Thesis submitted for the degree of Doctor of Philosophy
at the University of Oxford

Intergenic Long Noncoding RNAs Provide a Novel Layer of Post-transcriptional Regulation in Development and Disease

Jennifer Yihong Tan

KEBLE COLLEGE
OXFORD

TRINITY TERM
2014

Intergenic long noncoding RNAs provide a novel layer of post-transcriptional regulation in development and disease

Thesis submitted for the degree of Doctor of Philosophy at the University of Oxford
Jennifer Y. Tan, Keble College, Trinity Term 2014

ABSTRACT

Recent genome-wide sequencing projects revealed the pervasive transcription of intergenic long noncoding RNAs (lincRNAs) in eukaryotic genomes (reviewed in Ponting et al. 2009). For the vast majority of lincRNAs, their mechanisms of function remain largely unrecognized. However, the genome-wide signatures of functionality associated with many lincRNAs, including apparent evolutionary sequence conservation, spatial and temporal-restricted expression patterns, strong associations with epigenetic marks, and reported molecular and cellular functions, reinforce their biological relevance. My work investigates lincRNAs that post-transcriptionally regulate gene abundance by competing for the binding of common microRNAs (miRNAs) with protein-coding transcripts, termed competitive endogenous RNAs (ceRNAs) acting lincRNAs (lnceRNAs). First, I examine the biological relevance of this post-transcriptional regulation of gene abundance by ceRNAs. Next, I estimate the genome-wide prevalence of lnceRNAs in mouse embryonic stem cells (mESCs) and characterize their properties. Finally, using two specific examples of lnceRNAs, I show the contributions of lnceRNAs to human monogenic and complex trait diseases. Collectively, these results illustrate that lnceRNAs provide a novel layer of post-transcriptional regulation via a miRNA-mediated mechanism that contributes to organismal and cellular biology.

In dedication to my beloved grandfather, who has always been my source of knowledge and inspiration.

ACKNOWLEDGEMENTS

First of all, I thank my supervisors Prof. Chris Ponting and Dr. Ana Marques for introducing me to the world of long noncoding RNAs. Thank you for invaluable guidance, continuing encouragement and support. Most importantly, your dedications and enthusiasms to science have been a great source of inspiration to me.

I thank all past and present members of the Ponting group for creating a pleasant and stimulating research environment to work in. In particular, I thank Tamara Sirey, Kenny Roberts, Wilfried Hearty, Yang Li and Sergei Maslau for inspiring discussions and helpful feedbacks. I am also grateful for all my collaborators, without whose contributions, my work and submitted manuscripts for publication would not have been possible.

I am grateful for the Clarendon Fund, the Natural Sciences Engineering Research Council of Canada, the UK Medical Research Council, and the Sloane Robinson Clarendon Award from Keble College, Oxford for providing me with financial support.

On a personal level, I thank Chris Rands for always being there, Quinten Ferry for entertaining chats, A.C. and D.H. for moral support. I thank my best friends at home, Alice Wong, Carol Lee and Nancy Chow, for their continuing friendship and support.

Lastly, I thank my parents, grandparents, and pet dog Fluffy for their ongoing love and support during the past four years.

Table of Contents

CHAPTER 1

Introduction

1.1 PERVASIVE TRANSCRIPTION OF EUKARYOTIC GENOMES

The human genome was sequenced over a decade ago (Lander et al., 2001; Venter et al., 2001), yet we are only beginning to understand the complexity of functional information encoded within it. Genomes of eukaryotic species, such as *Caenorhabditis elegans* (approximately 18,000 protein-coding genes) (Wilson et al., 1994), *Drosophila melanogaster* (approximately 14,000 protein-coding genes) (Adams et al., 2000) and *Arabidopsis thaliana* (approximately 25,000 protein-coding genes) (Arabidopsis Genome, 2000), had been sequenced previously. Completion of the human genome led to the surprising revelation that it consists of a similar number of protein-coding genes to the other genomes (approximately 20,000-25,000, <2% of the genome) (International Human Genome Sequencing, 2004), now estimated at approximately 19,000-20,500 (1.1-1.2% of the genome) (Clamp et al., 2007; Church et al., 2009). As a result, it became apparent that the number of protein-coding genes alone is not sufficient to explain the differences in organismal complexity (Mattick, 2004a).

Recently, the Encyclopedia of DNA Elements (ENCODE) project, which catalogued the genome-wide transcription landscape across 147 cells lines and developmental stages, made the claim that at least 80% of the human genome

is associated with an active biochemical function, including signatures such as DNA-protein regulatory interactions (i.e. DNase1 hypersensitivity sites, transcriptional factor binding sites) and/or RNA transcription across noncoding regions (Bernstein et al., 2012). By considering only the portion of the human genome subjected to purifying selection, a more accurate estimate of approximately 8.2% (7.1-9.2%) of the genome exhibits signatures of sequence constraint, and is thus likely to be biologically functional (Meader et al., 2010; Rands et al., 2014).

Therefore, although mechanisms, including alternative splicing, RNA editing, and trans-splicing, have been suggested to allow a more varied proteome/transcriptome in a genome of limited size (reviewed in Keren et al. 2010), these mechanisms alone cannot account for the substantial difference between the amount of the genome under functional constraint (8.2%) and the portion that encodes protein-coding sequences (1.1-1.2%). Such observations suggest that a large proportion of functional information is embedded within noncoding regions of the genome to create a more diversified transcriptome, which contribute significantly to the eukaryotic genome's capacity to generate phenotypic complexity (Mattick, 2001).

Indeed, a more complex picture of the transcriptome emerged first from a series of tiling array experiments where most of the non-repetitive bases in the human genomes were covered by interleaving transcripts, including exonic, intronic and intergenic regions (Kapranov et al., 2002; Rinn et al., 2003; Bertone et al., 2004). Nearly half of the sequenced human cDNA clones have no apparent open reading frame (ORF) (Ota et al., 2004), whereas absence of a clear ORF

was also identified in nearly a third of frequently complete high quality cDNAs in mouse (FANTOM consortium) (Carninci et al., 2005). These large-scale expressed sequence tag (EST) and cDNA sequence datasets collected from multiple studies of RNA transcripts consistently indicated a widespread transcriptional activity in eukaryotic genomes beyond known protein-coding gene annotations.

Subsequently, the emergence of a second generation of sequencing technologies, referred to as next generation sequencing (NGS) technologies, circumvented the limited coverage frequently permitted by EST sequences and high rate of false positive results that are often associated with microarray-based studies of the transcriptome, largely generated from cross-hybridization between probes or high levels of background fluorescence (Ponting and Belgard, 2010). NGS technologies, such as whole-transcriptome shotgun sequencing (RNA-seq) experiments, which are able to generate millions of short sequencing reads in parallel, were used to sequence the transcriptome of various eukaryotic species (i.e. mouse (Mortazavi et al. 2008) and human (Cabili et al. 2010)). These experiments reinforced the notion that the transcribed noncoding proportion of mammalian genomes is significantly greater than what was previously thought (Figure 1.1).

Such large-scale studies in human and mouse have also been extended to other model organisms. For example, The Model Organism ENCyclopedia Of DNA Elements (modENCODE) project catalogues sequence-based functional elements, including noncoding RNA transcripts in *Drosophila melanogaster* (Roy et al., 2010) and *Caenorhabditis elegans* (Celniker et al., 2009; Gerstein

et al., 2010; Roy et al., 2010; Ulitsky and Bartel, 2013). Transcriptome profiles of other species, including nematode (Celniker et al., 2009), zebrafish (Ulitsky et al., 2011), frog (Tan et al., 2013), *Arabidopsis* (Liu et al., 2012a), and maize (Li et al., 2014), support the notion that pervasive transcription of the genome is an evolutionarily widespread phenomenon.

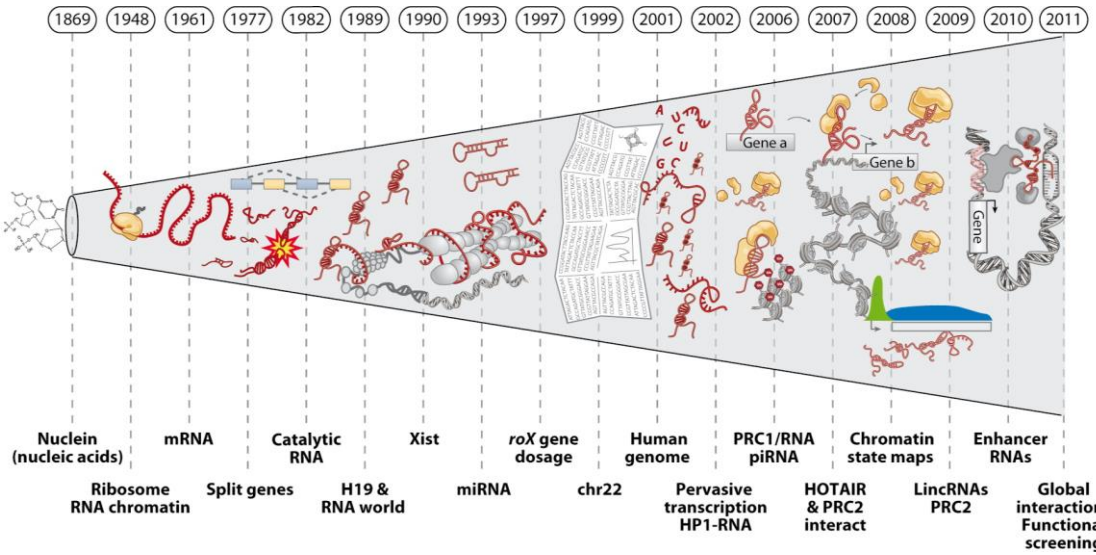


Figure 1.1 Timeline of discoveries of functional RNAs in biological regulation (Rinn and Chang, 2012).

1.2 NONCODING RNAS

Noncoding RNAs are initially defined as RNA products that are transcribed into RNA transcript but lack apparent open reading frames (ORFs). Noncoding RNAs can range in size from tens up to kilobases (kb) of nucleotides (nt) in length. They are often loosely classified based on their size into two broad classes by an arbitrary length cutoff of 200 nt based on the practicality of RNA

purification protocols (Kapranov et al., 2007): (1) small noncoding RNAs that are typically no longer than 30 nt (Jung et al., 2010) and (2) long noncoding RNAs (lncRNAs) with lengths from 200 bp up to several kbs long (Mattick, 2001; Costa, 2005).

1.2.1 Small noncoding RNAs

The broad class of small noncoding RNAs can be further classified based on their sequence, structure, and functional similarity. Structural housekeeping noncoding RNAs that serve as infrastructural component of complexes involved in protein synthesis and RNA processing include transfer RNAs (tRNAs), ribosomal RNAs (rRNAs), small nuclear RNAs (snRNAs) and small nucleolar RNAs (snoRNAs) have been known and well-studied for many years (reviewed in Eddy, 2001).

More recently, many classes of small noncoding RNAs were found with crucial regulatory, often post-transcriptional, roles. These include small interfering RNAs (siRNAs) (Xia et al., 2002), microRNAs (miRNAs) (Bartel, 2004), and piwi-interacting RNAs (piRNAs) (Houwing et al., 2007). Typically processed from longer transcript precursors, these small RNAs can engage in transcript silencing via RNA interference (RNAi) pathways. Through shared sequence complementarity with their target transcripts, they interact and guide various members of Argonaute proteins (AGOs) to bind and repress the transcript levels of target sequences (Peters and Meister, 2007; Carthew and Sontheimer, 2009; Ghildiyal and Zamore, 2009; Voinnet, 2009).

1.2.2 Long noncoding RNAs

In contrast to small noncoding RNAs that are relatively well-characterized, their longer counterparts, long noncoding RNAs (>200 bp, lncRNAs) (reviewed in Ponting et al. 2009), are a group of RNA transcripts whose functional roles, if any, remain largely unknown. Broadly, lncRNAs can be categorized based on their relative genomic location to neighbouring protein-coding sequences (Figure 1.2). Specifically, lncRNAs can overlap coding sequences, where they may originate from exons, introns, or the 5' or 3' untranslated regions (UTRs) of protein-coding genes (reviewed in Ponting et al. 2009). LncRNAs can also have no sequence overlap with annotated protein-coding gene models, and these transcripts are often referred to as intergenic long ncRNAs (lincRNAs) (Guttman et al., 2009; Ponting et al., 2009). The expression patterns and potential functional characteristics of lincRNAs are easier to interpret due to their transcriptional independence and absence of other confounding factors potentially associated with overlapping protein-coding gene loci (Ulitsky and Bartel, 2013).

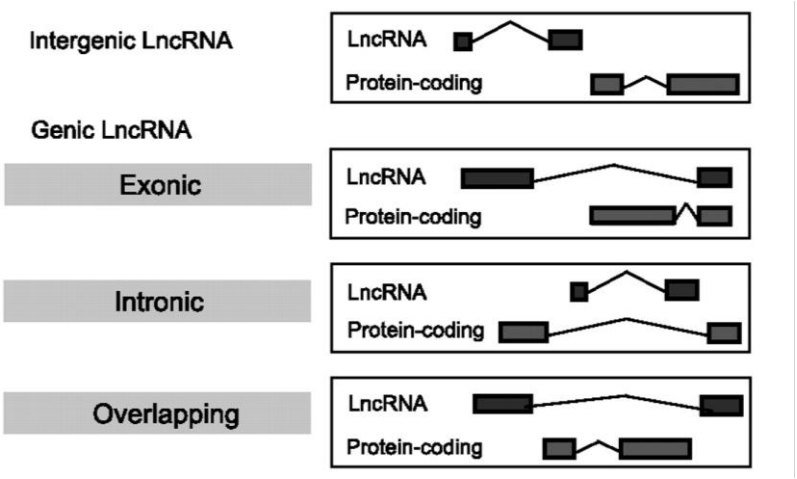


Figure 1.2 Classification of lncRNAs based on their relationship with protein-coding genes (Derrien et al., 2012).

1.3 MICRORNAS

1.3.1 Biogenesis and Turnover of miRNAs

Mature miRNAs are the product of a relatively complex biogenesis process (Figure 1.3) (Rodriguez et al., 2004; Saini et al., 2007). Individual primary miRNA transcripts (pri-miRNAs), which can encode one or more mature miRNAs (Cai et al., 2004), are transcribed predominantly by RNA Polymerase II (Pol ll) (Lee et al., 2004) and occasionally by RNA Pol III (Borchert et al., 2006). As for other Pol II transcripts, pri-miRNAs are also capped and polyadenylated (Cai et al., 2004; Lee et al., 2004). Pri-miRNAs can originate from intergenic promoters or overlap the introns or exons of both coding and noncoding transcripts (Cai et al., 2004). In addition, many miRNAs reside in clusters within the same polycistronic pri-miRNA transcripts (Lagos-Quintana et al., 2002).

While in the nucleus, pri-miRNAs are cleaved into hairpin stem-loop structures of approximately 70 nt in length, known as precursor transcripts (pre-miRNAs). Processed by a microprocessor complex, these hairpin structures contain a loop region and a ~33 nt stem region that is base paired completely or partially completely (Lee et al., 2003). Specifically, Drosha is an RNAse III endonuclease that is a component of the microprocessor complex (Lee et al., 2003) along with its co-factors, Partner of Drosha (Pasha) and DiGeorge Syndrome critical region 8 homolog (DGCR8) (Denli et al., 2004; Gregory et al., 2004). The Microprocessor complex recognizes the stem and the unpaired flanking regions of the hairpin. Subsequently, DGCR8 guides Drosha to cleave

the hairpin by its two RNase III domains (Zeng and Cullen, 2003, 2005), which leaves the stem with a 5' phosphate and a 3'OH on a 2-nt overhang (Figure 1.3).

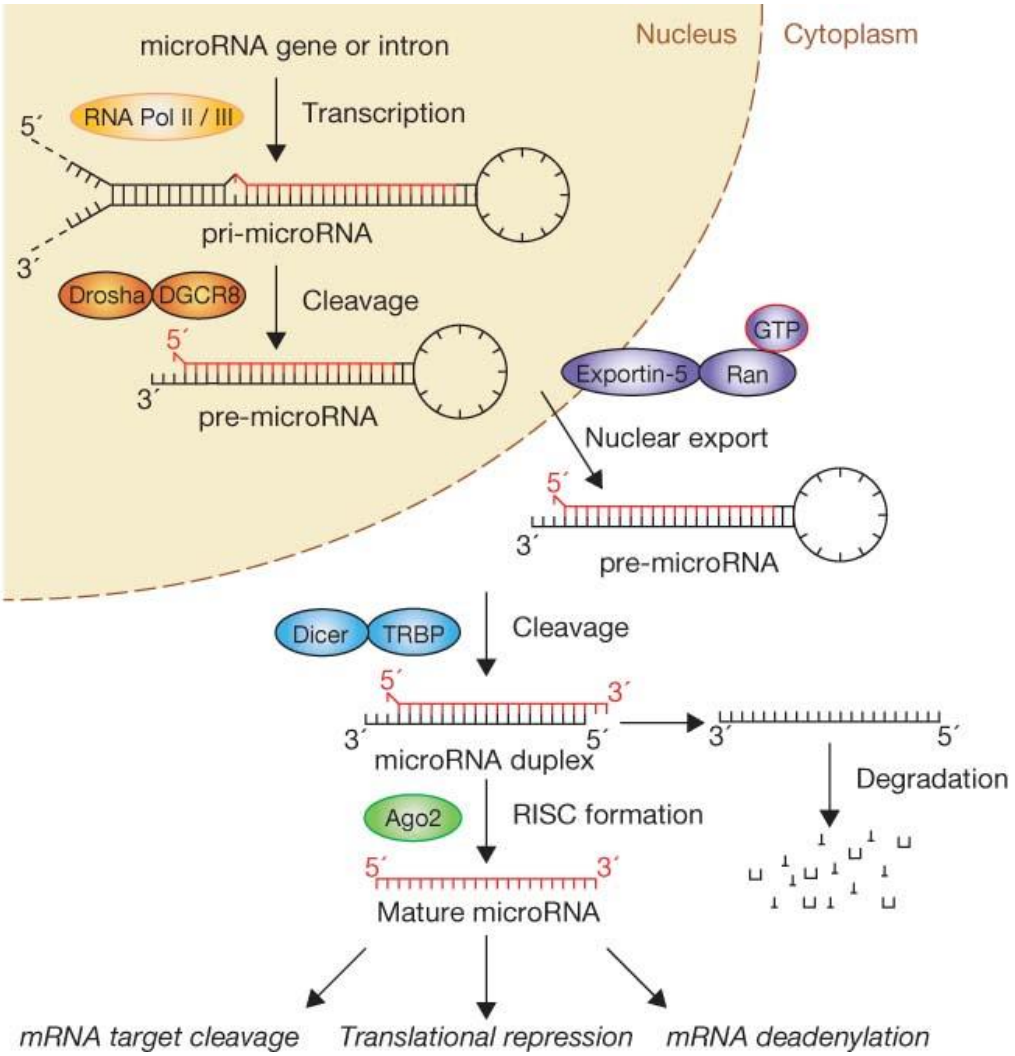


Figure 1.3 The mammalian miRNA biogenesis and processing pathway. In the nucleus, the primary miRNA transcript (pri-miRNA) is transcribed by mainly RNA polymerase II. After being cleaved by the microprocessor complex, Drosha–DGCR8, into precursor hairpin (pre-miRNA), it is exported to the cytoplasm. In the cytoplasm, the pre-miRNA hairpin is cleaved into its mature form by the RNase Dicer in complex with its cofactors. Along with Argonaute (AGO2) proteins, the functional strand of the mature miRNA is loaded into the RNA-induced silencing complex (RISC), where it guides RISC to silence target mRNAs through mRNA cleavage, translational repression or deadenylation. The passenger strand of the mature miRNA (black) is degraded. Figure is taken from Winter et al., 2009 and includes details and components of the pathway not described in the main text.

The processed pre-miRNAs hairpin structures are then exported from the nucleus into the cytoplasm (Lund et al., 2004), where they become cleaved by a second RNase III endonuclease, Dicer. Dicer removes stem loop of the pre-miRNA hairpin and produces a double stranded mature miRNA duplex of approximately 21-25 base pairs (bp) with a 5' phosphate and 3' OH on a 2 nt 3' overhangs (Gregory et al., 2004) (Figure 1.3). Following Dicer cleavage, one strand of the mature miRNA duplexes is incorporated into the RNA-induced silencing complex (RISC). Together, the mature miRNA-loaded RISC complex (miRISC) targets transcript sequences for post-transcriptional silencing. Loading of mature miRNAs into RISC requires the recognition of the miRNAs' 3' overhangs by the Argonaute (AGO) proteins within the RISC complex (Meister et al., 2004; Hutvagner and Simard, 2008). In principle, both strands of the miRNA duplex can become functional mature miRNAs (Marco et al., 2012). Despite this, for most miRNAs, only one strand (the guide strand) is preferentially loaded into the RISC complex. This strand is believed to be selected based on its higher thermodynamic stability (Khvorova et al., 2003; Schwarz et al., 2003; Krol et al., 2004). For a few miRNAs, both strands of the duplex are functional and are stably incorporated into the RISC complex independently.

1.3.2 Mechanisms of miRNA-Mediated Regulation of Gene Expression

MiRNA biogenesis is critical to the survival of eukaryotes. *Dicer*-null mouse (Bernstein et al., 2003) and zebrafish (Wienholds et al., 2003) embryos exhibit

developmental arrest and early embryonic death. Similarly, the deletion of one of the AGO family of proteins in flies, *dAGO1*, leads to their death during embryogenesis or early larval development stages (Kataoka et al., 2001; Okamura et al., 2004), whereas the deletion of two AGO-encoding genes, *alg-1* and *alg-2* in worms results in lethality early in development (Grishok et al., 2001).

MiRNA-mediated regulation of transcript abundance requires the recognition of target transcripts by the mature miRNA-loaded RISC complex (miRISC) (Bartel, 2009; Krol et al., 2010). The association between miRISC and its targets is dependent on Watson-Crick base pairing between the loaded miRNA and the miRNA response elements (MREs) embedded within the target sequence (Figure 1.4) (Bartel, 2004; Lewis et al., 2005; Bartel, 2009). The minimum sequence complementarity required for canonical miRNA-target association between a MRE and miRNA involves at least nts 2-8 from the 5' end of the mature miRNA (referred to as the 'seed' region) (Lewis et al., 2005; Bartel, 2009). Beyond that, the degree of sequence complementarity between the miRNA and its bound target appears, to some extent, to determine the mechanism of miRNA-mediated silencing (Bartel, 2009).

The steady-state level of miRNAs is determined by their rates of synthesis and decay. However, in contrast to our relatively extensive understanding of the processes underlying miRNA biogenesis, mechanisms that control miRNA decay and destabilization, which are likely to vary for different miRNAs, are less well understood (Liu, 2008; Kai and Pasquinelli, 2010; Gantier et al., 2011). Most mature miRNAs appear to be relatively stable and persist in the cell for

several days after their transcription and processing (Bail et al., 2010; Kai and Pasquinelli, 2010; Krol et al., 2010; Gantier et al., 2011). Target recognition, particularly through binding to highly complementary target sequences, can trigger miRNA degradation (Ameres et al., 2010b; Baccarini et al., 2011b; Wyman et al., 2011; Pasquinelli, 2012; Ruegger and Grosshans, 2012). Target-induced miRNA destabilization is often accompanied by the emergence of modified miRNA species with extended tails or trimmed ends and can induce miRNA degradation (Flynt et al., 2010). Found in various organisms, the target recognition-mediated regulation is likely an important contributor to the regulation of miRNA abundance (Ruegger and Grosshans, 2012).

Non-templated addition of uridines and adenosines is the predominant form of modifications at 3′ ends of mature miRNAs that can modulate miRNA stability (Li et al., 2005; Ramachandran and Chen, 2008; Ibrahim et al., 2010). While the addition of a polyadenylated (polyA) tail have been demonstrated to increase miRNA stability (Burroughs et al., 2010), the effect of urifylation on miRNA turnover is unclear. Pre-miRNA transcripts have been shown to be modified by the addition of a poly(U) tail which results in the destruction of the miRNA precursor (Heo et al., 2009). Furthermore, Uridylation of processed miRNAs is also associated with their decay following highly complementary target binding (Ameres et al., 2010a; Baccarini et al., 2011a). A further report has also described miRNA uridylation leading to miRNA deactivation, instead of its degradation (Jones et al., 2009). Therefore, 3’ uridylation may result in instability or inactivity depending on the miRNA species and their cellular contexts. On the other band, the specific cellular conditions and molecular mechanisms that trigger the degradation of miRNAs bound to their targets

through imperfect sequence complementarity are as yet largely unknown (Liu, 2008).

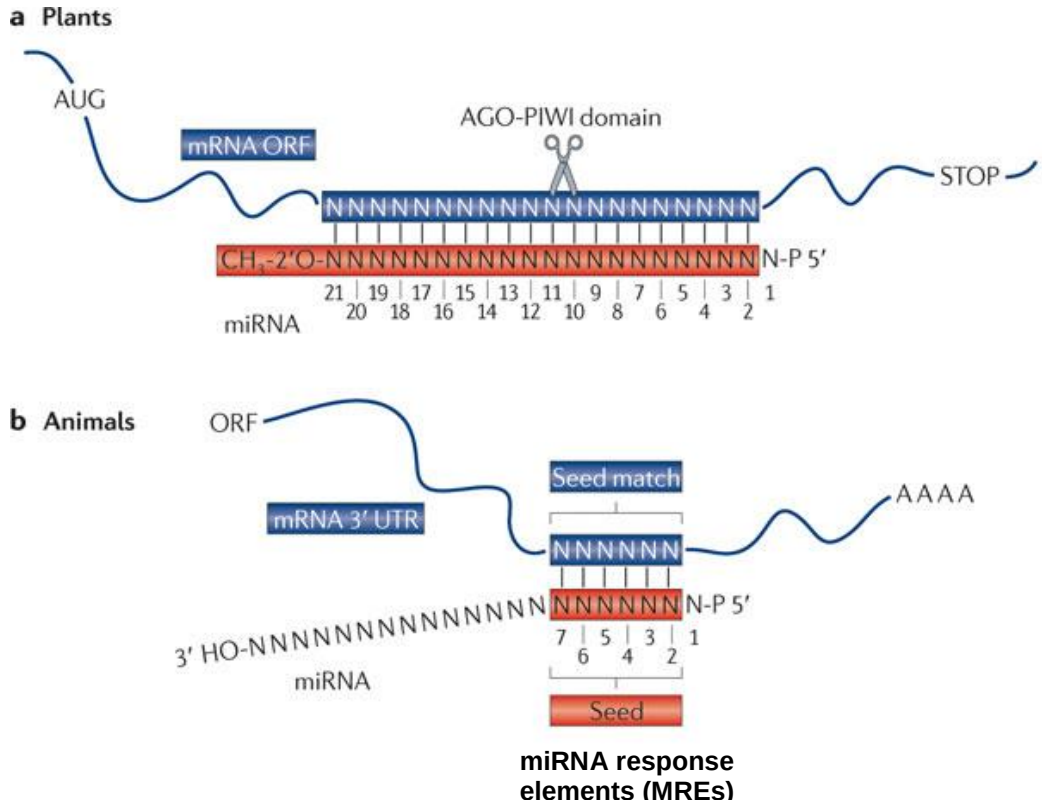


Figure 1.4 MicroRNA target recognition in (A) plants and (B) animals by Watson–Crick base pairing. (A) Plant miRNAs recognize complete or nearly complementary response elements within the target sequence, which are generally located within ORFs. (B) Animal miRNAs recognize partially complementary response elements within the target sequence, which are frequently found within 3′ UTRs of protein-coding genes. Sequence complementarity found within the 'seed' sequence, nts 2–8 at the 5’ end of the miRNA response element (MRE), may be sufficient to trigger silencing. Figure is taken from Huntzinger and Izaurralde, 2011 and includes details and components of the pathway not described in the main text.

Target recognition and mode of regulation by miRNAs is often different between organisms (Figure 1.4). In plants, most miRNA are highly complementary to their targets, with interactions often extending beyond the seed sequence (Jones-Rhoades et al., 2006). Target recognition via highly complementary miRNAs triggers the RNA interference (RNAi)-like pathway (Pillai et al., 2007), leads to transcript cleavage by the AGO protein (Llave et al., 2002; Rhoades et al., 2002), induces decapping or deadenylation of the target, and results in its degradation (Decker and Parker, 1993; Meister et al., 2004; Parker and Song, 2004; Behm-Ansmant et al., 2006a; Giraldez et al., 2006; Valencia-Sanchez et al., 2006; Wu et al., 2006; O'Carroll et al., 2007; Wakiyama et al., 2007).

In contrast to plant miRNAs, extensive sequence complementarity between miRNAs and their targets is rare in animals (Ambros, 2004). Instead, complementarity only between the miRNA seed sequence (nts 2-8 of the 5' of miRNA) and the target appears to be sufficient for target-recognition (Lewis et al., 2005; Bartel, 2009). This type of imperfect sequence complementarity results in bulges between the miRNA:target duplex, which are expected to preclude the AGO-mediated endonucleolytic cleavage of the transcript (Jones-Rhoades et al., 2006) and induce either exonucleolytic target degradation via mRNA deadenylation (Behm-Ansmant et al., 2006b; Giraldez et al., 2006; Wu et al., 2006; Eulalio et al., 2009) and/or decapping (Eulalio et al., 2007; Djuranovic et al., 2011) or translational inhibition of protein synthesis, which dramatically reduces the protein levels with little impact exerted on the respective mRNA targets (Djuranovic et al., 2011; Huntzinger and Izaurralde, 2011). Suppression of translational activities by miRNAs is illustrated by the

regulation of *LIN14* and *LIN28* levels, which bind the first identified miRNA, *lin-4*. *Lin-4* controls the temporal expression patterns of LIN14 and LIN28 proteins during *Caenorhabditis elegans* development (Lee et al., 1993; Wightman et al., 1993).

The predominant mechanism by which miRNAs regulate their target transcripts, through translational inhibition or mRNA degradation, has always been an area of active debate (Fabian et al., 2010; Djuranovic et al., 2011). Although target silencing at the translational level was initially suggested as the primary mechanism, where the degradation of mRNA targets occurs only as a by-product in animals (Nilsen, 2007; Bazzini et al., 2012; Djuranovic et al., 2012), studies have also reported mRNA decay to be the major consequence of miRNA regulation (Baek et al., 2008; Selbach et al., 2008; Guo et al., 2010). Collectively, both translational inhibition and transcriptional degradation are common mechanisms of miRNA-induced post-transcriptional repression in animals (Jones-Rhoades et al., 2006; Djuranovic et al., 2011) and the complete picture of the complex modes of miRNA actions in animals remains unclear (Figure 1.5).

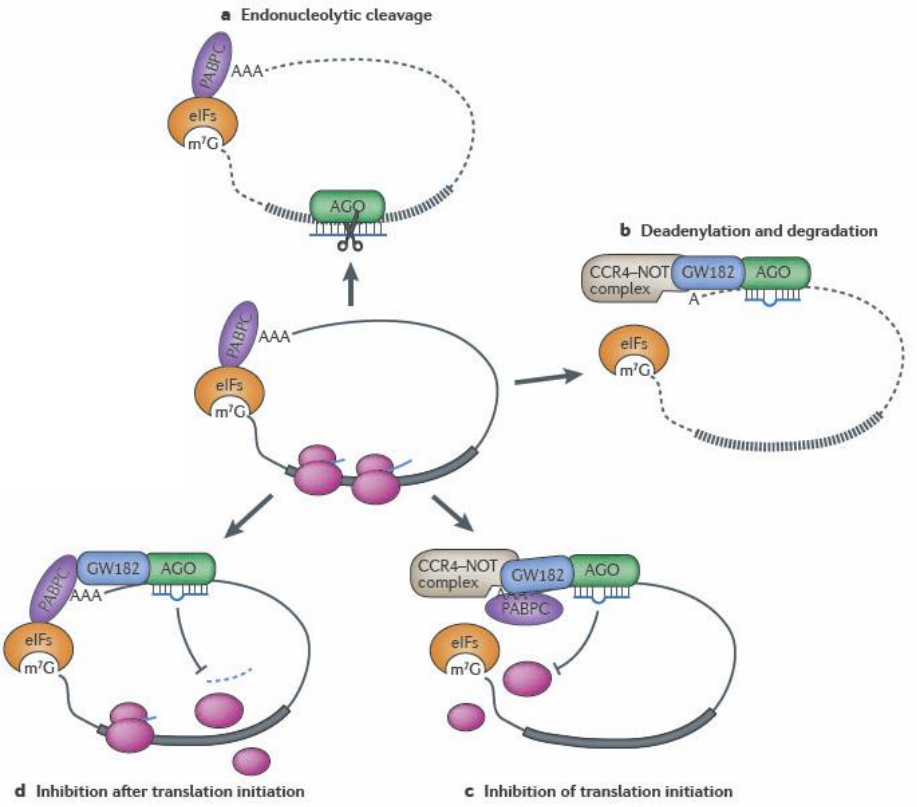


Figure 1.5 Possible pathways through which miRNAs post-transcriptionally regulate gene expression. (A) Perfect pairing between a miRNA and its target site induces endonucleolytic cleavage by AGO protein, leading to mRNA degradation. (B) Partial pairing of the miRNA complex to target MREs can result in mRNA deadenylation, followed by its degradation. (C) The miRISC can silence translation by blocking its initiation. (D) Translational inhibition induced by the miRISC can also occur in steps after translational initiation, such as by promoting ribosome drop-off or by stimulating proteolysis of the nascent peptide. Figure is adapted from Pasquinelli, 2012 and includes details and components of the pathway not described in the main text.

Transcripts can have multiple MREs for the same miRNA (Rajewsky and Socci, 2004) and transcriptome-wide analysis revealed that the impact of a miRNA on its target's product abundance is often directly proportional to its number of harboured MREs (Doench and Sharp, 2004), suggesting that their repeated occurrences frequently have a multiplicative effect. For example, adjacent MREs exert a significantly greater impact on gene expression than what would be expected from independently acting MREs, indicating a cooperative effect

from nearby sites (Doench and Sharp, 2004). In addition, the efficacy of miRNA regulation on gene expression is also affected by the relative location of the MREs within its target sequence (Grimson et al., 2007). With a few exceptions, miRNA targeting is more frequent and efficient within the 3' untranslated regions (UTRs) of protein-coding genes than within either the 5' UTRs or their ORFs (Gu et al., 2009). This is in part due to the likely displacement of the silencing miRISC complex by the translationary machinery as it progresses from the cap-ended side of the transcript (Gu et al., 2009). The lack of translation of long noncoding transcripts has led to the hypothesis that their abundance is more efficiently post-transcriptionally regulated by miRNAs (Ebert and Sharp, 2010; Hansen et al., 2013).

The ability for any given miRNA to regulate the expression levels of its targets is dependent on its subcellular co-localization. The majority of miRNAs are localized in the cytoplasmic fraction of the cell (Ambros, 2004; Bartel, 2004), as are the majority of protein-coding mRNAs (Kohler and Hurt, 2007a). Post-transcriptional regulation outside of the cell's transcriptional core, the nucleus, and later in the gene expression pipeline, may allow miRNAs to offset the stochastic fluctuations that eventually occur during upstream processes, such as transcription, splicing, and nuclear export (Bartel, 2004; Ebert and Sharp, 2012). Furthermore, despite their general cytoplasmic enrichment, some RISC-associated miRNAs have been shown to be able to re-enter the nucleus (Meister et al., 2004; Robb et al., 2005; Politz et al., 2006; Hwang et al., 2007; Weinmann et al., 2009). These nuclear miRNAs have been proposed to target and modulate the biogenesis and expression levels of primary miRNA transcripts and nuclear-retained RNA transcripts (Hansen et al., 2011; Chen et

al., 2012a; Tang et al., 2012; Liang et al., 2013). In addition to nuclear miRNAs, functional Dicer has also been shown to exist in the nucleus (White et al., 2014). Although the prevalence and mechanistic roles of nuclear miRNAs are less well understood than those of their cytoplasmic counterparts, these observations nonetheless expand the extent of miRNA regulation to virtually all levels of gene expression control.

1.3.3 Most miRNAs Confer Robustness to Gene Expression

Some miRNAs can function as molecular switches, regulating the levels of genes that are critical to the transition between distinctive cellular states (switch-like interactions) (Bartel, 2004; Sotiropoulou et al., 2009). For instance, the first miRNA discovered, *lin-4*, is a switch-like post-transcriptional regulator whose developmentally-regulated expression controls the timing of *C. elegans* larval development (Lee et al., 1993). *Lin-4* targets the transcript of *LIN-14*, a gene found to be expressed only at early stages of larval development. Lin-14 protein inhibits the transcription of genes involved in cell division and differentiation (Wightman et al., 1993). The expression of *lin-4* inhibits *LIN-14* later in development and activates stage-specific genes that promote the transitional switch from larval to adult in *C. elegans* (Ambros, 1989; Lee et al., 1993). The mutually exclusive temporal expression of *lin-4* and *lin-14* resembles a larval-to-adult developmental switch that controls the timing of *C. elegans* stage-specific post-embryonic development (Lee et al., 1993). In addition, rather than modulating the expression level of one crucial target, some miRNAs can function as switches by modulating the expression of several

genes (Lim et al., 2005). For example, ectopic expression of a brain-specific miRNA, *miR-124*, in cervical cancer derived HeLa cells induces a global shift in their expression profiles resulting in a repertoire of transcripts that closely resembles those found in brain tissues (Lim et al., 2005).

However, miRNAs capable of inducing switches between different cellular states are rare (Bartel, 2009) and the impact of most miRNAs on target product abundance is typically modest (Baek et al., 2008; Bartel, 2009). This is consistent with most miRNA:target interactions being either functionally neutral (Seitz, 2009) or contributing to the fine-tuning of gene expression (Bartel, 2004; Sevignani et al., 2006). By fine-tuning transcript levels, most miRNAs likely function to reinforce robust gene expression responses post-transcriptionally (Hornstein and Shomron, 2006; Ebert and Sharp, 2012). Such functions are likely to be non-essential in normal environmental conditions and might only become apparent under developmental or stress conditions (van Rooij et al., 2007; Brenner et al., 2010; Zheng et al., 2011). This might explain, at least in part, why despite complete loss of miRNA biogenesis being embryonic lethal (Wienholds et al., 2003; Kanellopoulou et al., 2005), knockout of most miRNAs in mice produces no overt morphological or behavioural phenotypes (Miska et al., 2007; Mukherji et al., 2011).

For some miRNAs, their contributions to cellular robustness are reinforced via autoregulatory feedback loops that directly link the output of miRNA actions with their transcriptional regulators (Linsley et al., 2007; Li et al., 2009; Osella et al., 2011). For example, the factors controlling cell fate decisions in mouse embryonic stem cells (mESCs) are a well-established example of one such

circuitry. Transcription of the precursor transcript of *miR-145* is regulated by one of the core transcription factors controlling cellular pluripotency, Oct4 (Xu et al., 2009). In turn, mature *miR-145* post-transcriptionally represses the levels of all three core transcription factors that govern ESC pluripotency, Nanog, Oct4 and Sox2 (Xu et al., 2009). The presence of such feedback circuitries is likely to allow rapid and accurate responses to homeostatic disturbance (Li et al., 2009; Osella et al., 2011).

1.3.4 Identification of miRNA:Target Interactions

Several experimental and computational approaches aiming to predict and validate functionally relevant miRNA-target interactions have been developed in recent years.

Computational approaches, designed to identify putative MREs within transcripts, have been extensively used to predict miRNA:target interactions (Stark et al., 2003; Chaudhuri and Chatterjee, 2007; Lindow and Gorodkin, 2007; Watanabe et al., 2007; Zhang and Verbeek, 2010; Witkos et al., 2011). These tools search for the presence of one or more features known to influence target recognition by miRNAs, namely: i) evidence of sequence complementarity, notably within the miRNA seed sequence; ii) evidence of other sequence features associated with this region of complementarity found in (i), such as the distance to the 3' end of the transcript or target secondary structure; iii) evolutionary conservation of the region of sequence

complementary; and iv) the thermodynamic stability of the putative miRNA-target duplex.

Large-scale prediction of miRNA targets was first published for *Drosophila melanogaster* (Stark et al., 2003). Subsequently, other algorithms, including miRanda (Enright et al., 2004; John et al., 2004) and TargetScan (Lewis et al., 2005; Grimson et al., 2007; Friedman et al., 2009; Garcia et al., 2011) were developed to computationally predict MREs. MiRanda (Enright et al., 2004; John et al., 2004) uses weighted target prediction scores to predict the degree of sequence complementarity, thermodynamics, and conservation. TargetScan (Lewis et al., 2005) relies on perfect (the 8mer site) or nearly-perfect (the 6mer site and the 7mer sites) matches between miRNA and seed regions, along with thermodynamics and evolutionary conservation of the predicted MREs (Grimson et al., 2007; Friedman et al., 2009; Garcia et al., 2011). Other popular softwares include: PicTar (Krek et al., 2005) use combinatorial methods to predict the binding of a single or a group of co-expressed miRNAs, and thus, accounts for potential cooperative effects from the presence and activities of a combination of miRNAs; TargetBoost (Saetrom et al., 2005) that predicts miRNA-target interaction rules using a machine learning approach built using a large set of validated miRNA targets (Boutla et al., 2003; Brennecke et al., 2003; Rajewsky and Socci, 2004) and a larger set of random sequences as negative training data (Rajewsky and Socci, 2004); and PITA (Miranda et al., 2006; Hammell et al., 2008) that takes into account target site accessibility by factoring in mRNA secondary structure.

Despite the large effort and the conceptually different approaches used to develop computational methods to predict miRNA targets, the overlap between predicted and experimental validated miRNA targets is low (Maziere and Enright, 2007). This demonstrates our yet incomplete understanding of the mechanisms and properties that underlie miRNA-mediated regulation. For example, on average, 3 in 10 interactions predicted by the algorithms do not occur *in vivo* (Lewis et al., 2003; Enright et al., 2004; Kiriakidou et al., 2004; Krek et al., 2005; Lewis et al., 2005). Several studies have compared different algorithms (Maziere and Enright, 2007; Barbato et al., 2009) and estimated their relative accuracy and sensitivity (Sethupathy et al., 2006; Martin et al., 2007) using experimentally validated miRNA targets. An emphasis on matches at seed regions limits the conventional miRNA recognition and binding rules. The implementation of additional features, such as secondary structure predictions and evolutionary conservation, has been shown to enhance the accuracy of some of these algorithms, including miRanda, TargetScan, PicTar and PITA (Sethupathy et al., 2006). In most cases, additional cellular context-specific requirements, such as the sequence of tissue specific isoforms, may vastly improve the accuracy and sensitivity of miRNA:target interaction predictions (Doench and Sharp, 2004; Farh et al., 2005; Grimson et al., 2007). Because they do not rely completely on our yet incomplete knowledge of the rules underlying miRNA:target recognition, machine learning approaches, similar to TargetBoost (Saetrom et al., 2005), may provide an advantage. Surprisingly, simpler approaches seem to have comparable precision to more complex methodologies (Selbach et al., 2008; Alexiou et al., 2009; Min and Yoon, 2010). In contrast, their computing times vary substantially with

algorithms that incorporate more parameters or metrics being more computationally expensive.

In parallel with the growth of computational miRNA target prediction tools, experimental techniques for the identification of miRNA-target interactions have also been developed (Rajewsky and Socci, 2004; Sethupathy et al., 2006; Rigoutsos, 2009). Transcriptome-wide analysis has been used to investigate the impact of manipulating the levels of individual miRNAs on global gene expression (Lim et al., 2005; Cole et al., 2008; Liu et al., 2008a; McLaughlin et al., 2008; Silber et al., 2008; Bader et al., 2010; Mallanna and Rizzino, 2010; Xiao et al., 2012). Reciprocally, arrays have also been used to pinpoint differentially expressed miRNAs resulting from changes in specific mRNA abundances or cellular context and disease states (Babak et al., 2004; Barad et al., 2004; Calin et al., 2004; Miska et al., 2004; Nelson et al., 2004; Thomson et al., 2004; Baskerville and Bartel, 2005; Liang et al., 2005; Liu et al., 2008a; Elkan-Miller et al., 2011). Subsequently, enriched miRNA response elements can be predicted using motif discovery programs, such as Sylamer (van Dongen et al., 2008) through its SylArray web interface (Bartonicek and Enright, 2010). More recently, proteomic approaches, such as stable isotope labeling with amino acids in cell culture (SILAC) followed by mass spectrometry have been employed to evaluate global changes in protein levels following changes in miRNA levels (Baek et al., 2008; Selbach et al., 2008; Yang et al., 2010; Ebner and Selbach, 2011; Kaller et al., 2011; Lossner et al., 2011; Yan et al., 2011; Bargaje et al., 2012; Bauer and Hummon, 2012; Huang et al., 2012). Despite providing important insights into the global impact on gene product abundances following miRNA perturbation, the fundamental limitation of such

analyses is that they do not allow the distinction between direct and indirect effects of miRNA regulation. As a result, these experimental techniques are not useful in the identification of miRNA targets (Johnson et al., 2007; Linsley et al., 2007; Baek et al., 2008).

MiRNA-target interaction can be identified by detecting transcripts that are physically associated with miRNA-incorporated RISC (Ule et al., 2003). RNA-binding protein Immunoprecipitation (RIP) was among one of the first techniques used to map RNA–protein interactions (Niranjanakumari et al., 2002). This technique requires the cross-linking of AGO-associated RNA using formaldehyde. Immunoprecipitation of the cross-linked sample using an AGO-specific antibody followed by extensive washes yields AGO-bound RNA (Keene et al., 2006). Following the digestion of unbound and exposed RNA regions, purified RNA from the AGO-protected regions can be detected by various methods, including microarray (RIP-Chip) (Keene et al., 2006; Tan et al., 2009; Dolken et al., 2010; Wang et al., 2010b; Wang et al., 2010c) and high-throughput sequencing (RIP-seq) (Kanematsu et al., 2013; Nie et al., 2013).

Similar to RIP, cross-linking immunoprecipitation-high-throughput sequencing (CLIP) also identifies miRNA binding sites by the immunoprecipitation of AGO-associated RISCs (Ule et al., 2003; Jensen and Darnell, 2008; Wang et al., 2009b). In contrast to RIP, in CLIP protocols, AGO-RISC associated RNA is cross-linked using UV radiation instead of formaldehyde (Ule et al., 2003; Jensen and Darnell, 2008). UV radiation creates stronger covalent bonds that are irreversible and thus, CLIP allows for more stringent purification conditions yielding more specific RNA-AGO interactions (Ule et al., 2003; Chi et al., 2009;

Hafner et al., 2010; Zisoulis et al., 2010). Like RIP, AGO-bound RNA from CLIP protocols can be sequenced using high-throughput methods (AGO HITS-CLIP) (Licatalosi et al., 2008; Darnell, 2010). Photoactivatable-ribonucleoside-enhanced CLIP (PAR-CLIP) is another CLIP technique that relies on the incorporation of photoreactive ribonucleoside analogs that allows the efficient cross-linking of photoreactive nucleoside-labeled cellular RNAs to interacting AGO-RISC under UV radiation (Hafner et al., 2010). The isolated MRE-containing RNA can then be deep-sequenced to reveal transcriptome-wide binding sites of miRNAs with relatively high resolution (Hafner et al., 2010; Ascano et al., 2012; Hafner et al., 2012; Chou et al., 2013; Erhard et al., 2013).

Despite the identification of regions within transcripts that are associated with miRISCs, these approaches fail to directly infer the identity of the specific miRNAs that were loaded into the target associated miRISC (Liu et al., 2013). This limitation can be partially circumvented by exogenously increasing the levels of particular miRNAs, using miRNA mimics in an attempt to identify targets of the specific miRNAs of interest (Thomson et al., 2011). However, this technique is also limited as it allows the testing of only one miRNA per cellular environment at a time. Moreover, critically, as transfection of miRNA mimics can result in changes in transcriptional programs, this approach might affect the repertoires of transcripts in the cell. Recently, a new method based on the principles of AGO-CLIP, cross-linking, ligation, and sequencing of hybrids (CLASH), was developed (Kudla et al., 2011; Helwak et al., 2013); CLASH aims to directly detect functional miRNA-target interactions by introducing an additional ligation step that results in the formation of miRNA:target chimeras (Kudla et al., 2011; Helwak et al., 2013). Recently, endogenous miRNA:target

chimeras generated in the absence of exogenous ligase were discovered and validated, demonstrating that systematic analysis of miRNA:target chimeras enables direct context-specific discovery of functional miRNA binding (Grosswendt et al., 2014).

The development of these and other high-throughput approaches to map miRNA:target associations allowed an exponential increase in the number of known *bona fide* interactions in different cells. Interestingly, all techniques have provided evidence that non-canonical (un-seeded) miRNA:target interactions are relatively common in animals (Grimson et al., 2007; Bartel, 2009). For example, 18% of the interactions mapped using CLASH suggest miRNA:target pairing that involved bulged or mismatched nucleotides (Helwak et al., 2013). Furthermore, apart from the 3' UTRs, many putative target:miRNA-seed matches were identified within the coding sequences of genes (Hendrickson et al., 2008a). Collectively, these observations demonstrate the flexibility of miRNA target recognition and highlight our relatively poor understanding of the rules that underlie miRNA:target interactions.

A few databases that compile experimentally validated interactions between miRNAs and their targets are now available. Examples include starBase (sRNA target base), a database of miRNA-mRNA interactions determined from a comprehensive set of AGO-CLIP-Seq (also known as HITS-CLIP-Seq), and Degradome-Seq data that explored the AGO-binding and cleavage sites in six organisms (Yang et al., 2011). Furthermore, motivated by the recent findings that many noncoding transcripts function as competitive endogenous RNAs (ceRNAs), including transcribed pseudogenes (Poliseno et al., 2010; Marques

et al., 2012), lincRNAs (Cesana et al., 2011), and circular RNAs (circRNAs) (Hansen et al., 2013; Memczak et al., 2013), several databases were also created with the aim at unravelling functional interactions between these transcripts and miRNAs. For instance, the DIANA-LncBase (Paraskevopoulou et al., 2013) integrates putative lncRNA MREs determined using HITS-CLIP and PAR-CLIP experimental data with MREs predicted using miRcode (Jeggari et al., 2012) on the GENCODE annotated lincRNAs.

1.4 INTERGENIC LONG NONCODING RNAS

Currently, there are thousands of lincRNAs annotated in the mouse (Ravasi et al., 2006; Ponjavic et al., 2007; Guttman et al., 2009; Guttman et al., 2010; Sigova et al., 2013) and human (Khalil et al., 2009; Jia et al., 2010; Orom et al., 2010; Cabili et al., 2011; Derrien et al., 2012; Sigova et al., 2013) genomes (Table 1.1). So far, catalogues have been found to be largely non-overlapping; for example GENCODE v7 lincRNAs (Derrien et al., 2012) overlapped 30-39% with previously published sets of human lincRNAs (Jia et al., 2010; Cabili et al., 2011).

Although it was initially controversial whether the vast amount of noncoding transcripts, including intergenic long noncoding RNAs (lincRNAs), generated from the pervasive transcription across noncoding regions of the genome represent biologically functional elements, or whether these noncoding RNA transcripts were the mere products of transcriptional noise (Mattick, 2004b; Wang et al., 2004; Struhl, 2007; Ebisuya et al., 2008), their frequent association with global signatures of functionality for a subset of these lincRNAs motivated continuous efforts from scientists in different research fields, including molecular biology, biochemistry, genetics, and computational genomics, to characterize and establish their molecular, cellular, and organismal functions.

Table 1.1 Catalogs of lincRNA Loci and Transcripts. Ulitsky and Bartel 2013 (Ulitsky and Bartel, 2013).

Reference		Data for Transcript Reconstruction	Genomic Features and Filters	Coding-Potential Filters	Number of lincRNAs
Mouse	Ravasi et al., 2006	cDNAs		Manual curation, ORF length, CRITICA	13,502 transcripts
	Ponjavic et al., 2007	cDNAs, CAGE		Manual curation, ORF length, BLAST, CRITICA	3,122 transcripts
	Guttman et al., 2009	Chromatin marks, tiling arrays	Collection of approximate exonic regions, chromatin domain ≥5 kb	CSF	1,675 loci (1,250 conservatively defined)
	Guttman et al., 2010	RNA-seq	Multi-exon only	CSF	1,140 lincRNA transcripts
	Sigova et al., 2013	RNA-seq, cDNAs, chromatin marks,	Antisense overlap with mRNA introns allowed, ≥100 nt mature length	CPC	1,664 loci
Human	Khalil et al., 2009	Chromatin marks, tiling arrays	Collection of approximate exonic regions, chromatin domain ≥5 kb	CSF	3,289 loci
	Jia et al., 2010	cDNAs	Overlap with mRNAs allowed		5,446 transcripts
	Orom et al., 2010	cDNAs	Restricted to loci >1 kb away from known protein-coding genes, ≥200 nt mature length	Manual curation based on length, conservation and other characteristics of the ORFs	3,019 transcripts from 2,286 loci
	Cabili et al., 2011	RNA-seq	Multi-exon only, ≥200 nt mature length	PhyloCSF, Pfam	8,195 transcripts (4,662 in the stringent set)
	Derrien et al., 2012	cDNAs	Overlap with mRNAs allowed (intergenic transcripts reported separately), ≥200 nt mature length	Manual curation based on length, conservation and other characteristics of the ORFs	14,880 transcripts from 9,277 loci, including 9,518 intergenic transcripts
	Sigova et al., 2013	RNA-seq, cDNAs, chromatin marks,	Antisense overlap with mRNA introns allowed, ≥100 nt mature length	CPC	3,548 loci from embryonic stem cells, and 3,986 loci from endodermal cells

Reference		Data for Transcript Reconstruction	Genomic Features and Filters	Coding-Potential Filters	Number of lincRNAs
Frog	Tan et al., 2013	RNA-Seq	>25 kb away from known protein-coding genes or on a different strand from the neighboring genes, ≥200 nt mature length	ORF length, BLAST, Pfam	6,686 transcripts from 3,859 loci
Zebrafish	Ulitsky et al., 2011	RNA-seq, cDNAs, 3P-seq, chromatin marks	Antisense overlap with mRNA introns allowed, ≥200 nt mature length	CPC	691 transcripts from 567 loci
	Pauli et al., 2012	RNA-seq	Stringent criteria for single exon, intron overlap with mRNA allowed, ≥160 nt mature length	ORF length, PhyloCSF, BLAST, Pfam	397 intergenic and 184 intronic overlapping transcripts
Fly	Tupy et al., 2005	cDNA		Manual curation based on ORF length, conservation and other characteristics, Ka/Ks test, QRNA	17 transcripts
	Young et al., 2012	RNA-seq	≥200 nt locus length		1,119 trancripts
Nematode	Nam and Bartel, 2012	RNA-seq, 3P-seq	≥100 nt mature length	CPC, RNAcode, ribosome profiling, polysome association	262 lincRNA transcripts from 170 loci
Arabidopsis	Liu et al., 2012a	cDNA, tiling arrays, RNA-seq	In part a collection of approximate exonic regions, >500 bp away from protein-coding genes, no overlap with transposable elements allowed, ≥200 nt mature length	ORF length	6,480 transcription units from tiling arrays, 278 transcripts from RNA-seq
Maize	Boerner and McGinnis, 2012	cDNA	Both sense overlap with introns and antisense overlap with mRNA or introns allowed, ≥200 nt mature length	ORF length	2,492 transcripts

1.4.1 Characterization of lincRNAs

Although lincRNAs represent a mixture of noncoding transcripts, categorized solely on their common lack of apparent open reading frame and length exceeding 200 nt, several general features of lincRNAs have been described (Cabili et al., 2011; Ulitsky et al., 2011; Derrien et al., 2012; Pauli et al., 2012). Similar to protein-coding messenger RNAs (mRNAs), a subset of lincRNAs are transcribed by RNA-polymerase II (Pol II), capped and polyadenylated (Carninci et al., 2005; Kapranov et al., 2007) and frequently spliced at canonical splicing sites (Chew et al., 2013), although at a seemingly reduced splicing efficiency compared to mRNAs (Tilgner et al., 2012). In addition, lincRNAs are enriched in the same chromatin marks as protein-coding genes: histone H3K4 trimethylation at their 5'-end and histone H3K36 trimethylation across the gene body (Mikkelsen et al., 2007; Ponjavic et al., 2007; Guttman et al., 2009; Khalil et al., 2009; Cabili et al., 2011; Derrien et al., 2012; Pauli et al., 2012).

Unlike protein-coding genes, lincRNAs are typically shorter in length, consisting of typically only 2-3 exons (Cabili et al., 2011; Derrien et al., 2012; Pauli et al., 2012). They are preferentially found within 10 kb of protein-coding genes (Bertone et al., 2004; Ponjavic et al., 2007; Jia et al., 2010; van Bakel et al., 2010; Cabili et al., 2011; Sigova et al., 2013) and share a similar distance distribution between lincRNAs and their closest protein-coding gene neighbor as that between adjacent protein-coding genes (Ulitsky et al., 2011; Ulitsky and Bartel, 2013).

In comparison to protein-coding mRNAs, lincRNAs are expressed at lower levels; for example, human lincRNAs were reported to be ten times lower in expression than that of mRNAs (Ravasi et al., 2006; Guttman et al., 2009; Guttman et al., 2010; Cabili et al., 2011; Ulitsky et al., 2011; Derrien et al., 2012; Pauli et al., 2012; Sigova et al., 2013; Ulitsky and Bartel, 2013)). Surprisingly, recent studies showed that lincRNAs are not generally unstable but rather show a wide variation in transcript stability similar to that observed for protein-coding mRNAs (Clark et al., 2012; Tani et al., 2012), suggesting that not all lincRNAs are preferentially targeted for degradation. Furthermore, this observation is consistent with the functional diversity of lincRNAs and their complex post-transcriptional regulation may explain their often temporal- and spatial-specific expression profiles (Cabili et al., 2011; Derrien et al., 2012).

1.4.2 Rapid turnover of lincRNA sequences and expression

In contrast to protein-coding mRNAs whose sequences are highly conserved, lincRNAs typically lack identifiable orthologs and are poorly conserved across sometimes closely related species (Ulitsky and Bartel, 2013). For example, only ~12% of lincRNA transcripts identified in human and mouse display significant sequence and expression conservation in both species (Church et al., 2009; Cabili et al., 2011) and <6% of zebrafish lincRNA transcripts have detectable sequence conservation with their potential mammalian lincRNA orthologs (Ulitsky et al., 2011).

However, despite their rapid sequence evolution relative to protein-coding mRNAs, weak signatures of purifying selection have been observed within the promoters and exons of lincRNA transcripts (Ponjavic et al., 2007). Transcribed lincRNAs are often more constrained in sequence than neutrally evolving neighbouring regions. These signatures of sequence constraint are nevertheless weaker than either coding or noncoding regions of mRNA exons (Ponjavic et al., 2007; Guttman et al., 2009; Khalil et al., 2009; Marques and Ponting, 2009; Ulitsky et al., 2011; Derrien et al., 2012).

While weak sequence conservation of lincRNA exons demonstrates they evolve under modest purifying selection, only a small subset appear to be conserved in expression across mammalian evolution (Church et al., 2009; Cabili et al., 2011; Kutter et al., 2012; Necsulea et al., 2014; Washietl et al., 2014). Specifically, >90% of protein-coding mRNAs, but only approximately 60% of lincRNAs that are expressed in the liver of mouse, appear to have orthologs that are also expressed in the livers of the rat, illustrating the rapid evolutionary turnover of lincRNAs (Kutter et al., 2012). In addition, 80% of human lincRNAs are primate-specific, while only 3% are found to be conserved across more than 300 million years (MY, i.e. from primates to tetrapods) (Necsulea et al., 2014). These observations suggest that a large portion of lincRNAs may have no biological relevance.

In contrast, examples exist where small stretches of highly conserved regions have been identified within lincRNAs, possibly representing conserved functional elements evolving under purifying selection, while the remaining parts of the noncoding transcript sequences show less conservation (Pang et

al., 2006). For instance, although no overall sequence constraint was found for the ~4.5 kb zebrafish lincRNA, *Cyrano*, this noncoding transcript is conserved in transcription from its syntenic loci in human and mouse (Ulitsky et al., 2011), except for a short stretch of 67 nt within the lincRNA transcript, a predicted miRNA response element (MRE) for *miR-7*. Surprisingly, both human and mouse orthologs of this lincRNA are able to rescue zebrafish *Cyrano* loss-of-function *in vivo* (Ulitsky et al., 2011), demonstrating that only a small fraction of a functional lincRNA transcript may be under purifying selection, such as those that represent functional regulatory motifs, although the overall sequence conservation of the complete transcript may be weak. Conserved functional features without high sequence similarity between orthologous lincRNA pairs in different species may be the case for many lincRNAs that exhibit strong secondary structure, genomic position and orientation conservation without detectable primary sequence conservation (Ulitsky et al., 2011).

1.4.3 Potential lincRNA functions

Despite only a relatively small number of lincRNAs having been functionally characterized to date, they have already been implicated to contribute at all levels of gene regulation by modulating levels of genomically adjacent or distally located gene products via *cis*- or *trans*-acting molecular mechanisms, respectively (reviewed in Kung et al., 2013) (Figure 1.6). In particular, lincRNAs have been implicated in (1) transcriptional regulation, including regulatory element control, chromatin remodeling and epigenetic control, as well as (2) post-transcriptional regulation where lincRNAs can act as molecular decoys for

other species of RNA transcripts (review in Kung et al., 2013). I will briefly describe a few examples of reported lincRNA regulatory roles.

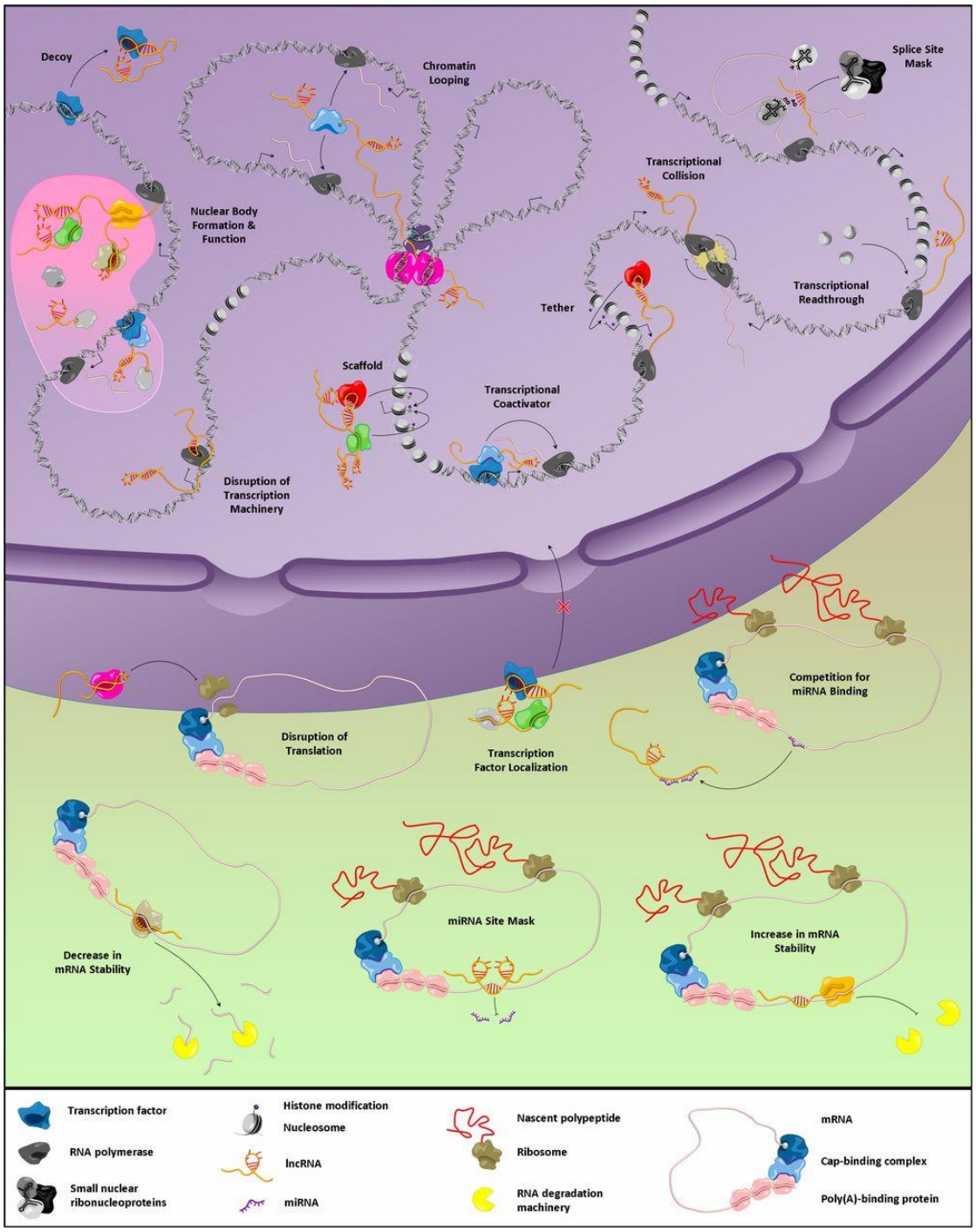


Figure 1.6 Mechanisms proposed for lincRNA. Taken from Kung et al., 2013 .

1.4.3.1 TRANSCRIPTIONAL REGULATION

LincRNAs and *Cis*-regulation of Local Gene Expression

In the nucleus, many lincRNAs appear to regulate near their site of synthesis the transcription of genomically adjacent genes in *cis* (Wang et al., 2011; Lai et al., 2013; Melo et al., 2013). However, it is vital to distinguish whether this regulation is dependent on the sequence of the noncoding transcripts (i.e. to mediate looping onto the promoter regions of their transcriptional targets), or whether it is simply the act of transcription of the lincRNAs that promote active expression of the nearby genes (i.e. by altering local chromatin status). In this thesis, I will be discussing only lincRNAs that possess functional regulatory roles transacted through transcript-dependent mechanisms.

An abundant class of *cis*-acting transcriptional regulating lincRNAs has been identified as being transcribed through gene enhancer regions, where they may function to positively promote the transcription of neighbouring target genes. Enhancers are genomic regulatory elements that activate the transcription in *cis* of target genes in frequently spatiotemporal-specific ways during development (Maston et al., 2006; Visel et al., 2009). Specifically, enhancers possess defined chromatin features, such that they often reside in DNase I sensitive genomic sites (i.e. regions of open chromatin and often containing protein-binding sites), and they are frequently associated with specific histone modifications, including H3K4me1 and H3K27ac, while being depleted for H3K4me3 (Natoli and Andrau, 2012).

A subset of enhancers can give rise to uni-directionally transcribed, polyadenylated, more stable and often spliced enhancer-associated lincRNAs (elncRNA) (De Santa et al., 2010; Orom et al., 2010; Natoli and Andrau, 2012; Marques et al., 2013), as well as bi-directionally transcribed and non-polyadenylated lincRNAs, termed enhancer lincRNAs (eRNAs) (Ntini et al., 2013). For instance, around 50% of detected enhancer-associated polyadenylated lincRNA species in mouse have been shown to be transcribed uni-directionally and exhibit enhancer-like chromatin signatures (i.e. high H3K4me1) at their transcription initiation regions; enhancer-like chromatin signatures have also been found within the gene body (Marques et al., 2013; Andersson et al., 2014). Similarly, recent systematic analysis performed by the FANTOM consortium showed that the majority of transcriptional enhancers produce bi-directionally transcribed eRNAs in various human primary cells where the expression levels of the eRNA transcripts correlate with the abundance of their neighbouring protein-coding genes (Kim et al., 2010; Andersson et al., 2014).

The growing numbers of eRNAs and elncRNAs have now been shown to function at their site of synthesis in a transcript-dependent manner to positively regulate expression of neighbouring protein-coding genes (Hah et al., 2013; Lam et al., 2013; Li et al., 2013; Melo et al., 2013; Mousavi et al., 2013), providing evidence that the transcription of such enhancer lncRNAs is a cause rather than consequence of enhancer activities (Orom and Shiekhattar, 2013).

LincRNA and *Trans*-regulation of Distal Gene Expression

In addition to lincRNAs involved in *cis*-regulatory mechanisms, lincRNAs have also been reported to regulate levels of gene products in *trans* by influencing the expression levels of distal genes. *Hotair*, a lincRNA transcribed from the *HoxC* genomic locus, was the first lincRNA shown to repress in *trans* the transcription of genes located on a distal chromosome, the *HoxD* gene cluster (Rinn et al., 2007). Subsequently, *Hotair* was shown to regulate chromatin states by interacting with chromatin modifying protein complexes, including a Polycomb-group protein, Polycomb Repressive Complex 2 (PRC2) through its 5' end (Rinn et al., 2007; Tsai et al., 2010). The reported genome-wide binding sites of *Hotair* to over 800 regions genome-wide, including within the *HoxD* cluster, are enriched within genes that become depressed upon *Hotair* depletion (Rinn et al., 2007; Chu et al., 2011). Therefore, lincRNAs may interact with chromatin and regulate multiple loci genome-wide.

Another lincRNA, *Paupar*, is transcribed from a conserved enhancer upstream of the gene encoding for the Pax6 transcription factor (Vance et al., 2014). *Paupar* was demonstrated to act both in *cis* to regulate *Pax6* expression and in *trans* by binding and interacting with chromatin at regulatory regions of target genes genome-wide to control large-scale transcriptional programs. This study illustrates that a lincRNA, *Paupar*, can possess dual functions in transcript-dependent manners by: (1) locally regulating the expression of its genomically neighbouring protein-coding gene, *Pax6*, and (2) distally regulating transcription of multiple target genes genome-wide (Vance et al., 2014).

LincRNAs in Dynamic Nuclear Organization

LincRNAs have also been shown to shape the nuclear organization of cells through interacting with various protein complexes, including chromatin regulators (reviewed in Rinn and Guttman, 2014).

One example is that individual lincRNAs can be involved in the regulation of chromosomal gene expression through dosage compensation, which equalizes the levels of genes expressed from the X chromosome between females (with two X chromosomes) and males (with only one X chromosome) (Lyon, 1961). In mammals, dosage compensation is achieved by inactivating the transcription from one of the two X chromosomes in female cells (Monk and Harper, 1979): X chromosome inactivation (XCI) is thought to involve the coating of one X chromosome by a 15 kb lincRNA, the X-inactive specific transcript (*Xist*) (Brockdorff et al., 1992). This lincRNA is transcribed from the X chromosome targeted for inactivation, where it spreads strictly in *cis* to coat the chromosome and hence prevents its transcription (Penny et al., 1996).

On a genome-wide scale, lincRNAs can contribute to the regulation of epigenetic landscape and gene expression, in *cis* or in *trans*, by binding and guiding chromatin remodeling complexes to specific genomic loci (Rinn et al., 2007; Bertani et al., 2011; Guttman et al., 2011; Wang and Chang, 2011). Some of the best-characterized and nuclear-enriched lincRNAs have demonstrated associations with chromatin factors, such as the *cis*-regulating *Xist* with YY1, a transcription factor proposed to bind and recruit *Xist* RNA to chromatin for transcription inactivation (Jeon and Lee, 2011), the *trans*-acting

Hotair suggested to function as a scaffold for the assembly of ribonucleoprotein complexes by interacting with different proteins through distinct interaction domains, PRC2 and CoREST (Tsai et al., 2010), and *Hottip*, a lincRNA transcribed from the 5' end of the *HoxA* locus, where the noncoding transcript, bound to the WDR5-MLL transcriptional activating complex, is brought in close proximity to the *HoxA* genes by chromosomal looping to activate transcription of the *HoxA* locus (Wang et al., 2011).

However, the precise mechanism by which lincRNAs recognize and form specific interactions with various chromatin regulator complexes remains largely unclear (Huarte et al., 2010; Murthy and Rangarajan, 2010). In addition, the specificity and relevance of such interactions has been questioned and remain debatable (Brockdorff, 2013).

1.4.3.2 POST-TRANSCRIPTIONAL REGULATION

Although the initially discovered and well-characterized lincRNAs, such as *Xist*, *Malat1*, and *Hotair* (Rinn and Chang, 2012), are almost exclusively found in the nucleus, genome-wide studies have also reported lincRNAs that reside predominantly in the cytoplasm (Ulitsky and Bartel, 2013; van Heesch et al., 2014) where they may interact with factors to assist protein localization and post-transcriptional regulation of mRNA translation and stability (Coccia et al., 1992; Kino et al., 2010; Ulitsky et al., 2011; Yoon et al., 2012; van Heesch et al., 2014)) (Figure 1.7).

Cytoplasmic lincRNAs are also predicted to bind proteins and influence their activities (Ulitsky and Bartel, 2013). For example, lincRNA binding has been shown to be involved in nuclear trafficking of transcription regulators, Tsl (Wang et al., 2008) and Nfat (Willingham et al., 2005). Additionally, cytoplasm lincRNAs have also been reported to act as protein decoys that titrate away protein regulators from their targets when the molecular stoichiometry relationship between the lincRNAs and protein regulators is appropriate (Kino et al., 2010; Liu et al., 2012b).

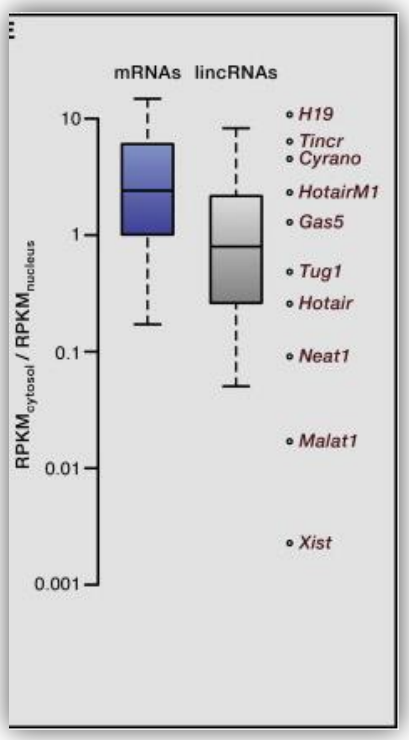


Figure 1.7 Relative subcellular localization of mRNAs and lincRNAs. Boxplots depicting the ratio of expression levels of genes between the cytosol and nucleus compartments (y-axis, log 10 scale) in HEK cells. Examples of known lincRNAs are listed on the right of the graph according to their abundance in the cellular compartments. Figure taken from Ulitsky and Bartel 2013.

Of particular interest for my work are lincRNAs that post-transcriptionally regulate transcript abundance by competing for the binding of shared miRNAs with protein-coding transcripts (Salmena et al., 2011). Termed competitive endogenous RNAs (ceRNAs), these coding or noncoding sequences share MREs with one or more other transcript(s). By competing for a limited pool of miRNAs, lincRNAs that act as ceRNAs can effectively titrate away miRNAs from their target transcripts and thus derepress them from miRNA-mediated

target translation inhibition and/or transcript degradation (Marques et al., 2011; Salmena et al., 2011) (Figure 1.8). These ceRNA-acting lincRNAs are discussed in further detail in the following chapters.

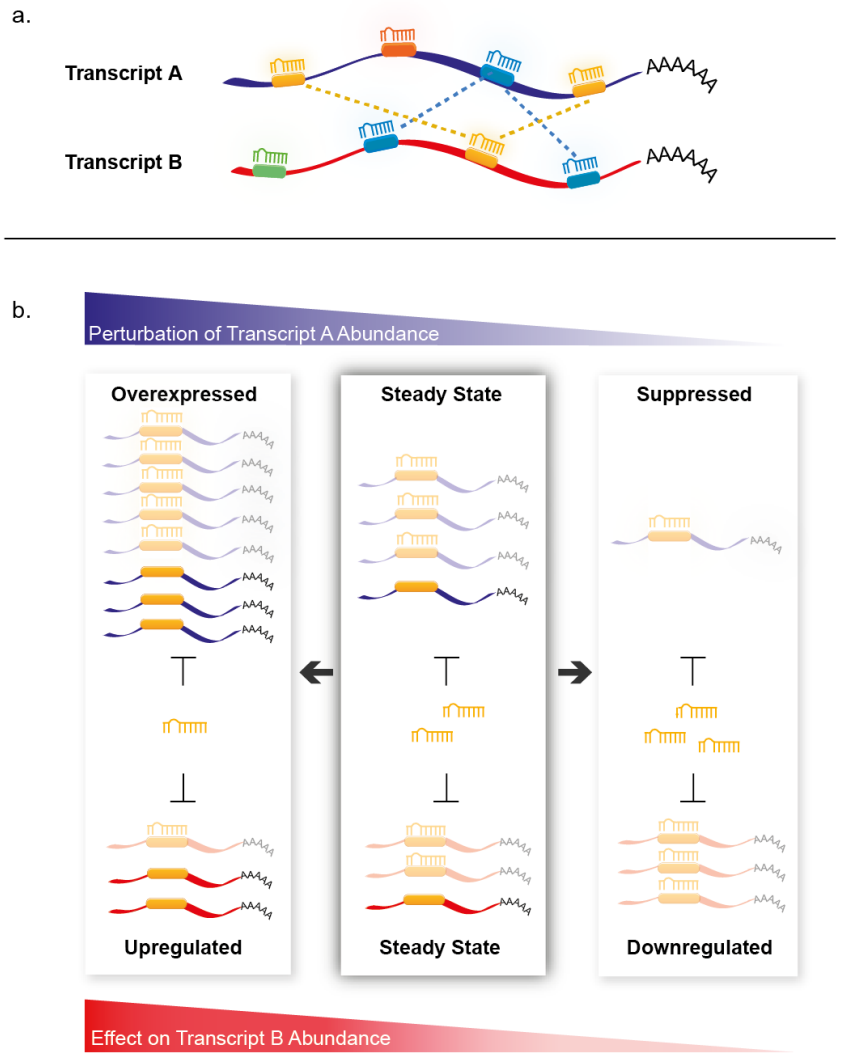


Figure 1.8 Competitive endogenous RNAs (ceRNAs). (A) Transcripts A (blue) and B (red) are a pair of ceRNAs sharing MREs (boxes) for two miRNAs (blue and yellow). The two ceRNAs can influence the expression level of each other through competitive miRNA binding (dotted lines). MiRNAs are represented as line structures bound to the MREs. (B) These transcripts co-regulate each other's expression level through competition for shared miRNA (yellow) binding on MREs. In the steady state (middle), ceRNAs and targeting miRNAs are in equilibrium. Overexpression of transcript A (left) reduces the concentration of free miRNAs, thereby increasing expression of transcript B. Decreased expression of transcript A (right) leads to an increase of available miRNAs to bind transcript B and consequently suppresses its expression level. Figure from Marques et al., 2011 .

1.5 MICRORNA-MEDIATED CROSSTALK BETWEEN RNA TRANSCRIPTS

According to current estimates, each human miRNA is able to recognize, and likely regulates, the abundance of hundreds of transcript targets (Lewis et al., 2005; Miranda et al., 2006). Thus, given that miRNA abundance is likely to be limited, transcripts regulated by the same miRNA have been proposed to compete for the binding of available miRNA and thus, be able to modulate each other's expression levels (Seitz, 2009).

Artificial miRNA sponges that contain several binding sites for specific miRNAs and are driven by strong promoter elements have been extensively used to manipulate endogenous miRNA levels (Ebert et al., 2007; Liu et al., 2008b; Brown and Naldini, 2009; Ebert and Sharp, 2010) in a variety of different animals (Asakawa and Kawakami, 2008; Kumar et al., 2008; Otaegi et al., 2011; Zhu et al., 2011) and cellular systems (Kumar et al., 2008; Bolisetty et al., 2009; Penna et al., 2011). These useful systems demonstrated that miRNA levels are also regulated by their targets and provided researchers with a tool to investigate miRNA-mediated functions and mechanisms.

The first naturally occurring evidence for such crosstalk interactions between endogenously expressed transcripts was described in plants (*Arabidopsis thaliana*) where *Induced by phosphate starvation 1* (*IPS1*) was found to compete for the binding of *miR-399* with an inorganic phosphate (Pi)-responsive genes, *PHO2* (Franco-Zorrilla et al., 2007). *IPS1* is a noncoding RNA whose expression in *Arabidopsis thaliana* is induced upon phosphorous

starvation (Franco-Zorrilla et al., 2007). This noncoding transcript contains a conserved MRE for *miR-399*, a miRNA which is also induced upon phosphorous starvation (Franco-Zorrilla et al., 2007). The nonreversible binding of *miR-399* to *IPS1* reduces the cellular availability of this miRNA, which in turn relieves *PHO2,* a protein-coding gene involved in phosphorous response, from *miR-399* mediated post-transcriptional repression (Franco-Zorrilla et al., 2007) (Figure 1.9). The competition for *miR-399* between *IPS1* and *PHO2* results in their co-expression *in vivo* (Figure 1.9). The *miR-399*-mediated crosstalk between *IPS1* and *PHO2* illustrates what is likely to be a more widespread mechanism of post-transcriptional regulation.

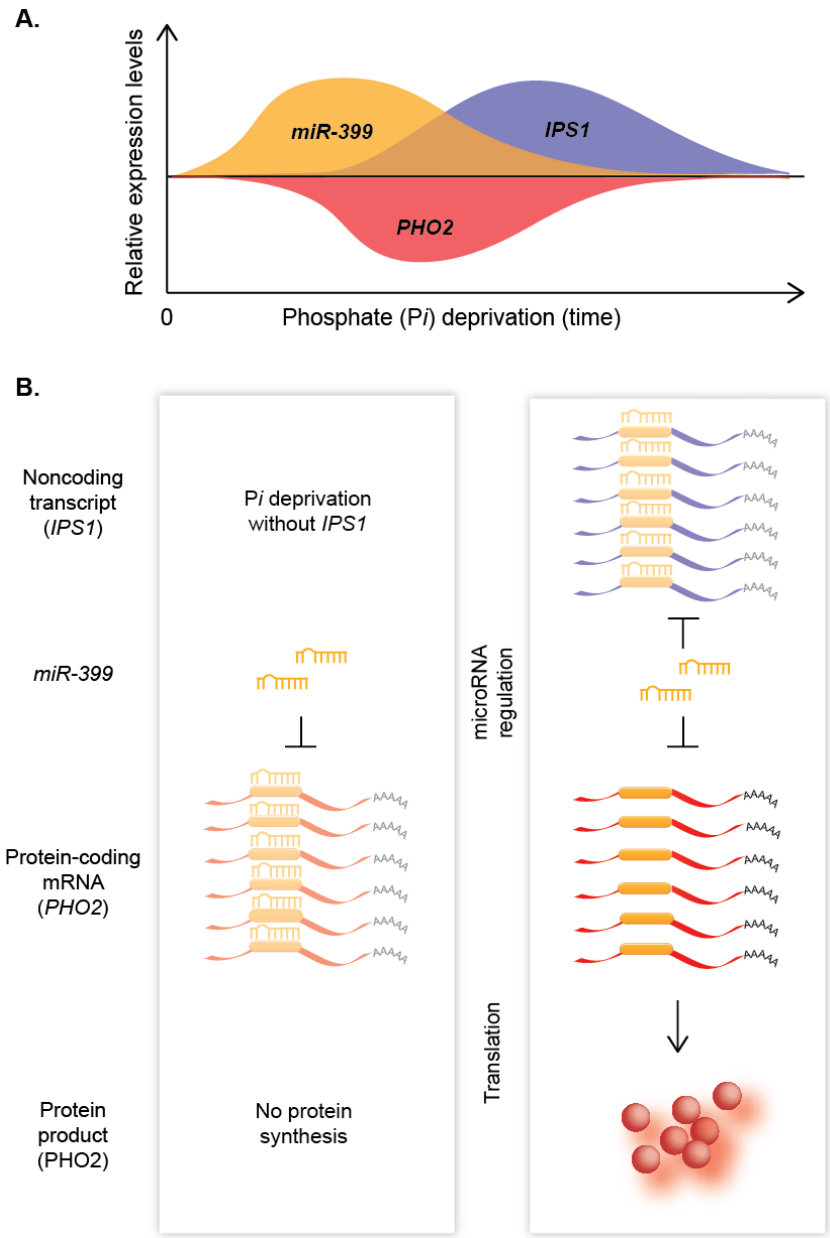


Figure 1.9 *IPS1* is an endogenous miRNA sponge for *miR-399* and regulates *PHO2*. A) Illustration of the changes in relative expression levels (y-axis) of a miRNA, *miR-399* (yellow), a noncoding RNA, *IPS1* (blue), and a protein-coding mRNA, *PHO2* (red), after the induction of phosphate (*Pi*) deprivation (x-axis, 0). *IPS1* is a molecular sponge for *miR-399* whose sponging activity reduces *miR-399* abundance and consequently, post-transcriptional derepresses levels of *PHO2*, another gene target of *miR-399*. **B)** Upon *Pi* deprivation and prior to *IPS1* induction (left panel), *miR-399* (yellow) post-transcriptionally represses *PHO2* (red) and suppress the levels of its protein product (red circles). *IPS1* expression (blue, right panel) sequesters *miR-399* and thereby, derepresses *PHO2* from its post-transcriptional regulation resulting in increased *PHO2* protein. Figure adapted from our review, Tan et al., 2014.

To date, my literature survey revealed at least 28 documented examples of endogenous transcripts that modulate the levels of other gene products via this miRNA-mediated mechanism (Table 1.2). Most reports were published following a series of articles (Poliseno et al., 2010; Salmena et al., 2011; Sumazin et al., 2011; Tay et al., 2011) (Figure 1.10) that described the miRNA-mediated interactions between *PTEN* and several coding and noncoding transcripts in cancerous cells. These studies were critical to illustrate how competition for shared miRNAs allows the reciprocal modulation of transcript levels and can lead to changes in cellular physiology and disease.

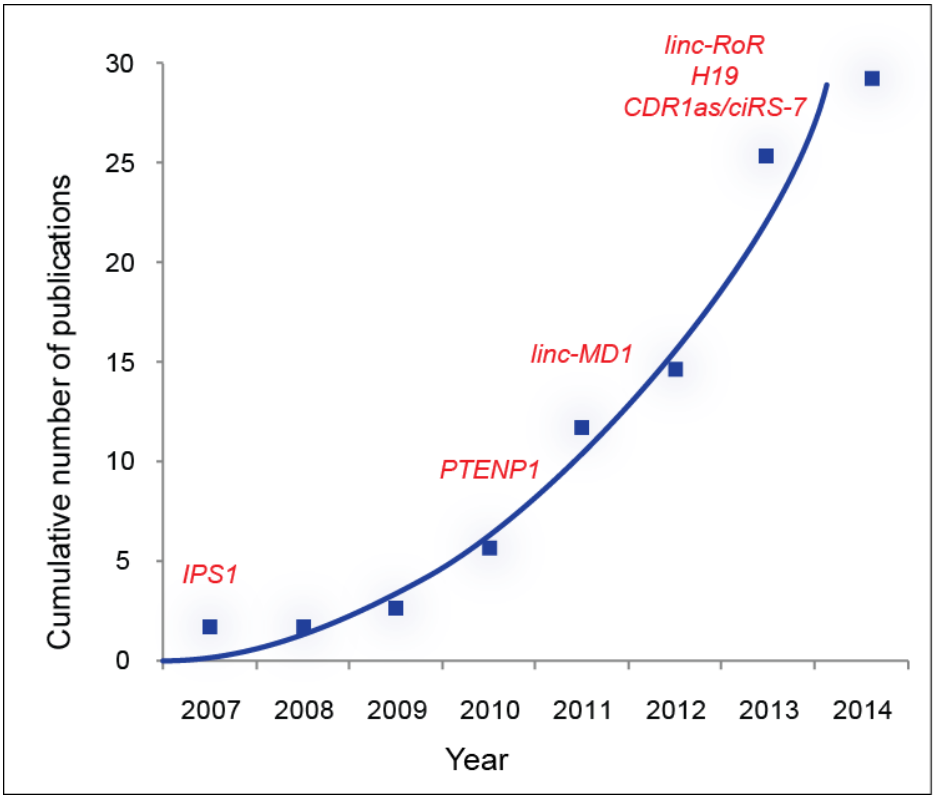


Figure 1.10 The cumulative number of manuscripts reporting on novel and experimentally characterized ceRNAs published from 2007 to 2014. Figure adapted from Tan et al., 2014.

Table 1.2 Examples of competitive endogenous RNA (ceRNA) interactions involving long noncoding transcripts characterized to date. This list does not contain references to competitive exogenous RNAs or poorly characterized examples of endogenous transcripts.

ceRNA	Gene class	miRNA(s) implicated	Crosstalking partner(s)	Disease(s)/ Development	Model organism(s)	References
IPS1 /At4 (TPSI family of ncRNAs)	Long noncoding RNA	*miR-399*	*PHO2*	Phosphate starvation	*Arabidopsis thaliana*	(Franco-Zorrilla et al., 2007)
PTENP1	Transcribed pseudogene	*miR-17-5p/ miR-20/miR-19*	*PTEN*	Prostate and colon cancers	Mouse/Human	(Poliseno et al., 2010)
KRAS1P	Transcribed pseudogene	*let-7/miR-143*	*KRAS*	Cell growth	Mouse/Human	(Poliseno et al., 2010)
PCas4	Unitary transcribed pseudogene	*miR-185*	*BCL2/IL17RD/ PNPLA3/SHISA7/T APBP*		Mouse/Human	(Marques et al., 2012)
ciRS-7 (CDR1as)	Circular RNA	*miR-7*	*SNCA/EGFR/IRS2*	Brain development	Mouse/Human	(Hansen et al., 2013; Memczak et al., 2013)
linc-RoR	Long noncoding RNA	*miR-145*	*Oct4/Nanog/Sox2*	Pluripotency maintenance and differentiation	Mouse	(Wang et al., 2013)
linc-MD1	Long noncoding RNA	*miR-133/-135*	*MAML1/MEF2C*	Muscle differentiation	Human/Mouse	(Cesana et al., 2011)
H19	Long noncoding RNA	*let-7*	*Dicer/Hmga2/Irs2/ Insr/CypB*	Muscle differentiation	Mouse/Human	(Kallen et al., 2013)
HULC	Long noncoding RNA	*miR-372*	*PRKACB*	Liver cancer	Human	(Wang et al., 2010a)
PTCSC3	Long noncoding RNA	*miR-574-5p*	Not characterized	Thyroid cancer	Human	(Fan et al., 2013)

1.5.1 Transcribed pseudogenes

The term competing endogenous RNAs was first used to describe the miRNA-mediated interaction between a transcribed retropseudogene, *PTENP1*, and its parental gene, the tumor suppressor, *PTEN* (Poliseno et al., 2010). Poliseno and colleagues found that *PTENP1* and *PTEN* transcripts harbour several homologous predicted MREs for 5 different miRNAs, *miR-20a*, *miR-16b*, *miR-21*, *miR-26a* and *miR-214*, and that the competition for the binding of these shared noncoding regulators allows the two transcripts to reciprocally regulate each other's levels (Poliseno et al., 2010).

In addition to *PTENP1*, another transcribed pseudogene, *KRAS1P*, was also demonstrated by Poliseno and colleagues to regulate the expression levels of its oncogenic ancestral gene, *KRAS*, by competing for the binding of *let-7* and *miR-143* (Poliseno et al., 2010). Most pseudogenes retain high sequence similarity with their ancestors. For instance, 71% of all human transcribed pseudogenes have remnant 3′ UTR sequences that are highly similar to their cognate ancestral genes (Kalyana-Sundaram et al., 2012). Therefore, if transcribed, their shared sequence similarity might allow them regulate the expression of their parental genes post-transcriptionally. Indeed, over 400 pseudogenes were identified to have parental genes that are already implicated in previously proposed ceRNA networks (Sumazin et al., 2011; Tay et al., 2011). Together, these observations suggest that some of the thousands of transcribed pseudogenes found in the mammalian genomes (Ohshima et al., 2003; Torrents et al., 2003; Zhang et al., 2003b) might also modulate the abundance of their protein-coding paralogs. More recently, the same functional

role was also demonstrated to be preserved for transcribed unitary pseudogenes in rodents, which will be discussed in more detail in **Chapter 3** (Marques et al., 2012).

1.5.2 Protein-coding mRNAs

In addition to *PTENP1*, the levels of *PTEN* in cancer cells are also modulated by other protein-coding ceRNAs through reciprocal ceRNA regulatory interactions (Figure 1.11). For example, miRNA-mediated crosstalk has been shown between *PTEN* and (i) *PTENP1*, vesicle-associated membrane protein associated protein A (*VAPA*), and CCR4-NOT transcription complex subunit 6-like (*CNOT6L*) in human colon and prostate cancer cells (Tay et al., 2011); (ii) *PTENP1*, *CNOT6L*, and zinc finger E-box binding homeobox 2 (*ZEB2*) in BRAF-induced melanoma cells (Karreth et al., 2011); and (iii) retinoblastoma protein (*RB1*), vascular endothelial growth factor A (*VEGFA*), and runt-related transcription factor 1 (*RUNX1*) in glioblastoma samples (Sumazin et al., 2011) (Figure 1.11). The partially non-overlapping ceRNA interactions in distinct cellular environments are likely a consequence of, among other things, the cell-specific expression of different miRNA repertoires and gene isoforms.

The complexity of the emerging ceRNA network underlying the post-transcriptional regulation of *PTEN* might be important to ensure robust and coordinated gene expression and to maintain homeostasis (Stark et al., 2005; Mukherji et al., 2011; Ebert and Sharp, 2012). However, this might also

represent a source of increased fragility and propensity for disease, as the loss of function in any of these PTEN modulators may potentially initiate a cascade of changes that eventually lead to the dysregulation of this tumor suppressor and to promote tumorigenesis.

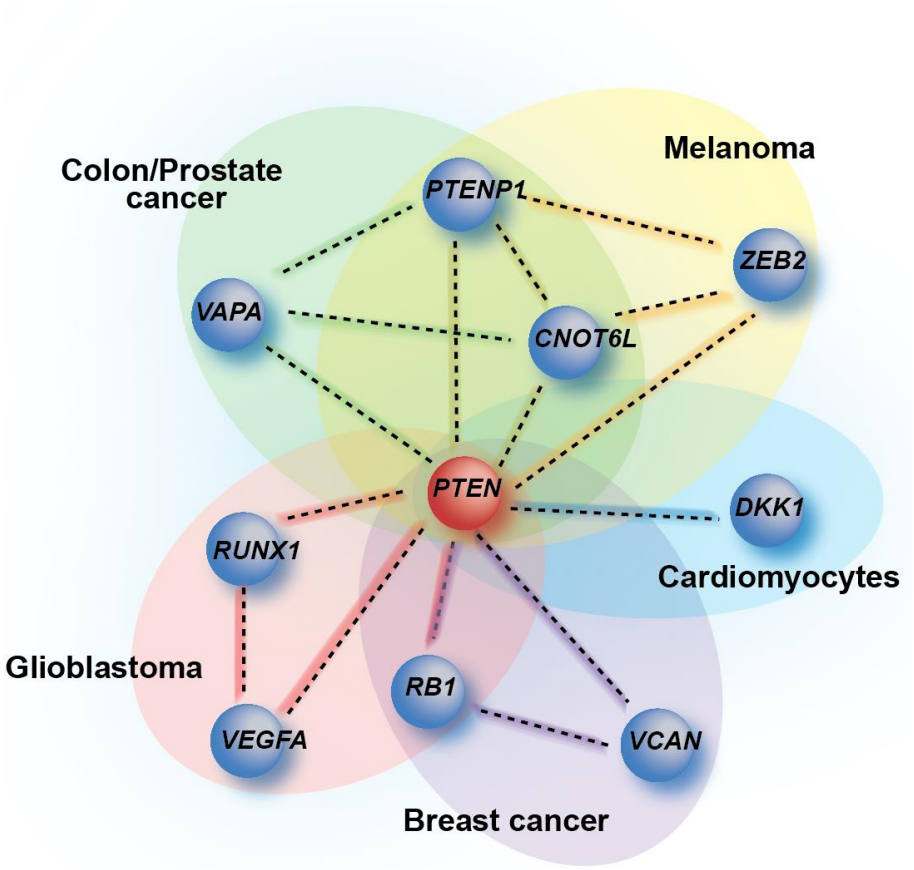


Figure 1.11 ceRNA network involving the tumor repressor gene *PTEN* (red) and its ceRNAs (blue) in 6 different types of cancers, including colon/prostate cancer (green shading), melanoma (yellow shading), cardiomyocytes (blue shading), breast cancer (purple shading) and glioblastoma (red shading). Dotted lines represent miRNA-mediated interactions between transcripts. Figure from Tan et al., 2014.

1.5.3 Circular noncoding RNAs

Recently, a noncoding circular RNA (circRNA), has also been shown to be an effective modulator of miRNA levels (Hansen et al., 2011; Hansen et al., 2013; Memczak et al., 2013). CircRNAs are abundant in the human genome (Salzman et al., 2012) and are potentially generated as a result of exon self-circularization (Jeck et al., 2013) and non-canonical modes of RNA splicing, such as exon scrambling (Cocquerelle et al., 1993; Salzman et al., 2012). Cerebellar degeneration-related protein 1 antisense (*CDR1as*) (Memczak et al., 2013) or circular RNA sponge for *miR-7* (*ciRS-7*) (Hansen et al., 2013) is a circRNA containing more than 70 conserved *miR-7* response elements (Hansen et al., 2011; Hansen et al., 2013; Memczak et al., 2013). *CDR1as*/*ciRS-7* is relatively highly expressed, with as many as 1,400 copies/cell. Similar to *miR-7*, the circRNA is predominantly found in neuronal tissues where it has been shown to sequester AGO2-bound *miR-7* (Hansen et al., 2013; Memczak et al., 2013). Consequently, by titrating *miR-7*, *CDR1as/ciRS-7* derepresses endogenous targets of *miR-7*, including *SCNA* (Junn et al., 2009), *EGFR* (Kefas et al., 2008), and *IRS2* (Jiang et al., 2010), which have been implicated in Parkinson's disease, cancer, and diabetes, respectively (Hansen et al., 2013). Interestingly, the linearization of *CDR1as/ciRS-7* leads to at least a 10-times fold reduction in the stability of its transcript (Memczak et al., 2013), suggesting that the efficiency of this transcript as a *miR-7* sponge is likely conferred by its circularity and abundance.

1.5.4 Intergenic long noncoding RNAs

Generally, long noncoding transcripts, including transcribed pseudogenes, circRNAs, and intergenic long noncoding RNAs (lincRNAs), have been proposed to act as miRNA sponges that effectively titrate available miRNAs from their endogenous targets, primarily due to their absence of apparent open-reading frames and translation (Su et al., 2013).

One lincRNA, termed *linc-RoR* for Regulator of Reprogramming, is particularly highly expressed during the reprogramming of induce pluripotency cells (iPSCs) and in undifferentiated ESCs (Loewer et al., 2010; Wang et al., 2013) (Figure 1.12). *Linc-RoR* overexpression increased the number iPSC colonies, where its repression reduced the success of somatic cell reprogramming (Loewer et al., 2010). Recently, *linc-RoR* was shown to compete for *miR-145* binding with key self-renewal transcription factors, including *Nanog*, *Oct4*, and *Sox2* (Wang et al., 2013). *Linc-RoR* overexpression was associated with increased levels of core pluripotency transcription factors. In contrast, the overexpression of a *linc-RoR* mutant, containing only mutated response elements for *miR-145*, had no impact on the abundance of these transcription factors (Wang et al., 2013). This is consistent with the hypothesis that *linc-RoR* regulates these transcripts via a *miR-145*-dependent mechanism. Interestingly, *linc-RoR* transcription is in turn regulated by Nanog, Oct4, and Sox2 proteins (Loewer et al., 2010), thus illustrating an autoregulatory feedback loop composed of transcription factors, a lincRNA and a miRNA within the regulatory circuitry that underlies cell fate decisions in ES cells (Wang et al., 2013). In addition to functioning as a ceRNA, *linc-RoR* was also proposed to promote cellular reprogramming by functioning

as a molecular scaffold for chromatin-modifying complexes (Zhang et al., 2013), hence suggesting the potential bifunctional roles of this lincRNA.

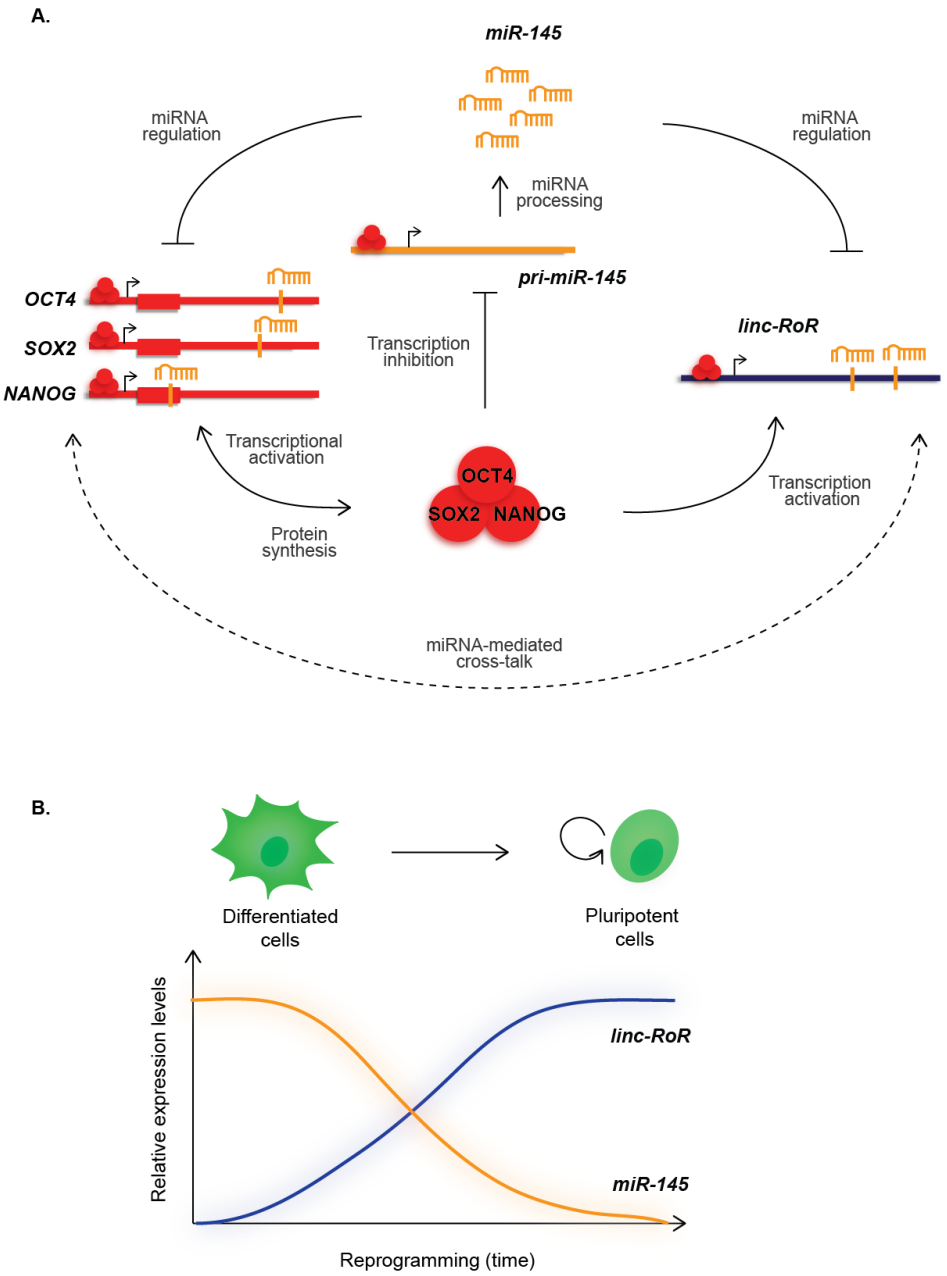


Figure 1.12 Contributions of *linc-RoR* to cellular reprogramming. (A) *miR-145* (yellow) recognizes *OCT4*, *SOX2*, and *NANOG* (red) mRNAs and *linc-RoR* (blue) via *miR-145* MREs within their sequences (yellow vertical lines) and post-transcriptionally regulates the abundances of OCT4, SOX2 and NANOG's protein products (red circles). These 3 core transcription factors are essential to activate their own mRNAs, *linc-RoR* and inhibit *pri-miR145* (red circles bound near the gene transcriptional start sites – represented as forward black arrows

on transcripts). Recently, *linc-RoR*, *OCT4*, *SOX2*, and *NANOG* were shown to crosstalk (dashed black line with double arrows) via the competition for *miR-145* binding. **(B)** This miRNA-mediated interaction was proposed to contribute to the reprogramming of terminally differentiated cells. Figure from Tan et al., 2014.

Similar to *linc-RoR*, *H19* is also a multi-functional lincRNA. The imprinted *H19* belongs to a highly conserved cluster of imprinted genes (Brannan et al., 1990; Rachmilewitz et al., 1992; Zhang and Tycko, 1992). This gene cluster also contains the paternally expressed insulin-like growth factor 2 (*Igf2*), whose reciprocal expression with the maternally expressed *H19* controls the timing of embryogenesis and postembryonic development (Elkin et al., 1995; Leighton et al., 1995; Arima et al., 1997). Interestingly, the *H19* RNA also encodes a conserved placenta-specific miRNA, *miR-675*. This miRNA appears to regulate placental growth, as it is exclusively expressed when placental growth halts, whereas *H19*-null placentas fail to stop growing (Keniry et al., 2012).

Although this lincRNA has been implicated in various human genetic disorders and cancers (Gabory et al., 2006; Matouk et al., 2007; Yoshimizu et al., 2008), its role and molecular mechanism of function during early embryonic development remains poorly understood. Recently, human *H19* was demonstrated to modulate the levels of the *let-7* miRNA family by acting as an endogenous sponge (Kallen et al., 2013). Similar to *lin-4*, *let-7* is another essential miRNA that regulates early developmental timing. Specifically, in *C. elegans*, the upregulation of *let-7* significantly represses heterochronic proteins that promote larval-specific cell fates and promotes the developmental transition from larval to adulthood (Pasquinelli et al., 2000; Reinhart et al., 2000; Thummel, 2001; Rougvie, 2005; Moss, 2007). The binding of *let-7* miRNAs to

H19 inhibited the transcript's activities in early embryo development (Kallen et al., 2013). Consistently, the knockdown of *H19* resulted in precocious muscle differentiation, a similar phenotype to that observed following *let-7* overexpression (Kallen et al., 2013).

The accelerated muscle differentiation process observed upon *H19* depletion suggests its role in controlling the timing of muscle differentiation, similar to that reported for the muscle-specific intergenic lincRNA, *linc-MD1*, which governs myoblast differentiation (Cesana et al., 2011). This muscle-specific intergenic long noncoding RNA contributes to the timing of muscle differentiation by acting as a molecular sponge for two muscle-specific miRNAs, *miR-133* and *miR-135* (Cesana et al., 2011) (Figure 1.13). Expressed exclusively during early stages of mouse myoblast differentiation, *linc-MD1* was found to compete for the binding of *miR-133* and *miR-135* with two key myogenic factors, *Maml1* and *Mef2c*, whose protein products activate muscle-specific gene expression (Cesana et al., 2011). Interestingly, the precursor transcript of *miR-133b* is derived from *linc-MD1*, thus suggesting an interesting, yet poorly understood, interplay between miRNA processing and modulation (Cesana et al., 2011). The function of mouse *linc-MD1* in myoblast differentiation is also conserved by its human orthologs, where its levels are significantly reduced in human myoblasts of Duchene Muscular Dystrophy (DMD) patients (Cesana et al., 2011), the most common form of muscular dystrophy (Blake et al., 2002). Interestingly, the ectopic expression of *linc-MD1* in DMD myoblasts rescued both *MAML1* and *MEF2C* synthesis and improved its delayed myoblast differentiation process compared to human primary myoblast controls (Cesana et al., 2011).

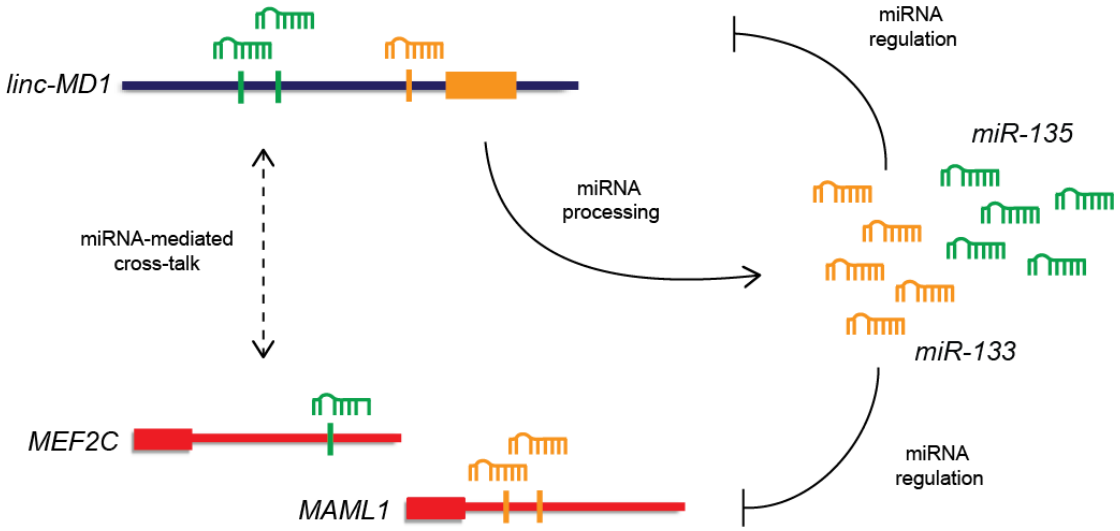


Figure 1.13 *Linc-MD1* encodes *miR-133* and is a miRNA sponge for this miRNA and *miR-135*. *Linc-MD1* (blue) is a muscle-specific lincRNA that competes with two protein-coding genes involved in the control of timing in myoblast differentiation, *MAML1* and *MEF2C* (red) for the binding (dashed black line with double arrows) of two muscle-specific miRNAs, *miR-133* (yellow) and *miR-135* (green). Not only is *linc-MD1* regulated by *miR-133*, it also encodes for the precursor of *miR-133b* (yellow box). Figure from Tan et al., 2014.

The contributions of ceRNAs to the control of developmental timing might be particularly significant and impactful when they involve transcripts whose levels are regulated through autoregulatory loops, as illustrated for *linc-RoR* (Wang et al., 2013), or when functionally-related genes are coordinately and cooperatively regulated by a common set of miRNAs, such as that demonstrated by *linc-MD1* (Cesana et al., 2011).

Finally, lincRNAs have also been found to contribute to cancer progression by participating in autoregulatory feedback loops within ceRNA networks. *Highly*

Up-regulated in Liver Cancer (*HULC*) was found to be highly abundant in hepatocellular carcinoma (Panzitt et al., 2007). This lincRNA is predominantly localized in the cytoplasm and competes for the binding of *miR-372* with *PRKACB*, which encodes a protein kinase that phosphorylates cAMP response binding protein (CREB) (Wang et al., 2010a). Phosphorylation is required for CREB activation, which in turn promotes the transcription activation of *HULC* mRNA (Wang et al., 2010a). In addition to being highly expressed in liver cancer, *HULC* is also abundant in metastasized colorectal carcinomas (Matouk et al., 2009). Overall, the presence of high levels of *HULC* in different cancers may suggest its wider contribution to tumorigenesis.

These examples illustrate how miRNA-mediated crosstalk between coding and long noncoding transcripts can contribute to complex circuitries that underlie cell pluripotency, development and disease. Importantly, these well-characterized ceRNAs provide important insights into miRNA-mediated molecular mechanisms of post-transcriptional regulation and what might be their contributions to different aspects of biology. Furthermore, the vast amount of emerging evidence of lincRNAs acting as ceRNAs, for which I term lnceRNAs, in diverse cellular environments suggests the high prevalence of this mechanism. With the current curated catalogue of characterized lnceRNAs, we may only be at the tip of an iceberg as further investigations on lnceRNAs continue.

1.6 THESIS SCOPE AND STRUCTURE

Although examples of individual ceRNA-acting long intergenic noncoding RNAs (lnceRNAs) have already been reported (Salmena et al., 2011), their prevalence and their roles as miRNA decoys remain largely unexplored. In my thesis, I aim to provide a broad perspective on the significance of such intergenic long noncoding sponges and to identify their associated characteristics. Ultimately, this will not only enhance our mechanistic understanding of these new players in gene regulation, but will also provide guidelines on how to predict or classify additional ceRNA-acting lincRNAs.

In my thesis, I aim to address the following questions:

1) What is the biological relevance of the regulation of gene product abundance by lincRNAs that act as ceRNAs (**Chapter 3**)?

2) What is the prevalence of lincRNAs that function as ceRNAs (**Chapter 4**)?

3) What are the contributions of lincRNAs that act as ceRNAs to diseases (**Chapters 5 and 6**)?

1.7 PUBLICATIONS

Chapter 1:

Tan, J.Y., and Marques, A.C.
The miRNA-Mediated Crosstalk between Transcripts Provides a Novel Layer of Posttranscriptional Regulation.
Advances in genetics. 2014. 85, 149-199.

Marques, A.C., **Tan, J.**, and Ponting, C.P.
Wrangling for microRNAs provokes much crosstalk.
Genome Biology. 2011. 12, 132.

Chapter 3:

Marques, A.C., **Tan, J.**, Lee, S., Kong, L., Heger, A., and Ponting, C.P.
Evidence for conserved post-transcriptional roles of unitary pseudogenes and for frequent bifunctionality of mRNAs.
Genome Biology. 2012. 13, R102.

Chapter 4:

Tan, J.Y., Sirey, T., Honti, F., Piovesan, A., Webber, C., Ponting, C.P., and Marques, A.C.
Extensive microRNA-mediated crosstalk between lncRNAs and mRNAs in mouse embryonic stem cells
Genome Research. 2014. Under review.

Chapter 5:

Tan, J.Y., Vance, K.W., Varela, M.A., Sirey, T., Watson, L.M., Curtis, H.J., Marinello, M., Alves, S., Steinkraus, B, Cooper, S., Nesterova, T., Brockdorff, N, Fulga, T, Brice, A, Sittler, A, Oliver, P.L., Wood, M.J., Ponting, C.P., and Marques, A.C.
Crosstalking noncoding RNAs contribute to cell-specific neurodegeneration in Spinocerebellar ataxia type 7.
Nature Structural Molecular Biology. 2014. Accepted.

CHAPTER 2

General Materials and Methods

2.1 TISSUE CULTURE

Mouse neuroblastoma (N2A) and human neuroblastoma (SH-SY5Y and SK-N-SH) cells were grown at 37^{o}C in a humidified atmosphere supplemented with 5% CO_2, in *Dulbecco's Modified Eagle Medium* (DMEM, Invitrogen) containing antibiotic penicillin/streptomycin, supplemented with 10% Fetal Calf Serum (FCS, Invitrogen).

Human retinoblastoma (WERI) cells were kindly provided by Dr. Alun Barnard, Dr. Michelle McClements, and Prof. Robert MacLaren. WERI cells were simi-adherent and were grown at 37^{o}C in 5% CO_2 in Iscove's Modified Dulbecco's Medium (IMDM, Invitrogen) supplemented with 2mM L-glutamine (Invitrogen), 55µM Beta-mercaptoethanol (Invitrogen), 10µg/mL Insulin (Sigma-Aldrich), 10% FCS and antibiotics.

Mouse DTCM23/49 XY embryonic stem (ES) cell lines (mESCs) were kindly provided by Dr. Tatyana Nesterova, Dr. Sarah Cooper and Dr. Bryony Graham and were grown as described previously (Nesterova et al., 2008). Deletion of Dicer's RNAse III domain was induced by culturing the cells in the presence of 800nM (Z)-4-Hydroxytamoxifen (4-OHT, Sigma); non-induced cells were treated with 0.1% ethanol and used as control.

SCA7 patient-derived fibroblast cell lines ($SCA7^{42Q/10Q}$, $SCA7^{49Q/10Q}$, $SCA7^{55Q/10Q}$) were obtained and maintained by Dr. Lauren Watson and Dr. Miguel Varela. These cell lines were incubated at 37 °C in 5% CO_2 in DMEM Glutamax (Gibco-BRL) supplemented with 10% FBS and antibiotics. Ethics approval for the establishment of patient fibroblast cultures was granted by the University of Cape Town (UCT) Faculty of Health Sciences Human Research Ethics Committee (HREC REF. 380/2009 and 434/2011), and was renewed annually.

2.2 GAIN- AND LOSS-OF-FUNCTION CONSTRUCT DESIGN

Target gene overexpression was executed using gene constructs containing sequence of the genes of interest driven by an exogenous promoter. Gene overexpression was achieved by cloning the sequence of the genes into the multiple cloning site (MCS) downstream of a CMV promoter on a *pcDNA3.1(+)* vector. Empty vector with no gene insert was used as an overexpression control.

Target gene knockdown was achieved using small interfering RNA (siRNA) or short hairpin RNA (shRNA) constructs. Multiple siRNA constructs were designed using the siRNA selection program from the Whitehead Institute (Yuan et al., 2004) to specifically target the gene of interest by selecting regions without substantial sequence similarity to other gene transcripts. As a control,

an oligo with randomly permutated nucleotides (scrambled control) was used that shows no significant sequence similarity to mRNAs from the genome.

Similarly, multiple shRNAs were made using sequences of the previously designed siRNAs or the non-specific scrambled control used as transfection control. Oligos of designed siRNAs and scramble control sequences were reverse complemented and the two arms of the hairpin linked by a loop sequence (TTCAAGAGA) with the adaptors required for cloning added to the ends of the oligos. Specifically, HPLC purified custom-made oligos (Sigma-Aldrich) were resuspended in water to a final concentration of 100 µM. For each shRNA, 10 µl of forward and reverse oligos were added to 160 µl of annealing buffer (10 mM Tris pH 8, 50 mM of NaCl) and incubated for 5 minutes at 95°C. After cooling to room temperature, oligos were phosphorylated using T4 polynucleotide kinase enzyme (New England BioLabs) and cloned downstream of a U6 promoter from a modified *pll3.7* vector (courtesy of Dr Esther Becker). All shRNAs were tested for their association with decreased levels of their target genes. The shRNA construct that achieved the greatest impact on target gene levels was used in subsequent experiments.

Construct designs are illustrated in Table 2.1. Primers used in Chapter 3 to 6 are listed in Table 2.2 – 2.5.

2.3 MUTAGENESIS

Directed mutagenesis of the gene of interest cloned within plasmid constructs were generated by PCR using 2 µl of reverse transcribed cDNA with 300 mM of primers, 1 U Expand High Fidelity DNA polymerase (Roche), 1.5 mM $MgCl_2$, 0.2 mM dNTPs, 5% DMSO and 10X buffer in 50 µl total volume. PCR reaction was carried out in a Veriti 96-well (Applied Biosystems) thermocycler as following: 94°C for 2 min, followed by 5 cycles with 15s at 94°C, 15s at a temperature gradient of 58 – 68°C and 2min at 94°C, followed by 15 cycles with 15s at 94°C, 30s at a temperature gradient of 55 – 65°C and 2 min at 72°C with 5s extension added after each cycle, followed by a terminal step at 72°C for 7min.

2.4 *IN VITRO* TRANSFECTION

Cultured cells were grown under standard conditions and 24 h before transfection, 1×10^5 cells/ml were plated in six-well cluster culture vessels (2mL per well). Overexpression or knockdown (shRNA) vectors with their respective control constructs, scrambled hairpin RNA (*sh-scramble*) and empty vector, respectively, at 1 µg were transfected using FuGENE 6 Transfection Agent (Roche) in triplicates according to the manufacturer's guideline unless otherwise specified. Gene silencing siRNAs (5 nM/well) (FlexiTube, Qiagen), and their negative control (*si-NC*, Cat. No. 1027280, Qiagen) were transfected using Lipofectamine RNAiMAX Reagent (Invitrogen).

The expression level of microRNAs (miRNAs) could be altered by introducing mimics or inhibitors of miRNAs of interest into the cellular system. A miRNA's mimic or inhibitor (50 nM/well) (mirVana, ABI), and their negative controls (*miR-NC*, Cat. No. 4464058, ABI) were transfected using Lipofectamine RNAiMAX Reagent (Invitrogen).

After transfection, cells were grown under standard conditions prior to harvesting at 48 h post-transfection unless otherwise specified.

2.5 RNA EXTRACTION AND QUANTIFICATION

Total cellular RNA was extracted from culture cells using the RNeasy kit (Qiagen) following the manufacturer's instructions. To quantify levels of mature miRNAs, total RNA was extracted using the miRNeasy kit (Qiagen). Genomic DNA was removed using the DNA-free kit (Ambion). RNA was reversed transcribed and cDNA was used to quantify gene expression changes, relative to *Gadph*, using sequence specific primers. Expression levels of genes of interest were estimated by real-time quantitative PCR (qRT-PCR) on a StepOneReal-Time PCR thermocycler (ABI) using SYBR green Master PCR mix (ABI) and gene loci specific primers in triplicate unless otherwise specified. Non-reverse transcribed RNA was used as negative amplification control.

For miRNA quantification, RNA was reverse transcribed using the TaqMan MicroRNA Reverse Transcription Kit (Invitrogen). For comparison of miRNA expression between cell lines or tissues, RNA was reverse transcribed using

the NCode VILO miRNA cDNA Synthesis Kit (Invitrogen), which was able to reverse transcribe both polyadenylated and non-polyadenylated RNA transcripts. MiRNA abundance was measured by qRT-PCR (SYBR green) using miRNA-specific TaqMan MicroRNA Assays (ABI) according to the manufacturer's instructions. The expression levels of miRNAs were normalized to that of 18S rRNA.

Primers used in Chapter 3 to 6 are listed in Table 2.2 – 2.5.

2.6 SUBCELLULAR FRACTIONATION

Subcellular fractionation of cultured cells was carried out using the PARIS kit (Invitrogen) following the manufacturer's instructions. After isolating the nuclear and cytoplasmic fractions from total cell lysates, RNA from each subcellular compartment was extracted and reverse transcribed as described previously. Using qRT-PCR, expression levels of all genes were measured independently in the cytoplasmic and nuclear fractions. Fold enrichment between the distinct compartments (ratio of expression level measured in the cytoplasmic/expression level measured in the nucleus) was reported after normalization to *Gapdh*. *Malat1* (Ji et al., 2003; Tripathi et al., 2010) was used as a control nuclear marker.

Primers used in Chapter 3 to 6 are listed in Table 2.2 – 2.5.

2.7 LUCIFERASE ASSAY

Luciferase assays were performed to assess the stability of genes of interest. Specifically, after potential binding of miRNA to their predicted miRNA response elements (MREs) within the genes, sequences of the genes of interest were cloned into the *XbaI* restriction site downstream of the luciferase reporter gene in the *pGL3-promoter* (*pGL3-pro*) vector (Promega). Each luciferase construct (2 µg) was co-transfected into cultured cells with 10 ng of pRL-*Renilla* luciferase control vector (Promega) and 50 nM of the mimics of the miRNA(s) of interest (mirVana, Invitrogen) or negative miRNA control mimics (miR-NC, Invitrogen) using the FuGENE 6 Transfection Agent (Roche). Transfection of pRL-*Renilla* luciferase control vector was used as control for transfection efficiency.

To test functional binding of miRNA to its predicted MREs, MREs harboured within genes of interest were disrupted by directed mutagenesis. Specifically, the MRE region(s) complementary to the miRNA seed, 5'-XXXYYY-3', were mutated by reversing the sequence to 5'-YYYXXX-3', where X and Y represent different DNA nucleotides and that 5'-YYYXXX-3' do not represent any seed for another known expressed miRNA in the cell line used. Comparison between the stability of the co-transfected luciferase constructs, which contains the wild-type (WT) gene sequence or that with mutated MRE(s), with mimics of miRNA(s) allowed direct analysis of functional binding of the miRNA to their predicted MRE(s) within the genes under study.

A luciferase assay was also carried out to test the location of functional regulatory elements. Putative regulatory elements of gene locus and a negative

control region (a region with no regulatory activity with similar genome composition as the predicted regulatory elements tested) were cloned into restriction sites upstream of the luciferase reporter gene in the *pGL3-enhancer* (*pGL3-enh*) vector (Promega). As a control, the same sequences were cloned into the same location in the reverse orientation. As described above, each luciferase construct (2 μg) was similarly co-transfected into cultured cells with 10 ng of pRL-*Renilla* luciferase control vector (Promega) using the FuGENE 6 Transfection Agent (Roche).

Following transfection, transfected cells were grown under standard conditions for 48 h before harvesting. Dual luciferase activity was measured using the Dual-luciferase reporter assay system (Promega) according to the manufacturer's guidelines on the FLUOstar OPTIMA (BMG Labtech) fluorescence plate reader. Luciferase activity was normalized against measured *Renilla* activity as proposed by the manufacturer.

Construct designs are illustrated in Table 2.1. Primers used in Chapter 3 to 6 are listed in Table 2.2 – 2.5.

2.8 CHROMATIN IMMUNOPRECIPITATION

All chromatin immunoprecipitation (ChIP) assays were performed by Dr. Keith Vance. Cultured cells were directly cross-linked for 10 min at 37 ^{o}C by adding 1% formaldehyde to the tissue culture medium. For animal tissue samples, dissected samples were dounced in PBS using a homogenizer to generate a

single cell suspension. Next, 1% final concentration formaldehyde was added and the samples are incubated for 10 min at room temperature with rotation. Cross-linking reactions were quenched using 0.125M glycine. Nuclei were isolated and chromatin sheared to approximately 500bp using a Bioruptor (Diagenode). Subsequently, 1 mg cross-links were immunoprecipitated using antibody (5 µg) specific for the protein of the interest or anti-rabbit IgG (5 µg, Millipore) as control overnight at 4^0C. Complexes were collected using Protein-A magnetic beads (Pierce), washed, eluted and the cross-links were reversed at 65°C overnight. DNA is precipitated, treated with Proteinase K (Roche) and purified using a PCR Purification Kit (Qiagen).

Genomic regions predicted to bind the protein of tested were tested using primers specifically targeting these regions that were designed with a similar nucleotide composition. Specific enrichment of binding by the protein of interest relative to IgG at the tested genomic regions was determined from three independent ChIP assays by qRT-PCR.

Primers used for ChIP-qPCR analyses in Chapter 5 are listed in Table 2.4.

2.9 GENE EXPRESSION PROFILING ACROSS TISSUES

Microarray gene expression data for genes of interest were obtained from Gene Expression Atlas (GNF) through BioGPS (http://biogps.org) for human (Su et al., 2004), and the correlation coefficient (Pearson's correlation, R^2) between the expression levels of pairs of genes was computed across multiple tissues or

cell lines where the gene loci are expressed (AD>20; Normalized Affymetrix microarray expression values using GCRMA algorithm (Su et al., 2004)).

Total RNA from 20 human normal adult tissues (adipose, bladder, brain, cervix, colon, esophagus, heart, kidney, liver, lung, ovary, placenta, prostate, skeletal muscle, small intestine, spleen, testes, thymus, thyroid, and trachea) was acquired from the FirstChoice Human Total RNA Survey Panel (Invitrogen). Total RNA from 11 mouse normal adult tissues (bladder, brain, colon, heart, kidney, liver, lung, pancreas, skeletal muscle, small intestine, and stomach) was obtained from Mouse Tissue Total RNA Panel (Amsbio). Total RNA from 9 mouse (postnatal day 5, pooled from 7 individuals) brain tissues (cerebellum, cortex, entorhinal cortex, hippocampus, hypothalamus, medulla, olfactory bulb, and striatum), courtesy of Dr. Tamara Sirey, was extracted using Trizol (Invitrogen)**.** RNA from mouse retina was extracted from wild-type (WT) C57BL/6 mice (2 pooled individuals) and human retinal RNA was extracted from human WERI retinoblastoma cells using RNeasy kit (Qiagen). RNA was reverse transcribed into cDNA and expression levels of genes were estimated by qRT-PCR using loci specific primers in triplicate as described previously. Non-reverse transcribed RNA was used as negative amplification control.

2.10 ABSOLUTE QUANTIFICATION OF TRANSCRIPT ABUNDANCE

Absolute quantification of transcript abundance was determined using digital droplet PCR (ddPCR). All experiments were performed according to the manufacturer's (Bio-Rad Laboratories Inc.) instructions with assistance from Bruno Steinkraus. Briefly, 20µl reactions were assembled with 2X ddPCR Supermix for Probes (Cat. No. 186-3023, Bio-Rad), 20X TaqMan Gene Expression Assay for genes of interest (Life Technologies), 1 µl of sample cDNA that was reverse transcribed from 2 ng/µl of RNA, and nuclease-free water. Droplets were generated with Droplet Generation Oil for Probes (Cat. No. 186-3030, Bio-Rad) on a QX-100 Droplet Generator, which was then carefully transferred to a 96-well plate and sealed with pierceable foil on a PX1 heat plate sealer (Bio-Rad). PCR was performed on a C1000 Touch with the following program: 95°C for 10min, 40 cycles of 94°C for 30s, 59°C for 1min, and 98°C 10min. Droplet fluorescence was detected on a QX100 Droplet Reader (Bio-Rad) in absolute quantification mode (ABS) with QuantaSoft Version 1.3.2.0 (Bio-Rad). Clear separation of droplet amplitudes was observed and thresholds were set to quantify positive droplets. Non-template controls were performed for each probe and technical triplicates were averaged. Absolute copy number of genes was quantified as copies/µL of cDNA after normalization to *Gapdh*.

2.11 TISSUE PREPARATION FOR RNA ANALYSES

Biopsies from neuronal tissues (retina, cerebellum, cortex, striatum, olfactory bulb, and spinal cord) were collected at 4°C using an adult mouse brain matrix slicer by Dr. Peter Oliver. Non-neuronal tissues (liver, lung and muscle) were dissected at 4°C. After dissection, biopsies were immediately frozen in liquid nitrogen and stored at -80°C until RNA extraction. Tissues were homogenized in RLT buffer with 1% β-mercaptoethanol (RNeasy kit, Qiagen) and prepared for RNA extraction, followed by reverse transcription to cDNA and gene expression detection by qRT-PCR as described previously.

2.12 *IN-SITU* HYBRIDIZATION

In-situ hybridization assays performed by Dr. Peter Oliver were used to detect expression of gene transcripts, to identify subcellular localization of the transcripts, and/or to compare their expression levels between tissue samples. Target sequences of interest were generated by PCR or qRT-PCR and cloned into *pCR4-TOPO* (Invitrogen). Digoxigenin (DIG)-labelled riboprobes were synthesized from linearized plasmid DNA. Tissue samples were snap frozen in OCT (VWR) and 14 µm sections were cut using a cryostat (Leica) and mounted onto Superfrost Plus slides (VWR). Probe hybridization, washing and signal detection using an alkaline-phosphatase conjugated anti-DIG antibody for genes of interest (mRNAs or lncRNAs) was carried out as previously described (Chodroff et al., 2010). Sense strand probes were also tested to obtain a

negative control signal (data not shown). For detection of small RNAs (i.e. miRNA), a DIG-labelled LNA probe (Exiqon) specific to the small RNA of interest was hybridized as above with some slight modifications (Deo et al., 2006). To ensure signals obtained prior to analysis were at sub-saturation levels, probes for mRNAs or lncRNAs are hybridized for 16 hours, whereas probes for small RNAs were hybridized for 4 hours.

Primers used in for *in-situ* hybridization analyses in Chapter 5 are listed in Table 2.4.

2.13 WESTERN BLOTTING

Western blots were performed to determine presence and relative expression levels of proteins of interest. First, whole cell extracts were lysed using RIPA buffer (Sigma) and the total protein concentration was determined using the BCA assay kit (Pierce). 100 µg of protein was loaded onto a 4-20% gradient Tris-Glycine gel (Invitrogen) and separated using gel electrophoresis. Proteins were then transferred onto a PVDF membrane in 1x Transfer buffer (48mM Tris, 39mM Glycine, 20% methanol) at 40V for 2 hrs. To determine successful transfer of total protein from the gel, the membrane was stained with Ponceau S (Sigma P7170-1L) according to the manufacturer's guideline and visualized for presence of proteins. Water was used to remove Ponceau S staining prior to western blot.

Subsequently, the membrane was blocked with 5% skimmed milk and incubated overnight in TBST buffer (0.9% NaCl, 100nM Tris, 1% Tween20) with primary antibodies at 4^{o}C to detect proteins of interest. Following incubation with primary antibodies, the membrane was washed 4 times for 30 min in TBST and incubated with biotinylated secondary antibodies that recognize and bind to the primary antibodies for 1 h, followed by multiple washes in TBST as above. The membrane was then incubated with horseradish peroxidase-conjugated streptavidin (ab7403, Abcam, 1:10,000) for an additional hour, followed by the same steps of stringent washes in TBST. Enhanced chemiluminesence detection of protein presence and relative abundance was performed as recommended by the manufacturer (Amersham). Finally, the membrane was stripped by washing with stripping buffer (15g Glycine, 1g SDS, and 10mL Tween20) twice for 10min, followed by two washes in PBS for 10 min and two washes in TBST for 5 min. The membrane was then blocked and re-probed with a positive control primary antibody, anti-α-tubulin (ab7291, Abcam, working dilution 1:5,000), followed by the procedure as described above [secondary antibody (ab64255), washes, horseradish peroxidase-conjugated streptavidin, washes, and detection].

2.14 GENOME-WIDE ANALYSIS OF MIRNA ABUNDANCE

To determine the genome-wide abundance of miRNAs, total RNA was extracted (as described previously) from cell line or tissue samples in mouse. Four biological replicates were assayed for each experimental condition unless

otherwise specified. For each sample, the expression level of a total of 611 mouse and murine virus-associated miRNAs was quantified using the nCounter miRNA Expression Assay (NanoString Technologies, Seattle, WA) (Geiss et al., 2008) according to the manufacturer's instructions by Bruno Steinkraus. Briefly, input RNA (105ng) and miRtag linkers were ligated prior to hybridization with barcoded reporter and biotinylated capture probes at 65°C for 16h. Samples were prepared for analysis on the nCounter Prep Station before data were collected at 555 FOV on the nCounter Digital Analyzer. Finally, generated data were analyzed using the NanoString Differential Expression (NanoStriDE) (Brumbaugh et al., 2011) interface and normalized using a set of house-keeping mRNAs, followed by Benjamini and Hochberg multiple test correction.

2.15 PREDICTION OF MIRNA RESPONSE ELEMENTS

Two algorithms were used for the prediction of miRNA response elements (MREs) within genes of interest. In **Chapter 5**, predicted MREs within target genes were downloaded from microRNA.org (all mirSVR scores) (Betel et al., 2008). In **Chapters 3, 4 and 6**, prediction of MREs within lncRNAs and protein-coding mRNAs were performed by TargetScan (Garcia et al., 2011).

2.16 STATISTICS

All expression correlation comparisons were determined using the Pearson's correlation test and all differential expression comparisons were determined using Student's *t*-test. Asterisks indicate the level of significance of the comparison between the expression of target transcripts (* $p<0.05$; ** $p<0.01$; *** $p<0.001$; NS [not significant] $p>0.05$). For each experimental analysis statistical values were calculated using data collected from three independent experiments.

Table 2.1 Construct table of vector backbones.

Overexpression constructs		
pcDNA3.1(+): 5428 bp	CMV promoter Multiple cloning site Ampicillin resistance gene	Bases 232-819 Bases 895-1010 Bases 4432-5428
Knockdown constructs (shRNA)		
pll3.7: 7650 bp	U6 promoter Multiple cloning site Ampicillin resistance gene	Bases 2616-2925 Bases 2935-2950 Bases 7517-6657
pTK-Hyg: 5040 bp	HSV TK promoter Hygromycin resistance gene Ampicillin resistance gene	Bases 2588-2835 Bases 1537-2574 Bases 3493-4803
Luciferase assay constructs		
pGL3-pro: 5010 bp	Promoter Luciferase gene (*luc*+) *XbaI* cloning site Ampicillin resistance gene	Bases 48-250 Bases 280-1932 Base 1934 Bases 3272-4132
pGL3-enh: 5064 bp	Promoter Multiple cloning site Luciferase gene (*luc*+) Ampicillin resistance gene	None Bases 1-58 Bases 88-1740 Bases 3329-4186
pRL-Renilla: 4166 bp	Promoter *Renilla* luciferase (*Rluc*) Ampicillin resistance gene	Bases 543-758 Bases 829-961 Bases 2480-3340

Table 2.2 Custom oligonucleotide sequences used in Chapter 3.

PCR primers		
	Forward (5’ to 3’)	**Reverse (5’ to 3’)**
ENSP00000352956	TGCCATGGTTCCTTGTCAAG	GGTTCGCTTTGTTCCTTCAG
Pbcas4	CAAGCTGTACAGGACGAAG	GTGACTAACATGGCAGGGTC
ENSP00000387075	GGTCATCAGACAAGACCTAC	GGAGCAACATCTGACAGTTG
ENSP00000358232	CTTGCTCCTGACTGTAGGTC	CTGATTGCCTCAGAGCTTTG
ENSP00000261047	GGCCAATCAACAAGATCAAG	GAACAGCGGTCATAATGAAG
qRT-PCR primers		
	Forward (5’ to 3’)	**Reverse (5’ to 3’)**
Pbcas4	AGCCCCTTGACTGTTAGCAT	CAAGGGAGCAAGAGAACCAA
GADPH	TGTGTCCGTCGTGGATCTGA	CCTGCTTCACCACCTTCTTGA
RACE primers		
5' outer	CTGCACCCTGGCATGCTAACAGTCAAG	
5' inner	GTGAGATCTCATGGCTGCTTTGCACAG	
3' outer	CTTGACTGTTAGCATGCCAGGGTGCAG	
3' inner	CTGTGCAAAGCAGCCATGAGATCTCAC	
siRNA oligos		
	Forward (5’ to 3’)	**Reverse (5’ to 3’)**
sh-pbcas4	GATCCCTGTTGTGACATGATTCAAGAGACCATGT CACAACAGGGATCCTTTTTC	TCGAGAAAAAGGATCCCTGTTGTGACATGGTCTC TTGAATCATGTCACAACAGGGATC
sh-scramble control	AGATACTCGATCCCGCGCTTCAAGAGAACGCGG GATCGAGTATCTCTTTTTC	TCGAGAAAAAGAGATACTCGATCCCGCGTTCTCTT GAAGCGCGGGATCGAGTATCT

Table 2.3 Custom oligonucleotide sequences used in Chapter 4.

Mouse qRT-PCR primers		
	Forward (5’ to 3’)	**Reverse (5’ to 3’)**
Dicer	CGATATTCGATCCTCCTGTGA	GCCAGCAAGCAGTCTTTTGT
Nanog	GAACTATTCTTGCTTACAAGGGTCTGC	GCATCTTCTGCTTCCTGGCAA
Oct4	CGTGGAGACTTTGCAGCCTG	GCTTGGCAAACTGTTCTAGCTCCT
Sox2	TGGAAACTTTTGTCCGAGA	GAAGCGTGTACTTATCCTTCTTCAT
Gapdh	TGTGTCCGTCGTGGATCTGA	CCTGCTTCACCACCTTCTTGA
miR-302a-3p	TAAGTGCTTCCATGTTTTGGTGA	
miR-200c-3p	TAACACTGTCTGGTAACGATGT	
18S rRNA	GTAACCCGTTGAACCCCATT	CCATCCAATCGGTAGTAGCG

Table 2.4 Custom oligonucleotide sequences used in Chapter 5.

shRNA oligos		
	Forward (5’ to 3’)	**Reverse (5’ to 3’)**
sh-lnc-SCA7-1	AATTACATACTGCGGACATTCAAGAGACGTCCG CAGTATGTAATTCTTTTTC	TCGAGAAAAAGAATTACATACTGCGGACGTCT CTTGAATGTCCGCAGTATGTAATT
sh-lnc-SCA7-2	TCTCGACCAATTTCAGCTTTCAAGAGACGCTGA AATTGGTCGAGACTTTTTC	TCGAGAAAAAGTCTCGACCAATTTCAGCGTCT CTTGAAAGCTGAAATTGGTCGAGA
sh-lnc-SCA7-3	CATATAGTTGACAGTGGTTTCAAGAGACCCACT GTCAACTATATGCTTTTTC	TCGAGAAAAAGCATATAGTTGACAGTGGGTCT CTTGAAACCACTGTCAACTATATG
sh-scramble	AGATACTCGATCCCGCGCTTCAAGAGAACGCG GGATCGAGTATCTCTTTTTC	TCGAGAAAAAGAGATACTCGATCCCGCGTTCT CTTGAAGCGCGGGATCGAGTATCT
Mouse qRT-PCR primers		
	Forward (5’ to 3’)	**Reverse (5’ to 3’)**
Atxn7	TCTCTCTCGGCCCATTCATC	GGGTTTGCTGCCACTGTTG
lnc-SCA7	GGGCCTCCTTCGAAGTTTTAG	TGTATCAAAGAGGCCCAATTGG
lnc-SCA7_ORF	ATTCCTGCCTGGGCTTCTG	TGGATGCCGAAATCCTTCAC
18S_rRNA	GTAACCCGTTGAACCCCATT	CCATCCAATCGGTAGTAGCG
Gapdh	TGTGTCCGTCGTGGATCTGA	CCTGCTTCACCACCTTCTTGA
Eny2	TTTCGGAGGGAAAAAACAAGAC	AGCTGGCCCTCAACCAATC
Kat2a	GCATTTGCTTCCGCATGTTTC	GCCCTTGACCTGCTCATTTG
Psmc1	ACAAGGTGCATGCTGTGATAGG	GGGCCTTTTCCACCTTCATC
Sorbs1	AGTCAGTCCCCATCCGTGTTC	ACCAGCCTCTAGCAGCACTTG
Supt3h	GGAGGCCTCTCCATGAAACTG	TCCCCGAAGCTGAGAGACTTC
Supt7l	CGTGGCCTTGGTCCATTAATC	ACAGGTCGGAGTCGAGGACTAG
Tada1	TCTCTGCTCCATGTGCCAATC	GAGCAGGCCAATAGGAATGC
Tada3	AGCCATTAAGCAGTCCCACTTC	GACCCTCCCTGGAATGTTCTG
Taf5l	CCCCTTCAGCCTGTACTTTGC	CGGGTACGTCCGATCAAAAGAC
Taf9	TCTCCACCCCCGAGAGATTTC	ATGGCTTGATCAGTGGCAAAG

Taf10	CACCCTAACCATGGAGGACTTG	TGCGGCTTCTTCACATTGATG
Taf12	TGGCCCTTCAGCGCTAATC	TTGGCCATGGAGCCTTGTG
Tafl	TGGCACTAGCCACCAACATC	TGCCATCCAGATAGGAGACATG
Trrap	AGGAGTGACGGAAACGAAATG	GAGCGGCTGGGATCAGATG
Usp22	GCGAATTGCGATGGAATGAC	CGAGTGATTTTCACACCGATTG
Ccdc101	TCCAGGGATTCGCCTGTTC	GCATCGAGCACAGGAAGAAAC
Hist2h2be	CATGAAGCGATCTTTCGAATCC	CACGACAGGAGGGAAAAAGC
Sap130	CATTCCCGTGGCAACAATC	GGGAAGCTCCAATGTGAAGTG
Sf3b3	GTTTGCTCTGCCACCCATAAG	TATCTCCCTGCTCCGTTTGG
Tada2b	TGAGCCTTTCCCCGTTTG	GAAAAGCTGATGGCAGGAACAC
Taf9b	GCACTCTGCAGTTCCTGTGAAC	GAGACCCGAATGCTCTGTTAGG
Trp53	GCCCTCATAGGGTCCATATCC	GCAGACAGGCTTTGCAGAATG
Human qRT-PCR primers		
	Forward (5’ to 3’)	**Reverse (5’ to 3’)**
ATXN7	ATGGATGGGCCAGTTTGG	CCATACCCCATTGCACTTGTC
lnc-SCA7	GGCTTCTTGGACAGTTTGG	TGTCTCAGAGAAGCCCATA
GAPDH	AATCCCATCACCATCTTCCA	TGGACTCCACGACGTACTCA
MALAT1	TGAGTTGGAAACAGGGAAGATG	GGCCTCGAACTCAGAAATCC
ChIP qRT-PCR primers		
	Forward (5’ to 3’)	**Reverse (5’ to 3’)**
miR124_1a_ChIP	AAGACAGATTTCTCATCGCA	GACGTCCGAGATTACAGAGG
miR124_1b_ChIP	CACGCAACTAGAGGCGCGA	TCACTGCGGAGAGCGCGAGG
miR124_2a_ChIP	ACCCAGCCGCAAGGGATGCT	GCTGCTTGCTCGGGCTTCCT
miR124_2b_ChIP	CTCTTCGCAGTAAAATCATA	CTGTTGCCAGAGATCTAGCA
miR124_2c_ChIP	GGATGAGTGACAAAGGTTTC	TGCATACATCTGCATGCCAT
miR124_3a_ChIP	TGTGTTCGCTGCACAGCG	GTGCCTCCCGCGCACACGCT

miR124_3b_ChIP	TCGCGCGGTGCGTGAGTGCG	TCCGCGAGGTCGGCACTGCG
miR124_3c_ChIP	TCTCCTCGAGCAGCTTCTCG	GAGCGCGCGTGCTCCGGCCT
miR124_3d_ChIP	GTCCCTACTGCAGGAACGCC	GTCCGCTGTGAACACGCAGA
NC_ChIP_na	CAACTAGAGTAGGCTAACTG	TCAAATGTACCCCTGCTTAC
NC_ChIP_nb	CATGCATGTGGAGATCAGAA	ATAGACAACTGATGCACTAC
Riboprobe cloning oligos		
	Forward (5’ to 3’)	**Reverse (5’ to 3’)**
Atxn7	CATTCCGTGAACTCTTTTAGG	TGTCAAACAGCCAGAGGTTAC
lnc-SCA7	CTTCAGCTGTTTGAAAACCACC	GGACAATAATAGAGACTGGTC
Mutagenesis oligos		
	Forward (5’ to 3’)	**Reverse (5’ to 3’)**
lncSCA7_mut_STOP	GGAGGAGATTTAGTTGGCCAACCTG	CAGGTTGGCCAACTAAATCTCCTCC
Atxn7_MREmut_1	CAGCACTCTGGACTCCACGATTCCGTTGAGTCTCTTTTTCTAACTCCTG	CAGGAGTTAGAAAAAGAGACTCAACGGAATCGTGGAGTCCAGAGTGCTG
Atxn7_MREmut_2	GGATCATTTCTGTAGCCTTTCCGTCCTTTTTCTTTGCCATCTGTCAG	CTGACAGATGGCAAAGAAAAAGGACGGAAAGGCTACAGAAATGATCC
lncSCA7_MREmut_1	GCATTTTGTGGGGACAAGGCCTTTCCGTTGGCTCATCAGAAGACCTG	CAGGTCTTCTGATGAGCCAACGGAAAGGCCTTGTCCCCACAAAATGC
lncSCA7_MREmut_2	GGAGGGAACTTTTTTCCATCTTCCGTCAATAATGTGAATGAG	CTCATTCACATTATTGACGGAAGATGGAAAAAAGTTCCCTCC
lncSCA7_MREmut_3	GTGTTTCTCTTATATAAGGCTTCCGTGCTAGTATATCTCCCTGCCTG	CAGGCAGGGAGATATACTAGCACGGAAGCCTTATATAAGAGAAACAC
lncSCA7_MREmut_4	GAGCCTTGAGACTGTCTGTTGTTTCCGTCTCAGAGTTTTGCAGCTCAGG	CCTGAGCTGCAAAACTCTGAGACGGAAACAACAGACAGTCTCAAGGCTC
lncSCA7_MREmut_5	GGCCATGGCCCACTCACTCTTTCCGTCAGATAGGGAGCTCAAATG	CATTTGAGCTCCCTATCTGACGGAAAGAGTGAGTGGGCCATGGCC
lncSCA7_MREmut_6	GTGACTTTAGTTTTCTTTCCGTTTCTGGTTAGAACATAAAGATG	CATCTTTATGTTCTAACCAGAAACGGAAAGAAAACTAAAGTCAC

Table 2.5 Custom oligonucleotide sequences used in Chapter 6.

RT-PCR primers		
	Forward (5' to 3')	**Reverse (5' to 3')**
lnc-ASD	CTACAGGCATCAAGCACTAGATGCCAAGGCC	GGCCTTGGCATCTAGTGCTTGATGCCTGTAG
qRT-PCR primers		
	Forward (5' to 3')	**Reverse (5' to 3')**
MSN	CTACCACTGTCTTCTTCAG	TTCTAGTATGTTTGGTGAAGG
lnc-ASD	CTTCTTCTTCAGGGTCCTGAG	TTCTAGTACATTTGGTGAAGCC
MECP2	TGTTAGCAGTGGGTCATGATG	CGATTTCATGTCAGTCAGAA
NIPBL	CAAACTGGTGCCTTGGGTAGAC	CGGCATACAGGGATCGTAAAG
CDKL5	CTCCAAGCAGCAGACCAAAG	TGAGGCCGAAGAGAGATGTAAC
FGD1	AGCAGCAGAGCACTCGAATG	GCCCCCTAACAGCTCATATACC
AHI1	GCAGCAAATTATCGGGAGAAG	TCCTGTTTCTGGGTTCCAAAC
TSC2	ACGAGAGACCCAAGAGGATACAG	CATCATGTCCAGACAGGTTTCC
NSD1	GGTCCAGACCCTTGTAGCTAAAG	GGTGTGACCTGATGAGGTGAAG
CREBBP	CTGGGTGACGACCCTTCAC	TGTCATAGTGCAGAACGCAAATC
miR-1253	AGAGAAGAAGAUCAGCCUGCA	
GADPH	AATCCCATCACCATCTTCCA	TGGACTCCACGACGTACTCA
MALAT1	CAAGCAACTTCTCTGCCACATC	GACCTCGACACCATCGTTACC
RNA stability qRT-PCR primers (5' to 3')		
MYC	GGCCCCCAAGGTAGTTATCC	TTTCCGCAACAAGTCCTCTTC
ATP5E	CTGATCTTCCTGCGGCTGAAC	TTGCACAGATCTGGGAGTATCG
siRNA oligos (5' to 3')		
si-lnc-ASD_1	GAAACTCAATTAAGAAGGT	
si-lnc-ASD_2	GGCACTCACTTGGTGATAT	
si-lnc-ASD_3	GGCTGTCAATTAACCTAAA	
si-lnc-ASD_4	GCTAGAATATTCCACTTCT	

CHAPTER 3

A MicroRNA-mediated post-transcriptional regulatory role is conserved in unitary pseudogenes after loss of protein-coding capability in their orthologous ancestral mRNAs.

3.1 ABSTRACT

Recent reports have highlighted instances of coding and noncoding RNAs that are able to regulate the abundance of other transcripts by competing for the post-transcriptional regulation of common microRNAs (miRNAs). The significance of this RNA-mediated function and whether it represents a biologically important mechanism or is a mere consequence of the competition by different transcripts for promiscuous miRNA binding, remain unknown. This study investigates the importance of this post-transcriptional RNA-mediated mechanism using rodent-specific 'unitary pseudogenes' – noncoding transcripts that once encoded proteins in the earliest eutherian ancestor, but have since lost their protein-coding capability, specifically during rodent evolution. By not using protein-coding mRNAs in the investigation, the study avoids the difficulty of separating the likely interdependency between RNA- and protein-mediated roles of these coding transcripts. Specifically, thirty-five percent of the rodent-specific unitary pseudogene loci have retained their active transcription in mouse. These loci also exhibit tissue expression profiles in mouse that are conserved with their human ancestral orthologs and this conservation appears

to be associated with the conservation of their MREs that preserved the post-transcriptional roles of their protein-coding ancestors. I used mouse *Pbcas4*, a transcribed unitary pseudogene, to demonstrate the importance and conservation of its competitive endogenous RNA (ceRNA) role. *Pbcas4* loss-of-function led to the down-regulation of a significantly higher proportion of genes, whose orthologs in human were positively correlated with *BCAS4*, the human protein-coding ortholog of *Pbcas4*. In addition, using one conserved miRNA, *miR-185*, that targets *Pbcas4/BCAS4*, I showed this miRNA mediates the conserved crosstalk between *Pbcas4*/*BCAS4* and their target protein-coding mRNAs in mouse and human. Together with genome-wide predictions, these results demonstrate that some transcribed unitary pseudogenes have conserved the post-transcriptional roles of their protein-coding ancestors after the loss of their protein-coding potential, supporting the biological relevance of their miRNA-mediated post-transcriptional regulatory function.

3.2 INTRODUCTION

The level of transcribed gene products can be regulated in a spatiotemporal manner post-transcriptionally. One such post-transcriptional regulation involves the recognition and binding of mature miRNAs to miRNA response elements (MREs) in their targets, which often leads to mRNA degradation or translational inhibition (Ambros, 2003; Wienholds and Plasterk, 2005; Bartel, 2009). This type of post-transcriptional regulation mediated by miRNAs is largely preserved in animal evolution and is widespread among eukaryotes (Sempere et al.,

2006). Since a miRNA can regulate a large number of transcripts (Grun et al., 2005; Krek et al., 2005), and target recognition is thought to result in decreased levels of free miRNAs (Bartel, 2009), transcripts harbouring MREs for the same miRNA, may in principle affect each other's abundance by competing for the binding to their shared miRNAs (**Chapter 1**). Recently, a new layer of post-transcriptional expression regulation was revealed that involves competition among coding as well as noncoding transcripts for binding to specific miRNAs, where transcripts that engage in such post-transcriptional crosstalk have been termed 'competitive endogenous RNAs' (ceRNAs) (Salmena et al., 2011).

Although several protein-coding transcripts have been shown to act as ceRNAs (Poliseno et al., 2010; Tay et al., 2011), the protein-coding and miRNA-mediated roles of mRNAs are not independent because: (1) targeting of miRNAs to a transcript's MREs can result in decrease levels of its encoded protein and mRNA abundance, and (2) levels of mRNAs, in turn, are able to regulate the levels of other transcripts through competition for miRNAs (Poliseno et al., 2010; Sumazin et al., 2011; Tay et al., 2011). It is this coupling between RNA- and protein-dependent functions of a transcript that renders the biological importance and implications of ceRNAs very difficult to determine using protein-coding ceRNAs. As a result, it remains unclear whether a transcript's MRE(s) might be sufficiently important for its miRNA decoy function to act autonomously of its protein-coding capability; for example, by conferring robustness to transcriptional networks or by buffering genetic noise (Wu et al., 2009).

This work set out to investigate the importance and contribution of this conserved, protein-coding independent, post-transcriptional regulatory mechanism to animal transcriptional regulation. Previous works have only focused on the post-transcriptional roles of transcribed pseudogenes that share MREs with their duplicated homologous transcripts (reviewed in Muro et al., 2011 and Pink et al., 2011). *PTENP1*, for example, is a processed (retroduplicated) pseudogene that acts as a ceRNA by modulating the expression level of its parental gene, *PTEN*, a known tumour suppressor, with which it shares several predicted MREs (He and Hannon, 2004).

To separate the RNA-mediated from protein-mediated roles of RNA transcripts, this study focused on protein-coding genes in human and mouse that lost their protein-coding capabilities in the rodent lineage (i.e. mice and rats), suggesting that the mammalian ancestor of these genes were protein-coding. These loci are termed unitary pseudogenes (Figure 3.1). In contrast to the high number of duplicated and retroduplicated pseudogenes (of which *PTENP1* is an example) in mammalian genomes (Jacq et al., 1977; Vanin et al., 1980), unitary pseudogenes are rarer and derived from the lineage-specific acquisition of disrupting mutations in the coding sequences of genes (Zheng et al., 2005; Karro et al., 2007; Zhu et al., 2007; Zhang et al., 2010).

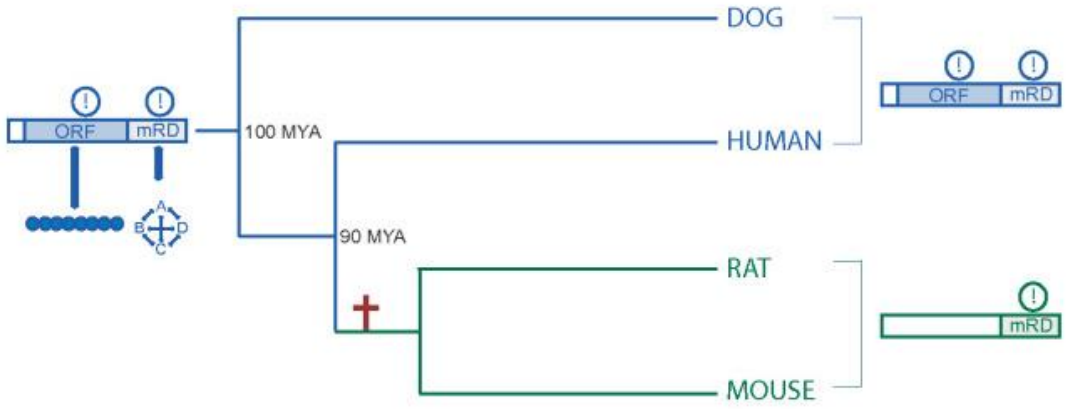


Figure 3.1 Identification of rodent-specific unitary pseudogenes. An ancestral transcript encoding a functional (shown by the exclamation mark) protein (vertical lines indicate an intact open reading frame (ORF)) and with a role as a miRNA decoy (mRD) has accumulated one or more mutations disabling its ORF (red cross) on the lineage leading to rodents. As a result the mouse or rat unitary pseudogene (green) has lost its coding potential while retaining its function as a competitive endogenous RNA while both functions are conserved in the protein-coding loci (blue) in human or dog. MYA (million years ago). Figure from Marques et al., 2012.

The loss of open reading frames (ORFs) implies that such rodent-specific unitary pseudogenes no longer encode a functional protein and that, if transcribed, their conserved expression is independent of their ancestral coding functions. This allows the dissection of the miRNA-mediated roles of the transcribed unitary pseudogenes from their ancestral protein-coding functions (Figure 3.1). In particular, it is possible to consider whether these noncoding roles have been conserved between humans and rodents since they last shared a common protein-coding ancestor approximately 90 million years ago. Conservation of ancestral post-transcriptional miRNA decoy functions would imply that the miRNA-mediated interactions between transcripts are biologically relevant, indicating the importance of this post-transcriptional regulatory mechanism to animal transcriptional regulation.

3.3 MATERIALS and METHODS

I performed all the work described below, except where noted otherwise.

Annotation of unitary pseudogenes in mouse

Annotation of rodent-specific unitary pseudogenes in mouse, mammalian protein-coding genes that were lost specifically in the rodent lineage, was performed by Dr. Ana Marques with assistance from Dr. Andreas Heger and Dr. Lesheng Kong (Figure 3.2). First, Transmap (Zhu et al., 2007) annotations of human protein-coding gene transcripts with orthologs in the mouse, rat and dog genomes were downloaded from UCSC (Dreszer et al., 2012), and protein-coding genes not annotated in EMSEMBL build 67 (Flicek et al., 2012) were removed to avoid potential mis-annotations. Next, the remaining human genes were aligned to their conserved syntenic regions (Zhu et al., 2007) in mouse and those with no overlap (by 1 base pair or more) to annotated mouse protein-coding gene were obtained (758).

Next, exhaustive 6-frame translated pairwise alignments were performed between the 758 human polypeptide templates and their regions of conserved synteny (extended by 5 kb upstream and downstream) in the mouse, rat and dog (used as an out-group) genomes using Exonerate (Slater and Birney, 2005). Sequences corresponding to the best alignments, alignments that contained at least 80% of the human template sequence and had partial or complete conservation of exon structures (Heger and Ponting, 2007), were classified into two groups: (1) sequences with conserved protein-coding

potential annotated in the syntenic regions in dog, but lost in regions of mouse and rat (48 unitary pseudogenes)(Marques et al., 2012) were catalogued as unitary pseudogenes, (2) otherwise, they were categorized as a conserved protein-coding gene (Heger and Ponting, 2007) (Figure 3.2). Unitary pseudogene predictions were visually inspected to ensure that: 1) frame-shifting indels or premature stop codon mutations were specific to both rodents; and 2) chromosomal gene order for genes immediately upstream and downstream of the lineage-specific unitary pseudogene was conserved in all four species (Schwartz et al., 2003; Fujita et al., 2011).

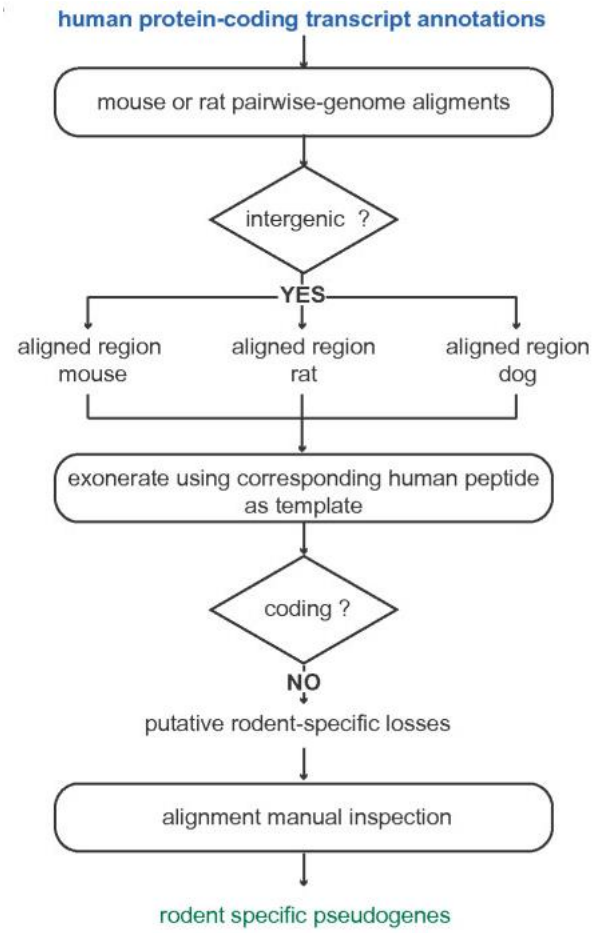


Figure 3.2 Flowchart for identification of rodent-specific unitary pseudogenes. Flowchart illustrating how the rodent-specific unitary pseudogenes were identified that have lost their protein-coding capability from their human orthologs. Figure produced by Dr. Ana Marques and taken from Marques et al., 2012.

Expression of mouse unitary pseudogenes

Transcript assembly of the transcribed mouse unitary pseudogenes was performed by Dr. Ana Marques. Publicly available polyA-selected RNA single-end sequencing reads for mouse and human adult tissues (testis, liver, heart, kidney, brain and cerebellum) (Brawand et al., 2011) were used to determine the presence of rodent-specific unitary pseudogenes in the mouse. For each tissue, reads were aligned using Tophat (Trapnell et al., 2009) and transcripts across the predicted genomic location of unitary pseudogenes were assembled *de novo* using Cufflinks (Trapnell et al., 2010). Unitary pseudogenes whose predicted genomic location (Zhu et al., 2007) overlapped at least one sequencing read in at least one tissue were considered to be expressed. The expression of unitary pseudogenes was confirmed in mouse using an independent data set of stranded paired-end polyA-selected RNA sequencing reads across 19 available mouse tissues and cell lines (Shen et al., 2012) by following the same steps as described above. Finally, a reference annotation was produced by combining the transcripts assembled in the different tissues using Cuffcompare (Trapnell et al., 2010).

Estimation of protein-coding potential

The lack of protein-coding capability of the *de novo* assembled transcripts (97 transcripts longer than 200 nt in length) of the 17 unitary pseudogenes considered to be expressed (Brawand et al., 2011) was assessed using two independent methods, PhyloCSF (Lin et al., 2011) and Coding Potential

Calculator (CPC) (Kong et al., 2007). PhyloCSF is a method used to determine the protein-coding potential of a region by using multi-species nucleotide sequence alignment of that region and by assessing the frequency of synonymous codon substitutions, conservative amino acid substitutions, and other missense and non-sense substitutions (Lin et al., 2011). The Coding Potential Calculator (CPC) (Kong et al., 2007) considers the length of all putative open reading frames (ORFs) and the homology of their encoded protein to known mammalian peptides, for example, if the mouse transcripts contain putative ORFs that are homologous but incomplete, due to the accumulation of deleterious mutations, to their protein-coding orthologs in humans.

Gene expression correlation

Dr. Ana Marques estimated gene expression levels (measured as total number of fragments per kilobase of sequence per million reads mapped, FPKM) across the six adult mouse and human tissues (Brawand et al., 2011) of unitary transcribed pseudogenes. For human and mouse gene annotations, the total numbers of reads that overlap protein-coding gene constitutive exons (ENSEMBL build 67) or expressed nucleotides within *de novo* annotated unitary pseudogene transcripts were independently normalized using TMM (edgeR package) (Robinson et al., 2010). To estimate the normalized library size for each species, 60% of expressed genes were considered to be transcribed at similar levels across tissues. For each tissue in human and mouse, the normalized mouse and human library size was used to calculate the expression level (measured as FPKM) at each locus.

The Pearson correlation coefficient was computed between the expression levels of each pair of mouse unitary pseudogene and orthologous protein-coding gene in humans across the six adult tissues considered. This correlation was then compared to the correlation in expression between non-orthologous pairs of randomly selected protein-coding genes and pairs of mouse-human one-to-one orthologs (ENSEMBL build 67). Only mouse protein-coding genes that have one-to-one orthologous relationship with human genes were considered in these calculations.

5' and 3' RACE

5’ and 3’ randomly amplification of cDNA ends (RACE) was performed by Dr. Ana Marques. Total RNA from mouse neuroblastoma (N2A) cells was extracted using the RNAeasy kit (Qiagen) followed by DNAse treatment with the DNA-free kit (Ambion), according to the manufacturer's instructions. cDNA was prepared using a RACE ready cDNA kit (Clontech). PCR amplifications were carried out using primers specific to the 5' and 3' ends of the transcript and 5' and 3' RACE outer primers provided by the manufacturer. PCR reaction products were further amplified using nested sequence primers and 5' and 3' RACE inner primers. The resulting product was purified using PCR cleanup kit (Qiagen), cloned into a TOPO vector (Invitrogen) and sequenced.

Transcriptome wide analysis of Pbcas4 knockdown

Short hairpin RNAs (shRNAs) specific to mouse *Pbcas4* were designed as described in **Chapter 2**. Mouse neuroblastoma (N2A) and human

neuroblastoma (SH-SY5Y) cells were used in this chapter. Cells were grown as described in **Chapter 2**. Transient transfection of 4 µg of shRNA constructs and scrambled controls were carried out using Lipofectamine Plus (Invitrogen) in triplicate. Cells were harvested 72 h post-transfection and their RNAs were extracted, reverse transcribed and assayed for expression qualification as described in **Chapter 2**. These extracted RNA samples were used for mRNA expression profiling using microarray as described below. Specifically, transfections of *Pbcas4-shRNA* led to a reproducible 50% decrease in expression of this unitary pseudogene in N2A cells. Oligos used in the experiments are listed in Table 2.1.

mRNA expression profiling using Affymetrix microarray was performed by Sheena Lee and subsequent analysis was carried out by Dr. Ana Marques. RNA integrity was assessed on a BioAnalyzer; all samples had an RNA Integrity Number (RIN) ≥ 7 (Agilent Laboratories, US). Sense single-stranded DNA was generated from 200 ng starting RNA with the Ambion WT Expression Kit according to the manufacturer's instructions and fragmented and labeled using the GeneChip® WT Terminal Labeling and Controls Kit. The distribution of fragment lengths was measured on a BioAnalyser. The labeled single-stranded DNA was hybridized to the Affymetrix Mouse Gene 1.0 ST Array (Affymetrix). Chips were processed on an Affymetrix GeneChip Fluidics Station 450 and Scanner 3000. CEL files were generated using Command Console (Affymetrix). Normalized and background corrected expression values were produced using the RMA method from the *affy* bioconductor package. *Limma*, from the bioconductor package, was then used to identify differentially expressed genes (Benjamini-Hochberg corrected $p<0.05$) between *Pbcas4* and

scrambled vector transfected cells. Only probes where variance between conditions exceeded 0.5 were considered.

These data are accessible through Gene Expression Omnibus accession number GSE38333.

Validation of post-transcriptional regulation by miR-185

Mouse and human neuroblastoma cells, N2A and SH-SY5Y, respectively, were prepared 24 h before transfection as described in **Chapter 2**. *miR-185* mimics and negative control miRNA mimic (50 nM; Applied Biosystems) were transfected using Lipofectamine RNAiMAX Reagent (Invitrogen). Cells were harvested 24 h post-transfection and RNA was extracted as previously described. Mature *miR-185* was reversed transcribed and quantified, following the manufacturer's instructions, using the TaqMan MicroRNA Reverse Transcription Kit and Taqman MicroRNA Assays (Applied Biosystems). Expression level of *miR-185* was normalized to *18S rRNA*. miRNAs were reverse transcribed and their expression levels were detected using quantitative real-time PCR (qRT-PCR) as described in **Chapter 2**.

Construct designs are illustrated in Table 2.1. Primers used in Chapter 3 are listed in Table 2.2.

3.4 RESULTS

3.4.1 A stringent catalogue of rodent-specific unitary pseudogenes

A set of 48 rodent-specific unitary pseudogenes, mammalian protein-coding genes whose protein-coding capability was lost specifically in the rodent lineage (i.e. mouse and rat) were identified. Briefly, Transmap (Zhu et al., 2007) annotations of human protein-coding gene transcripts with orthologs in all of mouse, rat and dog genomes were aligned to their conserved syntenic regions (Zhu et al., 2007) in mouse, and those with no overlap (by 1 base pair or more) to annotated mouse protein-coding genes were obtained (758). Next, Exonerate (Slater and Birney, 2005) was used to perform exhaustive 6-frame translated pairwise alignments between the 758 human polypeptide templates and their regions of conserved synteny (extended by 5 kb upstream and downstream) in the mouse, rat and dog (used as an out-group) genomes. Sequences corresponding to the best alignment (Heger and Ponting, 2007) were classified into two groups: (1) sequences with conserved protein-coding potential annotated in the syntenic regions in dog, but lost in regions of mouse and rat (48 unitary pseudogenes) (Marques et al., 2012) were catalogued as unitary pseudogenes, (2) otherwise, they were categorized as a conserved gene (Heger and Ponting, 2007) (Figure A3.1).

The expression of these rodent-specific unitary pseudogenes in mouse was determined using publicly available RNA-seq data across six adult mouse tissues (testis, liver, heart, kidney, brain and cerebellum) (Brawand et al., 2011). 17 of the 48 (35%) unitary pseudogenes showed evidence of

expression, where at least one overlapping RNA sequencing read mapped within their syntenic genomic locations in at least one mouse tissue (Table 3.1). The expression of these 17 unitary pseudogenes was supported by another RNA-seq data set of 19 mouse tissues and cell lines (Shen et al., 2012).

Subsequently, the lack of protein-coding capacity of these 17 unitary pseudogenes was confirmed by estimating the pairwise codon substitution frequencies of their transcripts (49) between mouse and rat using PhyloCSF (Lin et al., 2011) and by calculating their coding potentials using Coding Potential Calculator (CPC) (Kong et al., 2007). Findings from both tests indicated the 17 transcribed unitary pseudogenes are unlikely to have retained their protein-coding capacity.

Using PhyloCSF, only 3 out of the 97 transcripts were annotated as coding (94 annotated as noncoding) and the median PhyloCSF score between mouse and rat for these unitary pseudogene transcripts was -16.1, which is smaller than zero, as expected for noncoding regions (Lin et al., 2011). This is also substantially smaller than the corresponding score (25.4) for 1,000 randomly selected protein-coding gene transcript fragments matched in size to the unitary pseudogene transcripts ($p<10^{-16}$, two-tailed Mann-Whitney test; Figure 3.3). Furthermore, using the CPC (Kong et al., 2007), which considers the length of all putative open reading frames (ORFs) and the homology of their encoded protein to known mammalian peptides, 97% (94/97) of these transcripts contained putative ORFs that are homologous but incomplete, due to the accumulation of deleterious mutations, to their protein-coding orthologs in

Table 3.1 Genomic locations of the rodent-specific unitary pseudogenes in mouse (mm9).

ENSEMBL protein ID (build 67)	Gene	Genomic location (mm9)	Length	Number of sequencing reads	
				Brawand et al., 2011	Shen et al., 2012
ENSG00000161692	*DBF4B*	chr11: 102588988-102617359	1284	259	112
ENSG00000182952	*HMGN4*	chr13:23534848-23536620	499	210	156
ENSG00000177383	*MAGEF1*	chr16:21331982-21333435	1399	716	739
ENSG00000162086	*ZNF75A*	chr16:3761999-3777242	680	152	475
ENSG00000125731	*SH2D3A*	chr17:57403584-57414946	467	61	88
ENSG00000164296	*TIGD6*	chr18:61336748-61346946	182	64	33
ENSG00000124243	*BCAS4*	chr2:167942839-167998429	421	139	120
ENSG00000187609	*EXD3*	chr2:24950970-25031762	732	294	293
ENSG00000214402	*LCNL1*	chr2:25319909-25321240	625	217	6
ENSG00000145428	*RNF175*	chr3:83593293-83636364	190	103	27
ENSG00000126709	*IFI6*	chr4: 132494484-132497171	104	77	13
ENSG00000168152	*THAP9*	chr5:100858627-100865170	351	86	90
ENSG00000163257	*DCAF16*	chr5:46046524- 46060951	356	145	66
ENSG00000236287	*ZBED5*	chr7:118261148-118266213	673	181	306
ENSG00000171425	*ZNF581*	chr7:5005473-5006970	322	144	59
ENSG00000196653	*ZNF502*	chr9:122847141-122849738	454	127	87
ENSG00000076662	*ICAM3*	chr9:20897783-20905203	830	138	359

humans. Additionally, the fraction of unitary pseudogene transcripts annotated as coding is >20 times smaller than that found for 1,000 randomly selected protein-coding transcript fragments with matching size, a highly significant difference (656/1,000, $p<10^{-4}$, two-tailed Fisher's exact test).

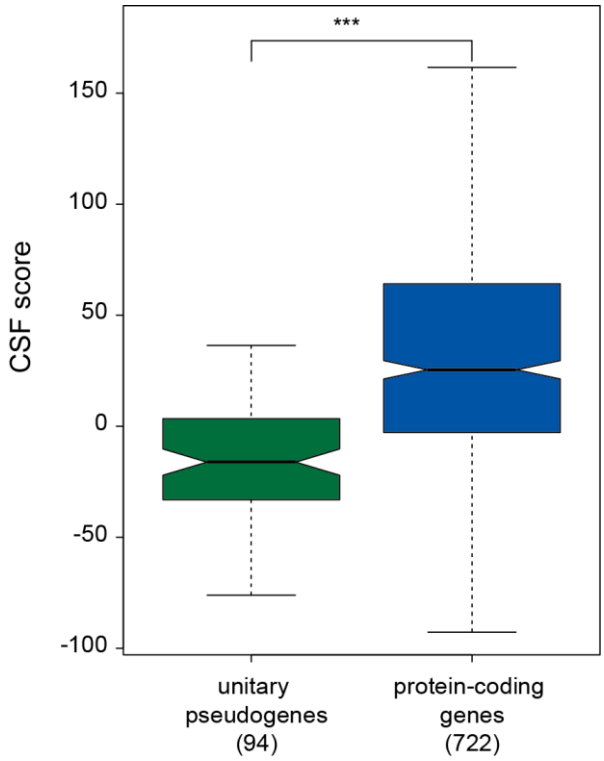


Figure 3.3 Codon substitution pattern of unitary pseudogenes. The coding substitution pattern of the unitary pseudogene (green) is significantly smaller (*** $p<0.001$) than that of protein-coding transcript fragments (blue) with matching size. Only transcripts with a sequence allowing reliable prediction of an open reading frame (94 and 722 unitary pseudogenes and protein-coding transcripts, respectively) were considered. Figure produced by Dr. Ana Marques and taken from Marques et al., 2012.

3.4.2 miRNA decoy functions are preserved in unitary pseudogenes after their loss of protein-coding potential

If a transcribed mouse unitary pseudogene has maintained the decoy roles of its protein-coding ancestral transcript, one would expect its expression profile and its expression correlation with crosstalking targets to be conserved and shared with its protein-coding human ortholog. First, gene expression levels were measured (as total number of fragments per kilobase of sequence per million reads mapped, FPKM) across the six adult mouse and human tissues (Brawand et al., 2011) of unitary transcribed pseudogenes (see section 3.3 Materials and Methods, Table 3.2). The Pearson correlation coefficient was computed between the expression levels of each mouse unitary pseudogene and orthologous protein-coding gene in humans across the six adult tissues considered. As expected for transcripts with preserved decoy roles, these 17 mouse unitary pseudogene-human gene ortholog pairs were found to be more highly correlated (median Pearson correlation coefficient=0.20) than randomly sampled pairs of protein-coding genes (median Pearson correlation coefficient=0; Figure 3.4) in their expression profiles, where the randomly selected genes are length and G+C content matched to the unitary pseudogenes. This implies the relative expression levels of these unitary pseudogene transcripts across the adult tissues tested were preserved, at least in part, after their loss of protein-coding capability.

Table 3.2 Expression level (FPKM) of transcribed unitary pseudogenes.

Name	Brain	Cerebellum	Heart	Kidney	Liver	Testis	median expression (FPKM)
BCAS4	0.75	0.43	0.54	0.58	0.60	0.35	0.56
DBF4B	0.54	0.47	0.48	0.038	0.053	0.035	0.26
DCAF16	0.71	0.37	0.08	0.29	0.77	1.13	0.54
EXD3	0.69	0.54	0.55	1.18	0.77	1.00	0.73
HMGN4	0.85	0.68	0.66	0.48	0.57	0.85	0.67
ICAM3	0.64	0.49	0.35	0.38	0.60	0.65	0.54
IFI6	0	0	0.79	0	0	0	0.79
LCNL1	0.33	0.35	0.31	0.15	0	0	0.32
MAGEF1	2.04	1.64	1.16	0.17	0.26	0.73	0.95
RNF175	0.62	1.19	0.11	0	0	0	0.62
SH2D3A	0.30	0.079	0.28	0.28	0.42	0.47	0.29
THAP9	0.39	0.58	0.36	0.35	0.52	0.53	0.46
TIGD6	0.29	0.81	0.72	0.31	0.32	0.65	0.48
ZBED5	0.61	0.53	0.83	0.60	0.71	1.03	0.66
ZNF502	0.48	0.37	0.49	0.078	0	0	0.43
ZNF581	0.58	0.65	0.59	0.11	0.16	0	0.58
ZNF75A	0.24	0.86	0.84	0.47	0	0.94	0.84

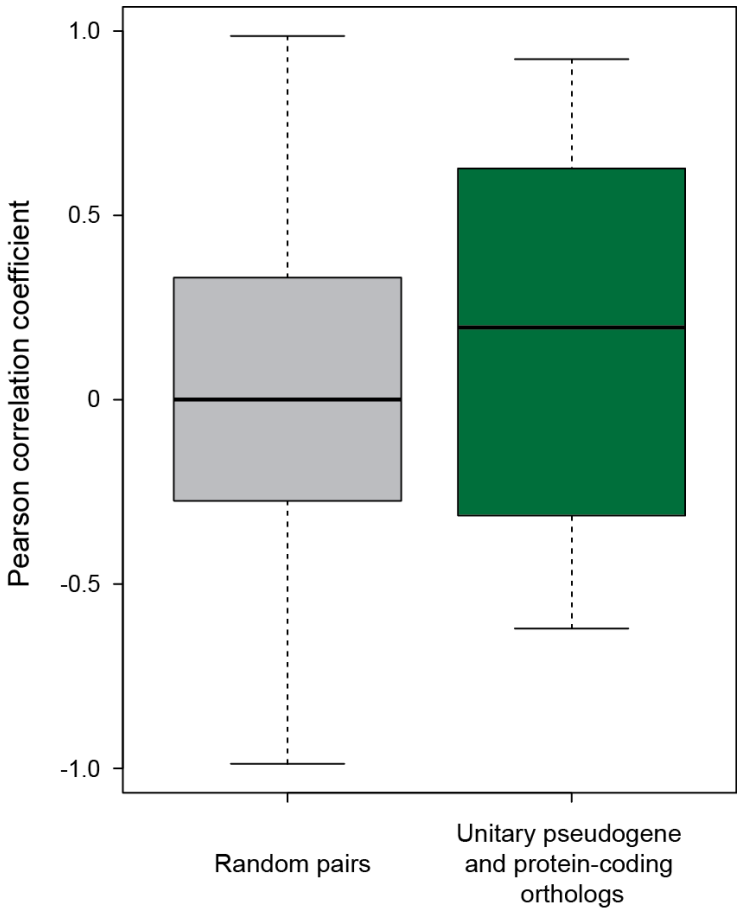


Figure 3.4 Tissue expression correlation between mouse and human loci. Distribution of mouse-human expression correlation (Pearson) between 1,000 mouse-human random pairs of non-orthologous protein-coding genes (grey) and mouse unitary pseudogene protein-coding orthologs (green). The *p*-value associated with the comparison between these distributions is 0.23. Figure produced by Dr. Ana Marques and taken from Marques et al., 2012.

Subsequently, the conservation of post-transcriptional regulatory networks involving the mouse unitary pseudogene or its human protein-coding gene orthologs was tested by asking whether a gene pair whose tissue expression values are significantly positively correlated in mouse also have significantly positively correlated expression in human. For each mouse gene, M_i, all mouse genes, m, whose expression was significantly correlated (empirical $p<0.05$) with M_i were identified. Similarly identified were the human one-to-one

orthologous gene, H_i, of mouse M_i, all human genes, *h*, whose expression was significantly correlated with H_i's expression values. Finally, the fraction, f_i, of all mouse genes in set *m* that have human one-to-one orthologs in set *h* with positively correlated expression levels was calculated (Figure 3.5A). When M_i and H_i are an orthologous pair of conserved protein-coding genes, the median fraction *fi* of *hi* with *mi* is 5.5% (Figure 3.5A). In contrast, when M_i is a mouse transcribed unitary pseudogene and H_i is its orthologous protein-coding gene, the median fraction f_i was 1.0% ($p<0.05$, two-tailed Mann-Whitney test; Figure 3.5B). When mouse M_i and human H_i genes were randomly paired, the median fraction f_i was significantly smaller (median=0, $p<4\times10^{-8}$, two-tailed Mann-Whitney test; Figure 3.5B).

This analysis revealed that a significant proportion of mouse unitary pseudogenes that are positively coexpressed with protein-coding genes in mouse are also positively correlated in expression between the human orthologs of the mouse unitary pseudogenes and their coexpressed protein-coding genes in human. This analysis provided evidence of the conservation of the coexpression networks involving orthologous mouse unitary pseudogene/human gene and their positively coexpressed protein-coding genes. It is hypothesized that this conservation reflects, at least in part, the preservation of post-transcriptional networks involving these rodent-specific unitary pseudogene transcripts after the loss of their ancestral protein-coding capability, likely due to the preservation of their ancestral regulatory elements.

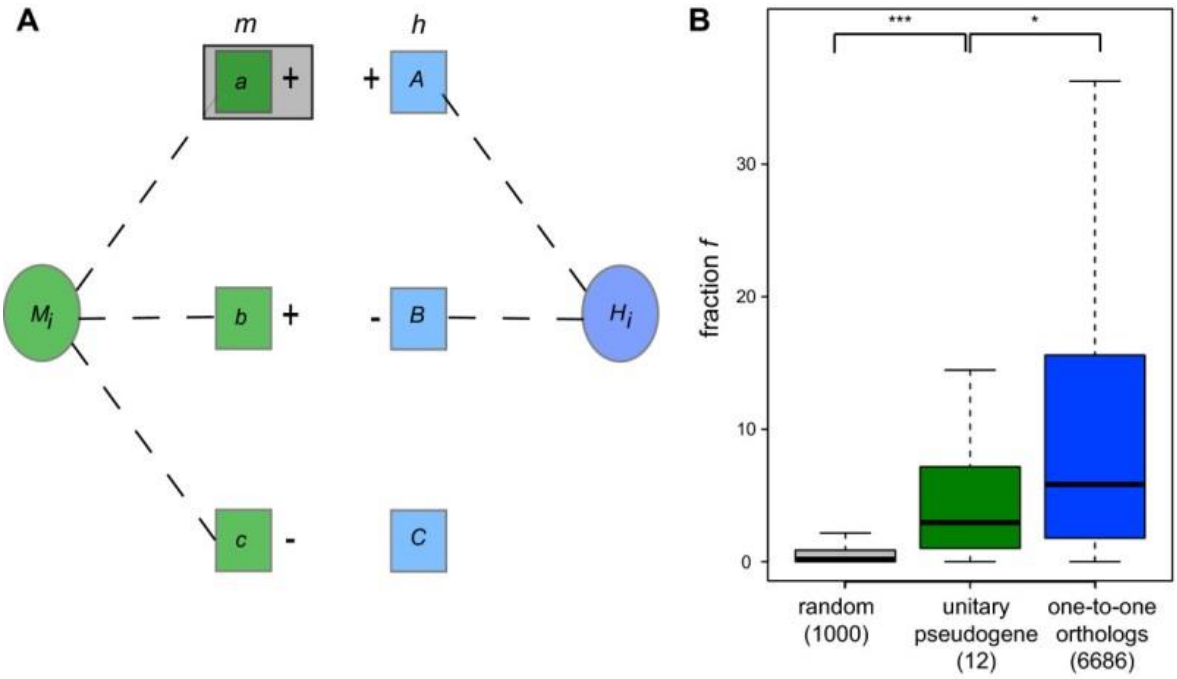


Figure 3.5 Conserved genetic interactions of transcribed unitary pseudogenes. (A) Correlation in expression was estimated between orthologous mouse (green) or human (blue) loci (*Mi* and *Hi*, respectively, ovals) and other protein-coding genes (rectangles) annotated in mouse or human genomes. Only genes exhibiting significantly (empirical $p<0.05$) positive (+) and negative (-) correlation with *Mi* or *Hi* tissue expression (dashed line) were retained. All mouse genes (for example, 'a' boxed) whose expression profiles are significantly correlated with that of *Mi*, and whose human one-to-one ortholog (for example, 'A') is also significantly correlated in the same direction with *Hi* were identified. (B) *f* is the fraction of mouse genes, *m*, that are significantly correlated in expression with a unitary pseudogene, *Mi*, whose human orthologs are also significantly correlated in expression (and in the same direction, either positively or negatively) with *Hi*, the human ortholog of *Mi*. *f* for human protein-coding gene and mouse unitary pseudogene pairs (green) is significantly higher (*** $p<0.001$) than for random non-orthologous mouse and human gene pairs (grey) and significantly lower (* $p<0.05$) for mouse and human one-to-one orthologs. No constitutively expressed exons (required to measure gene expression) were identified for *EXD3*, *THAP9*, *RNF175*, *DBF4B* and *ZBED*. Hence these genes were not considered in this analysis. Figure produced by Dr. Ana Marques and taken from Marques et al., 2012.

If two transcripts regulate each other's expression post-transcriptionally by competing for miRNA binding, one would expect their expression to be positively, rather than negatively, correlated (Schwartz et al., 2003). Of the 19,703 identified mouse unitary pseudogene:mouse gene pairs whose expression profiles are positively correlated, 1,340 (6.8%) have human orthologs whose expression profiles are also positively correlated (hereafter termed conserved and positively correlated quartets). In contrast, of the 13,579 negatively correlated pairs, a significantly lower proportion (607/13,579, 4.4%, $p<10^{-4}$, two-tailed Chi-square test) have human orthologs whose expression profiles are also negatively correlated. This higher level of preserved positive correlation is consistent with these transcripts forming parts of conserved ceRNA networks.

Next, the contribution of conservation of post-transcriptional networks by the preservation of orthologous miRNAs and their cognate MREs in mouse and human orthologous 3' untranslated regions (3' UTRs) was investigated. MREs predicted in mouse unitary pseudogenes (M_i) that were shared with protein-coding genes with which they had correlated expression were identified. To do this, 1,340 quartets of mouse and human loci that contain one of the mouse unitary pseudogenes (M_i) and its human protein-coding ortholog (H_i), and a pair of mouse and human orthologous genes that are each positively correlated in expression profile with M_i or H_i were considered (Figure 3.5A). For 22% of these conserved positively correlated quartets, at least one MRE predicted in M_i, the mouse unitary pseudogene, was also predicted in the 3' UTRs of each of the three other genes in the quartet. This was a significantly higher fraction (1.3 to 6.1%, $p<10^{-3}$, two-tailed Fisher's exact test; Figure 3.6) than found for

gene quartets associated either with significant negative expression correlations or with significant correlations that were in opposing directions in mouse and human (Figure 3.6).

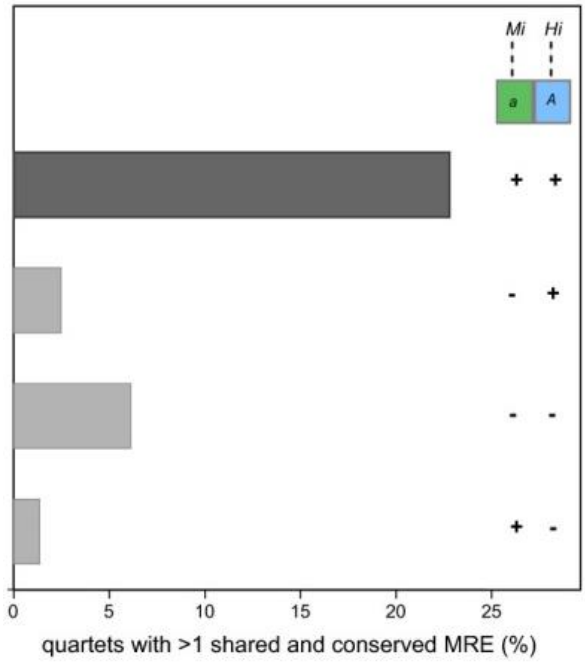


Figure 3.6 Conserved genetic interactions of transcribed unitary pseudogenes. Twenty-two percent of conserved significantly positively (++) correlated quartets (*Mi*-A-a-*Hi*, dark grey) shared at least one MRE for the same miRNA family across the four loci. This was a significantly higher fraction than found for quartets that are significantly negatively correlated (--) or significantly correlated (+- and -+) in different directions (light grey). Figure produced by Dr. Ana Marques and taken from Marques et al., 2012.

Next, the potential that unitary pseudogenes may be part of competitive endogenous RNA (ceRNA) networks that are conserved between mouse and human was investigated. For this to be true, some of the MREs predicted within mouse unitary pseudogene transcripts would be expected to also be found within the 3’ UTR of their human orthologous protein-coding genes. Consistent with this hypothesis, almost a fifth (17%) of MREs predicted within mouse unitary pseudogene transcripts were also identified within the 3' UTR of their human protein-coding orthologs. This value is significantly higher than expected in comparison to the common MREs predicted within 1,000 randomly selected pairs of mouse and protein-coding non-orthologous 3’ UTRs that were matched to the length and G+C content of the unitary pseudogenes (0, $p<10^{-4}$, two-tailed Mann-Whitney test) (Figure 3.7).

In addition, the conservation of post-transcriptional networks involving these rodent-specific unitary pseudogenes that serve as ceRNAs would imply similar MREs to be identified within the mouse unitary pseudogenes and their human orthologs, as well as within the 3’ UTRs of their positively expression correlated genes in mouse and human, respectively. Indeed, at least one predicted MRE within mouse unitary pseudogenes was also predicted within the 3’ UTRs of mouse genes that are positively correlated with the unitary pseudogenes in expression. The same was identified within the 3’ UTRs of their conserved orthologous human protein-coding gene pairs.

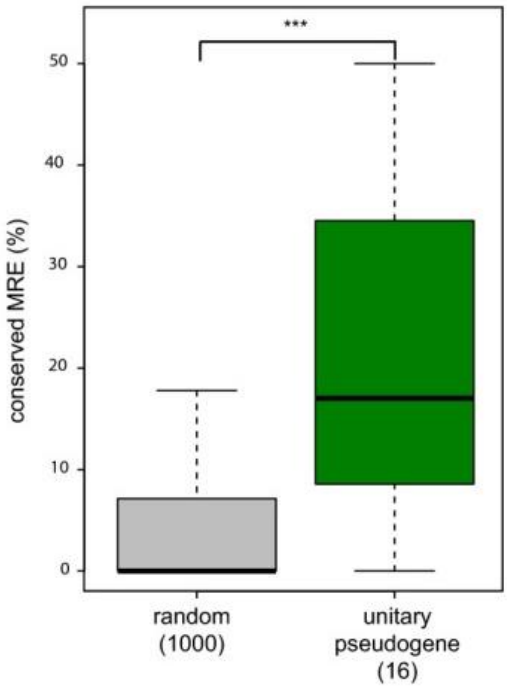


Figure 3.7 Conserved genetic interactions of transcribed unitary pseudogenes. Unitary pseudogenes (green) share significantly (*** p<0.001) more MREs with their human protein-coding ortholog 3' UTR than random pairs of mouse and human protein-coding genes (grey). MRE predictions were not available for *ZBED5* and this locus was not considered in this analysis. Figure produced by Dr. Ana Marques and taken from Marques et al., 2012.

Taken together, these results suggest that some transcribed rodent-specific unitary pseudogenes may have conserved their protein-coding ancestors' post-transcriptional roles and networks by acting as competitive endogenous RNAs.

3.4.3 BCAS4 pseudogene, Pbcas4, is a conserved competitive endogenous RNA

The computational analysis predicted that mouse transcribed unitary pseudogenes can serve as ceRNAs and that this post-transcriptional regulatory function is ancestral and shared with their orthologous human protein-coding genes. To provide support for this prediction, I chose one of the 17 mouse unitary pseudogenes, *Pbcas4*, based on its ubiquitous and relatively high expression in mouse adult tissues, for further investigation (Table 3.2).

Mouse *Pbcas4* is the transcribed unitary pseudogene of human *BCAS4* (Figure 3.8), which has protein-coding orthologs conserved from diptera to early branching vertebrates. The full-length transcript of mouse *Pbcas4*, as determined using rapid amplification of cDNA ends (RACE) in mouse neuroblastoma (N2A) cells corresponds only to the 3' UTR sequence of human *BCAS4* (Figure 3.8).

To investigate the transcriptome-wide effect of reducing the abundance of *Pbcas4* transcripts in N2A cells, short hairpin sequences (shRNAs) were designed and cloned to specifically target *Pbcas4*. Upon knockdown of *Pbcas4* by nearly 50%, microarray technology was used to profile transcript expression changes. Decreased levels of *Pbcas4* led to the differential expression of 165 genes, of which a significant majority (96) were down-regulated ($p<0.05$, binomial test).

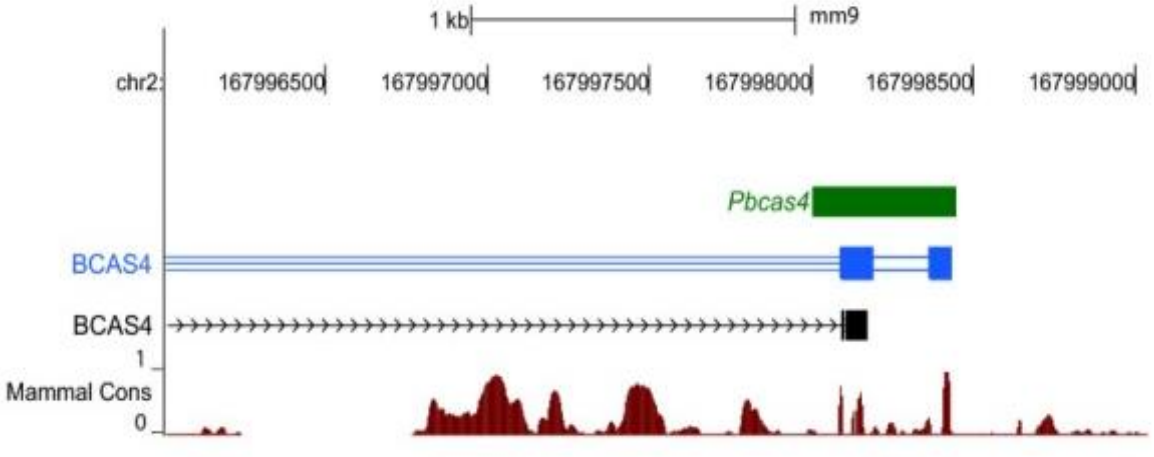


Figure 3.8 Genomic location of *Pbcas4*. Genome browser view of the transmap annotation (blue track) for the human breast carcinoma amplified sequence 4 (*BCAS4*) unitary pseudogene in mouse, *pbcas4*. The tBLASTn alignment available from UCSC between the BCAS4 peptide and the mouse genome is in grey. The full length transcript in N2A (green) is transcribed from chr2:167,998,005-167,998,447 (mm9) and aligns to the 3' UTR region of the human *BCAS4* gene. The mammalian conservation track on the bottom (UCSC genome browser) shows the degree of placental mammal base pair conservation (20 species). Figure produced by Dr. Ana Marques and taken from Marques et al., 2012.

If *Pbcas4* has a conserved function as a ceRNA, human orthologs of 57 protein-coding genes (those with one-to-one orthologs in human within the 96 down-regulated protein-coding genes) would be expected to exhibit positively correlated gene expression with *BCAS4*, *Pbcas4*'s human protein-coding ortholog. Indeed, the human orthologs of 41 (of 57, 72%) down-regulated mouse genes upon *Pbcas4* loss-of-function were positively correlated with *BCAS4*, whereas only 28 genes would be expected by chance (46% increase; $p<10^{-4}$, binomial test).

This finding that human orthologs of mouse protein-coding genes down-regulated upon mouse *Pbcas4* loss-of-function are significantly correlated in expression with human *BCAS4* suggests the ancestral post-transcriptional networks of *Pbcas4* may be preserved due to a miRNA-mediated mechanism which crosstalks between the unitary pseudogene and its ceRNA target genes. To test this hypothesis, miRNA response elements (MREs) of conserved miRNAs between mouse and human were predicted within *Pbcas4* and the 3' UTRs of gene differentially affected by *Pbcas4* loss-of-function using TargetScan (Garcia et al., 2011). Of the 12 MREs predicted in the full-length mouse *Pbcas4* transcript, two MREs, for *miR-185/882* and *miR-665*, are also predicted within the 3' UTR of its human orthologous *BCAS4*. Mouse genes containing predicted MREs for either *miR-185/882* or *miR-665*, where those two miRNAs are also predicted within their human orthologs, are nearly twice as likely (1.7-fold increase, $p<0.02$, Fisher's exact test; Figure 3.9) to be among the down-regulated genes upon *Pbcas4* knockdown (55/96, 58%) than among those that are up-regulated (22/69, 33%). This finding is consistent with *Pbcas4* sharing a miRNA decoy function with its human protein-coding ortholog.

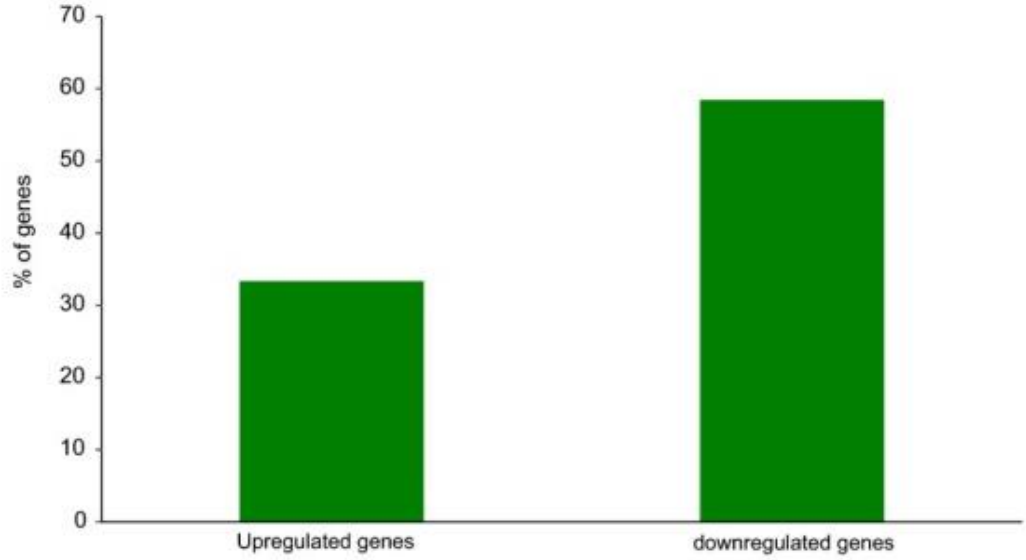


Figure 3.9 *Pbcas4* is a conserved competitive endogenous RNA. Knockdown of *Pbcas4* expression leads to down-regulation of protein-coding genes that have conserved human and mouse MREs for *miR-185/882* and *miR-665* families. The percentage of genes with conserved (mouse and human) MREs for either *miR-185/882* or *miR-665* (Y-axis) upon *Pbcas4* knockdown is 33% and 58% for genes up- and down-regulated, respectively. Figure produced by Dr. Ana Marques and taken from Marques et al., 2012.

Next, I selected five protein-coding gene candidates (*Bcl2*, *Ill7rd*, *Pnpla3*, *Shisa7* and *Tapbp*) whose expression levels were significantly down-regulated following *Pbcas4* loss-of-function and whose mouse and 3' UTR of their human protein-coding orthologs are both predicted to harbour *miR-185/882* and *miR-665* MREs (Table 3.3). I tested using quantitative real time-PCR (qRT-PCR) for the change in expression levels of *Pbcas4/BCAS4* and the five protein-coding genes after transfection of mouse and human neuroblastoma cells (N2A and SH-SY5Y, respectively), with mimics of *miR-185*, whose mature sequence is conserved between mouse and human. Because *miR-882* is not expected to be expressed in N2A cells (Landgraf et al., 2007) and because unlike *miR-185*, the mature sequence of *miR-665* differs, by a single nucleotide, between its mouse and human orthologs, I chose not to test these miRNAs.

Table 3.3 Selected protein-coding gene candidates for experimental validation of *miR-185* binding in mouse (red) and humans (grey). MREs within protein-coding genes are predicted by TargetScan (Garcia et al., 2011).

	Ensembl Gene ID	***Number of MREs***				***Fold change after Pcas4 loss-of-function***
		Total	***miR-665***	***miR-185/882***	***% of total predicted***	
Tapbp	ENSMUSG00000024308	62	1	2	4.84	0.60
	ENSG00000231925	424	6	4	2.36	
Pnpla3	ENSMUSG00000041653	164	3	3	3.66	0.64
	ENSG00000100344	217	1	1	0.92	
Shisa7	ENSMUSG00000053550	215	1	2	1.40	0.64
	ENSG00000187902	777	4	5	1.16	
Bcl2	ENSMUSG00000057329	263	1	2	1.14	0.67
	ENSG00000171791	814	3	2	0.61	
Il17rd	ENSMUSG00000040717	275	2	1	1.09	0.67
	ENSG00000144730	1017	1	3	0.39	

Consistent with the prediction that the expression levels of *Pbcas4* and the five protein-coding gene candidates are post-transcriptionally regulated by *miR-185*, a 68-fold increase in the abundance of this miRNA in N2A cells resulted in significantly reduced levels of mouse *Pbcas4* and each of the five predicted protein-coding transcript targets ($p<10^{-4}$, ANOVA, mean 0.32-fold reduction in expression (0.09- to 0.66-fold reduction), Figure 3.10A, B).

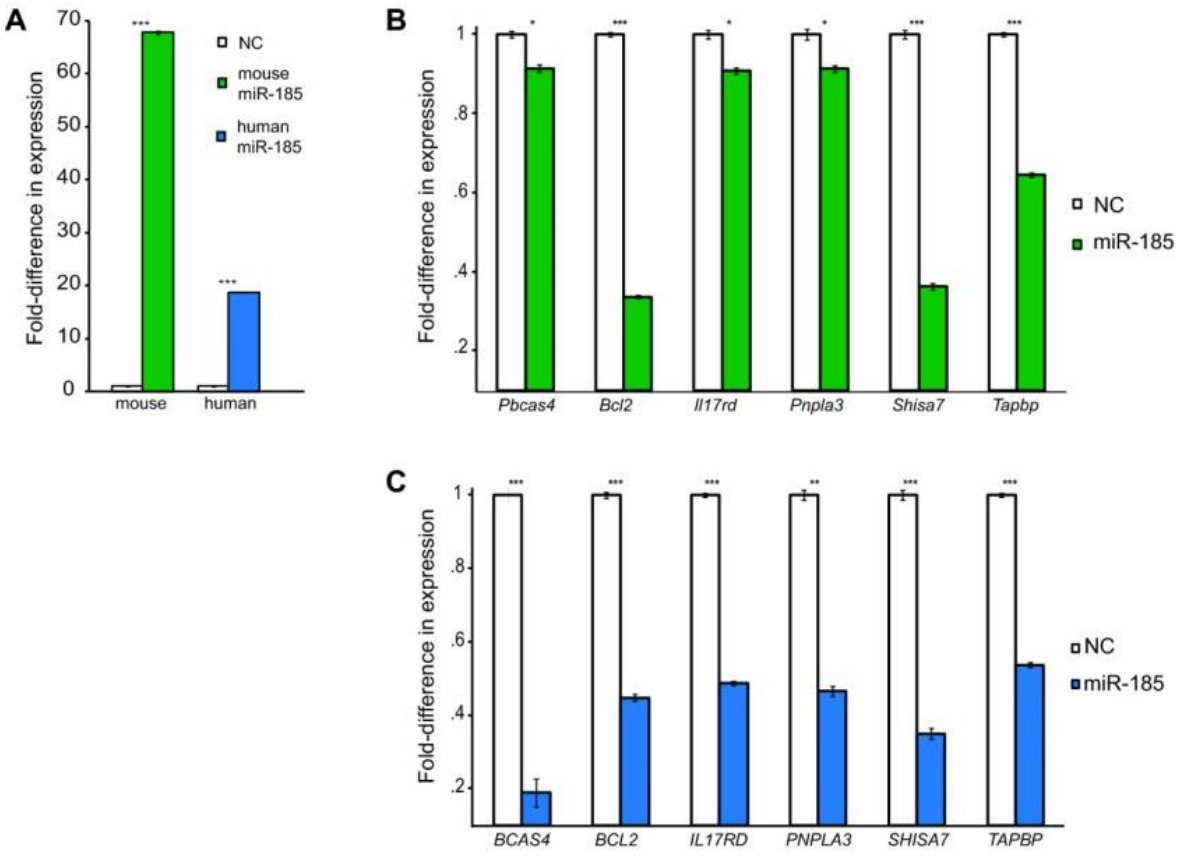


Figure 3.10 *miR-185* mediates conserved crosstalk between *Pbcas4* /*BCAS4* and their protein-coding genes in mouse and human. (A to C) An increased concentration of *miR-185* in mouse and human neuroblastoma cells (N2A and SH-SY5Y cells, respectively) (A) leads to significant down-regulation of *Pbcas4*, *BCAS4* and five mouse (B) and human (C) orthologous pairs that were predicted to compete for binding of *miR-185*. Asterisks indicate significance the level of the comparison (t-test) between the expression of target transcripts after transfection of negative control (set to 1) and *miR-185* mimic (** $p<0.01$; *** $p<0.001$). Figure produced by myself and taken from Marques et al., 2012.

To test whether post-transcriptional regulation by this miRNA is conserved in humans, I similarly transfected the *miR-185* miRNA mimics in human SH-SY5Y cells. Consistently, the transcript abundance of *BCAS4* and that of each of the five human genes (*BCL2*, *ILL7RD*, *PNPLA3*, *SHISA7* and *TAPBP*) was also significantly reduced upon a 19-fold increase in *miR-185* level ($p<10^{-4}$, ANOVA, mean 0.59-fold reduction in expression (0.49- to 0.81-fold reduction), Figure 3.10A, C). Together, these results reinforce the hypothesis that mouse *Pbcas4* has retained the post-transcriptional role as a ceRNA from its protein-coding gene ancestor.

3.5 DISCUSSION

Competition for miRNA-binding between transcripts with shared MREs has recently been demonstrated in animals and plants. MiRNA-mediated crosstalk with many non-homologous mRNAs as participants is likely to be complex and to contribute substantially to the regulation of a transcript's cellular concentration (Marques et al., 2011; Salmena et al., 2011). However, it has remained unclear whether a transcript's role as a miRNA 'decoy' is crucial for either its molecular or organismal function, or whether the biological importance of the decoy role is marginal, owing simply to the promiscuity of miRNA-binding. To address this issue, I used a genetic approach that complements the genomic analysis of a set of transcribed rodent-specific unitary pseudogenes. A substantial fraction (35%) of these transcripts have retained their transcriptional activity despite having lost their protein-coding capability. Transcription of the

unitary pseudogenes suggests conservation of their transcriptional or post-transcriptional regulatory mechanisms, particularly those that are independent of their ancestral coding function.

Using a set of 17 transcribed rodent-specific unitary pseudogenes, this work demonstrates that despite their loss of protein-coding potential, they tend to retain their ancestral gene expression levels and tissue expression patterns. Furthermore, the finding that the ancestral post-transcriptional networks of transcribed rodent-specific unitary pseudogenes are preserved due to, at least in part, the retention of MREs within competitive endogenous RNA (ceRNA) transcripts suggests that it may be a miRNA-mediated mechanism, which forms the crosstalk between ceRNAs, that underlies these preserved post-transcriptional networks. My work aimed to test this hypothesis for one transcribed unitary pseudogene, mouse *Pbcas4*, the unitary pseudogene of human *BCAS4*. My genetic analysis of *Pbcas4* supports that (1) the post-transcriptional regulatory roles of some loci can outlive their protein-coding functions, and (2) these loci are sufficiently important for selection against deleterious mutations to maintain their transcription. Therefore, the miRNA decoy functions of unitary pseudogenes are unlikely to be subordinate to their previous protein-coding functions, at least in the case of *Pbcas4*.

Transcription of duplicated and unitary pseudogenes in eukaryotes has been previously proposed to argue for their functionality (Zheng and Gerstein, 2007). Here, these unitary pseudogenes are shown to have retained the functions, namely their protein-independent and miRNA-dependent post-transcriptional roles of their orthologous protein-coding ancestors, demonstrating the

importance of post-transcriptional regulation through such miRNA-mediated crosstalk.

In addition to transcribed unitary pseudogenes, many intergenic long noncoding RNAs (lincRNAs) also exhibit evidence of active transcription and a large subset of these lincRNAs also demonstrate conservation in sequence and expression profiles across eukaryotic lineages, despite having no apparent or as yet unknown functionality (Church et al., 2009; Cabili et al., 2011). Crosstalks between lincRNAs and protein-coding transcripts through shared MREs have been described in numerous publications (Marques et al., 2011; Salmena et al., 2011; Tan and Marques, 2014) (**Chapter 1**). Therefore, it is plausible that other lincRNAs can participate in crosstalking ceRNA networks to post-transcriptionally regulate their MRE-sharing mRNA targets. This hypothesis has largely motivated my interest in investigating lincRNAs with putative regulatory functions via miRNA-mediated crosstalking mechanisms, which I will discuss in the following chapters of this thesis.

CHAPTER 4

Extensive microRNA-mediated crosstalk between lincRNAs and mRNAs in mouse embryonic stem cells

4.1 ABSTRACT

Recently a handful of intergenic long noncoding RNAs (lincRNAs) have been shown to compete with mRNAs for binding to miRNAs and to contribute to development and disease. Beyond these reports, little is yet known of the extent and functional consequences of miRNA-mediated regulation of mRNA levels by lincRNAs.

To gain further insights into lincRNA-mRNA miRNA-mediated crosstalk, the work described in this chapter reanalyzed transcriptome-wide changes induced by the targeted knockdown of each of over 140 lincRNA transcripts in mouse embryonic stem cells (mESCs). I predicted computationally that on average almost one fifth of the transcript level changes induced by lincRNAs are dependent on those miRNAs that are highly abundant in mESCs. I validated these findings experimentally by temporally profiling transcriptome-wide changes in gene expression following the loss of miRNA biogenesis in mESCs. Following the depletion of miRNAs, I showed that 63% of lincRNAs and their miRNA-dependent mRNA targets were up-regulated coordinately, consistent with their interaction being miRNA-mediated. These lincRNAs were preferentially located in the cytoplasm and the response elements for miRNAs they share with their targets have been preserved in mammals by purifying

selection. Lastly miRNA-dependent mRNA targets of each lincRNA tended to share common biological functions.

Post-transcriptional miRNA-mediated crosstalk between lincRNAs and mRNAs, in mESCs, is thus surprisingly prevalent, conserved in mammals and likely to contribute to critical developmental processes.

4.2 INTRODUCTION

Transcript abundance for large numbers of eukaryotic genes is modulated post-transcriptionally by microRNAs (miRNAs). The recognition and binding of a mature miRNA to response elements (MREs) present within the target transcript leads to its degradation or translational repression (Ambros 2003; Wienholds and Plasterk 2005; Bartel 2009). When a pair of transcripts is targeted by a particular miRNA, a change in the abundance of one transcript can modulate the level of the other transcript in the same direction (Franco-Zorrilla et al., 2007; Marques et al., 2011; Salmena et al., 2011; Tay et al., 2014). Transcripts engaging in such crosstalk are referred to as competitive endogenous RNAs (ceRNAs) (Salmena et al., 2011) (**Chapter 1**). Intricate networks of crosstalking RNAs are proposed to regulate coordinately the relative abundance of functionally-related transcripts (Sumazin et al., 2011; Ala et al., 2013; Han et al., 2013; Wehrspaun et al., 2014). This suggests a layer of post-transcriptional regulation that is overlaid upon other transcriptional programmes.

The miRNA-mediated crosstalk among transcripts can involve both coding and noncoding transcripts, including intergenic long noncoding RNAs (lincRNAs) (Das et al., 2014; Denzler et al., 2014). Thousands of lincRNAs have been annotated in eukaryotic genomes (Derrien et al., 2012; Ulitsky and Bartel, 2013), many of which are preferentially located in the cytoplasm (van Heesch et al., 2014), where they can engage in miRNA-mediated interactions with other transcripts. Both computational and experimental evidence support the extensive targeting of lincRNAs by miRNAs (Paraskevopoulou et al., 2013). Whilst a small number of lincRNAs are currently known to function as ceRNAs (Cesana et al., 2011; Fan et al., 2013; Wang et al., 2013), the full extent of lincRNAs possessing miRNA-dependent regulatory roles remains to be determined (Ulitsky and Bartel, 2013).

Efficient interaction among ceRNAs relies on a number of transcript-specific criteria, such as the relative abundances of miRNA and of crosstalking transcripts (Ebert and Sharp, 2010; Ala et al., 2013; Figliuzzi et al., 2013; Denzler et al., 2014). Most lincRNAs are found at relatively low steady-state levels compared to mRNAs (Cabili et al., 2011; Derrien et al., 2012) and thus it remains unclear how they effectively modulate the abundance of their mRNA targets in a miRNA-dependent manner. Nevertheless, some of lincRNAs with miRNA-dependent roles that have been identified don't have an unusually high number of predicted recognition elements for the miRNAs they share with their mRNA targets and are not especially abundant (Cesana et al., 2011; Wang et al., 2013). These ceRNAs are also no different to most other lincRNAs (Cabili et al., 2011; Derrien et al., 2012) with respect to their highly restricted spatial and temporal expression patterns. For example, *linc-MD1* is a muscle-specific

ceRNA that regulates transcript abundance of two key myogenic transcription factors, *Maml1* and *Mef2c*, which are required for activating muscle-specific gene expression (Cesana et al., 2011); in addition, *linc-RoR* competes for miR-145 binding with key self-renewal transcription factor transcripts, namely *Nanog*, *Oct4*, and *Sox2*, and is expressed during induced pluripotent stem cell (iPSC) reprogramming and in undifferentiated embryonic stem cells (ESCs) (Loewer et al., 2010; Wang et al., 2013). The narrow expression profile of these ceRNAs might specify the cells or tissues in which their activity exerts the greatest effect. Cell fate decisions, such as those involving *linc-RoR* or *linc-MD1*, regularly involve switch-like responses in the expression levels of key regulatory genes that result in coordinated changes in transcription profiles often driven by one or more key transcription factors. Several miRNAs have been found to contribute to the regulation of such switches (Mukherji et al., 2011) and lincRNAs have often been implicated in the regulation of the circuitry underlying cell-fate decisions (Jia et al., 2010; Cesana et al., 2011; Guttman et al., 2011; Sun et al., 2013; Wang et al., 2013). These findings suggest that relatively lowly abundant, yet specifically expressed, lincRNAs might function efficiently as ceRNAs regulating the transition between pluripotent and differentiated cell states. In such bi-stable states, small changes in miRNA levels induced by ceRNAs may have a greater impact on cellular homeostasis than in fully differentiated normal cells.

This study sought to determine the relative prevalence of miRNA-mediated changes induced by lincRNAs. This was done by taking advantage of publicly available and experimentally determined transcriptomic data on the impact of the knockdown, using short hairpin RNAs (shRNAs), for over 140 lincRNAs in

mouse embryonic stem cells (mESCs) (Guttman et al., 2011). This data previously provided support for the notion that some lincRNAs act as protein-binding scaffolds coordinating cell-type specific gene expression changes transcriptionally (Guttman et al., 2011). My analysis demonstrates that lincRNAs can also contribute to mESC fate decisions via post-transcriptional miRNA-mediated mechanisms.

4.3 MATERIALS and METHODS

I performed all the work described below, except where noted otherwise.

Tissue culture

Mouse DTCM23/49 XY embryonic stem cells (mESC) (Nesterova et al., 2008) were grown as described in **Chapter 2**. mESC cells were seeded at a density of 8.0 x 10^5 cells/dish in 10 cm^2 dishes and grown for 24 h prior to tamoxifen treatment. In triplicate, deletion of *Dicer*'s RNAse III domain was induced by culturing the cells in the presence of 800nM tamoxifen [(Z)-4-HydroxyTamoxifen (4-OHT), Sigma H7904]. Cells treated with 0.1% ethanol were used as control (3 replicates). *Dicer*-deficient colonies were selected and expanded from 10 cm^2 to T-75 75 cm^2 tissue culture flasks. Cells were passaged at 70-80% confluence (every 2-3 days) for 12 days.

Quantification of miRNA abundance

Mouse ESCs were harvested and total RNA was extracted using the miRNeasy kit (Qiagen 217004) in quadruplicate. A total of 611 mouse and murine virus-associated miRNAs were quantified using the nCounter miRNA Expression Assay (NanoString Technologies, Seattle, WA) (Geiss et al., 2008) as described in **Chapter 2**. Genome-wide miRNA abundance was normalized using a set of house-keeping mRNAs (Appendix Table A4.1). Unique miRNAs were grouped into miRNA families (as annotated by TargetScan, v6.2) (Garcia et al., 2011) and their expression levels (normalized counts) combined.

Primers used in Chapter 4 are listed in Table 2.3.

RNA sequencing, mapping and quantification of gene expression

Directional polyA-selected RNA sequencing libraries were prepared and sequenced (Illumina HiSEQ 2000) by BGI Tech Solutions (Hong Kong). Total cellular polyA-selected RNA samples at day 0 and day 12 after tamoxifen treatment were sequenced to a depth of approximately 100 million (minimum 93 million; maximum 123 million) 100 bp paired-end reads per sample. Approximately 33 million (minimum 27 million; maximum 41 million) 50 bp paired-end reads per total cellular RNA extracts at days 4, 8 and 10 were sequenced (Table 4.1). Cytosolic and nuclear RNA extracts were multiplexed and sequenced on one lane, yielding on average of approximately 51 million (minimum 45 million; maximum 69 million) 50 bp paired-end reads (Table 4.1).

Table 4.1 Design and sequencing of the experiment. All sequencing reads are pair-ended.

Lane	Sample	Average no. of reads	Read length	No. of biological replicates
1	Day 0 Total RNA	123 million 102 million 104 million	100 bp	3
	Day 12 Total RNA	109 million 110 million 93 million	100 bp	3
2	Day 4 Total RNA	37 million 34 million 28 million	50 bp	3
	Day 8 Total RNA	32 million 41 million 34 million	50 bp	3
	Day 10 Total RNA	36 million 27 million 32 million	50 bp	3
3	Day 0 cytosolic RNA	58 million 49 million 60 million	50 bp	3
	Day 0 nucleus RNA	46 million 51 million 45 million	50 bp	3
	Day 12 cytosolic RNA	51 million 45 million 48 million	50 bp	3
	Day 12 nucleus RNA	45 million 56 million 57 million	50 bp	3

Reads were aligned to the mouse reference genome (mm9) using TopHat (version 2.0.9) (Trapnell et al., 2009). Splice junctions from ENSEMBL build 70 (Flicek et al., 2012) were provided to facilitate read mapping across known splice junctions. Reads with paired mates mapping to distinct chromosomes were discarded. On average 91.0% (minimum 78%; maximum 99%) of RNA sequencing reads were successfully mapped to the mouse genome. To account for differences in RNA-sequencing depth across the five time points following *Dicer* loss-of-function (day 0, 4, 8, 10 and 12), I considered the smallest number of mapped reads (27 million, day 10), and randomly sampled the same number of mapped reads from the remaining samples collected at the five time points. The number of subsampled RNA sequencing reads covering constitutively expressed nucleotides of lincRNAs (Guttman et al., 2011) and ENSEMBL build 70 protein-coding gene and lincRNA annotations (Flicek et al., 2012) were estimated using CoverageBed [Bedtools version 2.17.0 (Quinlan and Hall, 2010)] and used to calculate the expression levels (as total fragments per kilobase of exon per million fragments mapped (FPKM)) across the different libraries at each of the five time points and for each of the replicates. No normalization was performed on subsampled RNA sequencing reads as the subsequent time course gene coexpression analyses using this data set should not be affected by, and thus require, data normalization.

To compare the abundance in the nuclear and cytoplasmic fraction of lincRNAs and their mRNAs targets before and after *Dicer* loss-of-function (day 0 and day 12), locus expression level was determined in each of the compartments independently as previously described and this was used to calculate the ratio between expression levels in the cytoplasm and nucleus.

Multidimensional scaling (MDS) analysis was performed using the edgeR package (Robinson et al., 2010).

The raw sequencing data and estimated transcript expression for the temporal profiling of mouse mESCs following loss of *Dicer* function have been deposited in the GEO under the accession number GSE58757.

Prediction of miRNA response elements

TargetScan (version 6.2) (Garcia et al., 2011) was used to predict MRE in sequences of mouse lincRNAs (Guttman et al., 2011) and the longest 3' UTRs of protein-coding mRNAs expressed in mouse ESCs (ENSEMBL build 70) (Flicek et al., 2012).

A conservative set of experimentally validated MREs was obtained by considering computationally predicted MREs overlapping (100% coverage) regions of the mouse genome enriched in Argonaute binding according to high throughput CLIP-sequencing analysis in mESCs (Leung et al., 2011). I considered the peaks as annotated in the original study (Leung et al., 2011).

Coexpression between lincRNAs and mRNA targets

For each lincRNA, I calculated its pairwise Pearson's correlation in expression across 12 days following Dicer loss-of-function, with all its mRNAs targets, defined as genes that are differentially expressed upon lincRNA loss-of-function experiment conducted by Guttman and colleagues (Guttman et al., 2011). I considered only lincRNAs with evidence of expression at day 0 and with more

than two putative ceRNA targets (ceRNAt, miRNA-dependent targets that were down-regulated upon lincRNA loss-of-function, which also share MREs with the lincRNA) or miRNA-independent targets; more than two targets are needed to estimate expression correlation between lincRNA and their target levels. For these lincRNAs (123), I calculated the median correlation coefficient for their ceRNAts and for their miRNA-independent targets and compared these value to what would be expected based on the median pairwise correlation between the lincRNA and 1,000 randomly selected groups of mESC-expressed genes sampled to the same size as ceRNAts or miRNA-independent sets.

Transcription factor analysis

I considered the 179 genes represented in the microarray used by Guttman and colleagues and annotated as transcription factors (TFs) in AnimalTFDB (Zhang et al., 2012).

Conservation of mouse lincRNAs expression in humans

PolyA-selected RNA-sequencing data from human embryonic stem cell (H1 hESC (Bernstein et al., 2012)) were mapped to the syntenic regions, in humans, of the mouse lincRNAs (obtained using LiftOver (Meyer et al., 2012) with parameters: -minMatch=0.2 -minBlocks=0.01). A mouse lincRNA with at least 5 sequencing reads covering 20% or more of the syntenic region in humans was considered to be conserved in expression. At this cut-off, both the median depth (0%) and coverage (0 reads) of the human syntenic regions by hESCs RNA sequencing reads for 10,000 randomly selected sets of intervals in

the mouse genome with the same length as the lincRNAs considered, but with no evidence of transcriptional activity in mESCs (no reads across the entire region) are zero.

Nucleotide substitution rates

Pairwise alignments of the different sequence features within the expression conserved lincRNAs (all, shared and non-shared MREs and non-MREs), between mouse (mm9) and human (hg19), were concatenated: shared MREs = 1,157 (6,942 bp), non-shared MREs = 198 (1,188 bp), and non-MREs = (111,623 bp). Mouse and human alignments between neighbouring and non-overlapping ancestral repeats (ARs), a good proxy for neutrally evolving sequence (Lunter et al., 2006), were used to simulate (1,000 times) sequence alignments, in mouse, of each of the considered sequences. Specifically, alignments of neighbouring ARs (within 1 Mb of the sequence under consideration) were concatenated and nucleotides within this concatenated AR sequence alignment were randomly selected to simulate 1,000 alignments with similar G+C content and size. Nucleotide substitution rates were estimated using the *REV* substitution model in *baseml* from the PAML package (Yang, 1997).

To obtain empirical *p*-values, the estimated nucleotide substitution rate across the concatenated alignment of the sequences of interest was compared to the estimates obtained for the corresponding simulated putatively neutral sequence alignments.

Integrated functional linkage network analysis

Functional similarity between ceRNAt or miRNA-independent target sets for each lnceRNA was estimated using an integrative phenotypic-linkage network of mouse protein-coding genes (Honti, 2014). For each lnceRNA, the median of functional linkages (measure of functional similarity) between lnceRNA targets in each group were calculated. These were then compared to a distribution of the same measures obtained from 1,000 random bootstrapped gene sets, which were gene length-matched and mESC-expressed, containing the same number of genes as that in the gene set of interest. Functional linkages between lnceRNA targets (nodes) are represented as edges connecting the nodes using Cytoscape (Shannon et al., 2003).

Gene Ontology (GO) enrichment analysis was performed using the functional classification tool Database for Annotation, Visualization, and Integrated Discovered (DAVID, (Huang et al., 2009)) using default parameters and all mESC expressed genes (FPKM > 0) as background. The list of significantly enriched GO terms (after Bonferroni correction, $p<0.05$) was summarized using REViGO with default parameters (Supek et al., 2011) and only non-redundant common ancestral terms were reported.

Statistics

All statistical analyses were done using the R package (R development core team, 2011). Asterisks indicate significance in the level of the comparison between the expression of target transcripts (* $p<0.05$; ** $p<0.01$; *** $p<0.001$;

NS (not significant) $p>0.05$). For each experimental analysis, statistical values were calculated using data collected from three or more independent experiments as described in **Chapter 2**.

4.4 RESULTS

4.4.1 Extensive miRNA-mediated crosstalk among lincRNAs and mRNAs

To investigate the extent of miRNA-dependent gene expression regulation of mRNAs by lincRNAs, I used a large set of experimentally determined expression profile changes in mouse embryonic stem cells (mESCs) induced by the targeted knock-down of 147 lincRNAs and 40 regulatory protein-coding gene controls (Guttman et al., 2011). mRNAs whose expression is differentially up- or down-regulated following knockdown, using short hairpin RNAs (shRNAs), of these noncoding and coding RNAs, were referred to as the 'targets' of the lincRNA or protein-coding gene controls (Figure 4.1). Depletion of lincRNAs resulted in expression level changes for an average of 163 targets, a similar number to that observed for protein-coding gene controls (on average 197 targets) (Guttman et al., 2011).

I considered whether some of these gene expression changes (Guttman et al., 2011) were a consequence of increased post-transcriptional repression of transcripts sharing miRNA response elements (MREs) with the depleted lincRNAs. In contrast with transcriptional regulation by lincRNAs that can lead to either activation or repression of their targets' expression, the primary consequence of competition among lincRNAs and mRNA targets for binding to the same miRNAs is a positive correlation between their transcripts' levels. I applied this signature to predict miRNA-dependent lincRNA-mRNA interactions.

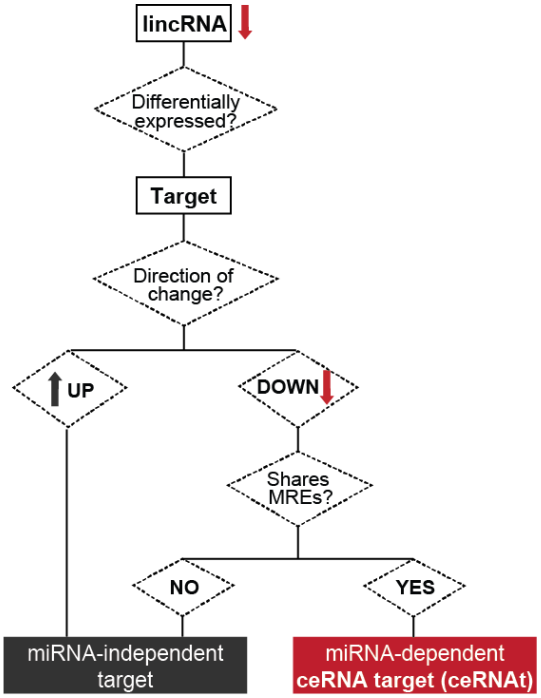


Figure 4.1 MiRNA-dependent regulation of mRNA abundance by a lincRNA. The classification of individual lincRNA's mRNA target as either competitive endogenous RNA (ceRNA) targets (ceRNAt) (red) or miRNA-independent targets (grey). Red and dark grey arrows represent down- and up-regulation respectively.

To predict the extent of miRNA-mediated regulation by lincRNAs, I first identified mESC-expressed miRNAs and then predicted which transcripts they bind and regulate. miRNA levels were quantified, in quadruplicate, using NanoString Technology (Appendix Table A4.1, see section 4.3 Materials and Methods) and subsequent analysis was performed considering only the 25% most highly expressed miRNAs (160 from 117 miRNA families) (Garcia et al., 2011), except where otherwise stated. MREs were predicted using TargetScan (version 6.2) (Garcia et al., 2011) across the entire sequence of the 147 lincRNAs (Guttman et al., 2011) and within the longest annotated 3'UTRs of

mouse protein-coding genes (ENSEMBL build 70, (Flicek et al., 2012), see section 4.3 Materials and Methods).

First, as a negative control, I considered the mRNA targets for each of the 40 regulatory protein-coding gene controls (Guttman et al. 2011). For each target, I calculated the density (number per kilobase (kb) of 3'UTR sequence) of predicted response elements for mESC-expressed miRNAs it shared with the transcription factor that had been identified in the original study as significantly altering its expression (Guttman et al., 2011). These transcription factors' targets were expected to be modulated transcriptionally, not post-transcriptionally (Guttman et al., 2011), and as expected, no significant differences were found in the densities of shared MREs between transcription factors and up- or down-regulated targets (Figure 4.2).

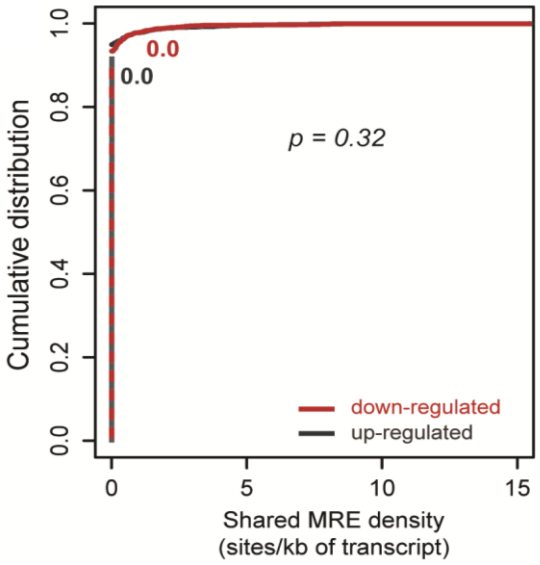


Figure 4.2 Transcription factor controls do not share significantly more MREs than targets that are negatively correlated. Median density of response elements for the top 25% most highly expressed miRNA families in mESCs shared between transcription factor controls and their down (0 sites/kb, red) and up-regulated (0 sites/kb, black) targets.

I then considered the mRNA targets for each of the 147 lincRNAs. In contrast to results for regulatory transcription factor controls, mRNAs that were down-regulated upon lincRNA knockdown shared a significantly higher number of predicted MREs with the lincRNA (median of 2.9 MREs/kb of 3'UTR) than up-regulated targets (median of 2.3 MREs/kb, $p<8x10^{-6}$, two tailed Mann-Whitney test, Figure 4.3). Similar results were obtained when MRE predictions were considered for all miRNAs expressed in mESCs or the 75% or 50% most highly expressed miRNAs in mESCs (Figure 4.4). These results were obtained using computationally predicted MREs, which are known to have relatively high false positive and negative rates (Maziere and Enright, 2007). Consequently, I next considered a stringent set of MREs that overlap (100% coverage) experimentally derived Argonaute-bound regions in mESCs (Leung et al., 2011). With these, I found that down-regulated mRNA targets contain over two-fold higher densities of MREs shared with their lincRNA (mean of $5.0x10^{-3}$ MREs/kb) than up-regulated mRNAs (mean of $2.1x10^{-3}$ MREs/kb, $p<0.05$, two tailed Mann-Whitney test).

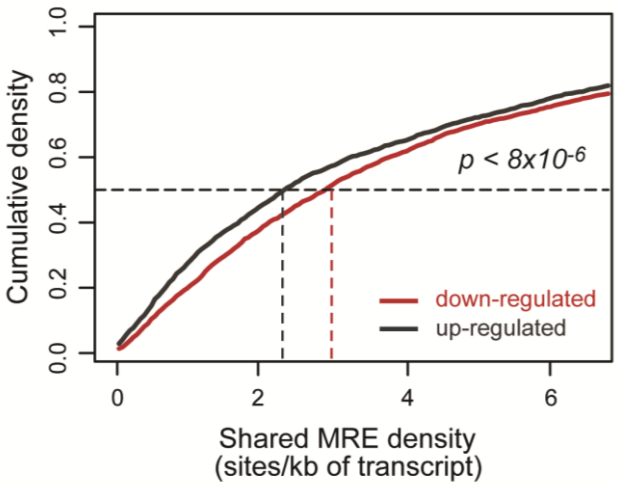


Figure 4.3. The median density of response elements for the top 25% most highly expressed miRNA families in mESCs shared between lincRNAs and their respective down-regulated targets (2.9 sites/kb of transcript, red), is significantly higher ($p<8x10^{-6}$, two-tailed Mann-Whitney test) than those shared with their up-regulated targets (2.3 sites/kb of transcript, black).

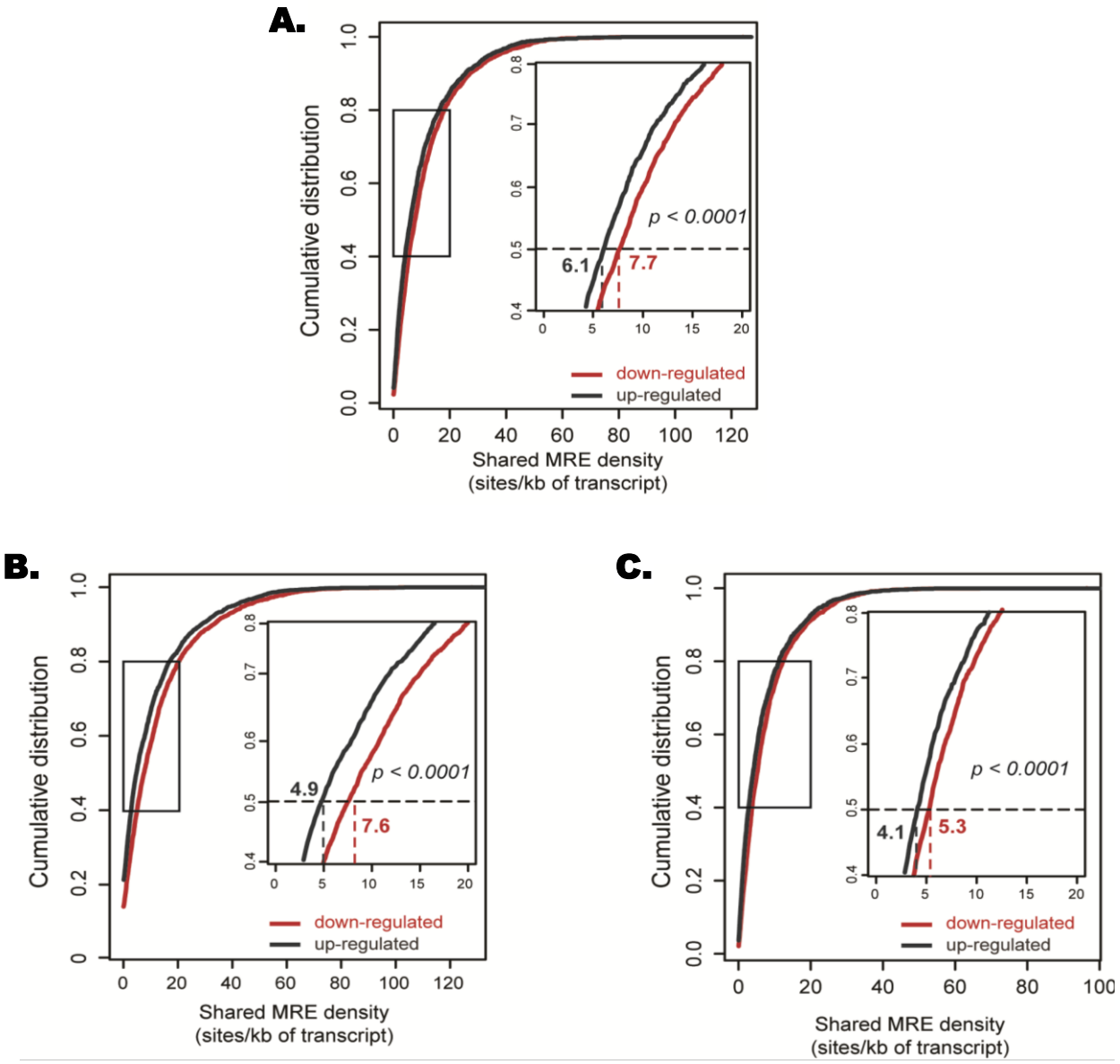


Figure 4.4 Positively correlated targets of lincRNAs share significantly more MREs than targets that are negatively correlated. Median density of response elements for (A) all expressed miRNA families, (B) the top 75%, and (C) the top 50% most highly expressed miRNAs in mESCs shared between lincRNAs and their down-(red) and up-regulated (black) targets. Median densities and two-tailed Mann-Whitney test *p*-values are depicted in the inserts.

These results are consistent with the abundance of some lincRNAs' mRNA targets being modulated through their competition for binding mESC-expressed miRNAs. On average, 19.3% (Table 4.2) of lincRNAs' targets exhibited the two signatures of miRNA-dependent crosstalk: 1) depletion of the lincRNA is associated with down-regulation of its mRNA target, and 2) coding and noncoding transcripts contain predicted MREs for the same miRNAs (Figure 4.1). Hereafter, I refer to mRNA targets of lincRNAs with these signatures as competitive endogenous RNA targets (ceRNAt).

What then of mRNAs whose changes in transcript abundance could not be explained by a ceRNA mechanism (Figure 4.1)? I hypothesized that some of these changes might be a consequence of secondary effects of transcriptional regulation mediated by transcription factor whose transcripts are primary targets of lincRNAs (i.e. TF ceRNAt). I found that for 87% (128 of 147) of lincRNAs, their knockdown was associated with significant down-regulation of mESC-expressed TFs with whom they share MREs. To test this hypothesis, I considered the 10 lincRNAs that have TF ceRNAt whose effects on gene expression upon knockdown were also experimentally determined by Guttman and colleagues (Guttman et al., 2011). Consistent with this indirect mode of mRNA regulation, the levels of twice as many miRNA-independent targets of these lincRNAs were affected in the same direction rather than in opposite directions upon knockdown of either the lincRNA or its respective TF ceRNAt (averages 15.8% versus 7.8%, respectively; $p<0.05$, two tailed Mann-Whitney test). This indicates that a proportion of gene expression changes that are not primarily miRNA-dependent may be explained through secondary transcriptional activation by TFs whose transcripts are ceRNAt.

Table 4.2 Number of ceRNAt and miRNA-independent target mRNAs per lincRNA predicted using the top 25% most highly expressed miRNAs.

lincRNA	ceRNAt		miRNA-independent target mRNAs		Total number of lincRNA-affected genes
	Number	%	Number	%	
linc1230	56	33.7%	110	66.3%	166
linc1235	2	2.1%	95	97.9%	97
linc1239	4	2.1%	186	97.9%	190
linc1242	16	5.0%	306	95.0%	322
linc1244	27	28.4%	68	71.6%	95
linc1245	114	23.1%	380	76.9%	494
linc1251	1	50.0%	1	50.0%	2
linc1252	42	33.6%	83	66.4%	125
linc1253	17	23.3%	56	76.7%	73
linc1256	45	22.6%	154	77.4%	199
linc1257	0	0.0%	58	100.0%	58
linc1259	47	32.2%	99	67.8%	146
linc1260	7	17.1%	34	82.9%	41
linc1261	31	11.6%	236	88.4%	267
linc1262	32	19.3%	134	80.7%	166
linc1267	110	34.6%	208	65.4%	318
linc1270	46	28.6%	115	71.4%	161
linc1274	0	0.0%	5	100.0%	5
linc1281	376	40.2%	560	59.8%	936
linc1282	62	28.4%	156	71.6%	218
linc1283	60	45.1%	73	54.9%	133
linc1289	0	0.0%	0	0.0%	0
linc1290	174	46.5%	200	53.5%	374
linc1293	4	7.4%	50	92.6%	54
linc1296	105	35.1%	194	64.9%	299
linc1300	37	31.1%	82	68.9%	119
linc1304	12	15.6%	65	84.4%	77
linc1307	39	19.7%	159	80.3%	198
linc1312	64	38.1%	104	61.9%	168
linc1313	9	6.4%	132	93.6%	141
linc1315	35	40.7%	51	59.3%	86
linc1316	65	46.1%	76	53.9%	141
linc1317	3	50.0%	3	50.0%	6
linc1327	115	31.0%	256	69.0%	371
linc1328	0	0.0%	7	100.0%	7
linc1331	13	8.4%	141	91.6%	154
linc1335	32	27.8%	83	72.2%	115
linc1337	0	0.0%	8	100.0%	8

lincRNA	ceRNAt		miRNA-independent target mRNAs		Total number of lincRNA-affected genes
	Number	%	Number	%	
linc1338	50	23.7%	161	76.3%	211
linc1346	94	28.5%	236	71.5%	330
linc1347	37	19.2%	156	80.8%	193
linc1349	9	11.5%	69	88.5%	78
linc1354	241	31.5%	524	68.5%	765
linc1356	96	37.5%	160	62.5%	256
linc1359	5	8.6%	53	91.4%	58
linc1361	7	10.0%	63	90.0%	70
linc1366	9	12.0%	66	88.0%	75
linc1368	23	16.3%	118	83.7%	141
linc1369	10	24.4%	31	75.6%	41
linc1382	48	51.1%	46	48.9%	94
linc1385	52	23.7%	167	76.3%	219
linc1386	5	9.3%	49	90.7%	54
linc1388	12	21.4%	44	78.6%	56
linc1389	17	31.5%	37	68.5%	54
linc1390	74	19.8%	299	80.2%	373
linc1391	7	12.3%	50	87.7%	57
linc1393	113	36.0%	201	64.0%	314
linc1400	12	19.4%	50	80.6%	62
linc1405	120	36.5%	209	63.5%	329
linc1406	2	100.0%	0	0.0%	2
linc1410	65	23.5%	212	76.5%	277
linc1411	42	34.7%	79	65.3%	121
linc1412	6	6.2%	91	93.8%	97
linc1413	8	4.1%	187	95.9%	195
linc1418	71	14.2%	430	85.8%	501
linc1419	14	8.9%	143	91.1%	157
linc1421	11	22.0%	39	78.0%	50
linc1422	6	16.2%	31	83.8%	37
linc1425	52	38.5%	83	61.5%	135
linc1427	125	49.2%	129	50.8%	254
linc1428	14	15.4%	77	84.6%	91
linc1434	11	12.4%	78	87.6%	89
linc1435	13	28.3%	33	71.7%	46
linc1448	17	6.7%	235	93.3%	252
linc1450	23	25.3%	68	74.7%	91
linc1454	21	10.6%	178	89.4%	199
linc1456	39	19.6%	160	80.4%	199
linc1457	4	3.2%	121	96.8%	125
linc1463	37	32.5%	77	67.5%	114
linc1465	3	4.6%	62	95.4%	65

lincRNA	ceRNAt		miRNA-independent target mRNAs		Total number of lincRNA-affected genes
	Number	%	Number	%	
linc1468	10	40.0%	15	60.0%	25
linc1470	17	16.5%	86	83.5%	103
linc1471	170	19.7%	694	80.3%	864
linc1473	8	7.5%	98	92.5%	106
linc1477	7	12.5%	49	87.5%	56
linc1483	13	25.5%	38	74.5%	51
linc1484	1	16.7%	5	83.3%	6
linc1490	7	11.7%	53	88.3%	60
linc1503	18	51.4%	17	48.6%	35
linc1505	67	22.5%	231	77.5%	298
linc1506	0	0.0%	5	100.0%	5
linc1510	25	22.5%	86	77.5%	111
linc1517	4	9.3%	39	90.7%	43
linc1524	5	3.3%	147	96.7%	152
linc1526	0	0.0%	35	100.0%	35
linc1536	58	24.5%	179	75.5%	237
linc1537	8	24.2%	25	75.8%	33
linc1540	75	46.9%	85	53.1%	160
linc1543	13	28.9%	32	71.1%	45
linc1547	17	9.3%	165	90.7%	182
linc1552	5	5.0%	96	95.0%	101
linc1555	110	34.0%	214	66.0%	324
linc1557	54	22.0%	192	78.0%	246
linc1558	18	15.4%	99	84.6%	117
linc1559	200	35.3%	366	64.7%	566
linc1562	31	23.3%	102	76.7%	133
linc1563	71	36.6%	123	63.4%	194
linc1572	24	35.8%	43	64.2%	67
linc1581	39	23.2%	129	76.8%	168
linc1582	17	19.1%	72	80.9%	89
linc1588	0	0.0%	20	100.0%	20
linc1589	0	0.0%	25	100.0%	25
linc1592	0	0.0%	1	100.0%	1
linc1595	0	0.0%	245	100.0%	245
linc1596	8	32.0%	17	68.0%	25
linc1598	1	3.1%	31	96.9%	32
linc1599	4	12.1%	29	87.9%	33
linc1600	2	7.7%	24	92.3%	26
linc1601	18	17.8%	83	82.2%	101
linc1602	17	13.7%	107	86.3%	124
linc1603	56	18.5%	246	81.5%	302
linc1604	13	5.4%	227	94.6%	240

lincRNA	ceRNAt		miRNA-independent target mRNAs		Total number of lincRNA-affected genes
	Number	%	Number	%	
linc1607	65	18.0%	297	82.0%	362
linc1608	0	0.0%	44	100.0%	44
linc1609	1	2.2%	45	97.8%	46
linc1610	0	0.0%	563	100.0%	563
linc1611	8	2.5%	318	97.5%	326
linc1612	16	9.8%	147	90.2%	163
linc1613	18	20.0%	72	80.0%	90
linc1614	3	3.0%	96	97.0%	99
linc1615	9	12.7%	62	87.3%	71
linc1616	15	8.2%	169	91.8%	184
linc1617	10	4.4%	216	95.6%	226
linc1618	1	3.1%	31	96.9%	32
linc1621	57	14.8%	327	85.2%	384
linc1622	18	11.3%	142	88.8%	160
linc1623	20	5.4%	351	94.6%	371
linc1624	0	0.0%	34	100.0%	34
linc1626	83	16.3%	427	83.7%	510
linc1627	5	4.8%	100	95.2%	105
linc1629	14	4.8%	275	95.2%	289
linc1630	1	2.6%	37	97.4%	38
linc1631	7	20.6%	27	79.4%	34
linc1632	25	36.2%	44	63.8%	69
linc1633	5	5.9%	80	94.1%	85
linc1634	9	6.8%	123	93.2%	132
linc1635	88	26.1%	249	73.9%	337

To illustrate this phenomenon, I considered the proposed miRNA-mediated crosstalk between *linc1471* and the mRNA encoding the transcription factor *Oct4*. Two of the 5 recognition elements predicted within the 3'UTR of *Oct4* were for *miR-421* and *miR-762,* for which MREs were also predicted in *linc1471* whose knockdown (Figure 4.5). The loss-of-function of *linc1471* led to a significant 21-fold decrease in *Oct4* mRNA abundance (Guttman et al., 2011). Some of *linc1471*'s targets, whose levels could not be explained by a ceRNA mechanism, may have resulted from secondary effect of miRNA-mediated decrease in *Oct4* expression. In these cases, one would expect that the levels of these genes to be affected upon knockdown of either *linc1471* or *Oct4* in the same direction. Indeed, miRNA-independent targets of *linc1471* that were also differentially expressed upon *Oct4* knockdown (by 14.8-fold, (Guttman et al. 2011)) are 5 times more likely to change in in the same (n=165) rather than in opposing directions (n=33) ($p<10^{-4}$, Fisher's exact test, Figure 4.5). Furthermore, of the *linc1471* targets that I hypothesized might be a secondary effect of *Oct4* changes, 21.2% show evidence that OCT4 protein binds at their promoter, in mESCs (Karwacki-Neisius et al., 2013), a significantly ($p<0.05$, Fisher's exact test) higher proportion than found when considering all other *linc1471* targets (13.7%).

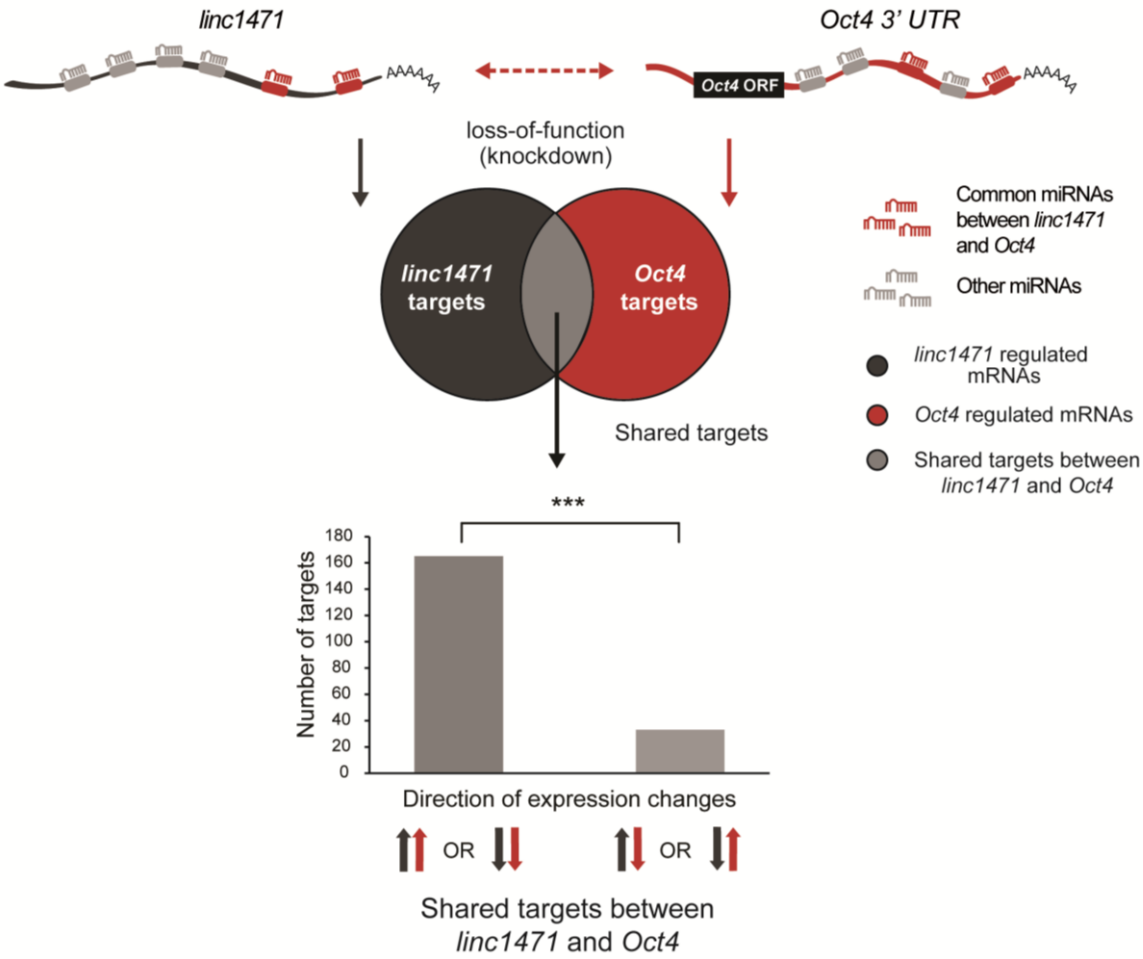


Figure 4.5 MiRNA-dependent regulation of mRNA abundance may be indirectly modulated by a lincRNA through transcription factor regulation. The transcription factor *Oct4* (red) and ceRNAt of *linc1471* (dark grey) were predicted to compete for binding of *miR-421* and *miR-762*, indicated with red boxes within transcript, and can regulate each other's abundance (dotted red arrow). MREs for miRNAs that are not common between the two genes are represented in light grey. Of the 198 *linc1471* targets that cannot be explained solely by a ceRNA mechanism, a significantly higher number (165, 83%) change in the same direction upon this lincRNA (dark grey) and *Oct4* (red) knockdown. Arrows indicate the direction of the observed expression changes for *linc1471* and *Oct4* knockdown. This indicates that the levels of 19.1% (165/864) of the gene transcripts possibly alter because of Oct4 level changes induced by miRNA mediated crosstalk with *linc1471*.

4.4.2 lincRNAs and their respective ceRNAt are co-ordinately up-regulated upon loss of miRNA biogenesis

Next, I undertook a more direct approach to experimentally validate lincRNA crosstalk with mRNAs via miRNAs. For this, I took advantage of mESCs in which a conditional mutation of a key gene in miRNA biogenesis, *Dicer*, was introduced. These cells contain a tamoxifen-inducible *Cre* recombinase that drives recombination between *loxP* sites flanking the *Dicer* RNase III domain (Nesterova et al., 2008). Loss of this domain (Figure 4.6A) and ablation of miRNA biogenesis occurred following tamoxifen addition. Global effects of tamoxifen on DNA synthesis, RNA polymerase activity and transcriptional regulation have been reported (Jordan and Koerner, 1975). These effects were not considered in the experimental design and can be accounted for by including an additional negative control where tamoxifen is introduced to a mESC cell line that does not contain the tamoxifen-inducible *Dicer* excision.

The time for complete miRNA removal, following *Dicer* loss-of-function, varied considerably owing to dependences on miRNA initial abundance and stability, as illustrated in Figure 4.6B-C for *miR-302* and *miR-200c*. Furthermore, I also profiled the changes in abundance of three transcription-factors with established roles in the maintenance of mESC pluripotency across the time course, specifically, *Nanog, Oct4,* and *Sox2*. I found no significant difference in the levels of these genes, suggesting that the loss of miRNA biogenesis did not affect substantially the pluripotent state of mESCs (Figure 4.6D-F). This temporal variation of miRNA abundance allowed me to investigate the miRNA-dependency of interactions between transcripts. In particular, I expected the

expression levels over time for a lincRNA and its ceRNAt to increase coordinately as the levels of miRNAs mediating their crosstalk diminish. No expression correlation was expected following loss of miRNA biogenesis between lincRNAs and mRNA targets that I predicted to be primarily regulated in a miRNA-independent manner, in particular for lincRNAs and mRNAs that do not share predicted MREs for mESC-expressed miRNAs (Figure 4.1).

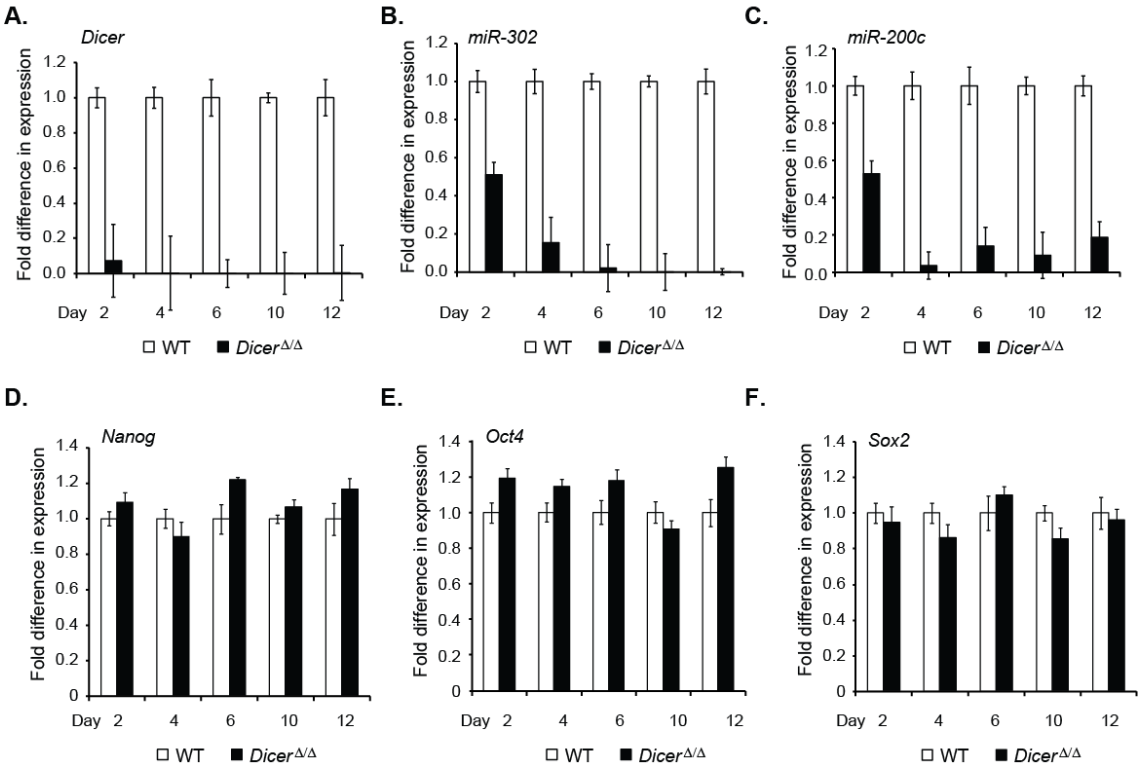


Figure 4.6 Change in expression levels of *Dicer*, two miRNAs, and several transcription factors upon loss of miRNA biogenesis. Loss of miRNA biogenesis (*Dicer*$^{\Delta/\Delta}$) is associated with decreased levels (black), relative to wild-type control (white), of (A) *Dicer*, mature (B) *miR-302* and (C) *miR-200c* abundance over a 12 day time course. In contrast, no change in abundance was observed for several transcription factors, specifically, (D) *Nanog*, (E) *Oct4* and (F) *Sox2*. Fold difference in expression relative to control was determined using qRT-PCR.

I determined the temporal variation of mRNA levels by collecting, in triplicate, and then sequenced poly-adenylated (polyA) RNA from mESCs prior to tamoxifen addition (day 0) and after the injection of tamoxifen at days 4, 8, 10 and 12 thereafter (Figure 4.7). On average, 88.0% (87.0%-89.6%) of RNA sequencing paired-end reads were mapped uniquely to the mouse genome (mm9). As expected, mRNA and lincRNA expression levels, in general, were clustered by time point (Figure 4.8). No significant decrease in *Myc* expression levels (in FPKM) was observed (Fold-change=0.20±0.04, FDR=0.57) in my experiment, in contrast to homozygous *Dgrc8*$^{-/-}$ (Melton et al., 2010) or *Dicer*$^{-/-}$ (Zheng et al., 2014) mESCs. In contrast to a recent analysis of mESC lincRNAs (Zheng et al., 2014), transcriptional regulation by *Myc* appeared not to contribute significantly to the changes in lincRNA or mRNA levels observed in my study.

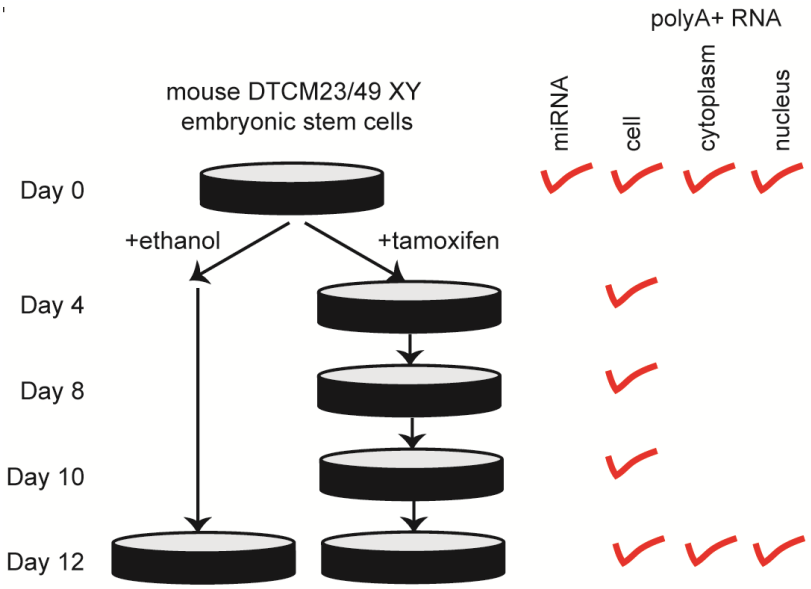


Figure 4.7 Long polyA selected RNA from total cellular extracts of DTCM23/49 XY mouse embryonic stem cells (mESCs) was collected on days 0, 4, 8, 10 and 12 following exposure to tamoxifen. Long PolyA selected RNA was also collected from the nuclear and cytoplasm of these cells before 0 and 12 days after treatment with tamoxifen. Total RNA used to quantify miRNA expression was also extracted before tamoxifen treatment.

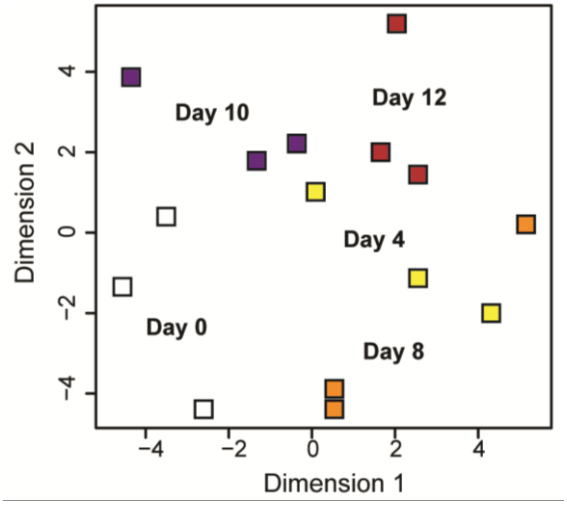


Figure 4.8 Multidimensional scaling (MDS) plot illustrating the separation of RNA sequencing data collected at each of the 5 time points (days 0-white, 4-yellow, 8-orange, 10-purple and 12-red) following Tamoxifen treatment of mESCs, each with 3 biological replicates.

I considered 123 of the initial set of 147 lincRNAs because these were expressed in the untreated mESCs that I analyzed and had multiple (>2) ceRNAt or miRNA-independent targets expressed in these cells.

As expected, lincRNA-ceRNAt pairs exhibited greater correlation in expression levels over time than lincRNA and miRNA-independent targets ($p<3x10^{-8}$, two tailed Mann-Whitney test, Figure 4.9A) consistent with the decreased abundance of miRNAs leading to coordinated changes in the abundance of transcripts that can regulate each other's levels via miRNA-mediated crosstalk. 76% (94 of 123) of the lincRNAs were better correlated in expression with their ceRNAt than with their miRNA-independent targets. For example, expression of *linc1405* was more highly correlated with its ceRNAt (166 targets, R=0.82) than with miRNA-independent targets (117 targets, R=0.35, Figure 4.9B).

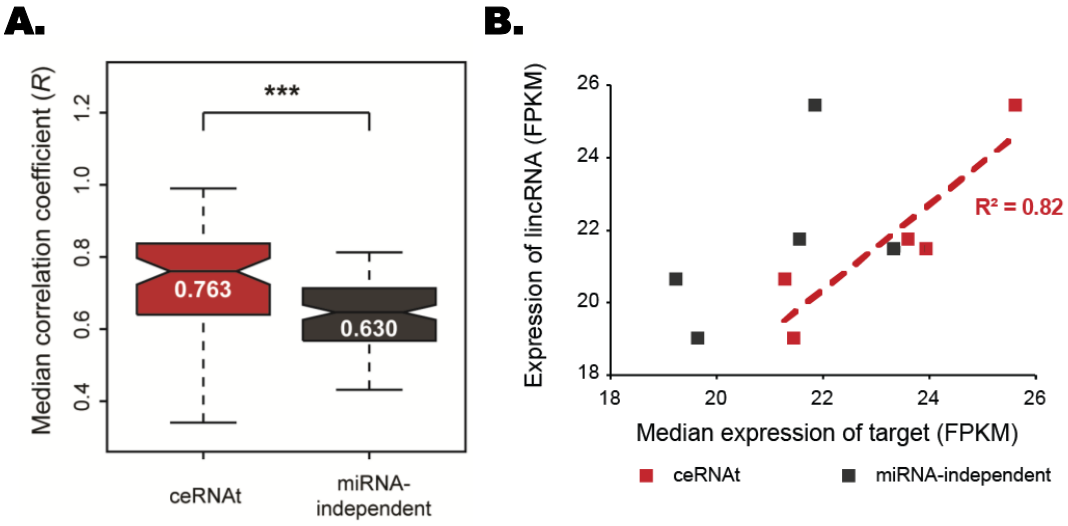


Figure 4.9 The expression levels of lincRNAs and their ceRNAt are positively correlated upon loss of miRNA biogenesis. (A) Over the 12 day time course following loss of miRNA biogenesis, lincRNAs' expression is significantly (p<3x10^{-8}, two-tailed Mann-Whitney test) better correlated with the levels of their ceRNAt (red, median R=0.763) than those of their miRNA-independent targets (grey, R=0.630). (B) *linc1405*'s expression (average across replicates of the gene expression measured as Fragment Per Kilobase of exon per Million reads mapped, FPKM, Y-axis) is better correlated with its ceRNAt (R=0.82, red) than with its miRNA-independent targets (R=0.35, grey). Pearson's correlation was calculated between the expression of *linc1405* and the median expression of all genes annotated as either ceRNAt or miRNA-independent targets at each time point (X-axis).

The expression level of 63% (78 of 123) of lincRNAs was significantly (p<0.05, empirical p value) better correlated with the levels of their ceRNAt, following *Dicer* knockout, than expected based on estimates for 1,000 sets of randomly selected mESC-expressed transcripts pairs of the same size (see section 4.3 Materials and Methods). These 78 lincRNAs, whose miRNA dependency of ceRNAt interactions was predicted experimentally, are hereafter referred to as long noncoding competitive endogenous RNAs or lnceRNAs (Table 4.3). Only 7 of 123 (~6%) lincRNAs were significantly correlated following *Dicer* conditional

excision with its miRNA-independent targets compared to randomly selected transcripts, consistent with chance observations (Figure 4.10, Table 4.4).

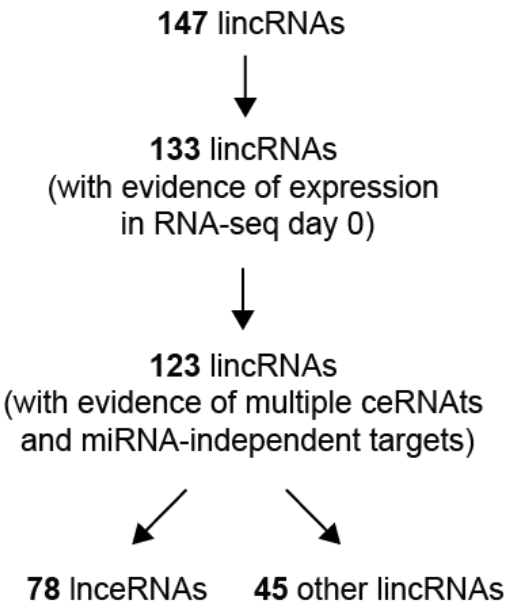


Figure 4.10 Flowchart illustrating filtering and classification of lincRNAs.

In summary, my results are consistent with 78 (63%) of the mESC-expressed lincRNAs investigated here interacting with their mRNA targets in a miRNA-dependent manner.

Table 4.3 Pairwise correlation coefficient between lnceRNAs and their ceRNAts.

lnceRNA	ceRNAts		miRNA-independent target mRNAs	
	Median correlation coefficient (R)	Median empirical *p* value	Median correlation coefficient (R)	Median empirical *p* value
linc1230	0.775	0.001	0.551	0.523
linc1239	0.891	0.001	0.654	0.641
linc1242	0.791	0.001	0.562	0.725
linc1245	0.787	0.001	0.432	1.000
linc1256	0.784	0.001	0.546	0.598
linc1259	0.861	0.001	0.638	0.339
linc1260	0.863	0.001	0.521	0.853
linc1261	0.682	0.001	0.549	0.071
linc1262	0.577	0.039	0.318	0.557
linc1267	0.575	0.001	0.442	0.645
linc1281	0.939	0.001	0.676	0.140
linc1282	0.892	0.001	0.648	0.218
linc1283	0.878	0.047	0.614	0.816
linc1290	0.916	0.001	0.668	0.032
linc1293	0.706	0.023	0.480	0.566
linc1307	0.913	0.001	0.660	0.239
linc1313	0.721	0.001	0.522	0.755
linc1316	0.728	0.001	0.592	0.094
linc1327	0.756	0.001	0.370	1.000
linc1331	0.957	0.001	0.669	0.690
linc1338	0.699	0.001	0.417	0.950
linc1346	0.949	0.001	0.693	0.083
linc1349	0.906	0.001	0.613	0.931
linc1354	0.966	0.001	0.655	1.000
linc1356	0.928	0.001	0.635	0.912
linc1359	0.845	0.001	0.570	0.690
linc1366	0.855	0.001	0.695	0.115
linc1368	0.728	0.001	0.485	0.840
linc1382	0.936	0.001	0.679	0.438
linc1385	0.781	0.001	0.592	0.082
linc1386	0.885	0.001	0.697	0.130
linc1389	0.677	0.001	0.485	0.505
linc1390	0.401	0.001	0.256	0.734
linc1391	0.956	0.001	0.685	0.126
linc1393	0.741	0.001	0.480	0.863
linc1400	0.710	0.009	0.574	0.223
linc1405	0.614	0.001	0.587	0.000
linc1410	0.906	0.001	0.647	0.534

lnceRNA	ceRNAts		miRNA-independent target mRNAs	
	Median correlation coefficient (R)	Median empirical *p* value	Median correlation coefficient (R)	Median empirical *p* value
linc1425	0.883	0.001	0.594	0.761
linc1427	0.756	0.001	0.544	0.298
linc1434	0.686	0.002	0.504	0.572
linc1435	0.781	0.003	0.630	0.375
linc1448	0.738	0.001	0.550	0.069
linc1450	0.868	0.001	0.576	0.903
linc1456	0.964	0.001	0.678	0.547
linc1463	0.751	0.001	0.559	0.629
linc1468	0.776	0.039	0.559	0.430
linc1470	0.860	0.001	0.593	0.748
linc1471	0.891	0.001	0.553	1.000
linc1473	0.666	0.001	0.528	0.246
linc1505	0.930	0.001	0.649	1.000
linc1510	0.959	0.001	0.673	0.751
linc1517	0.769	0.001	0.481	0.811
linc1536	0.829	0.001	0.553	0.620
linc1537	0.762	0.001	0.476	0.890
linc1540	0.592	0.001	0.453	0.290
linc1543	0.758	0.001	0.641	0.037
linc1547	0.882	0.001	0.609	0.662
linc1555	0.545	0.011	0.484	0.091
linc1557	0.938	0.001	0.657	0.461
linc1559	0.910	0.001	0.639	0.759
linc1562	0.715	0.001	0.559	0.134
linc1563	0.743	0.001	0.604	0.041
linc1572	0.661	0.001	0.579	0.698
linc1581	0.925	0.001	0.685	0.149
linc1582	0.679	0.001	0.480	0.719
linc1601	0.881	0.001	0.622	0.686
linc1602	0.560	0.001	0.416	0.784
linc1603	0.865	0.001	0.623	0.483
linc1617	0.966	0.001	0.665	0.786
linc1618	0.394	0.001	0.825	0.787
linc1621	0.767	0.001	0.630	0.509
linc1622	0.779	0.017	0.611	0.058
linc1623	0.929	0.001	0.659	0.601
linc1626	0.944	0.001	0.663	0.825
linc1629	0.826	0.001	0.601	0.515
linc1631	0.755	0.001	0.620	0.637
linc1635	0.734	0.010	0.573	0.101

Table 4.4 Pairwise correlation coefficient between non-lnceRNAs and their miRNA-independent target genes.

Non-lnceRNAs	ceRNAts		miRNA-independent target mRNAs	
	Median correlation coefficient (R)	Median empirical *p* value	Median correlation coefficient (R)	Median empirical *p* value
linc1244	0.787	0.092	0.761	0.142
linc1252	0.927	0.193	0.939	0.693
linc1253	0.572	0.793	0.738	0.142
linc1270	0.643	0.193	0.715	0.532
linc1296	0.572	0.977	0.732	0.013
linc1300	0.547	0.668	0.703	0.235
linc1304	0.750	0.281	0.878	0.378
linc1312	0.557	0.528	0.660	0.339
linc1315	0.639	0.950	0.889	0.004
linc1335	0.266	0.083	0.709	0.079
linc1347	0.707	0.483	0.849	0.058
linc1361	-0.062	0.535	-0.015	0.460
linc1369	0.924	0.295	0.851	0.967
linc1388	0.763	0.167	0.914	0.756
linc1411	0.567	0.667	0.838	0.006
linc1412	0.000	0.138	0.758	0.627
linc1413	0.000	0.138	0.760	0.628
linc1418	0.781	0.403	0.952	0.479
linc1419	0.910	0.281	0.529	0.256
linc1421	0.310	0.260	0.279	0.860
linc1422	0.783	0.673	0.631	0.376
linc1428	0.290	0.167	0.292	0.479
linc1454	0.974	0.100	0.822	0.738
linc1477	-0.157	0.846	0.838	0.201
linc1483	0.342	0.871	0.721	0.435
linc1490	0.887	0.235	0.698	0.130
linc1503	0.266	0.695	0.709	0.431
linc1524	0.087	0.899	0.026	0.356
linc1552	0.728	0.060	0.509	0.676
linc1558	0.553	0.137	0.692	0.401
linc1596	0.852	0.001	0.885	0.094
linc1599	0.880	0.073	0.615	0.818
linc1604	0.957	0.085	0.754	0.389
linc1607	0.797	0.078	0.748	0.714
linc1611	0.738	0.100	0.698	0.130
linc1612	0.479	0.204	0.490	0.357
linc1613	0.072	0.204	0.069	0.353

Non-lnceRNAs	ceRNAts		miRNA-independent target mRNAs	
	Median correlation coefficient (R)	Median empirical *p* value	Median correlation coefficient (R)	Median empirical *p* value
linc1614	0.688	0.097	0.634	0.873
linc1615	0.941	0.082	0.855	0.943
linc1616	0.817	0.294	0.793	0.420
linc1627	0.613	0.783	0.803	0.364
linc1630	0.572	0.001	0.751	0.515
linc1632	0.072	0.781	0.868	0.637
linc1633	0.086	0.897	0.155	0.360
linc1634	0.347	0.481	0.834	0.982

4.4.3 lnceRNAs are enriched in the cytoplasm

I expected lnceRNAs to be enriched in the cytoplasm because post-transcriptional regulation of gene expression by miRNAs occurs preferentially in this subcellular compartment (Bartel, 2004). Gene expression in cytoplasmic and nuclear mESC fractions was determined by extracting and sequencing polyA-selected RNA in triplicate (see section 4.3 Materials and Methods) before (day 0) and after (day 12) loss of *Dicer* function. Expression data from the different experimental conditions were clearly separated using multidimensional scaling analysis (Figure 4.11). I estimated the expression in the cytosol relative to the nucleus as r = $\text{expression}_{\text{cytosol}}/\text{expression}_{\text{nucleus}}$ for each mESC expressed locus.

The 123 mESC-expressed lincRNAs (median r = 0.644) considered were more abundant in the nuclear fraction than mRNAs (median r = 0.828, p<0.05, two-tailed Mann-Whitney test, Figure 4.12A), as was seen previously for a large set of human lincRNAs (Derrien et al., 2012). However, relative to the set of all ENSEMBL-annotated lincRNAs expressed in mESCs they were more abundant in the cytoplasm (median r = 0.470, p<$2.2\text{x}10^{-16}$, two-tailed Mann-Whitney test, Figure 4.12A). This difference may reflect, at least in part, the increased efficiency of RNAi targeting of transcripts in the cytosol which, in the original study (Guttman et al., 2011), would have favoured selection of cytosol-enriched lincRNAs for transcriptome-wide profiling. The subset of lincRNAs previously reported to physically interact with chromatin (Guttman et al., 2011) were found to be significantly (median r = 0.572, p<0.05, two-tailed Mann-Whitney test) more abundant in the nucleus of mESC than the remainder of the lincRNAs

tested (median r=0.738), which is consistent with their chromatin-association and proposed transcriptional roles that take place in the nucleus (Figure 4.12B).

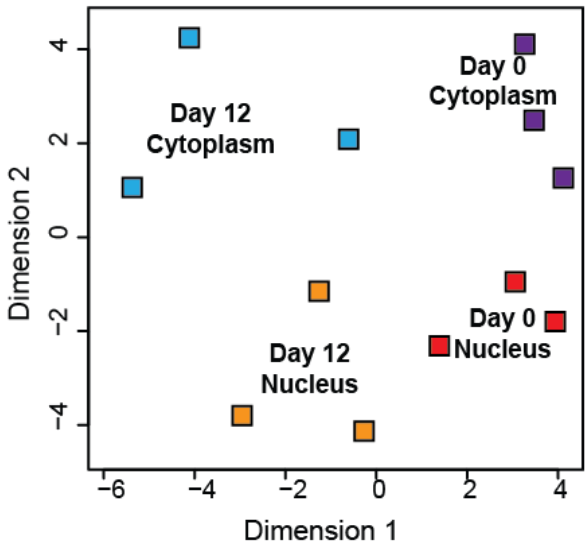


Figure 4.11 Multidimensional scaling (MDS) plot depicting RNA sequencing data of the cytoplasmic and nuclear subcellular fractions of mESCs before (day 0, cytoplasm-purple, nucleus-red) and 12 days after tamoxifen treatment (cytoplasm-blue, nucleus-orange).

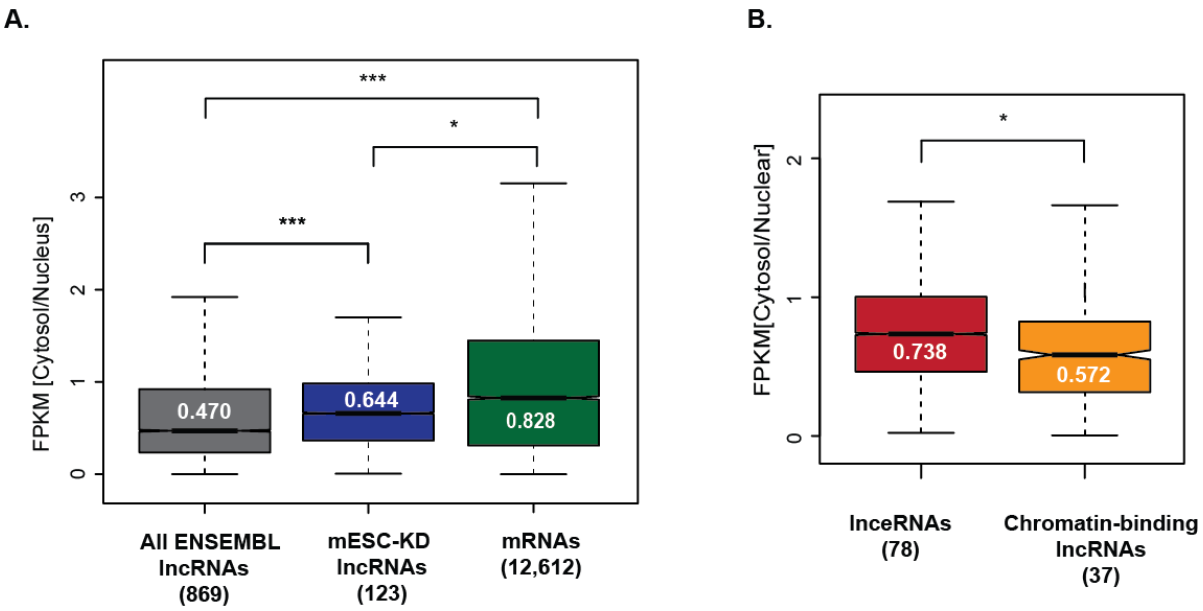


Figure 4.12 (A) ENSEMBL (build 70) annotated lincRNAs are significantly (p<9x10^{-4}, two-tailed Mann-Whitney test) more enriched in the nucleus (measured as the ratio between gene expression levels in the cytoplasmic and nuclear fractions, r=0.470, dark grey) than are the 123 mESC-expressed lincRNAs whose transcriptome-wide impact of knockdown was investigated by Guttman and colleagues (r=0.644, blue)(Guttman et al., 2011). Both ENSEMBL annotated lincRNAs and the mESC-expressed lincRNAs are significantly (p<0.05) more enriched in the nucleus than protein-coding mRNAs (r=0.828, green). (B) The 78 lnceRNAs (r=0.738, red) are significantly (p<0.05) more enriched in the cytosol than the 37 chromatin-binding lincRNAs (r=0.572, orange).

As expected, and relative to the remaining 45 lincRNAs with no evidence for miRNA-mediated regulatory roles (median r = 0.523), the 78 lnceRNAs were significantly more enriched in the cytosol (median r = 0.738, p<0.05, two-tailed Mann-Whitney test, Figure 4.13A, Table 4.5) with subcellular localizations that are statistically indistinguishable from those of mESC-expressed mRNAs (p=0.41, two-tailed Mann-Whitney test, Figure 4.13A).

Loss of miRNA biogenesis, and release from miRNA-mediated repression of transcripts, is expected to lead to increases in the cytoplasmic abundance of ceRNAt. Indeed, whilst their median nuclear abundances increased marginally, by only 3.3%, over 12 days, their cytoplasmic levels rose significantly, by 16.9% (p<2.2x10^{-16}, two-tailed Mann-Whitney test, Figure 4.13B, see section 4.3 Materials and Methods). For miRNA-independent targets of lnceRNAs (i.e. those that do not share predicted MREs or whose expression is not correlated with the lnceRNA), a small, but not significant 0.4% difference was observed between their nuclear and cytoplasmic abundances (p=0.06, Figure 4.13C). Similarly, a 0.6% non-significant difference was observed for targets of other non-ceRNA lincRNAs in the cytosol compared to the nucleus following loss of miRNA biogenesis (p=0.07, Figure 4.14A-B). These results are consistent with a model in which the observed coordinated up-regulation of ceRNAt is a consequence of the loss of post-transcriptional repression by miRNAs in the cytoplasm. The levels of lnceRNAs are also significantly increased in the cytoplasm relative to the nucleus (2.8%, p<2.2x10^{-6}, two-tailed Mann-Whitney test, Figure 4.14C).

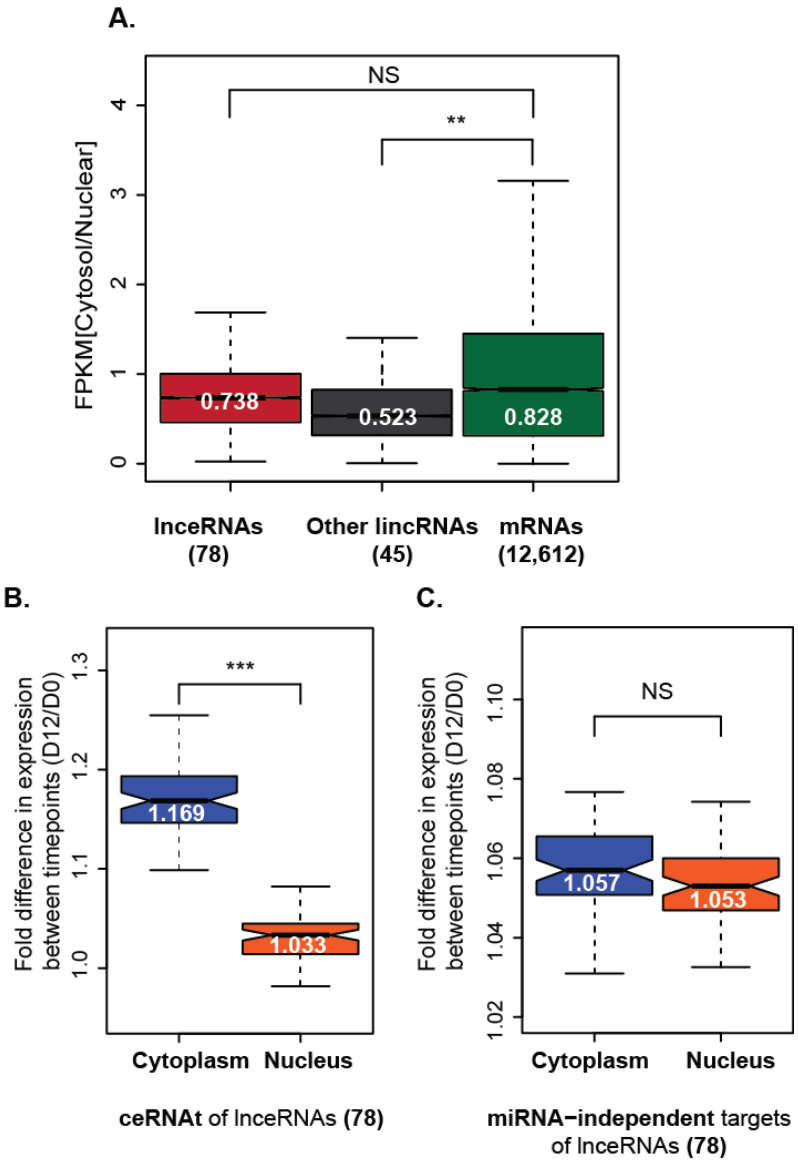


Figure 4.13 lnceRNAs are enriched in the cytoplasm. (A) LnceRNAs (n=78, red) have a similar subcellular distribution (measured as the ratio between gene expression (in FPKM) in the cytoplasmic and nuclear fraction, r = 0.738) to mESC-expressed mRNAs (ENSEMBL build 70, r = 0.828, green, p=0.41, two tailed Mann-Whitney test). That is in contrast with lincRNAs that are not coexpressed with their ceRNAt (n=45, grey), which are significantly more abundant in the nucleus (r = 0.523, p<0.01, two tailed Mann-Whitney test) than are mRNAs. The comparison between the relative abundance in the cytoplasm and nucleus of cells before (day 0) and after (day 12) *Dicer* loss-of-function revealed that ceRNAt (B) but not miRNA-independent targets (C) of lnceRNAs are significantly (p<2.2X10^{-16}, two tailed Mann-Whitney test) enriched in the cytoplasm (blue) relative to the nucleus (orange) of mESCs lacking miRNA biogenesis. Median fold differences are shown in the corresponding boxplot.

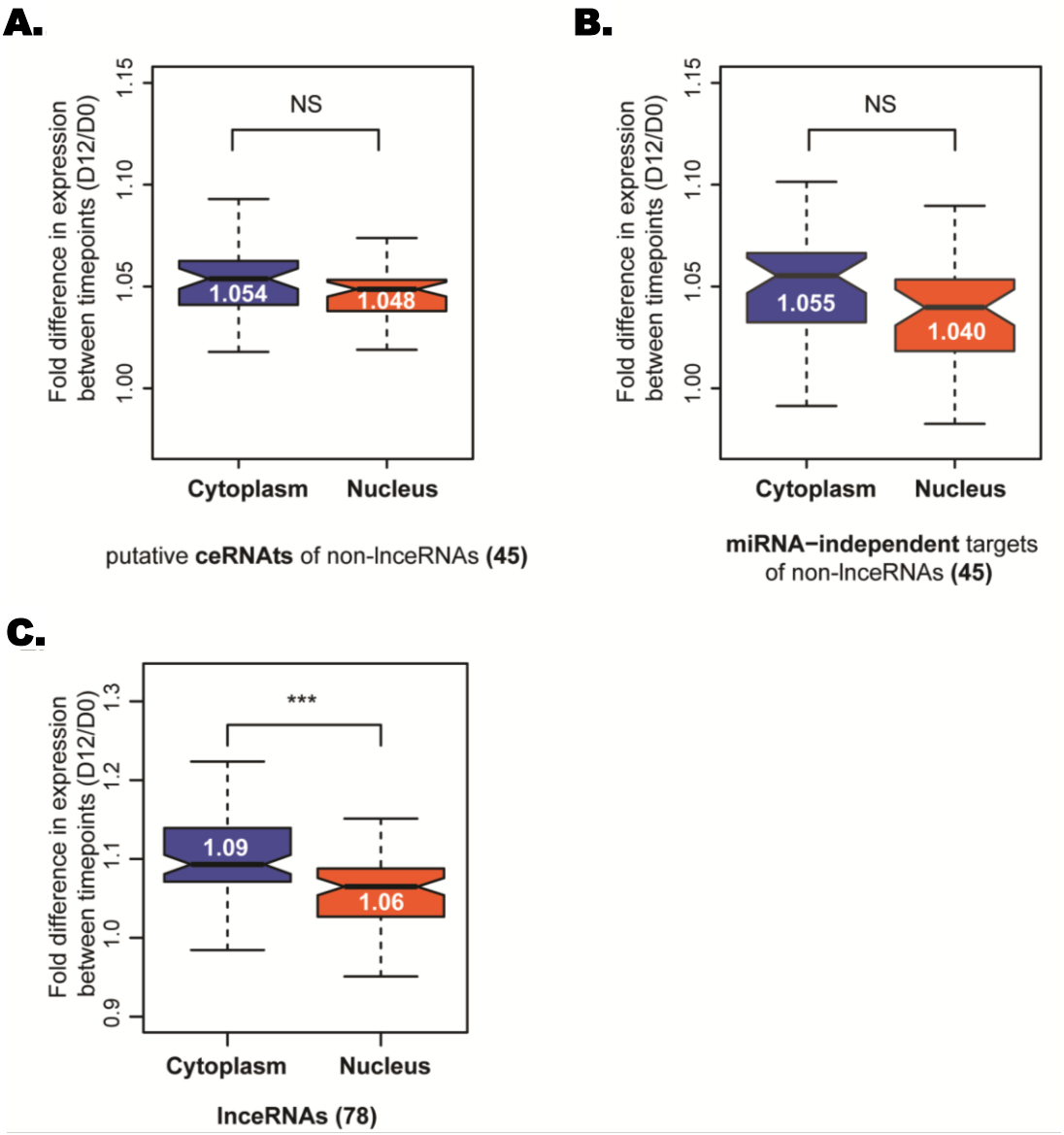


Figure 4.14 Gene expression patterns in the cytoplasm and nucleus of mESCs. (A and B) The comparison between the relative abundance in the cytoplasm (blue) and nucleus (orange) of cells before (day 0) and after (day 12) *Dicer* loss-of-function revealed that neither targets with ceRNAt signatures (putative ceRNAt) (A) nor miRNA-independent targets (B) of the lincRNAs not annotated as lnceRNAs (45 lincRNAs) are significantly enriched in a particular subcellular fraction (p=0.06 and p=0.07, respectively, two tailed Mann-Whitney test). (C) The levels of lnceRNAs are significantly ($p<3\times10^{-6}$, two-tailed Mann-Whitney test) increased in the cytoplasm (median=1.09) relative to the nucleus (median=1.06) on day 12 relative to that on day 0.

Table 4.5 Relative expression levels in the cytoplasmic and nuclear subcellular fractions of mESC lincRNAs. Relative expression levels of mESC-expressed lincRNAs, 78 lnceRNAs (red) and 45 non-lnceRNAs (grey), in the cytoplasmic (FPKM) and nuclear (FPKM) subcellular fractions.

lincRNAs	Ratio (Cytoplasmic FPKM/ Nuclear FPKM)	lincRNAs	Ratio (Cytoplasmic FPKM/ Nuclear FPKM)
linc1230	0.852	*linc1425*	0.462
linc1239	0.418	*linc1427*	1.650
linc1242	1.966	*linc1434*	0.644
linc1245	0.818	*linc1435*	0.563
linc1256	0.466	*linc1448*	0.281
linc1259	0.199	*linc1450*	0.329
linc1260	0.857	*linc1456*	1.373
linc1261	1.598	*linc1463*	0.990
linc1262	0.503	*linc1468*	0.750
linc1267	0.408	*linc1470*	1.069
linc1281	0.735	*linc1471*	0.931
linc1282	0.619	*linc1473*	0.022
linc1283	0.884	*linc1505*	1.186
linc1290	0.785	*linc1510*	0.493
linc1293	0.938	*linc1517*	0.767
linc1307	0.743	*linc1536*	0.493
linc1313	0.907	*linc1537*	0.530
linc1316	0.294	*linc1540*	0.264
linc1327	0.779	*linc1543*	1.688
linc1331	0.572	*linc1547*	0.269
linc1338	0.247	*linc1555*	0.741
linc1346	0.635	*linc1557*	0.541
linc1349	3.201	*linc1559*	0.853
linc1354	0.634	*linc1562*	0.280
linc1356	0.248	*linc1563*	1.269
linc1359	1.470	*linc1572*	0.228
linc1366	2.257	*linc1581*	1.002
linc1368	0.449	*linc1582*	1.226
linc1382	0.993	*linc1601*	0.657
linc1385	0.196	*linc1602*	2.421
linc1386	0.822	*linc1603*	1.209
linc1389	1.050	*linc1617*	0.559
linc1390	3.770	*linc1618*	0.431
linc1391	0.623	*linc1621*	0.652
linc1393	0.816	*linc1622*	0.232
linc1400	0.349	*linc1623*	0.482
linc1405	0.566	*linc1626*	1.110
linc1410	0.141	*linc1629*	1.329

lincRNAs	Ratio (Cytoplasmic FPKM/ Nuclear FPKM)
linc1631	0.982
linc1635	2.356
linc1244	0.050
linc1252	1.186
linc1253	0.890
linc1270	0.507
linc1296	0.693
linc1300	0.330
linc1304	0.207
linc1312	1.296
linc1315	1.150
linc1335	0.019
linc1347	0.391
linc1361	1.025
linc1369	0.350
linc1388	0.314
linc1411	0.868
linc1412	0.571
linc1413	0.690
linc1418	1.050
linc1419	0.623
linc1421	0.315
linc1422	0.138
linc1428	0.206

lincRNAs	Ratio (Cytoplasmic FPKM/ Nuclear FPKM)
linc1454	0.478
linc1477	0.271
linc1483	0.004
linc1490	0.775
linc1503	0.429
linc1524	0.533
linc1552	0.022
linc1558	0.495
linc1596	0.288
linc1599	0.628
linc1604	0.503
linc1607	0.308
linc1611	0.681
linc1612	0.663
linc1613	0.768
linc1614	0.800
linc1615	1.403
linc1616	0.824
linc1627	0.875
linc1630	0.332
linc1632	1.700
linc1633	1.111
linc1634	0.422

4.4.4 Recognition elements for miRNAs shared between lnceRNAs and ceRNAt have evolved under selective constraint in mammals

I reasoned that if the proposed miRNA-mediated regulation by lnceRNAs is an important layer of gene expression regulation, then their MRE sequences, in particular those shared between lnceRNAs and their ceRNAt, would show the signatures of purifying selection consistent with the action of natural selection to preserve this regulatory layer. Using publicly available polyA-selected RNA sequencing data for human H1 embryonic stem cells (Bernstein et al., 2012), I found evidence for conserved transcription (see section 4.3 Materials and Methods) for 59 of the 123 lincRNAs (48.0%). Of these, 64% (38) are lnceRNAs that I hereafter refer to as conserved lnceRNAs (Table 4.6). I then estimated the nucleotide substitution rate (d_{MRE}), between mouse and human, across response elements for mESC expressed and mammalian conserved miRNAs (62 miRNA families) predicted within the sequence of conserved lnceRNA (d_{MRE} = 0.352, Figure 4.15, see section 4.3 Materials and Methods). This rate was significantly and substantially suppressed compared with random samples of putatively neutrally-evolving sequence ($p<10^{-4}$, empirical *p*-value Figure 4.15). Next, I compared the rate estimated for MREs to what would be expected for non-MRE in conserved lnceRNA sequence that has been matched by length (see section 4.3 Materials and Methods). I also accounted for the observed difference in G+C content in predicted MREs (%G+C=41.3%) and non-MRE sequence (%G+C=42.4%, two-tailed Mann-Witney test, $p<8x10^{-5}$, Figure 4.16). The MRE nucleotide substitution rate normalized to a neutral rate (d_{MRE}/d_{AR} = 0.880) was significantly lower ($p<4x10^{-8}$, two-tailed Mann-Whitney test) than the

rate for non-MRE sequence (d_{nonMRE}/d_{AR} = 1.04, Figure 4.17A). This implies that MREs within conserved lnceRNAs evolved under stronger selective constraint than other lnceRNA regions. Consequently, despite the known low sensitivity of MRE prediction algorithms (Maziere and Enright, 2007), this unusually strong signature of purifying selection suggests that a fraction of the MRE sequences predicted are functional and conserved between mouse lnceRNAs and their human orthologs.

Finally, the substitution rates of lnceRNA MREs for miRNAs that are shared (median $d_{MRE\text{-}shared}/d_{AR}$ = 0.818) with their respective ceRNAt evolved under significantly greater constraint (p<2.2x10^{-16}, two-tailed Mann-Whitney test) than MREs within lnceRNAs that were not shared with their ceRNAt (median $d_{MRE\text{-}nonshared}/d_{AR}$ = 0.943, Figure 4.17B).

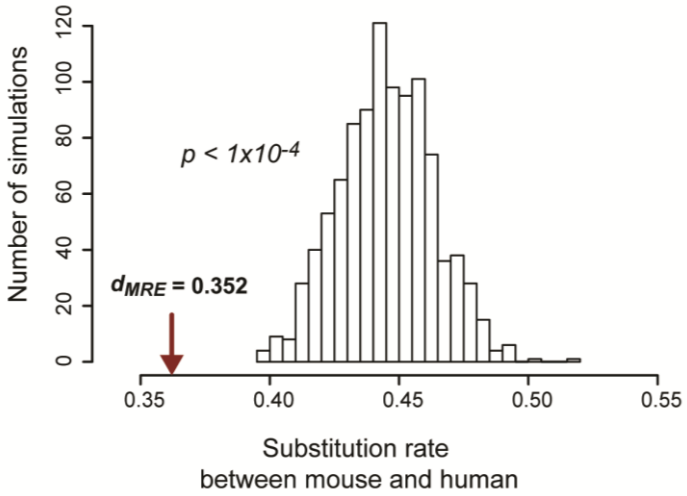


Figure 4.15 The nucleotide substitution rate between mouse and human lnceRNA's MREs is significantly lower (d_{MRE} = 0.352, p<1x10^{-4}, empirical *p*-value) than that of neighbouring neutrally evolving sequence with matching G+C content (ancestral repeats, ARs).

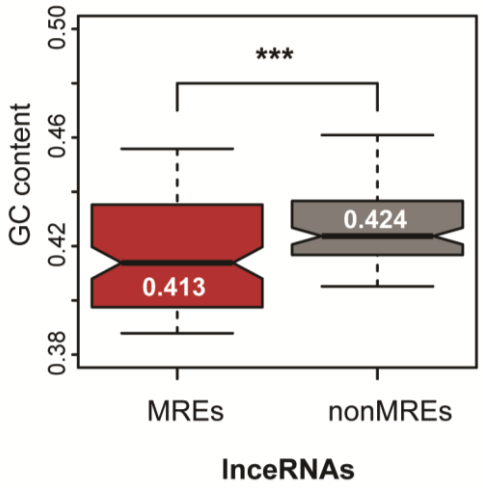


Figure 4.16. G+C content of MRE and nonMRE lnceRNA sequence. Significant difference (p<8x10^{-5}, two-tailed Mann-Whitney test) in G+C content between MREs (median = 0.413, red) and nonMREs (median = 0.424, grey).

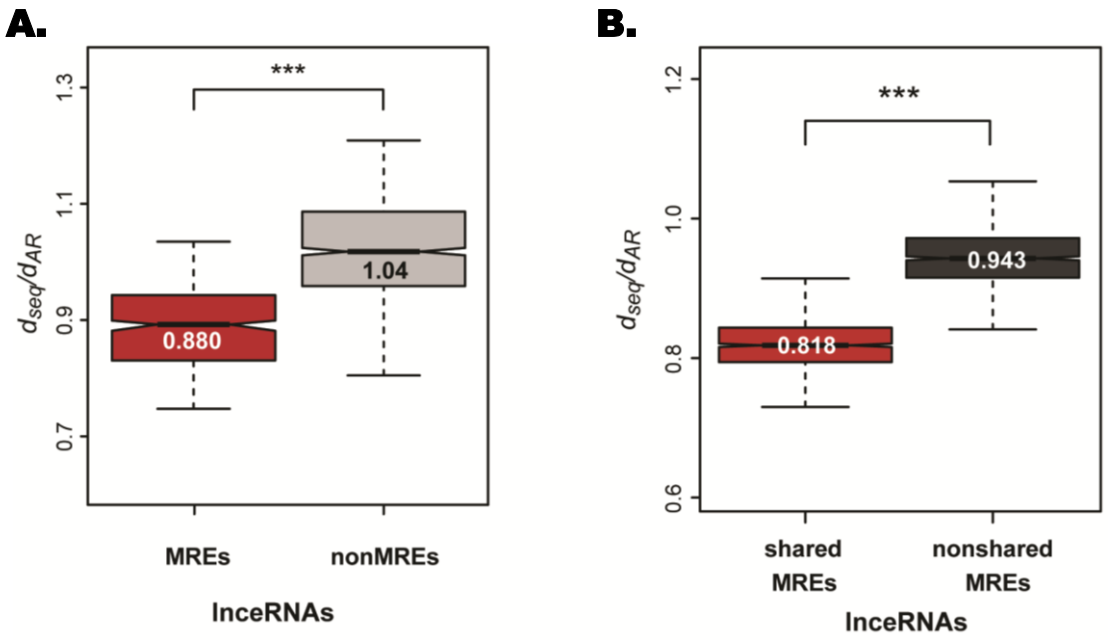


Figure 4.17 MiRNA response elements shared between lnceRNAs and ceRNAt are conserved through mammalian evolution. (A) The nucleotide substitution rate, between mouse and human, for predicted MREs within lnceRNAs (d_{MRE}/d_{AR}=0.880, red) is significantly lower (p<4x10^{-8}, two-tailed Mann-Whitney test) than for sequence within lnceRNAs not predicted to encode MREs (d_{nonMRE}/d_{AR} =1.04, grey). (B) The mouse-human substitution rate measured at MREs shared between lnceRNAs and their ceRNAt (red, $d_{MRE\text{-}shared}/d_{AR}$ = 0.818) is significantly lower (p<2.2x10^{-16}, two-tailed Mann-Whitney test) than that estimated at MREs for miRNAs not shared between lnceRNA and their targets (grey, $d_{MRE\text{-}nonshared}/d_{AR}$ = 0.943).

Table 4.6 Human conserved lnceRNAs. Genomic coordinates in mouse (mm9) human (hg19) of conserved lnceRNAs. Conserved lnceRNAs in human were estimated using human H1 embryonic stem cell RNAseq data (Bernstein et al. 2012) with at least 20% of the syntenic mouse transcript covered by 5 or more reads.

lnceRNA	Mouse lnceRNAs (mm9)	Human lnceRNAs (hg19)		
	Genomic coordinates	Genomic coordinates	Coverage (%)	Number of reads
linc1239	chr1: 162933175-162936722 (-)	chr1: 173868905-173872328 (+)	99.8%	2292
linc1256	chr11: 3539591-3548414 (-)	chr22: 31365894-31374840 (+)	100.0%	78132
linc1262	chr11: 82593024-82594596 (-)	chr17: 33303965-33307369 (-)	40.7%	33
linc1267	chr11: 120049396-120050818 (-)	chr17: 79317798-79319478 (-)	55.8%	23
linc1270	chr12: 72110268-72116930 (-)	chr14: 58763633-58765198 (-)	21.0%	310
linc1283	chr13: 98225757-98285626 (-)	chrX: 39646871-39647311 (-)	42.9%	168502
linc1290	chr15: 32170378-32174520 (-)	chr5: 9546399-9550350 (+)	100.0%	1256
linc1307	chr17: 35086206-35090401 (-)	chr6: 31802971-31805059 (+)	28.8%	16381
linc1316	chr18: 75518147-75526822 (-)	chr18: 46478649-46489126 (+)	20.2%	12
linc1327	chr2: 157368964-157373662 (-)	chr20: 36128472-36133956 (-)	90.0%	837
linc1331	chr2: 173421716-173484958 (-)	chr20: 56822228-56884326 (-)	98.9%	1275
linc1338	chr4: 90878362-90895209 (-)	chr9: 23643106-23662926 (-)	74.8%	38
linc1349	chr5: 33173807-33174458 (-)	chr22: 32112459-32113228 (-)	100.0%	270
linc1354	chr5: 100848961-100859422 (-)	chr4: 83816844-83821738 (-)	35.6%	1464
linc1359	chr6: 13804903-13847079 (-)	chr7: 112756787-112758795 (-)	88.1%	3559
linc1382	chr7: 49727950-49764427 (-)	chr7: 64437360-64439914 (-)	89.6%	774
linc1385	chr7: 67113886-67120883 (-)	chr15: 25240479-25247512 (+)	100.0%	11082
linc1389	chr7: 144566473-144603237 (-)	chr10: 131861645-131909092 (-)	60.4%	177
linc1393	chr8: 67551222-67562049 (-)	chr4: 165798234-165817519 (+)	90.0%	11990
linc1410	chrX: 50407668-50410215 (-)	chrX: 133681289-133683625 (-)	24.5%	18

lnceRNA	Mouse lnceRNAs (mm9)	Human lnceRNAs (hg19)		
	Genomic coordinates	Genomic coordinates	Coverage (%)	Number of reads
linc1425	chr1: 162965054-162969894 (+)	chr1: 173833395-173837295 (-)	52.5%	60570
linc1434	chr10: 81831057-81843206 (+)	chr19: 12575114-12576347 (-)	34.0%	14
linc1435	chr10: 83185191-83226402 (+)	chr12: 105724445-105765295 (+)	26.1%	443
linc1463	chr13: 62364089-62378052 (+)	chr1: 247319428-247320004 (-)	37.3%	218
linc1510	chr2: 119420000-119433459 (+)	chr15: 41576172-41598812 (+)	99.3%	27296
linc1526	chr4: 129488507-129489486 (+)	chr1: 32409773-32409933 (-)	47.8%	5
linc1540	chr5: 104946203-104983885 (+)	chr4: 88957482-89000621 (+)	21.3%	1488
linc1547	chr6: 52052320-52069791 (+)	chr7: 27030124-27071533 (+)	27.0%	93
linc1559	chr6: 133055321-133057678 (+)	chr12: 11323821-11325543 (+)	94.8%	4721
linc1563	chr7: 28595377-28608464 (+)	chr19: 40579461-40587825 (-)	75.0%	241
linc1572	chr7: 88673718-88676345 (+)	chr15: 83423427-83425889 (+)	81.2%	190
linc1582	chr8: 89996652-90050351 (+)	chr16: 49316017-49372824 (+)	48.3%	25
linc1601	chr11: 63893124-63896826 (+)	chr17: 13964525-13965136 (-)	86.3%	23
linc1603	chr11: 95720067-95721703 (+)	chr17: 47268930-47270262 (-)	35.8%	15
linc1618	chr2: 84581863-84584145 (-)	chr11: 57405837-57407497 (+)	100.0%	613
linc1623	chr4: 99320307-99322058 (-)	chr1: 63785736-63787490 (-)	87.5%	2795
linc1626	chr5: 115653674-115655290 (+)	chr12: 121067865-121069099 (-)	90.0%	63
linc1631	chr7: 73004067-73006879 (-)	chr15: 102030125-102032703 (+)	66.8%	42

4.4.5 ceRNAt of individual lnceRNAs tend to be functionally related

Finally, I investigated whether miRNA-dependent regulation by each one of the 78 lnceRNAs preferentially affects transcripts of functionally-related genes. For this, I took advantage of an integrative phenotypic-linkage network of mouse protein-coding genes (Honti, 2014). This network integrates gene-gene linkage information from diverse and complementary sources including gene coexpression, protein physical interaction, co-citation and gene functional annotation data. Relative to networks built using individual data types, this network exhibits improved coverage and accuracy (Honti, 2014). For each lnceRNA, I estimated the average link weight between genes classified as either ceRNAt or miRNA–independent targets; a higher link weight reflects the increased likelihood of two genes in the network being functionally related (Honti, 2014).

Strikingly, lnceRNAs' ceRNAt were found to be substantially more closely related to each other (median of average link weights = 0.718) than were miRNA-independent transcripts (median of average link weights = 0.225, two-tailed Mann-Whitney test, $p<5.3\text{x}10^{-5}$, Figure 4.18A). Similar results were obtained using another measure of functional relatedness, the sum of the linkage weights, after subsampling to the same target group size (Figure 4.18B, see section 4.3 Materials and Methods). These results argue that coordinated miRNA-mediated modulation of gene expression levels by lnceRNAs tends to affect predominantly functionally related protein-coding genes. This finding, together with the evolutionary constraint observed for shared MREs, argue that miRNA-mediated crosstalk is important for many normal biological processes.

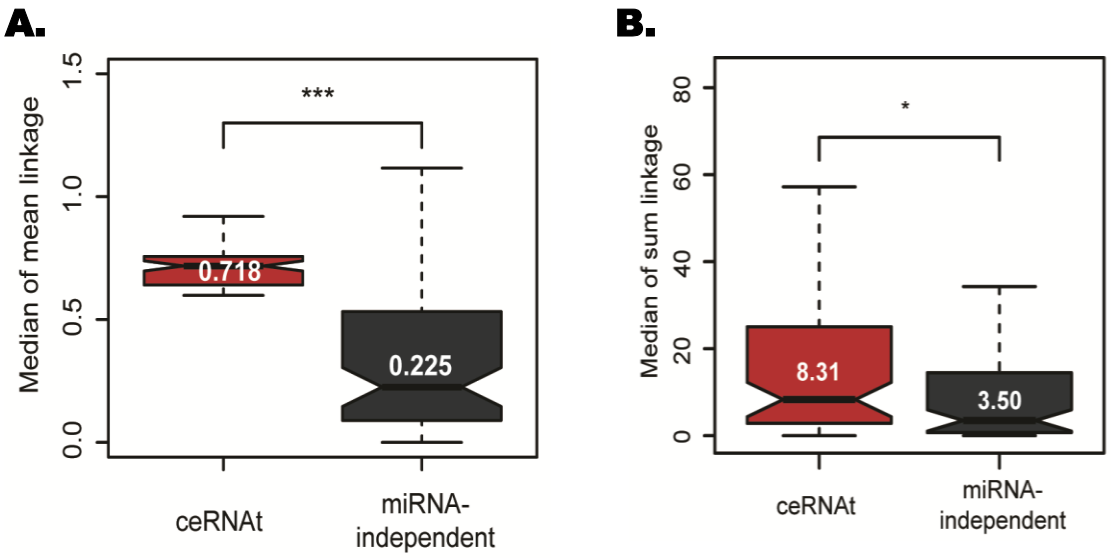


Figure 4.18. Functional relatedness of lnceRNA targets. (A) LnceRNAs' ceRNAts are significantly more functionally related (median of mean linkage=0.718, red, $p<6x10^{-5}$, two-tailed Mann-Whitney test) than are miRNA-independent targets (median of mean linkage=0.225, dark grey) in an integrated functional network (Honti, 2014). (B) LnceRNAs' ceRNAts (median of sum linkage=8.31, red) are significantly more functionally related (*, p=0.045, two-tailed Mann-Whitney test) than are miRNA-independent targets (median of mean linkage=3.50, dark grey) using the integrated functional network.

mESC-expressed lnceRNAs have, on average, 16.4 predicted MREs per kb of transcript that are specific to, on average, 13.9 different mESC-expressed miRNAs. This MRE density is over 5-fold higher than the density in 3' UTRs of their target mRNAs (2.9 MREs predicted per kb; $p<2.2x10^{-16}$, two-tailed Mann-Whitney test, Figure 4.19). A single lnceRNA might, therefore, be more likely than a mRNA to regulate post-transcriptionally the transcript abundance of many mRNAs via crosstalk with many miRNAs.

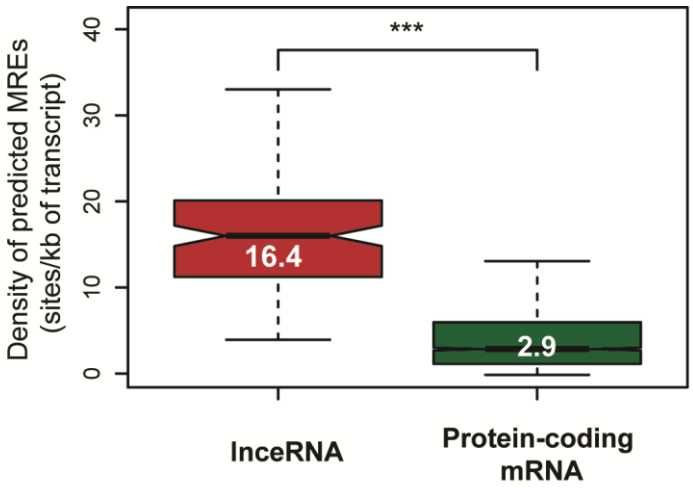


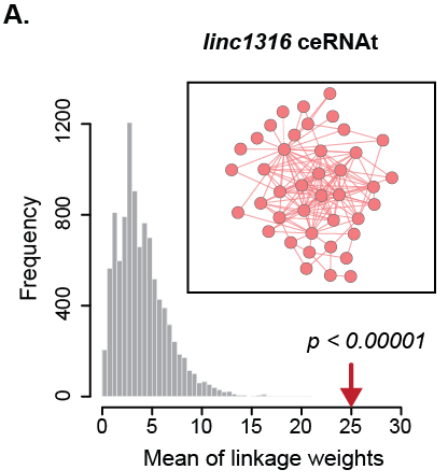
Figure 4.19 Distribution of MREs for lnceRNAs and protein-coding genes. LnceRNAs (red) have a significantly higher density of predicted MREs for mESC-expressed miRNAs than have protein-coding gene 3'UTRs (green).

I next sought to identify lnceRNAs that post-transcriptionally regulate the abundance of functionally related mRNAs. To do so, for each lnceRNA, I compared the mean of link weights for all its target transcripts to that expected from simulated data (see section 4.3 Materials and Methods). For 39 of 78 (50%) lnceRNAs I found that their ceRNAt were significantly more functionally related (p<0.05, empirical p value) than that expected from the simulated data.

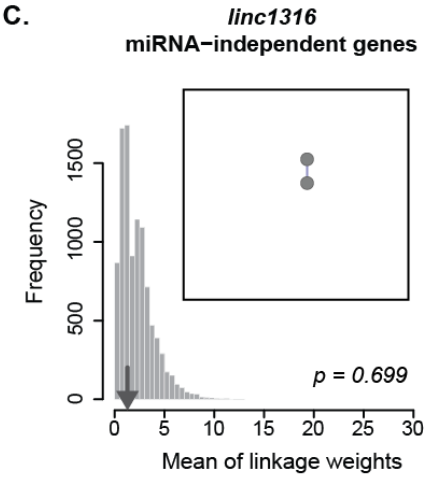
These mRNA targets were enriched in genes involved in the regulation of cell cycle, developmental process, cell signaling/communication or regulation of differentiation (Table 4.7) which is consistent with the originally proposed roles of these lnceRNAs in regulating the circuitry underlying cell state decision in

mESCs (Guttman et al., 2011). For example, depletion of *linc1316* led to changes in the levels of 62 ceRNAt and 57 miRNA-independent targets. This lnceRNA's ceRNAt are strongly functionally inter-related (Figure 4.20A) and are enriched in genes involved in the regulation of cellular signaling, developmental process and the cell cycle (Figure 4.20B). In contrast, miRNA-independent targets of *linc1316* are not functionally related and their annotations are not enriched in any particular GO Biological Process (Figure 4.20C) (Ashburner et al., 2000). Interestingly, *linc1316* harbours predicted binding sites for well-known miRNAs involved in the maintenance of ESC pluripotency, including *miR-290-295* and *miR-200* families (Peter, 2009; Melton et al., 2010; Lichner et al., 2011); and is likely to post-transcriptionally regulate transcriptional factors that promote stem cell self-renewal, such as *Hmga2* (Nishino et al., 2008) and *Myc* (Singh and Dalton, 2009) that are down-regulated upon its knockdown.

A.



C.



B.

GO Term	Description	Enrichment	FDR q-value
GO:0010468	regulation of gene expression	3.55	1.7×10^{-8}
GO:0031323	regulation of cellular metabolic process	2.80	6.5×10^{-8}
GO:0007389	pattern specification process	9.15	5.1×10^{-6}
GO:0007154	cell communication	8.07	7.0×10^{-4}
GO:0045595	regulation of cell differentiation	3.52	8.9×10^{-3}
GO:2000027	regulation of organ morphogenesis	9.86	1.4×10^{-2}
GO:0051726	regulation of cell cycle	4.44	2.9×10^{-2}

Figure 4.20 ceRNAt, but not miRNA-independent targets, of lnceRNA are functionally related. (A) The mean linkage weights of ceRNAt of *linc1316* is significantly higher ($p<1 \times 10^{-5}$, empirical *p* value) than expected based on 10,000 sets each with an equal number of randomly selected mESC-expressed genes. The insert panel illustrates the connectivity of functional similarities (red edges) within miRNA-dependent target genes (red nodes). (B) Table illustrating Biological Processes Gene Ontology (GO) annotations that are significantly enriched within the ceRNAt of *linc1316* relative to a background of mESC-expressed genes. (C) MiRNA-independent targets of *linc1316* are not significantly ($p=0.699$, empirical *p* value) different in functional similarity than expected based on 10,000 sets with an equal number of randomly selected mESC-expressed genes.

Table 4.7 Significantly enriched ceRNAt GO annotations.

Category	GO Term	Fold Enrichment	Bonferroni corrected *p* value
GO:0042128	nitrate assimilation	2.58	4.2E-18
GO:0048569	post-embryonic organ development	2.11	1.8E-02
GO:0050920	regulation of chemotaxis	2.02	4.9E-03
GO:0007093	mitotic cell cycle checkpoint	1.98	1.2E-02
GO:0006413	translational initiation	1.81	3.0E-02
GO:0022613	ribonucleoprotein complex biogenesis	1.75	6.6E-12
GO:0007059	chromosome segregation	1.64	5.7E-06
GO:0006259	DNA metabolic process	1.58	8.4E-54
GO:0034660	ncRNA metabolic process	1.44	6.4E-06
GO:0016055	Wnt receptor signaling pathway	1.42	3.0E-05
GO:0045596	negative regulation of cell differentiation	1.39	9.8E-07
GO:0051301	cell division	1.38	4.2E-08
GO:0048285	organelle fission	1.38	2.2E-04
GO:0060541	respiratory system development	1.36	2.9E-02
GO:0007423	sensory organ development	1.36	5.9E-06
GO:0055114	oxidation reduction	1.34	5.5E-26
GO:0033554	cellular response to stress	1.31	1.8E-13
GO:0007049	cell cycle	1.31	1.8E-13
GO:0007389	pattern specification process	1.28	1.8E-04
GO:0003002	regionalization	1.27	1.9E-02
GO:0046907	intracellular transport	1.26	4.5E-05
GO:0006396	RNA processing	1.25	1.3E-02
GO:0051716	cellular response to stimulus	1.24	1.7E-09
GO:0006915	apoptosis	1.21	8.5E-03
GO:0050793	regulation of developmental process	1.16	2.1E-02
GO:0048519	negative regulation of biological process	1.13	5.2E-05
GO:0009987	cellular process	1.05	3.5E-25
GO:0044237	cellular metabolic process	1.05	9.6E-07

4.5 DISCUSSION

Here, I have used experimental and computational genomics approaches to investigate the prevalence and properties of lnceRNAs. I focused my analyses on lincRNAs with proposed roles in the circuitry underlying pluripotent and differentiated cell states. Not only have most lnceRNAs described in this context been of this critical cellular transition, but their impact on gene expression profiles, when pools of miRNAs are limited and changes in their repertoires can lead to repression or activation of transcriptional programs, is likely to be greater than in fully differentiated cells or in homeostasis. Environmental or cellular stress, for example upon starvation or infection (Ebert et al., 2007; Franco-Zorrilla et al., 2007), may also provide similar opportunities for strong effects of small changes in target gene expression.

I integrated publicly available data on the transcriptome-wide impact of depleting for over 140 lincRNAs, in mESCs (Guttman et al., 2011) with in-house RNA-sequencing profiles of mESCs following conditional loss of *Dicer*, a key component of the miRNA biogenesis pathway. My analysis indicates that around a fifth of the expression changes induced by the knockdown of over 60% of these lincRNAs are miRNA-dependent. Furthermore, I predict that 87% of lincRNAs share miRNAs binding sites with transcription factor encoding mRNAs (Zhang et al., 2012) that are also down-regulated upon lincRNA knockdown, suggesting that the changes induced via this miRNA-mediated mechanism can lead to secondary effects of transcriptional regulation.

The properties of lnceRNAs and their interactions with their mRNA targets are consistent with established rules of post-transcriptional regulation by miRNAs. Namely, lnceRNAs are enriched in the cytoplasm and their interactions with their functionally related mRNA targets are dependent on the presence of miRNAs. The over 5-fold higher density, relative to 3'UTRs, of predicted MREs within lnceRNAs argues for these transcripts' enhanced ability to modulate their target levels post-transcriptionally. The increased evolutionary sequence constraint within MREs, in particular those shared between lnceRNAs and their ceRNAt, implies the conservation of these transcripts regulatory roles, suggesting they have important biological relevance.

Together, my results are consistent with a high prevalence of miRNA-mediated interactions between lincRNAs, particularly those enriched in the cytoplasm, and their mRNA targets. These findings suggest that this mechanism of lincRNA function, which hitherto has been relatively poorly studied, deserves a more prominent position to aid the understanding of cell fate determination.

CHAPTER 5

Crosstalking noncoding RNAs contribute to cell-specific neurodegeneration in Spinocerebellar ataxia type 7

5.1 ABSTRACT

What causes the tissue-specific pathology of genetic diseases resulting from mutations in ubiquitously expressed housekeeping genes? Specifically, in Spinocerebellar ataxia type 7 (SCA7), a neurodegenerative disorder caused by a CAG repeat expansion in *ATXN7* – an essential component of the mammalian transcription co-activation complex, STAGA – the factors underlying the characteristic progressive cerebellar ataxia and degeneration of the retinal macula observed in patients were unknown. I found that the ATXN7/STAGA complex is required for the transcription initiation of *miR-124*, the most abundant neuronal miRNA, which in turn mediates the post-transcriptional crosstalk between *lnc-SCA7,* a conserved long noncoding RNA, and *ATXN7*. In SCA7, mutations in *ATXN7* disrupt these regulatory interactions and result in a neuron-specific increase in *ATXN7* abundance. Strikingly, this increase is most prominent in the SCA7 disease-relevant tissues, namely the retina and cerebellum. These results demonstrate how noncoding RNA-mediated feedback regulation of a ubiquitously expressed housekeeping gene may contribute to specific neurodegeneration.

5.2 INTRODUCTION

Spinocerebellar ataxia type 7 (SCA7) is a rare inherited neurodegenerative disease that affects approximately 1 in 100,000 births (Gouw et al., 1998). Like other spinocerebellar ataxias, SCA7 is characterized by uncoordinated movement, abnormal gait, dysarthria, and dysphagia, due to selective neuronal death of Purkinje cells in the cerebellum (David et al., 1997). Unique to SCA7 is the degeneration of the retinal macula that leads to gradual loss of sight and eventually blindness (Aleman et al., 2000). SCA7 is caused by an in-frame CAG tri-nucleotide repeat expansion in the first coding exon of the Ataxin type 7 gene, *ATXN7* (David et al., 1997). Translation of the mutated *ATXN7* allele results in a polyglutamine (polyQ) tract expansion, the formation of protein aggregates and decreased protein activity (Holmberg et al., 1998).

ATXN7 is an essential component of the mammalian STAGA multi-protein complex (Helmlinger et al., 2004) whose chromatin remodeling activity facilitates transcriptional activation of multiple loci. *ATXN7* expression is not specific to the brain, but instead is found at similarly high levels in many non-neuronal tissues, such as the kidney, liver, and lung (David et al., 1997; Cancel et al., 2000). Why mutations in this ubiquitously expressed housekeeping gene lead to the degeneration of only retinal and cerebellar neurons has remained unresolved.

I investigated the hypothesis that regulation by noncoding RNAs, whose expression pattern is often spatially and developmentally restricted (Mattick, 2009), might contribute to SCA7's tissue-specific pathology. For instance,

aberrant expression of miRNAs has been previously associated with disease (reviewed in Sayed and Abdellatif, 2011): decreased expression of *miR-9/miR-9** has been found in the cortices of Huntington's disease patients (Packer et al., 2008), whereas inhibition of *miR-19*, *miR-101*, and *miR-130* increased the cytotoxicity of polyQ-expanded ATXN1 in Spinocerebellar ataxia type 1 (Lee et al., 2008). The importance of miRNAs for the survival of neuronal subtypes is clear from the observation that global loss of miRNAs in mouse, achieved by conditional knockout of *Dicer* (a gene required for miRNA biogenesis), led to progressive degeneration of specific types of neurons in the CNS, including retinal and cerebellar cells (Schaefer et al., 2007; Damiani et al., 2008). These particular phenotypes are driven, at least in part, by the loss of *miR-124*, the most abundantly expressed miRNA in the CNS (Lagos-Quintana et al., 2002) because targeted deletion of this miRNA partially phenocopied *Dicer* knockout and resulted in increased retinal and neuronal cell degeneration (Sanuki et al., 2011), the two cell types with the highest *miR-124* expression.

Similarly to miRNAs, long (>200 nucleotides) noncoding RNAs (lncRNAs), whose expression is often spatially and temporally restricted (Cabili et al., 2011; Derrien et al., 2012), have also been previously associated with neurodegeneration (Qureshi et al., 2010). Spinocerebellar ataxia type 8 (*SCA8*) (Koob et al., 1999), for instance, is caused by toxicity of a CUG tri-nucleotide repeat expansion in the lincRNA antisense to *ATXN8* as well as by loss-of-function of its overlapping polyQ-expanded mutant ATXN8 protein (Daughters et al., 2009). As in SCA8, ribonuclear inclusions have been found in other polyQ disorders, such as myotonic dystrophy and Huntington's Disease (La Spada and Taylor, 2010), suggesting that repeat expanded RNA may be

pathogenic. Long noncoding RNAs can also contribute to disease by regulating the transcript abundance for their overlapping disease-causing gene as found for *BACE1-AS*, in Alzheimer's disease (Faghihi et al., 2010), and for *SCAANT1*, a noncoding transcript overlapping *ATXN7* (Sopher et al., 2011). Several intergenic lncRNAs have been found to be dysregulated, compared to controls, in patients with Huntington's (Bithell et al., 2009) or Alzheimer's (Mus et al., 2007) diseases, and an intergenic lincRNA (lincRNA) was recently found to be correlated in expression with a strong intergenic risk allele for Parkinson's disease (Kumar et al., 2013). However, whether such associations reflect causal contributions by lincRNAs or else are consequences of the disease pathology remains undetermined.

I sought to establish the origin of the tissue-specific pathology of SCA7 by investigating the regulation of *ATXN7*, a ubiquitously expressed gene, by tissue-specifically expressed noncoding RNAs. My results demonstrate that the regulation of mutant ATXN7 abundance through crosstalking noncoding RNAs that are highly specific to the retina and the cerebellum contributes to the selective neurodegeneration observed in SCA7.

5.3 MATERIALS and METHODS

I performed all the work described below, except where noted otherwise.

Tissue culture

Mouse and human neuroblastoma (N2A and SH-SY5Y) cells, mouse DTCM23/49 XY embryonic stem (ES) cell lines (mESCs), human retinoblastoma (WERI) and fibroblast cells lines were cultured as described in **Chapter 2**.

The conditional and Dicer-deficient ES cells used in works included in this chapter were routinely maintained on a feeder layer of mitomycin-inactivated mouse primary embryonic fibroblasts. Prior to the analyses, feeder cells were depleted from the cultures by pre-plating trypsinized cells for 25 minutes and transferring ES-enriched cell suspension to a new gelatine-coated plate in a culturing medium supplemented with Leukemia Inhibitory Factor (Invitrogen) and 2i inhibitors (CHIR99021 at 3nM and PD0325901 at 1nM, Stemgent). After removal of the RNase III domain of Dicer, cells were kept for no longer than 6 passages.

SCA7 patient-derived fibroblast cell lines ($SCA7^{42Q/10Q}$, $SCA7^{49Q/10Q}$, $SCA7^{55Q/10Q}$) were maintained by Dr. Lauren Watson and Dr. Miguel Varela as described in **Chapter 2**.

Human and mouse gene expression profiling

Microarray gene expression data for *ATXN7* and *lnc-SCA7*, also known as *ATXN7L3B*, were obtained from Gene Expression Atlas (GNF) through BioGPS (http://biogps.org) for human (Su et al., 2004), and their correlation coefficient (Pearson's correlation, R^2) was computed across all available tissues or cells where both loci were expressed (AD>20; Normalized Affymetrix microarray expression values using GCRMA algorithm (Su et al., 2004)).

The relative expression levels of *Atxn7*, *lnc-SCA7*, *miR-124* are profiled in 20 human normal adult tissues, 11 mouse normal adult tissues, 9 (postnatal day 5) mouse brain tissues and human and mouse retinal cells as described in **Chapter 2**.

Western blotting

The relative levels of Atxn7 as result of *lnc-SCA7* expression level perturbation and the protein-coding ability of *lnc-SCA7* were tested by western blotting in N2A cells as described in **Chapter 2**. Primary antibodies used for Atxn7 and lnc-SCA7 detection were anti-Atxn7 (sc-21110, Santa Cruz Biotechnology, working dilution 1:200) and custom anti-lnc-SCA7/Atxn7l3 antibody (Amsbio, working dilution 1:100), respectively. Subsequently, biotinylated secondary antibodies were used for Atxn7 and lnc-SCA7/Atxn7l3 [ab6884, ab7089 (Abcam) respectively, working dilution 1:5,000].

Knockdown and over-expression constructs

Three knockdown small interfering RNA (siRNA) constructs, *si-lnc-SCA7s*, were designed to specifically target *lnc-SCA7* in mouse by selecting regions without substantial sequence similarity to either *Atxn7l3* or *Atxn7* as described in **Chapter 2**. As a control, I used an oligo with randomly permutated nucleotides (scrambled control) that showed no significant sequence similarity to mRNAs from the mouse genome. Oligos used are listed in Table 2.4.

Multiple short hairpin RNAs (shRNAs) generated to knockdown *lnc-SCA7* were made as described in **Chapter 2**. If all shRNAs tested were associated with decreased levels of *Atxn7*, the shRNA construct that had the greatest impact on *lnc-SCA7* levels was used in subsequent experiments. Transfection of N2A cells with the shRNA construct containing the randomly permutated control sequence (*sh-scramble*) was used as control.

Three stable *lnc-SCA7* knockdown and control N2A polyclonal cell lines were independently derived by co-transfection of the *sh-lnc-SCA7* knockdown and *sh-scramble* constructs with *pTK-Hyg*, a plasmid that contains the hygromycin-resistant gene (courtesy of Dr. Keith Vance). The *pTK-Hyg* vector is used as positive control for the integration of plasmid into the cell genome. Cells were grown in hygromycin-containing medium (Invitrogen, 200µg/mL) until high confluence was reached (after approximately 10 days). The medium was changed every 48 hours.

Overexpression constructs of *Atxn7* and *lnc-SCA7* were generated as described in **Chapter 2**. *Atxn7* 3' untranslated region (UTR) (nucleotide 2,876-6,582, ENST00000022257, ENSEMBL Build 70, *Atxn7*-WT) and the putative 3' noncoding untranslated sequences of mouse *lnc-SCA7* (nucleotide 599 – 3,607, *lnc-SCA7*-WT), the full-length mouse *lnc-SCA7* (*lnc-SCA7*-FULL) as well as *lnc-SCA7*-STOP were cloned downstream of a CMV promoter on the *pcDNA3.1(+)* vector. *lnc-SCA7*-STOP was generated by direct mutagenesis of position 14 of *lnc-SCA7*-FULL from C->A.

To disrupt *miR-124* binding sites within *lnc-SCA7*-WT and *Atxn7*-WT, all *miR-124* MRE regions (6 and 2 in total, respectively) complementary to the miRNA seed, 5'-TGCCTT-3', within these constructs were mutated by reversing the sequence, using direct mutagenesis, to 5'-TTCCGT-3' as described in **Chapter 2**. Empty *pcDNA3.1(+)* vector was used as transfection control in over-expression experiments.

Construct designs are illustrated in Table 2.1. Primers used are listed in Table 2.4.

Luciferase assays

lnc-SCA7 and *Atxn7* 3' UTRs were cloned into the *XbaI* restriction site downstream of the luciferase reporter gene in the *pGL3-promoter* (*pGL3-pro*) vector (Promega) as described in **Chapter 2**: *luc-lnc-SCA7*-WT and *luc-Atxn7*-WT, respectively. *luc-lnc-SCA7*-MUT and *luc-Atxn7*-MUT were generated by directed mutagenesis of all *miR-124* binding sites within *lnc-SCA7* (6) and

Atxn7 (2). All *miR-124* MRE regions complementary to the miRNA seed, 5'-TGCCTT-3', within the two transcripts were mutated by changing the sequence to 5'-TTCCGT-3'. Each luciferase construct (2 µg) was co-transfected with 10ng of pRL-*Renilla* luciferase control vector (Promega) and 50 nM mirVana *miR-124* or negative miRNA control mimics (miR-NC) (Invitrogen) using the FuGENE 6 Transfection Agent (Roche) as described in **Chapter 2**.

Putative promoter elements of the three *pri-miR-124s* (*miR-124_1*: chr14:65,205,705 – 65,207,200; *miR-124_2*: chr3:17,694,143 – 17,695,600; *miR-124_3*: chr2:180,627,439 – 180,628,900, mm9) and Negative control (NC chr14:65,183,839 – 65,185,271, mm9) were cloned into restriction sites (*miR-124_1*: *HindIII* and *KpnI*; *miR-124_2*: *HindIII* and *XhoI*; *miR-124_3*: *BglII*; NC: *HindIII* and *XhoI*) upstream of the luciferase reporter gene in the *pGL3-enhancer* (*pGL3-enh*) vector (Promega): *miR-124-1-prom-luc, miR-124-2-prom-luc miR-124-3-prom-luc* and *NC-promr-luc*. As control, the same sequences were cloned into the same location in the reverse orientation. As described in **Chapter 2**, each luciferase construct (2 µg) was similarly co-transfected with 10 ng of pRL-*Renilla* luciferase control vector (Promega) using the FuGENE 6 Transfection Agent (Roche).

Construct designs are illustrated in Table 2.1. Primers used are listed in Table 2.4.

Prediction of miRNA response elements

Predicted miRNA response elements (MREs) for the human *lnc-SCA7* (*ATXN7L3B*), *ATXN7* and the other 22 STAGA-subunit encoding mRNAs (obtained from neXtProt.com (Lane et al., 2012)) were downloaded from microRNA.org (all mirSVR scores) (Betel et al., 2008). The observed percentage of shared MREs between *lnc-SCA7* and STAGA encoding subunits was compared to the fraction of shared MREs found across 10,000 randomly selected sets of brain-expressed 23 genes that are not part of the STAGA complex (all mirSVR scores, microRNA.org). An empirical *p*-value was calculated by comparing the number of MREs shared between *lnc-SCA7* and the STAGA mRNAs with those shared between *lnc-SCA7* and the randomized sets of 23 brain-expressed genes. MREs predicted to be shared between *lnc-SCA7* and STAGA encoding mRNAs or randomly selected mRNAs were represented using Circos plots (Krzywinski et al., 2009) (circos.ca) where each node represents a transcript and each edge represents the percentage of shared MREs for a miRNA.

Genome-wide analysis of miRNA abundance

N2A cells were transfected with *sh-lnc-SCA7* or scrambled control (1 µg) and *lnc-SCA7*-WT, *lnc-SCA7*-MUT over-expression constructs or *pcDNA3.1(+)* empty vector control (1 µg). Genome-wide miRNA abundance from total RNA extracted from the transfected cells were assayed as described in **Chapter 2**.

Specifically, no significant difference in N2A's miRNA repertoires was detected following either knockdown or overexpression of *lnc-SCA7*. Although changes in *miR-124* levels were not genome-wide significant, the trends observed by qRT-PCR could be replicated in this experiment. I attribute the lack of genome significance to reduced power to detect transcripts expressed at *miR-124* levels using the standard protocol. I do not detect differences for highly abundant miRNAs, consistent with no global changes in miRNA repertoires.

Absolute quantification of lincRNA and mRNA abundance

Digital droplet PCR (ddPCR), as described in **Chapter 2**, was performed for the absolute quantification of *Atxn7* and *lnc-SCA7* using gene-specific 20X TaqMan Gene Expression Assay (*Atxn7*: Mm01315281_m1; *Atxn7l3b* (*lnc-SCA7*): Mm03015427_g1; *Gapdh*: Mm99999915_g1; Cat. No. 4351372, Life Technologies).

Chromatin Immunoprecipitation

Chromatin Immunoprecipitation (ChIP) assays were performed by Dr. Keith Vance as described in **Chapter 2** using wild-type N2As and three independently derived polyclonal N2A cell lines that stably expressing *sh-lnc-SCA7* or control shRNA (*sh-scramble*) for regions bound by GCN5 (within STAGA complex) by using anti-rabbit GCN5 antibody (Santa Cruz H-75) relative to that bound by anti-rabbit IgG control antibody (Millipore).

STAGA-binding was tested for five consecutive regions (250 bp in length) upstream of each transcriptional start site (TSS) of *miR-124* precursor

transcripts (*pri-miR-124*) annotated by the ENCODE consortium (UCSC browser) (Myers et al., 2011) that are sensitive to DNAse I treatment and are enriched in H3K27ac marks in the cerebellum within: chr14:65,205,705 – 65,207,200, chr3:17,694,143 – 17,695,600, and chr2:180,627,439 – 180,628,900 for *pre-miR-124-1*, *-2* and *-3*, respectively. The control region (chr14:65,183,839 – 65,185,271) was selected based on its lack of DNAse I and H3K27ac marks and its proximity to the predicted STAGA bound regions. Primers used to detect all regions were designed with a similar nucleotide composition. Specific enrichment of GCN5 relative to IgG was determined from three independent ChIP assays by qPCR. Primers used are listed in Table 2.4.

SCA7 knock-in mouse models

SCA7$^{100Q/5Q}$ knock-in mice carrying 100 CAG repeats on the pathological allele in the mouse Sca7 locus were kindly provided by Dr. H. Y. Zoghbi (Chen et al., 2012) to the lab of Alexis Brice and Annie Sittler and were maintained by Martina Marinello and Dr. Sandro Alves. Heterozygous SCA7$^{100Q/5Q}$ males were mated with SCA7$^{100Q/5Q}$ females. Genotyping was as described previously (Yoo et al., 2003). Homogenous SCA7$^{100Q/100Q}$ male animals (aged 28 weeks) were used for subsequent RNA quantification by qRT-PCR and by *in-situ* hybridization. The experiments were carried out in accordance with the Guide for the Care and Use of Laboratory Animals (National Research Council 1996), European Directive N°86/609 and the guidelines of the local institutional animal care and use committee. The study was approved (06/26/2010) by the local Institutional Review Board (Direction Générale pour la Recherche et l'Innovation).

SCA7$^{266Q/5Q}$ mice (Yoo et al., 2003) were obtained from the Jackson Laboratories (stock number 008682) and maintained by crossing heterozygous SCA7$^{266Q/5Q}$ with wild-type 5Q/5Q animals by Dr. Peter Oliver. Genotyping was performed as previously described (Yoo et al., 2003). Heterozygous SCA7$^{266Q/5Q}$ male animals (aged 5 weeks) were used for subsequent RNA quantification by qRT-PCR and *in-situ* hybridization. Experiments were carried out according to United Kingdom Home Office Animals (Scientific Procedures) Act 1986 and local ethical approval from the University of Oxford. In both cases, the mice were maintained on a 12 h light/dark cycle with access *ad libitum* to food and water. Only regions with the highest Gcn5 binding found in N2As was tested in the cerebellum of SCA7$^{100Q/100Q}$: namely 1a, 2b, 3c, na for *miR-124*-1, *miR-124*-2, *miR-124*-3 and negative control region, respectively. Primers used are listed in Table 2.4.

Tissue preparation for RNA analyses

Homozygous SCA7$^{100Q/100Q}$ mice at the late stage of disease (28 weeks of age, n = 4) and wild-type age- and sex-matched controls (n = 2) were used to perform qRT-PCR expression analyses. *In situ* hybridization and qRT-PCR was also carried out from heterozygous SCA7$^{266Q/5Q}$ mice (5 weeks of age, n = 3) and wild-type age- and sex- matched littermate controls (SCA7$^{5Q/5Q}$, n = 3). Biopsies from mice tissues were obtained and analyzed by Dr. Peter Oliver as described in **Chapter 2**.

In-situ hybridization

In-situ hybridization assays were performed by Dr. Peter Oliver as described in **Chapter 2** to visualize the expression and location of target sequences, specifically *Atxn7* and *miR-124*. Detection of *Atxn7* was carried out as described for general mRNAs and that of *miR-124* was performed as described for general small RNAs. Primers used are listed in Table 2.4.

SCA7 patient fibroblasts

Transcript abundance of SCA7 fibroblast cell lines against that of a control fibroblast cell line (10Q/10Q) was quantified by qRT-PCR using SYBR green Master PCR mix (Invitrogen) and target-specific primers in combination with a TaqMan (Invitrogen) probe (*ATXN7*: Cat. No. 4331182, Invitrogen; *GAPDH*: Cat. No. 4331182, Invitrogen; *lnc-SCA7* (Appendix Table A5.1). Results illustrated in Figure 5.20 were measured using sensitive TaqMan-based qRT-PCR. Ethics approval for the establishment of patient fibroblast cultures was granted by the University of Cape Town (UCT) Faculty of Health Sciences Human Research Ethics Committee (HREC REF. 380/2009 and 434/2011), and was renewed annually. Primers used are listed in Table 2.4.

Statistics

All expression correlation comparisons were determined using the Pearson's correlation test and all differential expression comparisons were determined using Student's *t*-test. Asterisks indicate the level of significance of the comparison between the expression of target transcripts (* $p<0.05$; ** $p<0.01$;

*** $p<0.001$; NS [not significant] $p>0.05$). For each experimental analysis statistical values were calculated using data collected from at least three independent experiments.

5.4 RESULTS

5.4.1 lnc-SCA7 is a post-transcriptional regulator of Atxn7 expression

Using publicly available gene expression data (Su et al., 2004), I identified a conserved retrotransposed gene, *lnc-SCA7*, also known as *ATXN7L3B*, whose expression pattern is significantly correlated with that of *ATXN7,* the gene mutated in SCA7, across 59 adult and developing human tissues and cell lines (Pearson's R^2=0.24, $p<0.05$, data not shown). Using quantitative reverse transcriptase PCR (qRT-PCR), I validated this correlation using 20 adult human tissues (Pearson's R^2=0.82, $p<0.001$, Figure 5.1A). Moreover, qRT-PCR expression levels of the mouse orthologs of *Atxn7* and *lnc-SCA7* were also significantly correlated across 11 mouse adult tissues and 9 postnatal central nervous system (CNS) regions (Figure 5.1B). The correlation between *lnc-SCA7* and *Atxn7* levels was significantly higher across mouse brain regions (Pearson's R^2=0.94, $p<0.001$) than elsewhere (Pearson's R^2=0.69, $p<0.05$, Figure 5.1B), suggesting the presence of an additional layer of regulation that controls the relative abundance of both transcripts in the brain. Correlated transcript abundance for these two genes suggests that they may be under similar transcriptional or post-transcriptional control, especially in the brain.

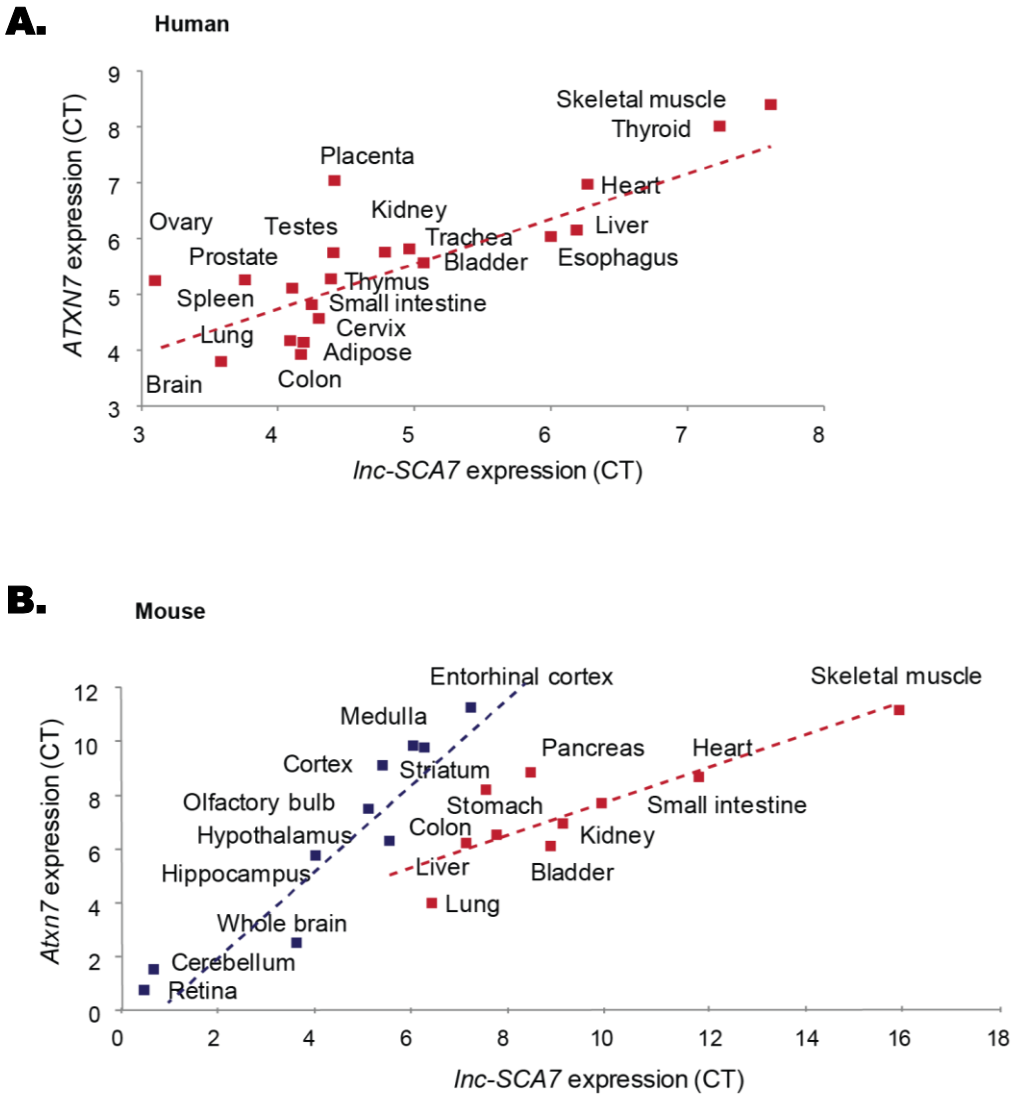


Figure 5.1 *lnc-SCA7* and *Atxn7* expression is correlated. (A) The levels of *lnc-SCA7* (x-axis) and *ATXN7* (y-axis) measured by quantitative reverse transcription PCR (qRT-PCR) [cross-threshold cycle (CT)] using transcript specific primers (Table 2.4) in 20 adult human tissues (adipose, bladder, brain, cervix, colon, esophagus, heart, kidney, liver, lung, ovary, placenta, prostate, skeletal muscle, small intestine, spleen, testes, thymus, thyroid, and trachea) are significantly correlated (R^2 = 0.82, p<0.001, Pearson's test). (B) Mouse *Atxn7* and *lnc-SCA7* expression levels, measured by qRT-PCR using transcript-specific primers (Table 2.4), across 11 mouse adult tissues (bladder, brain, colon, heart, kidney, liver, lung, pancreas, skeletal muscle, small intestine, and stomach) and across 9 brain regions (retina, cerebellum, cortex, entorhinal cortex, hippocampus, hypothalamus, medulla, olfactory bulb, and striatum) are significantly correlated (p<0.05, Pearson's test). The correlation is stronger across central nervous system regions (R^2 = 0.94, blue) than in non-CNS tissues (R^2 = 0.69, red).

lnc-SCA7 arose from retrotransposition of *Ataxin-7-like protein 3* (*Atxn7l3*), a distant paralog of *Atxn7*, in the common ancestor of placental mammals approximately 100 million years ago (Figure 5.2). I found no homology between the 1kb genomic regions upstream of *Atxn7l3* and *lnc-SCA7*, indicating that they are unlikely to have homologous promoters. *lnc-SCA7* inserted downstream of a pre-existing CpG island, and since its duplication accumulated frame-shifting deletions that resulted in premature stop-codons and a truncated open reading frame (ORF) (Figure 5.3A). The small conceptual polypeptide (97 amino acids) encoded by *lnc-SCA7* lacks the two annotated functional Atxn7l3 protein domains (Figure 5.3B). Despite both transcripts being expressed at similarly high levels in mouse neuroblastoma cells (N2A, Figure 5.3C) a custom antibody raised against the N-terminal protein sequence conserved between Atxn7l3 and the putative lnc-SCA7 protein (Figure 5.3A) detected translation of Atxn7l3 (predicted size 39kDa) but not of a polypeptide of the size expected for lnc-SCA7 protein (11kDa, Figure 5.3D). The transcript originating from this gene is thus unlikely to be translated into a stable protein product in these cells.

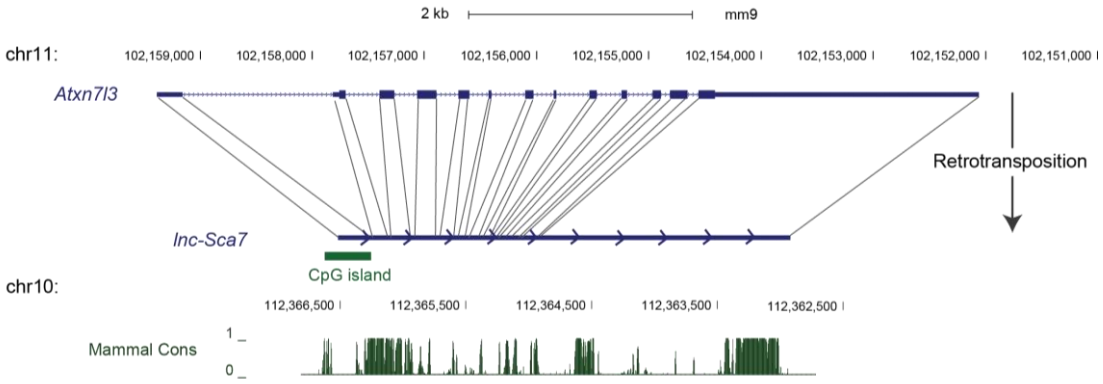


Figure 5.2 *lnc-SCA7* is a retropseudogene of the protein coding gene, *Atxn7l3*. UCSC track illustrating the genomic coordinates of *lnc-SCA7* and *Atxn7l3*. Blue boxes represent exonic sequences and black lines between the two genes denote regions of sequence homology, suggesting *lnc-SCA7* is derived from retrotransposition of *Atxn7l3*. The green box represents a stretch of CpG island, indicating that *lnc-SCA7* is likely transcribed from this CpG island, instead of from the same promoter element as its ancestral gene, *Atxn7l3*.

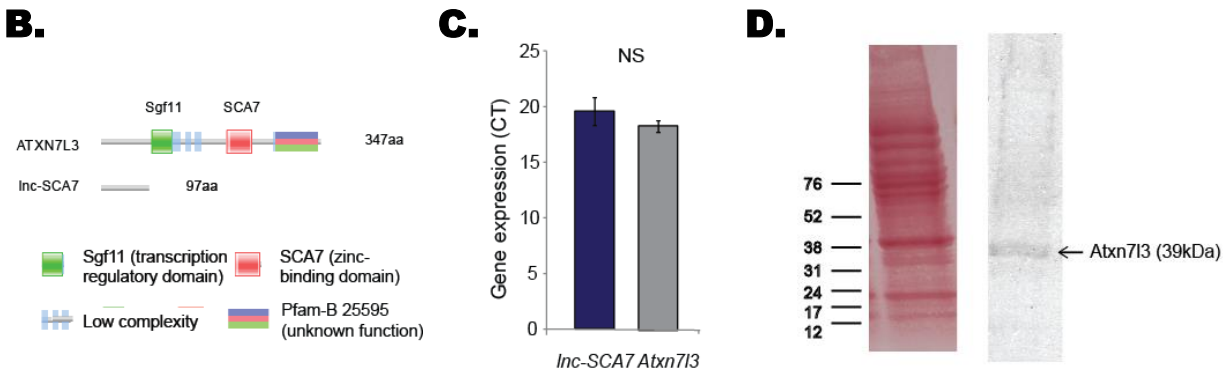
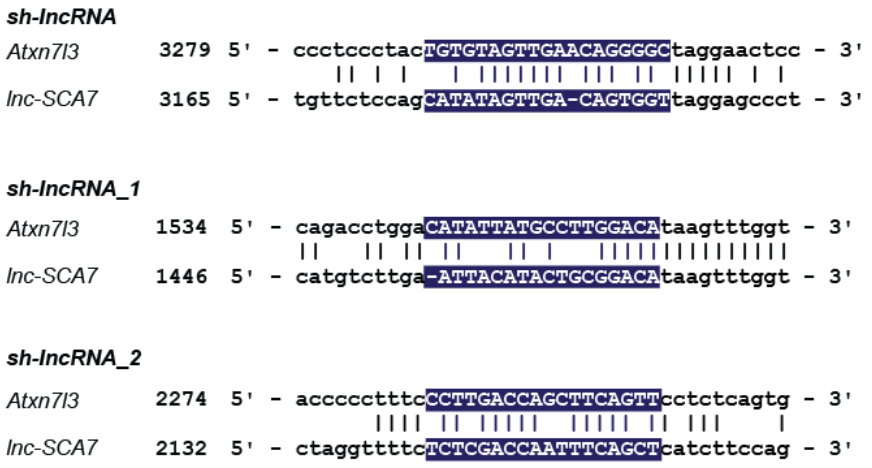


Figure 5.3 *lnc-SCA7* is a retropseudogene of *Atxn7l3*. (A) Pairwise sequence alignment between the open reading frame (ORF, blue) of mouse Atxn7l3 and the homologous region in lnc-SCA7. A frame-shifting deletion (black box) in lnc-SCA7 resulted in a premature stop codon (red box). The putatively coding region conserved between lnc-SCA7 and Atxn7l3 that was used to raise a custom antibody recognizing both Atxn7l3 and the putative peptide encoded by lnc-SCA7 is highlighted in red. The lnc-SCA7-STOP recombinant locus was generated by creating a premature stop codon in lnc-SCA7 by site-directed single nucleotide mutagenesis at nucleotide position 14 (C to A). Insertions and deletions in the pairwise alignment are represented by dashes. (B) Schematic diagram representing the protein domains within ATXN7L3 (ATXN7L3_HUMAN; Q14CW9) and lnc-SCA7 (ATXN7L3B) obtained from Pfam17. The putative protein product of lnc-SCA7 (97 amino acids) is shorter than ATXN7L3 (354 amino acids) and lacks the protein domains present in ATXN7L3, namely transcriptional regulation (Sgf11, green) and Zinc-binding (SCA7, red) secondary domains. (C) lnc-SCA7 (blue) and Atxn7l3 (grey) are expressed (y-axis) [cross-threshold cycle (CT)] at similar levels in N2A cells. Expression levels for the 2 genes were normalized relative to Gapdh. (D) In mouse neuroblastoma cells (N2A) whole cell protein lysate (left), the custom antibody raised against a conserved region between lnc-SCA7 and Atxn7l3 detected, by western blot, only a band at approximately 38kDa, likely to be Atxn7l3 (predicted size 39kDa, arrow). No detectable band was apparent at the expected size for lnc-SCA7 ORF (predicted size of 11kDa), suggesting that lnc-SCA7 is not translated into a stable protein (right).

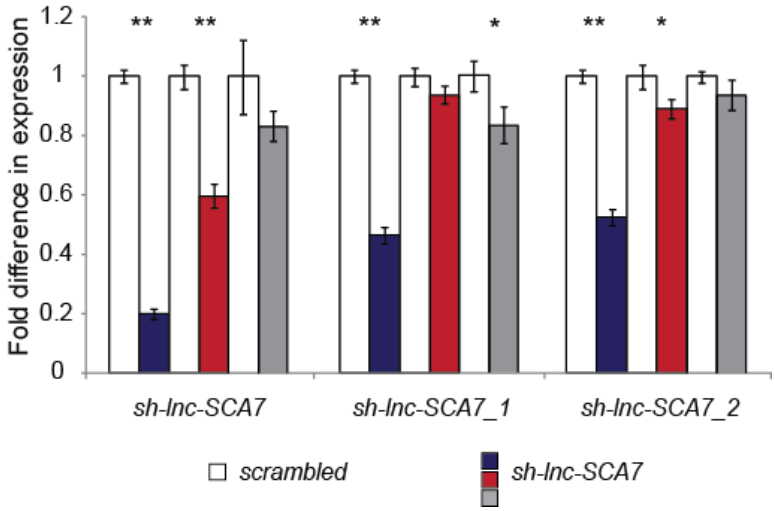
The coordinated expression between *lnc-SCA7* and *Atxn7* in both mouse and human prompted us to explore whether this long noncoding RNA regulates *Atxn7*'s transcript abundance. In N2A cells, *lnc-SCA7* depletion (up to 80%) using multiple target-specific short hairpin RNA constructs (*sh-lnc-SCA7*s, Figure 5.4A) significantly reduced *Atxn7* transcript (up to 40%, Figure 5.4B) and protein levels (by approximately 40%, Figure 5.4C). It was noted that *lnc-SCA7* knockdowns also marginally reduced *Atxn7l3* abundance, possibly as a result of off-target effect of the *sh-lnc-SCA7* constructs used (Figure 5.4B). Furthermore, over-expression (6.8-fold) in N2A cells of the region downstream of *lnc-SCA7*'s putative stop-codon (nucleotides 599–3607, hereafter termed *lnc-SCA7*-WT) significantly increased *Atxn7* transcript (2.3-fold, Figure 5.4D) and protein levels (Figure 5.4E). Comparable increases in *Atxn7* transcript levels were observed upon over-expression of either full-length *lnc-SCA7* sequence (*lnc-SCA7*-FULL) or a recombinant mutant carrying a premature stop codon by performing site direct mutagenesis on position 14 of *lnc-SCA7*-FULL from C->A (*lnc-SCA7*-STOP) (Figure 5.4D).

I concluded that, in mouse, *lnc-SCA7* modulates the abundance of *Atxn7* via a transcript-dependent mechanism that does not rely on the translation of its putative ORF. Given its cytoplasmic localization (Figure 5.5), I hypothesized that similarly to other lincRNAs (Tay et al., 2014) *lnc-SCA7* modulates *Atxn7* transcript levels post-transcriptionally by competing for the binding of shared miRNAs.

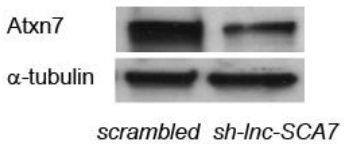
A.


sh-lncRNA
Atxn7l3 3279 5' - ccctccctactGTGTAGTTGAACAGGGGCtaggaactcc - 3'
lnc-SCA7 3165 5' - tgttctccagCATATAGTTGA-CAGTGGTtaggagccct - 3'
sh-lncRNA_1
Atxn7l3 1534 5' - cagacctggaCATATTATGCCTTGGACAtaagtttggt - 3'
lnc-SCA7 1446 5' - catgtcttga-ATTACATACTGCGGACAtaagtttggt - 3'
sh-lncRNA_2
Atxn7l3 2274 5' - acccccttttcCCTTGACCAGCTTCAGTTcctctcagtg - 3'
lnc-SCA7 2132 5' - ctaggttttcTCTCGACCAATTTCAGCTcatcttccag - 3'

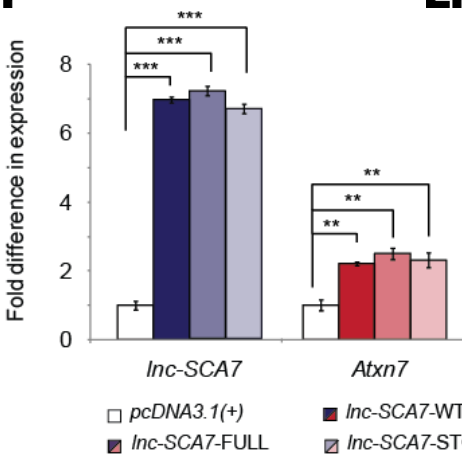
B.


Fold difference in expression
1.2
1
0.8
0.6
0.4
0.2
0
**
**
**
*
**
*
sh-lnc-SCA7
sh-lnc-SCA7_1
sh-lnc-SCA7_2
scrambled
sh-lnc-SCA7

C.


Atxn7
α-tubulin
scrambled sh-lnc-SCA7

D.


Fold difference in expression
8
6
4
2
0
***
***
***
**
**
**
lnc-SCA7
Atxn7
pcDNA3.1(+)
lnc-SCA7-FULL
lnc-SCA7-WT
lnc-SCA7-ST

E.

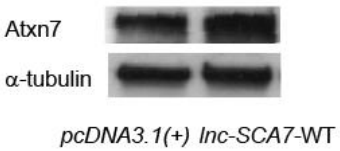

Atxn7
α-tubulin
pcDNA3.1(+) lnc-SCA7-WT

Figure 5.4 *lnc-SCA7* regulates *Atxn7* abundance. (A) Sequence alignment between regions in mouse *Atxn7l3* and *lnc-SCA7* used to design specific short hairpin RNAs (shRNAs) against *lnc-SCA7*. The initial base of the predicted binding sites in these transcripts is noted to the left of the alignment. Identical nucleotides between *Atxn7l3* and *lnc-SCA7* are denoted by vertical lines. Regions targeted by the siRNAs are highlighted in blue. (B) Knockdown of *lnc-SCA7* in N2A cells using each of the 3 designed shRNAs (*sh-lnc-SCA7*, *sh-lnc-SCA7_1*, and *sh-lnc-SCA7_2*) led to significantly decreased *lnc-SCA7* levels (blue) and was associated with decreased *Atxn7* (red) and *Atxn7l3* (grey) levels. (C) Fold difference in expression (y-axis) in N2A cells upon knockdown of *lnc-SCA7* (blue) was associated with a significant reduction in *Atxn7* (red) transcript abundance and reduced Atxn7 protein; α-tubulin was used as loading control. (D) Over-expressing the putative 3' noncoding region of *lnc-SCA7* (*lnc-SCA7-WT,* dark blue), full length *lnc-SCA7* (*lnc-SCA7-FULL*, blue) and recombinant *lnc-SCA7-STOP* (light blue) in N2A cells each led to significantly increased *Atxn7* levels (dark red, red and pink, respectively) relative to control (white). (E) Over-expression of *lnc-SCA7-WT* increased Atxn7 protein; α-tubulin was used as loading control.

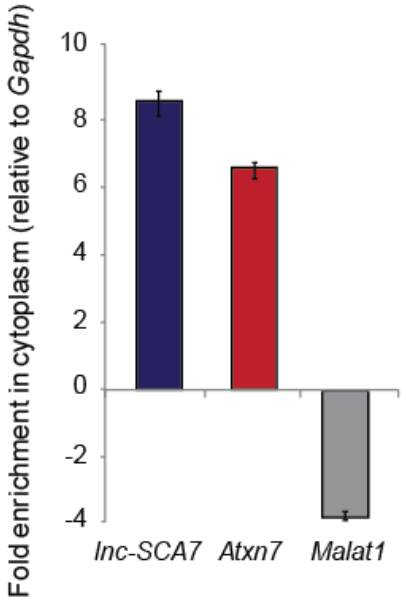


Figure 5.5 *lnc-SCA7* is enriched in the cytoplasm. *lnc-SCA7* (blue) and *Atxn7* mRNA (red) are found predominantly in the cytoplasm of N2A cells (y-axis). *Malat1* (light grey), a known nuclear transcript was used as control. Fold enrichments (expression level measured in the cytoplasmic/expression level measured in the nucleus) for all genes were measured and normalized against that of *Gapdh*, a cytoplasmic single exonic transcript.

5.4.2 miR-124 mediates the regulatory interaction between lnc-SCA7 and Atxn7

To test this hypothesis, I took advantage of a mouse embryonic stem (ES) cell (mESC) line kindly provided by Dr. Tatyana Nesterova and Dr. Sarah Cooper that is conditionally deficient for Dicer (*Dicer*$^{\Delta/\Delta}$), an essential component of the miRNA biogenesis pathway in mammals (Nesterova et al., 2008). Specifically, deletion of *Dicer*'s RNAse III domain in DTCM23/49 XY mESCs was induced by culturing the cells in the presence of tamoxifen [(Z)-4-Hydroxytamoxifen (4-OHT)] (*Dicer*$^{\Delta/\Delta}$), where non-induced cells were treated with 0.1% ethanol and used as control (*Dicer*$^{wt/wt}$). As for N2A cells, *lnc-SCA7* knockdown (24%) in wild-type mESCs significantly reduced the expression levels of *Atxn7* (18%, Figure 5.6A). In contrast, in *Dicer*-deficient mESCs similar level of *lnc-SCA7* knock-down had no significant effect on *Atxn7* abundance (Figure 5.6B). This is consistent with the regulation of *Atxn7* levels by *lnc-SCA7* being miRNA-dependent.

Of all brain-expressed miRNAs, only two, *miR-16* and *miR-124*, contain conserved (between mouse and human) predicted miRNA response elements (MREs) within the 3' UTRs of both *Atxn7* and *lnc-SCA7* (Figure 5.7A-B, Appendix Table A5.1). In contrast to *miR-16* that has no known role in the brain, *miR-124* is the most abundantly expressed miRNA in the CNS (Lagos-Quintana et al., 2002) and has well established roles in neuronal development (Visvanathan et al., 2007).

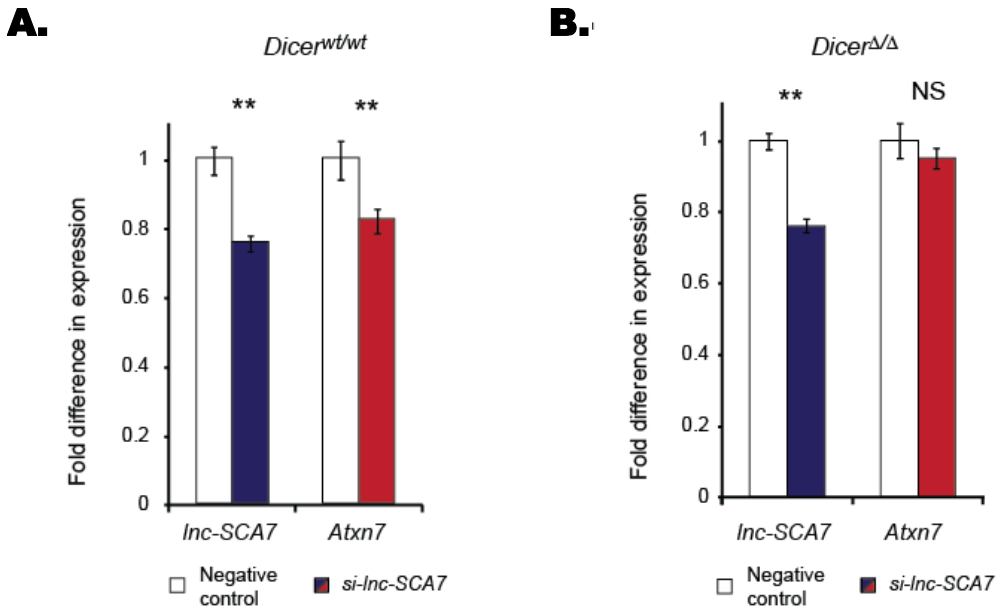


Figure 5.6 Post-transcriptional regulation by *lnc-SCA7* is *miR-124*-mediated. (A) Transfection of small interfering RNAs targeting *lnc-SCA7* (*si-lnc-SCA7*) significantly reduces its levels (blue) and leads to significant reduction in *Atxn7* abundance (red) in DTCM/D49 XY mouse ES cells (mESCs) not exposed to tamoxifen. (B) Upon transfection of mESCs lacking *Dicer* (*Dicer*$^{\Delta/\Delta}$), a similar reduction in *lnc-SCA7* levels (blue) had no significant effect on *Atxn7* levels (red). A non-specific scrambled sequence was used as control (white).

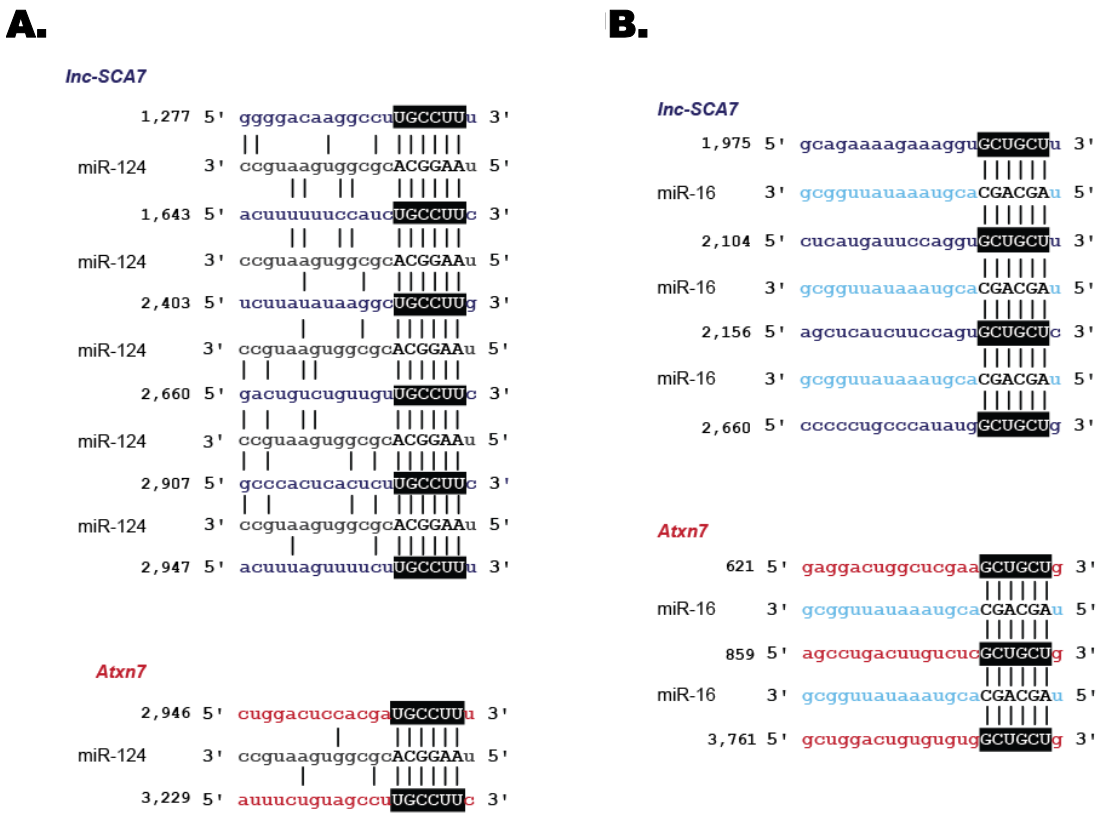


Figure 5.7 Sequence alignments between MREs of *lnc-SCA7* with their respective miRNA seeds. (A and B) Alignment of (A) *miR-124* and (B) *miR-16* and their respective miRNA response elements (MREs) within the 3' UTRs of mouse *lnc-SCA7* (top panels) or *Atxn7* (bottom panels). The initial base of the predicted binding sites in these transcripts is noted to the left of the alignment. Identical nucleotides between *miR-124/miR-16* with *lnc-SCA7* and *Atxn7* are denoted by vertical lines.

Transfection of N2A cells with *miR-124* mimics reduced *lnc-SCA7* and *Atxn7* expression levels relative to a non-specific miRNA negative control (by 68% and 81%, respectively; Figure 5.8A), while reduction of endogenous levels of *miR-124* (38% reduction) led to significant increases in both *lnc-SCA7* and *Atxn7* levels (by 1.8- and 2.5-fold, respectively; Figure 5.8B). In contrast, *miR-16* mimics failed to significantly alter the levels of these transcripts (Figure 5.9A). Together with *miR-16*'s considerably lower abundance (Figure 5.9B), these findings indicated that *miR-124*, but not *miR-16*, is likely to mediate crosstalk between *lnc-SCA7* and *Atxn7* in mouse neurons.

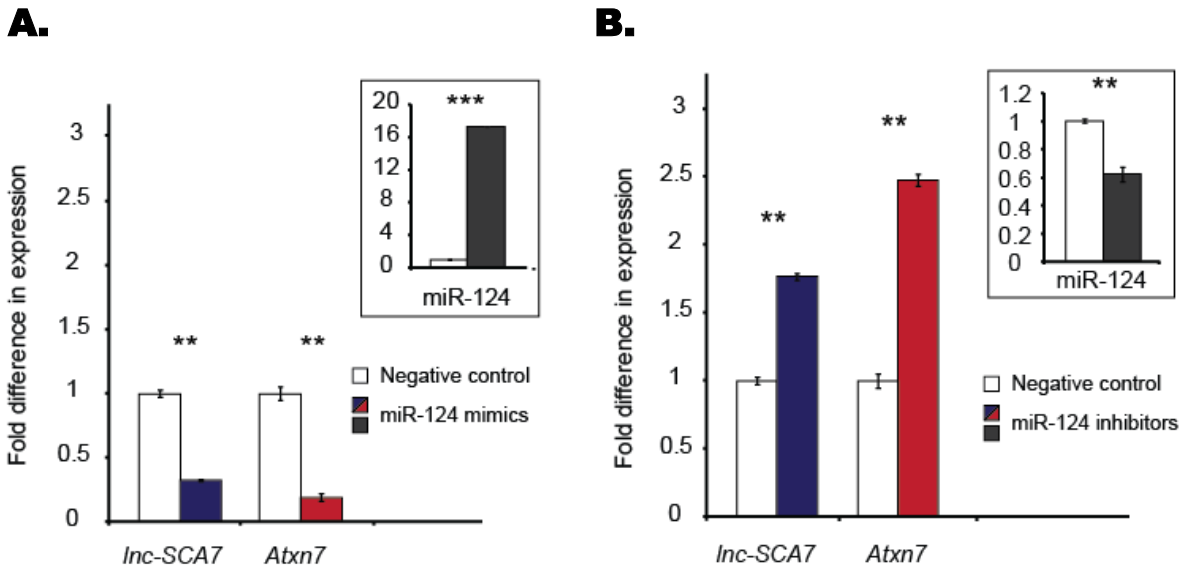


Figure 5.8 (A) Transfection of *miR-124* mimics in N2A cells (top-right insert, dark grey) results in decreased levels of both *lnc-SCA7* (blue) and *Atxn7* (red) relative to a negative miRNA transfection control (white). (B) Transfection of *miR-124* inhibitors in N2A cells (top-right insert, dark grey) results in increased levels of both *lnc-SCA7* (blue) and *Atxn7* (red) relative to a negative miRNA transfection control (white).

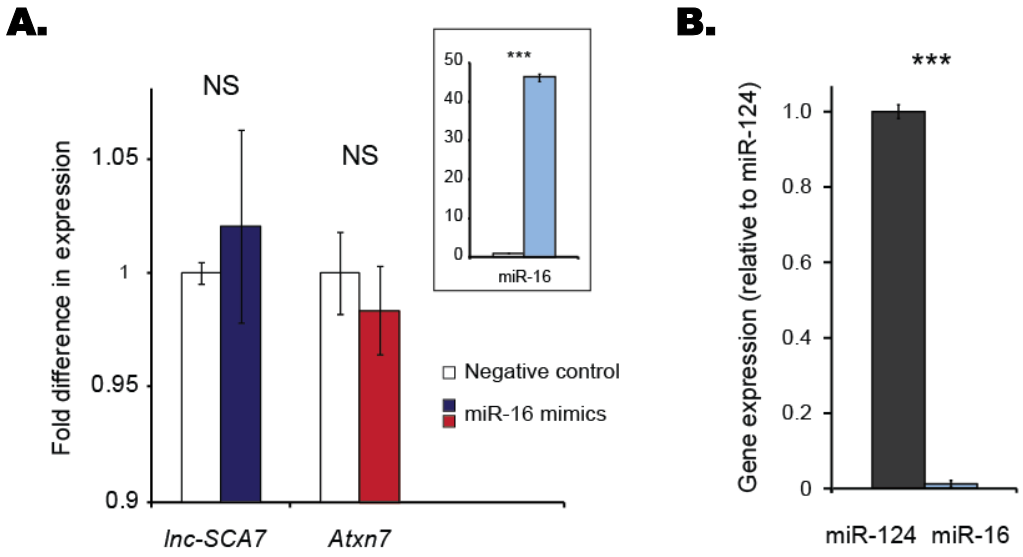


Figure 5.9 The ability of *lnc-SCA7* to regulate *Atxn7* abundance is *miR-124-,* but not *miR-16*-dependent. (C) Over-expression of *miR-16* mimics top right box, light blue) in N2A cells had no significant effect on the expression levels of either *lnc-SCA7* (NS, blue) or *Atxn7* (NS, red) relative to control (white). A non-specific sequence was used as control (white). (D) *miR-16* (light blue) is relatively more lowly expressed in N2A cells compared to *miR-124* (dark grey).

This conclusion was further supported by the reduction in reporter activity upon co-transfection of *miR-124* mimics and recombinant *lnc-SCA7* or *Atxn7* luciferase reporter constructs (by 77% and 58%, respectively; Figure 5.10A) being dependent on the presence of the predicted *miR-124* MREs in these transcripts. More specifically, inversion of the seed sequences of all *miR-124* MREs predicted within *lnc-SCA7* (6 MREs) and *Atxn7* (2 MREs) (hereafter referred to as *lnc-SCA7*-MUT and *Atxn7*-MUT, respectively) abolished the effect of *miR-124* on reporter activity (Figure 5.10A). As expected, neither *lnc-SCA7*-MUT nor *Atxn7*-MUT over-expression (7.7 and 9.7-fold, respectively) had a significant impact on *Atxn7* (Figure 5.10B) or *lnc-SCA7* abundance (Figure 5.10C), respectively, consistent with the ability of *lnc-SCA7* and *Atxn7* to modulate each other's abundance being *miR-124* dependent.

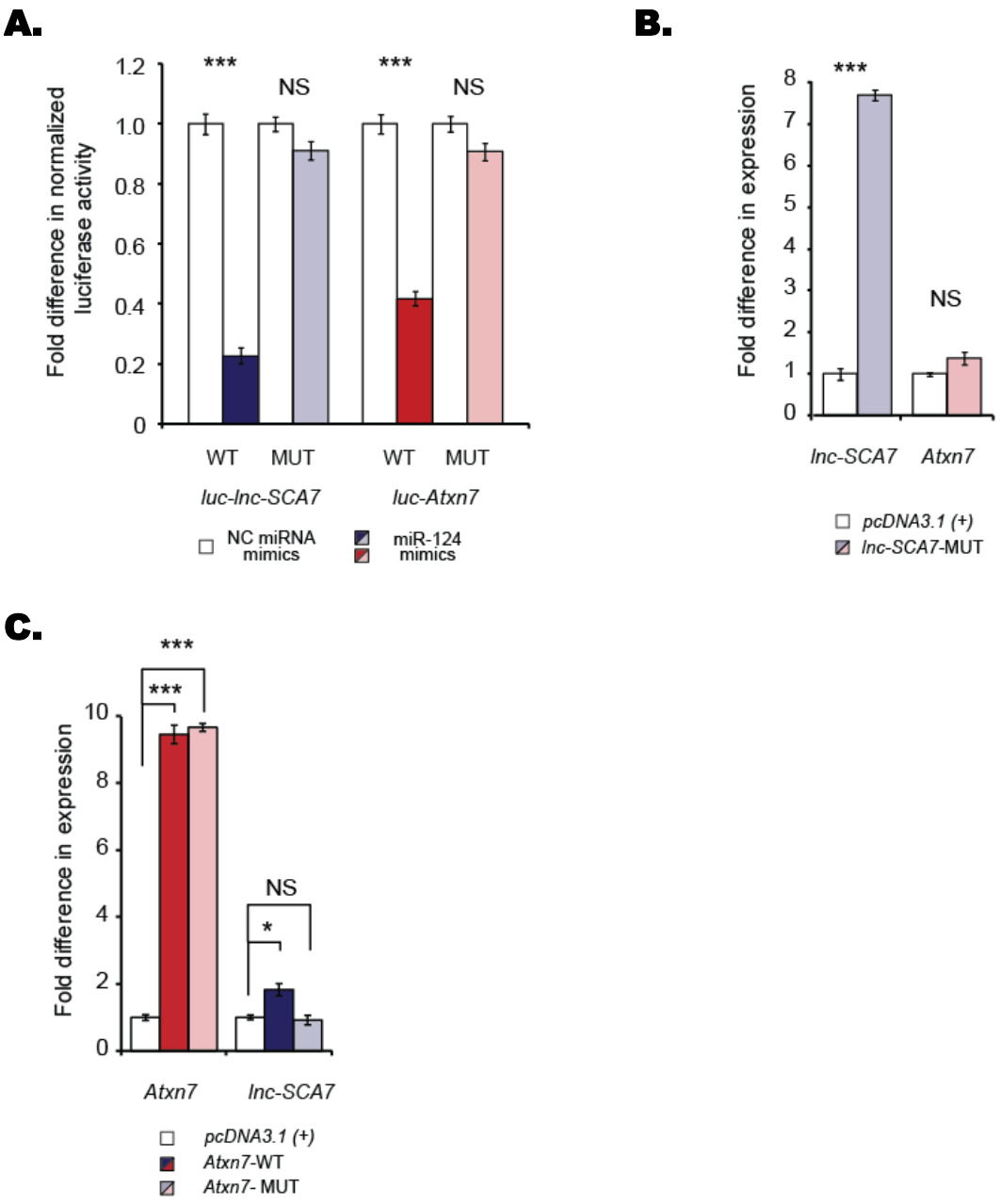
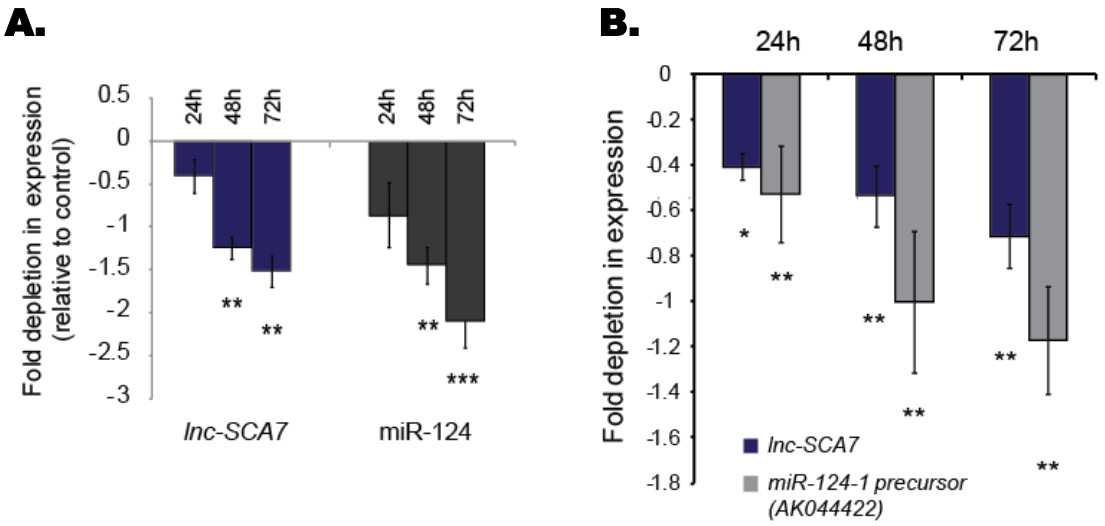


Figure 5.10 (A) Co-transfection of *miR-124* mimics with *luc-lnc-SCA7-WT* (blue) or *luc-Atxn7-WT* (red) luciferase reporter constructs, but not *luc-lnc-SCA7-MUT* (light blue) or *luc-Atxn7-MUT* (pink), resulted in a significant reduction in normalized luciferase activity, relative to control (NC, white) in N2A cells. (B) Over-expression of *lnc-SCA7-MUT* (light blue) was not associated with increased *Atxn7* abundance (pink), relative to control (white). (C) Over-expression of *Atxn7-WT* increased levels of *lnc-SCA7* (blue), whereas over-expression of *Atxn7-MUT* had no effect on *lnc-SCA7* (light blue), relative to control (white).

5.4.3 A novel negative feedback loop involving ATXN7 and miR-124

Reduction in *lnc-SCA7* levels surprisingly led to depletion of both mature (2.1-fold depletion, Figure 5.11A) and precursor *miR-124* levels (1.1-fold depletion, Figure 5.11B). Decreased levels (0.37-fold reduction) of *lnc-SCA7* in human neuroblastoma cells (SH-SY5Y) was associated with significantly decreased levels of both *ATXN7* (0.16-fold reduction) and *pri-miR-124a_1* (0.13-fold reduction), consistent with the conservation of regulatory interactions in humans (Figure 5.11C). Furthermore, and as seen in mouse neuroblastoma cells, decreased levels of endogenous miR-124 were associated with increased *lnc-SCA7* and *ATXN7* abundance in this cells (Figure 5.11D), consistent with the conservation of crosstalk between this transcripts. Changes in *miR-124* levels do not reflect a general effect of *lnc-SCA7* on miRNA expression or processing because genome-wide analysis of miRNA abundance revealed no significant differences in N2A miRNA repertoires following either knockdown or over-expression of *lnc-SCA7* (Data not shown due to thesis size limitations, see section 5.3 Materials and Methods).



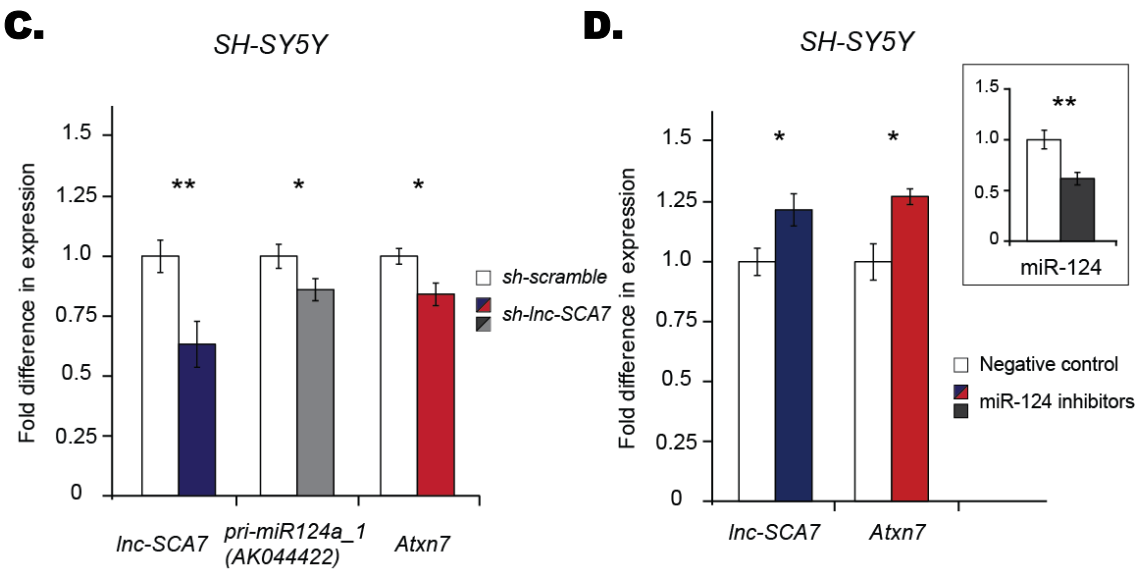


Figure 5.11 Transcription of *miR-124* precursors is STAGA-dependent. (A) Knockdown of *lnc-SCA7* (blue) in N2As led to a significant decrease in mature *miR-124* levels (dark grey) over a 72 hour time-course. (B) Knockdown of *lnc-SCA7* (blue) in N2A cells led to successive reductions in *pri-miR-124a_1* (light grey, *AK044422*) over the time course of 72h. Fold enrichment/depletion in expression, following *lnc-SCA7* knockdown, is calculated relative to transcript abundance following transfection with the scrambled control. (C) Transfection of miR-124 inhibitors in SH-SY5Y cells (top-right insert, dark grey) results in increased levels of both *lnc-SCA7* (blue) and *Atxn7* (red) relative to a negative miRNA transfection control (white). Fold enrichment/depletion in expression, following *lnc-SCA7* knockdown, is calculated relative to transcript abundance following transfection with the scrambled control.

This suggested that *lnc-SCA7* abundance correlates with the rate of transcription of *miR-124* precursor loci and that STAGA might be required for *miR-124* transcriptional initiation. To test this hypothesis, I first identified the putative promoters of the three *miR-124* precursor genes (*pri-miR-124*s) as their nearest upstream DNAse I hypersensitivity region in the mouse cerebellum that was marked with H3K27ac in that tissue (Figure 5.12A-C). STAGA is required for the transcriptional activation of *miR-124* loci. Each of

these promoters exhibited at least 8.5-fold higher reporter activity than the control antisense sequence (Figure 5.12D). In addition, as shown by the chromatin immunoprecipitation (ChIP) experiment, all of three promoter regions were bound by Gcn5, STAGA's histone acetyltransferase (Helmlinger et al., 2004), at 2.0- to 3.5-fold greater levels than IgG control (Figure 5.12E).

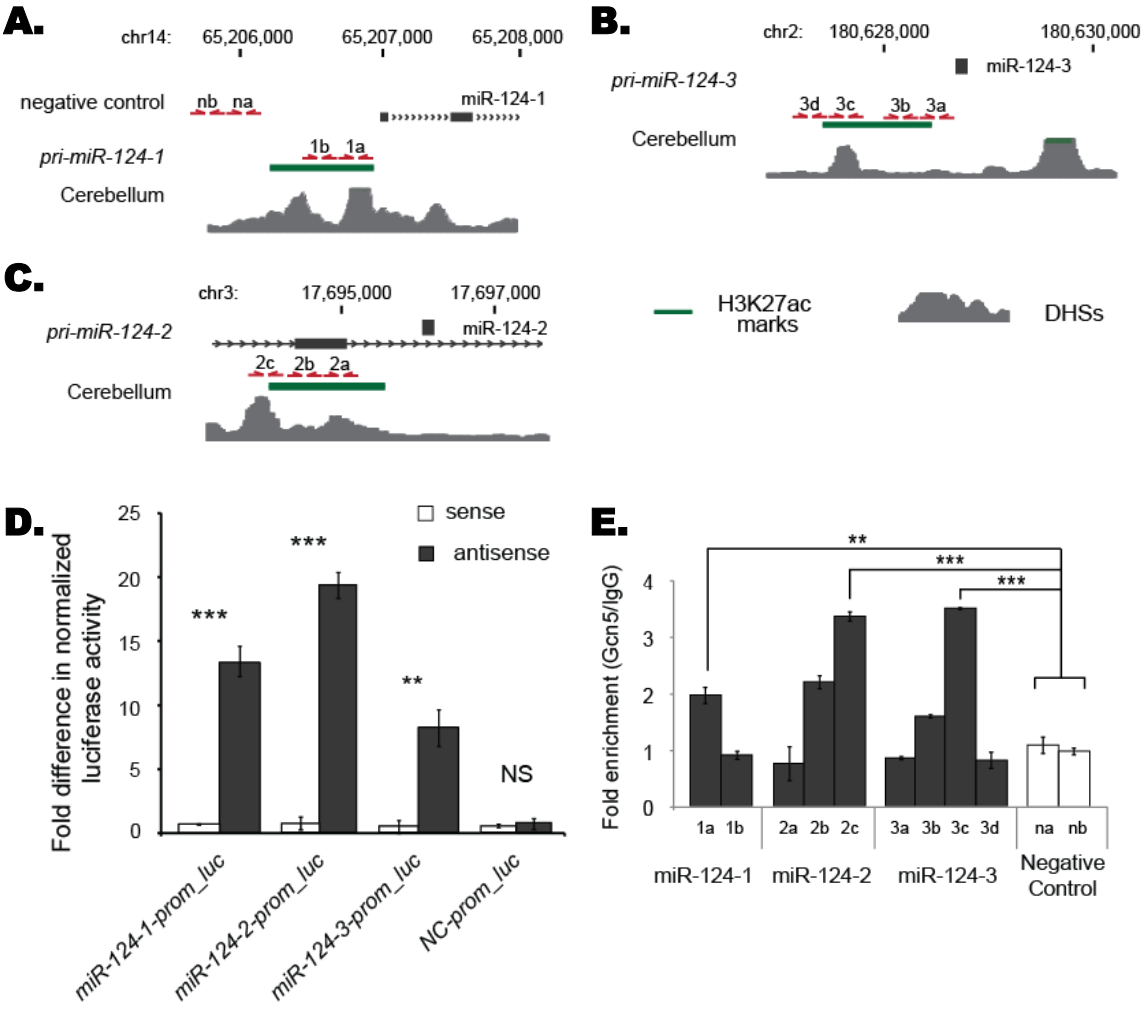


Figure 5.12 *lnc-SCA7* modulates *miR-124* abundance. (A-C) Genome browser view of the regions of the mouse genome (mm9) encoding the (A) negative control region and *pri-miR-124-1*, (B) *pri-miR-124-2* and (C) *pri-miR-124-3* putative promoter regions. These *pri-miR-124* regions are hypersensitive to DNase I treatment (grey, y-axis peak height correlates with number of ChIP-seq DHSI reads) in the cerebellum. Regions enriched in the cerebellum for

H3K27ac (green) were defined as possible promoters: *miR-124*-1 (chr14:65,205,705 – 65,207,200), *miR-124*-2 (chr3:17,694,143 – 17,695,600), *miR-124*-3 (chr2:180,627,439 – 180,628,900) and negative control region (chr14:65,183,839 – 65,185,271). Red arrows indicate the position of primers pairs used to test regions enriched in STAGA binding: 1b, 1a, 2a, 2b, 2c, 3d, 3c, 3b, 3a, na, and nb (Table 2.4). (D) Transfection of any of the 3 *miR-124-prom-luc* (dark grey) in N2A cells resulted in significantly increased normalized reporter activity relative to constructs for which these regions were cloned in the antisense orientation (white). No significant (NS) change was observed for the negative control region. (E) ChIP-qPCR revealed significant enrichment in Gcn5 binding, relative to IgG control, in the promoter regions of *pri-miR-124s* (dark grey; Negative control, NC, white).

Furthermore, while over-expression of *lnc-SCA7*-WT increased luciferase activity for all three *miR-124* promoters (1.5- to 1.8-fold), no significant changes in activity were detected following over-expression of *lnc-SCA7*-MUT (Figure 5.13A). Accordingly, *lnc-SCA7* knockdown decreased Gcn5 binding (Figure 5.13B) and luciferase activity for each of the three *miR-124* promoters (Figure 5.13C). Furthermore over-expression of *Atxn7*-WT led to a significantly higher increase on *miR-124* promoter activity (Figure 5.13D) and mature *miR-124* levels (Figure 5.13E) than did *Atxn7*-MUT. As quantified by digital droplet PCR (ddPCR), I found *lnc-SCA7*'s concentration in N2A and ES cells (320 and 145 copies/µL of cDNA, respectively) to be at least 20-times higher than that of *Atxn7* (16 and 5 copies/µL of cDNA, respectively), suggesting that this lincRNA can efficiently modulate the levels of Atxn7 via competition for shared miRNAs (Ala et al., 2013; Figliuzzi et al., 2013) (Figure 5.14).

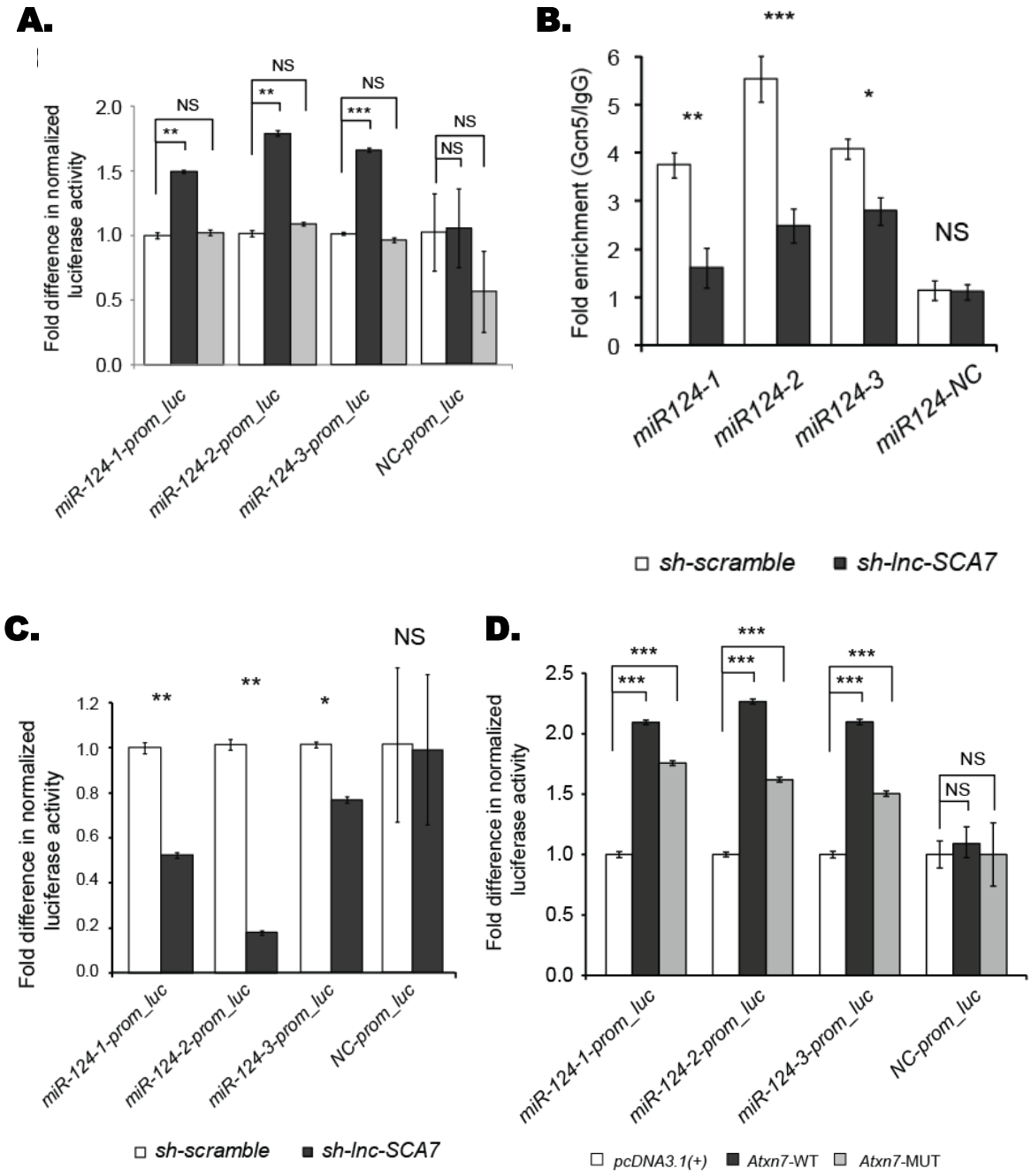


Figure 5.13 *lnc-SCA7* modulates *miR-124* levels by regulating promoter activity at their precursor genes. (A) Relative to control, co-transfection of *lnc-SCA7-WT* with all 3 *miR-124-prom-luc* reporter constructs resulted in significant increase in normalized reporter activity (dark grey), while co-transfection with *lnc-SCA7-MUT* had no significant effect on normalized luciferase activities (light grey). (B) ChIP-qPCR revealed significantly decreased enrichment in Gcn5 binding at *miR-124* promoters in N2As with stably knocked down *lnc-SCA7* (blue) relative to scrambled control (white). ChIP performed by Dr. Keith Vance followed qRT-PCR carried out by myself. (C) In N2As, and

relative to scrambled knockdown transfection control (white), co-transfection of *sh-lnc-SCA7* with recombinant luciferase constructs containing the *pri-miR-124* promoter regions resulted in significant decrease in normalized reporter activities for all *pri-miR-124* promoter regions (dark grey). No depletion was found in the negative control region, *NC-prom-luc*. *n = 3* biological replicates per condition. (D) In N2As, and relative to *Atxn7-MUT* (light grey), co-transfection of *Atxn7-WT* with recombinant luciferase constructs containing the *pri-miR-124* promoter regions resulted in significant increase in normalized reporter activities for all *pri-miR-124* promoter regions (dark grey).

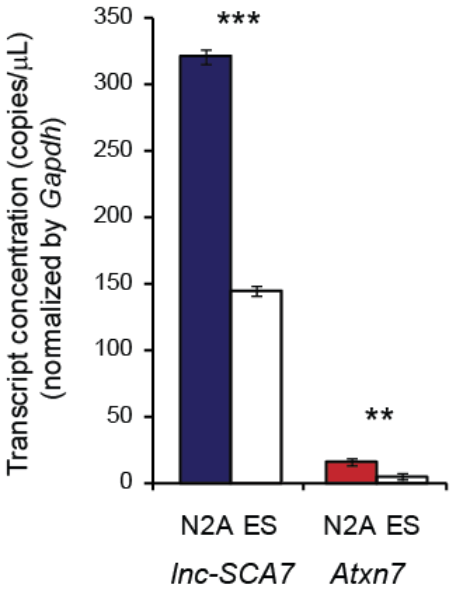


Figure 5.14 Absolute quantification of *lnc-SCA7* and *Atxn7* in mouse N2A and ES cells as measured by digital droplet PCR (ddPCR).

Next, I tested whether the miRNA-mediated crosstalk between *lnc-SCA7* and *Atxn7* extends to other mRNAs encoding other STAGA complex subunits. Twenty-two of these 23 mRNAs (Martinez et al., 2001) share with *lnc-SCA7* at least one predicted MRE in their 3' UTRs for a brain-expressed miRNA (Figure 5.15A); this proportion is significantly higher than that expected based on predicted shared MREs between the lincRNA and 10,000 gene sets, each containing 23 randomly sampled brain expressed mRNAs (permutation test, $p<10^{-4}$, Figure 5.15B). *miR-124* targeting sites are predicted for 10 of the 23 (44%) mouse mRNAs encoding STAGA subunits, of which 9 are also

conserved in human (Appendix Table A5.1). Moreover, I found that depletion of *lnc-SCA7* in N2A cells led to significantly (p<0.05) decreased levels for 17 of the 23 transcripts (Figure 5.16A) including *Atxn7* and *Atxn7l3*, as seen previously (Figure 5.4B). On the other hand, over-expression of *lnc-SCA7*-WT, but not *lnc-SCA7*-MUT, led to significant up-regulation of 16 out of 17 of these transcripts (p<0.05; Figure 5.16B).

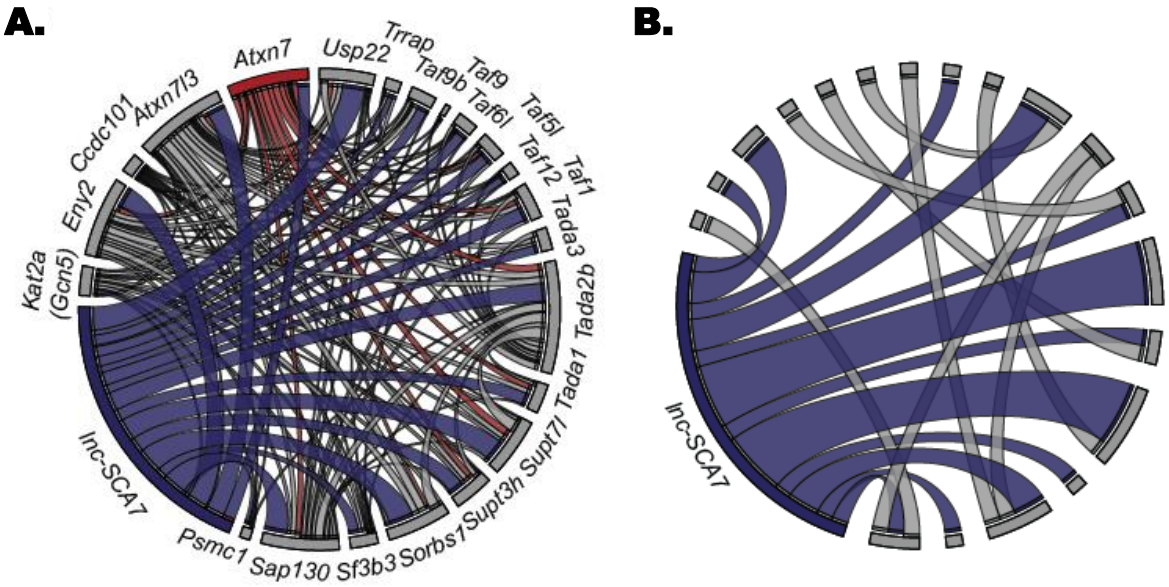


Figure 5.15 lnc-SCA7 post-transcriptionally regulates the abundance of STAGA-encoding mRNAs. (A) Circular visualization of predicted shared miRNAs between *lnc-SCA7* and STAGA encoding mRNAs. *lnc-SCA7* (blue), *Atxn7* (red) and remaining STAGA encoding mRNAs (grey) are represented individually on the outer circle. Each line connecting a pair of transcripts represents a response element for the same miRNA family (MRE) predicted in both transcripts; the thickness of the line indicates the proportion of predicted MREs in both transcripts that are shared between the two. (B) Representative circular visualization of the number of brain-expressed miRNA response elements shared between *lnc-SCA7* (blue) and the 3' UTRs of 10,000 randomly sampled sets of 23 brain-expressed protein-coding genes.

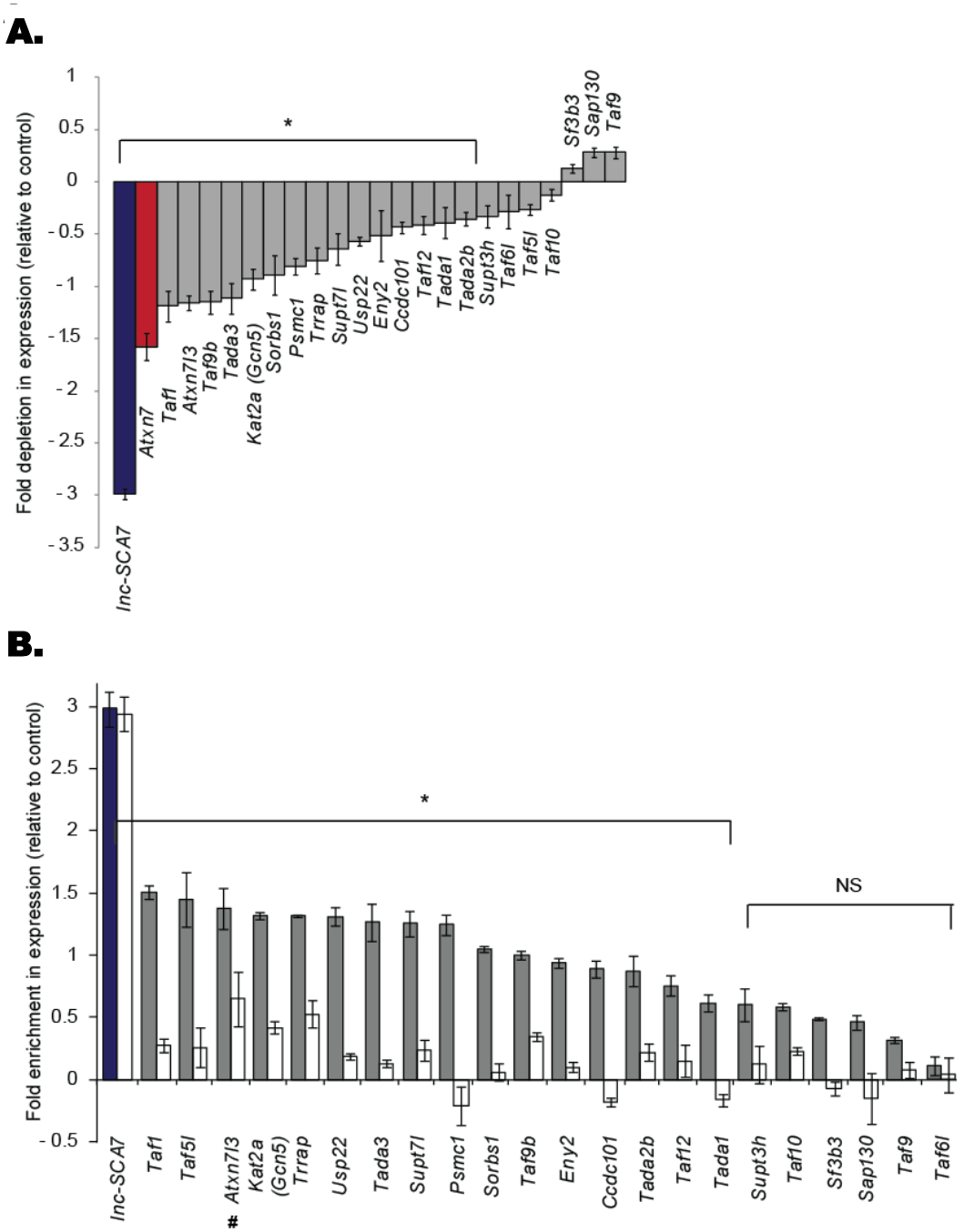


Figure 5.16 *Inc-SCA7* post-transcriptionally regulates the abundance of STAGA-encoding mRNAs. (A) Knockdown of *Inc-SCA7* (2.97-fold depletion, $p < 0.001$, blue) is associated with a significant ($p < 0.05$, *) decrease (≥0.36-fold depletion) in abundance of 16 out of the 23 mRNA encoding components of the STAGA complex (grey), including *Atxn7* (1.58-fold depletion, $p < 0.01$, red). (B) Overexpressing *Inc-SCA7-WT* (blue), but not *Inc-SCA7-MUT* (white), led to a significant increase in the abundance of 16 out of the 23 mRNA encoding components of the STAGA complex (dark grey), including *Atxn7* (Figure 5.4E, 5.10B). *Atxn7l3* (#) was significantly affected by over-expressing both constructs.

In summary, *Atxn7*/STAGA promotes *miR-124* transcription initiation; in turn this miRNA is key to the post-transcriptional crosstalk between *lnc-SCA7*, *Atxn7* and other STAGA mRNAs in the CNS (Figure 5.17), particularly in the tissues in which miR-124 is more highly expressed, namely the retina and the cerebellum (Figure 5.18A-B).

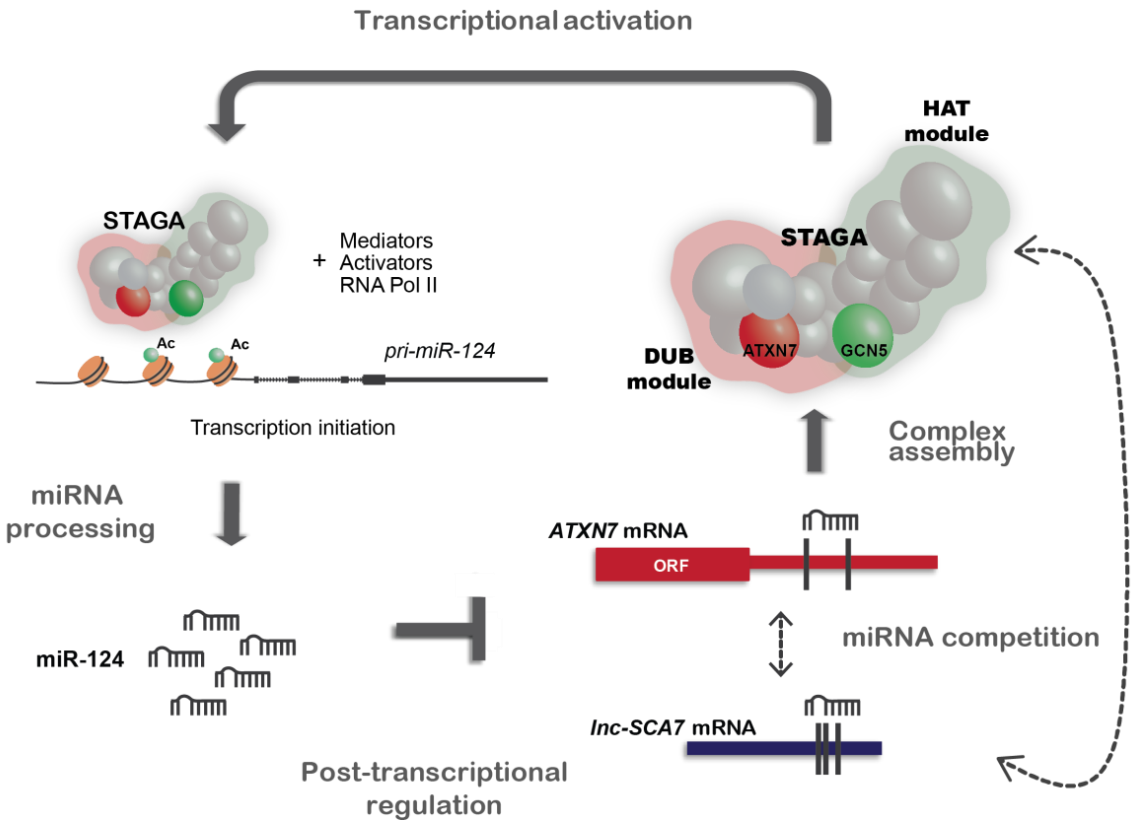


Figure 5.17 *Atxn7* abundance is controlled by a negative regulatory feedback loop involving two noncoding RNAs. *miR-124* (dark grey, bottom left corner) post-transcriptionally down-regulates *ATXN7* (red) and *lnc-SCA7* (blue), as well as other mRNA encoding STAGA complex subunits. The long noncoding RNA, *lnc-SCA7*, crosstalks (dashed arrows) via *miR-124* with *ATXN7* and other STAGA subunit encoding mRNAs, and thus acts as a post-transcriptional modulator of their transcript abundances. In turn, STAGA co-activates transcription of *miR-124* precursors, and thereby is a determining factor in mature *miR-124* abundance.

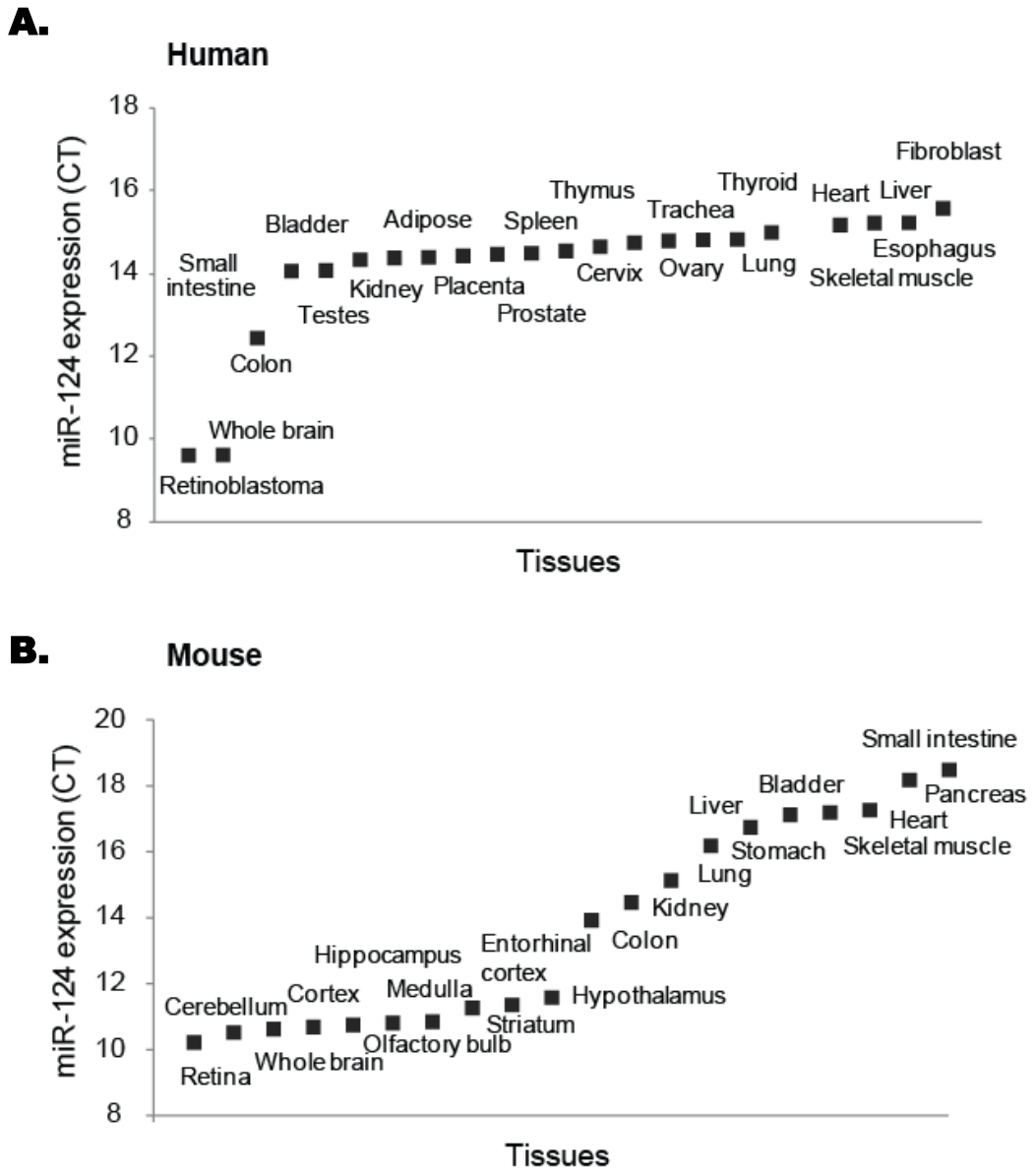


Figure 5.18 *miR-124* expression in human and mouse adult tissues. (A and B) *miR-124* expression level (y-axis) [(cross-threshold cycle (CT)] relative to *Gapdh*/*GAPDH* measured by qRT-PCR in (A) 22 human tissues, including a human fibroblast cell line derived from healthy human controls, and WERI retinoblastoma cells, and (B) 9 mouse CNS tissues, whole brain, and 10 mouse non-CNS tissues.

5.4.4 Noncoding RNAs mediate SCA7's tissue specific pathology

In SCA7, the polyQ-expanded ATXN7 protein is associated with decreased STAGA chromatin modification activity (McMahon et al., 2005; Palhan et al., 2005), and with reduced levels of transcripts from loci relying on ATXN7/STAGA transcriptional initiation (McCullough et al., 2012). The regulatory feedback loop (Figure 5.19A) revealed by my *in-vitro* analysis predicts that a decrease in STAGA activity in SCA7 would result in: (i) diminished *pri-miR-124* transcriptional initiation and, as a consequence, (ii) lowered mature *miR-124* and (iii) increased *lnc-SCA7* levels (Figure 5.19B).

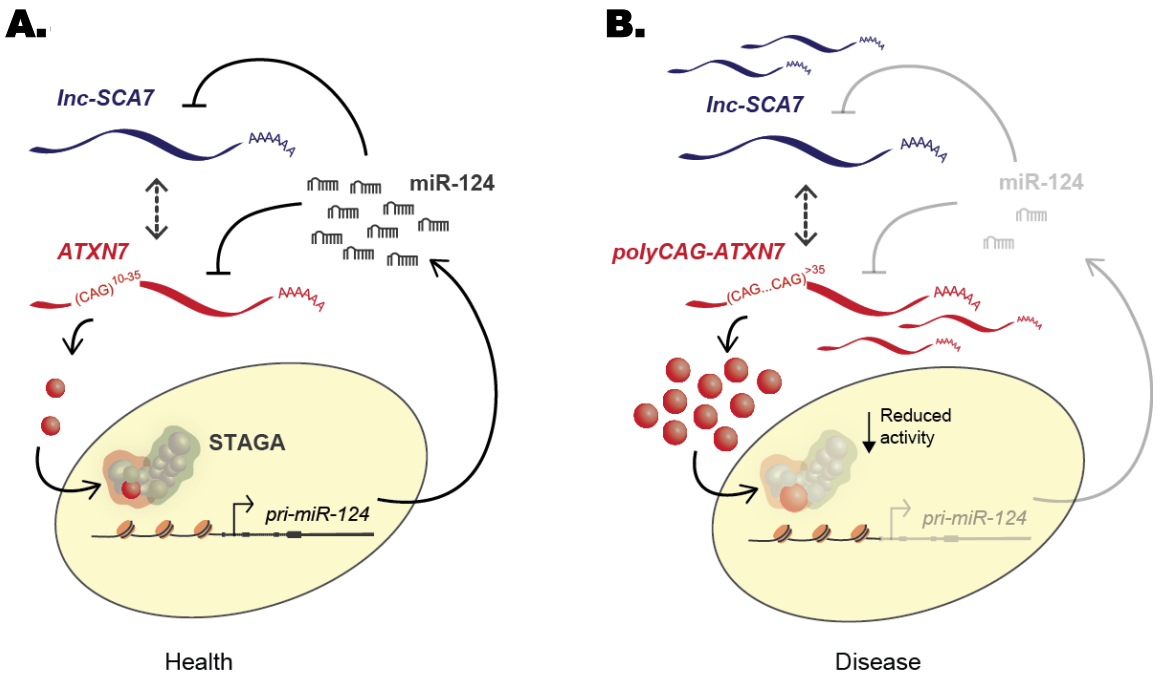


Figure 5.19 Crosstalking noncoding RNAs contribute to specific neurodegeneration in SCA7. (A) *miR-124* (dark grey) post-transcriptionally represses and mediates crosstalk (dashed arrows) between *ATXN7* (red) and *lnc-SCA7* (blue). STAGA co-activates transcription of *miR-124* precursors. Red circles represent ATXN7 protein. (B) In SCA7, incorporation of mutant polyQ-ATXN7 into STAGA reduces its activity, decreases *miR-124* abundance (light grey) and post-transcriptionally de-represses the miRNA's targets.

I validated these predictions in a human model of SCA7 by comparing the levels of *ATXN7*, *lnc-SCA7* and *miR-124* in fibroblasts derived from three SCA7 patients who carry 42, 49 or 55 polyQ *ATXN7* repeat expansions ($SCA7^{42Q/10Q}$, $SCA7^{49Q/10Q}$ and $SCA7^{55Q/10Q}$) against their levels in control fibroblasts (10 polyQ-repeats, $SCA7^{10Q/10Q}$). SCA7 patient-derived fibroblast cell lines ($SCA7^{42Q/10Q}$, $SCA7^{49Q/10Q}$, $SCA7^{55Q/10Q}$) were obtained from the University of Cape Town (UCT) and maintained by Dr. Lauren Watson and Dr. Miguel Varela. Expression levels of *miR-124* were reduced by two-fold, whereas transcript abundances of *lnc-SCA7* and *ATXN7* were increased substantially (by up to 1.8-fold and 5.2-fold, respectively) in these patients' cells (Figure 5.20A). Furthermore, decreased levels of endogenous miR-124 (0.38-fold reduction) in human fibroblasts associated with increase abundance of both *lnc-SCA7* and *ATXN7* (1.22- and 1.27-fold, respectively, Figure 5.20B), supporting the direct contribution of this miRNA to the post-transcriptional modulation of the 2 transcripts in fibroblasts.

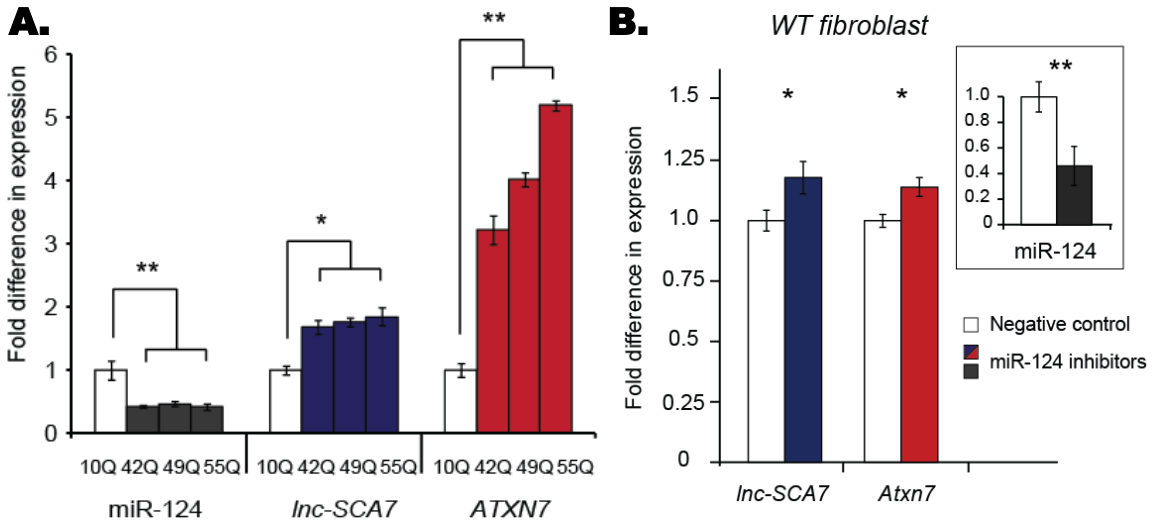


Figure 5.20 Contribution of noncoding RNAs to the tissue-specific pathology of SCA7 human patients. (A) *miR-124* abundance (dark grey) was significantly reduced and levels of *lnc-SCA7* (blue) and *ATXN7* (red) were successively increased in SCA7 patient fibroblasts with 42, 49, or 55 expanded ATXN7 polyQ repeats relative to healthy control (white). (B) Transfection of miR-124 inhibitors in human fibroblast cells (wild-type fibroblast with 10 polyQ repeats in ATXN7, top-right insert, dark grey) results in increased levels of both *lnc-SCA7* (blue) and *Atxn7* (red) relative to a negative miRNA transfection control (white).

Furthermore, I predicted that changes in *miR-124* abundance would result in the post-transcriptional de-repression of mutant *ATXN7* mRNA, and consequently, increased abundance of its encoded protein, with this effect being manifested more prominently in those tissues in which the miRNA most affects *ATXN7* gene abundance (Figure 5.19B). The crosstalk between *ATXN7* and *lnc-SCA7* in tissues where *miR-124* levels are high is likely enhanced, consistent with their observed higher correlation in the CNS (Figure 5.1B) and might amplify, post-transcriptionally, the impact of changes in miR-124 levels on ATXN7's abundance (Figure 5.19B). To test the validity of my predictions and the higher impact of noncoding de-regulation in the tissues primarily affected in disease, I took advantage of two established knock-in (KI) SCA7 mouse

models, homozygous SCA7$^{100Q/100Q}$ (Chen et al., 2012b) and heterozygous SCA7$^{266Q/5Q}$ (Yoo et al., 2003). The two models express full-length human *ATXN7* with 100 or 266 CAG (polyQ) repeats inserted into the endogenous mouse *Atxn7* locus.

Both mouse models exhibit typical ataxic symptoms, as well as the unique retinal degeneration specific to SCA7 (Yoo et al., 2003; Chen et al., 2012b). Consistent with what is seen in human patients (David et al., 1997), SCA7 mice that carry a higher number of polyQ repeats show earlier disease onset and increased disease severity (Yoo et al., 2003). I took advantage of the interval separating disease onset for the two SCA7 mouse models and tested our predictions using both symptomatic (aged 28 weeks, SCA7$^{100Q/100Q}$) and pre-symptomatic (aged 5 weeks, SCA7$^{266Q/5Q}$) animals.

First, I investigated whether mutations in ATXN7 decreases the binding of Gcn5/STAGA upstream of the promoter regions of the three *pri-miR-124*s with help from Dr. Keith Vance who performed all ChIP experiments. Indeed, in the cerebellum of the SCA7$^{100Q/100Q}$ mice, relative to wild-type matched control mice (SCA7$^{5Q/5Q}$), STAGA bound with significantly reduced levels (48-79% reduction) to all 3 *pri-miR-124* promoters (Figure 5.21A) as expected given the finding that *miR-124* loci are targets of the STAGA complex. Next I compared, between SCA7 KI mice and wild-type controls, the transcript abundances of *ATXN7*, *lnc-SCA7* and *miR-124* in nine tissues: the two CNS regions primarily affected in disease, namely the retina or cerebellum, four CNS regions largely unaffected by the disease (the striatum, olfactory bulb, cortex and spinal cord), and three non-CNS tissues (muscle, lung and liver). As expected, mature *miR-124* levels

were significantly decreased in the CNS of SCA7 mouse models where this miRNA is normally more highly expressed (26% to 77% reduction) resulting in elevated abundance of its targets, *lnc-SCA7* (by 1.4- to 3.8-fold) and *ATXN7* (by 1.3- to 2.6-fold, Figure 5.21B, Figure 5.22). As predicted, within the CNS the relative change between wild-type and SCA7 KI in *miR-124* levels and associated fold increases in the abundances of *lnc-SCA7* and *ATXN7* are highest in the two tissues primarily affected in disease, the retina and the cerebellum (Figure 5.21B, Figure 5.22).

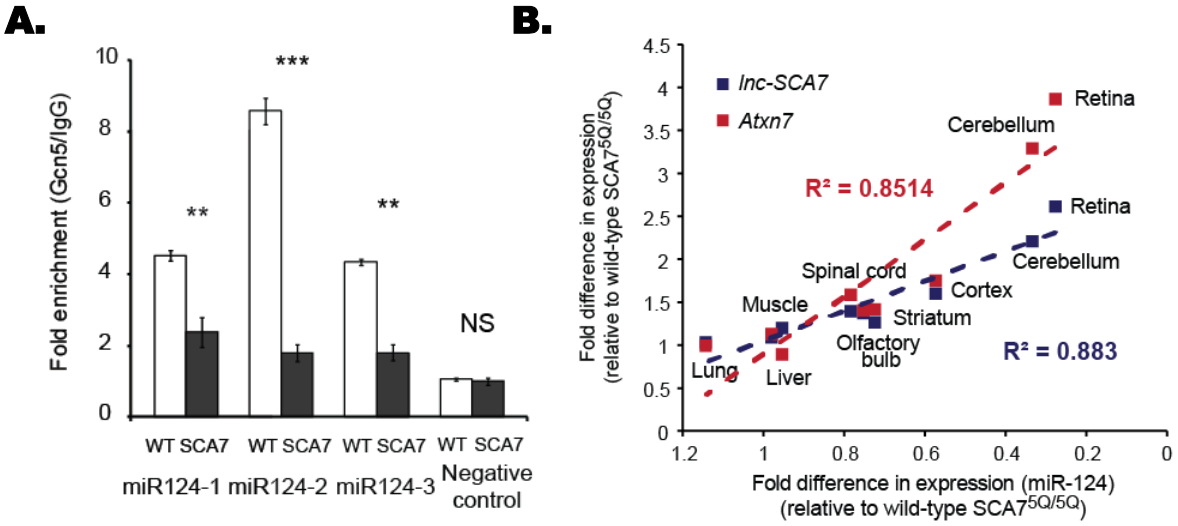


Figure 5.21 Contribution of noncoding RNAs to the tissue-specific pathology of SCA7 mouse models. (A) ChIP-qPCR revealed significantly decreased enrichment, relative to IgG control, in Gcn5 binding at *miR-124* promoters in SCA7$^{100Q/100Q}$ mice (dark grey) relative to controls animals (white). ChIP performed by Dr. Keith Vance followed qRT-PCR carried out by myself. (B) Correlation between the fold difference in expression levels between *lnc-SCA7* (Y-axis, blue) and *Atxn7* (y-axis, red) with *miR-124* (x-axis) in SCA7$^{266Q/5Q}$ mice as measured using qRT-PCR and relative to matched controls, SCA7$^{5Q/5Q}$.

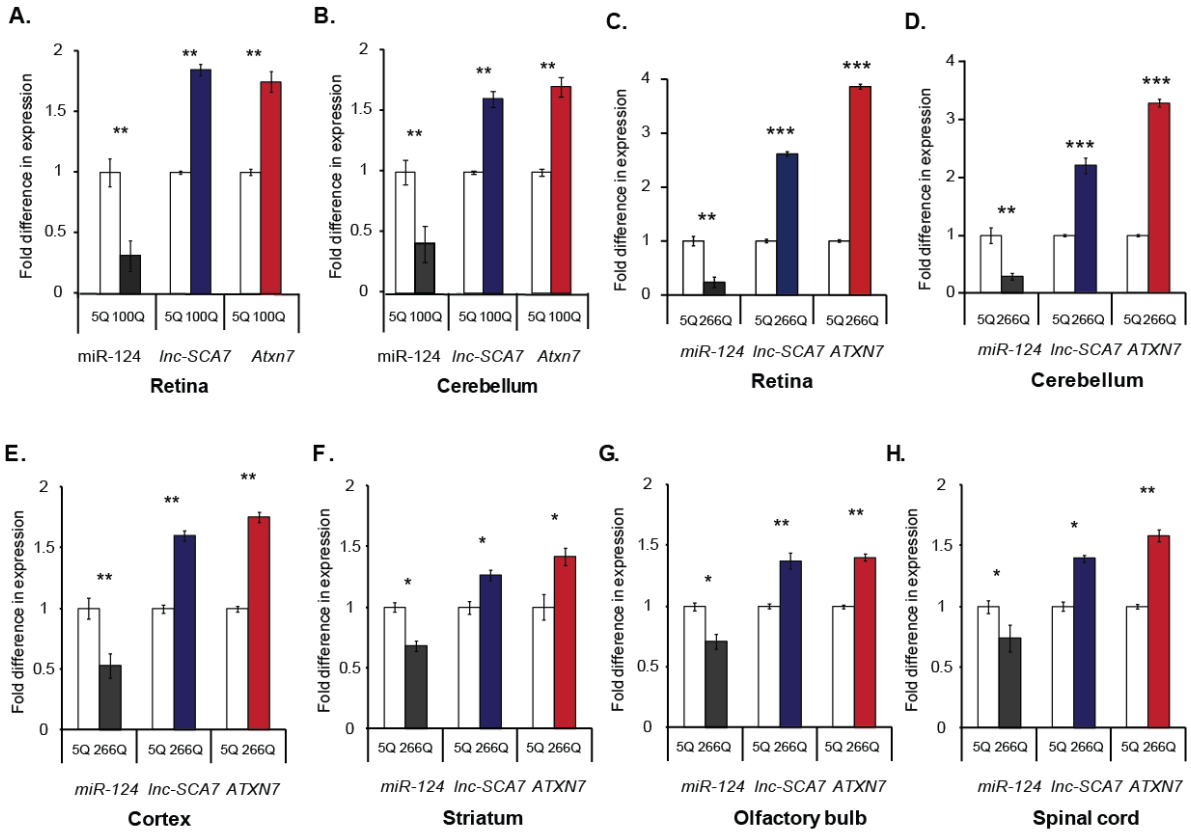


Figure 5.22 Expression levels of *Atxn7*, *lnc-SCA7*, *miR-124* and targets in CNS tissues of SCA7 mice. Fold difference in normalized expression (using *Gadph*) measured by qRT-PCR of *miR-124* (dark grey), *lnc-SCA7* (blue) and *ATXN7* (red) across tissues derived from SCA7 mouse models relative to matched controls (white): (A) Retina and (B) cerebellum of 28 weeks $SCA7^{100Q/100Q}$ mice, (C) retina, (D) cerebellum, (E) cortex, (F) striatum, (G) olfactory bulb and (H) spinal cord of $SCA7^{266Q/5Q}$ animals.

I tested the impact of these changes on the levels of known *miR-124* targets (Karginov et al., 2007; Makeyev et al., 2007; Hendrickson et al., 2008b; Agirre et al., 2009; Nakamachi et al., 2009; Yoo et al., 2009; Liu et al., 2011; Fang et al., 2012; Xia et al., 2012; Shi et al., 2013) and found that 8 and 12 out of the 13 transcripts tested were significantly (p<0.05) up-regulated in the cerebellum and retina, respectively (Figure 5.23). In contrast, in the three non-CNS tissues tested, muscle, lung and liver, I found no significant change in *ATXN7* or *lnc-SCA7* transcript levels in both SCA7 KI mice (Figure 5.24).

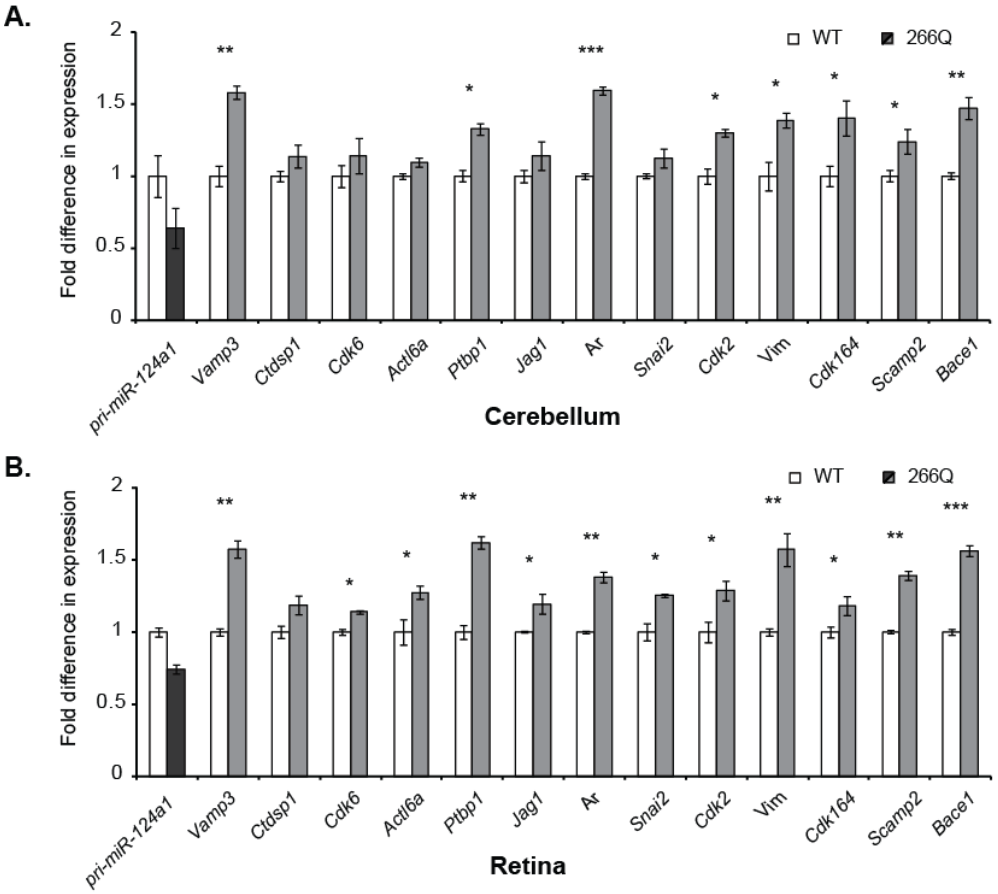


Figure 5.23 Expression levels of *miR-124* targets in the cerebellum and retina of SCA7 mice. (A and B) The fold-difference (y-axis) in expression in the (A) cerebellum or (B) retina of 5 week SCA7$^{266Q/5Q}$ mice (grey) relative to matched littermate SCA7$^{5Q/5Q}$ control mice (white) were elevated for 8 and 12 (of 13) known *miR-124* targets (light grey), respectively. Relative *pri-miR-124a1* (dark grey) levels were reduced in these tissues in SCA7 mice. * p<0.05, ** p<0.01, *** p<0.001, two tailed student's t-test.

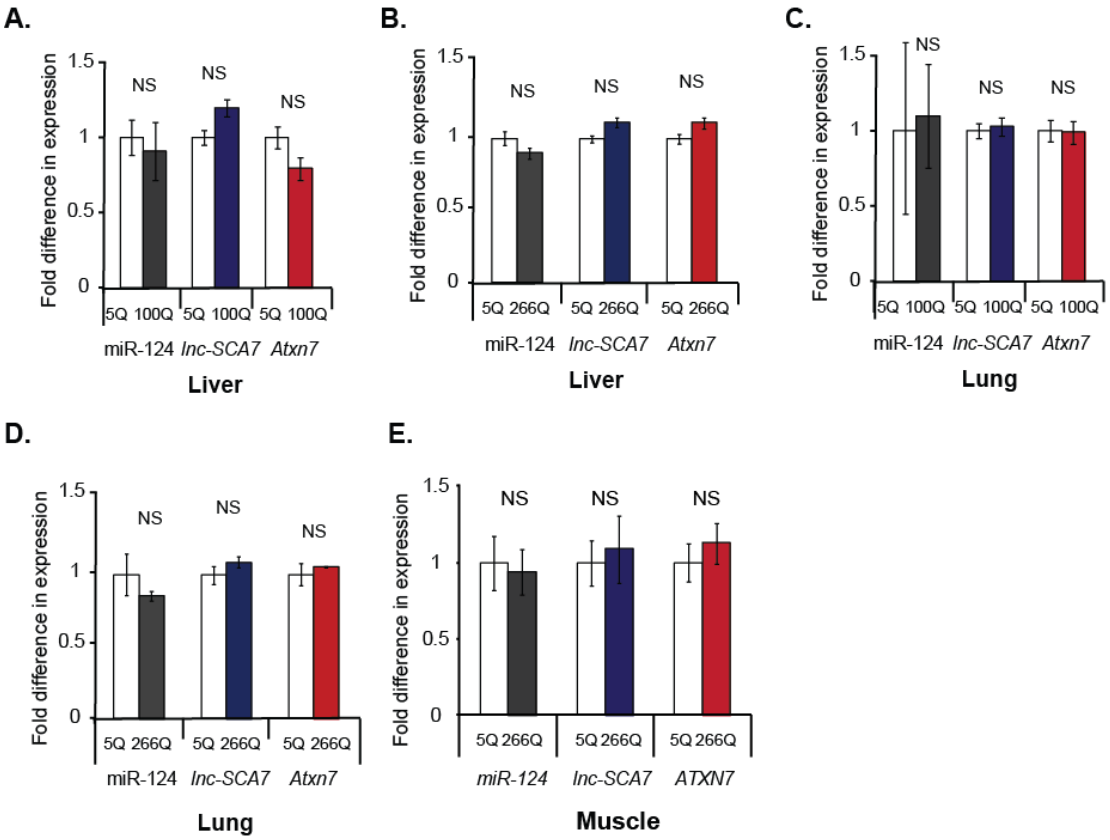


Figure 5.24 Expression levels of *Atxn7*, *lnc-SCA7*, *miR-124* and targets in non-CNS tissues of SCA7 mice. Fold difference in normalized expression (using *Gadph*) measured by qRT-PCR showed no significant differences (Not significant, NS) in the expression levels of *miR-124* (dark grey), *lnc-SCA7* (blue) and *ATXN7* (red) across tissues derived from SCA7 mouse models relative to controls. (A, B) livers, (C, D) lungs and (E) muscle of 28 week SCA7$^{100Q/100Q}$ or 5 week SCA7$^{266Q/5Q}$ mice relative to matched control mice (SCA7$^{5Q/5Q}$, white).

Absolute *ATXN7* and *lnc-SCA7* abundance, measured using droplet digital PCR in the retina, cerebellum, liver and lung of wild-type and SCA7 KI animals is strongly correlated with the levels estimated using standard qRT-PCR (Figure 5.25, Appendix Table A5.2). Interestingly, from *in-situ* hybridization experiments, I have also detected increased levels of mutant *ATXN7* mRNA and reduced

levels of *miR-124* in the retina and the cerebellum prior to SCA7 onset in the heterozygous SCA7$^{266Q/5Q}$ mice (Figure 5.26).

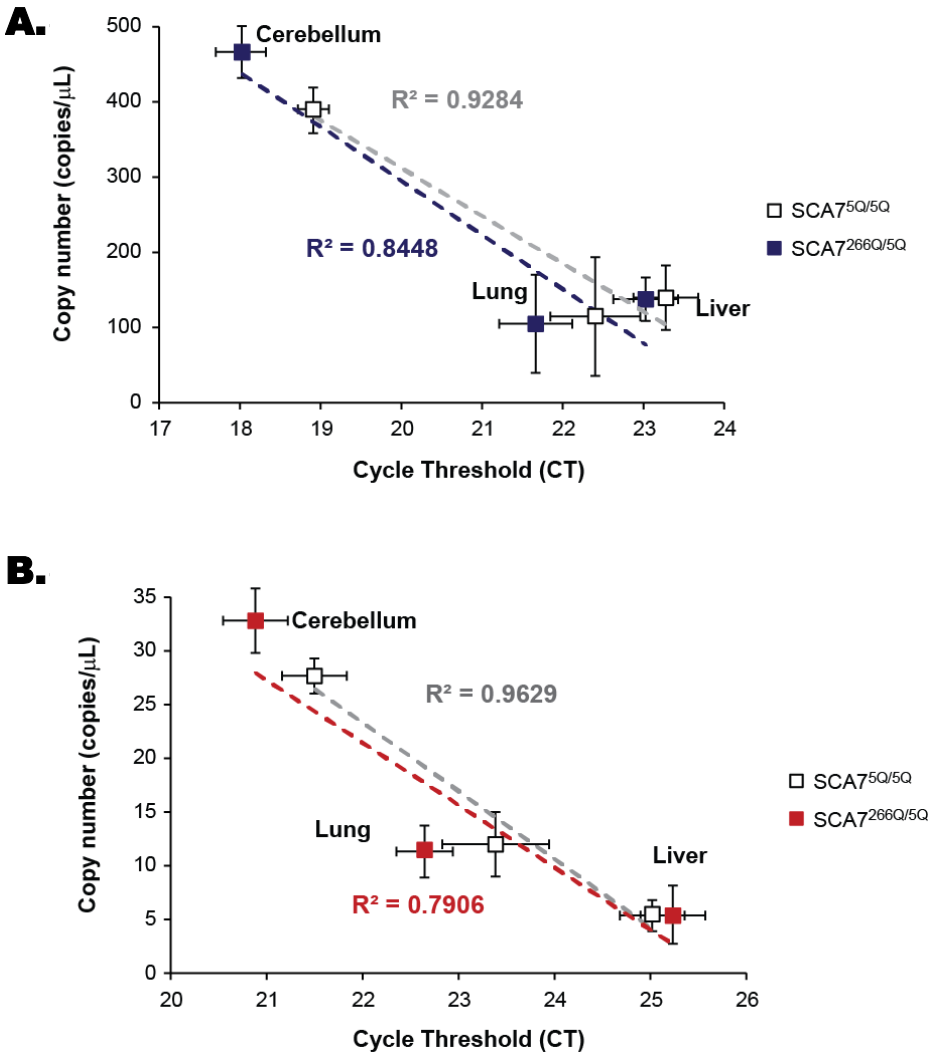


Figure 5.25 Expression levels of *Atxn7* and *lnc-SCA7* measured using digital droplet PCR correlates with standard qPCR measures. Expression levels of (A) *lnc-SCA7* (blue) and (B) *Atxn7* (red) quantified using ddPCR (y-axis, copy number per µL of cDNA) and qRT-PCR (x-axis, cycle threshold) in the cerebellum, lung, and liver of 5 week SCA7$^{266Q/5Q}$ and that of matched littermate SCA7$^{5Q/5Q}$ control mice (white) is strongly correlated.

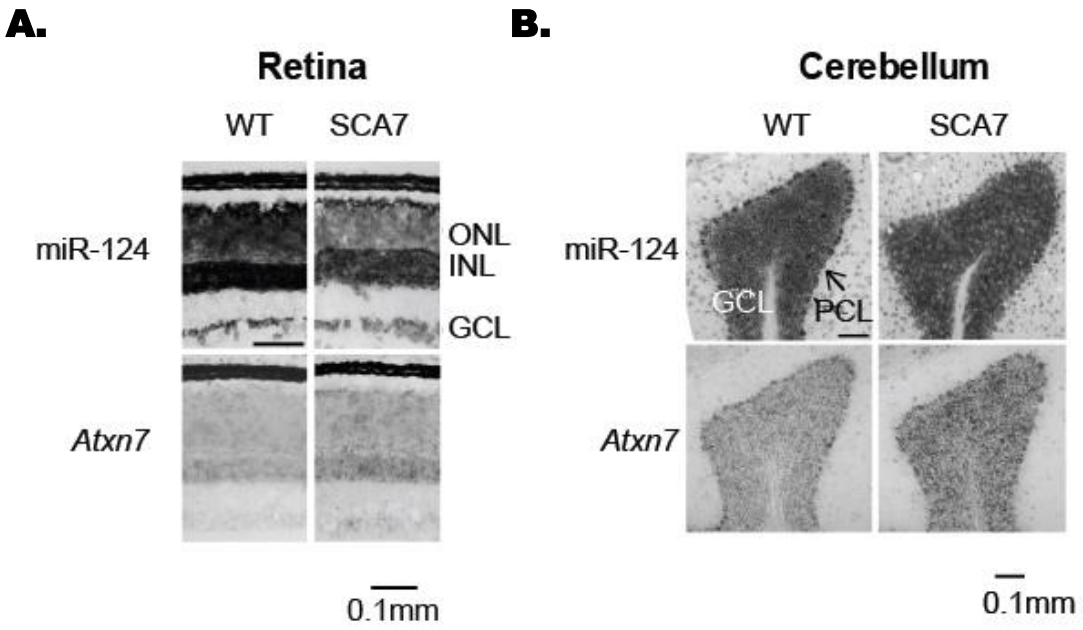
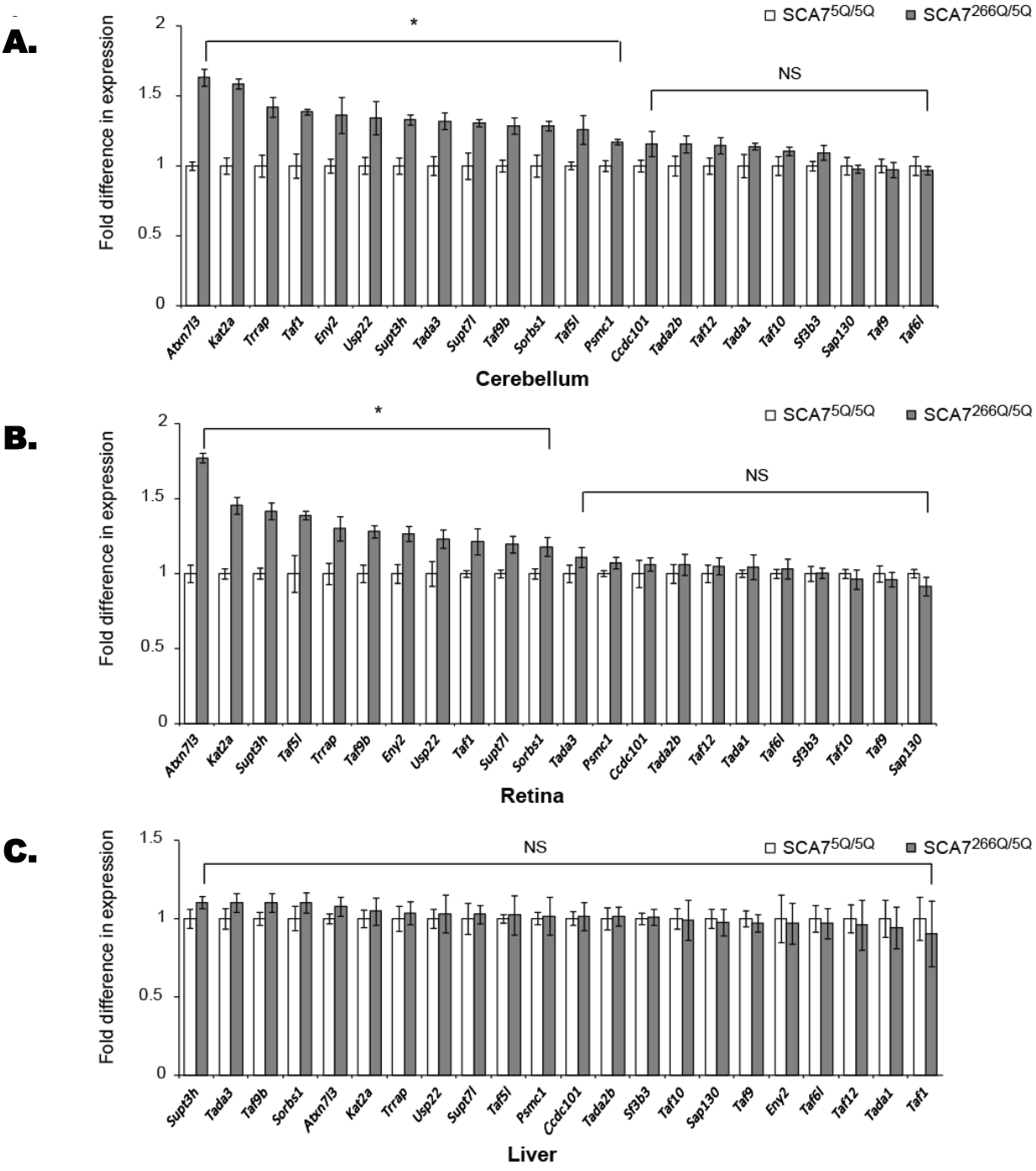


Figure 5.26 Contribution of noncoding RNAs to the tissue-specific pathology of SCA7$^{266Q/5Q}$ mouse model by *in-situ* hybridization. (A and B) RNA in-situ hybridization of *miR-124* and *Atxn7* in the retina and cerebellum of SCA7$^{266Q/5Q}$ mice and littermate SCA7$^{5Q/5Q}$ controls in the retina (A; ganglion cell layer (GCL), inner nuclear layer (INL), outer nuclear layer (ONL)) and cerebellum (B; granule cell layer (GCL), Purkinje cell layer (PCL)). Images produced by Dr. Peter Oliver.

In SCA7 mouse models, 75% and 63% STAGA-encoding subunits whose expression levels were affected *in vitro* by changes in *lnc-SCA7* levels (Figure 5.16) significantly increased in expression level in the cerebellum and retina, respectively (Figure 5.27A-B). Their abundance remained unchanged in the liver and lung of these mice (Figure 5.27C-D).


A.
SCA7^5Q/5Q
SCA7^266Q/5Q
Fold difference in expression
2
1.5
1
0.5
0
*
NS
Atxn7l3
Kat2a
Trrap
Taf1
Eny2
Usp22
Supt3h
Tada3
Supt7l
Taf9b
Sorbs1
Taf5l
Psmc1
Ccdc101
Tada2b
Taf12
Tada1
Taf10
Sf3b3
Sap130
Taf9
Taf6l
Cerebellum
B.
SCA7^5Q/5Q
SCA7^266Q/5Q
Fold difference in expression
2
1.5
1
0.5
0
*
NS
Atxn7l3
Kat2a
Supt3h
Taf5l
Trrap
Taf9b
Eny2
Usp22
Taf1
Supt7l
Sorbs1
Tada3
Psmc1
Ccdc101
Tada2b
Taf12
Tada1
Taf6l
Sf3b3
Taf10
Taf9
Sap130
Retina
C.
SCA7^5Q/5Q
SCA7^266Q/5Q
Fold difference in expression
1.5
1
0.5
0
NS
Supt3h
Tada3
Taf9b
Sorbs1
Atxn7l3
Kat2a
Trrap
Usp22
Supt7l
Taf5l
Psmc1
Ccdc101
Tada2b
Sf3b3
Taf10
Sap130
Taf9
Eny2
Taf6l
Taf12
Tada1
Taf1
Liver

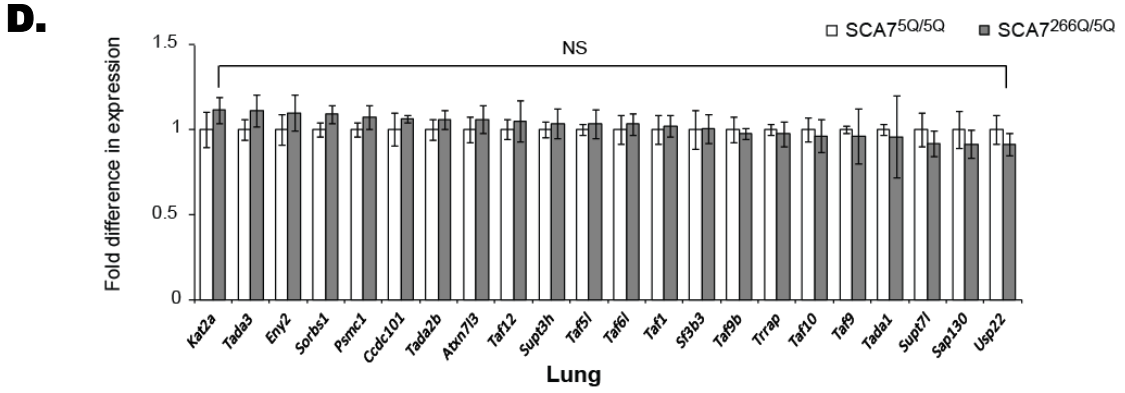


Figure 5.27 STAGA-encoding mRNAs abundance is increased in the cerebellum and retina, but not in liver and lung, of SCA7 mice. Expression levels of additional STAGA-encoding mRNAs, excluding *Atxn7*, in the (A) cerebellum, (B) retina, (C) liver and (D) lung of SCA7$^{266Q/5Q}$ mice in comparison to that of wild-type age and sex-matched SCA7$^{5Q/5Q}$ mice. * p<0.05; NS not significant.

Together, these findings are concordant with crosstalking noncoding RNAs contributing to the tissue-specific disease pathology of SCA7: differences in *miR-124*, *ATXN7* and *lnc-SCA7* levels are more pronounced in the two tissues, retina and cerebellum, that exhibit primary signs of cellular degeneration in SCA7 (David et al., 1997).

5.5 DISCUSSION

Why mutations in *ATXN7,* a ubiquitously expressed and housekeeping gene, lead to the specific degeneration of retinal and cerebellar neurons observed in SCA7 patients, has thus far remained unexplained. My *in vitro* and *in vivo* analysis in mouse suggests that post-transcriptional regulation by two crosstalking noncoding RNAs, *lnc-SCA7* and *miR-124*, mediates the tissue specificity of SCA7 pathology. Using two SCA7 mouse models, we showed that decreased levels of miR-124 owing to decreased transcriptional activation activity of the STAGA complex due to mutations in *ATXN7* lead to this gene's increased abundance particularly in the tissues where *miR-124* is most highly expressed, retina and cerebellum, and where it is expected to exert the strongest control on *ATXN7* transcript abundance, mirroring the cell-specific pathology of SCA7. Post-transcriptional crosstalk between *ATXN7* and *lnc-SCA7* is likely to amplify the impact of aberrant *miR-124* levels. Nuclear aggregates of mutant ATXN7 protein are often found in the retina and cerebellum of SCA7-affected individuals (Yoo et al., 2003) and although it remains unclear whether these inclusions are the underlying cause or else are a consequence of disease pathology, their presence and the mechanisms leading to their degradation are associated with increased cytotoxicity and disease severity (Zander et al., 2001). In SCA7 patients I propose first that increased levels of mutant ATXN7 protein result from its transcript's cell-specific post-transcriptional de-repression due to lowered *miR-124* abundance, and second that these high protein levels lead to increased cytotoxic nuclear inclusions, primarily in the retina and the cerebellum (Figure 5.19).

miR-124 has been previously associated with neurodegeneration. Its levels are known to be dysregulated in the brain of Huntington's disease patients (Packer et al., 2008). Furthermore, some Friedreich's Ataxia (FRDA) patients carry mutations in *miR-124* binding sites within the 3'UTR of their *FRDA* (*Frataxin*) gene and these are associated with mRNA upregulation *in vitro* (Bandiera et al., 2013). Here I have described how cerebellum and retina-specific decreases in *miR-124* levels in SCA7 patients contribute to cell-specific degeneration. I show that changes in *miR-124* levels also affect the abundance of other known miRNA targets including some that are involved in transcriptional regulation in the cerebellum (Chou et al., 2010) or in the control of retinal differentiation (Abou-Sleymane et al., 2006). I predict that dysregulation of *miR-124* targets contributes further to the neuronal specificity of the SCA7 phenotype.

While the contributions of miRNAs to transcriptional regulation and disease are becoming more recognized, those of their longer (>200nt) counterparts remain poorly understood. I have shown that a conserved lincRNA, *lnc-SCA7*, post-transcriptionally regulates the expression of *ATXN7* as well as many STAGA encoding genes.

Although the crosstalk between *lnc-SCA7* and *ATXN7* is likely to be ubiquitous, as highlighted by their correlated expression levels outside the CNS, it is greatly enhanced in the CNS, most prominently in the retina and the cerebellum, where *miR-124* is most abundant. This tissue specificity of the crosstalk between coding and noncoding transcripts supports a functional role for *lnc-SCA7* as a key modulator of *ATXN7* abundance in these tissues. Recently, deletion of the 12q21 chromosomal region containing the *lnc-SCA7* and *KCNC2* loci was

associated with familiar neurodevelopmental delay and cerebellar ataxia (Rajakulendran et al., 2013). Both genes were proposed to account for the complex neurodevelopmental phenotype observed in this family (Rajakulendran et al., 2013). Based on its most prominent crosstalk with *ATXN7* in the cerebellum, I propose that loss-of-function mutations in *lnc-SCA7* locus may underlie the ataxia-like symptoms of these individuals.

My results provide much needed insight into the contributions of noncoding RNAs to human disease, specifically those that confer the cell-specific disease pathology caused by mutations in ubiquitously expressed genes. Identifying additional noncoding RNAs that contribute to the tissue-specificity in other diseases should further improve our understanding of how RNA crosstalk modulates disease phenotypes.

CHAPTER 6

A primate-specific intergenic long noncoding, *lnc-ASD,* post-transcriptionally modulates the transcript abundance of several Autism Spectrum Disorder implicated genes.

6.1 ABSTRACT

Autism Spectrum Disease (ASD) is a complex neurodevelopmental disease caused by the crosstalk between multiple genetic and environmental factors. Hundreds of protein-coding genes have so far been associated with ASD but the genetics interactions between them remains unclear. I characterized an intergenic long noncoding RNA (lincRNA), *lnc-ASD*, whose expression levels were shown to correlate with a strong risk allele for ASD, and investigated how this lincRNA may modulate the transcript abundance of genes implicated in ASD. *Lnc-ASD* is primate-specific and acts as a competitive endogenous RNA (ceRNA) by competing for binding of a primate-specific microRNA (miRNA), *miR-1253,* with a set of transcripts whose genes have previously been implicated in autism. Here, I show that perturbation of *lnc-ASD* abundance *in vitro* affects the transcript levels of ASD-implicated genes and that this effect is dependent on the presence of *miR-1253* MREs within the lincRNA. Importantly, *lnc-ASD* contains an insertional polymorphism common in the human population that disrupts a *miR-1253* response element within the lincRNA that

affects its ability to modulate the transcript levels of these ASD-implicated genes. My results provide insights into how a lincRNA can post-transcriptionally regulates the transcript levels of multiple ASD-implicated genes and thus contribute to complex trait disorders.

6.2 INTRODUCTION

Autism Spectrum Disorders (ASD) are a group of complex neurodevelopmental disorders characterized by difficulties in social communication and interaction, restricted interests, repetitive activities, and unusual attachments to objects and routines (Piven et al., 1997; Abrahams and Geschwind, 2008; Geschwind, 2011). With an estimated median global prevalence of 1/160 (World Health Organization, 2013), ASD is the most common childhood neurodevelopmental disorder. These neurological disorders have early onset, typically within the first three years of life, and are four times more common in boys than in girls (Klauck, 2006). Although common characteristics, such as communication and social difficulties, are typically observed in ASD individuals, these disorders are manifested through a wide spectrum of symptoms, with diverse differences in the severity and symptoms of individual developmental impairments.

Although environmental factors, such as exposure of the fetus to hormones in the womb (Baron-Cohen et al., 2014), can contribute to the spectrum of ASD phenotypes (Enstrom et al., 2010; Hallmayer et al., 2011), the etiology of this disorder has a large genetic component (Rutter, 2000), as suggested by its

estimated heritability of 60%-90% and an approximately 15 to 30 times greater prevalence in siblings of ASD children (Szatmari, 1999; Ronald and Hoekstra, 2011). In addition, twin studies that evaluate autistic phenotypes between monozygotic (MZ) and dizygotic (DZ) twins, where higher concordance in MZ twins specifies genetic inheritance to be a predominant causative driver of the disease, reported an average of 64% concordance in MZ twins and 9% in DZ twins (Smalley et al., 1988). This suggests that interactions involving genetic factors contribute extensively to ASD.

Prior to the development of large-scale screens for genetic abnormalities, most studies focused on individual genetic loci implicated in disorders associated with behavioural phenotypes that can predispose people towards ASD, i.e. single gene variants that cause syndromic forms of ASD (reviewed in Huguet et al., 2013). In contrast to these genetic testing methods that examined single or few genetic regions, new technologies that aim to investigate the entire genome are largely unbiased as they are not focused on specific genes or candidates (reviewed in Bras et al., 2012). Array comparative genomic hybridization (CGH) and single-nucleotide polymorphism (SNP) arrays have been used in cytogenetic studies to investigate the burden of relevant inherited or *de novo* copy-number variants in large cohorts of ASD-affected individuals (reviewed in Miles 2011; Huguet et al., 2013). Currently, CGH identifies clinically relevant *de novo* cytogenetic abnormalities in 7%-10% of individuals with autism of unknown cause (reviewed in Miles, 2011). Subsequently, next-generation sequence technologies, such as whole-exome sequencing, allowed the estimation of the genome-wide contribution to ASD of *de novo* coding-

sequence mutations (reviewed in Huguet et al., 2013). From four large-scale exome sequencing studies, mutations in 65 associative genes were identified in ASD, accounting for a proposed 3.6-8.8% of individuals with ASD (Iossifov et al., 2012). The number of deleterious *do novo* mutations within these 65 genes were approximately two-fold higher in individuals with ASD than in their unaffected siblings (Huguet et al., 2013).

Genome-wide screening methodologies, including CGH arrays and exome sequencing, are not without limitations. Specifically, such technologies do not capture all exons of the genome (i.e. those with GC-rich sequences) and regions outside of annotated protein-coding genes (O'Roak et al., 2012a). In contrast to exome sequencing technologies, genome-wide association study (GWAS) allows the examination of common genetic variants, typically SNPs, across most of the genomic landscape (i.e. not only those within coding exons) that may be associated with a disease (Manolio, 2010; Maurano et al., 2012). Approximately 88% of disease-associated SNPs were mapped to non-coding regions, revealing a larger than anticipated role of for noncoding SNPs in diseases, although it remains unclear what proportion of these SNPs are causal variants (Hindorff et al., 2009). Common genetic variants found more frequently in ASD individuals compared to controls have been identified using GWAS, where each variant is assumed to contribute to increased disease risk (Freitag, 2007; Abrahams and Geschwind, 2008; Freitag et al., 2010). Importantly, in addition to protein-coding genes found to be implicated in ASD (Anney et al., 2010), GWAS have allowed the prediction of noncoding elements potentially involved in ASD pathophysiology.

In a recent study, Kerin and colleagues investigated a single GWAS variant (*rs4307059*) that was significantly (p=1x10^{-10}) associated with ASD (Wang et al., 2009a) and identified a noncoding RNA of ~4 kb expressed directly under the ASD association peak within chromosome 5p14.1 (Figure 6.1) (Kerin et al., 2012). This noncoding RNA overlaps, on the opposite strand, a retropseudogene sequence of the human protein-coding gene, moesin (*MSN*, chrX: 64,887,511-64,961,793, hg19, NM_002444), which has also been previously implicated in ASD (Garbett et al., 2008; Voineagu et al., 2011). The *rs4307059* genetic variant has been identified as a predictor for ASD-like behaviours, such as stereotyped conservation and reduced communication skills (St Pourcain et al., 2010), suggesting that this genetic locus may encode important information for communication phenotypes.

Named moesin pseudogene 1 (*MSNP1*, chr5: 25,909,414-25,913,399; NT_006576.16, GRCh38), this retropseudogene gene shares 94% sequence identity with *MSN* (Kerin et al., 2012). While no evidence of transcription was apparent for the retropseudogene (transcribed on the forward strand) in several human cell lines (HEK and neuroblastoma cell lines, SK-N-SH and SH-SY5Y) (Kerin et al., 2012), an intergenic long noncoding RNA (lincRNA) was found to be transcribed on the reverse strand of the *MSNP1* locus. This lincRNA was named in the original study as *MSNP1*-antisense (*MSNP1AS*) (Figure 6.1). To emphasize the association between this lincRNA and ASD, which goes beyond its ability to modulate MSN levels as I will later describe, I renamed *MSNP1AS* to lincRNA-associated-with-ASD, *lnc-ASD*.

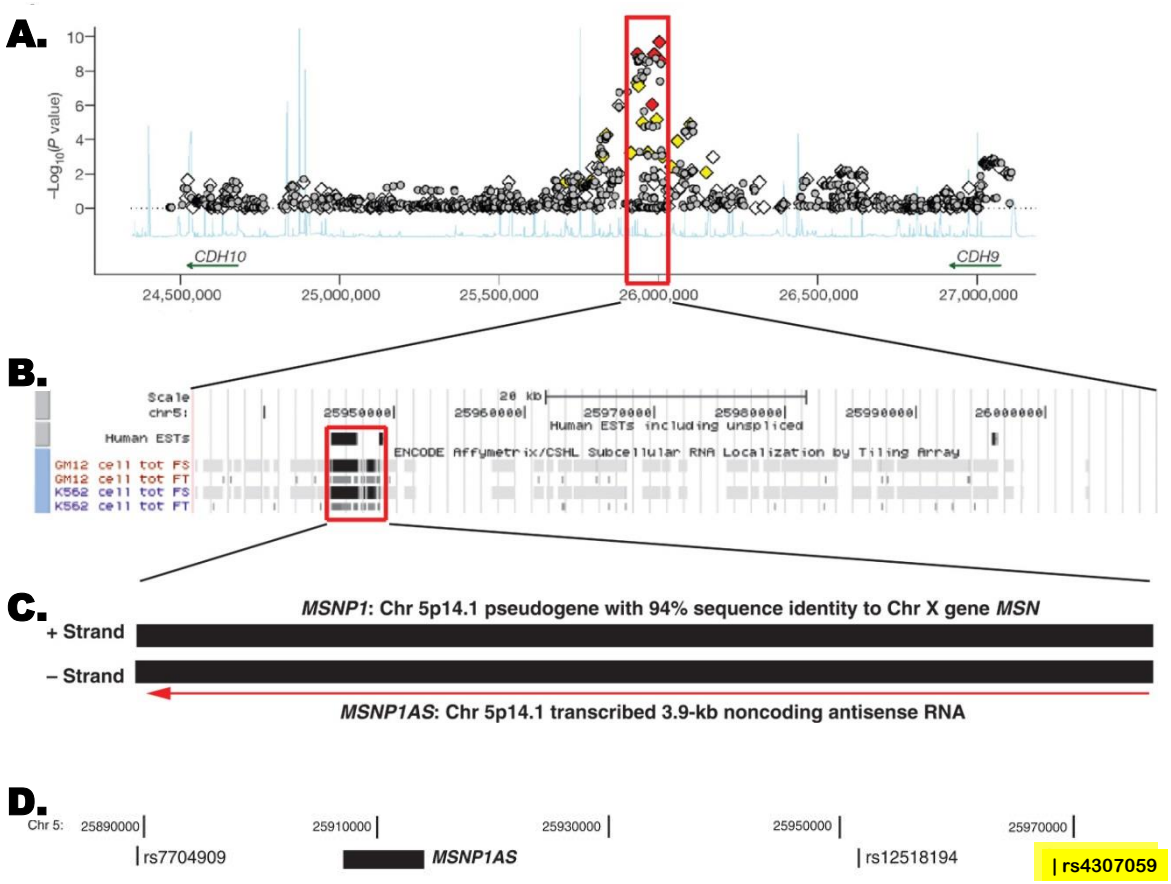


Figure 6.1 Genomic location of the GWAS significant genetic marker for increased ASD risk, which maps to an antisense transcript, *MSNP1AS* (*lnc-ASD*), of the pseudogene of *MSN*, *MSNP1*. (A) ASD-associated markers on chromosome 5p14.1 from the GWAS analysis (Wang et al., 2009a). (B) A noncoding RNA is transcribed from 5p14.1, shown by ESTs from the genome-wide ENCODE tiling array project. (C) The plus-strand of the region represents the pseudogene of the human protein-coding gene, moesin (*MSN*), moesin pseudogene 1 (*MSNP1*). The minus-strand produces a noncoding 3.9 kb RNA, *MSNP1*-antisense (*MSNP1AS*). (D) Genomic map of chromosome 5p14.1 illustrating the location of *MSNP1AS* relative to the genome-wide significant marker (*rs4307059*, highlighted) associated with increased ASD risk. Adapted from Kerin *et al.* 2012 .

In post-mortem brain samples of ASD individuals, the increase in levels of *MSNP1AS* was most pronounced, relative to controls, in individuals with the ASD-associated *rs4307059* T/T genotype compared to the C/C genotype, suggesting that this genetic variant may contribute (or is in linkage disequilibrium with a variant that contributes) to ASD risk (Kerin et al 2012, Figure 6.2A). In addition, the transcript levels of *MSNP1AS* were found to be significantly increased by 12.7-fold in the T/T genotype (Figure 6.2B) (Kerin et al., 2012). Similarly, the levels of its ancestral gene, *MSN*, were significantly increased (2.4-fold) (Figure 6.2C) in cases versus controls (Kerin et al., 2012). No correlation in levels was found for either *CDH9* or *CDH10*, two protein-coding genes on either side of *lnc-ASD*, with the ASD-associated *rs4307459* T/T genotype compared to the C/C genotype in cases versus controls, suggesting these two protein-coding genes do not account for the association of the genome-wide significant SNP to ASD (Kerin et al., 2012).

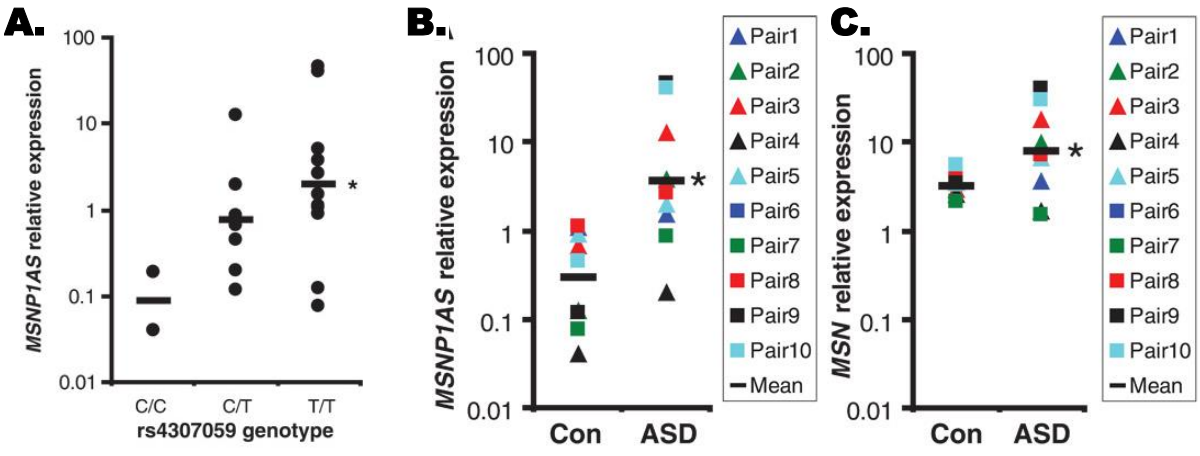


***Figure 6.2 Expression levels of MSNP1AS are increased and is* correlated with ASD-associated genotype *in post-mortem brain tissue of individuals with ASD*.** (A) Levels of *MSNP1AS* transcript is significantly increased in individuals with the ASD risk (*rs4307059*) associated genotype (T/T) compared to the C/C genotype. An intermediate level of *MSNP1AS* transcript was observed in individuals with heterozygous genotypes (C/T) that were not

significantly different from either homozygous genotypes (T/T and C/C). (B and C) Abundances of (B) *MSNP1AS* and (C) *MSN* are significantly increased in ASD individuals relative to controls. Adapted from Kerin *et al.* 2012 (Kerin et al., 2012).

Given the high sequence homology of *lnc-ASD* and *MSN*, it seemed likely this lincRNA arose in the human genome during the burst of retrotransposition that occurred during recent primate evolution (Zhang et al., 2003a). This study investigated the origin and evolution of *lnc-ASD* and provided important new insights into *lnc-ASD*'s molecular mechanism of function that I have further characterized. My work expanded the previous findings from Kerin et al. 2012 and provided new understanding of how this lincRNA may potentially contribute to the etiology of autism spectrum disorders.

6.3 MATERIALS and METHODS

I performed all the work described below, except where noted otherwise.

Identification of crosstalking miRNAs

A non-redundant list of 477 ASD-implicated genes was compiled by Dr. Allison Pioveson from publicly available data, including genes identified by genome-wide microarray, SNP array and exome sequencing data (Anney et al., 2010; Betancur, 2011; Neale et al., 2012; O'Roak et al., 2012b; Sanders et al., 2012; Kohler et al., 2014). The experimental candidates tested in the study are all associated with the syndromic forms of ASD: autism caused largely by genetic variants within single genes.

MiRNA response elements (MREs) within *lnc-ASD* and the list of 447 ASD-implicated genes for brain-expressed human miRNAs were predicted using TargetScan (version 6.1) (Garcia et al., 2011). Enrichment of predicted shared MREs between *lnc-ASD* and *MSN* was identified (see section 6.4 Results). The densities of predicted *miR-1253* MREs within the 477 ASD-implicated and 19774 protein-coding genes were calculated as the number of *miR-1253* MREs per kb of sequence.

Cloning and Mutagenesis

Directed mutagenesis of *lnc-ASD* overexpression constructs was generated by Dr. Esther Becker and myself as described in **Chapter 2**. Oligos used for PCR reactions are listed in Table 2.5.

Tissue culture, transfection, and gene expression profiling

Human neuroblastoma (SK-N-SH) cells were cultured as described in **Chapter 2**. Transfection of overexpression constructs of *lnc-ASD* and knockdown constructs of *lnc-ASD* (*si-lnc-ASD*) and that of *miR-1253* (*LNA-miR-1253*) with their respective controls (empty vector *pcDNA3.1(+)*, *si-NC*, and *LNA-NC*) were performed as described in **Chapter 2**. Subsequent extraction of total RNA and subcellular fractionated RNA, RNA reverse transcription, and gene expression profiling by qRT-PCR were performed as described in **Chapter 2**. Gene expression profiling of *lnc-ASD*, *MSN*, and *miR-1253* in 20 human tissues and SK-N-SH cell line was carried out as described in **Chapter 2**. qRT-PCR gene quantification was carried out in triplicate and the primers used are listed in Table 2.5.

Stability of RNA transcripts

To assess transcript stability, SK-N-SH cells (1.0×10^5 cells/ml) were seeded in 6-well dishes one day prior to the assay. RNA polymerase activity of SK-N-SH cells was blocked by adding 10 mg/mL actinomycin D (Sigma) in DMSO to the cell cultures, whereas control cells were treated with DMSO alone. Transcriptional inhibition assay of SK-N-SH cells was conducted for 24 h, where cells were harvested at times 0, 0.5, 2, 4, 8, 16, and 24 h. Cells were collected and total RNA extracted using miRNeasy and treated with DNase to remove genomic DNA following the manufacturer's instructions (Qiagen). The rapidly degraded *MYC* mRNA (Dani et al., 1984) and the relatively stable *ATP5E* mRNA were used as controls for the transcriptional inhibition assay

(Clark et al., 2012). Construct designs are illustrated in Table 2.1. Primers used in Chapter 6 are listed in Table 2.5.

Evolutionary analyses

Blat (Meyer et al., 2012) was used to search for homologs of human *lnc*-ASD (chr5: 25,909,414-25,913,399) in its syntenic genomic regions within other primate species by Dr. Ana Marques. Orthologs of *lnc-ASD* were considered to be present in non-human primate species if the genomic structure of human *lnc-ASD* was maintained within its synthetic genomic region in the non-human primate species.

The sequence constraint of *lnc-ASD* and *MSN* between the primate species was calculated by estimating their nucleotide substitution rate relative to that of randomly simulated (n=1,000) neighbouring ancestral repeat regions matched for G+C content and sequence length.

The insertional polymorphism (rs36113112, chr5: 25,910,483) within human *lnc-ASD* that disrupts a predicted *miR-1253* MRE was discovered from the cloning of *lnc-ASD* by Dr. Esther Becker and later confirmed using the 1000 Genome Project data by Dr. Ana Marques. The polymorphism corresponds to a 3 bp motif (GAA/AGA/AAG) that occurs in an AG-rich repetitive tract of ~12 nucleotides. The evolutionary analysis on the polymorphism was performed by Dr. Erika Kvikstad and Dr. Gerton Lunter.

6.4 RESULTS

6.4.1 lnc-ASD is primate-specific

First, the origin of *lnc-ASD* was investigated *in-silico* in mammals. Since the *lnc-ASD* locus in humans is located between two protein-coding genes, *CDH9* and *CDH10* (Figure 6.1A, 6.3), BLAT was used to search for a single exonic gene (a hallmark of retrotransposed genes) homologous to *MSN* that resides between the orthologs of human *CDH9* and *CDH10*. *MSN*, *CDH9* and *CDH10* are all found to be present in primate mammals. The lincRNA was found only in primate species that are closely related to the humans, including chimpanzee, orangutan, and likely gorilla (Figure 6.4). The poor quality of the gibbon genome does not allow the identification of the presence or absence of *lnc-ASD* in this species. However, in *Rhesus macaque*, no evidence of any locus homologous to *lnc-ASD* was found. Hence, we conclude *lnc-ASD* likely arose after the split between hominoids and old world monkeys.

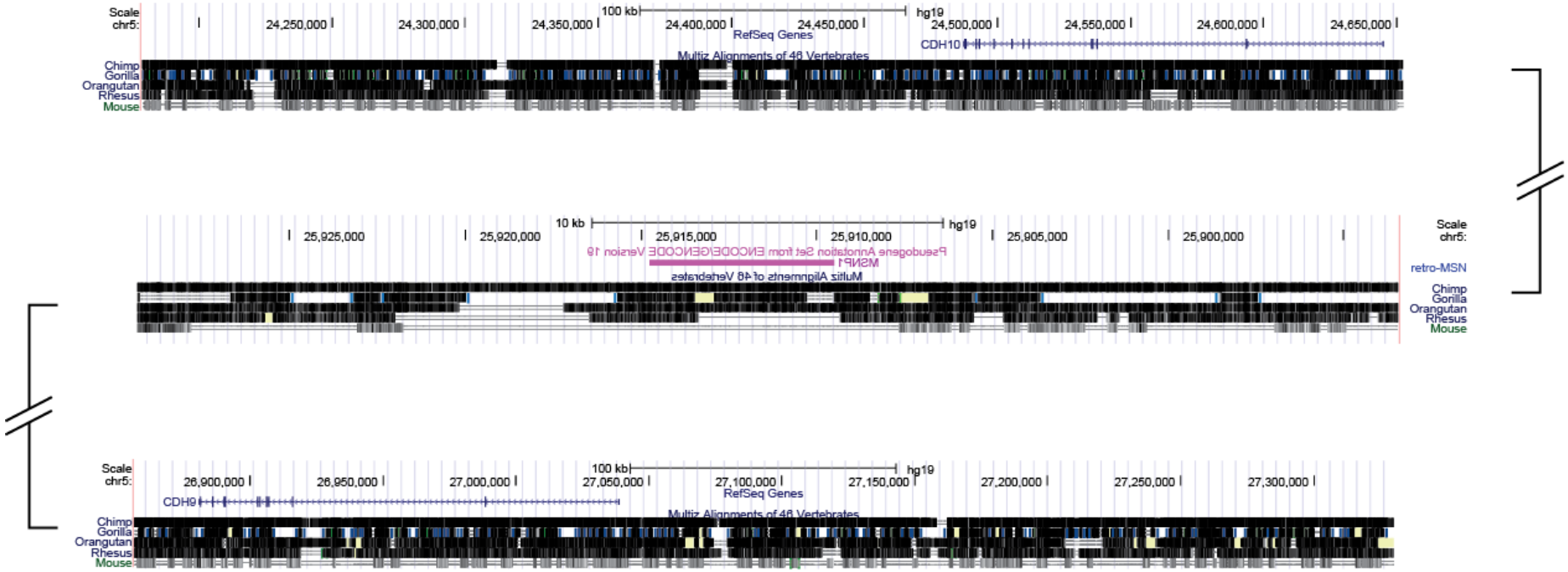


Figure 6.3 Genomic organization of the *lnc-ASD* locus across species. UCSC genome browser (Dreszer et al., 2012) depicting the genomic organization of *lnc-ASD* (chr5:25,909,414-25,913,399), which is located in between two protein-coding genes, *CDH10* (chr5:24487209-24645085; NM_006727) and *CDH9* (chr5:26880709-27038689; NM_016279) in human (hg19). Sequence conservation of the same region is illustrated in Chimpanzee (panTro4), Gorilla (gorGor3), Orangutan (ponAbe2), Rhesus macaque (rheMac3), and Mouse (mm10). No sequencing reads are mapped to the genomic location of *lnc-ASD* in genomes of the Rhesus macaque and the Mouse, illustrating the absence of the retrotransposed pseudogene in these species. The sequence assembly of the Gorilla genome is of insufficient quality to determine whether *lnc-ASD* is conserved in gorilla.

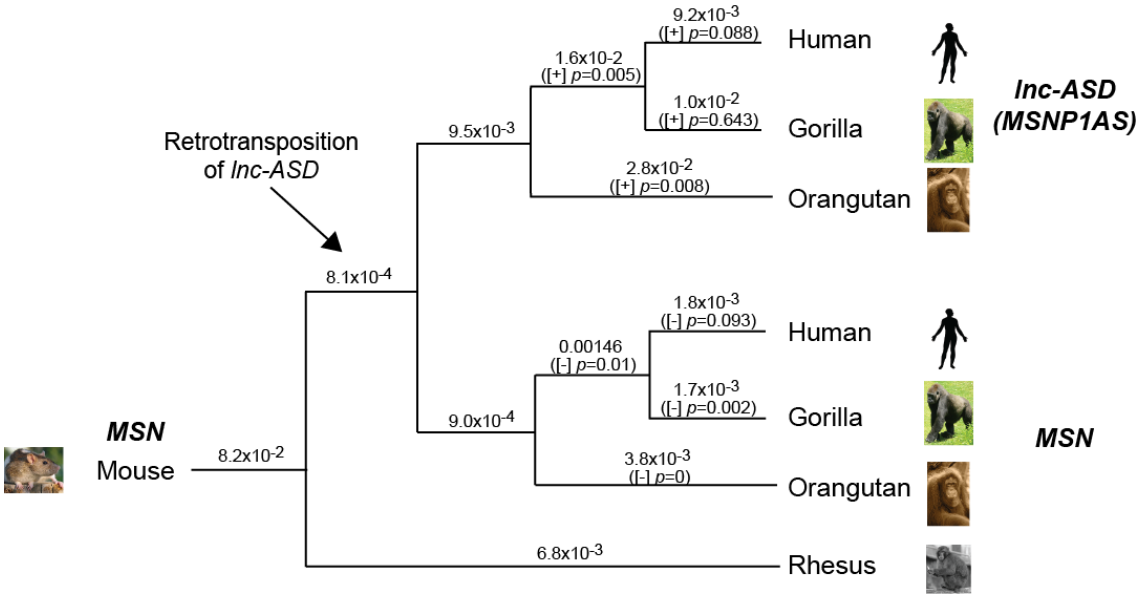


Figure 6.4 A phylogenetic tree of *lnc-ASD* and *MSN*. Evolutionary tree of *lnc-ASD* and *MSN* depicting that the lincRNA is found only in the primate lineage (Human, Chimpanzee, Gorilla and Orangutan), while its ancestral protein-coding gene, *MSN*, is present widely across placental mammals. Values on the tree branches indicate the substitution rates. Values in brackets denote the empirical *p* value of the substitution rate of the genes relative to that of neighbouring neutrally evolving sequences (estimated using neighbouring ancestral repeat sequences with similar G+C content and simulated 1,000 times to the same length as the gene of interest); $p<0.05$ denotes the sequence of interest is evolving at a significantly different rate than neighbouring neutral sequences: [-] indicates the sequence is under purifying selection, while [+] suggests the sequence is under accelerated evolution. This phylogenetic tree was built using DNA maximum likelihood (DNAML) program in the PHYLogeny Interference Package (PHYLIP) (Felsenstein, 1993). Sequence substitution rate was calculated using baseml of the PAML package (Yang, 1997).

The analysis of the sequence of *lnc-ASD* revealed that this lincRNA evolved rapidly. The substitution rate observed at this locus is significantly higher (p<0.05, empirical p test) than for neighboring matched neutrally-evolving sequences (Figure 6.4), implying it may be evolving at an accelerated rate. While the protein-coding ancestral *MSN*, accumulated significantly fewer substitutions than its neighbouring neutral sequences, consistent with its evolution under negative purifying selection (p<0.05, empirical p value) (Figure 6.4). Along with its described association with ASD (Kerin et al., 2012), the finding that *lnc-ASD* is specific to primates, and has evolved rapidly renders this lincRNA an attractive and atypical candidate to further investigate its potential contribution to ASD.

6.4.2 A prevalent polymorphism within lnc-ASD

Examination of human sequence population data (1000 Genome Project) revealed that *lnc-ASD* contains an insertional polymorphism that disrupts a predicted miRNA response element for *miR-1253* located at the 3' end of the lincRNA (Figure 6.5). The polymorphism corresponds to a 3 bp motif (GAA/AGA/AAG) within a longer AG-rich repetitive tract of ~12 nucleotides. This insertional polymorphism is estimated to segregate in the different populations of the 1000 genomes data at frequencies of 0.97 (JPTCHB), 0.69 (YRI), and 0.87 (CEU).

Investigation of the evolutionary history of the polymorphism within *lnc-ASD* revealed the number of copies of the motif was highly variable amongst primate species. By comparing the indel motif within *lnc-ASD* in human, chimpanzee,

gorilla and orangutan, to the same motif within the parental protein-coding gene, *MSN*, it appears that the insertional polymorphism has experienced at least two indel events within the last 16 million years (MY, divergence time between human and orangutan). Given that, the average indel mutation rate is 1.6×10^{-4} indel/site/MY (assuming a human-chimpanzee divergence time of 6 MY), the observation of 2 or more indel mutations within a sequence of 12 nt in length is not expected by chance. After partially controlling for sequence mutability observed within this repetitive sequence (i.e. repetitive sequences are more indel-prone), it appears that this polymorphism site may be evolving at a faster rate than expected under neutral evolution, although the locus did not show evidence of a selective sweep.

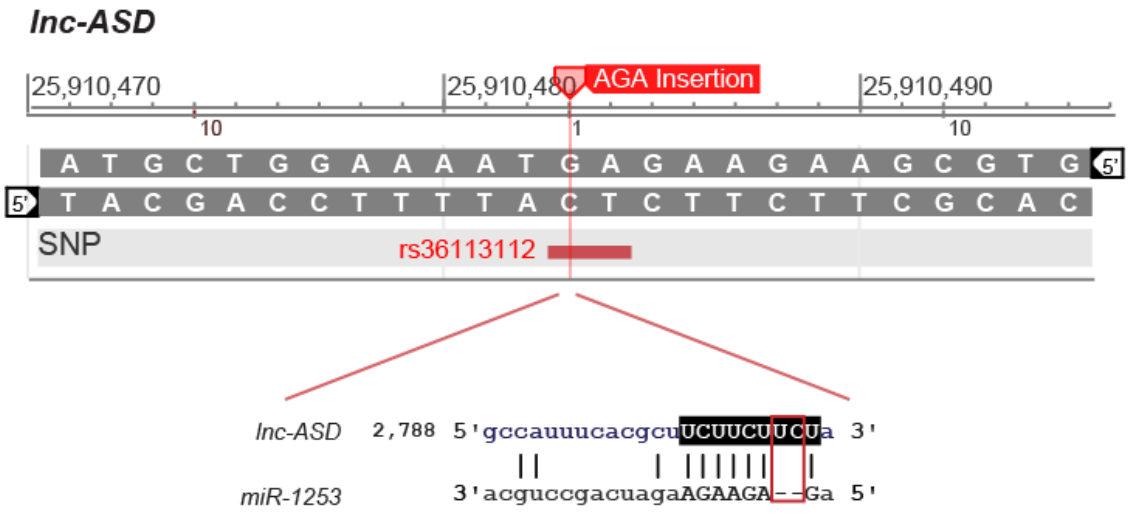


Figure 6.5 Insertional polymorphism at the 3' end of *lnc-ASD* disrupts a *miR-1253* MRE. The insertional mutation (rs36113112, chr5: 25,910,483) within *lnc-ASD* (chr5: 25,909,414-25,913,399) is shown on the 1000 Genome Browser (Genomes Project et al., 2010) at estimated frequencies of 0.97 (JPTCHB), 0.69 (YRI), and 0.87 (CEU). The polymorphism corresponds to a 3 bp motif (GAA/AGA/AAG) within a AG-rich repetitive tract of ~12 nucleotides. The polymorphism disrupts a MRE for *miR-1253* as shown by gaps aligned to the seed binding region of *miR-1253* (black box).

6.4.3 lnc-ASD and miR-1253 are each highly expressed in the brain and are predominated localized in the cytoplasm

Given the apparent accelerated rate of evolution at the disrupted *miR-1253* MRE within the *lnc-ASD* locus, its association with ASD, coupled with the previous findings that demonstrate the prevalence of ceRNA-acting lincRNAs (lnceRNAs) (**Chapter 4**) and their potential contribution to human disease (**Chapter 5**), I investigated the possibility that *lnc-ASD* serves as a miRNA decoy, specifically for *miR-1253*, for *MSN*, a protein-coding gene previously found to be associated with ASD (Garbett et al., 2008; Voineagu et al., 2011).

In order for *lnc-ASD* to act as a sponge for *miR-1253*, these noncoding transcripts should be found within the same subcellular location of the same cell/tissue types. First, I tested the subcellular localization of the lincRNA. In the human neuroblastoma cell line, SK-N-SH, *lnc-ASD* is predominantly localized in the cellular cytoplasm (3.9-fold enrichment relative to the nuclear fraction). A comparable enrichment in the cytoplasm was also observed for *MSN* (3.4-fold enrichment relative the nuclear fraction), which is expected for a protein-coding mRNA (Kohler and Hurt, 2007b) (Figure 6.6A). Primers used to quantify levels of *lnc-ASD* and *MSN* were designed with minimal sequence similarity between the genes (Figure 6.6B). *Malat1*, a nuclear-retained long noncoding RNA (Ji et al., 2003) was used as negative control (Figure 6.6A). *Lnc-ASD*'s cytoplasmic-enrichment indicates that it is unlikely to regulate transcription.

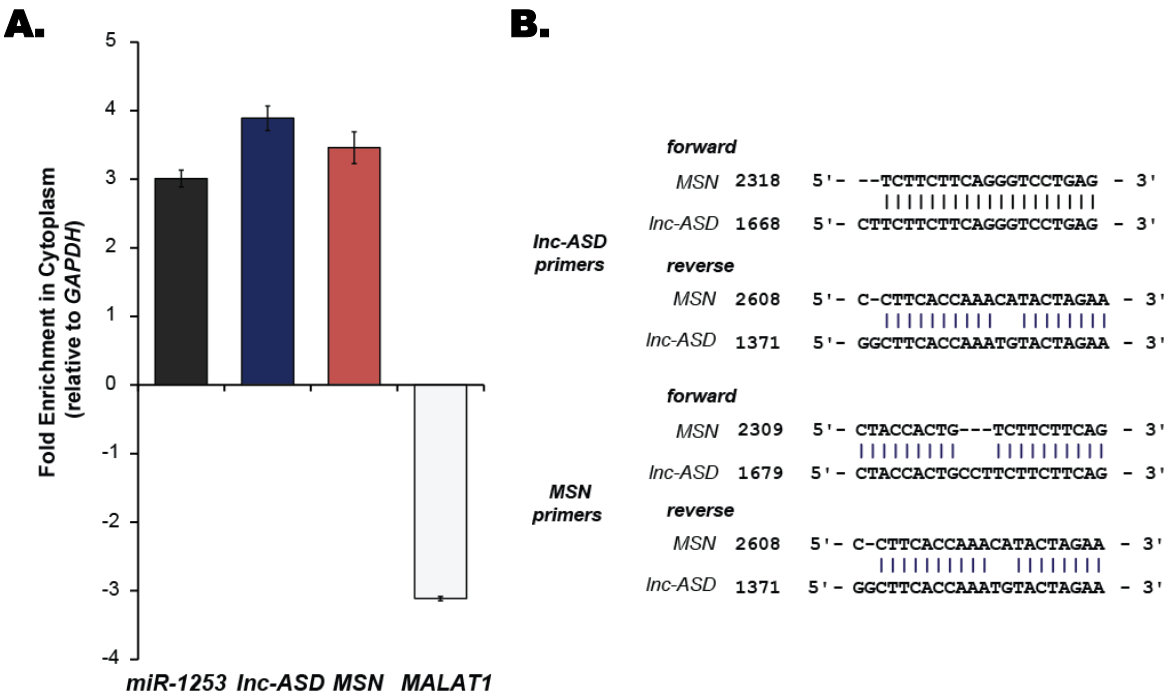


Figure 6.6 Subcellular localization of *lnc-ASD* in SK-N-SH cells. (A) All of *miR-1253* (dark grey), *lnc-ASD* (blue) and *MSN* (red) are predominantly enriched in the cytoplasm (3.0-fold, 3.9-fold, and 3.4-fold, respectively) relative to the expression of *GAPDH*. The nuclear-retained long noncoding RNA, *MALAT1* (3.1-fold enrichment in the nucleus, white), is used as control. (B) The qRT-PCR primers used to detect transcript levels of *lnc-ASD* and *MSN*. Because of the high sequence identity between *lnc-ASD* and *MSN* (94%), qRT-pCR primers were designed at regions that minimize sequence identity between the transcripts (Table 2.5).

Furthermore, consistent with the result reported in Kerin et al. 2012, I found a significant correlation between levels of *lnc-ASD* and its ancestral protein-coding gene, *MSN*, across a panel of 20 human tissues (R^2=0.68, p<0.001, Pearson's test, Figure 6.7). Specifically, both the lincRNA and *MSN* were most highly expressed in the brain, suggesting they may have functional roles there (Figure 6.7).

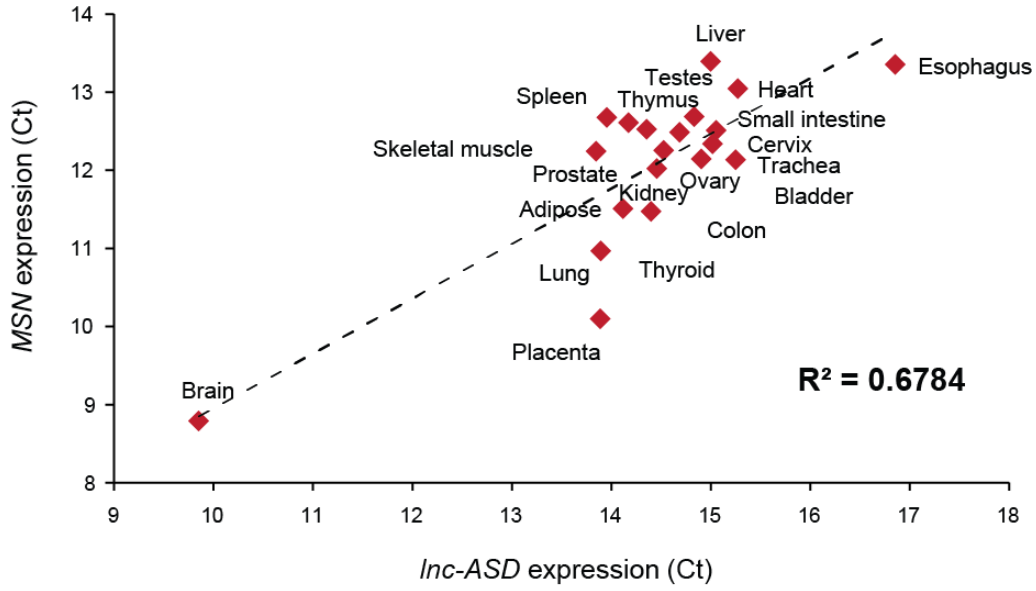


Figure 6.7 Expression levels of *lnc-ASD* and *MSN* are significantly correlated. The levels of *lnc-ASD* and *MSN* measured by quantitative real-time PCR (qRT-PCR) using sequence specific primers (Table 2.5) [cross-threshold cycle (Ct)] are significantly correlated (R^2=0.68, p<0.001, Pearson's test) in 20 human tissues (adipose, bladder, brain, cervix, colon, esophagus, heart, kidney, liver, lung, ovary, placenta, prostate, skeletal muscle, small intestine, spleen, testes, thymus, thyroid, and trachea).

Next, in addition to the observation that *lnc-ASD* is highly expressed in the cytoplasmic fraction of brain cells, I profiled the expression pattern of *miR-1253* across human tissues to test whether the miRNA is expressed in the same tissue and subcellular compartment as the lincRNA, which is necessary for miRNA-mediated crosstalk to occur between *lnc-ASD* and other *miR-1253* targets. Similar to *lnc-ASD* (Figure 6.7). *miR-1253* was also observed to be most highly expressed in the brain (Figure 6.8) and enriched in the cytoplasmic compartment of SK-N-SH cells (3.0-fold enrichment, Figure 6.6), supporting its post-transcriptional regulation of *lnc-ASD*. Furthermore, like *lnc-ASD*, *miR-1253* is also likely primate-specific, with no evidence of the presence of its precursor sequence (*pr-miR-1253*) in other mammalian lineages (*pri-miR-1253* is found in *Rhesus macaque* but not in mouse, data not shown).

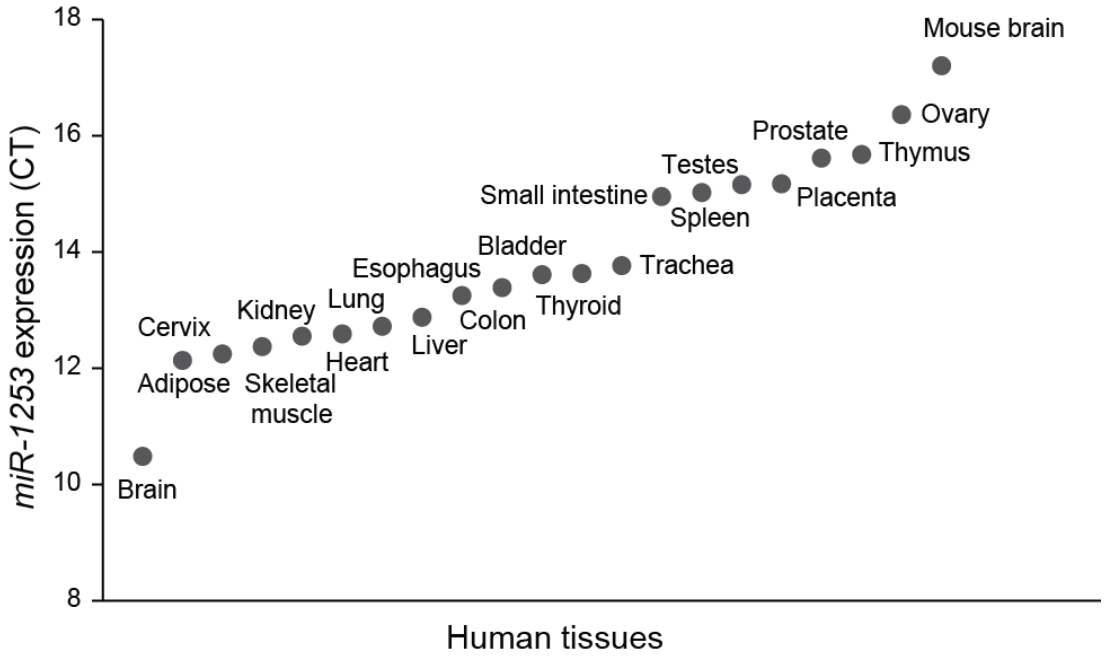


Figure 6.8 *miR-1253* expression in adult human tissues. *miR-1253* expression levels (y-axis) [(cross-threshold cycle (CT)] relative to *GAPDH* measured by qRT-PCR in 20 human tissues. *miR-1253* was also profiled for expressed in the mouse brain, where no expression (CT>35 prior to normalization to *GAPDH*) was found.

6.4.4 Crosstalk between lnc-ASD and MSN is miR-1253-dependent

The expression correlation between *lnc-ASD* and *MSN* (Figure 6.7), their sequence homology (Kerin et al., 2012), and the frequent human polymorphism within a *miR-1253* MRE within the lincRNA promoted my investigation of the possibility that *lnc-ASD* and *MSN* share sequence-dependent functional elements (i.e. MREs) and thus, are able to regulate each other's transcript levels through a competitive endogenous mechanism.

Using TargetScan (version 6.1)(Garcia et al., 2011), 229 MREs for 152 miRNA families were predicted to be shared between *lnc-ASD* and the 3' UTR of *MSN*. Out of these, the MREs for only two miRNA families, *miR-1253* and *miR-4778-3p*, were both present multiple times and were conserved between the two genes. Since *miR-4778-3p* showed no evidence of expression in human neuronal cell lines (Figure 6.9A), I decided to focus on *miR-1253* only for the rest of the analysis. Four and three *miR-1253* MREs were found within *lnc-ASD* and the 3' UTR of *MSN*, respectively (Figure 6.9B).

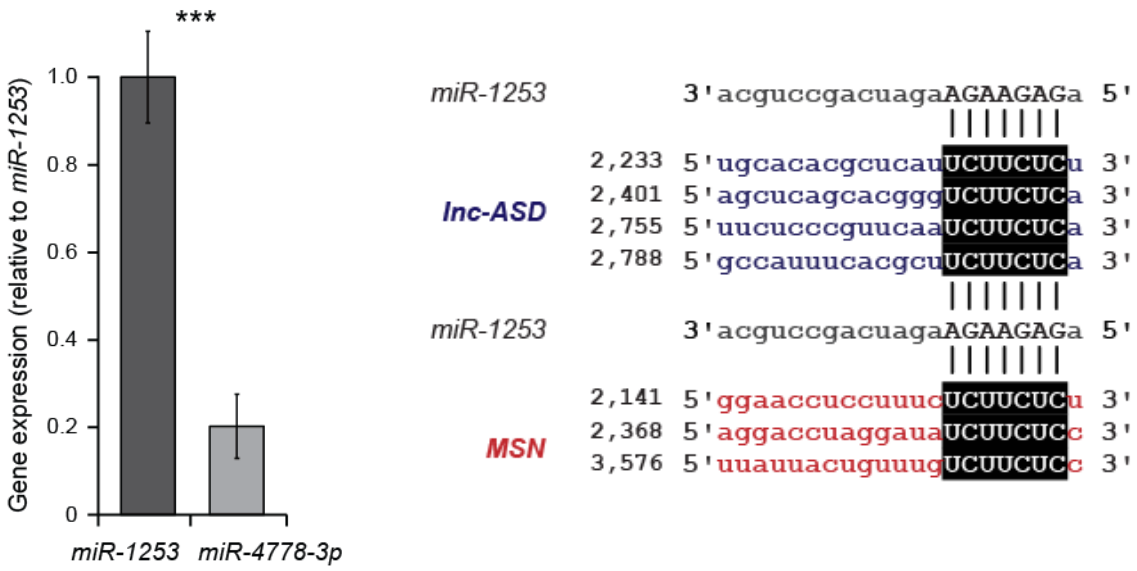


Figure 6.9 (A) *miR-4778-3p* (dark grey) is significantly (p<0.001) lower in expression levels in human SK-N-SH cells than *miR-1253* (80% lower, light grey). (B) Predicted *miR-1253* MREs within *lnc-ASD* and the 3' UTR of *MSN*.

To further dissect the potential *miR-1253*-mediated interaction between *lnc-ASD* and *MSN*, I measured the change in *miR-1253* levels after *lnc-ASD* knockdown by sequence-specific *si-lnc-ASDs* (Figure 6.10A) compared to non-specific negative siRNA control (*si-NC*). The knockdown of *lnc-ASD* (52% reduction) led to decreased levels of *MSN* (44% reduction) relative to control (*si-NC*) (Figure 6.10B). Furthermore, if *lnc-ASD* is targeted by *miR-1253* for post-transcriptional silencing, the down-regulation of the lincRNA is likely to free *miR-1253* from binding *lnc-ASD* and hence, increase the availability of the miRNA. As expected, reduced *lnc-ASD* abundance significantly (p<0.001) was associated with a 4.5-fold increase in *miR-1253* levels (Figure 6.10B).

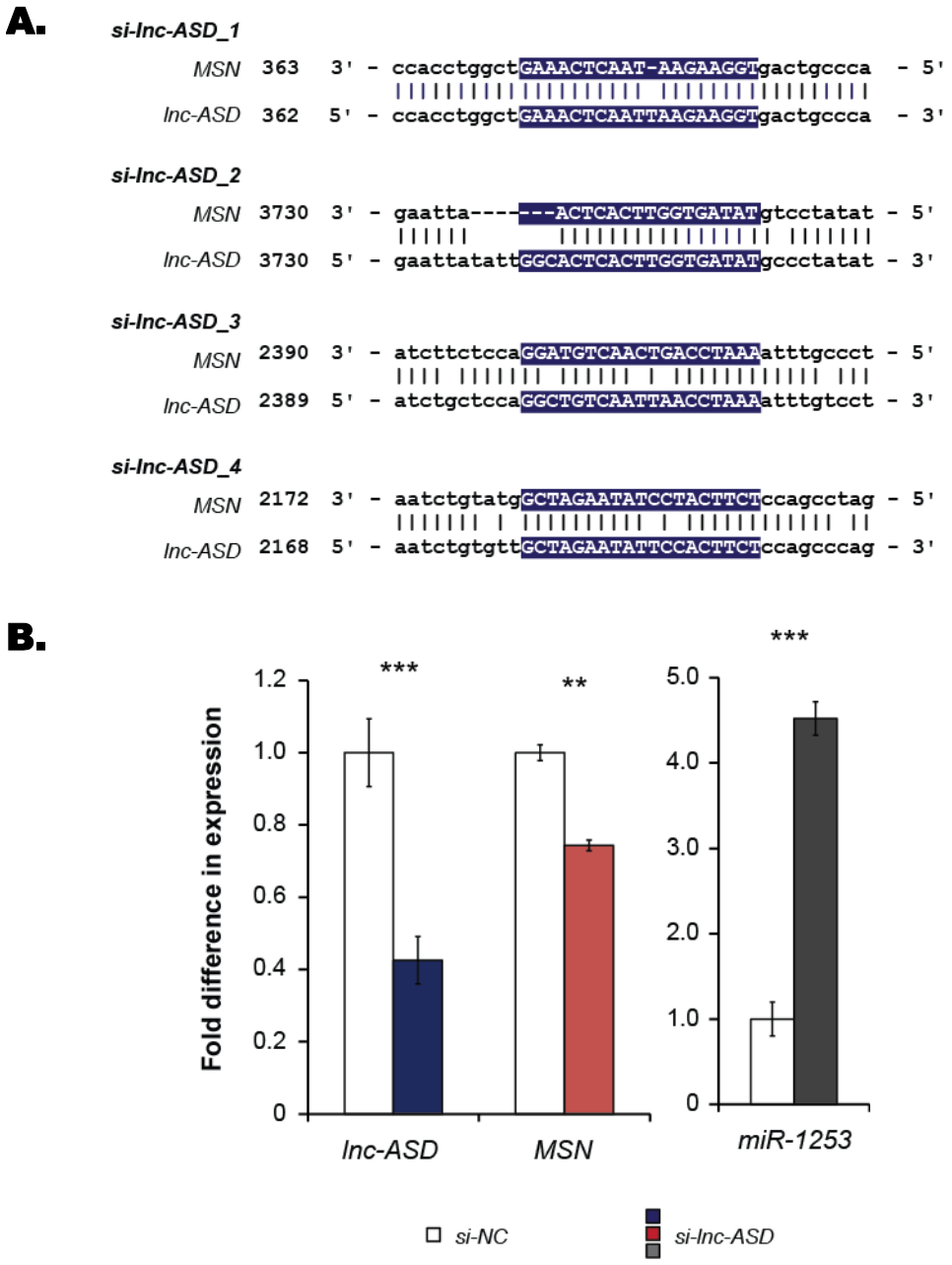


Figure 6.10 Knockdown of *lnc-ASD* down-regulate *MSN* levels and relieved *miR-1253* binding. (A) Alignment between small interference RNAs (siRNAs) used to knockdown endogenous levels of *lnc-ASD (si-lnc-ASD*) and their analogous sequences in *MSN* transcribed on the opposite strand as *lnc-ASD*, illustrating the target specificity of *si-lnc-ASDs*. (B) Knockdown of *lnc-ASD* by siRNAs, *si-lnc-ASD*, (52% reduction, $p<0.001$, blue) decreased the abundance of *MSN* transcript (44% reduction, $p<0.01$, red) and increased levels of *miR-1253* (4.5-fold, $p<0.001$, dark grey), compared to a non-specific siRNA control, *si-NC*.

Furthermore, if the post-transcriptional regulatory role of *lnc-ASD* is *miR-1253*-dependent, the human insertional polymorphism that is predicted to disrupt *miR-1253* binding at the 3' end of the lincRNA, if functional, should decrease the efficiency of crosstalk between this lincRNA and *MSN*. Overexpression of both *lnc-ASD* containing the common insertional mutation (i.e. with three *miR-1253* MREs instead of 4 *miR-1253* MREs), *lnc-ASD_3MREs*, and *lnc-ASD* that lack the polymorphism (i.e. with four *miR-1253* MREs), *lnc-ASD_4MREs*, increased the levels of *MSN* (1.7- and 2.7-fold, respectively) (Figure 6.11). However, the elevated levels of *MSN* was significantly (p<0.001) lower after *lnc-ASD_3MREs* than that observed for *lnc-ASD_4MREs* overexpression (Figure 6.11). In contrast, both *lnc-ASD_3MREs* and *lnc-ASD_4MREs* overexpression reduced *miR-1253* abundance (54% and 79% reduction, respectively), although the decrease in *miR-1253* levels was significantly (p<0.01) lower for *lnc-ASD_4MREs* overexpression (Figure 6.11). Therefore, the loss of one *miR-1253* MRE as a result of the insertional mutation frequent within the human population significantly reduced *lnc-ASD*'s ability to crosstalk and thus, post-transcriptionally regulate the levels of *MSN*.

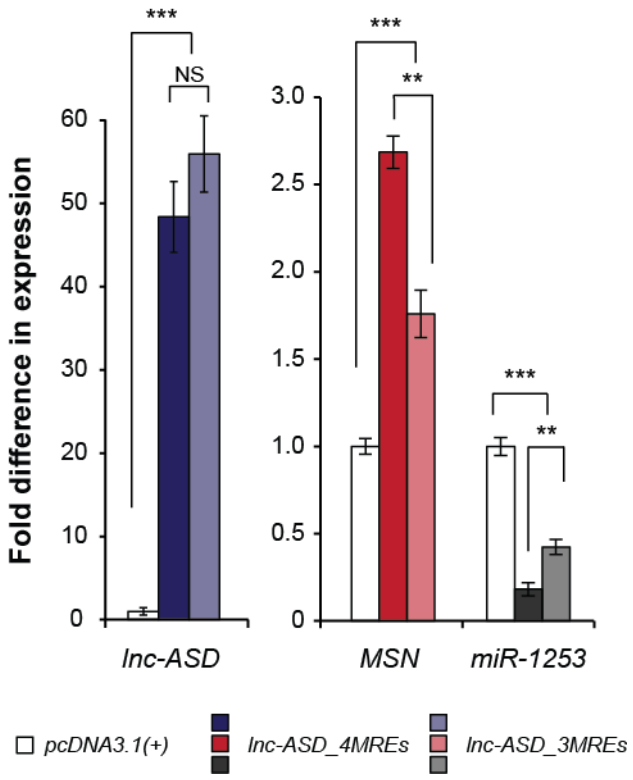


Figure 6.11 A common polymorphism disrupting the *miR-1253* MRE reduced *miR-1253*-mediated crosstalks between *lnc-ASD* on *MSN*. Overexpression of the wild-type *lnc-ASD* (*lnc-ASD_4MREs*, 48-fold, $p<0.001$, blue) significantly up-regulated *MSN* (2.7-fold, $p<0.001$, red) and down-regulated *miR-1253* (79% reduction, $p<0.001$, dark grey) levels relative to transfection of empty vector control (white). Overexpression of *lnc-ASD* that carried the frequent polymorphism that disrupts the last *miR-1253* MRE at its 3' end (*lnc-ASD_3MREs*, 56-fold, $p<0.001$, light blue) increased the levels of *MSN* (1.7-fold, $p<0.001$, light red) and decreased the levels of *miR-1253* (54% reduction, $p<0.001$, grey) relative to empty vector control (white). All expression levels were measured relative to that of *GAPDH*.

6.4.5 lnc-ASD post-transcriptionally modulates levels of multiple ASD-implicated genes

MiRNA-dependent crosstalk may occur between lincRNAs and several protein-coding transcripts that encode products that regulate functionally similar gene (**Chapter 5**). This led me to hypothesize *lnc-ASD* may crosstalk and modulate the levels of other ASD-implicated genes.

The 447 genes previously implicated in syndromic forms of ASD (Anney et al., 2010; Betancur, 2011; Neale et al., 2012; O'Roak et al., 2012b; Sanders et al., 2012; Kohler et al., 2014) have a significantly higher number of predicted *miR-1253* MREs per kilobase of sequence within their 3' UTRs (median density = 0.185 MREs/kb, 477 genes), as predicted by TargetScan version 6.1 (Garcia et al., 2011), than found within the 3' UTR of other neuronal-expressed protein-coding genes (median density = 0.137 MREs/kb, 3428 genes) (p<0.002, two-tailed Mann-Whitney test) (Figure 6.12).

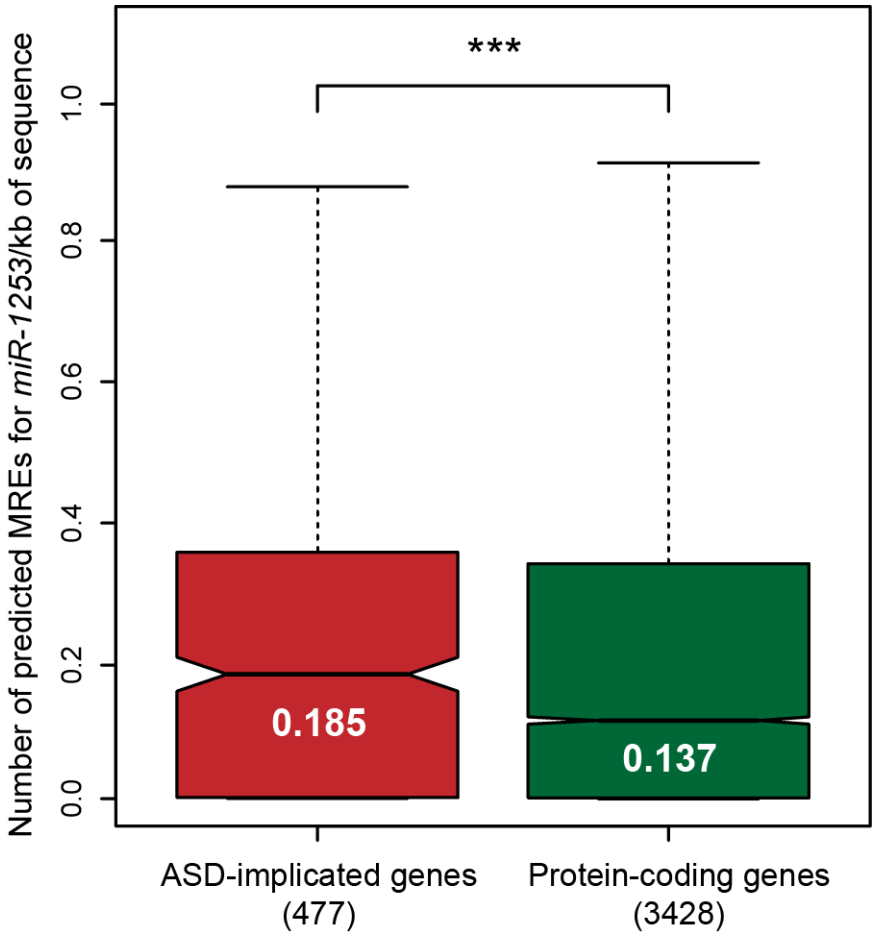


Figure 6.12 Crosstalking *lnc-ASD* and ASD-implicated protein-coding genes. *miR-1253* MREs are significantly (p<0.002, two-tailed Mann-Whitney test) more frequent in ASD-associated transcripts (median density=0.185 number of predicted *miR-1253* MREs/kb of gene length, 477 genes) than in the 3' UTR of all neuronal-expressed protein-coding transcripts (median density=0.137 number of predicted *miR-1253* MREs/kb of gene length, 3428 genes).

Next, I tested the prediction that *lnc-ASD* modulates the levels of other ASD-implicated genes. I selected six genes that when mutated give rise to syndromic forms of ASD and that have a high number of predicted MREs for miR1253, namely, *MECP2* (6 *miR-1253* MREs, Rett syndrome) (Amir et al., 1999), *NIPBL* (6 *miR-1253* MREs, Cornelia de Lange syndrome) (Krantz et al., 2004), *CDKL5*

(2 *miR-1253* MREs, Rett syndrome) (Weaving et al., 2004), *FGD1* (2 *miR-1253* MREs, Aarskog-Scott syndrome) (Orrico et al., 2005), *AHI1* (2 *miR-1253* MREs, Joubert syndrome) (Alvarez Retuerto et al., 2008) and *TSC2* (2 *miR-1253* MREs, tuberous sclerosis) (Smalley, 1998). As controls, I selected 2 genes whose implication in ASD are well established but lack predicted MREs for *miR-1253*, *NSD1* (Sotos syndrome) (Kurotaki et al., 2002) and *CREBBP* (Rubinstein-Taybi syndrome) (Barnby et al., 2005). In addition to the previously observed reduction in *MSN* (44% reduction) levels, knockdown of *lnc-ASD* (52% reduction) in SK-N-SH cells also significantly ($p<0.05$) decreased the levels of all 6 ASD-implicated genes tested that harbour *miR-1253* MREs (6% to 14% reduction) compared to non-specific negative siRNA control (*si-NC*) (Figure 6.13A). In contrast, no change in the levels of the two ASD-implicated genes used as negative controls was found for either *NSD1* or *CREBBP* (not significant, NS, $p>0.05$, Figure 6.13A). In addition, the overexpression of *lnc-ASD* (21-fold) significantly ($p<0.05$) up-regulated *MSN* (1.6-fold) levels, as well as the 6 ASD-implicated genes tested that harbour predicted *miR-1253* MREs (1.2-fold to 1.7-fold increase), but not the two genes that lack predicted *miR-1253* MREs (Figure 6.13B). Therefore, the ability of *lnc-ASD* to regulate *miR-1253* targets that are implicated in ASD but not ASD-implicated transcripts not regulated by *miR-1253* (i.e. *NSD1* and *CREBBP*) provides further evidence that *lnc-ASD* is involved in the pathophysiology of autism spectrum disorders possibly through a *miR-1253*-mediated mechanism, which underlies its genome-wide significant association with ASD (Kerin et al., 2012).

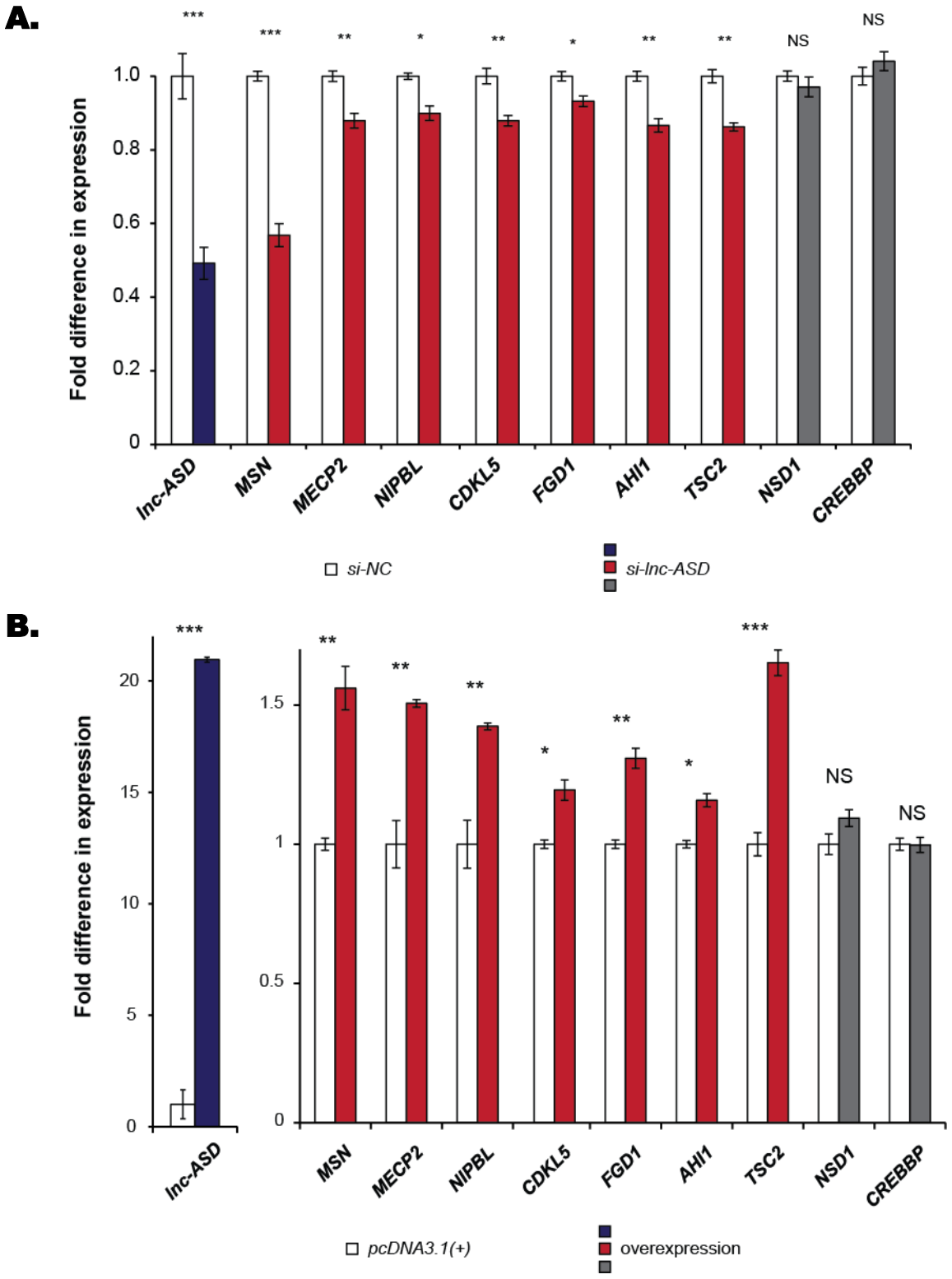


Figure 6.13 Overexpression and knockdown of *lnc-ASD* regulates transcript levels of *MSN* and 7 other genes implicated in ASD. (A) Knockdown of *lnc-ASD* (52% reduction, $p<0.001$, blue) significantly ($p<0.05$) increased the abundances of *MSN* (44% reduction, red) and 7 of the 8 additional ASD-implicated transcripts analyzed (6% to 14% reduction, red) compared to non-specific siRNA control (*si-NC*, white). No expression level difference were observed for the two ASD-implicated genes that do not harbour

miR-1253 MREs, *NSD1* and *CREBBP* (p>0.05, grey). (B) Overexpression of *lnc-ASD* (21-fold, p<0.001, blue) significantly (p<0.05) increased the abundances of *MSN* (1.6-fold, red) and 7 of the 8 additional ASD-implicated genes analyzed (1.2- to 1.7-fold, red) compared to empty vector control (*pcDNA3.1(+)*, white). No expression level differences were observed for the two ASD-implicated genes that do not harbour *miR-1253* MREs, *NSD1* and *CREBBP* (p>0.05, grey). All expression levels were measured relative to that of *GAPDH*.

Subsequently, to confirm whether crosstalks between *lnc-ASD* and ASD-implicated genes are *miR-1253*-dependent, I tested the effect of overexpressing the lincRNA with mutated *miR-1253* MREs (a total of four *miR-1253* MREs are predicted within *lnc-ASD*): (1) *lnc-ASD* containing mutations disrupting the last two *miR-1253* MREs at the 3' end of the lincRNA – *lnc-ASD_2MREs*; and (2) *lnc-ASD* containing mutations disabling all four *miR-1253* MREs – *lnc-ASD_0MRE*.

If *lnc-ASD*'s effect on ASD-implicated genes is mediated through *miR-1253*-dependent competition, then disrupting *miR-1253* MREs within the lincRNA sequence should partially, if not completely, disable *lnc-ASD*'s ability to crosstalk with *miR-1253* targeted genes. Mutations within *miR-1253* MREs were generated for the miRNA's 6 nt seed region, where no sequence homology was found between the mutated *miR-1253 MRE* and the seed region of any other human miRNAs. As shown previously, overexpression of *lnc-ASD* without the insertional polymorphism (i.e. four *miR-1253* MREs), *lnc-ASD_4MREs*, in SK-N-SH cells, significantly (p<0.001) increased the expression levels of *MSN* (2.3-fold) and all 6 additional ASD-implicated genes that harbour *miR-1253* MREs (1.4-fold to 2.0-fold) compared to transfection

control (Figure 6.14). The overexpression of *lnc-ASD* containing two mutated *miR-1253* MREs, *lnc-ASD_2MREs*, increased the levels of *MSN* (1.5-fold) and only 3 of the 6 additional ASD-implicated genes that harbours *miR-1253* MREs (1.3-fold to 1.5-fold); these increases were significantly lower than that observed previously for *lnc-ASD_4MREs* overexpression. Furthermore, no significant transcript differences in any of the ASD-implicated genes were observed upon overexpression of *lnc-ASD_0MRE* (Figure 6.14). Also, in all cases, no expression differences were observed for the two ASD-implicated genes that do not harbour *miR-1253* MREs (Figure 6.14).

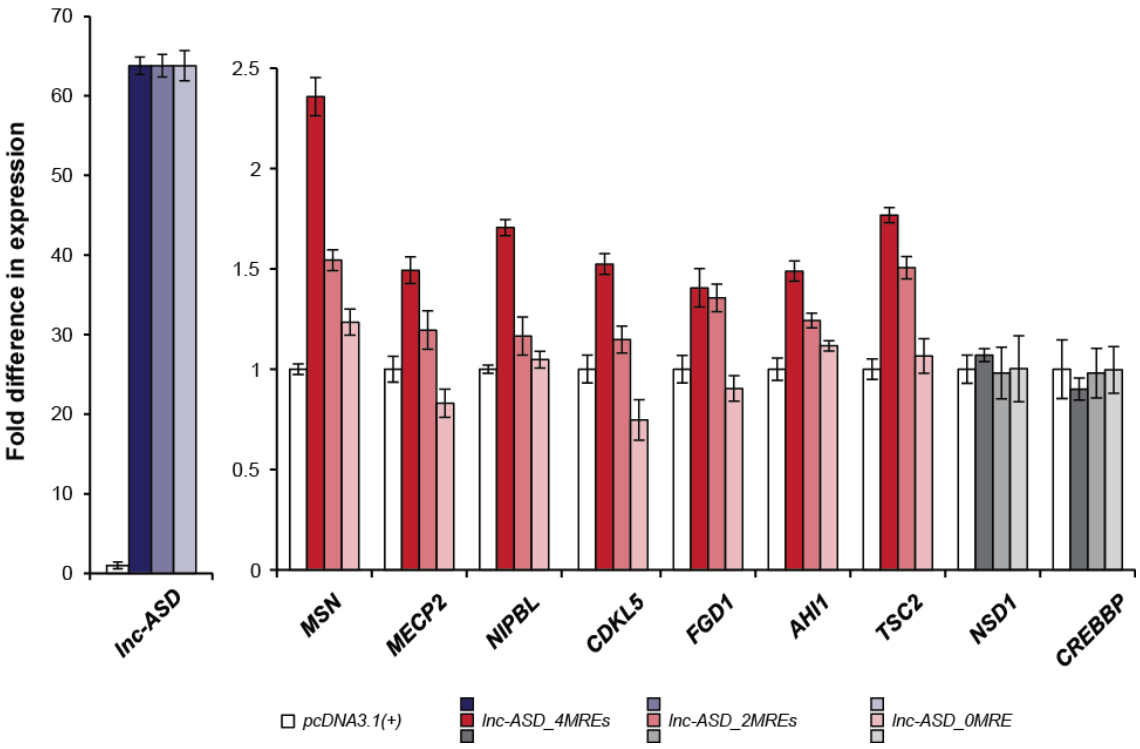


Figure 6.14 Regulation of *lnc-ASD* on *MSN* and ASD-implicated genes is dependent on *miR-1253* MREs. Overexpression of the wild-type *lnc-ASD* (*lnc-ASD_4MREs*, 64-fold, $p<0.001$, blue) significantly up-regulated *MSN* (2.3-fold, $p<0.001$, red) and all 8 ASD-implicated genes harbouring *miR-1253* MREs (1.4- to 2.0-fold, $p<0.001$, red), compared to transfection of empty vector control (white). Overexpression of *lnc-ASD* with two mutated *miR-1253* MREs at its 3' end (*lnc-ASD_2MREs*, 64-fold, $p<0.001$, light blue) increased the levels of *MSN* (1.5-fold, $p<0.001$, light red) and 3 additional ASD-implicated genes containing *miR-1253* MREs (1.3- to 1.5-fold, $p<0.05$, light red). Overexpression of *lnc-ASD* with all four *miR-1253* MREs mutated (*lnc-ASD_0MRE*, 64-fold, $p<0.001$,

lightest blue) had no effect on the levels of *MSN* (p>0.05, pink) and ASD-implicated genes (p>0.05, pink). No expression level difference were observed for the two ASD-implicated genes that do not harbour *miR-1253* MREs, *NSD1* and *CREBBP* (p>0.05, grey). All expression levels are measured relative to that of *GAPDH* and normalized between the independent transfections of *lnc-ASD_4MREs*, *lnc-ASD_2MREs*, and *lnc-ASD_0MRE*.

Finally, after establishing that the post-transcriptional regulatory role of *lnc-ASD* is *miR-1253*-dependent, I tested the effect of the human common polymorphism, which disabled *miR-1253* binding at its fourth MRE at the 3' end of the lincRNA, on ASD-implicated protein-coding genes. Consistent with the expectation that the polymorphism reduces the efficiency of crosstalks between *lnc-ASD* and ASD-implicated genes, overexpression of *lnc-ASD* that contains the common insertional mutation, *lnc-ASD_3MREs*, led to increased levels of all tested ASD-implicated genes (1.3- to 1.7-fold) at significantly (p<0.001) lowered rate than that observed for the overexpression of *lnc-ASD* with no polymorphism, *lnc-ASD_4MREs* (1.5- to 2.7-fold) (Figure 6.15). Therefore, the loss of one *miR-1253* MRE as a result of an insertional mutation had significantly reduced *lnc-ASD*'s ability to crosstalk and thus, post-transcriptionally regulate the expression levels of ASD-implicated genes.

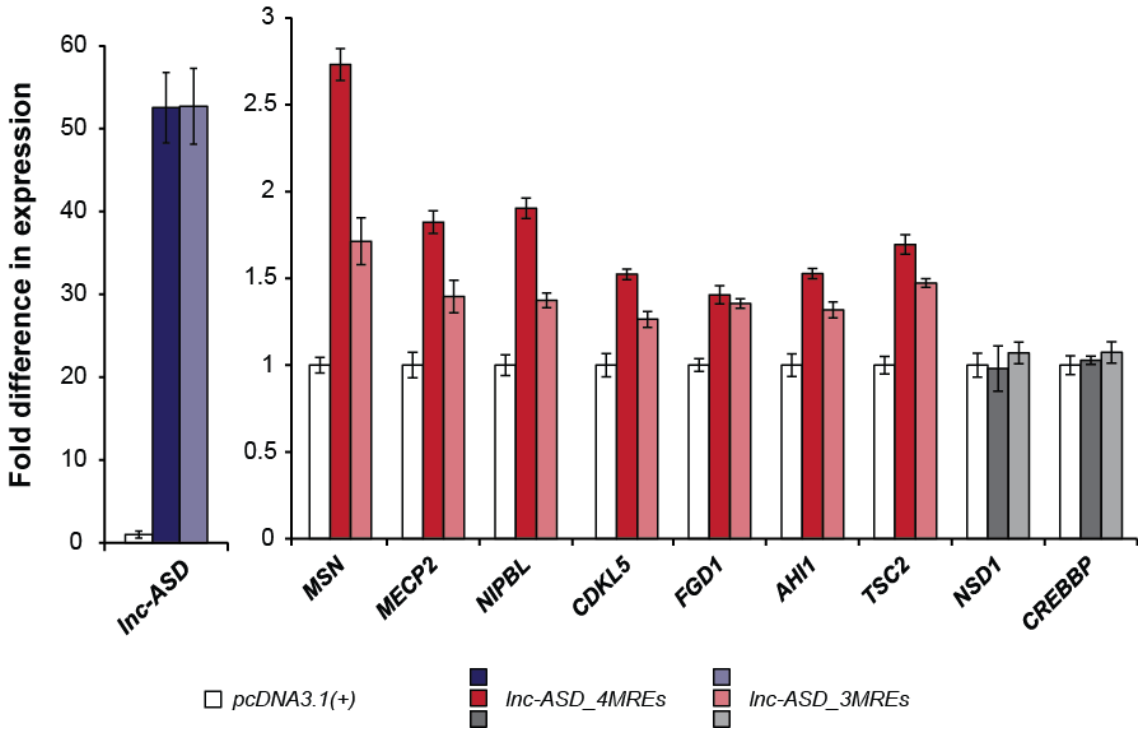


Figure 6.15 Effect of crosstalks through disrupted *miR-1253* MRE within *lnc-ASD* on *MSN* and ASD-implicated genes. Overexpression of the *lnc-ASD* without the common insertional polymorphism (*lnc-ASD_4MREs*, 48-fold, $p<0.001$, blue) significantly up-regulated *MSN* (2.7-fold, $p<0.001$, red) and all 8 ASD-implicated genes harbouring *miR-1253* MREs (1.5- to 1.9-fold, $p<0.001$, red), compared to transfection of empty vector control (white). Overexpression of *lnc-ASD* that carry the common polymorphism that disrupts the last *miR-1253* MRE at its 3' end (*lnc-ASD_3MREs*, 56-fold, $p<0.001$, light blue) increased the levels of *MSN* (1.7-fold, $p<0.001$, light red) and 6 additional ASD-implicated genes containing *miR-1253* MREs (1.3- to 1.5-fold, $p<0.05$, light red). No expression level differences were observed for the two ASD-implicated genes that do not harbour *miR-1253* MREs, *NSD1* and *CREBBP* ($p>0.05$, grey). All expression levels are measured relative to that of *GAPDH* and normalized between the independent transfections of *lnc-ASD_4MREs* and *lnc-ASD_3MREs*.

6.4.7 lnc-SCA7 and miR-1253 are rapidly turned over

Previously, I have shown evidence supporting *lnc-ASD*'s regulation on ASD-implicated protein-coding genes that harbour *miR-1253* MREs, both by artificially increasing the level of *lnc-ASD* (with a complete or partial set of functional *miR-1253* MREs), by reducing the abundance of *lnc-ASD* via RNA interference pathway (Figure 6.10, 6.13A) and by introducing exogenous copies of the lincRNA (Figure 6.11, 6.13B, 6.14), suggest *lnc-ASD*'s functional role as a competitive endogenous RNA (ceRNA) via *miR-1253*-mediated crosstalks. However, the expression of *lnc-ASD* is significantly lower than *MSN* or any of the ASD-implicated genes (16-fold lower than *MSN*, $p<0.001$, Figure 6.16).

To serve as a molecular decoy for *miR-1253* and post-transcriptionally regulate a number of relatively highly-expressed protein-coding genes implicated in ASD, the molecular stoichiometry between *lnc-ASD, miR-1253*, and its target mRNAs is an important issue to address. In this respect, it would be difficult to propose a realistic scenario where the lowly expressed *lnc-ASD* can simultaneously crosstalk with multiple relatively highly expressed protein-coding genes. However, previous findings that the overexpression of *lnc-ASD* reduced *miR-1253* abundance whereas the knockdown of *lnc-ASD* elevated *miR-1253* levels (Figure 6.10B, 6.11) suggested *lnc-ASD* was able to regulate *miR-1253* levels. Determining the turnover rate of *lnc-ASD* may be important to determine the inconsistency in stoichiometry requirement and miRNA-mediated post-transcriptional regulation of *lnc-ASD* where a rapidly turned over transcript may continually decoy miRNAs away from other target transcripts of the miRNAs.

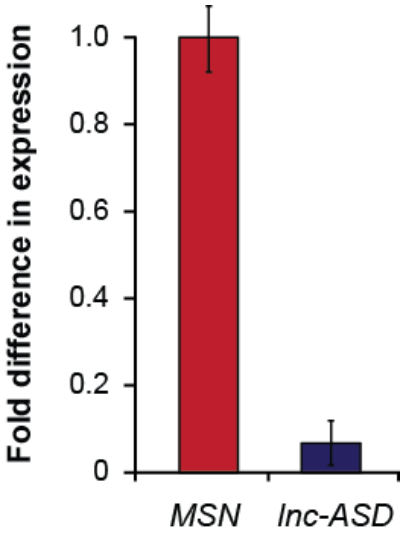


Figure 6.16 Expression level of *lnc-ASD* is significantly lower than that of *MSN*. The abundance of *lnc-ASD* in SK-N-SH cells is 16-fold lower (relative to *GAPDH*)(p<0.001) than its ancestral protein-coding gene, *MSN*.

Therefore, attempting to explain how despite its relatively low levels can *lnc-ASD* regulate its target mRNAs, I assessed the transcriptional turnover rate of the lincRNA by blocking RNA transcriptional initiation using *Actinomycin D* (*ActD*) (dissolved in Dimethyl sulfoxide – DMSO). *ActD* is an inhibitor of the cellular transcription machinery, which binds DNA at the transcription initiation complex and prevents elongation of the RNA chain by RNA polymerase (Sobell, 1985). SK-N-SH cells treated with only DMSO were used as transfection controls. *MYC*, a known protein-coding mRNA that is rapidly degraded (Dani et al., 1984) and *ATP5E*, a relatively stable protein-coding gene (Clark et al., 2012) were used as controls for *ActD* efficacy (Figure 6.17). Consistent with previous findings, over a course of 24 hours, *MYC* was unstable with a half-life of approximately 1 hour (similar to what has been reported in literature (Clark et al., 2012)). On the contrary, *ATP5E* is exceedingly stable, with up to 82% of the mRNA remaining 24 hours after transcription inhibition (Figure 6.17). As expected, *lnc-ASD* levels rapidly decreased with a half-life of approximately 1 hour (Figure 6.17), suggesting that *lnc-ASD* may be under constant regulatory suppression by miRNAs, which

subjects the lincRNA transcript to rapid degradation. In contrast, the ancestral protein-coding *MSN* decayed at a much slower rate, with a half-life of approximately 16 hours, indicating the difference in stoichiometry between *lnc-ASD* and *MSN* transcripts is likely owing to their distinct rates of turnover.

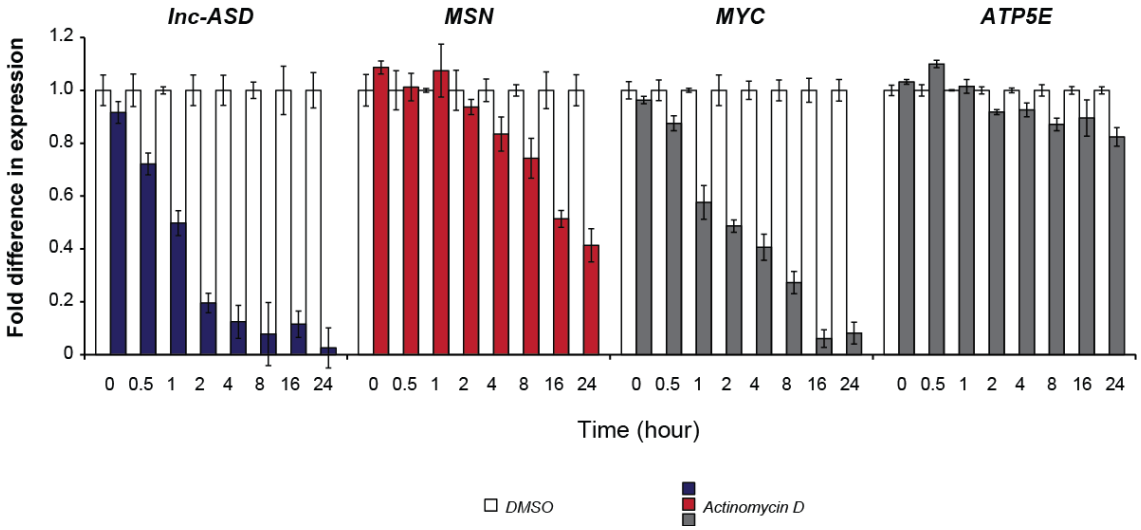
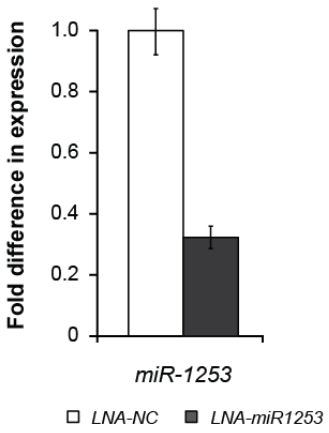


Figure 6.17 *lnc-ASD* is rapidly turned over. The stability of *lnc-ASD* (blue) and *MSN* (red) were assessed by blocking transcriptional initiation in SK-N-SH cells using *Actinomycin D* (in DMSO). RNA samples were extracted at 0, 0.5, 1, 2, 4, 8, 16 and 24 hours post-transcriptional inhibition. Cells treated with DMSO only were used as control (white). *Lnc-ASD* has a half-life of approximately 1 hour, while *MSN* has a half-life of approximately 16 hours. Expression levels of genes were measured relative to that of *GAPDH*. *MYC*, a known rapidly degraded mRNA (Dani et al., 1984) and *ATP5E*, an mRNA with a long half-life (Clark et al., 2012) were used as controls (grey).

In addition to assessing the turnover rate of the lincRNA and its ancestral protein-coding gene, I also tested the effect of reducing the endogenous levels of *miR-1253* on *lnc-ASD* and *MSN* transcript stability. If *miR-1253* post-transcriptionally regulates *lnc-ASD* and *MSN* by targeting these transcripts for degradation, one would expect a decrease in *miR-1253* levels to relieve its target genes from post-transcriptional silencing. Consistent with this expectation, after knocking down endogenous *miR-1253* levels using locked nucleic acid (LNA) designed specifically to target the miRNA (*LNA-miR-1253*, 68% reduction, $p<0.001$, Figure 6.18A), I observed a significantly decreased rate of transcriptional turnover for both *lnc-ASD* (half-life of approximately 4 hours compared to ~1 hour prior to *LNA-miR-1253* treatment, 75% reduction in turnover rate, $p<0.001$) and *MSN* (half-life of approximately 24 hours compared to ~16 hour prior to *LNA-miR-1253* treatment, 33% reduction in turnover rate, $p<0.001$), relative to non-specific LNA control (*LNA-NC*) (Figure 6.18 B and C). No significant differences in the half-lives of *MYC* and *ATP5E* were observed, as expected for genes that do not harbour *miR-1253* MREs (Figure 6.18 D and E).

A.



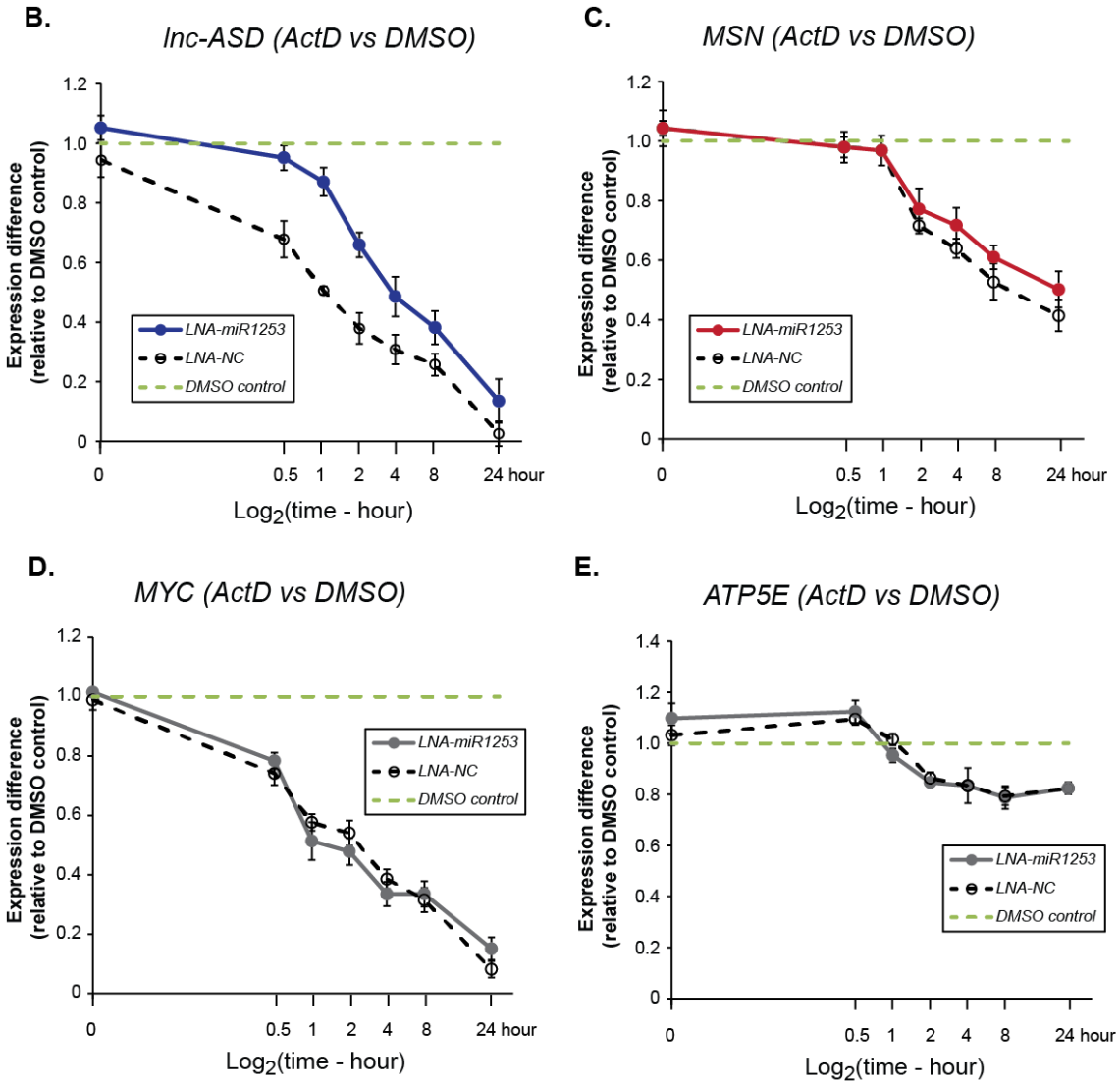


Figure 6.18 Stability of *lnc-ASD* is increased upon reduction of *miR-1253* levels. (A) Endogenous levels of *miR-1253* were reduced using locked nucleic acid (LNA) designed specifically to target *miR-1253* (*LNA-miR-1253*, 68% reduction, $p<0.001$) in SK-N-SH cells. Throughout the 24 hour time course, the levels of *miR-1253* was consistently higher in cells treated with *LNA-miR-1253* than the control cells that were treated with a non-specific LNA molecule (*LNA-NC*). (B to E) The stability of (B) *lnc-ASD* (blue), (C) *MSN* (red), (D) *MYC* (grey) and (E) *ATP5E* (grey) were assessed by blocking transcriptional initiation in SK-N-SH cells using *Actinomycin D* (dissolved in DMSO) with and without *LNA-miR-1253* treatment. RNA samples were extracted at 0, 0.5, 1, 2, 4, 8, 16 and 24 hours post-transcription inhibition. Cells treated with DMSO only were used as control (white). (B) The half-life of *lnc-ASD* increased from ~1 hour to ~4 hour by knocking down *miR-1253* with *LNA-miR-1253* (75% reduction in rate of transcription turnover, $p<0.001$). (C) The half-life of *MSN* increased from ~16

hour to ~24 hour by knocking down *miR-1253* with *LNA-miR-1253* (33% reduction in rate of transcription turnover, $p<0.001$). (D and E) No difference in the rate of transcription turnover was observed for that of (D) *MYC* and (E) *ATP5E*, two mRNAs with no predicted *miR-1253* MREs. Expression levels of genes were measured relative to that of *GAPDH*. X-axis is represented on the log 2 scale of time in minutes (black).

The rapid rate of transcriptional turnover of *lnc-ASD* might explain why despite being expressed at relatively low levels in comparison to its target mRNAs (Figure 6.15), *lnc-ASD* may nonetheless serve as a molecular miRNA decoy for *MSN* and other ASD-implicated genes. Importantly, reducing the levels of endogenous *miR-1253* increased the stability of *lnc-ASD* across a time course of 24 hours, hence reinforcing the hypothesis that the crosstalks between *lnc-ASD* and ASD-implicated genes are assembled via a mechanism mediated through *miR-1253* MREs.

6.5 DISCUSSION

In 2012, Kerin et al. reported an intergenic long noncoding RNA (lincRNA), *lnc-ASD* (or *MSNP1AS*), that is derived from a retropseudogene (*MSNP1*) of *MSN*, whose levels were correlated with a genome-wide significant nucleotide polymorphism enriched in ASD individuals (Wang et al., 2009a; Kerin et al., 2012). Increase in transcript levels of both *lnc-ASD* (12.7-fold) and *MSN* (2.3-fold) was found in post-mortem brain samples of ASD individuals compared to controls (Kerin et al., 2012). Because of their high sequence identity, Kerin and colleagues proposed that *lnc-ASD* regulates levels of *MSN* by directly binding to its mRNA transcripts, possibly subjecting the double-stranded RNA product to nonsense-mediated decay (NMD) (Kerin et al., 2012). This hypothesis was supported by decreased levels of MSN protein following *in vitro* overexpression of *lnc-ASD* (Kerin et al., 2012). However, this hypothesized mechanism of regulation cannot support the positive expression correlation between *MSN* and *lnc-ASD* observed in post-mortem brain samples of ASD individuals reported in the same study (Kerin et al., 2012).

In this chapter, I investigated an alternative mechanism by which *lnc-ASD* may contribute to the regulation of *MSN*, as well as extended the analysis on how this lincRNA regulate mRNAs encoding other ASD-implicated genes. Findings in this chapter also provide evidence that the lincRNA is specific to the primate lineage with an insertional polymorphism commonly present in the human population that disrupts the 3' miRNA response element (MRE) within the transcript specific for a primate-specific miRNA, *miR-1253*. Importantly, *lnc-ASD* appears to be evolving under faster than expected rate and the insertional

polymorphism that disrupts a *miR-1253* MRE within *lnc-ASD* also exhibits some signatures of accelerated evolution.

Here, I present supporting evidence that the association between *lnc-ASD* and ASD genotypes might be owing to its ability to post-transcriptionally regulate a network of ASD-implicated genes through crosstalks via *miR-1253* competition. Computational analysis of genes implicated in ASD demonstrated that genes that share at least one MRE for *miR-1253* with *lnc-ASD* were significantly more coexpressed than expected by chance in human prefrontal and temporal cortex (see Appendix A6.1). Interestingly, amongst the miRNAs whose MREs are enriched within *lnc-ASD*, *miR-1253* has MREs also harboured within 71% of candidate genes implicated in syndromic forms of ASD from past literature (Anney et al., 2010; Betancur, 2011; Neale et al., 2012; O'Roak et al., 2012b; Sanders et al., 2012; Kohler et al., 2014). In a human neuroblastoma cell line, by overexpression *lnc-ASD* sequences that harbour various numbers of *miR-1253* MREs, including the lincRNA sequence that contains the insertional polymorphism that disrupts a *miR-1253* binding site, I showed this lincRNA post-transcriptionally regulate *MSN*. Importantly, it also contributes to the regulation of the levels of multiple ASD-implicated genes through a likely *miR-1253*-mediated regulatory mechanism, and possibly mediating the assembly of a regulatory network of ceRNAs. In addition, *lnc-ASD* transcripts are particularly unstable and rapidly turned over, suggesting the lincRNA may be under frequent miRNA suppression and supporting its relative low abundance in comparison to its target mRNAs.

The evidence supporting that both *lnc-ASD* and *miR-1253* appear to be present only in primate species potentially suggest both noncoding RNAs having co-evolved to serve functions specifically in the primate lineage. Importantly, the disrupted *miR-1253* MRE within *lnc-ASD*, which decreased the amount of crosstalk between lincRNA and ASD-implicated genes, suggest this common polymorphism may be driven by an accelerated selection to reduce *lnc-ASD*'s ability to manipulate levels of *miR-1253*, which prevents crosstalks between ASD-implicated genes (Figure 6.19).

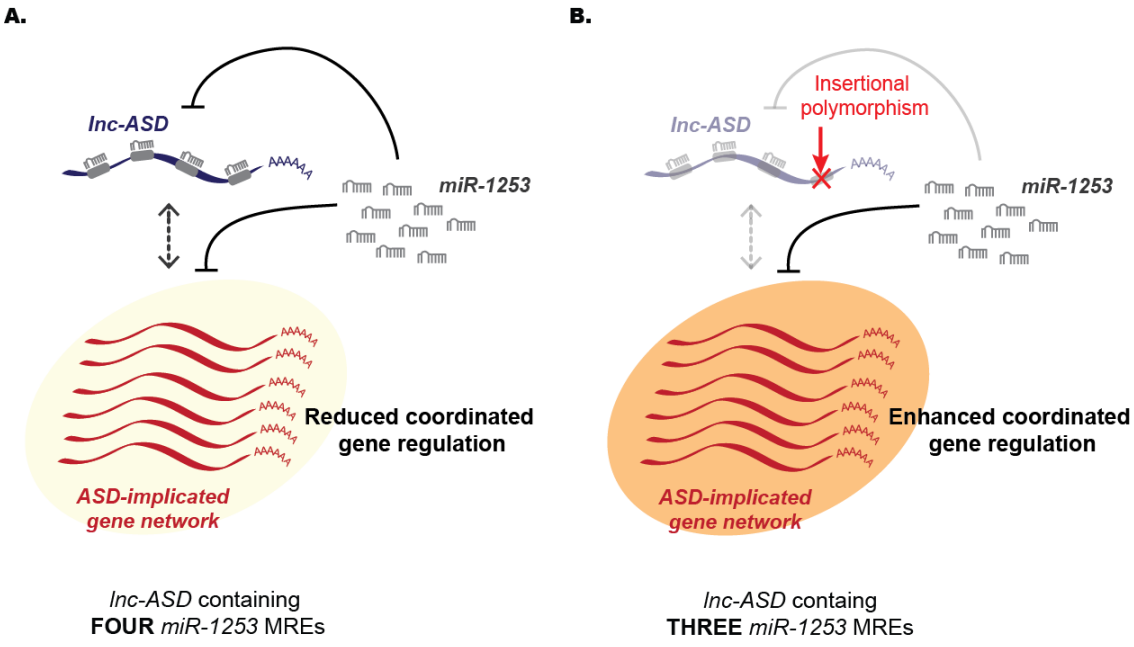


Figure 6.19 Working hypothesis of *lnc-ASD* and its potential contribution to increase risk of ASD. (A) *lnc-ASD* harbouring four *miR-1253* MREs display strengthened miRNA-mediated crosstalks with genes previously implicated in ASD compared to (B) *lnc-ASD* that contains a common insertional polymorphism in human, which disrupts the 3' *miR-1253* MRE within the lincRNA.

Recent genome-wide studies have begun to enhance our understanding of the multi-layered gene regulatory networks that underlie common neurodegenerative diseases, including ASD (Ma et al., 2009; Wang et al., 2009a; Weiss et al., 2009; Salyakina et al., 2010; Hussman et al., 2011). In particular, many disease-associated genetic signals were found to be located in genomic regions outside of annotated protein-coding genes (Jia et al., 2009; Musunuru et al., 2010), suggesting that noncoding mutations, in addition to protein-coding gene variants, may contribute to disease pathogenesis.

Current treatment of neurodegenerative diseases mainly target single cellular pathways downstream of disease initiation (Koller and Tse, 2004; Talbot, 2007; Schapira, 2009), which may happen at a much delayed stage past the effective therapeutic window and thus, renders treatment relatively ineffective. With an increased understanding of the multiple genetic factors, including those that lie in noncoding regions, that govern complex RNA regulatory networks, and the molecular mechanisms that underlie crosstalking interactions, novel therapeutics can potentially be more effectively designed to target the right genes at the appropriate time. Here, I demonstrated the primate-specific crosstalks between an intergenic long noncoding RNA, *lnc-ASD*, and ASD-implicated genes via a miRNA, *miR-1253*, mediated mechanism. The apparent accelerated evolution observed for *lnc-ASD* and the frequent insertional polymorphism that disrupts a *miR-1253* MRE within the lincRNA implies evolutionary selection may be driving to prevent *lnc-ASD* from interacting with *miR-1253*, which reduces crosstalks between ASD-implicated genes.

CHAPTER 7

Perspectives

The recent discovery that endogenously expressed transcripts, termed competitive endogenous RNAs (ceRNAs), can regulate each other's levels by competing for the binding to shared microRNAs (miRNAs) has revealed a new layer of post-transcriptional regulation. My thesis investigates the properties and prevalence of ceRNAs, with a specific focus on intergenic long noncoding RNAs (lincRNAs) that serve as ceRNAs (lnceRNAs), and their biological significance in development and diseases. My thesis provides important insights into (1) the significance of the ceRNA mechanism (**Chapter 3**), supported by the finding that rodent-specific unitary pseudogenes show preserved ceRNA function despite their loss of protein-coding ability; (2) the prevalence of lnceRNAs in mouse embryonic stem cells (mESCs) and the characterization of their properties (**Chapter 4**); (3) how a lnceRNA, *lnc-SCA7*, can contribute to the tissue-specific neurodegeneration observed in Spinocerebellar ataxia 7 (SCA7) patients (**Chapter 5**); and (4) how a primate-specific lnceRNA, *lnc-ASD*, modulates the levels of several autism spectrum disorders implicated genes (**Chapter 6**).

Despite previous encouraging discoveries of individual lnceRNAs (Tay et al., 2014), as well as findings described in this thesis, on the characterization and biological significance of lnceRNAs, the precise molecular rules that govern this miRNA-mediated crosstalking mechanism between coding and noncoding

transcripts and their overall relevance to normal physiology and disease remain relatively poorly characterized. In addition, most lincRNAs, possibly including those that act as lnceRNAs, are generally of low abundance relative to protein-coding genes (Cabili et al., 2011; Derrien et al., 2012) (**Chapter 4**). Given that lnceRNAs often post-transcriptionally regulate the levels of one or more protein-coding genes, it was proposed the disproportional stoichiometry of lnceRNA abundance relative to mRNAs might limit their ability to effectively modulate the abundance of their target mRNA(s), in a miRNA-dependent manner (Ebert and Sharp, 2010; Ala et al., 2013; Figliuzzi et al., 2013; Denzler et al., 2014).

Furthermore, the biological relevance of ceRNAs has recently been challenged (Broderick and Zamore, 2014). The authors showed that in adult hepatocytes, the levels of one transcript, *AldoA*, required to significantly alter the levels of a highly abundant miRNA, *miR-122*, and induce significant gene expression changes in its target genes are higher than the changes observed *in vivo*, even under extreme physiological or disease conditions (Denzler et al., 2014). This led to the proposal that ceRNAs interactions, in general, are not physiological relevant (Broderick and Zamore, 2014). This result is likely not generalizable as the study was conducted for only one potential ceRNA and for one specific and vastly abundant miRNA. Previous findings support the biological significance of lnceRNAs. For example, decrease in endogenous levels of lnceRNAs has been shown to affect levels of their mRNA targets (Cesana et al., 2011; Wang et al., 2013). Mutations within *PTENP1*, likely disrupting its miRNA-mediated crosstalk with the oncogenic *PTEN* tumour suppressor, has also been found in human melanoma patients (Karreth et al., 2011). The discrepancy between the finding of Denzler et al. (2014) and previous reports of biologically relevant lnceRNAs

indicates that further investigations are required to gain a complete understanding of the significance of lnceRNAs.

Contrary to the expectation that lnceRNAs should harbour a high number of crosstalking MREs and be found at similar levels relative to their crosstalking targets, some lincRNAs with miRNA-dependent roles share neither an unusually high number of predicted MREs with their mRNA targets nor are particularly abundant (Cesana et al., 2011; Wang et al., 2013; Tay et al., 2014). How can relatively lowly-expressed transcripts with only a few shared MREs modulate the expression levels of multiple abundantly expressed mRNAs and contribute to cell fate decisions? One answer could be that lowly-expressed lnceRNAs, or those with less effective MREs, can only efficiently influence cellular phenotypes if they crosstalk with (i) highly dosage sensitive transcripts and/or (ii) genes involved in controlling switch-like decisions (i.e. cell fate decisions) (de Giorgio et al., 2013), which may be frequently post-transcriptionally regulated by lnceRNAs. For instance, lnceRNAs might function to efficiently regulate the transition between pluripotent and differentiated cell states in undifferentiated cells, during which small changes in miRNA levels induced by these lnceRNAs may have important impact on cellular homeostasis (**Chapter 4**).

Alternatively, some lnceRNAs can amplify changes in transcript levels by acting as part of autoregulatory feedback loops (Xu et al., 2009; Wang et al., 2013). For example, in **Chapter 5,** I demonstrated that *lnc-SCA7* competes for *miR-124* with transcripts of *Atxn7*, the gene implicated in Spinocerebellar ataxia type 7 (SCA7), in a negative feedback loop. The disruption of this regulatory

circuitry, such as that caused by the mutated *Atxn7*, likely contributes to SCA7 etiology. In addition, other factors, such as miRNA-target affinity or miRNA turnover, might also contribute to efficient crosstalks mediated by lowly abundant lnceRNAs. For example, I have showed that *lnc-ASD*'s rapid turnover rate is dependent on the presence of its targeting miRNA, *miR-1253*. This suggests the unstable rate of decay of *lnc-ASD* transcripts is possibly owing to their regulation by *miR-1253* and it is the binding of *miR-1253* to *lnc-ASD* transcripts that enables the lincRNA to post-transcriptionally regulate multiple ASD-implicated protein-coding gene transcripts (**Chapter 6**).

A full understanding of the biological relevance of this mechanism will require deciphering the rules that determine the interactions between transcripts and their relative contributions to transcript level regulation. Mathematical modelling may be useful to address the stoichiometry requirement of lnceRNA, miRNA, and other crosstalking targets, by simulating biologically meaningful environments and conditions in which the crosstalk occurs, while accounting for factors that govern expression and degradation of the different classes of RNA transcripts. Further genetic and genomic evidence on many lnceRNAs and miRNAs in diverse cell types will be required to fully understand the genome-wide prevalence and physiological relevance of these lnceRNAs. Nevertheless, this novel mechanism of post-transcriptional regulation adds a new layer of complexity to genetic networks, and the putative associations between post-transcriptional regulation and diseases demonstrates the physiological importance of lnceRNAs and their contributions to biological phenomenon.

REFERENCES

Abou-Sleymane, G., Chalmel, F., Helmlinger, D., Lardenois, A., Thibault, C., Weber, C., Merienne, K., Mandel, J.L., Poch, O., Devys, D.*, et al.* (2006). Polyglutamine expansion causes neurodegeneration by altering the neuronal differentiation program. Human molecular genetics *15*, 691-703.

Abrahams, B.S., and Geschwind, D.H. (2008). Advances in autism genetics: on the threshold of a new neurobiology. Nat Rev Genet *9*, 341-355.

Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G., Scherer, S.E., Li, P.W., Hoskins, R.A., Galle, R.F.*, et al.* (2000). The genome sequence of Drosophila melanogaster. Science *287*, 2185-2195.

Agirre, X., Vilas-Zornoza, A., Jimenez-Velasco, A., Martin-Subero, J.I., Cordeu, L., Garate, L., San Jose-Eneriz, E., Abizanda, G., Rodriguez-Otero, P., Fortes, P.*, et al.* (2009). Epigenetic silencing of the tumor suppressor microRNA Hsa-miR-124a regulates CDK6 expression and confers a poor prognosis in acute lymphoblastic leukemia. Cancer research *69*, 4443-4453.

Ala, U., Karreth, F.A., Bosia, C., Pagnani, A., Taulli, R., Leopold, V., Tay, Y., Provero, P., Zecchina, R., and Pandolfi, P.P. (2013). Integrated transcriptional and competitive endogenous RNA networks are cross-regulated in permissive molecular environments. P Natl Acad Sci USA *110*, 7154-7159.

Aleman, T.S., Cideciyan, A.V., Huang, Y., Volpe, N., De Castro, E.B., Stevanin, G., Brice, A., and Jacobson, S.G. (2000). Retinal degeneration associated with spinocerebellar ataxia type 7 (SCA7). Invest Ophth Vis Sci *41*, S175-S175.

Alexiou, P., Maragkakis, M., Papadopoulos, G.L., Reczko, M., and Hatzigeorgiou, A.G. (2009). Lost in translation: an assessment and perspective for computational microRNA target identification. Bioinformatics *25*, 3049-3055.

Alvarez Retuerto, A.I., Cantor, R.M., Gleeson, J.G., Ustaszewska, A., Schackwitz, W.S., Pennacchio, L.A., and Geschwind, D.H. (2008). Association of common variants in the Joubert syndrome gene (AHI1) with autism. Human molecular genetics *17*, 3887-3896.

Ambros, V. (1989). A Hierarchy of Regulatory Genes Controls a Larva-to-Adult Developmental Switch in C-Elegans. Cell *57*, 49-57.

Ambros, V. (2003). MicroRNA pathways in flies and worms: growth, death, fat, stress, and timing. Cell *113*, 673-676.

Ambros, V. (2004). The functions of animal microRNAs. Nature *431*, 350-355.

Ameres, S.L., Horwich, M.D., Hung, J.H., Xu, J., Ghildiyal, M., Weng, Z., and Zamore, P.D. (2010a). Target RNA-directed trimming and tailing of small silencing RNAs. Science *328*, 1534-1539.

Ameres, S.L., Horwich, M.D., Hung, J.H., Xu, J., Ghildiyal, M., Weng, Z.P., and Zamore, P.D. (2010b). Target RNA-Directed Trimming and Tailing of Small Silencing RNAs. Science *328*, 1534-1539.

Amir, R.E., Van den Veyver, I.B., Wan, M., Tran, C.Q., Francke, U., and Zoghbi, H.Y. (1999). Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2. Nat Genet *23*, 185-188.

Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T.*, et al.* (2014). An atlas of active enhancers across human cell types and tissues. Nature *507*, 455-461.

Anney, R., Klei, L., Pinto, D., Regan, R., Conroy, J., Magalhaes, T.R., Correia, C., Abrahams, B.S., Sykes, N., Pagnamenta, A.T.*, et al.* (2010). A genome-wide scan for common alleles affecting risk for autism. Human molecular genetics *19*, 4072-4082.

Arabidopsis Genome, I. (2000). Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature *408*, 796-815.

Arima, T., Matsuda, T., Takagi, N., and Wake, N. (1997). Association of IGF2 and H19 imprinting with choriocarcinoma development. Cancer Genet Cytogen *93*, 39-47.
Asakawa, K., and Kawakami, K. (2008). Targeted gene expression by the Gal4-UAS system in zebrafish. Dev Growth Differ *50*, 391-399.
Ascano, M., Hafner, M., Cekan, P., Gerstberger, S., and Tuschl, T. (2012). Identification of RNA-protein interaction networks using PAR-CLIP. Wires Rna *3*, 159-177.
Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T.*, et al.* (2000). Gene Ontology: tool for the unification of biology. Nat Genet *25*, 25-29.
Babak, T., Zhang, W., Morris, Q., Blencowe, B.J., and Hughes, T.R. (2004). Probing microRNAs with microarrays: tissue specificity and functional inference. RNA *10*, 1813-1819.
Baccarini, A., Chauhan, H., Gardner, T.J., Jayaprakash, A.D., Sachidanandam, R., and Brown, B.D. (2011a). Kinetic analysis reveals the fate of a microRNA following target regulation in mammalian cells. Curr Biol *21*, 369-376.
Baccarini, A., Chauhan, H., Gardner, T.J., Jayaprakash, A.D., Sachidanandam, R., and Brown, B.D. (2011b). Kinetic Analysis Reveals the Fate of a MicroRNA following Target Regulation in Mammalian Cells. Curr Biol *21*, 369-376.
Bader, A.G., Brown, D., and Winkler, M. (2010). The Promise of MicroRNA Replacement Therapy. Cancer research *70*, 7027-7030.
Baek, D., Villen, J., Shin, C., Camargo, F.D., Gygi, S.P., and Bartel, D.P. (2008). The impact of microRNAs on protein output. Nature *455*, 64-U38.
Bail, S., Swerdel, M., Liu, H.D., Jiao, X.F., Goff, L.A., Hart, R.P., and Kiledjian, M. (2010). Differential regulation of microRNA stability. Rna *16*, 1032-1039.
Bandiera, S., Cartault, F., Jannot, A.S., Hatem, E., Girard, M., Rifai, L., Loiseau, C., Munnich, A., Lyonnet, S., and Henrion-Caude, A. (2013). Genetic variations creating microRNA target sites in the FXN 3'-UTR affect frataxin expression in Friedreich ataxia. PloS one *8*, e54791.
Barad, O., Meiri, E., Avniel, A., Aharonov, R., Barzilai, A., Bentwich, I., Einav, U., Glad, S., Hurban, P., Karov, Y.*, et al.* (2004). MicroRNA expression detected by oligonucleotide microarrays: System establishment and expression profiling in human tissues. Genome Res *14*, 2486-2494.
Barbato, C., Arisi, I., Frizzo, M.E., Brandi, R., Da Sacco, L., and Masotti, A. (2009). Computational Challenges in miRNA Target Predictions: To Be or Not to Be a True Target? J Biomed Biotechnol.
Bargaje, R., Gupta, S., Sarkeshik, A., Park, R., Xu, T., Sarkar, M., Halimani, M., Roy, S.S., Yates, J., and Pillai, B. (2012). Identification of Novel Targets for miR-29a Using miRNA Proteomics. PloS one *7*.
Barnby, G., Abbott, A., Sykes, N., Morris, A., Weeks, D.E., Mott, R., Lamb, J., Bailey, A.J., Monaco, A.P., and International Molecular Genetics Study of Autism, C. (2005). Candidate-gene screening and association analysis at the autism-susceptibility locus on chromosome 16p: evidence of association at GRIN2A and ABAT. Am J Hum Genet *76*, 950-966.
Baron-Cohen, S., Auyeung, B., Norgaard-Pedersen, B., Hougaard, D.M., Abdallah, M.W., Melgaard, L., Cohen, A.S., Chakrabarti, B., Ruta, L., and Lombardo, M.V. (2014). Elevated fetal steroidogenic activity in autism. Molecular psychiatry.
Bartel, D.P. (2004). MicroRNAs: Genomics, biogenesis, mechanism, and function. Cell *116*, 281-297.
Bartel, D.P. (2009). MicroRNAs: target recognition and regulatory functions. Cell *136*, 215-233.
Bartonicek, N., and Enright, A.J. (2010). SylArray: a web server for automated detection of miRNA effects from expression data. Bioinformatics *26*, 2900-2901.
Baskerville, S., and Bartel, D.P. (2005). Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes. Rna *11*, 241-247.

Bauer, K.M., and Hummon, A.B. (2012). Effects of the miR-143/-145 microRNA cluster on the colon cancer proteome and transcriptome. J Proteome Res *11*, 4744-4754.
Bazzini, A.A., Lee, M.T., and Giraldez, A.J. (2012). Ribosome profiling shows that miR-430 reduces translation before causing mRNA decay in zebrafish. Science *336*, 233-237.
Behm-Ansmant, I., Rehwinkel, J., Doerks, T., Stark, A., Bork, P., and Izaurralde, E. (2006a). MRNA degradation by miRNAs and GW182 requires both CCR4 : NOT deadenylase and DCP1 : DCP2 decapping complexes. Gene Dev *20*, 1885-1898.
Behm-Ansmant, I., Rehwinkel, J., Doerks, T., Stark, A., Bork, P., and Izaurralde, E. (2006b). mRNA degradation by miRNAs and GW182 requires both CCR4:NOT deadenylase and DCP1:DCP2 decapping complexes. Genes & development *20*, 1885-1898.
Bernstein, B.E., Birney, E., Dunham, I., Green, E.D., Gunter, C., and Snyder, M. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57-74.
Bernstein, E., Kim, S.Y., Carmell, M.A., Murchison, E.P., Alcorn, H., Li, M.Z., Mills, A.A., Elledge, S.J., Anderson, K.V., and Hannon, G.J. (2003). Dicer is essential for mouse development. Nat Genet *35*, 215-217.
Bertani, S., Sauer, S., Bolotin, E., and Sauer, F. (2011). The noncoding RNA Mistral activates Hoxa6 and Hoxa7 expression and stem cell differentiation by recruiting MLL1 to chromatin. Mol Cell *43*, 1040-1046.
Bertone, P., Stolc, V., Royce, T.E., Rozowsky, J.S., Urban, A.E., Zhu, X., Rinn, J.L., Tongprasit, W., Samanta, M., Weissman, S*., et al.* (2004). Global identification of human transcribed sequences with genome tiling arrays. Science *306*, 2242-2246.
Betancur, C. (2011). Etiological heterogeneity in autism spectrum disorders: more than 100 genetic and genomic disorders and still counting. Brain Res *1380*, 42-77.
Betel, D., Wilson, M., Gabow, A., Marks, D.S., and Sander, C. (2008). The microRNA.org resource: targets and expression. Nucleic acids research *36*, D149-153.
Bithell, A., Johnson, R., and Buckley, N.J. (2009). Transcriptional dysregulation of coding and non-coding genes in cellular models of Huntington's disease. Biochem Soc Trans *37*, 1270-1275.
Blake, D.J., Weir, A., Newey, S.E., and Davies, K.E. (2002). Function and genetics of dystrophin and dystrophin-related proteins in muscle. Physiol Rev *82*, 291-329.
Bolisetty, M.T., Dy, G., Tam, W., and Beemon, K.L. (2009). Reticuloendotheliosis Virus Strain T Induces miR-155, Which Targets JARID2 and Promotes Cell Survival. J Virol *83*, 12009-12017.
Borchert, G.M., Lanier, W., and Davidson, B.L. (2006). RNA polymerase III transcribes human microRNAs. Nat Struct Mol Biol *13*, 1097-1101.
Boutla, A., Delidakis, C., and Tabler, M. (2003). Developmental defects by antisense-mediated inactivation of micro-RNAs 2 and 13 in Drosophila and the identification of putative target genes. Nucleic acids research *31*, 4973-4980.
Brannan, C.I., Dees, E.C., Ingram, R.S., and Tilghman, S.M. (1990). The Product of the H19 Gene May Function as an Rna. Mol Cell Biol *10*, 28-36.
Bras, J., Guerreiro, R., and Hardy, J. (2012). Use of next-generation sequencing and other whole-genome strategies to dissect neurological disease. Nature reviews Neuroscience *13*, 453-464.
Brawand, D., Soumillon, M., Necsulea, A., Julien, P., Csardi, G., Harrigan, P., Weier, M., Liechti, A., Aximu-Petri, A., Kircher, M*., et al.* (2011). The evolution of gene expression levels in mammalian organs. Nature *478*, 343-348.
Brennecke, J., Hipfner, D.R., Stark, A., Russell, R.B., and Cohen, S.M. (2003). bantam encodes a developmentally regulated microRNA that controls cell proliferation and regulates the proapoptotic gene hid in Drosophila. Cell *113*, 25-36.
Brenner, J.L., Jasiewicz, K.L., Fahley, A.F., Kemp, B.J., and Abbott, A.L. (2010). Loss of Individual MicroRNAs Causes Mutant Phenotypes in Sensitized Genetic Backgrounds in C. elegans. Curr Biol *20*, 1321-1325.

Brockdorff, N. (2013). Noncoding RNA and Polycomb recruitment. RNA *19*, 429-442.
Brockdorff, N., Ashworth, A., Kay, G.F., McCabe, V.M., Norris, D.P., Cooper, P.J., Swift, S., and Rastan, S. (1992). The product of the mouse Xist gene is a 15 kb inactive X-specific transcript containing no conserved ORF and located in the nucleus. Cell *71*, 515-526.
Broderick, J.A., and Zamore, P.D. (2014). Competitive Endogenous RNAs Cannot Alter MicroRNA Function In Vivo. Mol Cell *54*, 711-713.
Brown, B.D., and Naldini, L. (2009). INNOVATION Exploiting and antagonizing microRNA regulation for therapeutic and experimental applications. Nat Rev Genet *10*, 578-585.
Brumbaugh, C.D., Kim, H.J., Giovacchini, M., and Pourmand, N. (2011). NanoStriDE: normalization and differential expression analysis of NanoString nCounter data. BMC bioinformatics *12*, 479.
Burroughs, A.M., Ando, Y., de Hoon, M.J., Tomaru, Y., Nishibu, T., Ukekawa, R., Funakoshi, T., Kurokawa, T., Suzuki, H., Hayashizaki, Y.*, et al.* (2010). A comprehensive survey of 3' animal miRNA modification events and a possible role for 3' adenylation in modulating miRNA targeting effectiveness. Genome Res *20*, 1398-1410.
Cabili, M.N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., and Rinn, J.L. (2011). Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes & development *25*, 1915-1927.
Cai, X.Z., 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.
Calin, G.A., Liu, C.G., Sevignani, C., Ferracin, M., Felli, N., Dumitru, C.D., Shimizu, M., Cimmino, A., Zupo, S., Dono, M.*, et al.* (2004). MicroRNA profiling reveals distinct signatures in B cell chronic lymphocytic leukemias. Proc Natl Acad Sci U S A *101*, 11755-11760.
Cancel, G., Duyckaerts, C., Holmberg, M., Zander, C., Yvert, G., Lebre, A.S., Ruberg, M., Faucheux, B., Agid, Y., Hirsch, E.*, et al.* (2000). Distribution of ataxin-7 in normal human brain and retina. Brain *123 Pt 12*, 2519-2530.
Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M.C., Maeda, N., Oyama, R., Ravasi, T., Lenhard, B., Wells, C.*, et al.* (2005). The transcriptional landscape of the mammalian genome. Science *309*, 1559-1563.
Carthew, R.W., and Sontheimer, E.J. (2009). Origins and Mechanisms of miRNAs and siRNAs. Cell *136*, 642-655.
Celniker, S.E., Dillon, L.A., Gerstein, M.B., Gunsalus, K.C., Henikoff, S., Karpen, G.H., Kellis, M., Lai, E.C., Lieb, J.D., MacAlpine, D.M.*, et al.* (2009). Unlocking the secrets of the genome. Nature *459*, 927-930.
Cesana, M., Cacchiarelli, D., Legnini, I., Santini, T., Sthandier, O., Chinappi, M., Tramontano, A., and Bozzoni, I. (2011). A Long Noncoding RNA Controls Muscle Differentiation by Functioning as a Competing Endogenous RNA. Cell *147*, 358-369.
Chaudhuri, K., and Chatterjee, R. (2007). MicroRNA detection and target prediction: integration of computational and experimental approaches. DNA Cell Biol *26*, 321-337.
Chen, X., Liang, H., Zhang, C.Y., and Zen, K. (2012a). miRNA regulates noncoding RNA: a noncanonical function model. Trends in biochemical sciences *37*, 457-459.
Chen, Y.C., Gatchel, J.R., Lewis, R.W., Mao, C.A., Grant, P.A., Zoghbi, H.Y., and Dent, S.Y. (2012b). Gcn5 loss-of-function accelerates cerebellar and retinal degeneration in a SCA7 mouse model. Human molecular genetics *21*, 394-405.
Chew, G.L., Pauli, A., Rinn, J.L., Regev, A., Schier, A.F., and Valen, E. (2013). Ribosome profiling reveals resemblance between long non-coding RNAs and 5' leaders of coding RNAs. Development *140*, 2828-2834.
Chi, S.W., Zang, J.B., Mele, A., and Darnell, R.B. (2009). Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. Nature *460*, 479-486.

Chodroff, R.A., Goodstadt, L., Sirey, T.M., Oliver, P.L., Davies, K.E., Green, E.D., Molnar, Z., and Ponting, C.P. (2010). Long noncoding RNA genes: conservation of sequence and brain expression among diverse amniotes. Genome Biol *11*, R72.
Chou, A.H., Chen, C.Y., Chen, S.Y., Chen, W.J., Chen, Y.L., Weng, Y.S., and Wang, H.L. (2010). Polyglutamine-expanded ataxin-7 causes cerebellar dysfunction by inducing transcriptional dysregulation. Neurochem Int *56*, 329-339.
Chou, C.H., Lin, F.M., Chou, M.T., Hsu, S.D., Chang, T.H., Weng, S.L., Shrestha, S., Hsiao, C.C., Hung, J.H., and Huang, H.D. (2013). A computational approach for identifying microRNA-target interactions using high-throughput CLIP and PAR-CLIP sequencing. BMC genomics *14*.
Chu, C., Qu, K., Zhong, F.L., Artandi, S.E., and Chang, H.Y. (2011). Genomic maps of long noncoding RNA occupancy reveal principles of RNA-chromatin interactions. Mol Cell *44*, 667-678.
Church, D.M., Goodstadt, L., Hillier, L.W., Zody, M.C., Goldstein, S., She, X., Bult, C.J., Agarwala, R., Cherry, J.L., DiCuccio, M.*, et al.* (2009). Lineage-specific biology revealed by a finished genome assembly of the mouse. Plos Biol *7*, e1000112.
Clamp, M., Fry, B., Kamal, M., Xie, X., Cuff, J., Lin, M.F., Kellis, M., Lindblad-Toh, K., and Lander, E.S. (2007). Distinguishing protein-coding and noncoding genes in the human genome. Proc Natl Acad Sci U S A *104*, 19428-19433.
Clark, M.B., Johnston, R.L., Inostroza-Ponta, M., Fox, A.H., Fortini, E., Moscato, P., Dinger, M.E., and Mattick, J.S. (2012). Genome-wide analysis of long noncoding RNA stability. Genome Res *22*, 885-898.
Coccia, E.M., Cicala, C., Charlesworth, A., Ciccarelli, C., Rossi, G.B., Philipson, L., and Sorrentino, V. (1992). Regulation and Expression of a Growth Arrest-Specific Gene (Gas5) during Growth, Differentiation, and Development. Mol Cell Biol *12*, 3514-3521.
Cocquerelle, C., Mascrez, B., Hetuin, D., and Bailleul, B. (1993). Mis-splicing yields circular RNA molecules. Faseb J *7*, 155-160.
Cole, K.A., Attiyeh, E.F., Mosse, Y.P., Laquaglia, M.J., Diskin, S.J., Brodeur, G.M., and Maris, J.M. (2008). A functional screen identifies miR-34a as a candidate neuroblastoma tumor suppressor gene. Mol Cancer Res *6*, 735-742.
Costa, F.F. (2005). Non-coding RNAs: new players in eukaryotic biology. Gene *357*, 83-94.
Damiani, D., Alexander, J.J., O'Rourke, J.R., McManus, M., Jadhav, A.P., Cepko, C.L., Hauswirth, W.W., Harfe, B.D., and Strettoi, E. (2008). Dicer inactivation leads to progressive functional and structural degeneration of the mouse retina. The Journal of neuroscience : the official journal of the Society for Neuroscience *28*, 4878-4887.
Dani, C., Blanchard, J.M., Piechaczyk, M., El Sabouty, S., Marty, L., and Jeanteur, P. (1984). Extreme instability of myc mRNA in normal and transformed human cells. Proc Natl Acad Sci U S A *81*, 7046-7050.
Darnell, R.B. (2010). HITS-CLIP: panoramic views of protein-RNA regulation in living cells. Wiley interdisciplinary reviews RNA *1*, 266-286.
Das, S., Ghosal, S., Sen, R., and Chakrabarti, J. (2014). lnCeDB: Database of Human Long Noncoding RNA Acting as Competing Endogenous RNA. PloS one *9*, e98965.
Daughters, R.S., Tuttle, D.L., Gao, W.C., Ikeda, Y., Moseley, M.L., Ebner, T.J., Swanson, M.S., and Ranum, L.P.W. (2009). RNA Gain-of-Function in Spinocerebellar Ataxia Type 8. Plos Genet *5*.
David, G., Abbas, N., Stevanin, G., Durr, A., Yvert, G., Cancel, G., Weber, C., Imbert, G., Saudou, F., Antoniou, E.*, et al.* (1997). Cloning of the SCA7 gene reveals a highly unstable CAG repeat expansion. Nat Genet *17*, 65-70.
de Giorgio, A., Krell, J., Harding, V., Stebbing, J., and Castellano, L. (2013). Emerging roles of competing endogenous RNAs in cancer: insights from the regulation of PTEN. Mol Cell Biol *33*, 3976-3982.
De Santa, F., Barozzi, I., Mietton, F., Ghisletti, S., Polletti, S., Tusi, B.K., Muller, H., Ragoussis, J., Wei, C.L., and Natoli, G. (2010). A large fraction of extragenic RNA pol II transcription sites overlap enhancers. Plos Biol *8*, e1000384.

Decker, C.J., and Parker, R. (1993). A turnover pathway for both stable and unstable mRNAs in yeast: evidence for a requirement for deadenylation. Genes & development *7*, 1632-1643.
Denli, A.M., Tops, B.B.J., Plasterk, R.H.A., Ketting, R.F., and Hannon, G.J. (2004). Processing of primary microRNAs by the Microprocessor complex. Nature *432*, 231-235.
Denzler, R., Agarwal, V., Stefano, J., Bartel, D.P., and Stoffel, M. (2014). Assessing the ceRNA Hypothesis with Quantitative Measurements of miRNA and Target Abundance. Mol Cell *54*, 766-776.
Deo, M., Yu, J.Y., Chung, K.H., Tippens, M., and Turner, D.L. (2006). Detection of mammalian microRNA expression by in situ hybridization with RNA oligonucleotides. Developmental dynamics : an official publication of the American Association of Anatomists *235*, 2538-2548.
Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., Guernec, G., Martin, D., Merkel, A., Knowles, D.G.*, et al.* (2012). The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome Res *22*, 1775-1789.
Djuranovic, S., Nahvi, A., and Green, R. (2011). A Parsimonious Model for Gene Regulation by miRNAs. Science *331*, 550-553.
Djuranovic, S., Nahvi, A., and Green, R. (2012). miRNA-mediated gene silencing by translational repression followed by mRNA deadenylation and decay. Science *336*, 237-240.
Doench, J.G., and Sharp, P.A. (2004). Specificity of microRNA target selection in translational repression. Gene Dev *18*, 504-511.
Dolken, L., Malterer, G., Erhard, F., Kothe, S., Friedel, C.C., Suffert, G., Marcinowski, L., Motsch, N., Barth, S., Beitzinger, M.*, et al.* (2010). Systematic Analysis of Viral and Cellular MicroRNA Targets in Cells Latently Infected with Human gamma-Herpesviruses by RISC Immunoprecipitation Assay. Cell Host Microbe *7*, 324-334.
Dreszer, T.R., Karolchik, D., Zweig, A.S., Hinrichs, A.S., Raney, B.J., Kuhn, R.M., Meyer, L.R., Wong, M., Sloan, C.A., Rosenbloom, K.R.*, et al.* (2012). The UCSC Genome Browser database: extensions and updates 2011. Nucleic acids research *40*, D918-923.
Ebert, M.S., Neilson, J.R., and Sharp, P.A. (2007). MicroRNA sponges: competitive inhibitors of small RNAs in mammalian cells. Nat Methods *4*, 721-726.
Ebert, M.S., and Sharp, P.A. (2010). Emerging roles for natural microRNA sponges. Curr Biol *20*, R858-861.
Ebert, M.S., and Sharp, P.A. (2012). Roles for MicroRNAs in Conferring Robustness to Biological Processes. Cell *149*, 515-524.
Ebisuya, M., Yamamoto, T., Nakajima, M., and Nishida, E. (2008). Ripples from neighbouring transcription. Nat Cell Biol *10*, 1106-1113.
Ebner, O.A., and Selbach, M. (2011). Whole cell proteome regulation by microRNAs captured in a pulsed SILAC mass spectrometry approach. Methods Mol Biol *725*, 315-331.
Eddy, S.R. (2001). Non-coding RNA genes and the modern RNA world. Nat Rev Genet *2*, 919-929.
Elkan-Miller, T., Ulitsky, I., Hertzano, R., Rudnicki, A., Dror, A.A., Lenz, D.R., Elkon, R., Irmler, M., Beckers, J., Shamir, R.*, et al.* (2011). Integration of Transcriptomics, Proteomics, and MicroRNA Analyses Reveals Novel MicroRNA Regulation of Targets in the Mammalian Inner Ear. PloS one *6*.
Elkin, M., Shevelev, A., Schulze, E., Tykocinsky, M., Cooper, M., Ariel, I., Pode, D., Kopf, E., Degroot, N., and Hochberg, A. (1995). The Expression of the Imprinted H19 and Igf-2 Genes in Human Bladder-Carcinoma. Febs Lett *374*, 57-61.
Enright, A.J., John, B., Gaul, U., Tuschl, T., Sander, C., and Marks, D.S. (2004). MicroRNA targets in Drosophila. Genome Biol *5*.

Enstrom, A.M., Onore, C.E., Van de Water, J.A., and Ashwood, P. (2010). Differential monocyte responses to TLR ligands in children with autism spectrum disorders. Brain, behavior, and immunity *24*, 64-71.
Erhard, F., Dolken, L., Jaskiewicz, L., and Zimmer, R. (2013). PARma: identification of microRNA target sites in AGO-PAR-CLIP data. Genome Biol *14*, R79.
Eulalio, A., Huntzinger, E., Nishihara, T., Rehwinkel, J., Fauser, M., and Izaurralde, E. (2009). Deadenylation is a widespread effect of miRNA regulation. RNA *15*, 21-32.
Eulalio, A., Rehwinkel, J., Stricker, M., Huntzinger, E., Yang, S.F., Doerks, T., Dorner, S., Bork, P., Boutros, M., and Izaurralde, E. (2007). Target-specific requirements for enhancers of decapping in miRNA-mediated gene silencing. Genes & development *21*, 2558-2570.
Fabian, M.R., Sonenberg, N., and Filipowicz, W. (2010). Regulation of mRNA translation and stability by microRNAs. Annual review of biochemistry *79*, 351-379.
Faghihi, M.A., Zhang, M., Huang, J., Modarresi, F., Van der Brug, M.P., Nalls, M.A., Cookson, M.R., St-Laurent, G., and Wahlestedt, C. (2010). Evidence for natural antisense transcript-mediated inhibition of microRNA function. Genome Biol *11*.
Fan, M., Li, X., Jiang, W., Huang, Y., Li, J., and Wang, Z. (2013). A long non-coding RNA, PTCSC3, as a tumor suppressor and a target of miRNAs in thyroid cancer cells. Experimental and therapeutic medicine *5*, 1143-1146.
Fang, M., Wang, J., Zhang, X., Geng, Y., Hu, Z., Rudd, J.A., Ling, S., Chen, W., and Han, S. (2012). The miR-124 regulates the expression of BACE1/beta-secretase correlated with cell death in Alzheimer's disease. Toxicology letters *209*, 94-105.
Farh, K.K.H., Grimson, A., Jan, C., Lewis, B.P., Johnston, W.K., Lim, L.P., Burge, C.B., and Bartel, D.P. (2005). The widespread impact of mammalian microRNAs on mRNA repression and evolution. Science *310*, 1817-1821.
Felsenstein, J. (1993). PHYLIP: Phylogeny Inference Package Version 3.57c (University of Washington, Seattle).
Figliuzzi, M., Marinari, E., and De Martino, A. (2013). MicroRNAs as a Selective Channel of Communication between Competing RNAs: a Steady-State Theory. Biophys J *104*, 1203-1213.
Flicek, P., Amode, M.R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S.*, et al.* (2012). Ensembl 2012. Nucleic acids research *40*, D84-D90.
Flynt, A.S., Greimann, J.C., Chung, W.J., Lima, C.D., and Lai, E.C. (2010). MicroRNA Biogenesis via Splicing and Exosome-Mediated Trimming in Drosophila. Mol Cell *38*, 900-907.
Franco-Zorrilla, J.M., Valli, A., Todesco, M., Mateos, I., Puga, M.I., Rubio-Somoza, I., Leyva, A., Weigel, D., Garcia, J.A., and Paz-Ares, J. (2007). Target mimicry provides a new mechanism for regulation of microRNA activity. Nat Genet *39*, 1033-1037.
Freitag, C.M. (2007). The genetics of autistic disorders and its clinical relevance: a review of the literature. Molecular psychiatry *12*, 2-22.
Freitag, C.M., Staal, W., Klauck, S.M., Duketis, E., and Waltes, R. (2010). Genetics of autistic disorders: review and clinical implications. European child & adolescent psychiatry *19*, 169-178.
Friedman, R.C., Farh, K.K.H., Burge, C.B., and Bartel, D.P. (2009). Most mammalian mRNAs are conserved targets of microRNAs. Genome Res *19*, 92-105.
Fujita, P.A., Rhead, B., Zweig, A.S., Hinrichs, A.S., Karolchik, D., Cline, M.S., Goldman, M., Barber, G.P., Clawson, H., Coelho, A.*, et al.* (2011). The UCSC Genome Browser database: update 2011. Nucleic acids research *39*, D876-882.
Gabory, A., Ripoche, M.A., Yoshimizu, T., and Dandolo, L. (2006). The H19 gene: regulation and function of a non-coding RNA. Cytogenet Genome Res *113*, 188-193.
Gantier, M.P., McCoy, C.E., Rusinova, I., Saulep, D., Wang, D., Xu, D., Irving, A.T., Behlke, M.A., Hertzog, P.J., Mackay, F.*, et al.* (2011). Analysis of microRNA turnover in mammalian cells following Dicer1 ablation. Nucleic Acids Res *39*, 5692-5703.

Garbett, K., Ebert, P.J., Mitchell, A., Lintas, C., Manzi, B., Mirnics, K., and Persico, A.M. (2008). Immune transcriptome alterations in the temporal cortex of subjects with autism. Neurobiol Dis *30*, 303-311.
Garcia, D.M., Baek, D., Shin, C., Bell, G.W., Grimson, A., and Bartel, D.P. (2011). Weak seed-pairing stability and high target-site abundance decrease the proficiency of lsy-6 and other microRNAs. Nat Struct Mol Biol *18*, 1139-U1175.
Geiss, G.K., Bumgarner, R.E., Birditt, B., Dahl, T., Dowidar, N., Dunaway, D.L., Fell, H.P., Ferree, S., George, R.D., Grogan, T.*, et al.* (2008). Direct multiplexed measurement of gene expression with color-coded probe pairs. Nature biotechnology *26*, 317-325.
Genomes Project, C., Abecasis, G.R., Altshuler, D., Auton, A., Brooks, L.D., Durbin, R.M., Gibbs, R.A., Hurles, M.E., and McVean, G.A. (2010). A map of human genome variation from population-scale sequencing. Nature *467*, 1061-1073.
Gerstein, M.B., Lu, Z.J., Van Nostrand, E.L., Cheng, C., Arshinoff, B.I., Liu, T., Yip, K.Y., Robilotto, R., Rechtsteiner, A., Ikegami, K.*, et al.* (2010). Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. Science *330*, 1775-1787.
Geschwind, D.H. (2011). Genetics of autism spectrum disorders. Trends in cognitive sciences *15*, 409-416.
Ghildiyal, M., and Zamore, P.D. (2009). Small silencing RNAs: an expanding universe. Nat Rev Genet *10*, 94-108.
Giraldez, A.J., Mishima, Y., Rihel, J., Grocock, R.J., Van Dongen, S., Inoue, K., Enright, A.J., and Schier, A.F. (2006). Zebrafish MiR-430 promotes deadenylation and clearance of maternal mRNAs. Science *312*, 75-79.
Gouw, L.G., Castaneda, M.A., McKenna, C.K., Digre, K.B., Pulst, S.M., Perlman, S., Lee, M.S., Gomez, C., Fischbeck, K., Gagnon, D.*, et al.* (1998). Analysis of the dynamic mutation in the SCA7 gene shows marked parental effects on CAG repeat transmission. Human molecular genetics *7*, 525-532.
Gregory, R.I., Yan, K.P., Amuthan, G., Chendrimada, T., Doratotaj, B., Cooch, N., and Shiekhattar, R. (2004). The Microprocessor complex mediates the genesis of microRNAs. Nature *432*, 235-240.
Grimson, A., Farh, K.K.H., Johnston, W.K., Garrett-Engele, P., Lim, L.P., and Bartel, D.P. (2007). MicroRNA targeting specificity in mammals: Determinants beyond seed pairing. Mol 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.
Grosswendt, S., Filipchyk, A., Manzano, M., Klironomos, F., Schilling, M., Herzog, M., Gottwein, E., and Rajewsky, N. (2014). Unambiguous Identification of miRNA:Target Site Interactions by Different Types of Ligation Reactions. Mol Cell *54*, 1042-1054.
Grun, D., Wang, Y.L., Langenberger, D., Gunsalus, K.C., and Rajewsky, N. (2005). MicroRNA target predictions across seven Drosophila species and comparison to mammalian targets. Plos Comput Biol *1*, 51-66.
Gu, S., Jin, L., Zhang, F.J., Sarnow, P., and Kay, M.A. (2009). Biological basis for restriction of microRNA targets to the 3 ' untranslated region in mammalian mRNAs. Nat Struct Mol Biol *16*, 144-150.
Guo, H., Ingolia, N.T., Weissman, J.S., and Bartel, D.P. (2010). Mammalian microRNAs predominantly act to decrease target mRNA levels. Nature *466*, 835-840.
Guttman, M., Amit, I., Garber, M., French, C., Lin, M.F., Feldser, D., Huarte, M., Zuk, O., Carey, B.W., Cassady, J.P.*, et al.* (2009). Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature *458*, 223-227.
Guttman, M., Donaghey, J., Carey, B.W., Garber, M., Grenier, J.K., Munson, G., Young, G., Lucas, A.B., Ach, R., Bruhn, L.*, et al.* (2011). lincRNAs act in the circuitry controlling pluripotency and differentiation. Nature *477*, 295-U260.

Guttman, M., Garber, M., Levin, J.Z., Donaghey, J., Robinson, J., Adiconis, X., Fan, L., Koziol, M.J., Gnirke, A., Nusbaum, C.*, et al.* (2010). Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. Nature biotechnology *28*, 503-510.
Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., Rothballer, A., Ascano, M., 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., Lianoglou, S., Tuschl, T., and Betel, D. (2012). Genome-wide identification of miRNA targets by PAR-CLIP. Methods *58*, 94-105.
Hah, N., Murakami, S., Nagari, A., Danko, C.G., and Kraus, W.L. (2013). Enhancer transcripts mark active estrogen receptor binding sites. Genome Res *23*, 1210-1223.
Hallmayer, J., Cleveland, S., Torres, A., Phillips, J., Cohen, B., Torigoe, T., Miller, J., Fedele, A., Collins, J., Smith, K.*, et al.* (2011). Genetic heritability and shared environmental factors among twin pairs with autism. Archives of general psychiatry *68*, 1095-1102.
Hammell, M., Long, D., Zhang, L., Lee, A., Carmack, C.S., Han, M., Ding, Y., and Ambros, V. (2008). mirWIP: microRNA target prediction based on microRNA-containing ribonucleoprotein-enriched transcripts. Nat Methods *5*, 813-819.
Han, K., Gennarino, V.A., Lee, Y., Pang, K., Hashimoto-Torii, K., Choufani, S., Raju, C.S., Oldham, M.C., Weksberg, R., Rakic, P.*, et al.* (2013). Human-specific regulation of MeCP2 levels in fetal brains by microRNA miR-483-5p. Genes & development *27*, 485-490.
Hansen, T.B., Jensen, T.I., Clausen, B.H., Bramsen, J.B., Finsen, B., Damgaard, C.K., and Kjems, J. (2013). Natural RNA circles function as efficient microRNA sponges. Nature *495*, 384-388.
Hansen, T.B., Wiklund, E.D., Bramsen, J.B., Villadsen, S.B., Statham, A.L., Clark, S.J., and Kjems, J. (2011). miRNA-dependent gene silencing involving Ago2-mediated cleavage of a circular antisense RNA. Embo J *30*, 4414-4422.
He, L., and Hannon, G.J. (2004). MicroRNAs: Small RNAs with a big role in gene regulation (vol 5, pg 522 2004). Nat Rev Genet *5*, 522-+.
Heger, A., and Ponting, C.P. (2007). Variable strength of translational selection among 12 drosophila species. Genetics *177*, 1337-1348.
Helmlinger, D., Hardy, S., Sasorith, S., Klein, F., Robert, F., Weber, C., Miguet, L., Potier, N., Van-Dorsselaer, A., Wurtz, J.M.*, et al.* (2004). Ataxin-7 is a subunit of GCN5 histone acetyltransferase-containing complexes. Human molecular genetics *13*, 1257-1265.
Helwak, A., Kudla, G., Dudnakova, T., and Tollervey, D. (2013). Mapping the Human miRNA Interactome by CLASH Reveals Frequent Noncanonical Binding. Cell *153*, 654-665.
Hendrickson, D.G., Hogan, D.J., Herschlag, D., Ferrell, J.E., and Brown, P.O. (2008a). Systematic Identification of mRNAs Recruited to Argonaute 2 by Specific microRNAs and Corresponding Changes in Transcript Abundance. PloS one *3*.
Hendrickson, D.G., Hogan, D.J., Herschlag, D., Ferrell, J.E., and Brown, P.O. (2008b). Systematic identification of mRNAs recruited to argonaute 2 by specific microRNAs and corresponding changes in transcript abundance. PloS one *3*, e2126.
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.
Hindorff, L.A., Sethupathy, P., Junkins, H.A., Ramos, E.M., Mehta, J.P., Collins, F.S., and Manolio, T.A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A *106*, 9362-9367.
Holmberg, M., Duyckaerts, C., Durr, A., Cancel, G., Gourfinkel-An, I., Damier, P., Faucheux, B., Trottier, Y., Hirsch, E.C., Agid, Y.*, et al.* (1998). Spinocerebellar ataxia

type 7 (SCA7): a neurodegenerative disorder with neuronal intranuclear inclusions. Human molecular genetics *7*, 913-918.
Honti, F., Meader, S., Webber, C. (2014). Unbiased Functional Clustering of Gene Variants with a Phenotypic-Linkage Network. PLOS Comp Bio *In press*.
Hornstein, E., and Shomron, N. (2006). Canalization of development by microRNAs. Nat Genet *38 Suppl*, S20-24.
Houwing, S., Kamminga, L.M., Berezikov, E., Cronembold, D., Girard, A., van den Elst, H., Filippov, D.V., Blaser, H., Raz, E., Moens, C.B.*, et al.* (2007). A role for Piwi and piRNAs in germ cell maintenance and transposon silencing in Zebrafish. Cell *129*, 69-82.
Huang, D.W., Sherman, B.T., and Lempicki, R.A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc *4*, 44-57.
Huang, T.C., Sahasrabuddhe, N.A., Kim, M.S., Getnet, D., Yang, Y., Peterson, J.M., Ghosh, B., Chaerkady, R., Leach, S.D., Marchionni, L.*, et al.* (2012). Regulation of Lipid Metabolism by Dicer Revealed through SILAC Mice. J Proteome Res *11*, 2193-2205.
Huarte, M., Guttman, M., Feldser, D., Garber, M., Koziol, M.J., Kenzelmann-Broz, D., Khalil, A.M., Zuk, O., Amit, I., Rabani, M.*, et al.* (2010). A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response. Cell *142*, 409-419.
Huguet, G., Ey, E., and Bourgeron, T. (2013). The genetic landscapes of autism spectrum disorders. Annual review of genomics and human genetics *14*, 191-213.
Huntzinger, E., and Izaurralde, E. (2011). Gene silencing by microRNAs: contributions of translational repression and mRNA decay. Nat Rev Genet *12*, 99-110.
Hussman, J.P., Chung, R.H., Griswold, A.J., Jaworski, J.M., Salyakina, D., Ma, D., Konidari, I., Whitehead, P.L., Vance, J.M., Martin, E.R.*, et al.* (2011). A noise-reduction GWAS analysis implicates altered regulation of neurite outgrowth and guidance in autism. Molecular autism *2*, 1.
Hutvagner, G., and Simard, M.J. (2008). Argonaute proteins: key players in RNA silencing. Nat Rev Mol Cell Bio *9*, 22-32.
Hwang, H.W., Wentzel, E.A., and Mendell, J.T. (2007). A hexanucleotide element directs microRNA nuclear import. Science *315*, 97-100.
Ibrahim, F., Rymarquis, L.A., Kim, E.J., Becker, J., Balassa, E., Green, P.J., and Cerutti, H. (2010). Uridylation of mature miRNAs and siRNAs by the MUT68 nucleotidyltransferase promotes their degradation in Chlamydomonas. Proc Natl Acad Sci U S A *107*, 3906-3911.
International Human Genome Sequencing, C. (2004). Finishing the euchromatic sequence of the human genome. Nature *431*, 931-945.
Iossifov, I., Ronemus, M., Levy, D., Wang, Z., Hakker, I., Rosenbaum, J., Yamrom, B., Lee, Y.H., Narzisi, G., Leotta, A.*, et al.* (2012). De novo gene disruptions in children on the autistic spectrum. Neuron *74*, 285-299.
Jacq, C., Miller, J.R., and Brownlee, G.G. (1977). A pseudogene structure in 5S DNA of Xenopus laevis. Cell *12*, 109-120.
Jeck, W.R., Sorrentino, J.A., Wang, K., Slevin, M.K., Burd, C.E., Liu, J., Marzluff, W.F., and Sharpless, N.E. (2013). Circular RNAs are abundant, conserved, and associated with ALU repeats. RNA *19*, 141-157.
Jeggari, A., Marks, D.S., and Larsson, E. (2012). miRcode: a map of putative microRNA target sites in the long non-coding transcriptome. Bioinformatics *28*, 2062-2063.
Jensen, K.B., and Darnell, R.B. (2008). CLIP: crosslinking and immunoprecipitation of in vivo RNA targets of RNA-binding proteins. Methods Mol Biol *488*, 85-98.
Jeon, Y., and Lee, J.T. (2011). YY1 tethers Xist RNA to the inactive X nucleation center. Cell *146*, 119-133.
Ji, P., Diederichs, S., Wang, W., Boing, S., Metzger, R., Schneider, P.M., Tidow, N., Brandt, B., Buerger, H., Bulk, E.*, et al.* (2003). MALAT-1, a novel noncoding RNA, and

thymosin beta4 predict metastasis and survival in early-stage non-small cell lung cancer. Oncogene *22*, 8031-8041.
Jia, H., Osak, M., Bogu, G.K., Stanton, L.W., Johnson, R., and Lipovich, L. (2010). Genome-wide computational identification and manual annotation of human long noncoding RNA genes. RNA *16*, 1478-1487.
Jia, L., Landan, G., Pomerantz, M., Jaschek, R., Herman, P., Reich, D., Yan, C., Khalid, O., Kantoff, P., Oh, W.*, et al.* (2009). Functional enhancers at the gene-poor 8q24 cancer-linked locus. Plos Genet *5*, e1000597.
Jiang, L., Liu, X.Q., Chen, Z.J., Jin, Y., Heidbreder, C.E., Kolokythas, A., Wang, A.X., Dai, Y., and Zhou, X.F. (2010). MicroRNA-7 targets IGF1R (insulin-like growth factor 1 receptor) in tongue squamous cell carcinoma cells. Biochemical Journal *432*, 199-205.
John, B., Enright, A.J., Aravin, A., Tuschl, T., Sander, C., and Marks, D.S. (2004). Human MicroRNA targets. Plos Biology *2*, 1862-1879.
Johnson, C.D., Esquela-Kerscher, A., Stefani, G., Byrom, N., Kelnar, K., Ovcharenko, D., Wilson, M., Wang, X.W., Shelton, J., Shingara, J.*, et al.* (2007). The let-7 MicroRNA represses cell proliferation pathways in human cells. Cancer Res *67*, 7713-7722.
Johnson, R., and Buckley, N.J. (2009). Gene dysregulation in Huntington's disease: REST, microRNAs and beyond. Neuromolecular medicine *11*, 183-199.
Jones-Rhoades, M.W., Bartel, D.P., and Bartel, B. (2006). MicroRNAs and their regulatory roles in plants. Annu Rev Plant Biol *57*, 19-53.
Jones, M.R., Quinton, L.J., Blahna, M.T., Neilson, J.R., Fu, S., Ivanov, A.R., Wolf, D.A., and Mizgerd, J.P. (2009). Zcchc11-dependent uridylation of microRNA directs cytokine expression. Nat Cell Biol *11*, 1157-1163.
Jordan, V.C., and Koerner, S. (1975). Tamoxifen (ICI 46,474) and the human carcinoma 8S oestrogen receptor. Eur J Cancer *11*, 205-206.
Jung, C.H., Hansen, M.A., Makunin, I.V., Korbie, D.J., and Mattick, J.S. (2010). Identification of novel non-coding RNAs using profiles of short sequence reads from next generation sequencing data. BMC genomics *11*, 77.
Junn, E., Lee, K.W., Jeong, B.S., Chan, T.W., Im, J.Y., and Mouradian, M.M. (2009). Repression of alpha-synuclein expression and toxicity by microRNA-7. Proc Natl Acad Sci U S A *106*, 13052-13057.
Kai, Z.S., and Pasquinelli, A.E. (2010). MicroRNA assassins: factors that regulate the disappearance of miRNAs. Nat Struct Mol Biol *17*, 5-10.
Kallen, A.N., Zhou, X.B., Xu, J., Qiao, C., Ma, J., Yan, L., Lu, L., Liu, C., Yi, J.S., Zhang, H.*, et al.* (2013). The Imprinted H19 LncRNA Antagonizes Let-7 MicroRNAs. Mol Cell *52*, 101-112.
Kaller, M., Liffers, S.T., Oeljeklaus, S., Kuhlmann, K., Roh, S., Hoffmann, R., Warscheid, B., and Hermeking, H. (2011). Genome-wide Characterization of miR-34a Induced Changes in Protein and mRNA Expression by a Combined Pulsed SILAC and Microarray Analysis. Mol Cell Proteomics *10*.
Kalyana-Sundaram, S., Kumar-Sinha, C., Shankar, S., Robinson, D.R., Wu, Y.M., Cao, X.H., Asangani, I.A., Kothari, V., Prensner, J.R., Lonigro, R.J.*, et al.* (2012). Expressed Pseudogenes in the Transcriptional Landscape of Human Cancers. Cell *149*, 1622-1634.
Kanellopoulou, C., Muljo, S.A., Kung, A.L., Ganesan, S., Drapkin, R., Jenuwein, T., Livingston, D.M., and Rajewsky, K. (2005). Dicer-deficient mouse embryonic stem cells are defective in differentiation and centromeric silencing. Gene Dev *19*, 489-501.
Kanematsu, S., Tanimoto, K., Suzuki, Y., and Sugano, S. (2013). Screening for possible miRNA-mRNA associations in a colon cancer cell line. Gene.
Kapranov, P., Cawley, S.E., Drenkow, J., Bekiranov, S., Strausberg, R.L., Fodor, S.P., and Gingeras, T.R. (2002). Large-scale transcriptional activity in chromosomes 21 and 22. Science *296*, 916-919.
Kapranov, P., Cheng, J., Dike, S., Nix, D.A., Duttagupta, R., Willingham, A.T., Stadler, P.F., Hertel, J., Hackermuller, J., Hofacker, I.L.*, et al.* (2007). RNA maps reveal new

RNA classes and a possible function for pervasive transcription. Science *316*, 1484-1488.
Karginov, F.V., Conaco, C., Xuan, Z., Schmidt, B.H., Parker, J.S., Mandel, G., and Hannon, G.J. (2007). A biochemical approach to identifying microRNA targets. Proc Natl Acad Sci U S A *104*, 19291-19296.
Karreth, F.A., Tay, Y., Pema, D., Ala, U., Tan, S.M., Rust, A.G., DeNicola, G., Webster, K.A., Weiss, D., Perez-Mancera, P.A.*, et al.* (2011). In Vivo Identification of Tumor-Suppressive PTEN ceRNAs in an Oncogenic BRAF-Induced Mouse Model of Melanoma (vol 147, pg 382, 2011). Cell *147*, 948-948.
Karro, J.E., Yan, Y.P., Zheng, D.Y., Zhang, Z.L., Carriero, N., Cayting, P., Harrrison, P., and Gerstein, M. (2007). Pseudogene.org: a comprehensive database and comparison platform for pseudogene annotation. Nucleic acids research *35*, D55-D60.
Karwacki-Neisius, V., Goke, J., Osorno, R., Halbritter, F., Ng, J.H., Weisse, A.Y., Wong, F.C., Gagliardi, A., Mullin, N.P., Festuccia, N.*, et al.* (2013). Reduced Oct4 expression directs a robust pluripotent state with distinct signaling activity and increased enhancer occupancy by Oct4 and Nanog. Cell Stem Cell *12*, 531-545.
Kataoka, Y., Takeichi, M., and Uemura, T. (2001). Developmental roles and molecular characterization of a Drosophila homologue of Arabidopsis Argonaute1, the founder of a novel gene superfamily. Genes Cells *6*, 313-325.
Keene, J.D., Komisarow, J.M., and Friedersdorf, M.B. (2006). RIP-Chip: the isolation and identification of mRNAs, microRNAs and protein components of ribonucleoprotein complexes from cell extracts. Nat Protoc *1*, 302-307.
Kefas, B., Godlewski, J., Comeau, L., Li, Y.Q., Abounader, R., Hawkinson, M., Lee, J.W., Fine, H., Chiocca, E.A., Lawler, S.*, et al.* (2008). microRNA-7 inhibits the epidermal growth factor receptor and the Akt pathway and is down-regulated in glioblastoma. Cancer research *68*, 3566-3572.
Keniry, A., Oxley, D., Monnier, P., Kyba, M., Dandolo, L., Smits, G., and Reik, W. (2012). The H19 lincRNA is a developmental reservoir of miR-675 that suppresses growth and Igf1r. Nat Cell Biol *14*, 659-665.
Keren, H., Lev-Maor, G., and Ast, G. (2010). Alternative splicing and evolution: diversification, exon definition and function. Nat Rev Genet *11*, 345-355.
Kerin, T., Ramanathan, A., Rivas, K., Grepo, N., Coetzee, G.A., and Campbell, D.B. (2012). A noncoding RNA antisense to moesin at 5p14.1 in autism. Science translational medicine *4*, 128ra140.
Khalil, A.M., Guttman, M., Huarte, M., Garber, M., Raj, A., Morales, D.R., Thomas, K., Presser, A., Bernstein, B.E., van Oudenaarden, A.*, et al.* (2009). Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. Proc Natl Acad Sci U S A *106*, 11667-11672.
Khvorova, A., Reynolds, A., and Jayasena, S.D. (2003). Functional siRNAs and miRNAs exhibit strand bias. Cell *115*, 209-216.
Kim, T.K., Hemberg, M., Gray, J.M., Costa, A.M., Bear, D.M., Wu, J., Harmin, D.A., Laptewicz, M., Barbara-Haley, K., Kuersten, S.*, et al.* (2010). Widespread transcription at neuronal activity-regulated enhancers. Nature *465*, 182-187.
Kino, T., Hurt, D.E., Ichijo, T., Nader, N., and Chrousos, G.P. (2010). Noncoding RNA Gas5 Is a Growth Arrest- and Starvation-Associated Repressor of the Glucocorticoid Receptor. Sci Signal *3*.
Kiriakidou, M., Nelson, P.T., Kouranov, A., Fitziev, P., Bouyioukos, C., Mourelatos, Z., and Hatzigeorgiou, A. (2004). A combined computational-experimental approach predicts human microRNA targets. Gene Dev *18*, 1165-1178.
Klauck, S.M. (2006). Genetics of autism spectrum disorder. Eur J Hum Genet *14*, 714-720.
Kohler, A., and Hurt, E. (2007a). Exporting RNA from the nucleus to the cytoplasm. Nat Rev Mol Cell Bio *8*, 761-773.
Kohler, A., and Hurt, E. (2007b). Exporting RNA from the nucleus to the cytoplasm. Nature reviews Molecular cell biology *8*, 761-773.

Kohler, S., Doelken, S.C., Mungall, C.J., Bauer, S., Firth, H.V., Bailleul-Forestier, I., Black, G.C., Brown, D.L., Brudno, M., Campbell, J.*, et al.* (2014). The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. Nucleic acids research *42*, D966-974.
Koller, W.C., and Tse, W. (2004). Unmet medical needs in Parkinson's disease. Neurology *62*, S1-8.
Kong, L., Zhang, Y., Ye, Z.Q., Liu, X.Q., Zhao, S.Q., Wei, L., and Gao, G. (2007). CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. Nucleic acids research *35*, W345-W349.
Koob, M.D., Moseley, M.L., Schut, L.J., Benzow, K.A., Bird, T.D., Day, J.W., and Ranum, L.P.W. (1999). An untranslated CTG expansion causes a novel form of spinocerebellar ataxia (SCA8). Nat Genet *21*, 379-384.
Krantz, I.D., McCallum, J., DeScipio, C., Kaur, M., Gillis, L.A., Yaeger, D., Jukofsky, L., Wasserman, N., Bottani, A., Morris, C.A.*, et al.* (2004). Cornelia de Lange syndrome is caused by mutations in NIPBL, the human homolog of Drosophila melanogaster Nipped-B. Nat Genet *36*, 631-635.
Krek, A., Grun, D., Poy, M.N., Wolf, R., Rosenberg, L., Epstein, E.J., MacMenamin, P., da Piedade, I., Gunsalus, K.C., Stoffel, M.*, et al.* (2005). Combinatorial microRNA target predictions. Nat Genet *37*, 495-500.
Krol, J., Loedige, I., and Filipowicz, W. (2010). The widespread regulation of microRNA biogenesis, function and decay. Nat Rev Genet *11*, 597-610.
Krol, J., Sobczak, K., Wilczynska, U., Drath, M., Jasinska, A., Kaczynska, D., and Krzyzosiak, W.J. (2004). Structural features of microRNA (miRNA) precursors and their relevance to miRNA biogenesis and small interfering RNA/short hairpin RNA design. J Biol Chem *279*, 42230-42239.
Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J., and Marra, M.A. (2009). Circos: an information aesthetic for comparative genomics. Genome Res *19*, 1639-1645.
Kudla, G., Granneman, S., Hahn, D., Beggs, J.D., and Tollervey, D. (2011). Cross-linking, ligation, and sequencing of hybrids reveals RNA-RNA interactions in yeast. P Natl Acad Sci USA *108*, 10010-10015.
Kumar, M.S., Erkeland, S.J., Pester, R.E., Chen, C.Y., Ebert, M.S., Sharp, P.A., and Jacks, T. (2008). Suppression of non-small cell lung tumor development by the let-7 microRNA family. Proc Natl Acad Sci U S A *105*, 3903-3908.
Kumar, V., Westra, H.J., Karjalainen, J., Zhernakova, D.V., Esko, T., Hrdlickova, B., Almeida, R., Zhernakova, A., Reinmaa, E., Vosa, U.*, et al.* (2013). Human disease-associated genetic variation impacts large intergenic non-coding RNA expression. Plos Genet *9*, e1003201.
Kung, J.T., Colognori, D., and Lee, J.T. (2013). Long noncoding RNAs: past, present, and future. Genetics *193*, 651-669.
Kurotaki, N., Imaizumi, K., Harada, N., Masuno, M., Kondoh, T., Nagai, T., Ohashi, H., Naritomi, K., Tsukahara, M., Makita, Y.*, et al.* (2002). Haploinsufficiency of NSD1 causes Sotos syndrome. Nat Genet *30*, 365-366.
Kutter, C., Watt, S., Stefflova, K., Wilson, M.D., Goncalves, A., Ponting, C.P., Odom, D.T., and Marques, A.C. (2012). Rapid Turnover of Long Noncoding RNAs and the Evolution of Gene Expression. Plos Genet *8*.
La Spada, A.R., and Taylor, J.P. (2010). Repeat expansion disease: progress and puzzles in disease pathogenesis. Nat Rev Genet *11*, 247-258.
Lagos-Quintana, M., Rauhut, R., Yalcin, A., Meyer, J., Lendeckel, W., and Tuschl, T. (2002). Identification of tissue-specific microRNAs from mouse. Curr Biol *12*, 735-739.
Lai, F., Orom, U.A., Cesaroni, M., Beringer, M., Taatjes, D.J., Blobel, G.A., and Shiekhattar, R. (2013). Activating RNAs associate with Mediator to enhance chromatin architecture and transcription. Nature *494*, 497-501.

Lam, M.T., Cho, H., Lesch, H.P., Gosselin, D., Heinz, S., Tanaka-Oishi, Y., Benner, C., Kaikkonen, M.U., Kim, A.S., Kosaka, M.*, et al.* (2013). Rev-Erbs repress macrophage gene expression by inhibiting enhancer-directed transcription. Nature *498*, 511-515.
Lander, E.S., Consortium, I.H.G.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M.*, et al.* (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860-921.
Landgraf, P., Rusu, M., Sheridan, R., Sewer, A., Iovino, N., Aravin, A., Pfeffer, S., Rice, A., Kamphorst, A.O., Landthaler, M.*, et al.* (2007). A mammalian microRNA expression atlas based on small RNA library sequencing. Cell *129*, 1401-1414.
Lane, L., Argoud-Puy, G., Britan, A., Cusin, I., Duek, P.D., Evalet, O., Gateau, A., Gaudet, P., Gleizes, A., Masselot, A.*, et al.* (2012). neXtProt: a knowledge platform for human proteins. Nucleic acids research *40*, D76-83.
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, Y., Ahn, C., Han, J.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., Kim, M., Han, J.J., Yeom, K.H., Lee, S., Baek, S.H., and Kim, V.N. (2004). MicroRNA genes are transcribed by RNA polymerase II. Embo J *23*, 4051-4060.
Lee, Y., Samaco, R.C., Gatchel, J.R., Thaller, C., Orr, H.T., and Zoghbi, H.Y. (2008). miR-19, miR-101 and miR-130 co-regulate ATXN1 levels to potentially modulate SCA1 pathogenesis. Nat Neurosci *11*, 1137-1139.
Leighton, P.A., Ingram, R.S., Eggenschwiler, J., Efstratiadis, A., and Tilghman, S.M. (1995). Disruption of Imprinting Caused by Deletion of the H19 Gene Region in Mice. Nature *375*, 34-39.
Leung, A.K., Young, A.G., Bhutkar, A., Zheng, G.X., Bosson, A.D., Nielsen, C.B., and Sharp, P.A. (2011). Genome-wide identification of Ago2 binding sites from mouse embryonic stem cells with and without mature microRNAs. Nat Struct Mol Biol *18*, 237-244.
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, J., Yang, Z., Yu, B., Liu, J., and Chen, X. (2005). Methylation protects miRNAs and siRNAs from a 3'-end uridylation activity in Arabidopsis. Curr Biol *15*, 1501-1507.
Li, L., Eichten, S.R., Shimizu, R., Petsch, K., Yeh, C.T., Wu, W., Chettoor, A.M., Givan, S.A., Cole, R.A., Fowler, J.E.*, et al.* (2014). Genome-wide discovery and characterization of maize long non-coding RNAs. Genome Biol *15*, R40.
Li, W., Notani, D., Ma, Q., Tanasa, B., Nunez, E., Chen, A.Y., Merkurjev, D., Zhang, J., Ohgi, K., Song, X.*, et al.* (2013). Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation. Nature *498*, 516-520.
Li, X., Cassidy, J.J., Reinke, C.A., Fischboeck, S., and Carthew, R.W. (2009). A MicroRNA Imparts Robustness against Environmental Fluctuation during Development. Cell *137*, 273-282.
Liang, H.W., Zhang, J.F., Zen, K., Zhang, C.Y., and Chen, X. (2013). Nuclear microRNAs and their unconventional role in regulating non-coding RNAs. Protein Cell *4*, 325-330.
Liang, R.Q., Li, W., Li, Y., Tan, C.Y., Li, J.X., Jin, Y.X., and Ruan, K.C. (2005). An oligonucleotide microarray for microRNA expression analysis based on labeling RNA with quantum dot and nanogold probe. Nucleic acids research *33*.
Licatalosi, D.D., Mele, A., Fak, J.J., Ule, J., Kayikci, M., Chi, S.W., Clark, T.A., Schweitzer, A.C., Blume, J.E., Wang, X.N.*, et al.* (2008). HITS-CLIP yields genome-wide insights into brain alternative RNA processing. Nature *456*, 464-U422.

Lichner, Z., Pall, E., Kerekes, A., Pallinger, E., Maraghechi, P., Bosze, Z., and Gocza, E. (2011). The miR-290-295 cluster promotes pluripotency maintenance by regulating cell cycle phase distribution in mouse embryonic stem cells. Differentiation; research in biological diversity *81*, 11-24.
Lim, L.P., Lau, N.C., Garrett-Engele, P., Grimson, A., Schelter, J.M., Castle, J., Bartel, D.P., Linsley, P.S., and Johnson, J.M. (2005). Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. Nature *433*, 769-773.
Lin, M.F., Jungreis, I., and Kellis, M. (2011). PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. Bioinformatics *27*, i275-282.
Lindow, M., and Gorodkin, J. (2007). Principles and limitations of computational microRNA gene and target finding. DNA Cell Biol *26*, 339-351.
Linsley, P.S., Schelter, J., Burchard, J., Kibukawa, M., Martin, M.M., Bartz, S.R., Johnson, J.M., Cummins, J.M., Raymond, C.K., Dai, H.Y*., et al.* (2007). Transcripts targeted by the MicroRNA-16 family cooperatively regulate cell cycle progression. Mol Cell Biol *27*, 2240-2252.
Liu, C., Mallick, B., Long, D., Rennie, W.A., Wolenc, A., Carmack, C.S., and Ding, Y. (2013). CLIP-based prediction of mammalian microRNA binding sites. Nucleic acids research *41*, e138.
Liu, C.G., Calin, G.A., Volinia, S., and Croce, C.M. (2008a). MicroRNA expression profiling using microarrays. Nat Protoc *3*, 563-578.
Liu, J. (2008). Control of protein synthesis and mRNA degradation by microRNAs. Curr Opin Cell Biol *20*, 214-221.
Liu, J., Jung, C., Xu, J., Wang, H., Deng, S., Bernad, L., Arenas-Huertero, C., and Chua, N.H. (2012a). Genome-wide analysis uncovers regulation of long intergenic noncoding RNAs in Arabidopsis. Plant Cell *24*, 4333-4345.
Liu, X., Li, D., Zhang, W., Guo, M., and Zhan, Q. (2012b). Long non-coding RNA gadd7 interacts with TDP-43 and regulates Cdk6 mRNA decay. Embo J *31*, 4415-4427.
Liu, X.S., Chopp, M., Zhang, R.L., Tao, T., Wang, X.L., Kassis, H., Hozeska-Solgot, A., Zhang, L., Chen, C., and Zhang, Z.G. (2011). MicroRNA Profiling in Subventricular Zone after Stroke: MiR-124a Regulates Proliferation of Neural Progenitor Cells through Notch Signaling Pathway. PloS one *6*.
Liu, Z., Sall, A., and Yang, D.C. (2008b). MicroRNA: an emerging therapeutic target and intervention tool. Int J Mol Sci *9*, 978-999.
Llave, C., Xie, Z., Kasschau, K.D., and Carrington, J.C. (2002). Cleavage of Scarecrow-like mRNA targets directed by a class of Arabidopsis miRNA. Science *297*, 2053-2056.
Loewer, S., Cabili, M.N., Guttman, M., Loh, Y.H., Thomas, K., Park, I.H., Garber, M., Curran, M., Onder, T., Agarwal, S*., et al.* (2010). Large intergenic non-coding RNA-RoR modulates reprogramming of human induced pluripotent stem cells. Nat Genet *42*, 1113-1117.
Lossner, C., Meier, J., Warnken, U., Rogers, M.A., Lichter, P., Pscherer, A., and Schnolzer, M. (2011). Quantitative proteomics identify novel miR-155 target proteins. PloS one *6*, e22146.
Lund, E., Guttinger, S., Calado, A., Dahlberg, J.E., and Kutay, U. (2004). Nuclear export of microRNA precursors. Science *303*, 95-98.
Lunter, G., Ponting, C.P., and Hein, J. (2006). Genome-wide identification of human functional DNA using a neutral indel model. Plos Comput Biol *2*, 2-12.
Lyon, M.F. (1961). Gene action in the X-chromosome of the mouse (Mus musculus L.). Nature *190*, 372-373.
Ma, D., Salyakina, D., Jaworski, J.M., Konidari, I., Whitehead, P.L., Andersen, A.N., Hoffman, J.D., Slifer, S.H., Hedges, D.J., Cukier, H.N*., et al.* (2009). A genome-wide association study of autism reveals a common novel risk locus at 5p14.1. Annals of human genetics *73*, 263-273.

Makeyev, E.V., Zhang, J., Carrasco, M.A., and Maniatis, T. (2007). The MicroRNA miR-124 promotes neuronal differentiation by triggering brain-specific alternative pre-mRNA splicing. Mol Cell *27*, 435-448.
Mallanna, S.K., and Rizzino, A. (2010). Emerging roles of microRNAs in the control of embryonic stem cells and the generation of induced pluripotent stem cells. Developmental biology *344*, 16-25.
Manolio, T.A. (2010). Genomewide association studies and assessment of the risk of disease. The New England journal of medicine *363*, 166-176.
Marco, A., Macpherson, J.I., Ronshaugen, M., and Griffiths-Jones, S. (2012). MicroRNAs from the same precursor have different targeting properties. Silence *3*, 8.
Marques, A.C., Hughes, J., Graham, B., Kowalczyk, M.S., Higgs, D.R., and Ponting, C.P. (2013). Chromatin signatures at transcriptional start sites separate two equally populated yet distinct classes of intergenic long noncoding RNAs. Genome Biol *14*, R131.
Marques, A.C., and Ponting, C.P. (2009). Catalogues of mammalian long noncoding RNAs: modest conservation and incompleteness. Genome Biol *10*.
Marques, A.C., Tan, J., Lee, S., Kong, L., Heger, A., and Ponting, C.P. (2012). Evidence for conserved post-transcriptional roles of unitary pseudogenes and for frequent bifunctionality of mRNAs. Genome Biol *13*, R102.
Marques, A.C., Tan, J., and Ponting, C.P. (2011). Wrangling for microRNAs provokes much crosstalk. Genome Biol *12*, 132.
Martin, G., Schouest, K., Kovvuru, P., and Spillane, C. (2007). Prediction and validation of microRNA targets in animal genomes. J Biosciences *32*, 1049-1052.
Martinez, E., Palhan, V.B., Tjernberg, A., Lymar, E.S., Gamper, A.M., Kundu, T.K., Chait, B.T., and Roeder, R.G. (2001). Human STAGA complex is a chromatin-acetylating transcription coactivator that interacts with pre-mRNA splicing and DNA damage-binding factors in vivo. Mol Cell Biol *21*, 6782-6795.
Maston, G.A., Evans, S.K., and Green, M.R. (2006). Transcriptional regulatory elements in the human genome. Annual review of genomics and human genetics *7*, 29-59.
Matouk, I.J., Abbasi, I., Hochberg, A., Galun, E., Dweik, H., and Akkawi, M. (2009). Highly upregulated in liver cancer noncoding RNA is overexpressed in hepatic colorectal metastasis. Eur J Gastroen Hepat *21*, 688-692.
Matouk, I.J., DeGroot, N., Mezan, S., Ayesh, S., Abu-Iail, R., Hochberg, A., and Galun, E. (2007). The H19 Non-Coding RNA Is Essential for Human Tumor Growth. PloS one *2*.
Mattick, J.S. (2001). Non-coding RNAs: the architects of eukaryotic complexity. EMBO reports *2*, 986-991.
Mattick, J.S. (2004a). The hidden genetic program of complex organisms. Scientific American *291*, 60-67.
Mattick, J.S. (2004b). RNA regulation: a new genetics? Nat Rev Genet *5*, 316-323.
Mattick, J.S. (2009). The genetic signatures of noncoding RNAs. Plos Genet *5*, e1000459.
Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., Brody, J.*, et al.* (2012). Systematic localization of common disease-associated variation in regulatory DNA. Science *337*, 1190-1195.
Maziere, P., and Enright, A.J. (2007). Prediction of microRNA targets. Drug Discov Today *12*, 452-458.
McCullough, S.D., Xu, X.J., Dent, S.Y.R., Bekiranov, S., Roeder, R.G., and Grant, P.A. (2012). Reelin is a target of polyglutamine expanded ataxin-7 in human spinocerebellar ataxia type 7 (SCA7) astrocytes. Proceedings of the National Academy of Sciences of the United States of America *109*, 21319-21324.
McLaughlin, J., Cheng, D., Singer, O., Lukacs, R.U., Radu, C.G., Verma, I.M., and Witte, O.N. (2008). Sustained suppression of Bcr-Abl-driven lymphoid leukemia by

microRNA mimics (vol 104, pg 20501, 2007). Proc Natl Acad Sci U S A *105*, 1774-1774.
McMahon, S.J., Pray-Grant, M.G., Schieltz, D., Yates, J.R., and Grant, P.A. (2005). Polyglutamine-expanded spinocerebellar ataxia-7 protein disrupts normal SAGA and SLIK histone acetyltransferase activity. Proc Natl Acad Sci U S A *102*, 8478-8482.
Meader, S., Ponting, C.P., and Lunter, G. (2010). Massive turnover of functional sequence in human and other mammalian genomes. Genome Res *20*, 1335-1343.
Meister, G., Landthaler, M., Patkaniowska, A., Dorsett, Y., Teng, G., and Tuschl, T. (2004). Human Argonaute2 mediates RNA cleavage targeted by miRNAs and siRNAs. Mol Cell *15*, 185-197.
Melo, C.A., Drost, J., Wijchers, P.J., van de Werken, H., de Wit, E., Oude Vrielink, J.A., Elkon, R., Melo, S.A., Leveille, N., Kalluri, R.*, et al.* (2013). eRNAs are required for p53-dependent enhancer activity and gene transcription. Mol Cell *49*, 524-535.
Melton, C., Judson, R.L., and Blelloch, R. (2010). Opposing microRNA families regulate self-renewal in mouse embryonic stem cells. Nature *463*, 621-626.
Memczak, S., Jens, M., Elefsinioti, A., Torti, F., Krueger, J., Rybak, A., Maier, L., Mackowiak, S.D., Gregersen, L.H., Munschauer, M.*, et al.* (2013). Circular RNAs are a large class of animal RNAs with regulatory potency. Nature *495*, 333-338.
Meyer, L.R., Zweig, A.S., Hinrichs, A.S., Karolchik, D., Kuhn, R.M., Wong, M., Sloan, C.A., Rosenbloom, K.R., Roe, G., Rhead, B.*, et al.* (2012). The UCSC Genome Browser database: extensions and updates 2013. Nucleic acids research.
Mikkelsen, T.S., Ku, M., Jaffe, D.B., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W., Kim, T.K., Koche, R.P.*, et al.* (2007). Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature *448*, 553-560.
Miles, J.H. (2011). Autism spectrum disorders--a genetics review. Genetics in medicine : official journal of the American College of Medical Genetics *13*, 278-294.
Min, H., and Yoon, S. (2010). Got target? Computational methods for microRNA target prediction and their extension. Experimental & molecular medicine *42*, 233-244.
Miranda, K.C., Huynh, T., Tay, Y., Ang, Y.S., Tam, W.L., Thomson, A.M., Lim, B., and Rigoutsos, I. (2006). A pattern-based method for the identification of microRNA binding sites and their corresponding heteroduplexes. Cell *126*, 1203-1217.
Miska, E.A., Alvarez-Saavedra, E., Abbott, A.L., Lau, N.C., Hellman, A.B., McGonagle, S.M., Bartel, D.P., Ambros, V.R., and Horvitz, H.R. (2007). Most Caenorhabditis elegans microRNAs are individually not essential for development or viability. Plos Genet *3*, 2395-2403.
Miska, E.A., Alvarez-Saavedra, E., Townsend, M., Yoshii, A., Sestan, N., Rakic, P., Constantine-Paton, M., and Horvitz, H.R. (2004). Microarray analysis of microRNA expression in the developing mammalian brain. Genome Biol *5*, R68.
Monk, M., and Harper, M.I. (1979). Sequential X chromosome inactivation coupled with cellular differentiation in early mouse embryos. Nature *281*, 311-313.
Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods *5*, 621-628.
Moss, E.G. (2007). Heterochronic genes and the nature of developmental time. Curr Biol *17*, R425-R434.
Mousavi, K., Zare, H., Dell'orso, S., Grontved, L., Gutierrez-Cruz, G., Derfoul, A., Hager, G.L., and Sartorelli, V. (2013). eRNAs promote transcription by establishing chromatin accessibility at defined genomic loci. Mol Cell *51*, 606-617.
Mukherji, S., Ebert, M.S., Zheng, G.X.Y., Tsang, J.S., Sharp, P.A., and van Oudenaarden, A. (2011). MicroRNAs can generate thresholds in target gene expression. Nat Genet *43*, 854-U860.
Muro, E.M., Mah, N., and Andrade-Navarro, M.A. (2011). Functional evidence of post-transcriptional regulation by pseudogenes. Biochimie *93*, 1916-1921.
Murthy, U.M., and Rangarajan, P.N. (2010). Identification of protein interaction regions of VINC/NEAT1/Men epsilon RNA. Febs Lett *584*, 1531-1535.

Mus, E., Hof, P.R., and Tiedge, H. (2007). Dendritic BC200 RNA in aging and in Alzheimer's disease. Proc Natl Acad Sci U S A *104*, 10679-10684.
Musunuru, K., Strong, A., Frank-Kamenetsky, M., Lee, N.E., Ahfeldt, T., Sachs, K.V., Li, X., Li, H., Kuperwasser, N., Ruda, V.M.*, et al.* (2010). From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. Nature *466*, 714-719.
Myers, R.M., Stamatoyannopoulos, J., Snyder, M., Dunham, I., Hardison, R.C., Bernstein, B.E., Gingeras, T.R., Kent, W.J., Birney, E., Wold, B.*, et al.* (2011). A User's Guide to the Encyclopedia of DNA Elements (ENCODE). Plos Biol *9*.
Nakamachi, Y., Kawano, S., Takenokuchi, M., Nishimura, K., Sakai, Y., Chin, T., Saura, R., Kurosaka, M., and Kumagai, S. (2009). MicroRNA-124a is a key regulator of proliferation and monocyte chemoattractant protein 1 secretion in fibroblast-like synoviocytes from patients with rheumatoid arthritis. Arthritis and rheumatism *60*, 1294-1304.
Natoli, G., and Andrau, J.C. (2012). Noncoding transcription at enhancers: general principles and functional models. Annu Rev Genet *46*, 1-19.
Neale, B.M., Kou, Y., Liu, L., Ma'ayan, A., Samocha, K.E., Sabo, A., Lin, C.F., Stevens, C., Wang, L.S., Makarov, V.*, et al.* (2012). Patterns and rates of exonic de novo mutations in autism spectrum disorders. Nature *485*, 242-245.
Necsulea, A., Soumillon, M., Warnefors, M., Liechti, A., Daish, T., Zeller, U., Baker, J.C., Grutzner, F., and Kaessmann, H. (2014). The evolution of lncRNA repertoires and expression patterns in tetrapods. Nature *505*, 635-640.
Nelson, P.T., Baldwin, D.A., Scearce, L.M., Oberholtzer, J.C., Tobias, J.W., and Mourelatos, Z. (2004). Microarray-based, high-throughput gene expression profiling of microRNAs. Nat Methods *1*, 155-161.
Nesterova, T.B., Popova, B.C., Cobb, B.S., Norton, S., Senner, C.E., Tang, Y.A., Spruce, T., Rodriguez, T.A., Sado, T., Merkenschlager, M.*, et al.* (2008). Dicer regulates Xist promoter methylation in ES cells indirectly through transcriptional control of Dnmt3a. Epigenet Chromatin *1*.
Nie, Z.M., Zhou, F., Li, D., Lv, Z.B., Chen, J., Liu, Y., Shu, J.H., Sheng, Q., Yu, W., Zhang, W.P.*, et al.* (2013). RIP-seq of BmAgo2-associated small RNAs reveal various types of small non-coding RNAs in the silkworm, Bombyx mori. BMC genomics *14*.
Nilsen, T.W. (2007). Mechanisms of microRNA-mediated gene regulation in animal cells. Trends Genet *23*, 243-249.
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.
Nishino, J., Kim, I., Chada, K., and Morrison, S.J. (2008). Hmga2 promotes neural stem cell self-renewal in young but not old mice by reducing p16Ink4a and p19Arf Expression. Cell *135*, 227-239.
Ntini, E., Jarvelin, A.I., Bornholdt, J., Chen, Y., Boyd, M., Jorgensen, M., Andersson, R., Hoof, I., Schein, A., Andersen, P.R.*, et al.* (2013). Polyadenylation site-induced decay of upstream transcripts enforces promoter directionality. Nat Struct Mol Biol *20*, 923-928.
O'Carroll, D., Mecklenbrauker, I., Das, P.P., Santana, A., Koenig, U., Enright, A.J., Miska, E.A., and Tarakhovsky, A. (2007). A Slicer-independent role for Argonaute 2 in hematopoiesis and the microRNA pathway. Gene Dev *21*, 1999-2004.
O'Roak, B.J., Vives, L., Fu, W., Egertson, J.D., Stanaway, I.B., Phelps, I.G., Carvill, G., Kumar, A., Lee, C., Ankenman, K.*, et al.* (2012a). Multiplex targeted sequencing identifies recurrently mutated genes in autism spectrum disorders. Science *338*, 1619-1622.
O'Roak, B.J., Vives, L., Girirajan, S., Karakoc, E., Krumm, N., Coe, B.P., Levy, R., Ko, A., Lee, C., Smith, J.D.*, et al.* (2012b). Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. Nature *485*, 246-250.
Ohshima, K., Hattori, M., Yada, T., Gojobori, T., Sakaki, Y., and Okada, N. (2003). Whole-genome screening indicates a possible burst of formation of processed

pseudogenes and Alu repeats by particular L1 subfamilies in ancestral primates. Genome Biol *4*.
Okamura, K., Ishizuka, A., Siomi, H., and Siomi, M.C. (2004). Distinct roles for argonaute proteins in small RNA-directed RNA cleavage pathways. Gene Dev *18*, 1655-1666.
Organization, W.H. (2013). Autism spectrum disorders & other developmental disorders: From raising awareness to building capacity. (Geneva, Switzerland).
Orom, U.A., Derrien, T., Beringer, M., Gumireddy, K., Gardini, A., Bussotti, G., Lai, F., Zytnicki, M., Notredame, C., Huang, Q.*, et al.* (2010). Long noncoding RNAs with enhancer-like function in human cells. Cell *143*, 46-58.
Orrico, A., Galli, L., Buoni, S., Hayek, G., Luchetti, A., Lorenzini, S., Zappella, M., Pomponi, M.G., and Sorrentino, V. (2005). Attention-deficit/hyperactivity disorder (ADHD) and variable clinical expression of Aarskog-Scott syndrome due to a novel FGD1 gene mutation (R408Q). American journal of medical genetics Part A *135*, 99-102.
Osella, M., Bosia, C., Cora, D., and Caselle, M. (2011). The Role of Incoherent MicroRNA-Mediated Feedforward Loops in Noise Buffering. Plos Comput Biol *7*.
Ota, T., Suzuki, Y., Nishikawa, T., Otsuki, T., Sugiyama, T., Irie, R., Wakamatsu, A., Hayashi, K., Sato, H., Nagai, K.*, et al.* (2004). Complete sequencing and characterization of 21,243 full-length human cDNAs. Nat Genet *36*, 40-45.
Otaegi, G., Pollock, A., and Sun, T. (2011). An Optimized Sponge for microRNA miR-9 Affects Spinal Motor Neuron Development in vivo. Frontiers in neuroscience *5*, 146.
Packer, A.N., Xing, Y., Harper, S.Q., Jones, L., and Davidson, B.L. (2008). The bifunctional microRNA miR-9/miR-9* regulates REST and CoREST and is downregulated in Huntington's disease. J Neurosci *28*, 14341-14346.
Palhan, V.B., Chen, S., Peng, G.H., Tjernberg, A., Gamper, A.M., Fan, Y., Chait, B.T., La Spada, A.R., and Roeder, R.G. (2005). Polyglutamine-expanded ataxin-7 inhibits STAGA histone acetyltransferase activity to produce retinal degeneration. Proc Natl Acad Sci U S A *102*, 8472-8477.
Pang, K.C., Frith, M.C., and Mattick, J.S. (2006). Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function. Trends Genet *22*, 1-5.
Panzitt, K., Tschernatsch, M.M., Guelly, C., Moustafa, T., Stradner, M., Strohmaier, H.M., Buck, C.R., Denk, H., Schroeder, R., Trauner, M.*, et al.* (2007). Characterization of HULC, a novel gene with striking up-regulation in hepatocellular carcinoma, as noncoding RNA. Gastroenterology *132*, 330-342.
Paraskevopoulou, M.D., Georgakilas, G., Kostoulas, N., Reczko, M., Maragkakis, M., Dalamagas, T.M., and Hatzigeorgiou, A.G. (2013). DIANA-LncBase: experimentally verified and computationally predicted microRNA targets on long non-coding RNAs. Nucleic Acids Res *41*, D239-D245.
Parker, R., and Song, H. (2004). The enzymes and control of eukaryotic mRNA turnover. Nat Struct Mol Biol *11*, 121-127.
Pasquinelli, A.E. (2012). NON-CODING RNA MicroRNAs and their targets: recognition, regulation and an emerging reciprocal relationship. Nat Rev Genet *13*, 271-282.
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.
Pauli, A., Valen, E., Lin, M.F., Garber, M., Vastenhouw, N.L., Levin, J.Z., Fan, L., Sandelin, A., Rinn, J.L., Regev, A.*, et al.* (2012). Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. Genome Res *22*, 577-591.
Penna, E., Orso, F., Cimino, D., Tenaglia, E., Lembo, A., Quaglino, E., Poliseno, L., Haimovic, A., Osella, S., De Pitta, C.*, et al.* (2011). microRNA-214 contributes to melanoma tumor progression through suppression of TFAP2C. Febs Journal *278*, 228-228.

Penny, G.D., Kay, G.F., Sheardown, S.A., Rastan, S., and Brockdorff, N. (1996). Requirement for Xist in X chromosome inactivation. Nature *379*, 131-137.
Peter, M.E. (2009). Let-7 and miR-200 microRNAs: guardians against pluripotency and cancer progression. Cell Cycle *8*, 843-852.
Peters, L., and Meister, G. (2007). Argonaute proteins: mediators of RNA silencing. Mol Cell *26*, 611-623.
Pillai, R.S., Bhattacharyya, S.N., and Filipowicz, W. (2007). Repression of protein synthesis by miRNAs: how many mechanisms? Trends Cell Biol *17*, 118-126.
Pink, R.C., Wicks, K., Caley, D.P., Punch, E.K., Jacobs, L., and Carter, D.R. (2011). Pseudogenes: pseudo-functional or key regulators in health and disease? RNA *17*, 792-798.
Piven, J., Palmer, P., Jacobi, D., Childress, D., and Arndt, S. (1997). Broader autism phenotype: evidence from a family history study of multiple-incidence autism families. The American journal of psychiatry *154*, 185-190.
Poliseno, L., Salmena, L., Zhang, J., Carver, B., Haveman, W.J., and Pandolfi, P.P. (2010). A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. Nature *465*, 1033-1038.
Politz, J.C.R., Zhang, F., and Pederson, T. (2006). MicroRNA-206 colocalizes with ribosome-rich regions in both the nucleolus and cytoplasm of rat myogenic cells. P Natl Acad Sci USA *103*, 18957-18962.
Ponjavic, J., Ponting, C.P., and Lunter, G. (2007). Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. Genome Res *17*, 556-565.
Ponting, C.P., and Belgard, T.G. (2010). Transcribed dark matter: meaning or myth? Human molecular genetics *19*, R162-168.
Ponting, C.P., Oliver, P.L., and Reik, W. (2009). Evolution and functions of long noncoding RNAs. Cell *136*, 629-641.
Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841-842.
Qureshi, I.A., Mattick, J.S., and Mehler, M.F. (2010). Long non-coding RNAs in nervous system function and disease. Brain Res *1338*, 20-35.
R, D.C.T. (2011). R: A Language and Environment for Statistical Computing. (Vienna, Austria.).
Rachmilewitz, J., Goshen, R., Ariel, I., Schneider, T., de Groot, N., and Hochberg, A. (1992). Parental imprinting of the human H19 gene. Febs Lett *309*, 25-28.
Rajakulendran, S., Roberts, J., Koltzenburg, M., Hanna, M.G., and Stewart, H. (2013). Deletion of chromosome 12q21 affecting KCNC2 and ATXN7L3B in a family with neurodevelopmental delay and ataxia. Journal of neurology, neurosurgery, and psychiatry *84*, 1255-1257.
Rajewsky, N., and Socci, N.D. (2004). Computational identification of microRNA targets. Developmental biology *267*, 529-535.
Ramachandran, V., and Chen, X. (2008). Degradation of microRNAs by a family of exoribonucleases in Arabidopsis. Science *321*, 1490-1492.
Rands, C.M., Meader, S., Ponting, C.P., and Lunter, G. (2014). 8.2% of the Human Genome Is Constrained: Variation in Rates of Turnover across Functional Element Classes in the Human Lineage. Plos Genet *10*, e1004525.
Ravasi, T., Suzuki, H., Pang, K.C., Katayama, S., Furuno, M., Okunishi, R., Fukuda, S., Ru, K., Frith, M.C., Gongora, M.M.*, et al.* (2006). Experimental validation of the regulated expression of large numbers of non-coding RNAs from the mouse genome. Genome Res *16*, 11-19.
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.
Rhoades, M.W., Reinhart, B.J., Lim, L.P., Burge, C.B., Bartel, B., and Bartel, D.P. (2002). Prediction of plant microRNA targets. Cell *110*, 513-520.

Rigoutsos, I. (2009). New Tricks for Animal MicroRNAs: Targeting of Amino Acid Coding Regions at Conserved and Nonconserved Sites. Cancer Res *69*, 3245-3248.
Rinn, J., and Guttman, M. (2014). RNA Function. RNA and dynamic nuclear organization. Science *345*, 1240-1241.
Rinn, J.L., and Chang, H.Y. (2012). Genome regulation by long noncoding RNAs. Annual review of biochemistry *81*, 145-166.
Rinn, J.L., Euskirchen, G., Bertone, P., Martone, R., Luscombe, N.M., Hartman, S., Harrison, P.M., Nelson, F.K., Miller, P., Gerstein, M.*, et al.* (2003). The transcriptional activity of human Chromosome 22. Genes & development *17*, 529-540.
Rinn, J.L., Kertesz, M., Wang, J.K., Squazzo, S.L., Xu, X., Brugmann, S.A., Goodnough, L.H., Helms, J.A., Farnham, P.J., Segal, E.*, et al.* (2007). Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. Cell *129*, 1311-1323.
Robb, G.B., Brown, K.M., Khurana, J., and Rana, T.M. (2005). Specific and potent RNAi in the nucleus of human cells. Nat Struct Mol Biol *12*, 133-137.
Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics *26*, 139-140.
Rodriguez, A., Griffiths-Jones, S., Ashurst, J.L., and Bradley, A. (2004). Identification of mammalian microRNA host genes and transcription units. Genome Res *14*, 1902-1910.
Ronald, A., and Hoekstra, R.A. (2011). Autism spectrum disorders and autistic traits: a decade of new twin studies. American journal of medical genetics Part B, Neuropsychiatric genetics : the official publication of the International Society of Psychiatric Genetics *156B*, 255-274.
Rougvie, A.E. (2005). Intrinsic and extrinsic regulators of developmental timing: from miRNAs to nutritional cues. Development *132*, 3787-3798.
Roy, S., Ernst, J., Kharchenko, P.V., Kheradpour, P., Negre, N., Eaton, M.L., Landolin, J.M., Bristow, C.A., Ma, L., Lin, M.F.*, et al.* (2010). Identification of functional elements and regulatory circuits by Drosophila modENCODE. Science *330*, 1787-1797.
Ruegger, S., and Grosshans, H. (2012). MicroRNA turnover: when, how, and why. Trends in biochemical sciences *37*, 436-446.
Rutter, M. (2000). Genetic studies of autism: from the 1970s into the millennium. Journal of abnormal child psychology *28*, 3-14.
Saetrom, O., Snove, O., and Saetrom, P. (2005). Weighted sequence motifs as an improved seeding step in microRNA target prediction algorithms. Rna *11*, 995-1003.
Saini, H.K., Griffiths-Jones, S., and Enright, A.J. (2007). Genomic analysis of human microRNA transcripts. P Natl Acad Sci USA *104*, 17719-17724.
Salmena, L., Poliseno, L., Tay, Y., Kats, L., and Pandolfi, P.P. (2011). A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language? Cell *146*, 353-358.
Salyakina, D., Ma, D.Q., Jaworski, J.M., Konidari, I., Whitehead, P.L., Henson, R., Martinez, D., Robinson, J.L., Sacharow, S., Wright, H.H.*, et al.* (2010). Variants in several genomic regions associated with asperger disorder. Autism research : official journal of the International Society for Autism Research *3*, 303-310.
Salzman, J., Gawad, C., Wang, P.L., Lacayo, N., and Brown, P.O. (2012). Circular RNAs Are the Predominant Transcript Isoform from Hundreds of Human Genes in Diverse Cell Types. PloS one *7*.
Sanders, S.J., Murtha, M.T., Gupta, A.R., Murdoch, J.D., Raubeson, M.J., Willsey, A.J., Ercan-Sencicek, A.G., DiLullo, N.M., Parikshak, N.N., Stein, J.L.*, et al.* (2012). De novo mutations revealed by whole-exome sequencing are strongly associated with autism. Nature *485*, 237-241.
Sanuki, R., Onishi, A., Koike, C., Muramatsu, R., Watanabe, S., Muranishi, Y., Irie, S., Uneo, S., Koyasu, T., Matsui, R.*, et al.* (2011). miR-124a is required for hippocampal axogenesis and retinal cone survival through Lhx2 suppression. Nat Neurosci *14*, 1125-U1177.

Sayed, D., and Abdellatif, M. (2011). MicroRNAs in development and disease. Physiol Rev *91*, 827-887.
Schaefer, A., O'Carroll, D., Tan, C.L., Hillman, D., Sugimori, M., Llinas, R., and Greengard, P. (2007). Cerebellar neuro degeneration in the absence of microRNAs. J Exp Med *204*, 1553-1558.
Schapira, A.H. (2009). Neurobiology and treatment of Parkinson's disease. Trends in pharmacological sciences *30*, 41-47.
Schwartz, S., Kent, W.J., Smit, A., Zhang, Z., Baertsch, R., Hardison, R.C., Haussler, D., and Miller, W. (2003). Human-mouse alignments with BLASTZ. Genome Res *13*, 103-107.
Schwarz, D.S., Hutvagner, G., Du, T., Xu, Z., Aronin, N., and Zamore, P.D. (2003). Asymmetry in the assembly of the RNAi enzyme complex. Cell *115*, 199-208.
Seitz, H. (2009). Redefining microRNA targets. Curr Biol *19*, 870-873.
Selbach, M., Schwanhausser, B., Thierfelder, N., Fang, Z., Khanin, R., and Rajewsky, N. (2008). Widespread changes in protein synthesis induced by microRNAs. Nature *455*, 58-63.
Sempere, L.F., Cole, C.N., McPeek, M.A., and Peterson, K.J. (2006). The phylogenetic distribution of metazoan microRNAs: Insights into evolutionary complexity and constraint. J Exp Zool Part B *306B*, 575-588.
Sethupathy, P., Megraw, M., and Hatzigeorgiou, A.G. (2006). A guide through present computational approaches for the identification of mammalian microRNA targets. Nat Methods *3*, 881-886.
Sevignani, C., Calin, G.A., Siracusa, L.D., and Croce, C.M. (2006). Mammalian microRNAs: a small world for fine-tuning gene expression. Mamm Genome *17*, 189-202.
Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res *13*, 2498-2504.
Shen, Y., Yue, F., McCleary, D.F., Ye, Z., Edsall, L., Kuan, S., Wagner, U., Dixon, J., Lee, L., Lobanenkov, V.V.*, et al.* (2012). A map of the cis-regulatory sequences in the mouse genome. Nature *488*, 116-120.
Shi, X.B., Xue, L., Ma, A.H., Tepper, C.G., Gandour-Edwards, R., Kung, H.J., and White, R.W.D. (2013). Tumor suppressive miR-124 targets androgen receptor and inhibits proliferation of prostate cancer cells. Oncogene *32*, 4130-4138.
Sigova, A.A., Mullen, A.C., Molinie, B., Gupta, S., Orlando, D.A., Guenther, M.G., Almada, A.E., Lin, C., Sharp, P.A., Giallourakis, C.C.*, et al.* (2013). Divergent transcription of long noncoding RNA/mRNA gene pairs in embryonic stem cells. Proc Natl Acad Sci U S A *110*, 2876-2881.
Silber, J., Lim, D.A., Petritsch, C., Persson, A.I., Maunakea, A.K., Yu, M., Vandenberg, S.R., Ginzinger, D.G., James, D., Costello, J.F.*, et al.* (2008). miR-124 and miR-137 inhibit proliferation of glioblastoma multiforme cells and induce differentiation of brain tumor stem cells. Bmc Med *6*.
Singh, A.M., and Dalton, S. (2009). The cell cycle and Myc intersect with mechanisms that regulate pluripotency and reprogramming. Cell Stem Cell *5*, 141-149.
Slater, G.S., and Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. BMC bioinformatics *6*, 31.
Smalley, S.L. (1998). Autism and tuberous sclerosis. Journal of autism and developmental disorders *28*, 407-414.
Smalley, S.L., Asarnow, R.F., and Spence, M.A. (1988). Autism and genetics. A decade of research. Archives of general psychiatry *45*, 953-961.
Sobell, H.M. (1985). Actinomycin and DNA transcription. Proc Natl Acad Sci U S A *82*, 5328-5331.
Sopher, B.L., Ladd, P.D., Pineda, V.V., Libby, R.T., Sunkin, S.M., Hurley, J.B., Thienes, C.P., Gaasterland, T., Filippova, G.N., and La Spada, A.R. (2011). CTCF

Regulates Ataxin-7 Expression through Promotion of a Convergently Transcribed, Antisense Noncoding RNA. Neuron *70*, 1071-1084.
Sotiropoulou, G., Pampalakis, G., Lianidou, E., and Mourelatos, Z. (2009). Emerging roles of microRNAs as molecular switches in the integrated circuit of the cancer cell. Rna *15*, 1443-1461.
St Pourcain, B., Wang, K., Glessner, J.T., Golding, J., Steer, C., Ring, S.M., Skuse, D.H., Grant, S.F., Hakonarson, H., and Davey Smith, G. (2010). Association between a high-risk autism locus on 5p14 and social communication spectrum phenotypes in the general population. The American journal of psychiatry *167*, 1364-1372.
Stark, A., Brennecke, J., Bushati, N., Russell, R.B., and Cohen, S.M. (2005). Animal microRNAs confer robustness to gene expression and have a significant impact on 3 ' UTR evolution. Cell *123*, 1133-1146.
Stark, A., Brennecke, J., Russell, R.B., and Cohen, S.M. (2003). Identification of Drosophila MicroRNA targets. Plos Biol *1*, 397-409.
Struhl, K. (2007). Transcriptional noise and the fidelity of initiation by RNA polymerase II. Nat Struct Mol Biol *14*, 103-105.
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. Proc Natl Acad Sci U S A *101*, 6062-6067.
Su, X.Q., Xing, J.D., Wang, Z.Z., Chen, L., Cui, M., and Jiang, B.H. (2013). microRNAs and ceRNAs: RNA networks in pathogenesis of cancer. Chinese J Cancer Res *25*, 235-239.
Sumazin, P., Yang, X., Chiu, H.S., Chung, W.J., Iyer, A., Llobet-Navas, D., Rajbhandari, P., Bansal, M., Guarnieri, P., Silva, J.*, et al.* (2011). An extensive microRNA-mediated network of RNA-RNA interactions regulates established oncogenic pathways in glioblastoma. Cell *147*, 370-381.
Sun, L., Goff, L.A., Trapnell, C., Alexander, R., Lo, K.A., Hacisuleyman, E., Sauvageau, M., Tazon-Vega, B., Kelley, D.R., Hendrickson, D.G.*, et al.* (2013). Long noncoding RNAs regulate adipogenesis. Proc Natl Acad Sci U S A *110*, 3387-3392.
Supek, F., Bosnjak, M., Skunca, N., and Smuc, T. (2011). REVIGO summarizes and visualizes long lists of gene ontology terms. PloS one *6*, e21800.
Szatmari, P. (1999). Heterogeneity and the genetics of autism. Journal of psychiatry & neuroscience : JPN *24*, 159-165.
Talbot, K. (2007). The study of rare diseases: butterfly collecting or an entree to understanding common conditions? Practical neurology *7*, 210-211.
Tan, J.Y., and Marques, A.C. (2014). The miRNA-Mediated Cross-Talk between Transcripts Provides a Novel Layer of Posttranscriptional Regulation. Advances in genetics *85*, 149-199.
Tan, L.P., Seinen, E., Duns, G., de Jong, D., Sibon, O.C.M., Poppema, S., Kroesen, B.J., Kok, K., and van den Berg, A. (2009). A high throughput experimental approach to identify miRNA targets in human cells. Nucleic acids research *37*.
Tan, M.H., Au, K.F., Yablonovitch, A.L., Wills, A.E., Chuang, J., Baker, J.C., Wong, W.H., and Li, J.B. (2013). RNA sequencing reveals a diverse and dynamic repertoire of the Xenopus tropicalis transcriptome over development. Genome Res *23*, 201-216.
Tang, R., Li, L.M., Zhu, D.H., Hou, D.X., Cao, T., Gu, H.W., Zhang, J., Chen, J.Y., Zhang, C.Y., and Zen, K. (2012). Mouse miRNA-709 directly regulates miRNA-15a/16-1 biogenesis at the posttranscriptional level in the nucleus: evidence for a microRNA hierarchy system. Cell Res *22*, 504-515.
Tani, H., Mizutani, R., Salam, K.A., Tano, K., Ijiri, K., Wakamatsu, A., Isogai, T., Suzuki, Y., and Akimitsu, N. (2012). Genome-wide determination of RNA stability reveals hundreds of short-lived noncoding transcripts in mammals. Genome Res *22*, 947-956.
Tay, Y., Kats, L., Salmena, L., Weiss, D., Tan, S.M., Ala, U., Karreth, F., Poliseno, L., Provero, P., Di Cunto, F.*, et al.* (2011). Coding-independent regulation of the tumor suppressor PTEN by competing endogenous mRNAs. Cell *147*, 344-357.

Tay, Y., Rinn, J., and Pandolfi, P.P. (2014). The multilayered complexity of ceRNA crosstalk and competition. Nature *505*, 344-352.
Thomson, D.W., Bracken, C.P., and Goodall, G.J. (2011). Experimental strategies for microRNA target identification. Nucleic acids research *39*, 6845-6853.
Thomson, J.M., Parker, J., Perou, C.M., and Hammond, S.M. (2004). A custom microarray platform for analysis of microRNA gene expression. Nat Methods *1*, 47-53.
Thummel, C.S. (2001). Molecular mechanisms of developmental timing in C-elegans and Drosophila. Dev Cell *1*, 453-465.
Tilgner, H., Knowles, D.G., Johnson, R., Davis, C.A., Chakrabortty, S., Djebali, S., Curado, J., Snyder, M., Gingeras, T.R., and Guigo, R. (2012). Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. Genome Res *22*, 1616-1625.
Torrents, D., Suyama, M., Zdobnov, E., and Bork, P. (2003). A genome-wide survey of human pseudogenes. Genome Res *13*, 2559-2567.
Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics *25*, 1105-1111.
Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature biotechnology *28*, 511-515.
Tripathi, V., Ellis, J.D., Shen, Z., Song, D.Y., Pan, Q., Watt, A.T., Freier, S.M., Bennett, C.F., Sharma, A., Bubulya, P.A.*, et al.* (2010). The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation. Mol Cell *39*, 925-938.
Tsai, M.C., Manor, O., Wan, Y., Mosammaparast, N., Wang, J.K., Lan, F., Shi, Y., Segal, E., and Chang, H.Y. (2010). Long noncoding RNA as modular scaffold of histone modification complexes. Science *329*, 689-693.
Ule, J., Jensen, K.B., Ruggiu, M., Mele, A., Ule, A., and Darnell, R.B. (2003). CLIP identifies Nova-regulated RNA networks in the brain. Science *302*, 1212-1215.
Ulitsky, I., and Bartel, D.P. (2013). lincRNAs: genomics, evolution, and mechanisms. Cell *154*, 26-46.
Ulitsky, I., Shkumatava, A., Jan, C.H., Sive, H., and Bartel, D.P. (2011). Conserved Function of lincRNAs in Vertebrate Embryonic Development despite Rapid Sequence Evolution. Cell *147*, 1537-1550.
Valencia-Sanchez, M.A., Liu, J., Hannon, G.J., and Parker, R. (2006). Control of translation and mRNA degradation by miRNAs and siRNAs. Genes & development *20*, 515-524.
van Bakel, H., Nislow, C., Blencowe, B.J., and Hughes, T.R. (2010). Most "dark matter" transcripts are associated with known genes. Plos Biol *8*, e1000371.
van Dongen, S., Abreu-Goodger, C., and Enright, A.J. (2008). Detecting microRNA binding and siRNA off-target effects from expression data. Nat Methods *5*, 1023-1025.
van Heesch, S., van Iterson, M., Jacobi, J., Boymans, S., Essers, P.B., de Bruijn, E., Hao, W., Macinnes, A.W., Cuppen, E., and Simonis, M. (2014). Extensive localization of long noncoding RNAs to the cytosol and mono- and polyribosomal complexes. Genome Biol *15*, R6.
van Rooij, E., Sutherland, L.B., Qi, X.X., Richardson, J.A., Hill, J., and Olson, E.N. (2007). Control of stress-dependent cardiac growth and gene expression by a microRNA. Science *316*, 575-579.
Vance, K.W., Sansom, S.N., Lee, S., Chalei, V., Kong, L., Cooper, S.E., Oliver, P.L., and Ponting, C.P. (2014). The long non-coding RNA Paupar regulates the expression of both local and distal genes. Embo J *33*, 296-311.
Vanin, E.F., Goldberg, G.I., Tucker, P.W., and Smithies, O. (1980). A mouse alpha-globin-related pseudogene lacking intervening sequences. Nature *286*, 222-226.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A.*, et al.* (2001). The sequence of the human genome. Science *291*, 1304-1351.
Visel, A., Rubin, E.M., and Pennacchio, L.A. (2009). Genomic views of distant-acting enhancers. Nature *461*, 199-205.
Visvanathan, J., Lee, S., Lee, B., Lee, J.W., and Lee, S.K. (2007). The microRNA miR-124 antagonizes the anti-neural REST/SCP1 pathway during embryonic CNS development. Genes Dev *21*, 744-749.
Voineagu, I., Wang, X., Johnston, P., Lowe, J.K., Tian, Y., Horvath, S., Mill, J., Cantor, R.M., Blencowe, B.J., and Geschwind, D.H. (2011). Transcriptomic analysis of autistic brain reveals convergent molecular pathology. Nature *474*, 380-384.
Voinnet, O. (2009). Origin, biogenesis, and activity of plant microRNAs. Cell *136*, 669-687.
Wakiyama, M., Takimoto, K., Ohara, O., and Yokoyama, S. (2007). Let-7 microRNA-mediated mRNA deadenylation and translational repression in a mammalian cell-free system. Gene Dev *21*, 1857-1862.
Wang, J., Zhang, J., Zheng, H., Li, J., Liu, D., Li, H., Samudrala, R., Yu, J., and Wong, G.K. (2004). Mouse transcriptome: neutral evolution of 'non-coding' complementary DNAs. Nature *431*, 1 p following 757; discussion following 757.
Wang, J.Y., Liu, X.F., Wu, H.C., Ni, P.H., Gu, Z.D., Qiao, Y.X., Chen, N., Sun, F.Y., and Fan, Q.S. (2010a). CREB up-regulates long non-coding RNA, HULC expression through interaction with microRNA-372 in liver cancer. Nucleic Acids Res *38*, 5366-5383.
Wang, K., Zhang, H., Ma, D., Bucan, M., Glessner, J.T., Abrahams, B.S., Salyakina, D., Imielinski, M., Bradfield, J.P., Sleiman, P.M.*, et al.* (2009a). Common genetic variants on 5p14.1 associate with autism spectrum disorders. Nature *459*, 528-533.
Wang, K.C., and Chang, H.Y. (2011). Molecular mechanisms of long noncoding RNAs. Mol Cell *43*, 904-914.
Wang, K.C., Yang, Y.W., Liu, B., Sanyal, A., Corces-Zimmerman, R., Chen, Y., Lajoie, B.R., Protacio, A., Flynn, R.A., Gupta, R.A.*, et al.* (2011). A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. Nature *472*, 120-124.
Wang, W.X., Wilfred, B.R., Hu, Y.L., Stromberg, A.J., and Nelson, P.T. (2010b). Anti-Argonaute RIP-Chip shows that miRNA transfections alter global patterns of mRNA recruitment to microribonucleoprotein complexes. Rna *16*, 394-404.
Wang, W.X., Wilfred, B.R., Xie, K., Jennings, M.H., Hu, Y.L., Stromberg, A.J., and Nelson, P.T. (2010c). Individual microRNAs (miRNAs) display distinct mRNA targeting "rules". Rna Biol *7*, 373-380.
Wang, X., Arai, S., Song, X., Reichart, D., Du, K., Pascual, G., Tempst, P., Rosenfeld, M.G., Glass, C.K., and Kurokawa, R. (2008). Induced ncRNAs allosterically modify RNA-binding proteins in cis to inhibit transcription. Nature *454*, 126-130.
Wang, Y., Xu, Z., Jiang, J., Xu, C., Kang, J., Xiao, L., Wu, M., Xiong, J., Guo, X., and Liu, H. (2013). Endogenous miRNA sponge lincRNA-RoR regulates Oct4, Nanog, and Sox2 in human embryonic stem cell self-renewal. Dev Cell *25*, 69-80.
Wang, Z., Tollervey, J., Briese, M., Turner, D., and Ule, J. (2009b). CLIP: Construction of cDNA libraries for high-throughput sequencing from RNAs cross-linked to proteins in vivo. Methods *48*, 287-293.
Washietl, S., Kellis, M., and Garber, M. (2014). Evolutionary dynamics and tissue specificity of human long noncoding RNAs in six mammals. Genome Res *24*, 616-628.
Watanabe, Y., Tomita, M., and Kanai, A. (2007). Computational methods for microRNA target prediction. Methods in enzymology *427*, 65-86.
Weaving, L.S., Christodoulou, J., Williamson, S.L., Friend, K.L., McKenzie, O.L., Archer, H., Evans, J., Clarke, A., Pelka, G.J., Tam, P.P.*, et al.* (2004). Mutations of CDKL5 cause a severe neurodevelopmental disorder with infantile spasms and mental retardation. Am J Hum Genet *75*, 1079-1093.

Wehrspaun, C.C., Ponting, C.P., and Marques, A.C. (2014). Brain-expressed 3'UTR extensions strengthen miRNA cross-talk between ion channel/transporter encoding mRNAs. Frontiers in genetics *5*, 41.
Weinmann, L., Hock, J., Ivacevic, T., Ohrt, T., Mutze, J., Schwille, P., Kremmer, E., Benes, V., Urlaub, H., and Meister, G. (2009). Importin 8 Is a Gene Silencing Factor that Targets Argonaute Proteins to Distinct mRNAs. Cell *136*, 496-507.
Weiss, L.A., Arking, D.E., Gene Discovery Project of Johns, H., the Autism, C., Daly, M.J., and Chakravarti, A. (2009). A genome-wide linkage and association scan reveals novel loci for autism. Nature *461*, 802-808.
White, E., Schlackow, M., Kamieniarz-Gdula, K., Proudfoot, N.J., and Gullerova, M. (2014). Human nuclear Dicer restricts the deleterious accumulation of endogenous double-stranded RNA. Nat Struct Mol Biol *21*, 552-559.
Wienholds, E., Koudijs, M.J., van Eeden, F.J.M., Cuppen, E., and Plasterk, R.H.A. (2003). The microRNA-producing enzyme Dicer1 is essential for zebrafish development. Nat Genet *35*, 217-218.
Wienholds, E., and Plasterk, R.H. (2005). MicroRNA function in animal development. Febs Lett *579*, 5911-5922.
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.
Willingham, A.T., Orth, A.P., Batalov, S., Peters, E.C., Wen, B.G., Aza-Blanc, P., Hogenesch, J.B., and Schultz, P.G. (2005). A strategy for probing the function of noncoding RNAs finds a repressor of NFAT. Science *309*, 1570-1573.
Wilson, R., Ainscough, R., Anderson, K., Baynes, C., Berks, M., Bonfield, J., Burton, J., Connell, M., Copsey, T., Cooper, J.*, et al.* (1994). 2.2 Mb of contiguous nucleotide sequence from chromosome III of C. elegans. Nature *368*, 32-38.
Winter, J., Jung, S., Keller, S., Gregory, R.I., and Diederichs, S. (2009). Many roads to maturity: microRNA biogenesis pathways and their regulation. Nat Cell Biol *11*, 228-234.
Witkos, T.M., Koscianska, E., and Krzyzosiak, W.J. (2011). Practical Aspects of microRNA Target Prediction. Curr Mol Med *11*, 93-109.
Wu, C.I., Shen, Y., and Tang, T. (2009). Evolution under canalization and the dual roles of microRNAs-A hypothesis. Genome Res *19*, 734-743.
Wu, L.G., Fan, J.H., and Belasco, J.G. (2006). MicroRNAs direct rapid deadenylation of mRNA. P Natl Acad Sci USA *103*, 4034-4039.
Wyman, S.K., Knouf, E.C., Parkin, R.K., Fritz, B.R., Lin, D.W., Dennis, L.M., Krouse, M.A., Webster, P.J., and Tewari, M. (2011). Post-transcriptional generation of miRNA variants by multiple nucleotidyl transferases contributes to miRNA transcriptome complexity. Genome Res *21*, 1450-1461.
Xia, H., Cheung, W.K., Ng, S.S., Jiang, X., Jiang, S., Sze, J., Leung, G.K., Lu, G., Chan, D.T., Bian, X.W.*, et al.* (2012). Loss of brain-enriched miR-124 microRNA enhances stem-like traits and invasiveness of glioma cells. J Biol Chem *287*, 9962-9971.
Xia, H., Mao, Q., Paulson, H.L., and Davidson, B.L. (2002). siRNA-mediated gene silencing in vitro and in vivo. Nature biotechnology *20*, 1006-1010.
Xiao, J., Yang, B., Lin, H., Lu, Y., Luo, X., and Wang, Z. (2012). Novel approaches for gene-specific interference via manipulating actions of microRNAs: Examination on the pacemaker channel genes HCN2 and HCN4 (Retraction of 212, pg 285, 2007). J Cell Physiol *227*, 877-877.
Xu, N., Papagiannakopoulos, T., Pan, G.J., Thomson, J.A., and Kosik, K.S. (2009). MicroRNA-145 Regulates OCT4, SOX2, and KLF4 and Represses Pluripotency in Human Embryonic Stem Cells. Cell *137*, 647-658.
Yan, G.R., Xu, S.H., Tan, Z.L., Liu, L., and He, Q.Y. (2011). Global identification of miR-373-regulated genes in breast cancer by quantitative proteomics. Proteomics *11*, 912-920.

Yang, J.H., Li, J.H., Shao, P., Zhou, H., Chen, Y.Q., and Qu, L.H. (2011). starBase: a database for exploring microRNA-mRNA interaction maps from Argonaute CLIP-Seq and Degradome-Seq data. Nucleic acids research *39*, D202-D209.
Yang, Y., Chaerkady, R., Kandasamy, K., Huang, T.C., Selvan, L.D.N., Dwivedi, S.B., Kent, O.A., Mendell, J.T., and Pandey, A. (2010). Identifying targets of miR-143 using a SILAC-based proteomic approach. Mol Biosyst *6*, 1873-1882.
Yang, Z. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. Computer applications in the biosciences : CABIOS *13*, 555-556.
Yoo, A.S., Staahl, B.T., Chen, L., and Crabtree, G.R. (2009). MicroRNA-mediated switching of chromatin-remodelling complexes in neural development. Nature *460*, 642-646.
Yoo, S.Y., Pennesi, M.E., Weeber, E.J., Xu, B., Atkinson, R., Chen, S., Armstrong, D.L., Wu, S.M., Sweatt, J.D., and Zoghbi, H.Y. (2003). SCA7 knockin mice model human SCA7 and reveal gradual accumulation of mutant ataxin-7 in neurons and abnormalities in short-term plasticity. Neuron *37*, 383-401.
Yoon, J.H., Abdelmohsen, K., Srikantan, S., Yang, X., Martindale, J.L., De, S., Huarte, M., Zhan, M., Becker, K.G., and Gorospe, M. (2012). LincRNA-p21 suppresses target mRNA translation. Mol Cell *47*, 648-655.
Yoshimizu, T., Miroglio, A., Ripoche, M.A., Gabory, A., Vernucci, M., Riccio, A., Colnot, S., Godard, C., Terris, B., Jammes, H.*, et al.* (2008). The H19 locus acts in vivo as a tumor suppressor. Proc Natl Acad Sci U S A *105*, 12417-12422.
Yuan, B., Latek, R., Hossbach, M., Tuschl, T., and Lewitter, F. (2004). siRNA Selection Server: an automated siRNA oligonucleotide prediction server. Nucleic acids research *32*, W130-134.
Zander, C., Takahashi, J., El Hachimi, K.H., Fujigasaki, H., Albanese, V., Lebre, A.S., Stevanin, G., Duyckaerts, C., and Brice, A. (2001). Similarities between spinocerebellar ataxia type 7 (SCA7) cell models and human brain: proteins recruited in inclusions and activation of caspase-3. Human molecular genetics *10*, 2569-2579.
Zeng, Y., and Cullen, B.R. (2003). Sequence requirements for micro RNA processing and function in human cells. RNA *9*, 112-123.
Zeng, Y., and Cullen, B.R. (2005). Efficient processing of primary microRNA hairpins by Drosha requires flanking nonstructured RNA sequences. J Biol Chem *280*, 27595-27603.
Zhang, A., Zhou, N.J., Huang, J.G., Liu, Q., Fukuda, K., Ma, D., Lu, Z.H., Bai, C.X., Watabe, K., and Mo, Y.Y. (2013). The human long non-coding RNA-RoR is a p53 repressor in response to DNA damage. Cell Res *23*, 340-350.
Zhang, H.M., Chen, H., Liu, W., Liu, H., Gong, J., Wang, H., and Guo, A.Y. (2012). AnimalTFDB: a comprehensive animal transcription factor database. Nucleic acids research *40*, D144-149.
Zhang, Y., and Tycko, B. (1992). Monoallelic expression of the human H19 gene. Nat Genet *1*, 40-44.
Zhang, Y., and Verbeek, F.J. (2010). Comparison and integration of target prediction algorithms for microRNA studies. Journal of integrative bioinformatics *7*.
Zhang, Z., Harrison, P.M., Liu, Y., and Gerstein, M. (2003a). Millions of years of evolution preserved: a comprehensive catalog of the processed pseudogenes in the human genome. Genome Res *13*, 2541-2558.
Zhang, Z.D., Frankish, A., Hunt, T., Harrow, J., and Gerstein, M. (2010). Identification and analysis of unitary pseudogenes: historic and contemporary gene losses in humans and other primates. Genome Biol *11*, R26.
Zhang, Z.L., Harrison, P.M., Liu, Y., and Gerstein, M. (2003b). Millions of years of evolution preserved: A comprehensive catalog of the processed pseudogenes in the human genome. Genome Res *13*, 2541-2558.
Zheng, D.Y., and Gerstein, M.B. (2007). The ambiguous boundary between genes and pseudogenes: the dead rise up, or do they? Trends Genet *23*, 219-224.

Zheng, D.Y., Zhang, Z.L., Harrison, P.M., Karro, J., Carriero, N., and Gerstein, M. (2005). Integrated pseudogene annotation for human chromosome 22: Evidence for transcription. Journal of molecular biology *349*, 27-45.
Zheng, G.X., Do, B.T., Webster, D.E., Khavari, P.A., and Chang, H.Y. (2014). Dicer-microRNA-Myc circuit promotes transcription of hundreds of long noncoding RNAs. Nat Struct Mol Biol.
Zheng, G.X.Y., Ravi, A., Calabrese, J.M., Medeiros, L.A., Kirak, O., Dennis, L.M., Jaenisch, R., Burge, C.B., and Sharp, P.A. (2011). A Latent Pro-Survival Function for the Mir-290-295 Cluster in Mouse Embryonic Stem Cells. Plos Genet *7*.
Zhu, J., Sanborn, J.Z., Diekhans, M., Lowe, C.B., Pringle, T.H., and Haussler, D. (2007). Comparative genomics search for losses of long-established genes on the human lineage. Plos Comput Biol *3*, e247.
Zhu, Q.B., Sun, W.Y., Okano, K., Chen, Y., Zhang, N., Maeda, T., and Palczewski, K. (2011). Sponge Transgenic Mouse Model Reveals Important Roles for the MicroRNA-183 (miR-183)/96/182 Cluster in Postmitotic Photoreceptors of the Retina. J Biol Chem *286*, 31749-31760.
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-U176.

CHAPTER 4 APPENDIX

Appendix Table A4.1 Expression levels of mESC-expressed miRNAs. Unique miRNAs were grouped into miRNA families.
* Only the top 25% most highly expressed miRNA families (red shading)

miRNA	seed	miRNA family	Normalized count	St. Dev.
mmu-miR-295	AAGUGCU	*miR-93/93a/105/106a/291a-3p/294/295/302abcde/372/373/428/519a/520be/520acd-3p/1378/1420ac*	66241.1	3875.7
mmu-miR-294	AAGUGCU	*miR-93/93a/105/106a/291a-3p/294/295/302abcde/372/373/428/519a/520be/520acd-3p/1378/1420ac*	19856.6	672.5
mmu-miR-293	GUGCCGC	*miR-293*	18749.7	1579.7
mmu-miR-720	UCUCGCU	*miR-720.m*	16455.5	2556.4
mmu-miR-292-3p	AAGUGCC	*miR-290-3p/292-3p/467a/1420b*	10372.4	611.9
mmu-miR-20a	AAAGUGC	*miR-17/17-5p/20ab/20b-5p/93/106ab/427/518a-3p/519d*	7935.7	687.2
mmu-miR-20b	AAAGUGC	*miR-17/17-5p/20ab/20b-5p/93/106ab/427/518a-3p/519d*	7935.7	687.2
mmu-miR-292-5p	CUCAAAC	*miR-290-5p/292-5p/371-5p/293*	4130.4	366.5
mmu-miR-291a-3p	AAGUGCU	*miR-93/93a/105/106a/291a-3p/294/295/302abcde/372/373/428/519a/520be/520acd-3p/1378/1420ac*	3975.0	395.0
mmu-miR-25	AUUGCAC	*miR-25/32/92abc/363/363-3p/367*	3430.2	206.6
mmu-miR-106a	AAAGUGC	*miR-17/17-5p/20ab/20b-5p/93/106ab/427/518a-3p/519d*	3396.3	359.5
mmu-miR-17	AAAGUGC	*miR-17/17-5p/20ab/20b-5p/93/106ab/427/518a-3p/519d*	3396.3	359.5
mmu-miR-34c	GGCAGUG	*miR-34ac/34bc-5p/449abc/449c-5p*	2653.1	245.7
mmu-miR-1945	CUUCGCG	*miR-1945*	2617.6	213.1
mmu-miR-291a-5p	AUCAAAG	*miR-291ab-5p*	2461.3	355.6

miRNA	seed	miRNA family	Normalized count	St. Dev.
mmu-miR-22	AGCUGCC	*miR-22/22-3p*	2316.5	156.8
mmu-miR-16	AGCAGCA	*miR-15abc/16/16abc/195/322/424/497/1907*	1988.0	170.3
mmu-miR-96	UUGGCAC	*miR-96/507/1271*	1865.8	237.0
mmu-miR-669f-3p	AUAUACA	*miR-669f-3p*	1713.0	250.6
mmu-miR-669f-5p	GUUGUGU	*miR-669pl/669af-5p/2304*	1713.0	250.6
mmu-miR-15b	AGCAGCA	*miR-15abc/16/16abc/195/322/424/497/1907*	1678.7	119.0
mmu-miR-21	AGCUUAU	*miR-21/590-5p*	1604.1	123.0
mmu-miR-130a	AGUGCAA	*miR-130ac/301ab/301b/301b-3p/454/721/4295/3666*	1584.6	161.6
mmu-miR-466a-3p	AUACAUA	*miR-297b-3p/466ade-3p/467g*	1353.8	121.3
mmu-miR-466b-3p	UACAUAC	*miR-466bcp-3p*	1353.8	121.3
mmu-miR-19a	GUGCAAA	*miR-19ab*	1287.5	130.2
mmu-miR-302d	AAGUGCU	*miR-93/93a/105/106a/291a-3p/294/295/302abcde/372/373/428/519a/520be/520acd-3p/1378/1420ac*	1268.4	100.0
mmu-miR-125a-5p	CCCUGAG	*miR-125a-5p/125b-5p/351/670/4319*	1128.8	114.4
mmu-miR-1224	UGAGGAC	*miR-1224-5p/1671*	982.4	85.6
mmu-miR-19b	GUGCAAA	*miR-19ab*	783.5	79.9
mmu-miR-183	AUGGCAC	*miR-183*	770.9	51.9
mmu-miR-130b	AGUGCAA	*miR-130ac/301ab/301b/301b-3p/454/721/4295/3666*	679.2	41.1
mmu-miR-543	AACAUUC	*miR-543*	666.3	30.7
mmu-miR-15a	AGCAGCA	*miR-15abc/16/16abc/195/322/424/497/1907*	654.4	72.3
mmu-miR-205	CCUUCAU	*miR-205/205ab*	610.1	24.8
mmu-miR-27a	UCACAGU	*miR-27abc/27a-3p*	607.4	49.8
mmu-miR-302a	AAGUGCU	*miR-93/93a/105/106a/291a-3p/294/295/302abcde/372/373/428/519a/520be/520acd-3p/1378/1420ac*	589.1	82.4

miRNA	seed	miRNA family	Normalized count	St. Dev.
mmu-miR-495	AACAAAC	*miR-495/1192*	577.6	39.6
mmu-miR-155	UAAUGCU	*miR-155*	570.2	46.6
mmu-miR-301a	AGUGCAA	*miR-130ac/301ab/301b/301b-3p/454/721/4295/3666*	535.9	34.3
mmu-miR-429	AAUACUG	*miR-200bc/429/548a*	519.7	47.3
mmu-miR-200b	AAUACUG	*miR-200bc/429/548a*	482.0	52.5
mmu-miR-367	AUUGCAC	*miR-25/32/92abc/363/363-3p/367*	448.1	26.8
mmu-miR-467c	AAGUGCG	*miR-467cd/1420ef*	405.8	30.8
mmu-miR-376b	UCAUAGA	*miR-376abd/376b-3p*	404.5	52.6
mmu-miR-136	CUCCAUU	*miR-136*	395.0	40.9
mmu-miR-28	AGGAGCU	*miR-28-5p/708/1407/1653/3139*	362.8	28.8
mmu-miR-467f	UAUACAC	*miR-467f/4789-5p*	351.3	26.8
mmu-miR-125b-5p	CCCUGAG	*miR-125a-5p/125b-5p/351/670/4319*	343.9	29.0
mmu-miR-99b	ACCCGUA	*miR-99ab/100*	337.1	25.3
mmu-miR-106b	AAAGUGC	*miR-17/17-5p/20ab/20b-5p/93/106ab/427/518a-3p/519d*	326.7	80.6
mmu-miR-1983	UCACCUG	*miR-412-3p/1983*	304.7	32.0
mmu-miR-369-3p	AUAAUAC	*miR-369-3p*	298.5	11.4
mmu-miR-23a	UCACAUU	*miR-23abc/23b-3p*	298.0	11.3
mmu-miR-191	AACGGAA	*miR-191*	283.6	23.9
mmu-miR-200c	AAUACUG	*miR-200bc/429/548a*	283.3	19.1
mmu-miR-669a-3p	CAUAACA	*miR-669ao-3p/669a-3-3p*	281.8	14.1
mmu-miR-669a-5p	GUUGUGU	*miR-669pl/669af-5p/2304*	281.8	14.1
mmu-miR-467a	AAGUGCC	*miR-290-3p/292-3p/467a/1420b*	280.9	30.9
mmu-miR-291b-3p	AAGUGCA	*miR-519a/519bc-3p/291b-3p/1347*	265.5	16.2
mmu-let-7a	GAGGUAG	*let-7/98/4458/4500*	255.0	30.0
mmu-miR-93	AAAGUGC	*miR-17/17-5p/20ab/20b-5p/93/106ab/427/518a-3p/519d*	254.1	28.9
mmu-miR-148a	CAGUGCA	*miR-148ab-3p/152*	231.8	25.3

miRNA	seed	miRNA family	Normalized count	St. Dev.
mmu-miR-467e	UAAGUGU	*miR-467eh*	228.7	26.6
mmu-miR-376c	ACAUAGA	*miR-376c/741-5p*	226.4	34.8
mmu-miR-410	AUAUAAC	*miR-410/344de/344b-1-3p*	208.4	19.2
mmu-miR-669o-3p	CAUAACA	*miR-669ao-3p/669a-3-3p*	206.8	9.1
mmu-miR-669o-5p	AGUUGUG	*miR-669o-5p*	206.8	9.1
mmu-miR-804	GUGAGUU	*miR-804*	203.9	11.0
mmu-miR-467h	UAAGUGU	*miR-467eh*	203.5	36.4
mmu-miR-669d	CUUGUGU	*miR-581/669d*	203.5	36.4
mmu-miR-669l	GUUGUGU	*miR-669pl/669af-5p/2304*	203.5	36.4
mmu-miR-377	UCACACA	*miR-377*	201.6	10.4
mmu-miR-323-3p	ACAUUAC	*miR-323/323-3p*	198.4	20.9
mmu-miR-126-3p	CGUACCG	*miR-126-3p*	198.0	22.5
mmu-miR-674	CACUGAG	*miR-674/674-5p/3473d*	193.5	11.2
mmu-miR-32	AUUGCAC	*miR-25/32/92abc/363/363-3p/367*	191.2	5.1
mmu-miR-103	GCAGCAU	*miR-103a/107/107ab*	189.0	19.4
mmu-miR-363-3p	AUUGCAC	*miR-25/32/92abc/363/363-3p/367*	187.5	20.3
mmu-miR-296-5p	GGGCCCC	*miR-296-5p*	178.8	13.6
mmu-miR-124	AAGGCAC	*miR-124/124ab/506*	169.8	22.2
mmu-miR-23b	UCACAUU	*miR-23abc/23b-3p*	169.5	10.1
mmu-miR-434-3p	UUGAACC	*miR-434-3p*	167.6	10.7
mmu-miR-30b	GUAAACA	*miR-30abcdef/30abe-5p/384-5p*	166.5	20.4
mmu-miR-376a	UCGUAGA	*miR-376a*	164.1	75.0
mmu-miR-29a	AGCACCA	*miR-29abcd*	159.5	9.6
mmu-miR-1196	AAUCUAC	*miR-1196*	159.0	9.6
mmu-miR-374	UAUAAUA	*miR-374ab*	157.3	20.5
mmu-miR-145	UCCAGUU	*miR-145*	157.2	7.6

miRNA	seed	miRNA family	Normalized count	St. Dev.
mmu-miR-200a	AACACUG	*miR-141/200a*	151.4	6.2
mmu-miR-382	AAGUUGU	*miR-382*	147.7	7.3
mmu-miR-423-3p	GCUCGGU	*miR-423-3p*	146.9	6.1
mmu-miR-151-5p	CGAGGAG	*miR-151-5p/151b*	145.5	8.6
mmu-miR-127	CGGAUCC	*miR-127/127-3p*	144.5	11.1
mmu-miR-532-5p	AUGCCUU	*miR-532-5p/511*	143.6	17.0
mmu-miR-1952	CUCCACC	*miR-1952*	143.0	25.4
mmu-miR-150	CUCCCAA	*miR-150/5127*	142.8	7.7
mmu-miR-431	GUCUUGC	*miR-431*	141.8	18.2
mmu-let-7g	GAGGUAG	*let-7/98/4458/4500*	141.3	5.6
mmu-miR-193b	ACUGGCC	*miR-193/193b/193a-3p*	138.8	14.3
mmu-miR-29b	AGCACCA	*miR-29abcd*	137.6	15.3
mmu-miR-1935	GGCAGAG	*miR-1935*	134.5	10.0
mmu-miR-297a	UGUAUGU	*miR-297ac/297b-5p*	134.1	9.2
mmu-miR-466f	CGUGUGU	*miR-466f*	134.1	9.2
mmu-miR-101b	ACAGUAC	*miR-101/101ab*	133.4	33.5
mmu-miR-134	GUGACUG	*miR-134/3118*	130.9	14.8
mmu-miR-433	UCAUGAU	*miR-433*	129.9	6.1
mmu-miR-18b	AAGGUGC	*miR-18ab/4735-3p*	127.7	9.3
mmu-miR-421	UCAACAG	*miR-421*	127.4	8.7
mmu-miR-340-5p	UAUAAAG	*miR-340-5p*	126.7	12.7
mmu-miR-142-3p	GUAGUGU	*miR-142-3p*	124.7	10.8
mmu-miR-541	AGGGAUU	*miR-541*	123.5	6.6
mmu-miR-466c-3p	UACAUAC	*miR-466bcp-3p*	122.3	3.3
mmu-miR-335-3p	UUUUCAU	*miR-335-3p*	118.8	10.7
mmu-let-7c	GAGGUAG	*let-7/98/4458/4500*	114.8	8.7

miRNA	seed	miRNA family	Normalized count	St. Dev.
mmu-miR-297c	UGUAUGU	*miR-297ac/297b-5p*	114.6	10.0
mmu-miR-291b-5p	AUCAAAG	*miR-291ab-5p*	114.1	22.6
mmu-miR-30d	GUAAACA	*miR-30abcdef/30abe-5p/384-5p*	114.0	10.0
mmu-miR-378	CUGGACU	*miR-378/422a/378bcdefhi*	113.7	3.0
mmu-miR-142-5p	AUAAAGU	*miR-142-5p*	111.9	18.4
mmu-miR-29c	AGCACCA	*miR-29abcd*	110.8	3.1
mmu-miR-872	AGGUUAC	*miR-872/1421acal*	106.9	14.2
mmu-miR-381	AUACAAG	*miR-300/381/539-3p*	104.5	10.6
mmu-miR-26b	UCAAGUA	*miR-26ab/1297/4465*	103.7	10.7
mmu-miR-30a	GUAAACA	*miR-30abcdef/30abe-5p/384-5p*	102.5	10.3
mmu-miR-484	CAGGCUC	*miR-344a-5p/484/3155/3155b*	97.6	15.6
mmu-let-7e	GAGGUAG	*let-7/98/4458/4500*	96.2	14.3
mmu-miR-423-5p	GAGGGGC	*miR-423a/423-5p/3184/3573-5p*	95.6	8.5
mmu-miR-1893	GCGCGGG	*miR-1893/2277-5p*	94.3	6.8
mmu-miR-466g	UACAGAC	*miR-466g*	94.1	9.7
mmu-miR-335-5p	CAAGAGC	*miR-335/335-5p*	93.7	15.6
mmu-miR-713	GCACUGA	*miR-713*	93.5	12.0
mmu-miR-148b	CAGUGCA	*miR-148ab-3p/152*	93.1	6.4
mmu-miR-350	UCACAAA	*miR-350*	92.8	13.4
mmu-miR-26a	UCAAGUA	*miR-26ab/1297/4465*	92.4	9.7
mmu-miR-129-1-3p	AGCCCUU	*miR-129-3p/129ab-3p/129-1-3p/129-2-3p*	91.8	4.6
mmu-miR-129-2-3p	AGCCCUU	*miR-129-3p/129ab-3p/129-1-3p/129-2-3p*	91.8	4.6

miRNA	seed	miRNA family	Normalized count	St. Dev.
mmu-miR-33	UGCAUUG	*miR-33ab/33-5p*	91.5	12.0
mmu-miR-154	AGGUUAU	*miR-154/872*	90.6	7.9
mmu-miR-152	CAGUGCA	*miR-148ab-3p/152*	90.3	11.9
mmu-miR-141	AACACUG	*miR-141/200a*	89.3	9.8
mmu-miR-329	ACACACC	*miR-329/329ab/362-3p*	89.0	14.3
mmu-miR-182	UUGGCAA	*miR-182*	88.5	19.6
mmu-miR-290-3p	AAGUGCC	*miR-290-3p/292-3p/467a/1420b*	87.4	19.9
mmu-miR-741	GAGAGAU	*miR-741*	84.4	10.2
mmu-miR-1198-3p	AGCUAGC	*miR-1198-3p*	84.3	6.7
mmu-miR-1198-5p	AUGUGUU	*miR-1198-5p*	84.3	6.7
mmu-miR-762	GGGCUGG	*miR-762/4492/4498*	83.5	16.4
mmu-miR-202-3p	GAGGUAU	*miR-202-3p*	83.3	11.7
mmu-miR-132	AACAGUC	*miR-132/212/212-3p*	82.7	6.0
mmu-miR-181a	ACAUUCA	*miR-181abcd/4262*	82.4	6.0
mmu-miR-487b	AUCGUAC	*miR-487b*	82.4	10.7
mmu-miR-1960	CAGUGCU	*miR-1419d/1960*	81.1	6.8
mmu-miR-195	AGCAGCA	*miR-15abc/16/16abc/195/322/424/497/1907*	81.0	10.2
mmu-miR-672	GAGGUUG	*miR-377/672*	80.4	8.1
mmu-miR-34a	GGCAGUG	*miR-34ac/34bc-5p/449abc/449c-5p*	79.4	10.7
mmu-miR-330	CUCUGGG	*miR-326/330/330-5p*	77.2	9.4
mmu-miR-468	AUGACUG	*miR-468*	76.8	9.7
mmu-miR-760-3p	GGCUCUG	*miR-760-3p/1842*	76.7	5.8
mmu-miR-760-5p	CCCUCAG	*miR-760-5p*	76.7	5.8

CHAPTER 5 APPENDIX

Appendix Table A5.1 Prediction of MREs for *miR-124* for all STAGA-encoding genes in both mouse and human (microrna.org, all miRSVR scores)(Betel et al., 2008). Manual miRNA seed matching (nucleotides 2-8) was performed for genes not found in the database, namely, mouse *lnc-SCA7* and *Tada2b*.

Gene symbol	Gene ID	*miR-124* binding	miRNA alignment	Alignment	Gene alignment	Gene start	Gene end	miRNA start	miRNA end	Species
lnc-SCA7	NM_001033474	Yes*	3' ccguaaguggcgcACGGAAu 5'	||||||	5' ggggacaaggccuUGCCUUu 3'	1277	1296	2	21	Mouse
lnc-SCA7	NM_001033474	Yes*	3' ccguaaguggcgcACGGAAu 5'	||||||	5' acuuuuuuccaucUGCCUUc 3'	1643	1662	2	21	Mouse
lnc-SCA7	NM_001033474	Yes*	3' ccguaaguggcgcACGGAAu 5'	||||||	5' ucuuauauaaggcUGCCUUg 3'	2403	2422	2	21	Mouse
lnc-SCA7	NM_001033474	Yes*	3' ccguaaguggcgcACGGAAu 5'	||||||	5' gacugucuguuguUGCCUUc 3'	2660	2679	2	21	Mouse
lnc-SCA7	NM_001033474	Yes*	3' ccguaaguggcgcACGGAAu 5'	||||||	5' gcccacucacucuUGCCUUc 3'	2907	2926	2	21	Mouse
lnc-SCA7	NM_001033474	Yes*	3' ccguaaguggcgcACGGAAu 5'	||||||	5' acuuuaguuuucuUGCCUUu 3'	2947	2966	2	21	Mouse
Atxn7l3	NM_001098836	Yes	3' ccguaaguggcgcACGGAAu 5'	||||||	5' ccuggacauauuaUGCCUUg 3'	208	227	2	21	Mouse
Atxn7	NM_139227	Yes	3' ccguaaguggcgcACGGAAu 5'	||||||	5' cuggacuccacgaUGCCUUu 3'	71	90	2	21	Mouse
Atxn7	NM_139227	Yes	3' ccguaaguggcgcACGGAAu 5'	||||||	5' auuucuguagccuUGCCUUc 3'	354	373	2	21	Mouse
Eny2	NM_175009	Yes	3' ccGUAAGUGGCGCACGGAAu 5'	:| || : :|||||||	5' cuUAAUCUGUAUGUGCCUUc 3'	448	467	2	21	Mouse
Kat2a	NM_020004	Yes	3' ccGUAAGUGGCGCA-CGGAAu 5'	:||||| :| || ||||	5' acUAUUCAGUGAGUAGCCUUc 3'	319	339	2	22	Mouse
Sorbs1	AK122487	Yes	3' ccguaAGUGGCGCACGGAAu 5'	|:| | :|||||||	5' cagggUUAGC-UGUGCCUUu 3'	738	757	2	21	Mouse

Gene symbol	Gene ID	*miR-124* binding	miRNA alignment	Alignment	Gene alignment	Gene start	Gene end	miRNA start	miRNA end	Species
Psmc1	NM_008947	No								Mouse
Supt7l	NM_028150	No								Mouse
Tada1	NM_030245	No								Mouse
Tada3	NM_133932	No								Mouse
Taf5l	NM_133966	No								Mouse
Taf9	NM_001001176	No								Mouse
Taf10	NM_020024	No								Mouse
Taf12	NM_025579	No								Mouse
Taf1	NM_001081008	Yes	3' ccguaaguggcgcACGGAAu 5'	||||||	5' cuuggauagacaaUGCCUUg 3'	398	417	2	21	Mouse
Taf1	NM_001081008	Yes	3' ccguaaguggcgcACGGAAu 5'	||||||	5' ccuuguuuguaccUGCCUUg 3'	942	961	2	21	Mouse
Taf1	NM_001081008	Yes	3' ccguaaguggcgcACGGAAu 5'	||||||	5' gcaugguggcacaUGCCUUu 3'	1049	1068	2	21	Mouse
Trrap	NM_001081362	Yes	3' ccguaaguggcgcACGGAAu 5'	||||||	5' ugagaccucucucUGCCUUg 3'	63	82	2	21	Mouse
Trrap	NM_001081362	Yes	3' ccguaaguggcgcACGGAAu 5'	||||||	5' gcagcuuccgagaUGCCUUu 3'	203	222	2	21	Mouse
Usp22	NM_001004143	Yes	3' ccguaaguggcgcACGGAAu 5'	||||||	5' --------ccuccUGCCUUg 3'	1	20	2	21	Mouse
Usp22	NM_001004143	Yes	3' ccguaaguggcgcACGGAAu 5'	||||||	5' cuccucauaccacUGCCUUu 3'	1108	1127	2	21	Mouse
Ccdc101	NM_029339	No								Mouse
Sap130	NM_172965	No								Mouse

Gene symbol	Gene ID	*miR-124* binding	miRNA alignment	Alignment	Gene alignment	Gene start	Gene end	miRNA start	miRNA end	Species
Sf3b3	NM_133953	No								Mouse
Tada2b	NM_001170454	Yes*	3' ccguaaguggcgcACGGAAu 5'	||||||	5' ucagaagacagcuUGCCUUu 3'	1700	1719	2	21	Mouse
Tada2b	NM_001170454	Yes*	3' ccguaaguggcgcACGGAAu 5'	||||||	5' gucagaugagggcUGCCUUu 3'	2229	2248	2	21	Mouse
Tada2b	NM_001170454	Yes*	3' ccguaaguggcgcACGGAAu 5'	||||||	5' agcauccugccugUGCCUUg 3'	2449	2468	2	21	Mouse
Tada2b	NM_001170454	Yes*	3' ccguaaguggcgcACGGAAu 5'	||||||	5' cgacaugacgacgUGCCUUu 3'	2781	2800	2	21	Mouse
Taf9b	NM_001001176	Yes	3' ccguaAGUGGCGCACGGAAu 5'	||| | | |||||	5' gggucUCAGCACUUGCCUUu 3'	634	653	2	21	Mouse
Taf9b	NM_001001176	Yes	3' ccguaaguggcgcACGGAAu 5'	||||||	5' auaacuuaaaaccUGCCUUg 3'	734	753	2	21	Mouse
Taf6l	AK144881	No								Mouse
lnc-SCA7	NM_001136262	Yes	3' ccguaaguggcgcACGGAAu 5'	||||||	5' uuggaacuguuacUGCCUUg 3'	2089	2108	2	21	Human
lnc-SCA7	NM_001136262	Yes	3' ccguaaguggcgcACGGAAu 5'	||||||	5' aguccaaucccucUGCCUUc 3'	2341	2360	2	21	Human
ATXN7L3	NM_020218	Yes	3' ccguaaguggcgcACGGAAu 5'	||||||	5' ccuggacauauuaUGCCUUg 3'	209	228	2	21	Human

Gene symbol	Gene ID	*miR-124* binding	miRNA alignment	Alignment	Gene alignment	Gene start	Gene end	miRNA start	miRNA end	Species
ATXN7	NM_001128149	Yes	3' ccguaaguggcgcACGGAAu 5'	||||||	5' uuauaaauaaaguUGCCUUu 3'	1938	1957	2	21	Human
ENY2	NM_020189	No								Human
KAT2A	NM_021078	Yes	3' ccguaaguggcgcACGGAAu 5'	||||||	5' cugaaggggccaUGCCUUg 3'	412	431	2	21	Human
PSMC1	NM_002802	No								Human
SORBS1	NM_001034954	Yes	3' ccguAAGUGGCGCA-CGGAAu 5'	| ||:| ||| ||||	5' aauuUGCAUCACGUCGCCUUc 3'	2432	2452	2	22	Human
SORBS1	NM_001034954	Yes	3' ccgUAAGUGGCGCACGGAAu 5'	| | :| || |||||	5' cuaACUGGCGGCAUGCCUUg 3'	2576	2595	2	21	Human
SUPT7L	NM_014860	Yes	3' ccguaAGU-GGCGCACGGAAu 5'	||| || : |||||	5' ucuccUCAUCCUUUUGCCUUu 3'	655	675	2	22	Human
SUPT7L	NM_014860	Yes	3' ccguaAGUGGCGCACGGAAu 5'	|::|: | |||||	5' augugUUGCUUCAUGCCUUg 3'	1030	1049	2	21	Human
SUPT7L	NM_014860	Yes	3' ccGUAAGUGGCGCACGGAAu 5'	|||| |::| |||||	5' uuCAUU-AUUGAAUGCCUUa 3'	1162	1181	2	21	Human
TADA1	NM_053053	Yes	3' ccguaaguggcgcACGGAAu 5'	||||||	5' agucuaaaauuucUGCCUUu 3'	399	418	2	21	Human
TADA3	NM_133480	No								Human
TAF5L	NM_001025247	No								Human
TAF9	NM_016283	No								Human
TAF10	NM_006284	No								Human
TAF12	NM_005644	No								Human
TAF1	NM_004606	Yes	3' ccguaaguggcgcACGGAAu 5'	||||||	5' guuuguuuguaccUGCCUUg 3'	1450	1469	2	21	Human
TRRAP	NM_003496	No								Human

Gene symbol	Gene ID	*miR-124* binding	miRNA alignment	Alignment	Gene alignment	Gene start	Gene end	miRNA start	miRNA end	Species
USP22	NM_015276	Yes	3' ccguaaguggcgcACGGAAu 5'	||||||	5' agggccuugcagaUGCCUUu 3'	433	452	2	21	Human
USP22	NM_015276	Yes	3' ccguAAGUGGCGCACGGAAu 5'	| | |:| ||||||	5' uggcUCCUCUGGGUGCCUUg 3'	818	837	2	21	Human
USP22	NM_015276	Yes	3' ccguaaguggcgcACGGAAu 5'	||||||	5' cucguauuuucuuUGCCUUc 3'	960	979	2	21	Human
USP22	NM_015276	Yes	3' ccguaaguggCG-CACGGAAu 5'	|| |||||||	5' ugugagcuggGCAGUGCCUUc 3'	2320	2340	2	22	Human
CCDC101	NM_138414	No								Human
SAP130	NM_001145928	Yes	3' ccguaaguggcgcACGGAAu 5'	||||||	5' ggacauagugagcUGCCUUc 3'	176	195	2	21	Human
SF3B3	NM_012426	Yes	3' ccguaaguggcgcACGGAAu 5'	||||||	5' guucccacgcugcUGCCUUu 3'	2258	2277	2	21	Human
SF3B3	NM_012426	Yes	3' ccguaaguggcgcACGGAAu 5'	||||||	5' caaaaaaucuuccUGCCUUg 3'	2603	2622	2	21	Human
SF3B3	NM_012426	Yes	3' ccguaaguggcgcACGGAAu 5'	||||||	5' ccaccucucacucUGCCUUc 3'	4555	4574	2	21	Human
SF3B3	NM_012426	Yes	3' ccguaaguggcgcACGGAAu 5'	||||||	5' ccugugguggccacUGCCUUg 3'	5665	5684	2	21	Human
TADA2B	NM_152293	Yes	3' ccGUAAGUGGCGCACGGAAu 5'	:| |:|| :|||||||	5' cuUAAUUAC--UGUGCCUUu 3'	1494	1513	2	21	Human

Gene symbol	Gene ID	*miR-124* binding	miRNA alignment	Alignment	Gene alignment	Gene start	Gene end	miRNA start	miRNA end	Species
TAF9B	NM_015975	Yes	3' ccguaaguggcgcACGGAAu 5'	||||||	5' ugagacacuuuucUGCCUUc 3'	257	276	2	21	Human
TAF9B	NM_015975	Yes	3' ccGUAAGUGG----CGCAC-GGAAu 5'	||||:||| |:||| ||||	5' auCAUUUACCAACUGUGUGACCUUa 3'	1656	1680	2	26	Human
TAF6L	NM_006473	No								Human

Appendix Table A5.2 Absolute quantification (in copy number per µL of cDNA) measured by digital droplet PCR (ddPCR) of *lnc-SCA7* and *Atxn7* in N2A and ES cells, as well as the cerebellum, lung, and liver of SCA7$^{266Q/5Q}$ and their matched littermate SCA7$^{5Q/5Q}$ control mice.

		lnc-SCA7 (copies/µL of cDNA)	***Atxn7 (copies/µL of cDNA)***
N2A		320	16
ES		145	5
Cerebellum	**SCA7$^{5Q/5Q}$**	390	28
	SCA7$^{266Q/5Q}$	470	33
Lung	**SCA7$^{5Q/5Q}$**	140	5.5
	SCA7$^{266Q/5Q}$	137	5.3
Liver	**SCA7$^{5Q/5Q}$**	115	12
	SCA7$^{266Q/5Q}$	105	12

CHAPTER 6 APPENDIX

A6.1 Supplementary Notes (Analyses performed by Dr. Allison Piovesan): Correlation between expression levels of MRE-sharing ASD-implicated genes in autism patients

Candidate gene-driven and genome-wide studies aimed to determine genes involved in autism spectrum disorders (ASD) have identified 447 protein-coding genes implicated in the syndromic forms of the disease. This analysis was aimed to identify a gene network in which ASD-implicated genes crosstalk through a microRNA (miRNA)-mediated mechanism and contribute to the disease etiology.

First, a non-redundant list of 447 ASD-implicated genes was obtained from publicly available data of previous studies (Anney et al., 2010; Betancur, 2011; Neale et al., 2012; O'Roak et al., 2012b; Sanders et al., 2012; Kohler et al., 2014). Furthermore, brain transcriptomic data from the prefrontal and temporal cortex of healthy and ASD affected adults with genotype information was obtained from Voineagu et al. 2011 (Appendix Table A6.1).

Appendix Table A6.1 Human postmortem frontal and temporal cortex samples of healthy and ASD-affected individuals. All Sample IDs and Platform IDs were obtained from microarray data (Illumina HumanRef-8 v3.0 Expression BeadChip) deposited at the GEO database.

Study ID	Sample ID	Tissue type	Disease state	Platform	Spots
A1…A6 (n=6)	GSM706412…17	Prefrontal cortex	Disease	GPL6883	24,526
A7…A15 (n=9)	GSM706428…36	Prefrontal cortex	Control	GPL6883	24,526
A16…A19 (n=4)	GSM706444…47	Temporal cortex	Disease	GPL6883	24,526
A20…A27 (n=8)	GSM706449…56	Temporal cortex	Control	GPL6883	24,526

Using TargetScan version 6.1 (Garcia et al., 2011), 317 of the 447 (71%) ASD-implicated genes were predicted to harbour at least one MRE for *miR-1253*, and 130 (29%) genes were not predicted to contain *miR-1253* MREs. Amongst the ASD-implicated genes, expression levels were available for 160 and 66 ASD-implicated genes with and without predicted *miR-1253* MREs, respectively in the healthy and ASD affected individuals (Appendix Table A6.1) (Voineagu et al., 2011).

Assessment of the correlation in available expression levels between pairs of all ASD-implicated genes with at least one predicted MRE for *miR-1253* (160 genes) and those without predicted *miR-1253* MREs (66 genes) revealed that only pairs of ASD-implicated that harbour *miR-1253* MREs were significantly correlated in expression levels (p=0 and p=0.002 for healthy and ASD-affected

adults, respectively, empirical *p* value) relative to 1,000 randomly simulated sets of 160 genes not implicated in ASD (Figure A6.1). In contrast, ASD-implicated genes with no predicted *miR-1253* MREs were not correlated in expression levels (p=0.15 and p=0.076 for healthy and ASD-affected adults, respectively) relative to 1,000 randomly simulated sets of 66 genes not implicated in ASD (Figure A6.2).

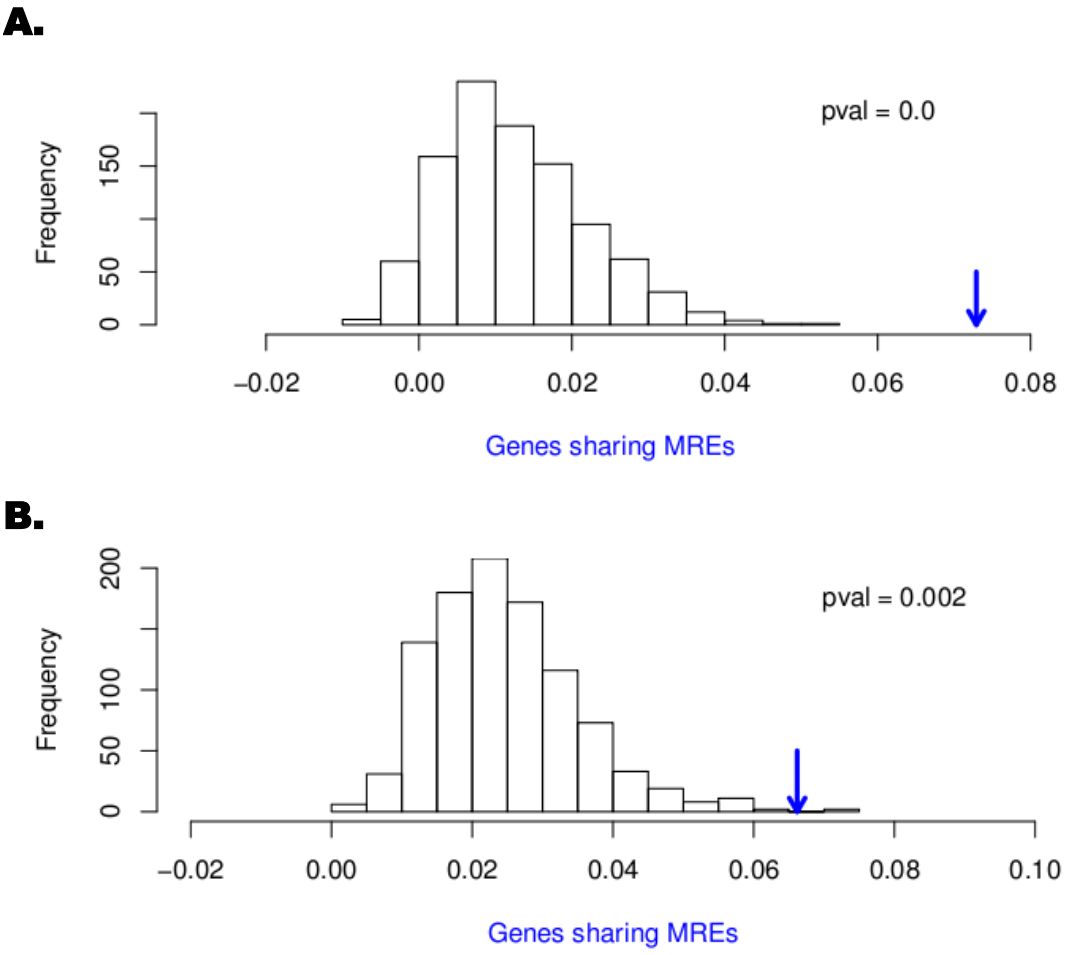
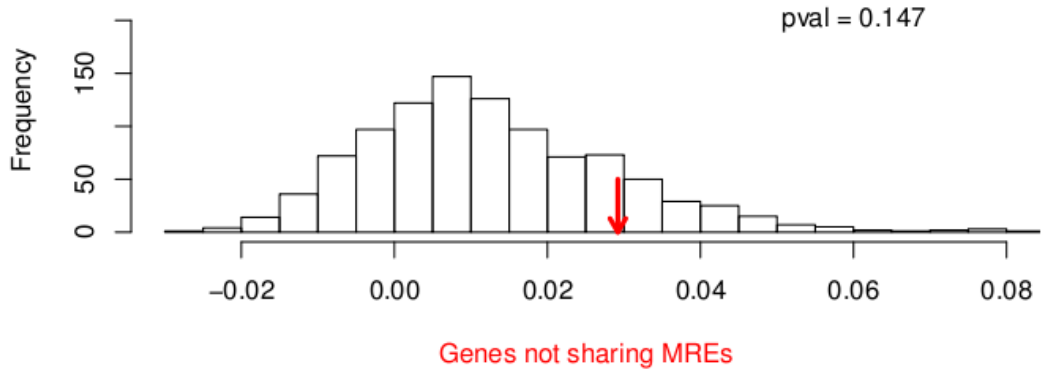


Figure A6.1 Pearson's correlation coefficients in pairwise expression of ASD-implicated genes in healthy and ASD-affected adult postmortem brain samples containing predicted *miR-1253* MREs. The arrows represent the median value of correlation coefficients of ASD-simulated genes containing at least one predicted MREs for *miR-1253* in (A) healthy and (B) ASD-affected individuals. The background histogram is constructed using the median value of correlation coefficients calculated for 1,000 randomly simulated sets of 160 genes that are not implicated in ASD.

A.



B.

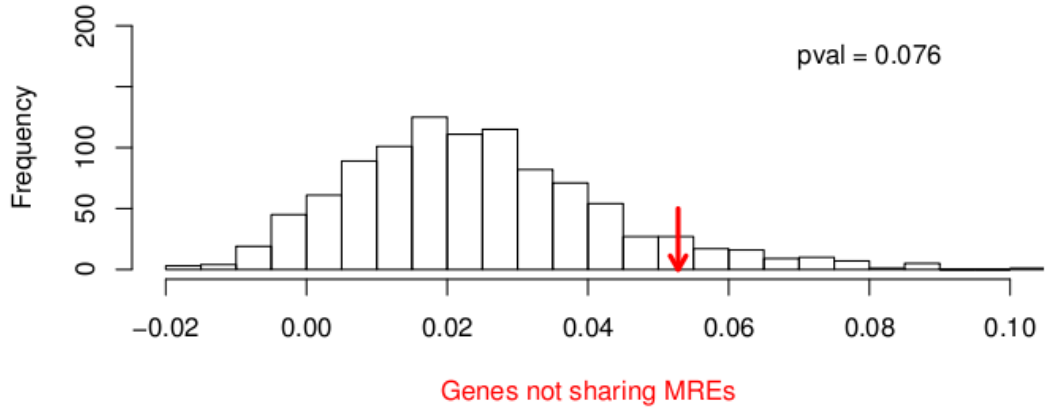


Figure A6.2 Pearson's correlation coefficients in pairwise expression of ASD-implicated genes in healthy and ASD-affected adult postmortem brain samples without predicted *miR-1253* MREs. The arrows represent the median value of correlation coefficients for ASD-simulated genes that do not contain predicted MREs for *miR-1253* in (A) healthy and (B) ASD-affected individuals. The background histogram is constructed using the median value of correlation coefficients calculated for 1,000 randomly simulated sets of 66 genes that not implicated in ASD.

The above computational analyses suggested that ASD-implicated genes with predicted MREs for *miR-1253* are more coexpressed in human prefrontal and temporal cortex of postmortem brain sample than expected by chance, in both healthy and ASD-affected individuals.