2010

Genome-Wide Decoding of mRNP and miRNA Maps

Sung Wook Chi

GENOME-WIDE DECODING OF MRNP AND MIRNA MAPS

A Thesis Presented to the Faculty of

The Rockefeller University

in Partial Fulfillment of the Requirements for

the degree of Doctor of Philosophy

by

Sung Wook Chi

June 2010

Genome-wide decoding of mRNP and miRNA maps

Sung Wook Chi, Ph.D.

The Rockefeller University 2010

The limited number of primary transcripts in the genome has promoted interest in the possibility that much of the complexity in the regulation of gene expression may be determined by RNA regulation controlled by RNA-binding proteins (RNABPs) and/or microRNAs (miRNAs). However, applying biochemical methods to understand such interactions in living tissues is major challenge. Here we developed a genome-wide means of mapping messenger ribonucleoprotein (mRNP) sites in vivo, by high-throughput sequencing of RNA isolated by crosslinking immunoprecipitation (HITS-CLIP). HITS-CLIP analysis of the neuron-specific splicing factor Nova provides genome-wide maps of Nova-RNA interactions in vivo and leads to a new finding that Nova may regulate the processesing of some miRNAs. Furthermore, HITS-CLIP analysis is extended to the problem of identifying miRNA targets, for which prediction is a major challenge since miRNA activity requires base pairing through only 6-8 "seed" nucleotides. By generating crosslinking of native Argonaute (Ago) protein-RNA complexes in mouse brain, Ago HITS-CLIP produced two simultaneous datasets—Ago-miRNA and Ago-mRNA binding sites—that were combined with bioinformatic analysis to identify miRNA-target mRNA interaction sites. We validated genome-wide interaction maps for miR-124, and generated additional maps for the 20 most abundant miRNAs present in P13

mouse brain. We also found that the relatively large number of Ago proteins bind in coding sequence, as well as introns, suggesting unexplored functions for miRNAs. Not all Ago mRNA clusters correspond to known seed sequence, leading to the discovery of putative new rules for miRNA-mRNA interactions. HITS-CLIP provides a general plaform to identify functional mRNP and miRNA binding sites in vivo and a solution to determining precise sequences for targeting clinically relevant sites of RNA regulation. In addition, overlaying mRNP maps with miRNA maps will be informative for the understanding of RNA regulations and complexity.

나의 사랑하는 아들 은성, 반려자 장은숙박사, 그리고 미래의둘째에게..

To my family... Alex, Eunnie and ...

Acknowledgements

First, I would like to thank and give all my love to Eunnie (Eun-sook) Jang for her support and advice as a friend, wife and scientist and to my lovely son, Alex Eunsung Chi, for bringing smile into my every day, which always encourage me to pass through any difficulty.

I would like to give all my gratitude and respect to my thesis advisor, Dr. Robert B. Darnell, for guiding and supporting me to think as a scientist. Bob is always inspirational, full of questions and insights, always intriguing me to organize ideas to draw final big pictures, and bringing the projects to the next level. I also would like to thank to Dr. Jennifer Darnell for helping me to have great time in the lab, by creating stimulating research environment together with Bob. I would also like to thank to my thesis committee members: Dr. Hermann Steller and Dr. C. David Allis for their discussion and valuable suggestions, Dr. Gregory J. Hannon for his constructive advice on miRNA projects and Dr. Michael Q. Zhang for his serving as my outside examiner.

I am grateful to the past and present members of the Darnell lab for their guidance and discussions. In particular, I would like to thank to Dr. Woong-yang Park and Taesun Eom for their help and guidance for my lab life here as Korean. I also thank Aldo Mele for his help with CLIP experiments, Julie Zang for her initiation of Ago CLIP project, Chaolin Zhang for bioinformatic discussion, Donny Licatalosi for Nova HITS-CLIP, John Fak for his help with microarray experiments and Kevin O'Donovan, Graeme Couture and Katie Staudt for their help with cdr2 project. I also thank to Scott for his help with high-throughput sequencing, Gabe for the specific discussion about the improvement of Ago CLIP.

I would like to thank to Dr. David Christini for directing Tri-institutional Program in Computational Biology and Medicine, Dr. Tae-Wan Kim and Dr. Hyung Don Ryoo for their precious advice on my Ph.D student life. I also thank to my best friend, Yong-seung Lee, for his patience with my often disturbing long-distance call. Finally, I send my all my love and gratitude to my parents and brothers in Korea for their love and support. Nothing would have been possible without them.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER I. GENERAL INTRODUCTION

RNA complexity

Each cell in our bodies contains a set of genes, from which information flows to precisely controlled genetic transcripts called messenger RNAs (mRNAs) that encode distinct functional outputs as translated proteins. The central dogma of molecular biology suggests that more complex organisms with more diverse function could be associated with larger numbers of genes (Mattick, 2003). However, genome sequencing projects reveal that the number of protein-coding genes in human (~22,808) is similar with lower organisms such as mouse (~22,010) (Waterston et al., 2002) and C. elegans (~19,427) (Consortium, 1998). This discrepancy between organismal complexity and number of genes could be explained by sophisticated regulatory mechanisms acting on RNA, where increased diversity leads to the amplification of genomic complexity (Sharp, 2009). As the coding capacity in DNA is increased by the different usage of promoters controlled by various DNA-binding proteins (Davuluri et al., 2008), the same extent of increase in RNA complexity is expected from alternative usage of exons, RNA editing and polyadenylation controlled by RNA-binding proteins (RNABPs). Furthermore, post-transcriptional regulation of the degradation (Houseley and Tollervey, 2009; Richter, 2008), translation and

localization of mRNA (Rodriguez et al., 2008) are also controlled by specific RNABPs through encoded RNA sequences beyond gene-coding information, such as RNA regulatory elements in introns for alternative splicing (Dredge and Darnell, 2003) or in 3'untraslated regions (3'UTRs) (Keene, 2007). Such regulation depends on the recognition of RNABPs to cis-acting RNA elements, which are more divergently encoded through evolution than the protein sequence of RNABPs, generating different RNA complexity leading to organismal diversity (Jelen et al., 2007).

More recently, identification of a small regulatory RNAs-protein complex, called RNA-inducing silencing complex (RISC) underscores the possibility of increased complexity by microRNAs (miRNAs) (Bartel, 2004), as each miRNA is believed to bind directly to many mRNAs to regulate their translation or stability (Filipowicz et al., 2008). Many miRNAs are evolutionarily conserved, although others are species-specific (including human miRNAs not conserved in chimpanzee), consistent with roles ranging from generating cellular to organismal diversity (Berezikov, 2006). Therefore, complex organisms are thought to have evolved along with a diversification of RNA output, mediated by regulatory RNA-protein and RNA-RNA interactions (Figure 1.1).

RNA regulation for complexity

RNA complexity is contributed by various RNA regulatory events controlled by RNABPs or non-coding RNAs including miRNAs (Figure 1.1). The diversity generated by such RNA regulation is likely to contribute to different functions in tissues with the same genomic coding capacity. By regulating RNA complexity through alternative splicing, alternative polyadenylation, RNA editing, nuclear export, localization, mRNA turnover and translational regulation, RNABPs (Lunde et al., 2007) and miRNAs (Ambros, 2004; He and Hannon, 2004) control a wide range of activities, including development, immune function and neuronal biology, consistent with its critical roles in cellular function and organismal diversity (Figure 1.1).

The role of RNABPs is especially critical in cellular functions in neurons, highly specialized cells with impressive complexity (Ule and Darnell, 2006). Neuronal RNABPs such as neuro-oncological ventral antigen (Nova) regulate neuron specific splicing to generate a versatile repertoire of proteins, functioning in synapse (Huang et al., 2005). In addition to increasing diversity of gene products, RNABPs also have special roles in spatial control of mRNAs, required by complicated structure and function of neurons (Figure 1.1) (Rodriguez et al., 2008). Neurons have small cell bodies relative to their extended projections,

which form complicated neuronal circuitry and require rapid response to the communication signal in synapse. To overcome such a long distance from nucleus to neuronal processes, RNABPs such as fragile-X mental retardation protein (FMRP) (Antar et al., 2004; Antar et al., 2005; De Diego Otero et al., 2002) and Zip code binding protein 1 (ZBP1) (Bassell et al., 1998; Zhang et al., 2001a) can deliver mRNAs from nucleus to neuronal processes, and may act by localizing with and suppressing the translation of target mRNAs to induce local protein translation in the presence of communication signals such as neurotransmitters. In addition to RNABPs, miRNAs such as miR-134 is also known to localize in dendrites to regulate synaptic plasticity through local translation (Schratt et al., 2006; Siegel et al., 2009). Therefore, it is perhaps not surprising that perturbations of RNA regulators are associated with variety of neurological diseases. Here a few examples of such RNA regulators are introduced to elucidate their diverse roles in RNA complexity and function.

-Alternative Splicing (Nova)

Nova is a target antigen in a paraneoplastic neurologic disorders (PND) termed paraneoplastic opsoclonus myoclonus ataxia (POMA), an autoimmune neurologic disease characterized by abnormal motor inhibition (Luque et al., 1991). Western blot analysis with POMA antisera detects three antigens, Nova-1,

Nova-2 and Supernova. Nova-1 is a 50-55 kD antigen expressed in the ventral midbrain, hindbrain, and ventral spinal cord (Buckanovich et al., 1993). Nova-2 is a 50 kD antigen expressed in neocortex, thalamus, inferior colliculus, and the external granule cell nuclei of the cerebellum (Yang et al., 1998). Supernova is an antigen of 70-75 kD, which may come from alternative start codon usage of Nova-2 gene. *Nova-1* null mice die at 7-10 days after birth with a profound motor failure (Jensen et al., 2000a). *Nova-2* null mice show similar post-natal lethal phenotype as *Nova-1* null mice (Ruggiu M., personal communication).

Nova proteins contain three functional hnRNP K-homology type (KH-type) RNA-binding domains and function as brain specific splicing factors which regulate the splicing of target RNAs (Buckanovich and Darnell, 1997). A specific sequence element, YCAY tetramer (where Y indicates a pyrimidine, U or C) was identified as a Nova RNA ligand by *in vitro* RNA selection (Jensen et al., 2000b), X-ray crystallography (Lewis et al., 2000) and further confirmed by *in vivo* cross-linking and immunoprecipitation (CLIP) (Ule et al., 2003), a method developed to screen *in vivo* protein-RNA interaction in the Darnell lab. In addition, custom splicing microarray from Affymetrix revealed that Nova regulates the splicing of target RNAs encoding proteins that function mostly at the synape (Huang et al., 2005). Based on splicing changes and location of YCAY motifs in ~50 validated

targets, it was possible to generalize the rule of Nova's action and more targets were identified by the prediction depending on the position of YCAY clusters (Ule et al., 2006). In detail, recruitment of Nova on exonic YCAY clusters changes the protein complexes assembled on pre-mRNA, blocking U1 snRNP binding, and inhibiting exon inclusion. However, binding of Nova on intronic YCAY clusters enhances spliceosome assembly and subsequently induces exon inclusion by means of removal of introns closest or harboring YCAY clusters (Ule et al., 2006). So far, 84 Nova targets have been identified by experimentally validating that their alternative splicing is dependent on Nova. In addition, long-term potentiation (LTP) of the slow inhibitory postsynaptic current (sIPC) was abolished in *Nova-2* null mice (Huang et al., 2005) and clustering of acetylcholine receptors (AChRs) at the neuromuscular junction (NMJ) is failed in *Nova-1*/*Nova-2* double knockout (DKO) mice but rescued in the DKO mice breeding with a Nova dependant isoform of agrin, containing inserts at a splice site called "Z" (Z(+) agrin) (Ruggiu et al., 2009). Such observations showing that RNA complexity, generated from the RNABP dependant splicing events, is important in neuronal function.

-Translational control (FMRP)

Fragile X syndrome is the most common inherited form of mental retardation disease, which is generally caused by the absence of the RNA binding protein called fragile X mental retardation protein (FMRP). This typically occurs because of expansion of a CGG repeat in the upstream region of corresponding gene, FMR1, leading to transcriptional silencing (Jin and Warren, 2000). Several RNA binding domains and corresponding ligands are identified in FMRP (Kaytor and Orr, 2001). Using *in vitro* RNA selection, it has been found that the RGG box, a RNA binding domain enriched in arginine and guanine residue, could interact with the specific RNA structure called a G-quartet, a planar conformation with four guanine residues (Darnell et al., 2001). In addition, the KH2 domain, the second RNA binding domain in tandem KH-type domain of FMRP, could recognize an RNA structure called a loop-loop pseudoknot or kissing complex (Darnell et al., 2005). Of note, a missense mutation (I304N) in the KH2 domain is associated with severe disease in a patient, and the KH2 RNA ligands compete FMRP off of brain polyribosomes, suggesting that kissing complexes may be FMRP target ligands for translational regulation (Darnell et al., 2005; Feng et al., 1997; Stefani et al., 2004). It has been thought that FMRP could bind its target mRNAs and regulate their local translation depending on the neuronal activity (Garber et al., 2006). So far, several candidate FMRP targets have been identified

from various approaches but few targets are well validated due to lack of a clear functional assay, as well as difficulties in detecting subtle change at the protein level of the targets and issues of potentail functional redundancy with the Fragile X related proteins FXR1P and FXR2P. One of the best validated target mRNAs is microtubule associated protein 1B (MAP1B), of which expression is slightly elevated in FMR1 null mice (Lu et al., 2004). In addition, the MAP1B fly homolog *Futsch,* is elevated in *Drosophila* Fmr1 null flies (Zhang et al., 2001b). This finding may partly explain the delayed dendritic spine maturation phenotype in fragile X mental retardation patients as well as in *Fmr1* knockout mice (Irwin et al., 2000; Nimchinsky et al., 2001). Taken together, it is believed that most FMRP targets are anticipated to have functions in synaptic development related with learning and memory. The Fragile X syndrome illustrates the critical role of RNABPs in neuronal function and pathogenesis of the disease, emphasizing the importance of RNA complexity generated from their different set of targets and locations where translation is under control.

-miRNA mediated post-transcriptional control

miRNAs are small non-coding RNAs which repress gene expression through post-transcriptional control. They are approximately ~22 nucleotides in length and hybridize to the 3'UTR of target mRNAs through partial base pairing (He

and Hannon, 2004). The first miRNA, lin-4, was discovered by screening of genetic mutants showing developmental defects in Caenorhabditis elegans and is known to have antisense complementarity to the 3' untranslated region of lin-14 (Lee et al., 1993). As target mRNAs usually remain intact after partial base pairing to miRNAs, it is believed that most miRNAs induce translational repression through incomplete hybridization, while in a subset of cases of perfect base pairing induces degradation of target mRNAs (Bartel, 2004).

The mechanism for miRNA-mediated translational repression is largely unknown. Each miRNA is believed to bind directly to many mRNAs to regulate their translation or stability and thereby control a wide range of activities, including development, immune function and neuronal biology (Ambros, 2004). Therefore, RNA regulation by miRNAs could generate great complexity at the post-transcriptional level, which may explain the gap between the great complexity of cellular functions and the limited number of primary transcripts. Furthermore, some miRNAs are species-specific, including ~50 miRNAs present in human that are not conserved in chimpanzee, although many miRNAs are evolutionarily conserved (Berezikov, 2006). Like neuronal RNABPs localized in dendrite for local translation control in neuron, neuronal miRNAs such as miR-134 are also reported to localize in dendrite to regulate activity dependent local

translation of target mRNA, generating another RNA complexity in terms of different space and location under control (Schratt et al., 2006; Siegel et al., 2009).

Systematic approaches for RNA regulation study

Completion of human genome projects (Lander et al., 2001) and recent technological advances in biological experiments such as microarray (Schena et al., 1995), high-throughput sequencing (Hutchison, 2007), cross-linking immnoprecipitation (CLIP) (Ule et al., 2003) and bioinformatics enable investigators to study systematic views of RNA regulation. Accordingly, genome-wide RNA maps can be generated based on the characteristics of RNA populations measured by these new methods depending on cellular context, giving genome-wide insights into new biological functions of RNA regulation (Ule et al., 2006). Here several examples of such methods are introduced to show the current states of technologies for studying genome-wide RNA regulations.

-Microarray

Systematic efforts to characterize RNA population began with microarrays, technologies developed to measure abundance of mRNAs with initial interest in mRNA level changes induced by transcriptional regulations (Schena et al., 1995). Microarrays have been widely used to characterize gene expression patterns in

many cells and tissues under diverse conditions, giving us systematic pictures of the transcriptional responses of the human genome to variety of physiological and pathological conditions (Brown and Botstein, 1999). Moreover, microarray technologies have also recently been applied to studying a post-transcriptional control of mRNAs, an important regulatory mechanism in gene expression with richer regulation than transcriptional control. For this, microarray has been modified into a variety of different platforms to elucidate such an RNA complexity, particularly of splice variants using probesets for alternative exons identified from previous experiments or genome sequencing data (Modrek and Lee, 2002).

Exon junction microarray is designed to have specific probesets for the sequences in exon junctions, including the junctions generated by alternative splicing events. RNA populations from different tissues were examined by the exon junction microarrays, leading to initial estimation that a large number of human genes (~3/4) are alternatively spliced (Johnson et al., 2003). Tissue-restricted patterns of alternative mRNA expression are somewhat explained by the evidence that they could be modulated by tissue specific RNABP (Ule and Darnell, 2006). Although exon junction microarray is specific and sensitive enough for detecting alternative mRNA isoforms, it is limited by the current state

of gene annotations, and unable to identify the novel splicing isoforms. In an effort to overcome this, exon microarrays, which interrogate complete sets of transcribed exons, were developed to measure the usage of individual exons not only detecting for novel alternative mRNA splicing isoforms but also for different polyadenylation (Bemmo et al., 2008; Clark et al., 2007). Furthermore, exon arrays also provide a more reliable way for estimating mRNA abundance using multiple probesets for each exon. Recently microarray platforms are further expanded to cover all genome sequences, called "tiling array", enable to detect and study all transcriptiome (Johnson et al., 2005). Although microarray technologies use sophisticated algorithms for designing probe sequences and also for data analysis, it is still has a limitation that experiments are dependent on nucleic acid hybridization which could be biased accordingly to sequence composition.

-High-throughput sequencing

Recent technology advances in sequencing enable to read billions of nucleotides in sequences in low-cost, opening the new age of sequencing that enable to begin personal genome analysis. The high-throughput sequencing, often called next-generation sequencing, was achieved by several platforms which parallelize the

process with different advanced reading technologies (Hutchison, 2007). Among them, 454, Solexa/Illumina, and SOLiD sequencing technologies are widely used.

454 sequencing is highly parallel pyrosequencing system, which reads a fiber-optic slide of individual wells, where DNA is amplified using an emulsion PCR method and is extended for sequencing in the solid support by DNA polymerase pyrophosphates (Margulies et al., 2005). During the extension, it adds one nucleotide species at a time and quantifying the number of nucleotides added to a given location by detecting the release of attached. This method can read total 400 million bases, generating up to 400 base pair reads, at 99% or better accuracy per run. Solexa/Illumina sequencing technology requires planar, optically transparent surfaces to physically attach DNA molecules in parallel and amplifies them as bridge but uses reversible terminator methods to read the sequencing by DNA polymerase (Bennett, 2004; Bennett et al., 2005). During the process, it adds one nucleotide at a time, detecting fluorescence dye-termination at each position and repeating removal of terminator group for another nucleotide extension, ensuring high accuracy and true base-by-base sequencing. Furthermore, it can be expanded to paired-end sequencing, which read sequences from the opposite ends by regenerating the templates after completion of the first read. This method can read total more than 20 billion bases,

generating ~75bp reads (~36bp reads for single-end sequencing). The SOLiD platform parallelizes the process by performing emulsion PCR with primer-coated beads for amplification and eliminating beads without template or with two different templates for enrichment (Shendure et al., 2005). For reading, it uses a DNA ligase to preferentially extend to the annealed oligonucleotide, one from the pool of all possible oligonucleotides of a fixed length, labeled according to the sequenced position with 2 base encoding. Due to the 2 base encoding, it provides sequencing accuracy greater than 99.94% and generates over 20 gigabases and 400M tags per run.

Taking advantage of the power of high-throughput sequencing technologies, complete RNA profiles are currently able to be produced (Wang et al., 2009). Such an application called "RNA-Seq", which uses high-throughput sequencing technologies to sequence complementary DNA (cDNA) in order to get information about a sample's RNA content, was used to assess mRNA complexity, providing analysis of alternative splicing and polyadenylation variants in difference tissues (Pan et al., 2008). The advantage of this method is to identify uncharacterized mRNA isofoms and novel non-coding RNAs, which can't be achieved by other methods such as microarray (Blencowe et al., 2009). Furthermore, high-throughput sequencing can be coupled with other methods,

which select and enrich the subset of RNA population, to identify and estimates the abundance of subset with improved signal:noise.

-Cross-linking and immunoprecipitation (CLIP)

Regulating the complexity of RNA has been shown to be mediated by variety of RNA binding protein and non-coding RNAs. To identify the functional targets and regulatory mechanism of these regulators, researchers developed variety of biochemical approaches for screening RNA targets (Darnell, 2006). But a critical caveat common to all of these approaches is their inability to definitively distinguish direct from indirect effects. For identifying direct targets through RNA-protein interactions, several biochemical methods have been developed based on immunoprecipitation (IP) of RNABPs followed by purification of bound RNA and characterization by RT-PCR or microarray analysis, which are proven to be useful to identify functional RNA targets *in vivo* (Keene et al., 2006). However, these methods are limited by poor specificity due to the low stringency of such co-IPs causing pulling down of secondary RNABPs, non-specifically bound RNAs, and problems of RNABP-RNA re-association *in vitro* (Mili and Steitz, 2004). To overcome such drawbacks in signal:noise ratio, a further improved method of identifying direct RNABP-RNA interactions *in vivo* has been developed, which is named cross-linking and immunoprecipitation

(CLIP) assay (Jensen and Darnell, 2008; Ule et al., 2005a; Ule et al., 2003). CLIP uses ultraviolet (UV) irradiation to covalently crosslink RNA-protein complexes that are in direct contact within the range of approximately single Angstrom distances in living cells, allowing them to be stringently purified. Partial RNA digestion reduced bound RNA to fragments that can be sequenced, providing a new opportunity to identify direct RNA-protein interactions.

-Bioinformatics

Bioinformatics has emerged as a powerful approach to compliment to current experimental technologies for analyzing RNA regulations. It enables one to mine general rules from massive data generated by high-throughput experiments and also to generate new hypotheses predicting RNA regulation by sequence analysis from the data. A fundamental analysis common to all of these bioinformatics analysis is to identify RNA regulatory elements to understand rules of RNA regulation. To identify such RNA elements, initially simple pattern-matching algorithms with exhaustive tree searches such as *RNAmot, RNAbob, RNAmotif* and *PatSan* algorithms were developed (Dsouza et al., 1997; Laferriere et al., 1994; Macke et al., 2001). However, they have limitations in sensitivity and also need an accurate consensus sequence for searching.

Two main approaches, comparative sequence analysis and energy minimization, which considers formation of RNA secondary structures, have been applied to improve the bioinformatics methods for RNA. Comparative sequence analysis uses phylogenetic information to build stochastic context free grammar models for RNA, considering base pairs from pairs of aligned sequences covarying in a *Watson-Crick* pair, and dynamic programming for searches (Eddy and Durbin, 1994; Sakakibara et al., 1994). The limitation of this approach is that it requires many diverse sequences and highly accurate multiple alignments. Energy minimization approaches have been developed utilizing calculated free energies of RNA secondary structures predicted by *Mfold* (Mathews et al., 1999; Zuker and Stiegler, 1981) or *RNAfold* (Hofacker, 2003) programs. These algorithms evaluate all feasible structures for a given RNA sequences, and find the one with minimal free energy based on dynamic programming. They are also able to calculate explicit partitional functions, the sum of exponentiated energies over all possible folds for analyzing probabilities of each base pairing (McCaskill, 1990), which can be utilized to identify structural RNA elements. One drawback of this approach is that it gives many different structures within the error range of the global optimum with minimal free energy. Integrated methods have been under developement for better performance and are expected to be applied to identify new functional RNA elements more precisely.

As the number and depth of available sequences and data from genome databases and high-throughput experiments are increasing, so too bioinformatics analysis need increasingly to be applied to RNA studies. Such analyses have led to great success in understanding RNA regulation. Sequence-based approaches considering sequence conservation across species have been used for studying pre-mRNA processing and identified regulatory elements for 5′ splicing sites (Lerner et al., 1980), consensus elements for alternative splicing (Castle et al., 2008) and the poly adenylation signal for 3′ end processing of transcripts (Beaudoing et al., 2000). Bioinformatics approaches are also important in identifying putative miRNAs, primarily for sequence conservation and RNA secondary structure to screen conserved sequences forming hairpin loop structures of pre-miRNAs (Bentwich et al., 2005). In addition, similar approaches are also critical in attempting to identify potential microRNA binding sites in 3′ UTRs (Rajewsky, 2006).

MicroRNAs and target regulations.

MicroRNAs (miRNA) are single stranded non-coding RNA molecules of approximately 21~23 nts (Lagos-Quintana et al., 2001). Functional miRNAs are generated from short stem-loop structures called pre-miRNA which is also processed from primary transcript called pri-miRNAs (He and Hannon, 2004;

Kim et al., 2009b). Mature miRNAs are partially complementary to targets and play an important role in down-regulate gene expression through controlling mRNA translation or degradation. Depending on the function of targets, miRNAs control a variety of cellular functions including development, immune function and neuronal biology. Dysregulation of miRNAs are reported to be associated with many diseases, such as cancer and neurodegenerative disease (Hammond, 2006; Kim et al., 2007). Therefore, idetification of targets is important to understand the function of each miRNA in cellular function and pathogenesis

-Generation of miRNAs

The genes producing miRNAs are much longer than the mature miRNAs and often contain multiple miRNAs as clusters. They may be encoded in independent genes or the parts of other genes such as introns and 3'UTRs (Murchison and Hannon, 2004). First, they are transcribed as primary transcripts called pri-miRNAs by RNA polymerase II or III (Borchert et al., 2006; Lee et al., 2004). Then, pri-miRNAs are processed into ~70 nucleoide stem-loop structures called pre-miRNAs, performed by microprocessor complex (Denli et al., 2004), consisting of RNase III enzyme Drosha (Lee et al., 2003) and Pasha/DGCR8 (Denli et al., 2004; Han et al., 2004). Processed pre-miRNAs are exported out of the nucleus via Exportin 5 in a Ran-GTP dependent manner (Bohnsack et al., 2004; Lund et al.,

2004; Yi et al., 2003). In the cytoplasm, pre-miRNAs are further processed to mature miRNAs by an RNase III enzyme called Dicer (Bernstein et al., 2001). Dicer cleaves a pre-miRNA stem-loop or a dsRNA, producing ~20-25 base pair RNA fragments with a 2 nucleotide 3' overhang (Zamore et al., 2000). Only one strand of two complementary short RNA fragments is selected to initiate the formation of the RNA-induced silencing complex (RISC). It is called the guide strand and is selected by the Argonaute (Ago) protein, a critical core component in RISC to repress translation of target mRNA (Gregory et al., 2005). Ago in the RISC complex forms an Ago-miRNA-mRNA ternary complex and shows RNase activity depending on the complementary between miRNA and target mRNA (Bartel, 2009). The remaining strand, called passenger strand, is cleaved quickly and is subsequently degraded by RISC as substrate (Matranga et al., 2005). The remaining mature miRNA recognizes complementary binding sites located in 3'UTRs of target mRNAs and functions as a guide for effectors, such as Argonaute proteins, to regulate mRNA stability or repress translation (Pillai et al., 2007).

-miRNAs and Argonaute proteins

Argonaute proteins (Ago, also called Eif2c) are the catalytic components of RISC, which responsible for the gene silencing called RNA interference (RNAi). Ago

binds small interfering RNA (siRNA) fragments and cleaves messenger RNA strands that are perfectly complementary to their bound siRNA (Carmell et al., 2002). But endogenously expressed miRNA, which are also bound by Ago, are usually partially complementary to target mRNA and regulate target gene expression via translational repression. Human has 4 Ago proteins (Ago1-4), which are ubiquitously expressed in many organisms (Hutvagner and Simard, 2008). Ago proteins are ~100kD proteins that contain three domains, namely PAZ, Mid and PIWI domains. The PAZ domain has been identified in Ago and Dicer, binding to single stranded RNAs (ssRNAs) recognizing the 3′-end of ssRNAs (Lingel et al., 2003). PAZ serves as a 3'-end binding module for siRNA transfer in the RNA silencing pathway, and as an anchoring site for the 3'-end of guide RNA within silencing effector complexes (Ma et al., 2004). The C-terminal PIWI domain, which is highly conserved, is known to have similar folding with RNaseH, suggesting that Ago proteins are the silencer in RISC complex with nuclease activity (Song et al., 2004). Among the four human AGO proteins, only AGO2 showed silencer activity and mutation in its PIWI domain abolished RISC activity (Liu et al., 2004; Meister et al., 2004). Ago ternary complexes, Ago-miRNA-mRNA, can accumulate in cytoplasmic foci, known as processing-bodies (P-bodies, cytoplasmic bodies or GW bodies), which are regions with high rates

of mRNA decay and translational repression (Jackson and Standart, 2007; Liu et al., 2005; Sen and Blau, 2005).

Recent X-Ray crystal structures of a ternary complex of Ago bound to both a 21-mer DNA guide strand (miRNA equivalent) and a 12 nt fragment of complementary RNA (target mRNA quivalent) suggest that these three molecules make close contacts (Wang et al., 2008a). The seed region (position 2 to 8) of the guide strand forms paired duplex with target RNA and both ends of the strand anchored in the Ago protein. In detail, the PAZ domain interacts with two terminal bases on the 3'end of guide DNA and a pocket between the middle domain (Mid) and the PIWI domain binds to the 5'-phosphate of the strand, widens on formation of ternary complex. Interestingly, guide-strand bases 11-18 are disordered in the crystal structure with target RNA in duplex, suggesting a high degree of mobility in this part.

-miRNA target predictions and identification.

The initial observation of a lin-4 miRNA target, lin-14 mRNA, showed some clues about the miRNA target sites. miRNA target sites are conserved sites in 3'UTR with partial complementarity to the corresponding miRNA sequence (Lee et al., 1993). Despite the increasing clues about miRNA target recognitions and

the need for target identification, it has not been easy to make a general determination about the specific RNA targets upon which they act. The problem stems from the observations that in animals most of miRNA target sites have partial complementarity. Due to the limited complementarity of target sequences, initial bioinformatics target-predictions produced great divergent results although they constrained predictions for evolutionarily conserved sequences in the 3'UTR and considered free energy of the entire miRNA-mRNA duplex (John et al., 2004; Kiriakidou et al., 2004; Lewis et al., 2003). The most significant contribution to target recognition is perfect Watson-Crick base pairing of miRNA seed region (position 2-7), and this information has been used to improve target predictions such as TargetScan (Lewis et al., 2005). The wide spread of miRNA seed matches in 3'UTR was also observed from the bioinformatics analysis with sequence conservarion, supporting the importance of miRNA seed region for its target recognition (Xie et al., 2005).

Most bioinformatics predictions of miRNA binding sites are based on miRNA seed sequences and conservation but also consider slightly different additional features to improve their performances. The PicTar algorithm uses a combinatorial approach, based on the assumption that differently co-expressed miRNAs coordinately regulate target (Krek et al., 2005). The PITA algorithm

considers target site accessibility, based on the assumption that free energy gained from the formation of miRNA-target hybridization should compensate the energy needed to unfold target RNA (Kertesz et al., 2007). A critical caveat common to all of these conservation based approaches is their inability to definitively identify species specific (non-conserved) miRNA target sites or target sites located in the regions other than the 3'UTR. In fact, sites outside of the 3' UTR can act as functional sites for miRNA regulation (Rosa et al., 2009; Tay et al., 2008). To overcome such limitations, RNA22, a pattern-based method which does not rely on conservation, was developed (Miranda et al., 2006). But it still has inherited the restriction that the initial pattern learning was trained from conserved sequences. Such bioinformatic efforts have greatly improved the ability to recognize bona fide miRNA binding sites. However, slightly different algorithms applied to the same sets of sequences still produce poor overlapping results among the predictions, which is a concern in terms of false positives and negatives (Rajewsky, 2006).

Experimental approaches combined with bioinformatics analysis also attempted to identify miRNA targets. After miRNAs were introduced into HeLa cells, microarray analysis revealed evidence for targets, showing decreased transcripts level depending on miRNA expression and enriched seed matches to

corresponding miRNAs (Lim et al., 2005). Furthermore, miRNA overexpression or knockdown studies, most recently in combination with proteomic studies, have led to the conclusion that individual miRNAs generally regulate a relatively small number of proteins (~40) at modest levels (< 2-fold) (Baek et al., 2008; Selbach et al., 2008). A limitation of such studies is their inability to definitively distinguish direct from indirect miRNA-target interactions. Biochemical methods were developed to identify the direct targets by identifying mRNAs coimmunoprecipitating with silencing complexes using microarray or sequencing (Beitzinger et al., 2007; Easow et al., 2007; Hammell, 2008; Hendrickson et al., 2008; Zhang, 2007). However those methods are not able to definitively define miRNA targets sites and raise the concern that immunoprecipitation of RNA-protein complexes may be susceptible to rearrangements and poor signal:noise ratios, requiring cross-linking strategies like CLIP (Ule et al., 2003) to get high resolution of binding sites and reduce the noise from the background. Furthermore methods based on CLIP may be able to provide new rules of miRNA target recognition beyond seed pairing.

-Mechanism of miRNA mediated post-transcriptional repression.

The mechanism of miRNA mediated post-transcriptional repression is still largely unknown. It has been believed that most miRNAs induce translational

repression rather than the degradation of target mRNAs, with exception of few targets harboring the perfect match sites to miRNA sequence (Yekta et al., 2004). Translational repression of lin-14 mRNA by lin-4 miRNA does not alter lin-14 mRNA association with polyribosomes (Olsen and Ambros, 1999), initially suggested that miRNA repression could happen after the translation initiation step. This hypothesis was further supported by the same observations with other target mRNAs (Maroney et al., 2006; Nottrott et al., 2006; Petersen et al., 2006). However, this hypothesis was then challenged by a report that target mRNAs are shifted to the top of polyribosome when their expression is repressed by miRNAs (Pillai et al., 2005). It was also observed that miRNAs can not repress M^7G-cap-independent translation via internal ribosome entry site (IRES), or recruit some translation initiation factors through RISC, suggesting that miRNAs interfere with mRNA translation at the initiation step (Chendrimada et al., 2007; Humphreys et al., 2005; Mathonnet et al., 2007; Pillai et al., 2005). In some cases miRNAs have been observed to mediate deadenylation and decay of targeted mRNAs together with its translational repression (Behm-Ansmant et al., 2006; Giraldez et al., 2006; Wakiyama et al., 2007; Wu et al., 2006). Regardless of the mechanism, it has been noted that reduced protein levels are frequently associated with decreased mRNA levels following miRNA suppression (Selbach et al., 2008) (Baek et al., 2008).

Other regulatory factors associated with RISC or target mRNA might explain the differences in such mechanisms. While incompletely understood, there is some evidence that RNABPs play role in miRNA dependant translational regulation. HuR, an AU-rich-element binding protein, binds to 3'UTR of the CAT1 mRNA after stress and the binding inhibits the CAT1 repression by miR-122 (Bhattacharyya et al., 2006b). A similar result was reported in a conserved RNABP, Dead end 1 (DnD1), which also protects mRNAs from miRNA-mediated repression through interacting with the same miRNA target (Kedde et al., 2007b). RNABPs may not only inhibit but also enhance miRNA mediated repression. It is reported that binding of HuR to 3'UTR of c-Myc recruits let-7-loaded RISC and represses c-Myc expression (Kim et al., 2009a). Furthermore, RNABPs may be involved in switching of miRNA mediated repression to activation. Upon cell cycle arrest, an AU-rich element (ARE) recruits miR-369-3 associated RISC together with the fragile X mental retardation-related protein 1 (FXR1) RNA binding protein to tumor necrosis factor-alpha (TNFalpha) mRNA and activates translation (Vasudevan et al., 2007). Taken together, these data suggest that effects of miRNA-mediated regulation on targets can be different depending on the context of RNABP associated with the target mRNA. However, a major limitation is lack of understanding of location of binding sites of both

RNABPs and miRNAs on target transcripts, which is pre-requisite to understanding mechanisms.

-Biological functions of microRNAs and targets

Since the first discovery of miRNA, lin-4, which was identified due to its critical roles in the timing of C. elegans larval development, more discoveries of miRNAs from genetic screening have implicated miRNAs in the regulation of biological processs and function (Johnston and Hobert, 2003; Lee et al., 1993). Several knockout mice of components in the miRNA processing pathway were generated. Dicer1-null mice are lethal in early development and their embryonic stem (ES) cells were not able to be generated, showing the importance of miRNA in maintaining stem cell populations during early mouse development (Bernstein et al., 2003). Conditional knockout of Dicer in Purkinje cells, neurons found in the layered structures of the mammalian cerebellum, was examined in mice and lead to cerebellar degeneration and development of ataxia by inducing Purkinje cell death, implicating the involvement of miRNAs in neurodegenerative disorders (Schaefer et al., 2007). Ago2-null mice are embryonic lethal and unable to response to siRNAs, indicating the critical role for miRNAs in mouse development and Ago2 in siRNA response (Liu et al., 2004). Loss-of-function studies of specific miRNAs in mice have also been used to elucidate their

functions and targets. miR-1-2 null mice showed heart defects with enrichment of seed matches to miR-1 in upregulated transcripts, elucidating the role of miR-1-2 in cardiogenesis (Zhao et al., 2007). By generating miR-150 null mice, it was found that miR-150 plays role in lymphocyte maturation through regulating its target, cMyb (Xiao et al., 2007).

It has also been reported that dysregulation of miRNAs are associated with human disease. miRNA profiles in cancer tissues are surprisingly informative, differentiating type and state of tumors more accurately than mRNA expression profiles (Lu, 2005). miRNAs in the miR-17-92 cluster are often amplified in human B-cell lymphomas and their expression is reported to accelerate tumour together with c-Myc as oncogene (He et al., 2005). The miR-34 family is known to be activated by p53 and to induce cell cycle arrest through target regulation, implying its role as tumor repressor together with p53 (He et al., 2007). Restoring the expression of miR-126 suppresses tumor and miR-335 inhibits metastasis through regulating its target, SOX4 (Tavazoie et al., 2008). Such reports showing tumor suppressor activity of miRNAs suggest potential therapeutic use in cancer. Interestingly, introducing of miR-26a using adeno-associated virus (AAV) into mouse model of hepatocellular carcinoma protects from disease progression, presumably through regulation of its targets, cyclins D2 and E2 (Kota et al., 2009).

Diseases other than cancer are also associated with miRNAs. miR-133b is reported to be absent in midbrain tissue of Parkinson's disease patient and have role in dopaminergic neuronal function through regulating its target, Pitx3 (Kim et al., 2007). Interestingly, mutation in the seed region of miR-96 is associated with progressive hearing loss in both human disease and mouse mutants, implicating the importance of seed region in target recognition and regulation of biological function (Lewis et al., 2009; Mencia et al., 2009)

Aim of the thesis

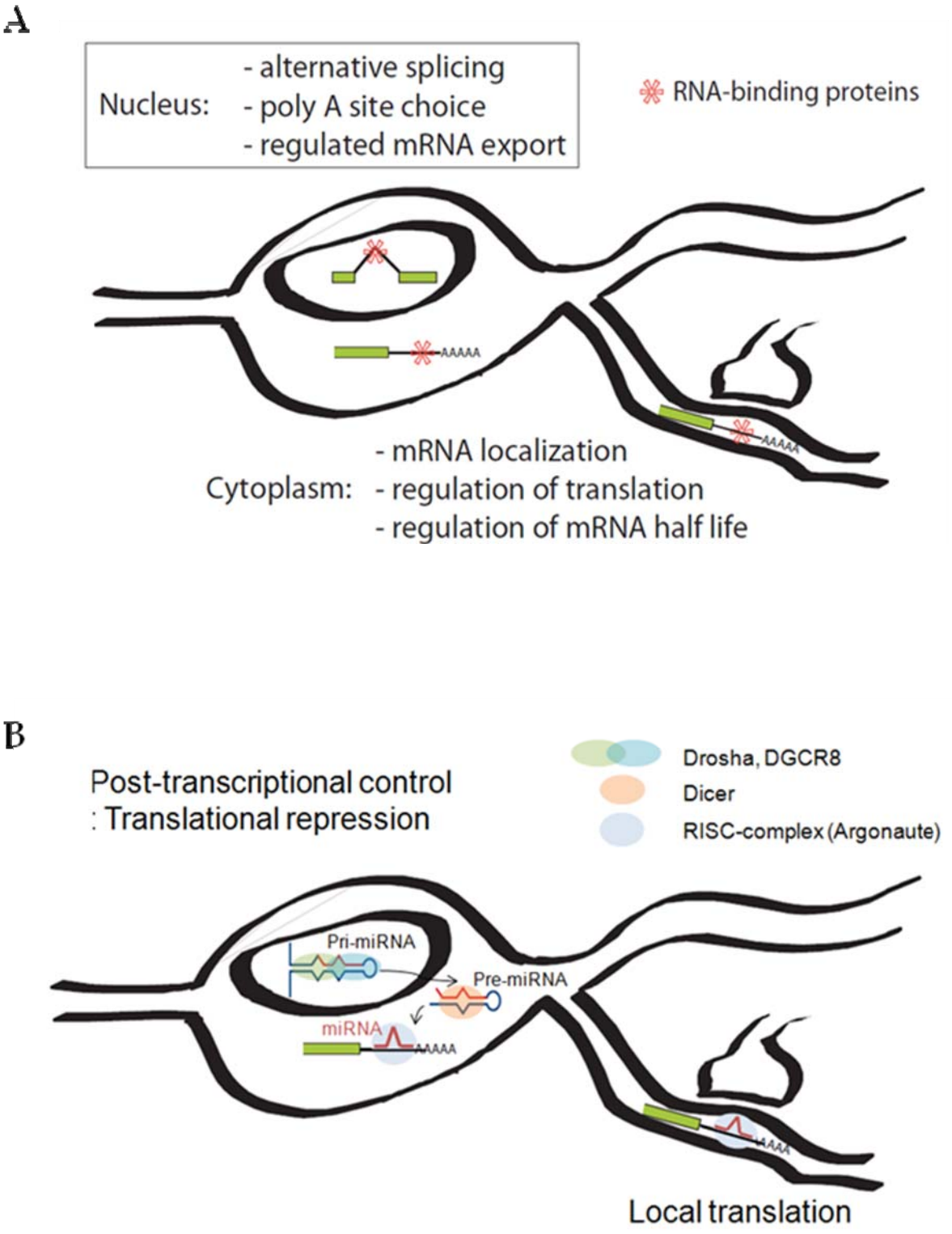
The overall goal of this study is to systematically define the functions of RNA regulation as it relates to RNA complexity. For this purpose, we have tried to develope a new genome-wide method for RNA regulation study called HITS-CLIP (High-throughput sequencing and cross-linking immunoprecipitation). We have applied this method to the study of two major RNA regulators, the RNA binding protein (Nova) and the Ago-miRNA RISC complex. Focusing on RNA regulation by miRNAs, a specific hypothesis regarding the regulatory function and mechanisms was generated and examined by combined computational-experimental approaches. Such studies provide a means to explore genome-wide RNA regulation, expand our understanding of RNA complexity and

provide advanced understanding of miRNA target regulations *in vivo*, which are critical for functional studies and also become increasingly important for RNA interference based therapeutics.

Figure 1.1. RNA complexity regulated by RNABPs and miRNAs.

A. RNA regulation by RNABPs in neuron. In the nucleus, RNABPs (red asterisks) regulate alternative splicing, alternative polyadenylation and mRNA export. through their direct interaction with primary mRNAs. In the cytoplasm, they also regulate mRNA stability, mRNA localization and local translation which occur in dendrites mostly through their interaction with 3'UTR of target mRNAs. Indeed, local translation is integral to aspects of synaptogenesis, and activity dependant neuronal function. B. Processing of miRNAs and their RNA regulations in neuron. In the nucleus, pri-miRNAs, which produced as long transcripts, are processed by Drosha (green) and DGCR8 (blue) complex (also called microprocessor) into pre-miRNAs, which are the hairpin-loop structured precursors of mature miRNAs..Then, pre-miRNAs are further processed by Dicer (orange) into mature miRNAs. In the cytoplasm, miRNAs are finally loaded on the RISC (sky blue) and recognized the target mRNAs to induce post-transcriptional gene silencing. There is a miRNA (miR-134) also known to localize to dendrite to regulate local translation in activity dependant manner like RNABPs.

Figure 1.1


A
Nucleus:
- alternative splicing
- poly A site choice
- regulated mRNA export
RNA-binding proteins
AAAAA
AAAAA
Cytoplasm:
- mRNA localization
- regulation of translation
- regulation of mRNA half life
B
Post-transcriptional control
: Translational repression
Drosha, DGCR8
Dicer
RISC-complex (Argonaute)
Pri-miRNA
Pre-miRNA
miRNA
AAAAA
AAAA
Local translation

CHAPTER II. EXPERIMENTAL PROCEDURES

Total tissue extract

Tissues were each Dounce homogenized in RIPA lysis buffer (0.5% NP-40, 0.5% deoxycholic acid, 0.1% SDS in PBS, 50% glycerol, and Complete protease inhibitor cocktail (Roche)) sonicated, and centrifuged at 20,000xg at 4°C for 15min. Protein concentration was determined by Bradford assay. 50ug of protein from each sample was boiled in SDS-sample loading buffer and continued for Western analysis.

RNA preparation

RNA from mouse tissues was extracted using Trizol LS Reagent (Invitrogen). 49:1 chloroform:isoamyl alcohol was added to facilitate aqueous and phenol-chloroform organic phase separation. After 15min 12,000xg centrifugation, the aqueous phase where RNA remained was collected and precipitated with ethanol at -20°C overnight. After precipitation, RNA was pelleted at 20,000xg for 20min at 4°C, washed with 75% ethanol, and then was dissolved in H_2O. Extracted RNA was then RQ1 DNase (Promega) treated at 37°C for 1hr and underwent a second round of phenol-chloroform extraction and ethanol precipitation.

Immunoblot Analysis

Samples were run on SDS polyacrylamide gels and transferred onto PVDF membranes (Millipore). Membranes were blocked for 1hr at room temperature in 10% non-fat milk in Western blot wash buffer (WBWB) (23mM Tris, pH 8.0, 190mM NaCl, 0.1% w/v BSA, 1mM EDTA, 0.5% Triton X-100, 0.02% SDS). Respective primary antibodies and horseradish peroxidase-labelled secondary antibodies (Jackson Immunoresearch) in 10% milk-WBWB were used during incubation from 1hr at room temperature to overnight at 4^{o}C. Blots were washed with WBWB 5min for 5 times after incubation with each antibody. Signals were detected by Chemiluminescence (Perkin Elmer) and quantified with Versadoc Imaging System (Bio-Rad).

RT-PCR and qRT-PCR

cDNA was generated by using random hexamers and Superscript III (Invitrogen). Once the number of PCR cycle was tested for the linear range, 32P-dCTP was added into the PCR reactions for the last two cycles and then run onto denaturing gels (6% polyarylamide/7M urea). It was exposed to X-ray film (Kodak). Quantitative RT-PCR (qRT-PCR) was performed by using a MyiQ single-color real time PCR detection system, using mouse tubulin as an internal control, and relative mRNA level (WT vs. DKO) was calculated by $\Delta\Delta$Ct values.

qRT-PCR was performed using Taqman (Applied Biosystems) following the manufacturer's protocol

Cell culture

All cell lines were maintained at 37ºC/5% CO2 in DMEM (Mediatech), 10% fetal bovine serum (Gemini) and pen-strep (50U/ml; Invitrogen).

Luciferase reporter assays

Dual-luciferase assays (Promega) were performed 24 h after transfection according to the manufacturer's protocol. Psi-check2 (Promega) was used for luciferase reporter and was cotransfected with a miR-124 and control miRNA from meridian miRNA Mimic (Dharmacon).Transfections were harvested 48 h after transfection, and the two luciferase activities consecutively assayed.

CLIP (*in vivo* Cross-link and Immunoprecipitation)

Ago HITS-CLIP was performed as described using a monoclonal Ago antibody, 2A8, generously supplied by Z. Mourelatos (Nelson et al., 2007). Experiments using the 7G1-1* antibody were undertaken after recognizing that this monoclonal antibody, obtained from batches generated by the Iowa Developmental Hybridoma (IDH) bank (Cat # 7G1-1), were a mixture of clones of

anti-FMRP antibody and anti-Ago antibody. Ago CLIP with the 7G1-1* antibody done in WT P13 mouse brain was in the presence of 200ug of blocking peptide (KHLDTKENTHFSQPN; mapped to FMRP amino acid 354 to 368), which has been demonstrated to block FMRP immunoprecipitation by 7G1-1 (Brown et al., 2001). 7G1-1 and 7G1-1* were obtained from the Developmental Studies Hybridoma Bank, developed under the auspices of the NICHD and maintained by the University of Iowa, Department of Biological Sciences, Iowa City, IA 52242. For immunoblot analysis, the anti-Ago 7G1-1* antibody was used to detect Ago proteins; results were confirmed independently with an independent anti-Ago antibody (Upstate:04-085). Other antibodies used (Figure 3.1) were generously supplied by others; anti-Ago1 and anti-Ago2 antibodies were from T. Tuschl, anti-FMRP (2F5) antibody was from J. Fallon. Nova HITS-CLIP was performed on mouse Nova2 WT and KO (CD1) brains as described (Ule et al., 2003). After PCR amplification, high throughput sequencing was performed (454 Life Sciences).

UV crosslinking of mouse brains

Cortex from P13 mice were harvested and sat in cold HBSS (50ml 10X Hank's balanced salt solution, Ca-Mg-free, (Invitrogen), 5ml 1M Hepes, pH 7.4, 445ml ddH_2O) until harvest was complete. Tissues were triturated with a 5ml pipette

and resuspended in 10 volumes of HBSS. Suspension was irradiated using 10ml per 10cm tissue culture plate six times for 400mJ/cm^2 in Stratalinker (Strategene model 2400). After irradiation, tissue suspension was collected and pelleted at 2500rpm 10min at 4^{o}C. Pellets were resuspended in HBSS (~2X pellet volume), distributed to Eppendorf tubes (1 brain per tube), quick pelleted at 4^{o}C. Pellets (~0.6ml/tube) were kept frozen at -80^{o}C until needed.

Immunoprecipitation

For each tube of crosslinked lysate 400ul protein A dynabeads (Dynal) were used. Beads were washed 3x with 0.1M Na-phosphate pH 8.0 and then were resuspended in 400ul of same wash buffer with 50ul of rabbit anti-mouse Fcγ bridging antibody, 2.4mg/ml (Jackson Immunoresearch), rotating at room temperature for 45min. After binding bridging antibody, beads were again washed 3x with 0.1M Na-phosphate pH 8.0 and then resuspended in the same buffer with 8ul of FMRP "specific" 7G1-1 antibody or 2ul of 2A8 , 5mg/ml rotating at 4^{o}C over night. Antibody bound beads were washed 3x with 1XPXL (0.1% SDS, 0.5% deoxycholate, 0.5% NP-40, PBS tissue culture grade, Ca-Mg-free). Each tube of crosslinked brain bits was lysed with 700ul of 1XPXL with 1X Complete EDTA-free protease inhibitor cocktail (Roche) and let sit on ice for 10min. 15ul RNAsin (Promega) and 20ul of RQ1 DNase (Promega) were added

to each tube and incubated in Thermomixer R (Eppendorf) at 37°C for 5min, 1000rpm. 10ul of 1:10,000 dilution of RNase A, 5mg/ml (USB) was used to limiting digest RNA at 37°C fro 10min, 1000rpm. Lysates were spun in pre-chilled ultra-microcentrifuge (polycarbonate tubes in TLA 120.2 rotor) at 30,000rpm for 20min at 4°C. Supernatants were carefully removed and added to prepared tubes of beads for immunoprecipitation for 1.5hrs at 4°C. Beads were washed with ice-cold buffer: 2x with 1XPXL, 2x with 5XPXL (0.1% SDS, 0.5% deoxycholate, 0.5% NP-40, 5XPBS tissue culture grade, Ca-Mg-free) and 2x with 1XPNK (50mM Tris-Cl pH 7.4, 10mM $MgCl_2$, 0.5% NP-40).

CIP treatment (on bead)

After immunoprecipitation, beads were resuspended in 80ul of 1X dephosphorylation buffer and 3ul of calf intestinal alkaline phosphatase (CIP) (Roche) were added. Beads were incubated in Thermomixer R at 37°C for 20min, programmed 1000rpm for 15sec every 4min. Beads were washed with 1x with 1XPNK, 1x with 1XPNK/EGTA (50mM Tris-Cl pH 7.4, 20mM EGTA, 0.5% NP-40) and 2x with 1XPNK.

Hot 3′ RNA linker ligation (on bead)

3′ RNA linker, L33 without 5′ phosphate, was 5′ radiolabeled with T4 phosphonucleotide kinase (PNK). 1ul of L33 no phosphate at 50pmol/ul, 15ul P^{32}-γ-ATP, 6ul of T4 PNK (NEB), and 2ul RNasin (Promega) in PNK buffer (NEB) were incubated at 37°C for 30min. The reaction was let go for additional 5min with 10ul 1mM ATP. Radiolabeled 3′RNA was then spun through a G25 column (Amersham) to remove free ATP. Each 10pmol of the labeled 3′ RNA linker were used for ligation reaction for one tube of beads, in addition, 8ul of 10X T4 RNA ligase buffer (Fermentas), 8ul of BSA (0.2ug/ul), 8ul 10mM ATP, 2ul T4 RNA ligase (Fermentas), 2ul of RNasin (Promega) and H_2O up to total volume of 80ul were added. On bead ligation reactions were incubated at 16°C for 1hr at Thermomixer R programmed 1000rpm for 15sec every 4min. After 1hr, 4ul of 3′RNA linker, L33 with 5′ phosphate at 20pmol/ul were added to each tube of reaction and incubation was continued over night. The next day beads were washed 3x with 1XPNK.

PNK treatment (on bead)

80ul of PNK mix (1ul 10mM ATP, 4ul T4 PNK enzyme (NEB), 2ul RNasin and 8ul of 10XPNK buffer (NEB), H_2O up to 80ul) were added to each tube. Mixture was incubated for 20min 1000rpm 15sec every 4min and washed 3x with 1XPNK.

SDS-PAGE and transfer to nitrocellulose

Each tube of beads was resuspended in 13ul of LDS NuPAGE loading buffer (Invitrogen), 4ul of 10X reducing agent (Invitrogen) and 13ul 1XPNK and incubated at 70°C for 10min at 1000rpm. Supernatant were taken off the beads for loading to Novex NuPAGE 10% Bis-Tris gel (Invitrogen). After running at 175V for 3hr, the protein gel was transferred to S&S BA-85 nitrocellulose using Novex wet transfer apparatus (Invitrogen). After transfer, nitrocellulose was quickly rinsed with PBS, blotted with Kimwipes, wrapped in plastic wrap and exposed to film. Most of radiolabeled unligated L33 linker would run close with the gel front. Free RNA ligated with L33 linker, which would have an average size of 70 nt, will migrate on the gel mostly between 10-30kD. Using 1:10000 RNase A dilution would generate RNAs of average size of 50 nt. As the tags contained 20 nt long L33 linker, their average size would be 70 nt. Thus with our RNase A digest, FMRP-RNA complex would migrate on average approximately 20kD higher than the molecular weight of FMRP alone due to the bound RNA tags. Because RNase digestion is random, RNA tag sizes could vary from 30-150 nt,.

RNA isolation and purification

Each band of nitrocellulose membrane was further cut into smaller pieces and proteinase K treated (200ul of 4mg/ml of proteinase K (Roche) in 1XPK buffer (100mM Tris-Cl pH 7.5, 50mM NaCl, 10mM EDTA)) at 37°C, 1100rpm for 20min. Then 200ul of 1XPK+7M urea solution were added and incubated for another 20min at 37°C 1100rpm. Finally, 400ul RNA phenol (Ambion) and 130ul of 49:1 $CHCl_3$:isoamyl alcohol were added and incubated at 37°C, 1100rpm for additional 20min. Tubes were spun at 20000xg in desktop microcentrifuge and aqueous phase is carefully taken. 0.5ul of glycogen (Ambion), 50ul 3M NaOAc pH 5.2 and 1ml of 1:1 ethanol:isopropanol were added and RNAs were precipitated overnight at -20°C.

5′ RNA linker ligation

RNAs were spun down, washed; pellet dried and dissolved in 6ul H_2O. RNA ligation was performed with 1ul 10X T4 RNA ligase buffer (Fermentas), 1ul BSA (0.2ug/ul), 1ul ATP (10mM), 0.1ul T4 RNA ligase, 3U (Fermentas), 1ul L51 RNA linker at 20pmol/ul and 6ul of RNA in H_2O at 16°C overnight.

DNase treatment and RT-PCR

To the ligation reaction, 77ul H_2O, 11ul 10X RQ1 DNase buffer, 5ul RQ1 DNase (Promega) and 5ul RNasin (Promega) were added and incubated at 37°C for 20min. 300ul H_2O, 300ul RNA phenol (Ambion) and 100ul $CHCl_3$ were added, vortexed, spun and aqueous layer taken. RNAs were precipitated with 50ul 3M NaOAc pH 5.2, 1ul glycogen (Ambion) and 1ml 1:1 ethanol:isopropanol over night at -20°C. Next day, RNAs were spun down, pellet was washed, dried, and resuspended in 8ul of H_2O. 8ul of RNA were mixed with 2ul of P33 at 5pmol/ul, incubated at 65°C for 5min, chilled, and spun. 3ul 3mM dNTPs, 1ul 0.1M DTT, 4ul 5X SuperScript RT buffer, 1ul RNasin, and 1ul SuperScript III (Invitrogen) were added and incubated at 50°C for 30min, 90°C for 5min and left at 4°C. PCR reactions were performed with 27ul Accuprime pfx supermix (Invitrogen), 1ul P51 primer, 5pmol/ul, 1ul of P33 primer, 5pmol/ul and 2ul of the RT reaction, cycled 20-25 cycles with 95°C 20sec, 61°C 30sec, and 68°C 20sec.

Polyacrylamide gel separation of RNA

A 10% denaturing polyacrylamide was poured and the entire PCR reaction was run. 3ul Amplisize Molecular Ruler (Biorad) were used. To visualize DNA, the gel was immersed in 10000-fold dilution of SYBR Gold (Molecular Probes) in TBE for 10min. DNA of 60-100 nts was cut out. DNA was extracted with Qiaex II kit

(protocol for polyacrylamide gel) (Qiagen) and resuspended in 15ul of H_2O. Finally they were re-PCR with 454 Adapter Sequences or Solexa Fusion Primers.

Linker and primer sequences

RNA linkers (from Dharmacon):

R51: 5′-OH AGG GAG GAC GAU GCG G 3′-OH

R33: 5′-P CGG UUG CGA GGU GAG UGA A 3′-puromycin

R33 no P: 5′-OH CGG UUG CGA GGU GAG UGA A 3′-puromycin

454 Adapter Sequences:

Adapter A: 5′-GCCTCCCTCGCGCCATC-3′

Adapter B: 5′-GCCTTGCCAGCCCGCTCAG-3′

DNA primers (from Operon):

P51: AGGGAGGACGATGCGG

P33: CTTCACTCACCTCGCAACCG

Solexa Fusion Primers (from Operon):

DSFP5:5'AATGATACGGCGACCACCGACTATGGATACTTAGTCAGGGAGG

ACGATGCGG-3'

DSFP3: 5'-CAAGCAGAAGACGGCATACGACCGCTGGAAGTGACTGACAC-3'

Solexa Sequencing Primer (from Operon):

SSP1: 5'-CTA TGG ATA CTT AGT CAG GGA GGA CGA TGC GG-3'

Ago CLIP in miR-124 transfected HeLa cell.

HeLa cells were transfected using Lipofectamine 2000 (InVitrogen) in 100 mm2 plates with 75 nM RNA duplexes, miR-124 and control miRNA from meridian miRNA Mimic (Dharmacon). After 24hr, cells were harvested and CLIP. Experiments were performed in two biological replicates using either 2A8 or 7G1-1 antibodies. Degenerate barcodes (4 nucleotides tags) were introduced in the 5' fusion linker to increase the complexity in unique tags and to avoid artifacts from preferential PCR duplication.

Exon arrays in P13 mouse brain.

For estimating the level of brain transcripts, total RNAs from three P13 mouse brains were extracted using Trizol and RNAeasy kit and mRNA was amplified and labeled by the method provided by Affymetrix. Mouse MoEx 1.0ST Arrays were used for measuring signal intensity of each exon in the samples. To process the signals from the array, quantile normalization and PM-GCBG (signal adjustment based on the background with similar GC content) were applied. The

IterPLIER method was used for selecting appropriate "core exon" probes to estimate gene-level intensities. The presence of transcript in the P13 brain was determined by the p-value derived from DABG method (detection above background, $P<0.05$) and genes with more than 100 in final total probe intensity were selected. Finally, median $\log_2$ values from three biological replicates are used to estimate level of transcripts in the P13 brain. Analyses were performed by using Affymetrix Power Tools.

HITS-CLIP data analysis

General bioinformatic analysis of CLIP tags.

CLIP tags were aligned to mm8 genome using BLAT and to miRNAs from miRBASE (http://microrna.sanger.ac.uk) using BLAST, further visualized and analyzed with UCSC genome browser (http://genome.ucsc.edu/) and Galaxy (http://galaxy.psu.edu). General bioinformatics analysis including *in silico* random CLIP were performed as Python script utilizing BioPython (http://biopython.org); linear regression analysis for motifs by MatrixREDUCE (http://bussemaker.bio.columbia.edu/software/MatrixREDUCE/); microarray data for miR-124 targets from GEO database (http://www.ncbi.nlm.nih.gov/geo); microarray analysis by BioConductor (http://www.bioconductor.org/); statistical tests by Scipy (http://www.scipy.org/).

Normalization of CLIP tags using in silico random CLIP algorithm.

Assuming that number of CLIP tags from transient nonspecific protein-RNA interactions are correlated with transcript abundance, then each transcript should have different threshold level to differentiate signal:noise. Therefore we estimated the false discovery rate (FDR) of CLIP tag clusters for each brain transcript using a simulation method named *in silico* random CLIP. Transcript abundance was measured by exon arrays (Affymetrix MoEx 1.0 ST, at the same age (P13) and tissue (mouse brain) used for HITS-CLIP). For simulating transcript abundance, a number for each brain transcript (N_n) was assigned using the normalized probe intensity from the microarray. RNase treatment was simulated by introducing cleavage at a random site in each population of transcript in given length (Y_n; determined by RefSeq annotation) and this process was repeated until an average length of 50 +/- 2 nt fragments was obtained. 50 nt was determined based on the observation of mean 50 nt size of Ago-mRNA CLIP tags. When normalized to length and abundance, the number of fragments per transcript $F_n = N_n \times Y_n / 50$. The total number of fragments per transcriptome Tot(F) is the sum of F_n for all transcripts.

To simulate immunoprecipitation of nonspecific and transient RNA-protein interactions, we first calculated the expected number of tags for a transcript

based on F_n and the total number of unique CLIP tags in that experiment. Thus for a given CLIP experiment, we calculated the expected number of tags in any given transcript (T_n), by multiplying total number of unique tags from actual CLIP experiments (Z) with fraction of fragments for that transcript (F_n/Tot(F)). We then randomly selected T_n fragments for each transcript, and these were aligned with their position in the given transcript. If the selected fragment was more than 36 nt, the 3′ end of the fragment was eliminated to leave only 36 nt to simulate an Illumina sequencing read. We calculated the maximum background cluster height (M_n, maximum number of overlapping tags) from the alignment, and repeated this simulation repeated 500 times for every transcript. FDR (P-value) was determined by counting the observed number of maximum clusters (M_n) from each of 500 repeats, and the cluster height giving $p<0.01$ in each transcript was used as threshold for normalization of HITS-CLIP cluster. In this way, the normalization for HITS-CLIP varied for each transcript according to its abundance, length, and its simulated background cluster height. The alogorithm was implemented in python script by utilizing the cubic spline modules in Scipy (interpolate.splrep(), interpolate.splev()).

Ranking of Ago-miRNA CLIP tags.

The ranking of miRNAs is based on the frequency of each ~22 nt sequences in Ago-miRNA tags. Where multiple members of family of miRNAs have identical seeds, we did analyses of the family in aggregate, summing the frequency of each member.

Normalization of Ago-mRNA clusters along transcripts

To normalize for differences in the length of transcripts in the CDS and 3′ UTR, the number of clusters at each position was divided by the total number of transcripts at that position (blue graph). The standard deviation of fraction in each position was estimated based on the assumption that prediction result is binomial distribution (drawn as light blue graph). To determine whether the different length of transcripts biased the position of clusters, the position of a cluster within each transcript was randomly redistributed.

Distribution of tags relative to cluster peaks

We plotted the distribution of Ago-mRNA tags within clusters. To compensate for differences in the number of CLIP tags present in different clusters, we randomly chose 30 tags from each of the 61 most robust clusters (single peak, BC = 5, and $\geq$ 30 tags per peak) or 135 robust clusters (BC=5, $\geq$ 30 tags per peak). The

peak of each cluster was determined by cubic spline interpolation analysis, and the distribution of tags relative to the peak was plotted). For the distribution of miRNA binding sites in 135 robust clusters, conserved seed sites (position 2-7) of top 30 miRNAs and bottom 30 miRNAs were searched in Ago CLIP cluster regions and plotted as the relative position from the peaks. If width of an Ago CLIP cluster was more than 100nt from the peak, we restricted the search within +/- 100nt window from the peaks.

To determine the distribution of conserved miR124 seed sites (6mers in position 2-8) in Ago-mRNA clusters (BC $\geq$ 2), we first determined the number of clusters as a function of relative position from the peak. Because of the variability in the width of AGO CLIP cluster, the number of predicted conserved mir-124 target sites in each relative position from the peak of the cluster was divided by the number of clusters in each position and indicated as fraction. Standard deviation of fraction in each position was estimated based on the assumption that prediction result is binomial distribution

Cluster peak analysis using cubic spline interpolation.

Based on the position of tags in genome, the number of unique tags in each nucleotide position was calculated for all AGO-mRNA CLIP clusters. Cubic

spline interpolation method was applied to interpolate tag density in the clusters by using Scipy (http://www.scipy.org/). By determining the derivative of the function at each point of the interpolation, locating the point where the derivative = 0, and confirming that the derivative changes from positive to negative around this point, the location and number of peaks per cluster were determined. Excess kurtosis was determined using Scipy (http://www.scipy.org/).

Prediction of miRNA binding sites

Based on the observation that enriched 6-8 mer seed sequences were present in the Ago footprint region, we identified all 6-mer sequences in miRNA seed position 1-8 to identify candidate 6-8 mer miRNA binding sites. For selecting conserved sites, Multiz8way results from UCSC genome browser (http://genome.ucsc.edu) were used to search for seed sites that were conserved across more than 4 of 5 species (human, mouse, rat, dog, chicken). Predicted miRNA binding sites were downloaded from databases available from the following websites, using precompiled batch results and default parameters set by the developers of each algorithm; miRBase (http://microrna.sanger.ac.uk/ , miRanda algorithm, Sep 2008 version), TargetScan 4.1 (http://www.targetscan.org/ downloading mm8 bed format files from the www

site), PicTar4 (predictions in human were obtained from UCSC genome browser and converted to mouse by the liftOver program), PITA (http://genie.weizmann.ac.il/pubs/mir07/index.html, TOP prediction), and RNA22[3] (http://cbcsrv.watson.ibm.com/rna22.html, precompiled data for the 3'UTR of mouse transcripts (with parameters: G=0 M=14 E=-25Kcal/mol) were mapped to mouse genome by ELAND (Efficient Large-Scale Alignment of Nucleotide Databases; provided by Illumina)).

Meta-analysis of microarray studies.

Raw data or normalized data from 5 microarray studies in miR-124 overexpression (Table 5.4) were obtained from GEO database (http://www.ncbi.nlm.nih.gov/geo). Raw data were normalized by the same method used in each study and finally log2 ratio for each transcript, which is the median value in all replicates and probes, was calculated only if its change is statistically significant ($P< 0.05$, t-test). For the meta-analysis of data to compare it with Ago CLIP, only transcripts showing robust change ($P<0.05$) at least 2 times in 5 studies without any significant opposite fold change were selected. Among these transcripts, 1278 brain expressed transcripts were further selected based on exon array result in P13 mouse brain (10743 genes, $p<0.05$ (DABG), normalized probe intensity >100). To compare the cumulative fraction depending on fold

change, Kolmogorov-Smirnov test was performed using Scipy (http://www.scipy.org/).

Gene ontology analysis

Gene ontology analysis was done using GoMiner (http://discover.nci.nih.gov/gominer/) or using DAVID (http://david.abcc.ncifcrf.gov/).

CHAPTER III. GENOME-WIDE ANALYSIS OF RNA-PROTEIN INTERACTIONS FOR NOVA

Introduction

Nova is an onconeural antigen targeted in patients with paraneoplastic opsoclonus myoclonus ataxia (POMA), a syndrome is characterized by abnormal motor inhibition (originally termed Ri) (Buckanovich et al., 1996). Nova contains three functional KH-type RNA-binding domains and functions as a brain specific splicing factor which regulates the splicing of target RNAs (Buckanovich and Darnell, 1997). A specific sequence element, YCAY tetramer (where Y indicates a pyrimidine, U or C) was identified as a Nova RNA ligand by *in vitro* RNA selection (Jensen et al., 2000b), X-ray crystallography (Lewis et al., 2000) and further confirmed by CLIP (*in vivo* cross-linking and immunoprecipitation) (Ule et al., 2003). In addition, studies with custom splicing microarrays from Affymetrix elucidated that Nova regulates the splicing of target RNAs encoding proteins that function mostly at the synapse (Ule et al., 2005b).

We have taken a different approach towards understanding protein–RNA interactions for studying Nova. A major breakthrough of the studies is

developing a crosslinking protocol that works in tissues, and can therefore be applied before protein purification (Jensen and Darnell, 2008; Ule et al., 2005a; Ule et al., 2003). This method, called "CLIP", involves covalent trapping of RNA-protein complexes in intact cells by irradiating them with UV-B light (~256 nM), and subsequently purification of the complexes of which RNA is digested into the size of a small nucleic acid fragment called "tag" under stringent conditions (Ule et al., 2003). CLIP has been used to study direct protein–RNA interactions extant in living cells, including identification of RNA targets for Nova and the discovery of hnRNPA1-dependent regulation of a microRNA (Guil and Caceres, 2007).

To comprehensively understand functional network of Nova target RNAs and yield genome wide insight into Nova-RNA interactions, a new high-throughput target identification system is attempted to be developed based on the application of CLIP method to high-throughput sequencing in this study. This method uses newly available pyrosequencing technology that sequences all tags isolated from CLIP procedure, by reading a fibre-optic slide of individual wells, where each tag is amplified using an emulsion method and is extended for sequencing in the solid support (Thomas et al., 2003). This method can read 25 million bases, at 99% or better accuracy, in one four-hour run. Here we

demonstrate that this approach confirms and generalizes the previous knowledge of Nova biology by generating massive information about *in vivo* target sites with high resolution and further uncovers global insight into RNA regulations by Nova, which regulates processing of some miRNAs. Development of HITS-CLIP is important in that it can be used as general platform to indentify functional RNA-protein interactions and yield genome-wide insight for the discovery of new biology.

Results

Development of HITS-CLIP for Nova.

To comprehensively understand genome-wide Nova action on target RNAs, a new high-throughput target identification system was developed based on massive parallel sequencing called HITS-CLIP (high-throughput sequencing CLIP; Figure 3.1). The Nova HITS-CLIP experiment, performed in P13 mouse brain, identified 15,300 unique tags from the ~75 kD band of Nova-RNA complex, 11,701 unique tags from the ~90kD band of high molecular weight Nova-RNA compelx (Super Nova-RNA complex), and 43,445 control unique tags from the supernatant after immunoprecipitation, which can be used as a control set (Figure 3.2). Derivatives of YCAY motif were identified as the most enriched

motifs (among all possible 4mers, 5mers and 6mers including degeneracy letter) in Nova HITS-CLIP tags by comparing its frequency in Nova sets with it in control sets or random shuffled sequence sets (Figure 3.3). For further identification, two critical values were calculated and considered in bioinformatics analysis. One is the log-odd ratio value derived from calculating motif frequency in Nova tags divided by the frequency in control tags. The other is the percentage of tags containing each specific motif (Figure 3.4). Interestingly, bioinformatics results of the motif identification showed few enriched motifs without YCAY, and most motifs were subset of an YCAY cluster (Figure 3.4C). In addition, bioinformatics analysis of YCAY motifs in Nova tags supports the hypothesis that the RNA binding element of Nova forms as YCAY clusters, multimers of the YCAY tetramer (Dredge and Darnell, 2003) (Figure 3.5). Taken together, these data support the biological relevance of Nova HITS-CLIP data to our previous knowledge of Nova biology, indicating that HITS-CLIP can be used as a platform to identify comprehensive *in vivo* binding sites of a given RNA binding protein and yield genome-wide insight into RNA regulations.

Nova HITS-CLIP tags are enriched in introns and non-coding RNAs

After alignment of Nova HITS-CLIP tags to the mouse genome (mm8, Build 36 assembly by NCBI and the Mouse Genome Sequencing Consortium), the

distribution of tags in the genome was examined based on the RefSeq and KnownGene annotation. Annotation of the tags (Figure 3.6A) showed that most tags (54%) are from introns, consistent with results from analysis of Nova's role in splicing. 19% of tags are from exonic regions and among them 15% of tags are from 3'UTRs, where Nova might regulate processing of 3'UTR such as polyadenylation, translation, or stability of the transcripts. The remaining tags are from ESTs (12%), repeated sequences (10%) or non-annotated regions (4%), which can be putative non-coding RNAs having important regulatory roles. In total 26% of tags are from putative non-coding RNAs, which is the second most enriched categories after intron (54%), suggesting that Nova may have function in non-coding RNA regulation (Figure 3.6A). Interestingly, a maternally expressed non-coding RNA, gtl2 has the highest number of HITS-CLIP tags (471 reads/100bp, 800 total tags), including downstream tags in many non-coding RNA clusters including miRNA clusters (Seitz et al., 2004) and C/D snoRNA clusters (Figure 3.6B-D). A genome wide YCAY cluster search was performed by measuring YCAY density (reads/100bp) and the highest peak of YCAY density in gtl2 intron overlaps with the highest peak of the tags (reads/100bp), suggesting specific binding and perhaps alternative splicing of gtl2 (Figure 3.7A). By performing RT-PCR analysis of brain RNAs from Nova DKO (Nova1 -/- and

Nova2 -/-) mice versus wild type mice, the 3'end exon of gtl2 was confirmed to be alternatively spliced depending on Nova expression (Figure 3.7A-B).

Nova regulates processing of some miRNAs in gtl2 region

Among miRNAs in gtl2 region, mir-770 locates in the gtl2 intron. To assess the effect of Nova on miR-770 levels, Q-PCR (quantitative real-time PCR)(Chen et al., 2005; Shi and Chiang, 2005) was used to measure each splicing form of gtl2 as well as each processed form of miR-770 (unprocessed pri-miR-770, processed pri-miR-770 and mature miR-770), generated during miRNA processing (Figure 3.7C-D). Every form generated during mir-770 processing was down-regulated in forebrain of Nova2 KO mice, showing that it correlates with the down-regulation of the gtl2 splicing variant harboring miR-770 in Nova2 null mice (Figure 3.7E). The process of this miR-770 containing variant may be enhanced by two regulatory mechanisms, Nova dependent alternative splicing and polyadenylation (Figure 3.7). Mir-380, another miRNA in a second miRNA cluster downstream of the gtl2 transcript, was also investigated by the same experimental method (RT-PCR and Q-PCR; Figure 3.9). Surprisingly, unprocessed pri-mir-380 is increased but mature mir-380 is decreased in the Nova2 KO, suggesting that Nova may regulate processing step of pri-mir-380, known to be performed by Drosha complex (Lee et al., 2003) (Figure 3.9). This

result correlates with recent observation of post-transcriptional regulation in miRNAs processing, which has been implicated in cancer (Thomson et al., 2006). In adition, more miRNAs in the gtl2 region (~20 miRNAs) or several regions other than gtl2 (~15 miRNAs) were found to have Nova-CLIP tags in their vicinity (≤1kb). Therefore, the effect of Nova on the processing of these miRNAs is worth further investigation by the same validation approaches, with the potential to find general rules for Nova dependant regulation of miRNA processing

Discussion

Genome-wide screens have been used to suggest correlations between the action of RNABPs and biologic diversity (Keene, 2007), but are unable to identify direct sites of RNA regulation. HITS-CLIP provides a general solution to this problem by generating a transcriptome-wide biochemical "footprints" of protein-RNA interactions in living tissues. This in turn allows a direct comparison of predicted (e.g. microarray or bioinformatically derived) and observed (HITS-CLIP) sites of action, and thereby provides a new platform for deriving functional RNABP maps. HITS-CLIP extends our transcriptome-wide understanding of Nova-RNA interactions, which was previously limited to bioinformatic analysis of YCAY

clusters within 200nt of alternate or bounding constitutive exons (Ule et al., 2006). Importantly, bioinformatics analysis on Nova HITS-CLIP map strongly correlates with our previous findings about Nova, suggesting that the map is sufficiently robust to predict protein-RNA regulation.

The unbiased nature of HITS-CLIP led to the unexpected identification of Nova binding in non-coding RNAs, including primary transcripts encoding miRNAs and the recognition of its role in regulating processing of miRNAs in the brain. The presence of such processing factors was postulated after the recognition of discrepancy between pri-miRNA and mature miRNA level and observation that post-transcriptional regulation in miRNAs processing is regulated in caner cell (Thomson et al., 2006).

In summary, HITS-CLIP offers a powerful new platform for studying RNA regulation *in vivo*. This genome-wide biochemical approach complements bioinformatic, microarray and genetic studies. HITS-CLIP is able to identify biologically relevant interactions, providing a focus on direct protein-RNA contacts as critical points for understanding RNABP function. The unbiased nature of the platform holds the potential for new discovery, including regulated RNA substrates. Identifying Nova as the potentail vertebrate factor to regulate

miRNA processing in mouse brain demonstrates that a single factor can regulate different aspects of tissue-specific RNA and miRNA metabolism

Contributions

Bioinformatic analysis, isolation of RNA from Nova-/- mice, RT-PCR and Q-PCR were done by Sung Wook Chi. Nova HITS-CLIP experiments were performed by Aldo Mele. HITS-CLIP analysis was developed by Sung Wook Chi, Robert B. Darnell, Donny D. Licatalosi, Jennifer C. Darnell and Aldo Mele. Jernej Ule discovered that gtl2 alternative splicing was Nova-dependent.

Figure 3.1. HITS-CLIP procedure.

CLIP takes advantage of the ability of UV-irradiation to directly penetrate intact cells or tissues and induce covalent crosslinks between RNA and proteins that are in direct contact (~1 Ångstrom distance). Once covalently bound, complexes can be purified under harsh conditions, with three key advantages: these conditions separate closely bound RNABP-RNABP complexes, reassociated RNAs, and background RNA. After purification, the CLIP method utilizes proteinase K to remove the RNABP, linker ligation and RT-PCR to analyze RNA sequencing, which can be done using high throughput sequencing methods "HITS-CLIP". (CLIP procedure figures are attributed to Jernej Ule)

Figure 3.1

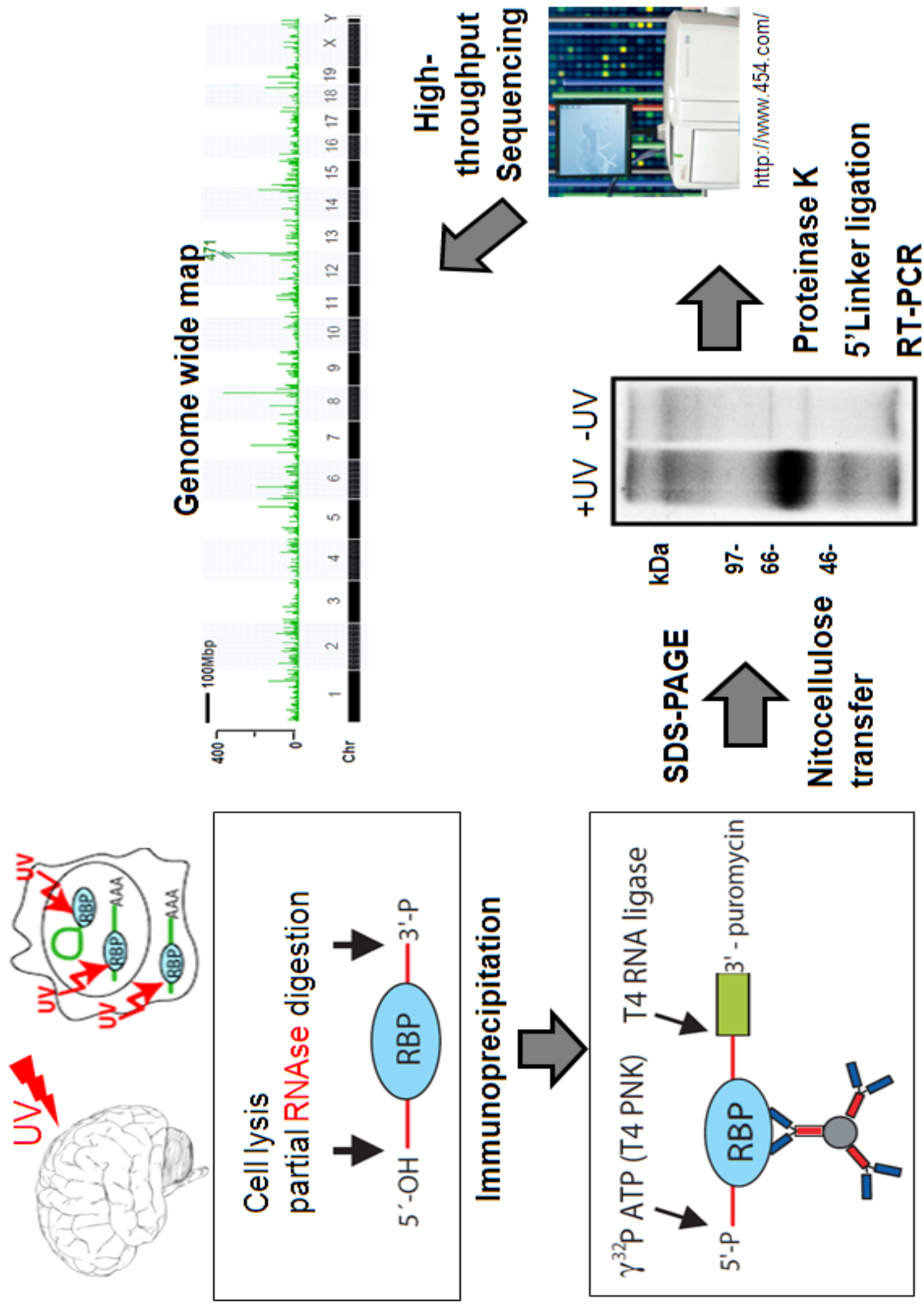
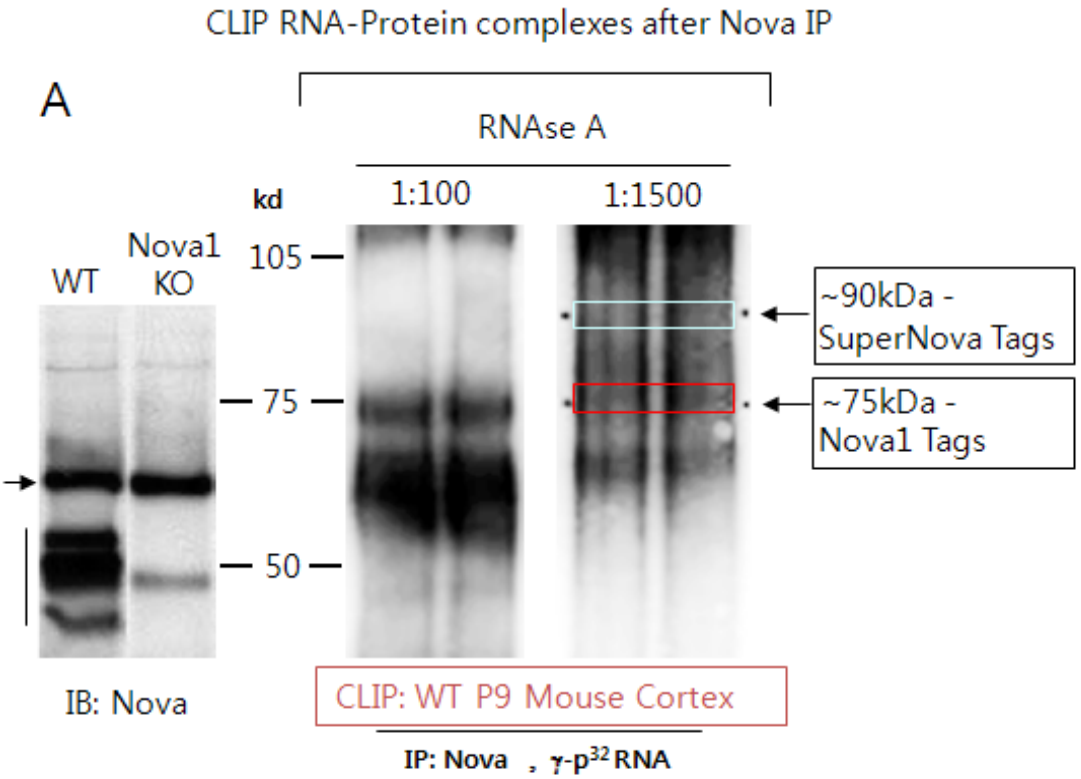

UV
UV
RBP
AAA
Cell lysis
partial RNAse digestion
5´-OH
RBP
3'-P
Immunoprecipitation
γ32P ATP (T4 PNK)
T4 RNA ligase
3' - puromycin
5'-P
RBP
SDS-PAGE
kDa
97-
66-
46-
Nitocellulose
transfer
+UV -UV
Proteinase K
5'Linker ligation
RT-PCR
High-
throughput
Sequencing
http://www.454.com/
Genome wide map
100Mbp
400
0
Chr
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 X Y

Figure 3.2. Nova HITS-CLIP

A. Postnatal day 9 mouse cortex tissue was UV irradiated, protein-RNA complexes immunoprecipitated with Nova antiserum and confirmed the size of two forms of Nova2 (~50kD Nova and ~70kD SuperNova). RNA-protein complexes from immunoprecipitation labeled with ^{32}P and separated by SDS-PAGE and transferred to nitrocellulose and visualized by autoradiography. CLIP revealed ~90kD SuperNova-RNA complex and ~75kD Nova1-RNA complex. B. From the nitrocellulose membrane, the two bands indicated with squared in panel A were cut out, the RNA purified, and amplified with RT-PCR.

Figure 3.2


CLIP RNA-Protein complexes after Nova IP
A
RNAse A
1:100
1:1500
kd
WT
Nova1 KO
105
75
50
~90kDa - SuperNova Tags
~75kDa - Nova1 Tags
IB: Nova
CLIP: WT P9 Mouse Cortex
IP: Nova , γ-p32 RNA

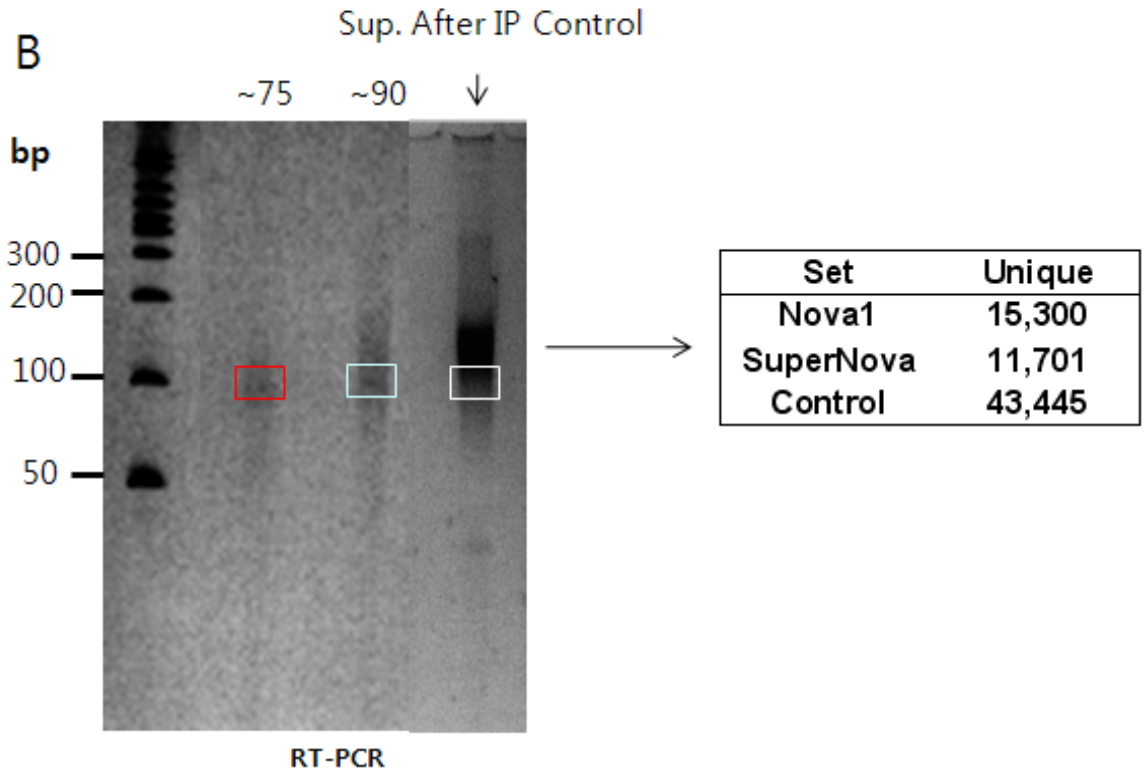
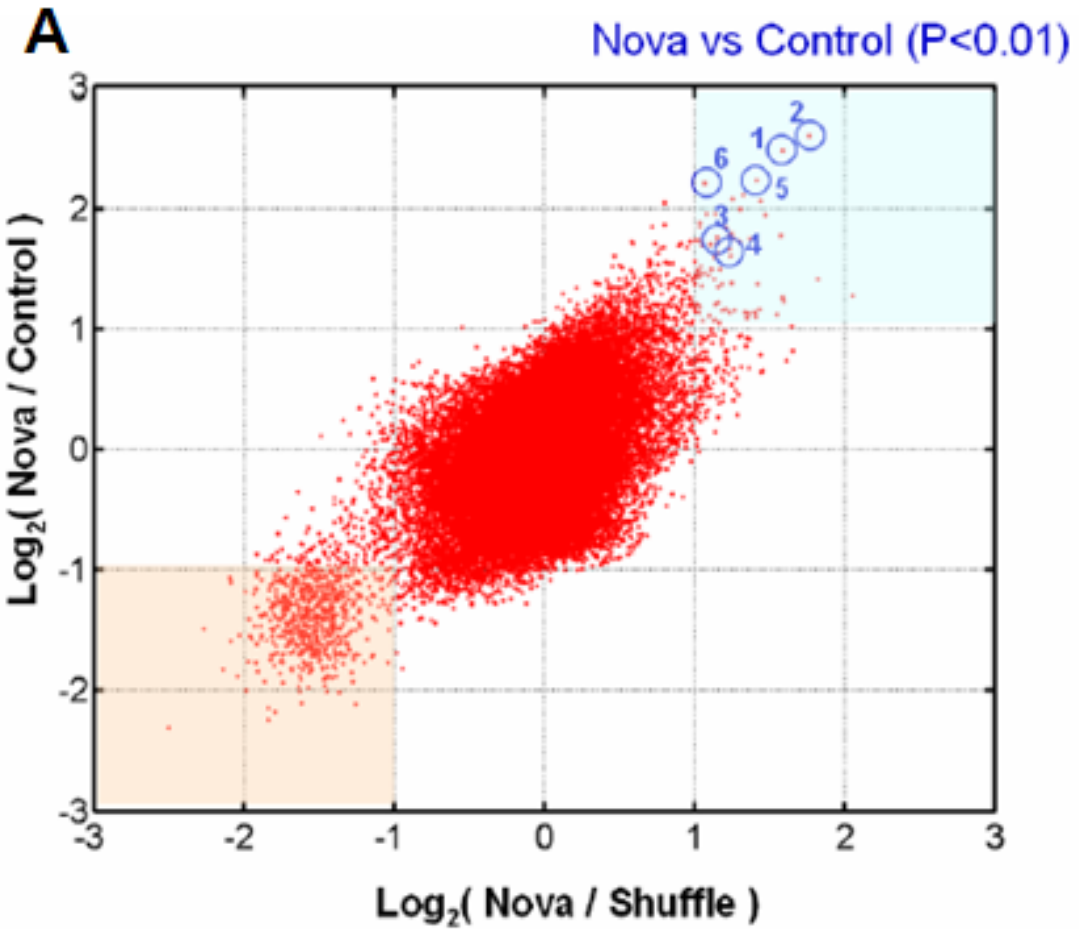

Sup. After IP Control
B
~75
~90
bp
300
200
100
50
RT-PCR
Set | Unique
Nova1 | 15,300
SuperNova | 11,701
Control | 43,445

Figure 3.3. Identification of enriched motifs in Nova HITS-CLIP tags.

A.The log-odds ratio of the number of the motif in Nova HITS-CLIP tags to the number of the motifs in shuffled sequence or control tags were plotted for all combinations of 4mer motifs including degenerate letters (n = 50626). Statistically significant motifs enriched in Nova HITS-CLIP tags are indicated as blue circle (Log2(Nova/Shuffle) > 1, Log2(Nova/Control) > 1, P < 0.01). B. Top 6 ranked motifs depending on p-value.

Figure 3.3



B

Rank	Motif	P-value	HITS-CLIP (%)
1	YCAY_	0.0057	78
2	_CAYY	0.0072	73
3	_CATH	0.0091	64
4	YCAT_	0.0085	60
5	_CATY	0.0045	58
6	_CATC	0.0053	39

Figure 3.4. Identification of enriched motifs in Nova HITS-CLIP tags.

A. The log-odds ratio of the number of the motif in Nova HITS-CLIP tags to the number of the motifs in control tags were plotted in x-axis and percentage of nova tags containing the motifs were plotted in y-axis for all the combination of 5mer motifs. Statistically significant motifs enriched in Nova HITS-CLIP tags are indicated as blue dots. Top 25 ranked motifs containing an YCAY motif B and without YCAY motif C.

Figure 3.4

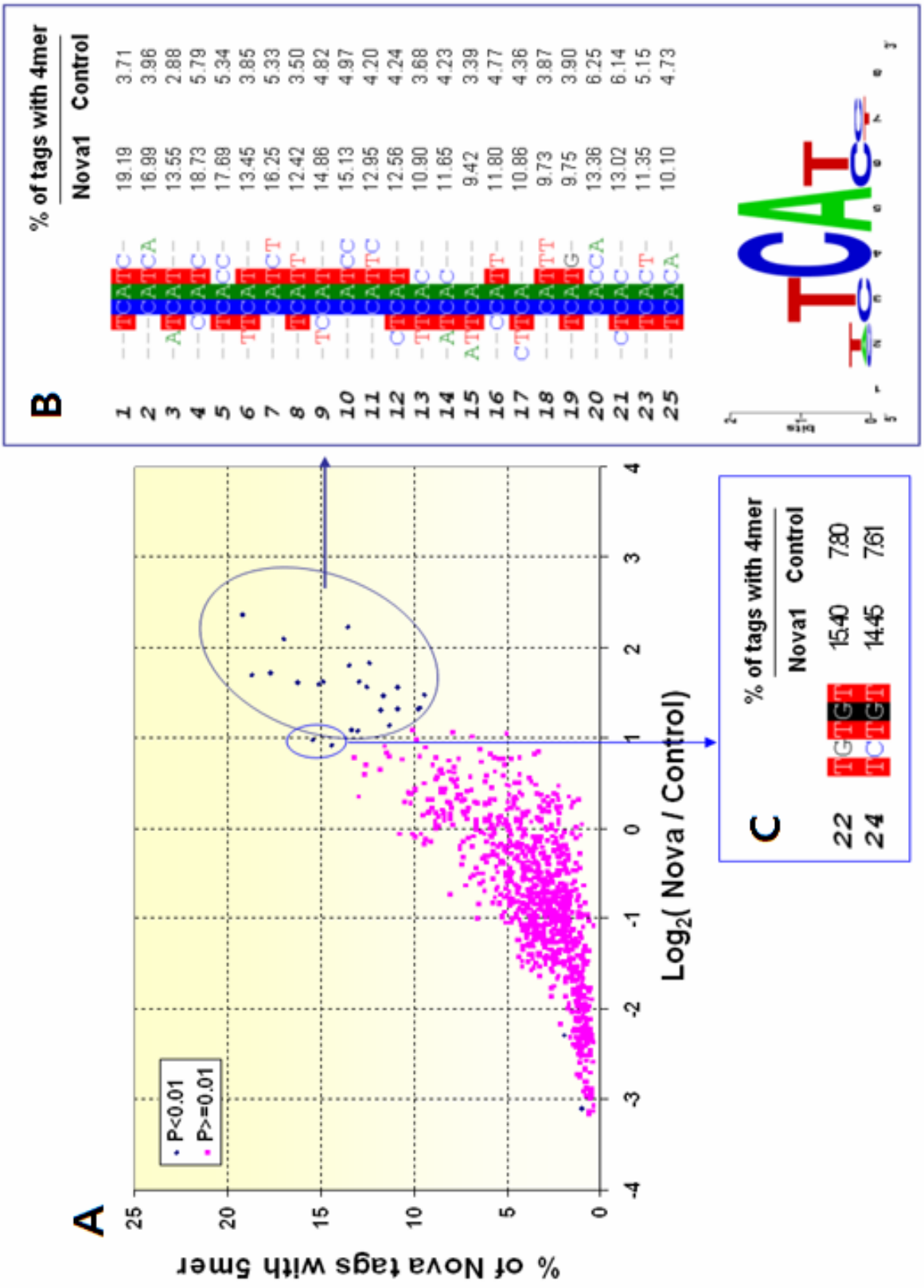

A
% of Nova tags with 5mer
Log2(Nova / Control)
P<0.01
P>=0.01
B
% of tags with 4mer
Nova1 Control
1 19.19 3.71
2 16.99 3.96
3 13.55 2.88
4 18.73 5.79
5 17.69 5.34
6 13.45 3.85
7 16.25 5.33
8 12.42 3.50
9 14.86 4.82
10 15.13 4.97
11 12.95 4.20
12 12.56 4.24
13 10.90 3.68
14 11.65 4.23
15 9.42 3.39
16 11.80 4.77
17 10.86 4.36
18 9.73 3.87
19 9.75 3.90
20 13.36 6.25
21 13.02 6.14
23 11.35 5.15
25 10.10 4.73
C
% of tags with 4mer
Nova1 Control
22 15.40 7.80
24 14.45 7.61

Figure 3.5. Identification of enriched motifs in Nova HITS-CLIP tags.

A. Percentage of the tags containing different numbers of YCAYs are calculated in Nova, control and shuffled HITS-CLIP tags. B. The shortest distance between YCAY motifs are calculated in the tags including Nova,control and shuffle HITS-CLIP tags. 65.7% of the Nova tags contains YCAY cluster with less than the 5 nucleotide distance. (34.1% in the control and 38.1% in the shuffled tags)

Figure 3.5

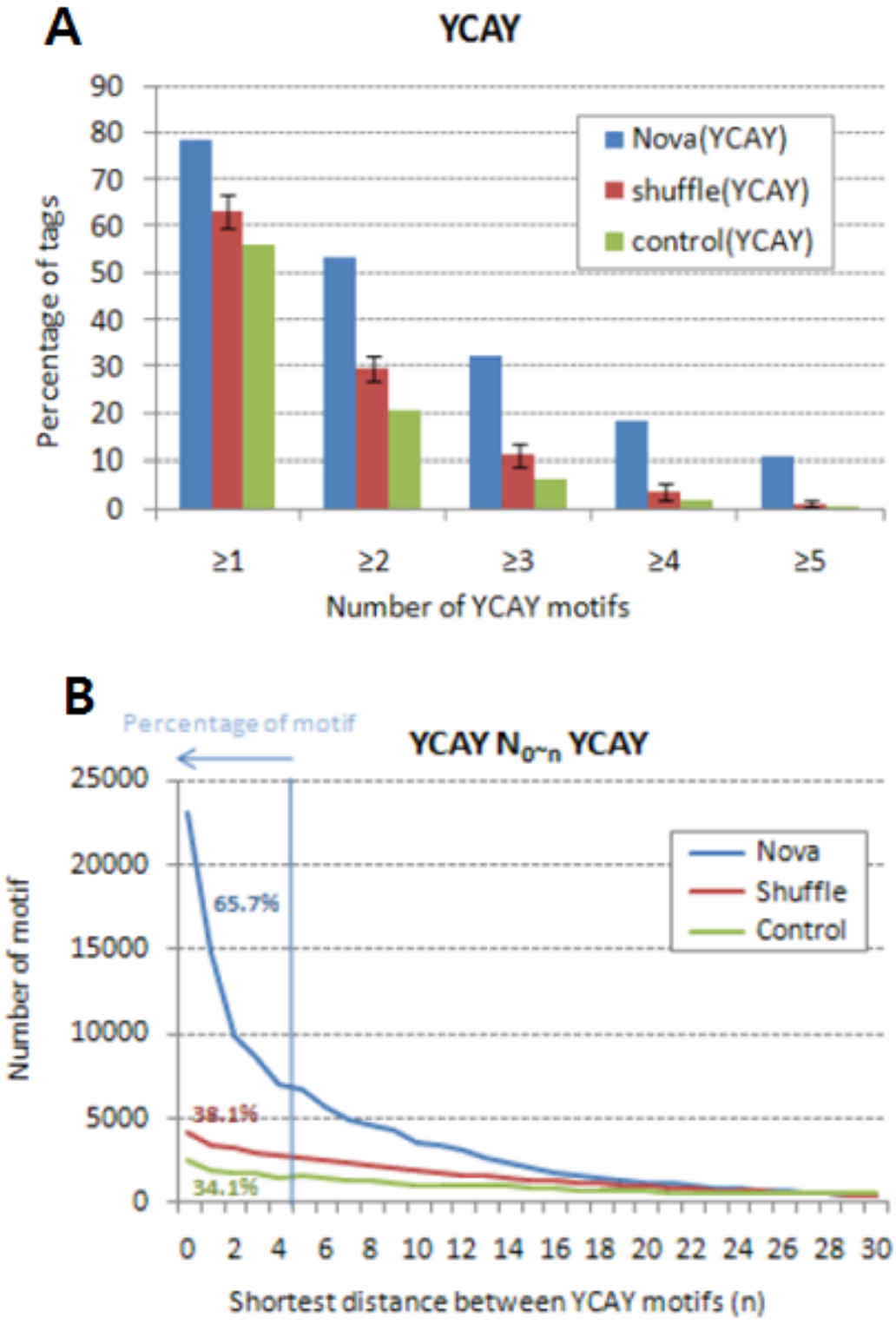

A
YCAY
Percentage of tags
90
80
70
60
50
40
30
20
10
0
Nova(YCAY)
shuffle(YCAY)
control(YCAY)
≥1
≥2
≥3
≥4
≥5
Number of YCAY motifs
B
Percentage of motif
YCAY N0~n YCAY
Number of motif
25000
20000
15000
10000
5000
0
65.7%
38.1%
34.1%
Nova
Shuffle
Control
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30
Shortest distance between YCAY motifs (n)

Figure 3.6. Genome wide distribution of Nova HITS-CLIP tags.

A. Pie chart representation of Nova HITS-CLIP annotation. Two databases (RefSeq and KnownGene) are used for the annotation of the tags. B. Number of Nova HITS-CLIP tags in 100bp windows are calculated. BLAT search is used to get the genomic coordination of mouse. C. Distribution of Nova HITS-CLIP tags in chromosome 12, where gtl2 is located. D. Distribution of the tags in gtl2 region (mm8, chr12:109,981,359-110,222,678) . Location of tags in genome is indicated as green bar together with bar graph with number of tags per 100bp in Y-axis. Transcript of gtl2 is displayed as blue arrow.Location of miRNAs in miRNA clusers are also indicated as red bar and C/D snoRNAs are represented as black bar.

Figure 3.6

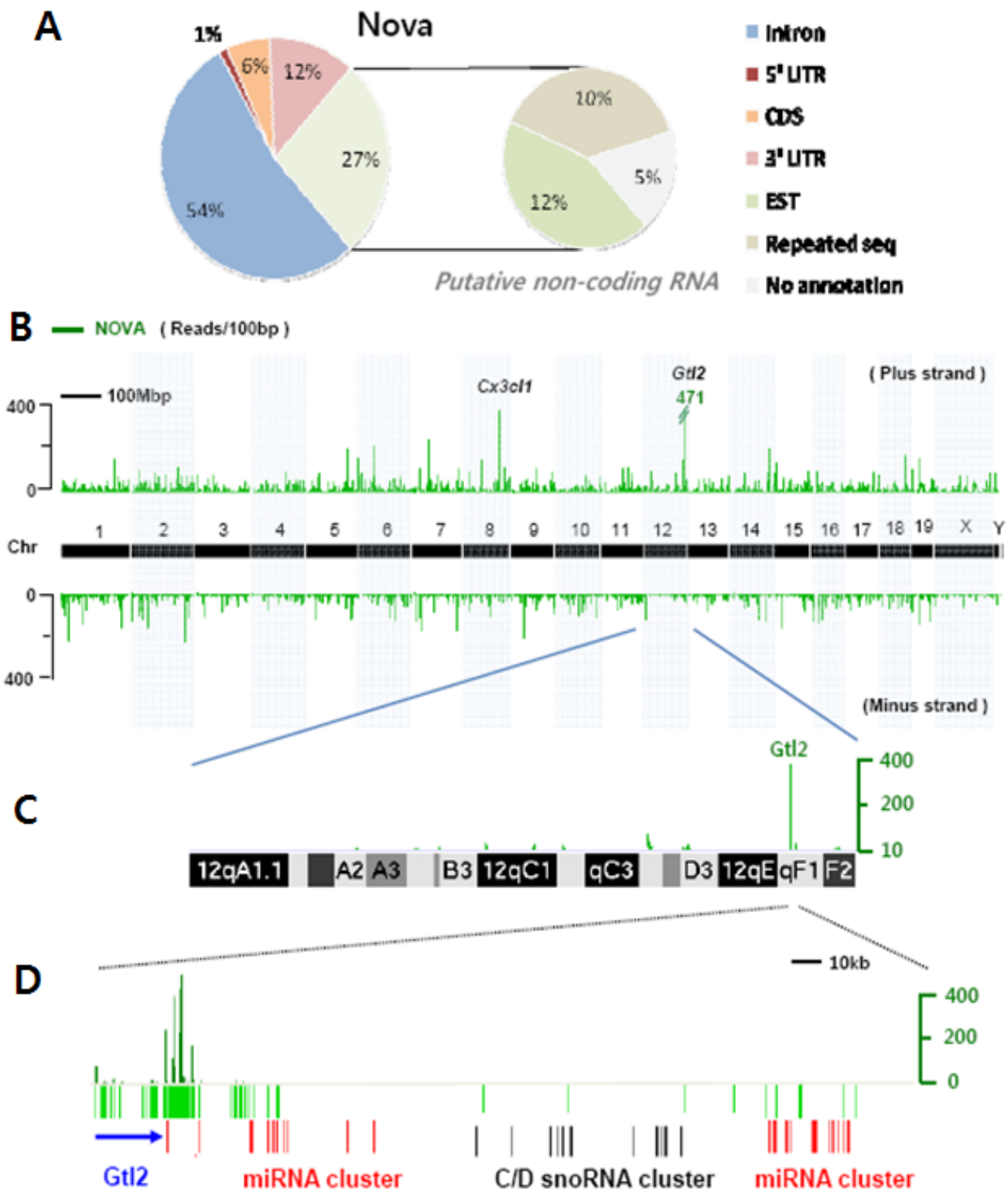

A
Nova
1%
6%
12%
27%
54%
10%
5%
12%
Putative non-coding RNA
Intron
5' UTR
CDS
3' UTR
EST
Repeated seq
No annotation
B
NOVA (Reads/100bp)
100Mbp
400
0
Cx3cl1
Gtl2
471
(Plus strand)
Chr
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 X Y
0
400
(Minus strand)
C
Gtl2
400
200
10
12qA1.1
A2
A3
B3
12qC1
qC3
D3
12qE
qF1
F2
D
10kb
400
200
0
Gtl2
miRNA cluster
C/D snoRNA cluster
miRNA cluster

Figure 3.7. Nova dependant alternative splicing of gtl2 and production of miR-770.

A. Gene structure of gtl2. Red or blue dot indicates the alternative splicing produt of gtl2 gene. Distribution of Nova-HITS tags are plotted as green and YCAY density (Reads/120bp) is also plotted as black. Location of mir-770 in gtl2 transcript is represented as red bar. B. Ratio of two-alternative splicing products (gtl2L and gtl2S) are measured by RT-PCR in the total RNA from wild type mouse brain (WT) versus Nova double knock-out mouse brain (DKO). C. Schematics of miRNA processing with location of PCR primers, used to measure the each processing product by quantitative real time PCR (Q-PCR). D. The amount of two alternative splicing forms of gtl2 is measured by Q-PCR. Ratio between values from WT and Nova2 knock out mouse forebrain (N2KO) is plotted as Y-axis. The amount of actin transcript is used for normalization. E. The amount of various products generated during miRNA processing is measured by Q-PCR (Un: unprocessed pri-miRNA, Pri: pri-miRNA, Ma: mature miRNA). Mir-25, which has no tags and no YCAY cluster in its vicinity, is used as control. For normalization, the amount of U6 is measured to obtain a ΔCT value.

Figure 3.7

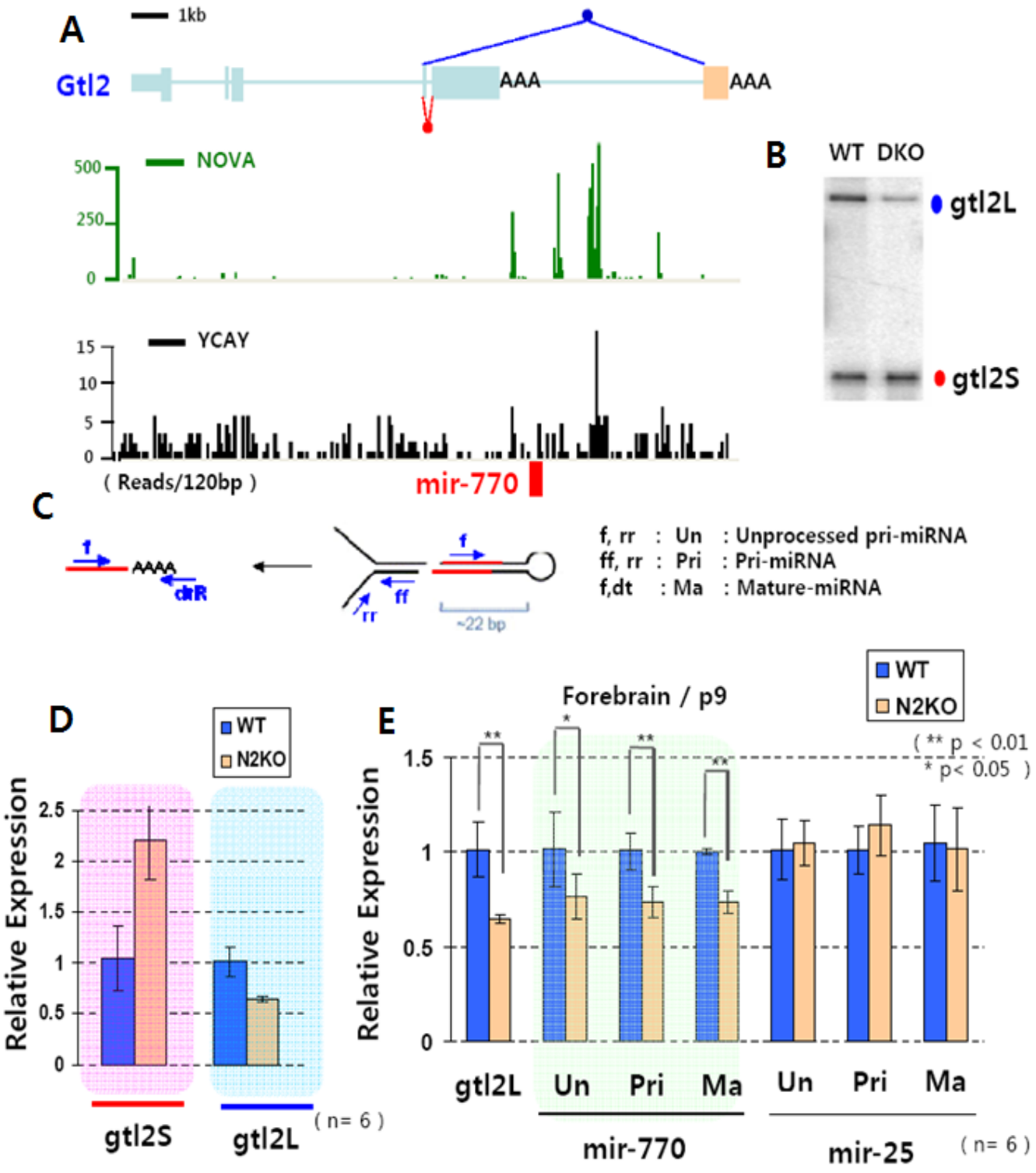
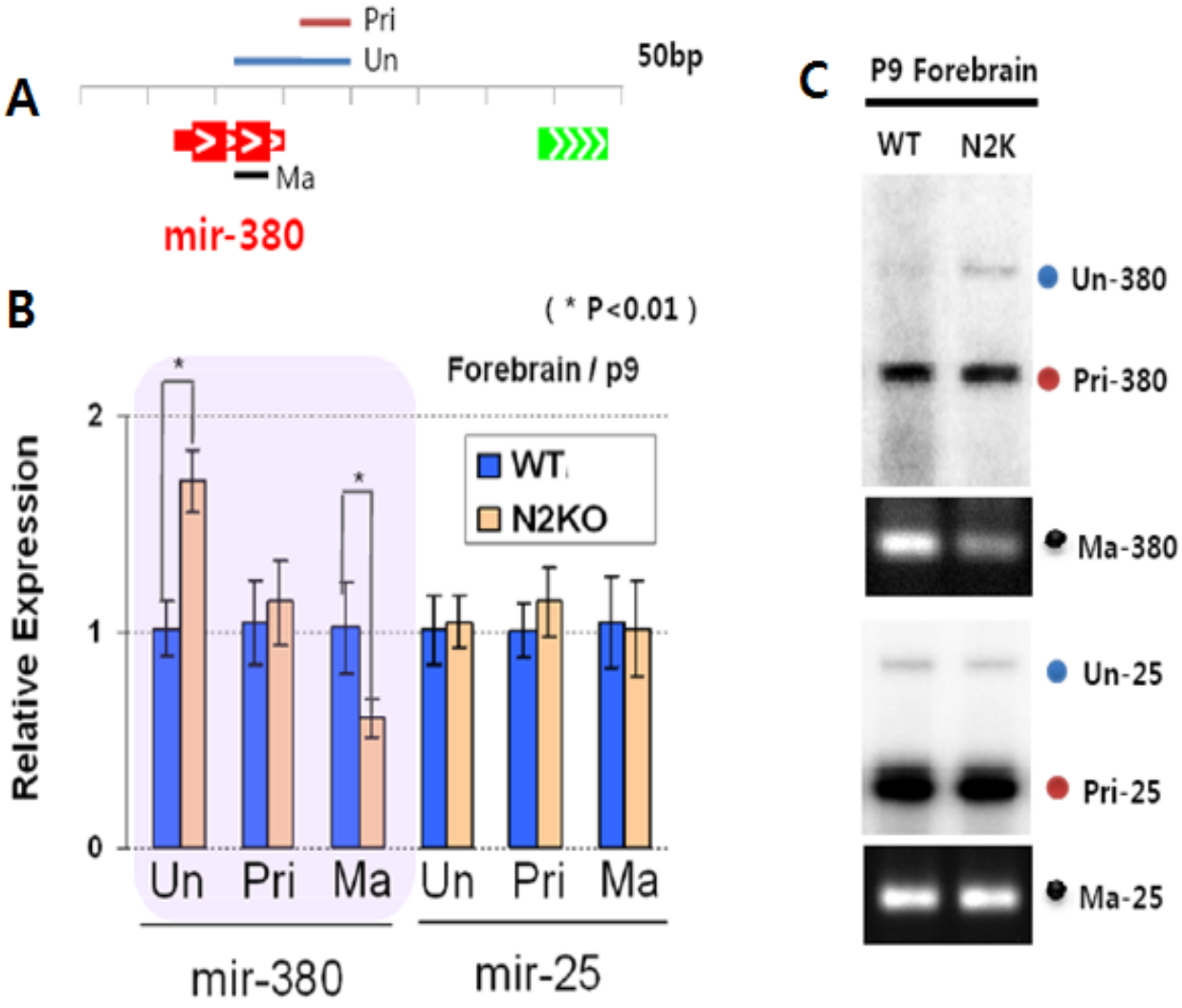

A
1kb
Gtl2
AAA
AAA
NOVA
500
250
0
YCAY
15
10
5
0
(Reads/120bp)
mir-770
B
WT DKO
gtl2L
gtl2S
C
f
AAAA
dtR
rr
ff
~22 bp
f, rr : Un : Unprocessed pri-miRNA
ff, rr : Pri : Pri-miRNA
f,dt : Ma : Mature-miRNA
D
WT
N2KO
Relative Expression
2.5
2
1.5
1
0.5
0
gtl2S
gtl2L
(n= 6)
E
Forebrain / p9
WT
N2KO
(** p < 0.01
* p< 0.05)
Relative Expression
1.5
1
0.5
0
gtl2L
Un
Pri
Ma
Un
Pri
Ma
mir-770
mir-25
(n= 6)

Figure 3.8. Nova dependant processing of miR-380.

A. Location of a Nova-CLIP tag in the primary transcript of miR-380 is indicated as green bar and miR-380 is displayed as red bar. PCR products amplified for measuring each processing product of miR-380 are indicated (Un: unprocessed pri-miRNA, Pri: pri-miRNA, Ma: mature miRNA). B. The amount of various products generated during miRNA processing is measured in miR-380 and miR-25. miR-25, which has no tags and no YCAY clusters in its vicinity, is used as control. For normalization, the amount of U6 is measured to obtain a ΔCT value. Ratio between the values from WT and Nova2 knock out mouse forebrain (N2KO) is plotted in Y-axis as relative expression. C. Various products generated during miR-380 and miR-25 processing are amplified by RT-PCR in the total RNA from wild type mouse forebrain (WT) versus Nova2 null mouse forebrain (N2K).

Figure 3.8


A
Pri
Un
50bp
Ma
mir-380
B
(* P<0.01)
Forebrain / p9
Relative Expression
2
1
0
WT
N2KO
Un
Pri
Ma
Un
Pri
Ma
mir-380
mir-25
C
P9 Forebrain
WT
N2K
Un-380
Pri-380
Ma-380
Un-25
Pri-25
Ma-25

CHAPTER IV. GENOME-WIDE INDENTIFICATION OF ARGONAUTE ASSOCIATED RNAS

Introduction

The limited number of primary transcripts in the mammalian genome has promoted interest in the possibility that much of the complexity in the regulation of gene expression may be determined by regulation of RNA metabolism and expression. MicroRNAs (miRNAs) underscore this possibility, as they directly bind to many mRNA to regulate their translation or stability (Filipowicz et al., 2008; Shyu et al., 2008), and thereby regulate a wide range of activities, including development, immune function and neuronal biology (Ambros, 2004; Bartel, 2004; He and Hannon, 2004). Many miRNAs are evolutionarily conserved, although others are species-specific (including human miRNAs not conserved in chimpanzees) (Berezikov, 2006), consistent with wide roles for miRNAs, from regulating basic cellular functions to generating organismal diversity.

We have developed HITS-CLIP to directly identify protein-RNA interactions in living tissues and in a genome-wide manner (Licatalosi et al., 2008; Ule et al., 2003). This method (Jensen and Darnell, 2008; Ule et al., 2005a) uses UV-

irradiation to covalently crosslink RNA-protein complexes that are in direct contact (~ single Angstrom distances) within cells, allowing RNA-protein complexes to be stringently purified. Partial RNA digestion reduces bound RNA to small fragments that can be sequenced by high throughput methods, yielding genome-wide information regarding the sites of the RNA-protein interactions (Licatalosi et al., 2008; Yeo, 2009). Recent X-Ray crystal structures of a ternary complex of a core RISC component, Argonaute (Ago), bound to both a 21-mer DNA guide strand (miRNA equivalent) and a 12 nt fragment of complementary mRNA (Wang et al., 2008a) suggest that Ago may make sufficiently close contacts with both nucleic acid strands to allow Ago HITS-CLIP to simultaneously identify Ago bound miRNAs and the nearby mRNA sites. Here we developed Ago HITS-CLIP analysis to identify Ago associated brain RNAs in genome wide manner. This provides a platform that can establish the direct targets upon which miRNAs act in a variety of biologic contexts, and the rules by which they do so.

Results

Immunoprecipitation of Argonaute complexes

HITS-CLIP experiments rely on a means of purifying RNA binding proteins (RNABPs) (Jensen and Darnell, 2008; Licatalosi et al., 2008; Ule et al., 2005a; Ule et al., 2003). To purify endogeneous Ago complexes, we performed immunoprecipitation (IP) using two different Ago antibodies, 7G1-1* and 2A8. Experiments using the 7G1-1* antibody were undertaken after recognizing that this monoclonal antibody, obtained from the Iowa Developmental Hybridoma (IDH) bank, were a mixture of clones of anti-FMRP antibody and anti-Ago antibody, an observation confirmed by mass- spectrometry sequencing of immunoprecipitated products in WT and FMRP KO brain (Figure 4.1), and confirmed independently (Z. Mourelatos, personal communication). Attempts to subclone independent hybridoma clones from a different frozen batch of cells (IDH) were not successful (as subsequent batches from IDH harbored only the original 7G1-1 clone). Specificity of Ago immunoprecipitation was confirmed in both experiments using 2A8 and 7G1-1 (Figure 4.1).

Development of Ago HITS-CLIP in mouse brain.

After confirming the specificity and efficiency of endogeneous Ago IP in mouse brain lysates (Figure 4.1), Ago HITS-CLIP was performed. To purify Ago bound

to mouse brain RNAs, we first UV-irradiated P13 neocortex and immunoprecipitated Ago under stringent conditions with strong detergent (0.1% SDS, 0.5% deoxycholate and 0.5% NP-40) and high salt wash (upto 685mM sodium chlrolide). Then we radiolabeled RNA, further purified crosslinked Ago-RNA complexes by SDS-PAGE and nitrocellulose transfer, and visualized complexes by autoradiography (Figure 4.2). HITS-CLIP with the 2A8 antibody (Nelson et al., 2007) was done in sibling P13 mice from freshly dissected neocortex and HITS-CLIP with the 7G1-1* antibody done in WT P13 mouse brain in the presence of blocking peptide (mapped to FMRP amino acid 354 to 368), which has been demonstrated to block FMRP immunoprecipitation by 7G1-1 (Brown et al., 2001). In addition, one experiment was done in sibling FMRP KO P13 mouse brain, although only tags that were also present in independent experiments were considered (Figure 4.1 and 4.6).

Interestingly, we observed complexes of two different modal sizes (~110 kD and ~130kD; Figure 4.1 and 4.2), suggesting that Ago (97kD) was crosslinked to two different RNA species. RT-PCR amplification revealed that the ~110 kD band harbored ~22 nt crosslinked RNAs and upper band both 22 nt and larger crosslinked RNAs (Figure 4.2C). Both products were sequenced with single molecule sequencing methods (Licatalosi et al., 2008) and found to correspond to

miRNAs and mRNAs, respectively (Table 4.1), suggesting that Ago might be sufficiently close to both miRNA and target mRNAs to form crosslinks to the ternary complex (Figure 4.2D). Such a result would allow the search for miRNA binding sites to be constrained to both the subset of miRNAs directly bound by Ago-CLIP and to the local regions of mRNAs to which Ago crosslinks, thus greatly reducing the rate of false positive predictions of miRNA binding.

Bioinformatic analysis of Ago-miRNA CLIP data

After aligning the tags from 110kD to miRNA sequences, toal ~22 x 10^6 miRNA tags are collected from 5 different biological replicates of Ago HITS-CLIP experiments (Table 4.1). We compared miRNAs by frequency with which they were detected in Ago-miRNA CLIP in five different experiments using two different antibodies (Figure 4.3). The set of Ago-crosslinked miRNAs were highly reproducible. Among biologic triplicates or 5 replicates done with two antibodies, the Pearson correlation coefficient was R2 > 0.9 and > 0.83, respectively, for Ago-miRNA CLIP (Figure 4.3 and 4.10B). We identified 454 unique miRNAs crosslinked to Ago in mouse brain, with Ago-miR-30e being the most abundant species (14% of total tags; Figure 4.3). The abundance of miRNAs was generally consistent with previous estimates of miRNA frequency in brain measured by cloning frequency (Landgraf, 2007) or bead-based cytometry (Lu,

2005), although the correlation with published results (R2 = 0.2 - 0.32; Figure 4.4) was not as high as among biologic replicates in our experiments. Some of the differences between our data and published results might be due to differences in the ages of brain used, regulation of Ago-mRNA interactions, and/or increased sensitivity allowed by the stringent conditions in CLIP which consequently improved signal:noise ratio

Bioinformatic analysis of Ago-mRNA CLIP data

After aligning the tags from the 130kD complex to mouse genome, toal ~1.5 x 10^6 unique tags are collected from 5 different biological replicates of Ago HITS-CLIP experiments (Table 4.1). To differentiate robust from non-specific or transient Ago-RNA interactions, we compared the results from biologic replicate experiments done with two different monoclonal antibodies (Figure 4.6). The data was further refined by comparison with in silico random CLIP, an algorithm designed to normalize for variation in transcript length, abundance, and simulated background cluster height (Figure 4.5). The processing algorithm, in silico random CLIP, simulates distribution of sequence tags across mRNAs based on its expression, providing a background level for the number of tags that would be expected by chance to simultaneously overlap one another, forming clusters (Figure 4.5). For in silico random CLIP algorithm, abundance of

transcripts in P13 mouse brain was measured by microarray and it is further confirmed that normalized probe intensity values from microarray can be used for in silico random CLIP simulation (Table 4.2). To facilitate the analysis of large numbers of Ago-mRNA CLIP tags (~1.5 x 10^6 unique tags; Table 4.1), we analyzed overlapping tags (clusters), which were normalized by in silico random CLIP and sorted by biologic complexity (Licatalosi et al., 2008) ("BC"; a measure of reproducibility between biologic replicates; Figure. 4.5-7). 1,463 robust clusters (BC=5; i.e. harboring CLIP tags in all five biologic experiments using both antibodies) mapped to 829 different brain transcripts, and 990 clusters had at least 10 tags (Figure. 4.7). We also estimated that sequencing depth was near saturation by simulation analysis (Figure 4.8). The set of Ago-crosslinked mRNAs were also highly reproducible. Among biologic triplicates or among 5 replicates done with two antibodies, the Pearson correlation coefficient was $R2 > 0.8$ and > 0.65, respectively, for Ago-mRNA CLIP (Figure 4.9 and 4.10A).

Genomic distribution of Ago-mRNA CLIP tags

The distribution of Ago-mRNA CLIP tags in the genome showed enrichment in transcribed sequences (Figure 4.10C). The pattern of tags in mRNAs mirrored functional assays with miRNAs (Grimson, 2007), which show no biologic activity when seed sites are present in the 5′ UTR (1% Ago-mRNA tags), and high

efficacy in 3′ UTRs (40% tags including 8% within 10kb downstream of annotated transcripts, regions likely to have unannotated 3′ UTRs (Licatalosi et al., 2008; Wang, 2008). In addition, an extensive set of tags were identified in other locations, including coding sequence (CDS; 25%), a site for which there is emerging evidence of miRNA regulation (Duursma et al., 2008; Forman et al., 2008; Lal, 2008; Shen et al., 2008; Tay et al., 2008), introns (12%), and non-coding RNAs (4%), suggesting the possibility that these sites may provide new insights into diverse biologic roles for Ago proteins. Within mRNAs, tags present in clusters were highly enriched in the 3′ UTR (~60%; Figure 4.11). Ago-mRNA clusters were highly enriched around stop codons (with a peak ~50 nt downstream) and at the 3′ end of transcripts (~70 nt upstream of presumptive poly(A) sites; $P<0.003$, Figure 4.10D), consistent with bioinformatic observations from microarray data (Grimson, 2007). Taken together, this data supports the biological relevance of Ago-mRNA clusters and suggests that they may be associated with functional binding sites.

Discussion

Here, we developed Ago HITS-CLIP to identify Ago associated RNAs, which are miRNAs and target mRNAs. Our Ago HITS-CLIP data provide a new approach promising to solve the difficult problem of identifying functionally relevant miRNA binding sites in living tissues. Several studies set the stage for the development of Ago HITS-CLIP as a way to possibly resolve the ambiguity of what RNAs miRNAs target and where they bind to them *in vivo*. Crystallographic structures of Ago-miRNA (Wang et al., 2008b) and Ago-target mRNA-miRNA ternary complexes (Wang et al., 2008a) have demonstrated close contacts between all three molecules, underscoring the suggestion that CLIP, which requires close protein-RNA contacts (Jensen and Darnell, 2008; Ule et al., 2005a), could identify Ago-RNA ternary complexes. High throughput experimental approaches, together with increasingly powerful bioinformatics technique shows robust and reproducible results across biological replicates of Ago HITS-CLIP and also shows relevant distribuction of mRNA tags along the genome and transcripts to miRNA biology. Such results indicate that Ago HITS-CLIP can be used as the platform to identify miRNA binding sites *in vivo*.and also would allow the search for miRNA binding sites to be constrained to both the subset of miRNAs directly bound by Ago and to the local regions of mRNAs

to which Ago crosslinked, potentially reducing the rate of false-positive predictions of miRNA binding.

Contributions

Ago HITS-CLIP analysis was developed by Sung Wook Chi, Julie B. Zang, Aldo Mele and Robert B. Darnell. Bioinformatic analysis, Ago HITS-CLIP experiments with the 2A8 antibody, and exon array experiments were done by Sung Wook Chi. *in silico* random CLIP algorithm was also developed by Sung Wook Chi. Characterization of 7G1-1* antibody and initial Ago CLIP experiments with the 7G1-1* antibody were performed by Julie B. Zang. Aldo Mele helped with all HITS-CLIP experiments. John J. Fak helped with microarray experiments. Scott Dewell helped with high-throughput sequencing (Illumina Genome Analyser).

Figure 4.1. Immunoprecipitation of Ago complex with 7G1-1* and 2A8

A. WT and FMRP KO mouse brain lysates were immunoprecipitated with either 7G1-1* (containing monoclonal antibodies to both FMRP and Ago) or 7G1-1 (containing only FMRP monoclonal antibody). Immunoprecipitates were analyzed by immunoblot probed with 7G1-1* (left panel), anti-Ago1 (right panel) and anti-Ago2 antibody (middle panel). Both 7G1-1 and 7G1-1* antibodies immunoprecipitated FMRP (not in FMRP KO). 7G-1* also IPed Ago1 and Ago2 at ~100 kD from both WT and KO recognized by anti-Ago1 and anti-Ago2 antibodies. B. 7G1-1* immunoprecipitate from KO brain was run on SDS PAGE and stained with Coomassie brilliant blue. Protein bands were analyzed with mass spectrometry. Ago proteins 1-4 were identified at ~100 kD. C. 7G1-1* antibody was pre-incubated with 200ug of FMRP peptide (354-KHLDTKENTHFSQPN-368; the epitope of 7G1-1* or an irrelevant control (FLAG peptide). FMRP peptide, but not control peptide, was sufficient to completely block FMRP immunoprecipitation from WT mouse brain and had no effect on Ago immunoprecipitation. D. 32P-labeled RNA crosslinked to Ago and FMRP immunoprecipiated with 7G1-1* from WT or FMRP KO mice as indicated. The antibody pulls down complexes of ~90kD, which correspond to FMRP, as indicated by the disappearance of this band when IPs performed from FMRP KO brains or in the presence of 200 ug of the competing peptide epitope. For HITS-CLIP experiments, two different animals were used, Brain "D" corresponds to CLIP of WT brain done in the presence of peptide competition and isolated the Ago band (D, lane 3), and Brain "E" corresponds to CLIP of FMRP KO brain (D, lane 2). E. Replicate Ago CLIP with 2A8 monoclonal antibody. Overdigestion of Ago RNA cross-linked complex showed the protein size of Ago (~95kD)

Figure 4.1

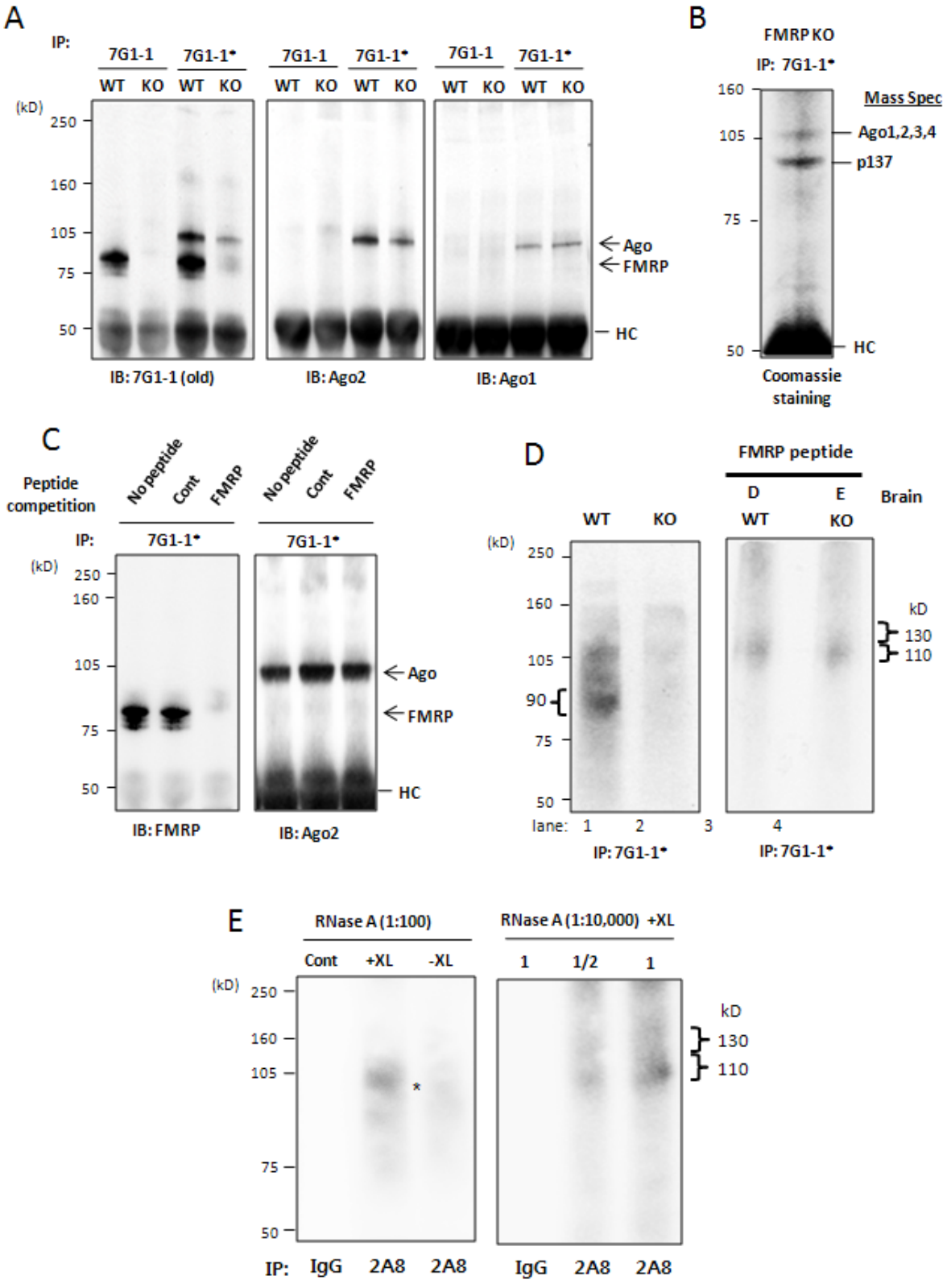

A
IP:
7G1-1
7G1-1*
WT KO WT KO
(kD)
250
160
105
75
50
Ago
FMRP
HC
IB: 7G1-1 (old)
IB: Ago2
IB: Ago1
B
FMRP KO
IP: 7G1-1*
Mass Spec
Ago1,2,3,4
p137
HC
Coomassie staining
C
Peptide competition
No peptide
Cont
FMRP
IP: 7G1-1*
IB: FMRP
IB: Ago2
D
FMRP peptide
D
E
Brain
WT KO
WT KO
kD
130
110
90
lane: 1 2 3 4
IP: 7G1-1*
IP: 7G1-1*
E
RNase A (1:100)
RNase A (1:10,000) +XL
Cont +XL -XL
1 1/2 1
kD
130
110
IP: IgG 2A8 2A8 IgG 2A8 2A8

Figure 4. 2. Argonaute HITS-CLIP.

A. Immunoblot (IB) analysis of Ago immunoprecipitates (IP) from P13 mouse neocortex using pre-immune IgG as a control or anti-Ago monoclonal antibody 2A8 were blotted with 7G1-1* antibody. B. Autoradiogram of 32P-labeled RNA crosslinked to mouse brain Ago purified by IP. RNA-protein complexes of ~110 kD and ~ 130 kD are seen with 2A8 but not control IP. C. PCR products amplified after linker (36 nt) ligation to RNA products excised from B. Products from the 110 kD RNA-protein complex were ~22 nt miRNAs, and those from 130 kD complexes were predominantly mRNAs. D. Illustration showing proposed interpretation of data in C. Ago (drawn based on structure 3F73 in PDB) (Wang et al., 2008a) binds in a ternary complex to both miRNA and mRNA, with sufficiently close contacts to allow UV-crosslinking to either RNA; mRNA tags will be in the immediate vicinity of miRNA binding sites. Putative crosslinking sites are indicated as stars.

Figure 4. 2

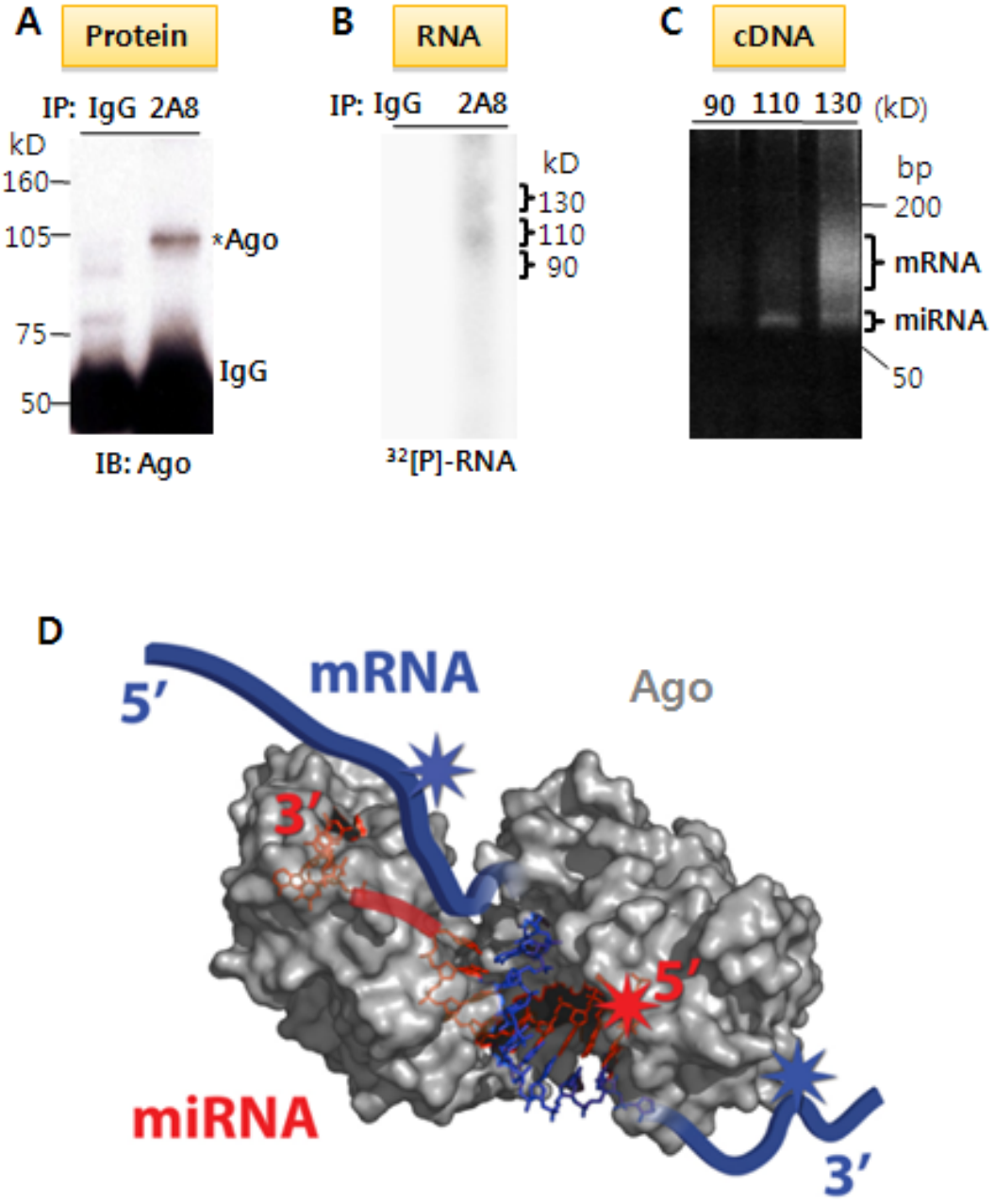

A
Protein
IP: IgG 2A8
kD
160
105
75
50
*Ago
IgG
IB: Ago
B
RNA
IP: IgG 2A8
kD
130
110
90
32[P]-RNA
C
cDNA
90 110 130 (kD)
bp
200
mRNA
miRNA
50
D
mRNA
Ago
5′
3′
5′
miRNA
3′

Figure 4.3. Reproducibility of Ago-miRNA CLIP results

A. Heat map comparing miRNAs by frequency with which they were detected in Ago-miRNA CLIP in five different experiments using two different antibodies. The number of tags in each experiment was normalized by rank quantile normalization and median value for each miRNA from 5 experiments is used as normalized tag frequency to determine a final ranking. A heat map was generated by the Treeview program (http://rana.lbl.gov/EisenSoftware.htm) with log2 ratio of normalized tag frequency to median. Red tags are ranked above median, green below, white at the median, grey is missing (no tags). % is the number of tags in total tags from 5 experiments. B. Correlation of Ago-miRNA CLIP results using different antibodies. C. The top 30 miRNAs are shown for each of 5 experiments. D. The bottom 30 miRNAs are shown.

Figure 4.3

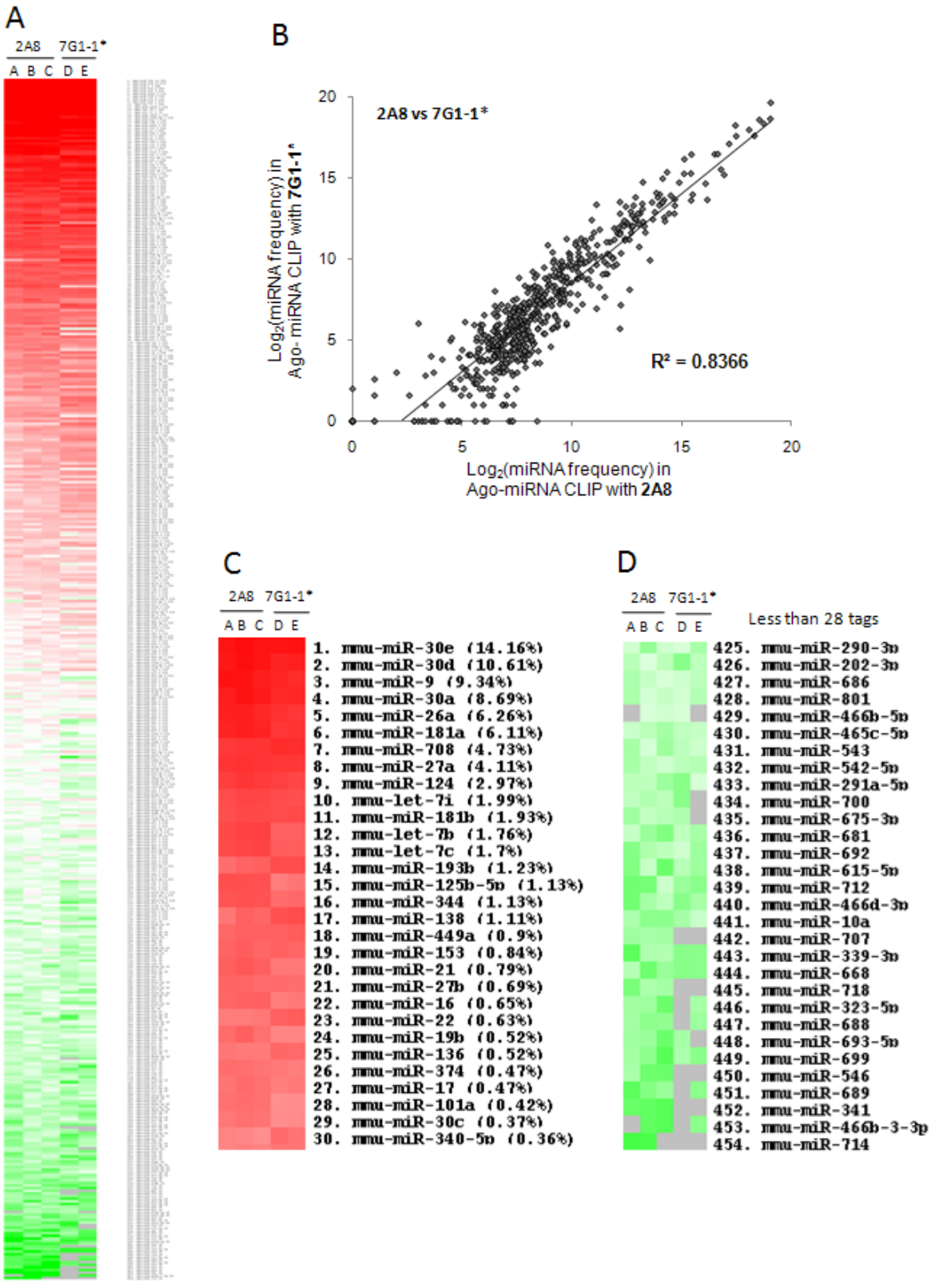

A
2A8
7G1-1*
A B C D E
B
2A8 vs 7G1-1*
Log2(miRNA frequency) in Ago- miRNA CLIP with 7G1-1*
Log2(miRNA frequency) in Ago-miRNA CLIP with 2A8
R² = 0.8366
0
5
10
15
20
C
2A8
7G1-1*
A B C D E
1. mmu-miR-30e (14.16%)
2. mmu-miR-30d (10.61%)
3. mmu-miR-9 (9.34%)
4. mmu-miR-30a (8.69%)
5. mmu-miR-26a (6.26%)
6. mmu-miR-181a (6.11%)
7. mmu-miR-708 (4.73%)
8. mmu-miR-27a (4.11%)
9. mmu-miR-124 (2.97%)
10. mmu-let-7i (1.99%)
11. mmu-miR-181b (1.93%)
12. mmu-let-7b (1.76%)
13. mmu-let-7c (1.7%)
14. mmu-miR-193b (1.23%)
15. mmu-miR-125b-5p (1.13%)
16. mmu-miR-344 (1.13%)
17. mmu-miR-138 (1.11%)
18. mmu-miR-449a (0.9%)
19. mmu-miR-153 (0.84%)
20. mmu-miR-21 (0.79%)
21. mmu-miR-27b (0.69%)
22. mmu-miR-16 (0.65%)
23. mmu-miR-22 (0.63%)
24. mmu-miR-19b (0.52%)
25. mmu-miR-136 (0.52%)
26. mmu-miR-374 (0.47%)
27. mmu-miR-17 (0.47%)
28. mmu-miR-101a (0.42%)
29. mmu-miR-30c (0.37%)
30. mmu-miR-340-5p (0.36%)
D
2A8
7G1-1*
A B C D E
Less than 28 tags
425. mmu-miR-290-3p
426. mmu-miR-202-3p
427. mmu-miR-686
428. mmu-miR-801
429. mmu-miR-466b-5p
430. mmu-miR-465c-5p
431. mmu-miR-543
432. mmu-miR-542-5p
433. mmu-miR-291a-5p
434. mmu-miR-700
435. mmu-miR-675-3p
436. mmu-miR-681
437. mmu-miR-692
438. mmu-miR-615-5p
439. mmu-miR-712
440. mmu-miR-466d-3p
441. mmu-miR-10a
442. mmu-miR-707
443. mmu-miR-339-3p
444. mmu-miR-668
445. mmu-miR-718
446. mmu-miR-323-5p
447. mmu-miR-688
448. mmu-miR-693-5p
449. mmu-miR-699
450. mmu-miR-546
451. mmu-miR-689
452. mmu-miR-341
453. mmu-miR-466b-3-3p
454. mmu-miR-714

Figure 4.4. Comparison of Ago-miRNA CLIP data with published profiles of brain expressed miRNAs.

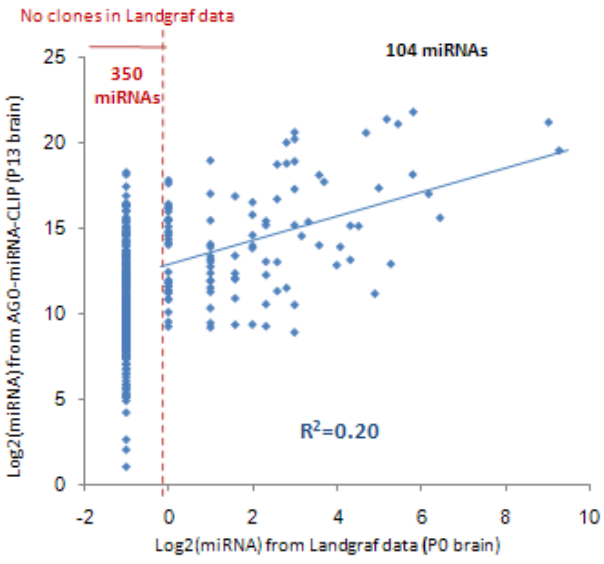
Generally the Ago-miRNA CLIP set of miRNAs corresponded to previously reported results, although there was a higher correlation between biologic replicates in our experiments than with previously published measures of miRNA frequency in brain, measured by cloning frequency12 (A,B,C) or bead-based cytometry13 (D,E,F) (R2 = 0.2 - 0.32, higher than Ago-miRNA CLIP in brain vs miRNA frequency in liver (0.06-0.12)). These differences might relate to different developmental ages of brain used for isolation; Landgraf et al analyzed P0 mouse brain, while Lu et al analyzed human brain. In addition, regulation of Ago interaction with miRNA could account for differences between crosslinked populations and whole miRNA populations, as the previously reported approaches used whole small RNA populations for analysis. It is also notable that a significant number of miRNAs identified by Ago HITS-CLIP were not found database derived from direct sequencing ((b); 350 miRNAs present in P13 brain Ago HITS-CLIP that were absent from whole sequencing of miRNAs in P0 brain), suggesting the possibility that this method has greater sensitivity than reported approaches.

Figure 4.4

A

	Landgraf et al	AGO miRNA CLIP	
miRNA	TOTAL CLONES NODE	Total clones in AGO-miRNA-CLIP	Rank in AGO-miRNA CLIP
mmu-miR-124	740	735430	9
mmu-miR-9	547	2309955	3
mmu-miR-126-3p	131	10238	103
mmu-miR-16	105	160843	22
mmu-miR-140	96	34537	62
mmu-miR-125b-5p	91	278206	15
mmu-miR-27a	84	1017631	8
mmu-miR-30e	82.5	3504225	1
mmu-miR-136	75	127192	25
mmu-miR-143	71	15878	89
mmu-miR-872	64	2224	195
mmu-miR-30a	60	2150543	4
mmu-miR-30d	47.5	2624768	2
mmu-miR-15a	41	35050	61
mmu-miR-379	39	7411	123
mmu-let-7c	35	419429	13
mmu-miR-29b	32	8154	116
mmu-miR-24	32	41024	57
mmu-miR-22	32	154156	23
mmu-miR-130a	28	14884	92

B



C

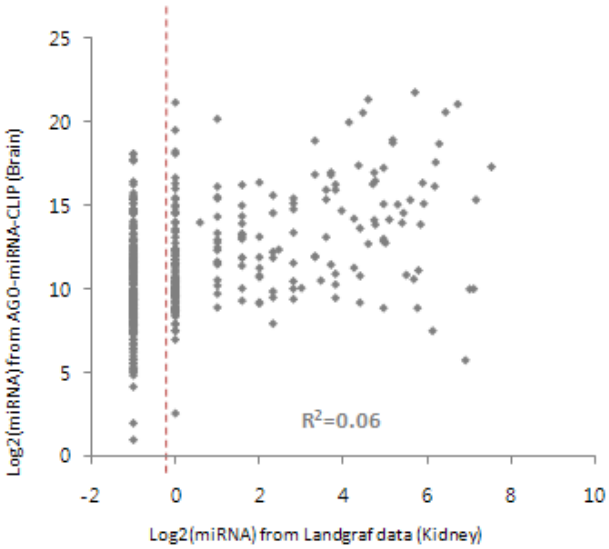
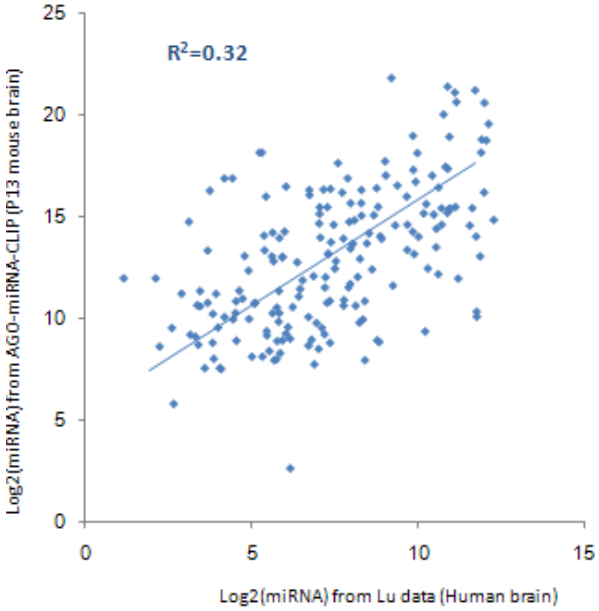


Figure 4.4 continues

D

	Lu et al	AGO miRNA CLIP	
miRNA	Expression level	Total clones in AGO-miRNA CLIP	Rank in AGO-miRNA CLIP
miR-338-5p	4913	28123	67
miR-124	4425	735430	9
let-7c	4229	419429	13
miR-181a	4054	1510285	6
let-7a	4009	71661	38
let-7b	3824	435438	12
miR-125b-5p	3767	278206	15
miR-29b	3701	8154	116
miR-29c	3455	1045	246
miR-29a	3439	1238	236
miR-219	3403	15966	88
miR-9	3339	2309955	3
miR-125a-5p	3148	42068	53
miR-128	2993	23182	79
miR-320	2353	3830	156
miR-26a	2262	1547706	5
miR-181c	2217	43566	49
miR-30a	2190	2150543	4
miR-24	2005	41024	57
miR-181b	1958	475794	11
let-7f	1949	41306	56
miR-16	1897	160843	22
miR-30d	1882	2624768	2
miR-137	1874	35776	58
miR-27b	1803	170907	21
miR-27a	1733	1017631	8

E



F

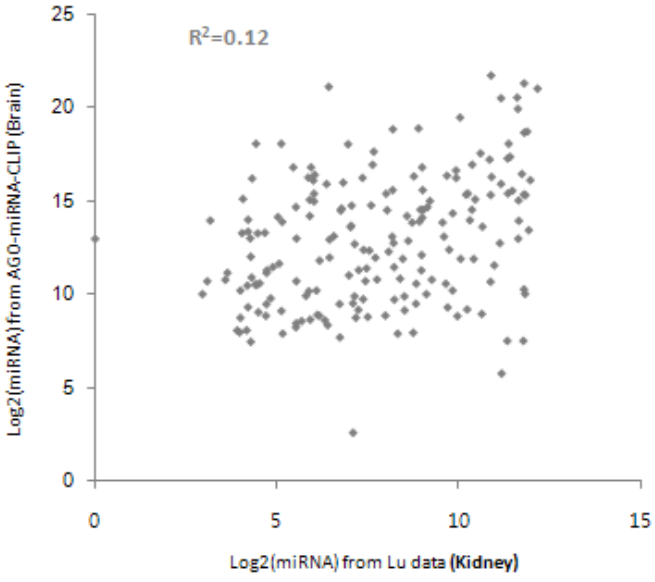


Figure 4.5. in silico random CLIP normalization.

A.Schematic of the in silico random CLIP normalization process and an example are shown. For example, thresholds for gene A were estimated by in silico random CLIP, and this estimate gives a normalization threshold of 3 (to yield an acceptable P value, < 0.01). Hence in the cartoon on bottom left, clusters of peak height < 3 were removed after normalization.

Figure 4.5

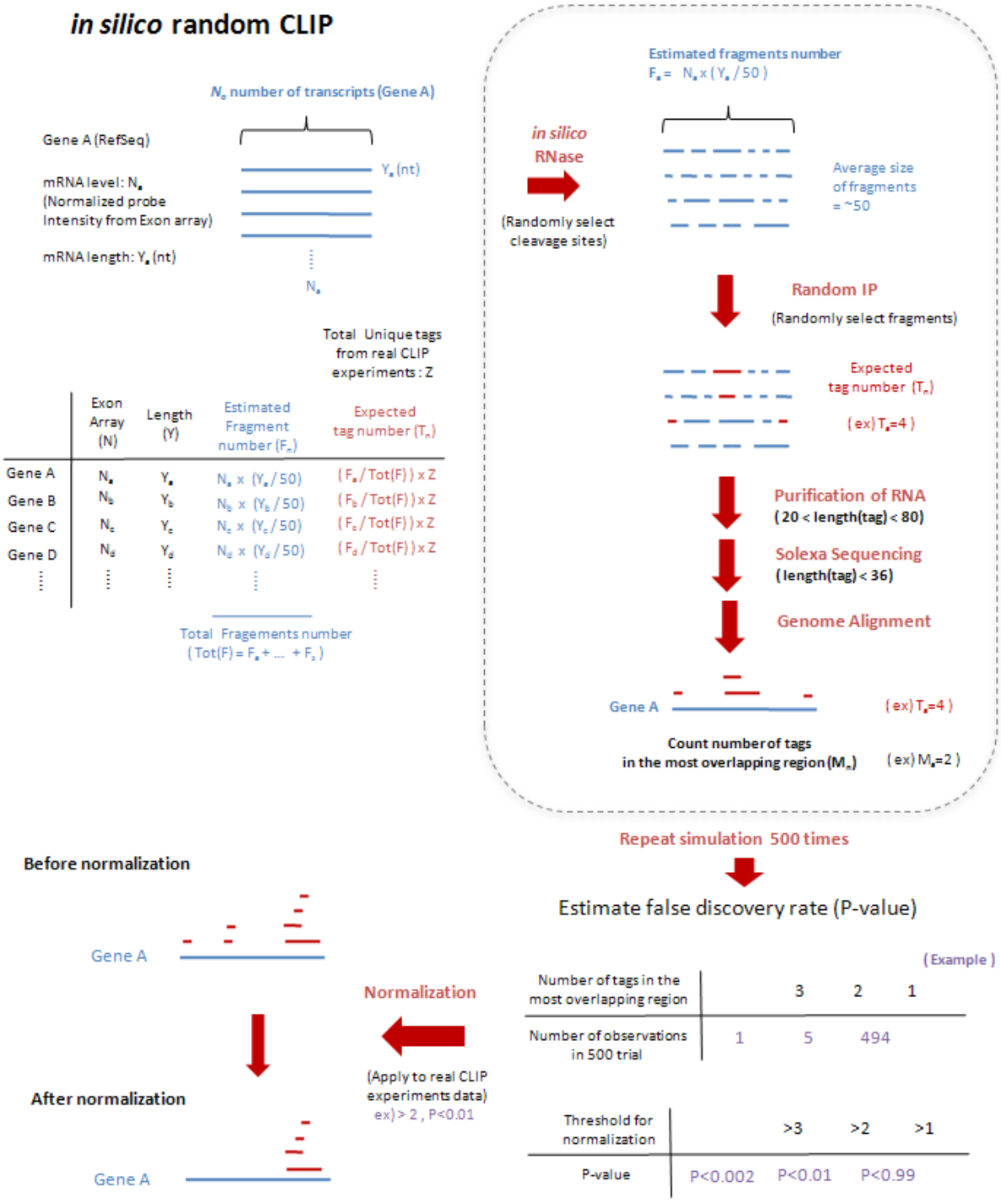

in silico random CLIP
N_a number of transcripts (Gene A)
Gene A (RefSeq)
mRNA level: N_a (Normalized probe Intensity from Exon array)
mRNA length: Y_a (nt)
Y_a (nt)
N_a
Total Unique tags from real CLIP experiments : Z
Exon Array (N)
Length (Y)
Estimated Fragment number (F_n)
Expected tag number (T_n)
Gene A
Gene B
Gene C
Gene D
N_a
N_b
N_c
N_d
Y_a
Y_b
Y_c
Y_d
N_a x (Y_a / 50)
N_b x (Y_b / 50)
N_c x (Y_c / 50)
N_d x (Y_d / 50)
(F_a / Tot(F)) x Z
(F_b / Tot(F)) x Z
(F_c / Tot(F)) x Z
(F_d / Tot(F)) x Z
Total Fragements number (Tot(F) = F_a + ... + F_z)
Estimated fragments number
F_a = N_a x (Y_a / 50)
in silico RNase
(Randomly select cleavage sites)
Average size of fragments = ~50
Random IP
(Randomly select fragments)
Expected tag number (T_n)
(ex) T_a=4
Purification of RNA
(20 < length(tag) < 80)
Solexa Sequencing
(length(tag) < 36)
Genome Alignment
Gene A
(ex) T_a=4
Count number of tags in the most overlapping region (M_n)
(ex) M_a=2
Repeat simulation 500 times
Estimate false discovery rate (P-value)
(Example)
Number of tags in the most overlapping region
3
2
1
Number of observations in 500 trial
1
5
494
Threshold for normalization
>3
>2
>1
P-value
P<0.002
P<0.01
P<0.99
Before normalization
Gene A
Normalization
(Apply to real CLIP experiments data)
ex) > 2 , P<0.01
After normalization
Gene A

Figure 4.6. Biologic complexity normalization.

Biologic complexity (BC) refers to overlapping tags between experiments. BC is determined by taking data from each Brain HITS-CLIP, normalizing by in silico random CLIP for mRNA abundance, and comparing data from at least one of each antibody immunoprecipitiation (2A8 and 7G1-1*). For example, for a cluster to have a biologic complexity of 2, it must have at least one tag from 2A8 and 7G1-1 Ago HITS-CLIP experiments. For Brain E, since the animal used was an FMRP KO, to be conservative, we only selected tags that were also present in WT (Brain D), to avoid confounding data that might result from the absence of FMRP.

Figure 4.6

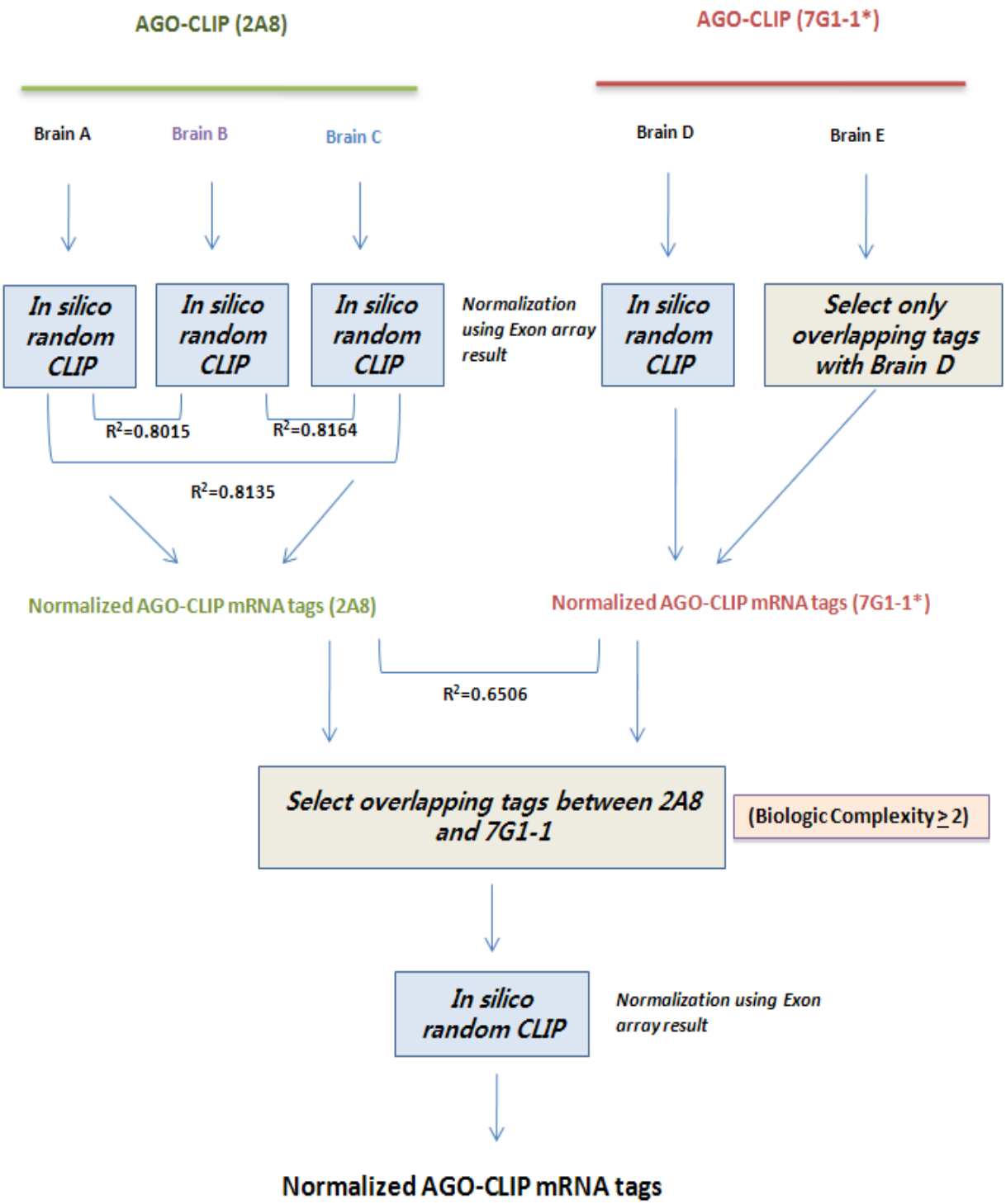

AGO-CLIP (2A8)
AGO-CLIP (7G1-1*)
Brain A
Brain B
Brain C
Brain D
Brain E
In silico random CLIP
In silico random CLIP
In silico random CLIP
Normalization using Exon array result
In silico random CLIP
Select only overlapping tags with Brain D
R²=0.8015
R²=0.8164
R²=0.8135
Normalized AGO-CLIP mRNA tags (2A8)
Normalized AGO-CLIP mRNA tags (7G1-1*)
R²=0.6506
Select overlapping tags between 2A8 and 7G1-1
(Biologic Complexity ≥ 2)
In silico random CLIP
Normalization using Exon array result
Normalized AGO-CLIP mRNA tags

Figure 4.7. Biologic complexity and peak height analysis of Ago mRNA clusters.

A. Table of Ago mRNA clusters sorted by biologic complexity and number of tags in the cluster peak (Peak height). B. Table of Ago clusters in genes transcribed in P13 brain, as determined by Affymetrix exon array. C. Plot of data in (A), demonstrating that data for samples with different biologic complexity converge at clusters with peak height > 20, suggesting that stringent data analysis can be obtained with peak height > 20. D. Comparison of distribution of clusters depending on peak height in in silico random CLIP vs. Ago-mRNA CLIP, plotted with different biologic complexity. This allows P values to be determined for selecting further stringent sets of clusters in a given global threshold (peak height) although those clusters were already normalized by in silico random CLIP.

Figure 4.7

A

Number of Ago-mRNA clusters

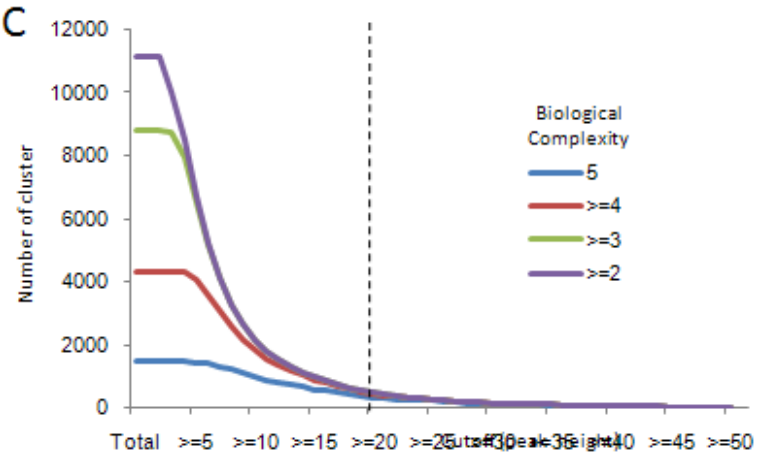
Cutoff (peak height)	Biological Complexity 5	>=4	>=3	>=2
Total	1463	4335	8775	11118
>=5	1448	4036	6521	6710
>=10	990	1844	2172	2173
>=15	587	878	959	959
>=20	336	456	479	479
>=25	223	265	272	272
>=30	134	151	151	151
>=35	94	102	102	102
>=40	60	63	63	63
>=45	31	32	32	32
>=50	20	20	20	20

B

complexity	>=5	>=4	>=3	>=2
Number of cluster	1463	4335	8775	11118
Number of genes	829	2081	3563	4280
Clusters / genes	1.77	2.08	2.46	2.60

P13 brain expressed gene: 10743 genes
(DABG : p<0.05 , Value > 100)

C



D

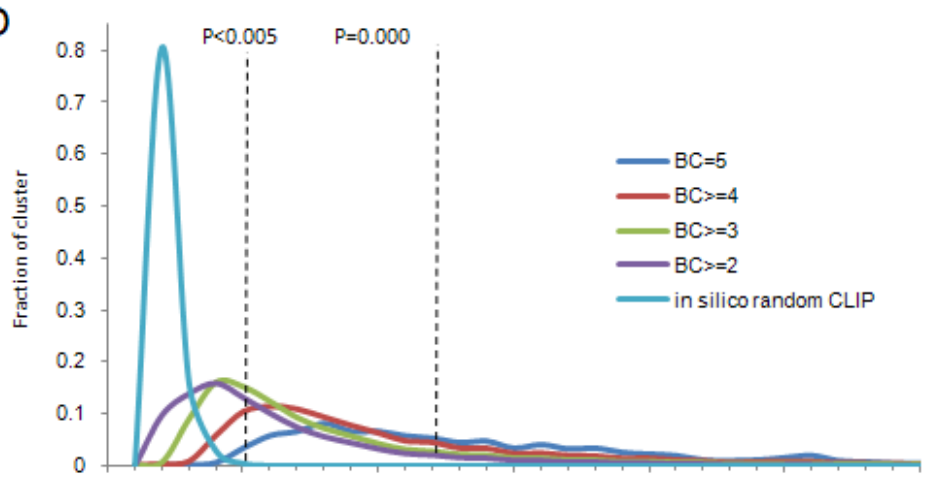


Figure 4.8. Simulations estimating the relationship between depth of HITS-CLIP sequencing and number of clusters/tags identified.

A. To estimate whether our sequencing depth was sufficiently saturated to give a global view of Ago HITS-CLIP targets, we undertook a simulation to see the effects of the last 10% of sequencing on our final Ago ternary cluster map. 10% of sequenced tags were removed randomly from each of 5 different AGO mRNA CLIP results, then those tags were processed by in silico normalization, alignment of tags using BLAT and cubic spline interpolation analysis. We then performed the simulation 10 times and compared the number of Ago ternary clusters (BC>=2) between the original set and the simulated one, focusing on the top 20 miRNAs clusters which we used for the Ago maps. The recovery rate of these simulation was ~92% for the Ago top 20 miRNA clusters (14411/15665) and ~95% for miR-124 (1487/1561), which are higher than the expected rate (90%). This increase of recovery rate was statistically significant (top 20; 1.02 = 92/90, P=2.08 x 10-11 and for miR-124; 1.06 = 95/90, P=2.55 x 10-7; Chi square test).

B. The simulation in (A) was repeated as a function of tags in the simulation. This was accomplished by repeating the simulation removing a range of tags, from 10-90%. We estimated the recovery of Ago miR-124 clusters (blue) and all top 20 Ago-miRNA-mRNA clusters as indicated. These results indicate that our sequencing depth is near the saturation point and is sufficient to show global view of miRNA target sites for the top 20 miRNAs. For example, even after removing 50% of tags, we had 88% recovery of the miR-124 clusters.

Figure 4.8

A

	Number of Ago mir-124 clusters	Number of Ago Top20 miRNA clusters
Ago-CLIP	1561	15665
Average # of clusters in 10 simulations	1487	14410.7
Standard deviation	2.05	16.57
simulation #1	1484	14401
simulation #2	1490	14417
simulation #3	1489	14401
simulation #4	1488	14440
simulation #5	1484	14413
simulation #6	1487	14408
simulation #7	1487	14388
simulation #8	1488	14412
simulation #9	1488	14393
simulation #10	1485	14434
recovery rate(%)	95	92

B

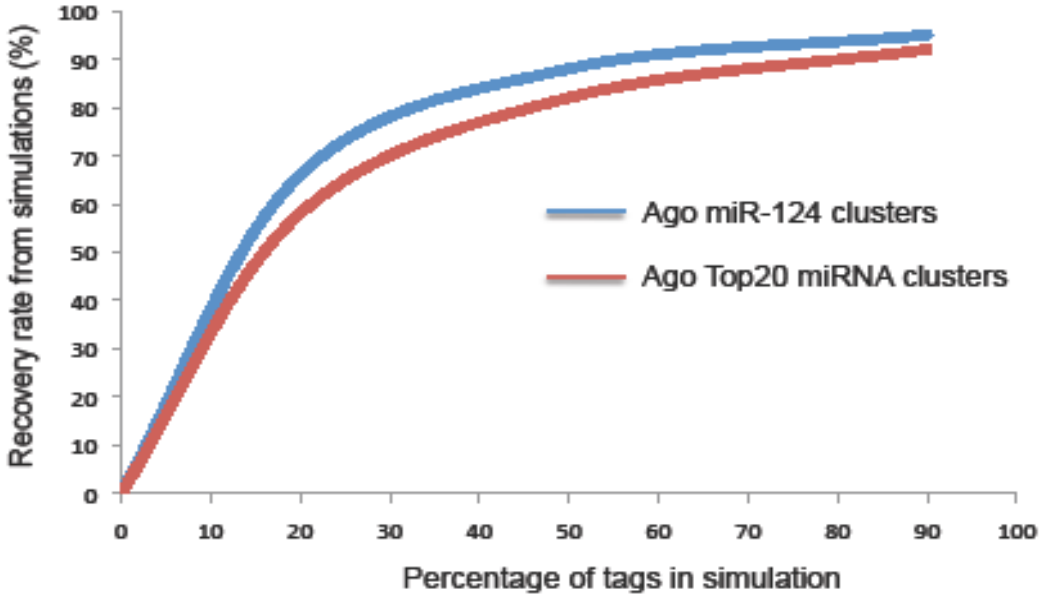


Figure 4.9. Reproducibility of Ago-mRNA CLIP results

A. To define reproducibility of Ago-mRNA clusters, we first identified inter-animal clusters10 with at least one tag in each of three biologic replicates. Pearson correlation coefficients represent the correlation between the normalized density of tags (log2 value after in silico random CLIP normalization) in each cluster. B. The analysis in (A) was repeated, comparing reproducibility with the two different Ago monoclonal antibodies. C. A genome-wide graph of data from (B).

Figure 4.9

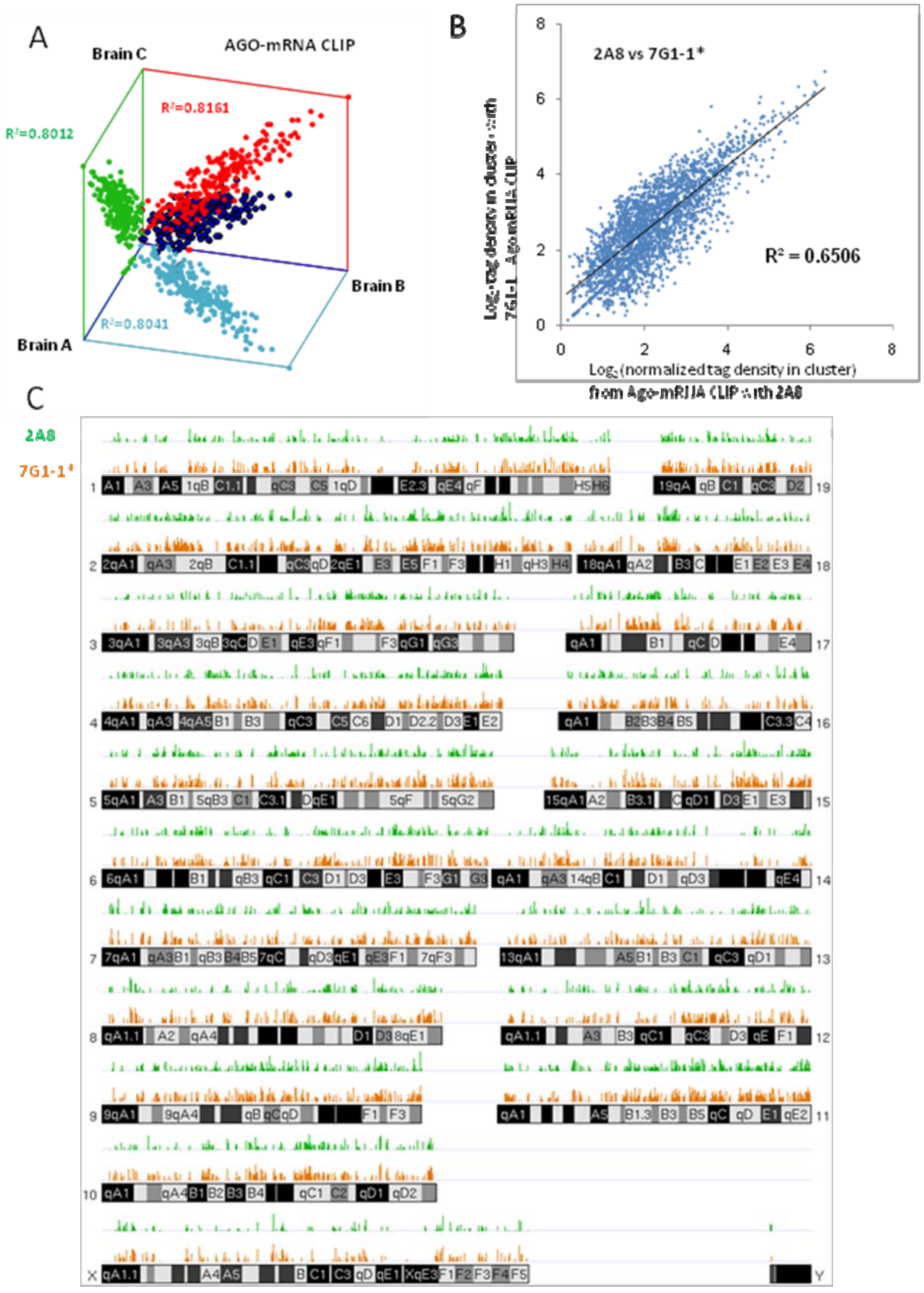

A
AGO-mRNA CLIP
Brain C
R²=0.8161
R²=0.8012
Brain B
R²=0.8041
Brain A
B
2A8 vs 7G1-1*
R² = 0.6506
Log₂(normalized tag density in cluster)
C
2A8
7G1-1*

Figure 4.10. Analysis of Ago HITS-CLIP tags and clusters

A-B. Reproducibility of all Ago-miRNA tags (A; graphed as log2(normalized miRNA frequency) per brain) or all tags within Ago-mRNA clusters. C. Location of reproducible Ago-mRNA tags (tags in clusters; BC>2) in the genome. Annotations are from RefSeq;"others" are unannotated EST transcripts, non-coding RNAs are from lincRNAs50 or FANTOM3. D. Top panel: The position of robust Ago-mRNA clusters (BC = 5) in transcripts is plotted relative to the stop codon and 3′ end (presumptive poly(A) site, as indicated). Data is plotted as normalized density relative to transcript abundance for Ago-mRNA clusters (blue) or control clusters (red), (standard deviations are shown in light colors). Regions with significant enrichment relative to control are indicated with black bars (> 3 standard deviations; P<0.003). Cluster enrichment ~1kb downstream from stop codon appears to be due to a large number of transcripts with ~1kb 3′ UTRs (data not shown). Bottom panel: All individual clusters (BC=5) are shown (each is a different color).

Figure 4.10

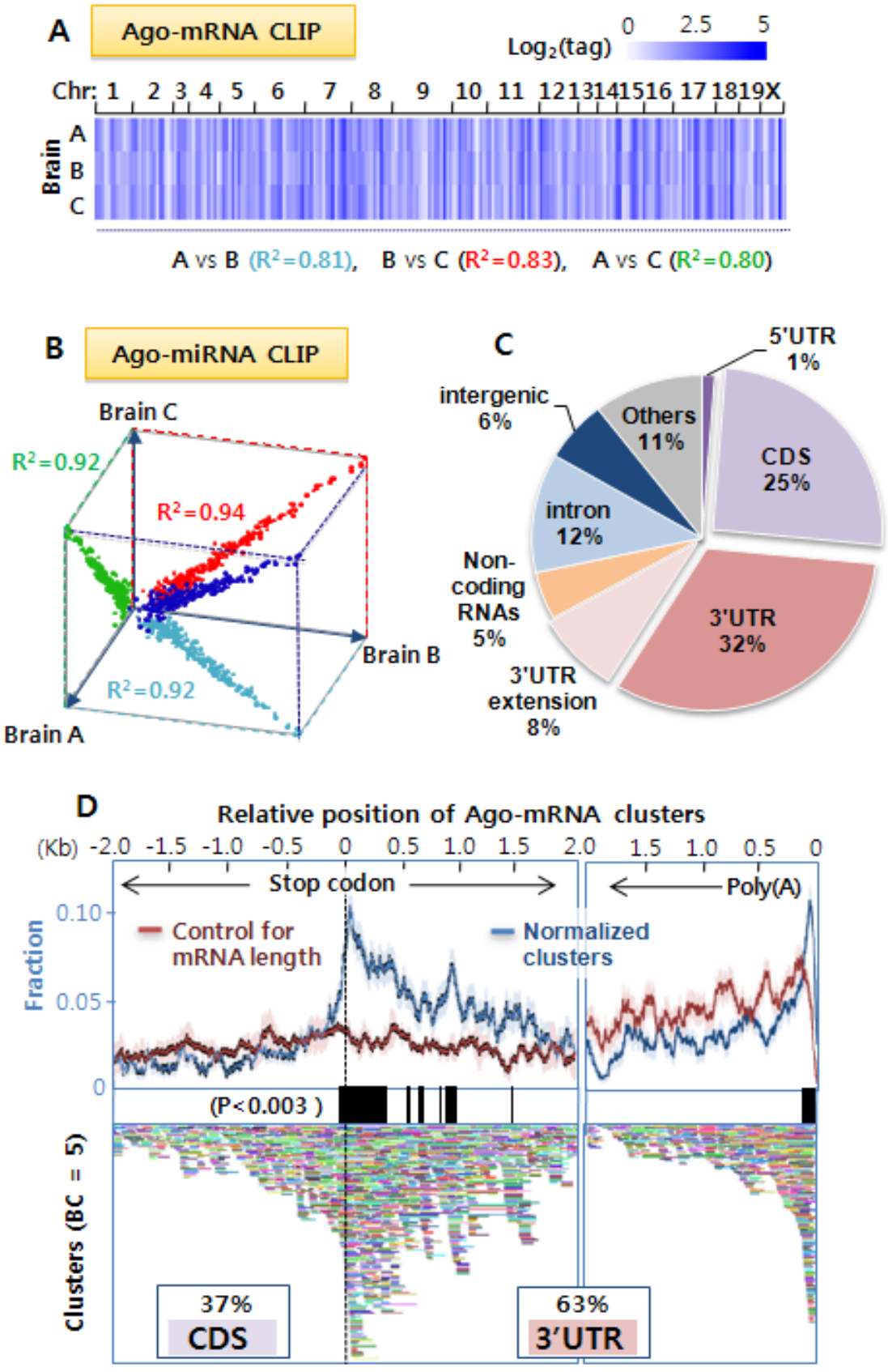

A
Ago-mRNA CLIP
Log2(tag)
0
2.5
5
Chr: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 X
Brain
A
B
C
A vs B (R²=0.81), B vs C (R²=0.83), A vs C (R²=0.80)
B
Ago-miRNA CLIP
Brain C
R²=0.92
R²=0.94
Brain B
R²=0.92
Brain A
C
5'UTR 1%
intergenic 6%
Others 11%
CDS 25%
intron 12%
Non-coding RNAs 5%
3'UTR 32%
3'UTR extension 8%
D
Relative position of Ago-mRNA clusters
(Kb) -2.0 -1.5 -1.0 -0.5 0 0.5 1.0 1.5 2.0 1.5 1.0 0.5 0
Stop codon
Poly(A)
Fraction
0.10
0.05
0
Control for mRNA length
Normalized clusters
(P<0.003)
Clusters (BC = 5)
37%
CDS
63%
3'UTR

Figure 4.11. Ago mRNA crosslinking sites within transcribed genes.

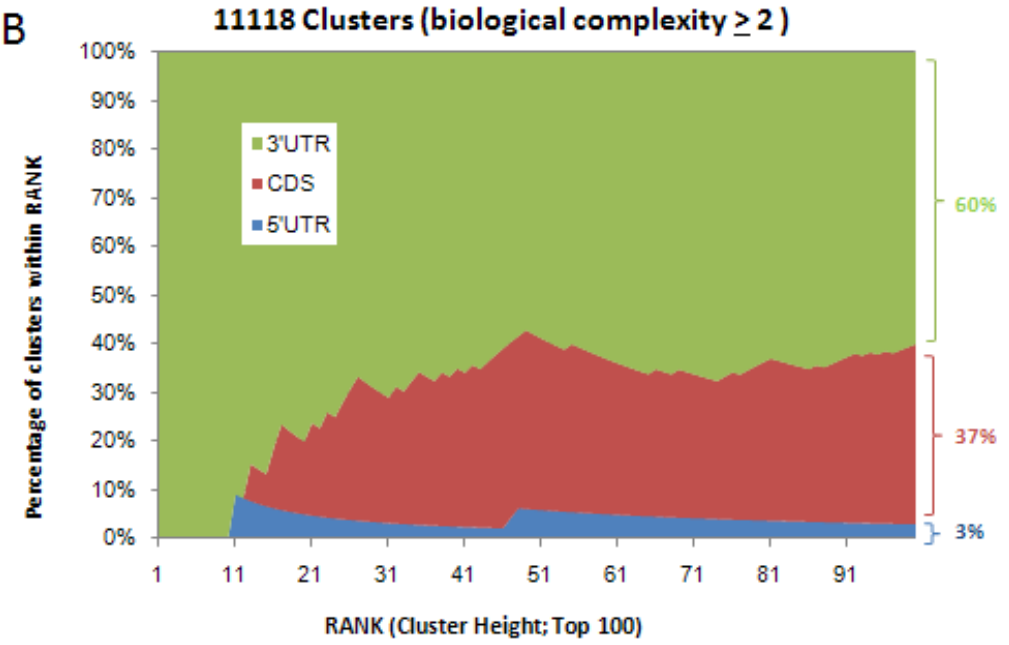
A. Ago mRNA CLIP tags and clusters according to their position in transcribed genes. The region of transcript (5′ UTR, coding sequence (CDS), or 3′ UTR) was determined by RefSeq annotation for all brain expressed transcripts (P13 mouse). The number of clusters refers to all clusters with BC > 2, and the number of tags refers to the number of tags within the cluster. Total length refers to the aggregate length of all transcribed brain sequences for each given transcribed region. Observed/Expected refers to the ratio of number of clusters or tags per expected number based on total length. P value was calculated by Chi square. Ago mRNA crosslinking sites in 3′UTR are ~1.5 fold enriched comparing to the expected number. B. The percentage of clusters in different regions of transcripts is shown according to their Rank (cluster height, defined as the maximum number of tags in each cluster). All clusters were analyzed, but only the top ranking 100 are shown. C. As in (B), with clusters BC = 5.

Figure 4.11

A **Enrichment of AGO-CLIP mRNA clustered tags in 3' UTR**

	Number of clusters (BC ≥ 2)	% Clusters	Number of tags	% tags	Total length (brain expressed transcripts)	% Total Length	Obs/Exp (number of clusters/ total length)	P-value	Obs/Exp (number of tags/ total length)	P-value
5'UTR	211	1.9	1,742	1.7	3,686,859	6.8	0.28	1.92×10^{-87}	0.25	0.00×10^{0}
CDS	4992	44.9	42,553	41.8	31,287,950	57.4	0.78	8.77×10^{-68}	0.73	0.00×10^{0}
3'UTR	5915	53.2	57,589	56.5	19,571,210	35.9	**1.48**	1.26×10^{-203}	**1.58**	0.00×10^{0}
Total	11,118	100	101,884	100	54,546,019	100				

B



C

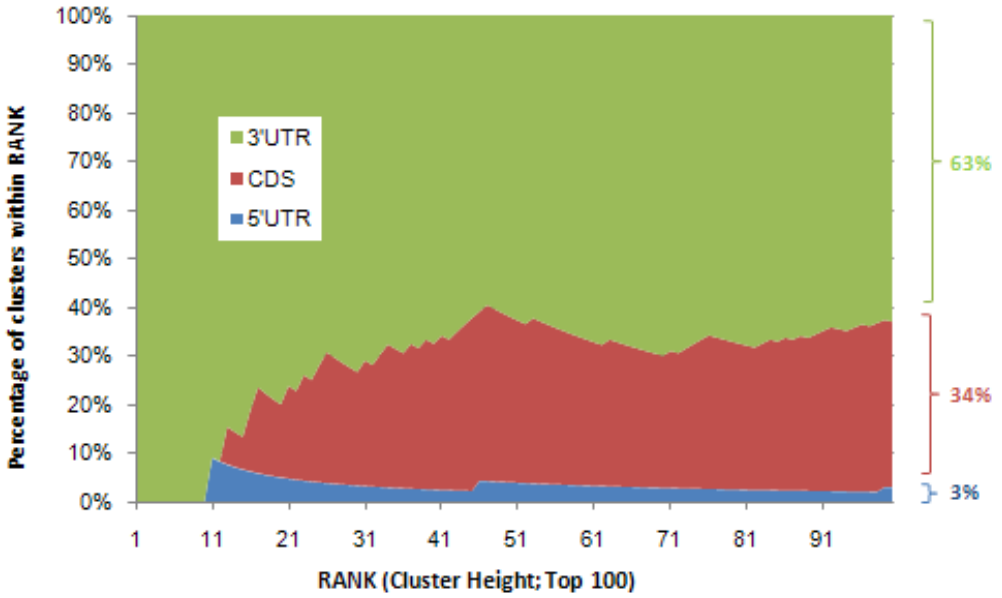


Table 4.1. High-throughput sequencing results from Ago HITS-CLIP.

Over 80% of RNAs isolated and sequenced from the upper 130 kD RNA-protein crosslinked band correspond to genomic sequences, and between 72 and 88% of RNAs from the lower 110 kD RNA-protein band correspond to miRNAs. Raw tags refer to number of Illumina tags after filtering step which company provided; BLAT refers to number of tags that could be aligned with genomic sequences from mm8, and BLAST from miRBASE, % is number of matched tags from BLAT search, using parameters (> 90% identity, Tag start <=3) or BLAST search result, filtered to include only matches with more than 95% identity. Unique tags are tags with different 5' ends after discarding those with identical genomic location miRBASE from May 2008 was used, which included 579 known mouse miRNAs (including passenger strand).

Table 4.1

A

From 130kD

Sample	Antibody	Age	Raw tags	BLAT (mm8)	% align	Unique tags
Brain A	2A8	P13	5,057,657	4,168,244	**82.4**	237,824
Brain B	2A8	P13	4,291,392	3,504,911	**81.7**	359,405
Brain C	2A8	P13	3,843,297	3,073,034	**80**	328,916
Brain D	7G1-1	P13	6,259,297	5,118,928	**81.8**	420,627
Brain E	7G1-1	P13	6,394,071	5,188,918	**81.2**	423,024

B

From 110kD

Sample	Antibody	Age	Raw tags	BLAST (miRBASE)	% align
Brain A	2A8	P13	4,918,586	3,717,702	**75.6**
Brain B	2A8	P13	4,160,480	3,004,104	**72.2**
Brain C	2A8	P13	3,490,474	2,576,516	**73.8**
Brain D	7G1-1	P13	4,853,688	4,275,690	**88.1**
Brain E	7G1-1	P13	5,409,367	4,765,957	**88.1**

Table 4.2. Verification of exon array usage for in silico random CLIP normalization.

To estimate the effect of varying (N) on the outcome of in silico CLIP, simulations were done with the gene Chad7 that has the minimum abundance used in our experiments and in in silico CLIP. The top table shows that Tn does not change as a function of abundance (Nn). The first column values indicate that the array value of every transcript was increased together with Chad7. All simulation has the same final result which selects Ago CLIP clusters of which height is more than 2. The middle table shows the maximum cluster heights over 500 simulations. No clusters are greater than 2 (Mn). The final FDR calculation based on the results of the middle table (e.g. 12/500=0.024) shows that for a cluster height of 3, the FDR does not change, with each FDR < 0.01. In conclusion, this simulation explains that the results are not affected by (N).

Table 4.2

Chad7 (NM 001081417)	(N_n) Exon array value	Length (Y_n)	Expected number of tags (T_n)
1x	100	9444	5
2x	200	9444	5
4x	400	9444	5
8x	800	9444	5
16x	1600	9444	5
32x	3200	9444	5
72x	7200	9444	5
100x	10000	9444	5
500x	50000	9444	5

Number of tags in the most overlapping region (Mn)

Simulation	3	2	1
1x	0	12	488
2x	0	52	448
4x	0	21	479
8x	0	57	443
16x	0	43	457
32x	0	20	480
72x	0	41	459
100x	0	61	439
500x	0	35	465

False discovery rate base on *in silico* CLIP simulation result

Simulation	3	2	1
1x	<0.002	0.024	0.976
2x	<0.002	0.104	0.896
4x	<0.002	0.042	0.958
8x	<0.002	0.114	0.886
16x	<0.002	0.086	0.914
32x	<0.002	0.04	0.96
72x	<0.002	0.082	0.918
100x	<0.002	0.122	0.878
500x	<0.002	0.07	0.93

CHAPTER V. DECODING OF MIRNA-MRNA INTERACTION MAPS

Introduction

MicroRNAs (miRNAs) are short (18-26 nt) sequences that act as post-transcriptional repressors of gene expression. 721 miRNAs have been reported in the human genome (miRBase release 14; 579 in mouse genome) and each is believed to bind directly to many mRNAs to regulate their translation or stability. Thus, miRNAs represent a key regulatory mechanism affecting numerous cellular activities, and are of particular interest in research about disease such as cancer (Kim et al., 2009b)

Despite the biologic importance of miRNAs, it has not been possible to make a general determination about the specific RNA targets upon which they act. The difficulty stems from the discovery that functional miRNA-mRNA interactions may require as few as 6 nucleotides (nt) of consecutive complementary seed sequences (Lim et al., 2005). Since such interactions are present on average every ~4 kb, it is possible that miRNAs bind to a very broad range of mRNAs, but the extent of their regulatory potential is not known. Bioinformatic analysis has

greatly improved the ability to recognize bona fide miRNA binding sites (Krek et al., 2005; Lewis et al., 2005; Rajewsky, 2006), principally by constraining searches for evolutionary conservation in 3′ UTR seed sequences. However, different algorithms still produce significantly divergent results, with significant false-positive rates (Bentwich, 2005). In addition, many miRNAs are encoded at multiple gene loci or are present in closely related miRNA families, complicating interpretation of loss of function studies in mammals (Ebert et al., 2007; Schaefer et al., 2007; Ventura, 2008; Xiao et al., 2007), although such studies have been informative for some miRNAs, particularly in invertebrates (Ambros, 2004; Johnston and Hobert, 2003b). Low sequence complexity and functional redundancy have complicated efforts at large-scale validation of individual miRNA-mRNA interaction sites *in vivo*. miRNA overexpression or knockdown studies, most recently in combination with proteomic studies (Baek et al., 2008; Selbach et al., 2008), have led to the conclusion that individual miRNAs generally regulate a relatively small number of proteins at modest levels (< 2-fold), although the false-positive rate of target predictions remains high (~up to 66%) (Baek et al., 2008), and the data sets analysed have been of limited size (~5,000 proteins). Similar high false-positive rates have been observed when miRNAs were co-immunoprecipitated with Ago proteins (Beitzinger et al., 2007; Easow et al., 2007; Hammell, 2008; Hendrickson et al., 2008; Zhang, 2007). A critical caveat

common to all of these studies is their inability to definitively distinguish direct from indirect miRNA-target interactions. At the same time, as therapeutic antisense strategies become more viable (van Rooij et al., 2008a; van Rooij et al., 2008b; Wheeler et al., 2007), knowledge of direct miRNA target sites has become increasingly important.

We used HITS-CLIP to covalently crosslink native Argonaute (Ago) protein-RNA complexes in mouse brain. This produced two simultaneous datasets—Ago-miRNA and Ago-mRNA binding sites. Here, these two Ago CLIP data were combined and decoded by bioinformatic analysis to identify miRNA-target mRNA interaction sites. We validated genome-wide interaction maps for miR-124, and generated additional maps for the 20 most abundant miRNAs present in P13 mouse brain. Ago HITS-CLIP provides a general platform for exploring the specificity and range of miRNA action *in vivo*, and identifies precise sequences for targeting clinically relevant miRNA-mRNA interactions.

Results

Decoding miRNA-mRNA interactions from Ago HITS-CLIP data

As a first estimate of whether a relationship between Ago-mRNA clusters and sites of miRNA action exists, we performed an unbiased search for all 6-8 nucleotides (nt) sequence motifs within the entire length of Ago-mRNA clusters, using linear regression analysis (Foat et al., 2005). The six most enriched motifs corresponded to seed sequences from the miRNAs most frequently crosslinked in Ago-miRNA CLIP (Table 5.1), with the most significant match corresponding to miR-124, a brain-specific miRNA ($P = 8.3 \times 10^{-58}$; this was not the same as the most frequently crosslinked Ago-miRNA (miR-30), perhaps reflecting over-representation of miR-124 seed sequences in the genome or contributions from rules of Ago binding that are not apparent). To more precisely define Ago-mRNA clusters, we analyzed 61 robust Ago-mRNA clusters (BC = 5; total tags > 30) using cubic spline interpolation analysis. We found that Ago bound within 45-60 nt of the cluster peaks 95% of the time (defined by analysis of overlapping Ago-mRNA tags, or by a minimum region common to all 61 clusters, tan and green graphs, respectively, Figure 5.1A), and we define this region as the Ago-mRNA footprint region. Linear regression again demonstrated a high correlation between the frequency of Ago-miRNAs and the frequency of their seed sequences in this Ago-mRNA footprint region (Figure 5.1B and Table 5.2).

We next examined the position and frequency of conserved seed sequences present in the top 30 miRNAs identified within the Ago-mRNA footprint. 171 seed sites were identified within 100 nt of 134 robust Ago-mRNA clusters, the majority of which (118) were located within the predicted 60 nt Ago-mRNA footprint (Figure 5.2A). The distribution of seed sequences appears to follow a leptokurtic distribution (one in which the peak is higher and sharper than a normal distribution; excess kurtosis (k) = 1.08, versus 0 in a normal distribution), illustrating the high resolution of CLIP to determine Ago-RNA interactions. As a control, seed sequences of a "negative" group of miRNAs (the lowest ranking miRNAs from the Ago-miRNA CLIP list) were uniformly distributed in these clusters (Figure 5.2A; k = -1.35, versus -1.2 in uniform distribution). Taken together, these results indicate that the Ago-mRNA footprint is rich in and may predict miRNA binding sites with enhanced specificity

Ago HITS-CLIP and miR-124

To further explore the relationship between the Ago footprint and miRNA binding we focused on miR-124 sites, a well studied brain-specific miRNA. We found a marked enrichment of conserved miR-124 seed sequences in Ago-mRNA clusters (Figure 5.2B). 86% of the predicted miR-124 binding sites were present within the Ago footprint region, again in a tight peak region showing leptokurtic

distribution (Figure 5.2B; k = 11.63). These results were more sensitive and sufficiently specific, relative to more stringent analyses (Figure 5.2A), such that we chose this threshold (BC > 2) for subsequent analyses. While some predicted seeds outside of the Ago footprint might correspond to false positives, we also noted small secondary peaks at ~ +/- 50 nt outside the Ago footprint, suggesting the possibility of cooperative secondary miR-124 binding sites in some transcripts, consistent with data suggesting miRNA cooperativity may occur within this range of distance (Grimson, 2007).

We searched published examples of miR-124 regulated transcripts for Ago-mRNA clusters with miR-124 seeds within the predicted 62 nt Ago footprint (termed Ago-miR-124 ternary clusters). We identified such ternary clusters 5 of 5 transcripts in which miR-124 seed sites had been well defined by functional studies (including mutagenesis of seed sequences in full length 3'UTR; Figure 5.3). In each transcript, there were many predicted miRNA target sites in the 3' UTR, but only a small number of Ago-mRNA ternary clusters found, suggesting that there may be a significant number of false positive predictions from bioinformatic algorithms (Figure 5.3). For example the 3' UTR of Itgb1 mRNA has ~50 predicted miRNA target sites, including two miR-124 sites, but only 5 Ago-mRNA ternary clusters (Figure. 5.3A). Using the Ago footprint to predict

miRNAs bound at these sites, we identified three as miR-124 sites, one of which was missed by computational predictions because the seed sequence is not well conserved across species (Figure. 5.3A); similar observations were made in the Ctdsp1 3′ UTR (Figure. 5.3C). Previous luciferase assays demonstrated that miR-124 suppression of Itgb1 (to 35% control levels) was partially reversed (to 85% control levels) by mutating both of the two predicted seed sequences (Yu et al., 2008); our observation of an Ago-miR-124 ternary cluster at this third non-conserved site may explain the partial rescue. Conversely, in the Polypyrimidine tract-binding protein 1 (Ptbp1) 3′ UTR, the absence of any Ago footprint at a predicted miR-9 seed site was consistent with prior studies which found this site to be non-functional (Makeyev et al., 2007)(Figure 5.3B). Additionally, Ptbp1 has seven predicted miR-124 seed sites, of which five were previously tested and only 2 found to be functional in luciferase assays (Makeyev et al., 2007); only these 2 sites harbored Ago-miR-124 ternary clusters (Figure 5.3B).

Validating Ago HITS-CLIP for miR-124

To extend these observations, we compiled brain-expressed transcripts from a meta-analysis of five published microarray experiments (Table 5.4), which sought to identify transcripts that were significantly downregulated by miR-124 overexpression in HeLa and other cell lines (Baek et al., 2008; Hendrickson et al.,

2008; Lim et al., 2005; Makeyev et al., 2007; Wang, 2006). Transcripts with predicted miR-124 seeds showed miR-124 dependent mRNA and protein suppression (Figure 5.5A-B), consistent with previous observations (Baek et al., 2008; Selbach et al., 2008). Remarkably, transcripts with Ago-miR-124 ternary clusters had a significant tendency to be downregulated at the RNA and protein level ($P<0.01$, relative to mir-124 seed prediction, Kolmogorov-Smirnov test, Figure 5.5A-B). We validated these studies experimentally by examining Ago-mRNA clusters that appeared de novo in HeLa cells (which do not express endogenous miR-124) after miR-124 transfection (Figure 5.4). Applying this data to the meta-analysis, mRNAs whose 3' UTRs harbored new Ago-miR-124 ternary clusters after miR-124 transfection showed an even greater enrichment in miR-124-dependent changes in transcript (Figure 5.5C) and protein (Figure 5.5D) levels ($P<0.01$, Kolmogorov-Smirnov test).

We next examined Ago-miR-124 ternary clusters present in validated individual transcripts identified from among 168 candidate miR-124 regulated transcripts (Table 5.3) (Lim et al., 2005). These targets had been further analyzed by Hannon and colleagues (Karginov et al., 2007) using a rigorous 3-part strategy to experimentally validate 22 of them (although miR-124 binding sites were not generally defined). We found that 16 of these 22 harbored Ago-miR-124 ternary

clusters in the 3′ UTR (Table 5.3). In 5 additional transcripts with low expression levels, Ago-mRNA CLIP-tags could be identified at predicted miR-124 seed sites. For transcripts of even moderate abundance (>700; average for P13 brain transcripts ~1,255), we identified all 10/10 predicted targets (Table 1). These data indicate that the sensitivity and specificity of detecting bona fide miR-124 targets is markedly enhanced by identification of Ago-miR-124 ternary clusters. We also examined this set of 22 targets (Karginov et al., 2007) for de novo Ago-mRNA clusters present in the 3′ UTR after miR-124 transfection (Table 5.3). Remarkably, from among many potential miR-124 seed sites, 17 de novo Ago-miR-124 ternary clusters appeared after transfection and 14 of 17 were at precisely the positions predicted from brain Ago miR-124 maps. Genome-wide, we identified 204 de novo Ago–miR-124 clusters with conserved seeds in mouse brain transcripts; of these 98 were independently identified as mouse brain Ago–miR-124 ternaryclusters. Taken together, these data demonstrate the Ago HITS-CLIP map predicts experimentally induced and functional sites of miRNA regulation that are not evident through bioinformatic analysis alone.

Predicting miRNA functional networks

Based on the robust correlation between previously validated miR-124 functional sites and Ago HITS-CLIP, we examined Ago-mRNA clusters to predict binding

maps for the 20 most abundant miRNAs identified in Ago-miRNA CLIP. These maps (Figure 5.7A) show Ago binds to target transcripts at very specific sites: on average there are only 2.6 Ago-mRNA clusters (BC $\geq$ 2) per transcript (2.12 per 3' UTR) and each miRNA binds an average of only 655 targets (Figure 5.6). To explore the potential of Ago HITS-CLIP maps to define miRNA regulated transcripts, we examined the functions encoded by the predicted targets of these 20 miRNAs using gene ontology analysis; comparison of these results with predictions made using GO analysis of TargetScan predictions (Figure 5.8) demonstrates that the FDR rate and 'quality' of the protein network deteriorate substantially when the Ago-mRNA map is not used (Figure 5.10). Target predictions from the Ago HITS-CLIP map suggest that diverse neuronal functions are regulated by different sets of miRNAs (Figure 5.7B). The largest set of miRNA associated functions, "neuronal differentiation" illustrates interwoven but distinct pathways predicted to be regulated by three miRNAs expressed in neurons (Figure. 5.7C). The Ago-RNA ternary map corresponds remarkably well with the current view of miR-124, miR-125 (Wu and Belasco, 2005) and miR-9 function, including actions to promote neurite outgrowth and differentiation by inhibiting Ago-miR-124 targets (Figure 5.7C) such as Itgb1 (Yu et al., 2008) (Figure 5.3A), Iqgap1 (Lim et al., 2005) (Figure 5.3E), Ptbp1 (Makeyev et al., 2007) (Figure 5.3B), and Sox9 (Cheng et al., 2009) (Figure 5.3F) and Ago-miR-9 targets

such as Fgfr1 (Leucht et al., 2008) (Figure 5.3G) and Foxg1 (Shibata et al., 2008) (Figure 5.3H).

Discussion

Ago-miRNA-mRNA ternary maps identify functionally relevant miRNA binding sites in living tissues, and were developed in the context of several recent studies. Crystallographic structures of Ago miRNA-mRNA ternary complexes (Wang et al., 2008a) demonstrated close contacts between all three molecules, consistent with the ability of CLIP (Ule et al., 2003), which requires close protein-RNA contacts to detect both Ago miRNA and mRNA interactions. The development of HITS-CLIP set the stage for generating and analyzing genome-wide RNA-protein maps in the brain (Licatalosi et al., 2008) and cultured cells (Yeo, 2009). High throughput experiments and bioinformatic analysis together generated genome-wide predictions of miRNA seed sequences, particularly of miR-124. These studies demonstrated that miR-124 simultaneously represses hundreds of transcripts (Baek et al., 2008; Hendrickson et al., 2008; Lim et al., 2005; Makeyev et al., 2007; Wang, 2006). and provided a genome-wide "gold standard" with which to compare Ago HITS-CLIP data.

Ago HITS-CLIP resolves some roadblocks that have arisen in efforts to understand miRNA action. It has been difficult to discriminate direct from indirect actions of miRNAs, and to extrapolate miRNA overexpression studies in tissue culture to organismal miRNA action. Target RNAs have previously been identified by immunoprecipitation, microarray analysis (Karginov et al., 2007) (Hendrickson et al., 2008) and reporter validation assays, with the concern that low stringency immunoprecipitation of non-crosslinked RNA-protein complexes, including Ago-miRNAs (Mili and Steitz, 2004), may purify indirect targets. This has spurred interest in efforts to explore miRNA-target identification by covalently crosslinking, using formaldehyde or 4-thio-uridine modified RNA in culture to identify transcripts complexed with Ago, miRNAs and additional proteins (Kirino and Mourelatos, 2008) (Vasudevan et al., 2007). HITS-CLIP offers a clear means of identifying direct Ago targets and identifying specific interaction sites, which in turn offers the possibility of specifically targeting miRNA activity.

Ago-HITS-CLIP complements bioinformatic approaches to miRNA target identification by restricting the sequence space to be analyzed to the ~45-60 nt Ago footprint. For highly conserved 3′ UTRs, such as those of the RNABPs Ptbp2, Nova1, and Fmr1, many miRNA sites are predicted using algorithms that

rely on sequence conservation, but each has only one Ago-mRNA CLIP cluster (Figure 5.9). In fact, miRNA selectivity is very high (Figure 5.3 and 5.9) such that on average transcripts have between 1-3 major Ago binding sites in a single tissue (Figure 5.6). Ago-mRNA binding sites themselves have no apparent sequence preference, suggesting that accessibility may rely on additional RNABPs. Such a mechanism, which may be assessed by overlaying HITS-CLIP maps of different RNABPs (Licatalosi et al., 2008), could provide a means of dynamically regulating miRNA binding and regulation.

By simultaneously generating binding maps for multiple miRNAs, Ago HITS-CLIP offers a new approach to understanding combinatorial control of target RNA expression. Both the FDR rate and evidence for such miRNA target protein networks deteriorate substantially when predicted maps are generated without experimental Ago HITS-CLIP data (Figure 5.10). At the same time, analysis of a single miRNA, miR-124, demonstrated that its expression not only induced Ago to bind miR-124 sites, but reduced or precluded Ago binding to sites occupied in untransfected cells (Table 5.3), perhaps reflecting competition between a limited capacity for miRNA binding on a given 3′ UTR. Such Ago occlusion has important mechanistic, experimental and clinical implications, where studies manipulating miRNA levels are envisioned.

Contributions

All bioinformatic analysis, transfection of miRNA mimics and Ago CLIP experiments were done by Sung Wook Chi under the supervision of Robert B. Darnell. Aldo Mele helped with HITS-CLIP experiments. Scott Dewell helped with high-throughput sequencing (Illumina Genome Analyser).

Figure 5.1. Distribution of mRNA tags

(a) Ago-mRNA cluster width and peaks. The peaks of 61 robust clusters (BC = 5, peak height >30, with single peaks) were determined, and the position of tags (brown lines and fraction plotted as brown graph) and width of individual clusters (green lines and fraction plotted as green graph) are shown relative to the peaks. The minimum region of overlap of all clusters (100%) was within -24 and +22 nt of cluster peaks, and > 95% were within -30 and +32 nt, suggesting that the Ago footprint on mRNA spans 62 nt (or, more stringently, 46 nt). (b) A linear regression model was used to compare miRNA seed matches enriched in the stringent Ago footprint region with the frequency of miRNAs experimentally determined by Ago-miRNA HITS-CLIP. For each miRNA in the database, 3 different six-mers (corresponding to positions 1-8) were compared with sequences in Ago-mRNA cluster peaks (-24 to +22) and analyzed by linear regression model according to the peak height. 22 values with $P < 0.05$ were selected.

Figure 5.1

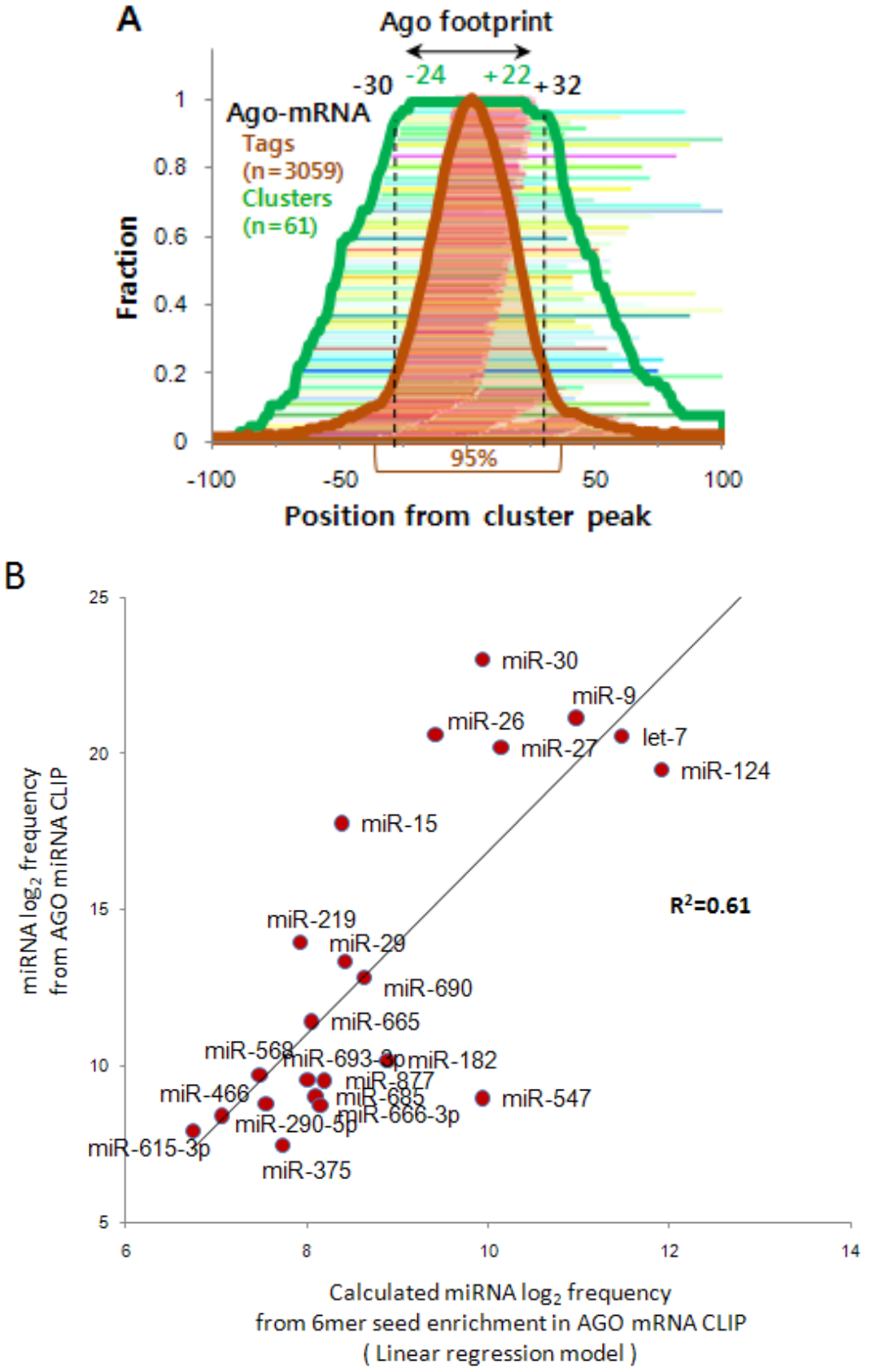

A
Ago footprint
-30
-24
+22
+32
Ago-mRNA
Tags
(n=3059)
Clusters
(n=61)
Fraction
1
0.8
0.6
0.4
0.2
0
95%
-100
-50
50
100
Position from cluster peak
B
miR-30
miR-9
miR-26
miR-27
let-7
miR-124
miR-15
R²=0.61
miR-219
miR-29
miR-690
miR-665
miR-568
miR-693-3p
miR-182
miR-877
miR-466
miR-685
miR-547
miR-666-3p
miR-290-5p
miR-615-3p
miR-375
miRNA log2 frequency
from AGO miRNA CLIP
25
20
15
10
5
6
8
10
12
14
Calculated miRNA log2 frequency
from 6mer seed enrichment in AGO mRNA CLIP
(Linear regression model)

Figure 5.2. Distribution of mRNA tags correlates with seed sequences of miRNAs from Ago CLIP.

A. The position of conserved core seed matches (position 2-7) from the top 30 Ago-bound miRNAs (independent of peak height; purple colors, including miR-124, red; or, bottom 30 miRNAs, expressed at extremely low levels in brain; black to gray colors) are plotted relative to the peak of 134 robust clusters (BC = 5, peak height >30). 118 of 171 seed matches are within the Ago 62 nt footprint B. The position of conserved miR-124 seed matches (bottom panel; each is represented by a different color) were plotted relative to the peak position of all Ago-mRNA clusters (BC > 2). Top panel; distribution of mir-124 seed matches (plotted relative to cluster peak, normalized to number of clusters; red graph); pale color indicates standard deviation. Excess kurtosis (k), a measure of the "peakedness" of the probability distribution of a real-valued random variable, indicates that seed sites are present in a sharp peak relative to a normal distribution.

Figure 5.2

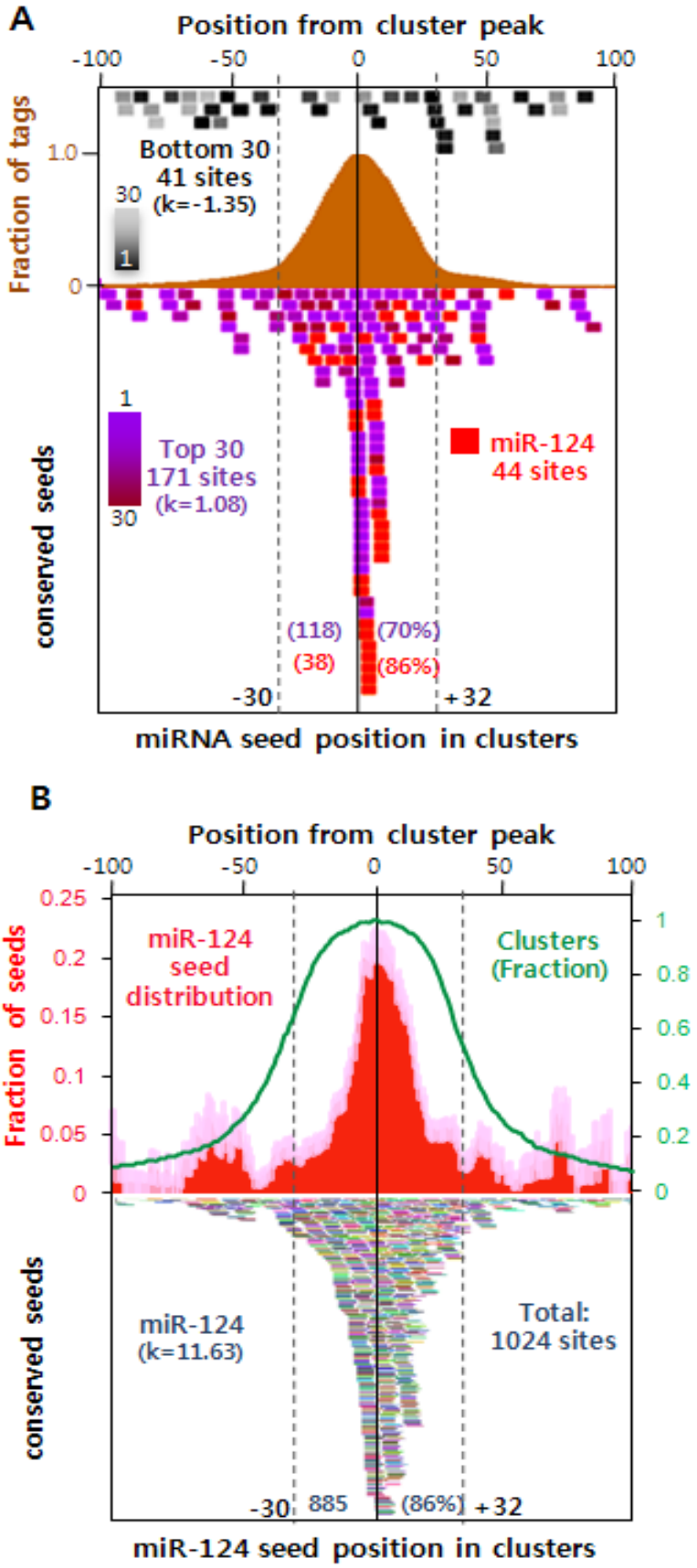

A
Position from cluster peak
-100
-50
0
50
100
Fraction of tags
1.0
0
Bottom 30
41 sites
(k=-1.35)
30
1
1
Top 30
171 sites
(k=1.08)
30
miR-124
44 sites
conserved seeds
(118)
(70%)
(38)
(86%)
-30
+32
miRNA seed position in clusters
B
Position from cluster peak
-100
-50
0
50
100
0.25
0.2
0.15
0.1
0.05
0
Fraction of seeds
miR-124
seed
distribution
Clusters
(Fraction)
1
0.8
0.6
0.4
0.2
0
conserved seeds
miR-124
(k=11.63)
Total:
1024 sites
-30
885
(86%)
+32
miR-124 seed position in clusters

Figure 5.3. Ago-miRNA ternary clusters in validated miR-124 mRNA targets.

A. Ago–mRNA CLIP tags (top panel: raw tags, one color per biologic replicate: second panel: ternary map of Ago-mRNA normalized clusters around top predicted miRNA sites) compared with predicted miRNA sites (using indicated algorithms) for the 3′ UTR of Itgb1 (third panel; colors indicate predicted top 20 miRNAs as in Fig. 5; grey bars indicate miRNAs ranking below the top 20 (per heat-map) in Ago-miRNA CLIP). All predicted miR-124 6-8 mer seeds (conserved (red) or non-conserved (yellow) sites) are shown. Bottom panel shows data from luciferase assays in which mutagenesis of predicted miR-124 seeds at the indicated positions had the indicated effects on restoring from miR-124 mediated suppression (35% of baseline luciferase levels). B. Ago-miRNA ternary maps compared with previously reported functional data for Ptbp1. C. Ago ternary map of Ctdsp1. D. Ago ternary map of Vamp3. E-F.I. Iqgap1 and Sox9 are previously identified as miR-124 targets. The fragments used for luciferase assay are shown with black bars. The Ago ternary map identified exact sites of validated miRNA binding within the fragments as shown. G-H. Validated miR-9 targets, Fgfr1 and FoxG1 are shown. Fgfr1 is validated as miR-9 target in Zebrafish.

Figure 5.3

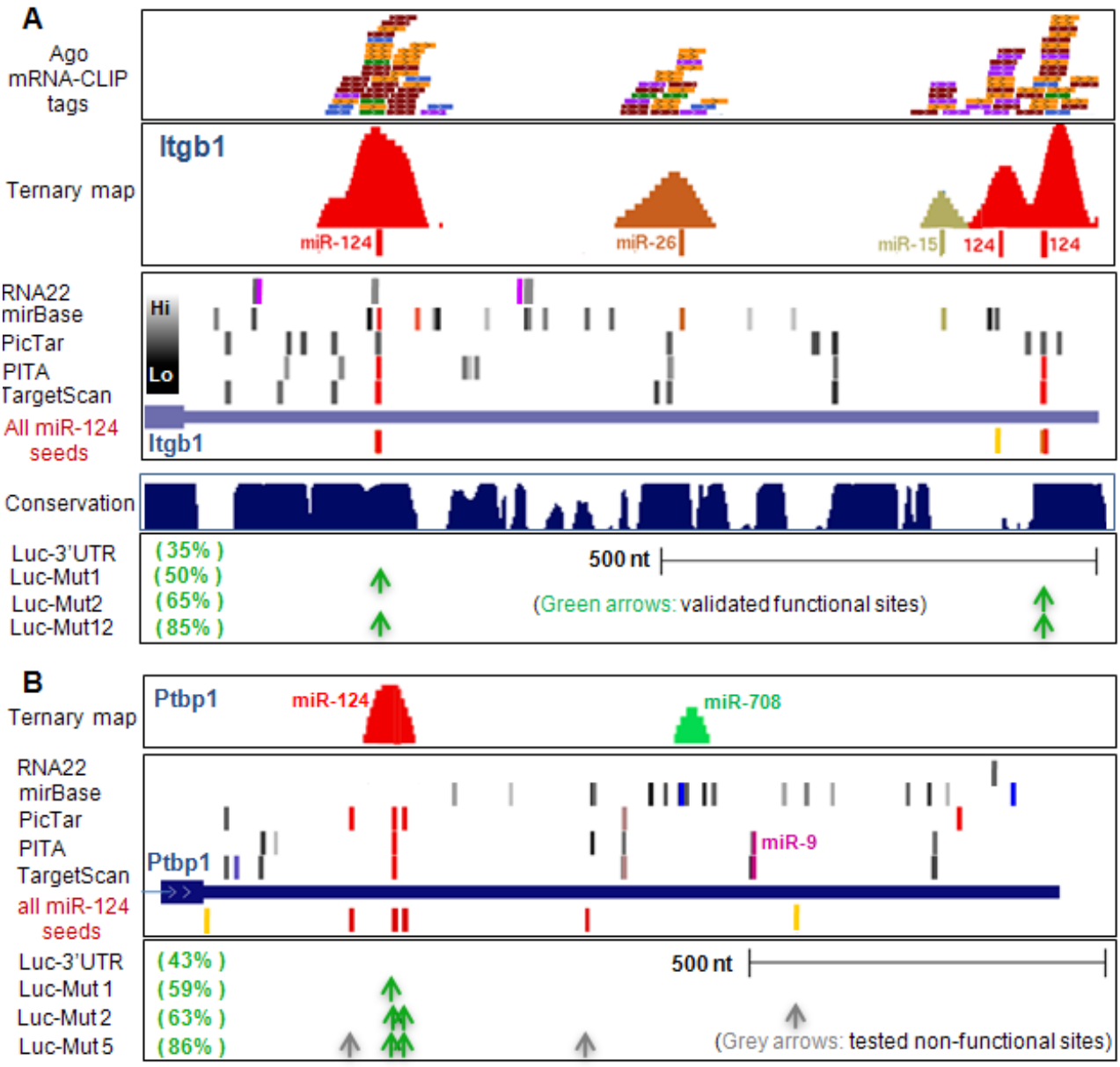

A
Ago mRNA-CLIP tags
Ternary map
Itgb1
miR-124
miR-26
miR-15
124
124
RNA22
mirBase
PicTar
PITA
TargetScan
All miR-124 seeds
Hi
Lo
Itgb1
Conservation
Luc-3'UTR
Luc-Mut1
Luc-Mut2
Luc-Mut12
(35%)
(50%)
(65%)
(85%)
500 nt
(Green arrows: validated functional sites)
B
Ternary map
Ptbp1
miR-124
miR-708
RNA22
mirBase
PicTar
PITA
TargetScan
all miR-124 seeds
Ptbp1
miR-9
Luc-3'UTR
Luc-Mut 1
Luc-Mut 2
Luc-Mut 5
(43%)
(59%)
(63%)
(86%)
500 nt
(Grey arrows: tested non-functional sites)

Figure 5.3 continues

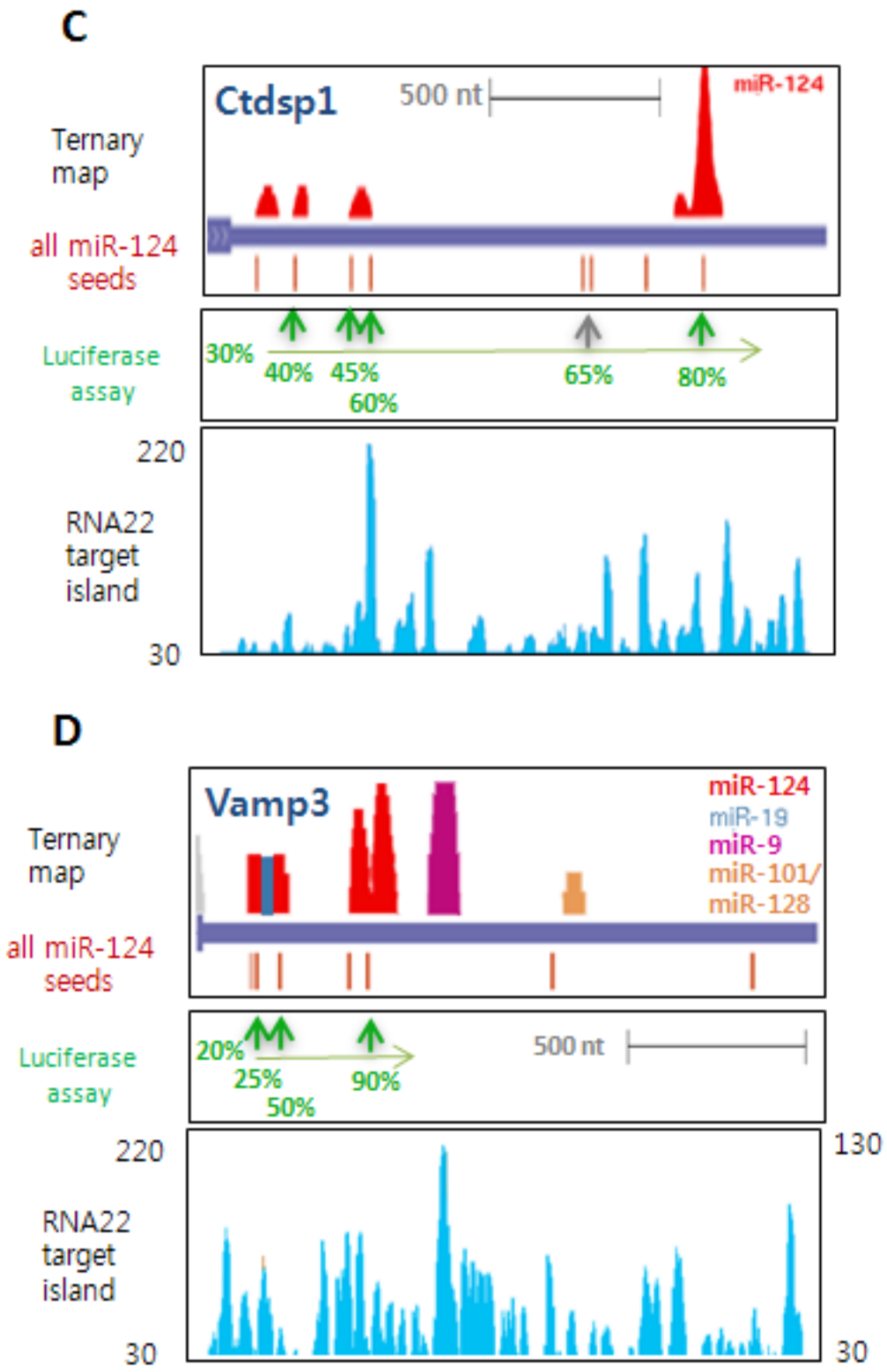


Figure 5.3 continues

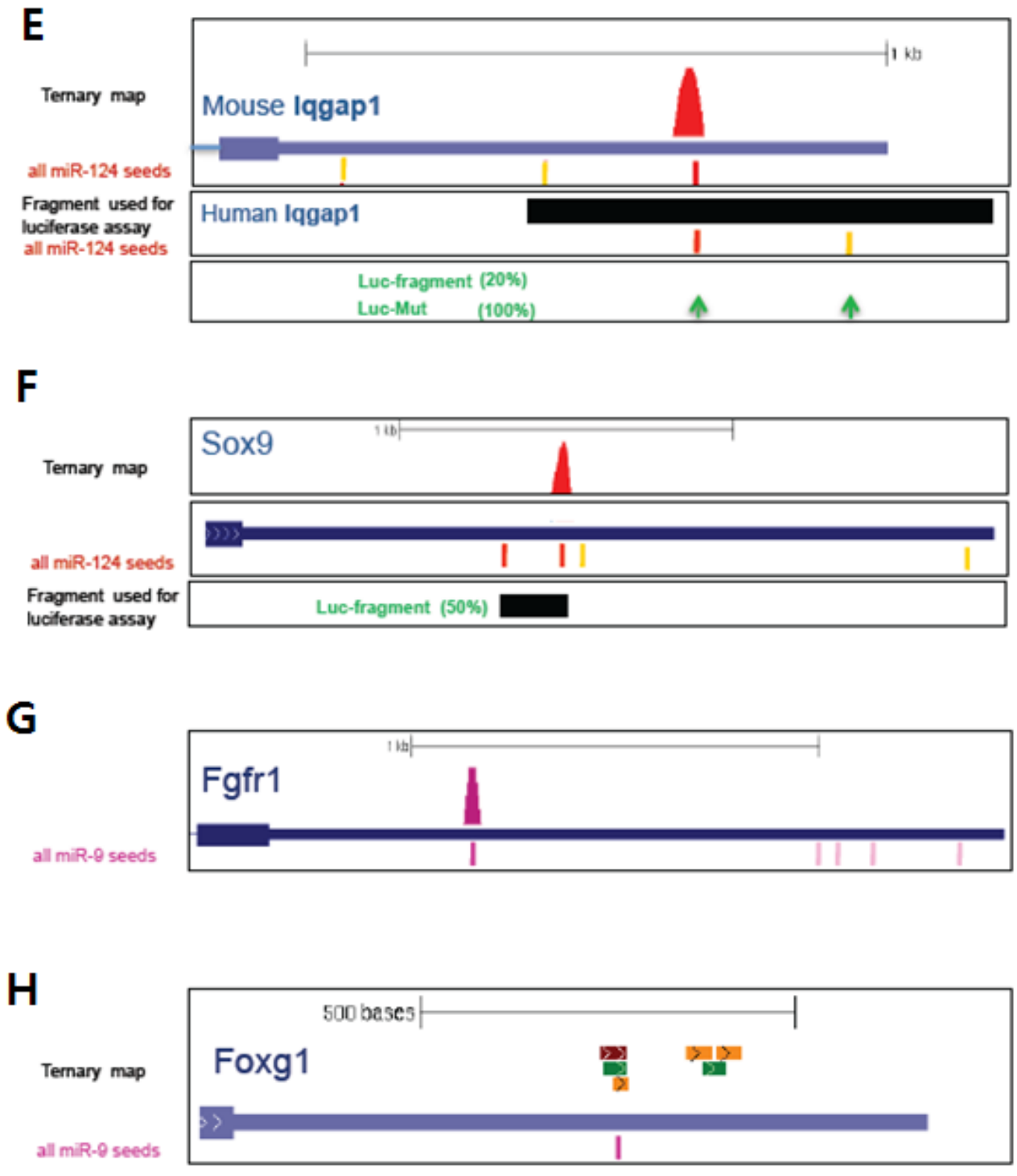
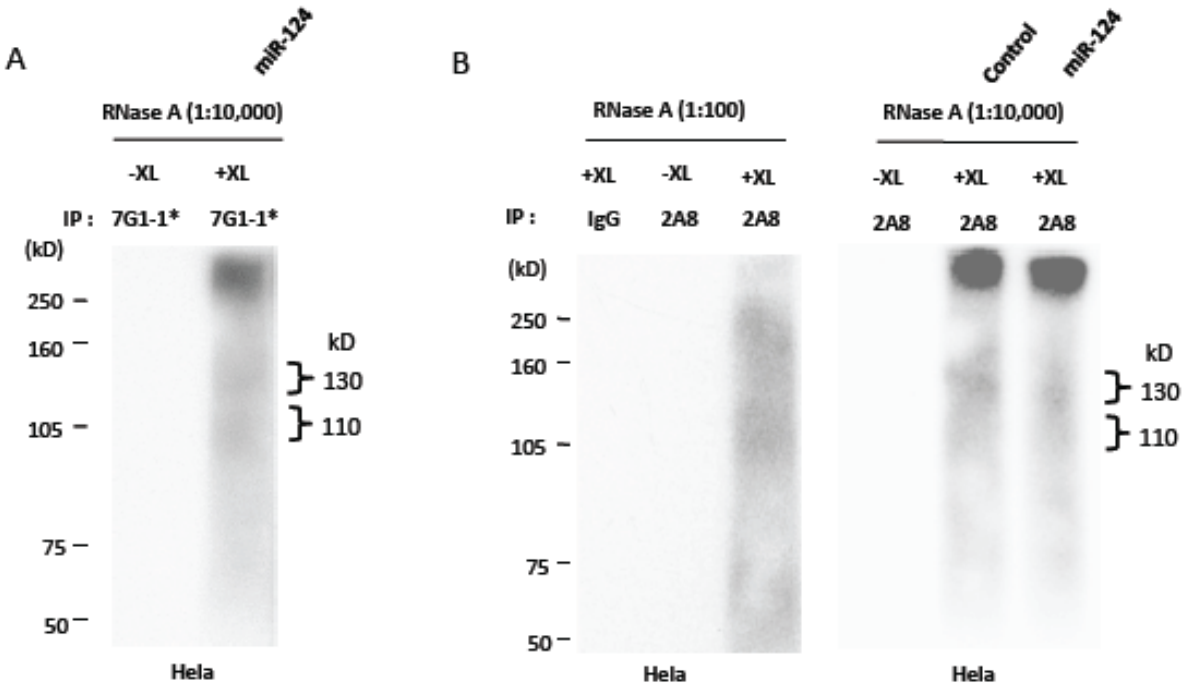


Figure 5.4. Argonaute HITS-CLIP in HeLa cells transfected with control or miR-124 microRNAs.

Argonaute HITS-CLIP in HeLa cells transfected with control or miR-124 microRNAs. A. 32P-labeled RNA crosslinked to Ago immunoprecipiated with 7G1-1* from miR-124 or control miRNA transfected HeLa cells are shown. Two Ago RNA cross-linked complexes (~130kD for Ago-mRNA and ~110kD for Ago-miRNA) are specific in UV-irradiated tissue (+XL) compared to non-irradiated brain (–XL). B. The same result as A in a CLIP experiment using 2A8. Also shown is a comparison of overdigestion of Ago RNA cross-linked complex (1:100 RNase A). C. HITS-CLIP sequencing results of RNAs isolated from the 130 kD RNA-protein complex in 8 HeLa experiments with two different miRNA transfections (miR-124 and control miRNA) and two different antibodies (2A8 and 7G1-1). Raw tags (prefiltered) refer to number of Illumina tags prior to default filtering; aligned tags refers to number of tags that could be aligned with genomic sequences from human genome database (hg18), derived from the combination of two alignment results using BLAT and ELAND (Efficient Large-Scale Alignment of Nucleotide Databases; provided by Illumina) program with same criteria used in AGO-CLIP in mouse brain (> 90% identity, Tag start <=3). Unique tags are tags with different 5′ ends or different degenerative 4 nucleotide barcode introduced in 5' fusion linker.

Figure 5.4



C

Sample	Transfected miRNA	Antibody	Raw tags (prefiltered)	Aligned tags (hg18)	% align	Unique tags
HeLa 1	miR-124	2A8	11,563,568	6,082,765	52.6	1,271,461
HeLa 2	miR-124	2A8	12,681,689	5,715,019	45.1	1,435,945
HeLa 3	miR-124	7G1-1	12,484,953	5,816,073	46.6	1,459,849
HeLa 4	miR-124	7G1-1	12,441,955	5,419,481	43.6	1,603,967
HeLa 5	control miRNA	2A8	10,802,378	5,951,833	55.1	1,184,079
HeLa 6	control miRNA	2A8	11,104,169	5,813,633	52.4	1,247,774
HeLa 7	control miRNA	7G1-1	12,204,108	6,916,662	56.7	1,272,437
HeLa 8	control miRNA	7G1-1	12,957,786	3,396,298	26.2	1,127,326

Figure 5.5. Meta-analysis of Ago-mRNA clusters in large-scale screens of miR-124 regulated targets.

A. Transcripts with predicted (conserved) miR-124 seeds (purple line) showed miR-124 suppression relative to all transcripts expressed in brain and cell lines (blue line) or those with no miR-124 seed sequences (green line). Transcripts with Ago-miR-124 ternary clusters (containing both miR-124 seed sequences and Ago-mRNA CLIP tags; red line) showed further miR-124 suppression. B. Similar results were seen when analyzing miR-124 dependent protein suppression (identified by SILAC), with discrimination by the presence of Ago-miR-124 ternary clusters especially evident where there were smaller numbers of transcripts showing larger changes (log2<-0.4; inset). C. Transcripts expressed in miR-124 transfected HeLa cells that harbor new Ago-miR-124 clusters (red line; or a subset of transcripts also harboring Ago-miR-124 clusters in mouse brain; yellow line), compared with previous analysis7 of regulated transcripts in miR-124 transfected HeLa cells. D. As in C, plotted for predicted protein levels, compared with prior data.

Figure 5.5

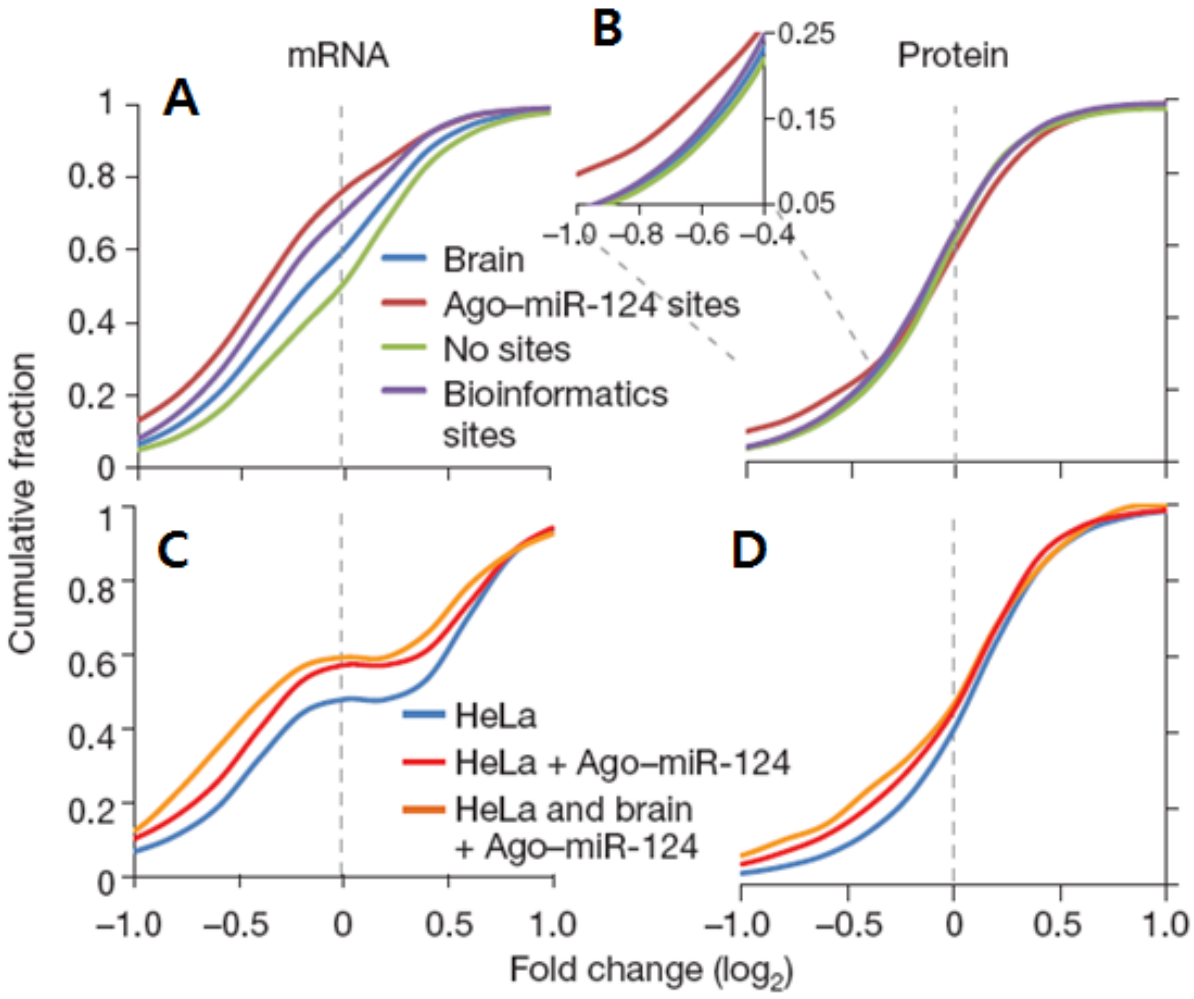

A
mRNA
B
0.25
0.15
0.05
−1.0 −0.8 −0.6 −0.4
Protein
Brain
Ago–miR-124 sites
No sites
Bioinformatics sites
C
D
HeLa
HeLa + Ago–miR-124
HeLa and brain + Ago–miR-124
Cumulative fraction
0
0.2
0.4
0.6
0.8
1
−1.0 −0.5 0 0.5 1.0
Fold change (log2)

Figure 5.6. AGO miRNA ternary map.

A. Table representing number of transcripts with given number of Ago-mRNA clusters (BC > 2) in 3'UTR, which shows average ~2 clusters per transcript in 3'UTR. The average number of clusters/transcript, Ago-regulated transcript, or 3' UTR of Ago-regulated transcripts are shown. B. Number of transcripts predicted as each top 20 Ago-miRNA targets (collapsed by family; see "pink" lines in Supplementary Table 3), based on Ago miRNA ternary map. C. Ago miRNA ternary map of top 20 Ago-miRNAs in mouse genome.

Figure 5.6

A

BC>=2

Number of Ago-mRNA clusters in 3'UTR	Number of 3'UTRs
1	1665
2	584
3	262
4	130
5	87
6	49
7	34
8	27
9	17
10	9
>10	43

Averages:
1.03 clusters/transcript in brain
2.60 clusters/Ago regulated transcript
2.12 clusters /3'UTR of Ago regulated transcript

B

miRNA family	Number of target transcripts
miR-124	1254
miR-9	1207
miR-27	1080
let-7	1060
miR-15	787
miR-181	740
miR-101	631
miR-34	626
miR-30	610
miR-26	602
miR-138	573
miR-17	536
miR-125	517
miR-708	498
miR-19	486
miR-153	455
miR-193	448
miR-21	358
miR-344	320
miR-221	315

Average 655 targets / miRNA

C

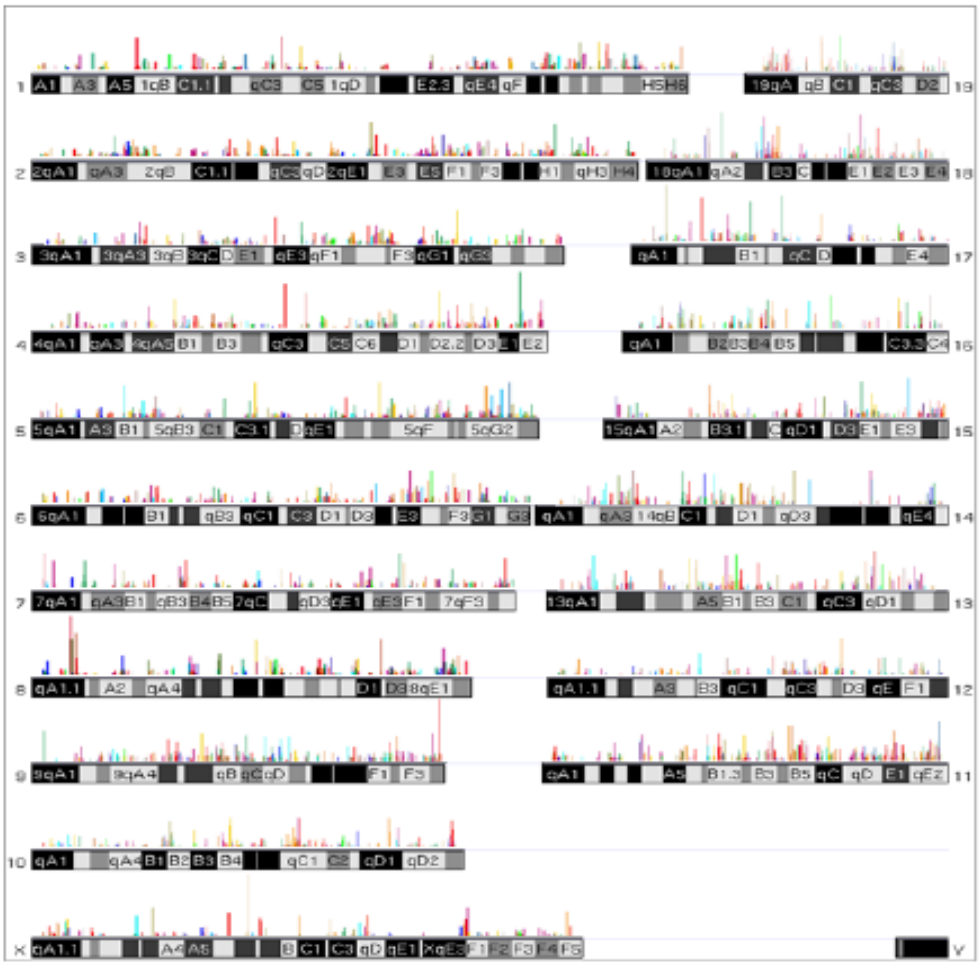


Figure 5.7. Analysis of Ago-miRNA ternary maps.

(a) Genome-wide views of Ago-miRNA ternary maps for the top 20 miRNAs from Ago HITS-CLIP (colors represent individual miRNA targets as indicated in (b)) are shown for the Itgb1 gene (top panel), the local gene region (middle panel; all transcripts are expressed in P13 brain except those outlined in grey boxes), showing tags in neighboring 3′ UTRs, and for all of chromosome 8 (bottom panel). (b) Heat map derived from gene ontology (GO) analysis of transcripts identified as targets of each of the top 20 miRNAs. Tree shows the hierarchical clustering of miRNAs based on GO. Significant clusters are outlined with black boxes. (c) Ago HITS-CLIP targets are shown for the significant pathways (neuronal differentiation/cytoskeleton regulation; based on FDR; for miR-124, miR-9 and miR-125 in mouse brain. Actin cytoskeleton pathways are shown based on KEGG database (http://www.genome.jp/kegg/).

Figure 5.7

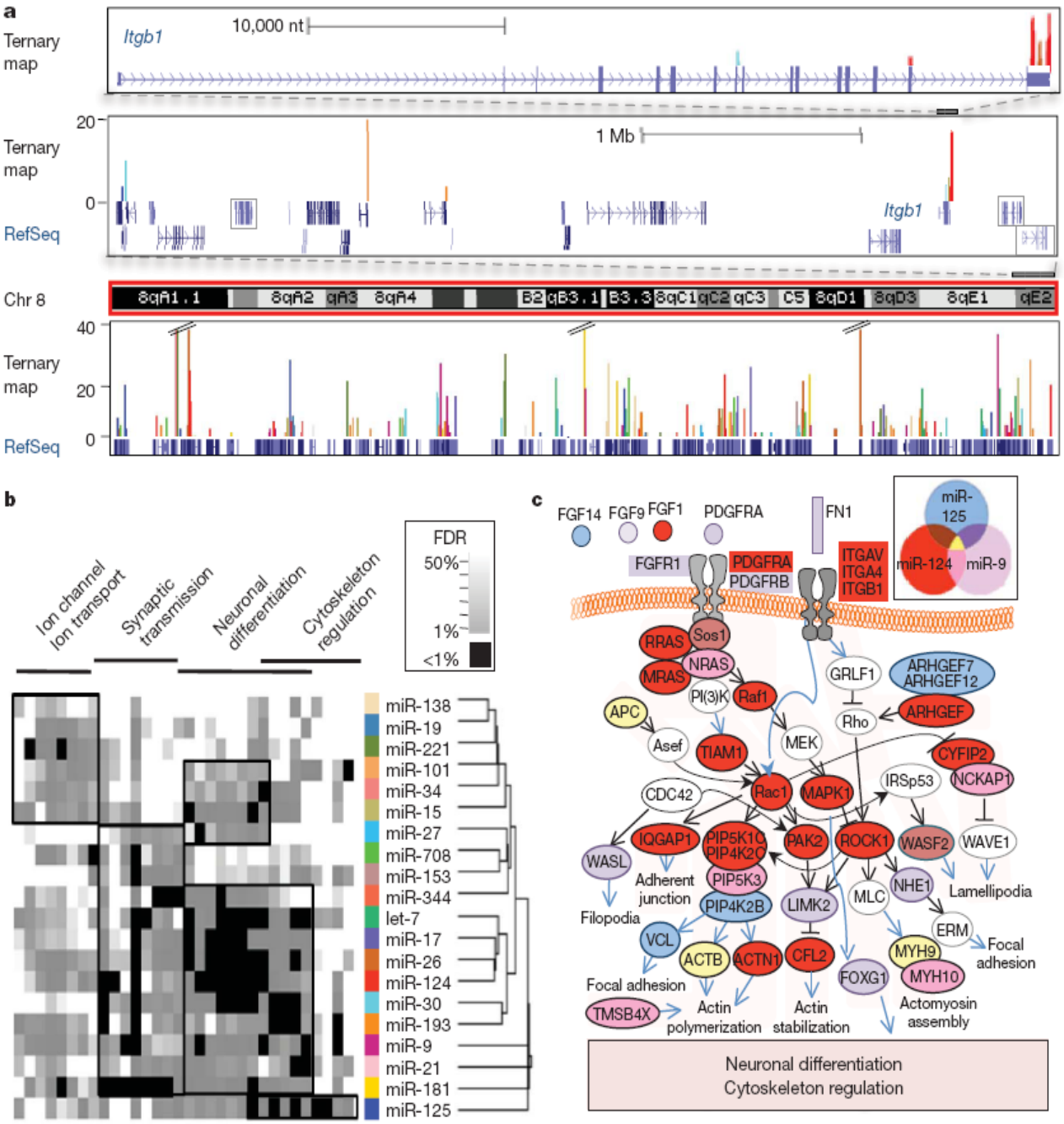

a
Ternary map
Itgb1
10,000 nt
20
Ternary map
0
1 Mb
RefSeq
Itgb1
Chr 8
8qA1.1
8qA2
qA3
8qA4
B2
qB3.1
B3.3
8qC1
qC2
qC3
C5
8qD1
8qD3
8qE1
qE2
40
Ternary map
20
0
RefSeq
b
Ion channel Ion transport
Synaptic transmission
Neuronal differentiation
Cytoskeleton regulation
FDR
50%
1%
<1%
miR-138
miR-19
miR-221
miR-101
miR-34
miR-15
miR-27
miR-708
miR-153
miR-344
let-7
miR-17
miR-26
miR-124
miR-30
miR-193
miR-9
miR-21
miR-181
miR-125
c
FGF14
FGF9
FGF1
PDGFRA
FN1
miR-125
miR-124
miR-9
FGFR1
PDGFRA
PDGFRB
ITGAV
ITGA4
ITGB1
RRAS
Sos1
NRAS
MRAS
PI(3)K
Raf1
GRLF1
ARHGEF7
ARHGEF12
ARHGEF
Rho
APC
Asef
TIAM1
MEK
CYFIP2
NCKAP1
IRSp53
CDC42
Rac1
MAPK1
IQGAP1
PIP5K1C
PIP4K2C
PAK2
ROCK1
WASF2
WAVE1
WASL
Adherent junction
PIP5K3
PIP4K2B
LIMK2
MLC
NHE1
Lamellipodia
Filopodia
VCL
ACTB
ACTN1
CFL2
ERM
MYH9
MYH10
Focal adhesion
FOXG1
Focal adhesion
TMSB4X
Actin polymerization
Actin stabilization
Actomyosin assembly
Neuronal differentiation
Cytoskeleton regulation

Figure 5.8. Gene ontology analysis of AGO ternary map.

This figure shows that target transcripts for each of the Ago top 20 miRNA ternary complexes that were used for gene ontology analysis. A. For each category, enrichment was compared with all P13 brain expressed transcripts using GoMiner (http://discover.nci.nih.gov/gominer/). False discovery rate (FDR) was calculated for each GO category and is represented as a different color as indicated. Hierarchical clustering of miRNAs and GO categories was performed using Cluster program and visualized as heat map and tree by Treeview (http://rana.lbl.gov/EisenSoftware.htm). GO categories were divided into two groups, one with little or no miRNA enrichment, and a second with a large amount of miRNA enrichment. The second group is shown here. B. As in (A), but using targets identified by TargetScan 4.1 rather than Ago HITS-CLIP, are shown for comparison. C-D. GO analysis using DAVID (http://david.abcc.ncifcrf.gov/) illustrating different results of the predicted actin cytoskeleton miRNA regulatory map for miR-124, miR-9 and miR-125 that are obtained with Ago HITS-CLIP and TargetScan 4.1. Taken together, these figures illustrate that both the FDR rate and evidence for a protein network deteriorate substantially when Ago-mRNA tags are not used.

Figure 5.8

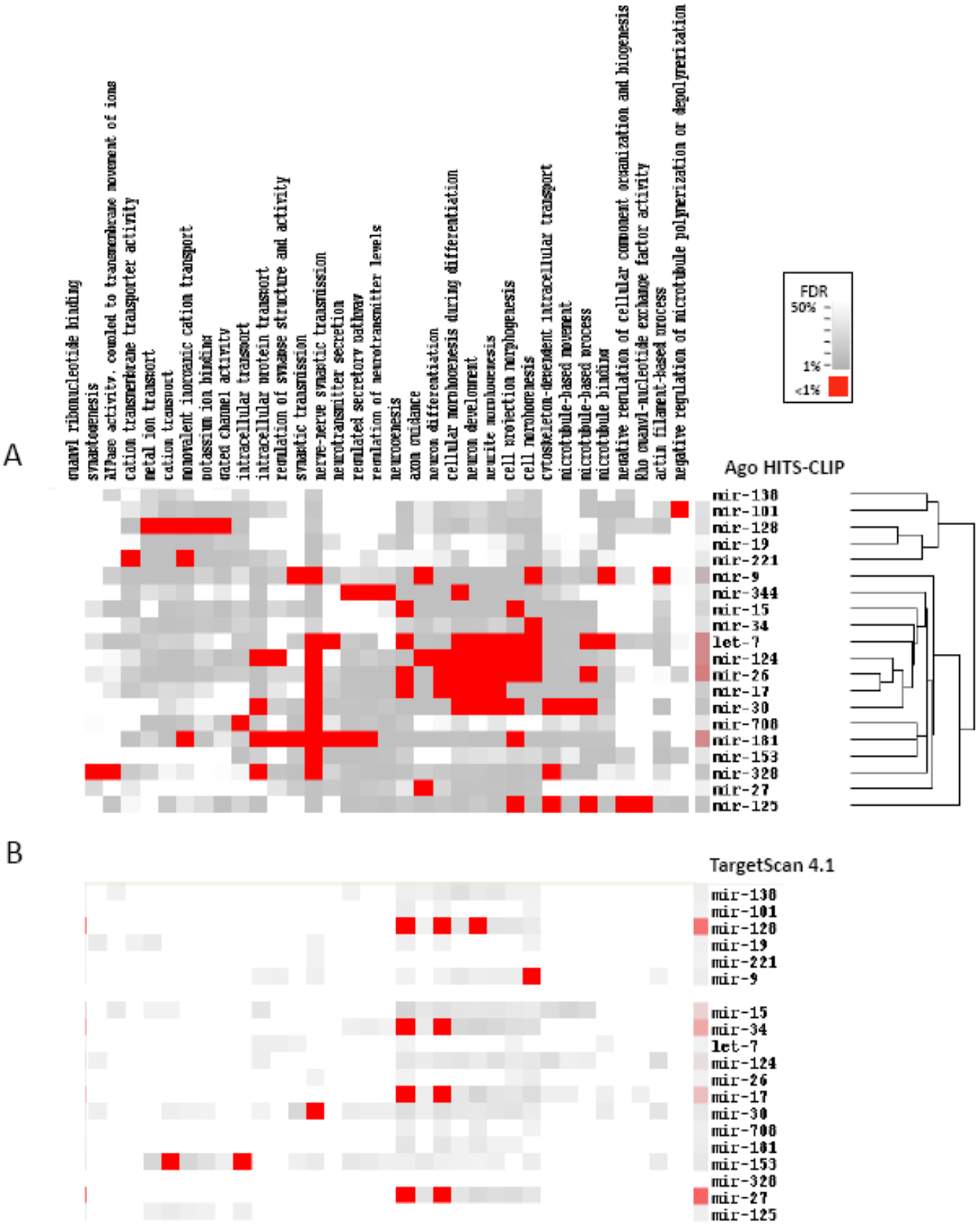

A
guanyl ribonucleotide binding
synaptogenesis
ATPase activity, coupled to transmembrane movement of ions
cation transmembrane transporter activity
metal ion transport
cation transport
monovalent inorganic cation transport
potassium ion binding
gated channel activity
intracellular transport
intracellular protein transport
regulation of synapse structure and activity
synaptic transmission
nerve-nerve synaptic transmission
neurotransmitter secretion
regulated secretory pathway
regulation of neurotransmitter levels
neurogenesis
axon guidance
neuron differentiation
cellular morphogenesis during differentiation
neuron development
neurite morphogenesis
cell projection morphogenesis
cell morphogenesis
cytoskeleton-dependent intracellular transport
microtubule-based movement
microtubule-based process
microtubule binding
negative regulation of cellular component organization and biogenesis
Rho guanyl-nucleotide exchange factor activity
actin filament-based process
negative regulation of microtubule polymerization or depolymerization
FDR
50%
1%
<1%
Ago HITS-CLIP
mir-130
mir-101
mir-128
mir-19
mir-221
mir-9
mir-344
mir-15
mir-34
let-7
mir-124
mir-26
mir-17
mir-30
mir-700
mir-181
mir-153
mir-320
mir-27
mir-125
B
TargetScan 4.1
mir-130
mir-101
mir-128
mir-19
mir-221
mir-9
mir-15
mir-34
let-7
mir-124
mir-26
mir-17
mir-30
mir-700
mir-181
mir-153
mir-320
mir-27
mir-125

Figure 5.9. Ago miRNA ternary map in miRNA targets.

This figure shows the same analysis as in Figure 3 on five of additional 3′ UTRs. A-C. Although many miRNA binding sites are predicted in those neuronal genes due to high conservation in 3′UTR, only one Ago-miRNA ternary cluster was identified in each transcript by AGO-HITS-CLIP in P13 mouse brain

Figure 5.9

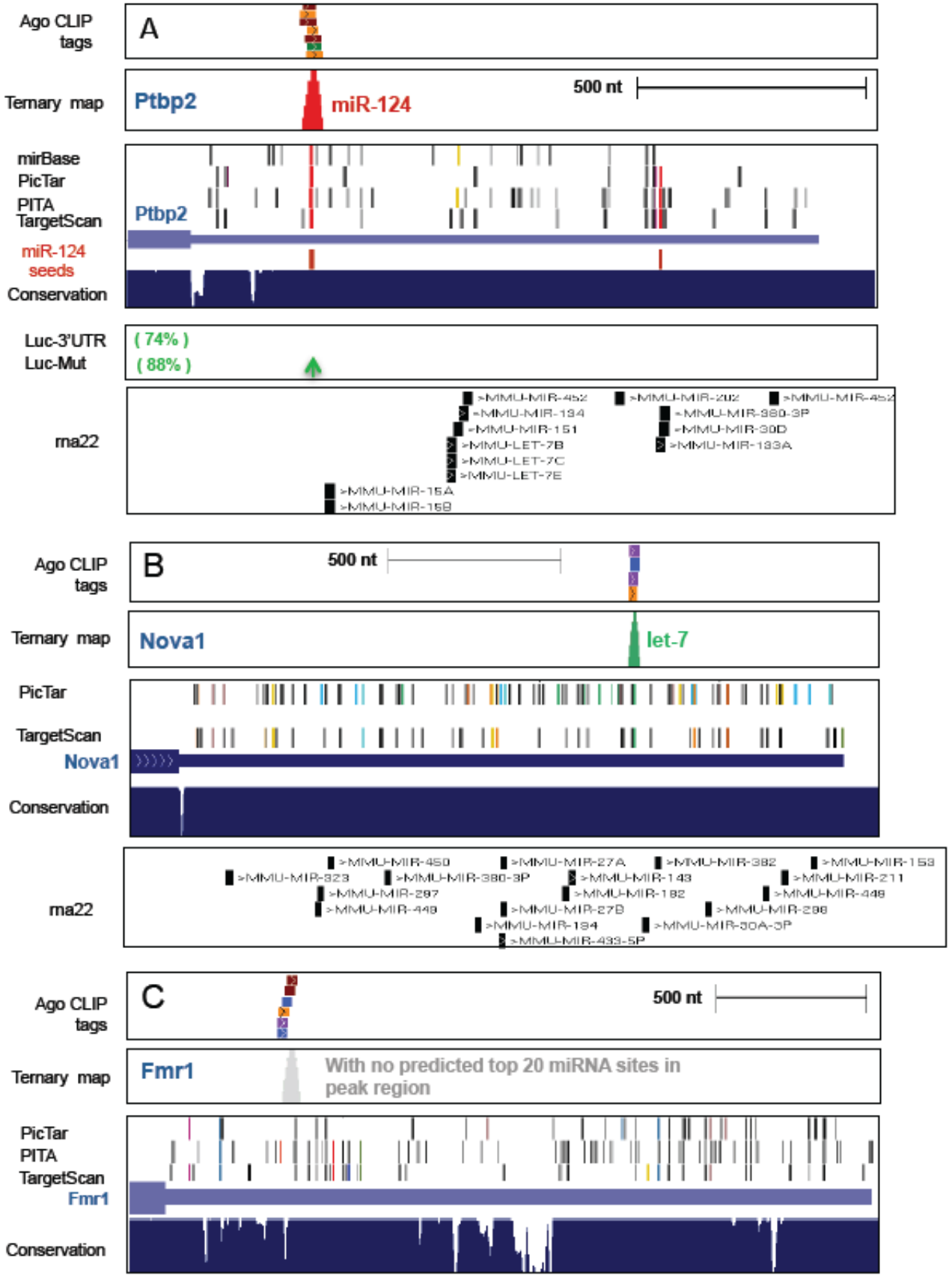

Ago CLIP tags
A
Ternary map
Ptbp2
miR-124
500 nt
mirBase
PicTar
PITA
TargetScan
Ptbp2
miR-124 seeds
Conservation
Luc-3'UTR
Luc-Mut
(74%)
(88%)
rna22
>MMU-MIR-452
>MMU-MIR-202
>MMU-MIR-194
>MMU-MIR-380-3P
>MMU-MIR-151
>MMU-MIR-30D
>MMU-LET-7B
>MMU-MIR-133A
>MMU-LET-7C
>MMU-LET-7E
>MMU-MIR-15A
>MMU-MIR-15B
Ago CLIP tags
B
500 nt
Ternary map
Nova1
let-7
PicTar
TargetScan
Nova1
Conservation
rna22
>MMU-MIR-450
>MMU-MIR-27A
>MMU-MIR-382
>MMU-MIR-153
>MMU-MIR-323
>MMU-MIR-380-3P
>MMU-MIR-143
>MMU-MIR-211
>MMU-MIR-297
>MMU-MIR-182
>MMU-MIR-448
>MMU-MIR-448
>MMU-MIR-27B
>MMU-MIR-298
>MMU-MIR-194
>MMU-MIR-30A-3P
>MMU-MIR-433-5P
Ago CLIP tags
C
500 nt
Ternary map
Fmr1
With no predicted top 20 miRNA sites in peak region
PicTar
PITA
TargetScan
Fmr1
Conservation

Figure 5.10. Gene ontology and pathway analysis of AGO mRNA ternary map

GO analysis using DAVID (http://david.abcc.ncifcrf.gov/) illustrating different results of the predicted actin cytoskeleton miRNA regulatory map for miR-124, miR-9 and miR-125 that are obtained with Ago HITS-CLIP (A) and TargetScan 4.1 (B). Taken together, these figures illustrate that both the FDR rate and evidence for a protein network deteriorate substantially when Ago-mRNA tags are not used.

Figure 5.10

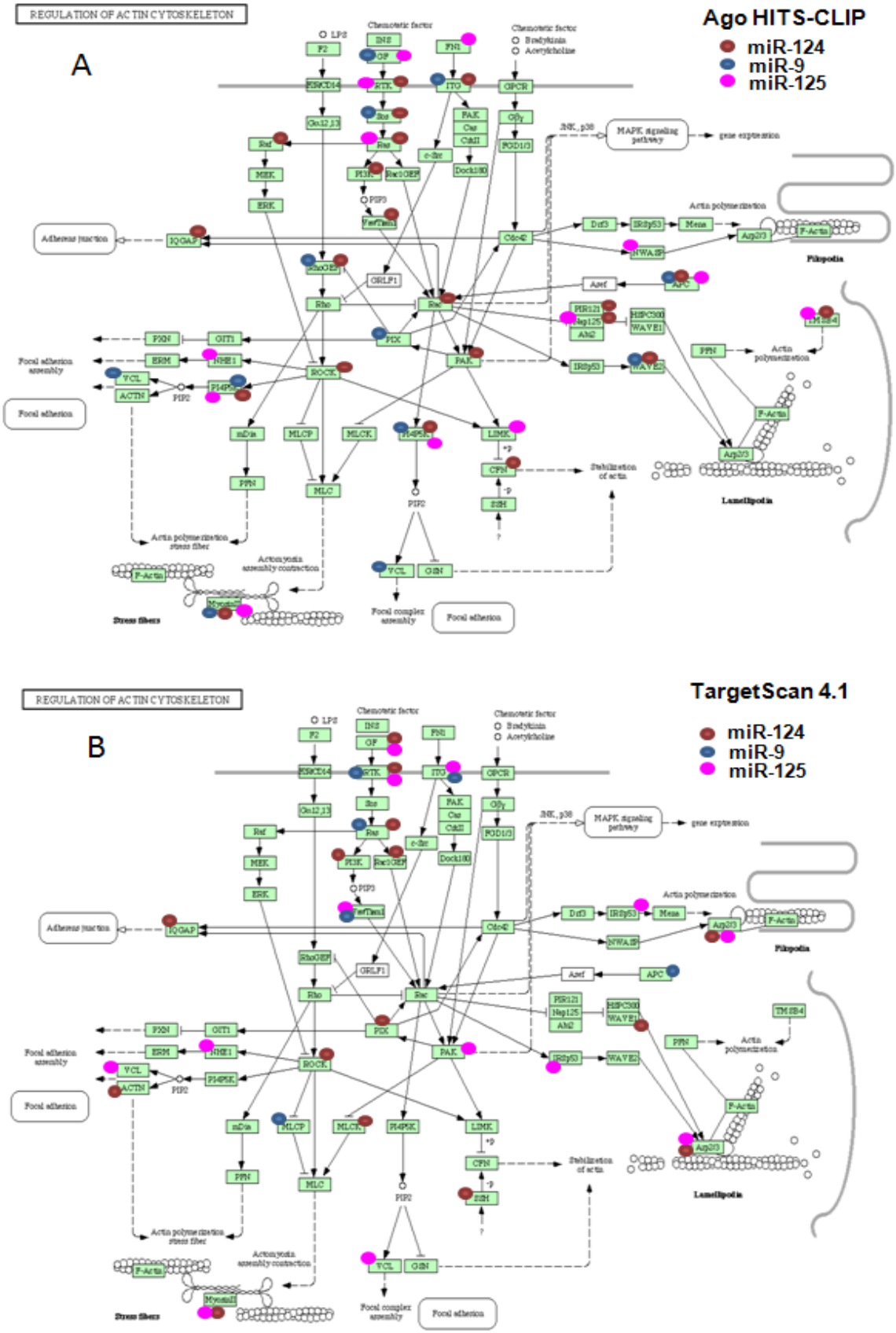

REGULATION OF ACTIN CYTOSKELETON
A
Ago HITS-CLIP
miR-124
miR-9
miR-125
REGULATION OF ACTIN CYTOSKELETON
B
TargetScan 4.1
miR-124
miR-9
miR-125

Table 5.1. Correlation of Ago-miRNA and Ago-mRNA CLIP data

A. We correlated all possible 6 to 8-mer motifs present in Ago-mRNA clusters (BC > 2) with cluster height (log2(maximum tag number) per cluster), using the MatrixREDUCE linear regression model: (http://bussemaker.bio.columbia.edu/software/MatrixREDUCE/11). The most enriched motifs selected by default parameters (except –max_gap=0) are shown. B. Since some miRNAs were missing from the analysis in (A), we analyzed the 7mer seeds (position 2-8) present in the top 10 miRNAs identified from Ago-miRNA CLIP. We compared with their observed (obs) versus expected (exp; from P13 brain transcripts) frequency in Ago-mRNA clusters (BC > 2). Seed sequences from 8 of the top 10 Ago-CLIP miRNAs were enriched in Ago-mRNA clusters. The miR124 seed match sequence is over-represented in the mouse genome and also in the mouse exons by about 50% (data not shown). This means that there are more miR-124 potential binding sites in the genome (we calculated this as 1.33 fold (51,461 total 6-mers for miR-124, vs. 38,752 for miR-30), and we interpreted this correlation as being consistent with a relative increase in miR-124 bound sites. It is also possible that there is underlying biologic regulation that we do not understand, or that there are additional rules of binding that are not apparent in current analyses.

Table 5.1

A

AGO-mRNA CLIP | AGO-miRNA CLIP

Top ranking motif	Regression coefficient	Number of tags	P-value	miRNA	Position of motif	Percentage in Ago-miRNA CLIP tags	Rank
TGCCTT	0.521	9153	8.3×10^{-58}	mir-124	2-7	2.97%	8
TACCTC	0.512	6403	1.6×10^{-40}	let-7	2-7	6.24%	5
CCAAAG	0.393	6878	3.1×10^{-32}	mir-9	2-7	9.33%	2
CTGTGA	0.388	4205	5.0×10^{-23}	mir-27	2-7	4.80%	6
TGTTTA	0.436	3206	7.9×10^{-21}	mir-30	3-8	34.01%	1
CTTGAA	0.356	2977	1.5×10^{-14}	mir-26	1-6	6.39%	4

B

AGO-miRNA CLIP | AGO-mRNA CLIP

Rank	7mer seed motif	miRNA	Number of Ago-miRNA CLIP tags (Total)	%	7mer seed obs/exp	Number of mRNA tags with seed (obs)	Number of clusters with seed
1	TGTTTAC	miR-30	8,417,151	34.29	3.304	1521	172
2	ACCAAAG	miR-9	2,309,955	9.41	2.175	1676	211
3	TGAATGT	miR-181	2,050,154	8.35	1.549	862	132
4	TACTTGA	miR-26	1,581,514	6.44	2.436	687	68
5	CTACCTC	let-7	1,549,064	6.31	4.000	2636	268
6	ACTGTGA	miR-27	1,188,538	4.84	1.190	969	137
7	AGCTCCT	miR-708	1,172,287	4.78	0.202	83	12
8	GTGCCTT	miR-124	735,430	3.00	5.187	3362	294
9	CACTGCC	miR-34	333,089	1.36	0.312	147	30
10	CTCAGGG	miR-125	330,186	1.35	1.187	506	65

Table 5.2. Generalized correlation of Ago-miRNA and Ago-mRNA CLIP.

A. Analysis as in Supplementary Table 2, but correlating all possible 6 to 8-mer motifs present in a narrowly defined Ago-mRNA footprint region (-24 to + 22 nt from peak, a region in which 100% of clusters are represented). All top ranking motifs are seed sequences from top 8 ranked miRNAs in AGO-miRNA CLIP B. For each miRNA in the database, 3 different six-mers (corresponding to positions 1-8) were compared with sequences in Ago-mRNA cluster peaks (-24 to +22) and analyzed by linear regression model according to the peak height. 22 values with $P < 0.05$ were selected.

Table 5.2

AGO-mRNA CLIP				AGO-miRNA CLIP			
Top ranking motif	Regression coefficient	Number of unique tags	P-value	miRNA	Matched seed position	Fraction in AGO-miRNA CLIP	Rank in AGO-miRNA CLIP
GTGCCTT	0.655	4581	9.01×10^{-31}	mir-124	2-8	2.97%	8
TGCCTT	0.421	9153	3.74×10^{-30}		2-7		
GTGCCT	0.247	6915	8.45×10^{-9}		3-8		
TACCTC	0.452	6403	2.13×10^{-25}	let-7	2-7	6.24%	5
TACCTCA	0.529	3558	1.82×10^{-13}		2-8		
CCAAAG	0.287	6878	9.22×10^{-14}	mir-9	2-7	9.33%	2
CCAAAGA	0.376	3549	2.06×10^{-12}		1-7		
TGTTTAC	0.404	2136	3.29×10^{-8}	mir-30	2-8	34.01%	1
GTTTAC	0.331	2964	7.19×10^{-8}		2-7		
CTGTGA	0.274	4205	1.47×10^{-9}	mir-27	2-7	4.80%	6
ACTGTGA	0.394	1665	3.01×10^{-7}		2-8		
TACTTGAA	0.676	877	3.89×10^{-8}	mir-26	1-8	6.39%	4
ACTTGA	0.290	2353	3.48×10^{-6}		2-7		

BC>=2, -24 ~ +22 region

Table 5.3. Ago-miR-124 ternary maps in brain and transfected HeLa cells.

Left panel: Comparison of 22 validated miR-124 targets44 with Ago-miR-124 ternary clusters in brain transcripts (BC > 2); miR-124 clusters are shown graphically above each 3′ UTR and corresponding conserved (red) or non-conserved (yellow) miR-124 seeds below) were identified in 16 of 22 validated transcripts (gene names colored red). In 5 additional transcripts (names colored orange) clusters were below normalization threshold, but individual Ago-mRNA CLIP tags are shown. Tom1l1 had no detectable Ago binding, although notably a brain-expressed antisense transcript (Cox11) is present. Previously reported changes in miR-124 dependent transcript levels from miR-124 transfected HeLa cells7, and levels of the same transcript ("tx") in mouse brain (normalized probe intensities) are shown. Right panels: Columns represent Ago-miR-124 ternary clusters in HeLa cells transfected with either miR-124 or a control microRNA ("-miR-124"). 17 de novo Ago-miR-124 ternary clusters (red; present only after miR-124 transfection) are shown, together with clusters present in both control and miR-124 transfected cells (purple), or present in control HeLa cells alone (blue).

Table 5.3

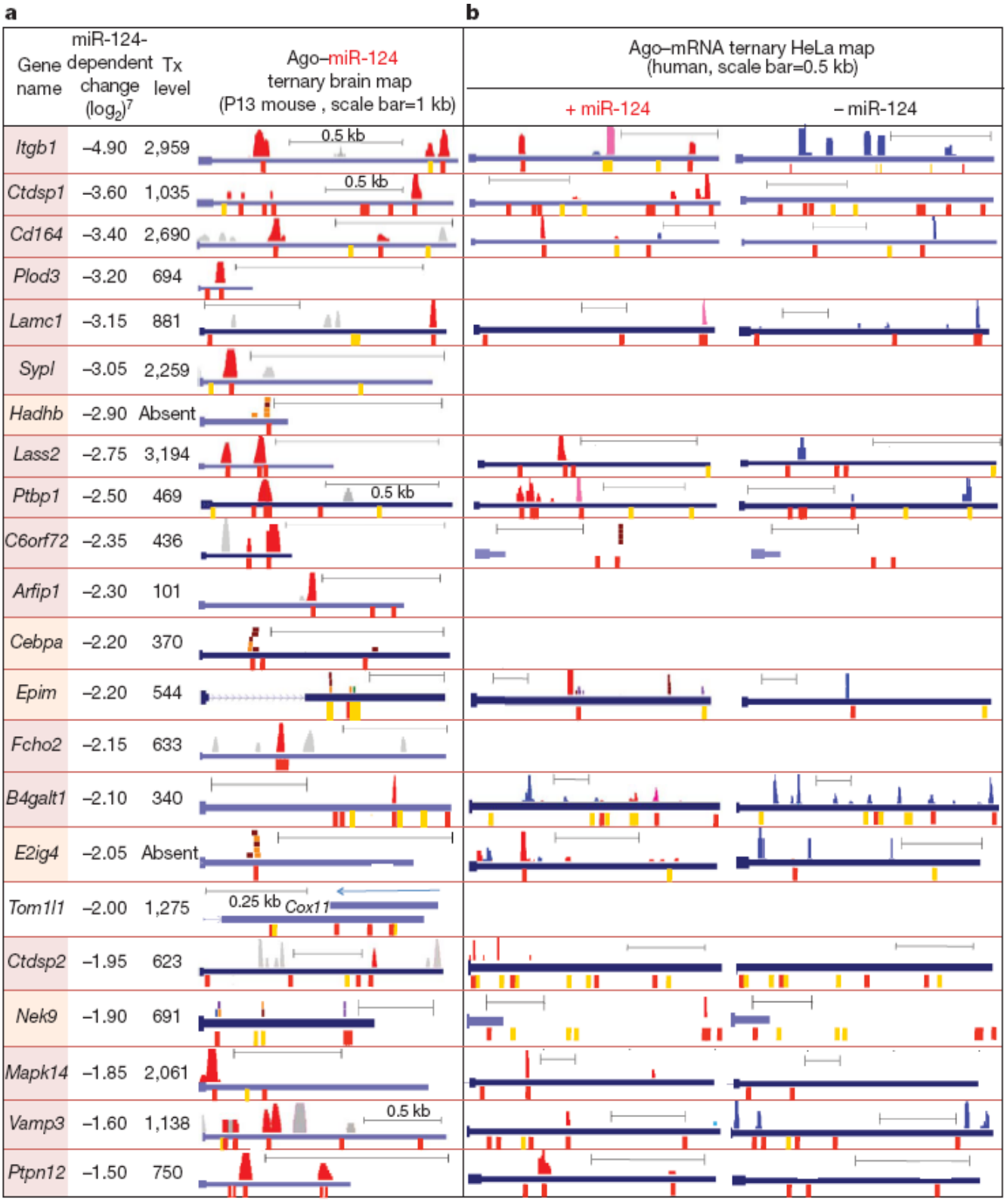

a
b
Gene name
miR-124-dependent change (log_2)[7]
Tx level
Ago–miR-124 ternary brain map (P13 mouse , scale bar=1 kb)
Ago–mRNA ternary HeLa map (human, scale bar=0.5 kb)
+ miR-124
– miR-124
Itgb1 –4.90 2,959
Ctdsp1 –3.60 1,035
Cd164 –3.40 2,690
Plod3 –3.20 694
Lamc1 –3.15 881
Sypl –3.05 2,259
Hadhb –2.90 Absent
Lass2 –2.75 3,194
Ptbp1 –2.50 469
C6orf72 –2.35 436
Arfip1 –2.30 101
Cebpa –2.20 370
Epim –2.20 544
Fcho2 –2.15 633
B4galt1 –2.10 340
E2ig4 –2.05 Absent
Tom1l1 –2.00 1,275
Ctdsp2 –1.95 623
Nek9 –1.90 691
Mapk14 –1.85 2,061
Vamp3 –1.60 1,138
Ptpn12 –1.50 750
0.5 kb
0.25 kb
Cox11

Table 5.4. Meta-analysis of five microarray studies in miR-124 overexpression.

Five previous microarray experiments in miR-124 tranfected cells used for meta analysis.

Table 5.4

GEO	experiments	Cell line	species	measure	Reference
GSE8498	overexpression	neuroblastoma CAD	mouse	microarray	Makeyev EV, Zhang J, Carrasco MA, Maniatis T. The MicroRNA miR-124 promotes neuronal differentiation by triggering brain-specific alternative pre-mRNA splicing. Mol Cell 2007 Aug 3;27(3): 435-48.
GSE6207	overexpression	HepG2	human	microarray	Wang X, Wang X. Systematic identification of microRNA functions by combining target prediction and expression profiling. Nucleic Acids Res 2006;34(5):1646-52
GSE2075	overexpression	Hela	human	microarray	Lim LP, Lau NC, Garrett-Engele P, Grimson A et al. Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. Nature 2005 Feb 17;433(7027):769-73
GSE11968	overexpression	Hela	human	SILAC-MS microarray	Baek D, Villén J, Shin C, Camargo FD et al. The impact of microRNAs on protein output. Nature 2008 Sep 4;455(7209):64-71
GSE11080	overexpression	HEK293T	human	microarray	David G. Hendrickson, Daniel J. Hogan, Daniel Herschlag, James E. Ferrell and Patrick O. Brown, Systematic Identification of mRNAs Recruited to Argonaute 2 by Specific microRNAs and Corresponding Changes in Transcript Abundance PLoS ONE, 2008 May 7;3(5):e2126

5 microarray (meta-analysis) : 1278 genes (more than two exp : $p<0.05$, brain expressed)

CHAPTER VI. ANALYSIS OF MIRNA-MRNA INTERACTIONS: BEYOND THE SEED

Introduction

miRNAs play critical roles in the regulation of gene expression. However, since miRNAs preferentially base pair with only 6-8 nucleotides of mRNA through their seed region (position 1-8), predicting target mRNAs is a major challenge (Bartel, 2009). High-throughput sequencing of RNAs isolated by crosslinking immunoprecipitation (HITS-CLIP) has identified functional protein-RNA interaction sites (Licatalosi et al., 2008). We used HITS-CLIP to covalently crosslink native Argonaute (Ago) protein-RNA complexes in mouse brain, generating two sets of biochemical results—Ago-miRNA CLIP and Ago-mRNA CLIP. By generating this collection, we constrained the sequence space that needed to be analyzed to predict the local regions of mRNAs to which Ago crosslinked and the potential miRNAs bound there, substantially reducing the rate of false-positive predictions of miRNA binding.

Defining Ago-mRNA interaction sites using HITS-CLIP has the potential to inform many different bioinformatic tools developed to predict miRNA target

interactions. The main remaining problem is mining the information from Ago interacting sites and developing bioinformatic methods to pair-up miRNAs with collected targets sites. In the previous work, we have chosen to focus on validating our results by using seed-driven approaches together with defining Ago footprints of mRNA tags and decoded Ago ternary maps showing miRNA-mRNA interactions. However, we recognized that not all Ago binding need to be constrained by seed rules. For example, 27% of Ago-mRNA clusters (BC $\geq$ 2) have no predicted 6-mer seed sites from among the top 20 Ago-miRNAs families. Such "orphan" Ago-mRNA clusters could be bound by miRNAs through other rules of binding.

Initially, seed-driven approaches for miRNA target prediction were based on evolutionary constraints (Lewis et al., 2005) and bioinformatic analysis (Xie et al., 2005). Although the seed match rule was further supported by large-scale experiments (Baek et al., 2008; Lim et al., 2005; Selbach et al., 2008) and greatly helped to characterize the miRNA targets and functions, there still is no evidence that any miRNA seed must be contiguously base-paired for a target site to be functional. Several groups reported that miRNAs can allow interactions with wobble (G:U pairing) or bulge in the seed region. In C. elegans, it was validated by both *in vivo* and *in vitro* assays that lin-41 has such wobble and bulge sites for

let-7 (Vella et al., 2004). In addition, GU wobbles in the seed region do not prevent endogenous lsy-6 from downregulating cog-1 (Didiano and Hobert, 2006). More recently, a similar result was also shown in a mammal (Tay et al., 2008); Four of the five target sites in the coding sequence (CDS) of Nanog, Oct4 and Sox2, which are shown to modulate ESC differentiation, contained wobble and bulge sites in the seed.

Although such bulges and wobbles are reported to be functional (Didiano and Hobert, 2006; Tay et al., 2008; Vella et al., 2004), rules based on these few reports have not been developed to allow miRNA target predicitons. Since such additional rules have the target sites predicted more often by chance, the prediction results generate more false-positives, which are more difficult to be filtered-out, than the seed based approaches. None of the atypical binding sites have been over-represented in the data from the large-scale experiments for miRNA target identifiactions or in the bioinformatic analysis for conserved motif discovery, indicating that their identification and functional study are challenging. Here, we further analyze Ago HITS-CLIP data focusing on the identification of such atypical rules of miRNA target recognition and their function in miRNA mediated gene silencing.

Result

Evaluation of the performance of Ago HITS-CLIP

First, we evaluated performance of our Ago HITS-CLIP analysis which used seed based approaches for predicting miRNA-mRNA interactions. False positive rates and specificity of our Ago-miR-124 ternary maps were assessed by comparing them with the results from meta-analysis of five published microarray experiments, which sought to identify transcripts that were significantly downregulated by miR-124 overexpression in HeLa and other cell lines (Baek et al., 2008; Hendrickson et al., 2008; Lim et al., 2005; Makeyev et al., 2007; Wang, 2006). Focusing on two sets – decreased transcripts enriched in direct targets (703 true set, negative fold change) and increased transcripts enriched in indirect targets (575 false set, positive fold change), the false positive rate was estimated as ~27% and specificity was ~92.5% (Figure 6.1), which achieved an approximately 1.5 fold improvement in the performance of prediction relative to bioinformatic prediction alone (~45% of false-positives and ~67.5% of specificity in the seed based prediction alone; Figure 6.1).

Next, we estimated the specificity of our method by comparing the number of conserved seed matches in the bottom 20 versus top 20 Ago-miRNAs in the 62nt Ago footprint region (Figure 6.2A). The false positive rate was estimated as ~10%

(~3.37 $\log_2$ fold ratio (top 20 vs bottom 20) = ~10 fold increase). In addition, a similar method was applied to the top20 miRNA seed matches in 62nt Ago footprint region (BC $\geq$ 2 ; Figure 6.2B). ~12.5% of false positive rate was estimated by comparing the number of conserved seed matches from the top 20 Ago-miRNA with the expected number by chance (among the P13 brain transcriptome) in the Ago footprint region (~3.02 $\log_2$ fold ratio (observed vs expected) = ~8 fold increase = ~12.5% false positive rate). By calculating false positive rates depending on BC and peak height, we further found that even under our lowest strigency conditions (BC $\geq$ 2) our observed/expected ratios were highly significant (Table 6.1). The least stringent assumption (non conserved 6-mer seeds present in Ago-mRNA clusters with BC $\geq$ 2) gives a maximum false positive rate of ~30%; the best assumption (conserved 8-mers) gives a false positive rate of ~5%, although clearly sensitivity is lower at this threshold. Taken together, these estimates of specificity and false positive rates (~93%, ~5-27%, respectively) in our method indicated that experimental Ago HITS-CLIP data outperformed other approaches.

Although we used seed-driven approaches to validate targets, not all Ago binding need be constrained by these rules. When we estimated false negative rates by comparing the number of Ago-mRNA clusters with no predicted seeds

in the top 20 Ago-miRNAs, relative to the proportion of the top 20 Ago-miRNAs in the brain, 27% of Ago-mRNA clusters (BC $\geq$ 2) have no predicted 6-mer seed sites from among the top 20 Ago-miRNAs families (comprising 88% of all Ago miRNAs). Our seed-driven approach yields an estimated false negative rate of ~15% (orphan clusters (27%) – Expected number of clusters from below top 20 (12%) = ~15%). This compares favorably with similar seed driven estimates of false negatives rates of 50% to 70% made by other investigators (Baek et al., 2008; Mourelatos, 2008; Selbach et al., 2008). However, we also recognized that even though this seed-based approach produced reasonable estimates for our false negative rate, currently unknown miRNA binding rules/sites might affect our false negative rate. Orphan clusters from false-negatives might be bound by miRNAs that follow other rules of binding, such as wobble/bulge matches (Didiano and Hobert, 2006; Tay et al., 2008; Vella et al., 2004) or perfect matches beyond seed (Grimson, 2007).

Bioinformatic evidences of miRNA-mRNA interactions beyond seed

Although we observed a high correlation between the frequency of seeds in mRNA tags and the frequency of miRNAs associated with Ago in the previous analysis (Fig 5.1B and 6.2; R^2 = ~0.61), there are some discrepancies showing erratic shapes in Figure 6.2. To check the possibility that such discrepancies were

caused by counting only the miRNA-mRNA interactions within seed region, we expanded the motif analysis to outside of seed region (Figure 6.3). The normalized frequency of 6mer matches to every position of top 50 miRNAs (Figure 6.3B) in the mouse brain showed that some miRNAs may interacts with other regions. For example, miR-708 (# 7 ranking) has enriched match sequences in position ~8-16 in Ago-mRNA tags instead of seed matches (Figure 6.3C). Hierarchical clustering of the same analysis, expanded to all brain miRNAs, showed several distinct patterns of potential miRNA-mRNA interaction regions and estimated that ~50% of miRNAs may interacts with mRNA through regions other than seed (Figure 6.3A). In addition, we could detect enrichment of 4mers in position between ~13 and ~18 of miR-124 although its fold ratio (~1.2 fold of observed (obs) versus expected (exp) frequency), is less than the seed matches (~1.4 fold of obs/exp; 14 fold of obs/exp in 7mer; Figure 6.4A-B). Such an observation correlates with 3′ supplementary or complementary sites whose importance was shown in previous analyses (Brennecke et al., 2005; Doench and Sharp, 2004; Grimson, 2007)

Focusing on Ago miR-124 clusters and bulge/wobble sites, we searched in a completely unbiased manner for all possible wobble and bulge variants of miR124 seeds in our Ago-mRNA footprints (Table 6.1). It is apparent from this

analysis that there are some potential new rules from the Ago HITS-CLIP data that are not intuitively obvious. This includes some but not all wobble positions; for example, wobble (4,5; GTGTTT) is slightly over-represented in the data, while wobble (4,5; GTTTTA) is very under-represented. Interestingly, G bulges at position 5-6 or 6-7 are highly enriched and A bulges at position 5-6 are moderately enriched while C bulges at 5-6 or 6-7 are not. We next examined the position and frequency of such conserved bulge sequences relative to the peaks of Ago-mRNA clusters and observed that G and A bulges showed many tags in a tight peak region, while others did not (Figure 6.5). However, a limitation on interpretation of the data is that miR-124 bulge sequences could match to a number of other miRNAs.

Identification of a bulge site in orphan clusters evident by miR-124 transfection

To examine whether miR-124 bulges are functionally important, we focused further analysis on *de novo* Ago-miR124 clusters, identified as clusters present in miR-124 tranfected HeLa but absent in the control HeLa cells. Linear regression analysis of the *de novo* miR-124 clusters identified G bulges together with seed motifs (Figure 6.6A). Motif analysis in orphan clusters (921 clusters with no miR-124 seeds (BC=4)) also discovered the same set of miR-124 G bulge sites

(Figure 6.6.C). However, A bulges at position 5-6 of miR-124, which initially appeared to be enriched in the brain set, were not identified in *de novo* miR-124 clusters. In fact, A bulges match a number of other miRNAs, implicating that the initial observation was an unavoidable artifact caused by heterogeneity of brain miRNAs. Focusing on the analysis of a robust subset from *de novo* miR-124 clusters, present in both mouse brain (BC$\geq$2) and miR124 transfected HeLa (BC=4), we found that among de novo miR-124 clusters, ~50% have classical seed sites, ~15% G bulge sites, and ~35% remain "orphan" clusters (Figure 6.6B).

Experimental evidence of the functional relevance of the bulge sites was achieved by luciferase reporter assay with various mutants of the G-bulge site in erythrocyte membrane protein band 4.1 (EPB41); only the G-bulge site showed statistically significant mir-124 dependant repression (~1.3 fold repression relative to the control, $P<0.05$, t-test) comparable to the seed site (~1.4 fold repression; Figure 6.7). Taken together, our data indicates that miRNAs may recognize a specific type of bulge, preferentially located in position 5-6 and these site are likely to be functional as validated bulge site in EPB41.

Discussion

Here we used Ago HITS-CLIP data to identify the rules of miRNA target recognition. We evaluated the performance of the Ago HITS-CLIP method in determining miRNA target sites in brain using seed based approaches and found that large number (30~50%) of Ago-mRNA clusters are orphans, which contain no seed matches to miRNAs associated with Ago, implicating that large numbers of target sites may be bound by miRNAs without seed pairing. In further bioinformatic analysis, we could observe enrichment of bulge matches to miR-124 seeds, perfect seed matches and some matches to sequences outside of the seed regions. Focusing on the analysis of de novo miR-124 clusters identified by Ago HITS-CLIP in miR-124 transfected HeLa cell, we confirmed that ~15% of miR-124 target sites contain G bulges between position 5 and 6 (5-6) in the seed and further validated that the G-bulge target site in EPB41 is functional by luciferase assay. Furthermore, we analyzed that 35% of de novo miR-124 clusters are still remain as orphans without any significant free energy change in hybridization (Figure 6.8A) although the free energies from seed sites and G bulges are significant (Figure 6.8A; relative to shuffled sequences). In addition, other miRNAs such as let-7 also has similar number of such bulge sites (5-6) with perfect seed sites in the brain (Figure 6.8B). In fact, a free energy calculation for any nucleotide bulge between position 5 and 6 gives the same value. However,

only a G bulge sequence can also make another 5 consecutive perfect match starting from position 2. The free energy calculation of this pair is lower than the free energy from the bulge pair. Therfore, preference of G position in the bulge may be caused by free energy contribution from the 5 consecutive perfect match, another optimal pairing from G bulge sequences. Similarily, A bulge in let-7 can make the same 5 consecutive perfect matches and it is the most enriched bulge in let-7 (14%). Taken together, we demonstrate that miRNAs can recognize functional target sites beyond seed and preferentially bind to certain type of bulge (5-6) sites. Understanding of such binding rules using Ago HITS-CLIP is increasingly important for studying miRNA functions and as well as designing siRNAs for basic research and therapeutics.

Contributions

All bioinformatic analysis, construction of luciferase reporter and luciferase assay were done by Sung Wook Chi.

Figure 6.1. Performace of Ago HITS-CLIP estimated by Meta-analysis of five microarray studies in miR-124 overexpression.

We compiled 1278 brain expressed transcripts from a meta-analysis of five microarray experiments which identified transcripts significantly downregulated by miR-124 overexpression in cell lines (see Supplementary Methods) The performance of Ago-miR-124 ternary cluster predictions was assessed by comparing it with miR-124 seed prediction in two sets – decreased transcripts enriched in direct targets (703 true set, negative fold change) and increased transcripts enriched in indirect targets (575 false set, positive fold change). The true positive rate or specificity was calculated as indicated in black box insets (Ago miR-124 ternary cluster prediction; 73% true positive rate and 92.5% specificity, miR-124 seed prediction; 55% true positive rate and 67.5% specificity).

Figure 6.1

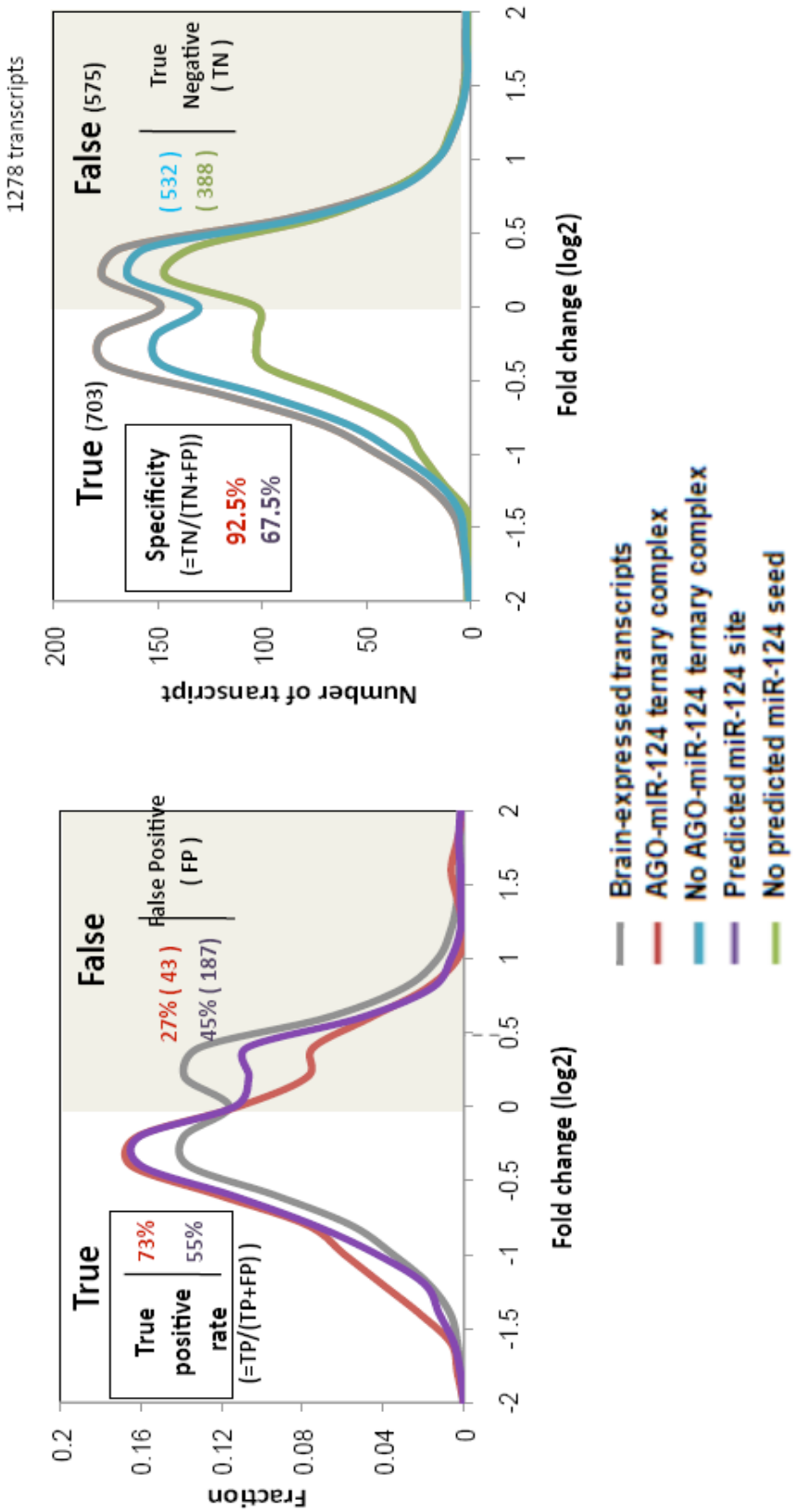
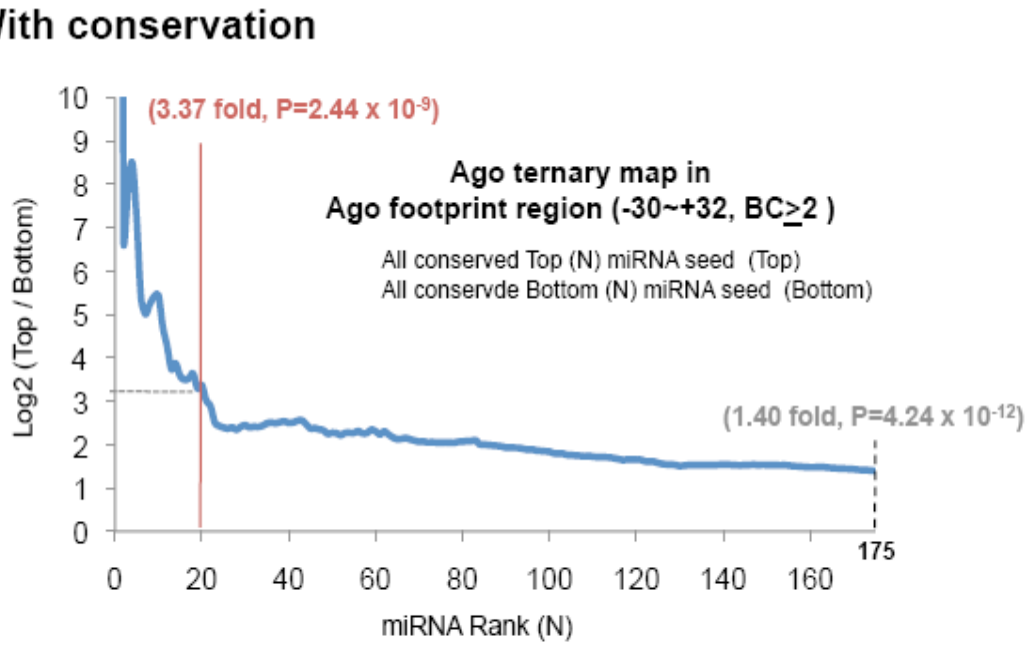

1278 transcripts
Number of transcript
200
150
100
50
0
False (575)
True Negative (TN)
(532)
(388)
True (703)
Specificity (=TN/(TN+FP))
92.5%
67.5%
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2
Fold change (log2)
Fraction
0.2
0.16
0.12
0.08
0.04
0
False
False Positive (FP)
27% (43)
45% (187)
True
True positive rate (=TP/(TP+FP))
73%
55%
Fold change (log2)
Brain-expressed transcripts
AGO-miR-124 ternary complex
No AGO-miR-124 ternary complex
Predicted miR-124 site
No predicted miR-124 seed

Figure 6.2. Statistical analysis of false positive rates for Ago HITS-CLIP analysis

A. Cumulative frequency of top (N) Ago-miRNA 6-mer seed sequences, compared to the frequency of the bottom (N) Ago-miRNA 6-mer seeds, in Ago footprint regions. B. Cumulative frequency as in (A), but comparing the observed vs. expected seed frequencies for the top 30 Ago-miRNA in Ago footprint regions. Both datasets suggest that the correlation between the top 20 miRNAs and Ago-mRNAs is robust as a group, and that correlations likely extend out significantly further. In addition, the shape of the cumulative curve (e.g. (B)) suggests variability among individual miRNA ternary maps. We used conserved seed sequences in calculation of cumulative frequencies.

Figure 6.2

A


With conservation
(3.37 fold, P=2.44 x 10-9)
Ago ternary map in
Ago footprint region (-30~+32, BC≥2)
All conserved Top (N) miRNA seed (Top)
All conservde Bottom (N) miRNA seed (Bottom)
(1.40 fold, P=4.24 x 10-12)
175
Log2 (Top / Bottom)
miRNA Rank (N)

B

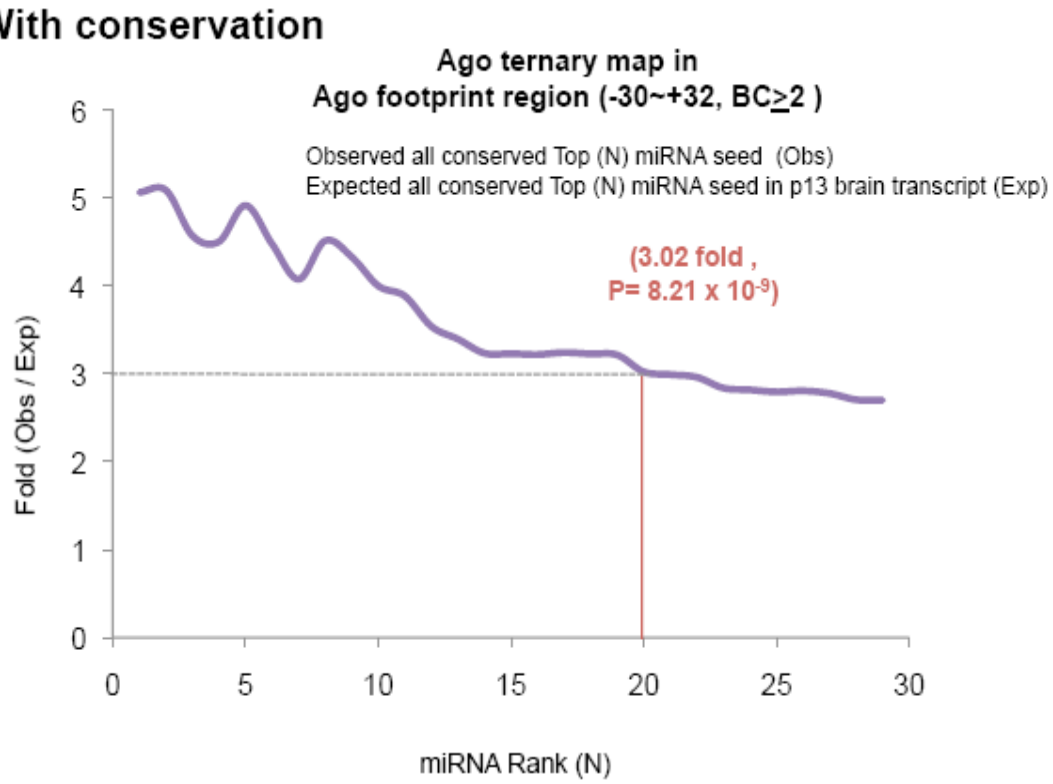

With conservation
Ago ternary map in
Ago footprint region (-30~+32, BC≥2)
Observed all conserved Top (N) miRNA seed (Obs)
Expected all conserved Top (N) miRNA seed in p13 brain transcript (Exp)
(3.02 fold ,
P= 8.21 x 10-9)
Fold (Obs / Exp)
miRNA Rank (N)

Figure 6.3. Analysis of miRNA matches from different positions in Ago mRNA tags

A.The frequencies of conserved 6mer matches from different positions of brain expressed miRNAs (confirmed to be expressed by Ago miRNA CLIP in P13 mouse brain) in the sequences of Ago mRNA clusters (BC≥2) are calculated, further normalized by mean value of all such frequencies in each miRNAs (log2 ratio centered by mean value; B), used for hierarchical clustering of the miRNAs and visualized as a heatmap and tree (A). miRNAs showing enrichment of their motifs in seed region (position 1-8) are indicated as group (a blue lined group or a pink colored tree). C.The same analysis as (A) but only for top 50 miRNAs (miRNA frequency ranking from the Ago CLIP) without performing hierarchical clustering

Figure 6.3

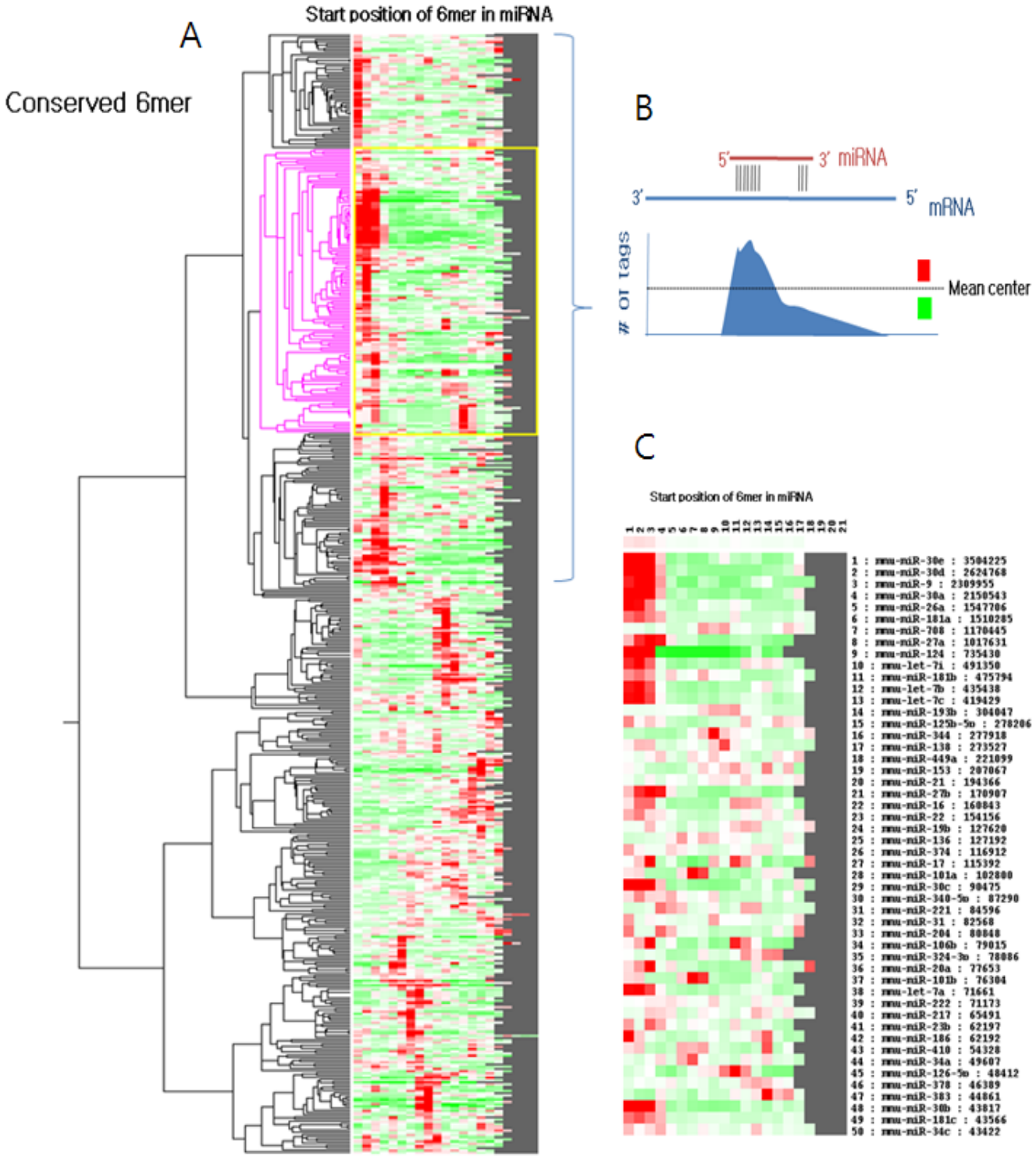

Start position of 6mer in miRNA
A
Conserved 6mer
B
5′
3′ miRNA
3′
5′ mRNA
# of tags
Mean center
C
Start position of 6mer in miRNA
1 : mmu-miR-30e : 3504225
2 : mmu-miR-30d : 2624768
3 : mmu-miR-9 : 2309955
4 : mmu-miR-30a : 2150543
5 : mmu-miR-26a : 1547706
6 : mmu-miR-181a : 1510285
7 : mmu-miR-708 : 1170445
8 : mmu-miR-27a : 1017631
9 : mmu-miR-124 : 735430
10 : mmu-let-7i : 491350
11 : mmu-miR-181b : 475794
12 : mmu-let-7b : 435438
13 : mmu-let-7c : 419429
14 : mmu-miR-193b : 304047
15 : mmu-miR-125b-5p : 278206
16 : mmu-miR-344 : 277918
17 : mmu-miR-138 : 273527
18 : mmu-miR-449a : 221099
19 : mmu-miR-153 : 207067
20 : mmu-miR-21 : 194366
21 : mmu-miR-27b : 170907
22 : mmu-miR-16 : 160843
23 : mmu-miR-22 : 154156
24 : mmu-miR-19b : 127620
25 : mmu-miR-136 : 127192
26 : mmu-miR-374 : 116912
27 : mmu-miR-17 : 115392
28 : mmu-miR-101a : 102800
29 : mmu-miR-30c : 90475
30 : mmu-miR-340-5p : 87290
31 : mmu-miR-221 : 84596
32 : mmu-miR-31 : 82568
33 : mmu-miR-204 : 80848
34 : mmu-miR-106b : 79015
35 : mmu-miR-324-3p : 78086
36 : mmu-miR-20a : 77653
37 : mmu-miR-101b : 76304
38 : mmu-let-7a : 71661
39 : mmu-miR-222 : 71173
40 : mmu-miR-217 : 65491
41 : mmu-miR-23b : 62197
42 : mmu-miR-186 : 62192
43 : mmu-miR-410 : 54328
44 : mmu-miR-34a : 49607
45 : mmu-miR-126-5p : 48412
46 : mmu-miR-378 : 46389
47 : mmu-miR-383 : 44861
48 : mmu-miR-30b : 43817
49 : mmu-miR-181c : 43566
50 : mmu-miR-34c : 43422

Figure 6.4. Motif analysis of Ago mRNA clusters

A. The frequencies of 7mer matches from the different positions of miR-124 (observed frequency; obs) are divided by expected frequencies (estimated frequencies in p13 brain transcriptome; exp) and plotted depending on the different position of miR-124. B. The same analysis performed in (A) with 4mer matches. Significantly enriched motifs (>1.2 fold) are highlighted as blue for seed region and red for other region. A top enriched motif from MEME analysis in Ago mRNA clusters (BC≥5; C) and orphan clusters (D). A bulge sequence is marked by red dot.

Figure 6.4

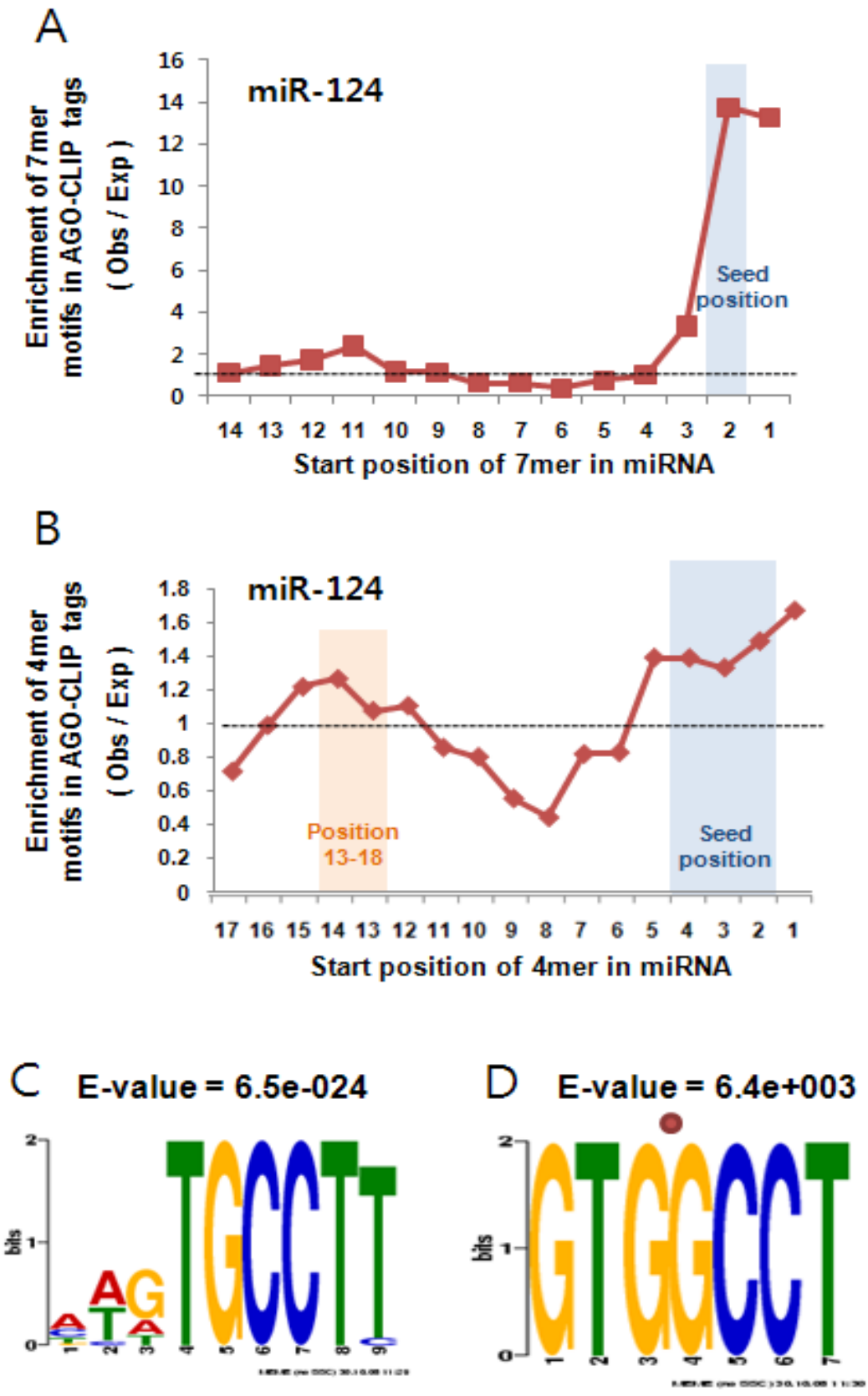

A
miR-124
Enrichment of 7mer motifs in AGO-CLIP tags (Obs / Exp)
16
14
12
10
8
6
4
2
0
Seed position
14 13 12 11 10 9 8 7 6 5 4 3 2 1
Start position of 7mer in miRNA
B
miR-124
Enrichment of 4mer motifs in AGO-CLIP tags (Obs / Exp)
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
Position 13-18
Seed position
17 16 15 14 13 12 11 10 9 8 7 6 5 4 3 2 1
Start position of 4mer in miRNA
C
E-value = 6.5e-024
bits
D
E-value = 6.4e+003
bits

Figure 6.5. Distribution of mRNA tags correlates with some bulge matches to miRNAs from Ago CLIP.

The position of conserved 7mer bulge matches to miR-124 (A-B with different sequence in bulge position; indicated as lower cases, bottom panel; each is represented by a different color) were plotted relative to peak position of all Ago-mRNA clusters (BC ≥ 2). Top panel; distribution of mir-124 seed matches (plotted relative to cluster peak, normalized to number of clusters; blue graph); pale color indicates standard deviation.

Figure 6.5

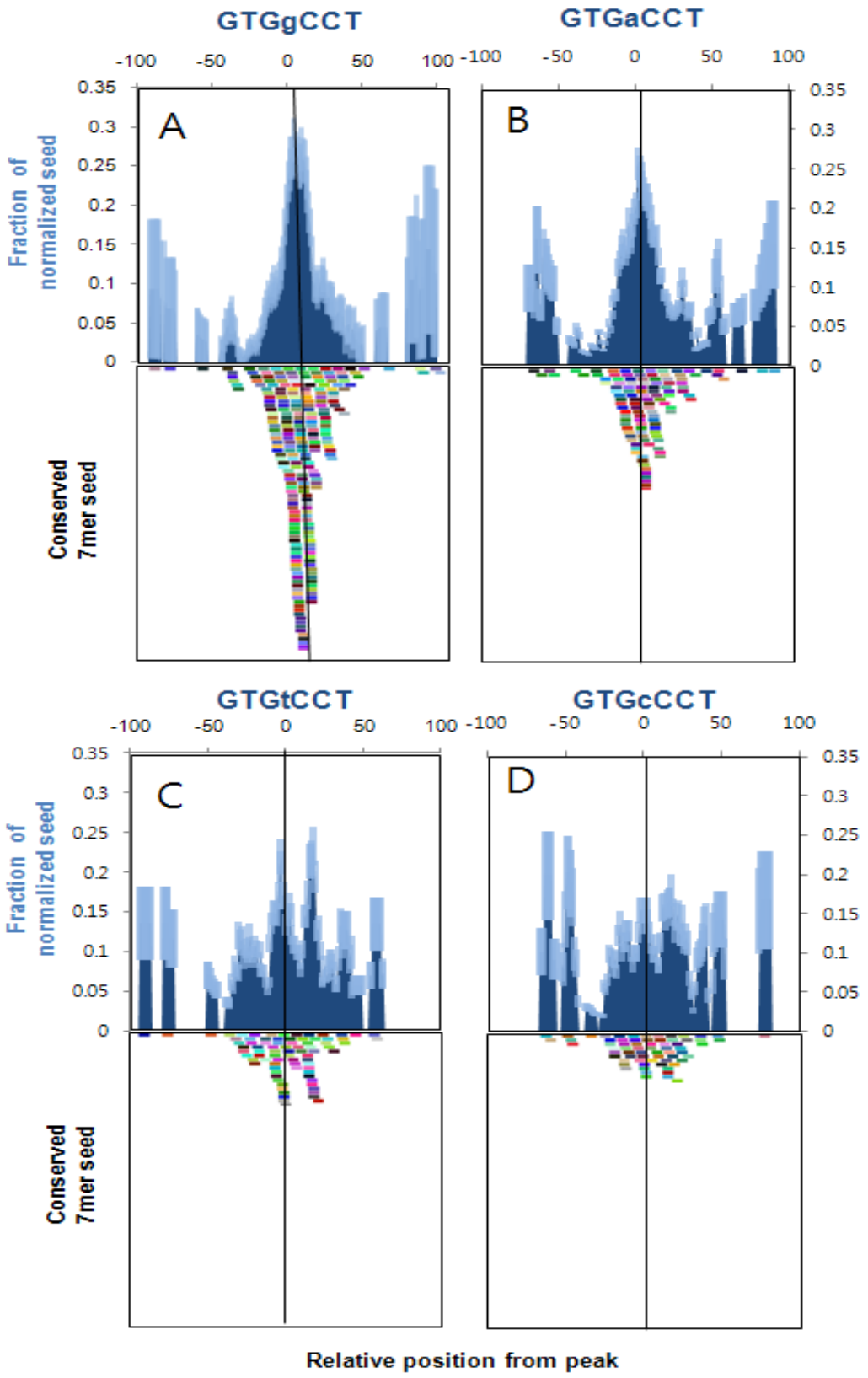

GTGgCCT
GTGaCCT
GTGtCCT
GTGcCCT
A
B
C
D
Fraction of normalized seed
Conserved 7mer seed
-100
-50
0
50
100
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
Relative position from peak

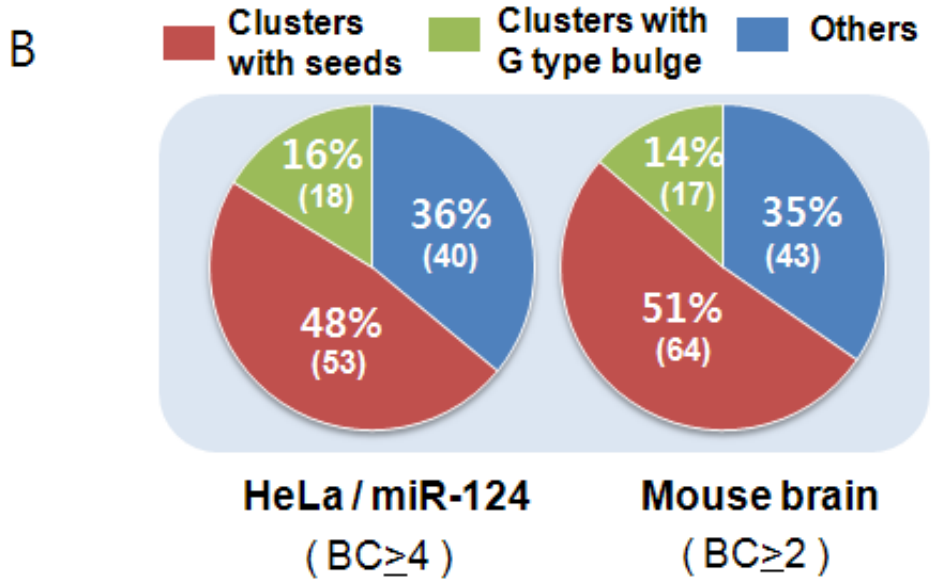
Figure 6.6. Identification of a bulge match to miR-124 in de novo miR-124 clusters from Ago CLIP.

A. We correlated all possible 6 to 8-mer motifs present in de novo Ago-mR124 clusters (BC=4, identified from Ago CLIP performed in miR-124 transfected versus control miRNA transfected HeLa) with cluster height (log2(maximum tag number) per cluster), using the MatrixREDUCE linear regression model: (http://bussemaker.bio.columbia.edu/software/MatrixREDUCE/). The most enriched motifs selected by default parameters (except –max_gap=0) are shown. B. Distribution of conserved robust Ago-miR124-clusters, clusters present in both mouse brain (BC≥2; right panel) and miR124 transfected HeLa (BC=4; left panel) is analyzed by three categories, clusters with miR124 seed matches, bulge matches (a bulge between position 5-6) and others.C. Motif discovery analysis using MEME program in orphan clusters (with no miR-124 seeds) identified the G bulge (G bulge between position 5-6).

Figure 6.6

A

Motif	Regression coefficient	p-value	miRNA	Matched seed position
TGCCTT	1.18659	3.96E-212	miR-124	2-7
GTGgCCT	1.39413	5.69E-150	miR-124	2-4,6-8
GTGCCT	0.602258	2.95E-54	miR-124	3-8
TGgCCT	0.722041	5.00E-38	miR-124	2-4,6-7



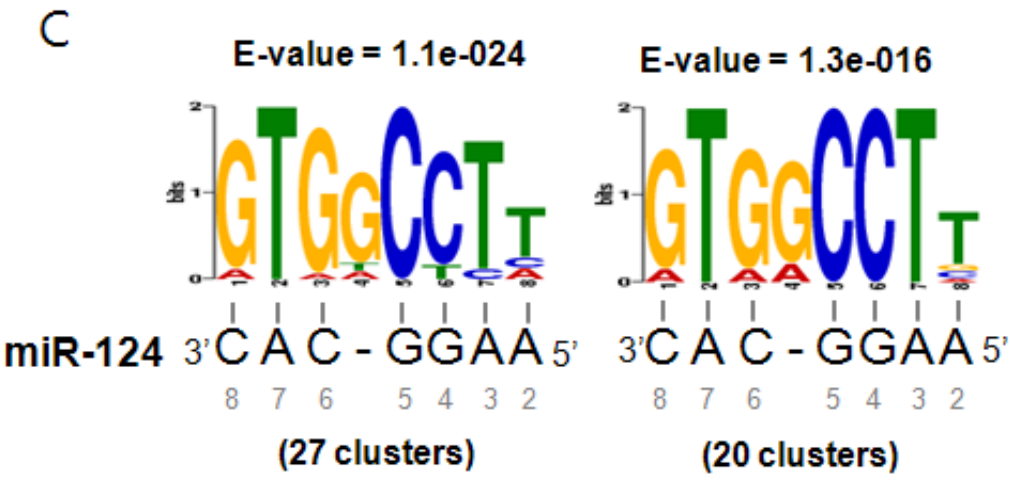


Figure 6.7. Validation of miR-124 bulge sites in EPB41 by luciferase reporter assays.

A.Conserved de novo Ago-miR-124 clusters (overlapping with Ago-mRNA clusters in P13 mouse brain) in 3'UTR of erythrocyte membrane protein band 4.1 (EPB41). The locations of miR-124 seed matches are indicated as red lines. Black bar indicated the region (104nt) containing a G bulge match, inserted into the luciferase reporter. Bottom panel shows wild type and mutant sequences used for the luciferase reporter assay. B. Normalized results (renilla reporter lucifease activity is normalized by firefily luciferase activity) from the luciferase reporter assay with the G-bulge sequence, and other mutants (perfect seed match and T bulge, A bulge and C bulge matches). The assays are performed in miR-124 tranfected versus control miRNA transfected HeLa cells (n=3; three replicates of experiments with different HeLa cell batches). Statistically significant results are indicated as asterics ($P<0.05$; t-test).

Figure 6.7

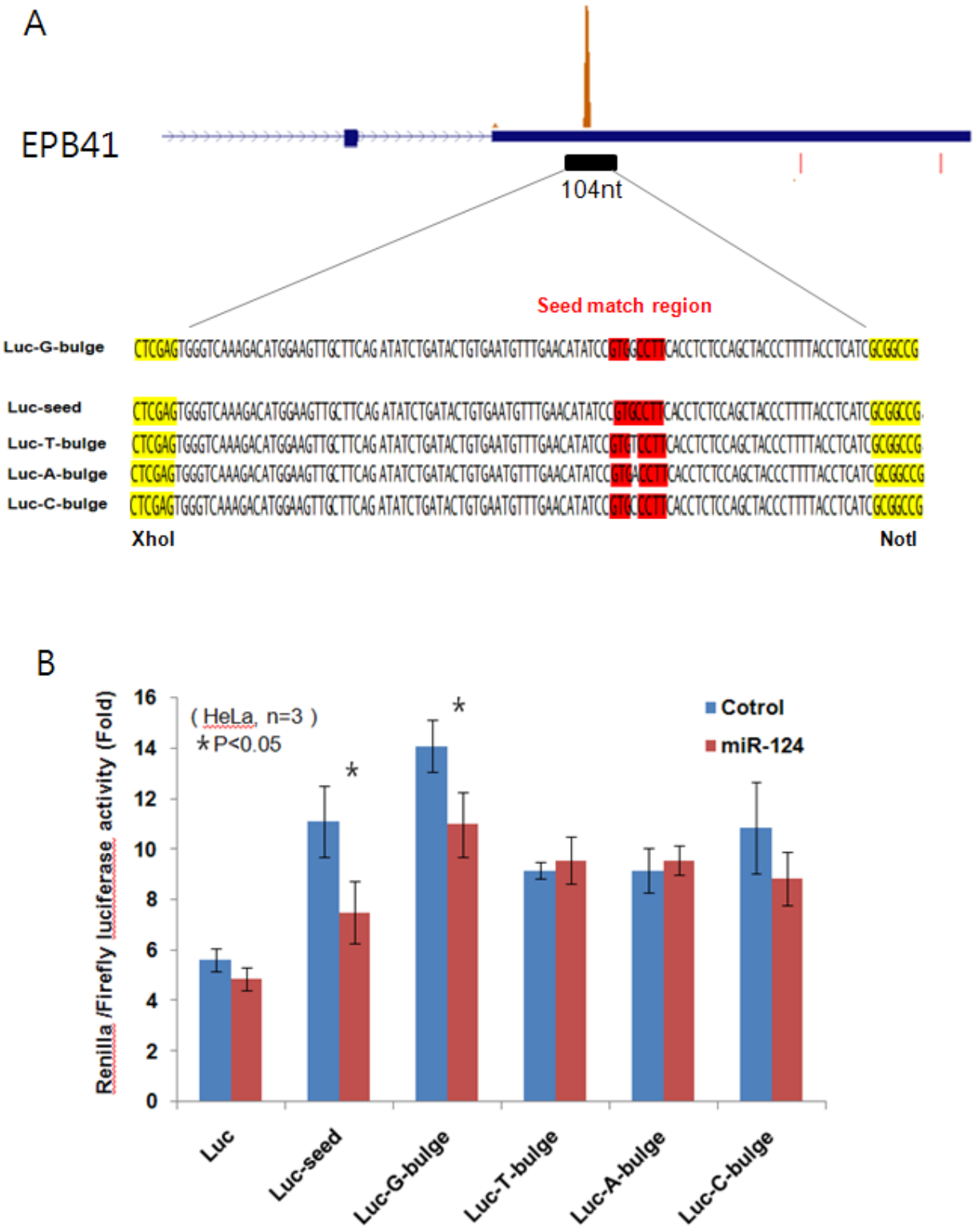
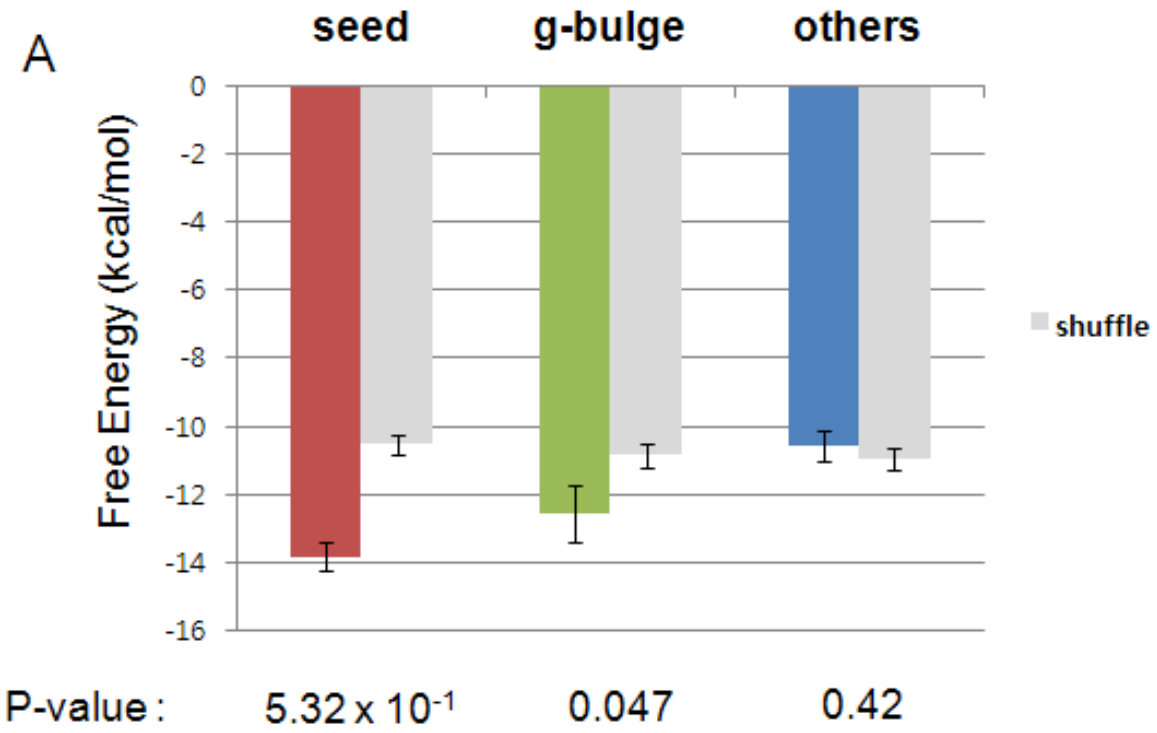

A
EPB41
104nt
Seed match region
Luc-G-bulge
CTCGAGTGGGTCAAAGACATGGAAGTTGCTTCAG ATATCTGATACTGTGAATGTTTGAACATATCCGTGCCTTCACCTCTCCAGCTACCCTTTTACCTCATCGCGGCCG
Luc-seed
CTCGAGTGGGTCAAAGACATGGAAGTTGCTTCAG ATATCTGATACTGTGAATGTTTGAACATATCCGTGCCTTCACCTCTCCAGCTACCCTTTTACCTCATCGCGGCCG
Luc-T-bulge
CTCGAGTGGGTCAAAGACATGGAAGTTGCTTCAG ATATCTGATACTGTGAATGTTTGAACATATCCGTGTCCTTCACCTCTCCAGCTACCCTTTTACCTCATCGCGGCCG
Luc-A-bulge
CTCGAGTGGGTCAAAGACATGGAAGTTGCTTCAG ATATCTGATACTGTGAATGTTTGAACATATCCGTG-CCTTCACCTCTCCAGCTACCCTTTTACCTCATCGCGGCCG
Luc-C-bulge
CTCGAGTGGGTCAAAGACATGGAAGTTGCTTCAG ATATCTGATACTGTGAATGTTTGAACATATCCGTGCCCTTCACCTCTCCAGCTACCCTTTTACCTCATCGCGGCCG
XhoI
NotI
B
Renilla /Firefly luciferase activity (Fold)
16
14
12
10
8
6
4
2
0
(HeLa, n=3)
*P<0.05
*
*
Cotrol
miR-124
Luc
Luc-seed
Luc-G-bulge
Luc-T-bulge
Luc-A-bulge
Luc-C-bulge

Figure 6.8. Free energy calculation of de novo Ago-miR-124 sites and analysis of buge matches to let-7 in Ago-mRNA clusters.

A.Calculations of free energies from three types of clusters (Figure 6.6B) annealing to miR-124. The free energies are calculated by RNAduplex (Hofacker, 2003)(average values are indicated as the same color in Figure 6.6B) and compared with the shuffled sequences (100 times; average values are indicated as gray bar). P-values from t-test are indicated. B.Numbers and fractions of the different bulge matches (a bulge between position 5 and 6) and the seed match to let-7 in Ago mRNA clusters (P13 mouse) are shown.

Figure 6.8


A
seed
g-bulge
others
Free Energy (kcal/mol)
0
-2
-4
-6
-8
-10
-12
-14
-16
shuffle
P-value :
5.32 x 10-1
0.047
0.42

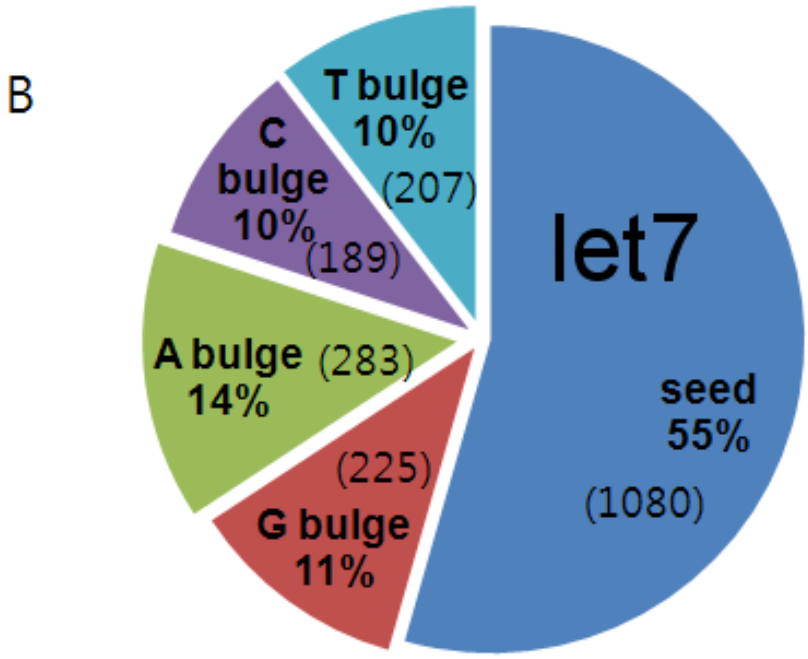

B
T bulge
10%
(207)
C
bulge
10%
(189)
A bulge
14%
(283)
G bulge
11%
(225)
let7
seed
55%
(1080)

Table 6.1. Biolologic complexity and peak height analysis of Ago mRNA clusters.

Effect of varying BC and peak height on Ago-HITS-CLIP map predictions. We evaluated enrichment of seeds (estimated by calculating Ago-miRNA seed sequences (observed vs. expected) in Ago-mRNA footprints from P13 mouse brain transcripts) for the #8 ranking (miR-124) and the #18 ranking (miR-19) miRNAs using the indicated thresholds in BC and peak height. We found that even under our lowest stringency conditions (BC > 2 or peak height > 2), our observed/expected ratios were highly significant. The least stringent assumption (6-mer seeds present in Ago-mRNA clusters with BC > 2) gives a maximum false positive rate of ~30%; the best assumption (conserved 8-mers) gives a false positive rate of ~5%, although clearly sensitivity is lower at this threshold.

Table 6.1

miR-124 6mers

BC	Obs/Exp	p-value	True positive rate(%)	Obs	Exp
>=2	3.2649573	0.00 x 100	76.5439348	2292	702.3587
>=3	3.7799929	0.00 x 100	79.07946684	2032	537.5671
>=4	4.071868	0.00 x 100	80.28339873	1180	289.7933
=5	5.0520888	0.00 x 100	83.47677917	453	89.66588

miR-124 6mers

peak	Obs/Exp	p-value	True positive Rate(%)	Obs	Exp
>=2	3.2649573	0.00 x 100	76.55310621	2292	702
>=10	5.3136153	0.00 x 100	84.16121417	832	156.5789
>=20	6.4373708	1.32 x 10-290	86.55438834	289	44.8941
>=30	6.6297288	1.08 x 10-124	86.89337425	118	17.79862
>=40	5.5610835	1.4E-42	84.7586149	50	8.991054
>=50	5.0924938	2.68E-15	83.58636006	19	3.730981

miR-19 6mers

BC	Obs/Exp	p-value	True positive rate (%)	Obs	Exp
>=2	1.1714286	1.69E-05	53.94736842	738	630
>=3	1.2227305	9.95E-07	55.01029036	590	482.5266
>=4	1.1302395	0.035681	53.05692112	294	260.1219
=5	1.428835	0.000119	58.82799722	115	80.48515

miR-19 6mers

peak	Obs/Exp	p-value	True positive rate(%)	Obs	Exp
>=2	1.1714286	1.69E-05	53.94736842	738	630
>=10	1.1384087	0.100824	53.23625524	160	140.5471
>=20	1.0174334	0.91188	50.43206914	41	40.29748
>=30	1.0640796	0.797852	51.55225689	17	15.97625
>=40	1.2390843	0.497008	55.33888625	10	8.070476
>=50	0.597198	0.46104	37.39035571	2	3.348973

Table 6.2. Wobbles and bulges matching to miR-124 seeds in Ago mRNA clusters.

Observed vs. expected frequencies for all possible wobble and bulges in miR-124 seed sequences present in Ago-mRNA clusters.

Table 6.2

Type of seed match	Motifs	mir-124 seed	Obs /Exp	p-value	Observed number (obs)	Expected number (exp)	seed matched miRNAs
Perfect match	GCCTTA	6mer (1-6)	5.798	0.00	459	79	miR-124
Perfect match	TGCCTT	6mer (2-7)	4.150	0.00	867	209	miR-124
Perfect match	GTGCCT	6mer (3-8)	3.796	1.15×10^{-283}	629	166	miR-124
Perfect match	TGCCTTA	7mer (1-7)	10.212	0	287	28	miR-124
Perfect match	GTGCCTT	7mer (2-8)	8.993	0	374	42	miR-124
Wobble (4)	GCTTTA	6mer (1-6)	1.853	4.07×10^{-16}	170	92	none
Wobble (4)	TGCTTT	6mer (2-7)	1.186	0.01	241	203	none
Wobble (4)	GTGCTT	6mer (3-8)	1.167	0.06	146	125	none
Wobble (4)	TGCTTTA	7mer (1-7)	1.879	5.54×10^{-7}	61	32	none
Wobble (4)	GTGCTTT	7mer (2-8)	1.369	0.02	52	38	none
Wobble (5)	GTCTTA	6mer (1-6)	1.237	0.05	85	69	miR-208
Wobble (5)	TGTCTT	6mer (2-7)	1.132	0.07	208	184	miR-743
Wobble (5)	GTGTCT	6mer (3-8)	1.165	0.03	192	165	none
Wobble (5)	TGTCTTA	7mer (1-7)	1.670	4.40×10^{-4}	46	28	miR-743
Wobble (5)	GTGTCTT	7mer (2-8)	1.478	2.74×10^{-3}	58	39	miR-743
Wobble (4,5)	GTTTTA	6mer (1-6)	1.032	0.72	130	126	none
Wobble (4,5)	TGTTTT	6mer (2-7)	0.919	0.18	257	280	none
Wobble (4,5)	GTGTTT	6mer (3-8)	1.219	4.07×10^{-4}	209	171	none
Wobble (4,5)	TGTTTTA	7mer (1-7)	1.310	0.03	67	51	none
Wobble (4,5)	GTGTTTT	7mer (2-8)	1.116	0.42	53	47	none
Bulge (A:3-4)	GTGCCAT	6mer (3-8)	1.096	0.59	35	32	miR-183
Bulge (A:3-4)	TGCCATT	6mer (2-7)	1.344	0.03	53	39	miR-183, miR-652, miR-685
Bulge (A:3-4)	GCCATTA	6mer (1-6)	1.567	0.01	29	19	miR-652, miR-685
Bulge (G:3-4)	GTGCCGT	6mer (3-8)	0.849	0.62	9	11	none
Bulge (G:3-4)	TGCCGTT	6mer (2-7)	1.309	0.35	12	9	miR-337-5p
Bulge (G:3-4)	GCCGTTA	6mer (1-6)	1.212	0.67	5	4	miR-337-5p
Bulge (C:3-4), (C:4-5), (C:5-6)	GTGCCCT	6mer (3-8)	1.170	0.26	52	44	none
Bulge (C:3-4), (C:4-5), (C:5-6)	TGCCCTT	6mer (2-7)	1.028	0.85	49	48	none
Bulge (C:3-4), (C:4-5), (C:5-6)	GCCCTTA	6mer (1-6)	1.167	0.50	19	16	none
Bulge (A:4-5)	GTGCACT	6mer (3-8)	1.471	0.02	38	26	miR-152, miR-291b-3p, miR-301, miR-130, miR-148, miR-721
Bulge (A:4-5)	TGCACTT	6mer (2-7)	2.581	2.69×10^{-20}	88	34	miR-152, miR-291b-3p, miR-301, miR-130, miR-148, miR-721, miR-17, miR-292-3p, miR-295, miR-294, miR-467, miR-20, miR-93, miR-105, miR-302, miR-106, miR-290-3p, miR-291a-3p
Bulge (A:4-5)	GCACTTA	6mer (1-6)	1.752	2.64×10^{-3}	28	16	miR-17, miR-292-3p, miR-295, miR-294,miR-467, miR-20, miR-93, miR-105, miR-302, miR-106, miR-290-3p, miR-291a-3p
Bulge (T:4-5)	GTGCTCT	6mer (3-8)	0.921	0.61	38	41	none
Bulge (T:4-5)	TGCTCTT	6mer (2-7)	1.060	0.65	61	58	none
Bulge (T:4-5)	GCTCTTA	6mer (1-6)	0.454	0.02	9	20	miR-335-5p
Bulge (G:4-5)	GTGCGCT	6mer (3-8)	0.139	1.09×10^{-3}	2	14	none
Bulge (G:4-5)	TGCGCTT	6mer (2-7)	0.585	0.12	8	14	none
Bulge (G:4-5)	GCGCTTA	6mer (1-6)	0.940	0.90	4	4	none
Bulge (A:5-6)	GTGACCT	6mer (3-8)	2.165	4.19×10^{-14}	91	42	miR-1193, miR-540-3p, miR-1898
Bulge (A:5-6)	TGACCTT	6mer (2-7)	1.826	2.53×10^{-8}	83	45	miR-540-3p, miR-1193, miR-1898, miR-493
Bulge (A:5-6)	GACCTTA	6mer (1-6)	2.112	3.22×10^{-6}	37	18	miR-493, miR-18
Bulge (T:5-6)	GTGTCCT	6mer (3-8)	1.240	0.09	60	48	miR-1197
Bulge (T:5-6)	TGTCCTT	6mer (2-7)	1.421	2.40×10^{-3}	74	52	miR-1197
Bulge (T:5-6)	GTCCTTA	6mer (1-6)	1.830	3.48×10^{-4}	34	19	none
Bulge (G:5-6), (G6-7)	GTGGCCT	6mer (3-8)	3.454	2.69×10^{-76}	197	57	none
Bulge (G:5-6), (G6-7)	TGGCCTT	6mer (2-7)	3.297	2.69×10^{-72}	203	62	none
Bulge (C:5-6)	GTGCCCT	6mer (3-8)	1.170	0.26	52	44	none
Bulge (C:5-6)	TGCCCTT	6mer (2-7)	1.028	0.85	49	48	none
Bulge (C:5-6)	GCCCTTA	6mer (1-6)	1.167	0.50	19	16	none
Bulge (A:6-7)	GTAGCCT	6mer (3-8)	1.899	7.30×10^{-5}	37	19	miR-690
Bulge (A:6-7)	TAGCCTT	6mer (2-7)	2.281	6.62×10^{-10}	53	23	miR-690
Bulge (T:6-7), (T:7-8)	GTTGCCT	6mer (3-8)	2.779	2.39×10^{-19}	71	26	miR-31
Bulge (C:6-7)	GTCGCCT	6mer (3-8)	0.687	0.26	9	13	none
Bulge (C:6-7)	TCGCCTT	6mer (2-7)	1.472	0.07	21	14	none

CHAPTER VII. GENERAL DISCUSSION

The value of HITS-CLIP

When the human genome was decoded, the lower than expected number of genes prompted interest in RNA complexity generated by RNA regulation such as alternative splicing, polyadenylation, and post-transcriptional control of gene expression. Such RNA regulation is often controlled by RNA-binding proteins (RNABPs) or non-coding RNAs including microRNAs (miRNAs), but their comprehensive study has lagged behind studies of DNA regulators because of lack of a precise genome-wide method for detecting RNA-protein interactions. To overcome such limitations, we have developed an unbiased, genome-wide method to characterize RNA–protein interactions in living tissues called HITS-CLIP (high-throughput sequencing of RNAs isolated by crosslinking immunoprecipitation). We demonstrated its potential by applying it to mouse brains and characterized the binding sites of the neuronal alternative splicing regulator, Nova, and decoded the target sites of miRNAs through the identification of Argonaute (Ago) associated RNAs.

Transcriptome-wide maps of functional RNABP-RNA interactions generated by HITS-CLIP yield new insights into rules by which RNA complexity is regulated. Our Nova HITS-CLIP analysis has generated compelling evidence that the position of RNABP-RNA interactions within primary transcripts dictates the functional outcome of alternative pre-mRNA processing events (Licatalosi et al., 2008). HITS-CLIP showed that binding of RNABPs within an alternative exon or the flanking upstream intronic sequence is generally associated with exon skipping, whereas binding of RNABPs to the downstream intronic sequence is generally associated with exonic splicing. These studies were presaged by bioinformatic analyses of Nova regulated RNAs (Ule et al., 2006), but such studies were unable to definitively distinguish direct from indirect Nova targets, one of the powerful advantages of HITS-CLIP.

The unbiased nature of the functional RNA-protein interaction maps can also allow completely unexpected discoveries. Based on the identification of Nova binding in miRNA-coding transcripts, we showed evidence that Nova may have an additional function in regulating processing of miRNAs. Nova HITS-CLIP analysis also led to the recognition of a role for Nova in regulating alternative polyadenylation, derived from the observation of Nova binding near poly(A) sites (Licatalosi et al., 2008). Furthermore, HITS-CLIP analysis was extended to

the study of ternary interactions between an RNABP (Ago), RNA and miRNAs, generating a genome-wide map of microRNA binding sites and yielding new rules of RNA regulations. Such findings show the power of HITS-CLIP method in yielding new insights into additional function of RNABPs and miRNAs regulating RNA complexity.

Regulation of miRNA processing by Nova

Transcripts encoding miRNAs called primary miRNAs (pri-miRNAs), are known to be first processed into precursor miRNAs (pre-miRNAs) by the multipotrein complexes called microprocessor, consists of Drosha and DGCR8, then into mature miRNAs (Kim et al., 2009b). However, the regulatory mechanisms of miRNA processing are not well known. There are a few reports that RNABPs have roles in the regulation of miRNA processing. It was reported that hnRNP A1 binds to pri-miR-18a and facilitates its processing (Guil and Caceres, 2007) and that KH-type splicing regulatory protein (KSRP) is a component of both Drosha and Dicer complexes, regulating the processing of subset of miRNAs (Trabucchi et al., 2009). In our study using Nova HITS-CLIP, we also showed that Nova indirectly upregulates miR-770 by inducing an alternative splicing variant containing miR-770 and also directly activate processing step of pri-miR-380

These findings raise some interesting questions and potential future directions for research. The first question is about the mechanism by which Nova activates Drosha in processing pri-miR-380. We can generate the hypothesis that miR-380 may require Nova to form a structure allowing recognition and processing by Drosha and DGCR8 complex (also known as microprocessor). In fact, it is known that microprocessor needs a specific structure, ~33 basepairs (bp) of pri-miRNA stem structure with flanking single stand RNA (ssRNA) located on ~11bp distance from the stem-ssRNA junction (Han et al, Cell, 2006). By recruiting Nova, pri-miR-380 may form this structure more easily and then be recognized and processed by the microprocessor. It has been shown that Nova can dimerize by the interaction through its KH domains (Lewis et al., Cell, 2000), suggesting a potential action whereby Nova may bind at two different locations on RNA and dimerize, and in this way facilitate a formation of an RNA loop of pri-miRNA, For example, two putative YCAY elements (CCAT, CCAC) are located at the each of strands forming stem structures of pre-miR-380. A further idea is the potential existence of other, as yet unidentified pri-miRNA structures, which might form and be processed only upon the binding of Nova. Furthermore, Nova might recruit other proteins which might have RNA chaperone activity for pri-miRNA structure. Further experiments to test this hypothesis might consist of determining direct interactions of pri-miR-380 with Nova, secondary structure of

the pri-miR-380 hairpin and efficiency of processing by microprocessor in the presence versus absence of Nova.

Evaluation of Ago HITS-CLIP analysis

Gene silencing induced by miRNAs in the regulation of gene expression is important for many biological function and pathogenesis of diseases. Although various efforts have been tried to determine miRNA targets based on computational approaches (Rajewsky, 2006) and were recently combined with large-scale experiments (Baek et al., 2008; Lim et al., 2005; Selbach et al., 2008), they are hampered by high false positive rates (~40-66%; detail in following discussion section) because the target site sequences are short and occur often in the genome. This problem has been overcome by our study using Ago HITS-CLIP method. We redesigned the HITS-CLIP method focusing on miRNA and mRNA interactions with Ago, a ubiquitous core component with endonuclease activity in RNA-induced silencing complex (RISC), and used this method to decode a precise map showing high resolution of miRNA binding sites in brain mRNA transcripts. The importance of this method is that it is generally applicable and it can be used for understanding the functional role of miRNAs and as well as determining target sites for RNA interference therapy on clinically relevant mRNAs.

The miRNA-mRNA interaction maps generated by Ago HITS-CLIP have the potential to advance our understanding of miRNA regulation. In the current setting, Ago HITS-CLIP generates two collections of tags, one sampling the set of miRNAs that are expressed and actively incorporated into Ago protein complex and the other sampling the mRNAs at the sites of interaction with one or more of these active miRNAs. By generating this collection, it substantially reduces set of miRNAs and search space for their binding sites by defining a ~45-60-nucleotide Ago footprint based on the distribution of mRNA tags (Figure 5.1). Although Ago HITS-CLIP analysis used the same "seed driven" approaches as other bioinformatics prediction, it outperformed bioinformatic predictions alone or combined with large-scale experiments (~93% specificity, ~13-27% false positive rates and ~15-25% false negative rates ; estimated by miR-124 datasets). The following discusses more details about the performance of Ago HITS-CLIP for miRNA-mRNA interaction maps (also called "Ago ternary maps") focusing on its specificity and sensitivity.

-Specificity of Ago ternary maps

To compare the specificity of Ago ternary maps from HITS-CLIP with other approaches, false positive rates from other seed based methods combining proteomic approaches were first estimated (Baek et al., 2008; Selbach et al., 2008).

They were estimated as ~40% (Mourelatos, 2008) or ~66% (based on their reports that two-thirds of the predicted targets appeared to be nonresponsive to miR-223 loss in neutrophils (Baek et al., 2008)). This higher number of false positives is also in agreement with findings from two other approaches using purification of miRNA-protein complexes (>50% (Hammell, 2008) or ~70% (Easow et al., 2007); based on reports about ~30% of seed-containing true positives). We previously estimated the performance of Ago HITS-CLIP analysis as ~93% of specificity and ~5-27% of false positive rates and bioinformatics alone as ~67.5% of specificity and ~45% of false-positives (detail analysis in chapter 6). Therefore we can conclude that the Ago ternary maps (with a false positive rate of ~6.8-27%; majorly estimated by miR-124 datasets) achieved a minimum of a 1.5-fold, and a maximum of a 10 fold improvement in false-positives relative to other approaches (~40-70%).

We further analyzed observed versus expected frequencies for each seed match from the top 30 miRNAs in Ago-mRNA clusters (BC$\geq$2). Some apparent false-positives are in this group, as estimated by observed versus expected seed enrichment (e.g. miR-125, miR-708 and miR-324-3p); this is also apparent in the erratic shape of the curves in Figure 6.1. However, even within the top 20 robust miRNA set, we do not know the rules of miRNA binding (other than "classical"

seed rules), and this impacts upon the false positive set. In addition, even without including seed conservation, if we examine observed versus expected 8-mer seed frequency of miR-125, one of the miRNAs showing apparent false-positives with 6mer seeds, then we again see statistically significant seed enrichment in Ago-mRNA clusters (data not shown). Moreover, this data makes biologic sense, since miR-125 is expressed in the brain, and the GO targets identified match the known biology (with roles in cytoskeletal regulation; Figure 5.7 and 5.8).

- Sensitivity of Ago ternary map

The issue of how to estimate sensitivity and false negatives is clearly important, but it also must be recognized that this is a difficult issue, because there are in fact no gold standard "true" datasets upon which to answer this question other than the dataset for miR-124. Therefore, we focused our attempts to estimate false negatives on the set of miR124 targets validated by Hannon and colleagues (Karginov et al., 2007). Unfortunately, this set is only 22 deep. Nonetheless, this analysis yielded a false negative rate of 6/22~25%. For transcripts of even moderate abundance (normalized probes intensity >700), we identified all 10/10 predicted targets (Table 5.3). Moreover, if we examine raw clusters (i.e. without

using in silico normalization), we identified 21/22 predicted targets (~0.05% false negative rate).

We also could address sensitivity from this analysis. For the 6 potential false negatives in Table 5.3, transcript levels were low (normalized probe intensity was < 700 for each transcript (excluding Tom1l1, which appears to have an unusual antisense transcript expressed in brain covering the 3′ UTR)). Of all 12 transcripts in Table 5.3 with transcript levels at or below 700, 7 were identified, while 5 were missed, suggesting that for the rarest transcripts, we detected 58% of true positives. Classifying our data in this fashion, we have very low false negative rate estimates (10/10 positives) for over 50% of the transcriptome (> 700; versus median expression level = 771, and average expression level ~1,255). We also compared the recovery of transcripts as a function of abundance versus our ability to identify miR-124 targets. This result (Figure 7.1) confirms those observed on a smaller scale in Table 5.3, which is that low abundance transcripts are (not surprisingly) recovered less efficiently.

In addition, we can also estimate the sensitivity of the Ago-mRNA map as it relates to miRNA abundance. We readily detected miRNAs present at less than 1% of the total population (Figure 4.3) and the seed signal for these miRNAs

showed approximately equal enrichment between the top of this list (e.g. #1, miR-181a, 6% of total Ago-miRNA, Figure 4.3, shows 1,190 sites, and a 1.7 fold enrichment in the Ago-mRNA footprint region) and the end of this list (#29, miR-23, <0.36%, which shows 1,036 sites, and a 1.4 fold enrichment). Therefore, the top 20 miRNAs provide a robust set of enriched seed sequences in the Ago-mRNA footprints (Figure 7.1), and suggests that Ago-HITS-CLIP is likely to work well for additional miRNAs beyond this set. Another way of estimating the false negative rate is by comparing the number of Ago-mRNA clusters with no predicted seeds in the top 20 Ago-miRNAs, relative to the proportion of the top 20 Ago-miRNAs in the brain (as described in chapter 6). 27% of Ago-mRNA clusters have no predicted seed matches among the top 20 Ago-miRNAs families. Such orphan clusters estimated ~15% of false negative rate but also indicated that there might be other rules of binding, which prompted further analysis for miRNA target recognition beyond seed pairing (chapter 6)

miRNA target recognitions beyond seed.

Although we observed high correlation between the frequency of seeds in mRNA tags and the frequency of miRNA associated with Ago, there appears to be a disparity: the hexanucleotide that appears most frequently in Ago-mRNA tags (seed match from #9 ranking miR-124) is not the one corresponding to the

miRNA that appears most frequently in AGO-miRNA tags (#1 ranking miR-30) (Figure 5.1B and Table 5.1). We found that a miR-124 seed match is ~1.33 fold more frequent than a miR-30 seed match (51,461 total 6-mers for miR-124, vs. 38,752 for miR-30 in P13 mouse brain transcriptome), meaning that there are more miR-124 potential binding sites in transcripts. We initially interepreted this correlation as being consistent with a relative increase in miR-124 bound sites. However, it is also possible that there is underlying biologic regulation that we do not understand, or that there are additional rules of binding that are not apparent in current seed centric analyses. Actually, we observed some apparent abundant miRNAs (e.g. #7 ranking miR-708) showed enrichment of sequence matches from regions other than seed (e.g. position ~8-16 in miR-708) in mRNA tags (Figure 6.3C). In general, we observed that ~50% of brain expressed miRNAs do not have their seed matches enriched in Ago-mRNA tags (Figure 6.3). Such miRNAs including miR-708 should be further investigated to confirm their new target recognition rules by analyzing de novo miRNA target sites identified by Ago HITS-CLIP in the cell where each of individual miRNA is transfected.

By analyzing de novo miR-124 clusters (clusters only identified by Ago HITS-CLIP in miR-124 transfected HeLa cell), we found that ~15% of miR-124 target sites contain G bulge (G between position 5 and 6 of seed match) in addition to

~50% of the seed match sites (Figure 6.6). Other nucleotide bulges (A, U, or C) at the same position (between position 5 and 6 of seed match) are neither statistically enriched in bioinformtic analysis nor responsive to miR-124 expression in luciferase reporter assays (Figure 6.6 and 6.7). In fact, a free energy calculation for any nucleotide bulge between position 5 and 6 gives the same value (-5.6 kcal/mol ; the free energy calculated for pairing it with miR-124 seed (position 2-8) by RNAduplex (Hofacker, 2003)). However, only a G bulge sequence can also make another 5 consecutive perfect match starting from position 2. The free energy calculation of this pair is -6.8kcal/mol, lower than the free energy from the bulge pair (-5.6kcal/mol). Based on computational analysis of some validated miRNA target sites, seed annealing is thought to initiate miRNA-mRNA duplex followed by further annealing which stabilizes thermodynamics, indicating that the free energy of seed pairing is more important than the free energy of the entire pairing (Rajewsky and Socci, 2004). Therfore, we can make a hypothesis that preference of G position in the bulge may be caused by free energy contribution from another optimal pairing from G bulge sequences and this hypothesis need to be further tested by luciferase reporter assays with mutant constructs. Although additional G bulge rule was accounted, ~35% of orphan clusters are still remained as de novo miR-124 target sites. Surprisingly, we could not find any consensus sequence or significant free

energy change by annealing with miR-124 in these remaining orphan clusters. Recently, it is reported that several miR-24 targets also contain functional "seedless" miRNA recognition elements without any consensus motifs and signigicant free energy changes (Lal et al., 2009) like our results, suggesting that our remaining orphans still may be associated with functional binding sites through complicated unknown rules, which should be further analyzed through more sophisticated bioinformatic analysis.

Improvement of Ago HITS-CLIP

The identification of the orphan clusters from Ago HITS-CLIP analysis prompted us to think of further improvement of bioinformatic analysis in current HITS-CLIP method. In the processing step of Ago-mRNA tags, we used a simulation approach called "*in silico* random CLIP" to simulate distributing sequence tags across messenger RNAs based on expression levels of mRNAs. We measured abundance of transcrtipts by performing microarray instead of "RNA-Seq" technique, high-throuput sequencing method for measuring expression levels of transcripts. Unfortunately, RNA-Seq was not established as a method until very recently, and was not available when Ago HITS-CLIP method was developed. Although RNA-Seq is better than microarray in terms of accuracy and sensitivity, current study suggests that microarray (exon array) is still highly correlated with

RNA-Seq data (probe intensity vs sequence freq(log_2)) (Marioni et al., 2008). However, it is very likely that there is bias related to how well the probes on the exon array can measure transcripts. As RNA-Seq is becoming an important tool going forward particularly as the depth of sequencing increases, *in silico* random CLIP algorithm should use RNA-Seq for measuring abundance of transcripts in the future. Using RNA-Seq will result in actual counts for the whole transcripts and be motivated as these counts should be more comparable across genes.

In addition to improve normalization process, development of a computational analysis will be required to further identify motifs or rules of miRNA recognition especially for remaining orphan clusters. As mentioned before, Ago HITS-CLIP method can be coupled to any computational method for finding targets and improve the quality of the predictions. Bioinformatic approaches should be improved to identify critical features of the target sequences such as positions of consensus sequences, combinatorial rules of regulatory RNA elements, secondary structures of the RNA binding site, and regulatory rules of target RNAs in terms of functional networks. Such rules and findings may achieved by using machine learning algorithms, learning from the various features obtained from various bioinformatic tools such as evolutional conservation (Multiz5way), enrichments of sequences (MEME), local RNA folding (mFold, RNAfold) and the

functions of target genes (EASE, GOminer) (Bailey and Elkan, 1994; Hofacker, 2003; Hosack et al., 2003; Thomas et al., 2003; Zeeberg et al., 2003; Zuker, 2003).

Our current Ago HITS CLIP analysis still depends on bioinformatic predictions to pair-up miRNAs with collected mRNA target sites, which can not definitively avoid from false-positives and false-negatives. If we can improve current Ago HITS-CLIP method to experimentally pair-up miRNA with mRNA target sequences, it would dramatically reduce false-postives. As miRNAs have 5′ phosphates which buried into Mid domain of Ago (Wang et al., 2008a), we may be able to only dephosphorylate 3′ phosphates of mRNA fragments generated by an RNaseA treatment and induce subsequent ligation of a protected 5′ phosphate of miRNA with a new 3′ hydroxyl group of an mRNA fragment. A prerequisite condition for this ligation might be crosslinking of both a miRNA and an mRNA fragment to Ago. In the beginning, we did not expect to see much evidence of Ago, miRNA and mRNA crosslinked together, because the efficiency of crosslinking is quite low which we estimated in Nova HITS-CLIP as on the order of 1% (Unpublished data). Thus the frequency of crosslinking Ago to both a miRNA and an mRNA would be estimated to be ~0.01%, an uncommon event.

However, we did in fact observed that the bigger complexes (130kD, Ago-mRNA complex) contained a subset of miRNAs (~22nt) and shorter mRNA tags (~30nt)

(Figure 4.2C) and further confirmed similar results from distribution analysis of length of tags in high-throuput sequencing data (Figure 7.2A-B). Although we amplified RT-PCR products from average ~50nt RNA-protein complex (130kD, crosslinked Ago-mRNA complex) and subsequently purified ~50nt PCR products for high-throughput sequencing, we still got ~10% of miRNA tags (~20nt) and also ~10% of putative co-crosslinked mRNA tags (~30nt) (Figure 7.2A-B). Such observations are absent in average ~50nt RNA-Nova complex. These observations implicates that actual co-crosslinking efficiency may be greater than ~10% in Ago HITS-CLIP. Therefore, these hypothesis are worth further experimentally considered and validated to improve the Ago HITS-CLIP method in the future. However, it is also possible that the ~20nt tags were smeared on the gel from the lower band, and that the ~30nt tags were stochastic variants running trapped with the larger complexes.

Ago mRNA clusters outside of 3′ UTR.

When Ago mRNA tags were overlaid with gene annotations, several patterns emerged (Figure 4.10C-D). As expected, a substantial portion (~40%) of Ago-bound tags were in 3′ UTRs where miRNA activity is known to have high efficacy. Some ~8% (one-fifth of the ~40%) were actually outside of the UTR but within 10kb downstream, regions likely to harbor unannotated 3′ UTRs.

Unsurprisingly, very few Ago-bound tags were in 5′ UTRs (~1.9% of clusters in transcripts; ~6.8% of expected frequency; Figure 4.11). Although those sites in 5'UTR are very few, they are likely to be functional based on the observation that they are biologically reproducible (BC=5) and the report that miRNA-guided repression is equally efficient when sites are located in the 5' or 3' UTRs of reporter transcripts (Lytle et al., 2007). It is worth testing the efficiency of these sites in 5'UTR by performing luciferase reporter assays.

However, a substantial fraction of tags fell in coding sequences (CDS) (~25%), introns (~12%), and non-coding RNAs (~4%), suggesting that miRNA activity occurs in these regions as well. Recently, there are several reports supporting functional miRNA target sites in CDS (Duursma et al., 2008; Forman et al., 2008; Lal, 2008; Shen et al., 2008; Tay et al., 2008). In C. elegans, it is reported that Ago transports siRNAs from the cytoplasm to the nucleus (Guang et al., 2008) and in human, miR-29b is shown to predominantly localize to the nucleus (Hwang et al., 2007). We also observed that Ago mRNA clusters in introns often contain seed matches from highly expressed miRNAs (data not shown). Therefore, Ago ternary complex may localize to nucleus, bind to introns of the targets and associated there to repress the targets possibly through endonuclease activity of Ago. By cleaving the targets at the pri-mRNA form, it may assure the repression

of the targets which need to be translationally repressed also in the cytoplasm. Ago-RNA bindings, detected in non-coding RNAs, could be the functional target sites of miRNA to degrade such non-coding RNAs which may have critical roles in biological function or the regulation of gene expression. However, they also could be considered as non-coding RNAs interacting with Ago to directly regulate the RISC activity, primary transcripts encoding miRNAs or other new class of small RNAs which could be loaded on RISC, since we can not discriminate whether RNAs from Ago CLIP experiments are effector (like miRNAs) or targets (like miRNA target mRNAs). It would be interesting to test such a hypothesis by reporter assay or biochemical assay in the future. In addition, these unexpected locations of miRNA binding may offer additional insights into the mechanisms of miRNA regulation, which also need to be further examined by HITS-CLIP.

miR-124 target regulation elucidated by Ago HITS-CLIP

We used Ago HITS-CLIP to identify a brain map of transcripts regulated by miR-124, a miRNA proposed to play a role in establishing and perhaps maintaining neuronal identity, in part by suppression of transcripts that encode proteins that suppress neuronal identity. The latter group has been suggested to include Cdtsp1, its target protein REST, and the RNABP Ptbp1, although screens to

distinguish direct from indirect targets have relied on bioinformatic analysis of miR-124 binding sites. Our data helps to confirm that Ctdsp1 and Ptbp1 are direct miR-124 targets, while refining the target sites to which Ago-miR-124 actually binds. The absence of Ago binding sites in REST at P13 mouse brain (data not shown) is consistent with REST playing an indirect (from an Ago perspective) role in neuronal identity. Interestingly, we found that overexpression of miR-124 can induce Ago to bind to validated miR-124 regulatory sites, but also that this binding often precludes Ago binding to sites that would otherwise be occupied, perhaps reflecting competition between a limited capacity for miRNA binding on a given 3′ UTR (Khan et al., 2009). More generally, the resolution of Ago regulated miR-124 pathways suggests that Ago HITS-CLIP can be used to define the role of miRNAs and the target transcripts implicated in a variety of biologic contexts, such as miR-133b in dopaminergic neurons (Kim et al., 2007), lsy-6 in neuronal left/right axis determination (Johnston and Hobert, 2003b), or the general role of miRNAs in cerebellar development (Schaefer et al., 2007).

Interplays between RNABPs and miRNAs

Ago HITS-CLIP offers dramatic improvement in identifying miRNA target sites especially for transcripts with highly conserved 3′ UTRs, which often have many

"predicted" miRNA binding sites because so many computational methods rely on conservation. For example, the 3' UTRs of the neuronal RNABPs Ptbp2, Nova1, and Fmr1 are highly conserved and have many predicted miRNA sites (in some cases over 50 using robust algorithms such as TargetScan or PicTar), but each has only single Ago-mRNA CLIP clusters (Figure 5.9). More generally, Ago HITS-CLIP demonstrates that the level of miRNA selectivity is very high such that on average transcripts have between 1-3 major Ago binding sites in a single tissue (Figure 5.3, 5.6 and 5.9). This implies that additional factors may contribute to the selectivity of miRNA-seed recognition. We find no evidence that Ago-mRNA binding sites themselves have a specific sequence preference (data not shown), suggesting that specific sites may be protected or made accessible by the binding of other sequence-specific RNABPs. Such a mechanism might provide a means of dynamically regulating miRNA binding and hence miRNA-mediated RNA regulation itself.

Based on the HITS-CLIP results and some reports about miRNA target regulation by RNABPs, we can generate models suggesting the mechanisms and effects of the interplay between RNABPs and Ago bindings on miRNA targets (Figure 7.2). RNABP bindings may be competitive to miRNA bindings (Figure 7.2A) like reports that HuR, an AU-rich-element (ARE) binding protein, binds

CAT1 and inhibits its repression by miR-122 (Bhattacharyya et al., 2006b). The same result was observed in Dead end 1 (DnD1) that it also protects target mRNAs from miRNA-mediated repression (Kedde et al., 2007b). RNA binding may activate miRNA mediated repression through cooperative interactions with target transcripts, through increasing local concentration of Ago complex (Figure 7.2B) or resolving a secondary structure hindering access of Ago complex to a miRNA target site (Figure 7.2C). This model is supported by the observation that binding of HuR to 3'UTR of c-Myc recruits let-7-loaded RISC and repress c-Myc expression (Kim et al., 2009a). Furthermore, RNABP may be involved in switching of miRNA mediated repression to activation (Figure 7.2D). Interestingly, it was shown that ARE recruits miR-369-3 associated with RISC together with a RNA binding protein called fragile X mental retardation-related protein 1 (FXR1) to tumor necrosis factor-alpha (TNFalpha) mRNA and activates translation (Vasudevan et al., 2007). However, a major limitation of this study is lack of understanding of location of binding sites of both RNABPs and miRNAs on target transcripts, which is prerequisite to understanding mechanisms. Such mechanisms, which may be assessed by overlaying HITS-CLIP maps of different RNABPs and Ago-ternary compelx with combination of genetic approach (RNABP null mice vs wild type) and experiments for measuring miRNA-mediated repression such as microarray (Lim et al., 2005) or SILAC (Baek et al.,

2008; Selbach et al., 2008), could provide a means of dynamically regulating miRNA binding and regulation (Figure 7.2).

Conclusion

In conclusion, we developed HITS-CLIP to comprehensively understand RNA complexity regulated by RNABPs and miRNAs. HITS-CLIP offers a number of new opportunities to explore RNABP and miRNA biology. Most significantly, we demonstrate that bioinformatic analysis of the data from a single set of experiments can identify the target sites on transcripts in a given biologic tissue with great specificity (~92.5%; ~15% of false negatve rates from Ago HITS-CLIP). Moreover, the ability to simultaneously assess multiple miRNAs offers a means of understanding combinatorial control of target RNA expression. Identification of specific binding sites also offers the possibility of specifically disrupting or enhancing miRNA or RNABP activity, for example with interfering or stable RNA analogues (Elmen et al., 2008). The unbiased nature of HITS-CLIP offers the possibility of confirming and discovering new biology; unexpected findings that the processes of some miRNAs may be regulated by Nova and the relatively large number of Ago complex bind in coding sequence, as well as some in introns, suggesting unexplored functions for Ago-miRNA complex. Moreover, not all Ago mRNA clusters correspond to known seed sequences (23% Ago foot

print regions have no conserved seed sites among all Ago-miRNAs) which lead to identifying miRNA recognitions beyond seed pairing. However, remaining ~35% of orphans from Ago miR-124 clusters suggest that further improvement of Ago HITS-CLIP analysis may help uncovering additional rules dictating Ago-miRNA-target interactions. Such specificity may be determined, in part, through additional cis-acting RNA-protein complexes (Bhattacharyya et al., 2006; Kedde et al., 2007a), and thus overlaying Ago HITS-CLIP maps with those for other RNABPs (Licatalosi et al., 2008) would be informative to understand RNA regulations and complexity

Figure 7.1. Ago ternary maps depending on the level of transcripts and distributional analysis of the length of CLIP tags.

A. Fraction of total p13 brain transcripts (blue line) or transcripts with Ago-miR-124 cluster (red line) according to their expression indicated as log2 ratio (normalized probe intensity vs median probe intensity (771)), showing that low abundance transcripts are recovered less efficiently by Ago HITS-CLIP. B. Distribution of mRNA tags from Ago HITS-CLIP (blue line) and Nova HITS-CLIP (red line) according to their length. Sequence data from 454 sequencing (B) or Solexa/Illumina genome analyzer (C) are plotted. Distinct peaks in the distributions are marked by arrows.

Figure 7.1

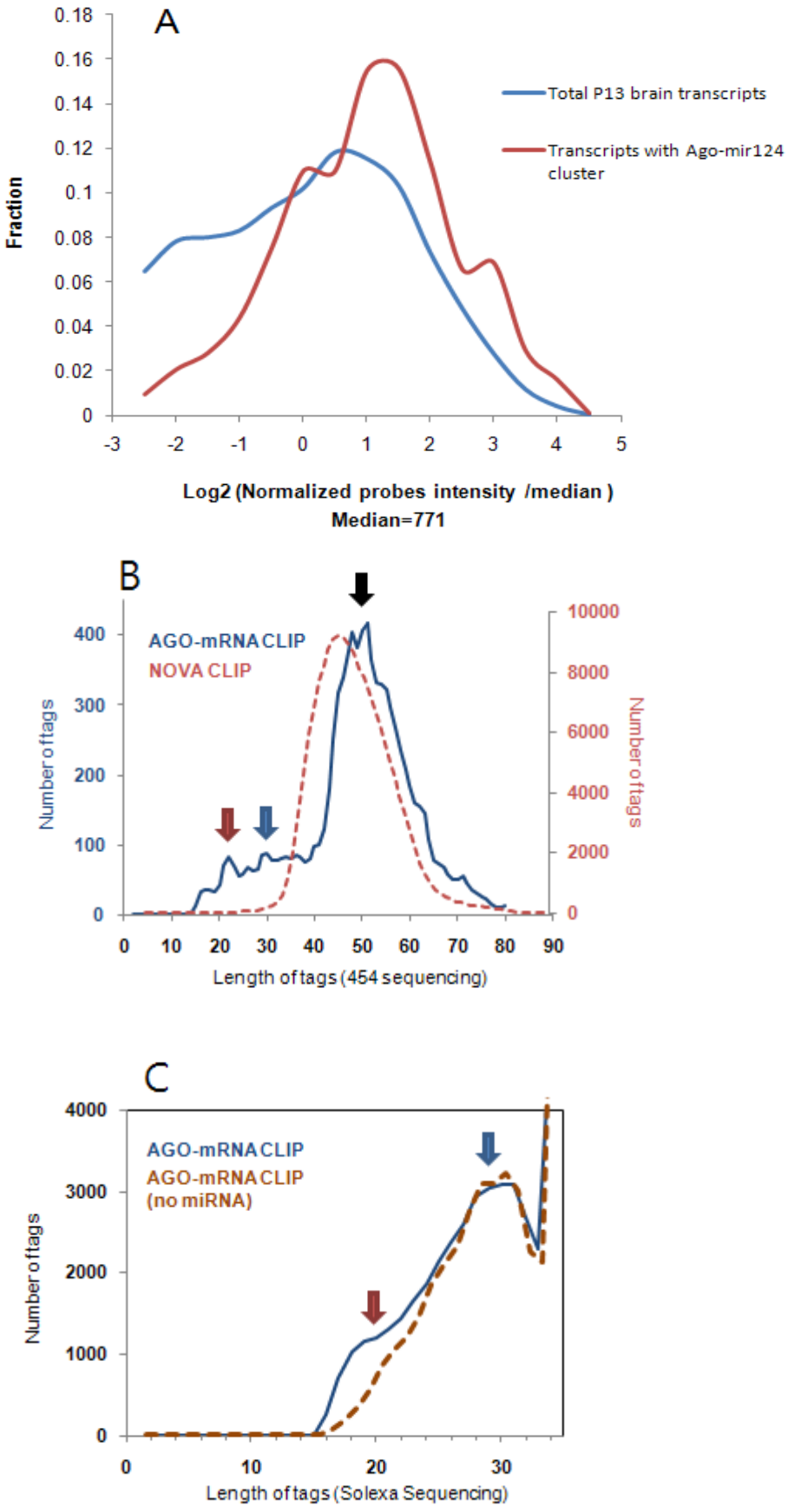
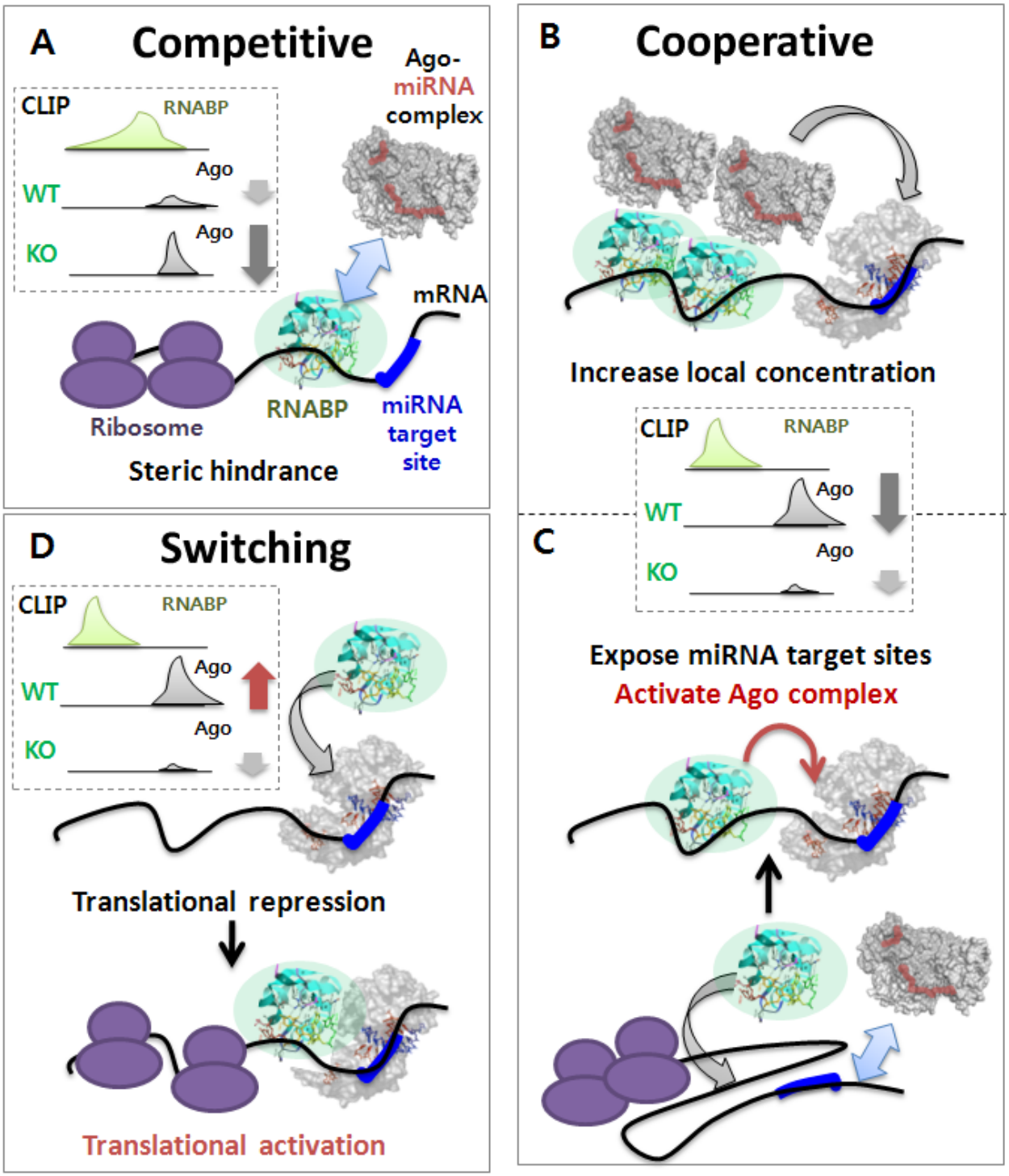

A
Total P13 brain transcripts
Transcripts with Ago-mir124 cluster
Fraction
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
-3
-2
-1
0
1
2
3
4
5
Log2 (Normalized probes intensity /median)
Median=771
B
AGO-mRNA CLIP
NOVA CLIP
Number of tags
Number of tags
400
300
200
100
0
10000
8000
6000
4000
2000
0
0
10
20
30
40
50
60
70
80
90
Length of tags (454 sequencing)
C
AGO-mRNA CLIP
AGO-mRNA CLIP
(no miRNA)
Number of tags
4000
3000
2000
1000
0
0
10
20
30
Length of tags (Solexa Sequencing)

Figure 7.2. Models of interplays between RNABPs and Ago-miRNA complexes

A. A model of competitive bindings between an RNABP and an Ago-miRNA complex. An RNABP-RNA interaction may inhibit Ago-miRNA from binding to an adjacent miRNA target site through steric hindrance. Therefore, the RNABP binding can protect the target transcript from miRNA-mediated repression. In RNABP null cells, more Ago-mRNA CLIP tags in the target site and enhanced miRNA mediated target repression are expected to be shown as indicated as box. B. A model of cooperative bindings between RNABPs and Ago-miRNA complexes, showing increased local concentration of Ago-miRNA complexes by RNABPs and subsequent Ago termary complex formation. If RNABPs can interact with Ago-miRNA complexes, binding of RNABPs to their interacting sites (high affinity bindings), adjacent to miRNA target sites, may increase the local concentration of Ago-miRNA complexes and facilitate their binding to the miRNA target sites (low affinity binding). C. The same cooperative model with different mechanisms. Secondary structure of an RNA transcript may hinder the access of Ago-miRNA complex to its miRNA binding site. The binding of an RNABP to the same transcript may refold the target RNA into the structure releasing a miRNA binding site. In addition, a neighboring RNA associated RNABP may further activate Ago-ternary complex. In RNABP null cells, less Ago-mRNA CLIP tags in the target site and less effect from miRNA mediated repression are expected to be shown. D. A model showing that an RNABP binding to Ago ternary complex can switch a miRNA-mediated target regulation from repression to activation. In the presence of an agonistic RNABP, more Ago-mRNA binding in the target site would lead to increased miRNA mediated translational acitivation.

Figure 7.2


A
Competitive
Ago-
miRNA
complex
CLIP
RNABP
Ago
WT
Ago
KO
mRNA
Ribosome
RNABP
miRNA
target
site
Steric hindrance
B
Cooperative
Increase local concentration
CLIP
RNABP
Ago
WT
Ago
KO
C
Expose miRNA target sites
Activate Ago complex
D
Switching
CLIP
RNABP
Ago
WT
Ago
KO
Translational repression
Translational activation

REFERENCES

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

Antar, L.N., Afroz, R., Dictenberg, J.B., Carroll, R.C., and Bassell, G.J. (2004). Metabotropic glutamate receptor activation regulates fragile x mental retardation protein and FMR1 mRNA localization differentially in dendrites and at synapses. J Neurosci *24*, 2648-2655.

Antar, L.N., Dictenberg, J.B., Plociniak, M., Afroz, R., and Bassell, G.J. (2005). Localization of FMRP-associated mRNA granules and requirement of microtubules for activity-dependent trafficking in hippocampal neurons. Genes, brain, and behavior *4*, 350-359.

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*, 64-71.

Bailey, T.L., and Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc Int Conf Intell Syst Mol Biol *2*, 28-36.

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.

Bassell, G.J., Zhang, H., Byrd, A.L., Femino, A.M., Singer, R.H., Taneja, K.L., Lifshitz, L.M., Herman, I.M., and Kosik, K.S. (1998). Sorting of beta-actin mRNA and protein to neurites and growth cones in culture. J Neurosci *18*, 251-265.

Beaudoing, E., Freier, S., Wyatt, J.R., Claverie, J.M., and Gautheret, D. (2000). Patterns of variant polyadenylation signal usage in human genes. Genome Res *10*, 1001-1010.

Behm-Ansmant, I., Rehwinkel, J., Doerks, T., Stark, A., Bork, P., and Izaurralde, E. (2006). mRNA degradation by miRNAs and GW182 requires both CCR4:NOT deadenylase and DCP1:DCP2 decapping complexes. Genes Dev *20*, 1885-1898.

Beitzinger, M., Peters, L., Zhu, J.Y., Kremmer, E., and Meister, G. (2007). Identification of human microRNA targets from isolated argonaute protein complexes. RNA Biol *4*, 76-84.

Bemmo, A., Benovoy, D., Kwan, T., Gaffney, D.J., Jensen, R.V., and Majewski, J. (2008). Gene expression and isoform variation analysis using Affymetrix Exon Arrays. BMC Genomics *9*, 529.

Bennett, S. (2004). Solexa Ltd. Pharmacogenomics *5*, 433-438.

Bennett, S.T., Barnes, C., Cox, A., Davies, L., and Brown, C. (2005). Toward the 1,000 dollars human genome. Pharmacogenomics *6*, 373-382.

Bentwich, I. (2005). Prediction and validation of microRNAs and their targets. FEBS Lett *579*, 5904-5910.

Bentwich, I., Avniel, A., Karov, Y., Aharonov, R., Gilad, S., Barad, O., Barzilai, A., Einat, P., Einav, U., Meiri, E.*, et al.* (2005). Identification of hundreds of conserved and nonconserved human microRNAs. Nat Genet *37*, 766-770.

Berezikov, E. (2006). Diversity of microRNAs in human and chimpanzee brain. Nature Genet *38*, 1375-1377.

Bernstein, E., Caudy, A.A., Hammond, S.M., and Hannon, G.J. (2001). Role for a bidentate ribonuclease in the initiation step of RNA interference. Nature *409*, 363-366.

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.

Bhattacharyya, S.N., Habermacher, R., Martine, U., Closs, E.I., and Filipowicz, W. (2006). Relief of microRNA-mediated translational repression in human cells subjected to stress. Cell *125*, 1111-1124.

Blencowe, B.J., Ahmad, S., and Lee, L.J. (2009). Current-generation high-throughput sequencing: deepening insights into mammalian transcriptomes. Genes Dev *23*, 1379-1386.

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

Borchert, G.M., Lanier, W., and Davidson, B.L. (2006). RNA polymerase III transcribes human microRNAs. Nature structural & molecular biology *13*, 1097-1101.

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

Brown, P.O., and Botstein, D. (1999). Exploring the new world of the genome with DNA microarrays. Nat Genet *21*, 33-37.

Brown, V., Jin, P., Ceman, S., Darnell, J.C., O'Donnell, W.T., Tenenbaum, S.A., Jin, X., Feng, Y., Wilkinson, K.D., Keene, J.D., *et al.* (2001). Microarray identification of FMRP-associated brain mRNAs and altered mRNA translational profiles in fragile X syndrome. Cell *107*, 477-487.

Buckanovich, R.J., and Darnell, R.B. (1997). The neuronal RNA binding protein Nova-1 recognizes specific RNA targets in vitro and in vivo. Mol Cell Biol *17*, 3194-3201.

Buckanovich, R.J., Posner, J.B., and Darnell, R.B. (1993). Nova, the paraneoplastic Ri antigen, is homologous to an RNA-binding protein and is specifically expressed in the developing motor system. Neuron *11*, 657-672.

Buckanovich, R.J., Yang, Y.Y., and Darnell, R.B. (1996). The onconeural antigen Nova-1 is a neuron-specific RNA-binding protein, the activity of which is inhibited by paraneoplastic antibodies. J Neurosci *16*, 1114-1122.

Carmell, M.A., Xuan, Z., Zhang, M.Q., and Hannon, G.J. (2002). The Argonaute family: tentacles that reach into RNAi, developmental control, stem cell maintenance, and tumorigenesis. Genes Dev *16*, 2733-2742.

Castle, J.C., Zhang, C., Shah, J.K., Kulkarni, A.V., Kalsotra, A., Cooper, T.A., and Johnson, J.M. (2008). Expression of 24,426 human alternative splicing events and predicted cis regulation in 48 tissues and cell lines. Nat Genet *40*, 1416-1425.

Chen, C., Ridzon, D.A., Broomer, A.J., Zhou, Z., Lee, D.H., Nguyen, J.T., Barbisin, M., Xu, N.L., Mahuvakar, V.R., Andersen, M.R., *et al.* (2005). Real-time quantification of microRNAs by stem-loop RT-PCR. Nucleic Acids Res *33*, e179.

Chendrimada, T.P., Finn, K.J., Ji, X., Baillat, D., Gregory, R.I., Liebhaber, S.A., Pasquinelli, A.E., and Shiekhattar, R. (2007). MicroRNA silencing through RISC recruitment of eIF6. Nature *447*, 823-828.

Cheng, L.C., Pastrana, E., Tavazoie, M., and Doetsch, F. (2009). miR-124 regulates adult neurogenesis in the subventricular zone stem cell niche. Nat Neurosci *12*, 399-408.

Clark, T.A., Schweitzer, A.C., Chen, T.X., Staples, M.K., Lu, G., Wang, H., Williams, A., and Blume, J.E. (2007). Discovery of tissue-specific exons using comprehensive human exon microarrays. Genome Biol *8*, R64.

Consortium, C.e.S. (1998). Genome sequence of the nematode C. elegans: a platform for investigating biology. Science *282*, 2012-2018.

Darnell, J.C., Fraser, C.E., Mostovetsky, O., Stefani, G., Jones, T.A., Eddy, S.R., and Darnell, R.B. (2005). Kissing complex RNAs mediate interaction between the Fragile-X mental retardation protein KH2 domain and brain polyribosomes. Genes Dev *19*, 903-918.

Darnell, J.C., Jensen, K.B., Jin, P., Brown, V., Warren, S.T., and Darnell, R.B. (2001). Fragile X mental retardation protein targets G quartet mRNAs important for neuronal function. Cell *107*, 489-499.

Darnell, R.B. (2006). Developing global insight into RNA regulation. Cold Spring Harb Symp Quant Biol *71*, 321-327.

Davuluri, R.V., Suzuki, Y., Sugano, S., Plass, C., and Huang, T.H. (2008). The functional consequences of alternative promoter use in mammalian genomes. Trends Genet *24*, 167-177.

De Diego Otero, Y., Severijnen, L.A., van Cappellen, G., Schrier, M., Oostra, B., and Willemsen, R. (2002). Transport of fragile X mental retardation protein via granules in neurites of PC12 cells. Mol Cell Biol *22*, 8332-8341.

Denli, A.M., Tops, B.B., Plasterk, R.H., Ketting, R.F., and Hannon, G.J. (2004). Processing of primary microRNAs by the Microprocessor complex. Nature *432*, 231-235.

Didiano, D., and Hobert, O. (2006). Perfect seed pairing is not a generally reliable predictor for miRNA-target interactions. Nature Struct Mol Biol *13*, 849-851.

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

Dredge, B.K., and Darnell, R.B. (2003). Nova regulates GABA(A) receptor gamma2 alternative splicing via a distal downstream UCAU-rich intronic splicing enhancer. Mol Cell Biol *23*, 4687-4700.

Dsouza, M., Larsen, N., and Overbeek, R. (1997). Searching for patterns in genomic data. Trends Genet *13*, 497-498.

Duursma, A.M., Kedde, M., Schrier, M., le Sage, C., and Agami, R. (2008). miR-148 targets human DNMT3b protein coding region. RNA *14*, 872-877.

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

Ebert, M.S., Neilson, J.R., and Sharp, P.A. (2007). MicroRNA sponges: competitive inhibitors of small RNAs in mammalian cells. Nature Methods *4*, 721-726.

Eddy, S.R., and Durbin, R. (1994). RNA sequence analysis using covariance models. Nucleic Acids Res *22*, 2079-2088.

Elmen, J., Lindow, M., Schutz, S., Lawrence, M., Petri, A., Obad, S., Lindholm, M., Hedtjarn, M., Hansen, H.F., Berger, U.*, et al.* (2008). LNA-mediated microRNA silencing in non-human primates. Nature *452*, 896-899.

Feng, Y., Absher, D., Eberhart, D.E., Brown, V., Malter, H.E., and Warren, S.T. (1997). FMRP associates with polyribosomes as an mRNP, and the I304N mutation of severe fragile X syndrome abolishes this association. Mol Cell *1*, 109-118.

Filipowicz, W., Bhattacharyya, S.N., and Sonenberg, N. (2008). Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? Nature Rev Genet *9*, 102-114.

Foat, B.C., Houshmandi, S.S., Olivas, W.M., and Bussemaker, H.J. (2005). Profiling condition-specific, genome-wide regulation of mRNA stability in yeast. Proc Natl Acad Sci U S A *102*, 17675-17680.

Forman, J.J., Legesse-Miller, A., and Coller, H.A. (2008). A search for conserved sequences in coding regions reveals that the let-7 microRNA targets Dicer within its coding sequence. Proc Natl Acad Sci USA *105*, 14879-14884.

Garber, K., Smith, K.T., Reines, D., and Warren, S.T. (2006). Transcription, translation and fragile X syndrome. Curr Opin Genet Dev *16*, 270-275.

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.

Gregory, R.I., Chendrimada, T.P., Cooch, N., and Shiekhattar, R. (2005). Human RISC couples microRNA biogenesis and posttranscriptional gene silencing. Cell *123*, 631-640.

Grimson, A. (2007). MicroRNA targeting specificity in mammals: determinants beyond seed pairing. Mol Cell *27*, 91-105.

Guang, S., Bochner, A.F., Pavelec, D.M., Burkhart, K.B., Harding, S., Lachowiec, J., and Kennedy, S. (2008). An Argonaute transports siRNAs from the cytoplasm to the nucleus. Science *321*, 537-541.

Guil, S., and Caceres, J.F. (2007). The multifunctional RNA-binding protein hnRNP A1 is required for processing of miR-18a. Nature structural & molecular biology *14*, 591-596.

Hammell, M. (2008). mirWIP: microRNA target prediction based on microRNA-containing ribonucleoprotein-enriched transcripts. Nature Methods *5*, 813-819.

Hammond, S.M. (2006). MicroRNAs as oncogenes. Curr Opin Genet Dev *16*, 4-9.

Han, J., Lee, Y., Yeom, K.H., Kim, Y.K., Jin, H., and Kim, V.N. (2004). The Drosha-DGCR8 complex in primary microRNA processing. Genes Dev *18*, 3016-3027.

Han, J., Lee, Y., Yeom, K.H., Nam, W., Heo, I., Rhee, J.K., Sohn, S.Y., Cho, Y., Zhang, B.T., and Kim, V.N. (2006) Molecular basis for the recognition of primary microRNAs by the Drosha-DGCR8 complex. Cell 125, 887-901.

He, L., and Hannon, G.J. (2004). MicroRNAs: small RNAs with a big role in gene regulation. Nature Rev Genet *5*, 522-531.

He, L., He, X., Lim, L.P., de Stanchina, E., Xuan, Z., Liang, Y., Xue, W., Zender, L., Magnus, J., Ridzon, D., *et al.* (2007). A microRNA component of the p53 tumour suppressor network. Nature *447*, 1130-1134.

He, L., Thomson, J.M., Hemann, M.T., Hernando-Monge, E., Mu, D., Goodson, S., Powers, S., Cordon-Cardo, C., Lowe, S.W., Hannon, G.J., *et al.* (2005). A microRNA polycistron as a potential human oncogene. Nature *435*, 828-833.

Hendrickson, D.G., Hogan, D.J., Herschlag, D., Ferrell, J.E., and Brown, P.O. (2008). Systematic identification of mRNAs recruited to argonaute 2 by specific microRNAs and corresponding changes in transcript abundance. PLoS ONE *3*, e2126.

Hofacker, I.L. (2003). Vienna RNA secondary structure server. Nucleic Acids Res *31*, 3429-3431.

Hosack, D.A., Dennis, G., Jr., Sherman, B.T., Lane, H.C., and Lempicki, R.A. (2003). Identifying biological themes within lists of genes with EASE. Genome Biol *4*, R70.

Houseley, J., and Tollervey, D. (2009). The many pathways of RNA degradation. Cell *136*, 763-776.

Huang, C.S., Shi, S.H., Ule, J., Ruggiu, M., Barker, L.A., Darnell, R.B., Jan, Y.N., and Jan, L.Y. (2005). Common molecular pathways mediate long-term potentiation of synaptic excitation and slow synaptic inhibition. Cell *123*, 105-118.

Humphreys, D.T., Westman, B.J., Martin, D.I., and Preiss, T. (2005). MicroRNAs control translation initiation by inhibiting eukaryotic initiation factor 4E/cap and poly(A) tail function. Proc Natl Acad Sci U S A *102*, 16961-16966.

Hutchison, C.A., 3rd (2007). DNA sequencing: bench to bedside and beyond. Nucleic Acids Res *35*, 6227-6237.

Hutvagner, G., and Simard, M.J. (2008). Argonaute proteins: key players in RNA silencing. Nat Rev Mol Cell Biol *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.

Irwin, S.A., Galvez, R., and Greenough, W.T. (2000). Dendritic spine structural anomalies in fragile-X mental retardation syndrome. Cereb Cortex *10*, 1038-1044.

Jackson, R.J., and Standart, N. (2007). How do microRNAs regulate gene expression? Sci STKE *2007*, re1.

Jelen, N., Ule, J., Zivin, M., and Darnell, R.B. (2007). Evolution of Nova-dependent splicing regulation in the brain. PLoS Genet *3*, 1838-1847.

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.

Jensen, K.B., Dredge, B.K., Stefani, G., Zhong, R., Buckanovich, R.J., Okano, H.J., Yang, Y.Y., and Darnell, R.B. (2000a). Nova-1 regulates neuron-specific alternative splicing and is essential for neuronal viability. Neuron *25*, 359-371.

Jensen, K.B., Musunuru, K., Lewis, H.A., Burley, S.K., and Darnell, R.B. (2000b). The tetranucleotide UCAY directs the specific recognition of RNA by the Nova K-homology 3 domain. Proc Natl Acad Sci U S A *97*, 5740-5745.

Jin, P., and Warren, S.T. (2000). Understanding the molecular basis of fragile X syndrome. Hum Mol Genet *9*, 901-908.

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

Johnson, J.M., Castle, J., Garrett-Engele, P., Kan, Z., Loerch, P.M., Armour, C.D., Santos, R., Schadt, E.E., Stoughton, R., and Shoemaker, D.D. (2003). Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. Science *302*, 2141-2144.

Johnson, J.M., Edwards, S., Shoemaker, D., and Schadt, E.E. (2005). Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments. Trends Genet *21*, 93-102.

Johnston, R.J., and Hobert, O. (2003). A microRNA controlling left/right neuronal asymmetry in Caenorhabditis elegans. Nature *426*, 845-849.

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.

Kaytor, M.D., and Orr, H.T. (2001). RNA targets of the fragile X protein. Cell *107*, 555-557.

Kedde, M., Strasser, M.J., Boldajipour, B., Oude Vrielink, J.A., Slanchev, K., le Sage, C., Nagel, R., Voorhoeve, P.M., van Duijse, J., Orom, U.A., *et al.* (2007a). RNA-binding protein Dnd1 inhibits microRNA access to target mRNA. Cell *131*, 1273-1286.

Kedde, M., Strasser, M.J., Boldajipour, B., Vrielink, J., Slanchev, K., le Sage, C., Nagel, R., Voorhoeve, P.M., van Duijse, J., and om, U.A. (2007b). RNA-binding protein Dnd1 inhibits microRNA access to target mRNA. Cell *131*, 1273-1286.

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

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.

Kertesz, M., Iovino, N., Unnerstall, U., Gaul, U., and Segal, E. (2007). The role of site accessibility in microRNA target recognition. Nat Genet *39*, 1278-1284.

Khan, A.A., Betel, D., Miller, M.L., Sander, C., Leslie, C.S., and Marks, D.S. (2009). Transfection of small RNAs globally perturbs gene regulation by endogenous microRNAs. Nat Biotechnol *27*, 549-555.

Kim, H.H., Kuwano, Y., Srikantan, S., Lee, E.K., Martindale, J.L., and Gorospe, M. (2009a). HuR recruits let-7/RISC to repress c-Myc expression. Genes & Development *23*, 1743.

Kim, J., Inoue, K., Ishii, J., Vanti, W.B., Voronov, S.V., Murchison, E., Hannon, G., and Abeliovich, A. (2007). A MicroRNA feedback circuit in midbrain dopamine neurons. Science *317*, 1220-1224.

Kim, V.N., Han, J., and Siomi, M.C. (2009b). Biogenesis of small RNAs in animals. Nat Rev Mol Cell Biol *10*, 126-139.

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. Genes Dev *18*, 1165-1178.

Kirino, Y., and Mourelatos, Z. (2008). Site-specific crosslinking of human microRNPs to RNA targets. RNA *14*, 2254-2259.

Kota, J., Chivukula, R.R., O'Donnell, K.A., Wentzel, E.A., Montgomery, C.L., Hwang, H.W., Chang, T.C., Vivekanandan, P., Torbenson, M., Clark, K.R., *et al.* (2009). Therapeutic microRNA delivery suppresses tumorigenesis in a murine liver cancer model. Cell *137*, 1005-1017.

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.

Laferriere, A., Gautheret, D., and Cedergren, R. (1994). An RNA pattern matching program with enhanced performance and portability. Comput Appl Biosci *10*, 211-212.

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

Lal, A. (2008). p16INK4a translation suppressed by miR-24. PLoS ONE *3*, e1864.

Lal, A., Navarro, F., Maher, C.A., Maliszewski, L.E., Yan, N., O'Day, E., Chowdhury, D., Dykxhoorn, D.M., Tsai, P., Hofmann, O., *et al.* (2009). miR-24 Inhibits cell proliferation by targeting E2F2, MYC, and other cell-cycle genes via binding to "seedless" 3'UTR microRNA recognition elements. Mol Cell *35*, 610-625.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., *et al.* (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860-921.

Landgraf, P. (2007). A mammalian microRNA expression atlas based on small RNA library sequencing. Cell *129*, 1401-1414.

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., 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., 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.

Lerner, M.R., Boyle, J.A., Mount, S.M., Wolin, S.L., and Steitz, J.A. (1980). Are snRNPs involved in splicing? Nature *283*, 220-224.

Leucht, C., Stigloher, C., Wizenmann, A., Klafke, R., Folchert, A., and Bally-Cuif, L. (2008). MicroRNA-9 directs late organizer activity of the midbrain-hindbrain boundary. Nat Neurosci *11*, 641-648.

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.

Lewis, H.A., Musunuru, K., Jensen, K.B., Edo, C., Chen, H., Darnell, R.B., and Burley, S.K. (2000). Sequence-specific RNA binding by a Nova KH domain: implications for paraneoplastic disease and the fragile X syndrome. Cell *100*, 323-332.

Lewis, M.A., Quint, E., Glazier, A.M., Fuchs, H., De Angelis, M.H., Langford, C., van Dongen, S., Abreu-Goodger, C., Piipari, M., Redshaw, N., *et al.* (2009). An

ENU-induced mutation of miR-96 associated with progressive hearing loss in mice. Nat Genet *41*, 614-618.

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

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.

Lingel, A., Simon, B., Izaurralde, E., and Sattler, M. (2003). Structure and nucleic-acid binding of the Drosophila Argonaute 2 PAZ domain. Nature *426*, 465-469.

Liu, J., Carmell, M.A., Rivas, F.V., Marsden, C.G., Thomson, J.M., Song, J.J., Hammond, S.M., Joshua-Tor, L., and Hannon, G.J. (2004). Argonaute2 is the catalytic engine of mammalian RNAi. Science *305*, 1437-1441.

Liu, J., Valencia-Sanchez, M.A., Hannon, G.J., and Parker, R. (2005). MicroRNA-dependent localization of targeted mRNAs to mammalian P-bodies. Nat Cell Biol *7*, 719-723.

Lu, J. (2005). MicroRNA expression profiles classify human cancers. Nature *435*, 834-838.

Lu, R., Wang, H., Liang, Z., Ku, L., O'Donnell W, T., Li, W., Warren, S.T., and Feng, Y. (2004). The fragile X protein controls microtubule-associated protein 1B translation and microtubule stability in brain neuron development. Proc Natl Acad Sci U S A *101*, 15201-15206.

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

Lunde, B.M., Moore, C., and Varani, G. (2007). RNA-binding proteins: modular design for efficient function. Nat Rev Mol Cell Biol *8*, 479-490.

Luque, F.A., Furneaux, H.M., Ferziger, R., Rosenblum, M.K., Wray, S.H., Schold, S.C., Jr., Glantz, M.J., Jaeckle, K.A., Biran, H., Lesser, M., *et al.* (1991). Anti-Ri: an antibody associated with paraneoplastic opsoclonus and breast cancer. Ann Neurol *29*, 241-251.

Lytle, J.R., Yario, T.A., and Steitz, J.A. (2007). Target mRNAs are repressed as efficiently by microRNA-binding sites in the 5' UTR as in the 3' UTR. Proc Natl Acad Sci U S A *104*, 9667-9672.

Ma, J.B., Ye, K., and Patel, D.J. (2004). Structural basis for overhang-specific small interfering RNA recognition by the PAZ domain. Nature *429*, 318-322.

Macke, T.J., Ecker, D.J., Gutell, R.R., Gautheret, D., Case, D.A., and Sampath, R. (2001). RNAMotif, an RNA secondary structure definition and search algorithm. Nucleic Acids Res *29*, 4724-4735.

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.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.J., Chen, Z., *et al.* (2005). Genome sequencing in microfabricated high-density picolitre reactors. Nature *437*, 376-380.

Marioni, J.C., Mason, C.E., Mane, S.M., Stephens, M., and Gilad, Y. (2008). RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res *18*, 1509-1517.

Maroney, P.A., Yu, Y., Fisher, J., and Nilsen, T.W. (2006). Evidence that microRNAs are associated with translating messenger RNAs in human cells. Nature structural & molecular biology *13*, 1102-1107.

Mathews, D.H., Sabina, J., Zuker, M., and Turner, D.H. (1999). Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. J Mol Biol *288*, 911-940.

Mathonnet, G., Fabian, M.R., Svitkin, Y.V., Parsyan, A., Huck, L., Murata, T., Biffo, S., Merrick, W.C., Darzynkiewicz, E., and Pillai, R.S. (2007). MicroRNA inhibition of translation initiation in vitro by targeting the cap-binding complex eIF4F. Science *317*, 1764.

Matranga, C., Tomari, Y., Shin, C., Bartel, D.P., and Zamore, P.D. (2005). Passenger-strand cleavage facilitates assembly of siRNA into Ago2-containing RNAi enzyme complexes. Cell *123*, 607-620.

Mattick, J.S. (2003). Challenging the dogma: the hidden layer of non-protein-coding RNAs in complex organisms. Bioessays *25*, 930-939.

McCaskill, J.S. (1990). The equilibrium partition function and base pair binding probabilities for RNA secondary structure. Biopolymers *29*, 1105-1119.

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.

Mencia, A., Modamio-Hoybjor, S., Redshaw, N., Morin, M., Mayo-Merino, F., Olavarrieta, L., Aguirre, L.A., del Castillo, I., Steel, K.P., Dalmay, T., *et al.* (2009). Mutations in the seed region of human miR-96 are responsible for nonsyndromic progressive hearing loss. Nat Genet *41*, 609-613.

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

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.

Modrek, B., and Lee, C. (2002). A genomic view of alternative splicing. Nat Genet *30*, 13-19.

Mourelatos, Z. (2008). Small RNAs: The seeds of silence. Nature *455*, 44-45.

Murchison, E.P., and Hannon, G.J. (2004). miRNAs on the move: miRNA biogenesis and the RNAi machinery. Curr Opin Cell Biol *16*, 223-229.

Nelson, P.T., De Planell-Saguer, M., Lamprinaki, S., Kiriakidou, M., Zhang, P., O'Doherty, U., and Mourelatos, Z. (2007). A novel monoclonal antibody against human Argonaute proteins reveals unexpected characteristics of miRNAs in human blood cells. RNA *13*, 1787-1792.

Nimchinsky, E.A., Oberlander, A.M., and Svoboda, K. (2001). Abnormal development of dendritic spines in FMR1 knock-out mice. J Neurosci *21*, 5139-5146.

Nottrott, S., Simard, M.J., and Richter, J.D. (2006). Human let-7a miRNA blocks protein production on actively translating polyribosomes. Nature structural & molecular biology *13*, 1108-1114.

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

Pan, Q., Shai, O., Lee, L.J., Frey, B.J., and Blencowe, B.J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet *40*, 1413-1415.

Petersen, C.P., Bordeleau, M.E., Pelletier, J., and Sharp, P.A. (2006). Short RNAs repress translation after initiation in mammalian cells. Molecular cell *21*, 533-542.

Pillai, R.S., Bhattacharyya, S.N., Artus, C.G., Zoller, T., Cougot, N., Basyuk, E., Bertrand, E., and Filipowicz, W. (2005). Inhibition of translational initiation by Let-7 MicroRNA in human cells. Science *309*, 1573-1576.

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.

Rajewsky, N. (2006). microRNA target predictions in animals. Nature Genet *38*, S8-S13.

Rajewsky, N., and Socci, N.D. (2004). Computational identification of microRNA targets. Dev Biol *267*, 529-535.

Richter, J.D. (2008). Breaking the code of polyadenylation-induced translation. Cell *132*, 335-337.

Rodriguez, A.J., Czaplinski, K., Condeelis, J.S., and Singer, R.H. (2008). Mechanisms and cellular roles of local protein synthesis in mammalian cells. Curr Opin Cell Biol *20*, 144-149.

Rosa, A., Spagnoli, F.M., and Brivanlou, A.H. (2009). The miR-430/427/302 family controls mesendodermal fate specification via species-specific target selection. Dev Cell *16*, 517-527.

Ruggiu, M., Herbst, R., Kim, N., Jevsek, M., Fak, J.J., Mann, M.A., Fischbach, G., Burden, S.J., and Darnell, R.B. (2009). Rescuing Z+ agrin splicing in Nova null mice restores synapse formation and unmasks a physiologic defect in motor neuron firing. Proc Natl Acad Sci U S A *106*, 3513-3518.

Sakakibara, Y., Brown, M., Hughey, R., Mian, I.S., Sjolander, K., Underwood, R.C., and Haussler, D. (1994). Stochastic context-free grammars for tRNA modeling. Nucleic Acids Res *22*, 5112-5120.

Schaefer, A., O'Carroll, D., Tan, C.L., Hillman, D., Sugimori, M., Llinas, R., and Greengard, P. (2007). Cerebellar neurodegeneration in the absence of microRNAs. J Exp Med *204*, 1553-1558.

Schena, M., Shalon, D., Davis, R.W., and Brown, P.O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science *270*, 467-470.

Schratt, G.M., Tuebing, F., Nigh, E.A., Kane, C.G., Sabatini, M.E., Kiebler, M., and Greenberg, M.E. (2006). A brain-specific microRNA regulates dendritic spine development. Nature *439*, 283-289.

Seitz, H., Royo, H., Bortolin, M.L., Lin, S.P., Ferguson-Smith, A.C., and Cavaille, J. (2004). A large imprinted microRNA gene cluster at the mouse Dlk1-Gtl2 domain. Genome Res *14*, 1741-1748.

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.

Sen, G.L., and Blau, H.M. (2005). Argonaute 2/RISC resides in sites of mammalian mRNA decay known as cytoplasmic bodies. Nat Cell Biol *7*, 633-636.

Sharp, P.A. (2009). The centrality of RNA. Cell *136*, 577-580.

Shen, W.F., Hu, Y.L., Uttarwar, L., Passegue, E., and Largman, C. (2008). MicroRNA-126 regulates HOXA9 by binding to the homeobox. Mol Cell Biol *28*, 4609-4619.

Shendure, J., Porreca, G.J., Reppas, N.B., Lin, X., McCutcheon, J.P., Rosenbaum, A.M., Wang, M.D., Zhang, K., Mitra, R.D., and Church, G.M. (2005). Accurate multiplex polony sequencing of an evolved bacterial genome. Science *309*, 1728-1732.

Shi, R., and Chiang, V.L. (2005). Facile means for quantifying microRNA expression by real-time PCR. Biotechniques *39*, 519-525.

Shibata, M., Kurokawa, D., Nakao, H., Ohmura, T., and Aizawa, S. (2008). MicroRNA-9 modulates Cajal-Retzius cell differentiation by suppressing Foxg1 expression in mouse medial pallium. J Neurosci *28*, 10415-10421.

Shyu, A.B., Wilkinson, M.F., and van Hoof, A. (2008). Messenger RNA regulation: to translate or to degrade. The EMBO journal *27*, 471-481.

Siegel, G., Obernosterer, G., Fiore, R., Oehmen, M., Bicker, S., Christensen, M., Khudayberdiev, S., Leuschner, P.F., Busch, C.J., Kane, C., *et al.* (2009). A functional screen implicates microRNA-138-dependent regulation of the depalmitoylation enzyme APT1 in dendritic spine morphogenesis. Nat Cell Biol *11*, 705-716.

Song, J.J., Smith, S.K., Hannon, G.J., and Joshua-Tor, L. (2004). Crystal structure of Argonaute and its implications for RISC slicer activity. Science *305*, 1434-1437.

Stefani, G., Fraser, C.E., Darnell, J.C., and Darnell, R.B. (2004). Fragile X mental retardation protein is associated with translating polyribosomes in neuronal cells. J Neurosci *24*, 7272-7276.

Tavazoie, S.F., Alarcon, C., Oskarsson, T., Padua, D., Wang, Q., Bos, P.D., Gerald, W.L., and Massague, J. (2008). Endogenous human microRNAs that suppress breast cancer metastasis. Nature *451*, 147-152.

Tay, Y., Zhang, J., Thomson, A.M., Lim, B., and Rigoutsos, I. (2008). MicroRNAs to Nanog, Oct4 and Sox2 coding regions modulate embryonic stem cell differentiation. Nature *455*, 1124-1128.

Thomas, J.W., Touchman, J.W., Blakesley, R.W., Bouffard, G.G., Beckstrom-Sternberg, S.M., Margulies, E.H., Blanchette, M., Siepel, A.C., Thomas, P.J., McDowell, J.C., *et al.* (2003). Comparative analyses of multi-species sequences from targeted genomic regions. Nature *424*, 788-793.

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

Trabucchi, M., Briata, P., Garcia-Mayoral, M., Haase, A.D., Filipowicz, W., Ramos, A., Gherzi, R., and Rosenfeld, M.G. (2009). The RNA-binding protein KSRP promotes the biogenesis of a subset of microRNAs. Nature *459*, 1010-1014.

Ule, J., and Darnell, R.B. (2006). RNA binding proteins and the regulation of neuronal synaptic plasticity. Curr Opin Neurobiol *16*, 102-110.

Ule, J., Jensen, K., Mele, A., and Darnell, R.B. (2005a). CLIP: a method for identifying protein-RNA interaction sites in living cells. Methods *37*, 376-386.

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.

Ule, J., Stefani, G., Mele, A., Ruggiu, M., Wang, X., Taneri, B., Gaasterland, T., Blencowe, B.J., and Darnell, R.B. (2006). An RNA map predicting Nova-dependent splicing regulation. Nature *444*, 580-586.

Ule, J., Ule, A., Spencer, J., Williams, A., Hu, J.S., Cline, M., Wang, H., Clark, T., Fraser, C., Ruggiu, M., *et al.* (2005b). Nova regulates brain-specific splicing to shape the synapse. Nat Genet *37*, 844-852.

van Rooij, E., Liu, N., and Olson, E.N. (2008a). MicroRNAs flex their muscles. Trends Genet *24*, 159-166.

van Rooij, E., Marshall, W.S., and Olson, E.N. (2008b). Toward microRNA-based therapeutics for heart disease: the sense in antisense. Circ Res *103*, 919-928.

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

Vella, M.C., Choi, E.Y., Lin, S.Y., Reinert, K., and Slack, F.J. (2004). The C. elegans microRNA let-7 binds to imperfect let-7 complementary sites from the lin-41 3[prime]UTR. Genes Dev *18*, 132-137.

Ventura, A. (2008). Targeted deletion reveals essential and overlapping functions of the miR-17-92 family of miRNA clusters. Cell *132*, 875-886.

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. Genes Dev *21*, 1857-1862.

Wang, E.T. (2008). Alternative isoform regulation in human tissue transcriptomes. Nature *456*, 470-476.

Wang, X. (2006). Systematic identification of microRNA functions by combining target prediction and expression profiling. Nucleic Acids Res *34*, 1646-1652.

Wang, Y., Juranek, S., Li, H., Sheng, G., Tuschl, T., and Patel, D.J. (2008a). Structure of an argonaute silencing complex with a seed-containing guide DNA and target RNA duplex. Nature *456*, 921-926.

Wang, Y., Sheng, G., Juranek, S., Tuschl, T., and Patel, D.J. (2008b). Structure of the guide-strand-containing argonaute silencing complex. Nature *456*, 209-213.

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet *10*, 57-63.

Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., *et al.* (2002). Initial sequencing and comparative analysis of the mouse genome. Nature *420*, 520-562.

Wheeler, T.M., Lueck, J.D., Swanson, M.S., Dirksen, R.T., and Thornton, C.A. (2007). Correction of ClC-1 splicing eliminates chloride channelopathy and myotonia in mouse models of myotonic dystrophy. J Clin Invest *117*, 3952-3957.

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

Wu, L., Fan, J., and Belasco, J.G. (2006). MicroRNAs direct rapid deadenylation of mRNA. Proceedings of the National Academy of Sciences of the United States of America *103*, 4034.

Xiao, C., Calado, D.P., Galler, G., Thai, T.H., Patterson, H.C., Wang, J., Rajewsky, N., Bender, T.P., and Rajewsky, K. (2007). MiR-150 controls B cell differentiation by targeting the transcription factor c-Myb. Cell *131*, 146-159.

Xie, X., Lu, J., Kulbokas, E.J., Golub, T.R., Mootha, V., Lindblad-Toh, K., Lander, E.S., and Kellis, M. (2005). Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. Nature *434*, 338-345.

Yang, Y.Y., Yin, G.L., and Darnell, R.B. (1998). The neuronal RNA-binding protein Nova-2 is implicated as the autoantigen targeted in POMA patients with dementia. Proc Natl Acad Sci U S A *95*, 13254-13259.

Yekta, S., Shih, I., and Bartel, D.P. (2004). MicroRNA-directed cleavage of HOXB8 mRNA (American Association for the Advancement of Science), pp. 594-596.

Yeo, G.W. (2009). An RNA code for the FOX2 splicing regulator revealed by mapping RNA-protein interactions in stem cells. Nature Struct Mol Biol *16*, 130-137.

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

Yu, J.Y., Chung, K.H., Deo, M., Thompson, R.C., and Turner, D.L. (2008). MicroRNA miR-124 regulates neurite outgrowth during neuronal differentiation. Exp Cell Res *314*, 2618-2633.

Zamore, P.D., Tuschl, T., Sharp, P.A., and Bartel, D.P. (2000). RNAi: double-stranded RNA directs the ATP-dependent cleavage of mRNA at 21 to 23 nucleotide intervals. Cell *101*, 25-33.

Zeeberg, B.R., Feng, W., Wang, G., Wang, M.D., Fojo, A.T., Sunshine, M., Narasimhan, S., Kane, D.W., Reinhold, W.C., Lababidi, S.*, et al.* (2003). GoMiner: a resource for biological interpretation of genomic and proteomic data. Genome Biol *4*, R28.

Zhang, H.L., Eom, T., Oleynikov, Y., Shenoy, S.M., Liebelt, D.A., Dictenberg, J.B., Singer, R.H., and Bassell, G.J. (2001a). Neurotrophin-induced transport of a beta-actin mRNP complex increases beta-actin levels and stimulates growth cone motility. Neuron *31*, 261-275.

Zhang, L. (2007). Systematic identification of C. elegans miRISC proteins, miRNAs, and mRNA targets by their interactions with GW182 proteins AIN-1 and AIN-2. Mol Cell *28*, 598-613.

Zhang, Y.Q., Bailey, A.M., Matthies, H.J., Renden, R.B., Smith, M.A., Speese, S.D., Rubin, G.M., and Broadie, K. (2001b). Drosophila fragile X-related gene regulates the MAP1B homolog Futsch to control synaptic structure and function. Cell *107*, 591-603.

Zhao, Y., Ransom, J.F., Li, A., Vedantham, V., von Drehle, M., Muth, A.N., Tsuchihashi, T., McManus, M.T., Schwartz, R.J., and Srivastava, D. (2007). Dysregulation of cardiogenesis, cardiac conduction, and cell cycle in mice lacking miRNA-1-2. Cell *129*, 303-317.

Zuker, M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res *31*, 3406-3415.

Zuker, M., and Stiegler, P. (1981). Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucleic Acids Res *9*, 133-148.