Published in final edited form as:

Methods. 2012 October ; 58(2): 94–105. doi:10.1016/j.ymeth.2012.08.006.

Genome-wide identification of miRNA targets by PAR-CLIP

Markus Hafner¹, Steve Lianoglou^{2,3}, Thomas Tuschl¹, and Doron Betel⁴

¹ Laboratory of RNA Molecular Biology, Howard Hughes Medical Institute, The Rockefeller University, New York, NY, USA

² Computational Biology Program, Memorial Sloan-Kettering Cancer Center, New York, NY, USA

³ Department of Physiology, Biophysics, and Systems Biology; Weill Cornell Medical College; New York, NY, USA

⁴ Department of Medicine and Institute for Computational Biomedicine; Weill Cornell Medical College; New York, NY, USA

Abstract

MiRNAs are short (20-23 nt) RNAs that are loaded into proteins of the Argonaute (AGO) family and guide them to partially complementary target sites on mRNAs, resulting in mRNA destabilization and/or translational repression. It is estimated that about 60% of the mammalian genes are potentially regulated by miRNAs, and therefore methods for experimental miRNA target determination have become valuable tools for the characterization of posttranscriptional gene regulation. Here we present a step-by-step protocol and guidelines for the computational analysis for the large-scale identification of miRNA target sites in cultured cells by photoactivatable ribonucleoside enhanced crosslinking and immunoprecipitation (PAR-CLIP) of AGO proteins.

Introduction

MicroRNAs (miRNAs), short 20-23 nucleotide (nt) RNAs, guide sequence-specific posttranscriptional gene regulation and are encoded in the genomes of animals, plants, and some viruses (1). They are expressed in specific spatiotemporal patterns and dysregulation or mutation of miRNA genes causes or contributes to several human diseases, mainly due to repression or loss of repression of mRNA targets (2, 3). Identification of miRNA targets is therefore crucial for understanding miRNA function in biology, however experimental assays to do so were limited until crosslinking methods coupled with next generation sequencing became available (4, 5).

miRNA target identification approach typically begins by profiling of miRNA abundance in the relevant tissues or cultured cells, followed by computational prediction of targets for the most abundant or dysregulated miRNAs. A panel of putative targets can be validated by luciferase assays and/or site directed mutagenesis. This general methodology is restricted to a few selected targets chosen by prior biological knowledge, experimental convenience or prediction ranking. It is now known that most miRNAs regulate hundreds of targets and many genes are concurrently regulated by multiple miRNAs. Elucidating this complex network of miRNA-mediated regulation requires large-scale identification of targets (6).

© 2012 Elsevier Inc. All rights reserved.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Approaches to accomplish this include the use of gene expression arrays and proteomics to measure global changes following miRNA transfection (7-9). These provide indirect measure of miRNA regulation and coupled with computational predictions provide a list of target candidates. Thus there is a need for genome-wide experimental approaches for *direct* target identification. The advent of large-scale sequencing technologies and new crosslinking methods for identifying protein-RNA interaction has enabled the systematic, genome-wide unbiased survey of all miRNA targets in the cell by crosslinking and immunoprecipitation experiments (CLIP). These methods are based on crosslinking the miRNA target RNAs to the Argonaute (AGO) proteins, key members of the RNA induced silencing complex (RISC), by irradiation with UV light. AGO-ribonucleoprotein (RNP) complexes can then be recovered by immunoprecipitation (IP), and the precise miRNA target sites of the co-purified target genes are identified by high-throughput sequencing (10-15).

CLIP approaches provide unprecedented genome-wide view of AGO binding sites that includes 3'UTR, coding regions and intergenic regions and illuminate novel forms of AGO-dependent regulation (15, 16). On the other hand these methods present a number of experimental and computational challenges in distinguishing true binding events from background signal. Furthermore, there are substantial differences among the various CLIP protocols, such as methods for controlling for background sequences derived from non-specific binding events, identification of crosslinked positions, or additional implementation of miRNA perturbation experiments to identify functional miRNA binding sites. These differences impact both the sample size and number of biological replicates required for each method, as well as the statistical modeling to identify binding events (4).

Here we describe in detail the use of photoactivatable ribonucleoside-enhanced CLIP (PAR-CLIP) protocol for the detection of endogenous miRNA targets in cell lines, based on the incorporation of 4-thiouridine (4SU) into nascent transcripts, followed by 365 nm UV crosslinking and IP of the AGO-proteins (Figure 1). The key characteristic of the PAR-CLIP method is the introduction of a 'T' to 'C' (T-to-C) mutation at the crosslinked positions during the cDNA library preparation. 4SU crosslinked to an oligopeptide is misrecognized during the reverse transcription step and as a result the altered 4SU base (encoded as 'T' in genomic sequence) is substituted with a 'C' in the resulting cDNA. This fortuitous event is a valuable indicator for the quality control of the experimental outcome and as meaningful signal for defining the AGO binding sites.

We provide a step-by-step protocol for performing the experiment and describe the subsequent computational analysis for assessment of the results and identification of the miRNA-directed AGO binding sites.

PAR-CLIP protocol

Material

Buffers—The following table is organized according to the time of use in the protocol.

Buffer	Composition
Growth medium FLAG/HA-AGO HEK293 cells NOTE: When using different cell lines, use appropriate growth medium. NOTE: Blasticidin is only necessary for doxycycline inducible FLAG/HA-AGO cell lines	DMEM, 10% FBS, 2 mM L-Glutamine, 100 U/ml penicillin, 100 U/ml streptomycin, 100 µg/ml hygromycin, 15 µg/ml blasticidin

Buffer	Composition
4-thiouridine (4SU) stock solution (1 M)	260.27 mg 4SU in 1 ml DMSO
Doxycycline stock solution (10 mg/ml)	10 mg doxycycline in 1 ml DMSO
1× NP40 lysis buffer NOTE: Prepare a stock of 5× buffer without DTT and protease inhibitors. Add DTT and protease inhibitor directly before the experiment.	50 mM HEPES, pH 7.5, 150 mM KCl, 2 mM EDTA, 1 mM NaF, 0.5% (v/v) NP40, 0.5 mM DTT, complete EDTA-free protease inhibitor cocktail (Roche)
Citrate-phosphate buffer	4.7 g/l citric acid, 9.2 g/l Na ₂ HPO ₄ , pH 5.0
IP-wash buffer NOTE: not every antibody will retain its binding ability in 300 mM KCl – adjust the salt concentration accordingly. If in doubt, use lysis buffer instead for washing. NOTE: add DTT and protease inhibitor directly before experiment	50 mM HEPES-KOH, pH 7.5, 300 mM KCl, 0.05% (v/v) NP40, 0.5 mM DTT, complete EDTA-free protease inhibitor cocktail (Roche)
High-salt wash buffer NOTE: not every antibody will retain its binding ability in 500 mM KCl – adjust the salt concentration accordingly. If in doubt, use lysis buffer instead for washing. NOTE: add DTT and protease inhibitor directly before experiment	50 mM HEPES-KOH, pH 7.5, 500 mM KCl, 0.05% (v/v) NP40, 0.5 mM DTT, complete EDTA-free protease inhibitor cocktail (Roche)
Dephosphorylation buffer NOTE: composition is the same as 1× NEB buffer 3	50 mM Tris-HCl, pH 7.9, 100 mM NaCl, 10 mM MgCl ₂ , 1 mM DTT
Phosphatase wash buffer	50 mM Tris-HCl, pH 7.5, 20 mM EGTA, 0.5% (v/v) NP40
Polynucleotide kinase (PNK) buffer without DTT	50 mM Tris-HCl, pH 7.5, 50 mM NaCl, 10 mM MgCl ₂
PNK buffer NOTE: add DTT directly before experiment	50 mM Tris-HCl, pH 7.5, 50 mM NaCl, 10 mM MgCl ₂ , 5 mM DTT
SDS PAGE loading buffer	10% glycerol (v/v), 50 mM Tris-HCl, pH 6.8, 2 mM EDTA, 2% SDS (w/v), 100 mM DTT, 0.1% bromophenol blue
2× Proteinase K buffer	100 mM Tris-HCl, pH 7.5, 150 mM NaCl, 12.5 mM EDTA, 2% (w/v) SDS
Acidic Phenol/Chloroform/IAA	25 ml acidic phenol, 24 ml chloroform, 1 ml isoamylalcohol
10× RNA ligase buffer without ATP	0.5 M Tris-HCl, pH 7.6, 0.1 M MgCl ₂ , 0.1 M 2-mercaptoethanol, 1 mg/ml acetylated BSA (Sigma, B-8894)
10× RNA ligase buffer with ATP	0.5 M Tris-HCl, pH 7.6, 0.1 M MgCl ₂ , 0.1 M 2-mercaptoethanol, 1 mg/ml acetylated BSA (Sigma, B-8894), 2 mM ATP
Formamide gel loading dye	50 mM EDTA, 0.05 % (w/v) bromophenol blue, formamide ad 100%
10× dNTP solution	2 mM dATP, 2 mM dCTP, 2 mM dGTP, 2 mM dTTP
10× PCR buffer	100 mM Tris-HCl pH 8.0, 500 mM KCl, 1% Triton-X-100, 20 mM MgCl ₂ , 10 mM 2-mercaptoethanol

Antibodies, enzymes and oligonucleotides

Antibodies—When using cell lines stably expressing FLAG/HA-tagged AGO1-4, we recommend using an anti-FLAG antibody. For immunoprecipitations of endogenous AGO proteins it is necessary to check for the suitability of the antibody you intend to use (not all antibodies will immunoprecipitate – some are only suitable for immunoblotting). Check at which salt concentrations in the wash buffers the antibody-antigen interactions will be disrupted, e.g. the anti-FLAG M2 antibody will bind its antigen in the presence of 500 mM

NaCl, while the anti-AGO2 (C1.9E8.2, Millipore) antibody should not be used with salt concentrations higher than 300 mM NaCl.

Experiment	Antibody
FLAG/HA-tagged AGO1-4 IP	anti-FLAG M2 (Sigma, F1804), mouse monoclonal 1 mg/ml
Endogenous AGO2 IP	anti-AGO2 C1.9E8.2 (Millipore), mouse monoclonal 1 mg/ml

Enzymes and other material—

Material	Supplier
Dynabeads Protein G	Invitrogen, 100.03D/04D
1.5 ml siliconized tubes	BioPlas 4165SL
5 μ m membrane syringe filter	Pall Acrodisc
RNase T1, concentration 1,000 U/ μ l	Fermentas, EN0541
Calf intestinal phosphatase (CIP)	New England Biolabs (NEB), M0290
T4 Polynucleotide Kinase (T4 PNK)	NEB, M0201
γ - ³² P-ATP, 10 mCi/ml, concentration 1.6 μ M	Perkin Elmer, NEG002Z001MC
NuPAGE Novex 4-12% BT Midi 1.0 gel	Invitrogen
20 $\times$ MOPS running buffer	Invitrogen
Protein size marker, Precision Marker Plus	BioRad, 161-0374
D-Tube Dialyzer Midi, cutoff 3.5 kDa	Novagen, 71506-3
Proteinase K	Roche, 03 115 879 001
Glycoblue or glycogen, 10 mg/ml	Ambion or Roche
Truncated and mutated RNA ligase 2, T4 Rnl2(1-249)K227Q, 1 mg/ml	NEB, M0351, plasmid for recombinant expression can also be obtained at addgene.org
T4 RNA ligase, T4 Rnl1, 1 mg/ml	Fermentas
SuperScript III Reverse Transcriptase	Invitrogen, 18080-044
Taq DNA polymerase, 5 U/ μ l	Various suppliers
Qiaquick gel purification kit	Qiagen

Oligonucleotide Sequences

- Pre-adenylated 3' adapter (DNA): AppTCGTATGCCGTCTTCTGCTTGT
- RNA size markers: 19 nt: 5' pCGUACGCGGGUUAACGA; 35-nt: 5' pCUCAUCUUGGUCGUACGCGGAUAGUUAAACUGU
- 5' adapter (RNA): GUUCAGAGUUCUACAGUCCGACGAUC
- 3' primer: CAAGCAGAAGACGGCATAACGA
- 5' primer: AATGATACGGCGACCACCGACAGGTTTCAGAGTTCTACAGTCCGA

Procedure

The protocol below describes the PAR-CLIP procedure for HEK293 cells stably or inducibly expressing FLAG/HA-tagged AGO proteins.

AGO PAR-CLIP will work with any cell line expressing detectable levels AGO proteins as long as there is an efficient antibody for immunoprecipitation available.

Expanding Cells

1. Expand cells in appropriate growth medium in 15-cm plates. We recommend using a number of cells that will result in 1.5-3 ml of wet pellet (for HEK293 cells approximately $100\text{--}200 \times 10^6$ cells or 10-20 15-cm cell culture plates) as a starting point. Grow cells to approximately 80% confluency.
2. 14 h before crosslinking add a) 4SU to a final concentration of $100 \mu\text{M}$ (1:1000 v/v of a 1 M 4SU stock solution) directly to the cell culture medium. NOTE: it is also possible to use $100 \mu\text{M}$ of 6-thioguanosine (6SG) as the photoactivatable ribonucleoside. 6SG has a lower crosslinking efficiency compared to 4SU and will result in a G to A mutation instead of a T to C mutation at the crosslinking site.
3. For tagged AGO proteins expressed inducibly only: induce expression of the FLAG/HA tagged RBP by addition of $1 \mu\text{g/ml}$ of doxycycline.

UV-Crosslinking

For adherent cells

1. Wash cells once with 10 ml ice-cold PBS per plate and remove PBS completely.
2. Place plates on a tray with ice and irradiate uncovered with a dose of 0.15 J/cm^2 of 365 nm UV light in a Spectrolinker XL-1500 (Spectronics Corporation) equipped with 365 nm light bulbs or similar device.
3. Scrape cells off with a rubber policeman in 1 ml PBS per plate, transfer to 50 ml centrifugation tubes and collect by centrifugation at 500g at 4°C for 5 min and discard the supernatant. 100×10^6 HEK293 cells (10 15-cm plates) will yield approx. 1 ml of wet cell pellet.
4. Unless you want to continue directly with cell lysis, snap freeze the cell pellet in liquid nitrogen and store at -80°C . Cell pellets can be stored for at least 12 months.

For cells grown in suspension

1. Collect cells by centrifugation at 500g at 4°C for 5 min.
2. Wash cells once with 20 ml ice-cold PBS and collect again by centrifugation at 4°C for 5 min.
3. Take up cells in 20 ml ice-cold PBS and transfer onto one 15-cm cell culture plate.
4. Place plates on a tray with ice and irradiate uncovered with a dose of 0.2 J/cm^2 of 365 nm UV light in a Spectrolinker XL-1500 (Spectronics Corporation) equipped with 365 nm light bulbs or similar device.
5. Transfer cells into a 50 ml centrifugation tube and collect by centrifugation at 500g for 5 min at 4°C and discard the supernatant.
6. Unless you want to continue directly with cell lysis, snap-freeze the cell pellet in liquid nitrogen and store at -80°C . Cell pellets can be stored for at least 12 months.

Cell lysis and RNaseT1 Digest

1. Take up cell pellet of crosslinked cells in 3 volumes of $1 \times$ NP40 lysis buffer and incubate on ice for 10 min.

2. Clear cell lysate by centrifugation at 13,000g at 4°C for 15 min. In the meantime, prepare the magnetic beads (see below).
3. Clear the lysate further by filtering it through a 5 µm membrane syringe filter.
4. Add RNase T1 to a final concentration of 1 U/µl and incubate in a water bath at 22°C for 15 min. Cool reaction subsequently for 5 min on ice before proceeding.

Immunoprecipitation and recovery of crosslinked target RNA fragments

General Guidelines for Using Magnetic Beads—Follow these guidelines throughout the sample preparation to prevent the magnetic beads from drying out.

1. Place the tube containing the beads on the magnetic stand for 1–2 minutes.
2. Add the buffer to the tube while the tube is on the magnetic separator.
3. Remove the tube from the magnetic separator, and resuspend the beads either by flicking the tube with your finger or by gently vortexing.
4. Centrifuge briefly to collect any beads that may remain in the tube cap.
5. Repeat steps 1 through 4 as required.

Preparation of Magnetic Beads

1. Transfer 10 µl of Protein G magnetic particles per ml cell lysate (for a typical experiment it should be approx. 60 – 100 µl of beads) to a 1.5 ml microtube. Wash beads twice with 1 ml of citrate-phosphate buffer. (NOTE: For experiments with less than 2.5 ml of cell lysate do not use less than 25 µl of magnetic beads as there will be minor loss of magnetic beads due to handling while performing the protocol).
2. Resuspend in twice the volume of citrate-phosphate buffer relative to the original volume of bead suspension.
3. Add 0.25 mg of antibody per ml bead suspension and incubate on a rotating wheel for 40 min at room temperature.
4. Wash beads twice in 1 ml of citrate-phosphate buffer to remove unbound antibody.
5. Resuspend beads in 1 volume of citrate-phosphate buffer.

Immunoprecipitation (IP), Second RNase T1 Digestion, and Dephosphorylation

1. Add 10 µl of freshly prepared antibody-conjugated magnetic beads per ml of partial RNase T1 treated cell lysate and incubate in 15 ml centrifugation tubes on a rotating wheel for 1 h at 4°C.
2. Collect magnetic beads on a magnetic particle collector for 15 and 50 ml centrifugation tubes and discard supernatant.
3. Add 1 ml of IP wash buffer and transfer to 1.5 ml microfuge tubes (NOTE: make sure that you do not exceed the maximal salt concentration at which the antibody recognizes its antigen. It might be necessary to determine the maximal tolerated salt concentration before starting the PAR-CLIP. If in doubt, use lysis buffer for washing).
4. Wash beads 2 times in 1 ml of IP wash buffer.
5. Take up cells in one bead volume of IP wash buffer.

6. Add RNaseT1 to a final concentration of 100 U/ μ l and incubate the bead suspension in a water bath for 15 min at 22 °C. Cool subsequently on ice for 5 min.
7. Wash beads 3 times in 1 ml of high salt wash buffer. (NOTE: make sure that you do not exceed the maximal salt concentration at which the antibody recognizes its antigen. It might be necessary to determine the maximal tolerated salt concentration before starting the PAR-CLIP. If in doubt, use lysis buffer for washing).
8. Resuspend beads in 1 volume of dephosphorylation buffer.
9. Add calf intestinal alkaline phosphatase to a final concentration of 0.5 U/ μ l, and incubate the suspension for 10 min at 37°C.
10. Wash beads twice in 1 ml of phosphatase wash buffer.
11. Wash beads twice in polynucleotide kinase (PNK) buffer without DTT (the DTT concentration necessary for the enzymatic reaction is high enough to damage the magnetic beads).
12. Resuspend beads in one original bead volume of PNK buffer.

Radiolabeling of RNA Segments Crosslinked to Immunoprecipitated Proteins

1. To the bead suspension described above, add γ -32P-ATP to a final concentration of 0.5 μ Ci/ μ l (1.6 μ M ATP) and T4 PNK to 1 U/ μ l in one original bead volume. Incubate the suspension for 30 min at 37°C.
2. Add non-radioactive ATP to obtain a final concentration of 100 μ M and incubate for another 5 min at 37°C.
3. Wash the magnetic beads 5 times with 800 μ l of PNK buffer without DTT.
4. Resuspend the beads in 70 μ l of SDS-PAGE loading buffer.
5. Incubate the radiolabeled suspension for 5 min in a heat block at 95°C to denature and release the immunoprecipitated AGO proteins with crosslinked RNA and vortex.
6. Remove the magnetic beads on the separator and transfer the supernatant to a clean 1.5 ml microtube.

SDS-PAGE and Electroelution of Crosslinked RNA-protein Complexes from Gel Slices

1. Load 40 μ l of the supernatant per well of a precast Novex Bis-Tris 4-12% polyacrylamide gel and run the gel for 55 min at 200 V. Keep 5 μ l of supernatant for a Western blot to confirm that the immunoprecipitation was efficient.
2. Disassemble the gel chamber and dismantle the gel, leaving it mounted on one plate. To facilitate the alignment of the gel to the phosphorimager paper printout, we recommend implanting three tiny radioactive gel pieces asymmetrically at three of the four corners of the gel. Radioactive gel pieces can be collected from gels that were previously used to purify radiolabeled synthetic oligonucleotides. Wrap the gel in plastic film (e.g. Saran wrap) to avoid contamination.
3. Expose the gel to a blanked phosphorimager screen for 1 h and visualize on a phosphorimager.
4. Align the gel on top of the phosphorimager printout using the implanted gel pieces for orientation. Cut out the bands that correspond to the expected size of RBP

(AGO2, approx. 100 kDa). (Pause point: you can freeze the excised bands and continue with the protocol at another time).

5. Add 800 μ l of H₂O to a D-Tube Dialyzer Midi Tube and let stand at room temperature for 5 min. Remove the water.
6. Transfer excised band to the dialyzer tube and add 800 μ l 1 $\times$ SDS running buffer.
7. Electroelute the crosslinked RNA-RBP complex in 1 $\times$ SDS running buffer at 100 V for 2 h.

Proteinase K Digestion

1. Add an equal volume of 2 $\times$ Proteinase K Buffer with respect to the electroeluate, followed by the addition of Proteinase K to a final concentration of 1.2 mg/ml. Incubate for 30 min at 55°C.
2. Recover the RNA by acidic phenol/chloroform/IAA extraction (25:24:1, pH 4.0) followed by a chloroform extraction. Add 1 μ l of glycogen (10 mg/ml stock) precipitate the RNA by adding 3 volumes of ethanol. Dissolve the pellet in 10 μ l water.

cDNA library preparation and deep sequencing

Carry the recovered RNA through a standard cDNA library preparation protocol originally described for cloning of small regulatory RNAs. Before generating the small RNA cDNA libraries following the steps described below, we strongly recommend reading the more detailed step-by-step procedure described in (Hafner et al. same METHODS issue). The main differences compared to that procedure are: (1) the use of a non-barcoded 3' adapter, (2) no spike-in of radioactive RNA size markers, and (3) no use of spike-in calibrator oligoribonucleotides.

Use siliconized tubes for all manipulations of the small RNAs. The minute amounts of small RNAs to be recovered after gel purification will readily adsorb to the walls of standard tubes.

3'adapter Ligation

1. Prepare the following reaction mixture for ligation of the 3'adenylated adapter, multiplying the volumes by number of ligation reactions to be performed plus one extra volume to account for pipetting error: 2 μ l of 10 $\times$ RNA ligase buffer without ATP, 6 μ l 50% DMSO and 1 μ l of 100 μ M adenylated 3' adapter. Add 9 μ l of the reaction mixture to each sample. Prepare one sample of 40 fmol of 1:100 dilution of 5'-³²P-labeled RNA size markers (for a standard PAR-CLIP experiment we use a mixture of the 19 nt and the 35 nt size marker). This controls for successful ligation and indicates the length of the bands that need to be cut out of the gel.
2. Denature the RNA by incubating the tubes for 1 min at 90°C. Place the tube immediately on ice for 2 min.
3. Add 1 μ l of Rnl2(1-249)K227Q ligase (1 μ g/ μ l), mix gently and incubate overnight on ice in the cold room.
4. Add 20 μ l of formamide gel loading solution.
5. Denature sample for 1 min at 90°C.
6. Load samples in two wells on a 15% denaturing acrylamide gel. Make sure to space different samples appropriately, typically at a two-well distance, to avoid cross-

contamination. Load the marker reactions on opposite ends of the gel, framing the PAR-CLIP samples, keeping a two-well distance from them to avoid cross-contamination.

7. Run the gel for 45 min at 30 W using 1× TBE buffer until the bromophenol blue dye is close to the bottom of the gel.
8. Dismantle the gel, leaving it mounted on one glass plate. Implant three tiny radioactive gel pieces asymmetrically at three of the four corners of the gel, to facilitate the alignment of the gel to the phosphorimager paper printout. Wrap the gel in plastic film (e.g. Saran wrap) to avoid contamination.
9. Expose the gel for at least 1 h to a phosphorimager screen. If the radioactivity of the recovered RNA is weak, you can expose the gel overnight, placing the cassette in a -20°C freezer.
10. Align the gel on top of a printout scaled to 100% according to the position of the three radioactive gel pieces. Cut out the 19 to 35 nt bands defined by the ligated products of the 19 nt and above the 35 nt marker (NOTE: Do not cut below the line delimited by the ligated 19 nt markers, as all sequenced reads that are shorter than 20 nt are not used in the bioinformatic analysis, as they have a higher probability of mismapping compared to longer reads).
11. Elute the ligation product from the gel slices overnight at 4°C with constant agitation with at least 3 to 4 volumes of 0.3 M NaCl.
12. Take off the supernatant, add 1 µl of glycogen solution, ethanol-precipitate and collect the small RNA as described above.

5' adapter Ligation

1. Dissolve the pellet in 9 µl water.
2. Prepare the following reaction mixture by combining 1 µl of 100 µM 5' adapter oligonucleotide, 2 µl of 10× RNA ligase buffer with ATP and 6 µl 50% aqueous DMSO. Add the 9 µl of this mixture to the sample.
3. Denature the RNA by incubation for 1 min at 90°C. Place the tube immediately on ice for 2 min.
4. Add 2 µl of Rnl1, mix gently, and incubate for 1 h at 37°C.
5. Add 20 µl of formamide gel loading solution and incubate the samples for 2 min at 95°C and load them in two adjacent wells of a 20-well 12% acrylamide gel. Make sure to space different samples appropriately, typically at a two-well distance, to avoid cross-contamination. Load the marker reactions on opposite ends of the gel, framing the PAR-CLIP samples, keeping a two-well distance from them to avoid cross-contamination.
6. Run the gel for 45 min at 30 W using 1× TBE buffer until the bromophenol blue dye is close to the bottom of the gel. Image the gel as described above and excise the new ligation product.
7. Elute the ligation product from the gel slices with 3 to 4 volumes of 0.3 M NaCl by constant shaking overnight at 4°C. Add 1 µl of 100 µM 3' primer during the elution as carrier to facilitate the recovery of the ligation product.
8. Ethanol-precipitate and collect the RNA.

Reverse Transcription

1. Dissolve the pellet in 5.6 μ l water. Prepare the following reaction mix: 1.5 μ l 0.1 M DTT, 3 μ l 5 $\times$ first-strand synthesis buffer, 4.2 μ l 10 $\times$ dNTPs.
2. Denature the RNA by incubating the tube for 30 s at 90°C and transfer the tube to a 50°C incubator.
3. Add 8.7 μ l of the prepared reaction mix to each sample and incubate for 3 min at 50°C. Add 0.75 μ l of Superscript III reverse transcriptase and incubate for 2 h at 50°C.
4. Add 85 μ l H₂O and mix well.

PCR Amplification

1. Prepare the following mix multiplied with number of samples: 40 μ l of the 10 $\times$ PCR buffer, 40 μ l 10 $\times$ dNTPs, 2 μ l of 100 μ M 5' primer, 2 μ l of 100 μ M 3' primer, 272 μ l H₂O. Perform a standard 100 μ l PCR with Taq polymerase (5 U/ μ l). 89 μ l of the reaction mix will be used in a pilot PCR reaction and rest of the mixture will be used for the large scale PCR.
2. To 89 μ l of the reaction mix add 10 μ l from the cDNA solution and 1 μ l of the Taq polymerase. Use the following cycle conditions: 45 s at 94°C, 85 s at 50°C, 60 s at 72°C.
3. To determine the necessary number of cycles for amplifying of the cDNA library, remove 10 μ l aliquots every third cycle starting with cycle number 12 up to cycle number 30. To remove aliquots from the PCR tube, temporarily pause the PCR cyclers at the end of the 72°C step.
4. Analyze the samples on a 2.5% agarose gel. The PCR-product might appear as a double band with the higher band running at the expected length of about 95-110 nt and a lower band at 65 nt corresponding to the direct ligation products of the 3' adapter and 5' adapter. Define the optimal cycle number for cDNA amplification, which should be within the exponential amplification phase of the PCR, approx. 5 cycles away from reaching the saturation level of PCR amplification.
5. Perform a 300 μ l PCR with the optimal cycle number and analyze the product again on a 2.5% agarose gel.
6. Concentrate the DNA by ethanol precipitation and take it up in 60 μ l of 1 $\times$ DNA loading dye.
7. Run the sample on two wells of 2.5% agarose gel stained with 0.4 μ g/ml ethidiumbromide to remove any amplified 5'adapter-3'adapter products that will also get sequenced and return insertless sequence reads. Take care not to overload the gel, as this will compromise the separating capacity of the agarose gel.
8. Visualize the DNA on a UV transilluminator and excise the band corresponding to 85-110 nt with a clean scalpel.
9. Purify the DNA using the Qiaquick gel extraction kit according to the instructions of the manufacturer. Use 30 μ l elution buffer to recover the cDNA.
10. Submit 10 μ l of the purified cDNA to Illumina sequencing. We recommend using 50 base sequencing either on a GAIIx or HiSeq machine. Our adapters are compatible with the Illumina small RNA sequencing kit.

Optional: Determination of Incorporation Levels of 4-thiouridine into Total RNA

Isolate total RNA from the cell line stably expressing the RBP of interest after growing in medium supplemented with 100 μ M 4SU 16 h prior to harvest. As a control, harvest cells grown without 4SU addition. Isolate total RNA by addition of 3 volumes of TRIzol reagent (Sigma) to the washed cell pellets following the manufacturer's instructions. Further purify total RNA using Qiagen RNeasy according to the manufacturer's instructions. It is important to add 0.1 mM dithiothreitol (DTT) to the wash buffers and subsequent enzymatic steps to prevent oxidation of 4SU during RNA isolation and analysis. Digest and dephosphorylate total RNA to single nucleosides for HPLC analysis as described before (17). Incubate 40 μ g of purified total RNA for 16 h at 37°C with 0.4 U bacterial alkaline phosphatase (e.g. Worthington Biochemical) and 0.09 U snake venom phosphodiesterase (e.g. Worthington Biochemical) in a 30 μ l volume. As a reference standard, use a synthetic 4SU-labeled RNA, (we standardly use CGUACGCGGAAUACUUCGA(4SU)U) and also subject it to complete enzymatic digestion. Separate the resulting mixtures of ribonucleosides by HPLC on a Supelco Discovery C18 (bonded phase silica 5 μ M particle, 250 $\times$ 4.6 mm) reverse phase column (Bellefont). HPLC buffers are 0.1 M triethylammonium acetate (TEAA) in 3% acetonitrile (A) and 90% acetonitrile in 0.1 M TEAA (B). Use an isocratic gradient: 0% B for 15 min, 0 to 10 % B for 20 min, 10 to 100% B for 30 min. Clean the HPLC column with a 5 min 100 % B wash between runs.

Computational Analysis

The objective of the computational analysis is to identify those AGO-binding sites that also show a miRNA signature and provide quality control measures for the experimental results. CLIP-based approaches provide a genome-wide view of the protein-RNA interaction sites however, in order to relate RNA binding to regulation of the targeted mRNAs additional analysis and experimentation is required. While sequence coverage of PAR-CLIP experiments generally correlates well with the mRNA expression levels (18) it does not necessarily correlate with the affinity of the studied protein to the target RNA and therefore read frequencies for binding sites from AGO PAR-CLIP experiments alone may not be predictive of the extent of miRNA-mediated downregulation of the target gene (19).

Identification of functional sites requires additional complementary miRNA perturbation experiments (either inhibition or over expression of the miRNA(s) (10, 13)) to identify changes in targeted mRNA abundance. These results are then correlated with the PAR-CLIP data to identify those AGO-bound targeted genes that had significant gene expression changes upon perturbation of the regulating miRNA.

In this section we describe the specific steps in the computational analysis of PAR-CLIP data from processing the raw sequence data to identification of potential miRNA regulators. The analysis consists of the following steps: i) preprocessing the sequenced reads, ii) aligning the data to the target genome and reference transcripts, iii) identifying AGO binding locations and iv) assigning the most likely miRNAs regulators to the AGO binding sites. We demonstrate these steps through the analysis of three PAR-CLIP experiments performed on HEK293 and human embryonic stem cells (hESCs). These experiments include: two replicate PAR-CLIP experiments from HEK293 using AGO2 antibody 11A9 (G.Meister) (*Kishore et al.*; SRA accessions SRR189784, SRR189785); one dataset from hESCs using native AGO2 antibody C1.9E8.2 (Milipore) (*Lipchina et al.*; SRA accession SRX103431); and two replicates anti-FLAG antibody M2 (Sigma) from HEK293 cells expressing FLAG/HA-tagged AGO2 (*Hafner et al.*; SRA accession SRX020784)(10, 13, 14). To date there is no single software package that performs the full analysis from sequence reads to identification of miRNA targets. Significant progress was made by the

introduction of the PARalyzer software which identifies AGO binding sites and performs motif enrichment analysis (20). Nevertheless, much of the analysis requires computational expertise and familiarity with software package for analysis of alignment data (samtools (21), bedtools (22), picard (23) , various Bioconductor packages (24) as well as libraries from other programming languages).

Preprocessing of Sequenced Reads

PAR-CLIP sequenced libraries should not be reduced (collapsed) to non-redundant reads to take advantage of read frequencies for scoring crosslinking efficiency. Our protocol yields complex small RNA cDNA libraries not subject to sequence-specific clonal amplification (25). Removing duplicate reads adversely affect the analysis by reducing the number of reads with T-to-C mutations and thereby reducing this informative signal.

When generating the cDNA library from the recovered RNA, we use the RNA fraction of 19 to 35 nucleotides in length and therefore the size of the sequenced inserts is shorter than the length of sequencing, which is commonly 50 bases for a standard Illumina sequencing experiment. As a result most of the sequenced reads include sequence stretches originating from the 3' adaptor at the 3' end of the reads, which need to be removed prior to alignment. There are multitude of programs with various degrees of functionalities that can be used to remove adaptor sequences including the fastx package, far [<http://sourceforge.net/apps/mediawiki/theflexibleadap/>], Trimmomatic [<http://www.usadellab.org/cms/index.php?page=trimmomatic>], cutadapt [<http://code.google.com/p/cutadapt/>], or many other *ad hoc* scripts.

For the purpose of the example analysis performed here all sequences from each library were used, Illumina adaptors were trimmed using the FASTX tools (fastx_clipper -a TCGTATGCCGCTTCTGCTTG).

Mapping of Sequenced Reads

Mapping sequenced reads to the target genome and reference transcripts can be performed similarly to other RNA-seq data using tools such as TopHat (26) (or bowtie (27)), bwa(28), SOAP3(29) or more customized tools (30). The choice of alignment parameters such as minimum read lengths, number of allowed mismatches and number of reported alignments for each reads can affect downstream analysis, specifically the number of reported AGO binding sites. Lenient setting of the mapping algorithm will increase sensitivity at the expense of specificity of binding sites detection and may require the refinement of binding detection by tuning the cluster detection parameters (see section *Identification of AGO Binding Sites*) and by restricting to specific genomic regions such as 3'UTRs. The recommended number of allowed mismatches is 2 and the minimum read lengths for alignment is 20 nucleotides. For the purpose of maximum sensitivity here we used minimum 10 nucleotides read length. Typical mapping yield of PAR-CLIP experiments range between 20-50% but can be lower (Table 1). Mapping algorithms do not differentiate between sequence mismatches that are intrinsic to the experimental protocol (such as PAR-CLIP or bisulfite sequencing) from mismatches due to sequencing errors which occur at low frequency and are randomly distributed. Future algorithmic development may take into consideration the experimental evidence (such as introduced mismatches) to improve mapping yields.

Assessing T-to-C Mutation Frequency

Before continuing with additional analysis the mapped sequenced reads should meet certain quality control benchmarks. In all CLIP experiments, fragments derived from non-crosslinked, abundant cellular RNAs, such as rRNAs and tRNAs are co-purified and

sequenced. Because they do not have an oligopeptide covalently attached, which causes the reverse transcriptase enzyme to stall particularly if the proteinase K digestion was incomplete, these unmodified sequences have a distinct advantage in the reverse transcription step, and may represent the majority of sequence reads. A key indication for a successful PAR-CLIP experiment is the significant increase of T-to-C substitution over all other base substitutions in the sequenced reads as expected from the crosslinking step. For this analysis we focused on regions of high coverage (minimum 5 read coverage in a window of 10 nucleotides) and measured the number of base substitutions between the reads and human genome (hg19). The choice of coverage and window size was somewhat arbitrary and was intended to focus the analysis on regions of high coverage. The abundance of non-coding RNAs as well as contamination of the IP from non-specific binding of AGO proteins may obscure the T-to-C mutation signal. In such cases excluding known non-coding RNAs (rRNA, tRNA, snoRNA, etc.; see Farazi et al. in this issue for details about obtaining annotated lists of small RNAs) or restricting the mutation analysis to mRNAs may direct the analysis to regions of biological interest and facilitates the detection of the T-to-C mutation signal.

There are a number of command-line tools that calculate sequence coverage such as samtools (21), GATK (31), and BEDtools(32) as well as a number of useful R packages that can be used for this purpose such as GenomicRanges and Rsamtools (24). Detection of mismatches between the genomic sequence and the sequenced reads requires parsing of the alignment information from the alignment files. There are a number of different file formats for storing alignments and identifying the mutations in aligned reads is entirely dependent on the file format produced by the aligners. For example, bowtie (27) output format includes a substitution string that describes the mismatches in the format of 7:G>T, 28:T>C indicating a change from G to T in position 7 and T to C in position 28 of the read. The SAM file format (and its corresponding binary BAM) is among the most commonly used alignment representation for high-throughput sequencing (21). In this format alignment details are encoded in the 'CIGAR' field and mismatch information is presented in the 'MD' field. Identification of the positions and identities of the mismatches requires reconstructing the alignment string by writing a specific program that can be in any number of the commonly used programming languages such as Perl, Python, Ruby, Java, C/C++, R and others using the appropriate SAM file APIs (21, 23, 33).

Figure 2 shows the mutation frequency of all possible 12 base substitutions for the three experiments. In all cases, the frequency 'T2C' and 'A2G' (the reverse complement) mutations are substantially over represented in the three datasets analyzed. It is important to note that crosslinking events are restricted to the modified 4SU bases (encoded as 'T' in genomic sequence) and therefore chemically only T-to-C mutations are introduced by the PAR-CLIP protocol. The A-to-G mutations are the result of the computational reverse complement conversion during the analysis of reverse strand genes. The unequal proportions of A2G and T2C mutations can be attributed to differences in the number of transcripts from each strand, unequal number of AGO binding events between the two stands, and background mutations that are not related to crosslinking. This step evaluates the quality and yield of the crosslinking procedure. The identification of high-coverage regions should be independent of mutation events in contrast to the identification of AGO binding sites (see next section) where the T-to-C mutations provide an important signal for the detection of binding sites.

Not all reads with T-to-C mutations indicate sequences crosslinked to AGO proteins as non-specific crosslinking events may obscure the results. This is evident in Hafner et al. dataset (Figure 2) where a large number of T-to-C mutations can be attributed to mapping of

fragments of abundant and noncrosslinked, but co-purified rRNAs to genes containing polymorphic variants.

Additional useful quality control measure is the enrichment for ‘G’ bases 5’ (upstream) to the sequenced read. This is a result of digestion with RNase T1, which cleaves after guanine position except when protected by RNA-binding protein. Sequence reads matching to multiple genomic locations may be assigned based on the presence of a ‘G’ upstream of the read.

Identification of AGO-Binding Sites

Immunoprecipitation experiments of DNA and RNA interacting proteins such as PAR-CLIP, ChIP-seq and methyl-seq are characterized by distinct sequencing profiles in which the DNA or RNA binding sites are identified as genomic regions with clusters of mapped reads. The hallmark T-to-C base substitution of the PAR-CLIP method provides an additional signal that can be used to filter signal from noise and help identify high-confidence binding sites. Hafner et al. and Lipchina et al. used a simple density measure to identify binding sites and required overlapping sequence reads with minimal coverage of 5 reads of which 20% needed to have a T-to-C change. To identify the miRNA binding site, a window of 41 (13) or 61 (10) nucleotides centered around the position of most frequent crosslinking was extracted and further analyzed. More recently a new method termed ‘PARalyzer’ was developed that is based on a kernel density classifier (20). Here interaction sites are identified as nucleotide positions where the density of T-to-C mutations over a predefined window size is higher than the density of non-mutated ‘T’s. The mutation density is defined as the sum of T-to-C mutation frequencies (or none-mutation events) within a fixed size window scaled by a Gaussian kernel function. Clusters are defined by extending from the high T-to-C mutated positions until coverage is below five reads. This method has the advantage of incorporating all crosslinked sites within a specified window and defining the clusters according to the read densities in contrast to predefined window size (Figure 3). A recently developed method termed wavCluster (available as an R package) uses a mixture model to identify experimentally derived T-to-C mutations from background events and peaks are identified by wavelet transformation of the read densities (34). Other approaches used in ChIP-seq analysis may also be adopted for cluster definition from PAR-CLIP experiments such as smoothing window (35) provided that the T-to-C mutation signals are used to identify in the putative binding sites.

Quality Control of AGO-binding Sites

The primary mode of miRNA-mediated regulation is through partial complementarity between the miRNA and the 3’UTR of the targeted mRNA. It may be expected therefore that the genomic distribution of PAR-CLIP clusters is strongly biased towards 3’UTR regions. To investigate this we annotated the AGO-binding sites identified by PARalyzer with their respective genomic categories based on the RefSeq gene annotation for hg19. Multi-isoform genes were reduced to a “unified gene model” by taking the union of all annotated exons from all isoforms of the gene. Using these models, the PAR-CLIP clusters were then annotated with the specific genomic region (CDS, intron, 5’UTR, 3’UTR) that the clusters overlap. Clusters that map overlapping genetic regions (e.g. the terminal CDS and 3’UTR) are assigned to the region of maximum overlap. In all three studies 3’UTR and non-coding regions have the highest densities of PAR-CLIP clusters measured as the number of clusters per Mb of genomic region (Figure 4). The strong enrichment in non-coding regions is attributed predominantly to miRNA transcripts that are also captured by the PAR-CLIP procedure (Figure 3b). Intergenic PAR-CLIP clusters often account for more than half of all PAR-CLIP sites even though their densities in those regions are lower than 3’UTR or even CDS.

Identifying miRNA Binding Sites in PAR-CLIP Clusters

Approximately half of the binding sites identified in the data sets fall into the coding regions. However, miRNA depletion and overexpression showed that these sites overall have a negligible effect on mRNA destabilization (13, 36, 37). Therefore we used the AGO binding sites detected by enrichment of T-to-C mutations in 3'UTRs to define the relevant genomic regions for searching miRNA target sites. Hafner et al. and Kishore et al. used *de novo* motif discovery tools such as cERMIT (38), MEME(39) and PhyloGibbs (40) to identify the enriched sequence motifs in the 3'UTR clusters and showed that the top-ranked motifs are complementary to the 5'-end of the highly expressed miRNAs. This analysis also revealed that the position of the most frequent T-to-C sites are typically 1-2 nucleotides upstream to the miRNA seed sequence complementary region.

De novo motif discovery is an unbiased approach to identify the regulating miRNAs and to validate that the identified AGO-binding sites are primarily directed by the endogenously expressed miRNAs. On the other hand it can require extensive computational resource and is often sensitive to the initiating parameters. Alternatively, one can perform a directed search for miRNA targets sites by a simple motif search for the complementary 7mer or 8mer seed sequence, which is typically sufficient to identify the most likely target sites, or by more advanced prediction methods that are more accurate and can identify non-canonical sites (41, 42).

If predictions are performed directly on the genomic sequence, and not on transcript sequences downloaded from NCBI or UCSC Genome Browser, then knowledge of the strand orientation of the mRNA transcript is required for correct target prediction. The library preparation protocol described in the experimental section, as well as all other cited small RNA library preparation procedures, preserve the strand orientation of the transcripts. Therefore, the strand orientation can be deduced from the strand assignment of the mapped sequence reads. In most cases a simple census of the mapped reads in each cluster will clearly indicate the strand orientation. Other cases where PAR-CLIP clusters map to overlapping bi-directional transcripts may require additional processing to identify the correct target site.

Most miRNAs can potentially target hundreds of different target mRNAs (43) and it is therefore expected that miRNA targeting is dependent on their abundance. In mammalian cells and tissues the top 100 expressed miRNAs typically account for more than 90% of all the miRNA content (13, 25), hence target prediction for the 100 most abundant miRNAs is likely sufficient to explain most cellular miRNA regulation. Using a simple string matching approach (available in Biostrings R package) we searched the PAR-CLIP clusters in 3'UTRs for sequence complementary to the seed sequences (positions 2-8) of the top 100 miRNA expressed in HEK293 or hESCs. The results indicate a general correlation between miRNA expression levels and the number of identified targets sites in the PAR-CLIP clusters (Figure 5) confirming that the identified PAR-CLIP clusters in 3'UTRs are endogenous miRNAs target sites.

A search for miRNA seed sequences in the AGO binding sites in the Hafner *et al.* study (13) showed that approximately 50% of the identified PAR-CLIP sites contain at least one 6-mer seed complementary sequence and can be found within the 3 nt upstream of the crosslinked site. However, depletion of the top 25 miRNAs expressed miRNAs in HEK293 and measurement of the effect of miRNA binding on mRNA stability showed that not all of the identified AGO binding contributed to mRNA destabilization. Thus, to identify miRNA-regulated transcripts complementary miRNA perturbation experiments coupled with gene expression measurement, such as by arrays or RNA-seq, are required to uniquely identify

the regulating miRNAs and to establish the extent of miRNA-mediated regulation of the target gene.

Conclusions

The increased interest in miRNA biology and implication in various diseases has propelled the development of genome-wide high-throughput methods for identifying miRNA targets such as PAR-CLIP (4), iCLIP (44), and HITS-CLIP (45). Here we described the experimental steps and outlined the computational analysis for the PAR-CLIP method which has been used to identify miRNA targets in HEK293 cells, Epstein-Barr virus miRNA targets (11), hESC miRNAs targets (10) as well as the RNA target sites of a growing number of RNA-binding proteins (13, 18, 46-49).

The genome-wide unbiased nature of the CLIP-based experiments has revealed a comprehensive view of AGO-binding sites that extend well beyond the 3'UTR regions and includes protein coding and intergenic regions. Not all identified AGO binding sites are explained by conventional seed matching pointing either to non-canonical target sites (12, 41), co-purification of other RNA binding proteins, or potentially binding of AGO to these sites in the absence of miRNAs (15).

The reproducibility of the results is an important validation of the experimental results and methodology. Both Kishore *et al.* and Hafner *et al.* (13, 14) reported high reproducibility among binding sites detected by PAR-CLIP particularly among the top ranked clusters (by read frequency). In addition, a number of studies also reported high overlap of miRNAs loaded into AGO1-4 isoforms (13, 36, 50) suggesting the use of AGO2 pull down is sufficient to generalize over all AGO targets. Nevertheless, with the increased capacity of sequencing and multiplexing of samples the cost of generating biological replications is becoming affordable and therefore additional replicate experiments will increase confidence in the experimental interpretation.

The methodologies introduced here will likely be modified at both the experimental side and the subsequent computational analysis. Accumulation of additional data will drive further developments of the computational analysis such as improved mapping algorithms, more refined cluster definition, integration of sequencing data with target prediction, and integration of miRNA perturbation experiments to correlate the CLIP evidence with the extent of miRNA-mediated regulation.

Acknowledgments

M.H. was funded by a fellowship of the Charles Revson, Jr. Foundation. T.T. is an HHMI investigator, and this work in his laboratory was supported by NIH grant 1RC1CA145442. T.T. and D.B. were supported by Starr Foundation grants.

References

1. Berezikov E. Nat Rev Genet. 2011; 12:846–60. [PubMed: 22094948]
2. Lu M, Zhang Q, Deng M, Miao J, Guo Y, Gao W, Cui Q. PLoS One. 2008; 3:e3420. [PubMed: 18923704]
3. de Pontual L, Yao E, Callier P, Faivre L, Drouin V, Cariou S, Van Haeringen A, Genevieve D, Goldenberg A, Oufadem M, Manouvrier S, Munnich A, Vidigal JA, Vekemans M, Lyonnet S, Henrion-Caude A, Ventura A, Amiel J. Nat Genet. 2011; 43:1026–30. [PubMed: 21892160]
4. Ascano M, Hafner M, Cekan P, Gerstberger S, Tuschl T. Wiley Interdiscip Rev RNA. 2012; 3:159–77. [PubMed: 22213601]

5. Licatalosi DD, Mele A, Fak JJ, Ule J, Kayikci M, Chi SW, Clark TA, Schweitzer AC, Blume JE, Wang X, Darnell JC, Darnell RB. *Nature*. 2008; 456:464–9. [PubMed: 18978773]
6. Arvey A, Larsson E, Sander C, Leslie CS, Marks DS. *Mol Syst Biol*. 2010; 6:363. [PubMed: 20404830]
7. Lim LP, Lau NC, Garrett-Engele P, Grimson A, Schelter JM, Castle J, Bartel DP, Linsley PS, Johnson JM. *Nature*. 2005; 433:769–73. [PubMed: 15685193]
8. Selbach M, Schwanhauser B, Thierfelder N, Fang Z, Khanin R, Rajewsky N. *Nature*. 2008; 455:58–63. [PubMed: 18668040]
9. Baek D, Villen J, Shin C, Camargo FD, Gygi SP, Bartel DP. *Nature*. 2008; 455:64–71. [PubMed: 18668037]
10. Lipchina I, Elkabetz Y, Hafner M, Sheridan R, Mihailovic A, Tuschl T, Sander C, Studer L, Betel D. *Genes Dev*. 2011; 25:2173–86. [PubMed: 22012620]
11. Skalsky RL, Corcoran DL, Gottwein E, Frank CL, Kang D, Hafner M, Nusbaum JD, Feederle R, Delecluse HJ, Luftig MA, Tuschl T, Ohler U, Cullen BR. *PLoS Pathog*. 2012; 8:e1002484. [PubMed: 22291592]
12. Chi SW, Hannon GJ, Darnell RB. *Nat Struct Mol Biol*. 2012; 19:321–7. [PubMed: 22343717]
13. Hafner M, Landthaler M, Burger L, Khorshid M, Hausser J, Berninger P, Rothballer A, Ascano M Jr, Jungkamp AC, Munschauer M, Ulrich A, Wardle GS, Dewell S, Zavolan M, Tuschl T. *Cell*. 2010; 141:129–41. [PubMed: 20371350]
14. Kishore S, Jaskiewicz L, Burger L, Hausser J, Khorshid M, Zavolan M. *Nat Methods*. 2011; 8:559–64. [PubMed: 21572407]
15. Leung AK, Young AG, Bhutkar A, Zheng GX, Bosson AD, Nielsen CB, Sharp PA. *Nat Struct Mol Biol*. 2011; 18:237–44. [PubMed: 21258322]
16. Zhang C, Darnell RB. *Nat Biotechnol*. 2011; 29:607–14. [PubMed: 21633356]
17. Andrus A, Kuimelis RG. *Curr Protoc Nucleic Acid Chem*. 2001 Chapter 10, Unit 10 6.
18. Hoell JI, Larsson E, Runge S, Nusbaum JD, Duggimpudi S, Farazi TA, Hafner M, Borkhardt A, Sander C, Tuschl T. *Nat Struct Mol Biol*. 2011; 18:1428–31. [PubMed: 22081015]
19. Wen J, Parker BJ, Jacobsen A, Krogh A. *RNA*. 2011; 17:820–34. [PubMed: 21389147]
20. Corcoran DL, Georgiev S, Mukherjee N, Gottwein E, Skalsky RL, Keene JD, Ohler U. *Genome Biol*. 2011; 12:R79. [PubMed: 21851591]
21. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. *Bioinformatics*. 2009; 25:2078–9. [PubMed: 19505943]
22. Quinlan AR, Hall IM. *Bioinformatics*. 2010; 26:841–2. [PubMed: 20110278]
23. Wysoker A, T. K. Fennell T. 2009; 2012 sourceforge.
24. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JY, Zhang J. *Genome Biol*. 2004; 5:R80. [PubMed: 15461798]
25. Farazi TA, Horlings HM, Ten Hoeve JJ, Mihailovic A, Halfwerk H, Morozov P, Brown M, Hafner M, Reyat F, van Kouwenhove M, Kreike B, Sie D, Hovestadt V, Wessels LF, van de Vijver MJ, Tuschl T. *Cancer Res*. 2011; 71:4443–53. [PubMed: 21586611]
26. Trapnell C, Pachter L, Salzberg SL. *Bioinformatics*. 2009; 25:1105–11. [PubMed: 19289445]
27. Langmead B, Trapnell C, Pop M, Salzberg SL. *Genome Biol*. 2009; 10:R25. [PubMed: 19261174]
28. Li H, Durbin R. *Bioinformatics*. 2009; 25:1754–60. [PubMed: 19451168]
29. Liu CM, Wong T, Wu E, Luo R, Yiu SM, Li Y, Wang B, Yu C, Chu X, Zhao K, Li R, Lam TW. *Bioinformatics*. 2012; 28:878–9. [PubMed: 22285832]
30. Berninger P, Gaidatzis D, van Nimwegen E, Zavolan M. *Methods*. 2008; 44:13–21. [PubMed: 18158128]
31. San Lucas FA, Wang G, Scheet P, Peng B. *Bioinformatics*. 2011; 28:421–2. [PubMed: 22138362]
32. Quinlan AR, Hall IM. *Bioinformatics*. 26:841–2. [PubMed: 20110278]
33. Barnett DW, Garrison EK, Quinlan AR, Stromberg MP, Marth GT. *Bioinformatics*. 2011; 27:1691–2. [PubMed: 21493652]

34. Sievers C, Schlumpf T, Sawarkar R, Comoglio F, Paro R. *Nucleic Acids Res.* 2012
35. Pepke S, Wold B, Mortazavi A. *Nat Methods.* 2009; 6:S22–32. [PubMed: 19844228]
36. Hendrickson DG, Hogan DJ, McCullough HL, Myers JW, Herschlag D, Ferrell JE, Brown PO. *PLoS Biol.* 2009; 7:e1000238. [PubMed: 19901979]
37. Grimson A, Farh KK, Johnston WK, Garrett-Engele P, Lim LP, Bartel DP. *Mol Cell.* 2007; 27:91–105. [PubMed: 17612493]
38. Georgiev S, Boyle AP, Jayasurya K, Ding X, Mukherjee S, Ohler U. *Genome Biol.* 2010; 11:R19. [PubMed: 20156354]
39. Bailey TL, Williams N, Mischel C, Li WW. *Nucleic Acids Res.* 2006; 34:W369–73. [PubMed: 16845028]
40. Siddharthan R, Siggia ED, van Nimwegen E. *PLoS Comput Biol.* 2005; 1:e67. [PubMed: 16477324]
41. Betel D, Koppal A, Agius P, Sander C, Leslie C. *Genome Biol.* 2010; 11:R90. [PubMed: 20799968]
42. Garcia DM, Baek D, Shin C, Bell GW, Grimson A, Bartel DP. *Nat Struct Mol Biol.* 2011; 18:1139–46. [PubMed: 21909094]
43. Friedman RC, Farh KK, Burge CB, Bartel DP. *Genome Res.* 2009; 19:92–105. [PubMed: 18955434]
44. Konig J, Zarnack K, Rot G, Curk T, Kayikci M, Zupan B, Turner DJ, Luscombe NM, Ule J. *J Vis Exp.* 2011
45. Darnell RB. *Wiley Interdiscip Rev RNA.* 2011; 1:266–86. [PubMed: 21935890]
46. Lebedeva S, Jens M, Theil K, Schwanhauser B, Selbach M, Landthaler M, Rajewsky N. *Mol Cell.* 2011; 43:340–52. [PubMed: 21723171]
47. Mukherjee N, Corcoran DL, Nusbaum JD, Reid DW, Georgiev S, Hafner M, Ascano M Jr, Tuschl T, Ohler U, Keene JD. *Mol Cell.* 2011; 43:327–39. [PubMed: 21723170]
48. Granneman S, Kudla G, Petfalski E, Tollervey D. *Proc Natl Acad Sci U S A.* 2009; 106:9613–8. [PubMed: 19482942]
49. Castello A, Fischer B, Eichelbaum K, Horos R, Beckmann BM, Strein C, Davey NE, Humphreys DT, Preiss T, Steinmetz LM, Krijgsveld J, Hentze MW. *Cell.* 2012
50. Burroughs AM, Ando Y, de Hoon MJ, Tomaru Y, Suzuki H, Hayashizaki Y, Daub CO. *RNA Biol.* 2011; 8:158–77. [PubMed: 21282978]
51. Rosa A, Spagnoli FM, Brivanlou AH. *Dev Cell.* 2009; 16:517–27. [PubMed: 19386261]

Figure 1. Outline of the PAR-CLIP methodology for biochemical identification of AGO binding sites

PAR-CLIP requires incorporation of photoactivatable thioribonucleosides into nascent transcripts and then crosslinking by using long-wavelength 365 nm UV. Isolation of crosslinked RNAAGO complexes is achieved by immunoprecipitation, and the crosslinked AGO proteins are further purified by SDS-PAGE. After recovery of the RNA from the purified radioactive band the RNA is carried through a small RNA cDNA library preparation protocol for sequencing. Reverse transcription through a crosslinked-modified RNA, followed by PCR amplification, leads to a characteristic mutation (T-to-C when using 4SU and G-to-A when using 6SG) that is used to identify the AGO binding sites.

Figure 2. Base substitution frequency

The frequency of the twelve possible base substitutions identified in high-coverage regions color-coded according to chromosomal location. In all cases ‘T2C’ and the reverse complement ‘A2G’ are the most frequent substitutions indicating a strong enrichment for crosslinked sites. Not all crosslinked sites are derived from bona fide target sites as indicated by the unusually large number of T-to-C mutations corresponding to putative rRNA as a result of co-purification of fragments of non-crosslinked, but abundant rRNA.

(a)

(b)

Figure 3. Example of AGO2 binding sites and loaded miRNAs identified by PAR-CLIP

(a) A schematic diagram of the sequencing coverage of LEFTY1 3'UTR, a well known target of miR-302 in hESC (51), identified by PAR-CLIP experiment from Lipchina et al. (10). Other expressed miRNAs such as miR-130 and miR-144 are also predicted to target this site. (b) miRNAs identified by the PAR-CLIP experiment from Hafner et al. T-to-C mutations (or corresponding A-to-G) frequencies and positions are marked by the yellow shaded bars.

Genomic distribution of PAR-CLIP clusters

Figure 4. Genomic annotation of PAR-CLIP clusters for the three datasets

PAR-CLIP cluster densities indicate a strong enrichment of clusters in 3'UTR position as expected from miRNA target recognition. Cluster densities are calculated by normalizing the number of clusters by the length of the respective genomic region. A large fraction of the clusters overlapping non-coding regions are derived from miRNA transcripts (shaded portions of the bar plot).

Figure 5. Correlation of miRNA expression with the number of PAR-CLIP target sites

Scatter plots of the Kishore et al. and Lipchina et al. datasets (13, 14) indicate an overall correlation between the log expression of the top 100 miRNAs in HEK293 and hESC cells with the log number of target sites identified in 3'UTR PAR-CLIP clusters. Red dots and lines indicate the mean and standard deviations of the 5-quantiles (i.e. 5 equal-size subsets of the data). Hafner et al. data were omitted from this analysis due to low number of 3'UTR PAR-CLIP clusters.

Table 1

Summary of read alignment statistics for the three datasets. Preprocessing of the sequences results in some loss of sequences as well as high mutation rates, which reduced mapping yield. Number of reads are reported as millions of reads and fraction from total reads percentages. All alignments were performed using bowtie (version 0.12.5) with ‘-v 2 -m 10 -best -strata’

Dataset	Total number of reads	Number of reads after adaptor trimming	Number of aligned reads
Hafner <i>et al.</i>	8.71M	5.28M (61%)	1.76M (20%)
Kishore <i>et al.</i>	47.62M	41.76M (88%)	19.38M (41%)
Lipchina <i>et al.</i>	101.52M	91.66M (90%)	17.11M (17%)