted and supplied with 0.1 μg/μl tRNA, to which 30 pmoles of biotinylated RNA was added. The RNA was incubated in the lysate for 2 hours at 4°C rocking, after which 40 μl of MyOne Streptavidin T1 beads was added to the mix. The mix was incubated for another hour at 4°C. The beads were washed for 3 times (10 minutes each) with 1 ml of the wash buffer (lysis buffer but with 300 mM KCl) on a magnetic rack. Finally, the beads were resuspended in 30 μl of sample buffer (4X, Biorad) and reducing agent (20X, Biorad) and boiled for 5 minutes at 95°C. The samples

were then run on 4-12% gradient Bis-Tris gels and stained with Sypro Red as described (Life Technologies, #S-12000) for protein detection and mass spectrometry or were transferred to a PVDF membrane for Western blotting.

The mass spectrometry analysis was done as follows: The eluates from RNA pull-downs done in three different cellular contexts (mouse adipose tissue, mouse adipocyte, and mESC lysates) using five different RRD+ isoforms and one RRD– isoform were run on a gel as described. The bands that are differential between RRD+ and RRD– isoforms were cut and processed for mass spectrometry. To identify proteins that preferentially co-precipitated with Firre in an RRD-dependent manner, we took the difference between the peptide counts of the RRD+ and RRD– isoforms. We repeated the differential analysis for each of the RRD+ isoforms and took the top 10% of the differential peptide count scores for each isoform identified in both mESC and adipocyte lysates.

The transfer for Western blotting was done in transfer buffer that was prepared in the following ratios: 100 ml of 10X TG (Biorad), 200 ml methanol and 700 ml ddH_2O. The membrane was activated in methanol first and equilibrated in transfer buffer prior to transfer. The transfer was done at 70 watts for 1 hour. After the transfer, the membrane was washed with methanol for blocking and incubated with the primary antibody diluted in 1:1000 in 0.1% Tween, 1% non-fat milk, and 1X PBS for 4-5 hours at RT or 4°C O/N. Following the primary antibody incubation (human hnRNPU (3G6): SantaCruz sc-32315 (reactivity with the species is validated and shown on the Santa Cruz website), mouse hnRNPU: Abcam ab20666 (validated by [55]), the membrane was washed 3 times with wash buffer (1X PBS supplemented with 0.1% Tween) and then incubated with the secondary antibody for 1 hour at RT. The membrane was washed again 3 times with wash buffer and then developed using SuperSignal West

Chemiluminescent Thermo Scientific reagents. 20% of the input lysate was used for all the Western blots (Fig. 6A,B, Supplementary Fig. 5D).

3.4.11 Endogenous RNA pull-down

The 23-25 bp oligos (Human RRD: TCCAGTGCTTGCTCCTGATGTCTC; Human Exon:CTAAAACAACGAGGACGCACCTGAGA;
Mouse:TCAAGCCCCGAAAGAAACTGGAAC,
Scramble: GCTCCCATACATTTCTTCGGCTCTTA)
were synthesized with 18S linker and desthiobiotin at the 5' end through IDT. The same protocol for RNA-pull downs above was followed, except instead of adding the biotinylated RNA, the DNA oligos were added to the lysate and an annealing step was followed. Annealing was done by incubating the lysate and the oligos at 37°C for 15 minutes, at room temperature for 15 minutes, and at 4°C for 6 hours-O/N. The rest of the steps were the same as above except for these: 1) during the incubation with the beads, heparin was spiked in at 0.5 μg/μL in the last half hour, 2) in the first wash 0.5 μg/μL heparin was spiked in again, and 3) instead of boiling the beads, the RNA and protein were eluted from the beads with 12.5 mM Biotin for 30 minutes at RT and 3 hours at 4°C. The 60% of the elution was used to extract RNA and 40% to run a protein gel.

3.4.12 RNA Immunoprecipitation

The protein lysate (1.5 mg) was incubated with 6-8 μg of the hnRNPU or IgG antibody (hnRNPU (3G6): Santa Cruz sc-32315, hnRNPU (H-94): Santa Cruz sc-25374 (validation shown on the Santa Cruz website), Anti-hnRNPU: Abcam ab20666, Mouse (G3A1) IgG1: Cell Signaling 5415) at 4°C for 2-3 hours. Then 45 μL of protein G Dynabeads (Life Technologies, #10003D) that were previously washed twice in 500 μL lysis buffer were added to the lysate and antibody mix. The lysate, antibody, and beads were incubated at 4°C for another 2 hours. The beads were washed 3 times (10 minutes each) with 1 mL of lysis buffer. The RNA was

extracted by adding 1 mL of TRIzol (Life Technologies, #15596-018) to the beads. For the total input RNA, 10% of the input lysate was mixed with 1 mL of TRIzol. For 1 ml of TRIzol, 200 µL of chloroform was added, and the mix was centrifuged at 4°C at 13,000 rpm for 15 minutes. The aqueous layer was then added 1 volume of isopropanol, 1/10 volume KOAc, and 1 µL of glycoblue and kept at -20°C for at least one hour. The samples were then centrifuged at 13,000 rpm at 4°C for 30 minutes. The supernatant was removed and the pellet was washed with 1 mL of ice-cold 70% EtOH twice (centrifuging 2 minutes each time at 4°C). The pellet was then resuspended in 15 µL of RNase-free water.

3.4.13 CLIP-Seq Analysis

We analyzed two human hnRNPU CLIP-Seq datasets generated independently by Huelga et al. (GSE34993)[49] and Xiao et al. (GSE34491)[50]. We downloaded fastq files and aligned using TopHat to hg19 and a custom transcriptome GTF consisting of UCSC coding genes, a recently published lncRNA catalog[58], and cloned Firre isoforms. We used the RNA-Seq differential expression software Cuffdiff to estimate read counts to Firre in all CLIP and total RNA datasets. We performed a Poisson-based statistical test for enrichment of aligned reads in the CLIP versus total RNA.

3.4.14 RNAi-mediated Knockdown of hnRNPU

mESCs were transfected by the reverse transfection method in 6-well plate format. Briefly, the lipofectamine RNAiMAX (6 µl/well, Life Technologies. # 13778030) and siRNA (50 nM final, Dharmacon On-Targetplus smart pool for mouse (L-051574-01-0005) and for human (L-013501-00-0005)) complexes were prepared in 400 µl of Opti-MEM and incubated at room temperature for 20-30 minutes, during which the mESCs were prepared for splitting. The split was done as follows: the cells were washed with 1X PBS and trypsinized for 3

minutes. The trypsin was quenched by 2I media; the cells were centrifuged for 5 minutes at 850 rpm at 4°C, resuspended in new 2I media, and counted. Approximately 280,000 cells were plated for each well of a 6-well plate. Immediately after, the Lipofectamine/oligo complexes were added to the wells. The media was changed after 24 hours and the cells were harvested after 96 hours for the complete knockdown of the protein (checked by qRT-PCR and by Western blot). The same protocol was used for HEK293s and HeLa cells; however, they were plated at a density of ~21,000 cells/cm^2.

3.4.15 RNA Extraction

RNA extraction was performed by adding 1 ml of TRIzol to each well of a 6-well plate. 200 µL of chloroform was added, and the mix was centrifuged at 4°C at 13,000 rpm for 15 minutes. The aqueous layer was processed on the RNeasy Mini columns (Qiagen, #74104). The RNA was reverse transcribed using the SuperScriptIII First-Strand Synthesis kit. The cDNA synthesis was performed at 25°C for 5 minutes, 50°C for 1 hour, and 70°C for 15 minutes. Then the cDNA (15 ng per well of the 384 qPCR plate) was added 1:1 to the SYBR and primer mix (100 nM) for qRT-PCR. All the primers used in qRT-PCR are shown in Supplementary Table 4.

3.4.16 RNA-Seq library preparation, sequencing, and analysis

200ng of extracted RNA from each of 2 Δ*Firre* and 3 wild-type JM8A male mouse ES cell cultures was used as input for the Illumina TruSeq library preparation kit, using manufacturer's guidelines. Libraries were individually barcoded and library size distribution and quality were assayed using a DNA High-sensitivity Chip on the Agilent Bioanaylzer 2100. Libraries were pooled and paired-end 35bp fragments were generated on an Illumina MiSeq sequencer to an average depth of 9 Million fragments per sample. Fragments were aligned to

the mouse genome (mm9) using Tophat2 with default options and the UCSC transcriptome as a reference. Aligned reads were quantified against all mouse UCSC genes using Cuffdiff2 with default options. Significantly differentially expressed (DE) genes were selected with an FDR of 5%. Circular representation of DE gene projections onto mm9 (Figure 5f) was generated using the Circos utility (http://circos.ca/). Since RNA-Seq data are heteroscedastic not normally distributed, we chose to evaluate pathway enrichment by using a preranked GSEA analysis. This was conducted by using the GSEA tool[59] with a list of all genes ranked by cuffdiff2 test statistic (KO/WT), against the c2.cp.reactome.v4.0.symbols.gmt gene set collection (MSigDB, Broad). Gene sets were selected as significantly enriched if the nominal p-value was less than 0.01.

3.4.17 Targeting and generation of conditional ΔFirre mESC

To generate ESCs specifically deficient in Firre, a two-step targeting strategy was used to introduce loxP sites in the 5' and 3' ends of the *Firre* locus. Targeting of only one allele was needed to obtain *Firre*-deficient ESCs as *Firre* resides on the X-chromosome and male ESCs (JM8) were used. To generate the *Firre* 3' targeting construct the following primers were used to amplify the homology arms, which were then cloned into the pEASY-FLIRT vector:

5' homology arm: gcggccgcCTATGGGTGCTCAAGTGGTTGCAG and gcggccgcCACTGGATCTTAAAGAGACATTTTC

3' homology arm: ggcgcgccGTAGGCAAGCCTGAGGAAAATTTC and ggcgcgccCGAGCTCTGGGGGTACTGGTTAGT.

In this targeting construct, the neo cassette that serves as a selection marker during the targeting process was flanked by two Frt sites. To generate the *Firre* 5' targeting construct the

following primers were used to amplify the homology arms, which were then cloned into the newly generated pEASY-Hygro vector:

5' homology arm: gtcgacCACCCAAACTGTGCAATTTTTA and gtcgacGTAGTTAAGAGCTCCTGTTGC

3' homology arm: gcggccgcACAGCCTTGGCTGAGATGTT and gcggccgcTGAGTCATCTTTCTGGCTTCAA.

In this construct, the hygro cassette that serves as a selection marker during the targeting process was flanked by two Loxp sites.

ES cells were maintained under standard conditions and targeted as previously described[37]. In brief, the vector targeting the 3' end of *Firre* was electroporated into C57BL/6 ES cells (JM8) and grown under selection with neomycin. Homologous recombinant ES cells identified by PCR analysis were subsequently electroporated with the vector targeting the 5' end of *Firre* and grown under selection with hygromycin. Double-targeted ES cells were identified by PCR analysis. To delete *Firre*, double-targeted ES cells were electroporated with a Cre recombinase expressing plasmid (pGK-Cre-bPA). PCR genotyping was used to identify clones, in which *Firre* had been deleted.

3.5 References

1 Huarte, M., and J. L. Rinn. 2010. 'Large non-coding RNAs: missing links in cancer?', *Hum Mol Genet*, 19: R152-61.

2 Rinn, J. L., and H. Y. Chang. 2012. 'Genome regulation by long noncoding RNAs', *Annu Rev Biochem*, 81: 145-66.

3 Carninci, P., T. Kasukawa, S. Katayama, J. Gough, M. C. Frith, N. Maeda, R. Oyama, T. Ravasi, B. Lenhard, C. Wells, R. Kodzius, K. Shimokawa, V. B. Bajic, S. E. Brenner, S. Batalov, A. R. Forrest, M. Zavolan, M. J. Davis, L. G. Wilming, V. Aidinis, J. E. Allen, A. Ambesi-Impiombato, R. Apweiler, R. N. Aturaliya, T. L. Bailey, M. Bansal, L. Baxter, K. W. Beisel, T. Bersano, H. Bono, A. M. Chalk, K. P. Chiu, V. Choudhary, A. Christoffels, D. R. Clutterbuck, M. L. Crowe, E. Dalla, B. P. Dalrymple, B. de Bono, G. Della Gatta, D. di Bernardo, T. Down, P. Engstrom, M. Fagiolini, G. Faulkner, C. F. Fletcher, T. Fukushima, M. Furuno, S. Futaki, M. Gariboldi, P. Georgii-Hemming, T. R. Gingeras, T. Gojobori, R. E. Green, S. Gustincich, M. Harbers, Y. Hayashi, T. K. Hensch, N. Hirokawa, D. Hill, L. Huminiecki, M. Iacono, K. Ikeo, A. Iwama, T. Ishikawa, M. Jakt, A. Kanapin, M. Katoh, Y. Kawasawa, J. Kelso, H. Kitamura, H. Kitano, G. Kollias, S. P. Krishnan, A. Kruger, S. K. Kummerfeld, I. V. Kurochkin, L. F. Lareau, D. Lazarevic, L. Lipovich, J. Liu, S. Liuni, S. McWilliam, M. Madan Babu, M. Madera, L. Marchionni, H. Matsuda, S. Matsuzawa, H. Miki, F. Mignone, S. Miyake, K. Morris, S. Mottagui-Tabar, N. Mulder, N. Nakano, H. Nakauchi, P. Ng, R. Nilsson, S. Nishiguchi, S. Nishikawa, F. Nori, O. Ohara, Y. Okazaki, V. Orlando, K. C. Pang, W. J. Pavan, G. Pavesi, G. Pesole, N. Petrovsky, S. Piazza, J. Reed, J. F. Reid, B. Z. Ring, M. Ringwald, B. Rost, Y. Ruan, S. L. Salzberg, A. Sandelin, C. Schneider, C. Schonbach, K. Sekiguchi, C. A. Semple, S. Seno, L. Sessa, Y. Sheng, Y. Shibata, H. Shimada, K. Shimada, D. Silva, B. Sinclair, S. Sperling, E. Stupka, K. Sugiura, R. Sultana, Y. Takenaka, K. Taki, K. Tammoja, S. L. Tan, S. Tang, M. S. Taylor, J. Tegner, S. A. Teichmann, H. R. Ueda, E. van Nimwegen, R. Verardo, C. L. Wei, K. Yagi, H. Yamanishi, E. Zabarovsky, S. Zhu, A. Zimmer, W. Hide, C. Bult, S. M. Grimmond, R. D. Teasdale, E. T. Liu, V. Brusic, J. Quackenbush, C. Wahlestedt, J. S. Mattick, D. A. Hume, C. Kai, D. Sasaki, Y. Tomaru, S. Fukuda, M. Kanamori-Katayama, M. Suzuki, J. Aoki, T. Arakawa, J. Iida, K. Imamura, M. Itoh, T. Kato, H. Kawaji, N. Kawagashira, T. Kawashima, M. Kojima, S. Kondo, H. Konno, K. Nakano, N. Ninomiya, T. Nishio, M. Okada, C. Plessy, K. Shibata, T. Shiraki, S. Suzuki, M. Tagami, K. Waki, A. Watahiki, Y. Okamura-Oho, H. Suzuki, J. Kawai, Y. Hayashizaki, Fantom Consortium, Riken Genome Exploration Research Group, and Group Genome Science. 2005. 'The transcriptional landscape of the mammalian genome', *Science*, 309: 1559-63.

4 Derrien, T., R. Johnson, G. Bussotti, A. Tanzer, S. Djebali, H. Tilgner, G. Guernec, D. Martin, A. Merkel, D. G. Knowles, J. Lagarde, L. Veeravalli, X. Ruan, Y. Ruan, T. Lassmann, P. Carninci, J. B. Brown, L. Lipovich, J. M. Gonzalez, M. Thomas, C. A. Davis, R. Shiekhattar, T. R. Gingeras, T. J. Hubbard, C. Notredame, J. Harrow, and R.

Guigo. 2012. 'The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression', *Genome Res*, 22: 1775-89.

5 Batista, P. J., and H. Y. Chang. 2013. 'Long noncoding RNAs: cellular address codes in development and disease', *Cell*, 152: 1298-307.

6 Guttman, M., and J. L. Rinn. 2012. 'Modular regulatory principles of large non-coding RNAs', *Nature*, 482: 339-46.

7 Clemson, C. M., J. N. Hutchinson, S. A. Sara, A. W. Ensminger, A. H. Fox, A. Chess, and J. B. Lawrence. 2009. 'An architectural role for a nuclear noncoding RNA: NEAT1 RNA is essential for the structure of paraspeckles', *Mol Cell*, 33: 717-26.

8 Maass, P. G., A. Rump, H. Schulz, S. Stricker, L. Schulze, K. Platzer, A. Aydin, S. Tinschert, M. B. Goldring, F. C. Luft, and S. Bahring. 2012. 'A misplaced lncRNA causes brachydactyly in humans', *J Clin Invest*, 122: 3990-4002.

9 Nie, L., H. J. Wu, J. M. Hsu, S. S. Chang, A. M. Labaff, C. W. Li, Y. Wang, J. L. Hsu, and M. C. Hung. 2012. 'Long non-coding RNAs: versatile master regulators of gene expression and crucial players in cancer', *Am J Transl Res*, 4: 127-50.

10 Pandey, R. R., T. Mondal, F. Mohammad, S. Enroth, L. Redrup, J. Komorowski, T. Nagano, D. Mancini-Dinardo, and C. Kanduri. 2008. 'Kcnq1ot1 antisense noncoding RNA mediates lineage-specific transcriptional silencing through chromatin-level regulation', *Mol Cell*, 32: 232-46.

11 Rinn, J. L., M. Kertesz, J. K. Wang, S. L. Squazzo, X. Xu, S. A. Brugmann, L. H. Goodnough, J. A. Helms, P. J. Farnham, E. Segal, and H. Y. Chang. 2007. 'Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs', *Cell*, 129: 1311-23.

12 Tsai, M. C., O. Manor, Y. Wan, N. Mosammaparast, J. K. Wang, F. Lan, Y. Shi, E. Segal, and H. Y. Chang. 2010. 'Long noncoding RNA as modular scaffold of histone modification complexes', *Science*, 329: 689-93.

13 Vallot, C., C. Huret, Y. Lesecque, A. Resch, N. Oudrhiri, A. Bennaceur-Griscelli, L. Duret, and C. Rougeulle. 2013. 'XACT, a long noncoding transcript coating the active X chromosome in human pluripotent cells', *Nat Genet*, 45: 239-41.

14 Zhao, J., B. K. Sun, J. A. Erwin, J. J. Song, and J. T. Lee. 2008. 'Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome', *Science*, 322: 750-6.

15 Kaneko, S., J. Son, S. S. Shen, D. Reinberg, and R. Bonasio. 2013. 'PRC2 binds active promoters and contacts nascent RNAs in embryonic stem cells', *Nat Struct Mol Biol*, 20: 1258-64.

16 Davidovich, C., L. Zheng, K. J. Goodrich, and T. R. Cech. 2013. 'Promiscuous RNA binding by Polycomb repressive complex 2', *Nat Struct Mol Biol*, 20: 1250-7.

17 Jeon, Y., K. Sarma, and J. T. Lee. 2012. 'New and Xisting regulatory mechanisms of X chromosome inactivation', *Curr Opin Genet Dev*, 22: 62-71.

18 Plath, K., S. Mlynarczyk-Evans, D. A. Nusinow, and B. Panning. 2002. 'Xist RNA and the mechanism of X chromosome inactivation', *Annu Rev Genet*, 36: 233-78.

19 Engreitz, J. M., A. Pandya-Jones, P. McDonel, A. Shishkin, K. Sirokman, C. Surka, S. Kadri, J. Xing, A. Goren, E. S. Lander, K. Plath, and M. Guttman. 2013. 'The Xist lncRNA exploits three-dimensional genome architecture to spread across the X chromosome', *Science*, 341: 1237973.

20 Simon, M. D., S. F. Pinter, R. Fang, K. Sarma, M. Rutenberg-Schoenberg, S. K. Bowman, B. A. Kesner, V. K. Maier, R. E. Kingston, and J. T. Lee. 2013. 'High-resolution Xist binding maps reveal two-step spreading during X-chromosome inactivation', *Nature*, 504: 465-9.

21 Fujita, M. 1982. 'Bioassembly lines: a general mechanism for maintaining the specific architecture of a cell, with an implication for the function of a small RNA', *J Theor Biol*, 99: 9-13.

22 Bickmore, W. A., and B. van Steensel. 2013. 'Genome architecture: domain organization of interphase chromosomes', *Cell*, 152: 1270-84.

23 Guetg, C., and R. Santoro. 2012. 'Formation of nuclear heterochromatin: the nucleolar point of view', *Epigenetics*, 7: 811-4.

24 Nickerson, J. A., G. Krochmalnic, K. M. Wan, and S. Penman. 1989. 'Chromatin architecture and nuclear RNA', *Proc Natl Acad Sci U S A*, 86: 177-81.

25 Bouvier, D., J. Hubert, A. P. Seve, and M. Bouteille. 1985. 'Nuclear RNA-associated proteins and their relationship to the nuclear matrix and related structures in HeLa cells', *Can J Biochem Cell Biol*, 63: 631-43.

26 Pederson, T., and J. S. Bhorjee. 1979. 'Evidence for a role of RNA in eukaryotic chromosome structure. Metabolically stable, small nuclear RNA species are covalently linked to chromosomal DNA in HeLa cells', *J Mol Biol*, 128: 451-80.

27 Umlauf, D., P. Fraser, and T. Nagano. 2008. 'The role of long non-coding RNAs in chromatin structure and gene regulation: variations on a theme', *Biol Chem*, 389: 323-31.

28 Shevtsov, S. P., and M. Dundr. 2011. 'Nucleation of nuclear bodies by RNA', *Nat Cell Biol*, 13: 167-73.

29 Caudron-Herger, M., K. Muller-Ott, J. P. Mallm, C. Marth, U. Schmidt, K. Fejes-Toth, and K. Rippe. 2011. 'Coding RNAs with a non-coding function: maintenance of open chromatin structure', *Nucleus*, 2: 410-24.

30 Caley, D. P., R. C. Pink, D. Trujillano, and D. R. Carter. 2010. 'Long noncoding RNAs, chromatin, and development', *ScientificWorldJournal*, 10: 90-102.

31 Sun, L., L. A. Goff, C. Trapnell, R. Alexander, K. A. Lo, E. Hacisuleyman, M. Sauvageau, B. Tazon-Vega, D. R. Kelley, D. G. Hendrickson, B. Yuan, M. Kellis, H. F. Lodish, and J. L. Rinn. 2013. 'Long noncoding RNAs regulate adipogenesis', *Proc Natl Acad Sci U S A*, 110: 3387-92.

32 Stanke, M., M. Diekhans, R. Baertsch, and D. Haussler. 2008. 'Using native and syntenically mapped cDNA alignments to improve de novo gene finding', *Bioinformatics*, 24: 637-44.

33 Bradley, R. K., A. Roberts, M. Smoot, S. Juvekar, J. Do, C. Dewey, I. Holmes, and L. Pachter. 2009. 'Fast statistical alignment', *PLoS Comput Biol*, 5: e1000392.

34 Raj, A., P. van den Bogaard, S. A. Rifkin, A. van Oudenaarden, and S. Tyagi. 2008. 'Imaging individual mRNA molecules using multiple singly labeled probes', *Nat Methods*, 5: 877-9.

35 Berezney, R., and D. S. Coffey. 1974. 'Identification of a nuclear protein matrix', *Biochem Biophys Res Commun*, 60: 1410-7.

36 Luderus, M. E., A. de Graaf, E. Mattia, J. L. den Blaauwen, M. A. Grande, L. de Jong, and R. van Driel. 1992. 'Binding of matrix attachment regions to lamin B1', *Cell*, 70: 949-59.

37 Cuddapah, S., R. Jothi, D. E. Schones, T. Y. Roh, K. Cui, and K. Zhao. 2009. 'Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains', *Genome Res*, 19: 24-32.

38 Chu, C., J. Quinn, and H. Y. Chang. 2012. 'Chromatin isolation by RNA purification (ChIRP)', *J Vis Exp*.

39 Simon, M. D., C. I. Wang, P. V. Kharchenko, J. A. West, B. A. Chapman, A. A. Alekseyenko, M. L. Borowsky, M. I. Kuroda, and R. E. Kingston. 2011. 'The genomic binding sites of a noncoding RNA', *Proc Natl Acad Sci U S A*, 108: 20497-502.

40 Mariner, P. D., R. D. Walters, C. A. Espinoza, L. F. Drullinger, S. D. Wagner, J. F. Kugel, and J. A. Goodrich. 2008. 'Human Alu RNA is a modular transacting repressor of mRNA transcription during heat shock', *Mol Cell*, 29: 499-509.

41 Lee, E. K., M. J. Lee, K. Abdelmohsen, W. Kim, M. M. Kim, S. Srikantan, J. L. Martindale, E. R. Hutchison, H. H. Kim, B. S. Marasa, R. Selimyan, J. M. Egan, S. R. Smith, S. K. Fried, and M. Gorospe. 2011. 'miR-130 suppresses adipogenesis by inhibiting peroxisome proliferator-activated receptor gamma expression', *Mol Cell Biol*, 31: 626-38.

42 Nukitrangsan, N., T. Okabe, T. Toda, M. Inafuku, H. Iwasaki, T. Yanagita, and H. Oku. 2011. 'Effect of Peucedanum japonicum Thunb on the expression of obesity-related genes in mice on a high-fat diet', *J Oleo Sci*, 60: 527-36.

43 Rubi, B., A. del Arco, C. Bartley, J. Satrustegui, and P. Maechler. 2004. 'The malate-aspartate NADH shuttle member Aralar1 determines glucose metabolic fate, mitochondrial activity, and insulin secretion in beta cells', *J Biol Chem*, 279: 55659-66.

44 Seo, J., E. S. Fortuno, 3rd, J. M. Suh, D. Stenesen, W. Tang, E. J. Parks, C. M. Adams, T. Townes, and J. M. Graff. 2009. 'Atf4 regulates obesity, glucose homeostasis, and energy expenditure', *Diabetes*, 58: 2565-73.

45 Choy, L., and R. Derynck. 2003. 'Transforming growth factor-beta inhibits adipocyte differentiation by Smad3 interacting with CCAAT/enhancer-binding protein (C/EBP) and repressing C/EBP transactivation function', *J Biol Chem*, 278: 9609-19.

46 Gohring, F., and F. O. Fackelmayer. 1997. 'The scaffold/matrix attachment region binding protein hnRNP-U (SAF-A) is directly bound to chromosomal DNA in vivo: a chemical cross-linking study', *Biochemistry*, 36: 8276-83.

47 Hasegawa, Y., N. Brockdorff, S. Kawano, K. Tsutui, K. Tsutui, and S. Nakagawa. 2010. 'The matrix protein hnRNP U is required for chromosomal localization of Xist RNA', *Dev Cell*, 19: 469-76.

48 Lobov, I. B., K. Tsutsui, A. R. Mitchell, and O. I. Podgornaya. 2001. 'Specificity of SAF-A and lamin B binding in vitro correlates with the satellite DNA bending state', *J Cell Biochem*, 83: 218-29.

49 Huelga, S. C., A. Q. Vu, J. D. Arnold, T. Y. Liang, P. P. Liu, B. Y. Yan, J. P. Donohue, L. Shiue, S. Hoon, S. Brenner, M. Ares, Jr., and G. W. Yeo. 2012. 'Integrative genome-wide analysis reveals cooperative regulation of alternative splicing by hnRNP proteins', *Cell Rep*, 1: 167-78.

50 Xiao, R., P. Tang, B. Yang, J. Huang, Y. Zhou, C. Shao, H. Li, H. Sun, Y. Zhang, and X. D. Fu. 2012. 'Nuclear matrix factor hnRNP U/SAF-A exerts a global control of alternative splicing by regulating U2 snRNP maturation', *Mol Cell*, 45: 656-68.

51 Wong, L. H., K. H. Brettingham-Moore, L. Chan, J. M. Quach, M. A. Anderson, E. L. Northrop, R. Hannan, R. Saffery, M. L. Shaw, E. Williams, and K. H. Choo. 2007. 'Centromere RNA is a key component for the assembly of nucleoproteins at the nucleolus and centromere', *Genome Res*, 17: 1146-60.

52 Delpretti, S., T. Montavon, M. Leleu, E. Joye, A. Tzika, M. Milinkovitch, and D. Duboule. 2013. 'Multiple enhancers regulate Hoxd genes and the Hotdog LncRNA during cecum budding', *Cell Rep*, 5: 137-50.

53 Donohoe, M. E., S. S. Silva, S. F. Pinter, N. Xu, and J. T. Lee. 2009. 'The pluripotency factor Oct4 interacts with Ctcf and also controls X-chromosome pairing and counting', *Nature*, 460: 128-32.

54 van de Nobelen, S., M. Rosa-Garrido, J. Leers, H. Heath, W. Soochit, L. Joosen, I. Jonkers, J. Demmers, M. van der Reijden, V. Torrano, F. Grosveld, M. D. Delgado, R. Renkawitz, N. Galjart, and F. Sleutels. 2010. 'CTCF regulates the local epigenetic state of ribosomal DNA repeats', *Epigenetics Chromatin*, 3: 19.

55 Price, A. L., N. C. Jones, and P. A. Pevzner. 2005. 'De novo identification of repeat families in large genomes', *Bioinformatics*, 21 Suppl 1: i351-8.

56 Arlotta, P., B. J. Molyneaux, J. Chen, J. Inoue, R. Kominami, and J. D. Macklis. 2005. 'Neuronal subtype-specific genes that control corticospinal motor neuron development in vivo', *Neuron*, 45: 207-21.

57 Tiveron, M. C., M. R. Hirsch, and J. F. Brunet. 1996. 'The expression pattern of the transcription factor Phox2 delineates synaptic pathways of the autonomic nervous system', *J Neurosci*, 16: 7649-60.

58 Kelley, D., and J. Rinn. 2012. 'Transposable elements reveal a stem cell-specific class of long noncoding RNAs', *Genome Biol*, 13: R107.

59 Subramanian, A., P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov. 2005. 'Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles', *Proc Natl Acad Sci U S A*, 102: 15545-50.

Chapter 4: Dissection of the evolutionary dynamics and sequence elements of Firre

4.1 Introduction

Significant progress has been made to understand the evolution and origin of lncRNAs. Initial studies, based on the analysis of single nucleotide substitutions, insertions, and deletions, have shown that lncRNA genes are more conserved than the background genomic average but less conserved than protein coding genes[1-5]. Recently, larger scale evolutionary studies have shown that 20% of the human lncRNAs are not even expressed beyond chimpanzee; however, mammalian and hominid-specific ones show a more tissue-specific expression, which is a conserved trait across those lncRNAs[6]. In fact, only 400 lncRNA genes have been found to be more than 300 million years old[7]. The fast turnover of lncRNAs can explain how they can contribute to lineage-specific gene expression programs. For instance, the transcription of lncRNAs *in cis* to the protein coding loci can impact their expression profile and thus influence the establishment of new regulatory networks[8,9]. Collectively, these studies suggest that although there is an abundance of lncRNAs in many species, they show rapid evolution and are conserved species-specifically, hinting at unique mechanisms of evolutionary dynamics.

The lack evolutionary conservation of lncRNAs might raise questions regarding their functionality; however, sole lack of primary sequence conservation does not imbue lack of function. Higher conservation at the promoters of lncRNAs suggests that their transcription is important[4,10]. Given that 20-40% of all protein coding genes have antisense RNA transcription[11-13], transcription of the lncRNA alone can have important implications. The transcription in the antisense direction will change the nucleosome dynamics within the locus, and the product of transcription, the tethered transcript, might change the epigenetic landscape of the region, both of which will impact the expression of the coding gene[14].

In addition to the conservation of transcription, functional conservation of a lncRNA can be mediated via its structure and binding properties to proteins or other RNAs. For instance, the steroid receptor RNA activator (SRA), which is a 870 nt lncRNA that acts as a co-activator for multiple nuclear receptors[15] but lacks primary sequence conservation, shows an incredible secondary structure conservation back to marsupials and monotremes[16]. Another type of conservation can be within the domains of lncRNAs, as exemplified by the Xist lncRNA. Although Xist is an essential RNA for X chromosome dosage compensation[17-21], it is not conserved and is only specific to eutherian mammals[22,23], except for its repetitive domains, specifically RepA[24,25]. Interestingly, RepA is the module that is found to be essential for the silencing of the X chromosome since it is the region that recruits the PRC2 components[26,27]. Overall, functional conservation of lncRNAs should be dissected using different criteria than those for mRNAs due to the different modes of action of lncRNAs.

In order to understand the fast sequence divergence of lncRNAs, the mechanisms of their generation should be considered. The creation of new protein coding genes has been studied in the context of gene duplications of homologous protein coding loci[28,29]; however, a variety of mechanisms have been proposed for lncRNAs. lncRNAs can be generated via[30]: 1) DNA or RNA-based duplication of existing sequences, 2) pseudogenization of protein coding gene, followed by loss of protein coding capacity, 3) exaptation of noncoding DNA, and 4) exaptation of transposable elements. Transposable elements (TEs) actually compose ~30% human lncRNA sequences[31], more specifically in the promoter regions of lncRNAs in human ES cells[32].

The interesting observation of the abundance of TEs in lncRNAs is important because, in fact, more than half of the human and mouse genomes are comprised of repetitive

sequences[33]. Tandem repeats (TR) and local repeats (LR) are also as abundant as TEs. Since the discovery of these sequences, they were considered as "junk" DNA and regarded as a threat to genomic stability. Intriguingly, most of these repetitive sequences in the genome lie in the noncoding regions, including lncRNAs[32,34,35]. To that end, it was suggested that TEs might be essential domains for lncRNA function, and they were termed Repeat Insertion Domains of lncRNAs (RIDLs)[36]. Tandem repeats, on the other hand, are generally depleted both in mRNA and lncRNAs with respect to the genomic average (Figure 4.1.1).

Closer examination of the repetitive elements of the genome soon revealed that these elements actually serve important functions on the protein, DNA, and RNA levels. For instance, local repetitive units; such as, tandem repeats of 1-2 codons, within protein sequences enhance the generation of protein motifs and drive eukaryotic protein evolution[37]. On the DNA level, repetitive sequences are important for the integrity of telomeres and centromeres during cell division[38]. Moreover, small RNAs produced from the repetitive regions regulate homologous recombination at double stranded breaks, maintaining genome stability[39]. Furthermore, repetitive motifs have been reported to be specific to certain protein complexes[40-42], one of which is RNA polymerase 2 (RNAP2). Alu element B2 is capable of specifically interacting with the catalytic core of RNAP2 and inhibiting transcription[42].

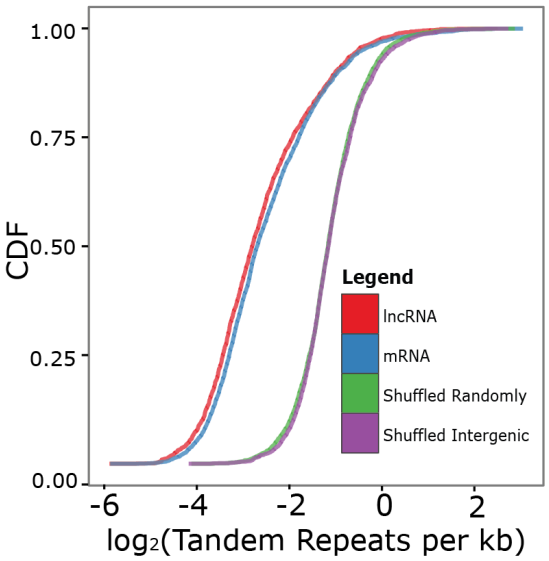


Figure 4.1.1: The distribution of tandem repeats across the genome.

The interaction of the repetitive motifs with specific regulatory or structural proteins renders them great candidates to nucleate compartments in the nucleus. These generic units can act as signals to physically integrate different parts of the genome in the complex environment of the nucleus. For example, Uchl1-antisense transcript harbors a SINEB2 element, which is responsible for regulating localization to the ribosome[43]. Similarly, Alu domain of the Anril lncRNA mediates the interaction with Polycomb, and in turn with its trans targets across chromosomes[44].

The roles of the repeat elements in the nucleus as discussed above present an interesting hypothesis for lncRNAs. lncRNAs harboring many of these elements possess the advantage of modularity, thus can play important regulatory and organizational roles in the nucleus. Due to experimental and computational rigor, repeat elements have been ignored, especially local repetitive motifs. In concordance with the analyses showing that the majority of the repetitive elements arise from the noncoding regions of the genome, we have analyzed the enrichment of local repetitive elements (LRs) in lncRNAs. To that end, we focused on one particular lncRNA, Firre[45], and interrogated the LRs in both human and mouse Firre loci. We found that one of the intronic LRs, R0, houses binding sites for critical transcription factors as well as CTCF, the binding of which is conserved across many species. On the other hand, further characterization of one of the exonic LRs, RRD, which we have previously discovered[45], showed us that RRD functions as an RNA nuclear localization signal in a species-specific manner via its protein partner hnRNPU. Collectively, our findings suggest that local repetitive motifs in lncRNAs can dictate the evolution of the function of lncRNAs, their binding partners, and their roles in the nucleus.

4.2 Results

4.2.1 lncRNAs are enriched in local repetitive motifs

To comprehensively survey the prevalence of repetitive elements in the human transcriptome (coding and noncoding), we calculated the number of local and tandem repeats present in each protein-coding gene annotated in GENCODE v19 and lncRNA in published catalogs (see Methods). We removed all known TEs and other simple repeats that have been previously cataloged in the human genome. We further controlled for differences in transcript length distributions of tested mRNA and lncRNAs. Once tandem and local repeat sequences were identified for each transcript, we performed additional controls by shuffling these sequences to see if they occur more often than the genomic average in a particular transcript and more often in lncRNAs or mRNAs. Specifically, we created two control sets, one by shuffling the lncRNA genes all around the genome and the other by shuffling the lncRNA genes to intergenic genomic regions (see Methods).

We found that lncRNA genes have significantly more local repeats per kb than protein coding genes as well as both the control sets (Mann Whitney Test, p-value <0.001) (Figure 4.2.1.1). In contrast, both lncRNAs and mRNAs have a similar number of tandem repeats per kb (Figure 4.1.1), and both have fewer tandem repeats than any of our control sets (Figure 4.1.1). Based on our analysis, we conclude that compared to mRNAs, lncRNAs have a higher density of local repeats but not tandem repeats.

4.2.2 FIRRE is a lncRNA with many unique local repetitive motifs

One of the chromosomes with many local repeat rich lncRNAs is the X chromosome (Figure 4.2.2.1). We find that protein coding genes on the X chromosome are not rich in repeats

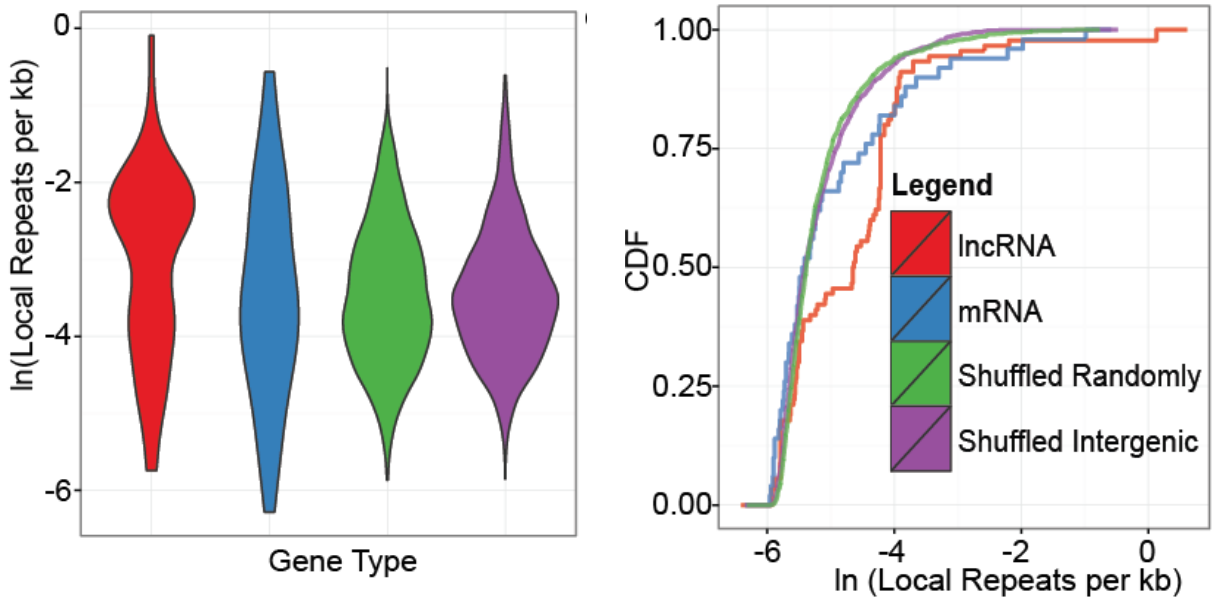


Figure 4.2.1.1: Local repeat distributions in lncRNAs and mRNAs compared to negative controls.

(Figure 4.2.2.1); where as, in contrast, lncRNAs are enriched for repeats. We observed a total of X of Y genes harboring local repeats on the X chromosome. These included X56, FIRRE, XIST, and many other annotated lncRNAs. Particularly, while *XIST* is enriched for tandem repeats, *X56* and *FIRRE* have numerous local repeats. Based on our previous characterization of the functional roles of FIRRE in nuclear organization, we hypothesized that these repeat regions may contribute to the function of FIRRE in the nucleus. We have previously shown that FIRRE is a strictly nuclear lncRNA that escapes X chromosome inactivation and is important for adipogenesis and nuclear organization. We found that RRD, a local repeat that is confined to the exons of FIRRE, is crucial for the nuclear localization and function of FIRRE as a nuclear organizer.

4.2.3 The Firre locus houses numerous conserved local repeats

To further investigate the repeat nature of the *FIRRE* locus, we used RepeatScout to find

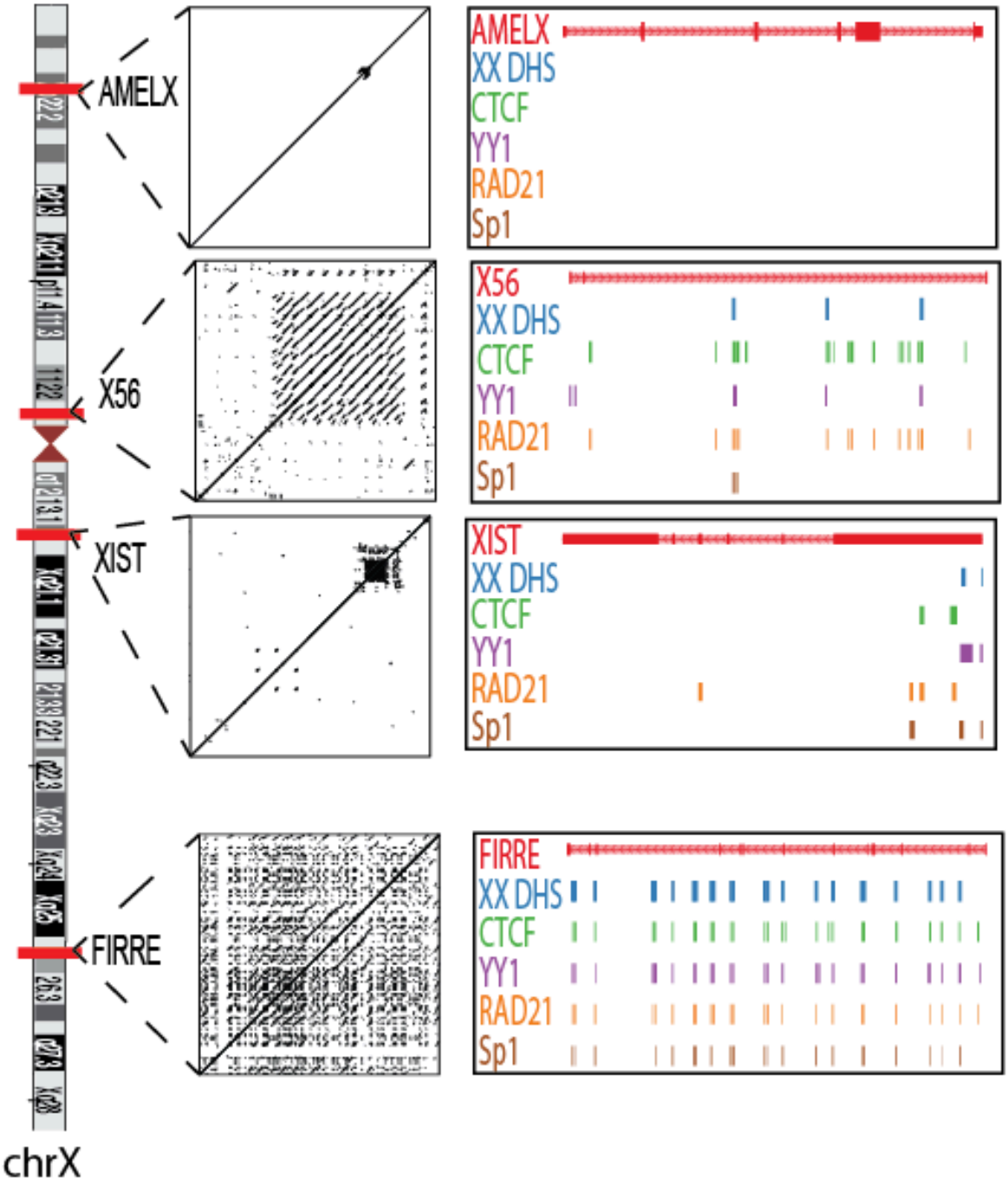


Figure 4.2.2.1: The repeat dot plots for coding (AMELX) and noncoding (X56, XIST, and FIRRE) regions on the X chromosome. Under the transcript structures, DNAse hypersensitive sites (blue), CTCF ChIP (green), YY1 ChIP (purple), RAD21 ChIP (orange), and Sp1 ChIP (brown).

the repeats and mapped them using RepeatMasker (see Methods). We found a total of 13 novel local repeat elements in the human *FIRRE* locus (Figure 4.2.3.1). These repeats ranged from 160-500 bp and occurred 7-50 times within the *FIRRE* locus. Interestingly, these repeats are unique to the *FIRRE* locus and are found nowhere else in the human genome. Furthermore, the previously studied RRD repeat was the only instance that consistently overlapped with the exonic sequences of FIRRE. In the human *FIRRE* locus, 5 out of 8 instances of RRD are contained within the exons (13 exons total); and in the mouse *Firre* locus, 11 out of 13 RRD occurrences are in the exons (23 exons total). Notably, all the new local repeats we found in *FIRRE* are DNA repeats that occur in the introns or in the promoter of FIRRE, possibly suggesting a role at the DNA level.

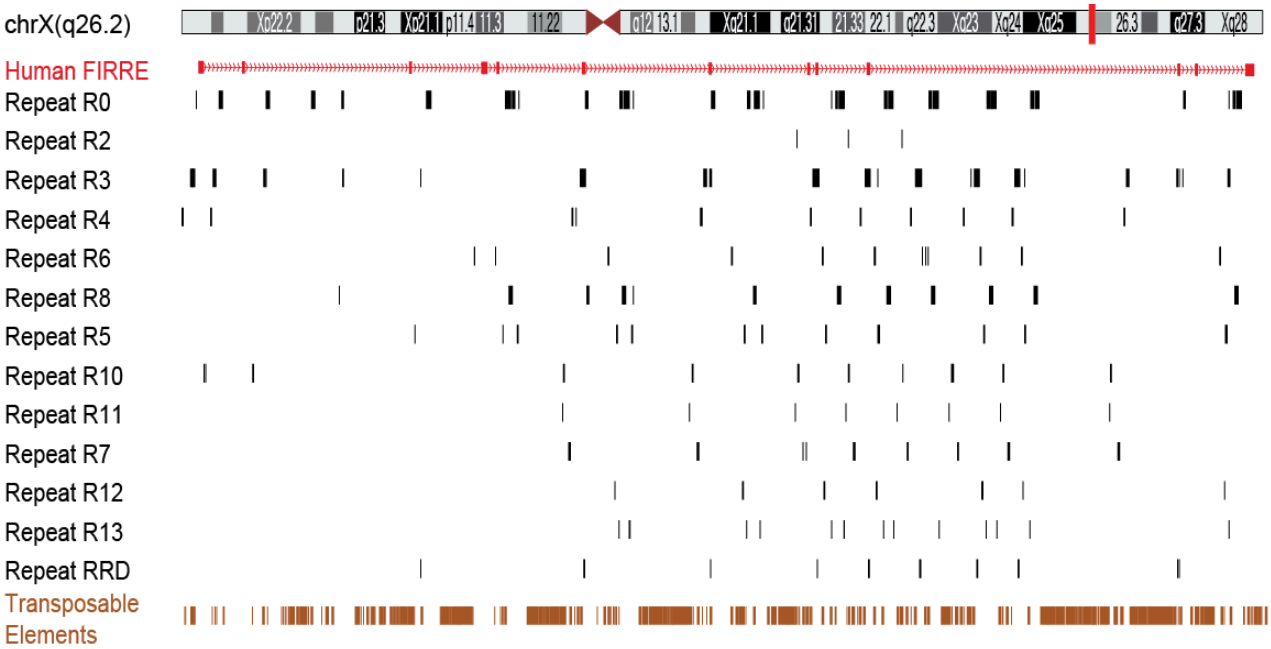


Figure 4.2.3.1: The human *FIRRE* locus along with the new local repeats and RRD. Exons marked by thick red bars in the FIRRE transcript.

We next compared the evolution of the 13 local repeats in *FIRRE* between the syntenic human and mouse loci. Briefly, we aligned the genomic regions between human and mouse and considered homology at 65% and above for sequence conservation. Our analysis revealed 5 out of 13 local repeats in human *FIRRE* are conserved in mouse (R0, R1, R3, R8, R10, and RRD) (Figure 4.2.3.2). Similar to the human *FIRRE* locus, mouse *Firre* has 4 of these repeats in the intronic regions and RRD, which is confined to the exonic sequences. Based on our analysis, we conclude that the human and mouse *FIRRE* loci contain many conserved local repeats.

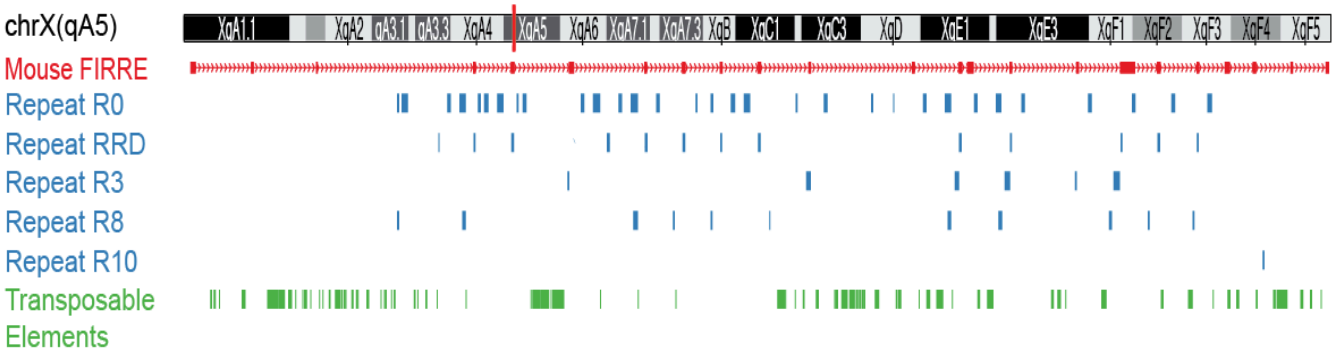


Figure 4.2.3.2: The mouse *Firre* locus along with new local repeats and RRD. Exons marked by thick red bars in the Firre transcript.

4.2.4 CTCF, YY1, Sp1 and RAD21 bind R0 across all occurrences

Since all the local repeats, except for RRD, are DNA repeats, we hypothesize that some of these repeats may influence the localization of transcription factors (TF) or other chromatin factors. In order to test this hypothesis, we first mapped all transcription factor motifs from the JASPAR database to these repeat elements. Briefly, we determined the enrichment for a given TF or other protein within a repeat region over its binding average on the genome. Out of the repeat motifs that we looked at, we, specifically, observed that repeat R0 has motifs for many

transcription factors like E2F3, ETS1, SP1, SP2, KLF5 and YY1.

In order to determine if any of these transcription factors binds the repeat motif, we analyzed ChIP-Seq data of YY1, Sp1 and Sp2 in human embryonic stem cells (hESC) from the Encylopedia of DNA Elements (ENCODE) consortium. Since, we are analyzing binding to repetitive regions of the genome, we took special consideration of the multi-mapping reads (see Methods). To this end, we aligned ENCODE ChIP-Seq data and performed a peak calling analysis and tested for enrichment. We observed that YY1 and Sp1 are both enriched at Repeat R0 in hESC (Poisson test - p-value < 0.001, Figure 4.2.4.1). However, we did not see binding of Sp2 at Repeat R0 in hESC (Figure 4.2.4.1). Interestingly, we also observed strong enrichment for CTCF binding at R0. We previously reported that there are many binding sites in the *FIRRE* locus[45]; and hereby found that the binding occurs specifically at every R0 sequence in hESCs (Poisson test - p-value < 0.001, Figure 4.2.4.1). In agreement with previous findings, YY1 co-occupies R0 sites with CTCF (Figure 4.2.4.1). Similarly, we also found that RAD21 is enriched at Repeat R0 in hESC (Poisson test - p-value < 0.001, Figure 4.2.4.1). To test whether the binding of CTCF, YY1, SP1, and RAD21 complexes at R0 is conserved across different cell types, we investigated ENCODE ChIP data for these factors in GM12878 and NHLF cells. In both cell lines, we observed similar binding patterns specifically at every R0 occurrence in the *FIRRE* locus.

Since Repeat R0 is conserved between human and mouse, we decided to test if the CTCF binding at R0 is also conserved across the two species. We analyzed CTCF ChIP-Seq data from mouse embryonic stem cells (mESC) from the ENCODE consortium and found that CTCF also binds at repeat R0 in the mouse *Firre* locus (Poisson test - p-value < 0.001, Figure 4.2.4.2). We further investigated CTCF binding in the *FIRRE* locus from macaque and rat and

again observed CTCF binding at repeat R0 (Poisson test - p-value < 0.001, 4.2.4.2). Collectively, these results suggest that R0 is a conserved local repeat in the *FIRRE* locus that exhibits strong and conserved binding of CTCF.

The binding of CTCF across the *Firre* locus shows an intriguing trend, with CTCF binding at the R0 motif, although R0 does not contain the CTCF binding motif, followed by RRD. Both human and mouse Firre loci, which are actively transcribed, show this pattern (Figure 4.2.4.3 and Figure 4.2.4.4).

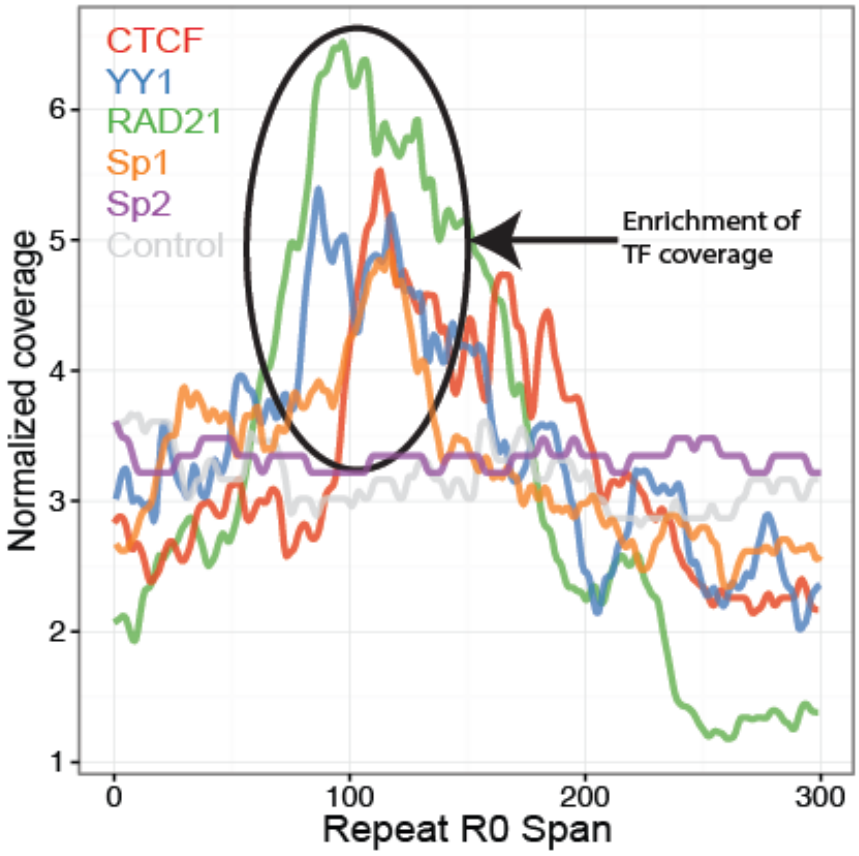


Figure 4.2.4.1: ChIP peaks for various transcription factors and CTCF shown across the R0 region. Normalized to input control.

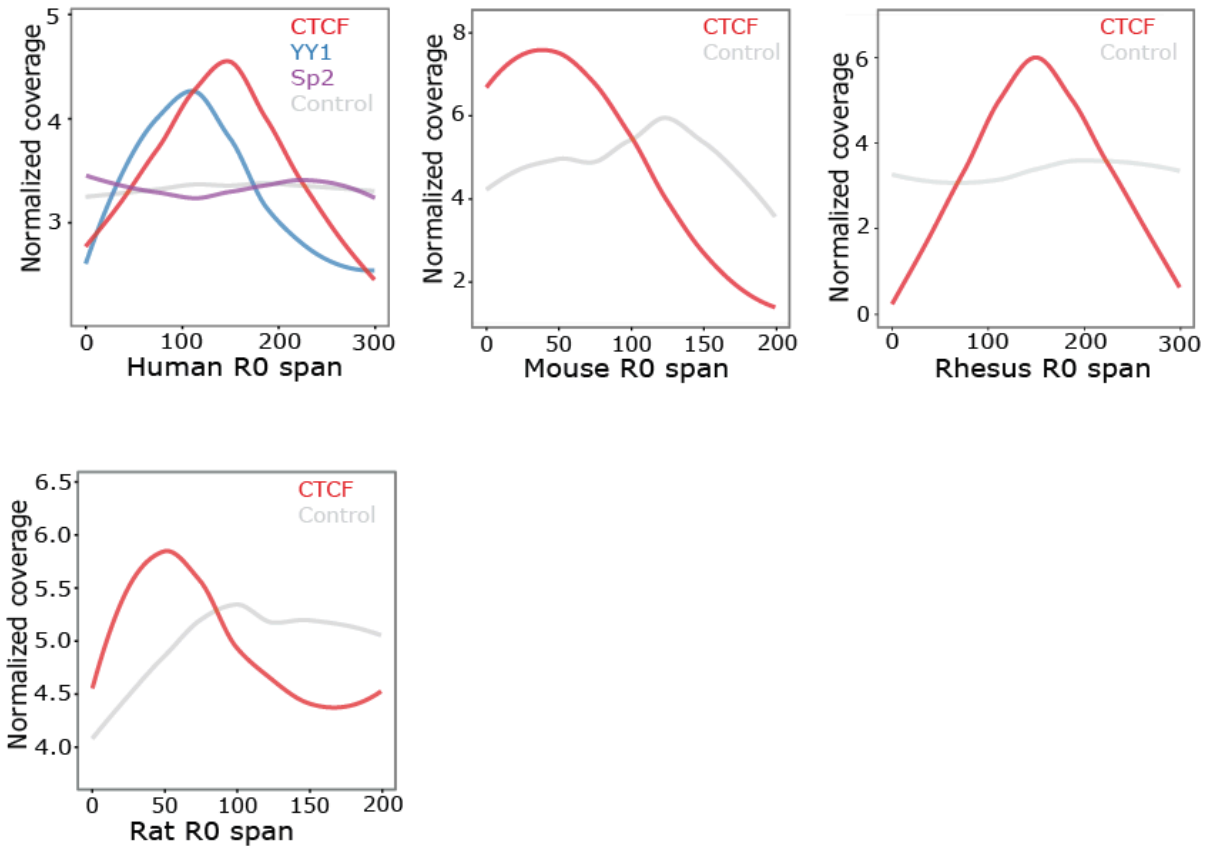


Figure 4.2.4.2: CTCF ChIP peaks across the R0 sequence in human, mouse, macaque, and rat species. Normalized to input.

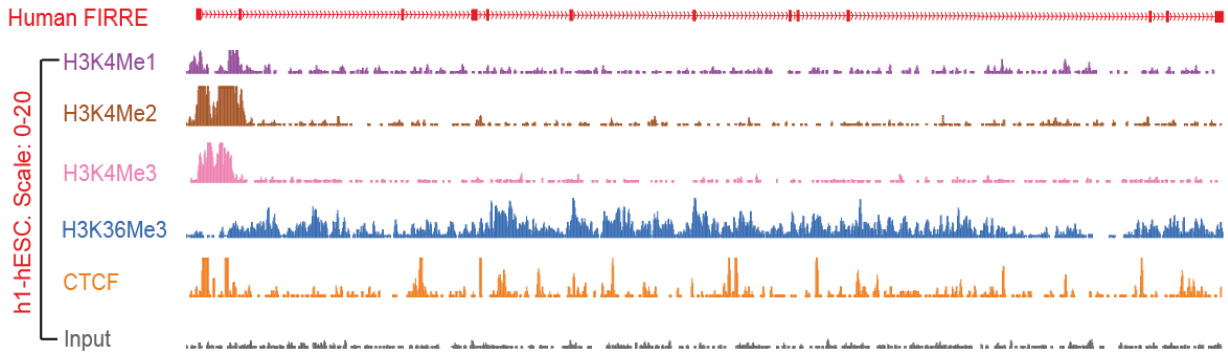


Figure 4.2.4.3: Actively transcribed human *FIRRE* locus, with CTCF marks shown at every R0 occurrence followed by RRD.

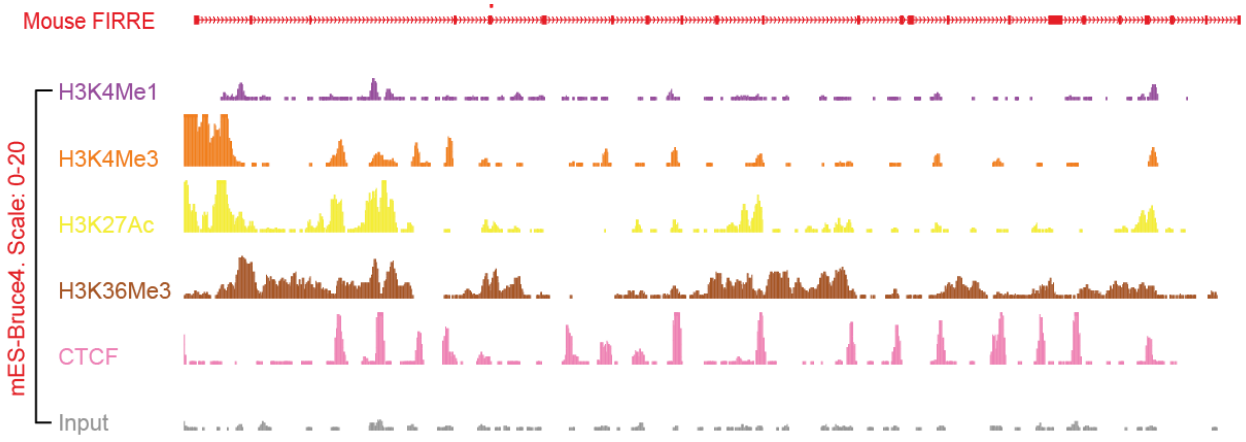


Figure 4.2.4.4: Actively transcribed mouse *Firre* locus, with CTCF marks shown at every R0 occurrence followed by RRD.

Interestingly, when we detect the occurrence of RRD following R0 (5 times), we also see a strong CTCF binding at R0 and RNA Polymerase 2 at RRD (Figure 4.2.4.5). CTCF ha been known to play a role in three-dimensional interactions and transcriptional regulation[46,47].

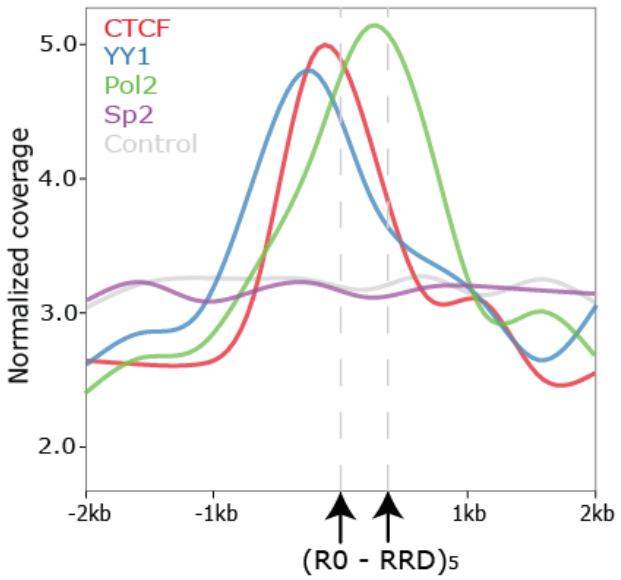


Figure 4.2.4.5: Occurrences of R0 and RRD in the *Firre* locus along with ChIP tracks for CTCF (red) and Pol2 (green).

4.2.5 RRD functions as a nuclear localization signal species-specifically

While RRD is conserved in human and mouse, they share only ~68% sequence identity. We also see this trend for the other local repeats; Repeat R0 for example shares ~65% sequence identity between human and mouse. Yet, these local repeats seem to share the same function of binding to multiple TFs and CTCF. To investigate this further, we decided to construct the evolutionary tree of FIRRE and each of the local repeats in different mammals using MAFFT and Neighbor Joining (see Methods). We see that FIRRE, RRD and R0 all show a clear split between primates and rodents (Figure 4.2.5.1).

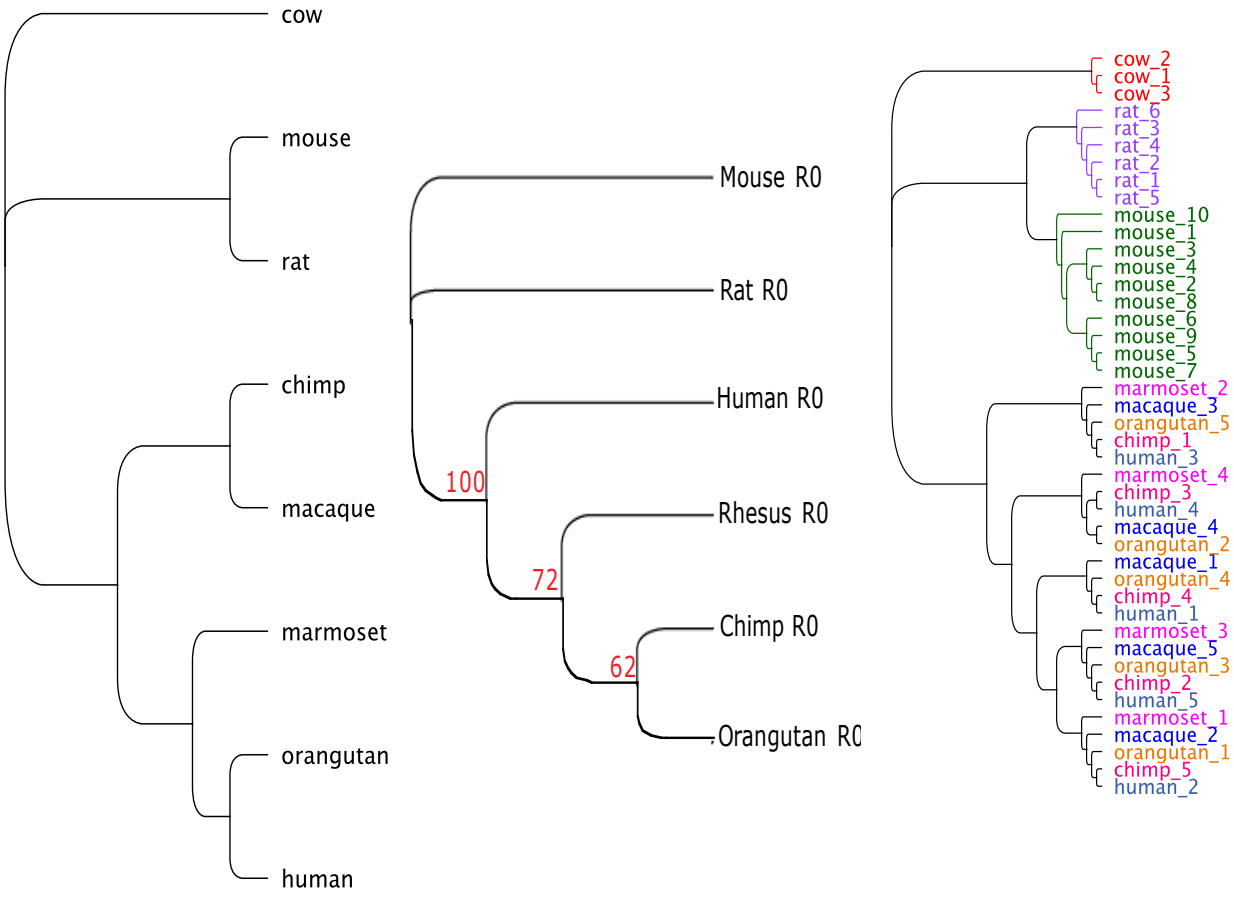


Figure 4.2.5.1: The evolutionary conservation of the Firre locus, R0 repeat and RRD, respectively.

The evolutionary divergence of the repeat sequences in the *FIRRE* locus lead us to question whether the divergence in sequences is representative of a divergence in functional roles. We first focused on RRD since it is the only repeat that is exclusively in the exons and wanted to further investigate how it might play a role in the function of mature Firre RNA. We have previously discovered that overexpression with an isoform of Firre without RRD results in the translocation of Firre transcripts into the cytoplasm[45]. Therefore, we wanted to test whether RRD is sufficient to localize RNA transcripts in the nucleus, and whether this functionality differs across species.

To determine if RRD localizes transcripts to the nucleus, we have made numerous constructs (Figure 4.2.5.2), in which RRD is appended to an otherwise cytoplasmic RNA Sox2. We performed these experiments in mouse lung fibroblasts (mLFs) because these cells do not endogenously express Sox2. Mouse Sox2 was cloned into a lentiviral expression vector, which we made and termed "lincXpress". To determine the localization of Sox2 after overexpression, we performed single molecule RNA FISH (smRNA-FISH) with exonic probes conjugated to Alexa 594 and targeting Sox2 as described[48]. We overexpressed Sox2 alone in mLFs and observed that ~80% of Sox2 transcripts localize in the cytoplasm. Next, we overexpressed Sox2 appended at its 3' end with RRDs from five different species: mouse, rat, macaque, chimp, and human. In

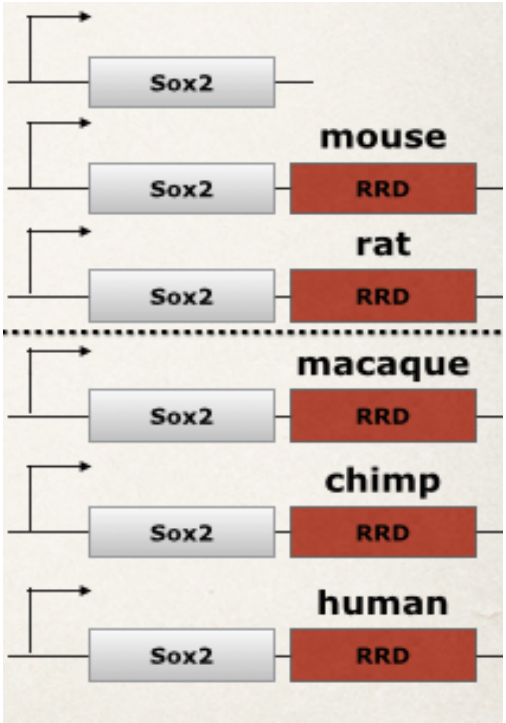


Figure 4.2.5.2: Lentiviral Sox2 constructs and variants used for viral transductions.

choosing these species, we aimed at having equal representations from the rodent and primate lineages based on the evolutionary divergence of RRD. Each construct was overexpressed in mLFs by viral transduction and visualized by smRNA-FISH (Figure 4.2.5.3). We quantified the percentage of transcripts that localized in the nucleus by using MATLAB scripts (Figure 4.2.5.4). As an independent control for overexpression variability, we checked the expression levels of each Sox2 variant by RT-PCR with primers targeting Sox2 (Figure 4.2.5.5), indicating the numbers of RNA transcripts were comparable across all conditions.

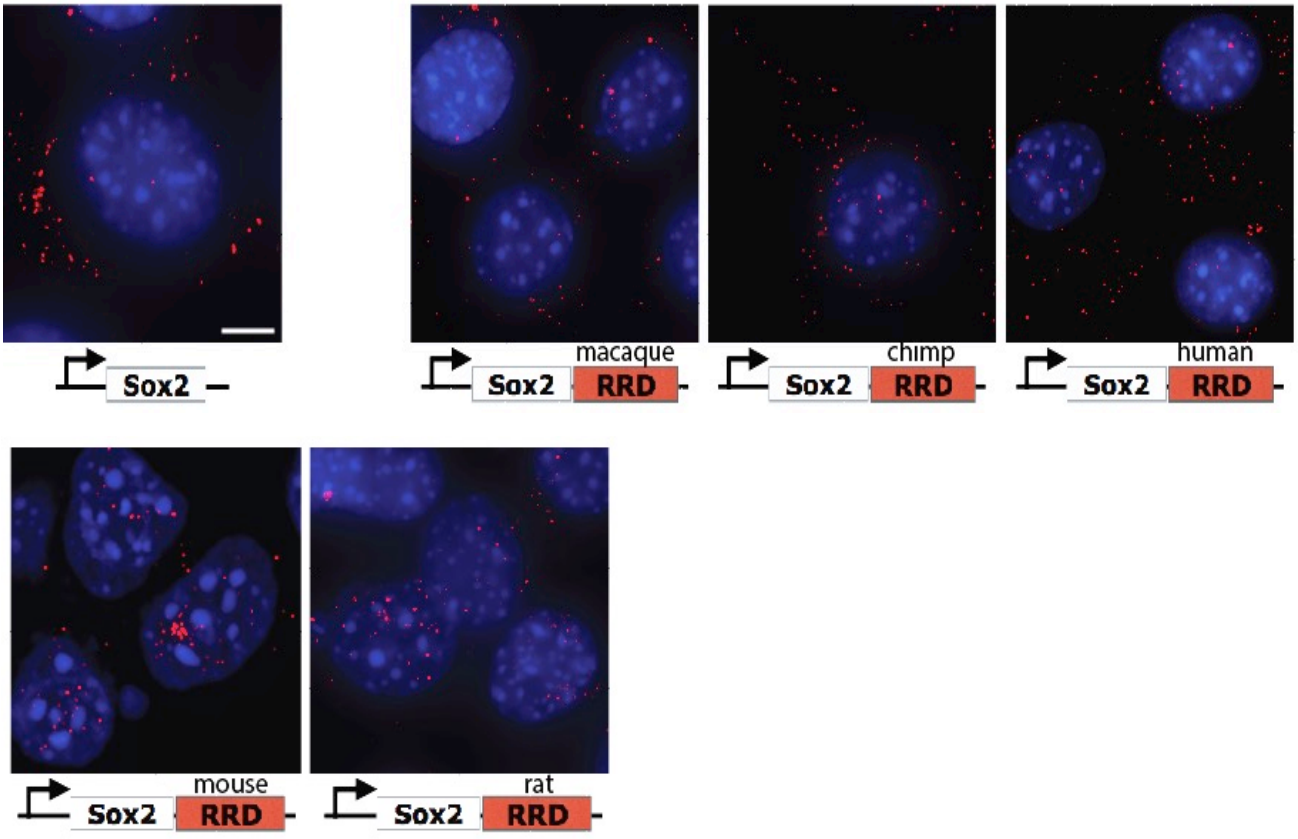


Figure 4.2.5.3: Viral overexpression of Sox2 or Sox2 appended with RRD from different species in mLFs. Visualized by RNA FISH targeting the exon of Sox2 labeled "red" (Alexa594). Nuclei by DAPI.

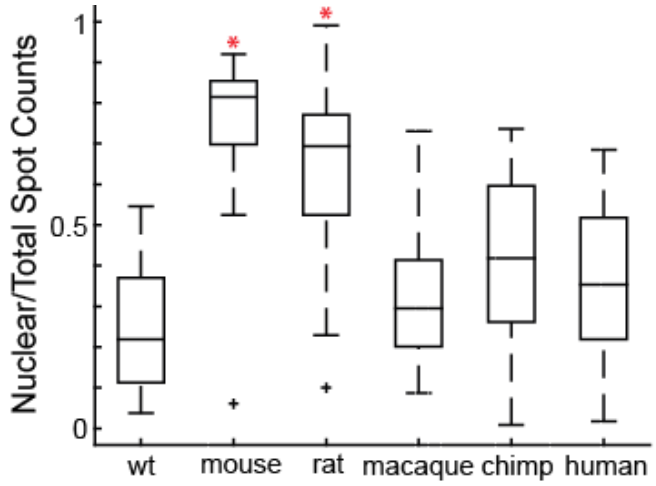


Figure 4.2.5.4: The quantification of the localization of Sox2 or Sox2+xRRD transcripts in mLFs.

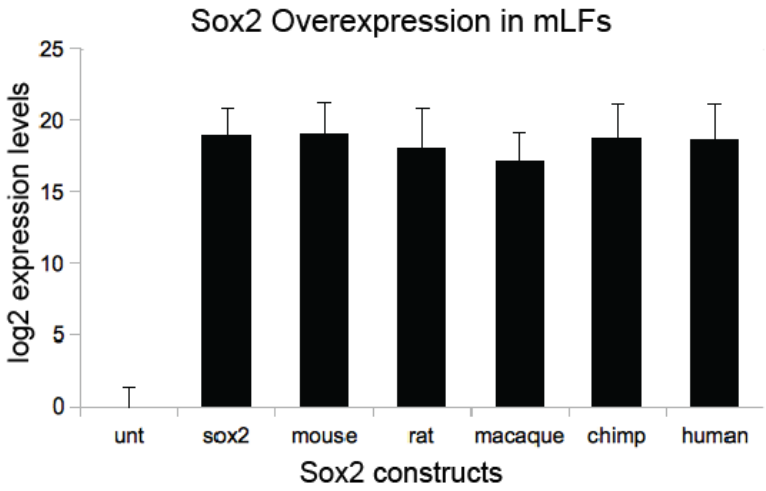


Figure 4.2.5.5: qRT-PCR measurement of expression levels of Sox2 across conditions in mLFs.

Strikingly, we observed an almost exclusively nuclear localization of Sox2 appended with mouse or rat RRD but not with macaque, chimp, or human RRDs. Our analyses revealed that the mouse and rat RRDs skew the distribution of Sox2 transcripts to be more nuclear: 80%

nuclear (Student's t-test, p<7.10e-9) and 65% (p<4.06e-14), respectively (Figure 4.2.5.3,4). In contrast, there was a significant reduction in the number of Sox2 transcripts that localized in the nucleus when macaque, chimp, or human RRD was added to the 3' end of Sox2 (28% (p<0.0286), 40% (p<3.08e-4), and 31% (p<0.0068), respectively) (Figure 4.2.5.3,4). We have further verified the distributions of transcripts by biochemical fractionation of nuclear and cytoplasmic compartments (Figure 4.2.5.6). Collectively, these results suggest that in mLFs, the rodent RRD sequences are sufficient to localize Sox2 in the nucleus; whereas, the primate lineage RRD sequences do not have an effect on the distributions.

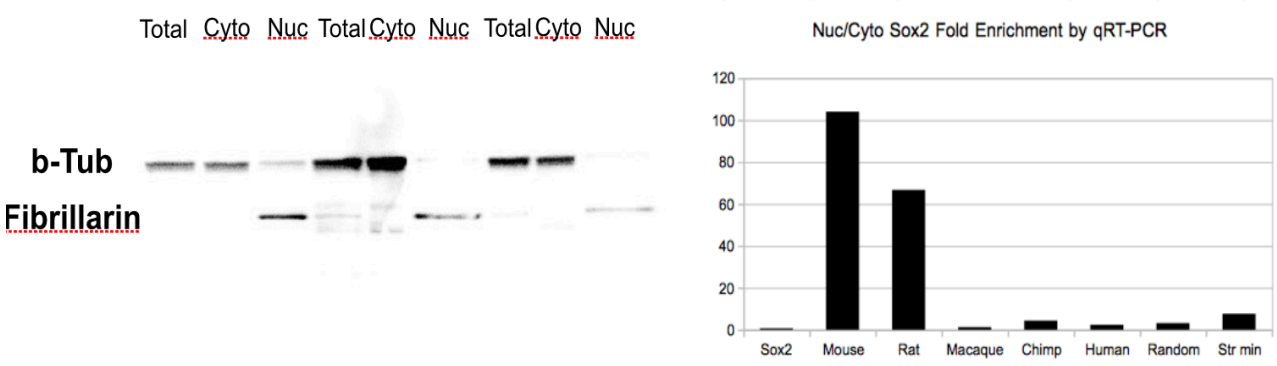


Figure 4.2.5.6: Biochemical fractionation of nuclear (fibrillarin) and cytoplasmic (b-tubulin) compartments of mLFs. Shown by RT-PCR for Sox2 or Sox2+RRD expression in each compartment (presented as nuclear/cytoplasmic ratio normalized to Sox2 alone overexpression).

Having detected a species-specific effect of RRD on the distribution of Sox2 in mLFs, we were intrigued by whether we would observe a reciprocal effect if we did the same experiments in human cells. We have chosen human foreskin fibroblasts since they also do not endogenously express Sox2. Similar to what was observed in mLFS, Sox2 alone localizes in

the cytoplasm (Figure 4.2.5.7). Surprisingly, the mouse (Student's t-test, $p<0.3221$) and rat ($p<0.6544$) RRDs result in a localization pattern similar to that of Sox2 alone; whereas, macaque, chimp, and human RRDs significantly alter the distribution of Sox2 RNAs to be more nuclear (Student's t-test, $p<1.49e-6$, $p<3.37e-7$, and $p<3.94e-9$, respectively) (Figure 4.2.5.7,8). Similarly, we have confirmed that the difference in the distribution of transcripts is not due a difference in the expression levels of each RNA species (Figure 4.2.5.9). Overall, our results show a divergence in sequence evolution between the rodent and primate lineages and a concordant functional alteration of nuclear localization in the respective species.

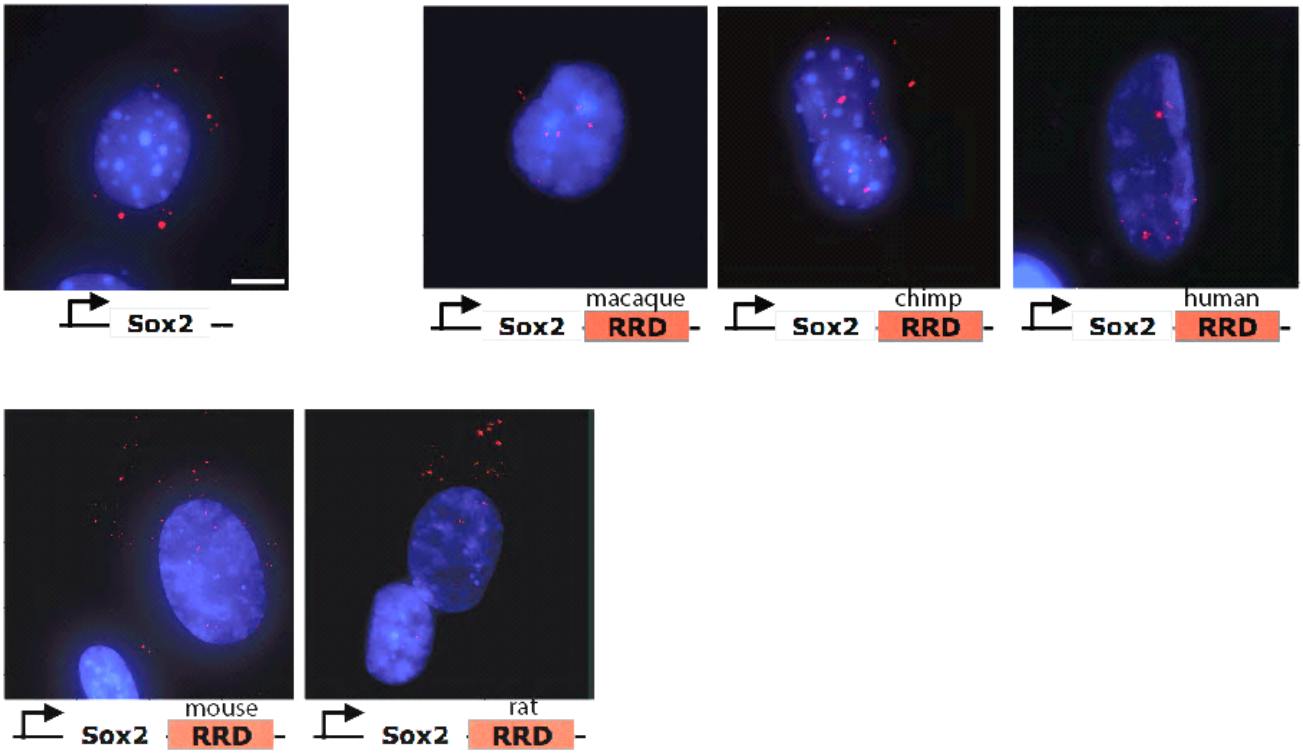


Figure 4.2.5.7: Viral overexpression of Sox2 or Sox2 appended with RRD from different species in hFFs. Visualized by RNA FISH targeting the exon of Sox2 labeled "red" (Alexa594). Nuclei by DAPI.

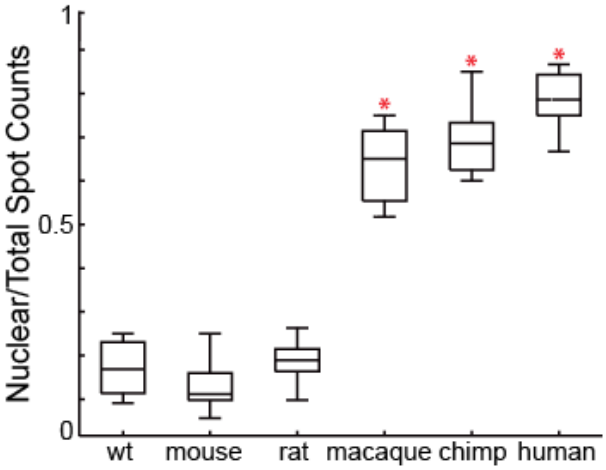


Figure 4.2.5.8: The quantification of the localization of Sox2 or Sox2+xRRD transcripts in hFFs.

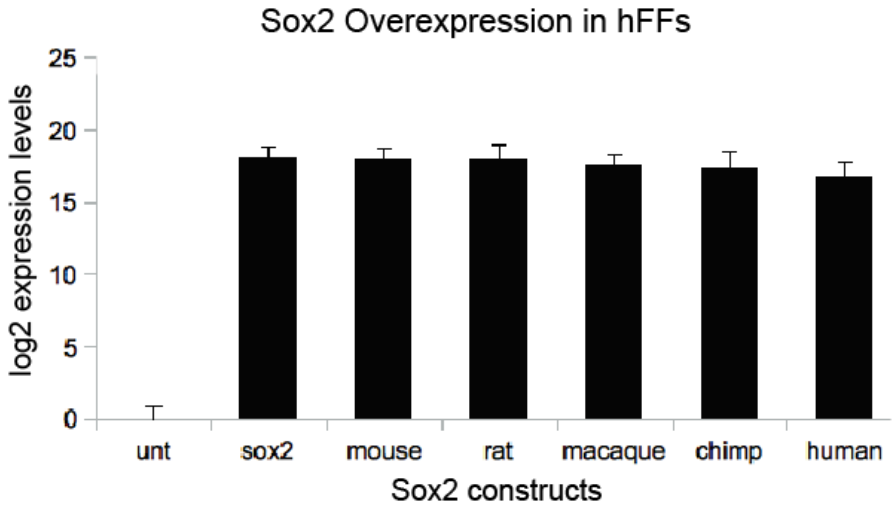


Figure 4.2.5.9: qRT-PCR measurement of expression levels of Sox2 across conditions in hFFs.

4.2.6 hnRNPU might play a role for the function of RRD

The species-specific distribution of RNA transcripts in mouse and human cells upon the addition of RRD from their respective species raised the question of whether there can be any responsible protein factors for the observed localization differences. We have previously found that hnRNPU binds Firre via RRD, and depletion of hnRNPU results in mislocalization of Firre transcripts into the cytoplasm in HEK293s and HeLa cells[45]. Furthermore, the loss of hnRNPU also affected the co-localization of Firre with its trans-chromosomal targets in the nucleus. Therefore, we hypothesized that hnRNPU might play an important role for how RRD affects the Sox2 distribution.

To test this hypothesis, we performed RNAi-mediated knockdown of hnRNPU in mLFs and hLFs and repeated Sox2+mouse RRD and Sox2+human RRD transductions, respectively (Figure 4.2.6.1). We found that the knockdown of hnRNPU had a dramatic effect on the nuclear localization of Sox2+mouse RRD in mLFs and Sox2+human RRD in hFFs, suggesting that hnRNPU could play a role in keeping transcripts with RRD in the nucleus. However, we are aware of the role of hnRNPU in nuclear organization; therefore, the alteration in the nuclear to cytoplasmic distribution of these transcripts can be an indirect effect, caused by loss of contacts with other proteins or change of organization in the sub-nuclear territories bound by hnRNPU.

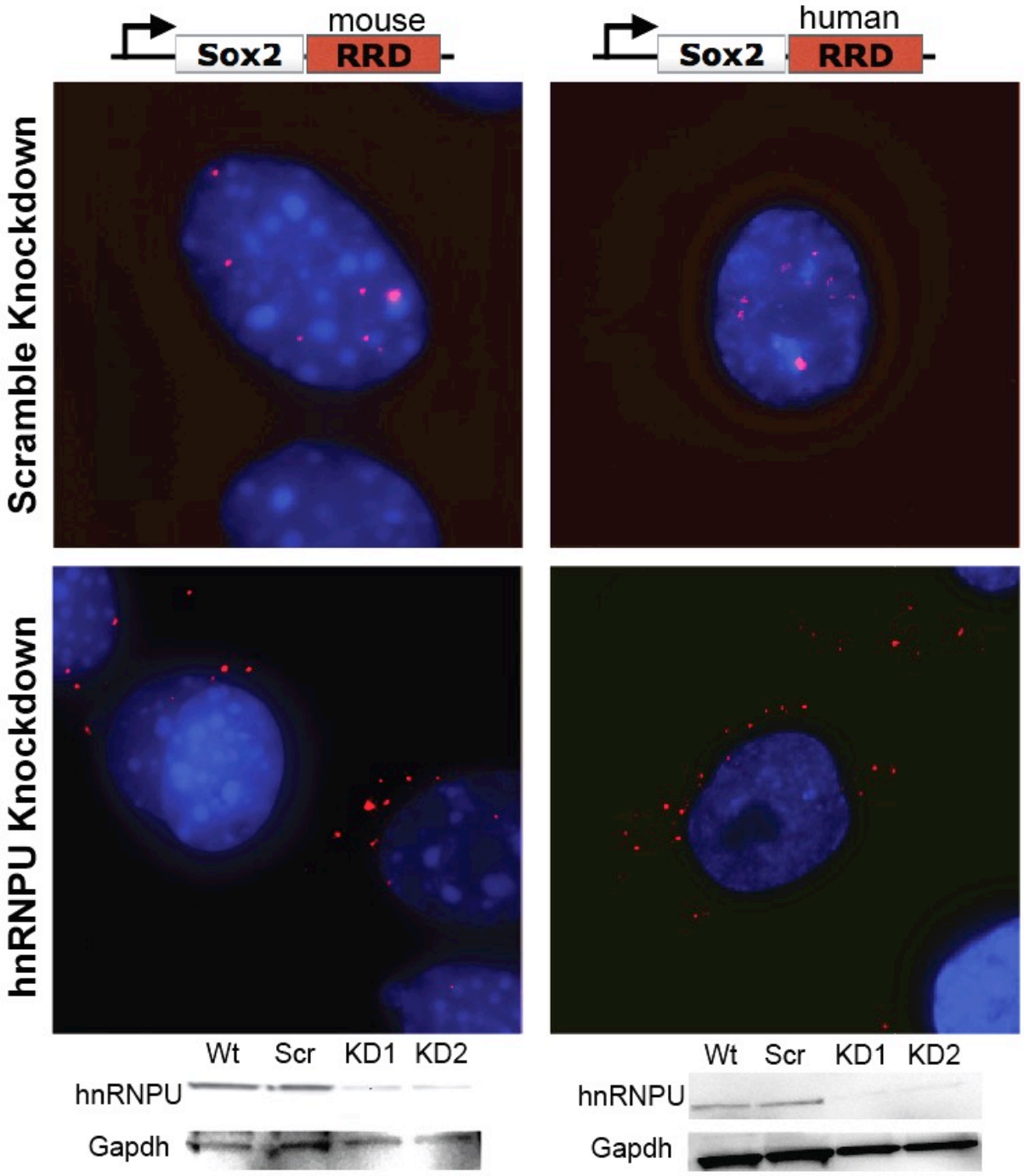
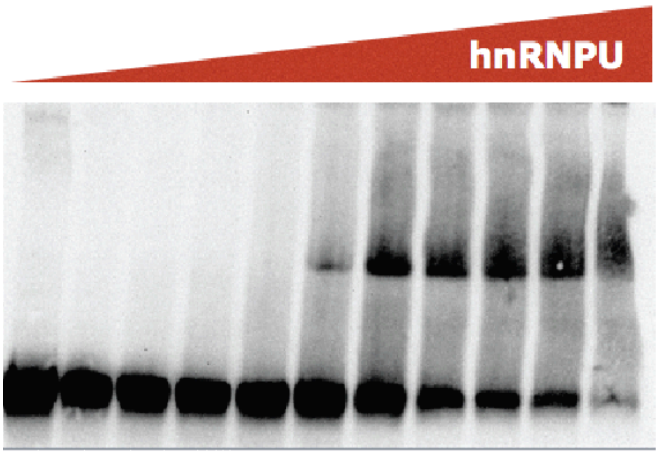


Figure 4.2.6.1: Viral overexpression of Sox2 + mouse RRD and Sox2 + human RRD in mLFs and hFFS, respectively, in hnRNPU knockdown conditions. Western blots showing protein levels.

4.2.7 hnRNPU binds RRD with high affinity and alters its structure

Having found that hnRNPU binds Firre via the RRD motif and affects the nuclear/cytoplasmic distribution of RRD+ transcripts in a species-specific manner, we further wanted to investigate the binding between hnRNPU and RRD. First, we wanted to determine the binding affinities of mouse hnRNPU:mouse RRD and human hnRNPU:human RRD interactions. Briefly, we purified human and mouse hnRNPU proteins using a BioEase tag affinity purification followed by AcTEV protease elution (see Methods). We tested the binding affinities via electromobility shift assay (EMSA) using human and mouse RRD RNA sequences. We found that the Kd of the mouse RRD and mouse hnRNPU interaction is 200±50 nM (Figure 4.2.7.1) and human RRD and human hnRNPU 180±25 nM (Figure 4.2.7.2); however, the interaction cross species has a significantly lower affinity (Figure 4.2.7.3). This finding indicated that there is a species-specific interaction between the protein and the RNA, suggesting that the localization difference between the mouse versus human RRD in their respective species might have been, in part, due to a species-specific co-evolution of binding.

We further wanted to assess the structure-function relationships driving the RRD and hnRNPU interaction. Specifically, we analyzed the structure of RRD with and without hnRNPU, using **s**elective 2'-**h**ydroxyl **a**cylation analyzed by **p**rimer **e**xtension followed by **seq**uencing (SHAPE-seq) (Methods). The structure along with the reactivity spectra of the human RRD is shown in Appendix 7. Then we wanted to determine whether we can detect any changes to the RNA structure in the presence of its protein binding partner. To that end, we repeated the SHAPE-seq experiment after incubating human RRD with human hnRNPU (previously described [49]).



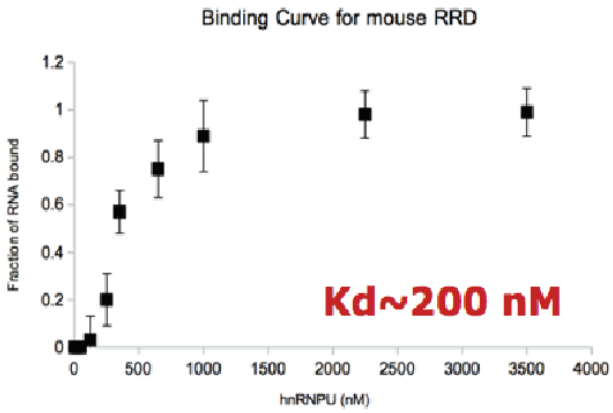


Figure 4.2.7.1: EMSA using purified mouse RRD and mouse hnRNPU. Kd ~200 nM.

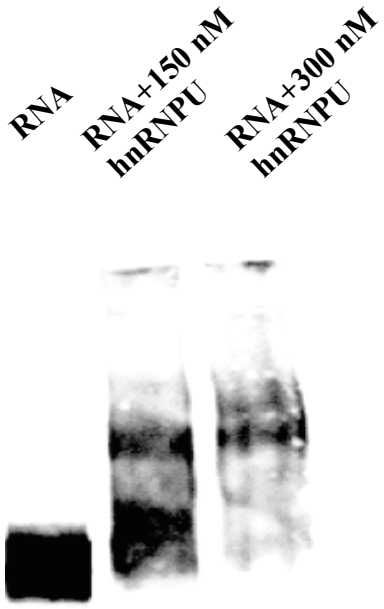


Figure 4.2.7.2: EMSA using purified human RRD and human hnRNPU. Kd ~180 nM.

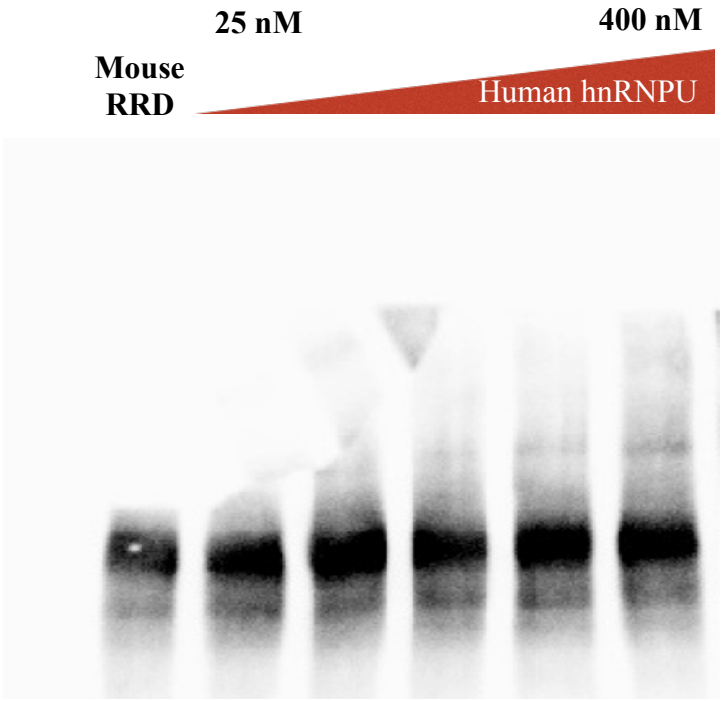


Figure 4.2.7.3: EMSA using purified mouse RRD and human hnRNPU.

The addition of human hnRNPU caused certain regions of RRD to be more reactive and others to be less reactive (Figure 4.2.7.5). Intriguingly, the stem loop, which becomes less reactive in human RRD upon human hnRNPU addition, contains the previously identified human hnRNPU binding motif, suggesting that the binding might be either mediated through this motif or the binding of hnRNPU causes a global alteration in the structure that results in the sequestering of the loop. Overall, our results show that there is a species-specific high affinity interaction between hnRNPU and RRD, which might play an important role for the recognition of RRD as a nuclear localization signal.

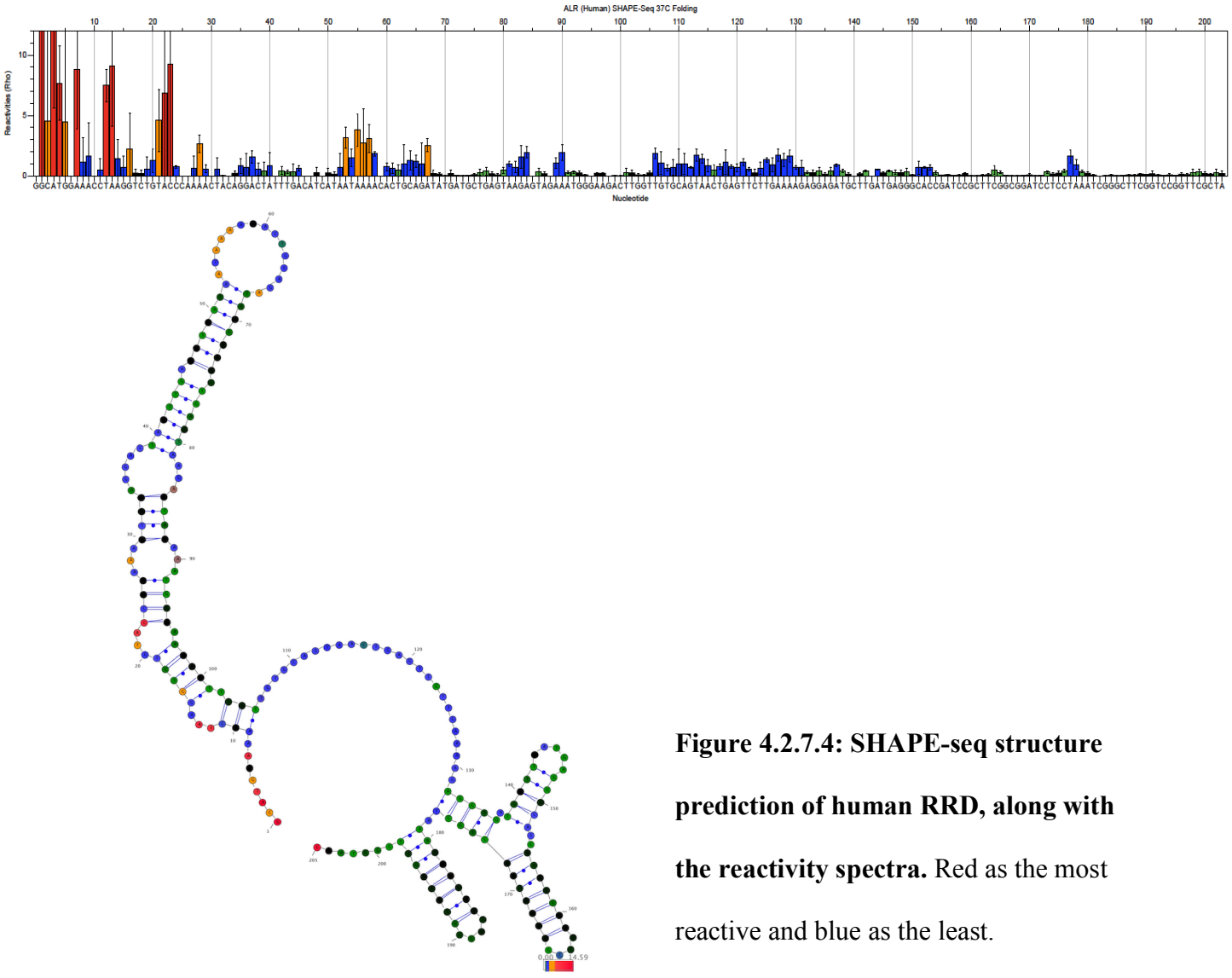


Figure 4.2.7.4: SHAPE-seq structure prediction of human RRD, along with the reactivity spectra. Red as the most reactive and blue as the least.

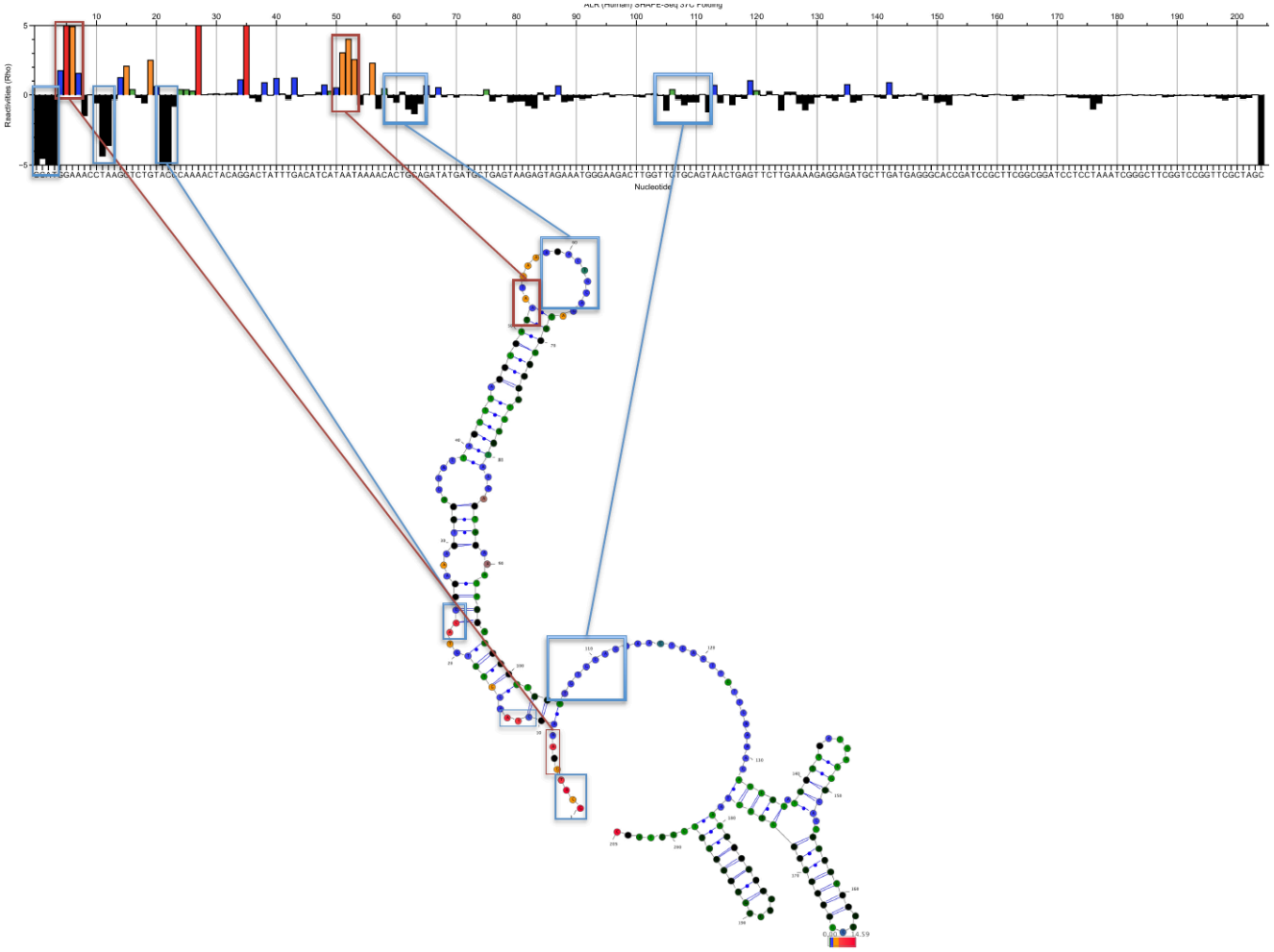


Figure 4.2.7.5: SHAPE-seq reactivity spectra differences when human hnRNPU is incubated with human RRD. The reactivities that increase are highlighted in orange boxes, the ones that decrease in blue boxes.

This work is being submitted to Cell Reports: **Hacisuleyman, E**., C. Shukla, J. Rinn. 'The role of local repetitive motifs in the function of the nuclear lncRNA Firre.'

4.3 Discussion

While it is becoming clearer that lncRNAs emerge as critical players in gene regulation and in the etiology of disease, the lack of primary sequence conservation has been a pressing concern for their functionality. Primary sequence of protein coding genes is under high selective pressure to preserve the codon structure; however, lncRNA sequence conservation is not high except for the promoter regions and splicing enhancers within exons[50]. This makes the orthology analysis of lncRNAs more difficult. In addition, lncRNAs are expressed at significantly lower levels than mRNAs, which blurs the boundaries between ease of detection versus evolutionary divergence. Finally, the hardest technological barrier is sequencing various tissues in multiple species other than mouse, human, and rat. This is of importance because lncRNAs are highly tissue-specific; thus, conservation analysis requires comparison across the same tissue types.

The reason that causes lncRNAs to diverge in their primary sequences might the abundance of repetitive units in the promoters as well as the gene bodies of lncRNAs. The repetitive units, especially specific classes of transposable elements (HERVHs), in the promoters of lncRNAs, regulate their stem cell specific expression pattern, establishing important regulatory networks. An example of this phenomenon is shown by the intergenic lncRNA ROR, which in its promoter houses an HERVH element that binds to core pluripotency proteins and thus modulates reprogramming of fibroblasts to induced pluripotent stem cells[51]. There might be a cell-type specific protein factor (transcription factor or a histone de/methylase) that recognizes the repetitive motif, which results in the up-regulation or down-regulation of the lncRNA target. It would be interesting to see if the tissue specific enhancer

ncRNAs[52] house any repetitive motifs that might be involved in the tissue specific activity of the RNA.

In accordance with the protein binding properties of repetitive motifs, we found that intronic R0 motifs in the *Firre* locus recruit important transcription factors and CTCF, and exonic RRD motifs bind RNAP2 and hnRNPU. Repetitive CTCF binding property is critical because we have previously shown that Firre plays a necessary role in the three dimensional interactions of multiple loci across chromosomes, forming a regulatory sub-nuclear domain[45]. Previously, a few CTCF motifs have been identified[53-56]. Interestingly, R0 does not contain the canonical CTCF binding motif; however, the ChIP studies that identified this motif might have been biased for open chromatin[57], or there might be auxiliary factors that determine the binding of CTCF, or CTCF might be forming a loop, which would be captured by ChIP but give incomplete understanding of CTCF sites. The latter probability is intriguing because we have identified the same CTCF binding tracks at the trans sites that Firre interacts with in both human and mouse ES cells, suggesting that these three dimensional interactions might be, in part, CTCF-mediated and conserved.

RRD, on the other hand, specifically binds RNAP2, suggesting, along with other evidence, that Firre is actively transcribed in males and females (from both X chromosomes). Given that the R0-RRD sequential genomic structure and the binding of CTCF and RNAP2, respectively, are maintained throughout the *Firre* locus, it is likely that CTCF might be actively playing a role in nucleosome repositioning or eviction and regulating active transcription, as has been previously suggested for other loci[58,59].

Besides binding to RNAP2, RRD RNA motif in Firre also associates with hnRNPU. This species-specific interaction results in the nuclear localization of Firre transcripts, as well

as any transcript that includes the RRD motif. The species-specific interaction suggests a model, in which the RNA and the protein partner have co-evolved from rodents to primates to maintain this functional unit. This finding presents an exciting possibility for deciphering a new RNA code for nuclear localization. Although many nuclear RNAs, in fact lncRNAs, have been studied in-depth, the specific domains that retain the RNAs in the nucleus remain unknown. Not only the repeat-mediated localization of a lncRNA in the nucleus, but also the repeat-mediated formation of a nuclear sub-compartment by a lncRNA offers new mechanisms of lncRNA functions. Furthermore, our data also emphasize the potential roles of local repetitive motifs for lncRNAs, which are normally not taken into account in genome-wide studies. Examination of these repetitive sequences will require additional computational and experimental analyses but will give us more insight into the evolution and regulation of lncRNAs and their mechanisms.

4.4 Materials and Methods

4.4.1 Pipeline for surveying the landscape of novel local and tandem repeats.

For each gene, we masked out the transposable elements annotated in the RepeatMasker file from the UCSC genome browser. Next, we used to RepeatScout to *de novo* find repeats in this repeat masked gene sequence. To get only the local repeats we used Tandem Repeat Finder to remove any tandem repeats from the set discovered by RepeatScout. Finally, to get all instances of a given local repeat, we mapped our local repeat catalog to the human genome using RepeatMasker. Separately, we used Tandem Repeat Finder to find all tandem repeats in the masked gene sequence and compile a catalog of tandem repeats.

4.4.2 Statistical Tests

The lncRNA annotation file was shuffled in two ways to get separate control sets. In the first case, the annotation file was shuffled to allow the new regions to be anywhere in the genome (shuffled). In the second case, the annotation file to only fall in unannotated intergenic regions of the genome in order to compare the repeat distribution of lncRNAs with other random intergenic regions (shuffled intergenic). Local and tandem repeats were found as described above in both these sets and the numbers in each set were compared separately to lncRNAs and mRNAs. To compare the number of repeats in any two sets, we used the Mann Whitney test.

4.4.3 Multi-mapping reads

While analyzing interactions at repetitive regions, it is very important to carefully interpret multi-mapping reads. We ask Segemehl to allow a large number (100,000) of seed alignments but only output 20 best alignments for each read. Downstream, in order to count the number of reads mapping to a particular region, we normalized the reads by the number of

locations they align to. For example, a read mapping to 20 positions in the genome, will be counted as $1/20^{th}$ at each position. Such an approach has been used in several papers previously to analyze reads at repetitive sequences.

4.4.4 ChIP-Seq analysis

First, we downloaded fastq files of ChIP-Seq reads generated by the ENCODE consortium from UCSC for CTCF, YY1, Sp1, Sp2 and Pol2. Next, we used the short read mapper Segemehl to map the reads to the genome paying special attention to the multi mapping reads. The alignments generated by Segemehl were used to plot coverage of the reads over a repeat region as well as compute enrichment over it. We computed the number of reads of the TF mapping to a given repeat and divided it by the average of the number of reads mapping to the given repeat if we randomly shuffled the reads around 100 times. This number was finally used to calculate a p-value for the enrichment assuming a Poisson background model.

4.4.5 Phylogenetic Trees

We built a multiple sequence alignment (MSA) of the input sequences using MAFFT run with default parameters. Using this MSA, we constructed a phylogenetic tree using a Neighbor Joining (NJ) method. In order to calculate the confidence for each branch, we used a bootstrapping approach and reported the branches with >50% confidence in the bootstraps.

4.4.6 Cloning Sox2 constructs

Lentiviral vector was modified in house for lncRNA overexpressions. Briefly, we removed the WPRE element, the SV40 promoter, and the blasticidin gene, keeping the gateway tails the same, to prevent any interference with the lncRNA structure and function. We termed this vector lncXpress. Then each consensus species RRD was amplified by PCR, and Gibson

tails were added to RRDs to clone at the 3' end of Sox2 using Gibson cloning.

4.4.7 Viral transductions

mLFs and hFFs were split into 12-well dishes (2 wells per condition to check for RNA levels and to do FISH), 80,000 per well, and equal volumes of the virus (same titer for each) was added at the time of the split. The untransduced control was used to normalize the overexpression values for RT-PCR. Each experiment was repeated 3 times on different days using different passage number for the cells.

Total RNA was extracted on day 3 to check for expression levels, and one well was split onto two-chamber cover glasses to grow overnight. The next day cells were fixed with 4% formaldehyde for 10 minutes at room temperature.

4.4.8 RNA FISH

The protocol outlined in *3.4.5* was followed. The Sox2 exon probes were conjugated to Alexa 594. The spots for transcripts were counted using StarSearch: http://rajlab.seas.upenn.edu/StarSearch/launch.html.

4.4.9 Reverse transcription and RT-PCR

For these steps, the protocols outlined in *3.4.2* and *3.4.15* were followed.

4.4.10 Biochemical fractionation

For this step, the protocol outlines in *3.4.4* was followed.

4.4.11 hnRNPU knockdown

For this step, the protocol outlined in *3.4.14* was followed.

4.4.12 hnRNPU purification

Human and mouse hnRNPU were cloned into the pLenti6/capTEV™ -NT-DEST1 vector, which has 6X His, TEV, and BioEase tags at the N terminal of the cDNA. The

constructs were transiently expressed in HEK293FTs (ATCC: CRL-1573), in 15 cm dishes, using 90 µl Lipofectamine 2000 and 40 µg DNA. HEK293s were grown as described in *3.4.9*, and the lysates were collected after 2 days by lysing the cells as described in *3.4.10*. The purification was done following the steps outlined in the NativePure Affinity Purification Kit (Invitrogen # BN3003, BN3006). The only modification was the lysis step; therefore, the following wash after the incubation of the lysate with the streptavidin beads was done using our lysis buffer instead of the one described in the kit. In addition, MyOne Streptavidin T1 beads (Life Technologies, #65601) were used instead of the Streptavidin agarose beads. Final modification was the protein concentration step. The protein samples were concentrated using the Millipore Amicon ultra-centrifugal filter units (30K). The purity of the samples and every fraction collected during the purification was checked by running 15-20 µg of the protein on 4-12% gradient Bis-Tris gels and staining with Sypro Red as described (Life Technologies, #S-12000).

4.4.13 Electrophoretic mobility shift assay

RRDs were *in vitro* transcribed as outlined in *3.4.10*. The shift assay was performed with the purified components according to the protocol outlined in LightShift RNA EMSA kit (Thermo Scientific, #20158).

4.4.14 SHAPE-seq

For SHAPE-seq, the previously published protocol was followed[49]. Briefly, RNAs were generated by *in vitro* transcription with T7 RNA polymerase and purified by running them on a denaturing gel (8% polyacrylamide, 7 M urea, 35 W, 3 h), followed by excision, passive elution, and ethanol precipitation. Then, all RNAs were folded and divided into 2 fractions, one of which was modified with 1-methyl-7-nitroisatoic anhydride (1M7) (6.5 mM final) and the

other with DMSO. The modification was pursued in 70 s at 37°C. For library preparation, linker sequences were added to RNAs, using T4 RNA ligase. After recovering the RNAs, RT was carried out as described in Mortimer et al[50]. The libraries were prepared and quality-checked using an Agilent Bioanalyzer 2100 high-sensitivity DNA chip to compare 9 and 12 cycle amplification, or using fluorescently labeled PCR primers to analyzed fragments by capillary electrophoresis. The sequencing was done according to Illumina protocols. Spats (spats.sourceforge.net) was used to analyze the sequencing results, which performs a bioinformatics read alignment and a maximum-likelihood-based signal decay correction to calculate SHAPE-Seq θ values for each nucleotide of an RNA[50-53]. The final structure prediction was calculated using *Fold* software package[54].

4.5 References

1 Johnsson, P., L. Lipovich, D. Grander, and K. V. Morris. 2014. 'Evolutionary conservation of long non-coding RNAs; sequence, structure, function', *Biochim Biophys Acta*, 1840: 1063-71.

2 Guttman, M., I. Amit, M. Garber, C. French, M. F. Lin, D. Feldser, M. Huarte, O. Zuk, B. W. Carey, J. P. Cassady, M. N. Cabili, R. Jaenisch, T. S. Mikkelsen, T. Jacks, N. Hacohen, B. E. Bernstein, M. Kellis, A. Regev, J. L. Rinn, and E. S. Lander. 2009. 'Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals', *Nature*, 458: 223-7.

3 Cabili, M. N., C. Trapnell, L. Goff, M. Koziol, B. Tazon-Vega, A. Regev, and J. L. Rinn. 2011. 'Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses', *Genes Dev*, 25: 1915-27.

4 Derrien, T., R. Johnson, G. Bussotti, A. Tanzer, S. Djebali, H. Tilgner, G. Guernec, D. Martin, A. Merkel, D. G. Knowles, J. Lagarde, L. Veeravalli, X. Ruan, Y. Ruan, T. Lassmann, P. Carninci, J. B. Brown, L. Lipovich, J. M. Gonzalez, M. Thomas, C. A. Davis, R. Shiekhattar, T. R. Gingeras, T. J. Hubbard, C. Notredame, J. Harrow, and R. Guigo. 2012. 'The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression', *Genome Res*, 22: 1775-89.

5 Haerty, W., and C. P. Ponting. 2013. 'Mutations within lncRNAs are effectively selected against in fruitfly but not in human', *Genome Biol*, 14: R49.

6 Washietl, S., M. Kellis, and M. Garber. 2014. 'Evolutionary dynamics and tissue specificity of human long noncoding RNAs in six mammals', *Genome Res*, 24: 616-28.

7 Necsulea, A., M. Soumillon, M. Warnefors, A. Liechti, T. Daish, U. Zeller, J. C. Baker, F. Grutzner, and H. Kaessmann. 2014. 'The evolution of lncRNA repertoires and expression patterns in tetrapods', *Nature*, 505: 635-40.

8 Kutter, C., S. Watt, K. Stefflova, M. D. Wilson, A. Goncalves, C. P. Ponting, D. T. Odom, and A. C. Marques. 2012. 'Rapid turnover of long noncoding RNAs and the evolution of gene expression', *PLoS Genet*, 8: e1002841.

9 Marques, A. C., J. Hughes, B. Graham, M. S. Kowalczyk, D. R. Higgs, and C. P. Ponting. 2013. 'Chromatin signatures at transcriptional start sites separate two equally populated yet distinct classes of intergenic long noncoding RNAs', *Genome Biol*, 14: R131.

10 Pang, K. C., M. C. Frith, and J. S. Mattick. 2006. 'Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function', *Trends Genet*, 22: 1-5.

11 Engstrom, P. G., H. Suzuki, N. Ninomiya, A. Akalin, L. Sessa, G. Lavorgna, A. Brozzi, L. Luzi, S. L. Tan, L. Yang, G. Kunarso, E. L. Ng, S. Batalov, C. Wahlestedt, C. Kai, J.

Kawai, P. Carninci, Y. Hayashizaki, C. Wells, V. B. Bajic, V. Orlando, J. F. Reid, B. Lenhard, and L. Lipovich. 2006. 'Complex Loci in human and mouse genomes', *PLoS Genet*, 2: e47.

12 Chen, J., M. Sun, W. J. Kent, X. Huang, H. Xie, W. Wang, G. Zhou, R. Z. Shi, and J. D. Rowley. 2004. 'Over 20% of human transcripts might form sense-antisense pairs', *Nucleic Acids Res*, 32: 4812-20.

13 Katayama, S., Y. Tomaru, T. Kasukawa, K. Waki, M. Nakanishi, M. Nakamura, H. Nishida, C. C. Yap, M. Suzuki, J. Kawai, H. Suzuki, P. Carninci, Y. Hayashizaki, C. Wells, M. Frith, T. Ravasi, K. C. Pang, J. Hallinan, J. Mattick, D. A. Hume, L. Lipovich, S. Batalov, P. G. Engstrom, Y. Mizuno, M. A. Faghihi, A. Sandelin, A. M. Chalk, S. Mottagui-Tabar, Z. Liang, B. Lenhard, C. Wahlestedt, Riken Genome Exploration Research Group, Group Genome Science, and Fantom Consortium. 2005. 'Antisense transcription in the mammalian transcriptome', *Science*, 309: 1564-6.

14 Kung, J. T., D. Colognori, and J. T. Lee. 2013. 'Long noncoding RNAs: past, present, and future', *Genetics*, 193: 651-69.

15 Lanz, R. B., N. J. McKenna, S. A. Onate, U. Albrecht, J. Wong, S. Y. Tsai, M. J. Tsai, and B. W. O'Malley. 1999. 'A steroid receptor coactivator, SRA, functions as an RNA and is present in an SRC-1 complex', *Cell*, 97: 17-27.

16 Novikova, I. V., S. P. Hennelly, and K. Y. Sanbonmatsu. 2012. 'Structural architecture of the human long non-coding RNA, steroid receptor RNA activator', *Nucleic Acids Res*, 40: 5034-51.

17 Brown, C. J., A. Ballabio, J. L. Rupert, R. G. Lafreniere, M. Grompe, R. Tonlorenzi, and H. F. Willard. 1991. 'A gene from the region of the human X inactivation centre is expressed exclusively from the inactive X chromosome', *Nature*, 349: 38-44.

18 Jeon, Y., and J. T. Lee. 2011. 'YY1 tethers Xist RNA to the inactive X nucleation center', *Cell*, 146: 119-33.

19 Penny, G. D., G. F. Kay, S. A. Sheardown, S. Rastan, and N. Brockdorff. 1996. 'Requirement for Xist in X chromosome inactivation', *Nature*, 379: 131-7.

20 Lucchesi, J. C., W. G. Kelly, and B. Panning. 2005. 'Chromatin remodeling in dosage compensation', *Annu Rev Genet*, 39: 615-51.

21 Zhang, L. F., K. D. Huynh, and J. T. Lee. 2007. 'Perinucleolar targeting of the inactive X during S phase: evidence for a role in the maintenance of silencing', *Cell*, 129: 693-706.

22 Romito, A., and C. Rougeulle. 2011. 'Origin and evolution of the long non-coding genes in the X-inactivation center', *Biochimie*, 93: 1935-42.

23 Ogawa, Y., B. K. Sun, and J. T. Lee. 2008. 'Intersection of the RNA interference and X-inactivation pathways', *Science*, 320: 1336-41.

24 Brown, C. J., B. D. Hendrich, J. L. Rupert, R. G. Lafreniere, Y. Xing, J. Lawrence, and H. F. Willard. 1992. 'The human XIST gene: analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus', *Cell*, 71: 527-42.

25 Brockdorff, N. 2002. 'X-chromosome inactivation: closing in on proteins that bind Xist RNA', *Trends Genet*, 18: 352-8.

26 Maenner, S., M. Blaud, L. Fouillen, A. Savoye, V. Marchand, A. Dubois, S. Sanglier-Cianferani, A. Van Dorsselaer, P. Clerc, P. Avner, A. Visvikis, and C. Branlant. 2010. '2-D structure of the A region of Xist RNA and its implication for PRC2 association', *PLoS Biol*, 8: e1000276.

27 Zhao, J., B. K. Sun, J. A. Erwin, J. J. Song, and J. T. Lee. 2008. 'Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome', *Science*, 322: 750-6.

28 Guerzoni, D., and A. McLysaght. 2011. 'De novo origins of human genes', *PLoS Genet*, 7: e1002381.

29 Murphy, D. N., and A. McLysaght. 2012. 'De novo origin of protein-coding genes in murine rodents', *PLoS One*, 7: e48650.

30 Marques, A. C., and C. P. Ponting. 2014. 'Intergenic lncRNAs and the evolution of gene expression', *Curr Opin Genet Dev*, 27: 48-53.

31 Kapusta, A., Z. Kronenberg, V. J. Lynch, X. Zhuo, L. Ramsay, G. Bourque, M. Yandell, and C. Feschotte. 2013. 'Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs', *PLoS Genet*, 9: e1003470.

32 Kelley, D., and J. Rinn. 2012. 'Transposable elements reveal a stem cell-specific class of long noncoding RNAs', *Genome Biol*, 13: R107.

33 de Koning, A. P., W. Gu, T. A. Castoe, M. A. Batzer, and D. D. Pollock. 2011. 'Repetitive elements may comprise over two-thirds of the human genome', *PLoS Genet*, 7: e1002384.

34 Chadwick, B. P. 2008. 'DXZ4 chromatin adopts an opposing conformation to that of the surrounding chromosome and acquires a novel inactive X-specific role involving CTCF and antisense transcripts', *Genome Res*, 18: 1259-69.

35 Snider, L., A. Asawachaicharn, A. E. Tyler, L. N. Geng, L. M. Petek, L. Maves, D. G. Miller, R. J. Lemmers, S. T. Winokur, R. Tawil, S. M. van der Maarel, G. N. Filippova,

and S. J. Tapscott. 2009. 'RNA transcripts, miRNA-sized fragments and proteins produced from D4Z4 units: new candidates for the pathophysiology of facioscapulohumeral dystrophy', *Hum Mol Genet*, 18: 2414-30.

36 Johnson, R., and R. Guigo. 2014. 'The RIDL hypothesis: transposable elements as functional domains of long noncoding RNAs', *RNA*, 20: 959-76.

37 Nishizawa, M., and K. Nishizawa. 1999. 'Local-scale repetitiveness in amino acid use in eukaryote protein sequences: a genomic factor in protein evolution', *Proteins*, 37: 284-92.

38 Blackburn, E. H. 1984. 'The molecular structure of centromeres and telomeres', *Annu Rev Biochem*, 53: 163-94.

39 Yang, Q., Q. A. Ye, and Y. Liu. 2015. 'Mechanism of siRNA production from repetitive DNA', *Genes Dev*, 29: 526-37.

40 Blackwell, B. J., M. F. Lopez, J. Wang, B. Krastins, D. Sarracino, J. R. Tollervey, M. Dobke, I. K. Jordan, and V. V. Lunyak. 2012. 'Protein interactions with piALU RNA indicates putative participation of retroRNA in the cell cycle, DNA repair and chromatin assembly', *Mob Genet Elements*, 2: 26-35.

41 Goodier, J. L., L. E. Cheung, and H. H. Kazazian, Jr. 2013. 'Mapping the LINE1 ORF1 protein interactome reveals associated inhibitors of human retrotransposition', *Nucleic Acids Res*, 41: 7401-19.

42 Mariner, P. D., R. D. Walters, C. A. Espinoza, L. F. Drullinger, S. D. Wagner, J. F. Kugel, and J. A. Goodrich. 2008. 'Human Alu RNA is a modular transacting repressor of mRNA transcription during heat shock', *Mol Cell*, 29: 499-509.

43 Carrieri, C., L. Cimatti, M. Biagioli, A. Beugnet, S. Zucchelli, S. Fedele, E. Pesce, I. Ferrer, L. Collavin, C. Santoro, A. R. Forrest, P. Carninci, S. Biffo, E. Stupka, and Gustincich. 2012. 'Long non-coding antisense RNA controls Uchl1 translation through an embedded SINEB2 repeat', *Nature*, 491: 454-7.

44 Holdt, L. M., S. Hoffmann, K. Sass, D. Langenberger, M. Scholz, K. Krohn, K.Finstermeier, A. Stahringer, W. Wilfert, F. Beutner, S. Gielen, G. Schuler, G. Gabel, H. Bergert, I. Bechmann, P. F. Stadler, J. Thiery, and D. Teupser. 2013. 'Alu elements in ANRIL non-coding RNA at chromosome 9p21 modulate atherogenic cell functions through trans-regulation of gene networks', *PLoS Genet*, 9: e1003588.

45 Hacisuleyman, E., L. A. Goff, C. Trapnell, A. Williams, J. Henao-Mejia, L. Sun, P. McClanahan, D. G. Hendrickson, M. Sauvageau, D. R. Kelley, M. Morse, J. Engreitz, E. S. Lander, M. Guttman, H. F. Lodish, R. Flavell, A. Raj, and J. L. Rinn. 2014. 'Topological organization of multichromosomal regions by the long intergenic noncoding RNA Firre', *Nat Struct Mol Biol*, 21: 198-206.

46 Junier, I., R. K. Dale, C. Hou, F. Kepes, and A. Dean. 2012. 'CTCF-mediated transcriptional regulation through cell type-specific chromosome organization in the beta-globin locus', *Nucleic Acids Res*, 40: 7718-27.

47 Lee, B. K., and V. R. Iyer. 2012. 'Genome-wide studies of CCCTC-binding factor (CTCF) and cohesin provide insight into chroma tin structure and regulation', *J Biol Chem*, 287: 30906-13.

48 Raj, A., P. van den Bogaard, S. A. Rifkin, A. van Oudenaarden, and S. Tyagi. 2008. 'Imaging individual mRNA molecules using multiple singly labeled probes', *Nat Methods*, 5: 877-9.

49 Loughrey, D., K. E. Watters, A. H. Settle, and J. B. Lucks. 2014. 'SHAPE-Seq 2.0: systematic optimization and extension of high-throughput chemical probing of RNA secondary structure with next generation sequencing', *Nucleic Acids Res*, 42.

50 Haerty, W., and C. P. Ponting. 2015. 'Unexpected selection to retain high GC content and splicing enhancers within exons of multiexonic lncRNA loci', *RNA*.

51 Loewer, S., M. N. Cabili, M. Guttman, Y. H. Loh, K. Thomas, I. H. Park, M. Garber, M. Curran, T. Onder, S. Agarwal, P. D. Manos, S. Datta, E. S. Lander, T. M. Schlaeger, G. Q. Daley, and J. L. Rinn. 2010. 'Large intergenic non-coding RNA-RoR modulates reprogramming of human induced pluripotent stem cells', *Nat Genet*, 42: 1113-7.

52 Vucicevic, D., O. Corradin, E. Ntini, P. C. Scacheri, and U. A. Orom. 2015. 'Long ncRNA expression associates with tissue-specific enhancers', *Cell Cycle*, 14: 253-60.

53 Chen, X., H. Xu, P. Yuan, F. Fang, M. Huss, V. B. Vega, E. Wong, Y. L. Orlov, W. Zhang, J. Jiang, Y. H. Loh, H. C. Yeo, Z. X. Yeo, V. Narang, K. R. Govindarajan, B. Leong, A. Shahab, Y. Ruan, G. Bourque, W. K. Sung, N. D. Clarke, C. L. Wei, and H. H. Ng. 2008. 'Integration of external signaling pathways with the core transcriptional network in embryonic stem cells', *Cell*, 133: 1106-17.

54 Holohan, E. E., C. Kwong, B. Adryan, M. Bartkuhn, M. Herold, R. Renkawitz, S. Russell, and R. White. 2007. 'CTCF genomic binding sites in Drosophila and the organisation of the bithorax complex', *PLoS Genet*, 3: e112.

55 Wendt, K. S., K. Yoshida, T. Itoh, M. Bando, B. Koch, E. Schirghuber, S. Tsutsumi, G. Nagae, K. Ishihara, T. Mishiro, K. Yahata, F. Imamoto, H. Aburatani, M. Nakao, N. Imamoto, K. Maeshima, K. Shirahige, and J. M. Peters. 2008. 'Cohesin mediates transcriptional insulation by CCCTC-binding factor', *Nature*, 451: 796-801.

56 Bartkuhn, M., T. Straub, M. Herold, M. Herrmann, C. Rathke, H. Saumweber, G. D. Gilfillan, P. B. Becker, and R. Renkawitz. 2009. 'Active promoters and insulators are marked by the centrosomal protein 190', *EMBO J*, 28: 877-88.

57 Teytelman, L., B. Ozaydin, O. Zill, P. Lefrancois, M. Snyder, J. Rine, and M. B. Eisen. 2009. 'Impact of chromatin structures on DNA processing for genomic analyses', *PLoS One*, 4: e6700.

58 Fu, Y., M. Sinha, C. L. Peterson, and Z. Weng. 2008. 'The insulator binding protein CTCF positions 20 nucleosomes around its binding sites across the human genome', *PLoS Genet*, 4: e1000138.

59 Davey, C., R. Fraser, M. Smolle, M. W. Simmen, and J. Allan. 2003. 'Nucleosome positioning signals in the DNA sequence of the human and mouse H19 imprinting control regions', *J Mol Biol*, 325: 873-87.

60 Mortimer, S. A., C. Trapnell, S. Aviran, L. Pachter, and J. B. Lucks. 2012. 'SHAPE-Seq: High-Throughput RNA Structure Analysis', *Curr Protoc Chem Biol*, 4: 275-97.

61 Lucks, J. B., S. A. Mortimer, C. Trapnell, S. Luo, S. Aviran, G. P. Schroth, L. Pachter, J. A. Doudna, and A. P. Arkin. 2011. 'Multiplexed RNA structure characterization with selective 2'-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq)', *Proc Natl Acad Sci U S A*, 108: 11063-8.

62 Aviran, S., C. Trapnell, J. B. Lucks, S. A. Mortimer, S. Luo, G. P. Schroth, J. A. Doudna, A. P. Arkin, and L. Pachter. 2011. 'Modeling and automation of sequencing-based characterization of RNA structure', *Proc Natl Acad Sci U S A*, 108: 11069-74.

63 Aviran, S., J.B Lucks, L. Pachter. 2011. 'RNA Structure characterization from chemical mapping experiments', *In: 49th Allerton Conference, UIUC Illinois*, doi: 10.1109/Allerton. 2011.6120379.

64 Reuter, J. S., and D. H. Mathews. 2010. 'RNAstructure: software for RNA secondary structure prediction and analysis', *BMC Bioinformatics*, 11: 129.

Chapter 5: Conclusions and future perspectives

5.1 RNA is a versatile molecule

Next generation sequencing approaches have revealed the complexity of the transcriptome across species. One of the most complex properties of the transcriptome turned out to be the nearly ubiquitous expression of long RNAs that do not code for proteins, which reformed the understanding of the function of eukaryotic genomes. Although for a long time efforts have focused on protein-coding genes to dissect cell type specific gene expression and differentiation, recently it became clear that this is a very incomplete view. The noncoding genome, in fact, provides an important layer of control at every step of gene regulation, from three-dimensional compaction of the genome to post-translational modification[1-15].

With the discovery of some of the proposed roles for lncRNAs, the central dogma shifted from being protein-centric to being more receptive to the versatility of RNA. In addition to facilitating the roles of proteins, such as by guiding them to specific loci, RNA itself can catalyze reactions. The unit of the ribosome that catalyzes the peptide bond formation, for instance, is composed entirely of RNA[16]. Detailed structural and kinetic studies have revealed that the ribosomal proteins act ac accessory components to aid in the proper folding of the ribosome not in its catalytic action[16-18], suggesting that there might have existed an RNA world that preceded the protein one. Ribosome is not the only example for catalytic RNAs; group I and II self-splicing introns are large and complex ribozymes that are responsible for catalyzing and regulating splicing events as well as creating genomic diversity by dispersing and migrating themselves throughout the genome[19-22]. Finally, another class of RNAs that prove that they can perform sophisticated functions without the assistance of proteins is the riboswitches. Riboswitches are regulatory RNAs that rearrange their structures upon binding to

a ligand and in return affect gene activation or repression by providing a binding site for RNAP2, creating a premature termination site, obscuring or introducing a translation initiation/termination sequence, or self-cleaving to induce decay[23-27]. All these studies highlight the role of RNA as a key regulator for the most important housekeeping functions of a cell.

5.2 More mechanistic dissection of lncRNAs is required

Despite the discovery of thousands of lncRNAs, few steps have been taken to mechanistically characterize and categorize them. Genome-wide computational analyses, annotations, and correlation studies present compelling but not convincing arguments to show that lncRNAs are functional. Xist has been a great example of how difficult but important it is to dissect sequences properties, binding partners, localization dynamics, and structural elements of a lncRNA to understand its function. Even after 25 years, there is very little agreement on how Xist exactly functions on the molecular level, suggesting that roles of lncRNAs are more intricate than just acting as guides for protein epigenetic regulatory complexes.

Recent studies have been very encouraging, as they have mostly focused on a mechanistic understanding of the lncRNAs of interest. For example, biochemical follow-up on sequencing efforts and loss of function studies of particular lncRNAs paved the way for the discovery of a new class of coding and noncoding RNAs: circular RNAs. It still remains to be determined in which contexts they are important, and whether they impact gene expression via titrating miRNAs, or competing with the splicing machinery, or by actually being translated into proteins[28-31]. In line with these studies, it is important to note that more effort should be spent to understand the role of lncRNAs in the cytoplasm although a majority of them have been found to play a role in the nucleus. Using a newly developed method, termed ribosome profiling, it was found that there is pervasive translation outside the annotated coding regions in

the genome, especially under stress conditions[32]. This critical finding suggests that although lncRNAs might not code for functional protein products[33], their association with the ribosome can still have regulatory outcomes for translation and RNA structure, stability, and localization, and studying the interaction of lncRNAs with the ribosome might enlighten certain aspects of translation initiation. Investigating cytoplasmic lncRNAs and their ribosome occupancy can also inform about the evolution of new noncoding sequences that are produced from coding genes that lost their coding capacity or vice versa. In addition to testing the full lncRNA sequences, what still remains as a challenge is to discover functional micro-peptides that are encoded within lncRNAs. In an exemplary study, Anderson et al. identified a conserved peptide, MLN, within a lncRNA and showed that the peptide is the functional unit that regulates Ca^{2+} uptake in the sarcoplasmic reticulum in skeletal muscle[34].

Similar mechanistic studies are being applied to the major fraction of lncRNAs, the nuclear lncRNAs. For instance, by studying one locus in depth, Ingrid Grummt's lab has identified a novel mechanism of action for lncRNAs: lncRNAs can form triplex interactions with double stranded DNA to recruit protein factors and regulate gene expression. They initially discovered this phenomenon in the context of a promoter-associated lncRNA that binds to the repetitive ribosomal DNA gene, upon which it recruits the DNA methyl transferase DNMT3b, which is incumbent upon the triplex interaction to bind and silence this locus[6]. This finding opened up the way for similar discoveries at many other loci in the genome, suggesting a direct role for the RNA in directing genome regulation.

Another biochemical characterization of nuclear lncRNAs proved that the functional outcome for transcriptional regulation is attributed to the RNA, not to transcription or the DNA locus. The repetitive SINE elements that diverge in sequence, mouse B2 and human ALU, have

been found to interact with RNAP2, and thereby inhibit transcription as a general mechanism during heat shock[35-39]. Specifically, B2 binds the DNA cleft and prevents the phosphorylation of the C-terminal domain on the largest subunit of RNAP2[37]. Similarly, ALU assembles into the pre-initiation complex and blocks RNA synthesis, suggesting conservation of structure and function without conservation of sequence[38]. Given the abundance of these repetitive elements in the genome, future work should focus on finding similar structural motifs, which would further inform mechanistic principles of transcription. These seminal biochemical and structural studies shed light on the evolutionary dynamics of repetitive motifs of lncRNAs, their regulatory roles in transcription, and the cross-talk of RNA polymerases via the production of lncRNAs.

5.3 Detailed examination of lncRNA mechanisms can reveal how cell identities are established

Intricate mechanistic efforts such as the ones discussed above are crucial to validate the claims of correlative studies, such as the argument that lncRNAs are expressed in a tissue and cell type specific manner. In a large pilot study, it was found that human lncRNAs promote pluripotency and specify neuronal differentiation pathways by associating with chromatin modifiers and transcription factors[40]. A similar study uncovered that lncRNAs are determinant of spatiotemporal dynamics during lineage specification in the neocortex[41]. However, a direct role for the lncRNA was not clear from these studies. On the other hand, in a follow-up study, Ng et al. showed how a lncRNA, rhabdomyosarcoma 2-associated transcript (RMST), which came up in these previous studies, specifies neuronal cell fate and is indispensible for neurogenesis. RMST specifically interacts with SOX2, one of the key factors that direct neuronal stem cell fate, and acts as its transcriptional co-regulator by providing specificity for

its promoter targets[42]. Structural studies will further inform the nature of SOX2:RMST interaction but this is one of the few mechanistic studies that clearly depicted a functional lncRNA with a brain specific expression pattern and a novel mechanism for a lncRNA that impacts transcriptional regulation and governs cell-fate determination.

The efforts to understand the functional roles of lncRNAs in specific cell or tissue types are fundamental since we still do not understand exactly how cell identities are established. DNA and histone methylation patterns are descriptive but do not exhume causative relationships; and protein factors are limited and ubiquitously expressed without any inherent specificity. Therefore, the question of how these structured cellular programs are carried out remains to be answered. Recent discoveries of the functions of the lncRNAs hint that they might bestow specificity to their protein partners and regulate genes directly *in cis* or *trans* in the cell types that are expressed. To that end, we and others aimed to find tissue-specific lncRNA regulators of metabolism, investigating their roles in adipogenesis, insulin metabolism, and maintenance of pancreatic beta cell identity[43-45]. We discovered Firre in addition to 9 other lncRNAs that are now being explored.

We focused our efforts to understand the mechanism of Firre, which paved the way for investigating a novel mechanism for a lncRNA in the nucleus and raised more questions than answers. We found that Firre RNA interacts with multiple genomic loci that play regulatory roles in adipogenesis and assembles them in a sub-compartment near its site of transcription on the X chromosome[8]. This assembly is critical for the proper regulation of one the genes that it targets, and it is being investigated how the other targets are affected in the absence of Firre. It is possible that the other genes alter their isoform preference. Furthermore, Firre mediates these interactions via its protein partner hnRNPU. This phenomenon was previously seen for *trans*

acting lncRNAs, such as Paupar, which acts both locally on the neighboring genes and distally on the regulatory elements by binding to PAX6[46], similar to the *cis* spread and *trans* effects of Firre. However, Paupar transcripts move to their targets sites; whereas, Firre brings them to its vicinity, suggesting that lncRNAs have various modes of action to modulate gene expression.

5.4 lncRNAs play key roles in nuclear organization

Our study raised more questions than answers because although more roles of nuclear lncRNAs are being uncovered, the specifics of how thousands of lncRNAs impact the precise organization in the nucleus are unknown. Neat1[47-50], Malat1[48,51,52], and Gomafu[53] are the best-characterized nuclear lncRNAs that are responsible for assembling paraspeckles, nuclear speckles, and specialized compartments, respectively. All three are nuclear-retained and conserved architectural lncRNAs that interact with specific proteins to form specialized bodies in the nucleus. In addition to affecting RNA-editing and splicing, much remains to be discovered for the *in vivo* relevance of these domains since the genetic deletions of these genes do not have any phenotypic outcome in mice. In addition to the structural lncRNAs, there are many others that affect different aspects of nuclear organization or dynamics. Another well-characterized lncRNA, CISTR-ACT exemplifies a *cis/trans* regulatory mechanism, by which the lncRNA regulates its *cis* target gene via chromosome looping and *trans* target via proximity, which results in brachydactyly upon chromosomal translocation in humans[54]. How this interaction within and across chromosomes is recruited to generate a finely tuned regulatory system, and how stochastic these interactions are still remain to be determined; however CISTR-ACT might represent a broader phenomenon of non-random and evolutionarily conserved and directed nuclear organization.

Enhancers, which are defined as units that activate their target gene's promoter

independent of orientation or location[55,56], also present an exciting area of research for lncRNAs since some of the enhancer regions are transcribed into RNAs (eRNAs)[57,58]. The mechanism that enhancers illicit a gene activation effect is thought to be by recruiting transcriptional activators and bringing them to the vicinity of the gene's promoter via chromosome looping[59-61]. Although this renders finding the targets of enhancers incredibly difficult, since there seems to be ~6 enhancers per promoter[62], eRNAs still offer a great opportunity to study physical demarcation of the genome and the cell type specific gene expression programs associated with it. Using chromatin conformation capture techniques (3C, 4C, 5C and Hi-C), it was shown that the expression of cell type specific genes correlates with high accuracy with the expression and contact frequency of cell type specific eRNAs, and these regulatory interactions cluster within subdomains in the nucleus[63-65]. Interestingly, eRNAs commonly bind cohesin, CTCF, and Mediator, all of which have been implicated in dynamic and developmentally regulated long-range interactions[66,67]. The physical contacts of eRNAs change as the cells differentiate or undergo stress, suggesting that they can mediate the organizational aspects in the nucleus that drive gene expression changes to determine cell fates. Future experiments will illuminate how the eRNAs provide specificity to the common transcriptional regulators, how their targets change upon external stimuli/differentiation, and whether all eRNA-promoter interactions are functional.

5.5 Physical partitioning of the nucleus is an organizational principle that drives proper gene regulation and cell type specific expression

Dissecting the mechanisms, by which lncRNAs impact the organization in the nucleus, is crucial because nuclear organization is key to understanding gene regulation and disease states. Considerable efforts have highlighted that chromosomes occupy defined territories and

preferentially pair up in the nucleus[68-71], and within those territories Mb-sized more structured units, termed topologically associated domains (TADs), exist[72]. TADs are highly conserved across cell types and species and determine timing of replication[73], leaving us with the pressing question of what is then different in the genome structure that results in massive distinctions. The answer lies in the sub-domains within the TADs: the position of the gene with respect to the nuclear periphery, association with the nuclear lamina, interactions within and across TADs, and alterations in the nucleosomal occupancy all factor in to determine specific regulatory outcomes. It was, in deed, predicted by polymer modeling that chromatin can acquire highly diverse configurations within a TAD although these are fluctuating rather than stable structures[74]. More molecular and genetic approaches are needed to perturb TADs (such as, via chromosomal rearrangements) in order to examine how each factor mentioned above affects the dynamics of this nonrandom organization.

Underlining the importance of genomic organization within TADs, Bau et al. showed that chromatin globules at the α-globin locus possess diverse shapes and long-range interactions in different cell lines that show a differential expression of the gene[75]. In line with these findings, Ling and colleagues found that the imprinting control region (ICR) of the *Igf2-H19* imprinted cluster interacts with the intergenic region between *Wsb1* and *Nf1* genes[76]. The maternal copy of the *H19* locus preferentially interacts with the paternal copy of the intergenic region, which is mediated by CTCF, and CTCF has previously been shown to regulate the maternal allele[77,78]. This preferential co-localization results in the up-regulation of the paternal alleles of the trans targets and suggests a potential imprinting mechanism. Intriguingly, we detect a monoallelic interaction pattern for Firre, which binds to only one allele of its trans

targets although they are expressed bi-allelically[8]. Whether the trans targets are the maternal or paternal copies remains as a future direction.

A dissection of the Xist locus gave insight into the integrity and infrastructure within TADs. In order to identify the major determinants of TAD organization, Giorgetti et al. computationally simulated the interactome of the *Tsix* TAD, associations of which are essential for the proper silencing of the X chromosome[74]. Then, they experimentally tested their computational predictions: *Xite/Tsix*, *Chic1*, and *Linx*, 3 of which are lncRNA loci, and found that these loci are the hot spots for maintaining the integrity of the TAD as well as its boundary separation from the neighboring TAD. Not so surprisingly, these loci harbor multiple CTCF and cohesin binding sites, similar to what has been observed for the *Fire* locus[8]. Furthermore, the separation between the two TADs is essential to ensure that Xist is activated from only one of the X chromosomes. Similar studies for other TADs will elucidate the necessary elements that help shape the genome.

5.6 lncRNAs and their sequence elements can elucidate the principles of nuclear organization

The identified roles and specific expression patterns of nuclear lncRNAs and the importance of nuclear organization to drive cell type specific gene expression programs merit a detailed dissection of lncRNA elements to understand nuclear dynamics. Our analysis of the Firre locus revealed a large number of repeats which bind to critical three-dimensional regulators, such as CTCF, and retain the Firre lncRNA in the nucleus. In addition, five of these repeats are conserved between human and mice, and two particular motifs show a higher conservation, similar to that of the *Firre* locus. The CTCF binding at the R0 motif, which is an intronic repeat, is conserved across human, macaque, mouse, and rat species. This finding is

supported by a recent study that embarked upon an evolutionary analysis of CTCF binding, which revealed that CTCF underlies the evolution of local chromosomal domains, suggesting a potential positive selection for the domain that Firre organizes.

Supporting the importance of repetitive motifs in lncRNA function and nuclear organization, Chen et al. found that Neat1 lncRNA regulates the retention of repeat-containing mRNAs in the nucleus[48]. Neat1 is up-regulated upon differentiation of human ES cells and assembles the paraspeckles. Nuclear-retention in the paraspeckles results in the editing of the mRNAs, providing quality control before translation and triggering other regulatory events.

Another example comes from *Plasmodium falciparum*, the transcriptome of which includes a large number of lncRNAs, especially generated from its telomere ends[79]. It was previously discovered that the precise organization in the nucleus of this protozoan is critical to its function[80]. Concordantly, Miranda et al. discovered two repetitive subtelomeric lncRNAs that nucleate a novel compartment in the nucleus[81]. This compartment consists of a multi-protein complex and interacts with histones through the stable and repetitive hairpin structures of the lncRNAs, revealing a new mechanism of gene regulation and nuclear function for *Plasmodium*.

The organization in the nucleus does not necessarily mean the formation of concentrated compartments but also the establishment of co-regulated domains that are functionally distinct. To that end, Lunyak et al. and Zuckerkandl and Cavalli proposed that transcription from interspersed repetitive sequences and their epigenetic regulation can determine boundaries across silenced and open chromatin domains[82,83].

The properties of lncRNAs make them great candidates to function in nuclear organization. The examples described above probably represent the tip of the iceberg. lncRNAs

with unique repeat motifs can specify distinct protein or DNA interactions, by means of sequence or structure. On the other hand, lncRNAs, with common repetitive motifs, can interact with the same protein factors, which will concentrate lncRNAs and associated loci in close proximity. lncRNAs, can also be tethered to their site of transcription and recruit specific proteins, by which they can seed, nucleate, and create novel bodies in the nucleus. Cajal bodies and the nucleolus exemplify this phenomenon, in which self-organization generates a high local concentration of particular DNA loci, RNAs, and proteins[84-87]. In addition, multiple repetitive motifs in the lncRNA structure render the lncRNAs modular, which allows them to scaffold and recruit distinct proteins or genomic loci. Therefore, dynamic nuclear compartments can be established by the activation or repression of lncRNAs, allowing spatial and temporal co-regulation.

5.7 Future perspectives for Firre

Many questions regarding intra and inter chromosomal interactions remain to be addressed: 1) Why do these sites associate in the nucleus; 2) What is the frequency of interactions; 3) Do these sites meet at specific time points during the cell cycle or during cellular transitions; 4) Do these chromosomal communications depend on protein factors, RNAs, transcription, or all, 5) Do these necessary factors initiate the contacts and form a sub-nuclear compartment or is the compartment pre-formed and the factors just maintain it?

In order to address all these questions and tackle the unknowns about the lncRNA field outlined in the sections above, dissecting the Firre lncRNA and its locus further might offer a good start.

One of the interesting properties of Firre is its isoform diversity. We cloned more than 50 different isoforms in mouse using cDNA from mouse pre-adipocytes, mature adipocytes,

and mES cells (Figure 3.2.1.12). This isoform diversity stems from the inclusion or exclusion of the RRD motif, ranging from one to five RRD repeats in various isoforms. Interestingly, in our clones we detected one isoform that is similar in length to most of the Firre isoforms (~1.1 kb); however, it does not house RRD. RRD is necessary to bind hnRNPU and localize Firre in the nucleus in a punctate manner, and our FISH studies did not identify an endogenously cytoplasmic Firre transcript, suggesting that either RRD-negative isoform is unstable or is expressed at very low levels, below the detection limit for FISH. Furthermore, our studies using different cell lines revealed that Firre isoforms are expressed cell type specifically. It would be informative to identify, for instance across a differentiation time course, whether different isoforms bind different proteins and how this results in a different regulatory behavior for Firre. Since the structure and the repetitive nature of each Firre isoform would be different, we believe that the *trans* targets as well as the protein binding partners will differ for each isoform, resulting in cell type specific phenotypes. The switch of RNA-protein interactions upon RNA structural changes was a previously established phenomenon[88]. Therefore, understanding this aspect about Firre is important in general for lncRNAs because they show efficient and alternative splicing, and alternatively spliced exons are known to evolve faster[89].

The isoform specific effects can also be analyzed by rescue experiments in *Firre* knockout mES cells that we generated. RNA-sequencing of wild type (wt) versus knockout (ko) cells revealed 1077 differentially expressed, including down-regulation of metabolism, mRNA processing, nuclear export pathways and up-regulation of extracellular matrix organization and cell surface receptor-ligand interaction factors ((Figure 3.2.6.1,2,3). In order to attribute these changes directly to the Firre RNA, rescue experiments should be performed, in which the Firre RNA should be overexpressed via lentiviral delivery in ko cells. Then, the

gene sets that show differential expression in wt versus ko comparison should be fed into the analysis of ko versus ko+Firre conditions. If the same gene sets that get down-regulated upon *Firre* knockout are up-regulated upon Firre overexpression in ko cells, that suggests that the RNA has a direct effect on those pathways. Furthermore, the overexpression can be performed using different isoforms of Firre to examine isoform specific effects on gene expression. It would be intriguing to compare RRD-positive versus RRD-negative isoform effects since RRD regulates nuclear localization and *trans* interactions of Fire.

Overexpression of a lncRNA by lentiviral delivery poses an impediment in dissecting the function of that lncRNA because lentiviral overexpression works by random integration into the genome. Therefore, if there is a *cis*-mediated effect, integration and expression from another locus might not result in the same phenotypic outcome. For example, this was observed for the HOTTIP lncRNA, which acts *in cis* due to its low copy number and can only activate transcription when it is actively recruited to the locus using Gal4-UAS system[90]. The *cis*-mediated *trans* effect seems to be the mechanism for Firre, which spreads in a 5 Mb window and then assembles its trans targets to its vicinity (Figure 3.2.4.1,2, 3.2.5.1,3). To that end, one of the post-docs in the lab, David M. Shechner, began developing a method using CRISPR/Cas9 to bring the naked RNA back to its locus. I joined in to further help establish this method, which turned into a publication that we have recently submitted (Appendix 8). We demonstrated that multiple RNA domains of up to 4.8 kb can be functionally appended onto the CRISPR scaffold and directed to specific loci in the genome. Future work should focus on using this method to deliver the Firre lncRNA back to its locus in ko cells. After delivery of the functional RNA unit, gene expression changes as well as the recovery of the *trans* sites should be examined since *trans* targets lose their co-localization in the absence of Firre (Figure

3.2.10.1,2). It can again be used to test the specific role of RRD by delivering RRD-positive and negative isoforms. This experiment will test the direct role of Firre RNA (and RRD) in bridging the interactions across chromosomes. Furthermore, this technique is applicable to many other lncRNAs and provides a powerful method with which to ectopically localize functional RNAs and ribonucleoprotein complexes at specified genomic loci, offering a tool to dissect and test the mechanistic hypotheses in the lncRNA field.

In line with the method that we developed, CRISPR/Cas9 can further be utilized to understand the effects of three-dimensional infrastructure of the nucleus and its dynamics on gene expression. Techniques, such as ChIP and FISH, reveal snapshots of a cell's lifetime and not the dynamics. By generating two guide RNAs with two separate stem loops that are specific for different protein modules fused to separate fluorophores (such as MS2 stem loops specific for MS2 binding protein with mCherry and PP7 for PP7 binding protein with GFP), one can target two genomic loci and watch how the two loci move and communicate under different conditions (e.g. differentiation or metabolic stress) using live cell imaging. Developing this method for Firre, for instance, will allow us to test what happens to the *trans* interactions if we differentiate mES cells into a specific lineage. Similarly, this tool can further be elaborated by designing a split GFP system, with each half of GFP attached to the protein module that would bind the guide RNA with the specific stem loop. If the two loci come in contact at any point, GFP would fluoresce. The cells with GFP fluorescence can be sorted by FACS (fluorescence activated cell sorting, flow cytometry) analyzed by single cell RNA-seq for effects on transcriptome. All these experiments would help answer: whether the sub-nuclear compartment is pre-set or only maintained by the necessary factors; how intra or inter-chromosomal contacts

behave under different conditions; and whether the nuclear structure precedes gene expression or vice versa.

In order to fully grasp the biological significance of the three-dimensional interactions after investigating them with the methods above, the *trans* interactions of Firre should also be examined in human ES cells. Following the work presented in Chapters 3 and 4, RAP and FISH analysis can be performed in human ES cells. Our analysis of the properties of the *trans* targets in both mouse and human cells revealed commonalities in transcription factor and CTCF binding, suggesting that contacts could have been preserved. If the same *trans* targets are found in human cells, then it might imply that the Firre compartment in the nucleus is an important regulatory subdomain that is conserved and CTCF-mediated in primates and rodents (Figure 4.2.4.1-4). If the *trans* targets are different, then they should be further investigated in the context of how Firre behaves or functions in human cells. This possibility would not be uncanny since Firre is a gene on the X chromosome and is previously shown to play a role in X chromosome activation (escapee genes) in human ES cells[91], which is a process that is considerably different in mouse ES cells[92].

All the future perspectives utilizing *Firre* ko cells described so far are contingent on the genomic locus or transcription from the locus not being critical for the functional roles of Firre. In order to fully differentiate between the role of RNA versus transcription or DNA elements, other methodologies need to be developed, in which the locus and transcription are left intact but RNA is depleted. The genetic deletion of the whole locus results in the abrogation of transcription and all of the genomic sequence along with potential regulatory elements. This consideration is especially critical for lncRNAs that are bi-directionally promoted with other genes, house small RNA genes, or overlap protein-coding genes. The details of different

strategies to analyze lncRNA loss-of-function and the importance of complementary methods to resolve the *in vivo* relevance of lncRNAs are discussed previously[93]. To that end, the best method to silence Firre would to use the CRISPR/Cas9 system to deliver an inactivation domain fused to dead Cas9 to the *Firre* locus, leaving the entire proper genetic context intact. This strategy would significantly deplete the levels of Firre without any off-targets, as was demonstrated for other lncRNAs[94], and allow an RNA-dependent analysis of three-dimensional interactions. Another method would be to put a polyA termination sequence within the first exon, allowing transcription and protecting the integrity of the genomic locus. Complementary efforts are needed to fully define the roles of lncRNAs.

An actual dissection of the *Firre* locus will not be possible without investigating the roles of the large number of unique repeats. Understanding the roles of these repeats, especially the conserved ones, will illuminate general principles about how they can act ac functional domains for lncRNAs. Majority of the repeats we discovered in the *Firre* locus are in the introns (Figure 4.2.3.1,2), which suggests that they might be important for the regulation of the DNA locus and not for the role of the mature transcript. In order to further understand the functional significance of these repeats, an examination similar to what we have followed for RRD should be employed (Figure 4.2.5.2-4, Figure 4.2.5.7,8, Figure 4.2.7.1-5). However, RRD is in the exons; therefore, a targeted deletion of the repeat DNA should be performed for the other repeats. Targeted deletions of each repeat can be realized using CRISPR/Cas9; we have proven the utility of this method targeting RRD. Following the deletion of each repeat, changes in gene expression, protein binding, transcript stability, nuclear localization, and *trans* targets can be analyzed. This approach will be the first instance of a comprehensive and systematic analysis of all the sequence elements in a lncRNA locus.

The studies outlined above can be scaled up to do a more thorough analysis of repetitive sequences and their roles in the three-dimensional organization of the nucleus by a novel application of the massively parallel reporter assay (MPRA)[95]. To further understand the molecular modality whereby the identified repeats confer nuclear localization, one can search for underlying sequences and structural motifs using MPRA. Briefly, mLFs will be transfected with a library of 10,000 oligos each containing a unique mutation within the repeat region. Each of the barcoded DNA oligo will be cloned at the 3'end of Sox2 in the Sox2 expression vector (as previously described in 4.2.5.2). After transfection, one can then proceed with biochemical fractionation into nuclear and cytoplasmic compartments followed by RNA-seq on both fractions, using the oligo's unique tag as readout for where each Sox2-oligo is localized within the cell. This tool makes it possible to examine thousands of mutations and separate those that maintain nuclear localization to find minimal required sequences. Moreover, this powerful assay can be leveraged to look for compensatory mutations that retain nuclear localization functionality for specific structural aspects of each repeat region.

As a parallel approach, one can determine the structure of each repeat of interest by SHAPE-seq (Figure 4.2.7.4,5). Once secondary structure predictions are made, similar experiments as above can be performed with constructs designed with compensatory mutations. Moreover, one can employ compensatory mutational analysis to test the effect of the secondary structure on nuclear localization. Finally, the secondary structures that are found to be necessary for nuclear localization can be used to search for similar motifs across the genome. Then these motifs can be tested for binding capacity to hnNRPU, which seems to be necessary for the proper localization of lncRNAs in the nucleus (Figure 3.2.8.2, Figure 4.2.6.1 and previous finding for Xist[96]). Collectively, these approaches will allow to dissect the functional

contribution of each repeat to Firre localization, nuclear organization and consequent changes in gene expression. A unique power arises from this multifaceted approach by comparing and contrasting common sequence and structural elements.

Finally, to assess the physiological relevance of Firre, a systematic phenotypic characterization of several mutant mouse models needs to be performed. We have already generated male and female mouse knockouts of *Firre*, in which the whole 110 kb locus is removed from the genome using the Cre-loxP system. It should be noted that a triple polyA transcription termination sequence should also be used to assess the true loss of function of the RNA without any effect on the genomic locus. We are currently in the process of investigating the effects of *Firre* deletion on adipose tissue and brain based on our previous work[43] and another recent study[97]. The work published by Abe at al. identified a chromosomal amplification comprising the Firre locus in the human disorder periventricular nodular heterotopia (PNH) with overlying Polymicrogyria (PMG)[97]. PNH and PMG are two etiologically heterogeneous brain malformations characterized by the presence of nodular masses of grey matter lining the lateral ventricles and by small cortical gyri separated by shallow grooves, respectively[98]. Both defects of neuronal progenitor migration/proliferation and of radial glia have been implicated in the genesis of PNH. Therefore, both loss and gain of function approaches for Firre are necessary to grasp its role in mammalian development and disease.

In order to study gain-of-function models of Firre, we have already generated a constitutive Firre transgenic strain by randomly integrating a BAC (RP23-225O12) encompassing the mouse *Firre* locus and its adjacent promoter. In addition, we are developing a Tet-on inducible Firre overexpression strain. This model will be generated by random

integration of the Firre cDNA under the control of a Tet-responsive minimal CMV promoter. We have already made the targeting vector by cloning a mouse Firre cDNA containing 5 RRD repeats in the pTRE2-puro vector and are in the process of generating this line. In addition, we will generate a Firre whole gene deletion/Firre inducible transgene double mutant strain to investigate potential ability of the Firre transgene to rescue phenotypes or gene expression patterns in deletion strains. By performing RNA-seq of tissues and cells in which Firre is deleted or amplified across numerous developmental and postnatal stages, one can gain a unique insight into the transcriptional programs perturbed. Finally, in all these strains, future work should focus on monitoring viability, fertility, developmental, and adult phenotypes, providing ground to dissect the molecular mechanism underlying the function of Firre *in vivo*.

5.8 References

1 Rinn, J. L., and H. Y. Chang. 2012. 'Genome regulation by long noncoding RNAs', *Annu Rev Biochem*, 81: 145-66.

2 Guttman, M., J. Donaghey, B. W. Carey, M. Garber, J. K. Grenier, G. Munson, G. Young, A. B. Lucas, R. Ach, L. Bruhn, X. Yang, I. Amit, A. Meissner, A. Regev, J. L. Rinn, D. E. Root, and E. S. Lander. 2011. 'lincRNAs act in the circuitry controlling pluripotency and differentiation', *Nature*, 477: 295-300.

3 Heo, J. B., and S. Sung. 2011. 'Vernalization-mediated epigenetic silencing by a long intronic noncoding RNA', *Science*, 331: 76-9.

4 Nagano, T., J. A. Mitchell, L. A. Sanz, F. M. Pauler, A. C. Ferguson-Smith, R. Feil, and P. Fraser. 2008. 'The Air noncoding RNA epigenetically silences transcription by targeting G9a to chromatin', *Science*, 322: 1717-20.

5 Pandey, R. R., T. Mondal, F. Mohammad, S. Enroth, L. Redrup, J. Komorowski, T. Nagano, D. Mancini-Dinardo, and C. Kanduri. 2008. 'Kcnq1ot1 antisense noncoding RNA mediates lineage-specific transcriptional silencing through chromatin-level regulation', *Mol Cell*, 32: 232-46.

6 Schmitz, K. M., C. Mayer, A. Postepska, and I. Grummt. 2010. 'Interaction of noncoding RNA with the rDNA promoter mediates recruitment of DNMT3b and silencing of rRNA genes', *Genes Dev*, 24: 2264-9.

7 Yao, H., K. Brick, Y. Evrard, T. Xiao, R. D. Camerini-Otero, and G. Felsenfeld. 2010. 'Mediation of CTCF transcriptional insulation by DEAD-box RNA-binding protein p68 and steroid receptor RNA activator SRA', *Genes Dev*, 24: 2543-55.

8 Hacisuleyman, E., L. A. Goff, C. Trapnell, A. Williams, J. Henao-Mejia, L. Sun, P. McClanahan, D. G. Hendrickson, M. Sauvageau, D. R. Kelley, M. Morse, J. Engreitz, E. S. Lander, M. Guttman, H. F. Lodish, R. Flavell, A. Raj, and J. L. Rinn. 2014. 'Topological organization of multichromosomal regions by the long intergenic noncoding RNA Firre', *Nat Struct Mol Biol*, 21: 198-206.

9 Hung, T., and H. Y. Chang. 2010. 'Long noncoding RNA in genome regulation: prospects and mechanisms', *RNA Biol*, 7: 582-5.

10 Bonasio, R., S. Tu, and D. Reinberg. 2010. 'Molecular signals of epigenetic states', *Science*, 330: 612-6.

11 Martianov, I., A. Ramadass, A. Serra Barros, N. Chow, and A. Akoulitchev. 2007. 'Repression of the human dihydrofolate reductase gene by a non-coding interfering

transcript', *Nature*, 445: 666-70.

12 Jeon, Y., and J. T. Lee. 2011. 'YY1 tethers Xist RNA to the inactive X nucleation center', *Cell*, 146: 119-33.

13 Gong, C., and L. E. Maquat. 2011. 'lncRNAs transactivate STAU1-mediated mRNA decay by duplexing with 3' UTRs via Alu elements', *Nature*, 470: 284-8.

14 Camblong, J., N. Iglesias, C. Fickentscher, G. Dieppois, and F. Stutz. 2007. 'Antisense RNA stabilization induces transcriptional gene silencing via histone deacetylation in S. cerevisiae', *Cell*, 131: 706-17.

15 Wang, X., S. Arai, X. Song, D. Reichart, K. Du, G. Pascual, P. Tempst, M. G. Rosenfeld, C. K. Glass, and R. Kurokawa. 2008. 'Induced ncRNAs allosterically modify RNA-binding proteins in cis to inhibit transcription', *Nature*, 454: 126-30.

16 Nissen, P., J. Hansen, N. Ban, P. B. Moore, and T. A. Steitz. 2000. 'The structural basis of ribosome activity in peptide bond synthesis', *Science*, 289: 920-30.

17 Ban, N., P. Nissen, J. Hansen, P. B. Moore, and T. A. Steitz. 2000. 'The complete atomic structure of the large ribosomal subunit at 2.4 A resolution', *Science*, 289: 905-20.

18 Schmeing, T. M., A. C. Seila, J. L. Hansen, B. Freeborn, J. K. Soukup, S. A. Scaringe, S. A. Strobel, P. B. Moore, and T. A. Steitz. 2002. 'A pre-translocational intermediate in protein synthesis observed in crystals of enzymatically active 50S subunits', *Nat Struct Biol*, 9: 225-30.

19 Nielsen, H., and S. D. Johansen. 2009. 'Group I introns: Moving in new directions', *RNA Biol*, 6: 375-83.

20 Cech, T. R. 1990. 'Self-splicing of group I introns', *Annu Rev Biochem*, 59: 543-68.

21 Koonin, E. V. 2006. 'The origin of introns and their role in eukaryogenesis: a compromise solution to the introns-early versus introns-late debate?', *Biol Direct*, 1: 22.

22 Centron, D., and P. H. Roy. 2002. 'Presence of a group II intron in a multiresistant Serratia marcescens strain that harbors three integrons and a novel gene fusion', *Antimicrob Agents Chemother*, 46: 1402-9.

23 Breaker, R. R., and G. F. Joyce. 2014. 'The expanding view of RNA and DNA function', *Chem Biol*, 21: 1059-65.

24 Tucker, B. J., and R. R. Breaker. 2005. 'Riboswitches as versatile gene control elements', *Curr Opin Struct Biol*, 15: 342-8.

25 Winkler, W. C., A. Nahvi, A. Roth, J. A. Collins, and R. R. Breaker. 2004. 'Control of gene

expression by a natural metabolite-responsive ribozyme', *Nature*, 428: 281-6.

26 Mironov, A. S., I. Gusarov, R. Rafikov, L. E. Lopez, K. Shatalin, R. A. Kreneva, D. A. Perumov, and E. Nudler. 2002. 'Sensing small molecules by nascent RNA: a mechanism to control transcription in bacteria', *Cell*, 111: 747-56.

27 Batey, R. T., S. D. Gilbert, and R. K. Montange. 2004. 'Structure of a natural guanine-responsive riboswitch complexed with the metabolite hypoxanthine', *Nature*, 432: 411-5.

28 Li, F., L. Zhang, W. Li, J. Deng, J. Zheng, M. An, J. Lu, and Y. Zhou. 2015. 'Circular RNA ITCH has inhibitory effect on ESCC by suppressing the Wnt/beta-catenin pathway', *Oncotarget*.

29 Chen, L. L., and L. Yang. 2015. 'Regulation of circRNA biogenesis', *RNA Biol*: 0.

30 Guo, J. U., V. Agarwal, H. Guo, and D. P. Bartel. 2014. 'Expanded identification and characterization of mammalian circular RNAs', *Genome Biol*, 15: 409.

31 AbouHaidar, M. G., S. Venkataraman, A. Golshani, B. Liu, and T. Ahmad. 2014. 'Novel coding, translation, and gene expression of a replicating covalently closed circular RNA of 220 nt', *Proc Natl Acad Sci U S A*, 111: 14542-7.

32 Ingolia, N. T., G. A. Brar, N. Stern-Ginossar, M. S. Harris, G. J. Talhouarne, S. E. Jackson, M. R. Wills, and J. S. Weissman. 2014. 'Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes', *Cell Rep*, 8: 1365-79.

33 Guttman, M., P. Russell, N. T. Ingolia, J. S. Weissman, and E. S. Lander. 2013. 'Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins', *Cell*, 154: 240-51.

34 Anderson, D. M., K. M. Anderson, C. L. Chang, C. A. Makarewich, B. R. Nelson, J. R. McAnally, P. Kasaragod, J. M. Shelton, J. Liou, R. Bassel-Duby, and E. N. Olson. 2015. 'A micropeptide encoded by a putative long noncoding RNA regulates muscle performance', *Cell*, 160: 595-606.

35 Allen, T. A., S. Von Kaenel, J. A. Goodrich, and J. F. Kugel. 2004. 'The SINE-encoded mouse B2 RNA represses mRNA transcription in response to heat shock', *Nat Struct Mol Biol*, 11: 816-21.

36 Espinoza, C. A., T. A. Allen, A. R. Hieb, J. F. Kugel, and J. A. Goodrich. 2004. 'B2 RNA binds directly to RNA polymerase II to repress transcript synthesis', *Nat Struct Mol Biol*, 11: 822-9.

37 Espinoza, C. A., J. A. Goodrich, and J. F. Kugel. 2007. 'Characterization of the structure, function, and mechanism of B2 RNA, an ncRNA repressor of RNA polymerase II

transcription', *RNA*, 13: 583-96.

38 Mariner, P. D., R. D. Walters, C. A. Espinoza, L. F. Drullinger, S. D. Wagner, J. F. Kugel, and J. A. Goodrich. 2008. 'Human Alu RNA is a modular transacting repressor of mRNA transcription during heat shock', *Mol Cell*, 29: 499-509.

39 Yakovchuk, P., J. A. Goodrich, and J. F. Kugel. 2009. 'B2 RNA and Alu RNA repress transcription by disrupting contacts between RNA polymerase II and promoter DNA within assembled complexes', *Proc Natl Acad Sci U S A*, 106: 5569-74.

40 Ng, S. Y., R. Johnson, and L. W. Stanton. 2012. 'Human long non-coding RNAs promote pluripotency and neuronal differentiation by association with chromatin modifiers and transcription factors', *EMBO J*, 31: 522-33.

41 Molyneaux, B. J., L. A. Goff, A. C. Brettler, H. H. Chen, J. R. Brown, S. Hrvatin, J. L. Rinn, and P. Arlotta. 2015. 'DeCoN: genome-wide analysis of in vivo transcriptional dynamics during pyramidal neuron fate selection in neocortex', *Neuron*, 85: 275-88.

42 Ng, S. Y., G. K. Bogu, B. S. Soh, and L. W. Stanton. 2013. 'The long noncoding RNA RMST interacts with SOX2 to regulate neurogenesis', *Mol Cell*, 51: 349-59.

43 Sun, L., L. A. Goff, C. Trapnell, R. Alexander, K. A. Lo, E. Hacisuleyman, M. Sauvageau, B. Tazon-Vega, D. R. Kelley, D. G. Hendrickson, B. Yuan, M. Kellis, H. F. Lodish, and J. L. Rinn. 2013. 'Long noncoding RNAs regulate adipogenesis', *Proc Natl Acad Sci U S A*, 110: 3387-92.

44 Ellis, B. C., L. D. Graham, and P. L. Molloy. 2014. 'CRNDE, a long non-coding RNA responsive to insulin/IGF signaling, regulates genes involved in central metabolism', *Biochim Biophys Acta*, 1843: 372-86.

45 Xu, B., I. Gerin, H. Miao, D. Vu-Phan, C. N. Johnson, R. Xu, X. W. Chen, W. P. Cawthorn, O. A. MacDougald, and R. J. Koenig. 2010. 'Multiple roles for the non-coding RNA SRA in regulation of adipogenesis and insulin sensitivity', *PLoS One*, 5: e14199.

46 Vance, K. W., S. N. Sansom, S. Lee, V. Chalei, L. Kong, S. E. Cooper, P. L. Oliver, and C. P. Ponting. 2014. 'The long non-coding RNA Paupar regulates the expression of both local and distal genes', *EMBO J*, 33: 296-311.

47 Chen, L. L., and G. G. Carmichael. 2009. 'Altered nuclear retention of mRNAs containing inverted repeats in human embryonic stem cells: functional role of a nuclear noncoding RNA', *Mol Cell*, 35: 467-78.

48 Clemson, C. M., J. N. Hutchinson, S. A. Sara, A. W. Ensminger, A. H. Fox, A. Chess, and J. B. Lawrence. 2009. 'An architectural role for a nuclear noncoding RNA: NEAT1 RNA is essential for the structure of paraspeckles', *Mol Cell*, 33: 717-26.

49 Sasaki, Y. T., T. Ideue, M. Sano, T. Mituyama, and T. Hirose. 2009. 'MENepsilon/beta noncoding RNAs are essential for structural integrity of nuclear paraspeckles', *Proc Natl Acad Sci U S A*, 106: 2525-30.

50 Sunwoo, H., M. E. Dinger, J. E. Wilusz, P. P. Amaral, J. S. Mattick, and D. L. Spector. 2009. 'MEN epsilon/beta nuclear-retained non-coding RNAs are up-regulated upon muscle differentiation and are essential components of paraspeckles', *Genome Res*, 19: 347-59.

51 Tripathi, V., J. D. Ellis, Z. Shen, D. Y. Song, Q. Pan, A. T. Watt, S. M. Freier, C. F. Bennett, A. Sharma, P. A. Bubulya, B. J. Blencowe, S. G. Prasanth, and K. V. Prasanth. 2010. 'The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation', *Mol Cell*, 39: 925-38.

52 Hutchinson, J. N., A. W. Ensminger, C. M. Clemson, C. R. Lynch, J. B. Lawrence, and A. Chess. 2007. 'A screen for nuclear transcripts identifies two linked noncoding RNAs associated with SC35 splicing domains', *BMC Genomics*, 8: 39.

53 Sone, M., T. Hayashi, H. Tarui, K. Agata, M. Takeichi, and S. Nakagawa. 2007. 'The mRNA-like noncoding RNA Gomafu constitutes a novel nuclear domain in a subset of neurons', *J Cell Sci*, 120: 2498-506.

54 Maass, P. G., A. Rump, H. Schulz, S. Stricker, L. Schulze, K. Platzer, A. Aydin, S. Tinschert, M. B. Goldring, F. C. Luft, and S. Bahring. 2012. 'A misplaced lncRNA causes brachydactyly in humans', *J Clin Invest*, 122: 3990-4002.

55 Banerji, J., L. Olson, and W. Schaffner. 1983. 'A lymphocyte-specific cellular enhancer is located downstream of the joining region in immunoglobulin heavy chain genes', *Cell*, 33: 729-40.

56 Gillies, S. D., S. L. Morrison, V. T. Oi, and S. Tonegawa. 1983. 'A tissue-specific transcription enhancer element is located in the major intron of a rearranged immunoglobulin heavy chain gene', *Cell*, 33: 717-28.

57 Li, W., D. Notani, Q. Ma, B. Tanasa, E. Nunez, A. Y. Chen, D. Merkurjev, J. Zhang, K. Ohgi, X. Song, S. Oh, H. S. Kim, C. K. Glass, and M. G. Rosenfeld. 2013. 'Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation', *Nature*, 498: 516-20.

58 Lai, F., U. A. Orom, M. Cesaroni, M. Beringer, D. J. Taatjes, G. A. Blobel, and R. Shiekhattar. 2013. 'Activating RNAs associate with Mediator to enhance chromatin architecture and transcription', *Nature*, 494: 497-501.

59 Maston, G. A., S. K. Evans, and M. R. Green. 2006. 'Transcriptional regulatory elements in the human genome', *Annu Rev Genomics Hum Genet*, 7: 29-59.

60 Nolis, I. K., D. J. McKay, E. Mantouvalou, S. Lomvardas, M. Merika, and D. Thanos. 2009. 'Transcription factors mediate long-range enhancer-promoter interactions', *Proc Natl Acad Sci U S A*, 106: 20222-7.

61 Dekker, J., K. Rippe, M. Dekker, and N. Kleckner. 2002. 'Capturing chromosome conformation', *Science*, 295: 1306-11.

62 Shen, Y., F. Yue, D. F. McCleary, Z. Ye, L. Edsall, S. Kuan, U. Wagner, J. Dixon, L. Lee, V. V. Lobanenkov, and B. Ren. 2012. 'A map of the cis-regulatory sequences in the mouse genome', *Nature*, 488: 116-20.

63 Li, G., X. Ruan, R. K. Auerbach, K. S. Sandhu, M. Zheng, P. Wang, H. M. Poh, Y. Goh, J. Lim, J. Zhang, H. S. Sim, S. Q. Peh, F. H. Mulawadi, C. T. Ong, Y. L. Orlov, S. Hong, Z. Zhang, S. Landt, D. Raha, G. Euskirchen, C. L. Wei, W. Ge, H. Wang, C. Davis, K. I. Fisher-Aylor, A. Mortazavi, M. Gerstein, T. Gingeras, B. Wold, Y. Sun, M. J. Fullwood, E. Cheung, E. Liu, W. K. Sung, M. Snyder, and Y. Ruan. 2012. 'Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation', *Cell*, 148: 84-98.

64 Ernst, J., P. Kheradpour, T. S. Mikkelsen, N. Shoresh, L. D. Ward, C. B. Epstein, X. Zhang, L. Wang, R. Issner, M. Coyne, M. Ku, T. Durham, M. Kellis, and B. E. Bernstein. 2011. 'Mapping and analysis of chromatin state dynamics in nine human cell types', *Nature*, 473: 43-9.

65 Ong, C. T., and V. G. Corces. 2011. 'Enhancer function: new insights into the regulation of tissue-specific gene expression', *Nat Rev Genet*, 12: 283-93.

66 Lin, Y. C., C. Benner, R. Mansson, S. Heinz, K. Miyazaki, M. Miyazaki, V. Chandra, C. Bossen, C. K. Glass, and C. Murre. 2012. 'Global changes in the nuclear positioning of genes and intra- and interdomain genomic interactions that orchestrate B cell fate', *Nat Immunol*, 13: 1196-204.

67 Kagey, M. H., J. J. Newman, S. Bilodeau, Y. Zhan, D. A. Orlando, N. L. van Berkum, C. C. Ebmeier, J. Goossens, P. B. Rahl, S. S. Levine, D. J. Taatjes, J. Dekker, and R. A. Young. 2010. 'Mediator and cohesin connect gene expression and chromatin architecture', *Nature*, 467: 430-5.

68 Cremer, T., G. Kreth, H. Koester, R. H. Fink, R. Heintzmann, M. Cremer, I. Solovei, D. Zink, and C. Cremer. 2000. 'Chromosome territories, interchromatin domain compartment, and nuclear matrix: an integrated view of the functional nuclear architecture', *Crit Rev Eukaryot Gene Expr*, 10: 179-212.

69 Cremer, M., J. von Hase, T. Volm, A. Brero, G. Kreth, J. Walter, C. Fischer, I. Solovei, C. Cremer, and T. Cremer. 2001. 'Non-random radial higher-order chromatin arrangements in nuclei of diploid human cells', *Chromosome Res*, 9: 541-67.

70 Dundr, M., and T. Misteli. 2001. 'Functional architecture in the cell nucleus', *Biochem J*, 356: 297-310.

71 Parada, L. A., P. G. McQueen, and T. Misteli. 2004. 'Tissue-specific spatial organization of genomes', *Genome Biol*, 5: R44.

72 Dixon, J. R., S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, M. Hu, J. S. Liu, and B. Ren. 2012. 'Topological domains in mammalian genomes identified by analysis of chromatin interactions', *Nature*, 485: 376-80.

73 Pope, B. D., T. Ryba, V. Dileep, F. Yue, W. Wu, O. Denas, D. L. Vera, Y. Wang, R. S. Hansen, T. K. Canfield, R. E. Thurman, Y. Cheng, G. Gulsoy, J. H. Dennis, M. P. Snyder, J. A. Stamatoyannopoulos, J. Taylor, R. C. Hardison, T. Kahveci, B. Ren, and D. M. Gilbert. 2014. 'Topologically associating domains are stable units of replication-timing regulation', *Nature*, 515: 402-5.

74 Giorgetti, L., R. Galupa, E. P. Nora, T. Piolot, F. Lam, J. Dekker, G. Tiana, and E. Heard. 2014. 'Predictive polymer modeling reveals coupled fluctuations in chromosome conformation and transcription', *Cell*, 157: 950-63.

75 Bau, D., A. Sanyal, B. R. Lajoie, E. Capriotti, M. Byron, J. B. Lawrence, J. Dekker, and M. A. Marti-Renom. 2011. 'The three-dimensional folding of the alpha-globin gene domain reveals formation of chromatin globules', *Nat Struct Mol Biol*, 18: 107-14.

76 Ling, J. Q., T. Li, J. F. Hu, T. H. Vu, H. L. Chen, X. W. Qiu, A. M. Cherry, and A. R. Hoffman. 2006. 'CTCF mediates interchromosomal colocalization between Igf2/H19 and Wsb1/Nf1', *Science*, 312: 269-72.

77 Bell, A. C., and G. Felsenfeld. 2000. 'Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene', *Nature*, 405: 482-5.

78 Hark, A. T., C. J. Schoenherr, D. J. Katz, R. S. Ingram, J. M. Levorse, and S. M. Tilghman. 2000. 'CTCF mediates methylation-sensitive enhancer-blocking activity at the H19/Igf2 locus', *Nature*, 405: 486-9.

79 Broadbent, K. M., D. Park, A. R. Wolf, D. Van Tyne, J. S. Sims, U. Ribacke, S. Volkman, M. Duraisingh, D. Wirth, P. C. Sabeti, and J. L. Rinn. 2011. 'A global transcriptional analysis of Plasmodium falciparum malaria reveals a novel family of telomere-associated lncRNAs', *Genome Biol*, 12: R56.

80 Scherf, A., J. J. Lopez-Rubio, and L. Riviere. 2008. 'Antigenic variation in Plasmodium falciparum', *Annu Rev Microbiol*, 62: 445-70.

81 Sierra-Miranda, M., D. M. Delgadillo, L. Mancio-Silva, M. Vargas, N. Villegas-Sepulveda, S. Martinez-Calvillo, A. Scherf, and R. Hernandez-Rivas. 2012. 'Two long non-coding RNAs generated from subtelomeric regions accumulate in a novel perinuclear

compartment in Plasmodium falciparum', *Mol Biochem Parasitol*, 185: 36-47.

82 Zuckerkandl, E., and G. Cavalli. 2007. 'Combinatorial epigenetics, "junk DNA", and the evolution of complex organisms', *Gene*, 390: 232-42.

83 Lunyak, V. V., G. G. Prefontaine, E. Nunez, T. Cramer, B. G. Ju, K. A. Ohgi, K. Hutt, R. Roy, A. Garcia-Diaz, X. Zhu, Y. Yung, L. Montoliu, C. K. Glass, and M. G. Rosenfeld. 2007. 'Developmentally regulated activation of a SINE B2 repeat as a domain boundary in organogenesis', *Science*, 317: 248-51.

84 Shevtsov, S. P., and M. Dundr. 2011. 'Nucleation of nuclear bodies by RNA', *Nat Cell Biol*, 13: 167-73.

85 Dundr, M., and T. Misteli. 2010. 'Biogenesis of nuclear bodies', *Cold Spring Harb Perspect Biol*, 2: a000711.

86 Misteli, T. 2001. 'The concept of self-organization in cellular architecture', *J Cell Biol*, 155: 181-5.

87 Kaiser, T. E., R. V. Intine, and M. Dundr. 2008. 'De novo formation of a subnuclear body', *Science*, 322: 1713-7.

88 Liu, N., Q. Dai, G. Zheng, C. He, M. Parisien, and T. Pan. 2015. 'N(6)-methyladenosine-dependent RNA structural switches regulate RNA-protein interactions', *Nature*, 518: 560-4.

89 Chen, F. C., S. S. Wang, C. J. Chen, W. H. Li, and T. J. Chuang. 2006.'Alternatively and constitutively spliced exons are subject to different evolutionary forces', *Mol Biol Evol*, 23: 675-82.

90 Wang, K. C., Y. W. Yang, B. Liu, A. Sanyal, R. Corces-Zimmerman, Y. Chen, B. R. Lajoie, A. Protacio, R. A. Flynn, R. A. Gupta, J. Wysocka, M. Lei, J. Dekker, J. A. Helms, and H. Y. Chang. 2011. 'A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression', *Nature*, 472: 120-4.

91 Horakova, A. H., S. C. Moseley, C. R. McLaughlin, D. C. Tremblay, and B. P. Chadwick. 2012. 'The macrosatellite DXZ4 mediates CTCF-dependent long-range intrachromosomal interactions on the human inactive X chromosome', *Hum Mol Genet*, 21: 4367-77.

92 Berletch, J. B., F. Yang, and C. M. Disteche. 2010. 'Escape from X inactivation in mice and humans', *Genome Biol*, 11: 213.

93 Bassett, A. R., A. Akhtar, D. P. Barlow, A. P. Bird, N. Brockdorff, D. Duboule, A. Ephrussi, A. C. Ferguson-Smith, T. R. Gingeras, W. Haerty, D. R. Higgs, E. A. Miska, and C. P. Ponting. 2014. 'Considerations when investigating lncRNA function in vivo', *Elife*, 3:

e03058.

94 Gilbert, L. A., M. A. Horlbeck, B. Adamson, J. E. Villalta, Y. Chen, E. H. Whitehead, C. Guimaraes, B. Panning, H. L. Ploegh, M. C. Bassik, L. S. Qi, M. Kampmann, and J. S. Weissman. 2014. 'Genome-Scale CRISPR-Mediated Control of Gene Repression and Activation', *Cell*, 159: 647-61.

95 Melnikov, A., X. Zhang, P. Rogov, L. Wang, and T. S. Mikkelsen. 2014. 'Massively parallel reporter assays in cultured mammalian cells', *J Vis Exp*.

96 Hasegawa, Y., N. Brockdorff, S. Kawano, K. Tsutui, K. Tsutui, and S. Nakagawa. 2010. 'The matrix protein hnRNP U is required for chromosomal localization of Xist RNA', *Dev Cell*, 19: 469-76.

97 Abe, Y., A. Kikuchi, S. Kobayashi, K. Wakusawa, S. Tanaka, T. Inui, S. Kunishima, S. Kure, and K. Haginoya. 2014. 'Xq26.1-26.2 gain identified on array comparative genomic hybridization in bilateral periventricular nodular heterotopia with overlying polymicrogyria', *Dev Med Child Neurol*, 56: 1221-4.

98 Wieck, G., R. J. Leventer, W. M. Squier, A. Jansen, E. Andermann, F. Dubeau, A. Ramazzotti, R. Guerrini, and W. B. Dobyns. 2005. 'Periventricular nodular heterotopia with overlying polymicrogyria', *Brain*, 128: 2811-21.

Table A1: The sequences of the primers used for qRT-PCR in Chapter 2

Clone ID	Fwd Primer	Rev Primer
AK079912	GGACAAGTTGCTCCTTCCTT	CAGAAGGCTTGTGTGTCAGA
Ak080070	GGGGTGGCCAAGCCTTCCTT	AGCTACACGGCTGCTGCTCC
AK040954	TCAGGAACCCAGTCCATAGT	TCCCTGGATTTAGGGTGTCT
AK045415	AGCATCCTTCTCGTCGTGCCT	AACACAGGACCCCAGGGTGG
AK041310	CGTGCAACGCCTGCTGTGAG	CGAGAGAGCGTGGGCCAGTT
AK019114	CAGGACCATCCAAAGCAGAT	CAGCAGGTGGATCTTTGTGA
AK016444	GGTCCTTAGGCAGAGTCTTG	TCCATGGAGCACAATAGCTG
AK079699	CAAAAGTGCCAGGTTTGGAC	GGGTTGGAAGTGTTCAGACA
AK165901	CTGCTGGAACTTTAACGGGA	CTCCTCCCACTTTCGTTTGT
AK133808	CTTGGAAGTTACCTCTCGGG	CCACCATGTGCTTGGAATTG
AK052674	TCTGTGCTCACAGTTTCCAG	GACCTCTCTCTGTTGGGTTC
AK045413	CTGGTCCGCCTACTTGAAAA	AGCTCCGTTTTGAGTTCTCC
AK136742	ACACGAACACACGCATACAA	TCATAACGACAGTGGTGCAG
AK050707	CCACGGCCAAGGTGTCCTCT	ACCTGGGGGAAAGGGTGCTC
AK005218	GGCAACGACTCAGAAAAAGC	TCTGAAGCAAGCCATGTTCT
AK161599	GGCAAGCCCTAAAGTTGAGA	ATAAACAGGCCCAAGAAGGG
AK017076	TGAACCTAGAAAACTGCCCC	CCTGTCTCTTCTCAGGTGTG
AK147324	GAGAATACAGCCCCAAGCAT	GCCAGTATCAGCAAGTCCAT
AK007571	TGCCAGCTCTGTGGTCCCTG	TAGGGAGTTGAGCGGCAGGC
AK047471	GCAAGTCCGCCTCACCGAGA	CAGCGTCGTTGTGTCGTGCG
Ak040027	TCATGCCGGCAGCCGAACTT	GGCTCCACTCCACACCTGCT
AK081581	AGCCATTTTGAAGCAGGGAA	CTCTGAAGGGTCAGGTGATG
Neat1	GCGAGGAGAAGCGGGGCTAA	CTGCCCCCATGTAGGCCTGT
AK030946	AGGTATGCTTCACCTCTCCT	CAAATTCAAGCAGGCAAGGG
AK029148	CAGCTGGGCCTGTGGCTAGT	TCTTCCTGCCTTGGCCTCCC
AK164174	CTGCCGGGCTGCTCCTACAT	AGCCACACCCAACTCGCTCA
Pparg	GTGCCAGTTTCGATCCGTAGA	GGCCAGCATCGTGTAGATGA
Cebpa	TGCGCAAGAGCCGAGATAAA	CCTTCTGTTGCGTCTCCACG
AdipoQ	CGATTGTCAGTGGATCTGACG	CAACAGTAGCATCCTGAGCCCT
FABP4	ACAAGCTGGTGGTGGAATGTG	CCTTTGGCTCATGCCCTTT
18S	GTAACCCGTTGAACCCCATT	CCATCCAATCGGTAGTAGC

Table A2: The sequences of siRNAs used in Chapter 2

Gene Name and Accession No	Sequence	
	sense（5'-3'）	antisense（5'-3'）
AK045415		
482	CGACGAGAAGGAUGCUAUATT	UAUAGCAUCCUUCUCGUCGTT
1157	GCUAGAGUAAGAAGUAGUATT	UACUACUUCUUACUCUAGCTT
1501	GCAACUUACUUGUCAUAAATT	UUUAUGACAAGUAAGUUGCTT
AK079699		
71	GGUUGUUUCUAAGUCACAATT	UUGUGACUUAGAAACAACCTT
735	GCUGCUGUCUGACCUAAUATT	UAUUAGGUCAGACAGCAGCTT
430	GCUUAUAAGUGUCCUGGUATT	UACCAGGACACUUAUAAGCTT
AK041310		
475	GCUCAAUGUUAAACCAUAATT	UUAUGGUUUAACAUUGAGCTT
236	CGUGCUACUUCCACAAGAATT	UUCUUGUGGAAGUAGCACGTT
155	GAUGACUCCUUUACAGUAATT	UUACUGUAAAGGAGUCAUCTT
AK133808		
608	CCUAGAACCUGAUAUUUAATT	UUAAAUAUCAGGUUCUAGGTT
853	GUUUGUCUGAAGACAGAUATT	UAUCUGUCUUCAGACAAACTT
348	GCGUAGAGCGGCGGUCUAATT	UUAGACCGCCGCUCUACGCTT
AK161599		
790	GAAUGUCUGUUCUACUAAATT	UUUAGUAGAACAGACAUUCTT
361	GGUGCGCACUAGGUUUCUATT	UAGAAACCUAGUGCGCACCTT
615	GGAAGCGCAAGAAACCGAATT	UUCGGUUUCUUGCGCUUCCTT
AK040954		
2119	GGAUAAUGAUGUAAAUAUATT	UAUAUUUACAUCAUUAUCCTT
1167	GAUAGAAGAGAAAGAUGAATT	UUCAUCUUUCUCUUCUAUCTT
474	AGCAGUUUGUUGUAGUCAATT	UUGACUACAACAAACUGCUTT
AK005218		
310	GGCUUGCUUCAGAAAGAAATT	UUUCUUUCUGAAGCAAGCCTT
143	AGCGGAGGCUGCAGGACAATT	UUGUCCUGCAGCCUCCGCUTT
51	GCUGCAGGUGUCCGUGCUATT	UAGCACGGACACCUGCAGCTT
AK007571		
591	CAUACUGUUCAUAGACUUATT	UAAGUCUAUGAACAGUAUGTT
136	GCUCAAGGGCAGAACUAAATT	UUUAGUUCUGCCCUUGAGCTT
436	GACCUUCUGAUGAACAUAATT	UUAUGUUCAUCAGAAGGUCTT
AK016444		
1334	GCAGCUAGCUCAGAGUUAATT	UUAACUCUGAGCUAGCUGCTT
1736	GAUGGAAUUUGCUGUUGAATT	UUCAACAGCAAAUUCCAUCTT
363	GAUGGAUCAAUCAAUUGAATT	UUCAAUUGAUUGAUCCAUCTT
AK017076		
1107	CCUCAGAGCUGCAGACAAATT	UUUGUCUGCAGCUCUGAGGTT
577	GGUAGCUGUCUAUGGACUATT	UAGUCCAUAGACAGCUACCTT
724	GAACAUCAGAUUUGUGUAATT	UUACACAAAUCUGAUGUUCTT
AK019114		
318	CCUUCACUAGCAAGGACAATT	UUGUCCUUGCUAGUGAAGGTT

Table A2 (continued)

189	GCUCUAUUGGAACUCUAUATT	UAUAGAGUUCCAAUAGAGCTT
279	GCUAGCAUGCAUAGCCUAATT	UUAGGCUAUGCAUGCUAGCTT
AK030946		
882	GGAUAAGGAUGGACAGAAATT	UUUCUGUCCAUCCUUAUCCTT
1100	GAAUGUCUUUCUAAACUAATT	UUAGUUUAGAAAGACAUUCTT
401	GCUCCUUGAGGAUGCUCUATT	UAGAGCAUCCUCAAGGAGCTT
AK040027		
106	GAAACAGACGAAUAAAUAATT	UUAUUUAUUCGUCUGUUUCTT
871	GGUAAUAAAGUAACUUAUATT	UAUAAGUUACUUUAUUACCTT
581	GGUUAAAGGUGAAAGUGCATT	UGCACUUUCACCUUUAACCTT
AK047471		
1237	CAUUAAGGUUAGAGGACAATT	UUGUCCUCUAACCUUAAUGTT
708	GCACAUUAUUUGCAGUAUATT	UAUACUGCAAAUAAUGUGCTT
1036	GGAAGUCCAAUCUUGUAAATT	UUUACAAGAUUGGACUUCCTT
AK050707		
608	GGAAGGUAAAUCUGCUCAATT	UUGAGCAGAUUUACCUUCCTT
334	CACUCUAUAAGUUCACUUATT	UAAGUGAACUUAUAGAGUGTT
87	GCUGAUUGGUGAACCUAGATT	UCUAGGUUCACCAAUCAGCTT
AK079912		
244	GAGUAGAAAUACUCCAGAATT	UUCUGGAGUAUUUCUACUCTT
35	GCUCUGACAUCUACUUCCATT	UGGAAGUAGAUGUCAGAGCTT
104	AGAAGACAGCAGGAGAUAATT	UUAUCUCCUGCUGUCUUCUTT
AK080070		
478	GGACAACGACAUGGUGUUATT	UAACACCAUGUCGUUGUCCTT
709	AGACAUAACUUCUACUCAATT	UUGAGUAGAAGUUAUGUCUTT
302	CCUUCACAGAUGACAAGAATT	UUCUUGUCAUCUGUGAAGGTT
AK081581		
1590	GGAUAAGGAUGGACAGAAATT	UUUCUGUCCAUCCUUAUCCTT
1347	GGAAGCUGCUGCAGCACAATT	UUGUGCUGCAGCAGCUUCCTT
756	GGGCUUGAAGAGAAGUCAATT	UUGACUUCUCUUCAAGCCCTT
AK136742		
507	CGCAGGUGUUGAUGAGGAATT	UUCCUCAUCAACACCUGCGTT
597	CAAUUUGCAGGACCAAGAATT	UUCUUGGUCCUGCAAAUUGTT
690	GGCUUACAGCAUUUCAAUATT	UAUUGAAAUGCUGUAAGCCTT
Malat1		
1082	GAUUGAAGCUAGCAAUCAATT	UUGAUUGCUAGCUUCAAUCTT
2677	GGUGUUAGGUAAUUGUUUATT	UAAACAAUUACCUAACACCTT
1762	GCUUCUGUGUAAAGAGAUATT	UAUCUCUUUACACAGAAGCTT
Neat1		
2890	GGGUCAUCUUACUAGAUAATT	UUAUCUAGUAAGAUGACCCTT
1689	GGUAGGGUUUGUGGUUUAATT	UUAAACCACAAACCCUACCTT
2015	GGAUCAAGCUUGGGAAUAATT	UUAUUCCCAAGCUUGAUCCTT
PPARg-1	CGCAUUCCUUUGACAUCAATT	UUGAUGUCAAAGGAAUGCGAG
PPARg -2	GGGCGAUCUUGACAGGAAATT	UUUCCUGUCAAGAUCGCCCTC

Table A3: Intronic and exonic RNA FISH probes for Firre

	Sequence	Sequence Name	Synthesis Scale	3' modification
Mouse Firre Intron Probes	ccatgtcttctccggtttac	Firre_intron1_1	10nmol delivered	3' Amine
	gaagccctttctatcccaag	Firre_intron1_2	10nmol delivered	3' Amine
	aagcttgcgtcttctgaaat	Firre_intron1_3	10nmol delivered	3' Amine
	aaaggatcgtataagcacgc	Firre_intron1_4	10nmol delivered	3' Amine
	tatgaatacacagccttggc	Firre_intron1_5	10nmol delivered	3' Amine
	aagtcttggatcaagcactg	Firre_intron1_6	10nmol delivered	3' Amine
	aatacagtgcttgcaaagga	Firre_intron1_7	10nmol delivered	3' Amine
	aggttcatttgacagagcac	Firre_intron1_8	10nmol delivered	3' Amine
	agattcctagggaatcctcc	Firre_intron1_9	10nmol delivered	3' Amine
	tcaggtgacttaggtctcag	Firre_intron1_10	10nmol delivered	3' Amine
	gcatgtagcacccagtttat	Firre_intron1_11	10nmol delivered	3' Amine
	ccttggatcccatctttgtc	Firre_intron1_12	10nmol delivered	3' Amine
	caataactggcctatcaggc	Firre_intron1_13	10nmol delivered	3' Amine
	ttcactgactctgagctaca	Firre_intron1_14	10nmol delivered	3' Amine
	gactcttggctgaatgaagg	Firre_intron1_15	10nmol delivered	3' Amine
	atttccccggagctaaattc	Firre_intron1_16	10nmol delivered	3' Amine
	ctgagacctaagtcacctga	Firre_intron1_17	10nmol delivered	3' Amine
	ggagcatcaattctgagtgt	Firre_intron1_18	10nmol delivered	3' Amine
	gcaaagtctcacctcatgaa	Firre_intron1_19	10nmol delivered	3' Amine
	tcaagggttaggtctcaagt	Firre_intron1_20	10nmol delivered	3' Amine
	gaggatatcctagaaggcct	Firre_intron1_21	10nmol delivered	3' Amine
	cctggtgtgcttaaaactct	Firre_intron1_22	10nmol delivered	3' Amine
	gacaatacagcagactagcg	Firre_intron1_23	10nmol delivered	3' Amine
	gttcaagtagtgagcaagca	Firre_intron1_24	10nmol delivered	3' Amine
Mouse Firre Exon Probes	taaaaggaccctagagctcc	Firre_exon1_1	10nmol delivered	3' Amine
	cacatgaagttctgttccca	Firre_exon1_2	10nmol delivered	3' Amine
	cgcacctgagacttttttaca	Firre_exon1_3	10nmol delivered	3' Amine
	ttttcttggctcgactgtcc	Firre_exon1_4	10nmol delivered	3' Amine
	ctccagtcctggttttgatc	Firre_exon1_5	10nmol delivered	3' Amine
	ttttttgccggcttcatcttc	Firre_exon1_6	10nmol delivered	3' Amine
	cataccttacaagagccgtg	Firre_exon1_7	10nmol delivered	3' Amine
	actttagagcatcctcaagg	Firre_exon1_8	10nmol delivered	3' Amine
	gtggttgtcttcagttctcc	Firre_exon1_9	10nmol delivered	3' Amine
	cacaaattcaagcaggcaag	Firre_exon1_10	10nmol delivered	3' Amine
	gctgctcaatgatcaagtct	Firre_exon1_11	10nmol delivered	3' Amine
	ccagagcatccatctgtatc	Firre_exon1_12	10nmol delivered	3' Amine
	tcatttcaaggtagcgactc	Firre_exon1_13	10nmol delivered	3' Amine
	tgggttctcaggacatatga	Firre_exon1_14	10nmol delivered	3' Amine

Table A3 (Continued)

	tctcagcatccagttctgag	Firre_exon1_15	10nmol delivered	3' Amine
	atagacaatgacaagcctgc	Firre_exon1_16	10nmol delivered	3' Amine
	cttgtgcatcctctctgaag	Firre_exon1_17	10nmol delivered	3' Amine
	aatactctgaagggtcctgt	Firre_exon1_18	10nmol delivered	3' Amine
	cttggtactggacctctgtc	Firre_exon1_19	10nmol delivered	3' Amine
	gtattttaccatggccagct	Firre_exon1_20	10nmol delivered	3' Amine
	gattacagacagcaggagca	Firre_exon1_21	10nmol delivered	3' Amine
	gctgctcaatgatcaagtct	Firre_exon1_22	10nmol delivered	3' Amine
	ggagactggacctgattttg	Firre_exon1_23	10nmol delivered	3' Amine
	catgcaatgctgtactccta	Firre_exon1_24	10nmol delivered	3' Amine
	tctagccttccttgtctctg	Firre_exon1_25	10nmol delivered	3' Amine
	cagtgtccactaactgtgtg	Firre_exon1_26	10nmol delivered	3' Amine
	gttctgagtgcagcaatctg	Firre_exon1_27	10nmol delivered	3' Amine
	atgacagtgtttgtttcccc	Firre_exon1_28	10nmol delivered	3' Amine
	cttcaaaatggctgaggaga	Firre_exon1_29	10nmol delivered	3' Amine
	aaaaaaggccaaatttcgcc	Firre_exon1_30	10nmol delivered	3' Amine
	tgcctacagaaggagagaaa	Firre_exon1_31	10nmol delivered	3' Amine
	aatactctgaagggtcaggt	Firre_exon1_32	10nmol delivered	3' Amine
	cttggtactggacctcagtc	Firre_exon1_33	10nmol delivered	3' Amine
	aggacaccagctagtagatt	Firre_exon1_34	10nmol delivered	3' Amine
	aaagaaactggaactgcagt	Firre_exon1_35	10nmol delivered	3' Amine
	ggcttccttgacttctcttc	Firre_exon1_36	10nmol delivered	3' Amine
	gctgttgtccttccaatgat	Firre_exon1_37	10nmol delivered	3' Amine
	cggtttcaaaactcacagct	Firre_exon1_38	10nmol delivered	3' Amine
	taagctcttgtgtggacctc	Firre_exon1_39	10nmol delivered	3' Amine
	tttacgacatagacgccatg	Firre_exon1_40	10nmol delivered	3' Amine
	tgtgattcccacagtagaga	Firre_exon1_41	10nmol delivered	3' Amine
	gatcacttgtgacccacttc	Firre_exon1_42	10nmol delivered	3' Amine
	gaacagcagtttgagaatcc	Firre_exon1_43	10nmol delivered	3' Amine
	agaccttctgaacactgaga	Firre_exon1_44	10nmol delivered	3' Amine
	ccttatccaggtgccttcta	Firre_exon1_45	10nmol delivered	3' Amine
	caaggcctctgatttctgtc	Firre_exon1_46	10nmol delivered	3' Amine
	tgaagaagtcagaatccaga	Firre_exon1_47	10nmol delivered	3' Amine
	caaaaagggttgagtcaagc	Firre_exon1_48	10nmol delivered	3' Amine
Human FIRRE Intron Probes	gtgcttatgcctcagaacat	FIRRE_intron1_1	10nmol delivered	3' Amine
	atttcaacatgcacagaggt	FIRRE_intron1_2	10nmol delivered	3' Amine
	tacaatatccctcttggggt	FIRRE_intron1_3	10nmol delivered	3' Amine
	tctgcttcagtgggaaattc	FIRRE_intron1_4	10nmol delivered	3' Amine
	tacttttgctcctggtttcc	FIRRE_intron1_5	10nmol delivered	3' Amine
	gggaaaatctgctctttcga	FIRRE_intron1_6	10nmol delivered	3' Amine

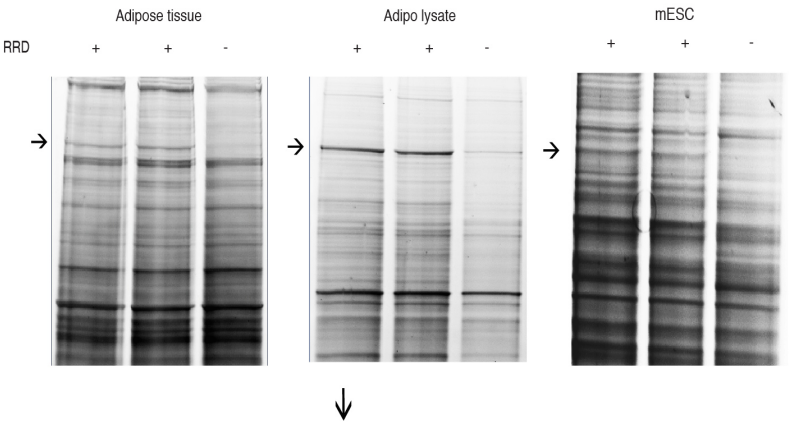
Table A3 (Continued)

	gggtagatatctaagcccca	FIRRE_intron1_8	10nmol delivered	3' Amine
	tgagatgtttcatctgcacc	FIRRE_intron1_9	10nmol delivered	3' Amine
	gctggctgctgttagatatc	FIRRE_intron1_10	10nmol delivered	3' Amine
	caaagccatggtagctgtta	FIRRE_intron1_11	10nmol delivered	3' Amine
	gtctcaatggcctcatcttc	FIRRE_intron1_12	10nmol delivered	3' Amine
	cctgatgtgctcagctttta	FIRRE_intron1_13	10nmol delivered	3' Amine
	ctatgctatgggcaagtgtc	FIRRE_intron1_14	10nmol delivered	3' Amine
	cgagaaaactctcacatgca	FIRRE_intron1_15	10nmol delivered	3' Amine
	ccatatgttacagcagtgca	FIRRE_intron1_16	10nmol delivered	3' Amine
	caggaaagtcaatttcccca	FIRRE_intron1_17	10nmol delivered	3' Amine
	ttgagaccattgagacatgc	FIRRE_intron1_18	10nmol delivered	3' Amine
	tggtcaagtcaaactgtgac	FIRRE_intron1_19	10nmol delivered	3' Amine
	cacacatctggctttcctag	FIRRE_intron1_20	10nmol delivered	3' Amine
	ccactctctgaccagatagt	FIRRE_intron1_21	10nmol delivered	3' Amine
	tggattccttttcaaaggct	FIRRE_intron1_22	10nmol delivered	3' Amine
	tgtgcgatcactttcaagag	FIRRE_intron1_23	10nmol delivered	3' Amine
	tggcgaatggatttcaatct	FIRRE_intron1_24	10nmol delivered	3' Amine
Human FIRRE Exon Probes	gtgactctgctaccacaaag	FIRRE_exon1_1	10nmol delivered	3' Amine
	caacctggtattgtctggtt	FIRRE_exon1_2	10nmol delivered	3' Amine
	agcacatggcatccttttat	FIRRE_exon1_3	10nmol delivered	3' Amine
	ctcctagctgcaagcaattt	FIRRE_exon1_4	10nmol delivered	3' Amine
	gctgatcccatttctcttca	FIRRE_exon1_5	10nmol delivered	3' Amine
	ctaacccttcttggtttggc	FIRRE_exon1_6	10nmol delivered	3' Amine
	gtcagctctcctagcaaaag	FIRRE_exon1_7	10nmol delivered	3' Amine
	tcttcactcggttttggttc	FIRRE_exon1_8	10nmol delivered	3' Amine
	ctccagcagacaagaacttt	FIRRE_exon1_9	10nmol delivered	3' Amine
	aagcgaggtcacaggtaatg	FIRRE_exon1_10	10nmol delivered	3' Amine
	aacagccatgtatggagaag	FIRRE_exon1_11	10nmol delivered	3' Amine
	tgcttcagaaagctggattt	FIRRE_exon1_12	10nmol delivered	3' Amine
	tctcaggtctcacatcacat	FIRRE_exon1_13	10nmol delivered	3' Amine
	tgcaacttgagttgcttctt	FIRRE_exon1_14	10nmol delivered	3' Amine
	ccttgcatgatacatgagga	FIRRE_exon1_15	10nmol delivered	3' Amine
	ggcaaagagcagaagataga	FIRRE_exon1_16	10nmol delivered	3' Amine
	cattttcagggacctcacag	FIRRE_exon1_17	10nmol delivered	3' Amine
	aaatctgggtagtattggcc	FIRRE_exon1_18	10nmol delivered	3' Amine
	accaactgtttcagtacacg	FIRRE_exon1_19	10nmol delivered	3' Amine
	actgaattccatctgtggtc	FIRRE_exon1_20	10nmol delivered	3' Amine
	accagatctcagtattccgt	FIRRE_exon1_21	10nmol delivered	3' Amine
	attagcaatggtgtttgcaa	FIRRE_exon1_22	10nmol delivered	3' Amine
	acctcagggatccatgtaat	FIRRE_exon1_23	10nmol delivered	3' Amine
	attgtcttgtcgtattgggg	FIRRE_exon1_24	10nmol delivered	3' Amine

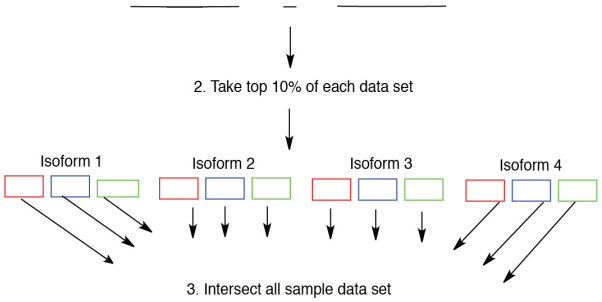
Table A3 (Continued)

	gcattgtttctgcagtgttc	FIRRE_exon1_25	10nmol delivered	3' Amine
	aaccaagtcttcccatttct	FIRRE_exon1_26	10nmol delivered	3' Amine
	ttcaagaactcagtttccgc	FIRRE_exon1_27	10nmol delivered	3' Amine
	ccctcatcaagcatctcctc	FIRRE_exon1_28	10nmol delivered	3' Amine
	gcacaaccaatcttctcatt	FIRRE_exon1_29	10nmol delivered	3' Amine
	tgggtacagaccttaggttt	FIRRE_exon1_30	10nmol delivered	3' Amine
	cagcatcatatctgcagtgt	FIRRE_exon1_31	10nmol delivered	3' Amine
	accaagtcttcccatttcta	FIRRE_exon1_32	10nmol delivered	3' Amine
	tcaagaactcagttactgca	FIRRE_exon1_33	10nmol delivered	3' Amine
	cctcatcaagcatctcctct	FIRRE_exon1_34	10nmol delivered	3' Amine
	ggaacagaccttagtgatcc	FIRRE_exon1_35	10nmol delivered	3' Amine
	tcagcatcgtatctgcagta	FIRRE_exon1_36	10nmol delivered	3' Amine
	tttctgcacaaccaagtctt	FIRRE_exon1_37	10nmol delivered	3' Amine
	caagaactcagtattggcca	FIRRE_exon1_38	10nmol delivered	3' Amine
	ctcatcaagcatctcctgtt	FIRRE_exon1_39	10nmol delivered	3' Amine
	agacgagtcaaggtaacaga	FIRRE_exon1_40	10nmol delivered	3' Amine
	cccatcttgggtcaatgaaa	FIRRE_exon1_41	10nmol delivered	3' Amine
	gtcttacattcctagtgcca	FIRRE_exon1_42	10nmol delivered	3' Amine
	agtttaaccgagggaaatcg	FIRRE_exon1_43	10nmol delivered	3' Amine
	acagagtggccttaacattg	FIRRE_exon1_44	10nmol delivered	3' Amine
	tagcacctcagttacacaga	FIRRE_exon1_45	10nmol delivered	3' Amine
	gcctaactacctctattggc	FIRRE_exon1_46	10nmol delivered	3' Amine
	cttaaatgaggtccacagca	FIRRE_exon1_47	10nmol delivered	3' Amine
	tgaaaagtgcttctgttgca	FIRRE_exon1_48	10nmol delivered	3' Amine

Appendix 4: RNA pull-downs in adipose and mESC lysates followed by differential mass spectrometry to identify the RRD-specific peptides



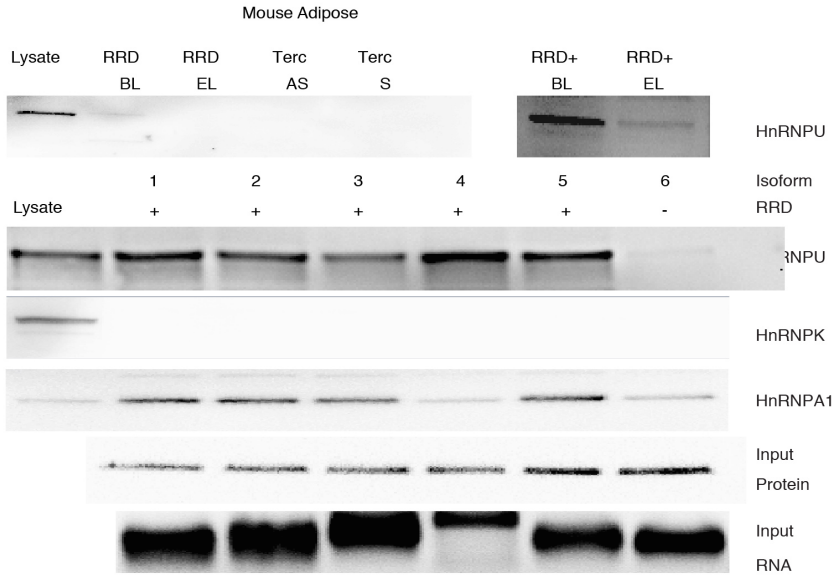
1. Take the difference between the ALR+ and ALR- peptide counts for each protein. Do this for each ALR + isoform (4) in every lysate (3). Repeat for all the isoforms.



2. Take top 10% of each data set

3. Intersect all sample data set

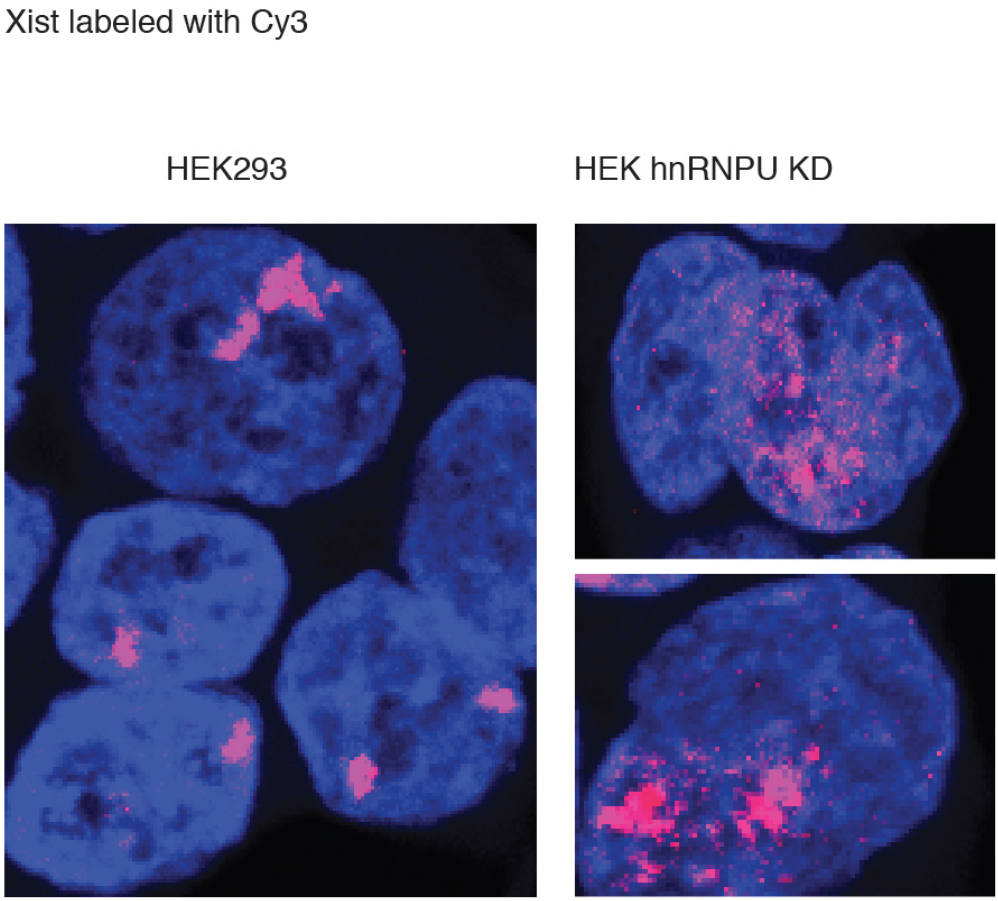
4. Finalize to the ones with highest unique peptide counts and unique spectra



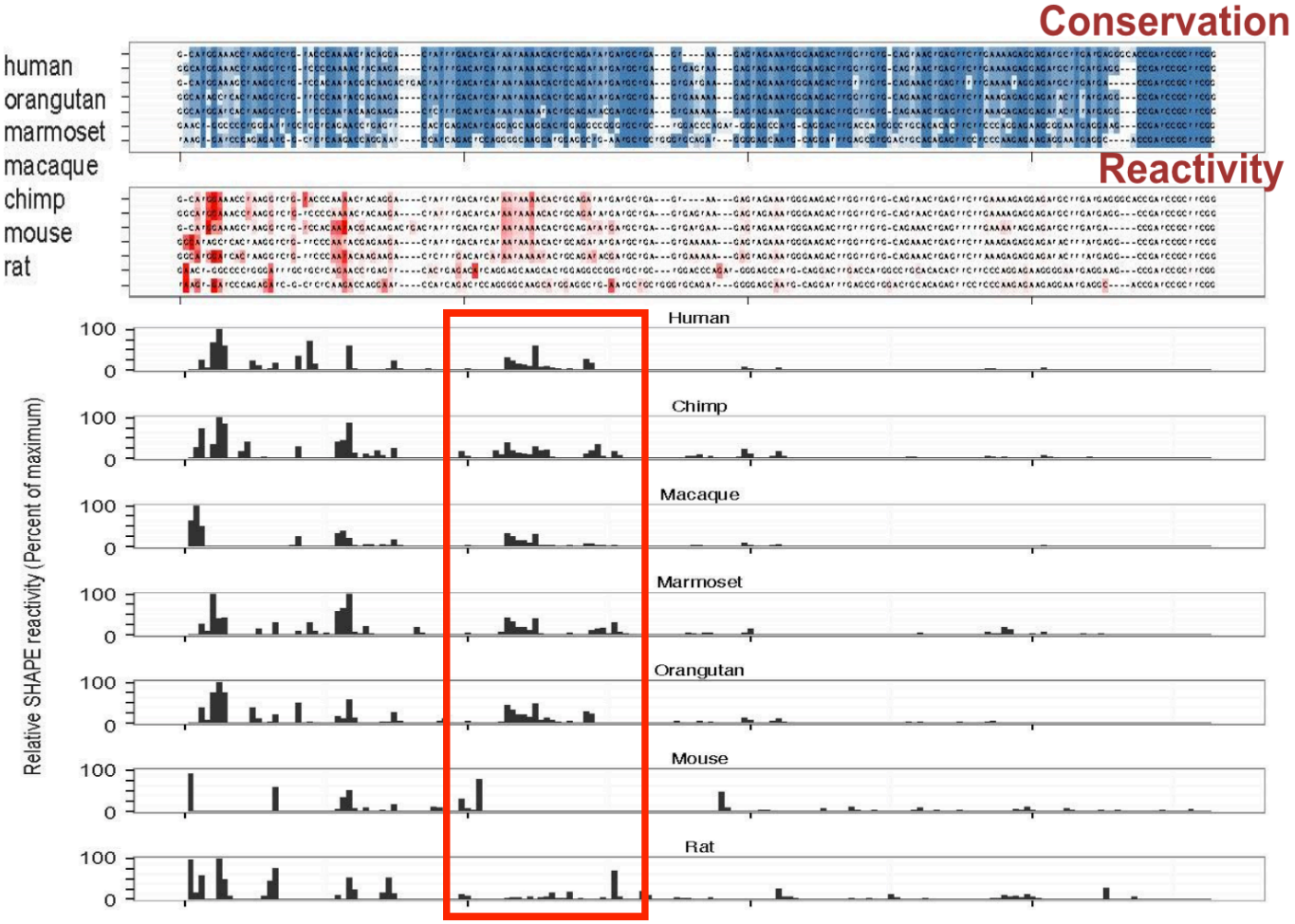
Appendix 5: Top mass spectrometry hits from RNA pull-downs

MASS SPEC HITS	Peptide_ A v
heterogeneous nuclear ribonucleoprotein U	25.0
heterogeneous nuclear ribonucleoproteins A1 isoform b	16.7
heterogeneous nuclear ribonucleoprotein U-like protein 2	11.7
zinc finger CCCH domain-containing protein 18 isoform b	11.0
fatty acid synthase	10.7
insulin-like growth factor 2 mRNA-binding protein 3	9.0
F AC T complex subunit SPT16	8.3
constitutive coactivator of P P AR-gamma-like protein 1	5.0

Appendix 6: RNA FISH targeting Xist in hnRNPU knockdown conditions



Appendix 7: SHAPE-Seq reactivity spectra for 7 species of RRD, red box marking the differential region between the rodents and primates



Appendix 8: CRISPR Display: A modular method for locus-specific targeting of long noncoding RNAs and synthetic RNA devices *in vivo*

Shechner, D., **E. Hacisuleyman**, S.T Younger, J.L Rinn. Nature Methods. In review.

Introduction

Noncoding RNAs (ncRNAs) are central components of diverse and fundamental processes in all kingdoms of life[1]. In eukaryotes, many well-established ncRNAs, and a number of newly identified mammalian long ncRNAs, are thought to help initiate or maintain regulatory processes within the nucleus[1-3]. However, mechanistic dissection of these putative nuclear regulators *in vivo* is often technically limited. For example, established knock-in and knockout strategies[4, 5] lack the throughput required for high-resolution structure/function analysis, and cannot easily separate roles performed by an RNA transcript from those performed by a functional DNA element or by a cryptically encoded peptide[5, 6].

Therefore, an experimental method that post-transcriptionally relocates a ncRNA transcript to an ectopic site would be an important tool for the study of natural ncRNA function. In addition, this method could provide a powerful engine for synthetic biology. Many natural RNA domains have been adapted as components in artificial regulators, reporters and scaffolds[7-12]; such devices can be further expanded using the almost limitless array of functional RNA motifs generated through *in vitro* selection, including aptamers and ribozymes[12-15]. Hence, the ability to target synthetic RNA devices to specific DNA loci would enable a wide range of novel synthetic biological methods.

Towards this goal, we aimed to develop a general ncRNA ectopic localization system. We reasoned that this could be achieved using an artificial protein "conduit" that is programmed to bind a ncRNA and target it to specific, also programmable, DNA locations. (**Fig. A8.1a**). This

strategy would facilitate reconstitution and functional dissection of natural ncRNAs, and would expand the purview of synthetic ncRNA devices by targeting them to specific genomic loci.

A potentially powerful source for such an RNA-localizing conduit is the *S. pyogenes* Cas9 nuclease (*Sp.* Cas9), an extremely high-affinity[16], programmable DNA-binding protein isolated from a type II CRISPR-associated system[17, 18]. The DNA locus targeted by Cas9 (and by its nuclease-deficient mutant, "dCas9" (Refs. 19-22)) precedes a three-nucleotide (nt) 5′-NGG-3′ "PAM" sequence, and matches a 15–22-nt guide sequence within a Cas9-bound RNA cofactor (**Fig. A8.1b**). Altering this guide is sufficient to reprogram (d)Cas9 targeting. In a multitude of CRISPR-based biotechnology applications[23-31], the guide is often presented in a so-called sgRNA, wherein the two natural Cas9 RNA cofactors (gRNA and tracrRNA)[17, 18] are fused via an engineered loop (**Fig. A8.1b**). Yet, despite recent work dissecting the determinants of Cas9 RNA recognition[32-36], it remained unclear if and where large, structured RNA domains could be implanted within CRISPR complexes while maintaining RNA-directed localization.

Here, we demonstrate that *Sp.* dCas9 can be co-opted to deploy a large RNA cargo to targeted DNA loci by directly linking that cargo to the sgRNA. We term this strategy, in which exogenous RNA domains are displayed on dCas9, CRISPR-Display, or "CRISP-Disp." With the appropriate expression system and insertion point, CRISP-Disp does not appear inherently limited by the size or sequence composition of its RNA cargo. This allows us to functionalize dCas9 complexes with structured RNA domains, natural lncRNAs of up to 4.8 kb in length, artificial RNA modules and pools of random sequences. Moreover, we demonstrate that these RNA-based functions can be simultaneously multiplexed using a shared pool of dCas9.
Collectively, we provide initial insights into the general utility of CRISP-Disp for both the study of natural ncRNAs and the construction of novel RNA-based devices.

Results

An experimental test bed for ncRNA localization methods

We first implemented a highly sensitive reporter assay in HEK293FT cells, modeled after established transcription activator reporter systems[19, 37], with which to survey potential ectopic ncRNA localization methods (**Supplementary Fig. 1a**). Our system comprises two constructs–Reporter and Normalizer–that encode mutually orthogonal pairs of fluorescent proteins and secreted luciferases, preceded by arrays of short "target" and "non-target" motifs to which artificial transcription activators (protein, RNA or RNP) can be directed (**Fig. A8.1b**). This system can be transiently transfected,wherein it exhibits extremely high sensitivity, or stably integrated into the host chromatin through lentiviral transduction, wherein it better approximates the behavior of endogenous genes.

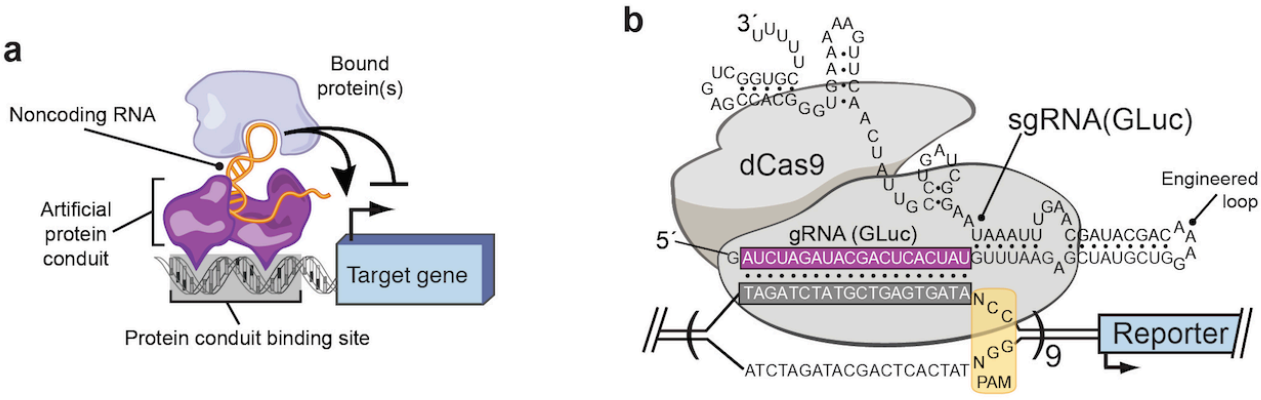


Figure A8.1. A dual reporter system for characterizing locus-specific ncRNA targeting strategies. (**a**) Schematic of a general, targeted ncRNA localization system: a ncRNA cargo (and potential associated proteins) is ectopically targeted to a DNA locus via a programmable protein conduit. (**b**) Overview DNA targeting by *S. pyogenes* dCas9, directed by a minimal sgRNA that targets the GLuc reporter. Modeled after (Ref. 25).

In our initial strategy the targeting "conduits" (**Fig. A8.1a**) were based upon transcription activator-like effectors (TALEs), a versatile class of customizable DNA-binding proteins[37]. While our reporter system responded robustly when coexpressed with a cognate TALE domain fused to established regulatory proteins[37], we were unable to coax the TALE into recruiting ncRNAs to a chromatin-integrated target locus (**Supplementary Figs. 2–3**). We ascribed this problem to the separable DNA- and RNA-binding activities of our TALE construct, which could be independently saturated without forming DNA•TALE•RNA ternary complexes[38].

To circumvent this issue, we turned to the *S. pyogenes* CRISPR-Cas9 system, which intrinsically couples its DNA- and RNA-binding activities[16, 18] (**Fig. A8.1b, Supplementary Fig. 1a**). Indeed, a nuclease-deficient Cas9 mutant (dCas9, Refs. 19-22), fused to the VP64 transcription activator (dCas~VP) robustly activated our reporter system in an RNA-dependent manner, as indicated by FACS and luminometric assays (**Supplementary Fig. 1b,c**). Critically, and in contrast with our TALE system (**Supplementary Figs. 2–3**), similar RNA-dependent targeting was observed with both transiently expressed and stably integrated reporters (**Supplementary Fig. 1d**, *right*).

Adapting CRISPR-Cas9 as an RNA display device

We next needed to establish if dCas9 could be co-opted to deploy a larger RNA cargo to a target DNA locus. However, it was unclear *a priori* where insertions within its minimal sgRNA scaffold would be tolerated, and how large they could be. To examine this, we devised five topologies (TOP1–4; INT) in which the sgRNA was appended with structured, 81–250 nt "accessory domains" that serve as proxies for larger RNAs in general, and within which were embedded cassettes of high-affinity stem-loops recognized by the PP7 phage coat protein[39] **(Fig.**

A8.2a; Supplementary Fig. 4). In TOP1 and TOP2, the sgRNA was placed at the 5′- and 3′-end of the accessory domain, respectively (**Fig. A8.2a**). In TOP3 and TOP4, the tracrRNA component of a natural crRNA•tracrRNA complex[18] was likewise appended. In INT, a smaller accessory domain was grafted into the internal sgRNA engineered loop (**Fig. A8.1b**), which makes no direct contacts with the dCas9 protein[40]. At 357 nt, the largest of these constructs is more than three times the length of a minimal sgRNA and adds three-fold more sequence than does the longest modified sgRNA previously reported[20, 29, 34, 35].

We subjected these sgRNA chimeras to two variations of our CRISPR transcription activator assay (**Fig. A8.2b**). In "direct activation" assays, they were coexpressed with dCas9~VP. Reporter gene activation indicates that the sgRNA variant forms a competent targeting complex with dCas9. In "bridged activation" assays, constructs were coexpressed with dCas9 and PP7~VP, a chimera of PP7 and VP64. Bridged activation should only occur if the accessory domain remains functional in the mature dCas9 complex. While similar experimental schemes have been used to develop CRISPR-based transcription activators[20-22, 35, 41], for our purposes the served only to survey the integrity of the accessory RNA domains.

Using transient reporters (**Fig. A8.2c**, *top*), we observed measurable direct activation with all five topologies. However, while the activities of TOP1, TOP2 and INT were reduced less than twofold from that of the minimal sgRNA, TOP3 and TOP4 were less robust, exhibiting ~3–11% the efficacies of their unimolecular counterparts. Critically, bridged activation was only observed with TOP1, TOP3 and INT, indicating that these constructs alone retained functional accessory domains in mature dCas9 complexes. FACS analysis corroborated these results (**Fig. A8.2d**). Using integrated reporters, results were qualitatively similar, although the low activities of TOP3 and TOP4 could not be significantly measured (**Fig. A8.2c**, *bottom*).

We hypothesized that the inability of TOP2 to induce bridged activation (**Fig. A8.2c,d**) might be due to degradation of its accessory domain, as has been observed with shorter sgRNA 5′-extensions[34, 42]. This hypothesis was supported by RNA immunoprecipitation (RIP) and qRT-PCR: while recovery of the sgRNA core and accessory domains from dCas9•TOP1 complexes was quantitative, from dCas9•TOP2 complexes the yield of accessory domain was nearly half that of the sgRNA core (**Fig. A8.2e)**. Although this did not indicate a complete loss of the accessory domain, we hypothesize that other factors–such as RNA folding, or the geometry of the dCas9•TOP2 complex–may contribute to the loss of TOP2 bridged activation. Furthermore, this result implies that structured 5′–additions might partially stabilize a modified sgRNA from degradation, as has been reported elsewhere[29].

Together, these results demonstrate that dCas9 can form productive targeting complexes with longer RNAs, and can efficiently present an RNA cargo to a DNA locus in at least two different topologies: on the sgRNA 3′ terminus (TOP1), or within the sgRNA engineered loop (INT) (**Fig. A8.1b**).

Targeting long RNAs to endogenous genomic loci

To illustrate the general applicability of CRISP-Disp, we tested whether results from our reporter system could be recapitulated at endogenous loci. We generated pools of sgRNA, TOP1 and INT constructs targeting the human *ASCL1, IL1RN, NTF3* and *TTN* promoters[21, 22, 41], and surveyed direct and bridged activation of these genes using qRT-PCR. These assays were performed in integrated GLuc reporter cells, allowing us to simultaneously monitor construct efficacy and target specificity. As shown in (**Fig. A8.2f**), activation of each locus paralleled the results obtained using our GLuc reporter, demonstrating CRISP-Disp enables deployment of large RNA domains to genomic loci. Furthermore, in all cases, activation of each locus was

specific to the gRNA sequences used, implying that the presence of an accessory RNA domain did not perturb the fidelity of dCas9 targeting.

To examine this in greater depth, we performed mRNA seq from reporter cells expressing dCas9~VP and GLuc-targeting sgRNA, TOP1 and INT constructs (**Supplementary Fig. 5**). As predicted, all three RNA constructs induced measurable and specific activation of the GLuc reporter locus (**Supplementary Fig. 5a–b**). However, changes in global gene expression induced by each RNA construct–a proxy for dCas9 off targeting–were essentially indistinguishable (**Supplementary Fig. 5b–c**). Two genes (*B3GNT8* and *RPS17*) were moderately activated by TOP1 and INT, respectively, but since neither gene is positioned within 10 kb an off-target site for our GLuc gRNA[43-45], we hypothesize that this activation was induced by expression of the modified sgRNA constructs themselves, and not by spurious mislocalization of dCas9~VP.

Collectively these results establish that CRISP-Disp with TOP1- and INT-like constructs should be generally functional at endogenous loci, and that addition of accessory RNA domains to the sgRNA scaffold does not significantly alter dCas9 fidelity.

CRISP-Disp with RNA Polymerase II transcripts

We next sought to engineer CRISP-Disp for use with long ncRNAs. This required replacing the conventional RNA Polymerase III (Pol III) expression system used above (U6, **Supplementary Fig. 1a**), which is limited in transcript length and sequence composition, with a Pol II promoter and terminator. Although this has been previously accomplished by nucleolytically excising sgRNAs from their primary transcripts[28, 46, 47], we feared that the cleavage products generated by these methods might be unstable. Therefore, using our five chimeric sgRNA constructs (**Fig. A8.2a**), we tested a variety of Pol II expression systems for the ability to generate nuclear-localized CRISP-Disp RNAs *de novo*. We surveyed noncanonical

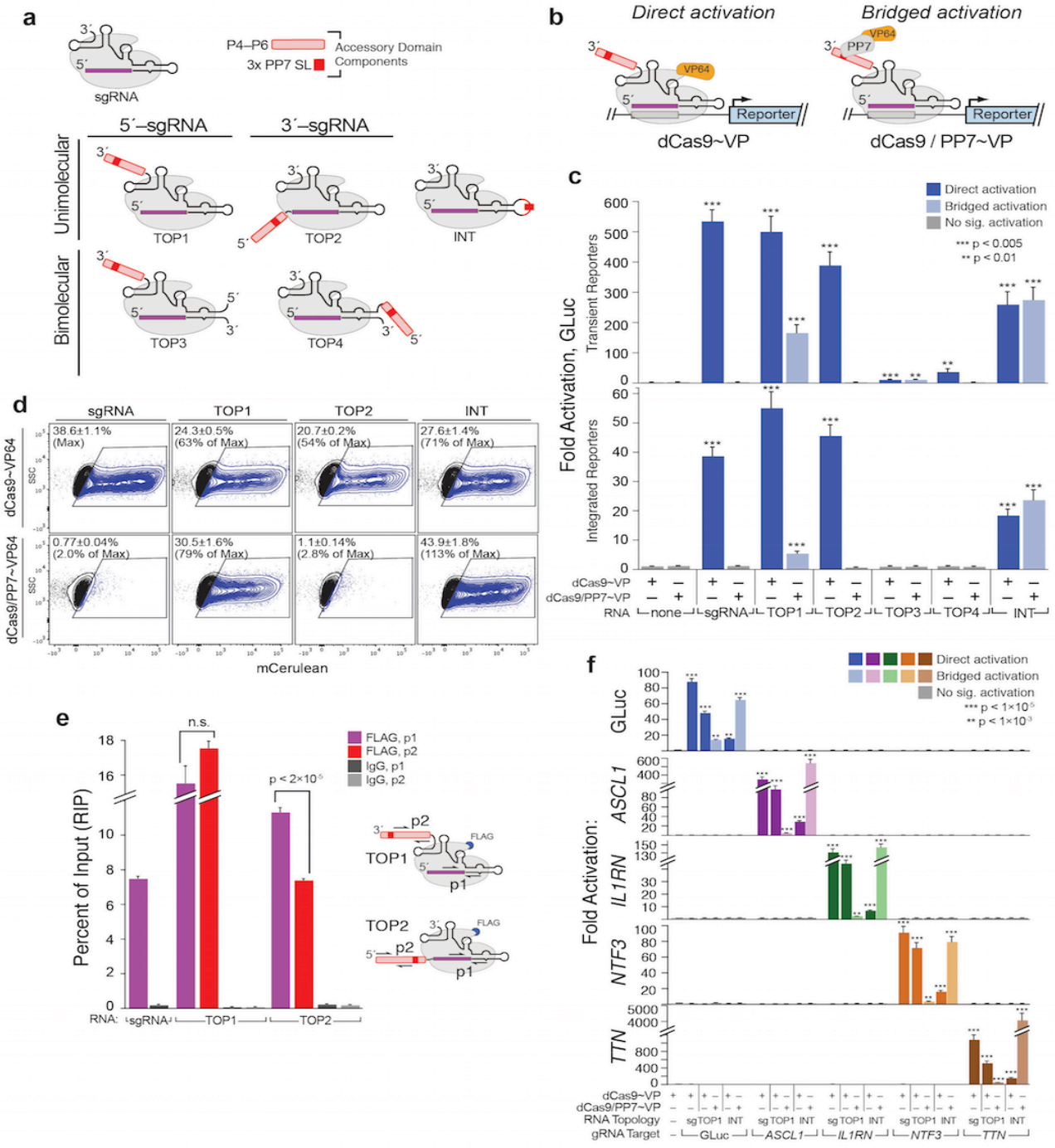


Figure A8.2. Large structured RNA domains can be functionally appended onto the sgRNA scaffold at multiple points. (**a**) Design of "TOP" topology constructs. Accessory RNA domains are detailed in (**Supplementary Fig. 4**). (**b**) Schematics summarizing direct activation

(dCas9~VP, *left*) and bridged activation (dCas9/PP7~VP, *right*) assays. (**c**) Luciferase reporter assays of the five topology constructs, as in (**Supplementary Fig. 1d**). Values are means ± standard deviation; n = 3. Student's one-tailed t-test, relative to negative controls (*far left*). Expression constructs for all components are detailed in (**Supplementary Fig. 1**). (**d**) FACS analyses on transient reporter assays, as in (**Supplementary Fig. 1c**). Means ± standard deviation; n = 3. (**e**) RIP/qRT–PCR of dCas9•TOP1 and dCas9•TOP2. qPCR primers target the core sgRNA and the accessory domain (p1 and p2, respectively, *right)*. Values are means ± standard deviation. n = 4., Student's one-tailed t-test. (**f**) Targeting minimal ("sg") and expanded ("TOP1" and "INT") sgRNAs to endogenous loci: *ASCL1*, *IL1RN*, *NTF3* and *TTN*. GLuc activation was measured by luciferase assays; endogenous gene activation was measured using qRT-PCR. Values are means ± standard deviation. n = 4, Student's one-tailed t-test, relative to negative control cells expressing dCasVP alone. Endogenous gene-targeting constructs were each mixed pools of four gRNAs[21, 22, 41].

promoters and terminators, as well as RNA motifs known to facilitate nuclear retention or import[48-51] (**Supplementary Fig. 6a**).

In general, such Pol II transcripts were markedly less effective CRISP-Disp components than their Pol III-driven counterparts, particularly when expressed from canonical promoter/terminator systems (**Supplementary Fig. 6b**). Proficiency was modestly restored by exploiting polyadenylation-independent terminators, such as the U1 snRNA 3′-Box[48, 52], and the RNA triplex-forming *MALAT1* ENE/MASC system[50] (**Supplementary Fig. 6b**). As predicted, for the "CMV/3′Box" backbone–which pairs the CMV promoter and U1 3′-Box–this rise in activity correlated with a lack of transcript polyadenylation and a concomitant increase in

nuclear transcript abundance (**Supplementary Fig. 7**). Surprisingly, expression from this backbone also enabled bridged activation in the TOP2 configuration (**Fig. A8.3a** and **Supplementary Fig. 6c–d**), which was not possible under Pol III expression (**Fig. A8.2c**). This was corroborated by RIP-qPCR: immunoprecipitation of dCas9•TOP2 isolated the sgRNA core and accessory domains in nearly stoichiometry yields, further indicating that the TOP2 long RNA chimera remained intact in complex with dCas9 (**Fig A8.3b**).

CRISP-Disp with artificial lncRNAs

To examine the length limitations on CRISP-Disp RNAs, as a proof-of-principle we next attempted to build dCas9 complexes with transcripts approaching the size of natural lncRNAs. We expanded our CMV/3′-Box TOP1 and TOP2 constructs by adding a second complete P4–P6 domain bearing a cassette of MS2 stem-loops (**Figs. A8.3c**, and **Supplementary Fig. 4**). The two P4–P6 domains in these "artificial lncRNA" constructs were positioned so as to bracket the sgRNA core ("Double TOP0,"), or contiguously on the sgRNA 3′- and 5′-terminus ("Double TOP1" and "Double TOP2," respectively). At 650 nt, these accessory domains are themselves nearly seven times longer than a minimal sgRNA. . Furthermore, constructs of this sort, which specifically bind two different cognate proteins, could prove useful in the design of chromatin-targeting lncRNA-like "scaffolds."[10]

All three constructs induced measurable direct activation of both transient and integrated GLuc reporters (**Fig. A8.3d** and **Supplementary Fig. 8**). Of these, Double TOP1–2 were more proficient, nearly rivaling the activities of their single-domain counterparts in transient assays (**Fig. A8.3a**). Moreover, all three constructs exhibited significant bridged activation. In transient reporter assays, luciferase activity monotonically increased upon coexpression with PP7~VP, MS2~VP or both, indicating that each construct retained both P4–P6 domains in mature dCas9

complexes (**Fig. A8.3d**). Although a qualitatively similar trend was observed for Double TOP1 using integrated reporters, the activities of all three constructs were hampered by the assay's limited dynamic range (**Supplementary Fig. 8)**. Therefore, to confirm that the 650 nt accessory domains in Double TOP1–2 remained intact during CRISP-Disp, we immunoprecipitated dCas9 and performed qRT–PCR, using a primer pair that spans the two P4–P6 domains. For each construct, we observed essentially stoichiometric yields of the sgRNA core and double P4–P6 accessory domains (**Fig. A8.3e**). We therefore infer that CRISP-Disp does not appear intrinsically limited by RNA length, a necessary prerequisite for its use in studying natural lncRNA function.

CRISP-Disp with natural lncRNA domains

A potentially powerful application of CRISP-Disp would be the ectopic localization of natural lncRNAs post-transcriptionally, since reconstitution of lncRNA activity at an ectopic site–unequivocal illustration that the RNA molecule *per se* is the functional element–is unattainable by existing methods[5, 6].

To demonstrate the plausibility of this approach, we first established that natural long ncRNAs could be incorporated into CRISP-Disp complexes. We generated Pol II-driven TOP1- and INT-like constructs appended with human lncRNA domains, spanning lengths of 87–4799 nt. These domains included the repressive NoRC-binding pRNA stem-loop[53], three enhancer-transcribed (eRNAs, Ref. 54), the repressive A-repeat domain ("RepA") of *Xist*[55] and putative transcription activator *HOTTIP*[56].

Each construct exhibited significant direct activation activity (with dCas9~VP) in both transient and integrated assays, indicating that each formed functional CRISP-Disp targeting complexes (**Fig. A8.3f**). This was supported by RIP-qPCR (targeting the sgRNA core), although

by this assay complexation efficiency appeared to decline monotonically with increasing lncRNA length (**Fig. A8.3g**). Furthermore, by surveying various intervals spanning each lncRNA domain, we observed that, relative to the sgRNA core, nearly quantitative yields of intact lncRNA domains were recovered for all constructs, indicating that each remained intact in the majority of CRISP-Disp complexes (**Fig. A8.3g**). Intriguingly, while some accessory domain loss was evident, this apparent degradation was not simply contingent upon RNA length. Perhaps this suggests that RNA structure might contribute to a particular lncRNA's overall stability, as has been observed with smaller accessory domains[34]. Regardless, we conclude that CRISP-Disp can accommodate ncRNA domains of up to several kilobases in length, including naturally-occurring lncRNAs.

Having established that lncRNAs can be incorporated into CRISP-Disp complexes, we next examined if these complexes could regulate our reporter. Encouragingly, most of the lncRNA constructs repressed or activated GLuc expression as suggested from existing studies[53-56], albeit quite modestly. Specifically, pRNA and RepA diminished normalized GLuc expression, while *TRERNA1*, ncRNA-a3 and *HOTTIP* induced moderate activation (**Fig. A8.3h**). While a full characterization of lncRNA ectopic reconstitution is beyond the scope of these initial studies, we assert that this represents an important proof-of-principle, demonstrating the plausibility of larger-scale lncRNA functional studies with CRISP-Disp.

CRISP-Disp with a diverse array of RNA species

We were particularly intrigued by the proficiency of U6-driven INT, which rivaled that of the minimal sgRNA in all conditions tested (**Fig. A8.2c,d,f**). In fact, bridged activation by INT constructs approached or exceeded the direct activation induced by minimal sgRNAs at all of the endogenous loci tested (**Fig. A8.2f**). Therefore, reasoning that the sgRNA "engineered loop"

(**Fig. A8.1b**) might provide a universal insertion point for exogenous RNA domains, we sought to explore the scope of sequences and structures tolerated at this position.

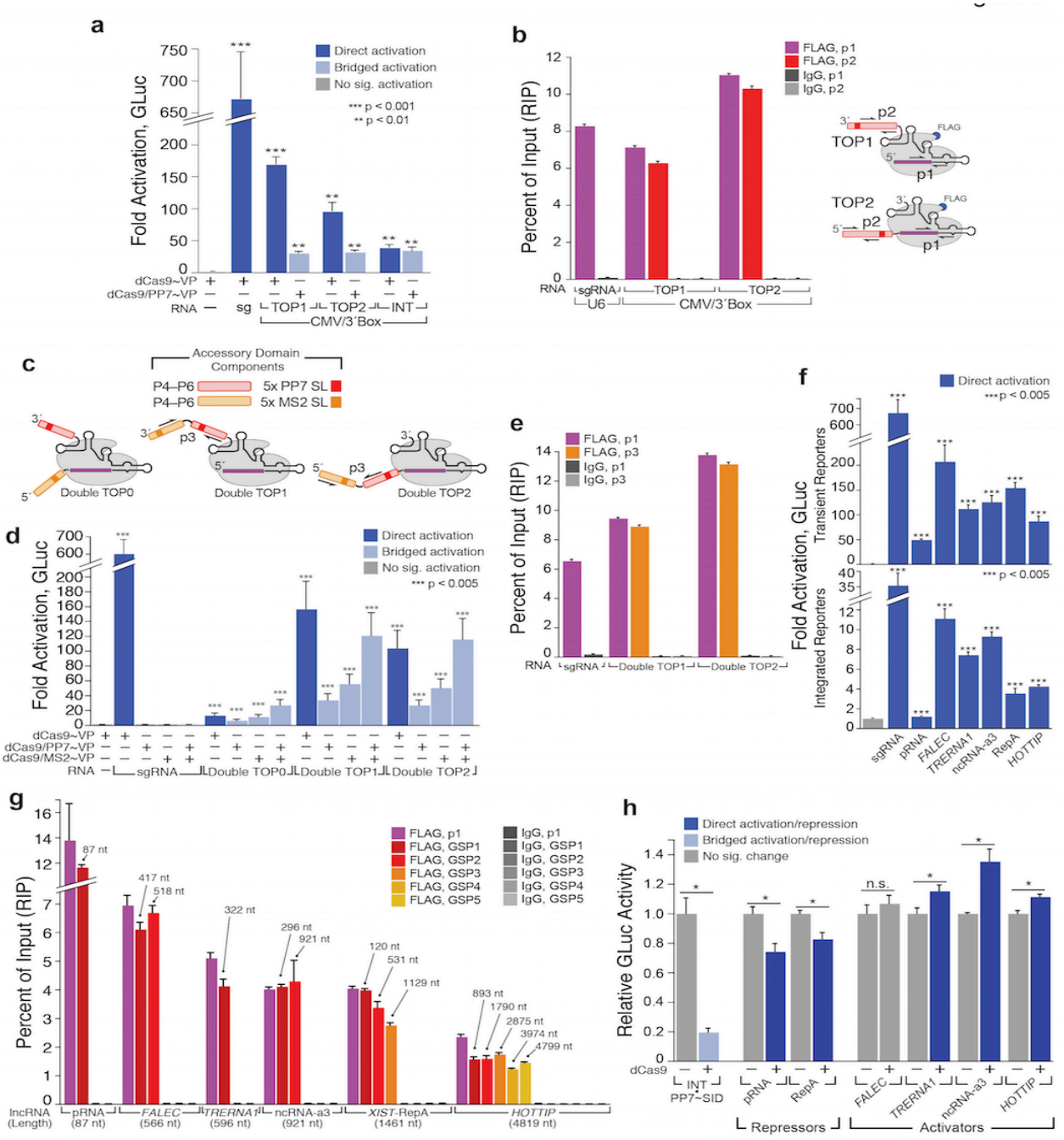


Figure A8.3. RNA polymerase II expression enables CRISP-Disp with artificial and natural lncRNAs. (**a–b**) Pol II expression restores function to the TOP2 accessory domain (**Fig.**

A8.2a). (**a**) Direct and bridged activation by the most effective topologies, using the CMV/3′ Box system (**Supplementary Figure 6**). Transient reporter assays are shown. Values are means ± standard deviation, *n = 3*; "sg," minimal sgRNA, driven from a U6 promoter. (**b**) RIP/qRT–PCR of dCas9 complexed with CMV/3′-Box TOP1 or TOP2, as in (**Fig. A8.2e**). (**c**–**e**) CRISP-Disp with "artificial lncRNAs." (**c**) Design of "Double TOP" constructs. Accessory domains are detailed in (**Supplementary Fig. 4**). Each P4–P6 domain is separated by a 25 nt unstructured linker, to produce 650 nt accessory domains. For Double TOP1 and Double TOP2, the position of the domain-spanning qPCR primer pair "p3", is indicated. (**d**) Direct and bridged activation assays, using transient reporters and Double TOP constructs. Error bars, means ± standard deviation. *n* = 3, Student's one-tailed t-test, relative to negative control cells expressing dCas9~VP alone (*far left*). Long ncRNAs were expressed from the CMV/3′Box backbone (**Supplementary Fig. 6**). Equivalent assays with integrated reporters fell below the detection limit for all but Double TOP1 (**Supplementary Fig. 8**) (**e**) RIP/qRT–PCR of dCas9•Double TOP1 and dCas9•Double TOP2. Immunopurified RNA was analyzed by qPCR primers targeting the sgRNA core, or which spanned the two P4–P6 monomers in the accessory domain (p1 and p3, respectively. Values are means ± standard deviation. *n* = 4. (**f**–**h**) CRISP-Disp with natural lncRNAs. (**f**) sgRNAs appended with a battery of lncRNA domains[53-56] form functional complexes with dCas9~VP. Direct activation assays using (*top*) transient and (*bottom*) integrated reporters are shown. The minimal *TIP5*-binding NoRC-associated RNA stem ("pRNA"[53]) was displayed internally, as in INT; all other domains were appended on the sgRNA 3′ terminus, as in TOP1. RNA constructs were expressed using the CMV/MASC system (**Supplementary Fig. 6**). Error bars, means ± standard deviation. *n* = 3, Student's one-tailed t-test, relative to negative control cells expressing dCas9~VP alone (*far left*). (**g**) LncRNA accessory domains remain intact

in CRISP-Disp targeting complexes. RIP/qRT–PCR of dCas9, complexed with a battery of constructs in (**f**). Immunopurified RNA was analyzed using qPCR primers targeting the sgRNA core (p1), or with sets of gene specific primers targeting intervals along the length the lncRNA domain (GSP1–GSP5). Above each primer set, the maximum distance between the qPCR amplicon and sgRNA core domain is indicated. Values are means ± standard deviation. $n = 4$. (**h**) Transient reporter assays with CRISP-Disp lncRNA constructs, grouped into putative repressors (*middle*) and activators (*right*). Values quoted are average (GLuc/CLuc), normalized relative to those of control cells expressing each lncRNA alone. For comparison, bridged repression with U6-driven INT, complexed with dCas9 and PP7~SID[59] is shown (*left, light blue*). Error bars, means ± standard deviation. $n = 3$, Student's one-tailed t-test, relative to negative control (*far left*) *, $p < 0.05$. None of the constructs tested–including INT•SID–perturbed the activity of integrated reporters (*not shown*).

To examine the influence of internal insert size on CRISP-Disp function, we first generated a series of INT-like constructs bearing one, three or five internal PP7 stem-loops, spanning 25–137 nt (**Fig. A8.4a**). Each construct induced robust Gluc activation in all assay formats (**Fig. A8.4a**), indicating that each formed a productive CRISP-Disp complex with an intact accessory domain. Notably, this fivefold expansion of insert size reduced activity only twofold, implying that yet larger internal insertions might be tolerated. To test this possibility, we appended a ~250 nt domain, equivalent to the accessory domains of TOP1–4 (**Fig. A8.2a, Supplementary Fig. 4**), via a flexible three-way junction at the internal insertion point (**Fig. A8.4b**). This construct also induced robust GLuc activation in all assay formats, indicating that even an insert 2.5-fold larger than–and structurally discontinuous with–the core sgRNA can be

easily accommodated. We predict that even larger and more structurally diverse species could be grafted internally.

Having established the viability of CRISP-Disp with large internal inserts, we hypothesized that a potentially vast portion of sequence space could be displayed at this position. To explore this possibility, we synthesized a pool of ~1.2×10^6 unique sgRNA variants displaying internal cassettes of 25 random nucleotides (**Figs. A8.4c, Supplementary Fig. 9**). In aggregate, this INT-N_{25}Pool activated GLuc expression at or beyond the level induced by the minimal sgRNA (**Fig. A8.4c**), implying that many of the variants formed productive CRISP-Disp complexes. To confirm this, we immunoprecipitated dCas9•INT-N_{25}Pool complexes and analyzed the copurified sgRNA sequences by deep sequencing (RIP-Seq, **Fig. A8.4d, Supplementary Figs. 9–10**). Fewer than 0.01% and 0.02% of the 1.2 million expressed sequence variants were significantly enriched or de-enriched in immunoprecipitated RNA samples, respectively; motif analysis of these variants revealed no clear sequence constraints influencing sgRNA•dCas9 complexation. Although a pool of this diversity represents a small (~1.1×10^{-9}), biased sampling of the total 25-nucleotide sequence space, we extrapolate that CRISP-Disp is not intrinsically limited by the sequence of an internal insert, provided that the modified sgRNA itself can be transcribed (**Supplementary Fig. 6a–b**).

Having established that the sgRNA scaffold can theoretically tolerate a wide variety of internally inserted structures, as a proof-of-principal we next generated a series of INT-like constructs displaying an array of functional RNA domains (**Fig. A8.4e***, left*). This compendium included motifs recognized by natural RNA-binding proteins[10, 36, 57], and artificial aptamers that bind proteins[14, 15] and small molecules[11] (**Supplementary Fig. 11**). All constructs exhibited significant direct activation (**Fig. A8.4e***, right*), indicating that all were viable CRISP-Disp

substrates, although their activities spanned a six-fold range. This variation did not appear to be caused by limiting RNA expression levels (**Supplementary Fig. 12**), implying that it may have arisen from an additional constraint–possibly RNA folding. Regardless, these results demonstrate that a diverse array of RNA structures can be functionally inserted into the sgRNA scaffold internally, although generation of a high-efficiency CRISP-Disp construct for a given motif may require some sequence optimization.

Next, we examined if INT-like constructs could be used to target structured RNA domains to endogenous loci. Having already demonstrated that the original INT design (which contains a cartridge of 3xPP7 stem-loops) could be targeted to an array of endogenous genes, *ASCL1*, *IL1RN*, *NTF3*, and *TTN*[21, 22, 41](**Fig. A8.2f**), we chose two additional domains for deployment to same loci. For this purpose, we selected the INT-S1 streptavidin aptamer[14], a smaller but potentially useful RNA-based device, and the INT-P4–P6[3xPP7] construct, the largest INT-like domain tested (**Fig. A8.4b,e**). As observed with TOP1 and INT constructs (**Fig. A8.2f**), activation of each genomic locus mirrored the results observed with our integrated GLuc reporters (**Fig A8.4f**). Moreover, in all cases, activation of each gene was specific for the gRNA sequences used. This was furthermore supported by RNA-Seq: GLuc activation by INT-P4–P6[3xPP7] induced changes in global gene expression that were essentially indistinguishable from those induced by the minimal sgRNA, TOP1 and INT constructs (**Supplementary Fig. 5b–c**).Together, these data demonstrate that INT-like constructs can be used to deploy novel RNA-based devices and large noncoding domains to specific loci genome-wide.

CRISP-Disp enables concomitant deployment of multiple functionalities

One potential advantage of CRISP-Disp is that it enables disparate functions (*e.g.* cutting,

activation, repression, imaging) to be simultaneously performed at multiple loci using a single toolset. This modularity could be achieved using an orthogonal set of high-affinity RNA•protein

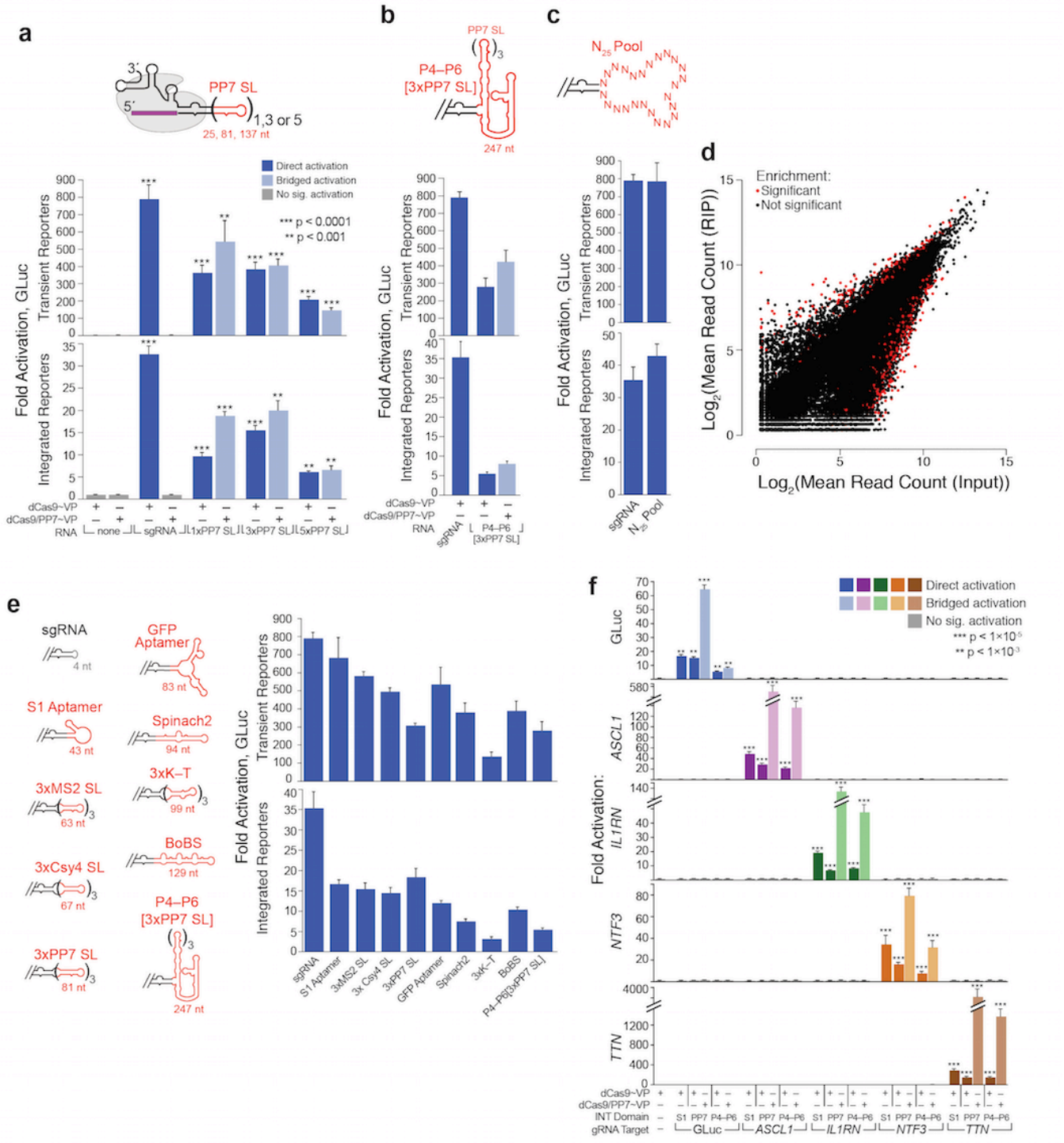


Figure A8.4. CRISPR Display with a compendium of structurally diverse RNA domains. (**a**) INT insert size has a modest effect on CRISP-Disp efficacy. Direct and bridged activation luciferase assays with constructs bearing internal cartridges of one, three or five PP7 stem-loops (insert lengths listed in red) (**b**) Functional INT inserts can be large and structurally discontinuous with the sgRNA core. (**c**) Assembly of functional CRISP-Disp INT complexes is independent of the sequence and structure near the insertion point. Direct activation assays with a mixed pool of ~1.2×10^6 unique INT–N_{25} variants (**Supplementary Fig. 9**). (**d**) dCas9 binds to nearly all expressed INT–N_{25} variants See also **(Supplementary Fig. 10**) (**e**) Assembling functional CRISP-Disp complexes bearing a wide assortment of natural and artificial RNA domains. *Left*: Cartoons depicting the INT constructs tested; insert lengths are listed below each in red. S1, an artificial streptavidin aptamer[14]; MS2 SL, cognate stem-loop for the MS2 phage coat protein[39]; Csy4 SL, cognate stem-loop for the *P. aeruginosa* Csy4 protein[36]; GFP aptamer as in[15]; Spinach2, a small-molecule-binding fluorescent aptamer[11]; K–T, a cognate kink-turn for the *A. fulgidus* L7Ae protein[57]; BoBS, "Bunch of Baby Spinach," (**Supplementary Fig. 11**). *Right*: direct activation activities of these constructs, sorted by insert length. Luciferase values are means ± standard deviation. n = 3. Student's one-tailed t-test, relative to a dCas9~VP alone negative controls. All RNA constructs were expressed from a human U6 promoter (**Supplementary Fig. 1a**). (**f**) Targeting INT-like constructs bearing RNA devices or large domains to endogenous loci. The INT-S1 aptamer ("S1") and INT-P4–P6[3x PP7] ("P4–P6") constructs were targeted to *ASCL1, IL1RN, NTF3 and TTN*. Data were generated and analyzed as in (**Fig. A8.2f**); those from the original INT-3xPP7 SL construct ("PP7") are included for comparison.

pairs: each protein would be appended with a different functional group, and targeted to distinct loci by sgRNAs that bear its cognate RNA motif. To demonstrate the plausibility of this scheme, we first confirmed that orthogonal RNA-binding proteins could be displayed on dCas9. We performed bridged activation assays using the well-established *A. fulgidus* L7Ae ribosomal protein and bacteriophage coat proteins MS2 and PP7, each fused to VP64 (Refs. 39, 57). As predicted, bridged activation was only observed when cognate sgRNA•protein pairs were coexpressed; no activation was observed with non-cognate complexes or with a minimal sgRNA (**Fig. A8.5a**).

We next sought to demonstrate the modularity of CRISP-Disp by simultaneously performing distinct functions at different loci. In this first proof-of-principle experiment, we bound dCas9 to multiple genomic targets but selectively activated only one (**Fig. A8.5b**). To accomplish this, we generated GLuc- and *NTF3*-targeting sgRNA variants (**Fig. A8.2f**) bearing internal cassettes of PP7 and MS2 stem-loops. We coexpressed orthogonally modified pairs of each targeting construct (*i.e.*, GLuc-PP7 with *NTF3*-MS2, or *vice versa*) in integrated GLuc reporter cells. When also coexpressed with dCas9~VP, we observed robust activation of both target genes, regardless of the sgRNA pair used (**Fig. A8.5b**, *left*), indicating that dCas9 had bound both loci under all conditions. However, when each sgRNA pair was coexpressed with dCas9 and PP7~VP, only the gene targeted by sgRNAs bearing PP7 stem-loops was activated (**Fig. A8.5b**, *middle*). The converse results were observed upon coexpression with dCas9 and MS2~VP (**Fig. A8.5b**, *right*). These data illustrate that CRISP-Disp enables simultaneous, modular control of gene expression, as has been recently demonstrated with constructs analogous to TOP1 (Ref. 34).

As a second demonstration of CRISP-Disp modularity, we sought to simultaneously perform two unrelated functions–transcription activation and live-cell imaging of genomic loci[25]–at different sites within the genome. To accomplish this, we first implemented a "bridged CRISPR-imaging" approach, in which the dCas9~eGFP fusion used in conventional CRISPR-imaging[25] was replaced by a ternary complex comprising dCas9, an MS2~mCherry fusion, and an INT-like sgRNA construct bearing a cassette of MS2 stem-loops (**Fig. A8.4e, Supplementary Fig. 13a)**.

In this proof-of-principle experiment, we targeted dCas9 to telomeres, as performed previously[25]. When dCas9, MS2~mCherry and the modified sgRNA were coexpressed in HEK293FT cells, we observed numerous (8–55; average of 26.6, in ~97 mCherry+ cells) fluorescent nuclear foci (**Supplementary Fig. 13b–c**). Critically, this signal was ablated by omission of dCas9, of the modified sgRNA, or by replacement of the MS2 stem-loop cassette with noncognate kink-turns (**Supplementary Fig. 13b**). Although our signal is less robust than that reported previously, it might be improved by stably expressing each construct at an optimal level[25].

To simultaneously activate one locus and image another, we employed our integrated GLuc reporter cells, performing bridged activation (using PP7~VP64, targeted by INT) at the reporter locus, and bridged imaging (using MS2~mCherry, as above) of telomeres (**Fig. A8.5c**, *top*). Upon coexpression of dCas9, PP7~VP64, MS2~mCherry and both modified sgRNAs, we observed both the induction of mCerulean CFP and the presence of mCherry nuclear foci (**Fig. A8.5c**, *lower right*). As predicted, omission of either PP7~VP64 or MS2~mCherry was sufficient to ablate the corresponding function, without perturbing the orthogonal function (**Fig. A8.5c**, *upper right* and *lower left, respectively*).

Collectively, these experiments demonstrate that CRISP-Disp allows multiple manipulations to be performed simultaneously at discrete loci, in a programmable manner (**Fig. A8.5b**, *bottom*). We note that more elaborate schemes can be achieved by expanding the repertoire of orthogonal RNA•Protein cognate pairs and fusion partners.

CRISP-Disp of autonomous RNA domains

Another advantage of CRISP-Disp is that it allows autonomously functional RNA domains, such as ribozymes, aptamers and artificial regulatory devices[7, 8, 10, 11], to be targeted to individual loci. As a preliminary illustration of this approach, we used CRISP-Disp to target "Spinach2," an artificial aptamer that binds to and induces fluorescence in a cell-permeable dye[11], to telomeres[25]. When we coexpressed a Spinach2-appended telomere-targeting sgRNA with dCas9 and treated cells with the Spinach2 ligand DFHBI-1T (Ref. 11), we observed numerous (10–20; average of 12, in ~20% of cells) nuclear fluorescent foci (**Fig. A8.5c**, *bottom right*, and **Supplementary Figure 14**). No fluorescent foci were observed in control experiments targeting the Spinach2 aptamer to the Gluc reporter (**Fig. A8.5c**, *bottom left*), or with either Spinach2 construct in the absence of dCas9 (**Fig. A8.5c**, *top*).

Although our experimental signal was less robust than that observed using the conventional dCas9~eGFP system[25], or our own MS2~mCherry system (**Figs. A8.5c** and **Supplementary Fig. 13**) it might be improved with future generations of the Spinach aptamer. Similar CRISP-Disp Spinach2 fusions might also prove useful for imaging the RNA component of CRISP-Disp complexes, which is not achievable using current technologies. Moreover, these experiments represent an important proof-of-principle, demonstrating that artificial RNA domains can be harnessed to imbue dCas9 with novel properties. Replacing Spinach2 with other

autonomously functional RNA modules could further expand the scope of CRISPR-based applications without requiring additional protein components.

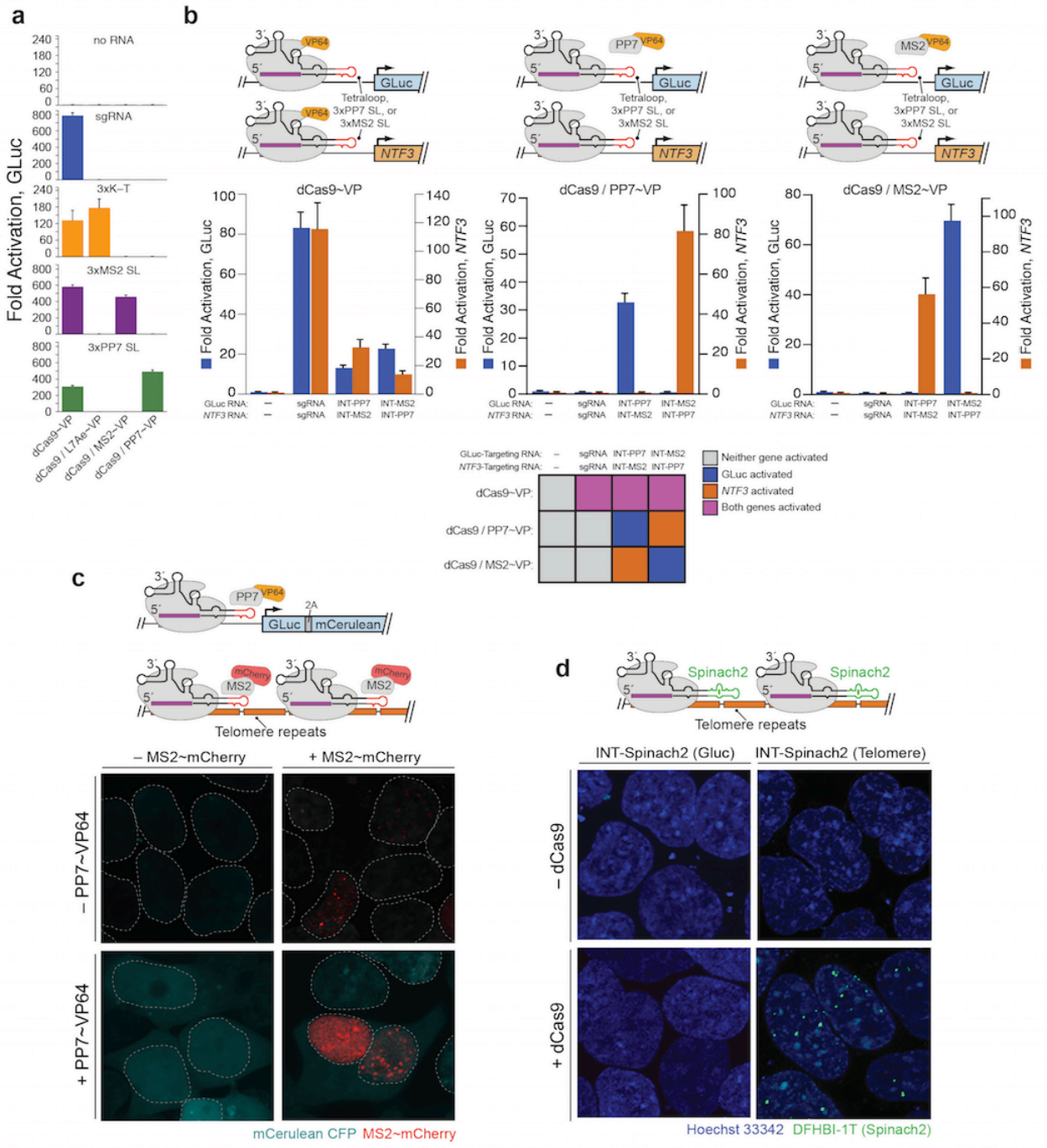


Figure A8.5. **CRISP-Disp expands the functional repertoire of CRISPR-based methods.** (**a**–**b**) CRISP-Disp enables modular, simultaneous control of multiple functions. (**a**) CRISP-Disp

enables orthogonality. RNA constructs are as defined in (**Fig. A8. 4e**); bridged activation assays employed L7Ae~VP, MS2~VP or PP7~VP. Values are means ± standard deviation, n = 3. Y-axes for "no RNA" and "3x K–T" data are different. (**b**) Performing distinct functions at multiple loci using a shared pool of dCas9. sgRNAs or INT derivatives bearing cassettes of PP7 and MS2 stem-loops ("INT-PP7," "INT-MS2") targeting GLuc and *NTF3* were simultaneously coexpressed in direct and bridged activation assays. *Left:* direct activation. *Middle*: bridged activation with PP7~VP. *Right:* bridged activation with MS2~VP. *Bottom:* schematic summarizing the results. GLuc activation was measured by luciferase assays, *NTF3* values by qRT-PCR; each is the mean ± standard deviation, n=4. (**c**) Simultaneous activation and imaging of distinct loci in integrated GLuc reporter cells, using a shared pool of dCas9. *Top:* schematic of the experimental design. INT derivatives bearing cassettes of PP7 and MS2 stem-loops targeting GLuc and telomeres, respectively, were simultaneously coexpressed with dCas9, PP7~VP64 and MS2~mCherry. 2A: a 2A "self-cleaving" peptide. *Middle* and *Bottom:* Confocal fluorescence images, at 63X magnification. All cells transiently expressed dCas9 and each INT-like sgRNA derivative. Additional fusion proteins (PP7~VP64 and MS2~mCherry) were transiently expressed as indicated. Dotted lines denote nuclear membranes. (**d**) CRISP-Disp allows locus-specific targeting of novel RNA-based functions. Aptamer-based imaging of DNA loci. *Top*: schematic of the experimental design. *Middle* and *Bottom:* Confocal fluorescence images, at 63X magnification. *Middle:* in the absence of dCas9. *Bottom:* in the presence of dCas9. The GLuc-targeting construct targets a site absent in the human genome. See also (**Supplementary Figure 14**).

Discussion

In the short time since its initial characterization[17, 18], *Sp.* CRISPR-Cas9 has already been coopted for a host of exquisitely powerful genome modification and regulatory technologies[23-31]. We envision that CRISPR-Display, which limits the function dCas9 to DNA-targeting and "outsources" all other roles to RNA-based domains, will provide the basis for an even wider array of methods (**Fig. A8.6**).

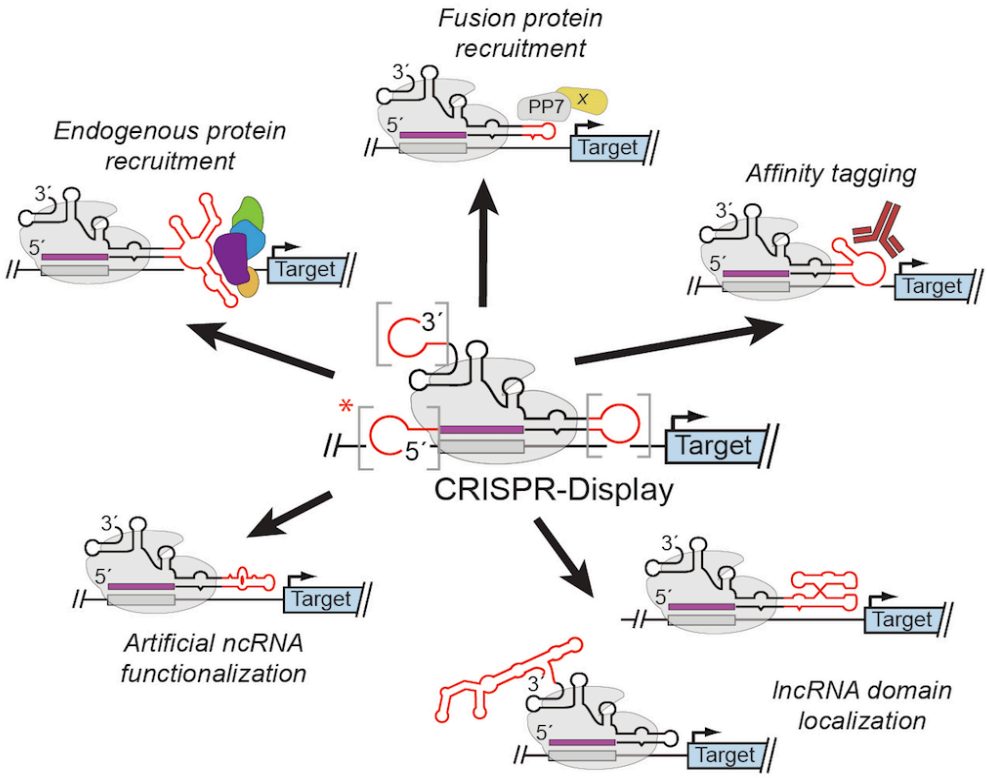


Figure A8.6. CRISPR-Display. Schematic summarizing some of the novel locus-targeted functionalities made possible by CRISPR-Display. Theoretically, multiple functions can be targeted to discrete sets of loci simultaneously. Note that accessory domains on the sgRNA 5′ end (*asterisk*) are not tolerated by most expression systems.

The present studies establish a framework for the implementation of CRISP-Disp, allowing us to propose preliminary "best practices" for construct design. By far the most robust insertion site for exogenous RNA domains is the "engineered loop," (**Fig. A8.1b**), which can tolerate diverse, large and structurally varied inserts. This site is therefore ideal if the RNA domain of interest does not require an exposed terminus, and would not be functionally constrained by placement within a stem-loop. Although we have not tested if other internal insertion points can accommodate exogenous sequences with this degree of modularity, we anticipate that other such amenable points exist, given the structural plasticity of the sgRNA core[32, 33], and its ability to accommodate short stem-loops at other positions[34, 35]. We furthermore hypothesize that certain internal inserts might perturb sgRNA folding, potentially in a guide-dependent manner, though we have not examined this exhaustively. In lieu of INT, display on the 3′ terminus, as in TOP1, is viable, although overall efficacy may be limited by the local structure near the attachment point[34]. In both cases, the conventional U6 expression system was the most robust we employed. Surprisingly, although the U6 promoter naturally drives a product of only ~100 nt, we were able to generate transcripts several times that length at high levels (**Supplementary Fig. 12**); the overall length limitation of this system remains unclear. For substantially larger structures, however, or those containing stretches of poly(uridine), several options are available (Refs. 28, 46, 47, and **Supplementary Fig. 6**). For example, the activities of constructs expressed from the U1 promoter (a "nonstandard" Pol II promoter) nearly rivaled those driven by U6. If a canonical 5′-cap is required, expression from the CMV/3′Box or CMV/MASC systems should suffice. If the ncRNA domain requires display on the sgRNA 5-terminus to function, then only the CMV/3′Box system appears sufficient.

Collectively, the ability of CRISPR-Disp to ectopically localize diverse RNA cargos (of up to at least 4.8 kb) to targeted DNA loci has several implications. First, CRISP-Disp has the unique ability to simultaneously target multiple distinct cargos to different genomic locations while maintaining cargo-specific functions at each target site. While such modularity can be alternatively achieved by simultaneously expressing multiple Cas9 orthologs[58], this is cumbersome and not immediately extensible. In contrast, the breadth of distinct functions accessible in a single CRISP-Disp experiment is theoretically limited only by the number of orthogonal RNA domains or RNA/binding-protein pairs available. For example, by appending a set of small, RNA-binding proteins (as in **Fig. A8.5a**) with different functional domains, one might simultaneously activate and repress transcription, epigenetically mark histones and DNA, induce double-stranded breaks and image discrete sets of target loci, using a common toolkit. Our preliminary experiments, which targeted discrete–though unrelated–genomic loci for binding, activation and imaging (**Fig. A8.5b–c**) provide a small glimpse into the potential power this approach holds. Similar experiments are now conceivably possible in which, for example, the dynamics of individual genomic loci are imaged while their regulatory factors are transcriptionally activated, or ectopically localized nearby.

Second, since CRISP-Disp is not intrinsically limited by RNA length (**Figs. A8.3, 4a,b**), it may provide a method for the locus-targeted reconstitution of natural regulatory RNAs, which could dramatically advance the study of lncRNA mechanism[1, 2, 6 3]. For example, CRISP-DISP could be used to bring a long noncoding RNA–or subdomains within that RNA–to a given target site, in order to determine if the RNA molecule alone is functionally sufficient when decoupled from the act of its transcription. The preliminary data presented here (**Fig. A8.3f–h**) demonstrate the broader plausibility of such experiments, although more thorough investigation into CRISP-

Disp lncRNA reconstitution will require consideration of several other variables, many of which may depend on the particular lncRNA under examination. *Bona fide* lncRNA reconstitution may require optimization of the RNA construct used, consideration of the local chromatin environment to which it is being delivered, or the presence of other factors absent in HEK293T cells. Furthermore, the assay used to survey reconstitution may require a more subtle readout than the reporter gene system we employed.

Third, CRISP-Disp expands the scope of Cas9-based methods by making available the broad functional repertoire of artificial ncRNAs and devices[7, 8, 10, 11, 13, 14]. While our initial illustration of this principal focused on imaging (**Fig. A8.5d**), other applications built from aptamers, ribozymes, sensors, processors and scaffolds[7, 8, 10, 11, 13, 14] are possible. For example, decorating loci with orthogonal RNA-based affinity tags[14] could enable multiplexed dissection of locus-specific proteomes, transcriptomes and higher-order chromatin structures. Additionally, CRISP-Disp with custom RNA scaffolds might allow enzymatic activities to be uniquely targeted to discrete subnuclear sites[10].

Finally, CRISP-Disp may provide a platform for the isolation of novel functional RNA domains. That the sgRNA scaffold can accommodate an expansive breadth of sequences inserted within its engineered loop (**Fig. A8.4**), including pools of random sequences (**Fig. A8.4c,d**), suggests that it could be used as the backbone for selection *in vitro*, or in living cells. Such selections might yield, for example, aptamers that sequester or recruit endogenous protein complexes to target loci, or ribozymes that tag nearby molecules with affinity tags or markers. Amassing a large repertoire of natural and artificial aptamers would allow, for example, the multiplexed retargeting of host proteins, genome-wide.

In summary, the proof-of-principle experiments presented here hint to the larger scope of novel applications CRISPR-Display enables. Given the diverse roles ncRNAs play in biology[1-3], and for which artificial RNAs have been engineered[7, 8, 10, 11, 13-15], we anticipate that this application list is, at best, preliminary.

Online Methods

Plasmid synthesis

Mammalian expression and reporter constructs were generated using standard restriction enzyme-based and ligation-independent cloning methods. Components were acquired as follows: The T7 promoter-targeting TALE[37] was the generous gift of Feng Zhang (Broad Institute). *Gaussia* and *Cypridina* luciferases were derived from pGLuc-Basic and pCLuc-Basic, respectively (New England Biolabs). dCas9 (*S. pyogenes* D10A/H841A Cas9) was isolated from Addgene plasmid 47754, the EF1a promoter from Addgene plasmid 11154, mCerulean from Addgene plasmid 23244, Venus from Addgene 15753 and the human Ubiquitin C promoter (hUBCPro) used to drive expression of L7Ae~VP, PP7~VP (**Supplementary Fig. 2a**, *bottom*) and MS2~mCherry (**Supplementary Fig. 13a**, *right*) from Addgene plasmid 17627. All other components were synthesized *de novo* from gBlocks or from smaller synthetic oligonucleotides (Integrated DNA Technologies). The backbone for Lentiviral reporter constructs was derived from pLenti6.3/TO/V5-DEST (Life Technologies), from which the Tet-reponsive promoter and Gateway cloning sites were removed. The backbone for the T7 TALE and MS2~VP constructs was derived from pcDNA3.1(+) (Life Technologies) in which the Neomycin expression cassette was removed. All other constructs were cloned into pNEB193 (New England Biolabs).

L7Ae, MS2 and PP7 were codon-optimized for expression in human cells and synthesized as gBlocks (Integrated DNA Technologies). The PP7 construct consists of two

tandem copies of the non-aggregating ΔFG mutant[39] joined by a flexible seven amino acid linker with the sequence GSTSGSG (**Supplementary Fig. 2a**, *bottom*). Similarly, the MS2 construct consists of two tandem copies of the non-aggregating V75E/A81G mutant [60] joined by the same linker. L7Ae was designed according to a published sequence[57].

INT-like internally appended constructs (**Figs. 4,5**) were cloned as follows. We first cloned an INT general-purpose cloning vector, "sgINTgpc," containing the following pertinent sequence:

*GATCTAGATACGACTCACTAT*GTTTAAGAGCTATGCTGCGAATACGAGAA**GTCTTC**TTTTTT**GA**
AGACAATCGTATTCGCAGCATAGCAAGTTTAAATAAGGCTAGTCCGTTATCAACTTGAAAAAGT
GGCACCGAGTCGGTGCTTTTTTT

…Wherein italicized nucleotides denote the GLuc-targeting protospacer sequence, underlined nucleotides denote an extended sgRNA stem1 (Ref. 25) and bold nucleotides denote two outward-facing *BbsI* restriction sites. This cassette is under expression of a human U6 promoter (*not shown*). Inserts cloned into this backbone had the general format: 5′–CGAG–[Insert]–CTCGT–3′, wherein underlined nucleotides denote the sticky ends used for cloning; the additional C following the insert restores base-pairing at the end of stem1. These inserts were generated by PCR and restriction digestion with *BbsI*, or by annealing synthetic, 5′-phosphorylated oligonucleotides (following the protocol used for the N_{25} pool, *below*). Inserts were ligated into *BbsI*-digested, gel-purified sgINTgpc using the Quick Ligation Kit (New England Biolabs).

All sgRNAs and derivatives were initially cloned bearing a GLuc-targeting protospacer. *ASCL1-, IL1RN-, NTF3-*, *TTN-* and telomere-targeting constructs (**Figs 2f, 4f** and **5b,c**) were derived from these parental constructs using an inverse-PCR method, using a forward primer that

anneals downstream of the protospacer and a reverse primer that anneals to the 3′-end of the U6 promoter. Namely, PCR products were amplified with primers of the general format:

Forward: TAGTA**GAAGAC**AAXXXXXXXXXXXXXGTTTAAGAGCTATGCTGCGAATACG
Reverse: TAGTA**GAAGAC**AAYYYYYYYYYYYYYGGTGTTTCGTCCTTTCCAC

…Wherein bold nucleotides denote *BbsI* restriction sites; X's denote nucleotides 9–21 of the new protospacer sequence; Y's denote the reverse complement of nucleotides 1–9 of the new protospacer; underlined nucleotides are reverse complementary to one another. PCR products were purified using the QIAgen PCR cleanup kit, digested with *BbsI* and *DpnI*, purified again and quantified by UV-vis spectroscopy. Products (25 ng, in 11 μL final) were self-ligated using the Quick Ligation Kit (New England Biolabs). All plasmid sequences were confirmed by Sanger sequencing (GeneWiz) prior to use.

Cloning the N_{25} Pool

Pool oligonucleotides (Integrated DNA Technologies) were as follows:

5′—[P]—CGAGNNNNNNNNNNNNNNNNNNNNNNNC—3′
5′—[P]—ACGAGNNNNNNNNNNNNNNNNNNNNNNN—3′

…Wherein 5′–[P] denotes a 5′ Phosphate, and N denotes an equimolar mixture of all four nucleotides. Oligonucleotides were resuspended in annealing buffer (10 mM Tris, pH 7.0, 50 mM NaCl) to 100 μM. 10 μL of each oligo were mixed in a 0.2 mL PCR tube; this mixture was heated to 95°C for 10 minutes and slowly annealed to 25°C over the course of two hours in a thermocycler. The reaction was snap-cooled on ice and diluted 100-fold with ice-cold annealing buffer. 1 μL of this diluted duplex mix was ligated into 25 ng of *BbsI*-cut sgINTgpc, in 12 μL final volume, using the Quick Ligation Kit (New England Biolabs). The entire reaction was transformed into 120 μL of XL10-Gold ultracompetent cells (Agilent), plated onto 12 LB Ampicillin plates and grown overnight at 37°C. Seven bacterial colonies were picked for Sanger

sequencing (**Supplementary Fig. 9**), and the remainder were pooled by scraping the plates into 100 mL of liquid LB(Amp). Bacteria were pelleted by ultracentrifugation, and the plasmid pool was harvested in a single plasmid maxi-prep (QIAgen) (**Supplementary Figs. 9–10**).

Cell Culture, stable and transient transgene expression

HEK 293FT cells (ATCC) were maintained on gelatinized plates in high glucose Dulbecco's modified Eagle's medium (DMEM, Gibco), supplemented with 10% FBS, 1x penicillin/streptomycin and 2 mM L–Glutamine (Gibco). Cells were grown at 37°C and 5% CO_2 in a humidified incubator. Lentiviral particles were generated using standard second generation packaging plasmids, in 293T cells. Integrated reporter cells were generated as follows: 250,000 HEK 293FT cells were plated per well of a gelatinized six-well dish and incubated overnight. Growth media was thereafter removed; cells were washed once in warmed PBS, and supplied with 1.7 mL fresh warmed media supplemented with 200 μL CLuc reporter lentivirus and 8 μg/mL polybrene. After 24 hr this process was repeated with a second dose of CLuc virus. Cells were subsequently passaged onto 10 cm gelatinized plates and selected with 2 μg/mL puromycin. CLuc reporter cells were then plated onto gelatinized six-well dishes and transduced with GLuc reporter lentivirus following the same transduction protocol. GLuc-transduced cells were not selected with hygromycin prior to use. A lentiviral variant of our EF1α-dCas9 construct (**Supplementary Fig. 1a**) was also used for aptamer-based imaging (**Fig. 5d**) following the same transduction protocol without antibiotic selection. To enrich for cells that expressed low levels of dCas9 (Ref. 25), we transiently transfected GLuc reporter, U6-INT and PP7~VP plasmids, as in analytical luciferase assays (*see below*), and collected $GLuc^+$ cells by FACS.

Transient transfections were per performed using Lipofectamine 2000 (Life Technologies), following the manufacturer's protocol. For luciferase assays, 125,000 cells in 0.6

mL media were plated per well of gelatinized 12-well dishes and incubated overnight. Transfection mixes contained 33 ng of each luciferase reporter plasmid (where appropriate), 59 ng of dCas9 or dCas9~VP plasmid, 66 ng of PP7~VP, L7Ae~VP or MS2~VP (where appropriate), 11.6 ng of U6–driven or 542 ng of Pol II-driven sgRNA variants. For experiments using TOP3 and TOP4 (**Fig. 2a, Supplementary Fig. 6b**), 11.6 ng of a separate U6-driven gRNA plasmid was also included. For FACS, (**Fig. 2d**, **Supplementary Figs. 1c, 6d**) transfection mixes also contained 10 ng of an mCherry cotransfection control. In all cases, the total transfected plasmid mass was brought to 750 ng per well using pNEB193 (New England Biolabs) in 18 μL final volume, with 2.25 μL Lipofectamine 2000.

For RNA immunoprecipitation (RIP) and RIP-Seq experiments, 2.1 million cells in 10 mL growth media were plated onto gelatinized 10 cm dishes and grown overnight. Transfection mixes were as described above, but all masses and volumes were scaled 15.7–fold to account for the increase in growth area and cell number. RIP transfection mixes included each luciferase reporter to independently monitor CRISP-Disp function.

To test CRISP-Disp function at endogenous loci (**Figs. 2f, 4f**), cells were plated in gelatinized 12-well dishes as in standard luciferase assays. Transfection mixes were similar to those described[21], and contained 500 ng dCas9 or dCas9~VP plasmid, 500 ng GLuc-Targeting sgRNA construct or 500 ng of a mix containing equal masses (125 ng each) of four *ASCL1-, IL1RN-, NTF3-* or *TTN*-targeting constructs[21, 22, 41]. Where appropriate, 556 ng of PP7~VP plasmid was also included. All mixes were brought to 1556 ng per well using pNEB193, in 38 uL final volume, with 4.7 μL Lipofectamine 2000.

For multiplexing experiments (**Fig 5b**) cells were plated in gelatinized 12-well dishes as above. Transfection mixes contained 250 ng dCas9 or dCas9~VP, 250 ng GLuc-targeting

sgRNA variant, 250 ng of a mix containing equal masses (62.5 ng each) of four *NTF3*-targeting constructs, and 278 ng of PP7~VP or MS2~VP, where appropriate. In all cases the total transfected mass was brought to 1028 ng using pNEB193, in 30 μL volume, with 3.1 μL Lipofectamine 2000.

In bridged imaging experiments (**Fig. 5c, Supplementary Fig. 13**), 80,000 cells in 1 mL growth media were plated per well of untreated Nunc Lab-Tek glass two-chamber slides (Thermo Scientific). Twenty-four hours thereafter, growth media was changed, and cells were transfected with 440 ng dCas9 and 235 ng of each modified sgRNA. Where appropriate, 440 ng of PP7~VP64 and/or 100 ng of MS2~mCherry were included. The total transfected mass was brought to 1500 ng with pNEB193, in 11.4 μL, with 4.5 μL Lipofectamine 2000, according to the manufacturer's protocol. For aptamer-based imaging (**Fig. 5d**), 80,000 dCas9-transduced ("+dCas9," *see above*) or untransduced ("–dCas9") cells in 1 mL growth media were plated per well of Nunc Lab-Tek glass two-chamber slides that had been treated as follows. Wells were coated with 100 μg/mL poly-L-lysine (Millipore) overnight at 4°C. The next day, wells were washed twice with ddH_2O, UV sterilized for five minutes in a biosafety cabinet, coated with 100 μg/mL rat collagen-I (Corning) and 50 μg/mL laminin (Life Technologies) for two hours at 37°C, and dried prior to plating cells. Transfections were performed 24 hours thereafter, with 600 ng (telomere- or GLuc-targeting) INT-spinach2 construct, 600 ng of pNEB193 and 4.5 ng of an mCherry cotransfection control, in a total volume of 11.4 μL, with 3.8 μL Lipofectamine 2000, according to the manufacturer's protocol. All live-cell imaging experiments were performed 48–72 hours post-transfection (*see below*).

Luciferase and FACS Assays

Luciferase assays were performed using the BioLux *Gaussia* and *Cypridina* Luciferase Assay kits (New England Biolabs), following the manufacturer's protocols. Growth media (200 μL) was harvested three days after transfection and, if not used immediately, was stored in the dark at 4°C in parafilm-sealed 96-well dishes. 20 μL of each experimental sample was manually pipetted into black-walled 96-well plates (Corning) and assayed using a FLUOstar OPTIMA Luminometer equipped with automatic injectors (BMG Labtech). *Gaussia* and *Cypridina* assays were performed in parallel. For each, a single empirically determined gain was applied to all samples within an experimental series. Each sample was injected with 50 μL of luciferase assay buffer and mixed for two seconds prior to data acquisition. Signal was integrated over 20 seconds using an open (unfiltered) top-down optic.

For each sample, experimental raw luciferase signals were background-subtracted, and the ratio of Luciferase values, (GLuc/CLuc), was calculated. Biological replicates (at least three per experiment) were used to calculate a mean value, <GLuc/Cluc>. Fold activation was then calculated relative to a control sample in which dCas~VP was expressed in the absence of an sgRNA construct:

$$Fold\ Activation = \frac{\left\langle \frac{GLuc}{CLuc} \right\rangle_{(Experimental\ Sample)}}{\left\langle \frac{GLuc}{CLuc} \right\rangle_{(dCas9\sim VP\ alone)}}$$

Statistical significance testing likewise used this dCas9~VP control as the basis of comparison.

For FACS assays, cells were propagated and transfected in gelatinized 12-well dishes, as described for luciferase assays, and analyzed three days after transfection. Cells were harvested by trypsinization, quenched by the addition of chilled growth media, diluted threefold in chilled staining media (Hank's Balanced Salt Solution (HBSS, Gibco), supplemented with 2% Donor

Bovine Serum (DBS, Atlanta Biologicals)), and pelleted at 200 g in a swinging bucket rotor. Cells were resuspended in chilled staining media and analyzed on a BD LSR II Flow Cytometer (BD Sciences), equipped with HcRed, CFP and YFP filters. Voltages, compensations and gates were empirically determined using unstained and single color controls, via standard methods. 100,000 $mCherry^+$ cells were recorded from each sample.

RNA Immunoprecipitation (RIP)

Cells were propagated on gelatinized 10-centimeter dishes, transfected as described above, and harvested three days after transfection. Thereafter, RIP was performed essentially as described previously[61]. Growth media was collected, and cells were washed twice with 10 mL room temperature PBS (Gibco). Cells were crosslinked by incubation in 0.1% (v/v) formaldehyde in PBS for 10 minutes at room temperature, under very gentle agitation. Crosslinking was quenched by the addition of Glycine to 133 mM and gentle agitation for an additional five minutes at room temperature, after which the liquid phase was aspirated. Crosslinked cells were washed twice with room temperature PBS, harvested by scraping, allotted into samples of 1×10^7 cells (typically three samples per 10 cm dish), and pelleted at 200 g in a swinging bucket rotor. PBS was aspirated and cell pellets were flash-frozen in liquid nitrogen and stored at –80°C until use.

Cell pellets were thawed on ice, gently resuspended into 1 mL of ice-cold RIPA(+) buffer (standard RIPA supplemented with 0.1 U/μL RNAseOUT (Life Technologies), 1x EDTA-free Proteinase Inhibitor Cocktail (Thermo Scientific) and 0.5 mM DTT), and lysed for 10 minutes at 4°C with end-over-end agitation. Samples were then sheared using a Branson Digital Sonifier 250 (Emerson Industrial Automation) at 10% amplitude for three 30-second intervals (0.7 seconds on + 1.3 seconds off), with 30-second resting steps between intervals. Samples were

held in ice-cold metal thermal blocks throughout sonication. Sheared samples were then clarified by ultracentrifugation and diluted with 1 mL each of ice-cold Native Lysis Buffer(+) (25 mM Tris, pH 7.4, 150 mM KCl, 5 mM EDTA, 0.5% (v/v) NP-40, supplemented with inhibitors and DTT, as above), filtered through a 0.45 μm syringe-mounted filter, and flash-frozen in liquid nitrogen before use.

Clarified lysates were thawed on ice and pre-cleared by incubation with buffer-equilibrated magnetic Protein G beads (Life Technologies) for 30 minutes at 4°C, with end-over-end rotation. 100 μL aliquots were removed and frozen, to serve as "input" normalization controls. Cleared lysates corresponding to 5×10^6 cells were then incubated with 6 μg rabbit anti-FLAG (SIGMA) or Rabbit normal IgG (Cell Signaling Technology), for two hours at 4°C with end-over-end rotation. Buffer-equilibrated magnetic Protein G beads were then added and the samples were again rotated end-over-end for one hour at 4°C. Beads were collected and twice washed with Native Lysis Buffer(+) for 10 minutes at 4°C, with end-over-end rotation. Immunoprecipitated RNA was thereafter isolated as described below.

RNA Isolation, Quantitative RT–PCR and mRNA Seq

Whole cell RNA (**Figs. 2f, 4f**, **5b and Supplementary Figs. 5 and 12**) and RNA from subcellular fractions (**Supplementary Fig. 7**) were isolated by extraction with Trizol and Trizol-LS Reagent (Life Technologies), respectively, following the manufacturer's protocols. RNA was precipitated with isopropanol using GlycoBlue (Life Technologies) as a carrier, and subsequently purified using RNEasy spin columns (QIAgen), following the manufacturer's "RNA Cleanup" protocol, with on-column DNase treatment.

RNA from RIP and RIP-Seq experiments (**Figs. 2e, 3b,e,f, 4d**) was isolated as follows. Following RIP (*see above*), protein G beads were suspended in 56 μL nuclease-free water, and

processed alongside input samples (56 μL; 5.6% of the total). All samples were brought to 100 μL with 3x Reverse Crosslinking Buffer (final concentrations: 1x PBS, 2% N-Lauroyl Sarcosine, 10 mM EDTA, 5 mM DTT, 0.4 U/μL RNaseOUT and 2 mg/mL proteinase K (Ambion)). Formaldehyde crosslinks were reversed by incubation in a thermocycler at 42°C for one hour, and then 55°C for one hour. RNA was thereafter purified using four volumes (400 μL) Agencourt RNAClean XP Beads (Beckman Coulter), following the manufacturer's protocol, and eluted into 30 μL nuclease-free water. Residual DNA was removed by treatment with 5 U RNase-free DNAs (RQ1, Promega) in 50 μL, following the manufacturer's protocol. RNA was subsequently purified using four volumes (200 μL) Agencourt RNAClean XP beads, eluted into 20 μL nuclease-free water, and stored at –20°C until use.

cDNA was synthesized using SuperScript III Reverse Transcriptase (Life Technologies), according to the manufacturer's protocol, priming from anchored oligo-dT_{21}, random hexamers (Life Technologies) or a gene specific primer (Integrated DNA Technologies), where appropriate. Target RNA abundance was quantified by qRT-PCR on a 7900HT Fast Real-Time PCR System (Applied Biosystems), using Rox-normalized FastStart Universal SYBR Green Master Mix (Roche) and gene-specific primers, in quadruplicate. Non-reverse-transcribed RNA was used as a negative control. "Clipped" data were processed using Realtime PCR Miner[62], to calculate C_T and primer efficiency values. Bulk gene expression measurements (**Figs. 2f, 4f**, **5b** and **Supplementary Fig. 12**) were normalized to a GAPDH internal control; RIP measurements were normalized to input RNA levels. In subcellular fractionation experiments (**Supplementary Fig. 7**), the yield of RNA in each compartment was quantified relative to the unfractionated input level, as in RIP experiments. Data analysis was performed using standard methods.

For global gene expression analysis (**Supplementary Fig. 5**), Poly(A)$^+$ mRNA seq libraries were prepared using the TruSeq RNA sample preparation kit, v2 (Illumina) as described[61]. Libraries were pooled and subjected to 50 cycles of paired end sequencing, followed by 25 cycles of indexing, on two lanes of an Illumina HiSeq 2500 (FAS Center for Systems Biology, Harvard). For characterization of gene expression, sequencing reads were mapped to a custom gene set comprising UCSC known human genes (hg19), appended with dCas9, GLuc, CLuc and sgRNA constructs, using TopHat2 with default options[63]. Differential analysis of gene expression was assessed using Cuffdiff2 with default options[64]. Genes plotted in (**Supplementary Fig. 5**) were restricted to the top 75% of expressed genes, based on FPKM values.

Error Propagation and Reproducibility

For Luciferase and qRT-PCR assays, experimental uncertainties were propagated as described previously[65]. Namely, given S, the sum or difference of values A, B, uncertainly was calculated using the formula:

$$\sigma_S = \sqrt{(\sigma_A)^2 + (\sigma_B)^2}$$

…wherein σ_A and σ_B are the measurement errors of A and B, respectively. For P, the product or quotient of values A and B, uncertainty was calculated using the formula:

$$\sigma_P = P \times \sqrt{\left(\frac{\sigma_A}{A}\right)^2 + \left(\frac{\sigma_B}{B}\right)^2}$$

The uncertainty of other functions, F(x), was calculated using the first derivative approximation:

$$\sigma_{f(x)} = \sigma_x \times f'(x)$$

Sample sizes were determined in accordance with standard practices used in similar experiments in the literature; no sample-size estimates were performed to ensure adequate power to detect a

prespecified effect size. Experiments were neither randomized nor blinded to experimental conditions. No samples were excluded from analysis.

Subcellular Fractionation

Cytoplasmic and nuclear fractions (**Supplementary Fig. 7**) were isolated as described in[66, 67]. Briefly, cells were grown and transfected in gelatinized 10-cm dishes, as described for RIP experiments, above. Three days after transfection, cells were harvested by trypsinization, quenched with growth media, pelleted and washed thrice with ice-cold PBS. Cells were gently resuspended in five packed cell pellet volumes ("cv's") of ice-cold Cyto Extract Buffer(+) (20 mM Tris, pH 7.6, 0.1 mM EDTA, 2 mM $MgCl_2$, supplemented with 0.5 U/μL RNAseOUT and 1x EDTA-free Proteinase Inhibitor Cocktail), and swollen by incubation at room temperature for two minutes, and on ice for ten minutes more. Cells were then lysed by addition of CHAPS to 0.6% final, gentle pipetting, and two passages through a syringe equipped with a 20G needle. Lysate was clarified by centrifugation at 500g in a tabletop microcentrifuge at 4°C; 70% of the resulting supernatant was retrieved as the cytoplasmic fraction. The pellet, corresponding to nuclei and cell debris, was washed twice by gentle resuspension into five cv's of Nuclear Wash Buffer(+) (Cyto Extract Buffer, supplemented to 0.6% CHAPS and with inhibitors, as above), followed by centrifugation at 500g. Washed nuclei were gently resuspended into two cv's of Nuclei Resuspension Buffer(+) (10 mM Tris, pH 7.5, 150 mM NaCl, 0.15% (v/v) NP–40, supplemented with inhibitors, as above) layered onto a cushion of five cv's Sucrose Buffer(+) (10 mM Tris, pH 7.5, 150 mM NaCl, 24% (w/v) Sucrose, plus inhibitors), and pelleted at 14,000 rpm in a tabletop microcentrifuge at 4°C. The resulting pelleted nuclei were resuspended into two cv's of ice-cold PBS and pelleted at 500g. We confirmed the success of our fractionations by two methods: western blotting and qRT-PCR. In western blots, aliquots of whole cell lysate, the

cytoplasmic fraction and PBS-suspended nuclei were probed using antibodies against (α/β)-Tubulin and Fibrillarin (Cell Signaling Technology). For qPCR, extracted RNA (*see above*) was quantified using primers against the cytoplasmic ncRNA *SNHG5* and the nuclear ncRNA *XIST*.

N_{25} RNA Library Preparation, Sequencing and Analysis

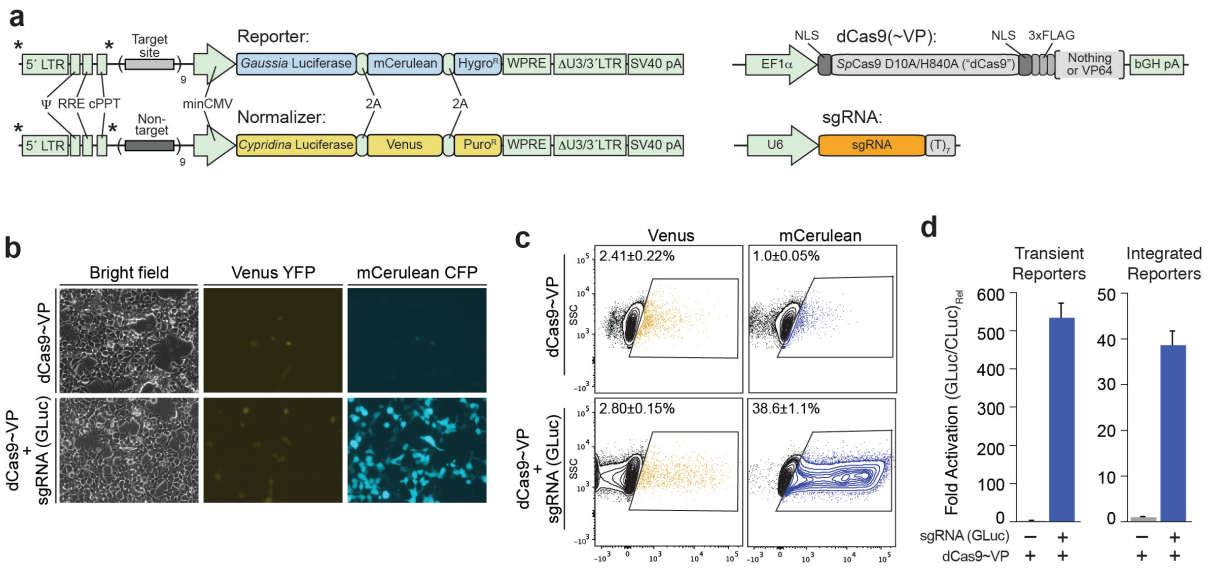
For the N_{25} RIP-Seq experiment (**Fig 4d**), cell growth, transfection, RIP and RNA preparation were performed as described above, in triplicate. Seven deep sequencing libraries were prepared: one from the starting plasmid pool, three from replicates of the input RNA, and three from replicates of the immunoprecipitated RNA. The plasmid pool library was generated directly via PCR, using 5 ng of plasmid template in a 50 μL reaction, amplified through 19 cycles of PCR with Pfu Ultra II HS polymerase (Agilent), according to the manufacturer's protocol. Gene-specific PCR primers that bracketed the N_{25} insertion site, appended with standard Illumina adapters and indexes, were used (**Supplementary Fig. 9**). For each input and RIP library, 10 ng RNA was reverse-transcribed in 20 μL as described above, using a gene specific primer. Each cDNA reaction was used in its entirety as PCR template, using the same primer design as was used for the plasmid pool, but with different Illumina indexes. The pools were amplified in 200 μL, through 26 cycles of PCR with Pfu Ultra II HS polymerase (Agilent), according to the manufacturer's protocol. The resulting deep sequencing libraries were purified twice with 1.0 volume of Agencourt AMPure XP Beads (Beckman Coulter), according the manufacturer's protocol, and eluted in EB Buffer (QIAgen). The plasmid pool library contained traces of high molecular weight contaminants (*not shown*) that were removed by "reverse selection:" the sample was treated with 0.65 volumes of AMPure XP Beads, and the unbound fraction was retained. The integrity and concentration of each final library was measured using a "DNA High Sensitivity" assay on an Agilent 2100 model Bioanalyzer (**Supplementary Fig. 9**).

Libraries were denatured in 50 mM NaOH, diluted in buffer HT1 (Illumina) and combined to yield a 20 pM pool, according to standard protocols. This pool was doped with TailorMix Indexed PhiX Control Library (SeqMatic), at a ratio of 7:3 N_{25}:PhiX, and sequenced on two lanes of an Illumina HiSeq 2500 (FAS Center for Systems Biology, Harvard) for 150 cycles, followed by 25 cycles of indexing. Random insert sequences were extracted from raw sequencing reads by removing the constant sequences abutting each side of the insertion point. The number of occurrences of each random sequence within each individual sample was then tabulated. Sequence counts were used to calculate enrichment using DESeq2 (Ref. 68).

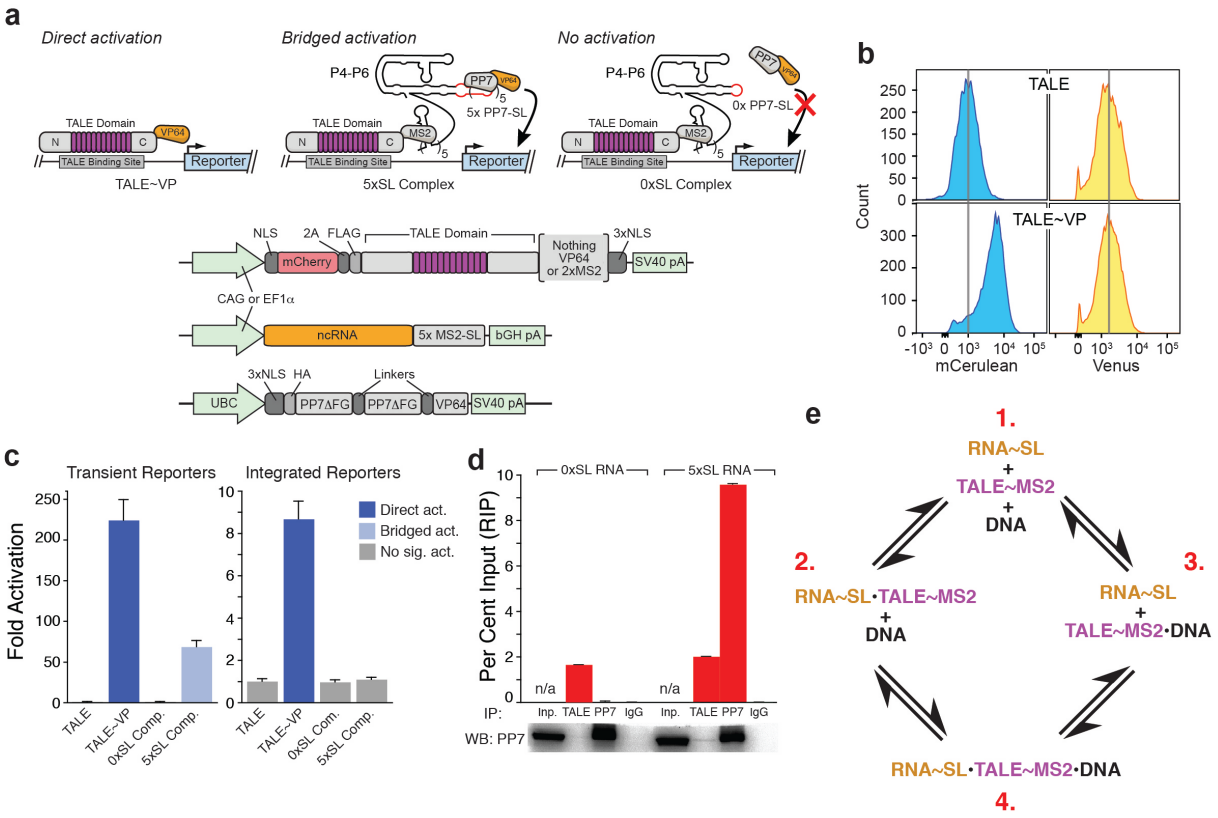
Live Cell Imaging

Images in (**Supplementary Fig. 1b**) were collected on an Axio Observer D1 system (Zeiss), equipped with eYFP and eCFP filters. Live fluorescence images (**Fig. 5c**–d, and **Supplementary Figs. 13–14**) were taken with an LSM 700 Inverted Confocal Microscope (Harvard Center for Biological Imaging), with an aperture setting of 1 A.U., using the DAPI filter for Hoechst 33342, the CFP filter for mCerulean, the mCherry filter for mCherry and the FITC filter for DFHBI-1T, where appropriate. In bridged imaging experiments (**Fig. 5c**, **Supplementary Fig. 13**), cells were imaged two days post-transfection, in their growth media. Images are max-merges of 37–47 Z-stacks, taken with a step size of 0.33 μm, at 63X magnification. For aptamer-based imaging (**Fig. 5d** and **Supplementary Fig. 14**), growth media was replaced with imaging media (Fluorobrite DMEM (Life Technologies), 25 mM HEPES, 5 mM MgSO4, 1 μg/ml Hoechst 33342 (Life Technologies), and 20 μM DFHBI-1T (Lucerna)) for 30 minutes at 37°C. Images in (**Fig. 5c–d**) are max-merges of 20–30 Z-stacks, taken with a step size 0.35 μm, at 63X magnification.

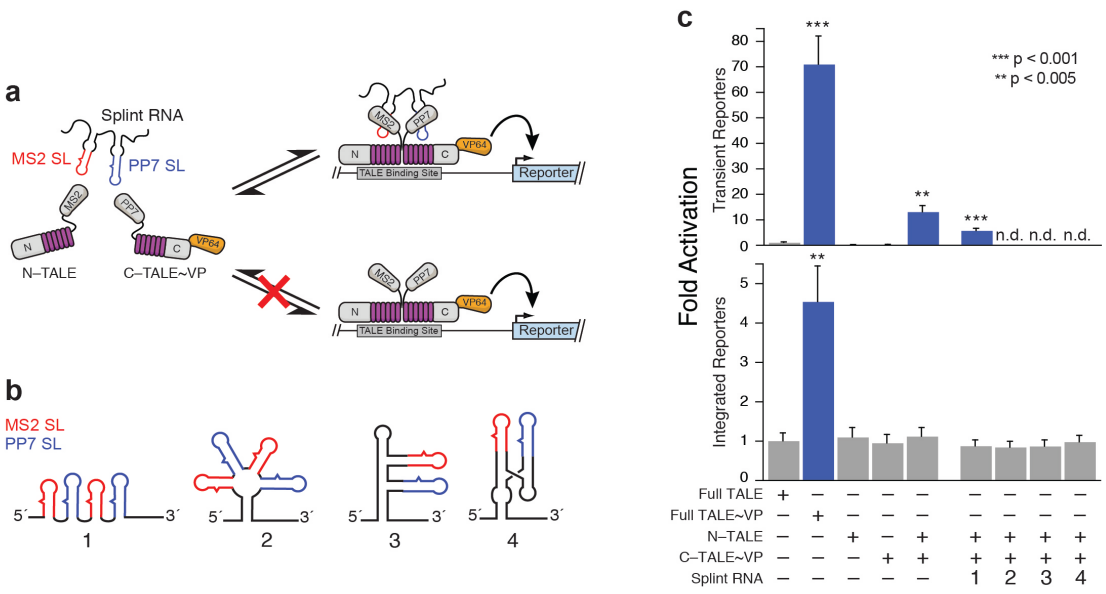
Supplementary Figures



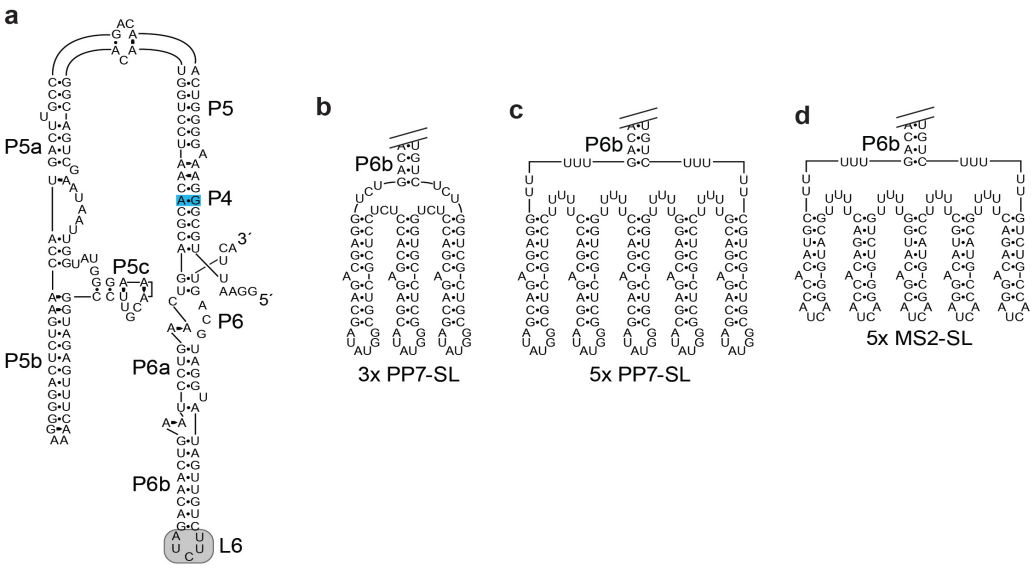
Supplementary Figure 1: Transcirption activator assay design and positivve control experiments. (a) Expression contructs. **(b-d)** Reporter system visualization in HEK293s, in live cells (b), by FACS (c), by luciferase assay (d) (n=3).



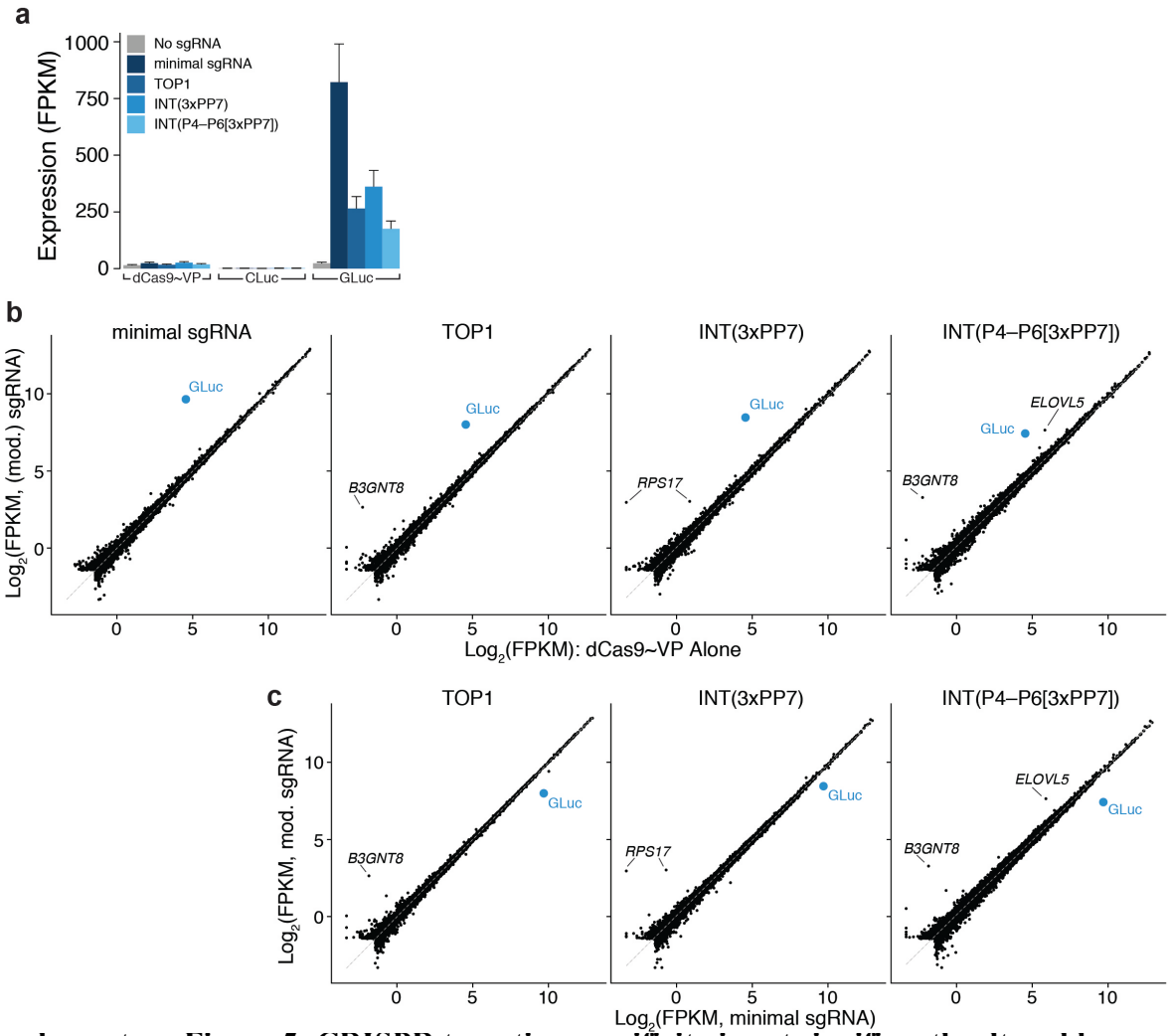
Supplementary Figure 2: A proposed ncRNA ectopic localization system based on TALE two-hybrids. (a) System schematic and controls. **(b)** FACS plots demonstrating direct activation by the TALE system. Top: baseline reporter expresion in the presence of the unmodified TALE domain. Bottom: direct activtion by the TALE-VP. **(c)** Luciferase assays using transient (left) and integrated (right) reporters (n=3). **(d)** Components of the TALE system form the expected binary and ternary complexes. Each control RNA was coexpressed with TALE-MS2 and PP7-VP and immunoprecipitated with anti-FLAg (TALE) or anti-HA (PP7-VP) antibodies. Bottom panel shows the western blots. Inputs correspond to the 2.5% of the starting sample; immunoprecipitates correspond to 12.5% of the recovered material. **(e)** Thermodynamic scheme summarizing the binding events within th TALE two-hybrid system, including the formation of the binary complexes.



Supplementary Figure 3: A split TALE approach to couple DNA-binding to RNA-binding in the TALE two-hybrid system. (a) System schematic. **(b)** Cartoons summarizing Splint RNAs tested. Each construct was under U6 promoter; the splinting cassette preceded a 30-nt unstructured RNA common to all constructs. **(c)** Luciferase assays, demonstrating failure of the split TALE approach in its current design (n=3).

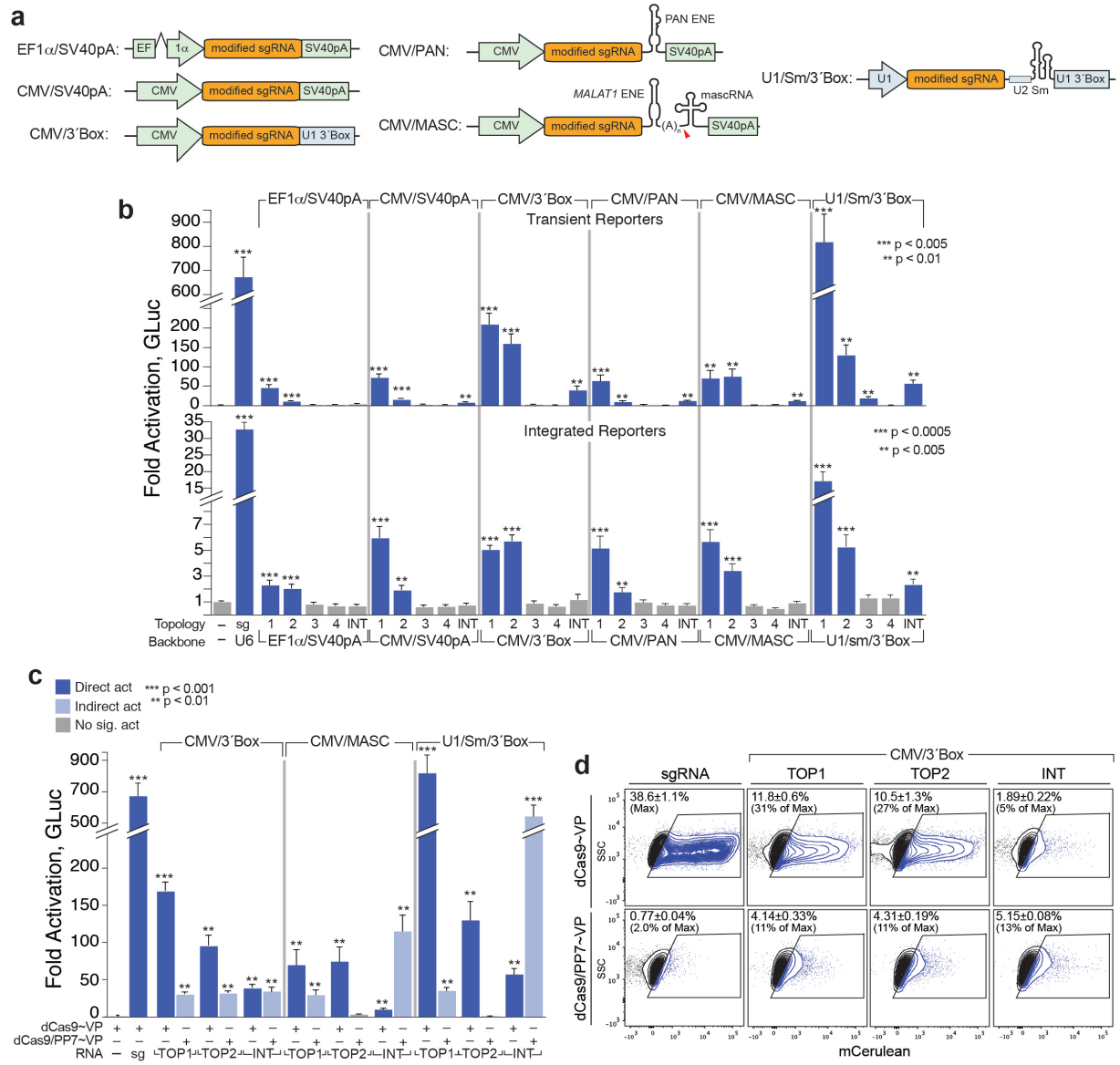


Supplementary Figure 4: Secondary structures of TOP1-4 and double TOP0-2 accesory domains.



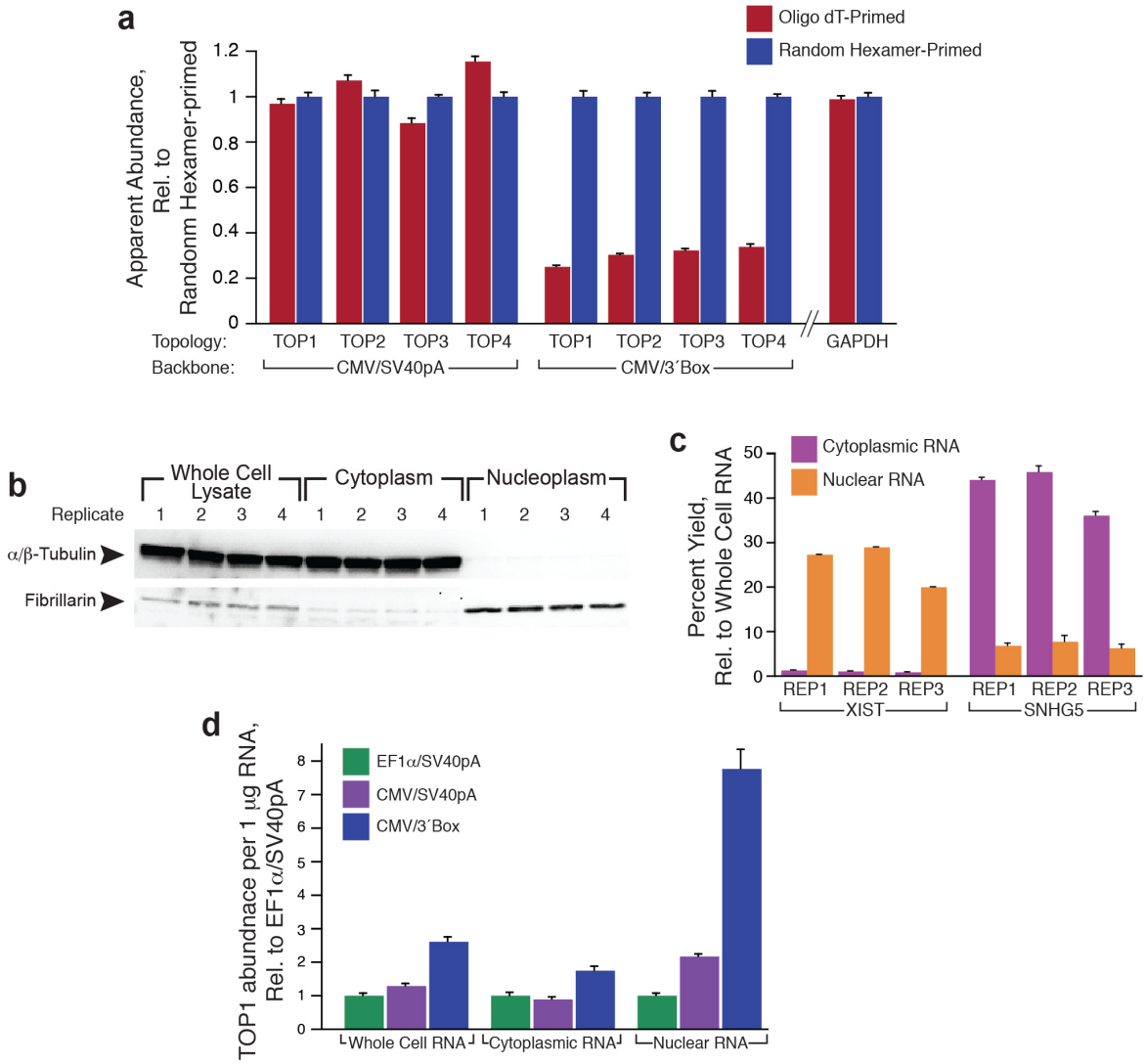
Supplementary Figure 5: CRISPR targeting specificity is not significantly altered by appending the sgRNA core with accesory RNA domains. (a) Specific activation of the Gluc-reporter by sgRNA and modified sgRNA constructs corroborates observations from luciferase and FACS assays. Values for the CLuc normalizer and dCas9-VP are shown. **(b)** Differential gene expression plots for the minimal and each modified sgRNA construct, relative to cells expressing dCas9-VP alone. All measured by poly(A)+ mRNA-Seq from Gluc reporter HEK293FT cells transiently expressing dCas9-VP alone or dCas9-VP with Gluc targeting minimal sgRNA, TOP1, INT(3xPP7) or INT(P4-P6[3xPP7]) constructs. (c) Differential gene expression plots for each modified sgRNA, relative to the minimal sgRNA. Off-targets genes are

indicated on the plots; no PAM-adjacent off-target binding sites complementary to the GLuc gRNA were observed within these loci, suggesting sporadic activation.

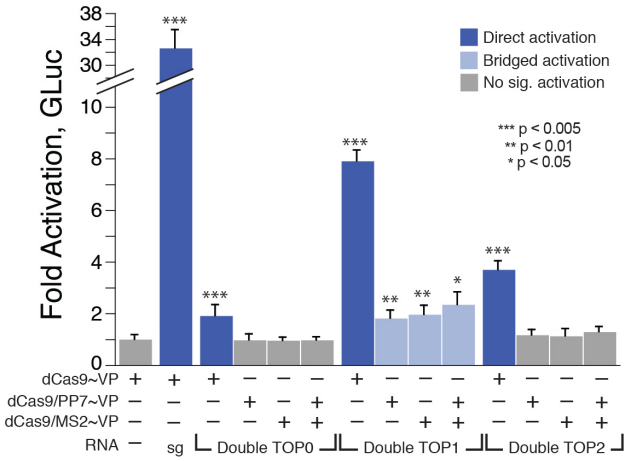


Supplementary Figure 6: Surveying pol II expression systems for CRISPR-Display. (a) Schematics of RNA pol II expression system for modified sgRNA constructs. **(b)** Direct activation by Pol II-driven TOP constructrs, measured by luciferase assays. **(c)** Direct and bridged expression by the most effective constructs. Transient reporter assays are shown**.** "sg"

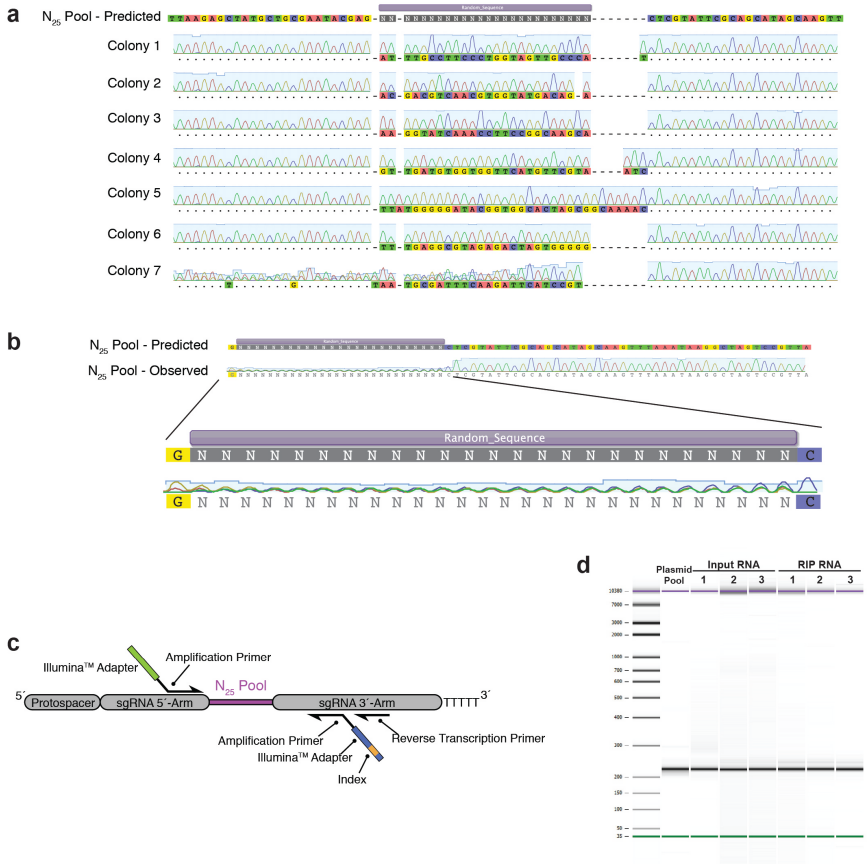
minimal sgRNA driven from U6. **(d)** FACS analysis on the transient reporter assays with CMV/3'Box constructs as in Fig.2d (n=3).



Supplementary Figure 7: The CMV/3'Box system generates non-polyadenylated, nuclear-localized transcripts. (a) cDNA synthesized by random hexamers or oligo-dT primers. GAPDH control is shown. **(b-c)** Subcellular fractionation shown by Western blot analysis and qRT-PCR controls for XIS T and SNHG5. **(d)** Cells expressing TOP1 from each Pol II backbone were fractionated and analyzed by fractionation (n=3).

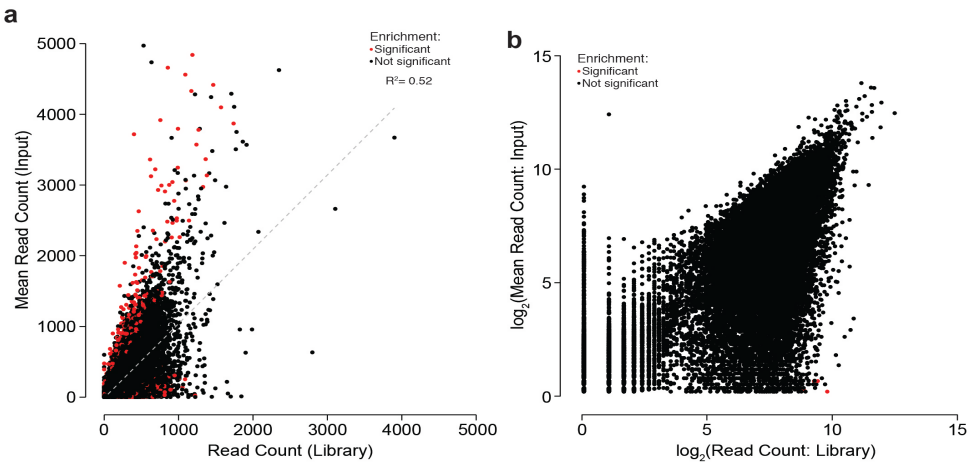


Supplementary Figure 8: Integrated reporter luciferase assays with "Double TOP" constructs. Double TOP constructs were expressed using CMV/3'Box system. "sg" minimal sgRNA driven from U6. Relative to cells expressing dCas9-VP alone (n=3, Student's t-test).

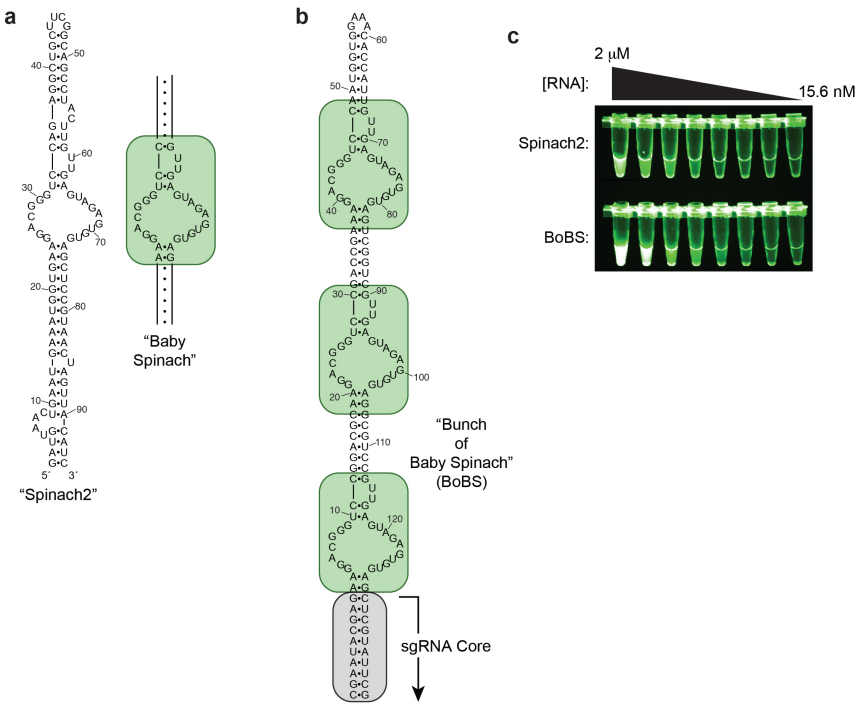


Supplementary Figure 9: Synthesizing and sequencing the INT-N25 pool. (a) Sequences and chromatograms of 7 clones. **(b)** Seq chromatogram of the aggregate INT pool. **(c)** Schematic for

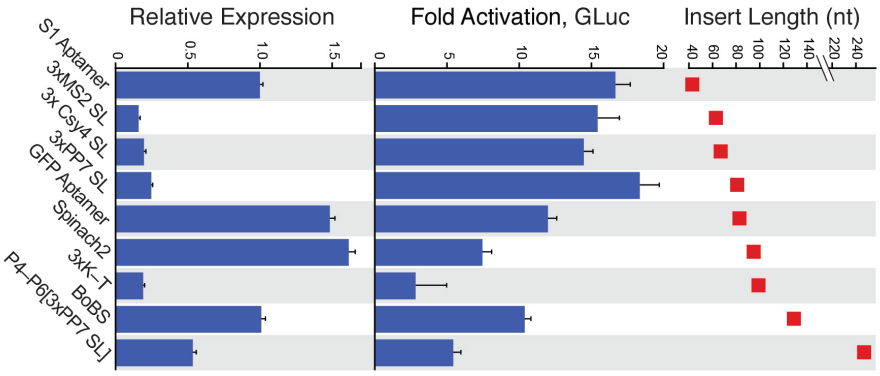
the design of primers used to generate targeted deep sequencing libraries. **(d)** Bioanalyzer traces of the final sequencing libraries.



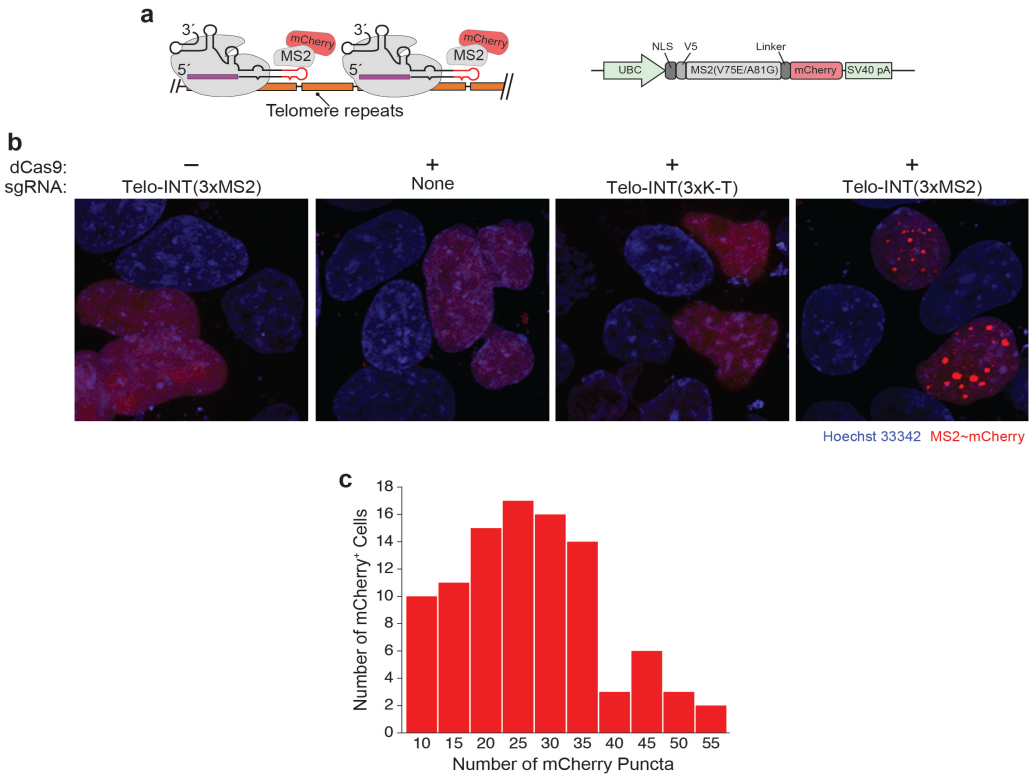
Supplementary Figure 10: Sequence diversity and expression of the INT-N25 pool. (a) Read counts of 25mers in the plasmid pool vs. mean read counts in input. 783,612 unique sequences: 524 (0.07%) significantly enriched and 7,011 (0.9%) depleted in input. Sequences containing more than 5 consecutive U's act as Pol III termination signal and are among the depleted sequences**. (b)** Same data as in (a) on log scale.



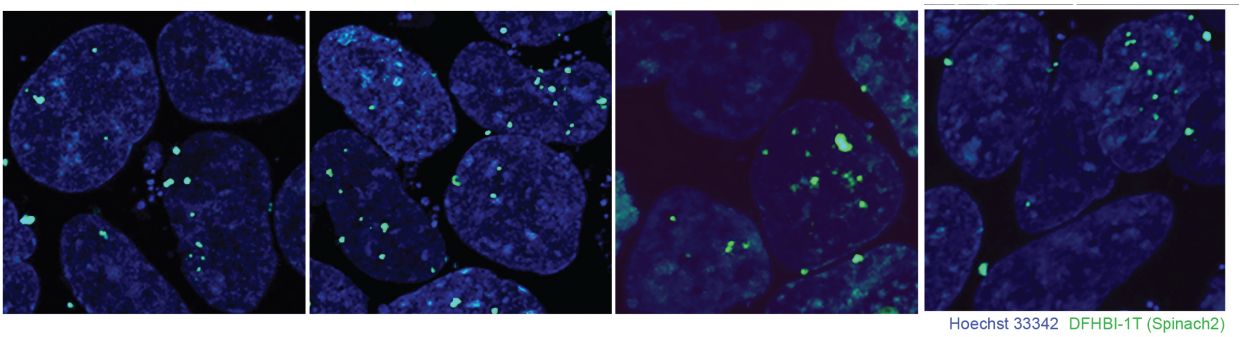
Supplementary Figure 11: Design of "Bunch of Baby Spinach" (BoBS) construct. (a) Secondary structures of the 94 nt Spinach2 and minimal Spinach aptamer. A 29 nt double G-quadruplex core binds DFHBI-1T, the fluorophore ligand is boxed in green. **(b)** Design of the 129 nt BoBS construct**. (c)** In vitro transcribed INT-Spinach2 and INT-BoBS, incubated with 20 uM DFHBI-1T (Lucerna Technologies) and imaged using UV trans-illuminator.



Supplementary Figure 12: Efficacy of INT-like CRISP-Disp constructs partially varies with length and expression level. Quantified by RT-qPCR using universal primer pINT that targets sgRNA core outside of the INT insertion site. Normalized to GAPDH then to INT-S1 Aptamer.



Supplementary Figure 13: Bridged imaging of genomic loci with CRISPR-Display. (a) Experimental design. **(b)** All cells express MS2-mCherry, in addition to the indicated constructs. 3xK-T; sgRNA appended with three kink-turns. (63X magnification). **(c)** Histogram of observed fluorescent puncta in 97 mCherry+ cells.



Supplementary Figure 14: Additional aptamer-based live cell images. dCas9 cells, transfected with telomere targeting sgRNA internally appended with Spinach2 (63x mag).

References

1. Cech, T.R. & Steitz, J.A. The noncoding RNA revolution-trashing old rules to forge new ones. *Cell* **157**, 77-94 (2014).

2. Rinn, J.L. & Chang, H.Y. Genome regulation by long noncoding RNAs. *Annual review of biochemistry* **81**, 145-166 (2012).

3. Ulitsky, I. & Bartel, D.P. lincRNAs: genomics, evolution, and mechanisms. *Cell* **154**, 26-46 (2013).

4. Chow, J.C. et al. Inducible XIST-dependent X-chromosome inactivation in human somatic cells is reversible. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 10104-10109 (2007).

5. Sauvageau, M. et al. Multiple knockout mouse models reveal lincRNAs are required for life and brain development. *eLife* **2**, e01749 (2013).

6. Bassett, A.R. et al. Considerations when investigating lncRNA function in vivo. *eLife* **3**, e03058 (2014).

7. Liang, J.C., Bloom, R.J. & Smolke, C.D. Engineering biological systems with synthetic RNA molecules. *Molecular cell* **43**, 915-926 (2011).

8. Chappell, J. et al. The centrality of RNA for engineering gene expression. *Biotechnology journal* **8**, 1379-1395 (2013).

9. Carothers, J.M., Goler, J.A., Juminaga, D. & Keasling, J.D. Model-driven engineering of RNA devices to quantitatively program gene expression. *Science* **334**, 1716-1719 (2011).

10. Delebecque, C.J., Lindner, A.B., Silver, P.A. & Aldaye, F.A. Organization of intracellular reactions with rationally designed RNA assemblies. *Science* **333**, 470-474 (2011).

11. Song, W., Strack, R.L., Svensen, N. & Jaffrey, S.R. Plug-and-play fluorophores extend the spectral properties of Spinach. *Journal of the American Chemical Society* **136**, 1198-1201 (2014).

12. Auslander, S. et al. A general design strategy for protein-responsive riboswitches in mammalian cells. *Nat Meth* **11**, 1154-1160 (2014).

13. Chen, X., Li, N. & Ellington, A.D. Ribozyme catalysis of metabolism in the RNA world. *Chemistry & biodiversity* **4**, 633-655 (2007).

14. Walker, S.C., Good, P.D., Gipson, T.A. & Engelke, D.R. The dual use of RNA aptamer sequences for affinity purification and localization studies of RNAs and RNA-protein complexes. *Methods in molecular biology* **714**, 423-444 (2011).

15. Tome, J.M. et al. Comprehensive analysis of RNA-protein interactions by high-throughput sequencing-RNA affinity profiling. *Nature methods* **11**, 683-688 (2014).

16. Sternberg, S.H., Redding, S., Jinek, M., Greene, E.C. & Doudna, J.A. DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. *Nature* **507**, 62-67 (2014).

17. Garneau, J.E. et al. The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature* **468**, 67-71 (2010).

18. Jinek, M. et al. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816-821 (2012).

19. Gilbert, L.A. et al. CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. *Cell* **154**, 442-451 (2013).

20. Mali, P. et al. CAS9 transcriptional activators for target specificity screening and paired nickases for cooperative genome engineering. *Nature biotechnology* **31**, 833-838 (2013).

21. Maeder, M.L. et al. CRISPR RNA-guided activation of endogenous human genes. *Nature methods* **10**, 977-979 (2013).

22. Perez-Pinera, P. et al. RNA-guided gene activation by CRISPR-Cas9-based transcription factors. *Nature methods* **10**, 973-976 (2013).

23. Mali, P., Esvelt, K.M. & Church, G.M. Cas9 as a versatile tool for engineering biology. *Nature methods* **10**, 957-963 (2013).

24. Hsu, P.D., Lander, E.S. & Zhang, F. Development and Applications of CRISPR-Cas9 for Genome Engineering. *Cell* **157**, 1262-1278 (2014).

25. Chen, B. et al. Dynamic imaging of genomic loci in living human cells by an optimized CRISPR/Cas system. *Cell* **155**, 1479-1491 (2013).

26. Shalem, O. et al. Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science* **343**, 84-87 (2014).

27. Wang, T., Wei, J.J., Sabatini, D.M. & Lander, E.S. Genetic screens in human cells using the CRISPR-Cas9 system. *Science* **343**, 80-84 (2014).

28. Nissim, L., Perli, S.D., Fridkin, A., Perez-Pinera, P. & Lu, T.K. Multiplexed and Programmable Regulation of Gene Networks with an Integrated RNA and CRISPR/Cas Toolkit in Human Cells. *Molecular cell* **54**, 698-710 (2014).

29. Ryan, O.W. et al. Selection of chromosomal DNA libraries using a multiplex CRISPR system. *eLife* **3** (2014).

30. Gilbert, L.A. et al. Genome-Scale CRISPR-Mediated Control of Gene Repression and Activation. *Cell* (2014).

31. Citorik, R.J., Mimee, M. & Lu, T.K. Sequence-specific antimicrobials using efficiently delivered RNA-guided nucleases. *Nature biotechnology* (2014).

32. Briner, Alexandra E. et al. Guide RNA Functional Modules Direct Cas9 Activity and Orthogonality. *Molecular cell* **56**, 333-339 (2014).

33. Wright, A.V. et al. Rational design of a split-Cas9 enzyme complex. *Proceedings of the National Academy of Sciences of the United States of America* **112**, 2984-2989 (2015).

34. Zalatan, J.G. et al. Engineering complex synthetic transcriptional programs with CRISPR RNA scaffolds. *Cell* **160**, 339-350 (2015).

35. Konermann, S. et al. Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex. *Nature* **517**, 583-588 (2015).

36. Sternberg, S.H., Haurwitz, R.E. & Doudna, J.A. Mechanism of substrate selection by a highly specific CRISPR endoribonuclease. *Rna* **18**, 661-672 (2012).

37. Zhang, F. et al. Efficient construction of sequence-specific TAL effectors for modulating mammalian transcription. *Nature biotechnology* **29**, 149-153 (2011).

38. Douglass, E.F., Jr., Miller, C.J., Sparer, G., Shapiro, H. & Spiegel, D.A. A comprehensive mathematical model for three-body binding equilibria. *Journal of the American Chemical Society* **135**, 6092-6099 (2013).

39. Chao, J.A., Patskovsky, Y., Almo, S.C. & Singer, R.H. Structural basis for the coevolution of a viral RNA-protein complex. *Nature structural & molecular biology* **15**, 103-105 (2008).

40. Jinek, M. et al. Structures of Cas9 endonucleases reveal RNA-mediated conformational activation. *Science* **343**, 1247997 (2014).

41. Chavez, A. et al. Highly efficient Cas9-mediated transcriptional programming. *Nature methods* (2015).

42. Ran, F.A. et al. Double nicking by RNA-guided CRISPR Cas9 for enhanced genome editing specificity. *Cell* **154**, 1380-1389 (2013).

43. Kuscu, C., Arslan, S., Singh, R., Thorpe, J. & Adli, M. Genome-wide analysis reveals characteristics of off-target sites bound by the Cas9 endonuclease. *Nature biotechnology* (2014).

44. Tsai, S.Q. et al. GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. *Nature biotechnology* **33**, 187-197 (2015).

45. Wu, X. et al. Genome-wide binding of the CRISPR endonuclease Cas9 in mammalian cells. *Nature biotechnology* (2014).

46. Gao, Y. & Zhao, Y. Self-processing of ribozyme-flanked RNAs into guide RNAs in vitro and in vivo for CRISPR-mediated genome editing. *Journal of integrative plant biology* **56**, 343-349 (2014).

47. Tsai, S.Q. et al. Dimeric CRISPR RNA-guided FokI nucleases for highly specific genome editing. *Nature biotechnology* **32**, 569-576 (2014).

48. Fuke, H. & Ohno, M. Role of poly (A) tail as an identity element for mRNA nuclear export. *Nucleic acids research* **36**, 1037-1049 (2008).

49. Conrad, N.K., Mili, S., Marshall, E.L., Shu, M.D. & Steitz, J.A. Identification of a rapid mammalian deadenylation-dependent decay pathway and its inhibition by a viral RNA element. *Molecular cell* **24**, 943-953 (2006).

50. Wilusz, J.E. et al. A triple helix stabilizes the 3' ends of long noncoding RNAs that lack poly(A) tails. *Genes & development* **26**, 2392-2407 (2012).

51. Battle, D.J. et al. The SMN complex: an assembly machine for RNPs. *Cold Spring Harbor symposia on quantitative biology* **71**, 313-320 (2006).

52. Cuello, P., Boyd, D.C., Dye, M.J., Proudfoot, N.J. & Murphy, S. Transcription of the human U2 snRNA genes continues beyond the 3' box in vivo. *The EMBO journal* **18**, 2867-2877 (1999).

53. Mayer, C., Neubert, M. & Grummt, I. The structure of NoRC-associated RNA is crucial for targeting the chromatin remodelling complex NoRC to the nucleolus. *EMBO reports* **9**, 774-780 (2008).

54. Orom, U.A. et al. Long noncoding RNAs with enhancer-like function in human cells. *Cell* **143**, 46-58 (2010).

55. Minks, J., Baldry, S.E., Yang, C., Cotton, A.M. & Brown, C.J. XIST-induced silencing of flanking genes is achieved by additive action of repeat a monomers in human somatic cells. *Epigenetics & chromatin* **6**, 23 (2013).

56. Wang, K.C. et al. A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. *Nature* **472**, 120-124 (2011).

57. Saito, H. et al. Synthetic translational regulation by an L7Ae-kink-turn RNP switch. *Nature chemical biology* **6**, 71-78 (2010).

58. Esvelt, K.M. et al. Orthogonal Cas9 proteins for RNA-guided gene regulation and editing. *Nature methods* **10**, 1116-1121 (2013).

59. Cong, L., Zhou, R., Kuo, Y.C., Cunniff, M. & Zhang, F. Comprehensive interrogation of natural TALE DNA-binding modules and transcriptional repressor domains. *Nature communications* **3**, 968 (2012).

60. LeCuyer, K.A., Behlen, L.S. & Uhlenbeck, O.C. Mutants of the bacteriophage MS2 coat protein that alter its cooperative binding to RNA. *Biochemistry* **34**, 10600-10606 (1995).

61. Kelley, D.R., Hendrickson, D.G., Tenen, D. & Rinn, J.L. Transposable elements modulate human RNA abundance and splicing via specific RNA-protein interactions. *Genome Biol* **15**, 537 (2014).

62. Zhao, S. & Fernald, R.D. Comprehensive algorithm for quantitative real-time polymerase chain reaction. *Journal of computational biology : a journal of computational molecular cell biology* **12**, 1047-1064 (2005).

63. Kim, D. et al. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* **14**, R36 (2013).

64. Trapnell, C. et al. Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nature biotechnology* **31**, 46-53 (2013).

65. Shechner, D.M. & Bartel, D.P. The structural basis of RNA-catalyzed RNA polymerization. *Nature structural & molecular biology* **18**, 1036-1042 (2011).

66. Rosner, M. & Hengstschlager, M. Detection of cytoplasmic and nuclear functions of mTOR by fractionation. *Methods in molecular biology* **821**, 105-124 (2012).

67. Bhatt, D.M. et al. Transcript dynamics of proinflammatory genes revealed by sequence analysis of subcellular RNA fractions. *Cell* **150**, 279-290 (2012).

68. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol* **11**, R106 (2010).