UNIVERSITY OF CALIFORNIA, SAN DIEGO

Post-transcriptional regulation by the pluripotency associated RNA-binding protein LIN28

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

in

Biomedical Sciences

by

Melissa L. Wilbert

Committee in charge:

Professor Gene W. Yeo, Chair
Professor Neil C. Chi
Professor Christopher K. Glass
Professor Lawrence S. B. Goldstein
Professor Amy E. Pasquinelli

2014

The Dissertation of Melissa L. Wilbert is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____

Chair

University of California, San Diego

2014

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to acknowledge all past and present members of the Yeo lab for their help and support both scientifically and personally that has enabled me to achieve this goal. It has been this wonderful group that has raised me up from the beginning of graduate school into the scientist I am today. In particular I want to thank our founding group of students and researchers who crammed into our original lab space. Stephanie C. Huelga has been invaluable for her expertise during manuscript and presentation preparation, always improving the esthetics of our work, staying cool under pressure, and enriching the lab and my life. I would like to thank Michael T. Lovci for providing computational support and mentoring that forced him to function more like a postdoc and friend rather than fellow graduate student. I am grateful that I could always turn to Thomas J. Stark as a model for scientific diligence and dedication, and friendship no matter what the hour in lab. Thanks also goes to David A. Nelles for the energy and enthusiasm he brought to our group that always brightened the mood.

I would like to thank Katlin B. Massirer for being my first female mentor in science and really caring for the wellbeing of our lab family. Katannya Kapeli became my second female postdoc mentor, offering me tremendous support during completion of our publication, even as she was going through her own transitions into the lab. Thanks to Kasey R. Hutt for his postdoc mentorship; always giving sound advice and detailing considerations for analysis, even when that meant more factors that I wanted to regard. I also want to thank Jason L. Nathanson for being a

constant source of a friendly ear and advice. I want to acknowledge Tiffany Y. Liang and Anthony Q. Vu for their indefatigable skill, dedication, and attitude while turning out quality data that has been priceless for our successful studies.

I acknowledge all of the undergraduates that have helped to keep the Yeo lab running. In particular, my mentee Stella X. Chen who struggled on the front lines with me through the development of the LIN28 project, exposing herself to the often stressful and unpredictable scientific process, and coming through with poise and fortitude. Thanks also to Bernice Y. Yan for lending a skillful, cheerful hand to my projects and throughout the lab. I thank Chau Ly for allowing me to mentor her as I gained as much benefit for myself as I offered her, especially through her support during the final push to finish this dissertation.

I appreciate and am inspired by the faculty, students and staff of the Biomedical Sciences program and their incredible efforts to improve science education. I want to acknowledge Shannon Muir for her perseverance throughout graduate school that has touched and motivated me as woman, scientist, and friend. I thank Anne E. Conway, for her understanding, adaptability, diligence and kindness she has demonstrated as a collaborator and friend. She was instrumental in the completion of this dissertation through her encouragement and assitance with editing.

I am forever grateful to all the mentors and teachers that have touched my life. Credit is due to our leader and mentor, Gene W. Yeo who has tirelessly built our group and fueled the exciting research that we do. I have gained so much from

his example of how to stay at the forefront of research, always with an eye on the horizon. I would also like to acknowledge the members of my committee for their mentorship and contributions towards my training. I will be forever grateful for the opportunities I have been afforded at UCSD, in particular through training with the California Institute of Regenerative Medicine and UCSD Genetics Training Program. I would like to thank Bruce Hamilton for his advocacy of good science and good teaching, and all his efforts that make the Genetics Training Program a success. I acknowledge Lawrence Goldstein for his inspiring advocacy for stem cell research. My previous mentor and friend from the Rochester Institute of Technology, Gary Skuse, continued to provide me with invaluable support and mentorship during my doctoral training, for which I express my sincerest appreciation.

I would like to acknowledge all of my friends and family for their love and support, for always encouraging me and helping me pick myself up; I could not have made it without you. I especially thank my mother for always encouraging my academic pursuits and my sister for her unwavering companionship. I am humbled, honored, and sincerely touched by all the people who have taken time and energy from their lives to help me succeed. Thank you.

I would like to acknowledge my co-authors on manuscripts that have contributed to the content of this dissertation. Chapter 1, in part, is an adaptation of material that appears in “Genome-wide approaches in the study of microRNA biology” by Melissa L. Wilbert and Gene W. Yeo, as published in *Wiley*

Interdisciplinary Reviews: Systems Biology and Medicine December 2010. Chapter 3, in part Figure 19, is also an adaptation from this publication. Chapter 2, in full, is an adaptation of material that appears in "LIN28 binds messenger RNAs at GGAGA motifs and regulates splicing factor abundance to affect alternative splicing" by Melissa L. Wilbert, Stephanie C. Huelga, Katannya Kapeli, Thomas J. Stark, Tiffany Y. Liang, Stella X. Chen, Bernice Y. Yan, Jason L. Nathanson, Kasey R. Hutt, Michael T. Lovci, Hilal Kazan, Anthony Q. Vu, Katlin B. Massirer, Quaid Morris, Shawn Hoon and Gene W. Yeo, as published in *Molecular Cell* October 2012. Chapter 3, in part, is an adaptation of material that appears in "The RNA binding protein IG2BP1/IMP1 promotes survival and adhesion in human pluripotent stem cells by stabilizing target RNAs" by Anne E. Conway, Melissa L. Wilbert and Gene W. Yeo and "LIN28 has slight effects on mRNA translation" by Melissa L. Wilbert, James J. Moresco, and Gene W. Yeo, both in preparation for publication.

VITA

EDUCATION

PhD in Biomedical Sciences **2008 –2014**
University of California San Diego
GPA: 3.96 (out of 4.0)
Doctoral thesis committee: Dr. Gene W. Yeo (chair), Dr. Neil C. Chi, Dr. Christopher K. Glass, Dr. Lawrence S. B. Goldstein, and Dr. Amy E. Pasquinelli

B.S./M.S. in Bioinformatics **2003 –2008**
Rochester Institute of Technology
GPA: B.S. 3.8/ M.S. 4.0 (out of 4.0)
GRE: Quantitative 800 (94^{th} percentile) / Verbal 610 (87^{th} percentile)

RESEARCH EXPERIENCE

University of California San Diego - Doctoral Dissertation **2008 – 2014**
PI: Dr. Gene W. Yeo

Lead transcriptome-wide studies of RNA-binding proteins (RBPs) in the context of human pluripotent stem cells and neurogenesis. Focused on the key developmental regulators and concofetal proteins LIN28 and IMP family RBPs, and microRNA (miRNA) regulation in human embryonic stem cells (hESCs). Described in detail LIN28 interactions with, and regulation of, thousands of mRNA transcripts in a similar sequence and structural context as LIN28 binding to let-7 miRNAs. This published work was recommended by the 'Faculty of 1000' as a top article in Biology and Medicine. Reannotation of the *C. elegans* transcriptome using deep RNA-sequencing resulted in thousands of reassigned untranslated regions and contribution as second-author to the 'Article of the Month' in *Nature Structural and Molecular Biology* for February 2010. Experienced in reprogramming adult human fibroblasts into induced pluripotent stem cells (iPSCs), and culture, characterization and manipulation of iPSCs and hESCs. Pioneered the development of custom biochemical protocols and computational algorithms for the generation and analyses of high-throughput sequencing and proteomics data. Designed and executed complex, high-throughput projects while anticipating and responding to sources of biases and limitations introduced during experimental and analytical stages. Collaborated with colleagues in multiple disciplines to effectively apply new techniques to published and ongoing work in the Yeo laboratory. Contributed to successful fellowship and grant applications and renewals, and participated in the peer-review process of submitted manuscripts.

University of California, San Diego - Graduate Rotation **Spring 2008**

PI: Dr. Lawrence S. B. Goldstein

Successfully generated induced iPSC lines from adult human dermal fibroblasts using retroviral induction of the transcription factors Oct4, Sox2, c-Myc, and Klf-4. My interest in genetic variation lead me to establish a sequencing-based project that identified the presence of genomic instabilities in iPSCs in collaboration with Dr. Kun Zhang and colleagues, published in *Nature*, March 2011.

University of California, San Diego - Graduate Rotation **Winter 2008**

PI: Dr. Andrew J. McCammon

Utilized AutoDock4 and the NCI Discovery Set in prediction of small molecule interactions that could potentially disrupt *Trypanosoma brucei,* which is cause of African sleeping sickness. Specifically, interactions between *T. brucei* RNA editing ligase and bound RNA were modeled.

Rochester Institute of Technology - Master's Thesis **2006-2008**

Co-Advisors: Dr. Gary R. Skuse & Dr. Michael V. Osier

Committee Member: Dr. David A. Lawlor

Maximum likelihood analyses and the software package HyPhy were used to identify differential selective pressures acting on the neuraminidase (NA) and hemagglutinin (HA) proteins of Influenza A H3N2 isolated from humans in geographically distinct datasets. Developed a novel approach to gain insight into functional adaptation of the virus to various climates, its seasonality, and global spread.

Diffinity Genomics - Research Associate **2006-2007**

PI: Dr. Lewis Rothberg, University of Rochester

Designed and executed an independent research project in development of a colorimetric assay for detection of genomic DNA using gold colloidal nanoparticles.

Roswell Park Cancer Institute - Research Apprentice **2006-2007**

PI: Dr. Norma Nowak, Genomics & Microarray Core

Aided in platform design for array-based comparative genomic hybridization (aCGH). Consensus regions involved in chromosomal insertion/ deletion syndromes were identified with OMIM, and BAC clone coverage assigned to these regions using the UCSC Genome Browser. Optimized microarray data management through assessment of GeneDirector/ ARM microarray management software.

GRANTS AND FELLOWSHIPS

2012-2013	California Institute for Regenerative Medicine fellowship
2009-2012	UCSD NIH genetics training grant
2012	ISSCR conference travel grant
2011 & 2012	UCSD Stem Cell Program ISSCR conference travel grant
2009-2010	Genentech graduate fellowship

PUBLICATIONS - *First author papers underlined, *denotes co-authors*

Wilbert ML, Huelga SC, Kapeli K, Stark TJ, Liang TY, Chen SX, Yan BY, Nathanson JL, Hutt KR, Lovci MT, Kazan H, Vu AQ, Massirer KB, Morris Q, Hoon S & Yeo GW. LIN28 binds messenger RNAs at GGAGA motifs and regulates splicing factor abundance to affect alternative splicing. *Mol Cell*. 2012 Oct.

Gore AJ, Li Z, Fung H, Young J, Agarwal S, Antosiewicz-Bourget J, Canto I, Giorgetti A, Israel MA, Kiskinis E, Lee J, Loh Y, Manos PD, Montserrat N, Panopoulos AD, Ruiz S, **Wilbert ML**, Yu J, Kirkness EF, Belmonte JC, Rossi DJ, Thomson JA, Eggan K, Daley GQ, Goldstein LS, & Zhang K. Somatic coding mutations in human induced pluripotent stem cells. *Nature.* 2011 March.

Wilbert ML, & Yeo GW. Genome-wide approaches in the study of microRNA biology. *Wiley Interdiscip Rev Syst Biol and Med*. 2010 Dec.

Zisoulis DG*, Lovci MT*, **Wilbert ML**, Hutt KR, Liang TY, Pasquinelli AE, & Yeo GW. Comprehensive discovery of endogenous Argonaute binding sites in Caenorhabditis elegans. *Nat Struct Mol Biol*. 2010 Feb. Epub 2010 Jan 10.

SCIENTIFIC PRESENTATIONS

June 2013	Sanford Consortium for Regenerative Medicine Seminar Series La Jolla, CA – Invited talk Transcriptome-wide maps of stem cell enriched RNA-binding proteins
May 2013	Department of Biomedical Genetics Seminar Series University of Rochester Medical Center –Rochester, NY – Invited talk Integrated regulation of human pluripotent stem cells by RNA-binding proteins
April 2013	The Thomas H. Gosnell School of Life Sciences Seminar Series Rochester Institute of Technology, Rochester NY – Invited talk The role of RNA-binding proteins in the post-transcriptional control of human embryonic stem cells

SCIENTIFIC PRESENTATIONS - CONTINUED

February 2013 Biomedical Sciences Incoming Student Recruitment
La Jolla, CA – Invited talks
The role of the RNA-binding protein LIN28 in human embryonic stem cells

November 2012 California Institute for Regenerative Medicine Fellows Retreat
San Francisco, CA – Poster
LIN28 binds messenger RNAs at GGAGA motifs and affects splicing factor abundance

October 2012 Stem Cell Meeting on the Mesa, Salk Institute for Biological Studies
La Jolla, CA - Poster
LIN28 binds messenger RNAs at GGAGA motifs and affects splicing factor abundance

September 2012 San Diego Center for Systems Biology NextGen Gene Expression Analysis Workshop, Salk Institute for Biological Studies, La Jolla, CA – Invited talk
LIN28 binds messenger RNAs at GGAGA motifs and affects splicing factor abundance: Lessons from RNA sequencing approaches

September 2012 Biomedical Sciences Program Retreat
La Quinta, CA – Invited talk
LIN28 binds messenger RNAs at GGAGA motifs and affects splicing factor abundance

August 2012 UCSD Neurodegenerative Interest Group
San Diego, CA – Talk
LIN28 binds messenger RNAs at GGAGA motifs and affects splicing factor abundance

June 2012 International Society for Stem Cell Research (ISSCR)
Yokohama, Japan – Selected talk
LIN28 interacts with discrete binding motifs in messenger RNAs and regulates alternative splicing through modulation of splicing factors

May 2012 The RNA Society Annual Meeting
Ann Arbor, Michigan – Poster
LIN28 interacts with GGAGA motifs in messenger RNAs

SCIENTIFIC PRESENTATIONS - CONTINUED

April 2012 — American Society for Biochemistry and Molecular Biology
San Diego, CA – Selected talk
Nucleotide-level resolution of LIN28 binding sites in mature messenger RNA sequences reveals regulation of a network of splicing factors and downstream alternative splicing patterns

March 2012 — UCSD Cellular and Molecular Medicine East Symposium
Sanford Consortium for Regenerative Medicine,
La Jolla, CA – Talk
LIN28 binding and regulation of mRNA transcripts

December 2011 — Salk Institute Stem Cell Interest Group Monthly Meeting
La Jolla, CA – (Talk)
LIN28 binds GGAGA motifs in messenger RNAs and regulates alternative splicing through splicing factors

June 2011 — International Society for Stem Cell Research (ISSCR)
Toronto, Canada – Poster
Nucleotide-level resolution of LIN28 RNA binding sites reveals a network of post-transcriptional control in human embryonic stem cells

TEACHING AND MENTOR EXPERIECE

2012 — Teaching Assistant for "Genes, Embryos, and Development" at UCSD

2011-2012 — Mentor of a UCSD biochemistry undergraduate researcher

ABSTRACT OF THE DISSERTATION

Post-transcriptional regulation by the pluripotency associated
RNA-binding protein LIN28

by

Melissa L. Wilbert

Doctor of Philosophy in Biomedical Sciences

University of California, San Diego, 2014

Professor Gene W. Yeo, Chair

The field of stem cell biology is moving forward at an unprecedented rate in part due to the discovery that adult somatic cells can be reprogrammed to a pluripotent stem cell like state. The factors first used in reprogramming were transcription factors such as OCT4, SOX2 and NANOG, and the RNA-binding protein LIN28. Like transcription factors, RNA-binding proteins (RBPs) control vast networks of gene

targets to direct pathways in the cell; however, for RBPs this is accomplished through post-transcriptional binding to RNA transcripts. Only recently has it been possible to survey the transcripome-wide RNA binding interactions of a protein, through isolation of endogenous RBP-RNA complexes paired with high-throughput sequencing technologies. Using cross-linking followed by immunoprecipitation of protein-RNA complexes and sequencing of isolated transcripts (CLIP-seq) we have identified LIN28 binding sites throughout the human transcriptome. The resolution of our data enabled us to define characteristic LIN28 mRNA interactions at GGAGA rich motifs within unpaired regions of hairpin loops. This binding pattern mimics interactions described for LIN28 binding within let-7 family microRNA precursors. The ability to consider LIN28 targets on a global scale enabled the identification of RNA processing factors, in particular splicing factors, as prevalent functions encoded by LIN28 bound RNAs. This information helped to accurately predict which of the thousands of LIN28 targets would be functionally regulated. We found evidence that LIN28 increases the protein production of splicing factors resulting in massive rearrangement of RNA transcripts through downstream splicing changes. Subsequent transcriptome-wide studies of LIN28 have confirmed these findings despite differences in the pool of direct targets defined by individual reports. Taken together, we understand that LIN28 can bind to a wide network of transcripts, influencing development through these direct RNA interactions and via downstream effects. Combinatorial approaches in the study of LIN28 using changes in RNA-levels, protein production, strength of CLIP-seq binding, and ontological classification of gene targets have extracted meaningful

information about mechanisms of LIN28 regulation. We expect that application of similar methods will enable studies of additional RBPs. For example, in the study of other stem cell enriched proteins like the IGFII-mRNA binding proteins (IG2BP or IMP). Furthermore, the overlap of other regulatory networks hold promise of highlighting novel hubs of regulation that may be exploited in reprogramming or directed differentiation. The next step is to use these connections to explain how genetic changes within an individual can affect RBP function and result in disease. We can apply *in vitro* modeling of development using directed differentiation to iteratively test how the connection of LIN28 to its target transcripts impacts its role in development and disease.

CHAPTER 1 - INTRODUCTION

RNA-BINDING PROTEINS CONTROL THE TRANSCRIPTOME

The fate of the transcriptome ultimately determines the fate of the cell, and it is RNA-binding proteins (RBPs) that critically control processing of these transcripts (Dreyfuss et al., 2002). Regulation of eukaryotic messenger RNA (mRNA) begins at the point of transcription, or co-transcriptionally. RBPs control the processing of the transcribed precursor mRNA (pre-mRNA) to affect splicing and modifications such as polyadenylation and methylation. Next, RBPs mediate the transport of the processed mRNA into the cytoplasm and to the final location of translation (Keene, 2007; Martin and Ephrussi, 2009). Lastly, protein translation from mRNA transcripts can be enhanced or hindered by RBPs (Moore and Proudfoot, 2009; Sonenberg and Hinnebusch, 2009).

The transcriptome consists of many classes of RNAs beyond protein-coding mRNAs such as small nucleolar RNAs (snoRNA), micro-RNAs (miRNAs), long non-coding RNAs (lncRNAs), short interfering RNAs (siRNAs), Piwi-interacting (pi-RNAs) and likely more yet to be classified (Brown et al., 1992; Carninci, 2009; Lee et al., 1993; Maxwell and Fournier, 1995; Siomi et al., 2011; Wightman et al., 1993). The vast majority of these will rely upon RBPs for production or function. This general description only begins to brush the surface of the complex network of regulatory functions and outcomes controlled through RNA-protein formations, referred to as ribonucleoprotein complexes (RNPs). Due to their function as critical RNA regulators a number of RBPs have been associated with neurological disorders

and cancer (Baltz et al., 2012; Castello et al., 2012; Lukong et al., 2008). Specifically the action of RBPs such as TDP-43 and FUS/TLS have been shown to play roles in the formation of cytoplasmic aggregates that are hallmarks of amyotrophic lateral sclerosis (ALS, also known as Lou Gerig's Disease), Alzheimer's, and other neurological disorders (Arai et al., 2006; Kapeli and Yeo, 2012; Lagier-Tourenne et al., 2010; Neumann et al., 2006). In addition, the re-expression of early developmental proteins has been documented in various types of cancers leading to the classification of some RBPs, including LIN28 and IMP1, as oncofetal regulators (Bell et al., 2013; Zhou et al., 2013).

RNA-binding proteins are an abundant class of proteins in terms of the number encoded in the genome, and also their level of expression across all cell types. For example, various members of the heterogeneous nuclear ribonucleoparticle protein family (hnRNP) of RBPs alone are differentially expressed and localized, and even vary in isoform preference across a wide-range of cell types (Kamma et al., 1995; Kamma et al., 1999). To add to this complexity the hnRNPs, like other RBP protein families, are known to have both compensatory and antagonistic roles (Huelga et al., 2012). This family of proteins alone begins to paint a picture of the intricate relationship between RBPs. Although a comprehensive review of all RNA and RBP biology is beyond the scope of this dissertation, the study of one RBP, LIN28 provides insights into regulation of various RNAs by a widely important, developmentally regulated protein. Its regulation of different classes of RNA transcripts, for example miRNAs and mRNAs, with differential outcome in response to target binding,

provides an array of examples that demonstrate how we address questions of mRNP biology. Furthermore, the understanding of LIN28 as the center of a complex genetic circuit made of transcriptional regulation by DNA-binding factors, cooperation with other RBPs, and downstream effectors on thousands of RNA targets highlights the integration of RNA regulatory networks. In the following sections miRNAs and LIN28 regulation are discussed in depth within in the context of foundational RBP and transcriptome studies.

LIN28A AND LIN28B FAMILY RNA-BINDING PROTEINS

The *lin-28* gene was first studied in *C. elegans* (Ambros and Horvitz, 1984), but it is conserved throughout animal species to mice (Lin28) and humans (LIN28). Higher vertebrates have two paralogs of this gene, LIN28A and LIN28B. *C. elegans* mutants lacking *lin-28* are precocious, such that their seam cells perform an L3 specific pattern of cell division at the expense of L2 patterns (Ambros and Horvitz, 1984). Furthermore, Ambros and Horvitz observed that *lin-28* mutant worms expressed L4 state-specific events up to 2 stages early (Ambros and Horvitz, 1984). Importantly, mutant animals with these early developmental defects also exhibited lasting dysfunctions at the V5 adult stage. This early work established LIN28 as a heterochronic gene, that is, one that regulates the timing of developmental events. This critical regulation by LIN28 during development is a feature conserved from worms to humans. Indeed, LIN28 is an important factor in development as *Lin28a* knockout mice fail to thrive and more than 93% die within one day of birth (Shinoda et al.,

2013; Zhu et al., 2010). Ectopic LIN28 expression on the other hand, leads to overgrowth and delayed sexual development (Zhu et al., 2010). In humans genetic variations at the *Lin28B* locus have been associated with differences in height and timing of puberty (He et al., 2009; Lettre et al., 2008; Ong et al., 2009; Ong et al., 2011; Perry et al., 2009; Sulem et al., 2009). These examples illustrate the far-reaching effect of LIN28 on organisms. Below we discuss these and other phenotypes that dependent on LIN28 function.

LIN28 expression is almost entirely confined to the undifferentiated cells of early tissue development, such as embryonic stem cells and embryonic carcinoma cells (Yang and Moss 2003; Richards et al., 2004; Wu and Belasco, 2005). This expression profile made LIN28 an early candidate for a 'stemness' marker (Richards et al., 2004). This concept was driven home when Yu and colleagues used LIN28, along with the transcription factors OCT4, SOX2, and NANOG to reprogram human somatic cells into induced pluripotent stem cells (iPSCs) (Yu et al., 2007). Following these reports there has been intense efforts to better understand the function, mechanism, and networks by which LIN28 acts. Although LIN28A and LIN28B have been suggested to have redundant roles in many aspects, more careful studies have identified differences in their expression patterns and affects on some target RNAs (Gaytan et al., 2013; Pisokonova et al., 2012; Wilbert et al., 2012). In particular, LIN28A expression seems to be more influential and prevalent in undifferentiated cells and early tissues, while LIN28B activity was found to be important in the regulation of adult tissues (Gaytan et al., 2013; Shinoda et al., 2013). The expression

of the mRNAs profiled across human tissues is shown in Figure 1. Looking at these values we see that appreciable *LIN28A* levels are only found in adult skeletal muscle as was originally reported in mouse studies (Yang and Moss, 2003).

Protein Structure

The LIN28 protein contains two RNA-binding domains: a cold-shock domain (CSD) and a set of retroviral-like CHCC zinc-finger knuckles through which it can bind to nucleic acids (Moss et al., 1997). It is the only animal protein known to contain this combination of RNA-binding domains. Cold-shock domains, like the one found within LIN28, are known to bind single-stranded RNAs (ssRNAs) with poly-pyrimidine tracks (Phadtare and Inouye 1999; Max et al. 2006, 2007; Nam et al. 2011; Mayr et al. 2012). Although CSDs are most commonly found in DNA-binding proteins the motif is also similar to the RNA-binding motif of RNP-1 and is found in a number of RNA-binding proteins (Landsman, 1992). The zinc-finger domain (ZFD) of LIN28 is a retrovirus-like set of two zinc-finger knuckles that resemble the ZFD of the HIV-1 nucleocapsid protein NCp7 (De Guzman et al., 1998). A flexible linker peptide joins this set of RNA-binding domains, presumably allowing LIN28 to accommodate binding to RNA targets that vary slightly in sequence and structure (Nam et al., 2011).

In vertebrates both LIN28 paralogs, LIN28A and LIN28B, have this protein structure. There is high homology between the LIN28A and LIN28B paralogs, with 76% amino acid identity reported for their protein-coding regions (Piskonova et al., 2011). Clustal Omega alignment of the LIN28A and LIN28B resulted in a 66.84%

amino acid sequence identity (Figure 2). Alignment of the nucleotide sequences of the LIN28A and LIN28B 3'UTR resulted in a 55.96% identity between these regions (compared to 57.73% identity for the entire mRNA sequences). Interestingly, the different LIN28A proteins are more similar across different species than they are to the LIN28B paralog in the same organism (Guo et al., 2006). The term LIN28 has been either used to refer to the LIN28A protein, or to the family of both proteins. Alternative names for LIN28A in human are Protein lin-28 homolog A, CSDD1, ZCCHC1, or Zinc finger CCHC domain containing protein 1. LIN28B in humans is named Protein lin-28 homolog B, or CSDD2. Lower species, like the nematode *C. elegans* also have a conserved *LIN28* protein (*lin-28*) that seems most closely related to the vertebrate *LIN28B* gene. These are small ~ 28kDa proteins; LIN28A is 209 AA and LIN28B is slightly bigger at 250 AA (Figure 2). There is also a truncated, small isoform of LIN28B, LIN28BS (180 AA, 23 kDa), which is missing half of the CSD and N-terminus end of the protein. The zinc finger knuckles of LIN28BS are still intact and likely still bind RNA; however, a functional comparison with the full length LIN28B has not been conducted.

Cellular localization

Cellular compartmentalization of proteins and RNAs controls their access to co-factors, thereby affecting possible outcomes of binding events. The primary localization of LIN28 expression is within the cytoplasm (Balzer and Moss, 2007; Guo et al., 2006; Polesskaya et al., 2007). Point mutations in both the CSD and CCHC-zinc fingers result in Lin28 localization entirely to the nucleus (Balzer and Moss, 2007).

This suggests that it is the RNA targets of LIN28 that in some way dictate its nuclear export and ultimate destination in the cytoplasm. For example, Lin28 will also localize to processing bodies (P-bodies) only when both of its RNA-binding domains are intact (Balzer and Moss, 2007). P-bodies are sites of translational repression and lack any ribosome components or translation factors. Small RNAs can guide their target mRNA to P-bodies. It is possible that this is a way small RNAs play a role in Lin28 regulation, that is, by determining the ultimate location of Lin28-bound-RNAs. However, some localization events of LIN28 are likely to be RNA independent as evidence by its aggregation in stress granules under stress conditions despite CCHC and CSD mutations (Balzer and Moss, 2007).

Localization of LIN28-mRNA complexes could also be directed by other RBPs bound to the same RNA target within a functional mRNP complex or through protein-protein interactions with LIN28 itself. LIN28 localization to sites of active translation within polysome fractions has been associated with direct or indirect interactions with components of the translational machinery such as translation initiation factors and elongation factors, although the cause-and-effect of these interactions has not been tested (Balzer and Moss, 2007; Polesskaya et al., 2007). There is evidence that the RBP Musashi1 (Msi1), which also interacts with LIN28 (Polesskaya et al., 2007), does have an active role in increasing LIN28 localization to the nucleus during differentiation (Kawahara et al., 2011). LIN28 can still associate with target transcripts with mutations in a single RNA-binding domain; however, functional consequence of binding is diminished (Balzer et al., 2007; Jin et al., 2011).

This demonstrates that LIN28 is promiscuous in binding contacts, yet selective in those specific interactions that trigger a functional impact on the bound RNA. This provides clues for interpretation of transcriptome-wide LIN28 binding maps discussed in Chapters 2 and 3.

More recently, mouse Lin28b expressed in human cells, and endogenously in human cancer cells (H1299) was described primarily within the nucleus (Piskonova et al., 2011). These findings may have been overstated as some Lin28B expression can still clearly be seen within the cytoplasmic faction of Lin28B expressing cells (Figure 2C, Piskonova et al., 2011). Endogenous and exogenously expressed LIN28B and LIN28A were shown to be primarily cytoplasmic in HEK293 cells, with about 30% of LIN28B found in the nuclear fraction (Hafner et al., 2013). Some nuclear localization of LIN28B is expected, in particular nucleolar localization, due to the presence of both a functional nuclear and nucleolar localization signal in its C-terminus end and flexible linker region, respectively (Piskonova et al., 2011). The functionality of a sequence in LIN28A closely related to the LIN28B nucleolar localization signal has not been tested. However, there is evidence that Lin28a is actively transported to the nucleus (Balzer and Moss, 2007). In other cell types, such as differentiated myotubes 10-15% of LIN28A was found localized to the nucleus (Polesskaya et al., 2007), or transported to the nucleus in a cell-cycle dependant manner (Guo et al., 2006). Further studies will be required to fully describe how selective mechanisms control LIN28 localization and how this relates to the regulation of individual target RNAs.

LET-7 MICRORNAS ARE CRITICALLY CONTROLLED BY LIN28

A major component of LIN28 regulation is its role in a double-negative feedback loop with the let-7 microRNA (miRNA). More specifically, LIN28 negatively regulates let-7 biogenesis and let-7 negatively regulates LIN28 expression. This aspect of LIN28 biology is tied to the reciprocal expression of these regulators during differentiation and turmorigenesis (Rybak et al., 2008; Viswanathan et al., 2008; Viswanathan et al., 2009). Below we discuss the foundations and our understanding of this of this relationship in more detail.

MicroRNAs are a large class of post-transcriptional regulators

Of the recently identified small non-coding RNAs, microRNAs (miRNAs) have been intensely studied; in large due to the overreaching impact they have on gene regulation. MicroRNAs (miRNAs) are a class of ~21-23 nucleotide long non-coding RNAs that have critical roles in diverse biological processes that encompass development, proliferation, apoptosis, stress response, and fat metabolism. The field of miRNA biology emerged with the discovery that the gene *lin-4*, which controls developmental timing in the nematode *Caenorhabditis elegans*, surprisingly did not code for protein, but instead acted as a ~22 nt RNA transcript (Chalfie et al., 1981; Lee et al., 1993). Experiments showed that this small RNA molecule regulated its first known target, *lin-14,* by base-pairing to the 3′ untranslated region (3′ UTR) of *lin-14* mRNA via partial sequence complementation (Lee et al., 1993; Wightman et al., 1993). The downregulation of lin-28 in *C. elegans* also depends upon a short *lin-4*

target sequence in its 3' UTR (Moss and Tang 2003). This work established the dogmatic understanding of miRNAs as post-transcriptional regulators that target the 3′ UTRs of protein-coding genes to repress gene expression. Since then, miRNAs have been shown to play a variety of regulatory roles and target other genic regions in addition to 3′ UTRs (Brennecke et al., 2005; Easow et al., 2007; Grimson et al., 2007; John et al., 2004; Lytle et al., 2007; Nakamoto et al., 2005; Place et al., 2008; Vasudevan et al., 2007).

The discovery that the second characterized miRNA, *let-7* (Reinhart et al., 2000) was evolutionarily conserved across *bilateria (*Pasquinelli et al., 2000) and that its expression was regulated through the course of development, sparked concerted efforts to identify other miRNAs and elucidate their functions. A framework to study the consequences of miRNA regulation emerged from findings that mRNA 3′ UTRs widely contain evolutionarily conserved elements that are complementary to the 5′ end of certain miRNAs and mediate repression by these miRNAs (Brennecke et al., 2005; Doench and Sharp, 2004; Lee et al., 1993; Lewis et al., 2005; Lewis et al., 2003; Moss et al., 1997; Reinhart et al., 2000; Slack et al., 2000; Wightman et al., 1991; Wightman et al., 1993). For example, within the 3' UTR of the protein-coding gene lin-28 there are let-7 complementary sites that first suggested let-7 recognizes and regulates lin-28 mRNA (Pasquinelli et al., 2000; Reinhart et al., 2000). The finding that a mutation in the miRNA *let-7* partially suppresses the *lin-28* mutant phenotype supported this idea (Reinhart et al., 2000). The short stretch of nucleotides at the 5' end of the mature miRNA that pairs with target sequences is known as the ‘seed sequence’ (Lewis et al.,

2005; Lewis et al., 2003). Families of miRNAs share high sequence homology throughout their length, but in particular have matching seed sequences, and hence often target the same sites within mRNA. In all there are twelve let-7 family miRNAs from eight individual loci have been annotated in humans, including let-7a-1, -2, -3; let-7b; let-7c; let-7f-1, -2; let-7g; let-7i; miR-98) (Griffiths-Jones et al., 2006; Johnson et al., 2005).

Subsequent studies have solidified our understanding that let-7 miRNA family members bind to the 3' UTR of LIN28 mRNA and repress its protein production (Morita and Han, 2006; Rybak et al., 2009; Rybak et al., 2008; Wu and Belasco, 2005; Zisoulis et al., 2010) (Melton et al., 2010). Another miRNA, miR-125b, is also capable of binding to its seed complementary site and repressing Lin28 expression (Wu and Belasco, 2005). MiR-125b and miR-125a are the human homologues of *C. elegans lin-4* that acts in lin-28 regulation to control hypodermal stem cell self-renewal (Lagos-Quintana et al., 2002). In hematopoietic lineage specification of K562 human cells, miR-181 can directly bind its target sequence in the LIN28 3' UTR thereby repressing its expression and enabling megakaryocytic (MK) differentiation (Li et al., 2012b).

From these foundations, the field of miRNA biology has quickly progressed. Once researches understood that each miRNA could target many mRNAs, and thereby have a profound effect on cellular physiology, they sought to define these networks of targets. It has been estimated that miRNAs collectively regulate ~30% of all human genes (Lewis et al., 2005). Consequently, misregulation of miRNAs or disruption of

their target sites in genes has been implicated in a variety of human diseases ranging from cancer metastasis to neurological disorders (Chou et al., 2013; Goodall et al., 2013; Mishra et al., 2008). With the development and availability of genomic technologies and computational approaches, the field of miRNA biology has advanced tremendously over the last decade. High-throughput sequencing and accompanying computational algorithms have enabled transcriptome profiling in response to miRNA expression and identification of miRNA target sites through its association with the RISC complex (see below). These global miRNA studies were some of the first genome-wide approaches applied to the study of mRNPs, and have allowed for the discovery of new miRNAs, the characterization of their targets, and a systems-level view of their impact on cellular function (see (Wilbert et al., 2012).

MicroRNA biogenesis relies on the action of RNA-binding proteins

Each step of miRNA biogenesis and function is regulated and made possible by RNA-binding proteins. MiRNAs begin as primary transcripts (pri-) transcribed by RNA polymerase II either independently, in polycistronic clusters, or within the introns of protein-coding genes (Bartel, 2004; Lagos-Quintana et al., 2001; Lau et al., 2001; Lee et al., 2002b). Primary miRNA transcripts are typically one the order of kilobases long and are capped, spliced, and polyadenylated (see Kim 2005 for review). Pri-mRNAs are processed into precursor (pre-) miRNAs by the Microprocessor proteins Drosha and DGCR8/Pasha (Cai et al., 2004; Lee et al., 2003). Drosha is a RNase III endonuclease RNA-binding protein that cleaves the pri-miRNA leaving a 5'

phosphate and ~2 nt 3' overhang on the RNA ends (Basyuk et al., 2003; Lee et al., 1993). The resulting pre-miRNA forms a characteristic hairpin structure and is exported into the cytoplasm by Exportin-5 and Ran-GTP (Bohnsack et al., 2004; Lund et al., 2004; Yi et al., 2003). In the cytoplasm Dicer, another endonuclease III protein, processes the pre-miRNA into a miRNA/miRNA* duplex (Grishok et al., 2001; Hutvagner et al., 2001; Jaskiewicz and Filipowicz, 2008; Ketting et al., 2001; Knight and Bass, 2001). Dicer was first identified as a RBP that controls biogenesis of small interfering RNAs (siRNAs)(Bernstein et al., 2001). The duplex miRNA/miRNA* duplex is unwound and the freed mature miRNA is loaded onto the RNA induced silencing complex (RISC) (Hammond et al., 2001; Mourelatos et al., 2002; Tabara et al., 1999) The miRNA/miRNA* are now more commonly referred to as the 5' and 3' arm transcripts, since the either or both can become the functioning mature miRNA, depending on conditions. The main effecter of the RISC complex in humans is Argonaute 2 (Ago2) and this protein interacts most closely with the miRNA. The bound mature miRNA guides the Ago2/RISC complex to target mRNA through partial sequence complementarity. Once bound, this complex catalyzes repression of gene expression by destabilization through deadenylation or decapping, and degradation of the mRNA, or prevention of translational initiation or elongation (Olsen and Ambros, 1999). More recently, it has also been demonstrated that miRNA targeting can cause up regulation of mRNA translation (Vasudevan et al., 2007).

LIN28 binds to the let -7 miRNA and regulates its biogenesis

While the basic lifecycle of a miRNA is described above, numerous other RNA and protein factors can be involved in miRNA regulation. In particular, each miRNA family or even family member can have its own processing co-factors, often dependent on cellular context. One of the earliest associations made between a specific miRNA and RBP regulator was the ability of LIN28 to regulate let-7 family microRNAs. As described above, the let-7 miRNA represses LIN28 expression; however, early genetic studies in *C. elegans* also suggested that LIN28 could act in a reciprocal manner to suppress *let-7* (Slack and Ruvkun, 1997). This sort of feedback mechanism is common in biological systems, in particular when signals must be tightly controlled as in the case of development. Proper development depends on the succession of many cell-fate choices that ultimately determine the shape, size, and function of an organism. The timing of these developmental events is critical and thus tightly controlled through both temporal and special mechanisms. “Heterochrony”, differences in developmental timing, often lead to gross abnormalities in body patterning. These differences are related to the precocious and retarded expression of cell fate specifiers which result from mutations in the heterochronic genes that control this process.

Both LIN28 and let-7 are heterochronic genes that regulate developmental timing across species (Ambros and Horvitz, 1984 (Moss et al., 1997; Reinhart et al., 2000); however, it was observed that they had antagonistic patterns of expression. That is, let-7 is expressed later in development in differentiated cells types, while

LIN28 expression is primarily confined to early development and multipotent cells (Moss et al., 1997). Indeed, Pasquinelli and colleagues first observed that later in animal development when most cells have high let-7 expression, the cells with the lowest expression were also the most immature cells (bone marrow cells) (Pasquinelli et al., 2000). The direct connection between these two regulators was established through an effort to explain the observed contradiction between expression of the precursor let-7, but not mature let-7, in mESCs and hESCs (Suh et al., 2004; Thomson et al., 2006). The consistent levels of pre-let-7, but not mature let-7, across differentiated and undifferentiated cells provided evidence that let-7 biogenesis is regulated at the Dicer processing step and motivated a search for the responsible co-factors (Wulczyn et al, 2007).

Using a region of the let-7 miRNA pri/precursor loop that was conserved in sequence among the let-7 family members as bait, proteins bound to the transcript were analyzed using mass spectrometry. This approach in mouse embryonal carcinoma P19 cells revealed Lin28a and Lin28b as major protein interactors with the let-7 terminal loop (Newman et al., 2008; Viswanathan et al., 2008). Studies in mammalian embryonic stem cells, and comparisons of differentiated versus multipotent cell lines, demonstrated that LIN28A and LIN28B interaction with the terminal loop of the let-7 hairpin interferes with Drosha processing of pri-let-7 miRNAs (Newman et al., 2008; Viswanathan et al., 2008; Piskounova 2008, Hagan 2009). Rybak and colleagues reported that Lin28 regulation of let-7 occurs in the cytoplasm where Dicer processing of the miRNA was affected (Rybak et al, 2008;

Heo et al., 2008; Heo 2009). These early reports also demonstrated the *in vitro* ability of LIN28B to block let-7 processing in the same manner as LIN28A and suggested that these proteins have equivalent role in this process. Mutational studies by Balzer and Moss made it clear that the mechanism of blocking let-7 processing went beyond simple binding of Lin28 and likely involved interactions with other co-factors. This was evidenced by the ability of LIN28 with point mutations in the zinc-finger knuckles to bind precursor miRNAs but not affect miRNA biogenesis (Balzer and Moss, 2007). Other single point mutations in the CSD or ZFD did reduce Lin28 binding to let-7 and had proportional effects on the ability to block let-7 processing demonstrating that both RNA-binding domains play a role in this regulation (Piskounova 2008).

The mechanism of this regulation was further elucidated by researchers who showed that Lin28a in the cytoplasm mediates the addition of multiple uracil (U) residues to the 3' end of the pre-let-7 miRNA (Heo et al., 2008). This uridylation results in a pre-let-7 with an elongated 3' tail that cannot be processed by Dicer. Furthermore, this modified polyuridylated pre-let-7 is less stable and more readily degraded than the unmodified form (Heo et al, 2008). The terminal uridyly transferase responsible for this reaction was identified as TUTase4 (TUT4 or ZCCHC11), a non-canonical poly (A) polymerase (Heo et al., 2009; Hagan et al., 2009). The function of Zcchc11 in Lin28-mediated let-7 repression relies on a single C2H2-type zinc finger domain within the protein and is redundant with another TUTase, Zcchc6 (TUTase7) (Thornton et al., 2012). In contrast to polyuridylation, in the absence of LIN28 binding

monouridylation of pre-let-7 by Tut7/4 or another TUTase, Tut2 (Papd4/GLd4), generates a 2 nt 3' overhang which is an ideal substrate for Dicer processing (Heo et al., 2012).

Interestingly, Lin28B was found to function in a TUT4-independent mechanism, instead acting within the nucleus to block Drosha/Microprocessor processing (Piskonova et al., 2011). Currently the mechanism for this sequestration within the nucleus is unknown, although the active nuclear localization of LIN28B through its nuclear and nucleolar localization signals are likely to be contributing factors (Piskounova et al., 2011). The nuclear localization of LIN28A was observed during neural differentiation of embryonic stem cells was contributed to recruitment by the RBP Mushashi1 (Msi1); however, whether this protein plays a role in LIN28B localization remains to be tested (Kawahara et al., 2011). Differences in let-7 processing and variation in LIN28A and LIN28B expression in human cancer cells and tissues (see below) has important implications for therapeutic approaches. For example, inhibition of Zcchc11 in LIN28A expressing cells reduced their tumorigenicity and invasiveness, but had no affect in LIN28B expressing cells since it does not rely on this enzyme for function (Piskonova et al., 2011).

Lehrbach and colleagues reported that the mechanism of let-7 uridylation is conserved in *C. elegans* and mediated by a potential Zcchc11 ortholog, PUP2 (Lehrbach et al., 2009). More recently in *C. elegans*, *lin-28* has been shown to interact with primary *let-7* (*pri-let-7*) transcripts, thereby blocking Drosha processing and

preventing mature let-7 expression independent of an uridylation mechanism (Van Wynsberghe et al., 2010).

Studies in *Lin28a/b* knockout mice show that Lin28a, but not Lin28b, affects let-7 miRNAs in the early embryo (Shinoda et al., 2013). The primary role of Lin28a in regulating embryonic, but not adult let-7 levels is consistent with its expression mainly confined to fetal tissues. Knockout of Lin28b, which has higher expression in adult tissues, affected let-7 levels in adult skeletal muscle, but not in neonatal tissue where Lin28a had an effect. Together this evidence supports a model where Lin28a acts to regulate let-7 biogenesis during embryogenesis and this role decreases during development while Lin28b maintains control of let-7 during adulthood (Shinoda et al., 2013).

The specific site of LIN28 binding to pre/pri-let-7 miRNAs has been reported as a conserved cytosine in the loop (Piskounova et al., 2008), and a GGAG sequence downstream of this residue (Heo et al., 2009). This GGAG sequence is sufficient to induce repression of an unrelated miRNA, miR-16 and is necessary for terminal uridylation of pre-let-7 (Heo et al., 2009). The crystal structure of Lin28 in complex with *let-7* sequences provided definitive evidence that the Lin28 CCHC zinc finger knuckles dimerize at this GGAG sequence (Nam et al., 2011). Since this position is adjacent to the Dicer cleavage site LIN28 binding here would prevent its activity. The CSD domain binds another structured motif with a similar GNGAY sequence (Nam et al., 2011). It has been proposed that the initial binding of Lin28 ZFKs at the conserved GGAG site acts to disrupt base pairing of the *let-7* stem making it impervious to Dicer

cleavage. This would expose additional GU rich motifs, allowing for multiple LIN28 molecules, as up to three have been predicted to bind to a single pre-let-7a RNA molecule (Hafner et al., 2013). A number of other miRNAs with GGAG motifs have also been identified as regulated by Lin28 through terminal uridylation; however, this regulation appears subtle compared to the effect on let-7; examples include pre-miR-107, 143 and 200c (Heo et al., 2009). Further reports have demonstrated that RNAs corresponding to the loop of pre-let-7 or the mature let-7 sequence or can compete for LIN28 binding, indicating interactions with both these sequences (Rybak et al., 2009).

LIN28 IN HUMAN PLURIPOTENT CELLS

LIN28 is expressed in an oncofetal manner and can drive pluripotency *in vivo* and *in vitro*. These abilities are often linked to its repression of let-7 biogenesis and the release of downstream oncogenes from repression. However, let-7 independent effects of LIN28 expression likely contribute to its role in tumorgenesis. For example, LIN28 can bind and directly enhance translation of mRNAs encoding cell-cycle regulators and other growth factors to promote cellular proliferation (as discussed below). The powerful role of LIN28 in embryonic and cancer stem cells drives our desire to better understand how it contributes to pluripotency.

LIN28 overexpression drives tumorigenesis

The relationship between LIN28 and let-7 is intimately linked to the role of those regulators in development and tumorigenesis. Let-7 is classified as a tumor

suppressor due to its action in down regulating a number or oncogenes including RAS, Hmga2, and MYC, (Johnson et al., 2005; Mayr et al., 2007; Sampson et al., 2007). In many instances, the downregulation of let-7 has been attributed re-expression of LIN28 in adult tissues where it is not normally found. Most commonly the model that follows is that LIN28 upregulation results in repression of mature let-7 expression, releasing let-7 target oncogenes (such as RAS, Hmga2, and MYC) from repression allowing their overexpression and activation of tumorigenesis. One of the first mechanistic studies looking at LIN28 expression itself in cancers found approximately 15% of primary tumors across various tissue types had re-expression of LIN28 where it was not normally found in healthy tissues (Viswanathan et al., 2009). The authors demonstrated that overexpression of LIN28A/B can promote malignant transformation of cells (Wang et al., 2010; West et al., 2009). Furthermore, the expression of LIN28A and LIN28B was assocaited with more advaced disease and poor clinical prognosis, indicating the theraputic importance of these proteins (Viswanathan et al., 2009). The overexpression of LIN28 was critically linked to let-7 repression in disease tissues, and rescued expression of let-7 in LIN28 overexpressing cells rescued repression of ongogene targets and cellular overgrowth phenotype (Viswanathan et al., 2009).

An upstream driver of LIN28A/B transcription that has often been attributed to their re-expression in cancer is the trancription factor Myc, which has been shown to bind to the Lin28B promoter (Chang et al., 2009). In a differenet breast cancer model *Lin28b* expression was transctivated by NF-kB and Src (Iliopoulos et al., 2009). Release from a negative regulation, Tristetraprolin (TTP), might lead to an increase in

LIN28A mRNA as an addition mechanism of LIN28 overexpression (Kim et al., 2012).

A number of studies have reported LIN28A/B overexpression in a range of tumor types (for review see Thornton and Gregory, 2012). Embryonic carcinoma cells, such as P19 cells, were some of the first used to study Lin28A activity and continue to be a useful model for retrinoic-acid-induced neurogenesis (see below) (Hagan et al., 2009; Heo et al., 2008; Heo et al., 2009). Overexpression of LIN28A and LIN28B have been documented in ovarian cell lines, usually mutually exclusive of oneanother (IGROV1, A2780)(Lu et al., 2009; Peng et al., 2010; Piskounova et al., 2011). Primary prostate tumors samples (Nadiminty et al., 2012), germ cell malignancies (Cao et al., 2011a; Cao et al., 2011b; West et al., 2009) and human colon adenocarcinomas (Iliopoulos et al., 2009; King et al., 2011a; King et al., 2011b; Piskounova et al., 2011) have also been documented to express LIN28A or LIN28B at high levels. Their overexpression was further impliacted in breast (Piskounova et al., 2011; Sakurai et al., 2012), lung (Chang et al., 2009; Pan et al., 2011; Piskounova et al., 2011; Viswanathan et al., 2009; Zhang et al., 2012) and esophageal (Hamano et al., 2012) cancer cells and tumors.

Primary tumors and cell lines of hepatocellular carcinoma, including HepG2, Huh7, Hep3B, HCC36 and HA22T were commonly found overexpressing LIN28B and coorelated this with poor prognosis (Cairo et al., 2010; Guo et al., 2006; Wang et al., 2010). LIN28B has also been particularly indicated in blood cancers such as chronic myelogenous leukemis (K562) (Piskounova et al., 2011; Viswanathan et al.,

2009), B cell lymphoma (P-49306) (Chang et al., 2009), and medulloblastoma primary tumors (Rodini et al., 2012). In the skin cancer cell line SK_MEL_28 LIN28B expression was necessary for increased tumor growth in zenograph models (Piskounova et al., 2011). A number of these reports associate increased LIN28 expression with poor prognosis and suggest that LIN28 can serve as a marker of distrinct tumorigenic subpopulations. For example, LIN28 has been proposed as a marker of particular germ cell tumors (Cao et al., 2011a; Cao et al., 2011b). The Huang group has described a sub-population of stem cell-like ovarian cancer cells that highly expressed LIN28 and OCT4 expression (Peng et al., 2010). An intriguing possibility that has arisen from these studies is the suggestion that LIN28A/B expression may be a feature of a subpopulation of tumor initiating, cancer stem cells (CSCs) (Zhang et al., 2012).

Often parallels can be drawn between regulation that contributes to the ability of tumorigenic cells and embryonic stem cells to proliferate rapidly and indefinitely. This holds true for the dichotomy between let-7 and LIN28 expression, and the role of LIN28 in promoting growth of pluripotent cells. Below we discuss the action of LIN28 in embryonic stem cells, how it contributes to cellular proliferation and development, and how its regulation of mRNA transcripts contributes to these functions.

The study of LIN28 in human embryonic stem cells

Human embryonic stem cells are derived from the inner cell mass (ICM) of 5-day-old embryos and have the ability to differentiate into any cell type of the three germ layers (Reubinoff et al., 2000; Richards et al., 2002; Thomson et al., 1998). Studies have shown that Lin28a and Lin28b contribute to maintenance of pluripotency of the ICM and epiblast through repression of let-7 miRNA (Melton et al., 2010; Suh et al., 2010). Single-cell expression analysis demonstrates that when ICM cells are cultured to establish embryonic stem cell lines their attainment of self-renewal capacity coincides with increased LIN28 expression (Tang et al., 2010).

It was the first studies of the transcriptome in human embryonic stem cell lines that readily identified LIN28 mRNA as highly expressed in these cells (Richards et al., 2004). The observed downregulation of LIN28 upon ES cell differentiation, and conservation of these expression patterns in mouse ES cells helped define LIN28 as a marker for undifferentiated stem cells (Richards et al., 2004). Interestingly, despite the impact on cellular phenotype and contribution to overall development, LIN28 expression is dispensable for hESC self-renewal, as knockdown of LIN28 in H9 cells did not lead to a significant morphological change in hESCs or a decrease in OCT4 or TRA-1-60 levels (Darr and Benvenisty, 2009). This is consistent with our own studies in these cells (data not shown).

However, other studies in H1 hESCs and human embryonic carcinoma cells (PA-1) reported that OCT4 expression was decreased upon LIN28 knockdown, and furthermore that LIN28 showed binding affinity for the OCT4 mRNA (Qiu et al.,

2010). The basic approach used in these early works to identify RNA transcripts bound by LIN28 protein was to immunoprecipitate (IP) the LIN28 protein either via an epitope tag or antibody to the endogenous protein and then isolate and reverse transcribe the associated RNAs. Use of primers specific to candidate genes can then determine if a particular mRNA is more enriched in a LIN28 IP versus a control protein not expected to bind specific RNAs, for example IgG. Using these IP techniques in mouse ES cells Xu and colleagues also did not find Oct4 to be a strong Lin28 mRNA target (Xu and Huang, 2009; Xu et al., 2009). Furthermore, Cho and colleagues recently reported that with 70% knockdown of Lin28a in mESCs Myc, Pou5f1 and Nanog mRNA increase ~ 25% (Cho et al., 2009). Although this small increase in pluripotency factors was deemed insignificant, the trend is in conflict with reports that Lin28a knockdown in mESCs and hESCs decreases Oct4 expression.

LIN28 contributes to stem cell growth through regulation of the cell cycle. The expression of LIN28 in mouse (Xu et al., 2009) or human (Qiu et al., 2010) ES cells promotes cellular proliferation and binds mRNAs encoding proteins responsible for cell-cycle control including Histone H2a, Cyclin A, Cyclin B, and Cdk4. Knockdown of Lin28 in mouse ES cells resulted in a shift toward away from the G2/M phase of the cell-cycle, indicating LIN28 facilitates a rapid cell-cycle and proliferation. In seeming contradiction to this, overexpression of LIN28 in human ES cells also caused a decrease in the rate of cell cycle and knockdown had no affect (Darr and Benvenisty, 2009). LIN28 overexpression was also shown to reduce apoptosis in H9 cells but did not increase it with knockdown in mouse ES cells (Darr and Benvenisty, 2009;

Polesskaya et al., 2007). Despite some discrepancies in reports, the consensus from the bulk of LIN28 studies it that its expression contributes to cell growth and proliferation.

Perhaps somewhat counter intuitively, LIN28-overexpressing cells had increased differentiation (via a decreases in TRA-1-60 expression compared to control)(Darr and Benvenisty, 2009). However, this phenomenon also occurs with differentiation drive by overexpression of OCT4 or SOX2. Overexpression of Oct-3/4 causes differentiation to the primitive endoderm and mesoderm (Niwa et al., 2000). NANOG helps to maintain pluripotency by preventing differentiation to the primitive endoderm (Chambers et al., 2003; Mitsui et al., 2003; Wang et al., 2003). Its overexpression drives cells towards the ectodermal fate (Darr et al., 2006). In the absence of Nanog the epiblast layer that ectodermal cells are derived from fails to develop and mice cannot develop past the blastocyst stage (Wang et al., 2003). Since these pluripotency factors are also lineage specifiers we can think of the state of pluripotency as a balance between counteracting forces of lineage specification. The ability of these pluripotency regulators to also drive differentiation when introduced in access amounts may be related to the redundancy observed in overexpression and knockdown phenotypes of LIN28 (Darr and Benvenisty, 2009).

LIN28 IS A REPROGRAMMING FACTOR FOR iPSC INDUCTION

These early studies of LIN28 in embryonic stem cells established its place of importance among the transcription factor pluripotency associated genes like OCT4. This fact was driven home when LIN28, in combination with OCT4, SOX2, and

NANOG were used to successfully reprogram adult fibroblasts into an embryonic-like stem cell fate (Yu et al., 2007). Since these foundational works, LIN28 overexpression has contributed to reprogramming protocols used in a wide variety of animals for the generation of species-specific iPSCs (Deng et al., 2012; Nethercott et al., 2011; Song et al., 2013; Tomioka et al., 2010).

These core reprogramming factors themselves are known to bind the LIN28 promoter to drive its transcription (Marson et al., 2008). Sequencing of single cells during reprogramming has demonstrated that early gene expression events are stochastic, among which Lin28 expression is a better indicator of cells that go on to fully form iPSCs than previous standards of Oct4 or Fgf4 (Buganim et al., 2012). Using Bayes probabilistic network modeling to represent conditional dependencies between gene expression profiles in these cells these authors identified Sox2 expression as a critical switch initiating hierarchical activation of key pluripotency genes, including Lin28 (Buganim et al., 2012). Nuclear association of the LIN28 and Sox2 proteins adds to evidence that these factors could function together (Cox et al., 2010).

Study of the role of LIN28 in reprogramming provided evidence that its contribution is to promote proliferation of cells (Hanna et al., 2009). Similar to the role of LIN28 in cancer, many of its affects in reprogramming may be contributed to its repression of let-7 biogenesis and downstream affects. Let-7 itself is essential for differentiation; its expression alone can rescue differentiation defective Dgcr8-/- mESCs in the absense of other miRNAs (Melton et al., 2010). Removing let-7 activity

via an anti-sense oligonucleotide improves reprogramming efficiency in mouse fibroblasts (Melton et al., 2010). The intertwined role of let-7 and LIN28 in regulating metabolism has also been proposed to contribute to stem cell pluripotency. Embryonic and induced pluripotent stem cells are known to shift in glycolytic activity to an aerobic state, similar to the Warburg effect (Folmes et al., 2011). Evidence supports the theory that by binding to mRNAs encoding mitochondrial enzymes and counteracting let-7 activity LIN28 likely affects mitochondrial oxidation and downstream metabolites to support metabolism of pluripotent cells (Shyh-Chang and Daley, 2013; Shyh-Chang et al., 2013a).

Induced pluripotent stem cells (iPSCs) as a model system

Although the function of LIN28 in iPSC generation is of great interest to the scientific community, this new system of study presents some challenges to transcritome-wide RBP characterization. This technology has already begun to revolutionize approaches to disease modeling and drug screening by making multipotent populations of patient cells available for *in vitro* modeling. However, high-throughput exome sequencing of 22 iPSC lines from five different reprogramming methods all revealed evidence of reprogramming-associated mutations suggesting these cells inherently acquire polymorphisms in cancer genes during the reprogramming process (Gore et al., 2011). These early questions about the genetic stability of iPSCs suggested this would be a potentially volatile system to carry out our foundational studies of LIN28 function. Even with better-studied hESCs it is common practice to use cells within 50 passages due to the accumulation of *in vitro* artifacts,

and in particular chromosomal abnormalities. Being able to reliably assign RNA transcripts to their genomic source is a critical aspect of transcriptome studies and is dependent upon accurate gene annotations. Even in a steady genetic background we are only beginning to appreciate the complexity of the transcriptome. For examplein, in the well-studied model systems *C. elegans* RNA sequencing revealed alternative unannotated untranslated regions for 40% of the genes in the genome (Zisoulis et al., 2010). These lessons are important to keep in mind especially while investigating cell-types and systems where previously unobserved transcripts and modifications are likely to exist.

LIN28 mRNA REGULATION AND DEVELOPMENTAL CONTROL

The early studies of LIN28 in stem cells established some of the first known mRNA targets of LIN28 such as OCT4, CDK4, and H2a, and the understanding that a common function of LIN28 in binding mRNA is to enhance translation (Qiu et al., 2010; Xu and Huang, 2009; Xu et al., 2009). The ability of LIN28 to regulate mRNA translation and, in some cases RNA transcript levels, is critical for its impact during early differentiation. These roles outside the regulation of let-7 are particularly obvious in the specification of skeletal muscle and neural lineages.

Myogenesis relies on LIN28 enhanced translation

Studies in differentiating skeletal muscle cells lead to the discovery of the insulin-like growth factor 2 (IGF-2) as one of the first known mRNA targets of Lin28

(Polesskaya et al., 2007). In most adult tissues Lin28 expression is silenced; however, its expression remains low in skeletal and cardiac muscle (Yang and Moss, 2003)(see also Figure 1). Interestingly, Lin28 expression actually increases during differentiation of primary myoblasts and regeneration of skeletal muscle fibers (Polesskaya et al., 2007), in sharp contrast to its general decrease during differentiation (Richards et al., 2004; Lee et al., 2005). RNA levels of IGF-2 were unchanged under the influence of Lin28; similar to what was reported for its stem cell mRNA targets, indicating its main activity was not to affect transcript levels. Instead, Lin28 regulation resulted in increased protein production of IGF-2 and was involved in an increase in the number of ribosomes bound to IGF-2 mRNA, indicative of an active translational state. This work was the first to establish LIN28 as a "translational enhancer" (Polesskaya et al., 2007). This is in agreement with observations made in P19 and mouse ES cells that Lin28 protein is enriched in the polysome fraction where translating ribonucloprotein complexes segregate during sucrose gradient centrifugation (Balzer and Moss, 2007; Cho et al., 2010; Qui et al, 2009).

Protein interactions with translational machinery including the 5' cap-binding protein, poly(A)-binding protein, eIF3ß translation initiation factor, and EF1-α/-α2 translation elongation factors supported the idea that Lin28 binding was directly associated with active translational units. In mouse myoblast C2C12 cells the ability of Lin28 to drive IGF-2 translation supports the growth and differentiation of muscle cells. This impact is so significant that Lin28 expression is necessary for efficient differentiation of myoblasts to myotubes (Polesskaya et al., 2007). It is possible that

defects in skeletal muscle differentiation due to Lin28 deficiency are the underlying cause of the cardiac ventricular septal defects noted as the only significant histological defect in Lin28a KO mice that died perinatally (Shinoda et al., 2013).

LIN28 acts with catalytic protein intermediates to affect translation

Mass spectometery analysis of LIN28 interacting proteins in mouse ES cells also identified RNA-independent interactions with components of the translational machinery including PABP and translation initiation factor eIF3beta (Qui et al., 2009). In addition, this study identified RNA helicase A (RHA) bound to Lin28 protein and showed that expression of this helicase was necessary for full stimulation of Oct4 mRNA translation. RNA helicase A (RHA) is a ubiquitously expressed member of the highly conserved DEAD-box family of RNA helicase proteins that acts in various RNA processing events including splicing, nuclear export, translation, and RNA interference (Bleichert and Baserga, 2007; Robb and Rana, 2007). Functional interaction of LIN28 and RHA was also described in human ES and embryoinc carcinoma PA-1 cells and specified that the C-terminus of LIN28 had the most influence on direct binding to RHA protein (Peng et al., 2011). A second group simultaneously confirmed the significance of the LIN28 C-terminus for direct interactions with RHA, proposing a model where LIN28 directs the RHA protein to mRNA targets (Jin et al., 2011). Importantly, LIN28 with mutations in either RNA-binding domain can still bind RNA, but these mutations affect its ability to associate with RHA and subsequently affect translation from these mRNAs.

Drosha has also been proposed as a co-factor in LIN28 mRNA regulation. One report suggests that some LIN28 targets contain particular sequence and structural element centered around an “A” bulge with flanking G:C base paired residues that signal Drosha-mediated destabilization of the RNA target (Qiao et al., 2012). This rule has not been demonstrated on a large-scale with many LIN28 target transcripts. A global study of Lin28A and Lin28b targets in mouse ES cells demonstrated that select mRNAs marked for translation in endoplasmic reticulum (ER)-bound ribosomes are selectively repressed by Lin28 binding (Cho et al., 2012). This indicates the outcome of Lin28 mRNA regulation may also be dependent on cellular localization of translation and co-factors.

Neurogenesis is regulated by LIN28

One lineage LIN28 has been particularly associated with is the neurogoenesis. In the mouse embryo, Lin28 mRNA expression co-localized with SOX2 in the developing neural tube before neural differentiation begins (Blazer et al., 2010); however, these levels are almost completely diminished by E 10.5 of mouse development (Darr et al., 2009; Balzer et al., 2010). The functional association of LIN28 and SOX2 was explained in neural progenitor cells where SOX2 acts to acetylate the LIN28 promoter to maintain an optimal level of its expression in these cells (Cimadamore et al., 2012). LIN28 is also downregulated during retinoic acid induced differentiation of P19 embryoinc carcinoma cells, a model of neurogliogensis (Wu and Belasco, 2005). This was attributed to the negative influence of miR-125b binding to the LIN28 3' UTR (Wu and Belasco, 2005). Constitutive expression of

Lin28a or Lin28b during this process pushes cells towards the neural lineage at the expense of glialgenesis (Balzer et al., 2010). Importantly, there was evidence that LIN28-impacted change in lineage regulation occurred prior to any effect on let-7 processing. This data supports the conclusion that LIN28 has important regulatory roles beyond its control of let-7 biogenesis. Stem cells expressing LIN28 under an inducible reporter formed teratomas with an overrepresentation of primitive neural tissue (West et al., 2009), something that has been associated with high-grade, aggressive tertocarcinomas in humans and indicates a propensity of LIN28 to drive neurogenesis. In neural stem cells LIN28 is down regulated by let-7 (Rybak et al., 2008). Within the developing mouse brain, Lin28 has been shown to regulate a particular set of mRNAs, downstream of Brain-Derived Neurotrophic Factor (BDNF) (Huang et al., 2012). BDNF indirectly aids to promote the translation of a set of genes important for the differentiation of excitatory neurons through positive regulation of Lin28 expression. Thus, although LIN28 is not expressed in adult neurons, it plays a role in earlier in their development.

Metabolism and Angiogenesis

LIN28 expression early in development also affects the ability to maintain normal glucose homeostasis. This was exhibited by the impaired glucose tolerance and susceptibility to high-fat diet induced diabetes in Lin28 deficient transgenic mice (Shinoda et al., 2013; Zhu et al., 2010; Zhu et al., 2011). Overexpression of human *LIN28B* or mouse *Lin28a* improves the ability of these mice to respond to increased insulin levels providing protection from diabetes. This antidiabetic phenotype is

phenocopied by let-7 repression with locked nucleic acid (LNA)-modified antimiRs (Frost and Olson, 2011; Zhu et al., 2011). Furthermore, *let-7* overexpression increases insulin resistance and impaired glucose tolerance (Zhu et al., 2011). This phenotype was linked to the insulin-PI3K-mTOR pathway. Let-7 is known to regulate several components of this signaling pathway that the authors used to explain downstream effects of Lin28 deficiency. However, despite the inverse correlation between the affect of Lin28 and let-7 on glucose and insulin tolerance, there is some evidence that the phenotype driven by Lin28 is not dependent upon the mechanism driven by let-7 under all conditions. Mice with muscle-specific knockout of Lin28a maintain levels of let-7 while still displaying defects in glucose tolerance and insulin resistance, suggesting a let-7 independent mechanism (Zhu et al., 2011). However, more recent evidence suggests that these changes in metabolism are the result of LIN28 function during early embryogenesis, thus making adult levels of let-7 irrelevant to its effects (Shinoda et al., 2013). Conditional knockout of Lin28 at different developmental states demonstrated that fetal difficiency, but not adult or neonatal, impacted glucose metabolism. This reveals that a role of Lin28 during development is to prime later adult metabolic function (Shinoda et al., 2013).

A rare minority of *Lin28a* and *Lin28b* knockout (KO) mice survive past birth, enabling better comparison of these protein paralogs in development (Shinoda et al., 2013). *Lin28a* KO animals have decreased bone mineral density and fat mass, and proportionately smaller organ sizes, indicating metabolic dysfunction as early as E13.5. In contrast, Lin28b KO mice do not exhibit growth defects until after birth with

evidence of postnatal dwarfism only in male animals. This provides an interesting example of gender-specific growth regulation by *Lin28b* that remains to be explained. Induced let-7 expression specifically in skeletal muscle is sufficient to drive postnatal dwarfism in male and female mice demonstrating its role in this phenotype (Shinoda et al., 2013). These changes in animal growth were associated with lower rates of glycolysis in the early embryo. Although LIN28 is not expressed in adult fat tissues, its regulation has an effect on fat mass in LIN28 transgenic mice, likely resulting from these defects in the 'programming' of metabolic function in the embryo (Shinoda et al., 2013). Early influence of Lin28a can also explain the absence of dwarfism in muscle specific Lin28a KO mice, whereas Lin28b muscle specific KO had the same phenotype as total KO (Shinoda et al., 2013). The ability of fetal Lin28b expression to program adult muscle metabolism and growth was contributed in part to an affect on the Tsc1-mTORC1 pathway. The ability of early expression of LIN28 to affect long-term epigenetic changes in metabolic function provide support for the generally accepted "Baker hypothesis" that fetal and infant nutrition have lifelong affects on the risk for type 2 diabetes mellitus (T2DM) and other metabolic disorders (Hales and Barker, 1992, 2001). It also begins to explain the association of polymorphisms in *LIN28* and its promoter region with T2DM in women (Zhang et al., 2013).

CONCLUSIONS

As discussed above, LIN28 expression is critical for normal development across animal species (Ambros and Horvitz, 1984; Zhu et al., 2010). Its role in

regulating let-7 biogenesis often accounts for downstream effects on cellular phenotype; however, this relationship alone cannot explain all of the effects of LIN28 expression. Examples of this can be seen during neurogenesis when LIN28 acts on mRNA targets before affecting let-7 and in stem cells with LIN28 enhances translation of cell-cycle regulators (Balzer et al., 2010; Xu et al., 2009). Perhaps the most convincing evidence of the extent that LIN28 acts through let-7 independent mechanisms comes from studies in *C. elegans*. The Moss group used mutant nematodes to show that even without *let-7, lin-28* mutant animals still skipped one symmetric division in the seam cell lineage; thus this precocious phenotype does not depend on *let-7* accumulation (Vadla et al., 2012). The knowledge of only a handful of LIN28 mRNA targets greatly limited our ability to define what pathways and mechanisms contribute to its let-7 independent regulation. The first published study to identify transcriptome-wide targets of LIN28 revealed that it interacts with hundreds or even thousands of protein-coding transcripts (Peng et al., 2011). These foundations motivated us to carefully define LIN28 binding sites throughout the human transcriptome in order to shed light on its network of direct and indirect mechanisms of regulation.

METHODS

Sequence information and annotations

Amino acid and nucleotide sequences of LIN28A (gene id: 94536796; NBCI reference sequence NM_024674.4; Homo sapiens lin-28 homolog A) and LIN28B (gene id: 392513692; NBCI reference sequence NM_001004317.3; Homo sapiens lin-28 homolog B) were obtained from the National Center for Biotechnology Information resource (http://www.ncbi.nlm.nih.gov/). The sequence of protein functional domains was identified and visualized at UniProtKB for protein references Q6ZN17 (LN28B_HUMAN) and Q9H9Z2 (LN28A_HUMAN) (Magrane and Consortium, 2011).

Multiple sequence alignments

The Clustal Omega (Clustal O 1.2.0) tool at EMBL-EBI was used to generate multiple sequence alignments of the amino acid and nucleotide sequences of LIN28A and LIN28B (Sievers et al., 2011). The resulting alignment aims to represent a biologically meaningful alignment of divergent protein sequences. Percent identity results were generated with Clusal2.1.

LIN28A and LIN28B mRNA expression data

The GeneAtlas U133A dataset of tissue-specific mRNA expression from high-density oligionucleotide arrays from a panel of 79 human tissues (Su et al., 2004) was

visualized using the gene expression/ activity chart tool of BioGPS (Wu et al., 2013; Wu et al., 2009).

AUTHORS' CONTRIBUTIONS

Chapter 1, in part, is an adaptation of material that appears in "Genome-wide approaches in the study of microRNA biology" by Melissa L. Wilbert and Gene W. Yeo, as published in *Wiley Interdisciplinary Reviews: Systems Biology and Medicine* December 2010. M. Wilbert wrote and prepared this publication, with editing and guidance from Gene W. Yeo.

FIGURES

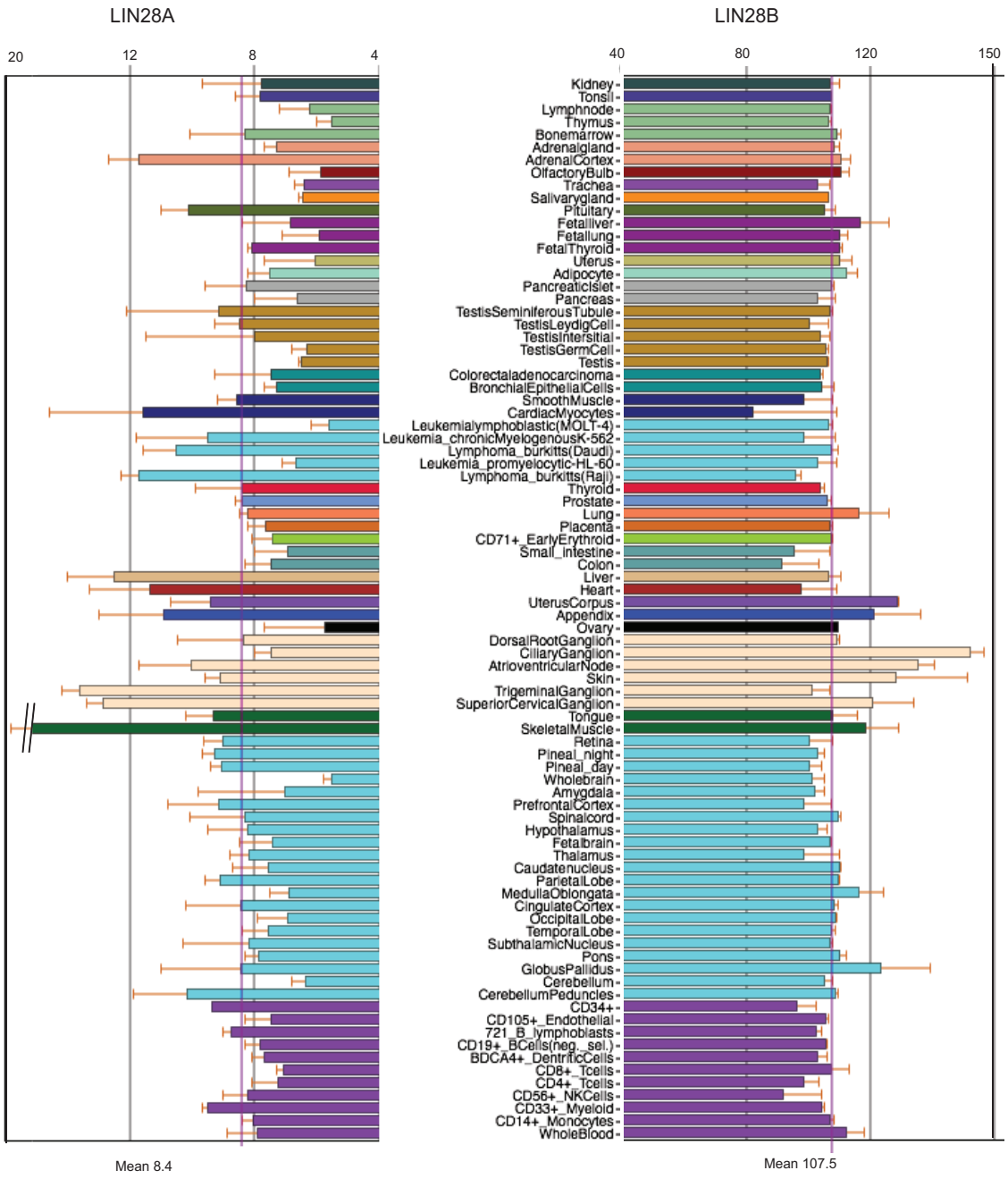


Figure 1. RNA expression levels of LIN28A and LIN28B in human tissues.
Visulalization of the GeneAtlas U133A dataset of tissue-specific mRNA expression from high-density oligionucleotide arrays from a panel of 79 human tissues.

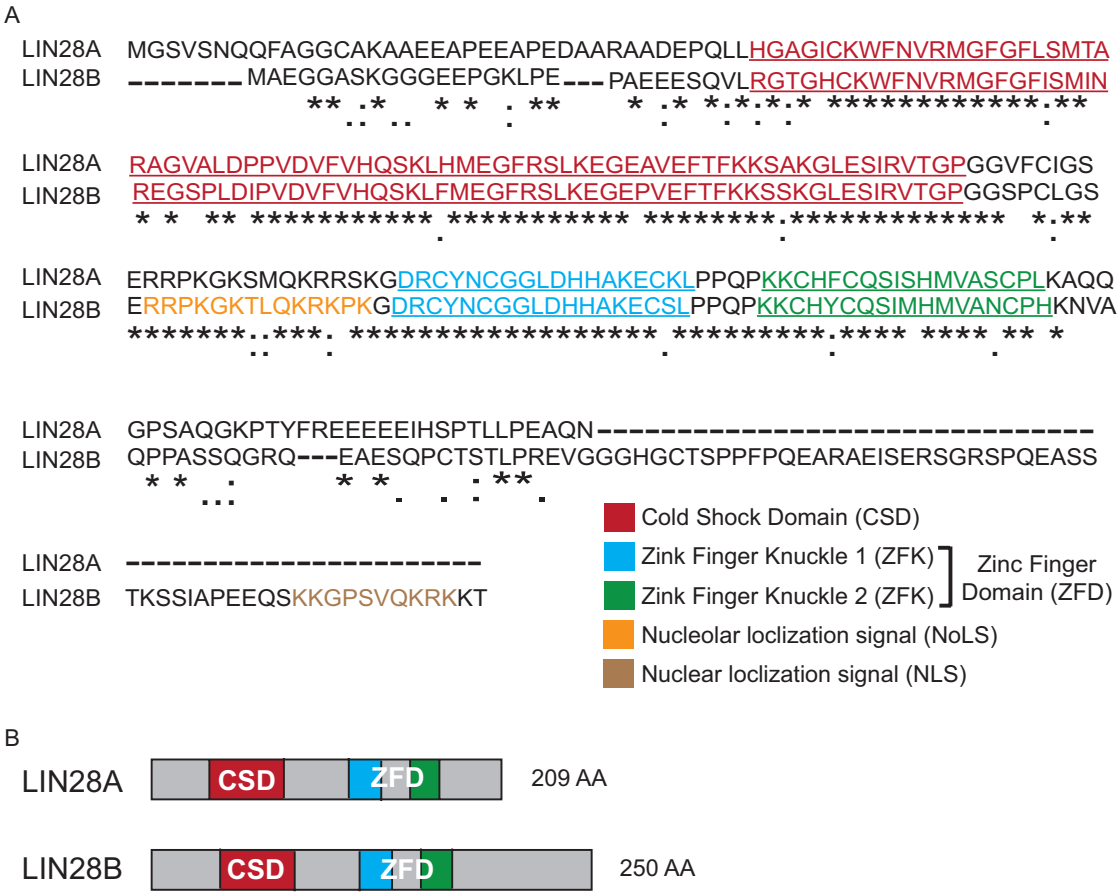


Figure 2. Comparison of LIN28A and LIN28B human protein sequences

(A) Multiple sequence alignment of human LIN28A and LIN28B amino acid (AA) sequences. A perfectly matched residue is annotated with an asterisk (*), alignment of residues with strongly similar properties are annotated with a colon (:), and alignment of residues with weakly similar properties are annotated with a period (.). The functional domains within LIN28A and LIN28B are color coded as described in the key. (B) Cartoon of the LIN28A and LIN28B peptide sequences with the Cold Shock Domain (CSD) in red and Zinc Finger Domains (ZFD) in blue and green.

CHAPTER 2 - LIN28 BINDS MESSENGER RNAS AT GGAGA MOTIFS AND REGULATES SPLICING FACTOR ABUNDANCE

ABSTRACT

LIN28 is a conserved RNA binding protein implicated in pluripotency, reprogramming and oncogenesis. Previously shown to act primarily by blocking let-7 microRNA (miRNA) biogenesis, here we elucidate distinct roles of LIN28 regulation via its direct messenger RNA (mRNA) targets. Through cross-linking and immunoprecipitation coupled with high-throughput sequencing (CLIP-seq) in human embryonic stem cells and somatic cells expressing exogenous LIN28, we have defined discrete LIN28 binding sites in a quarter of human transcripts. These sites revealed that LIN28 binds to GGAGA sequences enriched within loop structures in mRNAs, reminiscent of its interaction with let-7 miRNA precursors. Among LIN28 mRNA targets, we found evidence for LIN28 autoregulation and also direct but differing effects on the protein abundance of splicing regulators in somatic and pluripotent stem cells. Splicing-sensitive microarrays demonstrated that exogenous LIN28 expression causes widespread downstream alternative splicing changes. These findings identify important regulatory functions of LIN28 via direct mRNA interactions.

INTRODUCTION

Post-transcriptional regulation of gene expression is fundamentally important to a multitude of cellular processes, including development, homeostasis and differentiation. RNA binding proteins (RBPs) interact directly with RNA transcripts in cells to exert various forms of regulation such as alternative splicing, turnover,

localization and translation (Glisovic et al., 2008). Altered expression levels of RBPs often results in genetic diseases and cancer (Lukong et al., 2008). Among these key proteins is LIN28A (herein referred to as LIN28). Conserved across bilaterian animals, LIN28 is highly expressed early in development and is selectively downregulated during differentiation (Moss et al., 1997; Yang and Moss, 2003). Consistent with this pattern of expression, LIN28 has been shown to be important in the maintenance of embryonic stem (ES) cell pluripotency and efficacy of induced pluripotent stem cell (iPSC) derivation (Moss et al., 1997; Newman and Hammond, 2010; Yu et al., 2007). Of the factors used in reprogramming, LIN28 is unique in its classification as an RBP, rather than as a transcription factor. Notably, aberrant upregulation of LIN28 has been found in a range of different cancer cells and primary tumor tissues (Cao et al., 2011a; Viswanathan et al., 2009; West et al., 2009).

LIN28 and its only paralog in humans, LIN28B, block the processing of let-7 microRNAs (miRNAs) by binding to the terminal loop of the let-7 precursor (pre-let-7) hairpin via a cold-shock domain (CSD) and two retroviral-like CHCC zinc-finger knuckles (Hagan et al., 2009; Heo et al., 2008; Heo et al., 2009; Nam et al., 2011; Piskounova et al., 2008). Subsequent reports have described several modes of interaction between LIN28 and primary, precursor, and mature forms of let-7 miRNAs (Desjardins et al., 2011; Nam et al., 2011; Rybak et al., 2008; Van Wynsberghe et al., 2011; Viswanathan et al., 2008; Zisoulis et al., 2012). In the context of a negative feedback loop, mature let-7 miRNAs have also been shown to repress LIN28 protein expression (Reinhart et al., 2000; Rybak et al., 2008).

Thus far, the regulation of let-7 miRNAs is the best-studied mechanism by which LIN28 controls gene regulatory networks. Reactivation of LIN28 in cancerous tissues has been proposed to cause downregulation of let-7 and subsequent activation of oncogenes such as *K-RAS*, *C-MYC*, and *HMGA2* (Bussing et al., 2008). Similarly, LIN28 expression can convey resistance to diet-induced diabetes by releasing let-7 repression of insulin-PI3K-mTOR pathway genes *IGF1R*, *INSR*, and *IRS2* (Zhu et al., 2011). However, changes in LIN28 expression have also been shown to have phenotypic consequences independent of altered let-7 levels. For example, transgenic mice with muscle-specific deletion of LIN28 exhibited impaired glucose uptake and insulin sensitivity, despite unchanged let-7 levels (Zhu et al., 2011). Other transgenic mice aberrantly expressing LIN28 show phenotypes of greater organ mass even in adult tissues where let-7 was unaffected (Zhu et al., 2010). Furthermore, during neurogliogenesis, constitutive expression of LIN28 has been shown to favor differentiation towards the neural lineage at the expense of glial cell development, prior to any influence on let-7 levels (Balzer et al., 2010). In ES cells, LIN28 has a positive influence on proliferation, in part by binding to and increasing the translation of mRNAs encoding cell-cycle regulators (Peng et al., 2011; Xu et al., 2009). These findings strongly suggest that regulation of other RNA transcripts, beyond let-7 miRNAs, is an equally important function of this protein. Until now, the lack of precise genome-wide LIN28 binding sites in RNA targets has represented a significant hurdle in our understanding of its regulatory network of target genes.

To generate a LIN28 protein-RNA interaction map, we used UV cross-linking and immunoprecipitation followed by high-throughput sequencing (CLIP-seq), which resulted in the discovery of LIN28 binding sites in over 6,000 gene targets. These sites were recapitulated in human ES (hES) cells and in a somatic cell line stably expressing LIN28. The resolution afforded by CLIP-seq enabled us to discover a GGAGA motif enriched in LIN28 binding sites within mRNA sequences. This motif occurs preferentially within predicted hairpins and other unpaired loop structures, similar to its context within pre-let-7. Among its mRNA targets, we find that LIN28 preferentially binds to transcripts encoding RNA processing and splicing factors. In fact, we demonstrate that exogenous expression of LIN28 in somatic cells, independent of altered let-7 miRNA levels, enhances the translation of a subset of RBPs that are known to regulate alternative splicing, namely hnRNP F, TIA-1, FUS/TLS and TDP-43. We showed that binding sites within these mRNAs were sufficient to enhance the activity of reporter constructs. Alternative inclusion of LIN28 binding sites within TDP-43 mRNA also revealed an interesting coupling between alternative splicing and translation control of this transcript. As a consequence of this direct regulation of splicing factors, LIN28 expression in somatic cells results in widespread alteration of splicing patterns. Depletion of LIN28 and LIN28B in hES cells also resulted in protein level changes of splicing factors. Furthermore, LIN28 and LIN28B exerted different effects on their targets in hES cells, hinting at further complexity in target regulation.

RESULTS

LIN28 binding sites found within thousands of human genes

In hES cells where LIN28 is expressed at high levels, we performed CLIP-seq with an antibody that specifically recognizes the endogenous protein (Figure 3A and Figure 4A). To model the reactivation of LIN28 expression observed in many cancer cells, we generated a stable Flp-In HEK293 cell line that constitutively expresses a C-terminal V5-tagged human LIN28 protein at physiological levels, but 5-6 fold below that of endogenous LIN28 in hES cells (LIN28-V5 293 cells; Figure 4B). We performed CLIP-seq on these cells, in this case with a V5 antibody (Figure 3 and Figure 4C). LIN28-bound RNA fragments from transcripts expressed in hES and LIN28-V5 293 cells were represented by 4.8 and 2.8 million sequenced reads that mapped to non-repetitive regions of the human genome, respectively (Table 1). comparable to previously published CLIP-seq experiments performed with hES cells (Yeo et al., 2009).

High-confidence LIN28 clusters (binding sites) were defined using a published computational procedure (Polymenidou et al., 2011; Zisoulis et al., 2010). We found that 5,969 and 6,061 protein-coding genes in hES and LIN28-V5 293 cells contained at least one LIN28 cluster (Table 1). Despite differences in the variety and copy number of transcripts expressed between these two cell types, we found that over half (4,111) of the genes with at least one cluster in hES cells (69%) were also targets in the LIN28-V5 293 cells (68%) (Figure 3B). Thus, when expressed in somatic cells, LIN28 binds a significant portion of its mRNA targets that are naturally found in hES

cells. In comparison with the 1,259 mRNA transcripts previously identified as LIN28 targets in hES cells (Peng et al., 2011) using a RNA immunoprecipitation (RIP) technique (which suffers from the caveat that the absence of cross-linking allows re-association of RNAs and RBPs after cell lysis (Mili and Steitz, 2004)), an average of 67% of the previously identified targets were detected in our CLIP-seq experiments (Figure 4D and E). While 82% of the 273 highest ranked RIP targets (Peng et al., 2011) were identified in our CLIP-seq datasets, more than 85% of the transcripts we have identified are not previously described as LIN28 targets (Figure 4F and G).

LIN28 binding sites are enriched within exons and 3′ untranslated regions of mRNAs

LIN28 was observed to bind in multiple locations within mRNA transcripts in hES and LIN28-V5 293 cells. Each target gene had ~3.5 significant clusters, approximately 35 nucleotides or less in length, totaling 26,279 hES and 15,028 LIN28-V5 293 binding sites. Within mRNAs that were expressed in both cell types, 26% of LIN28 hES clusters overlapped with a cluster identified in LIN28-V5 293 cells by at least 1 nucleotide (Figure 3B), comprising 47% of LIN28-V5 clusters. This was 4.3-fold higher than expected (6%) when LIN28 hES clusters were compared to randomly located clusters within the same genic regions (Figure 3B; $p < 10^{-4}$, hypergeometric test). To illustrate the concordance of LIN28 binding sites in hES and LIN28-V5 293 cells, clusters from both CLIP-seq experiments were found in overlapping positions within the 3′ untranslated region (3′UTR) of the gene encoding the heterogeneous nuclear ribonucleoparticle protein F (*hnRNP F*) (Figure 3C). As a

testament to the specificity of LIN28 binding, reads from a CLIP-seq experiment for the splicing factor RBFOX2 in hES cells (Yeo et al., 2009) were sparse in this region (Figure 3C). Indeed, only 4% of all LIN28 and RBFOX2 clusters in hES gene targets overlapped (Figure 3B).

We observed significant enrichment of LIN28 binding within coding exons and 3′UTRs, compared to the expected percentage of these regions in the transcriptome (Figure 3D). Less than 7% of LIN28 CLIP-seq clusters were found within intronic regions, indicating that LIN28 largely interacts with sequences within mature mRNA transcripts, consistent with the dominant localization of LIN28 protein in the cytoplasm (Balzer and Moss, 2007). In addition, LIN28 binding sites were found uniformly distributed across exons and 3′UTRs (Figure 5). The concordance between our hES and LIN28-V5 293 datasets suggests that when aberrantly expressed, LIN28 interacts with similar loci within mRNAs as it does in transcripts expressed in ES cells.

CLIP-seq confirms binding of LIN28 to pre-miRNAs

We identified 32 and 56 pre-miRNAs in hES and LIN28-V5 293 cells that featured LIN28 CLIP-seq reads, 15 of which were common between the two cell types (Table 2). Of the 17 pre-miRNA targets unique to hES cells, the majority were miRNAs that are more abundant in hES relative to LIN28-V5 293 cells. Similarly, more than half of the pre-miRNA targets in LIN28-V5 293 cells were more highly expressed in these cells, as compared to hES cells (Table 2). This suggests that LIN28

target specificity depends in part upon differences in cell type specific expression levels of miRNAs. Consistent with previous publications, we found evidence of LIN28 binding within all let-7 family members, such as let-7a-1, let-7f, let-7g, let-7i and miR-98 pre-miRNAs (Figure 6A, B and Table 2) (Hagan et al., 2009; Heo et al., 2008; Piskounova et al., 2008). CLIP-seq reads centered on the let-7 precursor loop fall precisely within the reported LIN28 interaction site at a GGAGA motif (Figure 6A and 2B) (Heo et al., 2009; Nam et al., 2011). To minimize the contribution of let-7 regulation in our study, we have selected a LIN28-V5 293 cell line with expression of LIN28 that did not alter the levels of highly abundant mature let-7a (Figure 4B, Figure 6C and D, and Figure 7), as confirmed both by Northern blot analysis (Figure 6D) and by deep sequencing the small RNA fraction of these cells (Figure 6E). Nevertheless, LIN28 targets let-7f, let-7g, let-7i and miR-98, which are expressed ~10-100 fold lower than let-7a, were reduced in the presence of LIN28-V5 expression (Figure 6E). CLIP-seq also identified other LIN28-interacting miRNAs, such as miR-302 family members (Figure 6F and Table 2), consistent with a previous report (Balzer et al., 2010). Of these many LIN28-interacting miRNAs, only the levels of let-7 family members appear to be directly affected by LIN28 binding in this system.

LIN28 binds mRNA sequences at GGAGA(U) motifs

The resolution of binding sites identified by CLIP-seq was exploited to identify motifs that characterize the interaction of LIN28 with mRNA sequences. The pentamer with the strongest statistical enrichment in LIN28 binding sites from both hES and LIN28-V5 293 cells was GGAGA ($p < 10^{-4}$, Z-score analysis) (Figure 8A).

Despite occurring two-fold higher than control clusters, this exact pentamer was neither necessary nor sufficient for LIN28 interaction, as only 13% (or 8%) of LIN28 hES (or LIN28-V5 293) clusters contained the sequence GGAGA. Nevertheless, this sequence element is enriched even in binding sites within lowly expressed transcripts, showing that we have captured LIN28 interaction with genes expressed across a wide spectrum of levels (Figure 9). HOMER, a *de novo* differential motif discovery algorithm (Heinz et al., 2010) confirmed statistically significant enrichment for degenerate GGAGA (LIN28 hES) and GGAGAU (LIN28-V5) motifs (Figure 8B; $P < 10^{-46}$). These motifs were prominently located at the center of LIN28 clusters in hES and LIN28-V5 293 cells in both coding exons and also within 3′UTRs (Figure 8C and D, and Figure 10A and B), confirming that this signal is not attributed to nucleotide biases within coding regions.

Although the sequence GGAG has been reported as the functional binding site of LIN28 in the terminal loop of let-7 miRNAs (Heo et al., 2009), we observed that the full sequence GGAGAU is conserved across let-7 pre-miRNA family members at this location. Crystal structures of mouse Lin28 in complex with let-7 pre-miRNAs confirmed that the zinc-finger knuckles of LIN28 interact with this GGAG motif (Nam et al., 2011), and also provided evidence that the CSD binds another discrete structural element within the precursor terminal loop containing the consensus motif NGNGAYNN (Y = pyrimidine; N = any base), which constitutes the expanded sequence GGAGAU that we have identified. Thus, we conclude that LIN28 interacts with a consensus GGAGA(U) motif within miRNA, as well as mRNA, sequences.

LIN28 shows a preference for unpaired mRNA regions of secondary structure

Since LIN28-miRNA interactions occur in the context of RNA secondary structures, we hypothesized that LIN28 might also interact with its motifs within a structural context in mRNA transcripts. As previously performed for a range of RBPs (Kazan et al., 2010; Li et al., 2010; Zisoulis et al., 2010), we applied the algorithm RNAplfold (Bernhart et al., 2006) to analyze LIN28-bound mRNA regions for structural features. Using RNA folding simulations, we calculated the likelihood for each position in two variants of our consensus motif, GGAG or GNGAY, to base-pair within stretches of ~200 nucleotides. These calculations enabled us to assign a probability that the motif frequently occurs in a hairpin, external, internal, or multi-loop, or is base-paired. Our results indicate a significant preference for GGAG and GNGAY motifs within LIN28 clusters to reside in hairpin and other loop structures both in exons and in 3′UTRs relative to instances of these motifs in control clusters (Figure 8E and F, Figure 10C and D, and Table 3). We also concluded that GGAG and GNGAY motifs within LIN28 binding sites are less frequently base-paired (Figure 8E and F, Figure 10C and D, and Table 3). While complex structures with an ‘A’ bulge in a handful of genes have been suggested to interact with LIN28 (Lei et al., 2011), our results demonstrate that LIN28 preferentially interacts directly with mRNA transcripts at GGAGA(U) sequence motifs within regions of unpaired secondary structure.

LIN28 binds to its own mRNA as a mode of autoregulation

CLIP-seq in hES cells provided evidence that LIN28 binds within its own mRNA, primarily in its 3′UTR where there were 13 significant clusters, the majority of which harbored GGAGA motifs (Figure 11A). Previous studies have suggested that LIN28 may bind to its own mRNA; however, experimental support was not presented (Polesskaya et al., 2007). We confirmed this interaction by RIP analysis of LIN28 in the HUES6 hES cell line (to complement our independent CLIP-seq experiments using the H9 line) (Figure 11B). Quantitative RT-PCR using primers recognizing the 3′UTR of the endogenous *LIN28* mRNA showed a three-fold increase in steady state mRNA in LIN28-V5 293 cells compared to control Flp-In-293 cells Figure 11C).

To evaluate if the LIN28-bound sequence within the *LIN28* 3′UTR is sufficient to enhance expression levels in a heterologous context, the region containing the highest density of LIN28 clusters (Figure 11A, “Cloned Region”) was inserted downstream of a luciferase reporter. Co-transfection of the reporter with a plasmid expressing LIN28-GFP demonstrated that this region of the *LIN28* 3′UTR is sufficient to enhance luciferase activity, whereas transfection of a control plasmid had no effect (Figure 11D). As it is thought that LIN28 can be regulated by let-7, we noted that the increased luciferase activity might be due to a relief from repression by let-7. However, neither deletion nor mutation of the let-7 complementary site (as performed by Mayr and colleagues (Mayr et al., 2007)) within the *LIN28* 3′UTR reporter construct increased luciferase levels beyond those observed from LIN28-GFP overexpression. Therefore, we conclude that LIN28 directly enhances its own

expression level by binding to sites within its 3′UTR, revealing a mechanism of positive feed-forward regulation by LIN28. The transcription factors OCT4, SOX2 and NANOG, which are required for propagation of undifferentiated ES cells and are important for reprogramming, also collaborate to autoregulate themselves in feed-forward loops (Boyer et al., 2005). Our results suggest that LIN28 exhibits the same ability to affect its own protein levels.

LIN28 directly regulates the protein levels of RNA binding proteins

To explore potential pathways affected by LIN28, Gene Ontology (GO) analysis identified "regulation of RNA metabolic processes" (1386 target genes), "RNA splicing" (234), and "RNA localization" (87) as statistically significant RNA-related categories enriched among LIN28 target genes, as well as categories consistent with its known roles in cellular proliferation and neurogenesis (Figure 12A). To specifically address if RBPs were enriched as LIN28 targets, we analyzed a compiled set of 443 RBPs for the presence of LIN28 clusters (Huelga et al., 2012). Out of these RBPs, 248 (56%) and 236 (53%) were found to be direct targets of LIN28 in hES and LIN28-V5 293 cells, respectively, ($p < 10^{-4}$, hypergeometric test).

To establish if direct LIN28 targets, such as genes encoding RBPs, were regulated by LIN28 at the RNA level, we conducted triplicate microarray gene expression analysis of LIN28-V5 293 and control Flp-In-293 cells (Figure 13A). Our results indicated that genes with altered expression levels were not enriched for binding relative to unchanged genes (at a $p < 0.01$ cutoff, chi-square test; Figure 13B),

suggesting that direct targets of LIN28 are neither frequently nor significantly affected at the steady-state mRNA level when LIN28 is expressed. This result was also recapitulated with deep sequencing of cDNAs (RNA-seq) from hES cells transduced with lentivirus encoding an shRNA targeting LIN28 (Figure 13B, C, D and E).

To determine if LIN28 targets were instead controlled at the level of translation, we first evaluated the protein level of cyclin B1. We observed higher levels of cyclin B1 in the LIN28-V5 293 compared to control Flp-In-293 cells (Figure 12B and Figure 13F), consistent with published results indicating that murine cyclin B1 decreases upon LIN28 depletion in mouse ES cells (Xue et al., 2009). Next we selected a number of LIN28 targets, focusing on RBPs which have published roles in regulating splicing, including *FUS/TLS*, *hnRNP F*, *TDP-43* and *TIA-1*. These genes all increased by at least two-fold at the protein level in LIN28-V5 cells compared to control cells, but were unaltered at the mRNA level (as measured by the microarrays) (Figure 12B).

Since higher levels of LIN28 reduced let-7f expression (Figure 6E), we introduced let-7f mimics (artificial mature miRNA duplexes) that were insensitive to LIN28 regulation into LIN28-V5 293 cells to determine if the levels of these RBPs were higher due to lack of let-7f. Compared to a control mimic, the protein levels of IMP2, a known let-7 target (Yun et al., 2011) was effectively downregulated in the presence of the let-7f mimic (Figure 12C). We also noted that the paralog LIN28B protein was downregulated upon increased let-7f expression, suggesting that LIN28B is likely regulated by let-7f (Guo et al., 2006). Importantly, FUS/TLS, hnRNP F, TDP-

43, and TIA-1 were unaffected in their protein levels by let-7f expression (Figure 12C and Figure 13G), supporting the conclusion that these RBPs are directly regulated by LIN28-mRNA interactions, and not through let-7f.

Next we set out to determine whether specific LIN28-bound regions of target genes were sufficient to convey LIN28-dependent translational regulation. We cloned mRNA regions from *hnRNP F* (coding region; Figure 14A) and *FUS/TLS* (coding region and 3′UTR; Figure 14B) that contained LIN28 binding sites in both hES and LIN28-V5 293 cells downstream of a luciferase reporter. Consistent with our western blot results (Figure 12B), co-expression of LIN28-GFP, but not a control plasmid, significantly enhanced luciferase activity ($p < 0.001$, Figure 12D), confirming that LIN28 binding sites are sufficient to increase translational regulation of *hnRNP F* and *FUS/TLS*.

Within the 3′UTR of *TDP-43,* we observed LIN28 binding sites overlapping with purine-rich (GGAGA) motifs in a retained intronic region (Figure 12E). This region was previously reported to be bound and spliced by TDP-43 itself, thereby eliciting nonsense-mediated decay (NMD) to reduce its mRNA levels (Polymenidou et al., 2011). We hypothesized that when this 3′UTR-embedded intron remains unspliced and the TDP-43 mRNA is exported to the cytoplasm, the LIN28 protein could interact with binding sites in the 3′UTR to enhance translation of the mRNA. However, a spliced TDP-43 3′UTR would not contain LIN28 binding sites, and thus would not be affected by LIN28 expression. To test this hypothesis, we utilized two reporter constructs containing different arrangements of the homologous mouse *TDP-43*

3′UTR downstream of a luciferase open reading frame (Polymenidou et al., 2011) (Figure 12E). The first reporter, referred to as "short," contained the spliced 3′UTR, which removed the majority of LIN28 binding sites. The second reporter, referred to as "long", harbors an unspliced region of the *TDP-43* 3′UTR homologous to the human region containing LIN28 binding sites. Co-transfection of these reporter constructs demonstrated that the reporter containing the LIN28 binding sites was significantly enhanced at the translational level when LIN28-GFP was overexpressed; however, the spliced "short" construct was not (Figure 12F). Deletion of one of the four LIN28 GGAGA binding motifs within the "long" reporter reduced its translational output by ~15% in the presence of LIN28-GFP expression, suggesting that site-specific interactions of LIN28 contributes to its ability to enhance translation (Figure 14C). We conclude that LIN28 regulates TDP-43 protein levels by interacting with specific binding sites within a retained intron in the *TDP-43* 3′UTR. Importantly, if this intron is spliced these binding sites are not available for control of protein levels, offering an interesting example of a coupling between the regulation of splicing and translation.

Increased levels of LIN28 in somatic cells causes widespread changes in alternative splicing

If LIN28 regulates the translation of many splicing factors, we expect that LIN28 misregulation will result in changes in alternative splicing (AS). To test this, we subjected total RNA from LIN28-V5 293 cells and control Flp-In-293 cells to

splicing-sensitive microarray (HJAY) analysis. We identified 1,985 differentially regulated AS events in the presence of LIN28 expression, out of 14,643 events detected on the array (Figure 15A). These events are comprised of isoform changes in approximately 1,965 genes. This number of AS events is comparable to the numbers regulated by well-studied splicing factors such as hnRNP proteins, RBFOX2 and HuR (Huelga et al., 2012; Mukherjee et al., 2011; Venables et al., 2009). Since we found little evidence of LIN28 binding to intronic regions (Figure 3D), we reasoned that LIN28 likely interacts with cytoplasmic, mature mRNA transcripts, which suggests that the observed AS events are most likely the downstream result of LIN28 regulation of splicing factors. We successfully validated a number of these AS changes by semi-quantitative RT-PCR with an 85% validation rate (Figure 15B and Figure 16). As an interesting example, we validated the alternative splicing of a 63 nucleotide (nt) cassette exon 23a in the neurofibrimin 1 (NF1) gene, which is skipped upon expression of LIN28-V5 (Figure 15B). As a known negative regulator of the *Ras* signaling pathway, accurate control of NF1 isoforms are important in cancer and neuronal differentiation (Patrakitkomjorn et al., 2008), thereby providing a glimpse into signaling pathways that LIN28 may affect through regulation of AS.

To analyze the extent of alternative splicing events affected due to the regulation of a single splicing factor by LIN28, we overexpressed a plasmid harboring the open reading frame of TDP-43 fused to a C-terminal GFP in Flp-In-293 cells, reproducing the upregulation of TDP-43 upon LIN28 expression observed in LIN28-V5 293 cells (Figure 15C). We subjected total RNA to splicing-sensitive microarray

analysis, identifying a total of 865 AS events that changed, including 526 differentially spliced cassette exons (Figure 15A and D). Of the cassette exons affected by stable LIN28-V5 expression in our cell line, we identified a significantly overlapping subset of 113 cassettes (13%) that were also affected upon upregulation of TDP-43 ($p < 10^{-5}$, hypergeometric test), with 70% of the cassette events changing in the same direction (Figure 15D). Of the hundreds of splicing factors that LIN28 is predicted to regulate, LIN28 affects a statistically significant overlapping set of alternative splicing events with at least one splicing factor, TDP-43.

Decreased levels of LIN28 and LIN28B in embryonic stem cells modulates translation of RNA binding proteins

We were surprised to find that depletion of LIN28 in hES cells resulted in less than half of the number of AS events as in LIN28-V5 293 cells, and that few of these events were reciprocal (Figure 17A and Figure 18A). In addition, despite the high concordance between hES and LIN28-V5 293 cells of the location of LIN28 binding sites on target mRNAs, its splicing factor targets did not display a decrease in protein levels expected upon knockdown of LIN28 (Figure 18C). Given that LIN28B, the paralog of LIN28, was significantly enhanced when LIN28 was depleted in hES cells (Figure 18B) and that LIN28 and LIN28B interact with a common set of mRNAs encoding splicing factors (Figure 17B), we hypothesized the LIN28B may compensate for loss of LIN28. To address this relation between LIN28 and LIN28B, we electroporated hES cells with siRNAs that individually depleted LIN28 and LIN28B, as well as both proteins simultaneously (Figure 18C). Interestingly, we observed that

hnRNP F increases at the protein level with depletion of LIN28B, TDP-43 is downregulated when either LIN28 or LIN28B was depleted but not further downregulated by depletion of both, and FUS/TLS was reduced only when both LIN28 and LIN28B were concurrently depleted. Therefore, LIN28 and LIN28B may exhibit synergistic (FUS/TLS), and both repressive (hnRNP F) and enhancing (TDP-43, FUS/TLS) effects on translation of their mRNA targets in stem cells. Our observations that LIN28 and LIN28B have differing effects on their targets, and that LIN28 levels affect LIN28B expression (Figure 18B), reveal another layer of complexity ripe for future investigation. These studies will be important to address the extent of this functional overlap between LIN28 and LIN28B, and to identify co-factor complexes that underlie differences in cell type and gene-specific regulation by these proteins.

DISCUSSION

Systematic, genome-wide identification of thousands of LIN28 binding sites revealed that more than 6,000 genes are targets of LIN28 in hES cells and in somatic cells where LIN28 was exogenously introduced. We report the identification of a GGAGA(U) motif within LIN28 mRNA binding sites which resembles the sequence and structural context of the interaction with let-7 miRNA precursors. We also provide evidence of LIN28 autoregulation by direct binding to its own mRNA. Independent of prerequisite alteration of let-7 levels, we find that LIN28 binds to mRNA regions within transcripts that code for splicing factors, including *TDP-43, FUS/TLS, TIA-1,* and *hnRNP F* and controls their protein abundance. Upregulation of protein levels of

these targets in response to an increase in LIN28 in somatic cells leads to widespread changes in alternative splicing patterns. Surprisingly, downregulation of LIN28 in hES cells does not always result in reciprocal changes for these RBPs. Furthermore, LIN28B does not in general compensate for lack of LIN28 function, despite also interacting with mRNAs encoding these RBPs, and has different, or sometimes synergistic, effects on these targets. This cell type specific control of gene regulatory targets by LIN28 presents an alternative mechanism through which LIN28 and LIN28B expression can shape cell fate and homeostasis.

Aside from alternative splicing, the RBP targets of LIN28 are also involved in other RNA processing steps, expanding the breadth of known effects of LIN28 on gene regulation. Both TDP-43 and FUS/TLS regulate mRNA transport, translation, turnover and miRNA processing, and disruption of either protein leads to amyotrophic lateral sclerosis (Lagier-Tourenne et al., 2010). TIA-1 is a central player in the formation of stress granules, which safeguards selected mRNAs by controlling their translation and stability during cellular stress (Kedersha and Anderson, 2002). Our finding that LIN28 regulates TIA-1 expression provides another link between LIN28 and RNA regulation through control of stress granule formation (Balzer and Moss, 2007). HnRNP F protein, as well as the structurally similar hnRNP H1 protein, has been observed to co-immunoprecipitate with LIN28 (Polesskaya et al., 2007). Of note, hnRNP F and H1 (Caputi and Zahler, 2001) are known to recognize GGGA sequences in RNA. With our finding that LIN28 also binds GNGAY motifs, it is possible that these hnRNP proteins and LIN28 regulate a common set of targets. To summarize, our

genome-wide study reveals avenues by which LIN28 impacts gene regulatory networks through direct regulation of its mRNA targets, and provides a valuable framework for future characterization of the molecular roles of LIN28 and LIN28B in biological pathways.

METHODS

Cell culture and stable cell line generation

HEK293 cells containing an integrated Flp-In site (Flp-In-293; Life Technologies) were cultured in DMEM media supplemented with 10% FBS and 2mM L-glutamine and passaged using TrypLE (Life Technologies). The LIN28 open reading frame (*Homo sapiens*, GenBank: DQ896719) was cloned from a Gateway pENTR221 vector (Open Biosystems) into the Gateway pEF5/FRT/V5 destination vector (Life Technologies) to generate the V5-tagged LIN28. To generate LIN28-V5 293 stable cell lines, pEF5/FRT/LIN28-V5 plasmid was co-transfected along with the FLP Recombinase expressing plasmid pOG44 into Flp-In-293 cells using Lipofectamine-2000 (Life Technologies) according to the manufacturer's instructions. Recombination and insertion of LIN28-V5 at the Flp-In site confers stable integration and expression of the LIN28-V5 fusion protein and hygromycin resistance. Stably transected clones were selected and propagated in media supplemented with 75-100 μg/ml hygromycin B (Life Technologies) and several independent clonal cell lines were established. Cell line LIN28-V5 #6 expressed an intermediate level of LIN28-V5 expression and was used for all LIN28-V5 experiments unless otherwise indicated.

Human ES cell lines H9 and HUES6 were grown in feeder-free conditions with mTeSR media (STEMCELL Technologies) and on Matrigel (BD Biosciences) for support. Cells were passaged manually or with Dispase (BD Biosciences) every 5-7 days.

RNA IP experiments

For RNA immunoprecipitation (RIP), HUES6 or Flp-In-293 cells were lysed in 1X RIPA buffer (10X Millipore RIPA: 0.5M Tris-HCl, pH 7.4, 1.5M NaCl, 2.5% deoxycholic acid, 10% NP-40, 10mM EDTA) with 1X Roche Complete Protease Inhibitors, EDTA Free. RIP was performed as previously described (Van Wynsberghe et al., 2011). Briefly, lysates were pre-cleared with Protein G Dynabeads (Life Technologies) and then immunoprecipitation was performed for 4h with antibodies against LIN28 (Abcam ab46020), LIN28B (Cell Signaling 4196) or IgG (Caltag Laboratories 10500C), using beads pre-bound with either antibody. Immunprecipitated material was treated with Proteinase K (Life Technologies), and RNA was extracted with TRIzol (Life Technologies) and subsequently treated with RQ1 DNase (Promega).

Northern blot analysis

Northern blots were performed as described (Stark et al., 2011). The let-7a miRNA was detected using the StarFire probe system from Integrated DNA Technologies (sequence AACTATACAACCTACTACCTCA).

Western blot analysis

Cells were rinsed twice with cold 1X PBS on ice, and then suspended in 1X RIPA buffer (Millipore) supplemented with 1X Roche Complete Protease Inhibitors, EDTA Free. Cells were lysed by repeated pipetting, followed with sonication for 5-7 min in a Diagenode BioRuptor. Lysates were then centrifuged at 14,000 RPM for 20 min at 4˚C. Protein concentrations of the resulting supernatants were quantified using the BCA quantification assay (Thermo Pierce). For Western analysis, 15-25 μg of protein lysate was separated on 10% or 4-12% Bis-Tris gels using the NuPAGE system (Life Technologies) and transferred to Hybond-P membrane (Amersham Biosciences). Membrane incubations with anti-GAPDH (1:10,000; Abcam ab8245), anti-LIN28 (1:5,000; Abcam ab46020), anti-LIN28B (1:1,000; Cell Signaling 4196), anti-FUS/TLS (1:1,000; Santa Cruz Biotechnologies SC-47711), anti-TDP43 (1:1,000; Aviva ARP35837_P050), anti-Tia1 (1:200; Santa Cruz Biotechnologies SC-1751), anti-cyclin B1 (1:200; Abcam ab72), anti-hnRNP F (1:200; Santa Cruz Biotechnologies SC-10045), and IMP2 (1:1,000; MBL RN008P) were performed overnight. Secondary antibodies were used at 1:10,000 (anti-rabbit Calbiochem 401393 or Cell Signaling 7074, anti-mouse Cell Signaling 7076, anti-Goat Promega V-4771) and chemiluminescence reagents (Thermo Pierce) according to manufacturers' recommendations.

RNA extractions, RT-PCR and qRT-PCR validation

All RNA extractions were performed with TRIzol reagent (Life Technologies) and DNase treated with Turbo DNase (Ambion) according to the manufacturers' instructions, unless otherwise specified. To generate cDNA, reverse transcription was performed using 1-2 μg of total RNA, oligo(dT) primer, random hexamers and Superscript III reverse transcriptase (Life Technologies) according to the manufacturer's instructions. PCR amplification was preformed for 32-36 cycles using template cDNA diluted 1:15 and products were run on a 1-2% agarose gel stained with SYBR safe (Life Technologies). Primers were designed against the LIN28 3′UTR to assay for immunoprecipitation (PCR) and measure expression level differences (qPCR). For qPCR of the LIN28 3′UTR (Figure 11C), RNA from three biological replicates of the cell line LIN28-V5 #4 (Figure 4B) were extracted and quantified in technical triplicates. For AS validations, primers were designed to span exons flanking alternative cassettes. All primer sequences will be made available upon request. The intensity of PCR products was quantified using ImageJ software (http://imagej.nih.gov/ij/). For quantitative RT-PCR, SYBR Green (Applied Biosystems) was used according to the manufacturer's recommendations on a 7900HT real time PCR machine with supporting software. An unpaired Student's t-test was used to determine significant changes in relative mRNA expression level between experimental and control conditions after normalization to GAPDH.

Luciferase assays

The spliced (long) and unspliced retained intron (short) forms of the mouse *TDP-43* 3′UTR were previously cloned into the psiCHECK-2 vector (Promega) (Polymenidou et al., 2011). A TDP-43 3′UTR deletion mutant lacking one of four LIN28 binding sites was constructed using primers to generate the deletion PCR fragment and hybridize the PCR fragment into the psiCHECK-2 vector using a published approach (Heckman and Pease, 2007). A portion of the human 3′UTR of *LIN28* ('Cloned Region' Figure 11A), and mRNA sequences of *FUS/TLS* and *hnRNP F* (Figure 14Aand B) were amplified using cDNA derived from LIN28-V5 293 cells and with primers that added XhoI and NotI restriction sites. Amplified products were inserted into the XhoI and NotI sites of the psiCHECK-2 vector downstream of *Renilla* luciferase. To disrupt let-7 binding to the 3′UTR of *LIN28*, the let-7 seed region was removed (ΔLet-7) or mutated (Mut) within a psiCHECK-2 construct containing the cloned portion of *LIN28* 3'UTR and using a protocol described by Heakman and Pease, 2007. Transfection of Flp-In-293 cells in 12-well cell culture plates was performed using 400ng of reporter plasmid and 400ng of pcDNA3.1 (Life Technologies) or LIN28-GFP (Balzer and Moss, 2007) using Fugene 6 (Roche Applied Science) according to manufacturer's instructions. Cells were harvested 48 hours later and luciferase activities were assayed using the Dual-Luciferase Reporter system (Promega). Renilla activity was normalized to firefly activity, which is used as the internal control. Five individual luciferase assays were performed, and for each assay, transfections were performed in triplicates.

Lentiviral shRNA-mediated and siRNA-mediated knockdown of LIN28 and LIN28B

To achieve knockdown of LIN28, we utilized an shRNA construct targeting human LIN28 in the pLKO.1 vector (TRCN0000102579; Open Biosystems). As a control, a pLKO.1 vector containing an shRNA toward GFP was used (Open Biosystems). Lentivirus expressing these constructs was prepared in 293T cells, as previously described (Yeo et al., 2009). Human ES cells were treated with Accutase (STEMCELL Technologies) to generate single-cell suspensions, and infected with LIN28 or matched control GFP virus. The media was changed daily and cells were harvested 72 hours after infection.

To deplete LIN28 (LIN28A) and LIN28B, we utilized On-TARGET*plus* SMARTpool siRNAs from Dharmacon (LIN28A: L-018411-01-0005, LIN28B: L-028584-01-0005, On-TARGET*plus* Non-Targeting Pool: D-001810-10-05). For siRNA treatments in hES cells, cells were pretreated for 4 hours with Rock inhibitor (Millipore) and then dissociated into a single cell suspension with Accutase (STEMCELL Technologies) for 7 minutes. Cells were electroporated using the Amaxa system, Human Stem Cell Nucleofector Kit 2, with a total of 1.2 x 10^6 cells treated in each condition at a final siRNA concentration of 300nM. Experimental conditions consisted of 150 nM control non-targeting siRNA pool and 150 nM siRNA LIN28 or LIN28B pool, or 150 nM of siRNA LIN28 and 150 nM siRNA LIN28B, and control conditions were 300 nM of control non-targeting siRNAs. Media was changed daily and cells harvested 48 hours post-nucelofection.

TDP-43 overexpression

Flp-In-293 cells were grown to ~70% confluency in 6-well cell culture plates and transfected with 4μg of TDP-43-GFP (Liu-Yesucevitz et al., 2010) or control pEGFP-C2 (Clontech) plasmid using Lipofectamine-2000 (Life Technologies) according to manufacturer's instructions. After 48 hours cells were harvested, and RNA and protein lysates extracted and quantified as above.

Let-7f expression

Rescue of let-7f expression levels in LIN28-V5 293 cells was achieved via replicate transfections of cells freshly seeded at a density of 0.4-1.6 x 10^5 cells per well of a 24-well cell culture plate with a final concentration of 5 nM human let-7f mimic (miScript syn-hsa-let-7f; Qiagen MSY0000067) using Lipofectamine RNAiMax (Life Technologies) according to the manufacturers' recommendation. Control tranfections were similarly prepared via treatment with a control miRNA mimic with no known homology to any mammalian gene (AllStars Negative Control; Qiagen 1027280). Cells were harvested 48 hours post-transfection and protein lysates quantified as above.

RNA-Seq

We used one round of polyA selection with Oligo $(dT)_{25}$ Dynabeads (Life Technologies) according to manufacturer's recommendations to isolate mRNA from 3-9 μg total RNA. This mRNA was then fragmented with the Ambion RNA

Fragmentation kit (AM8740) according to the manufacturer's instructions. The fragmented RNA was phosphorylated with T4 polynucleokinase (PNK), and subjected to the small RNA-seq protocol v1.5 (Illumina) as per manufacturer's instruction and bands corresponding to 110-160bp fragments extracted for sequencing.

Small RNA-Seq

Small RNA libraries were generated from total RNA isolated from H9, untreated Flp-In-293, and LIN28-V5 293 cell line. H9 RNA was treated as described in Illumina's Small RNA Digital Gene Expression v1.5 protocol, using the supplied adapters. RNA extracted from Flp-In-293 and LIN28-V5 293 cells was treated as described in Illumina's Small RNA Digital Gene Expression v1.5 protocol, with the exception of custom oligonucleotide adapters and primers used which matched those of Illumina's TruSeq Small RNA Sample Preparation protocol. For all samples, 2µg total RNA was ligated with a 3' pre-adenylated adaptor using T4 RNA ligase 2, truncated (NEB), followed by ligation of a 5' adaptor using T4 RNA ligase 1 (NEB). First-strand cDNA synthesis was performed using Superscript II (Life Technologies). DNA was amplified using 13 cycles of PCR with Phusion polymerase (Finnzymes). DNA products (H9: 85-115bp; Flp-In-293 and LIN28-V5 293: 135-160bp) were purified using a 6% non-denaturing TBE acrylamide gel (Life Technologies). Flp-In-293 and LIN28-V5 293 libraries contained the Illumina Truseq barcode indices 6 and 5, respectively. DNA products were eluted by diffusion, filtered using Spin-X columns (Costar), and precipitated. DNA was quantified using Quant-It PicoGreen (Life

Technologies). If barcoded, libraries were mixed, and 17.5fmol of cDNA per flow cell lane was sequenced on the Illumina GAII for 36 cycles.

CLIP-seq data processing and cluster generation

Read mapping from CLIP-seq experiments and data processing was performed as published (Polymenidou et al., 2011). Briefly, reads were processed and mapped to the human genome (hg18 http://genome.ucsc.edu; Bowtie version 0.12.2, with parameters -q -p 4 -e 70 -y -l 25 -n 2 -m 5 --best --strata) and assigned to 21,605 genes (as annotated previously (Yeo et al., 2009)). Significant clusters of reads were calculated as previously described (Polymenidou et al., 2011) with a local cutoff that was determined using a gene-specific frequency calculated as the number of reads overlapping that gene, divided by the mRNA length ($P < 0.01$ Bonferroni correction). A smaller sliding window of 40bp was used to determine where the number of reads exceeded both the local (gene) and global (transcriptome) cutoffs. Sets of control clusters were generated by taking LIN28 CLIP-seq clusters and selecting a same sized sequence a random distance from the transcript start of the target gene, to control for differences in gene expression. Control clusters were also confined to the same genic regions as LIN28 CLIP-seq clusters. For all control datasets 10 iterations of randomly selected controls were generated. CLIP-seq generated sequenced reads were also mapped to the human genome build hg19 to be comparable with miRBase 16 miRNA annotations using Bowtie (version 0.12.2) with parameters -q -p 4 -e 70 -y -l 25 -n 2 -m 15 --best --strata. Raw sequence files and processed CLIP-seq cluster files can be

accessed at the Gene Expression Omnibus (GEO) under the "HTS" SubSeries of SuperSeries GSE39873.

RNA structure calculations

The prediction algorithm RNAplfold from the Vienna RNA Package was used to calculate the probability of secondary structure around LIN28 binding sites with the parameters -W 240 -L 160 -u 1. Input sequences were obtained from a database of mRNA annotations, as we observed the vast majority of LIN28 targets occurred within mature mRNA transcripts. The original source code was modified to obtain separate probabilities for each position of a motif to be in a hairpin loop, external loop, internal loop, or multi-loop structure or paired. P-values of rejecting the hypothesis that the two independent cumulative distributions were drawn from the same underlying continuous population were computed by the two-sample Kolgomorov-Smirnov test.

RNA-seq data processing and gene expression analysis

Strand-specific RNA-seq reads were mapped to our annotated gene structure database (Bowtie version 0.12.2, with parameters –q -e 70 –y –l 25 –n 2 –m 5 –k 5 --best --strata). We measured gene expression as the number of reads uniquely mapped to exons of a gene, per kilobase of exon sequence for that gene, normalized by the total number of million mapped reads to genes (RPKM). Differentially expressed genes were identified using a Z-score analysis as previously described with a cutoff of $Z < -2$ (downregulated) or $Z > 2$ (upregulated) (Polymenidou et al., 2011). Raw

sequence files and processed RPKM calculations can be accessed at GEO under the "HTS" SubSeries of SuperSeries GSE39873.

Small RNA-seq data processing and mature miRNA expression analysis

Small RNA reads were mapped to the human (hg19) genome using Bowtie short read aligner (Langmead et al., 2009) and associated with coordinates of known miRNAs from mirBase18 (Griffiths-Jones, 2004; Kozomara and Griffiths-Jones, 2011). We required exact 5' end alignment to the canonical miRNAs for read alignments. An expression value for each miRNA was calculated using the metric RPM; reads mapped to the mature miRNA normalized to the number of million mapped reads. Changes in miRNA expression were calculated by Z-score analyses of the $\log_2$ fold change (RPM LIN28-V5 over RPM Flp-In-293 cells) for all miRNAs with an RPM >= 1. Mature miRNAs with an absolute Z-score >= 2 and an RPM > 1 in both cell types were considered significantly changed. Raw sequence files and processed RPM calculations can be accessed at GEO under the "HTS" SubSeries of SuperSeries GSE39873.

Splicing array analysis for splicing and RNA expression changes

Microarray data analysis for LIN28-V5 293 cells, untreated Flp-In-293 cells, TDP-43 overexpression in Flp-In-293 cells and H9 hES cells with control or LIN28 knockdown conditions were performed using a previously described method (Sugnet et al., 2006), with cutoff of q-value < 0.05 and an absolute separation score > 0.5 to identify alternative splicing events. Comparisons of LIN28 dependent AS events and

TDP-43 dependent AS events were performed using in-house perl scripts. Expression changes were determined by a 2-fold or greater fold change in normalized probe values as previously described (Sugnet et al., 2006). Raw array CEL files and normalized probe intensities for each experiment can be accessed at GEO under the "splicing array" SubSeries of SuperSeries GSE39873.

Motif analysis

Motif analysis was performed as previously described (Yeo et al., 2009) using LIN28 clusters and the randomly distributed set of control clusters counting all possible pentamers. De novo motif finding was also applied using the HOMER v3.4 differential motif discovery algorithm (Heinz et al., 2010). This algorithm was used to search 50bp from the center of all clusters for the most enriched motifs in real versus control clusters. The HOMER script findMotifsGenome.pl was run with the following command and parameters: findMotifsGenome.pl hg18r -len 3,4,5,6,7,8,9,10 -norevopp.

Gene ontology analysis

The **D**atabase for **A**nnotation, **V**isualization and **I**ntegrated **D**iscovery (DAVID Bioinformatic Resources 6.7; http://david.abcc.ncifcrf.gov/) was used to generate gene ontology associations and assign functional categories to genes. The set of transcripts expressed in H9 hES cells, as detected by microarray, were used as background.

Accession numbers

All Illumina sequencing and splicing array data is accessible through the Gene Expression Omnibus (GEO) accession number GSE39873.

AUTHORS' CONTRIBUTIONS

Chapter 2, in full, is an adaptation of material that appears in "LIN28 binds messenger RNAs at GGAGA motifs and regulates splicing factor abundance to affect alternative splicing" by Melissa L. Wilbert, Stephanie C. Huelga, Katannya Kapeli, Thomas J. Stark, Tiffany Y. Liang, Stella X. Chen, Bernice Y. Yan, Jason L. Nathanson, Kasey R. Hutt, Michael T. Lovci, Hilal Kazan, Anthony Q. Vu, Katlin B. Massirer, Quaid Morris, Shawn Hoon and Gene W. Yeo, as published in *Molecular Cell* October 2012. SCH assisted in splicing analysis and manuscript preparation. TJS performed northern blot experiments. CLIP-seq libraries were generated by T.Y.L. Support for computational analyses were provided by K.R.H. and MTL. Luciferase constructs were generated by KK and TJS; in vitro assay using these constructs were performed by KK and MLW. SXC assisted with HEK293 culture and RNA extractions. Other tissue culture and in vitro manipulations (knockdown, overexpression etc.) were performed by MLW. Lentivirus expressing shRNAs were generated by AQV and MLW. Splicing validations were performed by MLW, BYY, and SXC. Small RNA-seq libraries were generated by JLN and analyzed by MLW and JLN. RNA-seq libraries were generated by MLW. Sequencing on the Illumina GA and GAII analyzer were performed by SH. RIP of LIN28 in HUES6 cells was performed

by KBM and PCR analysis performed by MLW. Structural predictions were run by HK using experimental and control datasets generated by MLW. All other analyses were performed by MLW. Manuscript preparation and design of experiments were assisted by GWY.

FIGURES

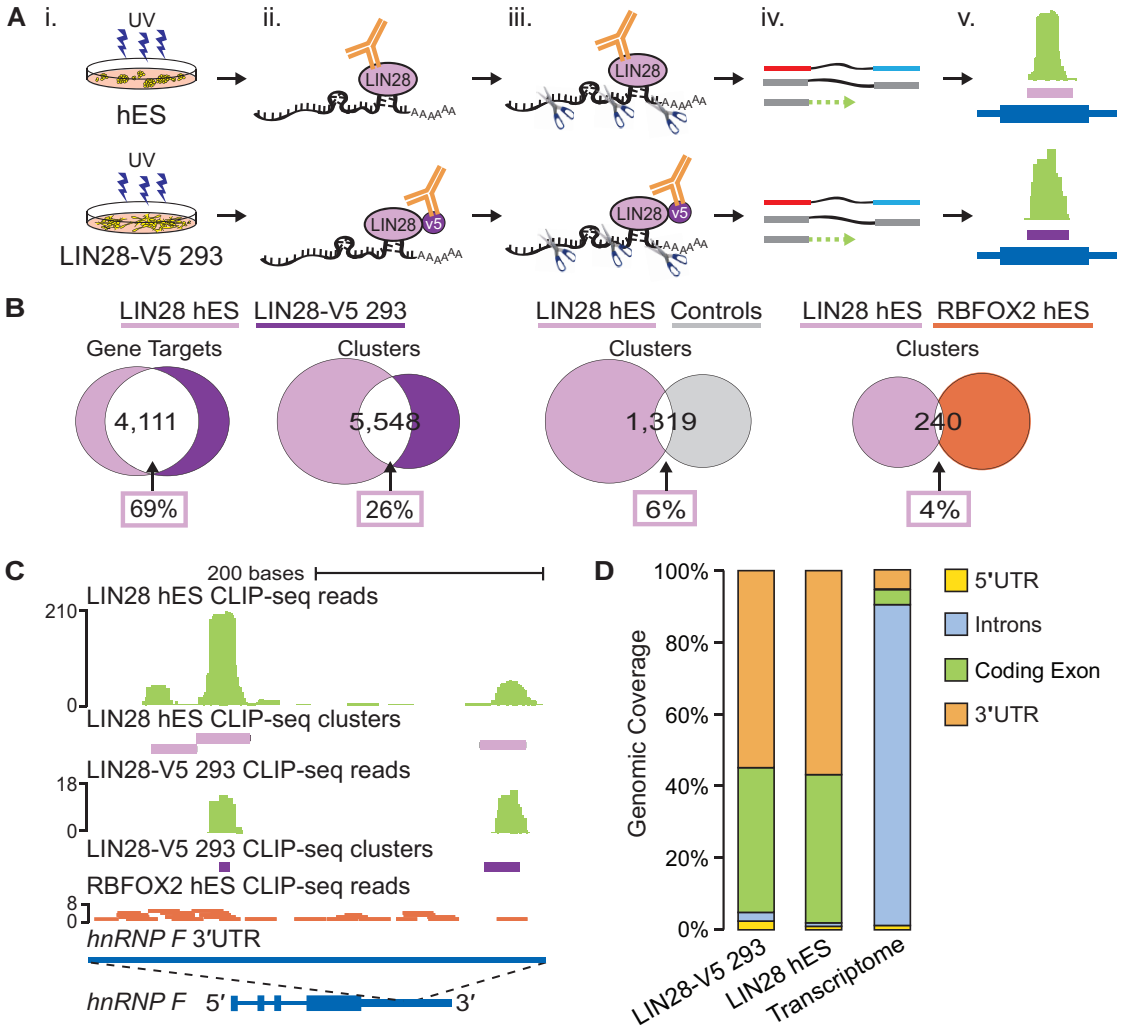


Figure 3. CLIP-seq identifies LIN28 binding sites in thousands of human genes.
(A) CLIP-seq experimental approach performed using H9 human ES (hES) and LIN28-V5 293 cells. (i) UV cross-linking; (ii) immunoprecipitation of LIN28 protein-RNA complexes; (iii) micrococcal nuclease treatment, SDS PAGE gel size selection, and protease digestion; (iv) cDNA library preparation and high-throughput sequencing; and (v) cluster identification (in purple) based on the density of reads (in green) mapped to genes (in dark blue). (B) Venn diagrams illustrating the number of LIN28 target genes and clusters in common between hES and LIN28-V5 293 cells. For comparison, randomly located clusters in the same genes and genic regions as LIN28-V5 293 clusters and clusters from RBFOX2 in hES cells were used. The percentage of LIN28 hES gene targets or clusters in common with each comparison dataset are indicated within boxes. (C) LIN28 binding sites identified within the 3′UTR of the hnRNP F gene in both hES and LIN28-V5 293 cells. Clusters are depicted by purple rectangles representing the highest density of CLIP-seq reads (graphed as continuous densities in green). Individual RBFOX2 hES CLIP-seq reads are shown in red for comparison. The scale to the left indicates the height of aligned reads. (D) LIN28 binding enrichment in coding exons and 3′UTR sequences in both hES and LIN28-V5 293 cells, as compared to the observed percentage of nucleotides in the annotated transcriptome.

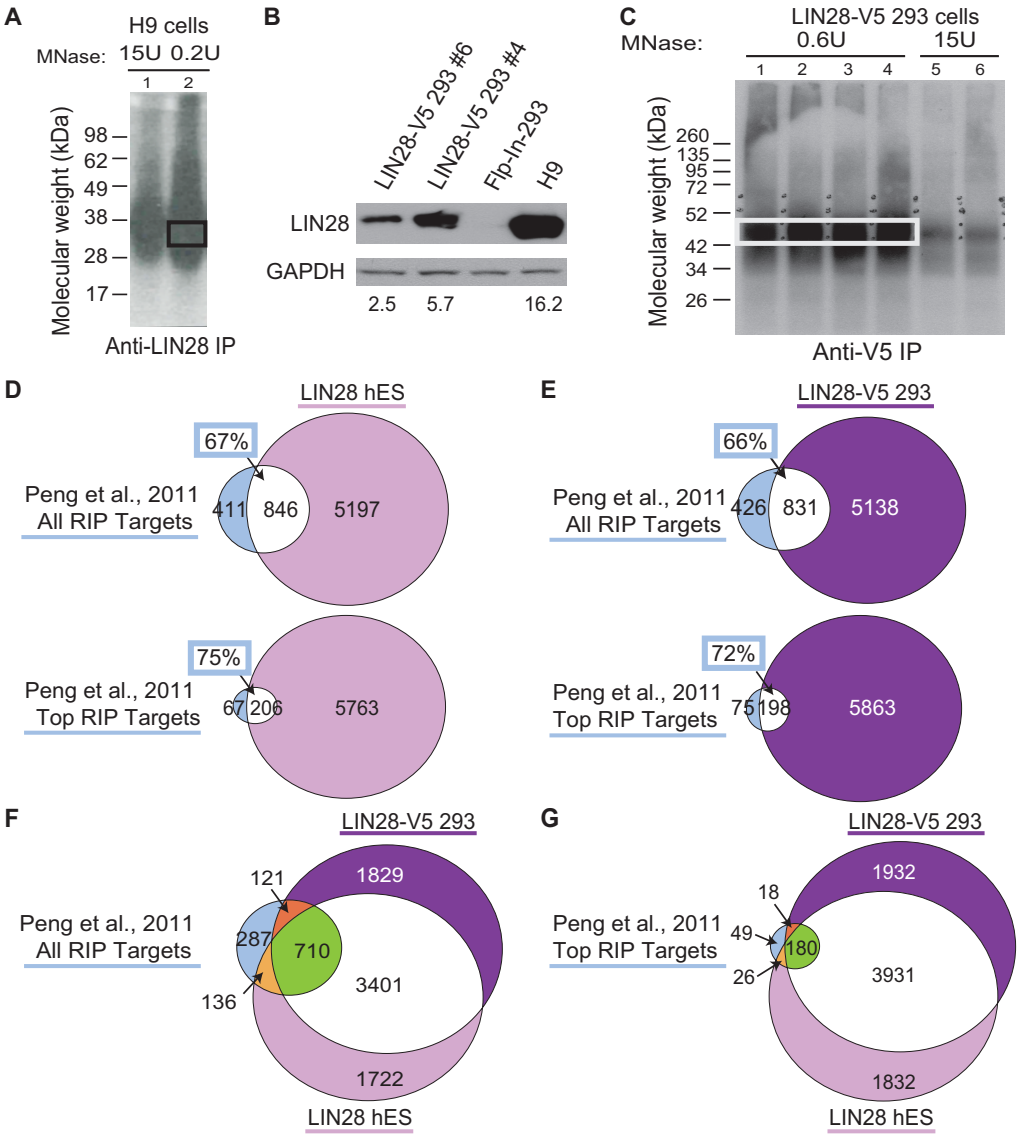


Figure 4. CLIP-seq identifies LIN28 binding sites in two cell lines.
(A) SDS-PAGE gel size selection of LIN28 bound to radioactively labeled RNA targets at two different micrococcal nuclease (MNase) concentrations in H9 hES cells. The bands between 28 and 38 kDa were excised (boxed). (B) Flp-In-293 cells were used to generate HEK293 cell lines stably expressing LIN28-V5 fusion protein. Western blot analysis with an antibody recognizing LIN28 shows LIN28-V5 expression in two stable cell lines, Flp-In-293 cells, and H9 hES cells. The cell line LIN28-V5 #6 was used for all experiments unless otherwise noted. GAPDH served as a loading control. Blot intensities were quantified using ImageJ software; values relative to GAPDH are provided. (C) SDS-PAGE gel size selection of LIN28 bound to radiolabeled RNA targets at two different MNase concentrations in LIN28-V5 293 cells. The bands between 42 and 52 kDa were excised (boxed) and materials from lanes 1-4 were combined. (D and E) Venn diagram comparisons of LIN28 gene targets defined by previous RIP-seq experiments (Peng et al., 2011) and those identified by CLIP-seq in (D) hES or (E) LIN28-V5 293 cells. All RIP targets (top) or a subset of the targets ranked highly by Peng et al. (bottom) are compared. The percentage of RIP-seq targets in common with each CLIP-seq dataset are given (blue boxes). (F and G) Venn diagram comparisons of LIN28 hES and LIN28-V5 293 target genes found in common between (F) all, or (G) highly ranked RIP-seq targets.

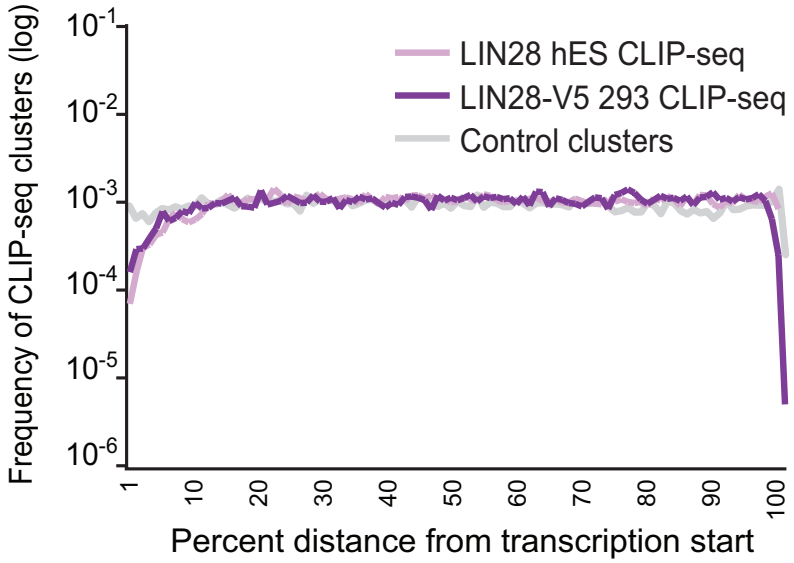


Figure 5. Distribution of LIN28 CLIP-seq clusters across mRNAs
Positional distribution of LIN28 CLIP-seq and randomly distributed control clusters across mature mRNA transcripts, displayed as an aggregate model by percent from transcription start site.

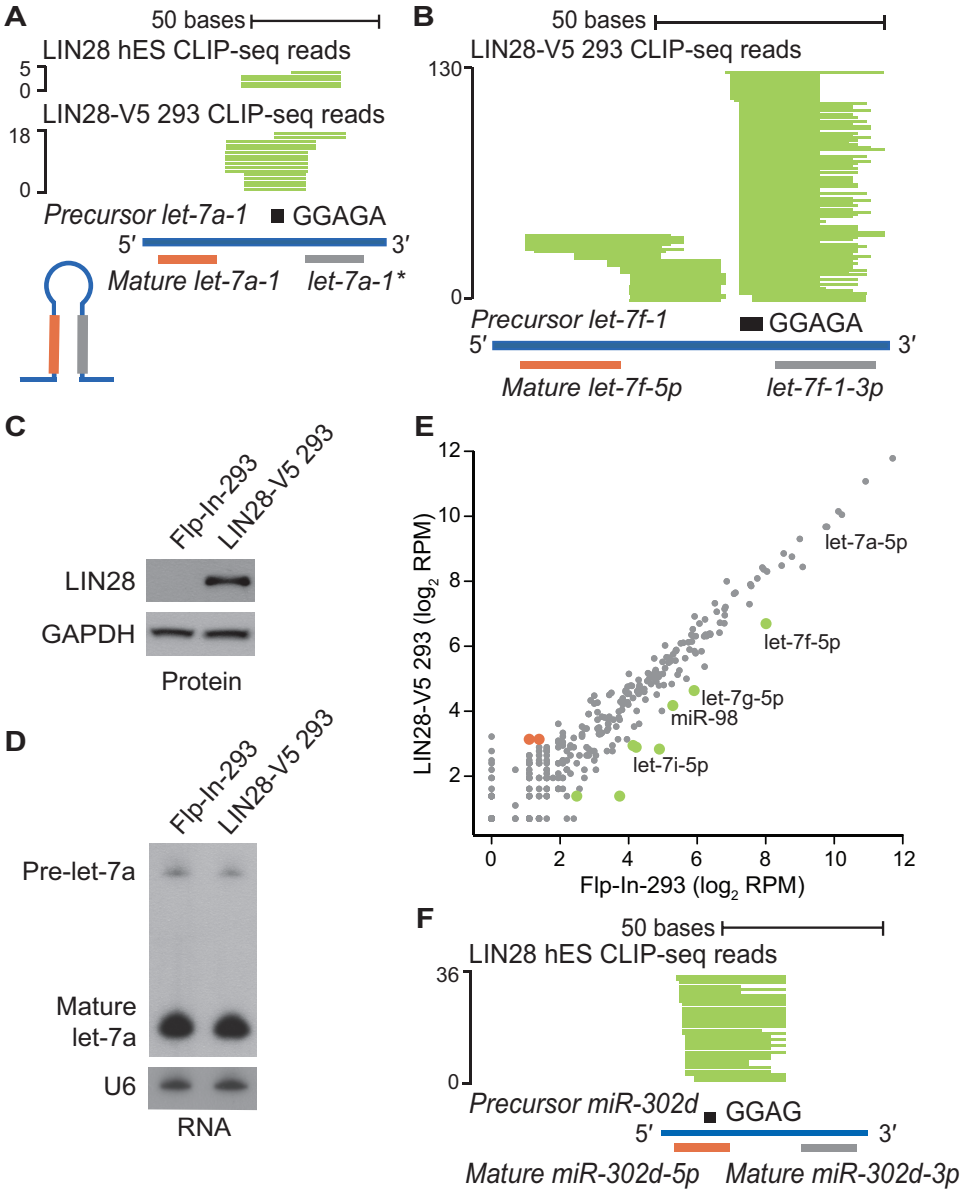


Figure 6. CLIP-seq defines LIN28 binding sites within miRNA precursors.
(A, B) Individual LIN28 CLIP-seq reads (in green) aligned to (A) precursor miRNA let-7a-1 and (B) precursor miRNA let-7f-1, with the mature miRNA boundaries depicted below. The sequence GGAGA in the hairpin loop is depicted as a black rectangle. The scale to the left indicates the number of aligned reads. (C) Western blot analysis of LIN28 protein levels in control Flp-In-293 and LIN28-V5 293 cells. GAPDH serves as a loading control. (D) Northern blot analysis of the human let-7a miRNA in control Flp-In-293 and LIN28-V5 293 cells. The U6 snRNA serves as a loading control. (E) Scatter plot comparing the log2 RPM (reads per million mapped) for expressed mature miRNAs in control Flp-In-293 and LIN28-V5 293 cells (gray), showing significantly upregulated (red) and downregulated (green) miRNAs. (F) LIN28 CLIP-seq reads (in green) aligned to precursor miR-302d, centered on the motif GGAG (black rectangle).

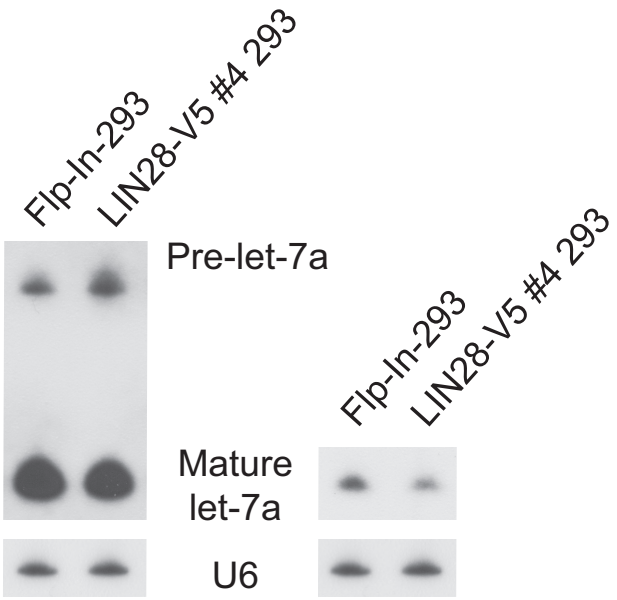


Figure 7. Further overexpression of LIN28 in HEK293 cells prevents let-7 biogenesis
Northern blot analysis of the human let-7a miRNA in control Flp-In-293 and LIN28-V5 #4 293 cells. The LIN28-V5 #4 293 line has higher expression of LIN28-V5 than the LIN28-V5 #6. A shorter exposure of the mature let-7a bands is shown on the right. The U6 snRNA serves as a loading control.

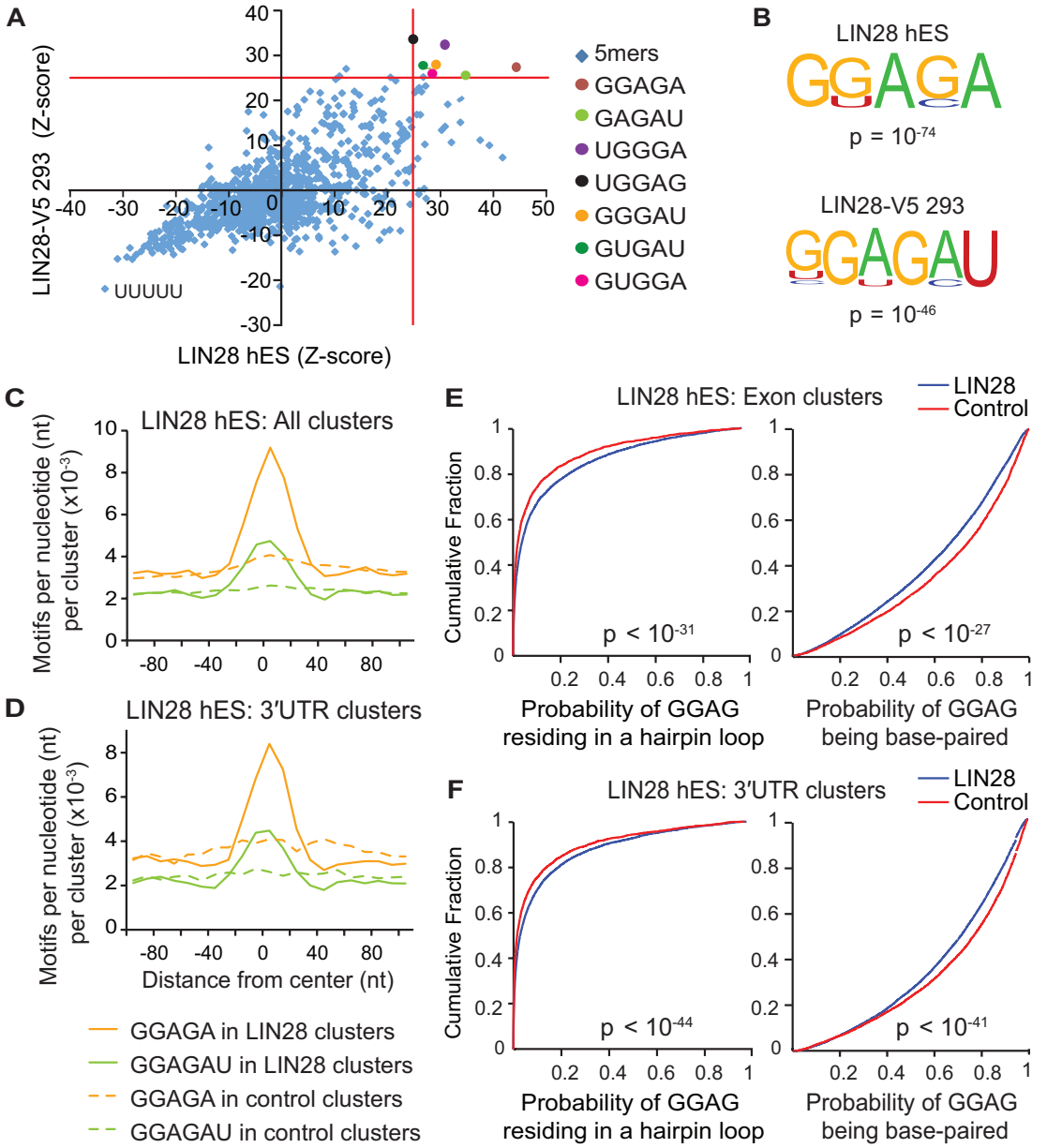


Figure 8. LIN28 binds GGAGA(U) motifs in mRNA sequences within hairpin loop structures.
(A) Scatter plot comparing pentamer Z-scores in hES and LIN28-V5 293 cells. Pentamers overrepresented ($p < 10\text{-}4$) in both cell-types are highlighted by colored circles, and defined on the right. (B) Consensus motifs within LIN28 clusters identified by the HOMER algorithm (Heinz et al., 2010) in hES and LIN28-V5 293 cells with corresponding p-values shown below the motif. (C and D) The positional frequency of consensus motifs GGAGA and GGAGAU relative to the center of (C) all LIN28 hES clusters and (D) clusters only in 3′UTRs. Dashed lines correspond to the positional frequency of these motifs within randomly distributed control clusters from the same type of genic region. (E and F) Cumulative distribution plots display the probability that each nucleotide of a GGAG sequence found within LIN28 hES clusters (blue) or control clusters (red) resides in a predicted hairpin loop (left panel) or base-paired region (right panels) of mRNA; (E) considering clusters only in exons, or (F) clusters only in 3′UTRs (p-values calculated by two-sample Kolgomorov-Smirnov test).

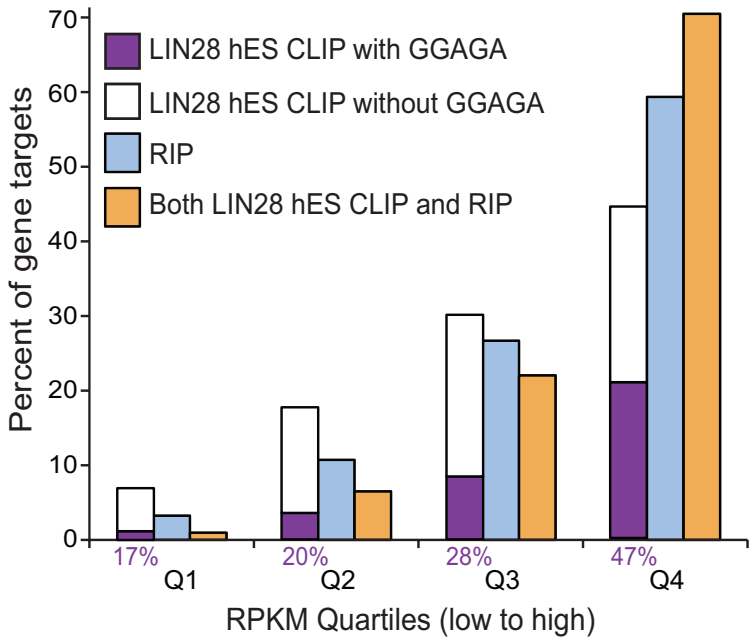


Figure 9. LIN28 binds GGAG(A) sequences in transcripts expressed at different levels
Genes that are LIN28 CLIP-seq and/or RIP-seq (Peng et al., 2011) targets in H9 hES cells were divided into four quartiles based on RPKM values from RNA-seq measurements in untreated H9 cells (see Figure 13D). The percent of LIN28 hES CLIP-seq clusters containing the sequence GGAGA are given.

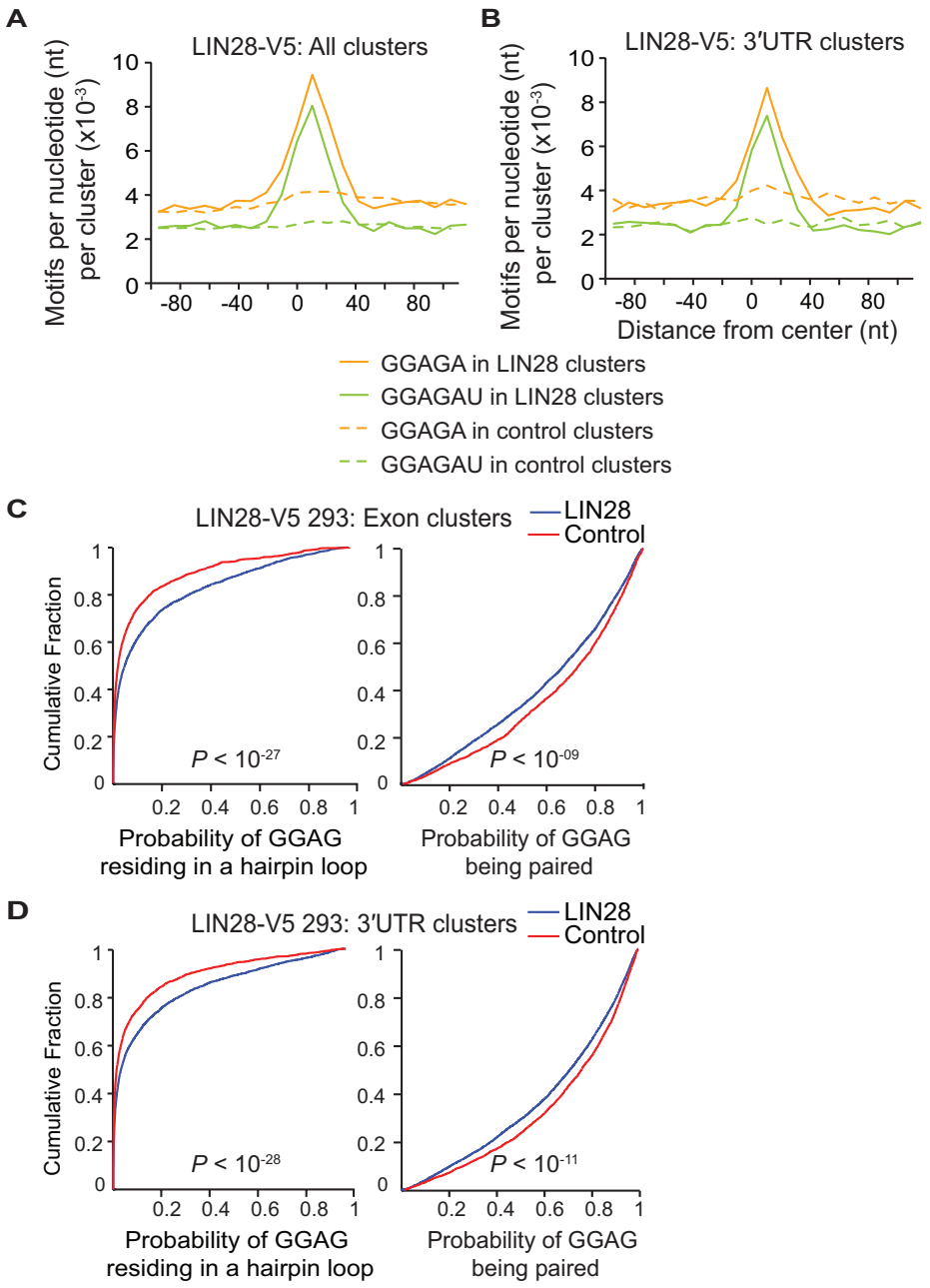


Figure 10. LIN28 binds GGAG(A) sequences in mRNA transcripts within unpaired regions of secondary structure

The positional frequency of consensus motifs GGAGA and GGAGAU relative to the center of (A) all LIN28-V5 293 clusters and (B) clusters only in 3′UTRs. Dashed lines correspond to the positional frequency of these motifs within randomly distributed control clusters from the same type of genic region. (C-D) Cumulative distribution plots display the probability that each nucleotide of a GGAG sequence found within LIN28-V5 293 clusters (blue) or control clusters (red) resides in a predicted hairpin loop (left panel) or a base-paired region (right panels) of mRNA; (C) considering clusters only in exons, or (D) clusters only in 3′UTRs (p-values calculated by two-sample Kolgomorov-Smirnov test).

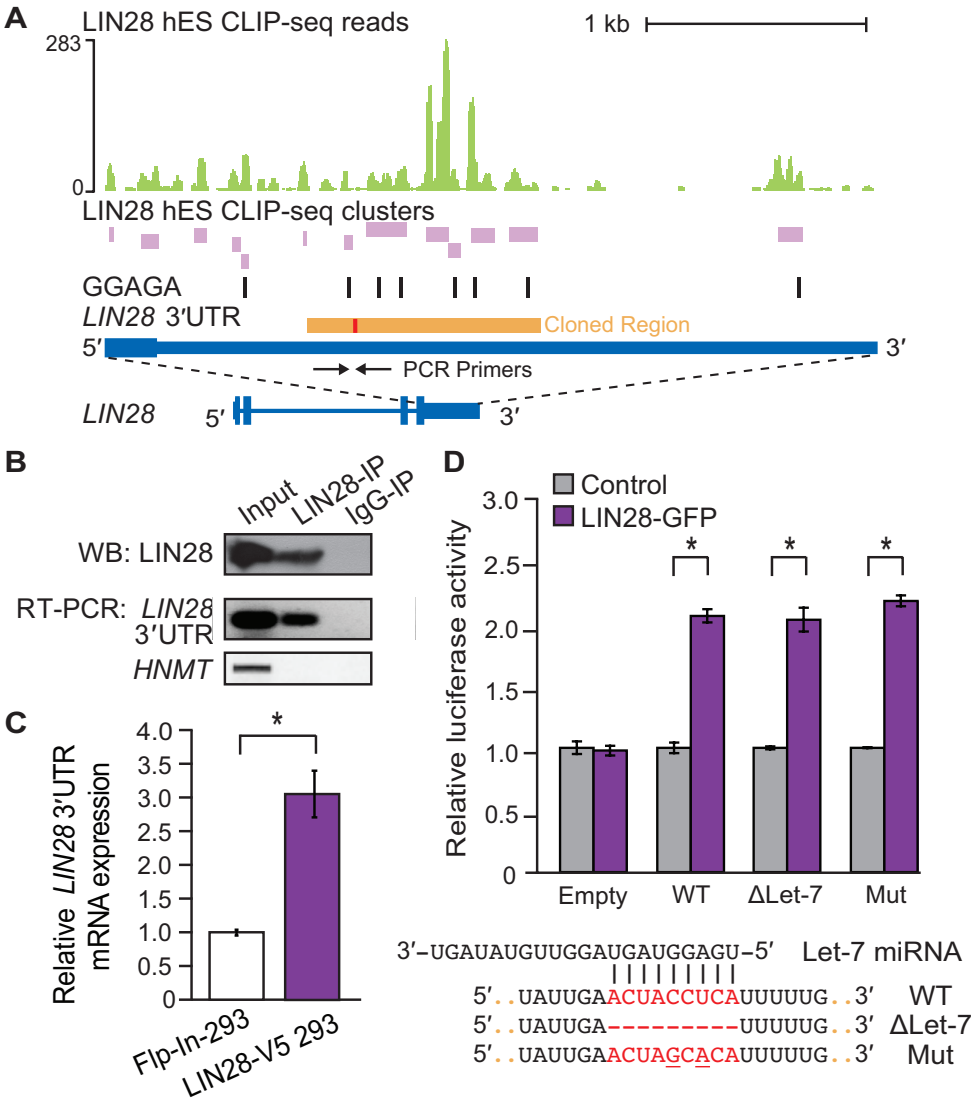


Figure 11. LIN28 binds to its own 3′UTR to positively autoregulate.
(A) LIN28 H9 hES CLIP-seq reads (graphed as continuous densities in green) and clusters (in purple) falling within the LIN28 3′UTR. Instances of GGAGA motifs within clusters are shown (black boxes). The scale to the left indicates the height of aligned reads. A portion of the LIN28 3′UTR (orange) containing a let-7 binding site (red) was cloned downstream of a luciferase open-reading frame (ORF) reporter (see (D)). (B) Western blot (WB) analysis using an antibody recognizing endogenous LIN28 in lysates after immunoprecipitation (IP) of LIN28 and bound RNA transcripts in HUES6 hES cells. IgG was used as an IP control. RNA isolated from the IP was also used for RT-PCR experiments to confirm IP of the endogenous LIN28 3′UTR (primers shown as arrows in (A)), and a negative control, HMNT, that is not bound by LIN28. (C) Quantitative RT-PCR analysis showing increased mRNA levels of endogenous LIN28 in LIN28-V5 293 relative to control Flp-In-293 cells, and normalized to GAPDH levels (*$p < 0.05$, Student's t-test, error bars ± s.d.). (D) Relative luciferase activity of reporters containing a portion of the wild-type (WT) LIN28 3'UTR (as depicted in A), or deletion (ΔLet-7) of or mutations (Mut) within a sequence complementary to the let-7f miRNA, when co-transfected in Flp-In-293 cells with a LIN28-GFP expression vector (purple) or with an unrelated control vector (grey) (*$p < 0.001$, Student's t-test, error bars ± s.d.). A control luciferase reporter lacking the partial LIN28 3′UTR (Empty) was unaffected by LIN28-GFP.

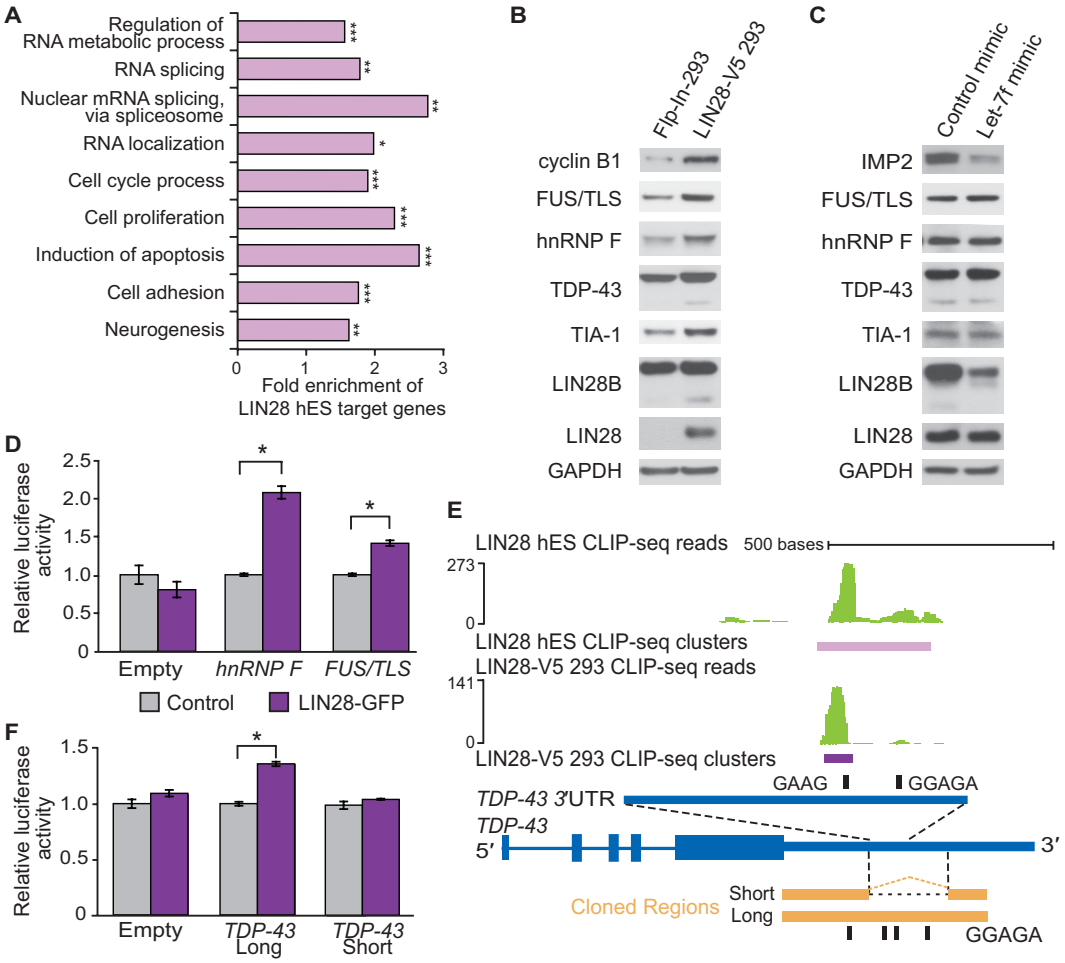
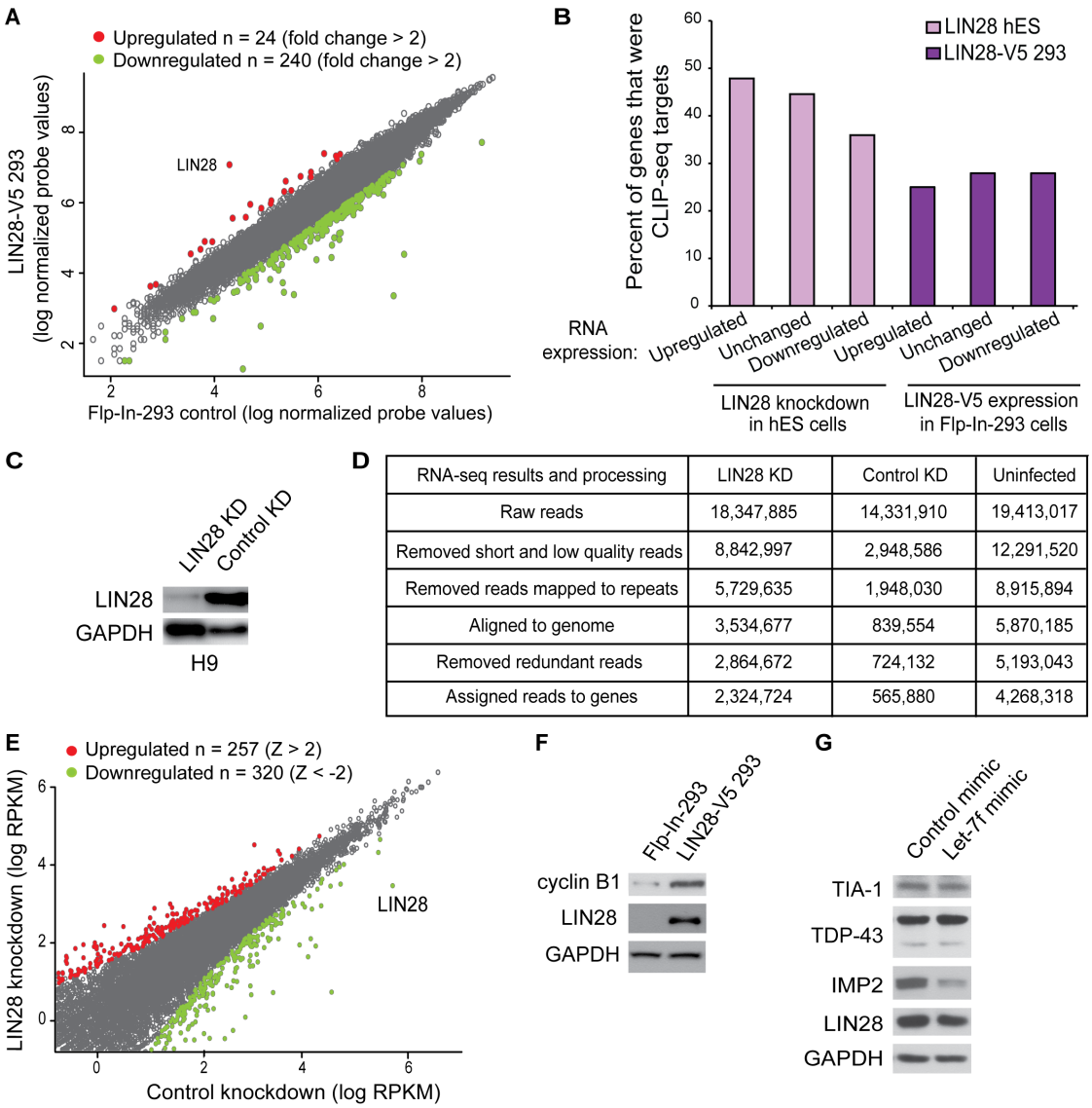


Figure 12. LIN28 binds and regulates splicing factors.

(A) Enriched Gene Ontology (GO) terms for hES LIN28 target genes were identified using the DAVID algorithm (Huang et al., 2009). Statistical comparisons to all genes with transcripts expressed in H9 hES cells were made (*p < 10-10, **p < 10-15, ***p < 10-40). (B) Western blot analysis of splicing factors in control Flp-In-293 and LIN28-V5 293 cells. Cyclin B1 is shown as a positive control (Xu et al., 2009). All membranes were probed for LIN28 and GAPDH (used as a loading control). (C) Western blot analysis of splicing factors in LIN28-V5 293 cells transfected with a let-7f miRNA mimic or control mimic. The let-7 target IMP2 was used as a positive control. (D) Relative luciferase activity of reporters containing cloned portions of the hnRNP F or FUS/TLS LIN28-bound RNA regions co-transfected into Flp-In-293 cells with a LIN28-GFP expression vector (purple) or with a control vector (grey) (*p < 0.001, Student's t-test, error bars ± s.d.). A control luciferase reporter lacking a LIN28-bound region (Empty) was unchanged by LIN28-GFP. (E) LIN28 CLIP-seq reads (in green) and clusters (in purple) mapped to an intronic region within the 3′UTR of the human TDP-43 gene (in blue). The scale to the left indicates the height of aligned reads. Portions of the homologous mouse TDP-43 3'UTR that contain (long) or lack (short) the intronic region that harbors the majority of LIN28 binding sites are shown aligned (in orange). These regions were inserted downstream of a luciferase reporter as previously described (Polymenidou et al., 2011). Instances of GGAGA and GAAG motifs in the respective organisms are shown (black rectangles). (F) Relative luciferase activity of reporters containing the TDP-43 3'UTR with LIN28 binding sites (long), and the TDP-43 3'UTR without LIN28 binding sites (short) co-transfected into Flp-In-293 cells with a LIN28-GFP expression vector (purple) or a control vector (grey) (*p < 0.001, Student's t-test, error bars ± s.d.). A control luciferase reporter lacking any LIN28-bound region (Empty) was unchanged by LIN28-GFP.



RNA-seq results and processing	LIN28 KD	Control KD	Uninfected
Raw reads	18,347,885	14,331,910	19,413,017
Removed short and low quality reads	8,842,997	2,948,586	12,291,520
Removed reads mapped to repeats	5,729,635	1,948,030	8,915,894
Aligned to genome	3,534,677	839,554	5,870,185
Removed redundant reads	2,864,672	724,132	5,193,043
Assigned reads to genes	2,324,724	565,880	4,268,318

Figure 13. Expression analyses upon misregulation of LIN28 in hES and Flp-In-293 cells.
(A) Scatter-plot of log-normalized probe values from splicing-sensitive microarray analyses representing upregulated (red), downregulated (green) and unchanged (gray) genes upon LIN28 expression (LIN28-V5) in Flp-In-293 cells. Up and downregulated genes were determined by 2-fold or greater change in normalized probe values. (B) Bar plots showing the percentages of genes affected at the mRNA level upon LIN28 depletion (hES cells) or expression (LIN28-V5 293 cells) that were also LIN28 CLIP-seq targets in the corresponding cell-type. (C) Western blot analysis with an antibody recognizing LIN28 in H9 hES cells treated with lentivirus carrying an shRNA targeting LIN28 or a GFP control. GAPDH is used as loading control. (D) RNA extracted from three replicate infections were combined and subjected to strand-specific RNA-seq and the resulting sequenced reads were processed. Read counts for each processing step are shown. (E) Scatter-plot of log-normalized RPKM values representing upregulated (red, Z-score > 2), downregulated (green, Z-score < 2) and unchanged (gray) genes upon LIN28 depletion in hES cells. (F) Cyclin B1 and LIN28 western blot with GAPDH as loading control. Only representative controls were used in Figure 12B. (G) Western blots with GAPDH as loading control. Only representative controls were used in Figure 12C.

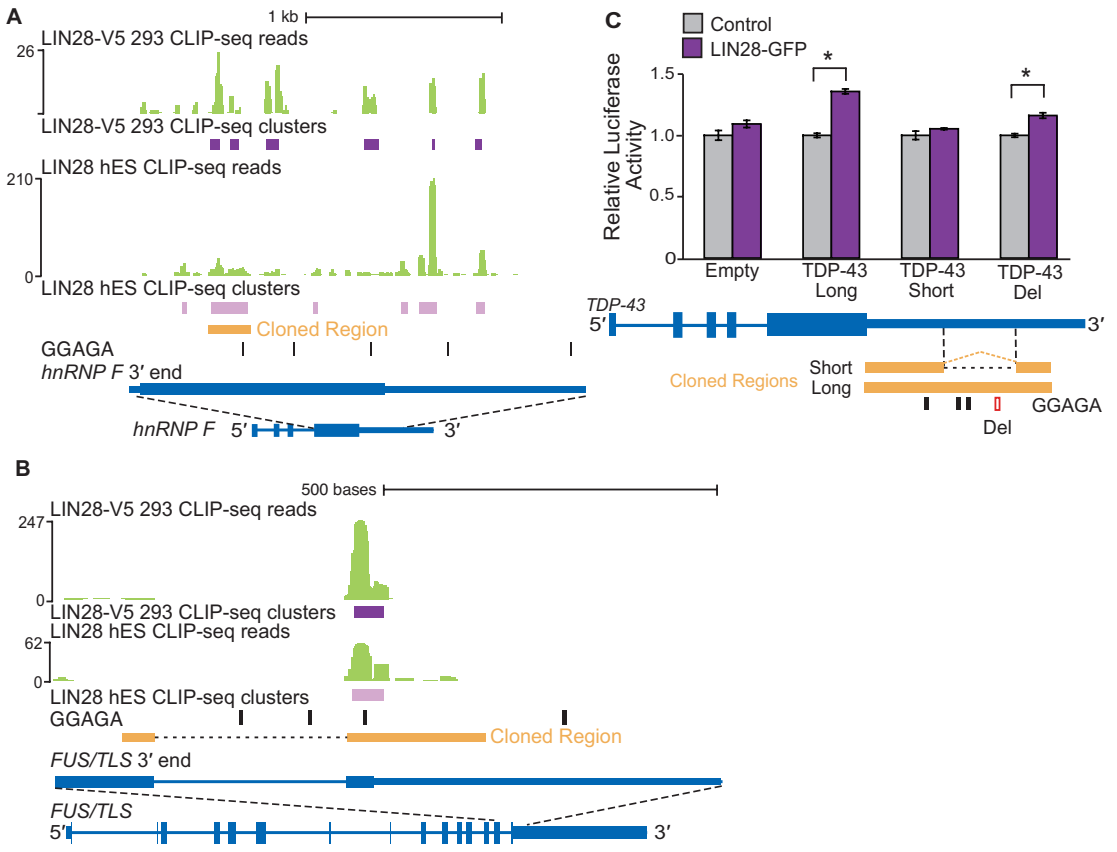


Figure 14. LIN28 CLIP-seq identifies sites of regulation within mRNA transcripts.
(A,B) LIN28 CLIP-seq reads (in green) and clusters (in purple) mapped to regions within the hnRNP F (A) and FUS/TLS genes. Instances of GGAGA motifs in the respective genes are shown (black rectangles). The scale to the left indicates the height of aligned reads. A region of mRNA from each gene ("Cloned Region") was cloned downstream of a Renilla reporter gene, within a construct that expresses Firefly luciferase as an internal transfection control. (C) A schematic (below) of human TDP-43 (in blue) depicts regions in its 3'UTR homologous to the mouse TDP-43 3'UTR (in orange). The relative luciferase activity was determined for reporters containing LIN28 binding sites (long), the spliced TDP-43 3'UTR without LIN28 binding sites (short), and the TDP-43 3'UTR with LIN28 binding sites but a deletion of the most 3' GGAGA motif (Del) co-transfected into Flp-In-293 cells with a LIN28-GFP expression vector or a control vector (*$p < 0.005$, Student's t-test, error bars ± s.d.). The positions of the GGAGA motif (black rectangles) and the deleted most 3′ GGAGA motif (red rectangle) are shown. A control luciferase ORF reporter lacking any LIN28-bound region (Empty) was unaffected by co-transfection with LIN28-GFP.

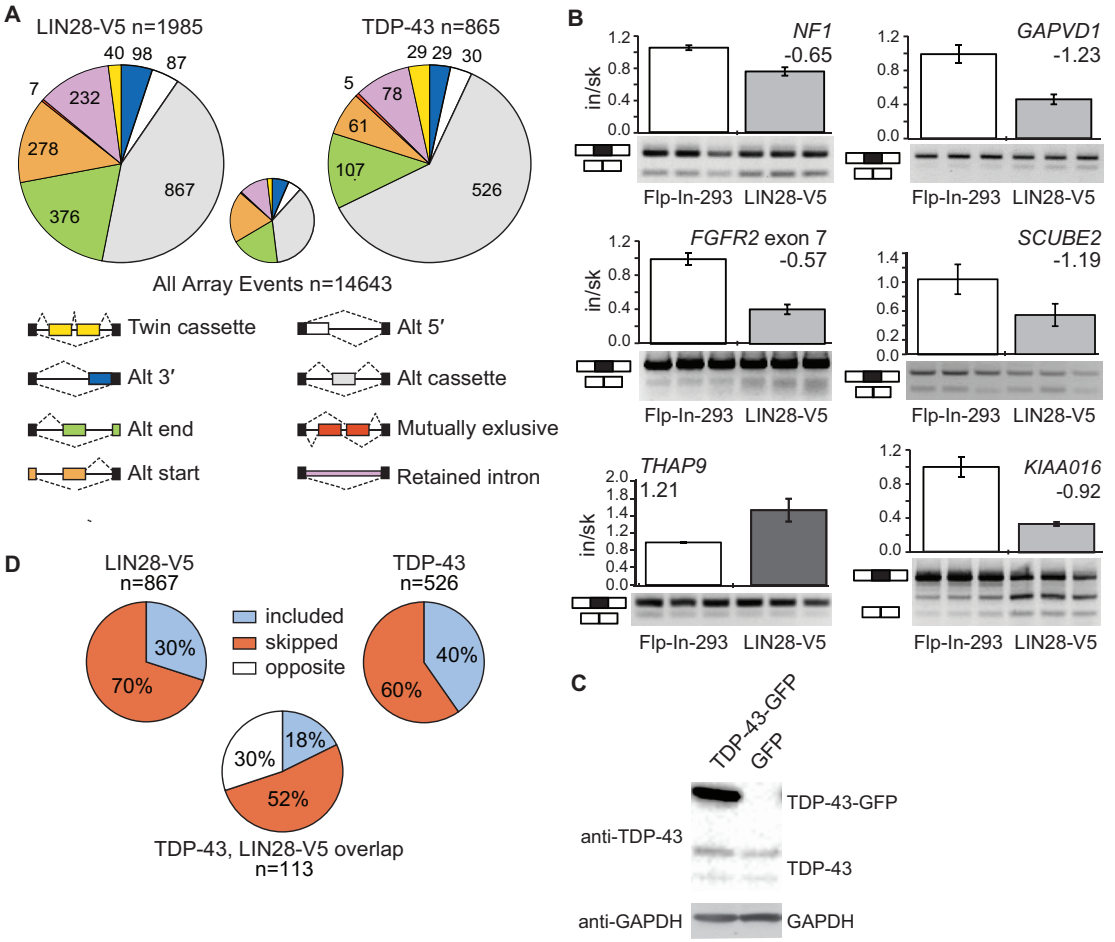


Figure 15. LIN28 expression in somatic cells results in thousands of alternative splicing events, in part through regulation of TDP-43 levels.

(A) The pie charts display the number of each type of alternative splicing event changed upon overexpression of LIN28-V5 (left) or TDP-43 (right) in Flp-In-293 cells, as detected by splicing-sensitive microarray analyses. The small pie chart (center) represents the distribution of alternative splicing event types detected on the microarray. (B) RT-PCR validations of alternative cassette events detected by microarray analysis. All plots show significant differences between control Flp-In-293 and LIN28-V5 293 cells ($p < 0.05$, Students t-test). Bars represent an average and error bars represent the standard deviation across biological triplicates. (C) Western blot analysis using antibodies against TDP-43 and GAPDH (loading control) in Flp-In-293 cells transfected with a TDP-43-GFP or control GFP expression vector. (D) The percent of included versus skipped alternative cassette exons upon overexpression of LIN28-V5 (left) or TDP-43 (right) in Flp-In-293 cells. For the alternative cassette exons that changed in both conditions (n = 113), the percent of exons affected in the same (where the exon is included or skipped in both conditions) or opposite direction are shown below.

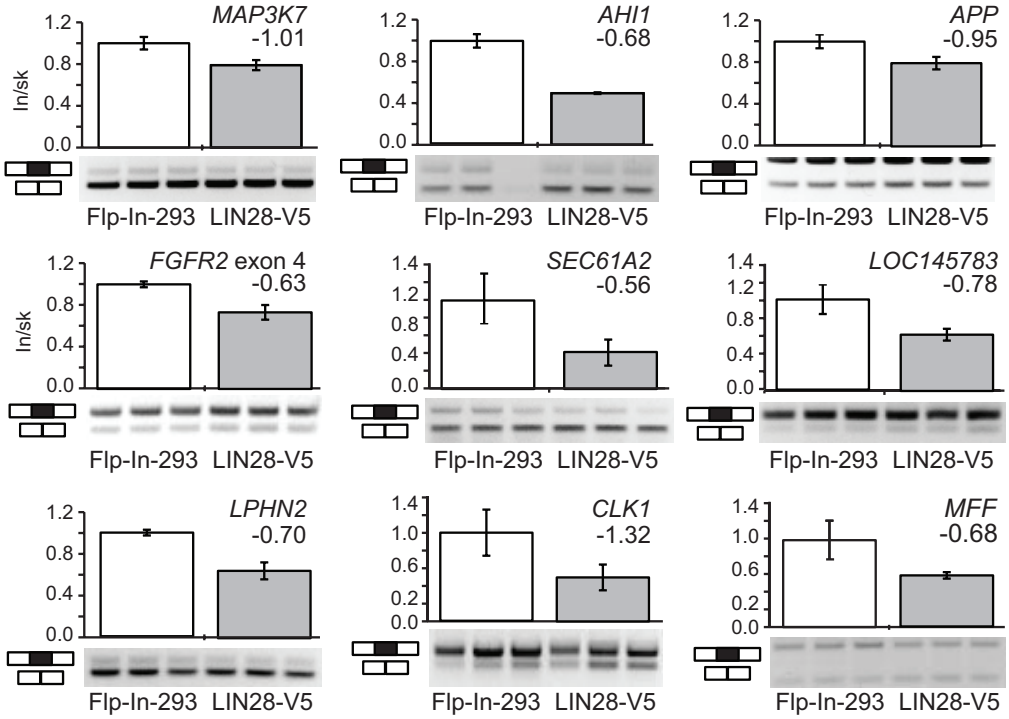


Figure 16. Validation of LIN28 regulated alternative splicing events in HEK293 cells.
Additional RT-PCR validations of alternative cassette events detected by microarray analysis. All plots show significant differences between control Flp-In 293 and LIN28-V5 293 cells ($p < 0.05$, Students t-test). Bar graphs represent the average value and error bars display the standard deviation across biological triplicates.

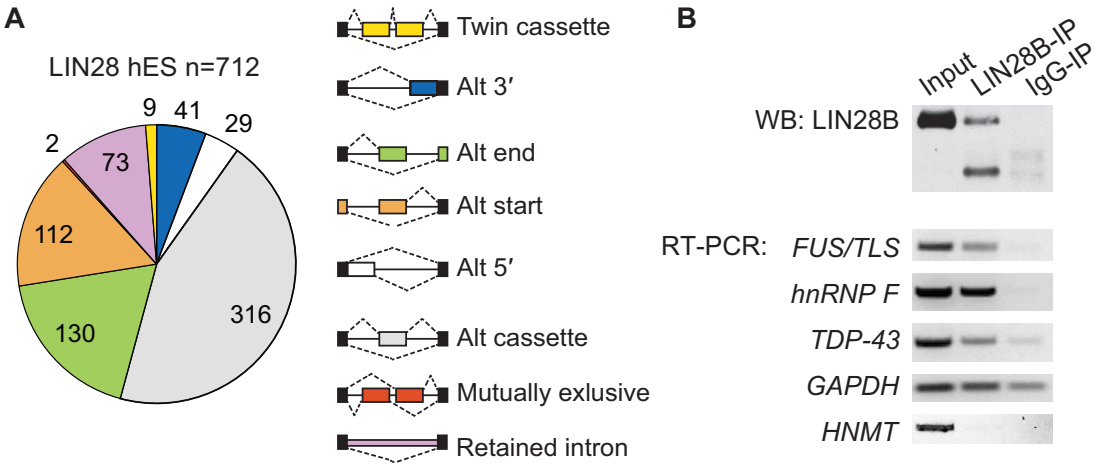


Figure 17. LIN28 regulates alternative splicing in hESCs.
(A) The pie charts display the number of each type of alternative splicing event changed upon knockdown of LIN28 in H9 hES cells (different event types depicted to the right). (B) Western blot (WB) analysis using an antibody recognizing endogenous LIN28B in lysates after immunoprecipitation (IP) of LIN28B and bound RNA transcripts in control Flp-In-293 cells. IgG was used as an IP control. RNA isolated from the IP was used for RT-PCR experiments with primers targeting splicing factor targets. HNMT (not bound by LIN28B) and GAPDH (not sufficiently enriched by LIN28B IP above IgG IP) serve as negative controls.

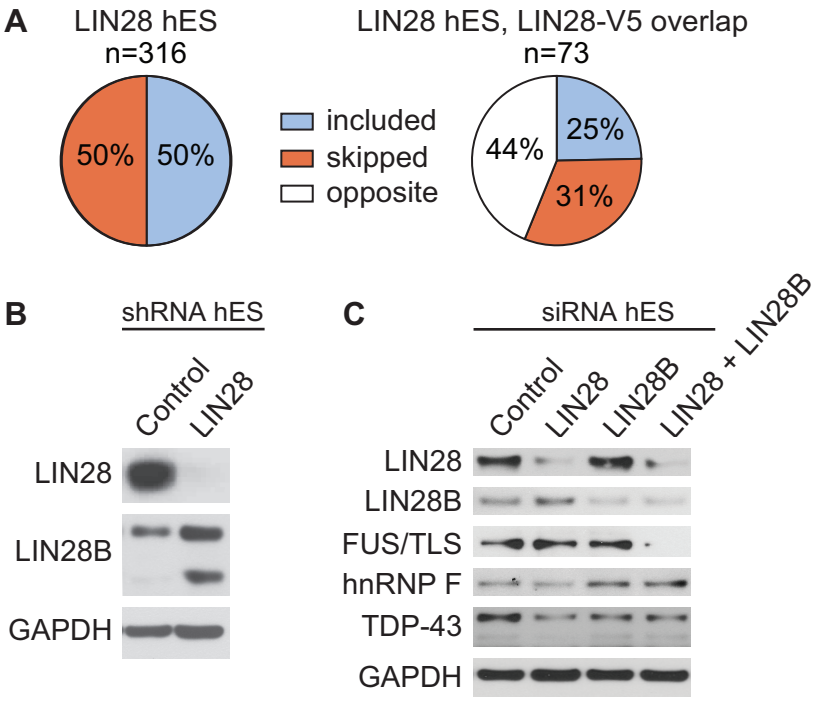


Figure 18. LIN28 and LIN28B affect splicing factors differently in human ES cells.
(A) The percent of included versus skipped alternative cassette exons upon depletion of LIN28 in hES cells is shown. Of the alternative cassette events changed in both the LIN28 hES and LIN28-V5 293 experiments (n = 73), the percent of exons affected in the same (where the exon is included or skipped in both conditions) or opposite direction are shown below. The direction of cassette exon splicing changes due to LIN28 depletion in hES cells is flipped to correspond to LIN28 overexpression. (B) Western blot analysis of LIN28B levels upon shRNA-mediated depletion of LIN28 in hES cells. GAPDH serves as a loading control. (C) Western blot analysis of LIN28, LIN28B, and splicing factors upon siRNA-mediated depletion of LIN28, LIN28B, or both in hES cells.

TABLES

Table 1. CLIP-seq sequencing results and processing.

Processing for mRNA targets	LIN28 hES CLIP 1	LIN28 hES CLIP 2	LIN28-V5 293 CLIP
Raw reads	1,756,073	12,476,453	12,476,453
Removed short and low quality reads	1,474,739	9,132,443	7,952,475
Removed reads mapped to repeats	980,415	5,473,826	4,230,107
Aligned to genome	649,989	4,194,424	2,839,810
Removed redundant reads	581,457	2,521,717	1,255,931
Assigned reads to genes	2,373,148		880,288
Significant clusters	26,279		15,028
Gene targets	5,969		6,601
Processing for miRNA targets of LIN28			
Reads mapped to miRNA precursors	925		5,393
Precursor miRNAs targeted	36		58

Table 2. CLIP-seq reads mapped to precursor miRNAs and mature miRNA expression values.
Reads per million (RPM) returned from small-RNA seq for mature miRNAs with evidence of CLIP-seq binding to the corresponding precursors. Significantly changed miRNAs upon LIN28-V5 overexpression are highlighted in green (downregulated) and red (upregulated).

Precursor miRNA	**V5 CLIP**	**ES CLIP**	**Mature miRNA**	**LIN28-V5 RPM**	**Flp-In-293 RPM**	**H9 RPM**
let-7a-1	18	5	let-7a-5p	15,917	17,172	1,976
let-7a-3	19	0	let-7a-5p	15,917	17,172	1,976
let-7b	69	0	let-7b-5p	148	295	13
let-7c	22	0	let-7c	215	501	128
let-7d	24	0	let-7d-3p	0	3	0
			let-7d-5p	4	12	30
let-7e	6	0	let-7e-3p	0	1	0
			let-7e-5p	345	458	57
let-7f-1	130	0	let-7f-5p	808	3,009	778
let-7f-2	8	0	let-7f-5p	808	3,009	778
let-7g	134	0	let-7g-5p	103	369	91
let-7i	69	0	let-7i-5p	18	68	59
mir-101-2	13	0	miR-101-5p	0	0	3
mir-103-2	17	0	miR-103a-2-5p	0	0	5
miR-106b	25	0	miR-106b-3p	108	86	1
			miR-106b-5p	168	136	719
miR-10a	133	0	miR-10a-3p	4	3	0
			miR-10a-5p	131,067	121,222	242
mir-1182	8	0	mir-1182-5p	0	0	0
miR-1226	13	0	miR-1226-3p	5	4	4
miR-1244-1	12	136	miR-1244	0	0	0
miR-1244-2	20	97	miR-1244	0	0	0
miR-1244-3	8	70	miR-1244	0	0	0
miR-1248	368	5	miR-1248	0	0	4
miR-125a	80	0	miR-125a-3p	5	13	14
			miR-125a-5p	1,268	898	102
miR-1291	70	0	miR-1291	0	0	0
miR-1306	7	0	miR-1306-5p	5	5	1

Table 2 continued. CLIP-seq reads mapped to precursor miRNAs and mature miRNA expression values.
Reads per million (RPM) returned from small-RNA seq for mature miRNAs with evidence of CLIP-seq binding to the corresponding precursors. Significantly changed miRNAs upon LIN28-V5 overexpression are highlighted in green (downregulated) and red (upregulated).

Precursor miRNA	**V5 CLIP**	**ES CLIP**	**Mature miRNA**	**LIN28-V5 RPM**	**Flp-In-293 RPM**	**H9 RPM**
miR-1307	6	0	miR-1307-3p	172	55	40
			miR-1307-5p	149	117	94
miR-16-2	0	7	miR-16-2-3p	2	2	1
			miR-16-5p	4,749	4,823	1,454
miR-181b-1	6	0	miR-181b-3p	7	15	0
			miR-181b-5p	314	200	10
miR-182	8	0	miR-182-5p	15,891	17,871	6,125
miR-1825	0	6	miR-1825-5p	0	0	0
miR-186	6	0	miR-186-3p	0	1	0
			miR-186-5p	681	355	361
miR-193b	0	5	miR-193b-3p	99	50	160
			miR-193b-5p	2	3	5
miR-19a	5	0	miR-19a-3p	325	385	1,615
miR-19a	5	0	miR-19a-5p	2	3	4
miR-19b-1	19	1	miR-19b-1-5p	5	7	111
miR-198	0	50	miR-198-5p	0	0	0
miR-200c	0	13	miR-200c-3p	7	7	2,983
			miR-200c-5p	0	0	5
miR-20a	14	5	miR-20a-3p	2	0	6
			miR-20a-5p	1,471	1,856	68,443
miR-20b	0	5	miR-20b-3p	0	0	62
			miR-20b-5p	23	40	40,183
miR-25	11	0	miR-25-3p	10,957	8,064	1,064
			miR-25-5p	14	10	10
miR-26a-1	2	6	miR-26a-5p	4,024	3,097	1,738
miR-29b-2	23	0	miR-29b-2-5p	0	0	0
miR-301a	0	9	miR-301a-3p	347	372	100
			miR-301a-5p	5	5	22

Table 2 continued. CLIP-seq reads mapped to precursor miRNAs and mature miRNA expression values.
Reads per million (RPM) returned from small-RNA seq for mature miRNAs with evidence of CLIP-seq binding to the corresponding precursors. Significantly changed miRNAs upon LIN28-V5 overexpression are highlighted in green (downregulated) and red (upregulated).

Precursor miRNA	**V5 CLIP**	**ES CLIP**	**Mature miRNA**	**LIN28-V5 RPM**	**Flp-In-293 RPM**	**H9 RPM**
miR-302a	0	9	miR-302a-3p	2	1	21,846
			miR-302a-5p	11	1	6,716
miR-302b	0	24	miR-302b-3p	14	3	509,541
			miR-302b-5p	2	0	116
miR-302c	0	20	miR-302c-3p	4	0	40,942
			miR-302c-5p	0	1	817
miR-302d	0	36	miR-302d-3p	18	7	24,480
			miR-302d-5p	0	0	147
miR-30a	18	0	miR-30a-3p	360	219	238
			miR-30a-5p	150	100	136
miR-30e	15	5	miR-30e-3p	811	431	1,467
			miR-30e-5p	126	192	727
miR-3137	5	0	miR-3137	0	0	1
miR-324	6	0	miR-324-3p	2	2	3
			miR-324-5p	9	9	3
miR-34a	8	0	miR-34a-3p	0	3	0
			miR-34a-5p	122	152	35
miR-3607	0	20	miR-3607-3p	2	1	6
			miR-3607-5p	0	0	4
miR-3620	100	1	miR-3620	2	0	0
miR-3648	20	0	miR-3648-5p	0	0	0
miR-3652	28	0	miR-3652-5p	0	0	0
miR-3654	27	53	miR-3654	0	0	0
miR-3658	0	6	miR-3658-5p	0	0	0
miR-367	0	111	miR-367-3p	0	0	12,338
			miR-367-5p	0	0	5
miR-3676	5	1	miR-3676-3p	0	0	6
			miR-3676-5p	0	0	0
miR-3687	14	0	miR-3687-5p	0	0	0

Table 2 continued. CLIP-seq reads mapped to precursor miRNAs and mature miRNA expression values.
Reads per million (RPM) returned from small-RNA seq for mature miRNAs with evidence of CLIP-seq binding to the corresponding precursors. Significantly changed miRNAs upon LIN28-V5 overexpression are highlighted in green (downregulated) and red (upregulated).

Precursor miRNA	**V5 CLIP**	**ES CLIP**	**Mature miRNA**	**LIN28-V5 RPM**	**Flp-In-293 RPM**	**H9 RPM**
miR-374a	14	0	miR-374a-3p	441	250	7,953
			miR-374a-5p	121	75	261
miR-581	6	0	miR-581-5p	0	0	0
miR-448	0	5	miR-448	2	0	71
miR-598	51	0	miR-598	34	38	1,026
miR-611	0	7	miR-611-5p	0	0	0
miR-615	144	0	miR-615-3p	550	387	0
			miR-615-5p	2	1	0
miR-632	5	0	miR-632-5p	0	0	0
miR-652	1062	10	miR-652-3p	81	46	115
			miR-652-5p	9	0	1
miR-663	457	35	miR-663a	0	0	1
miR-671	1815	3	miR-671-3p	45	41	0
			miR-671-5p	12	20	79
miR-877	8	0	miR-877-5p	34	14	283
miR-92a-1	0	10	miR-92a-1-5p	32	30	91
miR-93	38	0	miR-93-3p	4	5	5
			miR-93-5p	2,093	1,233	8,148
miR-98	14	0	miR-98	65	198	5
miR-1247	0	0	miR-1247-5p	4	12	0
miR-34c	0	0	miR-34c-5p	4	42	1307
miR-1	0	0	miR-1	19	62	286
miR-100	0	0	miR-100-5p	17	134	97
miR-1910	0	0	miR-1910	23	4	0
miR-548k	0	0	miR-548k	23	3	16

Table 3. Probabilities of LIN28 bound motifs to occur within regions of secondary structure.
The probability that each nucleotide of a GGAG or GNGAY motif sequence found within LIN28-V5 293 or LIN28 hES clusters to reside within structural elements or base-paired regions as compared to control clusters, p-values calculated by two-sample Kolgomorov-Smirnov test.

	LIN28 hES exon clusters		LIN28-V5 293 exon clusters		LIN28 hES 3′UTR clusters		LIN28-V5 293 3′UTR clusters	
	GGAG (p)	GNGAY (p)	GGAG (p)	GNGAY (p)	GGAG (p)	GNGAY (p)	GGAG (p)	GNGAY (p)
Paired	1.12E-27	9.18E-04	1.97E-09	1.87E-03	9.58E-41	7.81E-31	6.24E-11	7.15E-04
Hairpin loop	1.91E-31	4.94E-08	8.60E-27	2.55E-12	2.43E-44	1.11E-24	3.61E-28	1.03E-11
Internal loop	2.16E-09	2.38E-04	5.02E-01	9.47E-02	7.43E-21	1.11E-36	6.80E-01	3.19E-04
Multi-loop	5.18E-19	1.28E-04	4.76E-06	2.75E-02	1.37E-22	4.91E-19	3.50E-05	1.38E-09
External loop	8.90E-10	4.05E-03	6.61E-08	6.01E-07	1.70E-19	3.60E-10	5.70E-08	2.28E-19

ACKNOWLEDGMENTS

I would like to thank Jonathan Scolnick for critical reading of the manuscript, D. Cleveland for the *TDP-43* luciferase constructs, E. Moss for the LIN28-GFP construct, and B. Wolozin for the TDP-43-GFP construct. We thank L. Shiu, J.P.D and M. Ares for assistance with splicing array analysis. M. Wilbert and T. Stark were supported in part by the UCSD Genetics Training Program through an institutional training grant from the National Institute of General Medical Sciences, T32 GM008666. S. Huelga was funded by a National Science Foundation Graduate Research Fellowship. M Lovci was supported by a fellowship from Genentech. This work was supported by grants to G. Yeo from the US National Institutes of Health (HG004659, GM084317 and NS075449) and the California Institute for Regenerative Medicine (RB1-01413 and RB3-05009). G. Yeo is an Alfred P. Sloan Research Fellow.

CHAPTER 3 - COMPARATIVE ANALYSIS OF TRANSCRIPTOME REGULATION NETWORKS

ABSTRACT

The availability of transcriptome-wide binding data for RNA-binding proteins (RBPs) has increased our understanding of their far-reaching influence on cell fate. Access to many related datasets of RNA maps presents both opportunities and challenges for the interpretation of RBP function. In particular, intense interest in the pluripotency associated RBP LIN28 has resulted in several studies of its global RNA targets and mechanisms in post-transcriptional regulation. Review of the major findings of these reports and comparisons of their sets of target genes for both LIN28A and LIN28B show that despite differences in total gene target sets conclusions regarding their characteristic binding patterns are consistent.

Keeping these lessons in mind, we interrogated a potentially related family of developmentally regulated RBPs, the IMP proteins (Igf2- mRNA binding protein or IG2BP). Like the LIN28s, the IMPs are expressed in an oncofetal expression pattern and are highly expressed in human pluripotent stem cells. Our studies suggest that IMP1 and IMP2 bind a large subset of the same RNA transcripts, but in different contexts. These results provide insight into the molecular function of these two family members, and how they may act in distinct but coordinate mechanisms to direct the fate of their target RNAs. Differences in IMP1 and IMP2 binding patterns and target sequence motifs suggest divergent mechanisms of RNA regulation for these similar but not equivalent proteins.

As we explore connections between different post-transcriptional networks, the question remains how these relate to underlying transcriptional regulation. It has been commonly observed that regulators within a certain pathway will commonly affect other factors in control of the same cellular process. Indeed, the reprogramming factors OCT4, SOX2, and NANOG co-occupy the same promoter sites in mouse and human ES cells. To test if this tendency to control the same gene targets extends to post-transcriptional regulators of pluripotency we compared LIN28 and IMP target genes to ChIP-seq DNA binding sites. Our initial studies suggest that LIN28 supports its own network of mRNA targets, possibly stabilizing gene expression through pathways unique from these transcription factors. However, it remains to be experimentally investigated if any of the co-regulated genes by OSN and LIN28 are vital to the contribution of LIN28 to the reprogramming process.

INTRODUCTION

Transcriptome-wide studies have defined LIN28 mRNA interactions

Access to global methods for the characterization of cellular systems has expanded our appreciation of regulation from one-to-one connections into a view of genome-wide control. This has been particularly true in the study of RNA-binding proteins (RBPs), especially with advances in technologies based on immunoprecipitation of proteins and associated RNAs to define direct binding events. This can be accomplished using the RNA transcript as bait or more commonly with an antibody to a particular protein of interest (Castello et al., 2012; Yeo et al., 2009). The

combination of these basic techniques with advances in high-throughput sequencing can provide information about precise interaction sites as well as the global view of pathways and mechanisms under RBP control.

The early identification of LIN28 as a ‘stemness marker’ and its ability to drive de-differentiation of adult somatic cells into an induced pluripotency stem cell state has directed the application of the latest approaches to identify the targets and pathways that make this function possible (Richards et al., 2004; Yu et al., 2007). In particular CLIP-seq (cross-linking followed by immunoprecipitation and high-throughput sequencing) can be applied in stem cell models to reliably identify sites of RBP transcript binding (Yeo et al., 2009). This protocol, as applied to LIN28-RNA target finding is described in Figure 19. Several recent studies have applied variations of this approach to LIN28A and LIN28B family RBPs. Six related LIN28A transcriptome-wide studies have been published including RIP-Chip in A2780 cells; RIP-seq in hESCs; CLIP-Seq in hES, mES, and HEK293 cells, and PAR-CLIP in HEK293 cell (Cho et al., 2012; Hafner et al., 2013; Li et al., 2012a; Peng et al., 2011; Wilbert et al., 2012). LIN28B CLIP-seq datasets were also generated by PAR-CLIP in HEK293 cells, as well as by iDoPAR-CLIP, which associates RNA binding sites with a particular RNA-binding domain of the protein (Graf et al., 2013; Hafner et al., 2013). All of these approaches rely on the basic concept that immunoprecipitation (IP) can be used to selectively isolate mRNP complexes from living cells, and that RNA targets of a RBP can then be identified through subsequent steps. Distinct aspects of these protocols include the method of cross-linking, choice of cellular system, choice

of antibody, selection of bound RNAs, sequencing library preparation, and finally the analytical approaches to identify RBP target sites. The methods applied to defined functional consequences of binding events and overall mechanism of protein action are still evolving. These datasets provide a valuable resource to evaluate various ways in which the study of mRNP interactions can be addressed. Regardless of approach, we expect that conclusions resulting from robust biological signals should be consistent.

IMP RBPs are part of the network of stem cell enriched, developmental regulators

Expression of LIN28 is part of the precise network of control necessary for normal development. It is well understood that control of cell fate requires an intricate network of regulators. To understand how the gene targets it regulates fit into the wider context of stem cell control we compared its regulatory network with those of ES enriched RBPs and transcription factors. The approach of studying groups of regulators based on their expression patterns is not new; it was the unique expression patterns and important developmental roles that lead to the fruitful studies of the first pluripotency transcription factors (Scholer et al., 1990; Nicholas et al. 1998; Chambers et al., 2003)**.** Following this logic, along with other evidence from the literature, we selected the stem cell enriched and developmentally regulated IMP (Igf2 mRNA binding protein) family proteins to compare with LIN28 bound transcripts. The IMP family of RBPs are highly conserved from *Drosophila* to mammals and contain six RNA binding motifs (two RRM domains and four KH domains) (Hammer et al., 2005). The *IMP* and *LIN28* genes follow similar expression patterns during

development with high expression confined mainly to ES cells. However, unlike LIN28, the IMP proteins are also enriched in immature neural populations (Figure 20). The IMP or IGF2BP proteins are named in part for their regulation of IGF2 mRNA, which was also the first mRNA found to be translationally regulated by LIN28 (Polesskaya et al., 2007). There is evidence that this is not the only common RNA target of LIN28 and IMP. In particular, the IMP family of proteins co-IP with LIN28 in an RNA dependent manner in proliferating, but not differentiating, human muscle C2C12 cells (Polesskaya et al., 2007). A number of other published reports of IMP and LIN28 proteins physically interacting with each other, as well as with components of the translational machinery (Hafner et al., 2013; Polesskaya et al., 2007; Weinlich et al., 2009). LIN28 and IMP are also related through the miRNA regulatory network, since both proteins are downregulated by let-7 family miRNAs. Furthermore, transcriptome-wide studies have defined LIN28A and LIN28B binding sites in IMP1-3 mRNAs (Hafner et al., 2013; Wilbert et al., 2012). The roles of LIN28 and IMP in development also have parallels, as both are important regulators of neurogenesis, metabolism, and tumorigenes (Balzer et al., 2010; Bell et al., 2013; Christiansen et al., 2009; Nielsen et al., 2001; Perycz et al., 2011; Shyh-Chang and Daley, 2013; Zhou et al., 2013). However, it is not known how widely the RNA targets of these proteins overlap or if interactions are confined only to certain transcripts. To address this question and that of redundancy between RBP family members, we generated RNA binding maps of IMP1 and IMP2. These datasets will also provide insight into the cell-type specificity of RBPs and their binding targets.

Pluripotency is controlled by an overlapping network of regulators

Combining our knowledge of these post-transcriptional regulators with that of transcription factors helps to illustrate how many forces influence the expression of a single gene. Reinforcing gene expression programs through multiple feedback mechanisms helps to ensure precise control required during development. Previous work has demonstrated that transcription factors and miRNAs form central hubs of control in human and mouse embryonic stem cells (Marson et al., 2008). In particular several studies have demonstrated that the reprogramming factors OCT4, SOX2, and NANOG occupy the same sites in the genome (Boyer et al., 2005). OCT4 is a POU domain transcription factor specifically expressed in pluripotent stem cells and its expression is required for maintenance of these cells in the inner cell mass (ICM) of the developing blastocyst (Nichols et al., 1998; Pan et al., 2002). OCT4 can heterodimerize with SOX2, an HMG-box transcription factor, and together they contribute to gene expression changes in ESCs (Botquin et al., 1998; Nishimoto et al., 1999; Yuan et al., 1995). More generally, Sox2 and Oct4 were shown to have widely overlapping target sites in human ES cells which were attributed to binding at *cis* Sox2-Oct4 regulatory elements (Boyer et al., 2005). OCT4 binds the octamer ATGCAAAT and often binds in conjunction with SOX2 at sites with a neighboring SOX binding element (Chambers and Smith, 2004; Pesce and Scholer, 2001). Interestingly where OCT4 and SOX2 DNA-binding sites have been identified, NANOG binding is almost always present as well (Boyer et al., 2005; Marson et al., 2008). Nanog is a homeobox domain transcription factor that also plays a role in maintaining pluripotency; however, it does so in part by preventing differentiation to

the primitive endoderm (Chambers et al., 2003; Mitsui et al., 2003; Wang et al., 2003). This protein is downregulated with embryonic development, maintaining expression in the germ cells and genital ridge as late as E11.5 in mouse embryos (Chambers et al., 2003).

The ability of these factors to support or limit each other's expression levels is a critical component of this pluripotency network. Indeed, these core transcription factors can bind to each others promoters, for example at a composite OCT4/SOX2 motif upstream of the Nanog transcription start site that promotes Nanog expression in pluripotent cells (Kuroda et al., 2005; Rodda et al., 2005). Oct4, Sox2, and Nanog also occupy the promoter region of Lin28A (Marson et al., 2008). The phenomenon has been observed when key transcriptional regulators of a cellular process often regulate other factors in that same pathway, as is seen with these pluripotency factors (Guenther et al., 2005; Lee et al., 2002a; Odom et al., 2004). Similarly, RNA-binding proteins have been shown to preferentially bind mRNA transcripts encoding other RBPs, as was described for LIN28 in our studies (Huelga et al., 2012; Wilbert et al., 2012). However, it has not been studied whether this phenomenon holds true across different classes of regulators that control similar cellular processes. Specifically, we asked if RNAs regulated at the transcriptional level by OCT4, SOX2, and NANOG are subsequently bound by LIN28. Although our analyses reveal that the majority of LIN28 targets were not associated with OSN bound promoters, the subset of genes in common between these regulators are potentially powerful points of control in this network. Our ability to associate single genes with the cascade of regulators directing

cell fate provides useful connections for the understanding of development and how disruption of these pathways result can in disease.

RESULTS

LIN28 CLIP-seq reports produce divergent sets of gene targets

Original approaches to mRNP target identification relied on microarray technologies to detect and quantify isolated transcripts. The hybridization of long transcripts on cDNA arrays limited the resolution of these studies and scope of interrogated target sequences. Pairing high-throughput sequencing with CLIP (CLIP-seq or HITS-CLIP) greatly improved the ability of early studies to define and characterize RNA targets of the miRNA machinery (Chi et al., 2009; Zisoulis et al., 2010). The preparation of size selected cDNAs corresponding to only small RNA fragments protected by the RBP from RNAse digestion are a feature that distinguishes a basic RIP-seq (RNA-immunoprecipitation) approach from CLIP-seq. RIP-seq was employed by Peng and colleagues to determine LIN28 target sites in hESCs (Peng et al., 2011). The combination of RIP-seq data and low-throughput study of LIN28 targets were only successful in defining 95 base pair or larger regions responsive to LIN28 expression (Pend et al., 2011). Furthermore, evidence suggested that multiple binding regions in a single RNA could be active LIN28 targets. For example, in the Oct4 mRNA in mouse ES cells LIN28 binding and activity was detected in the middle of the open reading frame (369 nt long), and the 5' third of the ORF (Qui et al., 2004). These results left questions as to the general binding preference of LIN28 within

protein-coding transcripts. Due to these complexities of *in vivo* binding the ability to use CLIP-seq to generate transcriptome-wide maps at a nucleotide resolution of RBP-RNA interactions were necessary to better characterize LIN28 targeting and function.

The influential roles of LIN28 during early development, neurogenesis, oncogenesis, and somatic cell reprogramming have fueled scientific interests in the mechanism and networks through which this protein acts (Zhu et al., 2010; Balzer et al., 2010; Thornton and Gregory et al., 2012; Yu et al., 2007). In response, a number of groups have published LIN28-RNA interactome datasets obtained through IP-seq (immunoprecipitation coupled with sequencing) based experiments. The typical CLIP-seq approach begins with cells or tissues, cross-linking to form stable complexes, IP of mRNP complexes, digestion of unbound RNAs and size selection of the protected RNA fragments, followed by cDNA library preparation. Here we review some of the variables that contribute to differences in these protocols as they apply to transcriptome-wide LIN28 RNA-binding site definition. The basic approaches employed by each of these recent studies, and their major findings are detailed in

.

We wanted to compare the total target gene sets defined by each other these approaches to build a picture of the impact cellular system, approach, and protein family member (LIN28A versus LIN28B) have on overall conclusions of LIN28 binding and function. Previous comparisons between sets of these studies showed that about 70% of the same transcripts were bound between hESCs and HEK293 cells (Wilbert et al., 2012), and ~50% in common between HEK293 and A2780 (breast

cancer) cell lines (Li et al., 2012). To begin additional comparisons we started by asking how a variation in parameters of our own approach would affect the target genes defined. Using LIN28A hES CLIP-seq sequencing information generated in Wilbert et al., 2012, we re-analyzed this data removing any reads that has the same start and stop location as another in an effort to safeguard against potential PCR duplication artifacts. This more stringent analysis filtered out a number of genes, resulting in an overall smaller dataset that had a higher percent of targets in common with the LIN28-V5 target set generated in HEK293 cells (Figure 21A). This is evidence that our most confident hESC LIN28 binding sites are those also recapitulated in a second cell type adding to the growing body of evidence that LIN28 is a promiscuous binder, not highly dependant on co-factors for binding and able to bind its target transcripts wherever they are expressed. The re-processing of the LIN28 hES dataset also demonstrates how a single parameter can drastically shift the population of significant gene targets.

Our LIN28-V5 (LIN28A) HEK293 dataset (Wilbert et al., 2012) and LIN28A PAR-CLIP by the Tuschl group had the most similar experimental setup (Hafner et al., 2013). However, when we compared overlaps of these gene targets (Figure 21B) there was no correlation between the two HEK293 LIN28A datasets versus the hESC dataset. Furthermore, the HEK293 LIN28B set of gene targets shared the same relative amount overlap between any of LIN28A target sets. Proportionately more targets were in common between the LIN28A HEK293 gene set (Hafner et al., 2013) and either of two LIN28B HEK293 gene targets sets than there were between the two LIN28B sets

(Figure 21C). In the original study of LIN28A and LIN28B in the same overexpression system in HEK293 cells the authors reported that almost all LIN28A target mRNAs (1,674/1,803; 93%) were also bound by LIN28B (Hafner et al., 2013). From this cumulative information it is likely LIN28A and LIN28B do bind an overlapping set of gene targets; but the variance in these transcriptome-wide methods make gross comparisons less informative.

The largest LIN28A gene set from mouse ES cells (~60% of the mouse transcriptome Cho et al., 2012) had about ~60% of target in common to RIP-seq and CLIP-seq hESC (Figure 21D) or to LIN28A and LIN28B targets in HEK293 cells (Figure 21E). Comparing the junctions of genes in common further confirmed these sets were made from different groups of genes overall. That is, a single consensus LIN28 target gene set cannot be drawn from the intersection of these multiple data sources. Therefore, to define which of these targets are confident binding sites and those that have functional response we must rely on rules learned from each target set as a whole.

The CLIP-seq approach is compatible with LIN28 biology to define binding motifs

Traditional IP-based methods suffer from artifacts of spurious mRNP interactions that occur in cell lysates but not *in vivo* (Mili and Steitz, 2004). The incorporation of cross-linking into these protocols to bind together RNAs and closely associated RBPs helps to prevent this problem and enables stringent selection steps for the mRNP complex(es) of interest. A popular method of cross-linking is the use of UV

irradiation to form covalent bonds between proteins and nucleic acids within one Angstrom apart (Ule et al., 2003). A caveat of UV cross-linking it that is relies upon the close proximity of nucleic acid bases and aromatic amino acid side chains since these are the photo-reactive groups of the RNA and protein molecules. Due to this, proteins that interact with the phosphate backbone of RNA, like eIF4AIII a DEAD-box helicase component of the exon junction complex, are less efficiently cross-linked (Singh et al., 2013). However, successful application of CLIP-seq to this protein shows that with the use of sufficient starting material and careful sample processing traditional UV cross-linking can be successful even for this type of RBP-RNA interaction (Sauliere et al., 2012). The use of formaldehyde is another popular cross-linking agent as it does not rely on these criteria and instead non-specifically preserves native RNA-protein interactions (Niranjanakumari et al., 2002).

Studies of the interactions between LIN28 and the let-7 family miRNA precursors have collectively led to our understanding that the site of binding occurs at conserved GGAGAU and degenerate GNGAY (Y = pyrimidine) sequences within the hairpin loop. This suggested that LIN28 would be an ideal candidate for transcriptome-wide CLIP-seq study since these nucleotides are effectively cross-linked, as are sites within unpaired loop regions of ssRNA. However, it was initially unknown whether LIN28 would also target this sequence-structural element in mRNAs or other target transcripts. The first CLIP-seq experiment for LIN28A in hESCs revealed that this is indeed the case (Wilbert et al., 2012). That is, LIN28 preferentially binds mRNAs at GGAGA motifs embedded within unpaired regions of

loop structures. A second study in mouse ES cells recapitulated these findings for the Lin28a homologue at a degenerate motif AAGNNG or AAGNG, that encompass the human LIN28 mRNA binding motif GGAGA (Cho et al., 2012). A UGUG motif was also observed, albeit less frequency. The most commonly observed tetramer from LIN28B studies was GGAG (G/A)GG(G/C)(A/U)G (Graf et al., 2013).

PAR-CLIP utilizes photoactivatable-ribonucleosides to enhance cross-linking efficiency and mark the precise site of binding (Hafner et al., 2010; Hafner et al., 2013). Incubating cells with the photoactivatable ribonucleoside 4-thiouridine (4SU) or 6-thioguanosine (6SG) prior to UV-irradiation results in their incorporation into newly transcribed transcripts. The effect of this is that cross-linked sites can be distinguished by thymidine to cytidine (T to C), or guanosine to adenosine (G to A) transitions in the cDNA prepared from extracted RNA. These single base pair changes in sequence are then used as evidence to define cross-linked sites and specific RBP bound sites. Despite differences in CLIP and PAR-CLIP they generally produce similar results. Hafner and colleagues have applied PAR-CLIP and the tool PARalyser (Corcoran et al. 2011) to identify peaks of LIN28A or LIN28B binding. Their criteria for T-to-C transitions were that each peak had at least two independent transitions, at least 5 reads with T-to-C conversions and overall conversion rate ≥0.25. In their 500 top binding sites, as ranked by cross-linking transitions, found a pyrimidine-rich degenerate motif AYYHY (Y = U,C and H = A,C,U) in 93% of all LIN28B binding sites. Since the target site definition was always centered on T-to-C transitions this may account for a higher frequency of these bases in the motif defined. Even without

the use of photoactivatable ribonucleoside Chao and colleagues were able to utilize mutations induced by cross-linking to define nucleotide sites where the protein interacted, almost always centered on a G (Cho et al., 2012). In these comparisons the PAR-CLIP approach does not enhance LIN28 target site definition and motif finding.

Sequence information reveals structure of LIN28 mRNA interactions

Although *in silico* predictions of RNA secondary structure are notoriously variable, information from LIN28 CLIP-seq binding has enabled the study of structural constraints on its mRNA targeting (Cho et al., 2012; Wilbert et al., 2012). Cho and colleagues calculated the simple Watson-Crick base-paring occurrences for nucleotides surrounding LIN28 bound GGAG motifs and found evidence for hairpin structures at these sites. More sophisticated modeling by Wilbert and collaborators using the RNAplfold algorithm to calculate the probability of secondary structures also produced evidence that LIN28 preferentially binds to unpaired regions of secondary structure with mRNA transcripts (Wilbert et al., 2012). Conclusions from these studies are useful when applied to understanding and studying LIN28 functional binding to individual RNAs at specific sites. For example, the necessity of transcripts to form secondary structures before they can be targeted by LIN28 may help to explain why luciferase reporter constructs shorter than 95 base pairs were unable to stimulate translation in the presence of Lin28 expression (Peng et al., 2011). To aid in the design of ideally positioned LIN28 responsive sites for luciferase reporters we included secondary structure predictions in our criteria, in addition to the presence of

strong CLIP-seq binding and some variation of the LIN28 binding motif GGAGAY and/or GAU rich sequences (Figure 22).

Individual RNA-binding domains can be studied with PAR-CLIP

While basic CLIP-seq and PAR-CLIP studies provide comparable information about RBP binding, an adaptation of PAR-CLIP that enables IP of individual RBP binding domains provides a more detailed description of RBP-RNA interactions. Previous studies with CSD and/or ZKD mutations have demonstrated that both are essential for proper regulation of miRNA and mRNA targets of LIN28 (Balzer et al., 2010; Balzer and Moss, 2007). In order to define transcriptome-wide binding patterns of the LIN28B protein Graf and colleagues used another modified version of the protein, this time with a cleavable peptide within the flexible linker region of LIN28B and tags on either C- or N-terminus allowing the selective IP of either domain and bound RNAs (Graf et al., 2013). A useful aspect in the design of this experimental system is the ability of the protein to bind under endogenous conditions with both functional binding domains before their disassociation. This is particularly relevant when we consider lessons from LIN28-let-7 interactions that suggest initial binding of the ZKD helps to open up the RNA structure and allow for subsequent binding of the CSD, and possibly additional molecules of LIN28 (Nam et al., 2011). Their results shows that both domains interact together on the same RNA within close proximity to each other, likely with the CSD upstream of the ZKD (5' to 3' orientation on the mRNA) (Graf et al., 2013). Here, we see a valuable adaptation of the original CLIP-seq protocol to facilitate understanding of individual RBP binding domains. However,

it remains possible that deeper analyses of CLIP-seq data may be able to produce similar findings.

In vitro assays complement CLIP-seq studies

As discussed above, the application of CLIP-seq based technologies to LIN28 studies have returned consistent information about sequence and structure of mRNA transcripts it interactions with. However, some information my be lost in these analyses and warrant complementary systems of study. For example, *in vitro* binding assays by the Tuschl group determined that ssRNAs ≥18 nt with $(U)_{18}$, $(GU)_{18}$, or $(AG)_{18}$ sequences were optimal LIN28 targets (Hafner et al., 2013). In our own studies we concluded that poly(U) sequences were the least prevalent in our LIN28 target sites. This discrepancy likely results from several inherent limitations of the CLIP-seq approach. One difficulty for all sequence studies is mapping reads to repetitive regions of the genome and areas of low sequence complexity. Since we mask repetitive regions of the genome when we assign LIN28 CLIP-seq reads those areas are not reported and our ability to identify target sites within these sequences decreases. This is an example where a complementary approach can complete our understanding of RBP-RNA interactions. In this case, electrophoretic mobility shift assays (EMSA) were able to confirm motif preferences revealed by CLIP-seq and also to extract additional information about low-complexity sequences that the high-throughput approach may miss (Hafner et al., 2013).

LIN28 has subtle effects on mRNA

The objective in defining direct binding sites of RBPs is to enable study of direct regulatory mechanisms versus downstream effects. The most accessible approach for transcriptome-wide study of RBP function is to assay impact on RNA levels in response to changes in RBP expression. The overall conclusion from LIN28 studies is that this protein has only a small affect, if any, on transcript levels. For example, only 152 genes in A2780 breast cancer cells had greater than 1.5 fold change in mRNA levels upon LIN28 knockdown (Li et al., 2012a). By comparing knockdown to overexpression of LIN28B in HEK293 cells (as opposed to a control base line condition), Hafner and colleagues detected a slight but significant increase in LIN28 bound RNA transcript levels, in the presence of LIN28B (Hafner et al., 2013).

Our own studies demonstrated that there is not a significant enrichment of LIN28 bound transcripts within groups of genes up, down, or unchanged in response to LIN28 expression changes (Wilbert et al., 2012). In response to the Hafner report (Hafner et al., 2013) that did find a slight but significant ability of LIN28B expression to stabilize mRNA levels, we revisited our RNA-level changes in response to LIN28A. RNA-seq gene expression changes from LIN28A knockdown in hESCs demonstrated that if we disregard direction of change, genes with changes in RNA expression level were enriched for LIN28 bound targets, as compared to non-targets (chi-square = 57, P = 4.11e-14). However, the vast majority of LIN28 hES targets (n = 5,491, 92%) did not exhibit changes at the RNA level. In addition, cumulative distribution analysis of LIN28 CLIP-seq bound genes versus non-targets demonstrated

that the amount of change for LIN28 targets was less than non-targets (two-sample Kolgomorov-Smirnov test, p = 2.7e-64) (Figure 23A). Together, these results indicate that while some genes with changes at the RNA level upon LIN28 knockdown may indeed be functional LIN28 targets, the primary mechanism of LIN28 regulation is not an affect on RNA steady state levels.

In agreement with this, results from microarray experiments in the presence of LIN28-V5 expression showed that only 24 genes had increased and 240 decreased RNA expression levels (by 2-fold or more) upon LIN28-V5 expression in 293 cell**s**, leaving 94% of targets unchanged. Comparing the genes that decreased upon LIN28 knockdown with those that responded to LIN28-V5 expression few reciprocal events were found (<10 in common). This is due in part to the differences in the assays used, but also possibly due to differences in cell-type and directionality of LIN28 modulation. In support of the idea that LIN28 may have different effects upon overexpression versus knockdown, for example, cumulative distribution analysis of LIN28-V5 CLIP-seq target genes in HEK293 cells demonstrated greater decrease in RNA levels than non-targets (two-sample Kolgomorov-Smirnov test, p = 1.42e-110) (Figure 23B).

Another piece of information that complicates our understanding of LIN28 function upon overexpression in HEK293 cells is the apparent ability of LIN28 to regulate translation in a dose-dependent manner. We demonstrate that different levels of LIN28-V5 exogenously expressed in HEK293 (Flp-In-293) cells results in an increase in protein production of the splicing factors HNRNP F and TDP-43 (Figure

24). However, once a saturating level of LIN28-V5 is reached the impact on its targets is a decrease in protein expression. The mechanism behind this response, and whether it is the direct result of more LIN28 binding or the result of feedback mechanisms remains to be tested. Dose-dependent effects of LIN28 have been reported in knockout mouse models where *Lin28a+/-b-/-* mice had an increased effect on postnatal dwarfism as compared to *Lin28a+/+b-/-* (Shinoda et al., 2013). More drastically, while mice haploinsufficient in one Lin28 paralog survive with knockout of the other, double kockout embryos died by E12.5. The impact of dose on LIN28 regulation indicates discrepancies reported in its regulation of OCT4 in ES cells could be the result of variance in the level of LIN28 knockdown (Cho et al., 2012; Darr and Benvenisty, 2009; Qiu et al., 2010; Xu and Huang, 2009; Xu et al., 2009).

Despite the limited number of RNA expression level changes, looking at those genes that were regulated is still informative of the downstream pathways that LIN28 may affect. For instance, upon LIN28 knockdown we see a reduction in RNA levels of genes important for "bone development" (GO:0060348, $P = 0.03$), and "muscle system process" (GO:0003012, $P = 0.003$), and an increases in those regulating "neuron apoptosis" (GO:0043523, $P = 0.08$), as indicated by enrichment in GO analyses, processes that this protein is known to affect (Polyesskaya 2007, Balzer 2010) (Table 5A). Interestingly upon LIN28-V5 overexpression, it was also downregulated gene targets that had the most significant enrichment in gene ontology and group of factors included transcriptional regulators and DNA binding factors (Table 5B). This could be evidence of antagonistic roles between LIN28 and DNA-

binding factors, a concept that would contrast our hypothesis that LIN28 might cooperate with OCT4, SOX2, and NANOG through common mechanisms in reprogramming.

LIN28 affects protein levels

Lack of convincing RNA-level control by LIN28 and evidence that it cooperates with translational machinery in active polysomes made the study of its global impact on translation of target mRNAs a priority. Two accessible approaches for changes in proteome measurements are mass spectrometry or the more recently developed method of ribosomal profiling. A commonly used method for quantitative measurement of peptides is the use of stable isotope labeling by amino acids in cell culture (SILAC). This approach uses mass spectrometry to measure mass ratios of protein peptides from control cells versus SILAC labeled cells (Ong,S.E. 2002; Zhu,H. 2002). Early miRNA studies successfully employed SILAC to build models of miRNA regulation (Vinther,J. 2006; 239 Baek,D. 2008; 211 Selbach,M. 2008). Using pulsed SILAC (pSILAC) Graf and colleagues demonstrated that LIN28B coding-sequence binding promotes translation of targets in HEK293 cells. Similar SILAC experiments form the Tuschl group confirmed this positive impact on protein production from LIN28B bound transcripts (Hafner et al., 2013). However, the power to detect these changes required comparison of knockdown and overexpression data (as opposed to an unaffected control). Furthermore, the relative influence on mRNA levels was not accounted for.

Another approach for genome-wide affects on translation is the use of ribosomal occupancy on RNA transcripts as a surrogate for translational state. Ribosomal profiling, or Ribo-seq, relies on the protection of small ~30nt sequences by the translating ribosome when lysates are treated with RNase (Guo et al., 2010; Ignolia et al., 2009). Isolating and sequencing these transcripts generates a map 'footprints' where ribosomes were loaded on the mRNA (Figure 25). This method has the advantage that impact of RNA-level changes can readily be subtracted through the incorporation of RNA-seq information from the transcriptome. The Kim group used this technique in mouse ES cells to detect Lin28a dependent changes in translational state (Cho et al., 2012). Their work identified a group of endoplasmic reticulum proteins that are repressed at the level of translation by Lin28a in the peri-nuclear region.

We conducted our own studies of protein abundance in response to LIN28-V5 expression in HEK29 cells. Unlabeled lysates from control HEK293 or cells expressing LIN28-V5 were combined with a SILAC labeled control lystate that was used to for quantification for normalization between the samples. Four replicate biological samples were prepared for the LIN28-V5 cells and the peptides detected in each were largely from the same population (Figure 26) enabling quantification of ~2,000 proteins. Three replicate control samples from HEK293 (Flp-In-293) cells were used in this analysis. Using protein expression changes from this data, along with RNA level changes from LIN28-V5 versus control profiled on mircoarrays (Wilbert et al., 2012), k-means clustering was performed to define groups of genes related by

relative RNA and protein level changes (Figure 27A). We created a ranking system for CLIP-seq binding based on the percent of reads that fell in clusters versus all reads found within a transcript. The concept is that more reliable RBP targets are those with the majority of reads within significant binding sites, as opposed to randomly dispersed reads. Using groups of genes defined in Figure 27A cumulative distribution analysis was used to test if any were significantly enriched for CLIP-seq binding (Figure 27B). Two gene sets were significantly different than all genes profiled. One group (Figure 27A, group i) had little change at the RNA-level but large decrease in protein expression in the presence of LIN28 expression. This group was under-enriched for CLIP-seq binding indicating these were transcripts less bound by LIN28 (Figure 27B, group i). The second group (Figure 27A, group ii) had little RNA-level changes detected, but a slight increase at the protein level. These genes had significantly more CLIP-seq reads within significant binding sites. However, the average fold-change in protein for this group was only about a ~ 1.2 fold. The biological significance of this small amount of change is difficult to gauge, but nonetheless is consistent with the slight effects confirmed by western blot analysis of LIN28 target genes (Cho et al., 2012; Graf et al., 2012; Hafner et al., 2013; Wilbert et al., 2012). From this we conclude that LIN28A binding can have a slight but significant effect of increasing protein production from target transcripts, consistent with reports described above for LIN28B.

LIN28 and let-7 share an overlapping network of gene targets

Among the mRNA targets of LIN28 we defined in hESCs (Wilbert et al., 2012), we identified an overrepresentation of predicted let-7 target genes (n = 355 TargetScan predicted targets, Z-score = 14.25). This regulation generates a potential a feed-forward model where LIN28 performs synergistic regulation of its targets, directly via translational enhancement, and indirectly, via elimination of let-7 repression. This is in keeping with other feed-forward loops observed in pluripotent cells, for example between transcriptional activation of LIN28 by c-Myc, which in turns blocks let-7 processing, thereby releasing let-7 repression of c-Myc (Kim et al., 2009). These antagonistic mechanisms of LIN28 and let-7 provide reciprocal means of regulation and evolve in order to provide robustness to cellular systems. This precise control is especially important in the regulation of critical developmental pathways. Our understanding of counter regulation by LIN28/let-7 of the same gene targets may explain why muscle specific knockout of Lin28a in transgenic mice phenocopies induction of let-7 in transgenic mice, and yet let-7 levels were unaffected in the knockout muscle (Zhu et al., 2011). Specifically, Lin28a transgenic mice exhibited insulin-sensitivity through regulation of the insulin-PI3K-mTOR pathway that rendered them resistant to high-fat-diet induced diabetes. Several genes shown to be critical in this effect, for example mTOR, AKT1 and AKT2, all of which are direct LIN28 targets by CLIP-seq. In our studies of LIN28 impact on RNA-level changes in response to LIN28 knockdown, we wanted to know if changes in let-7 might be impacting our results. However, of the transcripts downregulated upon LIN28 knockdown in hES cells, only 13 of these genes were predicted let-7 targets by

TargetScan, indicating that expression of this miRNA most likely did not contribute significantly to mRNA level changes observed. While not every gene with LIN28 or let-7 binding sites will be a functional target under all cellular conditions, the overlap in target genes suggests a feed-forward mechanism between LIN28 and let-7 by which LIN28 directly binds to and supports translation of mRNA targets, while simultaneously blocking let-7 mediated repression of these genes.

IMP family RBPs present an opportunity for comparative studies in hESCs

To emphasize the utility of our model of RBP interrogation, we identified another family of RNA-binding proteins that are highly expressed early in development and play important roles in early cell fate specification. We noted that among other evidence in the literature that LIN28A and IMP1 may be related that they also share an oncofetal expression pattern (Figure 20). *IMP1* expression is confined to early, undifferentiated cells and *IMP2* is expressed ubiquitously across early and late cell types. This expression profile was not accurately represented by RNA levels alone in Figure 20. Here, we draw parallels between the oncofetal expression pattern of IMP1 and LIN28A, versus evidence that LIN28B and IMP2 may be more influential later in adult tissues (Gaytan et al., 2013; Shinoda et al., 2013). Furthermore, there is *in vivo* and *in vitro* evidence that members of both these protein families have important roles in metabolism, oncogenesis, and neurogenesis. By studying another family of RBPs in pluripotent stem cells we are able to compare and contrast differences within and between RBP families.

To define the transcriptome-wide binding sites of the IMP1 and IMP2 proteins we used the approach of cross-linking and immunoprecipitation of RNP complexes followed by high-throughput sequencing of bound RNA (CLIP-seq) as in (Wilbert et al., 2012). Significantly bound sites where defined where clusters of reads passed both transcript and transcriptome cutoffs determined by a Poission statistic. This analysis found 27,130 IMP1 bound regions in 9,297 gene transcripts, and 5,584 IMP2 bound regions in 3,029 gene transcripts. We found this set of target genes to be largely the same, with more than 85% of IMP2 targets (2,597/ 3,029) found within the set of IMP1 bound genes ($p < 0.001$, by hypergeometric test) (Figure 28A). A random intersection of these datasets, within the background of expressed transcripts in hESCs would only be expected to result in about half this overlap (mean=1,303; std= 24). This overlap was not entirely unexpected since almost identical target sets of the IMP proteins were reported previously; however, these sites were generated using the PAR-CLIP approach in HEK293 cell lines overexpressing these proteins individually along with a FLAG-tag used in immunoprecipitation (Hafner et al., 2010) (Figure 28B). Furthermore IMP2 has been shown to act within the nucleus to help load IMP1 onto target mRNAs (Pan et al., 2007). These comparisons produced results similar to our overlaps of LIN28A and LIN28B targets. That is, IMP1 and IMP2 family members seem to be able to target similar sets of transcripts in different cell-types (Figure 28C and D), and both family members bind a common set of genes (Figure 28A and B). In particular, the almost identical target set of IMP2 binding in hES and HEK293 cells might be related to the fact that it is more commonly expressed across different cells

and therefore may be conditioned to bind and regulate the same targets in different environments.

IMP1 and IMP2 bind RNA transcripts in different contexts

Despite the correspondence of gene targets between IMP1 and IMP2, the distributions of their binding sites were strikingly different (Figure 29). IMP1 clusters were found distributed in 3' UTR and intron regions of target RNAs, while IMP2 clusters were almost entirely found in 3' UTRs. This suggests IMP1 binding in the nucleus when nascent pre-mRNAs are being processed. Previous studies have demonstrated that this is indeed the case. The strong preference for IMP2 binding in the 3' UTR does not preclude the possibility that it also interacts with RNAs in the nucleus as has been reported (Pan et al., 2007); however, if introns are present in IMP2 bound transcripts our data shows direct binding events do not take place within them. When we examined the clusters within common targets from each dataset to see if they overlapped in genomic coordinates by one nucleotide or more, we found only 8.4% of IMP1 clusters in the same position as IMP2 clusters (26.5% of IMP2 clusters). At least a fraction of the non-overlapping IMP1 clusters can be contributed to those found within introns. The possibility remains that IMP1 and IMP2 binding sites are related in some manner within 3' UTR regions. This discrepancy in IMP1 and IMP2 binding adds to evidence that the three paralogs of IMP all have different functions, related in part to differences in their expression patterns and RNA-binding properties (Wachter et al., 2013).

Inspecting the 5' UTR more closely we found that IMP1 clusters were enriched above what would be expected by random in the transcriptome (Figure 30). This is in contrast to LIN28A CLIP-seq clusters that are underrepresented in this region in our data and others. We defined two classes of 5' UTR exons as constitutive or variably transcribed as would occur in the case of alternative transcription start sites. IMP1 clusters occurred more frequently in variably transcribed 5' UTR exons. One known example of translational control occurring at the 5' UTR region is the regulation of IGF2 (Insulin-like growth factor II), a known target of IMP1 (Nielsen et al., 1999). Alternative isoforms of IGF2 transcripts have different 5' UTR regions and are translationally regulated by different mechanisms. Some IGF2 transcript 5' UTRs have IMP1 biding sites this protein uses to control localization of the IGF2 translation. Further experimentation is required to test if accessibility of IMP1 binding sites are widely controlled through regulation of alternative transcription start sites or to test the potential role of IMP1 in regulating transcription state site usage. Although IMP2 only showed a slight preference for binding in 5' UTRs as compared to the transcriptome background, these sites are also enriched within variably transcribed exons. Interestingly, when we assigned IMP1 PAR-CLIP binding sites (Hafner et al., 2010) to our mutually exclusive gene annotations, this inclination for 5' UTR binding was recapitulated, indicating differences in their analyses prevented them from detecting this enrichment (see discussion).

The most enriched motifs found at IMP1 and IMP2 binding sites also showed marked differences. Specifically, when the most enriched hexamers in IMP1 bound

clusters was GGACUG (Z-score = 49, $p < 0.0001$), while IMP2 sites were enriched for AAUAAA (Z-score = 27, $p < 0.0001$). The sequence AAUAAA enriched within IMP2 bound sites is the signal for polyadenylation. This is consistent with almost exclusive localization of IMP2 binding events to the 3' UTR. It is interesting to speculate how this might be related to IMP2 regulation, either in determining polyA site usage or possibly affecting the ability of IMP2 to bind and regulate tissue specific transcripts. We know that polyadenylation usage is one approach used by ubiquitously transcribed genes to achieve tissue-specific protein expression (Lianoglou et al., 2013).

The sequence 'GGACUG' makes up the core of the known IMP 'zipcode' binding sequence described in the 3' UTR of B-actin conserved from chick to humans and within the cMYC "CRD" sequence bound by IMP (Ross et al., 1997; Bernstein et al., 1992). CLIP-seq binding data at the ß-actin 3' UTR shows that we are able to perfectly identify this site within IMP clusters (Figure 31). It is worthwhile to note that when we increased the stringency of our parameters and considered only unique reads (removing possible effects from PCR duplication) this exact site no longer maintained significant signal. Here we present an example of the effect variations in CLIP-seq data processing can have on binding site definition. Structural studies from Chao and colleagues demonstrate the spacing of the sequence CGGACUG between ACA motifs within 7-30 nucleotides on either side represents ideal binding conditions for the KH3 and KH4 domains of IMP1 (Chao et al., 2010).

We were curious to know how IMP1 and IMP2 targets might be related to regulation in stem cells. We found evidence that these proteins bind to various stem cell enriched transcripts, such as LIN28 and SOX2. Within the SOX2 3' UTR we found an ideally situated IMP binding motif with 'GGACU' positioned between two ACA rich sites (Figure 32). Downstream from this was an instance of IMP2 binding at the poly(A) site. We noted that these sites were recapitulated in the PAR-CLIP data from HEK293 cells (Hafner et al., 2010). Knockdown of IMP1 in hESCs results in an increase in SOX2 expression (Conway et al., in preparation). To understand if the roles of IMP1 and IMP2 in post-transcriptional regulation in hESCs are connected to LIN28 regulation through direct RNA targets we compared the genes sets with significant CLIP-seq clusters from all three proteins (Figure 33). IMP1 target genes included as great a proportion of LIN28 bound genes as it did those of IMP2, suggesting related pathways of action. Not surprisingly among the genes commonly bound we found a large proportion of RNA regulating factors, and genes involved in neurogenesis and cell mobility. Further functional studies of gene expression changes in response to these RBP regulators will help to elucidate their possible cooperation or antagonist impacts on bound transcripts.

Control of pluripotency is established through interconnected networks of regulators

It is well understood that control of cell fate requires an intricate network of regulators. We wanted to know how RBPs integrate into this network, and if we could test the hypothesis that core pluripotency transcription factors OCT4, SOX2, and NANOG (OSN) widely drive transcription of RNAs important for pluripotency that

are subsequently affected downstream by LIN28 binding. Since the transcriptional regulators of pluripotency have been extensively studied, we were able to use published data to define OCT4, SOX2, and NANOG co-regulated genes. Chromatin immunoprecipitation coupled with high-throughput sequencing (ChIP-seq) is a well-tested method to determine the DNA-binding sites of transcription factors and other DBPs (Johnson et al., 2007). We utilized high quality ChIP-Seq datasets generated for OCT4, SOX2, NANOG, KLF4, p300 and TAF1 by the Ecker and Ren groups (Lister et al., 2009). The protein p300 is a histone acetyltrasferase that acetylates all four core histones in the nucelosome and that commonly binds at enhancers, leading to its use as a marker for these regions (Heintzman et al., 2009; Ogryzko et al., 1996). TAF1 is used as a surrogate for the transcription initiation complex that it is part of.

The advantage of using data from a sequencing-based approach like this over an array platform is a less biased sampling of the genome and potentially less biased approach for assigning DNA binding peaks to genes they regulate. We used genomic coordinates of ChIP-Seq peaks from Lister et al., 2009, and converted these hg18 coordinates to hg19 to be compatible with the most updated genome information and our CLIP-seq datasets described above (Lister et al., 2009). Next we used ENSEMBL gene annotations and assigned transcription factors binding sites within 800 base pairs downstream to 200 base pairs upstream of the transcription start site (TSS). Using this window we assigned 3889 OCT4 bound sites to 1,538 genes, 5682 SOX2 bound sites to 1,538 genes, 25,071 NANOG bound sites to 1,538 genes, 3794 KLF4 bound sites to 1,538 genes, 3093 p300 bound sites to 1,538 genes, and 12,362 TAF1 bound sites to

1,538 genes (Figure 34A). The -800, +200 bp window we chose was rather conservative since it is known that transcriptional regulators can bind much further from transcriptional start sites that they affect. To test whether we had assigned binding events to the majority of transcription start site proximal promoters, we expanded our window around the TSS and asked how many gene targets were now assigned to each transcription factor. Expanding our window to 2000 base pairs downstream and 500 base pairs upsteam increased the number of targeted genes by less than 1% for the pluripotency associated TFs (Figure 34A). Looking even further away from 10kb to 2kb around the TSS increased OCT4 target gene numbers by almost 13% to 1,762 sites bound compared to 1,538 bound sites within 800-200 bp. These results tell us that we have identified that majority of promoter proximal sites and that a population of TF sites exists outside of these regions.

Since it is known that miRNAs and other non-coding RNAs are targets of these TFs and RBPs in the regulation of pluripotency networks, we wanted to know how many of our bound genes were protein-coding (Marson et al., 2008; Xu et al. 2009). We found up to 26% of the TF binding sites were assigned to gene types other than coding mRNAs, indicating a significant portion of targets were indeed proximal to other categories of transcripts (Figure 34B). A smaller portion of the gene targets in common between OSN were non-coding (13%), and almost all (44/55) of the commonly targeted transcripts by OSN and LIN28 were protein-coding. We conclude that we have effectively assigned TF DNA-binding sites to promoter proximal regions of primarily protein-coding genes. Our numbers and relative proportions of assigned

gene targets are comparable to a similar study by Boyer and colleagues that found only 3% of OCT4 (623), 7% of SOX2 (1271), and 9% of NANOG (1687) bound sites within promoter regions known protein-coding genes (Boyer et al., 2005). Bound mRNA information from CLIP-Seq was used to define 2,782 LIN28, 9,297 IMP1, and 3,029 IMP2 target genes under stringent filtering parameters.

Previous microarray-base data in human ES cells reported that about half of the OCT4 target sites in the genome are also SOX2 sites due to their overlapping binding sites (Boyer et al., 2005). Furthermore that more than 90% of the OCT4-SOX2 sites were also bound by NANOG in this study (Boyer et al., 2005). Curiously, we did not see as strong of an overlap between OCT4 and SOX2 with only about 15% of total gene targets in common (Figure 34C and Figure 35A). However, almost all of these genes were also NANOG targets (216/227) (Figure 34B and C, and Figure 35). Looking at the set of OSN target genes the most prevalent biological function was transcriptional regulation (~40 genes) in keeping with the expected enrichment of transcription factors to be among these common targets (Boyer et al., 2005). Importantly the most significant group of transcriptional regulators were involved in 'negative regulation of transcription' in a DNA dependent function ($p<0.0001$, Benjamini corrected value=0.033). These factors may work together with OSN to repress transcription of lineage specific factors, which was an outcome reported for a group of genes with OSN promoter occupancy (Boyer et al., 2005). We also identified another functional group of proteins commonly bound by OSN that surpassed genome-wide statistical cutoffs. This was a group of 14 proteins involved in "negative

regulation of RNA metabolic process" ($p<0.0001$, Benjamini corrected value=0.029), indicating the pluripotency transcription factors may also shut down competing RNA post-transcriptional regulation, in addition to lineage specific transcriptional regulators.

Only 20% of OSN bound gene promoters (45/ 216) had evidence of LIN28 binding to mRNAs transcribed from these loci, demonstrating that only a small subset of LIN28 targets are a part of this network (Figure 34B, D, E, and F). None of the tertiary comparisons of two transcription factors and LIN28 target genes indicated enriched coordination through the same targets either (Figure 34D, E, and F). To investigate the possibility that any one of the TFs is dominant in driving transcription of LIN28 target genes, or the developmentally regulated IMP1 and IMP2 target genes, binary comparisons of gene targets for all these proteins were performed (Figure 35). The percent of genes in common between any two proteins was largely proportional to the overall dataset sizes. Returning to the original ChIP-seq datasets and larger network of LIN28 targets as published for the hg18 genome build did not significantly alter these conclusions, with the exception of finding better correlation between TAF1 and LIN28, as expected since a gene must be transcriptionally active (indicated by TAF1 binding) in order for its mRNA to be available for LIN28 targeting. We may have expected to see a larger percent of SOX2 and LIN28 genes in common since these proteins may act together in the nucleus and together coordinate neural differentiation (Cimadamore et al., 2012; Cox et al., 2010). Further functional information will be required to determine what, if any, influence the small hubs of

overlap between these RBPs and TFs have on commonly regulated pathways. Looking to the list of genes in common between OSN and LIN28 (Table 6) suggests some possible effects of coordinated regulation could generally impact cell cycle (CDK6), translation (EEF1A1), or cellular metabolism (AGPAT4, UST, PLOD2, CHCT9). However, our initial analyses show that widespread overlap of common gene targets does not seem to occur between stem cell enriched transcriptional and post-transcriptional regulators.

DISCUSSION

Different CLIP-seq datasets are a rich information source for RBP studies

Despite minor difference in the details of each approach the large-scale studies recently completed for LIN28 suggest the same distinguishing features of its interactions with RNAs. However, the overall gene target sets from each individual experiment were not completely overlapping. Differences in LIN28 CLIP-seq datasets can arise from a number of sources. One obvious difference in the Chu et al., 2012 dataset is that it was generated from mouse ES cells rather than human. It has long been recognized that human and mouse ES cells are not equivalent in gene expression and differentiation potentials (Rizzino, 2002). For example stage-specific embryonic antigen 1 (SSEA1) is a marker of undifferentiated mouse ES cells, but is only expressed in differentiated human cells (Thompson et al., 1998 (Henderson et al., 2002; Reubinoff et al., 2000). In addition, mouse ES cells are critically controlled by leukemia inhibitor factor (LIF) and BMP signaling pathways (Niwa et al., 1998; Ying

et al., 2003). However, under the same conditions BMP4 will cause hESC differentiation and LIF alone cannot maintain the pluripotent state of these cells (Daheron et al., 2004; Xu et al., 2002). Instead, important factors for hESCs are FGF and a balance of TGFb/Activin and BMP signaling (James et al., 2005; Vallier et al., 2005; Xu et al., 2005). However, these differences alone cannot account for the large group of gene targets unique to Lin28a mESC as defined by Cho and colleagues (Cho et al., 2012).

More likely, variations in LIN28 RNA-bound maps reflect promiscuous binding of the protein to a large set of transcripts. The ability to identify essentially the same sequence and structure characteristics for LIN28-mRNA interactions indicates that the bulk of transcripts detected by any of the given studies represent bona fide interactions. This cumulative knowledge supports a more general role for LIN28 to act in the formation of RNA scaffolds and facilitate in the binding of other co-factors.

The underlying algorithms, assumptions, and filters used in analysis of global datasets can all influence the conclusions drawn from this information. For example, Graf and colleagues used a priority system when assigning LIN28 binding sites to genic regions. If multiple Refseqs diverged in annotation at a site (e.g. one transcript has a coding region, the other a UTR) these authors assigned a single annotation for that position; choosing in order of preference CSD over 3' UTR or 5' UTR, and any of these regions above intron. On the other hand our gene models used an aggregate system to unambiguously assign annotations based on the intersection of multiple transcripts. That is, at each position of the genome we store information from all

overlapping isoforms such that we could distinguish between uniquely defined regions (e.g. coding-exon in all transcripts) versus variable regions, for example that result from alternative transcript ends (see Figure 36 and Figure 37). Either approach will have its advantages and limitations. By extracting more information we have the potential to identify subtle events that would be missed otherwise. For example, in the analysis of IMP1 enrichment in 5' UTRs we could distinguish between exons that were always included in gene isoforms, versus variable areas not always included. Variable exons represent alternative transcription start sites, or less commonly alternative splicing events in the 5' UTR. This distinction uses binding information to provide clues about IMP regulation and motivates subsequent studies to address to potential cause and effect relationship of IMP1 binding to alternative transcript starts. Similarly, IMP2 was found to preferentially bind 3' UTRs and poly(A) signals suggesting possible mechanisms of regulation upstream or downstream of IMP2 binding. Together these examples demonstrate the depth of information that can be gleaned from RBP binding patterns, something that we are continuing to study and exploit.

For example, another challenge for CLIP-seq studies will be to use this information to determine the orientation of RBP binding domains. It will be interesting to know if adaptations like individual-domain CLIP (iDo-PAR-CLIP) (Graf et al., 2013) are necessary to draw these conclusions or not. Most basically we will be curious to know the prevalence and distribution of the KH3-KH4 binding sequences ‘GGACU’ with surrounding ‘ACA’ sites as previously identified in ß-actin,

and described within the SOX2 mRNA here (Figure 32). As suggested here, and evidenced through CLIP-seq studies from other groups, we expect that there is greater information content yet to be extracted from even basic RBP binding maps.

Combinatorial approaches improve predictions of RNA regulation from LIN28 binding maps

One of the remaining questions in these studies of RNA-binding proteins is to understand what distinguishes the function of binding interactions. In the studies of LIN28 described here several groups have been able to overlay multiple experimental datasets to extract relevant information about its functional targets and the downstream pathways and mechanisms it controls, as in regulation of splicing factor abundance (Wilbert et al., 2012) and translation of endoplasmic reticulum associated proteins (Cho et al., 2012). This subtle effect of LIN28 on RNA and protein levels has somewhat limited further conclusions between binding and functional data; however, improvements in the sensitivity of high-throughput assay applied to these measurements will likely warrant a revisit of these associations.

The success of combinatorial approaches suggests that additional permutations with other sources of data, for example presence of co-factors or cellular localization, will facilitate discoveries from transcriptome-wide studies. For example, RNA-helicase A has been shown to be critical for LIN28 mediated translation enhancement of certain mRNA targets. Using information about the transcripts that are commonly bound by both these RBPs might shed light on the positive versus negatively regulated targets of LIN28. Localization of LIN28 mRNPs is another deterministic factor that

can give information about translational state (e.g. stalled translation within P-bodies). Interrogation of LIN28-RNA interactions within particular subcellular compartments is another layer of information that should help us understand its control of particular transcripts.

Some groups of regulators may interact with LIN28 and a subset of its targets to drive certain pathways. For example, the group of gene targets in common between LIN28 and the IMP proteins suggest coordinated roles in neural differentiation and metabolism. Future experiments will be required to test the function and significance of direct IMP and LIN28 binding to the same transcripts. Evidence suggests that the coordination between these post-transcriptional regulators is more likely than the possible link between transcription factors and RBPs that we explored for the reprogramming factors. Instead of widespread control of the same genes, our results indicate that LIN28 does not preferentially bind transcripts driven by OSN, thereby contributing to pluripotency through different pathways or indirect influence on OSN targets. Despite these global observations it remains to be tested if the sub-network of OSNL bound genes could contribute to critical control of stem cell metabolism. Recent findings that LIN28 can enhance tissue repair though changes in glycolysis and oxidative phosphorylation add support to this hypothesis (Shyh-Chang et al., 2013b).

AUTHORS' CONTRIBUTIONS

Chapter 3, in part, is an adaptation of material that appears in "The RNA binding protein IG2BP1/IMP1 promotes survival and adhesion in human pluripotent stem cells by stabilizing target RNAs" by Anne E. Conway, Melissa L. Wilbert and Gene W. Yeo and "LIN28 has slight effects on mRNA translation" by Melissa L. Wilbert, James J. Moresco, and Gene W. Yeo, both in preparation for publication. Proteomics lysates were analyzed by James J. Moresco in the Department of Chemical Physiology, at 10550 North Torrey Pines Road, SR11, The Scripps Research Institute, La Jolla, California 92037. JJM was supported by the National Center for Research Resources (5P41RR011823-17), National Institute of General Medical Sciences (8 P41 GM103533-17), and National Institute on Aging (R01AG027463-04). Anne E. Conway collaborated in the preparation of IMP1 CLIP-seq libraries and editing of this dissertation. Chapter 3, Figure 19, is an adaptation of material that appears in "Genome-wide approaches in the study of microRNA biology" by Melissa L. Wilbert and Gene W. Yeo, as published in *Wiley Interdisciplinary Reviews: Systems Biology and Medicine* December 2010.

FIGURES

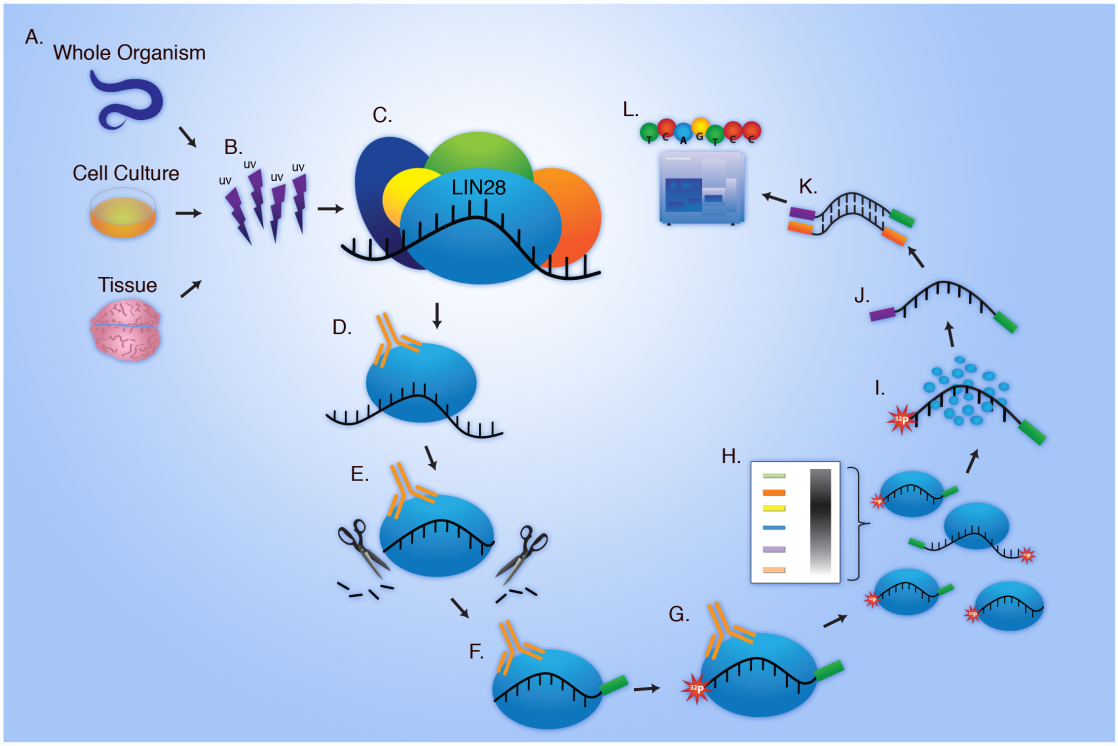


Figure 19. CLIP-seq method overview.
The CLIP-seq method for identifying RBP target transcripts can begin with different types of starting material, (A) either whole organisms, cell culture optionally grown in the presence of photoreactable ribonucleosides, or tissue samples. (B) UV-irradiation is used to covalently cross-link proteins with interacting nucleotides. (C) Proteins of the mRNP complex associated with the RBP of interest (LIN28 shown) bound to target RNA. (D) Immunoprecipitation of the protein can be used to co-precipitate bound RNA transcripts. (E) Unbound RNA transcripts are degraded, for example with MNase treatment. (F) A 3'q linker for adaptation to the Illumina sequencing platform is ligated to precipitated RNAs. (G) Radiolabeling and (H) SDS-PAGE are used to purify and select RBP–RNA complexes of interest. (I) Proteins bound to RNA transcripts are degraded before (J) ligation of the 5_ sequencing linker and (K) preparation of a cDNA library via RT-PCR and followed by further PCR amplification. (L) Sequencing is typically performed on the Illumina/Solexa system; reads returned correspond to RNA transcript originally bound in the LIN28-mRNP complex.

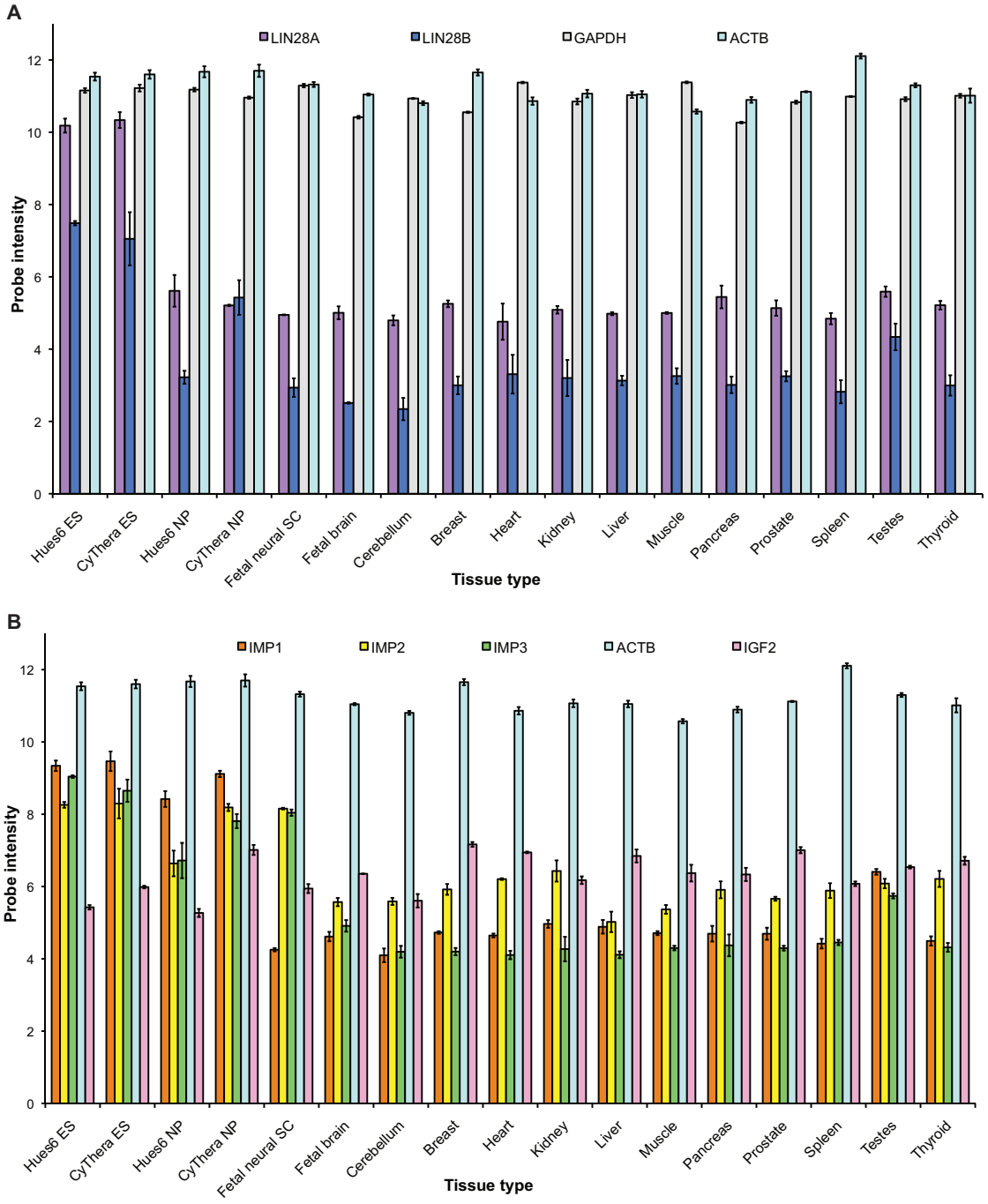


Figure 20. LIN28 and IMP RNA expression in human stem cells, progenitor cells, and adult tissues. Primary fetal and adult human tissue samples and cell lines were subjected to custom microarray platforms profiling mRNA expression. Expression levels are represented by probe intensity normalized across the array. SC = stem cells; ES = embryonic stem cells; NP = neural progenitor cells.

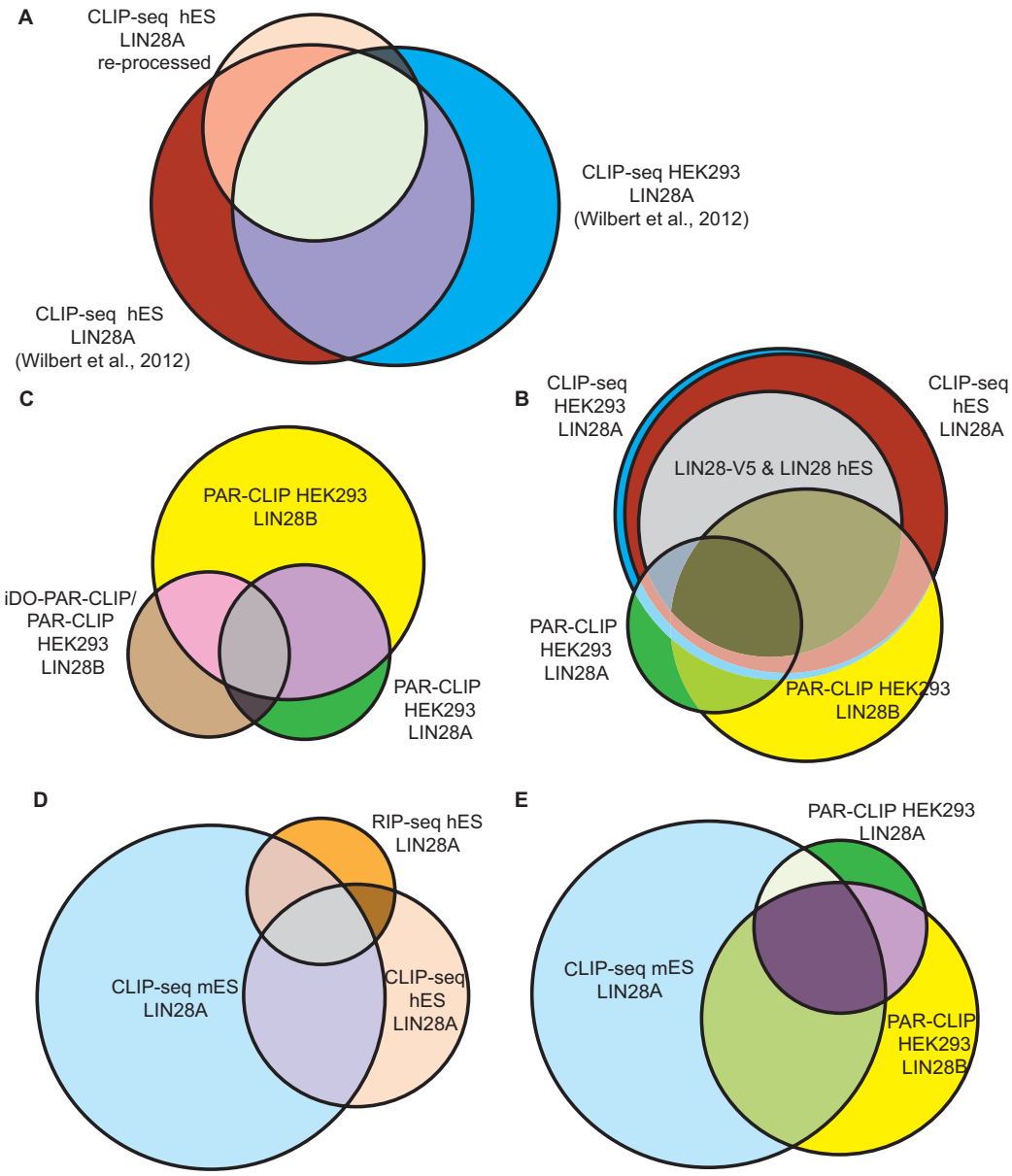


Figure 21. Overlaps between LIN28 transcriptome-wide binding datasets

Datases of LIN28A and LIN28B RNA targets were obtained from in-house and published sources (see methods). (A) Overlap of gene targets identified by CLIP-seq in hESCs and HEK293 cells, and more stringent analysis of the LIN28 hES binding data (re-processed) that produced a smaller subset of targets. (B) LIN28A CLIP-seq target genes in common between hES and HEK293 (Wilbert et al., 2012)) shown in grey. The total target sets (HEK293 blue; hESCs red) are also given disregarding their actual overlap. LIN28A and LIN28B PAR-CLIP data from Hafner et al., 2013 are shown in green and yellow respectively. (C) PAR-CLIP LIN28A and LIN28B targets in HEK293 cells as in B, compared to LIN28B individual domain and CLIP (iDO-PAR-CLIP) (Graf et al., 20133). (D) LIN28A gene targets from CLIP-seq in hESCs (as reprocess in A); CLIP-Seq in hES, mES (Cho et al., 2012) and RIP-seq in hES (Peng et al., 2011). (E) PAR-CLIP LIN28A and LIN28B targets in HEK293 cells as in B and CLIP-seq in hESCs.

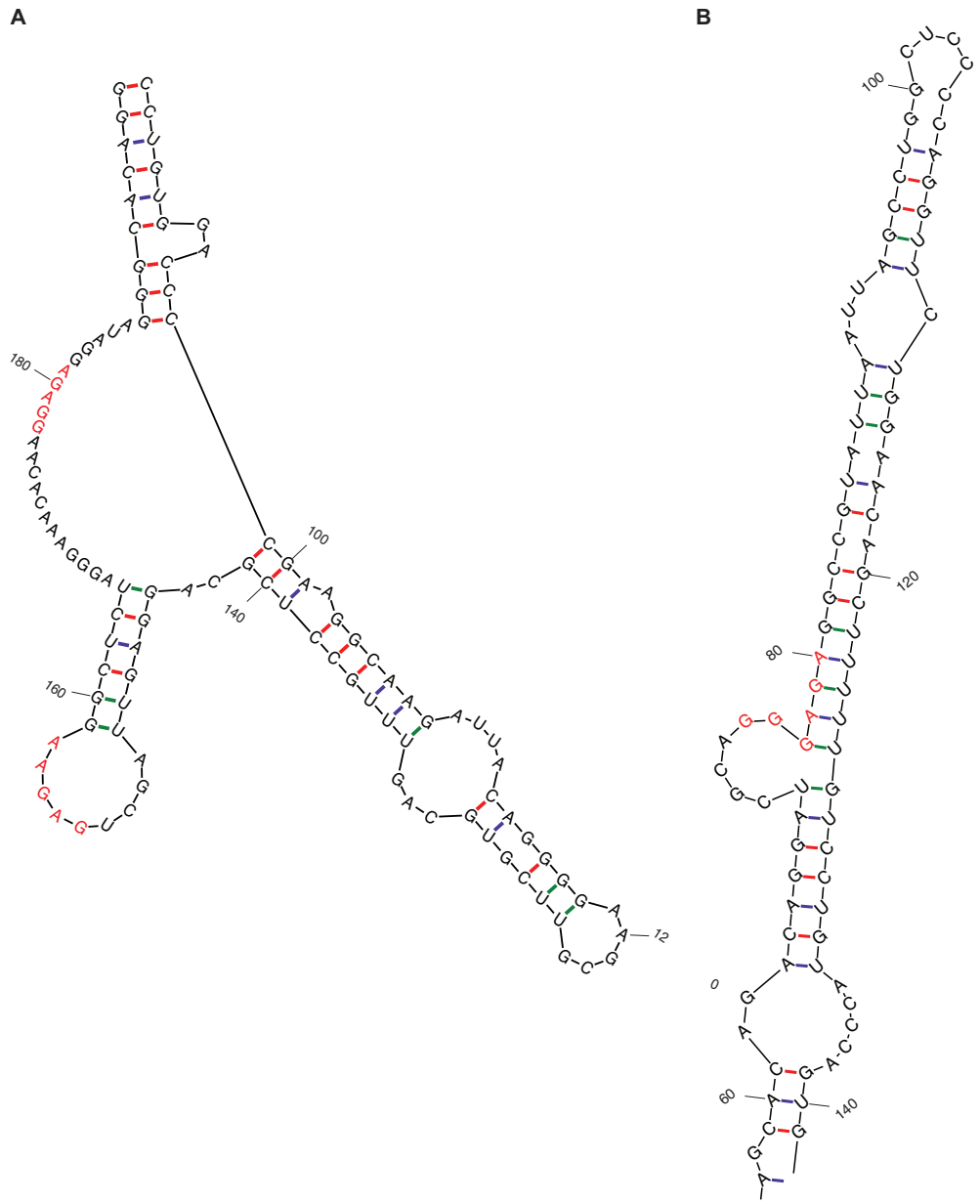


Figure 22. Secondary structure predictions of LIN28 bound mRNAs.
(A) The predicted secondary structure folding of a portion of the hnRNP F 3' UTR and FUS/TLS last coding exon with 3' UTR used in design of LIN28 responsive luciferase reporters. The LIN28 recognition motifs GAGAA and GGAGA are highlighted in red. Numbering corresponds to relative location in the sequence used in construct design for Wilbert et al., 2012.

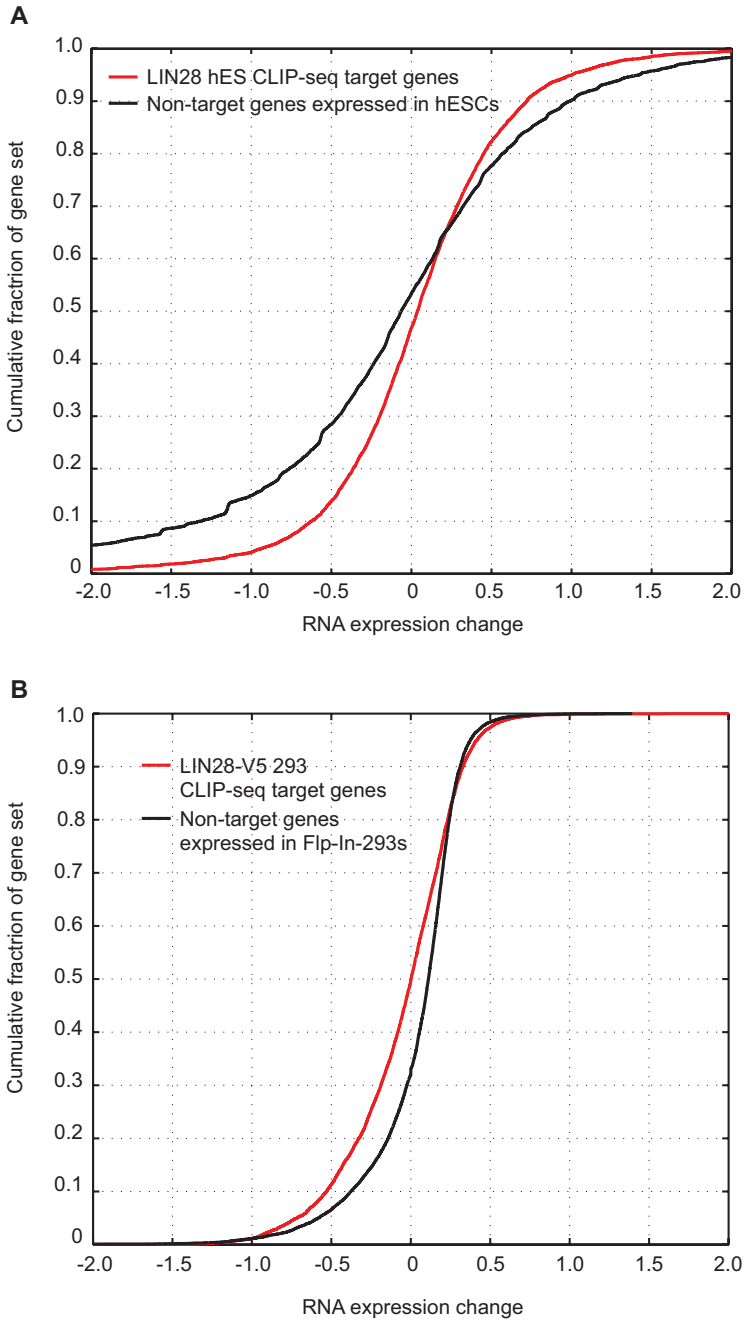


Figure 23. Cumulative distribution plots of RNA level changes in response to LIN28.
(A) RNA expression levels were assayed using RNA-seq in hESCs upon control or LIN28A knockdown. The change in expression for genes with LIN28 CLIP-seq binding (Wilbert et al., 2012) or not were graphed as a cumulative fraction of the entire gene set. (B) RNA expression levels were assayed with mircoarray analysis in HEK293 cells with LIN28-V5 overexpression. The change in expression for genes with LIN28 CLIP-seq binding (Wilbert et al., 2012) or not were graphed as a cumulative fraction of the entire gene set.

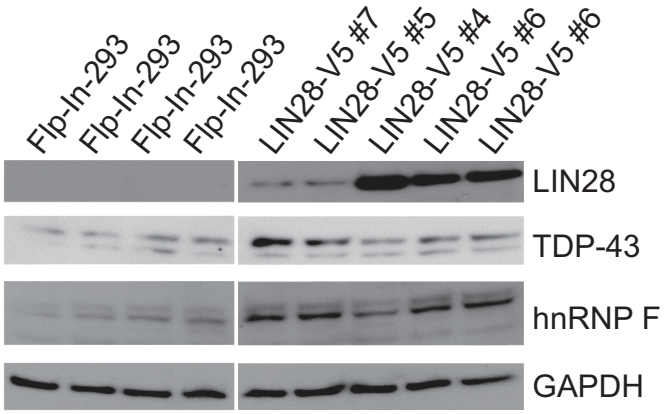


Figure 24. LIN28 targets respond in a dose dependent manner.
Western blot of the splicing factor proteins TDP-43 and hnRNP-F in Flp-In-293 cells (HEK293 with FLP recombination site) and Flp-In-293 cells with stable expression of a V5 tagged LIN28 protein. Clonal LIN28-V5 lines vary in LIN28-V5 expression from high (LIN28-V5 #4), to mid (LIN28-V5 #6, two replicates shown), to low (LIN28-V5 #7 & LIN28-V5#5).

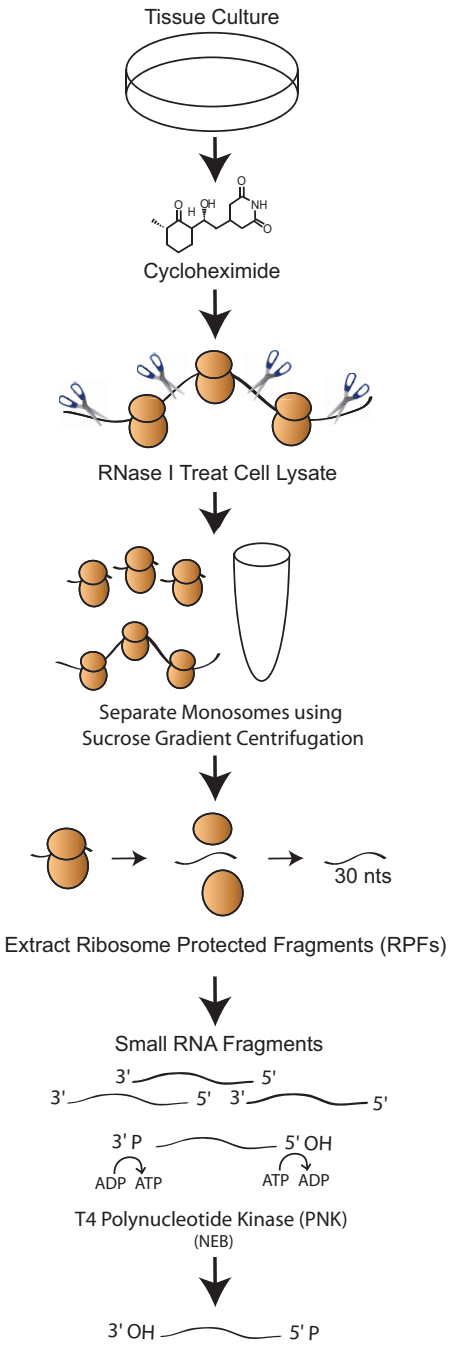


Figure 25. Ribosomal profiling method overview.
Cells are treated in culture with cycloheximide to freeze translating ribosomes on mRNA transcripts. RNAse I treatment degrades RNA not protected the by ribosome. Sucrose gradient centrifugation isolates individual monosomes, from which ~30nt protected RNA fragments can be isolated. Modification of the ends of these short transcripts makes them compatible with downstream sequencing protocol, such as the Illumina small RNA-seq kit.

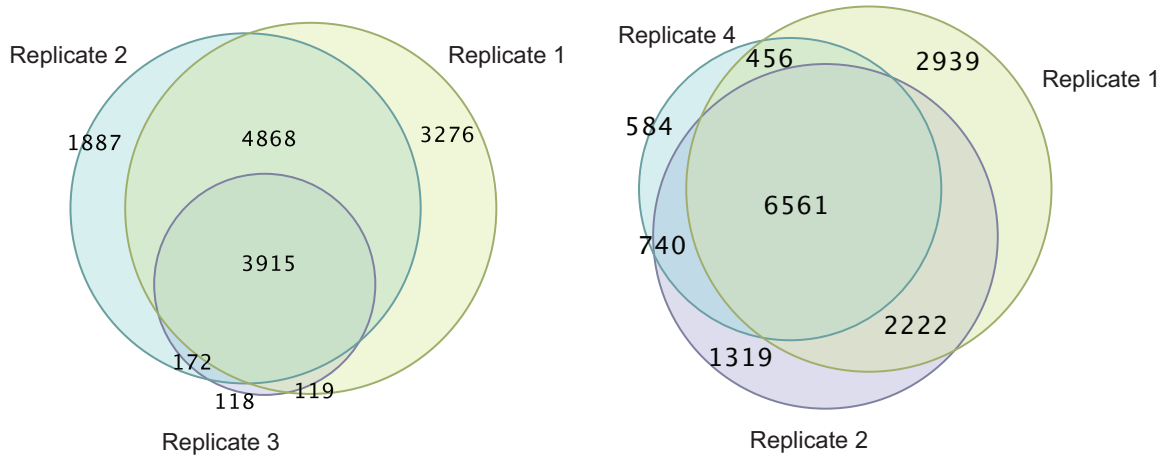


Figure 26. Overlap of peptides quantified in replicate mass spectral analyses.
Peptides were analyzed using tandem mass spectrometry in replicate samples of LIN28-V5 expressing HEK293 cells. Sets of peptides detected in biological replicate samples 1,2, and 3 (left) and 4 (right).

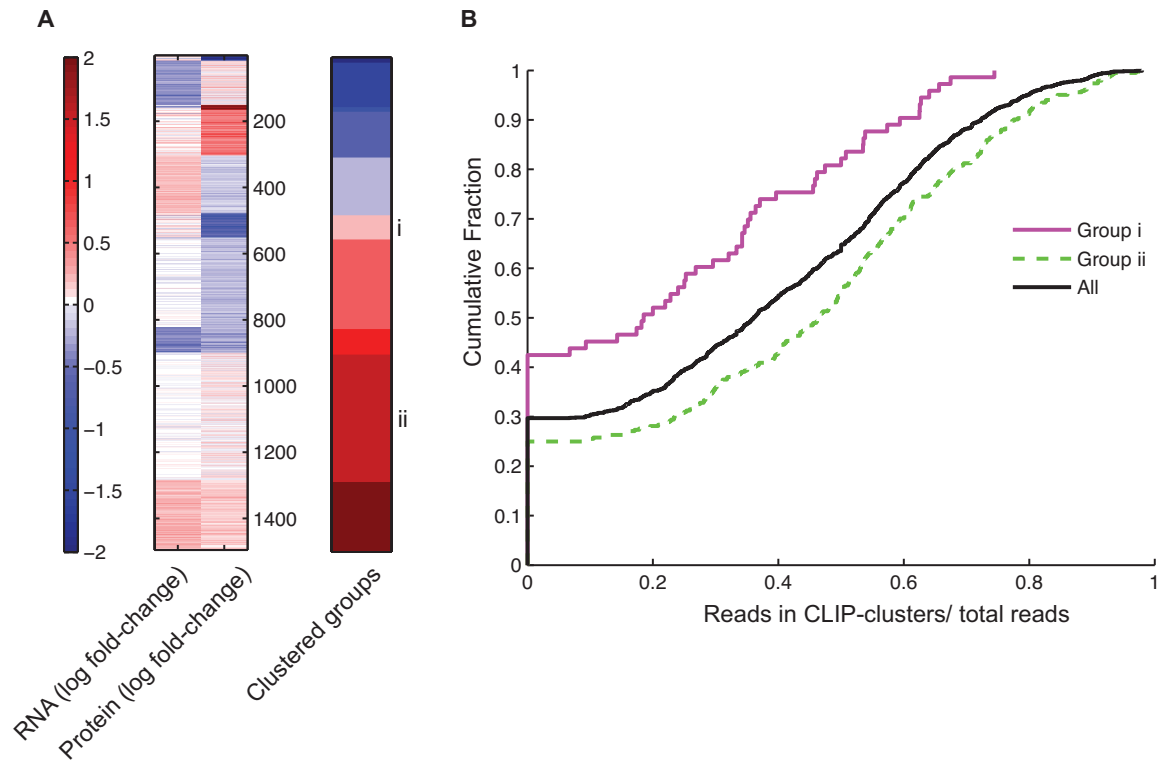


Figure 27. Clustering of genes by correlation of RNA and protein expression change
(A) K-means clustering based on RNA fold-change and protein fold-change in LIN28-V5 expressing HEK293 cells versus control. (B) Gene groups i and ii were graphed as a cumulative distribution relative to the strength of CLIP-seq data as measured by the reads in clusters/ total reads per gene.

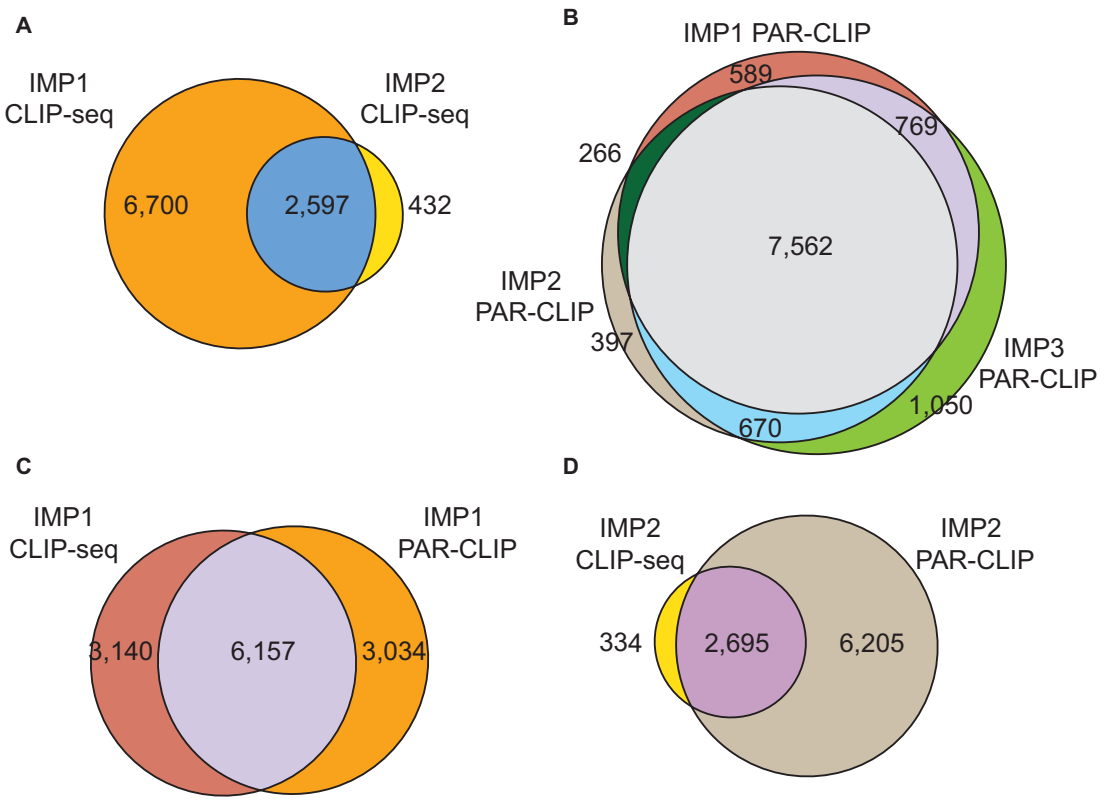


Figure 28. Overlap of gene targets of the IMP family proteins determined by CLIP-seq and PAR-CLIP CLIP-seq datasets of endogenous IMP1 and IMP2 binding generated in hESCs (A) were compared to published PAR-CLIP datasets of exogenously expressed IMP1-3 targets in HEK293 cells (B). IMP1 gene targets (C) and IMP2 gene targets (D) were largely in common.

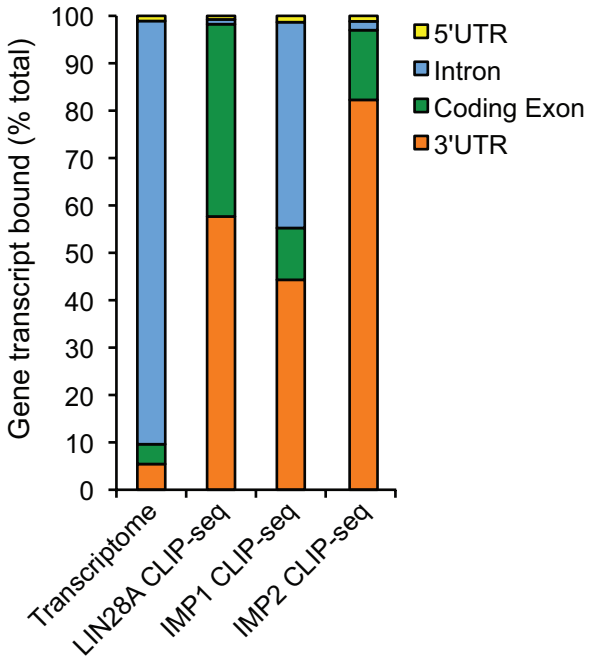


Figure 29. CLIP-seq binding cover of genic regions.
The total nucleotide occupancy within annotated 5' UTR, intron, coding-exon, and 3' UTR regions was calculated for the entire transcriptome (background control), and LIN28A, IMP1, and IMP2 CLIP-seq bound sites, and graphed as a percent of the entire dataset.

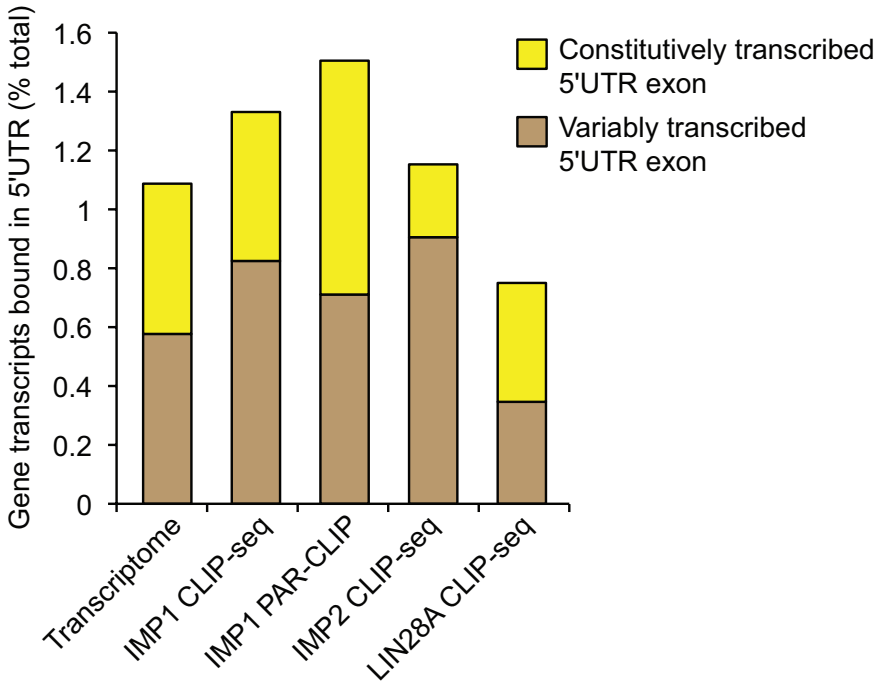


Figure 30. 5' UTR cover of CLIP-seq libraries

The total nucleotide occupancy within annotated 5' UTR exons either constitutively included in transcripts, or exons variably transcribed, (sometimes not part of the transcript) as in the case of alternative transcription start sites. The percent of total nucleotide over is graphed as a percent of the entire dataset. PAR-CLIP data from Hafner et al., 2010 was mapped to our gene annotations. The entire transcriptome serves as a background control.

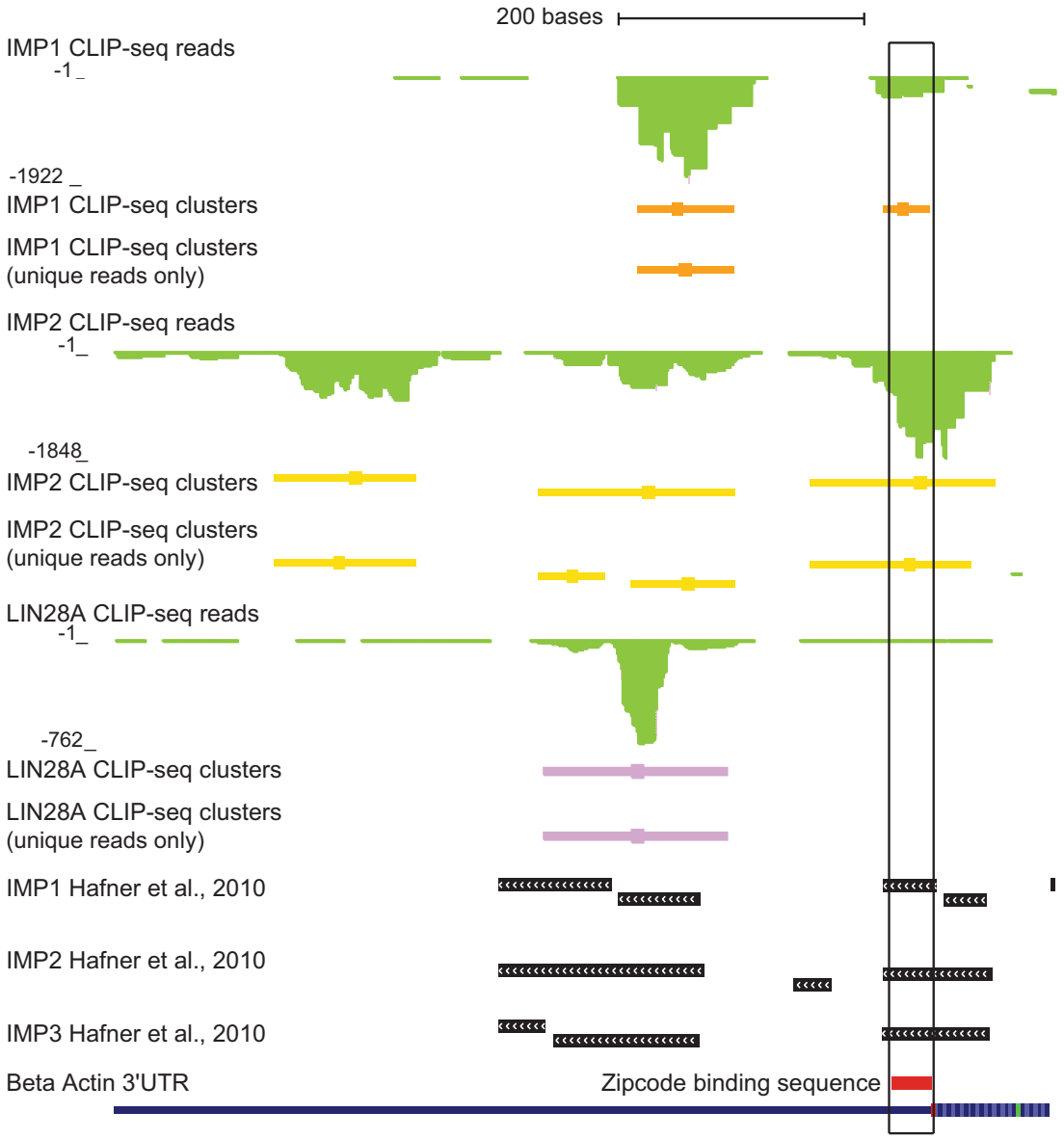


Figure 31. CLIP-seq identifies the IMP zipcode binding sequence in ACTB.
Genome browser view of CLIP-seq reads (green) mapped to the 3' UTR of beta-actin (encoded on the reverse strand). Significant binding sites (CLIP-seq clusters) of IMP1 (orange), IMP2 (yellow), LIN28 (purple), and IMP1-3 (black)(PAR-CLIP sites) are depicted. The zipcode binding sequence defined at the active site of IMP binding by previous studies if shown in red.

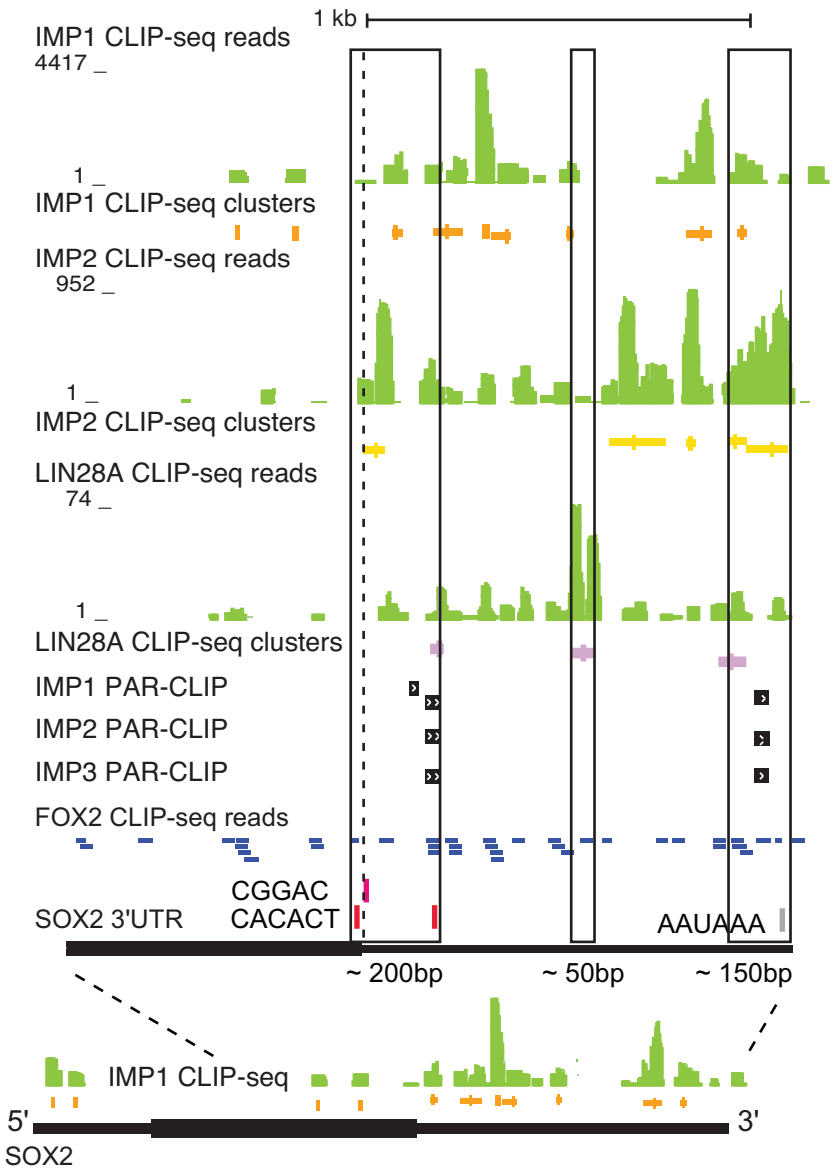


Figure 32. The IMP family RBPs bind the 3' UTR of SOX2.
Genome browser view of CLIP-seq reads (green) mapped to the 3' UTR of SOX2. Significant binding sites (CLIP-seq clusters) of IMP1 (orange), IMP2 (yellow), LIN28 (purple), and IMP1-3 (black)(PAR-CLIP sites) are depicted. FOX2 CLIP-seq reads, not enriched, are in blue. The IMP1 binding motifs CGGAC and surrounding CACACU are situated by the last coding exon. The polyA signal motif AAUAAA is found overrepresented in IMP2 CLIP-seq sites. Below IMP1 reads and clusters mapped across the entire SOX2 gene.

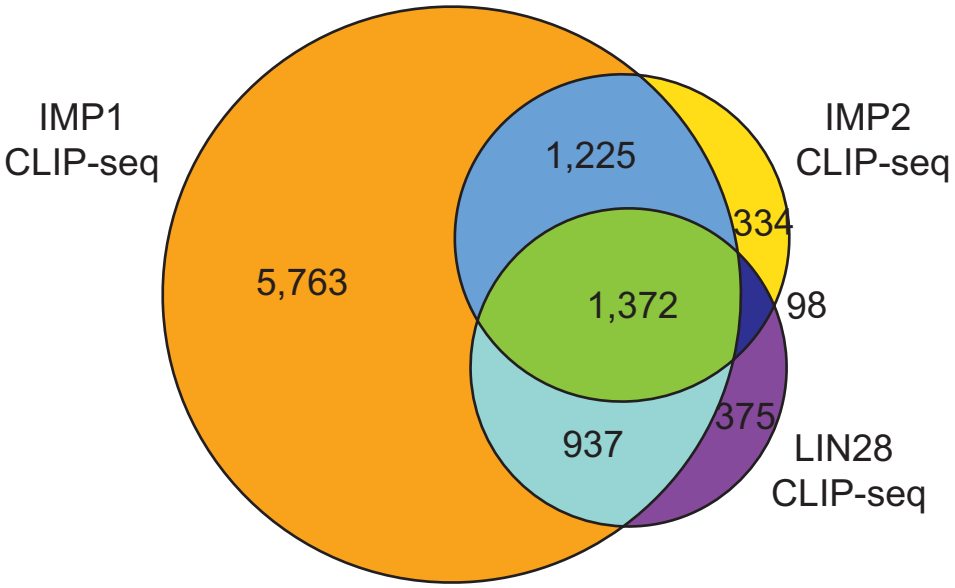


Figure 33. Interaction of IMP1, IMP2, and LIN28 RNA bound gene targets
CLIP-seq binding data was used to define significantly bound RNAs of these RBPs in hESCs. The number of targets in common and unique to each dataset are displayed by venn diagram.

A

	DNA bound sites	-800 bp to +200bp	-2 kb to +500bp	-10 kb to +2 kb
OCT4	3,889	1,538	1,549 (0.7%)	1,762 (12.7%)
SOX2	5,682	1,636	1,639 (0.2%)	1,722 (5.0%)
NANOG	25,071	4,770	4,773 (0.1%)	5,020 (5.0%)
KLF4	3,794	1,626	1,631 (0.3%)	1,826 (11.0%)
p300	3,093	849	853 (0.5%)	898 (5.5%)
TAF1	12,362	3,668	3,737 (1.8%)	4,681 (21.6%)

B

	All genes	Protein-coding genes
OCT4	1538	1293 (84%)
SOX2	1636	1237 (76%)
NANOG	4770	3520 (74%)
KLF4	1626	1368 (84%)
p300	849	637 (75%)
TAF1	3668	3174 (87%)
LIN28	2782	2594 (93%)
IMP1	9297	8055 (87%)
IMP2	3029	2923 (97%)
OSN	216	187 (87%)
OSN LIN28	45	44 (98%)
OSN IMP1	128	122 (95%)
OSN IMP2	41	41 (100%)

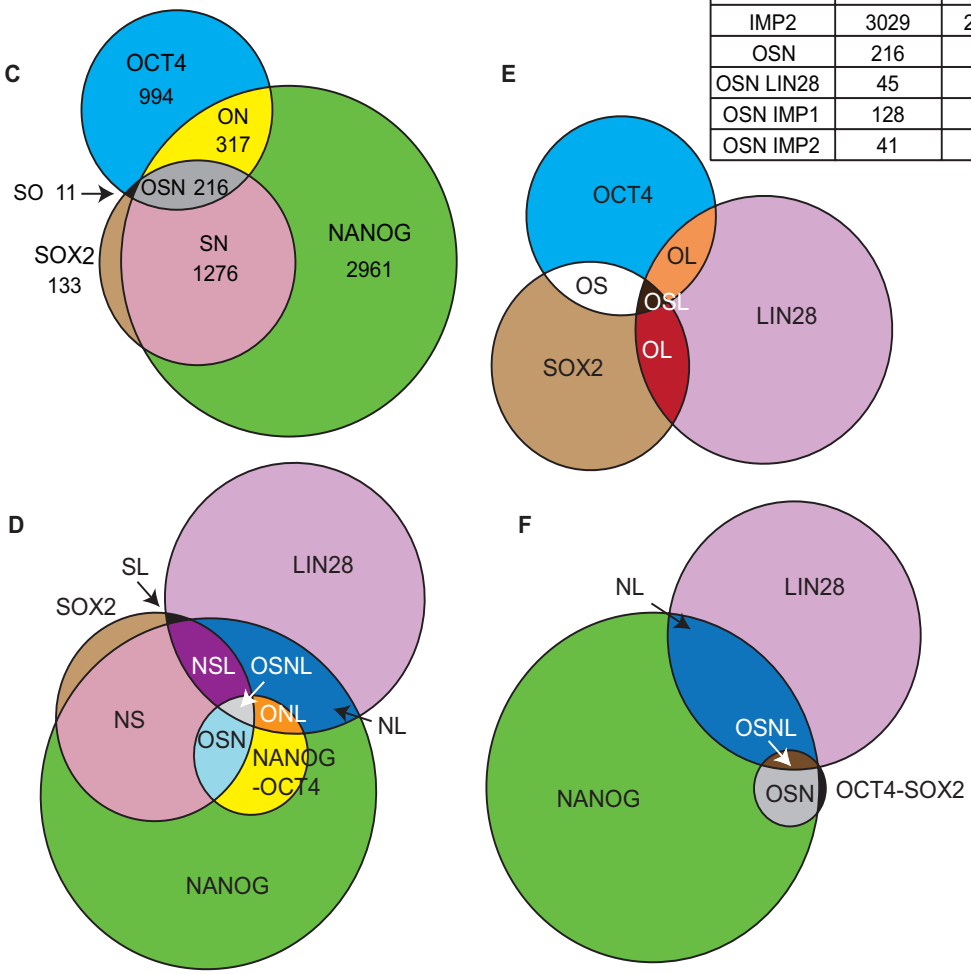
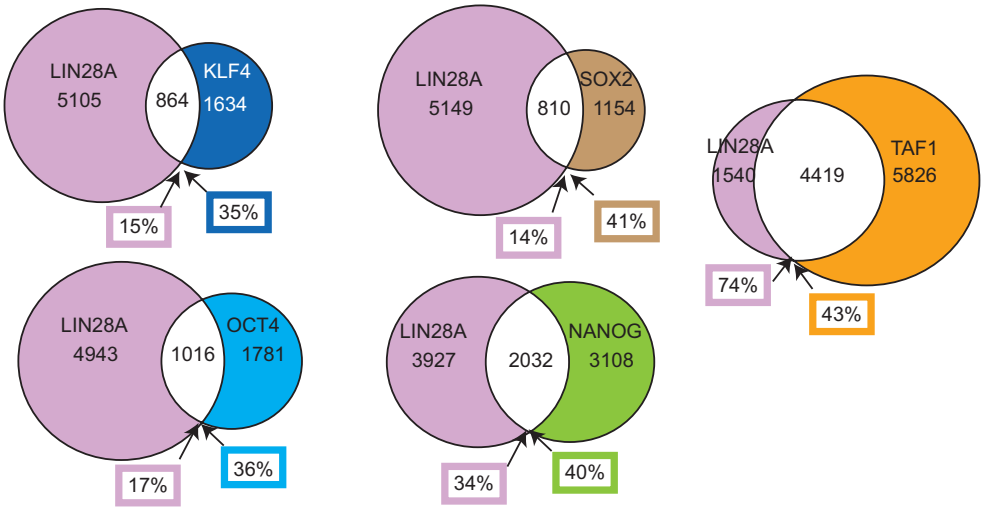


Figure 34. Intersection of hESC enriched transcription factor and RBPs gene targets.
(A) The number of gene promoters associated with DNA-binding protein occupied sites based on coordinates published by Lister et al., 2009, mapped to the human genome build hg19. (B) The number of DNA bound and RNA bound sites associated with annotated genes or protein-coding genes only (% of protein-coding gene out of all target genes identified). CLIP-seq RNA bound sites were generated with unique reads only mapped to hg19. (C-F) Venn diagrams of the gene targets in common between OCT4 (O), SOX (2), NANOG (N), and LIN28 (L). In (D) the subset of NANOG-OCT4 and (F) SOX2-OCT4 target genes are represented as independent groups.

A

	OCT4	SOX2	NANOG	KLF4	p300	TAF1	LIN28	IMP1	IMP2
OCT4	1538	227 15%/ 14%	521 34%/ 11%	411 27%/ 25%	138 9.0%/ 16%	660 43%/ 18%	222 14%/ 8.0%	776 51%/ 8.3%	270 18%/ 8.9%
SOX2	227 14%/ 15%	1636	1492 91%/ 31%	202 12%/ 12%	627 38%/ 74%	329 20%/ 9.0%	261 16%/ 9.4%	830 51%/ 8.9%	292 18%/ 9.6%
NANOG	521 11%/ 34%	1492 31%/ 91%	4770	521 11%/ 32%	838 18%/ 99%	913 19%/ 25%	739 15%/ 27%	2344 49%/ 25%	798 17%/ 26%
KLF4	411 25%/ 27%	202 12%/ 12%	521 32%/ 11%	1626	113 6.9%/ 13%	518 32%/ 14%	233 14%/ 8.4%	818 50%/ 8.8%	288 18%/ 9.5%
p300	138 16%/ 9.0%	627 74%/ 38%	838 99%/ 18%	113 14%/ 6.9%	849	183 22%/ 5.0%	156 18%/ 5.6%	443 52%/ 4.8%	151 18%/ 5.0%
TAF1	660 18%/ 43%	329 9.0%/ 20%	913 25%/ 19%	521 14%/ 32%	183 5.0%/ 22%	3668	707 19%/ 25%	2012 55%/ 22%	660 22%/ 27%

B



C

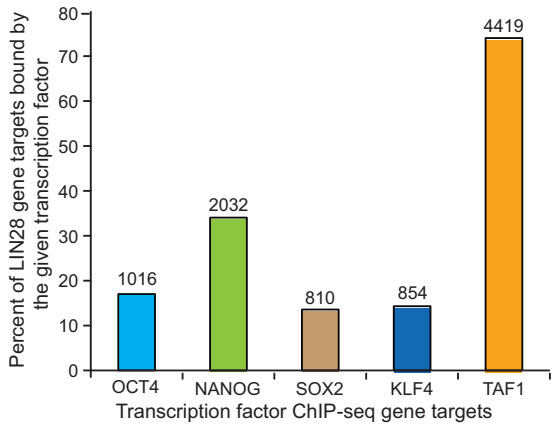
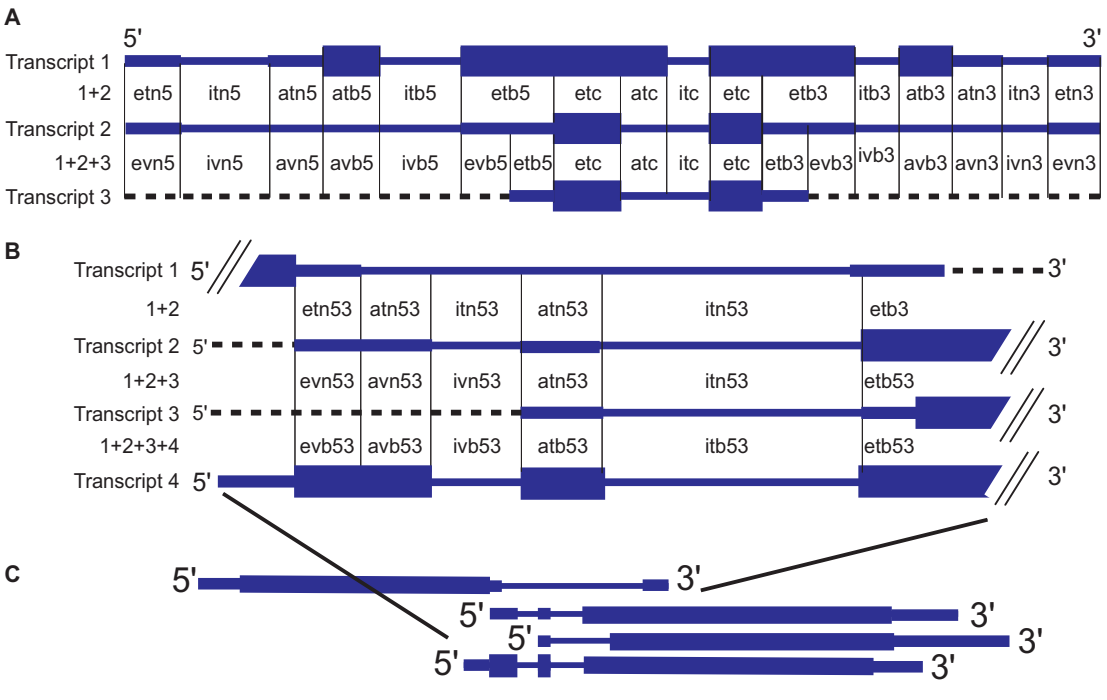


Figure 35. Binary comparisons of transcription factor and RNA-binding protein target genes in hESCs. (A) The number of genes found in ChIP- or CLIP-seq datasets for the given proteins in common as associate based on locations in hg19 genome coordinates. The percent found in common of total targets of the first (row) and second (column) datasets are given (e.g. row %/ column %). (B) Using original hg18 datasets the overlap in (A) were repeated for the transcription factors and published LIN28 mRNA targets (5,969 genes). The percent of overlap for either protein is given inside the respective colored box. (C) The information in (B) displayed as a bar chart for clarity.



e = exon - This region is exonic for all known genes
i = intron – This region is intronic for all known genes
a = alternatively spliced –This region is exonic for 1 or more known genes, and intronic for 1 or more known genes

t = transcribed – This region is transcribed for all known genes
v = variable transcription region – This region is transcribed for 1 or more known genes, and not transcribed for a different 1 or more known genes

o = outside transcription region – This region is outside the transcription region of a known gene, this is used within the program to calculate variable transcription regions

c = coding – This region is inside the coding region for all known genes
n = non-coding – This region is outside the coding region for all known genes
b = both coding and non-coding – This region is inside the coding region of 1 or more known genes, and it is outside the coding region of a different 1 or more known genes

5 = 5′ – This region is in the 5′UTR of 1 or more of the known genes
3 = 3′ – This region is in the 3′UTR of 1 or more of the known genes

Figure 36. Intersection of transcripts and their functional region exclusive annotations (FREA)
Three representative transcripts encoded from the same region of the genome. Each transcript contributes information to the annotations of gene regions. Abbreviations of each regions type are given in (D). (B) Examples of overlapping transcript ends and transcript annotations that would result from genes arranged in the genome as in (C). (D) Definitions used to create mutually exclusive categories based on splicing, transcription, and translation

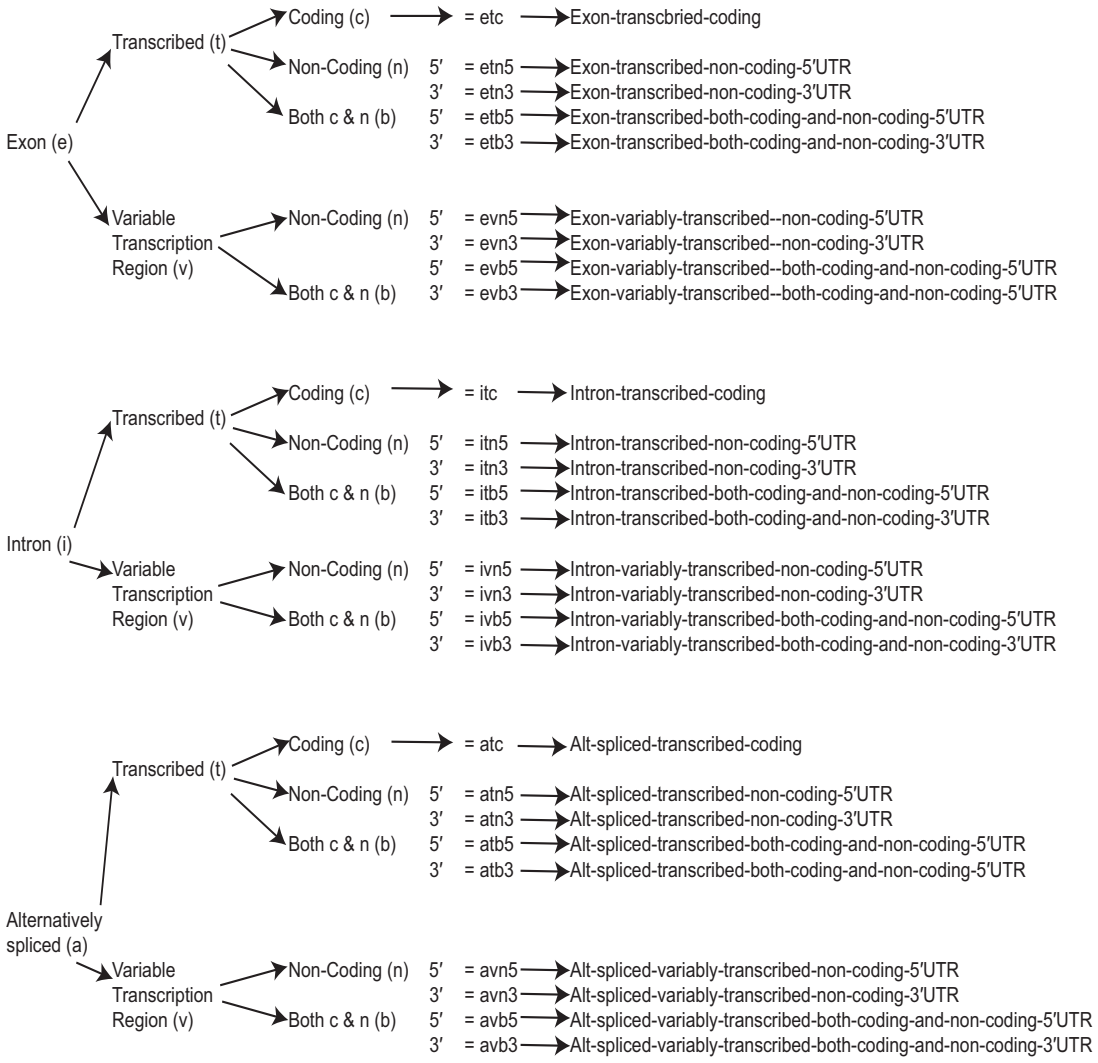


Figure 37. Flow chart of FREA dependencies.
Every region of the transcriptome was defined based on overlapping refseqs, mRNAs, and cDNA information at every position. Regions were given a mutually exclusive annotation encompassing all variations of exon, intron, transcription, and translation. The flow chart above described how a region is defined along with abbreviations used in Figure 36.

Table 4. Review of LIN28A and LIN28B transcriptome studies.

Cell-type	Protein targeted	CLIP approach	Reference	Gene targets/ binding sites	Major findings
hESC H1 (WA01, WiCell)	LIN28A (ab40620)	RIP-seq (IgG IP control)	Peng et al., *Stem Cells*, 2011	1,259 mRNAs total, 268 top genes	Top target RNP proteins and splicing factors, translational regulators, cellular metabolism; LIN28 KD move targets to the monosome; LIN28 works with RNA helicase A to positively regulate translation of target mRNAs
A2780 mESCs R1	LIN28A (ab46020)	RIP-mRNA array (IgG IP control)	Li et al., *J. Biol Chem*, 2012	1,707 mRNAs in A2780 cells; 2,806 mRNAs in mESCs	801 common transcripts were enriched in IPs from human breast cancer cells (A2780) and mESCs; LIN28A promotes cancer cell proliferation through control of cell cycle regulators, some also let-7 affected e.g. CCND1; few mRNAs respond to LIN28 KD
hESC H9, HEK293 (Flp-In 293)	LIN28A (ab46020), LIN28-V5 (anti-V5)	CLIP-seq, MNase	Wilbert et al., *Molecular Cell*, 2012	~6,000 mRNAs, ~26,000 LIN28 hESC and ~15,000 LIN28-V5 293 sites	LIN28A binds many miRNAs and thousands of mRNAs in a similar sequence GGAGA(U) and structural context as let-7; LIN28 binds largely overlapping set of targets in hESC and 293 cells; LIN28 regulates translation of splicing factor transcripts resulting in thousands of splicing changes downstream; LIN28 performs auto-regulation and affects LIN28B levels
mESCs	Lin28A (ab46020, 35L33G, 2J3)	CLIP-seq, RNase A	Cho et al., *Cell*, 2012	~13,000 mRNAs, 516,259 putative binding sites	Average of 38.5 Lin28 binding sites per mRNA; consensus motif AAGNGG, most common motif AAGGAG, UGUG novel enriched motif; Lin28 binding correlated with increase in translation upon KD (Ribo-seq), no affect on RNA levels; Lin28 negatively regulates levels of ER-associated proteins; Lin28 localizes peri-nuclear region
Flp-In 293 T-Rex	FLAG/HA-LIN28B; FLAG/HA-LIN28A (anti-FLAG)	PAR-CLIP, RNase T1	Hanfer et al., *RNA*, 2013	LIN28A ~1,800 mRNAs, ~2,500 bound sites; LIN28B ~3,800 mRNAs ~9,500 bound sites	Most bound transcripts in common, 60% LIN28A binding sites overlap a LIN28B site; AYYHY (Y=C,U; H = A,C,U) binding primarily in CDS, and 3'UTR; binds ssRNA but can also bind motifs within disrupted paired regions and secondary structure; EMSA shows U, GU, and AG polyers optimal LIN28A ligands; top ranked targets, RBPs, slight but significant increase in RNA and protein levels (array, SILAC); suggests LIN28 proteins act as part of mRNP scaffolds
Flp-In 293 T-Rex	FLAG/HA-LIN28B; FLAG/HA-LIN28B-HIS (anti-FLAG)	PAR-CLIP; iDo-PAR-CLIP (individual domain CLIP)	Graf et al., *RNA Biology*, 2013	~10,500 targets, high confidence set of 1,527 protein-coding genes, 2,540 sequence clusters	LIN28B binds 51% 3' UTR and 44% CDS; let-7 miRNAs and 3 others; 3 UTR binding motif (G/A)GG(G/C)(A/U)G, CDS motif AAG(A/G)(A/U)G; (A/U/G)GGGAG ZKD motif, UUUUCC CSD motif; 5' to 3' orientation of CSD to ZFD; LIN28B acts as a positive regulator of protein synthesis measured by pSILAC shotgun proteomics; CDS binding influences protein synthesis more than 3' UTR only; analysis biased towards CSD and 3' UTR annotations; LIN28B KD in HEK293 cells impairs cell-cycle (shift to G2/M phase upon KD) and proliferation; no control for RNA change

Table 5. Enriched GO categories for genes downregulated by LIN28 overexpression or knockdown.

A

Downregulated upon LIN28 knockdown		All genes (320)		LIN28hES CLIP-seq targets (115)	
GO categories	GO term	Number of target genes in GO term	Corrected *P*-value	Number of target genes in GO term	Corrected *P*-value
Molecular function	Herapin binding	6	1.86×10^{-6}	5	1.58×10^{-11}
	Growth factor activity	11	7.04×10^{-9}		
Cellular component	Nucleus	24	1×10^{0}		
Biological process	Cell adhesion	17	5.35×10^{-4}	14	4.61×10^{-12}
	Muscle development	9	8.32×10^{-5}	5	9.99×10^{-6}
	Neurogensis	13	1.89×10^{-5}	6	1.89×10^{-5}

B

Downregulated upon LIN28 expression		All genes (240)		LIN28v5 CLIP-seq targets (67)	
GO categories	GO term	Number of target genes in GO term	Corrected *P*-value	Number of target genes in GO term	Corrected *P*-value
Molecular function	DNA binding	24	1.90×10^{-6}	12	2.33×10^{-6}
	Zinc ion binding	40	1.69×10^{-9}	12	2.10×10^{-4}
	Helicase activity	5	1.86×10^{-5}		
Cellular component	Nucleus	49	2.20×10^{-5}	23	5.39×10^{-6}
Biological process	DNA replication	5	7.11×10^{-5}		

Table 6. Gene targets in common between LIN28, OCT4, SOX2, and NANOG in hESCs

Gene Name	Description	Gene Name	Description
ETV1	ets variant 1	GULP1	GULP, engulfment adaptor PTB domain containing 1
GRAMD1B	GRAM domain containing 1B	PAM	peptidylglycine alpha-amidating monooxygenase
AGPAT4	1-acylglycerol-3-phosphate O-acyltransferase 4	NLGN4X	neuroligin 4, X-linked
NEDD4L	neural precursor cell expressed, developmentally down-regulated 4-like, E3 ubiquitin protein ligase	PLOD2	procollagen-lysine, 2-oxoglutarate 5-dioxygenase 2
PPP1R12A	protein phosphatase 1, regulatory subunit 12A	CACNA2D1	calcium channel, voltage-dependent, alpha 2/delta subunit 1
SEMA3A	sema domain, immunoglobulin domain (Ig), short basic domain, secreted, (semaphorin) 3A	CHST9	carbohydrate (N-acetylgalactosamine 4-0) sulfotransferase 9
FAT1	FAT atypical cadherin 1	MMP16	matrix metallopeptidase 16 (membrane-inserted)
LAPTM4B	lysosomal protein transmembrane 4 beta	EEF1A1	eukaryotic translation elongation factor 1 alpha 1
CDK6	cyclin-dependent kinase 6	HK1	hexokinase 1
UST	uronyl-2-sulfotransferase	AMOTL1	angiomotin like 1
QKI	QKI, KH domain containing, RNA binding	KIF5C	kinesin family member 5C
TFG	TRK-fused gene	ANTXR1	anthrax toxin receptor 1
TGIF2	TGFB-induced factor homeobox 2	CDH2	cadherin 2, type 1, N-cadherin (neuronal)
LRAT	lecithin retinol acyltransferase (phosphatidylcholine--retinol O-acyltransferase)	AKIRIN1	akirin 1
ABCC4	ATP-binding cassette, sub-family C (CFTR/MRP), member 4	PCCA	propionyl CoA carboxylase, alpha polypeptide
HIP1	huntingtin interacting protein 1	BCOR	BCL6 corepressor
PXDN	peroxidasin homolog (Drosophila)	MYO5A	myosin VA (heavy chain 12, myoxin)
DIAPH1	diaphanous-related formin 1	TLK1	tousled-like kinase 1
DUSP6	dual specificity phosphatase 6	ALPK2	alpha-kinase 2
TMEM132B	transmembrane protein 132B	CAPZA2	capping protein (actin filament) muscle Z-line, alpha 2
CUL4A	cullin 4A	LEFTY1	left-right determination factor 1
SMAD4	SMAD family member 4	KIAA0319L	KIAA0319-like

METHODS

ChIP-seq data analysis

Transcription factor-DNA interactions sites were identified by ChIP-Seq in H1 cells by Lister et al., 2010. These sites were the result of combining at least two biological replicates. These published hg18 coordinates were converted to BED formatted files and then to hg19 using the liftOver tool from the UCSC Genome Browser utilities (http://genome.ucsc.edu/). We mapped transcription factor bound regions to genes transcription state sites (TSS) based on ENSEMBL gene annotations within a defined window upstream and downstream of the TSS.

LIN28 CLIP-seq and PAR-CLIP datasets

Comparisons of LIN28 RNA maps from various groups were made through conversion of all datasets to ENSEMBL gene IDs and overlap of these gene lists. Published coordinates of LIN28 bound sites within RNA transcripts were obtained from Hafner et al., 2010 (LIN28A and LIN28B PAR-CLIP in HEK293 cells) and assigned to genes using the same annotations and algorithm used for our in-house LIN28 bound clusters. LIN28 hESC CLIP-seq data from Wilbert et al., 2012 was used as published but reassigned to target ENSEMBL IDs (where specified), or re-processed with more stringent parameters to remove all redundant reads mapped the identical start/stop locations. This refined dataset was compared to the other published studies. Lin28a bound RefSeq annotations reported by Cho and colleagues (Cho et al., 2012) were mapped to their human homologues. Published gene symbols from Peng et

al., 2011 and Graf et al., 2013 were converted to their associated ENSEMBL ID using annotations downloaded from http://www.ensembl.org/bioma.

PAR-CLIP data

Published PAR-CLIP datasets of IGF2BP binding sites were obtained from Hafner et al., 2010. The liftOver tool from the UCSC Genome Browser utilities (http://genome.ucsc.edu/) was used to map the published NCBI36/hg18 coordinates to the latest GRch37/hg19 genome build.

CLIP-seq

The CLIP-seq protocol was performed as in (Wilbert et al., 2012) using antibodies against endogenous IMP1 (Santa Cruz Biotechnology, sc-21026) and IMP2 (MBL, RN008P) in the human embryonic stem cell line H9 and HUES6, respectively.

Pre-processing of reads was used to as a quality control step to remove sequences of low quality and polynucleotide run-ons. The rationale for removing ploynucleotide run-ons is based on the high likelihood of these sequences resulting from errors in sequencing or not map to the genome, as in the case of polyA tails that are added post-transctionally. Sequence resulting from adapters and barcodes are also removed at this step. This processing was performed using an in house tools or cutadapt (code.google.com/p/cutadapt), both of which provide similar options, for example for dealing with partial adapter matches and adapter matches occurring anywhere in the sequence. Clustering parameters included removing reads with the same start and stop site (trim), use of an mRNA transcript length as background, and

searching within windows of 50-200bp. Sets of control clusters were generated by taking LIN28 CLIP-seq clusters and selecting a same sized sequence a random distance from the transcript start of the target gene, to control for differences in gene expression. Control clusters were also confined to the same genic regions as LIN28 CLIP-seq clusters. For all control datasets 10 iterations of randomly selected controls were generated.

Proteomics sample preparation LIN28-V5 HEK293 cells

To harvest cells each 10cm dish was rinsed with ice-cold PBS and then lysed with 500 μl 8M Urea/20 mM Tris-HCl/protein phosphatase inhibitor, pH 8.0. Dissociate the lysate using three 15 s pulses with a microprobe tip at 20W, while keeping lysates as cool as possible. Centrifugation at 10,000g for 15 minutes was used to clear the lysate, which was then quantified using the BCA quantification assay (Thermo Pierce). We removed 1mg of protein, adjusted it to a final concentration of 10mM DTT and then incubated it at 60°C for 30 minutes. After incubation 5mM DTT was added to quench alkylation and then each sample was diluted 4x with 100mM Tris-HCL-1mM $CaCl_2$, then combined with 10μg trypsin (1:100).

Proteomics sample processing and analysis

Equal amounts of heavy SILAC and light proteins were combined and digested (Veenstra et al., 2000). Peptides were separated by MudPIT (Wolters et al., 2001) and analyzed with a LTQ-Velos-Orbitrap. Tandem mass spectra were extracted from raw files using RawExtract 1.9.9 (McDonald et al., 2004a, b) and were searched against a

UniProt human database with reversed sequences using ProLuCID (Xu et al., 2006). Peptide candidates were filtered using DTASelect (McDonald et al., 2004a, b; Tabb et al., 2002) and relative quantification was determined using Census (Park et al., 2008). Peptides reliably detected in 3 or 4 (experimental condition LIN28-V5 expressing cells) or 2of 3 (control HEK Flp-IN-293 cells) were included in further calculations.

Gene ontology analysis

The **D**atabase for **A**nnotation, **V**isualization and **I**ntegrated **D**iscovery (DAVID Bioinformatic Resources 6.7; http://david.abcc.ncifcrf.gov/) was used to generate gene ontology associations and assign functional categories to genes. The set of transcripts from the human genome hg19 was used as background.

Human mRNA expression arrays

Normalized probe intensity values for triplicate microarray experiments using GeneChip Human Exon 1.0 ST arrays on human tissues and cell lines were obtained from published (Yeo et al., 2007) and public sources http://www.affymetrix.com/support/technical/sample_data/exon_array_data.affx. Mean and standard deviation of expression values were calculated using Perl scripts.

RNA expression changes

RNA expression values and changes upon expression of LIN28 in HEK293 cells (LIN28-V5 293 cells) or knockdown in hES H9 cells were obtained from Wilbert et al., 2012.

FUTURE DIRECTIONS

Functional annotation of the genome is central to our ability to connect genetic information with the impact variations have on the organism. The necessity for the basic understanding of where DNA-binding factors and RNA-binding factors interact with target transcripts, and the outcome of these contacts has fueled investments in large, multi-group efforts towards this goal (2011). Understanding the exact point of interaction between RNAs and regulators provides valuable information on two scales. Firstly, this information allows us to predict how a change in nucleotide sequence will affect the ability of that transcript to be properly regulated. This information alone can inform us of disease susceptibility or drug response. Secondly, connecting mRNP binding and regulatory maps with other –omics datasets, both static and context dependent, improves our ability to make predictions of how defects in RBP function affect downstream pathways controlling cellular function. Combining advances in these data mining techniques, with the ability to model human development through *in vitro* differentiation of pluripotent cells, opens new opportunities for rapid discovery and conformation of functional polymorphisms affecting human health and development.

CONNECTING LIN28 RNA REGULATORY MAPS TO DISEASE ASSOCIATED POLYMORPHISMS

Our CLIP-seq studies and those of other groups enable transcriptome-wide identification of RBP-RNA interactions at the level of nucleotide contacts. This

resolution allows us to determine if single nucleotide polymorphisms (SNPs) occur within RBP binding sites and has the potential to connect base pair changes to functional outcome. A change in nucleotide sequence can affect the ability of an RBP to bind and regulate a target transcript, resulting in a cascade of downstream effects and possibly a disease phenotype. This phenomenon has been documented in AGO-miRNA binding sites. Because of the constraints surrounding miRNA base pairing to mRNA targets, a change in the sequence of the miRNA or mRNA target due to genetic variation can disrupt miRNA targeting. However, SNPs can also create miRNA target sites, or affect the miRNA transcript itself (Iwai and Naraba, 2005). These polymorphisms that affect miRNA function are known as miRSNPs or miR-polymorphisms and directly result in disease phenotypes and instances of drug resistance (Abelson et al., 2005; Calin et al., 2005; Calin et al., 2004; Iwai and Naraba, 2005; Mayr et al., 2007; Mishra et al., 2008; Nicoloso et al., 2010). One such SNP in LIN28 (rs3811463) upstream of a let-7 binding site has been shown to reduce binding by the miRNA, leading to an increase in LIN28 protein expression (Chen et al., 2011). Separate studies have found significant association with this allele in breast cancer and type 2 diabetes mellitus (Zhang et al., 2013). Similarly, a polymorphism in *Lin28b* was identified by another genome-wide association study (GWAS) as a risk factor in neuroblastoma (Diskin et al., 2012). It is likely that this SNP affects the ability of LIN28B to properly regulate neural crest progenitors, permitting a transformation into neuroblastoma. By making this connection we can explain the contribution of LIN28

to these phenotypes and potentially indicate the appropriate points of therapeutic intervention in the regulatory cascade.

The connection of LIN28 to type 2 diabetes was also suggested through let-7 targets known to be associated with and contribute to that phenotype (Zhu et al., 2011). However, less is understood about how the direct mRNA targets of LIN28 affect its contribution to disease and development. It remains to be tested if we can identify LIN28 contribution to diseases through SNPs in its protein-coding target genes. Similar to the miRSNP scenario polymorphisms in RBP target sites can affect their binding and function. It may be possible to use the CLIP-seq approach itself to screen for changes in RBP binding associated with patient specific genotypes (Figure 38). Using correlations between RBP binding sites and the findings of GWAS it may possible to connect polymorphisms associated with disease to our RBP-RNA maps. Adding information about either these direct functional interactions or downstream connections between genes can add power to GWAS studies. Inspection of our data revealed that 6 out of 10 genes associated with bone density in humans by the Framingham Osteroporosis study (Kiel et al., 2007), namely CNTNAP2, GRP98, NRG1, VAMP1, RMBS3, and CTNNBL1, were either strong LIN28 binding targets and/or had LIN28-dependent alternative splicing changes. Lin28a transgenic mice were found to have significantly greater bone mineral content and density than wild type littermates (Zhu et al., 2011). The direct and indirect influence of LIN28 on these genes involved with bone formation may explain its role in this phenotype. Whether

LIN28 binding to these targets is directly affected by the sequence changes, as proposed in the miRSNP scenario, remains to be investigated.

NETWORK MODELING OF REGULATION DURING *IN VITRO* DIFFERENTIATION

Classic approaches to genome-wide association studies (GWAS) have historically been limited in their power to associate polymorphisms with disease. Due to this, the simple overlap of RBP binding sites with known disease causing SNPs alone could lack the necessary information to connect a single change to subsequent cascade of disease etiology. Adding additional information content, for example knowledge of signaling pathways or physical protein-protein interactions, is being aggressively explored to expand our ability to draw biological knowledge from network connections. For example, physical protein-protein contacts provide signaling information and information about the architecture of the cell (2011; Stelzl et al., 2005; Tarassov et al., 2008; Yu et al., 2008). Entire complexes of molecules involved in regulation or specific types of modifiers, for example kinase-substrate interactions, has also been added to these approaches (Linding et al., 2007; Ptacek et al., 2005). Genetic networks describe combined effects of mutations in a phenotype, often lethality (Butland et al., 2008; Horn et al., 2011; Linding et al., 2007; Tong et al., 2001; Typas et al., 2008). These static metrics can, and are, being combined to build models that explain how the influence of single factors are carried through a network to affect the system as a whole. Our combinatorial interrogation of LIN28 regulated

RNA networks in relation to other stem cell regulators warrant further investigation utilizing the methods applied to these types of connections.

More recently, the incorporation of dynamic measurements in network mapping creates a picture of interactions across multiple conditions, species or times (reviewed in Ideker and Krogan, 2012). The measurement of genetic interactions in the presence or absence of a pharmacological agent is one example (Bandyopadhyay et al., 2010). This approach has the potential to identify relatively small hubs of regulation that have the greatest influence, a condition that may describe the small subset of LIN28 and OCT4, SOX2, and NANOG target genes defined in Chapter 3. Genetic studies using mouse models have indicated that LIN28 regulation during embryogenesis is essential for proper metabolic function later in the adult animal (Shinoda et al., 2013; Zhu et al., 2011). This provides definitive evidence that this key heterochonic regulator affects tissue specific function long before that tissue has completed development. Thus, LIN28 expression sets a cascade of events in motion that continues over time and through development until the ultimate effects are observed in adult animals. This concept expands our understanding of regulatory networks across the dimension of time and development. When we look for associations between mutational loads in a disease for example, we may consider building information on relationships that include pathways connected by roles across development. The utility of defining functional interactions within a changing system, like development, holds the greatest potential to better define the subtle molecular changes that contribute to the programming of lifelong cellular function.

Advances in pluripotent stem cell technologies now enable defined *in vitro* differentiation to lineages in which LIN28 is required, such as neurogenesis and skeletal muscle. Overlapping information from LIN28 and IMP regulation with that of SOX2 in neural development would be a particularly interesting model as all three of these regulators have been shown to impact this lineage, and to physically interact with and regulate each other (Polesskaya et al., 2007). Interestingly, LIN28, commonly thought of as a ubiquitous pluripotency marker, also displays an increase in expression from the early blastocyst to day E 8.5 of early embryonic development in the mouse, about a day after peak OCT4 expression (Darr and Benvenisty, 2009). This may indicate a critical time point in LIN28 driven regulation and warrant exploration of its action around this stage of early development. LIN28 and IMP2 regulation have also been commonly associated with metabolic changes and regulation of each other either through mRNA binding or indirectly through let-7 (Hafner et al., 2013) motivating the study of their potentially interconnected roles in programming metabolism.

The ability to reprogram adult somatic cells into induced pluripotent stem cells (iPSCs) allow us to interrogate how different genetic backgrounds contribute to subtle changes in development that underlie defects during adulthood. Using these models we can test how LIN28 regulation (possibly in conjunction with IMP proteins or other co-factors) interacts with different genetic backgrounds, for example in cells from patients with type 2 diabetes. Furthermore, breakthroughs in techniques to selectively modify specific points in the human genome make it possible to accurately compare

the impact of mutations within the same background of an individual's unique genome. These approaches use engineered nucleases that cleave DNA in a site-specific manner allowing targeted mutagenesis and have been developed using zinc-finger nucleases (ZNs), transcription activator-like effector nucleases (TALENs) and CRISPR-Cas9-derived RNA-guided endonucleases (RGENS) to interrupt protein-coding regions or non-coding transcripts like miRNAs (Cho et al., 2013; Kim et al., 2009; Miller et al., 2011; Sung et al., 2013; Urnov et al., 2005). These techniques are already being used in conjunction with differentiation of human pluripotent cells into various metabolic cell types to demonstrate phenotypes such as insulin-resistance, hypoglycemia, and motor-neuron death (Ding et al., 2013). The significant advances in our understanding of LIN28, in particular its role in mRNA binding and post-transcriptional control, along with these approaches will enable rapid prototyping of network models affected downstream of LIN28 that contribute to disease and development.

CONCLUSIONS

The promises of personalized medicine rely on our ability to interpret how genetic and epigenetic changes contribute to ones health. To make this connection we can use protein and genetic interaction maps to provide information about how pathways are structured and how complex biological instructions are carried out. Our studies of LIN28 in human ES cells gives important insight into how this protein is connected to regulation of cell fate through its mRNA targets. The ability of LIN28 to

bind the same transcripts in different cell-types implies our RNA maps will be useful for the study of LIN28 in various systems. Now that we have access to this static information the challenge remains to connect the regulation of LIN28 to networks that act over the course of development. This will improve our power to discern how its functions in early life program metabolic and other regulation later in the adult. Our ability to connect regulatory networks between drivers of transcription and post-transcriptional regulators, as well as non-coding RNA intermediates like miRNAs, will greatly enhance our ability to predict downstream effects of subtle genetic or epigenetic changes between individuals. The application of this information to models in stem cells and directed differentiation allow us to iteratively test predictions made from modeling regulatory networks. Understanding these interactions will facilitate our ability to determine ideal points for therapeutic intervention and enable better control of *in vitro* differentiation that may be applied to tissue engineering for regenerative medicine applications.

FIGURES

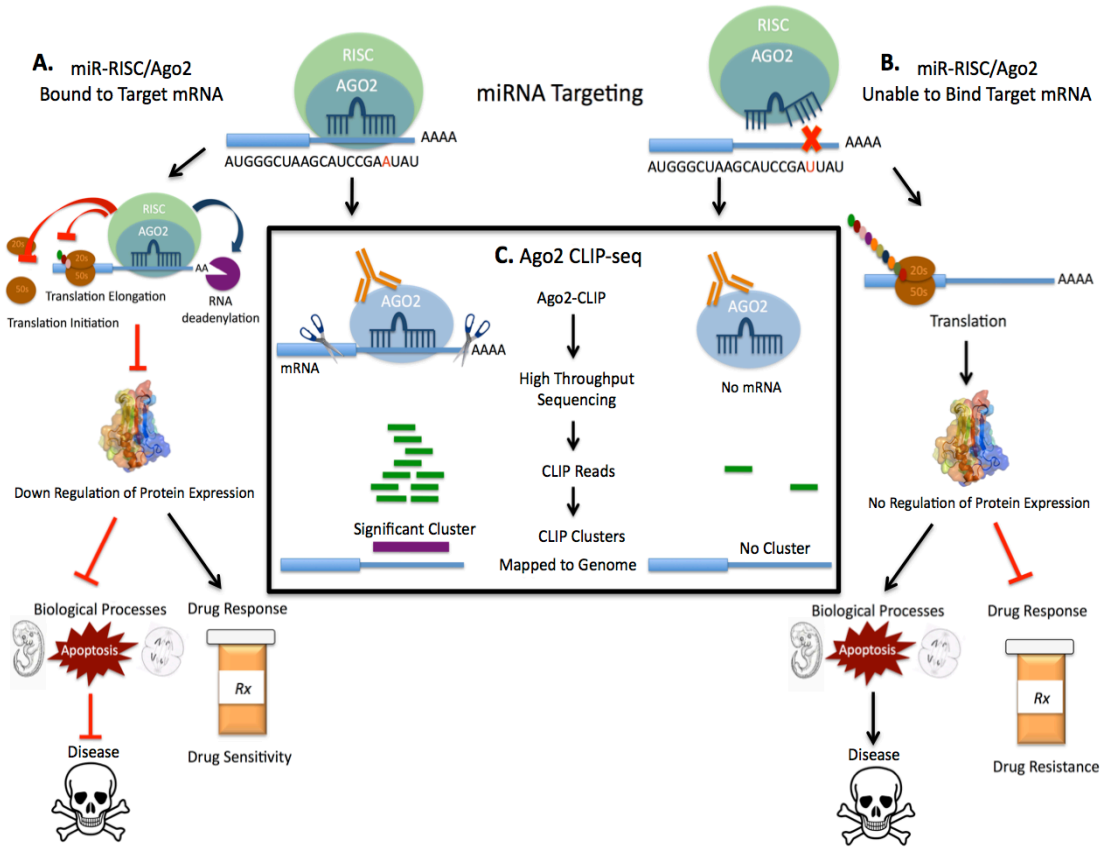


Figure 38. Impact of miRSNPs on miRNA Targeting and Function
A) MiRNAs traditionally target the 3'UTR of protein-coding genes causing down regulation of gene expression by destabilization through deadenylation or decapping, and degradation of the mRNA, or prevention of translational initiation or elongation This regulation allows for precise control of a number of biological processes, such as development, apoptosis and proliferation. It also keeps proteins at basal levels such that drugs are able to effectively target them. B) When a polymorphism in the miRNA or target site prevents miRNA binding, the miR-RISC/Ago2 complex is unable to regulate target mRNA, often leading to overexpression of the protein. This can have aberrant downstream effects on a number of biological processes leading to a disease phenotype. Similarly, overexpression of a target protein can result in drug resistance. C) CLIP-seq with immunoprecipitation of the Ago2 protein can detect miRNA binding or lack of miRNA binding. When a miRNA is able to bind its target mRNA, immunoprecipitation of Ago2 pulls down the bound miRNA and mRNA. However, if a polymorphism prevents miRNA binding to target mRNA, immunoprecipitation of Ago2 will only pull down the miRNA. Subsequent processing through our CLIP-seq protocol and sequencing of the RNAs gives reads corresponding to target mRNA. These reads can be mapped to the genome and and computational analyses are used to identify significant clusters. These clusters represent the mRNA targets site of miRNAs.

REFERENCES

(2011). Evidence for network evolution in an Arabidopsis interactome map. Science *333*, 601-607.

Abelson, J.F., Kwan, K.Y., O'Roak, B.J., Baek, D.Y., Stillman, A.A., Morgan, T.M., Mathews, C.A., Pauls, D.L., Rasin, M.R., Gunel, M.*, et al.* (2005). Sequence variants in SLITRK1 are associated with Tourette's syndrome. Science *310*, 317-320.

Ambros, V., and Horvitz, H.R. (1984). Heterochronic mutants of the nematode Caenorhabditis elegans. Science *226*, 409-416.

Arai, T., Hasegawa, M., Akiyama, H., Ikeda, K., Nonaka, T., Mori, H., Mann, D., Tsuchiya, K., Yoshida, M., Hashizume, Y.*, et al.* (2006). TDP-43 is a component of ubiquitin-positive tau-negative inclusions in frontotemporal lobar degeneration and amyotrophic lateral sclerosis. Biochemical and biophysical research communications *351*, 602-611.

Baltz, A.G., Munschauer, M., Schwanhausser, B., Vasile, A., Murakawa, Y., Schueler, M., Youngs, N., Penfold-Brown, D., Drew, K., Milek, M.*, et al.* (2012). The mRNA-bound proteome and its global occupancy profile on protein-coding transcripts. Molecular cell *46*, 674-690.

Balzer, E., Heine, C., Jiang, Q., Lee, V.M., and Moss, E.G. (2010). LIN28 alters cell fate succession and acts independently of the let-7 microRNA during neurogliogenesis in vitro. Development *137*, 891-900.

Balzer, E., and Moss, E.G. (2007). Localization of the developmental timing regulator Lin28 to mRNP complexes, P-bodies and stress granules. RNA biology *4*, 16-25.

Bandyopadhyay, S., Mehta, M., Kuo, D., Sung, M.K., Chuang, R., Jaehnig, E.J., Bodenmiller, B., Licon, K., Copeland, W., Shales, M.*, et al.* (2010). Rewiring of genetic networks in response to DNA damage. Science *330*, 1385-1389.

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

Basyuk, E., Suavet, F., Doglio, A., Bordonne, R., and Bertrand, E. (2003). Human let-7 stem-loop precursors harbor features of RNase III cleavage products. Nucleic acids research *31*, 6593-6597.

Bell, J.L., Wachter, K., Muhleck, B., Pazaitis, N., Kohn, M., Lederer, M., and Huttelmaier, S. (2013). Insulin-like growth factor 2 mRNA-binding proteins

(IGF2BPs): post-transcriptional drivers of cancer progression? Cellular and molecular life sciences : CMLS *70*, 2657-2675.

Bernhart, S.H., Hofacker, I.L., and Stadler, P.F. (2006). Local RNA base pairing probabilities in large sequences. Bioinformatics *22*, 614-615.

Bleichert, F., and Baserga, S.J. (2007). The long unwinding road of RNA helicases. Molecular cell *27*, 339-352.

Bohnsack, M.T., Czaplinski, K., and Gorlich, D. (2004). Exportin 5 is a RanGTP-dependent dsRNA-binding protein that mediates nuclear export of pre-miRNAs. RNA *10*, 185-191.

Botquin, V., Hess, H., Fuhrmann, G., Anastassiadis, C., Gross, M.K., Vriend, G., and Scholer, H.R. (1998). New POU dimer configuration mediates antagonistic control of an osteopontin preimplantation enhancer by Oct-4 and Sox-2. Genes & development *12*, 2073-2090.

Boyer, L.A., Lee, T.I., Cole, M.F., Johnstone, S.E., Levine, S.S., Zucker, J.P., Guenther, M.G., Kumar, R.M., Murray, H.L., Jenner, R.G.*, et al.* (2005). Core transcriptional regulatory circuitry in human embryonic stem cells. Cell *122*, 947-956.

Brennecke, J., Stark, A., Russell, R.B., and Cohen, S.M. (2005). Principles of microRNA-target recognition. PLoS biology *3*, e85.

Brown, C.J., Hendrich, B.D., Rupert, J.L., Lafreniere, R.G., Xing, Y., Lawrence, J., and Willard, H.F. (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*, 527-542.

Buganim, Y., Faddah, D.A., Cheng, A.W., Itskovich, E., Markoulaki, S., Ganz, K., Klemm, S.L., van Oudenaarden, A., and Jaenisch, R. (2012). Single-cell expression analyses during cellular reprogramming reveal an early stochastic and a late hierarchic phase. Cell *150*, 1209-1222.

Bussing, I., Slack, F.J., and Grosshans, H. (2008). let-7 microRNAs in development, stem cells and cancer. Trends in molecular medicine *14*, 400-409.

Butland, G., Babu, M., Diaz-Mejia, J.J., Bohdana, F., Phanse, S., Gold, B., Yang, W., Li, J., Gagarinova, A.G., Pogoutse, O.*, et al.* (2008). eSGA: E. coli synthetic genetic array analysis. Nature methods *5*, 789-795.

Cai, X., Hagedorn, C.H., and Cullen, B.R. (2004). Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs. RNA *10*, 1957-1966.

Cairo, S., Wang, Y., de Reynies, A., Duroure, K., Dahan, J., Redon, M.J., Fabre, M., McClelland, M., Wang, X.W., Croce, C.M.*, et al.* (2010). Stem cell-like micro-RNA signature driven by Myc in aggressive liver cancer. Proceedings of the National Academy of Sciences of the United States of America *107*, 20471-20476.

Calin, G.A., Ferracin, M., Cimmino, A., Di Leva, G., Shimizu, M., Wojcik, S.E., Iorio, M.V., Visone, R., Sever, N.I., Fabbri, M.*, et al.* (2005). A MicroRNA signature associated with prognosis and progression in chronic lymphocytic leukemia. The New England journal of medicine *353*, 1793-1801.

Calin, G.A., Sevignani, C., Dumitru, C.D., Hyslop, T., Noch, E., Yendamuri, S., Shimizu, M., Rattan, S., Bullrich, F., Negrini, M.*, et al.* (2004). Human microRNA genes are frequently located at fragile sites and genomic regions involved in cancers. Proceedings of the National Academy of Sciences of the United States of America *101*, 2999-3004.

Cao, D., Allan, R.W., Cheng, L., Peng, Y., Guo, C.C., Dahiya, N., Akhi, S., and Li, J. (2011a). RNA-binding protein LIN28 is a marker for testicular germ cell tumors. Human pathology *42*, 710-718.

Cao, D., Liu, A., Wang, F., Allan, R.W., Mei, K., Peng, Y., Du, J., Guo, S., Abel, T.W., Lane, Z.*, et al.* (2011b). RNA-binding protein LIN28 is a marker for primary extragonadal germ cell tumors: an immunohistochemical study of 131 cases. Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc *24*, 288-296.

Caputi, M., and Zahler, A.M. (2001). Determination of the RNA binding specificity of the heterogeneous nuclear ribonucleoprotein (hnRNP) H/H'/F/2H9 family. The Journal of biological chemistry *276*, 43850-43859.

Carninci, P. (2009). Molecular biology: The long and short of RNAs. Nature *457*, 974-975.

Castello, A., Fischer, B., Eichelbaum, K., Horos, R., Beckmann, B.M., Strein, C., Davey, N.E., Humphreys, D.T., Preiss, T., Steinmetz, L.M.*, et al.* (2012). Insights into RNA biology from an atlas of mammalian mRNA-binding proteins. Cell *149*, 1393-1406.

Chalfie, M., Horvitz, H.R., and Sulston, J.E. (1981). Mutations that lead to reiterations in the cell lineages of C. elegans. Cell *24*, 59-69.

Chambers, I., Colby, D., Robertson, M., Nichols, J., Lee, S., Tweedie, S., and Smith, A. (2003). Functional expression cloning of Nanog, a pluripotency sustaining factor in embryonic stem cells. Cell *113*, 643-655.

Chambers, I., and Smith, A. (2004). Self-renewal of teratocarcinoma and embryonic stem cells. Oncogene *23*, 7150-7160.

Chang, T.C., Zeitels, L.R., Hwang, H.W., Chivukula, R.R., Wentzel, E.A., Dews, M., Jung, J., Gao, P., Dang, C.V., Beer, M.A.*, et al.* (2009). Lin-28B transactivation is necessary for Myc-mediated let-7 repression and proliferation. Proceedings of the National Academy of Sciences of the United States of America *106*, 3384-3389.

Chao, J.A., Patskovsky, Y., Patel, V., Levy, M., Almo, S.C., and Singer, R.H. (2010). ZBP1 recognition of beta-actin zipcode induces RNA looping. Genes & development *24*, 148-158.

Chen, A.X., Yu, K.D., Fan, L., Li, J.Y., Yang, C., Huang, A.J., and Shao, Z.M. (2011). Germline genetic variants disturbing the Let-7/LIN28 double-negative feedback loop alter breast cancer susceptibility. PLoS genetics *7*, e1002259.

Cho, J., Chang, H., Kwon, S.C., Kim, B., Kim, Y., Choe, J., Ha, M., Kim, Y.K., and Kim, V.N. (2012). LIN28A is a suppressor of ER-associated translation in embryonic stem cells. Cell *151*, 765-777.

Cho, S.W., Kim, S., Kim, J.M., and Kim, J.S. (2013). Targeted genome engineering in human cells with the Cas9 RNA-guided endonuclease. Nature biotechnology *31*, 230-232.

Chou, J., Shahi, P., and Werb, Z. (2013). microRNA-mediated regulation of the tumor microenvironment. Cell Cycle *12*, 3262-3271.

Christiansen, J., Kolte, A.M., Hansen, T., and Nielsen, F.C. (2009). IGF2 mRNA-binding protein 2: biological function and putative role in type 2 diabetes. Journal of molecular endocrinology *43*, 187-195.

Cimadamore, F., Shah, M., Amador-Arjona, A., Navarro-Peran, E., Chen, C., Huang, C.T., and Terskikh, A.V. (2012). SOX2 modulates levels of MITF in normal human melanocytes, and melanoma lines in vitro. Pigment cell & melanoma research *25*, 533-536.

Cox, J.L., Mallanna, S.K., Luo, X., and Rizzino, A. (2010). Sox2 uses multiple domains to associate with proteins present in Sox2-protein complexes. PloS one *5*, e15486.

Daheron, L., Opitz, S.L., Zaehres, H., Lensch, M.W., Andrews, P.W., Itskovitz-Eldor, J., and Daley, G.Q. (2004). LIF/STAT3 signaling fails to maintain self-renewal of human embryonic stem cells. Stem Cells *22*, 770-778.

Darr, H., and Benvenisty, N. (2009). Genetic analysis of the role of the reprogramming gene LIN-28 in human embryonic stem cells. Stem Cells *27*, 352-362.

Darr, H., Mayshar, Y., and Benvenisty, N. (2006). Overexpression of NANOG in human ES cells enables feeder-free growth while inducing primitive ectoderm features. Development *133*, 1193-1201.

De Guzman, R.N., Turner, R.B., and Summers, M.F. (1998). Protein-RNA recognition. Biopolymers *48*, 181-195.

Deng, Y., Liu, Q., Luo, C., Chen, S., Li, X., Wang, C., Liu, Z., Lei, X., Zhang, H., Sun, H.*, et al.* (2012). Generation of induced pluripotent stem cells from buffalo (Bubalus bubalis) fetal fibroblasts with buffalo defined factors. Stem cells and development *21*, 2485-2494.

Desjardins, A., Yang, A., Bouvette, J., Omichinski, J.G., and Legault, P. (2011). Importance of the NCp7-like domain in the recognition of pre-let-7g by the pluripotency factor Lin28. Nucleic acids research.

Ding, Q., Lee, Y.K., Schaefer, E.A., Peters, D.T., Veres, A., Kim, K., Kuperwasser, N., Motola, D.L., Meissner, T.B., Hendriks, W.T.*, et al.* (2013). A TALEN genome-editing system for generating human stem cell-based disease models. Cell stem cell *12*, 238-251.

Diskin, S.J., Capasso, M., Schnepp, R.W., Cole, K.A., Attiyeh, E.F., Hou, C., Diamond, M., Carpenter, E.L., Winter, C., Lee, H.*, et al.* (2012). Common variation at 6q16 within HACE1 and LIN28B influences susceptibility to neuroblastoma. Nature genetics *44*, 1126-1130.

Doench, J.G., and Sharp, P.A. (2004). Specificity of microRNA target selection in translational repression. Genes & development *18*, 504-511.

Dreyfuss, G., Kim, V.N., and Kataoka, N. (2002). Messenger-RNA-binding proteins and the messages they carry. Nature reviews Molecular cell biology *3*, 195-205.

Easow, G., Teleman, A.A., and Cohen, S.M. (2007). Isolation of microRNA targets by miRNP immunopurification. RNA *13*, 1198-1204.

Folmes, C.D., Nelson, T.J., Martinez-Fernandez, A., Arrell, D.K., Lindor, J.Z., Dzeja, P.P., Ikeda, Y., Perez-Terzic, C., and Terzic, A. (2011). Somatic oxidative

bioenergetics transitions into pluripotency-dependent glycolysis to facilitate nuclear reprogramming. Cell metabolism *14*, 264-271.

Frost, R.J., and Olson, E.N. (2011). Control of glucose homeostasis and insulin sensitivity by the Let-7 family of microRNAs. Proceedings of the National Academy of Sciences of the United States of America *108*, 21075-21080.

Glisovic, T., Bachorik, J.L., Yong, J., and Dreyfuss, G. (2008). RNA-binding proteins and post-transcriptional gene regulation. FEBS letters *582*, 1977-1986.

Goodall, E.F., Heath, P.R., Bandmann, O., Kirby, J., and Shaw, P.J. (2013). Neuronal dark matter: the emerging role of microRNAs in neurodegeneration. Frontiers in cellular neuroscience *7*, 178.

Gore, A., Li, Z., Fung, H.L., Young, J.E., Agarwal, S., Antosiewicz-Bourget, J., Canto, I., Giorgetti, A., Israel, M.A., Kiskinis, E*., et al.* (2011). Somatic coding mutations in human induced pluripotent stem cells. Nature *471*, 63-67.

Graf, R., Munschauer, M., Mastrobuoni, G., Mayr, F., Heinemann, U., Kempa, S., Rajewsky, N., and Landthaler, M. (2013). Identification of LIN28B-bound mRNAs reveals features of target recognition and regulation. RNA biology *10*, 1146-1159.

Griffiths-Jones, S. (2004). The microRNA Registry. Nucleic acids research *32*, D109-111.

Griffiths-Jones, S., Grocock, R.J., van Dongen, S., Bateman, A., and Enright, A.J. (2006). miRBase: microRNA sequences, targets and gene nomenclature. Nucleic acids research *34*, D140-144.

Grimson, A., Farh, K.K., Johnston, W.K., Garrett-Engele, P., Lim, L.P., and Bartel, D.P. (2007). MicroRNA targeting specificity in mammals: determinants beyond seed pairing. Molecular cell *27*, 91-105.

Grishok, A., Pasquinelli, A.E., Conte, D., Li, N., Parrish, S., Ha, I., Baillie, D.L., Fire, A., Ruvkun, G., and Mello, C.C. (2001). Genes and mechanisms related to RNA interference regulate expression of the small temporal RNAs that control C. elegans developmental timing. Cell *106*, 23-34.

Guenther, M.G., Jenner, R.G., Chevalier, B., Nakamura, T., Croce, C.M., Canaani, E., and Young, R.A. (2005). Global and Hox-specific roles for the MLL1 methyltransferase. Proceedings of the National Academy of Sciences of the United States of America *102*, 8603-8608.

Guo, Y., Chen, Y., Ito, H., Watanabe, A., Ge, X., Kodama, T., and Aburatani, H. (2006). Identification and characterization of lin-28 homolog B (LIN28B) in human hepatocellular carcinoma. Gene *384*, 51-61.

Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., Rothballer, A., Ascano, M., Jr., Jungkamp, A.C., Munschauer, M.*, et al.* (2010). Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. Cell *141*, 129-141.

Hafner, M., Max, K.E., Bandaru, P., Morozov, P., Gerstberger, S., Brown, M., Molina, H., and Tuschl, T. (2013). Identification of mRNAs bound and regulated by human LIN28 proteins and molecular requirements for RNA recognition. RNA *19*, 613-626.

Hagan, J.P., Piskounova, E., and Gregory, R.I. (2009). Lin28 recruits the TUTase Zcchc11 to inhibit let-7 maturation in mouse embryonic stem cells. Nat Struct Mol Biol *16*, 1021-1025.

Hales, C.N., and Barker, D.J. (1992). Type 2 (non-insulin-dependent) diabetes mellitus: the thrifty phenotype hypothesis. Diabetologia *35*, 595-601.

Hales, C.N., and Barker, D.J. (2001). The thrifty phenotype hypothesis. British medical bulletin *60*, 5-20.

Hamano, R., Miyata, H., Yamasaki, M., Sugimura, K., Tanaka, K., Kurokawa, Y., Nakajima, K., Takiguchi, S., Fujiwara, Y., Mori, M.*, et al.* (2012). High expression of Lin28 is associated with tumour aggressiveness and poor prognosis of patients in oesophagus cancer. British journal of cancer *106*, 1415-1423.

Hammer, N.A., Hansen, T., Byskov, A.G., Rajpert-De Meyts, E., Grondahl, M.L., Bredkjaer, H.E., Wewer, U.M., Christiansen, J., and Nielsen, F.C. (2005). Expression of IGF-II mRNA-binding proteins (IMPs) in gonads and testicular cancer. Reproduction *130*, 203-212.

Hammond, S.M., Boettcher, S., Caudy, A.A., Kobayashi, R., and Hannon, G.J. (2001). Argonaute2, a link between genetic and biochemical analyses of RNAi. Science *293*, 1146-1150.

Hanna, J., Saha, K., Pando, B., van Zon, J., Lengner, C.J., Creyghton, M.P., van Oudenaarden, A., and Jaenisch, R. (2009). Direct cell reprogramming is a stochastic process amenable to acceleration. Nature *462*, 595-601.

He, C., Kraft, P., Chen, C., Buring, J.E., Pare, G., Hankinson, S.E., Chanock, S.J., Ridker, P.M., Hunter, D.J., and Chasman, D.I. (2009). Genome-wide association

studies identify loci associated with age at menarche and age at natural menopause. Nature genetics *41*, 724-728.

Heckman, K.L., and Pease, L.R. (2007). Gene splicing and mutagenesis by PCR-driven overlap extension. Nature protocols *2*, 924-932.

Heintzman, N.D., Hon, G.C., Hawkins, R.D., Kheradpour, P., Stark, A., Harp, L.F., Ye, Z., Lee, L.K., Stuart, R.K., Ching, C.W.*, et al.* (2009). Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature *459*, 108-112.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Molecular cell *38*, 576-589.

Henderson, J.K., Draper, J.S., Baillie, H.S., Fishel, S., Thomson, J.A., Moore, H., and Andrews, P.W. (2002). Preimplantation human embryos and embryonic stem cells show comparable expression of stage-specific embryonic antigens. Stem Cells *20*, 329-337.

Heo, I., Ha, M., Lim, J., Yoon, M.J., Park, J.E., Kwon, S.C., Chang, H., and Kim, V.N. (2012). Mono-uridylation of pre-microRNA as a key step in the biogenesis of group II let-7 microRNAs. Cell *151*, 521-532.

Heo, I., Joo, C., Cho, J., Ha, M., Han, J., and Kim, V.N. (2008). Lin28 mediates the terminal uridylation of let-7 precursor MicroRNA. Molecular cell *32*, 276-284.

Heo, I., Joo, C., Kim, Y.K., Ha, M., Yoon, M.J., Cho, J., Yeom, K.H., Han, J., and Kim, V.N. (2009). TUT4 in concert with Lin28 suppresses microRNA biogenesis through pre-microRNA uridylation. Cell *138*, 696-708.

Horn, T., Sandmann, T., Fischer, B., Axelsson, E., Huber, W., and Boutros, M. (2011). Mapping of signaling networks through synthetic genetic interaction analysis by RNAi. Nature methods *8*, 341-346.

Huelga, S.C., Vu, A.Q., Arnold, J.D., Liang, T.Y., Liu, P.P., Yan, B.Y., Donohue, J.P., Shiue, L., Hoon, S., Brenner, S.*, et al.* (2012). Integrative genome-wide analysis reveals cooperative regulation of alternative splicing by hnRNP proteins. Cell reports *1*, 167-178.

Hutvagner, G., McLachlan, J., Pasquinelli, A.E., Balint, E., Tuschl, T., and Zamore, P.D. (2001). A cellular function for the RNA-interference enzyme Dicer in the maturation of the let-7 small temporal RNA. Science *293*, 834-838.

Ideker, T., and Krogan, N.J. (2012). Differential network biology. Molecular systems biology *8*, 565.

Iliopoulos, D., Hirsch, H.A., and Struhl, K. (2009). An epigenetic switch involving NF-kappaB, Lin28, Let-7 MicroRNA, and IL6 links inflammation to cell transformation. Cell *139*, 693-706.

Iwai, N., and Naraba, H. (2005). Polymorphisms in human pre-miRNAs. Biochemical and biophysical research communications *331*, 1439-1444.

James, D., Levine, A.J., Besser, D., and Hemmati-Brivanlou, A. (2005). TGFbeta/activin/nodal signaling is necessary for the maintenance of pluripotency in human embryonic stem cells. Development *132*, 1273-1282.

Jaskiewicz, L., and Filipowicz, W. (2008). Role of Dicer in posttranscriptional RNA silencing. Current topics in microbiology and immunology *320*, 77-97.

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

Johnson, D.S., Mortazavi, A., Myers, R.M., and Wold, B. (2007). Genome-wide mapping of in vivo protein-DNA interactions. Science *316*, 1497-1502.

Johnson, S.M., Grosshans, H., Shingara, J., Byrom, M., Jarvis, R., Cheng, A., Labourier, E., Reinert, K.L., Brown, D., and Slack, F.J. (2005). RAS is regulated by the let-7 microRNA family. Cell *120*, 635-647.

Kapeli, K., and Yeo, G.W. (2012). Genome-wide approaches to dissect the roles of RNA binding proteins in translational control: implications for neurological diseases. Frontiers in neuroscience *6*, 144.

Kawahara, H., Okada, Y., Imai, T., Iwanami, A., Mischel, P.S., and Okano, H. (2011). Musashi1 cooperates in abnormal cell lineage protein 28 (Lin28)-mediated let-7 family microRNA biogenesis in early neural differentiation. The Journal of biological chemistry *286*, 16121-16130.

Kazan, H., Ray, D., Chan, E.T., Hughes, T.R., and Morris, Q. (2010). RNAcontext: a new method for learning the sequence and structure binding preferences of RNA-binding proteins. PLoS computational biology *6*, e1000832.

Kedersha, N., and Anderson, P. (2002). Stress granules: sites of mRNA triage that regulate mRNA stability and translatability. Biochem Soc Trans *30*, 963-969.

Keene, J.D. (2007). RNA regulons: coordination of post-transcriptional events. Nature reviews Genetics *8*, 533-543.

Ketting, R.F., Fischer, S.E., Bernstein, E., Sijen, T., Hannon, G.J., and Plasterk, R.H. (2001). Dicer functions in RNA interference and in synthesis of small RNA involved in developmental timing in C. elegans. Genes & development *15*, 2654-2659.

Kiel, D.P., Demissie, S., Dupuis, J., Lunetta, K.L., Murabito, J.M., and Karasik, D. (2007). Genome-wide association with bone mass and geometry in the Framingham Heart Study. BMC medical genetics *8 Suppl 1*, S14.

Kim, C.W., Vo, M.T., Kim, H.K., Lee, H.H., Yoon, N.A., Lee, B.J., Min, Y.J., Joo, W.D., Cha, H.J., Park, J.W.*, et al.* (2012). Ectopic over-expression of tristetraprolin in human cancer cells promotes biogenesis of let-7 by down-regulation of Lin28. Nucleic acids research *40*, 3856-3869.

Kim, H.J., Lee, H.J., Kim, H., Cho, S.W., and Kim, J.S. (2009). Targeted genome editing in human cells with zinc finger nucleases constructed via modular assembly. Genome research *19*, 1279-1288.

King, C.E., Cuatrecasas, M., Castells, A., Sepulveda, A.R., Lee, J.S., and Rustgi, A.K. (2011a). LIN28B promotes colon cancer progression and metastasis. Cancer research *71*, 4260-4268.

King, C.E., Wang, L., Winograd, R., Madison, B.B., Mongroo, P.S., Johnstone, C.N., and Rustgi, A.K. (2011b). LIN28B fosters colon cancer migration, invasion and transformation through let-7-dependent and -independent mechanisms. Oncogene *30*, 4185-4193.

Knight, S.W., and Bass, B.L. (2001). A role for the RNase III enzyme DCR-1 in RNA interference and germ line development in Caenorhabditis elegans. Science *293*, 2269-2271.

Kozomara, A., and Griffiths-Jones, S. (2011). miRBase: integrating microRNA annotation and deep-sequencing data. Nucleic acids research *39*, D152-157.

Kuroda, T., Tada, M., Kubota, H., Kimura, H., Hatano, S.Y., Suemori, H., Nakatsuji, N., and Tada, T. (2005). Octamer and Sox elements are required for transcriptional cis regulation of Nanog gene expression. Molecular and cellular biology *25*, 2475-2485.

Lagier-Tourenne, C., Polymenidou, M., and Cleveland, D.W. (2010). TDP-43 and FUS/TLS: emerging roles in RNA processing and neurodegeneration. Human molecular genetics *19*, R46-64.

Lagos-Quintana, M., Rauhut, R., Lendeckel, W., and Tuschl, T. (2001). Identification of novel genes coding for small expressed RNAs. Science *294*, 853-858.

Lagos-Quintana, M., Rauhut, R., Yalcin, A., Meyer, J., Lendeckel, W., and Tuschl, T. (2002). Identification of tissue-specific microRNAs from mouse. Current biology : CB *12*, 735-739.

Landsman, D. (1992). RNP-1, an RNA-binding motif is conserved in the DNA-binding cold shock domain. Nucleic acids research *20*, 2861-2864.

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

Lau, N.C., Lim, L.P., Weinstein, E.G., and Bartel, D.P. (2001). An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans. Science *294*, 858-862.

Lee, R.C., Feinbaum, R.L., and Ambros, V. (1993). The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell *75*, 843-854.

Lee, T.I., Rinaldi, N.J., Robert, F., Odom, D.T., Bar-Joseph, Z., Gerber, G.K., Hannett, N.M., Harbison, C.T., Thompson, C.M., Simon, I.*, et al.* (2002a). Transcriptional regulatory networks in Saccharomyces cerevisiae. Science *298*, 799-804.

Lee, Y., Ahn, C., Han, J., Choi, H., Kim, J., Yim, J., Lee, J., Provost, P., Radmark, O., Kim, S.*, et al.* (2003). The nuclear RNase III Drosha initiates microRNA processing. Nature *425*, 415-419.

Lee, Y., Jeon, K., Lee, J.T., Kim, S., and Kim, V.N. (2002b). MicroRNA maturation: stepwise processing and subcellular localization. The EMBO journal *21*, 4663-4670.

Lei, X.X., Xu, J., Ma, W., Qiao, C., Newman, M.A., Hammond, S.M., and Huang, Y. (2011). Determinants of mRNA recognition and translation regulation by Lin28. Nucleic Acids Res.

Lettre, G., Jackson, A.U., Gieger, C., Schumacher, F.R., Berndt, S.I., Sanna, S., Eyheramendy, S., Voight, B.F., Butler, J.L., Guiducci, C.*, et al.* (2008). Identification of ten loci associated with height highlights new biological pathways in human growth. Nature genetics *40*, 584-591.

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

Lewis, B.P., Shih, I.H., Jones-Rhoades, M.W., Bartel, D.P., and Burge, C.B. (2003). Prediction of mammalian microRNA targets. Cell *115*, 787-798.

Li, N., Zhong, X., Lin, X., Guo, J., Zou, L., Tanyi, J.L., Shao, Z., Liang, S., Wang, L.P., Hwang, W.T.*, et al.* (2012a). Lin-28 homologue A (LIN28A) promotes cell cycle progression via regulation of cyclin-dependent kinase 2 (CDK2), cyclin D1 (CCND1), and cell division cycle 25 homolog A (CDC25A) expression in cancer. The Journal of biological chemistry *287*, 17386-17397.

Li, X., Quon, G., Lipshitz, H.D., and Morris, Q. (2010). Predicting in vivo binding sites of RNA-binding proteins using mRNA secondary structure. RNA *16*, 1096-1107.

Li, X., Zhang, J., Gao, L., McClellan, S., Finan, M.A., Butler, T.W., Owen, L.B., Piazza, G.A., and Xi, Y. (2012b). MiR-181 mediates cell differentiation by interrupting the Lin28 and let-7 feedback circuit. Cell death and differentiation *19*, 378-386.

Linding, R., Jensen, L.J., Ostheimer, G.J., van Vugt, M.A., Jorgensen, C., Miron, I.M., Diella, F., Colwill, K., Taylor, L., Elder, K.*, et al.* (2007). Systematic discovery of in vivo phosphorylation networks. Cell *129*, 1415-1426.

Lister, R., Pelizzola, M., Dowen, R.H., Hawkins, R.D., Hon, G., Tonti-Filippini, J., Nery, J.R., Lee, L., Ye, Z., Ngo, Q.M.*, et al.* (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. Nature *462*, 315-322.

Liu-Yesucevitz, L., Bilgutay, A., Zhang, Y.J., Vanderweyde, T., Citro, A., Mehta, T., Zaarur, N., McKee, A., Bowser, R., Sherman, M.*, et al.* (2010). Tar DNA binding protein-43 (TDP-43) associates with stress granules: analysis of cultured cells and pathological brain tissue. PloS one *5*, e13250.

Lu, L., Katsaros, D., Shaverdashvili, K., Qian, B., Wu, Y., de la Longrais, I.A., Preti, M., Menato, G., and Yu, H. (2009). Pluripotent factor lin-28 and its homologue lin-28b in epithelial ovarian cancer and their associations with disease outcomes and expression of let-7a and IGF-II. Eur J Cancer *45*, 2212-2218.

Lukong, K.E., Chang, K.W., Khandjian, E.W., and Richard, S. (2008). RNA-binding proteins in human genetic disease. Trends in genetics : TIG *24*, 416-425.

Lund, E., Guttinger, S., Calado, A., Dahlberg, J.E., and Kutay, U. (2004). Nuclear export of microRNA precursors. Science *303*, 95-98.

Lytle, J.R., Yario, T.A., and Steitz, J.A. (2007). Target mRNAs are repressed as efficiently by microRNA-binding sites in the 5' UTR as in the 3' UTR. Proceedings of the National Academy of Sciences of the United States of America *104*, 9667-9672.

Magrane, M., and Consortium, U. (2011). UniProt Knowledgebase: a hub of integrated protein data. Database : the journal of biological databases and curation *2011*, bar009.

Marson, A., Levine, S.S., Cole, M.F., Frampton, G.M., Brambrink, T., Johnstone, S., Guenther, M.G., Johnston, W.K., Wernig, M., Newman, J.*, et al.* (2008). Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells. Cell *134*, 521-533.

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

Maxwell, E.S., and Fournier, M.J. (1995). The small nucleolar RNAs. Annual review of biochemistry *64*, 897-934.

Mayr, C., Hemann, M.T., and Bartel, D.P. (2007). Disrupting the pairing between let-7 and Hmga2 enhances oncogenic transformation. Science *315*, 1576-1579.

McDonald, W.H., Tabb, D.L., Sadygov, R.G., MacCoss, M.J., Venable, J., Graumann, J., Johnson, J.R., Cociorva, D., and Yates, J.R., 3rd (2004a). MS1, MS2, and SQT-three unified, compact, and easily parsed file formats for the storage of shotgun proteomic spectra and identifications. Rapid communications in mass spectrometry : RCM *18*, 2162-2168.

McDonald, W.H., Tabb, D.L., Sadygov, R.G., MacCoss, M.J., Venable, J., Graumann, J., Johnson, J.R., Cociorva, D., and Yates, J.R., 3rd (2004b). MS1, MS2, and SQT-three unified, compact, and easily parsed file formats for the storage of shotgun proteomic spectra and identifications. Rapid Commun Mass Spectrom *18*, 2162-2168.

Melton, C., Judson, R.L., and Blelloch, R. (2010). Opposing microRNA families regulate self-renewal in mouse embryonic stem cells. Nature *463*, 621-626.

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

Miller, J.C., Tan, S., Qiao, G., Barlow, K.A., Wang, J., Xia, D.F., Meng, X., Paschon, D.E., Leung, E., Hinkley, S.J.*, et al.* (2011). A TALE nuclease architecture for efficient genome editing. Nature biotechnology *29*, 143-148.

Mishra, P.J., Mishra, P.J., Banerjee, D., and Bertino, J.R. (2008). MiRSNPs or MiR-polymorphisms, new players in microRNA mediated regulation of the cell: Introducing microRNA pharmacogenomics. Cell Cycle *7*, 853-858.

Mitsui, K., Tokuzawa, Y., Itoh, H., Segawa, K., Murakami, M., Takahashi, K., Maruyama, M., Maeda, M., and Yamanaka, S. (2003). The homeoprotein Nanog is required for maintenance of pluripotency in mouse epiblast and ES cells. Cell *113*, 631-642.

Moore, M.J., and Proudfoot, N.J. (2009). Pre-mRNA processing reaches back to transcription and ahead to translation. Cell *136*, 688-700.

Morita, K., and Han, M. (2006). Multiple mechanisms are involved in regulating the expression of the developmental timing regulator lin-28 in Caenorhabditis elegans. The EMBO journal *25*, 5794-5804.

Moss, E.G., Lee, R.C., and Ambros, V. (1997). The cold shock domain protein LIN-28 controls developmental timing in C. elegans and is regulated by the lin-4 RNA. Cell *88*, 637-646.

Mourelatos, Z., Dostie, J., Paushkin, S., Sharma, A., Charroux, B., Abel, L., Rappsilber, J., Mann, M., and Dreyfuss, G. (2002). miRNPs: a novel class of ribonucleoproteins containing numerous microRNAs. Genes & development *16*, 720-728.

Mukherjee, N., Corcoran, D.L., Nusbaum, J.D., Reid, D.W., Georgiev, S., Hafner, M., Ascano, M., Jr., Tuschl, T., Ohler, U., and Keene, J.D. (2011). Integrative regulatory mapping indicates that the RNA-binding protein HuR couples pre-mRNA processing and mRNA stability. Molecular cell *43*, 327-339.

Nadiminty, N., Tummala, R., Lou, W., Zhu, Y., Zhang, J., Chen, X., eVere White, R.W., Kung, H.J., Evans, C.P., and Gao, A.C. (2012). MicroRNA let-7c suppresses androgen receptor expression and activity via regulation of Myc expression in prostate cancer cells. The Journal of biological chemistry *287*, 1527-1537.

Nakamoto, M., Jin, P., O'Donnell, W.T., and Warren, S.T. (2005). Physiological identification of human transcripts translationally regulated by a specific microRNA. Human molecular genetics *14*, 3813-3821.

Nam, Y., Chen, C., Gregory, R.I., Chou, J.J., and Sliz, P. (2011). Molecular Basis for Interaction of let-7 MicroRNAs with Lin28. Cell.

Nethercott, H.E., Brick, D.J., and Schwartz, P.H. (2011). Derivation of induced pluripotent stem cells by lentiviral transduction. Methods Mol Biol *767*, 67-85.

Neumann, M., Sampathu, D.M., Kwong, L.K., Truax, A.C., Micsenyi, M.C., Chou, T.T., Bruce, J., Schuck, T., Grossman, M., Clark, C.M.*, et al.* (2006). Ubiquitinated TDP-43 in frontotemporal lobar degeneration and amyotrophic lateral sclerosis. Science *314*, 130-133.

Newman, M.A., and Hammond, S.M. (2010). Lin-28: an early embryonic sentinel that blocks Let-7 biogenesis. The international journal of biochemistry & cell biology *42*, 1330-1333.

Nichols, J., Zevnik, B., Anastassiadis, K., Niwa, H., Klewe-Nebenius, D., Chambers, I., Scholer, H., and Smith, A. (1998). Formation of pluripotent stem cells in the mammalian embryo depends on the POU transcription factor Oct4. Cell *95*, 379-391.

Nicoloso, M.S., Sun, H., Spizzo, R., Kim, H., Wickramasinghe, P., Shimizu, M., Wojcik, S.E., Ferdin, J., Kunej, T., Xiao, L.*, et al.* (2010). Single-nucleotide polymorphisms inside microRNA target sites influence tumor susceptibility. Cancer research *70*, 2789-2798.

Nielsen, F.C., Nielsen, J., and Christiansen, J. (2001). A family of IGF-II mRNA binding proteins (IMP) involved in RNA trafficking. Scandinavian journal of clinical and laboratory investigation Supplementum *234*, 93-99.

Nielsen, J., Christiansen, J., Lykke-Andersen, J., Johnsen, A.H., Wewer, U.M., and Nielsen, F.C. (1999). A family of insulin-like growth factor II mRNA-binding proteins represses translation in late development. Molecular and cellular biology *19*, 1262-1270.

Niranjanakumari, S., Lasda, E., Brazas, R., and Garcia-Blanco, M.A. (2002). Reversible cross-linking combined with immunoprecipitation to study RNA-protein interactions in vivo. Methods *26*, 182-190.

Nishimoto, M., Fukushima, A., Okuda, A., and Muramatsu, M. (1999). The gene for the embryonic stem cell coactivator UTF1 carries a regulatory element which selectively interacts with a complex composed of Oct-3/4 and Sox-2. Molecular and cellular biology *19*, 5453-5465.

Niwa, H., Burdon, T., Chambers, I., and Smith, A. (1998). Self-renewal of pluripotent embryonic stem cells is mediated via activation of STAT3. Genes & development *12*, 2048-2060.

Odom, D.T., Zizlsperger, N., Gordon, D.B., Bell, G.W., Rinaldi, N.J., Murray, H.L., Volkert, T.L., Schreiber, J., Rolfe, P.A., Gifford, D.K.*, et al.* (2004). Control of pancreas and liver gene expression by HNF transcription factors. Science *303*, 1378-1381.

Ogryzko, V.V., Schiltz, R.L., Russanova, V., Howard, B.H., and Nakatani, Y. (1996). The transcriptional coactivators p300 and CBP are histone acetyltransferases. Cell *87*, 953-959.

Olsen, P.H., and Ambros, V. (1999). The lin-4 regulatory RNA controls developmental timing in Caenorhabditis elegans by blocking LIN-14 protein synthesis after the initiation of translation. Developmental biology *216*, 671-680.

Ong, K.K., Elks, C.E., Li, S., Zhao, J.H., Luan, J., Andersen, L.B., Bingham, S.A., Brage, S., Smith, G.D., Ekelund, U.*, et al.* (2009). Genetic variation in LIN28B is associated with the timing of puberty. Nature genetics *41*, 729-733.

Ong, K.K., Elks, C.E., Wills, A.K., Wong, A., Wareham, N.J., Loos, R.J., Kuh, D., and Hardy, R. (2011). Associations between the pubertal timing-related variant in LIN28B and BMI vary across the life course. The Journal of clinical endocrinology and metabolism *96*, E125-129.

Pan, G.J., Chang, Z.Y., Scholer, H.R., and Pei, D. (2002). Stem cell pluripotency and transcription factor Oct4. Cell research *12*, 321-329.

Pan, L., Gong, Z., Zhong, Z., Dong, Z., Liu, Q., Le, Y., and Guo, J. (2011). Lin-28 reactivation is required for let-7 repression and proliferation in human small cell lung cancer cells. Molecular and cellular biochemistry *355*, 257-263.

Park, S.K., Venable, J.D., Xu, T., and Yates, J.R., 3rd (2008). A quantitative analysis software tool for mass spectrometry-based proteomics. Nature methods *5*, 319-322.

Pasquinelli, A.E., Reinhart, B.J., Slack, F., Martindale, M.Q., Kuroda, M.I., Maller, B., Hayward, D.C., Ball, E.E., Degnan, B., Muller, P.*, et al.* (2000). Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. Nature *408*, 86-89.

Patrakitkomjorn, S., Kobayashi, D., Morikawa, T., Wilson, M.M., Tsubota, N., Irie, A., Ozawa, T., Aoki, M., Arimura, N., Kaibuchi, K.*, et al.* (2008). Neurofibromatosis type 1 (NF1) tumor suppressor, neurofibromin, regulates the neuronal differentiation of PC12 cells via its associating protein, CRMP-2. J Biol Chem *283*, 9399-9413.

Peng, S., Chen, L.L., Lei, X.X., Yang, L., Lin, H., Carmichael, G.G., and Huang, Y. (2011). Genome-wide studies reveal that Lin28 enhances the translation of genes important for growth and survival of human embryonic stem cells. Stem Cells *29*, 496-504.

Peng, S., Maihle, N.J., and Huang, Y. (2010). Pluripotency factors Lin28 and Oct4 identify a sub-population of stem cell-like cells in ovarian cancer. Oncogene *29*, 2153-2159.

Perry, J.R., Stolk, L., Franceschini, N., Lunetta, K.L., Zhai, G., McArdle, P.F., Smith, A.V., Aspelund, T., Bandinelli, S., Boerwinkle, E.*, et al.* (2009). Meta-analysis of genome-wide association data identifies two loci influencing age at menarche. Nature genetics *41*, 648-650.

Perycz, M., Urbanska, A.S., Krawczyk, P.S., Parobczak, K., and Jaworski, J. (2011). Zipcode binding protein 1 regulates the development of dendritic arbors in hippocampal neurons. The Journal of neuroscience : the official journal of the Society for Neuroscience *31*, 5271-5285.

Pesce, M., and Scholer, H.R. (2001). Oct-4: gatekeeper in the beginnings of mammalian development. Stem Cells *19*, 271-278.

Piskounova, E., Polytarchou, C., Thornton, J.E., LaPierre, R.J., Pothoulakis, C., Hagan, J.P., Iliopoulos, D., and Gregory, R.I. (2011). Lin28A and Lin28B inhibit let-7 microRNA biogenesis by distinct mechanisms. Cell *147*, 1066-1079.

Piskounova, E., Viswanathan, S.R., Janas, M., LaPierre, R.J., Daley, G.Q., Sliz, P., and Gregory, R.I. (2008). Determinants of microRNA processing inhibition by the developmentally regulated RNA-binding protein Lin28. The Journal of biological chemistry *283*, 21310-21314.

Place, R.F., Li, L.C., Pookot, D., Noonan, E.J., and Dahiya, R. (2008). MicroRNA-373 induces expression of genes with complementary promoter sequences. Proceedings of the National Academy of Sciences of the United States of America *105*, 1608-1613.

Polesskaya, A., Cuvellier, S., Naguibneva, I., Duquet, A., Moss, E.G., and Harel-Bellan, A. (2007). Lin-28 binds IGF-2 mRNA and participates in skeletal myogenesis by increasing translation efficiency. Genes & development *21*, 1125-1138.

Polymenidou, M., Lagier-Tourenne, C., Hutt, K.R., Huelga, S.C., Moran, J., Liang, T.Y., Ling, S.C., Sun, E., Wancewicz, E., Mazur, C.*, et al.* (2011). Long pre-mRNA depletion and RNA missplicing contribute to neuronal vulnerability from loss of TDP-43. Nature neuroscience *14*, 459-468.

Ptacek, J., Devgan, G., Michaud, G., Zhu, H., Zhu, X., Fasolo, J., Guo, H., Jona, G., Breitkreutz, A., Sopko, R.*, et al.* (2005). Global analysis of protein phosphorylation in yeast. Nature *438*, 679-684.

Qiao, C., Ma, J., Xu, J., Xie, M., Ma, W., and Huang, Y. (2012). Drosha mediates destabilization of Lin28 mRNA targets. Cell Cycle *11*, 3590-3598.

Qiu, C., Ma, Y., Wang, J., Peng, S., and Huang, Y. (2010). Lin28-mediated post-transcriptional regulation of Oct4 expression in human embryonic stem cells. Nucleic acids research *38*, 1240-1248.

Reinhart, B.J., Slack, F.J., Basson, M., Pasquinelli, A.E., Bettinger, J.C., Rougvie, A.E., Horvitz, H.R., and Ruvkun, G. (2000). The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. Nature *403*, 901-906.

Reubinoff, B.E., Pera, M.F., Fong, C.Y., Trounson, A., and Bongso, A. (2000). Embryonic stem cell lines from human blastocysts: somatic differentiation in vitro. Nature biotechnology *18*, 399-404.

Richards, M., Fong, C.Y., Chan, W.K., Wong, P.C., and Bongso, A. (2002). Human feeders support prolonged undifferentiated growth of human inner cell masses and embryonic stem cells. Nature biotechnology *20*, 933-936.

Richards, M., Tan, S.P., Tan, J.H., Chan, W.K., and Bongso, A. (2004). The transcriptome profile of human embryonic stem cells as defined by SAGE. Stem Cells *22*, 51-64.

Rizzino, A. (2002). Embryonic stem cells provide a powerful and versatile model system. Vitamins and hormones *64*, 1-42.

Robb, G.B., and Rana, T.M. (2007). RNA helicase A interacts with RISC in human cells and functions in RISC loading. Molecular cell *26*, 523-537.

Rodda, D.J., Chew, J.L., Lim, L.H., Loh, Y.H., Wang, B., Ng, H.H., and Robson, P. (2005). Transcriptional regulation of nanog by OCT4 and SOX2. The Journal of biological chemistry *280*, 24731-24737.

Rodini, C.O., Suzuki, D.E., Saba-Silva, N., Cappellano, A., de Souza, J.E., Cavalheiro, S., Toledo, S.R., and Okamoto, O.K. (2012). Expression analysis of stem cell-related genes reveal OCT4 as a predictor of poor clinical outcome in medulloblastoma. Journal of neuro-oncology *106*, 71-79.

Rybak, A., Fuchs, H., Hadian, K., Smirnova, L., Wulczyn, E.A., Michel, G., Nitsch, R., Krappmann, D., and Wulczyn, F.G. (2009). The let-7 target gene mouse lin-41 is a stem cell specific E3 ubiquitin ligase for the miRNA pathway protein Ago2. Nature cell biology *11*, 1411-1420.

Rybak, A., Fuchs, H., Smirnova, L., Brandt, C., Pohl, E.E., Nitsch, R., and Wulczyn, F.G. (2008). A feedback loop comprising lin-28 and let-7 controls pre-let-7 maturation during neural stem-cell commitment. Nature cell biology *10*, 987-993.

Sakurai, M., Miki, Y., Masuda, M., Hata, S., Shibahara, Y., Hirakawa, H., Suzuki, T., and Sasano, H. (2012). LIN28: a regulator of tumor-suppressing activity of let-7 microRNA in human breast cancer. The Journal of steroid biochemistry and molecular biology *131*, 101-106.

Sampson, V.B., Rong, N.H., Han, J., Yang, Q., Aris, V., Soteropoulos, P., Petrelli, N.J., Dunn, S.P., and Krueger, L.J. (2007). MicroRNA let-7a down-regulates MYC and reverts MYC-induced growth in Burkitt lymphoma cells. Cancer research *67*, 9762-9770.

Sauliere, J., Murigneux, V., Wang, Z., Marquenet, E., Barbosa, I., Le Tonqueze, O., Audic, Y., Paillard, L., Roest Crollius, H., and Le Hir, H. (2012). CLIP-seq of eIF4AIII reveals transcriptome-wide mapping of the human exon junction complex. Nat Struct Mol Biol *19*, 1124-1131.

Shinoda, G., Shyh-Chang, N., Soysa, T.Y., Zhu, H., Seligson, M.T., Shah, S.P., Abo-Sido, N., Yabuuchi, A., Hagan, J.P., Gregory, R.I.*, et al.* (2013). Fetal deficiency of lin28 programs life-long aberrations in growth and glucose metabolism. Stem Cells *31*, 1563-1573.

Shyh-Chang, N., and Daley, G.Q. (2013). Lin28: primal regulator of growth and metabolism in stem cells. Cell stem cell *12*, 395-406.

Shyh-Chang, N., Daley, G.Q., and Cantley, L.C. (2013a). Stem cell metabolism in tissue development and aging. Development *140*, 2535-2547.

Shyh-Chang, N., Zhu, H., Yvanka de Soysa, T., Shinoda, G., Seligson, M.T., Tsanov, K.M., Nguyen, L., Asara, J.M., Cantley, L.C., and Daley, G.Q. (2013b). Lin28 enhances tissue repair by reprogramming cellular metabolism. Cell *155*, 778-792.

Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Soding, J.*, et al.* (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Molecular systems biology *7*, 539.

Singh, G., Ricci, E.P., and Moore, M.J. (2013). RIPiT-Seq: A high-throughput approach for footprinting RNA:protein complexes. Methods.

Siomi, M.C., Sato, K., Pezic, D., and Aravin, A.A. (2011). PIWI-interacting small RNAs: the vanguard of genome defence. Nature reviews Molecular cell biology *12*, 246-258.

Slack, F., and Ruvkun, G. (1997). Temporal pattern formation by heterochronic genes. Annual review of genetics *31*, 611-634.

Slack, F.J., Basson, M., Liu, Z., Ambros, V., Horvitz, H.R., and Ruvkun, G. (2000). The lin-41 RBCC gene acts in the C. elegans heterochronic pathway between the let-7 regulatory RNA and the LIN-29 transcription factor. Molecular cell *5*, 659-669.

Sonenberg, N., and Hinnebusch, A.G. (2009). Regulation of translation initiation in eukaryotes: mechanisms and biological targets. Cell *136*, 731-745.

Song, H., Li, H., Huang, M., Xu, D., Gu, C., Wang, Z., Dong, F., and Wang, F. (2013). Induced pluripotent stem cells from goat fibroblasts. Molecular reproduction and development.

Stark, T.J., Arnold, J.D., Spector, D.H., and Yeo, G.W. (2011). High-resolution profiling and analysis of viral and host small RNAs during human cytomegalovirus infection. Journal of virology.

Stelzl, U., Worm, U., Lalowski, M., Haenig, C., Brembeck, F.H., Goehler, H., Stroedicke, M., Zenkner, M., Schoenherr, A., Koeppen, S.*, et al.* (2005). A human protein-protein interaction network: a resource for annotating the proteome. Cell *122*, 957-968.

Su, A.I., Wiltshire, T., Batalov, S., Lapp, H., Ching, K.A., Block, D., Zhang, J., Soden, R., Hayakawa, M., Kreiman, G.*, et al.* (2004). A gene atlas of the mouse and human protein-encoding transcriptomes. Proceedings of the National Academy of Sciences of the United States of America *101*, 6062-6067.

Sugnet, C.W., Srinivasan, K., Clark, T.A., O'Brien, G., Cline, M.S., Wang, H., Williams, A., Kulp, D., Blume, J.E., Haussler, D.*, et al.* (2006). Unusual intron conservation near tissue-regulated exons found by splicing microarrays. PLoS computational biology *2*, e4.

Suh, M.R., Lee, Y., Kim, J.Y., Kim, S.K., Moon, S.H., Lee, J.Y., Cha, K.Y., Chung, H.M., Yoon, H.S., Moon, S.Y.*, et al.* (2004). Human embryonic stem cells express a unique set of microRNAs. Developmental biology *270*, 488-498.

Suh, N., Baehner, L., Moltzahn, F., Melton, C., Shenoy, A., Chen, J., and Blelloch, R. (2010). MicroRNA function is globally suppressed in mouse oocytes and early embryos. Current biology : CB *20*, 271-277.

Sulem, P., Gudbjartsson, D.F., Rafnar, T., Holm, H., Olafsdottir, E.J., Olafsdottir, G.H., Jonsson, T., Alexandersen, P., Feenstra, B., Boyd, H.A.*, et al.* (2009). Genome-wide association study identifies sequence variants on 6q21 associated with age at menarche. Nature genetics *41*, 734-738.

Sung, Y.H., Baek, I.J., Kim, D.H., Jeon, J., Lee, J., Lee, K., Jeong, D., Kim, J.S., and Lee, H.W. (2013). Knockout mice created by TALEN-mediated gene targeting. Nature biotechnology *31*, 23-24.

Tabara, H., Sarkissian, M., Kelly, W.G., Fleenor, J., Grishok, A., Timmons, L., Fire, A., and Mello, C.C. (1999). The rde-1 gene, RNA interference, and transposon silencing in C. elegans. Cell *99*, 123-132.

Tabb, D.L., McDonald, W.H., and Yates, J.R., 3rd (2002). DTASelect and Contrast: tools for assembling and comparing protein identifications from shotgun proteomics. Journal of proteome research *1*, 21-26.

Tang, F., Barbacioru, C., Bao, S., Lee, C., Nordman, E., Wang, X., Lao, K., and Surani, M.A. (2010). Tracing the derivation of embryonic stem cells from the inner cell mass by single-cell RNA-Seq analysis. Cell stem cell *6*, 468-478.

Tarassov, K., Messier, V., Landry, C.R., Radinovic, S., Serna Molina, M.M., Shames, I., Malitskaya, Y., Vogel, J., Bussey, H., and Michnick, S.W. (2008). An in vivo map of the yeast protein interactome. Science *320*, 1465-1470.

Thomson, J.A., Itskovitz-Eldor, J., Shapiro, S.S., Waknitz, M.A., Swiergiel, J.J., Marshall, V.S., and Jones, J.M. (1998). Embryonic stem cell lines derived from human blastocysts. Science *282*, 1145-1147.

Thomson, J.M., Newman, M., Parker, J.S., Morin-Kensicki, E.M., Wright, T., and Hammond, S.M. (2006). Extensive post-transcriptional regulation of microRNAs and its implications for cancer. Genes & development *20*, 2202-2207.

Thornton, J.E., Chang, H.M., Piskounova, E., and Gregory, R.I. (2012). Lin28-mediated control of let-7 microRNA expression by alternative TUTases Zcchc11 (TUT4) and Zcchc6 (TUT7). RNA *18*, 1875-1885.

Thornton, J.E., and Gregory, R.I. (2012). How does Lin28 let-7 control development and disease? Trends in cell biology *22*, 474-482.

Tomioka, I., Maeda, T., Shimada, H., Kawai, K., Okada, Y., Igarashi, H., Oiwa, R., Iwasaki, T., Aoki, M., Kimura, T.*, et al.* (2010). Generating induced pluripotent stem cells from common marmoset (Callithrix jacchus) fetal liver cells using defined

factors, including Lin28. Genes to cells : devoted to molecular & cellular mechanisms *15*, 959-969.

Tong, A.H., Evangelista, M., Parsons, A.B., Xu, H., Bader, G.D., Page, N., Robinson, M., Raghibizadeh, S., Hogue, C.W., Bussey, H.*, et al.* (2001). Systematic genetic analysis with ordered arrays of yeast deletion mutants. Science *294*, 2364-2368.

Typas, A., Nichols, R.J., Siegele, D.A., Shales, M., Collins, S.R., Lim, B., Braberg, H., Yamamoto, N., Takeuchi, R., Wanner, B.L.*, et al.* (2008). High-throughput, quantitative analyses of genetic interactions in E. coli. Nature methods *5*, 781-787.

Urnov, F.D., Miller, J.C., Lee, Y.L., Beausejour, C.M., Rock, J.M., Augustus, S., Jamieson, A.C., Porteus, M.H., Gregory, P.D., and Holmes, M.C. (2005). Highly efficient endogenous human gene correction using designed zinc-finger nucleases. Nature *435*, 646-651.

Vadla, B., Kemper, K., Alaimo, J., Heine, C., and Moss, E.G. (2012). lin-28 controls the succession of cell fate choices via two distinct activities. PLoS genetics *8*, e1002588.

Vallier, L., Alexander, M., and Pedersen, R.A. (2005). Activin/Nodal and FGF pathways cooperate to maintain pluripotency of human embryonic stem cells. Journal of cell science *118*, 4495-4509.

Van Wynsberghe, P.M., Kai, Z.S., Massirer, K.B., Burton, V.H., Yeo, G.W., and Pasquinelli, A.E. (2011). LIN-28 co-transcriptionally binds primary let-7 to regulate miRNA maturation in Caenorhabditis elegans. Nat Struct Mol Biol *18*, 302-308.

Vasudevan, S., Tong, Y., and Steitz, J.A. (2007). Switching from repression to activation: microRNAs can up-regulate translation. Science *318*, 1931-1934.

Veenstra, T.D., Martinovic, S., Anderson, G.A., Pasa-Tolic, L., and Smith, R.D. (2000). Proteome analysis using selective incorporation of isotopically labeled amino acids. Journal of the American Society for Mass Spectrometry *11*, 78-82.

Venables, J.P., Klinck, R., Koh, C., Gervais-Bird, J., Bramard, A., Inkel, L., Durand, M., Couture, S., Froehlich, U., Lapointe, E.*, et al.* (2009). Cancer-associated regulation of alternative splicing. Nat Struct Mol Biol *16*, 670-676.

Viswanathan, S.R., Daley, G.Q., and Gregory, R.I. (2008). Selective blockade of microRNA processing by Lin28. Science *320*, 97-100.

Viswanathan, S.R., Powers, J.T., Einhorn, W., Hoshida, Y., Ng, T.L., Toffanin, S., O'Sullivan, M., Lu, J., Phillips, L.A., Lockhart, V.L.*, et al.* (2009). Lin28 promotes

transformation and is associated with advanced human malignancies. Nature genetics *41*, 843-848.

Wang, S.H., Tsai, M.S., Chiang, M.F., and Li, H. (2003). A novel NK-type homeobox gene, ENK (early embryo specific NK), preferentially expressed in embryonic stem cells. Gene expression patterns : GEP *3*, 99-103.

Wang, Y.C., Chen, Y.L., Yuan, R.H., Pan, H.W., Yang, W.C., Hsu, H.C., and Jeng, Y.M. (2010). Lin-28B expression promotes transformation and invasion in human hepatocellular carcinoma. Carcinogenesis *31*, 1516-1522.

Weinlich, S., Huttelmaier, S., Schierhorn, A., Behrens, S.E., Ostareck-Lederer, A., and Ostareck, D.H. (2009). IGF2BP1 enhances HCV IRES-mediated translation initiation via the 3'UTR. RNA *15*, 1528-1542.

West, J.A., Viswanathan, S.R., Yabuuchi, A., Cunniff, K., Takeuchi, A., Park, I.H., Sero, J.E., Zhu, H., Perez-Atayde, A., Frazier, A.L.*, et al.* (2009). A role for Lin28 in primordial germ-cell development and germ-cell malignancy. Nature *460*, 909-913.

Wightman, B., Burglin, T.R., Gatto, J., Arasu, P., and Ruvkun, G. (1991). Negative regulatory sequences in the lin-14 3'-untranslated region are necessary to generate a temporal switch during Caenorhabditis elegans development. Genes & development *5*, 1813-1824.

Wightman, B., Ha, I., and Ruvkun, G. (1993). Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans. Cell *75*, 855-862.

Wilbert, M.L., Huelga, S.C., Kapeli, K., Stark, T.J., Liang, T.Y., Chen, S.X., Yan, B.Y., Nathanson, J.L., Hutt, K.R., Lovci, M.T.*, et al.* (2012). LIN28 binds messenger RNAs at GGAGA motifs and regulates splicing factor abundance. Molecular cell *48*, 195-206.

Wolters, D.A., Washburn, M.P., and Yates, J.R., 3rd (2001). An automated multidimensional protein identification technology for shotgun proteomics. Analytical chemistry *73*, 5683-5690.

Wu, C., Macleod, I., and Su, A.I. (2013). BioGPS and MyGene.info: organizing online, gene-centric information. Nucleic acids research *41*, D561-565.

Wu, C., Orozco, C., Boyer, J., Leglise, M., Goodale, J., Batalov, S., Hodge, C.L., Haase, J., Janes, J., Huss, J.W., 3rd*, et al.* (2009). BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. Genome biology *10*, R130.

Wu, L., and Belasco, J.G. (2005). Micro-RNA regulation of the mammalian lin-28 gene during neuronal differentiation of embryonal carcinoma cells. Molecular and cellular biology *25*, 9198-9208.

Xu, B., and Huang, Y. (2009). Histone H2a mRNA interacts with Lin28 and contains a Lin28-dependent posttranscriptional regulatory element. Nucleic acids research *37*, 4256-4263.

Xu, B., Zhang, K., and Huang, Y. (2009). Lin28 modulates cell growth and associates with a subset of cell cycle regulator mRNAs in mouse embryonic stem cells. Rna *15*, 357-361.

Xu, R.H., Chen, X., Li, D.S., Li, R., Addicks, G.C., Glennon, C., Zwaka, T.P., and Thomson, J.A. (2002). BMP4 initiates human embryonic stem cell differentiation to trophoblast. Nature biotechnology *20*, 1261-1264.

Xu, R.H., Peck, R.M., Li, D.S., Feng, X., Ludwig, T., and Thomson, J.A. (2005). Basic FGF and suppression of BMP signaling sustain undifferentiated proliferation of human ES cells. Nature methods *2*, 185-190.

Xu, T., Venable, J.D., Park, S.K., Cociorva, D., Lu, B., Liao, L., Wohlschlegel, J., Hewel, J., and Yates, J.R. (2006). ProLuCID, a fast and sensitive tandem mass spectra-based protein identification program. Molecular & Cellular Proteomics *5*, S174-S174.

Xue, Y., Zhou, Y., Wu, T., Zhu, T., Ji, X., Kwon, Y.S., Zhang, C., Yeo, G., Black, D.L., Sun, H.*, et al.* (2009). Genome-wide analysis of PTB-RNA interactions reveals a strategy used by the general splicing repressor to modulate exon inclusion or skipping. Molecular cell *36*, 996-1006.

Yang, D.H., and Moss, E.G. (2003). Temporally regulated expression of Lin-28 in diverse tissues of the developing mouse. Gene expression patterns : GEP *3*, 719-726.

Yeo, G.W., Coufal, N.G., Liang, T.Y., Peng, G.E., Fu, X.D., and Gage, F.H. (2009). An RNA code for the FOX2 splicing regulator revealed by mapping RNA-protein interactions in stem cells. Nat Struct Mol Biol *16*, 130-137.

Yeo, G.W., Xu, X., Liang, T.Y., Muotri, A.R., Carson, C.T., Coufal, N.G., and Gage, F.H. (2007). Alternative splicing events identified in human embryonic stem cells and neural progenitors. PLoS computational biology *3*, 1951-1967.

Yi, R., Qin, Y., Macara, I.G., and Cullen, B.R. (2003). Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs. Genes & development *17*, 3011-3016.

Ying, Q.L., Nichols, J., Chambers, I., and Smith, A. (2003). BMP induction of Id proteins suppresses differentiation and sustains embryonic stem cell self-renewal in collaboration with STAT3. Cell *115*, 281-292.

Yu, H., Braun, P., Yildirim, M.A., Lemmens, I., Venkatesan, K., Sahalie, J., Hirozane-Kishikawa, T., Gebreab, F., Li, N., Simonis, N.*, et al.* (2008). High-quality binary protein interaction map of the yeast interactome network. Science *322*, 104-110.

Yu, J., Vodyanik, M.A., Smuga-Otto, K., Antosiewicz-Bourget, J., Frane, J.L., Tian, S., Nie, J., Jonsdottir, G.A., Ruotti, V., Stewart, R.*, et al.* (2007). Induced pluripotent stem cell lines derived from human somatic cells. Science *318*, 1917-1920.

Yuan, H., Corbi, N., Basilico, C., and Dailey, L. (1995). Developmental-specific activity of the FGF-4 enhancer requires the synergistic action of Sox2 and Oct-3. Genes & development *9*, 2635-2645.

Zhang, J., Zhang, L., Fan, R., Guo, N., Xiong, C., Wang, L., Jin, S., Li, W., and Lu, J. (2013). The polymorphism in the let-7 targeted region of the Lin28 gene is associated with increased risk of type 2 diabetes mellitus. Molecular and cellular endocrinology *375*, 53-57.

Zhang, W.C., Shyh-Chang, N., Yang, H., Rai, A., Umashankar, S., Ma, S., Soh, B.S., Sun, L.L., Tai, B.C., Nga, M.E.*, et al.* (2012). Glycine decarboxylase activity drives non-small cell lung cancer tumor-initiating cells and tumorigenesis. Cell *148*, 259-272.

Zhou, J., Ng, S.B., and Chng, W.J. (2013). LIN28/LIN28B: an emerging oncogenic driver in cancer stem cells. The international journal of biochemistry & cell biology *45*, 973-978.

Zhu, H., Shah, S., Shyh-Chang, N., Shinoda, G., Einhorn, W.S., Viswanathan, S.R., Takeuchi, A., Grasemann, C., Rinn, J.L., Lopez, M.F.*, et al.* (2010). Lin28a transgenic mice manifest size and puberty phenotypes identified in human genetic association studies. Nature genetics *42*, 626-630.

Zhu, H., Shyh-Chang, N., Segre, A.V., Shinoda, G., Shah, S.P., Einhorn, W.S., Takeuchi, A., Engreitz, J.M., Hagan, J.P., Kharas, M.G.*, et al.* (2011). The Lin28/let-7 axis regulates glucose metabolism. Cell *147*, 81-94.

Zisoulis, D.G., Kai, Z.S., Chang, R.K., and Pasquinelli, A.E. (2012). Autoregulation of microRNA biogenesis by let-7 and Argonaute. Nature *486*, 541-544.

Zisoulis, D.G., Lovci, M.T., Wilbert, M.L., Hutt, K.R., Liang, T.Y., Pasquinelli, A.E., and Yeo, G.W. (2010). Comprehensive discovery of endogenous Argonaute binding sites in Caenorhabditis elegans. Nat Struct Mol Biol *17*, 173-179.