UC San Diego

UC San Diego Electronic Theses and Dissertations

Title

High-throughput computation to uncover novel mechanisms of RNA regulation

Permalink

https://escholarship.org/uc/item/2fk2n12x

Author

Lovci, Michael Thomas

Publication Date

2014

Supplemental Material

https://escholarship.org/uc/item/2fk2n12x#supplemental

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA, SAN DIEGO

High-throughput computation to uncover novel mechanisms of RNA regulation

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

in

Biomedical Sciences

by

Michael T. Lovci

Committee in charge:

Professor Gene W. Yeo, Chair
Professor Neil C. Chi
Professor Lawrence S. B. Goldstein
Professor Alysson R. Muotri
Professor Amy E. Pasquinelli

2014

©

Michael T. Lovci, 2014

All rights reserved.

The Dissertation of Michael T. Lovci is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Chair

University of California, San Diego

2014

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SUPPLEMENTAL FILES

Appendix 1: miRNAs detected by ALG-1 CLIP-seq

Appendix 2: RBFOX CLIP-seq library statistics

Appendix 3: Gene ontology results

Appendix 4: Gene expression analysis

Appendix 5: Splicing analysis

ACKNOWLEDGEMENTS

I would like to acknowledge Professor Gene W. Yeo and my thesis committee for their guidance. To all past and present members of the Yeo Lab, thank you for sharing your thoughts and science with me. To my parents, friends and family, thank you for your support, there is no way I could have done this without you.

Thank you also to Melissa Wilbert and Stephanie Huelga for their help decoding the intricacies of Microsoft Word. Their help was essential in the preparation of this dissertation.

Chapter 2, in full, is an adaptation of material that appears in "RNA-seq Analysis of Gene Expression and Alternative Splicing by Double-Random Priming Stratagy" by Lovci MT, Li HR, Fu XD, and Yeo GW, as published in *Methods in Mol Biol*. in 2011. Parts of this work also appear in Li H, Lovci MT, Kwon Y-S, Rosenfeld MG, Fu X-D, Yeo GW. Determination of tag density required for digital transcriptome analysis: Application to an androgen-sensitive prostate cancer model. *PNAS*. 2008 Dec. *MTL and HL contributed equally to this work.

Chapter 3, in full, is an adaptation of material that appears in "Comprehensive discovery of endogenous Argonaute binding sites in Caenorhabditis elegans" by Zisoulis DG, Lovci MT, Wilbert ML, Hutt KR, Liang YL, Pasquinelli AE, Yeo GW, as published in *Nat Struct Mol Biol*. in Jan. 2010. *MTL and DGZ contributed equally to this work.

Chapter 4, in full, is an adaptation of material that appears in "Rbfox proteins regulate alternative mRNA splicing through evolutionarily conserved RNA bridges" by Michael T Lovci, Dana Ghanem, Henry Marr, Justin Arnold, Sherry Gee, Marilyn Parra, Tiffany Y Liang, Thomas Stark, Lauren T Gehman, Shawn Hoon, Katlin Massirer, Gabriel A. Pratt, Douglas L Black, Joe Gray, John G Conboy and Gene W Yeo as published in *Nat Struct Mol Biol* in Dec. 2013.

VITA

Education

Ph.D. Biomedical Sciences. University of California, San Diego. September 2014.

B.S. General Biology. University of California, San Diego. June 2009.

Honors & Awards

The Milton H. Saier, Sr. Memorial Award for Outstanding Original Research Achievement in Computational Biology; 2009

National Science Foundation GK12 Fellowship (UCSD Socrates); 2012-2013

ARCS Fellowship; 2013

Professional Experience

Graduate student, Biomedical Sciences Program

University of California, San Diego

Principal Investigator: Gene Yeo, PhD

September 2009 - September 2014

Laboratory assistant, Department of Cellular and Molecular Medicine.

University of California, San Diego

Principal Investigator: Gene Yeo, PhD

October 2008 - September 2009

Laboratory assistant. Laboratory of Genetics.

Salk Institute for Biological Sciences

Principal Investigator: Fred Gage, PhD

September 2005 - September 2008

Student outreach counselor, TRIO Outreach Program, UC San Diego.

September 2005 - June 2006

Publications

1. Rbfox proteins regulate alternative mRNA splicing through evolutionarily conserved RNA bridges. Lovci MT*, Ghanem D, Marr H, Arnold J, Gee S, Parra M, Liang TY, Stark TJ, Gehman LT, Hoon S, Massirer KB, Pratt GA, Black DL, Gray JW, Conboy JG, Yeo GW. Nature Structural and Molecular Biology. 2013. Citation count: 2

2. Functional Genomic Analysis of the let-7 Regulatory Network in Caenorhabditis elegans. Hunter SE, Finnegan EF, Zisoulis DG, Lovci MT, Melnik-Martinez KV, et al. PLoS Genetics. 2013. Citation count:2

3. LIN28 binds messenger RNAs at GGAGA motifs and regulates splicing factor abundance. Wilbert MW, Huelga SC, Kapeli K, Stark TJ, Liang TY, Chen SX, Yan BY, Nathanson JL, Hutt KR, Lovci MT, Kazan H, Vu AQ, Massirer KB, Morris Q, Hoon S, Yeo GW. Molecular Cell. 2012. Citation count: 24

4. RNA-seq Analysis of Gene Expression and Alternative Splicing by Double-Random Priming Strategy. Lovci MT*, Li HR, Fu XD, Yeo GW. Methods in Molecular Biology. 2011. Citation count: 3

5. Comprehensive discovery of endogenous Argonaute binding sites in Caenorhabditis elegans. Zisoulis DG, Lovci MT*, Wilbert ML, Hutt KR, Liang YL, Pasquinelli AE, Yeo GW. Nature Structural and Molecular Biology. 2010. Citation count: 102

6. L1 Retrotransposition in Human Neural Progenitor Cells. Coufal NG, Garcia-Perez JL, Peng GE, Yeo GW, Mu Y, Lovci MT, Maria M, O'Shea KS, Moran JV, Gage, FH. Nature. 2009. Citation count: 202

7. Deep sequencing identifies new and regulated microRNAs in Schmidtea mediterranea. Lu Y-C, Smielewska M, Palakodeti D, Lovci MT, Aigner S, Yeo GW, Graveley BR. RNA. 2009. Citation count: 27

8. Determination of tag density required for digital transcriptome analysis: Application to an androgen-sensitive prostate cancer model. Li H, Lovci MT*, Kwon Y-S, Rosenfeld MG, Fu X-D, Yeo GW. PNAS. 2008. Citation count: 53

In chronological order. * indicates first-author or co-first author. Citation counts updated May 6, 2014.

Presentations & Posters

RBFOX1 and RBFOX2 CLIP-seq reveals that RBFOX proteins regualte alternative splicing from binding sites deep in introns via RNA-RNA interactions which bridge bound RNA-binding proteins from exon-distal to exon-proximal regions

Talk - Biomedical Sciences Lunch Talk UC, San Diego June 5, 2013

Talk - UCSD RNA club UC, San Diego Apr 20, 2012

Talk - SCRM Neurodegeneration Seminar UC, San Diego May 7, 2012

Poster - RNA Society Meeting UM, Ann Arbor June 30, 2012

Comprehensive Discovery of Endogenous Argonaute Binding Sites at Nucleotide-level Resolution in Animals

Talk - CMM In-House Seminar UC, San Diego Jun 8, 2009

Talk - RNA Society Meeting UW, Madison May 30, 2009

Talk - Undergraduate Research Symposium UC, San Diego May 15, 2009

Service & Teaching

• I provided assistance and direction for "hands-on" exercises in the graduate course "Quantitative Methods in Genetics", 2008-2011. I developed curriculum and guided students through RNA-seq analyses.

• I developed course curriculum and led classes for the Salk Institute's middle school outreach program.

• I assisted my PI in assessing several publications sent to him for peer-review

• As part of the NSF GK12 Socrates Fellowship I designed inquiry-based curriculum for high-school juniors to explore concepts of molecular biology and bioinformatics.

• I provided mentorship for an undergraduate bioinformatics project.

• I gave guidance and mentorship to incoming graduate students and technicians in bioinformatics practices.

• I designed interactive activities for the public at the San Diego Science Festival to demonstrate stem cell biology.

Laboratory & Computational Skills

Programming Languages

Perl, Python, C/C++, R, MATLAB, bash, Scala, Javascript

Bioinformatics Tools

RNA-STAR, GSNAP, SAMtools, parallel processing with SGE/PBS/Hadoop, bowtie, BLAST, UCSC Genome Browser, BLAT, Multiz, Phastcons, BEDTools and others

Laboratory Techniques

PCR, DNA/RNA Isolation, western blot, molecular cloning, northern blot, sterile technique, somatic cell reprogramming, neural differentiation of pluripotent stem cells.

General Software Experience

Microsoft Office, iWork, Adobe Creative Suite.

ABSTRACT OF THE DISSERTATION

High-throughput computation to uncover novel mechanisms of RNA regulation

by

Michael T. Lovci

Doctor of Philosophy in Biomedical Sciences

University of California, San Diego, 2014

Professor Gene W. Yeo, Chair

My thesis work is aimed at integrating high-throughput biochemical readouts to understand the effect that particular RNA-binding proteins have on their targets' metabolism. I studied several RNA-binding proteins in diverse model organisms. Along

the way, I have gained insight into the characteristics of target sites for two families of proteins: Argonaute and RBFOX proteins and learned about novel mechanisms they use to control their targets' fate.

In this dissertation, I will present three of my articles that summarize the methods and content of my work. In the first chapter I give an overview of methods I used to quantify the ideal depth of coverage for sequencing experiments in order to sufficiently capture a desired level of complexity. This included the development of novel computational approaches to quantify RNA abundance and splicing with sequencing tools. This was applied to a human model of prostate cancer, LNCaP cells upon stimulation with an androgen compound.

To closely examine mechanisms of miRNA regulation, we generated genome wide maps of the Argonaute protein ALG-1's binding in *C. elegans.* We found that there exists a large potential for non-canonical associations of ALG-1/miRNA complexes with their targets. Among the surprises we encountered, ALG-1 binds in coding exons but does not seem to repress gene expression when bound there, and we also found an auto-regulatory repression by ALG-1 on miRNA pathway components.

Finally, I turned my attention to the RBFOX family of proteins that is known for their role in RNA splicing regulation. In the work presented in chapter 4, we elucidated a new molecular mechanism whereby RBFOX proteins can regulate RNA splicing from very distant sites. These RNA-bridges, as we have called them because they link RNA regulators with regulated sites via RNA structures, appear to be a common feature of alternatively spliced exons and the regulation of RNA structures like this may be important for dictating splicing outcomes. The application of this knowledge that distant

binding sites are functional and that they are mediated by RNA structures is immediately relevant to the design of novel therapeutics for diseases that arise from defects in RNA splicing.

To put the goals of my thesis broadly, I approach two questions: what are the mechanisms of RNA-binding protein targeting, and what are the effects of RNA-binding proteins on their direct targets?

CHAPTER 1 – INTRODUCTION

The abundance and relative quantities of biomolecules a cell uses at a given moment determines its ability to perform specific metabolic processes and the way that it responds to its environment. Research of the past decade established that predictable differences in gene expression exist among tissues and through life. The combined effect of differential usage of these genes controls the complement of biomolecules that comprise a cell. Mechanisms have evolved that tightly regulate the output, activity and localization of innumerable chemical reactions and macromolecules. In response to this daunting complexity, scientists have intensely studied gene expression by creating maps of the patterns of usage for each gene in order to establish differences that determine tissue identity and intricate sub-structures of complicated organs. Molecular biology is the field devoted to obtaining and dissecting the patterns of macromolecule expression to understand the basic processes that lead to development and disease.

A field born in the 1950s, molecular biologists have at their disposal increasingly accurate assays, developed to identify the presence and interactions between biomolecules. In addition to survey tools, reductionist techniques have been developed to tear apart and re-engineer biological systems for our benefit, including PCR and, more recently, CRISPR. These techniques have been aided by computation and instrumentation that provide measurements that remain sensitive even as they measure orders of magnitude more molecules than was previously possible. As advances in biochemistry techniques enabled annotation of almost all protein-coding regions in the human genome, and a large component of the regulatory portions of the genome, all these data revealed a

Gordian Knot of information that is difficult or impossible to interpret with standard methods or by one person alone.

Collaborative efforts around the world have set out to catalog and share these data in order to advance of our understanding of the content in the genome. This undertaking is no small task, but as technologies emerge to meet this complex system's demands for scale and acuity, so do new tools we to combat new problems. Advances in other industries that examine patterns in large datasets have greatly contributed to a shift toward high-throughput approaches to molecular biology. These data analysis tools, combined with the ever-increasing set of nucleotide sequences from hundreds of species, our understanding of the blueprints that guide the construction of life is constantly becoming more complete. We are beginning to decode the purpose for swaths of supposed 'junk' DNA.

Our current understanding of the genome provides that only a minor proportion, less than two percent, consists of nucleotides that can translate into protein-coding sequence. Most of these sequences are conserved through evolution. Besides protein-coding regions there are highly conserved stretches around and inside genes, these nucleotide sequences demark the sites for interactions with regulatory proteins, DNA-binding proteins and RNA-binding proteins, and can also control the secondary structure of DNA and RNA. In addition to these *cis*-acting regions within RNA transcripts, our genome contains a large number of *trans*-acting non-protein-coding RNA species that serve critical enzymatic roles. Examples of these include micro-RNAs (miRNAs), short guide RNAs that direct nucleolysis, and long non-coding RNAs that can be chemical scaffolds or ribozymes.

In the search for functional elements in the genome, we have also learned that conserved stretches are also not the only functional segments in the genome. In addition to portions that are preserved in form and function through evolution there are also segments of genomes that are highly dynamic between species and even between generations. For example, each animal species contains a unique complement of endogenous retroviral elements that interrupt evolutionarily conserved stretches of genomes and accelerate genome rearrangement over time. The functions of these non-conserved elements and others like them are mostly unknown but emerging functions include roles in regulating RNA splicing, RNA structures and chromosomal instability.

It stands as the task of our time to integrate the knowledge of the content of a genome with biochemical readouts of cell state in order to gain grasp over how information, coded in the genome, ultimately leads to the development of complex organisms. We can use these data to understand what the most important parts of the genome are, and begin to decipher their purpose. More than simple atlases, integrative approaches to these disparate data types have been foundational in forging a new understanding of the mechanisms of development and disease. They reveal combinatorial hierarchies of regulation and lead to the elucidation of integrated disease and developmental pathways. This knowledge is critical for our ability to interject and control these systems for public health and synthetic applications.

CHAPTER 2 - RNA-SEQ ANALYSIS OF GENE EXPRESSION AND ALTERNATIVE SPLICING BY DOUBLE-RANDOM PRIMING STRATEGY

ABSTRACT

Transcriptome analysis by deep sequencing, more commonly known as RNA-seq is becoming the method of choice for gene discovery and quantitative splicing detection. We recently published a double-random priming RNA-seq approach capable of generating strand-specific information (Li et al., 2008). Poly(A)+ RNA from a treated and an untreated sample was utilized to generate RNA-seq libraries that were sequenced on the Illumina GA1 analyzer. Statistical analysis of ~10 million sequence reads generated from both control and treated cells suggests that this tag density is sufficient for quantitative analysis of gene expression. We were also able to detect a large fraction of reads corresponding to annotated alternative exons, with a subset of the reads matching known and detecting new splice junctions. In this chapter, we provide a detailed bench ready protocol for the double-random priming method and provide user-friendly templates for the curve-fitting model described in the paper to estimate the tag density needed for optimal detection of regulated gene expression and alternative splicing.

INTRODUCTION

We have devised a procedure based on double-random priming and solid phase selection to produce libraries for high-throughput sequencing on the Illumina Genome Analyzer. In order to sequence these libraries, P1 and P2 adapter sequences must be

added to the ends of the DNA of interest. In this protocol, double poly(A)-selected RNA is first primed with an oligonucleotide that contains a random octamer and the P1 adapter sequence. This first primer also carries a biotin moiety at the 5′ end, which allows for capture of extended cDNA product on streptavidin beads. A second random primer linked to the other sequencing primer (P2) adapter sequence, is next added to the cDNA bound to the streptavidin-coated magnetic beads. After extensive washes, potential P2 dimers are eliminated and the second random primed products are released from the beads by heat, leaving behind unused P1 primer, P1-extended cDNA, and potential P1 dimers. The released products are PCR-amplified, gel purified to enrich for amplicons in the size range of 100–300 nt, quantified, and subjected to sequencing (from the P1 primer side) on the Illumina/Solexa flowcell. This procedure has the following advantages compared to previous published protocols. First, it provides strand-specific information, as opposed to other methods that convert RNA to cDNA before primer addition. Second, sequencing a short region right after the first random priming reaction avoids cDNA artifacts resulting from extension by the hairpins formed after the first strand synthesis (Perocchi et al., 2007), which may account for artifactual “antisense transcripts” seen in previous large-scale mRNA sequencing and tiling analysis, (Carninci et al., 2005; Cheng et al., 2005). Third, the built-in random primer region retains the molecular memory for originally primed products allowing computational elimination of sequenced reads amplified by PCR, because all PCR products from the same initial amplicon will have identical sequences in the randomized region. This strategy permits the use of PCR amplification without distorting the representation of the transcriptome, a feature critical for quantitative analysis on a limited population of cells.

TRANSCRIPT DATABASES FOR GENE-EXPRESSION AND ALTERNATIVE SPLICING DETECTION

In order to utilize RNA-seq reads to quantitatively measure gene expression, it is imperative to first define our concept of genes. To that end, we have developed detailed annotations of the genome based on publicly available annotations downloaded from the University of California, Santa Cruz (UCSC)(Karolchik et al., 2003). We have also parsed the genome to create alignable sequence databases for the use with data generated from high-throughput sequencing and the purpose of aligning sequencing reads to spliced mRNA transcripts. Basic notes on the acquisition and processing of data like this are outlined here. Please review our previously published work for more detailed information (Yeo et al., 2007a).

Building an aggregate gene model

Genome sequences of human (hg17) and annotation for protein-coding genes were obtained from the UCSC Genome Browser Database. The lists of known human genes (knownGene containing 43,401 entries) and known isoforms (knownIsoforms containing 43,286 entries in 21,397 unique isoform clusters) with annotated exon alignments to human hg17 genomic sequence were processed as follows. Known genes that were mapped to different isoform clusters were discarded. All mRNAs aligned to the human genome that were >300 bases long were clustered together with the known isoforms. For the purposes of measuring differential gene expression, all genes were considered. For the purposes of inferring alternative splicing, genes containing < 3 exons were removed from further consideration. Exons with canonical splice signals (GT-AG,

AT-AC, GC-AG) were retained, resulting in a total of 213,736 exons. Of these, 92% of all exons were constitutive exons, 7% had evidence of exon-skipping, 1% exons were mutually exclusive alternative events, 3% exons had alternative 3′ splice sites, and 2% exons had alternative 5′ splice sites (Figure 1). A total of 2.7 million spliced ESTs were mapped onto the 17,478 high-quality gene clusters to identify alternative splicing. To eliminate redundancies in this analysis, final annotated gene regions were clustered together so that any overlapping portion of these databases was defined by a single genomic position.

Building an exon-junction database

Exons with canonical splice signals (GT-AG, AT-AC, GC-AG) were used to create an exon-junction database (EJDB). For each protein-coding gene, the 35 bases at the 3′ end of each exon were concatenated with the 35 bases at the 5′ end of the downstream exon. This was repeated, joining every exon of a gene to every exon downstream. This approach produced 1,929,065 theoretical splicing junctions. An equal number of "impossible" junctions was generated by joining the 35-base exon junction sequences, in reverse order.

METRICS FOR DIFFERENTIAL GENE EXPRESSION

Alignment

MosaikAligner (Hillier et al., 2008), using a maximum of 2 mismatches over 95% alignment of the tag (34 nt) and a hash size of 15, was used to align reads to the human genome (hg17). However, since the publication of this work there have been several new

alignment algorithms made available that offer other options for this step (such as QPalma (De Bona et al., 2008), Bowtie (Langmead et al., 2009), or RazerS (Weese et al., 2009)). To determine the number of reads contained within proteincoding genes, promoter, and intergenic regions, we arbitrarily defined promoter regions as 3 kb upstream of the transcriptional start site of the gene and intergenic regions as unannotated regions in the genome. Analysis and design of CLIP-sequencing experiments Alignments to our EJDB were also done using the same alignment algorithm and mapping requirements, with the added requirement that reads map at least 4nt across the exon-exon junction.

Evaluation of differential gene expression

Differentially expressed transcripts were identified by enumerating the number of reads that mapped within the spliced mRNA transcript in untreated and hormone-treated cells, using the total number of reads mapped to exons in each condition as a basis for determining significance by the χ_2 statistic. The χ_2 statistic was calculated for genes with $\geq$5 reads in each experimental condition and the value of the χ_2 statistic was computed using a 2x2 square with the reads within a particular gene in both conditions on the top row and the reads not within that gene in both conditions on the bottom row. After the number of reads mapped in each condition and the statistical significance was determined, each gene can be plotted as a scatter plot as in Figure 2.

Detection of alternative splicing

Alternative splicing was detected by using reads mapped across exon junctions. We were able to detect both annotated and novel splice junctions. The type of exon-exon junction (i.e. constitutive or alternative) was determined based on our aggregate gene model (see above). False-discovery rate (FDR) was assessed by mapping reads to a set of "impossible" junctions which were created by reversing the order of exons in the EJDB (for example if exons 1 and 2 of a particular gene are in the EJDB joined 1- > 2, the impossible version of this would be the same exons joined in the reverse order, 2- > 1).

Power curve analysis

To establish the depth of sequencing required to examine several transcriptome features, we devised a method to predict not only the number of reads required to analyze a particular feature, but also the number of features observable at that sequencing depth. Reads were randomly sampled into subsets representing 10%, 20% etc. of the total number of sequence reads available using custom perl scripts. These were aligned as described above and the number of features detected was assessed. To determine the number of sequence reads required to reach a user-defined threshold for saturation, the percentage change in discovering additional features was determined as follows:

$$T(n) = sn$$

$$C(n) = [F(n) - F(n-1)] / [F(n-1)]$$

where T(n) is the number of reads, s is the sampling size (in our case, 2 million reads), n is a constant multiplier, *C(n)* is the empirical change in number of features detected, and *F(n)* is the number of empirical features detected at n. A scatter plot of *C(n)* to *T(n)* was

fitted with a power curve of the form $c(n) = a \times T(n)^b$ and an exponential curve of the form $c(n) = ae^{bT(n),}$ where $c(n)$ is the change estimated by the curve fitting. The equation that had the best fit, indicated by r_2, was used to extrapolate the tag density required to achieve a defined change in the number of features detected. The number of estimated features was calculated by

$$f(n) = \sum_{i=m}^{n} f(i-1) + f(i-1) \times c(i)$$

where m is user-defined (in our case, m = 6). This will compute the predicted number of features observable based on observed change in feature detection, extrapolated from an area in the middle of the curve. Figure 3 depicts one such fitted curve.

These calculations can be done easily using the "Data Analysis" toolpack for Microsoft Excel. An example worksheet that calculates features using data from three independent samplings (labeled X, Y, and Z) can be downloaded from http://gonzo.ucsd.edu/yeolab/PUBLIC_RESOURCES/EXAMPLE.xls

METHODS

Double-random priming method

The bench-ready protocol is described as follows:

Reverse Transcription

1. Add 1 ul Adaptor 1 to 10 pg-5 ug of total RNA, 1 ul dNTP mix, and RNase-free water to 13 ul per reaction.

2. Heat mixture to 65º C for 5 minutes and incubate on ice for at least 1 minute.

3. Add 4 ul RT-buffer.

4. Incubate at 50º C for 30-60 minutes.

5. Add water to 100 ul and inactivate the reaction by heating at 70º C for 15 minutes.

6. To remove the free biotin-labeled oligos, add 500 ul of Qiagen PCR purification buffer before transferring mixture to a Qiagen purification column. Wash the Qiagen column once with binding buffer and twice with wash buffer. Elute with 50 ul of Qiagen elution buffer to a clean tube.

1st Primer Blocking and Random Primer Extension Reaction

1. Transfer the eluted solution to PCR tubes. Add 15 ul terminal transferase buffer, 3ul ddNTP mix and DI water to 150 ul. Add 2 ul terminal transferase enzyme.

Incubate at 37º C for 1 hour.

2. Add EDTA to 20 mM.

3. Add 5 ul beads and incubate at room temperature for 20 minutes. Collect the beads with a magnetic stand and discard the supernatant. Remove the tubes containing beads from the magnetic stand. Wash the beads with 100 ul NaOH solution by pipetting gently up and down onto the beads. Incubate 5 minutes at room temperature.

4. Collect the beads with the magnetic stand and wash with DI water twice removing tubes from magnetic stand between washes.

5. Off the magnetic stand, add 1 ul Adaptor 2 to the beads, 5 ul PCR buffer, 1 μl dNTPs, add DI water to 49 ul. Add 1 ul Taq DNA polymerase (5U).

6. Incubate at 25º C for 1 hour. Heat to 72º C for 30 seconds and then raise the temperature to 75º C for 5 minutes. Add EDTA to 10 mM to stop the polymerization reaction.

7. Collect the beads and wash the beads twice with 150 ul wash buffer, removing tubes from magnetic stand during washes.

8. On the stand, add 20 ul water and heat for 5 minutes at 95º C. Collect eluate from the beads containing the extended DNA.

9. Amplify the extended DNA with PCR using Solexa adaptors 1 and 2 as primers (without poly(T)+ or N(8)).

10. Run the library on an agarose gel and excise the band corresponding to 75-125nt. Gel extract the band to elute DNA library.

11. Quantify DNA using PicoGreen or quantitative PCR prior to sequencing.

Notes:

1. Ensure that the beads do not dry out throughout the protocol.

2. Ensure that the area used to perform experiments with RNA are free of RNAase contaminants.

3. Check the quality of adaptors by running them on an agarose gel (there should be one band) and be sure that they are PAGE-purified.

4. When washing beads on the magnetic stand, it is useful to spin the tubes in the stand to get them to transfer from one side of the tube to the other, the beads tend to stick to the wall of the tube and this makes washes faster and more thorough.

FIGURES

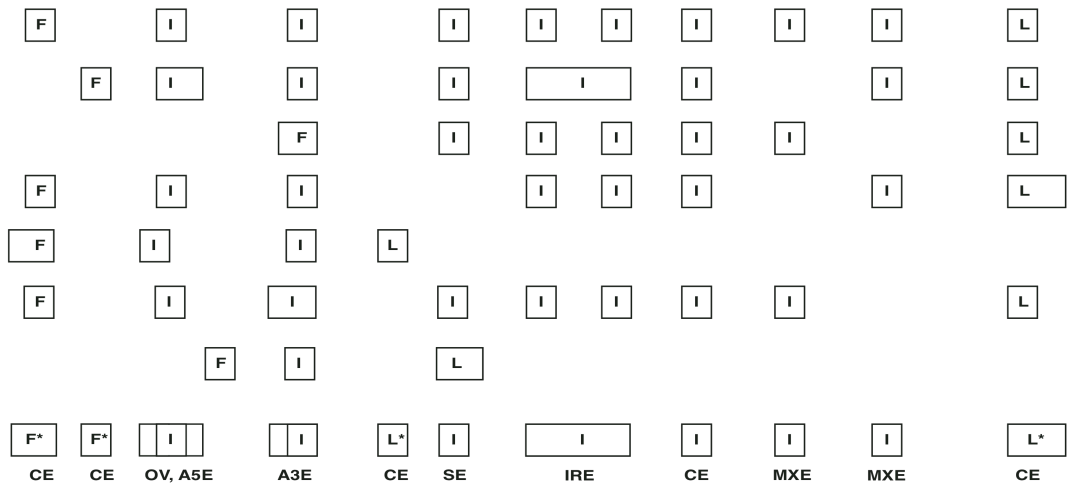


Figure 1. Cartoon depicting construction of an aggregate gene model.

Exons are depicted as boxes labeled as either Internal (I) or First (F) or Last (L). Region classifications are listed on the bottom of the schematic. Classifications of splicing were defined as follows: OV (overlap), SE (skipped exons), A5E/3E (alternative 5'/3' exons), CE (constitutive exons), MXE (mutually skipped exons),IRE (intron retentions).

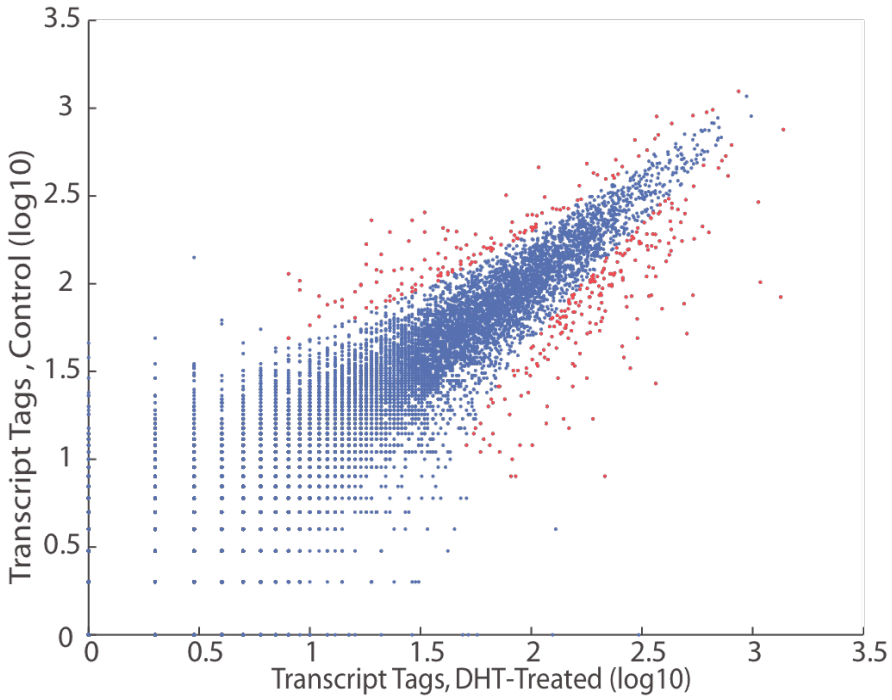


Figure 2. Digital analysis of androgen-regulated gene expression in LNCaP cells.

Scatter plot of gene expression in mock-treated and DHT-induced cells. Differential expressed genes were labeled red based on χ_2 ($P < 0.01$).

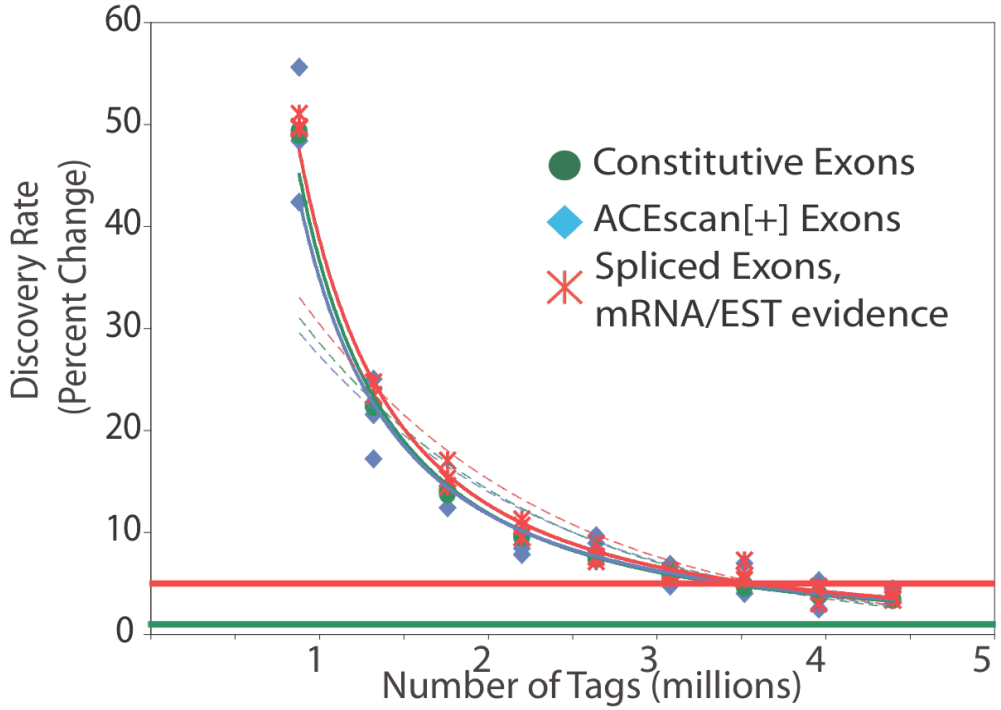


Figure 3. Curve fitting the change in the number of exons and splice junctions detected against increasing tag densities.

Dashed line indicated exponential curve; solid line indicated power curve. Decline in the rate of identifying additional exons as a function of increasing tag density.

ACKNOWLEDGEMENTS

The authors responsible for the content and preparation of this work were Michael T. Lovci, Gene W. Yeo and Xiang-dong Fu. I would like to thank members of the Yeo and Fu labs for critical reading of this manuscript. This manuscript was supported by grants to Gene W. Yeo and Xiang-dong Fu from the US National Institutes of Health (HG004659 and GM084317 and GM052872) for funding this research and the development of this protocol.

CHAPTER 3 - COMPREHENSIVE DISCOVERY OF ENDOGENOUS ARGONAUTE BINDING SITES IN *CAENORHABDITIS ELEGANS*

ABSTRACT

MicroRNAs (miRNAs) regulate gene expression by guiding Argonaute proteins to specific target mRNA sequences. Identification of bona fide miRNA target sites in animals is challenging because of uncertainties regarding the base-pairing requirements between miRNA and target as well as the location of functional binding sites within mRNAs. Here we present the results of a comprehensive strategy aimed at isolating endogenous mRNA target sequences bound by the Argonaute protein ALG-1 in *C. elegans*. Using cross-linking and ALG-1 immunoprecipitation coupled with high-throughput sequencing (CLIP-seq), we identified extensive ALG-1 interactions with specific 3′ untranslated region (UTR) and coding exon sequences and discovered features that distinguish miRNA complex binding sites in 3′ UTRs from those in other genic regions. Furthermore, our analyses revealed a striking enrichment of Argonaute binding sites in genes important for miRNA function, suggesting an autoregulatory role that may confer robustness to the miRNA pathway.

INTRODUCTION

miRNAs function as ~22-nucleotide (nt) RNAs that target messenger RNAs (mRNAs) for degradation or translational repression (Bartel, 2009; Chekulaeva and Filipowicz, 2009). A single miRNA can potentially repress hundreds of genes by binding with partial sequence complementarity to mRNAs (Lim et al., 2005; Sood et al., 2006).

By combinatorial regulation of thousands of genes, the miRNA pathway critically influences many developmental programs as well as cellular homeostasis, the disruption of which leads to human disease. Thus, an outstanding challenge has been to distinguish biologically relevant miRNA-target interactions. To date, identification of miRNA target sites has been dependent largely on computational methods that have limited capability for predicting specific and physiologically relevant targets. Addressing this need, several studies have reported biochemical approaches to isolate targets by immunoprecipitation of miRNA effector complexes containing miRNA–mRNA duplexes (Beitzinger et al., 2007; Easow et al., 2007; Hendrickson et al., 2008; Karginov et al., 2007; Zhang et al., 2007; Zhang et al., 2009). Despite reduction of the search space for miRNA target sites from within all transcribed genes to a subset of immunoprecipitated RNAs, the identification of miRNA binding sequences is still not directly obtained and usually relies on subsequent computational searches for complementary sites within the precipitated transcripts. Here we have narrowed the regions recognized by miRNA effector complexes to approximately 100-nt sequences. The identification and analysis of sequences directly associated with Argonaute protein *in vivo* in *C. elegans* has enabled the discovery of distinct features related to this core component of the miRNA-induced silencing complex (miRISC) as well as its interaction with mRNA target sites.

RESULTS

ALG-1 CLIP-seq in C. elegans identifies known miRNA targets.

As miRNAs guide Argonaute proteins to specific complementary sequences in mRNAs, we applied the CLIP-seq (also referred to as HITS-CLIP) method (Licatalosi et

al., 2008; Sanford et al., 2009; Yeo et al., 2009) to capture and identify the miRNA and target-site sequences bound by the miRNA complex (miRISC) in developing worms. A recent application of this approach in mouse brain resulted in a map of Argonaute binding sites in this tissue (Chi et al., 2009). *C. elegans* offers several advantages for applying the CLIP-seq procedure to detect global Argonaute protein–RNA interactions. A single Argonaute protein, ALG-1, is largely responsible for miRNA function, and viable *alg-1* genetic mutants exist (Grishok et al., 2001). A short but well-established list of miRNA targets expected to be bound by ALG-1 at discrete positions is available (Abbott et al., 2005; Abrahante et al., 2003; Chang et al., 2004; Grosshans et al., 2005; Hayes et al., 2006; Hillier et al., 2009; Johnson et al., 2005; Johnston and Hobert, 2003; Lall et al., 2006; Lee et al., 1993; Lin et al., 2003; Moss et al., 1997; Reinhart et al., 2000; Slack et al., 2000; Vella et al., 2004; Wightman et al., 1993; Yoo and Greenwald, 2005). Of these targets, extensive studies have confirmed that *lin-41* is regulated by *let-7* miRNA during the fourth larval (L4) stage via two clustered sequences, *let-7* complementary sites 1 and 2 (LCS1 and LCS2) (Reinhart et al., 2000; Slack et al., 2000; Vella et al., 2004). We used this example to optimize the CLIP-seq method to detect bona fide ALG-1 binding sites (Figure 4). Synchronized L4-stage wild-type (WT) worms and *alg-1(gk214)* mutants (hereafter referred to as *alg-1*(–)), which lack the anti–ALG-1 antibody epitope sequence, were treated with UV irradiation to stabilize *in vivo* protein-RNA interactions (Figure 4a). A custom antibody specific for the *C. elegans* ALG-1 protein (Figure 4b) was used to enrich for ALG-1 complexes expected to include miRNA and target RNA species. Immunoprecipitated complexes were processed for isolation of sequences protected by ALG-1 protein from nuclease digestion.

We obtained 3,864,848 and 5,127,241 reads from WT and *alg-1*(–) CLIP-seq libraries, respectively, out of which 1,651,523 (42.7%) and 695,895 (13.6%) mapped uniquely to the repeat-masked *C. elegans* genome (Figure 5a). Using MIResque, a microRNA prediction algorithm designed to analyze small-RNA reads obtained from high-throughput sequencing (S. Aigner and G.W. Yeo, unpublished data), 136 previously reported miRNAs, 37 of which represent the 'star' strand, and 1 novel miRNA gene were identified in the WT library (Figure 4). Heterogeneity in the terminal sequences, which primarily consisted of lost nucleotides from the 3′ ends, could be due to the cloning method, but in two cases we found base additions to the 5′ ends that altered the seed sequence. Identification of the distinct pool of miRNAs bound to ALG-1, which included a large number of star sequences, enabled selective analysis of pairing capacity between miRNAs and mRNA sequences associated with ALG-1 at the stage of sample collection.

To correctly assign CLIP-seq reads to authentic transcribed regions, we reannotated the 5′ and 3′ untranslated regions (UTRs) of gene loci using publicly available 36-bp reads obtained from high-throughput sequencing of poly(A)-selected cDNA libraries from the L3 and L4 stages of *C. elegans* larval development (Hillier et al., 2009). Reads that mapped upstream and downstream of currently annotated genes were used to redefine the 5′ and 3′ UTRs. In total, the 5′ and/or 3′ ends of 8,231 genes (40% of genes in the genome) were reannotated by our analysis. The median (average ± s.d.) lengths of bases extended for 5′ and 3′ UTRs are 56 (391 ± 621) and 215 (543 ± 700) nt, respectively. This substantial change in the landscape of *C. elegans* gene predictions was important for defining the genic location of ALG-1 binding sites and for choosing control sequences for computational analyses (see below).

To distinguish authentic and specific ALG-1 binding sites, we developed a new version of our CLIP-cluster identification algorithm (Yeo et al., 2009). Briefly, for each of the three biological replicates of the ALG-1 CLIP-seq experiments (WT or *alg-1*(–)), we first defined 'regions' in each gene by extending the sequencing reads to account for the length of the RNA fragments in our CLIP libraries (Figure 5a). To retain biologically reproducible regions while accounting for the different number of sequenced reads in each replicate library, we weighted regions that overlapped across replicate experiments by the fraction of reads in the region relative to all the reads in that experiment mapping within the gene. Regions that passed our stringent threshold corresponded to being reproducible in at least two of three replicate experiments (see Methods). Reads within accepted regions were further integrated from replicates to form a 'cluster', and clusters containing more reads than statistically expected were kept for further analyses. Finally, clusters that overlapped by at least 25% between WT and *alg-1*(–) were removed as potential sources of false positives, such as reads from highly abundant rRNA and protein-coding genes (see *act-5* gene in Figure 5b).

In total, 5,310 WT and 826 *alg-1*(–) clusters were identified, 4,806 of which were unique to WT (Figure 5a), representing 3,093 genes, approximately one-fifth of the annotated *C. elegans* protein-coding genes expressed at this stage in development. Over half of these genes contained a single cluster (Figure 5c). The CLIP-seq results provided a significantly refined and biologically based dataset for identifying miRNA target sites and studying ALG-1 binding properties. Compared to the entire transcriptome, 3′ UTRs only, or 3′ UTRs of mRNAs from miRISC immunoprecipitates, we greatly reduced the search space for functional regions, by a factor of 47, 20 or 5, respectively (see Methods

for calculation). The tracks for the reannotated gene regions, WT and *alg-1*(–) reads and clusters are available at the UCSC genome browser (http://genome.ucsc.edu) under 'ALG1 CLIP-seq' within the 'Regulation' section in the ce6 genome.

Isolation of sequences containing well-established miRNA target sites demonstrates the sensitivity of the ALG-1 CLIP-seq method. Extensive genetic and reporter gene experiments have pointed to LCS1 and LCS2 in the *lin-41* 3′ UTR as critical sequences for miRNA regulation of this gene (Reinhart et al., 2000; Slack et al., 2000; Vella et al., 2004). Our ALG-1 CLIP-seq results identified a series of reads forming a significant cluster that maps directly on top of the closely spaced LCS1-LCS2 region (Figure 6a). Notably, regulation of *lin-41* by *let-7* miRNA results in substantial mRNA degradation (Bagga et al., 2005). Thus, the detection of *lin-41* by ALG-1 CLIP-seq demonstrates the sensitivity of this method for detecting miRNA targets regardless of regulatory mechanism. The first discovered miRNA target, *lin-14*, is regulated by *lin-4* miRNA via multiple 3′ UTR complementary elements (LCEs)(Lee et al., 1993; Wightman et al., 1993). We identified three significant clusters that encompass the proposed LCEs 1–3, 5, and 6–7, respectively (Figure 6b). Another cluster, toward the end of the *lin-14* 3′ UTR, is consistent with evidence that this gene is also regulated by other miRNAs (Chendrimada et al., 2007; Reinhart et al., 2000). Multiple *let-7* binding sites have been predicted to mediate regulation of *hbl-1* and *daf-12* (Abrahante et al., 2003; Grossman et al., 2005; Lin et al., 2003), and clusters cover a select few of the LCSs in the 3′ UTRs of these genes (Figure 6c,d). Thus, ALG-1 CLIP-seq provides direct biochemical evidence for predicted miRNA target sites and reveals regions of greater relative occupancy by miRISC within a regulated 3′ UTR.

Of 13 well-established miRNA target genes in *C. elegans*, all but 3 were found to contain at least one significant 3′-UTR cluster (Table 1). Moreover, the clusters include the cognate miRNA target site for 9 of these 10 genes. The majority of these miRNA target genes were also found to be enriched in ALG-1 interacting proteins 1 and 2 (AIN-1 and AIN-2, members of the GW182 family of proteins) immunoprecipitation experiments (Zhang et al., 2007). Beyond showing miRISC association with specific endogenous mRNAs, the ALG-1 CLIP-seq dataset contributes nucleotide-level resolution of the actual target region (Table 1 and Table 2).

Genomic and sequence properties of ALG-1 binding sites.

Although most genetic and computational studies support a bias for the location of miRNA target sites in 3′ UTRs, functional interaction of miRISC at other genic positions has also been demonstrated (Bartel, 2009; Duursma et al., 2008; Miranda et al., 2006; Shen et al., 2008; Tay et al., 2008). To study the global distribution of ALG-1 binding in *C. elegans* protein-coding genes, we mapped the positions of clusters relative to the length of targeted mRNAs. We observed a distinct profile of CLIP-derived cluster (CDC) occupancy proximal to the 3′ ends of spliced mRNAs from WT but not *alg-1*(–) worms (Figure 7a). Notably, the frequency of clusters throughout the composite gene model was higher in WT than in *alg-1*(–) worms, showing that ALG-1 binding extends to other genic regions (Figure 7a). Furthermore, the CDC distribution, as a percentage of 3′-UTR length, was not enriched proximal to the stop codon or poly(A) sites, in contrast to the bias for predicted miRNA target sites residing near either end of mammalian 3′ UTRs1 (Figure 7b). In fact, the fraction of clusters that mapped a given distance from the

stop codon largely mirrored the distribution of 3′-UTR lengths in *C. elegans* (Figure 7). In total, 1,656 (34.5%) of CDCs were located in 3′ UTRs, 2,473 (51.5%) in coding exons, 602 (12.5%) in introns and 75 (1.6%) in 5′ UTRs.

To characterize the sequence properties of ALG-1 binding sites, we subjected CDCs and a control set of random derived clusters (RDCs) to a battery of computational analyses. In order to perform as equitable a comparison as possible, we minimized biases due to GC content, evolutionary conservation, genic region and length of the bound region when selecting RDCs (Figure 8a,b). Furthermore, RDCs were selected from genes depleted of ALG-1 binding sites. Caveats of this approach are that a chosen RDC may actually be bound by ALG-1 at a different developmental stage or that the target mRNA may be present at such low abundance that it is not detected. Our ability to detect the *lin-41* LCS1-LCS2 region (Figure 6a and Figure 4c) despite strong downregulation of this mRNA at the L4 stage33 suggests that this second point is a minor issue. In spite of the potential limitations for assigning RDCs, our results corroborate expected properties and identify new features associated with miRISC binding to endogenous sequences (CDCs) on a global scale.

Preferential evolutionary conservation is a common feature used to predict miRNA target sites (40–42). Indeed, we observed substantially higher conservation levels within CDCs compared to RDCs in 3′ UTRs (Figure 10a) and a similar trend for coding exon and intron regions (Figure 11a). Also, consistent with the observation that functional miRNA target sites are frequently located in RNA sequences of higher accessibility (in other words, less secondary structure) (Hammell et al., 2008; Kertesz et al., 2007; Long et al., 2007; Robins et al., 2005; Zhao et al., 2005), the ALG-1–bound

regions (CDCs), as well as the 100-nt upstream and downstream flanking sequences, were significantly more accessible than RDCs in the 3′ UTRs ($P < 10{-}10$) (Figure 10b). However, this was not true for CDCs in the other genic regions (Figure 11b). It has also been suggested that a high local AU content is responsible for the more accessible 3′ UTR sites targeted by miRNAs (Baek et al., 2008; Grimson et al., 2007). Thus, we analyzed the nucleotide composition within and 100 nt upstream and downstream of 3′ UTR CDCs to search for motifs statistically enriched relative to RDCs. Unexpectedly, the ten most enriched 5- to 7-mers in 3′ UTR CDCs and their flanking regions are almost exclusively composed of CU nucleotides ($P < 10^{-4}$), revealing alternative sequence elements that may mediate miRNA–ALG-1 target recognition and regulation in *C. elegans* (Figure 10c). Moreover, this striking pattern was not associated with CDCs from 5′ UTR, coding exon or intron regions (Figure 11c) or with clusters from *alg-1*(–) animals, indicating that the CU-rich motifs are a specific characteristic of ALG-1–bound regions in 3′ UTRs.

Multiple computational prediction methods and extensive reporter validation assays point to the miRNA ‘seed’ (defined as perfect pairing between miRNA bases 2–7 and the target site) as a primary determinant for specific target recognition (Brennecke et al., 2005; Krek et al., 2005; Lewis et al., 2005; Lewis et al., 2003). Indeed, the top ten most highly cloned miRNAs in our immunoprecipitations have significantly more frequent seed pairs within the 3′-UTR CDCs than do the least-cloned miRNAs ($P < 0.0045$; Figure 12a), and this general trend was also observed when all cloned miRNAs were analyzed (Figure 12b). To globally assess whether the seed or any other mature

miRNA region showed enriched pairing capacity to CDCs, we used the complete set of miRNAs associated with ALG-1 in our experiments (Table 2).

. We calculated the number of conserved hexamers present in CDCs that have perfect complementarity to regions within cloned mature miRNAs or have one conserved G•U wobble pair (Figure 10d–g). Our analysis shows that a statistically significant ($P < 10^{-6}$) number of conserved hexamers from 3′-UTR CDCs were complementary to bases 1–6, 2–7 and 3–8 of miRNAs compared to controls: conserved RDC hexamers paired to miRNAs or CDC hexamers paired to scrambled miRNAs. We observed the strongest signal for bases 2–7 (seed), with the allowance of one G•U pair generating the highest number of sites within the 3′-UTR clusters (Figure 10d). This trend was not observed for CDCs in the other genic regions (Figure 10e and Figure 12). Unexpectedly, CDCs within coding exons showed statistically significant pairing to the central region of miRNAs ($P < 10^{-6}$) (Figure 10e). We extended our analysis to include not only perfect conservation with or without G•U pairs but also cases in which there would be a G• U pair in one of the two *Caenorhabditis* species but a perfect match in the other species (G•U and G-C, respectively, 'semiconserved hexamers') (Figure 12). This enabled us to assess the specific contributions of conservation of sequence versus pairing capacity to the patterns of miRNA complementarity to sites in CDCs. Overall, perfect conservation coupled with one G•U contributed most substantially to the pairing capacity of miRNAs to 3′-UTR and coding-exon CDCs. The percentage of 3′-UTR clusters containing a perfectly conserved seed match allowing zero and only one G•U base pair was 55% and 63%, respectively, in comparison to 30% and 41% in RDCs ($P < 10^{-4}$). These results indicate that, although

seed pairing is an important determinant of miRNA–mRNA interaction in the 3′ UTRs, other pairing conformations may contribute significantly to ALG-1 binding *in vivo*.

We also analyzed the pairing capacity of the archetypical miRNAs *let-7* and *lin-4* within 3′-UTR CDCs (Figure 10f,g). Despite the caveat that a number of these CDCs may not be targeted only by *let-7* or *lin-4*, we observed a strong pairing capacity for the *let-7* seed at positions 2–7. Unexpectedly, we observed a significant enrichment for pairing at positions 14–19 when allowing a single G•U base pair at the 3′ end of the *let-7* miRNA (Figure 10f). Notably, *lin-4* also showed significant 3′ base-pairing at the same positions, but the strongest signal at the 5′ end of the miRNA was at positions 4–9 (Figure 10g). The ability of *lin-4* to base-pair with potential target sites at positions other than the canonical 2–7 may indicate that individual miRNAs show specific pairing preferences with different outcomes for gene regulation.

Expression and functional biases of ALG-1 mRNA targets.

Regulation by miRNAs can result in substantial degradation of target mRNA levels or translational repression with little, if any, mRNA destabilization (Chekulaeva and Filipowicz, 2009). Given that the overwhelming majority of clusters reside in the 3′ UTR and coding exons, we sought to investigate whether the location of clusters affects mRNA levels. To test whether genes bound by ALG-1 at the 3′ UTR and coding exons were subject to regulation at the mRNA level, we performed microarray experiments comparing WT to *alg-1*(–) L4-stage worms. Consistent with previous reports that miRNA regulation can result in substantial target-mRNA degradation in *C. elegans* (Bagga et al., 2005; Ding and Grosshans, 2009), *lin-41*, *lin-14*, *lin-28* and many other established

miRNA targets were upregulated in the *alg-1*(–) mutant worms (Appendix 1 and Table 2).

. Notably, genes containing 3′-UTR clusters were strongly upregulated in *alg-1*(–) mutants compared to genes that had no ALG-1–bound sites (Figure 14a). In contrast, no relationship was detected between mRNA expression levels and genes with clusters in coding exons (Figure 14a). These findings suggest that the mechanism of target regulation may be different for genes with ALG-1 binding sites in 3′ UTRs versus coding exons.

We next asked if genes with ALG-1 binding sites or expression changes in *alg-1*(–) compared to WT worms were enriched ($P < 0.05$) in particular functional classes based on the "Topomap" categories, which group co-regulated genes from extensive microarray datasets (Kim et al., 2001). Notably, several functional categories were distinctly associated with genes that contained CDCs in 3′ UTRs versus coding exons and were up- or downregulated in *alg-1*(–) mutants (Figure 14b). For example, genes belonging to the functional classes "Protein kinases" and "Cell biology" are enriched for containing 3′-UTR CDCs and being upregulated in *alg-1*(–) worms. Genes in the "Histone" category are also associated with upregulation but tend to have CDCs in their coding exons. This difference in locality of miRNA binding may be related to the typically short and nonpolyadenylated status of histone mRNAs (Dominski and Marzluff, 2007). Some functional categories included genes with CDCs in the 3′ UTR and coding exons and/or up- and downregulated genes. The overlap in categories is not surprising given the large fraction of genes with ALG-1–bound regions and the likely widespread direct and indirect effects on mRNA expression by the miRNA pathway. Our results

reveal biological pathways targeted *in vivo* by ALG-1 in developing worms and indicate that some gene categories tend to be differentially bound and regulated by ALG-1.

miRNA pathway genes are enriched in ALG-1 targets.

During our analyses of categories of genes bound by ALG-1, we discovered a strong enrichment for genes implicated in the miRNA pathway. CDCs in the 3′ UTR in the *alg-1* gene indicate autoregulation of this core miRNA factor (Figure 14c). Additionally, significant clusters were identified in the 3′ UTRs of *ain-1* and *ain-2*, and mRNA levels of these genes and of the *alg-1* homolog *alg-2* were found to be upregulated in *alg-1*(–) worms (Table 3). The potential cross-regulation of these miRNA effector genes may explain the nonlethal phenotype associated with loss of any single one of these genes (Grishok et al., 2001; Zhang et al., 2007). To investigate the extent of ALG-1 regulation of miRNA pathway genes, we analyzed two published lists of genes specifically connected to miRNA function by proteomic and genetic evidence (Dominski and Marzluff, 2007; Zhang et al., 2007). We observed that this network of miRNA pathway genes showed statistically significant enrichment in ALG-1 CDCs (30 out of 39 genes identified by proteomics and 15 out of 44 identified by genetics), compared to an expectation of ~16% ($P < 10^{-4}$; Table 3). We speculate that cross-regulation of these genes may confer robustness to the miRNA pathway by relaxing repression of miRNA cofactors to compensate for insufficiencies in major components such as ALG-1.

ALG-1–bound regions as a resource for miRNA target predictions.

A number of different algorithms (mirWIP, rna22, PicTar, TargetScan, PITA and miRanda) are available for predicting miRNA target sites in *C. elegans* genes (Dominski and Marzluff, 2007; Hammell et al., 2008; Kertesz et al., 2007; Lall et al., 2006; Miranda et al., 2006; Ruby et al., 2006). Most of these prediction methods use a common set of criteria (seed, conservation and energy requirements), except for PITA, which does not require conservation, and rna22, which uses a different set of parameters. Because predictions are typically available for 3′-UTR sequences, we assessed the ability of these methods to detect predicted miRNA target sites within the ALG-1–bound 3′-UTR CDCs (Figure 15, tracks for the predicted sites from these algorithms are available under 'ALG1 CLIP-seq' within the 'Regulation' section in the ce6 genome). Although 93% of the 3′-UTR CDCs contained a miRNA target site predicted by at least one of the algorithms (1,539 CDCs), only 3% of the CDCs had at least one site predicted by all 6 programs (52 CDCs). As an example, five of the six target prediction programs list potential miRNA target sites, largely disparate in both location and number, in the *alg-1* 3′ UTR (Figure 14c); our results narrow the regions recognized and bound by ALG-1 *in vivo* at the specific developmental stage tested. The prominent disparity among prediction methods has been previously noted (Hammell et al., 2008) and emphasizes the value of the ALG-1 CLIP-seq as a tool to improve miRNA target identification.

CONCLUSIONS

We present a global snapshot of an endogenous miRISC RNA binding profile in whole animals. We demonstrate that binding of the core miRNA effector protein

Argonaute is strongly enriched at the 3′ ends of transcripts, although substantial numbers of CDCs also reside within the 5′ UTR, coding exonic and intronic regions of genes as well. A striking signature of the ALG-1–bound 3′-UTR CDCs emerged: the regions showed greater sequence conservation and accessibility, they contained and were flanked by CU-rich motifs, they were enriched for sequences complementary to the 5′-end seed regions of miRNAs and they were associated with upregulation of mRNA expression in the *alg-1*(–) mutant background. Although some of these characteristics were shared with clusters in other genic regions, the marked overall differences in 3′ UTR versus other regions suggests that separate rules may regulate ALG-1 binding to distinct positions within an mRNA. The importance of context could underlie the conflicting conclusions that have been drawn about the ability of miRNAs to target different regions in mRNAs (Gu et al., 2009; Kloosterman et al., 2004; Lytle et al., 2007) and, in some cases, the failure of reporter assays to demonstrate miRNA regulation of genes bound and regulated by *alg-1* (see Table 2).

. In addition to providing a map of ALG-1 interaction sites for the *C. elegans* protein-coding genes potentially under miRNA regulation in late larval development (see Figure 15 and Methods), compared to previously available methods our strategy substantially reduced the search space by factors of 5, 20 and 47 for identifying direct miRNA target sites. Although we detected a strong signal for pairing to the miRNA seed region in 3′-UTR ALG-1–bound sites, ~40% of the ALG-1 clusters lacked conserved seed pairing capacity, indicating that more flexible base-pairing rules may guide a large fraction of miRNA target recognition *in vivo*. Furthermore, the observation of different patterns for *let-7* or *lin-4* miRNA paired to sites within ALG-1–bound sequences raises

the possibility of individual miRNA pairing rules. The discovery of miRNA pathway genes as an exceptional class of genes bound and regulated by endogenous *alg-1* suggests that cross-regulation of miRNA cofactors contributes substantially to this essential posttranscriptional control mechanism. In conclusion, our analyses and data provide a framework and a rich resource for understanding *in vivo* miRNA–mRNA interactions in a context-specific manner.

METHODS

Accession codes

Microarray CEL files have been deposited at the Gene Expression Omnibus database repository under accession number GSE19138.

Cross-linking and immunoprecipitation coupled to sequencing

For three independent experiments using WT and *alg-1*(–) worms, we harvested approximately 40,000 L4-stage worms and irradiated them with UV-B (3 kJ m^{-2}). We lysed the irradiated worms by sonication. For the immunoprecipitation, we used a custom antibody and performed the rest of the CLIP-seq method essentially as described (Yeo et al., 2009). Each step of the protocol as we applied is described below.

UV Irradiation and Lysis. For three independent experiments using WT and *alg-1*(-) worms, approximately 40,000 staged L4 worms were harvested and irradiated with UV-B (3kJ/m^2) in 100 mm plates using the Spectrolinker XL-1000 (Spectronics Corporation). The irradiated worms were lysed by sonication in Homogenization Buffer [100mM NaCl (Fisher Scientific), 25mM HEPES (Sigma), 250uM EDTA (Fisher

Scientific), 2mM Dithiothreitol (DTT) (Fisher Scientific), 0.1% NP-40 (Amersham Biosciences), 24units/ml RNAsin (Promega) and Protease Inhibitors (Complete Mini, Roche)]. Sonication was performed using the Sonic Dismembrator Model 100 (Fisher Scientific) with five 10-second pulses (18 watts RMS output power). Lysates were centrifuged at 16,000g for 15min at 4°C and supernatants were collected. Protein concentration of the lysate supernatants was adjusted to 1mg/ml.

Immunoprecipitation. Protein G Sepharose beads (GE Healthcare) were washed 3 times with the Homogenization Buffer with gentle shaking and centrifugation for 30sec at 1,000g. Lysates were pre-cleared with 50ul of beads (100ul of 50:50 beads/Homogenization Buffer) for 1h at 4°C, centrifuged at 1,000g for 30 sec, and the supernatant was incubated with anti-ALG-1 antibody [7ug of antibody for 1ml of protein lysate (1mg/ml)] overnight at 4°C with gentle shaking. 50ul of beads (100ul of 50:50 beads/Homogenization Buffer) were added the next day and incubated for 1h at 4°C with gentle shaking. The beads were collected with centrifugation and washed: i) twice with Wash Buffer [1X PBS (20 mM Tris HCl, pH 7.4, 137 mM NaCl), 0.1% sodium dodecyl sulfate (SDS, Fisher Scientific), 0.5% sodium deoxycholate (Fisher Scientific), and 0.5% NP-40], ii) twice with High Salt Wash Buffer [5X PBS, 0.1% SDS, 0.5% sodium deoxycholate, and 0.5% NP-40] and iii) twice with Polynucleotide Kinase Buffer (PNK Buffer) [50mM Tris-Cl pH 7.4 (Fisher Scientific), 10mM $MgCl_2$ (Fisher Scientific) and 0.5% NP-40].

Fragmentation of the RNA tag. Trimming of the RNA tag was performed by Micrococcal Nuclease treatment. Beads were incubated with 500ul of MN Reaction

Buffer [50mM Tris-Cl pH 7.9, 5mM $CaCl_2$ (Fisher Scientific) containing 1ng of Micrococcal Nuclease (NEB) for a total of 10 min at 4°C with intermittent shaking on a Thermomixer R (Eppendorf) (1200rpm for 1min and then 1200rpm for 15 sec every 3 minutes). The beads were then washed with ice-cold buffers: i) twice with PNK+EGTA Buffer [50mM Tris-Cl pH 7.4, 20mM EGTA (Fisher Scientific) and 0.5% NP-40] in order to inactivate the Micrococcal Nuclease activity, ii) twice with Wash Buffer and iii) twice with PNK Buffer.

Alkaline Phosphatase Treatment. The 3' phophoryl group of the fragmented RNA tag was removed by Alkaline Phosphatase. The beads were incubated for 10 minutes at 37°C in the Thermomixer with intermittent shaking (1200rpm for 15sec every 3 min) in an 80ul 1X NEB Buffer 3 solution containing 30 units of Calf Intestinal Phosphatase (NEB). The beads were then washed: i) twice with PNK+EGTA Buffer to quench the phosphatase activity, ii) twice with PNK Buffer and finally iii) twice with 0.1mg/ml Bovine Serum Albumin (BSA, NEB) solution to increase the efficiency of the following ligation reaction.

3′ RNA Linker Ligation. The RNA linker 5'-UCG UAU GCC GUC UUC UGC UUG-3' with a puromycin modification at the 3' end (PAGE purified, Dharmacon) to avoid self-circularization was linked to the mRNA/miRNA present in the ALG-1 complexes by the T4 RNA Ligase (Fermentas). The RNA linker is compatible with the Illumina 1G sequencing system. 160 pmoles of the linker were added to the beads in a 80ul reaction according to the manufacturer's instructions and allowed to incubate

overnight at 16°C with gentle shaking (1300rpm every 5min for 15 sec in the Thermomixer).

Polynucleotide Kinase Treatment. The 5' hydroxyl of the RNA tag due to the Micrococcal Nuclease activity was converted to a 5’ phosphoryl group by Polynucleotide Kinase. The beads were washed three times with PNK Buffer and incubated in 80 ul PNK Buffer (NEB) with 40 units of T4 PNK enzyme (NEB) in the presence of P^{32}-g-ATP (1mCi). The samples were incubated for 10min at 37°C with intermittent shaking (1000rpm every 4 min for 15 sec). Cold ATP was added to the reaction at a final concentration of 1.25mM and incubated for five additional minutes. The reaction was terminated with three washes of PNK+EGTA Buffer.

Sodium Dodecyl Sulfate Polyacrylamide Gel Electrophoresis. The ALG-1/mRNA/miRNA complexes were eluted from the beads by incubation for 10 min at 70°C in Nupage LDS Buffer (Invitrogen) without a reducing agent. Total input and supernatant controls were also treated the same way. The samples were loaded onto a native 10% Bis-Tris Gel (Invitrogen) with MOPS SDS Running Buffer (Invitrogen). Next, the samples were transferred to a nitrocellulose membrane of 0.45um pore size (Biorad) with Nupage Transfer Buffer (Invitrogen) using a wet transfer apparatus (Biorad) at 40V for 4h. The membrane was exposed to MS autoradiogram film (Kodak) and the ALG-1/RNA complexes appeared as a diffused radioactive band at approximately 150kDa.

RNA Isolation and Purification. The band corresponding to the ALG-1/RNA complexes was cut out of the membrane and protein was degraded by Proteinase K (Invitrogen). 200ul of a 4mg/ml Proteinase K solution in PK Buffer (100mM Tris-Cl pH 7.5, 50mM NaCl, 10mM EDTA) were incubated for 20min at 37°C to deactivate any RNAses present and was then added to each of the isolated pieces of nitrocellulose membrane for another 20min at 37°C with rigorous shaking (Thermomixer, 1200rpm). 200ul of a 7M urea solution in PK buffer were added to the samples and incubated for an additional 20 min at 37°C with rigorous shaking (1000rpm). The samples were then subjected to phenol/chloroform extraction followed by ethanol precipitation.

5' RNA Linker Ligation and DNAse treatment. RNA was resuspended and ligated to the Illumina 1G Sequencing system-compatible 5' RNA Linker 5'-AAU GAU ACG GCG ACC ACC GA-3' that harbors a biotin modification at the 5' end (Page Purified, Ambion) to avoid self-circularization. The RNA ligation was performed by T4 RNA Ligase (Fermentas) in a 10ul reaction according to the manufacturer's instructions in the presence of 20pmoles of 5' RNA linker at 16°C and allowed to go overnight. The RNA samples were subjected to DNase treatment, to eliminate potential DNA contamination, in an 80ul reaction in DNAse I Buffer (Ambion) and 5units of DNAse I (Promega) for 20min at 37°C followed by phenol/chloroform extraction and ethanol precipitation.

Reverse Transcription Polymerase Chain Reaction (RT-PCR). The ligated RNA was resuspended and reverse transcribed in a 20ul reaction using Superscript III (Invitrogen), according to the manufacturer's instructions, with 20pmoles of P3 Primer

5’-CAA GCA GAA GAC GGC ATA CG A-3’ [PAGE purified (IDT DNA)]. Polymerase Chain Reaction was performed using the Phusion High Fidelity Polymerase (NEB) in a 30 ul reaction, following the manufacturer’s recommendations, with 10pmoles of P3 and P5 [5’-AAT GAT ACG GCG ACC ACC GA-3’, PAGE Purified (IDT DNA)] primers. The PCR product was analyzed on an 11% PAGE gel (Pasquinelli *et al*, 2003), the bands corresponding to 75-150 nt were isolated and the DNA was extracted with 50mM NaCl and 10mM Tris-Cl pH 8.0. A fraction of the PCR product was cloned into the sequencing vector TOPO-4 PCR using the TOPO-cloning kit (Invitrogen) and sequences were analyzed to verify that that the PCR product did not contain linker concatamers.

Detection of specific mRNA transcripts by PCR

The primers used for mRNA transcript detection after PCR were: for *lin-41* LCS region, A588 Primer (5’-TAA TAC GAC TCA CTA TAG GGC ACC TCT TTT CCT CAA ATTG C-3’) and A25 Primer (5’-GAG GCA GAA TGG TTG TAT AA-3’) and for *lin-41* coding region, A261 (5’- TTG C AGC AATC GAT GAA GAC AAC-3’) and A262 (5’- AGT GGG CCA TGT GCC AAG AAT AG-3’).

Western blot Analysis

The nitrocellulose membrane was blocked with 5% nonfat dry milk in Tris-buffered saline (TBS)-Tween (T) (20 mM Tris HCl, pH 7.6, 137 mM NaCl, and 0.1% Tween 20) at 37°C for 1h. The blot was then probed with a polyclonal antibody P0345 (Open Biosystems), in 5% nonfat dry milk-TBS-T at 4°C overnight. The P0345 antibody was raised against an oligopeptide sequence present only in the ALG-1 protein but not

the other 26 Argonaute homologs in *C. elegans* and, thus, specifically recognizes the ALG-1 protein. The secondary antibody was horseradish peroxidase-conjugated anti-rabbit IgG (Jackson Imunochemicals) and the nitrocellulose membrane was visualized using enhanced chemiluminescence reagents (GE Healthcare) followed by exposure to MS Film (Kodak).

Microarray analysis

We prepared three independent sets of RNA from WT and *alg-1*(–) worms at the L4 stage, labeled them as per manufacturer's instructions (Affymetrix) and hybridized them to Affymetrix *C. elegans* gene microarrays. To assign a value for differential gene expression between the two groups, we computed a *t*-statistic as described before[59]; to assign a value for differential gene expression between the two groups, a t-statistic was computed as in (Yeo et al., 2007b) using the equation:

$$t_{WT,ALG-1} = \frac{\mu_{WT} - \mu_{ALG-1}}{\sqrt{\frac{((n_{WT}-1)\sigma^2 + (n_{ALG-1}-1)\sigma^2)(n_{WT}+n_{ALG-1})}{(n_{WT}n_{ALG-1})(n_{WT}+n_{ALG-1}-2)}}}$$

where μ is the mean expression for n replicate probes with a variance of σ^2. Using the equation in this form gives a negative t-statistic for genes that are up-regulated (higher mean value for μ_{ALG-1})**.** Multiple hypothesis testing was corrected by controlling for the false discovery rate (Benjamini-Hochberg). For the analysis of the relationship between the presence of ALG-1 binding sites and mRNA expression levels (Fig. 4A) the Wilcoxon rank-sum test was used.

Improved UTR annotation

We obtained publicly available stage-specific *C. elegans* RNA-sequence libraries from published work[32], aligned them to the genome of *C. elegans* and then assigned them to the composite gene loci, as described in detail below. To control for RNA transcripts from noncoding and unannotated transcribed regions, we applied a 3-kb cutoff such that only reads within 3 kb of an annotated gene end could be assigned to the gene.

In summary, genome sequences of *C. elegans* (ce6) and annotation for protein-coding genes were obtained from the University of California, Santa Cruz (UCSC) Genome Browser. Sanger mRNAs and expressed sequence tags were aligned to known *C. elegans* genes and known isoforms to generate a composite gene locus. Stage-specific *C. elegans* RNA-sequence libraries were obtained from Hillier et al. (Hillier et al., 2009). Reads from 16 RNA-seq library runs from L3 and L4 stage worms were combined since our definition of L4 stage worms fell between the time-points for the definitions of L3 and L4 stages used in Hillier et al. (Hillier et al., 2009). The 36 base pair reads were aligned to genome sequences of *C. elegans* using bowtie version 0.9.9.2 with options -m 5 -k 5 (Langmead et al., 2009). Genome aligned reads were then assigned to the composite gene loci with the following rules: reads aligned to a gene locus were assigned to that gene and any overlapping genes; reads aligned in the intergenic regions between gene loci were preferentially assigned to the 3' end of the gene(s); reads aligned in the intergenic regions between annotated 5’ ends of gene loci were assigned to the closest gene. We based the 3' preference rule on the fact that the poly-A selection used in preparing the RNA-seq libraries resulted in a bias with more reads sequenced from the 3' end of genes. In addition, the current annotated *C. elegans* 3'UTRs are significantly

longer than annotated 5'UTRs ($P<0.05$, one-tailed t-test). In addition, to control for RNA transcripts from non-coding and un-annotated transcribed regions, a 3kb cutoff was applied such that only reads within 3kb of an annotated gene end could be assigned to the gene. Out of the combined set of 21,416,772 L3 and L4 RNA-Seq reads, 20,047,952 were assigned to genes. Next, reads assigned to each gene were analyzed to determine if extension of the annotated 5' and/or 3' gene end was supported by the RNA-seq data. For each gene, two representative reads from each sequencing run were considered, the one assigned farthest upstream of the gene and the one assigned farthest downstream. Considering 5' and 3' ends independently, if reads were assigned past the annotated gene end in more than half of the sequencing runs, the gene end was extended to the average location of these reads.

Defining experimentally reproducible regions

We aligned reads from ALG-1 CLIP-seq to the repeat-masked *C. elegans* genome (ce6) and extended 50 bases in the 3′ direction to account for the size of the gel-extracted PCR product. Reads that overlapped within and across experiments formed contiguous 'regions.' We assigned a score to each nucleotide within a region, and we gave more weight to the region from the experiment with the most reads based on the assumption that there was a higher likelihood of detecting real interactions, even for weakly abundant RNAs. After assigning weights to each nucleotide, we considered a region to be replicated across biological experiments only if at least one of the nucleotides had a score greater than a user-defined cutoff.

In more detail, for this process after removal of primer adaptor sequences from the Illumina 1G Sequencing data, the reads from ALG-1 CLIP-seq were aligned to the

repeat-masked *C. elegans* genome (ce6) using bowtie version 0.9.9.2 with options -l 25 -n 3 -e 200 -y -k 10 --best -m 10 -p 4. Reads that were aligned to the *C. elegans* genome were extended 50 bases in the 3' direction to account for the size of the gel-extracted PCR product. Reads in the same strand as our re-annotated gene loci were analyzed as follows. Reads that overlapped within and across experiments formed contiguous 'regions.' Each nucleotide within a region was assigned a score, $S_c = \sum_{i=1:3} w_i\gamma_i$ where c was condition wild-type (WT) or *alg-1(-)* (MT), i was the biological replicate experiment in condition c (3 each), w_i was the "weight" associated with the experiment and γ_i was the indicator function (1 if the region in experiment i has reads, 0 otherwise). We defined w_i as the number of reads aligned in a given experiment divided by the total number of reads aligned in all experiments from that condition (WT or MT). This method gave higher weight to the region from the experiment with the most reads based on the assumption that there was a higher likelihood of detecting real interactions even for lowly abundant RNAs. After weights were assigned to each nucleotide, a region was considered replicated across biological experiments only if at least one of the nucleotides had a score S_c greater than a user-defined cutoff of 0.5, after testing various cutoffs of 0.3334, 0.4, 0.5, and 0.6667.

Computational identification of ALG-1 binding sites

For finding peaks, we considered only the regions that had at least one nucleotide satisfying the user-defined cutoff within WT or *alg-1*(–) samples. We calculated significant peaks by first determining read-number cutoffs using the Poisson distribution. The Poisson distribution assumes all intervals are independent and have equal probability

of an occurrence happening. We determined a global and local cutoff by assigning the cutoff value using the whole transcriptome frequency as the global cutoff and using a gene-specific frequency for the local cutoff. The gene-specific frequency was simply the number of reads overlapping a gene divided by the pre-mRNA length of that gene. After finding these cutoffs, we used a sliding window the size of the interval to determine where the actual read numbers exceed both the global and local cutoffs. At each significant interval, we attempted to extend the region by adding in the next read and recalculating the significance of this new interval. If the probability was still significant, and the distance between this extension and the previous interval was sufficiently small, this read was included and the peak width was updated. This extension was empirically limited to two times the size of the minimal interval. We identified clusters independently for CLIP-seq performed on WT and *alg-1*(–) strains. Next, we considered WT clusters that overlapped with *alg-1*(–) by 25% either as abundant unbound RNA cloned independently of ALG-1 interaction or as PCR artifacts and removed these clusters. We considered WT clusters that did not overlap with *alg-1*(–) clusters for more than 25% of their length as bona fide ALG-1–interacting loci and termed them CLIP-derived clusters (CDCs).

Significant peaks were calculated by first determining read number cutoffs using the Poisson distribution,

$$f(k;\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$$

where λ was the frequency of reads mapped over an interval of nucleotide sequence, k was the number of reads being analyzed for significance, and f (k;λ) returned

the probability that exactly k reads would be found. For any desired p-value, p-cutoff, a read number cutoff was calculated by summing the probabilities for finding k or more tags, and determining the minimum value of k that satisfies i=k such that f(k;λ)>p-cutoff. The frequencyλwas calculated simply by dividing the total number of mapped reads by the number of non-overlapping intervals present in the transcriptome. The interval size was chosen based on the average size of the CLIP product, which includes only the cloned RNA fragments but not any ligated adapters. The Poisson distribution assumed all intervals were independent and have equal probability of an occurrence happening. A global and local cutoff was determined by assigning the cutoff value using the whole transcriptome frequency as the global cutoff, and using a gene-specific frequency for the local cutoff. The gene-specific frequency was simply the number of reads overlapping that gene divided by the pre-mRNA length of that gene. After finding these cutoffs, a sliding window the size of the interval is used to determine where the actual read numbers exceed both the global and local cutoffs. At each significant interval, we attempt to extend the region by adding in the next read, and recalculating the significance of this new interval. If the probability was still significant, and the distance between this extension and the previous interval was sufficiently small, this read was included and the peak width was updated. This extension was empirically limited to two times the size of the minimal interval. We identified clusters independently for CLIP-seq performed on wild-type (WT) and *alg-1*(-) mutant (MT) strains. Next, we considered WT clusters that overlapped with MT by 25% as abundant unbound RNA cloned independently of ALG-1 interaction, or PCR artifacts and removed these clusters. WT clusters that did not overlap with MT clusters for more than 25% of their length were considered *bona fide* ALG-1

interacting loci and termed CLIP-derived Clusters (CDCs). A BED file containing the genomic coordinates for CDCs will be available after publication.

Assigning CDCs to Functional Genic Regions

After clusters were located to a composite model of gene regions as in (Yeo et al., 2007a), the genomic coordinates for each CDC were compared to the extended RefSeq database to determine to which functional genic region the CDC most likely belongs. A randomly chosen RefSeq gene was used to assign clusters to a functional genic region based on where the center of the cluster mapped on that randomly-chosen RefSeq gene.

Generation of randomly derived clusters (RDCs)

We divided genes into quartiles based on the number of reads aligned uniquely to each gene locus. We compared binding sites within genes to 'background regions' in genes expressed within the same quartile. We further divided the transcriptome into functional regions: 5′ UTR, coding exon, intron and 3′ UTR. For each region, we determined the average evolutionary conservation level by the algorithm PhastCons_{60} and GC content. We then divided each region into quartiles based on their conservation level and GC content. For each bona fide binding site of length L that is contained within an unambiguously assigned functional region (see above) of conservation level C and GC content G, we picked a background binding site at random of length L from the transcriptome that fell in the same functional region of conservation level C' and GC content G', where C' and C as well as G' and G are in the same conservation and GC quartile, respectively. We implemented controls for C' and G' on the level of the whole

genic region. For the determination of motifs and conservation levels for CDCs vs RDCs, we did not control GC-content or conservation levels, respectively. For further details see below.

In order to control for gene expression, genes were divided into quartiles based on the number of reads aligned uniquely to each gene loci. Binding sites within genes were compared to 'background regions' in genes expressed within the same quartile. To generate sets of background regions comparable to the ALG-1 binding sites in terms of nucleotide biases and conservation patterns, we further divided the transcriptome into functional regions – 5'UTR, coding exon, intron and 3'UTR. For each region, we determined the average evolutionary conservation level (expressed from 0 to 1 as computed by the algorithm PhastCons (Siepel et al., 2005) and GC content (expressed from 0 to 1 as the ratio of #(G+C)/#(A+C+G+T). Each region was then divided into quartiles based on their conservation level and GC content. For example, a coding exon can be in the first quartile (0-0.25) in terms of conservation, and in the last quartile (0.75-1.0) in terms of GC content. For each *bona fide* binding site of length L that is contained within an unambiguously assigned functional region (see above) of conservation level C and GC content G, a background binding site of length L is picked at random from the transcriptome that falls in the same functional region of conservation level C' and GC content G', where C' and C, and G' and G are in the same conservation and GC quartile, respectively. Importantly, the controls for C' and G' are implemented on the level of the whole genic region (i.e. a whole exon or a whole intron) rather than the randomly selected portion of that region of length L. A total of 10 independent sets of background binding sites were generated for each ALG-1 binding site. For the determination of

motifs and conservation levels for CDCs vs RDCs, we did not control for GC-content or conservation levels, respectively.

Assessing miRNA-target base-pairing rules

We used all the miRNAs that were sequenced in the WT library in order to analyze base-pairing to target regions (CDCs) as compared to background regions (RDCs). We assessed the number of CDC sites complementary to every adjacent position of each miRNA for two definitions of conservation: (i) 'exact' conservation and (ii) 'semiconservation' and for two different definitions of 'binding capacity': (i) Watson-Crick base-pairing and (ii) G•U mismatches, allowing a single G•U base pair in a 6-mer site (see inset in Figure 13). For further details see below.

All miRNAs that were sequenced in the WT library were utilized to analyze base-pairing to target regions (CDCs) as compared to background regions (RDCs). The number of CDC sites complementary to every adjacent position of each miRNA was assessed for two definitions of conservation: (1)"Exact" conservation and (2) "semi-conservation" (see below) and for two different definitions of "binding capacity": (1) Watson-Crick base-pairing and (2) G:U-mismatches, allowing a single G-U base pair in a 6-mer site (see inset in Figure 12). To control for the level of potential binding expected by the relative stringency of each definition of miRNA binding, mature miRNA sequences were shuffled and compared to CDCs.

Prediction of miRNA target sites within ALG-1 bound regions by published algorithms

To compare different miRNA target algorithms, it was necessary to have all of the target predictions in terms of genome coordinates from the same build of the *C. elegans* genome. Six target prediction algorithms were used: (1) TargetScan, (2) PicTar, (3) mirWIP, (4) rna22, (5) PITA and (6) miRanda. Each algorithm was converted to ce6 genome coordinates as follows: (1) TargetScan (Conserved and Non-Conserved) predictions are given as 3'UTR coordinates on RefSeq genes, these coordinates were converted to ce4 genome coordinates using RefSeq gene annotations from the UCSC genome browser. ce4 genome coordinates were converted to ce6 coordinates using the liftOver utility from the UCSC genome browser. (2) PicTar predictions are given as ce2 genome coordinates, they were first converted to ce4 genome coordinates then to ce6 genome coordinates using the liftOver utility. (3) mirWIP predictions are given as relative gene coordinates. These predictions correspond to "final" binding sites that have passed strict thresholds on sites scores, family scores, and total target scores as defined by mirWIP [(M. Hammel, personal communication and (Hammell et al., 2008)]. Gene coordinates are converted to ce4 coordinates using gene annotations from the UCSC genome coordinates. ce4 genome coordinates were converted to ce6 coordinates using the liftOver utility from the UCSC genome browser. (4) RNA22 predictions are given as a predicted targeted gene along with the predicted pairing of miRNA to the target RNA. To determine the ce6 coordinates, the target RNA sequence was aligned to the ce6 genome using the program BLASTn. (5) PITA all and PITA top predictions are given as ce6 genome coordinates, no conversion was necessary. (6) Miranda predictions are given as ce4 genome coordinates. ce4 coordinates were converted to ce6 coordinates with the

liftOver utility. After all the prediction algorithms were expressed in terms of ce6 genome coordinates, we analyzed the 3'UTR CDCs for overlap with predicted miRNA target sites by each of the programs.

Conservation levels

Conservation was determined by parsing with perl scripts the multiz alignments of the *C. elegans* (May 2008, ce6) to the *C. brenneri* (Feb 2008, caePb1) genomes downloaded from the UCSC genome browser. Unless otherwise stated, aligned nucleotides were determined to be either perfectly conserved or not and the conservation of a region was defined as the fraction of bases that are conserved to the length of the region. For analyses of base-pairing of microRNA sequences to complementary mRNA sequences, changes between C and U among species were considered conserved because these changes maintain the capacity to base pair to a G nucleotide in the miRNA. Box-plots represent the distribution of values as follows: the central mark indicates the median, the edges of the box are the 25th and 75th percentiles and the whiskers extend to the most extreme datapoints that the algorithm considers not to be outliers, and the outliers are plotted individually.

Accessibility

The algorithm RNAplfold from the Vienna RNA package (Bernhart et al., 2006) was utilized to compute local pair probabilities for base pairs with parameters -u 8 -W 80 -L 40. Accessibility was represented as the log (base 10) average probability of being

unpaired within the window. Kolmogorov-Smirnov two-sample test was used to compare the distributions between CDCs and RDCs.

Calculation of miRNA search space

Identifying a miRNA binding site within the entire C. elegans transcriptome requires searching T (32,022,560nt), where T is the length of the transcriptome. By restricting to the 3'UTR regions U (5,309,907nt), where U is the length of all 3'UTRs in the transcriptome, computational algorithms reduce the search space by ~6-fold (T/U). ALG-1 immunoprecipitation within a specific stage reduces the search space to Tip (7,408,519nt) corresponding to ~4-fold reduction (T/ Tip), where Tip is the length of the immunoprecipitated transcripts. Immunoprecipitation approaches followed by searching within 3'UTRs (Hammell et al., 2008; Zhang et al., 2007; Zhang et al., 2009) reduces the search space to Uip (1,341,377nt), where Uip is the length of 3'UTRs from immunoprecipitated RNAs. This corresponds to reductions of ~6-fold (Tip/Uip) or ~4-fold (U/Uip). The use of CLIP-seq focuses the search space to TCLIP (676,696nt) and UCLIP (265,232nt) where TCLIP and UCLIP are the total sizes of CLIP regions in the transcriptome (including coding exons) and 3'UTRs only. Compared to the entire transcriptome, the search space is reduced by 47-fold (T/TCLIP). Compared to 3'UTRs, the search space is reduced by 20-fold (U/UCLIP) Compared to normal immunoprecipitation approaches, this reduces the search space by ~5-fold (Uip/UCLIP), while retaining a high sensitivity and specificity for identifying experimentally verified miRNA targets in worms.

Generation of C. elegans gene expression map (Topomap)

We compared lists of genes with 3'UTR CDCs, coding exon CDCs, genes up-regulated or down-regulated in alg-1(-) versus WT worms, and genes that have 3'UTR or coding exon CDCs and are also up-regulated or down-regulated in alg-1(-) versus WT worms to the C. elegans gene expression map data (Kim et al., 2001) by uploading these gene lists to http://elegans.uky.edu/gl/cgi-bin/gl_mod.cgi?action=compare2. p-values used were Holm-Bonferroni corrected probabilities associated with the "representation factor."

Coincidence of CLIP-seq detected genes and miRNA pathway genes

To determine the significance of co-occurrence of genes detected by ALG-1 CLIP-seq and the genes implicated in the miRNA pathways by Zhang *et al*, 2007 and Parry *et al*, 2007 a Z-score was computed by the following:

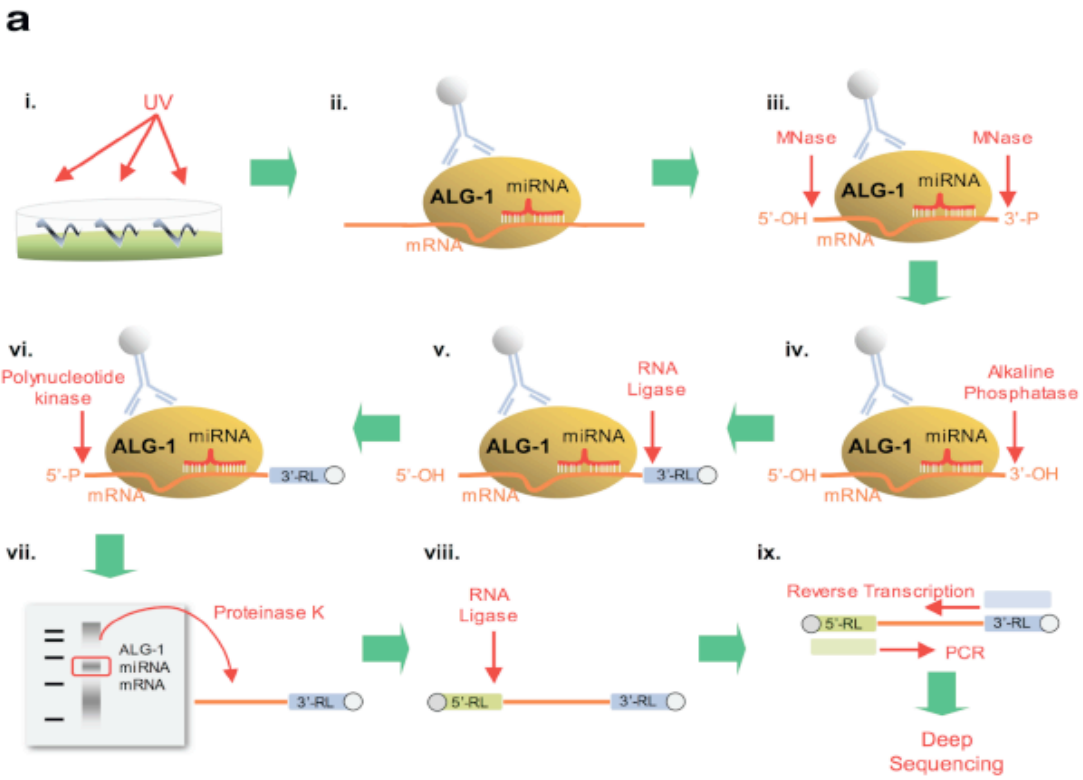
$$\frac{|y - \mu|}{\sigma}$$

where y is the number of genes that were detected by both CLIP-seq and either Zhang *et al* or Parry *et al*, μ is the mean number of genes that match either Zhang *et al* or Parry *et al* from a randomly selected, equally sized set of genes the same length as the real number of CLIP-seq, over 1000 iterations and σ is the standard deviation of these 1000 iterations. P-values were derived from the Z-distributions.

FIGURES

Figure 4. Key steps of the ALG-1 CLIP-seq protocol.

(a) Outline of the experimental approach: i) stabilization of ALG-1/RNA interactions with UV irradiation *in vivo*, ii) immunoprecipitation of the ALG-1/RNA complexes iii) trimming of the ALG-1-associated RNA by micrococcal nuclease (MNase), iv) removal of the residual 3′ phosphoryl group, v) 3′ ligation with a 3′ puromycin-modified RNA linker, vi) 5′-labeling using P^{32}-γ-ATP, vii) native SDS-PAGE to isolate the band corresponding to ALG-1/RNA complexes, degradation of proteins with proteinase K and RNA extraction, viii) ligation with a 5′ biotin-modified RNA linker and ix) RT-PCR amplification and parallel sequencing using the Illumina 1G system. (b) Western Blot analysis of protein lysates from wild-type (WT) and *alg-1*(*gk214)* [*alg-1*(-)] animals using a rabbit polyclonal antibody against *C. elegans* ALG-1. (c) Detection by PCR of the *lin-41* 3′UTR region containing the *let-7* complementary sites 1 and 2 (LCS1 & 2) in the WT and *alg-1*(-) CLIP-seq DNA libraries prior to comprehensive sequencing. RT-PCR amplification of *lin-41* coding region was used as a negative control.


a
i.
UV
ii.
ALG-1
miRNA
mRNA
iii.
MNase
MNase
ALG-1
miRNA
5'-OH
3'-P
mRNA
iv.
Alkaline
Phosphatase
ALG-1
miRNA
5'-OH
3'-OH
mRNA
v.
RNA
Ligase
ALG-1
miRNA
5'-OH
3'-RL
mRNA
vi.
Polynucleotide
kinase
ALG-1
miRNA
5'-P
3'-RL
mRNA
vii.
Proteinase K
ALG-1
miRNA
mRNA
3'-RL
viii.
RNA
Ligase
5'-RL
3'-RL
ix.
Reverse Transcription
5'-RL
3'-RL
PCR
Deep
Sequencing

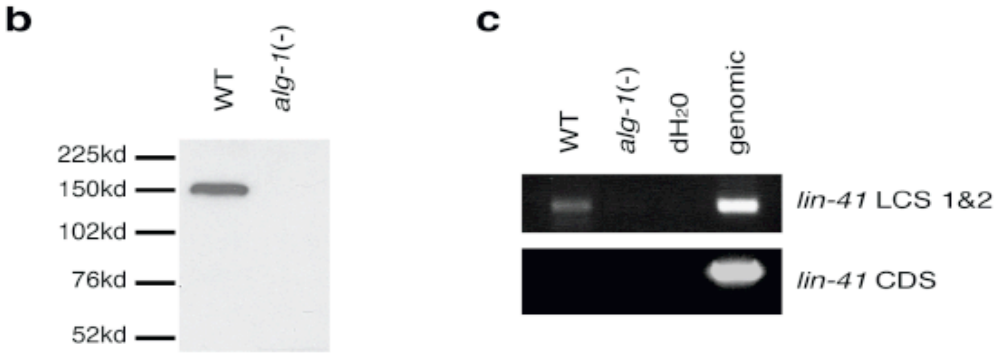
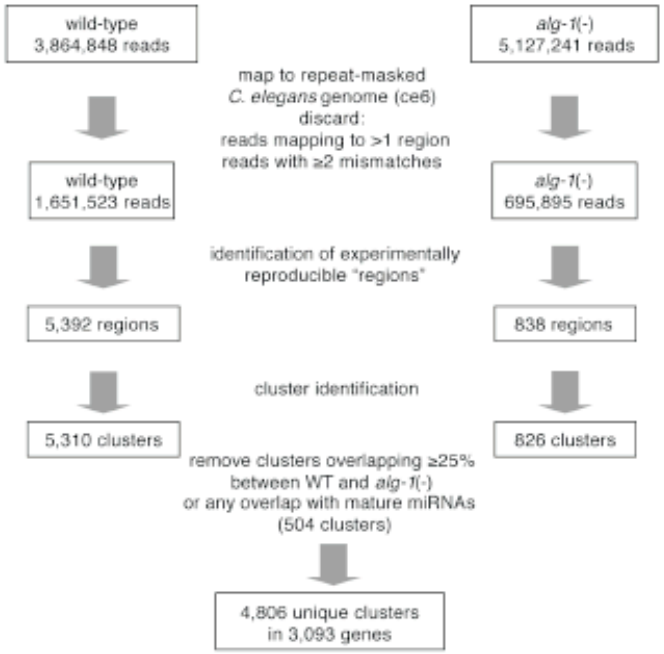

b
WT
alg-1(-)
225kd
150kd
102kd
76kd
52kd
c
WT
alg-1(-)
dH2O
genomic
lin-41 LCS 1&2
lin-41 CDS

Figure 5. Key points of the bioinformatics analysis.

(a) Workflow of the sequence analysis of the wild-type (WT) and *alg-1*(-) CLIP-seq libraries. The number of elements after each filtering and processing stage is shown. (b) Three clusters from the *alg-1*(-) and wildtype (WT) samples overlap in the *act-5* gene, flagging the regions as not dependent on ALG-1 binding, and were removed from further analysis (grey boxes). (c) The number of genes that contain the corresponding number of unique CLIP-derived clusters.

a



c

Number of Clusters	Number of Genes
1	2,171
2	548
3	203
4	86
5	40
6	17
7	11
8	5
9	4
≥10	8
Total 4,806	Total 3,093

b

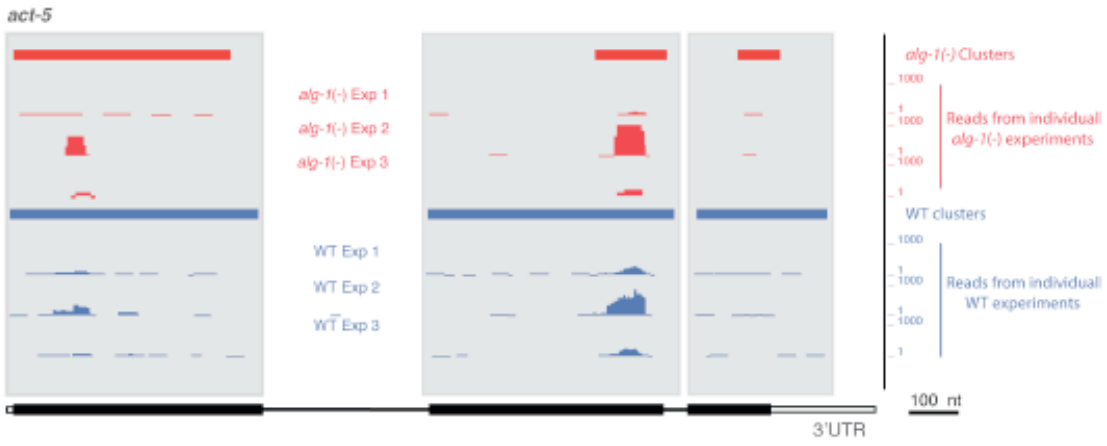
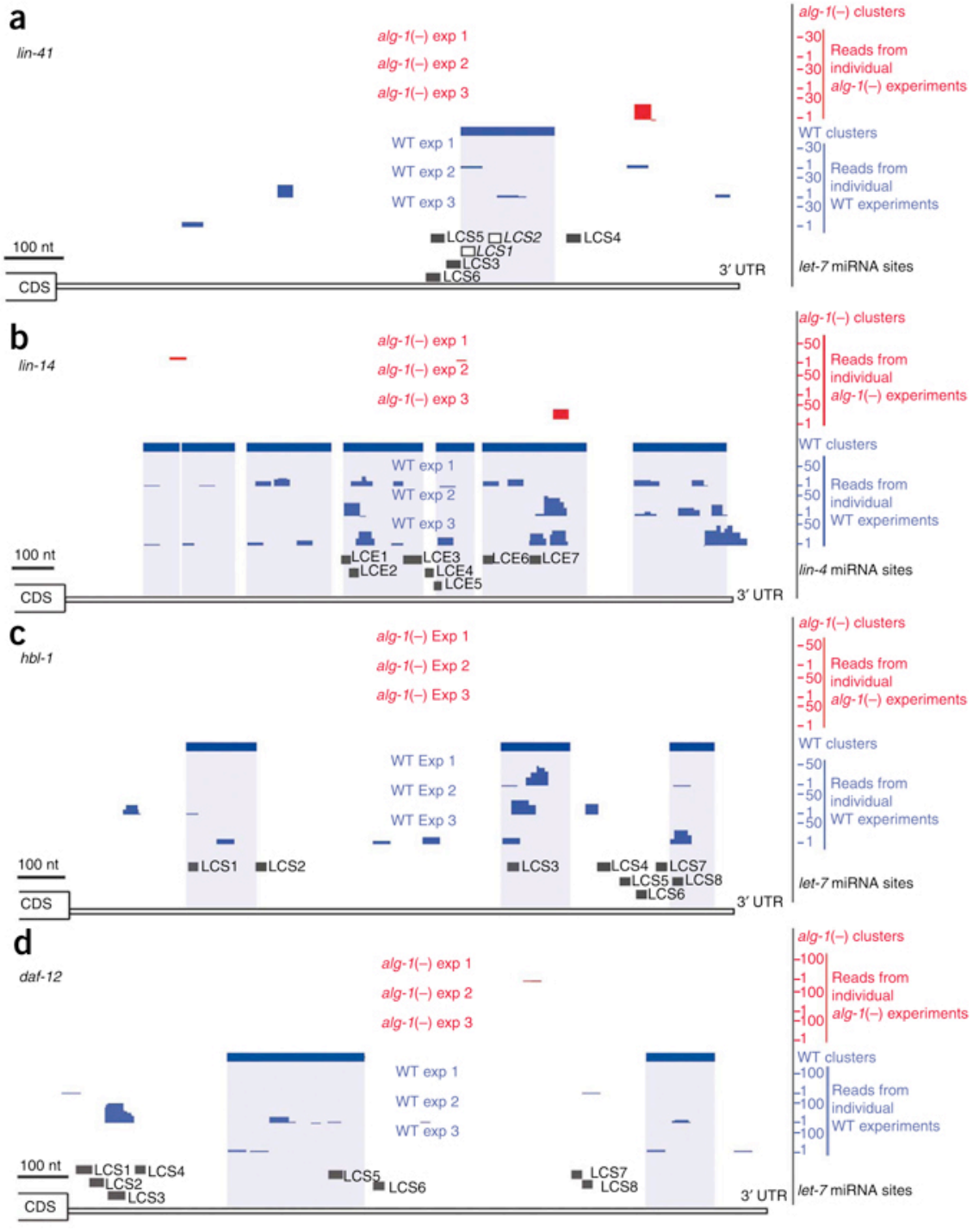


Figure 6. MicroRNA targets identified by ALG-1 CLIP-seq in L4-stage worms.

Graphical depictions of the number and location of reads from *alg-1*(–) (red upper tracks) and WT (blue lower tracks) from three biological replicates, CLIP-derived clusters (solid rectangular boxes over the reads) and putative miRNA binding sites in the 3′ UTR of mRNA transcripts (LCS, *let-7* complementary sequence; LCE, *lin-4* complementary element). (**a**) *lin-41* 3′ UTR. Of the six predicted LCSs, LCS1 and LCS2 (open boxes) have experimentally validated *let-7* sites (Reinhart et al., 2000; Slack et al., 2000; Vella et al., 2004). (**b**) *lin-14* 3′ UTR. Deletion of all the predicted LCE sites or of LCEs 1–5 results in misregulation of *lin-14* expression (Lee et al., 2003; Wightman et al.,1993). (**c**) *hbl-1* 3′ UTR (Abrahante et al., 2003; Lin et al., 2003; Abbott et al., 2005). The sites for *let-7* miRNA (LCSs 1–8) binding have been predicted but not experimentally tested. (**d**) *daf-12* 3′ UTR. Deletion of the predicted LCSs 1–4 or 5–8 in reporter constructs leads to misregulation of reporter gene expression (Grosshans et al., 2005).


a
lin-41
alg-1(–) exp 1
alg-1(–) exp 2
alg-1(–) exp 3
WT exp 1
WT exp 2
WT exp 3
LCS5
LCS2
LCS1
LCS3
LCS6
LCS4
100 nt
CDS
3′ UTR
alg-1(–) clusters
Reads from individual alg-1(–) experiments
WT clusters
Reads from individual WT experiments
let-7 miRNA sites
b
lin-14
alg-1(–) exp 1
alg-1(–) exp 2
alg-1(–) exp 3
WT exp 1
WT exp 2
WT exp 3
LCE1
LCE2
LCE3
LCE4
LCE5
LCE6
LCE7
100 nt
CDS
3′ UTR
alg-1(–) clusters
Reads from individual alg-1(–) experiments
WT clusters
Reads from individual WT experiments
lin-4 miRNA sites
c
hbl-1
alg-1(–) Exp 1
alg-1(–) Exp 2
alg-1(–) Exp 3
WT Exp 1
WT Exp 2
WT Exp 3
LCS1
LCS2
LCS3
LCS4
LCS7
LCS5
LCS8
LCS6
100 nt
CDS
3′ UTR
alg-1(–) clusters
Reads from individual alg-1(–) experiments
WT clusters
Reads from individual WT experiments
let-7 miRNA sites
d
daf-12
alg-1(–) exp 1
alg-1(–) exp 2
alg-1(–) exp 3
WT exp 1
WT exp 2
WT exp 3
LCS1
LCS4
LCS2
LCS3
LCS5
LCS6
LCS7
LCS8
100 nt
CDS
3′ UTR
alg-1(–) clusters
Reads from individual alg-1(–) experiments
WT clusters
Reads from individual WT experiments
let-7 miRNA sites

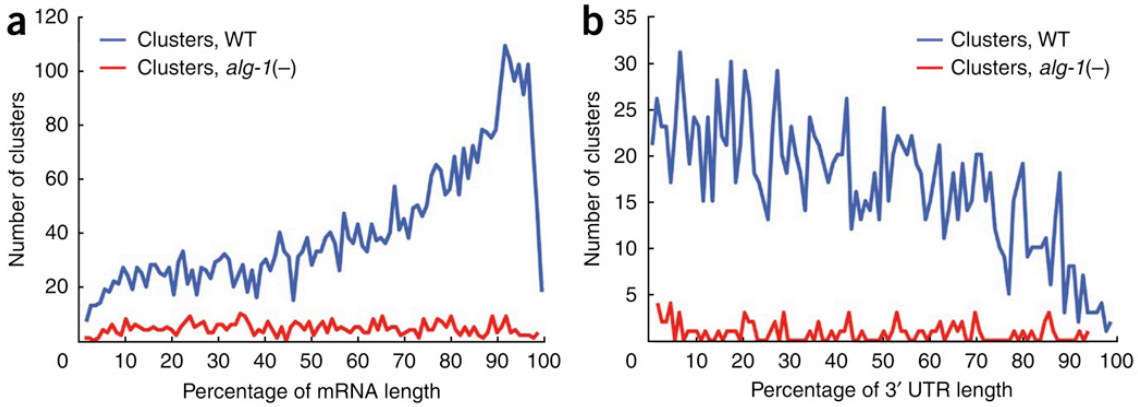


Figure 7. Relative position of ALG-1 binding sites across protein-coding genes.

(a) Distribution of WT (blue) and *alg-1*(−) (red) clusters across a composite mRNA length. Cluster position is depicted as a percentage of the gene region, from the beginning to the end of spliced transcripts. (b) Distribution of WT (blue) and *alg-1*(−) (red) clusters across the 3′ UTR region. Cluster position is depicted as a percentage of the region from the annotated translational stop codon to the end of transcripts, as defined by our reannotation of *C. elegans* genes.

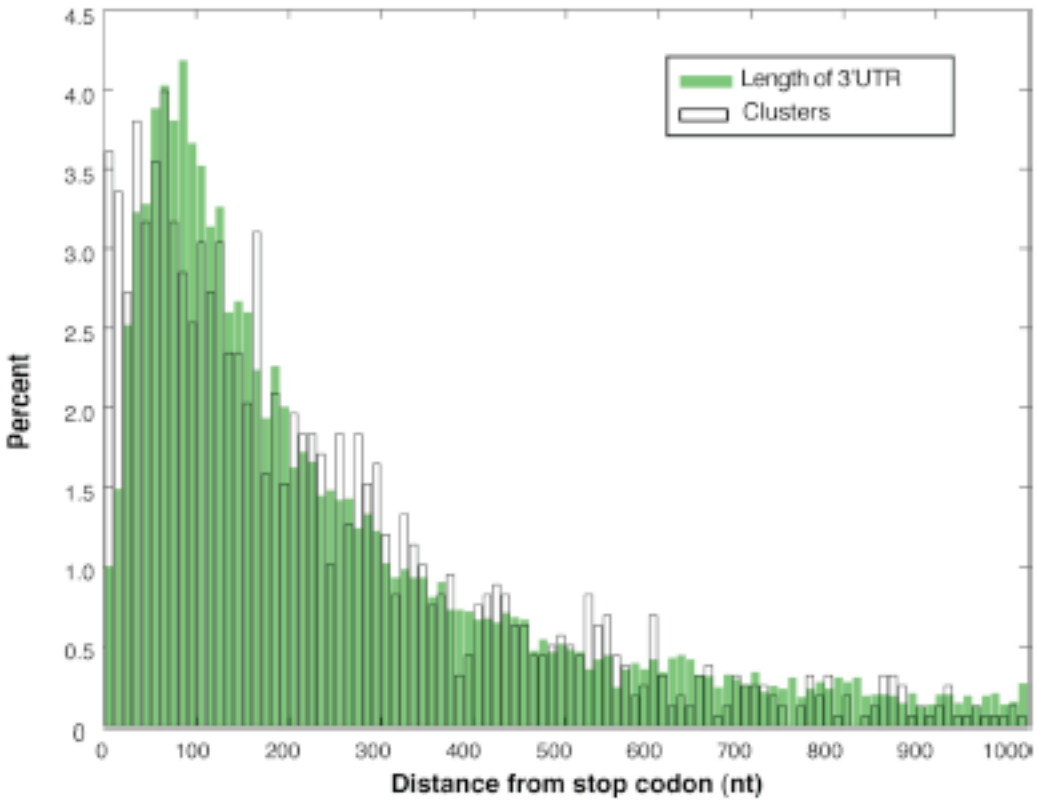


Figure 8. Cluster locations and 3′UTR lengths relative to the translation stop codon.

The x-axis of the histogram depicts the distance in nucleotides from the stop codon for all protein-coding genes with at least one CDC and the y-axis depicts the fraction of either clusters (open bars) or 3′UTR ends (green bars).

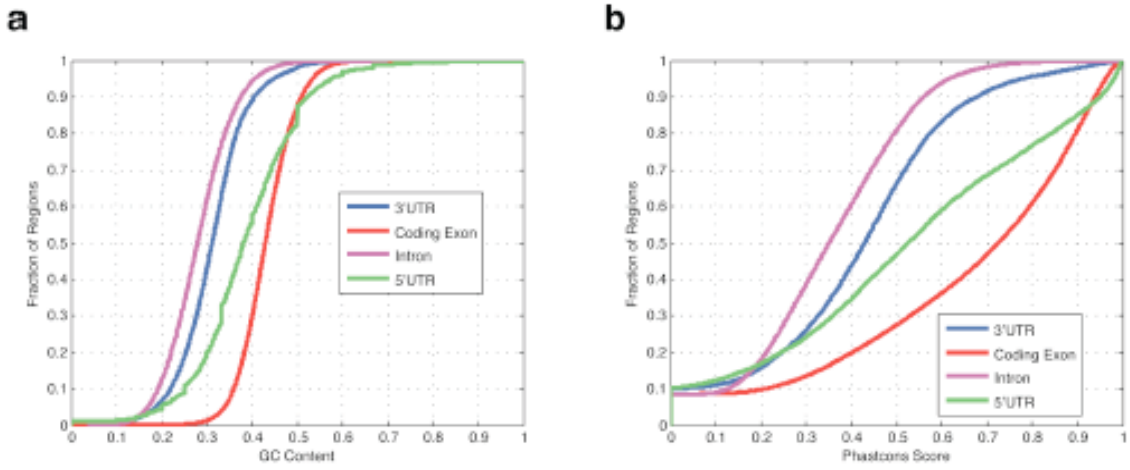


Figure 9. Criteria for Random-derived cluster (RDC) selection.

Cumulative distribution functions representing GC content (a) and conservation as determined by Phastcons scores 3 among 5 nematode species (b) of 5′UTRs, 3′UTRs, coding exons and introns in the *C.elegans* genome.

Figure 10. Attributes enriched in ALG-1 binding sites within 3′ UTRs.

(**a**) Box plots of the conservation levels measured as the fraction of perfectly conserved nucleotides between genome-wide alignments of *C. elegans* and HG004659 and GM084317 *brenneri* in CLIP-derived clusters (CDCs) and randomly derived clusters (RDCs). CDCs are significantly more conserved than RDCs as assessed by the Kolgomorov-Smirnov two-sample test ($P < 10^{-36}$). (**b**) Box plots of RNA accessibility, measured as the average probability of being unpaired, of CDC and RDC and their corresponding flanking sequences (100 nt upstream or downstream). CDCs and flanking sequences are significantly more accessible than RDCs in the same locations as assessed by the Kolgomorov-Smirnov two-sample test ($P < 10^{-10}$). (**c**) The ten most enriched *k*-mers ($k = 5, 6, 7$) within or 100 nt upstream or downstream of CDCs, compared to RDCs, are shown along with the range of *Z*-scores for the specific categories. (**d–g**). The number of conserved hexamers within CDCs (solid line) and RDCs (dashed line) that base-pair to miRNA or scrambled miRNA regions (dotted line), allowing for zero (orange) or only one G•U base pair (black). Error bars in dashed and dotted lines represent the s.d. among ten independent sets of RDCs and scrambled miRNAs, respectively. Hexamers within 3′-UTR CDCs and RDCs (**d**) or coding-exon CDCs and RDCs (**e**) that base-pair to cloned miRNAs or shuffled versions of cloned miRNAs. Hexamers within 3′-UTR CDCs and RDCs that base-pair to the *let-7* or shuffled *let-7* miRNA (**f**) and *lin-4* or shuffled *lin-4* miRNA (**g**). Regions of the miRNA(s) that have statistically enriched numbers of complementary hexamers within CDCs when compared to RDCs or shuffled miRNAs are denoted by * ($P < 0.01$) and ** ($P < 10^{-6}$) as measured by a *Z*-test.

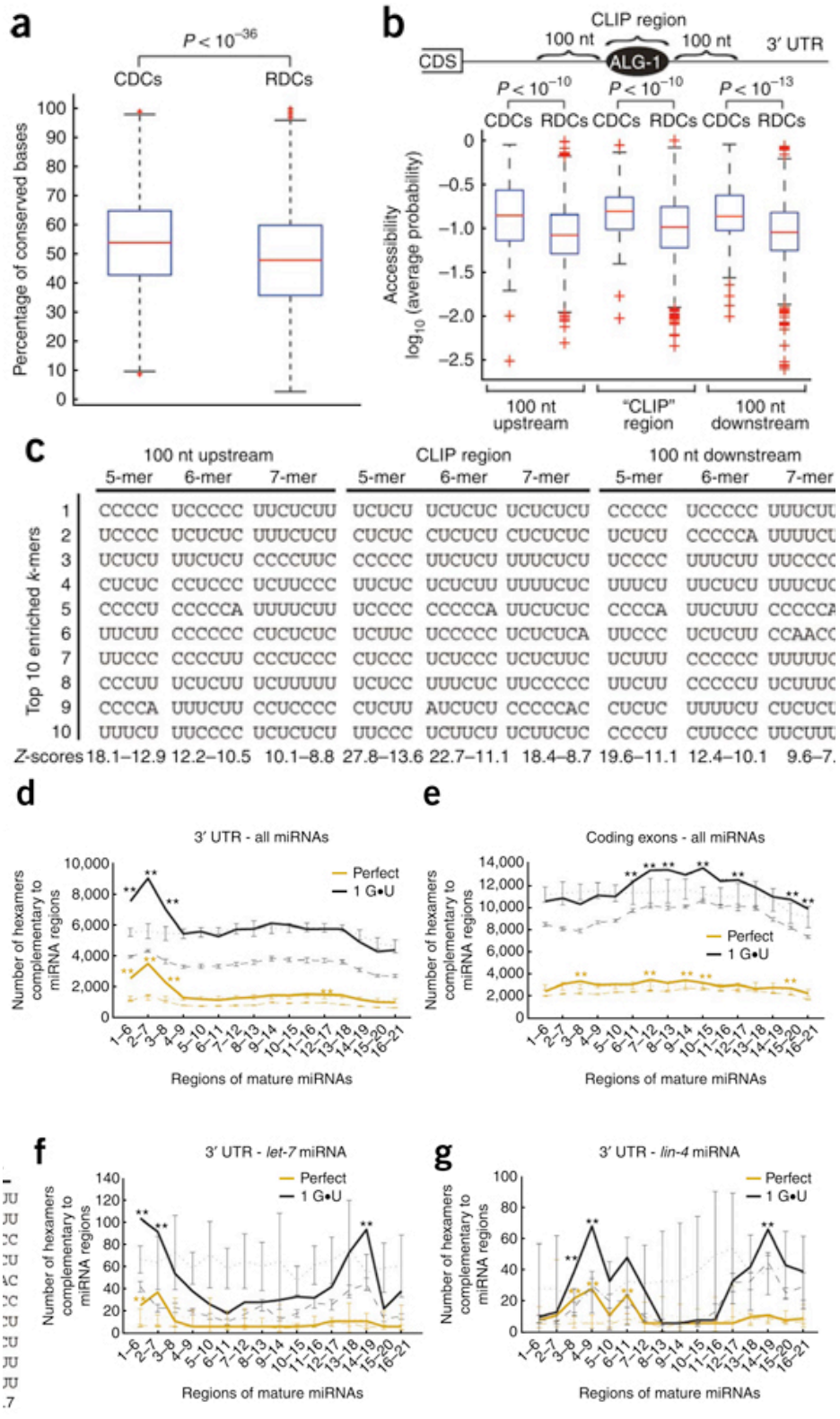
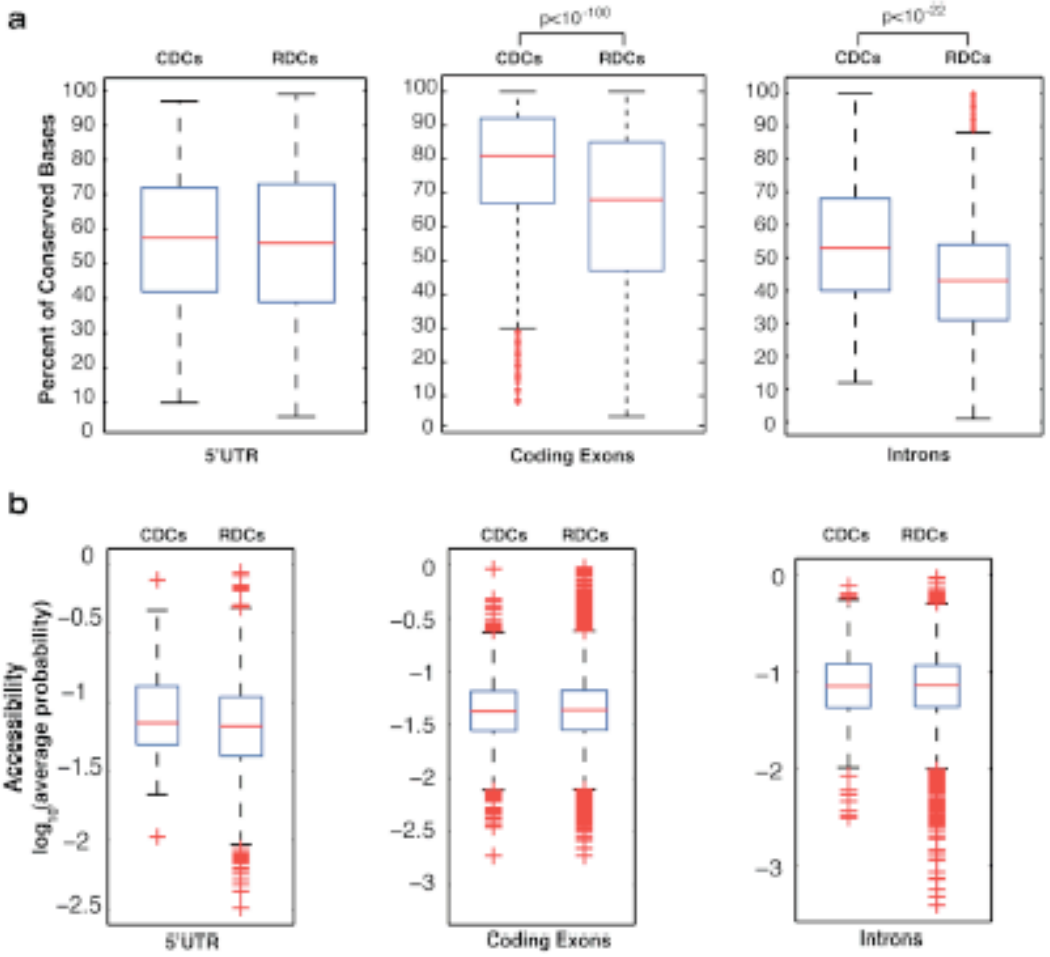

a
P < 10−36
CDCs
RDCs
Percentage of conserved bases
b
CLIP region
100 nt
100 nt
3′ UTR
CDS
ALG-1
P < 10−10
P < 10−10
P < 10−13
CDCs RDCs CDCs RDCs CDCs RDCs
Accessibility
log10 (average probability)
100 nt upstream
"CLIP" region
100 nt downstream
c
100 nt upstream
CLIP region
100 nt downstream
5-mer 6-mer 7-mer 5-mer 6-mer 7-mer 5-mer 6-mer 7-mer
Top 10 enriched k-mers
Z-scores 18.1–12.9 12.2–10.5 10.1–8.8 27.8–13.6 22.7–11.1 18.4–8.7 19.6–11.1 12.4–10.1 9.6–7.
d
3′ UTR - all miRNAs
Perfect
1 G•U
Number of hexamers complementary to miRNA regions
Regions of mature miRNAs
e
Coding exons - all miRNAs
Perfect
1 G•U
Number of hexamers complementary to miRNA regions
Regions of mature miRNAs
f
3′ UTR - let-7 miRNA
Perfect
1 G•U
Number of hexamers complementary to miRNA regions
Regions of mature miRNAs
g
3′ UTR - lin-4 miRNA
Perfect
1 G•U
Number of hexamers complementary to miRNA regions
Regions of mature miRNAs

Figure 11. Attributes enriched in CLIP-derived clusters (CDCs) present in the 5′UTR, coding exons and introns of ALG-1-bound genes.

(a) Conservation levels of CDCs and RDCs. Box-plots depicting the distributions of the log (base 10) probability of being unpaired, with the central mark indicating the median, the edges of the box are the 25th and 75th percentiles and the whiskers that extend to the most extreme datapoints that the algorithm considers not to be outliers, and the outliers are plotted individually. Conservation levels are measured between *C. elegans* and *C. brenneri*. CDCs in coding exon and introns are significantly conserved ($p<10^{-22}$ by the Kolgomorov-Smirnov two-sample test). (b) Accessibility of ALG-1-bound regions. CDCs in the 5′UTR, coding exons or introns are not statistically different than RDCs in the same regions. (c) The ten most enriched *k*-mers (k = 5, 6, 7) within CDCs, compared to RDCs. Z-score range represents maximum and minimum Z-scores for the *k*-mers shown.



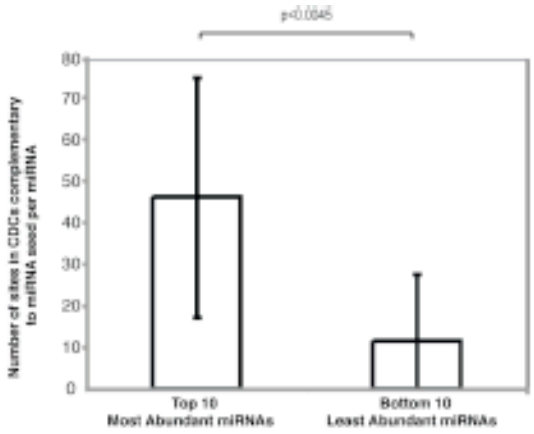
c

Top 10 Enriched k-mers	5'UTR 5-mers	5'UTR 6-mers	5'UTR 7-mers	Coding Exons 5-mers	Coding Exons 6-mers	Coding Exons 7-mers	Introns 5-mers	Introns 6-mers	Introns 7-mers
1	AGAAG	UCUUCU	GAAUGGU	CCAGG	CCCAGG	CCCAGGA	UCUCU	GGUAAG	GUGUGUG
2	UGCUG	UAGGCU	UUAGGCU	CCCAG	CCAGGA	CCCCAGG	CGUCG	ACGUCG	UGGUAAG
3	CUUCU	UGCUUC	UAGGCUU	CAGGA	CCCCAG	GCCCCAG	UCCUC	GCCGUC	UGUGUGU
4	CUGCU	UACGCU	CCGUCAG	AGGUA	AGCCCC	UUUUUUU	UCAUC	GUCGUC	GGUGAGU
5	UCUUC	GCCACC	CAUCAUA	GCCCC	GCCCCA	AGCCCCA	GAUCG	GUGAGU	CACGUCG
6	GUCUU	CGAAUG	AGGCUUA	CCCCA	UUUUUU	GAGCCCC	UCGUC	CGUCGU	GGUAAGA
7	GAAGG	GUUGAG	UAUGAUC	GUAAG	GUAAGU	GGAGCCC	AGGAG	AGGAGA	GUCGUCG
8	ACAGG	AAGAAG	GGCUUAG	ACCAG	AGGUAA	GUAAGUU	CAAGG	CAGGUG	CAAGUCG
9	CGGGU	CACAGG	UACGCUU	CCAAG	GCCAAG	CCAGGAG	CAGGU	UGUGUG	CAGAUCG
10	AUGGU	CGGGUA	CCAACUG	AGGAC	GAGCCC	AGGAGCC	CGAUC	UGGUAG	GGAGGAG
Z-scores	5.4-3.3	5.8-4.4	6.5-5.5	29.7-15.8	32.7-16.6	35.8-19.2	8.7-6.9	10.2-6.5	8.3-6.6

Figure 12. Number of seed matches in 3′UTR CDCs.

(a).The number of seed matches for the most and least abundant annotated miRNAs. The top and bottom ten miRNAs ranked by expression, were analyzed separately for seed pairing to 3′UTR CDCs. The height of each bar represents the median value for the top/bottom ten miRNAs, the error bars represent the standard deviation. The top ten cloned miRNAs exhibit a median of 46 conserved hexamers that are perfectly complementary to bases 2-7 of the microRNAs, relative to a median of 12 conserved hexamers for the 10 least cloned ones ($p<0.0045$) with a Wilcoxon rank-sum test). The top/bottom ten miRNAs and the number of reads and seed matches are shown in the table. (b) Scatter-plot showing the correlation of the conserved seed matches (bases 2-7) of all the cloned miRNAs in the 3' UTR CDCs and the cloning frequency of these miRNAs (see Appendix 1). Each dot on the plot represents an individual miRNA.

a



Top 10 Most Abundant miRNAs	Number of reads	Number of seed matches
miR-58	303168	93
miR-52	22101	50
miR-71	16512	63
miR-80	10536	93
miR-228	9949	23
miR-1	8799.5	84
miR-79	8381	42
miR-84	5673	34
miR-244	4573	15
miR-229	4108	34

Bottom 10 Least Abundant miRNAs	Number of reads	Number of seed matches
miR-56	1	0
miR-785	3	55
miR-242	3	5
miR-793	7	25
miR-1820	8	12
miR-1821	8	8
miR-799	12	5
miR-41	17	24
miR-251	19	11
miR-39,miR-40	19	24

b

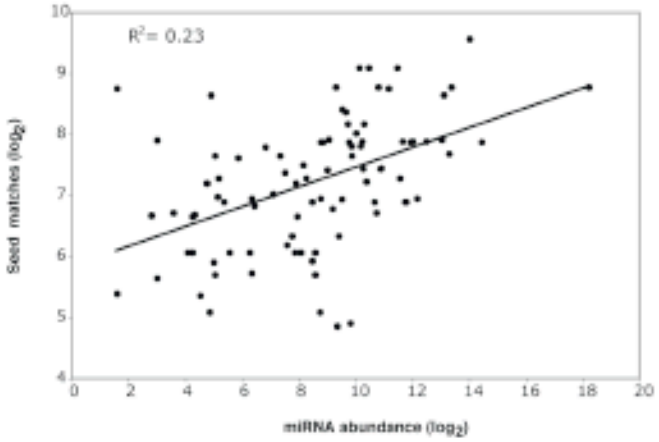
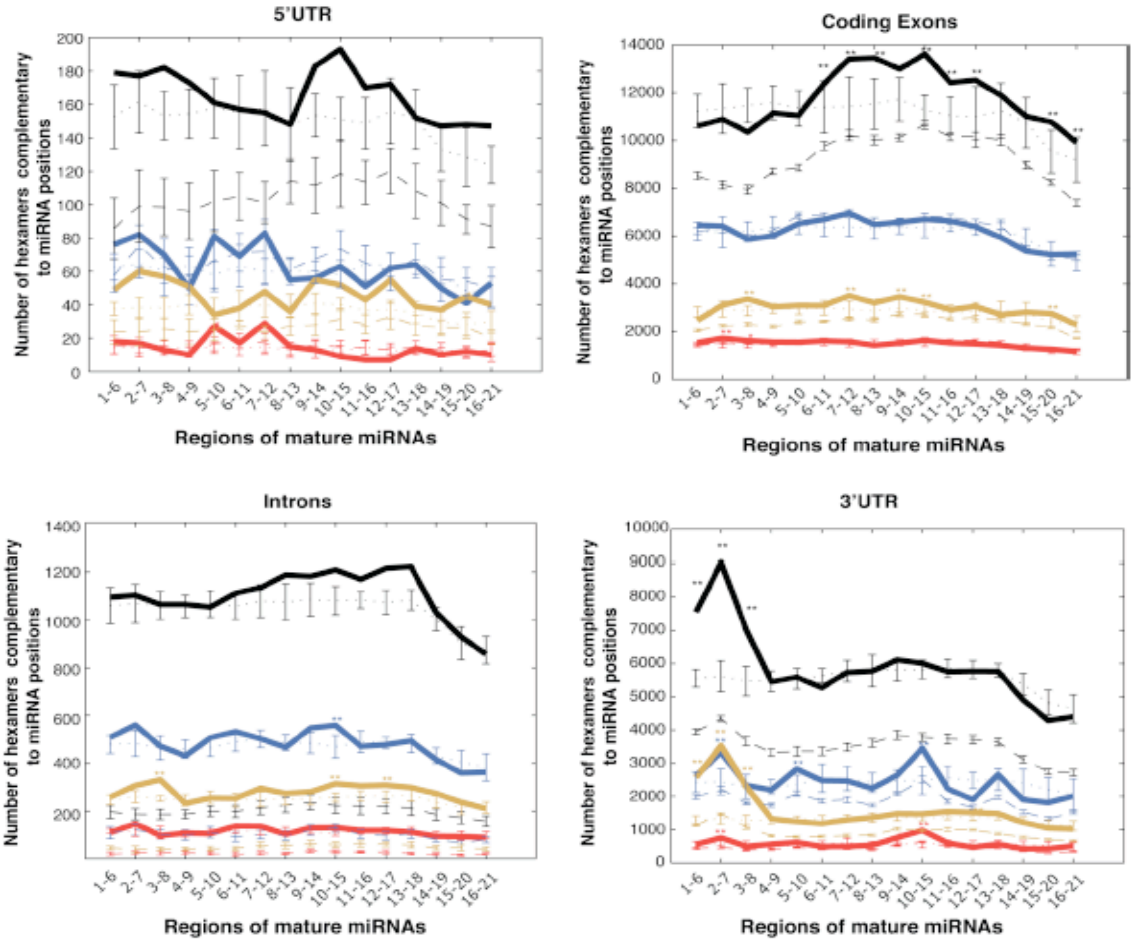


Figure 13. Number of conserved hexamers in 5′UTR, coding exons, introns and 3′UTR CDCs that are complementary to regions of cloned or scrambled mature miRNAs.

CDC (solid line) and RDC (dashed line) hexamers that pair to cloned miRNA positions or scrambled miRNA positions (dotted line). Perfectly conserved base-pairing with no G:U base-pairs allowed (orange); "Semi-conserved" base-pairing with G:U base-pairing allowed only in *C. brenneri* (red); Perfectly conserved base-pairing allowing for one conserved G:U base-pair (black); and "Semi-conserved" base-pairing with G:U base-pairing allowed only in *C. elegans*. Error bars in dashed and dotted lines represent the standard deviation among ten independent sets of RDCs and scrambled miRNAs, respectively (*: $p<.01$; **: $p<10^{-6}$).


5'UTR
Coding Exons
Introns
3'UTR
Number of hexamers complementary to miRNA positions
Regions of mature miRNAs
1-6
2-7
3-8
4-9
5-10
6-11
7-12
8-13
9-14
10-15
11-16
12-17
13-18
14-19
15-20
16-21

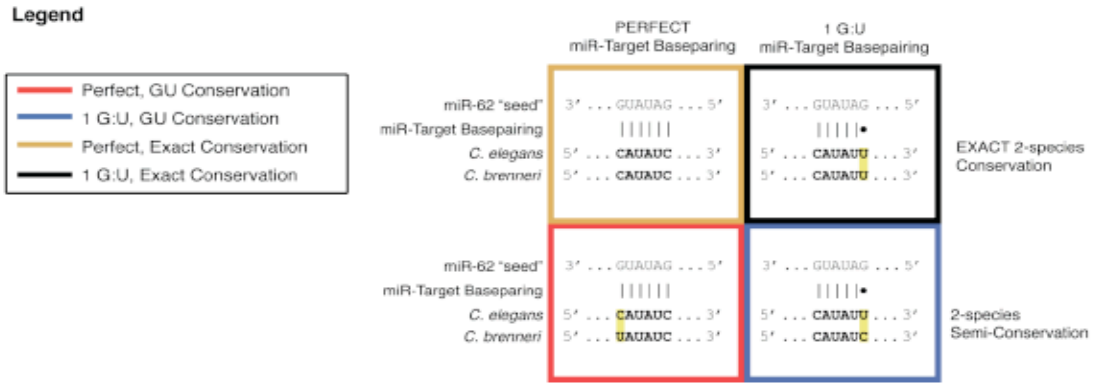

Legend
Perfect, GU Conservation
1 G:U, GU Conservation
Perfect, Exact Conservation
1 G:U, Exact Conservation
PERFECT
miR-Target Baseparing
1 G:U
miR-Target Basepairing
miR-62 "seed"
miR-Target Baseparing
C. elegans
C. brenneri
3' . . . GUAUAG . . . 5'
5' . . . CAUAUC . . . 3'
5' . . . UAUAUC . . . 3'
5' . . . CAUAUU . . . 3'
EXACT 2-species Conservation
2-species Semi-Conservation

Figure 14. Relationship between ALG-1 binding and mRNA expression levels.

(a) Effects of ALG-1 binding on mRNA levels. Box plots representing the differential expression (as a *t*-statistic) of genes from biological replicate microarray experiments comparing *alg-1*(−) to WT L4-stage worms. Genes are divided into those that contained no CDCs and those that contained CDCs only within 3′ UTRs or coding exons. Compared to genes with no CDCs or coding-exon CDCs, genes with 3′-UTR CDCs are significantly more upregulated in *alg-1*(−) relative to WT as assayed by the Wilcoxon rank-sum test ($P < 10^{-4}$). (b) Functional enrichment of genes that have CDCs only within 3′ UTR or coding exons that are up- or downregulated in *alg-1*(−) worms using significantly enriched ($P < 0.05$ in at least one row; Holm-Bonferroni corrected) functional categories defined by the *C. elegans* Topomap algorithm (Kim et al., 2001). The intensity on the heat-map denotes $-\log_{10}(p$ value). Genes represented by these functional categories can be divided in a matrix (right) depending on the location of the CDCs (3′ UTRs or coding exons), and whether the genes are up- or downregulated in the *alg-1*(−) mutants relative to WT worms. Several categories occupy multiple cells in the matrix, for example “Cell structure,” “Collagen,” “Cell adhesion,” “Protein expression,” “RNA binding” and “Germ line–enriched.” (c) UCSC Genome Browser view depicting clusters in the 3′ UTR of the *alg-1* gene (blue, WT clusters; red, *alg-1*(−) clusters, none present) and the predicted miRNA binding sites by the various algorithms.

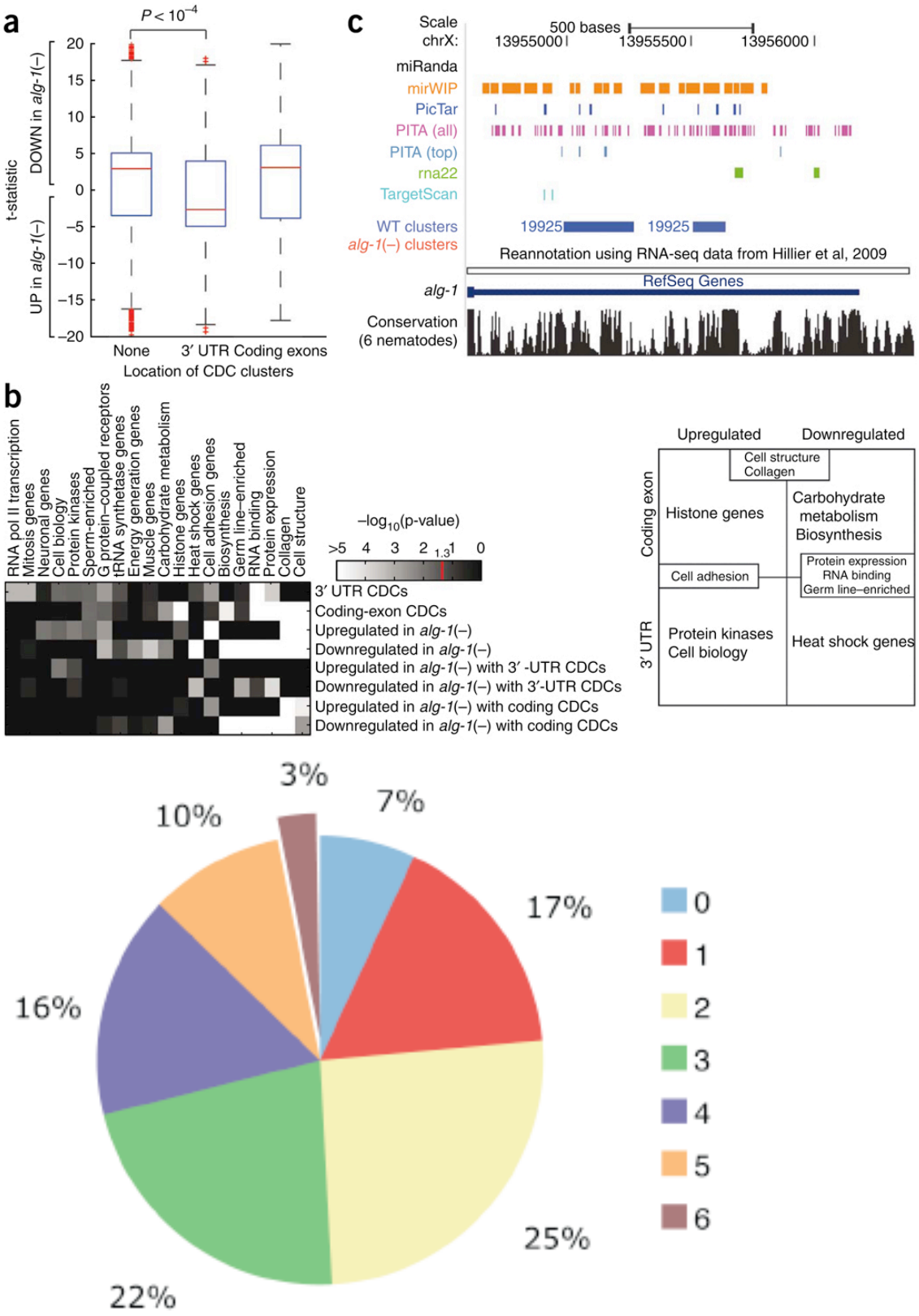

a
P < 10−4
t-statistic
DOWN in alg-1(–)
UP in alg-1(–)
20
15
10
5
0
−5
−10
−15
−20
None
3′ UTR
Coding exons
Location of CDC clusters
c
Scale
chrX:
500 bases
13955000
13955500
13956000
miRanda
mirWIP
PicTar
PITA (all)
PITA (top)
rna22
TargetScan
WT clusters
alg-1(–) clusters
19925
19925
Reannotation using RNA-seq data from Hillier et al, 2009
RefSeq Genes
alg-1
Conservation
(6 nematodes)
b
RNA pol II transcription
Mitosis genes
Neuronal genes
Cell biology
Protein kinases
Sperm-enriched
G protein–coupled receptors
tRNA synthetase genes
Energy generation genes
Muscle genes
Carbohydrate metabolism
Histone genes
Heat shock genes
Cell adhesion genes
Biosynthesis
Germ line–enriched
RNA binding
Protein expression
Collagen
Cell structure
−log10(p-value)
>5 4 3 2 1.3 1 0
3′ UTR CDCs
Coding-exon CDCs
Upregulated in alg-1(–)
Downregulated in alg-1(–)
Upregulated in alg-1(–) with 3′ -UTR CDCs
Downregulated in alg-1(–) with 3′-UTR CDCs
Upregulated in alg-1(–) with coding CDCs
Downregulated in alg-1(–) with coding CDCs
Upregulated
Downregulated
Coding exon
3′ UTR
Cell structure
Collagen
Histone genes
Carbohydrate
metabolism
Biosynthesis
Cell adhesion
Protein expression
RNA binding
Germ line–enriched
Protein kinases
Cell biology
Heat shock genes
3%
7%
10%
17%
16%
25%
22%
0
1
2
3
4
5
6

Figure 15. Percent of 3′UTR CDCs with predicted miRNA target sites.

Pie chart depicting the number of clusters that simultaneously contain predicted miRNA binding sites by six published target prediction algorithms, namely miRanda, mirWIP, PicTar, PITA (ALL no flank and TOP no flank combined), rna22, TargetScan (Conserved and Non-conserved combined). Note only the presence of a predicted target site(s), but not necessarily the same site(s), within each 3′UTR CDC was analyzed.

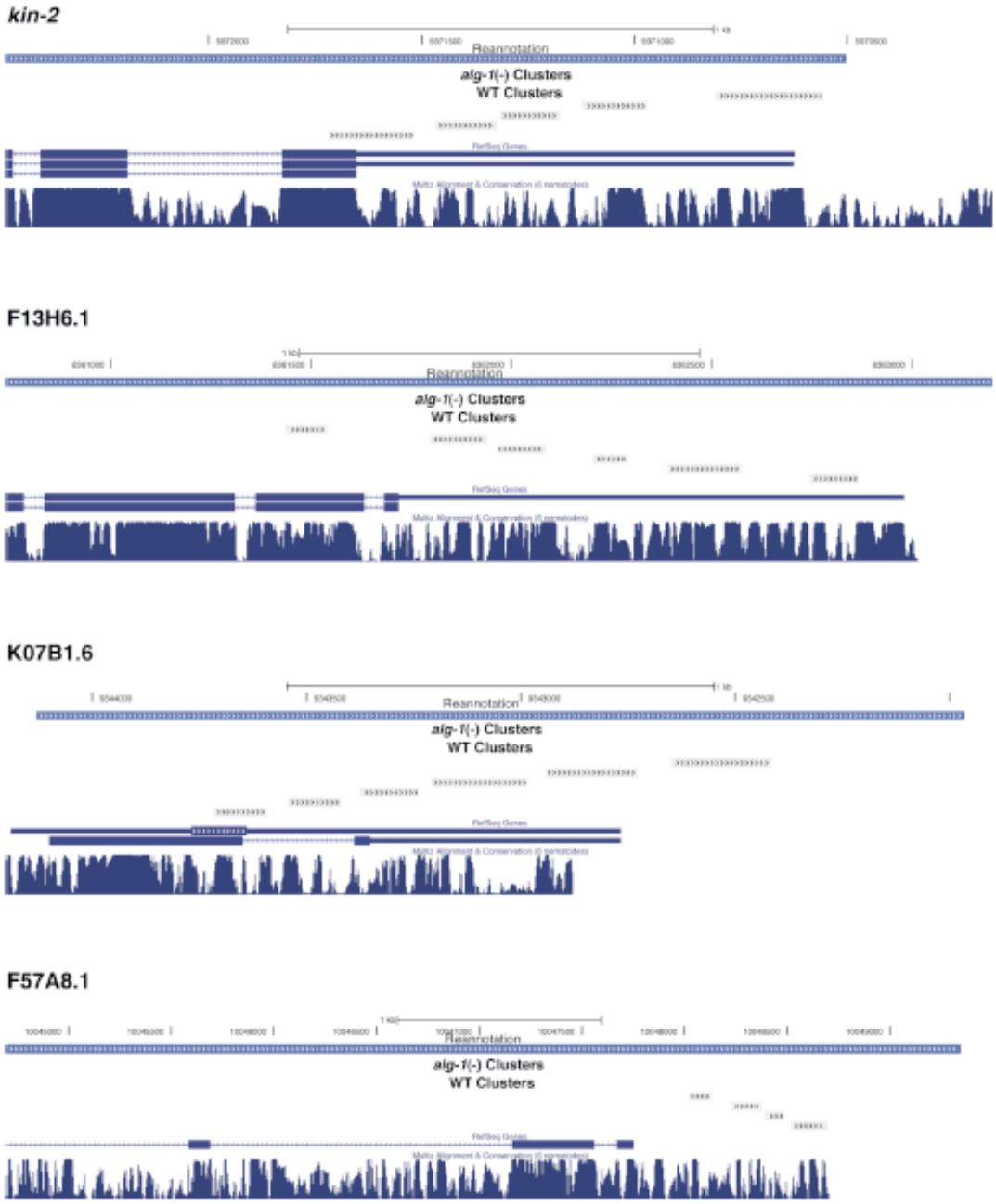


Figure 16. UCSC genome browser tracks depicting the genes with the highest number of clusters in 3′UTRs.

The tracks display the boundaries of the reannotated gene loci based on RNA-seq data from Hillier et al., 2009, WT clusters and *alg-1*(-) clusters (if present). The *lin-14* gene, which contains the most 3′UTR clusters, is displayed in Figure 4b.

TABLES

Table 1. ALG-1 CLIP-seq

[a] mRNA transcripts enriched (>0.5 average percentile rank) in AIN-1 or AIN-2 immunoprecipitations ($p<0.01$).
[b] t-statistic difference of >2.5 with $p<0.01$.
[c] genetic suppression (G) or GFP/lacZ reporter (R) evidence of post-transcriptional regulation.

R-1 : 3' UTR mediates post-transcriptional regulation and mutation of target site(s) alters regulation.

R-2 : 3' UTR regulation is lost in animals bearing mutations in the miRNA.

R-3 : miRNA and target site(s) are required for post-transcriptional regulation.

[#] regulation disrupted upon deletion of LCS1-4 or LCS5-8.
[d] detected by qRT-PCR in the AIN-2 IP.
[e] a region that includes the cognate miRNA site is covered by sequencing tags but only in one experiment, thus not achieving statistical significance.

Table 1. ALG-1 CLIP-seq

miRNA :: mRNA target	AIN-IP[a]	ALG-1 CLIP-seq		mRNA *alg-1* / WT[b]	Evidence for miRNA targeting[c]	References
		cluster in 3'UTR	cognate miRNA site			
lin-4 :: lin-14	●	●	●	UP	G/R-3	(Lee et al., 1993; Wightman et al., 1993)
lin-4 :: lin-28	●			UP	G/R-3	(Moss et al., 1997)
let-7/mir-48/84/241 :: hbl-1	●	●	●	UP	G/R-2	(Abbott et al., 2005; Abrahante et al., 2003; Lin et al., 2003)
let-7 :: lin-41	●[d]	●	●	UP	G/R-3	(Slack et al., 2000; Vella et al., 2004)
let-7/mir-84 :: let-60	●	●	●	NC	G/R-2	(Johnson et al., 2005)
let-7 :: daf-12	●	●	●	UP	G/R-2[#]	(Grosshans et al., 2005)
let-7 :: pha-4	●	●[e]	●[e]	UP	G/R-2	(Grosshans et al., 2005)
let-7 :: lss-4	●	●	●	UP	G/R-2	(Grosshans et al., 2005)
let-7 or *mir-273 :: die-1*	●	●	●	UP	G/R-3	(Chang et al., 2004; Grosshans et al., 2005)
let-7 :: nhr-25	●			UP	G	(Hayes et al., 2006)
let-7 :: T14B1.1	●	●	●	UP	R-1	(Lall et al., 2006)
lsy-6 :: cog-1		●		NC	G/R-3	(Johnson et al., 2003)
mir-61 :: vav-1		●	●	NC	G/R-3	(Yoo and Greenwald, 2005)

Table 2. ALG-1 CLIP-seq cluster information

[a] mRNA transcripts enriched (>0.5 average percentile rank) in AIN-1 or AIN-2 immunoprecipitations ($p<0.01$).

[b] t-statistic difference of >2.5 with $p<0.01$.

[c] genetic suppression (G) or GFP/lacZ reporter (R) evidence of post-transcriptional regulation.

R-0 : post-transcriptional regulation by 3' UTR not detected.

R-1 : 3' UTR mediates post-transcriptional regulation.

R-2 : 3' UTR mediates post-transcriptional regulation and mutation of target site(s) alters regulation.

R-3 : 3' UTR regulation is lost in animals bearing mutations in the miRNA.

R-4 : miRNA and target site(s) are required for post-transcriptional regulation.

[#] GFP fused to 3'UTR showed no regulation in *lsy-6*-expressing cells (ASEL) compared to ASER.

[$] GFP reporter fused to 3'UTR was not expressed compared to *unc-54* 3'UTR control.

[&] miRNA target site mutation did not affect regulation.

[^] regulation disrupted upon deletion of LCS1-4 or LCS5-8.

[d] detected by qRT-PCR in the AIN-2 IP.

[e] a region that includes the cognate miRNA site is covered by sequencing tags but only in one experiment, thus not achieving statistical significance.

[f] microarray probes in the coding sequence show up-regulation; in the 3'UTR they show down-regulation.

Table 2. ALG-1 CLIP-seq cluster information

miRNA :: mRNA target	**AIN-IP[a]**	**ALG-1 CLIP-seq cluster**		**mRNA[b]**	**Evidence for miRNA targeting[c]**	**References**
		in 3'UTR	**includes miRNA site**	***alg-1* / WT**		
lin-4 :: lin-14	●	●	●	UP	G/R-4	(Lee et al., 1993 ; Wightman et al., 1993)
lin-4 :: lin-28	●			UP	G/R-4	(Moss et al., 1997)
let-7/mir-48/84/241 :: hbl-1	●	●	●	UP	G/R-3	(Abbott et al., 2005; Abrahante et al., 2003; Lin et al., 2003)
let-7 :: C35E7.4				UP	R-2[$]	(Lall et al., 2006)
let-7 :: ccr-4		●	●	UP	R-0	(Lall et al., 2006)
let-7 :: ceh-16				UP	R-1[$]	(Lall et al., 2006)
let-7 :: daf-12	●	●	●	UP	G/R-3[^]	(Grosshans et al., 2005)
let-7 :: daf-16	●	●	●	UP	R-0	(Lall et al., 2006)
let-7 :: dpy-2		●	●	UP	R-0	(Lall et al., 2006)
let-7 :: lin-41	●[d]	●	●	UP	G/R-4	(Slack et al., 2000; Vella et al., 2004)
let-7 :: lss-4	●	●	●	UP	G/R-3	(Grosshans et al., 2005)

Table 2 Continued. ALG-1 CLIP-seq cluster information

miRNA :: mRNA target	**AIN-IP[a]**	**ALG-1 CLIP-seq cluster**		**mRNA[b]**	**Evidence for miRNA targeting[c]**	**References**
		in 3'UTR	**includes miRNA site**	***alg-1* / WT**		
let-7 :: nhr-17	●	●	●	DOWN	R-0	(Lall et al., 2006)
let-7 :: nhr-25	●			UP	G	(Hayes et al., 2006)
let-7 :: nhr-4	●			NC	R-1	(Lall et al., 2006)
let-7 :: oig-2				NC	R[$&]	(Lall et al., 2006)
let-7 :: pha-4	●	●[e]	●[e]	UP	G/R-3	(Grosshans et al., 2005)
let-7 :: sma-1				UP	R-0	(Lall et al., 2006)
let-7 :: T14B1.1	●	●	●	UP	R-2	(Lall et al., 2006)
let-7 :: uba-1	●	●		UP	R-1[$]	(Lall et al., 2006)
let-7 :: unc-129				NC	R-1	(Lall et al., 2006)
let-7 or *mir-273 :: die-1*	●	●	●	UP	G/R-4	(Chang et al., 2004; Grosshans et al., 2005)
let-7/mir-84 :: let-60	●	●	●	NC	G/R-3	(Johnson et al., 2005)
lsy-6 :: cog-1		●		NC	G/R-4	(Johnson et al., 2003)
lsy-6 :: C02B8.4	●			NC	R[#]	(Didiano and Hobert, 2006)
lsy-6 :: C27H6.3				DOWN	R[#]	(Didiano and Hobert, 2006)

Table 2 Continued. ALG-1 CLIP-seq cluster information

miRNA :: mRNA target	AIN-IP[a]	ALG-1 CLIP-seq cluster in 3'UTR	ALG-1 CLIP-seq cluster includes miRNA site	mRNA[b] *alg-1* / WT	Evidence for miRNA targeting[c]	References
lsy-6 :: C48D5.2	●			UP	R[#]	(Didiano and Hobert, 2006)
lsy-6 :: F40H3.4				NC	R[#]	(Didiano and Hobert, 2006)
lsy-6 :: F55G1.12				NC	R[#]	(Didiano and Hobert, 2006)
lsy-6 :: F59A6.1				NC	R[#]	(Didiano and Hobert, 2006)
lsy-6 :: R07E3.5	●			NC	R[#]	(Didiano and Hobert, 2006)
lsy-6 :: T04C9.2	●			NF	R[#]	(Didiano and Hobert, 2006)
lsy-6 :: T05C12.8				NC	R[#]	(Didiano and Hobert, 2006)
lsy-6 :: T14G12.2				NC	R[#]	(Didiano and Hobert, 2006)
lsy-6 :: T20G5.9				NC	R[#]	(Didiano and Hobert, 2006)
lsy-6 :: T23E1.1				NC	R[#]	(Didiano and Hobert, 2006)
lsy-6 :: ZK637.13				NC	R[#]	(Didiano and Hobert, 2006)
mir-61 :: vav-1		●	●	NC	G/R-4	(Yoo and Greenwald, 2005)
mir-1 :: mef-2	●	●	●[e]	UP/DOWN[f]	R-3	(Simon et al., 2008)
mir-1 :: unc-29	●			NC	R-3[&]	(Simon et al., 2008)

Table 3. Genes connected to miRNA function by proteomic evidence from Zhang et al, 2007.
[a] t-statistic difference of >2.5 with p<0.01

Locus	Gene name	mRNA[a] *alg-1* / WT	ALG-1 CLIP-seq clusters	Description
C06G1.4	*ain-1*	UP	5	ALG-1 INteracting protein family member (ain-1)
F48F7.1	*alg-1*	DOWN	3	Argonaute (plant)-Like Gene family member
T07D3.7	*alg-2*	UP	0	Argonaute (plant)-Like Gene family member
B0041.2	*ain-2*	UP	2	ALG-1 INteracting protein family member (ain-2)
Y106G6H.2	*pab-1*	NC	5	PolyA Binding protein family member (pab-1)
Y55B1AR.1	*lec-6*	NC	0	gaLECtin family member (lec-6)
Y18D10A.17	*car-1*	NC	3	Cytokinesis, Apoptosis, RNA-associated family
H28O16.1		DOWN	5	ATP synthase alpha and beta subunits, ATP synthase Alpha chain, C terminal
C02A12.4	*lys-7*	UP	0	LYSozyme family member (lys-7)
Y73B6BL.6	*sqd-1*	DOWN	4	homologous to Drosophila SQD (squid) protein
F18H3.3	*pab-2*	UP	2	PolyA Binding protein family member (pab-2)
T23G11.3	*gld-1*	DOWN	2	defective in Germ Line Development family member, RNA binding
Y37E3.7	*rla-1*	NC	0	Ribosomal protein, Large subunit, Acidic (P1)
K07H8.6b	*vit-6*	DOWN	0	VITellogenin structural genes (yolk protein
B0403.4	*tag-320*	UP	2	protein disulfide-isomerase family member
C07A12.4	*pdi-2*	UP	4	Protein Disulfide Isomerase family member
C05E4.9	*gei-7*	UP	6	GEX Interacting protein family member (gei-7) (isocitrate lyase)
C12C8.1	*hsp-70*	DOWN	1	heat shock protein 70
F44E5.5		NF	2	Heat shock hsp70 proteins
Y41E3.10		DOWN	3	Elongation factor 1 (beta\/delta chain)

Table 3 Continued. Genes connected to miRNA function by proteomic evidence from Zhang *et al*, 2007.
[a] t-statistic difference of >2.5 with p<0.01

Locus	**Gene name**	**mRNA[a] *alg-1* / WT**	**ALG-1 CLIP-seq clusters**	**Description**
M88.5a		DOWN	5	hnRNP K
F10G7.2	*tsn-1*	NC	3	Tudor Staphylococcal Nuclease homolog family, RISC component
ZK455.1	*aco-1*	DOWN	2	ACOnitase family member (aco-1)
M110.4	*ifg-1*	DOWN	7	eukaryotic initiation factor
B0393.1	*rps-0*	DOWN	2	40S ribosomal protein
C09D4.5	*rpl-19*	NC	2	60S ribosomal protein L19
C44B12.5		NC	0	
C04C3.3		DOWN	2	pyruvate dehydrogenase
C44B7.10		NC	3	
R11A5.4		DOWN	4	phosphoenolpyruvate carboxykinase
F54D8.3	*alh-1*	NC	1	Aldehyde dehydrogenase family member
F01G10.1		DOWN	0	transketolase
F21F8.7	*asp-6*	UP	2	aspartic protease.
Y22F5A.4	*lys-1*	UP	3	lysozyme family member
C17G10.5	*lys-8*	UP	0	lysozyme family member
R12H7.2	*asp-4*	UP	0	aspartyl protease
F43D9.4	*sip-1*	UP	0	Heat shock hsp20 proteins
Y22D7AL.5		DOWN	7	Heat Shock Protein family member
C37H5.8	*hsp-6*	DOWN	3	heat shock 70 protein member

Table 3 Continued. Genes connected to miRNA function by genetic evidence from Parry *et al*, 2007.
[a] t-statistic difference of >2.5 with p<0.01.

Locus	Gene name	mRNA[a] *alg-1* / WT	ALG-1 CLIP-seq clusters (.5)	Brief Description
K12H4.8	*dcr-1*	NC	1	dead box helicase, dicer related family member
Y59A8B.1	*dpy-21*	UP	0	dosage compensation complex protein
F32B6.3		DOWN	0	Human HPRP18 protein like
F02E9.2	*lin-28*	UP	0	cold shock domain
W04D2.6		NC	0	ortholog of human RBM25, RNA recognition domain
Y34D9A.4	*spd-1*	DOWN	0	SPindle Defective family member, homolog of human PRC
W09C5.2	*unc-59*	NC	0	Cell division protein
ZC64.3	*ceh-18*	UP	1	*C.elegans* Homeobox family member (ceh-18)
F13H6.1		NC	7	
H20J04.3		UP	0	
C50E10.4	*sop-2*	NC	0	SAM domain-containing protein, ETS transcription factors related
Y53G8AR.9		NC	0	
F56B3.4		DOWN	0	
C54H2.5	*sft-4*	UP	0	
B0207.6		NC	0	
F59E10.1	*orc-2*	NC	0	origin recognition complex, second largest subunit ORC2
Y55F3AM.4	*atg-3*	NC	0	ortholog of the autophagic budding yeast protein Atg3p
C15C6.1		UP	0	
Y39A1A.13		NC	0	
Y87G2A.1		NC	1	
R08D7.3	*eif-3.D*	DOWN	1	translation initiation factor 3 subunit d

Table 3 Continued. Genes connected to miRNA function by genetic evidence from Parry *et al*, 2007.
[a] t-statistic difference of >2.5 with p<0.01.

Locus	**Gene name**	**mRNA[a] *alg-1* / WT**	**ALG-1 CLIP-seq clusters (.5)**	**Brief Description**
F01F1.7	*ddx-23*	NC	0	RNA helicase
F37E3.1	*ncbp-1*	NC	0	nuclear CAP-binding complex subunit protein
ZK742.1	*xpo-1*	NC	1	importin beta, nuclear transport factor
B0336.2	*arf-1.2*	UP	2	ADP-ribosylation factor
F57H12.1	*arf-3*	UP	1	ADP-ribosylation factor related protein
C17H12.1	*dyci-1*	DOWN	2	dynein intermediate chain
ZK593.5	*dnc-1*	NC	0	dynactin
ZK154.3	*mec-7*	NC	2	beta-tubulin
C47B2.3	*tba-2*	NC	4	alpha-tubulin
Y19D2B.1		NC	1	
C54D1.6	*bar-1*	NC	0	beta-catenin, transcriptional coactivator
F38H4.9	*let-92*	NC	5	serine/threonine protein phosphatase
W10C8.2	*pop-1*	NC	2	HMG box-containing protein, transcription factor
ZK1236.3		NF	0	SOp-2 Related (ectopic expression of Hox genes)
W09C2.1	*elt-1*	NC	0	Erythroid-Like Transcription factor family member
Y110A7A.14	*pas-3*	DOWN	0	endopeptidase
Y51H7C.6	*cogc-4*	UP	0	subunit of lobe A of the conserved oligomeric Golgi complex (COGC)
B0285.1		NC	0	serine/threonine kinase (CDC2/CDKX subfamily)
F25B4.6		UP	1	hydroxymethlyglutaryl-CoA synthase

Table 3 Continued. Genes connected to miRNA function by genetic evidence from Parry *et al*, 2007.
[a] t-statistic difference of >2.5 with p<0.01.

Locus	**Gene name**	**mRNA[a] *alg-1* / WT**	**ALG-1 CLIP-seq clusters (.5)**	**Brief Description**
eY50D7A.11		NF	0	
ZC581.1	*nekl-2*	UP	0	NEK (NEver in mitosis Kinase) Like family member, ser/thr-protein kinase
F48C1.4		NC	0	
Y110A7A.11		DOWN	0	

ACKNOWLEDGEMENTS

Amy E. Pasquinelli and Gene W.Yeo designed and directed the project; A.E.P., Dimitrios G. Zisoulis, G.W.Y. and Michael T. Lovci wrote the paper; D.G.Z. and Tiffany Y. Liang performed the experiments; M.T.L., Melissa L. Wilbert, Kasey R. Hutt, T.Y.L. and G.W.Y. performed the bioinformatics analyses. The authors thank Gary Ruvkun, Xiang-dong Fu, W. McGinnis and members of our laboratories for critical reading of the manuscript. We thank B. Hehli, S. Hunter and S. Bagga for technical assistance, Victor Ambros and M. Hammell for providing the list of mirWIP predictions and David Bartel for helpful advice. MLW is supported by the Genetics Training Program at the University of California, San Diego and a graduate fellowship from Genentech. This work was supported by grants from the US National Institutes of Health (GM071654-01 to A.E.P. and HG004659 and GM084317 to G.W.Y.), the Keck, Searle, V., Emerald and Peter Gruber Foundations (A.E.P.) and the Stem Cell Program at the University of California, San Diego (G.W.Y.).

CHAPTER 4 - RBFOX PROTEINS REGULATE ALTERNATIVE MRNA SPLICING VIA EVOLUTIONARILY CONSERVED RNA-BRIDGES

ABSTRACT

Alternative splicing (AS) enables programmed diversity of gene expression across tissues and development. We show here that binding in distal intronic regions (>500nt from any exon) by developmentally-important Rbfox splicing factors is extensive and an active mode of splicing regulation. Similarly to exon-proximal sites, distal sites contain evolutionarily conserved GCAUG sequences and are associated with AS activation and repression upon modulation of Rbfox levels in human and mouse experimental systems. As a proof of principle, we validated the activity of two specific Rbfox enhancers in *KIF21A* and *ENAH* distal introns and demonstrate that a conserved long-range RNA-RNA base-pairing interaction (an RNA-bridge) is necessary for Rbfox-mediated exon inclusion in the *ENAH* gene. Thus we demonstrate a novel RNA-mediated mechanism for AS control by distally bound RNA-binding proteins.

INTRODUCTION

The variety of alternative mRNA isoforms in higher eukaryotic transcriptomes indicates that a complex interplay among *cis*-elements and *trans*-factors exists to regulate splicing decisions. Splicing factors such as RNA binding proteins (RBPs) often bind as complexes within precursor messenger RNA (pre-mRNA) sequences to promote or repress splice site recognition (Black, 2003; Matlin et al., 2005; Wang et al., 2008). Variation within *trans*-factors or their binding sites leads to phenotypic diversity across mammalian evolution, and inherited or somatic genetic defects in these sites cause human

diseases. The recent application of genome-scale immunoprecipitation and high-throughput sequencing in mammalian cells provides insights into the networks of interactions among RBPs and their RNA substrates (Hafner et al., 2010; Hoell et al., 2011; Huelga et al., 2012; Konig et al., 2010; Lagier-Tourenne et al., 2012; Licatalosi et al., 2008; Polymenidou et al., 2011; Tollervey et al., 2011; Wilbert et al., 2012; Yeo et al., 2009; Zarnack et al., 2013). It has long been known that splicing factors bind within constitutive and alternative exons and their proximal intronic regions to alter splicing (Yeo et al., 2008; Ule et al., 2006; Yeo et al., 2007; Xue et al., 2009; Zhang et al., 2008). Sequence information within 500 nt of alternative exons and their neighboring flanking exons has been extensively studied to derive a computational splicing regulatory code (Barash et al., 2010). However, the genome-wide maps of RNA binding by proteins also reveal a large fraction of binding sites that are located much farther than 500 nt from potential target exons.

Published studies of distally located sequences that affected splicing allow only limited conclusions. The regulatory elements previously considered distal are often relatively close to the regulated exon or flanking exons and not inconsistent with existing models of splicing regulation. Few distal intronic enhancers have been demonstrated biochemically (Guo and Kawamoto, 2000) or proposed on the basis of conservation (Lapuk et al., 2010). For instance, the decoy 3' splice acceptor site sequence in the caspase-2 gene is located only ~200nt downstream from the regulated exon (Cote et al., 2001) and motifs that enhance splicing of *FN1* exon EIIIB are less than 500nt from the downstream exon (Baraniak et al., 2003; Lim and Sharp, 1998). Another complicating aspect of these studies is that the splicing factors recognizing the distal sequences are not

always known. For example, it is not known which RBPs bind a 526nt segment of the intronic sequence downstream of an exon in the *MYPT1* gene (Dirksen et al., 2003). Similarly, it is unclear how a *cis*-element in the first intron of the equine β-casein gene increases the inclusion of all weak exons in its pre-mRNA (Lenasi et al., 2006). Finally, the mechanisms by which a splicing factor might act on a distant exon are largely obscure. For example, a distal Rbfox motif was found to affect exon N30 in non-muscle myosin heavy chain B gene, but how this occurred was unexplored (Guo and Kawamoto, 2000; Kim et al., 2011).

To examine the genome-wide relevance of distal regulatory sites in splicing, we examined the *Rbfox* family of RNA binding proteins *in vivo* and in human cell lines. These proteins control tissue-specific AS of exons in brain, muscle, epithelial and mesenchymal cells, and embryonic stem cells (Gallagher et al., 2011; Gehman et al., 2012; Gehman et al., 2011; Venables et al., 2013; Yeo et al., 2009; Yeo et al., 2007b), and their binding sites are exceptionally highly conserved in sequence and position across vertebrate evolution (Minovitsky et al., 2005; Yeo et al., 2007b). RBFOX proteins interact with proteins mutated in spinal cerebellar ataxia types 1 and 2 (Lim et al., 2006; Shibata et al., 2000), and patients with mutations mapping to the *RBFOX1* gene locus exhibit a range of neurological deficits, such as mental retardation, epilepsy and autism spectrum disorder (ASD) (Bhalla et al., 2004; Davis et al., 2012; Martin et al., 2007; Matlin et al., 2005; Sebat et al., 2007). In muscle, post-transcriptional down-regulation of *RBFOX1* expression plays a role in the pathology of facioscapulohumeral muscular dystrophy (FSHD) (Pistoni et al., 2013). Moreover, animal models with knockout or knockdown of Rbfox protein expression reveal extensive defects in both neuronal and

muscle physiology (Gehman et a., 2011; Gehman et al., 2012; Gallagher et al., 2011), further suggesting that this class of RNA binding proteins plays critical roles in normal development.

Here we used genome-wide cross-linking, immunoprecipitation and sequencing (CLIP-seq) assays in mammalian brain to show that more than half of Rbfox binding sites are located distally (>500nt) from exons and that these distal sites are preserved through evolution. We used RNA-seq measurements of AS to show that distal Rbfox binding sites and distal conserved Rbfox motifs are preferentially associated with exons that are differentially spliced in experiments modeling Rbfox loss and gain. We experimentally demonstrated that these distal Rbfox binding sites directly control splicing in both endogenous genes and in minigene splicing reporters. Finally, we showed a novel mechanism by which long-range RNA-RNA secondary structures mediate distal splicing regulation by Rbfox. These results indicate distal intronic regions are rich reservoirs of highly conserved RNA *cis*-elements critical for splicing regulation.

RESULTS

Rbfox interacts in vivo with conserved and distal GCAUGs.

To generate genome-wide maps of Rbfox protein-RNA interactions *in vivo*, protein-RNA complexes from adult mouse brain were covalently cross-linked by ultraviolet irradiation and immunoprecipitated with antibodies specific for either Rbfox1 or Rbfox2 proteins (Figure 17a). Isolated RNA fragments representing Rbfox protein binding sites were sequenced and processed, resulting in 2,071,607 (Rbfox1) and 2,451,256 (Rbfox2) non-redundant alignments (Appendix 2) to the mouse (mm9)

genome. Our CLIP-seq cluster-finding algorithm, CLIPper (available here: http://github.com/YeoLab/clipper) was used to delineate clusters of reads representing regions in the transcriptome significantly associated with Rbfox binding. We identified 10,062 Rbfox1 clusters in 3,490 genes and 7,466 Rbfox2 clusters in 2,672 genes, with 1,901 genes containing both Rbfox1 and Rbfox2 clusters (Figure 17b).

The majority of Rbfox clusters (62%) were found within distal intronic regions, which we defined as intronic space >500nt from any annotated exon (Figure 18a). Rbfox clusters were also identified within 3´UTRs, consistent with previous observations[3], and only a minority of clusters (8%) was located in proximal introns. Second, the authenticity of distal clusters as *bona fide* Rbfox binding sites was supported by a *de novo* motif search with the HOMER algorithm, which demonstrated that both proximal and distal clusters exhibited statistically significant enrichment of the UGCAUG motif (Rbfox1: $P < 10^{-205}$ and $P < 10^{-48}$ respectively, Rbfox2: $P < 10^{-48}$ and $P < 10^{-90}$ respectively), compared to the appropriate backgrounds selected from similar genic regions (Figure 18b and Figure 17c for other highly ranked motifs). An alternative method of statistical analysis (Z-score)(Yeo et al., 2009) confirmed a significant enrichment for UGCAUG ($P < 10^{-10}$) and hexamers that contained GCAUG, for both proximal and distal clusters (Figure 17d). In addition, a GU-rich element previously observed in CLIP studies of RBFOX2 in human embryonic stem cells (Yeo et al., 2009) was present, likely representing other proteins that interact synergistically with Rbfox proteins. Third, we evaluated the evolutionary conservation of GCAUG sequences within Rbfox CLIP-defined binding sites. While only a small fraction (<15%) of bound sites contained a GCAUG sequence that was evolutionarily conserved between mouse and human (Figure

18), GCAUG sequences conserved across multiple genomes (mouse and human, rat or dog) were ~3.5 times more likely to be occupied in vivo by Rbfox than GCAUG sequences present only in the mouse genome (Figure 18d). Nevertheless, a statistically significant ($P < 0.05$) number of distal (and proximal) binding sites contained conserved GCAUG motifs, as compared to clusters of similar sizes distributed randomly in distal introns (Figure 18e). Fourth, as Rbfox1 and Rbfox2 proteins interact with the same sequence motif, we measured the correspondence in their binding sites as a measure of functional relevance. Notably, the level of overlap between distal Rbfox1 and Rbfox2 binding sites was similar to the proximal sites (Figure 17e). Furthermore, the overlap between both proximal and distal Rbfox1 and Rbfox2 binding sites increased as a function of degree of GCAUG site conservation within clusters (Figure 17f).

Lastly, the ontologies of genes bound in distal intronic regions were similar to, but also extended the categories associated with genes that contain exon-proximal binding sites (Figure 18f, Figure 17g for the entire list of statistically significant gene ontology categories and Appendix 3 for the list of genes within each category). Many of these genes bound by Rbfox in distal intronic regions, such as *Shank1* previously implicated in autism (Figure 19a)(Sato et al., 2012), are clearly important for neuronal function. Rbfox was observed to bind both proximal and distal regions downstream of the seizure-associated exon in *Snap25* (Figure 19b)(Johansson et al., 2008) and the stress axis-related exon in the *Kcnma1* gene, whose inclusion results in K^+ channels that are more sensitive to Ca^{2+} (Figure 19c)(Xie and McCobb, 1998). We also identified distal Rbfox binding sites a kilobase away from the auto-regulated exon encoding an RNA-recognition motif in each of the *Rbfox1* and *Rbfox2* genes, in addition to the previously

known proximal sites (Figure 19d and e)(Damianov and Black, 2010). Our genome-wide protein-RNA interaction maps of Rbfox proteins in mouse brains indicated that distal Rbfox sites have sequence conservation properties, gene targets and other features that support their functional roles in RNA regulation and disease etiology.

UGCAUG is enriched and conserved distal to alternative exons.

The hypothesis that distal Rbfox splicing enhancers or repressors regulate developmentally-important mammalian exons predicts that distal GCAUG motifs, as has been noted for proximal GCAUG motifs (Yeo et al., 2009; Zhang et al., 2008), are evolutionarily conserved and preferentially enriched in introns flanking AS exons. We tested whether the highly conserved distal regions have two properties expected for AS control regions: enrichment for known splicing regulatory motifs, and preferential association of these motifs with alternative exons more than constitutive exons. To achieve this, we modified a computational strategy (see Methods; Yeo et al., 2007) to score hexamers for their statistical enrichment in highly conserved intronic regions relative to lowly conserved intronic regions. We separately scored hexamer enrichment in conserved intronic regions flanking alternative cassette (single-exclusion) exons relative to regions flanking constitutively spliced exons then plotted these two scores against each other. This strategy infers function from evolutionary conservation and relevance for AS regulation (as opposed to another function) from proximity to alternative exons. First, we computationally identified 655,467 highly conserved regions in the human transcriptome, of average length 51 bases, excluding repetitive DNA or RNA elements, microRNAs, snRNA, rRNAs, and transcription factor binding sites identified by the ENCODE

consortium (Figure 20a) (Maher et al., 2012). Although <1% of intronic space meets the criteria for high conservation, almost half (42%) of all highly conserved regions are located within introns, with a great number of these falling in distal regions (35% of the total, or 225K regions; Figure 20b, Figure 21a).

We found that the *cis*-element composition of proximal conserved regions around AS exons (Figure 20c) was not identical to that of distal regions (Figure 20d and Figure 21b). For example, a CU-rich motif, which is a substrate for the PTB family of splicing factors, is enriched in conserved regions proximal to AS exons but depleted in distal conserved regions (Figure 20c and d; yellow triangles), thus suggesting these are under positive evolutionary selection in proximal regions, but negative selection in distal regions. Another motif (CAAUAA) was found to be highly conserved in both proximal and distal regions, but preferentially depleted around AS exons, compared to constitutively spliced exons (Figure 20c and d; light blue triangles). The Rbfox binding motif, UGCAUG (Figure 20c and d, dark blue circles), was consistently the most enriched conserved *cis*-element associated with AS exons in both proximal and distal regions ($P < 0.01$ by both criteria).

Distal Rbfox sites are associated with Rbfox-regulated exons.

Having determined that *in vivo* distal intronic Rbfox binding sites contain conserved GCAUG motifs, and that conserved distal intronic regions flanking annotated AS are enriched in Rbfox binding motifs in general, we next investigated their role in Rbfox-dependent AS regulation. Rbfox-regulated exons were identified by strand-specific RNA-seq from homogenized whole mouse brain isolated from nestin-conditional *Rbfox1*$^{-/-}$ and *Rbfox2*$^{-/-}$ knockout (KO) animals and paired wild-type control (Kim et al.,

2011; Gehman et al., 2012), and human 293T cells ectopically expressing either RBFOX1, RBFOX2, RBFOX3 or empty-vector control plasmids. Both loss of Rbfox1 and Rbfox2, as well as ectopic expression of RBFOX in human 293T cells, had virtually no effect on overall gene expression levels (Figure 22 a-e and Appendix 4).

To estimate the level of Rbfox-dependent exon usage, we calculated percent-spliced-in (psi, Ψ) values for annotated AS exons. We identified 620 and 934 (379 in common) mouse exons that were alternatively spliced in brain (change in the absolute value of Ψ or $|\Delta\Psi| \geq 5\%$) upon loss of Rbfox1 and Rbfox2, respectively (Figure 22F and G, Appendix 5 for the list of regulated exons). The degree of differential splicing upon Rbfox loss correlated well with RT-PCR measurements in the publications describing these knockout mice (Figure 22h and i)(Kim et al., 2011; Gehman et al., 2012). In mouse brains, of the exons that were differentially spliced ($|\Delta\Psi| \geq 5\%$) in both experiments, only about half (210 of 379) changed in the same direction; in contrast, ectopic expression of each RBFOX in 293T cells resulted in AS of hundreds of cassette exons (Figure 22g) but the regulated changes in exon inclusion in these cell-lines were more positively correlated (Figure 22j). RNA splicing components such as *Mbnl1, Mbnl2, Prpf18, Cwc22, Rsrc1, Raly, Thrap3, Rnps1, Clk4* were themselves alternatively spliced upon *Rbfox1* and *Rbfox2* knockout, suggesting that some of the AS changes measured are indirect. Notably, categories of genes that undergo regulation by AS are more closely related to categories of genes bound in proximal and distal intronic regions by Rbfox1 and Rbfox2, as compared to ones bound in 3´UTRs by Rbfox1 and Rbfox2 (Figure 18f and Figure 17e), suggesting a separable biological function of Rbfox1 and Rbfox2 in 3´UTR-mediated gene regulation.

Cassette exons were divided into differentially included ($\Delta\Psi > 5\%$), excluded ($\Delta\Psi < -5\%$) or unaffected ($|\Delta\Psi| < 2\%$) categories by Rbfox loss (in knockout mice compared to wild-type sibling pairs) or by RBFOX gain (in 293T cells ectopically expressing RBFOX compared to an empty-vector control) (Figure 22f-j). We determined the proportion of Rbfox-regulated and unaffected AS exons that had CLIP evidence for Rbfox binding or a GCAUG motif, at different levels of evolutionary conservation, in the proximal or distal ("PI" and "DI" columns of (Figure 23a-d), upstream (Figure 23a, c) or downstream (Figure 23b, d) intronic regions. Upon Rbfox2 loss in mouse brain, a statistically significantly higher fraction of differentially included AS exons (blue bars) than unaffected exons (grey bars) contain a conserved GCAUG motif in the upstream proximal region ($P < 1x10^{-3}$; Figure 23a); while a higher fraction of excluded AS exons (golden bars) contain conserved GCAUG motifs in the downstream proximal region ($P < 4x10^{-6}$; Figure 23b). Interestingly, exons included upon Rbfox2 loss are depleted of CLIP-defined binding sites in downstream proximal intronic regions ($P < 4x10^{-3}$; Figure 23b). As expected, inverse effects were observed when RBFOX2 was ectopically expressed in 293T cells ($P < 3x10^{-4}$, $P < 3x10^{-3}$ for downstream of included and upstream of excluded exons, respectively; Figure 23c and d). RBFOX1 and RBFOX3 experiments had similar, but less dramatic effects on splicing (Figure 24a and Figure 26b). Therefore, Rbfox interaction within proximal intronic regions was associated with Rbfox regulation, confirming previous "position-dependent" rules: that upstream Rbfox binding is associated with repression of exon recognition, while downstream binding correlates with exon inclusion (Yeo et al., 2009; Zhange et al., 2008).

We next examined the association of distal Rbfox interaction with splicing changes. In Rbfox2 loss, a statistically higher fraction of excluded exons than unaffected exons contained upstream, distal conserved motifs ($P < 1x10^{-2}$; Figure 23a) and CLIP-defined Rbfox2 binding sites ($P < 2x10^{-2}$; Figure 23a). Interestingly, unlike what we found in proximal regions, a higher proportion of differentially included exons than unaffected exons contained downstream, distal conserved motifs ($P < 5x10^{-4}$; Figure 23b). Also, a higher fraction of excluded than unaffected exons contained downstream, distal CLIP-defined Rbfox binding sites ($P < 9x10^{-3}$ for Rbfox1 CLIP; $P < 2x10^{-2}$ for Rbfox2 CLIP; Figure 23b). In RBFOX2 ectopic expression in human cells, we found that a higher fraction ($P < 4x10^{-2}$; Figure 23d) of excluded than unaffected exons had downstream distal conserved motifs.

As further support for the hypothesis that distal Rbfox sites can elicit regulatory effects on AS, the cumulative distributions of $\Delta\Psi$ values for mouse exons with either up- or down-stream, distal conserved GCAUG motifs are statistically significantly different ($P < 0.05$ and $P < 0.006$ for up- and down-stream motifs, respectively; two-sample Kolmogorov-Smirnov (KS) test) compared to the distribution of $\Delta\Psi$ values for exons without intronic conserved GCAUG motifs; Figure 23e). In ectopic expression experiments, only the distribution of $\Delta\Psi$ values for exons with downstream conserved motifs was significantly different from background (Figure 23f). While the directionality of AS changes mediated by distal sites is likely more complex than proximal sites, these two complementary approaches demonstrate that conserved GCAUG motifs and *in* vivo distal Rbfox binding sites are active splicing regulatory elements associated with Rbfox-dependent AS changes.

Distal Rbfox sites regulate AS in vitro and in vivo.

To demonstrate proof-of-principle that distal Rbfox motifs regulate AS, we investigated exons from two human genes that exhibit RBFOX2-dependent AS (Lapuk et al., 2010): *KIF21A* exon 23 (E23) and *ENAH* (also called MENA) exon 11a (E11a). Biochemical evidence demonstrates interaction at distal intronic sites flanking these exons with Rbfox1 in mouse brains and Rbfox2 in both mouse brains and human 293T cells (Figure 25a, d). Furthermore, binding sites are remarkably conserved across genomes (Figure 25b, e). The distal UGCAUG motifs in both exons are at least 500 base pairs away from any exon, allowing us to assess their functionality at long distances.

KIF21A is a member of the kinesin super-family that is over-expressed in Down syndrome (Salemi et al., 2013) and mutated in congenital fibrosis of the extraocular muscles type 1 (Heidary et al., 2008). Inclusion of a 21nt exon (E23) in *KIF21A* is RBFOX2-dependent and analysis of its downstream flanking intron revealed both proximal and distal conserved RBFOX motifs. The two distal sites are located 42nt apart in a short region of homology ~3.3kb downstream of the alternative exon and ~550nt upstream of the next exon. To assess whether these distal conserved sites function in the context of endogenous transcripts to alter splicing we tested the ability of antisense morpholino oligonucleotides (MOs) designed against these sites to alter E23 splicing in HS578T cells (Figure 25c, upper panel). Inclusion of E23 was reduced from ~39% inclusion in mock-treated cells to ~18% inclusion in cells treated with an MO against the first distal site, indicating that this UGCAUG motif regulates E23 splicing from a distance. An MO complementary to the second distal site had no effect, either because this site doesn't regulate splicing in this cell line, or because the site's physical

conformation inhibits MO efficacy. As a control, an MO directed against a heterologous event in the cytoskeletal gene, *EPB41* (Parra et al., 2011), altered splicing of its intended target transcript but did not affect E23 splicing (Figure 25c, lower panel). We conclude that one of the distal RBFOX motifs downstream of E23 is a strong distal splicing enhancer, and is required for optimal splicing even in the presence of conserved proximal sites.

ENAH E11a exhibits dramatically reduced inclusion during epithelial-mesenchymal transition (Warzecha et al., 2010), and is spliced in a breast cancer subtype-specific manner (Lapuk et al.,2010). E11a splicing is regulated by RBFOX2 (Yeo et al., 2009; (Kim et al., 2011; Gehman et al., 2012), and our genome-wide CLIP assays identified binding sites for Rbfox1 and Rbfox2 1.8kb downstream of E11a in mouse brain, in human 293T cells (Figure 25d) and in human embryonic stem cells (Yeo et al., 2009).

In contrast to *KIF21A* E23, the only conserved GCAUG sequence motifs in the intron downstream of *ENAH* E11a are located distally, ~1.8kb from the regulated exon. A group of three motifs is well conserved in at least 34 mammalian genomes, and can also be found at orthologous positions in avian genomes (Figure 25d). The absence of conserved GCAUG sequences in the proximal flanking intron suggested that the conserved distal sites mediate RBFOX-dependent splicing enhancer activity. We investigated the function of these distal RBFOX sites in endogenous transcripts in an *in vivo* context using vivo-MOs (vMOs, MOs with chemical modifications that improve their efficiency *in vivo*)(Morcos et al., 2008). E11a was partially included in *ENAH* mRNA isolated from kidney and liver of mice treated with a saline control (Figure 25f,

upper panel, mock). Injection of a single vMO complementary to two of the conserved RBFOX motifs (α-*ENAH* vMO) greatly reduced E11a inclusion in both tissues (Figure 25f, upper panel). Control vMOs that target a heterologous event (α-*EPB41* vMO1 and vMO2)(Parra et al., 2011) only showed splicing changes in their intended targets. We conclude that RBFOX proteins regulate AS of E11a under physiological conditions from distal conserved binding sites.

To further validate the role of distal RBFOX sites and exclude off-target effects of the vMO, we transfected human breast cancer cell-line HCC1954 with various three-exon minigene splicing reporters representing the E11-E11a-E12 region of the human *ENAH* gene (Figure 25g). In strong support of our hypothesis that these distal sites are functional, inclusion of E11a was dramatically reduced to almost complete exon skipping by mutating the three conserved distal GCAUG sequences (Figure 25h, lanes 1 and 2). In contrast, mutation of two non-conserved GCAUG sequences (open ovals) had little effect on splicing (Figure 25h, lane 3). Co-precipitation of biotinylated RNA containing two of the distal RBFOX sites with *in vitro* translated RBFOX2 confirmed protein-RNA binding, which was lost when we mutated RBFOX sites (Figure 25i). In summary, the above experiments functionally demonstrate that distal evolutionarily conserved GCAUG elements control splicing of exon E11a in the *ENAH* gene and E23 of the *KIF21A* gene and, more globally, there is strong statistical association between the presence of distal RBFOX sites and RBFOX-regulated exons. Next we approached the mechanism of this molecular phenomenon.

A long-range RNA-bridge mediates AS regulation by RBFOX.

Distal Rbfox binding sites must require a mechanism for recruitment to an alternative exon, in order to productively enhance spliceosomal activity. We reasoned that RNA secondary structure might provide a "bridge" that links distal *cis*-elements with their exon targets. To investigate a role for such RNA-bridges we first scanned all cassette and constitutive exons for potential RNA-RNA interactions between exon-proximal and exon-distal intronic segments using RNAhybrid (Figure 26a). By requiring candidate RNA-bridges to be evolutionarily conserved, as expected for developmentally important structures, we pared the initial list of more than two million RNA-bridges to approximately twenty-four thousand conserved RNA-bridges. These were significantly enriched near alternatively spliced exons, compared to constitutively-spliced exons ($P < 0.05$ by χ^2 test; Figure 26b). Moreover, the relative enrichment of RNA-bridges near alternative exons increased as a function of duplex stability, a variable not dependent on conservation levels, suggesting that conserved RNA-bridges are more common and more thermodynamically stable around alternatively spliced exons than around constitutive exons (Figure 26b, red line).

RNA-bridges that may play a role in distal Rbfox regulation were further enriched by identifying structures that had a distal arm within 50nt of a conserved GCAUG site (BL score ≥ 0.3; see Methods). By this approach, we found 699 predicted RNA-bridges around 125 exons (i.e. there are multiple potential bridges per exon). Focusing only on RNA-bridges around exons that were altered upon ectopic expression of RBFOX ($|\Delta\Psi| >$ 5% in any experiment) resulted in 162 predicted RNA-bridges around 19 exons. Among these we found an RNA-bridge connecting *ENAH* E11a to the distal site characterized

above (Figure 26c), but notably we did not predict an RNA-bridge around *KIF21A*'s AS event that met the strict filtering criteria we applied. Some AS events, including those in the *HnRNPR* gene (not shown) and *ENAH* had predicted RNA bridges that met the above criteria and also had strong biochemical evidence for RBFOX2 binding from our iCLIP in 293T cells (Figure 25a).

Downstream of *ENAH* E11a, our computational scan for RNA structures retrieved two overlapping RNA-bridges predicted to connect a conserved region 30-120nt immediately proximal to E11a to a similarly conserved region 10-100nt upstream of the distal RBFOX cluster (Figure 26c). The proximal and distal arms of the structure were separated by a putative loop of 1.1-1.7kb (mammals) or 0.6kb (chicken), and each arm consisted of two subdomains. This structure was conserved in most mammalian genomes, and also in the chicken genome, with evidence for compensatory mutations that maintain structure but not primary sequence (Figure 26c). We hypothesized that this stem loop structure could bridge or recruit the distal RBFOX sites close to E11a, in effect recapitulating the classical example of Rbfox splicing regulation proximal to alternative exons.

To probe the role of this structure in regulating E11a splicing, we generated minigene constructs that (1) contained disrupted RNA-RNA interactions in each of the structure's subdomains or (2) compensatory mutations that rescue RNA-RNA interactions but not primary sequence or (3) a structure that had a perfectly complementary RNA-bridge (Figure 26d). We found a dramatic reduction in E11a inclusion upon mutation of any subdomain (referred to as STEM-a and STEM-b) of the predicted stem structure (Figure 26e; lanes 2, 3, 5 and 6), even though the RBFOX sites

remained intact. Combining the two STEM-a mutations created construct STEM-a comp, which restored base-pairing (Figure 26e; lane 4) and rescued E11a inclusion to half of the normal level. Better rescue of E11a inclusion was observed in the STEM-b compensatory mutation (Figure 26e; lane 8). Consistent with these results, we also found that a double mutant construct (STEM-ab prox) completely abrogated inclusion (data not shown), while double-compensatory mutation (STEM-ab comp) recovered more than half of E11a inclusion (Figure 26e; lane 8) Finally, a mutation that extended the original base pairing by creating a perfect 42nt stem actually increased E11a splicing efficiency above normal levels (Figure 26e; lane 9). Based on these results, we theorize that AS regulation can be mediated by long-range RNA-RNA interactions between paired sequences that form an "RNA-bridge" to recruit a distal RBFOX site close to its target exon (Figure 26g).

DISCUSSION

Aberrant regulation of post-transcriptional RNA processing networks is increasingly recognized as a major cause of human genetic disease. Establishing the consequence of RBP interactions will be key to understanding the molecular basis of human diseases and the basic mechanisms that drive cellular processes. Our protein-RNA interaction maps reveal that Rbfox proteins not only bind proximal, but also distally to exons and furthermore that these sites are active. Minimally, this suggests there is a vast and untapped trove of information that could be used to predict the inclusion of exons according to cell state or environment. Aside from reaffirming the positional rules where proximal binding of Rbfox upstream of an exon suppresses, and downstream enhances

exon usage, we have identified hundreds of distal sites that are associated with Rbfox-regulated splicing. To our knowledge, we provide the first evidence that long-range RNA-RNA interactions can function over kilobase distances in vivo to mediate activity of distal enhancers, and that such RNA-bridges may be common components of distal regulatory mechanisms in AS control.

RNA secondary structures have long been known to alter splicing patterns in yeast (Goguel and Rosbash, 1993; Plass et al., 2012; Rogic et al., 2008), Drosophila (Kreahling and Graveley, 2005; Raker et al., 2009) and mammalian pre-mRNAs (Licatalosi et al., 2008; Matlin et al., 2005; Pervouchine et al., 2012). These have most often been observed to loop out exons or splice sites to induce their skipping (Kreahling and Graveley, 2005; Nasim et al., 2002), also reviewed in (McManus and Graveley, 2011). Early work in yeast also showed that intra-intronic base pairing interactions could enhance the splicing of long introns (Goguel and Rosbash, 1993; Rogic et al., 2008). In mammals, secondary structures have been shown to alter the activity of regulatory proteins by blocking or removing their binding sites (Warf et al., 2009), but we here provide the first evidence for their role as chaperones for RBP-mediated long-range splicing regulation. Our data indicate that RNA-bridges can function by dramatically shortening the distance between a distally bound splicing regulator and its target exon; indeed, close examination of iCLIP-seq data in Figure 26d revealed evidence for Rbfox2 interactions at the proximal stem region of the bridge in *ENAH* intron 11a that lacks GCAUG motifs and would not be expected to bind Rbfox2 directly. In summary, RNA-RNA interactions within introns can have major effects on splicing and provide a versatile mechanism for juxtaposing splicing controllers with synergistic or antagonistic

relationships. Beyond that, it adds and an additional layer of AS regulation via promotion or inhibition of RNA-bridge formation, for example, through RNA-editing.

Our results suggest that long-distance regulation of splicing is more far-reaching than the small number of earlier reports might imply. Abundant regulatory information located deeper within introns also represents an under-appreciated source of disease-causing mutations. As the number of sequenced human genomes increases, our catalogs of RNA binding sites and conserved regions will enable nucleotide-level, functional association of natural and disease variation with AS. We conclude that future studies of AS networks in normal development and in disease must consider *both* proximal and distal RNA binding sites and non-canonical molecular mechanisms that lead to distal enhancer activity in order to accurately predict RNA splicing with high accuracy. As we have demonstrated with vMOs, these *cis*-regulatory elements provide an important opportunity for targeted rationally-designed therapeutic interventions, for example such as those which are proving effective and specific in the treatment of splicing-related diseases.

METHODS

Accession Codes

Raw *.fastq* files are available from SRA under accessions SRP029987 and SRP030031.

CLIP-seq library generation and analysis

CLIP-seq libraries were constructed as previously described[7]. Fresh brains from 8-week-old female C57Bl/6 mice were rapidly dissociated by forcing through a cell strainer with a pore size of 100 μm (BD Falcon) before ultraviolet cross-linking. Using antibodies against the proteins *Rbfox1* (custom generated in the Black lab, UCLA) and *Rbfox2* (Bethyl Laboratories), we generated CLIP-seq libraries that were sequenced on the Illumina GAIIx platform. Raw reads were subjected to custom trimming scripts that removed adapter sequences and truncated reads at low-quality bases or 10-mer homopolymers. Reads were filtered through a catalog of consensus genomic elements by mapping with bowtie (Langmead et al., 2009)(parameters: -q -p 1 -e 100 -l 20). Reads that did not map to repetitive elements were mapped to the human (hg19) or mouse (mm9) reference genomes using GSNAP (Wu and Nacu, 2010). GSNAP was supplied with the location of exon-junctions from Ensembl and UCSC genes and mapped with the following parameters: -t 2 -N 1 -n 10 -Q -B 5. CLIP-seq reads from replicate libraries were combined after quality checks and the analysis of clusters from each library revealed an enriched GCAUG motif. CLIP-seq reads were collapsed to remove PCR artifacts using samtools rmdup (Li et al., 2009). A new cluster-identification algorithm, CLIPper (CLIP-seq peak enrichment; http://github.com/YeoLab/clipper), was developed to identify clusters representing binding sites for *Rbfox1* and *Rbfox2*. For each pre-mRNA, we determined the minimum height of CLIP-seq reads required to satisfy a user-defined false discovery rate (FDR) of 0.05, based on our previous method (Yeo et al., 2009). Next, we interpolated the heights of reads across the length of the pre-mRNA using cubic splines. From the fitted curve, peaks, centers and widths that represented

clusters were identified. The number of reads expected within each cluster was estimated by the Poisson distribution using the total number of reads that mapped within the entire length of the pre-mRNA as described in (Zisoulis et al., 2010). In addition to this "pre-mRNA" cutoff, for each cluster, we also recalculated these cutoffs for reads within 1kb and included any putative clusters that were significant by either this local analysis or by the gene-wide analysis described above. Clusters were assigned to genic regions based on the following order of priority: Exon > 3′UTR > 5′UTR > Proximal Intron > Distal Intron. Background regions to compute statistically significant motifs were selected four times for each cluster by deriving a random "cluster" of equal length in a randomly selected location in the same type of genic region in any gene. The software packages pybedtools and bedtools (Dale et al., 2011; Quinlan and Hall, 2010) were used to enumerate overlaps between clusters and motifs. To determine the statistical significance of the extent of overlap, regions were randomly located ten times, keeping the ratio of locations in the different genic regions the same. A Z-score with corresponding *P* value was computed from the shuffled mean and standard deviation.

De novo motif analysis for Rbfox clusters

Parameters supplied for motif finding with HOMER's findMotifs program were: -p 4 -rna -S 10 -len 5,6,7,8,9. Cluster sequences and background selected from the same genic region as real CLIP clusters were supplied as a fasta file.

Multi-species alignments

Alignments were obtained directly from MultiZ tracks (hg19 46-way and mm9 30-way alignments) available from the UCSC Genome Browser and avian genome alignments, as displayed in Figures 25 and 26, were manually adjusted for local accuracy where supplied alignments were unsatisfactory.

Branch-length (BL) scoring to measure conservation levels of GCAUG motifs

Rbfox motifs were scored by walking through every position of the transcriptome and summing the edit distance from a TGCATG motif across all aligned orthologs. Edit distances were multiplied by the branch length (supplied by UCSC) to a given ortholog after we used a sigmoid function to severely penalize edit distances < 0.8. Then we expressed this score as a fraction (0 to 1) of the maximum score possible if all orthologs contained an aligned perfect match to TGCATG. Lastly, we excluded positions where the target genome (either mouse or human) did not contain a GCATG pentamer.

De novo motif analysis for conserved regions

Parameters supplied for motif finding with HOMER's findMotifsGenome program were: "-size given -S 100 -rna -float -bits -len 6 -nlen 0 -noweight -h -minlp 0". Background locations supplied were lowly-conserved regions flanking constitutive exons.

Identification and word frequency analysis of conserved regions

Contiguous regions within the human genome with phastCons score, S were divided into categories of low ($0 \leq S \leq 0.3$), moderate ($0.3 < S \leq 0.9$) and high ($0.9 < S \leq 1.0$) levels of evolutionary conservation. If a region is separated by only 1 nt with another region, both regions are combined. Conservatively, regions longer than 2kb or shorter than 10nt were eliminated from further consideration. Regions that overlapped RepeatMasker (Smit et al., 1996-2010) annotated repeats or ribosomal RNA and microRNA genes were removed. We also removed any conserved regions overlapping with transcription-factor binding sites as defined by experiments conducted by the ENCODE Consortium (obtained from the UCSC genome browser here: http://hgdownload-test.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeRegTfbsClustered/wgEncodeRegTfbsClustered.bed.gz). We found that ~20% of our highly-conserved regions overlapped with a transcription factor binding site. After defining conserved regions, they were assigned to functional genic categories (exon, 5′ untranslated region, 3′ untranslated region, proximal intron or distal intron). To compute the coverage of each genic region, we divided the number of nucleotides represented by each category of conservation with the total number of nucleotides in that region. Alternative and constitutive splicing annotations were defined using available transcripts from Ensembl compiled with previously published methods (Yeo et al., 2007a). In total, we analyzed 23,982 annotated protein-coding genes, 18,551 cassette (or skipped) exons and 164,920 constitutive exons. A χ^2 enrichment score for each motif was computed with the counts for each hexamer as follows: (1, x-axis of Figure 20) counts in highly conserved regions relative to lowly

conserved regions and, (2, y-axis of Figure 20) counts in highly conserved regions associated with exons that were alternatively spliced, compared to hexamers in highly conserved regions associated with constitutively spliced exons. Background for each test was the number of other hexamers in these regions. Enrichment scores are multiplied by the direction of enrichment, which was determined as the sign of the difference between ratios of hexamer counts over background. The software package HOMER (Heinz et al., 2010)was utilized to identify degenerate motifs that were significantly enriched in highly conserved regions around cassette exons versus lowly conserved regions around constitutive exons, as inset into Figure 20c. To identify 6-mers that were similar to the HOMER-derived motif, we computed a Pearson correlation coefficient between the 6-mer and motif. A correlation coefficient of greater than 0.8 constituted a similar motif.

RNA-seq library generation and analysis

Total RNA from whole brain of *Rbfox1* and *Rbfox2* nestin-specific knockout one-month old male and wild-type sib-pairs was extracted by Trizol as per manufacturer's instructions. Total RNA from human 293T cells after ectopic expression of N-terminal FLAG-tagged mouse *Rbfox1* (NP_067452)(Underwood et al., 2005), human RBFOX2 (AAL67150) (Underwood et al., 2005) and mouse *Rbfox3* (NM_001024931) in pcDNA3.1 (Life Technologies) 48 hours after transfection using Lipofectamine 2000 (Life Technologies) was subjected to Trizol extraction. RBFOX1, RBFOX2 and RBFOX3 levels in 293T cells were measured by qRT-PCR and were evaluated to be 6-, 4- and 4-fold higher than cells transfected with control vector pcDNA3.1 expressing only FLAG, respectively. RNA-seq libraries were prepared using 8 ug of total RNA, and

subjected to polyA-selection. After strand-specific dUTP library preparation, as described in (Parkhomchuk et al., 2009), cDNA corresponding to 150–225 nt fragments was single-end sequenced with Illumina GAIIx to 101nt. Raw reads were subjected to custom trimming scripts which removed adapter sequences and truncated reads at low-quality bases or 10-mer homopolymers. Reads were filtered through a catalog of consensus genomic elements by mapping with bowtie (Langmead et al., 2009)(parameters: -q -p 1 -e 100 -l 20). Reads that did not map to repetitive elements were mapped to the human (hg19) or mouse (mm9) reference genomes using GSNAP (Wu et al., 2010). We supplied GSNAP with the location of exon-junctions from Ensembl and UCSC genes and mapped with the following parameters: -t 2 -N 1 –n 10 –Q –B 5. Mapped reads were used to quantify gene-expression and splicing measurements as was done in Polymenidou et a., 2011 with improvements that allowed us to use spliced reads mapped to the genome instead of an exon junction database. Our RNA-seq analysis verified reductions in mRNA levels of *Rbfox1* (by 63%) and *Rbfox2* (by 84%) in the *Rbfox1* and *Rbfox2* knockout mice, respectively. Also, as was previously observed by western blot (Gehman et al., 2012; Gehman et al., 2011), we found reciprocal compensatory increases of 21% for *Rbfox2* mRNA upon *Rbfox1* loss and 10% for *Rbfox1* upon *Rbfox2* loss. Percent-spliced-in (Ψ) values were calculated as half the number of reads mapped to all inclusion isoforms over the number of reads mapped to all exclusion isoforms plus half the number of reads mapped to all inclusion. Only exons with evidence for alternative splicing from Ensembl annotations (human: GRCh37 v65, mouse: NCBI37) were evaluated for differential splicing.

Gene ontology.

Gene ontology analysis consisted of a hypergeometric test comparing the fraction of genes in each ontology category that appeared relevant (bound, differentially expressed or alternatively spliced) in a particular high-throughput experiment to the fraction of all expressed (RPKM > 0.5 in the species-appropriate RNA-seq experiment) genes in that category. Reported *P* values are Bonferroni-corrected for multiple-hypothesis testing and clustered along rows and columns using a Euclidean distance metric.

Distal association of Rbfox sites with regulated exons

We categorized cassette exons into classes undergoing three types of regulation. Exons were either included ($\Delta\Psi > 5\%$), excluded ($\Delta\Psi < -5\%$) or unaffected ($|\Delta\Psi| < 2\%$) according to the RNA-seq experiments described above. Furthermore, exons were required to have ≥30 or ≥50 reads mapped across exon-junctions showing evidence for inclusion and/or exclusion in human and mouse experiments, respectively, and a flanking intron ≥ 1.5 kB in at least one direction. We tested several features we expected to be associated with exons that were regulated upon *Rbfox* depletion or ectopic expression. In mouse experiments, the features we examined were: (Li et al., 2008) *Rbfox* GCAUG motifs (2) conserved *Rbfox* GCAUG motifs with a BL score ≥ 0.1 (3) conserved *Rbfox* GCAUG motifs with a BL score ≥ 0.4 (4) *Rbfox1* CLIP clusters (5) *Rbfox2* CLIP clusters. The features we tested in human were: (Li et al., 2008) *Rbfox* GCAUG motifs (2) conserved *Rbfox* GCAUG motifs with a BL score ≥ 0.1 (3) conserved *Rbfox* GCAUG motifs with a BL score ≥ 0.2. We calculated significance by a Fisher's exact test

comparing the proportion of changing cassette exons with at least one of a given feature in a given region to the proportion of non-changing cassette exons with that feature in that region.

Splicing reporter construction

The wild type minigene contains a 7.3kb fragment of the human ENAH gene encompassing the exon 11-11a-12 region, plus ~50nt of upstream and downstream intron sequence upstream. Primers used to amplify the E11-E12 region were as follows:

Forward primer:

5′-tggaattctgcagatGTCTGGCATTGTGCAAATTAGA-3′;

Reverse primer:

5′-gccactgtgctggatCATTCAGGATCCATGTCAAAGA-3′.

The lower case nucleotides provided 15 nt overlaps with EcoRV-linearized pcDNA3.1 vector. Insert and vector were assembled together using the In-Fusion Advantage kit according to the manufacturer's instructions (Clontech). In-Fusion technology was also used to introduce mutations at the deep intron RBFOX2 sites. First, the entire wild type splicing reporter, except a small region containing the wild type RBFOX sites, was PCR-amplified so as to generate a linearized construct opened at the intron enhancer region. Complementary 39-mer oligonucleotides containing the three mutated RBFOX sites, and 15nt overlaps with the linearized splicing reporter, were annealed together and inserted into the vector by In-Fusion methods. Primers used to amplify the wild type minigene:

Forward primer:

5′-TTAAAAATTTGACTGTTTCCACAATTG-TTTATTACA-3′;

Reverse primer:

5′-TCAGTCTAACAGTCAATCCATCACCACCACCACCAC-3′.

Oligonucleotides containing the mutated RBFOX sites (underlined):

Forward:

5′-TGACTGTTAGACTGAATTAAATTTTTAAAAATTTGACTG-3′;

Reverse:

5′-CAGTCAAATTTTT-AAAAATTTAATTCAGTCTAACAGTCA-3′.

The non-conserved RBFOX sites were modified using multi-site mutagenesis to mutate both non-conserved sites in one reaction, using the following primers in which the mutated RBFOX motifs are represented in upper case:

Primer 1: 5′-catggtTGACTGtgcctgtgggaggctg-3′;

Primer 2: 5′-ggaataggtAGACTGagtgaatatatgaaataacatcc-3′.

Splicing analysis of endogenous transcripts and minigene reporters

Splicing reporter minigenes were transfected into human HCC1954 or T47D cells using FuGene HD Transfection Reagent (Roche). Total RNA was extracted from cells with Qiagen's RNeasy Mini Kit, and then reverse-transcribed into cDNA using random primers and the Superscript III First Strand Synthesis System (Invitrogen). Subsequent PCR analysis was performed using AccuTaq polymerase (Sigma). Amplification of minigene E11a inclusion and exclusion products as a measure of splicing efficiency was done using primers in intron 10 (forward primer 5′-GCATTGTGCAAATTAGAGTCCTT-3′) and intron 12 (reverse primer 5′-

CAGGATCCATGTCAAAGATATGC-3′). Amplification of mouse endogenous ENAH transcripts was performed using the primers:

Forward primer: 5′-GCTGAGAAGGGATCAACAATAG-3′;

Reverse primer: 5′-GCTCTGCTTCAGCCTGTCATAG-3′

Splicing of human KIF21A was assayed using primers:

Forward primer: 5′-GAAATAACCAGTGCTACCCAAAAC-3′

Reverse primer: 5′-GTTTAAAGGAGCATCCTCATCAGT-3′

Morpholino treatments

25 nucleotide antisense morpholinos sequences were obtained from Gene Tools, LLC (Philomath, OR). The *vivo* morpholino for mouse experiments contains a covalently linked octa-guanidine dendrimer as a delivery moiety to facilitate entry into cells in vivo. The enhancer blocking sequence for ENAH was as follows:

5′TAATT**CATGCT**AC**CATGCA**ATCCAC-3′;

underlined sequences are complementary to two of the conserved RBFOX binding motifs. Mice received tail vein morpholino injections at 15mg/kg on two consecutive days, then RNA was purified from selected tissues on the third day. Tissues were rinsed in 1X PBS and snap frozen in ethanol/dry ice bath and stored at -80C. For KIF21A experiments, morpholinos (without a covalently linked octa-guanidine dendrimer) blocking the distal RBFOX sites were delivered to HS578T cells in 12 well plates using 5 µl of morpholino and 6 µl of endoporter (Gene Tools, LLC). RNA was extracted 48 hours after treatment. Morpholino sequences were as follows:

KIF21 distal1 5′-**CATGCA**ACAGCTC7gftres5fTGTAAACTAATA -3′

KIF21A distal2 5′- ACACATCAG**CATGCA**GCTCATTCAC-3′3

Image cropping

Gel images were cropped to show bands of the expected size in RT-PCR and biotin pull-down assays. Figure 27 shows uncropped images.

FIGURES

Figure 17. CLIP-seq to identify Rbfox binding sites, motifs and gene ontology analyses.

(a) Autoradiograph of Rbfox protein-RNA complexes from mouse brain immunoprecipitated with Rbfox-specific antibodies and trimmed with optimized unit (U) concentrations of micrococcal nuclease (MNase). (b) Venn diagram showing the number of genes bound by Rbfox1 and Rbfox2 in mouse brain. (c) List of the top three motifs significantly enriched in the Rbfox1 CLIP compared to appropriate background controls. Similar results were found for Rbfox2 (not shown). (d) Box-plots of hexamer Z-scores, comparing Rbfox1 CLIP clusters to randomly located clusters in each genic region. GCAUG and UGCAUG, the known Rbfox motifs, and a GU-rich 6-mer were enriched in all genic regions except 5′UTRs. (e) Venn diagrams showing the number of Rbfox1 (red) and Rbfox2 (blue) CLIP-seq clusters that overlap (yellow) each other by ≥ 1nt ("All clusters"). Clusters are then restricted to those within 50 nt of a GCAUG motif in mouse ("GCAUG in mouse"), or where the GCAUG motif is conserved ("Conserved GCAUG") at increasing branch-length (BL) scores (the higher the score, the more conserved the GCAUG sequence across multiple genomes). Analyses of proximal ("PI") and distal intronic ("DI") clusters are displayed on the left and right columns. (f) Bar plots represent the fraction of clusters that overlap relative to the Rbfox1 (red) or Rbfox2 (blue) clusters corresponding to the different restrictions in (e). (g) Dendrogram showing hierarchical clustering with Euclidean distance of statistically significant gene ontology terms (in at least one experimental condition) using negative log10 *P* values that measure the enrichment of genes within the ontology terms for various experimental conditions.

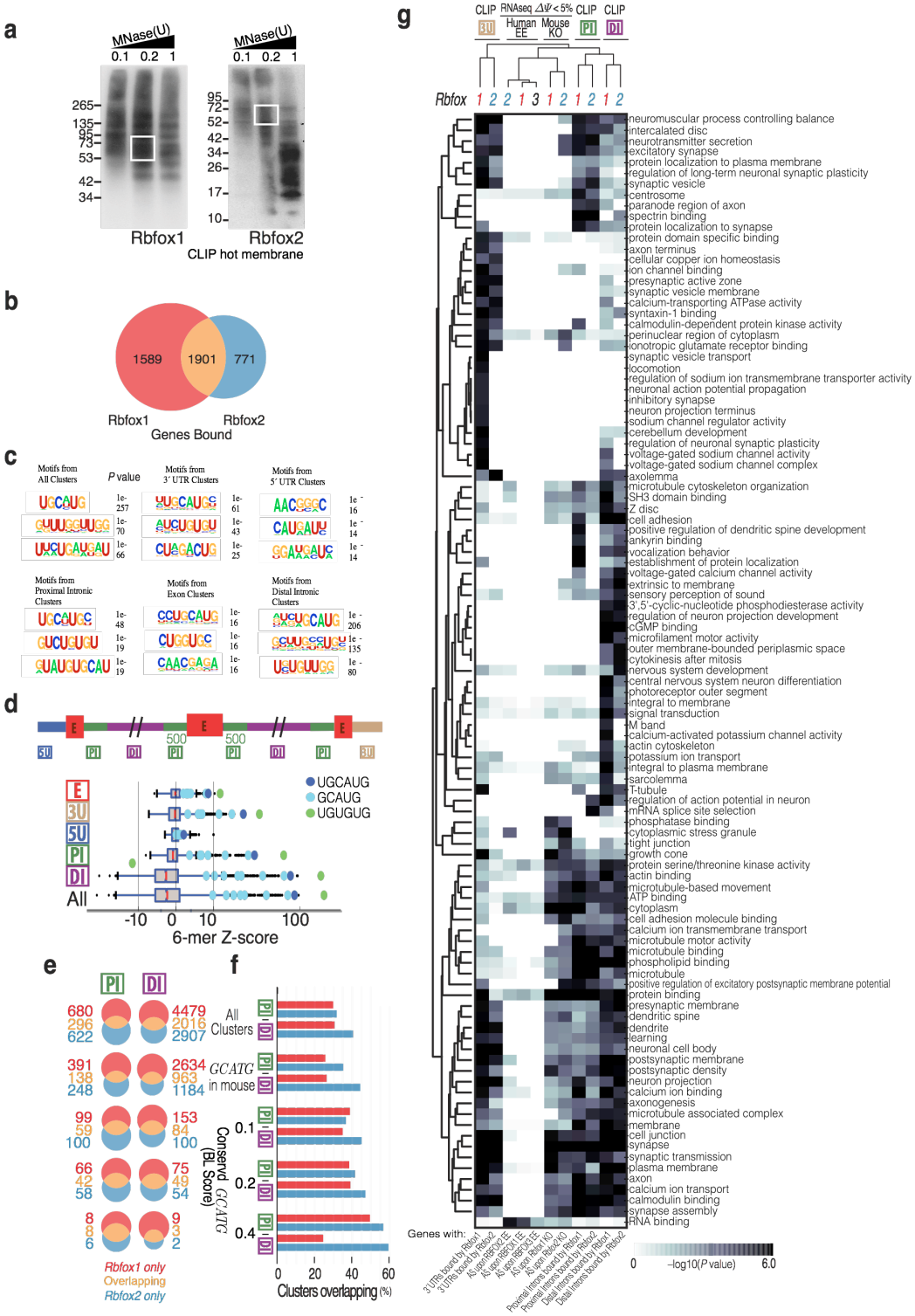

a
MNase(U)
0.1 0.2 1
265
135
95
73
53
42
34
Rbfox1
MNase(U)
0.1 0.2 1
95
72
52
42
34
26
17
10
Rbfox2
CLIP hot membrane
b
1589
1901
771
Rbfox1
Rbfox2
Genes Bound
c
Motifs from All Clusters
P value
Motifs from 3′ UTR Clusters
Motifs from 5′ UTR Clusters
Motifs from Proximal Intronic Clusters
Motifs from Exon Clusters
Motifs from Distal Intronic Clusters
d
E
3U
5U
PI
DI
500
UGCAUG
GCAUG
UGUGUG
All
-10 0 10 100
6-mer Z-score
e
PI
DI
680
296
622
4479
2016
2907
All Clusters
391
138
248
2634
963
1184
GCATG in mouse
99
59
100
153
84
100
66
42
58
75
49
54
8
8
6
9
3
2
Rbfox1 only
Overlapping
Rbfox2 only
f
Conserved GCATG (BL Score)
0.1
0.2
0.4
0 20 40 60
Clusters overlapping (%)
g
CLIP
RNAseq ΔΨ<5%
Human EE
Mouse KO
CLIP
CLIP
3U
PI
DI
Rbfox 1 2 2 1 3 1 2 1 2 1 2
neuromuscular process controlling balance
intercalated disc
neurotransmitter secretion
excitatory synapse
protein localization to plasma membrane
regulation of long-term neuronal synaptic plasticity
synaptic vesicle
centrosome
paranode region of axon
spectrin binding
protein localization to synapse
protein domain specific binding
axon terminus
cellular copper ion homeostasis
ion channel binding
presynaptic active zone
synaptic vesicle membrane
calcium-transporting ATPase activity
syntaxin-1 binding
calmodulin-dependent protein kinase activity
perinuclear region of cytoplasm
ionotropic glutamate receptor binding
synaptic vesicle transport
locomotion
regulation of sodium ion transmembrane transporter activity
neuronal action potential propagation
inhibitory synapse
neuron projection terminus
sodium channel regulator activity
cerebellum development
regulation of neuronal synaptic plasticity
voltage-gated sodium channel activity
voltage-gated sodium channel complex
axolemma
microtubule cytoskeleton organization
SH3 domain binding
Z disc
cell adhesion
positive regulation of dendritic spine development
ankyrin binding
vocalization behavior
establishment of protein localization
voltage-gated calcium channel activity
extrinsic to membrane
sensory perception of sound
3',5'-cyclic-nucleotide phosphodiesterase activity
regulation of neuron projection development
cGMP binding
microfilament motor activity
outer membrane-bounded periplasmic space
cytokinesis after mitosis
nervous system development
central nervous system neuron differentiation
photoreceptor outer segment
integral to membrane
signal transduction
M band
calcium-activated potassium channel activity
actin cytoskeleton
potassium ion transport
integral to plasma membrane
sarcolemma
T-tubule
regulation of action potential in neuron
mRNA splice site selection
phosphatase binding
cytoplasmic stress granule
tight junction
growth cone
protein serine/threonine kinase activity
actin binding
microtubule-based movement
ATP binding
cytoplasm
cell adhesion molecule binding
calcium ion transmembrane transport
microtubule motor activity
microtubule binding
phospholipid binding
microtubule
positive regulation of excitatory postsynaptic membrane potential
protein binding
presynaptic membrane
dendritic spine
dendrite
learning
neuronal cell body
postsynaptic membrane
postsynaptic density
neuron projection
calcium ion binding
axonogenesis
microtubule associated complex
membrane
cell junction
synapse
synaptic transmission
plasma membrane
axon
calcium ion transport
calmodulin binding
synapse assembly
RNA binding
Genes with:
3′UTRs bound by Rbfox1
3′UTRs bound by Rbfox2
AS upon RBFOX1 EE
AS upon RBFOX2 EE
AS upon Rbfox1 KO
AS upon Rbfox2 KO
Proximal Introns bound by Rbfox1
Proximal Introns bound by Rbfox2
Distal Introns bound by Rbfox1
Distal Introns bound by Rbfox2
0
−log10(P value)
6.0

Figure 18. Characteristics of Rbfox binding in distal intronic regions.

(a) Pie charts depict the fraction of Rbfox1 and Rbfox2 clusters, defined from CLIP-seq, within different genic regions, compared to the length distribution of each region across the transcriptome. A schematic of genic region definitions used in this paper is represented below. (b) *De novo* sequence motifs enriched above background that were similar to the canonical Rbfox motif are listed with their associated *P* value. (c) Pie charts depict the fraction of Rbfox1 clusters that contain (within 200nt) the sequence GCAUG, conserved in 1 (teal), 2 (burgundy) or 4 (orange) species (species abbreviations: Mm, *Mus musculus*; Hs, *Homo sapiens*; Rn, *Rattus norvegicus*; Cf, *Canis familiaris*). (d) Bar plots show the fraction of GCUAG motifs conserved in 1 (teal), 2 (burgundy) or 4 (orange) species occupied by Rbfox1 (overlapping within 200nt). (e) Bar charts show the fraction of Rbfox1 CLIP-seq clusters that contain (within 200nt) GCAUG motifs conserved in 1, 2, or 4 species (as in (c)) is shown separately for all, proximal and distal regions. (f) A heat map shows a portion of gene ontology categories represented by distal binding (see Figure 18e). The intensity of gray corresponds to the negative log10(*P* value) of a hypergeometric test for enrichment in gene ontology categories representing by genes bound in proximal intron (green boxed "PI"), distal intron (purple boxed "DI") or 3'UTR (brown boxed "3U"); or genes exhibiting AS in the RNA-seq experiments in the *Rbfox1* or *Rbfox2* knockout (KO) or *RBFOX* ectopic expression (EE).

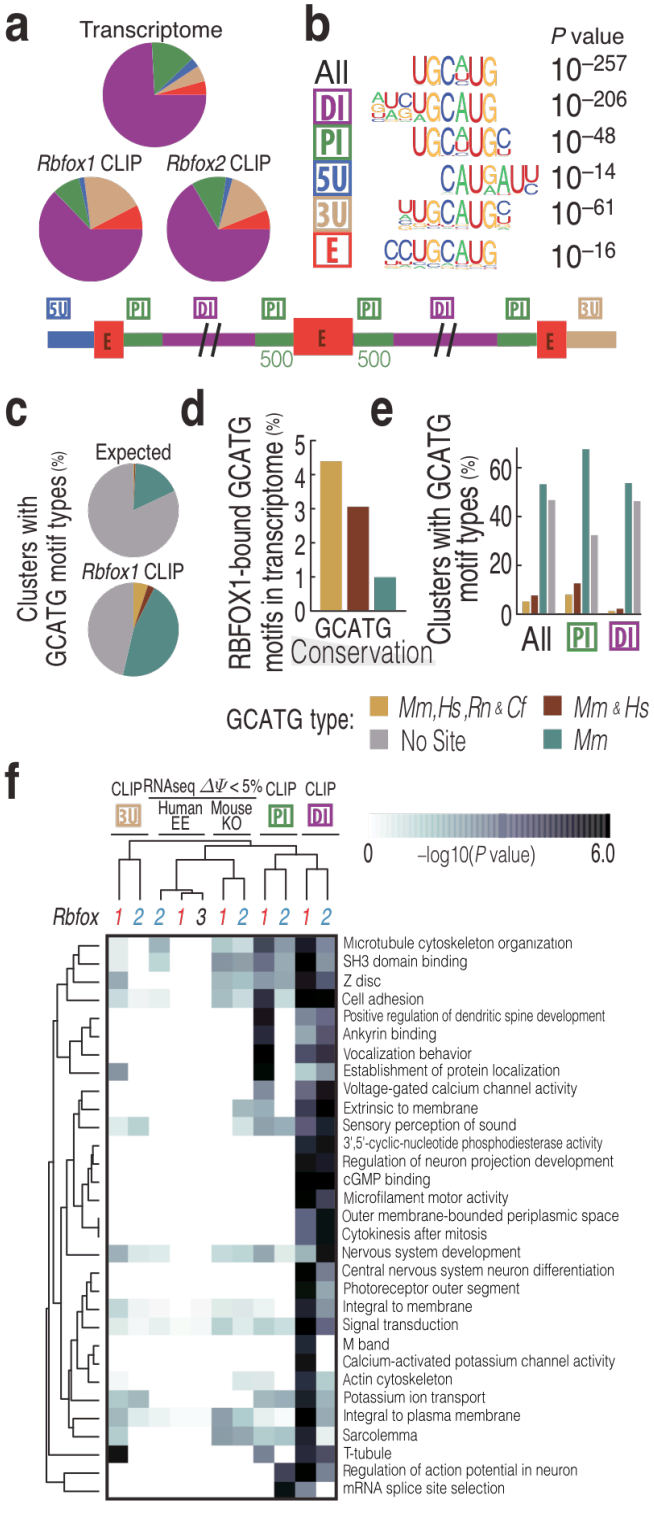

a
Transcriptome
Rbfox1 CLIP
Rbfox2 CLIP
b
P value
All
DI
PI
5U
3U
E
10−257
10−206
10−48
10−14
10−61
10−16
500
500
c
Clusters with GCATG motif types (%)
Expected
Rbfox1 CLIP
d
RBFOX1-bound GCATG motifs in transcriptome (%)
GCATG
Conservation
e
Clusters with GCATG motif types (%)
All
PI
DI
GCATG type:
Mm,Hs,Rn & Cf
Mm & Hs
No Site
Mm
f
CLIP
RNAseq ΔΨ < 5%
Human EE
Mouse KO
CLIP
CLIP
3U
PI
DI
0
−log10(P value)
6.0
Rbfox 1 2 2 1 3 1 2 1 2 1 2
Microtubule cytoskeleton organization
SH3 domain binding
Z disc
Cell adhesion
Positive regulation of dendritic spine development
Ankyrin binding
Vocalization behavior
Establishment of protein localization
Voltage-gated calcium channel activity
Extrinsic to membrane
Sensory perception of sound
3',5'-cyclic-nucleotide phosphodiesterase activity
Regulation of neuron projection development
cGMP binding
Microfilament motor activity
Outer membrane-bounded periplasmic space
Cytokinesis after mitosis
Nervous system development
Central nervous system neuron differentiation
Photoreceptor outer segment
Integral to membrane
Signal transduction
M band
Calcium-activated potassium channel activity
Actin cytoskeleton
Potassium ion transport
Integral to plasma membrane
Sarcolemma
T-tubule
Regulation of action potential in neuron
mRNA splice site selection

Figure 19. Genome browser views of selected alternatively spliced genes containing distal RBFOX binding sites.

(a-e) Exons and introns are indicated in blue (UCSC genes) or red (Ensembl Genes) with thick lines representing exons and thin lines overlaid with arrows representing introns. The direction of the arrows denotes the direction of transcription. Read density from Rbfox1 and Rbfox2 CLIPseq (orange and green tracks respectively) on several genes is shown with positive and negative values corresponding to the direction of transcription of aligned reads. *Shank1* (a), *Snap25* (b), *Kcnma1* (c), *Rbfox1* (d) and *Rbfox2* (e) possess alternatively spliced exons (red arrows). PhastCons evolutionary conservation scores for placental mammals are displayed at the bottom of each panel in dark green. Genomic GCATG sites are indicated above the conservation track in black. Distal and proximal highly conserved regions are designated along the top with purple and green filled rectangles, respectively. Scale bars at the top define the size of region displayed.

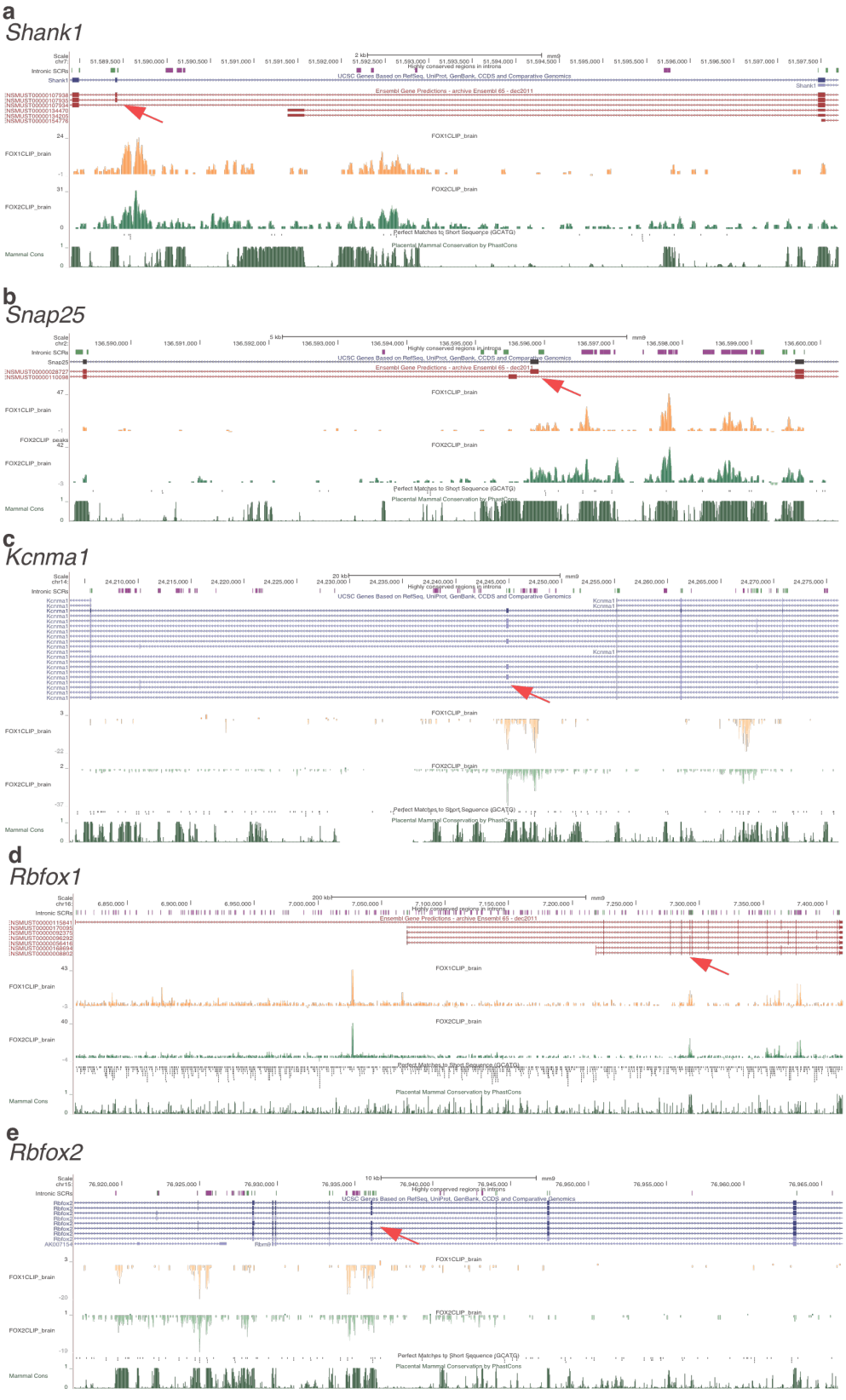

a
Shank1
b
Snap25
c
Kcnma1
d
Rbfox1
e
Rbfox2
FOX1CLIP_brain
FOX2CLIP_brain
Mammal Cons
Placental Mammal Conservation by PhastCons

Figure 20. The Rbfox binding motif UGCAUG is the most enriched hexamer in conserved regions in distal intronic space around alternatively spliced exons.

(a) A flow-chart of our computational strategy to identify highly conserved regions within the human transcriptome. (b) Pie chart showing the distribution of highly conserved regions in different gene regions. (c - d) Scatter-plot of enrichment scores for 4096 hexamers (grey points) in proximal (c) and distal (d) intronic regions. The y-axis indicates the enrichment of each word within intronic regions proximal to cassette relative to constitutive exons. The x-axis indicates the enrichment of each word within highly conserved, relative to lowly conserved regions. The five most enriched motifs in highly conserved regions proximal to AS exons compared to lowly conserved regions proximal to constitutive exons judged by an alternative *de novo* approach is inset in (c) (additional motif information is in Figure 22b). Words similar to these five motifs are highlighted with filled shapes, overlaid onto the scatter-plot. UGCAUG was significantly enriched in both proximal and distal conserved intronic regions flanking AS exons.

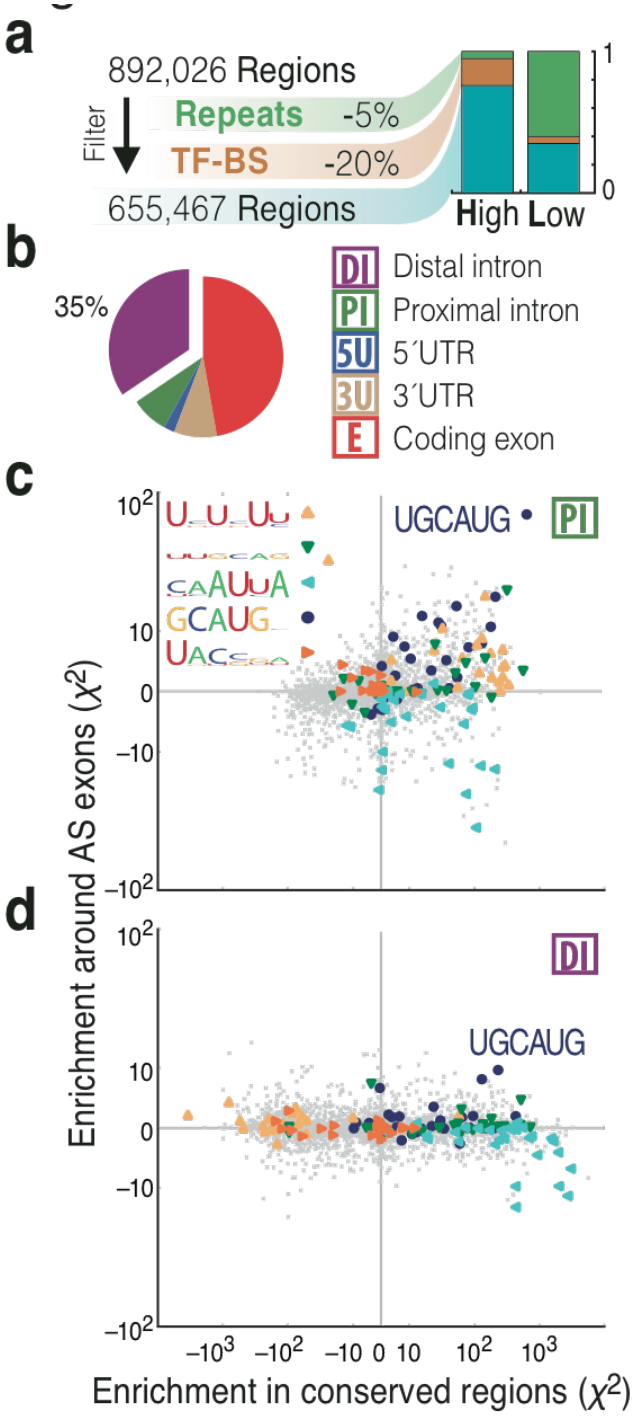

a
892,026 Regions
Filter
Repeats -5%
TF-BS -20%
655,467 Regions
1
0
High Low
b
35%
DI Distal intron
PI Proximal intron
5U 5′UTR
3U 3′UTR
E Coding exon
c
UGCAUG
PI
d
DI
UGCAUG
Enrichment around AS exons (χ2)
Enrichment in conserved regions (χ2)

Figure 21. Discovery and characterization of conserved regions.

(a) Contiguous regions within the human transcriptome were divided into three categories based on their degree of evolutionary conservation as determined by phastCons scores, *S*. Conserved regions that overlap known repetitive elements and transcription factor binding sites were removed. Pie charts illustrate the distribution of the three categories of lowly (L), moderately (M) and highly (H) conserved regions of approximately similar lengths (box-plots on the right) within different genic regions (5′UTR, exon, proximal intron, distal intron and 3′UTR) in protein-coding genes. Bar charts show the fraction of total nucleotides in each genic region covered by a highly-conserved region. (b) Results from a *de novo* motif search using HOMER (Heinz et al., 2010) for enriched hexamer motifs is shown for distal and proximal highly-conserved regions from (a). Up to 10 motifs are shown with their associated *P* values indicated to the right.

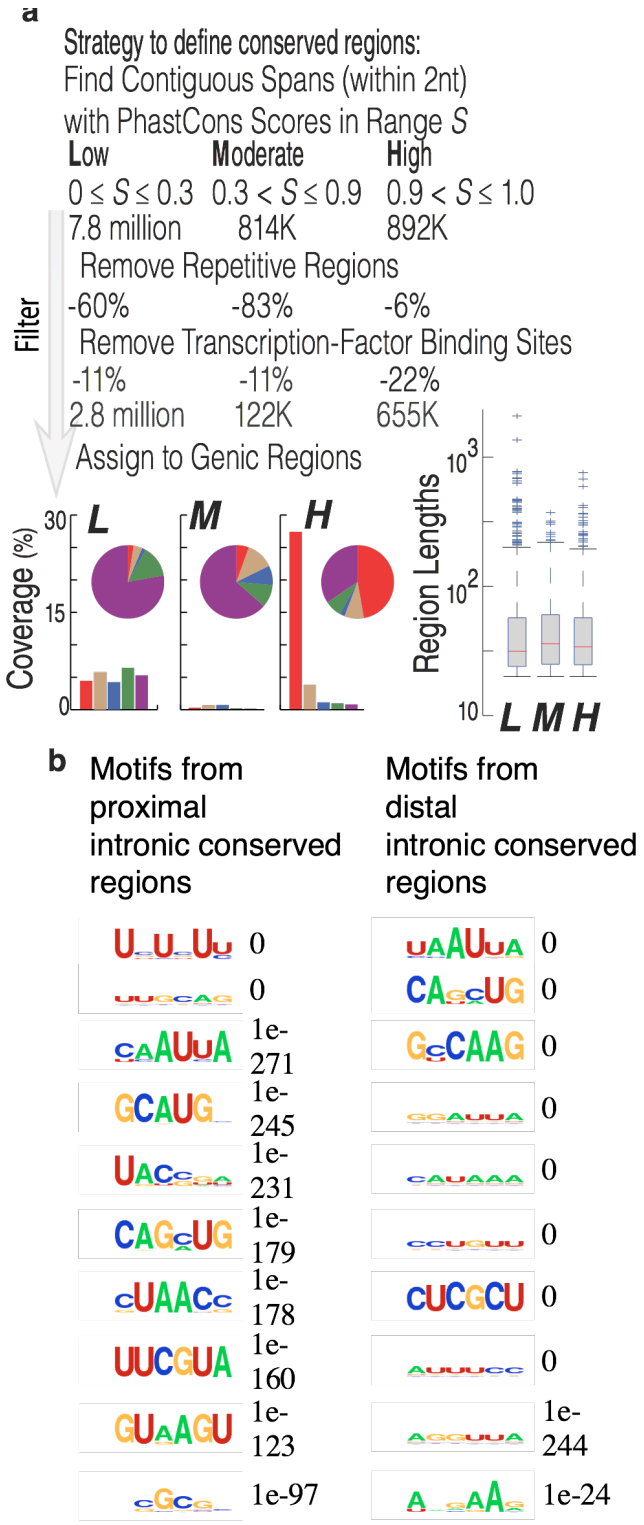

a
Strategy to define conserved regions:
Find Contiguous Spans (within 2nt)
with PhastCons Scores in Range S
Low
Moderate
High
0 ≤ S ≤ 0.3
0.3 < S ≤ 0.9
0.9 < S ≤ 1.0
7.8 million
814K
892K
Remove Repetitive Regions
-60%
-83%
-6%
Remove Transcription-Factor Binding Sites
-11%
-11%
-22%
2.8 million
122K
655K
Assign to Genic Regions
Filter
Coverage (%)
30
15
0
L
M
H
Region Lengths
10^3
10^2
10
L M H
b
Motifs from proximal intronic conserved regions
Motifs from distal intronic conserved regions
0
0
1e-271
1e-245
1e-231
1e-179
1e-178
1e-160
1e-123
1e-97
0
0
0
0
0
0
0
1e-244
1e-24

Figure 22. Identification of Rbfox-dependent expression and splicing changes by RNA-seq.

(a, b) Scatter plots of RPKM (reads per kilobase mapped in exons per million reads) values between *Rbfox1* and *Rbfox2* knockout (KO; y-axis) compared to wildtype (WT; x-axis) sibling pairs showed that 46 and 6 genes are significantly up-regulated (red points), and 11 and 9 genes are down-regulated (green points) upon loss of Rbfox1 and Rbfox2 in mouse brain, respectively (P <0.001). (c - e) Scatter plots of RPKM values comparing ectopic expression (EE) of *RBFOX1*, *RBFOX2* and *RBFOX3* (y-axes) to a plasmid control (x-axis). (f) Scatter plots of percent-spliced-in (Ψ) values of cassette exons comparing *Rbfox1* and *Rbfox2* KO (y-axes) to WT (x-axes); and ectopic expression (EE) of *RBFOX1*, *RBFOX2* and *RBFOX3* to plasmid control. Mouse exons that are alternatively spliced ($|\Delta\Psi| \geq 5\%$) upon loss of Rbfox1 (in red), Rbfox2 (in blue) and either (orange) are marked on top two plots. Human exons that are alternatively spliced upon EE of RBFOX1 (in red), RBFOX2 (in blue) and RBFOX3 (in green) are marked on the bottom three plots. A linear fit of the data (black line) and associated R_2 value are shown. (g) Venn diagrams show the number of cassette exon splicing events that are in common upon loss of Rbfox1 and Rbfox2 in mouse (top) and upon ectopic expression of RBFOX proteins in human 293T cells (bottom), considering exons with $|\Delta\Psi| \geq 5\%$ (any change, left), $\Delta\Psi \geq 5\%$ (included, right top) and $\Delta\Psi \leq -5\%$ (excluded, right bottom). *RBFOX1*-, *RBFOX2*- and *RBFOX3*-specific events are colored red, blue and green, with orange and dark green representing shared events. (h) A scatter plot shows $\Delta\Psi$ values from RT-PCR (Gehman et al., 2011; Gehman et al., 2012) (x-axis) and RNA-seq-derived RPKM (y-axis) measurements of exon inclusion for manually validated cassette splicing exon events in mouse brain for *Rbfox1* (red) and *Rbfox2* (blue). A linear fit to each set of points and associated R_2 value are shown. (i) Bar-plots show a comparison between $\Delta\Psi$ values using RT-PCR or RNA-seq. RTPCR values were obtained from (Matlin et al., 2005; Wang et al., 2008).

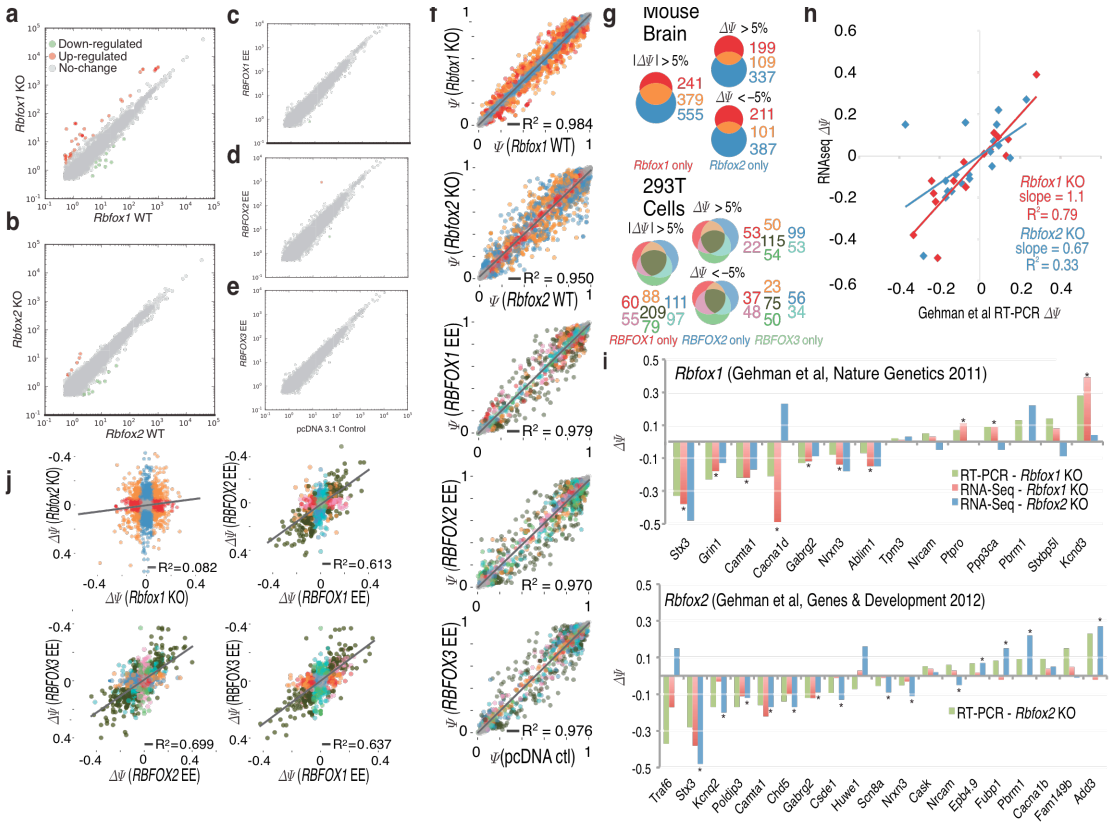

a
Down-regulated
Up-regulated
No-change
Rbfox1 KO
Rbfox1 WT
b
Rbfox2 KO
Rbfox2 WT
c
RBFOX1 EE
d
RBFOX2 EE
e
RBFOX3 EE
pcDNA 3.1 Control
f
Ψ (Rbfox1 KO)
R² = 0.984
Ψ (Rbfox1 WT)
Ψ (Rbfox2 KO)
R² = 0.950
Ψ (Rbfox2 WT)
Ψ (RBFOX1 EE)
R² = 0.979
Ψ (RBFOX2 EE)
R² = 0.970
Ψ (RBFOX3 EE)
R² = 0.976
Ψ(pcDNA ctl)
g
Mouse Brain
|ΔΨ| > 5%
241
379
555
ΔΨ > 5%
199
109
337
ΔΨ < −5%
211
101
387
Rbfox1 only
Rbfox2 only
293T Cells
|ΔΨ| > 5%
ΔΨ > 5%
ΔΨ < −5%
RBFOX1 only
RBFOX2 only
RBFOX3 only
h
RNAseq ΔΨ
Gehman et al RT-PCR ΔΨ
Rbfox1 KO
slope = 1.1
R² = 0.79
Rbfox2 KO
slope = 0.67
R² = 0.33
i
Rbfox1 (Gehman et al, Nature Genetics 2011)
ΔΨ
RT-PCR - Rbfox1 KO
RNA-Seq - Rbfox1 KO
RNA-Seq - Rbfox2 KO
Stx3
Grin1
Camta1
Cacna1d
Gabrg2
Nrxn3
Ablim1
Tpm3
Ncam
Ptpro
Ppp3ca
Pbrm1
Stxbp5l
Kcnd3
Rbfox2 (Gehman et al, Genes & Development 2012)
RT-PCR - Rbfox2 KO
Traf6
Stx3
Kcnq2
Poldip3
Camta1
Chd5
Gabrg2
Csde1
Huwe1
Scn8a
Nrxn3
Cask
Nrcam
Epb4.9
Fubp1
Pbrm1
Cacna1b
Fam149b
Add3
j
ΔΨ (Rbfox2 KO)
ΔΨ (Rbfox1 KO)
R²=0.082
ΔΨ (RBFOX2 EE)
ΔΨ (RBFOX1 EE)
R²=0.613
ΔΨ (RBFOX3 EE)
ΔΨ (RBFOX2 EE)
R²=0.699
ΔΨ (RBFOX1 EE)
R²=0.637

Figure 23. Both proximal and distal Rbfox motifs regulate splicing.

(a - d) Bar plots depict the fraction of cassette exons that have evidences (listed below the first column of each panel) for direct Rbfox regulation within proximal or distal intronic regions, up or downstream (listed above each panel). Exons were classified as differentially included ($\Delta\Psi \geq 5\%$; blue, on the positive y-axis), excluded ($\Delta\Psi \leq -5\%$; goldenrod reflected on the negative y-axis) or not changing ($-2\% < \Delta\Psi < 2\%$; grey, on the positive and y-axis and also reflected on the negative y-axis) according to *Rbfox2* RNA-seq experiments in mouse (a, b) and human (c,d). Ties with one or two "*" symbols mark significance at $P < 0.05$ and $P < 0.001$, respectively, for a Fisher's exact test comparing the relative proportion of changed versus unaffected exons which possess a particular feature in the indicated intronic region. A full accounting of analyses using other RNA-seq experiments and other features is in Supplementary Figure 5. (e - f) Cumulative distributions of $\Delta\Psi$ values for mouse (E) and human (G) cassette exons which have conserved GCAUG motifs (BL score > 0.2 in mouse BL score > 0.1 in human) in proximal (left) or distal (right) regions upstream (purple) or downstream (orange) or with no motifs at all in that region (grey).

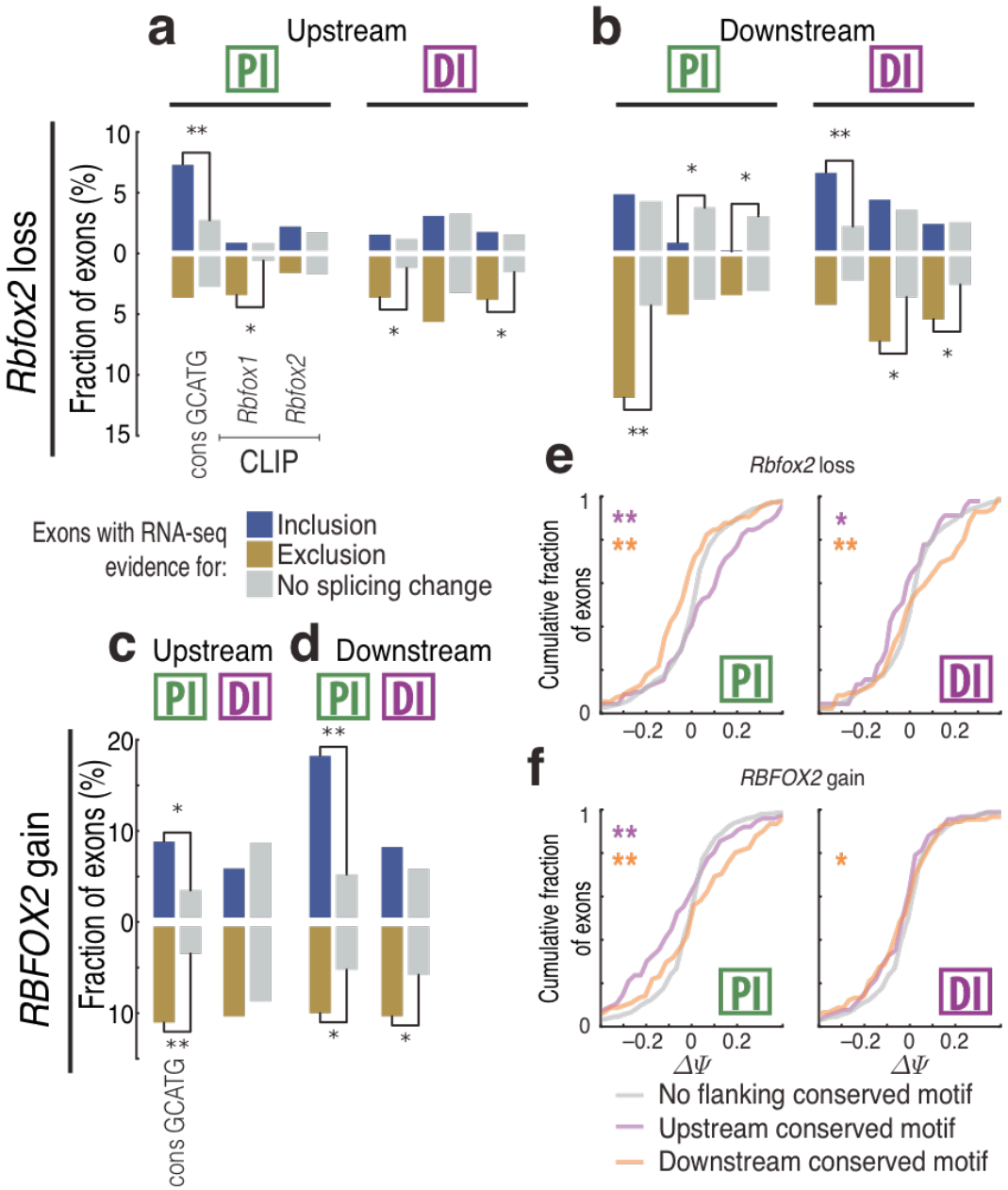

a
Upstream
PI
DI
b
Downstream
PI
DI
Rbfox2 loss
Fraction of exons (%)
10
5
0
5
10
15
cons GCATG
Rbfox1
Rbfox2
CLIP
Exons with RNA-seq evidence for:
Inclusion
Exclusion
No splicing change
c
Upstream
d
Downstream
RBFOX2 gain
Fraction of exons (%)
20
10
0
10
cons GCATG
e
Rbfox2 loss
Cumulative fraction of exons
1
0
−0.2
0
0.2
f
RBFOX2 gain
ΔΨ
No flanking conserved motif
Upstream conserved motif
Downstream conserved motif

Figure 24. Distal conserved regions containing Rbfox sites control splicing of upstream alternative exons.

(a, d) Genomic regions showing *KIF21A* exon 23 (a) and *ENAH* exon 11a (d) and neighboring intronic and exonic regions. The location of GCAUG sequences in both genes is marked by orange bars. Rbfox protein-RNA binding sites that overlap distal conserved GCAUG motifs are outlined with a black box, and match the highest density of CLIP-seq reads (graphed as continuous densities) for RBFOX2 in 293T cells (iCLIP; green track), Rbfox1 in mouse brain (CLIP-seq; red track), and Rbfox2 in mouse brain (CLIP-seq; blue track). PhastCons scores of evolutionary conservation are represented as continuous densities in dark green. (b, e) Phylogenetic conservation of UGCAUG (red) and GCAUG (blue) elements within the highlighted distal intronic regions in the *KIF21A* and *ENAH* gene (boxed in (a) and (d)). (c, f) Cartoon representations of binding sites for MOs or vMOs targeted to block distal RBFOX sites in *KIF21A* (c) and *ENAH* (f) are shown. Bottom panels show RT-PCR analysis of *KIF21A* and *ENAH* splicing in the presence or absence of morpholinos. Ψ is listed below each lane. (g) Three-exon minigenes consisting of E11-E11a-E12 are illustrated. Ovals represent conserved (filled) or non-conserved (unfilled) UGCAUG (red) and GCAUG (blue) sequences. (h) RT-PCR analyses of minigene-derived transcripts. Ψ is listed below each lane, as quantified by image densitometric analysis. (i) Pull-down assay measuring in vitro-translated RBFOX2 protein binding to biotinylated RNA containing consensus UGCAUG motifs (lane 1) or mutated RBFOX motifs (lane 2).

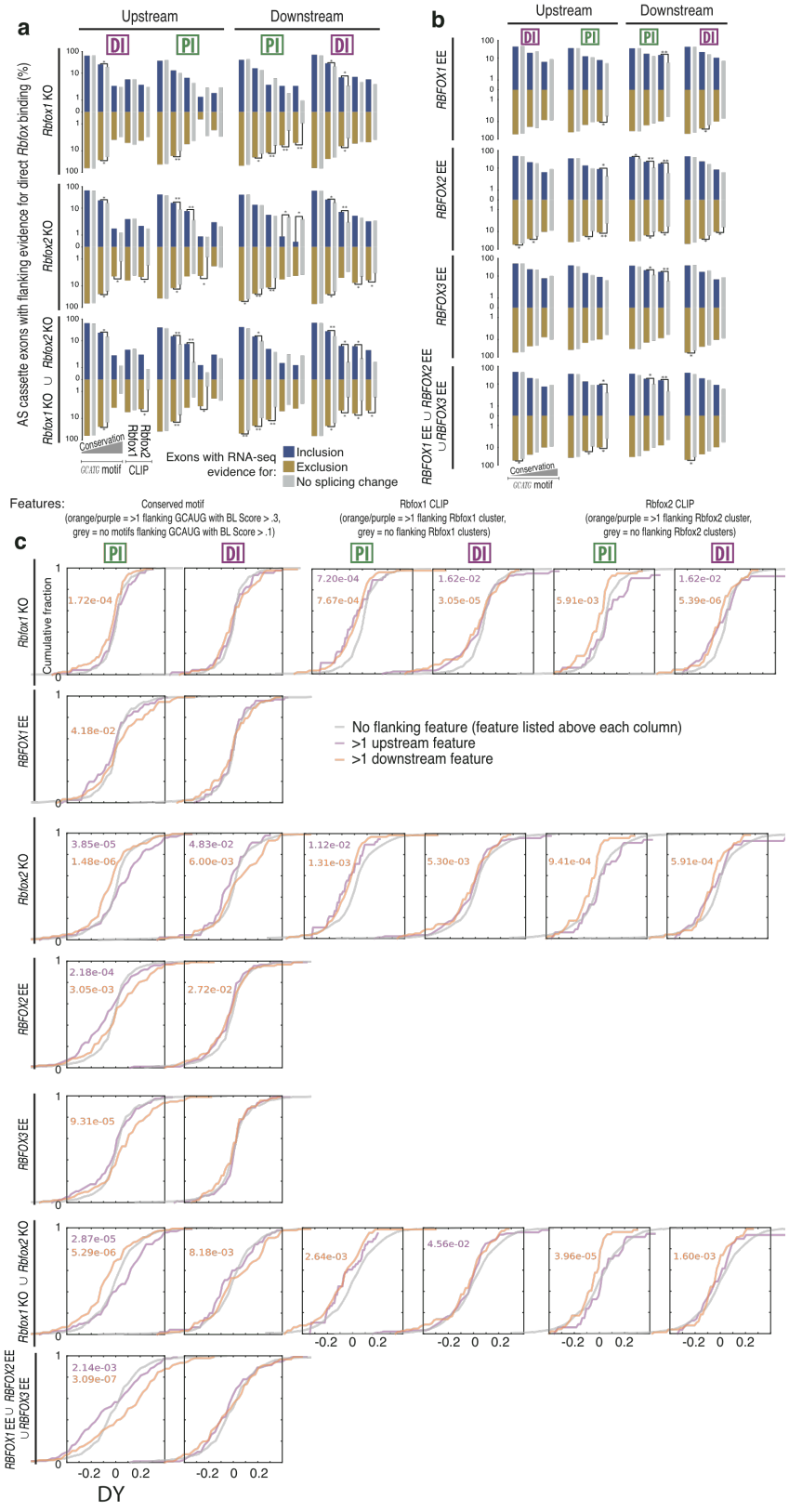

a
Upstream
Downstream
DI
PI
AS cassette exons with flanking evidence for direct Rbfox binding (%)
Rbfox1 KO
Rbfox2 KO
Rbfox2 KO ∪ Rbfox1 KO
Conservation
GCAUG motif
Rbfox2 CLIP
Rbfox1 CLIP
Exons with RNA-seq evidence for:
Inclusion
Exclusion
No splicing change
b
RBFOX1 EE
RBFOX2 EE
RBFOX3 EE
RBFOX1 EE ∪ RBFOX2 EE ∪ RBFOX3 EE
Features:
Conserved motif
(orange/purple = >1 flanking GCAUG with BL Score > .3,
grey = no motifs flanking GCAUG with BL Score > .1)
Rbfox1 CLIP
(orange/purple = >1 flanking Rbfox1 cluster,
grey = no flanking Rbfox1 clusters)
Rbfox2 CLIP
(orange/purple = >1 flanking Rbfox2 cluster,
grey = no flanking Rbfox2 clusters)
c
Cumulative fraction
1.72e-04
7.20e-04
7.67e-04
1.62e-02
3.05e-05
5.91e-03
1.62e-02
5.39e-06
4.18e-02
No flanking feature (feature listed above each column)
>1 upstream feature
>1 downstream feature
3.85e-05
1.48e-06
4.83e-02
6.00e-03
1.12e-02
1.31e-03
5.30e-03
9.41e-04
5.91e-04
2.18e-04
3.05e-03
2.72e-02
9.31e-05
2.87e-05
5.29e-06
8.18e-03
2.64e-03
4.56e-02
3.96e-05
1.60e-03
2.14e-03
3.09e-07
-0.2 0 0.2
DY

Figure 25. Distal association of *Rbfox* sites with *Rbfox*-dependent alternatively spliced exons.

(a, b) The fractions of alternatively spliced exons compared to unchanged that have GCAUG motifs (at increasing levels of conservation) and CLIP-defined Rbfox1 and Rbfox2 sites either upstream or downstream, for both distal ("DI") and proximal ("PI") regions, are represented as vertical bars. *Rbfox1* and *Rbfox2* knockout (KO) mice compared to sibling WT pair are represented in (a). Ectopic expression (EE) of *RBFOX1*, *RBFOX2* and *RBFOX3* are represented in (B). The fractions are represented on the log-scale (y-axis). Blue bars represent exons with RNA-seq evidence for inclusion ($\Delta\Psi > 5\%$); goldenrod bars represent exons with RNA-seq evidence for exclusion ($\Delta\Psi < -5\%$); grey bars represent exons with no change by RNA-seq ($-2\% < \Delta\Psi < 2\%$). Ties with "*" symbols indicate statistically significant differences when categories of exons are compared (*$P < 0.05$, **$P < 0.001$; Fisher's exact test). The maximal change for each exon was used when combining evidence from all experiments, as represented by their unions (U; y-axis) (the final rows for (a) and (b)). (c) The cumulative distributions of $\Delta\Psi$ values for exons with >1 GCAUG motif or CLIP-defined Rbfox binding site within either upstream or downstream, for both distal ("DI") and proximal ("PI") regions were compared to exons with no flanking conserved GCAUG motifs. *P* values by the two-sample Kolgomorov- Smirnov tests are indicated when statistically significant ($P < 0.05$).

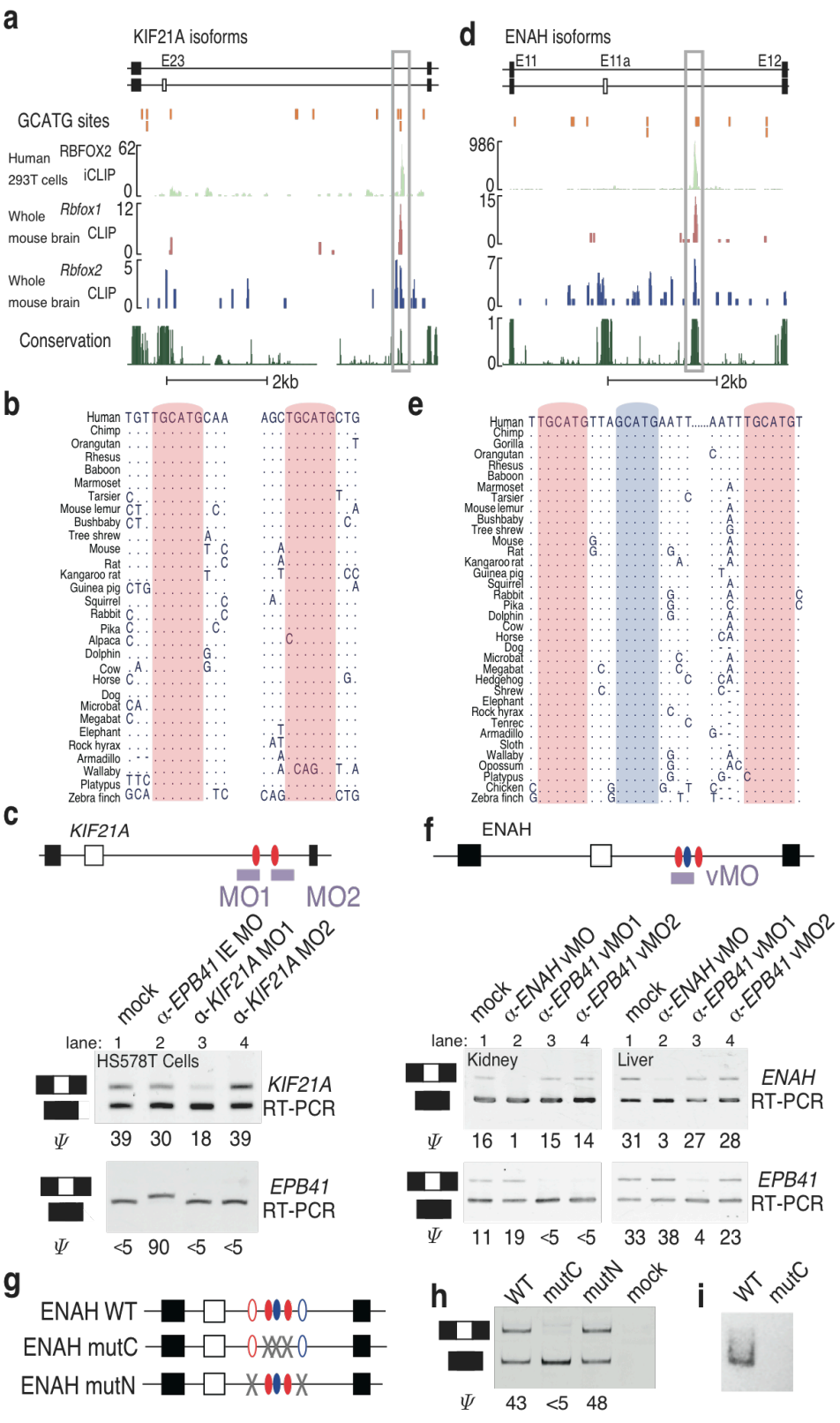
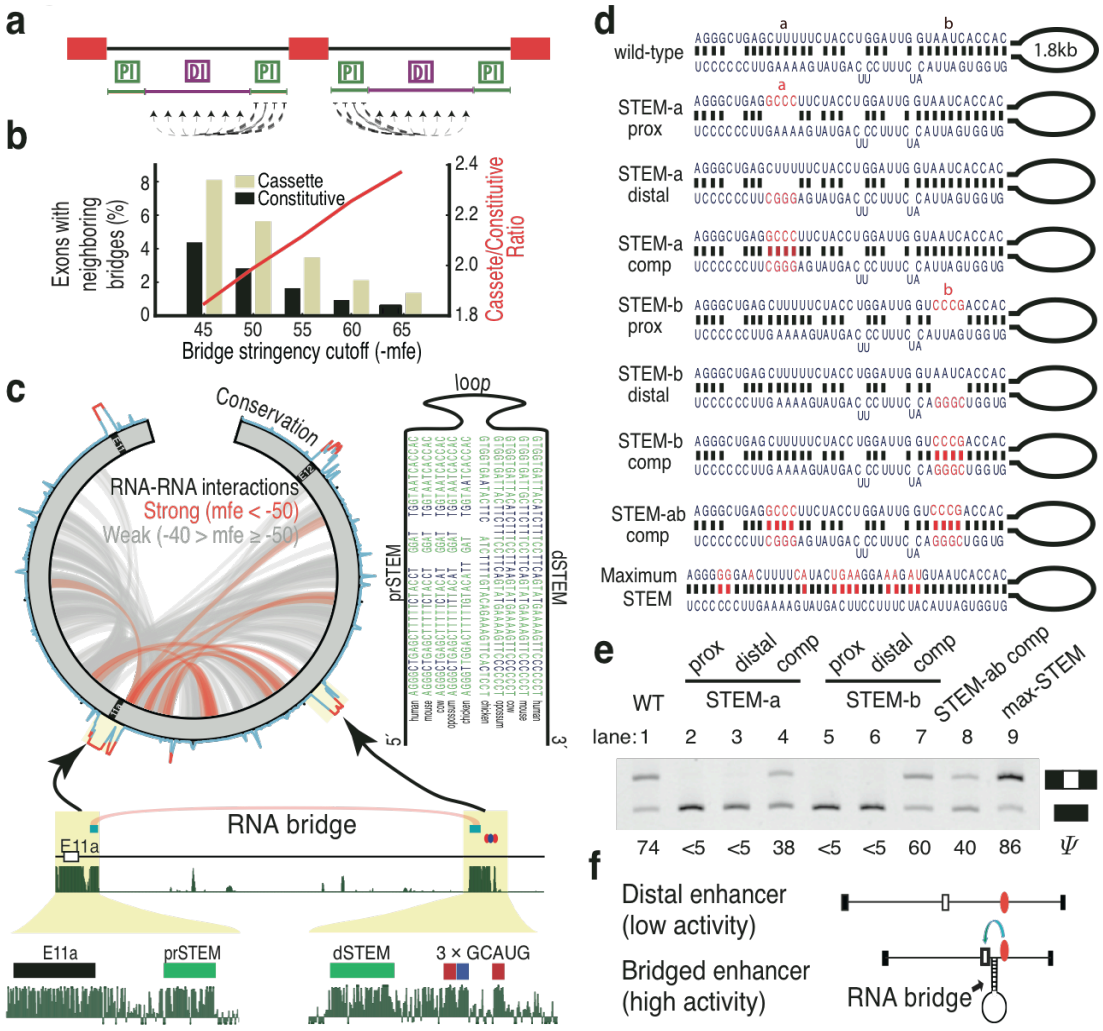

a
KIF21A isoforms
E23
GCATG sites
Human 293T cells RBFOX2 iCLIP
Whole mouse brain Rbfox1 CLIP
Whole mouse brain Rbfox2 CLIP
Conservation
2kb
d
ENAH isoforms
E11
E11a
E12
2kb
b
Human TGTTGCATGCAA AGCTGCATGCTG
e
Human TTGCATGTTAGCATGAATT.....AATTTGCATGT
c
KIF21A
MO1
MO2
mock
α-EPB41 IE MO
α-KIF21A MO1
α-KIF21A MO2
lane: 1 2 3 4
HS578T Cells
KIF21A RT-PCR
Ψ 39 30 18 39
EPB41 RT-PCR
Ψ <5 90 <5 <5
f
ENAH
vMO
mock
α-ENAH vMO
α-EPB41 vMO1
α-EPB41 vMO2
lane: 1 2 3 4 1 2 3 4
Kidney
Liver
ENAH RT-PCR
Ψ 16 1 15 14 31 3 27 28
EPB41 RT-PCR
Ψ 11 19 <5 <5 33 38 4 23
g
ENAH WT
ENAH mutC
ENAH mutN
h
WT
mutC
mutN
mock
Ψ 43 <5 48
i
WT
mutC

Figure 26. An RNA-bridge between ENAH E11a and a conserved distal RBFOX site is necessary for exon inclusion.

(a) A schematic of the strategy used to find RNA-bridges. Regions proximal to exons were tested for the ability to pair to all positions in the distal region in the same intron. (b) The fraction of cassette (yellow) and constitutive (black) exons with neighboring predicted RNA-bridges as the negative minimum free energy (mfe) threshold for defining an RNA-bridge is made more stringent. The ratio between these fractions is depicted as a red line. (c) All predicted RNA-RNA interactions within E11-E11a-E12 (green) of the *ENAH* pre-mRNA are displayed in a circos plot (left side). Paired regions are classified as strong, (mfe $\leq$ –50 kcal per mol; red) or weak (mfe > –50 kcal per mol; grey). PhastCons conservation scores are illustrated on the circumference of the circle. A stem-loop structure that is conserved to avian genomes is shown with base-paired nucleotides indicated with green font. The location of this RNA-bridge is shown in detail at the bottom of the panel. Vertebrate conservation from phastCons is represented as continuous density in dark green. (d) Wild-type and mutated (red letters) RNA duplex structures are shown. (e) RT-PCR analysis of *ENAH* structural mutants in transfected T47D cells. Labels above each numbered lane correspond to experiments using each of the structures in (d). Ψ is listed below each lane. (f) Model illustrating the function of the RNA-bridge to position Rbfox sites (red ovals) close to an exon to regulate splicing.


a
PI
DI
PI
PI
DI
PI
b
Exons with neighboring bridges (%)
Cassette
Constitutive
Cassete/Constitutive Ratio
Bridge stringency cutoff (-mfe)
c
Conservation
RNA-RNA interactions
Strong (mfe < -50)
Weak (-40 > mfe ≥ -50)
loop
prSTEM
dSTEM
RNA bridge
E11a
E11a
prSTEM
dSTEM
3 × GCAUG
d
wild-type
1.8kb
STEM-a prox
STEM-a distal
STEM-a comp
STEM-b prox
STEM-b distal
STEM-b comp
STEM-ab comp
Maximum STEM
e
prox
distal
comp
prox
distal
comp
STEM-ab comp
max-STEM
WT
STEM-a
STEM-b
lane: 1 2 3 4 5 6 7 8 9
74 <5 <5 38 <5 <5 60 40 86 Ψ
f
Distal enhancer (low activity)
Bridged enhancer (high activity)
RNA bridge

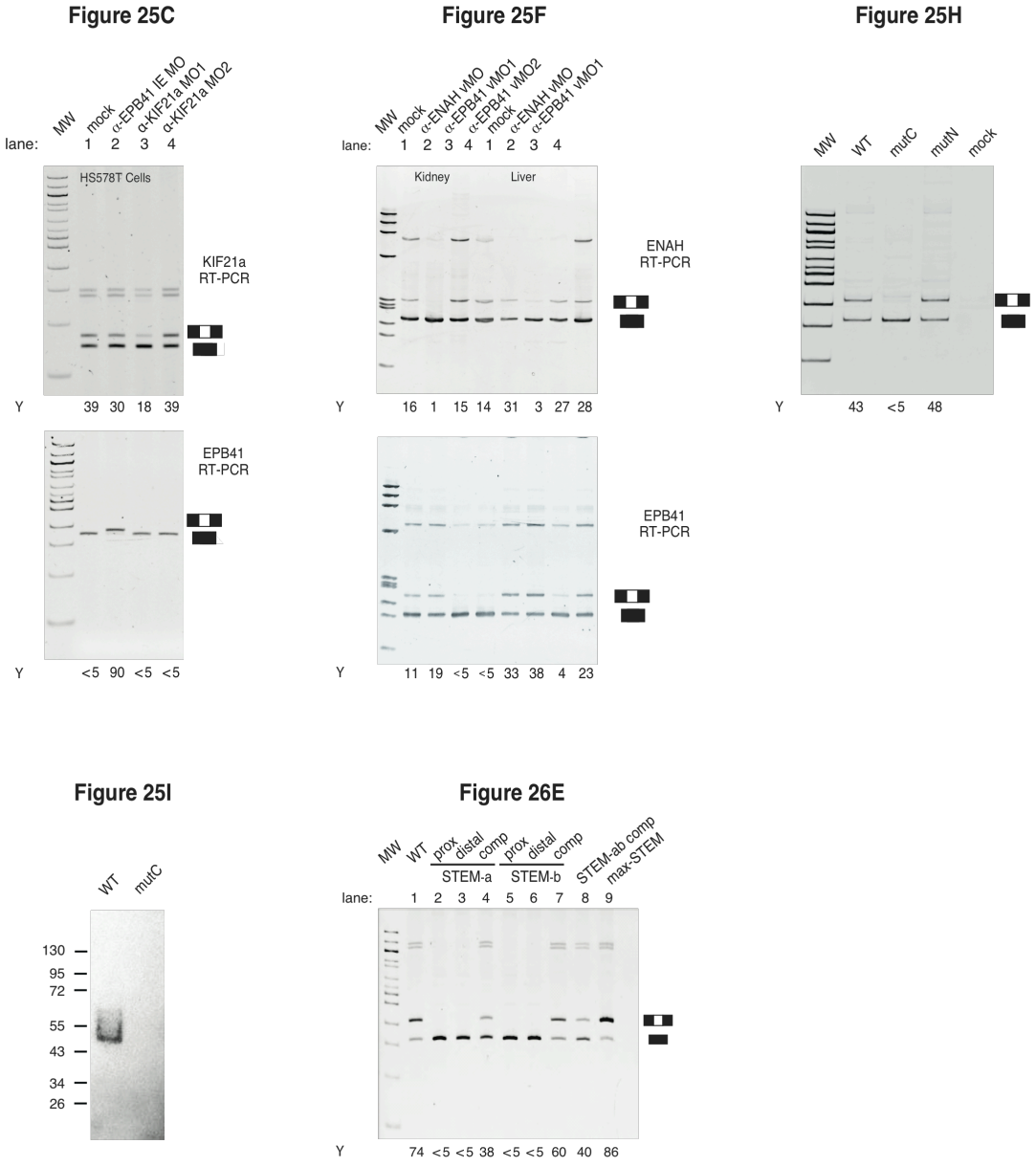


Figure 27. Un-cropped gel images.

Un-cropped gel images are shown for RT-PCR and biotin pull-down assays. These include slow-migrating bands, suspected heteroduplexes, which are excluded from main figures.

ACKNOWLEDGEMENTS

I would like to thank Amy E. Pasquinelli, Neil. C. Chi, Karl Willert and Lawrence Goldstein, and members of the Yeo, Conboy and Goldstein labs for critical reading of the manuscript. Michael T. Lovci is supported as a National Science Foundation GK12 Fellow. This work was supported by grants from the National Institute of Health, to Gene W. Yeo (U54 HG007005, R01 HG004659, R01 GM084317 and R01 NS075449), to John G. Conboy (HL045182 and DK094699) and partially supported by grants to Joe W. Gray (CA112970 and CA126551). J.G.C. also acknowledges support from DK032094. This work was also supported by the Director, Office of Science, and Office of Biological & Environmental Research of the U.S. Department of Energy under Contract No. DE-AC02-05CH1123. M.T.L. and G.W.Y. are grateful for a gift from Patrick Yang at Genentech that supported M.T.L. G.W.Y. is supported as an Alfred P. Sloan Research Fellow. M.T.L. and Gabriel A. Pratt conducted the bioinformatics analyses. Dana Ghanem, Henry Marr, Justin A. Arnold, Sherry Gee, Marilyn Parra, Tiffany Y. Liang, Thomas J. Stark, Shawn Hoon and Katlin B. Massirer conducted biological experiments. Lauren T. Gehman and Douglas L. Black generated the Rbfox mutant mice and isolated brain RNA. J.G.C., J.W.G. and G.W.Y. designed the study. M.T.L., D.G., J.G.C. and G.W.Y. wrote the manuscript with input from all authors.

REFERENCES

Abbott, A.L., Alvarez-Saavedra, E., Miska, E.A., Lau, N.C., Bartel, D.P., Horvitz, H.R., and Ambros, V. (2005). The let-7 MicroRNA family members mir-48, mir-84, and mir-241 function together to regulate developmental timing in Caenorhabditis elegans. Developmental cell *9*, 403-414.

Abrahante, J.E., Daul, A.L., Li, M., Volk, M.L., Tennessen, J.M., Miller, E.A., and Rougvie, A.E. (2003). The Caenorhabditis elegans hunchback-like gene lin-57/hbl-1 controls developmental time and is regulated by microRNAs. Developmental cell *4*, 625-637.

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.

Bagga, S., Bracht, J., Hunter, S., Massirer, K., Holtz, J., Eachus, R., and Pasquinelli, A.E. (2005). Regulation by let-7 and lin-4 miRNAs results in target mRNA degradation. Cell *122*, 553-563.

Baraniak, A.P., Lasda, E.L., Wagner, E.J., and Garcia-Blanco, M.A. (2003). A stem structure in fibroblast growth factor receptor 2 transcripts mediates cell-type-specific splicing by approximating intronic control elements. Molecular and cellular biology *23*, 9327-9337.

Barash, Y., Calarco, J.A., Gao, W., Pan, Q., Wang, X., Shai, O., Blencowe, B.J., and Frey, B.J. (2010). Deciphering the splicing code. Nature *465*, 53-59.

Bartel, D.P. (2009). MicroRNAs: target recognition and regulatory functions. Cell *136*, 215-233.

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

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

Bhalla, K., Phillips, H.A., Crawford, J., McKenzie, O.L., Mulley, J.C., Eyre, H., Gardner, A.E., Kremmidiotis, G., and Callen, D.F. (2004). The de novo chromosome 16 translocations of two patients with abnormal phenotypes (mental retardation and epilepsy) disrupt the A2BP1 gene. Journal of human genetics *49*, 308-311.

Black, D.L. (2003). Mechanisms of alternative pre-messenger RNA splicing. Annual review of biochemistry *72*, 291-336.

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

Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M.C., Maeda, N., Oyama, R., Ravasi, T., Lenhard, B., Wells, C., Kodzius, R., Shimokawa, K., Bajic, V.B., Brenner, S.E., Batalov, S., Forrest, A.R., Zavolan, M., Davis, M.J., Wilming, L.G., Aidinis, V., Allen, J.E., Ambesi-Impiombato, A., Apweiler, R., Aturaliya, R.N., Bailey, T.L., Bansal, M., Baxter, L., Beisel, K.W., Bersano, T., Bono, H., Chalk, A.M., Chiu, K.P., Choudhary, V., Christoffels, A., Clutterbuck, D.R., Crowe, M.L., Dalla, E., Dalrymple,

B.P., de Bono, B., Della Gatta, G., di Bernardo, D., Down, T., Engstrom, P., Fagiolini, M., Faulkner, G., Fletcher, C.F., Fukushima, T., Furuno, M., Futaki, S., Gariboldi, M., Georgii-Hemming, P., Gingeras, T.R., Gojobori, T., Green, R.E., Gustincich, S., Harbers, M., Hayashi, Y., Hensch, T.K., Hirokawa, N., Hill, D., Huminiecki, L., Iacono, M., Ikeo, K., Iwama, A., Ishikawa, T., Jakt, M., Kanapin, A., Katoh, M., Kawasawa, Y., Kelso, J., Kitamura, H., Kitano, H., Kollias, G., Krishnan, S.P., Kruger, A., Kummerfeld, S.K., Kurochkin, I.V., Lareau, L.F., Lazarevic, D., Lipovich, L., Liu, J., Liuni, S., McWilliam, S., Madan Babu, M., Madera, M., Marchionni, L., Matsuda, H., Matsuzawa, S., Miki, H., Mignone, F., Miyake, S., Morris, K., Mottagui-Tabar, S., Mulder, N., Nakano, N., Nakauchi, H., Ng, P., Nilsson, R., Nishiguchi, S., Nishikawa, S., Nori, F., Ohara, O., Okazaki, Y., Orlando, V., Pang, K.C., Pavan, W.J., Pavesi, G., Pesole, G., Petrovsky, N., Piazza, S., Reed, J., Reid, J.F., Ring, B.Z., Ringwald, M., Rost, B., Ruan, Y., Salzberg, S.L., Sandelin, A., Schneider, C., Schonbach, C., Sekiguchi, K., Semple, C.A., Seno, S., Sessa, L., Sheng, Y., Shibata, Y., Shimada, H., Shimada, K., Silva, D., Sinclair, B., Sperling, S., Stupka, E., Sugiura, K., Sultana, R., Takenaka, Y., Taki, K., Tammoja, K., Tan, S.L., Tang, S., Taylor, M.S., Tegner, J., Teichmann, S.A., Ueda, H.R., van Nimwegen, E., Verardo, R., Wei, C.L., Yagi, K., Yamanishi, H., Zabarovsky, E., Zhu, S., Zimmer, A., Hide, W., Bult, C., Grimmond, S.M., Teasdale, R.D., Liu, E.T., Brusic, V., Quackenbush, J., Wahlestedt, C., Mattick, J.S., Hume, D.A., Kai, C., Sasaki, D., Tomaru, Y., Fukuda, S., Kanamori-Katayama, M., Suzuki, M., Aoki, J., Arakawa, T., Iida, J., Imamura, K., Itoh, M., Kato, T., Kawaji, H., Kawagashira, N., Kawashima, T., Kojima, M., Kondo, S., Konno, H., Nakano, K., Ninomiya, N., Nishio, T., Okada, M., Plessy, C., Shibata, K., Shiraki, T., Suzuki, S., Tagami, M., Waki, K., Watahiki, A., Okamura-Oho, Y., Suzuki, H., Kawai, J., and Hayashizaki, Y. (2005). The transcriptional landscape of the mammalian genome. Science *309*, 1559-1563.

Chang, S., Johnston, R.J., Jr., Frokjaer-Jensen, C., Lockery, S., and Hobert, O. (2004). MicroRNAs act sequentially and asymmetrically to control chemosensory laterality in the nematode. Nature *430*, 785-789.

Chekulaeva, M., and Filipowicz, W. (2009). Mechanisms of miRNA-mediated post-transcriptional regulation in animal cells. Current opinion in cell biology *21*, 452-460.

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, J., Kapranov, P., Drenkow, J., Dike, S., Brubaker, S., Patel, S., Long, J., Stern, D., Tammana, H., Helt, G., Sementchenko, V., Piccolboni, A., Bekiranov, S., Bailey, D.K., Ganesh, M., Ghosh, S., Bell, I., Gerhard, D.S., and Gingeras, T.R. (2005). Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. Science *308*, 1149-1154.

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

Cote, J., Dupuis, S., Jiang, Z., and Wu, J.Y. (2001). Caspase-2 pre-mRNA alternative splicing: Identification of an intronic element containing a decoy 3' acceptor site. Proceedings of the National Academy of Sciences of the United States of America *98*, 938-943.

Dale, R.K., Pedersen, B.S., and Quinlan, A.R. (2011). Pybedtools: a flexible Python library for manipulating genomic datasets and annotations. Bioinformatics *27*, 3423-3424.

Damianov, A., and Black, D.L. (2010). Autoregulation of Fox protein expression to produce dominant negative splicing factors. RNA *16*, 405-416.

Davis, L.K., Maltman, N., Mosconi, M.W., Macmillan, C., Schmitt, L., Moore, K., Francis, S.M., Jacob, S., Sweeney, J.A., and Cook, E.H. (2012). Rare inherited A2BP1 deletion in a proband with autism and developmental hemiparesis. American journal of medical genetics Part A *158A*, 1654-1661.

De Bona, F., Ossowski, S., Schneeberger, K., and Ratsch, G. (2008). Optimal spliced alignments of short sequence reads. Bioinformatics *24*, i174-180.

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

Ding, X.C., and Grosshans, H. (2009). Repression of C. elegans microRNA targets at the initiation level of translation requires GW182 proteins. The EMBO journal *28*, 213-222.

Dirksen, W.P., Mohamed, S.A., and Fisher, S.A. (2003). Splicing of a myosin phosphatase targeting subunit 1 alternative exon is regulated by intronic cis-elements and a novel bipartite exonic enhancer/silencer element. The Journal of biological chemistry *278*, 9722-9732.

Dominski, Z., and Marzluff, W.F. (2007). Formation of the 3' end of histone mRNA: getting closer to the end. Gene *396*, 373-390.

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.

Gallagher, T.L., Arribere, J.A., Geurts, P.A., Exner, C.R., McDonald, K.L., Dill, K.K., Marr, H.L., Adkar, S.S., Garnett, A.T., Amacher, S.L., and Conboy, J.G. (2011). Rbfox-regulated alternative splicing is critical for zebrafish cardiac and skeletal muscle functions. Developmental biology *359*, 251-261.

Gehman, L.T., Meera, P., Stoilov, P., Shiue, L., O'Brien, J.E., Meisler, M.H., Ares, M., Jr., Otis, T.S., and Black, D.L. (2012). The splicing regulator Rbfox2 is required for both cerebellar development and mature motor function. Genes & development *26*, 445-460.

Gehman, L.T., Stoilov, P., Maguire, J., Damianov, A., Lin, C.H., Shiue, L., Ares, M., Jr., Mody, I., and Black, D.L. (2011). The splicing regulator Rbfox1 (A2BP1) controls neuronal excitation in the mammalian brain. Nature genetics *43*, 706-711.

Goguel, V., and Rosbash, M. (1993). Splice site choice and splicing efficiency are positively influenced by pre-mRNA intramolecular base pairing in yeast. Cell *72*, 893-901.

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

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

Grosshans, H., Johnson, T., Reinert, K.L., Gerstein, M., and Slack, F.J. (2005). The temporal patterning microRNA let-7 regulates several transcription factors at the larval to adult transition in C. elegans. Developmental cell *8*, 321-330.

Gu, S., Jin, L., Zhang, F., Sarnow, P., and Kay, M.A. (2009). Biological basis for restriction of microRNA targets to the 3' untranslated region in mammalian mRNAs. Nature structural & molecular biology *16*, 144-150.

Guo, N., and Kawamoto, S. (2000). An intronic downstream enhancer promotes 3' splice site usage of a neural cell-specific exon. The Journal of biological chemistry *275*, 33641-33649.

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

Hammell, M., Long, D., Zhang, L., Lee, A., Carmack, C.S., Han, M., Ding, Y., and Ambros, V. (2008). mirWIP: microRNA target prediction based on microRNA-containing ribonucleoprotein-enriched transcripts. Nature methods *5*, 813-819.

Hayes, G.D., Frand, A.R., and Ruvkun, G. (2006). The mir-84 and let-7 paralogous microRNA genes of Caenorhabditis elegans direct the cessation of molting via the conserved nuclear hormone receptors NHR-23 and NHR-25. Development *133*, 4631-4641.

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

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.

Hillier, L.W., Marth, G.T., Quinlan, A.R., Dooling, D., Fewell, G., Barnett, D., Fox, P., Glasscock, J.I., Hickenbotham, M., Huang, W., Magrini, V.J., Richt, R.J., Sander, S.N., Stewart, D.A., Stromberg, M., Tsung, E.F., Wylie, T., Schedl, T., Wilson, R.K., and Mardis, E.R. (2008). Whole-genome sequencing and variant discovery in C. elegans. Nature methods *5*, 183-188.

Hillier, L.W., Reinke, V., Green, P., Hirst, M., Marra, M.A., and Waterston, R.H. (2009). Massively parallel sequencing of the polyadenylated transcriptome of C. elegans. Genome research *19*, 657-666.

Hoell, J.I., Larsson, E., Runge, S., Nusbaum, J.D., Duggimpudi, S., Farazi, T.A., Hafner, M., Borkhardt, A., Sander, C., and Tuschl, T. (2011). RNA targets of wild-type and mutant FET family proteins. Nature structural & molecular biology *18*, 1428-1431.

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

Johansson, J.U., Ericsson, J., Janson, J., Beraki, S., Stanic, D., Mandic, S.A., Wikstrom, M.A., Hokfelt, T., Ogren, S.O., Rozell, B., Berggren, P.O., and Bark, C. (2008). An ancient duplication of exon 5 in the Snap25 gene is required for complex neuronal development/function. PLoS genetics *4*, e1000278.

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

Johnson, S.M., Lin, S.Y., and Slack, F.J. (2003). The time of appearance of the C. elegans let-7 microRNA is transcriptionally controlled utilizing a temporal regulatory element in its promoter. Developmental biology *259*, 364-379.

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. Proceedings of the National Academy of Sciences of the United States of America *104*, 19291-19296.

Karolchik, D., Baertsch, R., Diekhans, M., Furey, T.S., Hinrichs, A., Lu, Y.T., Roskin, K.M., Schwartz, M., Sugnet, C.W., Thomas, D.J., Weber, R.J., Haussler, D., and Kent, W.J. (2003). The UCSC Genome Browser Database. Nucleic acids research *31*, 51-54.

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

Kim, K.K., Kim, Y.C., Adelstein, R.S., and Kawamoto, S. (2011). Fox-3 and PSF interact to activate neural cell-specific alternative splicing. Nucleic acids research *39*, 3064-3078.

Kim, S.K., Lund, J., Kiraly, M., Duke, K., Jiang, M., Stuart, J.M., Eizinger, A., Wylie, B.N., and Davidson, G.S. (2001). A gene expression map for Caenorhabditis elegans. Science *293*, 2087-2092.

Kloosterman, W.P., Wienholds, E., Ketting, R.F., and Plasterk, R.H. (2004). Substrate requirements for let-7 function in the developing zebrafish embryo. Nucleic acids research *32*, 6284-6291.

Konig, J., Zarnack, K., Rot, G., Curk, T., Kayikci, M., Zupan, B., Turner, D.J., Luscombe, N.M., and Ule, J. (2010). iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. Nature structural & molecular biology *17*, 909-915.

Kreahling, J.M., and Graveley, B.R. (2005). The iStem, a long-range RNA secondary structure element required for efficient exon inclusion in the Drosophila Dscam pre-mRNA. Molecular and cellular biology *25*, 10251-10260.

Krek, A., Grun, D., Poy, M.N., Wolf, R., Rosenberg, L., Epstein, E.J., MacMenamin, P., da Piedade, I., Gunsalus, K.C., Stoffel, M., and Rajewsky, N. (2005). Combinatorial microRNA target predictions. Nature genetics *37*, 495-500.

Lagier-Tourenne, C., Polymenidou, M., Hutt, K.R., Vu, A.Q., Baughn, M., Huelga, S.C., Clutario, K.M., Ling, S.C., Liang, T.Y., Mazur, C., Wancewicz, E., Kim, A.S., Watt, A., Freier, S., Hicks, G.G., Donohue, J.P., Shiue, L., Bennett, C.F., Ravits, J., Cleveland, D.W., and Yeo, G.W. (2012). Divergent roles of ALS-linked proteins FUS/TLS and TDP-43 intersect in processing long pre-mRNAs. Nature neuroscience *15*, 1488-1497.

Lall, S., Grun, D., Krek, A., Chen, K., Wang, Y.L., Dewey, C.N., Sood, P., Colombo, T., Bray, N., Macmenamin, P., Kao, H.L., Gunsalus, K.C., Pachter, L., Piano, F., and Rajewsky, N. (2006). A genome-wide map of conserved microRNA targets in C. elegans. Current biology : CB *16*, 460-471.

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

Lapuk, A., Marr, H., Jakkula, L., Pedro, H., Bhattacharya, S., Purdom, E., Hu, Z., Simpson, K., Pachter, L., Durinck, S., Wang, N., Parvin, B., Fontenay, G., Speed, T., Garbe, J., Stampfer, M., Bayandorian, H., Dorton, S., Clark, T.A., Schweitzer, A., Wyrobek, A., Feiler, H., Spellman, P., Conboy, J., and Gray, J.W. (2010). Exon-level microarray analyses identify alternative splicing programs in breast cancer. Molecular cancer research : MCR *8*, 961-974.

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.

Lenasi, T., Peterlin, B.M., and Dovc, P. (2006). Distal regulation of alternative splicing by splicing enhancer in equine beta-casein intron 1. RNA *12*, 498-507.

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

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

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078-2079.

Li, H., Lovci, M.T., Kwon, Y.S., Rosenfeld, M.G., Fu, X.D., and Yeo, G.W. (2008). Determination of tag density required for digital transcriptome analysis: application to an androgen-sensitive prostate cancer model. Proceedings of the National Academy of Sciences of the United States of America *105*, 20179-20184.

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., Darnell, J.C., and Darnell, R.B. (2008). HITS-CLIP yields genome-wide insights into brain alternative RNA processing. Nature *456*, 464-469.

Lim, J., Hao, T., Shaw, C., Patel, A.J., Szabo, G., Rual, J.F., Fisk, C.J., Li, N., Smolyar, A., Hill, D.E., Barabasi, A.L., Vidal, M., and Zoghbi, H.Y. (2006). A protein-protein interaction network for human inherited ataxias and disorders of Purkinje cell degeneration. Cell *125*, 801-814.

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.

Lim, L.P., and Sharp, P.A. (1998). Alternative splicing of the fibronectin EIIIB exon depends on specific TGCATG repeats. Molecular and cellular biology *18*, 3900-3906.

Lin, S.Y., Johnson, S.M., Abraham, M., Vella, M.C., Pasquinelli, A., Gamberi, C., Gottlieb, E., and Slack, F.J. (2003). The C elegans hunchback homolog, hbl-1, controls temporal patterning and is a probable microRNA target. Developmental cell *4*, 639-650.

Long, D., Lee, R., Williams, P., Chan, C.Y., Ambros, V., and Ding, Y. (2007). Potent effect of target structure on microRNA function. Nature structural & molecular biology *14*, 287-294.

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

Martin, C.L., Duvall, J.A., Ilkin, Y., Simon, J.S., Arreaza, M.G., Wilkes, K., Alvarez-Retuerto, A., Whichello, A., Powell, C.M., Rao, K., Cook, E., and Geschwind, D.H. (2007). Cytogenetic and molecular characterization of A2BP1/FOX1 as a candidate gene for autism. American journal of medical genetics Part B, Neuropsychiatric genetics : the official publication of the International Society of Psychiatric Genetics *144B*, 869-876.

Matlin, A.J., Clark, F., and Smith, C.W. (2005). Understanding alternative splicing: towards a cellular code. Nature reviews Molecular cell biology *6*, 386-398.

McManus, C.J., and Graveley, B.R. (2011). RNA structure and the mechanisms of alternative splicing. Current opinion in genetics & development *21*, 373-379.

Minovitsky, S., Gee, S.L., Schokrpur, S., Dubchak, I., and Conboy, J.G. (2005). The splicing regulatory element, UGCAUG, is phylogenetically and spatially conserved in introns that flank tissue-specific alternative exons. Nucleic acids research *33*, 714-724.

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.

Morcos, P.A., Li, Y., and Jiang, S. (2008). Vivo-Morpholinos: a non-peptide transporter delivers Morpholinos into a wide array of mouse tissues. BioTechniques *45*, 613-614, 616, 618 passim.

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

Nasim, F.U., Hutchison, S., Cordeau, M., and Chabot, B. (2002). High-affinity hnRNP A1 binding sites and duplex-forming inverted repeats have similar effects on 5' splice site selection in support of a common looping out and repression mechanism. RNA *8*, 1078-1089.

Parkhomchuk, D., Borodina, T., Amstislavskiy, V., Banaru, M., Hallen, L., Krobitsch, S., Lehrach, H., and Soldatov, A. (2009). Transcriptome analysis by strand-specific sequencing of complementary DNA. Nucleic acids research *37*, e123.

Parra, M.K., Gee, S., Mohandas, N., and Conboy, J.G. (2011). Efficient in vivo manipulation of alternative pre-mRNA splicing events using antisense morpholinos in mice. The Journal of biological chemistry *286*, 6033-6039.

Perocchi, F., Xu, Z., Clauder-Munster, S., and Steinmetz, L.M. (2007). Antisense artifacts in transcriptome microarray experiments are resolved by actinomycin D. Nucleic acids research *35*, e128.

Pervouchine, D.D., Khrameeva, E.E., Pichugina, M.Y., Nikolaienko, O.V., Gelfand, M.S., Rubtsov, P.M., and Mironov, A.A. (2012). Evidence for widespread association of mammalian splicing and conserved long-range RNA structures. RNA *18*, 1-15.

Pistoni, M., Shiue, L., Cline, M.S., Bortolanza, S., Neguembor, M.V., Xynos, A., Ares, M., Jr., and Gabellini, D. (2013). Rbfox1 downregulation and altered calpain 3 splicing by FRG1 in a mouse model of Facioscapulohumeral muscular dystrophy (FSHD). PLoS genetics *9*, e1003186.

Plass, M., Codony-Servat, C., Ferreira, P.G., Vilardell, J., and Eyras, E. (2012). RNA secondary structure mediates alternative 3'ss selection in Saccharomyces cerevisiae. RNA *18*, 1103-1115.

Polymenidou, M., Lagier-Tourenne, C., Hutt, K.R., Huelga, S.C., Moran, J., Liang, T.Y., Ling, S.C., Sun, E., Wancewicz, E., Mazur, C., Kordasiewicz, H., Sedaghat, Y., Donohue, J.P., Shiue, L., Bennett, C.F., Yeo, G.W., and Cleveland, D.W. (2011). Long pre-mRNA depletion and RNA missplicing contribute to neuronal vulnerability from loss of TDP-43. Nature neuroscience *14*, 459-468.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841-842.

Raker, V.A., Mironov, A.A., Gelfand, M.S., and Pervouchine, D.D. (2009). Modulation of alternative splicing by long-range RNA structures in Drosophila. Nucleic acids research *37*, 4533-4544.

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

Robins, H., Li, Y., and Padgett, R.W. (2005). Incorporating structure to predict microRNA targets. Proceedings of the National Academy of Sciences of the United States of America *102*, 4006-4009.

Rogic, S., Montpetit, B., Hoos, H.H., Mackworth, A.K., Ouellette, B.F., and Hieter, P. (2008). Correlation between the secondary structure of pre-mRNA introns and the efficiency of splicing in Saccharomyces cerevisiae. BMC genomics *9*, 355.

Ruby, J.G., Jan, C., Player, C., Axtell, M.J., Lee, W., Nusbaum, C., Ge, H., and Bartel, D.P. (2006). Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in C. elegans. Cell *127*, 1193-1207.

Salemi, M., Barone, C., Romano, C., Ridolfo, F., Scavuzzo, C., Cantarella, R.A., Salluzzo, M.G., Calogero, A.E., Romano, C., and Bosco, P. (2013). KIF21A mRNA expression in patients with Down syndrome. Neurological sciences : official journal of the Italian Neurological Society and of the Italian Society of Clinical Neurophysiology *34*, 569-571.

Sanford, J.R., Wang, X., Mort, M., Vanduyn, N., Cooper, D.N., Mooney, S.D., Edenberg, H.J., and Liu, Y. (2009). Splicing factor SFRS1 recognizes a functionally diverse landscape of RNA transcripts. Genome research *19*, 381-394.

Sato, D., Lionel, A.C., Leblond, C.S., Prasad, A., Pinto, D., Walker, S., O'Connor, I., Russell, C., Drmic, I.E., Hamdan, F.F., Michaud, J.L., Endris, V., Roeth, R., Delorme, R., Huguet, G., Leboyer, M., Rastam, M., Gillberg, C., Lathrop, M., Stavropoulos, D.J., Anagnostou, E., Weksberg, R., Fombonne, E., Zwaigenbaum, L., Fernandez, B.A., Roberts, W., Rappold, G.A., Marshall, C.R., Bourgeron, T., Szatmari, P., and Scherer, S.W. (2012). SHANK1 Deletions in Males with Autism Spectrum Disorder. American journal of human genetics *90*, 879-887.

Sebat, J., Lakshmi, B., Malhotra, D., Troge, J., Lese-Martin, C., Walsh, T., Yamrom, B., Yoon, S., Krasnitz, A., Kendall, J., Leotta, A., Pai, D., Zhang, R., Lee, Y.H., Hicks, J., Spence, S.J., Lee, A.T., Puura, K., Lehtimaki, T., Ledbetter, D., Gregersen, P.K., Bregman, J., Sutcliffe, J.S., Jobanputra, V., Chung, W., Warburton, D., King, M.C., Skuse, D., Geschwind, D.H., Gilliam, T.C., Ye, K., and Wigler, M. (2007). Strong association of de novo copy number mutations with autism. Science *316*, 445-449.

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

Shibata, H., Huynh, D.P., and Pulst, S.M. (2000). A novel protein with RNA-binding motifs interacts with ataxin-2. Human molecular genetics *9*, 1303-1313.

Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S., Weinstock, G.M., Wilson, R.K., Gibbs, R.A., Kent, W.J., Miller, W., and Haussler, D. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome research *15*, 1034-1050.

Simon, D.J., Madison, J.M., Conery, A.L., Thompson-Peer, K.L., Soskis, M., Ruvkun, G.B., Kaplan, J.M., and Kim, J.K. (2008). The microRNA miR-1 regulates a MEF-2-dependent retrograde signal at neuromuscular junctions. Cell *133*, 903-915.

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

Sood, P., Krek, A., Zavolan, M., Macino, G., and Rajewsky, N. (2006). Cell-type-specific signatures of microRNAs on target mRNA expression. Proceedings of the National Academy of Sciences of the United States of America *103*, 2746-2751.

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.

Tollervey, J.R., Curk, T., Rogelj, B., Briese, M., Cereda, M., Kayikci, M., Konig, J., Hortobagyi, T., Nishimura, A.L., Zupunski, V., Patani, R., Chandran, S., Rot, G., Zupan, B., Shaw, C.E., and Ule, J. (2011). Characterizing the RNA targets and position-dependent splicing regulation by TDP-43. Nature neuroscience *14*, 452-458.

Underwood, J.G., Boutz, P.L., Dougherty, J.D., Stoilov, P., and Black, D.L. (2005). Homologues of the Caenorhabditis elegans Fox-1 protein are neuronal splicing regulators in mammals. Molecular and cellular biology *25*, 10005-10016.

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'UTR. Genes & development *18*, 132-137.

Venables, J.P., Brosseau, J.P., Gadea, G., Klinck, R., Prinos, P., Beaulieu, J.F., Lapointe, E., Durand, M., Thibault, P., Tremblay, K., Rousset, F., Tazi, J., Abou Elela, S., and Chabot, B. (2013). RBFOX2 is an important regulator of mesenchymal tissue-specific splicing in both normal and cancer tissues. Molecular and cellular biology *33*, 396-405.

Wang, E.T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S.F., Schroth, G.P., and Burge, C.B. (2008). Alternative isoform regulation in human tissue transcriptomes. Nature *456*, 470-476.

Warf, M.B., Diegel, J.V., von Hippel, P.H., and Berglund, J.A. (2009). The protein factors MBNL1 and U2AF65 bind alternative RNA structures to regulate splicing. Proceedings of the National Academy of Sciences of the United States of America *106*, 9203-9208.

Weese, D., Emde, A.K., Rausch, T., Doring, A., and Reinert, K. (2009). RazerS--fast read mapping with sensitivity control. Genome research *19*, 1646-1654.

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

Wilbert, M.L., Huelga, S.C., Kapeli, K., Stark, T.J., Liang, T.Y., Chen, S.X., Yan, B.Y., Nathanson, J.L., Hutt, K.R., Lovci, M.T., Kazan, H., Vu, A.Q., Massirer, K.B., Morris, Q., Hoon, S., and Yeo, G.W. (2012). LIN28 binds messenger RNAs at GGAGA motifs and regulates splicing factor abundance. Molecular cell *48*, 195-206.

Wu, T.D., and Nacu, S. (2010). Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics *26*, 873-881.

Xie, J., and McCobb, D.P. (1998). Control of alternative splicing of potassium channels by stress hormones. Science *280*, 443-446.

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

Yeo, G.W., Van Nostrand, E.L., and Liang, T.Y. (2007a). Discovery and analysis of evolutionarily conserved intronic splicing regulatory elements. PLoS genetics *3*, e85.

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

Yoo, A.S., and Greenwald, I. (2005). LIN-12/Notch activation leads to microRNA-mediated down-regulation of Vav in C. elegans. Science *310*, 1330-1333.

Zarnack, K., Konig, J., Tajnik, M., Martincorena, I., Eustermann, S., Stevant, I., Reyes, A., Anders, S., Luscombe, N.M., and Ule, J. (2013). Direct competition between hnRNP C and U2AF65 protects the transcriptome from the exonization of Alu elements. Cell *152*, 453-466.

Zhang, L., Ding, L., Cheung, T.H., Dong, M.Q., Chen, J., Sewell, A.K., Liu, X., Yates, J.R., 3rd, and Han, M. (2007). Systematic identification of C. elegans miRISC proteins, miRNAs, and mRNA targets by their interactions with GW182 proteins AIN-1 and AIN-2. Molecular cell *28*, 598-613.

Zhang, L., Hammell, M., Kudlow, B.A., Ambros, V., and Han, M. (2009). Systematic analysis of dynamic miRNA-target interactions during C. elegans development. Development *136*, 3043-3055.

Zhao, Y., Samal, E., and Srivastava, D. (2005). Serum response factor regulates a muscle-specific microRNA that targets Hand2 during cardiogenesis. Nature *436*, 214-220.

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